Modelos Lineares Generalizados

Módulo #1. Uriel Moreira Silva

urielmoreirasilva@ufmg.br

DEST-ICEx UFMG

2024/1

EXEMPLO PRÁTICO I

Exemplo Prático I

- 48 indicadores macroeconômicos, de diferentes categorias
  - coletados da *Penn World Table* (Feenstra et al., 2015)

- Disponíveis para 183 países, entre os anos de 1950 a 2019

- Os dados desse exemplo estão disponíveis em “Módulo 1/code/_dta/pwt1001.dta”

Exemplo Prático I

- Nosso objetivo nesse exemplo é estimar uma *função de produção* do tipo *Cobb-Douglas* para todos os países, utilizando dados do ano de 2010

- Em essência, uma função de produção mede a capacidade máxima que um país (ou uma empresa) pode produzir para uma dada quantidade de insumos
  - os insumos em geral são *capital* e *força de trabalho*

DEST
UFMG

Exemplo Prático I

- A função de produção do tipo Cobb-Douglas em geral toma a forma

$$Y = AL^{\alpha}K^{\beta},$$

em que

Y é o nível de produção total

A é a produtividade total dos fatores ("tecnologia")

L é o tamanho da força de trabalho

K é a quantidade de capital investido

Exemplo Prático I

- De acordo com algumas teorias macroeconômicas, é razoável supor que os "parâmetros de produtividade" satisfazem $0 < \alpha < 1$ e $0 < \beta < 1$
  - veremos mais adiante que α e β são na verdade parâmetros de *elasticidade* do produto Y em resposta à uma mudança no fator de produção correspondente

- Uma outra hipótese comum acerca desse modelo é supor que $\alpha + \beta = 1$
  - nesse caso, diz-se que a função de produção é *homogênea* (de grau 1)

- Na sequência, estimaremos α e β através de um modelo de regressão, e testaremos essas hipóteses para o caso brasileiro

Exemplo Prático I

- Para estimarmos uma função de produção do tipo Cobb-Douglas através de um modelo de regressão, primeiramente coletamos uma amostra $\boldsymbol{Y} := (Y_1, \dots, Y_n)^T$, $\boldsymbol{K} := (K_1, \dots, K_n)^T$ e $\boldsymbol{L} = (L_1, \dots, L_n)^T$, onde $n = 144$

- Para cada um dos $i = 1, \dots, n$ países, um modelo de regressão natural para o nosso problema seria, então,

$$Y_i = A L_i^{\alpha} K_i^{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim iid\ N(0, \sigma^2)$$

DEST
UFMG

Exemplo Prático I

- Como o modelo de regressão definido no slide anterior não é um Modelo Linear Normal, ainda não temos ferramentas adequadas para estimá-lo

- Podemos tentar redefinir então nosso modelo como

$$Y_i = A L_i^{\alpha} K_i^{\beta} \cdot \varepsilon_i, \qquad \varepsilon_i \sim LN(0, \sigma^2),$$

onde $LN(\mu, \sigma^2)$ denota uma distribuição log-Normal com log-média μ e log-variância σ^2

DEST UFMG

Exemplo Prático I

- Nessa forma, como $\varepsilon_i^* := \log(\varepsilon_i) \sim N(0, \sigma^2)$, então

$$\log(Y_i) = \log\left(AL_i^{\alpha} K_i^{\beta} \cdot \varepsilon_i\right)$$

$$\Rightarrow \log(Y_i) = \log(A) + \alpha \log(L_i) + \beta \log(K_i) + \log(\varepsilon_i)$$

$$\Rightarrow \log(Y_i) = \log(A) + \alpha \log(L_i) + \beta \log(K_i) + \varepsilon_i^*$$

DEST UFMG

Exemplo Prático I

- O modelo anterior é um modelo de regressão linear Normal na forma "log-log", isto é, tal que

$$\log(Y_i) = \sum_{k=1}^{p+1} \beta_k \log(X_{ik}) + \varepsilon_i, \qquad \varepsilon_i \sim iid\ N(0, \sigma^2)$$

onde no nosso contexto, especificamente, temos

$$\boldsymbol{X}_i \coloneqq (1, X_{i1}, X_{i2})^T = (1, \log(L_i), \log(K_i))^T$$

$$\boldsymbol{\beta} \coloneqq (\beta_0, \beta_1, \beta_2)^T = (\log(A), \alpha, \beta)^T$$

DEST UFMG

Exemplo Prático I

- Na forma de MLG, podemos reescrever o modelo anterior como

$$\log(Y_i)\,|\boldsymbol{X}_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = g^{-1}(\eta_i)$$

$$\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta}$$

DEST UFMG

Exemplo Prático I

- Note que esse modelo é na verdade um modelo *log-linear*, pois o mesmo resultado também vale para $\boldsymbol{X}_i$ (não necessariamente log-transformado)

- Mais especificamente, seja

$$Y_i = \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right) \cdot \varepsilon_i, \qquad \varepsilon_i \sim iid\ LN(0, \sigma^2)$$

DEST UFMG

Exemplo Prático I

- Temos então, mais uma vez que

$$\log(Y_i)\,|\boldsymbol{X}_i \sim N(\mu_i, \sigma^2)$$

- Em particular, o caso log-log pode ser obtido notando que $\exp\left(\log\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right)\right) = \boldsymbol{X}_i^T\boldsymbol{\beta}$, ou ainda tomando $\log(\boldsymbol{X}_i) \coloneqq (1, \log(X_{i1})\,, \dots, \log(X_{n1}))^T$

- Generalizaremos esses resultados mais adiante, quando formos estudar outras funções de ligação no Modelo Normal

DEST
UFMG

Exemplo Prático I

- O modelo anterior é um Modelo Linear Normal, para o qual já sabemos fazer estimação e inferência

- Note que, além disso, a transformação logarítmica também nos permite estimar e realizar inferência sobre um *Modelo Log-Normal* com o mesmo ferramental, uma vez que, de maneira equivalente,

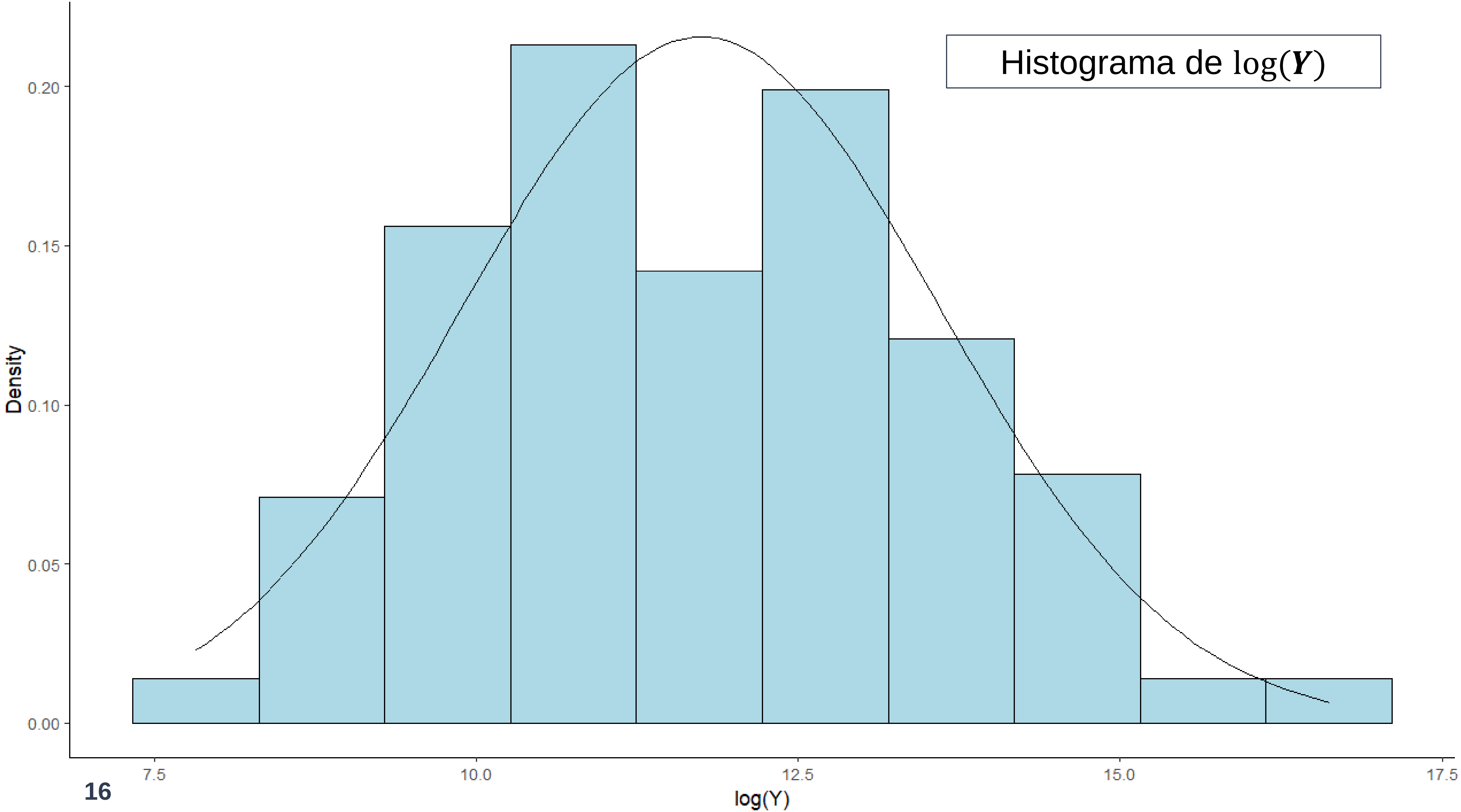
$$Y_i|\boldsymbol{X}_i \sim LN(\mu_i, \sigma^2)$$

