Lista 2

MI406-Regressão

Caio Gomes Alves

1 Questão 1:

1.1 Pergunta

Considere os pontos $x_1 = 1, x_2 = 2, ..., x_{10} = 10$. Gere 10.000 (dez mil) simulações das variáveis respostas Y_i para o modelo descrito acima. Para cada simulação gerada, registre os seguintes valores:

- Estimativas de Mínimos Quadrados β_0 e β_1.
- Intervalo de confiança de 95% para β_0 e β_1.
- Estatística do teste para as hipóteses $\beta_0 = 5$ vs $\beta_0 \neq 5$.
- Estatística do teste para as hipóteses $\beta_1 = 2$ vs $\beta_1 \neq 2$.
- Estatística do teste para as hipóteses $\beta_1 = 1.8$ vs $\beta_1 \neq 1.8$.

1.2 Resposta

Inicialmente é preciso gerar os dados:

```r
# Seed para reprodutibilidade:
set.seed(5050)

# Valores dos x:
x <- 1:10

# 10.000 simulações dos Y_i:
y_sim <- replicate(10000, 5 + 2 * x + rnorm(10, 0, 1))
```

O resultado é uma matriz com 10 linhas e 10.000 colunas. Para nos auxiliar, iremos criar alguns objetos intermediários para armazenar as informações de maneira eficiente:

```r
# Vetores para as estimativas pontuais dos betas:
beta_0 <- vector(mode = "numeric", 10000)
beta_1 <- vector(mode = "numeric", 10000)
sigma_2 <- vector(mode = "numeric", 10000)

# Listas para os intervalos de confiança:
ic_beta_0 <- vector(mode = "list", 10000)
ic_beta_1 <- vector(mode = "list", 10000)

# Vetores para as estatísticas para os testes de hipótese:
estat_teste_beta_0 <- vector(mode = "numeric", 10000)
```

```r
estat_teste_beta_1_2 <- vector(mode = "numeric", 10000)
estat_teste_beta_1_1.8 <- vector(mode = "numeric", 10000)
```

Para além disso, iremos utilizar a seguinte função para calcular as somas dos produtos corrigidos pela média:

```r
func_soma <- function(a, b) {
    sum((a - mean(a)) * (b - mean(b)))
}
```

Agora, podemos utilizar um loop para a estimação dos β_0 e β_1, além das demais informações solicitadas:

```r
# Loop para fazer os ajustes e popular os vetores e listas:
for (i in 1:10000) {
    # Estimativas pontuais dos Betas:
    beta_1[i] <- func_soma(y_sim[, i], x)/func_soma(x, x)
    beta_0[i] <- mean(y_sim[, i]) - beta_1[i] * mean(x)
    # Estimativa da variância pelos resíduos:
    sigma_2[i] <- sum((y_sim[, i] - (beta_0[i] + beta_1[i] *
        x))^(2))/(length(x) - 2)
    # Valor da distribuição t para criação dos intervalos
    # de confiança:
    t_alpha <- qt(1 - 0.05/2, length(x) - 2)
    # Calcula os intervalos de confiança dos betas:
    ic_beta_0[[i]] <- c(`2.5 %` = beta_0[i] - (sqrt((sigma_2[i] *
        sum(x^(2)))/(length(x) * func_soma(x, x))) * t_alpha),
        `97.5 %` = beta_0[i] + (sqrt((sigma_2[i] * sum(x^(2)))/(length(x) *
            func_soma(x, x))) * t_alpha))
    ic_beta_1[[i]] <- c(`2.5 %` = beta_1[i] - sqrt((sigma_2[i])/(func_soma(x,
        x))) * t_alpha, `97.5 %` = beta_1[i] + sqrt((sigma_2[i])/(func_soma(x,
        x))) * t_alpha)
    # Estatística de teste para Beta_0 = 5:
    estat_teste_beta_0[i] <- (beta_0[i] - 5)/sqrt((sigma_2[i] *
        sum(x^(2)))/(length(x) * func_soma(x, x)))
    # Estatística de teste para Beta_1 = 2:
    estat_teste_beta_1_2[i] <- (beta_1[i] - 2)/sqrt(sigma_2[i]/func_soma(x,
        x))
    # Estatística de teste para Beta_1 = 1.8:
    estat_teste_beta_1_1.8[i] <- (beta_1[i] - 1.8)/sqrt(sigma_2[i]/func_soma(x,
        x))
}
```

Por fim, transformemos as listas que armazenam os valores inferiores e superiores dos intervalos de confiança em matrizes, para que a manipulação seja mais direta:

```r
ic_beta_0 <- do.call(rbind, ic_beta_0)
ic_beta_1 <- do.call(rbind, ic_beta_1)
```

2 Questão 2:

2.1 Pergunta

Gere um gráfico para visualizar a distribuição de β_0 e β_1 ao longo das simulações e apresente a média e variância dessas estatísticas. Que valores de média e variância você esperaria obter?

