Regressão Linear

Teoria e Prática

curso-r

2021-01-09

Agenda

- O que é e quando usar
- Parâmetro vs estimador
- Teste de Hipóteses e valor-p
- Interpretação dos parâmetros
- Desempenho: EQM e EPR
- Outliers
- Regressão Linear Múltipla
- Preditores Categóricos
- Transformações Não Lineares dos Preditores
- Interações
- Multicolinearidade
- Fazendo Predições
- Interpretabilidade da Predição
- Limitações da Regressão Linear
- Exercício para casa
- Sobreajuste (overfitting)
- Regularização
- LASSO e Seleção de Preditores
- Validação Cruzada

Referências

- Aprendizagem de Máquinas: Uma Abordagem Estatística (Rafael Izbicki e Thiago Mendonça, 2020)
- Introduction to Statistical Learning (Hastie, et al)
- Ciência de Dados: Fundamentos e Aplicações

Usos da Regressão Linear ("duas culturas")

- Predição
- Inferência

Predição

Em muitas situações X está disponível facilmente mas, Y não é fácil de descobrir. (Ou mesmo não é possível descobrí-lo).

$$\hat{Y} = \hat{f}\,(X)$$

é uma boa estimativa. Neste caso não estamos interessados em como é a estrutura $\hat{f}$ desde que ela apresente predições boas para Y.

Usos da Regressão Linear ("duas culturas")

- Predição
- Inferência

Inferência

Em inferência estamos mais interessados em entender a relação entre as variáveis explciativas X e a variável resposta Y.

Por exemplo:

- Quais são as variáveis que estão mais relacionadas com a respostas?
- Qual a relação entre a resposta e cada um dos preditores?

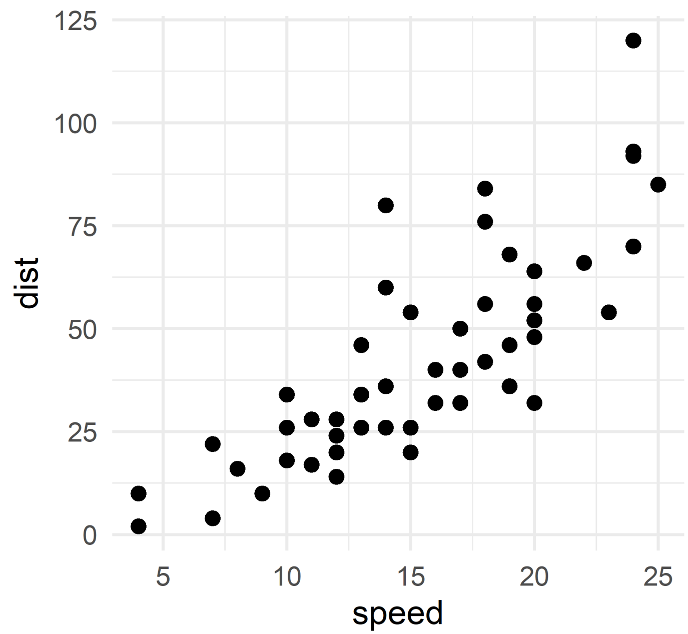
O que é e quando usar

Regressão Linear Simples

$$y = \beta_0 + \beta_1 x$$

Exemplo:

$$dist = \beta_0 + \beta_1 speed$$



Ver ISL página 61 (Simple Linear Regression).

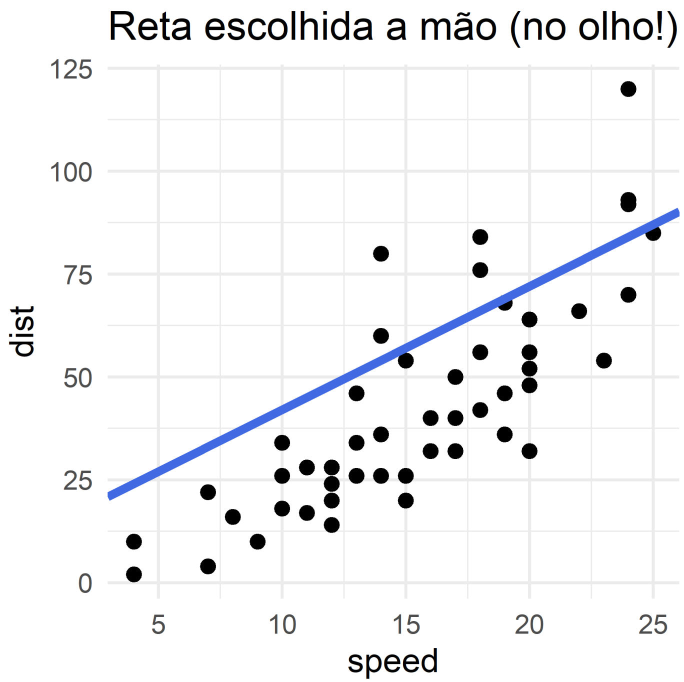
O que é e quando usar

Regressão Linear Simples

$$y = \beta_0 + \beta_1 x$$

Exemplo:

$$dist = \beta_0 + \beta_1 speed$$



Ver ISL página 61 (Simple Linear Regression).

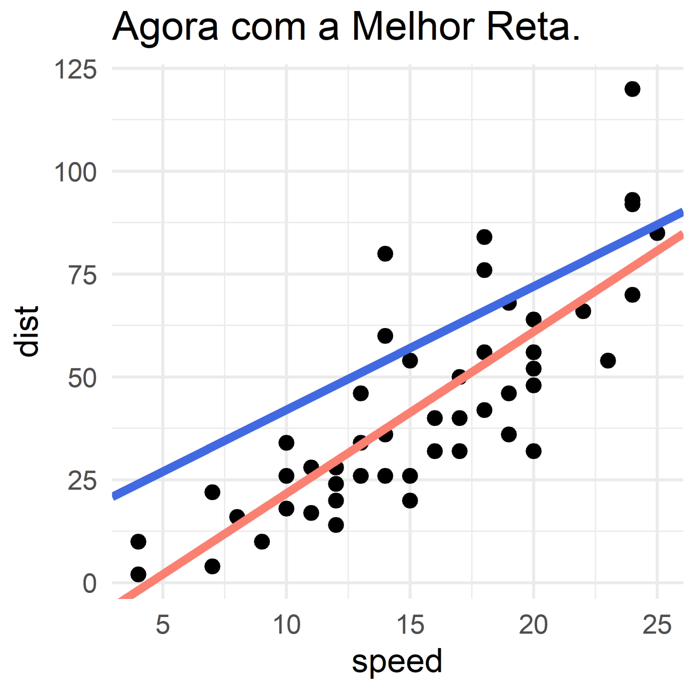
O que é e quando usar

Regressão Linear Simples

$$y = \beta_0 + \beta_1 x$$

Exemplo:

$$dist = \beta_0 + \beta_1 speed$$



Ver ISL página 61 (Simple Linear Regression).

O que é e quando usar

Regressão Linear Simples

$$y = \beta_0 + \beta_1 x$$

```r
# ajuste de uma regressão linear simples no R
melhor_reta <- lm(dist ~ speed, data = cars)
melhor_reta
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

Ver ISL página 61 (Simple Linear Regression).

"Melhor Reta" segundo o quê?

Queremos a reta que **erre menos**.

Uma medida de erro: **E**rro **Q**uadrático **M**édio.

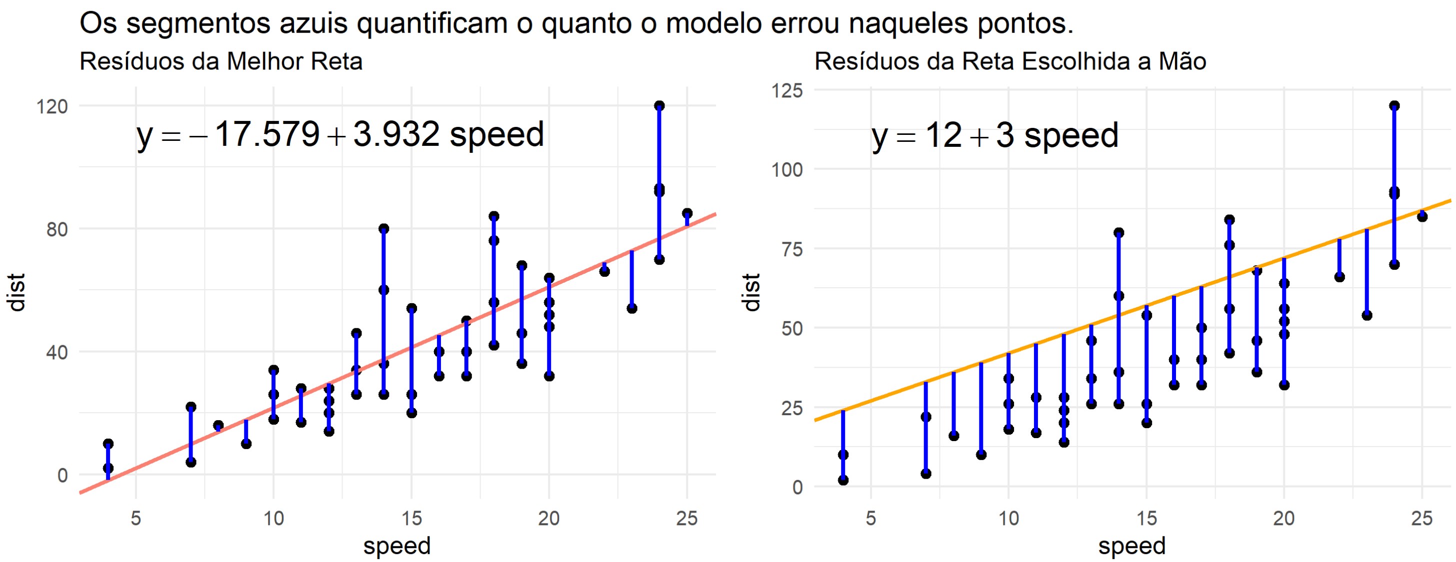
$$EQM = \frac{1}{N} \sum (y_i - \hat{y}_i)^2$$

"Melhor Reta" segundo o quê?

Queremos a reta que **erre menos**.

Uma medida de erro: **E**rro **Q**uadrático **M**édio.

$$EQM = \frac{1}{N}\sum(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$$



"Melhor Reta" segundo o quê?

Queremos a reta que **erre menos**.

Uma medida de erro: **E**rro **Q**uadrático **M**édio.

$$EQM = \frac{1}{N}\sum(y_i - \hat{y}_i)^2$$

Ou seja, nosso **objetivo** é

Encontrar $\hat{\beta}_0$ e $\hat{\beta}_1$ que nos retorne o ~menor~ EQM.

... que é o mesmo que dizer "encontrar a melhor reta que explique os dados".

OBS: o EQM é a nossa **Função de Custo**.

```
# lembrete: exercício 1 do script!
```

Qual o valor ótimo para β_0 e β_1?

No nosso exemplo, a nossa **HIPÓTESE** é de que

$$dist = \beta_0 + \beta_1 speed$$

Então podemos escrever o Erro Quadrático Médio como

$$EQM = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum(y_i - (\hat{\beta}_0 + \hat{\beta}_1 speed))^2$$

Com ajuda do Cálculo é possível mostrar que os valores ótimos para β_0 e β_1 são

$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

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

Já que vieram do EQM, eles são chamados de **Estimadores de Mínimos Quadrados**.

```
# lembrete: exercício 2 do script!
```

Depois de estimar...

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

Exemplo:

$$\hat{dist} = \hat{\beta}_0 + \hat{\beta}_1 speed$$

Colocamos um ^ em cima dos termos para representar "estimativas". Ou seja, $\hat{y}_i$ é uma estimativa de y_i.

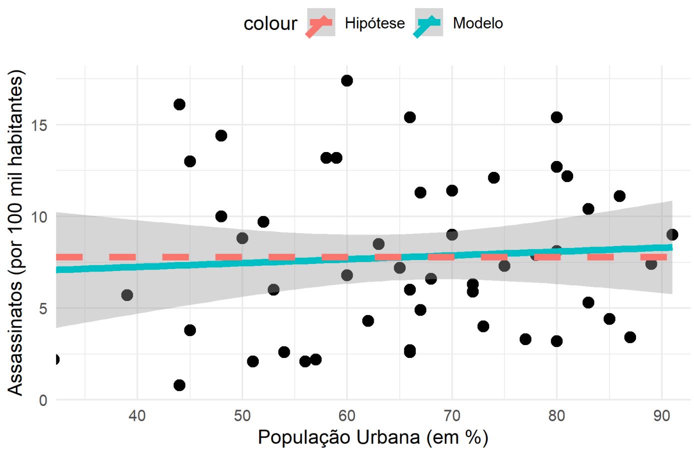
No nosso exemplo,

- $\hat{\beta}_0$ é uma estimativa de β_0 e vale `-17.579`.
- $\hat{\beta}_1$ é uma estimativa de β_1 e vale `3.932`.
- $\hat{dist}$ é uma estimativa de $dist$ e vale `-17.579 + 3.932 x speed`.

