



Regressão Linear

Machine Learning

Prof. Neylson Crepalde

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Pense nos dados

Advertisement

Algumas perguntas importantes:

1. Existe uma relação entre o investimento em propaganda e vendas?
2. Quão forte é a relação entre o investimento em propaganda e vendas?
3. Qual *media* contribui para as vendas?
4. Quão precisa é a estimação do efeito de cada *media* sobre as vendas?
5. Quão precisamente podemos prever novas vendas?
6. A relação é linear?
7. Existe sinergia entre as diversas *media* de propaganda?

Na regressão linear, podemos prever o valor de uma variável quantitativa Y (dependente) a partir de um preditor X . Matematicamente, podemos representar essa relação da seguinte maneira:

$$Y \approx \beta_0 + \beta_1 X.$$

onde \approx representa "é aproximadamente modelado como". β_0 e β_1 são constantes do modelo que representam, respectivamente o *intercepto* e o *slope* (inclinação). Esses coeficientes (ou parâmetros, no caso da população, estimadores, no caso da amostra) são desconhecidos e serão estimados. A partir deles, podemos prever novas vendas.

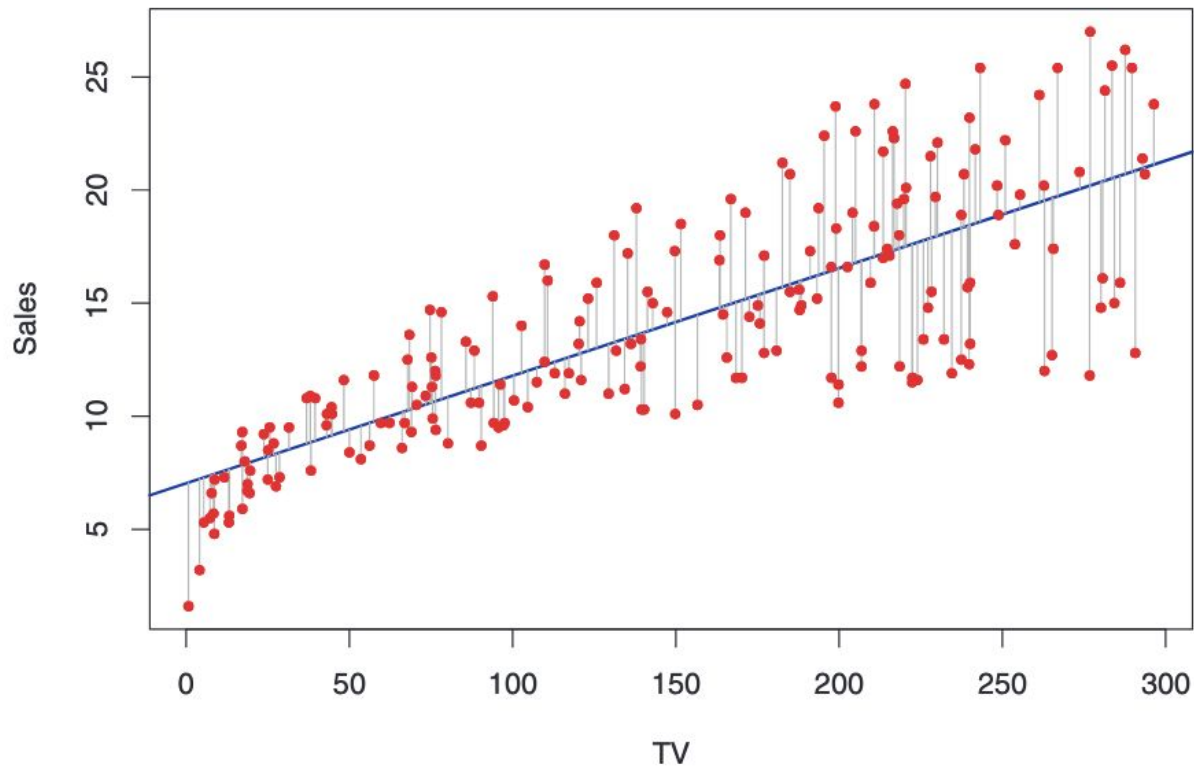


FIGURE 3.1. For the **Advertising** data, the least squares fit for the regression of **sales** onto **TV** is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

O método dos mínimos quadrados ordinários visa minimizar a soma dos quadrados dos erros (*Residual Sum of Squares*).

