

Aprendizado supervisionado 1

Especialização em Ciência de dados, IM/UFRJ
Thaís C O Fonseca
DME/UFRJ

Ementa

- Introdução ao modelo linear geral;
Regressão múltipla;
- Regressão na família exponencial;
Dados binários e regressão logística;
- Regressão com penalização L1 e L2;
- Regressão por splines.



Fonte: Bayesian Reasoning and Machine Learning, David Barber

Bibliografia sugerida

- The Elements of Statistical Learning, T. Hastie, R. Tibshirani, J.H. Friedman. Springer, 2001.
- Bayesian reasoning and Machine Learning, D. Barber.
- R4DS: R for Data Science, G. Grolemund and H. Wickham
- Information Theory, Inference, and Learning Algorithms, MacKay, 2005.

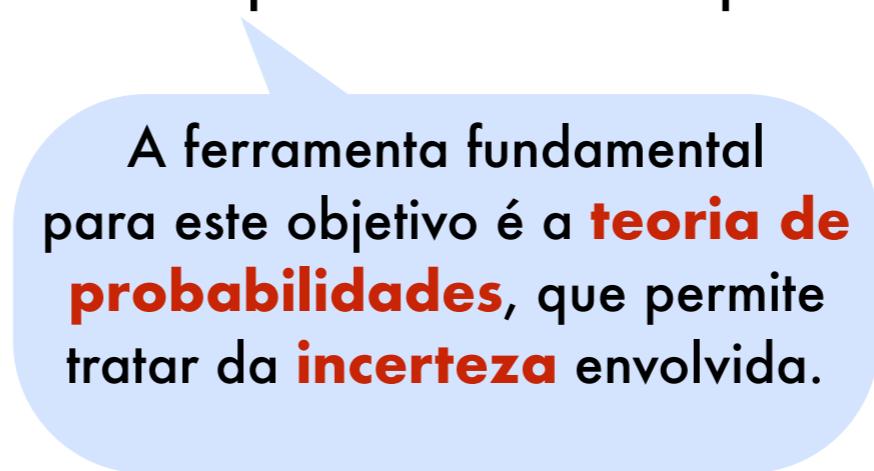
Introdução: aprendizado supervisionado

Dados

- Nos tempos modernos, dados estão disponíveis em grande quantidade e em diversos níveis de resolução.
- Alguns exemplos englobam bioinformática, astronomia, física, monitoramento ambiental, comércio, transações financeiras, reconhecimento de imagens.
- Objetivo geral da análise de dados: saber **processar** e **extrair informação de valor** dos dados.

O que é aprendizado de máquinas

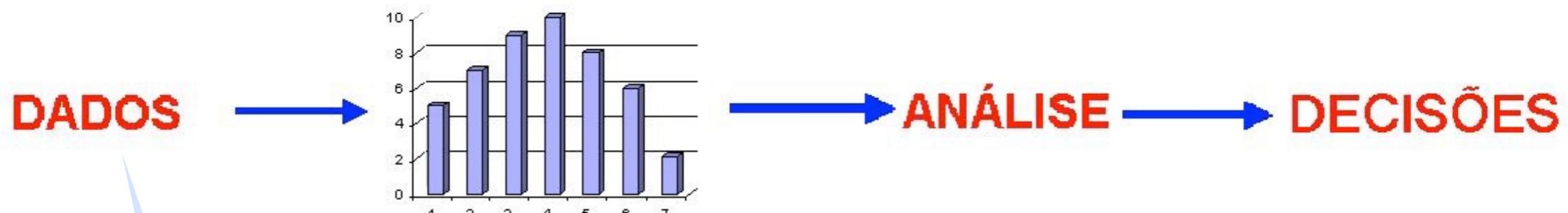
- É o estudo de métodos baseados em **processamento de dados**.
- Estes métodos devem ser capazes de imitar e entender processos e ajudar na tomada de decisões em diversas áreas. *Fonte: Bayesian reasoning and machine learning, Barber.*
- Questões importantes para atingir estes objetivos:
 - Como comprimir e processar dados de alta dimensão?
 - Como fazer previsões de um processo complexo, por exemplo, em tempos futuros?



A ferramenta fundamental para este objetivo é a **teoria de probabilidades**, que permite tratar da **incerteza** envolvida.

O que é estatística

- É uma ciência que se dedica à coleta, análise e interpretação de dados.
- Pode ser pensada como a ciência de aprendizagem de dados.



Note que dados podem ser vistos apenas como números se não olhamos o contexto.

De forma simplificada pode ser dividida em 3 áreas: **análise descritiva, probabilidade e inferência**.

Aprendendo com dados

Dados

Aprendendo
com dados



X e Y
disponíveis

Somente X
disponível

Aprendizado
supervisionado

- Regressão
- Classificação
- Árvores e florestas

Aprendizado não
supervisionado

- Análise de cluster
- Componentes principais

Aprendizado supervisionado

- Definição:**
- (i) Para dados $D = \{(x_i, y_i), i = 1, \dots, N\}$ o objetivo é aprender a relação entre o input x e o output y tal que, para novos dados x_0 a predição de y_0 seja acurada;
 - (ii) O par $(x_0, y_0) \notin D$ mas assume-se que foram obtidos do mesmo processo gerador;
 - (iii) Para especificar acurácia de forma mais explícita define-se uma função de perda $L(y^{pred}, y^{true})$ ou, a utilidade $U = -L$.

Aprendizado supervisionado

- Exemplo: Considere um conjunto de dados de 10000 faces humanas representadas por um vetor \mathbf{x} (informação para os pixels da imagem). Para essas 10000 faces temos o output homem ou mulher.
- Objetivo: prever o gênero de uma nova face para a qual temos \mathbf{x}_0 mas não temos o output.



Dado de treinamento x dado de teste

- Em muitos contextos, utilizar todos os dados disponíveis para ajuste de um modelo pode levar a previsões ruins fora da amostra.
- Com objetivo de contornar esse problema, usualmente consideramos a divisão dos dados em duas partes.

Dado de treinamento

Usado apenas para ajustar o modelo assumido para o problema.

Dado de teste

Usado apenas para verificar a performance preditiva do modelo treinado.

Exemplo:

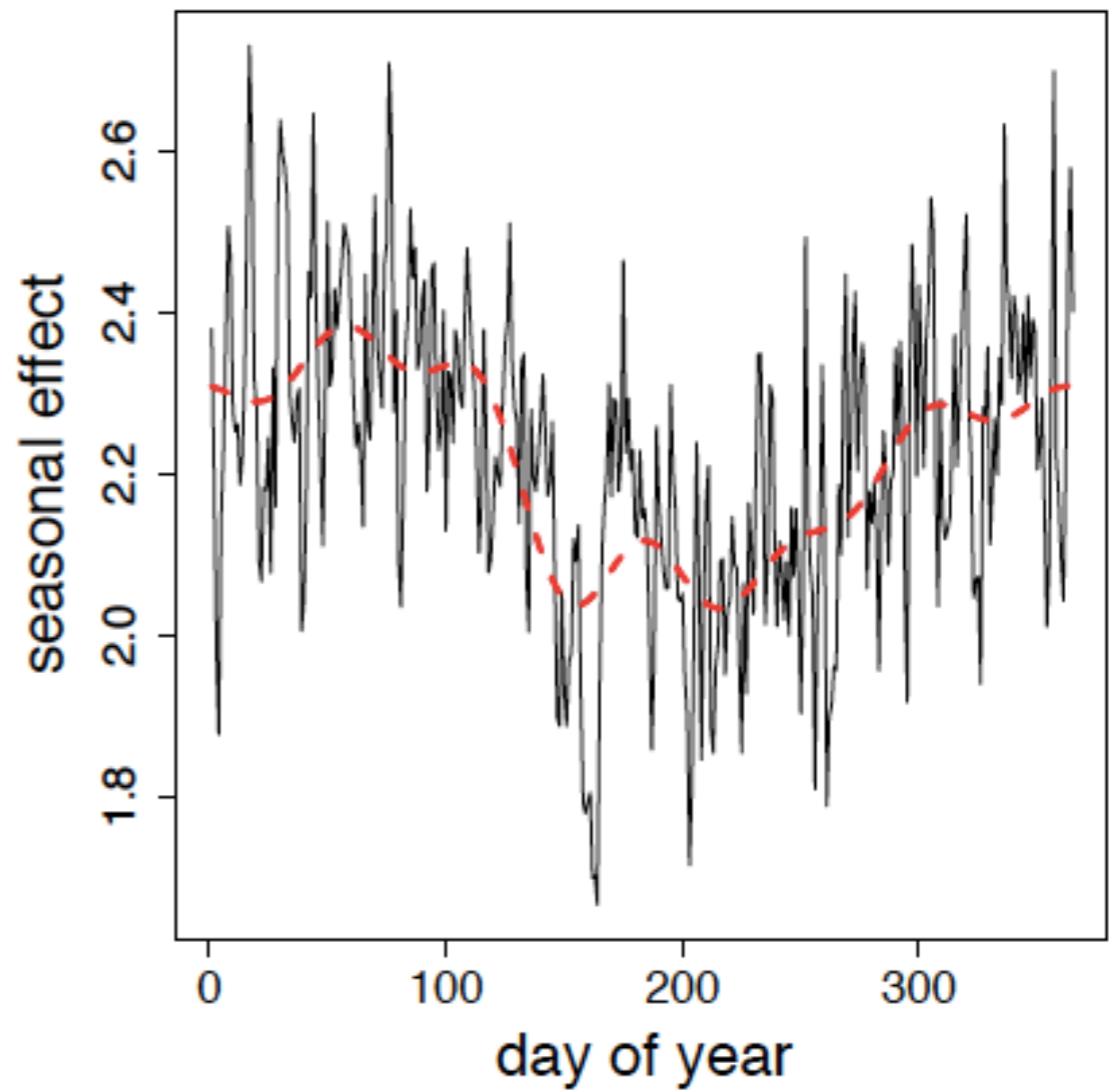
- Um pai decide explicar ao filho o que é um carro esporte.
- Considerando difícil explicar com palavras, o pai leva o filho para uma ponte próxima a uma rodovia e para cada carro esporte que ele vê ele grita para o filho: veja um carro esporte!
- Após 10 minutos o pai pergunta ao filho se ele entendeu o que é um carro esporte.
- O filho responde: claro!!
- Então passa um fusca vermelho e o filho grita: um carro esporte!
- E o pai surpreso pergunta: por que você acha isso?
- E o filho conclui: porque todos os carros esporte são vermelhos!

Sobre aprendizado supervisionado

- Esse exemplo mostra como um modelo pode ter uma boa performance na amostra de treinamento e ser péssimo para previsão.
- Este exemplo também mostra claramente o que é um processo de aprendizado supervisionado. Se o filho já soubesse o que é um carro esporte não seria necessário um “treinamento”.

Exemplo: série temporal

- Exemplo: série temporal diária de velocidade do vento e modelo ajustado.
- Temos inputs $t=1,\dots,N$ e outputs y_1,\dots,y_N .
- Isto é, estimativa e previsão em séries temporais representam um tipo de aprendizado supervisionado.
- A previsão em tempos futuros usando o modelo ajustado para os dados de treinamento (t, y_t) .



Classificação e regressão

Problema de classificação

Quando o output é uma variável categórica chamamos o output de label e o problema de aprendizado supervisionado é dito problema de classificação.

Problema de regressão

Quando o output é quantitativo usualmente dizemos que o aprendizado supervisionado é um problema de regressão.

- Observação: podemos transformar uma variável quantitativa em categórica e tratar o problema de regressão como um problema de classificação.
- Ex: considere interesse em prever o preço futuro de um produto em um supermercado. Este pode ser visto como um problema de regressão. Porém, se os preços são categorizados em 3 categorias passamos a ter um problema de classificação.

Aprendizado não supervisionado

- Definição:** (i) Para dados $D = \{x_i, i = 1, \dots, N\}$ o objetivo é descrevê-los de forma parcimoniosa;
- (ii) Uma função objetivo é definida para quantificar a acurácia da descrição;
- (iii) Do ponto de vista probabilístico, estamos interessados na distribuição $p(\mathbf{x})$.

Exemplo: Um supermercado deseja dividir os consumidores em perfis de consumo. Para isso coleta dados de 1000 check-outs, onde consta uma matriz (esparsa) para cada consumidor com itens consumidos ou não consumidos para 10000 produtos do supermercado.

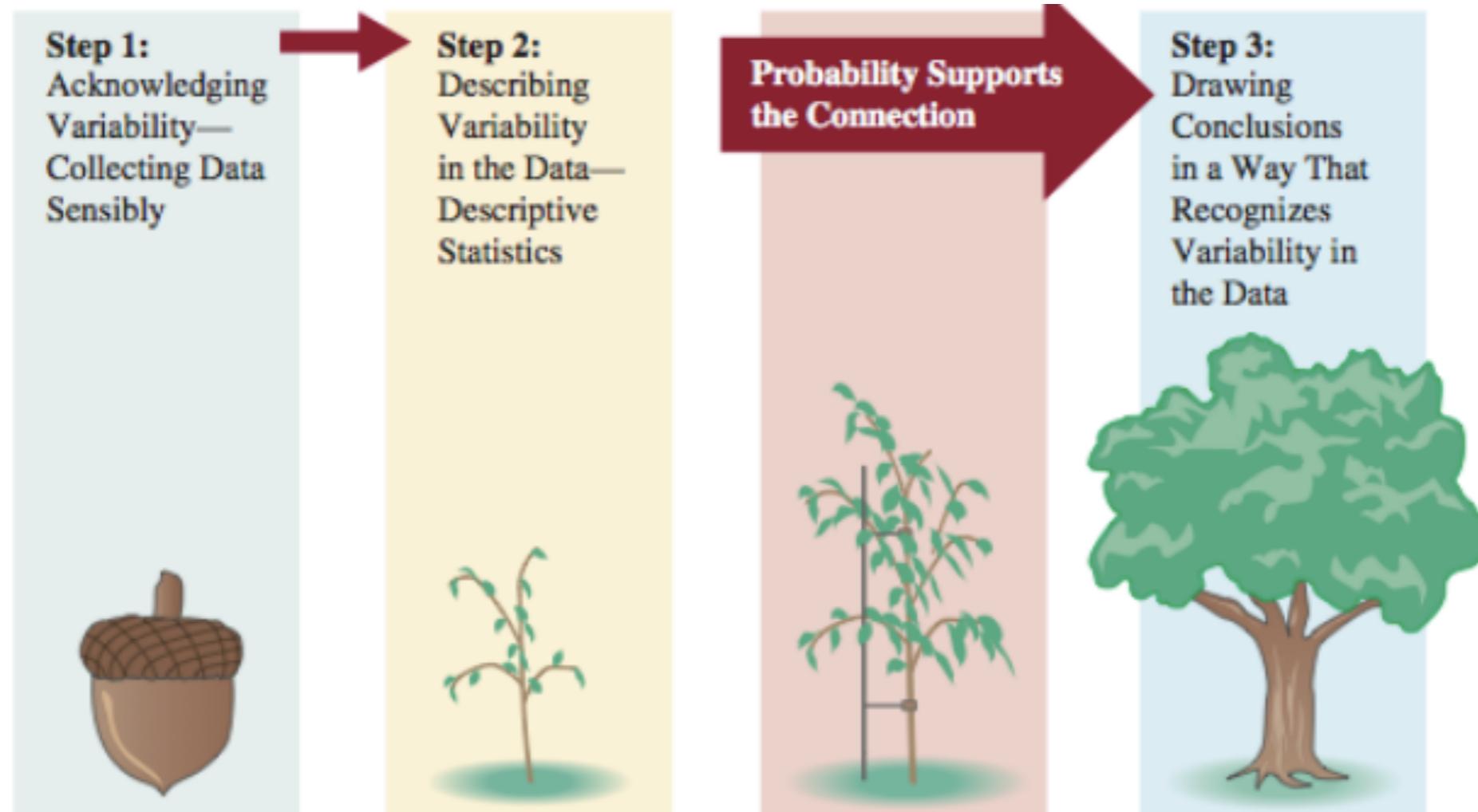
coffee	1	0	0	1	0	0	0	...
tea	0	0	1	0	0	0	0	...
milk	1	0	1	1	0	1	1	...
beer	0	0	0	1	1	0	1	...
diapers	0	0	1	0	1	0	1	...
aspirin	0	1	0	0	1	0	1	...

