Trabalho Final da disciplina FLP0468 - Métodos Quantitativos de Pesquisa na Ciência Política IV (2024)

Giovanna Claudino de Almeida Silva

2025-08-16

Replicação do artigo "A redução da maioridade penal diminui a violência? Evidências de um estudo comparado"

O artigo *"A redução da maioridade penal diminui a violência? Evidências de um estudo comparado"*, publicado em 2016 por Rodrigo Lins, Dalson Figueiredo Filho e Lucas Silva, investiga a hipótese de que a redução da maioridade penal contribui para a diminuição da violência. O estudo adota um desenho metodológico que combina análise espacial, estatística descritiva e modelos de regressão linear, a partir de dados coletados de 197 países. Com foco em duas variáveis principais — maioridade penal, coletada por Hazel (2008) e por Grand Valley (2012), e idade de responsabilidade criminal, coletada por Hazel (2008) e Cipriani (2009) — a pesquisa busca compreender as relações entre os limites legais de imputabilidade e os indicadores de violência, representados pela taxa de homicídios. Os resultados sugerem que os países com maioridade penal mais alta tendem a apresentar níveis menores de violência. Os coeficientes das regressões apontam para uma relação inversa entre essas variáveis, indicando que a redução da maioridade penal pode estar associada a um aumento na taxa de homicidios.

Avaliação inicial da base de dados

Há muitos dados faltantes nas variáveis analisadas, principalmente em: gini, desemprego, responsabilidade criminal da Hazel e maioridade penal da Hazel. Esse desencontro de informações é problemático, pois pode gerar vieses no modelo. Por exemplo, a regressão da Hazel aparenta estar mais concentrada em países com IDHs médios a altos — e, se realmente estiver enviesada, é provavelmente porque ela conseguiu poucos dados de países africanos, como é mencionado no artigo. Para garantir a validade do modelo, é necessário que a amostra tenha um tamanho mínimo, de forma a capturar toda a variabilidade dos dados. Quanto maior o número de observações (N), menor a probabilidade de viés, como explica a Lei dos Grandes Números. Essa limitação na amostra representa um erro crucial, que inviabiliza extrapolações do modelo para previsões ou inferências causais.

É importante mencionar que foram utilizadas várias medidas de maioridade penal e responsabilidade criminal. Embora algumas dessas medidas apresentem menos dados faltantes, nenhuma possui uma base de dados suficientemente grande, o que seria o ideal. Segundo Peduzzi et al. (1996), o tamanho ideal de uma base de dados deve ser de 10 eventos (ou observações) por variável. Os autores demonstraram que, com menos de 10 eventos por variável, os coeficientes de regressão tendem a ser enviesados, as estimativas variam de forma excessiva, e os limites de confiança não têm cobertura adequada. No nosso caso, considerando quatro variáveis por regressão, o número ideal de observações seria 40. Embora a base de dados original possua 197 observações, a quantidade real será menor devido aos valores ausentes (NAs). Para determinar a quantidade exata de observações por regressão, verificaremos a quantidade de resíduos mais adiante. Caso o número de resíduos seja inferior a 40, a base de dados não terá o tamanho ideal.

Além disso, as demais variáveis também apresentam tamanhos variáveis, dependendo da quantidade de dados faltantes nas variáveis de maioridade penal e responsabilidade criminal. Isso ocorre porque, ao haver um NA em qualquer uma dessas variáveis, toda a linha é descartada, eliminando também observações válidas de outras variáveis. Fiz testes utilizando as duas variáveis da taxa de homicídios, e as duas variáveis de índice de Gini. Apesar da recomendação, troquei a variável ginet_solt pela variável ginmar_solt, já que com essa segunda variável eu obtive valores iguais as regressões apresentadas no artigo — e o intuito desse trabalho é replicar o artigo e avaliar sua qualidade.

```
url <- "https://raw.githubusercontent.com/gioclaudino/replicacao-maioridade-penal/main/maioridade.tab"

data_lins16 <- fread(url) %>%
  clean_names() %>%
  mutate(acr_cipriani = ifelse(acr_cipriani == 999, NA, acr_cipriani))

data_lins16$idh[155] <- 0.313

#Redução da base de dados para as variáveis de interesse

data_lins16 <- data_lins16 %>%
  select(
    -homi_number,
    -homi_number_unodc,
    -homi_rate,
    -ginet_solt,
    -cname) %>%
  rename(
    gini = ginmar_solt,
    taxa_homicidios = homi_rate_unodc,
    desemprego = desemprego_longo,
    rcrim_hz = acr_hazel,
    rcrim_cp = acr_cipriani,
    mpenal_hz = acm_hazel,
    mpenal_gv = acm_gv
    )

#Problemas com os dados: quantidade de NAs
#Há muitos dados faltantes, principalmente nas variáveis: gini, desemprego, rcrim_hz, mpenal_hz
#Contando a quantidade de NAs

na_por_coluna <- sapply(data_lins16, function(x) sum(is.na(x)))
na_por_coluna
```

```
##            gini      desemprego             idh taxa_homicidios        rcrim_hz
##             140             155              28              74             100
##        rcrim_cp       mpenal_hz       mpenal_gv
##               4             143              77
```

```
reg_mpenal_hz <- lm (taxa_homicidios ~ mpenal_hz + idh + gini + desemprego, data = data_lins16)
reg_mpenal_gv <- lm (taxa_homicidios ~ mpenal_gv + idh + gini + desemprego, data = data_lins16)
reg_rcrim_hz <- lm (taxa_homicidios ~ rcrim_hz + idh + gini + desemprego, data = data_lins16)
reg_rcrim_cp <- lm (taxa_homicidios ~ rcrim_cp + idh + gini + desemprego, data = data_lins16)
```

Os critérios de avaliação da regressão múltipla

Avaliação do modelo através do summary

Quando utilizamos a função summary em uma regressão múltipla gerada com lm, recebemos um resumo das principais informações daquela regressão. Primeiramente, aparece um resumo dos resíduos, que são os valores mínimo, 1º quartil, mediana, 3º quartil e máximo da distribuição desses resíduos. Essas informações ajudam a identificar possíveis assimetrias ou outliers nos resíduos, que podem afetar a validade do modelo — resíduos bem distribuídos indicam que o modelo se ajusta adequadamente aos dados.

O segundo item é referente à regressão em si, com os coeficientes estimados para o intercepto e para cada variável independente. Esses coeficientes indicam o efeito médio de uma unidade de aumento na variável explicativa sobre a variável dependente, mantendo as demais constantes. Para cada coeficiente, são fornecidos o erro-padrão, que mede a precisão da estimativa, o valor t, que testa se o coeficiente é significativamente diferente de zero, e o p-valor, que indica a significância estatística desse teste — valores baixos de p (geralmente abaixo de 0,05) sugerem que o coeficiente é relevante para o modelo.

Depois, temos o R^2 e o R^2 ajustado. No caso desse trabalho, o R^2 ajustado é mais relevante, pois ele é o indicado para regressões múltiplas, já que ele ajusta o R^2 básico (que mede a proporção da variância explicada pelo modelo) para o número de variáveis incluídas. Um R^2 ajustado alto sugere que o modelo explica bem a variabilidade da variável dependente, considerando o número de preditores. Em regressões múltiplas, ele é um dos principais indicadores da qualidade do modelo.

Intervalo de confiança

O intervalo de confiança é uma faixa de valores que estima a incerteza sobre um parâmetro, como dos coeficientes da regressão. Ele indica o intervalo dentro do qual esperamos que o valor real do parâmetro se encontre com uma determinada confiança (por exemplo, 95%). O intervalo de confiança pode ser calculado no R com o código confint(modelo, level = 0.95), onde "modelo" seria a regressão gerada na função lm. Esse código retorna os limites inferior e superior para os coeficientes do modelo, com base no nível de confiança especificado.

Checagem dos pressupostos

Homoscedasticidade

A homoscedasticidade é uma condição em regressões lineares em que os resíduos, ou diferenças entre os valores observados e estimados, apresentam variância constante em todos os níveis da variável explicativa. Essa característica é fundamental para garantir a precisão dos erros padrão e a validade das inferências estatísticas. A violação desse pressuposto, conhecida como heteroscedasticidade, pode comprometer a confiabilidade das estimativas, subestimando ou superestimando a variabilidade dos erros em diferentes partes do conjunto de dados. Isso pode indicar que variáveis relevantes estão ausentes do modelo ou que o modelo não está capturando adequadamente a relação entre as variáveis.

A verificação da homoscedasticidade será feita por meio de um teste e um gráfico. Primeiro, o gráfico Scale-Location avalia a distribuição da raiz quadrada dos resíduos padronizados em relação aos valores previstos. O ideal é que os pontos estejam distribuídos de forma aleatória e constante, com uma linha vermelha horizontal — a presença de qualquer padrão na distribuição dos resíduos representa uma violação do pressuposto. Segundo, o teste Breusch-Pagan (função bptest() do pacote lmtest em R) é um teste de hipóteses que avalia a homoscedasticidade. A hipótese nula assume homoscedasticidade, em um nível de significância de 0,05, enquanto valores menores indicam heteroscedasticidade. Lembrando que, esse pressuposto não pode ser avaliado diretamente, pois não temos o valor real do erro, e usamos os resíduos para fazer uma aproximação. Para utilizar o teste Breusch-Pagan, é importante que a regressão não tenha violado o pressuposto da normalidade.

Normalidade

A normalidade dos resíduos é um pressuposto importante da regressão linear que assume que os resíduos possuem uma distribuição normal com média 0. Esse pressuposto é essencial para a validade de testes estatísticos e intervalos de confiança. A violação da normalidade pode indicar problemas no modelo ou nas variáveis utilizadas, comprometendo a precisão das inferências.

Para verificar a normalidade, utiliza-se o gráfico Q-Q plot e o teste Shapiro-Wilk. O Q-Q plot compara os quantis dos resíduos observados com os quantis teóricos de uma distribuição normal; se os pontos seguem a linha diagonal, a normalidade é satisfeita. O teste Shapiro-Wilk testa a hipótese nula de que os resíduos seguem distribuição normal. Um p-valor maior que 0,05 sugere que a normalidade não foi violada. Se houver desvios significativos no gráfico ou um p-valor pequeno, é recomendável revisar o modelo ou considerar transformações das variáveis.

Linearidade

A linearidade entre a variável dependente e as independentes é o fundamento de um modelo de regressão linear. Esse pressuposto garante que o modelo captura adequadamente a relação entre as variáveis. Se a relação for não linear, o uso de um modelo linear não é apropriado.

A linearidade é avaliada com o gráfico "Residuals vs Fitted" e a função pair.panels(). No primeiro, os resíduos devem se distribuir aleatoriamente em torno da linha horizontal. Já no pair.panels, analisa-se a relação de cada variável independente com a dependente; uma linha reta vermelha sugere linearidade. Caso a linearidade seja violada, como evidenciado por padrões nos gráficos, é necessário ajustar o modelo com transformações ou considerar alternativas como modelos não lineares.