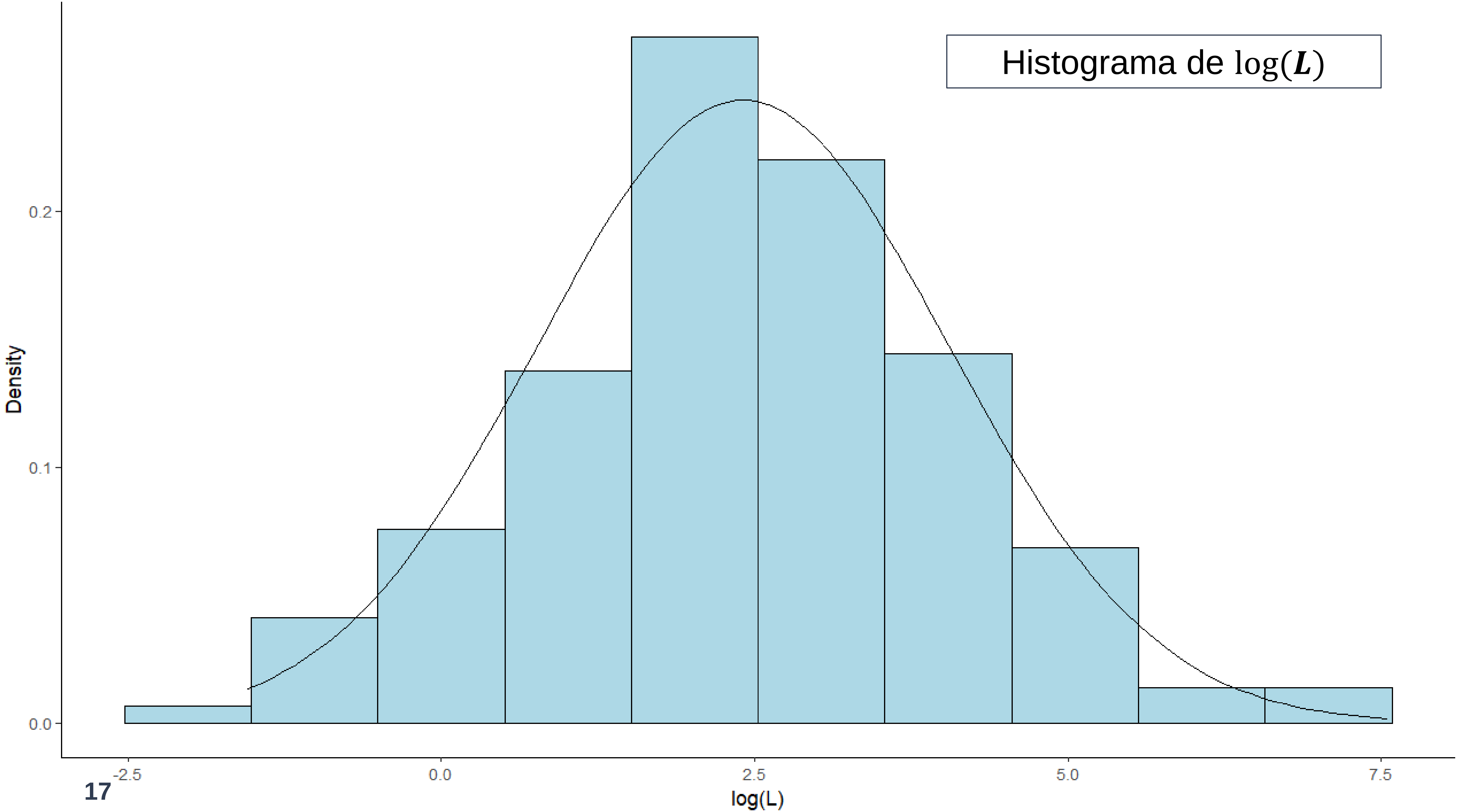
- Esse é um resultado importante tanto do ponto de vista teórico quanto prático, uma vez que a distribuição log-Normal *não está* na Família Exponencial

Descritivas básicas

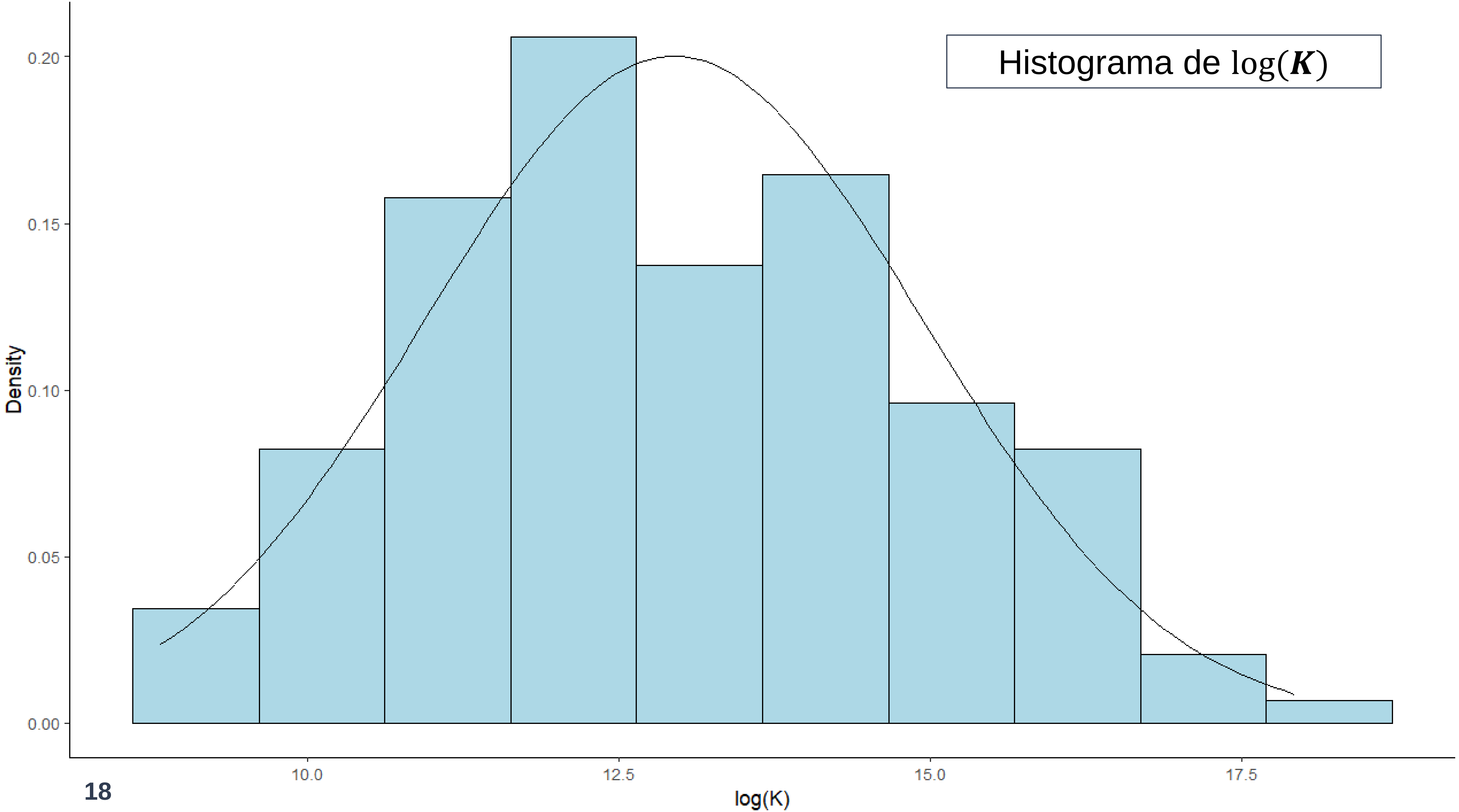
```
       Y                    L                     K           
 Min.   :    2510   Min.   :   0.2128   Min.   :    6791  
 1st Qu.:   28951   1st Qu.:   4.1034   1st Qu.:   91627  
 Median :  111455   Median :  10.0709   Median :  350857  
 Mean   :  673823   Mean   :  50.3454   Mean   : 2642759  
 3rd Qu.:  452522   3rd Qu.:  32.0512   3rd Qu.: 1928975  
 Max.   :16651722   Max.   :1906.5248   Max.   :61035284  
```


DEST UFMG


Histograma de log(Y)
Density
0.00
0.05
0.10
0.15
0.20
7.5
10.0
12.5
15.0
17.5
log(Y)



Histograma de $\log(L)$
Density
0.2
0.1
0.0
-2.5
0.0
2.5
5.0
7.5
log(L)



Histograma de $\log(\boldsymbol{K})$
Density
0.00
0.05
0.10
0.15
0.20
10.0
12.5
15.0
17.5
log(K)