```
# Exercício: se speed for 15 m/h, quanto que
# seria a distância dist esperada?
```

Teste de Hipóteses e valor-p

Exemplo: relação entre População Urbana e Assassinatos.



Modelo proposto:

$$y = \beta_0 + \beta_1 x$$

Hipótese do pesquisador:

> "Assassinatos não estão relacionados com a proporção de população urbana de uma cidade."

Tradução da hipótese em termos matemáticos:

$$H_0 : \beta_1 = 0 \quad vs \quad H_a : \beta_1 \neq 0$$

Se a hipótese for verdade, então o β_1 deveria ser zero. Porém, os dados disseram que $\hat{\beta}_1 = 0.02$.

0.02 é diferente de 0.00?

0.02 é diferente de 0.00?

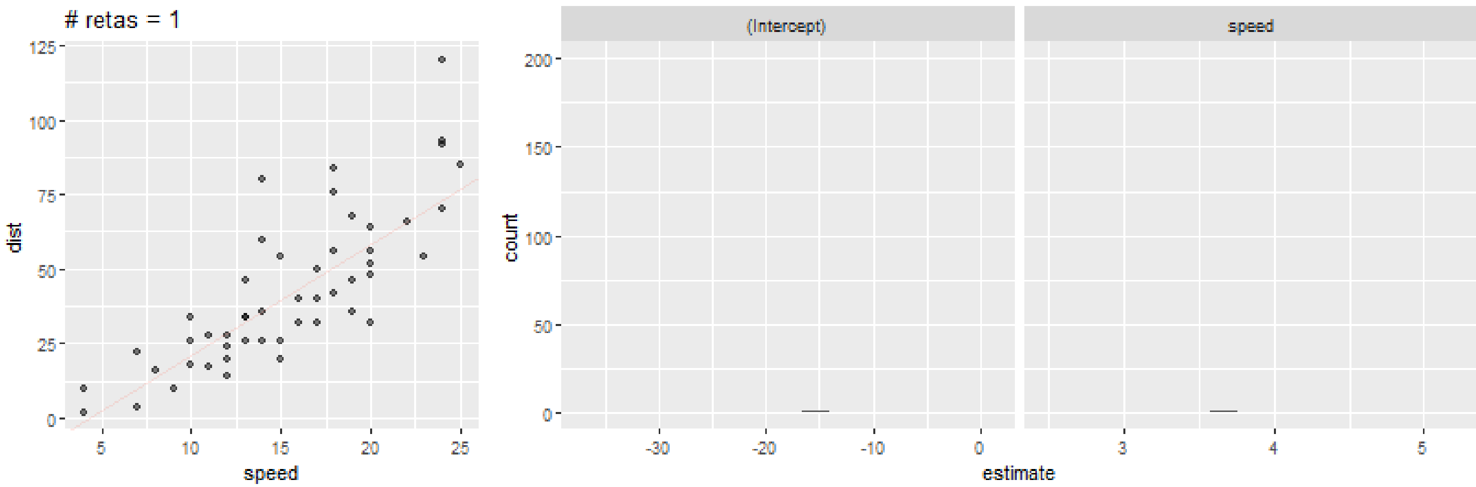
Saída do R

```
##
## Call:
## lm(formula = Murder ~ UrbanPop, data = USArrests)
##
## 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 *
## UrbanPop     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
```

0.02 é diferente de 0.00?

Conceito importante: Os estimadores ($\hat{\beta}_0$ e $\hat{\beta}_1$ no nosso caso) têm distribuições de probabilidade.

Simulação de 1000 retas (ajustadas com dados diferentes).



0.02 é diferente de 0.00?

Conceito importante: Os estimadores ($\hat{\beta}_0$ e $\hat{\beta}_1$ no nosso caso) têm distribuições de probabilidade.

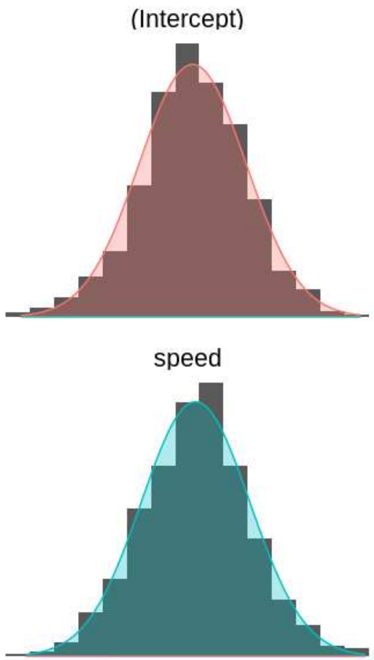
A Teoria Assintótica nos fornece o seguinte resultado:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \overset{a}{\sim} t(N-2)$$

Em que

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{EQM}{\sum(x_i - \bar{x})^2}}$$

Usamos essas distribuições assintóticas para testar as hipóteses.



0.02 é diferente de 0.00?

Conceito importante: Os estimadores ($\hat{\beta}_0$ e $\hat{\beta}_1$ no nosso caso) têm distribuições de probabilidade.

A Teoria Assintótica nos fornece o seguinte resultado:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \overset{a}{\sim} t(N-2)$$

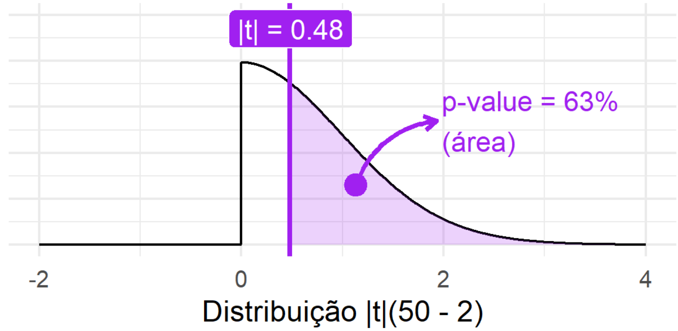
Em que

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{EQM}{\sum(x_i - \bar{x})^2}}$$

Usamos essas distribuições assintóticas para testar as hipóteses.

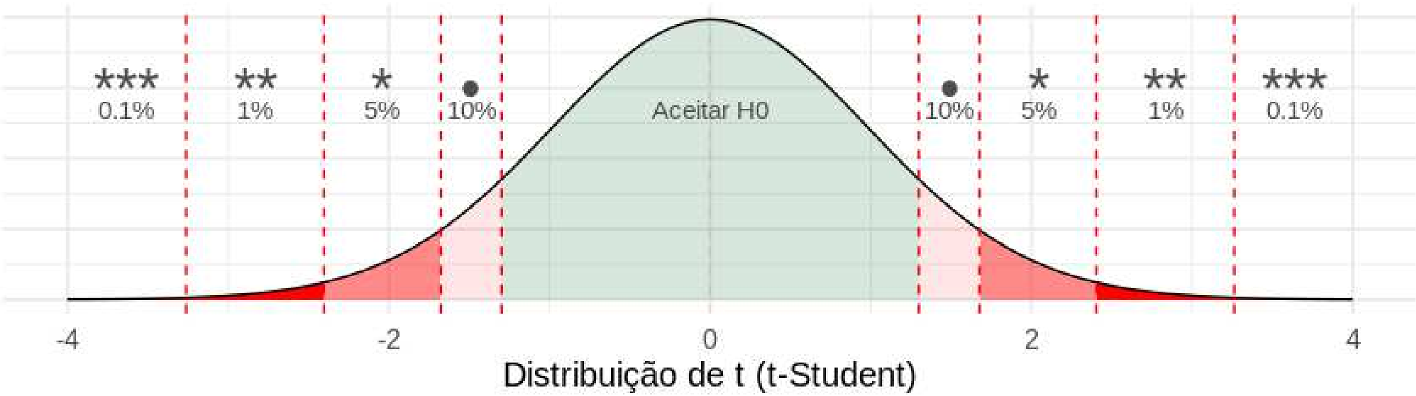
No nosso exemplo, a hipótese é $H_0 : \beta_1 = 0$, então

$$t = \frac{0.02 - 0}{0.04} = 0.48$$



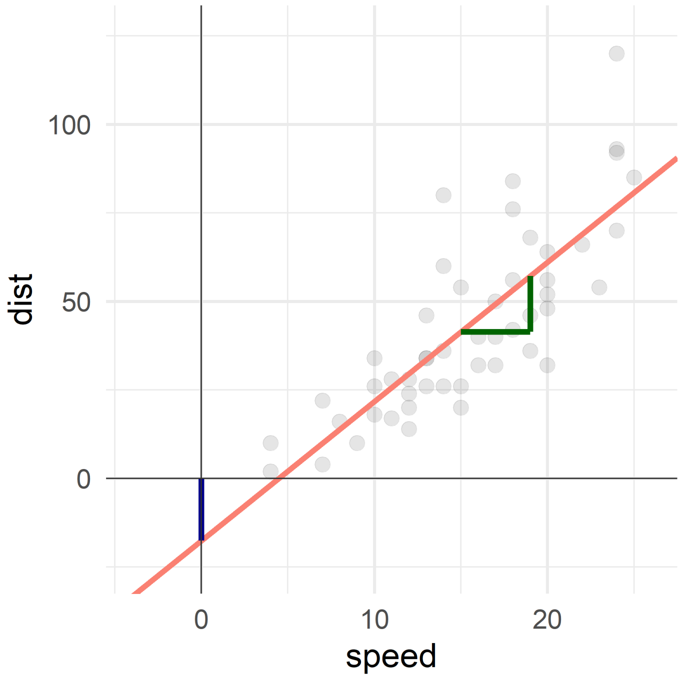
0.02 é diferente de 0.00?

Então agora podemos tomar decisão! Se a estimativa cair muito distante da distribuição t da hipótese 0, decidimos por **rejeitá-la**. Caso contrário, decidimos por **aceitá-la** como verdade.




```
## NO R:
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.41594    2.90669   2.207   0.0321 *
## UrbanPop     0.02093    0.04333   0.483   0.6312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretação dos parâmetros



$$y = \beta_0 + \beta_1 x$$

Interpretações matemáticas

β_0 é o lugar em que a reta cruza o eixo Y.

β_1 é a derivada de Y em relação ao X. É quanto Y varia quando X varia em 1 unidade.

Interpretações estatísticas

β_0 é a distância percorrida esperada quando o carro está parado (X = 0).

β_1 é o efeito médio na distância por variar 1 ml/h na velocidade do carro.

Teste de Hipóteses e valor-p

Exercício 3 do script:

No R, use a função `summary(melhor_reta)` (ver slide 9) para decidir se `speed` está associado com `dist`. Descubra o valor-p associado.

lembrete: o banco de dados se chama `cars`.

Exercício 4 do script:

Interprete o parâmetro β_1.

Intervalo de confiança para β_1

Intervalo de 95%

$$[\hat{\beta} - 1,96 * \hat{\sigma}_{\beta_1}, \hat{\beta} + 1,96 * \hat{\sigma}_{\beta_1}]$$

Intervalo de 90%

$$[\hat{\beta} - 1,64 * \hat{\sigma}_{\beta_1}, \hat{\beta} + 1,64 * \hat{\sigma}_{\beta_1}]$$

Intervalo de 1 - α%

$$[\hat{\beta}_1 - q_\alpha * \hat{\sigma}_{\beta_1}, \hat{\beta} + q_\alpha * \hat{\sigma}_{\beta_1}]$$

em que q_α é o quantil da $Normal(0,1)$.

Ver ISL página 66 (Assessing the Accuracy of the Model).

O modelo está bom?

EQM e EPR

ERP significa *E*rro *P*adrão dos *R*esíduos e é definido como

$$EPR = \frac{\sum (y_i - \hat{y}_i)^2}{N-2} = \frac{SQR}{N-2}$$

O **2** no denominador decorre do fato de termos **2 parâmetros** para estimar no modelo.

- Se $y_i = \hat{y}_i \quad \rightarrow EPR = 0 \downarrow$
- Se $y_i >> \hat{y}_i \rightarrow EPR = alto \uparrow$
- Se $y_i << \hat{y}_i \rightarrow EPR = alto \uparrow$

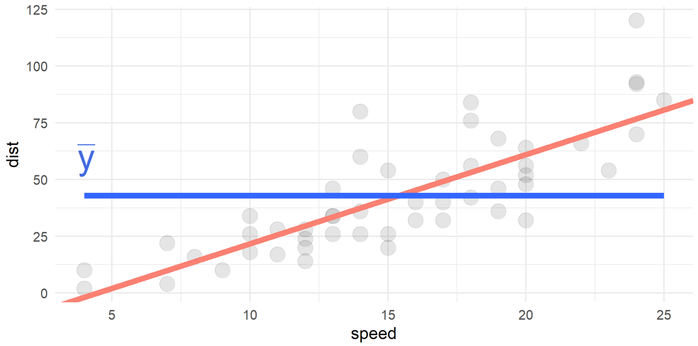
Problema: Como sabemos se o EPR é grande ou pequeno?

Ver ISL página 68 (Assessing the Accuracy of the Model).

O modelo está bom?

R-quadrado (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SQR}{SQT}$$



$R^2 \approx 1 \rightarrow SQR << SQT$.
$R^2 \approx 0 \rightarrow reta$ em cima da $reta$.

Problema do R^2 é que ele sempre aumenta conforme novos preditores vão sendo incluídos.

Ver ISL página 68 (Assessing the Accuracy of the Model).

O modelo está bom?

