Preparado para: AT autoridade tributária e aduaneira

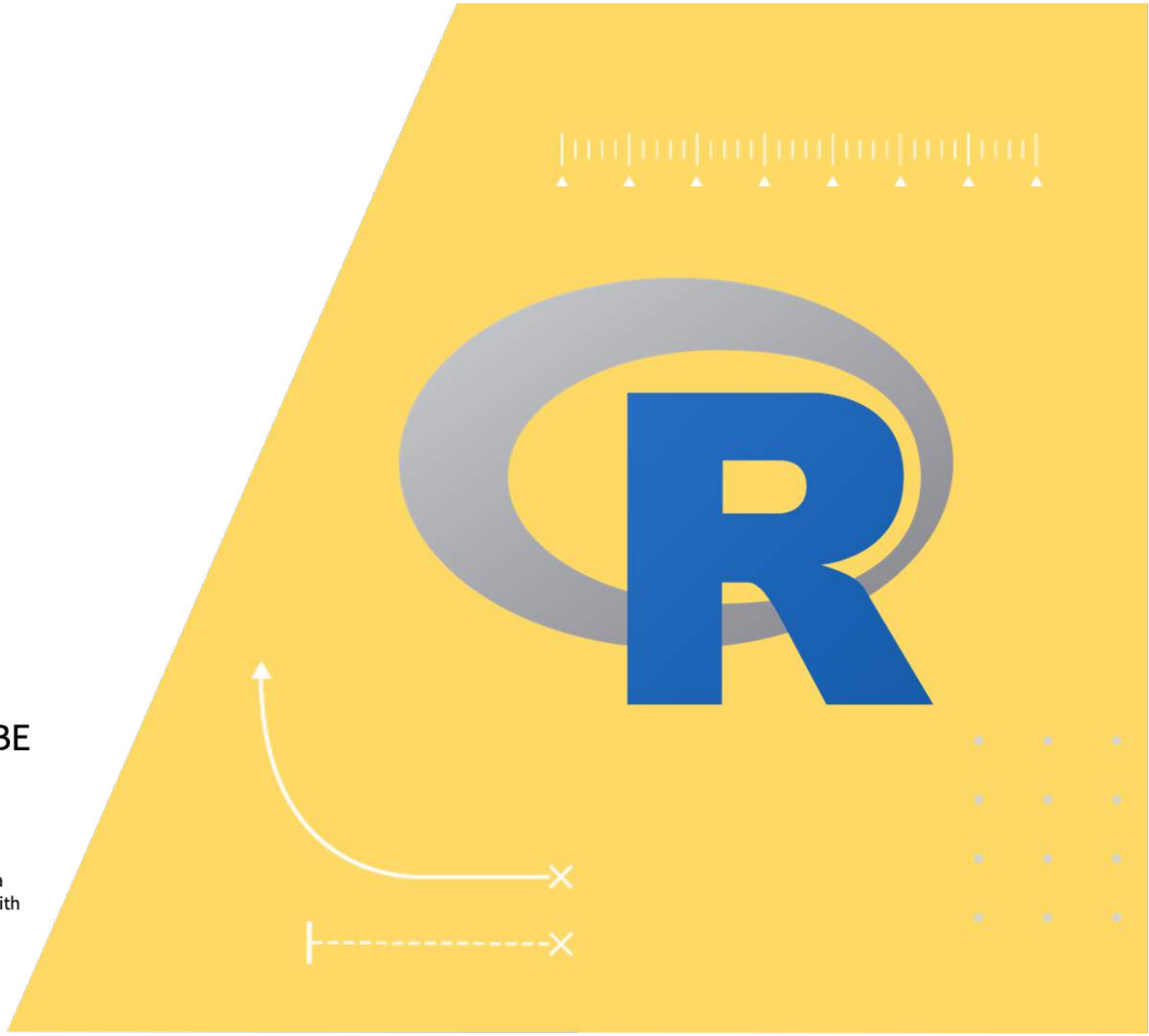
REFORM/SC2022/126
DELIVERABLE 4
MÓDULO 4
REGRESSÃO LINEAR

DESIGNING A NEW VALUATION MODEL
FOR RURAL PROPERTIES IN PORTUGAL

Parte I

Formador: Luís Teles Morais | Nova SBE

Lisboa, 27 junho 2023



This project is carried out with funding by the European Union via the Structural Reform Support Programme and in cooperation with the Directorate General for Structural Reform Support of the European Commission

AARC NOVA SCHOOL OF BUSINESS & ECONOMICS LOBO VASQUES esri INNERLANDS

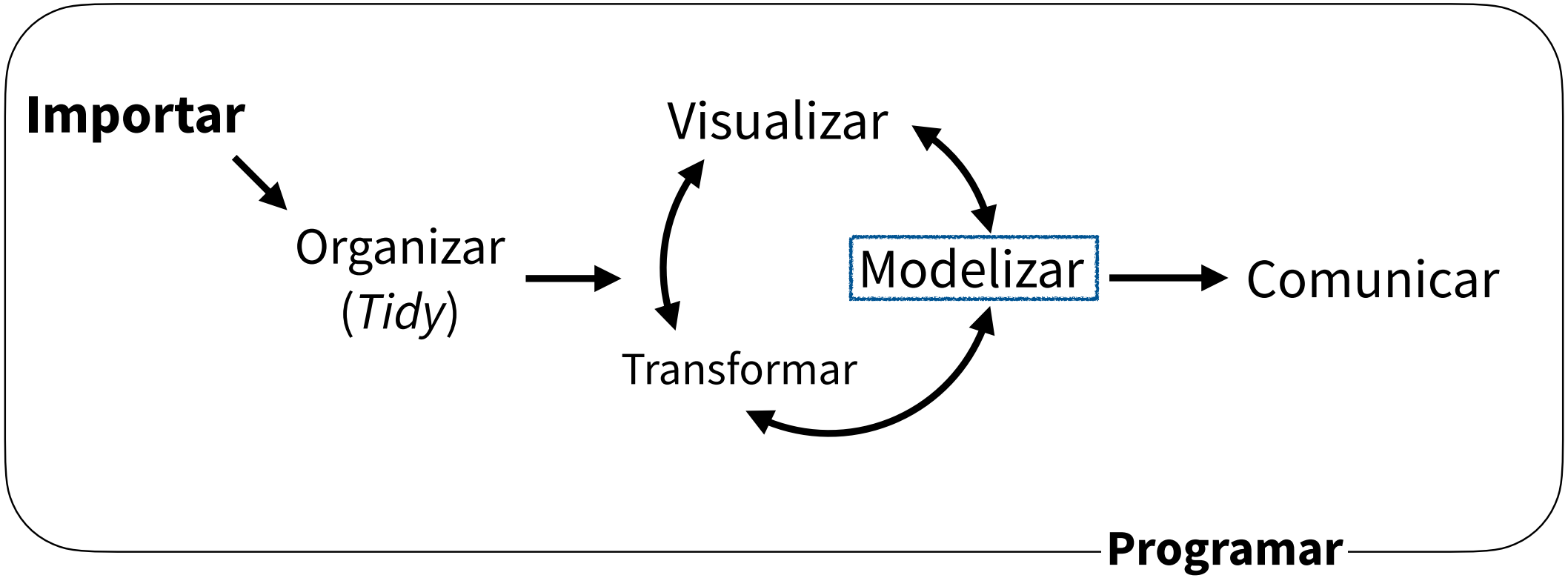
Quiz 3

Programa

Módulos	Duração
Módulo 1 – Introdução ao R: - O que é o R? - Como instalar e configurar o R. - Sintaxe básica e comandos. - Tipos de dados, objetos e classes.	**4 Horas**
Módulo 2 – Gestão e tratamento de dados em R: - Carregar dados no R. - Perceber as estruturas de dados e *subsetting*. - Limpeza de dados: *missing values, outliers* e transformações - Juntar bases de dados	**8 Horas**
Módulo 3 – Estatística básica em R: - Estatísticas descritivas: medidas de dispersão central e variação. - Distribuições probabilísticas: variáveis discretas e contínuas. - Testes de hipóteses.	**8 Horas**

Módulos	Duração
Módulo 4 – Regressão Linear: - O modelo classico linear. - Estimação de parametros segundo o MMQ. - Testes de hipóteses: significância estatística e ajuste do modelo. - Modelo de regressão múltipla. - Testar as premissas: multicolinearidade, heteroscedasticidade e normalidade dos resíduos. - Critérios de seleção dos modelos.	**12 Horas**
Módulo 5 – O modelo: - Estrutura do modelo e premissas – Perceber o modelo (4 Hours). - Uso e tratamento dos dados (4 Hours). - Descrição do modelo (4 Hours). - Aplicação do modelo a cada piloto (12 Hours). - Aplicação autónoma do modelo a uma região (8 Hours).	**32 Horas**

Ciência de dados



Alguns conceitos

Visualizar

Estatística **descritiva**

- Analisar um conjunto de dados, reduzindo-o a medidas sumárias simples

Modelizar

Inferência estatística

- Calcular ou estimar algo que não podemos observar diretamente, a partir dos dados existentes

O que é um modelo?

O que é um **modelo**?

Modelos lineares

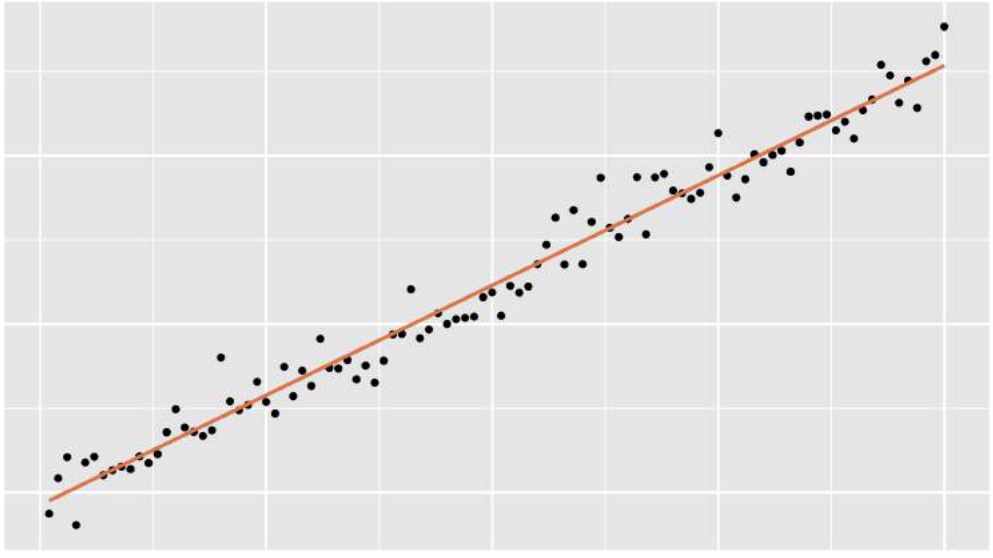
- Usamos **modelos** para analisar a relação entre diversas variáveis (aleatórias).
  Objetivo:
  - realizar previsões (através de inferência estatística)...
  - sobre aspetos da realidade desconhecidos (parâmetros)...
  - a partir de informação conhecida (amostras).

