UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

ANTÔNIO ARTHUR SILVA DE LIMA
FRANCISCO GUSTAVO BRAGA BATISTA
ROMULO BARROS DE FREITAS

ANÁLISE DE RESÍDUOS EM MODELOS DE REGRESSÃO LINEAR SIMPLES

FORTALEZA
2023

ANTÔNIO ARTHUR SILVA DE LIMA
FRANCISCO GUSTAVO BRAGA BATISTA
ROMULO BARROS DE FREITAS

ANÁLISE DE RESÍDUOS EM MODELOS DE REGRESSÃO LINEAR SIMPLES

Relatório apresentado ao curso de Bacharelado em Estatística do Centro de Ciências da Universidade Federal do Ceará, como parte dos requisitos para a aprovação na disciplina de Modelos de Regressão I no semestre de 2023.2.
Prof.: Ronald Targino Nojosa.

FORTALEZA
2023

Sumário

Lista de Figuras

1 Introdução

A análise de resíduos desempenha papel fundamental em tópicos de Estatística e Ciência de Dados, sendo amplamente utilizada em Modelos de Regressão Linear (*MRL*) e em algoritmos de aprendizado de máquina.

Sua principal contribuição, é revelar o quão bem os modelos se ajustam, ou o quão adequado é aquele modelo, para o problema em questão.

Assim, uma maneira de encontrar tais resultados, é testar se as suposições inicialmente feitas para a construção do modelo, são satisfeitas, o que pode ser realizado através da análise residual.

2 Modelo de Regressão Linear Simples (MRLS)

O Modelo de Regressão Linear Simples, é uma técnica estatística que visa compreender e quantificar a relação entre uma variável dependente e uma única variável independente. O *MRLS* busca encontrar a linha reta que melhor se ajusta aos pontos de dados, de forma a minimizar a diferença entre as observações reais e as previsões feitas por essa reta.

A relação entre a variável dependente e a variável independente é representada por uma equação linear conhecida como "função de regressão", expressa na forma de uma equação:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

onde:

- β_0 e β_1 são os parâmetros;
- β_0 é o coeficiente linear;
- β_1 é o coeficiente angular;
- Y_i é a variável resposta/dependente;
- X_i é a variável explicativa/independente;
- ϵ_i é a fonte de variação/erro aleatório.

Para que os resultados do *MRLS* sejam confiáveis e interpretáveis, é crucial entender e reconhecer as suposições subjacentes que sustentam o seu uso. Estas suposições desempenham um papel fundamental na validade das estimativas e inferências derivadas do modelo.

2.1 Suposições do MRLS

- A função de regressão é linear nos parâmetros;
- Os valores de X são fixos e independentes;
- $E(\epsilon_i) = 0$;
- Os erros não são correlacionados, ou seja: $Cov(\epsilon_i, \epsilon_j) \underset{\forall i \neq j}{=} E(\epsilon_i, \epsilon_j) = 0$;

- Homocedasticidade: $V(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$;
- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$;
- $\epsilon_i \sim N(0; \sigma^2)$.

3 Análise residual

A análise de resíduos busca investigar se as suposições inicialmente feitas para a criação do MRLS são satisfeitas. Ou seja, permite avaliar a qualidade do ajuste do modelo e identificar possíveis violações das suposições feitas primordialmente.

Para isso, diversos métodos podem ser empregados, dentre eles, métodos gráficos e testes de hipótese. Entretanto, é necessário entender previamente o que são os resíduos, bem como seus tipos e significado.

3.1 Tipos de resíduos

3.1.1 Resíduo ordinário

A diferença entre o valor observado e o valor ajustado é chamado de **resíduo ordinário**:

$$e_i = y_i - \hat{y}_i,$$

que pode ser vista como a diferença entre o desvio de y_i em relação à $\bar{y}$ e o desvio de $\hat{y}_i$ em relação à $\bar{y}$:

$$e_i = y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}),$$

ou seja, mede a distância entre o valor observado e o valor previsto.

3.1.2 Resíduo padronizado

Outro tipo de resíduo comumente utilizado é o chamado **resíduo padronizado**, que, como o nome sugere, é uma padronização dos resíduos ordinários, e possui a seguinte forma:

$$z_i = \frac{e_i}{\sqrt{QM_{Res}}},$$

onde $\sqrt{QM_{Res}}$ é o erro padrão de $\hat{\sigma^2} = QM_{Res}$. Este resíduo é uma alternativa ao ordinário, que altera levemente a escala dos gráficos e facilita ver se há ou não normalidade (se sim, 95% dos resíduos devem estar entre -2 e 2).

