Preparado para: AT autoridade tributária e aduaneira

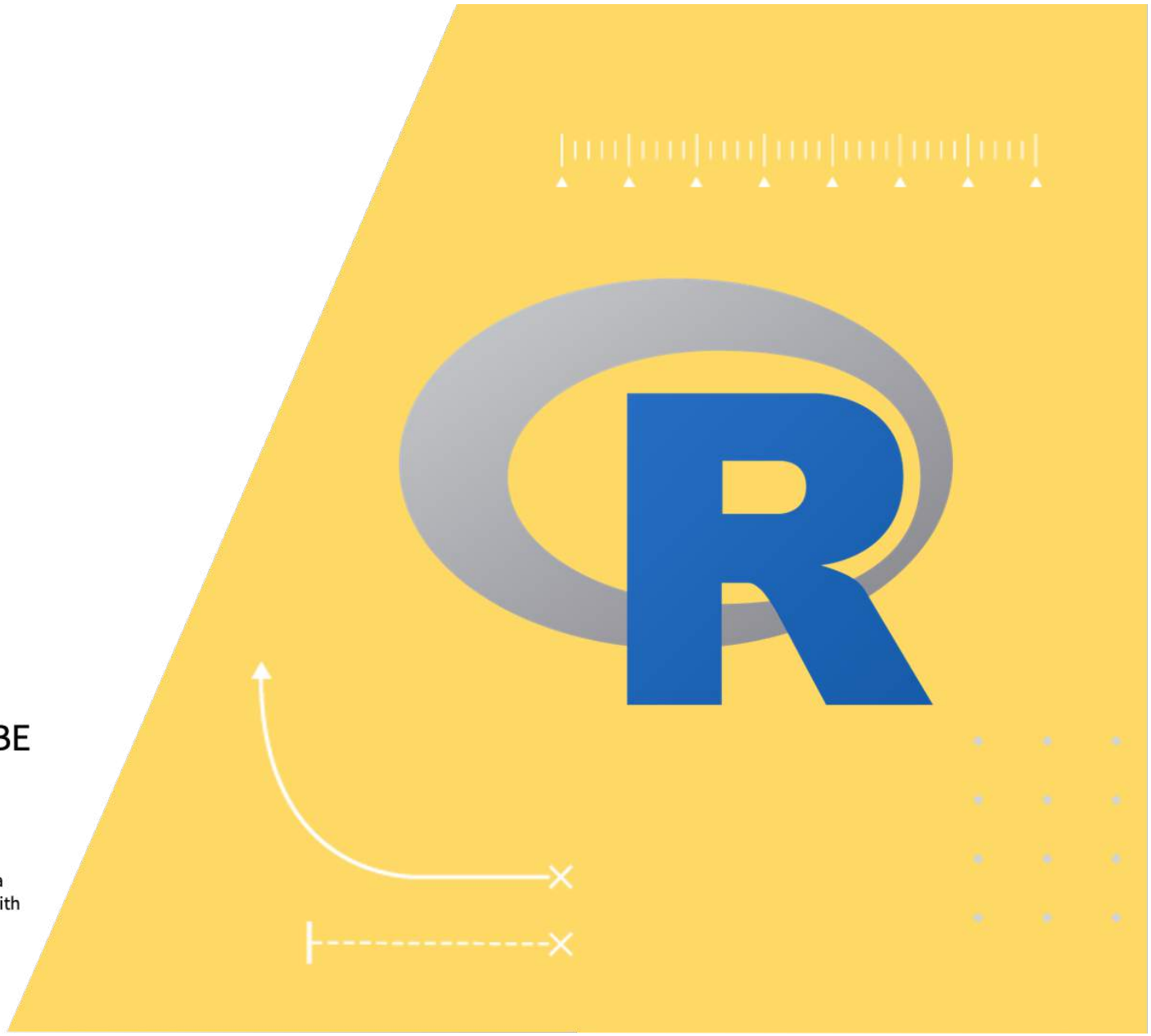
REFORM/SC2022/126
DELIVERABLE 4
MÓDULO 4
REGRESSÃO LINEAR

DESIGNING A NEW VALUATION MODEL
FOR RURAL PROPERTIES IN PORTUGAL

Parte III

Formador: Luís Teles Morais | Nova SBE

Lisboa, 29 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

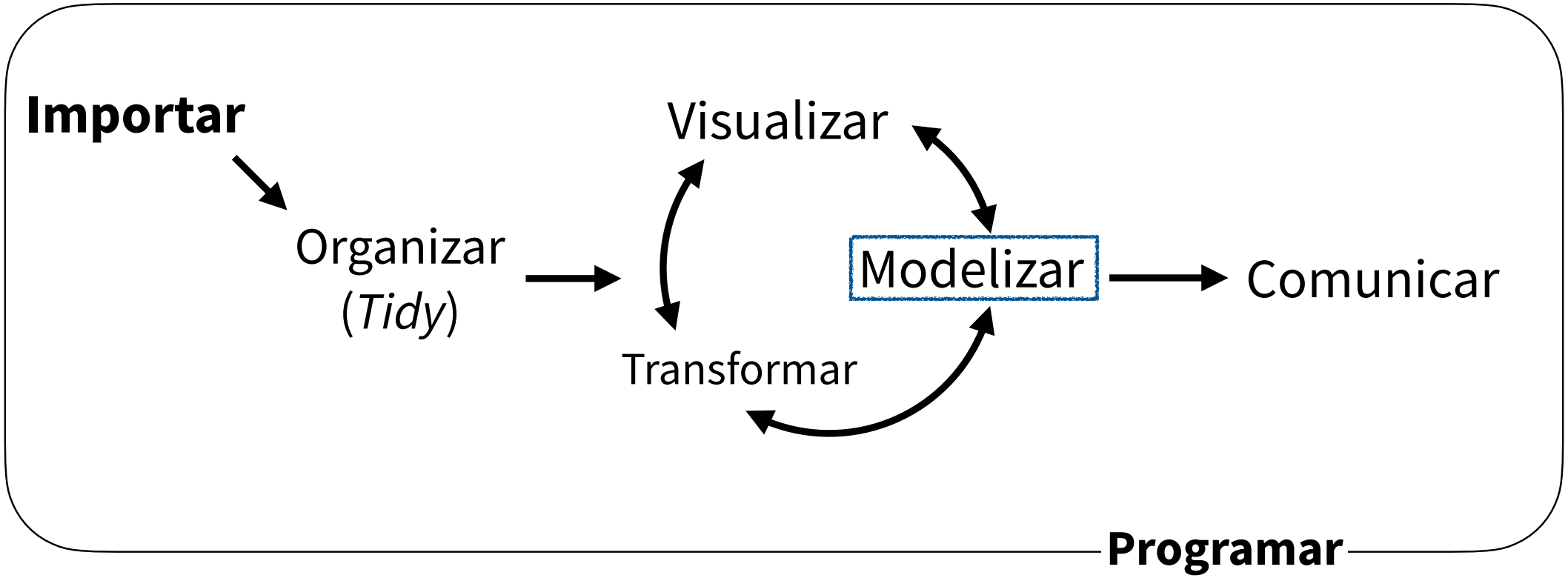
AARC NOVA SCHOOL OF BUSINESS & ECONOMICS LOBO VASQUES esri INNERLANDS

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



lm()

linear **m**odel

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)

