Atividade Avaliativa 3 - Análise de Regressão Prática

Alisson Rosa e Vítor Perira

Sumário

1 Introdução

A proposta do respectivo trabalho é predizer o produto interno bruto per capita (PIB per capita) de 26 estados do Brasil no ano de 2019, os dados foram extraidos de planinhas 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 2019;

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

População e Geográfia do Estado: Área e Densidade Demográfica;

Condição de Vida: Pobreza;

Educação : IDHe.

2 Análise Descritiva

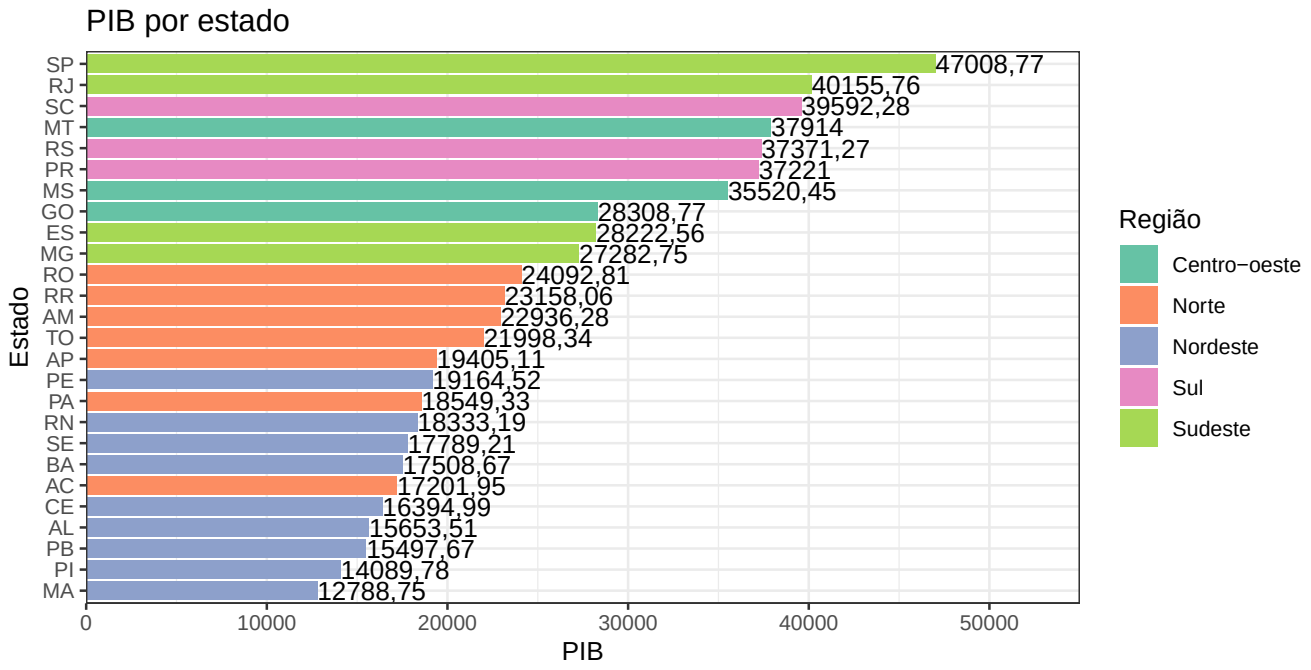
Nesta seção veremos um breve resumo das variáveis de estudo, com medidas descritivas, medidas de dispresão e gráficos de dispersão.

Começaremos por uma tabela resumo, com informações sobre as covariáveis:

Tabela 1: Resumo das variáveis:

	n	Média	Desvio Padrão	Mediana	Minímo	Máximo
PIB	26	25121,530	9778,952	22467,310	12788,750	47008,770
IDHe	26	0,714	0,048	0,714	0,636	0,828
Área	26	327117,686	377295,180	231019,532	21926,908	1559168,117
Densidade Demográfica	26	58,706	82,438	31,215	2,660	394,620
Pobreza	26	0,120	0,074	0,129	0,017	0,263

Note pelo seguinte gráfico que os estados da região Norte estão bem próximos da média do PIB per capita por estado (25121,53), SP e RJ estão bem acima e os estados da região nordeste estão bem abaixo.

