Pengantar Statistika Bayes - STA1312

Metode Komputasi Bayesian - Part 2

(Contoh Penerapan Komputasi Bayesian pada Regresi Linear)

Dr. Kusman Sadik, S.Si, M.Si
Program Studi Statistika dan Sains Data IPB

Tahun Akademik 2023/2024

Regresi Linear dalam Bayesian

- Perhatikan persamaan regresi linear sederhana berikut ini:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Asumsi yang digunakan adalah:

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{dan} \quad y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

- Metode Bayesian digunakan untuk membuat inferensi mengenai parameter model, yaitu β_0 dan β_1.

- Sehingga diperlukan sebaran prior bagi β_0 dan β_1.

- Selanjutnya inferensi didasarkan pada sebaran posterior bagi β_0 dan β_1.

Regresi Linear dalam Bayesian (*cont*.)

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Bayes' rule (parameter estimation version) tells us how to calculate the posterior distribution:

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

This is the generic form for parameters θ and data x. In our particular case, the unknown parameters are β_0 and β_1, and the data are the y values of the data points. The data also consist of a number N of points and the x-values, but we shall assume that these on their own provide no information about the slope and intercept (it would be a bit strange if they did). So the x-values and the number of points N act like prior information that lurks "in the background" of this entire analysis. The y-values are our data in the sense that we will obtain our likelihood by writing down a probability distribution for the y-values given the parameters.

Sebaran Prior bagi β_0 dan β_1

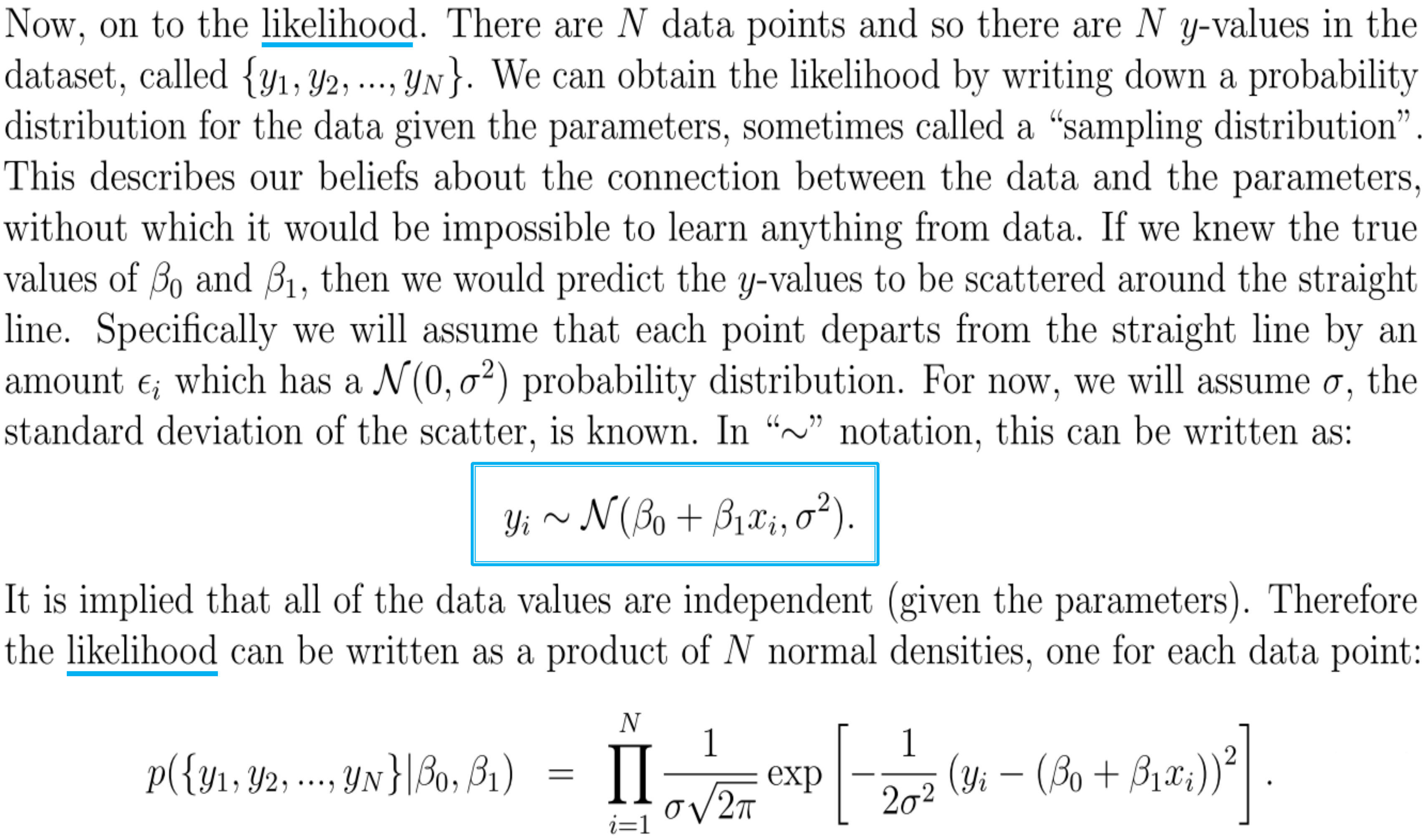
Therefore, Bayes' rule *for this problem* (i.e. with the actual names of our parameters and data, rather than generic names) reads:

$$\boxed{p(\beta_0, \beta_1 | y_1, y_2, ..., y_N) \propto p(\beta_0, \beta_1) p(y_1, y_2, ..., y_N | \beta_0, \beta_1)}$$

We can now say some things about Bayesian linear regression by working analytically. For starters, let's assume uniform priors for both β_0 and β_1, and that the prior for these two parameters are independent. The probability density for a uniform prior distribution can be written simply as:

$$p(\beta_0, \beta_1) \propto 1.$$

Fungsi Likelihood



Now, on to the likelihood. There are N data points and so there are N y-values in the dataset, called $\{y_1, y_2, ..., y_N\}$. We can obtain the likelihood by writing down a probability distribution for the data given the parameters, sometimes called a "sampling distribution". This describes our beliefs about the connection between the data and the parameters, without which it would be impossible to learn anything from data. If we knew the true values of β_0 and β_1, then we would predict the y-values to be scattered around the straight line. Specifically we will assume that each point departs from the straight line by an amount ϵ_i which has a $\mathcal{N}(0, \sigma^2)$ probability distribution. For now, we will assume σ, the standard deviation of the scatter, is known. In "$\sim$" notation, this can be written as:

$$\boxed{y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).}$$

It is implied that all of the data values are independent (given the parameters). Therefore the likelihood can be written as a product of N normal densities, one for each data point:

$$p(\{y_1, y_2, ..., y_N\} | \beta_0, \beta_1) \quad = \quad \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right].$$

