Regressão Linear e Mínimos Quadrados aplicados às notas do IDEB

Álgebra Linear

Hugo de Araújo, Marcéli Melchiors

Novembro, 2021

Fundação Getúlio Vargas - Escola de Matemática Aplicada (FGV/EMAp)

Introdução

- O trabalho consiste em um modelo linear para as notas do IDEB de cada unidade federativa brasileira.
- Uso de regressão linear feita pelo método de mínimos quadrados.
- Cálculo do coeficiente de determinação (R^2) para verificar se o modelo se ajusta bem aos dados.

Sobre os métodos

- Modelo: $b_i = Dt_i + C$, onde b_i é a nota em função de t_i, o ano. Tendo os pontos (t_i, b_i), a solução utilizando mínimos quadrados, sendo erro = e, é dada por:
- $e = \sum_{i=1}^{n} (y(t_i) - b_i)^2 = \sum_{i=1}^{n} (C + Dt_i - b_i)^2$
- $A = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$ $b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ $\hat{x} = \begin{bmatrix} C \\ D \end{bmatrix}$
- $e = ||Ax - b||^2 \Rightarrow \hat{x} = (A^T A)^{-1} A^T b \Rightarrow A^T A \hat{x} = A^T b$
- $A^T A = \begin{bmatrix} m & \sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} t_i & \sum_{i=1}^{n} t_i^2 \end{bmatrix}$, $A^T b = \begin{bmatrix} \sum_{i=1}^{n} b_i \\ \sum_{i=1}^{n} t_i b_i \end{bmatrix}$
- $R^2 = \dfrac{SEQ_m - SEQ}{SEQ_m}$.

Base de dados: IDEB

- O Índice de Desenvolvimento da Educação Básica (IDEB) é um indicador social da qualidade da educação brasileira criado em 2005 pelo Ministério da Educação.
- As notas variam de 0 a 10 e são computadas com base no fluxo escolar e nas provas do Saeb (Sistema de Avaliação da Educação Básica).
- O IDEB serve para nortear as políticas públicas e encontrar exemplos de sucesso na educação brasileira.
- Para este trabalho, utilizou-se as notas do IDEB para o 5º ano do Ensino Fundamental de 2005 a 2019, no nível estadual.

Ánálise dos Dados

- Primeiramente, plotou-se um gráfico de dispersão entre o IDEB do estado e o tempo, em anos.
- Pela diferença da escala das variáveis, o logaritmo natural foi aplicado nas variáveis e mais um plot foi feito.
- Por fim, plotou-se um terceiro gráfico contendo a regressão encontrada pelo método dos mínimos quadrados e outra encontrada pela biblioteca Scikit-Learn.
- As regressões foram feitas para as 27 unidades federativas.

Gráfico do Acre

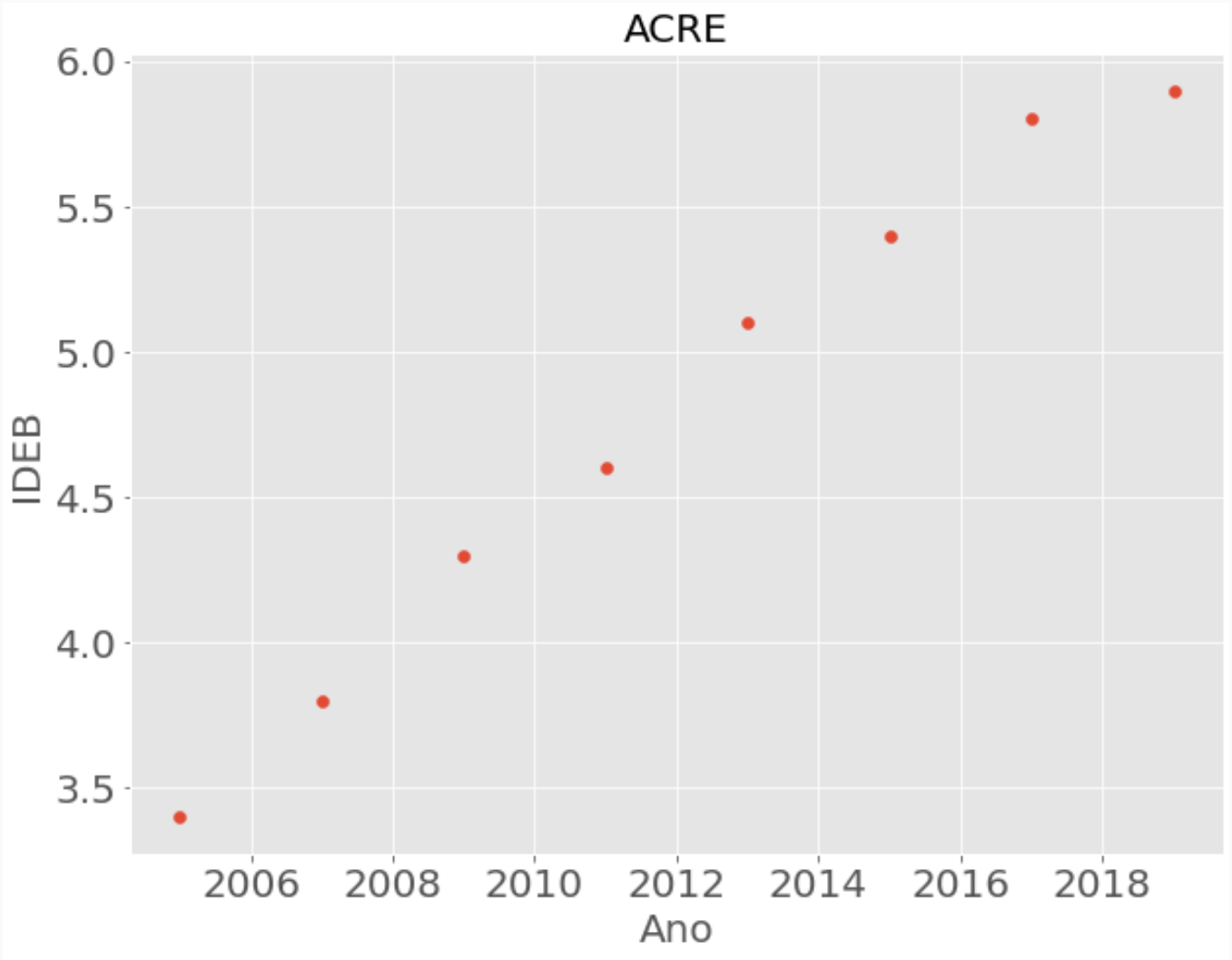

Figura 1: Gráfico de Dispersão para o estado do Acre.