3.1.3 Resíduo estudentizado

O último resíduo a ser visto é chamado **resíduo estudentizado**, possuindo a seguinte forma:

$$r_i = \frac{e_i}{\sqrt{QM_{Res}(1 - v_{ii})}}, \quad v_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}$$

Este tipo de resíduo torna-se uma opção mais poderosa quando queremos identificar se há *outliers* nos dados que possam influenciar o modelo ajustado de forma significativa.

3.2 Gráficos e interpretações

Métodos gráficos fazem parte da grande maioria de problemas científicos, e em análise residual não é diferente. Para verificar se as suposições iniciais são atendidas no *MRL* ajustado, pode-se criar gráficos de dispersão com (x_i, e_i), $(\hat{y}_i, e_i)$, (x_i, z_i) e (x_i, r_i), atentando-se a se os resíduos:

1. são aleatórios;
2. se concentram em torno do 0;
3. são homocedásticos.

Além dos gráficos de dispersão, o *qqplot* pode ser utilizado, mostrando se os resíduos estão muito afastados ou próximos da linha de correlação normal.

A figura 3.1, do Bussab e Morettin [3], mostra alguns tipos de gráficos da forma (x_i, e_i), que poderiam ser obtidos a partir de uma análise residual.

É possível notar facilmente que os gráficos (b) e (c) apontam que um *MRLS* não é o mais adequado, pois não há linearidade presente. Já nos gráficos (e), (f) e (g), é possível notar heterocedasticidade (variância muda com os valores de x), enquanto em (h) não há centralidade em torno do 0. Em (d), é possível ver a presença de um *outlier*, que requer uma investigação maior para testar sua influência no modelo, e finalmente, o gráfico (a) reflete as condições ideais apontadas anteriormente para se ter um bom ajuste.

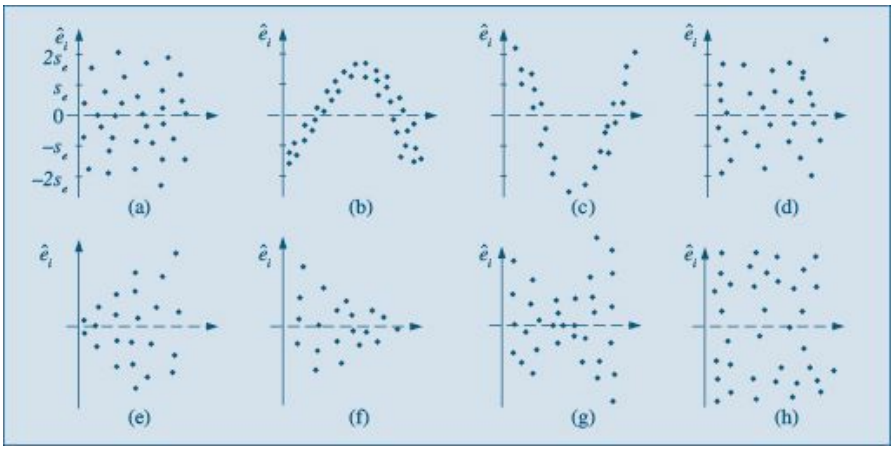

Figura 3.1: Interpretação gráfica dos resíduos

Há ainda o gráfico do tipo *Residuals vs Leverage*, que junto do gráfico de dispersão de (x_i, r_i), busca encontrar outliers com significativa influência sobre o modelo ajustado, por meio da distância de Cook, onde pontos fora dessa distância estabelecida, são considerados como pontos de influência.

Além dos gráficos, testes de hipótese também podem ser feitos para testar a normalidade, por exemplo, utilizando o teste de Shapiro-Wilk, ou criando intervalos de confiança para os parâmetros μ e σ^2.

4 Exemplos

A fim de melhor ilustrar o processo de análise de resíduos, trazemos duas aplicações de uso de *MRLS*, onde buscamos mensurar a qualidade de tais modelos, tendo um deles boa adequação, e outro com pouca adequação, para representar ambas as situações.

4.1 Caso 1

A seguir, serão apresentados alguns exemplos gráficos de análise residual utilizando a base de dados *USArrests*, do software gratuito conhecido como R, disponível para download no endereço www.rproject.org. O conjunto de dados escolhido traz estatísticas sobre crimes de assalto, assassinato e estupro praticados nos 50 estados dos EUA no ano de 1973. Além disso, o dataset traz o percentual da população urbana de cada estado.

