Regression - The Linear Model

Manuel A. Rivas
Department of Biomedical Data Science
BDS215
Stanford University
rivaslab.stanford.edu

RIVASLAB

Review concepts learned

- Multilevel (Hierarchical) Modeling
- Mixture Models

Motivating examples for today

- Regression models

Linear model

Linear regression

Linear regression is one of the most widely used statstical tools.

In today's lecture we will focus on Bayesian model building and inference for normal linear models.

Linear model

Linear regression

Linear regression is one of the most widely used statstical tools.

In today's lecture we will focus on Bayesian model building and inference for normal linear models.

Learning objective: To set up the relevant Bayesian models and draw samples from posterior distributions for parameters θ and future observables $\tilde{y}$.

Motivation

- In many scientific studies concern relations among two or more observable quantities. A common question is: how does a quantity, y, vary as a function of another quantity or vector of quantities, x?

Motivation

- In many scientific studies concern relations among two or more observable quantities. A common question is: how does a quantity, y, vary as a function of another quantity or vector of quantities, x?

  We are interested in the conditional distribution of y, given x, denoted as $p\left(y|\theta, x\right)$.

Motivation

- In many scientific studies concern relations among two or more observable quantities. A common question is: how does a quantity, y, vary as a function of another quantity or vector of quantities, x?

  We are interested in the conditional distribution of y, given x, denoted as $p\left(y|\theta, x\right)$.

  - y is called the *response* or *outcome variable*

Motivation

- In many scientific studies concern relations among two or more observable quantities. A common question is: how does a quantity, y, vary as a function of another quantity or vector of quantities, x?

We are interested in the conditional distribution of y, given x, denoted as $p\left(y|\theta, x\right)$.

- y is called the *response* or *outcome variable*
- $x = (x_1, \ldots, x_k)$ are called explanatory variables

Motivation

- In many scientific studies concern relations among two or more observable quantities. A common question is: how does a quantity, y, vary as a function of another quantity or vector of quantities, x?

We are interested in the conditional distribution of y, given x, denoted as $p\left(y|\theta, x\right)$.

- y is called the *response* or *outcome variable*
- $x = (x_1, \ldots, x_k)$ are called explanatory variables

- Sometimes, a single variable x_j may be of primary interest and consider it the *treatment* variable, labeling the other components of x as the *control* variables

The Linear Model

Let's rewrite to better accommodate notation.

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).
- $\mathbb{X}$ is an $n \times k$ matrix of known coefficients

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).
- $\mathbb{X}$ is an $n \times k$ matrix of known coefficients
- β is a $k \times 1$ vector of parameters

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).
- $\mathbb{X}$ is an $n \times k$ matrix of known coefficients
- β is a $k \times 1$ vector of parameters
- ϵ isan $n \times 1$ vector of random errors

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).
- $\mathbb{X}$ is an $n \times k$ matrix of known coefficients
- β is a $k \times 1$ vector of parameters
- ϵ isan $n \times 1$ vector of random errors

The simplest and most widely used version of the linear model is the *normal linear model*

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).
- $\mathbb{X}$ is an $n \times k$ matrix of known coefficients
- β is a $k \times 1$ vector of parameters
- ϵ isan $n \times 1$ vector of random errors

The simplest and most widely used version of the linear model is the *normal linear model*

- The elements are assumed to have *zero* mean (critical and often overlooked in practice)

The Linear Model

Let's rewrite to better accommodate notation.

$$\mathbb{y} = \mathbb{X}\beta + \epsilon$$

- $\mathbb{y}$ is an $n \times 1$ vector of observations (n is the number of individuals in the study, for example).
- $\mathbb{X}$ is an $n \times k$ matrix of known coefficients
- β is a $k \times 1$ vector of parameters
- ϵ isan $n \times 1$ vector of random errors

The simplest and most widely used version of the linear model is the *normal linear model*

- The elements are assumed to have *zero* mean (critical and often overlooked in practice)
- Assumed to be uncorrelated and to have common variance σ^2, which becomes an additional parameter

The Linear Model

Question: What are the parameters?

The Linear Model

Question: What are the parameters?

