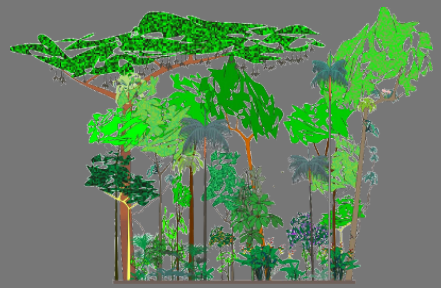
Modelos Lineares

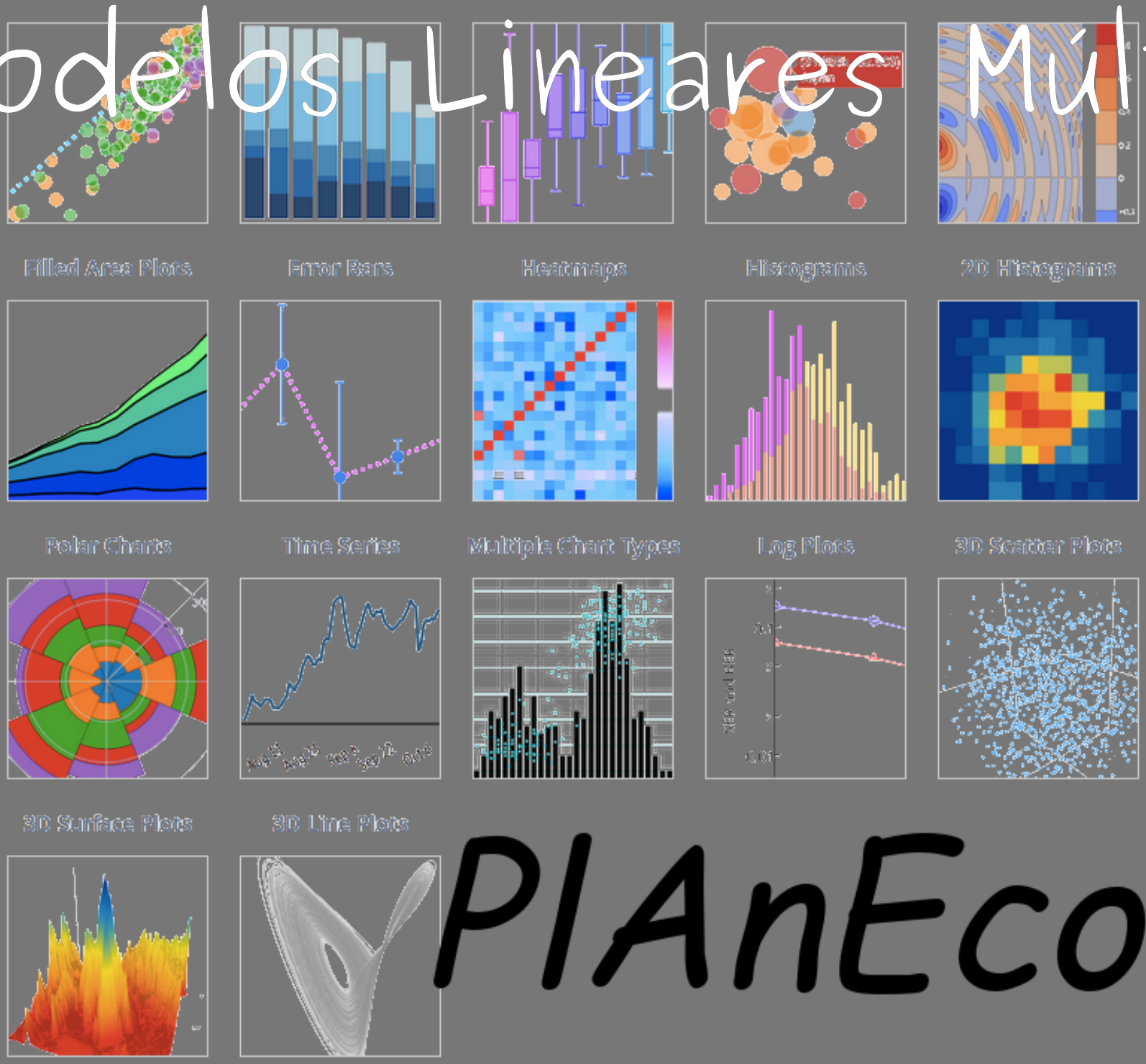
múltiplas preditoras

Alexandre Adalardo de Oliveira

PlanECO 2019



Modelos Lineares Múltipl



PlAnEco

Conceitos

Conceitos

- preditoras contínuas e categóricas
- interação entre preditoras
- matriz do modelo (algebra linear)
- simplificação do modelo
- colinearidade