Multicolinearidade

A multicolinearidade ocorre quando variáveis independentes estão altamente correlacionadas, dificultando a estimativa dos efeitos individuais no modelo. Isso compromete a interpretação dos coeficientes e aumenta a variância das estimativas. Para verificar, utiliza-se a função pair.panels() para calcular os coeficientes de correlação de Pearson entre pares de variáveis; valores absolutos acima de 0,8 indicam alta correlação. Além disso, o fator de inflação da variância (VIF), calculado com a função vif, deve ser inferior a 10 para indicar ausência de multicolinearidade.

Esperança dos Resíduos Igual a 0

A esperança dos resíduos deve ser, em média, igual a 0, refletindo que a soma dos erros é balanceada em torno do valor estimado. Isso garante que o modelo está ajustado corretamente. Embora a média dos resíduos sempre seja 0 devido ao método dos mínimos quadrados, problemas podem ser identificados analisando o histograma ou gráfico de densidade dos resíduos. Os resíduos devem ser distribuídos de forma aleatória, com maior densidade em torno de 0. Uma distribuição enviesada pode indicar falhas no modelo ou variáveis ausentes.

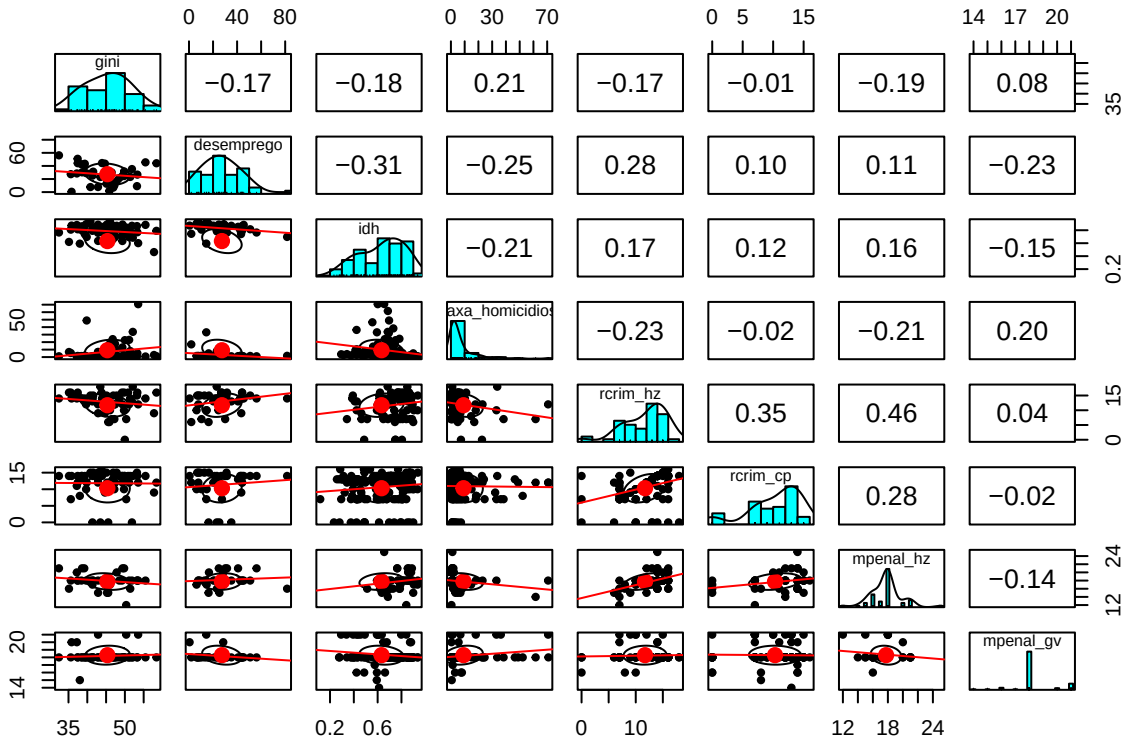
Independência dos Resíduos

A independência dos resíduos implica que os erros não estão correlacionados, sendo particularmente importante em dados temporais ou espaciais. Violar esse pressuposto pode comprometer a validade do modelo. Para verificar, utiliza-se o gráfico de dispersão dos resíduos contra um identificador (id) e o teste de Durbin-Watson (para esse teste, é importante que a regressão não tenha violado o pressuposto da normalidade). No gráfico, os resíduos devem ser distribuídos de forma aleatória. No teste Durbin-Watson, valores próximos de 2 indicam independência (intervalo ideal: 1 a 3). Um p-valor maior que 0,05 no teste sugere que a autocorrelação não é significativa.

Outliers

Outliers são valores extremos que podem influenciar negativamente o ajuste do modelo, atuando como pontos de alavancagem. Embora nem sempre comprometam a validade do modelo, podem distorcer a reta de regressão. Para detectá-los, utiliza-se o gráfico "Residuals vs Leverage" e a função rstandard para medir resíduos padronizados. Valores acima de 3 ou abaixo de -3 são considerados outliers. A exclusão de outliers deve ser feita com cautela, avaliando sua influência no modelo e justificando cientificamente sua remoção.

```
#PAINEL QUE APRESENTA A MULTICOLINEARIDADE E A LINEARIDADE DE TODAS AS VARIÁVEIS ENTRE SI NO data_lins1

pairs.panels(data_lins16, lm=TRUE)
```



Avaliação da regressão de maioridade penal da Hazel

- Análise dos Coeficientes: Os únicos coeficientes que demonstraram relevância estatística foram o intercepto e o IDH. A variável de interesse do artigo, mpenal_hz, tem relação negativa com a taxa de homicídios — quando a taxa aumenta em uma unidade, a maioridade penal diminui em 0,364. O modelo consegue explicar 25% da variabilidade da variável dependente. Essa regressão teve 29 resíduos, ou seja, 29 observações.
- Testando os Pressupostos
  - Multicolinearidade: A análise da correlação de Pearson mostrou que as variáveis independentes não apresentam correlações altas. O VIF não indicou problemas de multicolinearidade, com valores bem abaixo de 10.

- Intervalo de Confiança: O intervalo de confiança das variáveis IDH e do intercepto é grande, indicando incerteza nas estimativas. O coeficiente de desemprego apresentou o menor intervalo, ou seja, é o mais preciso.
- Linearidade: A variável Gini, IDH e a de maioridade penal apresentaram uma pequena relação linear (apesar de que a reta do IDH está sendo influenciada por um ponto isolado, que faz a relação parecer mais linear). No entanto, o gráfico "Residuals vs Fitted" sugere violação do pressuposto de linearidade, pois a linha vermelha não segue a linha pontilhada.
- Normalidade dos Resíduos: O teste de Shapiro-Wilk (p=0,0116) rejeitou a hipótese de normalidade. O gráfico Q-Q também confirmou que a distribuição dos resíduos se aproximou da distribuição normal só no meio da reta, variando muito nas extremidades. Logo, além de violar o pressuposto, também invalida os resultados dos testes de homoscedasticidade e de independência — logo, a análise deve ser centralizada nos gráficos.
- Homoscedasticidade: O gráfico Scale-location indicou a violação deste pressuposto, pois a linha vermelha não foi horizontal e os pontos não estavam distribuídos aleatoriamente. O teste de Breusch-Pagan, apesar de ser inválido nesse contexto, também rejeitaria a hipótese nula de que há homoscedasticidade — ou seja, o pressuposto foi violado.
- Outliers: No gráfico "Residuals vs Leverage", dois pontos (102 e 182) estavam próximos dos limites de outliers, mas o summary dos resíduos padronizados (rstandard) mostrou que nem o máximo nem o mínimo ultrapassaram os valores 3 e -3, ou seja, não há violação desse pressuposto.
- Independência dos Resíduos: A análise dos resíduos visualmente não indicou violação da independência dos resíduos. O gráfico mostrou distribuição aleatória, o que sugere que os resíduos são independentes entre si.
- Esperança dos Resíduos = 0: O histograma e a média dos resíduos indicaram que a esperança é próxima de 0, o que não confirma o pressuposto, mas era o que se esperava acontecer.

```r
#TESTES PARA A REGRESSÃO DE MAIORIDADE PENAL DE HAZEL

#Resumo da regressão
summary(reg_mpenal_hz)
```

```
##
## Call:
## lm(formula = taxa_homicidios ~ mpenal_hz + idh + gini + desemprego,
##     data = data_lins16)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1951 -0.8836 -0.2558  0.4244  4.2489
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.74336    6.32480   3.122  0.00464 **
## mpenal_hz    -0.36462    0.19702  -1.851  0.07656 .
## idh         -13.77780    5.68912  -2.422  0.02337 *
## gini          0.02525    0.05333   0.473  0.64017
## desemprego   -0.02511    0.02785  -0.902  0.37614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.563 on 24 degrees of freedom
##   (168 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.3644, Adjusted R-squared:  0.2584
## F-statistic: 3.439 on 4 and 24 DF,  p-value: 0.02337
```

