Regressão Linear Simples Via Inferência Bayesiana

Ana Beatriz Tozzo Martins
Orientador: Prof. PhD. Paulo Justiniano Ribeiro Jr.

30 de outubro de 2007

Consideremos um modelo de regressão linear simples tal que $y_i \sim N(\beta x_i; 1)$. Este trabalho tem por objetivo fazer inferência sobre β obtendo uma amostra da posteriori usando reamostragem ponderada. Para isto, são utilizados dados que fazem parte da Tese de doutorado de Edson A. A. Silva.

```
> soja <- read.table("/home/anab/MetComputInt/DadosRegLinear.txt",
+     dec = ",", header = TRUE)

> nomes <- names(soja)
> nomes

[1] "X"    "Y"    "P"    "PH"   "K"    "MO"   "PROD" "SB"
```

A matriz de correlação é obtida a partir dos comandos seguintes e apresentada a seguir:

```
> SOJA <- data.frame(soja)
> SOJA.corr <- cor(SOJA)

> require(xtable)

> print(xtable(SOJA.corr))
```

Podemos observar uma forte correlação (89%) entre as variáveis PH e SB. Considerando PH a variável independente e SB a variável dependente, tem-se:

```
> lm(soja$SB ~ soja$PH)

Call:
lm(formula = soja$SB ~ soja$PH)

Coefficients:
(Intercept)      soja$PH
     -66.97        23.83
```

	X	Y	P	PH	K	MO	PROD	SB
X	1.00	−0.01	0.12	0.02	0.01	0.08	−0.02	−0.00
Y	−0.01	1.00	0.31	−0.53	0.08	0.55	−0.30	−0.50
P	0.12	0.31	1.00	−0.23	0.28	0.24	−0.05	−0.19
PH	0.02	−0.53	−0.23	1.00	−0.07	−0.50	0.04	0.89
K	0.01	0.08	0.28	−0.07	1.00	0.31	0.12	0.01
MO	0.08	0.55	0.24	−0.50	0.31	1.00	−0.04	−0.47
PROD	−0.02	−0.30	−0.05	0.04	0.12	−0.04	1.00	0.07
SB	−0.00	−0.50	−0.19	0.89	0.01	−0.47	0.07	1.00

e a estimativa de máxima verossimilhança para o parâmetro β é $23,83$.
Como no experimento foram utilizados dois tipos de tratamentos, criamos uma nova variável PC assumindo os valores 0 e 1.

```
> PC <- rep(c(rep.int(0:1, 8), rep.int(1:0, 8)), 8)
> soja <- cbind(soja, PC)
> class(soja)
```

```
[1] "data.frame"
```

Selecionamos os dados originados do mesmo tipo de tratamento, dados para $PC = 0$.