Answer: $\theta = \left(\beta_1, \ldots, \beta_k, \sigma^2\right)$

Key statistical modeling issues in normal linear model framework

Key statistical modeling issues in normal linear model framework

1. Defining the variables x and y (possibly using transformations)

Key statistical modeling issues in normal linear model framework

1. Defining the variables x and y (possibly using transformations)

2. Setting up a prior distribution on the model parameters that accurately reflect substantive knowledge

Key statistical modeling issues in normal linear model framework

1. Defining the variables x and y (possibly using transformations)

2. Setting up a prior distribution on the model parameters that accurately reflect substantive knowledge

Statistical inference problem is to estimate the parameters θ, conditional on $\mathbb{X}$ and $\mathbb{y}$.

Notation for a basic normal linear model

Notation for a basic normal linear model

$$\mathbb{y}|\beta, \sigma^2, \mathbb{X} \sim \mathcal{N}\left(\mathbb{X}\beta, \sigma^2\mathbf{I}\right)$$

Notation for a basic normal linear model

$$\mathbb{y}|\beta, \sigma^2, \mathbb{X} \sim \mathcal{N}\left(\mathbb{X}\beta, \sigma^2\mathbf{I}\right)$$

Multivariate normal distribution where as usual $\mathbf{I}$ represents the identity matrix.

Notation for a basic normal linear model

$$\mathrm{y}|\beta, \sigma^2, \mathbb{X} \sim \mathcal{N}\left(\mathbb{X}\beta, \sigma^2\mathbf{I}\right)$$

Multivariate normal distribution where as usual $\mathbf{I}$ represents the identity matrix.

The likelihood becomes

$$f\left(\mathrm{y}|\beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{\frac{-n}{2}} exp\left[\left(\mathrm{y} - \mathbb{X}\beta\right)'\left(\mathrm{y} - \mathbb{X}\beta\right) / \left(2\sigma^2\right)\right]$$

Notation for a basic normal linear model

$$\mathrm{y}|\beta, \sigma^2, \mathbb{X} \sim \mathcal{N}\left(\mathbb{X}\beta, \sigma^2\mathbf{I}\right)$$

Multivariate normal distribution where as usual $\mathbf{I}$ represents the identity matrix.

The likelihood becomes

$$f\left(\mathrm{y}|\beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{\frac{-n}{2}} exp\left[\left(\mathrm{y} - \mathbb{X}\beta\right)'\left(\mathrm{y} - \mathbb{X}\beta\right) / \left(2\sigma^2\right)\right]$$

$\left(\mathrm{y} - \mathbb{X}\beta\right)'\left(\mathrm{y} - \mathbb{X}\beta\right)$ is referred to as the *quadratic form*

Least squares estimator of β

$$\hat{\beta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathrm{y}$$

Least squares estimator of β

$$\hat{\beta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathrm{y}$$

Residual sum of squares

$$S = \left(\mathrm{y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathrm{y} - \mathbb{X}\hat{\beta}\right)$$

The posterior distribution

We focused on examples of normal distributions with unknown mean and variances in previous lectures.

The posterior distribution

We focused on examples of normal distributions with unknown mean and variances in previous lectures.

Question: We are interested in the posterior distribution of...

The posterior distribution

$$\beta, \sigma^2$$

The posterior distribution

$$\beta, \sigma^2$$

We factor the join posterior distribution for β and σ^2 as

$$p\left(\beta, \sigma^2 | y\right) = p\left(\beta | \sigma^2, y\right) p\left(\sigma^2 | y\right)$$

The posterior distribution

Here, we are interested in determining first, the posterior distribution for β, conditioning on σ^2, and then the marginal posterior distribution for σ^2.

The conditional posterior distribution of β, given σ^2

$$\beta|\sigma^2, \mathbb{y} \sim \mathcal{N}\left(\hat{\beta}, \mathbf{V}_\beta \sigma^2\right),$$

Least squares estimator of β

$$\hat{\beta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathrm{y}$$

$\mathbf{V}_\beta$

$$\mathbf{V}_\beta = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}$$

Marginal posterior distribution of σ^2

$$\sigma^2|\mathrm{y} \sim IG\left(\frac{n-k}{2}, s^2\right),$$

Marginal posterior distribution of σ^2

$$\sigma^2|\mathrm{y} \sim IG\left(\frac{n-k}{2}, s^2\right),$$

What you usually do in practice is to draw inferences by simulation

1. Draw simulations of σ^2
2. Then, draw simulations of $\beta|\sigma^2$

non-Bayesian estimates of β and σ^2

- $\hat{\beta}$
- s^2

We obtain the classical standard error estimate for β by setting σ^2 to s^2

Case study : Trees

Plotting Outcome (Volume) as a function of Covariates (Girth, Height)