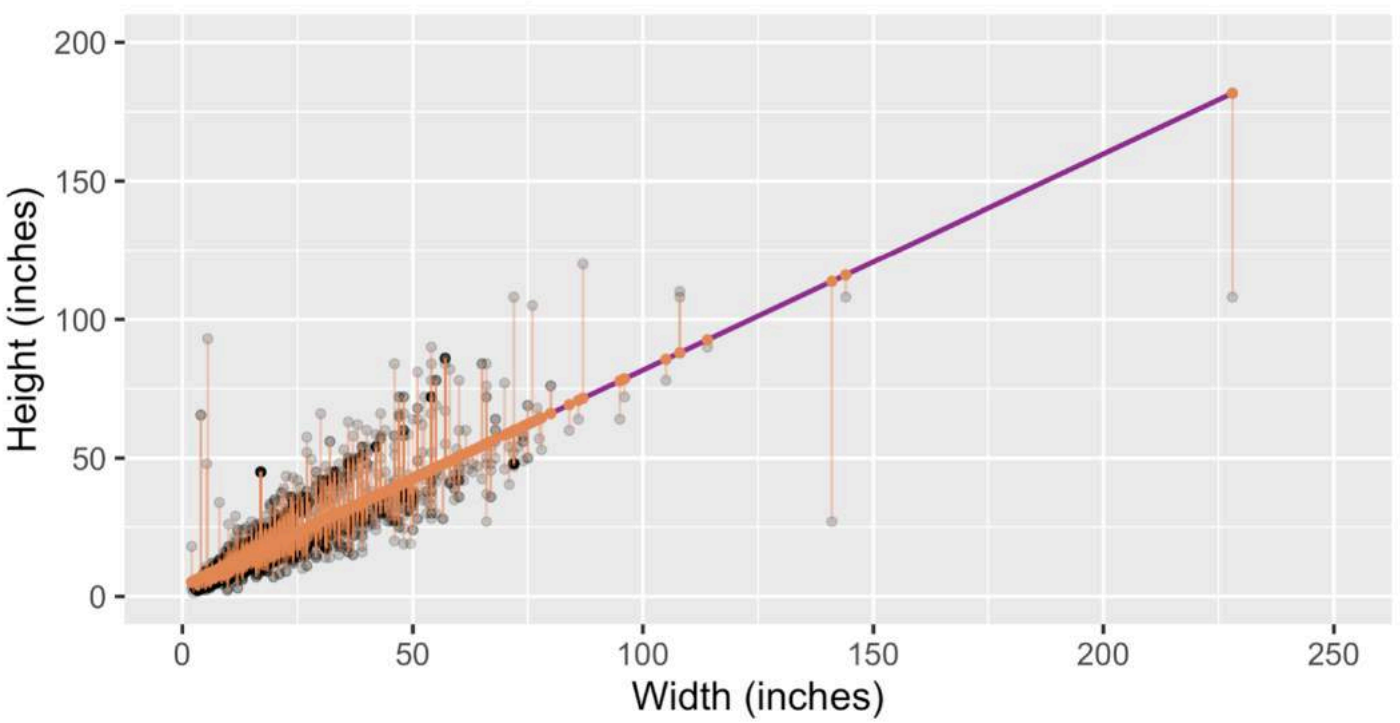
- *Ordinary Least Squares:* minimizar os resíduos

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

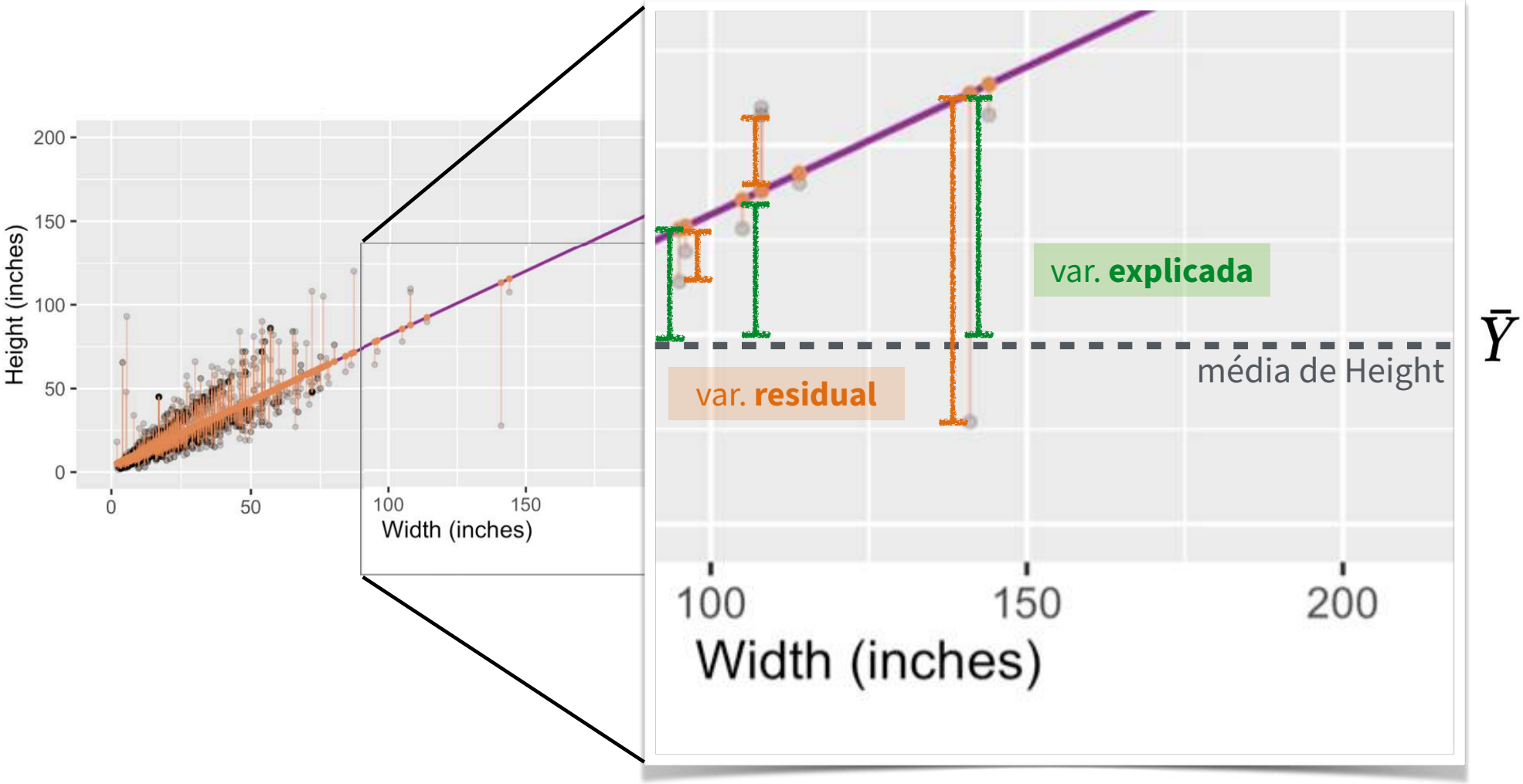
- Mostra-se que:

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

Minimizar os resíduos



Minimizar os resíduos



“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

Hipóteses de um modelo OLS

Recorde: pelo CLT, para 1 variável a média da amostra é um bom estimados da média da população — ou seja, não é enviesado.

$$Y = \alpha + \beta X + \varepsilon \quad \longrightarrow \quad Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

Da mesma forma, os **estimadores OLS** são bons estimadores dos parâmetros, sob certas hipóteses.

Hip. OLS

- Linearidade
- Resíduos não correlacionados com a variável independente (**exogeneidade**)
- Amostra i.i.d. - independente e identicamente distribuída

Hipóteses de um modelo OLS

Recorde: pelo CLT, para 1 variável a média da amostra é um bom estimados da média da população — ou seja, não é enviesado.

$$Y = \alpha + \beta X + \varepsilon \longrightarrow Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

Da mesma forma, os **estimadores OLS** são bons estimadores dos parâmetros, sob certas hipóteses.

Y , X variáveis aleatórias $\longrightarrow$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

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

também são variáveis aleatórias, com a sua própria distribuição

Significância estatística

Sob estas hipóteses, podemos ter uma ideia da significância do coeficiente de declive: traduz ele realmente uma relação entre as variáveis?

