Regressão Linear

Agenda

- Introdução
- Regressão Linear Simples
- Regressão Linear Múltipla
- Considerações adicionais

Introdução

Modelo de aprendizado de máquina **supervisionado**

Resposta **quantitativa**

Bom **ponto de partida** antes de se utilizar modelos mais complexos

Motivação

Qual a relação entre a e b?

A relação entre a e b é linear?

Qual o fator mais importante para se prever b?

Regressão Linear Simples

Regressão Linear Simples

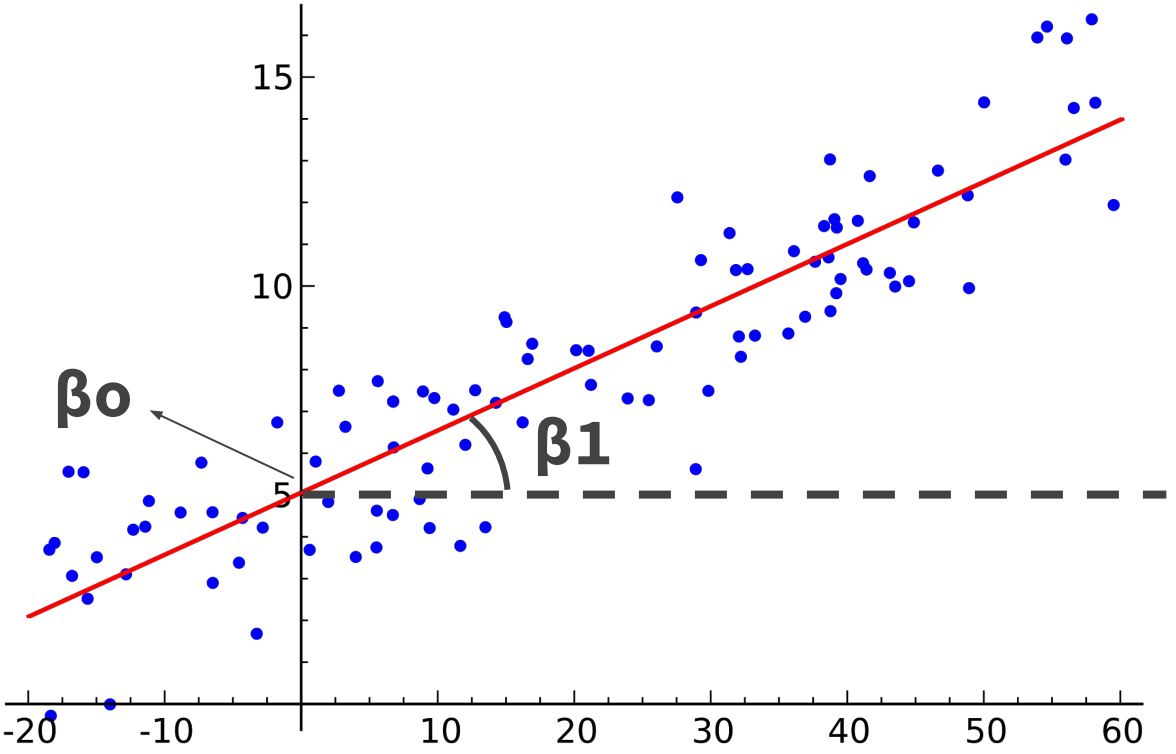
Prever Y utilizando **uma única** variável

$$Y \approx \beta o + \beta 1 * X$$

βo: constante, interceptação da reta com o eixo vertical

β1: inclinação da reta

Regressão Linear Simples



Regressão Linear Simples

Generalização: $$Y \approx \beta_0 + \beta_1 X$$

Predição: $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimando os coeficientes

Incógnitas: βo e β1

Suponha que $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ seja a predição de um valor de y baseado em um valor de x

Erro associado a essa predição: $e_i = y_i - \hat{y}_i$

Real Predição

Erro total: $e_1^2 + e_2^2 + \cdots + e_n^2$

Estimando os coeficientes

Com $e_i = y_i - \hat{y}_i$ e $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, temos que

$$e_1^2 + e_2^2 + \cdots + e_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Estimando os coeficientes