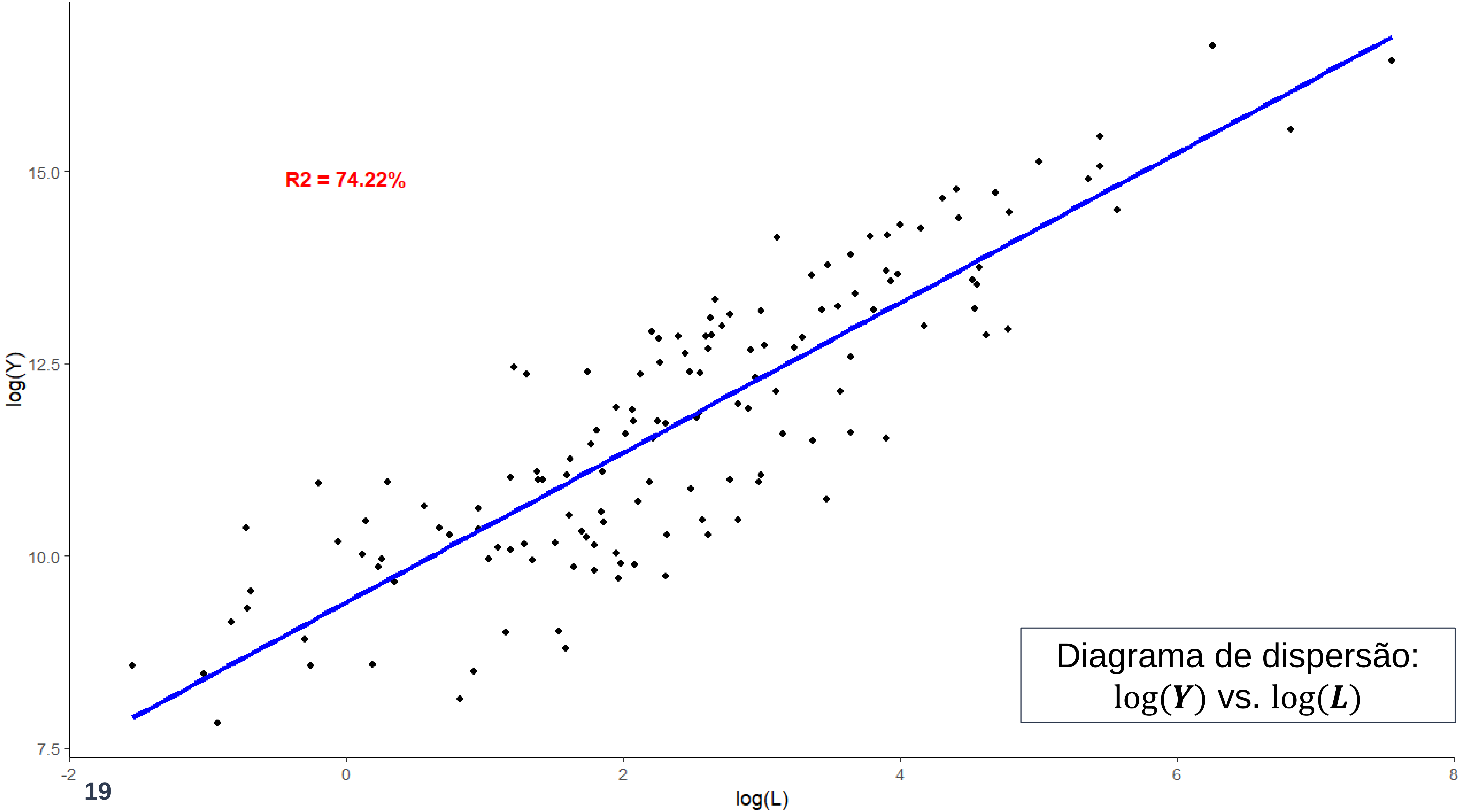
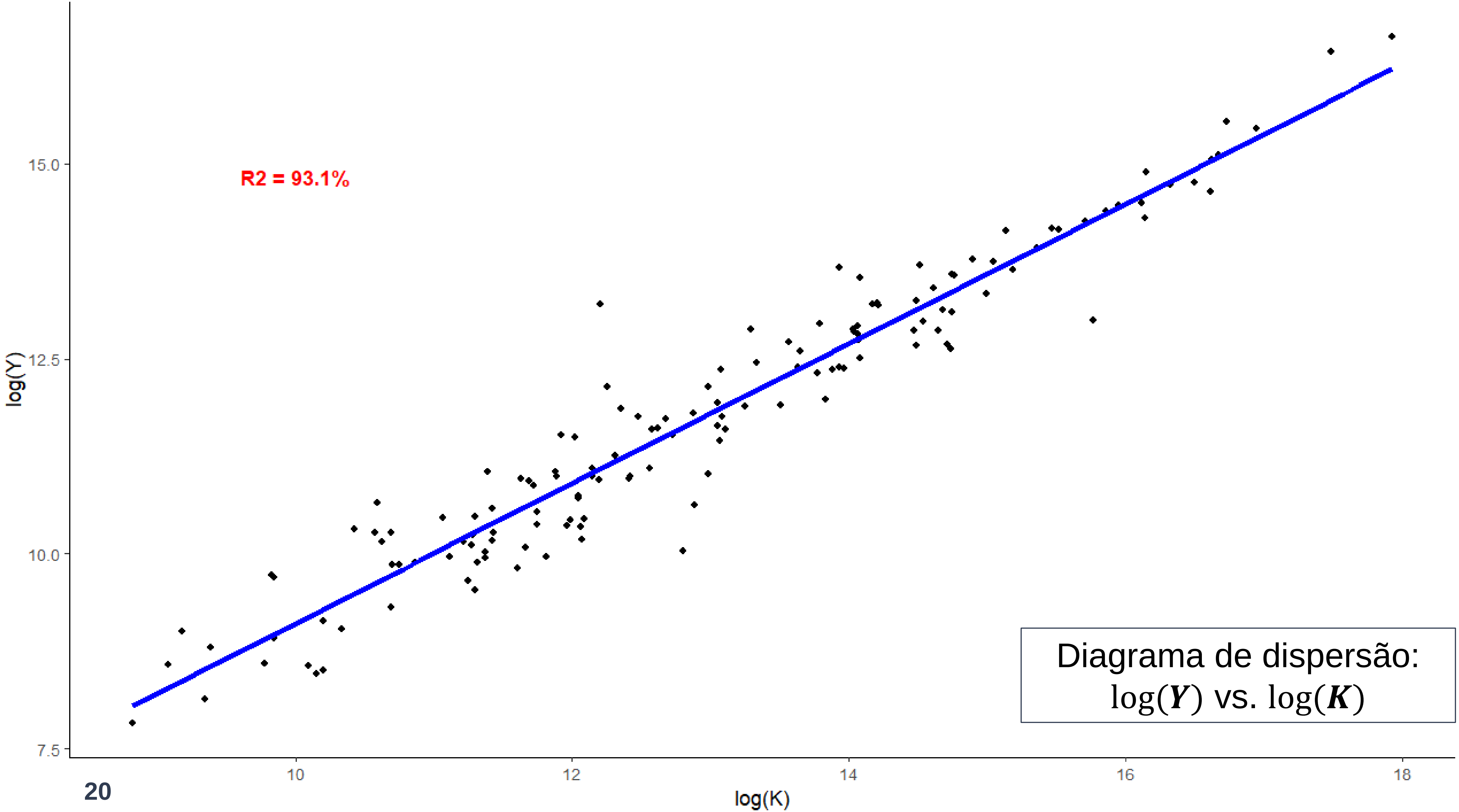


Diagrama de dispersão:
$\log(\boldsymbol{Y})$ vs. $\log(\boldsymbol{L})$


R2 = 93.1%
15.0
12.5
10.0
7.5
log(Y)
10
12
14
16
18
log(K)
Diagrama de dispersão:
log(Y) vs. log(K)

Regressão de $\log(\boldsymbol{Y})$ em $\log(\boldsymbol{L})$ e $\log(\boldsymbol{K})$

```
Call:
lm(formula = log(Y) ~ log(L) + log(K), data = dta)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.47113 -0.16854  0.00069  0.19596  1.68084 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.89912    0.29555   6.426 1.88e-09 ***
log(L)       0.28971    0.03315   8.738 6.28e-15 ***
log(K)       0.70675    0.02729  25.903  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3943 on 141 degrees of freedom
Multiple R-squared:  0.9552,  Adjusted R-squared:  0.9546 
F-statistic:  1504 on 2 and 141 DF,  p-value: < 2.2e-16
```