Gráfico modelado do Acre

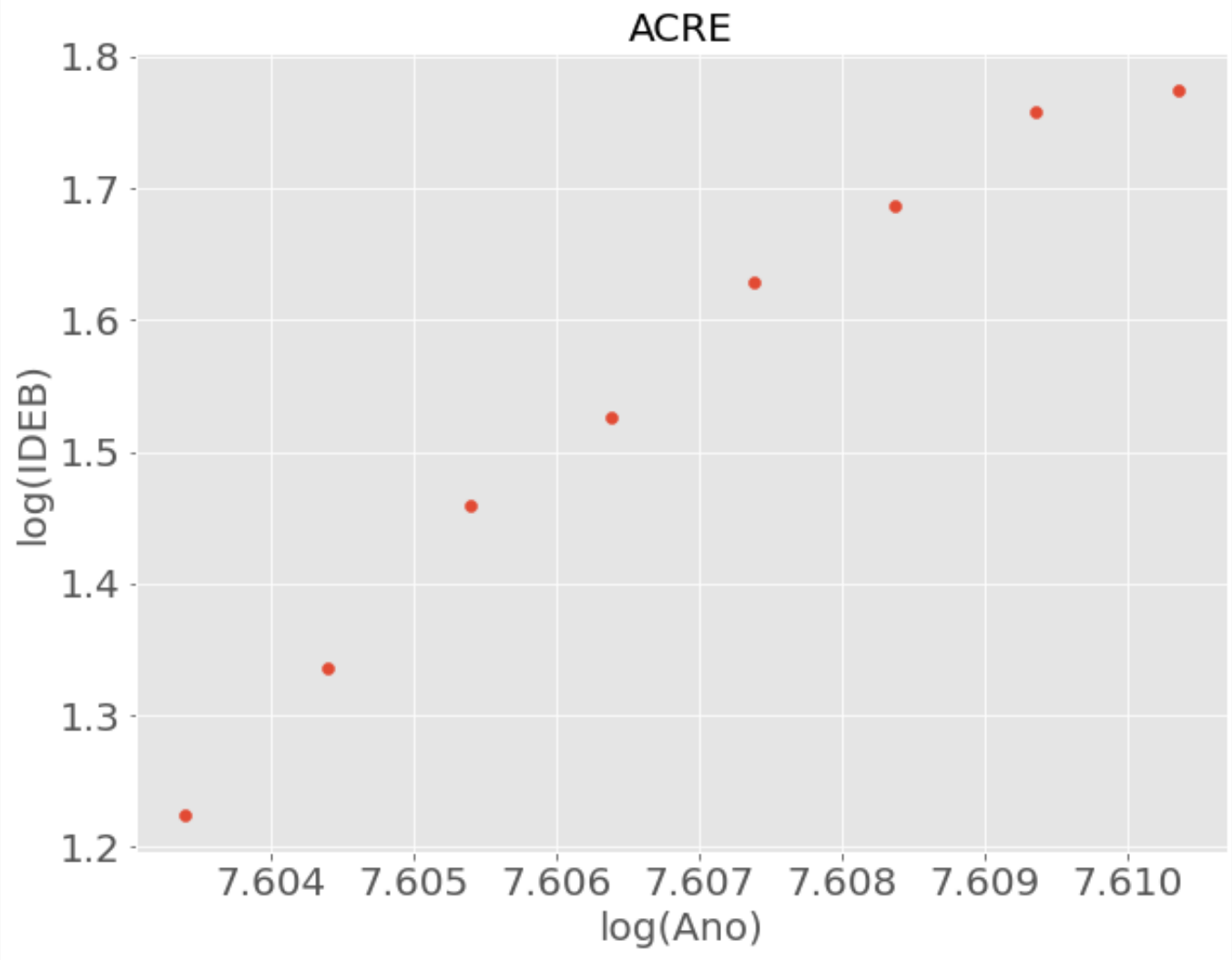

Figura 2: Gráfico de Dispersão para o estado do Acre contendo o logaritmo natural.

Gráfico do Acre com eixos do NumPy e Scikit-Learn

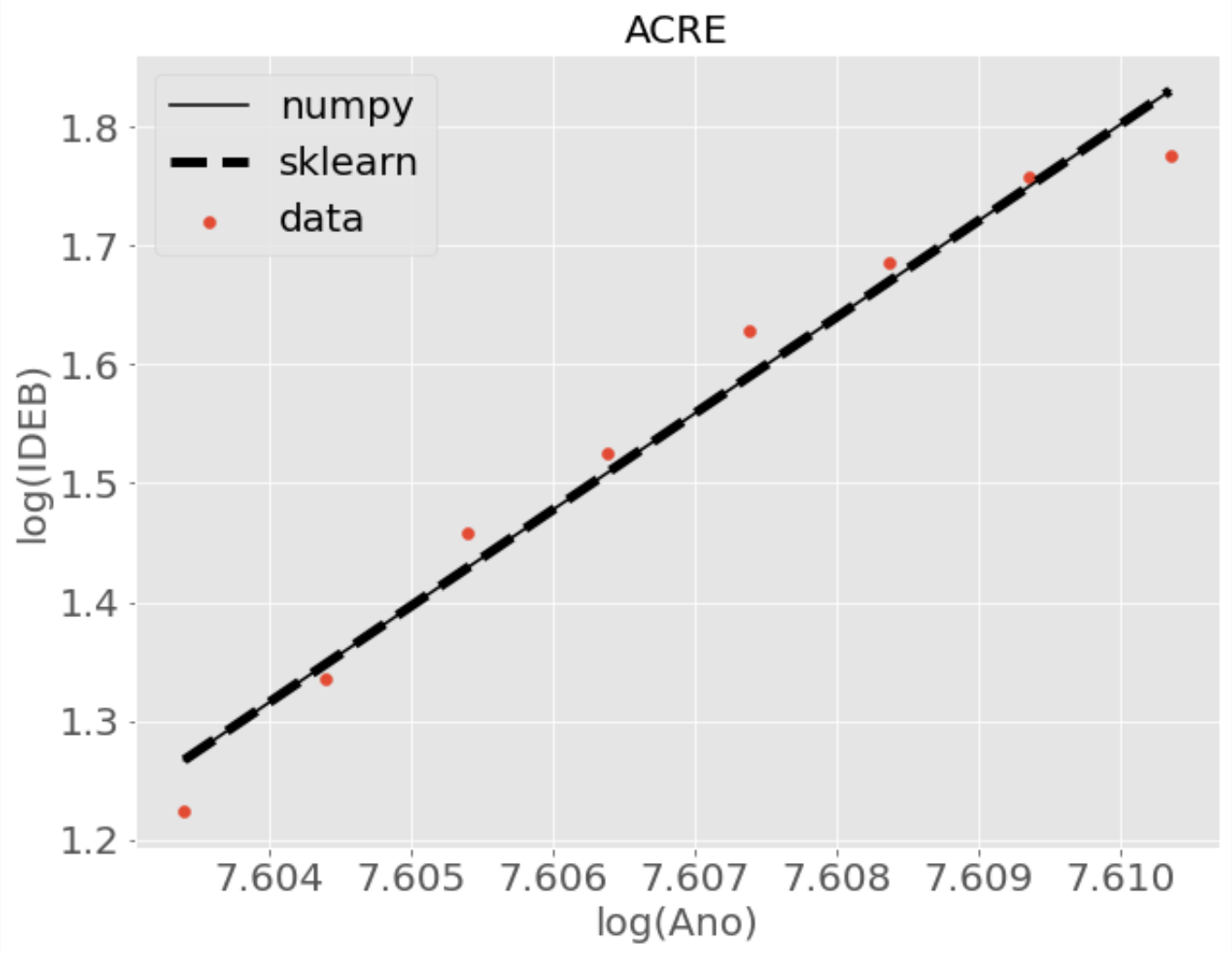

Figura 3: Gráfico de Dispersão para o estado do Acre comparando as regressões pela fórmula e pelo Scikit-Learn.