Modelo Linear Simples

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

$$\epsilon = N(0, \sigma)$$

Modelo Linear Múltiplo

$$y = \alpha + \sum \beta_i x_i + \epsilon$$

$$\epsilon = N(0, \sigma)$$

Retomando o Modelo Linear

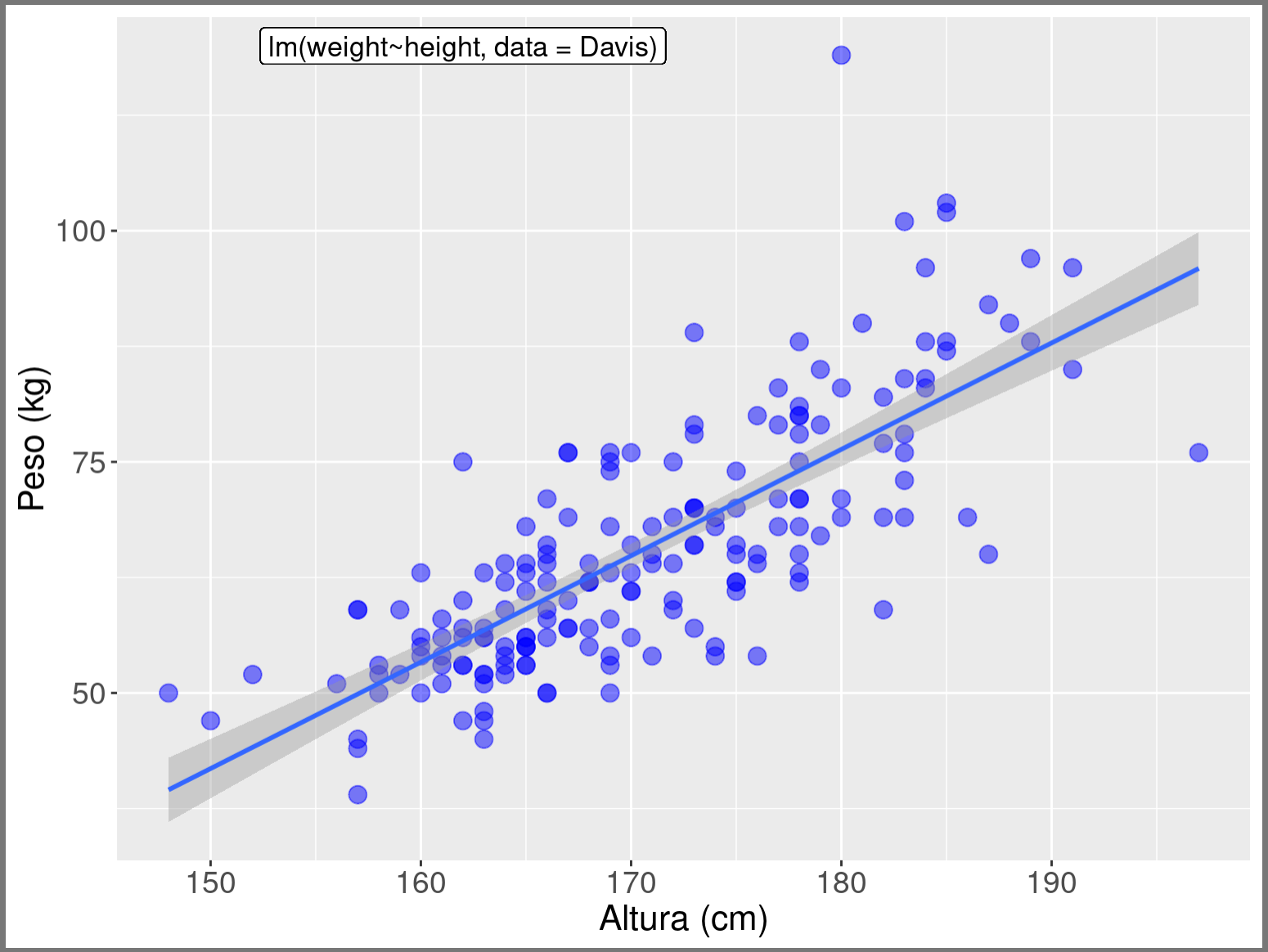
Davis 1990

	sex	weight	height	repwt	repht
1	M	77	182	77	180
2	F	58	161	51	159
3	F	53	161	54	158
4	M	68	177	70	175
5	F	59	157	59	155
194	F	51	156	51	158
195	F	62	164	61	161
196	M	74	175	71	175
197	M	83	180	80	180
