Atividade 6

Análise de Resíduos e identificando observações atípicas

Paulo Ricardo Seganfredo Campana

23 de setembro de 2023

Problema:

Um hospital está implementando um programa para melhorar a qualidade do serviço e a produtividade. Como parte deste programa, o gerenciamento hospitalar está tentando medir e avaliar a satisfação do paciente. Os dados que foram obtidos de uma amostra aleatória de 25 pacientes recém-dispensados. A variável resposta é *satisfaction*, uma medida subjetiva da satisfação do paciente, em escala crescente. Enquanto as potenciais variáveis regressoras são *age*, a idade do paciente, e *severity*, um índice que mede a gravidade da doença do paciente. Estes dados são do livro do *Montgomery* podem ser encontrados no pacote *MPV* do **R** como o nome *table.b17*.

Depois de algumas análises, concluímos que a melhor equação de regressão parece ser:

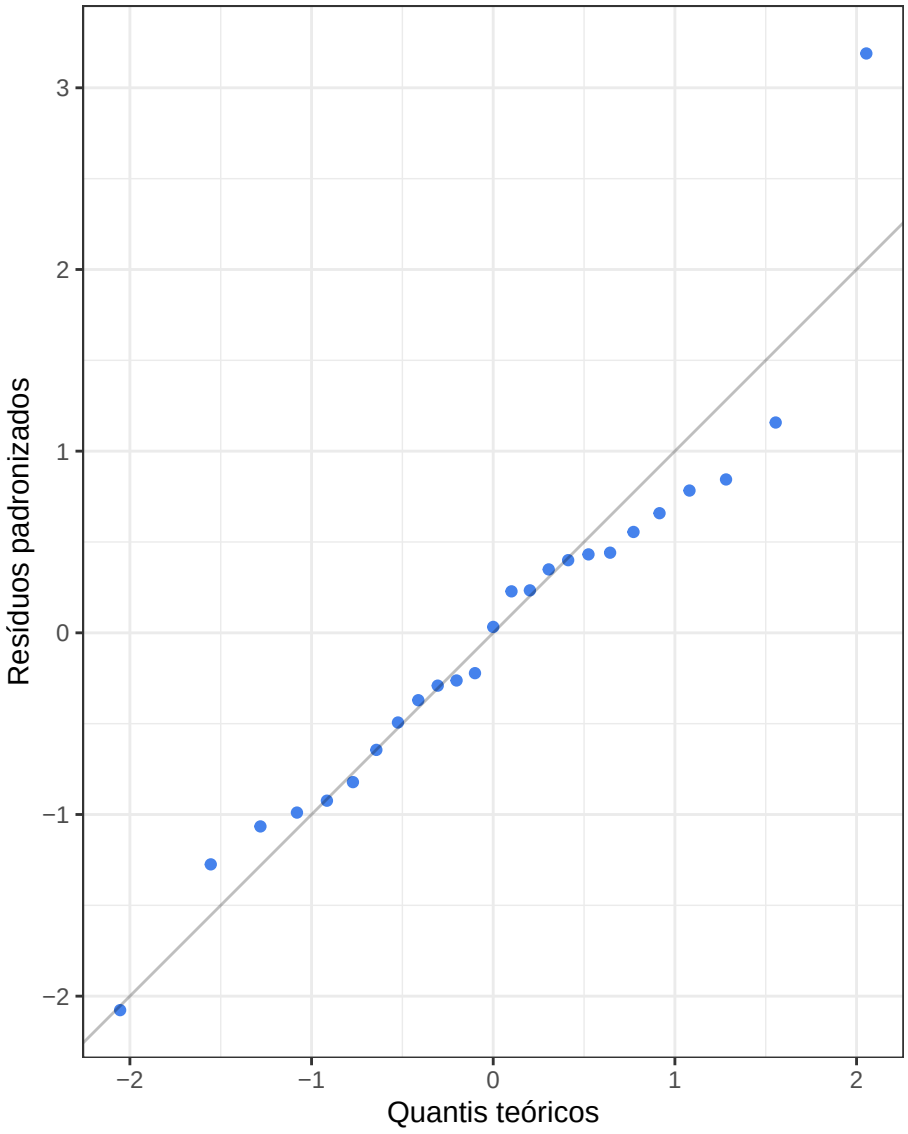
$$\text{satisfação} = \beta_0 + \beta_1 \times \text{idade} + \beta_2 \times \text{gravidade},$$