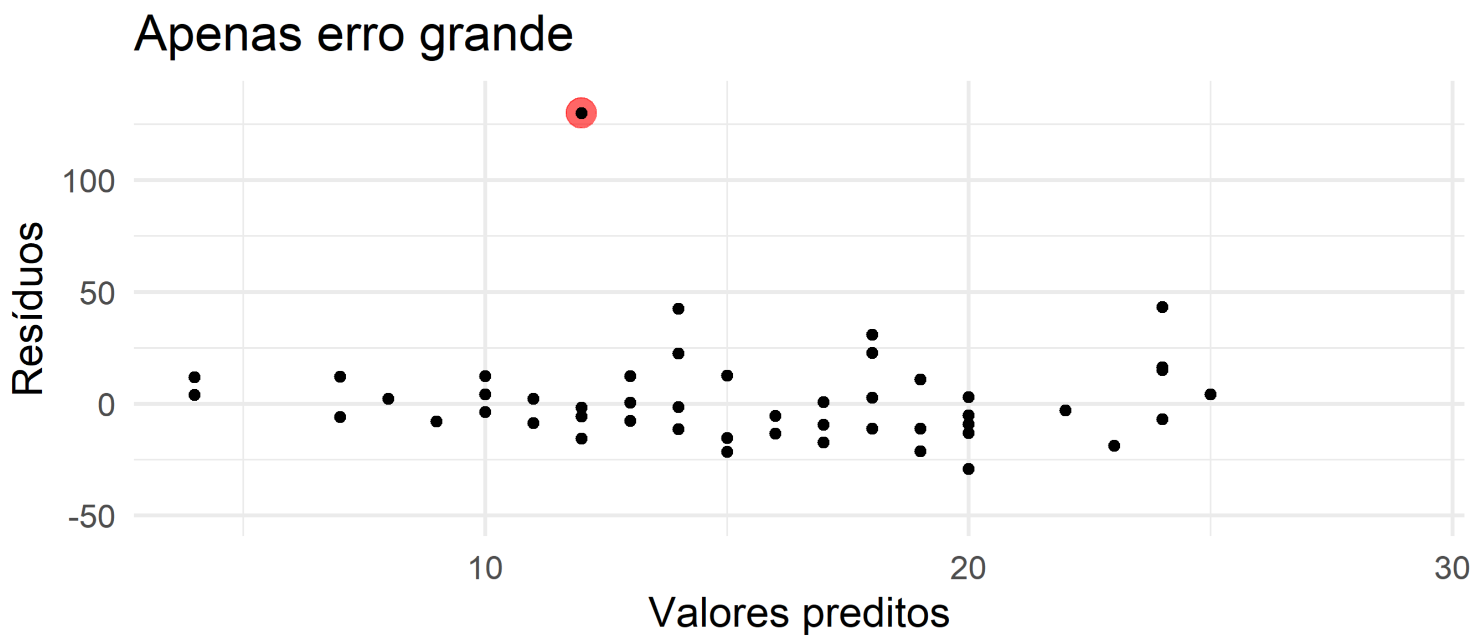
R-quadrado ajustado

$$R^2 = 1 - \frac{SQR}{SQT}\frac{N-1}{N-p}$$

Em que p é o número de parâmetros do modelo (no caso da regressão linear simples, $p = 2$).

```
# lembrete: exercícios 5 e 6 do script!
```

Outliers

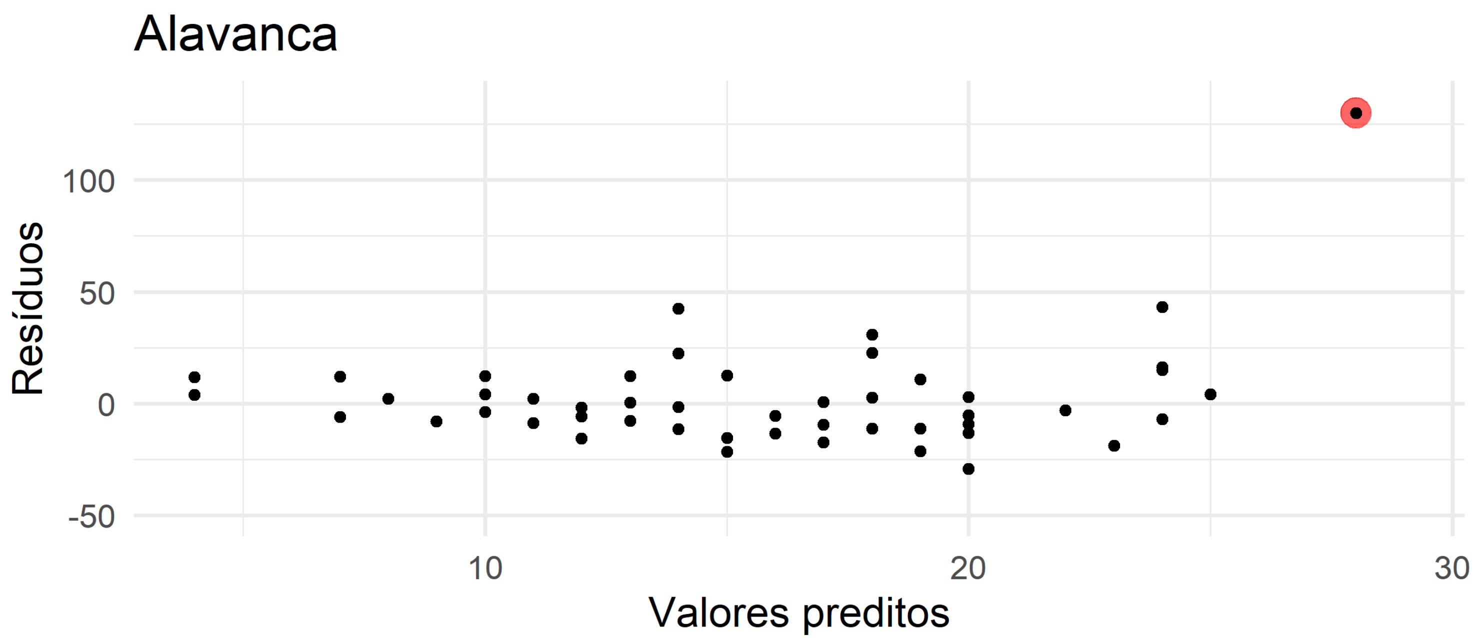


Resíduo = $y_i - \hat{y}_i$ (observado - esperado).

```
# lembrete: faça o exercício 7 do script
```

Ver ISL página 96 (Outliers).

Outliers

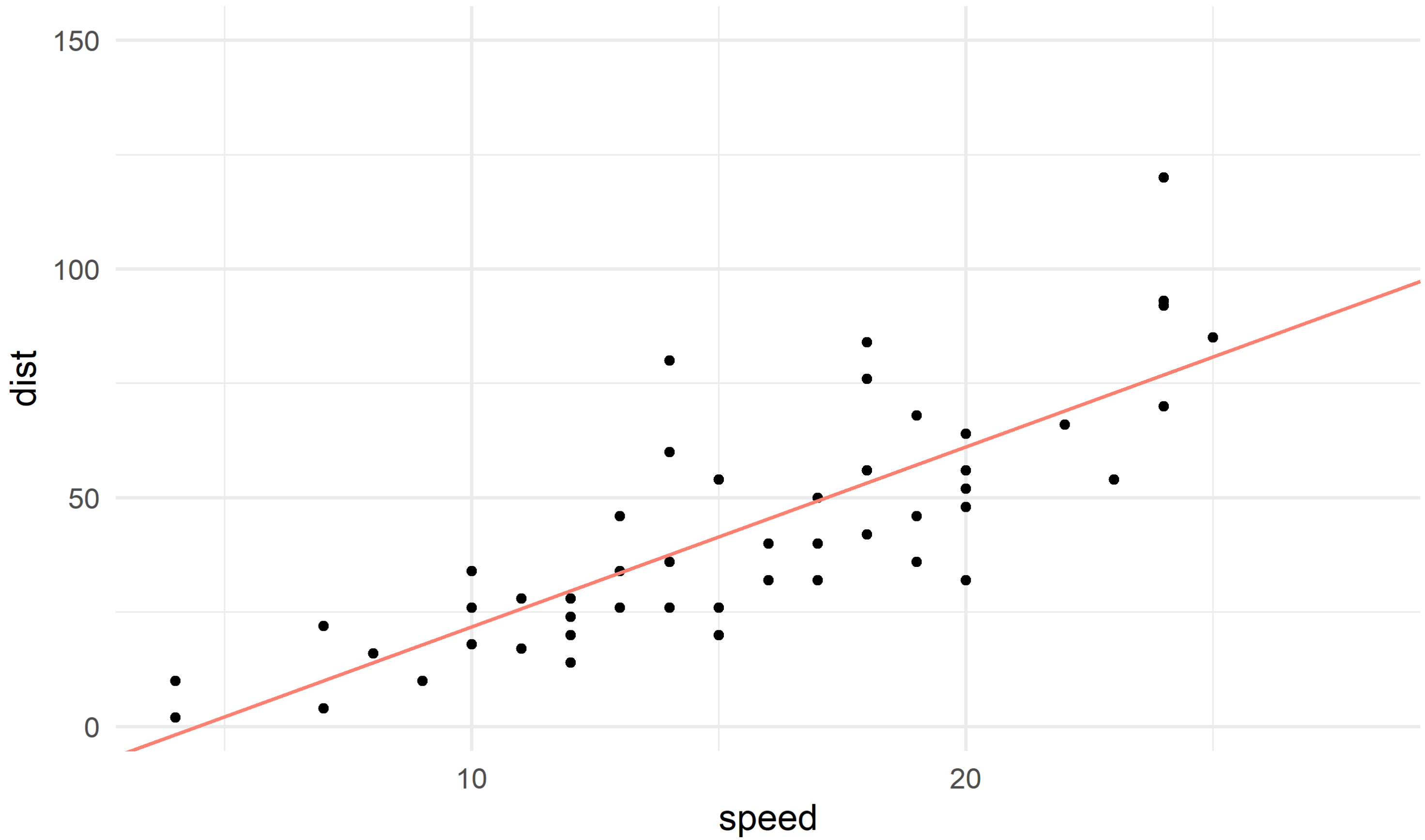


Resíduo = $y_i - \hat{y}_i$ (observado - esperado).

```
# lembrete: faça o exercício 7 do script
```

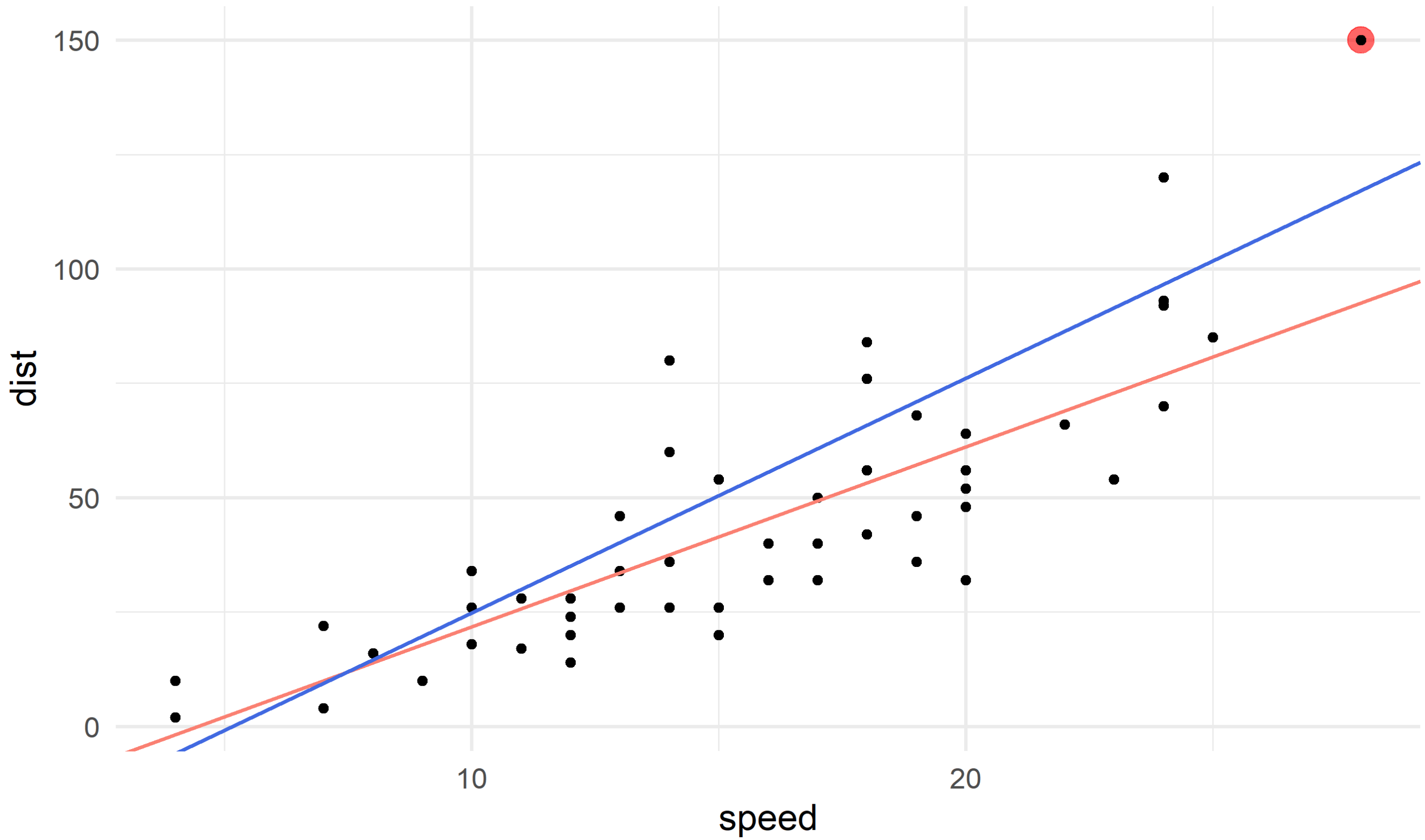
Ver ISL página 96 (Outliers).

Outliers



Ver ISL página 96 (Outliers).

Outliers



Ver ISL página 96 (Outliers).

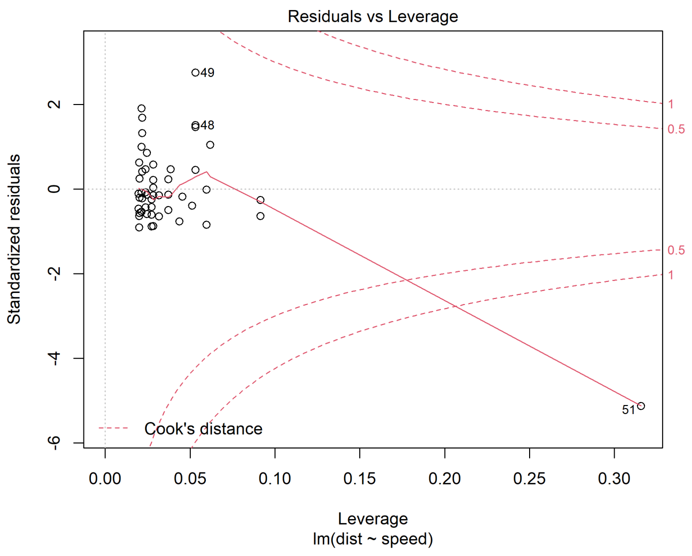
Outliers

Distância de Cook

A distância de Cook mede o efeito de excluir uma dada observação.