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

A partir desse método de estimação, os coeficientes podem ser estimados por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Acurácia do modelo

Para mensurar o quanto nossas estimativas estão próximas ou não dos parâmetros populacionais, usamos o Erro Padrão (*Standard Error*).

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Os Erros Padrão podem ser usados para estimar os intervalos de confiança da seguinte maneira:

$$\left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1) \right]$$

Testes de hipótese

Os Erros Padrão podem ser usados para computar os testes de hipótese. Testamos a hipótese nula de que não há relação entre X e Y e a hipótese alternativa de que há relação entre X e Y. O teste de Hipótese é expresso da seguinte maneira:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0,$$

Para testar a hipótese nula, verificamos se o estimador é suficientemente distante de zero de modo que temos confiança de que ele é de fato diferente de zero. Para isso, computamos a estatística t :

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

que mede o número de desvios padrão que o estimador Beta1 está afastado de 0. Consequentemente, podemos computar a probabilidade de observar um número igual a t ou maior em valores absolutos assumindo Beta1 = 0. Chamamos essa probabilidade de *p-value*. De modo geral, interpretamos o *p-value* da seguinte maneira:

Um *p-value* pequeno indica uma baixa probabilidade de observar uma associação tão substancial entre o preditor e a dependente por acaso, na ausência de uma relação real entre X e Y. Desse modo, quando obtemos um *p-value* pequeno (< 0.05), rejeitamos a hipótese nula e inferimos que há associação entre X e Y. Do contrário, não podemos rejeitar a hipótese nula e consideramos que não há relação entre X e Y.

Lab: Regressão Simples com *Advertisement Data*

Acurácia do modelo

O RSE (*Residual Standard Error*) - Erro Padrão do Resíduo - é uma estimativa do desvio padrão do erro. Ela pode ser definida por:

$$\text{RSE} = \sqrt{\frac{1}{n-2} \text{RSS}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

RSE é considerado uma medida de *falta de ajuste* do modelo. Quanto menor o RSE melhor o modelo se ajusta aos dados.

R² representa a proporção da variância explicada. É uma medida de quanto da variância de Y o modelo explica. R² pode ser calculado assim:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

onde:

$$TSS = \sum (y_i - \bar{y})^2 \quad \& \quad RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$


Regressão Linear Múltipla

Na prática, não é comum utilizarmos a regressão simples pois é conhecido que qualquer fenômeno que possamos escolher para investigação pode ter múltiplas correlações. Em face ao argumento de que poderíamos estimar várias regressões simples para testar o efeito de várias variáveis, podemos afirmar que essa abordagem não é satisfatória pois nenhum efeito estimado leva em conta as outras variáveis mensuradas. Por esse motivo, a estimação de correlações simples entre variáveis pode levar a um viés grande.

A regressão múltipla possui a seguinte equação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

onde os vários X representam as diversas variáveis que vamos inserir no modelo. Cada preditor vai possuir o seu *slope*, ou seja, seu efeito sobre Y levando em conta, controlando, pelos demais preditores (*Ceteris Paribus*).



Lab: Regressão Simples e Múltipla com *Advertisement Data*

Algumas questões importantes

1. Pelo menos um dos preditores é útil para prever a variável dependente?
2. Todos os preditores ajudam a explicar Y ou apenas um subconjunto desses preditores?
3. Quão bem o modelo se ajusta aos dados?
4. Dado um set de preditores, qual resposta nós daríamos e quão acurada seria essa resposta?