A identificação dos nomes das variáveis, a classificação do arquivo, o cabeçalho dos dados, estatísticas descritivas e a correlação entre SB e PH são apresentadas a seguir:

```
> names(soja1)
```

```
[1] "X"    "Y"    "P"    "PH"   "K"    "MO"   "PROD" "SB"   "PC"
```

```
> class(soja1)
```

```
[1] "data.frame"
```

```
> head(soja1)
```

```
       X   Y   P  PH    K    MO PROD    SB PC
1    5.6 3.6 4.4 6.3 0.53 47.52 2.56 80.63  0
3   24.8 4.6 3.5 5.4 0.47 51.79 2.35 61.47  0
5   44.0 1.6 3.6 5.6 0.52 59.70 2.99 71.61  0
7   63.2 2.6 2.9 5.2 0.47 48.74 2.91 61.11  0
9   82.4 3.6 3.0 5.7 0.36 51.79 3.00 70.67  0
11 101.6 3.6 2.7 5.6 0.38 51.17 2.36 67.44  0
```

```
> summary(soja1)
```

```
       X                Y                P               PH
 Min.   :  1.60   Min.   :  1.60   Min.   :2.200   Min.   :4.300
 1st Qu.: 38.60   1st Qu.: 29.60   1st Qu.:3.100   1st Qu.:4.800
 Median : 75.60   Median : 57.60   Median :3.700   Median :5.100
 Mean   : 75.85   Mean   : 57.29   Mean   :3.918   Mean   :5.113
 3rd Qu.:112.60   3rd Qu.: 85.60   3rd Qu.:4.500   3rd Qu.:5.400
 Max.   :149.60   Max.   :113.60   Max.   :9.100   Max.   :6.500
       K                MO              PROD             SB              PC
 Min.   :0.1700   Min.   :36.55   Min.   :1.190   Min.   :14.89   Min.   :0
 1st Qu.:0.2600   1st Qu.:48.74   1st Qu.:2.390   1st Qu.:46.08   1st Qu.:0
 Median :0.3150   Median :53.61   Median :2.750   Median :55.48   Median :0
 Mean   :0.3291   Mean   :52.89   Mean   :2.708   Mean   :54.83   Mean   :0
 3rd Qu.:0.3800   3rd Qu.:56.97   3rd Qu.:2.993   3rd Qu.:63.54   3rd Qu.:0
 Max.   :0.7400   Max.   :68.35   Max.   :3.700   Max.   :82.56   Max.   :0
```

```
> cor(soja1$SB, soja1$PH)

[1] 0.8964243
```

Construindo o diagrama de dispersão e ajustando a reta de regressão:

```
> par(mar = c(3.5, 3.5, 0.5, 0.5), mgp = c(2, 0.8, 0))
> plot(soja1$PH, soja1$SB, xlab = "PH", ylab = "SB", cex.axis = 1,
+     cex.lab = 1.1)
> abline(lm(soja1$SB ~ soja1$PH))
```

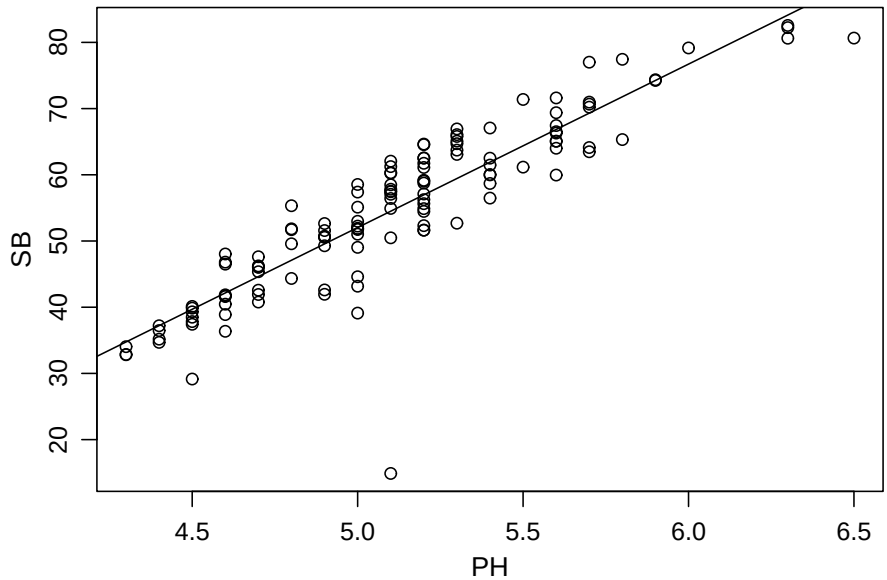

Figura 1: Diagrama de dispersão de PH versus SB e reta de regressão ajustada

Geramos agora 1000 valores usando uma priori vaga $N(0; 4)$ e em seguida apresentamos o histograma dos valores simulados.

```
> n = 1000
> beta = rnorm(n, 20, 10)
```

```
> par(mar = c(3.5, 3.5, 0.5, 0.5), mgp = c(2, 0.8, 0))
> hist(beta, main = "", cex.axis = 1, cex.lab = 1.1)
```