Source: http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-11/code-8/

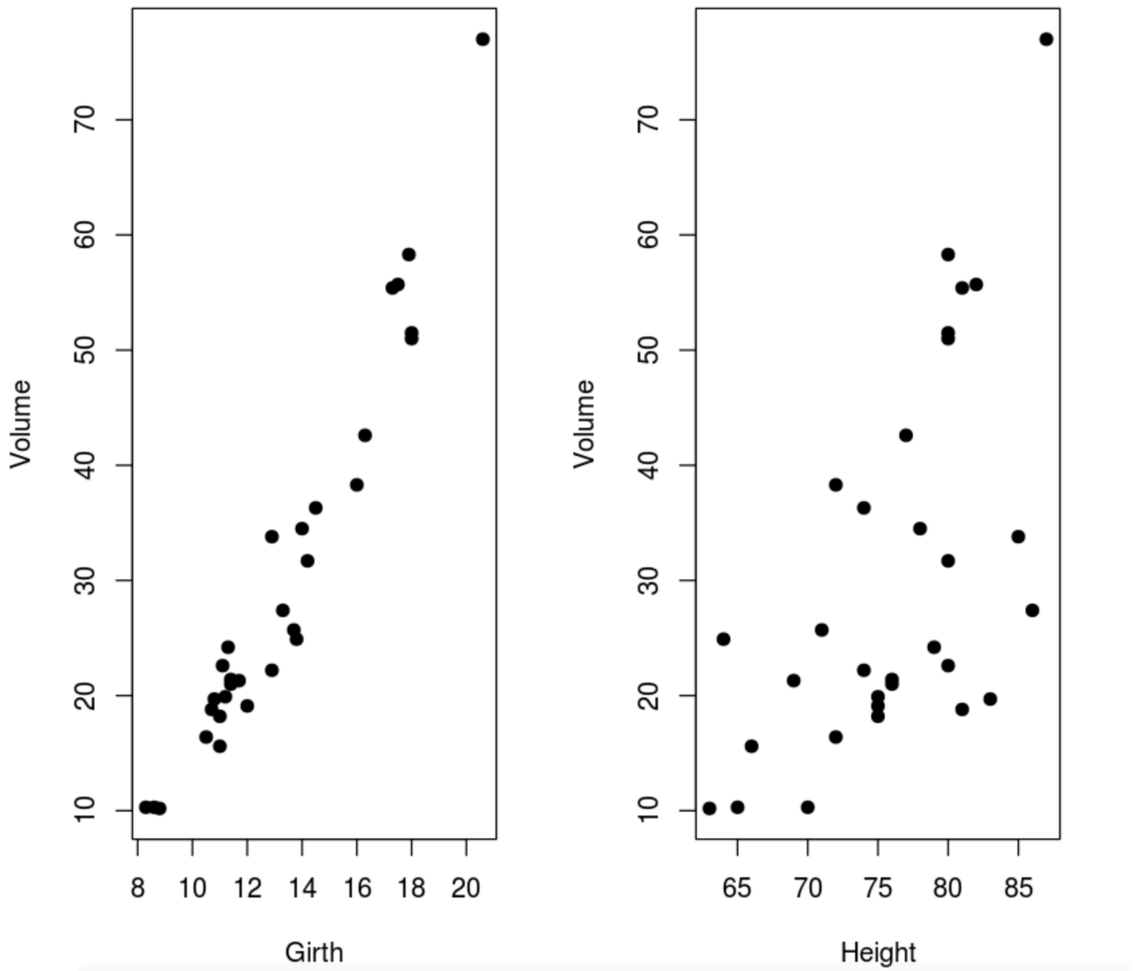
```
data(trees)
attach(trees)
dim(trees)
trees
```

Case study : Trees

31 3

Girth	Height	Volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2

Case study : Trees



Case study : Trees

```
## MLE fit of Regression Model
model <- lm(Volume~Girth+Height)
summary(model)
```

