

# Atividade Avaliativa 3 - Análise de Regressão Prática

Alisson Rosa e Vítor Pereira

## Sumário

|     |  |   |
|-----|--|---|
| 1   | Introdução                                   | 1 |
| 2   | Análise Descritiva                           | 1 |
| 3   | Análise de Influência                        | 4 |
| 4   | Verificação dos pressupostos                 | 6 |
| 4.1 | Teste de hipótese dos pressupostos . . . . . | 6 |
| 5   | Ajuste final                                 | 6 |
| 6   | Comentário                                   | 7 |

## 1 Introdução

A proposta do respectivo trabalho é prever o produto interno bruto (PIB) de 26 estados do Brasil no ano de 2019, os dados foram extraídos de planilhas disponíveis no site do Instituto Brasileiro de Geografia e Estatística (IBGE) para isso utiliza-se como variáveis explicativas (covariáveis): **Pobreza**: Que fornece a taxa de extrema pobreza no ano 2010;

**Densidade Demográfica**: Informa a densidade demográfica de cada estado no ano de 2019;

**Área** : Refere-se a área em km de cada estado no ano de 2019;

**índice de Desenvolvimento Humano (IDH) Educacional**: Refere-se ao IDH educacional no de ano de 2017, a escolha das covariáveis foram para conter três eixos:

**População e Geografia do Estado**: Área e Densidade Demográfica;

**Condição de Vida**: Pobreza;

e **Educação** : IDH;

Como modelos de preditivos foi utilizado uma regressão linear, florestas aleatórias (rf) e os k-vizinhos mais próximos (knn)

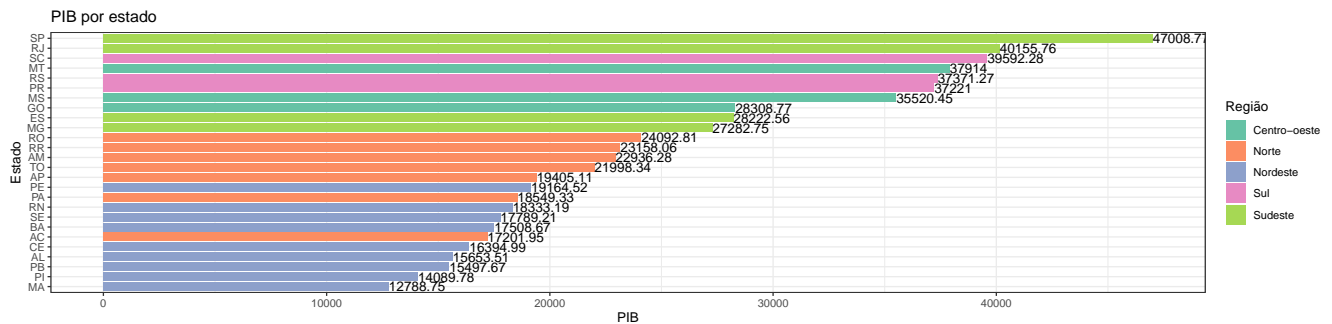
## 2 Análise Descritiva

Vejamos um breve resumo das variáveis de estudo :

Tabela 1: Resumo das variáveis:

|                       | n  | Média    | Desvio Padrão | Mediana  | Minímo   | Máximo   |
|-----------------------|----|----------|---------------|----------|----------|----------|
| PIB                   | 26 | 2.51e+04 | 9.78e+03      | 2.25e+04 | 1.28e+04 | 4.70e+04 |
| IDHe                  | 26 | 7.14e-01 | 4.80e-02      | 7.14e-01 | 6.36e-01 | 8.28e-01 |
| Área                  | 26 | 3.27e+05 | 3.77e+05      | 2.31e+05 | 2.19e+04 | 1.56e+06 |
| Densidade Demográfica | 26 | 5.87e+01 | 8.24e+01      | 3.12e+01 | 2.66e+00 | 3.95e+02 |
| Pobreza               | 26 | 1.20e-01 | 7.40e-02      | 1.29e-01 | 1.70e-02 | 2.63e-01 |