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{pEQM}$$

```
modelo <- lm(dist ~ speed, data
plot(modelo)
cooks.distance(modelo)
```



Ver Distância de Cook na Wikipedia.

Outliers

Diagnóstico

- Visualização univariada (histograma, boxplot);
- Comparação do valor com o desvio padrão;
- Distância de Cook

Tratamentos

- Transformações log(), etc;
- Categorização;
- Remover os valores extremos (raramente boa ideia);

Regressão Linear Múltipla

Regressão Linear Simples

$$y = \beta_0 + \beta_1 x$$

Regressão Linear Múltipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

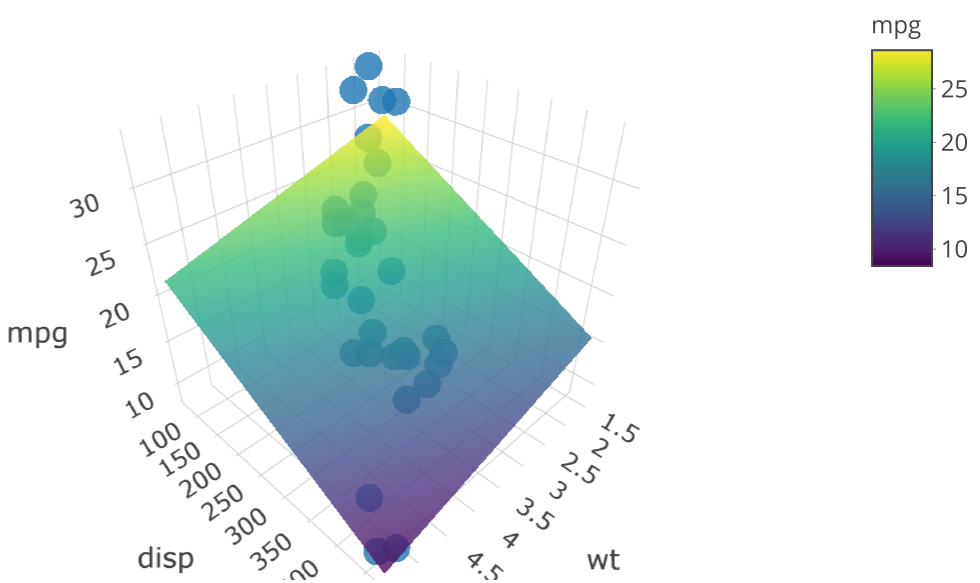
```
# ajuste de uma regressão linear múltipla no R
modelo_boston <- lm(medv ~ lstat + age, data = Boston)
summary(modelo_boston)
#             Estimate Std.Error t value Pr(>|t|)
# (Intercept) 33.22    0.73       45.4    < 2e-16 ***
# lstat       -1.03    0.04      -21.4    < 2e-16 ***
# age          0.03    0.01        2.8    0.00491 **
```

Ver ISL página 71 (Multiple Linear Regression).

Regressão Linear Múltipla

Exemplo: Plano em vez de reta

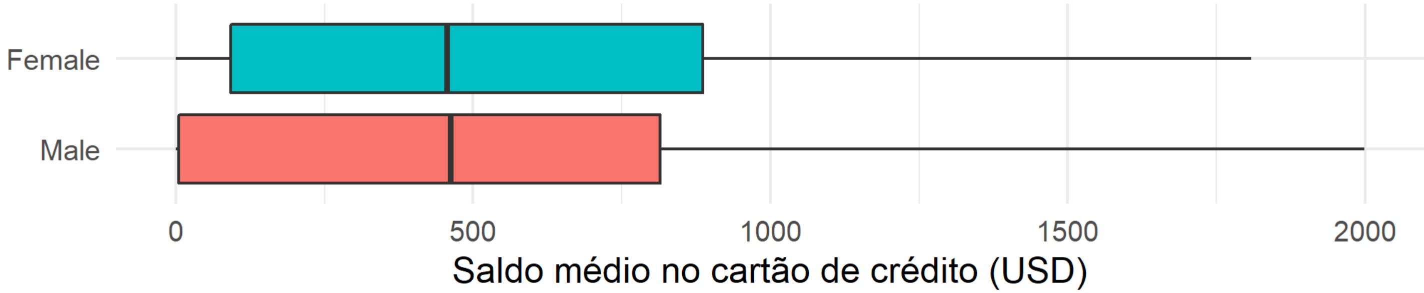
Modelo: `lm(mpg ~ disp + wt, data = mtcars)`



Preditores Categóricos

Preditor com apenas 2 categorias

Saldo médio no cartão de crédito é diferente entre homens e mulheres?



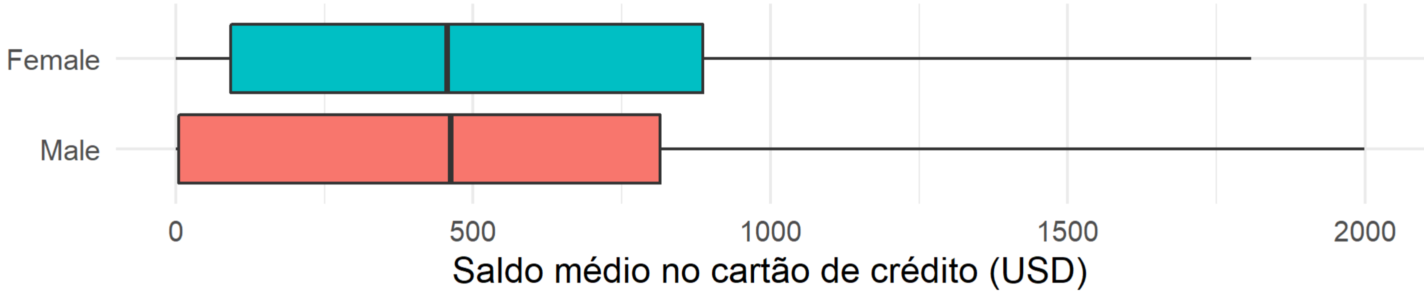

```
summary(lm(Balance ~ Gender, data = Credit))
# Coefficients:
#               Estimate  Std.Error  t value Pr(>|t|)
# (Intercept)     509.80  33.13       15.389   <2e-16 ***
# GenderFemale     19.73  46.05        0.429    0.669
```

Ver ISL página 84 (Predictors with Only Two Levels).

Preditores Categóricos

Preditor com apenas 2 categorias

Saldo médio no cartão de crédito é diferente entre homens e mulheres?



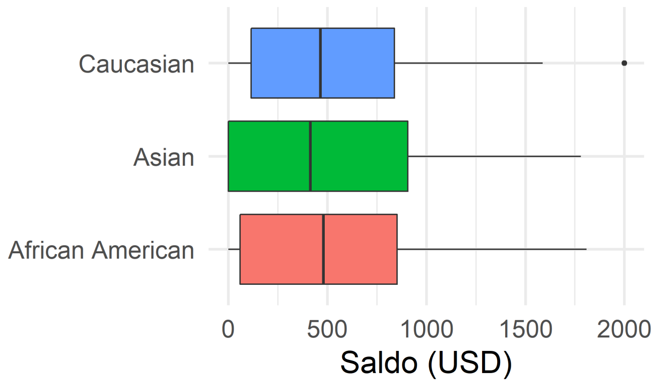
$$y_i = \beta_0 + \beta_1 x_i \quad \text{em que} \quad x_i = \begin{cases} 1 & \text{se a i-ésima pessoa for } \texttt{female} \\ 0 & \text{se a i-ésima pessoa for } \texttt{male} \end{cases}$$

```
# lembrete: exercícios 8, 9 e 10 do script!
```

Ver ISL página 84 (Predictors with Only Two Levels).

Preditores Categóricos

Preditor com 3 ou mais categorias




```
summary(lm(Balance ~ Ethnicity,
#                      Estimate  S
# (Intercept)            531.00  4
# EthnicityAsian         -18.69  6
# EthnicityCaucasian     -12.50  5
```

Modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Em que

$$x_{1i} = \begin{cases} 1 & \text{se for } \texttt{Asian} \\ 0 & \text{caso contrário} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{se for } \texttt{Caucasian} \\ 0 & \text{caso contrário} \end{cases}$$

Preditores Categóricos

Preditor com 3 ou mais categorias

"One hot enconding" ou "Dummies" ou "Indicadores".

Ethnicity	(Intercept)	EthnicityAsian	EthnicityCaucasian
Caucasian	1	0	1
Asian	1	1	0
Asian	1	1	0
Asian	1	1	0
Caucasian	1	0	1
Caucasian	1	0	1
African American	1	0	0
Asian	1	1	0

Preditores Categóricos

Preditor com 3 ou mais categorias

Interpretação dos parâmetros:

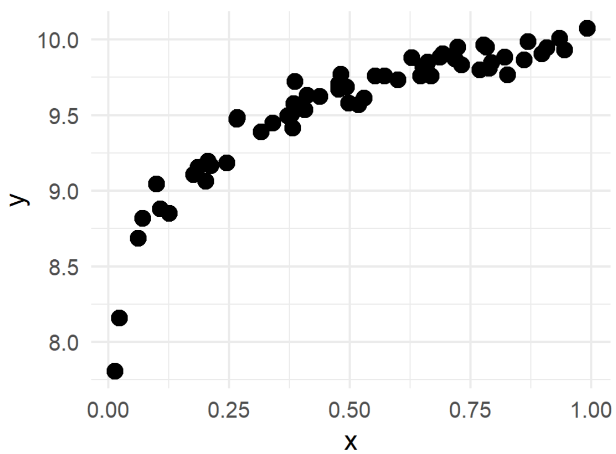
$$y_i = \begin{cases} \beta_0 & \text{se for } \texttt{Afro American} \\ \beta_0 + \beta_1 & \text{se for } \texttt{Asian} \\ \beta_0 + \beta_2 & \text{se for } \texttt{Caucasian} \end{cases}$$

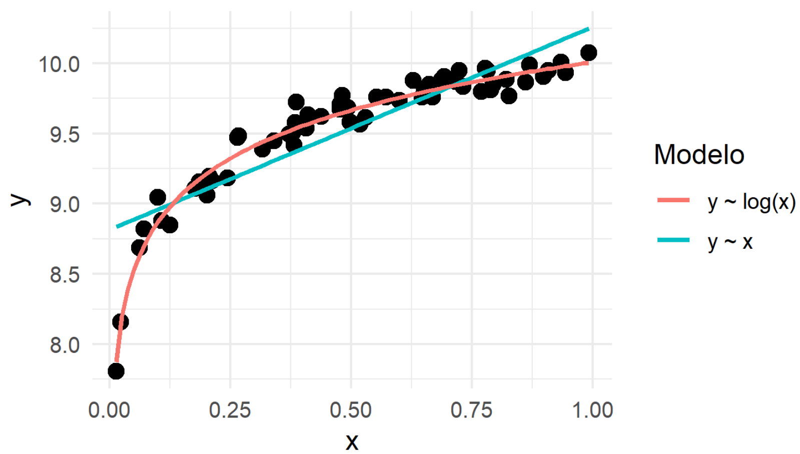
```
# interprete cada um dos três parâmetros individualmente.
# lembrete: exercício 11 do script!
```

Transformações Não Lineares dos Preditores

Exemplo: log

Modelo real: $y = 10 + 0.5 log(x)$ Modelo proposto: $y = \beta_0 + \beta_1 log(x)$





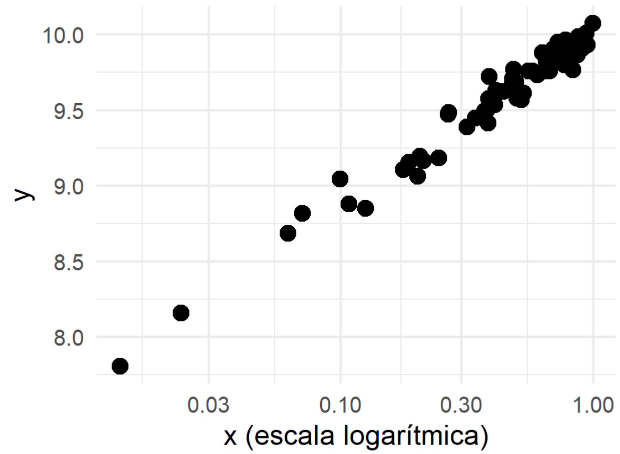
Outras transformações comuns: raíz quadrada, polinômios, Box-Cox, ...