Objetivo: **minimizar** o erro total

Após um pouco de cálculo…

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimando os coeficientes

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad , \text{ sendo:}$$

$\bar{y}$ média dos valores de y

$\bar{x}$ média dos valores de x

Significado dos coeficientes

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Preço da casa = 50.000 + 20.000 * área

A **cada** metro quadrado adicionado, o valor de uma casa aumenta em **R$20000,00**

Significância dos coeficientes

$$Y \approx \beta_0 + \beta_1 X \quad \Rightarrow \quad Y = \beta_0 + \beta_1 X + \epsilon$$

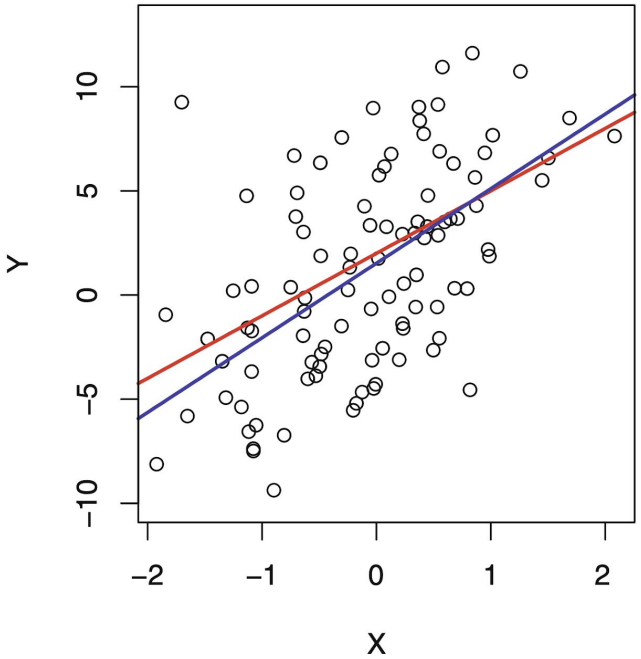
Termo de erro

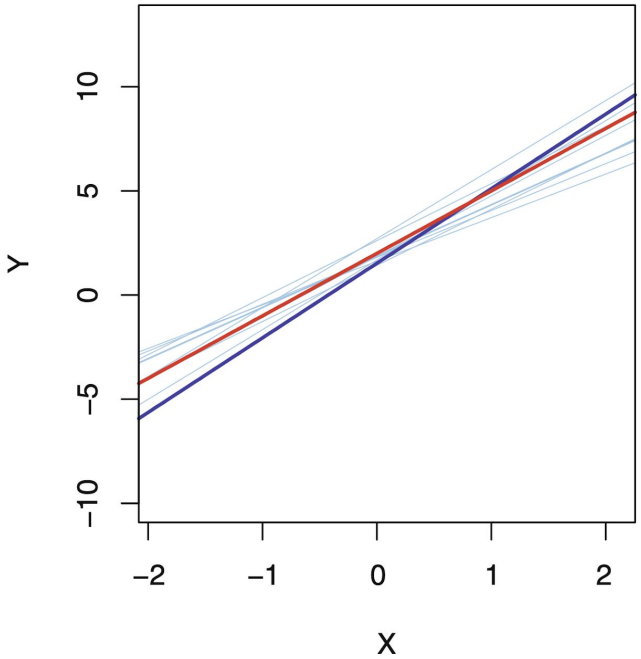
Por quê?

- A relação não é totalmente linear
- Há outros fatores que causam impacto em Y
- Erros nos dados (ruído)

Significância dos coeficientes

$$Y = 2 + 3X + \epsilon$$





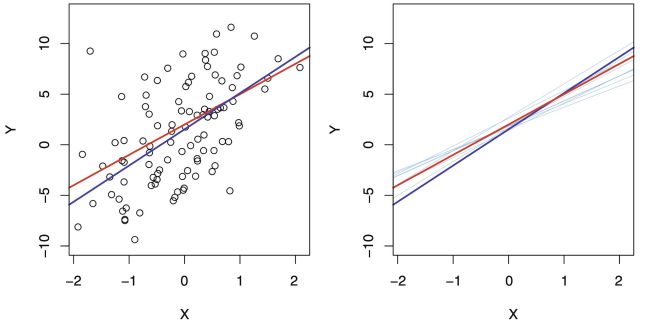
Relação verdadeira (população)

Relação estimada (amostra)

Geralmente não sabemos a relação verdadeira nem dispomos de todos os dados!