Sebaran Posterior bagi $\boldsymbol{\beta_0}$ dan $\boldsymbol{\beta_1}$

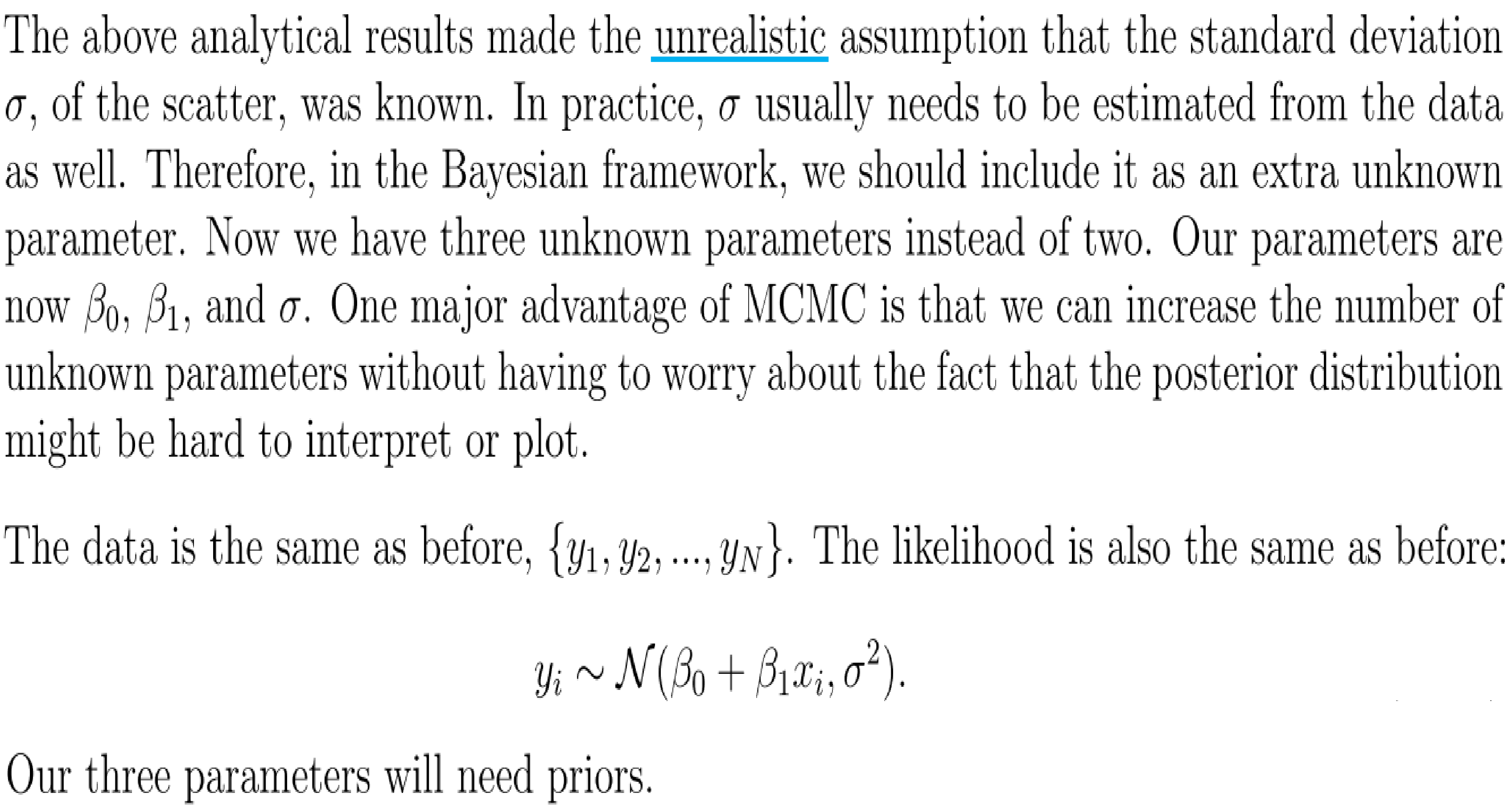
$$p(\{y_1, y_2, ..., y_N\}|\beta_0, \beta_1) \;=\; \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right].$$

Remember, when we combine the likelihood with the prior using Bayes' rule, we can usually ignore any constant factors which do not depend on the parameters. This allows us to ignore the first part of the product, outside the exponential (since we are assuming σ is known).

$$\begin{aligned}
p(\beta_0, \beta_1|y_1, y_2, ..., y_N) &\propto p(\beta_0, \beta_1)p(y_1, y_2, ..., y_N|\beta_0, \beta_1) \\
&\propto 1 \times \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right].
\end{aligned}$$

We have just found the expression for the posterior distribution for β_0 and β_1. This is a distribution for two parameters (i.e. it is bivariate).

Penggunaan MCMC



The above analytical results made the unrealistic assumption that the standard deviation σ, of the scatter, was known. In practice, σ usually needs to be estimated from the data as well. Therefore, in the Bayesian framework, we should include it as an extra unknown parameter. Now we have three unknown parameters instead of two. Our parameters are now β_0, β_1, and σ. One major advantage of MCMC is that we can increase the number of unknown parameters without having to worry about the fact that the posterior distribution might be hard to interpret or plot.

The data is the same as before, $\{y_1, y_2, ..., y_N\}$. The likelihood is also the same as before:

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

Our three parameters will need priors.

Pengujian Hipotesis bagi β_0 dan β_1

- Pengujian hipotesis bagi β_0 dan β_1 dapat menggunakan pendekatan sebaran ***t-student***, yaitu:

$$\frac{\hat{\beta}_j - \beta_j}{st.dev(\hat{\beta}_j)} \sim t_{(n-2)}$$

dengan *j* = 0 atau 1.

- $\hat{\beta}_j$ merupakan penduga Bayes (***posterior mean***) sedangkan $st.dev(\hat{\beta}_j)$ merupakan galat baku bagi $\hat{\beta}_j$ (***posterior standard deviation***).
- Nilai $\hat{\beta}_j$dan $st.dev(\hat{\beta}_j)$ dapat diperoleh melalui metode **MCMC**.
- ***Credible-interval*** (1 – α)100% juga dapat menggunakan sebaran *t-student* tersebut, yaitu:

$$\hat{\beta}_j \pm \left(t_{(\alpha/2;\ n-2)}\right)\left(st.dev(\hat{\beta}_j)\right)$$

Implementasi dalam Program R

Studi Kasus 1:

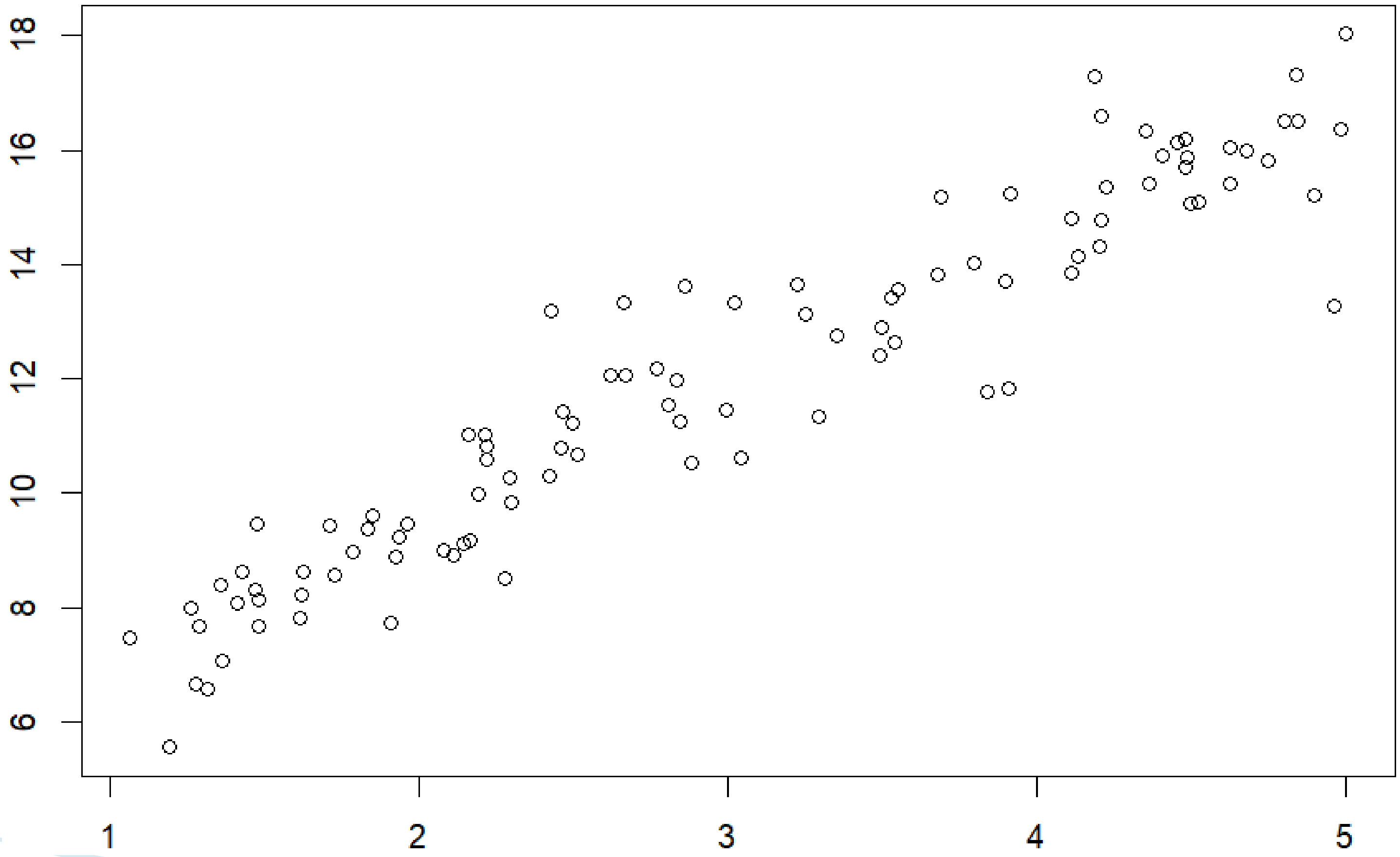
- Misalkan model yang digunakan $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, dengan $\varepsilon_i \sim N(0, \sigma^2)$.
- Data y_i dibangkitkan dengan x_i ~ Uniform(1, 5), $\beta_0 = 4.5$, $\beta_1 = 2.5$, dan $\varepsilon_i \sim N(0,1)$ serta banyak data *n* = 100.
- Sebaran prior yang digunakan bagi β_0 dan β_1 adalah "*flat*" yakni Uniform(0,1).
- Implementasi MCMC Bayesian pada Program R menggunakan *function* **`bayes.lin.reg()`** dalam *package* "**`Bolstad`**".
- Berikan penjelasan terkait inferensi Bayes dari *output* **`bayes.lin.reg()`** tersebut.
- Lakukan pengujian hipotesis untuk H_0: β_1 = 0 vs H_1: β_1 ≠ 0 pada taraf uji α = 0.05.
- Berdasarkan model yang diperoleh tentukan nilai prediksi bagi *y* apabila diketahui nilai *x* adalah 4, 1, dan 2.

Studi Kasus 1: *(cont.)*

```
> library(Bolstad)
>
> ## adjust plot margins
> par(mar = c(2, 2, 2, 2))
>
> ## Pembangkitan data x dan y
> set.seed(1312)
> x = runif(100, min=1, max=5)
> y = 4.5 + 2.5*x + rnorm(100)
>
> data.frame(x,y)
      x    y
1   2.3 10.3
2   2.2  9.2
3   5.0 18.0
.
.
99  2.8 12.2
100 1.3  7.7
```

Studi Kasus 1: *(cont.)*

```
> plot(x,y,xlab ="Nilai X",ylab="Nilai Y")
```



Studi Kasus 1: *(cont.)*

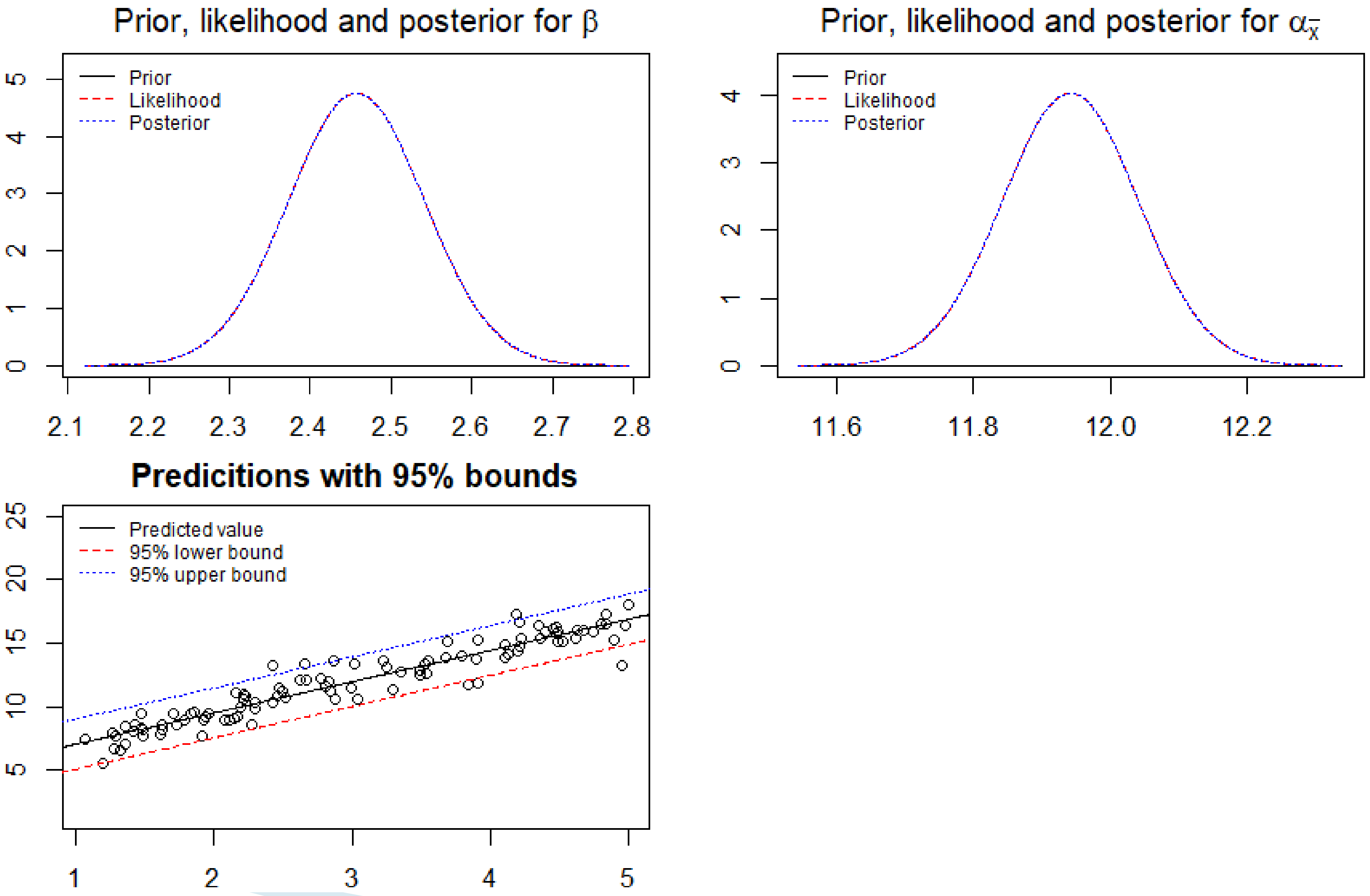
```
> ## Menggunakan prior flat untuk slope.prior (beta1)
> ## Menggunakan prior flat untuk intcpt.prior (beta0)
>
> bayes.lin.reg(y,x, slope.prior = "flat",
                intcpt.prior = "flat",
                alpha = 0.05, plot.data = TRUE)

Standard deviation of residuals:  0.99

            Posterior Mean Posterior Std. Deviation
            -------------- ------------------------
Intercept:  11.94          0.099036
Slope:       2.46          0.084019
```