Modelos: contexto de estatística e machine learning

Dados e distribuições de probabilidade

- Tipicamente a distribuição dos dados não é completamente conhecida.
- Os dados devem ser usados para aprender sobre essas distribuições.
- Devemos considerar diversas distribuições de probabilidade que podem ser usadas em cada problema.
- E estimar os parâmetros desconhecidos que parametrizam estas distribuições.

Resumindo



Fonte: *Introduction to Statistics and Data Analysis*, R. Peck, C. Olsen e J. L. Devore, 5a edição

Família paramétrica

- Considere a família paramétrica

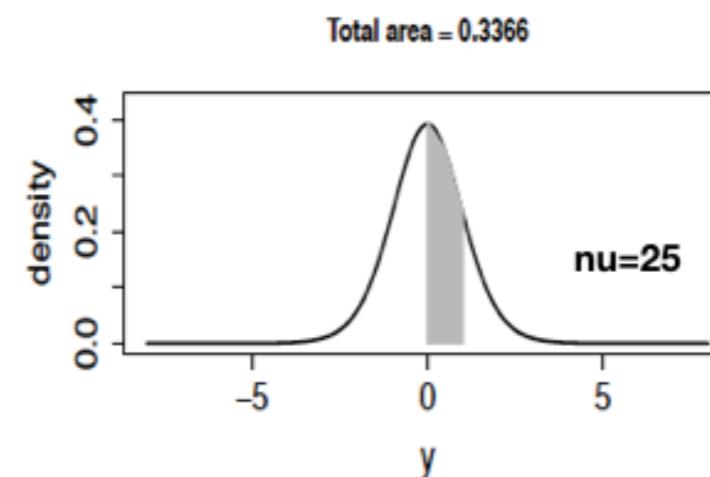
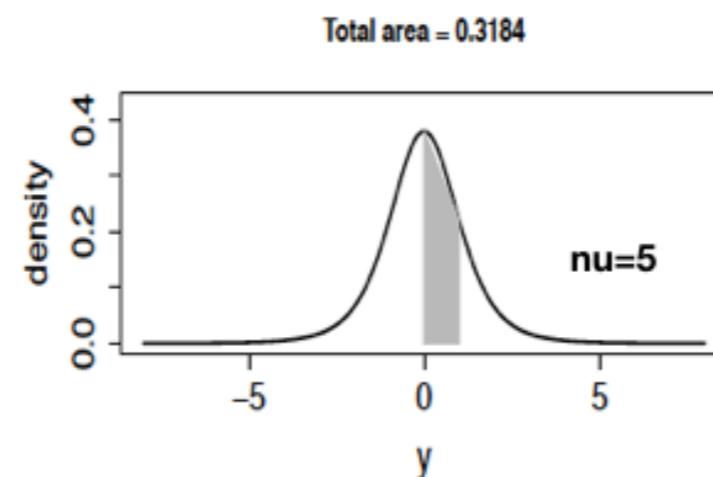
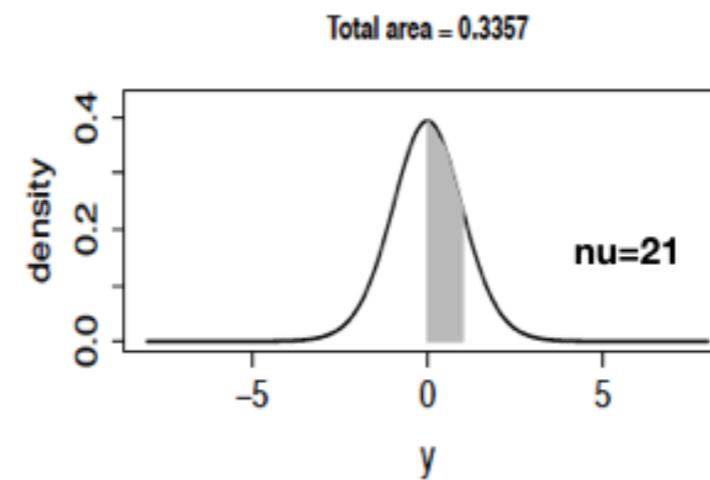
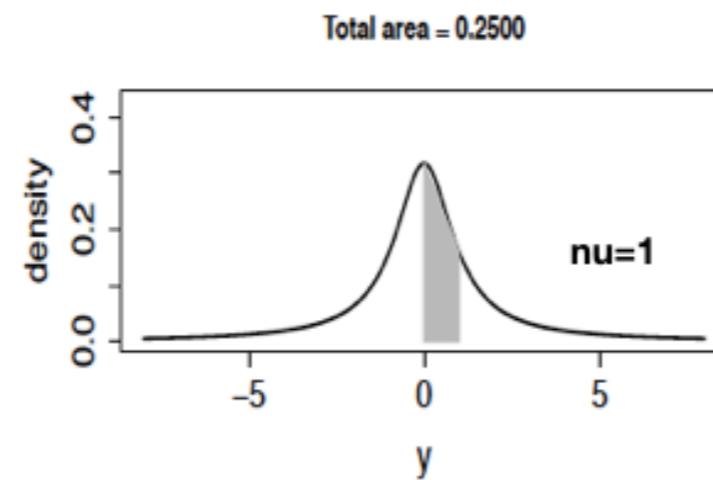
$$\mathcal{M} = \{f(\cdot \mid \theta), \theta \in \mathcal{H}\}$$

- Exemplo: A t-student é uma família paramétrica que depende dos parâmetros graus de liberdade, locação e escala.

$$f(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right) \nu^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right) \pi^{1/2} \sigma} \left\{ \nu + \left(\frac{y - \mu}{\sigma} \right)^2 \right\}^{-(\nu+1)/2}, \quad y \in \mathfrak{R}.$$

Distância entre distribuições

Note que a distância no domínio paramétrico é 4 para os dois exemplos. Mas as distribuições são bem distintas.



Como medir distâncias entre distribuições?

Divergências

- A divergência entre duas distribuições pode ser medida por

$$D_H(P, Q) = \int_{\Re} (\sqrt{f(y, P)} - \sqrt{f(y, Q)})^2 dy,$$

- No exemplo, para essa medida temos que a distância no primeiro caso é **0.11** e no segundo caso é **5.5 10⁻⁵**.
- A divergência de Kullback-Leibler (KL) é dada por

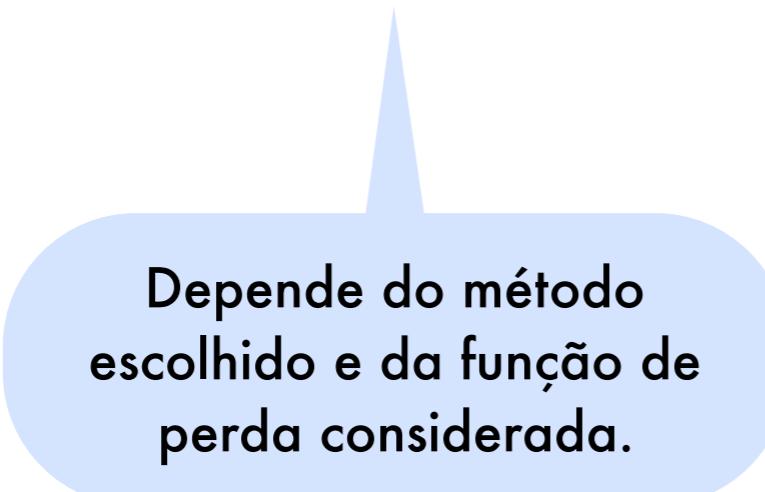
$$KL(q \mid p) = \int log(q(y)) - log(p(y)) q(y) dy$$

Algumas distribuições de probabilidade

- Bernoulli
- Uniforme discreta
- Binomial
- Poisson
- Binomial negativa
- Normal
- Gama
- Beta
- Lognormal
- Laplace (exponencial dupla)
- Dirichlet

Learning distribution

- Dada uma distribuição $f(y | \theta)$ parametrizada por θ e dados $\{y_1, \dots, y_N\}$, aprender significa inferir θ que **melhor** se ajusta aos dados.
- O que significa **melhor**?



Depende do método
escolhido e da função de
perda considerada.

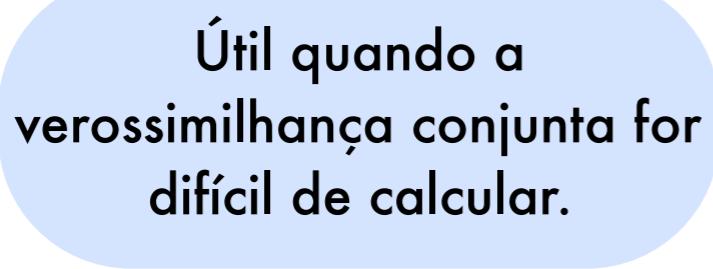
Verossimilhança, posteriori e pseudo-verossimilhança

- Máximo a posteriori: $\theta^{MAP} = \arg \max_{\theta} p(\theta | y)$, onde

$$p(\theta | y) = f(y | \theta)p(\theta)/g(y).$$

- Máximo da verossimilhança: $\theta^{ML} = \arg \max_{\theta} f(y | \theta)$.
- Máximo usando pseudo-verossimilhança: se y_i é um vetor $y_i = (y_{i1}, \dots, y_{id})$ então

$$\theta^{MPL} = \arg \max_{\theta} \sum_{i=1}^n \sum_{j=1}^p \log(f(y_{i,j} | y_{i,-j}, \theta))$$



Útil quando a verossimilhança conjunta for difícil de calcular.

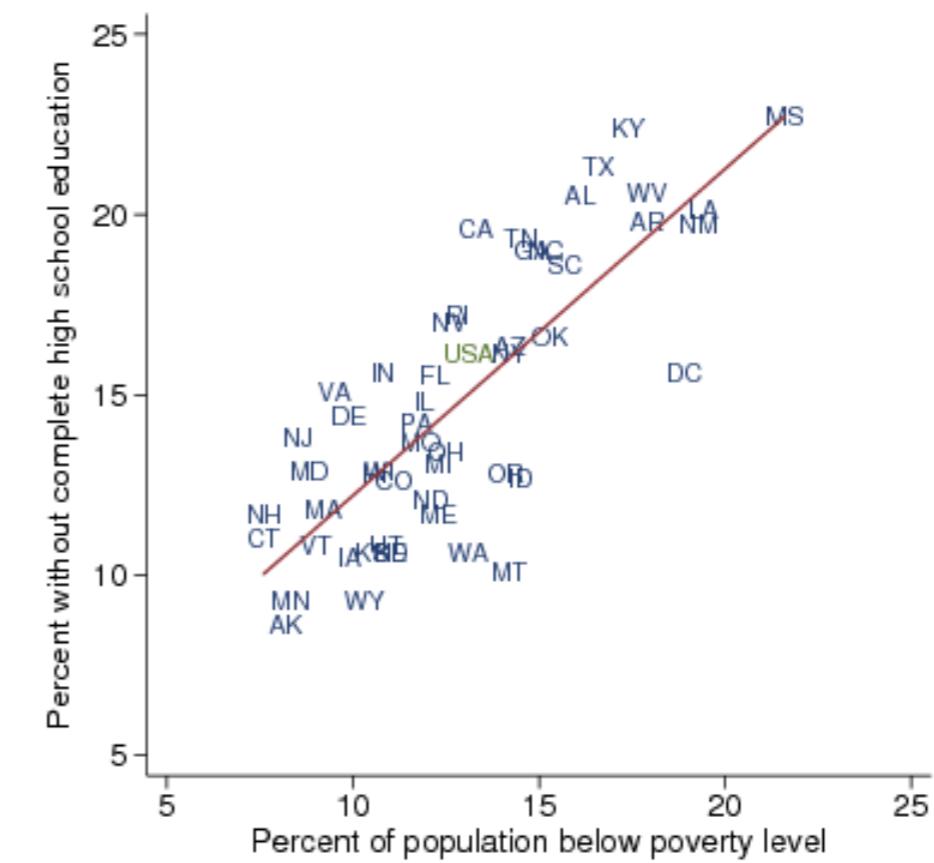
Modelos

- De forma geral, o objetivo é representar um conjunto de dados em dimensões menores.
- Dessa forma, iremos dividir a variabilidade dos dados em duas partes:

Padrão de interesse + Erro

- O padrão de interesse pode ter várias camadas.
- Em geral, técnicas de machine learning criam algoritmos baseados em bom poder preditivo.
- Mas note que o uso de uma ferramenta de machine learning pode não ser adequada para tempo futuros (long term), devendo então ser feita uma revisão da ferramenta (avaliação dinâmica).

Modelos lineares



Ideia básica

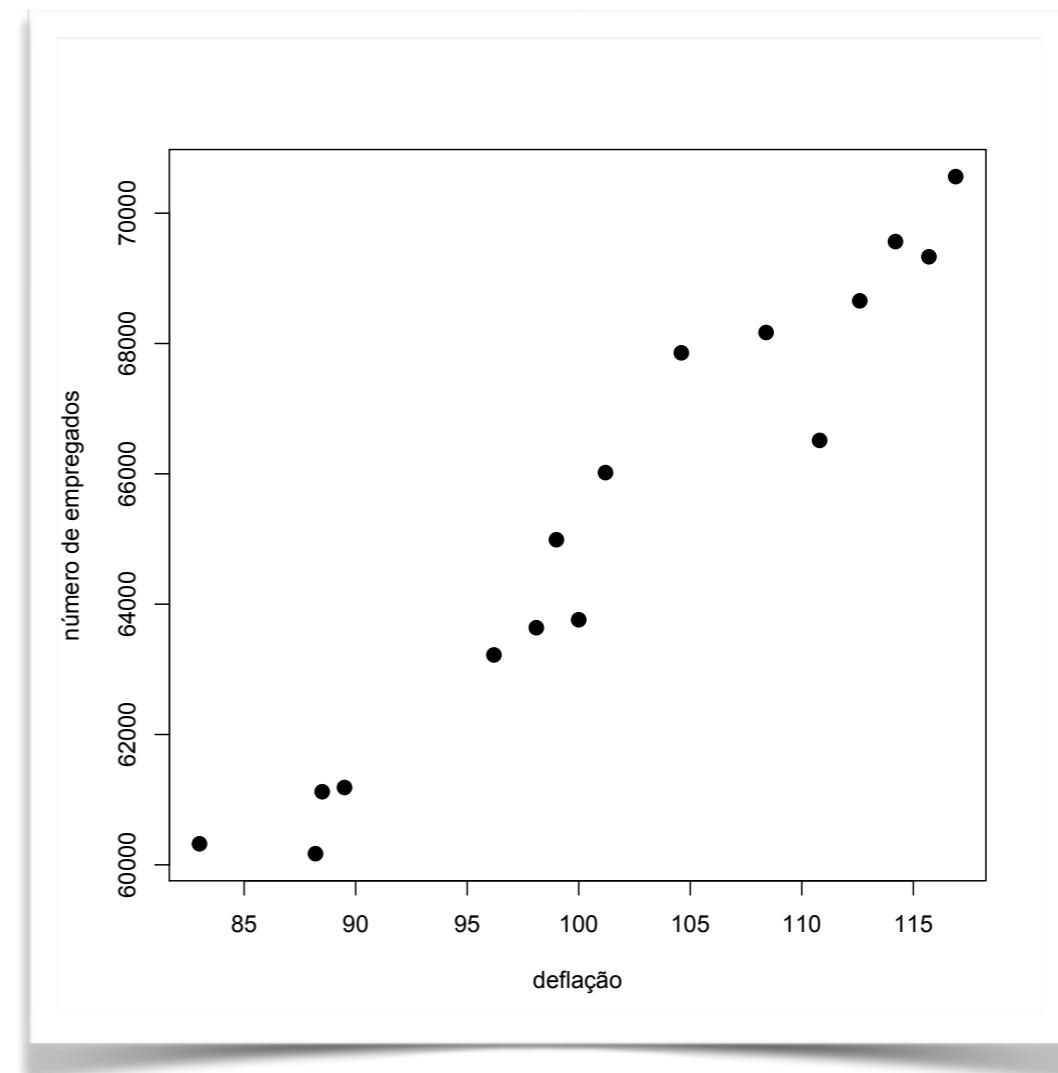
- Após observar dados de treinamento $\{(\mathbf{x}_i, y), i = 1, \dots, N\}$ para um input \mathbf{x}_i e output y_i , deseja-se ajustar um modelo linear para relacionar input e output.
- De forma simplificada podemos dizer

$$y(x) = a + bx + \epsilon.$$

- As constante a e b devem ser encontradas segundo algum critério de otimização de uma função perda.