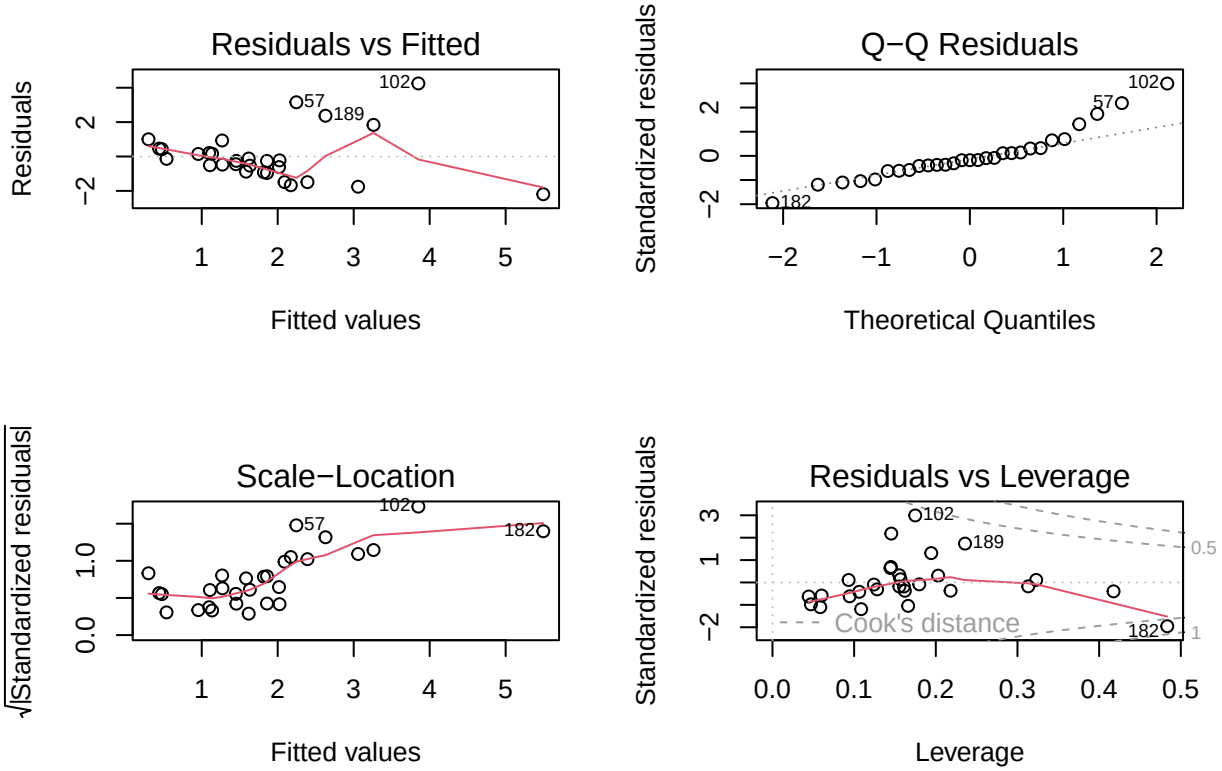
```
#Medida de multicolinearidade
vif(reg_mpenal_hz)
```

```
##  mpenal_hz        idh       gini desemprego
##   1.179338   1.217309   1.166489   1.448512
```

```
#Intervalo de confiança dos coeficientes
confint(reg_mpenal_hz, level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept)   6.68960550 32.79711398
## mpenal_hz    -0.77125457  0.04200603
## idh         -25.51956960 -2.03603069
## gini         -0.08482279  0.13532372
## desemprego   -0.08259530  0.03236565
```

```
#Gráfico que apresenta: linearidade, normalidade, homoscedasticidade, e outliers
par(mfrow=(c(2,2)))
plot(reg_mpenal_hz)
```



```r
#Teste de normalidade
shapiro.test(residuals(reg_mpenal_hz))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(reg_mpenal_hz)
## W = 0.90305, p-value = 0.0116
```

```r
#Teste de homoscedasticidade
bptest(reg_mpenal_hz)
```

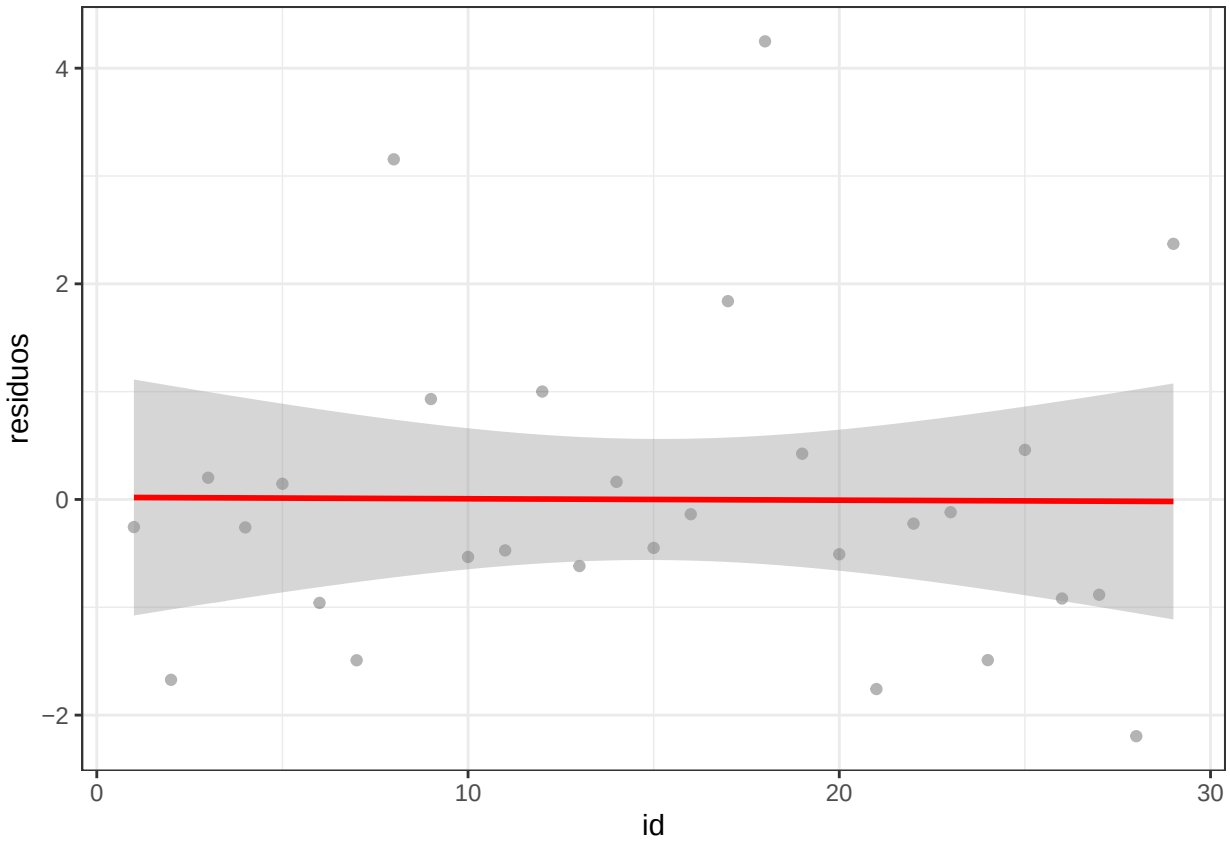
```
##
##  studentized Breusch-Pagan test
##
## data:  reg_mpenal_hz
## BP = 10.877, df = 4, p-value = 0.02798
```

```r
#Outliers
summary(rstandard(reg_mpenal_hz))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.954260 -0.583183 -0.178719 -0.004181  0.304174  2.992863
```

```r
#Gráfico de independência
residuos_mpenal_hz <- data.frame (
  residuos = residuals(reg_mpenal_hz)) %>%
  mutate(
    id = row_number())

residuos_mpenal_hz %>%
  ggplot(aes(x = id, y = residuos)) +
  geom_point(
    alpha = 0.3) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



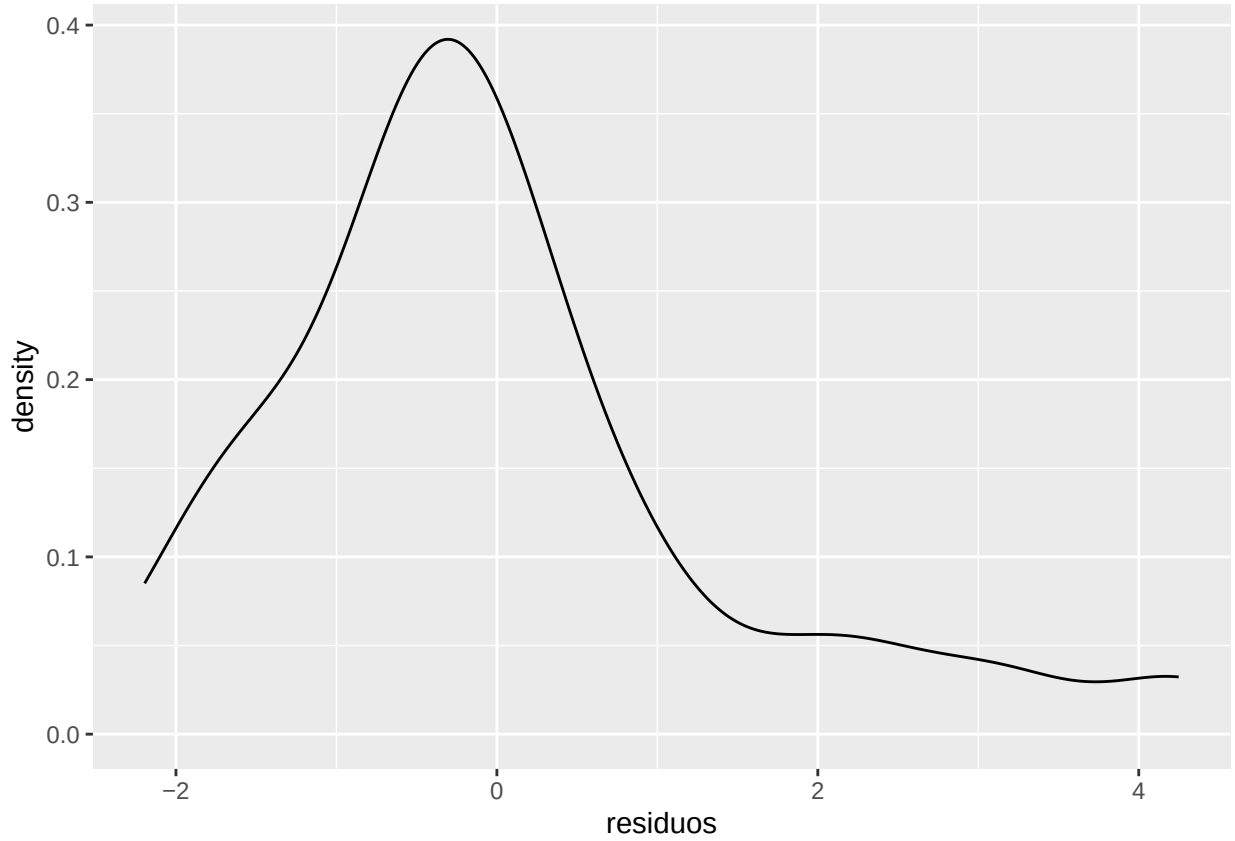

```
#Teste de independencia
durbinWatsonTest(reg_mpenal_hz)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.08793468      1.727119     0.4
##  Alternative hypothesis: rho != 0
```

```
#Gráfico de densidade e média dos resíduos, pressuposto da esperança = 0
residuos_mpenal_hz %>%
   summarise(
    media = round(mean(residuos), 10))
```