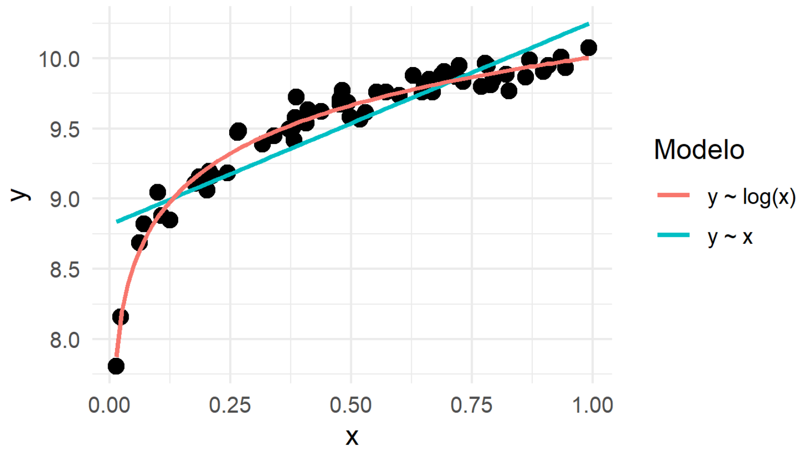
```
# lembrete: exercício 13 do script!
```

Transformações Não Lineares dos Preditores

Exemplo: log

Modelo real: $y = 10 + 0.5log(x)$ Modelo proposto: $y = \beta_0 + \beta_1 log(x)$



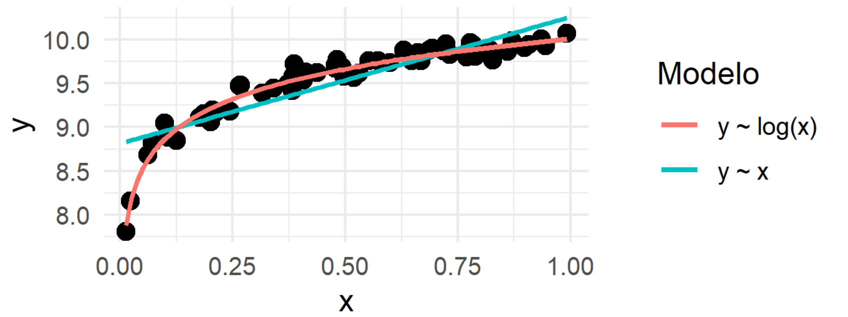


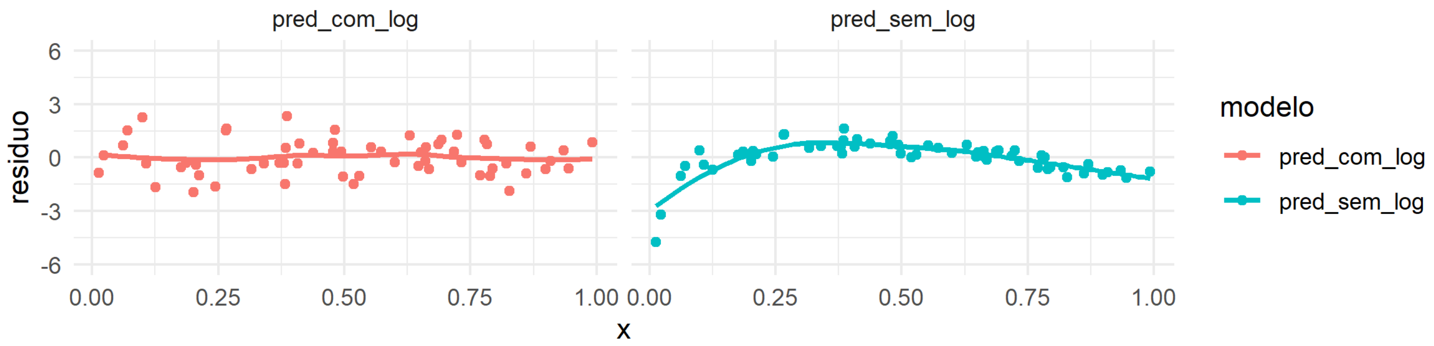
Outras transformações comuns: raíz quadrada, polinômios, Box-Cox, ...

```
# lembrete: exercício 13 do script!
```

Transformações Não Lineares dos Preditores

Gráfico de Resíduos

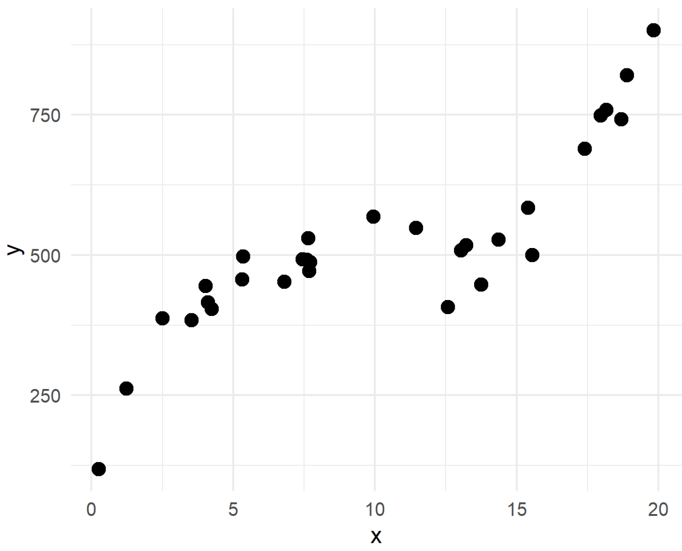




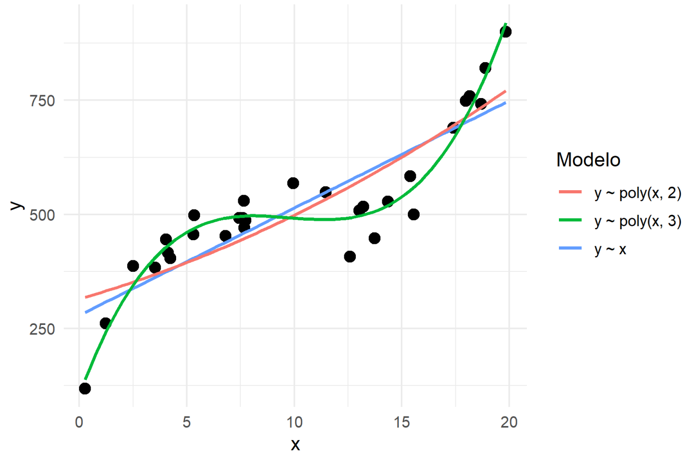
Transformações Não Lineares dos Preditores

Exemplo: Regressão Polinomial

Modelo real:
$y = 500 + 0.4(x - 10)^3$



Modelo proposto:
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

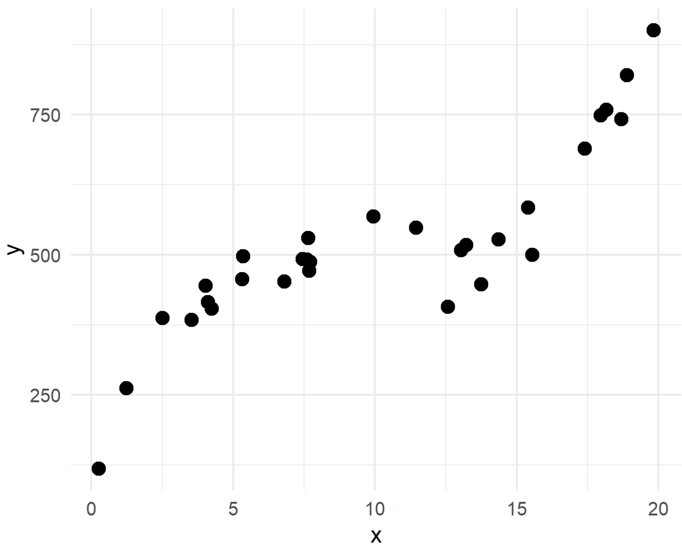



```
# lembrete: exercício 14 do script!
```

Transformações Não Lineares dos Preditores

Exemplo: Regressão Polinomial

Modelo real:
$y = 500 + 0.4(x - 10)^3$



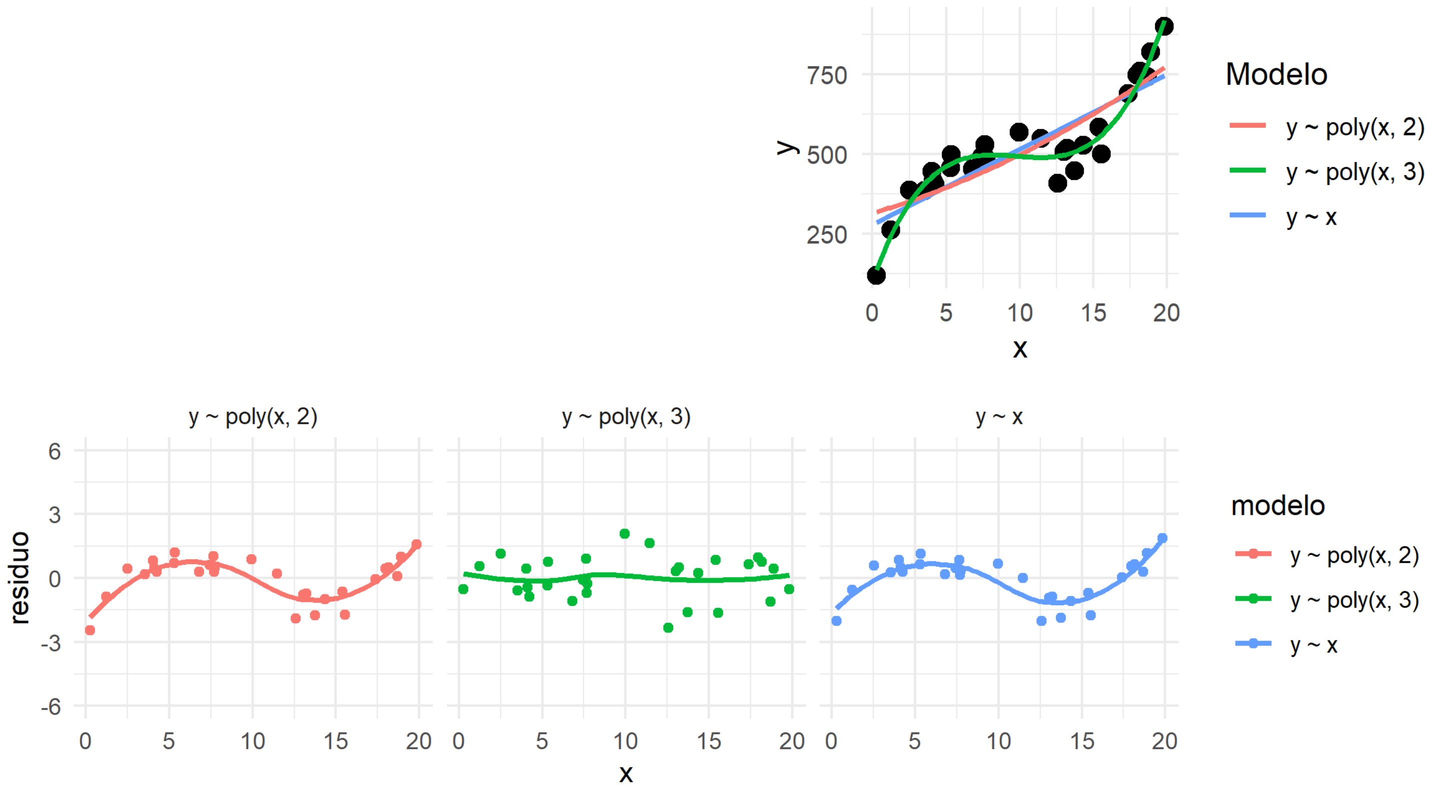
Modelo proposto:
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

y	x	x2	x3
456.5	5.3	28.2	149.7
492.5	7.4	55.4	412.2
548.4	11.5	131.3	1503.9
758.7	18.2	329.9	5993.0
444.7	4.0	16.3	65.6
748.3	18.0	322.8	5800.8
820.5	18.9	357.0	6744.3

```
# lembrete: exercício 13 do script!
```

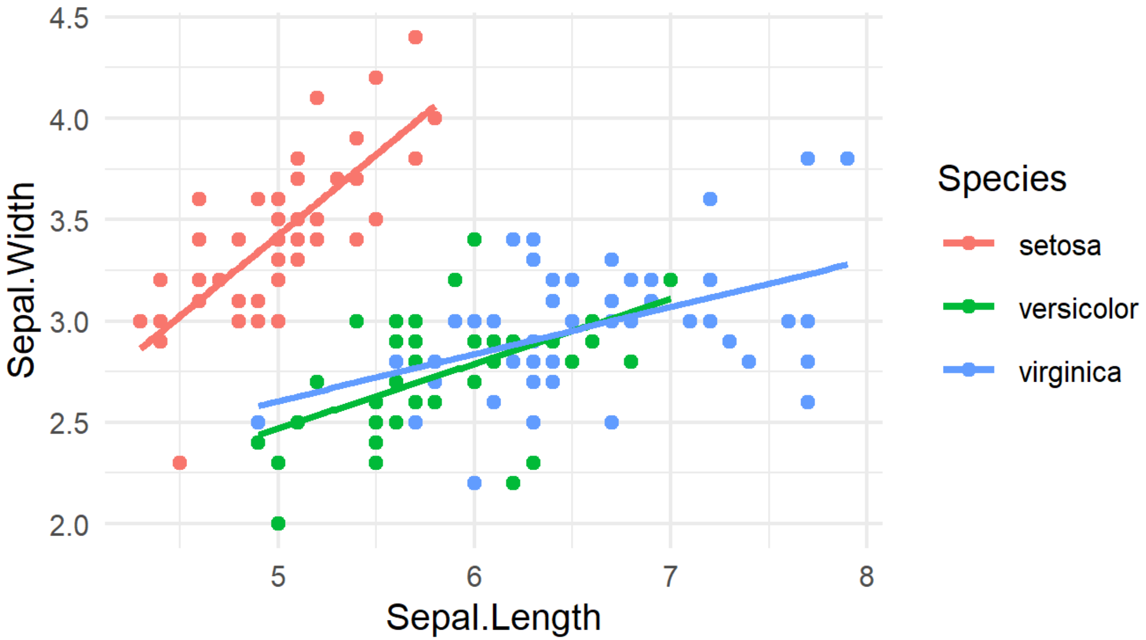
Transformações Não Lineares dos Preditores