2.2 Resposta

2.2.1 Para $\hat{\beta}_0$:

Temos que $\mathbb{E}(\hat{\beta}_0) = \beta_0 = 5$ e a variância desse estimador é dado por:

$$\begin{aligned} \text{Var}(\hat{\beta}_0) &= \frac{\sigma^2 \sum_{i=1}^{n}(x_i^2)}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{1 \times 385}{10 \times 82.5} \\ &= 0.4\bar{6} \end{aligned}$$

Podemos ver se os valores dos betas ajustados se aproximam disso:

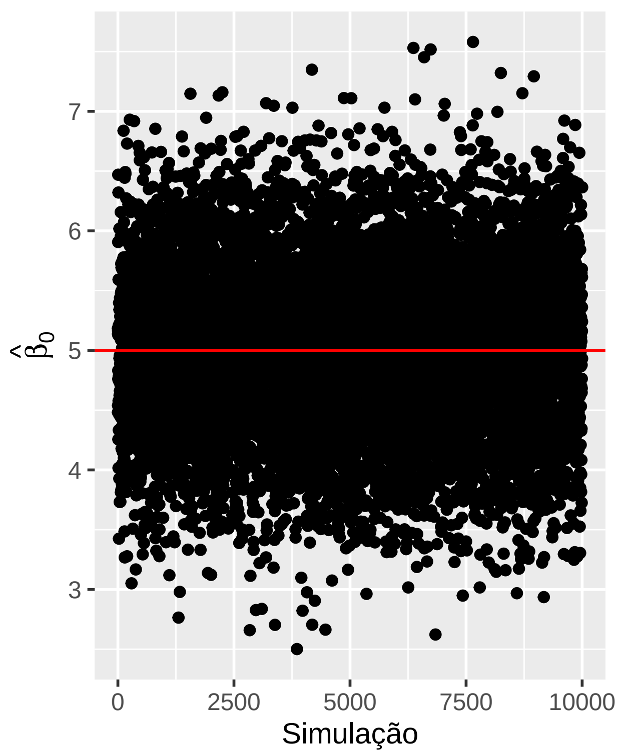
```
# Média das estimativas:
mean(beta_0)
```

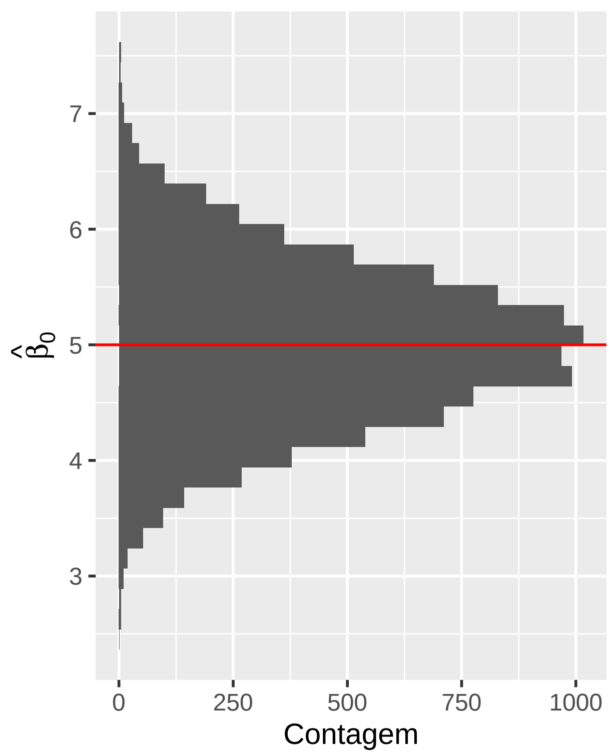
```
## [1] 4.996607
```

```
# Variância das estimativas:
var(beta_0)
```

```
## [1] 0.4647289
```

