SME0809 - Inferência Bayesiana - Exercício de Regressão Linear

Grupo 13 - Francisco Miranda - 4402962 - Heitor Carvalho - 11833351

05/11/2021

Exercício de Regressão (Agricultura X Urbanismo)

Os dados a seguir mostram a escolaridade (X) e o rendimento (Y) de 5 pessoas ocupadas na agricultura e 5 pessoas ocupadas nos setores "urbanos" (indústria ou serviços). Define-se uma variável binária Z que é igual a zero para pessoas ocupadas na agricultura e é igual a um nos demais casos.

Você diria que existe diferença entre o rendimento dos trabalhadores dos diferentes setores? (Faça uma análise bayesiana)

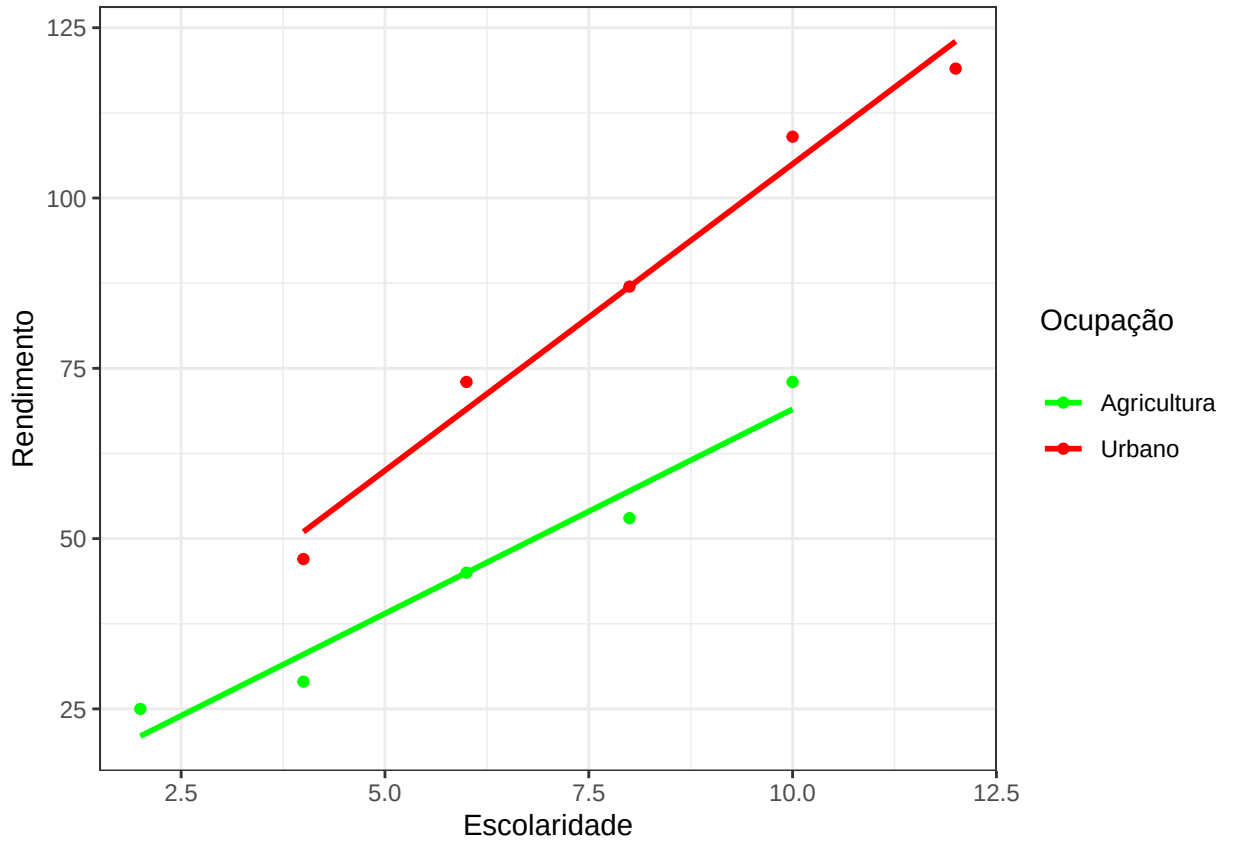
Antes de começarmos, precisamos separar nosso dataset em duas partes, uma apenas com os valores de $z = 0$ e outra com os valores para $z=1$.

```
z <- c(rep(0,5),rep(1,5))

x <- c(seq(2,10,2),seq(4,12,2))

y <- c(25,29,45,53,73,47,73,87,109,119)

dad <- data.frame(z=z,x=x,y=y)

#valores para z = 0
dad_0 <- dad |> filter(z == 0)
dad_1 <- dad |> filter(z == 1)
```