ou seja, a satisfação do paciente sendo explicada por sua idade e pela gravidade da sua doença. Tomando esta equação de regressão e considerando 5% de significância, responda as questões abaixo:

```
library(tidyverse)
data <- MPV::table.b17[1:3]
fit1 <- lm(Satisfaction ~ Age + Severity, data)
data_resid <- data |>
    mutate(
        residuals = residuals(fit1),
        rstandard = rstandard(fit1),
        rstudent = rstudent(fit1),
        predict = predict(fit1),
        hatvalues = hatvalues(fit1)
    )
```

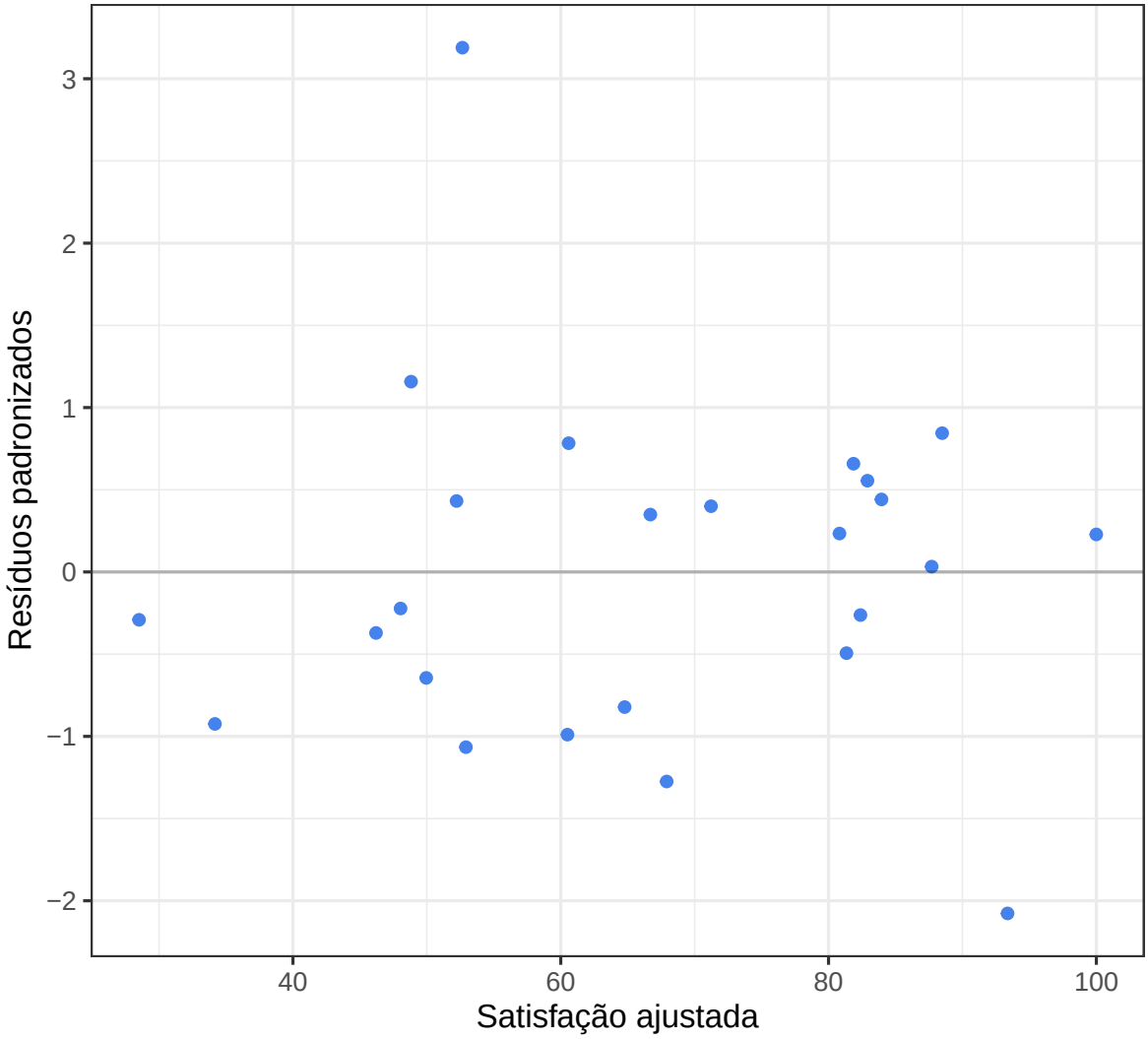
1. Realize a análise residual para o modelo ajustado. Comente TODOS os resultados.

