An Introduction to Bayesian Statistics

An-Ting Jhuang and Christina Knudson

UnitedHealth Group R&D and University of St. Thomas

February 22, 2020

Overview

1. Bayesian Framework
   - Introduction
   - Bayesian vs frequentist
   - Bayes reasoning

2. Four examples in R
   - Coin flip example
   - Linear regression example
   - Logistic regression example
   - MCMC diagnostics

Preparation for Running R

- Make sure you install the following packages: `devtools`, `ggplot2`, `HDInterval`, `MCMCpack`, `mcmc`, `mcmcse`, `Rcpp`, `RcppArmadillo`, `stableGR`.
- Please download the R codes and R Markdown document at: `http://cknudson.com/Presentations/Rexamples_CSP.Rmd` and `http://cknudson.com/Presentations/Rexamples_CSP.pdf`

The Bayesian Framework

Figure 1: Thomas Bayes, 1701-1761

- Parameters are not fixed: they are random variables with distributions.

The Bayesian Framework

Figure 1: Thomas Bayes, 1701-1761

- Parameters are not fixed: they are random variables with distributions.
- Shed light on the distribution of the parameters by combining data and prior info (past experiments, similar experiments, professional opinions)

The Bayesian Framework

Figure 1: Thomas Bayes, 1701-1761

- Parameters are not fixed: they are random variables with distributions.
- Shed light on the distribution of the parameters by combining data and prior info (past experiments, similar experiments, professional opinions)
- If you collect more data, you can incorporate it to shed more light on the parameters' distribution.

Bayesian vs Frequentist

Table 1: Comparison between frequentist and Bayesian

Characteristic	Frequentist	Bayesian
Parameter	Unknown constant	Random variable

Bayesian vs Frequentist

Table 1: Comparison between frequentist and Bayesian

Characteristic	Frequentist	Bayesian
Parameter	Unknown constant	Random variable
Intervals	We are 95% confident the mean is between 0 and 3.	There is a 95% chance the mean is between 1 and 4.

Bayes Reasoning

A particular pregnancy test can detect 98% of pregnancies. A person takes this test, which says they're pregnant. What's the probability this person is actually pregnant?

Bayes Reasoning

A particular pregnancy test can detect 98% of pregnancies. A person takes this test, which says they're pregnant. What's the probability this person is actually pregnant?

Bayes' Theorem

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

Bayes Reasoning

A particular pregnancy test can detect 98% of pregnancies. A person takes this test, which says they're pregnant. What's the probability this person is actually pregnant?

Bayes' Theorem

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

Posterior Distribution

$$\begin{aligned} f(\text{parameter}|\text{data}) &= \frac{f(\text{data}|\text{parameter})f(\text{parameter})}{f(\text{data})} \\ &\propto f(\text{data}|\text{parameter})f(\text{parameter}) \end{aligned}$$

Terminology

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

$$f(\text{parameter}|\text{data}) \propto f(\text{data}|\text{parameter}) f(\text{parameter})$$

- **Prior**: the parameters' probability distribution prior to collecting data

Terminology

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

$$f(\text{parameter}|\text{data}) \propto f(\text{data}|\text{parameter}) f(\text{parameter})$$

- **Prior**: the parameters' probability distribution prior to collecting data
- **Likelihood**: the data's probability distribution given the parameters

Terminology

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

$$f(\text{parameter}|\text{data}) \propto f(\text{data}|\text{parameter}) f(\text{parameter})$$

- **Prior**: the parameters' probability distribution prior to collecting data
- **Likelihood**: the data's probability distribution given the parameters
- **Posterior**: the parameters' probability distribution after accounting for collected data (this combines prior information and the data)

Example 0: Coin Flip

We have a coin and wonder how many heads we'll get after several tosses. How do we solve this problem using Bayesian and frequentist methods?

- Notation: let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses.

Example 0: Coin Flip

We have a coin and wonder how many heads we'll get after several tosses. How do we solve this problem using Bayesian and frequentist methods?

- Notation: let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses.
- Statistical question: what's the estimate of θ?

Example 0: Coin Flip

We have a coin and wonder how many heads we'll get after several tosses. How do we solve this problem using Bayesian and frequentist methods?