Gráficos: Alagoas, Amapá e Amazonas

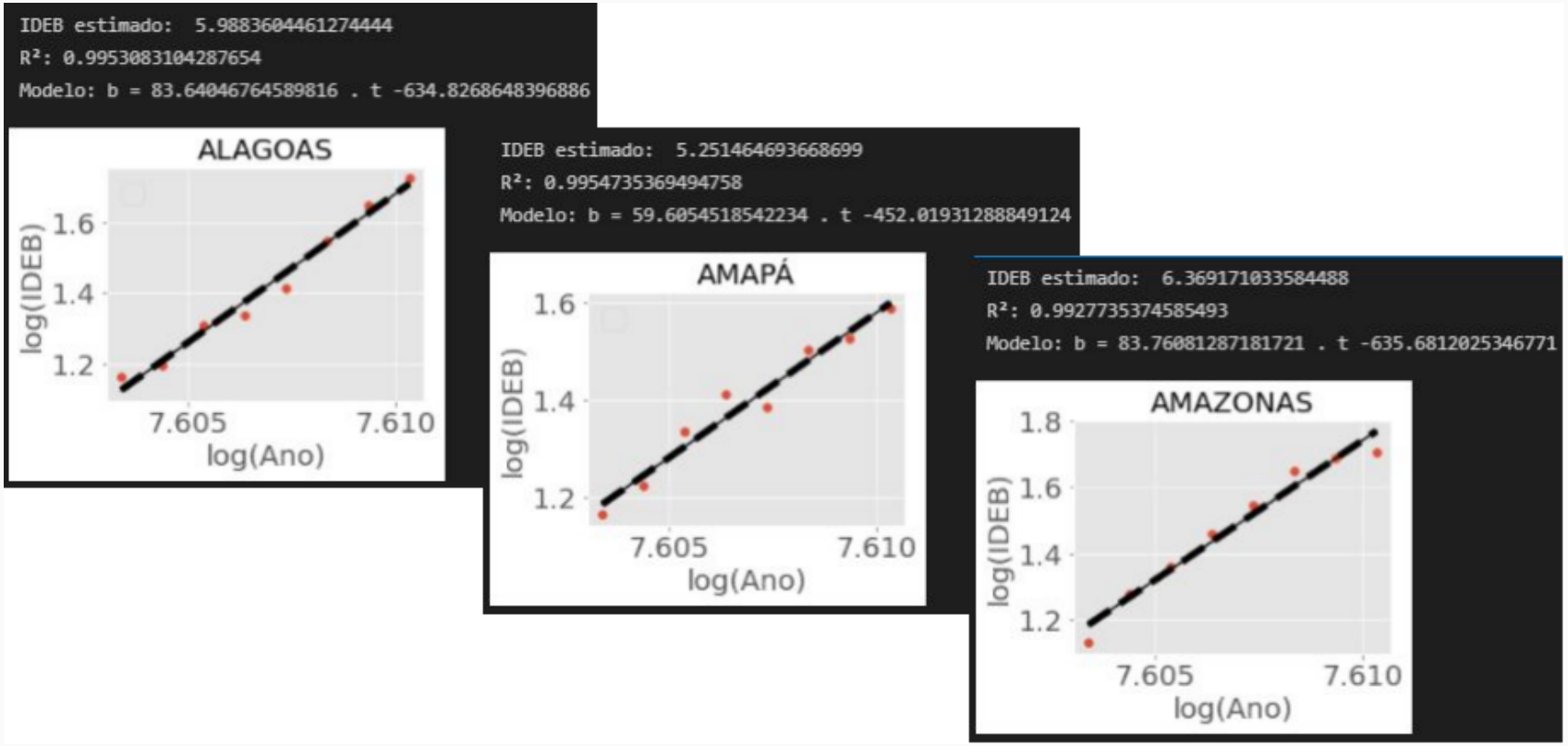

Figura 4: Gráfico de Dispersão para os estados de Alagoas, Amapá e Amazonas.