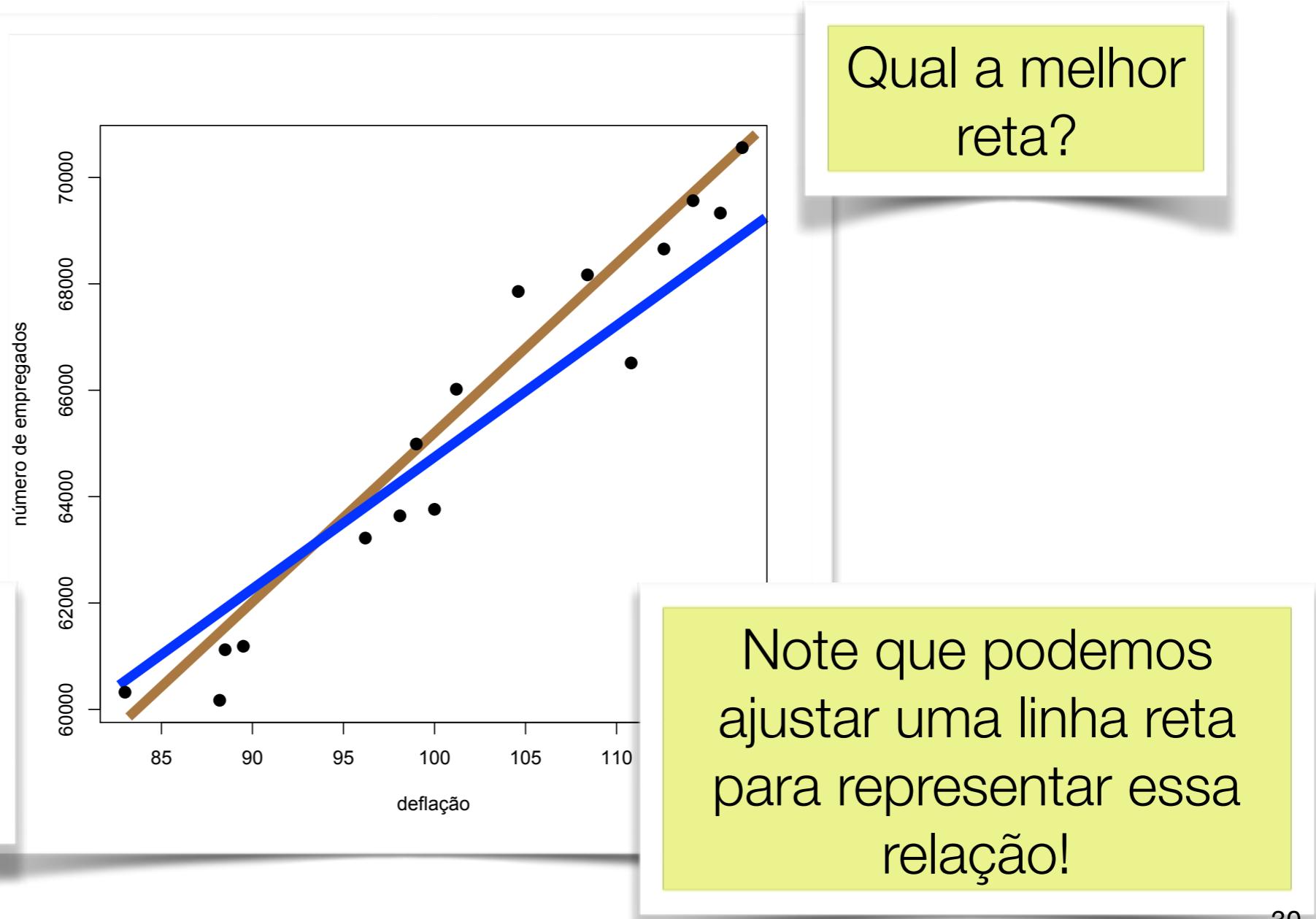
Objetivo do modelo de regressão

- Objetivo: investigar e modelar a relação entre variáveis.
- Exemplo 1: X é um índice econômico de deflação (queda de preços). Y é o número de pessoas empregadas.
- Existe relação entre X e Y?



Relação entre x e y

- Esse gráfico é conhecido como scatter plot e é usado para visualizar a relação entre 2 variáveis (x e y).

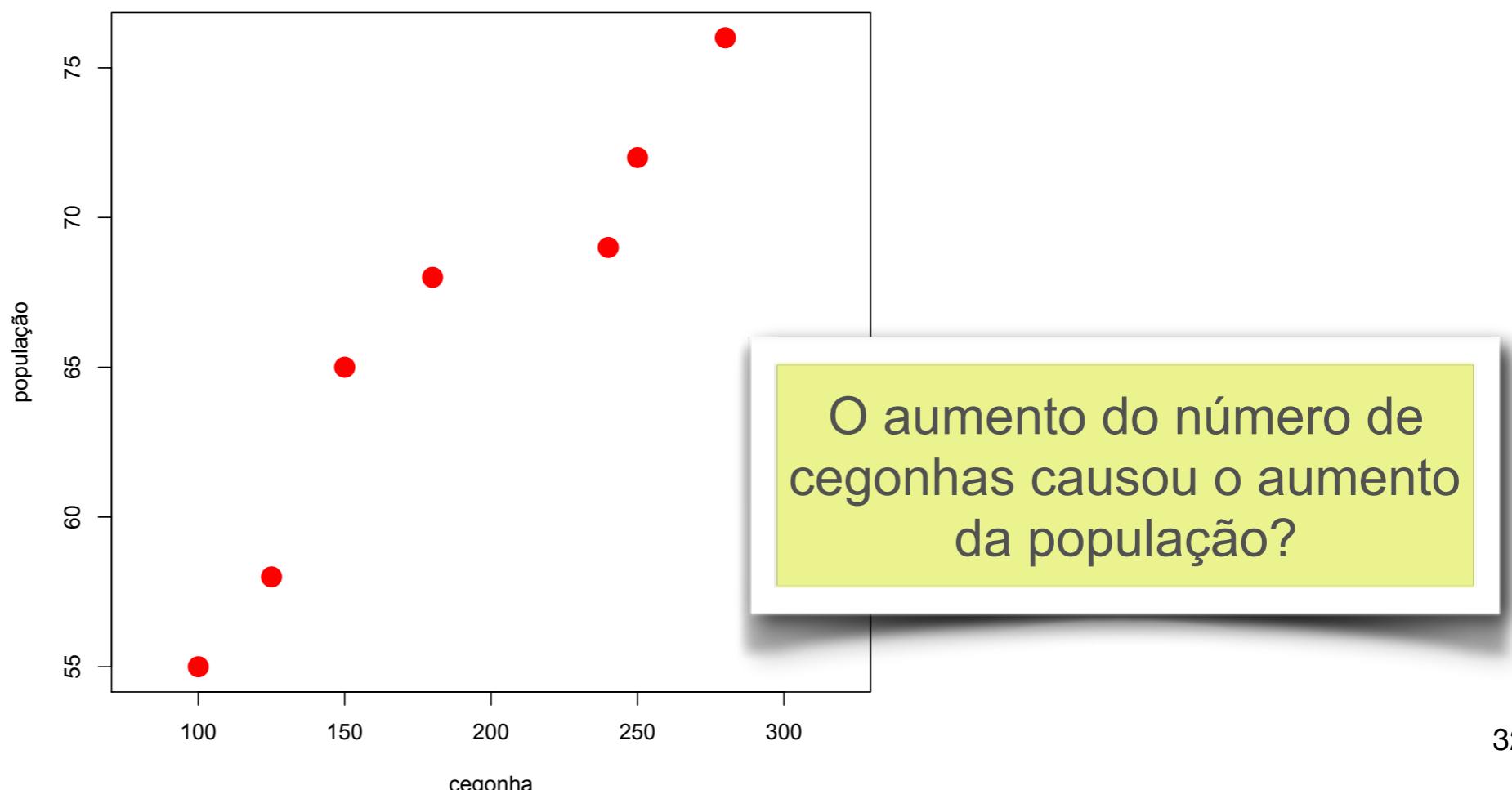


Algumas definições

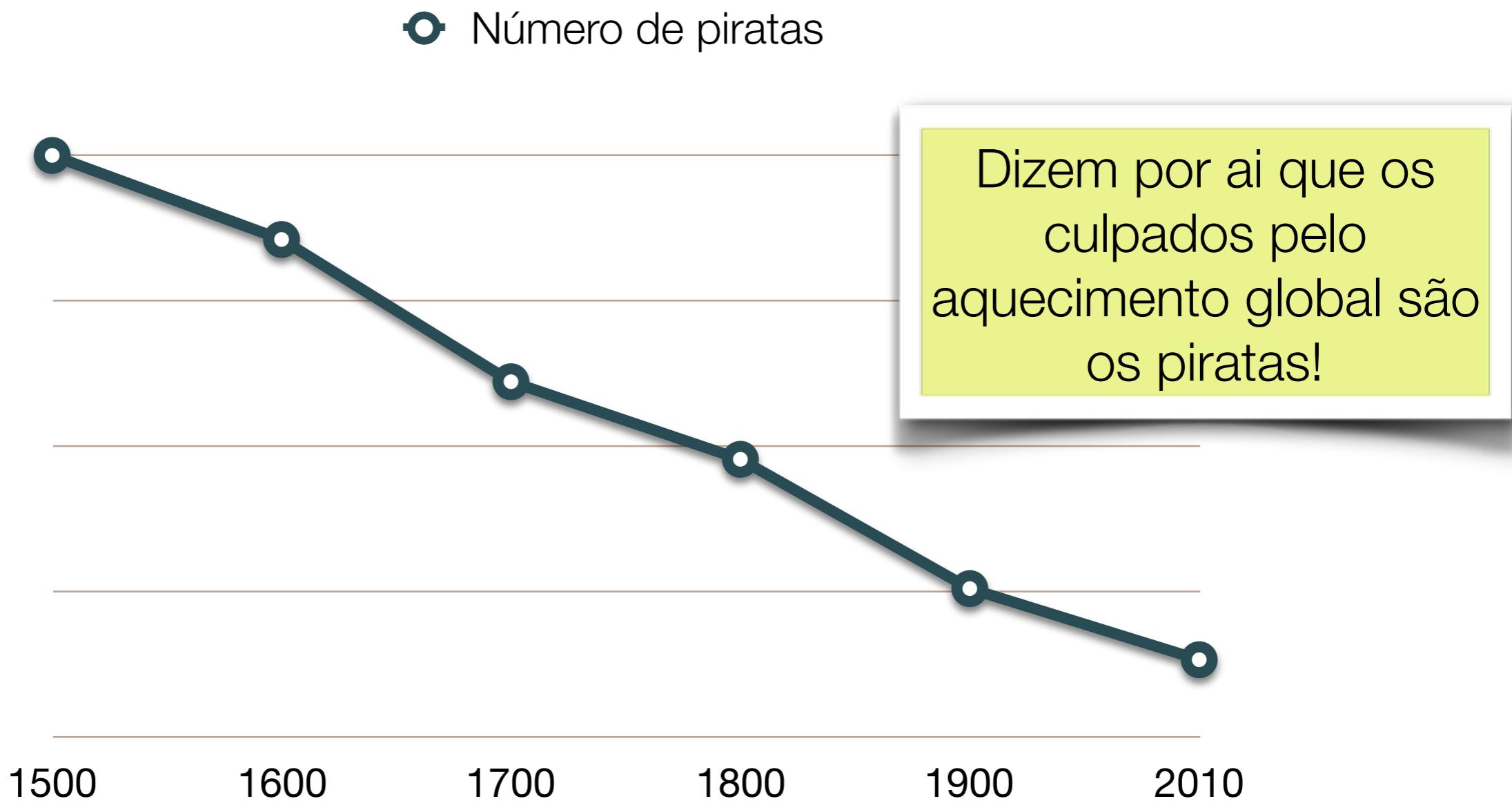
- X: variável explicativa (frequentemente chamada input ou feature em machine learning);
- Y: variável resposta (frequentemente chamada output em machine learning);
- Regressão linear simples: modelo que representa a relação entre duas variáveis x e y através de uma reta;
- Regressão linear múltipla: modelo que representa a relação entre y e mais de uma variável explicativa de forma linear nos parâmetros.

Importante: causa e efeito

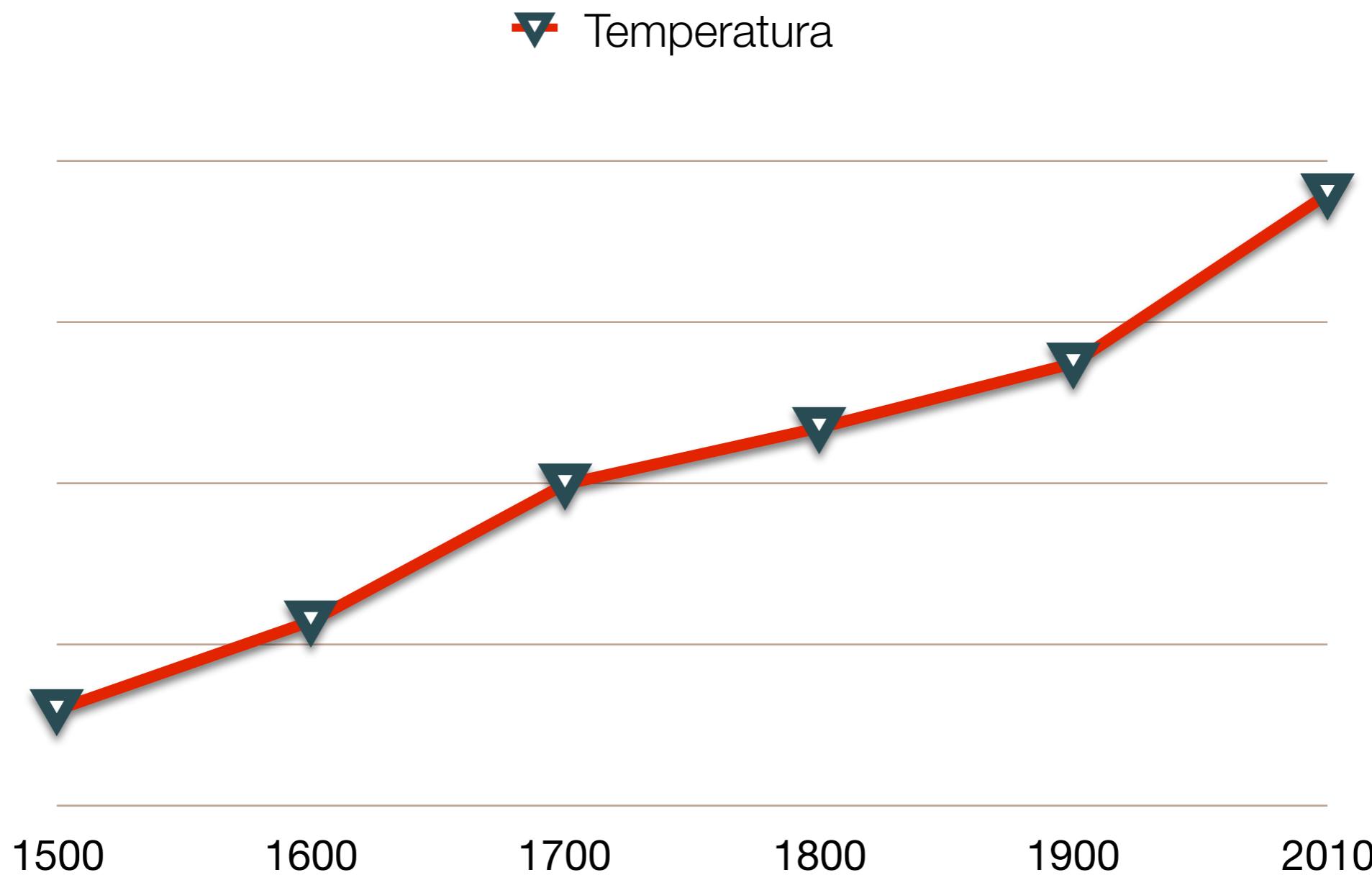
- Note que um modelo de regressão não implica em relação causa e efeito!
- Exemplo: (Box, Hunter & Hunter, Statistics for Experimenters, p.8). O gráfico mostra a população de Oldemberg, Alemanha, no fim de cada um dos 7 anos (Y) contra o número de cegonhas (pássaros) naquele ano (X).



