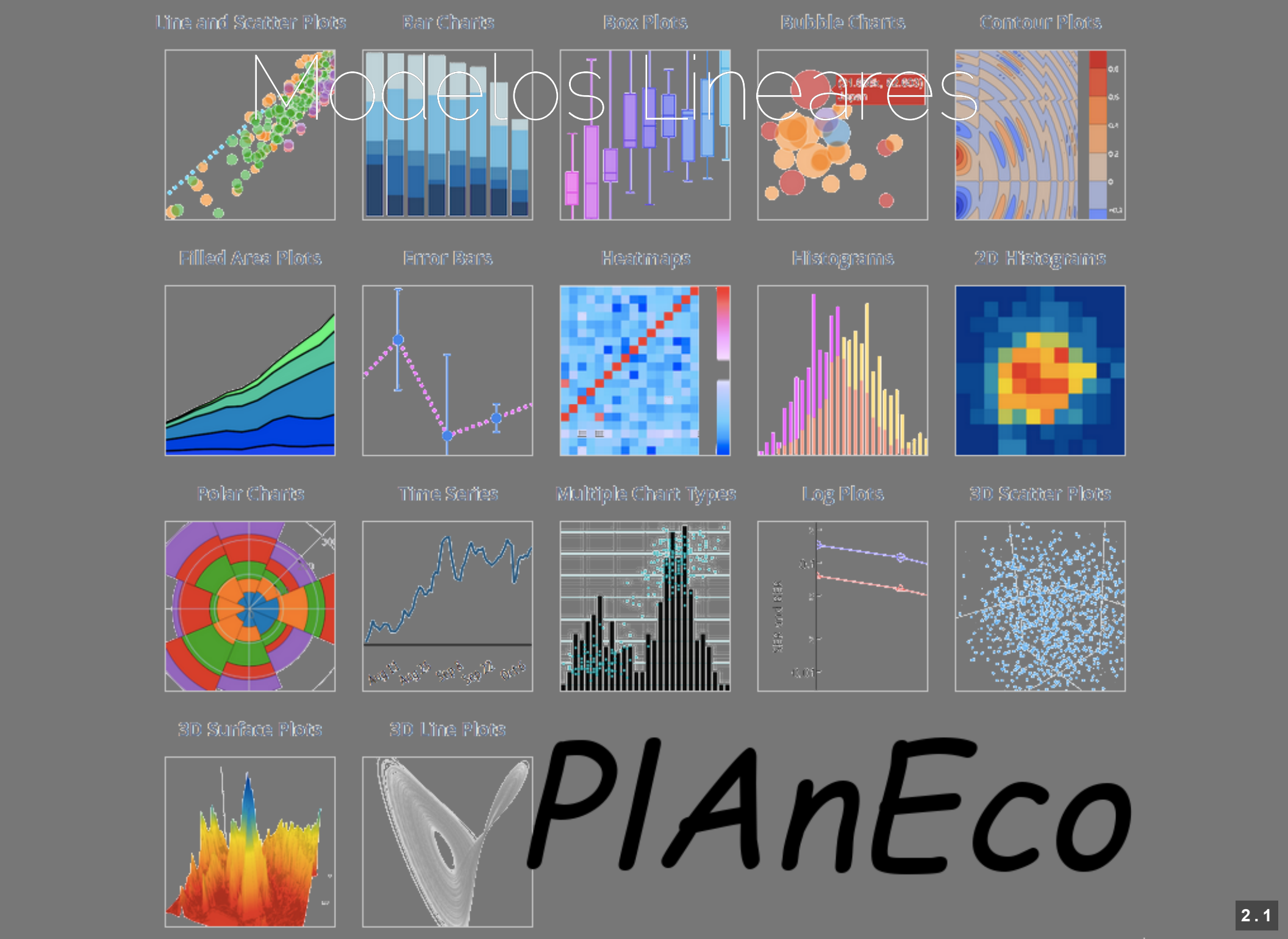
Modelos Lineares

unificação metodológica

Alexandre Adalardo de Oliveira

PlanECO 2018


Line and Scatter Plots
Bar Charts
Box Plots
Bubble Charts
Contour Plots
Modelos Lineares
Filled Area Plots
Error Bars
Heatmaps
Histograms
2D Histograms
Polar Charts
Time Series
Multiple Chart Types
Log Plots
3D Scatter Plots
3D Surface Plots
3D Line Plots
PlAnEco

Conceitos

- razão das variâncias
- dummy variables (indicadoras)
- matriz do modelo
- interação entre preditoras
- ANOVA é similar a uma regressão!

Testes Clássicos

Tipo de Variável		Estatística Clássica	
Resposta	Preditora	Teste	Hipótese
Categórica	Categórica	Qui-quadrado	independência
Contínua	Categórica (2 niveis)	Teste t	$\mu_1 = \mu_2$
Contínua	Categórica	Anova	$\mu_1 = \mu_2 = \mu_n$
Contínua	1 Contínua	Regressão	$\beta_1 = 0$
Contínua	>1 Contínua	Reg. múltipla	$\beta_1 = 0 ; \beta_n = 0$
Contínua	Cont + Categ	Ancova	$\beta_1 = \beta_2 ; \alpha_1 = \alpha 2$
Proporção	Contínua	Reg. Logística	$logit(\beta_1) = 1$

Modelo Linear Simples

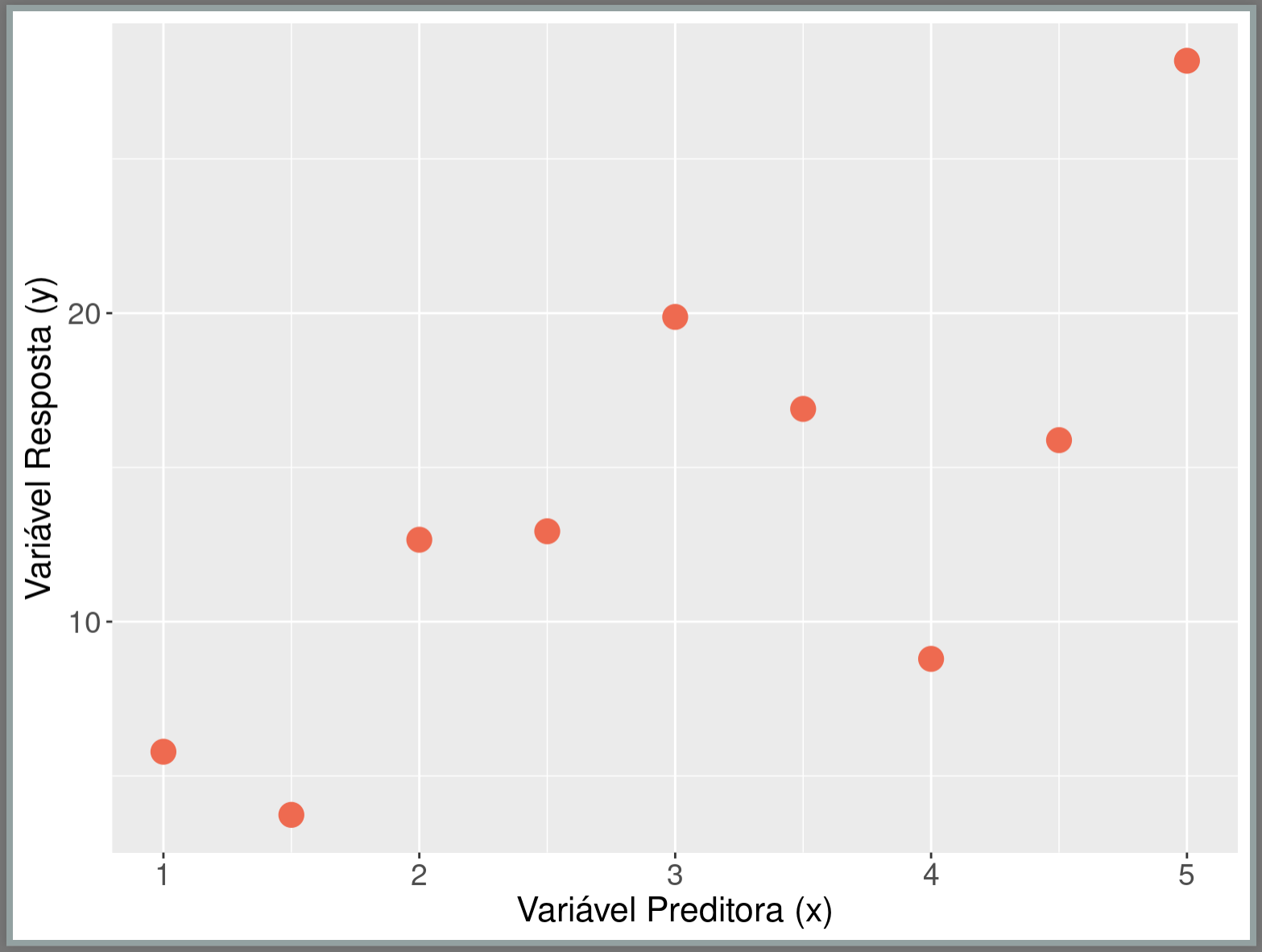
$$y = \alpha + \beta x + \epsilon$$
$$\epsilon = N(0, \sigma)$$

Simulando dados

$$y = \alpha + \beta x + \epsilon$$
$$y_i = 1.2 + 3.5 x_i + \epsilon_i$$
$$\epsilon_i = N(0, 5)$$