Modelos lineares

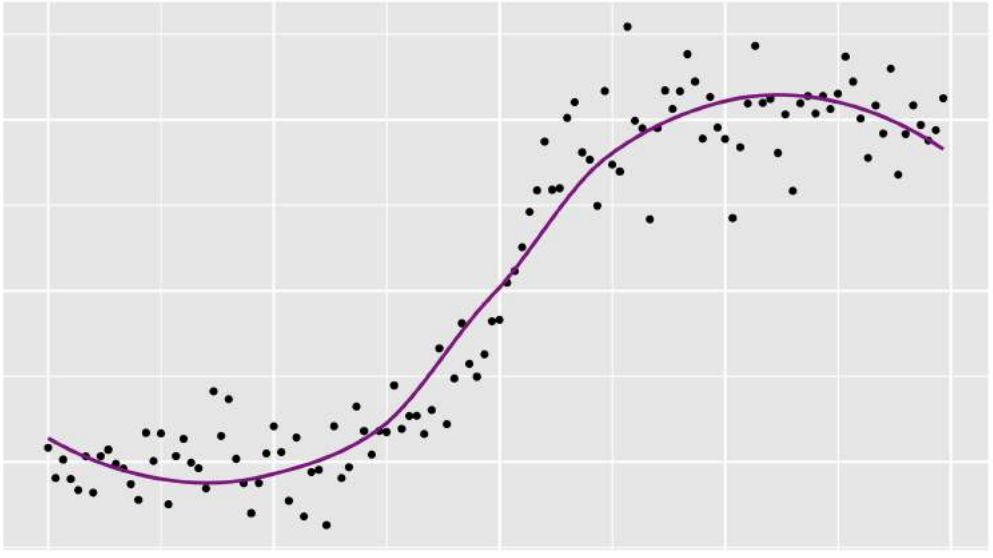
- Usamos **modelos** para analisar a relação entre diversas variáveis (aleatórias). Objetivo:
  - realizar previsões (através de inferência estatística)...
  - sobre aspetos da realidade desconhecidos (parâmetros)...
  - a partir de informação conhecida (amostras).
- Aqui vamos ater-nos aos modelos lineares, ou de **regressão linear**
- Ter presente que a realidade muitas vezes não é linear...
  - Linearidade: hipótese simplificadora

Modelos lineares

Linear



Não-linear



Dados: quadros de Paris

Leilões em Paris no séc. XVIII

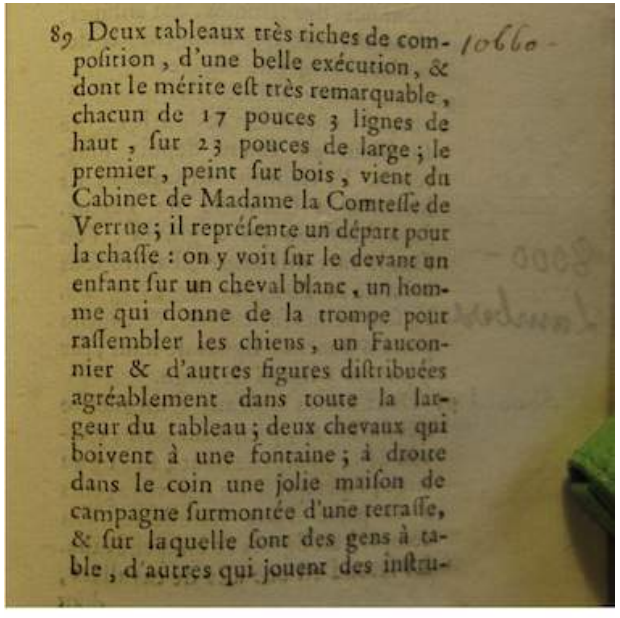
Pierre-Antoine de Machy, Public Sale at the Hôtel Bullion, Musée Carnavalet, Paris (séc. XVIII)

paris-paintings.xlsx

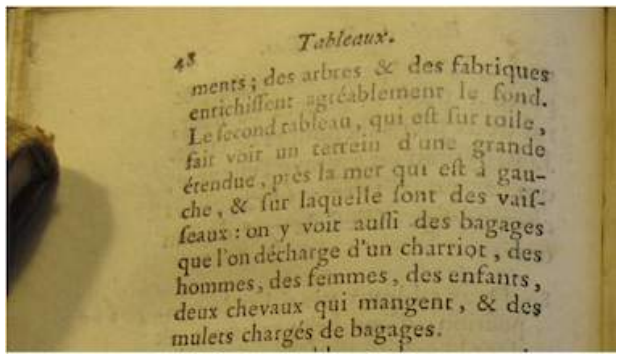
- Fonte: Catálogos impressos de 28 leilões de arte em Paris, 1764 - 1780
- Créditos: Sandra van Ginhoven and Hilary Coe Cronheim (U. Duke)

Départ pour la chasse

Fonte primária dos dados



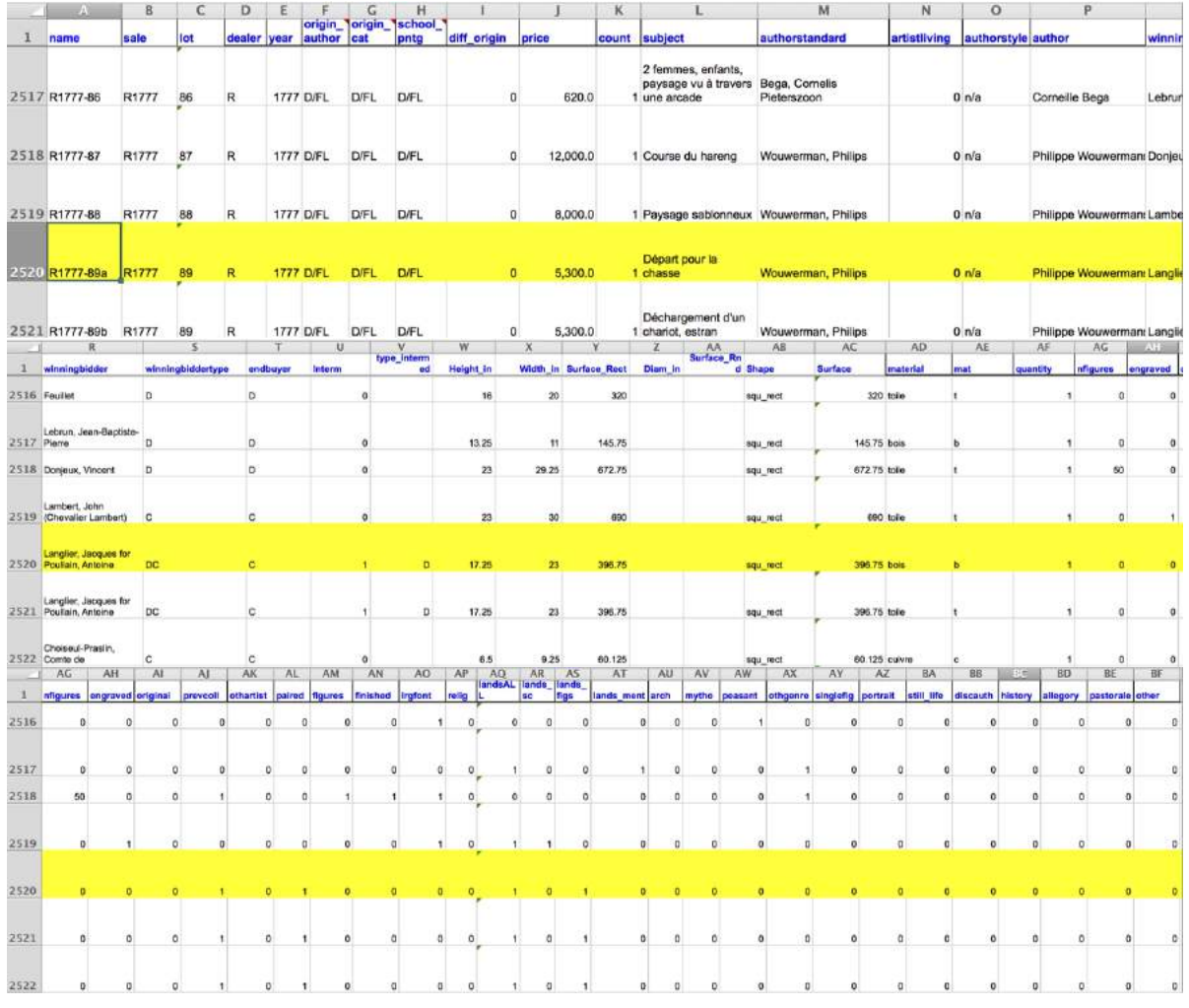
89 Deux tableaux très riches de compoſition, d'une belle exécution, & dont le mérite eſt très remarquable, chacun de 17 pouces 3 lignes de haut, ſur 23 pouces de large; le premier, peint ſur bois, vient du Cabinet de Madame la Comteſſe de Verrue; il repréſente un départ pour la chaſſe : on y voit ſur le devant un enfant ſur un cheval blanc, un homme qui donne de la trompe pour raſſembler les chiens, un Fauconnier & d'autres figures diſtribuées agréablement dans toute la largeur du tableau; deux chevaux qui boivent à une fontaine; à droite dans le coin une jolie maiſon de campagne ſurmontée d'une terraſſe, & ſur laquelle ſont des gens à table, d'autres qui jouent des inſtru-



Tableaux.

ments; des arbres & des fabriques enrichiſſent agréablement le fond. Le ſecond tableau, qui eſt ſur toile, fait voir un terrein d'une grande étendue, près la mer qui eſt à gauche, & ſur laquelle ſont des vaiſſeaux : on y voit auſſi des bagages que l'on décharge d'un charriot, des hommes, des femmes, des enfants, deux chevaux qui mangent, & des mulets chargés de bagages.