```
##   media
## 1     0
```

```
residuos_mpenal_hz %>%
  ggplot(aes(x= residuos)) +
  geom_density()
```



Avaliação da regressão de maioridade penal da Grand Valley State University

- Análise dos Coeficientes: Os únicos coeficientes que demonstraram relevância estatística foram o desemprego e o IDH — sendo que esse último demonstrou grande relevância. O IDH possui uma relação negativa com a taxa de homicídios, diminuindo em 28,973 a cada uma unidade da variável dependente. A variável de interesse do artigo, mpenal_gv, tem relação negativa com a taxa de homicídios — quando a taxa aumenta em uma unidade, a maioridade penal diminui em 0,224. O modelo consegue explicar 30% da variabilidade da variável dependente. Essa regressão teve 37 resíduos, ou seja, 37 observações.
- Testando os Pressupostos
  - Multicolinearidade: A análise da correlação de Pearson mostrou que as variáveis independentes não apresentam correlações altas. O VIF não indicou problemas de multicolinearidade, com valores bem abaixo de 10.
  - Intervalo de Confiança: O intervalo de confiança das variáveis IDH e do intercepto é grande, indicando incerteza nas estimativas. O coeficiente de desemprego continuou sendo o que apresentou o menor intervalo, ou seja, é o mais preciso.
  - Linearidade: As observações sobre o pairs.panels continuam iguais a da regressão anterior. O gráfico "Residuals vs Fitted" sugere violação do pressuposto de linearidade, pois a linha vermelha não segue a linha pontilhada.
  - Normalidade dos Resíduos: O teste de Shapiro-Wilk (p=0,0002) rejeitou a hipótese de normalidade. Porém, vemos no gráfico Q-Q que a distribuição dos resíduos se aproximou da distribuição

normal, mas que há um ponto que está muito distante, provavelmente influenciando o teste. Logo, irei considerar que o pressuposto não foi violado, principalmente se confirmarmos que há um outlier muito influente (ponto 114).

- Homoscedasticidade: O gráfico Scale-location indicou a violação deste pressuposto, pois a linha vermelha não foi horizontal e os pontos não estavam distribuídos aleatoriamente. O teste de Breusch-Pagan também rejeitou a hipótese nula de que há homoscedasticidade (p=0,006) — ou seja, o pressuposto foi violado.
- Outliers: No gráfico "Residuals vs Leverage", dois pontos aparecem próximos dos limites de outliers, principalmente o ponto 114. O summary dos resíduos padronizados (rstandard) mostrou que o valor máximo dos resíduos foi 4,464, muito acima do valor ideal de 3. Nesse caso, além de haver uma violação do pressuposto, seria indicado investigar esse caso, para analisar se é plausível removê-lo — pois podemos afirmar que ele está agindo como um ponto de alavancagem.
- Independência dos Resíduos: A análise dos resíduos visualmente e a partir do teste não indicou violação da independência dos resíduos. O gráfico mostrou distribuição aleatória, o que sugere que os resíduos são independentes entre si.
- Esperança dos Resíduos = 0: O histograma e a média dos resíduos indicaram que a esperança é próxima de 0, o que não confirma o pressuposto, mas era o que se esperava acontecer.

```
#TESTES PARA A REGRESSÃO DE MAIORIDADE PENAL DE GRAND VALLEY STATE UNIVERSITY

#Resumo da regressão
summary(reg_mpenal_gv)
```

```
##
## Call:
## lm(formula = taxa_homicidios ~ mpenal_gv + idh + gini + desemprego,
##     data = data_lins16)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7813 -1.3517 -0.2813  0.9336  9.5668
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.80042   24.76404   1.365 0.181805
## mpenal_gv    -0.22444    1.31017  -0.171 0.865065
## idh         -28.97369    7.71838  -3.754 0.000696 ***
## gini         -0.01190    0.07449  -0.160 0.874118
## desemprego   -0.10327    0.03275  -3.153 0.003500 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.501 on 32 degrees of freedom
##   (160 observations deleted due to missingness)
## Multiple R-squared:  0.3821, Adjusted R-squared:  0.3048
## F-statistic: 4.947 on 4 and 32 DF,  p-value: 0.003213
```

```
#Medida de multicolinearidade
vif(reg_mpenal_gv)
```

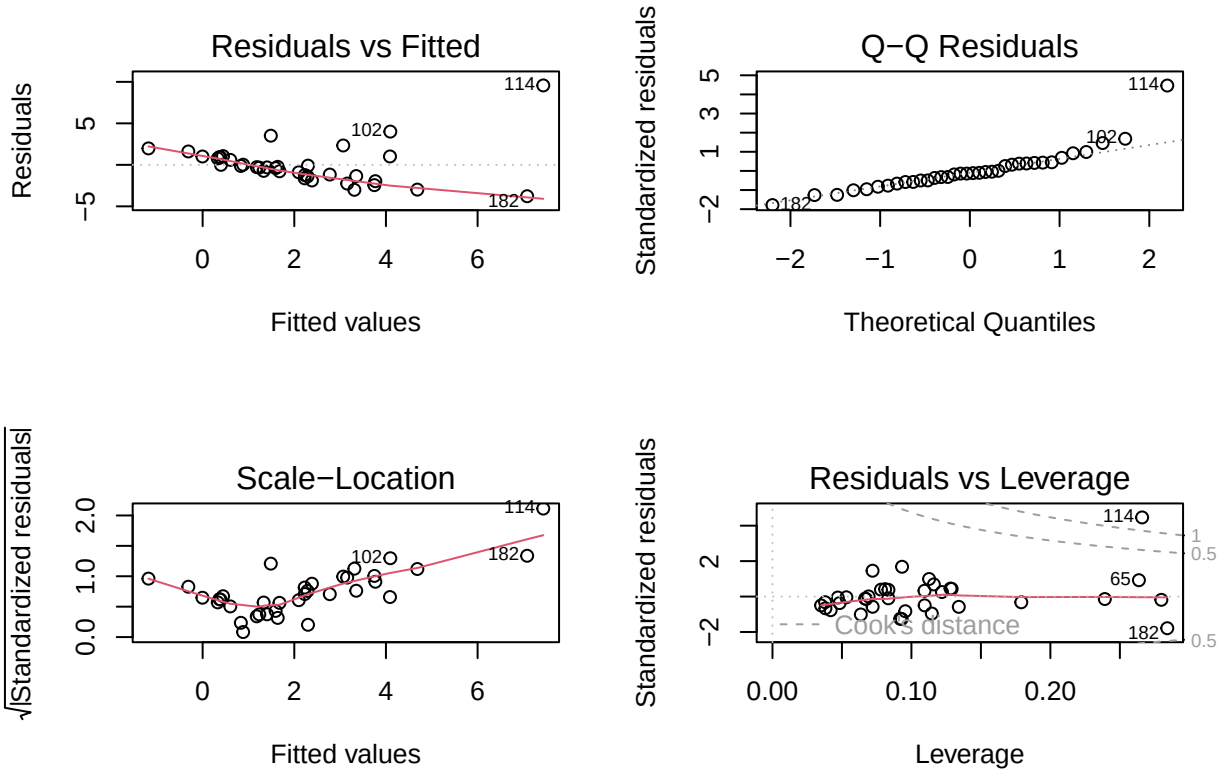
```
##  mpenal_gv        idh       gini desemprego
##   1.067810   1.102253   1.126522   1.161630
```

```r
#Intervalo de confiança dos coeficientes
confint(reg_mpenal_gv, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept) -16.6422676  84.24311640
## mpenal_gv    -2.8931648   2.44429415
## idh         -44.6955093 -13.25187479
## gini         -0.1636235   0.13983142
## desemprego   -0.1699787  -0.03655631
```

```r
#Gráfico que apresenta: linearidade, normalidade, homoscedasticidade, e outliers
par(mfrow=(c(2,2)))
plot(reg_mpenal_gv)
```

```
## Warning: not plotting observations with leverage one:
##   19
```




```r
#Teste de normalidade
shapiro.test(residuals(reg_mpenal_gv))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(reg_mpenal_gv)
## W = 0.86144, p-value = 0.0002944
```

```r
#Teste de homoscedasticidade
bptest(reg_mpenal_gv)
```

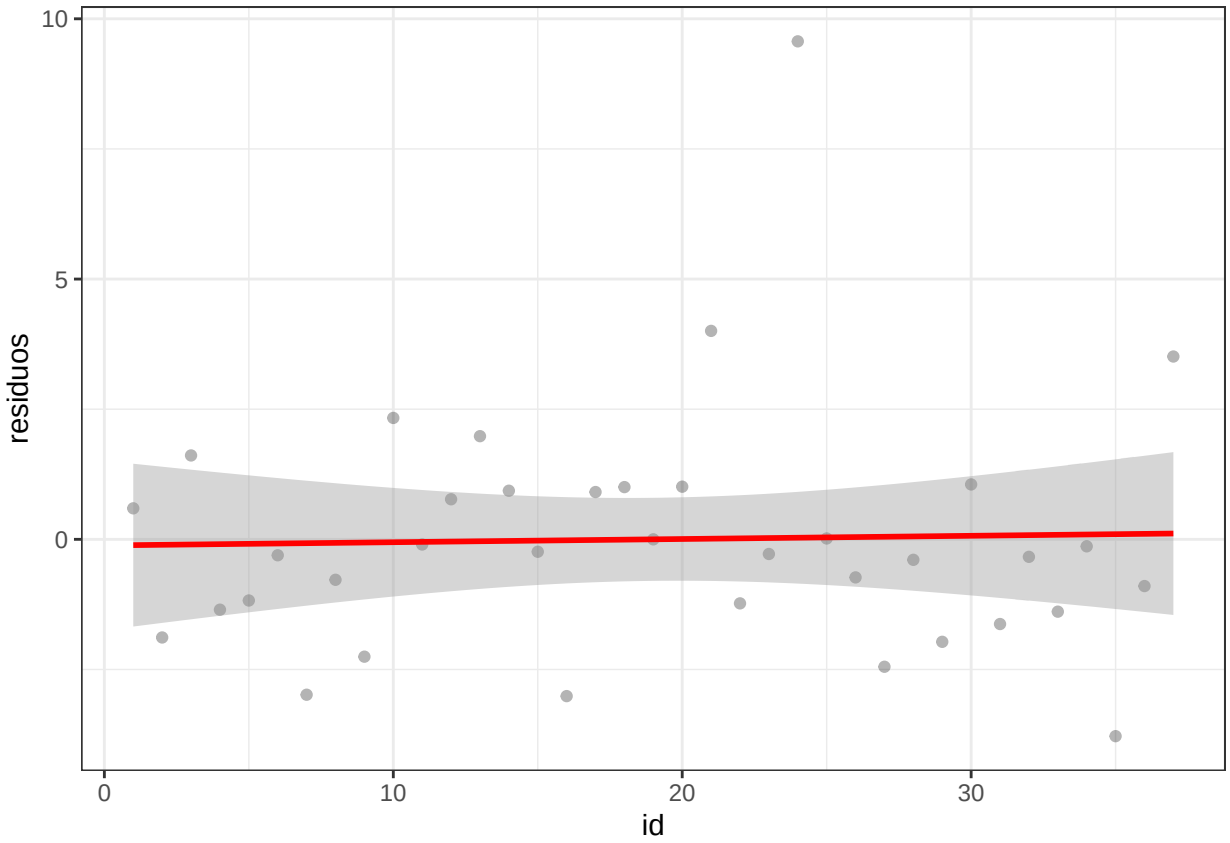
```
##
##  studentized Breusch-Pagan test
##
## data:  reg_mpenal_gv
## BP = 14.293, df = 4, p-value = 0.006416
```

```r
#Outliers
summary(rstandard(reg_mpenal_gv))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -1.78672 -0.57766 -0.12782  0.01076  0.39619  4.46442        1
```