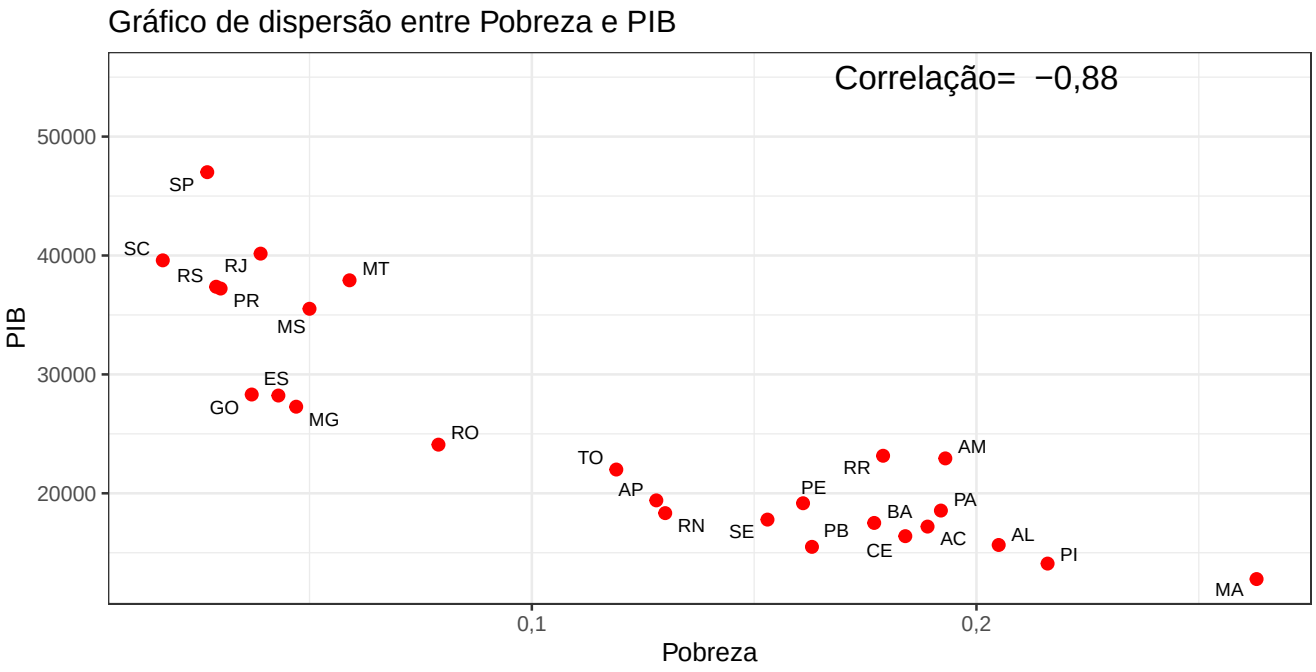


A titulo de curiosidade vejamos o PIB por região

Tabela 2: PIB por Região

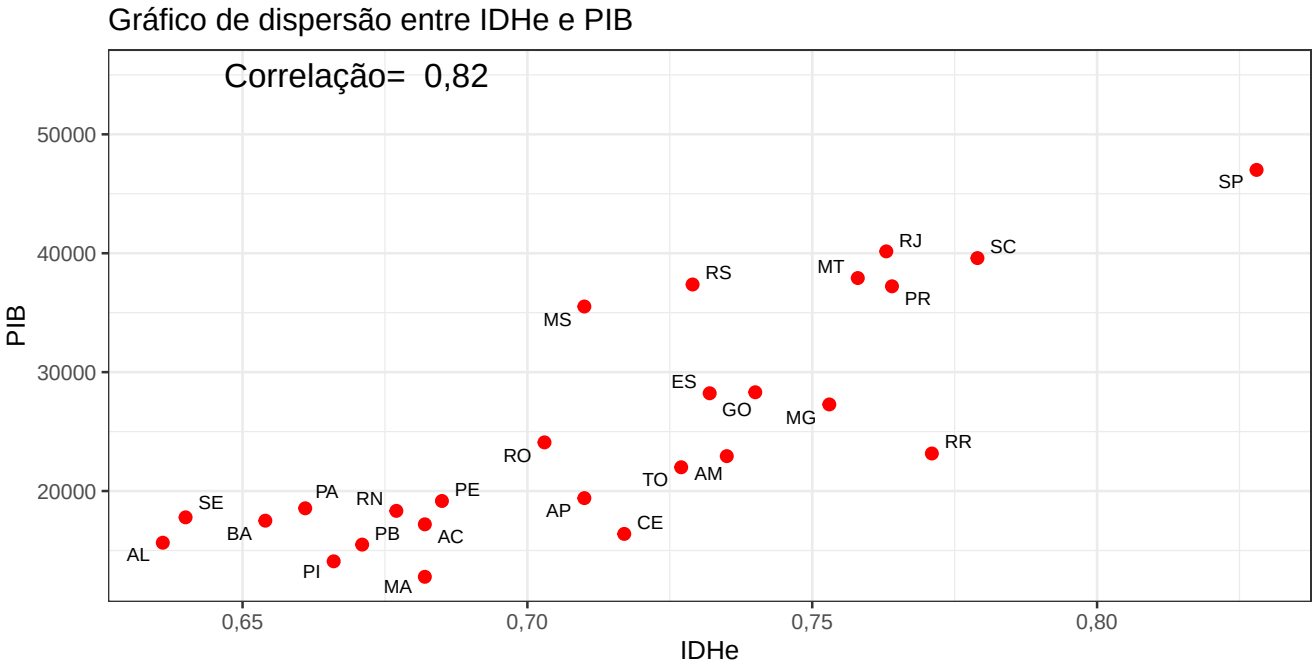
Região	Média PIB	Desvio padrão	Quantidade
Centro-oeste	33914	5000	3
Norte	21049	2643	7
Nordeste	16358	2070	9
Sul	38062	1328	3
Sudeste	35667	9566	4

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



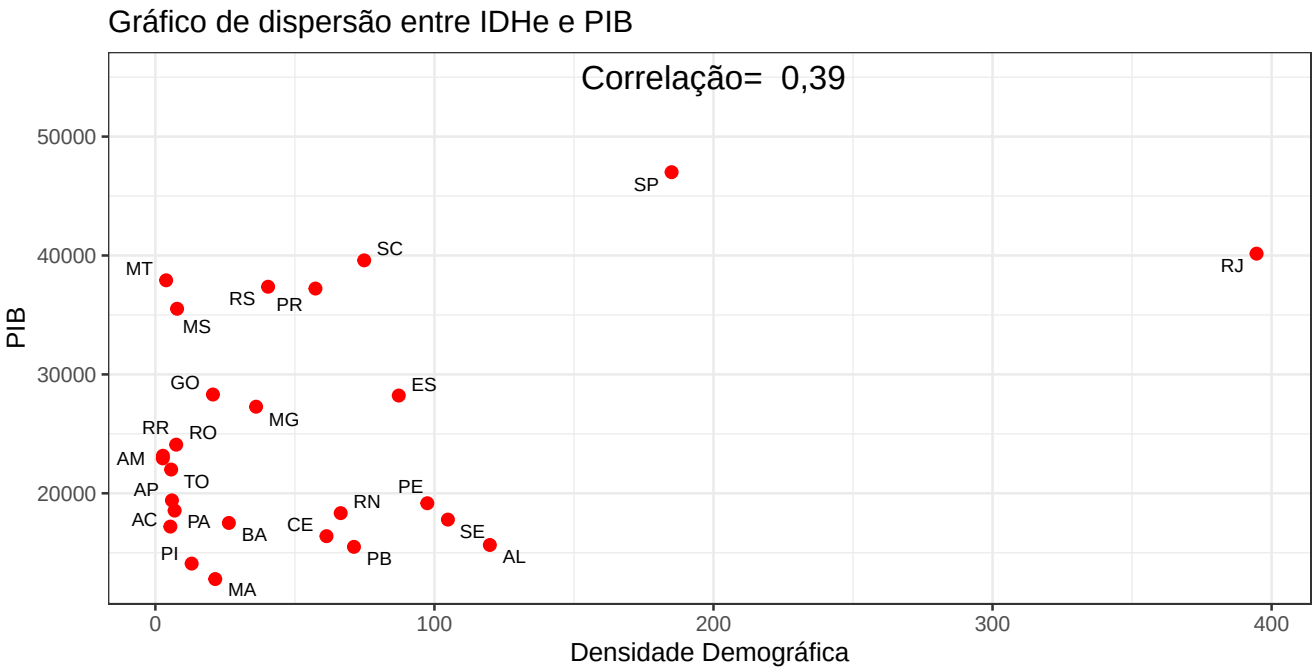
Pode-se ver pelo gráfico e pela correlação de -0,88 que quanto maior for a taxa de pobreza do estado, menor será seu PIB.