Duas pinturas de composição muito rica, de bela execução, e cujo mérito é notável, cada uma com 17 polegadas e 3 linhas de altura, 23 polegadas de largura; o primeiro, pintado em madeira, provém do acervo de Madame la Comtesse de Verrue; representa uma **partida para a caça**: mostra à frente uma criança num cavalo branco, um homem que toca a corneta para reunir os cães, um falcoeiro e outras figuras bem distribuídas pela largura da pintura; dois cavalos bebendo de uma fonte; à direita, ao canto, uma bela casa de campo encimada por um terraço, onde estão pessoas à mesa, outras que tocam instrumentos; árvores e tecidos enriquecem agradavelmente o fundo.



	name	sale	lot	dealer	year	origin_author	origin_cat	school_pntg	diff_origin	price	count	subject	authorstandard	artistliving	authorstyle	author	winnir
2517	R1777-86	R1777	86	R	1777	D/FL	D/FL	D/FL	0	620.0	1	2 femmes, enfants, paysage vu à travers une arcade	Bega, Cornelis Pieterszoon	0	n/a	Cornelle Bega	Lebrur
2518	R1777-87	R1777	87	R	1777	D/FL	D/FL	D/FL	0	12,000.0	1	Course du hareng	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Donjeu
2519	R1777-88	R1777	88	R	1777	D/FL	D/FL	D/FL	0	8,000.0	1	Paysage sablonneux	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Lambe
2520	R1777-89a	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Départ pour la chasse	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langli
2521	R1777-89b	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Déchargement d'un chariot, estran	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langli

	winningbidder	winningbiddertype	endbuyer	interm	type_interm ed	Height_in	Width_in	Surface_Rect	Diam_in	Surface_Rnd	Shape	Surface	material	mat	quantity	nfigures	engraved
2516	Feuillet	D	D	0		16	20	320			squ_rect	320	toile	t	1	0	0
2517	Lebrun, Jean-Baptiste-Pierre	D	D	0		13.25	11	145.75			squ_rect	145.75	bois	b	1	0	0
2518	Donjeux, Vincent	D	D	0		23	29.25	672.75			squ_rect	672.75	toile	t	1	50	0
2519	Lambert, John (Chevalier Lambert)	C	C	0		23	30	690			squ_rect	690	toile	t	1	0	1
2520	Langlier, Jacques for Poullain, Antoine	DC	C	1	D	17.25	23	396.75			squ_rect	396.75	bois	b	1	0	0
2521	Langlier, Jacques for Poullain, Antoine	DC	C	1	D	17.25	23	396.75			squ_rect	396.75	toile	t	1	0	0
2522	Choiseul-Praslin, Comte de	C	C	0		6.5	9.25	60.125			squ_rect	60.125	cuivre	c	1	0	0

	nfigures	engraved	original	prevcoll	othartist	paired	figures	finished	lrgfont	relig	landsALL	landsc	landsfigs	lands_ment	arch	mythe	peasant	othgenre	singlefig	portrait	still_life	discauth	history	allegory	pastorale	other
2516	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2517	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2518	50	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2519	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2520	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2521	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2522	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

- $>$ 3000 quadros, preços, e detalhes digitalizados $\rightarrow$ $>$ 60 variáveis

readxl::read_excel()

Importar dados a partir de Excel

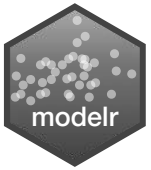
```
read_excel("ficheiro.xlsx", sheet = "data", …)
```

Nome de caminho do ficheiro a importar
A partir do *working directory*

```
getwd()
```

Nome de folha dentro do livro Excel

Outras opções



Importar os dados

- Importe os dados na folha **data** do livro Excel **paris-paintings.xlsx**:
  - Para um objeto com o nome **pp**
  - Garantindo que quaisquer destes: "n/a", "", "NA" é interpretado corretamente como valor NA
- Experimente usar os menus do R primeiro...
- ... e depois usando a função **read_excel**. Qual é a forma mais rápida?

Importe os dados

```
pp <- readxl::read_excel("data/paris-paintings.xlsx", sheet = "data",
                         na = c("n/a", "", "NA"))

pp
```

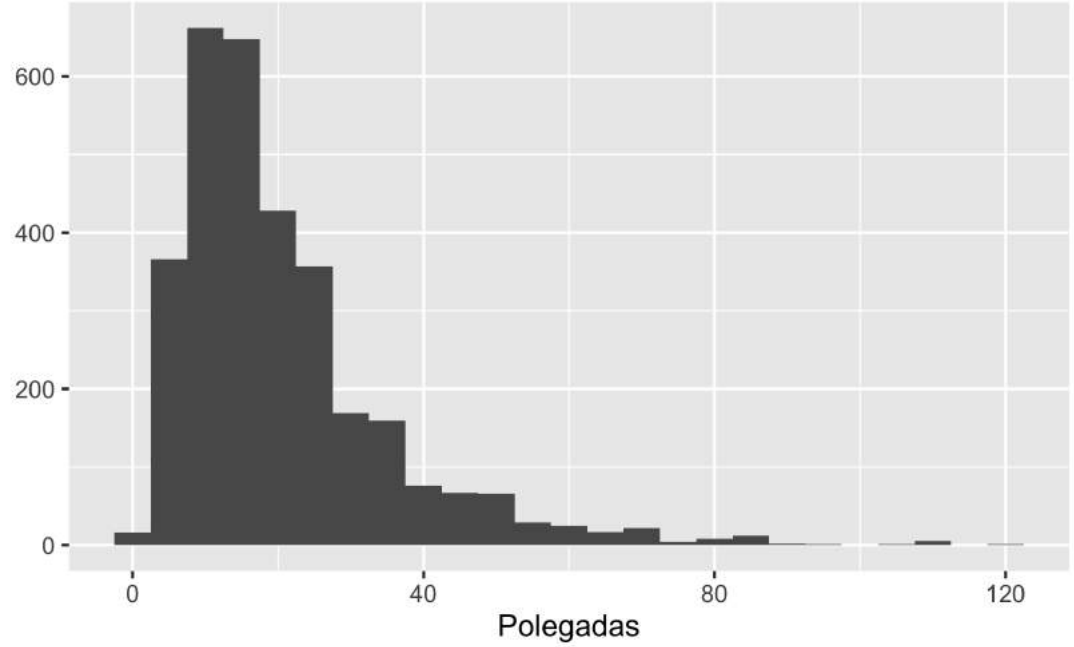
```
## # A tibble: 3,393 × 61
##   name     sale  lot   position dealer  year origin_author
##   <chr>    <chr> <chr>    <dbl> <chr>  <dbl> <chr>
## 1 L1764-2  L1764 2       0.0328 L       1764 F
## 2 L1764-3  L1764 3       0.0492 L       1764 I
## 3 L1764-4  L1764 4       0.0656 L       1764 X
## 4 L1764-5a L1764 5       0.0820 L       1764 F
## 5 L1764-5b L1764 5       0.0820 L       1764 F
## 6 L1764-6  L1764 6       0.0984 L       1764 X
## # ℹ 3,387 more rows
## # ℹ 54 more variables: origin_cat <chr>, school_pntg <chr>,
## #   diff_origin <dbl>, logprice <dbl>, price <chr>, count <dbl>,
## #   subject <chr>, authorstandard <chr>, artistliving <dbl>,
## #   authorstyle <chr>, author <chr>, winningbidder <chr>,
## #   winningbiddertype <chr>, endbuyer <chr>, Interm <dbl>,
## #   type_intermed <chr>, Height_in <dbl>, Width_in <dbl>, …
```

Modelo: largura x altura?

Distribuições (univariadas)

Altura

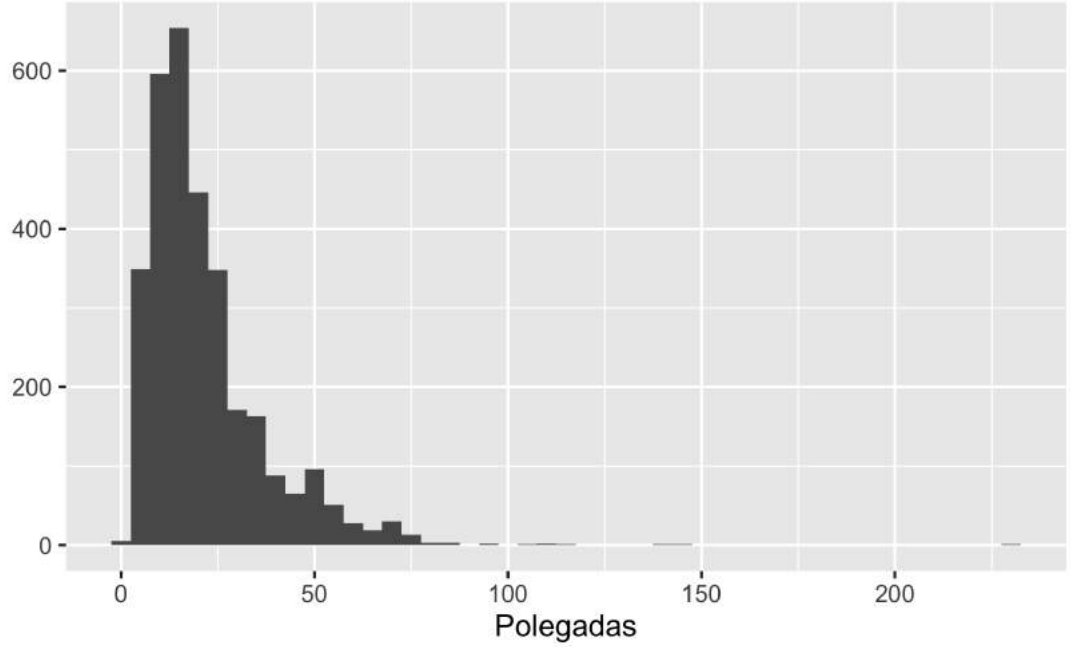
```
ggplot(data = pp, aes(x = Height_in)) +
  geom_histogram(binwidth = 5) +
  labs(x = "Polegadas", y = NULL)
```



Distribuições marginais (univariadas)