Ao analisarmos o gráfico 4.1 que tem como objetivo explicar o número de prisões por assassinatos nos EUA através do percentual de população urbana de cada estado, notamos que as duas variáveis não possuem uma correlação expressiva. Fato este, que é comprovado quando calculamos a correlação de X: Percentual de População Urbana e Y: Número de prisões por assassinatos:

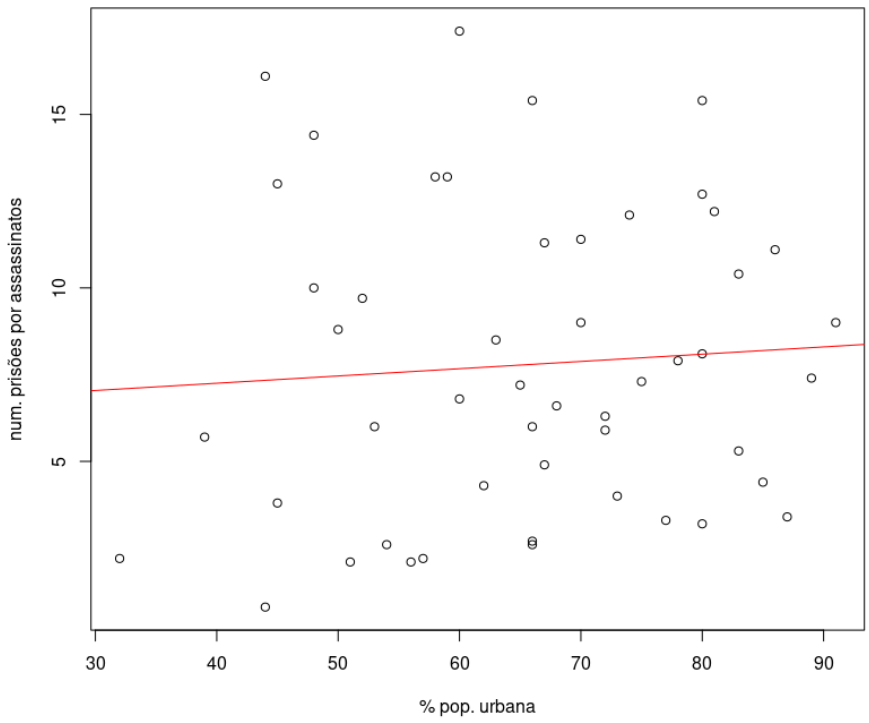

Figura 4.1: Gráfico de disperção

```
# Correlação entre as duas variáveis:
[1] 0.06957262
```

Em seguida, quando fazemos um modelo de regressão linear simples das duas variáveis no software R, obtemos a seguinte saída:

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-6.537 -3.736 -0.779  3.332  9.728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.41594    2.90669   2.207   0.0321 *
x            0.02093    0.04333   0.483   0.6312
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.39 on 48 degrees of freedom
Multiple R-squared:  0.00484,Adjusted R-squared:  -0.01589
F-statistic: 0.2335 on 1 and 48 DF,  p-value: 0.6312
```

Quando construímos um intervalo de confiança com 95%, notamos que β_1 pode ser nulo, implicando que, utilizando este modelo, não teríamos regressão. Além disso, com base no nível descritivo do teste, não rejeitamos a hipótese de que β_1 seja nulo

```
                2.5 %     97.5 %
(Intercept)  0.57164534 12.2602396
x           -0.06617904  0.1080484
```

Iniciando a análise residual, na figura 4.2, temos um gráfico de dispersão do tipo (x_i, e_i), que nos mostra que a suposição de normalidade não está satisfeita, uma vez que parece haver uma tendência dos valores observados se concentrarem na parte direita do gráfico:

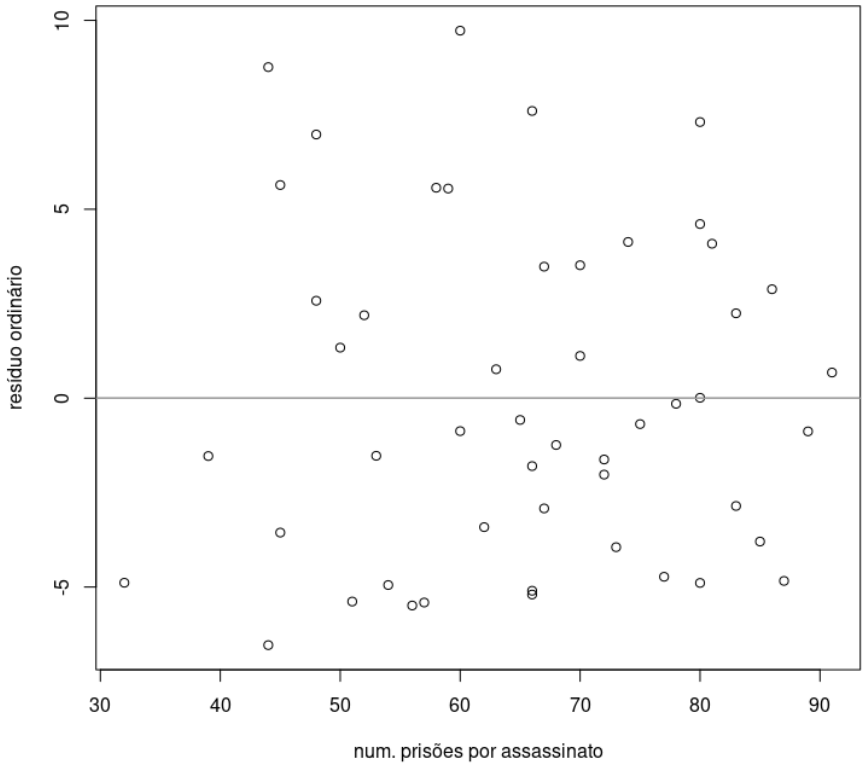

Figura 4.2: Valores Observados Vs Resíduos Ordinários

Ao realizarmos o teste de normalidade *Shapiro-Wilk* nos resíduos ordinários notamos um p-valor de 0,02688. Ou seja, para níveis de confiança maiores que 2,688%, rejeitamos a hipótese de normalidade dos resíduos.

