Regressão Linear aplicada ao IPARDES - Índice de Gini

Luiz Francisco, Mateus Fernandes, Mateus Luzzi. *GRRs*

junho/2023

Resumo

Introdução

Neste trabalho, iremos aplicar métodos de Regressão Linear para analisar a relação entre o Índice de Gini e um conjunto de variáveis explicativas relacionadas ao desenvolvimento socioeconômico dos municípios.

O Índice de Gini, criado pelo matemático italiano Conrado Gini, é um instrumento para medir o grau de concentração de renda em determinado grupo. Ele aponta a diferença entre os rendimentos dos mais pobres e dos mais ricos. Numericamente, varia de zero a um (alguns apresentam de zero a cem). O valor zero representa a situação de igualdade, ou seja, todos têm a mesma renda. O valor um (ou cem) está no extremo oposto, isto é, uma só pessoa detém toda a riqueza. Na prática, o Índice de Gini costuma comparar os 20% mais pobres com os 20% mais ricos. No Relatório de Desenvolvimento Humano 2004, elaborado pelo Pnud, o Brasil aparece com Índice de 0,591, quase no final da lista de 127 países. Apenas sete nações apresentam maior concentração de renda.[1]

Em resumo, o Índice de Gini é uma medida que permite avaliar o grau de desigualdade de renda em determinado grupo populacional, sendo utilizado como um indicador importante para medir a distribuição de renda.

As variáveis explicativas selecionadas para este estudo são:

1. População Estimada (IBGE) - Residentes em 01/07: A população estimada de um município pode influenciar o Índice de Gini, uma vez que a concentração de renda está relacionada à distribuição da população. Municípios com maior população podem apresentar maior desigualdade de renda devido a diversos fatores, como a presença de grandes centros urbanos ou desequilíbrios regionais.

2. Índice Ipardes de Desempenho Municipal (IPDM): O IPDM é um indicador composto que mede o desempenho dos municípios em diferentes áreas, como educação, saúde, infraestrutura, entre outros. A inclusão dessa variável se baseia na premissa de que municípios com melhores indicadores de desempenho tendem a apresentar menor desigualdade de renda.

3. Produto Interno Bruto (PIB) a Preços Correntes (R$ 1.000,00): O PIB é uma medida amplamente utilizada para medir a atividade econômica de um determinado local. A inclusão dessa variável se baseia na suposição de que municípios com maior PIB podem apresentar menor desigualdade de renda, uma vez que um maior nível de atividade econômica pode gerar oportunidades de emprego e renda para a população local.

[1] Fonte: www.ipea.gov.br/desafios/index.php?option=com_content&id=2048:catid=28

4. Produto Interno Bruto (PIB) per Capita (R$ 1,00): Essa variável corresponde ao PIB dividido pela população do município, representando uma medida de renda média por habitante. A inclusão dessa variável se justifica pelo pressuposto de que municípios com maior renda per capita tendem a apresentar menor desigualdade de renda, uma vez que a riqueza gerada é mais distribuída entre a população.

5. Taxa de Analfabetismo de 15 anos ou mais (%): A taxa de analfabetismo é um indicador que reflete o acesso à educação e pode estar associado à desigualdade de renda. Municípios com altas taxas de analfabetismo podem apresentar maior desigualdade, uma vez que a falta de educação formal pode limitar as oportunidades de emprego e renda.

6. Taxa de Distorção Idade Série no Ensino Fundamental (%): Essa taxa indica a proporção de alunos que apresentam defasagem idade-série no ensino fundamental, ou seja, que estão em séries inferiores àquelas correspondentes à sua idade. A inclusão dessa variável se baseia na suposição de que a distorção idade-série pode estar relacionada à desigualdade de oportunidades educacionais, o que por sua vez pode influenciar a desigualdade de renda.

7. Taxa de Distorção Idade Série no Ensino Médio (%): Essa taxa é similar à anterior, porém avalia a distorção idade-série no ensino médio. A inclusão dessa variável se justifica pelo mesmo motivo mencionado anteriormente: a distorção idade-série pode indicar desigualdades educacionais que podem influenciar a desigualdade de renda.

8. PEA (10 anos e mais) - Total: Essa variável representa a População Economicamente Ativa, ou seja, o número de pessoas em idade ativa que estão empregadas ou em busca de emprego. A inclusão dessa variável se baseia na suposição de que municípios com maior participação da população na força de trabalho podem apresentar menor desigualdade de renda, uma vez que a existência de mais oportunidades de emprego pode contribuir para a redução das disparidades.

Por meio da análise de regressão linear, pretende-se investigar a relação entre o Índice de Gini e essas variáveis explicativas, com o objetivo de compreender os fatores socioeconômicos que podem estar associados à desigualdade de renda nos municípios.

Material e métodos

Extração

Os dados utilizados neste estudo foram extraídos do Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES)[2]. Foram selecionadas oito variáveis previamente mencionadas, que abrangem diferentes aspectos socioeconômicos, e que podem estar relacionadas ao Índice de Gini. Essas variáveis foram coletadas para os 399 municípios do Estado do Paraná.