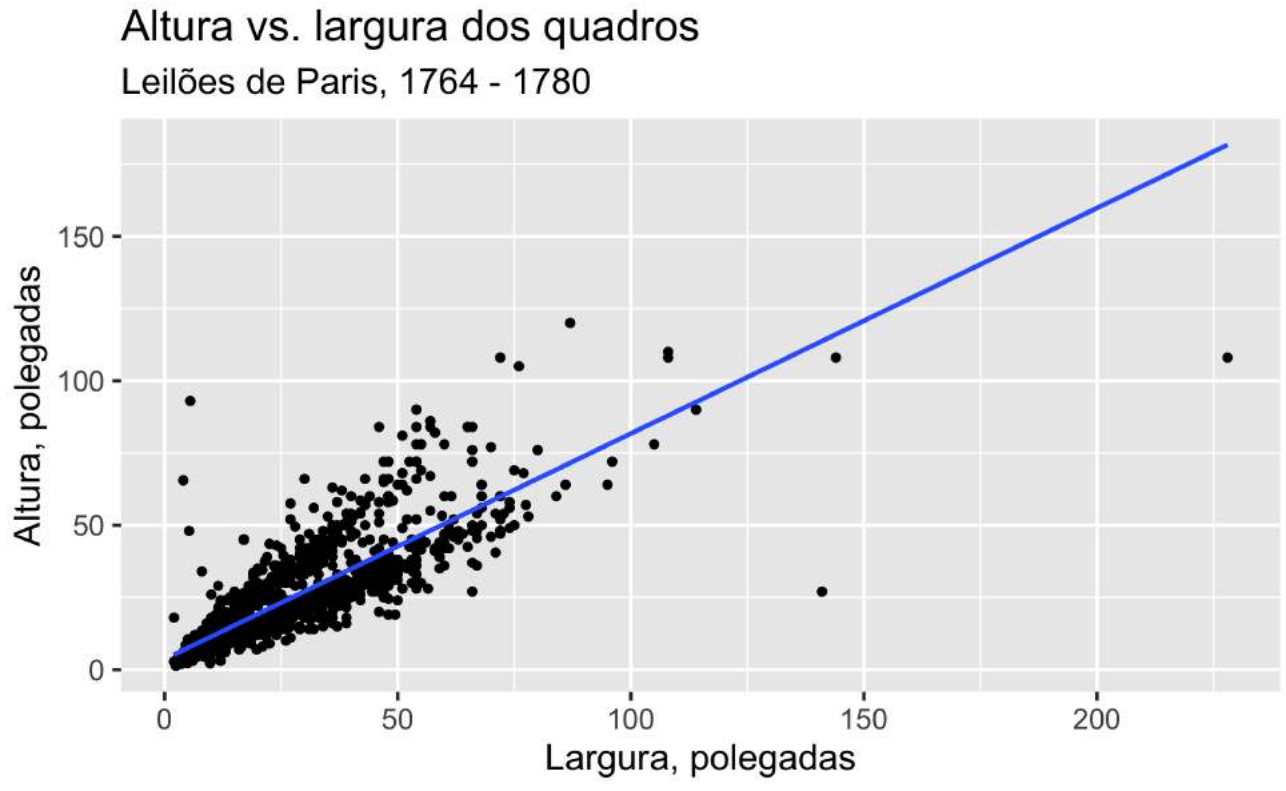
Largura

```
ggplot(data = pp, aes(x = Width_in)) +
  geom_histogram(binwidth = 5) +
  labs(x = "Polegadas", y = NULL)
```



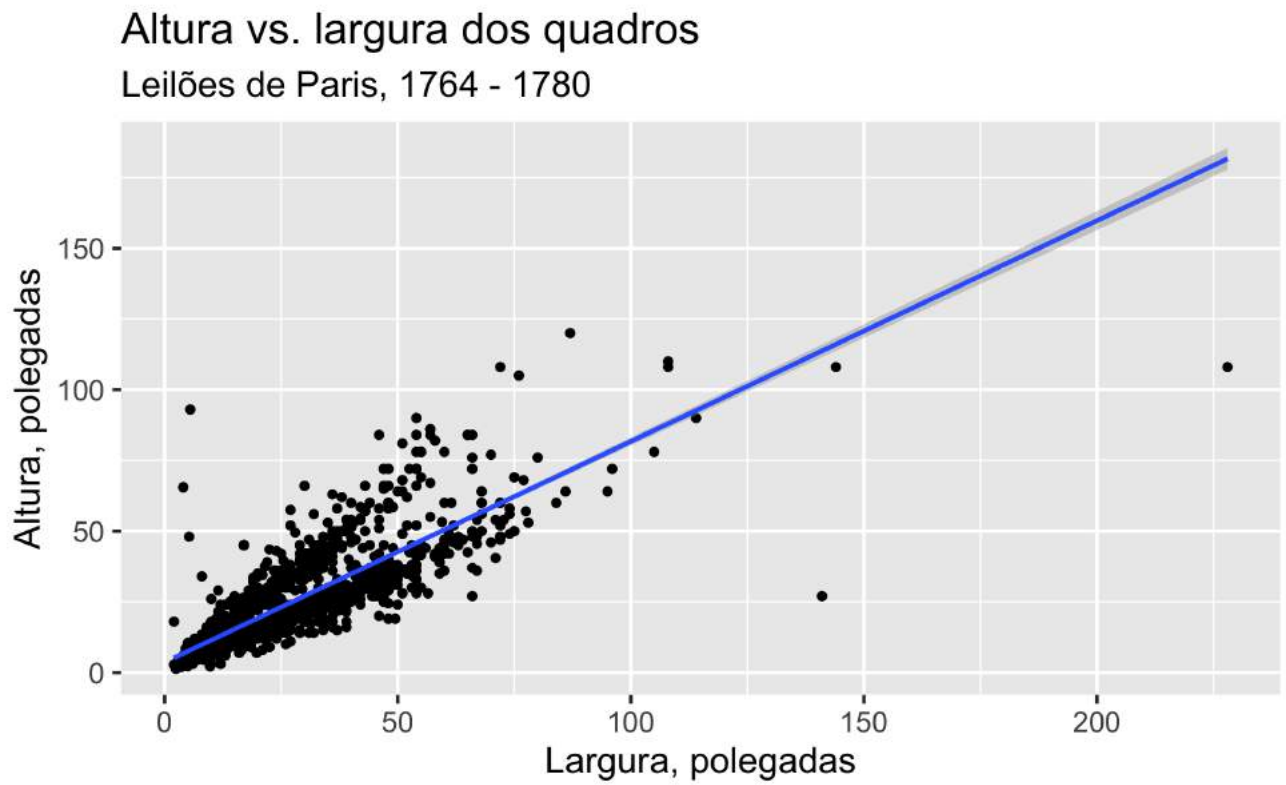
Altura vs largura (multivariada)

Plot Code



... com uma medida de incerteza

Plot Code



- Intervalo de confiança

Vocabulário

- **Variável dependente** ou **resposta** Variável cujo comportamento queremos entender / variabilidade queremos explicar a partir de outra(s) -- eixo **yy**

Vocabulário

- **Variável dependente** ou **resposta** Variável cujo comportamento queremos entender / variabilidade queremos explicar a partir de outra(s) -- eixo **yy**
- **Variáveis independentes** ou **explicativas** Outras váriaveis que utilizamos para explicar o comportamento da variável dependente -- eixo **xx**

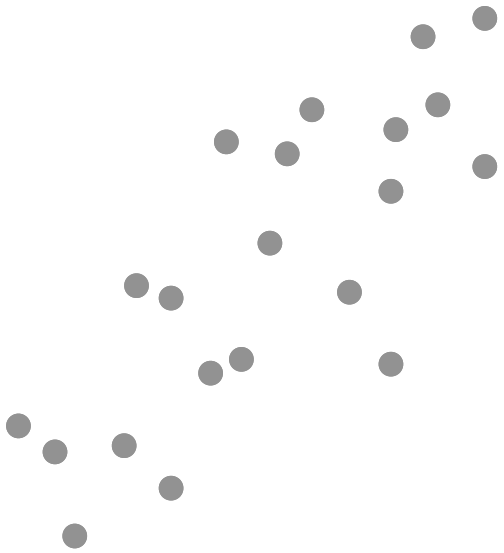
Vocabulário

- **Variável dependente** ou **resposta** Variável cujo comportamento queremos entender / variabilidade queremos explicar a partir de outra(s) -- eixo **yy**
- **Variáveis independentes** ou **explicativas** Outras váriaveis que utilizamos para explicar o comportamento da variável dependente -- eixo **xx**
- **Valor estimado** ou previsto ou ajustado ($\hat{y}$): o output do **modelo**
  - O modelo dá o valor médio (ou esperado) da váriavel dependente, *condicional*, i.e. para um determinado valor, da variável independente

Vocabulário

- **Variável dependente** ou **resposta** Variável cujo comportamento queremos entender / variabilidade queremos explicar a partir de outra(s) -- eixo **yy**
- **Variáveis independentes** ou **explicativas** Outras váriaveis que utilizamos para explicar o comportamento da variável dependente -- eixo **xx**
- **Valor estimado** ou previsto ou ajustado ($\hat{y}$): o output do **modelo**
  - O modelo dá o valor médio (ou esperado) da váriavel dependente, *condicional*, i.e. para um determinado valor, da variável independente
  - **Resíduo**: Mede a distância entre um valor observado (numa amostra) e o valor estimado (com base num determinado modelo)
  - Resíduo = Valor observado - Valor estimado
  - Indica quão próximo está o modelo de "acertar" num determinado ponto dos dados, ou por outra, por quanto é que o modelo "falha"

Modelos



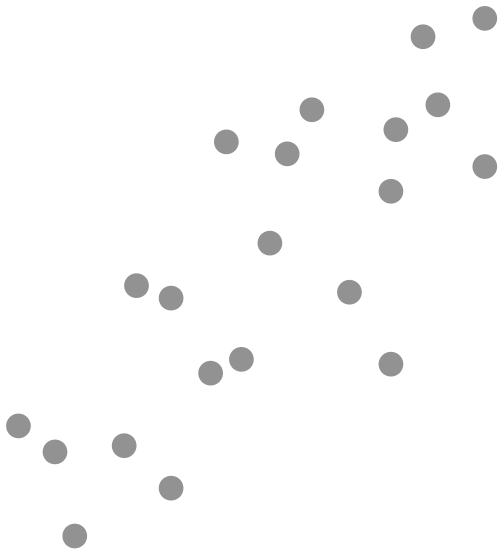
Estimação em R

Dados

Model Function

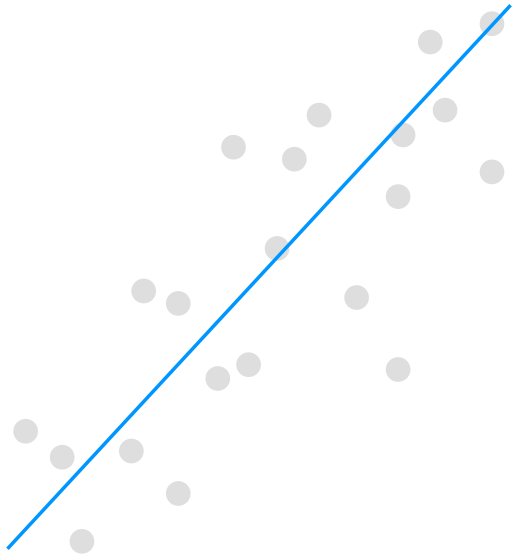
Modelos

Qual é o **modelo** que parece melhor descrever os dados?