Gráfico de Resíduos



Interações

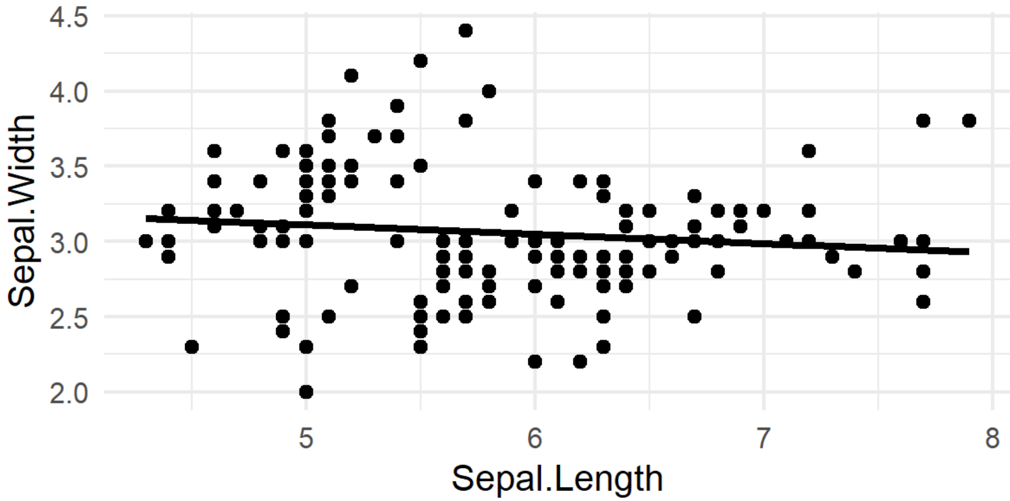
Interação entre duas variáveis explicativas: `Species` e `Sepal.Length`



Interações

Modelo proposto (Matemático): Seja `y = Sepal.Width` e `x = Sepal.Length`,

$$y = \beta_0 + \beta_1 x$$



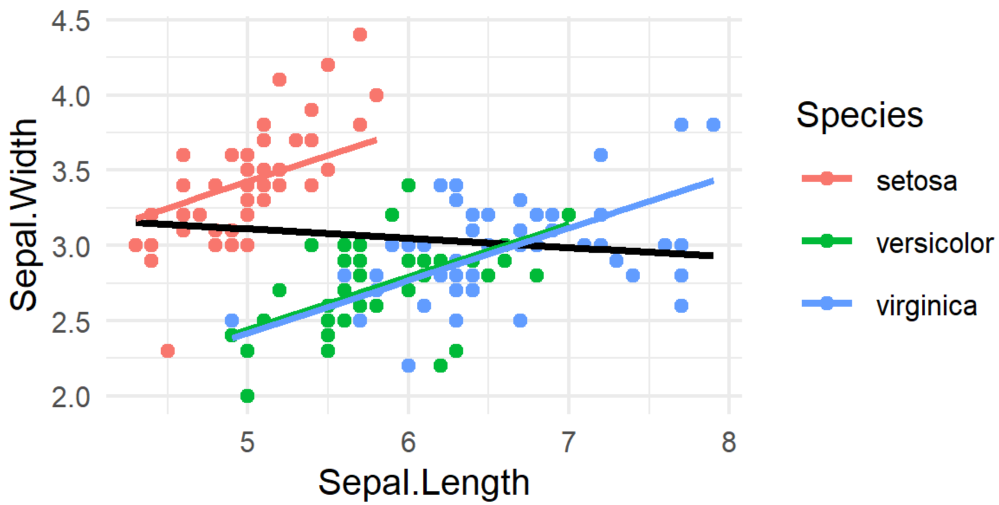
Modelo proposto (em R): `Sepal.Width ~ Sepal.Length`

```
# lembrete: exercícios 14 ao 17 do script!
```

Interações

Modelo proposto (Matemático): Seja `y = Sepal.Width` e `x = Sepal.Length`,

$$y = \beta_0 + \beta_1 x + \beta_2 I_{versicolor} + \beta_3 I_{virginica}$$



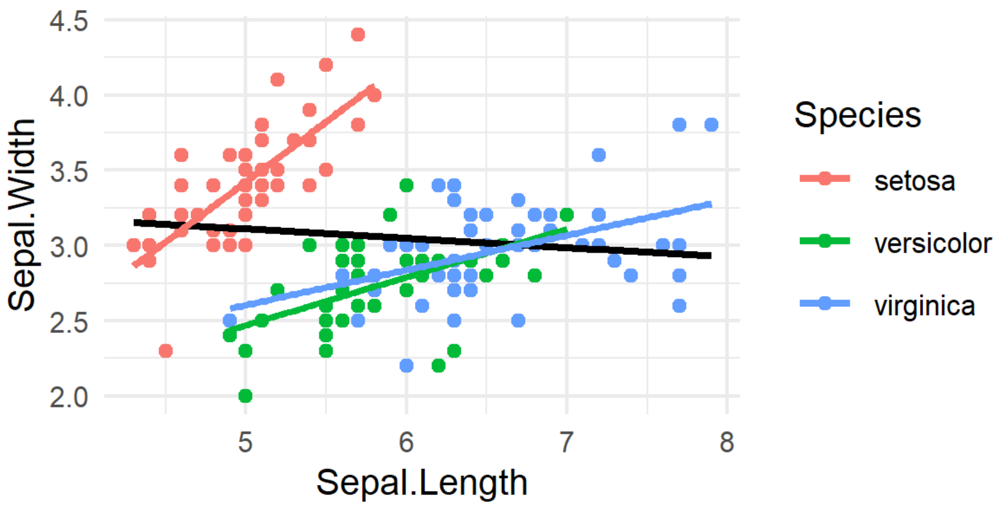
Modelo proposto (em R): `Sepal.Width ~ Sepal.Length + Species`

```
# lembrete: exercícios 14 ao 17 do script!
```

Interações

Modelo proposto (Matemático): Seja `y = Sepal.Width` e `x = Sepal.Length`,

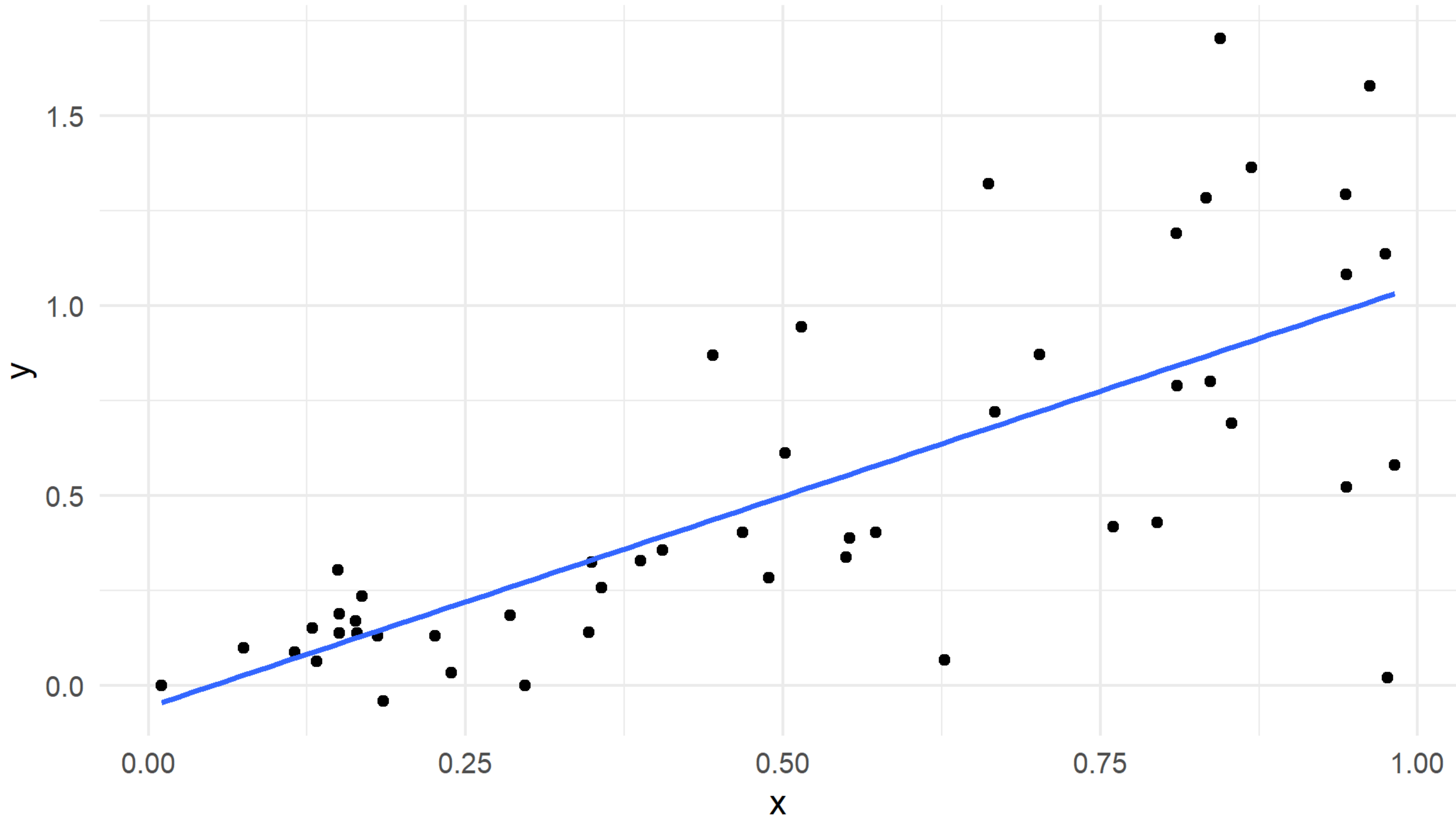
$$y = \beta_0 + \beta_1 x + \beta_2 I_{versicolor} + \beta_3 I_{virginica} + \beta_4 x I_{versicolor} + \beta_5 x I_{virginica}$$



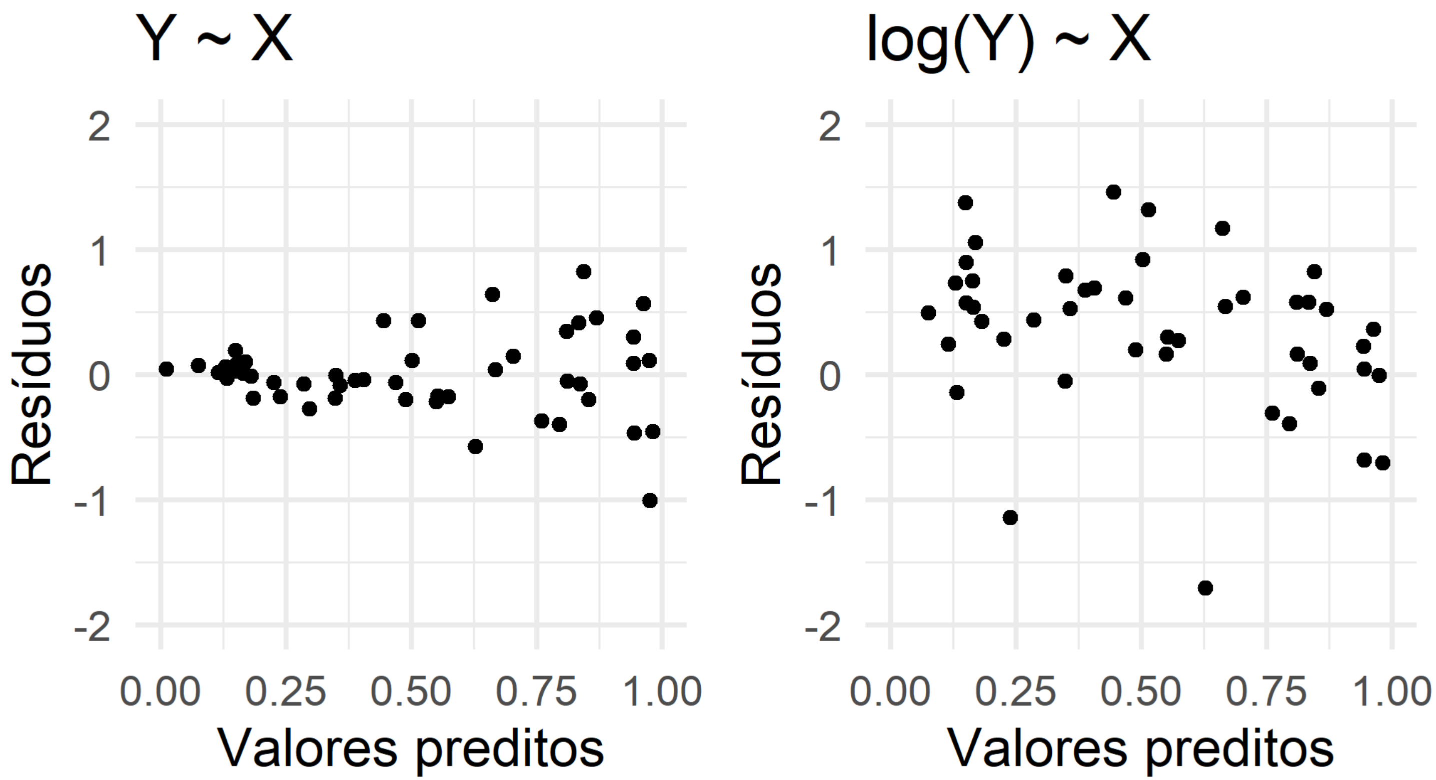
Modelo proposto (em R): `Sepal.Width ~ Sepal.Length * Species`

```
# lembrete: exercícios 14 ao 17 do script!
```