DEST
UFMG

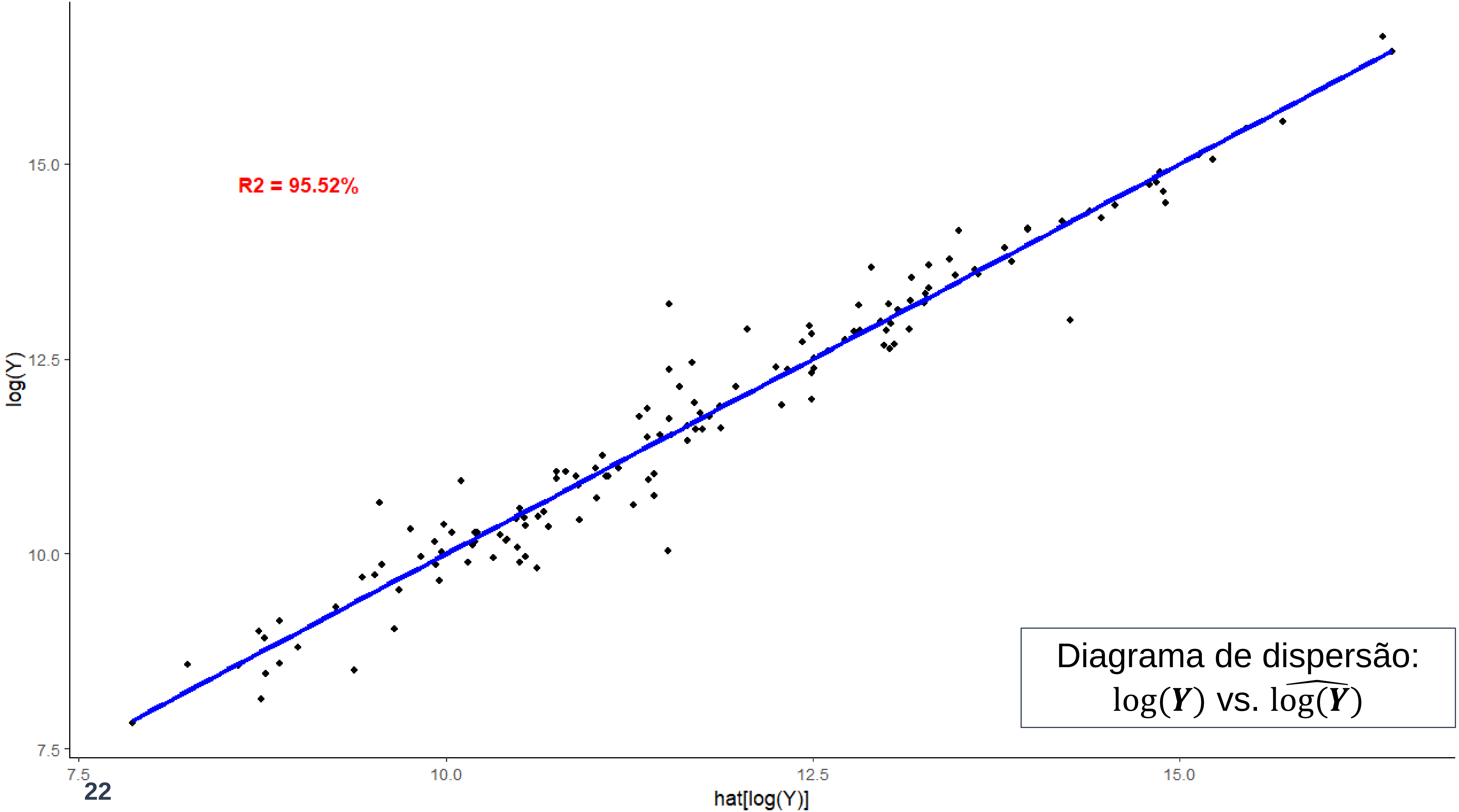


Diagrama de dispersão:
$\log(\boldsymbol{Y})$ vs. $\widehat{\log(\boldsymbol{Y})}$

Teste de $H_0: \alpha + \beta = 1$

```
Linear hypothesis test

Hypothesis:
log(L)  + log(K) = 1

Model 1: restricted model
Model 2: log(Y) ~ log(L) + log(K)

  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    142 21.930
2    141 21.925  1 0.0048175 0.031 0.8605
```

Exemplo Prático I

- Explicaremos agora a razão pela qual α e β (e, em geral, quaisquer coeficientes em um modelo de regressão log-log) são *elasticidades*

- Primeiramente, note que, de $Y_i = AL_i^{\alpha} K_i^{\beta}$, vem

$$\frac{\partial Y_i}{\partial L_i} = \alpha A L_i^{\alpha-1} K_i^{\beta} = \frac{\alpha}{L_i} A L_i^{\alpha} K_i^{\beta} = \frac{\alpha}{L_i} Y_i$$

DEST UFMG

Exemplo Prático I

- Dessa forma, a *variação relativa infinitesimal* de Y_i com respeito à uma variação relativa infinitesimal em L_i é dada por

$$\frac{\frac{\partial Y_i}{Y_i}}{\frac{\partial L_i}{L_i}} = \frac{\partial Y_i}{\partial L_i}\frac{L_i}{Y_i} = \frac{\alpha}{L_i}Y_i\frac{L_i}{Y_i} = \alpha$$

- Assim, uma variação de 1% no tamanho da força de trabalho L_i geraria aproximadamente uma variação de $\alpha\%$ no PIB total Y_i
  - o mesmo vale para qualquer outro coeficiente angular estimado em um modelo de regressão log-log

DEST UFMG

EXEMPLO PRÁTICO II

DEST
UFMG

Exemplo Prático II

- 5 variáveis: total de óbitos por homicídio, população residente, PIB, Gini e taxa de desemprego
  - coletados do sistema TABNET, do DATASUS <https://datasus.saude.gov.br/informacoes-de-saude-tabnet/>

- Disponíveis para todos os 5,565 municípios do Brasil
  - de acordo com o Censo de 2010

DEST UFMG

Exemplo Prático II

- As fontes dos dados são:
  - óbitos por homicídio: *Sistema de Informações sobre Mortalidade* (SIM – Ministério da Saúde)
  - população, Gini, PIB e taxa de desemprego: Censo de 2010 (IBGE)

- Os dados desse exemplo estão disponíveis em diferentes arquivos .csv na pasta “Módulo 1/code/_dta”
  - cada arquivo tem um prefixo correspondente à fonte dos dados: por exemplo, os dados do SIM são nomeados "sim [...].csv"

DEST UFMG

Exemplo Prático II

- Nosso objetivo principal aqui é modelar as *taxas de homicídios* (por 100.000 hab.), $R_i \coloneqq 10^5 \cdot (Y_i/N_i)$, como função de PIB_i, $Gini_i$ e $Desemp_i$

- Aqui, para o i-ésimo município, temos

  Y_i = número de óbitos por homicídio

  N_i = população total

  PIB_i = produto interno bruto (em milhares de reais)

  $Gini_i$ = coeficiente de Gini

  $Desemp_i$ = taxa de desemprego

Exemplo Prático II

- Nosso modelo Poisson para as taxas de mortalidade, na forma de MLG, é dado por

$$Y_i|\boldsymbol{X}_i \sim Poisson(\mu_i)$$

$$\mu_i = \exp(\eta_i)$$

$$\eta_i = O_i + \boldsymbol{X}_i^T \boldsymbol{\beta}$$

onde $\boldsymbol{X}_i \coloneqq (PIB_i, Gini_i, Desemp_i)^T$ e $O_i \coloneqq \log(N_i/10^5)$, $i = 1, \ldots, n$

DEST
UFMG

Exemplo Prático II

- Lembre que, com uma função logarítmica para ligar o preditor linear η_i à média condicional de Y_i, i.e. assumindo que

$$\mu_i = \exp(\eta_i) \Leftrightarrow \eta_i = \log(\mu_i)\,,$$

o papel do offset $O_i \coloneqq \log(N_i/10^5)$ é fazer com que o modelo (que é um modelo de *contagem*) esteja definido corretamente para as *taxas* (nesse caso específico, por 100.000 habitantes)

Exemplo Prático II

- Em detalhes,

$$\log(\mathbb{E}[R_i|\boldsymbol{X}_i]) = \log\left(\mathbb{E}\left[\frac{Y_i}{N_i}|\boldsymbol{X}_i\right]\right) = \log(\mathbb{E}[Y_i|\boldsymbol{X}_i]) - \log(N_i)$$

$$\Leftrightarrow \log(\mathbb{E}[Y_i|\boldsymbol{X}_i]) = \log(\mathbb{E}[R_i|\boldsymbol{X}_i]) + \log(N_i)$$

Exemplo Prático II

e, como $\mu_i \coloneqq \mathbb{E}[Y_i | \boldsymbol{X}_i]$ e $O_i \coloneqq \log(N_i)$, segue que

$$\log(\mu_i) = O_i + \boldsymbol{X}_i^T \boldsymbol{\beta}$$

Exemplo Prático II

- Dessa forma, o *preditor linear das contagens* é

$$\eta_i = \log(\mu_i) = O_i + \boldsymbol{X}_i^T \boldsymbol{\beta}$$

e a esperança condicional das contagens é dada por

$$\mu_i = \exp(\eta_i) = \exp\left(O_i + \boldsymbol{X}_i^T \boldsymbol{\beta}\right) = N_i \cdot \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right)$$

Exemplo Prático II

- Por outro lado, o *preditor linear das taxas* é apenas

$$\boldsymbol{X}_i^T \boldsymbol{\beta}$$

uma vez que a esperança condicional das taxas é

$$\frac{\mu_i}{N_i} = \frac{\exp(\eta_i)}{N_i} = \frac{\exp\left(O_i + \boldsymbol{X}_i^T \boldsymbol{\beta}\right)}{N_i} = \frac{N_i \cdot \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right)}{N_i} = \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right)$$

