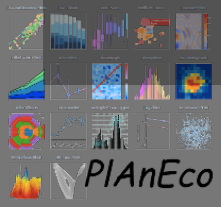
Modelos Lineares

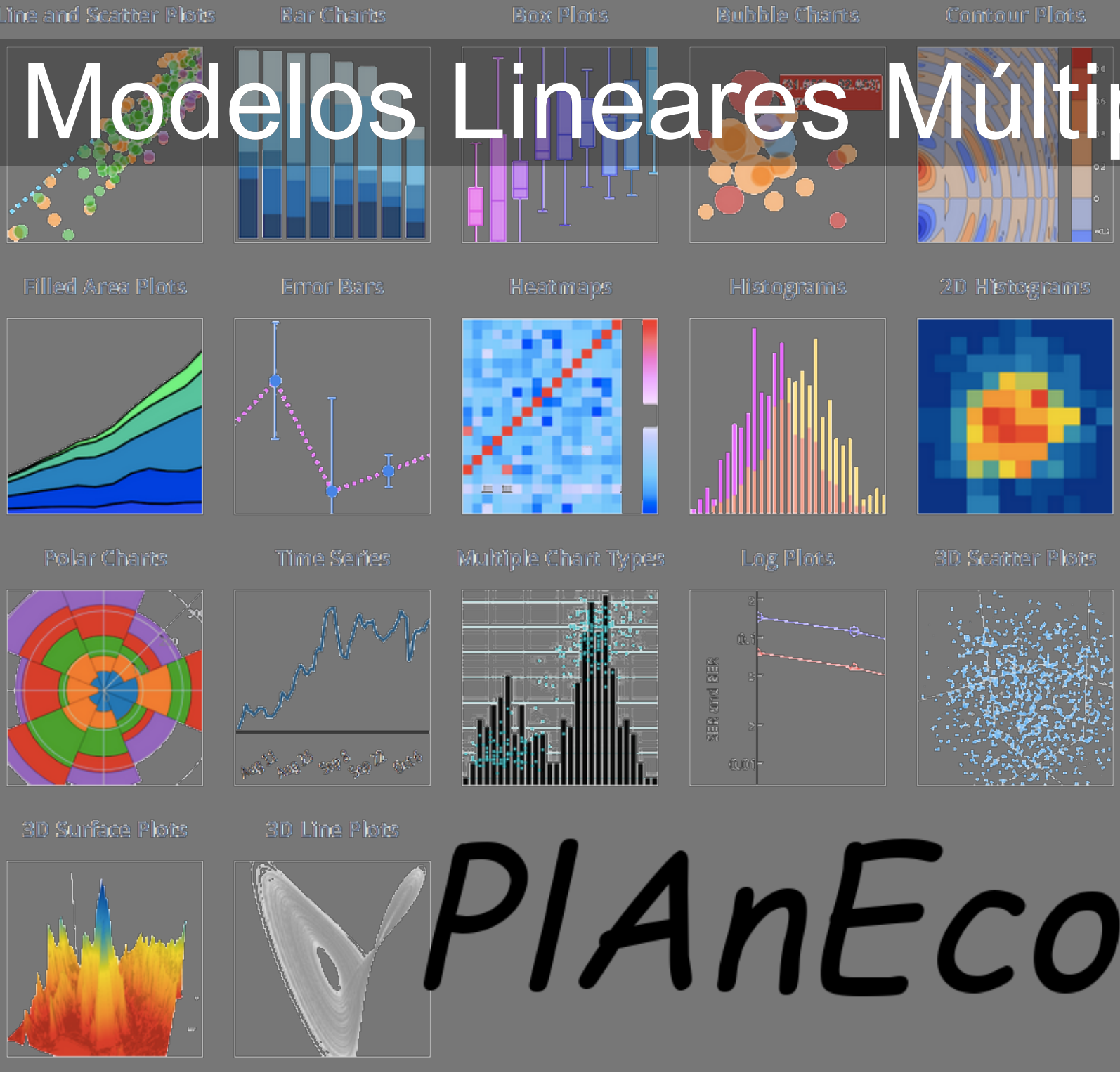
múltiplas preditoras

Alexandre Adalardo de Oliveira

PIAnEco 2024



Modelos Lineares Múltiplos



PlAnEco

Conceitos

- predito pelo modelo
- interação entre preditoras
- interpretação de variável dummy
- simplificação do modelo
- colinearidade

Modelo Linear

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

Modelo Linear Múltiplo

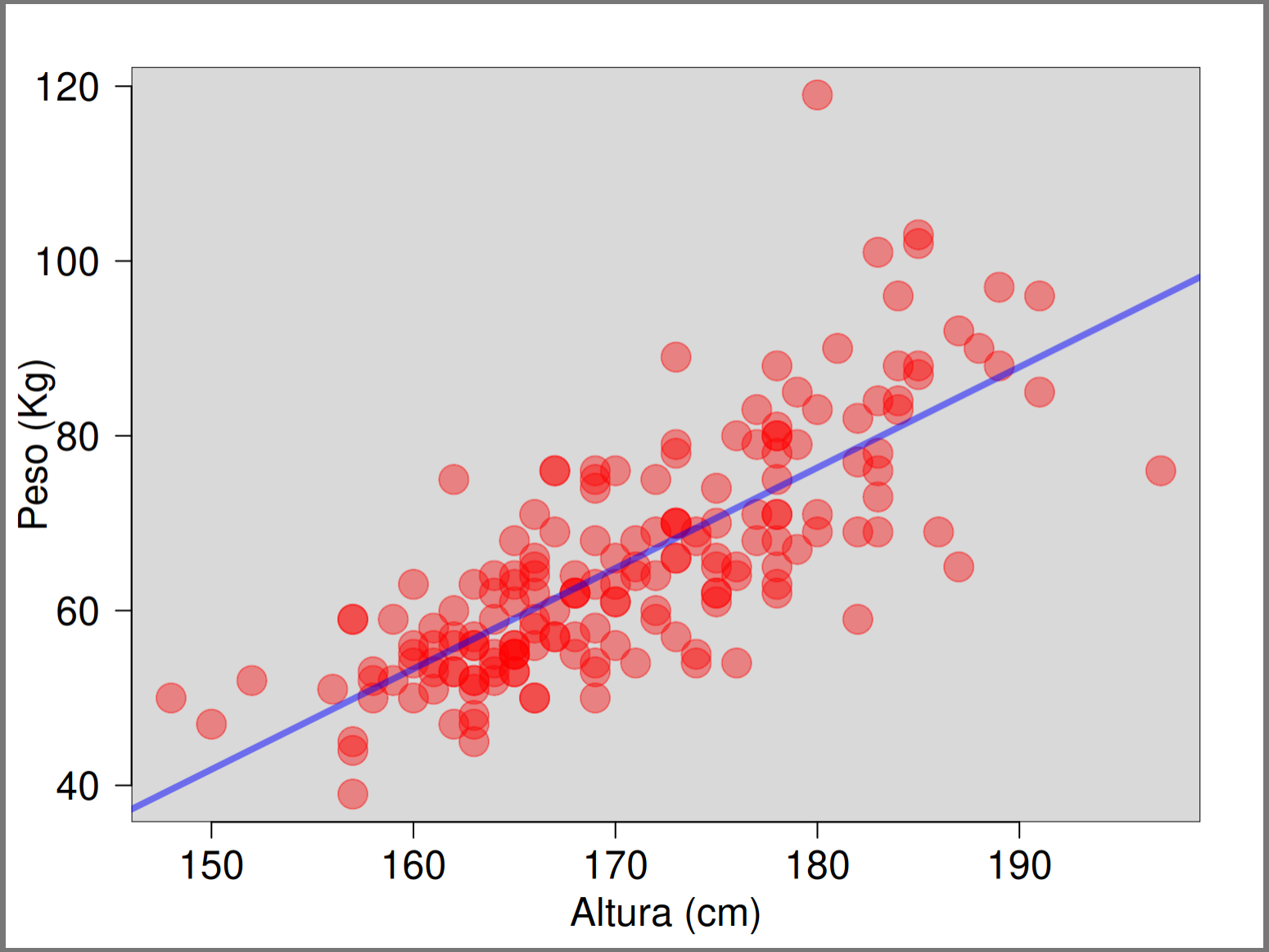
$$y = \alpha + \sum \beta_i x_i + \epsilon$$
$$\epsilon = N(0, \sigma)$$

Retomando o Modelo Linear

Davis (1990)

sex	weight	height
M	77	182
F	58	161
F	53	161
M	68	177
F	59	157
F	51	156
F	62	164
M	74	175
M	83	180
M	90	181
M	79	177

weight ~ heigth



Modelo Linear

```
summary(lmdavis)
```

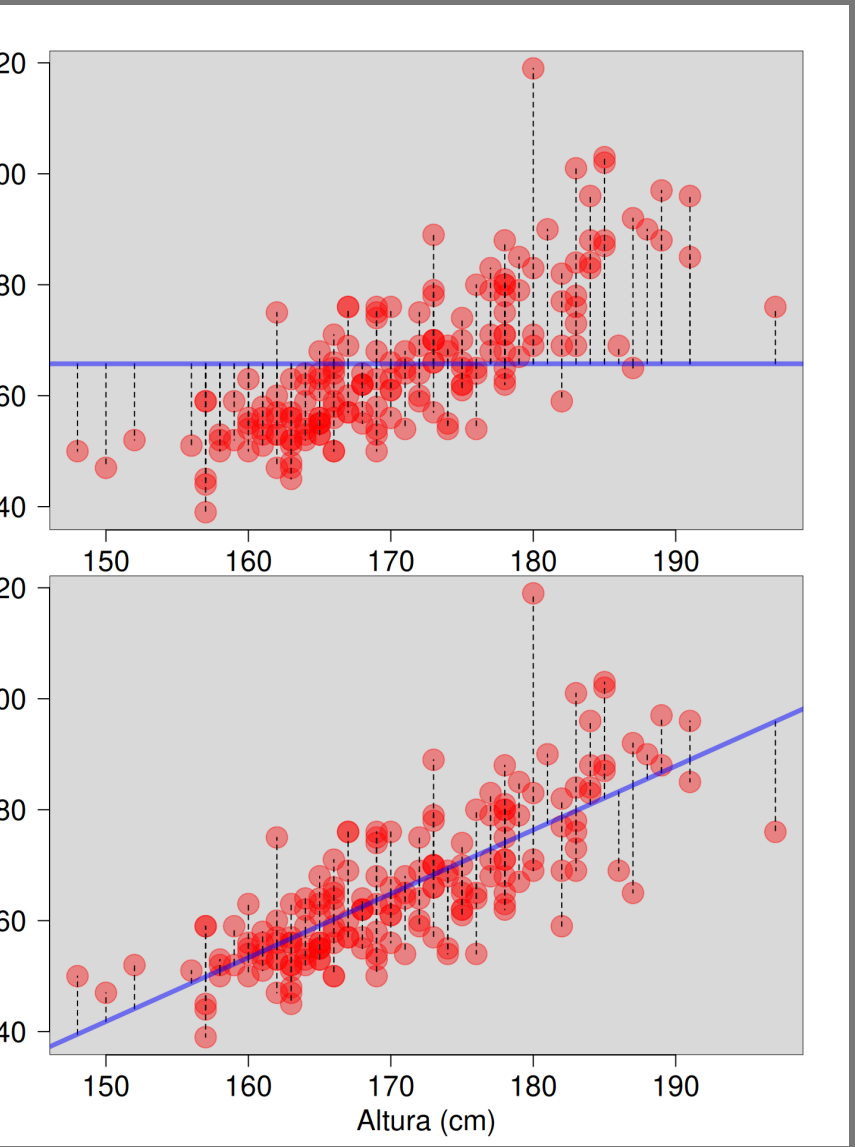
```
Call:
lm(formula = weight ~ height, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.928  -5.406  -0.651   4.891  42.641 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -130.84185   12.30184  -10.64   <2e-16 ***
height         1.15112    0.07193   16.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.635 on 178 degrees of freedom
Multiple R-squared:  0.5899,    Adjusted R-squared:  0.5876 
F-statistic: 256.1 on 1 and 178 DF,  p-value: < 2.2e-16
```

Partição da Variação

```
anova(lmdavis)
```

```
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq F value    Pr(>F)
height      1  19095 19095.0  256.08 < 2.2e-16 ***
Residuals 178  13273    74.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo Mínimo



lm(weigt