1

Consideremos a hipótese nula

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

com relação à alternativa:

$$H_a : \text{at least one } \beta_j \text{ is non-zero.}$$

Podemos testá-la calculando a estatística de teste F que mede, em linhas gerais, quantas vezes o resultado do modelo é melhor que a média:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

2 Decidindo sobre variáveis importantes

Parte do nosso curso será destinado exclusivamente a este problema. Idealmente nós tentaríamos todas as combinações de variáveis e verificaríamos qual dos modelos possui o melhor ajuste através de qualquer medida escolhida (AIC, BIC, R2, RSS). Entretanto, existe um total de 2^p modelos. Se há apenas, 2 variáveis, ótimo. $2^2 = 4$ modelos. Entretanto, se há no banco 30 variáveis, temos um total de $2^{30} = 1.073.741.824$ modelos. IMPOSSÍVEL!!

Três abordagens são mais conhecidas:

1. *Forward Selection*: Começar com um modelo vazio e adicionar variáveis gradualmente testando o ajuste a cada adição;
2. *Backward Selection*: Começar com todas as variáveis e retirar gradualmente enquanto testa o ajuste;
3. *Mixed Selection*: Uma mistura de *forward* e *backward selection*.

3 Ajuste

Observar:

- R^2
- RSE

4 Predições

Para prever a variável Y é necessário apenas realizar a equação de regressão uma vez que os estimadores Betas são conhecidos. Entretanto, há um erro associado aos estimadores. Portanto, para uma melhor acurácia dos resultados, é possível computar um intervalo de confiança. Normalmente utilizamos 95% de confiança.

É importante lembrar que o modelo linear assume uma relação linear entre as variáveis o que por vezes pode não ser verdade. Isso introduz um viés no modelo. Estudaremos posteriormente maneiras de corrigir esse viés. Por ora, assumiremos que a equação linear está correta.

Mesmo que os parâmetros reais fossem conhecidos, seria impossível uma predição completamente acertada tendo em vista o componente aleatório do modelo. Para saber qual é o tamanho da variação que Y predito pode ter, podemos computar os intervalos de predição.

Preditores Qualitativos

Ao inserir uma variável **qualitativa binária** no modelo (sexo, por exemplo), podemos transformá-la em uma variável que assume os valores de 0 e 1 da seguinte maneira:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

e utilizá-la desse modo de modo que o modelo pode ser reescrito da seguinte maneira:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Agora, β_0 pode ser interpretado como a média da variável Y para homens, $\beta_0 + \beta_1$ como a média da variável Y para mulheres e β_1 como a diferença média entre homens e mulheres.

Quando trabalhamos com variáveis qualitativas com **mais de duas categorias**, adotamos o mesmo procedimento: transformamos cada categoria da variável como uma binária (0,1) deixando uma delas de fora para servir como categoria de referência. Pensando na variável **raça**, definimos

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases} \quad (3.28)$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases} \quad (3.29)$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases} \quad (3.30)$$

Desse modo, podemos interpretar b_0 como a média de Y para os negros (*African American*), b_1 como a diferença média de Y entre Asiáticos e Negros e b_2 como a diferença média de Y entre Brancos e Negros.

Exercício!!

Termos interativos

Os termos interativos são uma maneira de identificar se existe sinergia entre os preditores, ou seja, se os valores de um preditor interferem no efeito de outro preditor sobre Y . Os termos interativos são estimados multiplicando preditores de modo que a nova equação de regressão possua a seguinte forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Se voltarmos no modelo dos dados *Advertisement*, podemos testar a interação entre o investimento em TV e rádio da seguinte forma:

$$\begin{aligned} \text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon. \end{aligned}$$

Os termos interativos também podem ser testados com variáveis qualitativas ou combinações de uma variável qualitativa e uma quantitativa. Voltemos ao banco de dados *credit*. Podemos estimar a diferença do efeito da variável *Income* para estudantes e não estudantes utilizando um termo interativo:

$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if not student} \end{cases} \end{aligned}$$