Exemplo: piratas x aquecimento global



Exemplo: piratas x aquecimento global



Exemplo: piratas x aquecimento global



Será que a diminuição do número de piratas está causando o aumento da temperatura nos oceanos?

Número de piratas

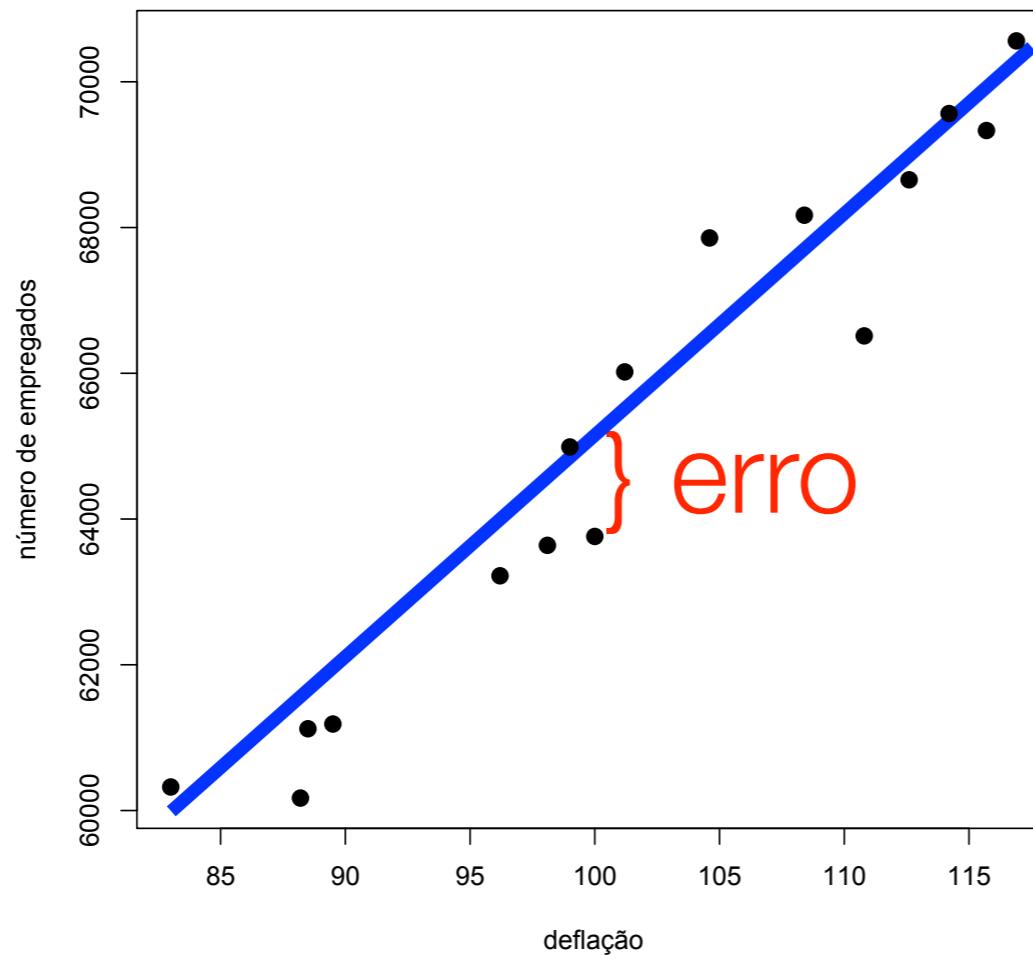
Perhaps pirates throw
ice cubes into the water
to cool the Earth down?



Exemplo: piratas x aquecimento global

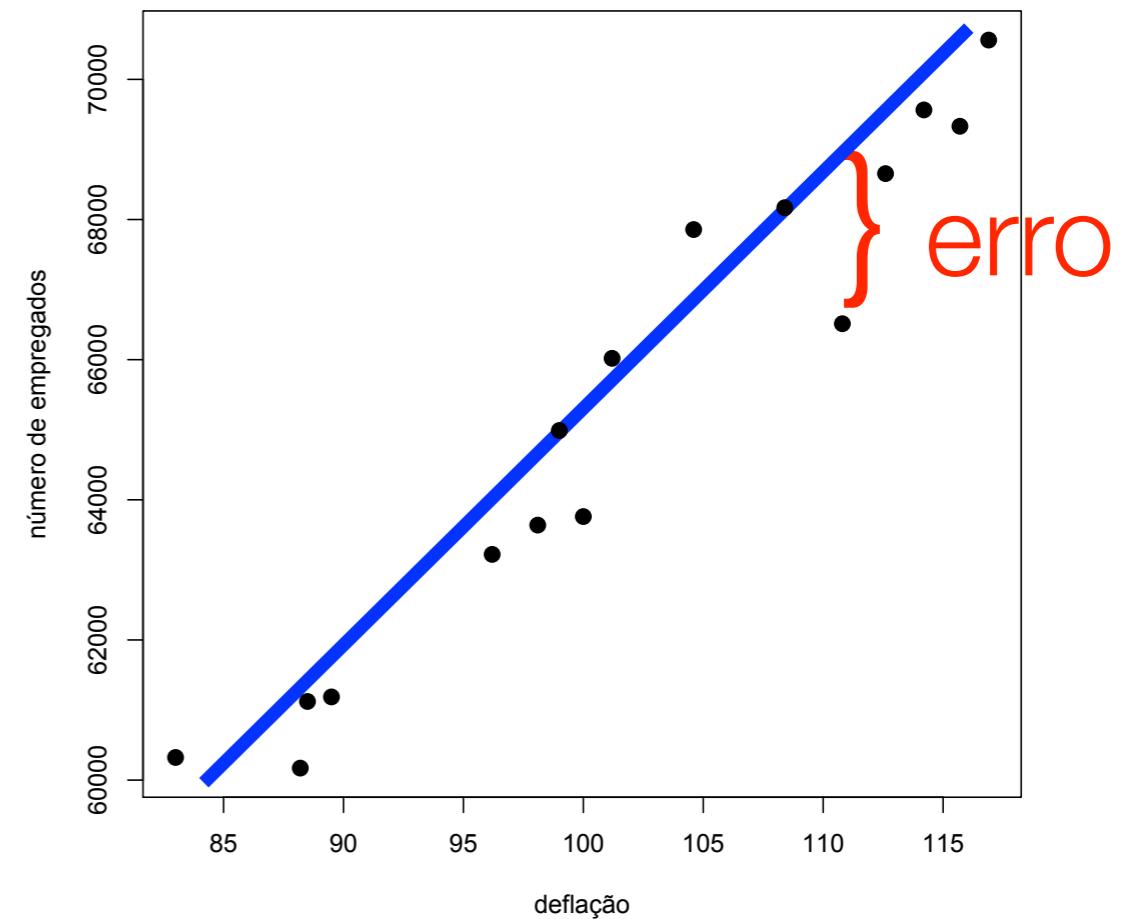
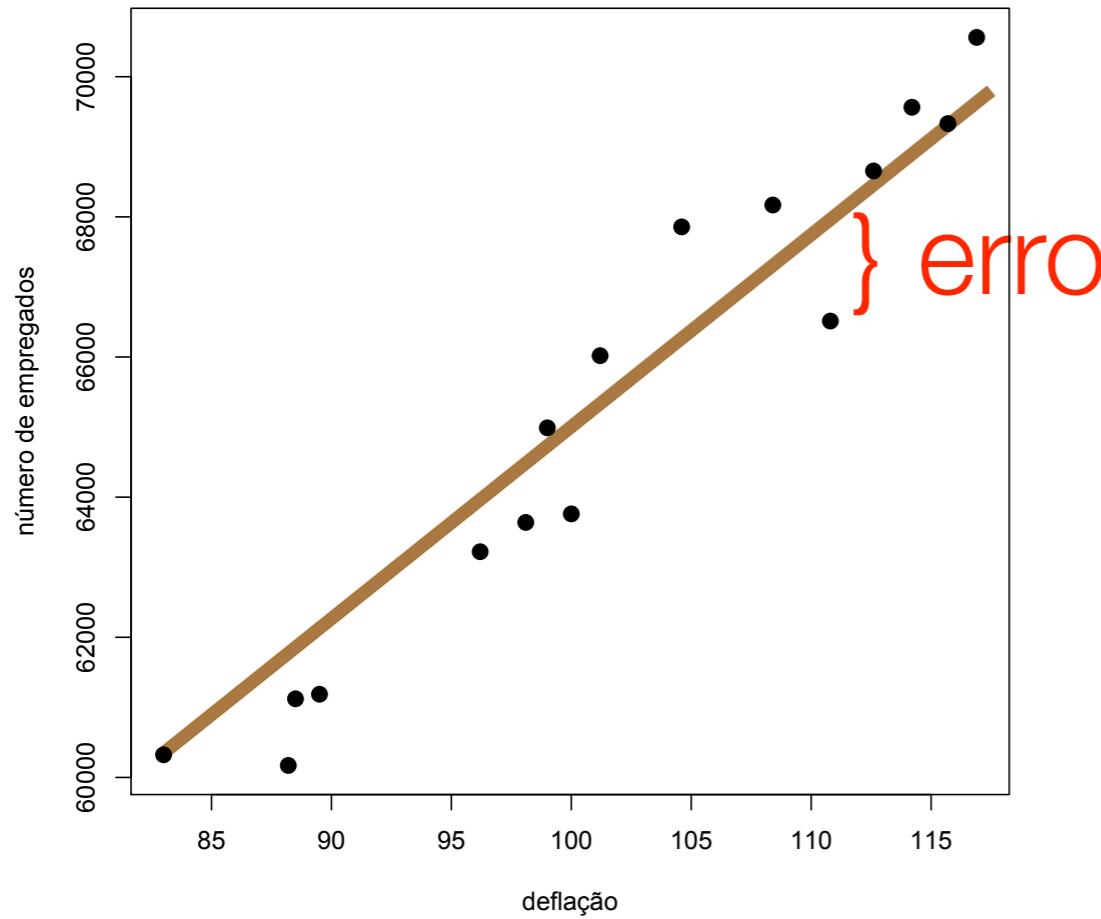
- O número de piratas provavelmente diminuiu ao longo do tempo devido a segurança mais rigorosa de mares e fronteiras!
- E a temperatura possivelmente está subindo devido a ação do homem (poluição etc).
- Não existe uma relação de causa e efeito nesse caso!

Erro do modelo ajustado



- Note que os dados (x,y) não caem exatamente na reta.
- Isso ocorre devido ao erro.

Método de mínimos quadrados



- O método encontra a reta que minimiza a soma dos erros ao quadrado.

Voltando a pergunta: Qual a melhor reta?

Método de mínimos quadrados

Considere observações

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

Queremos encontrar a reta ajustada

$$\hat{y} = a + bx$$

que minimiza a soma de quadrados dos erros. Então queremos encontrar a e b de forma a minimizar

$$S(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

Estamos utilizando uma função de risco quadrático. E iremos encontrar a e b que minimizem o risco.

Outras funções de risco (tal como valor absoluto) poderiam ser usadas.

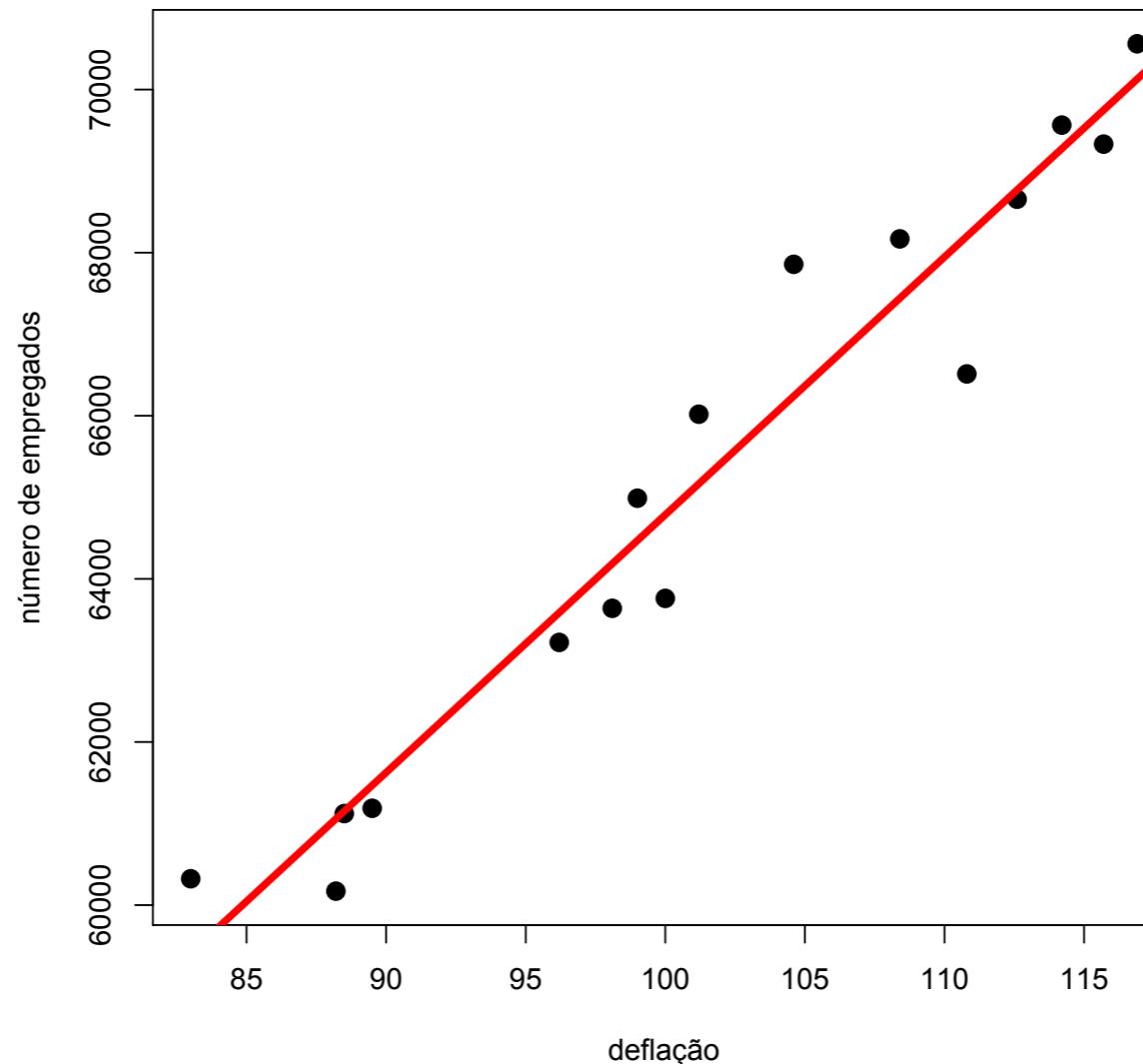
Mínimos quadrados

Solução de mínimos quadrados (regressão linear simples)

$$\frac{\partial}{\partial a} S(a, b) = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \rightarrow a = \bar{y} - b\bar{x}$$

$$\frac{\partial}{\partial b} S(a, b) = -2 \sum_{i=1}^n x_i(y_i - a - bx_i) = 0 \rightarrow b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