   a) Construa um gráfico para verificar a suposição de normalidade. Parece haver algum problema com esta suposição?



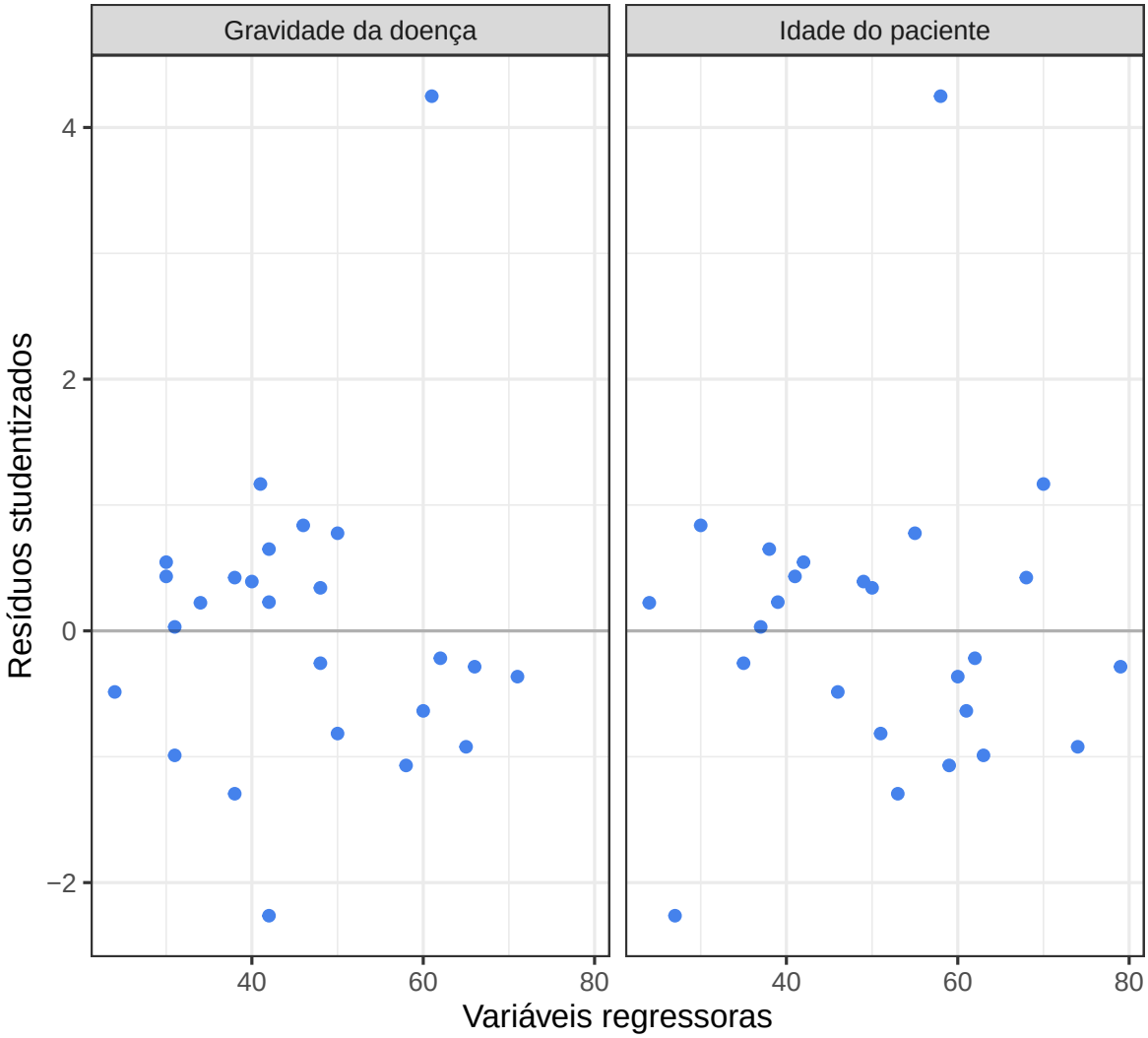
Não, os valores dos resíduos padronizados parecem seguir distribuição normal porém com alguns desvios não aleatórios nas caudas da distribuição e um ponto fora do padrão, distante a mais de 3 desvios padrões, o que não é esperado para apenas 25 obervações.