DEST
UFMG

Exemplo Prático II

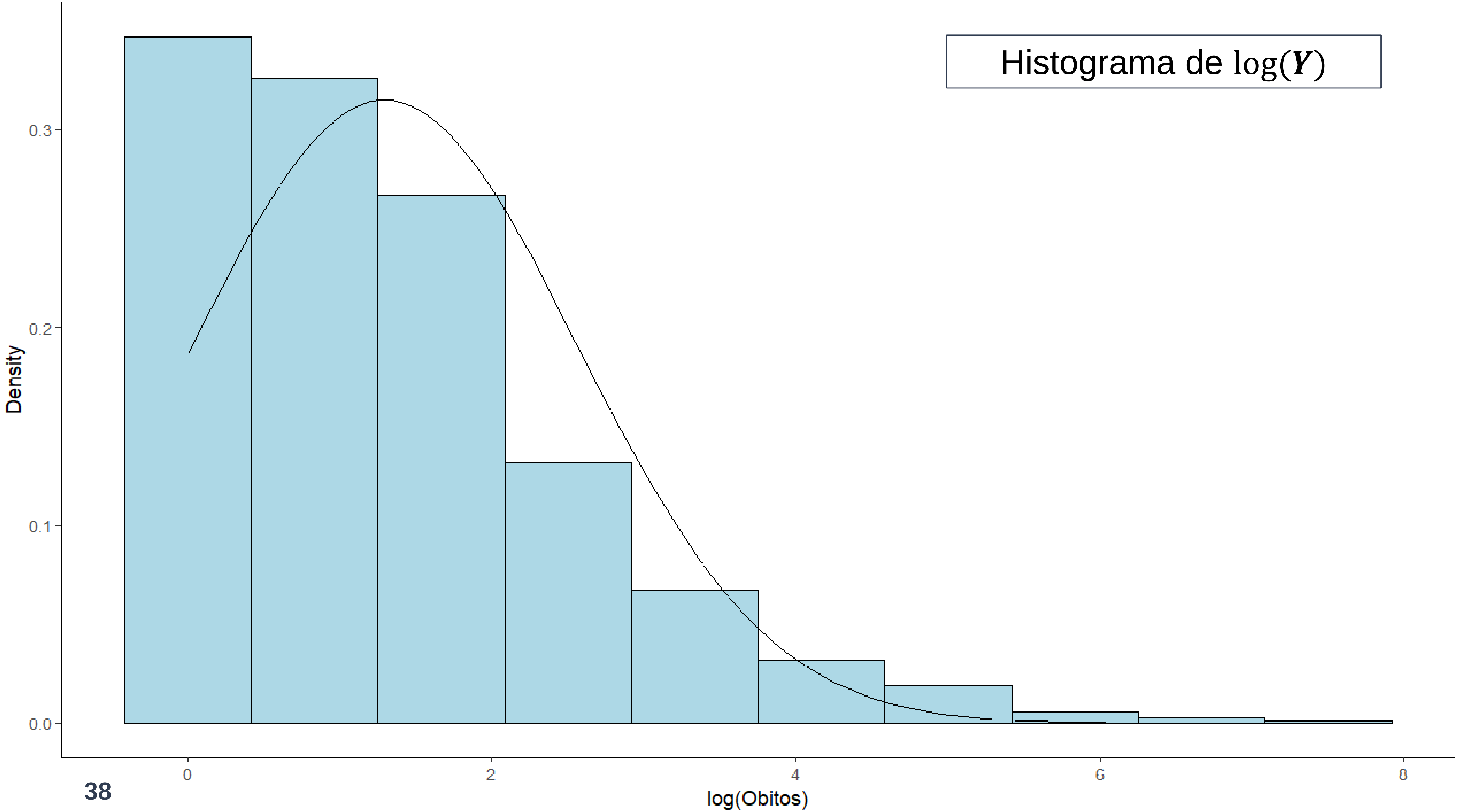
- Finalmente, note que o fator de normalização da população (100.000) foi omitido da expressão anterior

- Tais fatores podem ser sempre inclusos diretamente no offset, uma vez que, como,

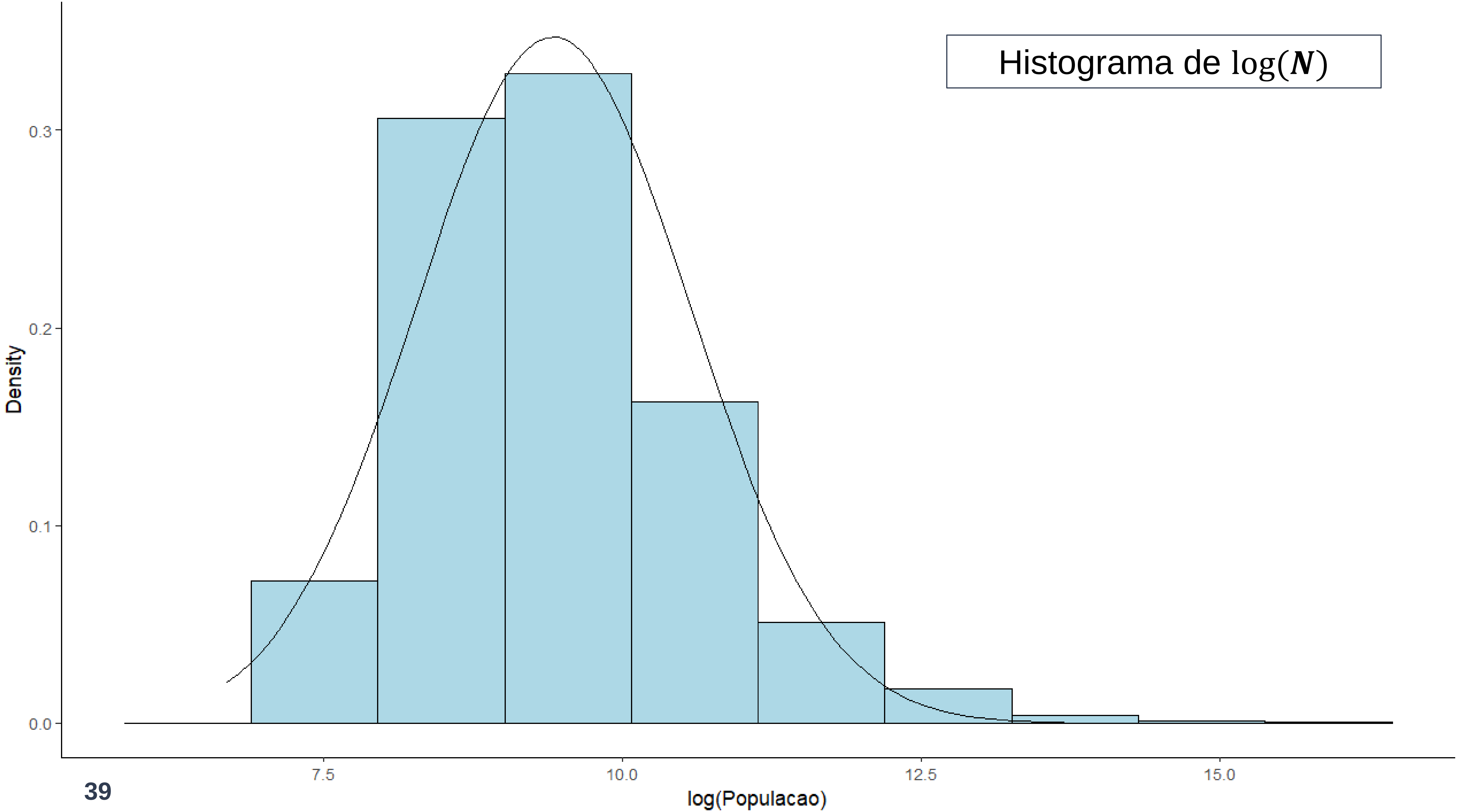
$$R_i = 10^5 \cdot \frac{Y_i}{N_i} = \frac{Y_i}{N_i/10^5},$$

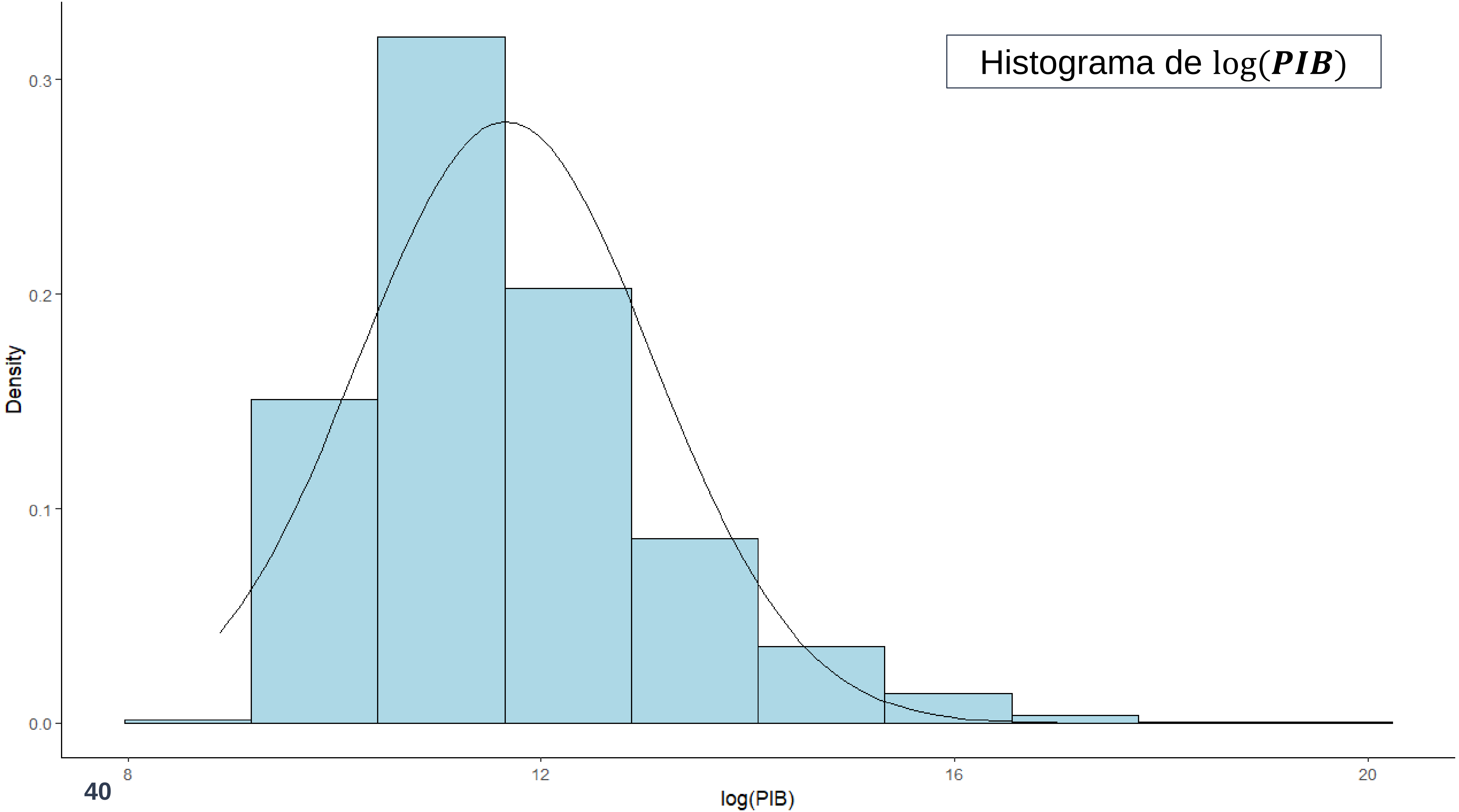
basta então tomar $O_i \coloneqq \log(N_i/10^5)$