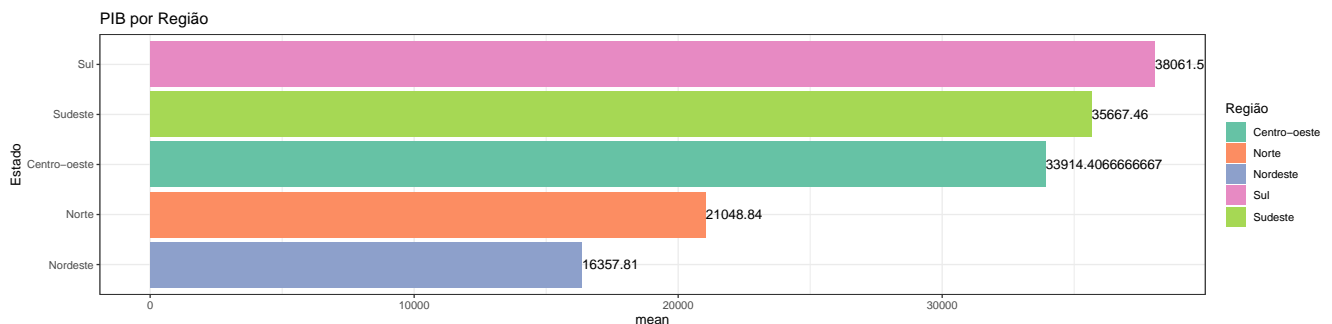
Note pelo seguinte gráfico que os estados da região Norte estão bem próximos da média do PIB por estado ( $2.512 \times 10^4$ ), SP e RJ estão bem acima e os estados da região nordeste estão bem abaixo.



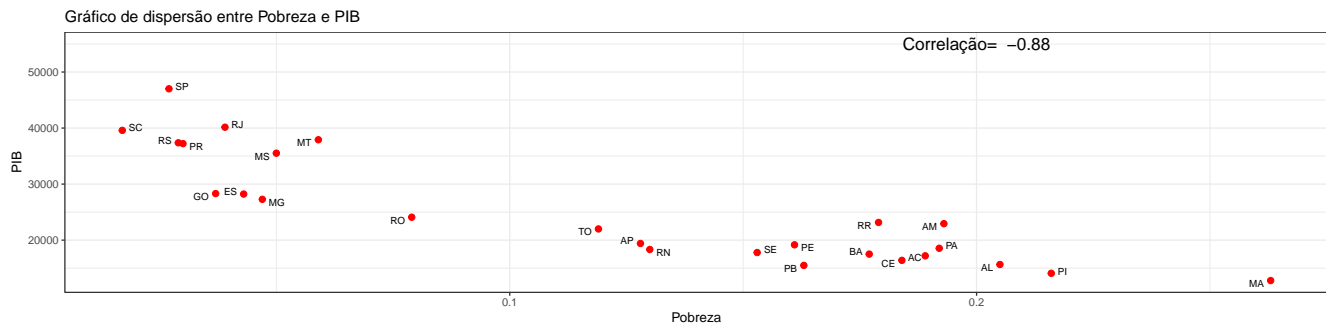
A título de curiosidade vejamos o PIB por região

Tabela 2: PIB por Região

| Região       | Média PIB | Desvio padrão | Quantidade |
|--------------|-----------|---------------|------------|
| Centro-oeste | 33914     | 5000          | 3          |
| Norte        | 21049     | 2643          | 7          |
| Nordeste     | 16358     | 2070          | 9          |
| Sul          | 38062     | 1328          | 3          |
| Sudeste      | 35667     | 9566          | 4          |