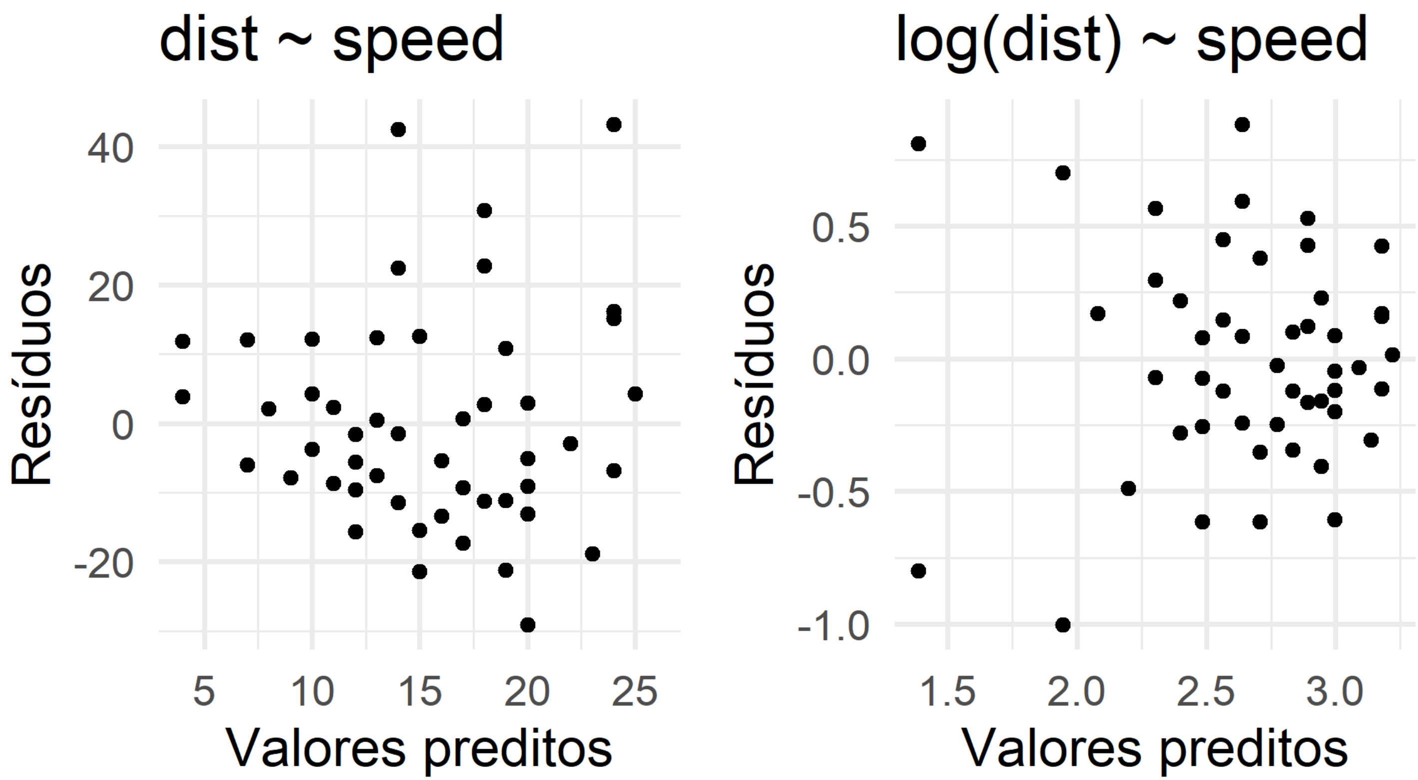
Heterocedasticidade



Heterocedasticidade



Heterocedasticidade



Heterocedasticidade

Problema

- O estimador $\hat{\sigma}_{\beta_1} = \sqrt{\frac{EQM}{\sum(x_i - \bar{x})^2}}$ deixa de ter as melhores propriedades. Poderíamos ter conclusões estranhas para β_1.

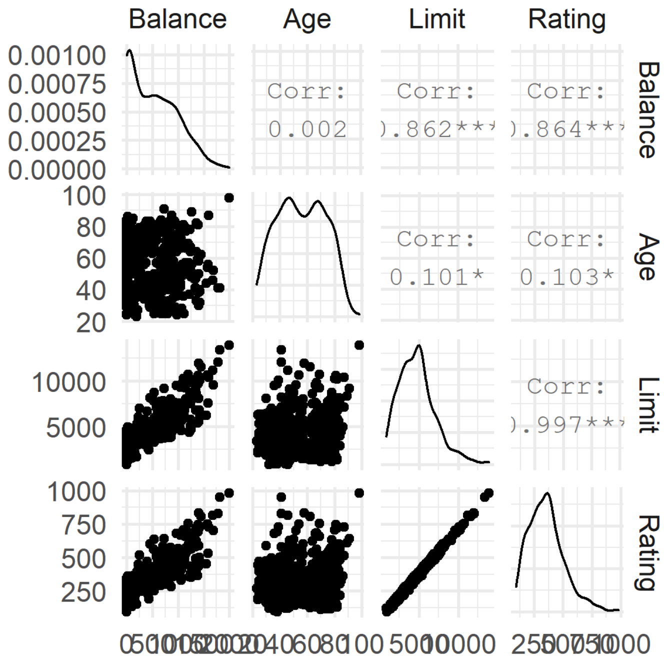
Diagnóstico

- Visualização dos resíduos;

Tratamentos

- Transformações na variável resposta. log(y), sqrt(y), 1/y, etc;

Multicolinearidade



Modelo 1: sem colineares

term	estimate	std.error	statistic	p.value
(Intercept)	-173.41	43.83	-3.96	0
Limit	0.17	0.01	34.50	0
Age	-2.29	0.67	-3.41	0

Modelo 2: com colineares

term	estimate	std.error	statistic	p.value
(Intercept)	-377.54	45.25	-8.34	0.00
Limit	0.02	0.06	0.38	0.70
Rating	2.20	0.95	2.31	0.02

Problema: Instabilidade numérica, desvios padrão inflados e interpretação comprometida.

Soluções: eliminar uma das variáveis muito correlacionadas ou Consultar o VIF (Variance Inflation Factor)

Multicolinearidade

VIF (Variance Inflation Factor)

Detecta preditores que são combinações lineares de outros preditores.

Procedimento: Para cada preditor X_j,

1) Ajusta regressão linear com as demais: `lm(X_j ~ X_1 + ... + X_p)`.

2) Calcula-se o R-quadrado dessa regressão e aplica a fórmula abaixo

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

3) Remova o preditor se VIF maior que 5 (regra de bolso).

```
# lembrete: exercícios 18 do script!
```

Ver ISL página 101.

Fazendo Predições

Modelo

```
modelo_iris <- lm(Sepal.Width ~ Sepal.Length * Species, data = iris)
```

Suponha que dados novos chegaram:

```
dados_novos <- tibble(Sepal.Length = 10, Species = "setosa")
```

Utilizando a função `augment()` do pacote `broom`

```
augment(modelo_iris, newdata = dados_novos)
##   Sepal.Length Species .fitted .se.fit
## 1           10 setosa     7.42   0.553
```

O valor estimado de `Sepal.Width` foi de `7.42` +/- `0.55`.

Fazendo Predições

Modelo

```
modelo_iris <- lm(Sepal.Width ~ Sepal.Length * Species, data = iris)
```

Suponha que dados novos chegaram:

```
dados_novos <- data.frame(Sepal.Length = 10, Species = "setosa")
```

Utilizando a função `predict()`

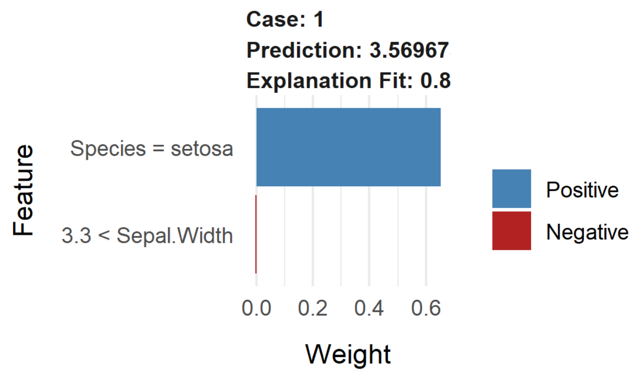
```
dados_novos %>%
  mutate(S.W.Est = predict(modelo_iris, newdata = dados_novos))
##   Sepal.Length Species S.W.Est
##             10 setosa     7.42
```

Interpretabilidade da Predição

LIME: Local Interpretable Model-agnostic Explanations

```
# fazendo lm com caret::train() pq o lime soh aceita caret
modelo_iris <- train(Sepal.Width ~ Sepal.Length * Species, data = ir
explicador <- lime(iris, modelo_iris)
explicacoes <- lime::explain(dados_novos, explicador, n_features = 2)
plot_features(explicacoes)
```

```
# lembrete: exercícios 19
# do script!
```



Ver LIME for R página 101.

(Opcional) Abordagem Probabilística

Do ponto de vista probabilístico, modela-se o problema como uma amostra de N indivíduos, todos independentes entre si e com distribuição Normal.

$$Y_i|x_i \sim N(\mu_i, \sigma^2),\ i = 1, \dots, N$$

E então, supõem que a média de Y dado o valor de x seja linear:

$$\mu = E[Y|x] = \beta_0 + \beta_1 x$$

Assim, gostaríamos de achar β_0 e β_1 que fizessem dessa amostra a mais verossímil possível.

Daí entra o conceito de verossimilhança, que é a probabilidade conjunta dos dados acontecerem:

$$P(Y_1, \dots, Y_N|x) \overset{\text{indep}}{=} P(Y_1|x_1)P(Y_2|x_2)\dots P(Y_N|x_N)$$

continua...

(Opcional) Abordagem Probabilística

Se tirarmos o $logarítmo$ dessa probabilidade conjunta, teremos:

$$logP(Y_1, \ldots, Y_N|x) \overset{\text{indep}}{=} logP(Y_i|x_1) + logP(Y_2|x_2) + \cdots + logP(Y_N|x_N)$$

Que podemos escrever de forma mais sussinta usando um somatório:

$$\ell = logP(Y_1, \ldots, Y_N|x) = \sum_{i=1}^{N} logP(Y_i|x_i)$$

Essa expressão que chamamos de ℓ é conhecida como log-verossimilhança (log-likelihood no inglês).

continua…

(Opcional) Abordagem Probabilística

Já que assumimos que $Y_i|x_i$ segue uma distribuição $N(\beta_0 + \beta_1 x_i, \sigma^2)$, temos que:

$$\ell = \sum_{i=1}^{N} log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y_i - \mu_i)^2 \right] \right)$$

Que depois de simplificar (e deixando as constantes de fora), fica

$$\ell = -\frac{1}{N} \sum (y_i - \mu_i)^2 = -\frac{1}{N} \sum (y_i - {\color{red}(\beta_0 + \beta_1 speed)})^2 = -EQM$$

Ou seja, maximizar a verossimilhança é equivalente a minimizar o EQM como vínhamos fazendo.

Questões importantes

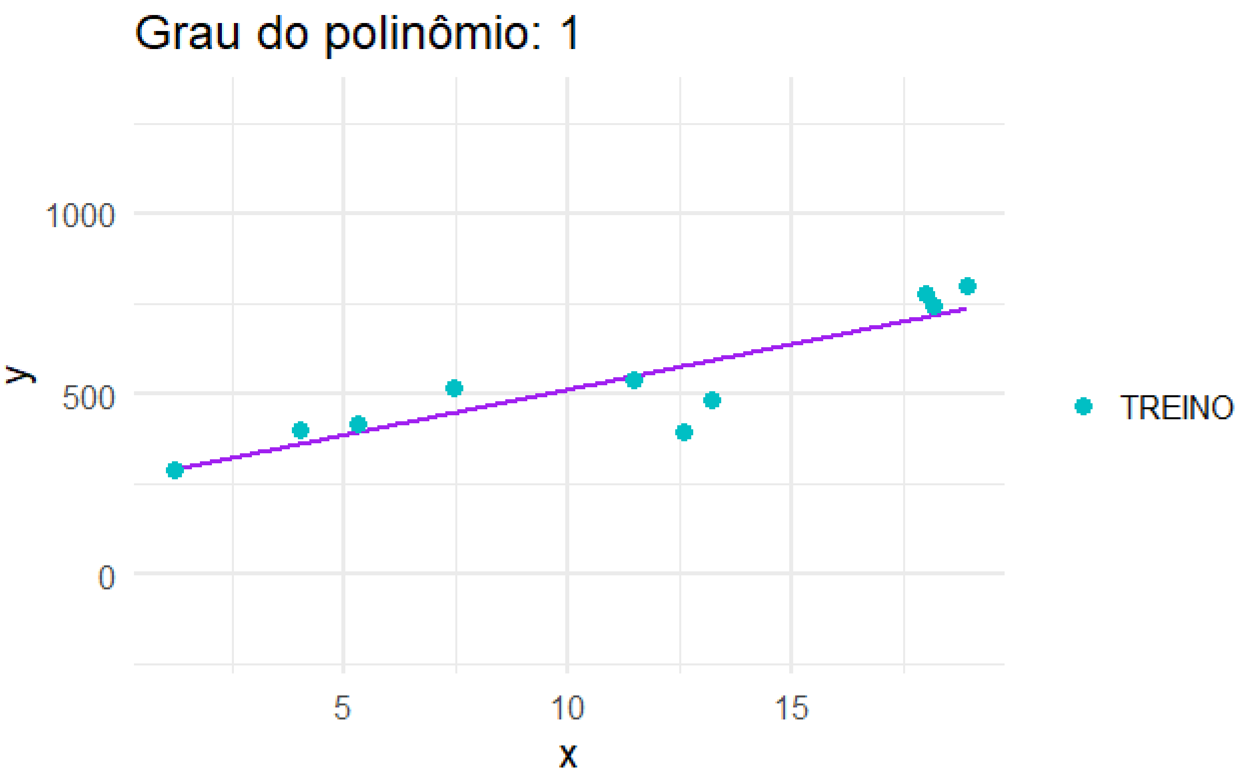
Questões que usualmente estamos interessados quando ajustamos uma regressão linear.

- Pelo menos um dos preditores $X1, X2, \ldots, X_p$ é útil para prever/explicar?
- Todos os preditores são úteis ou apenas um subconjunto deles que é?
- O quão bem o modelo se ajusta aos dados?

Ver ISL página 75 (Some Important Questions).