O período da extração compreende os anos de 2017 a 2022 para todas as variáveis, exceto a Taxa de Analfabetismo e o Índice de Gini, cujo últimos dados disponíveis se referem ao ano de 2010. A escolha desse período é relevante para analisar o impacto das variáveis explicativas sobre o Índice de Gini em um intervalo de tempo recente, permitindo uma compreensão atualizada da desigualdade de renda nos municípios paranaenses.

Os dados foram obtidos por meio da consulta com os parâmetros já enunciados e o arquivo retorno pode ser visto a seguir:

```
##   Localidade           Variavel      X2010       X2019       X2020 X2021 X2022
## 1     Abatiá               Gini       0,44                                   
## 2     Abatiá               IPDM     0,5641      0,6518      0,6516             
## 3     Abatiá           Populacao                 7.457       7.408 7.360
```

[2]Fonte: http://www.ipardes.gov.br/imp/

```
## 4     Abatiá       PIB Corrente 80.853,972 188.630,643 239.760,090
## 5     Abatiá     PIB per Capita     10.414      25.296      32.365
## 6     Abatiá      Analfabetismo      16,76
## 7     Abatiá Distorcao Fundamen       20,4        14,2        14,4  13,0  14,0
## 8     Abatiá    Distorcao Medio       37,9        21,4        18,5  28,9  23,3
## 9     Abatiá                PEA      4.141
```

Pré-processamento e tratamento

Ao analisar as variáveis fornecidas, podemos observar que nem todos os anos selecionados possuem dados disponíveis. Para lidar com essa situação, uma abordagem comum é realizar uma transformação nos dados, criando uma nova coluna que receberá o valor mais recente disponível para cada variável. Dessa forma, garantimos que todas as observações tenham pelo menos um valor válido.

Para destacar as modificações realizadas nas variáveis do dataframe, foram aplicadas as seguintes transformações:

1. PEA: Foi realizada uma modificação na variável "PEA" para tratá-la como uma taxa em relação à variável "População". Essa transformação é justificada pelo fato de que a População Economicamente Ativa (PEA) é um subgrupo da população total. Para realizar essa modificação, cada valor da variável "PEA" foi dividido pelo valor correspondente da variável "População". Dessa forma, os valores da variável "PEA" passaram a representar a proporção da população economicamente ativa em relação à população total.

2. População: Após a transformação acima, a variável "População" foi transformada com a aplicação do logaritmo em cada valor. O uso do logaritmo permite uma interpretação diferente dos valores, enfatizando diferenças proporcionais em vez de diferenças absolutas. Essa transformação pode ser útil em análises estatísticas, especialmente quando os dados estão distribuídos de forma assimétrica.

3. PIB Corrente: A variável "PIB Corrente" foi transformada aplicando a função logarítmica em cada valor. Essa modificação foi realizada para tratar assimetrias e variações proporcionais nos valores do PIB Corrente. A aplicação do logaritmo ajuda a reduzir a amplitude dos dados, permitindo uma interpretação mais precisa e facilitando a análise comparativa dos valores do PIB Corrente entre os municípios do Paraná.

4. PIB per Capita: Similarmente à variável anterior, a variável "PIB per Capita" também passou por uma transformação logarítmica. Essa modificação foi aplicada para lidar com a ampla dispersão dos valores e para permitir uma melhor interpretação comparativa dos valores do PIB per Capita entre os municípios do Paraná.

5. Distorção Fundamental: A variável "Distorção Fundamental" foi modificada dividindo cada valor por 100. Essa transformação foi realizada para expressar a distorção em relação à educação fundamental como uma proporção decimal. Ao dividir os valores por 100, os valores da variável "Distorção Fundamental" agora representam a proporção da distorção em relação à educação fundamental em uma escala de 0 a 1.