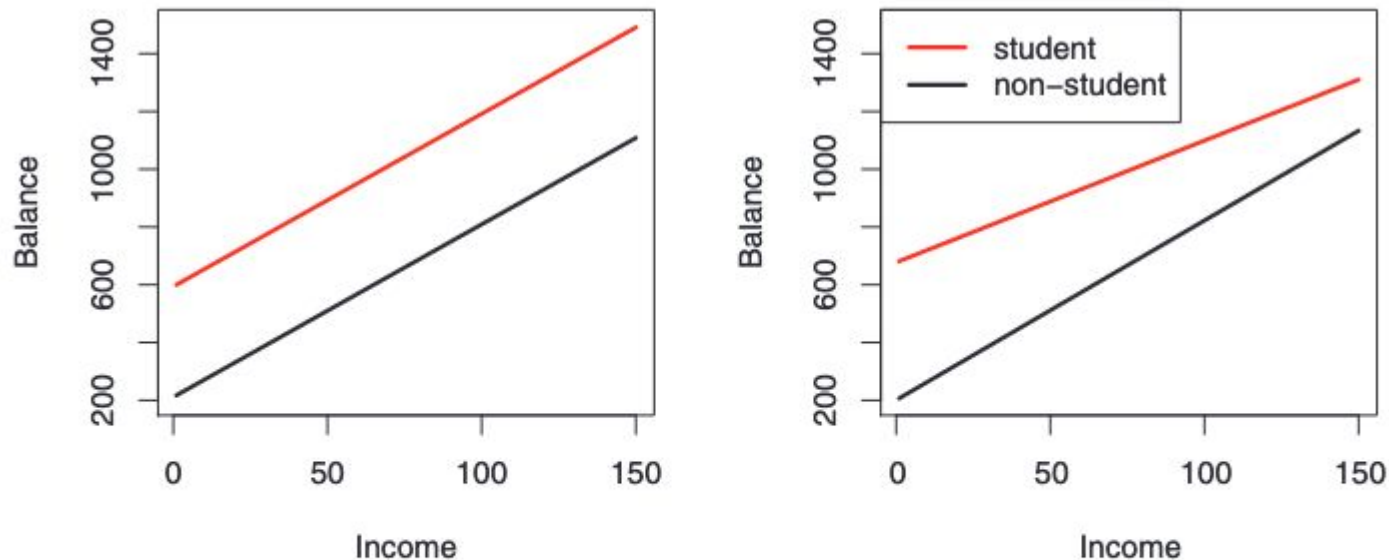


FIGURE 3.7. For the **Credit** data, the least squares lines are shown for prediction of **balance** from **income** for students and non-students. Left: The model (3.34) was fit. There is no interaction between **income** and **student**. Right: The model (3.35) was fit. There is an interaction term between **income** and **student**.

Problemas com o modelo de regressão

Alguns problemas podem surgir quando trabalhamos com a regressão linear. Os mais comuns deles são:

1. Não linearidade entre a variável resposta e os preditores;
2. Correlação dos termos do erro;
3. Variância do erro não constante;
4. Outliers;
5. Pontos de High Leverage;
6. Colinearidade

1. Não linearidade

Os gráficos de resíduos são úteis para verificar quando a relação entre a variável resposta e os preditores não é linear.