Case study : Trees

```
Call:
lm(formula = Volume ~ Girth + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

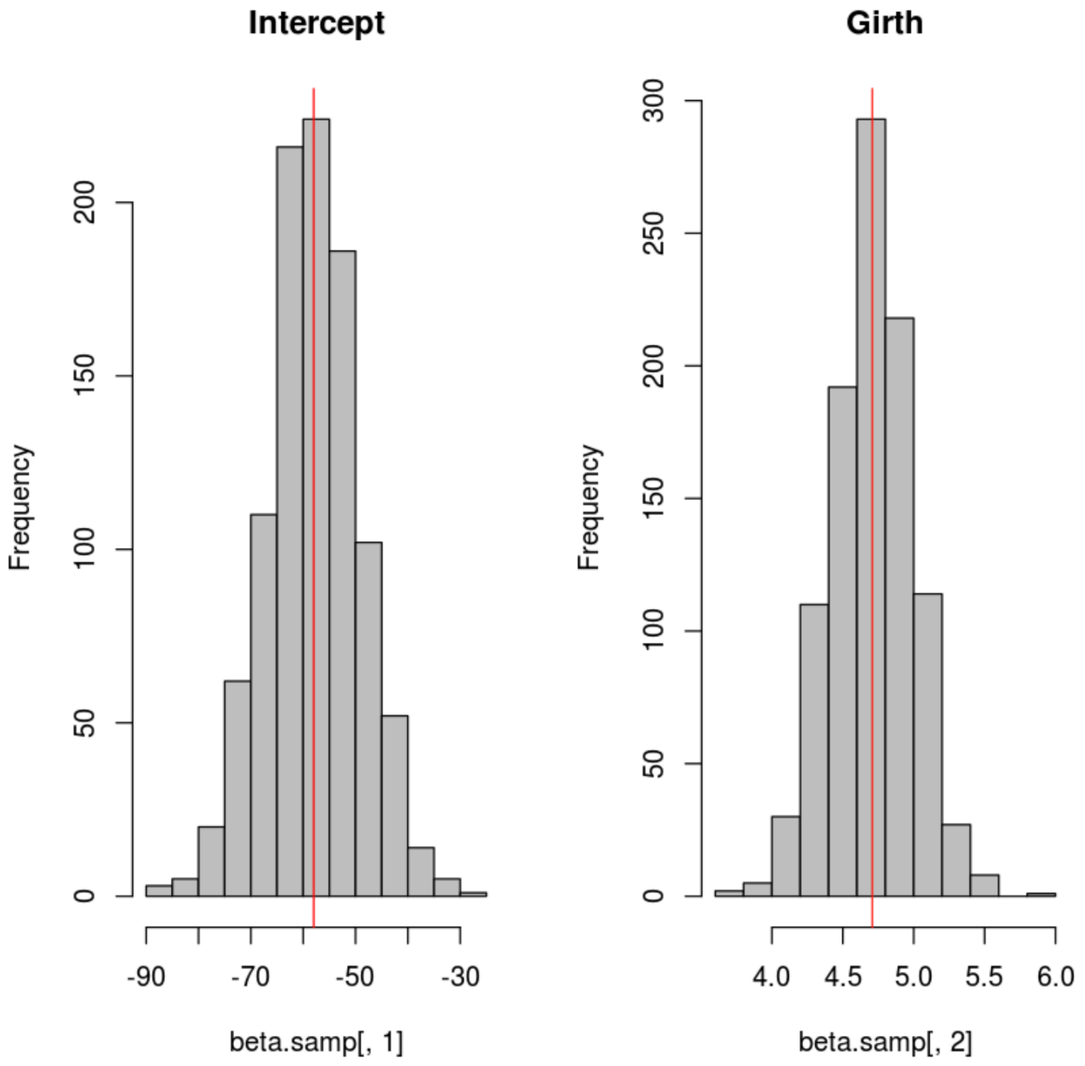
Sampling from posterior distribution of coefficients beta and variance sigsq