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0$$

Sob hipóteses OLS, numa amostra grande:

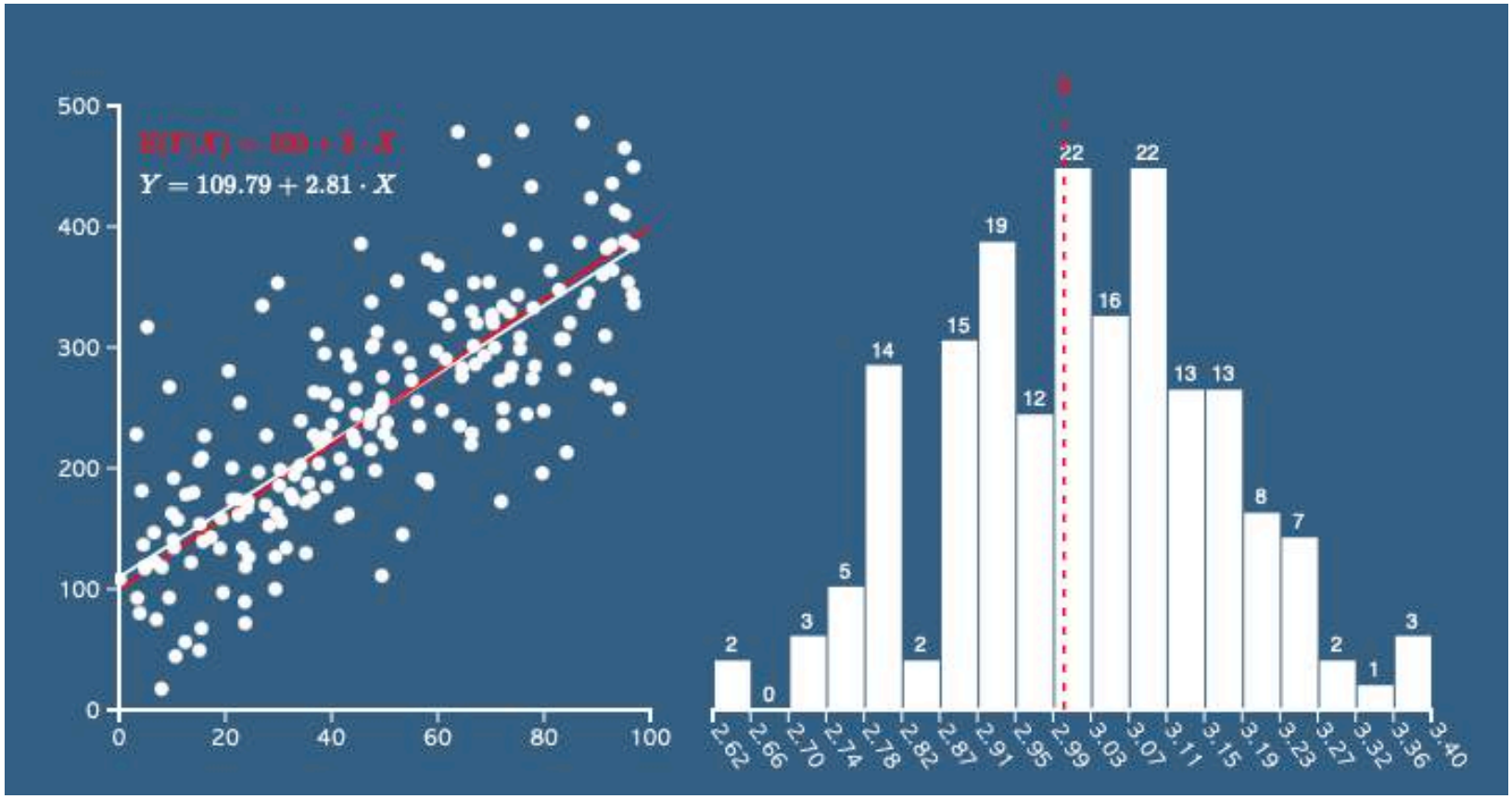
- $\hat{\beta}$ acerta em média (é centrada)
- A variância, ou o erro padrão, de $\hat{\beta}$ tem uma distribuição Normal pelo T. do Limite Central

→

Podemos avaliar a estatística-t:

$$t = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}$$

Hipóteses de um modelo OLS



https://www.econometrics-with-r.org/4.5-tsdotoe.html

Significância estatística

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0$$

Sob hipóteses OLS, numa amostra grande, mostra-se que

$$t = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)} \sim N(0,1)$$

Interpretação de modelos: **queremos erros-padrão pequenos em relação a ^β**

Quando | **t** | **> 1.96** (= valor crítico):

- rejeita-se H0 (= **β é diferente de zero)**
- com um nível de significância de **5%**

Significância estatística

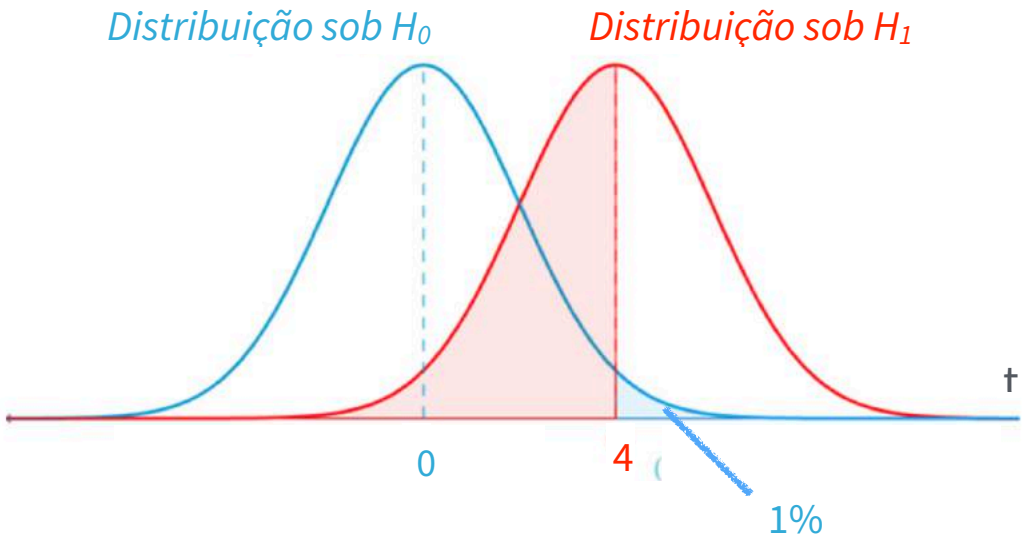
$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0$$

Interpretação de modelos: **queremos p-value pequenos**

O mesmo teste pode ser visto rapidamente com **p-values:**

$$t = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)} \sim N(0,1)$$

(para amostras grandes)



valor da **estatística *t* grande**
<=> **valor-p pequeno**

- **p-value = 2% < nível de significância = 5%**
  - rejeita-se H0 => **<u>estimativa de β é significativa</u>**

Variáveis **categóricas**

Variável binária

```
## # A tibble: 3,393 × 3
##    name      Height_in landsALL
##    <chr>         <dbl>    <dbl>
##  1 L1764-2          37        0
##  2 L1764-3          18        0
##  3 L1764-4          13        1
##  4 L1764-5a         14        1
##  5 L1764-5b         14        1
##  6 L1764-6           7        0
##  7 L1764-7a          6        0
##  8 L1764-7b          6        0
##  9 L1764-8          15        0
## 10 L1764-9a          9        0
## 11 L1764-9b          9        0
## 12 L1764-10a        16        1
## 13 L1764-10b        16        1
## 14 L1764-10c        16        1
## 15 L1764-11         20        0
## 16 L1764-12a        14        1
## 17 L1764-12b        14        1
## 18 L1764-13a        15        1
## 19 L1764-13b        15        1
## 20 L1764-14         37        0
## # … with 3,373 more rows
```

- **landsall**
  - ***1****: tem características de paisagem*
  - ***0:*** *não tem qualquer característica de paisagem*
- **Qual seria outro tipo de dados no R (mais correcto) que esta variável poderia assumir?**

```
modelo_bi <- pp_rect %>%
                lm(price ~ Surface + landsALL, data = .)

modelo_bi %>%
  tidy()
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  622.      51.1         12.2  2.23e-33
2 Surface        0.197    0.0326       6.04 1.75e- 9
3 landsALL     138.      73.3          1.88 6.00e- 2
```

Média de landsAll?