```
Shapiro-Wilk normality test

data:  residuos
W = 0.94747, p-value = 0.02688
```

Na figura 4.3 temos um gráfico de dispersão do tipo $(\hat{y}_i, e_i)$, que nos mostra que a suposição de homoscedasticidade foi violada, já que os pontos apresentam um tendência a se concentrar na parte direita do gráfico.

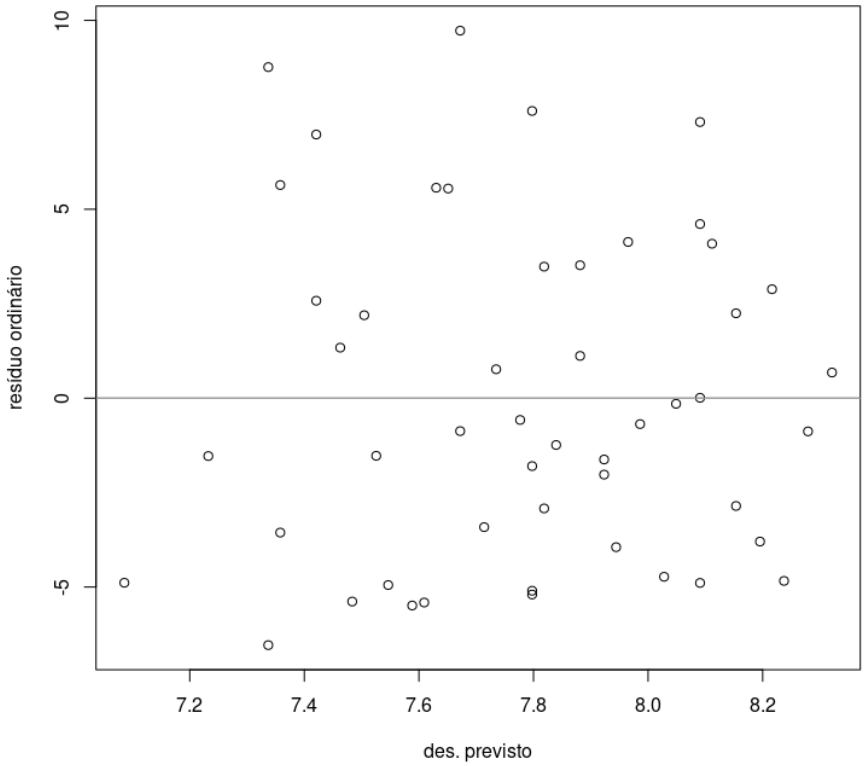

Figura 4.3: Resíduos Ordinários Vs Desvio Previsto

Quando analisamos o gráfico da figura 4.4 notamos que alguns dos pontos estão fora da reta teórica da distribuição normal. Além disso, o gráfico apresenta um formato de **S** característica de que a distribuição dos erros padronizados tem caudas leves.

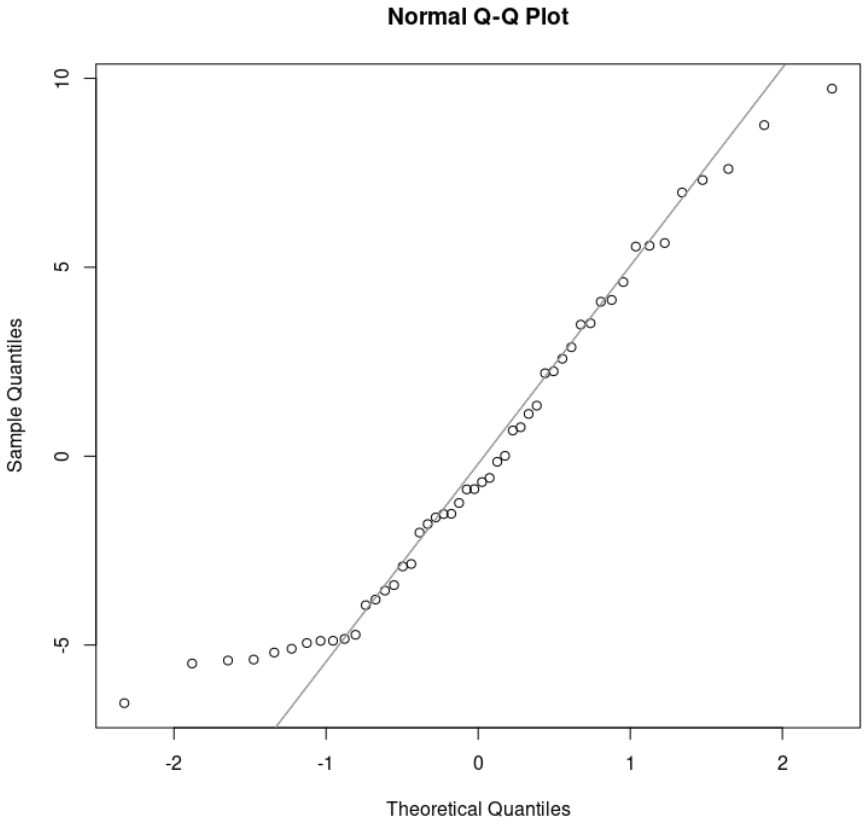