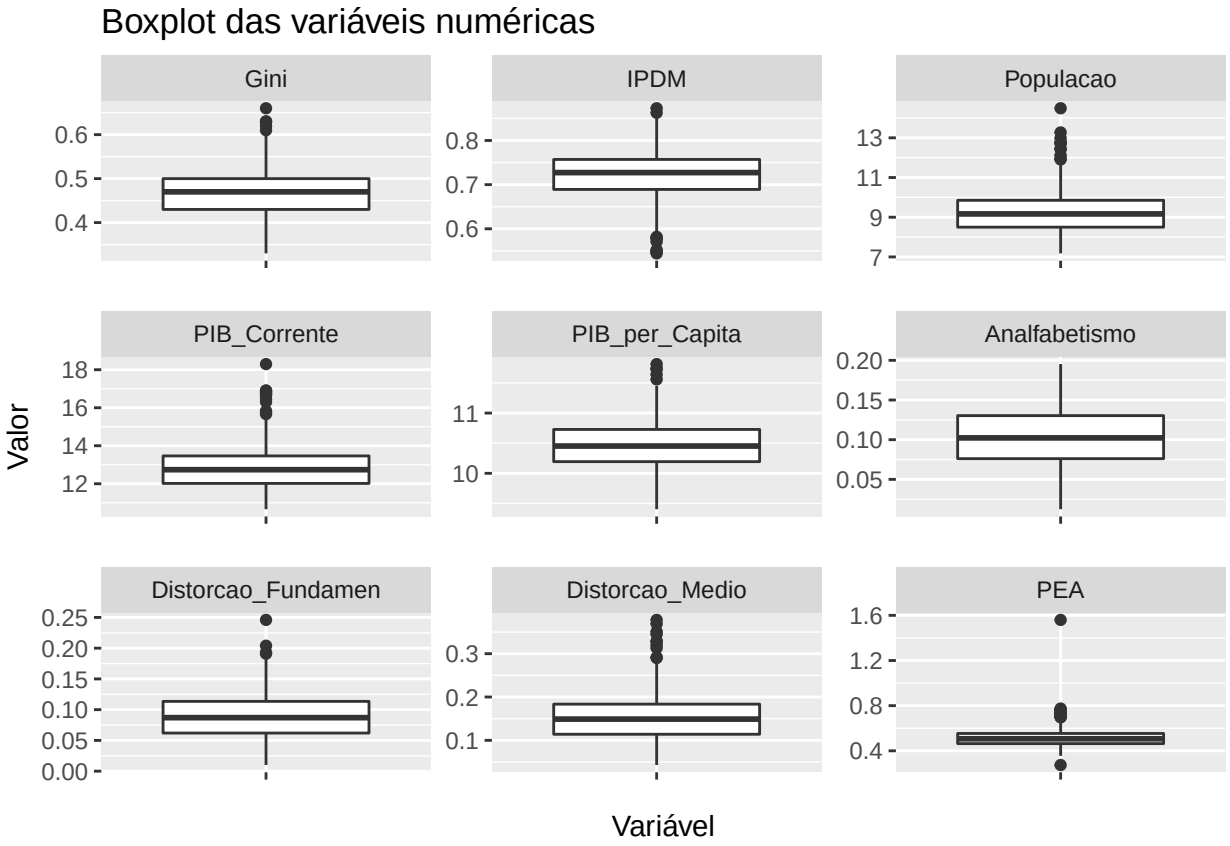
6. Distorção Médio: A variável "Distorção Médio" também foi modificada dividindo cada valor por 100. Essa transformação foi realizada para expressar a distorção em relação à educação média como uma proporção decimal. Ao dividir os valores por 100, os valores da variável "Distorção Médio" agora representam a proporção da distorção em relação à educação média em uma escala de 0 a 1.

7. Analfabetismo: A variável "Analfabetismo" foi modificada dividindo cada valor por 100. Essa transformação foi realizada para expressar a taxa de analfabetismo como uma proporção decimal. Ao dividir os valores por 100, os valores da variável "Analfabetismo" agora representam a proporção da taxa de analfabetismo em uma escala de 0 a 1.

Essas transformações foram realizadas com o objetivo de fornecer uma perspectiva mais adequada e comparativa das variáveis, possibilitando uma análise mais aprofundada dos dados do dataframe.

Análise Descritiva

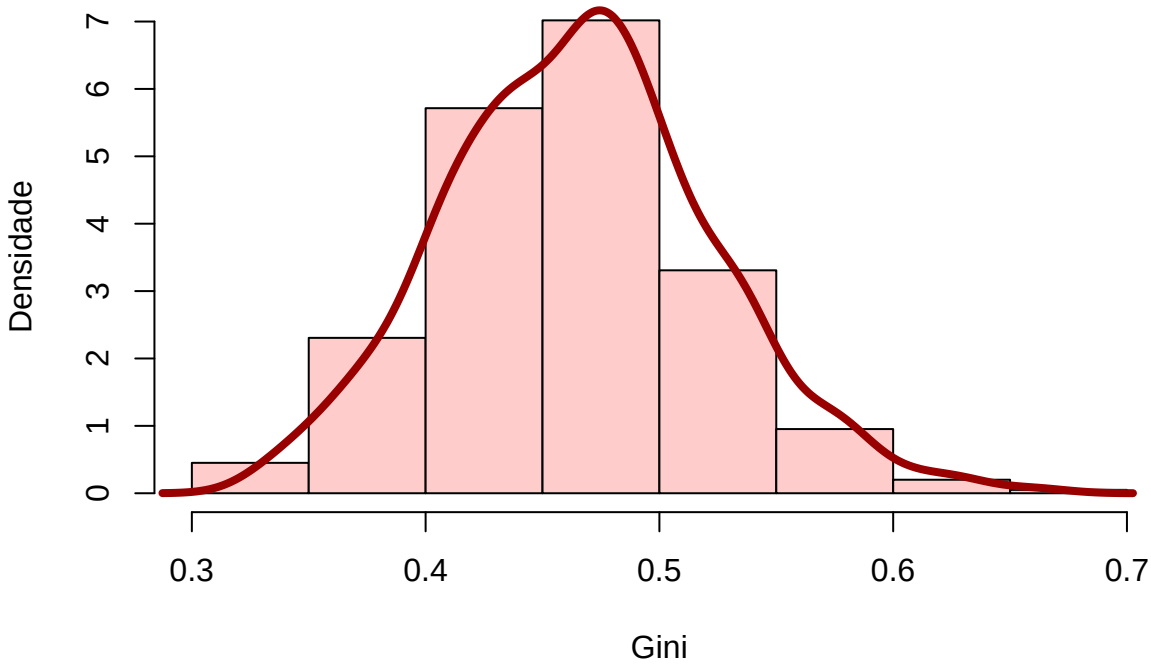
A visualização abaixo apresenta o boxplot das variáveis numéricas do dataframe. Cada variável é representada em um painel separado, permitindo uma análise comparativa dos seus valores.