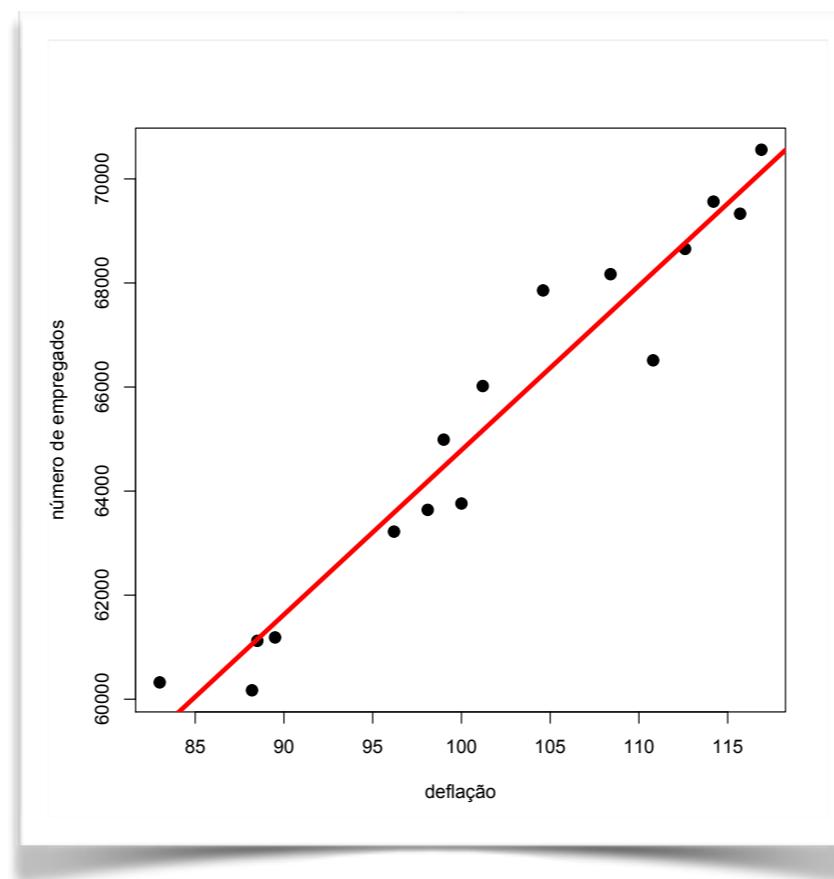
Exemplo 1: reta de mínimos quadrados



Reta de mínimos quadrados $\hat{y} = 33180.94 + 316.05x$

Usos do modelo de regressão

- **Descrição dos dados:**
 - A reta $a + b x_i$ pode ser utilizada como resumo dos dados.
 - Gráfico: scatter plot com a reta ajustada.



Usos do modelo de regressão

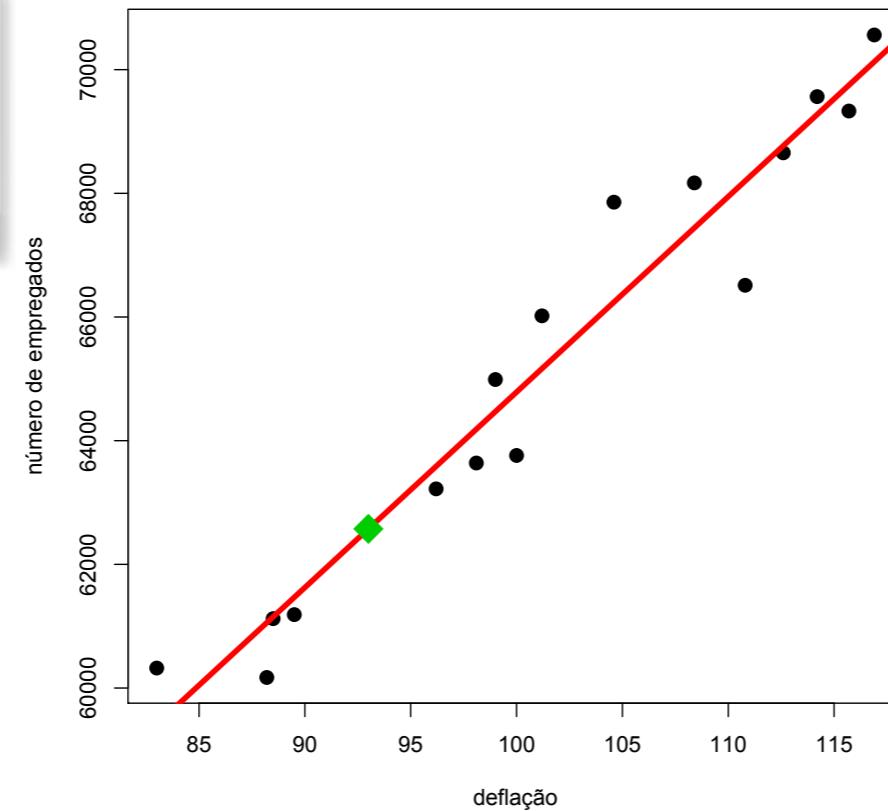
- **Estimação dos parâmetros**
 - Os coeficientes de regressão a e b fornecem informação sobre o problema em estudo.
 - No exemplo, a reta é $y = 33180.94 + 316.05 x$, para 0 de deflação temos número de empregados de 33180.94.
 - Se aumentamos em 1 unidade a deflação, aumentamos em 316.05 o número de pessoas empregadas.

Usos do modelo de regressão

- **Previsão**
 - Dados um novo valor x_0 podemos encontrar o valor predito nesse ponto:

$$\hat{y}_0 = a + bx_0$$

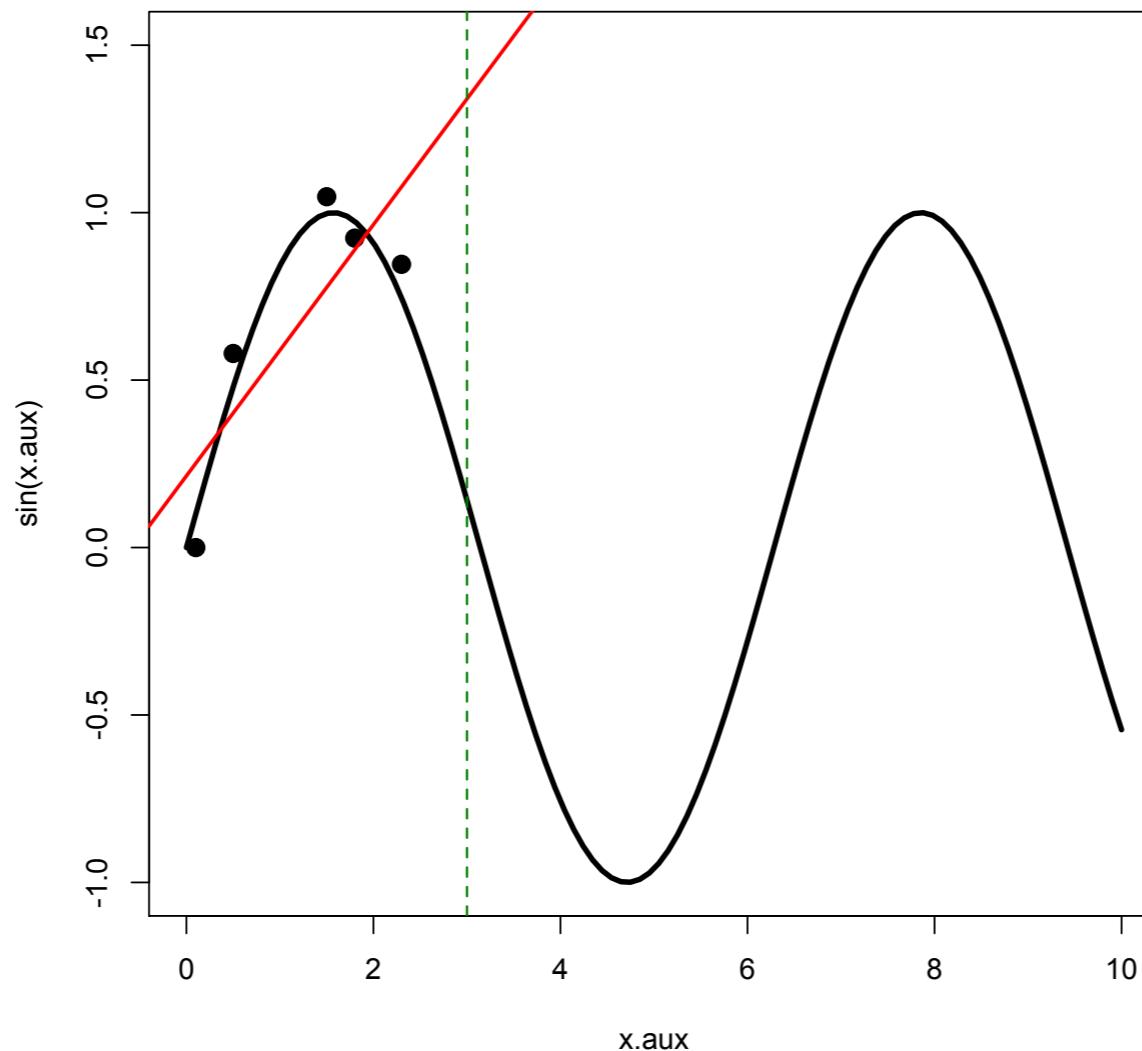
Exemplo: $x_0 = 100$
então $y_0 = 64786.26$



Chamaremos interpolação se x_0 está entre $\min\{x_i\}$ e $\max\{x_i\}$.
E extrapolação se x_0 está fora desses limites!

Importante: extrapolação

- Extrapolação pode ser perigoso!



- A reta ajustada pode ser uma boa aproximação somente na região onde os dados foram observados.

Modelo básico

Considere o modelo de regressão para representar a relação entre a covariável x e a resposta y .

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

onde β_0 é o intercepto, β_1 é o coeficiente angular e ϵ é o erro.

Suposições do modelo: o erro é uma variável aleatória com as seguintes propriedades.

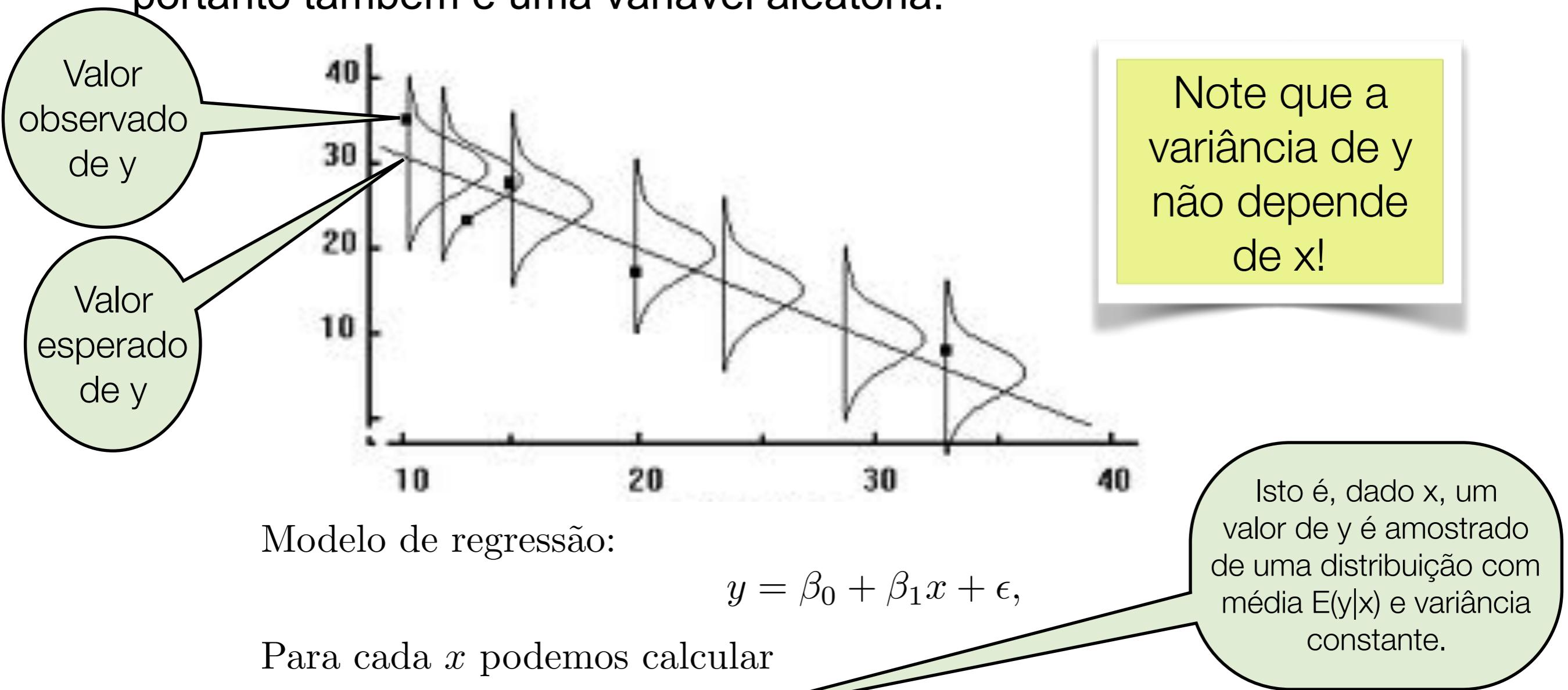
- O erro tem média 0, isto é, $E(\epsilon) = 0$;
- O erro tem variância desconhecida constante, isto é, $Var(\epsilon) = \sigma^2$;
- Os erros são não correlacionados, isto é, $Corr(\epsilon_i, \epsilon_j) = 0$ se $i \neq j$.

Chamado
vício em
machine
learning.

Os erros têm média 0, variância
constante e são não correlacionados!

Variável resposta

- No modelo de regressão a variável resposta y é função dos erros e portanto também é uma variável aleatória.



Modelo de regressão:

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

Para cada x podemos calcular

$$E(y | x) = \beta_0 + \beta_1 x, \text{ e } Var(y | x) = \sigma^2$$

Interpretação dos coeficientes

- Vimos que para cada x temos

$$E(y | x) = \beta_0 + \beta_1 x.$$

- (i) Para $x = 0$ temos $E(y | x) = \beta_0$, então β_0 é o valor esperado de y quando x assume o valor 0.
- (ii) Seja x^* e $x^* + 1$ dois possíveis valores de x . Podemos obter

$$E(y | x^* + 1) = \beta_0 + \beta_1(x^* + 1)$$

$$E(y | x^*) = \beta_0 + \beta_1 x^*$$

Resultando em

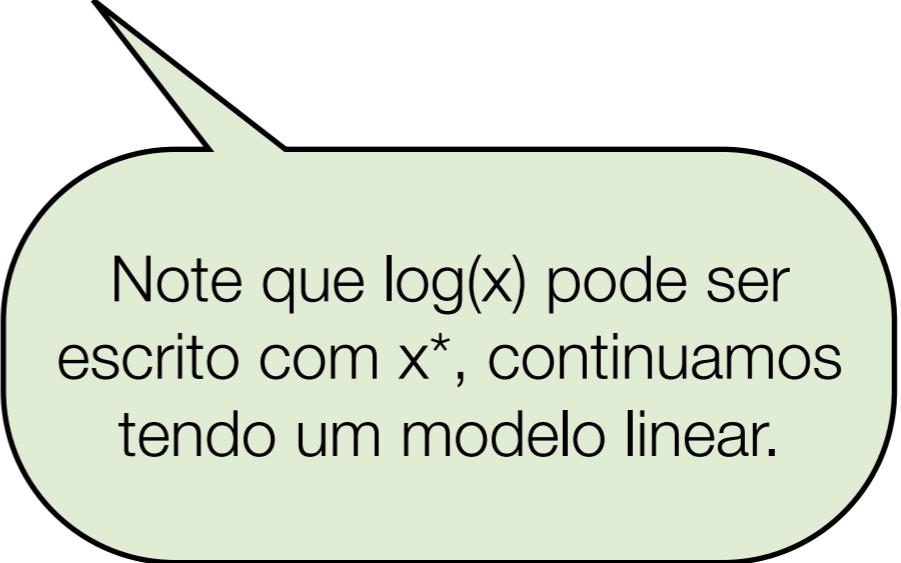
$$E(y | x^* + 1) - E(y | x^*) = \beta_1,$$

então β_1 é a mudança no valor esperado de y quando acrescentamos uma unidade em x .