Studi Kasus 1: *(cont.)*

```
            Posterior Mean Posterior Std. Deviation
            -------------- ------------------------
Intercept:  11.94          0.099036
Slope:       2.46          0.084019
```

Lakukan pengujian hipotesis untuk H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$ pada taraf uji $\alpha = 0.05$.

Studi Kasus 1: *(cont.)*



Studi Kasus 1: *(cont.)*

```
> ## Memprediksi y berdasarkan data x yang baru: x = 4, 1, 2
>
> bayes.lin.reg(y,x, slope.prior = "flat",
                intcpt.prior = "flat",
                pred.x=c(4, 1, 2))

x        Predicted y    SE
------   -----------   -----------
4          14.41        0.9989
1           7.04        1.0093
2           9.50        0.9988
```

Studi Kasus 2:

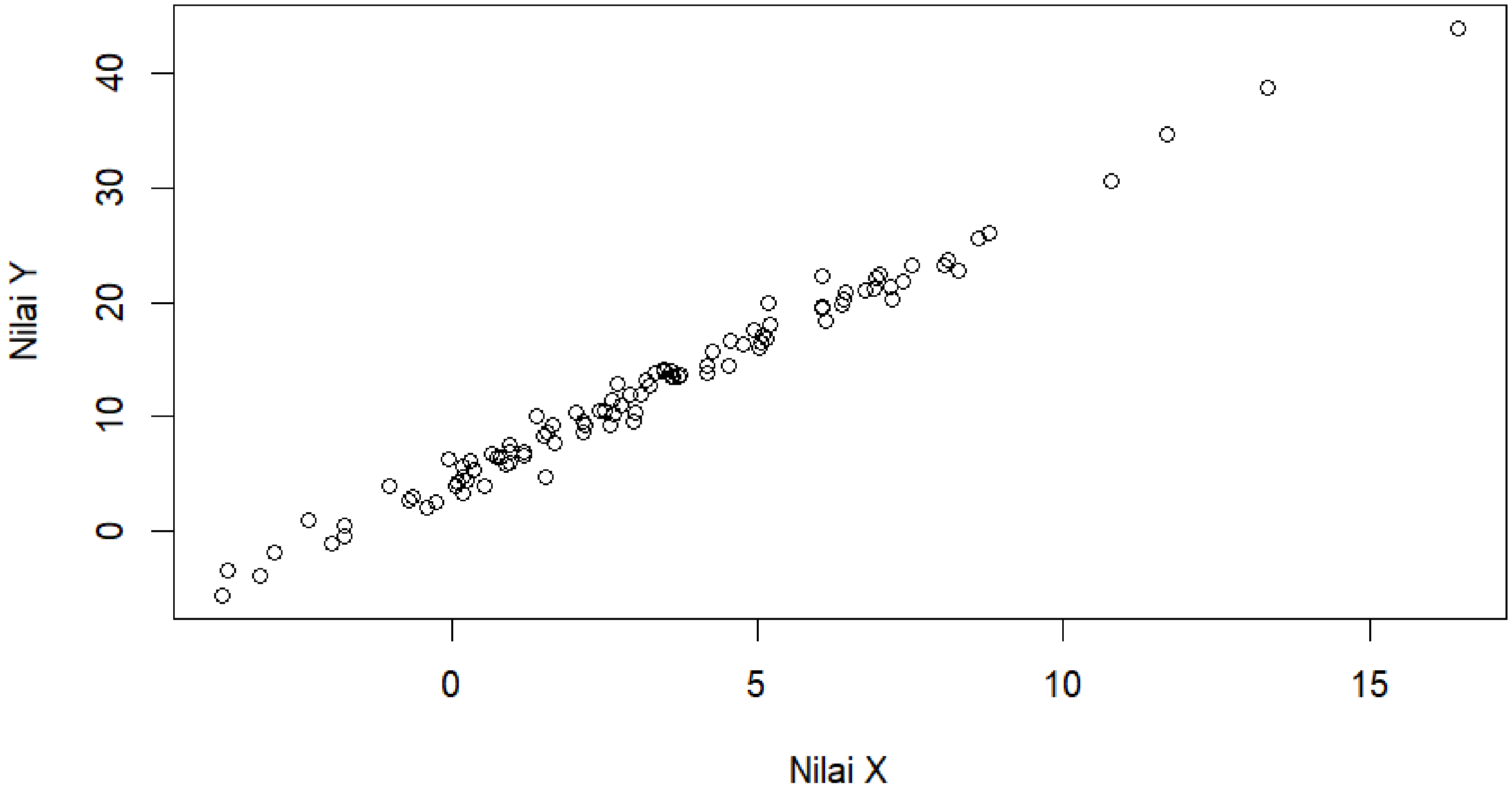
- Misalkan model yang digunakan $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, dengan $\varepsilon_i \sim N(0, \sigma^2)$.
- Data y_i dibangkitkan yang mana x_i ~ Normal(μ = 3, σ = 4), $\beta_0 = 4.5$, $\beta_1 = 2.5$, dan $\varepsilon_i \sim N(0, 1)$.
- Sebaran prior yang digunakan bagi β_0 adalah Normal(μ = 0, σ = 2) dan bagi β_1 adalah Normal(μ = 1, σ = 3).
- Implementasi MCMC Bayesian pada Program R menggunakan *function* `bayes.lin.reg()` dalam *package* "`Bolstad`".
- Berikan penjelasan terkait inferensi Bayes dari *output* `bayes.lin.reg()` tersebut.
- Berdasarkan model yang diperoleh tentukan nilai prediksi bagi *y* apabila diketahui nilai *x* adalah 3, -1, dan 6.