```
beta.hat <- model$coef
  n <- length(Volume)
  k <- length(beta.hat)
  s2 <- (n-k)*summary(model)$sigma^2
  V.beta <- summary(model)$cov.unscaled
```

Sampling from posterior distribution of coefficients beta and variance sigsq

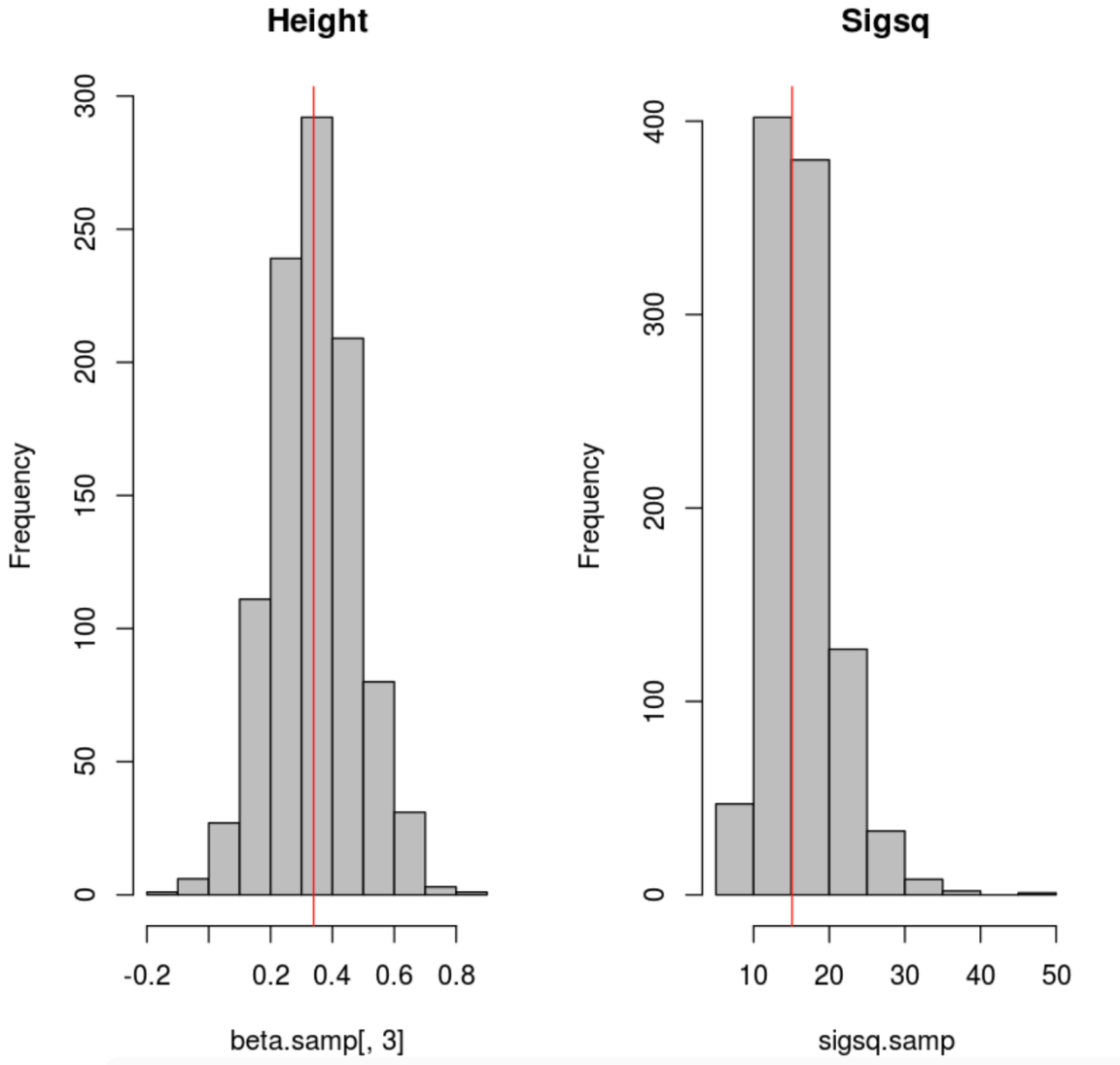
```
numsamp <- 1000
beta.samp <- matrix(NA,nrow=numsamp,ncol=k)
sigsq.samp <- rep(NA,numsamp)
for (i in 1:numsamp){
    temp <- rgamma(1,shape=(n-k)/2,rate=s2/2)
    cursigsq <- 1/temp # sampling from IG
    curvarbeta <- cursigsq*V.beta # sigma^2 *V_{beta}
    curvarbeta.chol <- t(chol(curvarbeta))
    z <- rnorm(k,0,1)
    curbeta <- beta.hat+curvarbeta.chol%*%z # sampling from normal
    sigsq.samp[i] <- cursigsq
    beta.samp[i,] <- curbeta
  }
```

Sampling from posterior distribution of coefficients beta and variance sigsq



RIVASLAB

Sampling from posterior distribution of coefficients beta and variance sigsq



RIVASLAB

Posterior predictive distribution for new data

Suppose we apply the regression model to a new set of data, for which we have observed the matrix $\tilde{\mathbb{X}}$ of explanatory variables, and we wish to predict the outcomes, $\tilde{\mathbb{y}}$.