Como interpretar?

```
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  622.       51.1        12.2  2.23e-33
2 Surface        0.197     0.0326      6.04 1.75e- 9
3 landsALL     138.       73.3         1.88 6.00e- 2
```

- **landsall** (=1)**:** Espera-se que um quadro com algumas características de paisagem tenha em média um preço superior em 138 Francos, a um quadro sem qualquer característica de paisagem **e com idêntica área de superfície (tudo o resto constante...).**

```
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  622.       51.1        12.2  2.23e-33
2 Surface        0.197     0.0326      6.04 1.75e- 9
3 landsALL     138.       73.3         1.88 6.00e- 2
```

Preço médio estimado de um quadro paisagem:
price = 622 + 0.197 * Surface **+ 138** (ex.: Surface = 100 -> price = **780**)

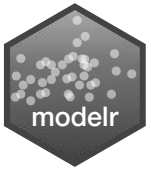
Preço médio estimado de um quadro não-paisagem (retrato?):
price = 622 + 0.197 * Surface (ex: price = 780-138 = **642**)

wages

CHARACTER

income <dbl>	height <dbl>	weight <dbl>	age <dbl>	marital <chr>	sex <chr>	education <dbl>
19000	60	155	53	married	female	13
35000	70	156	51	married	female	10
105000	65	195	52	married	male	16
40000	63	197	54	married	female	14
75000	66	190	49	married	male	14
102000	68	200	49	divorced	female	18
70000	64	160	54	divorced	female	12
60000	69	162	55	divorced	male	12
150000	69	194	54	divorced	male	13
115000	64	145	53	married	female	16

1–10 of 5,266 rows | 1–7 o... Previous 1 2 3 4 5 ... 100 Next



factors

Variável categórica: discreta/qualitativa, com uma certa ordem (dada pelos levels/níveis).

```
sexes <- factor(x = c("male", "female", "male"),
                levels = c("male", "female", "other"))
sexes
## male   female male
## Levels: male female other
```

Pode sempre alterar-se os níveis recriando o factor:

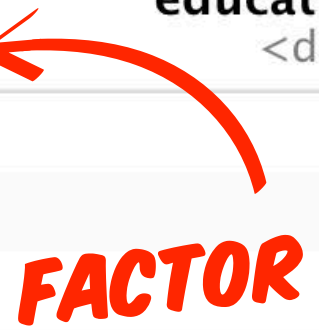
```
sexes <- factor(sexes, levels = c("male", "female", "other"))

sexes
## male   female male
Levels: female male other

unclass(sexes)
## 2 1 2
## attr(,"levels")
## "female" "male"   "other"
```

```
wages <-
  wages %>%
  mutate(sex = factor(sex, levels = c("male", "female")))
```

income <dbl>	height <dbl>	weight <dbl>	age <dbl>	marital <chr>	sex <fctr>	education <dbl>
19000	60	155	53	married	female	13
35000	70	156	51	married	female	10
105000	65	195	52	married	male	16
40000	63	197	54	married	female	14
75000	66	190	49	married	male	14
102000	68	200	49	divorced	female	18
70000	64	160	54	divorced	female	12



Variável categórica (>2 níveis)

```
## # A tibble: 3,393 × 3
##    name      Height_in school_pntg
##    <chr>         <dbl> <chr>
##  1 L1764-2          37 F
##  2 L1764-3          18 I
##  3 L1764-4          13 D/FL
##  4 L1764-5a         14 F
##  5 L1764-5b         14 F
##  6 L1764-6           7 I
##  7 L1764-7a          6 F
##  8 L1764-7b          6 F
##  9 L1764-8          15 I
## 10 L1764-9a          9 D/FL
## 11 L1764-9b          9 D/FL
## 12 L1764-10a        16 X
## 13 L1764-10b        16 X
## 14 L1764-10c        16 X
## 15 L1764-11         20 D/FL
## 16 L1764-12a        14 D/FL
## 17 L1764-12b        14 D/FL
## 18 L1764-13a        15 D/FL
## 19 L1764-13b        15 D/FL
## 20 L1764-14         37 F
## # … with 3,373 more rows
```

- **school_pntg**
  - ***"F"**: França*
  - ***"I":** Itália*
  - ***"D/FL":** Alemanha/Flandres*
  - *…*
- **Como introduzir num modelo?**

Variável categórica (>2 níveis)

school_pntg	D_FL	F	G	I	S	X
A	0	0	0	0	0	0
D/FL	1	0	0	0	0	0
F	0	1	0	0	0	0
G	0	0	1	0	0	0
I	0	0	0	1	0	0
S	0	0	0	0	1	0
X	0	0	0	0	0	1

■ **1 variável para cada categoria?**

Variáveis “dummy” (*burras*)

school_pntg	D_FL	F	G	I	S	X
A	0	0	0	0	0	0
D/FL	1	0	0	0	0	0
F	0	1	0	0	0	0
G	0	0	1	0	0	0
I	0	0	0	1	0	0
S	0	0	0	0	1	0
X	0	0	0	0	0	1

- **Não!** N.º categorias — 1
- Definir categoria **baseline**
- Cada variável **dummy** representa o efeito dessa categoria por comparação com a baseline

O que aconteceria se colocássemos variáveis dummy para todas as categorias?

Experimente

1. Volte a estimar com os dados **pp_rect**, um modelo linear do preço de venda do quadro (**price**) com três variáveis explicativas: a altura, a área de superfície e o ano **year**.
2. Agora estime o mesmo modelo mas acrescentando variáveis "dummy" para incorporar diferenças médias entre as diferentes escolas (**school_pntg**).
   - O que aconteceu ao R2?
   - Interprete as novas estimativas obtidas

```
modelo_4sc <- pp_rect %>% lm(price ~ Surface +Height_in + year +
                               factor(school_pntg), data = .)
modelo_4sc %>% tidy()
```

```
# A tibble: 10 × 6
   term                      estimate  std.error statistic  p.value
   <chr>                        <dbl>      <dbl>     <dbl>    <dbl>
 1 (Intercept)              -95306.     12542.        -7.60  3.93e-14
 2 Surface                      -0.0238     0.0609    -0.390 6.96e- 1
 3 Height_in                    26.7        4.63       5.77  8.92e- 9
 4 year                         53.4        7.00       7.63  3.21e-14
 5 factor(school_pntg)D/FL    1407.       1344.        1.05  2.96e- 1
 6 factor(school_pntg)F        538.       1345.        0.400 6.89e- 1
 7 factor(school_pntg)G        249.       1589.        0.157 8.75e- 1
 8 factor(school_pntg)I        537.       1348.        0.398 6.90e- 1
 9 factor(school_pntg)S       2165.       1525.        1.42  1.56e- 1
10 factor(school_pntg)X        384.       1377.        0.279 7.80e- 1
```

Quando as hipóteses OLS **falham**

Hipóteses de um modelo OLS

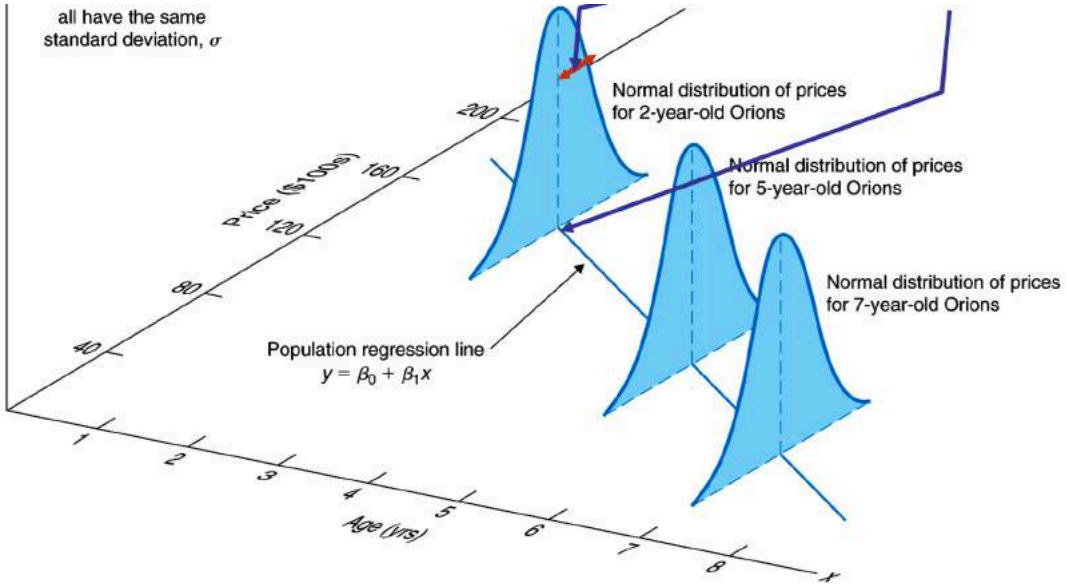
$$Y = \alpha + \beta X + \varepsilon \quad \longrightarrow \quad Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

Os **estimadores OLS** são bons estimadores dos parâmetros,
sob certas hipóteses.