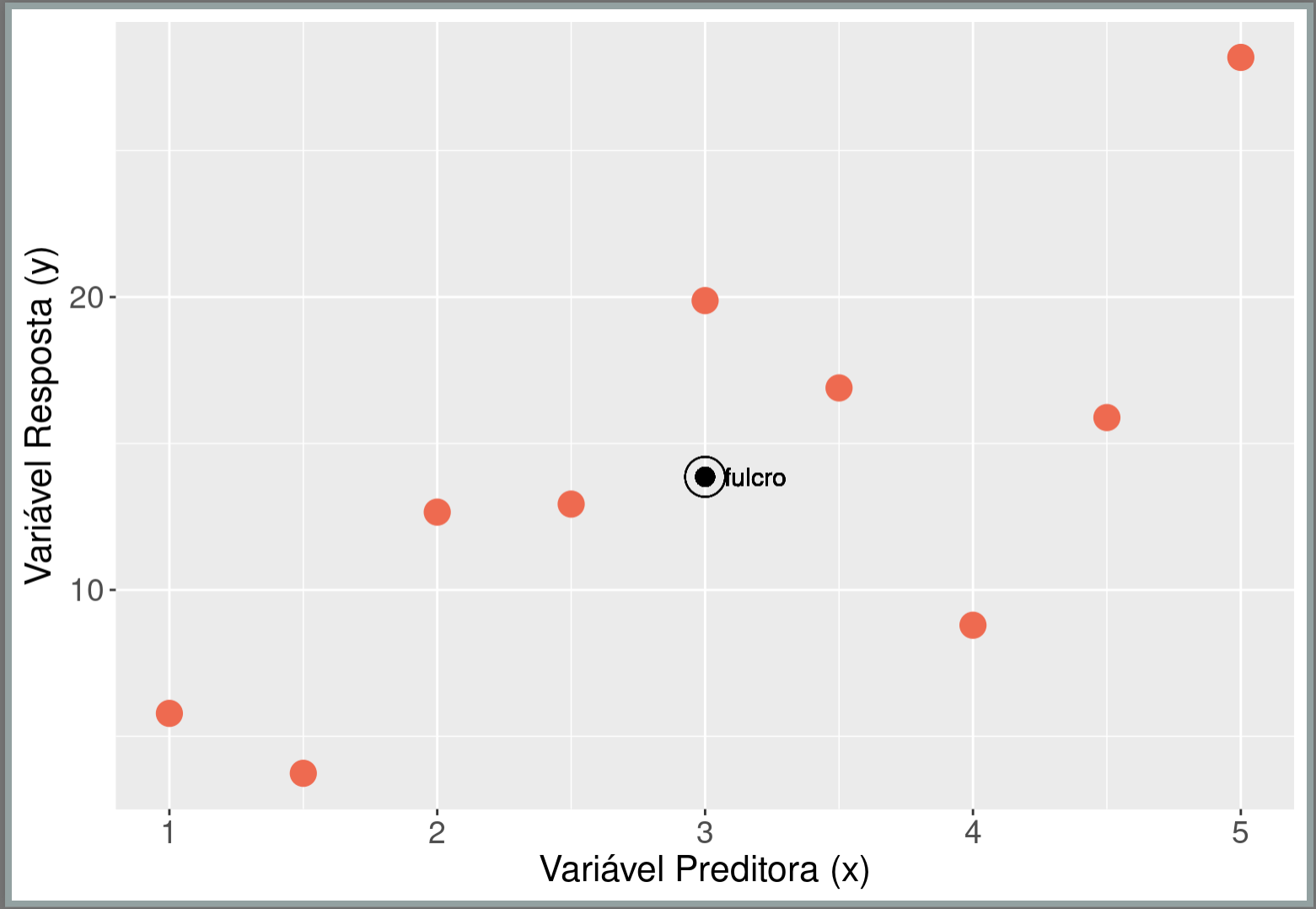
x1	y0	res	y1
1.0	4.70	1.08	5.78
1.5	6.45	-2.71	3.74
2.0	8.20	4.46	12.66
2.5	9.95	2.98	12.93
3.0	11.70	8.18	19.88
3.5	13.45	3.45	16.90
4.0	15.20	-6.41	8.79
4.5	16.95	-1.07	15.88
5.0	18.70	9.48	28.18

Modelo Linear: dados simulados

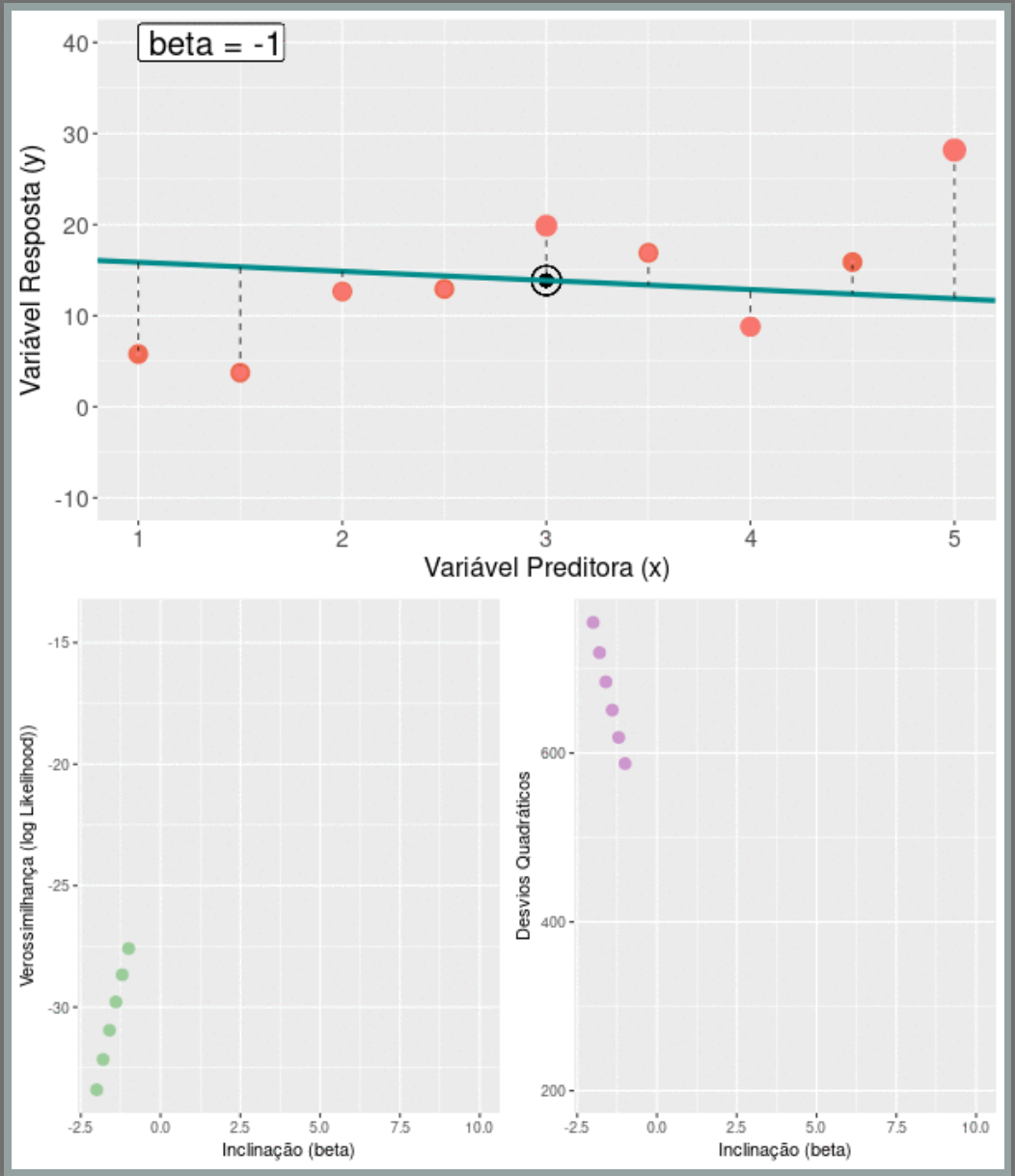


Mínimos Quadrados

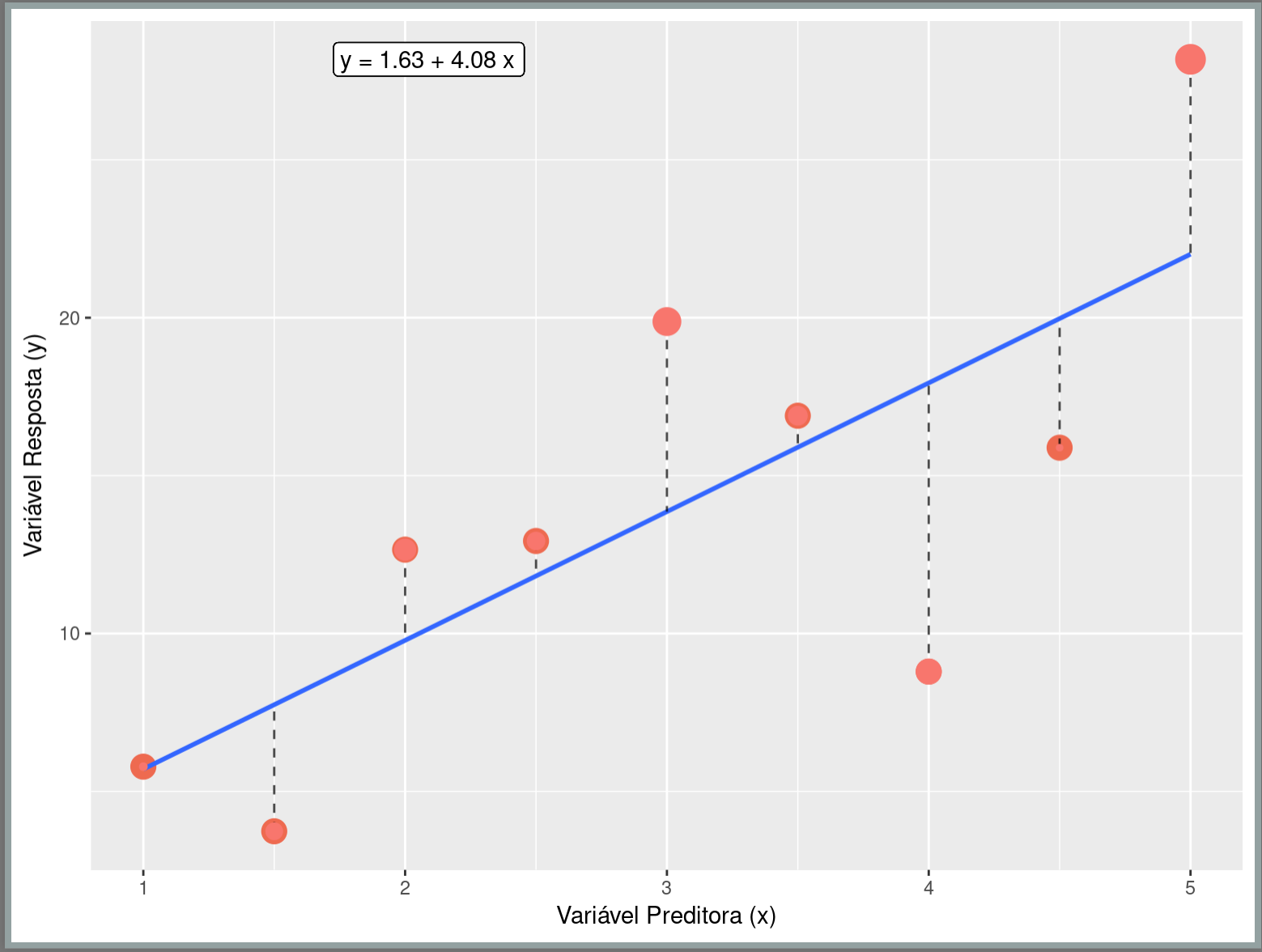
Ponto de Fulcro



Mínimos Quadrados



Estima Parâmetros



Modelo Linear

$$y_i = 1.2 + 3.5x_i + \epsilon_i$$
$$\epsilon_i = N(0, 5)$$

```
##
## Call:
## lm(formula = y1 ~ x1, data = xy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1424 -4.0088  0.9982  2.8714  6.1706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)    1.632      4.520   0.361   0.728
## x1             4.076      1.384   2.945   0.021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 5.361 on 7 degrees of
## Multiple R-squared:  0.5534, Adjusted R-squared
## F-statistic: 8.672 on 1 and 7 DF,  p-value: 0.0
```