```r
#Gráfico de independência
residuos_mpenal_gv <- data.frame (
  residuos = residuals(reg_mpenal_gv)) %>%
  mutate(
    id = row_number())

residuos_mpenal_gv %>%
  ggplot(aes(x = id, y = residuos)) +
  geom_point(
    alpha = 0.3) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



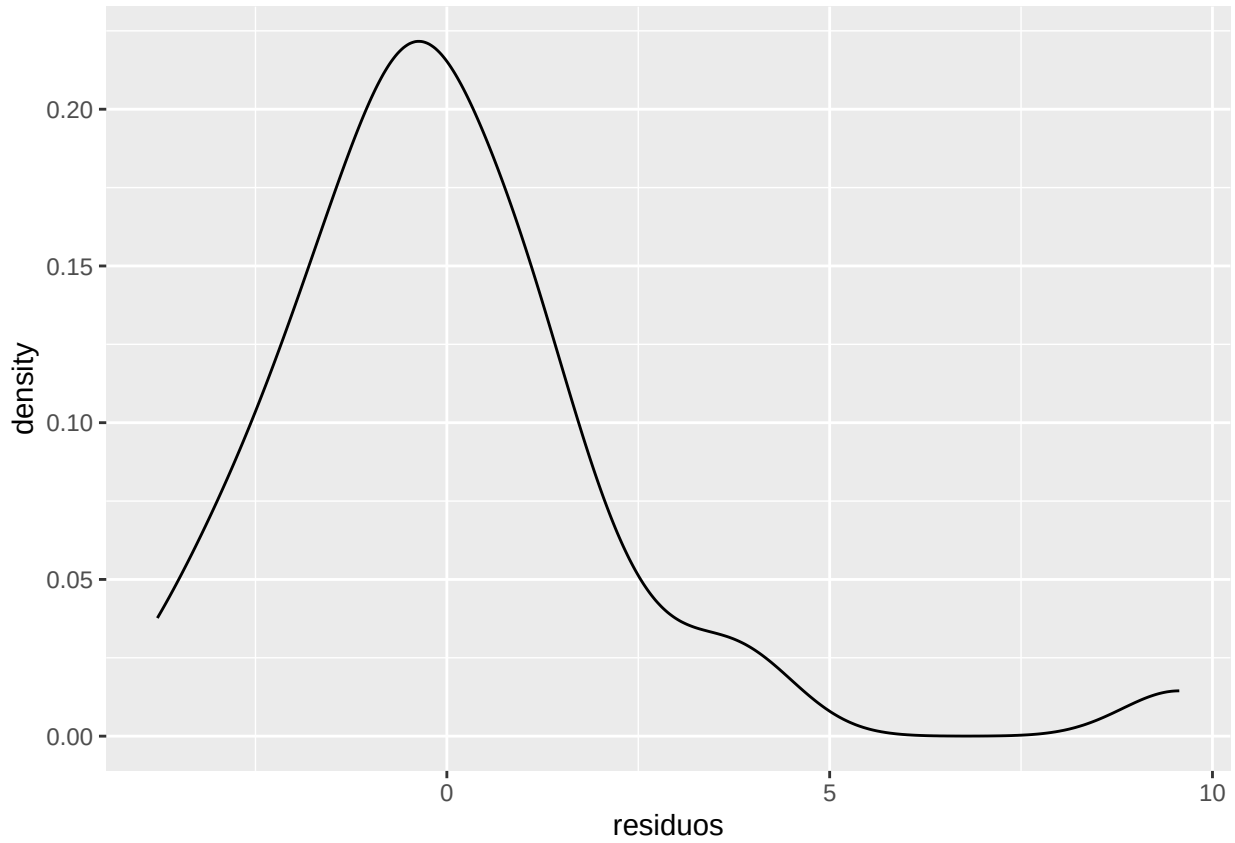

```
#Teste de independencia
durbinWatsonTest(reg_mpenal_gv)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.0215553      1.979726   0.928
##  Alternative hypothesis: rho != 0
```

```
#Gráfico de densidade e média dos resíduos, pressuposto da esperança = 0
residuos_mpenal_gv %>%
   summarise(
    media = round(mean(residuos), 10))
```

```
##   media
## 1     0
```

```
residuos_mpenal_gv %>%
  ggplot(aes(x= residuos)) +
  geom_density()
```



Avaliação da regressão de responsabilidade criminal da Hazel

- Análise dos Coeficientes: Os únicos coeficientes que demonstraram relevância estatística foram o intercepto, o IDH (alta relevância) e o desemprego. O IDH possui uma relação negativa com a taxa de homicídios, diminuindo em 28,574 a cada uma unidade da variável dependente. A variável de interesse do artigo, rcrim_hz, tem relação negativa com a taxa de homicídios — quando a taxa aumenta em uma unidade, a responsabilidade criminal diminui em 0,271. O modelo consegue explicar 37% da variabilidade da variável independente. Essa regressão teve 37 resíduos, ou seja, 37 observações.
- Testando os Pressupostos
  - Multicolinearidade: A análise da correlação de Pearson mostrou que as variáveis independentes não apresentam correlações altas. O VIF não indicou problemas de multicolinearidade, com valores bem abaixo de 10.
  - Intervalo de Confiança: O intervalo de confiança das variáveis IDH e do intercepto continuam sendo as maiores, indicando incerteza nas estimativas. O coeficiente de desemprego continuou sendo o mais preciso.
  - Linearidade: No pairs.panels, temos as mesmas conclusões sobre as variáveis Gini e IDH, mas a variável de responsabilidade criminal que usamos nessa regressão apresenta uma relação linear mais evidente. No entanto, o gráfico "Residuals vs Fitted" sugere violação do pressuposto de linearidade, pois a linha vermelha não segue a linha pontilhada.
  - Normalidade dos Resíduos: O teste de Shapiro-Wilk (p=0,0055) rejeitou a hipótese de normalidade. Porém, vemos no gráfico Q-Q que a distribuição dos resíduos se aproximou da distribuição normal, mas que há um ponto que está muito distante, provavelmente influenciando o teste.

Logo, irei considerar que o pressuposto não foi violado, principalmente porque o ponto 114 já foi demonstrado como um outlier.

- Homoscedasticidade: O gráfico Scale-location indicou a violação deste pressuposto, pois a linha vermelha não foi horizontal e os pontos não estavam distribuídos aleatoriamente. O teste de Breusch-Pagan também rejeitou a hipótese nula de que há homoscedasticidade (p=0,0002) — ou seja, o pressuposto foi violado.
- Outliers: No gráfico "Residuals vs Leverage", um ponto está próximo do limite inferior de outliers, e outro ponto ultrapassou o limite superior. No summary dos resíduos padronizados (rstandard), vemos que o valor máximo ultrapassou o valor indicado de 3 (4,302), o que indica um outlier influente — o mesmo ponto que afetou a regressão anterior.
- Independência dos Resíduos: A análise dos resíduos visualmente não indicou violação da independência dos resíduos. O gráfico mostrou distribuição aleatória, o que sugere que os resíduos são independentes entre si.
- Esperança dos Resíduos = 0: O histograma e a média dos resíduos indicaram que a esperança é próxima de 0, o que não confirma o pressuposto, mas era o que se esperava acontecer.

```
#TESTES PARA A REGRESSÃO DE RESPONSABILIDADE CRIMINAL DE HAZEL

#Resumo da regressão
summary(reg_rcrim_hz)
```

```
##
## Call:
## lm(formula = taxa_homicidios ~ rcrim_hz + idh + gini + desemprego,
##     data = data_lins16)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3266 -1.4022 -0.2095  0.6973  8.1287
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.399665   7.445155   4.352 0.000129 ***
## rcrim_hz     -0.271311   0.138934  -1.953 0.059635 .
## idh         -28.574970   7.242636  -3.945 0.000408 ***
## gini         -0.009776   0.068860  -0.142 0.887989
## desemprego   -0.086084   0.032188  -2.674 0.011692 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.365 on 32 degrees of freedom
##   (160 observations deleted due to missingness)
## Multiple R-squared:  0.4474, Adjusted R-squared:  0.3783
## F-statistic: 6.476 on 4 and 32 DF,  p-value: 0.0006173
```

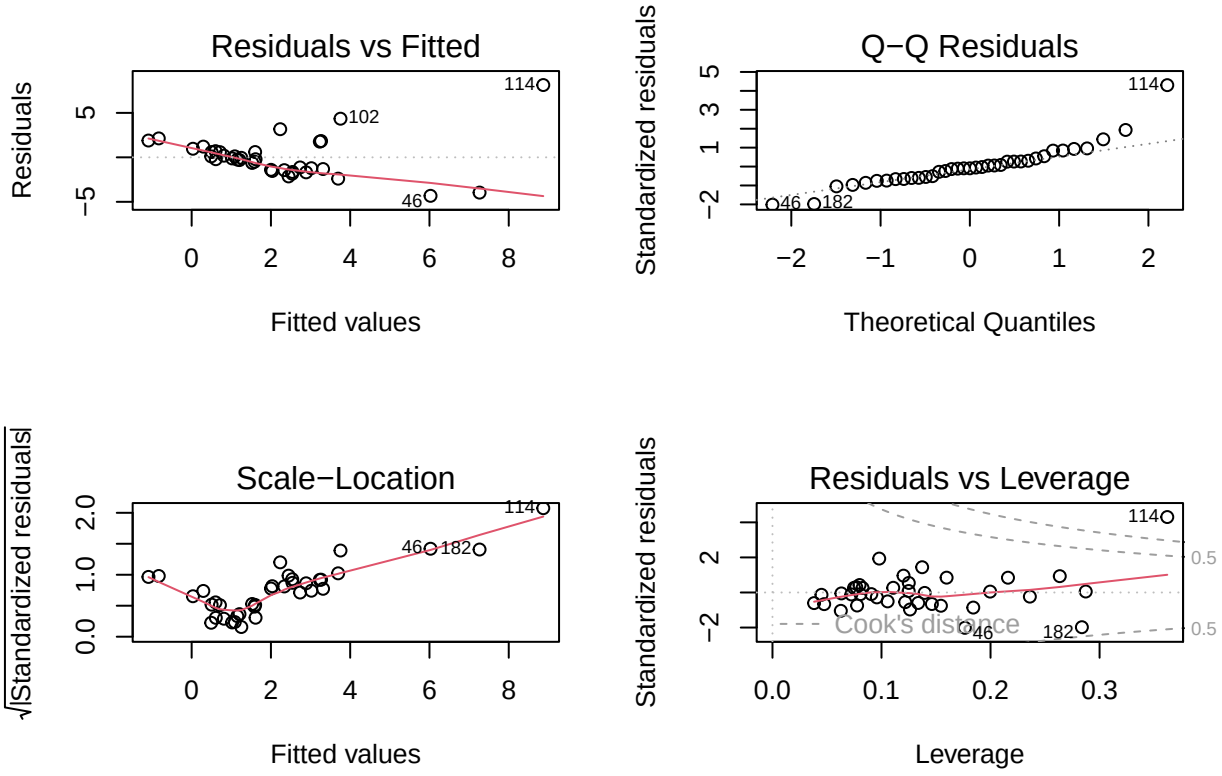
```
#Medida de multicolinearidade
vif(reg_rcrim_hz)
```

```
##   rcrim_hz        idh       gini desemprego
##   1.088636   1.085228   1.076449   1.254607
```

```r
#Intervalo de confiança dos coeficientes
confint(reg_rcrim_hz, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept)  17.2343792  47.56495004
## rcrim_hz     -0.5543092   0.01168758
## idh         -43.3277360 -13.82220322
## gini         -0.1500391   0.13048616
## desemprego   -0.1516486  -0.02051937
```

```r
#Gráfico que apresenta: linearidade, normalidade, homoscedasticidade, e outliers
par(mfrow=(c(2,2)))
plot(reg_rcrim_hz)
```




```r
#Teste de normalidade
shapiro.test(residuals(reg_rcrim_hz))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(reg_rcrim_hz)
## W = 0.9098, p-value = 0.005583
```

```r
#Teste de homoscedasticidade
bptest(reg_rcrim_hz)
```