- Notation: let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses.
- Statistical question: what's the estimate of θ?
- Frequentist: $\hat{\theta}_{\text{ML}} = y/n$

Example 0: Coin Flip

We have a coin and wonder how many heads we'll get after several tosses. How do we solve this problem using Bayesian and frequentist methods?

- Notation: let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses.
- Statistical question: 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

Example 0: Coin Flip

- Let's choose prior $\theta \sim \text{Beta}(\alpha, \beta)$ and assume the data are binomial ($y|\theta \sim \text{Bin}(n, \theta)$). Then the θ's distribution post-data collection is:

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

Example 0: Coin Flip

- Let's choose prior $\theta \sim \text{Beta}(\alpha, \beta)$ and assume the data are binomial ($y|\theta \sim \text{Bin}(n, \theta)$). Then the θ's distribution post-data collection is:

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

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

Example 0: Coin Flip

- Let's choose prior $\theta \sim \text{Beta}(\alpha, \beta)$ and assume the data are binomial ($y|\theta \sim \text{Bin}(n, \theta)$). Then the θ's distribution post-data collection is:

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

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

Example 0: Coin Flip

Let's choose $\alpha = \beta = 2$ (so that the prior mean is $\frac{1}{2}$).

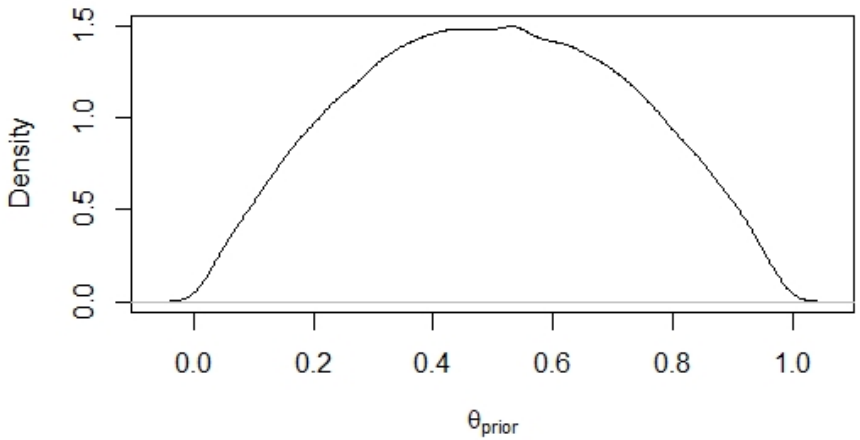


Figure 2: Prior density of θ

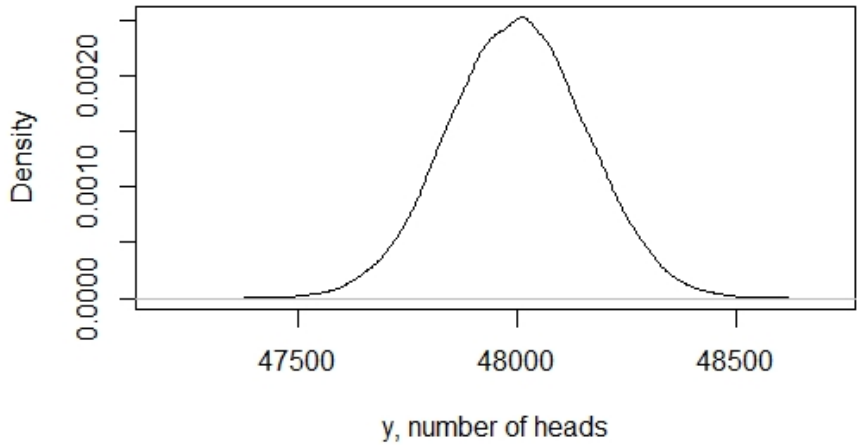
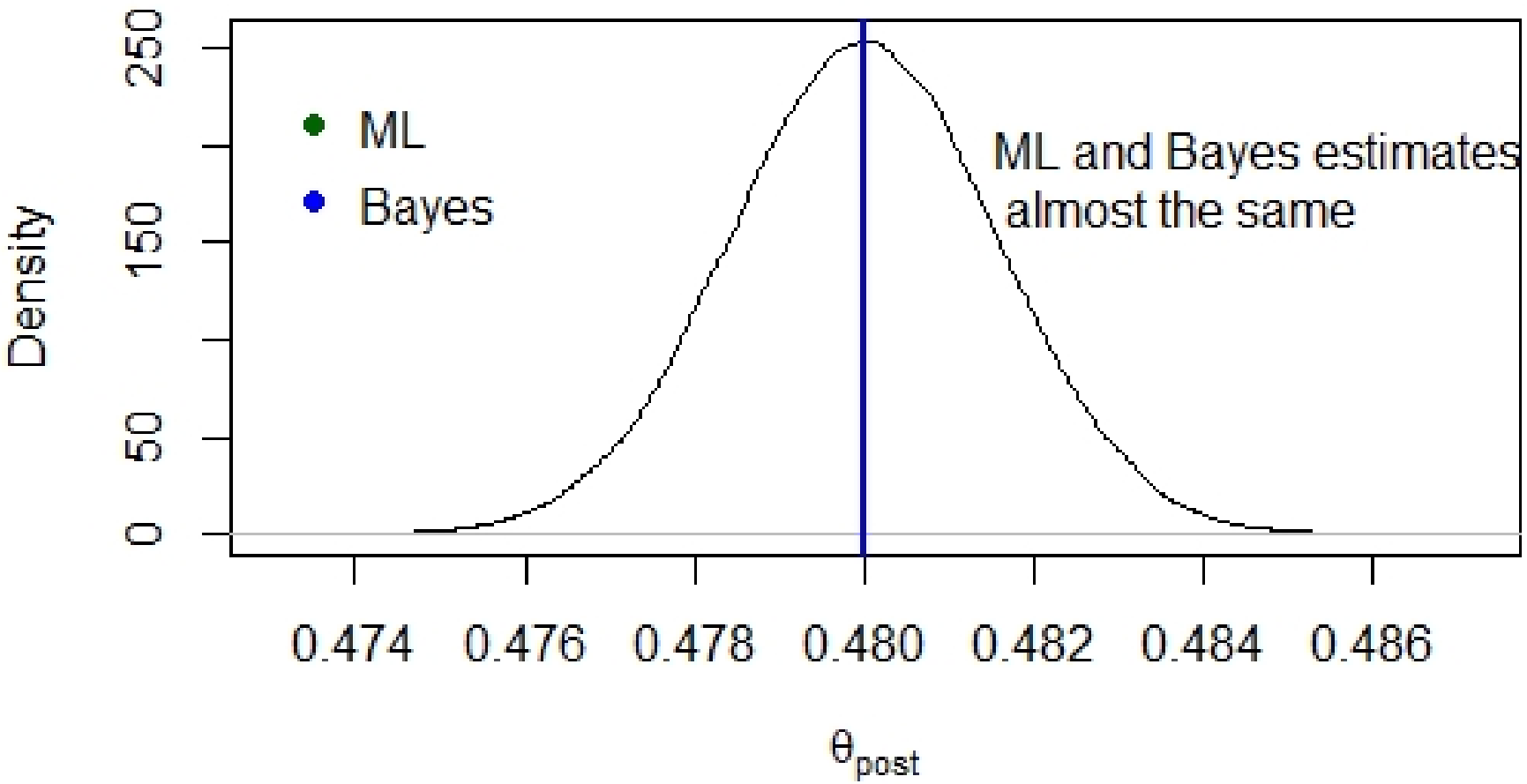


Figure 3: Likelihood of $y|\theta$

Example 0: Coin Flip

Then, using $\alpha = \beta = 2$,



Example 1: Linear Regression

- **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.

Example 1: Linear Regression

- **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 perceived temperature is linearly related to number of registered riders.

Example 1: Linear Regression

- **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 perceived temperature is linearly related to number of registered riders.