Gráficos: Bahia, Ceará e Distrito Federal

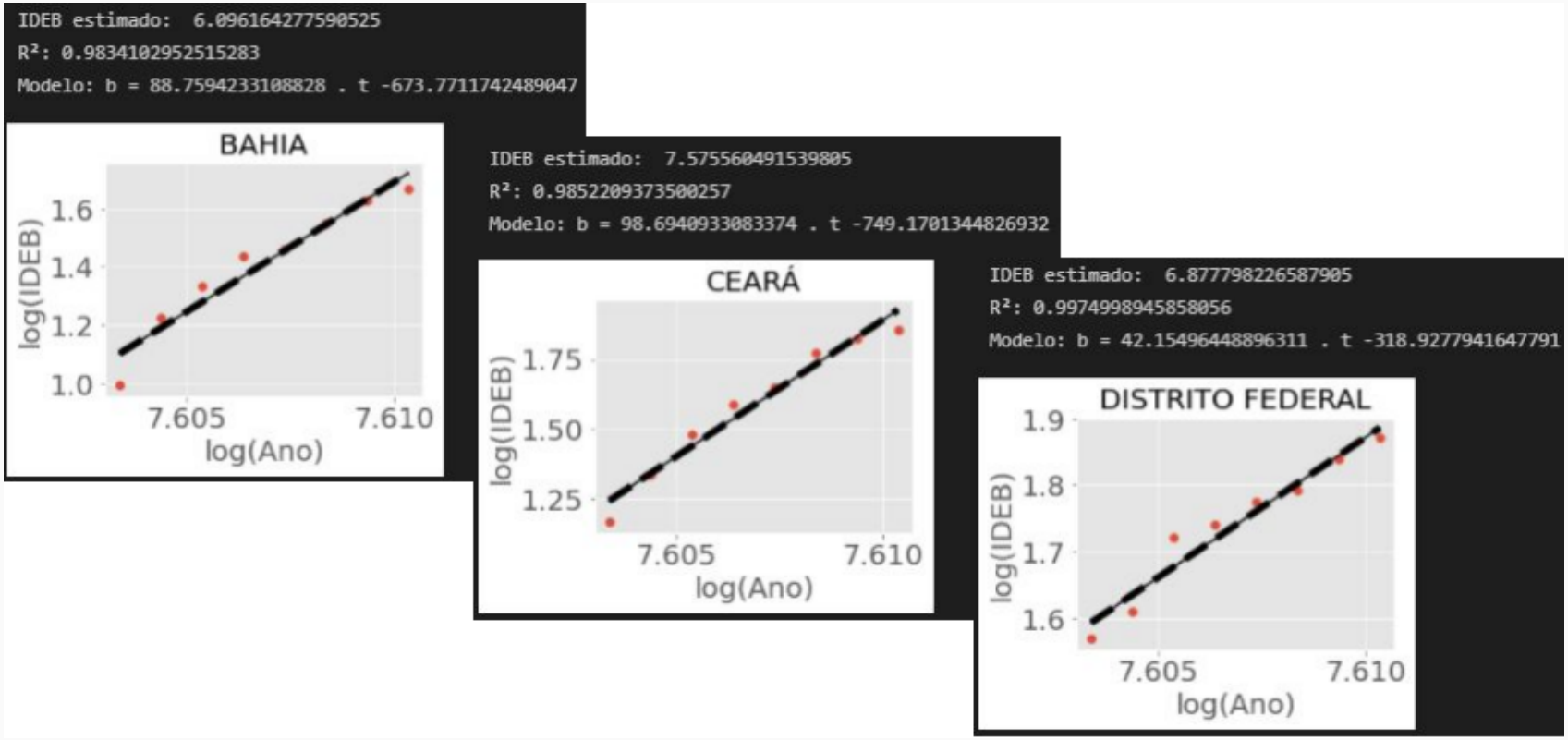

Figura 5: Gráfico de Dispersão para os estados da Bahia, Ceará e o Distrito Federal.

Gráficos: Espiríto Santo, Goiás e Maranhão

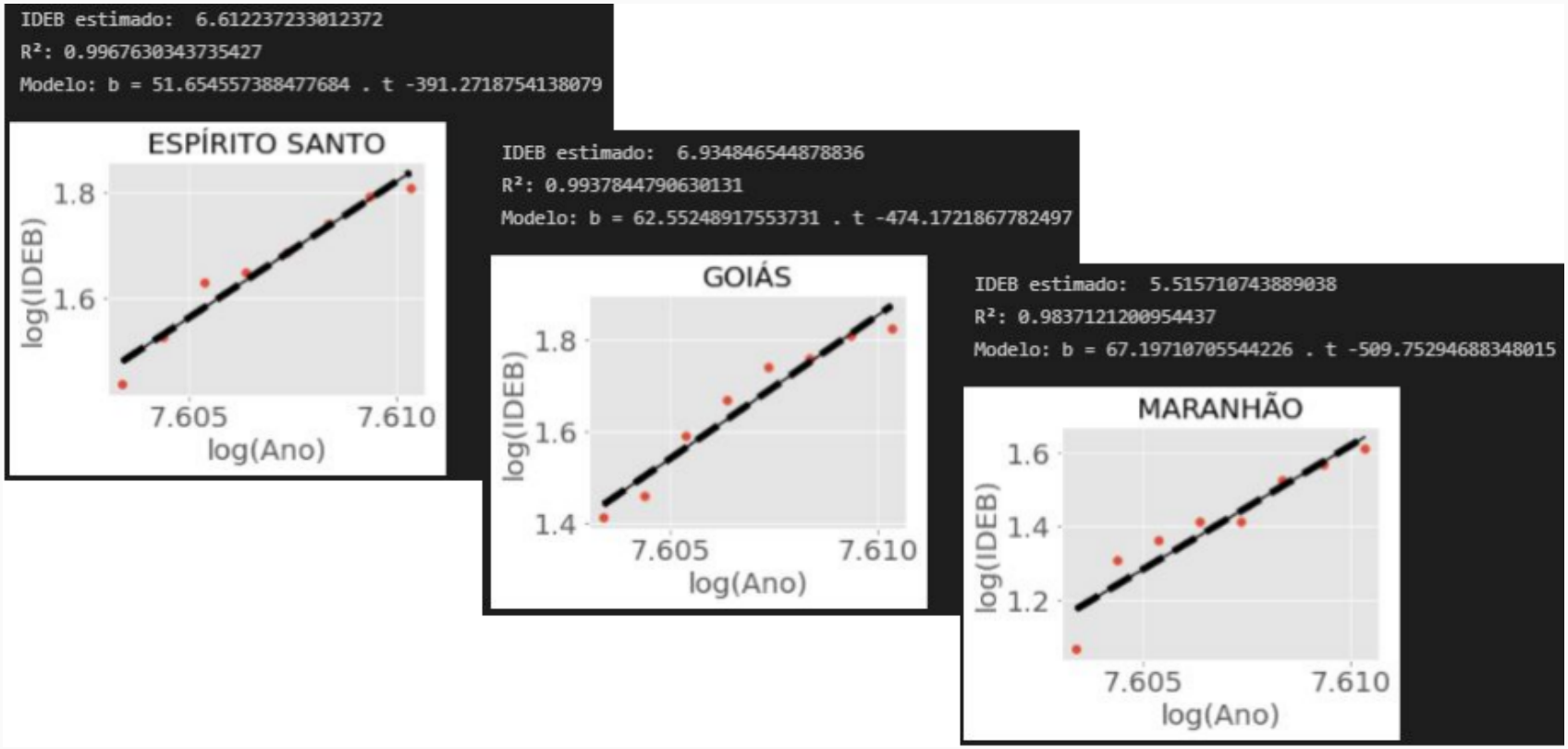

Figura 6: Gráfico de Dispersão para os estados do Espiríto Santo, Goiás e Maranhão.