Hip. OLS

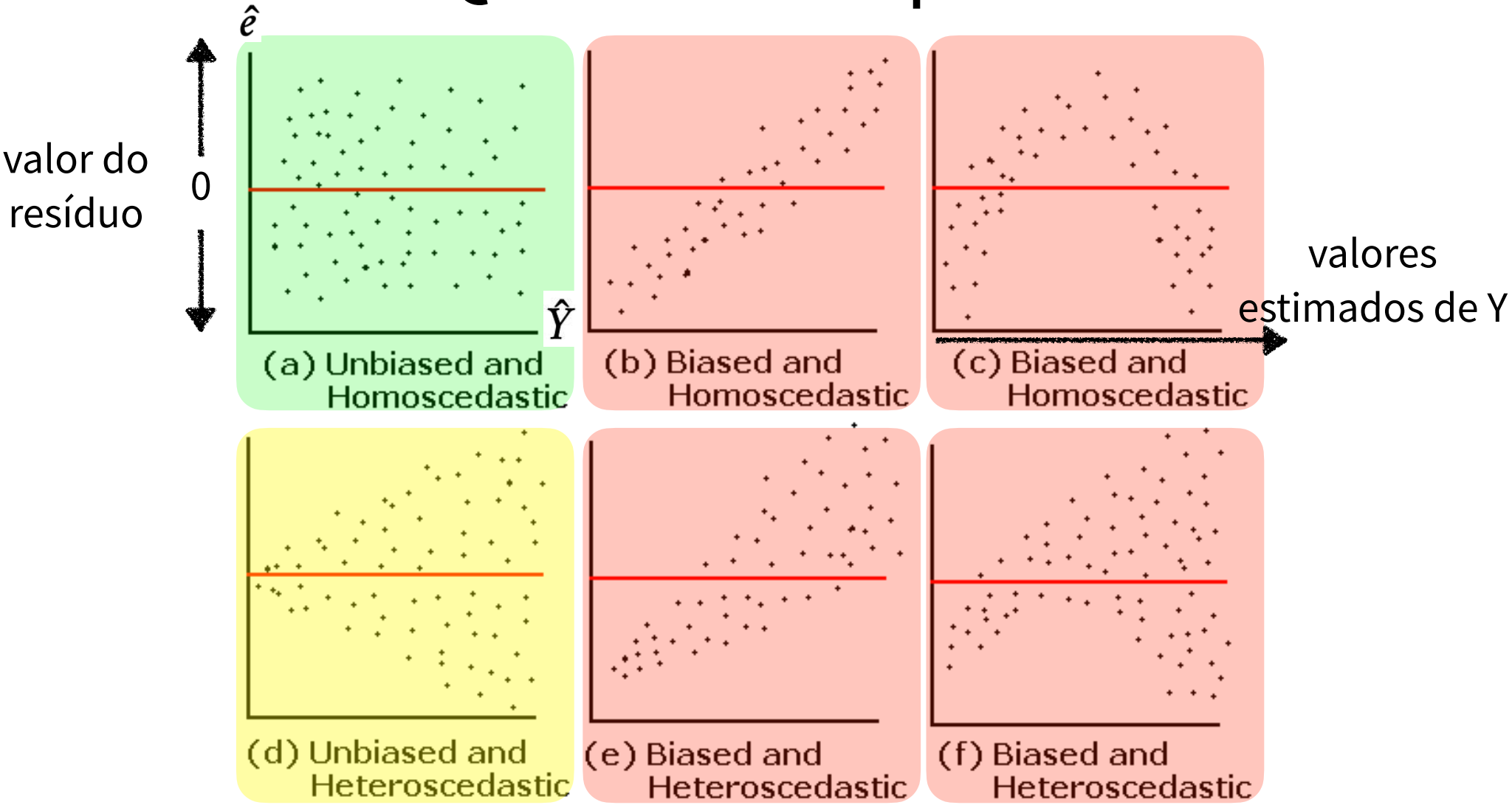
- **Linearidade** nos parametros
- Resíduos não correlacionados com a variável independente (**exogeneidade**)
- **Amostra i.i.d**. - independente e identicamente distribuída
- (Normalidade dos resíduos)

Analisar os resíduos



- Quando modelo é válido os resíduos devem ter distribuições (aproximadamente) normais com a mesma variância, e média zero
- **Check: gráfico dos resíduos**

Quando as hipóteses falham…

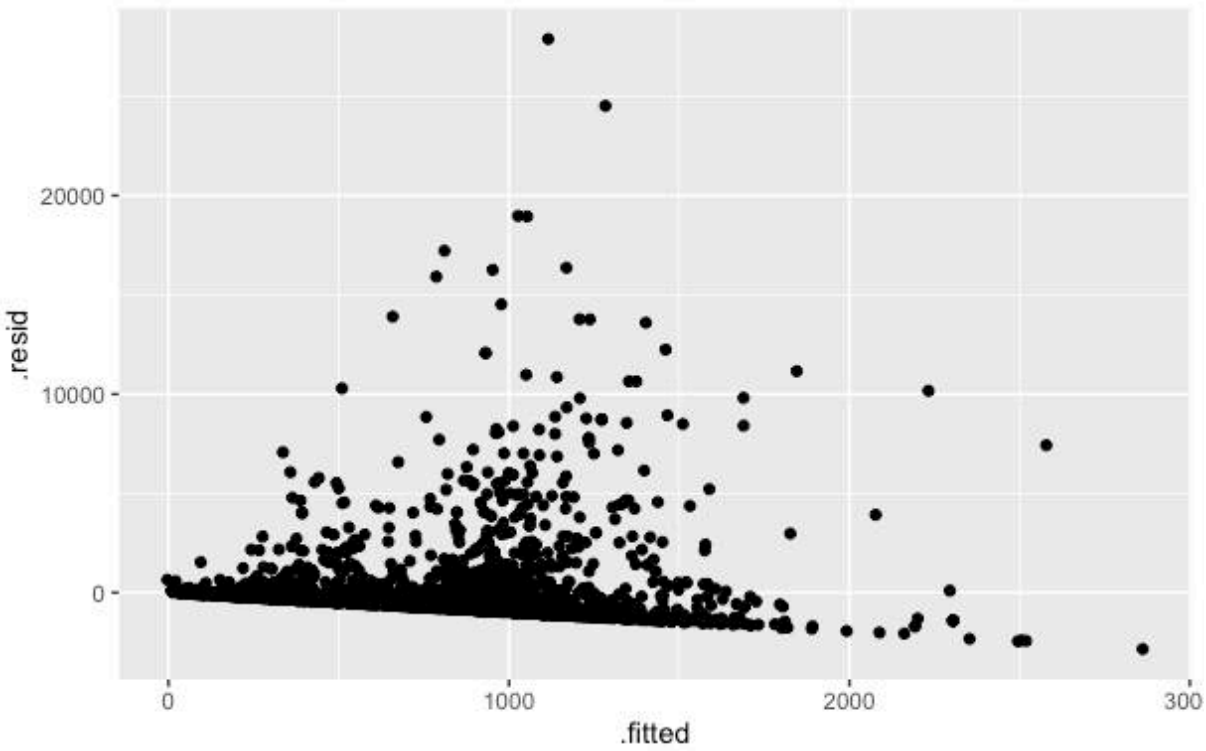


Experimente

1. Volte a estimar (ou recupere o modelo já estimado) com os dados **pp_rect**, um modelo linear do preço de venda do quadro (**price**) com três variáveis explicativas: a altura, a área de superfície e o ano **year**.
   - Produza o gráfico de resíduos deste modelo