Modelo Linear

Incerteza na estimativa

$$y_i = 1.2 + 3.5x_i + \epsilon_i$$

Coeficientes estimados

```
## (Intercept)          x1
##    1.632390    4.075949
```

Intervalo de confiança

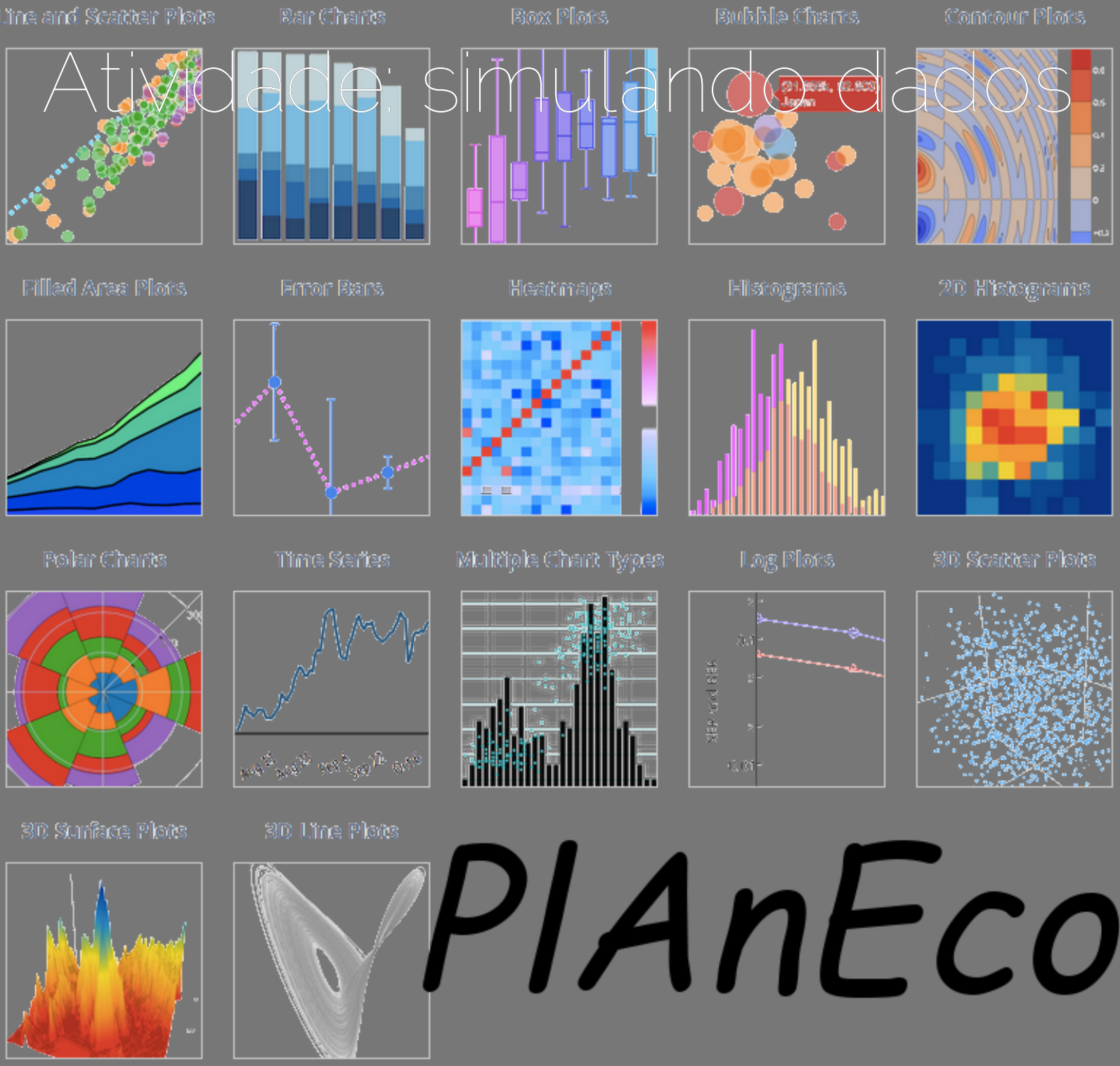
```
##                   2.5 %    97.5 %
## (Intercept) -9.0566136 12.321394
## x1           0.8031237  7.348775
```

Resumo do Modelo Linear

$$y_i = 1.2 + 3.5x_i + \epsilon_i$$
$$\epsilon_i = N(0, 5)$$

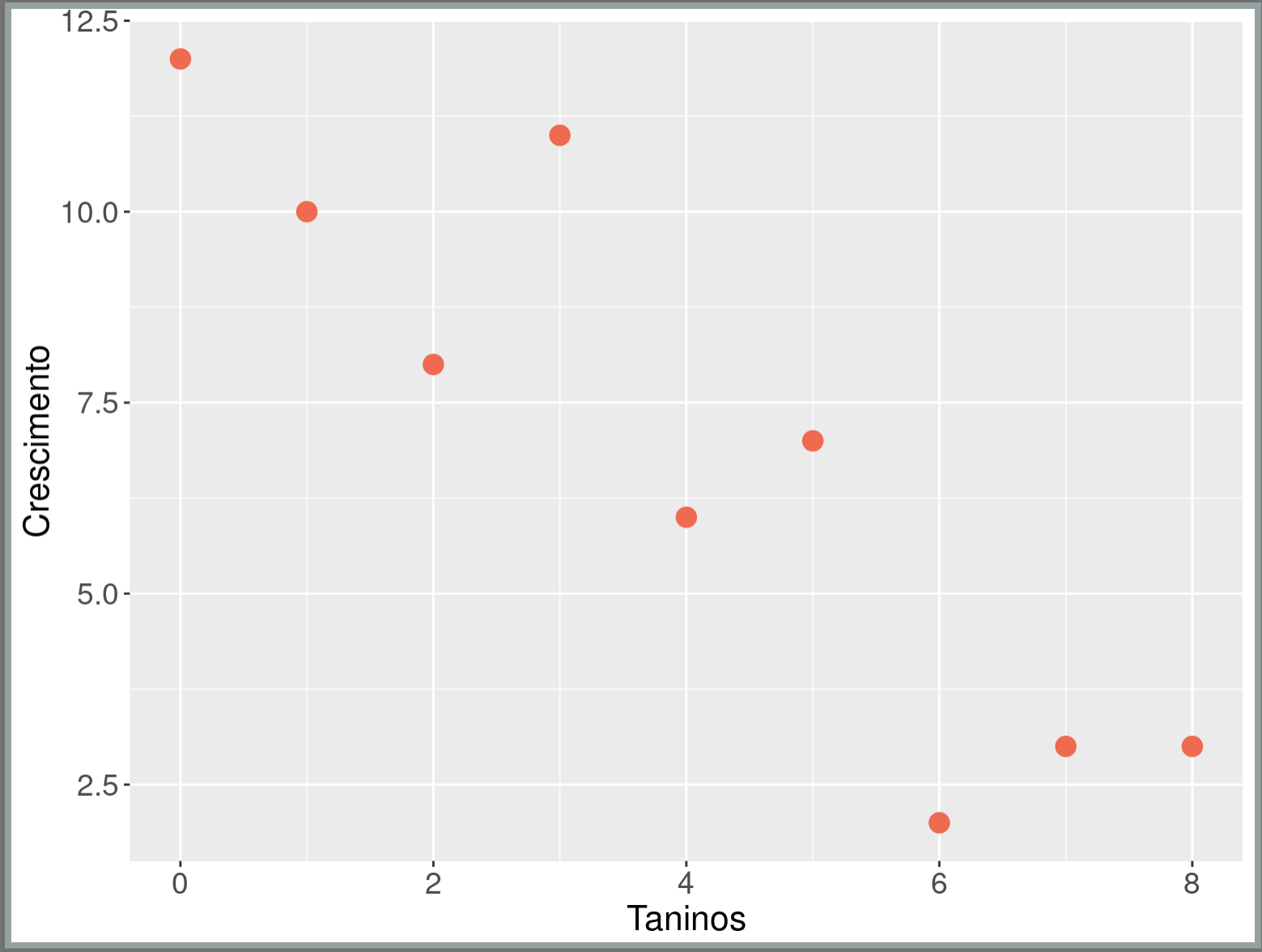
```
##
## Call:
## lm(formula = y1 ~ x1, data = xy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1424 -4.0088  0.9982  2.8714  6.1706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)    1.632      4.520   0.361   0.728
## x1             4.076      1.384   2.945   0.021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 5.361 on 7 degrees of
## Multiple R-squared:  0.5534, Adjusted R-squared
## F-statistic: 8.672 on 1 and 7 DF,  p-value: 0.0
```