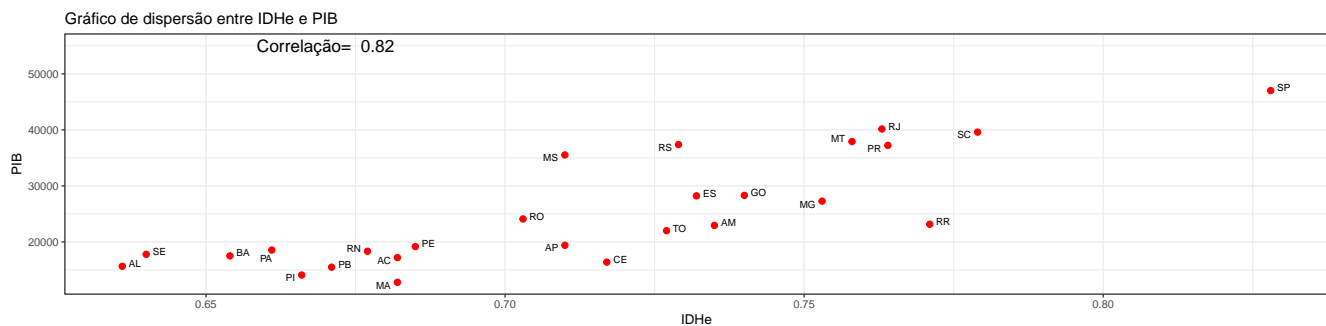


Para o eixo condição de vida, perceba a relação entre **Pobreza** e o **PIB** pelo seguinte gráfico de dispersão:



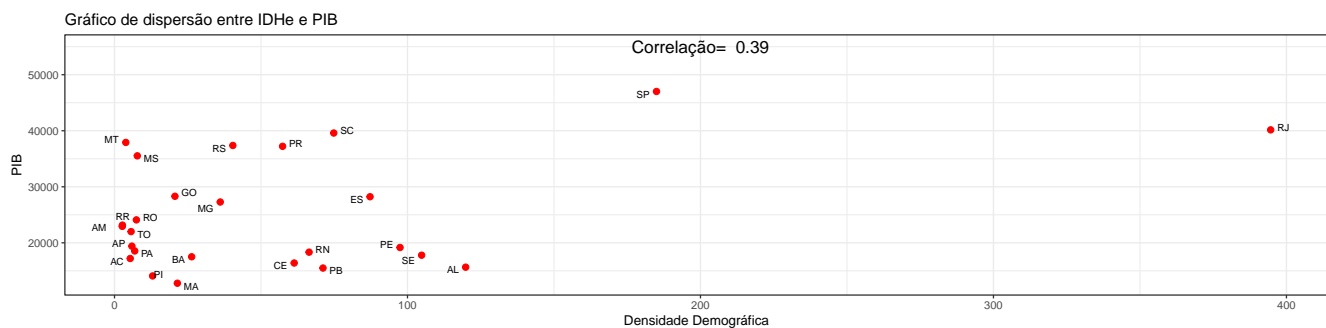
Pode-se ver pelo gráfico e pela correlação de -0.88 que quanto maior for a taxa de pobreza do estado, menor será seu PIB.

Para o eixo Educação, perceba a relação entre **IDHe** e o **PIB** pelo seguinte gráfico de dispersão:



Pode-se ver pelo gráfico e pela correlação de 0.822 que quanto maior for a IDHe, maior será seu PIB.

E para o eixo População e Geografia do Estado tem-se o gráfico de dispersão:



Duas observações se destacam das outras, pois possuem uma densidade demográfica bastante superior a média, sendo elas São Paulo e Rio de Janeiro, pelo gráfico de dispersão não fica muito claro o comportamento da relação entre PIB e Densidade Demográfica, a correlação de 0.391 indica que é uma correlação positiva entretanto fraca.