Dados

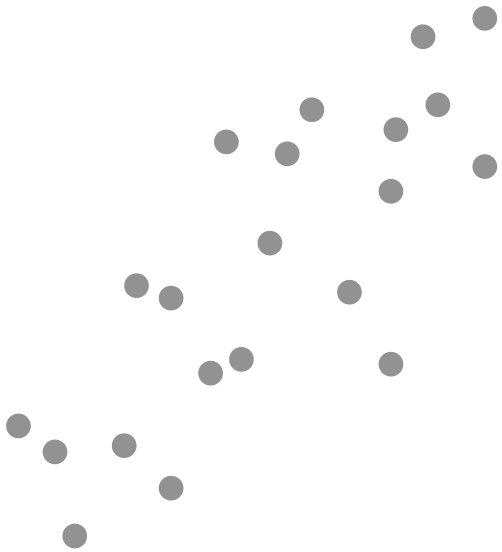
Estimação em R



Model Function

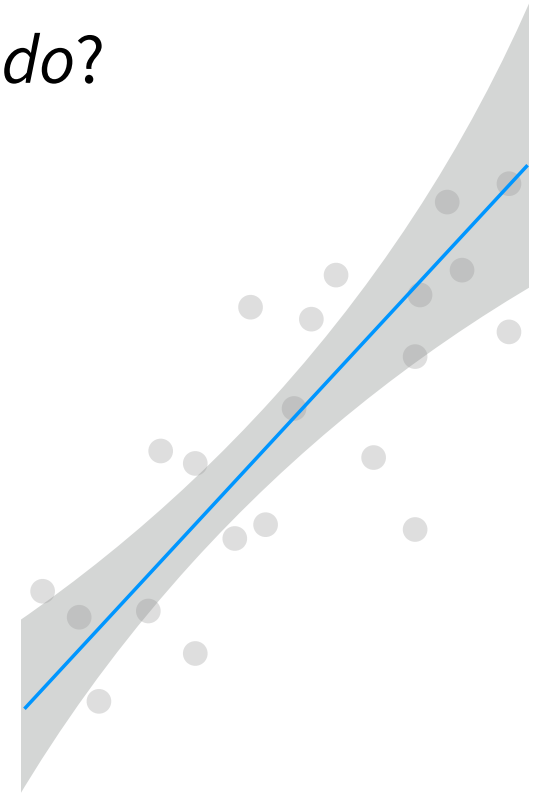
Modelos

Qual é a **incerteza** associada ao modelo *estimado*?



Dados

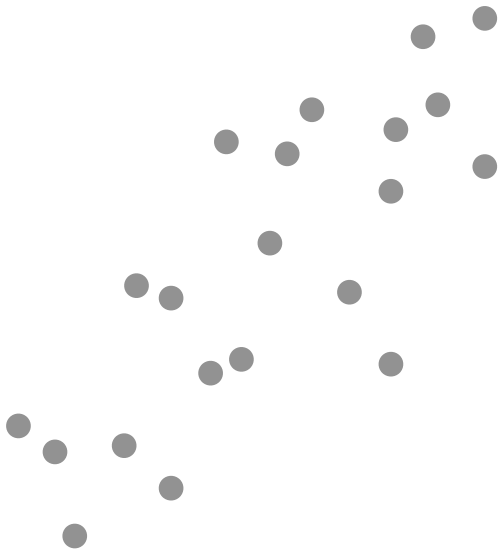
Estimação em R



Model Function

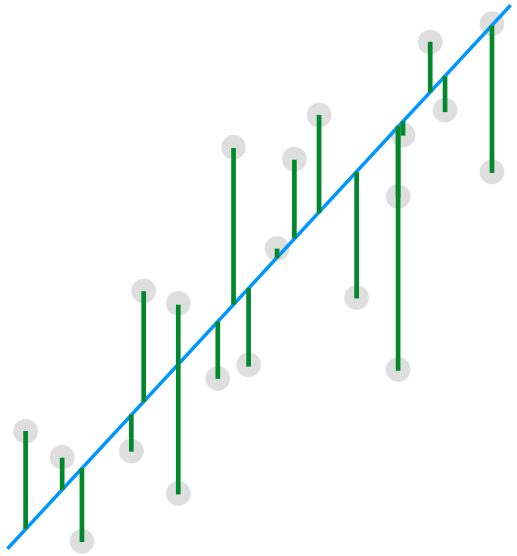
Modelos

Que parte da variação é que o modelo consegue ou *não* explicar => quais os **resíduos** do modelo?



Dados

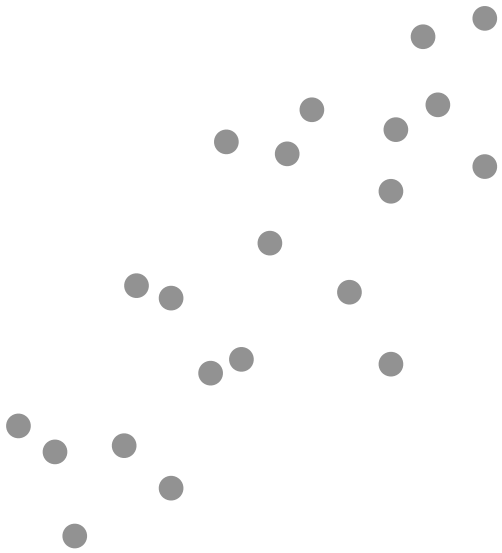
Estimação em R



Model Function

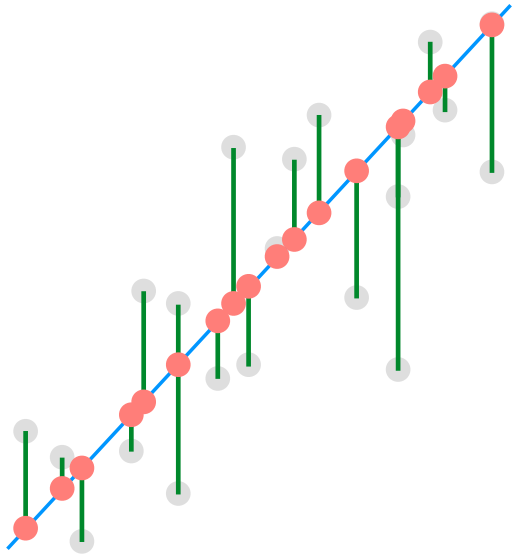
Modelos

Que parte da variação é que o modelo consegue ou *não* explicar => quais os **valores estimados**?