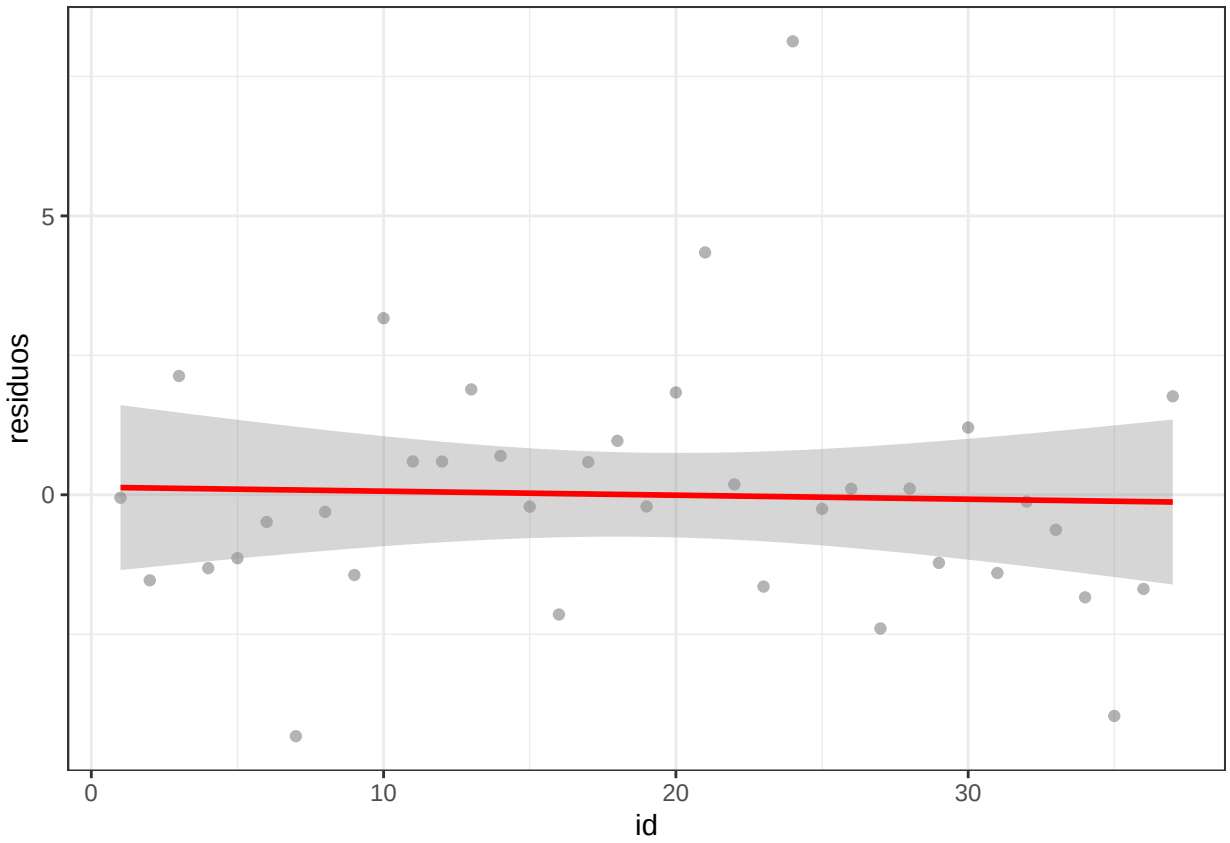
```
##
##  studentized Breusch-Pagan test
##
## data:  reg_rcrim_hz
## BP = 21.508, df = 4, p-value = 0.0002511
```

```r
#Outliers
summary(rstandard(reg_rcrim_hz))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.01555 -0.60442 -0.09287  0.01461  0.30680  4.30255
```

```r
#Gráfico de independência
residuos_rcrim_hz <- data.frame (
  residuos = residuals(reg_rcrim_hz)) %>%
  mutate(
    id = row_number())

residuos_rcrim_hz %>%
  ggplot(aes(x = id, y = residuos)) +
  geom_point(
    alpha = 0.3) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



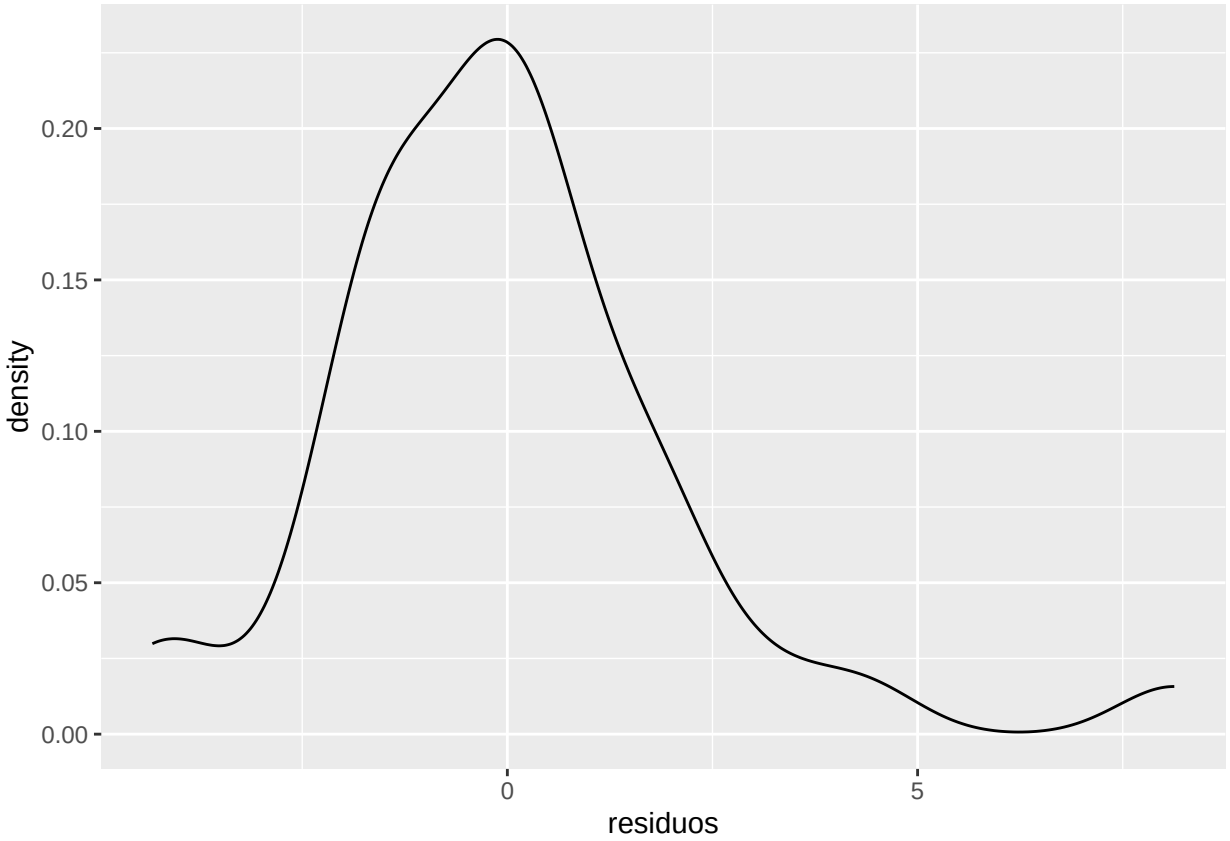

```
#Teste de independencia
durbinWatsonTest(reg_rcrim_hz)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.003940992      1.974688   0.914
##  Alternative hypothesis: rho != 0
```

```
#Histograma e média dos resíduos, pressuposto da esperança = 0
residuos_rcrim_hz %>%
   summarise(
    media = round(mean(residuos), 10))
```

```
##   media
## 1     0
```

```
residuos_rcrim_hz %>%
  ggplot(aes(x= residuos)) +
  geom_density()
```



Avaliação da regressão de responsabilidade criminal de Cipriani

- Análise dos Coeficientes: Os únicos coeficientes que demonstraram relevância estatística foram o intercepto, o IDH (alta relevância) e o desemprego (relevância maior do que nas regressões anteriores, mas ainda menor que o IDH). O IDH e o desemprego possuem uma relação negativa com a taxa de homicídios, diminuindo, respectivamente, 29,454 e 0,102 a cada uma unidade da variável dependente. A variável de interesse do artigo, rcrim_cp, tem relação negativa com a taxa de homicídios — quando a taxa aumenta em uma unidade, a responsabilidade criminal diminui em 0,048. O modelo consegue explicar 31% da variabilidade da variável independente. Essa regressão teve 37 resíduos, ou seja, 37 observações.
- Testando os Pressupostos
  - Multicolinearidade: A análise da correlação de Pearson mostrou que as variáveis independentes não apresentam correlações altas. O VIF não indicou problemas de multicolinearidade, com valores bem abaixo de 10.
  - Intervalo de Confiança: O intervalo de confiança das variáveis IDH e do intercepto continuam sendo as maiores, indicando incerteza nas estimativas. O coeficiente de desemprego continuou sendo o mais preciso.
  - Linearidade: No pairs.panels, temos as mesmas conclusões sobre as variáveis Gini e IDH, mas a variável de responsabilidade criminal do Cipriani é a que menos apresenta uma relação linear, já que a linha está completamente reta. O gráfico "Residuals vs Fitted" confirma a violação do pressuposto de linearidade, pois a linha vermelha não segue a linha pontilhada.

- Normalidade dos Resíduos: O teste de Shapiro-Wilk (p=0,0002) rejeitou a hipótese de normalidade. Porém, vemos no gráfico Q-Q que a distribuição dos resíduos se aproximou da distribuição normal, mas que há um ponto que está muito distante, provavelmente influenciando o teste. Logo, irei considerar que o pressuposto não foi violado, principalmente porque o ponto 114 já foi demonstrado como um outlier.
- Homoscedasticidade: O gráfico Scale-location indicou a violação deste pressuposto, pois a linha vermelha não foi horizontal e os pontos não estavam distribuídos aleatoriamente. O teste de Breusch-Pagan também rejeitou a hipótese nula de que há homoscedasticidade (p=0,006) — ou seja, o pressuposto foi violado.
- Outliers: No gráfico "Residuals vs Leverage", um ponto que ultrapassou o limite superior. No summary dos resíduos padronizados (rstandard), vemos que o valor máximo ultrapassou o valor indicado de 3 (4,488), o que indica um outlier influente — o mesmo ponto que afetou a regressão anterior.
- Independência dos Resíduos: A análise dos resíduos visualmente não indicou violação da independência dos resíduos. O gráfico mostrou distribuição aleatória, o que sugere que os resíduos são independentes entre si.
- Esperança dos Resíduos = 0: O histograma e a média dos resíduos indicaram que a esperança é próxima de 0, o que não confirma o pressuposto, mas era o que se esperava acontecer.