Significância dos coeficientes

Por que tantas retas?



- Obter informações de uma população a partir de uma amostra!
- Do mesmo jeito que podemos ter uma noção da média de uma amostra a partir da média da população, fazemos o mesmo com βo e β1

Significância dos coeficientes

Continuando a analogia...

- A média de vários $\hat{\mu}$ obtido de uma mesma fonte de dados é uma boa aproximação para μ
- O quão longe um $\hat{\mu}$ pode estar longe de μ ? $\frac{\sigma^2}{n}$ Desvio padrão

(na média)

Significância dos coeficientes

Continuando a analogia...

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

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

O que isso significa? **Na média**, o quanto nossas estimativas dos coeficientes diferem dos coeficientes verdadeiros!

Premissa: desvio padrão deve ser independente dos erros associados

Significância dos coeficientes

Intervalos de confiança

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

O que isso significa? **Com 95% de confiança**, os valores verdadeiros dos coeficiente estarão nesses intervalos

Significância dos coeficientes

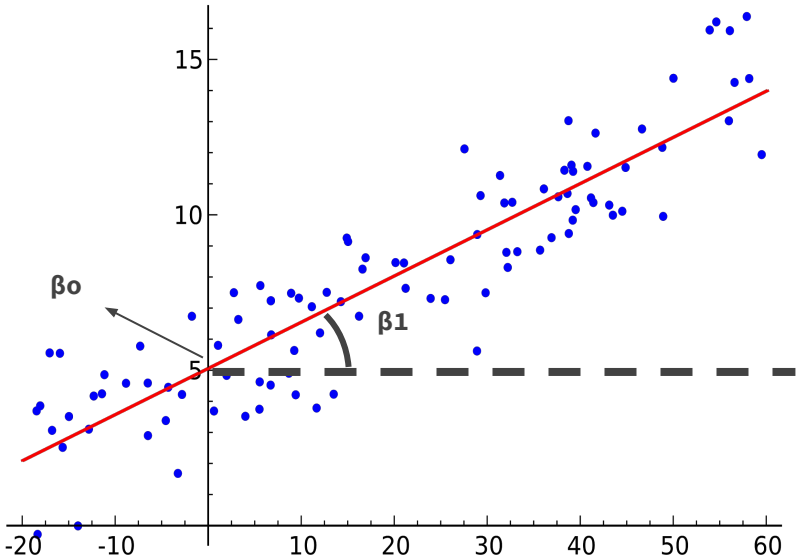
Testes de hipótese

H0: não há relação entre X e Y

H1: há relação entre X e Y

H0: β1 = 0

H1: β1 ≠ 0



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Significância dos coeficientes

Testes de hipótese

H0: β1 = 0

H1: β1 ≠ 0

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

Se não há relação entre X e Y, será uma distribuição de t com n-2 graus de liberdade

Com o valor de t, estimamos o p-valor.

Quanto **menor** o p-valor, **maior** a chance de se rejeitar H0!

Qualidade do modelo

O quão bem o meu modelo representa os dados?

- RSE *(residual standard error)*
- R2

Qualidade do modelo

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

Amostras

Valor real

Valor estimado

Bom modelo: RSE **pequeno!**

Qualidade do modelo

R2: medida de **proporção de variância explicada**

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Valor real — y_i (numerador)

Valor estimado — $\hat{y}_i$

Valor real — y_i (denominador)

Média de y — $\bar{y}$

O que isso significa?
O quanto os valores de Y podem ser explicados por X

Quanto maior o R2, mais *explicabilidade* tem o seu modelo

Regressão Linear Múltipla

Regressão Linear Múltipla

Prever Y utilizando **mais de uma** variável

Novas variáveis

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

Preço da casa $= \beta_0 + \beta_1 \times$ Área $+ \beta_2 \times$ Quartos $+ \beta_3 \times$ Vagas de garagem $+ \epsilon$