A princípio, vamos verificar o modelo linear ajustado pelo mínimos quadrados, utilizando-se da função **lm** do R.

```
# z <- as.factor(z)
# z <- fct_recode(z, Agricultura = '0', Urbano = '1')

dad |>
  ggplot(aes(x = x, y = y, color = as.factor(z))) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs( x = "Escolaridade", y = "Rendimento", color = "Ocupação\n") +
  scale_color_manual(labels = c("Agricultura", "Urbano"), values = c("green", "red")) +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



De início, podemos verificar a partir do gráfico que os tralhadores urbanos apresentam um incremento maior no **Rendimento**, dado uma variação unitária na **Escolaridade**.

Podemos calcular os resultados do modelo linear usando **summary(lm)**, para a posterior comparação.

```
summary(lm(dad_0$y~dad_0$x))
```

```
##
## Call:
## lm(formula = dad_0$y ~ dad_0$x)
##
## Residuals:
##          1          2          3          4          5
##  4.000e+00 -4.000e+00 -3.553e-15 -4.000e+00  4.000e+00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.0000     4.8442   1.858  0.16017
## dad_0$x       6.0000     0.7303   8.216  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.619 on 3 degrees of freedom
## Multiple R-squared:  0.9574, Adjusted R-squared:  0.9433
## F-statistic:  67.5 on 1 and 3 DF,  p-value: 0.003774
```

```
summary(lm(dad_1$y~dad_1$x))
```

```
##
## Call:
## lm(formula = dad_1$y ~ dad_1$x)
##
## Residuals:
##          1          2          3          4          5
## -4.000e+00  4.000e+00 -4.441e-16  4.000e+00 -4.000e+00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.0000     6.1968   2.421  0.09412 .
## dad_1$x       9.0000     0.7303  12.324  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.619 on 3 degrees of freedom
## Multiple R-squared:  0.9806, Adjusted R-squared:  0.9742
## F-statistic: 151.9 on 1 and 3 DF,  p-value: 0.001151
```

Regressão Linear Bayesiana

Vamos criar um arquivo de texto, que receberá o nosso modelo.Em seguida, faremos duas iterações, uma para $\mathbf{z = 0}$, correspondendo ao Trabalhador Agrícola e outra para $\mathbf{z = 1}$, referente ao Trabalhador Urbano.

Assumiremos para β_0, β_1 prioris vagas tal que: $\beta_0, \beta_1 \sim \mathcal{N}(0, 10^{-6})$

```
modelo <- function(){
  #definindo a distribuição dos dados
  for (i  in 1:N) {
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1*x[i]
  }
#modelagem da incerteza a priori
#usando uma priori vaga

  beta0 ~ dnorm(0.0, 1.0E-6)
  beta1 ~ dnorm(0.0, 1.0E-6)
  tau ~ dgamma(0.001, 0.001)
  sigma2 <- 1/tau #o OPENBUGS trabalha com a precisão, que é o inverso da variancia
}
```

Salvando o modelo em um arquivo de texto.

```
mod <- file.path(tempdir(), "mod0.txt")

write.model(modelo, mod)
```

Executando o modelo para $z = 0$:

```r
#Entrada dos dados

#modelo para z = 0
x <- dad_0$x
y <- dad_0$y

N <- length(x)
data <- list('N' = N, 'x' = x, "y" = y)

#atribuindo os valores iniciais para os nós estocásticos
params <- c('beta0','beta1',"tau","sigma")

inits <- list(
  list(beta0 = 1, beta1 = 0.5, tau = 1),
  list(beta0 = 2,beta1=2, tau=10),
  list(beta0 =-1,beta1=2, tau=0.5))

#obtendo as cadeias de markov
out <- bugs(data, inits, params, modelo, codaPkg = TRUE,
            n.chains = 3, n.iter = 14000, n.thin = 1,
            n.burnin = 4000, debug = T)
out.coda <- read.bugs(out)
```