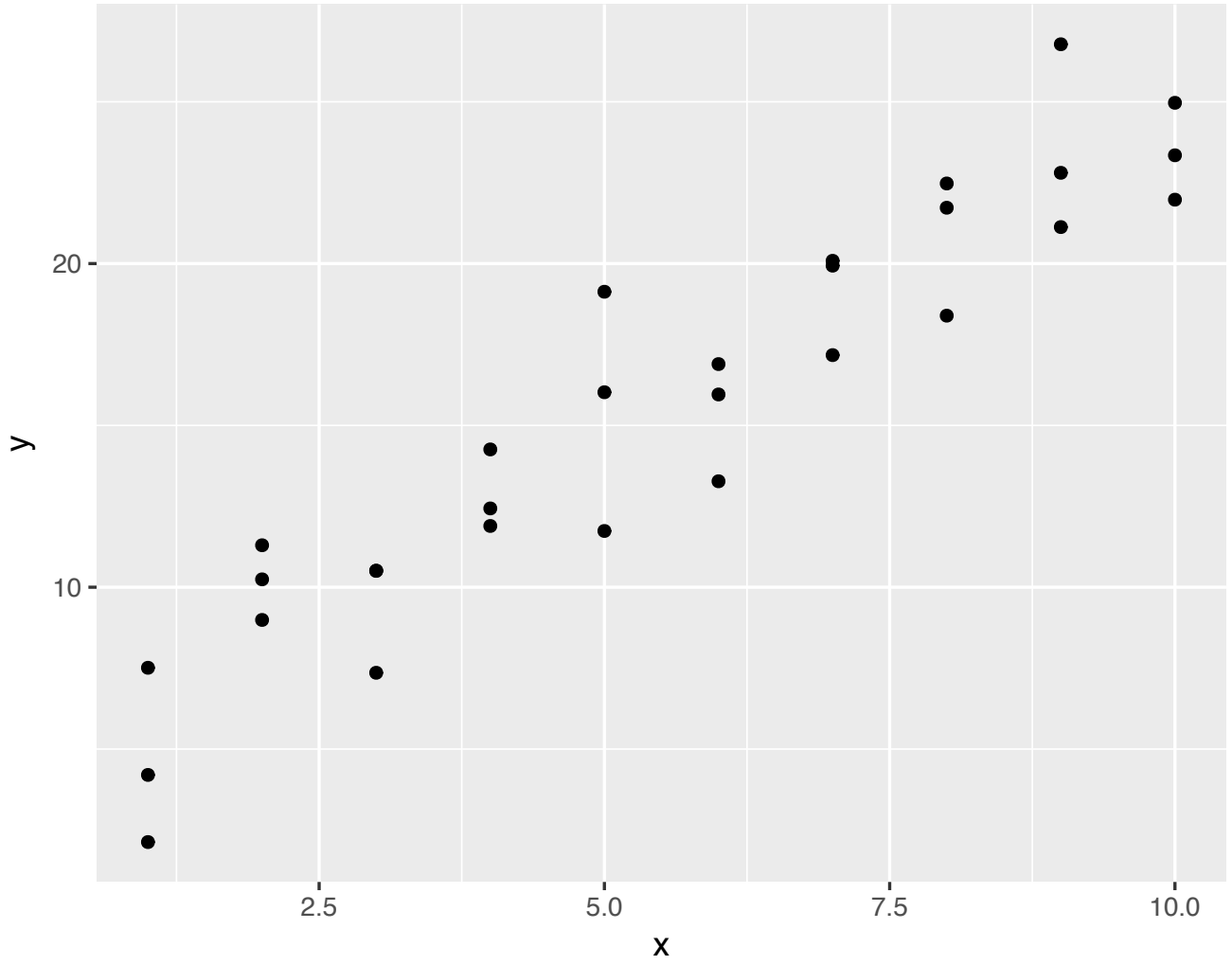
Dados

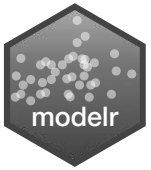
Estimação em R



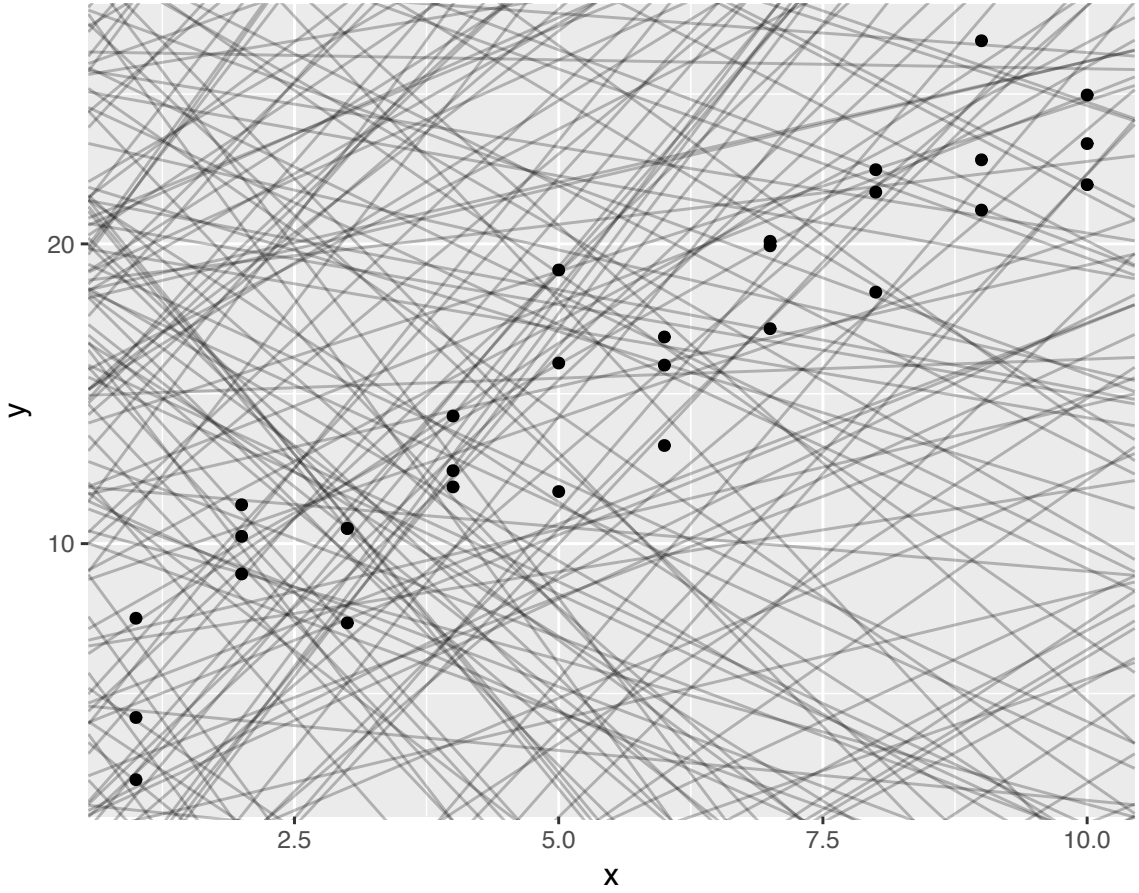
Model Function

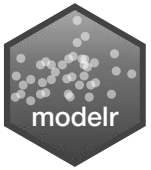
Exemplo visual



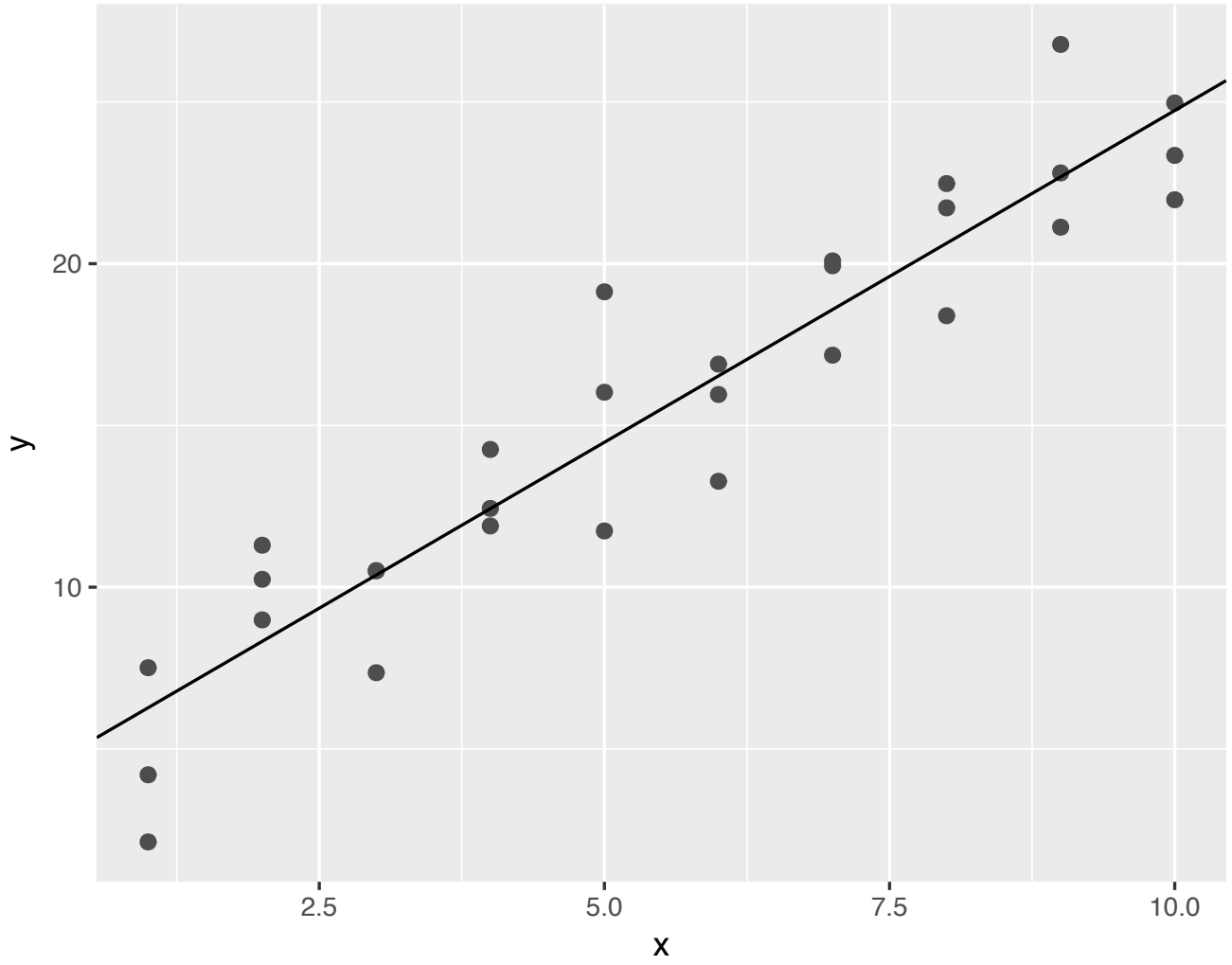


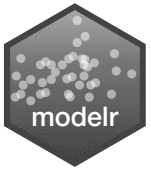
Existem infinitas *funções* possíveis para modelos estatísticos com estes dados