```
anova(davisNull,
```

```
Analysis of Vari

Model 1: weight
Model 2: weight
  Res.Df   RSS D
1    179 32368
2    178 13273
---
Signif. codes:
```

$p_{valor} =$

$p_{valor} = 2$

$R^2 =$

Interpretação dos coeficientes

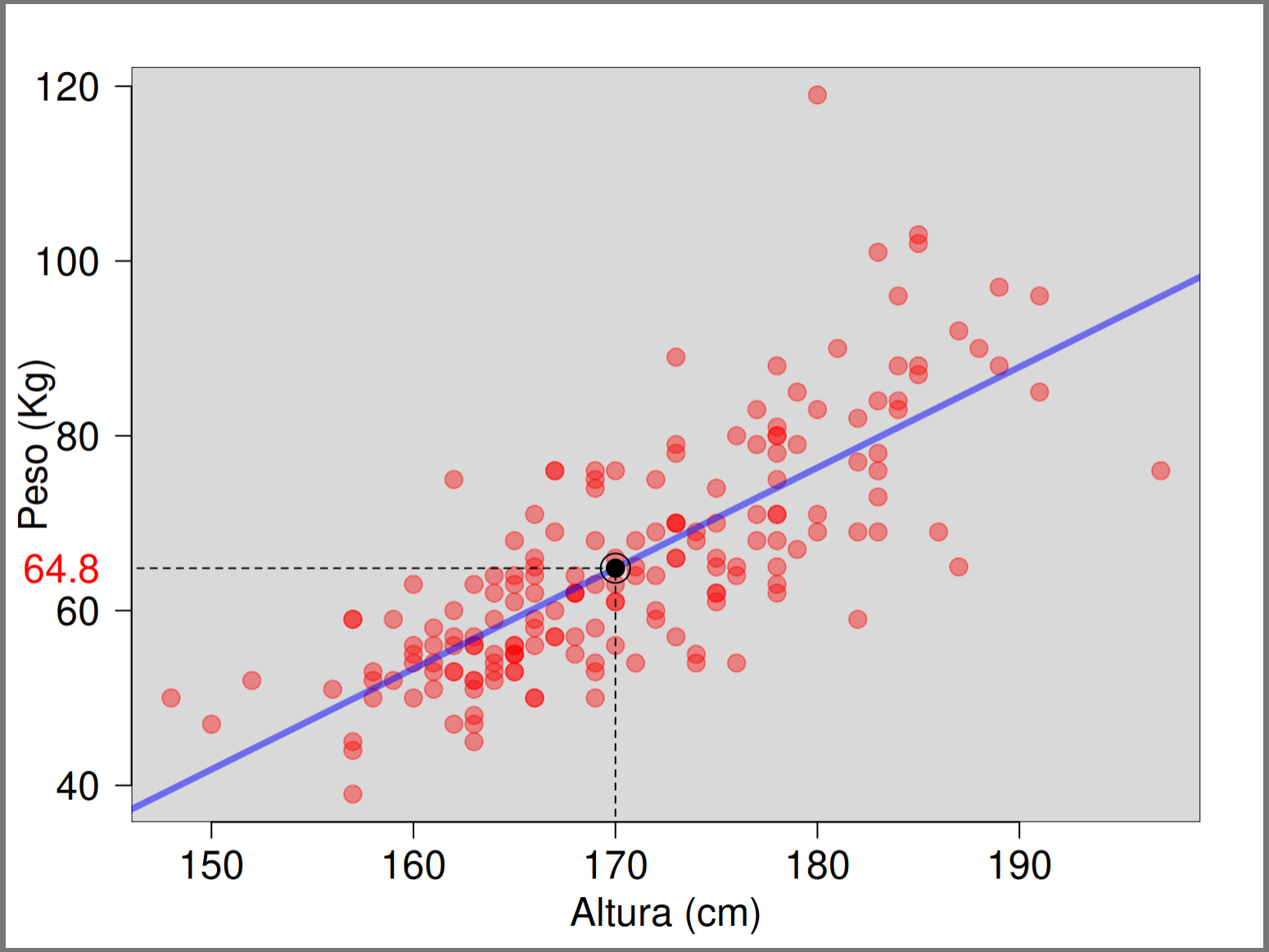
```
(Intercept)      height
-130.841846    1.151118
```

Uma pessoa com 170 cm de altura

$$y_{170} = -130.84 + 1.15 * 170$$

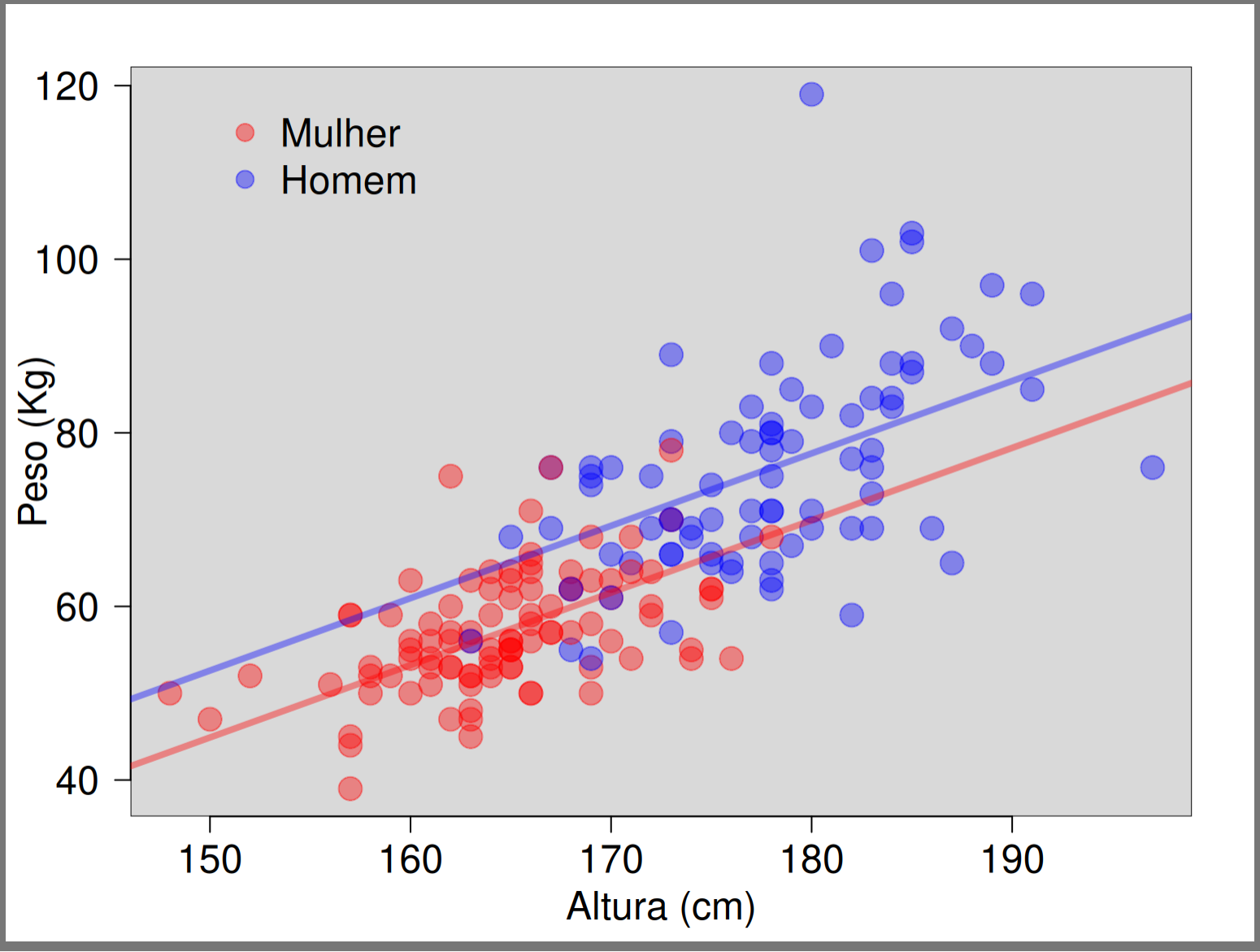
$$y_{170} = 64.85$$

Predito pelo modelo



Multiplas preditoras

lm(weight ~ height + sex, data = Davis)



Resumo do Modelo

sexo

dummy: mulher = 0, homem = 1

```
summary(lmdavis01)
```

```
Call:
lm(formula = weight ~ height + sex, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.302  -4.808  -0.335   5.239  41.366 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -80.2107    16.8415  -4.763 3.96e-06 ***
height        0.8341     0.1021   8.169 5.71e-14 ***
sexM          7.7070     1.8345   4.201 4.20e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.258 on 177 degrees of freedom
Multiple R-squared:  0.6271,    Adjusted R-squared:  0.6229 
F-statistic: 148.8 on 2 and 177 DF,  p-value: < 2.2e-16
```

Interpretando o modelo

lm(weight ~ height + sex, data = Davis)

```
(Intercept)      height        sexM
-80.2107328   0.8340964   7.7070166
```

Mulher (sex = 0)

$$w_f = \hat{\alpha} + \hat{\beta}_s * sex + \hat{\beta}_h * height$$

$$w_f = \hat{\alpha} + \hat{\beta}_h * height$$

$$w_f = -80.2 + 0.83 * height$$

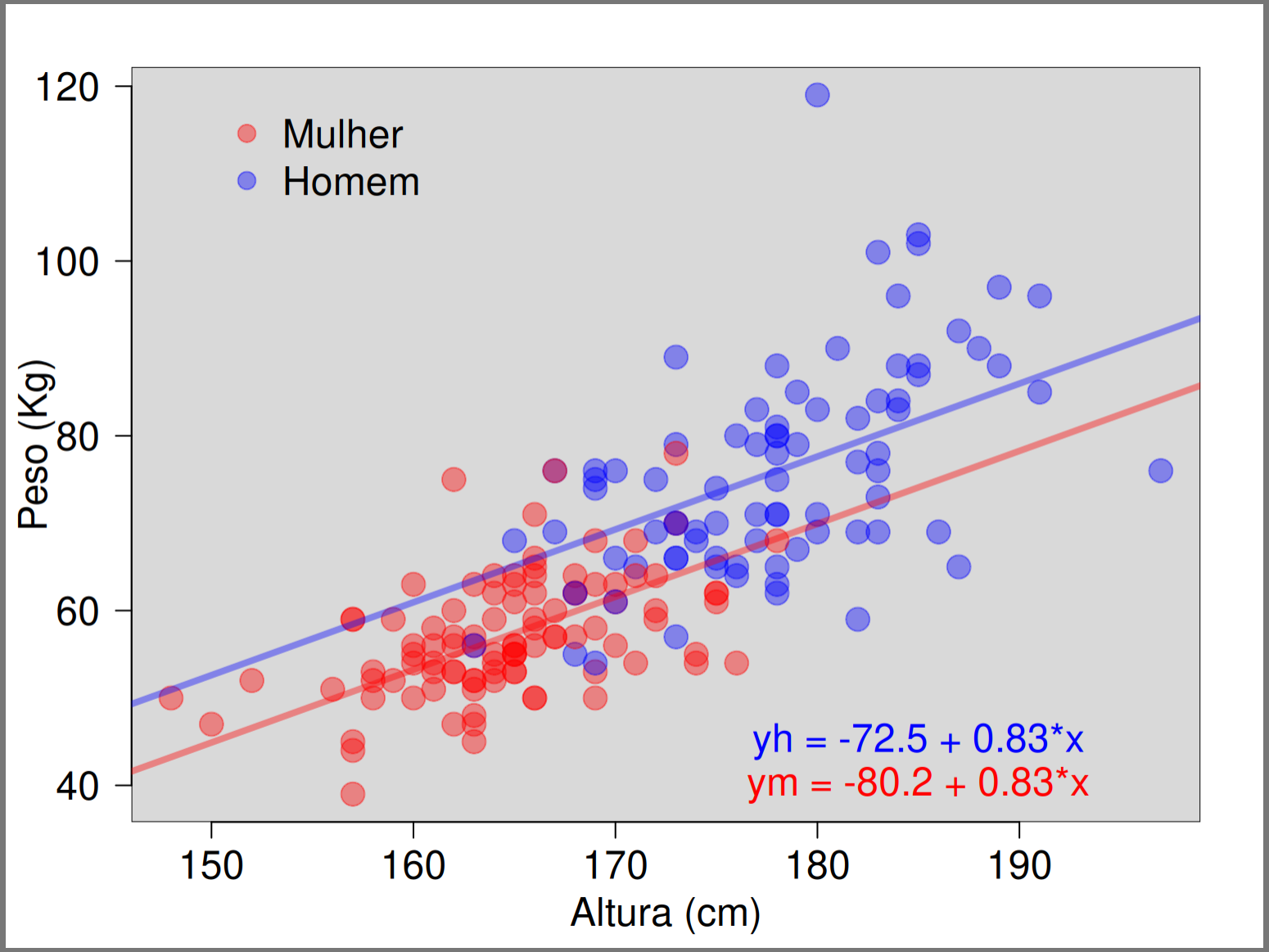
Homem (sex=1)

$$w_m = \hat{\alpha} + \hat{\beta}_s * sex + \hat{\beta} * height$$

$$w_m = \hat{\alpha} + \hat{\beta}_s + \hat{\beta}_h * height$$

$$w_m = -72.5 + 0.83 * height$$

weight ~ height + sex

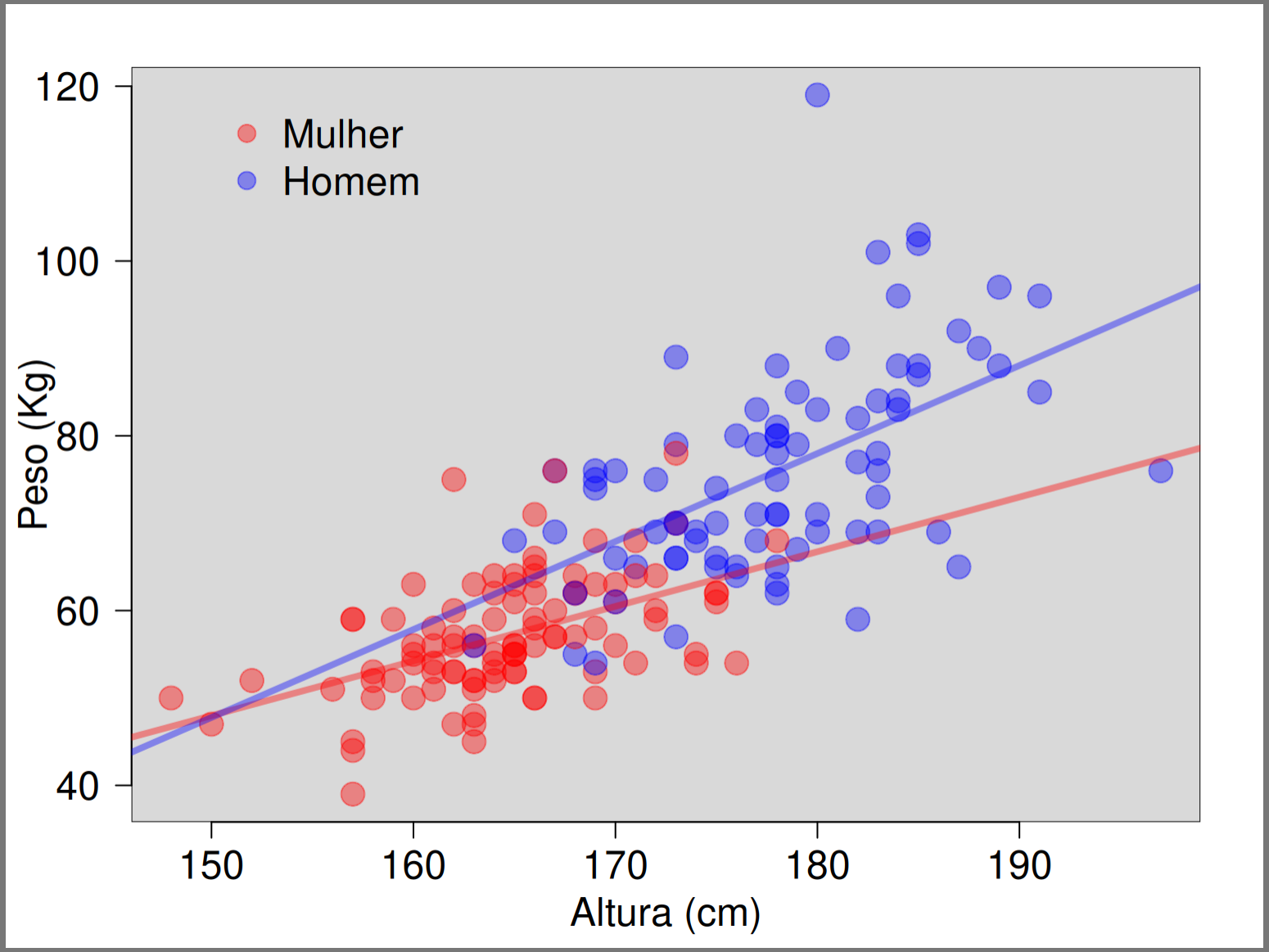


weight ~ height + sex + height:sex

```
lmdavisfull <- lm(weight ~ height + sex + sex:height, data = Davis)
lmdavisfull <- lm(weight ~ height * sex, data=Davis)
```

```
Call:
lm(formula = weight ~ height + sex + sex:height, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.990  -4.548  -0.926   4.821  41.023 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -45.7988    24.8453  -1.843   0.0670 .  
height        0.6252     0.1507   4.148 5.22e-05 ***
sexM        -57.4326    34.8293  -1.649   0.1009    
height:sexM   0.3815     0.2037   1.873   0.0628 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.2 on 176 degrees of freedom
Multiple R-squared:  0.6344,    Adjusted R-squared:  0.6282 
F-statistic: 101.8 on 3 and 176 DF,  p-value: < 2.2e-16
```

Interação: height:sex



lm(weight ~ height + sex + sex:height)