Os valores estão bem próximos dos valores exatos. Podemos ver a distribuição das estimativas ao longo das simulações a partir de dois gráficos: o primeiro indica a posição pontual de cada uma das 10.000 simulações e o segundo é um histograma de todas as estimativas em conjunto. Em ambos a linha em vermelho indica o verdadeiro valor de β_0:





2.2.2 Para $\hat{\beta}_1$:

Temos que $\mathbb{E}(\hat{\beta}_1) = \beta_1 = 2$ e a variância desse estimador é dado por:

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{1}{82.5} \\
&= 0.01\overline{2}
\end{aligned}
$$

Podemos ver se os valores dos betas ajustados se aproximam disso:

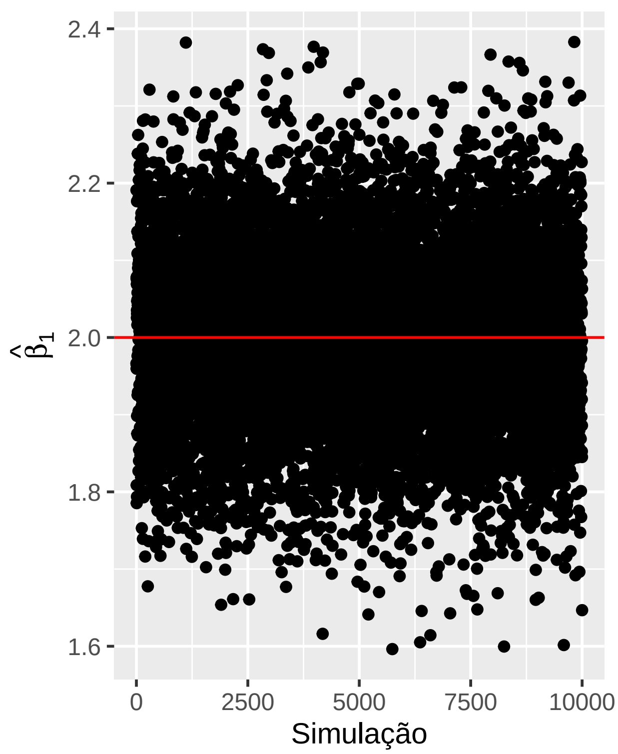
```r
# Média das estimativas:
mean(beta_1)
```

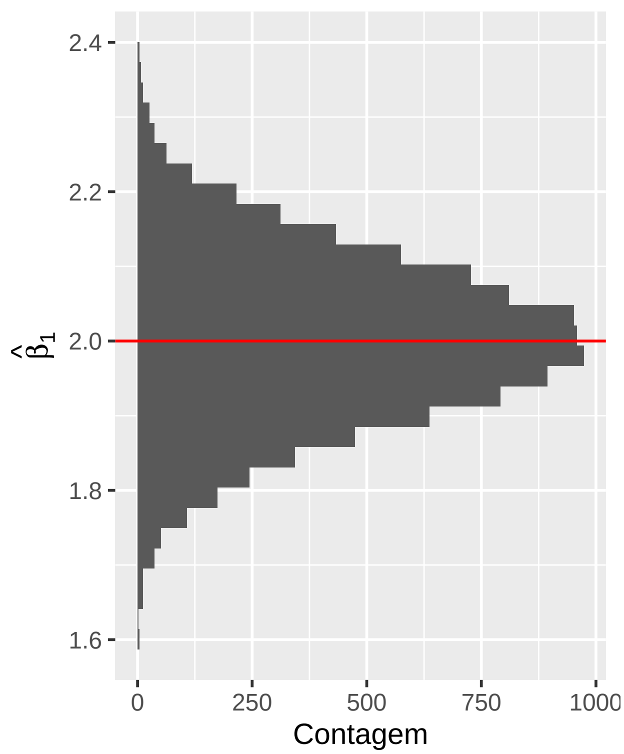
```
## [1] 2.00069
```

```r
# Variância das estimativas:
var(beta_1)
```

```
## [1] 0.01225903
```

Os valores estão bem próximos dos valores exatos. Podemos ver a distribuição das estimativas ao longo das simulações a partir de dois gráficos: o primeiro indica a posição pontual de cada uma das 10.000 simulações e o segundo é um histograma de todas as estimativas em conjunto. Em ambos a linha em vermelho indica o verdadeiro valor de β_1:





3 Questão 3:

3.1 Pergunta

Em quantas simulações o valor verdadeiro esteve dentro do intervalo de confiança para β_0 e β_1, isoladamente? Em quantas simulações ambos os intervalos continham o valor verdadeiro? Que quantidades você esperaria em cada um dos casos?