É importante destacar um ponto interessante no boxplot da variável PEA (População Economicamente Ativa). Observa-se um possível outlier, um valor discrepante com o restante dos dados. O município de Altamira do Paraná tem a variável PEA com valor 1.55, sendo que esta deveria ser um valor entre 0 e 1, como proposto anteriormente pela transformação, já que a População Economicamente Ativa de um município é subconjunto de sua População. Após uma investigação mais detalhada nos dados estatísticos do município, foi confirmado que a população desse município vem diminuindo ano após ano. Como resultado, a População estimada pelo IBGE para o ano corrente é maior do que a variável PEA, que foi medida pela última vez em 2010. Para lidar com esse outlier, foi feita a imputação por mediana.

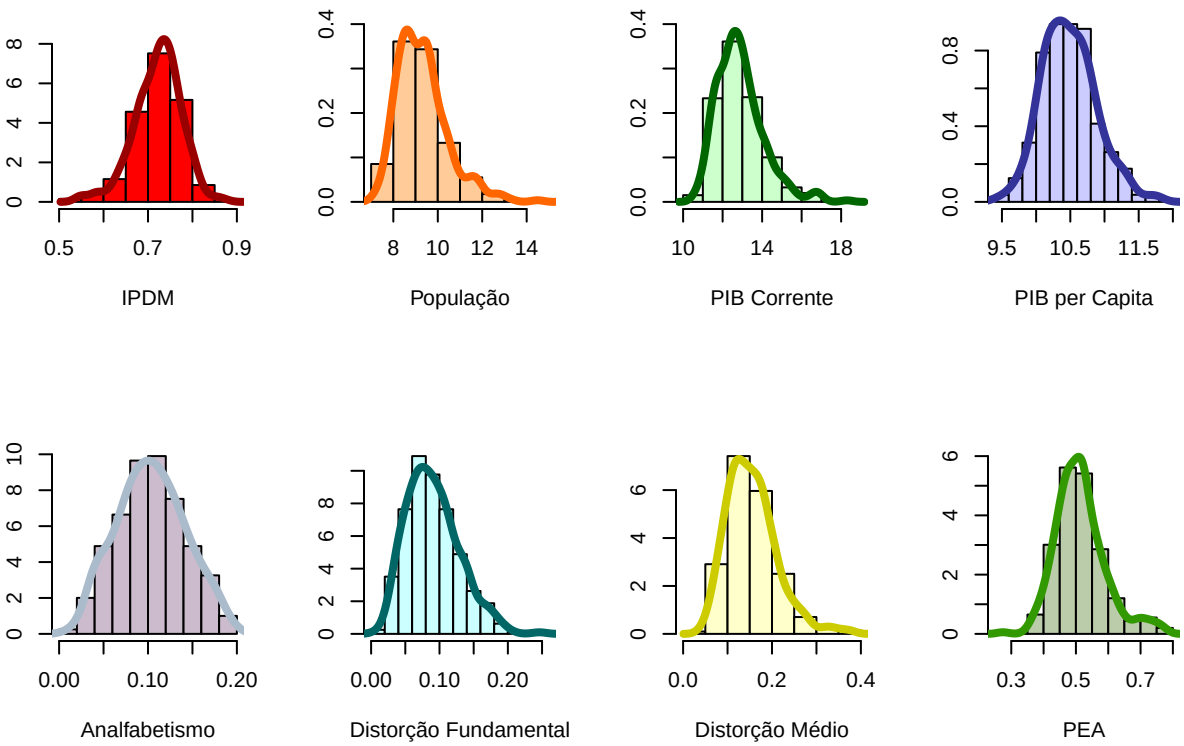
Histogramas

Abaixo temos o histograma para a variável resposta Gini.