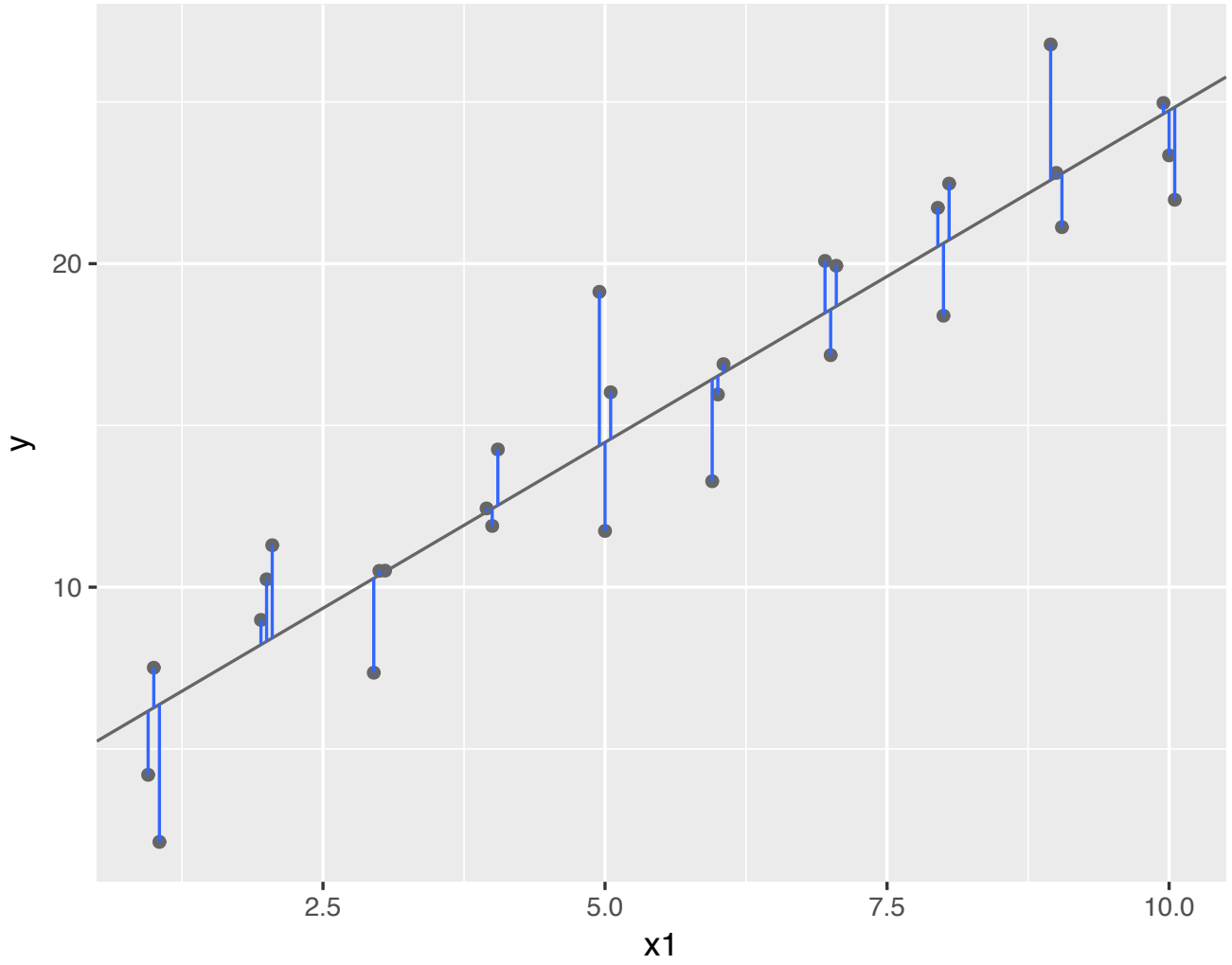


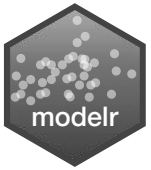
O modelo mais ajustado aos dados





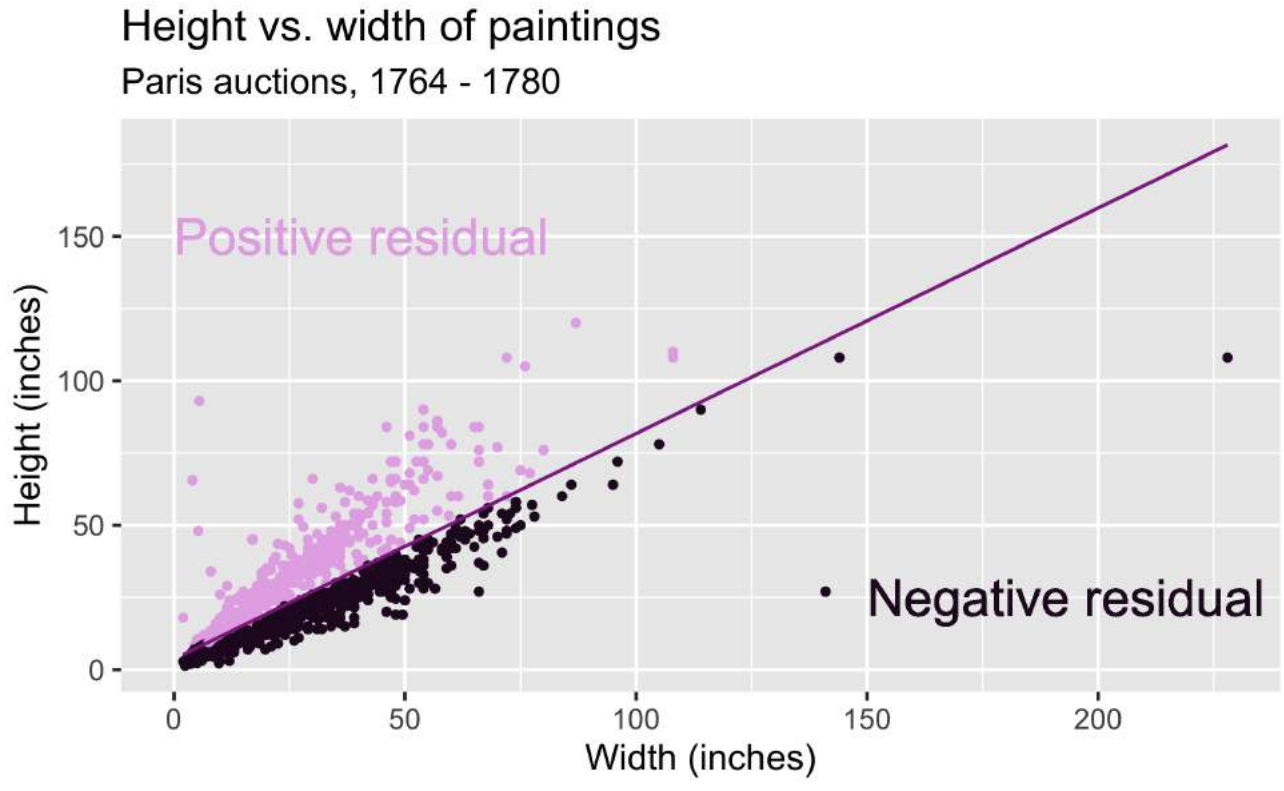
Os resíduos



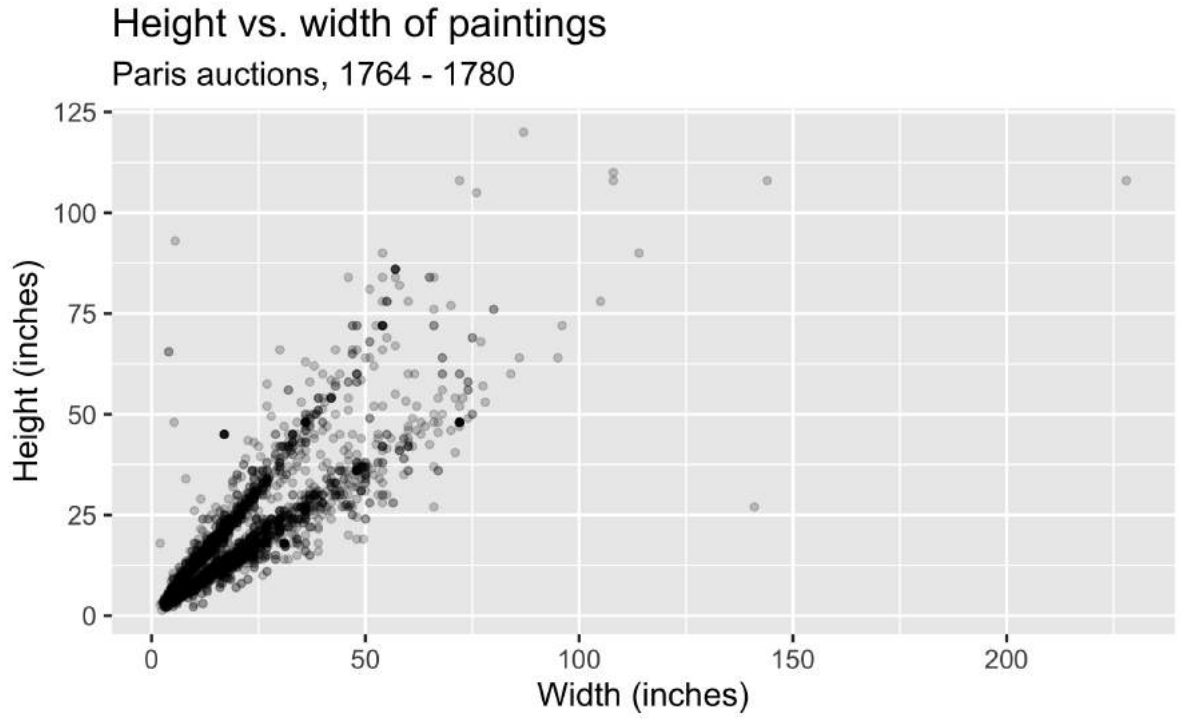


Resíduos

Plot Code



- O que mudou no gráfico?
- Que padrão nos dados se tornou agora aparente?
- O que pode significar, em termos estatísticos? E na realidade?

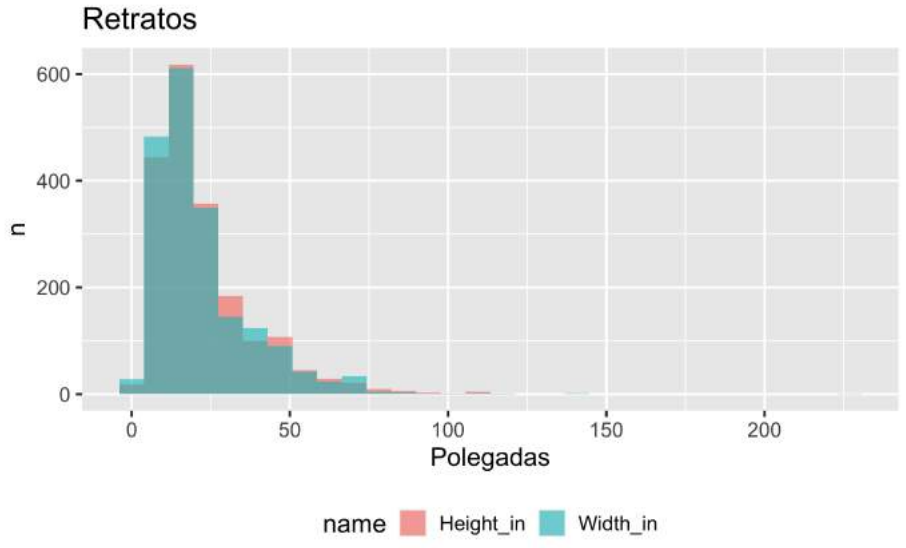


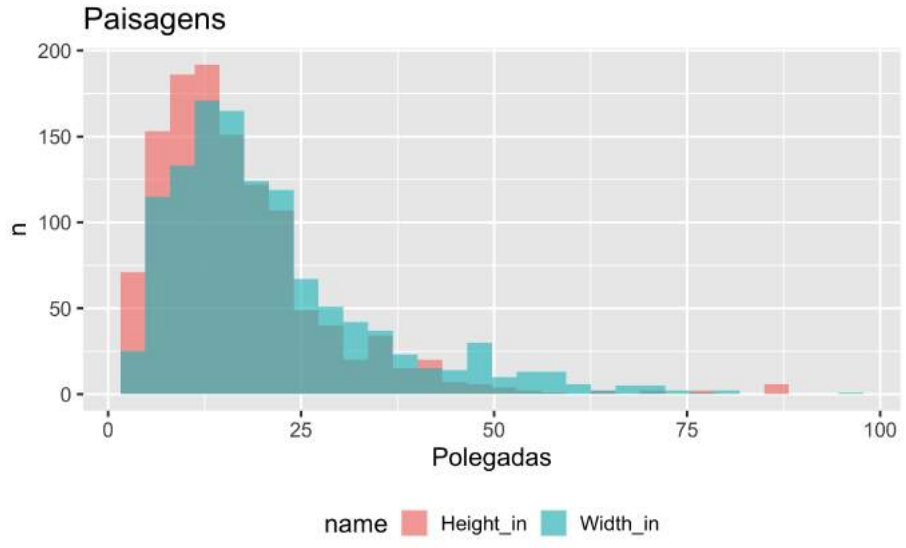
Paisagens vs. retratos

- A pintura de paisagem é a representação de paisagens - cenários naturais como montanhas, vales, árvores, rios e florestas composição
  - Habitualmente, largura $>$ altura
- Na pintura de retratos a intenção é retratar um sujeito humano:
  - Habitualmente, largura $<$ altura