```
(Intercept)      height        sexM height:sexM
-45.7988220   0.6252035 -57.4326307   0.3815088
```

Mulher (sex = 0)

$$w = \hat{\alpha} + \hat{\beta}_s * sex + \hat{\beta}_h * height + \hat{\beta}_{s:h} * sex * height$$
$$w_f = \hat{\alpha} + \hat{\beta}_h * height$$
$$w_f = -45.80 + 0.62 * height$$

Homem (sex = 1)

$$w = \hat{\alpha} + \hat{\beta}_s * sex + \hat{\beta}_h * height + \hat{\beta}_{h:s} * sex * height$$
$$w_h = \hat{\alpha} + \hat{\beta}_s + (\hat{\beta}_h + \hat{\beta}_{h:s}) * height$$
$$w_h = -103.23 + 1.01 * height$$

Predição do modelo

Uma mulher de 161 cm de altura

$$w = \hat{\alpha} + \hat{\beta}_s sex + \hat{\beta}_h height + \hat{\beta}_{s:h} sex * height$$

$$sex = 0$$

```
(Intercept)      height        sexM height:sexM
-45.7988220   0.6252035 -57.4326307   0.3815088
```

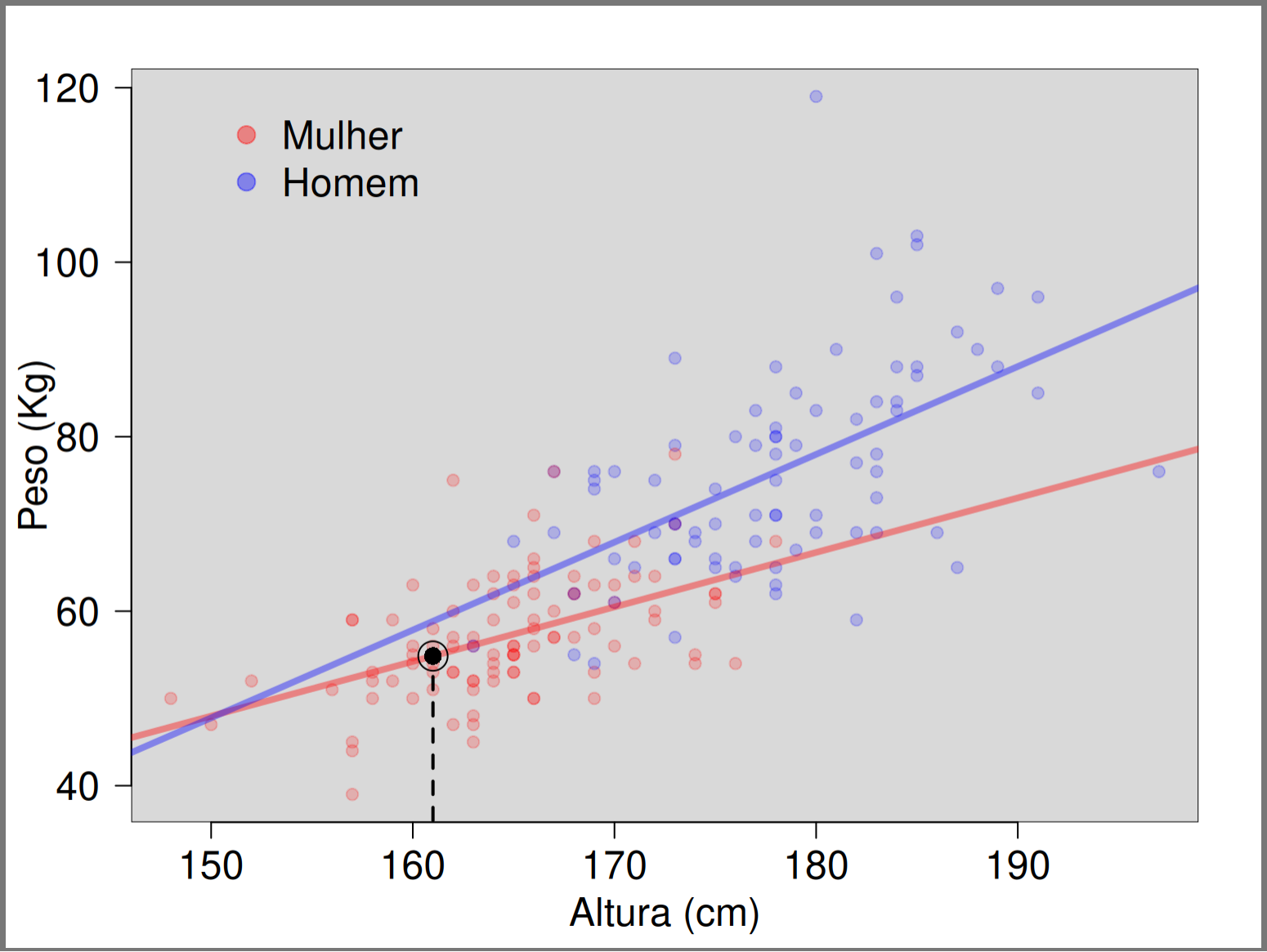
$$w = \hat{\alpha} + \hat{\beta}_h height$$

```
 -45.8 + 0.625 * 161
```

```
[1] 54.85893
```

Predito pelo modelo

- Uma mulher com 161cm de altura tem peso 54.86 Kg.



Predito do Modelo

Homem com 182 cm

$$w = \hat{\alpha} + \hat{\beta}_s sex + \hat{\beta}_h height + \hat{\beta}_{s:h} sex * height$$
$$sex = 1$$

```
(Intercept)      height        sexM height:sexM
-45.7988220   0.6252035 -57.4326307   0.3815088
```

$$w = \hat{\alpha} + \hat{\beta}_s + \hat{\beta}_h * height + \hat{\beta}_{s:h} * height$$

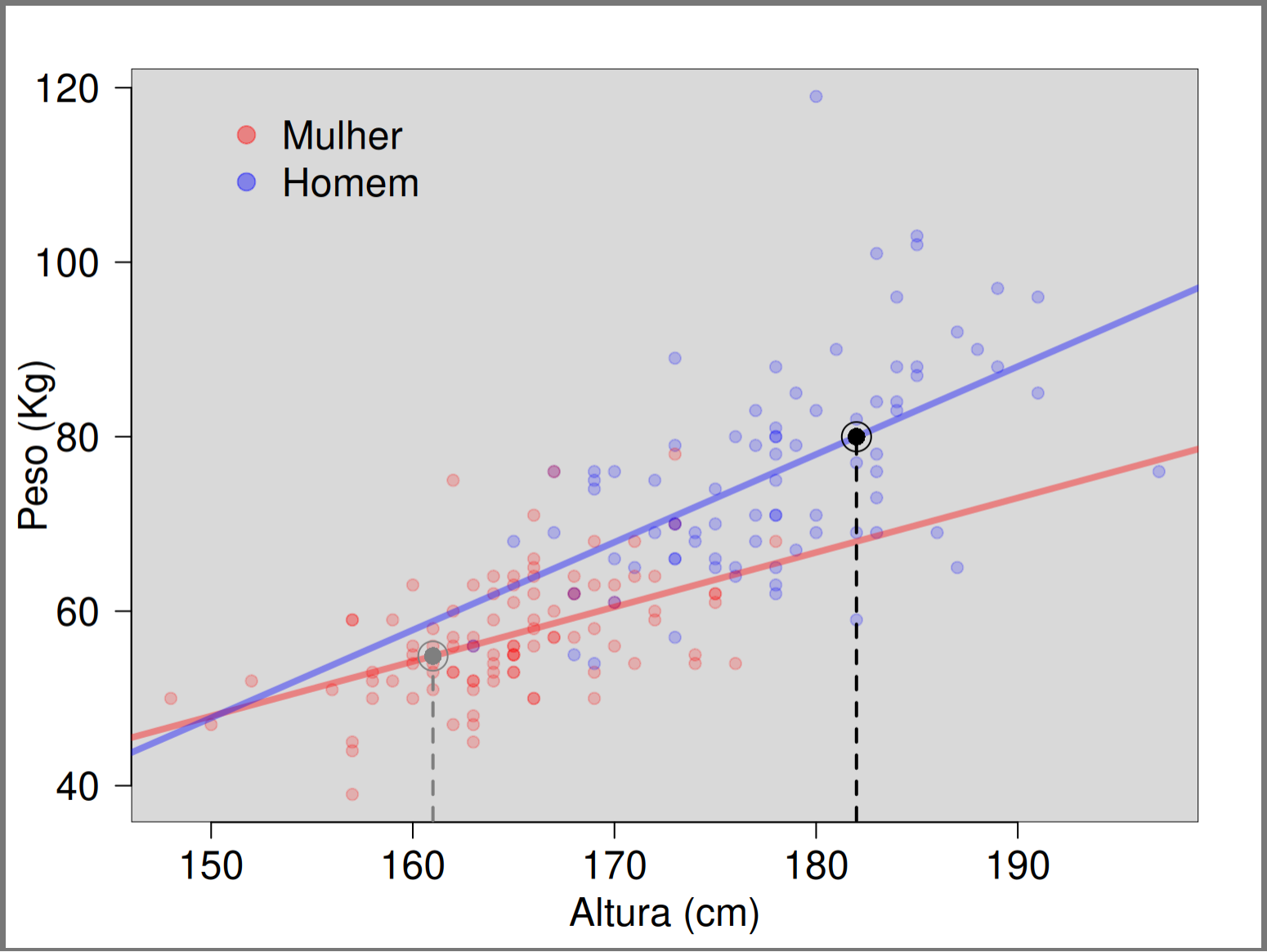
$$w = \hat{\alpha} + \hat{\beta}_s + (\hat{\beta}_h + \hat{\beta}_{s:h}) * height$$