```
## Abstracting beta0 ... 10000 valid values
## Abstracting beta1 ... 10000 valid values
## Abstracting deviance ... 10000 valid values
## Abstracting tau ... 10000 valid values
## Abstracting beta0 ... 10000 valid values
## Abstracting beta1 ... 10000 valid values
## Abstracting deviance ... 10000 valid values
## Abstracting tau ... 10000 valid values
## Abstracting beta0 ... 10000 valid values
## Abstracting beta1 ... 10000 valid values
## Abstracting deviance ... 10000 valid values
## Abstracting tau ... 10000 valid values
```

O modelo resultante apresenta o seguinte resumo estatístico:

	mean	sd	val2.5pc	median	val97.5pc	sample
beta0	8.817	7.998	-8.805	8.985	24.37	30000
beta1	6.026	1.207	3.776	6.001	8.538	30000
deviance	31.39	3.674	27.26	30.41	40.99	30000
tau	0.04645	0.03828	0.002897	0.03677	0.1465	30000

Deviance information

	Dbar	Dhat	DIC	pD
y	31.39	27.51	35.27	3.878
total	31.39	27.51	35.27	3.878

Figure 1: Modelo z = 0

Executando o modelo para $z = 1$:

```r
#Lendo os dados para z = 1
#modelo para z = 0
x <- dad_1$x
y <- dad_1$y

N <- length(x)
data <- list('N' = N, 'x' = x, "y" = y)

#atribuindo os valores iniciais para os nós estocásticos
params <- c('beta0','beta1',"tau","sigma")

inits <- list(
  list(beta0 = 1, beta1 = 0.5, tau = 1),
  list(beta0 = 2,beta1=2, tau=10),
  list(beta0 =-1,beta1=2, tau=0.5))

#obtendo as cadeias de markov
out <- bugs(data, inits, params, modelo, codaPkg = TRUE,
            n.chains = 3, n.iter = 14000, n.thin = 1,
            n.burnin = 4000, debug = T)
out.coda <- read.bugs(out)
```

```
## Abstracting beta0 ... 10000 valid values
## Abstracting beta1 ... 10000 valid values
## Abstracting deviance ... 10000 valid values
## Abstracting tau ... 10000 valid values
## Abstracting beta0 ... 10000 valid values
## Abstracting beta1 ... 10000 valid values
## Abstracting deviance ... 10000 valid values
## Abstracting tau ... 10000 valid values
## Abstracting beta0 ... 10000 valid values
## Abstracting beta1 ... 10000 valid values
## Abstracting deviance ... 10000 valid values
## Abstracting tau ... 10000 valid values
```

Analogamente, o modelo para $z = 1$ apresenta o seguinte resumo estatístico:

Summary statistics

	mean	sd	val2.5pc	median	val97.5pc	sample
beta0	14.73	10.22	-6.204	15.03	34.2	30000
beta1	9.03	1.207	6.718	8.999	11.38	30000
deviance	31.4	3.676	27.24	30.44	40.66	30000
tau	0.04636	0.0383	0.002909	0.03663	0.1468	30000

Deviance information

	Dbar	Dhat	DIC	pD
y	31.4	27.51	35.29	3.886
total	31.4	27.51	35.29	3.886

Figure 2: Modelo z = 1

Comparando a abordagem bayesiana com a tradicional implementada pela função **lm**, concluímos que os coeficientes encontrados são basicamente, idênticos, conforme as tabelas abaixo.

```r
lm_mod0 <- c(9, 6)
lm_mod1 <- c(15, 9)
bayes_mod0 <- c(8.817, 6.026)
bayes_mod1 <- c(14.73, 9.03)

#tabela com os valores dos coeficientes
df = data.frame(Parameter = c("$\\beta_{0}$", "$\\beta_{1}$"),
                "z0" = lm_mod0,"z1" = lm_mod1)
kable(t(df), caption = "Coeficientes Mínimos quadrados para valores de z = 0 e z = 1")
```

Table 1: Coeficientes Mínimos quadrados para valores de z = 0 e z = 1

Parameter	β_0	β_1
z0	9	6
z1	15	9

```r
df2 = data.frame(Parameter = c("$\\beta_{0}$", "$\\beta_{1}$"),
                "z0" = bayes_mod0,"z1" = bayes_mod1)
kable(t(df2), caption = "Coeficientes do modelo Bayesiano para valores de z = 0 e z = 1")
```

Table 2: Coeficientes do modelo Bayesiano para valores de z = 0 e z = 1

Parameter	β_0	β_1
z0	8.817	6.026
z1	14.73	9.03

Por fim, podemos concluir que para os trabalhadores do setor Urbano, a variação do **Rendimento (y)** por unidade de variação de **Escolaridade (x)**, é cerca de 1.6× maior que para os trabalhadores do setor Agrícola.