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 Geográfia 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

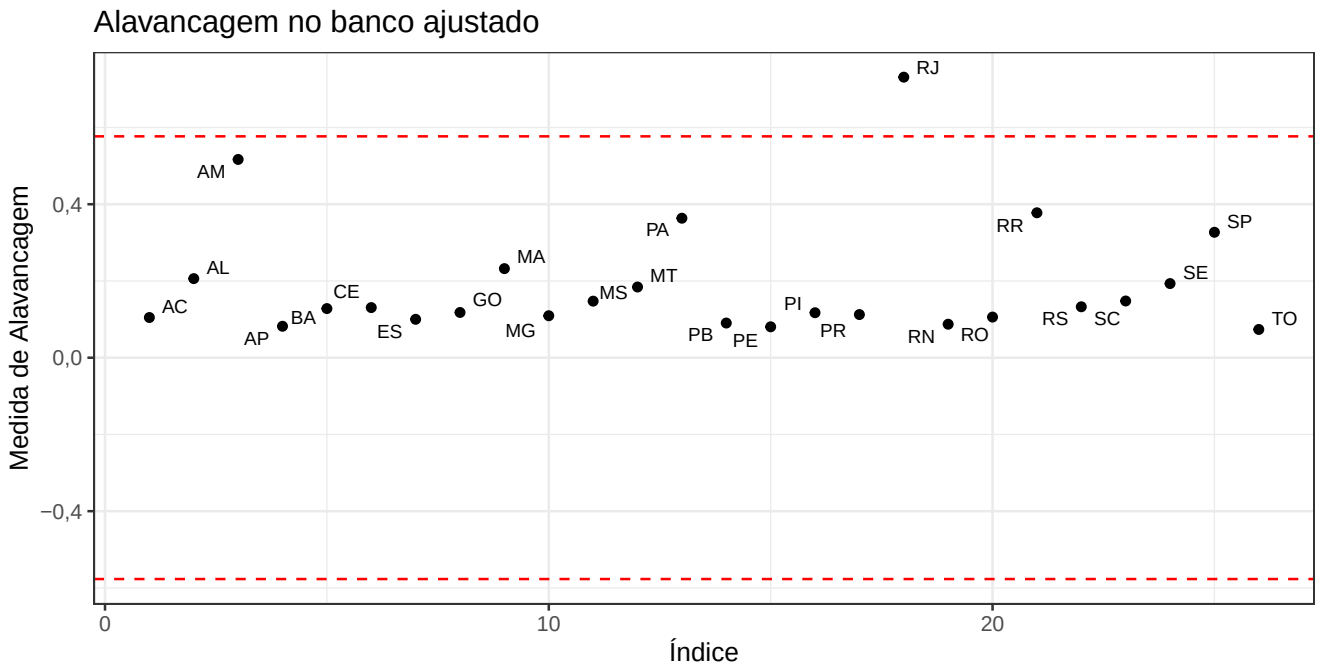
Podemos notar nos valores observados das variáveis que existe um pouco de correlação nas covariáveis, testaremos mais a frente a existência de multicolineriadade.

3 Análise de Influência

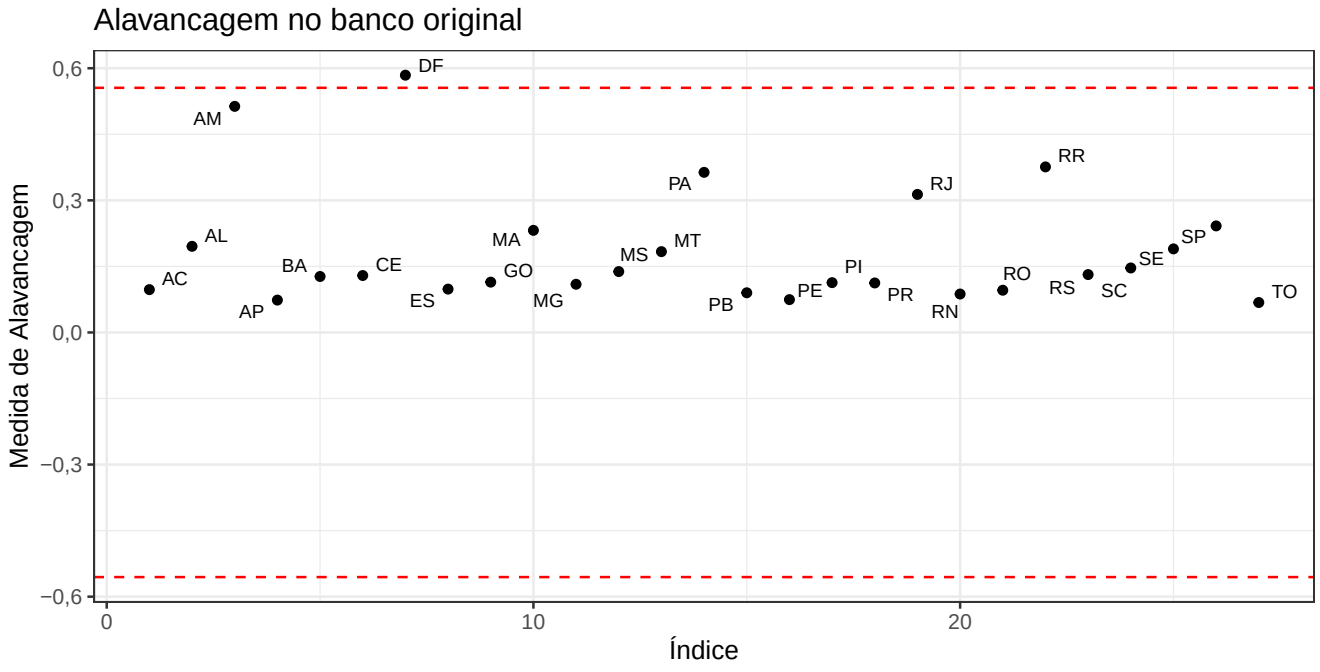
É possível notar que toda a análise descritiva foi realizada utilizando apenas os 26 estados brasileiros, sem o Distrito Federal, pois em uma Análise de Influência prévia foi analisado que esse estado é um ponto influente no nosso modelo, como será demonstrado nesta seção e no modelo de regressão linear.