199	M	90	181	91	178
200	M	79	177	81	178

Davis (1990)

Variável	Descrição	Tipo
sex	sexo	categórica dois níveis (M, F)
weight	peso	continua (kg)
height	altura	continua (cm)
repwt	peso reportado	continua (kg)
repht	altura reportada	continua (cm)

peso ~ weight


lm(weight~height, data = Davis)
Peso (kg)
100
75
50
150
160
170
180
190
Altura (cm)

Modelo Linear

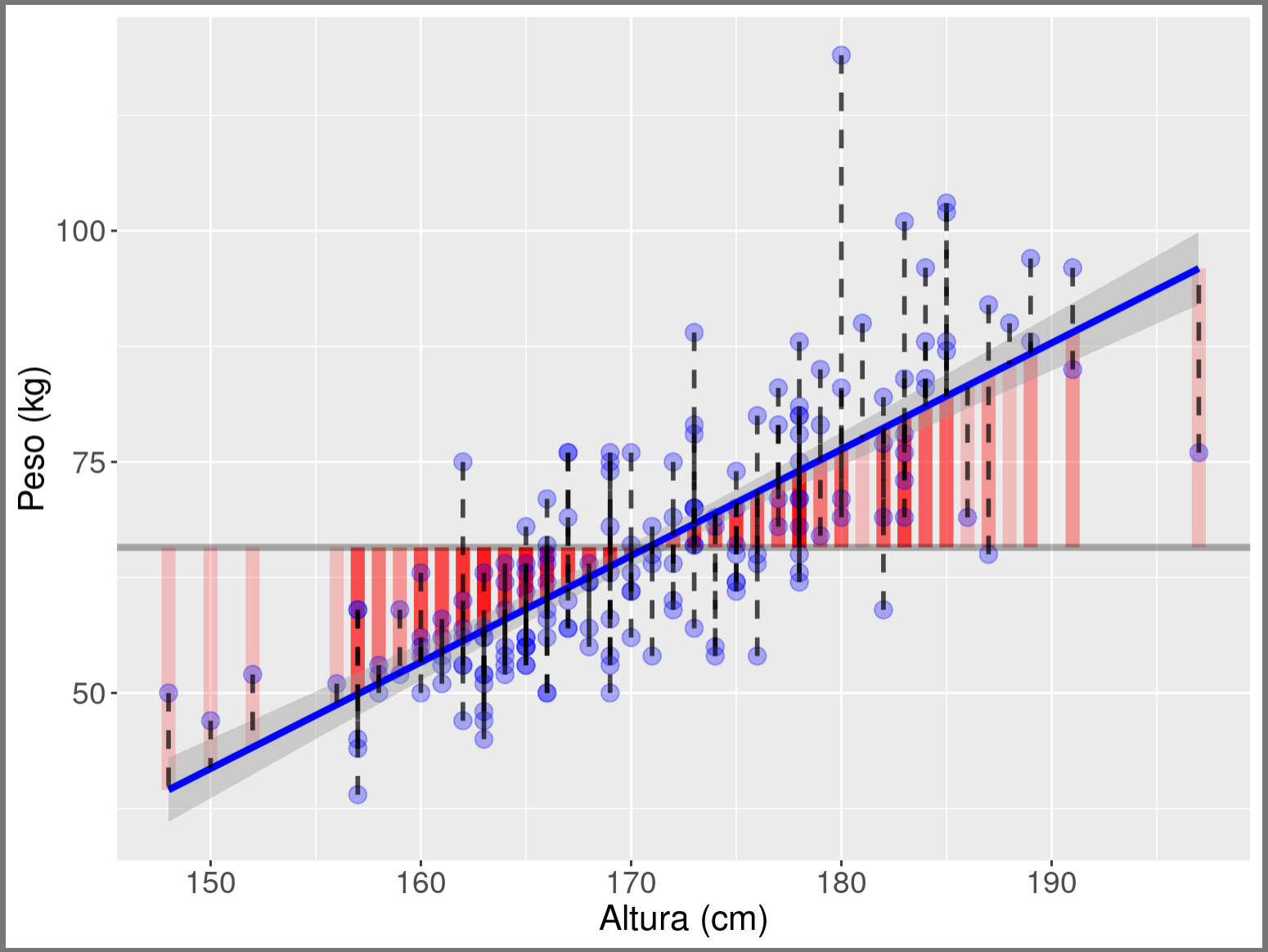
```
summary(lmdavis)
```

```
##
## 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
```