```
-45.8 -57.4 + 0.625 * 182 + 0.381 * 182
```

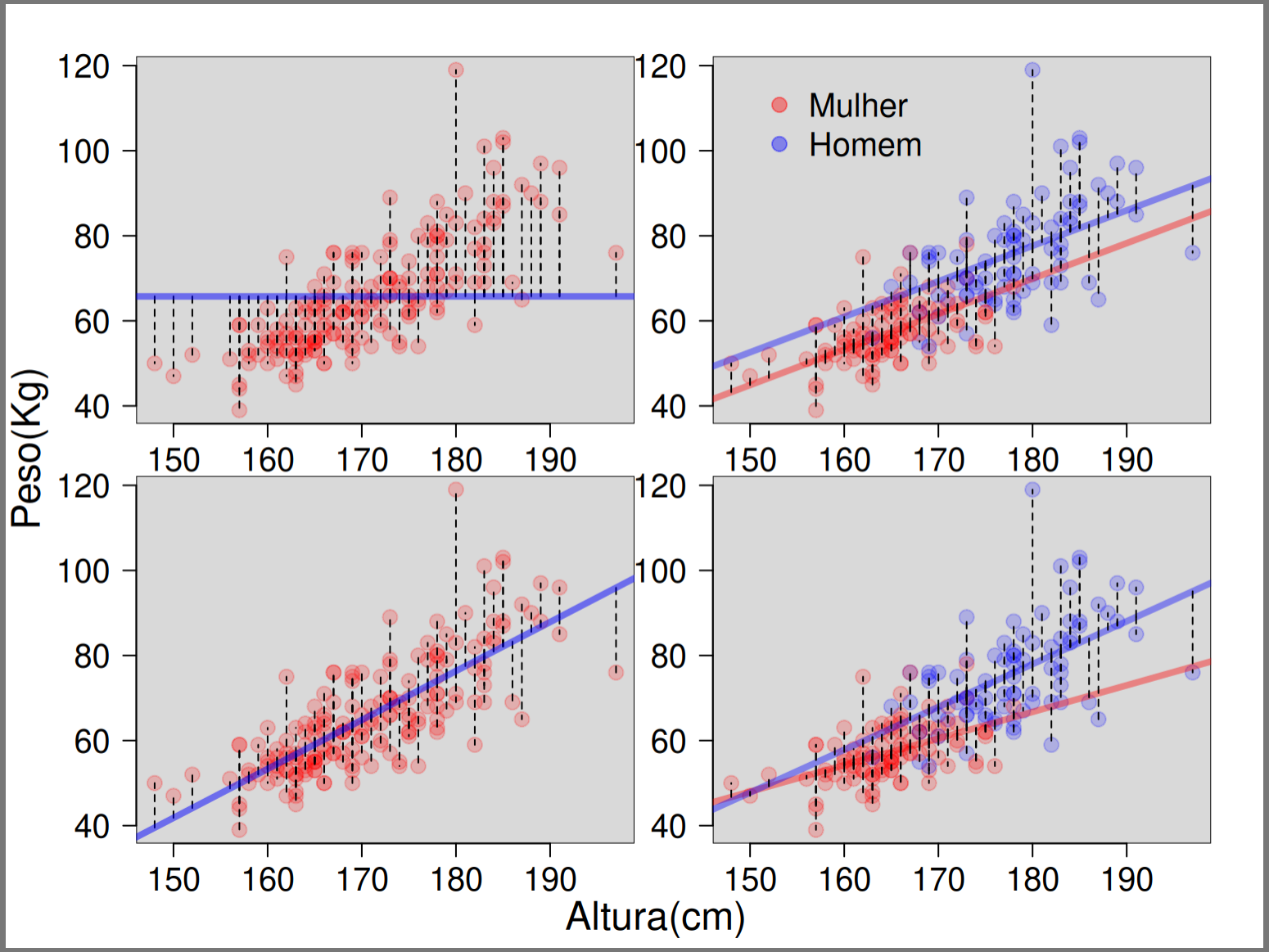
```
[1] 79.892
```

Predito pelo modelo

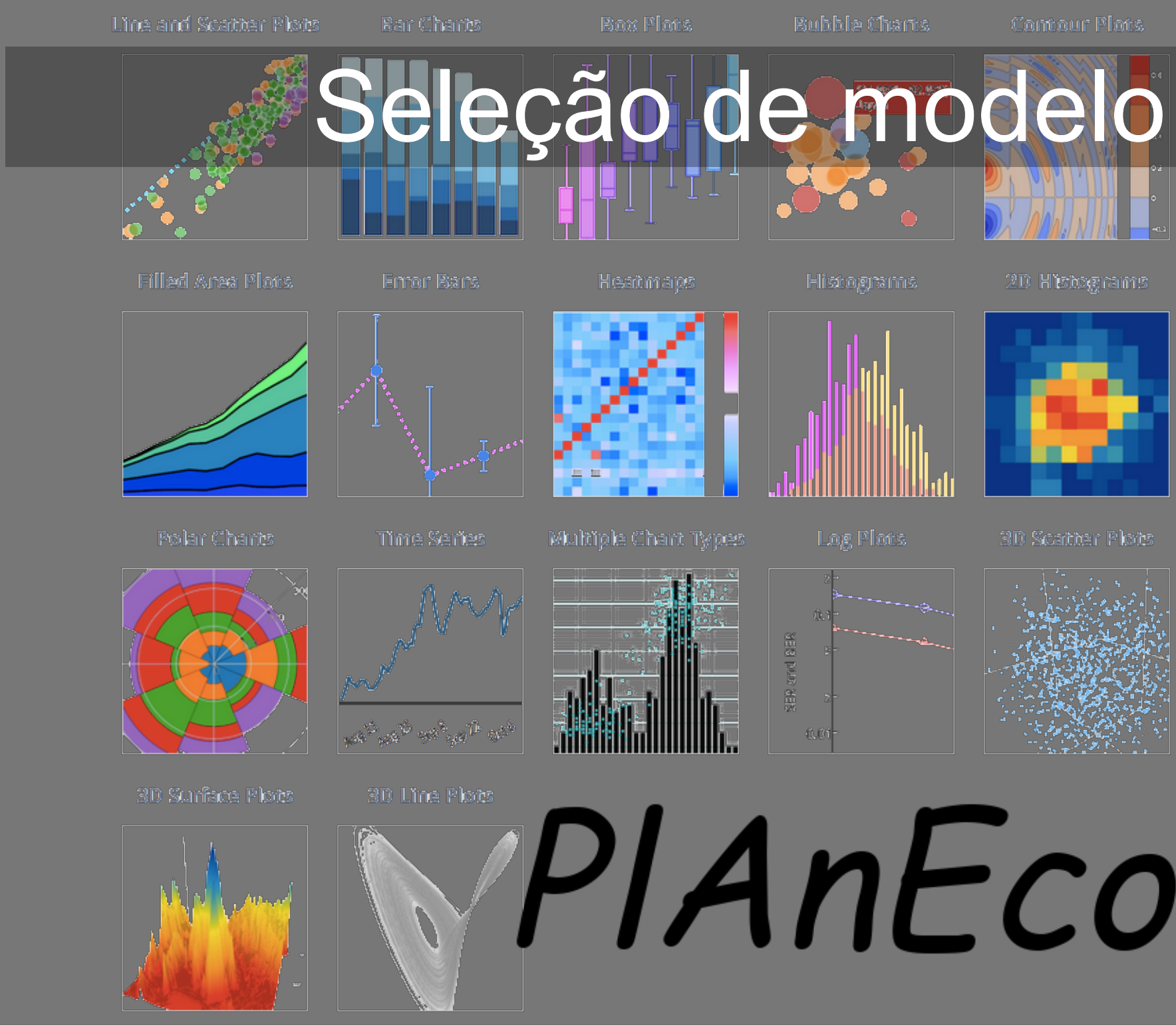
- Um homem com 182cm de altura tem peso 79.99 kg.



Qual o melhor modelo?

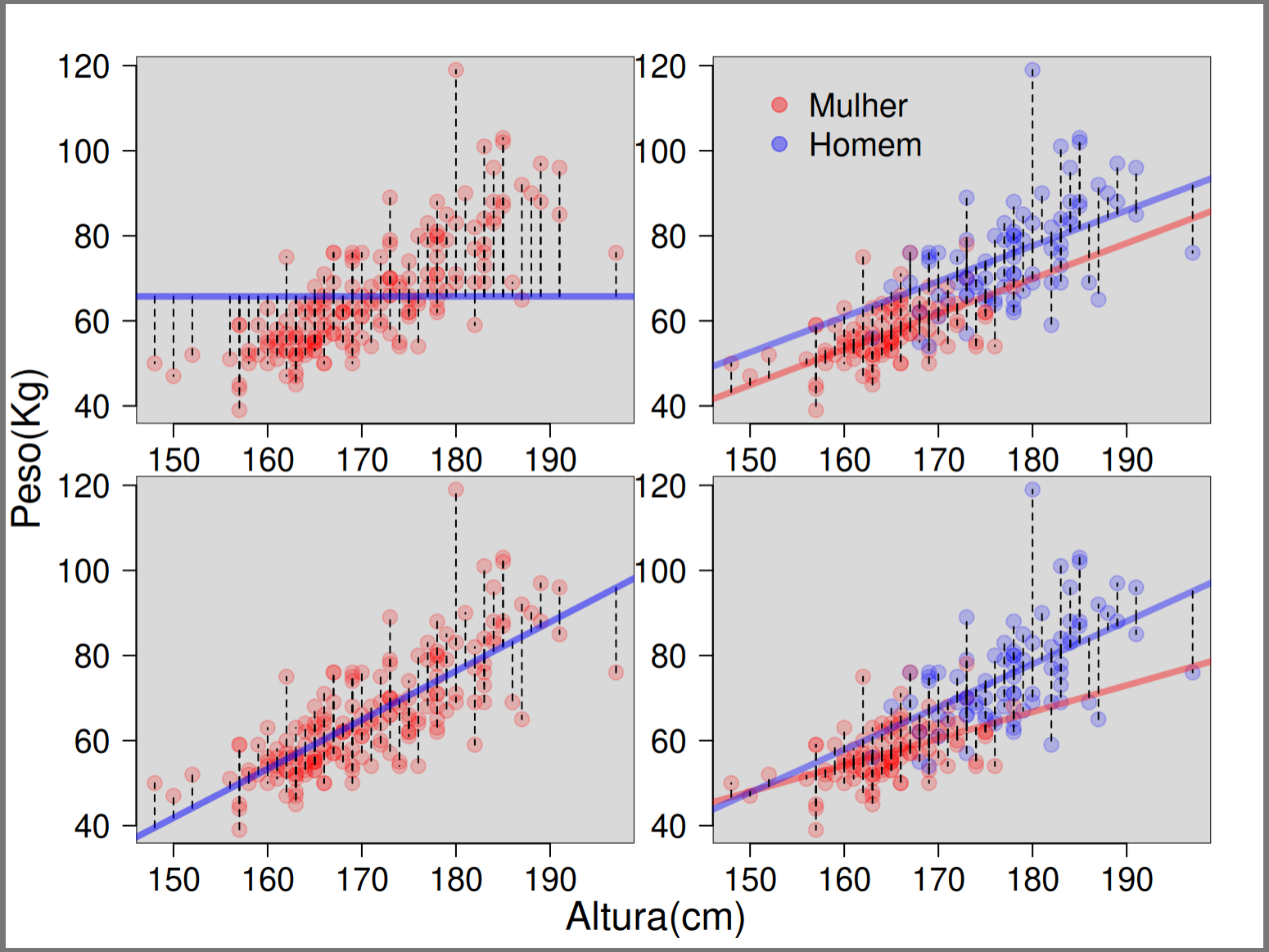


Seleção de modelo



PlAnEco

Qual o melhor modelo?



Tipo de Seleção

Teste de hipótese

Modelos aninhados: o mais simples está contido no mais complexo.

ANOVA (Resíduos)

Razão da Variância (F)

Deviance (Generalização)

Distância ao modelo saturado

$$D_m = 2 * (LL_s - LL_m)$$

Razão de verossimilhança

$$D = 2 * (LL_{ratio})$$

Aproximação ao X^2

Princípio da parcimônia (Navalha de Occam)

- mínimo número de parâmetros
- linear é melhor que não-linear
- menos pressupostos
- mínimo adequado
- explicações mais simples são preferíveis

Método do modelo cheio ao mínimo adequado

1. ajuste o modelo máximo (cheio)
2. simplifique o modelo:
    - inspecione os coeficientes (summary)
    - remova termos não significativos um de cada vez
3. ordem de remoção de termos:
    - interação não significativos (maior ordem)
    - termos quadráticos ou não lineares
    - variáveis explicativas não significativas
    - agrupar níveis de fatores sem diferença
    - sempre confira a ordem de retirada de termos de mesmo nível

Simplificação do modelo:

Critério para a tomada de decisão (Variância/Deviance)

Compare os modelos

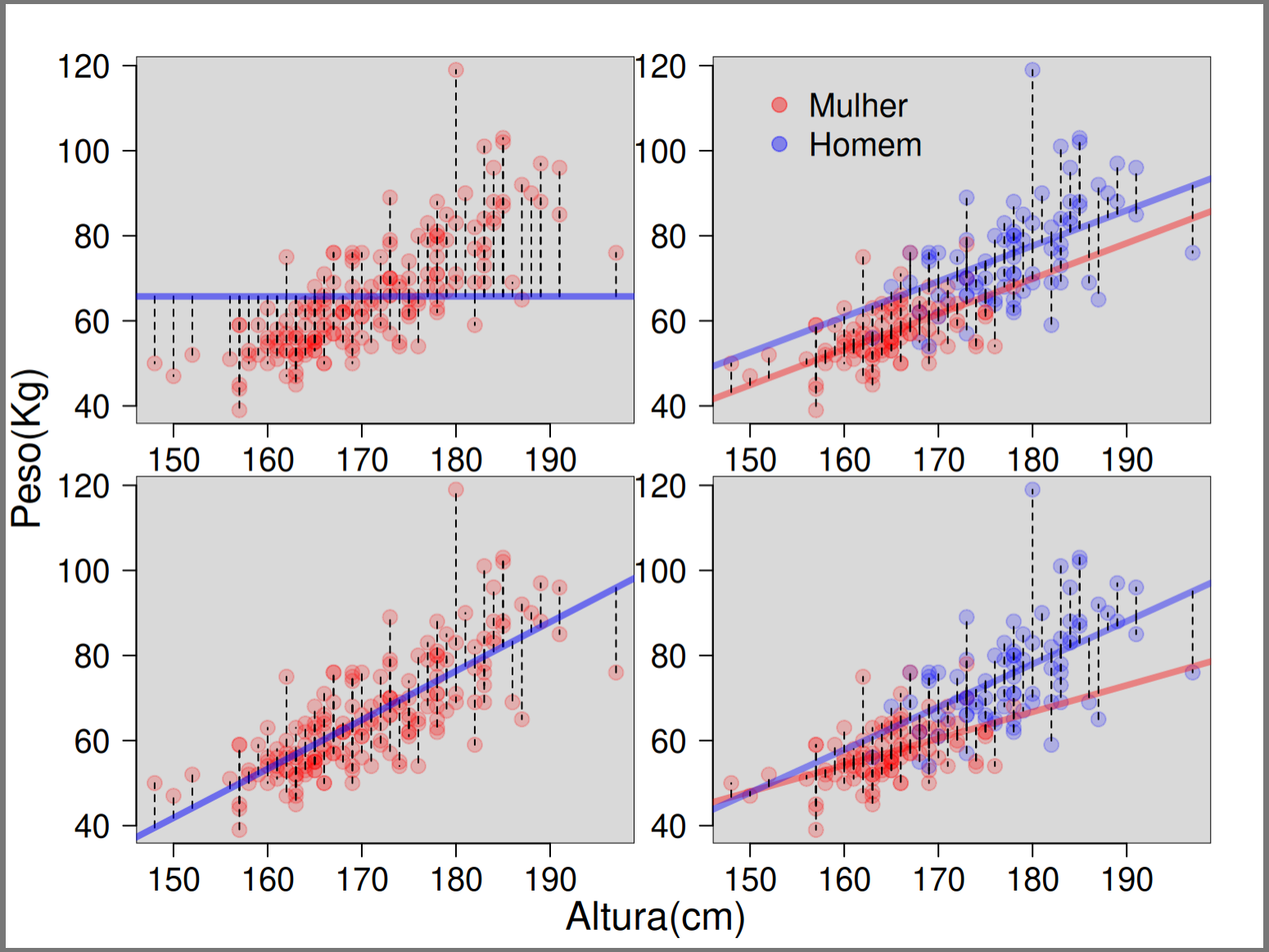
A diferença não é significativa

```
* retenha o modelo mais simples
* continue simplificando
```

A difereça é significativa

```
* retenha o modelo complexo
* confira a ordem de retirada dos termos
* modelo MINÍMO ADEQUADO
```

Simplificando Modelo



Modelo cheio

```
Call:
lm(formula = weight ~ height + sex + sex:height, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.990  -4.548  -0.926   4.821  41.023 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -45.7988    24.8453  -1.843   0.0670 .  
height        0.6252     0.1507   4.148 5.22e-05 ***
sexM        -57.4326    34.8293  -1.649   0.1009    
height:sexM   0.3815     0.2037   1.873   0.0628 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.2 on 176 degrees of freedom
Multiple R-squared:  0.6344,    Adjusted R-squared:  0.6282 
F-statistic: 101.8 on 3 and 176 DF,  p-value: < 2.2e-16
```

Simplificando Modelo

weight ~ height + sex + sex:height

weight ~ height + sex

```
Analysis of Variance Table

Model 1: weight ~ height + sex
Model 2: weight ~ height + sex + sex:height
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    177 12069
2    176 11833  1    235.82 3.5075 0.06275 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simplificando Modelo

weight ~ height + sex

weight ~ height

```
Analysis of Variance Table

Model 1: weight ~ height
Model 2: weight ~ height + sex
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    178 13273
2    177 12069  1    1203.5 17.65 4.204e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simplificando Modelo

weight ~ height + sex

weight ~ sex

```
Analysis of Variance Table

Model 1: weight ~ sex
Model 2: weight ~ height + sex
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    178 16619
2    177 12069  1    4550.1 66.728 5.713e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo Mínimo Adequado

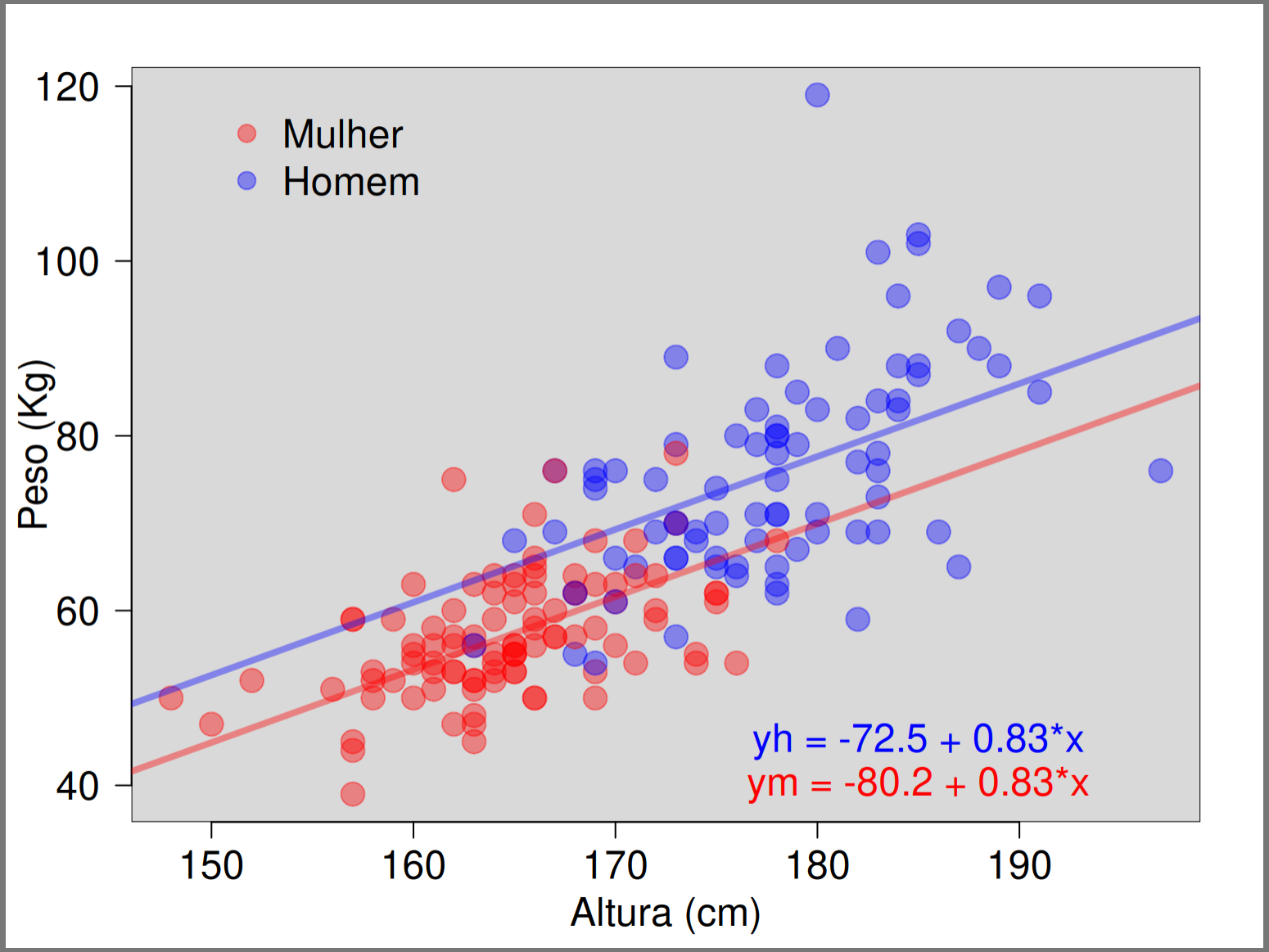
```
Call:
lm(formula = weight ~ height + sex, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.302  -4.808  -0.335   5.239  41.366 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -80.2107    16.8415  -4.763 3.96e-06 ***
height        0.8341     0.1021   8.169 5.71e-14 ***
sexM          7.7070     1.8345   4.201 4.20e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.258 on 177 degrees of freedom
Multiple R-squared:  0.6271,    Adjusted R-squared:  0.6229 
F-statistic: 148.8 on 2 and 177 DF,  p-value: < 2.2e-16
```