Então é necessário ver se existem observações atípicas no conjunto dados, que podem estar influenciando a análise:

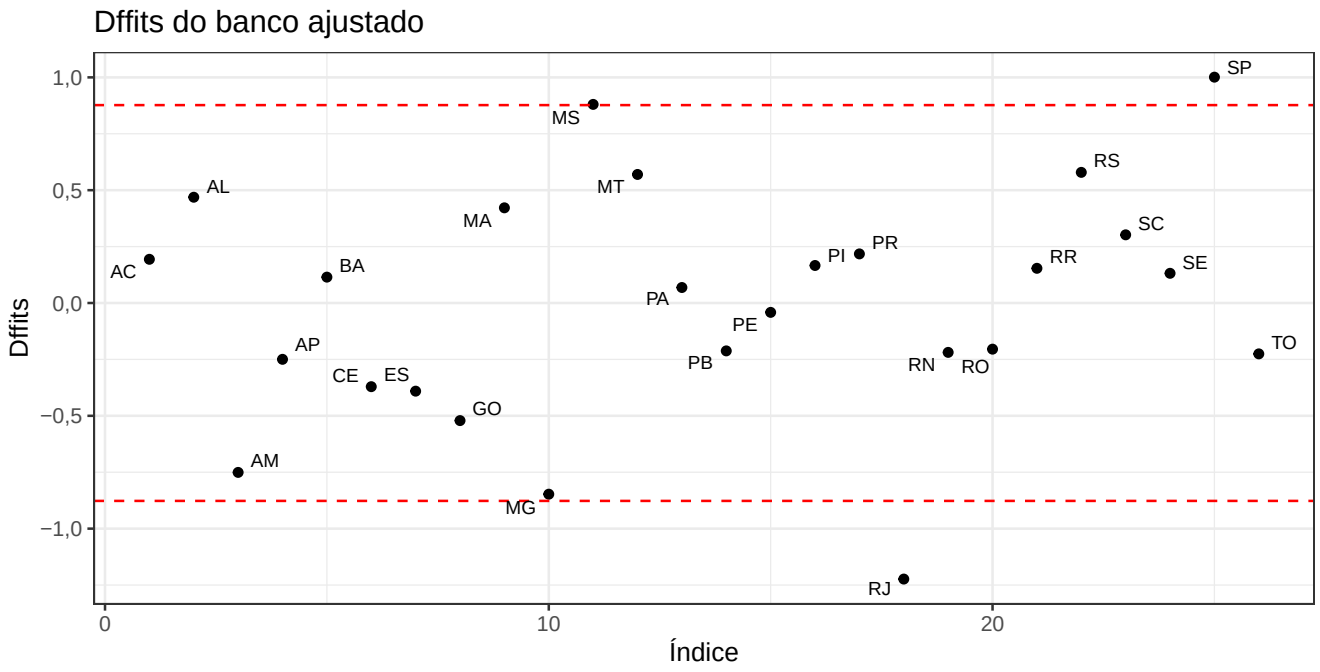
No gráfico a seguir vemos as medidas de alavancagem, que informam se uma observação é discrepante em termos de covariável, nota-se que apenas uma observação está um pouco fora dos limites pré-estabelecidos:



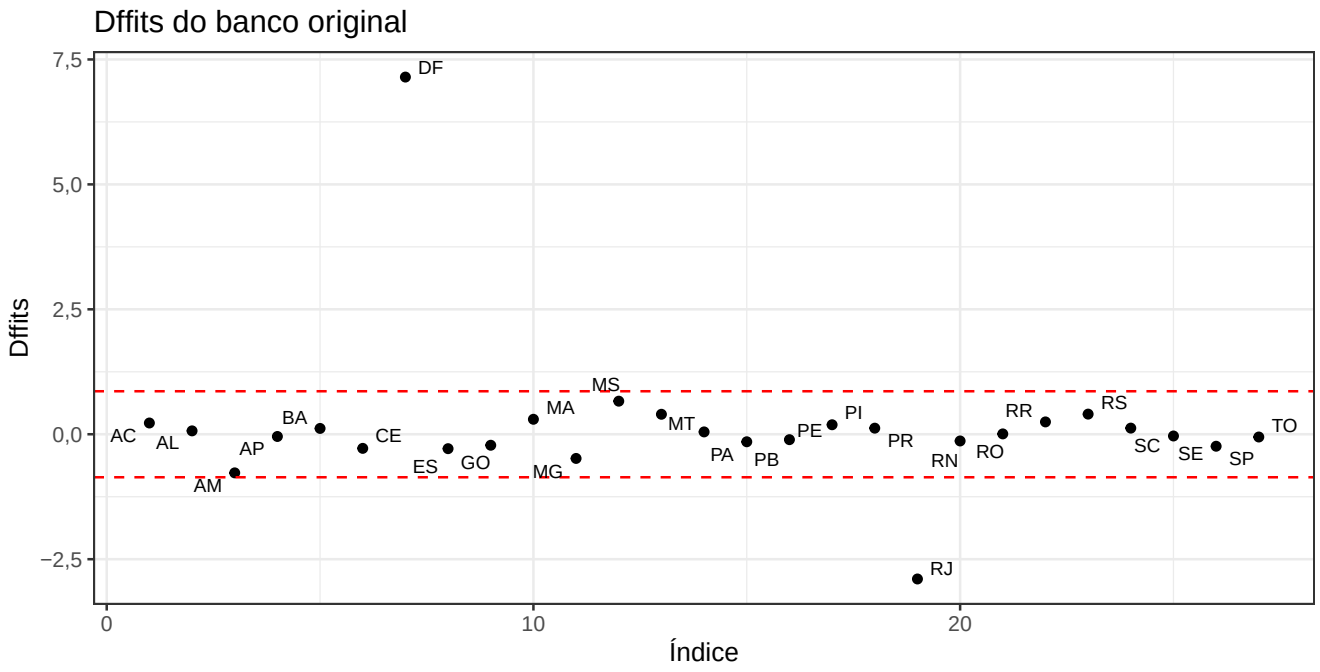
A Alanvancagem no banco original, fica com apenas o DF acima dos limites, mas bem próximo:



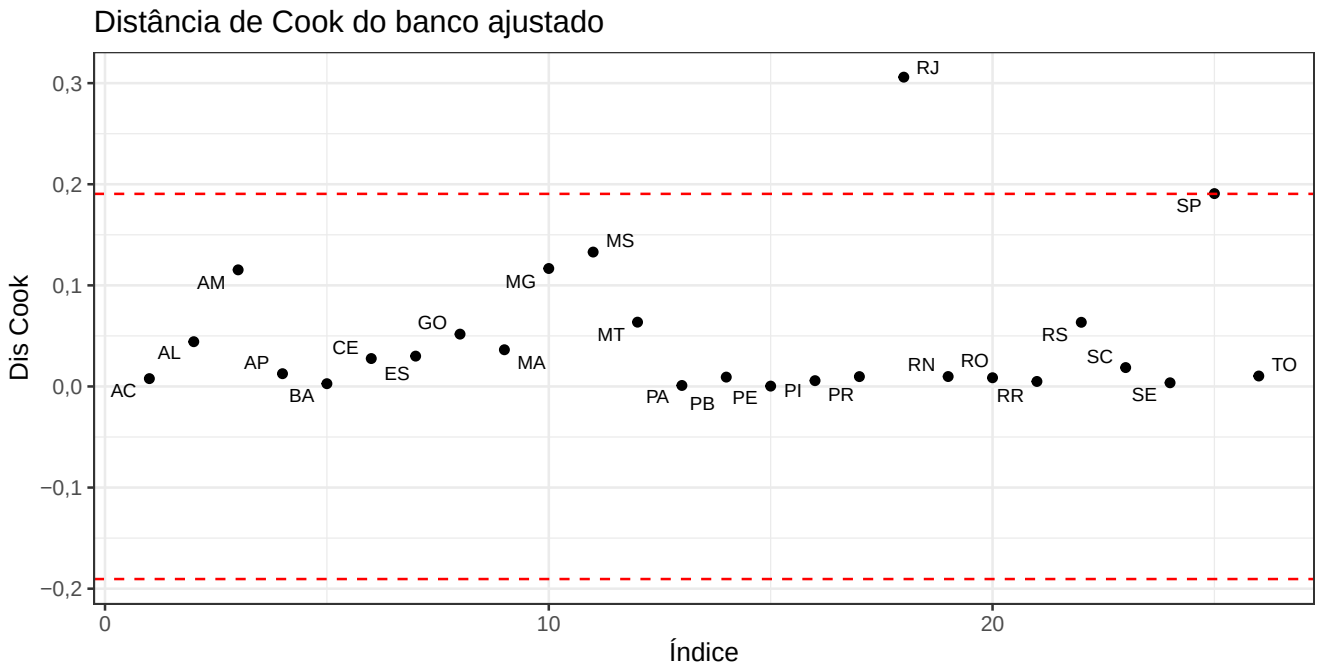
No diagnóstico dffits, que informam o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$, percebe-se somente uma observação levemente fora dos limites:



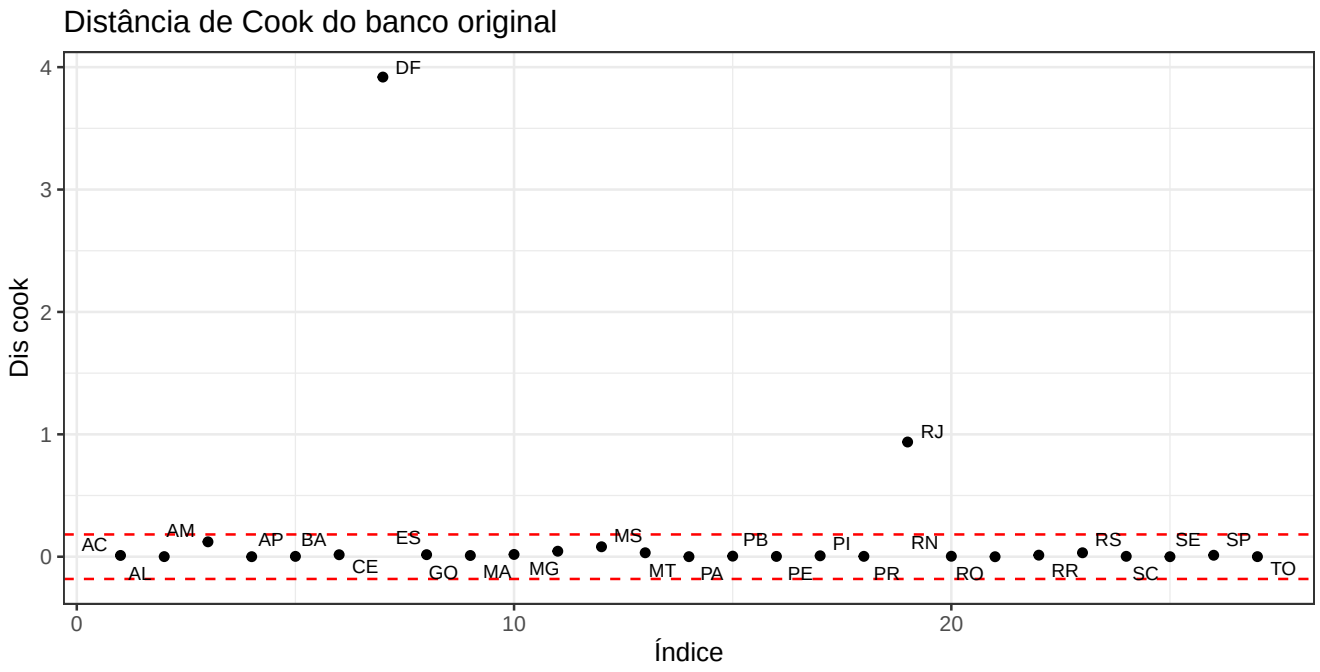
No Dffits do banco original, podemos perceber que o DF é um candidato a ponto de influencia, por estar muito acima do limite, achatando o gráfico e o RJ um pouco abaixo do limite:



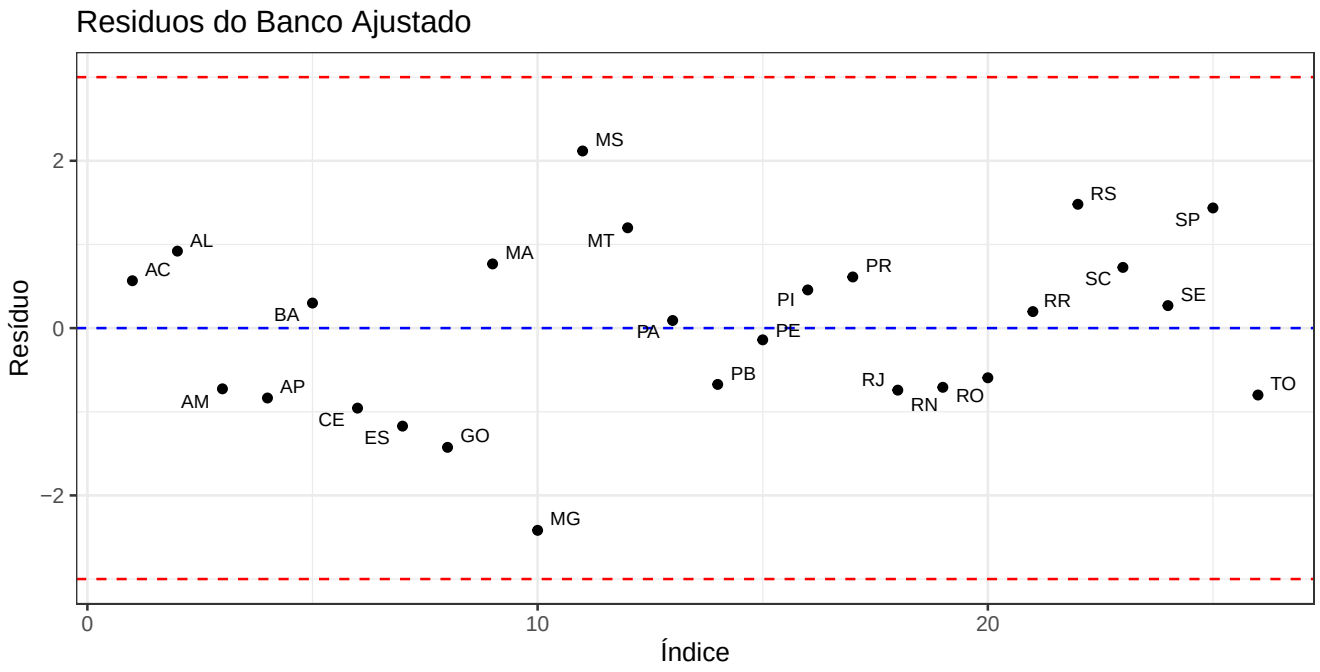
Tem-se também a distância de cook, que fornece a influência da observação i sobre todos os n valores ajustados, novamente com apenas o RJ acima dos limites estipulados:



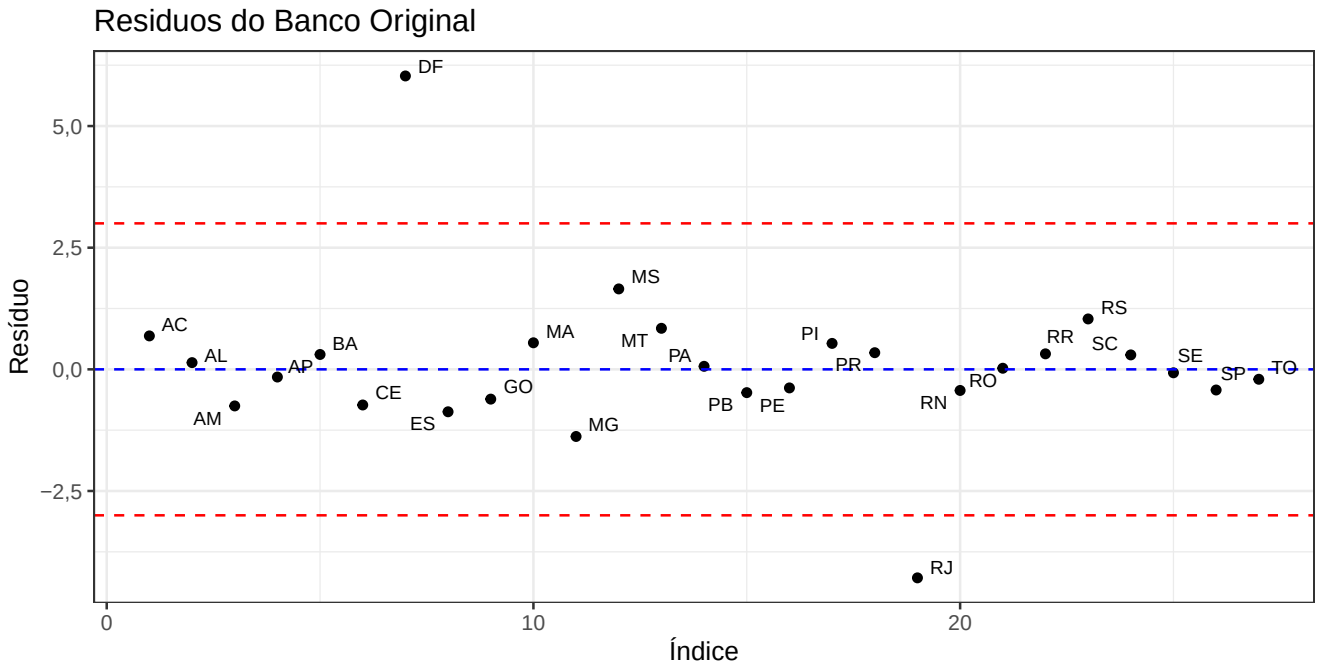
Na Distância de Cook do banco inicial podemos ver que o DF reincidente como um ponto extremamente fora dos limites achatando o gráfico, e o RJ também está acima:



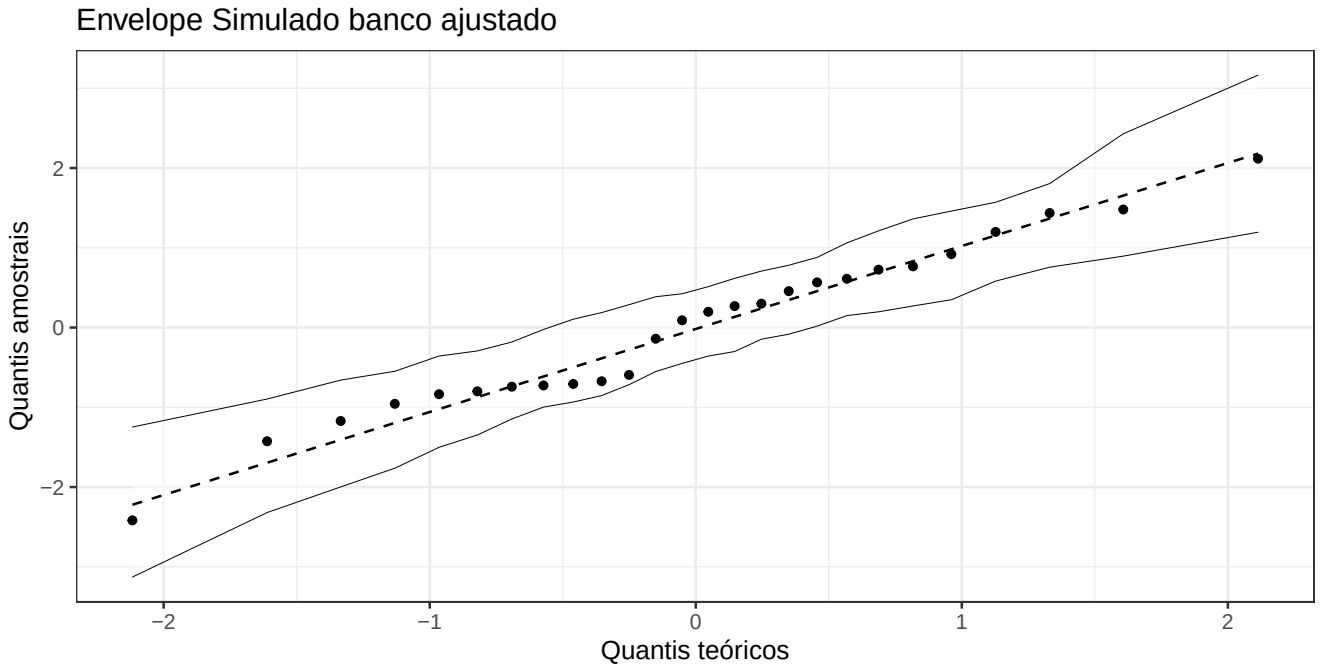
O gráfico de resíduos também é importante para verificarmos visualmente a média dos resíduos e se existe algum valor fora do limite de 3 desvios padrões, pois esses possui baixíssima probabilidade de serem observados, no gráfico abaixo verificamos que todos os estados estão dentro dos limites:



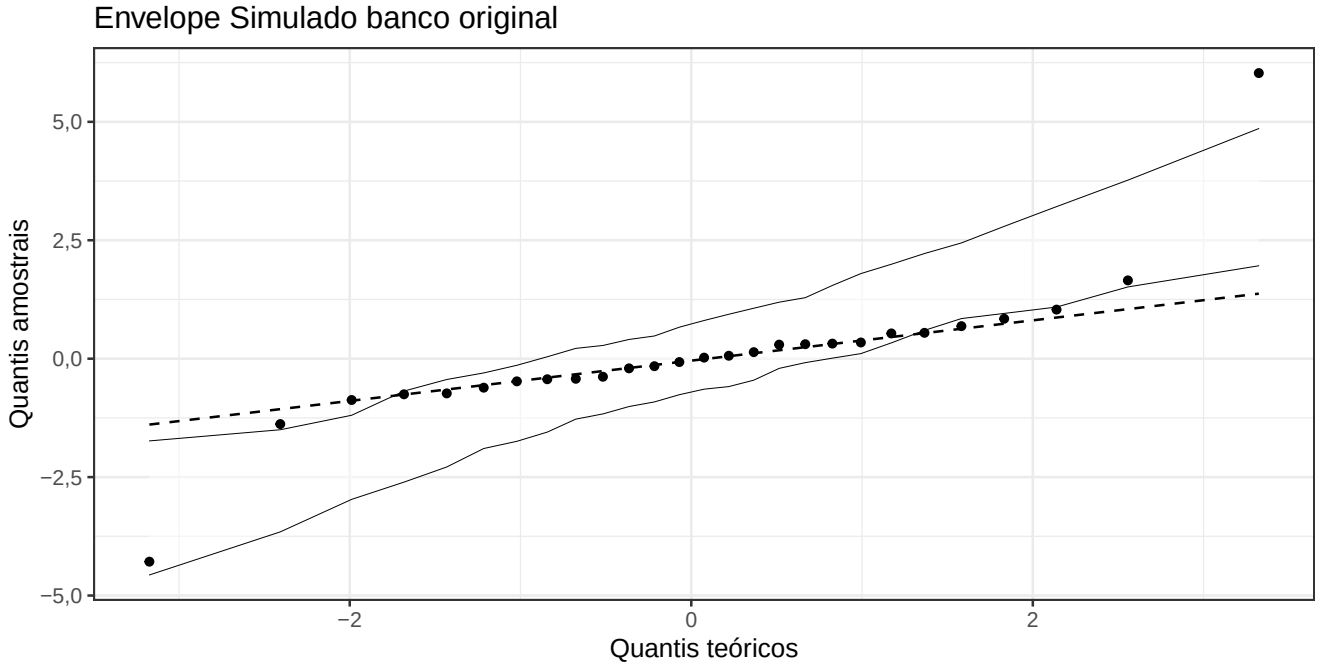
Abaixo percebemos que também no gráfico de resíduos studentizados, o DF é um candidato a ponto influente do modelo, assim como o RJ, porém no modelo ajustado, o RJ não fica fora do limite, tornando assim mais uma evidência da influência do DF no modelo:



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 dentros das bandas simuladas:



Ao extremo contrário do envelope simulado do banco original, em que podemos perceber pontos na borda e fora 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 ajustado obtem-se um p-valor igual a 0,065 que nos informa que não existem evidências contra a hipótese suposta.