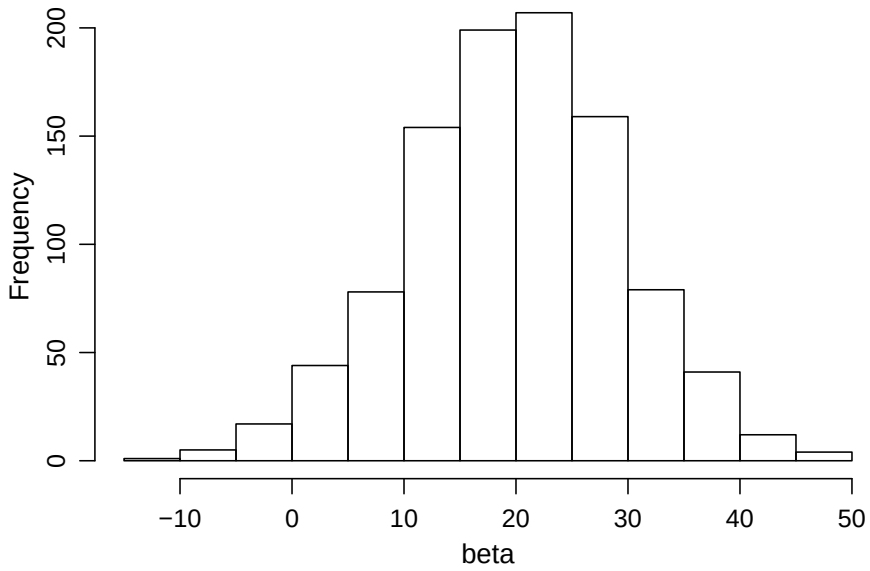

Figura 2: Distribuição dos valores de β simulados

O próximo passo é o cálculo dos pesos ou probabilidades

$$w_i = \frac{p(\theta_i|\mathbf{x})/q(\theta_i)}{\sum_{j=1}^{1000} p(\theta_j|\mathbf{x})/q(\theta_j)}, \qquad i = 1, ..., 1000,$$

que torna-se

$$w_i = \frac{p(\mathbf{x}|\theta_i)}{\sum_{j=1}^{1000} p(\mathbf{x}|\theta_j)}, \qquad i = 1, ..., 1000$$

ao considerar a priori como densidade auxiliar, $q(\theta) = p(\theta)$.

Calculamos a verossimilhança que, neste caso, será o logaritmo neperiano de L para evitar resultados nulos.

```
> ln.L = sapply(beta, function(b) (-0.5 * (sum((soja1$SB - b *
+     soja1$PH)^2))))
```

Com os resultados de $ln\, L$ calculamos os pesos w_i:

```
> w = ln.L/sum(ln.L)
> sum(w)
```

```
[1] 1
```