3.2 Resposta

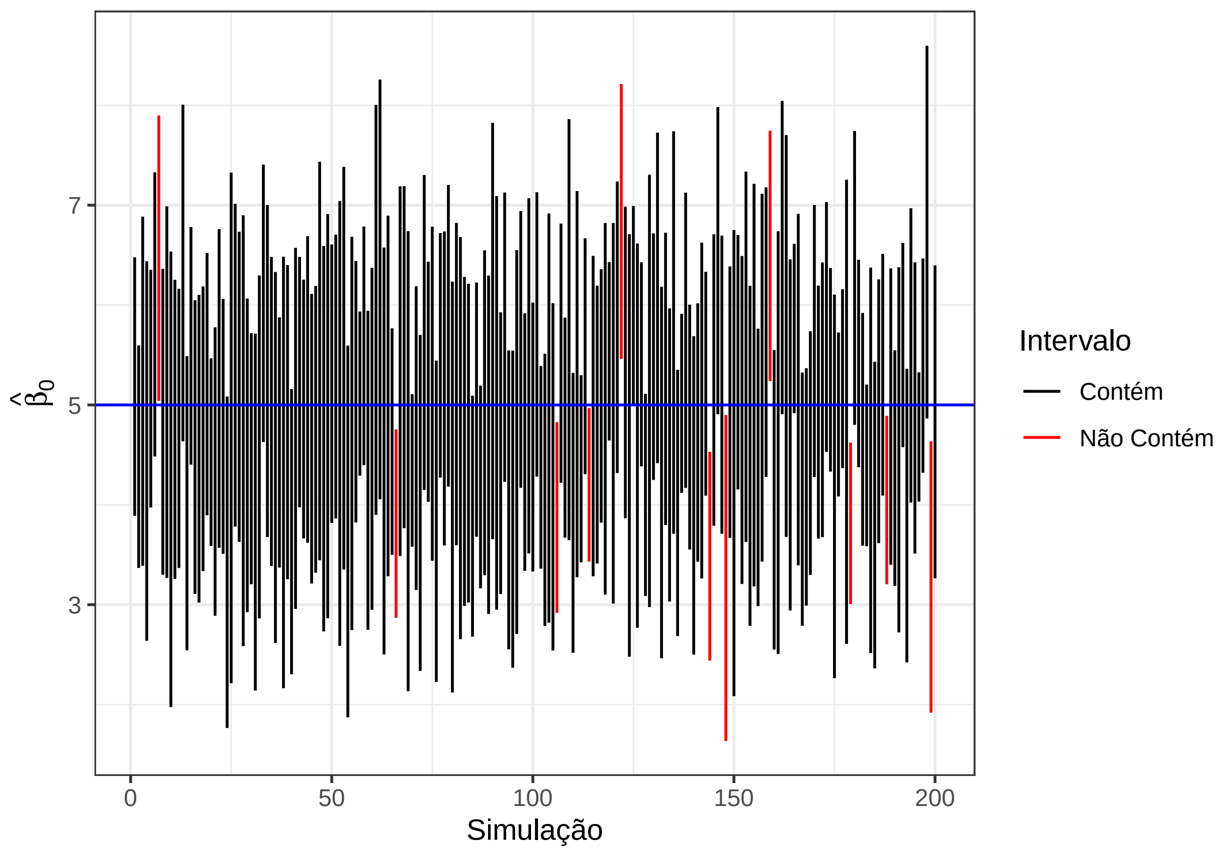
3.2.1 Para $\hat{\beta}_0$:

Podemos encontrar em quantos intervalos de confiança o valor verdadeiro de β_0 a partir do seguinte código:

```
10000 - sum(ic_beta_0[, 1] > 5 | ic_beta_0[, 2] < 5)
```

```
## [1] 9502
```

Assim, a quantidade de intervalos que não contém é de 498, aproximadamente 5% (que era o valor esperado, já que o intervalo é construído utilizando o valor de $\alpha = 0.05$ para a distribuição t). Podemos, a título de exemplo, mostrar os intervalos construídos para as 200 primeiras simulações:



Dos 200 primeiros intervalos, apenas 11 não contém o verdadeiro valor de β_0 (o esperado eram 10, o que é bem próximo do observado).