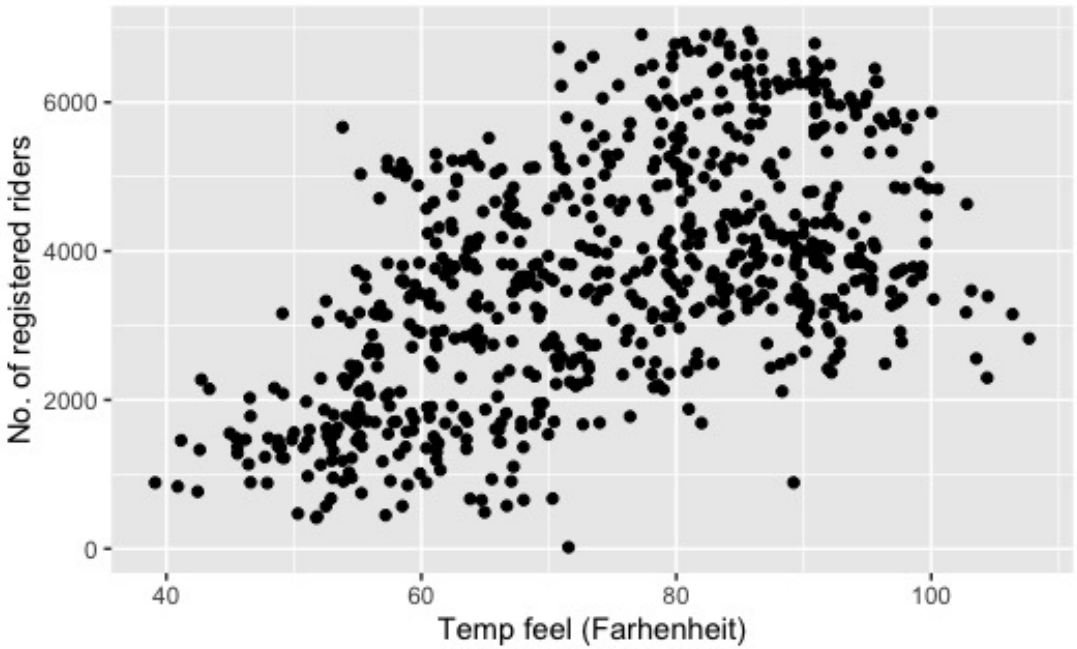


Figure 5: Scatter plot of perceived temperature and number of registered riders

Example 1: Linear Regression

- 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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Example 1: Linear Regression

- 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: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Example 1: Linear Regression

- 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: $\beta_0, \beta_1 \overset{indep}{\sim} N(\mu_{\beta_k}, \sigma^2_{\beta_k}), \sigma^2 \sim \text{InvGamma}(a, b)$

Example 1: Linear Regression

- 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: $\beta_0, \beta_1 \overset{indep}{\sim} N(\mu_{\beta_k}, \sigma^2_{\beta_k}), \sigma^2 \sim \text{InvGamma}(a, b)$
- Take $\mu_{\beta_k} = \hat{\beta}_{k,OLS}, \sigma^2 = 10^2, a = b = 0.05$.

Example 1: Linear Regression

- Fit the model using ordinary least square and Bayesian methods $\Rightarrow$ very close regression lines: $\hat{y} = -667.9 + 57.9x$.