Estimando os coeficientes

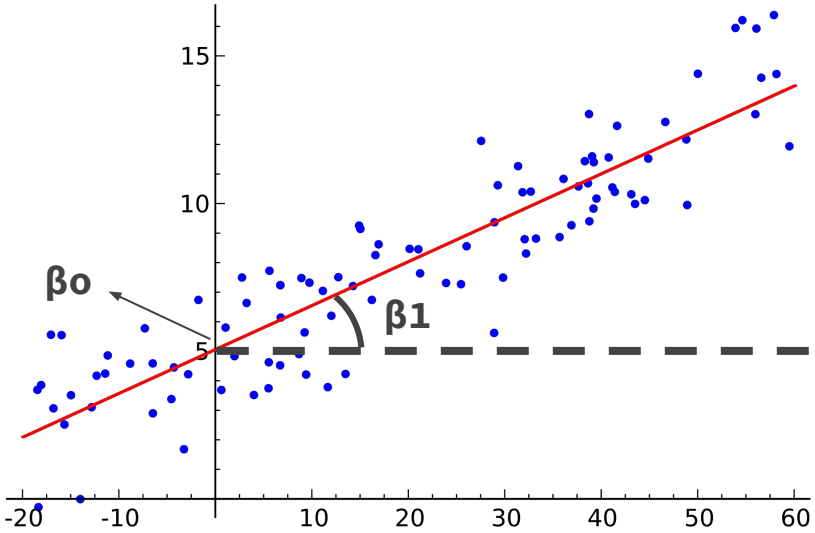
Incógnitas: βo, β1, … , βp

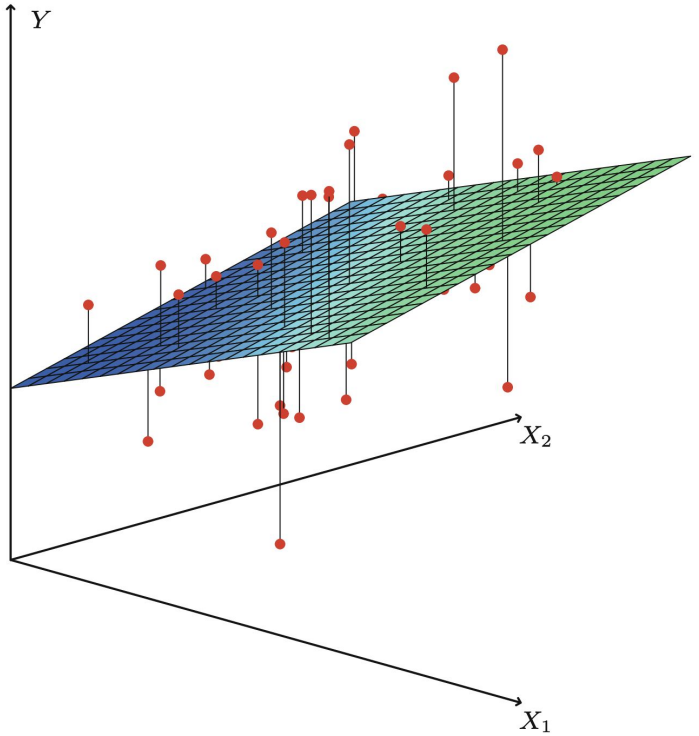
Erro associado à essa predição: $e_i = y_i - \hat{y}_i$

Real Predição

Objetivo: minimizar $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$

Estimando os coeficientes





Perguntas importantes

- Pelos menos uma das variáveis X1, X2,..., Xp é útil para prever Y?
- Todas as variáveis ajudam a explicar Y ou apenas uma parte delas é útil?
- O quão bem o meu modelo representa os dados?

Significância dos coeficientes

Testes de hipótese

H0: $\beta 1 = \beta 2 = \ldots = \beta p$

H1: Pelo menos um dos coeficientes é maior que 0

F-estatística

Significância dos coeficientes

Média de y

Predição

$$F = \frac{\left(\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2\right)/p}{\sum(y_i - \hat{y}_i)^2/(n - p - 1)}$$

Real

Variáveis

O que isso significa?
Se não há relação entre Y e as variáveis, F≈1

Se **F>1**, pode-se esperar que **há relação** entre as variáveis e Y!