No entanto considerando o banco de dados original obtem-se um p-valor de 0, então rejeitamos a hipótese nula e concluímos que não há evidências para considerar o modelo do banco de dados original, contendo o Distrito Federal, corretamente especificado. Assim usaremos o modelo ajustado.

Por que o Distrito Federal é um ponto influente?

Porque é a capital do páis, sendo assim a única das 27 unidades federativas que não é um estado, foi estritamente planejada por Juscelino Kubitschek para ser o polo político e diplomátido do país, assim também sendo uma maneira de desenvolver outras regiões do país, no caso a Região Centro-Oeste. Ao contrário de todos os outros estados, que em sua grande maioria não tiveram nenhum planejamento estatal.

Notamos que por ser uma unidade federativa com a menor área, porém com grande relevância política, com o maior IDH e a maior densidade demográfica, sendo essas nossas covariáveis, pois em sua maioria é composto por pessoas ligadas a instituições públicas.

Agora iremos verificar os seguintes pressupostos, para utilização correta do modelo de regressão linear:

Amedia dos erros é zero, Homoscedasticidade dos erros, Não-autocorrelação, Ausência de Multicolinearidade e Normalidade dos erros.

Primeiramente vamos testar se os erros (ϵ) 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.

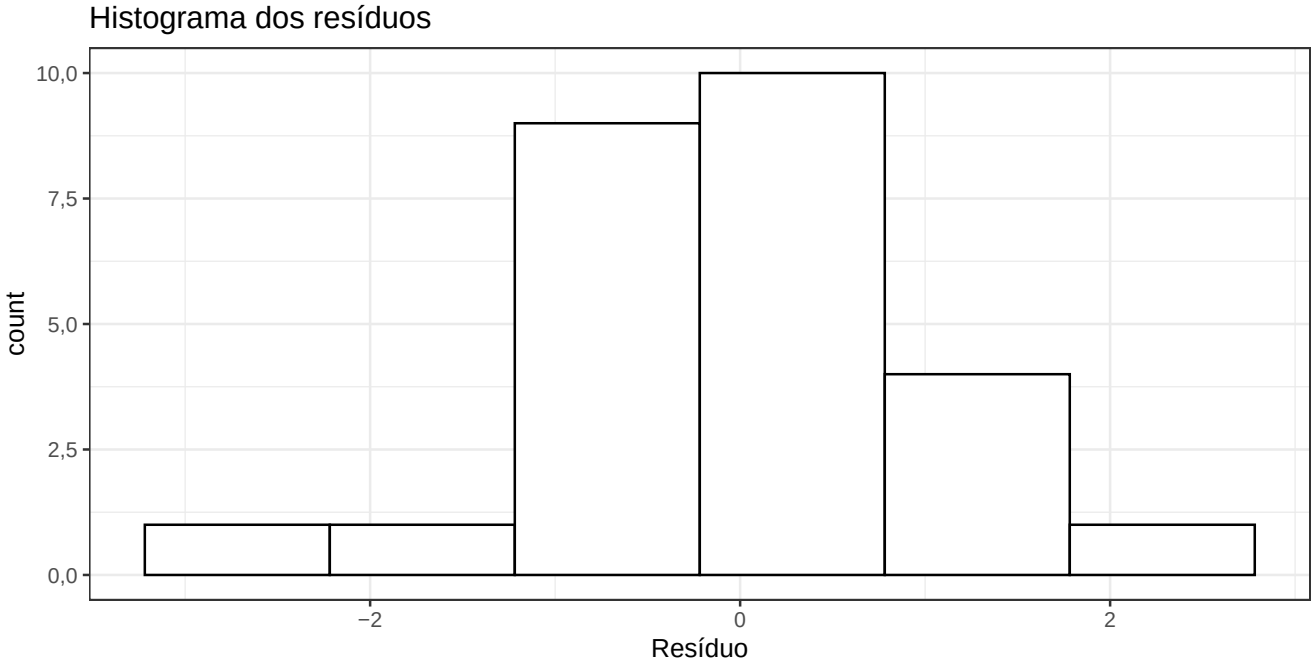
Segundo precisamos testar a hipótese de variância constante dos erros, usaremos o Teste de Bressch-Pagan, que tem por hipótese:

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

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, assim não rejeitando a hipótese nula, que é a existência normalidade dos erros.

Como podemos analisar no histograma dos resíduos abaixos, que se assemelha a distribuição normal:



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 e o máximo para não multicolinearidade é 5, 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 indicío 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 Conclusão

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

Tabela 4: Resumo do modelo final

Coeficientes	Estimativa	Erro Padrão	p-valor
(Intercept)	-18296,206	15527,248	0,252
IDHe	70391,625	20444,077	0,002
Área	0,004	0,002	0,046
'Densidade Demográfica'	22,830	9,292	0,023
Pobreza	-80336,960	13190,734	0,000

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 per capita dos estados é explicada pelas covariáveis propostas.

Assim um estado com IDHe máximo (1) adicionaria 70.391,625 reais ao PIB per capita, sendo o maior IDHe existente 0,828 e o maior PIB per capita 47008,77, da mesma maneira que um estado com a Incidência de Pobreza máxima (1), sendo o maior existente 0,263, reduziria 80336,96 reais na variável. A influência da área é de apenas 0,004, ou seja, a cada 1 km^2 adiciona 0,004 reais ao PIB per capita, assim como a influência da densidade demográfica é que a cada 1 hab/km^2 adiciona 22,830 reais.

6 Apêndice

Temos conteúdos extras ao trabalho principal que serão linkados nessa seção, que estão relacionados ao objetivo, mas não correspondem ao fundamental do projeto, porém são interessantes para amplificar a análise do PIB per capita.

6.1 Apêndice A: Modelagem

Neste apêndice serão analisados outros modelos preditivos junto com o regressão linear, como os modelos de florestas aleatórias (rf) e os k-vizinhos mais próximos (knn), que pode ser acessado clicando aqui.

6.2 Apêndice B: Análise Descritiva, influência e pressupostos Extra

Neste apêndice serão feitas análises extras que são interessante para a compreensão da variável desfecho, mas não necessariamente fundamentais, que pode ser acessado clicando aqui.

6.3 Apêndice C: Código Completo

O código completo pode ser acessado aqui.