Modelos lineares

- Por que dizemos que o modelo é linear?
- O modelo abaixo é linear?

$$y = \beta_0 + \beta_1 \log(x) + \epsilon$$



Note que $\log(x)$ pode ser escrito com x^* , continuamos tendo um modelo linear.

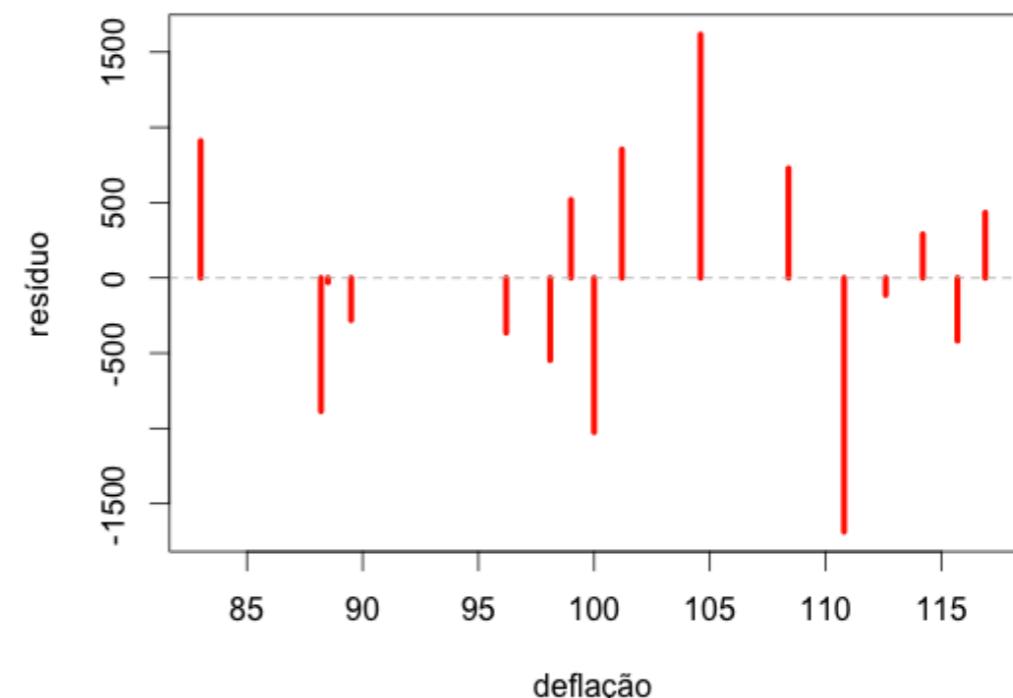
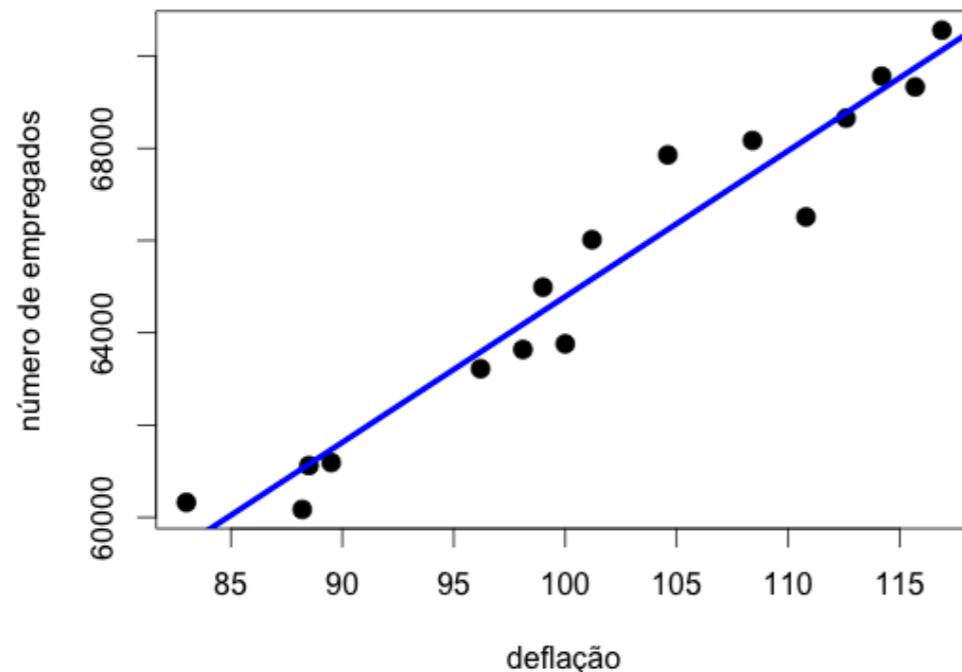
Alguns conceitos importantes

- **Reta ajustada:** É um estimador pontual para a média de y dado x ,
 $E(y | x) = \beta_0 + \beta_1 x$.

$$\hat{y} = a + bx.$$

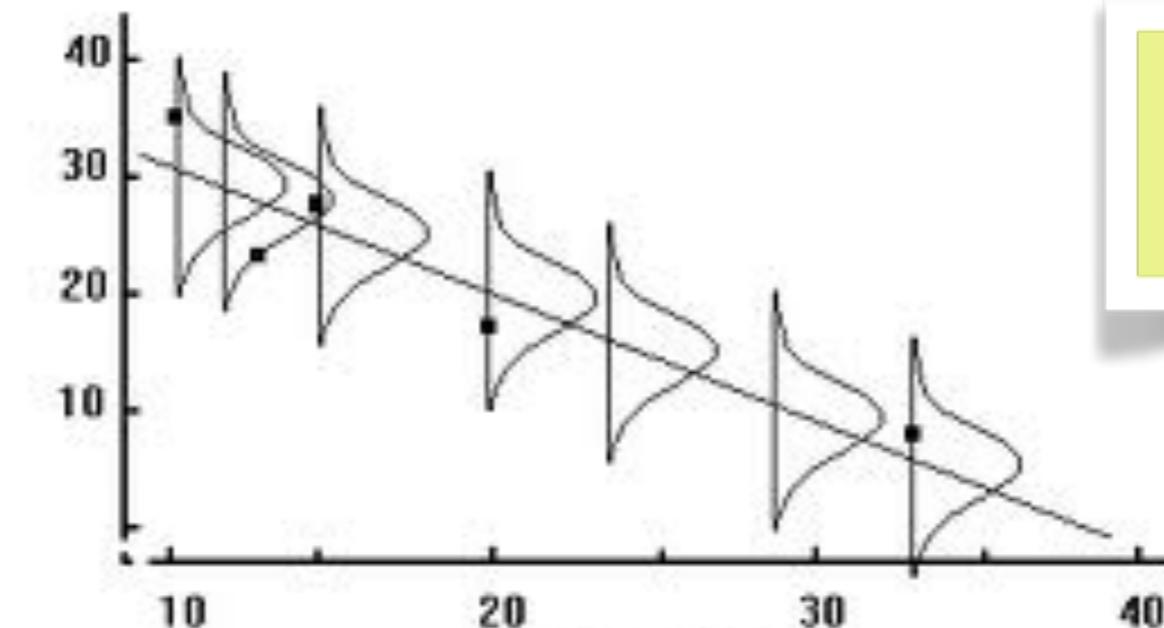
- **Resíduo:** São vistos como realizações da variável aleatória ϵ .

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i, \quad i = 1, \dots, n.$$



Distribuição dos erros

- Vimos que no modelo de regressão a variável resposta y é função dos erros e , portanto, também é uma variável aleatória.



Mas qual a distribuição dos erros? E qual a distribuição de y ?

- Existe uma tendência dos erros, em muitas situações, de terem distribuição normal.
- Isso se deve ao Teorema Central do Limite: se o erro é a soma de vários erros, vindos de diversas fontes, então a distribuição da soma pode ser aproximada pela normal!

Vamos supor que os erros tem distribuição Normal!

Distribuição dos y_i 's

Vamos supor que os erros são não correlacionados com

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n.$$

O modelo básico de regressão é tal que

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Podemos calcular:

- $E(y_i | x_i) = \beta_0 + \beta_1 x_i;$
- $V(y_i | x_i) = V(\epsilon_i) = \sigma^2;$

Além disso, temos que y_i é uma combinação linear de uma variável aleatória com distribuição normal, então y_i também tem distribuição normal.

Conclusão:

$$y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Resumo de resultados de regressão

- Os EMV dos coeficientes da regressão coincidem com os estimadores de mínimos quadrados (Apêndice I).
- A significância da regressão pode ser testada ao assumirmos esse modelo para os dados (Apêndice II).
- Podemos obter intervalos de confiança para os coeficientes e para previsões (Apêndice III).

Decomposição da variabilidade

Ideia: separar a variabilidade total de y em duas partes:

ERRO e REGRESSÃO

Podemos escrever:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Resultando em

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

Medida
da Variabilidade
total
SS_T

Medida da
Variabilidade do
erro
SS_{res}

Medida da
Variabilidade da
regressão
SS_{reg}

= 0

Outra maneira de testar a significância da regressão

- Decomposição da variância:

$$SS_T = SS_{res} + SS_{reg}$$

- Podemos mostrar que

$$SS_{reg} = \hat{\beta}_1^2 S_{xx}$$

- Isso sugere que para testar se a regressão (β_1) é significativa basta testar se SS_{reg} próximo de 0 ou grande.

Isso é feito usando o teste F.

Coeficiente de determinação R^2

Podemos analisar a importância relativa da regressão com relação a variabilidade total.

$$R^2 = \frac{SS_{reg}}{SS_T} = 1 - \frac{SS_{res}}{SS_T}$$

- Se $SS_{res} \approx 0$ temos R^2 grande (próximo de 1) e grande importância da regressão.
- Se SS_{res} grande, então SS_{reg} pequeno e R^2 também pequeno (próximo de 0).

Em resumo, R^2 indica quanto da variabilidade de y é explicada pela regressão.

Exemplo de modelagem (regressão)

Exemplo: consumo de combustível

- Vamos começar com um exemplo simples (`data(mpg)` do R):

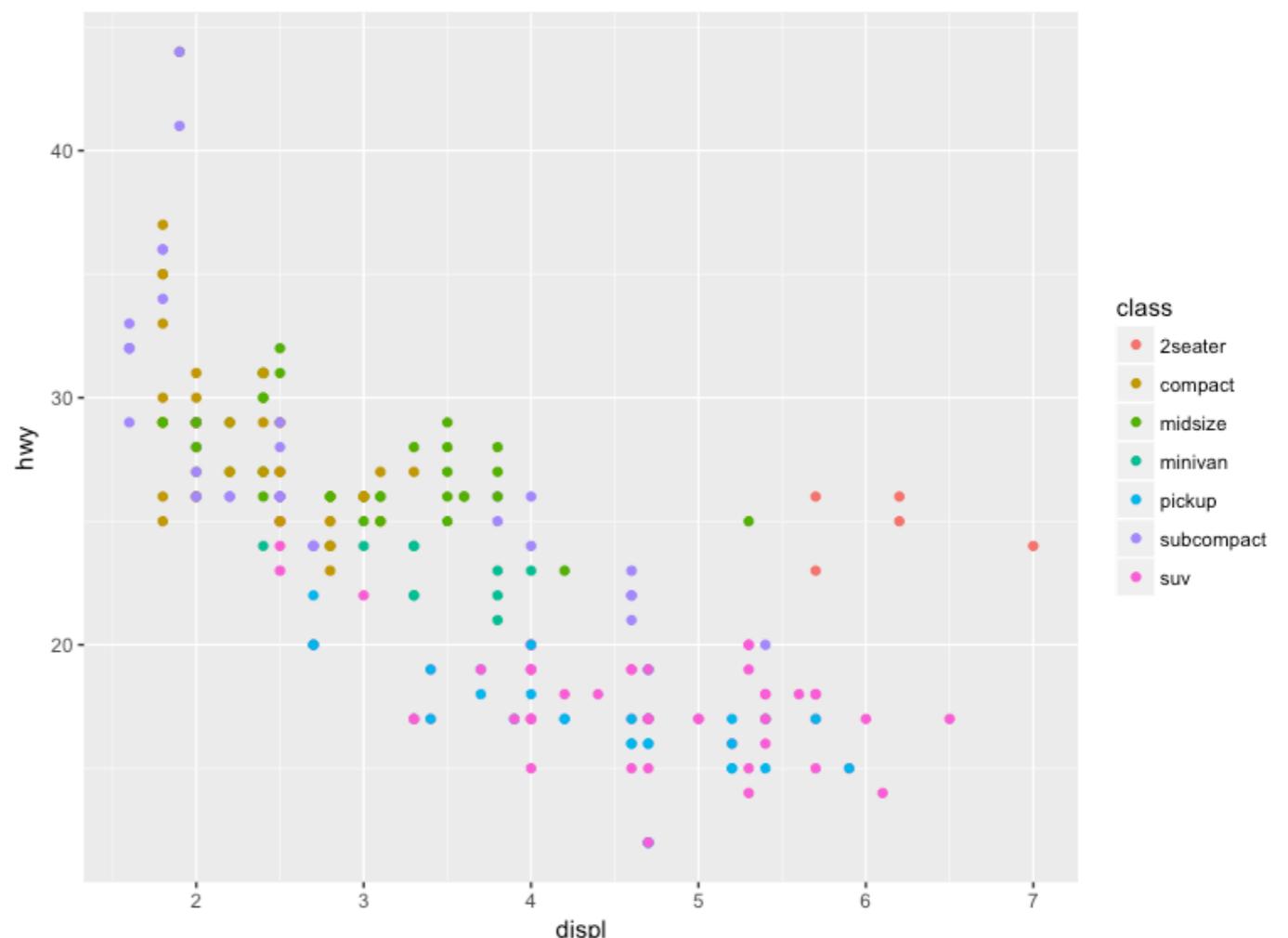
Carros com motores grandes usam mais combustível por quilômetro rodado (são menos eficientes) do que carros com motores pequenos?

Dados: dados para 234 automóveis e 11 variáveis de 1999 à 2008 para 38 modelos de carros o USA Environmental Protection Agency (EPA).

Fonte: <http://fueleconomy.gov>.

Eixo x: tamanho do motor

Eixo y: Eficiência de consumo de combustível



Exemplo: consumo de combustível

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
	<chr>	<chr>	<dbl>	<int>	<int>	<chr>	<chr>	<int>	<int>	<chr>	<chr>
1	audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
2	audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
3	audi	a4	2	2008	4	manual(m6)	f	20	31	p	compact
4	audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
5	audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
6	audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact

- 234 linhas e 11 colunas: Marca, modelo, tamanho do motor em litros, ano, número de cilindros do motor, câmbio, motorista, milhas por galão (na cidade), milhas por galão (na estrada), tipo de combustível, tipo de carro.
- Note que na tabela há alguma indicação sobre os tipos de variáveis. Porém cuidado deve ser tomado pois transformações podem ser necessárias.
- Iremos considerar nesta análise apenas tamanho do motor (em litros) e eficiência (milhas por galão na cidade).