```r
#TESTES PARA A REGRESSÃO DE RESPONSABILIDADE CRIMINAL DE CIPRIANI

#Resumo da regressão
summary(reg_rcrim_cp)
```

```
##
## Call:
## lm(formula = taxa_homicidios ~ rcrim_cp + idh + gini + desemprego,
##     data = data_lins16)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8159 -1.2526 -0.1829  1.0408  9.5826
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.52904    7.85380   3.887 0.000480 ***
## rcrim_cp     -0.04813    0.09347  -0.515 0.610140
## idh         -29.45416    7.64767  -3.851 0.000531 ***
## gini         -0.00863    0.07255  -0.119 0.906063
## desemprego   -0.10249    0.03266  -3.138 0.003640 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.492 on 32 degrees of freedom
##   (160 observations deleted due to missingness)
## Multiple R-squared:  0.3866, Adjusted R-squared:  0.3099
## F-statistic: 5.042 on 4 and 32 DF,  p-value: 0.002887
```

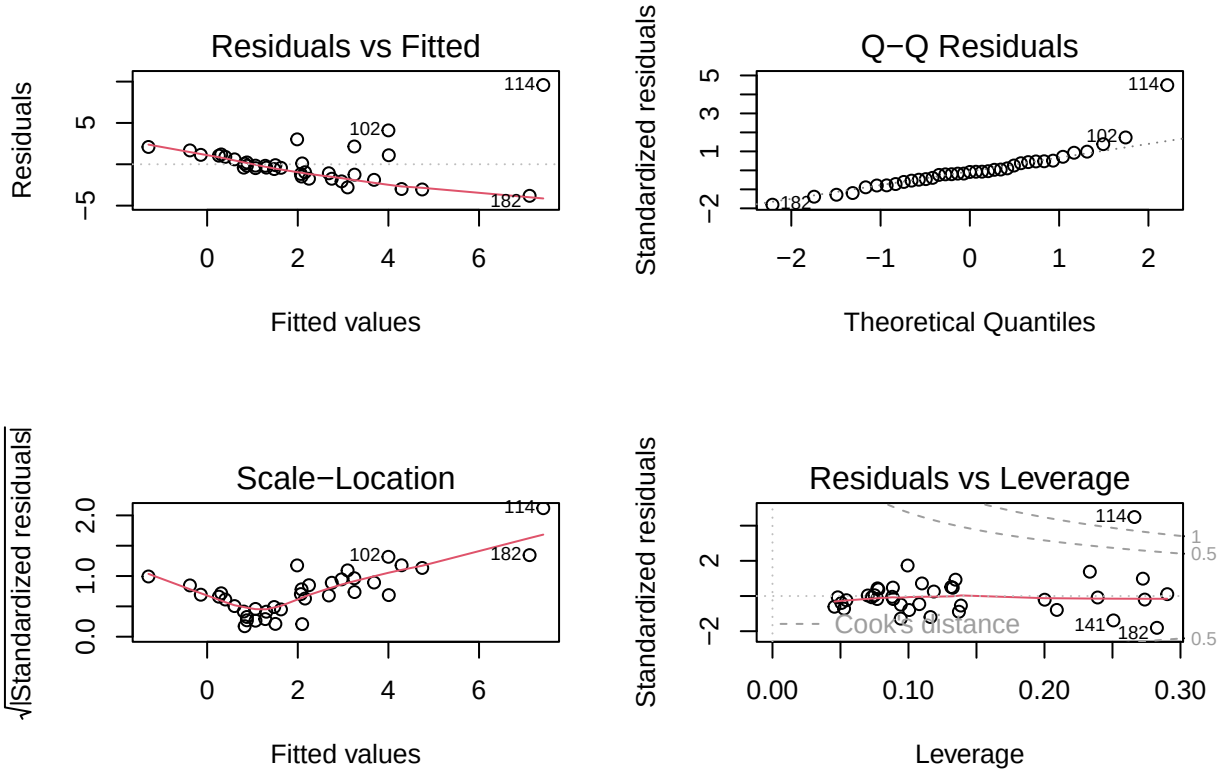
```r
#Medida de multicolinearidade
vif(reg_rcrim_cp)
```

```
##   rcrim_cp        idh       gini desemprego
##   1.010685   1.090117   1.076685   1.163637
```

```r
#Intervalo de confiança dos coeficientes
confint(reg_rcrim_cp, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept)  14.5313704  46.52672040
## rcrim_cp     -0.2385288   0.14226513
## idh         -45.0319442 -13.87636711
## gini         -0.1564203   0.13916021
## desemprego   -0.1690109  -0.03596211
```

```r
#Gráfico que apresenta: linearidade, normalidade, homoscedasticidade, e outliers
par(mfrow=(c(2,2)))
plot(reg_rcrim_cp)
```




```r
#Teste de normalidade
shapiro.test(residuals(reg_rcrim_cp))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(reg_rcrim_cp)
## W = 0.86147, p-value = 0.000295
```

```r
#Teste de homoscedasticidade
bptest(reg_rcrim_cp)
```

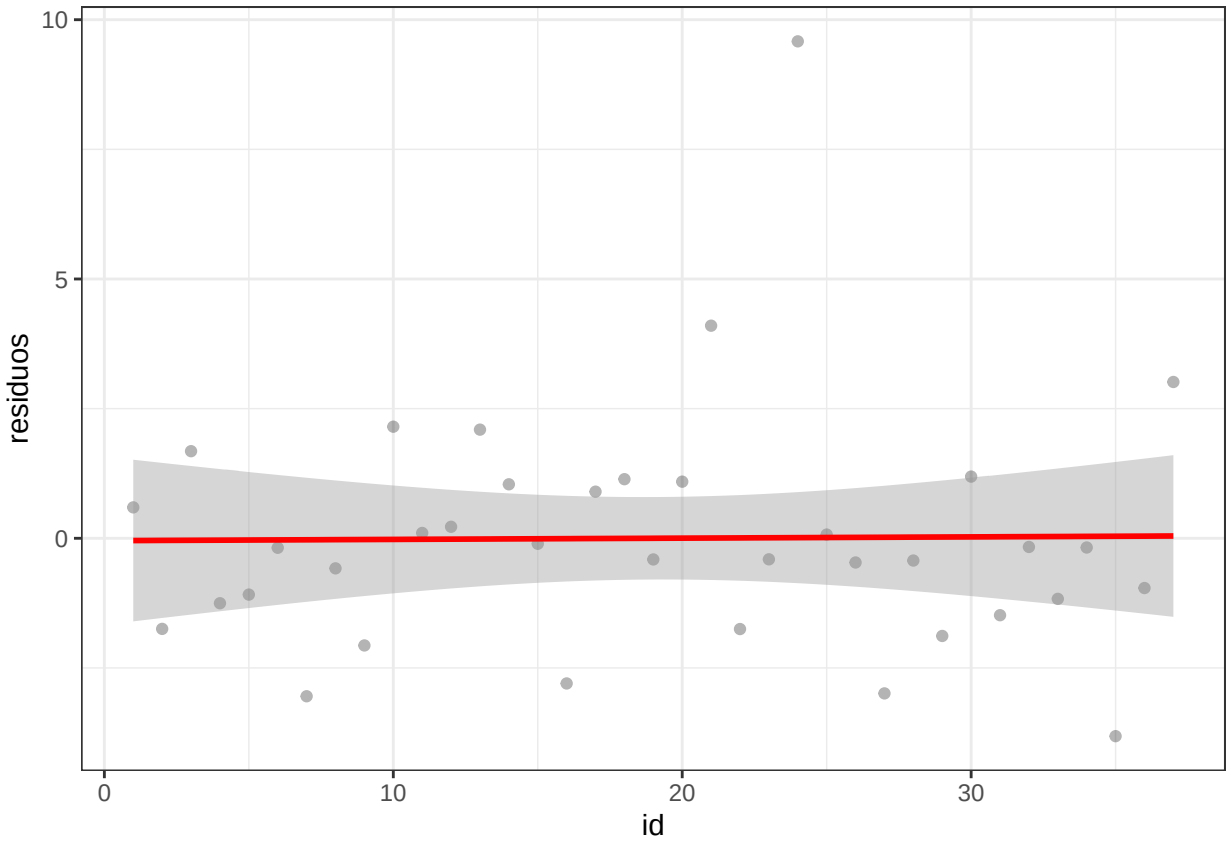
```
##
##  studentized Breusch-Pagan test
##
## data:  reg_rcrim_cp
## BP = 14.432, df = 4, p-value = 0.006038
```

```r
#Outliers
summary(rstandard(reg_rcrim_cp))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.807873 -0.541534 -0.084135  0.008076  0.434769  4.488554
```

```r
#Gráfico de independência
residuos_rcrim_cp <- data.frame (
  residuos = residuals(reg_rcrim_cp)) %>%
  mutate(
    id = row_number())

residuos_rcrim_cp %>%
  ggplot(aes(x = id, y = residuos)) +
  geom_point(
    alpha = 0.3) +
  geom_smooth(method = "lm", color = "red") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