Atividade: simulando dados

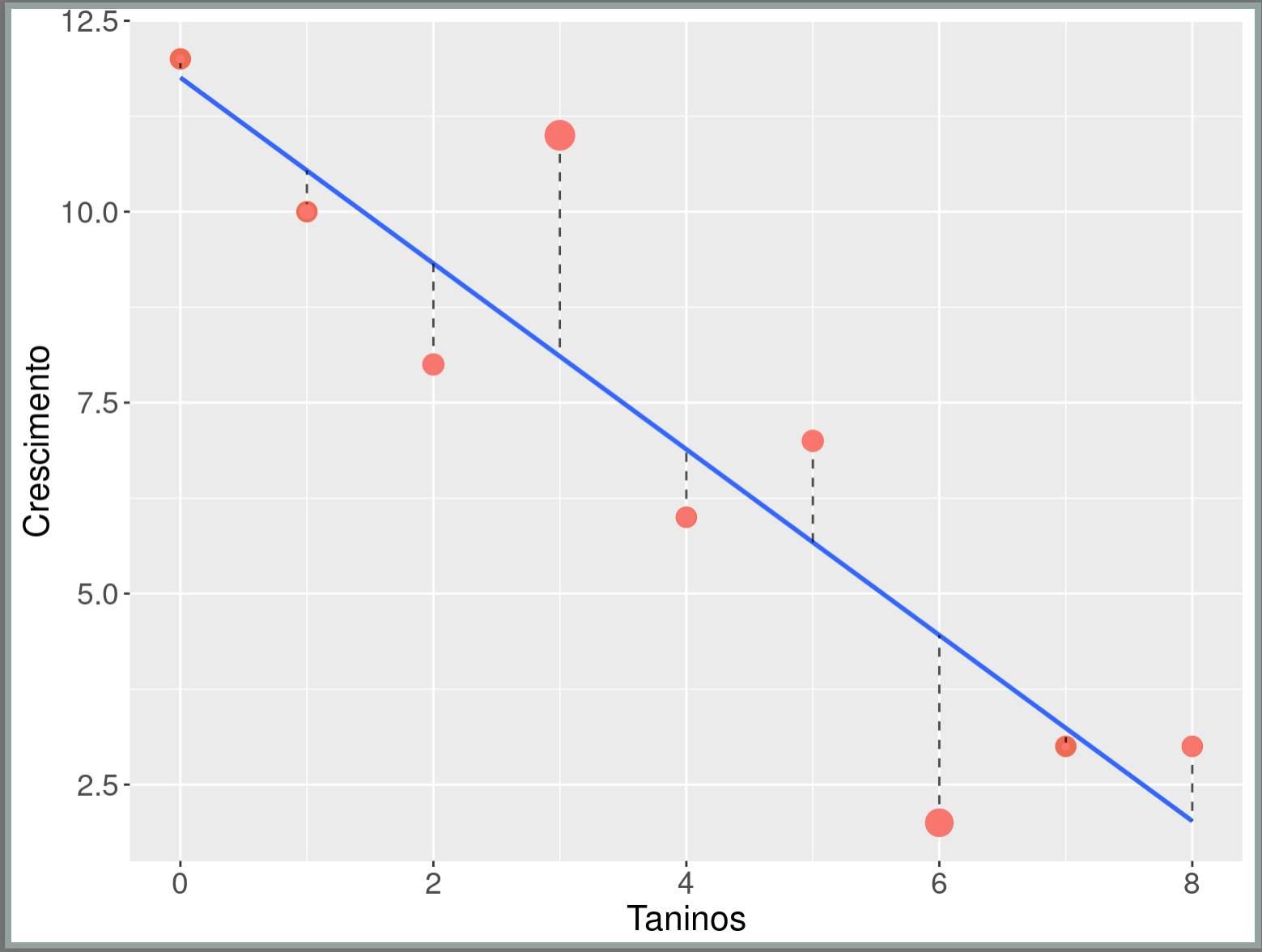


Dieta de Lagarta

Exemplo: dieta de lagarta



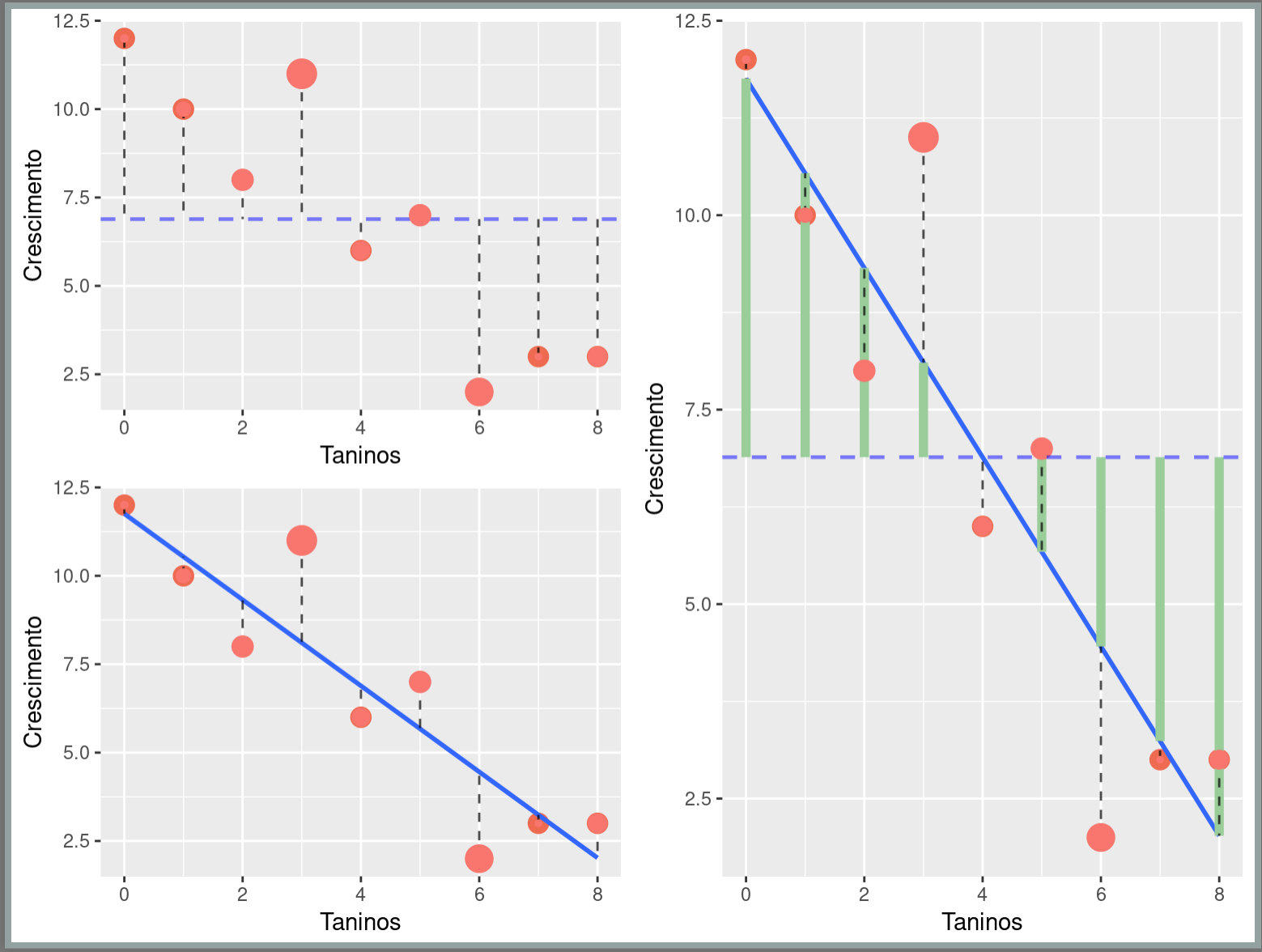
Modelo Linear: lagartas



Modelo Linear: dieta de lagarta

```
##
## Call:
## lm(formula = growth ~ tannin, data = lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4556 -0.8889 -0.2389  0.9778  2.8944
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)  11.7556     1.0408  11.295 9.54e-0
## tannin       -1.2167     0.2186  -5.565 0.00084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 1.693 on 7 degrees of
## Multiple R-squared:  0.8157, Adjusted R-squared
## F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.0
```