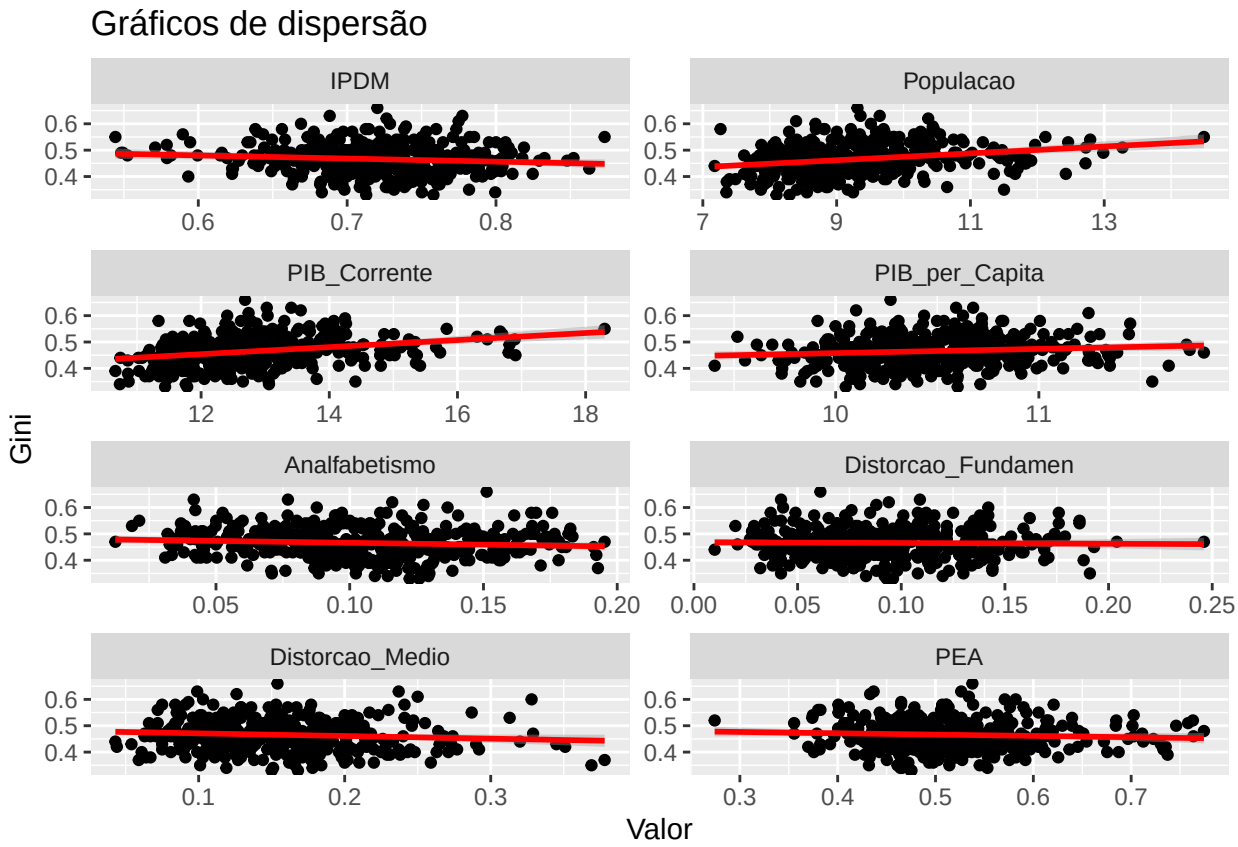


Marginalmente, a variável resposta Gini apresenta uma distribuição aproximadamente simétrica. Isso significa que há uma concentração de valores em torno de um pico central no histograma, indicando uma distribuição relativamente uniforme dos índices de Gini no conjunto de dados.

Abaixo, tem-se os histogramas das variáveis explicativas. Nota-se que, assim como a variável resposta, as variáveis IPDM, PIB per Capita, Analfabetismo e PEA também exibem uma distribuição aproximadamente simétrica.



Abaixo temos um gráfico de dispersão com retas de regressão ajustadas para cada uma das variáveis explicativas em relação à variável resposta Gini. O gráfico foi organizado em um grid, com cada variável representada em um painel separado. O eixo x representa os valores das variáveis explicativas e o eixo y representa o valor da variável resposta Gini.



Pontos a destacar:

1. IPDM: É a variável que apresenta uma reta de regressão ajustada com uma inclinação negativa mais acentuada. Isso indica na direção de que há uma relação negativa entre essa variável e o Gini, ou seja, quanto maior o valor de IPDM, menor tende a ser o valor de Gini.
2. PIB_Corrente e População: São as variáveis que apresentam uma reta de regressão ajustada com uma inclinação positiva mais acentuada. Isso sugere uma relação positiva entre essas variável e o Índice de Gini.
3. Demais variáveis: Não são observadas tendências claras em relação à variável resposta Gini. Isso sugere que essas variáveis podem ter uma influência limitada ou não linear sobre o Gini.