Figura 4.4: QQ Plot dos Resíduos

Quando analisamos a figura 4.5 que leva em consideração os resíduos padronizados notamos que nenhum resíduo tem influência sobre o modelo, já que nenhum dos pontos observados está fora da linha de Cook.

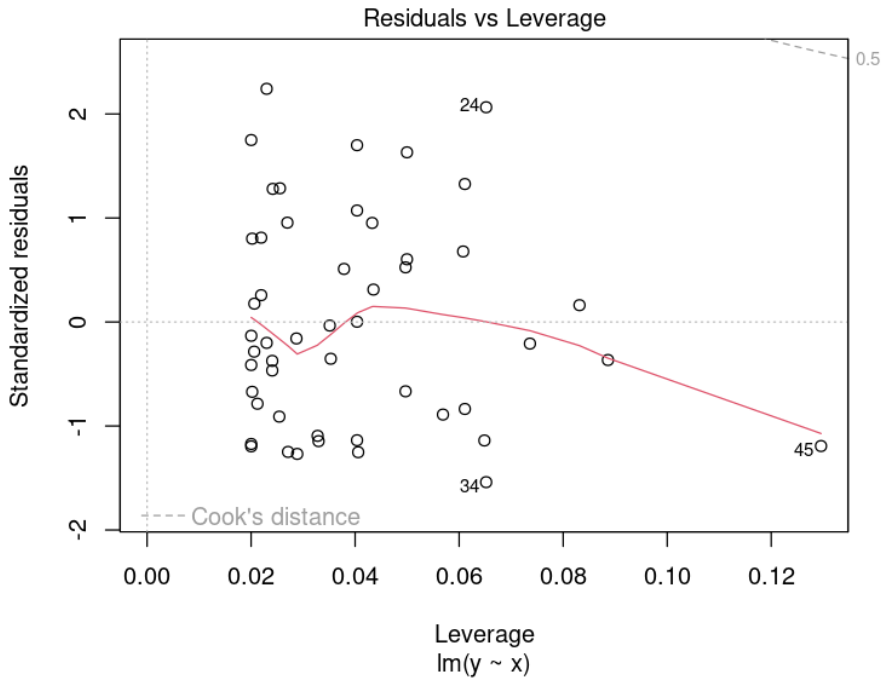

Figura 4.5: Gráfico do tipo *Residuals vs Leverage*.

4.2 Caso 2

Para exemplificar um modelo ajustado com bom encaixe, utilizamos as notas de aula 1 e os códigos em R desenvolvidos em sala de aula, onde se criou um *MRLS* para explicar o desempenho de carros através do seu peso.

O modelo apresentou as seguintes estatísticas, e buscamos por meio da análise de resíduos, saber se há um bom ajuste:

```
Call:
lm(formula = des ~ peso)

Residuals:
   Min     1Q Median     3Q    Max
-2.278 -1.498  0.207  1.488  2.736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.199      2.534   7.970 2.28e-05 ***
peso          -4.964      1.380  -3.597  0.00577 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.866 on 9 degrees of freedom
Multiple R-squared:  0.5898,Adjusted R-squared:  0.5442
F-statistic: 12.94 on 1 and 9 DF,  p-value: 0.005773
```

Assim, baseado nessas estatísticas, rejeitamos as hipóteses de que β_1 e β_0 são nulos.

Agora, medindo a qualidade do ajuste, a figura 4.6, apresenta resíduos que são aleatórios, centrados em torno de 0 e homocedásticos.

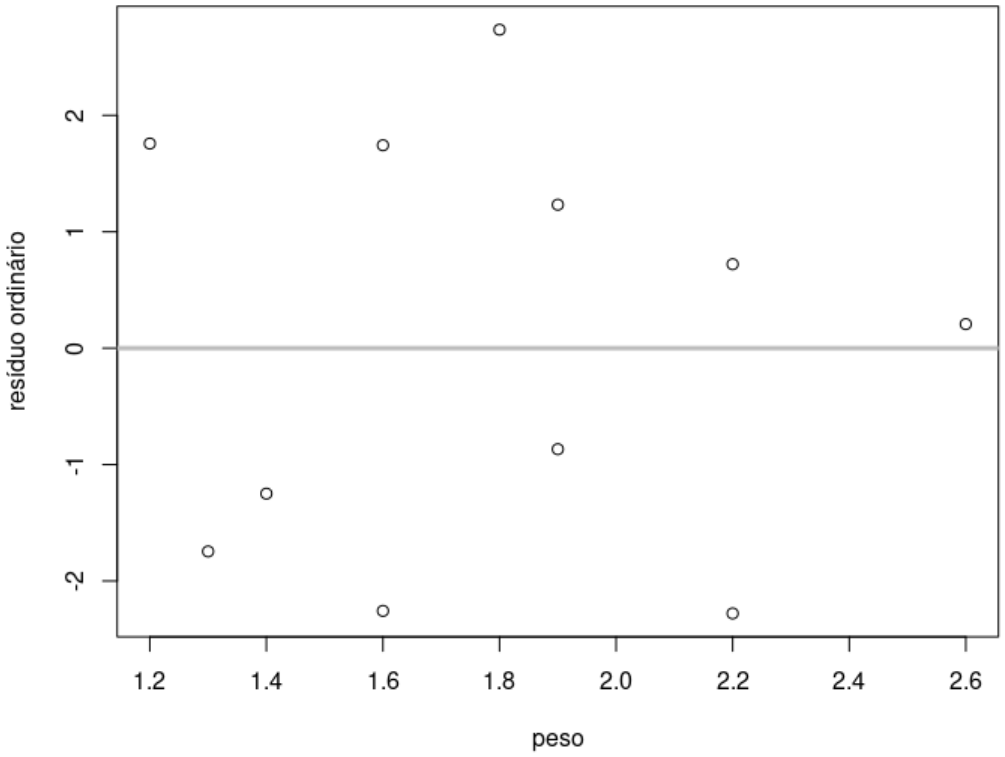

Figura 4.6: Gráfico do tipo (x_i, e_i).

Já a figura 4.7 também revela que os resíduos obedecem às suposições iniciais.