Partição da Variância: lagarta



Partição da Variância

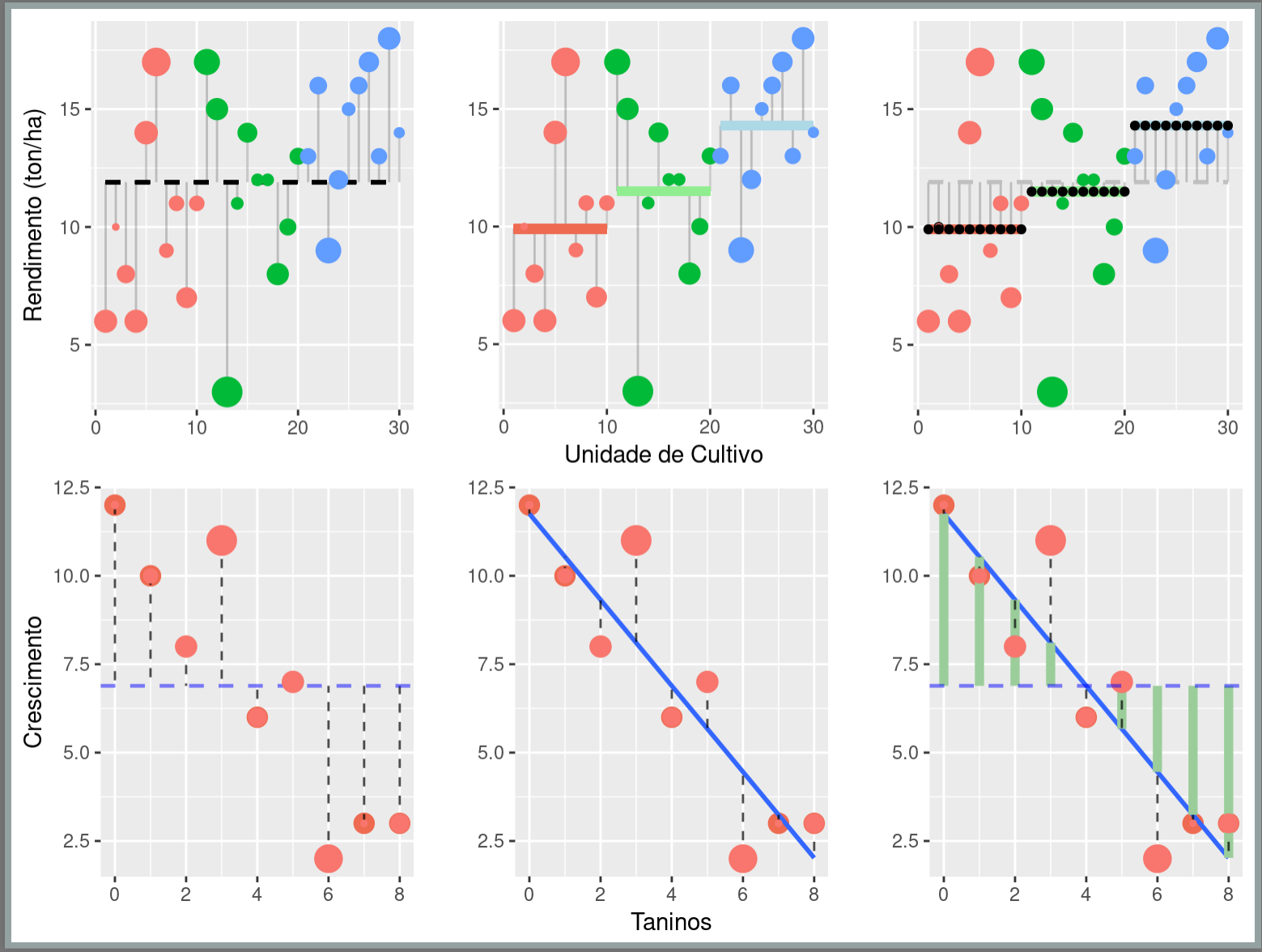
Anova

$$SQ_{total} = SQ_{entre} + SQ_{intra}$$

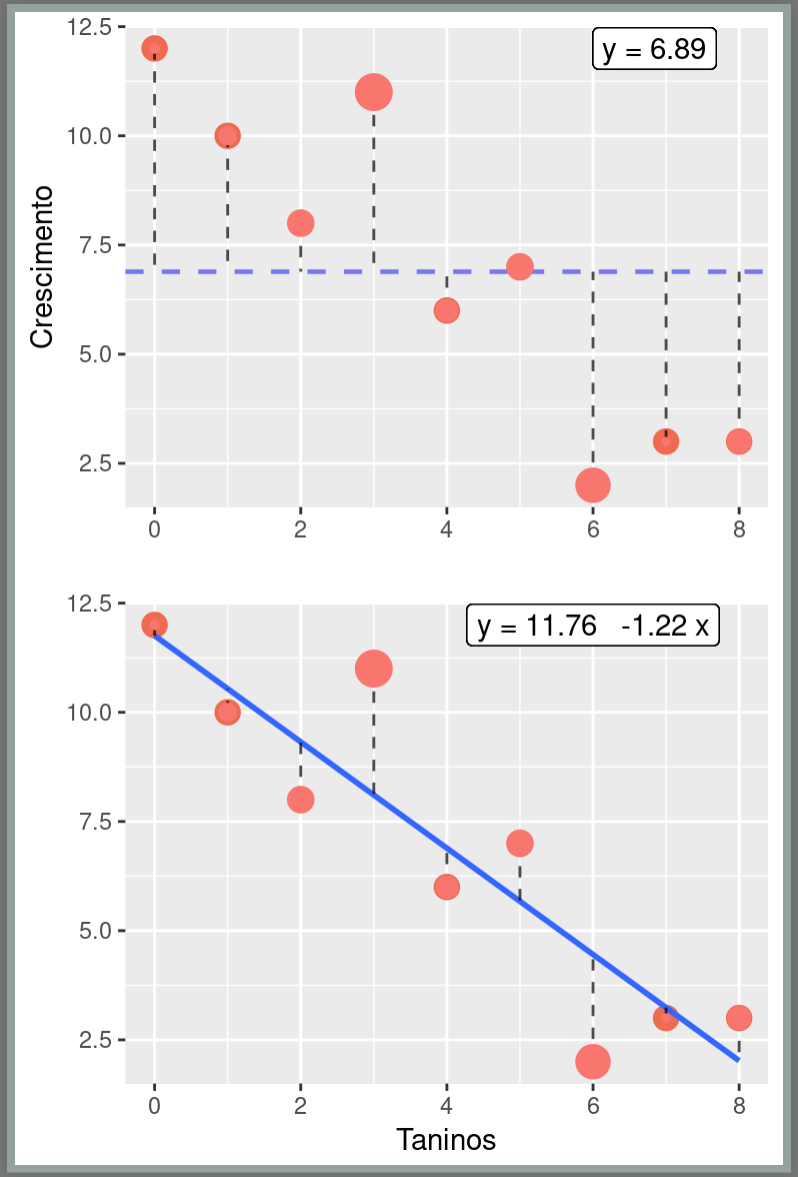
Modelo Linear

$$SQ_{total} = SQ_{mod} + SQ_{erro}$$

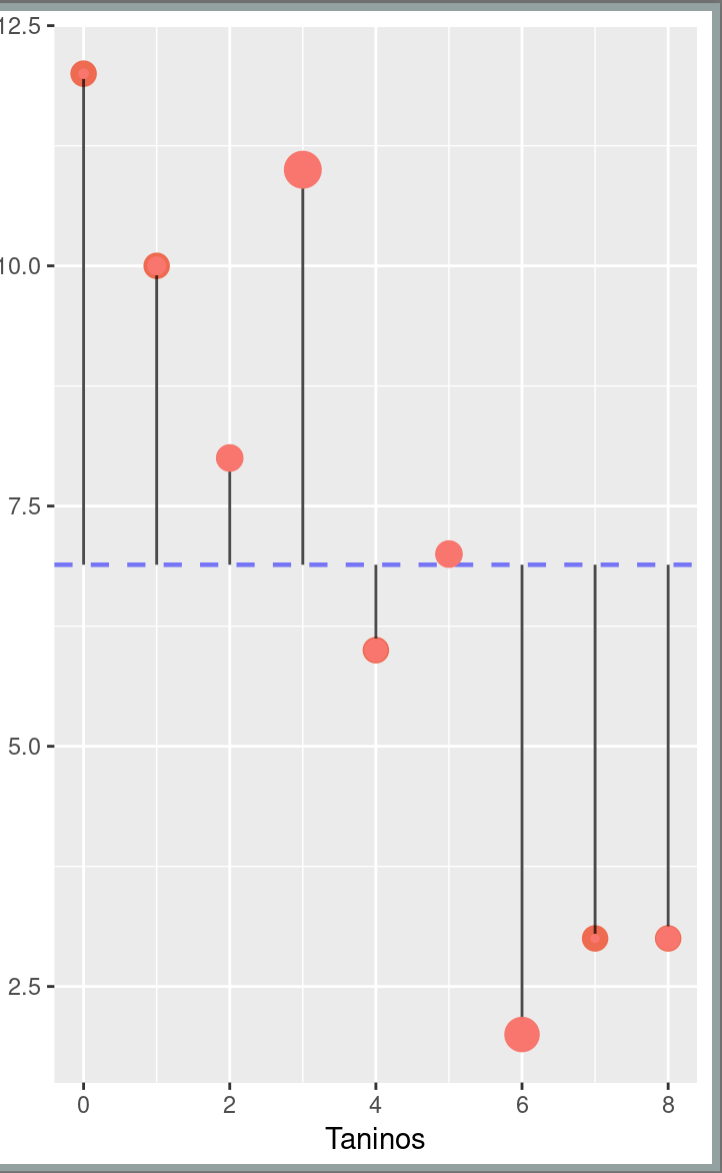
Anova x Modelo Linear



Partição da Variância: modelo linear



Desvios Quadráticos: total

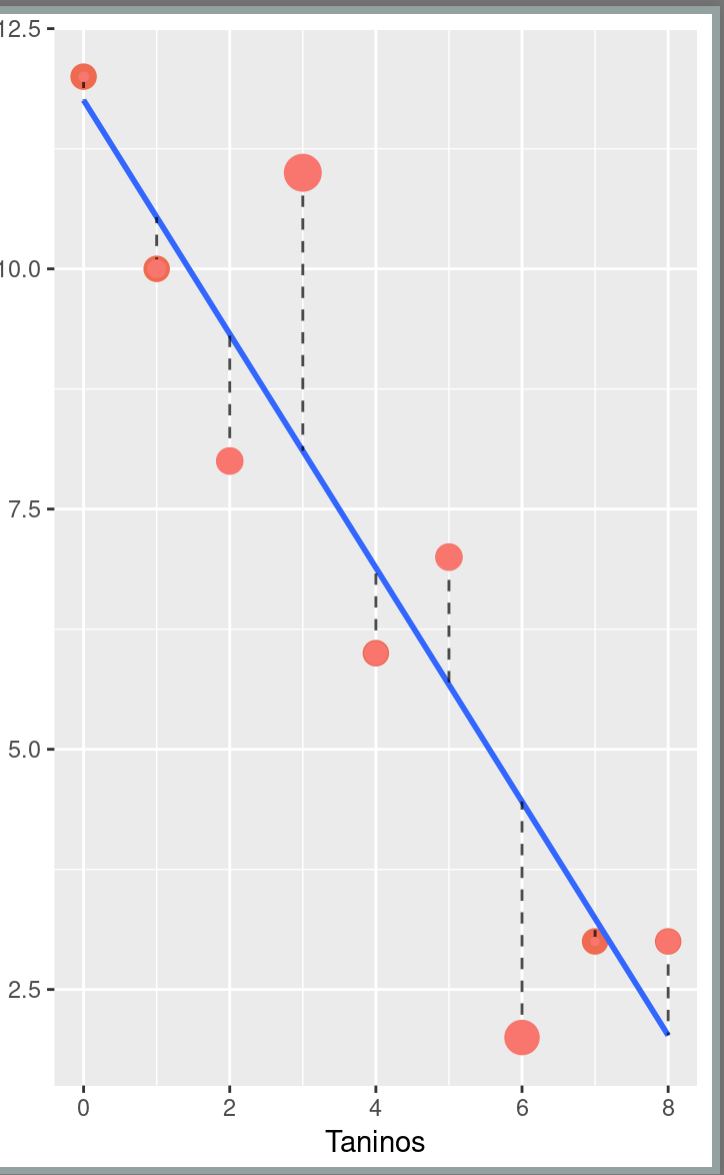


$$SQ_{total} =$$

tannin	growth	me
0	12	
1	10	
2	8	
3	11	
4	6	
5	7	
6	2	
7	3	
8	3	

```
## [1] 10
```

Desvios Quadráticos: resíduo



$$SQ_{error} = \sum$$

tannin	growth	resid
0	12	
1	10	-(
2	8	-
3	11	2
4	6	-(
5	7	
6	2	-2
7	3	-(
8	3	(

```
## [1] 20
```

Desvios Quadraticos do Modelo

Incluir um gráfico com o residuo, a variação total e a variação explicada pelo modelo (o segmento entre o modelo e a média geral.

- Explicar o grau de liberdade da regressão como sendo: dois valores (intercepto e inclinação) para definir o modelo, menos um para definir a grande média.