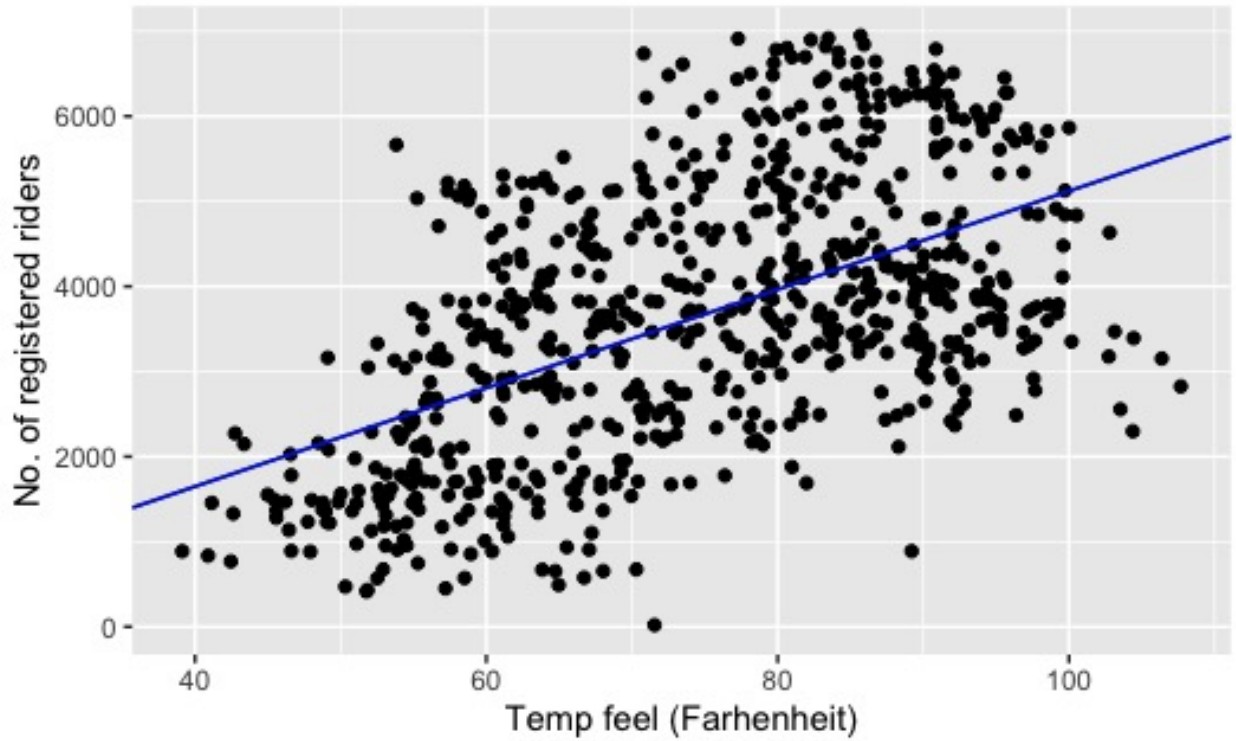


Figure 6: Regression lines and scatter plot of perceived temperature and number of registered riders

Example 1: Linear Regression

- Is perceived temperature significantly related to number of registered riders?

Example 1: Linear Regression

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

Example 1: Linear Regression

- Is perceived 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 7: Coefficient summary of ordinary least square estimates

Example 1: Linear Regression

```
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 8: Coefficient summary of Bayesian estimates

Example 1: Linear Regression

```
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 8: 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 9: Quantile summary of Bayesian estimates

Example 1: Linear Regression

- In frequentist method, a 95% confidence interval of β_1 is $\hat{\beta}_1 \pm t_{0.025,729}\text{SE}(\hat{\beta}_1) \approx (51.4, 64.4)$.

Example 1: Linear Regression

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

Example 1: Linear Regression

- In frequentist method, a 95% confidence interval of β_1 is $\hat{\beta}_1 \pm t_{0.025,729}\text{SE}(\hat{\beta}_1) \approx (51.4, 64.4)$.
- 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	$P(56.6 \leq \beta_1 \leq 59.2) = 0.95$

Example 1: Linear Regression

- In frequentist method, a 95% confidence interval of β_1 is $\hat{\beta}_1 \pm t_{0.025,729}\text{SE}(\hat{\beta}_1) \approx (51.4, 64.4)$.
- 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 perceived temperature has a significant association with number of registered riders.

Prior Distributions

What if we change priors of the parameters $\beta_0, \beta_1, \sigma^2$ in the Bayesian regression model?

Prior Distributions

What if we change priors of the parameters $\beta_0, \beta_1, \sigma^2$ in the Bayesian regression model?

Priors can be chosen in MANY ways, according to many different criteria.

Prior Distributions

What if we change priors of the parameters $\beta_0, \beta_1, \sigma^2$ in the Bayesian regression model?

Priors can be chosen in MANY ways, according to many different criteria.

In the previous example, we used "conjugate" priors (priors that combine nicely with the likelihood to produce a recognizable posterior distribution).

Prior Distributions

- By mathematical property,

Prior Distributions

- By mathematical property,
  - Conjugate prior: leads to a posterior from the same parametric family as the prior

Prior Distributions

- 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

Prior Distributions

- 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
- By reasoning,

Prior Distributions

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

Prior Distributions

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

Prior Distributions

- 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
- 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

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 (MCMC)
3. Assess the posterior approximation (e.g. sufficient samples)
4. Use the MCMC samples to conduct inference

You can either code it yourself or use a package from 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?