Gráficos: Mato Grosso, Mato Grosso do Sul e Minas Gerais

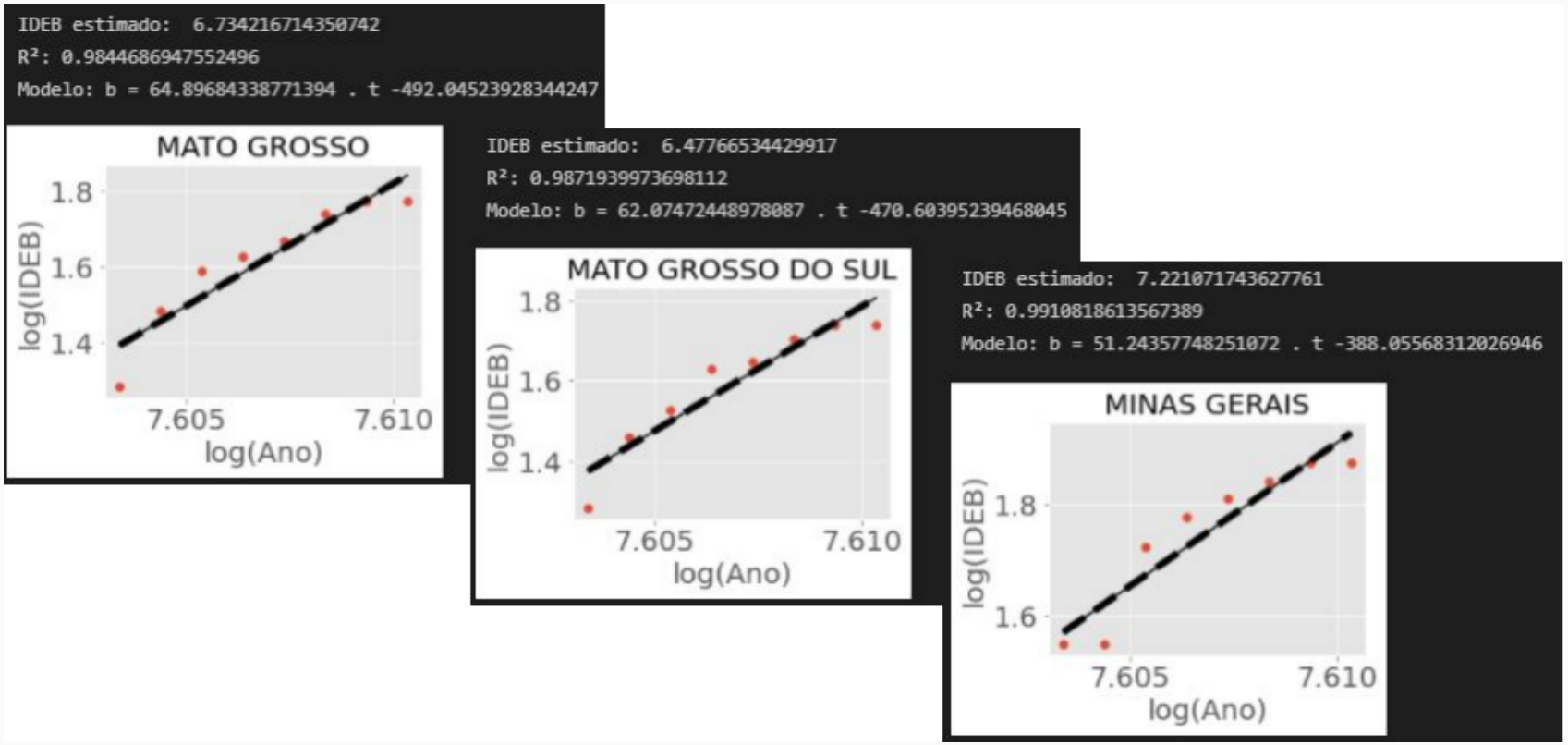

Figura 7: Gráfico de Dispersão para os estados do Mato Grosso, Mato Grosso do Sul e Minas Gerais.

Gráficos: Pará, Paraíba e Paraná

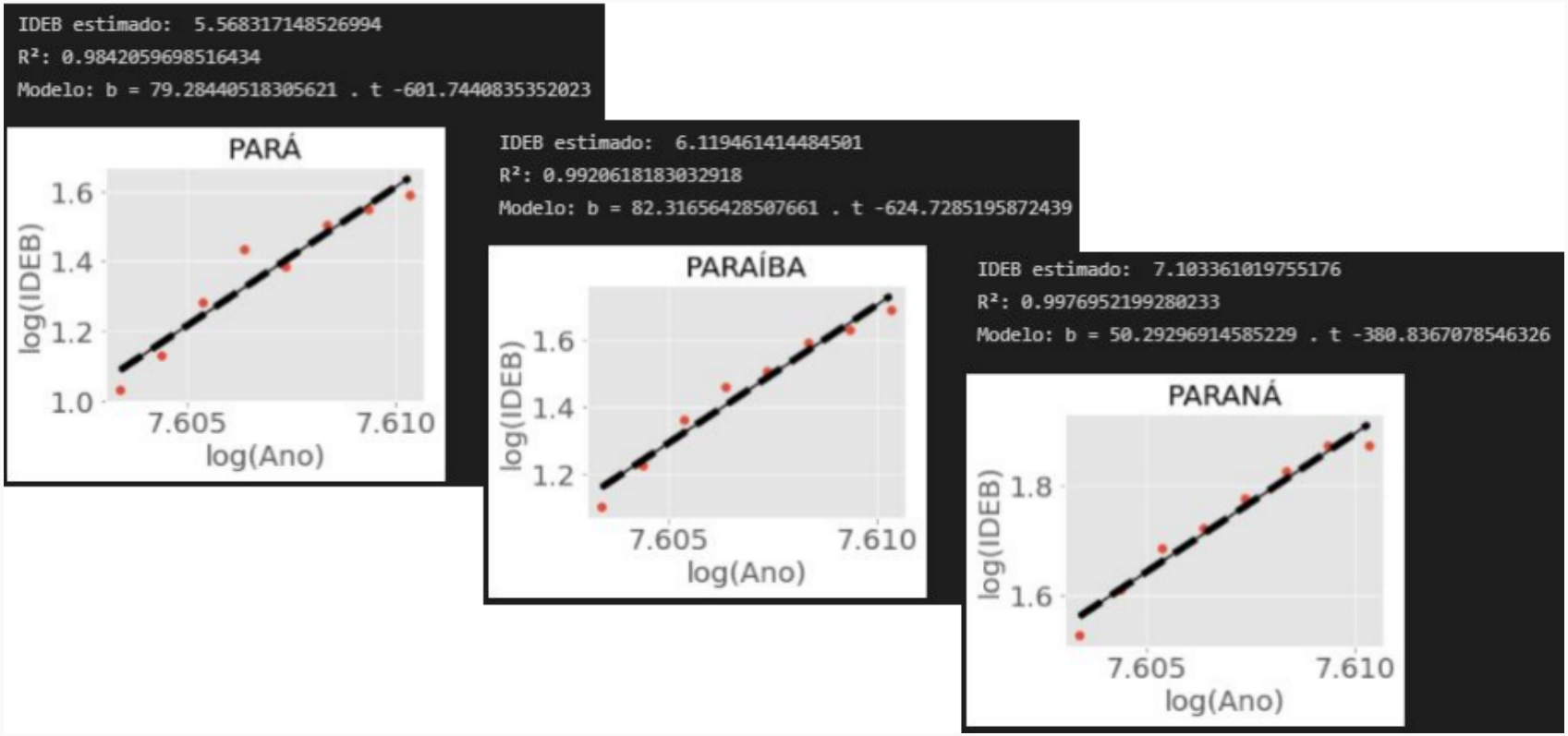

Figura 8: Gráfico de Dispersão para os estados do Pará, Paraíba e Paraná.