Lógica do Modelo Linear

$$SQ_{total} = SQ_{mod} + SQ_{erro}$$

$$SQ_{mod} = SQ_{total} - SQ_{erro}$$

Desvios Quadraticos do Modelo

```
## [1] 88.81667
```

Tabela de Anova

Fonte	SQ	GL	MQ
Modelo Linear	88.82	1	88.82
Erro	20.07	7	2.87
Total	108.89	8	

Modelo Linear

Teste de hipótese: F

Fonte	SQ	GL	MQ	F	pvalor
Modelo Linear	88.82	1	88.82	30.97	0.00085
Erro	20.07	7	2.87		
Total	108.89	8			

Modelo Linear: lagarta

Tabela de Anova no R

```
anova(lmlag)
```

```
## Analysis of Variance Table
##
## Response: growth
##           Df Sum Sq Mean Sq F value    Pr(>F)
## tannin     1 88.817  88.817  30.974 0.0008461 *
## Residuals  7 20.072   2.867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
```

Fonte	SQ	GL	MQ	F	pvalor
Modelo Linear	88.82	1	88.82	30.97	0.00085
Erro	20.07	7	2.87		
Total	108.89	8			

Coeficiente de Determinação

$$R^2 = \frac{SQ_{mod}}{SQ_{total}}$$

Fonte	SQ	GL	MQ	F	pvalor
Modelo Linear	88.82	1	88.82	30.97	0.00085
Erro	20.07	7	2.87		
Total	108.89	8			

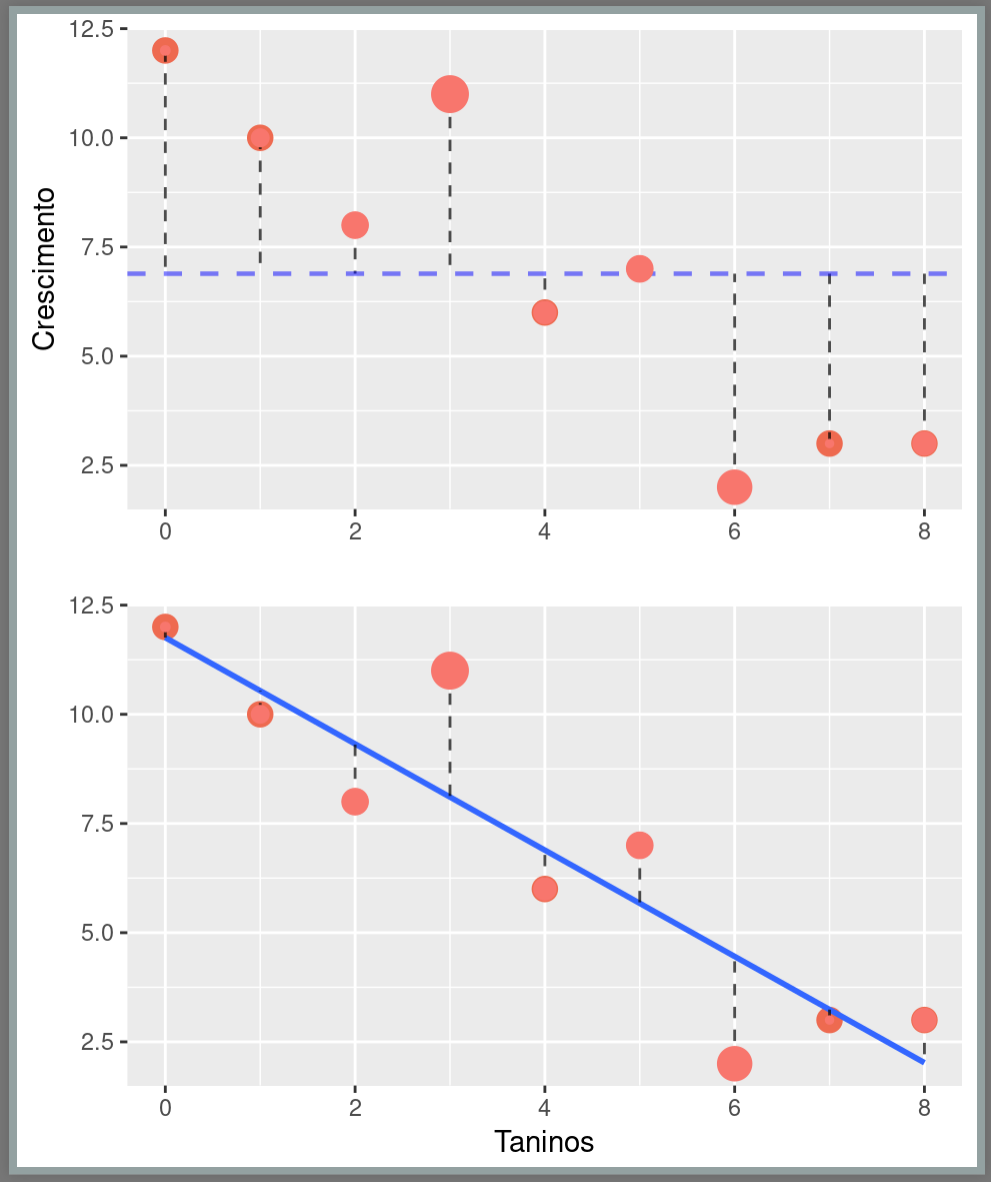
$$R^2$$

```
## [1] 0.816
```

Resumo do Modelo

```
##
## Call:
## lm(formula = growth ~ tannin, data = lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4556 -0.8889 -0.2389  0.9778  2.8944
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)  11.7556     1.0408  11.295 9.54e-0
## tannin       -1.2167     0.2186  -5.565 0.00084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 1.693 on 7 degrees of
## Multiple R-squared:  0.8157, Adjusted R-squared
## F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.0
```

Comparando Modelos

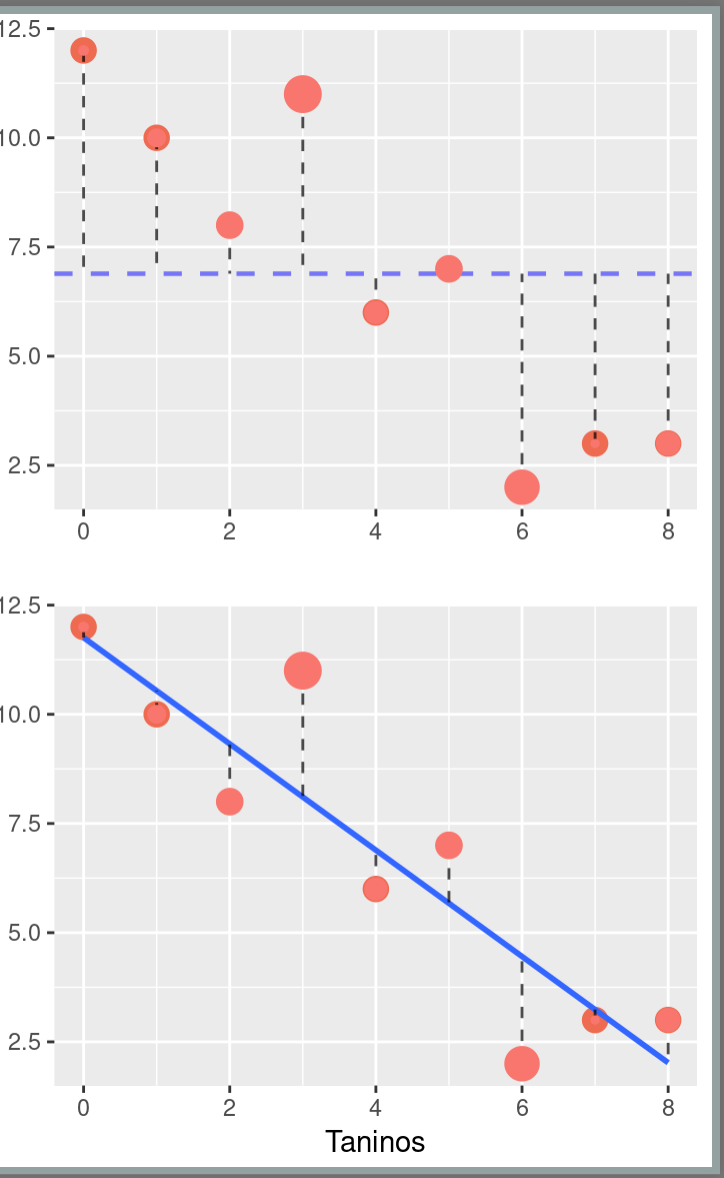


Modelo Mínimo (nulo)

```
nullag <- lm(growth ~ 1, data = lag)
summary(nullag)
```

```
##
## Call:
## lm(formula = growth ~ 1, data = lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8889 -3.8889  0.1111  3.1111  5.1111
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)    6.889      1.230   5.602 0.00050
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 3.689 on 8 degrees of
```

Comparando Modelos




```
anova(nullag, lmlag)
```

```
## Analysis of Variance Table
##
## Model 1: growth ~ 1
## Model 2: growth ~ tannin
##   Res.Df     RSS Df Sum of Sq
## 1      8 108.889
## 2      7  20.072  1    88.817
## ---
## Signif. codes:  0 '***' 0.001
```

Comparando Modelos

Anova do modelo: anova(lmlag)

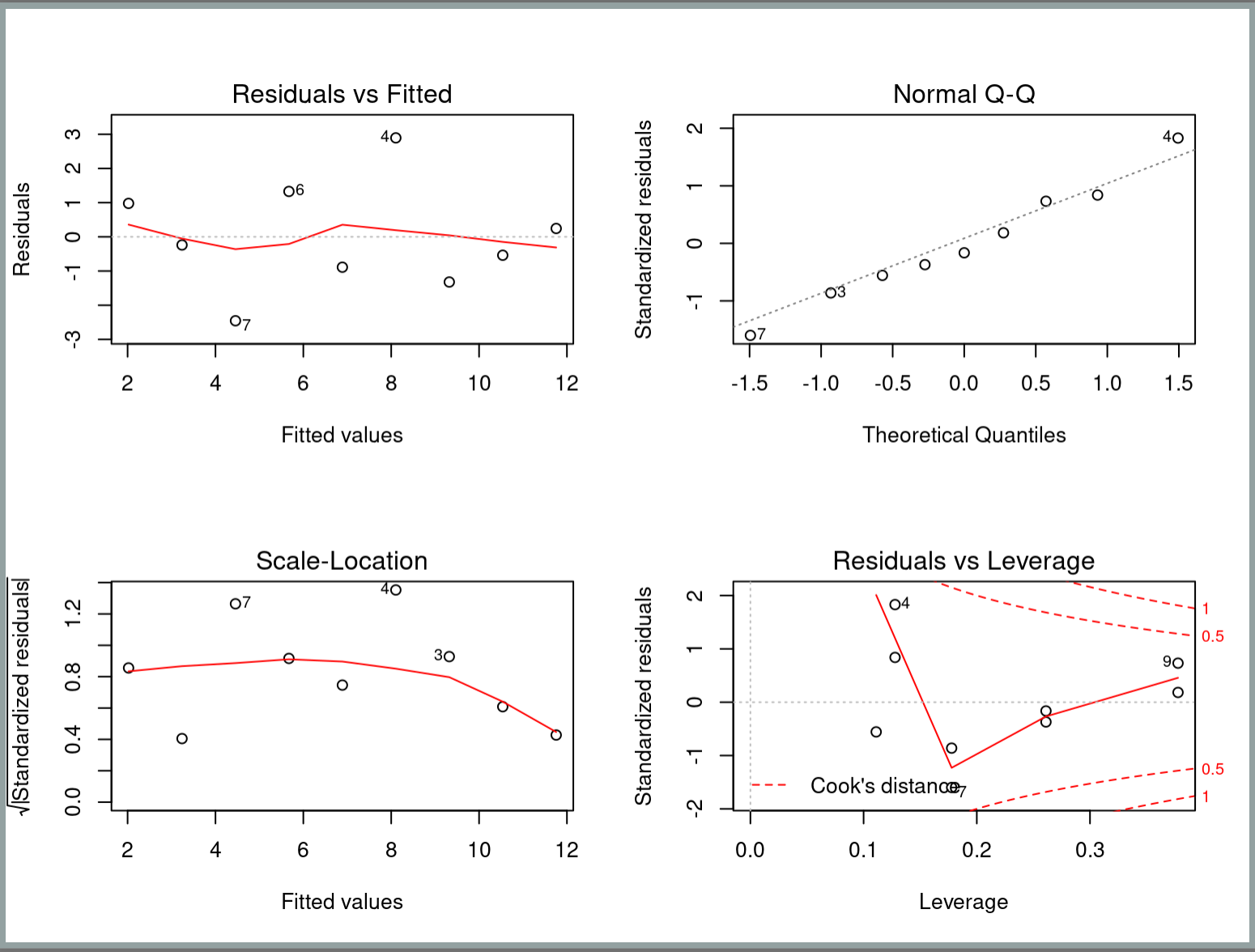
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tannin	1	88.81667	88.81667	30.97398	0.0008461
Residuals	7	20.07222	2.86746		