Tabela 3: Correlação entre as variáveis

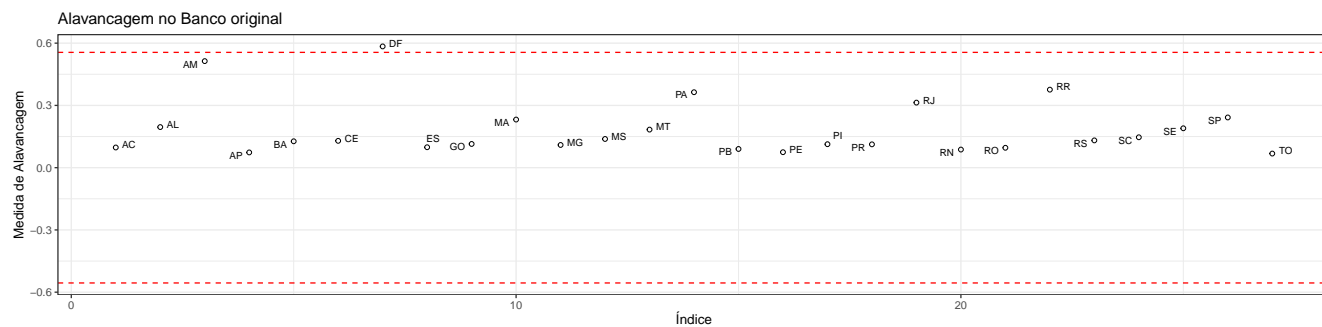
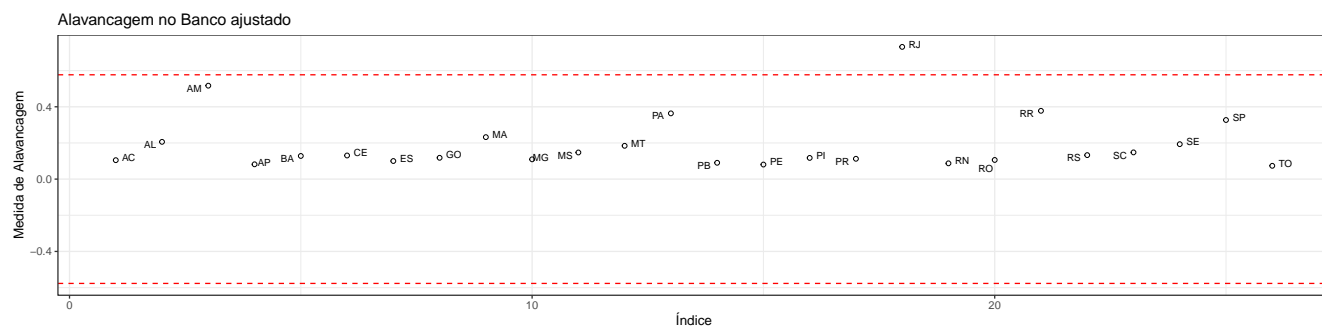
|                       | PIB    | IDHe   | Área   | Densidade Demográfica | Pobreza |
|-----------------------|--------|--------|--------|-----------------------|---------|
| PIB                   | 1.000  | 0.822  | 0.016  | 0.391                 | -0.880  |
| IDHe                  | 0.822  | 1.000  | 0.058  | 0.237                 | -0.692  |
| Área                  | 0.016  | 0.058  | 1.000  | -0.376                | 0.158   |
| Densidade Demográfica | 0.391  | 0.237  | -0.376 | 1.000                 | -0.292  |
| Pobreza               | -0.880 | -0.692 | 0.158  | -0.292                | 1.000   |

Podemos notar nos valores observados das variáveis que existe um pouco de correlação nas covariáveis, testaremos mais a frente a existência de multicolinearidade.