Gráficos: Pernambuco, Piauí e Rio de Janeiro

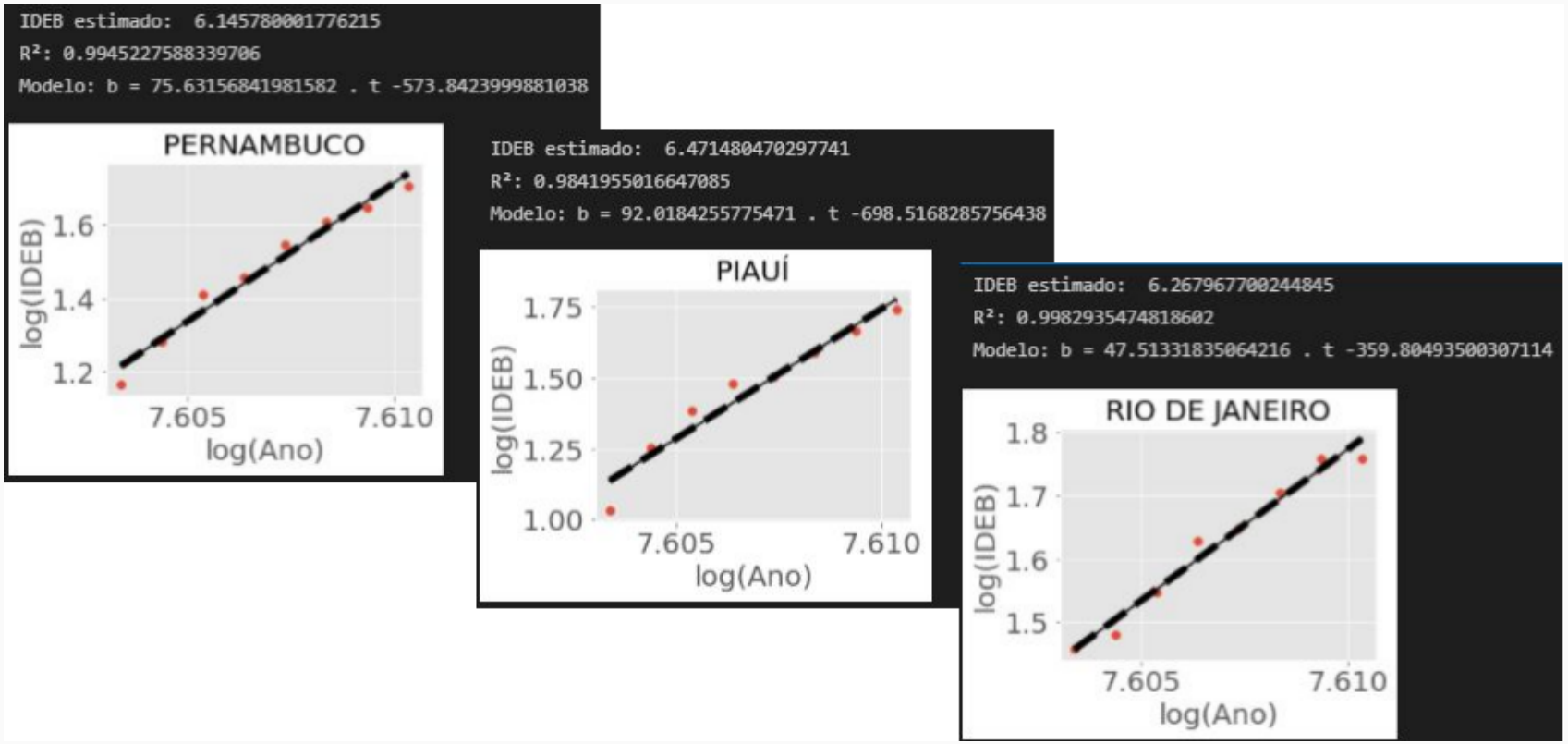

Figura 9: Gráfico de Dispersão para os estados de Pernambuco, Piauí e Rio de Janeiro.

Gráficos: Rio Grande do Norte, Rio Grande do Sul e Rondônia

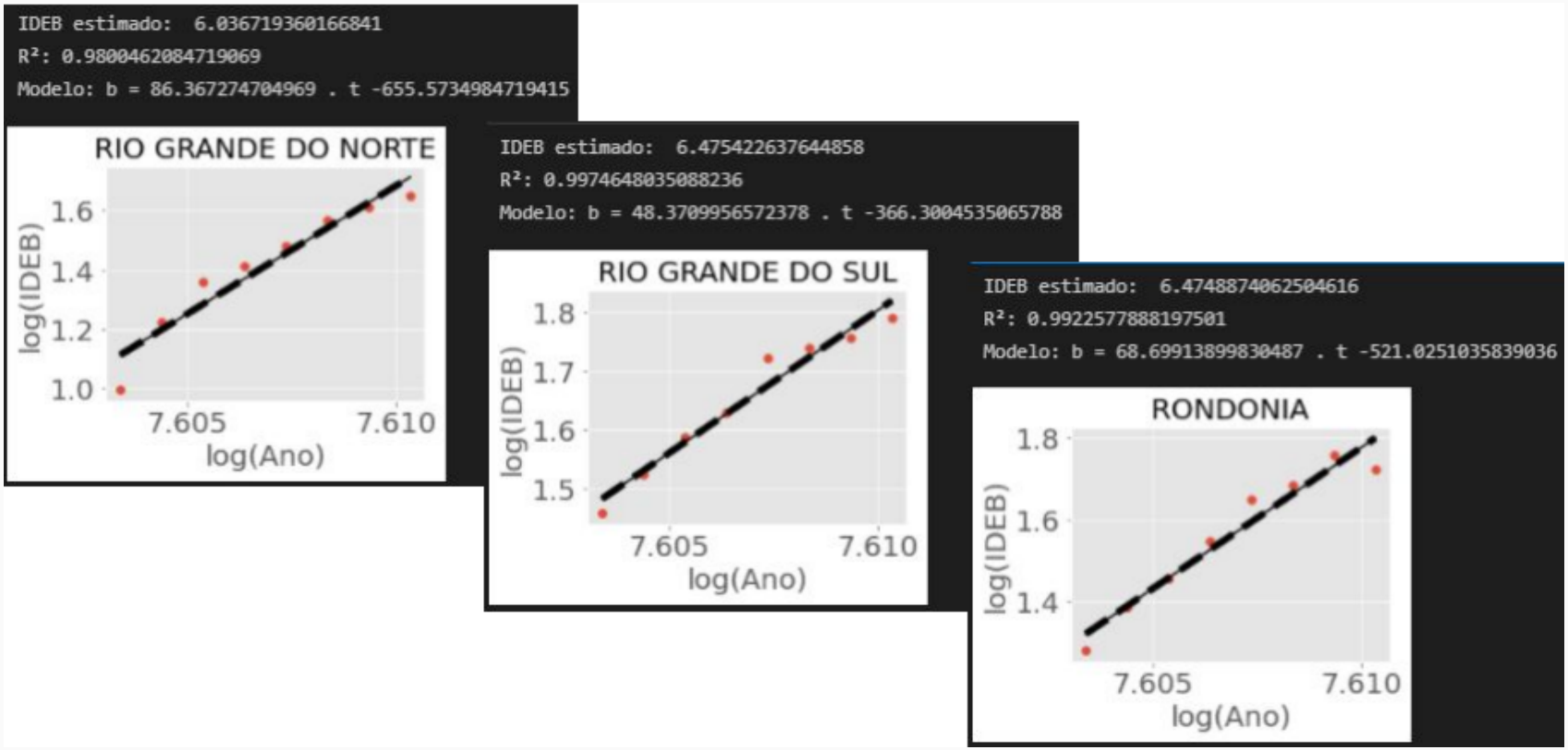

Figura 10: Gráfico de Dispersão para os estados do Rio Grande do Norte, Rio Grande do Sul e Rondônia.