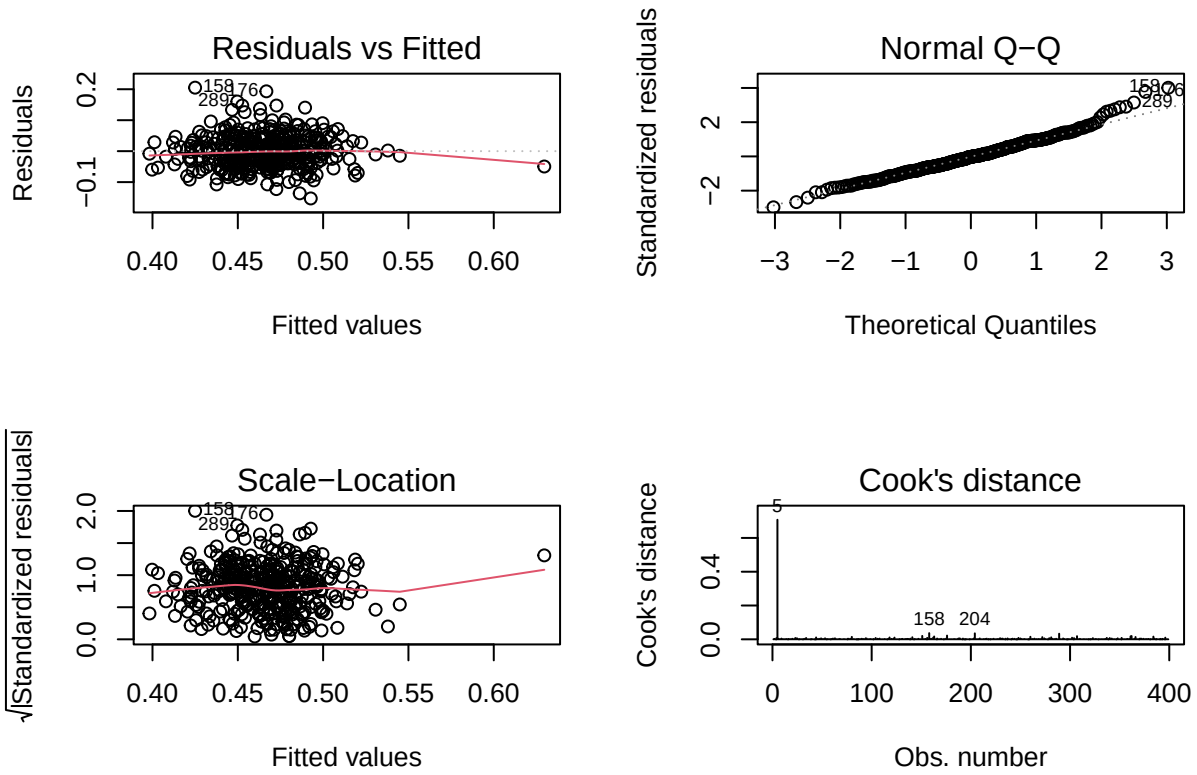
Modelo

Inicialmente, ajusta-se o modelo de regressão linear múltipla com todas as variáveis de forma aditiva e todas as observações.

A expressão do modelo proposto é dada por:

$$y_i|x_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = E(y_i|x_i) = \beta_0+\beta_1 ipdm+\beta_2 popul+\beta_3 pib\_corr+\beta_4 pib\_cap+\beta_5 analf+\beta_6 dist\_fund+\beta_7 dist\_medio+\beta_8 pea$$



O gráfico do canto esquerdo superior traz os resíduos ordinários versus os valores ajustados. Percebe-se os pontos aleatoriamente distribuídos e uma linha aproximadamente constante ao decorrer do gráfico. Porém, um ponto a direita se destaca. Refere-se ao município de Altamira do Paraná. Pode ser um indicativo de que a imputação pela média não foi uma boa escolha, ou que o munícipio está com outros dados também incorretos, pois esse munícipio também aprenseta a maior distância de Cook e é o mais isolado a direita no gráfico dos resíduos padronizados.

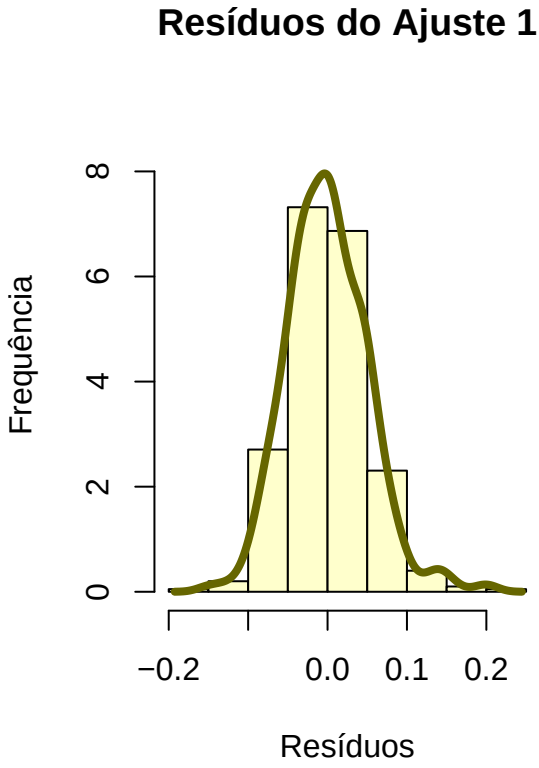
No canto direito superior mostra os quantis teóricos da distribuição Normal padrão contra os resíduos padronizados. Nesse caso, indicam uma suposta normalidade, com uma fuga considerável na extremidade superior.

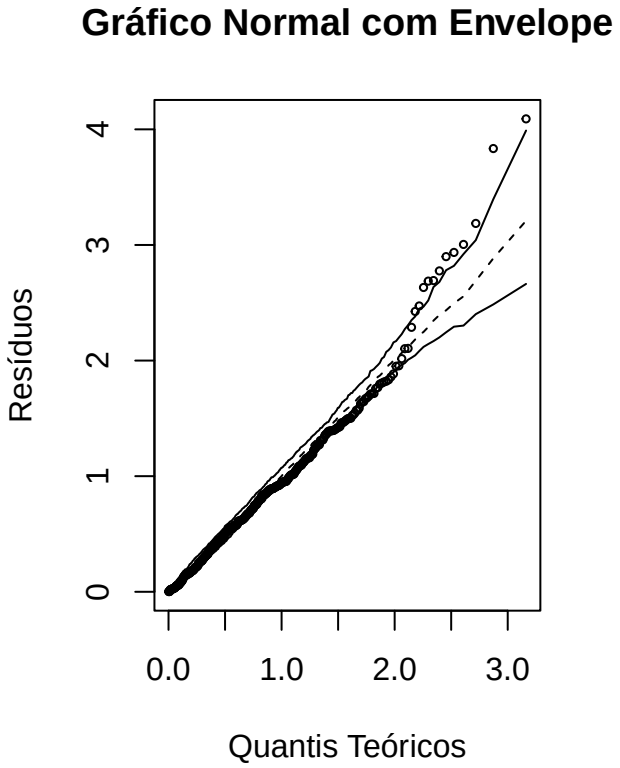
Ao inferior esquerdo apresenta-se a raiz quadrada dos resíduos padronizados versus os valores ajustados. É possível inferir uma variância constante por conta da linha central ser aproximadamente linear.