Studi Kasus 2: *(cont.)*

```
> library(Bolstad)
>
> ## adjust plot margins
> par(mar = c(2, 2, 2, 2))
>
> ## Pembangkitan data x dan y
> set.seed(1312)
> x = rnorm(100, mean = 3, sd = 4)
> y = 4.5 + 2.5*x + rnorm(100)
>
> data.frame(x,y)
        x       y
1     1.176   6.68
2    16.438  44.05
3     2.707  12.84
.
.
99    2.186   9.35
100   8.796  26.03
```

Studi Kasus 2: *(cont.)*

```
> plot(x,y,xlab ="Nilai X",ylab="Nilai Y")
```

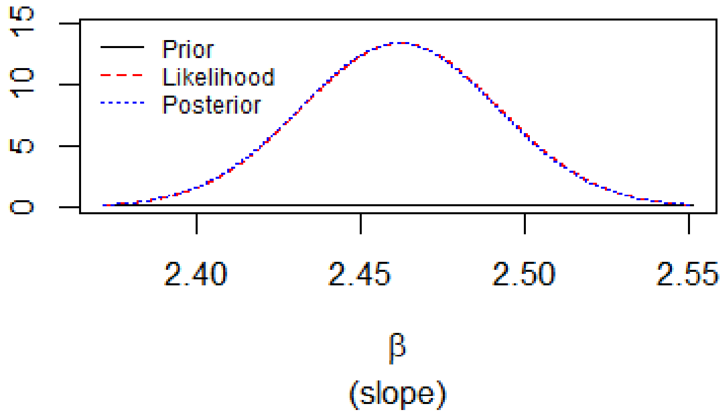


Studi Kasus 2: *(cont.)*

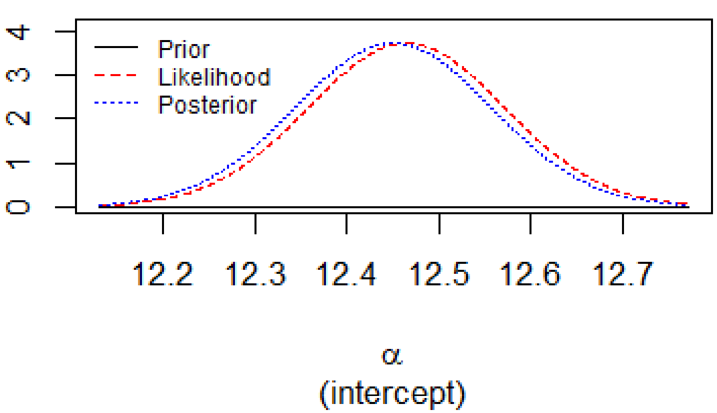
```
> ## Menggunakan slope.prior Normal(mean=0, sd=2)
> ## Menggunakan intcpt.prior Normal(mean=1,sd=3)>

> bayes.lin.reg(y,x, slope.prior = "normal",
                intcpt.prior = "normal",
                0, 2, 1, 3, alpha = 0.05,
                plot.data = TRUE)

Standard deviation of residuals:  1.07


            Posterior Mean Posterior Std. Deviation
            -------------- ------------------------
Intercept:  12.45          0.10690
Slope:       2.46          0.02995
```