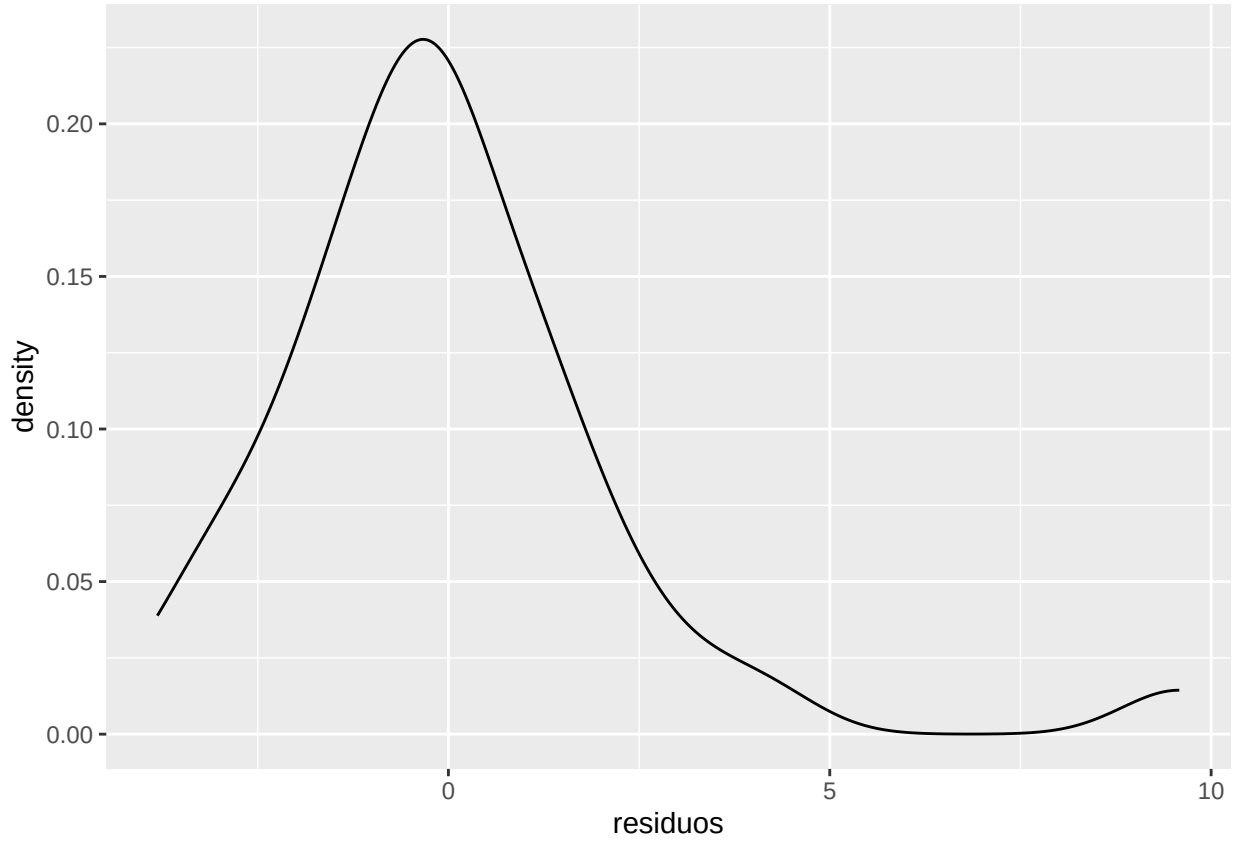

```
#Teste de independencia
durbinWatsonTest(reg_rcrim_cp)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.04235545      2.037214    0.91
##  Alternative hypothesis: rho != 0
```

```
#Histograma e média dos resíduos, pressuposto da esperança = 0
residuos_rcrim_cp %>%
   summarise(
    media = round(mean(residuos), 10))
```

```
##   media
## 1     0
```

```
residuos_rcrim_cp %>%
  ggplot(aes(x= residuos)) +
  geom_density()
```

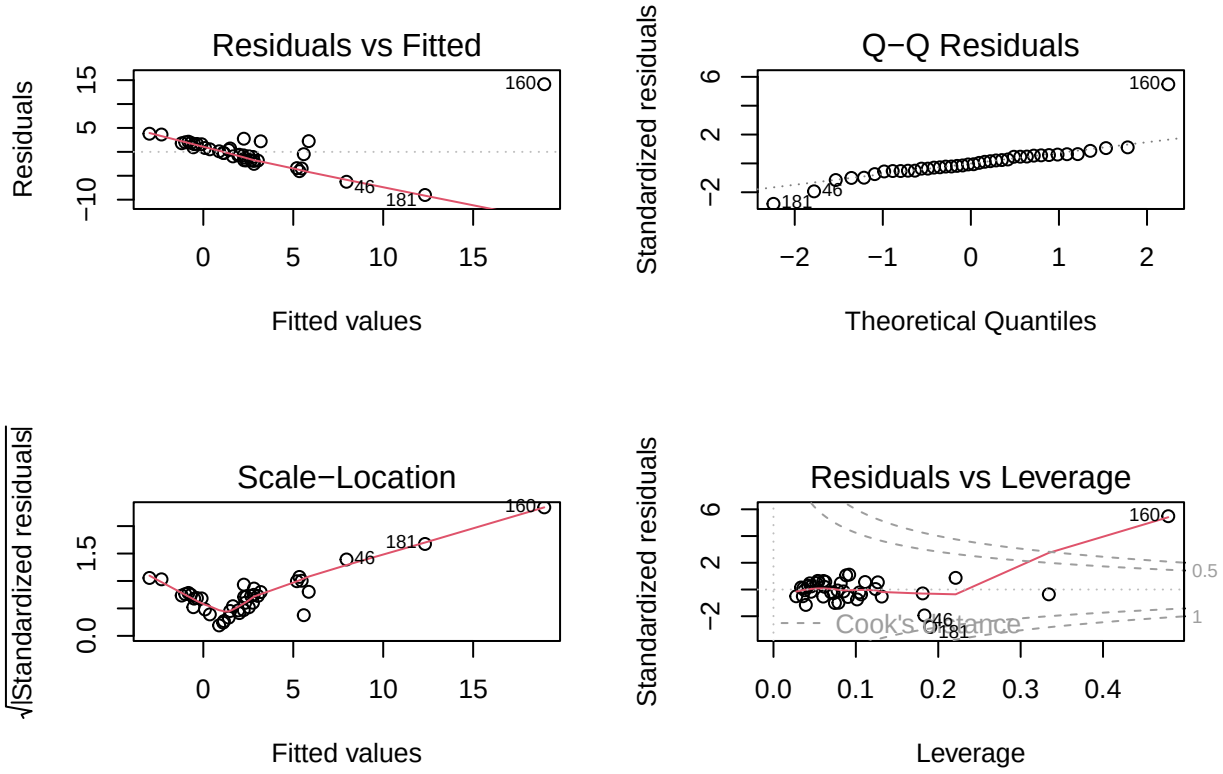



```
#Teste com as variáveis de maior relevância, e retirando o outlier 114

data_lins16teste <- data_lins16 [-114, ]

testereg <- lm(taxa_homicidios ~ idh + desemprego + rcrim_hz, data = data_lins16teste)
summary(testereg)
```

```
##
## Call:
## lm(formula = taxa_homicidios ~ idh + desemprego + rcrim_hz, data = data_lins16teste)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.022 -1.763 -0.226  1.691 14.148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.77090    8.19380   7.051 2.82e-08 ***
## idh         -57.46455    8.81090  -6.522 1.40e-07 ***
## desemprego   -0.13670    0.04013  -3.407  0.00163 **
## rcrim_hz     -0.27162    0.21777  -1.247  0.22035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.576 on 36 degrees of freedom
```

```
##   (156 observations deleted due to missingness)
## Multiple R-squared:  0.5648, Adjusted R-squared:  0.5285
## F-statistic: 15.57 on 3 and 36 DF,  p-value: 1.175e-06
```

```r
par(mfrow=(c(2,2)))
plot(testereg)
```




```r
#Os NAs das variáveis de maioridade penal e responsabilidade criminal alteram as características das ou

data_mpenal_hz <- model.frame(reg_mpenal_hz)
data_mpenal_gv <- model.frame(reg_mpenal_gv)
data_rcrim_hz <- model.frame(reg_rcrim_hz)
data_rcrim_cp <- model.frame(reg_rcrim_cp)

print(colMeans(data_lins16, na.rm = TRUE))
```

```
##           gini      desemprego             idh taxa_homicidios        rcrim_hz
##      45.3400157      27.3857143       0.6329941       9.1609756      11.7113402
##        rcrim_cp       mpenal_hz       mpenal_gv
##      10.3056995      17.7592593      18.2916667
```

```r
print(colMeans(data_mpenal_hz, na.rm = TRUE))
```

```
## taxa_homicidios       mpenal_hz             idh            gini      desemprego
##       1.8310345      18.3103448       0.8441034      44.0484362      28.6000000
```

```r
print(colMeans(data_mpenal_gv, na.rm = TRUE))
```

```
## taxa_homicidios       mpenal_gv             idh            gini      desemprego
##       2.1189189      18.0540541       0.8363243      44.6773664      27.7594595
```

```r
print(colMeans(data_rcrim_hz, na.rm = TRUE))
```

```
## taxa_homicidios        rcrim_hz             idh            gini      desemprego
##       2.1189189      13.1081081       0.8363243      44.6773664      27.7594595
```

```r
print(colMeans(data_rcrim_cp, na.rm = TRUE))
```

```
## taxa_homicidios        rcrim_cp             idh            gini      desemprego
##       2.1189189      11.3513514       0.8363243      44.6773664      27.7594595
```

Conclusões

A regressão que conseguiu explicar maior parte da variabilidade da variável dependente foi a de responsabilidade criminal da Hazel, mesmo assim com um valor muito baixo (37%). Selecionei as variáveis que apresentaram maior relevância estatística entre as regressões — IDH, desemprego e responsabilidade criminal da Hazel — e fiz uma regressão que consegui explicar 53% da variabilidade da variável dependente. Apesar disso, todas as regressões falharam na maioria das checagens, violando pressupostos essenciais para a validação do uso do modelo linear (nenhuma regressão passou no pressuposto da linearidade e da homoscedasticidade, por exemplo). Isso não significa que não haja uma relação entre as variáveis independentes com a dependente, já que vimos indícios dessas relações no pairs.panels, mas sim que o modelo de regressão linear não é adequado para esses dados.

Todas as regressões tinham uma base de dados menor do que o mínimo para criar um bom modelo, o que pode explicar os altos intervalos de confiança, e o p-valor tão baixo — os valores obtidos não trazem nenhuma confiança de que não estão enviesados, pois não tem o tamanho mínimo para essa garantia. A grande quantidade de NAs pode também influenciar na presença de outliers, já que tem poucos dados para balancear esse efeito. Vimos que a linha 114 alterou fortemente as estatísticas (que é a observação referente ao México), mas ao fazer um teste retirando ela, vemos que outros pontos aparecem influenciando fortemente as estatísticas. Comparando as médias das variáveis na base de dados original e em cada regressão, podemos ver algumas diferenças que podem decorrer da quantidade de NAs. A taxa de homicídio na base de dados foi de 9,160, e nas regressões o valor da taxa de homicídios mais próximo foi 2,118. Além disso, todas as variáveis tiveram uma média de IDH bem alta, o que também demonstra um viés dentro das regressões.

Concluo que esse artigo não tem capacidade de comprovar e nem negar a sua hipótese de que a redução na maioridade penal diminui os níveis de violência, por conta da falta de dados e da má adequação dos modelos de regressão, já que os pressupostos necessários para o uso da regressão linear não foram atendidos.