Gráficos: Roraima, Santa Catarina e São Paulo

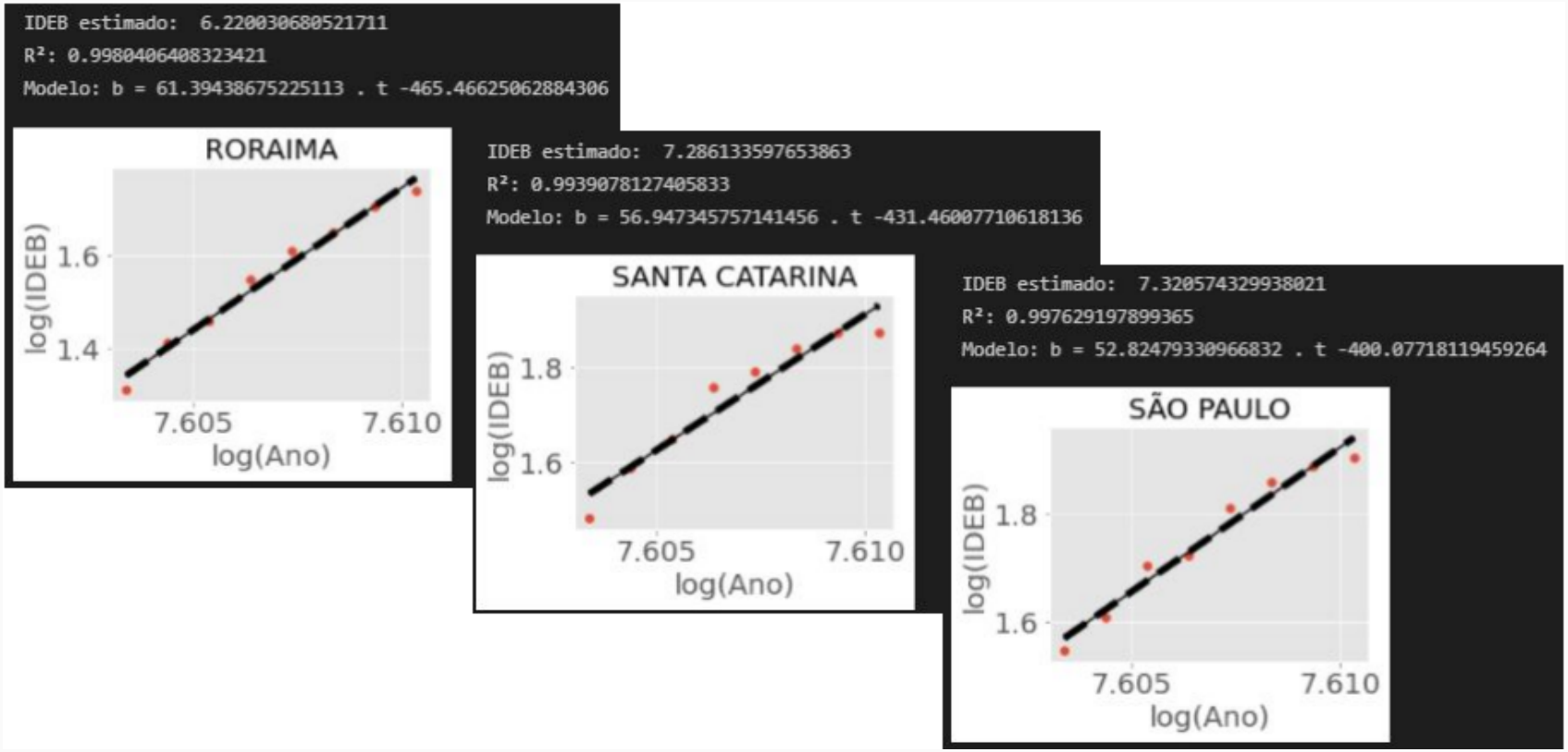

Figura 11: Gráfico de Dispersão para os estados de Roraima, Santa Catarina e São Paulo.

Gráficos: Sergipe e Tocantins

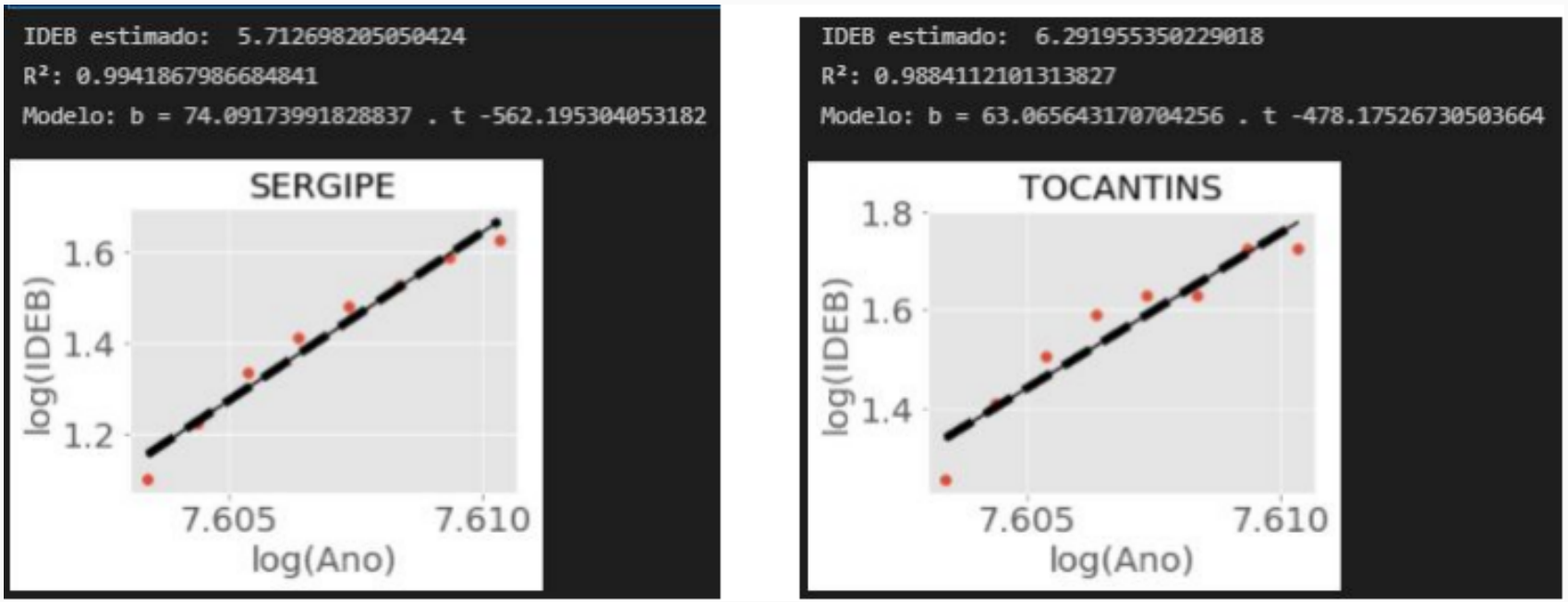

Figura 12: Gráfico de Dispersão para os estados de Sergipe e Tocantins.