Anova da comparação de modelos: anova(nullag, lmlag)

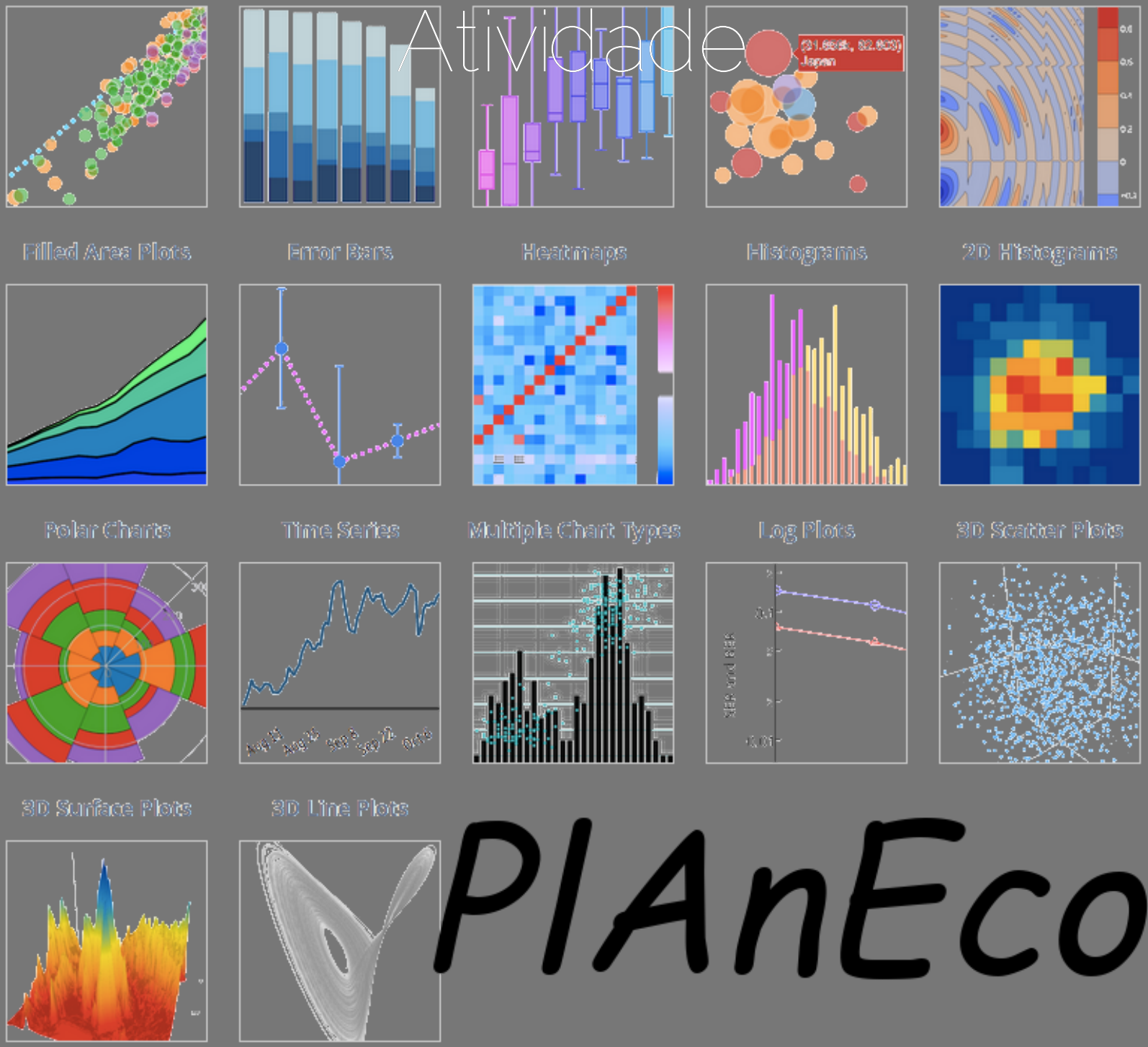
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
8	108.88889				
7	20.07222	1	88.81667	30.97398	0.0008461

NÃO DESESPERE, ESPERE! KEEP CALM!!