- If β and σ^2 were known exactly, the vector $\tilde{\mathbb{y}}$ would have a normal distribution with mean $\tilde{\mathbb{X}}\beta$ and variance matrix $\sigma^2\mathbf{I}$.
- Instead, it is summarized by our posterior distribution.

Posterior predictive distribution for new data

Suppose we apply the regression model to a new set of data, for which we have observed the matrix $\tilde{\mathbb{X}}$ of explanatory variables, and we wish to predict the outcomes, $\tilde{\mathbb{y}}$.

- If β and σ^2 were known exactly, the vector $\tilde{\mathbb{y}}$ would have a normal distribution with mean $\tilde{\mathbb{X}}\beta$ and variance matrix $\sigma^2\mathbf{I}$.
- Instead, it is summarized by our posterior distribution.

To draw a random sample $\tilde{\mathbb{y}}$ from its posterior predictive distribution

- We first draw $\left(\beta, \sigma^2\right)$ from the joint posterior distribution
- Then, we draw $\tilde{\mathbb{y}} \sim \mathcal{N}\left(\tilde{\mathbb{X}}\beta, \sigma^2\mathbf{I}\right)$

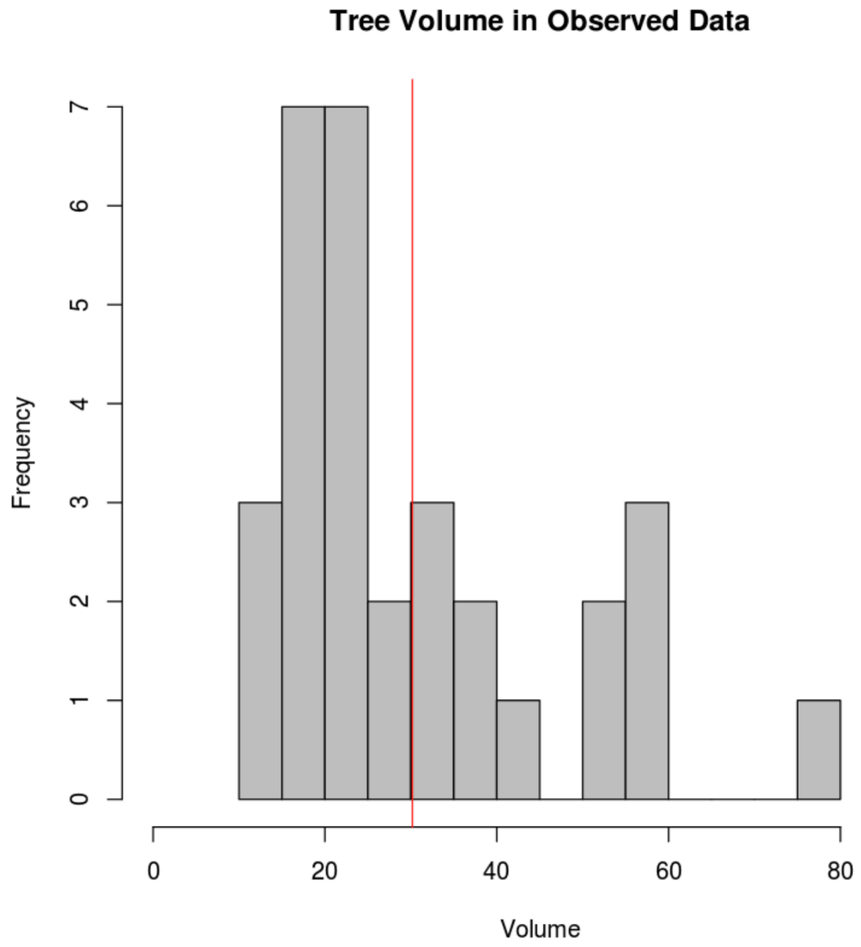
Posterior Predictive Sampling for new tree with girth = 18 and height = 80