Dataframe

Dataframe com os dados obtidos: Coeficiente Angular, R^2 e a previsão do IDEB para 2021

	Estado	Coef angular	R^2	IDEB 2021
1	Acre	80.960	0.9946	6.75
2	Alagoas	83.640	0.9953	5.98
3	Amapá	59.600	0.9954	5.25
4	Amazonas	83.760	0.9927	6.36
5	Bahia	88.750	0.9834	6.09
6	Ceará	98.694	0.9973	7.57
7	Distrito Federal	42.150	0.9974	6.87
8	Espírito Santo	51.650	0.9967	6.61
9	Goiás	62.550	0.9937	6.93
10	Maranhão	67.190	0.9837	5.51
11	Mato Grosso	64.890	0.9844	6.73
12	Mato Grosso do Sul	62.070	0.9871	6.47
13	Minas Gerais	51.240	0.9910	7.22
14	Pará	79.280	0.9842	5.56
15	Paraíba	82.310	0.9920	6.11
16	Paraná	50.290	0.9976	7.10
17	Pernambuco	75.630	0.9945	6.14
18	Piauí	92.010	0.9841	6.47
19	Rio de Janeiro	47.510	0.9982	6.26
20	Rio grande do Norte	86.360	0.9800	6.03
21	Rio Grande do Sul	48.370	0.9974	6.47
22	Rondônia	68.690	0.9922	6.47
23	Roraima	61.390	0.9980	6.22
24	Santa Catarina	56.940	0.9939	7.28
25	São Paulo	52.820	0.9976	7.32
26	Sergipe	74.090	0.9941	5.71
27	Tocantins	63.060	0.9884	6.29

Figura 13: Dataframe com os dados obtidos.

Dataframe odenado por R^2

	Estado	Coef angular	R^2	IDEB 2021
19	Rio de Janeiro	47.510	0.9982	6.26
23	Roraima	61.390	0.9980	6.22
16	Paraná	50.290	0.9976	7.10
25	São Paulo	52.820	0.9976	7.32
7	Distrito Federal	42.150	0.9974	6.87
21	Rio Grande do Sul	48.370	0.9974	6.47
6	Ceará	98.694	0.9973	7.57
8	Espírito Santo	51.650	0.9967	6.61
3	Amapá	59.600	0.9954	5.25
2	Alagoas	83.640	0.9953	5.98
1	Acre	80.960	0.9946	6.75
17	Pernambuco	75.630	0.9945	6.14
26	Sergipe	74.090	0.9941	5.71
24	Santa Catarina	56.940	0.9939	7.28
9	Goiás	62.550	0.9937	6.93
4	Amazonas	83.760	0.9927	6.36
22	Rondônia	68.690	0.9922	6.47
15	Paraíba	82.310	0.9920	6.11
13	Minas Gerais	51.240	0.9910	7.22
27	Tocantins	63.060	0.9884	6.29
12	Mato Grosso do Sul	62.070	0.9871	6.47
11	Mato Grosso	64.890	0.9844	6.73
14	Pará	79.280	0.9842	5.56
18	Piauí	92.010	0.9841	6.47
10	Maranhão	67.190	0.9837	5.51
5	Bahia	88.750	0.9834	6.09
20	Rio grande do Norte	86.360	0.9800	6.03

Figura 14: Ordenação decrescente dos estados por R^2.

Taxa de crescimento

Dataframe odenado por coeficiente angular

	Estado	Coef angular	R^2	IDEB 2021
6	Ceará	98.694	0.9973	7.57
18	Piauí	92.010	0.9841	6.47
5	Bahia	88.750	0.9834	6.09
20	Rio grande do Norte	86.360	0.9800	6.03
4	Amazonas	83.760	0.9927	6.36
2	Alagoas	83.640	0.9953	5.98
15	Paraíba	82.310	0.9920	6.11
1	Acre	80.960	0.9946	6.75
14	Pará	79.280	0.9842	5.56
17	Pernambuco	75.630	0.9945	6.14
26	Sergipe	74.090	0.9941	5.71
22	Rondônia	68.690	0.9922	6.47
10	Maranhão	67.190	0.9837	5.51
11	Mato Grosso	64.890	0.9844	6.73
27	Tocantins	63.060	0.9884	6.29
9	Goiás	62.550	0.9937	6.93
12	Mato Grosso do Sul	62.070	0.9871	6.47
23	Roraima	61.390	0.9980	6.22
3	Amapá	59.600	0.9954	5.25
24	Santa Catarina	56.940	0.9939	7.28
25	São Paulo	52.820	0.9976	7.32
8	Espírito Santo	51.650	0.9967	6.61
13	Minas Gerais	51.240	0.9910	7.22
16	Paraná	50.290	0.9976	7.10
21	Rio Grande do Sul	48.370	0.9974	6.47
19	Rio de Janeiro	47.510	0.9982	6.26
7	Distrito Federal	42.150	0.9974	6.87

Figura 15: Ordenação decrescente dos estados por crescimento.

Estimativa do IDEB para 2021

Dataframe odenado pela estimativa do IDEB para 2021

	Estado	Coef angular	R^2	IDEB 2021
6	Ceará	98.694	0.9973	7.57
25	São Paulo	52.820	0.9976	7.32
24	Santa Catarina	56.940	0.9939	7.28
13	Minas Gerais	51.240	0.9910	7.22
16	Paraná	50.290	0.9976	7.10
9	Goiás	62.550	0.9937	6.93
7	Distrito Federal	42.150	0.9974	6.87
1	Acre	80.960	0.9946	6.75
11	Mato Grosso	64.890	0.9844	6.73
8	Espírito Santo	51.650	0.9967	6.61
22	Rondônia	68.690	0.9922	6.47
12	Mato Grosso do Sul	62.070	0.9871	6.47
21	Rio Grande do Sul	48.370	0.9974	6.47
18	Piauí	92.010	0.9841	6.47
4	Amazonas	83.760	0.9927	6.36
27	Tocantins	63.060	0.9884	6.29
19	Rio de Janeiro	47.510	0.9982	6.26
23	Roraima	61.390	0.9980	6.22
17	Pernambuco	75.630	0.9945	6.14
15	Paraíba	82.310	0.9920	6.11
5	Bahia	88.750	0.9834	6.09
20	Rio grande do Norte	86.360	0.9800	6.03
2	Alagoas	83.640	0.9953	5.98
26	Sergipe	74.090	0.9941	5.71
14	Pará	79.280	0.9842	5.56
10	Maranhão	67.190	0.9837	5.51
3	Amapá	59.600	0.9954	5.25

Figura 16: Ordenação decrescente dos estados pela estimativa para 2021.

Conclusão

- A regressão pelo Scikit-Learn coincide com a feita pelo NumPy, confirmando a funcionalidade do método dos mínimos quadrados.
- Pelo cálculo do R^2, que variou entre 0.980 (correspondente ao Rio Grande do Norte) e 0.998 (correspondente ao Rio de Janeiro), podemos concluir que o modelo se ajusta bem aos dados.
- Os valores estimados pela fórmula do Rio de Janeiro, Distrito Federal e Rio Grande do Sul se assemelharam à meta estipulada pelo governo para 2021.
- Por outro lado, o Rio Grande do Norte e o Ceará apresentaram grande diferenças entre esses valores.
- Isso mostra que, para estipular uma meta para o IDEB, não basta apenas observar sua regressão, mas também o contexto de cada localidade.