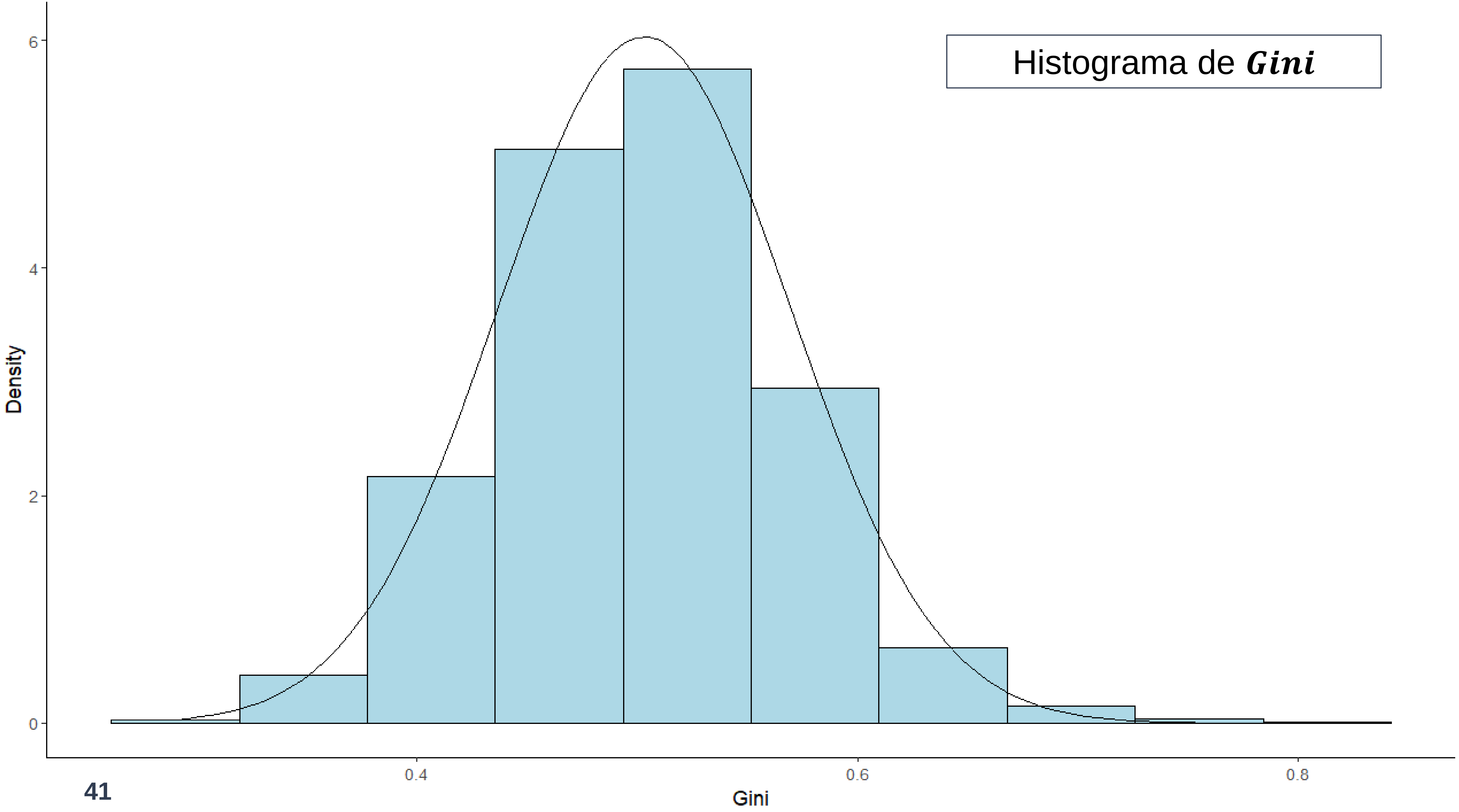
```
    Obitos          Populacao              PIB                  Gini       
Min.   :   1.00   Min.   :    1096   Min.   :    10041   Min.   :0.2907  
1st Qu.:   1.00   1st Qu.:    9072   1st Qu.:    64194   1st Qu.:0.4742  
Median :   3.00   Median :   17544   Median :   140952   Median :0.5145  
Mean   :  13.79   Mean   :   49690   Mean   :  1044734   Mean   :0.5152  
3rd Qu.:   7.00   3rd Qu.:   34366   3rd Qu.:   398007   3rd Qu.:0.5538  
Max.   :1811.00   Max.   :11253503   Max.   :446958815   Max.   :0.8082  

  Tx_Desemp          Tx_Mort         PIB_per_capita   
Min.   : 0.100   Min.   :  1.664   Min.   :  2.263  
1st Qu.: 4.540   1st Qu.: 11.474   1st Qu.:  5.065  
Median : 6.420   Median : 19.395   Median :  9.113  
Mean   : 6.924   Mean   : 24.446   Mean   : 12.798  
3rd Qu.: 8.670   3rd Qu.: 31.712   3rd Qu.: 15.315  
Max.   :29.410   Max.   :153.787   Max.   :311.919  
```

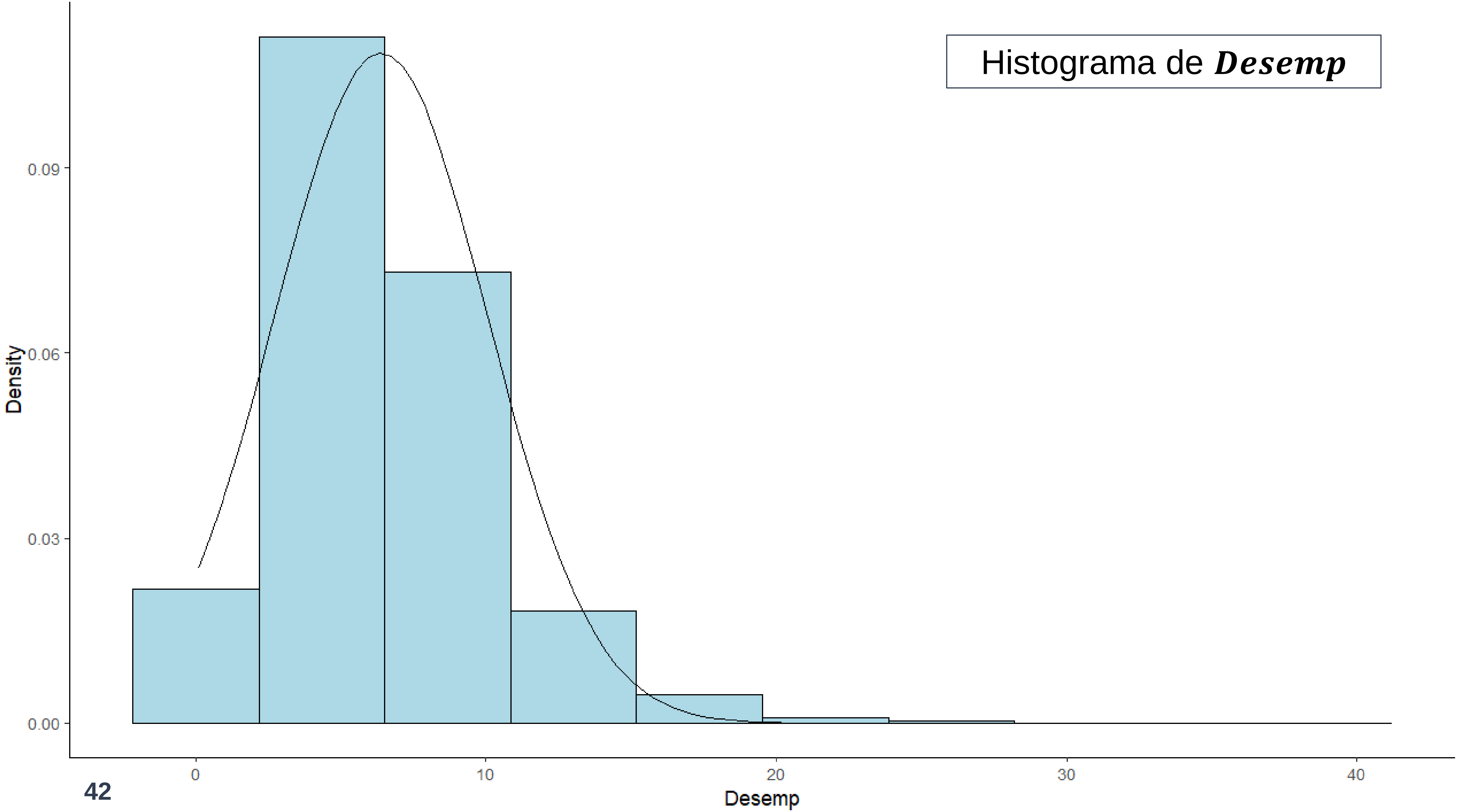


Histograma de $\log(\boldsymbol{Y})$
Density
0.3
0.2
0.1
0.0
0
2
4
6
8
log(Obitos)