Com o valor de F, estimamos o p-valor.

Quanto **menor** o p-valor, **maior** a chance de se rejeitar H0!

Escolhendo variáveis

Conseguimos rejeitar a hipótese nula, mas quais variáveis são as importantes?

Modelo 1: Y ~ X1

Modelo 2: Y ~ X2

Modelo 3: Y ~ X1 e X2

Para um número p de variáveis, há 2^p modelos, o que pode ser **computacionalmente custoso!**

Escolhendo variáveis

Forward selection

Começa com um modelo nulo e adiciona variáveis de acordo com a menor soma residual dos quadrados

Backward selection

Começa com todas as variáveis e vai retirando a variável com o maior p-valor

Mixed selection

Começa com um modelo nulo e adiciona variáveis de acordo com a menor soma residual. Caso uma variável fique com um p-valor alto, ela é retirada

Qualidade do modelo

O quão bem o meu modelo representa os dados?

- RSE *(residual standard error)*
- R2

As mesmas métricas da regressão linear simples

Qualidade do modelo

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

Variáveis

Predição

Bom modelo: RSE **pequeno!**

Qualidade do modelo

R2 sempre irá aumentar caso haja adição de uma nova variável, fazendo com que o fit dos dados de treino seja melhor (não necessariamente nos dados de teste).

Deve-se buscar mudanças *sensíveis* no valor de R2 com a adição de uma nova variável.

Considerações adicionais

Preditores qualitativos

Em uma regressão linear, é possível existir uma variável qualitativa

Exemplo: tipo de moradia (apartamento, flat, casa...)

Preditores qualitativos

Exemplo: tipo de moradia (apartamento ou casa)

$$x_i = \begin{cases} 1 & \text{Se moradia é uma casa} \\ 0 & \text{Se moradia é apartamento} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{Se moradia é uma casa} \\ \beta_0 + \epsilon_i & \text{Se moradia é apartamento} \end{cases}$$

Preditores qualitativos

β'

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{Se moradia é uma casa} \\ \beta_0 + \epsilon_i & \text{Se moradia é apartamento} \end{cases}$$

Nesse caso…

β0 representa o preço médio dos apartamentos
β0 + β1 representa o preço médio das casas

Premissas

Premissa aditiva

O efeito da variável Xj em Y é independente dos valores das outras variáveis.

Premissa linear

A mudança no valor de Y em relação ao aumento em uma unidade em Xj é constante

Como obedecer as premissas?

Premissa aditiva

Adição de termos de interação: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

Premissa linear

Utilizar transformações não lineares nos dados: ao quadrado, raiz, log…

Possíveis problemas

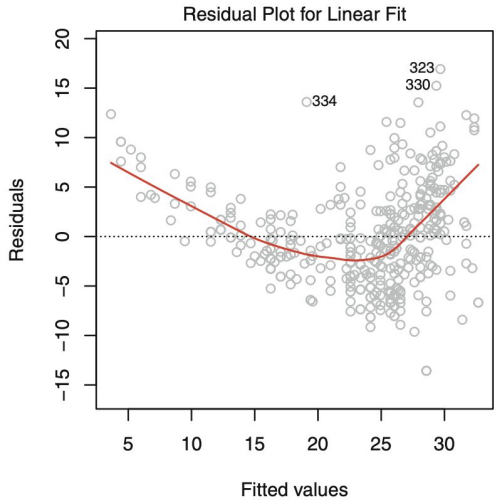
---

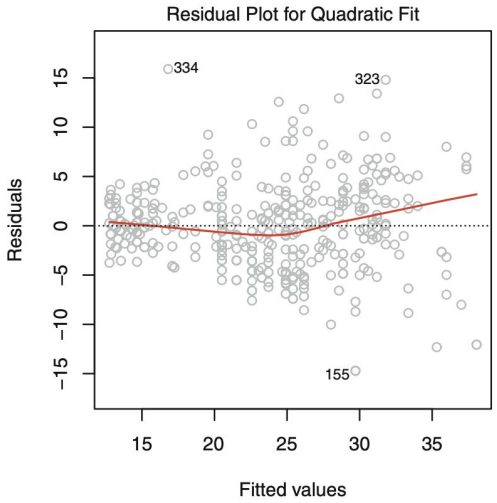
1. *Não linearidade entre variável alvo e variáveis preditoras*
2. *Variância não constante dos termos de erro*
3. *Outliers*
4. *Valores incomuns de variáveis*
5. *Colinearidade*

