Modelos de Regressão Linear (MRL): Técnicas de Diagnóstico e Aplicações

Ben Dêivide

31 de maio de 2016

Modelo de Regressão linear múltipla (MRLM)

Definição 1 (Modelo de regressão linear múltipla)

Seja uma variável aleatória Y *e* X_1, X_2, …, X_p *um conjunto de variáveis fixas, o modelo de regressão linear múltipla é definido por*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad (1)$$

sendo $i = 1, 2, \ldots, n.$

Notação matricial

O modelo em (1) em notação matricial:

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times 1)}{\boldsymbol{X\theta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \qquad p' = p + 1 \tag{2}$$

em que:

Notação matricial

O modelo em (1) em notação matricial:

$$\underset{(n \times 1)}{\boldsymbol{Y}} = \underset{(n \times 1)}{\boldsymbol{X\theta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}} \qquad p' = p + 1 \tag{2}$$

em que:

$$\boldsymbol{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix},$$

Notação matricial

O modelo em (1) em notação matricial:

$$\underset{(n \times 1)}{\boldsymbol{Y}} = \underset{(n \times 1)}{\boldsymbol{X\theta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}} \qquad p' = p + 1 \tag{2}$$

em que:

$$\boldsymbol{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\theta}_{p' \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

Notação matricial

O modelo em (1) em notação matricial:

$$\underset{(n \times 1)}{\boldsymbol{Y}} = \underset{(n \times 1)}{\boldsymbol{X\theta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}} \qquad p' = p + 1 \tag{2}$$

em que:

$$\boldsymbol{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\theta}_{p' \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \text{ e}$$

Notação matricial

O modelo em (1) em notação matricial:

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times 1)}{\boldsymbol{X\theta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \qquad p' = p+1 \tag{2}$$

em que:

$$\boldsymbol{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\theta}_{p'\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \text{ e}$$

$$\boldsymbol{X}_{n\times p'} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}.$$

Pressuposições do MRLM

Para $i = 1, 2, \ldots, n$, então

1. $E[\varepsilon_i] = 0$,

Pressuposições do MRLM

Para $i = 1, 2, \ldots, n$, então

1. $E[\varepsilon_i] = 0$,
2. $Var[\varepsilon_i] = \sigma^2$ (Homocedasticidade)

Pressuposições do MRLM

Para $i = 1, 2, \ldots, n$, então

1. $E[\varepsilon_i] = 0$,
2. $Var[\varepsilon_i] = \sigma^2$ (Homocedasticidade)
3. ε_i são independentes

Pressuposições do MRLM

Para $i = 1, 2, \ldots, n$, então

1. $E[\varepsilon_i] = 0$,
2. $Var[\varepsilon_i] = \sigma^2$ (Homocedasticidade)
3. ε_i são independentes
4. $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ (Normalidade)

Pressuposições do MRLM

Para $i = 1, 2, \ldots, n$, então

1. $E[\varepsilon_i] = 0$,
2. $Var[\varepsilon_i] = \sigma^2$ (Homocedasticidade)
3. ε_i são independentes
4. $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ (Normalidade)
5. De forma matricial: $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{I}\sigma^2)$

Resíduo observado

Definição 2 (Resíduo observado)

Seja o modelo expresso em (2), o resíduo observado é dado por:

$$\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \tag{3}$$

sendo $\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}$ *o estimador de mínimos quadrados dado por* $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. □

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Considerado $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Considerado $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Considerado $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
- $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Considerado $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
- $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$
- Pode-se verificar que

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Considerado $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
- $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$
- Pode-se verificar que
  - $E[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}$

Consequências do resíduo observado

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Considerado $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
- $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$
- Pode-se verificar que
  - $E[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}$
  - $Var[\hat{\boldsymbol{\varepsilon}}] = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$