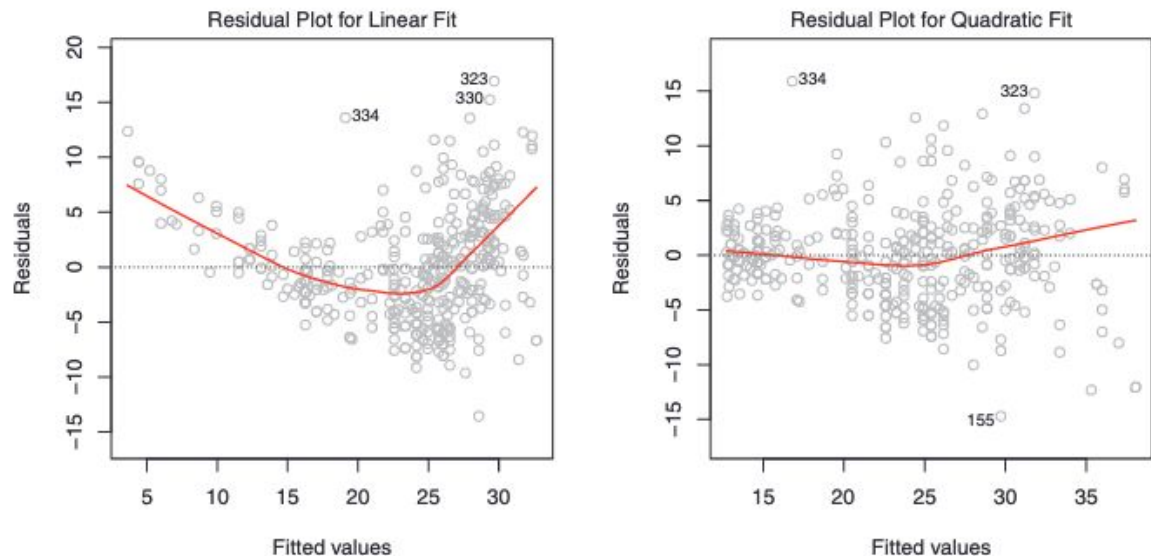


FIGURE 3.9. Plots of residuals versus predicted (or fitted) values for the **Auto** data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of **mpg** on **horsepower**. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of **mpg** on **horsepower** and **horsepower**². There is little pattern in the residuals.



2. Correlação dos termos do erro

A regressão linear está ancorada no pressuposto de que os termos do erro são independentes (independência das observações).

Termos de erro correlacionados (como no caso de séries temporais) podem causar uma subestimação do erro padrão e, conseqüentemente, subestimação dos p-values e, conseqüentemente, subestimação do intervalo de confiança!

3. Variância do erro não constante

A variância do erro pode também ser verificada a partir dos gráficos de resíduo.

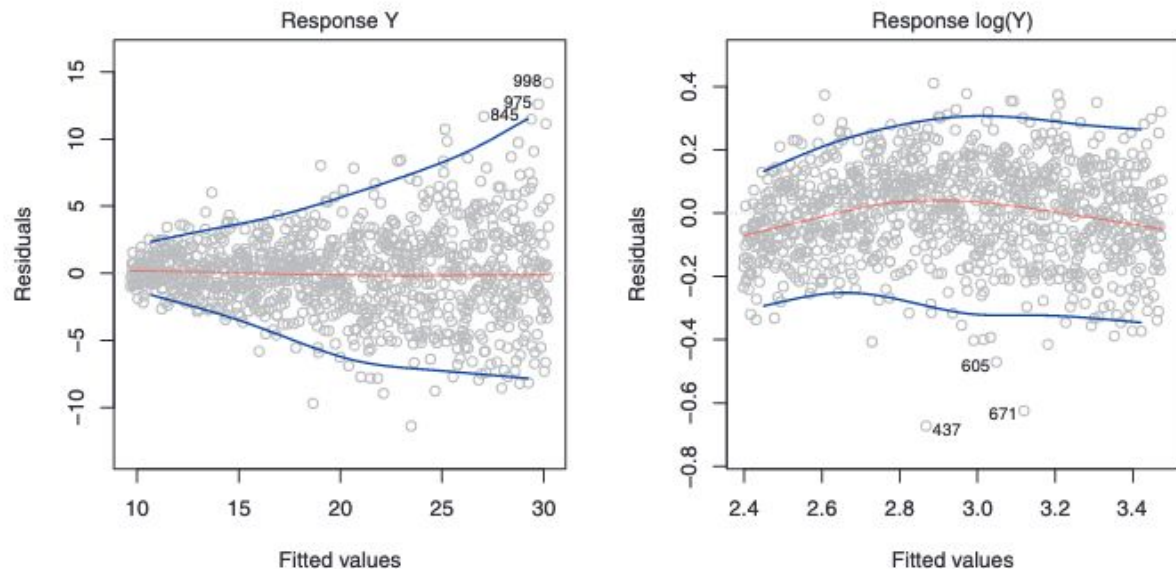


FIGURE 3.11. Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The response has been log transformed, and there is now no evidence of heteroscedasticity.