b) Construa um gráfico dos resíduos padronizados versus a resposta prevista. É possível identificar algum problema de adequação no modelo ao analisar este gráfico?



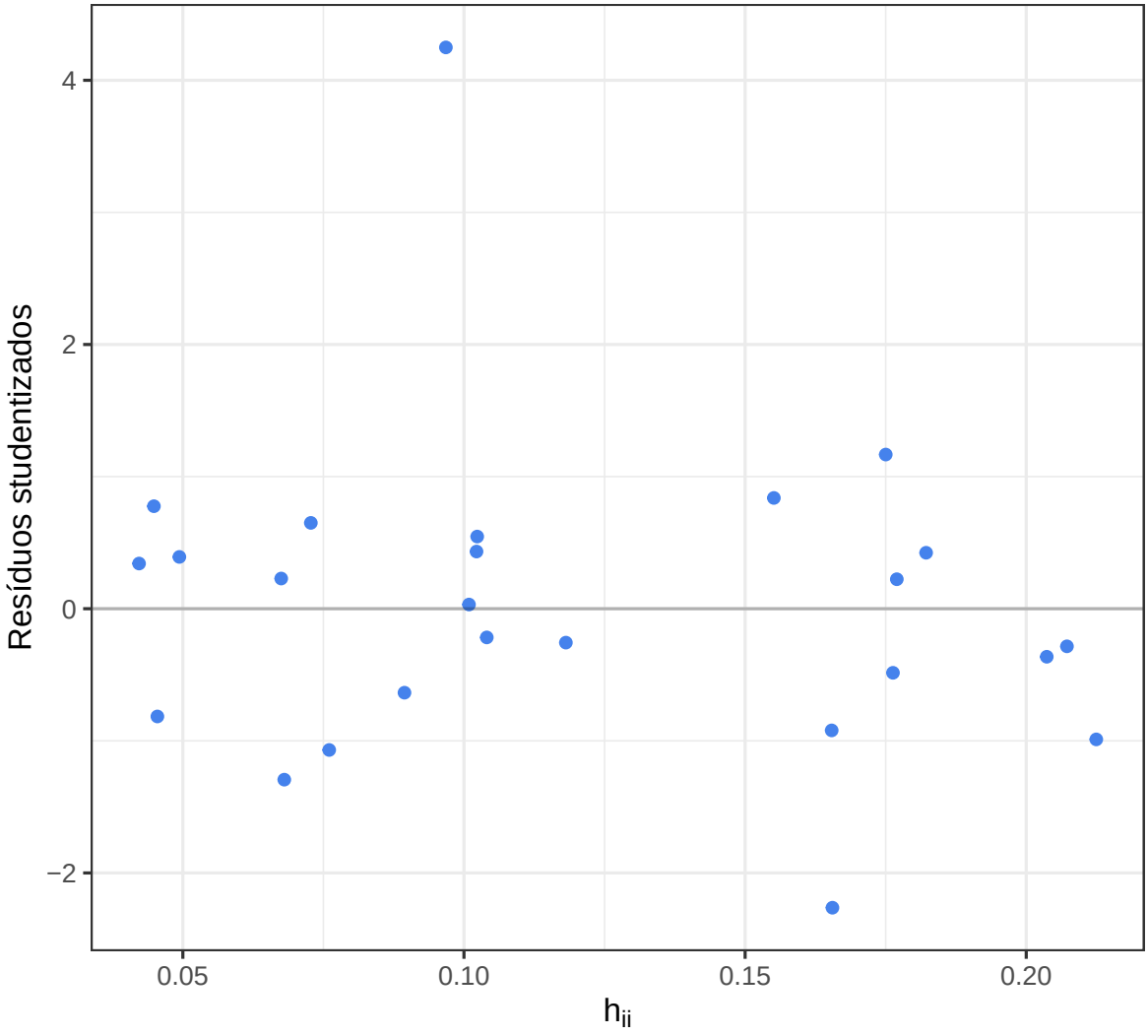
Com exceção do ponto mencionado acima, não podemos notar nenhum comportamento não aleatório na distribuição dos resíduos padronizados, sua variância parece ser constante para diferentes valores da satisfação.