Implicações do Resíduo observado $\hat{\boldsymbol{\varepsilon}}$

- $\hat{\boldsymbol{\varepsilon}} \sim N_n[\mathbf{0}, (\boldsymbol{I} - \boldsymbol{H})\sigma^2] \Rightarrow \hat{\varepsilon}_i \sim N[0, (1 - h_{ii})\sigma^2]$
- h_{ii} é o elemento da diagonal da matriz $\boldsymbol{H}$ (leverage)
- $0 \leq h_{ii} \leq 1$
- Os $\hat{\varepsilon}_i$ não são independentes
- A variância dos $\hat{\varepsilon}_i$ são heterogêneas

Resíduo Padronizado

Definição 3 (Resíduo Padronizado)

Seja o modelo expresso em (2) *e o resíduo observado* $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, *então o resíduo estudentizado internamente é definido por:*

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{S^2(1 - h_{ii})}}, \quad i = 1, 2, \ldots, n, \tag{4}$$

em que h_{ii} *é o* i*-ésimo elemento da diagonal de* $\boldsymbol{H}$, *e* S^2 *expresso por:*

$$S^2 = \frac{1}{n - p'}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}). \tag{5}$$

Resíduo Estudentizado

Definição 4 (Resíduo Estudentizado)

Seja o modelo expresso em (2) *e o resíduo observado* $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, *então o resíduo estudentizado externamente é definido por:*

$$r_i^* = \frac{\hat{\varepsilon}_i}{\sqrt{S_{(i)}^2(1 - h_{ii})}}, \quad i = 1, 2, \ldots, n, \tag{6}$$

em que h_{ii} *é o* i*-ésimo elemento da diagonal de* $\boldsymbol{H}$, *e* $S_{(i)}^2$ *é o estimador de* σ^2, *dado por:*

$$S_{(i)}^2 = \frac{(n - p')S^2 - \hat{\varepsilon}_i^2/(1 - h_{ii})}{n - p' - 1} = S^2\left(\frac{n - p' - 1}{n - p' - r_i^2}\right). \tag{7}$$

O índice (i) *indica que a* i*-ésimo observação será omitida para estimar* σ^2.

Resíduos recursivos

Definição 5 (Resíduos recursivos)

Considere o modelo expresso em (2), *sendo* $\boldsymbol{y}_r$ *e* $\boldsymbol{x}_r'$ *as* r*-ésimas linhas de* $\boldsymbol{Y}$ *e* $\boldsymbol{X}$, *respectivamente. Considere ainda* $\boldsymbol{X}_r$ *as primeiras* r *linhas de* $\boldsymbol{X}$ *e* $\hat{\boldsymbol{\theta}}_r$ *o estimador de quadrados mínimos usando as primeiras* r *observações. Então, o resíduo recursivo é definido por*

$$w_r = \frac{\boldsymbol{y}_r - \boldsymbol{x}_r'\hat{\boldsymbol{\theta}}_{r-1}}{[1 - \boldsymbol{x}_r'(\boldsymbol{X}_{r-1}'\boldsymbol{X}_{r-1})^{-1}\boldsymbol{x}_r]^{1/2}}, \quad r = p' + 1, \ldots, n. \tag{8}$$

□

Implicações desses três tipos de resíduo

Tabela 1. Quadro resumo das características dos tipos de resíduos.

Pressuposições	$\hat{\varepsilon}$	r_i	r_i^*	w_r
Independência	Não	Não	Não	Sim
Homocedasticidade	Não	Sim	Sim	Sim
Teste exatos sob Normalidade	Não	Não	Não	Sim

OBS.: As observações estão baseadas em Cook (1982) e Rawlings et. al. (1998).