Modelo Mínimo Adequado

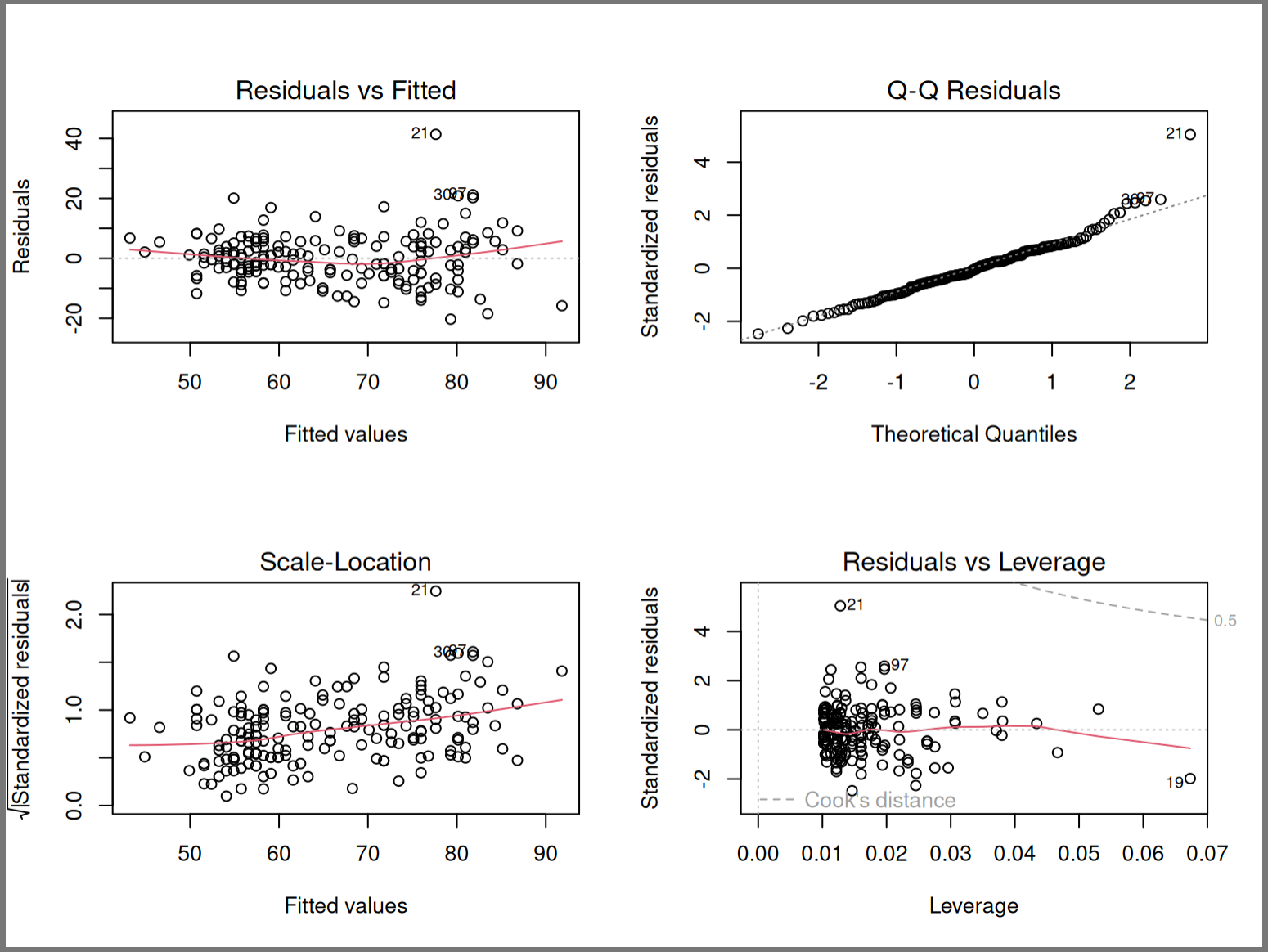
```
(Intercept)      height        sexM
-80.2107328   0.8340964   7.7070166
```

```
                2.5 %     97.5 %
(Intercept) -113.44661 -46.974852
height         0.63259   1.035603
sexM           4.08671  11.327323
```

Modelo Mínimo Adequado



Diagnóstico do Modelo:



Outros tipos de Seleção

Teoria da Informação (AIC)

Baseado no cálculo da verossimilhança, proporcional à probabilidade da realização dos dados e penalizado pelo número de parâmetros.

Distância de Kullback-Leibler

Distância ao modelo verdadeiro

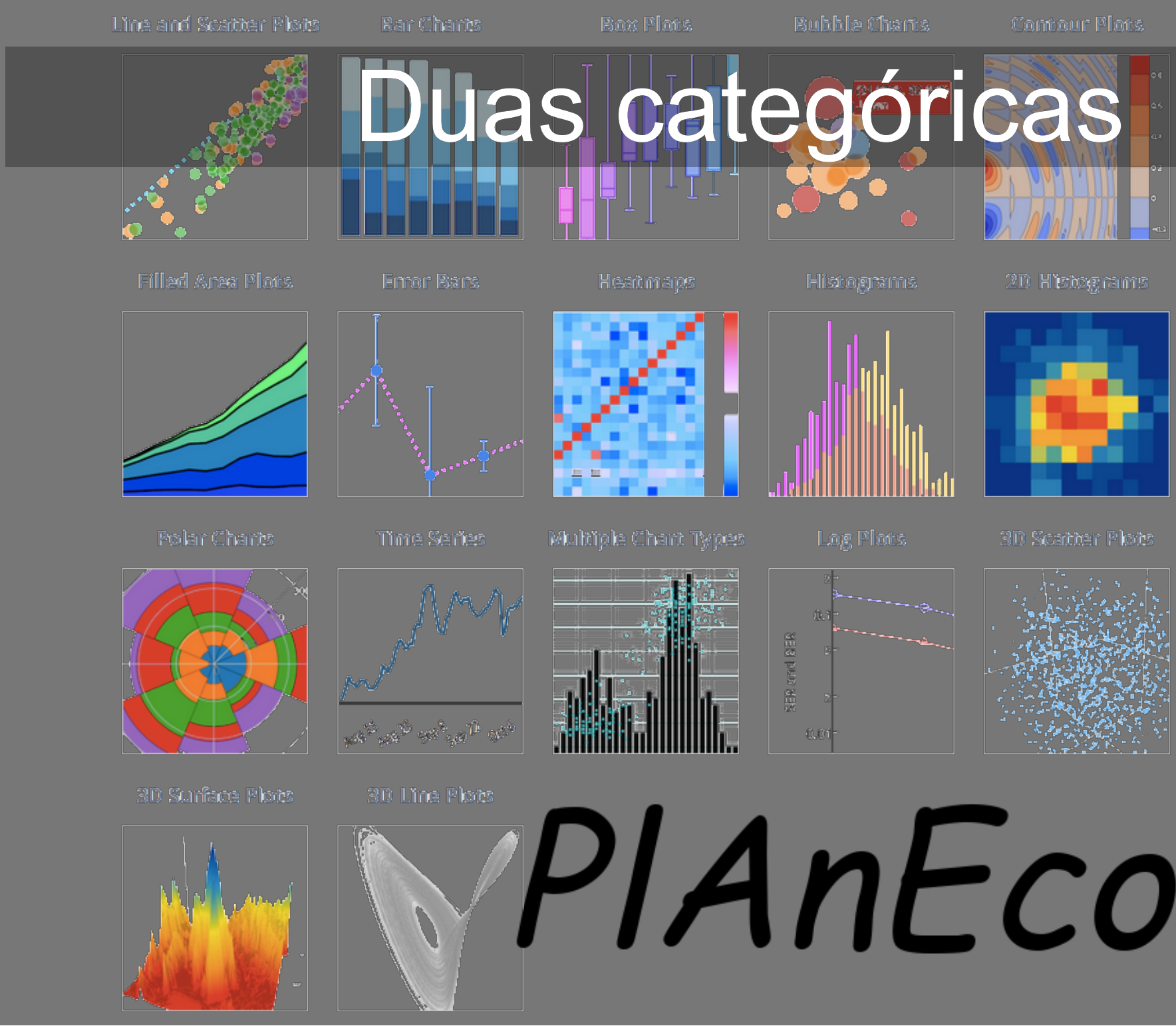
$$AIC = -2LL + 2k$$

Inferência Bayesiana (Teorema Bayes)

Atualização da probabilidade posteriori, baseado em uma probabilidade priori

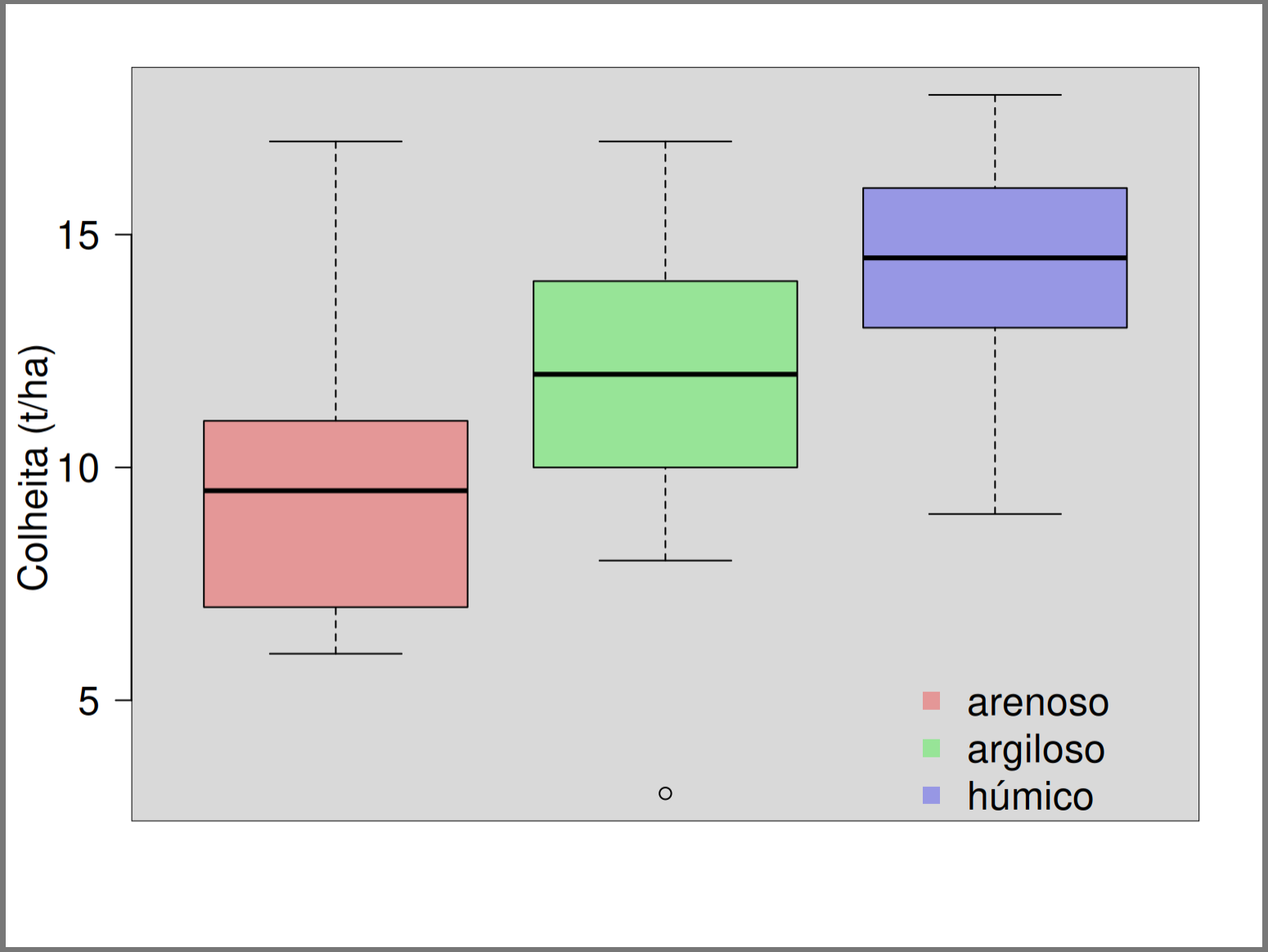
$$P(H|dados) \sim L(dados|\theta) * P(prior)$$