Studi Kasus 2: *(cont.)*

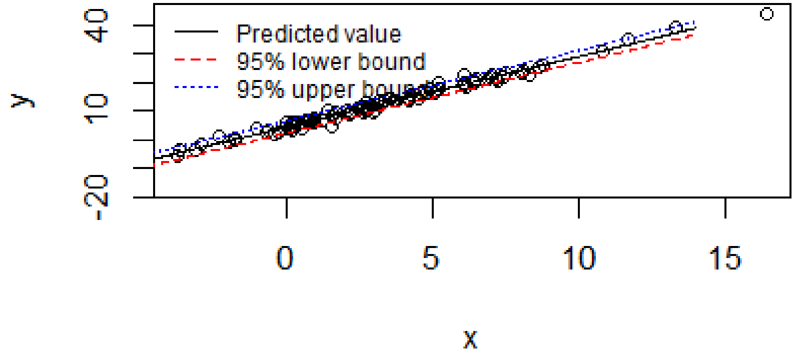
Prior, likelihood and posterior for β



Prior, likelihood and posterior for $\alpha_{\bar{x}}$



Predicitions with 95% bounds

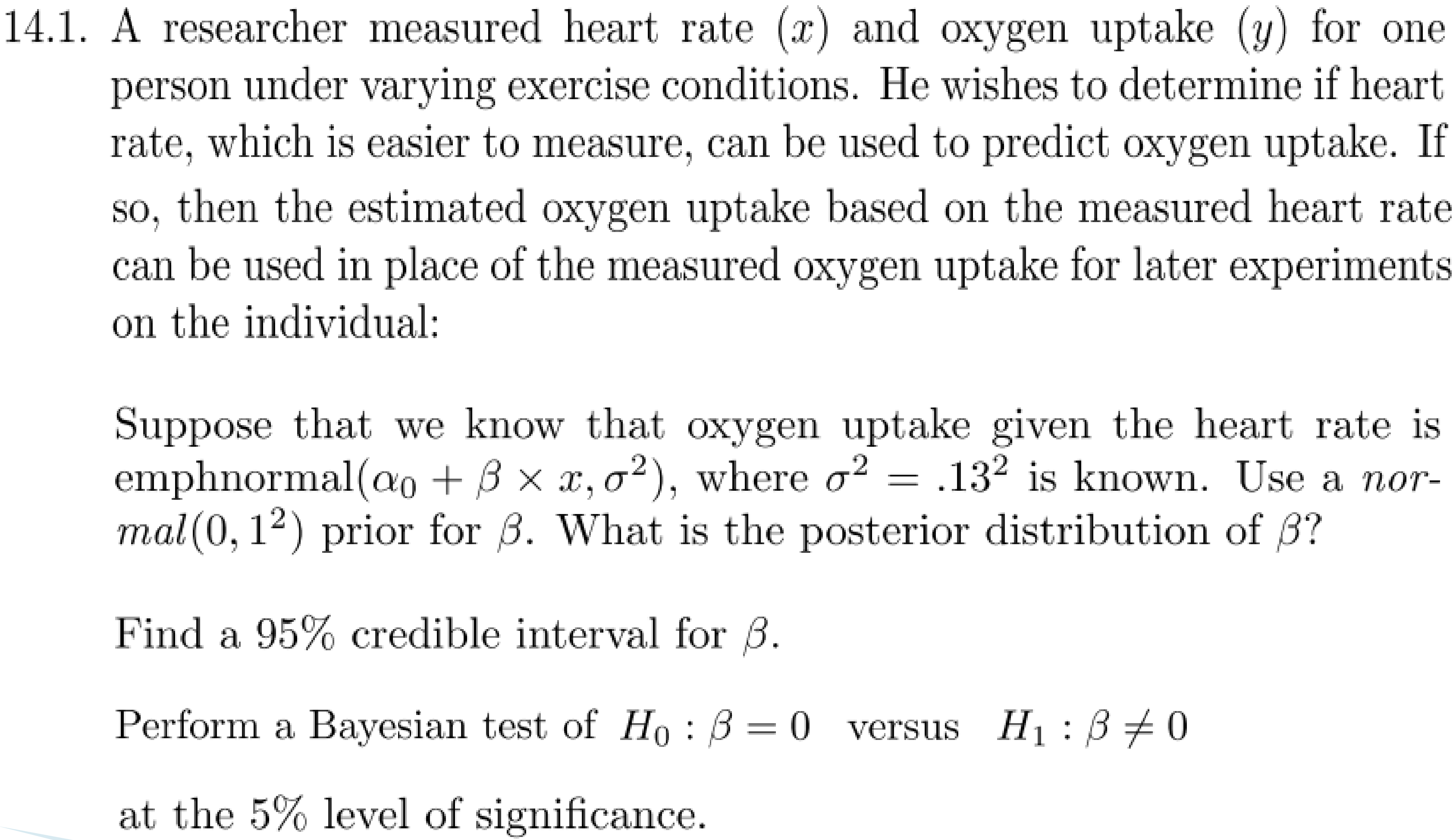


Studi Kasus 2: *(cont.)*

```
> ## Memprediksi y berdasarkan data x yang baru: x = 3, -1, 6
>
> bayes.lin.reg(y,x, slope.prior = "normal",
               intcpt.prior = "normal",
               0, 2, 1, 3, pred.x=c(3, -1, 6))


 x      Predicted y     SE
------  -----------  -----------
 3         11.81        1.0750
-1          1.97        1.0825
 6         19.20        1.0781
```

Studi Kasus 3:

Bolstad dan Curran (2017), hlm. 303



14.1. A researcher measured heart rate (x) and oxygen uptake (y) for one person under varying exercise conditions. He wishes to determine if heart rate, which is easier to measure, can be used to predict oxygen uptake. If so, then the estimated oxygen uptake based on the measured heart rate can be used in place of the measured oxygen uptake for later experiments on the individual:

Suppose that we know that oxygen uptake given the heart rate is emphnormal$(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = .13^2$ is known. Use a *normal*$(0, 1^2)$ prior for β. What is the posterior distribution of β?

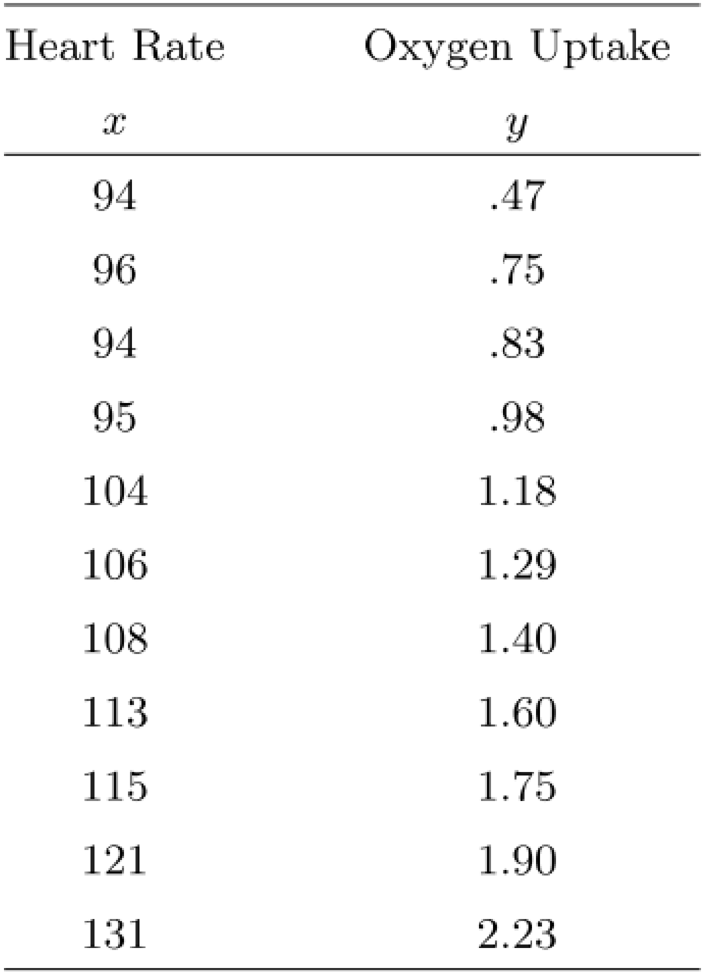
Find a 95% credible interval for β.

Perform a Bayesian test of $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$

at the 5% level of significance.

Studi Kasus 3:

Bolstad dan Curran (2017), hlm. 303



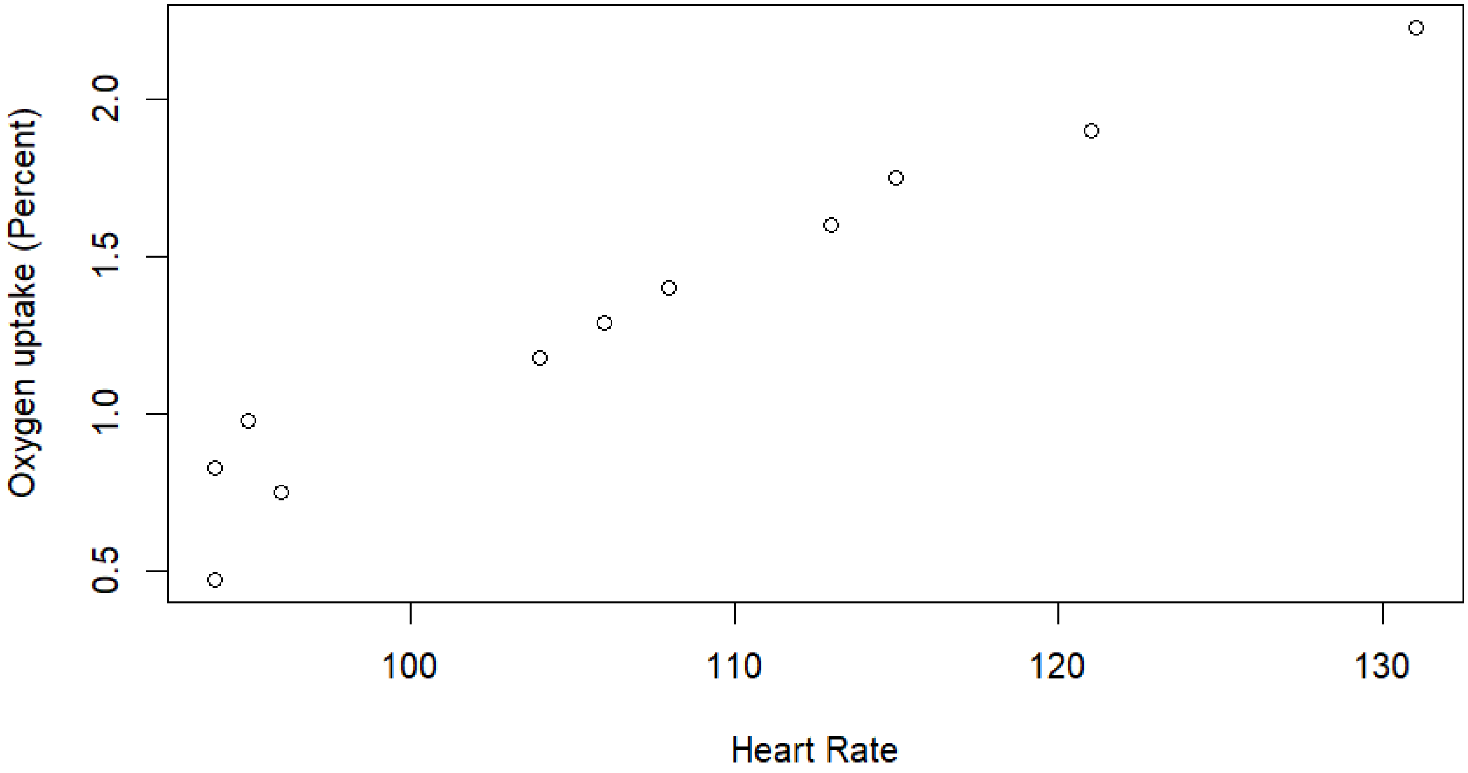
Heart Rate	Oxygen Uptake
x	y
94	.47
96	.75
94	.83
95	.98
104	1.18
106	1.29
108	1.40
113	1.60
115	1.75
121	1.90
131	2.23