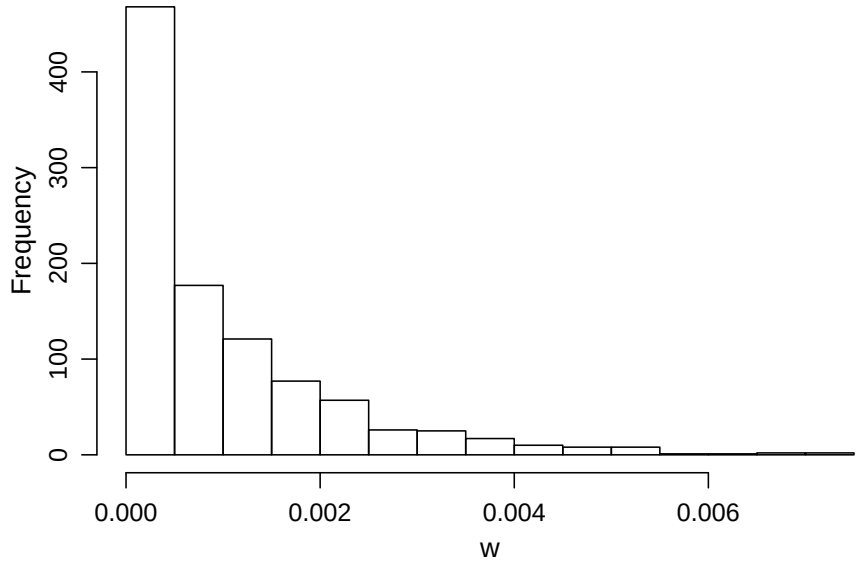

Figura 3: Distribuição dos pesos

```
> par(mar = c(3.5, 3.5, 0.5, 0.5), mgp = c(2, 0.8, 0))
> hist(w, main = "", cex.axis = 1, cex.lab = 1.1)
```

Reamostramos, agora, 500 betas com probabilidades w. Serão escolhidos os 500 valores de beta correspondentes as 500 maiores probabilidades w.

```
> m = 500
> beta.resample = sample(beta, size = m, rep = T, prob = w)
> summary(beta.resample)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -7.735  24.490  29.710  28.450  34.590  46.140
```

A figura 4, à esquerda, mostra o comportamento dos β's reamostrados e, à direita, observamos que os valores de β foram gerados em região de média densidade.

```
> par(mfrow = c(1, 2), mar = c(3.5, 3.5, 0.5, 0.5), mgp = c(2,
+     0.8, 0))
> hist(beta.resample, main = "", cex.axis = 1, cex.lab = 1.1)
> curve(dnorm(x, 20, 10), from = -15, to = 55, ylab = "priori",
+     xlab = expression(beta), cex.axis = 1, cex.lab = 1.1)
> rug(beta.resample)
```

O estimador $\hat{\beta}$ de β é a média dos valores reamostrados, dado por 28.45.

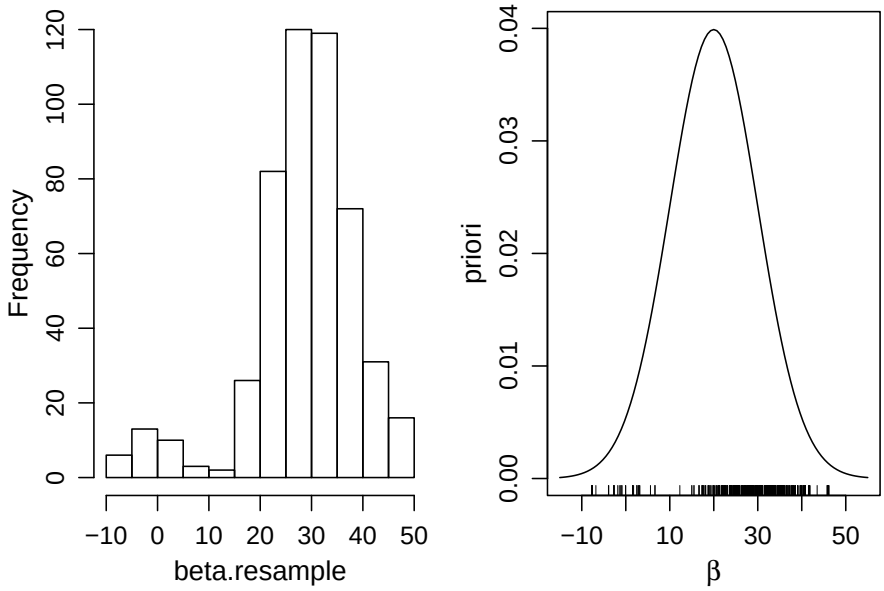

Figura 4: Distribuição dos valores de β reamostrados (esquerda), Densidade a priori e valores reamostrados (direita).