```
Xstar <- c(1,18,80)  # new tree with girth = 18 and height = 80

  Ystar.samp <- rep(NA,numsamp)
  for (i in 1:numsamp){
     Ystar.hat <- sum(beta.samp[i,]*Xstar)
     Ystar.samp[i] <- rnorm(1,mean=Ystar.hat,sd=sqrt(sigsq.samp[i]))
  }
```

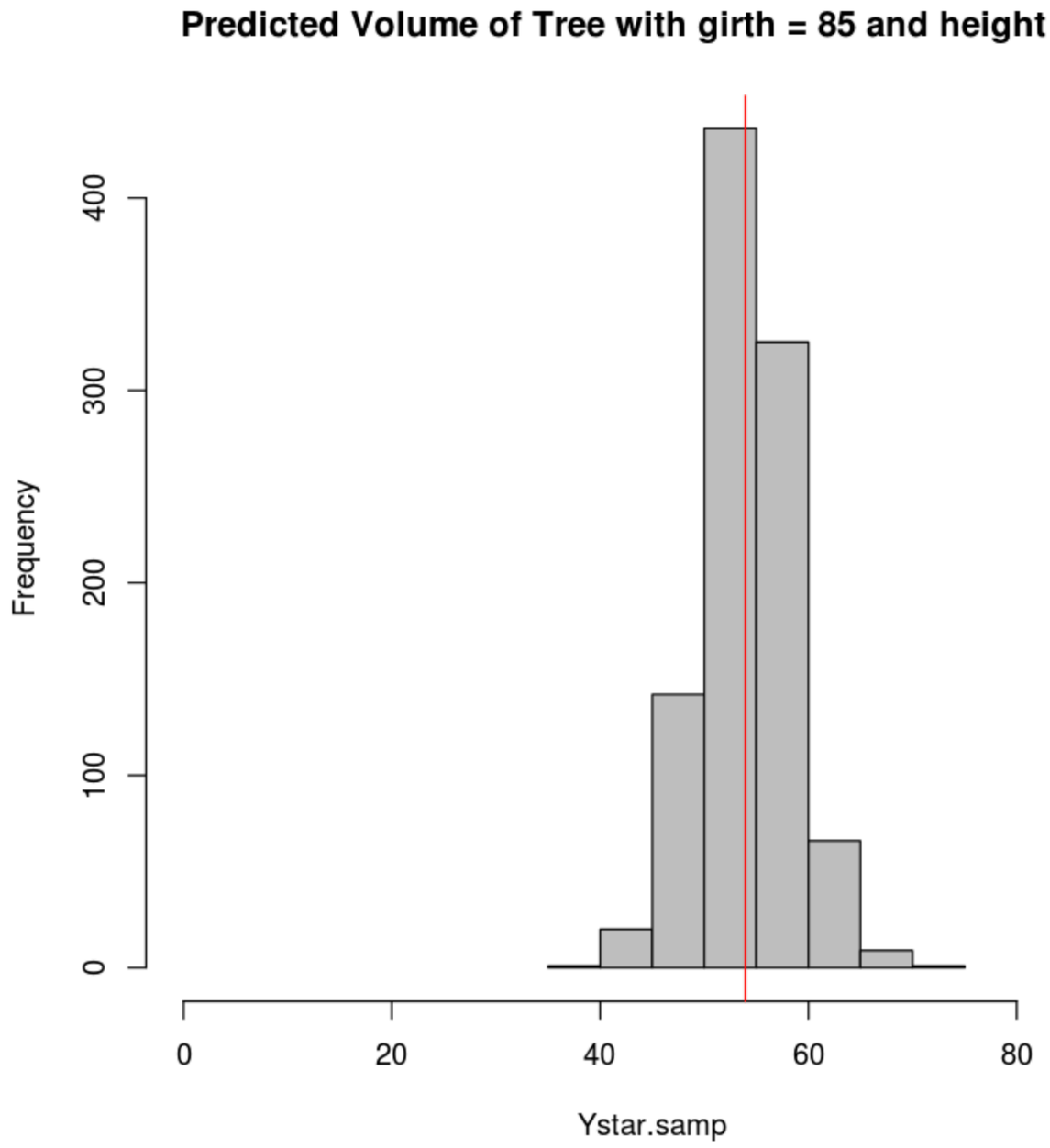
Posterior Predictive Sampling for new tree with girth = 18 and height = 80