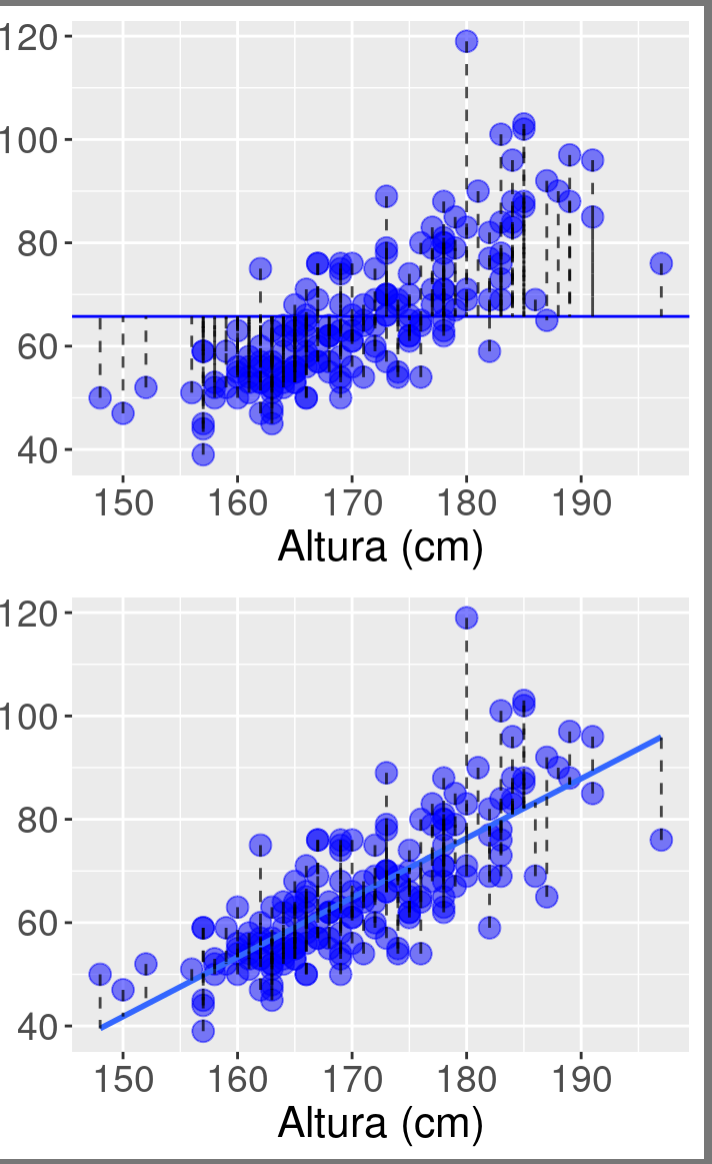
```
anova(lmdavis)
```

```
## 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
```