Studi Kasus 3: *(cont.)*

```
> library(Bolstad)
>
> ## adjust plot margins
> par(mar = c(2, 2, 2, 2))
>
> ## data x dan y
> OU = c(0.47,0.75,0.83,0.98,1.18,1.29,1.40,
         1.60,1.75,1.90,2.23)
> HR = c(94,96,94,95,104,106,108,113,
         115,121,131)
```

Studi Kasus 3: *(cont.)*

```
> plot(HR,OU,xlab="Heart Rate",ylab="Oxygen uptake
      (Percent)")
```



Studi Kasus 3: *(cont.)*

```
> bayes.lin.reg(OU,HR,slope.prior = "normal",
                intcpt.prior = "flat",0,1,sigma=0.13)


Known standard deviation:  0.13


            Posterior Mean Posterior Std. Deviation
            -------------- ------------------------
Intercept:  1.307          0.039196
Slope:      0.043          0.003372
```

Studi Kasus 3: *(cont.)*

```
Posterior Mean Posterior Std. Deviation
            -------------- ------------------------
Intercept:  1.307          0.039196
Slope:      0.043          0.003372
```

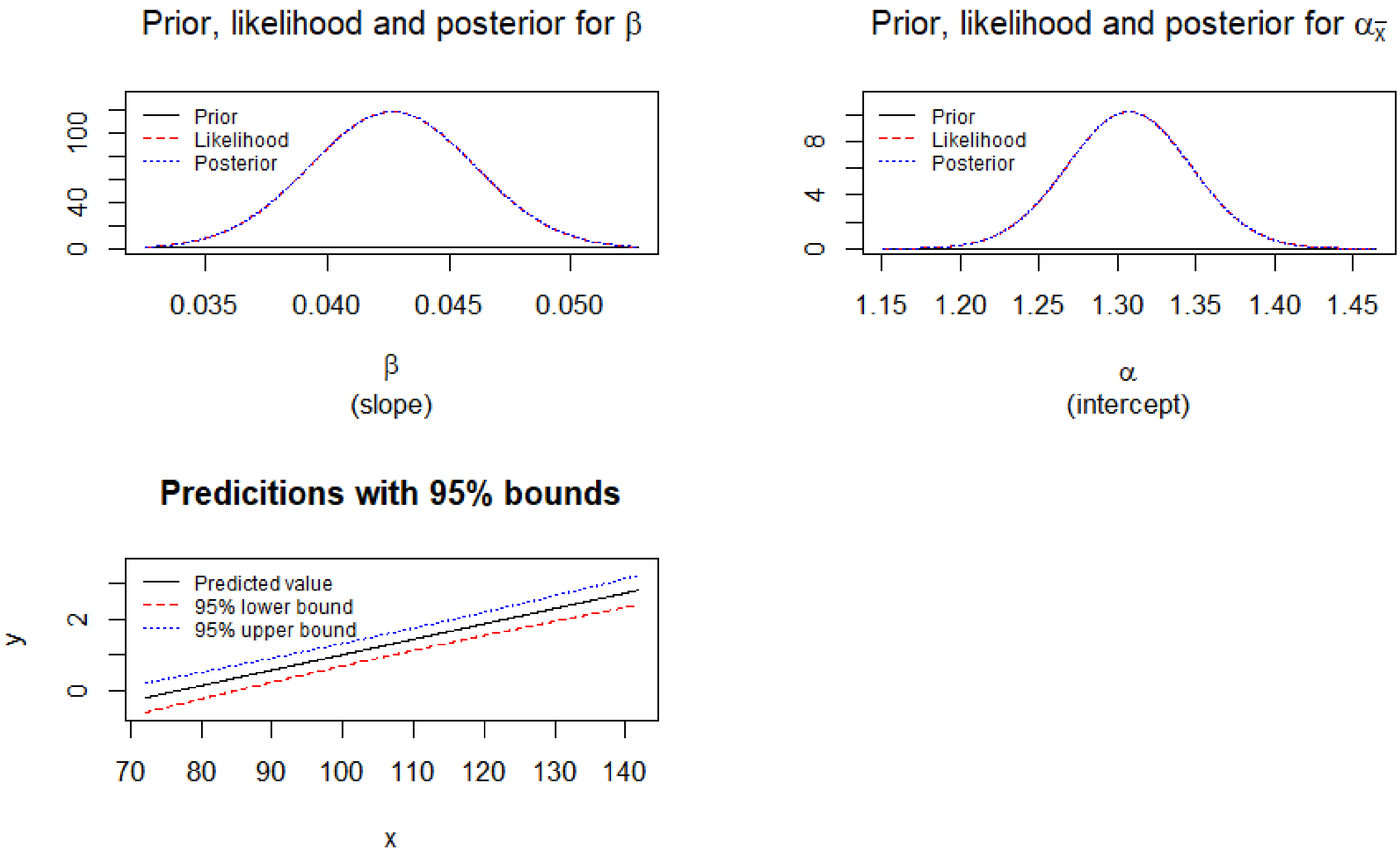
Find a 95% credible interval for β.

Perform a Bayesian test of

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0$$

at the 5% level of significance.