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:

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

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

At each step, a sampler can update

- A single parameter

  e.g. univariate Gibbs, variable-at-a-time Metropolis-Hastings

Approximating the Posterior via MCMC

At each step, a sampler can update

- A single parameter

  e.g. univariate Gibbs, variable-at-a-time Metropolis-Hastings

- All the parameters

  e.g. random walk Metropolis-Hastings

Approximating the Posterior via MCMC

At each step, a sampler can update

- A single parameter
  e.g. univariate Gibbs, variable-at-a-time Metropolis-Hastings
- All the parameters
  e.g. random walk Metropolis-Hastings
- Some of the parameters
  e.g. 2 of the 3 parameters

Approximating the Posterior via MCMC

- Some packages (such as `mcmc`) focus on the MCMC (independent of the model/context) and are therefore **more general**.

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**.

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.

Example 2: Logistic Regression with a Non-Conjugate Prior

- Use the simulated dataset **logit** in the `mcmc` package.

Example 2: Logistic Regression with a Non-Conjugate Prior

- Use the simulated dataset **logit** in the `mcmc` package.
- Fit a logistic regression of the response y on the predictor x_1.

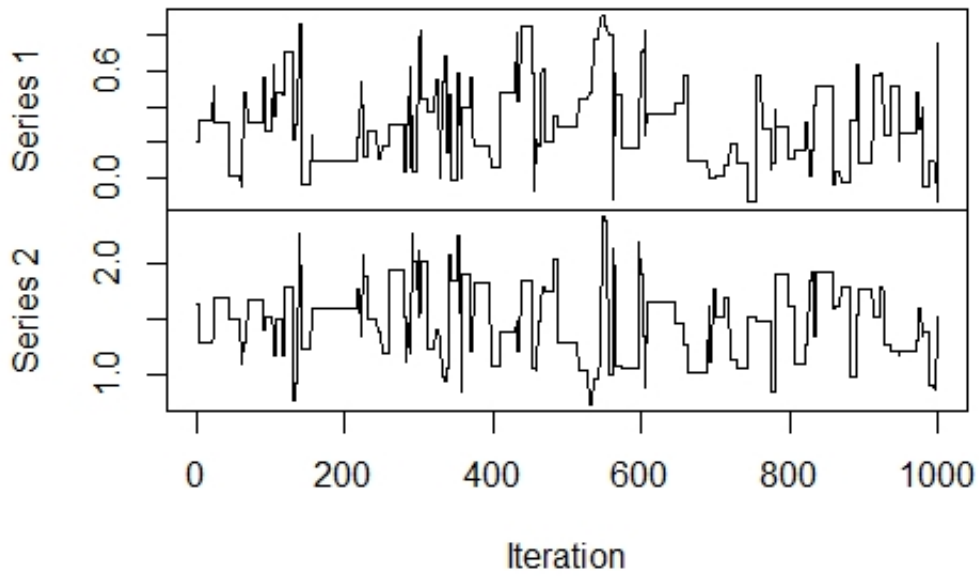


Figure 10: Trace plot of $\hat{\beta}_0$ and $\hat{\beta}_1$

Example 2: Logistic Regression with a Non-Conjugate Prior

- Let's play with another dataset **titanic.complete** in the `stableGR` package.

Example 2: Logistic Regression with a Non-Conjugate Prior

- Let's play with another dataset **titanic.complete** in the `stableGR` package.
- It's a Titanic passenger survival data with complete cases only. There are 8 variables including if a passenger survived, fare, age, gender and so on.

Example 2: Logistic Regression with a Non-Conjugate Prior

- Let's play with another dataset **titanic.complete** in the `stableGR` package.
- It's a Titanic passenger survival data with complete cases only. There are 8 variables including if a passenger survived, fare, age, gender and so on.
- **Objective**: examine if the fare passengers paid is linearly related to log odds of survival

Example 2: Logistic Regression with a Non-Conjugate Prior

- Let's play with another dataset **titanic.complete** in the `stableGR` package.
- It's a Titanic passenger survival data with complete cases only. There are 8 variables including if a passenger survived, fare, age, gender and so on.
- **Objective**: examine if the fare passengers paid is linearly related to log odds of survival
- Let's run it in R!

Assessing the Posterior Approximation

Did the MCMC sampler run long enough to create a suitably-detailed approximation of the posterior? Gelman-Rubin (1992):

$$\text{psrf} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{between-chain variance}}{\text{within-chain variance}}}$$

`psrf` decreases to 1 as chain length increases.