Histograma de $\log(N)$
Density
0.0
0.1
0.2
0.3
7.5
10.0
12.5
15.0
log(Populacao)


Histograma de $\log(\boldsymbol{PIB})$
Density
0.0
0.1
0.2
0.3
8
12
16
20
log(PIB)


Histograma de $Gini$
Density
Gini
0
2
4
6
0.4
0.6
0.8


Histograma de $Desemp$
Density
0.09
0.06
0.03
0.00
0
10
20
30
40
Desemp

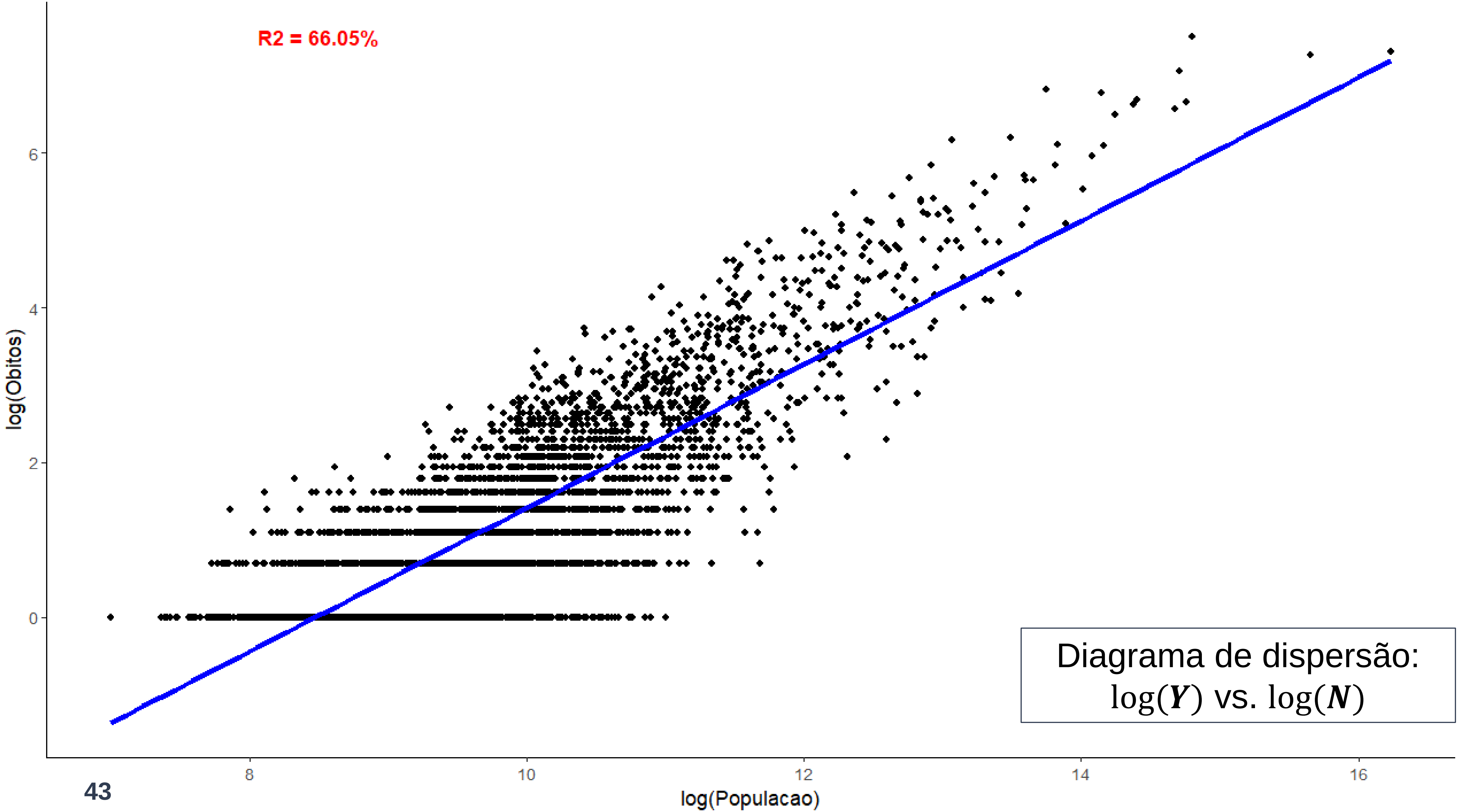
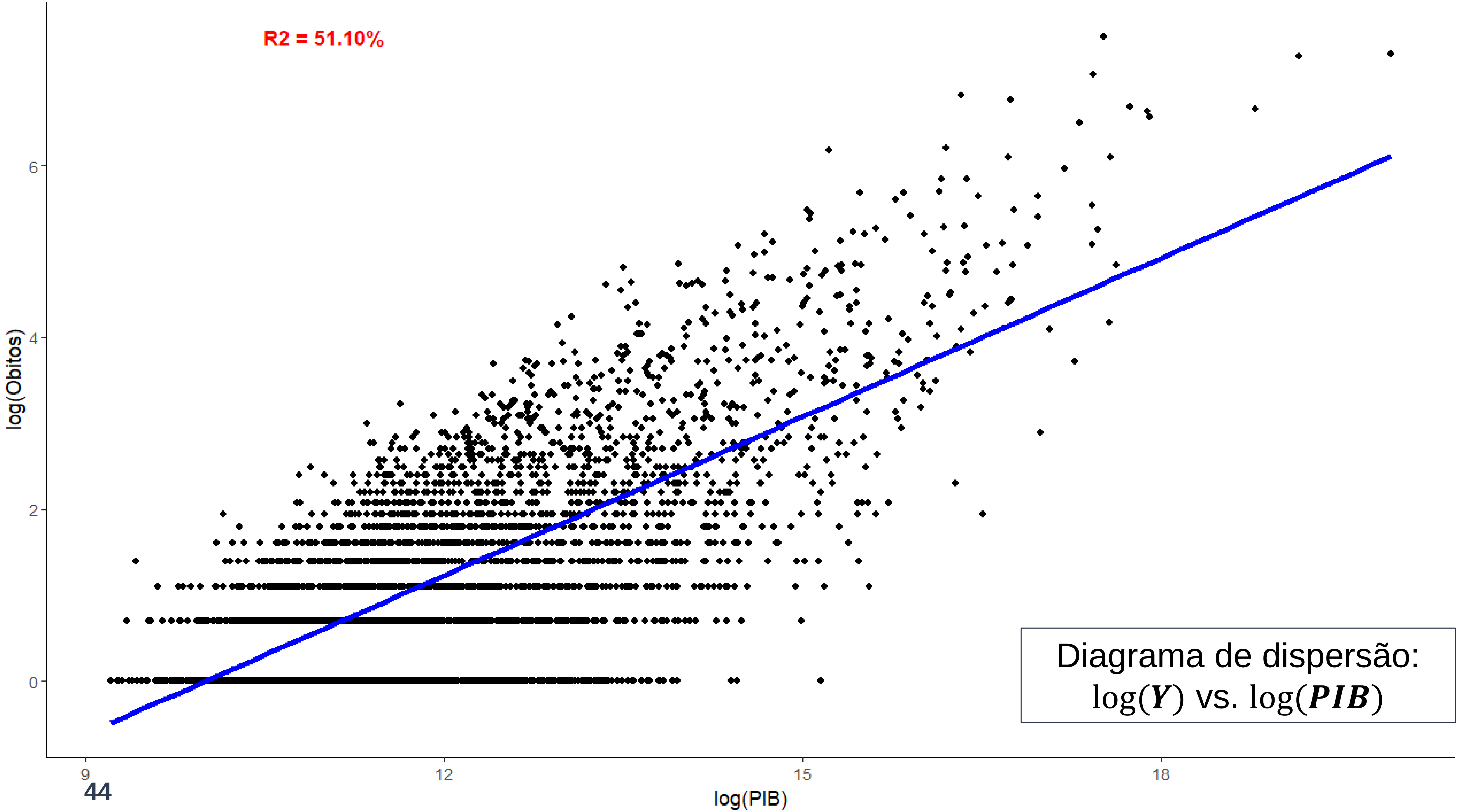


Diagrama de dispersão:
$\log(\boldsymbol{Y})$ vs. $\log(\boldsymbol{N})$


R2 = 51.10%
log(Obitos)
log(PIB)
0
2
4
6
9
12
15
18
Diagrama de dispersão:
log(Y) vs. log(PIB)