E por fim, o gráfico do canto inferior direito apresenta os valores da distância de Cook para cada observação. Destaca-se a observação 5, referente ao município de Altamira do Paraná.

Os gráficos de resíduos indicam aparente homocedasticidade, e uma observação atípica.

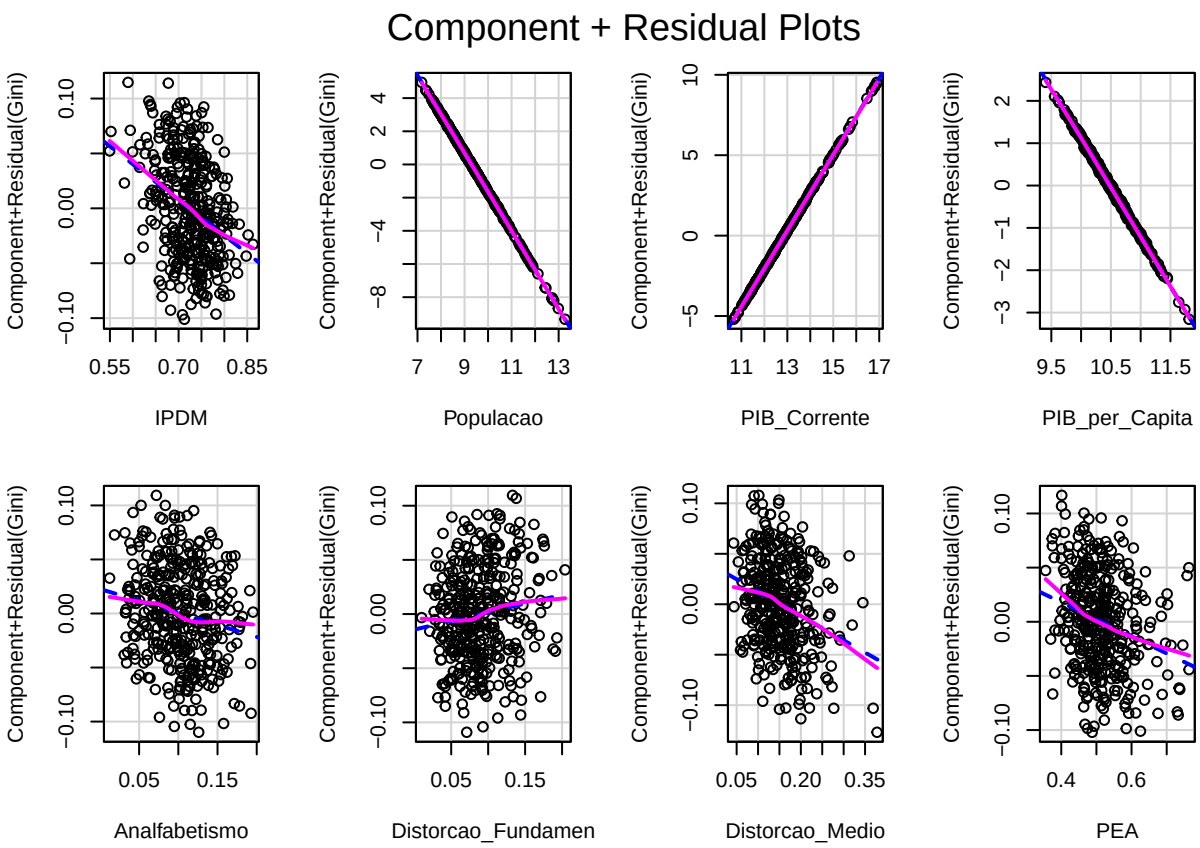
Verificação de normalidade dos resíduos no ajuste 1





Ao realizar o teste Shapiro-Wilk, onde H_0: Os resíduos têm distribuição normal e H_1: Os resíduos não têm distribuição normal; temos evidência para rejeitar a hipótese nula e, portanto, admitir não normalidade dos dados. O que confere com a análise gráfica apresentada acima.

Para avaliar a influência de cada observação no modelo, são calculadas as seguintes medidas de influência: DFBETAS, DFFITS, COVRATIO, distância de Cook e elementos da matriz H. Em seguida, foi criado um subconjunto de dados que exclui as observações identificadas como influentes de acordo com as medidas calculadas anteriormente. Abaixo tem-se os gráficos de comparação de resíduos ajustados para esse novo modelo.