Studi Kasus 3: *(cont.)*



Prior, likelihood and posterior for β

Prior, likelihood and posterior for $\alpha_{\bar{x}}$

Predicitions with 95% bounds

Materi Praktikum

1

Bolstad dan Curran (2017), hlm. 305 → Gunakan Program R

14.2. A researcher is investigating the relationship between yield of potatoes (y) and level of fertilizer (x.) She divides a field into eight plots of equal size and applied fertilizer at a different level to each plot. The level of fertilizer and yield for each plot is recorded below:

Fertilizer Level	Yield
x	y
1	25
1.5	31
2	27
2.5	28
3	36
3.5	35
4	32
4.5	34

1

Bolstad dan Curran (2017), hlm. 305

(a) Plot a scatterplot of yield versus fertilizer level.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose that we know that yield given the fertilizer level is emphnormal$(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = 3.0^2$ is known. Use a $normal(2, 2^2)$ prior for β. What is the posterior distribution of β?

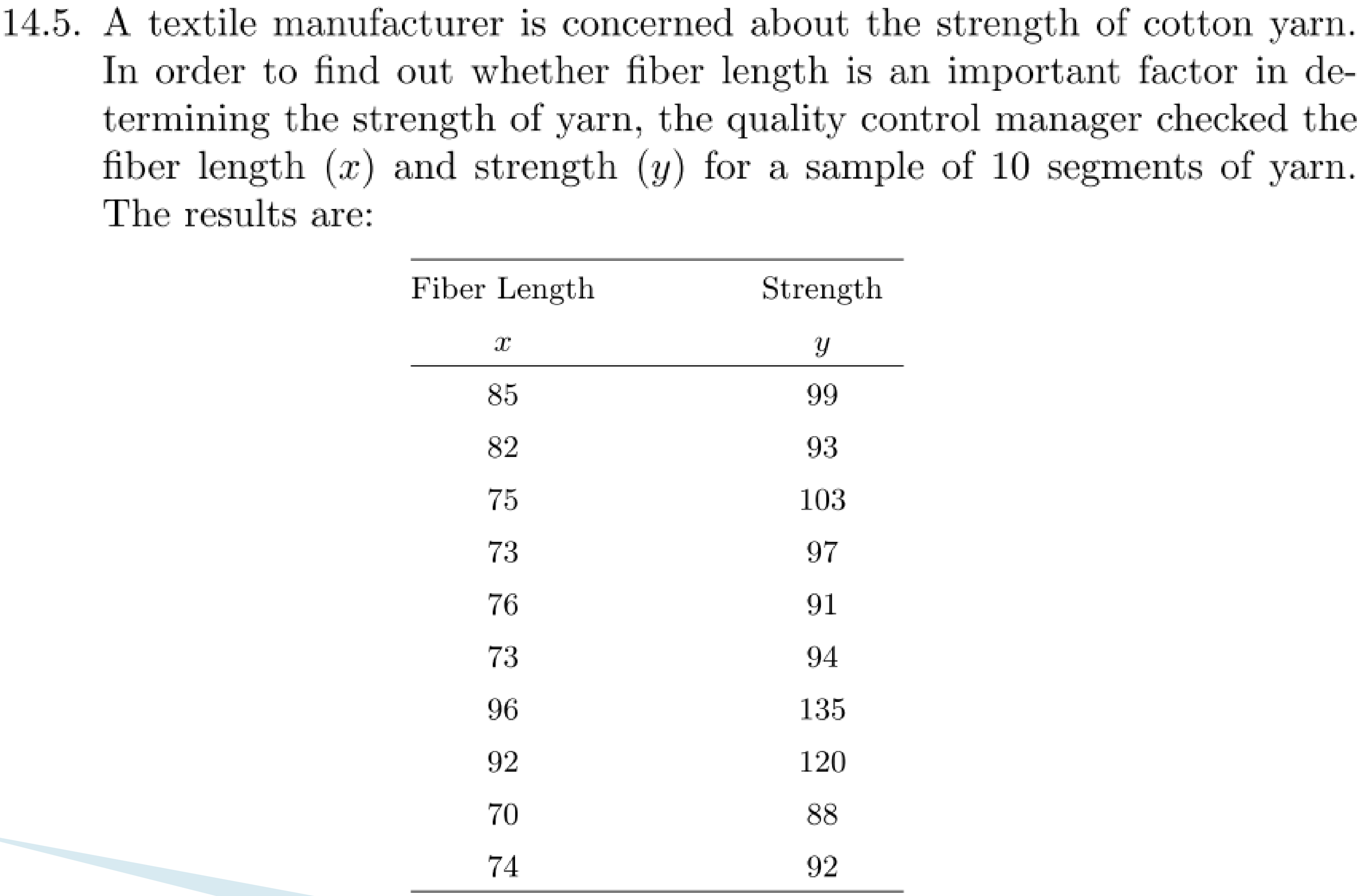
(f) Find a 95% credible interval for β.

(g) Perform a Bayesian test of

$$H_0 : \beta \leq 0 \quad \text{versus} \quad H_1 : \beta > 0$$

at the 5% level of significance.

2 Bolstad dan Curran (2017), hlm. 307 → Gunakan Program R



14.5. A textile manufacturer is concerned about the strength of cotton yarn. In order to find out whether fiber length is an important factor in determining the strength of yarn, the quality control manager checked the fiber length (x) and strength (y) for a sample of 10 segments of yarn. The results are:

Fiber Length x	Strength y
85	99
82	93
75	103
73	97
76	91
73	94
96	135
92	120
70	88
74	92

2 Bolstad dan Curran (2017), hlm. 307

(a) Plot a scatterplot of strength versus fiber length.

(b) Calculate the parameters of the least squares line.

(c) Graph the least squares line on your scatterplot.

(d) Calculate the estimated variance about the least squares line.

(e) Suppose we know that the strength given the fiber length is emphnormal$(\alpha_0 + \beta \times x, \sigma^2)$, where $\sigma^2 = 7.7^2$ is known. Use a $normal(0, 10^2)$ prior for β. What is the posterior distribution of β.

(f) Find a 95% credible interval for β.

(g) Perform a Bayesian test of

$$H_0 : \beta \leq 0 \quad \text{versus} \quad H_1 : \beta > 0$$

at the 5% level of significance.

(h) Find the predictive distribution for y_{11}, the strength of the next piece of yarn which has fiber length $x_{11} = 90$.

(i) Find a 95% credible interval for the prediction.

Pustaka

1. Reich BJ dan Ghosh SK. (2019). *Bayesian Statistical Methods*. Taylor and Francis Group.

2. Bolstad WM dan Curran JM. (2017). *Introduction to Bayesian Statistics 3th Edition*. John Wiley and Sons, Inc.

3. Lee, Peter M. (2012). *Bayesian Statistics, An Introduction 4th Edition*. John Wiley and Sons, Inc.

4. Albert, Jim. (2009). *Bayesian Computation with R*. Springer Science and Business Media.

5. Ghosh JK, Delampady M, dan Samanta T. (2006). *An Introduction to Bayesian Analysis, Theory and Methods*. Springer Science and Business Media.

Terima Kasih