Duas categóricas



PlAnEco

Retomando solos



Experimento plausível

Acrescentar mais um fator

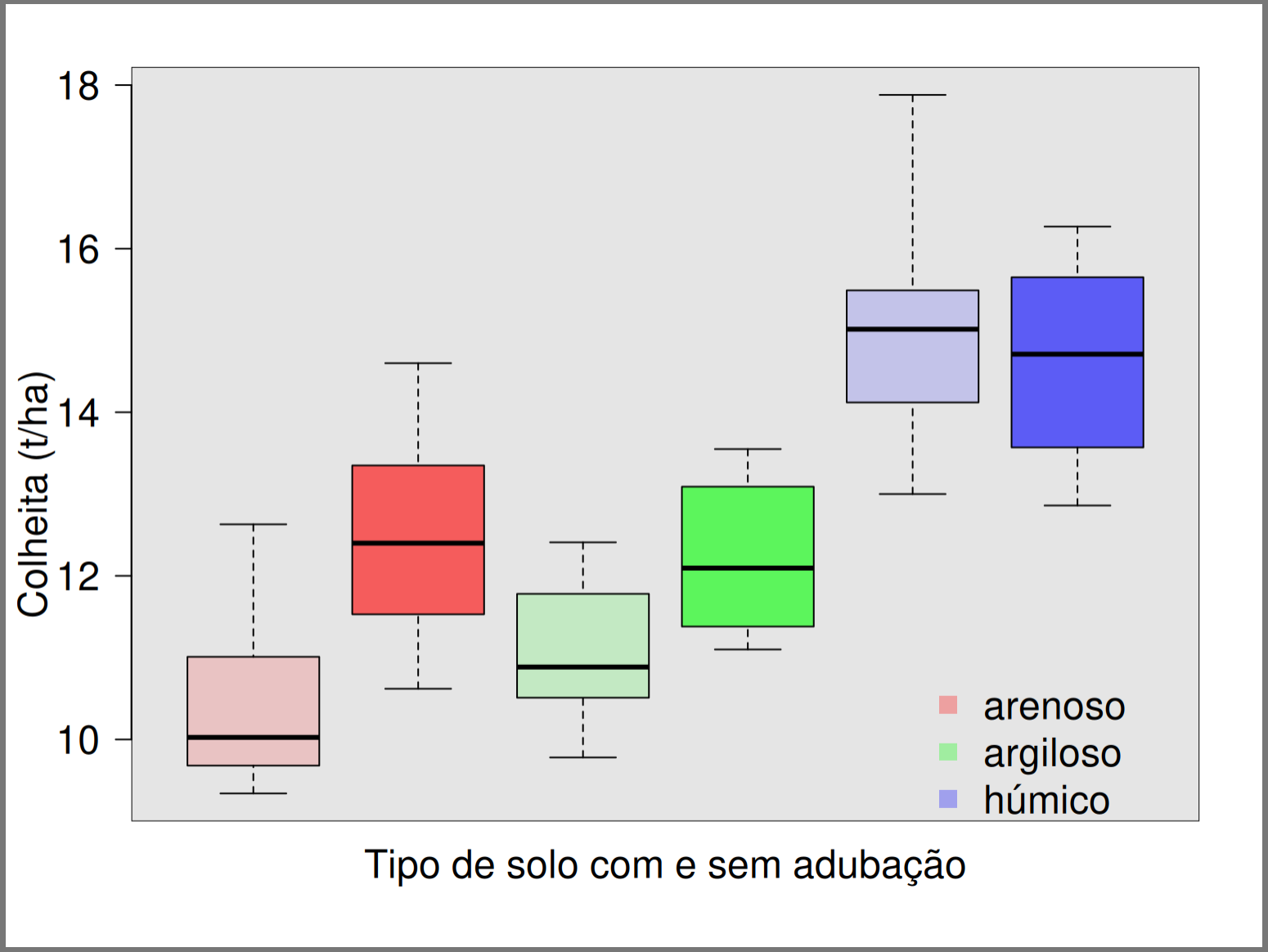
Adubação orgânica: categórica de dois níveis

```
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-10 ***
soloarg        1.600      1.529   1.047  0.30456
solohum        4.400      1.529   2.878  0.00773 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-squared:  0.2392,    Adjusted R-squared:  0.1829
F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.02495
```

Experimento plausível

solo	adubo	colhe
are	nao	9.34
are	nao	12.63
are	nao	11.94
are	nao	9.83
are	sim	11.56
are	sim	12.49
are	sim	11.39
are	sim	13.35
arg	nao	10.87
arg	nao	11.50
arg	nao	12.11
arg	nao	11.78

Solos e adubo



Modelo cheio

```
Call:
lm(formula = colhe ~ solo + adubo + solo:adubo, data = cropsim)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1060 -0.7712 -0.1970  0.8015  2.7740 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       10.4270     0.3485  29.923  < 2e-16 ***
soloarg            0.6700     0.4928   1.360 0.179609    
solohum            4.6790     0.4928   9.495 4.15e-13 ***
adubosim           2.0230     0.4928   4.105 0.000138 ***
soloarg:adubosim  -0.9260     0.6969  -1.329 0.189532    
solohum:adubosim  -2.4810     0.6969  -3.560 0.000783 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.102 on 54 degrees of freedom
Multiple R-squared:  0.7289,    Adjusted R-squared:  0.7037 
F-statistic: 29.03 on 5 and 54 DF,  p-value: 3.563e-14
```

Modelo sem interação

```
Analysis of Variance Table

Model 1: colhe ~ solo + adubo
Model 2: colhe ~ solo + adubo + solo:adubo
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     56 81.287
2     54 65.569  2    15.718 6.4724 0.003022 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo Mínimo adequado

```
Call:
lm(formula = colhe ~ solo + adubo + solo:adubo, data = cropsim)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1060 -0.7712 -0.1970  0.8015  2.7740 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       10.4270     0.3485  29.923  < 2e-16 ***
soloarg            0.6700     0.4928   1.360 0.179609    
solohum            4.6790     0.4928   9.495 4.15e-13 ***
adubosim           2.0230     0.4928   4.105 0.000138 ***
soloarg:adubosim  -0.9260     0.6969  -1.329 0.189532    
solohum:adubosim  -2.4810     0.6969  -3.560 0.000783 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.102 on 54 degrees of freedom
Multiple R-squared:  0.7289,    Adjusted R-squared:  0.7037 
F-statistic: 29.03 on 5 and 54 DF,  p-value: 3.563e-14
```

Predito pelo modelo

(Intercept)	soloarg	solohum	adubosim	soloarg:adubosim	solohum:adubosim
10.43	0.67	4.68	2.02	-0.93	-2.48

Solo arenoso

- adubo não: 10.43
- adubo sim: 10.43 + 2.02 = 12.45

Solo argiloso

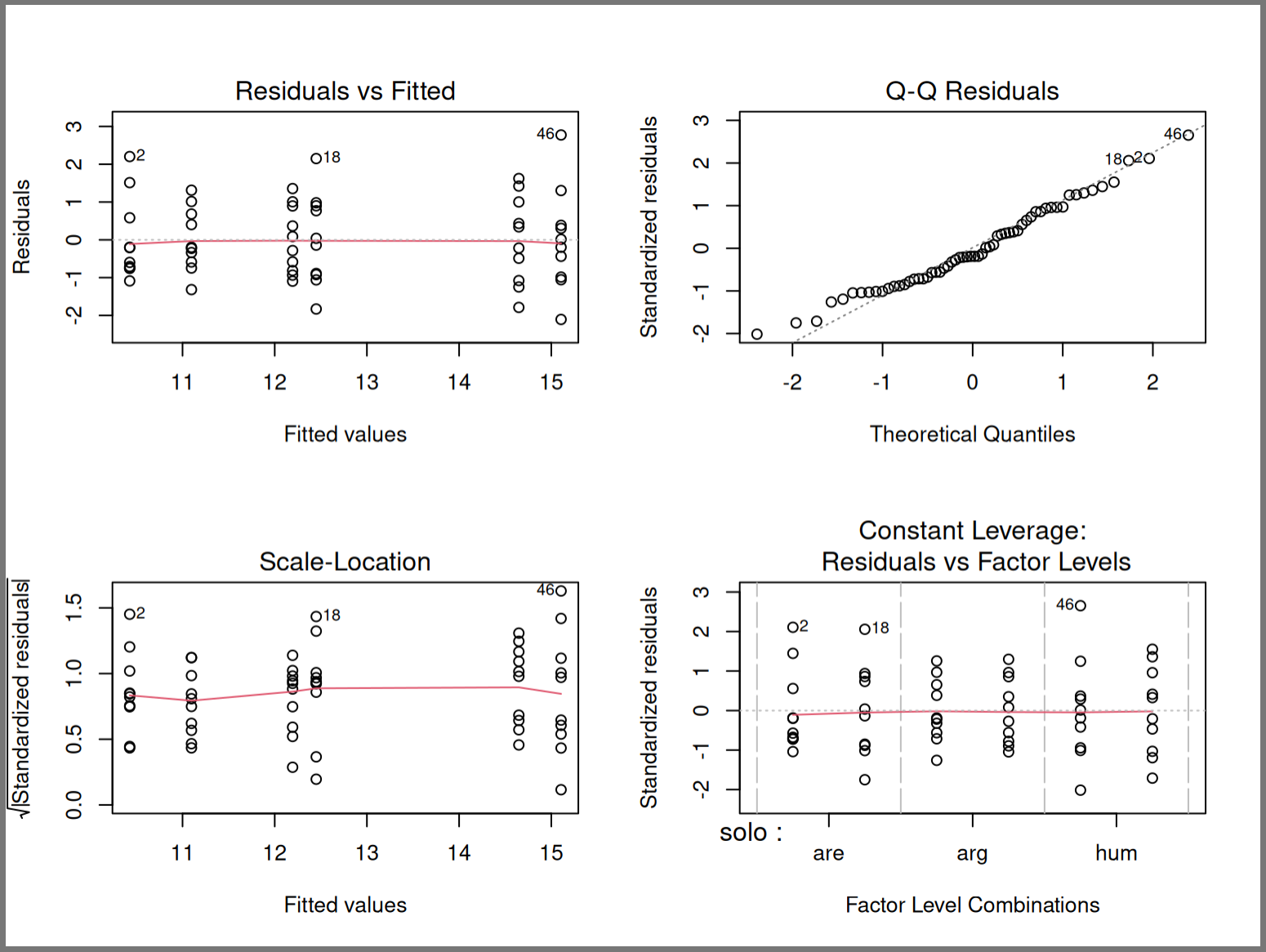
- adubo não: 10.43 + 0.67 = 11.10
- adubo sim: 10.43 + 0.67 + 2.02 - 0.93 = 12.19

Solo húmico

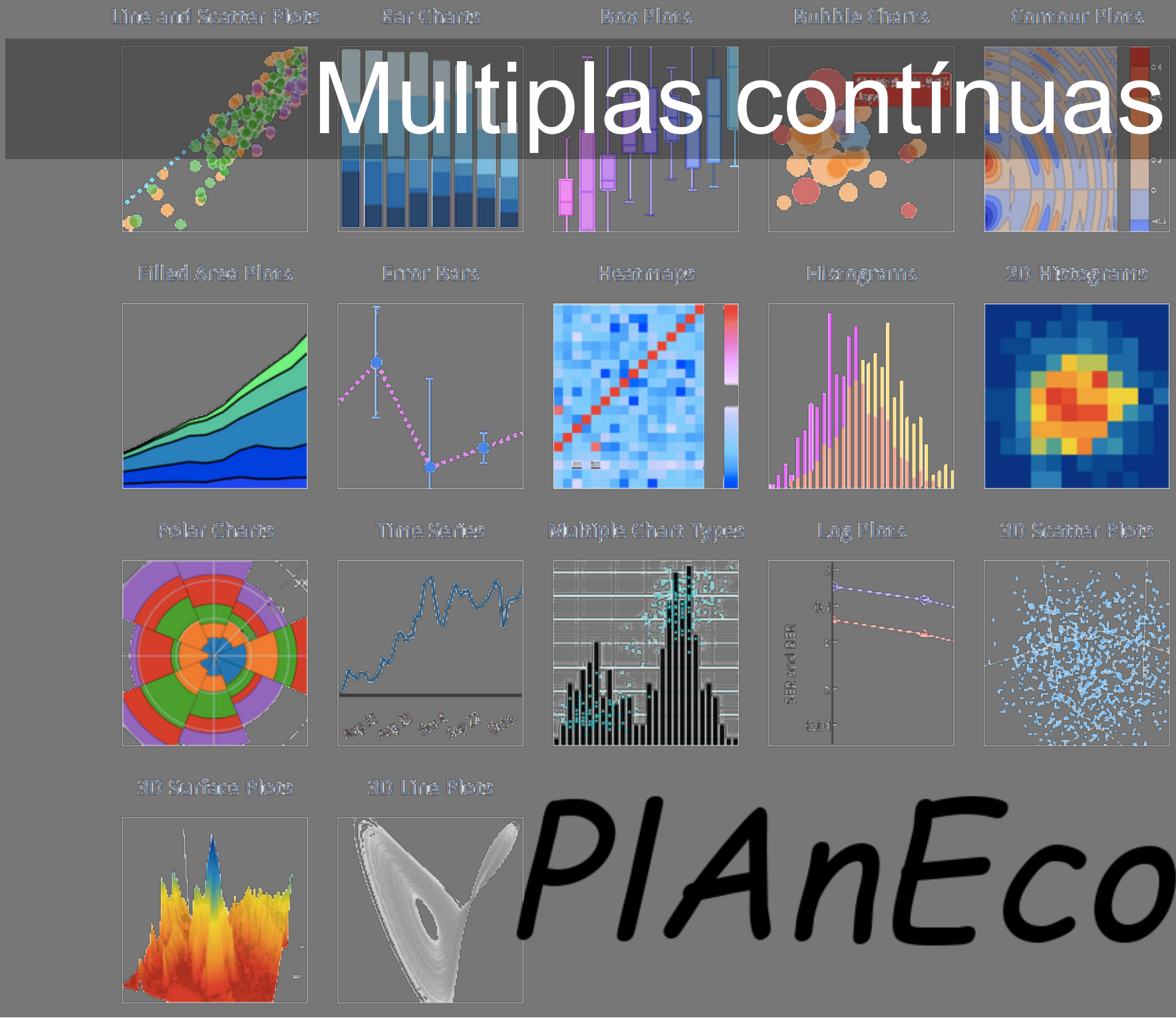
- adubo não: 10.43 + 4.67 = 15.10
- adubo sim: 10.43 + 4.67 + 2.02 - 2.48 = 14.64

Diagnóstico do modelo