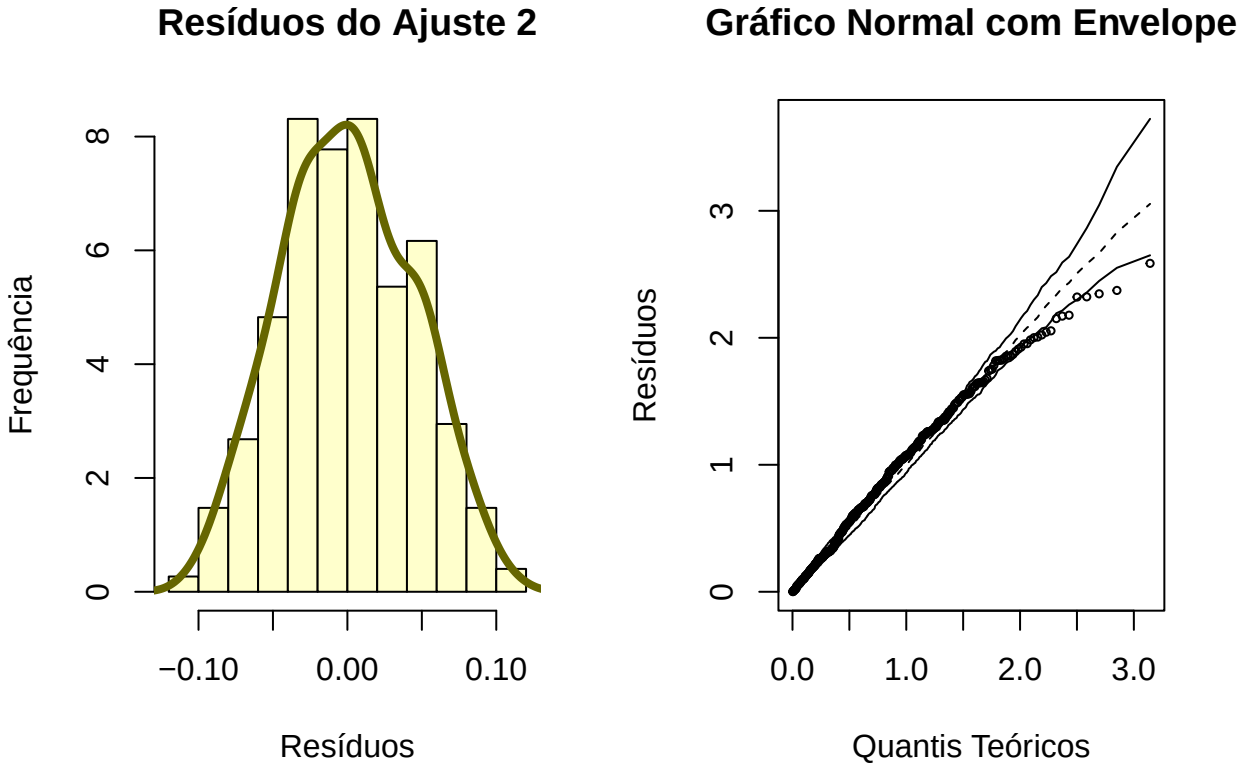
Através desse gráfico, é possível avaliar a adequação do modelo, identificar possíveis padrões nos resíduos e verificar se há violações das suposições do modelo de regressão. Destacam-se as seguintes fortes relações: crescente quanto a variável *PIB_Corrente*, e decrescente quanto as variáveis *Populacao* e *PIB_per_Capita*. As fracas relações entre as variáveis *Analfabetismo* e *Distorcao_Fundamen* e os resíduos ajustados do modelo de regressão sugerem que essas variáveis podem não contribuir de forma significativa para a explicação da variabilidade dos dados. Para fins de simplificação do modelo essas variáveis foram excluídas da regressão múltipla. A decisão se mostrou condizente com os dados ao se analisar que o coeficiente de determinação (R-quadrado) não sofreu alterações significativas. Essa constatação fortalece a justificativa para a exclusão dessas variáveis, pois indica que elas têm pouca influência na previsão da variável resposta.

Sendo assim, a expressão do novo modelo, excluindo as variáveis citadas, é dada por:

$$y_i|x_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = E(y_i|x_i) = \beta_0 + \beta_1 ipdm + \beta_2 popul + \beta_3 pib\_corr + \beta_4 pib\_cap + \beta_5 dist\_medio + \beta_6 pea$$

Verificação de normalidade dos resíduos no ajuste 2



Ao analisar os gráficos dos resíduos do novo ajuste , observa-se uma melhora na adequação desses resíduos à distribuição normal. O histograma dos resíduos apresenta uma distribuição mais simétrica e a curva de densidade está mais próxima de uma distribuição normal. Além de que a fuga dos pontos mais superiores no gráfico com envelope é menor. Isso sugere que o modelo ajustado está capturando de forma mais precisa o padrão dos dados e reduzindo a presença de desvios significativos.

Essa sugestão de melhor adequação à distribuição normal é confirmada pelo teste de *Shapiro-Wilk* realizado nos resíduos do ajuste1. O resultado do teste apresentou um *p-valor* de 0,06207, o que indica que não há evidências suficientes para rejeitar a hipótese nula (H0) de que os resíduos seguem uma distribuição normal. Embora o p-valor esteja ligeiramente acima do nível de significância de 0,05 adotado, essa diferença é pequena e pode ser considerada dentro da margem de erro. Portanto, com base nos resultados do teste, é razoável admitir a hipótese nula e considerar que os resíduos se aproximam de uma distribuição normal.

Resultados e discussão

Considerando o último ajuste, temos os seguintes resultados:

1. Intercept:

Conclusão