Não linearidade entre variável alvo e variáveis preditoras

Uso do gráfico de resíduos

Resíduos possuem um padrão



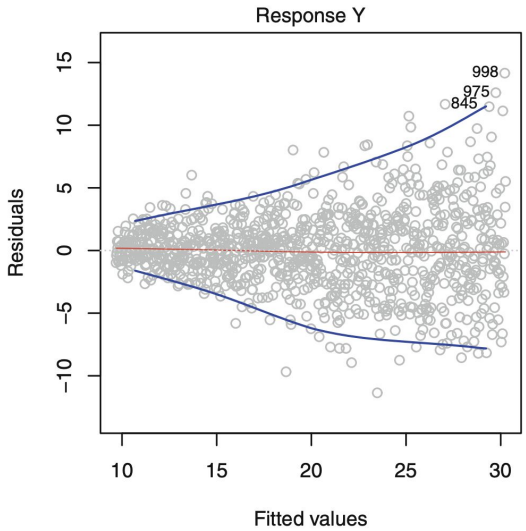


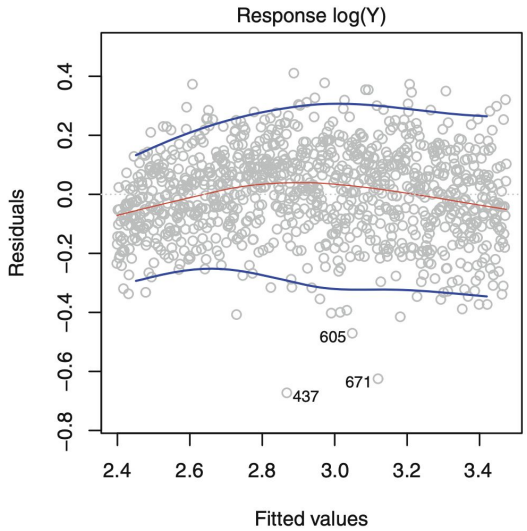
Padrão reduzido pela adição de uma variável ao quadrado

Variância não constante dos termos de erro

Uso do gráfico de resíduos

Resíduos possuem heterocedasticidade (formato de funil)