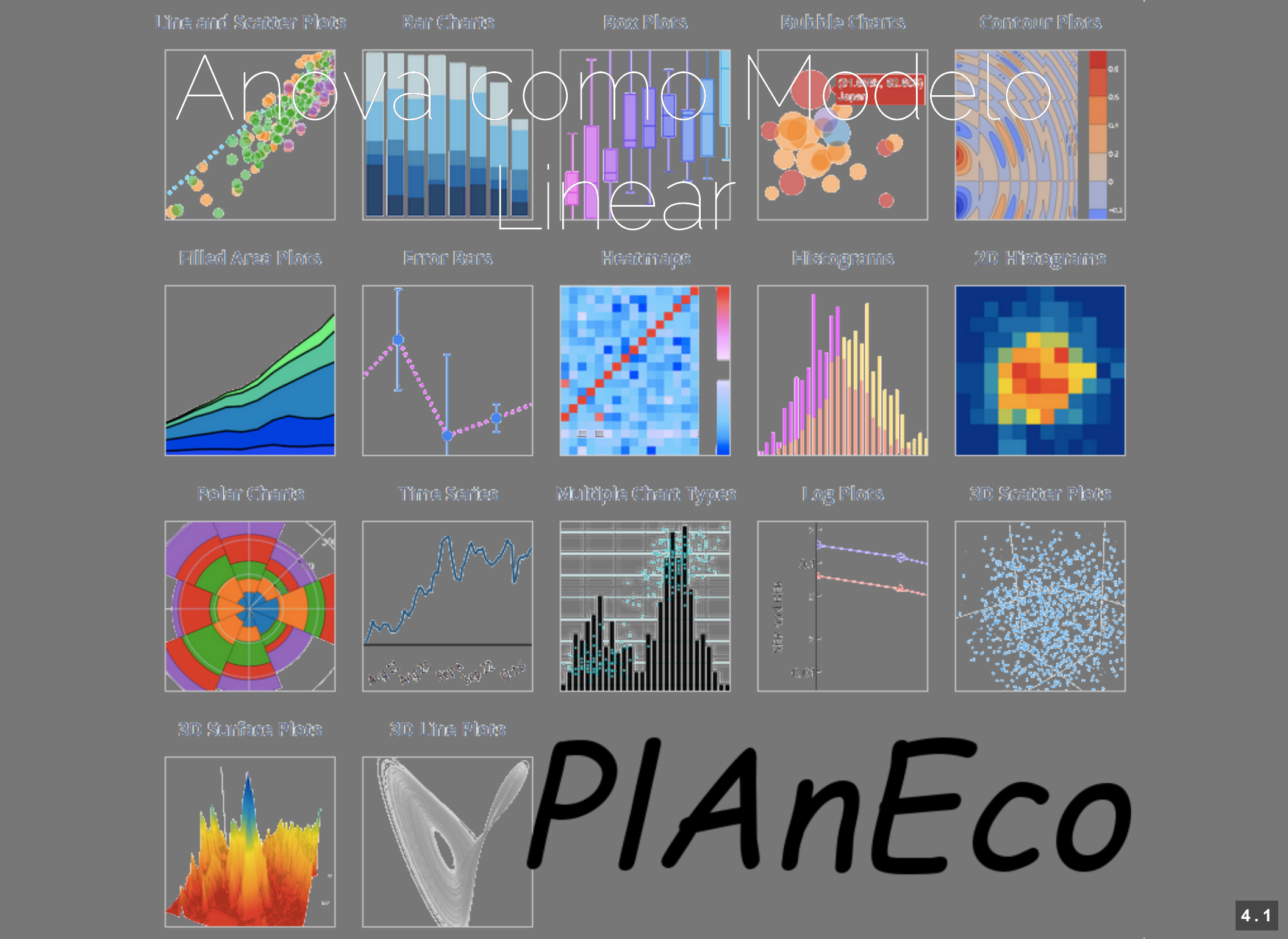
Diagnóstico do Modelos



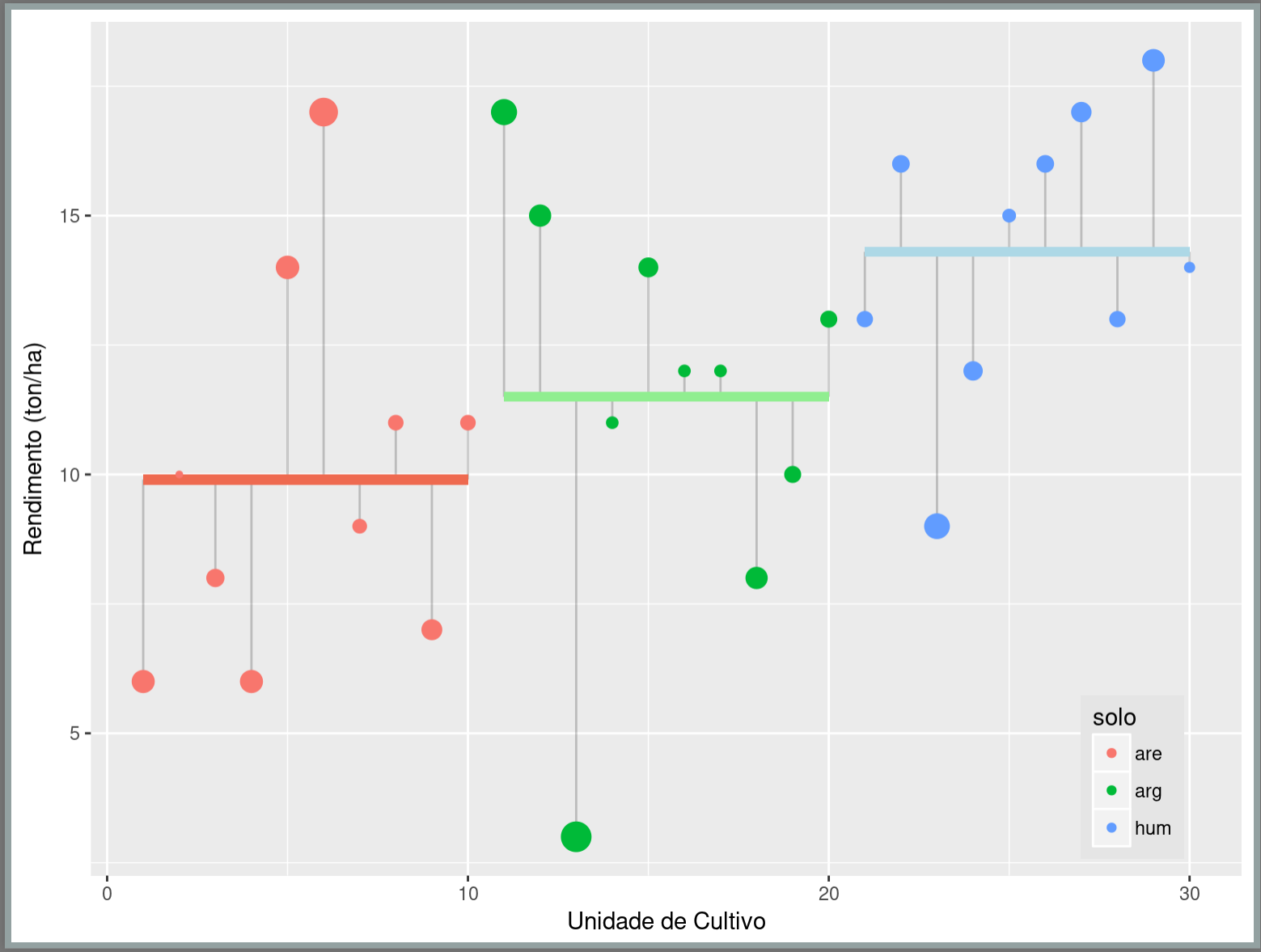
Atividade



PlAnEco


Line and Scatter Plots
Bar Charts
Box Plots
Bubble Charts
Contour Plots
Filled Area Plots
Error Bars
Heatmaps
Histograms
2D Histograms
Polar Charts
Time Series
Multiple Chart Types
Log Plots
3D Scatter Plots
3D Surface Plots
3D Line Plots
Anova como Modelo Linear
PlAnEco
4.1

Modelo linear: variável categórica?!



Rendimento Colheita: dados

	solo	colhe
1	are	6
2	are	10
3	are	8
4	are	6
5	are	14
11	arg	17
12	arg	15
13	arg	3
14	arg	11
15	arg	14
21	hum	13
22	hum	16
23	hum	9
24	hum	12
25	hum	15

Variavel Dammy ou Indicadora

	colhe	solo	arg	hum
1	6	are	0	0
2	10	are	0	0
3	8	are	0	0
4	6	are	0	0
5	14	are	0	0
11	17	arg	1	0
12	15	arg	1	0
13	3	arg	1	0
14	11	arg	1	0
15	14	arg	1	0
21	13	hum	0	1
22	16	hum	0	1
23	9	hum	0	1
24	12	hum	0	1
25	15	hum	0	1

Número de níveis do fator menos 1 (intercepto)

Modelo linear: dummy

odelo

$$= \alpha_{d_1} + \beta_2 x_{d_2} + \beta_3 x_{d_3}$$

tercepto:

$$_1 = \bar{x}_1$$

oeficientes:

$$= \bar{x}_2 - \bar{x}_1$$

$$= \bar{x}_3 - \bar{x}_1$$

	colhe	s
1	6	a
2	10	a
3	8	a
4	6	a
5	14	a
11	17	a
12	15	a
13	3	a
14	11	a
15	14	a
21	13	b
22	16	b
23	9	b
24	12	b
25	15	b

Variável Dummy

```
##
## Call:
## lm(formula = colhe ~ arg + hum, data = croplin)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
##   -8.5    -1.8     0.3     1.7     7.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)    9.900      1.081   9.158 9.04e-1
## arg            1.600      1.529   1.047   0.3045
## hum            4.400      1.529   2.878   0.0077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 3.418 on 27 degrees of
## Multiple R-squared:  0.2392, Adjusted R-squared
## F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.
```

Modelo Linear Normal

```
##
## Call:
## lm(formula = colhe ~ solo, data = crop)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
##   -8.5    -1.8     0.3     1.7     7.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|
## (Intercept)    9.900      1.081   9.158 9.04e-1
## soloarg        1.600      1.529   1.047   0.3045
## solohum        4.400      1.529   2.878   0.0077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 3.418 on 27 degrees of
## Multiple R-squared:  0.2392, Adjusted R-squared
## F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.
```

Coeficientes do modelo

```
coef(lmdum)
```

```
## (Intercept)         arg         hum
##         9.9         1.6         4.4
```

```
tapply(crop$colhe, crop$solo, mean)
```

```
##  are  arg  hum
##  9.9 11.5 14.3
```

$$y = \hat{\alpha}_{d_1} + \hat{\beta}_2 x_{d_2} + \hat{\beta}_3 x_{d_3}$$

	colhe	solo	arg	hum
1	6	are	0	0
2	10	are	0	0
3	8	are	0	0
11	17	arg	1	0
12	15	arg	1	0
13	3	arg	1	0
21	13	hum	0	1
22	16	hum	0	1
23	9	hum	0	1

Modelo

$$y = \alpha_{d_1} + \beta_2 x_{d_2} + \beta_3 x_{d_3}$$

Intercepto:

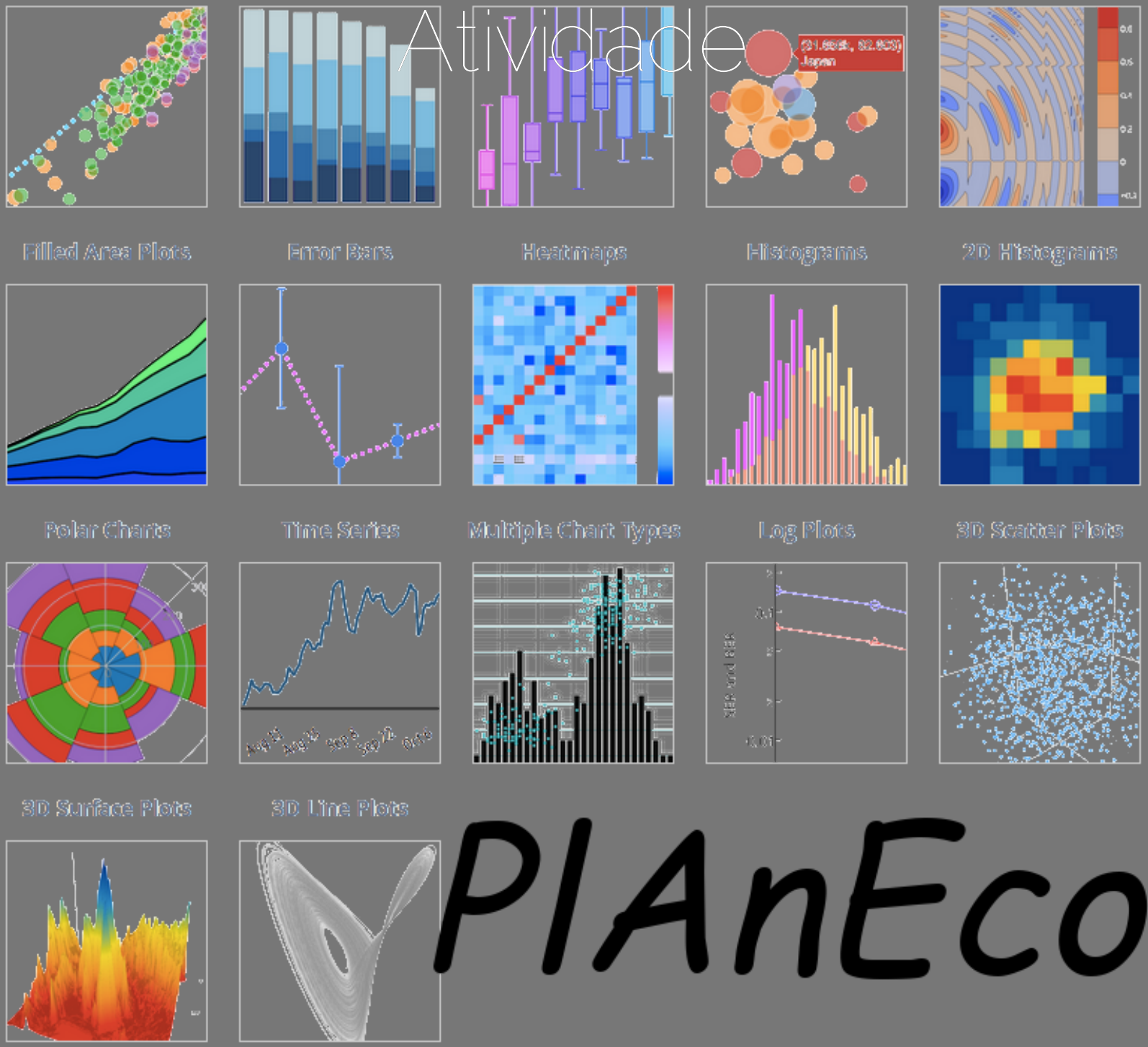
$$\alpha_{d_1} = \bar{x}_1$$

Coeficientes:

$$\beta_2 = \bar{x}_2 - \bar{x}_1$$

$$\beta_3 = \bar{x}_3 - \bar{x}_1$$

Atividade



PlAnEco