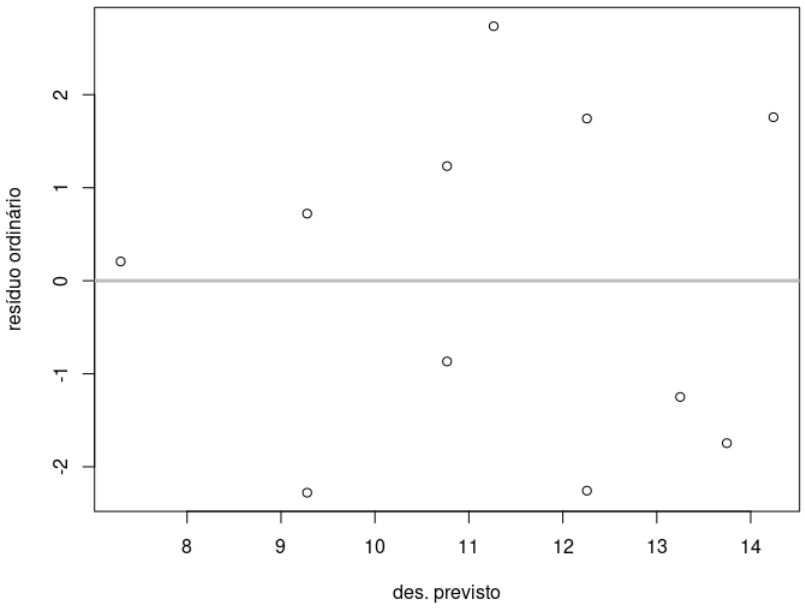

Figura 4.7: Gráfico do tipo $(\hat{y}_i, e_i)$.

Também foi criado um *qqplot* para mostrar quão próximos e/ou distantes os resíduos estão de uma linha reta normal, como mostra a figura 4.8.

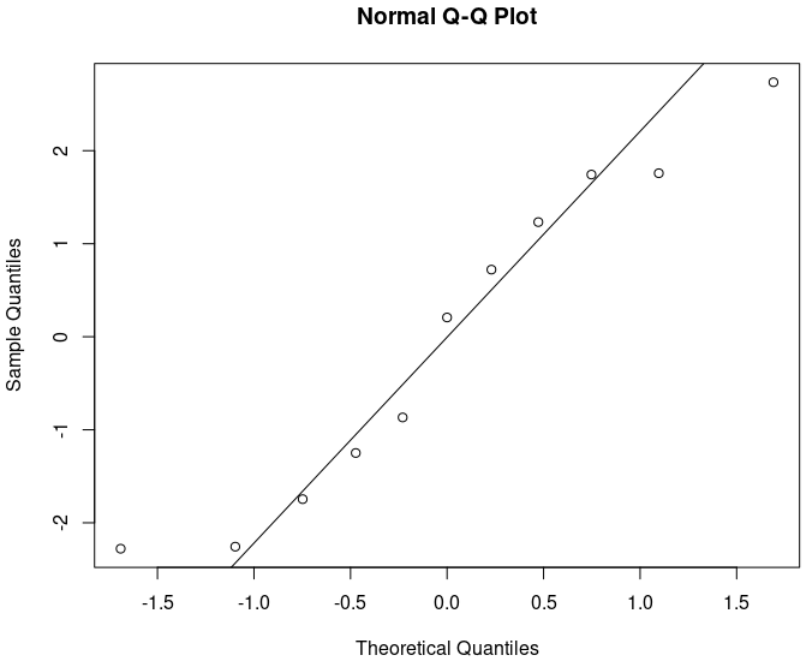

Figura 4.8: Gráfico do tipo QQ dos resíduos.

Dessa forma, podemos ver que os resíduos estão bem próximos à reta, sugerindo que os mesmos têm distribuição normal.

Finalmente, temos que na figura 4.9, nenhum resíduo encontra-se fora da linha de Cook, revelando que nenhum ponto específico tem influência significativa sobre o modelo.

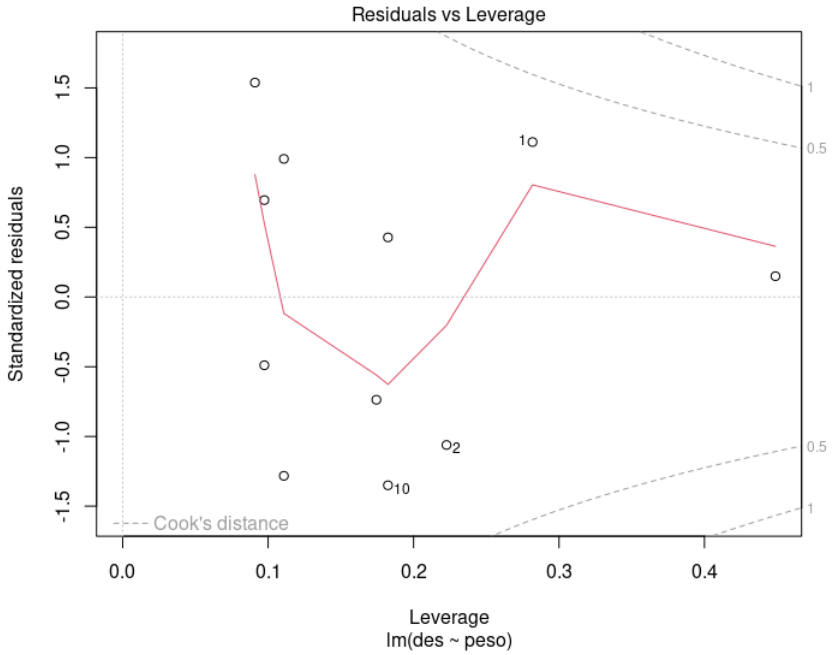

Figura 4.9: Gráfico do tipo *Residuals vs Leverage*.

Ainda, o teste de *Shapiro-Wilk*, aplicado sobre os resíduos, nos leva a não rejeitar a hipótese nula de que os resíduos são normalmente distribuídos.