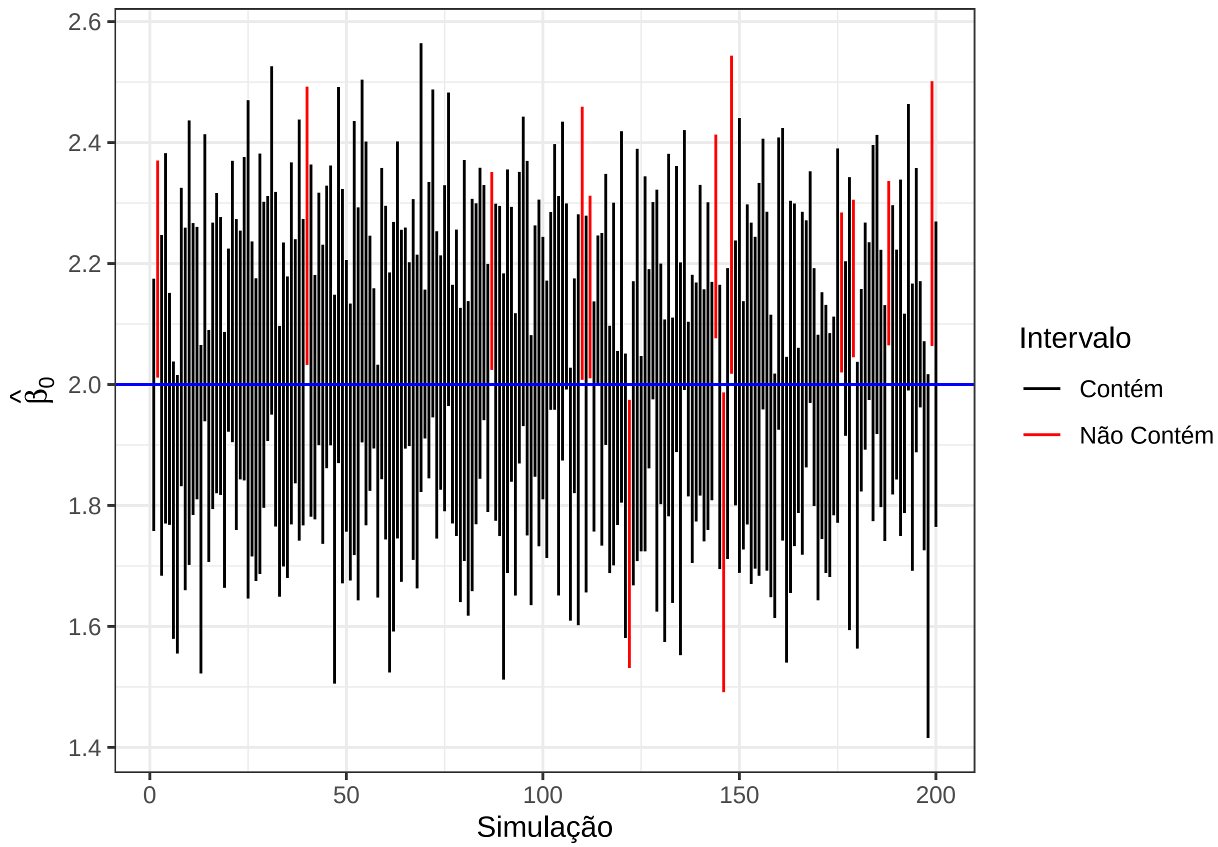
3.2.2 Para $\hat{\beta}_1$:

Podemos encontrar em quantos intervalos de confiança o valor verdadeiro de β_1 a partir do seguinte código:

```
10000 - sum(ic_beta_1[, 1] > 2 | ic_beta_1[, 2] < 2)
```

```
## [1] 9466
```

Assim, a quantidade de intervalos que não contém é de 534, aproximadamente 5% (que era o valor esperado, já que o intervalo é construído utilizando o valor de $\alpha = 0.05$ para a distribuição t). Podemos, a título de exemplo, mostrar os intervalos construídos para as 200 primeiras simulações:



Dos 200 primeiros intervalos, apenas 13 não contém o verdadeiro valor de β_1 (o esperado eram 10, o que é bem próximo do observado).