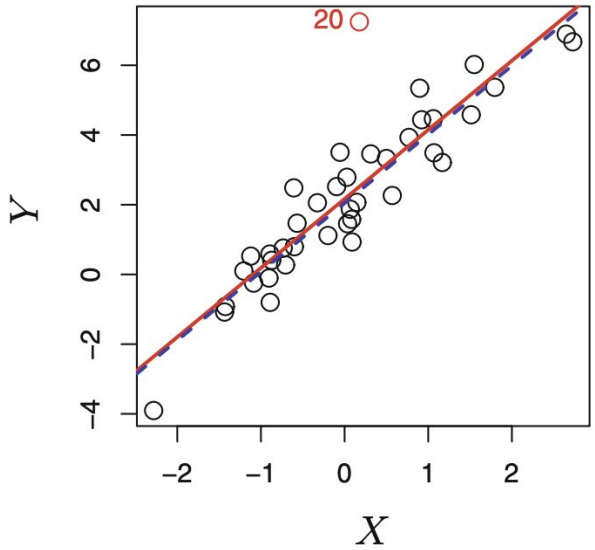


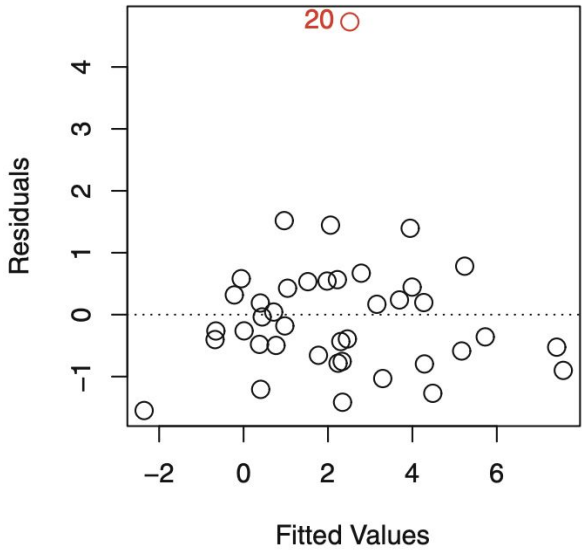


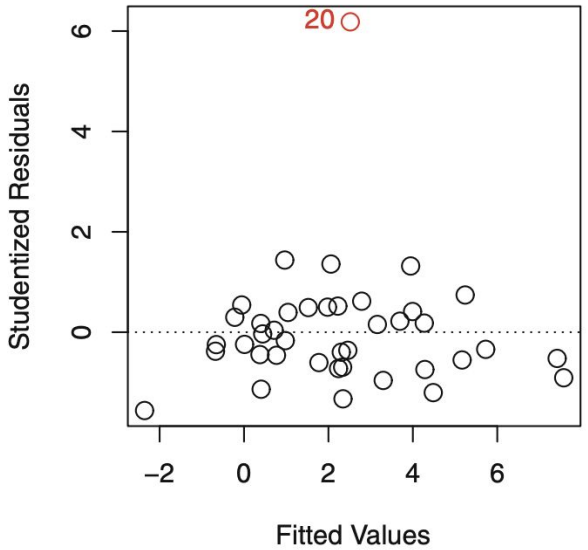
Heterocedasticidade reduzida pela transformação logarítmica de Y

Outliers

Valor predito é muito diferente do valor real (ideal: remover)