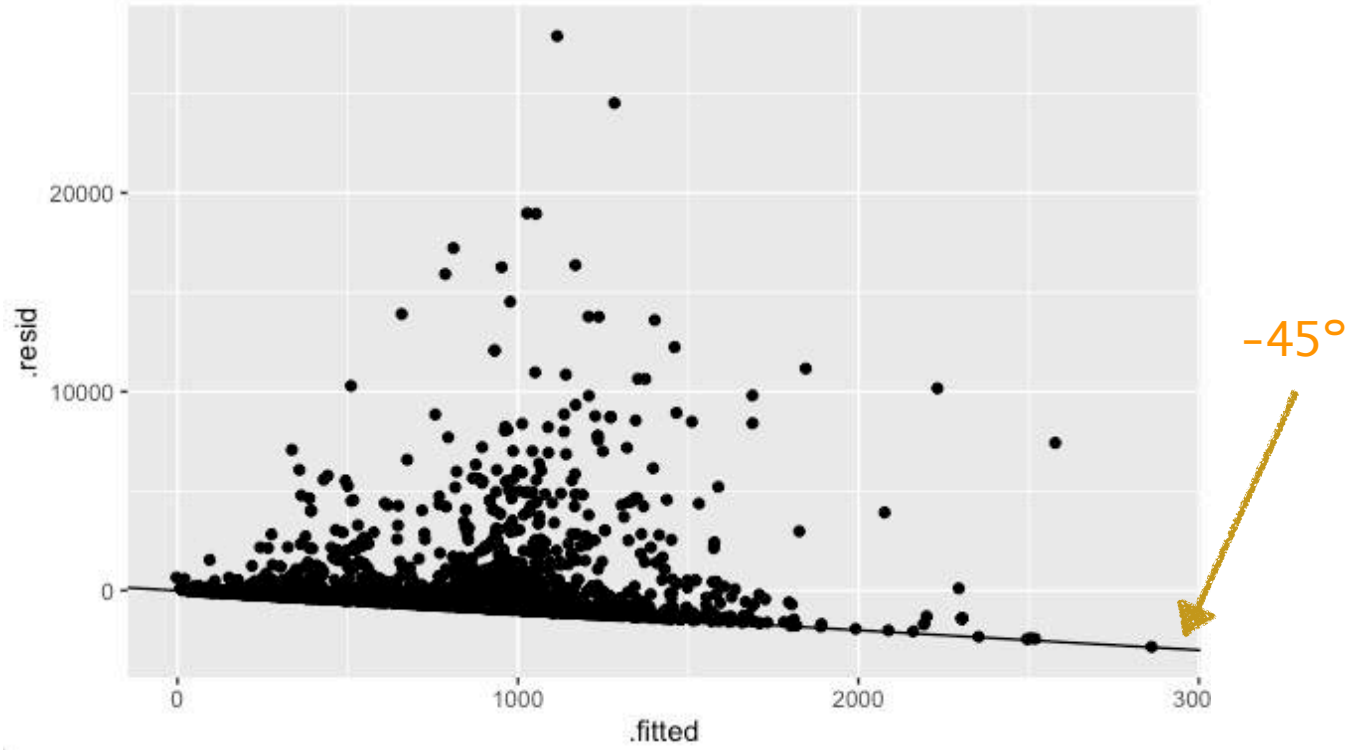
```
modelo_4 <- pp_rect %>% lm(price ~ Surface + Height_in + year, data = .)


modelo_4 %>% augment() %>%
          ggplot(aes(x = .fitted, y = .resid)) + geom_point()
```

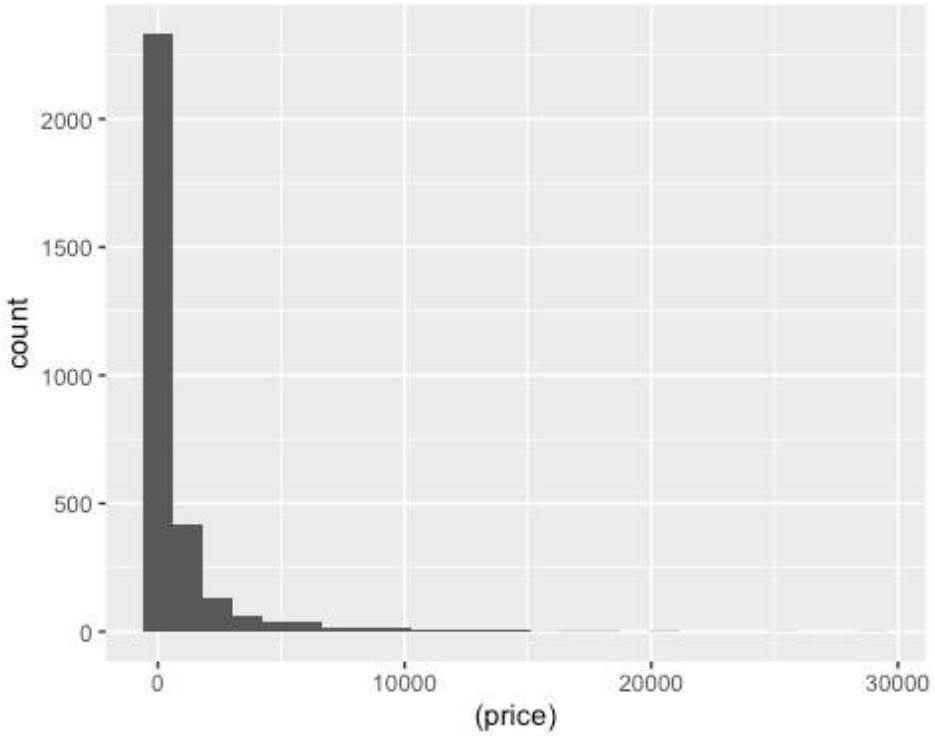


```
modelo_4 <- pp_rect %>% lm(price ~ Surface + Height_in + year, data = .)


modelo_4 %>% augment() %>%
          ggplot(aes(x = .fitted, y = .resid)) + geom_point() +
               geom_abline(aes(slope = -1, intercept = 0))
```