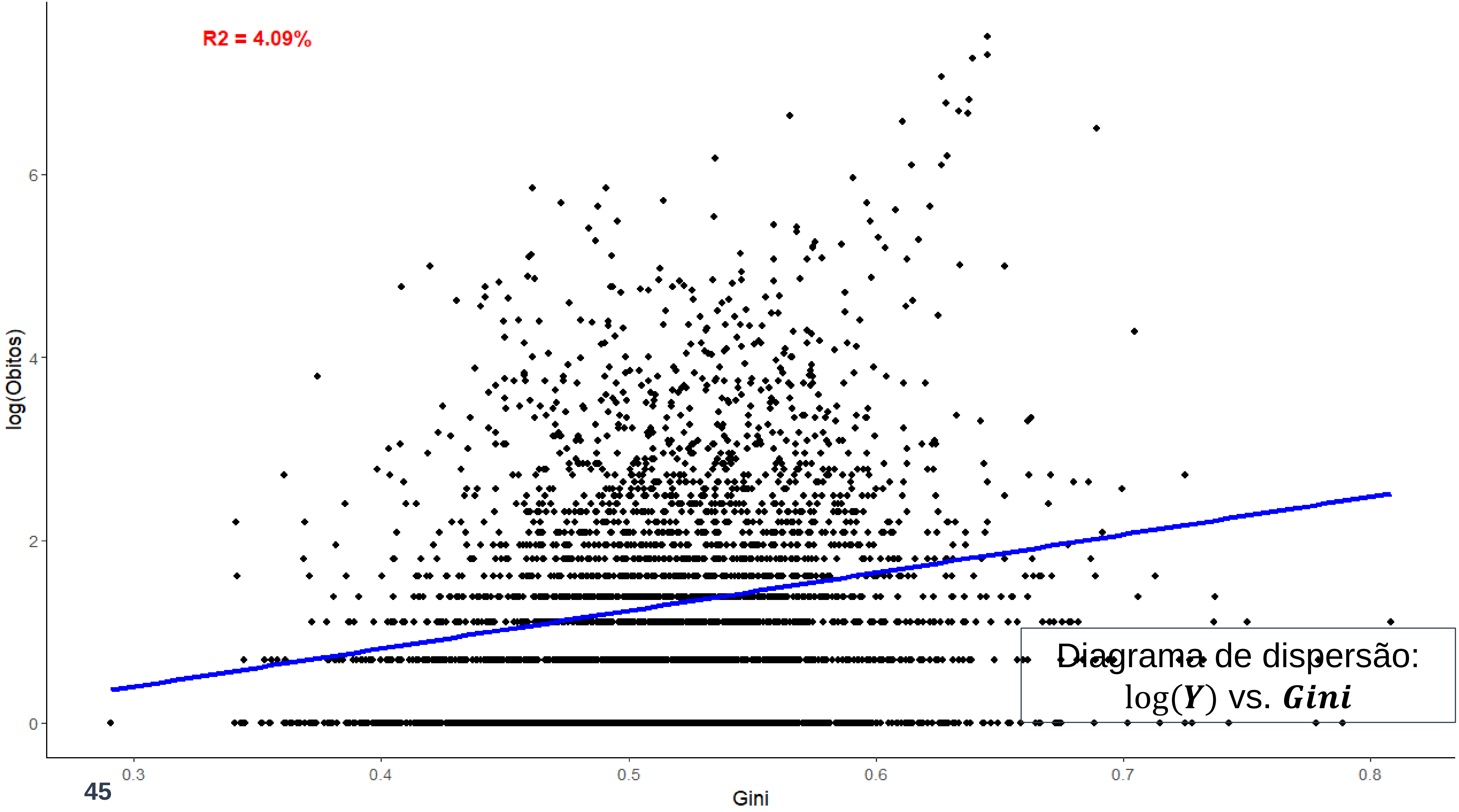
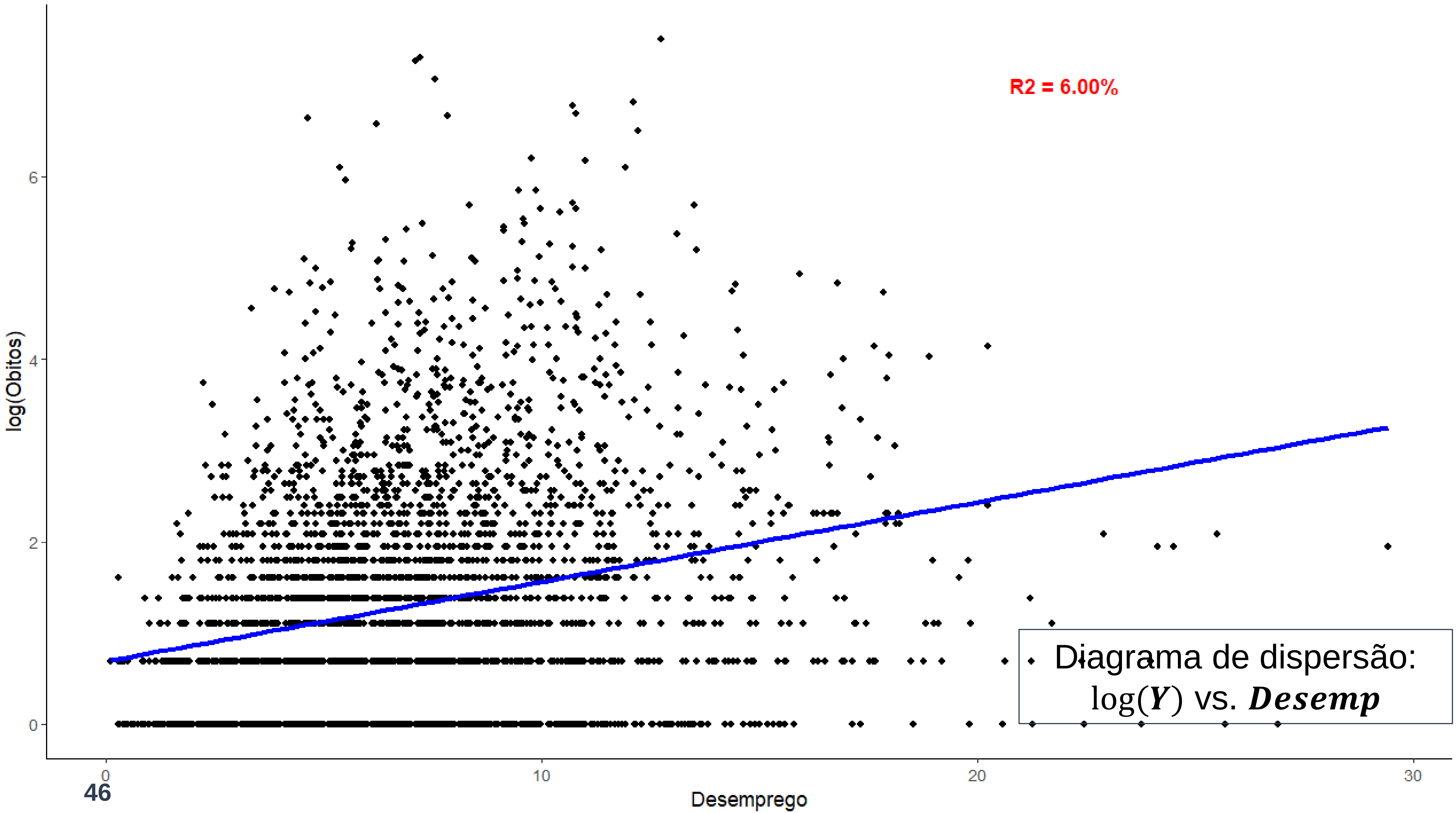


Diagrama de dispersão:
$\log(\boldsymbol{Y})$ vs. $\boldsymbol{Gini}$


R2 = 6.00%
log(Obitos)
0
2
4
6
0
10
20
30
Desemprego
Diagrama de dispersão:
$\log(Y)$ vs. $\boldsymbol{Desemp}$

Resultados MLG Poisson

```
Call:
glm(formula = Obitos ~ log_PIB_per_capita + Gini + Tx_Desemp,
    family = poisson(link = log), data = dta, offset = model.offset)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.226606   0.040194  55.396  < 2e-16 ***
log_PIB_per_capita -0.043269   0.006147  -7.039 1.93e-12 ***
Gini                1.006104   0.070270  14.318  < 2e-16 ***
Tx_Desemp           0.080911   0.001422  56.896  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23153  on 3555  degrees of freedom
Residual deviance: 19355  on 3552  degrees of freedom
AIC: 30774

Number of Fisher Scoring iterations: 5
```


DEST UFMG

Riscos Relativos

	Estimate	2.5 %	97.5 %
(Intercept)	9.2683531	8.5658963	10.0276655
log_PIB_per_capita	0.9576541	0.9461885	0.9692633
Gini	2.7349245	2.3831087	3.1388636
Tx_Desemp	1.0842747	1.0812518	1.0872960

REFERÊNCIAS BIBLIOGRÁFICAS

Referências Bibliográficas

- Paula, G. A. (2023). *Modelos de regressão: com apoio computacional*. São Paulo: IME-USP. Disponível em https://www.ime.usp.br/~giapaula/textoregressao.htm

- Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" *American Economic Review*, 105(10), 3150-3182, available for download at www.gddc.net/pwt

DEST
UFMG