c) Construa um gráfico dos resíduos studentizados versus os valores das variáveis regressoras. Faça uma analise destes gráficos.



Utilizando agora as variáveis regressoras ao invés da variável resposta, obtemos um gráfico similar onde não é possível detectar a presença de variância não constante dos erros, porém agora com os resíduos studentizados, os pontos com resíduos altos são mais extremos devido a distribuição t apresentar caudas mais densas.

d) Construa um gráfico dos resíduos versus os valores de h_{ii}. O que é possível verificar ao analisar este gráfico?



Podemos ver que as observações com resíduos altos não possuem h_{ii} diferente de outras observações com resíduos normais, isso significa que os mesmos estão dentro da região conjunta dos dados.

2. Realize uma análise de diagnóstico para o modelo final. Comente os resutados.

e) No gráfico obtido na letra **b)** foi possível identificar algum possível ***outlier***?

Sim, houve um ponto em que seu resíduo padronizado foi acima de 3, indicando um possível outlier.

f) No gráfico obtido na letra **d)** foi possível identificar algum possível **ponto de alavanca**?

Não, não há nenhuma observação nos dados com h_{ii} muito elevado comparado com os demais, significando que não há pontos muito distantes da região conjunta dos dados para serem considerados pontos de alavanca.

```
data_detect <- data |>
    mutate(
        cooks = cooks.distance(fit1),
        dffits = dffits(fit1),
        dfbetas_int = dfbetas(fit1)[ ,1],
        dfbetas_age = dfbetas(fit1)[ ,2],
        dfbetas_sev = dfbetas(fit1)[ ,3]
    )
```

g) Através da análise da distância de Cook, é possível identificar alguma possível **observação influente**?

```
data_detect |>
    filter(cooks > 1)
```

Satisfaction	Age	Severity	cooks	dffits	dfbetas_int	dfbetas_age	dfbetas_sev

Não, nenhuma observação tem distância de Cook maior que 1.

h) Faça agora detecção de pontos de influência utilizando as medidas ***DFFITS*** e ***DFBETAS***.

```
data_detect |>
    filter(dffits > 2 * sqrt(ncol(data) / nrow(data)))
```