Assessing the Posterior Approximation

Did the MCMC sampler run long enough to create a suitably-detailed approximation of the posterior? Gelman-Rubin (1992):

$$\text{psrf} = \sqrt{\frac{\text{chain length} - 1}{\text{chain length}} + \frac{\text{between-chain variance}}{\text{within-chain variance}}}$$

`psrf` decreases to 1 as chain length increases.

```
> titanic.mod <- MCMClogit(Survived ~ Fare,
                           data = titanic.complete)
> stable.GR(titanic.mod)$psrf
[1] 1.000414 1.000401
```

Thanks to batch means variance estimation, `psrf` can be calculated whether we have one chain or multiple chains!

Assessing the Posterior Approximation

In a multivariate setting, it's better to check the multivariate psrf.

- Assesses joint convergence rather than component-wise.
- Like `psrf`, `mpsrf` decreases to 1 as chain length increases.

```
> stable.GR(titanic.mod)$mpsrf
[1] 1.00044
```

Assessing the Posterior Approximation

In a multivariate setting, it's better to check the multivariate psrf.

- Assesses joint convergence rather than component-wise.
- Like `psrf`, `mpsrf` decreases to 1 as chain length increases.

```
> stable.GR(titanic.mod)$mpsrf
[1] 1.00044
```

Is this low enough? Use `target.psrf` from `stableGR`.

```
> target.psrf(p=2, m=1)
$`psrf`
[1] 1.000066
```

Previously, the recommended threshold for (m)psrf was 1.1. We now know that this almost always results in premature termination of the MCMC sampler.

Assessing the Posterior Approximation

Equivalently, `n.eff` from `stableGR` checks whether sampler ran long enough.

```
> n.eff(titanic.mod)
$n.eff
[1] 1020.501
```

Effective sample size: number of uncorrelated samples that produce the same precision as the MCMC sample.

```
$converged
[1] FALSE
```

Did our sample achieve the target psrf?

```
$n.target
73778
```

If not, `n.target` approximates target Monte Carlo sample size to hit the target psrf.

Assessing the Posterior Approximation

We have to draw more posterior samples because the MCMC doesn't converge and `n.target` isn't achieved.

Assessing the Posterior Approximation

We have to draw more posterior samples because the MCMC doesn't converge and `n.target` isn't achieved.

```
newmod <- MCMClogit(Survived ~ Fare,
                    data = titanic.complete, mcmc=80000)
n.eff(newmod)
$n.eff
[1] 8825.746

$converged
[1] TRUE

$n.target
NULL
```

Using the MCMC Samples

Get basic model info (estimates and Monte Carlo standard errors):

```
> mcse.mat(titanic.mod)
                      est           se
(Intercept) -0.90241027 3.484705e-03
Fare         0.01607569 7.539261e-05
```

As the chain length increases, the Monte Carlo standard error decreases.

Monte Carlo SE measures the variability from run to run.

Using the MCMC Samples

It can be useful to understand the the covariance between the parameters:

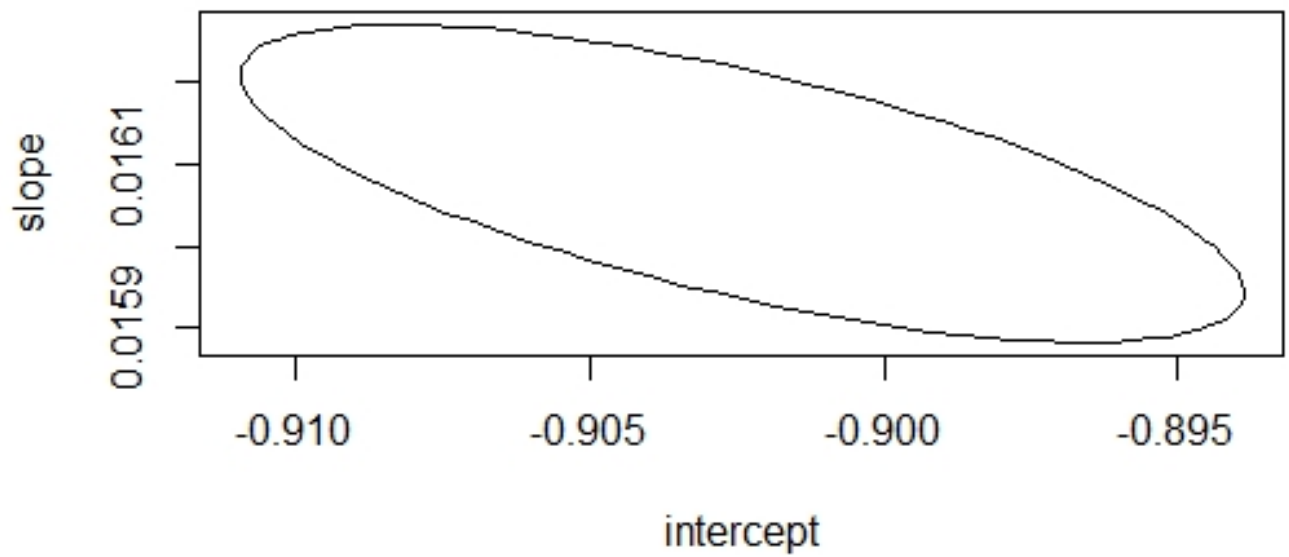
```
> mcse.multi(titanic.mod)
$cov
              [,1]           [,2]
[1,]   0.12143170 -1.878170e-03
[2,]  -0.00187817  6.176029e-05

$est
(Intercept)        Fare
-0.90241027 0.01607569
```

Using the MCMC Samples

Plotting joint confidence regions visualizes the dependence between a pair of parameters:

```
> mcerror <- mcse.multi(titanic.mod, blather = TRUE)
> plot(confRegion(mcerror, level = .95),
  xlab = "intercept", ylab = "slope", type = "l")
```

`confRegion` from package `mcmcse`



Using the MCMC Samples

Calculate quantiles and credible intervals using `quantile`:

```
> quantile(titanic.mod[,2], c(.025, .975))
      2.5%      97.5%
0.01149369 0.02106026
```

There is a significant linear relationship between how much passengers paid and their survival.

Using the MCMC Samples

Calculate quantiles and credible intervals using `quantile`:

```
> quantile(titanic.mod[,2], c(.025, .975))
      2.5%      97.5%
0.01149369 0.02106026
```

There is a significant linear relationship between how much passengers paid and their survival.

Or find the shortest (highest posterior density) credible intervals:

```
> hdi(titanic.mod)
      (Intercept)       Fare
lower -1.1204212 0.01137472
upper -0.7028561 0.02087360

attr(,"credMass")
[1] 0.95
```

Using the MCMC Samples

Monte Carlo standard errors for quantiles (`mcmcse`):

```
> mcse.q.mat(titanic.mod, method = "bm", q=.025)
                      est           se
(Intercept) -1.11859163 0.0081269853
Fare         0.01149369 0.0001210395
```

Wrapping Up

Basic Bayesian Steps

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

Wrapping Up

Basic Bayesian Steps

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

Some R Packages

- `MCMCpack` for making specific Bayesian models
- `mcmc` for general MCMC sampling
- `stableGR` for assessing the posterior approximation with Gelman-Rubin diagnostic (psrf) and effective sample size
- `mcmcse` for conducting multivariate analyses on the posterior
- `HDInterval` for finding credible intervals

Feel free to contact us at:

knud8583@stthomas.edu
An-TingJhuang@uhg.com

Enjoy the rest of your weekend!

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.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457-472.

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`

Christina P. Knudson and Dootika Vats (2019). stableGR: A Stable Gelman-Rubin Diagnostic for Markov Chain Monte Carlo. R package version 1.0. `https://CRAN.R-project.org/package=stableGR`.

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.

Vats, D., Flegal, J. M, and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321-337.

Vats, D. and Knudson, C. Revisiting the Gelman-Rubin diagnostic, arXiv:1812.09384 (under review).

Commands for Installing R Package stableGR from Github

```
  #get some required packages
install.packages("Rcpp")
install.packages("RcppArmadillo")
install.packages("devtools")

  #install mcmcse from github  (rather than CRAN)
library(devtools)
install_github("dvats/mcmcse")

  #install stableGR package
install_github("knudson1/stableGR/stableGR")
library(stableGR)
```

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,

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

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