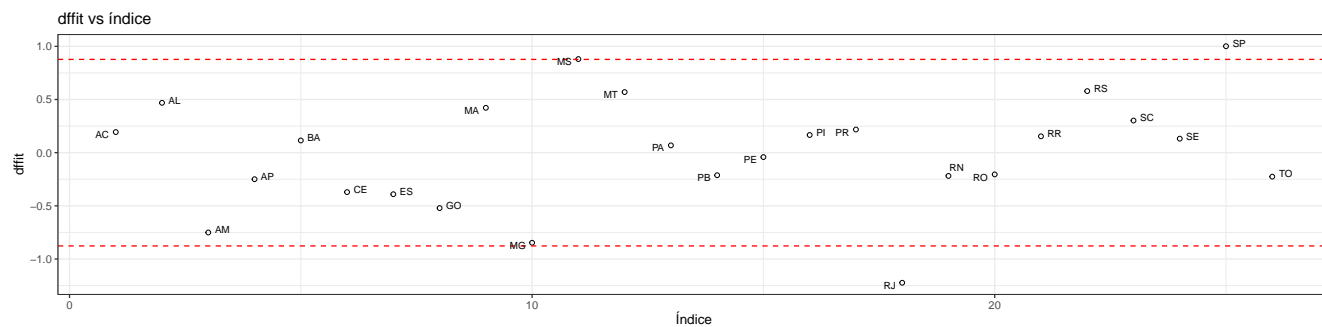
### 3 Análise de Influência

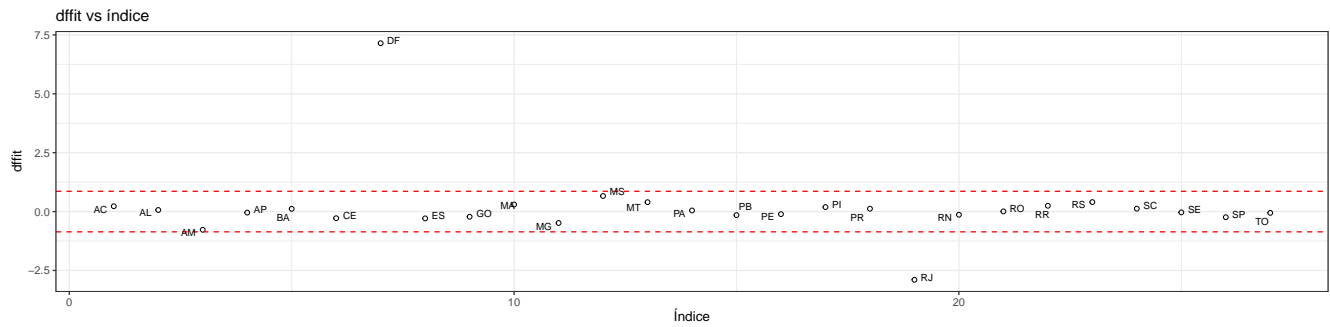
É possível notar que toda a análise descritiva foi realizada utilizando apenas os 26 estados brasileiros, sem o Distrito Federal, pois em uma Análise de Influência prévia foi demonstrado que esse estado é um ponto influente no nosso modelo, como será demonstrado nesta seção e no modelo de regressão linear.

No gráfico a seguir vemos as medidas de alavancagem, que informam se uma observação é discrepante em termos de covariável, nota-se que apenas uma observação está um pouco fora dos limites pré-estabelecidos.

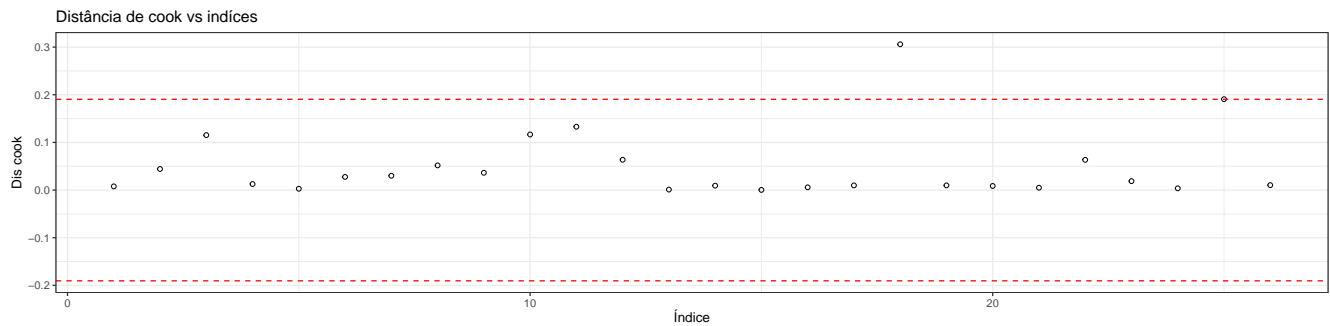


Temos dffits, que informam o grau de influência que a observação  $i$  tem sobre o valor seu próprio valor ajustado  $\hat{y}_i$ , percebe-se somente uma observação levemente fora dos limites