4. Outliers

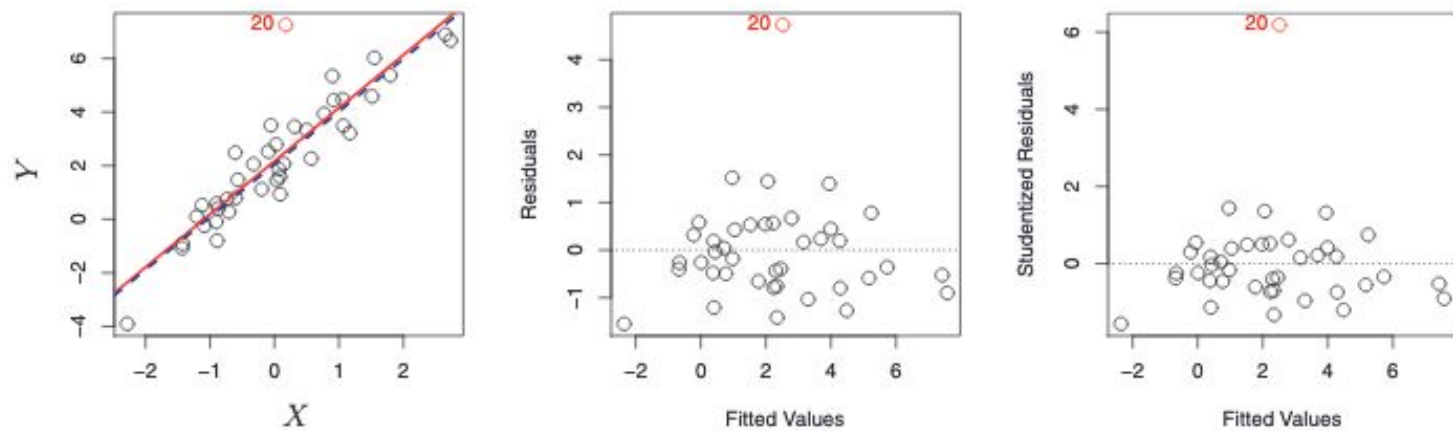


FIGURE 3.12. Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual of 6; typically we expect values between -3 and 3 .



5. Pontos com High Leverage

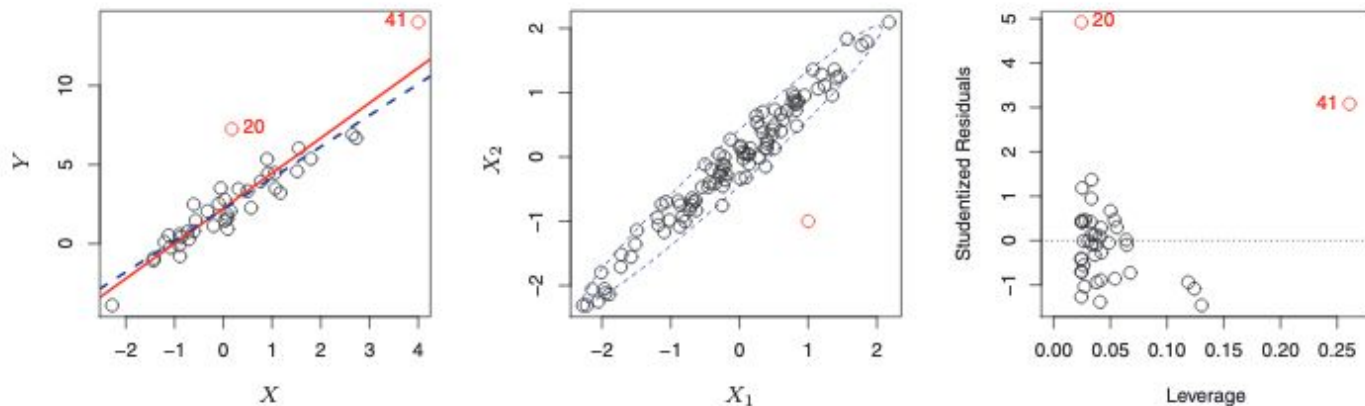


FIGURE 3.13. Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.



6. Colinearidade

Chamamos multicolinearidade o caso em que os preditores de um modelo possuem alta correlação entre si. Isso pode ser verificado com a medida VIF (*Variance Inflation Factor*). Adotamos o valor crítico de 10.



Atividade Avaliativa - Regressão Linear