Resumo gráfico no R

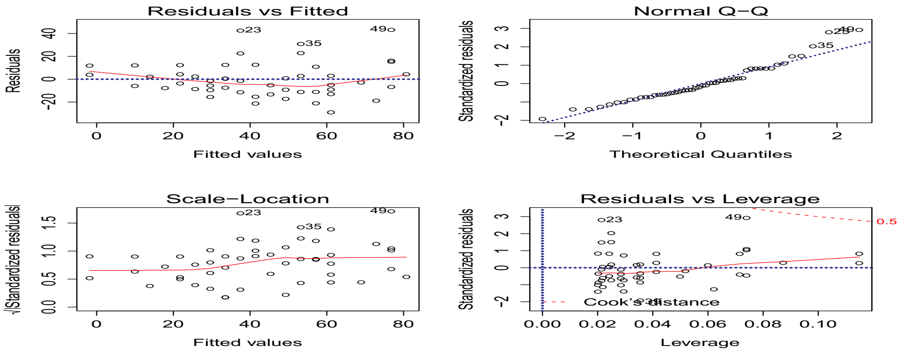

Figura: Tipos de gráficos para avaliar as pressuposições.

Gráfico: Resíduos estudentizados vs valores ajustados

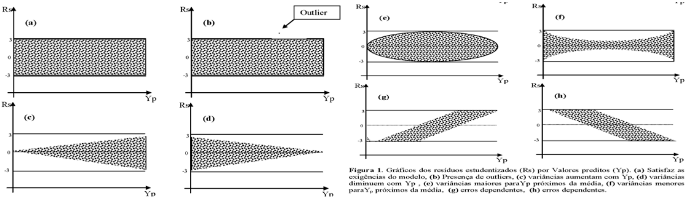

Figura: Gráfico do resíduo estudentizado apresentando padrões de problemas nas pressuposições.

Gráfico: Resíduos estudentizados x Valores ajustados

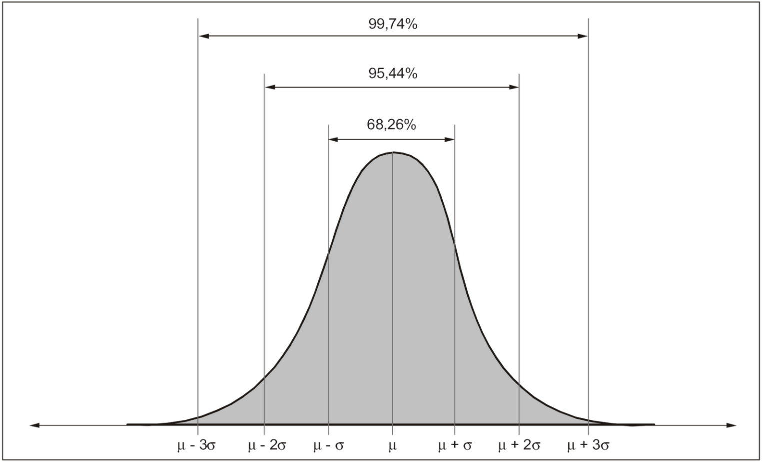

Figura: Gráfico de uma distribuição normal padrão

Normalidade

- Teste de normalidade:

$$
\begin{aligned}
H_0 &: \text{Os } \epsilon' s \text{ têm distribuição normal} \\
H_1 &: \text{Os } \epsilon' s \text{ não seguem uma distribuição normal}
\end{aligned}
$$

  - Teste Shapiro-Wilk

Variância

- Teste de variâncias:

$$
\begin{aligned}
H_0 &: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 \\
H_1 &: \text{pelo menos um dos } \sigma_i^2 \text{ diferente, } i = 1, 2, \ldots, n.
\end{aligned}
$$

  - Teste de Breusch-Pagan

Independência

- Teste de independência:

$$H_0 \quad : \quad \rho = 0$$
$$H_1 \quad : \quad \rho \neq 0$$

  - Teste de Durbin-Watson

Leverage

Definição 6 (Ponto influente)

Uma observação é um ponto influente se a sua exclusão causa uma mudança substancial nos valores ajustados do modelo de regressão.

Leverage

Definição 6 (Ponto influente)

Uma observação é um ponto influente se a sua exclusão causa uma mudança substancial nos valores ajustados do modelo de regressão.