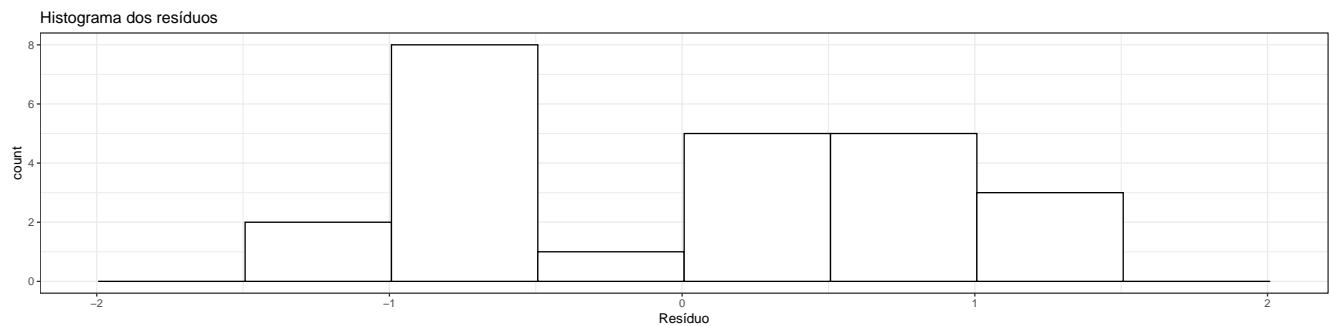
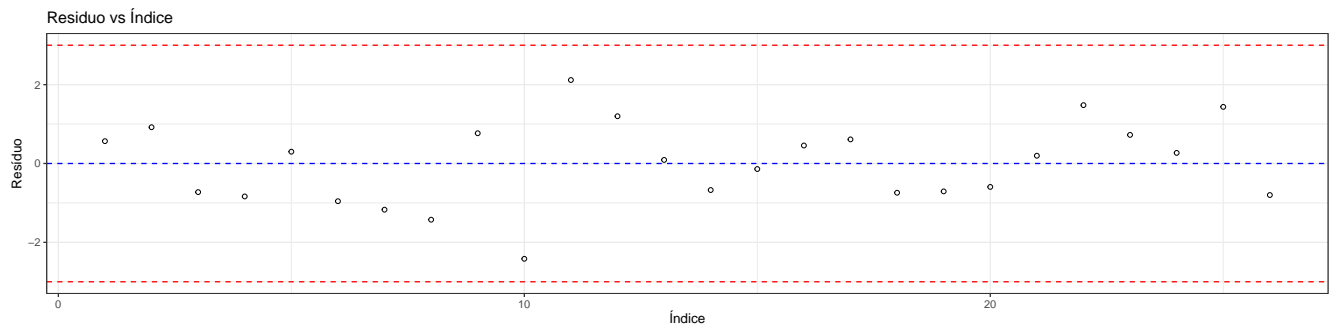




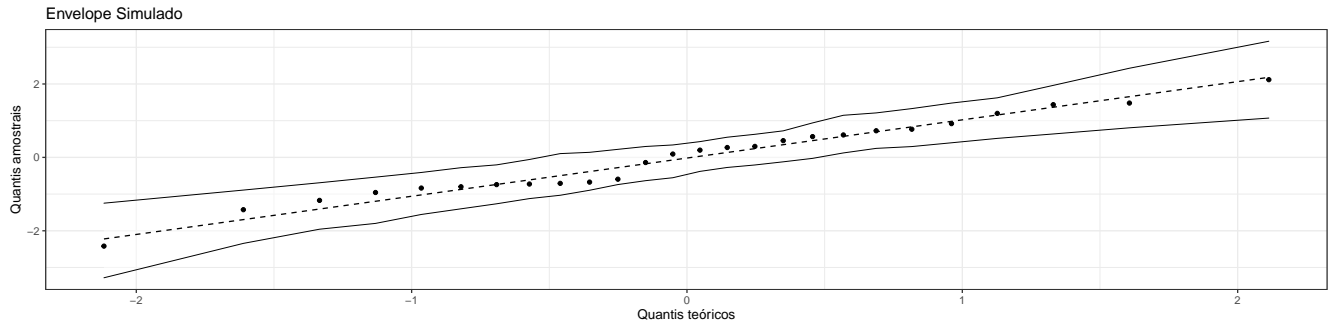
Tem-se também a distância de cook, que fornece a influência da observação  $i$  sobre todos os  $n$  valores ajustados