Exemplo: consumo de combustível

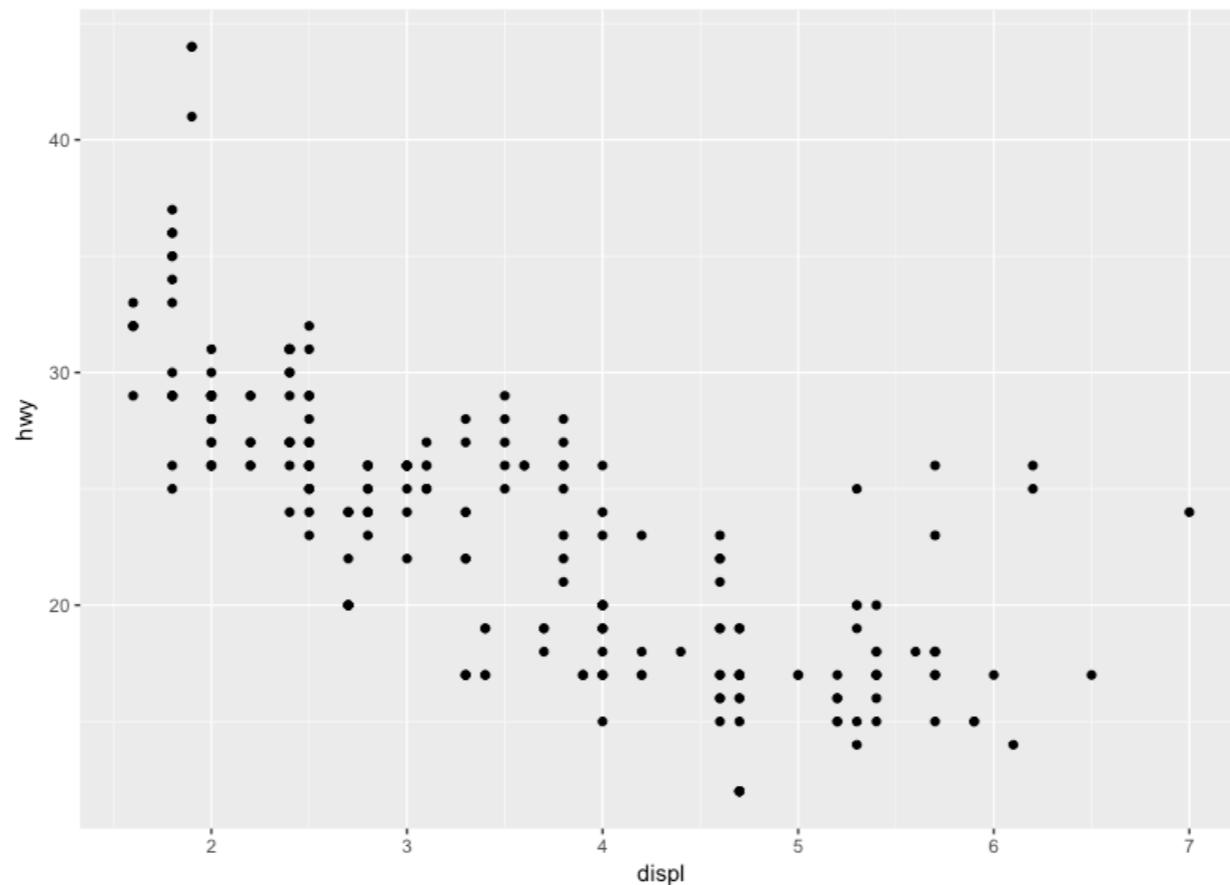
- Duas variáveis importantes no banco de dados:
 - **displ**: tamanho do motor em litros.
 - **hwy**: eficiência do carro em milhas por galão (cidade). Um carro pouco eficiente consome mais combustível do que um carro eficiente que viaja maiores distâncias com o mesmo combustível.

No R:

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy))
```

Exemplo: consumo de combustível

- Antes de responder à pergunta inicial, podemos tornar essa pergunta mais precisa.
- Como é a relação entre tamanho do motor e eficiência no consumo de combustível? É positiva? Negativa? Linear? Não linear?



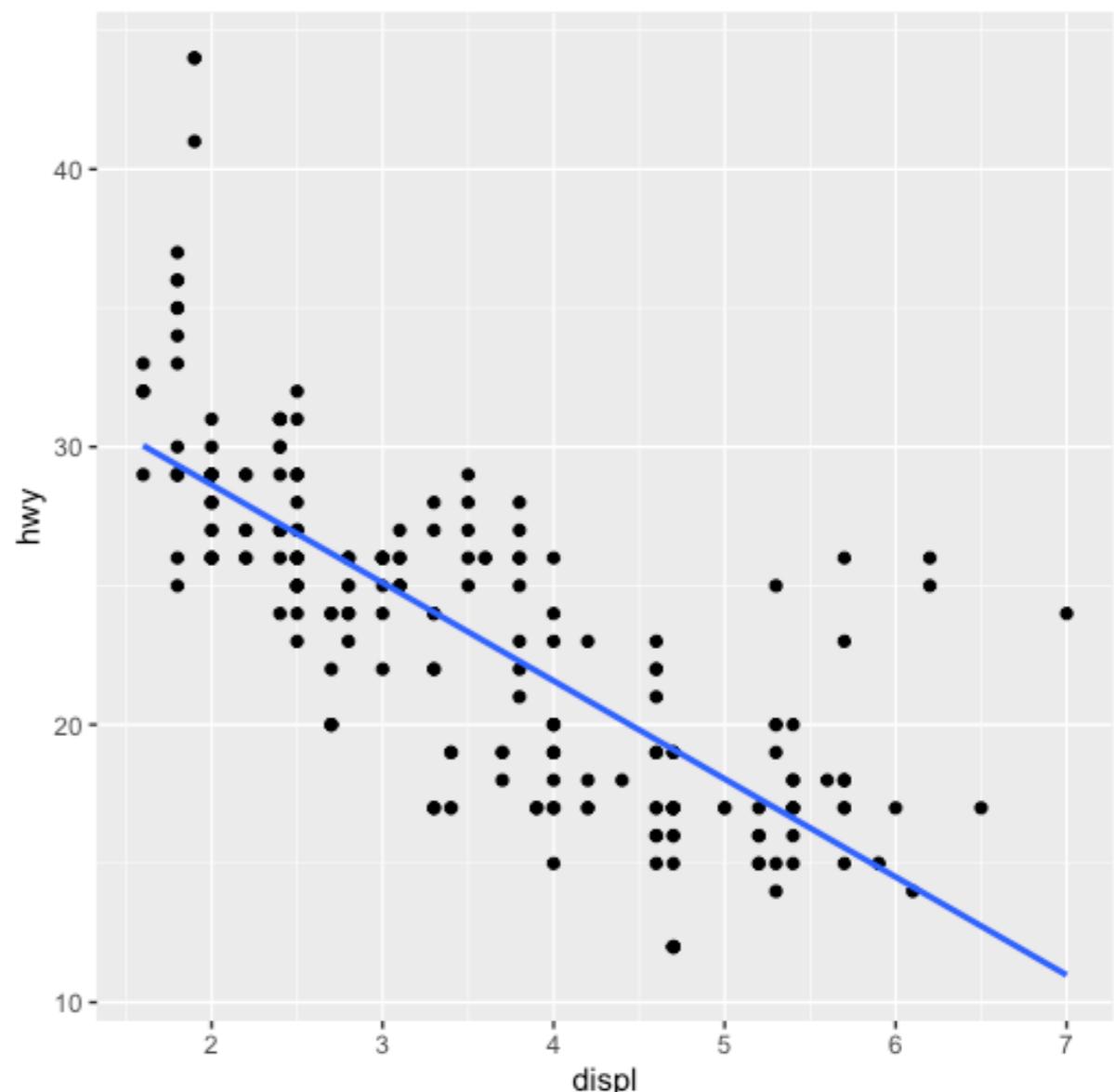
Exemplo: consumo de combustível

O ajuste parece adequado?

Para 1 unidade que aumentamos no tamanho do motor reduzimos a eficiência em 3.53 unidades.

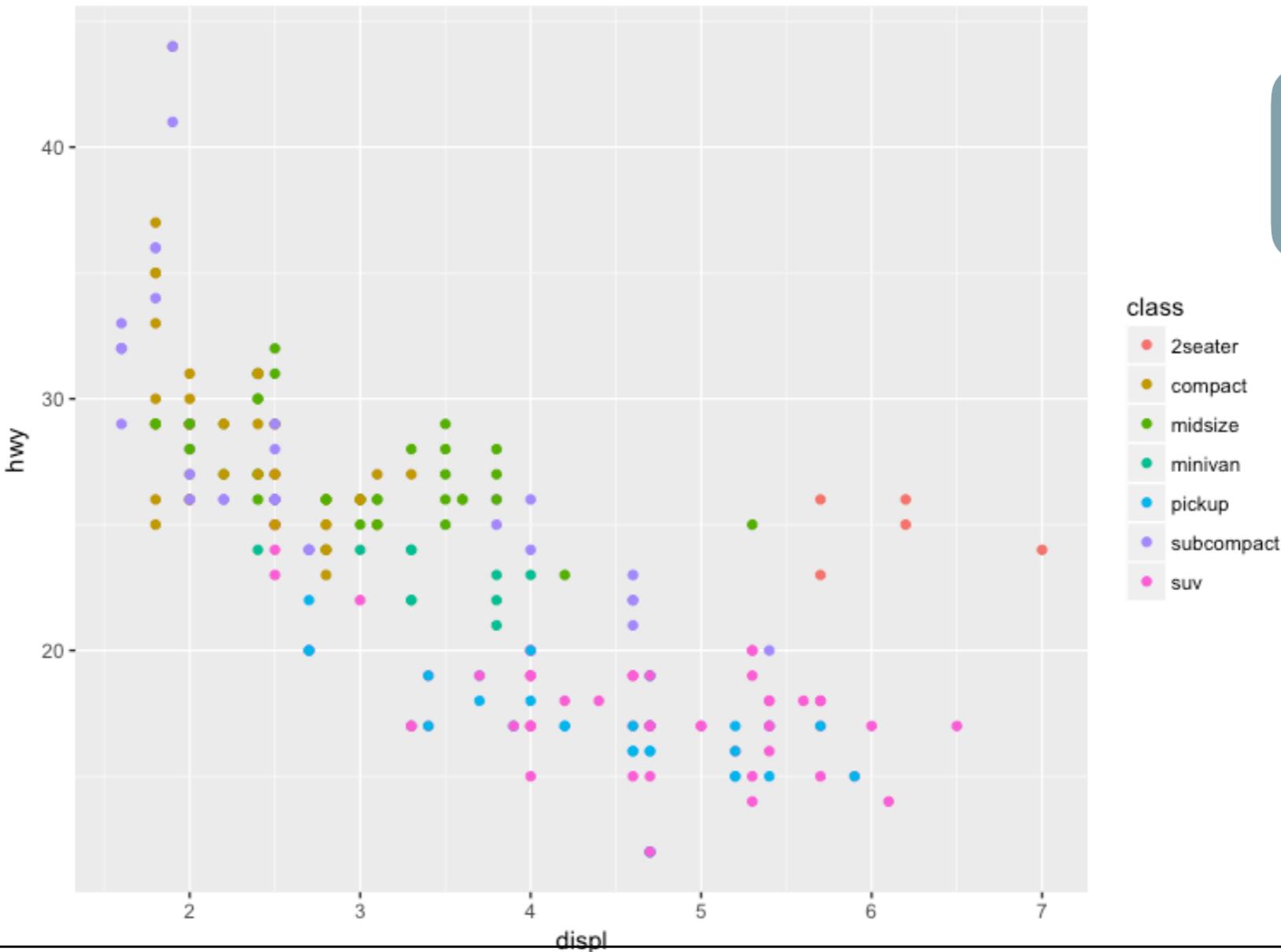
```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.6977    0.7204   49.55 <2e-16 ***
displ      -3.5306    0.1945  -18.15 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.836 on 232 degrees of freedom
Multiple R-squared:  0.5868, Adjusted R-squared:  0.585
F-statistic: 329.5 on 1 and 232 DF,  p-value: < 2.2e-16
```



Existem padrões importantes não revelados por essa análise simples?

Exemplo: consumo de combustível



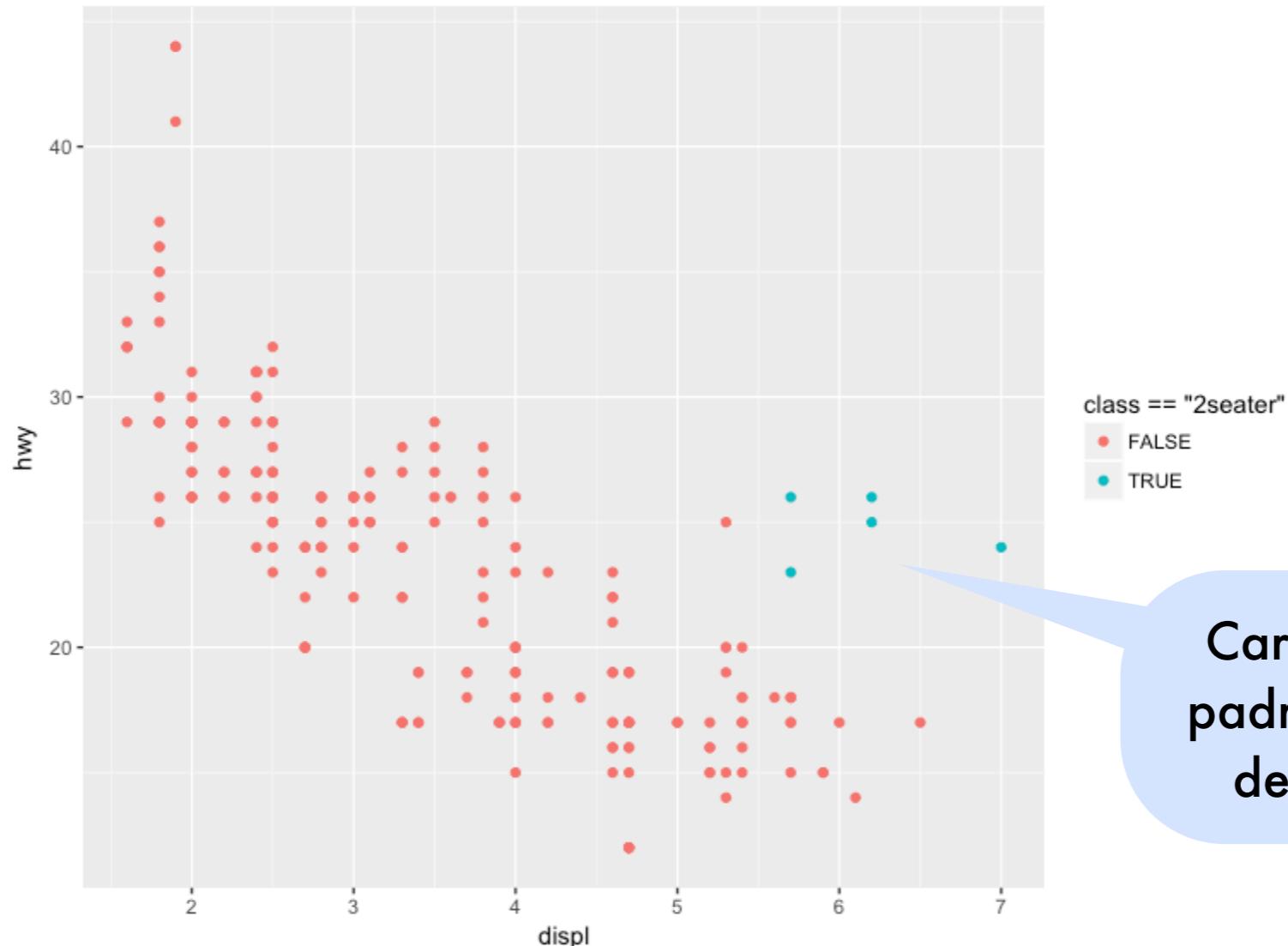
Algum padrão chama a atenção?

Os carros ditos 2 seater tem motores maiores e tem eficiência maior do que os de motores de tamanho similar.

No R:

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy), color = class)
```

Exemplo: consumo de combustível



Carros 2-seater tem
padrão diferente dos
demais. Outliers?