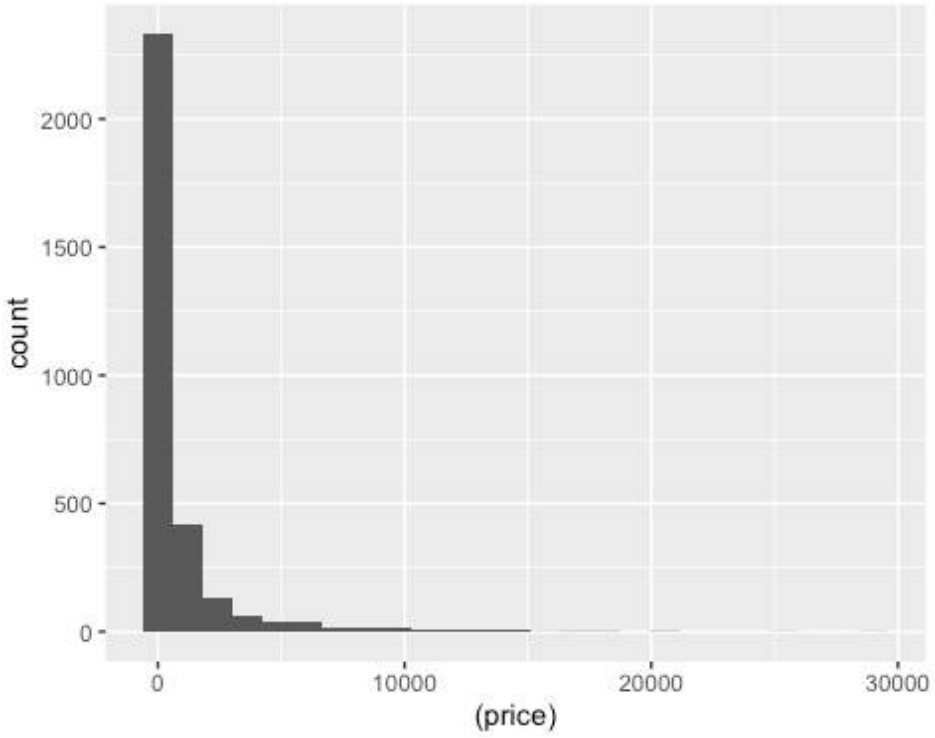


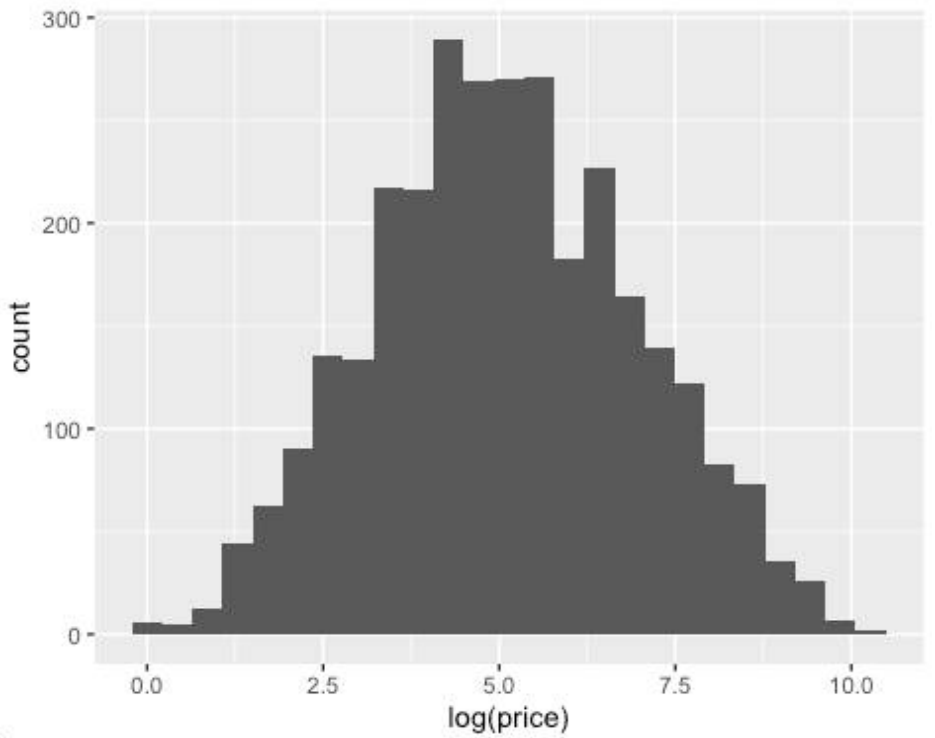
```
pp_rect %>% ggplot(aes(x = price)) +
    geom_histogram(bins = 25)
```



```
pp_rect %>% ggplot(aes(x = price)) +
    geom_histogram(bins = 25)
```

```
pp_rect %>% ggplot(aes(x = log(price))) +
    geom_histogram(bins = 25)
```





Linearidade

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

- Modelo linear nos parâmetros
- Exige relação de proporcionalidade entre Y e X
- **Por vezes falha mas pode ser recuperado com transformações simples das variáveis de interesse — e.g. logaritmo**

Linearidade

	X	
Y	X	$\log X$
Y	*linear*	*linear-log*
	$\hat{Y}_i = \alpha + \beta X_i$	$\hat{Y}_i = \alpha + \beta \log X_i$
$\log Y$	*log-linear*	*log-log*
	$\log \hat{Y}_i = \alpha + \beta X_i$	$\log \hat{Y}_i = \alpha + \beta \log X_i$

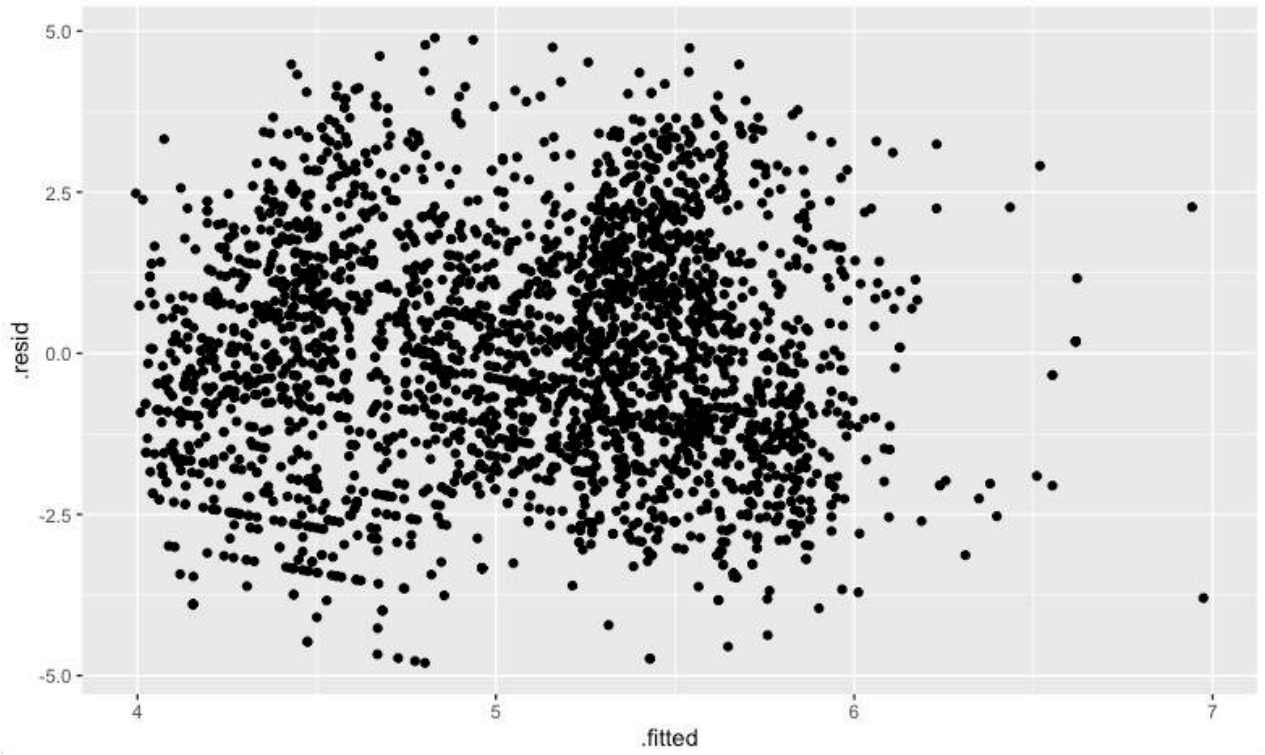
- Interpretação: variação percentual
- e.g. log-linear: + 1 unid. de X => + β (x 100) % de Y

Experimente

1. Volte a estimar com os dados **pp_rect**, um modelo linear do preço de venda do quadro (**price**) com três variáveis explicativas: a altura, a área de superfície e o ano **year**.
2. Agora estime o mesmo modelo mas numa versão log-linear, ou seja com **log(price)** como variável dependente.
   - Produza o gráfico de resíduos deste modelo
   - O que aconteceu ao R2?
   - Interprete as novas estimativas obtidas

```
modelo_4l <- pp_rect %>% lm(log(price) ~ Surface + Height_in + year, data
= .)


modelo_4l %>% augment() %>%
          ggplot(aes(x = .fitted, y = .resid)) + geom_point()
```



```
modelo_4l <- p_rect %>%
        lm(log(price) ~ Surface +Height_in + year, data = .)

modelo_4 %>% tidy() %>%
    cross_join((modelo_4 %>% glance %>% select(r.squared)))
```

```
# A tibble: 4 x 6
  term             estimate  std.error statistic  p.value r.squared
  <chr>               <dbl>      <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept) -178.         11.5           -15.5   6.36e-52    0.0825
2 Surface        0.0000512   0.0000569       0.900 3.68e- 1    0.0825
3 Height_in      0.0125      0.00424         2.96  3.10e- 3    0.0825
4 year           0.103       0.00651        15.9   1.25e-54    0.0825
```

```
# A tibble: 4 × 6
  term             estimate  std.error statistic  p.value r.squared
  <chr>               <dbl>      <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept) -110909.      12576.         -8.82  1.90e-18    0.0373
2 Surface          0.0392       0.0620     0.633 5.27e- 1    0.0373
3 Height_in       15.8          4.61       3.43  6.04e- 4    0.0373
4 year            62.8          7.09       8.86  1.30e-18    0.0373
```

```
# A tibble: 4 × 6
  term             estimate  std.error statistic  p.value r.squared
  <chr>               <dbl>      <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept) -178.         11.5          -15.5   6.36e-52    0.0825
2 Surface          0.0000512    0.0000569     0.900 3.68e- 1    0.0825
3 Height_in        0.0125       0.00424       2.96  3.10e- 3    0.0825
4 year             0.103        0.00651      15.9   1.25e-54    0.0825
```

Interpretação log/linear

```
# A tibble: 4 x 6
  term           estimate  std.error statistic  p.value r.squared
  <chr>             <dbl>      <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept) -178.         11.5         -15.5   6.36e-52    0.0825
2 Surface        0.0000512   0.0000569     0.900 3.68e- 1    0.0825
3 Height_in      0.0125      0.00424       2.96  3.10e- 3    0.0825
4 year           0.103       0.00651      15.9   1.25e-54    0.0825
```

- **Declive - Height_in: Tudo o resto constante (*ceteris paribus*)**, por cada polegada adicional em altura, espera-se que o preço do quadro seja, em média, **1.25% mais elevado**.
  Quadros maiores tendem a ter preços mais elevados.