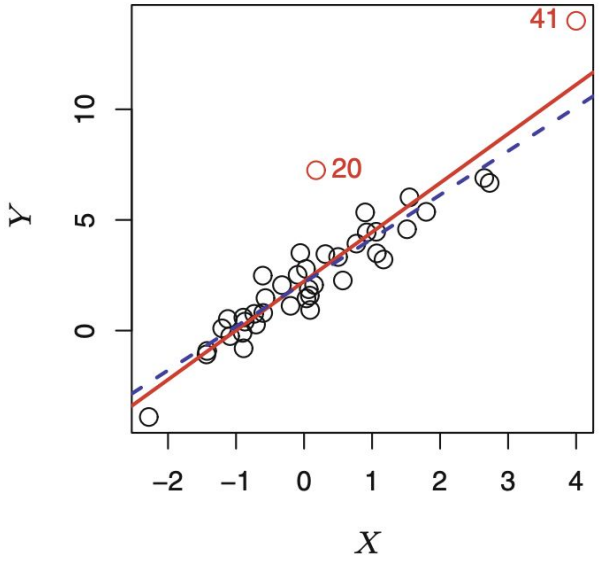


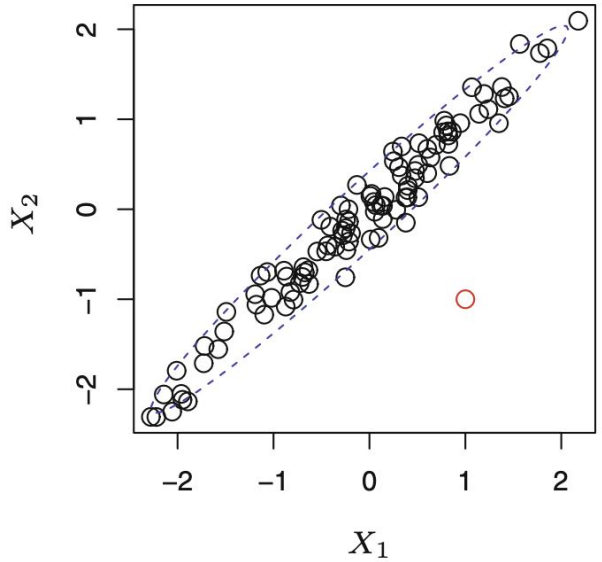


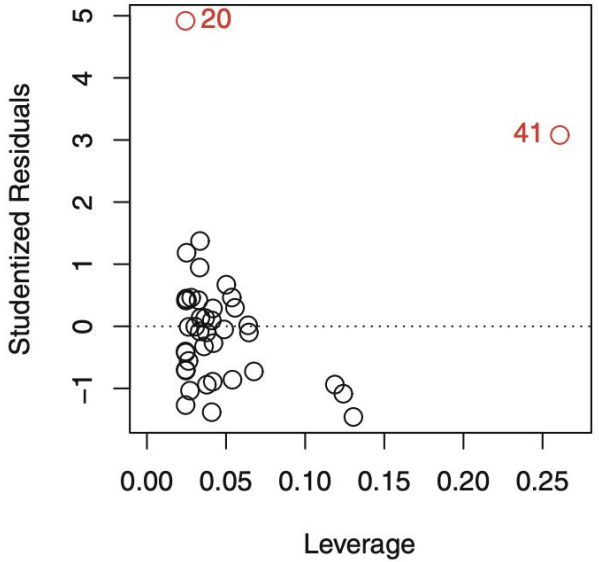


Valores incomuns de variáveis

Valor de X é muito diferente do usual (ideal: remover)





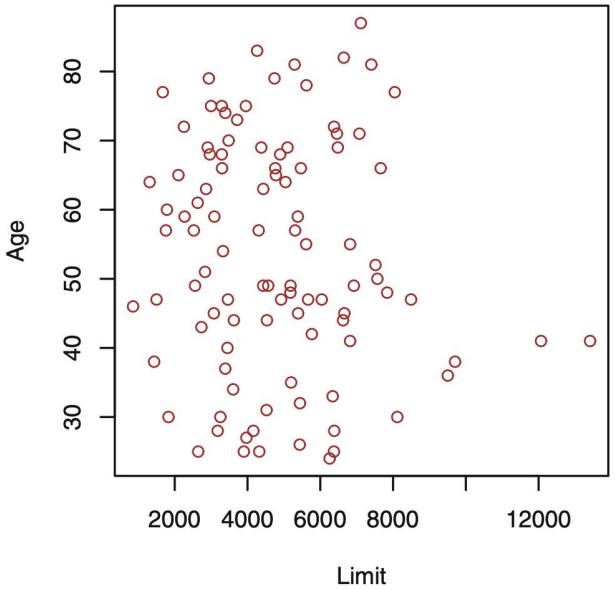


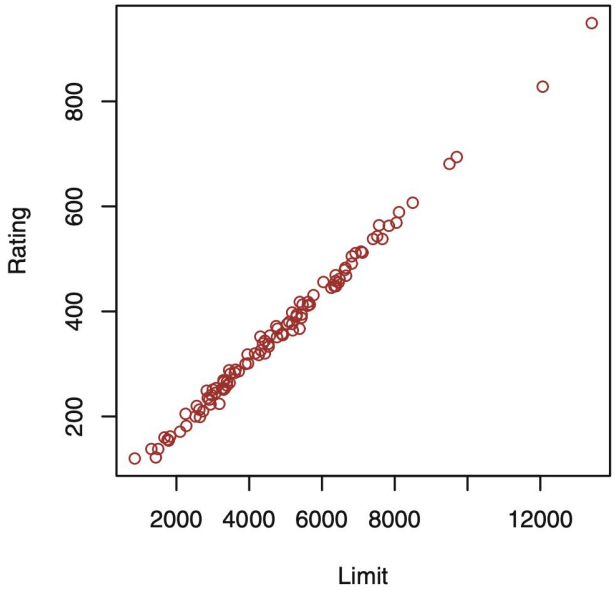
Colinearidade

Variáveis muito parecidas entre si

Ideal: escolher uma ou fazer uma combinação das variáveis colineares

Variáveis não colineares





Variáveis colineares

Colinearidade

Multicolinearidade: colinearidade entre mais de 2 variáveis

Quanto maior o VIF (fator de inflação da variância), maior a colinearidade

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j | X_{-j}}}$$

R2 da regressão de Xj pelas outras variáveis

Dúvidas?