4 Questão 4

4.1 Pergunta:

Considerando um nível de significância $\alpha = 0.05$, em quantas simulações a hipótese $\beta_0 = 5$ seria rejeitada? Que número você esperaria?

4.2 Resposta:

Podemos encontrar a quantidade de simulações em que rejeitam a hipótese $\beta_0 = 5$ comparando o valor absoluto da estatística de teste obtida com a distribuição $t(8)$:

```
sum(abs(estat_teste_beta_0) > qt(1 - 0.05/2, length(x) - 2))
```

```
## [1] 498
```

O valor esperado é de 5% das simulações (ou seja, 500 das 10.000). Essa quantidade coincide com quantos intervalos contém o verdadeiro valor de β_0, visto que ambos são calculados com o mesmo quantil da distribuição $t(8)$, considerando um $\alpha = 0.05$.

5 Questão 5

5.1 Pergunta:

Considerando um nível de significância $\alpha = 0.05$, em quantas simulações a hipótese $\beta_1 = 2$ seria rejeitada? Que número você esperaria?

5.2 Resposta:

Podemos encontrar a quantidade de simulações em que rejeitam a hipótese $\beta_1 = 2$ comparando o valor absoluto da estatística de teste obtida com a distribuição $t(8)$:

```
sum(abs(estat_teste_beta_1_2) > qt(1 - 0.05/2, length(x) - 2))
```

```
## [1] 534
```

O valor esperado é de 5% das simulações (ou seja, 500 das 10.000). Essa quantidade coincide com quantos intervalos contém o verdadeiro valor de β_1, visto que ambos são calculados com o mesmo quantil da distribuição $t(8)$, considerando um $\alpha = 0.05$.

6 Questão 6

6.1 Pergunta:

Considerando um nível de significância $\alpha = 0.05$, em quantas simulações a hipótese $\beta_1 = 1.8$ seria rejeitada? Que número você esperaria?

6.2 Resposta:

Podemos encontrar a quantidade de simulações em que rejeitam a hipótese $\beta_1 = 1.8$ comparando o valor absoluto da estatística de teste obtida com a distribuição $t(8)$:

```r
sum(abs(estat_teste_beta_1_1.8) > qt(1 - 0.05/2, length(x) -
    2))
```

```
## [1] 3634
```

Aqui, como o valor da hipótese está mais "longe" do verdadeiro valor que gera os dados (que conhecemos, neste caso), esperávamos que a quantidade de simulações que rejeitariam a hipótese nula aumentasse, o que realmente ocorreu.

7 Questão 7

7.1 Pergunta:

Considerando um nível de significância $\alpha = 0.05$, em quantas simulações pelo menos uma das hipóteses $\beta_0 = 5$ ou $\beta_1 = 2$ seria rejeitada? Que número você esperaria? O que você pode concluir a partir disso?

7.2 Resposta:

A quantidade de simulações em que, pelo menos uma das duas hipóteses seria rejeitada é dada por:

```r
sum((abs(estat_teste_beta_1_2) > qt(1 - 0.05/2, length(x) - 2)) |
    (abs(estat_teste_beta_0) > qt(1 - 0.05/2, length(x) - 2)))
```

```
## [1] 716
```

Caso as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ fossem independentes, esperávamos que a seguinte proporção fosse dada pela seguinte expressão:

$$
\begin{aligned}
P\left(\hat{\beta}_0 \neq 5, \hat{\beta}_1 = 2 | \hat{\beta}_0 = 5, \hat{\beta}_1 = 2\right) + P\left(\hat{\beta}_0 = 5, \hat{\beta}_1 \neq 2 | \hat{\beta}_0 = 5, \hat{\beta}_1 = 2\right) + P\left(\hat{\beta}_0 \neq 5, \hat{\beta}_1 \neq 2 | \hat{\beta}_0 = 5, \hat{\beta}_1 = 2\right) \\
= 0.05 \times 0.95 + 0.05 \times 0.95 + 0.05 \times 0.05 \\
= 0.0975
\end{aligned}
$$

Porém, como as estimativas não são independentes, esses intervalos não serão ortogonais e a quantidade de simulações em que pelo menos uma das hipóteses será rejeitada é menor do que caso elas fossem independentes.