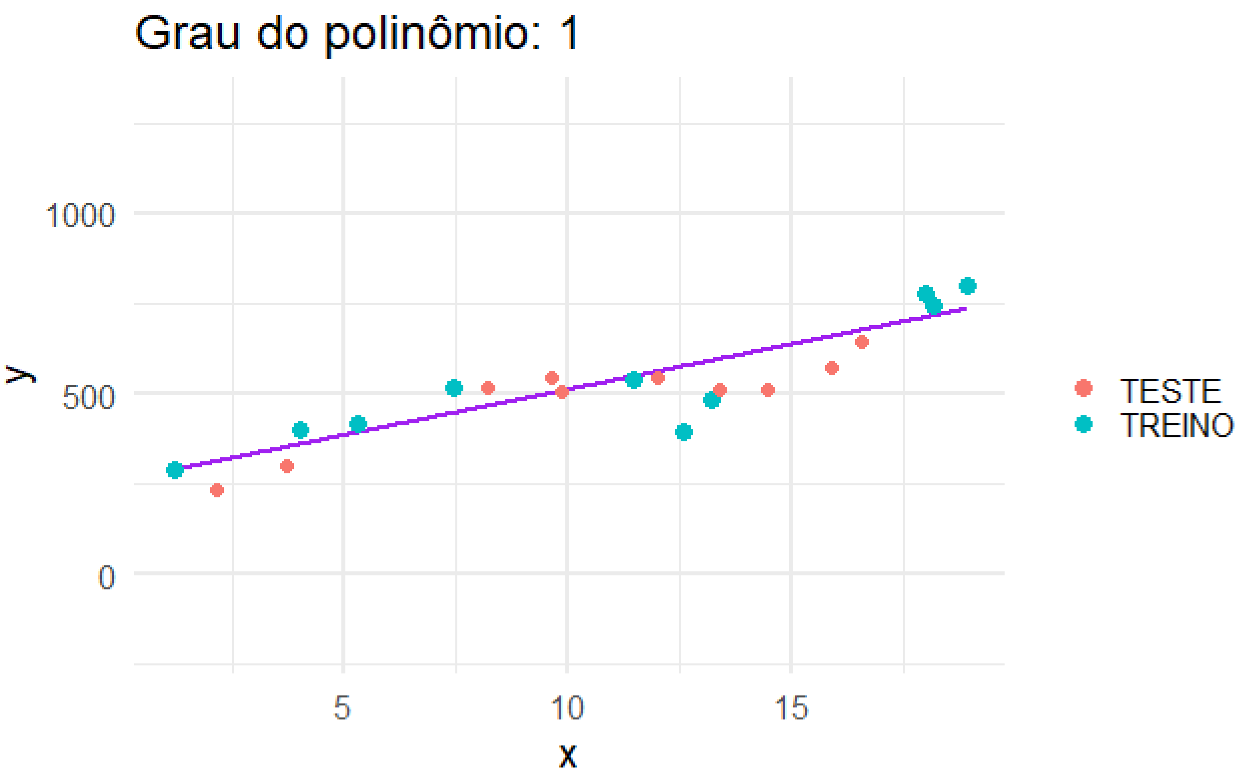
Sobreajuste (overfitting)

Modelo real é de **grau 3**



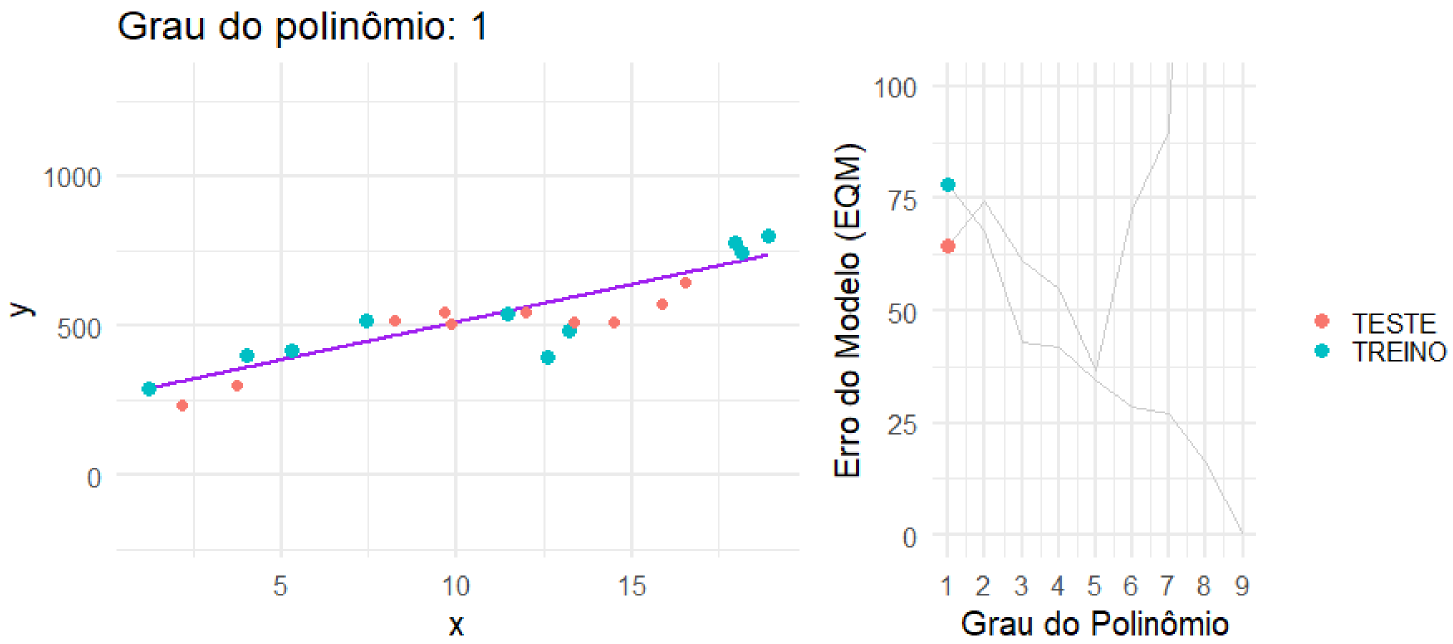
Sobreajuste (overfitting)

Modelo real é de **grau 3**



Sobreajuste (overfitting)

Modelo real é de **grau 3**



Ver ISL página 61 (Simple Linear Regression).

Regularização

Relembrando o nossa **função de custo** EQM.

$$EQM = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}))^2$$

Regularizar é "não deixar os $\beta's$ soltos demais".

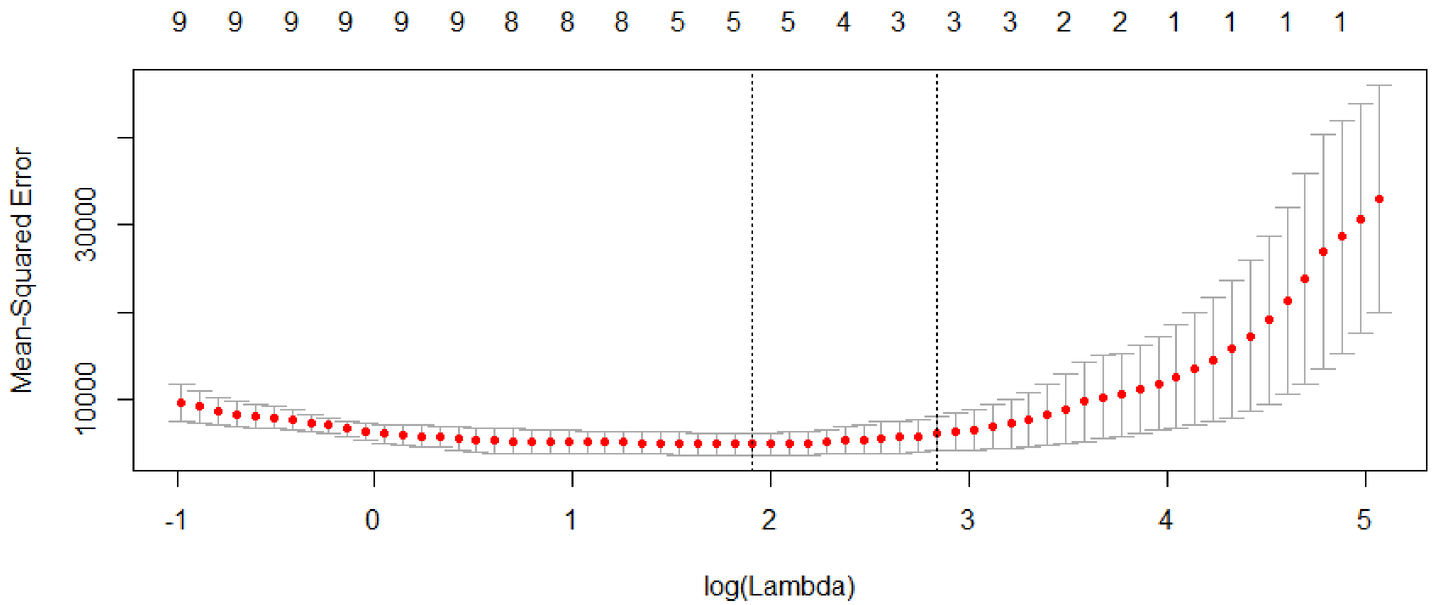
$$EQM_{regularizado} = EQM + \lambda \sum_{j=1}^{p} |\beta_j|$$

Ou seja, **penalizamos** a função de custo se os $\beta's$ forem muito grandes.

Ver ISL página 203 (Linear Model Selection and Regularization).

LASSO e Seleção de Preditores

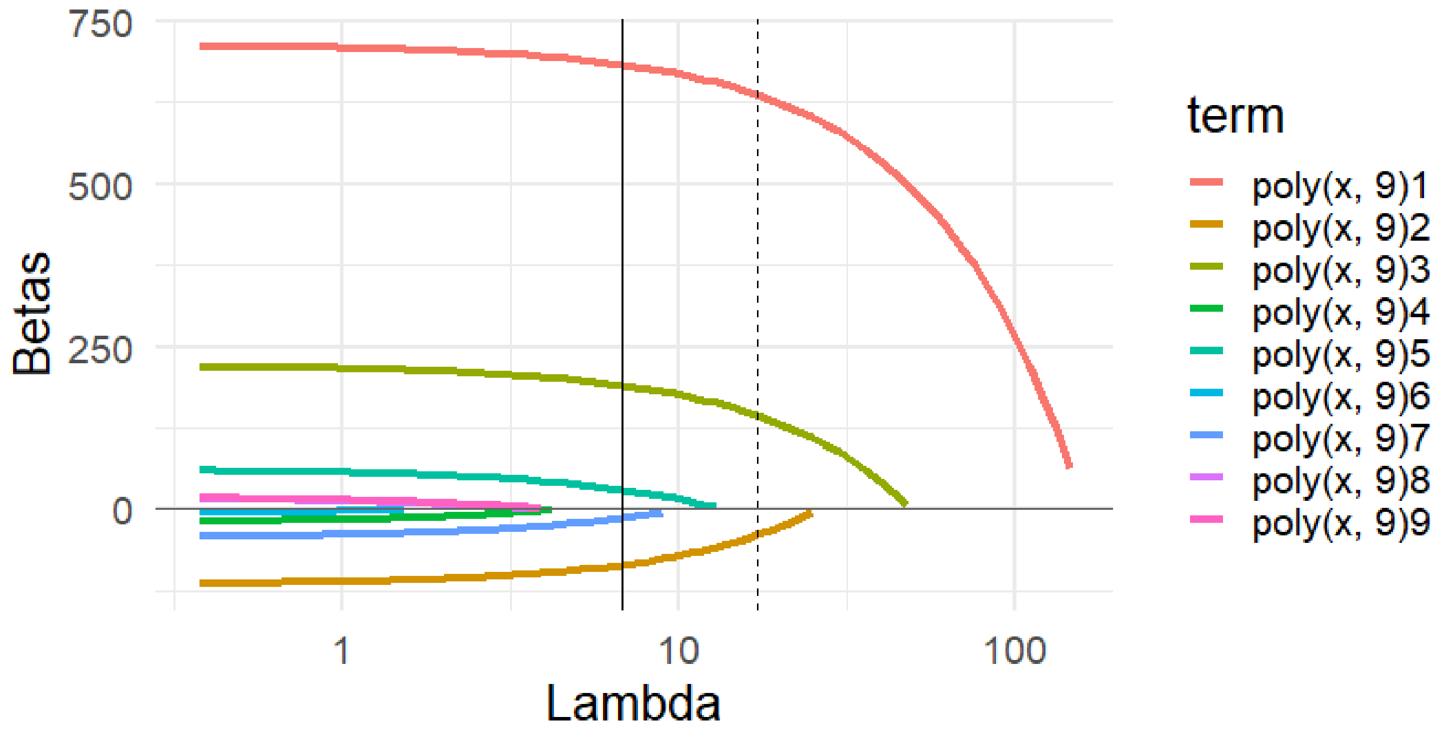
Conforme aumentamos o λ, forçamos os $\beta's$ a serem cada vez menores.



Ver ISL página 219 (The LASSO).

LASSO e Seleção de Preditores

Conforme aumentamos o λ, forçamos os $\beta's$ a serem cada vez menores.



Ver ISL página 219 (The LASSO).

Validação Cruzada

```
## #  5-fold cross-validation
## # A tibble: 5 x 6
##   splits          id    n_treino n_teste regressao eqm_teste
##   <list>          <chr>    <dbl>   <dbl> <list>        <dbl>
## 1 <split [40/10]> Fold1       40      10 <lm>           17.0
## 2 <split [40/10]> Fold2       40      10 <lm>           13.0
## 3 <split [40/10]> Fold3       40      10 <lm>           12.2
## 4 <split [40/10]> Fold4       40      10 <lm>           15.7
## 5 <split [40/10]> Fold5       40      10 <lm>           19.6
```

```
## [1] 15.494
```

ERRO DE VALIDAÇÃO CRUZADA:

$$EQM_{cv} = \frac{1}{10}\sum_{i=1}^{10} EQM_{Fold_i} = 14,671$$

Validação Cruzada

fold	speed	dist
fold 1	4	2
fold 1	4	10
fold 2	7	4
fold 2	7	22
fold 3	8	16
fold 3	9	10
fold 4	10	18
fold 4	10	26
fold 5	10	34
fold 5	11	17

Limitações da Regressão Linear

- Variável resposta Não Normal
- Variável resposta Positiva
- Variável resposta Categórica
- Relação funcional não linear entre X e Y
- Muitas interações para testar entre as preditoras