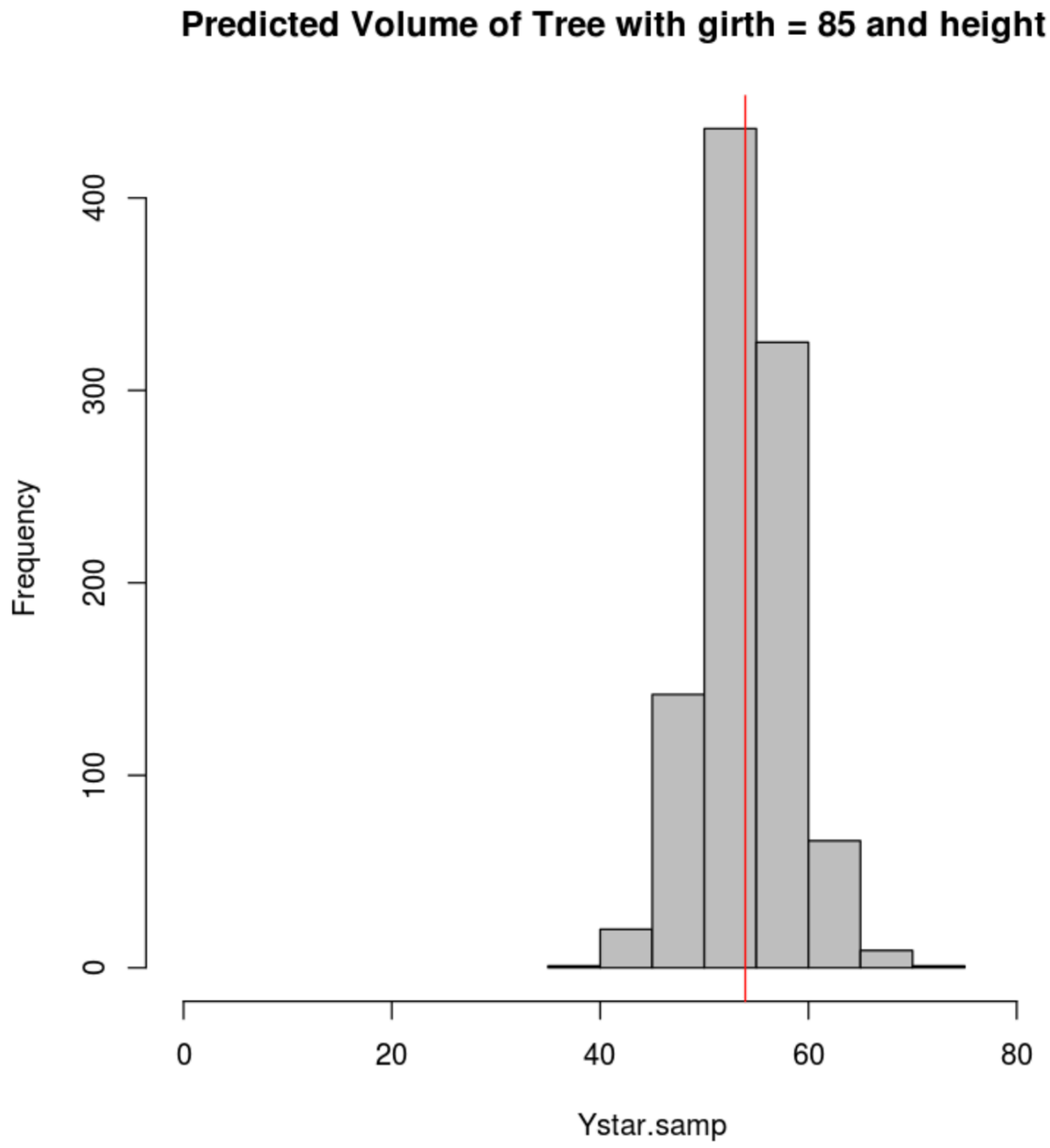
```
Xstar <- c(1,18,

  Ystar.samp <-
  for (i in 1:nu
     Ystar.hat <
     Ystar.samp[
  }
```

```
i]))
```



RIVASLAB


Predicted Volume of Tree with girth = 85 and height = 80
Frequency
0
100
200
300
400
0
20
40
60
80
Ystar.samp


Predicted Volume of Tree with girth = 85 and height = 80
Frequency
0
100
200
300
400
0
20
40
60
80
Ystar.samp