O gráfico de resíduos também é importante para verificarmos visualmente a média dos resíduos e se existe algum valor fora de 3 desvios padrões, pois esses possui baixíssima probabilidade de serem observados.



E por último tem-se o gráfico de envelope simulado, que informa se a distribuição proposta para os dados está em conforme com os valores observados, percebe-se todos os valores dentro das bandas simuladas



## 4 Verificação dos pressupostos

Precisamos primeiramente testar se os modelos estão corretamente especificados, faremos pelo teste Reset que tem como hipótese nula que o modelo está corretamente especificado, fazendo o teste para o modelo de regressão obtem-se um p-valor igual a 0.065 que nos informa que não existem evidências contra a hipótese suposta, portanto, É necessário ver se existem observações atípicas no conjunto dados, que podem estar influenciando a análise:

### 4.1 Teste de hipótese dos pressupostos

Primeiramente vamos testar se os erros ( $\epsilon$ ) possuem média zero, para isso usaremos um teste t que tem como hipótese:

$$H_0 : E(\epsilon_i) = 0$$

Obtem-se um p-valor » 0.1, portanto manteremos a hipótese de média nula dos erros.

Segundo precisamos testar a hipótese de variância constante dos erros, usaremos o Teste de Bressch-Pagan, que tem por hipótese:

$$H_0 : Var(\epsilon_i) = \sigma^2$$

Obtém-se um p-valor de 0.062 que também informa que não possuímos evidências amostrais contra a hipótese proposta.

Agora fazemos o teste de normalidade, utilizando o teste de Jarque-Bera, obtivemos um p-valor de 0.928 que também informa que não existem evidência contra normalidade dos erros

Como informado no início também é necessário testar se existe multicolinearidade, para tal usa-se fatores de inflação da variância (vif) para detectar, é ideal é que vif=1, obtemos 2.073, 1.26, 1.272, 2.088 para as variáveis  $x_1, x_2, x_3$  e  $x_4$  respectivamente.

E por último é necessário testar a existência de autocorrelação, usaremos o Teste de Durbin-Watson, que tem por hipótese, que existe não existe correlação, após aplicação do teste obtém-se um p-valor de 0.53 i.e, não existem evidências contra a hipótese de autocorrelação.

Logo, pelo testes anteriores não existem evidências contra os pressupostos teóricos, com isso podemos estabelecer inferência para os parâmetros do modelo

## 5 Ajuste final

Tem-se portanto como resumo do modelo final a seguinte tabela:

Tabela 4: Resumo do modelo final

| Coefficientes           | Estimativa | Erro Padrão | p-valor |
|-------------------------|------------|-------------|---------|
| (Intercept)             | -1.83e+04  | 1.55e+04    | 0.2518  |
| IDHe                    | 7.04e+04   | 2.04e+04    | 0.0024  |
| Área                    | 4.30e-03   | 2.00e-03    | 0.0463  |
| ‘Densidade Demográfica’ | 2.28e+01   | 9.29e+00    | 0.0228  |
| Pobreza                 | -8.03e+04  | 1.32e+04    | 0.0000  |

Que informa que o intercepto não são significativos a 1%, tem-se um p-valor « 0.001 do teste F e  $R^2$  dado por 0.899 que informa que aproximadamente 89.9% da variação do PIB dos estados é explicada pelas covariáveis propostas.

## 6 Comentário

O código completo pode ser acessado aqui.