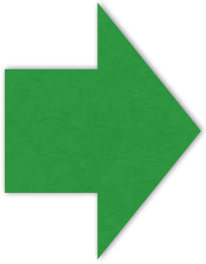
Endogeneidade & variáveis omitidas

Violação da hipótese da exogeneidade:

- **correlação entre X** (variáveis independentes) **e os resíduos**

Caso típico:

- **resíduos "escondem" uma variável omitida**
  - correlacionada com X
  - importante para determinar variável dependente Y



Estimador de **β enviesado**
ou descentrado: não acerta, em média, no parâmetro da população.

$$\hat{\beta}_1 \rightarrow \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

Experimente

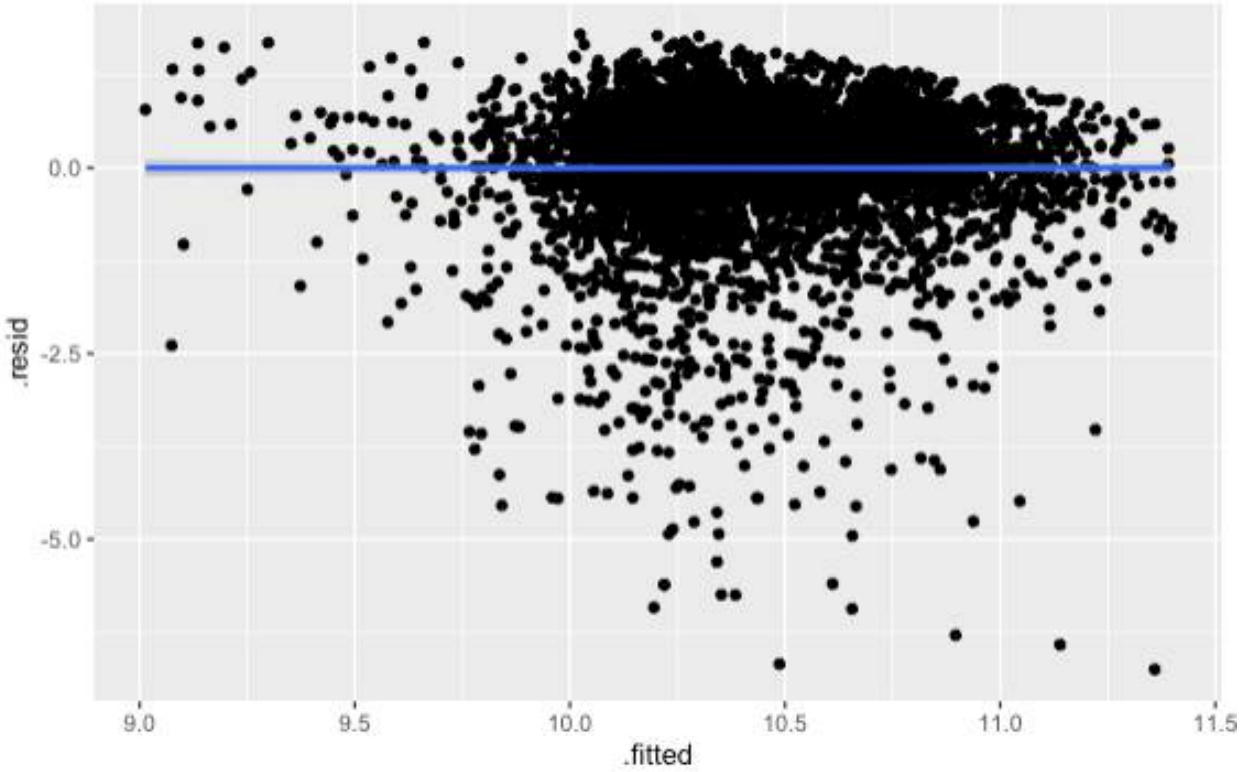
1. Carregue os dados wages.xlsx (processando corretamente quaisquer NAs).
2. Estime um modelo que explique o logaritmo do **income** com as variáveis **education** e o logaritmo de **weight**.
   - Como interpreta o coeficiente de **weight**? Os resultados fazem sentido?
   - Analise o gráfico de resíduos do modelo. As estimativas obtidas são confiáveis? Porquê?

```
wm <- wages_data %>%
        lm(log(income) ~ (education) + log(weight) , data = .)

wm %>% tidy()
```

```
# A tibble: 3 x 5
  term         estimate std.error statistic   p.value
  <chr>           <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)    6.38     0.326        19.6  2.84e- 82
2 education      0.122    0.00529      23.1  1.23e-112
3 log(weight)    0.458    0.0602        7.62 3.04e- 14
```

```
wm %>% augment() %>%
          ggplot(aes(x = .fitted, y = .resid)) + geom_point()  +
             geom_smooth(method = 'lm')
```



```
wm <- wages_data %>%
        lm(log(income) ~ (education) + log(weight) + factor(sex),
data = .)

wm %>% tidy()
```

```
# A tibble: 4 × 5
  term            estimate std.error statistic   p.value
  <chr>              <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)        8.06     0.341       23.6  3.73e-117
2 education          0.129    0.00521     24.7  9.91e-128
3 log(weight)        0.0796   0.0648       1.23 2.19e-  1
4 factor(sex)male    0.409    0.0290      14.1  1.80e- 44
```

Multicolinearidade

Quando as variáveis independentes estão fortemente correlacionadas:

- Estimativas instáveis
- Impossível distinguir efeito de diferentes variáveis

Forma fácil de inspecionar:

- Matriz de correlações
- -> Todos os coeficientes de correlação cruzados

```
pp_rect %>% select(Surface, Height_in, year) %>%
cor()
```

```
             Surface  Height_in       year
Surface    1.0000000  0.8563879 -0.1130731
Height_in  0.8563879  1.0000000 -0.1611672
year      -0.1130731 -0.1611672  1.0000000
```

```
wm <- wages_data %>%
        lm(log(income) ~ (education) + log(weight) +
            (sex == 'male') + (sex == 'female)),  data = .)

wm %>% tidy()
```

```
# A tibble: 5 × 5
  term                    estimate std.error statistic    p.value
  <chr>                      <dbl>     <dbl>     <dbl>      <dbl>
1 "(Intercept)"              8.06      0.341      23.6   3.73e-117
2 "education"                0.129     0.00521    24.7   9.91e-128
3 "log(weight)"              0.0796    0.0648      1.23  2.19e-  1
4 "sex == \"male\"TRUE"      0.409     0.0290     14.1   1.80e- 44
5 "sex == \"female\"TRUE"    NA        NA         NA     NA
```

- **Multicolinearidade perfeita**: *male* = 1 - *female*
  - => estimadores indefinidos

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

```
wm <- wages_data %>%
        lm(log(income) ~ (education) + log(weight) +
            (sex == 'male') + (sex == 'female) - 1),  data = .)

wm %>% tidy()
```

```
# A tibble: 4 × 5
  term                estimate std.error statistic   p.value
  <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
1 education             0.129    0.00521      24.7  9.91e-128
2 log(weight)           0.0796   0.0648        1.23 2.19e-  1
3 factor(sex)female     8.06     0.341        23.6  3.73e-117
4 factor(sex)male       8.47     0.352        24.0  5.23e-121
                        = 8.06 + 0.409
```

intercept

- “Dummies” redistribuem a constante por diferentes grupos

Obrigado
e até à próxima!

luis.morais@novasbe.pt