No R:

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy), color = class == "2seater")
```

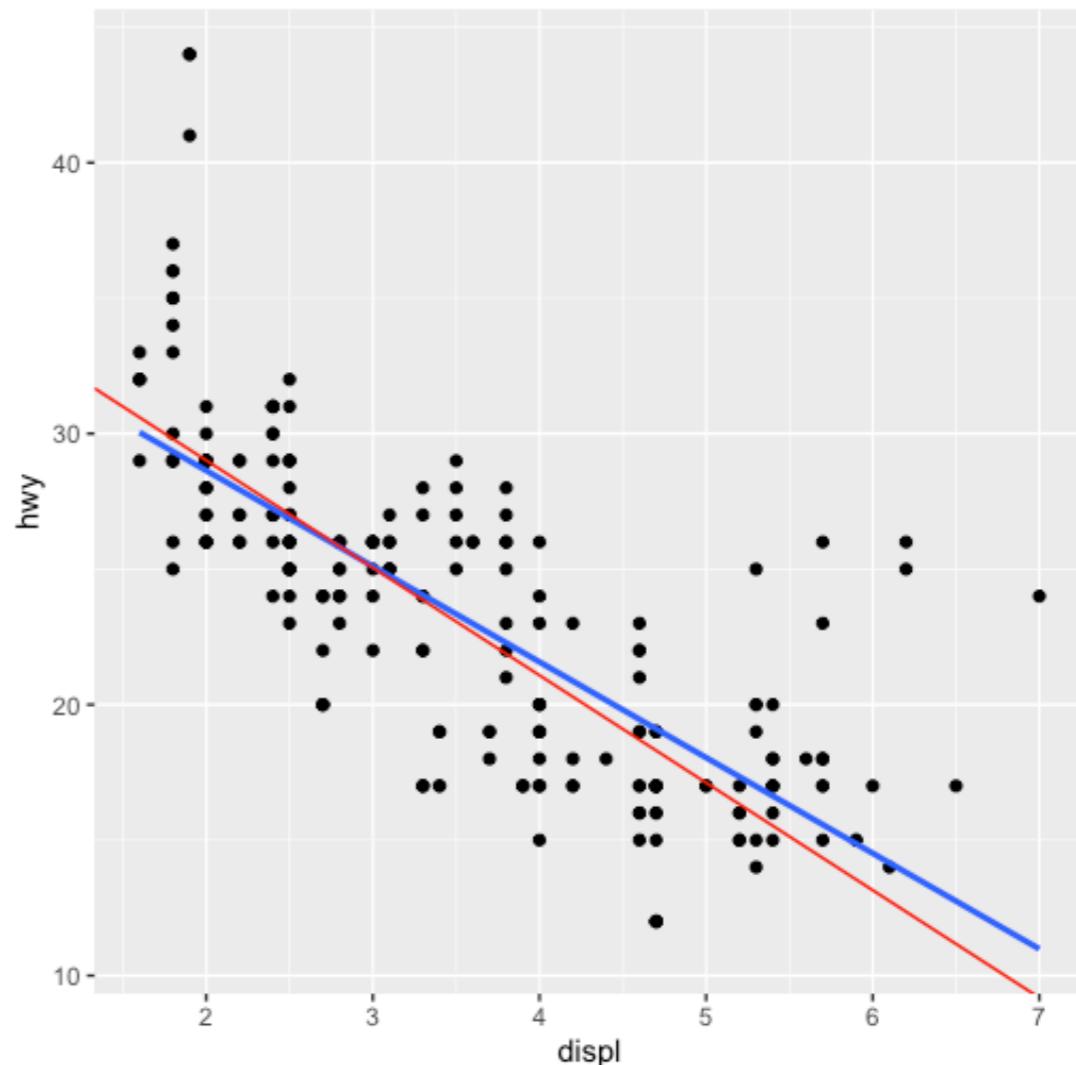
Exemplo: consumo de combustível

Ajuste de regressão excluindo os automóveis da classe “2-seater”.

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.9460    0.6718   54.99 <2e-16 ***
displ      -3.9658    0.1850  -21.43 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.466 on 227 degrees of freedom
Multiple R-squared:  0.6692, Adjusted R-squared:  0.6678
F-statistic: 459.3 on 1 and 227 DF,  p-value: < 2.2e-16
```

O R^2 aumentou, passando de 58.7% para 66.9%.

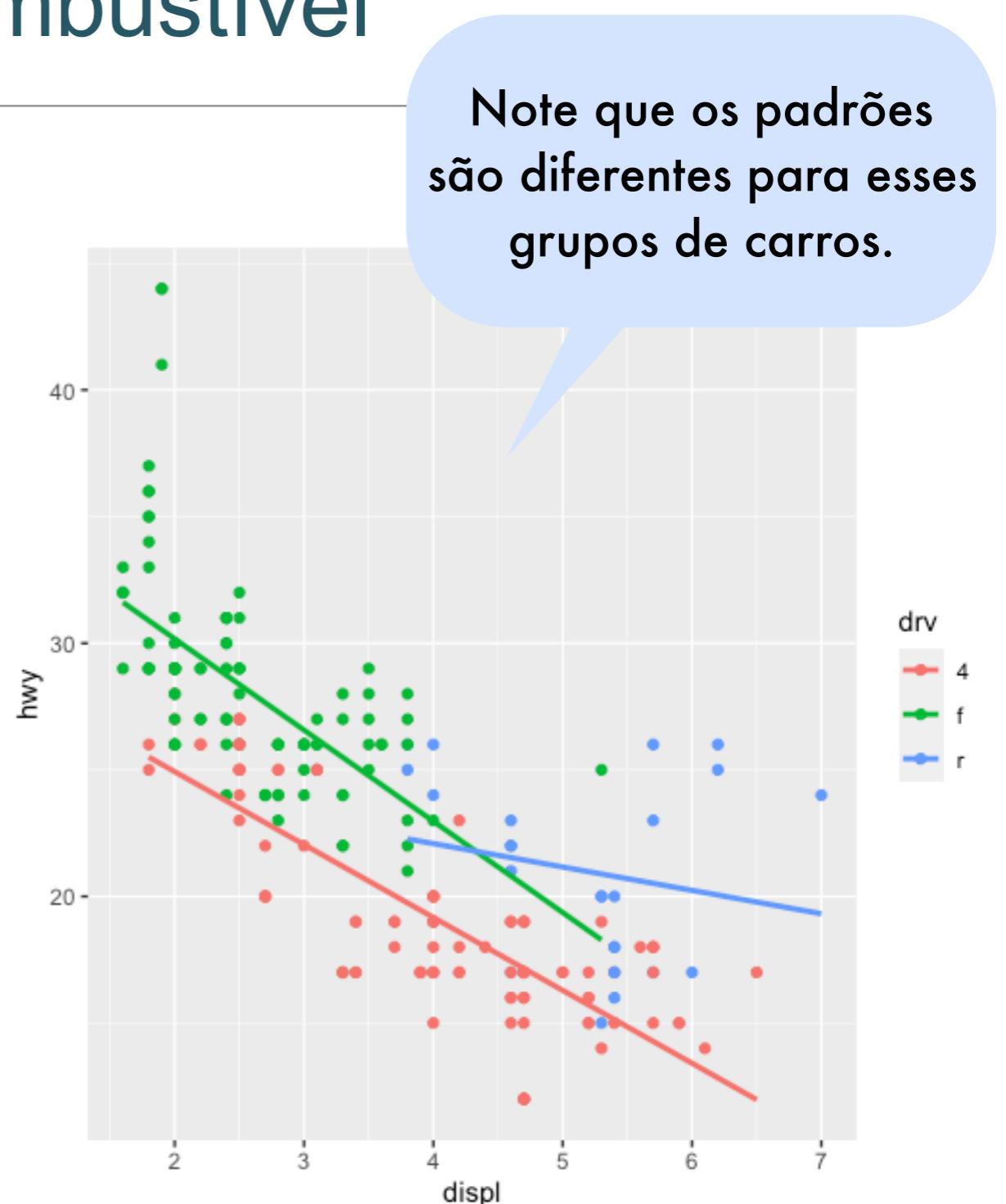


Exemplo: consumo de combustível

A variável `drv` tem categorias:

- 4 - four-wheel drive,
- f - front-wheel drive,
- r - rear-wheel drive.

O modelo de regressão linear múltipla poderá acomodar esses diferentes padrões.



Fontes de variabilidade

- Podemos procurar por variabilidade em cada variável de estudo ou entre variáveis sendo estudadas conjuntamente.
- No exemplo do consumo de combustível, padrões diferentes podem ser revelados ao olharmos as outras variáveis da base. Veremos essa análise mais a frente.

Apêndices

Apêndice I - Estimação por máxima verossimilhança

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2; y, x) &= f(y_1, \dots, y_n; x) \\ &= \prod_{i=1}^n f(y_i; x_i) \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 \right\} \end{aligned}$$

Para encontrar o EMV devemos maximizar a função

$$\log L(\beta_0, \beta_1, \sigma^2; y, x) = -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Os valores de $\beta_0, \beta_1, \sigma^2$ que maximizam a função acima são

Os EMV dos coeficientes da regressão são iguais aos estimadores de mínimos quadrados!

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

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

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}$$

Esse estimador para a variância é viciado!!

Apêndice II - Teste de hipóteses

Após obter os coeficientes estimados $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\sigma}^2$ devemos testar a significância da regressão.

Perguntas importantes:

1. O intercepto é significativo ou um modelo mais simples com $\beta_0 = 0$ seria mais adequado?
2. A variável x é realmente importante no modelo? Ou seja, a regressão é significativa? Neste caso, usar o modelo mais simples com apenas uma média global seria mais adequado?

Usaremos teste de hipóteses para responder essas perguntas.

Para a pergunta 1, as hipóteses seriam:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

Para a pergunta 2, as hipóteses seriam:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Teste t para o coeficiente angular

Vimos que

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{S_{xx}},$$

isto é, $\hat{\beta}_1$ é combinação linear dos y'_i s.

Como y_i tem distribuição normal, segue que $\hat{\beta}_1$ também tem distribuição normal.

Além disso, vimos que

$$E(\hat{\beta}_1) = \beta_1$$

e também que

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

Conclusão,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Teste t para o coeficiente angular

Temos que a variável padronizada tem distribuição normal padrão

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{S_{xx}}} \sim N(0, 1)$$

Podemos mostrar que

$$S = \frac{SS_{res}}{\sigma^2} \sim \chi^2_{(n-2)}$$

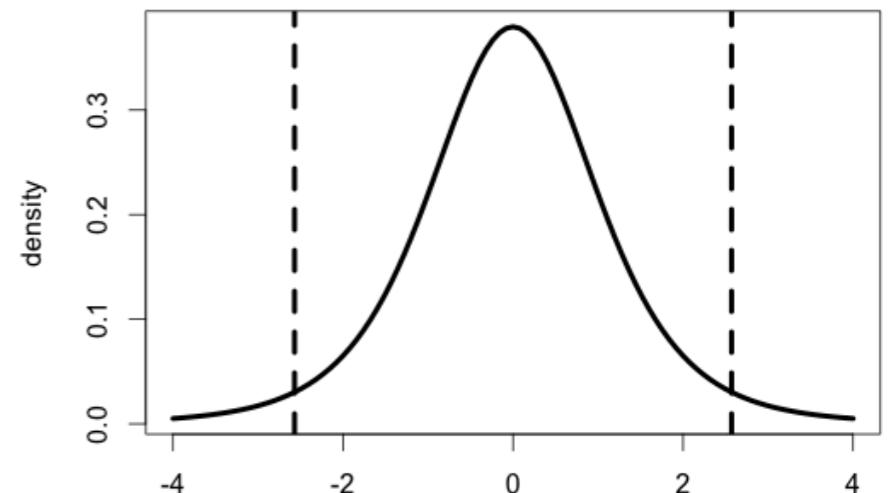
A estatística T é tal que

$$T = \frac{Z}{\sqrt{S/(n-2)}} \sim t_{n-2}.$$

Suponha que deseja-se testar

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Basta verificar se $|T_0| > qt(1 - \alpha/2, n - 2)$.



Rejeite H_0 se T_0 observado nas caudas!

Teste t para o intercepto

De forma análoga temos que

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma \sqrt{1/n + \bar{x}^2/S_{xx}}} \sim N(0, 1)$$

Podemos mostrar que

$$S = \frac{SS_{res}}{\sigma^2} \sim \chi^2_{(n-2)}$$

A estatística T é tal que

$$T = \frac{Z}{\sqrt{S/(n-2)}} \sim t_{n-2}.$$

Suponha que deseja-se testar

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

Basta verificar se $|T_0| > qt(1 - \alpha/2, n - 2)$.

Rejeite H_0 se T_0 observado nas caudas!

Apêndice III: Intervalos de confiança

Vimos que a estatística de teste para β_1 é

$$T = \frac{(\hat{\beta}_1 - \beta_1)/\sqrt{S_{xx}}}{\sqrt{SS_{res}/(n-2)}} \sim t_{n-2}.$$

Então podemos calcular

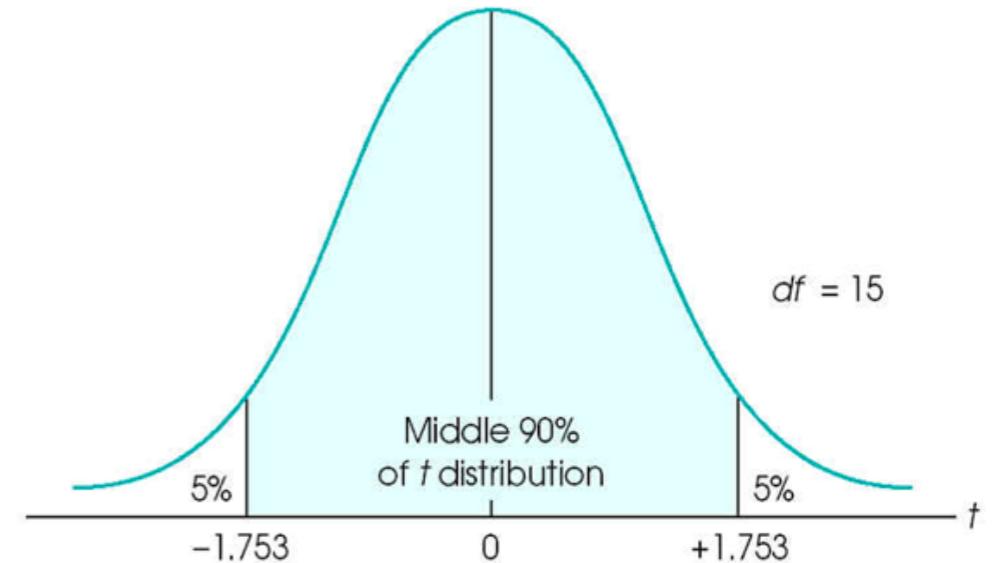
$$P(|T| \leq qt((1 + \gamma)/2, n - 2)) = \gamma.$$

Que resulta no intervalo

$$IC(\beta_1, \gamma) = \left(\hat{\beta}_1 - \sqrt{\frac{SS_{res}}{(n-2)S_{xx}}} qt^*, \hat{\beta}_1 + \sqrt{\frac{SS_{res}}{(n-2)S_{xx}}} qt^* \right),$$

onde

$$qt^* = qt((1 + \gamma)/2, n - 2).$$



De forma análoga podemos obter o intervalo para o intercepto!

Erro padrão.

Intervalo para a resposta média e previsão

- A resposta média é dada por $E(y | x_0) = \beta_0 + \beta_1 x_0$. E o estimador pontual para a resposta média é

$$\hat{E}(y | x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

- Uma nova observação prevista é $y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$. Cujo estimador pontual é

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

- Porém os intervalos são diferentes:

$$IC(E(y | x_0), \gamma) = \left(\hat{E}(y | x_0) - \sqrt{\frac{SS_{res}}{(n-2)} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} qt^*, \hat{E}(y | x_0) + \sqrt{\frac{SS_{res}}{(n-2)} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} qt^* \right)$$

$$IC(y_0, \gamma) = \left(\hat{y}_0 - \sqrt{\frac{SS_{res}}{(n-2)} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} qt^*, \hat{y}_0 + \sqrt{\frac{SS_{res}}{(n-2)} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} qt^* \right).$$

Maior erro padrão.