Satisfaction	Age	Severity	cooks	dffits	dfbetas_int	dfbetas_age	dfbetas_sev
82	58	61	0.363	1.39	-0.632	-0.139	0.97

```r
data_detect |>
    filter(
        abs(dfbetas_int) > 2 / sqrt(nrow(data)) |
        abs(dfbetas_age) > 2 / sqrt(nrow(data)) |
        abs(dfbetas_sev) > 2 / sqrt(nrow(data))
    )
```

Satisfaction	Age	Severity	cooks	dffits	dfbetas_int	dfbetas_age	dfbetas_sev
75	27	42	0.285	-1.008	-0.639	0.864	-0.328
82	58	61	0.363	1.391	-0.632	-0.139	0.970
59	70	41	0.095	0.538	-0.062	0.462	-0.328
52	63	31	0.088	-0.514	-0.105	-0.383	0.423

Com DFFITS, foi identificado um ponto de influência, aquele que possui resíduo altíssimo, além desse, DFBETAS identificou outros 3 pontos que exercem influência muito grande nos valores dos coeficientes do modelo.

3. Caso você tenha detectado a presença de observações atípicas, tais como ***Outliers***, **pontos de alavanca** e de **influência**, cheque o efeito dessas obervações nos principais resultados do ajuste. Para isso, ajuste um novo modelo excluindo as observações da base de dados e compare os resultados obtidos com àqueles obtidos com o uso da base completa. Comente os resultados!

Seja o primeiro modelo aquele feito com todas as observações, o segundo modelo retirando a observação flagrada pela medida DFFITS e o terceiro modelo sem as 4 observações identificadas por DFBETAS.

```r
fit2 <- lm(
    Satisfaction ~ Age + Severity,
    data_detect,
    subset = dffits < 2 * sqrt(ncol(data) / nrow(data))
)
fit3 <- lm(
    Satisfaction ~ Age + Severity,
    data_detect,
    subset =
        abs(dfbetas_int) < 2 / sqrt(nrow(data)) &
        abs(dfbetas_age) < 2 / sqrt(nrow(data)) &
        abs(dfbetas_sev) < 2 / sqrt(nrow(data))
)
```

Tabela 4: Variáveis dos modelos de regressão

Variáveis	Estimativa	Erro padrão	Estatística t	p-valor
Primeiro modelo				
(Intercepto)	139.923	8.100	17.27	2.78×10^{-14}
Idade	-1.046	0.157	-6.65	1.09×10^{-6}
Gravidade	-0.436	0.179	-2.44	0.0233
Segundo modelo				
(Intercepto)	143.766	6.147	23.39	1.60×10^{-16}
Idade	-1.030	0.118	-8.72	2.01×10^{-8}
Gravidade	-0.566	0.138	-4.11	0.000496
Terceiro modelo				
(Intercepto)	149.401	4.812	31.05	4.37×10^{-17}
Idade	-1.190	0.113	-10.52	4.06×10^{-9}
Gravidade	-0.498	0.121	-4.11	0.000659

Sobre os coeficientes da regressão, há algumas pequenas diferenças em seus valores entre os modelos, podemos notar que o erro padrão e o p-valor das estimativas dos coeficientes diminui ao retirar aquelas observações atípicas.

Tabela 7: Métricas de performance dos modelos de regressão

Modelo	MSE	R^2	R^2_a	Estatística F	p-valor
Primeiro modelo	9.682	0.809	0.792	46.77	1.19×10^{-08}
Segundo modelo	7.267	0.895	0.885	89.75	5.14×10^{-11}
Terceiro modelo	5.459	0.948	0.942	163	2.93×10^{-12}

Com essa retirada, também observamos que o erro padrão dos resíduos diminui e o coeficiente de determinação aumenta, indicando um ajuste melhor, o teste F trás a informação de que estes modelos são mais significativos.

Porém, estas estatísticas foram calculadas sem as observações retiradas e não com o conjunto de dados original, se estas observações foram de fato comuns e não atípicas, o segundo e terceiro modelo apresentarão resultados piores na predição de tais valores.