```
> shapiro.test(reg_mod$residuals)

Shapiro-Wilk normality test

data:  reg_mod$residuals
W = 0.92833, p-value = 0.3942
```

Objetivando não estender o trabalho, não foram apresentados explicitamente os gráficos de dispersão de (x_i, z_i) e (x_i, r_i). Entretanto, os códigos utilizados na geração do mesmos são encontrados nos **Anexos**, sendo fácil replicá-los no software R.

Outro ponto importante a ser salientado a respeito da análise, é que o tamanho da amostra, por se pequeno, pode induzir a interpretações gráficas errôneas, e por esse motivo, faz-se fundamental também utilizar testes de hipótese e outros métodos numéricos.

5 Anexos

5.1 Código: caso 1

```
# Base utilizada para o exemplo: USArrests:

y = USArrests$Murder # variavel resposta
x = USArrests$UrbanPop # variavel explicativa

m = lm(y~x)
summary(m)
plot(m)

plot(x, y, xlab = '% pop. urbana', ylab = 'num. prisões por assassinatos')
abline(m, col = 'red')
summary(m)
residuos = m$residuals
shapiro.test(residuos)
m$fitted.values

cor(x, y)
cor(y, x)
# Residuos vs valores ajustados: (y^_i; e_i)
plot(m$fitted.values, m$residuals, xlab = 'des. previsto',
     ylab = 'resíduo ordinário')
abline(h = 0, col = 'grey')

# Residuos vs Variavel explicativa:
plot(x, m$residuals, xlab = "num. prisões por assassinato",
     ylab = 'resíduo ordinário')
abline(h = 0, col = 'gray')

# Homocedasticidade (variancia dos erros e contante) (variavel homogenea)

## QQ plot
qqnorm(m$residuals)
qqline(m$residuals, col = 'gray')

summary(m)
plot(m)

# Residuos padronizados:
residuos = m$residuals
SQRES = sum((y-m$fitted.values)**2)
QMRES = SQRES/(length(y)-2)

residuos_padronizados = residuos/sqrt(QMRES)
plot(residuos_padronizados, pch = 19,
     ylab = 'Resíduos Padronizados', xlab = 'Valores')
```

```
abline(h = 0, col = 'green4')
shapiro.test(residuos_padronizados)

# Outra forma:
predito = m$fitted.values

cbind(predito, residuos)
plot(predito, residuos, pch = 19,
     main = 'Diagrama de Disperção', xlab = 'Valores Preditos',
     ylab = 'Resíduos')
abline(h=0, col = 'green4', lwd = 2)

hist(residuos, prob = T, main = 'Histograma dos Resíduos Padronizados',
     col = 'green4')
hist(residuos_padronizados, prob = T)

# Residuo Estudentizado:
vii = 1/length(x) + ((x - mean(x))**2)/(sum((p - mx)**2)
residuos_est = residuos/(sqrt(QMRES*(1-vii)))
plot(residuos_est, pch = 19, ylab = 'Resíduos Estudentizados', xlab = 'Valores')

confint(m)
```

5.2 Código: caso 2

```
#### Análise de Resíduos #####

# Peso(p) # variavel explicativa
peso = c(1.2, 1.3, 1.4, 1.6, 1.6, 1.8, 1.9, 1.9, 2.2, 2.2, 2.6)

# Desempenho(d) # variavel resposta
des = c(16, 12, 12, 14, 10, 14, 12, 9.9, 10, 7, 7.5)

# gráficos
plot(peso,reg_mod$residuals, ylab = 'resíduo ordinário')
abline(h=0, col='grey', lwd=3)
plot(reg_mod$fitted.values,reg_mod$residuals, xlab = 'des. previsto',
ylab = 'resíduo ordinário')
abline(h=0, col='grey', lwd=3)
qqnorm(reg_mod$residuals)
qqline(reg_mod$residuals)

# resíduos padronizados e estudentizados
residuos_pdr = residuos/Sdres
residuos_std = residuos/(Sdres*sqrt(1-(1/n + (peso - xb)^2/Sxx)))

plot(peso,residuos_pdr, ylab='resíduo padronizado')
```

```
abline(h=0, col='grey', lwd=3)
plot(peso,residuos_std, ylab='resíduo studentizado')
abline(h=0, col='grey', lwd=3)

# teste de normalidade
shapiro.test(reg_mod$residuals)

# plotando tudo direto
plot(reg_mod)
```

Referências

[1] Manuel António Matos. "Manual operacional para a regressão linear". Em: *Faculdade de Engenharia da Universidade do Porto* 63 (1995).

[2] Douglas C Montgomery, Elizabeth A Peck e G Geoffrey Vining. *Introduction to linear regression analysis*. 4th ed. John Wiley, 2006.

[3] Pedro A Morettin e Wilton O Bussab. *Estatística básica*. Saraiva Educação SA, 2017.

[4] Ronald Targino Nojosa. *Modelo de Regressão I: notas de aula - parte 1*.

[5] Ronald Targino Nojosa. *Modelo de Regressão I: notas de aula - parte 2*.

[6] *Residual Analysis - Scaler Topics*. https://www.scaler.com/topics/data-science/residual-analysis/. (Acessado em 01/09/2023).

[7] J.M. Singer, J.S. Nobre e F.M.M. Rocha. *Análise de Dados Longitudinais (versão parcial preliminar)*. https://www.ime.usp.br/~jmsinger/MAE0610/Singer&Nobre&Rocha2018jun.pdf. (Acessado em 01/09/2023).

[8] *Studentized Residuals*. https://online.stat.psu.edu/stat462/node/247/. (Acessado em 02/09/2023).

[9] *What is a Residuals vs. Leverage Plot? (Definition & Example) - Statology*. https://www.statology.org/residuals-vs-leverage-plot/. (Acessado em 01/09/2023).