Modelos, oportunidades e riscos

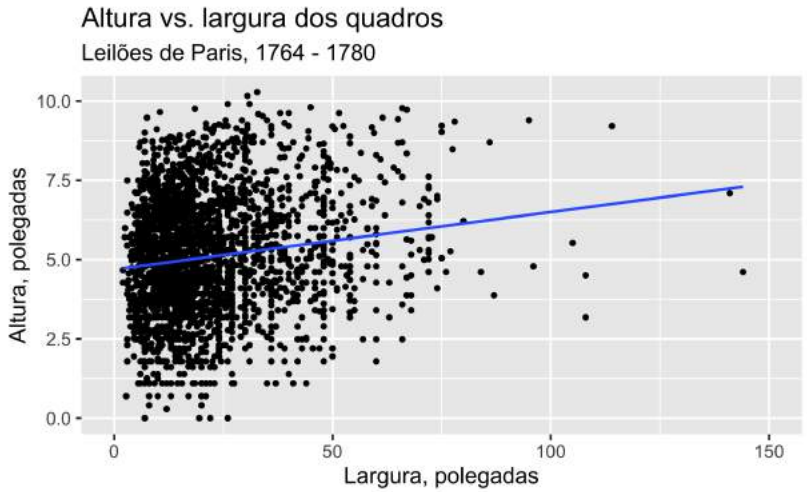
- **Oportunidade**: os modelos estatísticos podem revelar padrões que não se conseguem desvelar num gráfico
  - sobretudo em modelos de regressão múltipla, i.e. com diversas variáveis explicativas em simultâneo (próxima aula)





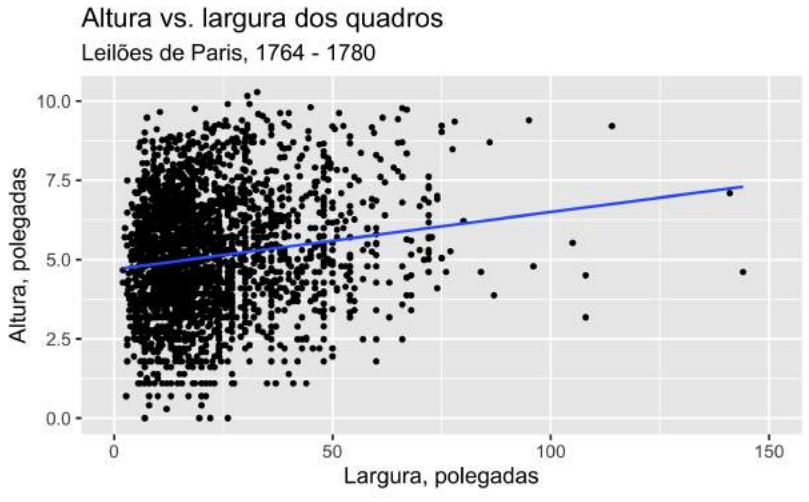
Modelos, oportunidades e riscos

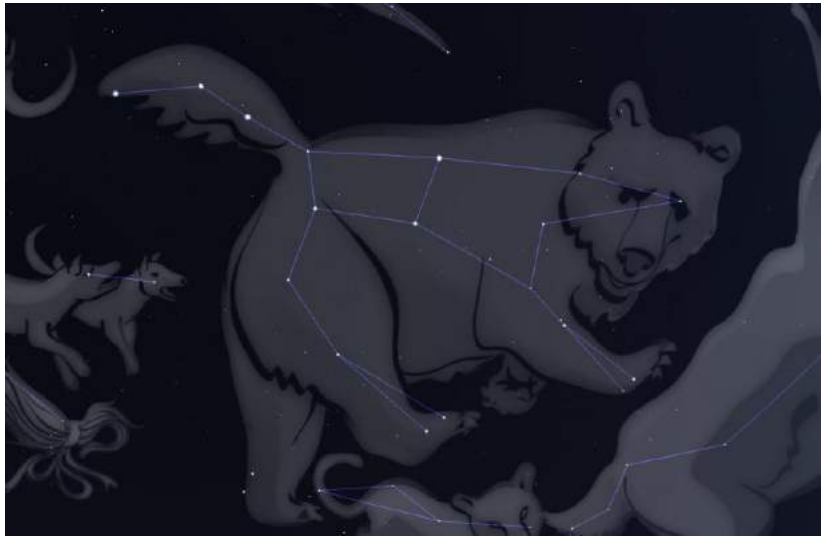
- **Risco**: um modelo pode indicar que os dados têm uma determinada estrutura (uma distribuição) que não corresponde à realidade -- regressão espúria



Modelos, oportunidades e riscos

- **Risco**: um modelo pode indicar que os dados têm uma determinada estrutura (uma distribuição) que não corresponde à realidade -- regressão espúria





OLS (MMQ)

Definição de um modelo linear

$$Y = \alpha + \beta X + \varepsilon$$

- Ex.: Y altura, X largura
- α - constante (ordenada na origem
- β - coeficiente de regressão / declive
- ε - erro do modelo

porque não resíduo?

Definição de um modelo linear

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

não observados

- Ex.: Y altura, X largura
- α - constante (ordenada na origem
- β - coeficiente de regressão / declive
- ε - erro do modelo

Estimar $\hat{\alpha}$ $\hat{\beta}$

- Hip.: linearidade
- Parâmetros e estimativas a partir dos dados i = 1, 2, …, N

Método dos mínimos quadrados (OLS)

- *Ordinary Least Squares:* minimizar os resíduos

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} \left[Y_i - \left(\hat{\alpha} + \hat{\beta} X_i \right) \right]^2$$

- Mostra-se que:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}$$

parece familiar?

“Bondade do ajustamento”

- R2: medida de ajustamento do modelo aos dados

Fonte da variação	Soma dos quadrados
Variação explicada	$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
Variação residual	$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
Variação total	$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

TSS=ESS+SSR

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$
$$0 \leq R^2 \leq 1$$

- R2 = 1: toda a variação dos dados pode ser explicada pelo modelo
- R2 = 0: vice-versa

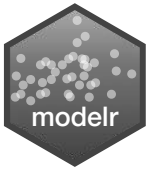
ggpmisc::stat_poly_eq()

Estimativas de regressão no gráfico

```
ggplot(data, aes(…)) + geom … +
    stat_poly_eq(use_label("R", "P", "n", "eq", …), method = 'lm')
```

Resultados a mostrar
(e.g. **R**2, **P**-value, **n**.º observações, **eq**uação da regressão)

Método de estimação (e.g. linear model, lm) tem **de bater certo com linha**



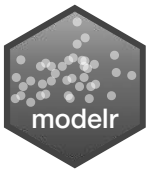
Experimente

- Vamos utilizar um pacote novo: **ggpmisc** - extensões ao ggplot
- Construa um gráfico com os dados dos quadros de Paris com:
  - a nuvem de pontos de largura (eixo xx) e altura (eixo yy)
  - linha de regressão linear SEM intervalo de confiança
  - a equação da regressão, o R2 e o n.º obs.

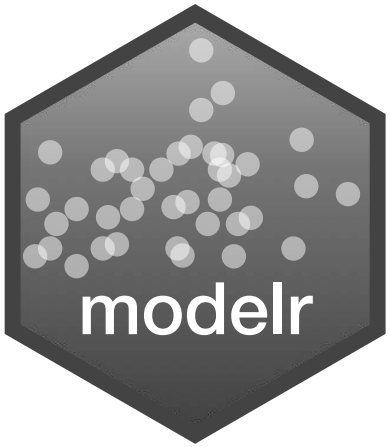
Modelos em R

Funções em R para estimar diferentes modelos

function	package	fits
lm()	**stats**	**linear models**
glm()	stats	generalized linear models
gam()	mgcv	generalized additive models
glmnet()	glmnet	penalized linear models
rlm()	MASS	robust linear models
rpart()	rpart	trees
randomForest()	randomForest	random forests
xgboost()	xgboost	gradient boosting machines

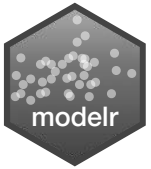


modelr



Funções tidy para trabalhar com modelos no *tidyverse*

```
library(tidyverse)
library(modelr)
```

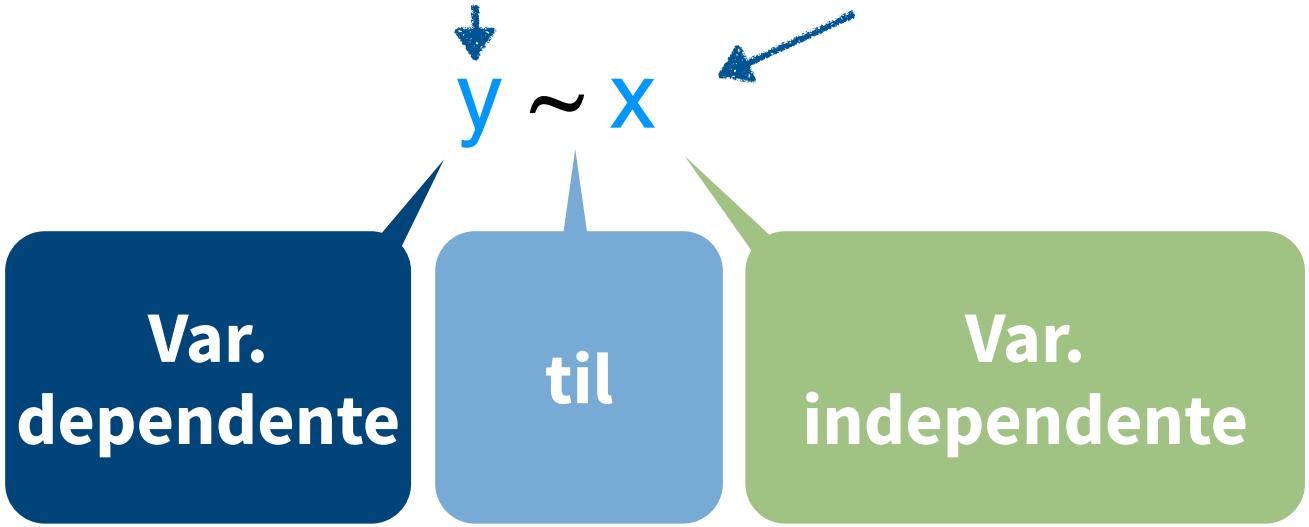
lm()

linear **m**odel

fórmulas no R

A equação de um modelo define-se no R em fórmulas, onde apenas é necessário indicar as variáveis dependente e independentes

$$y = \alpha + \beta x + \epsilon$$



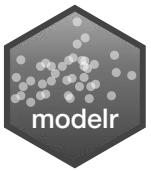
lm()

Função base de modelos lineares:

```
modelo <- lm(y ~ x, data = babynames)
```

Fórmula (equação da regressão a estimar)

Tabela de dados (tibble ou data.frame) onde as variáveis do modelo se localizam



%>% .

Utiliza-se o ponto final quando queremos passar uma tabela a uma função, noutro local que não o 1.º argumento

```
mod_e <- wages %>%
  lm(log(income) ~ education, data = .)
```

wages will be passed to here

Experimente

Corra 2 modelos lineares:

`mass` = α + β x `height`, com os dados starwars

`price` = α + β x `Width_in`

e examine os outputs, atribuindo-os aos objetos *modelo_star* e *modelo_quadros,* respetivamente

Obrigado
e até à próxima!

luis.morais@novasbe.pt