Resíduo observado:

$$\hat{\varepsilon}_i \sim N[0, (1 - h_{ii})\sigma^2]$$

- $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
- h_{ii} é o elemento da diagonal da matriz $\boldsymbol{H}$ (leverage)
- $0 \leq h_{ii} \leq 1$

Distância de Cook

Estatística da distância de Cook:

$$D_i = \frac{r_i^2}{p'}\left(\frac{h_{ii}}{1-h_{ii}}\right),$$

em que:

- r_i é o resíduo padronizado
- p' o número de parâmetros do modelo
- h_{ii} é o leverage

Resíduo Estudentizado

Rawlings et. al. (1998) mostra que

$$r_i^* = \frac{\hat{\varepsilon}_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}} \sim t_{n-p'-1} \tag{9}$$

Teste para verificar um ponto influente é:

- Se $|r_i^*| > t_{n-p'-1;\alpha/2}$, então a i-ésima observação é influente com uma significância α.

Aplicação

Problema: Buscando conhecer o rendimento dos automóveis, o estudo baseou-se em determinar uma relação entre a velocidade (mph) de um automóvel e a distância (pés) percorrida até a parada após um sinal. Foram coletadas 50 observações ao longo do tempo. Esse banco de dados foi retirado do pacote `datasets` do R.

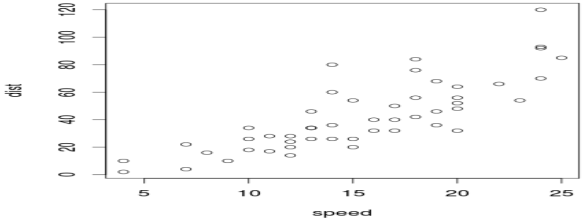

Figura: Gráfico de dispersão.

Dados

```
> # DESCRICAO:   Os dados sao da velocidade dos
> #              carros e a distancia até a
> #              parada apos um sinal
> dados <- cars; cars
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
.      .    .
.      .    .
50    25   85
```

Modelo

```
> # Analise de Regressao linear simples:
> reg <- lm(rev(cars))
> reg <- lm(dist ~ speed, data = cars)
> summary(reg)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
```

Residuos

```
rob <- residuals(reg) # residuos observado
rsi <- rstandard(reg) # res padronizado
rse <- rstudent(reg)  # res estudentizado
rr  <- recresid(reg)  # residuo recursivo
```

Teste de Normalidade (Shapiro-Wilk)

```
> shapiro.test(rob) # Residuo observado
W = 0.94509, p-value = 0.02152

> shapiro.test(rsi) # Residuo Estudentizado
W = 0.94518, p-value = 0.0217

> shapiro.test(rse) # Residuo Padronizado
W = 0.93418, p-value = 0.00798

> shapiro.test(rr)  # Residuo recursivo
W = 0.95962, p-value = 0.09745
```

Teste de Homocedasticidade (Breusch-Pagan)

```
> library(car)
> ncvTest(reg) # Residuo observado
Chisquare = 4.650233    Df = 1     p = 0.03104933

> library(lmtest)
> bptest(reg) # Residuo Estudentizado
studentized Breusch-Pagan test
BP = 3.2149, df = 1, p-value = 0.07297
```

Teste de Independência (Durbin-Watson)

```
> #Independencia dos residuos
> library(car)
> durbinWatsonTest(reg)
lag Autocorrelation D-W Statistic p-value
1       0.1604322      1.676225   0.212
Alternative hypothesis: rho != 0
```

Análise Gráfica

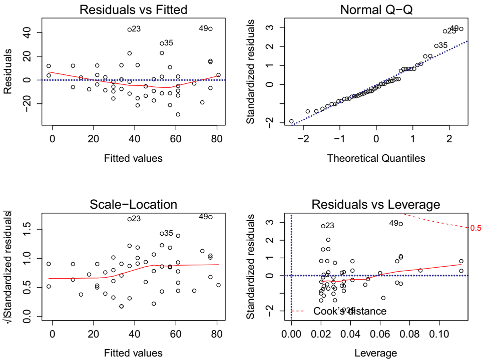

Figura: Análise Gráfica dos dados velocidade e distância.

Ponto influente (Resíduo Estudentizado)