```
oldpar = par(mfrow= c(2,2))
plot(lmadubo)
```


Residuals vs Fitted
Residuals
Fitted values
2
18
46
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
solo :
are
arg
hum
Factor Level Combinations

Multiplas contínuas



PlAnEco

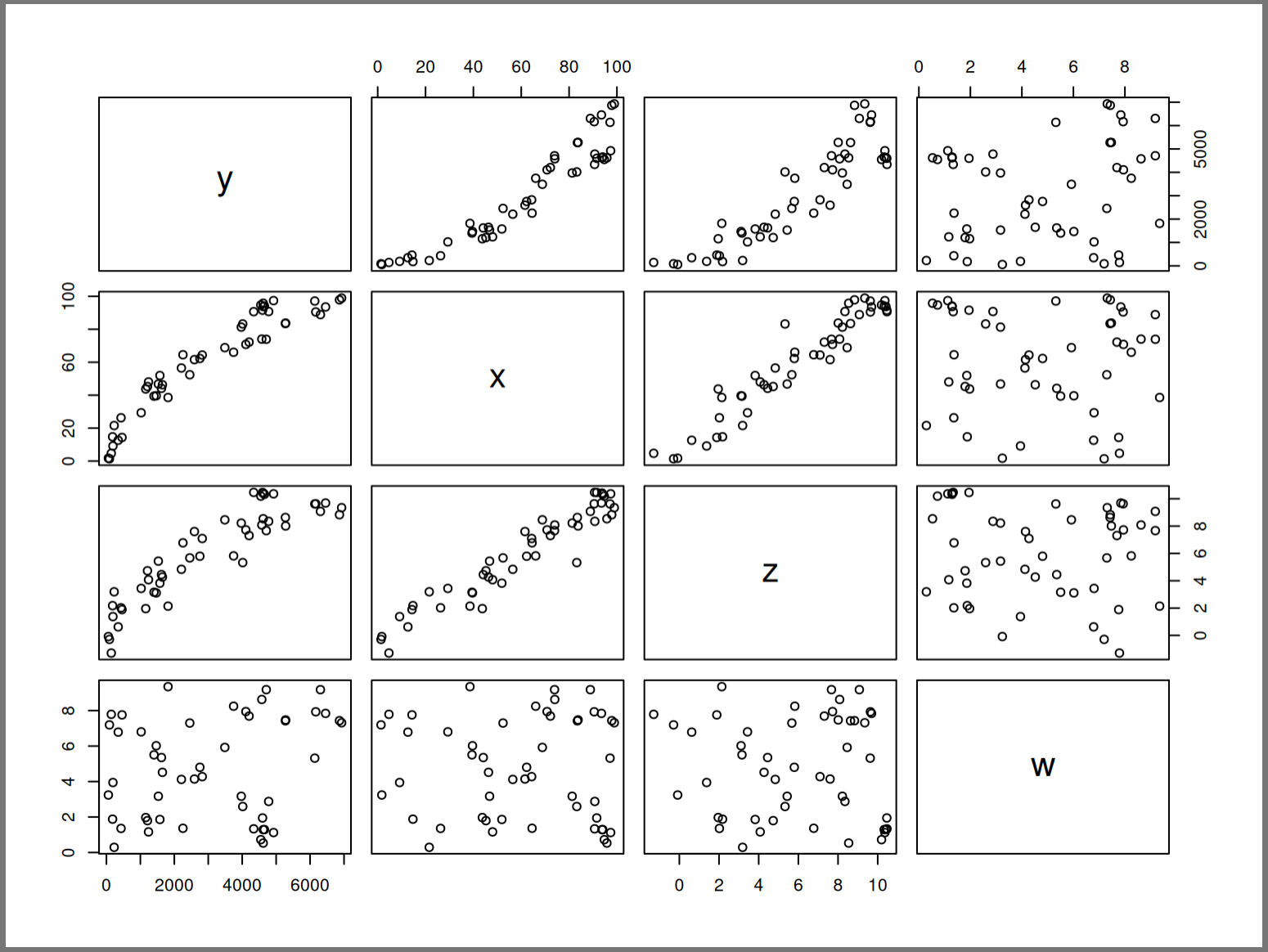
Simulando dados

$$y \sim x + z + w \ldots$$

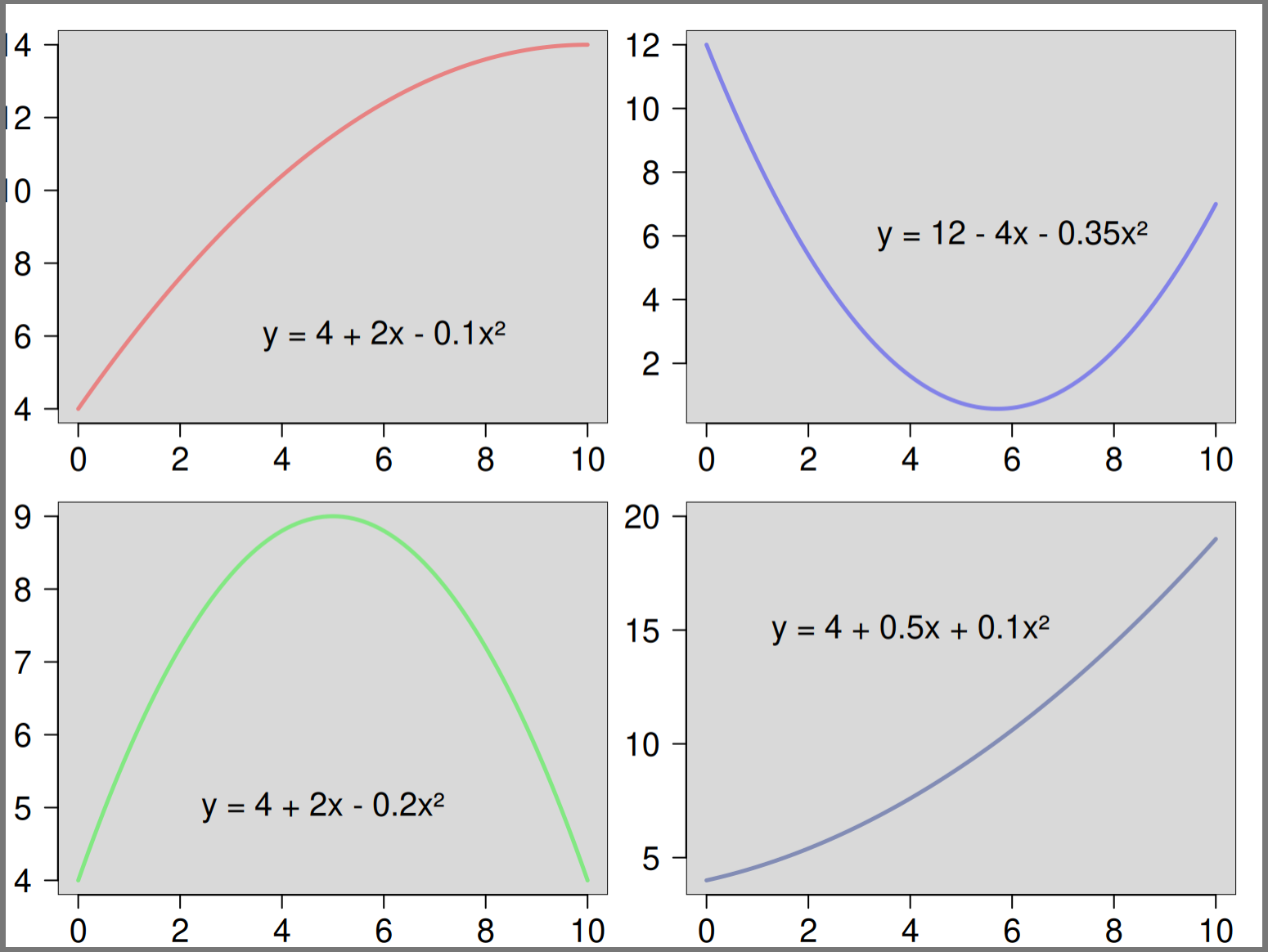
Quais variáveis estão relacionadas à resposta?

y	x	z	w
88.98744	1.390885	-0.2913806	7.1937859
56.79278	1.726080	-0.0846613	3.2408595
143.14914	4.705672	-1.2925959	7.7880950
192.00850	9.161318	1.3762292	3.9444100
347.20778	12.631249	0.6231300	6.7859287
459.51276	14.331993	1.8886494	7.7582504
181.53773	14.732307	2.1780680	1.8786904
228.08316	21.558238	3.1909274	0.2908582
431.00516	26.287454	2.0198190	1.3571380
1023.92200	29.327814	3.4377365	6.8016418

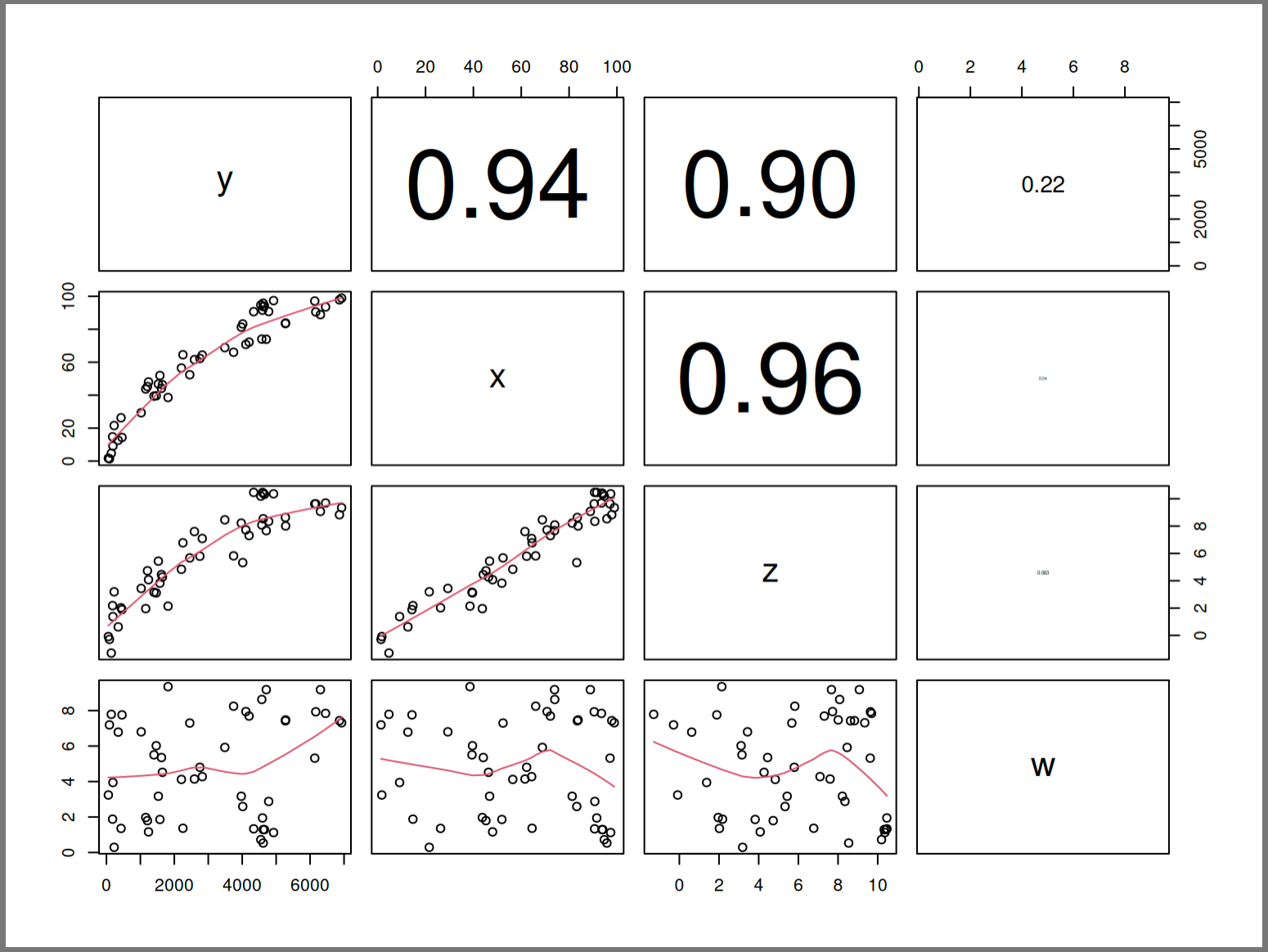
Análise Exploratória



Polinômios



Correlação entre preditoras



Colinearidade: soluções

- reter apenas uma das variáveis colineares
- reduzir as dimensões das variáveis colineares (PCA)
- reter mas ficar atento:
  - estimativas dos coeficientes
  - ordem dos termos do modelo

Definir os termos do modelo cheio

- x
- z
- w
- x^2
- z^2
- $x : z$
- $x : w$
- $z : w$
- $x : z : w$

Modelo Cheio

```
Call:
lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + z:w + 
    z:w:x, data = yxzw)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.274  -8.693  -0.498   6.395  44.494 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.7908718 17.1643176   1.503 0.140800    
x           -3.0593264  0.8043883  -3.803 0.000478 ***
w            4.0019121  2.9446161   1.359 0.181741    
z            7.6029451  8.8429234   0.860 0.395036    
I(x^2)       0.5281215  0.0187834  28.116  < 2e-16 ***
I(z^2)       0.7742739  1.6037353   0.483 0.631874    
x:w          2.9661379  0.1055799  28.094  < 2e-16 ***
x:z         -0.2990845  0.3374012  -0.886 0.380681    
w:z          0.4198608  1.0893709   0.385 0.701971    
x:w:z       -0.0009584  0.0101688  -0.094 0.925378    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.89 on 40 degrees of freedom
```

Simplificando o modelo

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + z:w + z:w:x
Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + w:z
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     40 7720.1
2     41 7721.8 -1   -1.7146 0.0089 0.9254
```

Avalia o modelo retido

```
Call:
lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + w:z,
    data = yxzw)

Residuals:
    Min      1Q  Median      3Q     Max
-25.323  -8.808  -0.614   6.423  44.259

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.59758   11.44903   2.148 0.037635 *
x           -3.04368    0.77750  -3.915 0.000334 ***
w            4.21361    1.88107   2.240 0.030574 *
z            8.02429    7.53696   1.065 0.293260
I(x^2)       0.52810    0.01855  28.464  < 2e-16 ***
I(z^2)       0.76719    1.58249   0.485 0.630400
x:w          2.96159    0.09279  31.917  < 2e-16 ***
x:z         -0.30308    0.33065  -0.917 0.364706
w:z          0.35674    0.84875   0.420 0.676454
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 41 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
```

Retirando w:z

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + w:z
Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 7721.8
2     42 7755.0 -1   -33.271 0.1767 0.6765
```

Retirando x:z

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + w:z
Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:w + w:z
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 7721.8
2     42 7880.0 -1   -158.24 0.8402 0.3647
```

Retirando x:w

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + w:z
Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:z + z:w
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     41   7722
2     42 199581 -1   -191859 1018.7 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sem w:z

```
Call:
lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z, data = y

Residuals:
    Min      1Q  Median      3Q     Max 
-24.371  -9.431  -0.188   6.552  43.777 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.44091   11.33024   2.157   0.0368 *  
x           -3.22912    0.63389  -5.094 7.84e-06 ***
w            4.09867    1.84275   2.224   0.0316 *  
z           10.06374    5.71047   1.762   0.0853 .  
I(x^2)       0.52961    0.01802  29.383  < 2e-16 ***
I(z^2)       0.84006    1.55747   0.539   0.5925    
x:w          2.99897    0.02625 114.240  < 2e-16 ***
x:z         -0.32768    0.32223  -1.017   0.3150    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.59 on 42 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 1.658e+05 on 7 and 42 DF,  p-value: < 2.2e-16
```

Simplificando

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z
Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:w
  Res.Df  RSS Df Sum of Sq      F Pr(>F)
1     42 7755
2     43 7946 -1   -190.94 1.0341  0.315
```

Avalia o modelo

```
Call:
lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w, data = yxzw)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.506  -9.553  -0.384   7.003  44.330 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.076039  11.317500   2.216   0.0321 *  
x           -2.911350   0.551731  -5.277 4.07e-06 ***
w            3.966437   1.838890   2.157   0.0366 *  
z            7.359415   5.055452   1.456   0.1527    
I(x^2)       0.511880   0.004588 111.564  < 2e-16 ***
I(z^2)      -0.692429   0.393440  -1.760   0.0855 .  
x:w          2.997727   0.026233 114.271  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.59 on 43 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 1.933e+05 on 6 and 43 DF,  p-value: < 2.2e-16
```

Retirando z^2

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w
Model 2: y ~ x + w + z + I(x^2) + x:w
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     43 7946.0
2     44 8518.4 -1   -572.37 3.0974 0.08553 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avalia o modelo retido

```
Call:
lm(formula = y ~ x + w + z + I(x^2) + x:w, data = yxzw)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.634  -8.562  -0.765   6.866  46.135 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.296874  11.321016   2.588    0.013 *  
x           -2.188280   0.376927  -5.806 6.51e-07 ***
w            2.692380   1.730220   1.556    0.127    
z           -0.773120   2.098829  -0.368    0.714    
I(x^2)       0.505046   0.002502 201.863  < 2e-16 ***
x:w          3.017220   0.024341 123.958  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.91 on 44 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 2.214e+05 on 5 and 44 DF,  p-value: < 2.2e-16
```

Retirando z

```
Analysis of Variance Table

Model 1: y ~ x + w + z + I(x^2) + x:w
Model 2: y ~ x + w + I(x^2) + x:w
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 8518.4
2     45 8544.6 -1   -26.269 0.1357 0.7144
```

Avalia o modelo retido

```
Call:
lm(formula = y ~ x + w + I(x^2) + x:w, data = yxzw)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.914  -8.309  -1.179   7.016  45.459 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.155901  11.205360   2.602   0.0125 *  
x           -2.260714   0.318466  -7.099 7.27e-09 ***
w            2.765990   1.702057   1.625   0.1111    
I(x^2)       0.504995   0.002474 204.121  < 2e-16 ***
x:w          3.016414   0.024008 125.641  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.78 on 45 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 2.821e+05 on 4 and 45 DF,  p-value: < 2.2e-16
```

Simplificando?

```
Analysis of Variance Table

Model 1: y ~ x + w + I(x^2) + x:w
Model 2: y ~ x + w + I(x^2)
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1     45    8545
2     46 3005951 -1  -2997406 15786 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

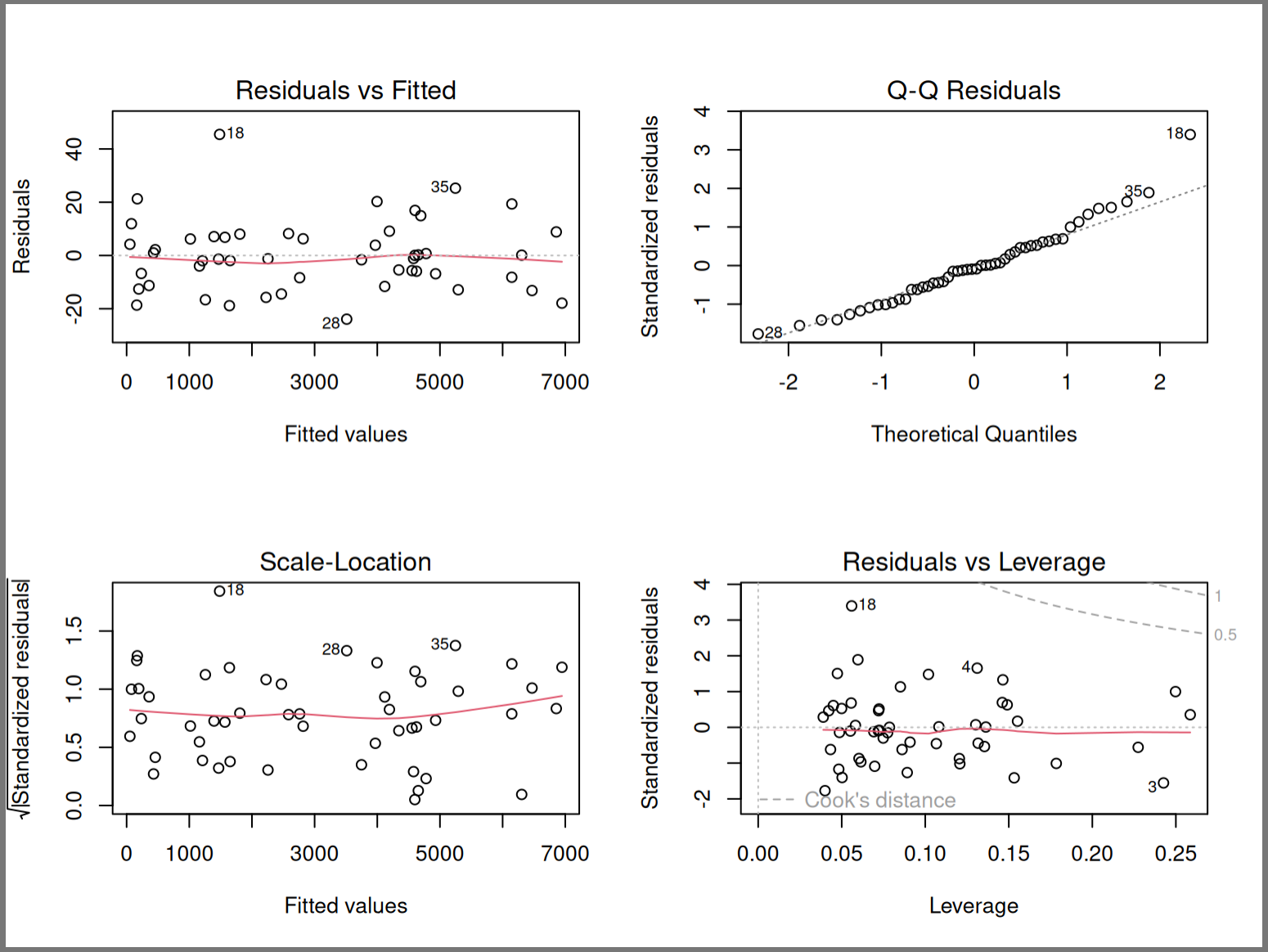
```
Analysis of Variance Table

Model 1: y ~ x + w + I(x^2) + x:w
Model 2: y ~ x + w + x:w
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1     45    8545
2     46 7920005 -1  -7911461 41665 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo Mínimo Adequado

$$y \sim x + w + x^2 + x : w$$

```
Call:
lm(formula = y ~ x + w + I(x^2) + x:w, data = yxzw)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.914  -8.309  -1.179   7.016  45.459 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.155901  11.205360   2.602   0.0125 *  
x           -2.260714   0.318466  -7.099 7.27e-09 ***
w            2.765990   1.702057   1.625   0.1111    
I(x^2)       0.504995   0.002474 204.121  < 2e-16 ***
x:w          3.016414   0.024008 125.641  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.78 on 45 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 2.821e+05 on 4 and 45 DF,  p-value: < 2.2e-16
```

DIAGNÓSTICO DO MODELO



Estimativa do Modelo

(Intercept)	x	w	I(x^2)	x:w
29.1559	-2.260714	2.76599	0.5049954	3.016414

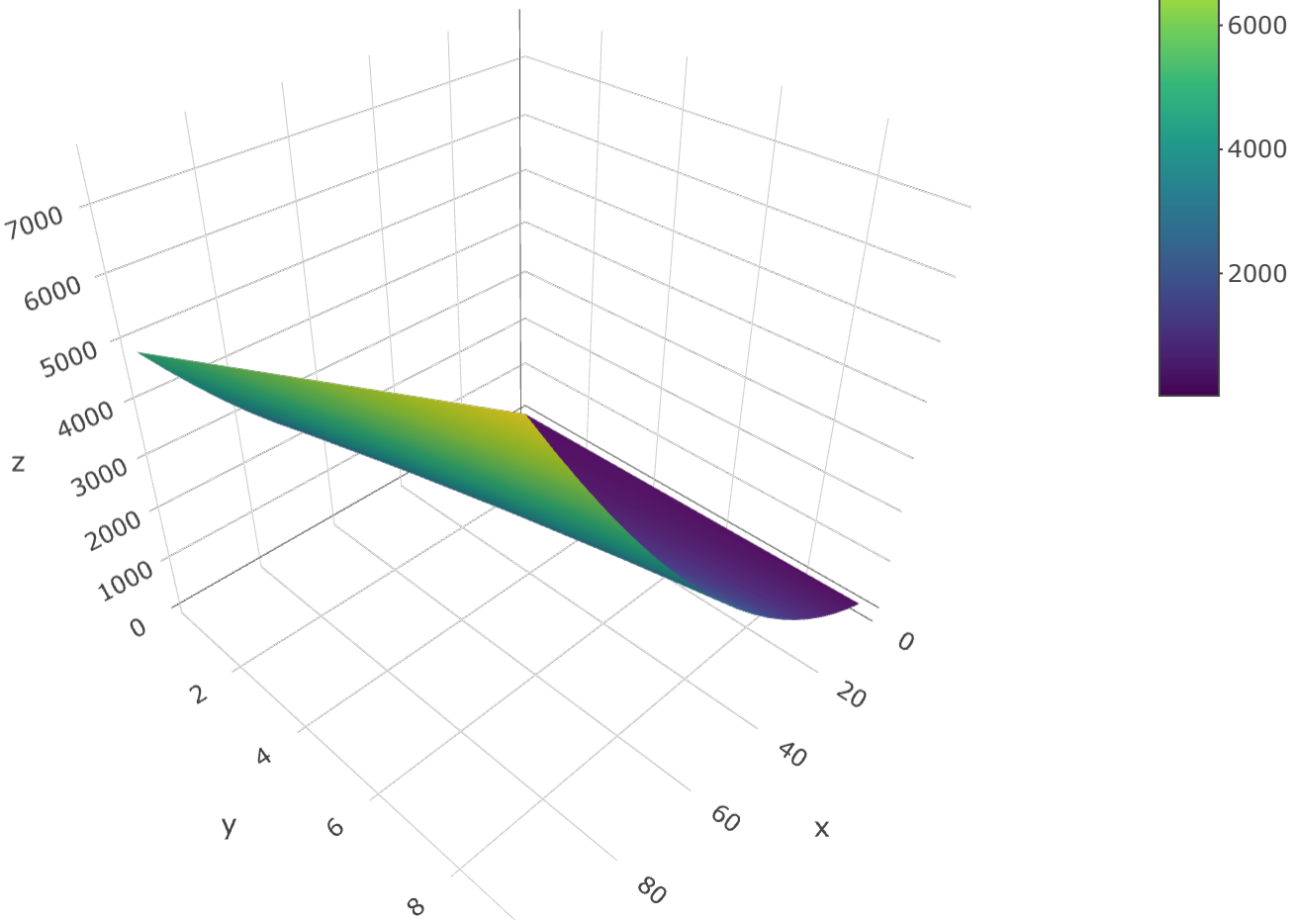
$$y = 12 - 1.6 * x + 0.5 * x^2 + 4.2 * w + 3 * x * w + rnorm(50, 0, 15$$

O que gerou os dados:

```
x = sort(runif(50,1,100))
z = x/10 + rnorm(50,0,1)
w = runif(50,0,10)
y = 12 - 1.6 * x + 0.50 * x^2 -  4.2 * w  - 3 *x * w + rnorm(50,
        0, 10)
```

(Intercept)	x	w	I(x^2)	x:w
29.1559	-2.260714	2.76599	0.5049954	3.016414

Gráfico do modelo



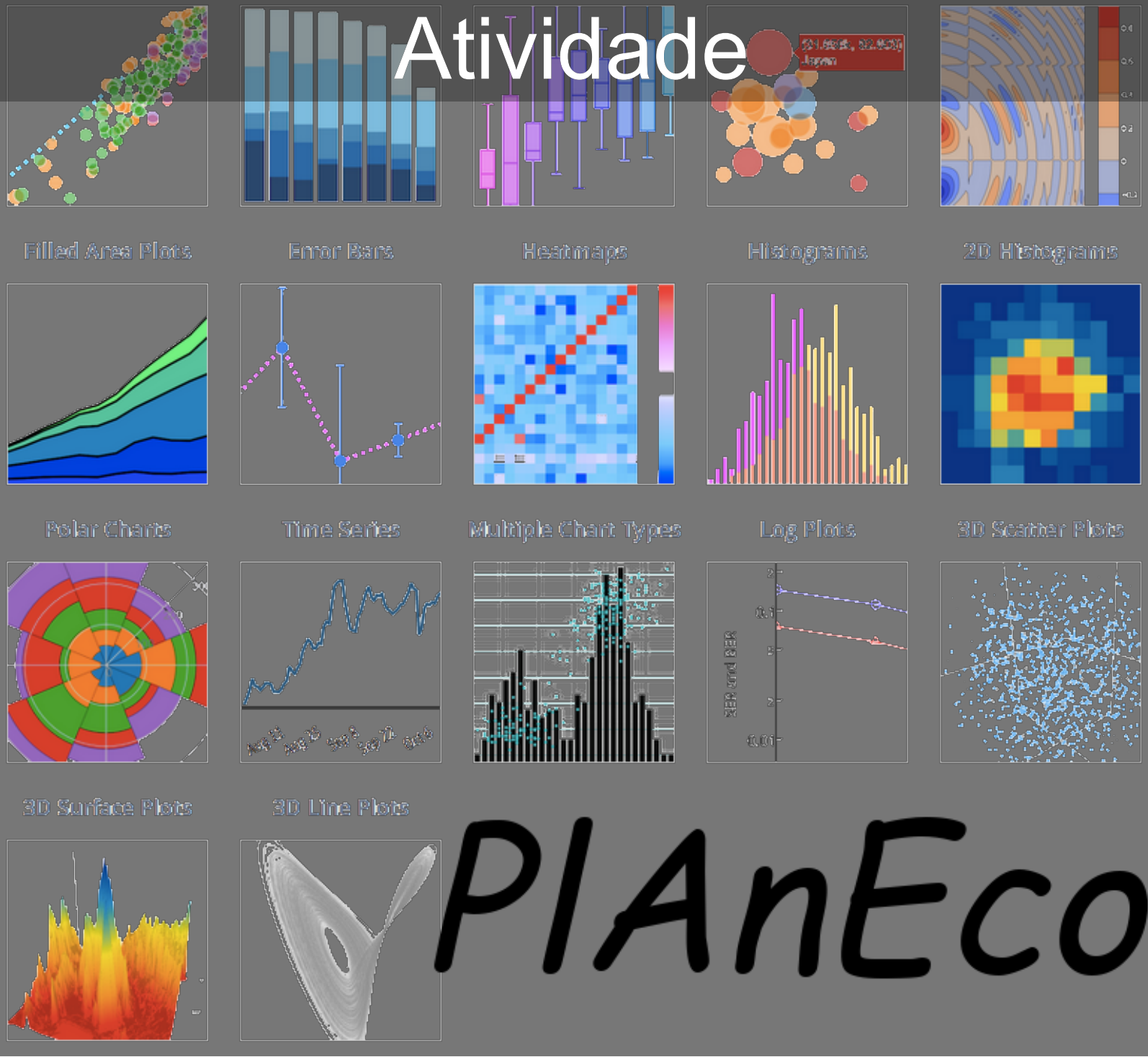
Problema da colinearidade

```
Call:
lm(formula = y ~ w + z + I(z^2) + z:w, data = yxzw)

Residuals:
     Min       1Q   Median       3Q      Max 
-1279.94  -412.04   -68.01   219.94  2058.01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  134.681    514.396   0.262  0.79465   
w              7.094     76.927   0.092  0.92693   
z             88.564    143.217   0.618  0.53944   
I(z^2)        31.596     10.101   3.128  0.00308 **
w:z           31.568     10.877   2.902  0.00572 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 638 on 45 degrees of freedom
Multiple R-squared:  0.9145,    Adjusted R-squared:  0.9069 
F-statistic: 120.4 on 4 and 45 DF,  p-value: < 2.2e-16
```

Modelo Linear Múltiplo:

- quais variáveis incluir
- relações não lineares
- interações entre variáveis
- correlação entre variáveis preditoras (colinearidade)
- saturação do modelo (complexidade)

Atividade



PlAnEco