Partição da Variância



lm(peso ~ altura)



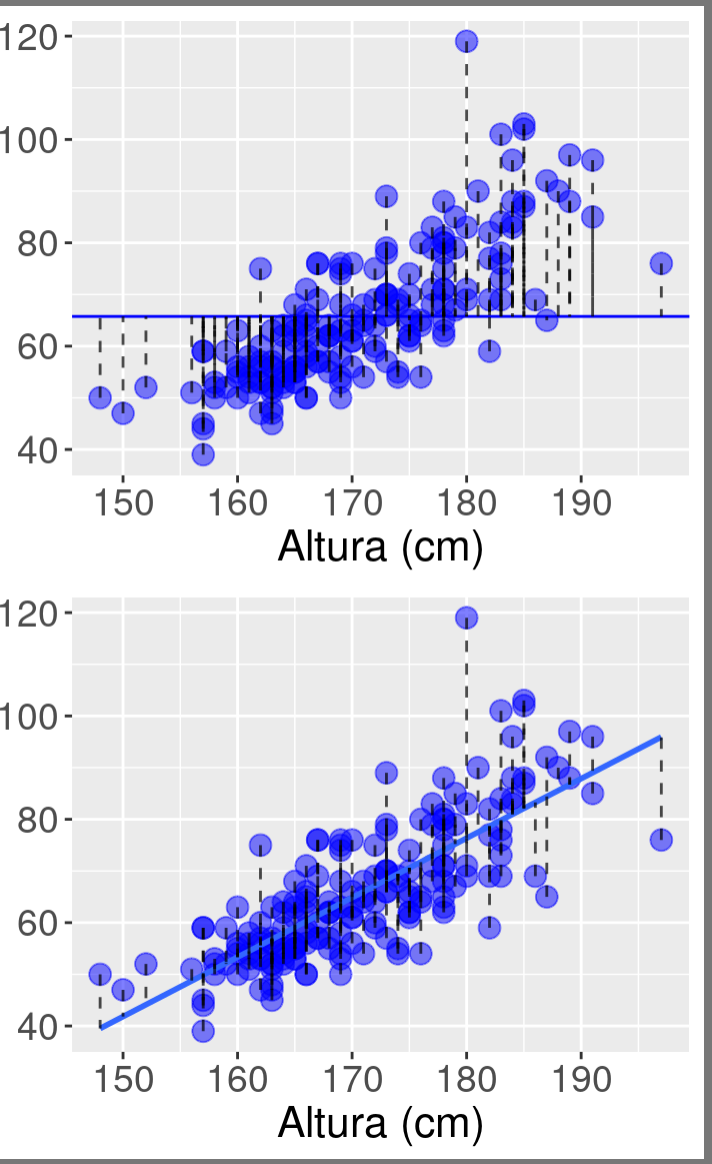

```
sum((Davis$weigh
  mean(Davis$wei
```

```
## [1] 32367.75
```

```
sum((Davis$weigh
  predict(lmdavi
```

```
## [1] 13272.71
```

lm(peso ~ altura)




```
anova(davisNull,
```

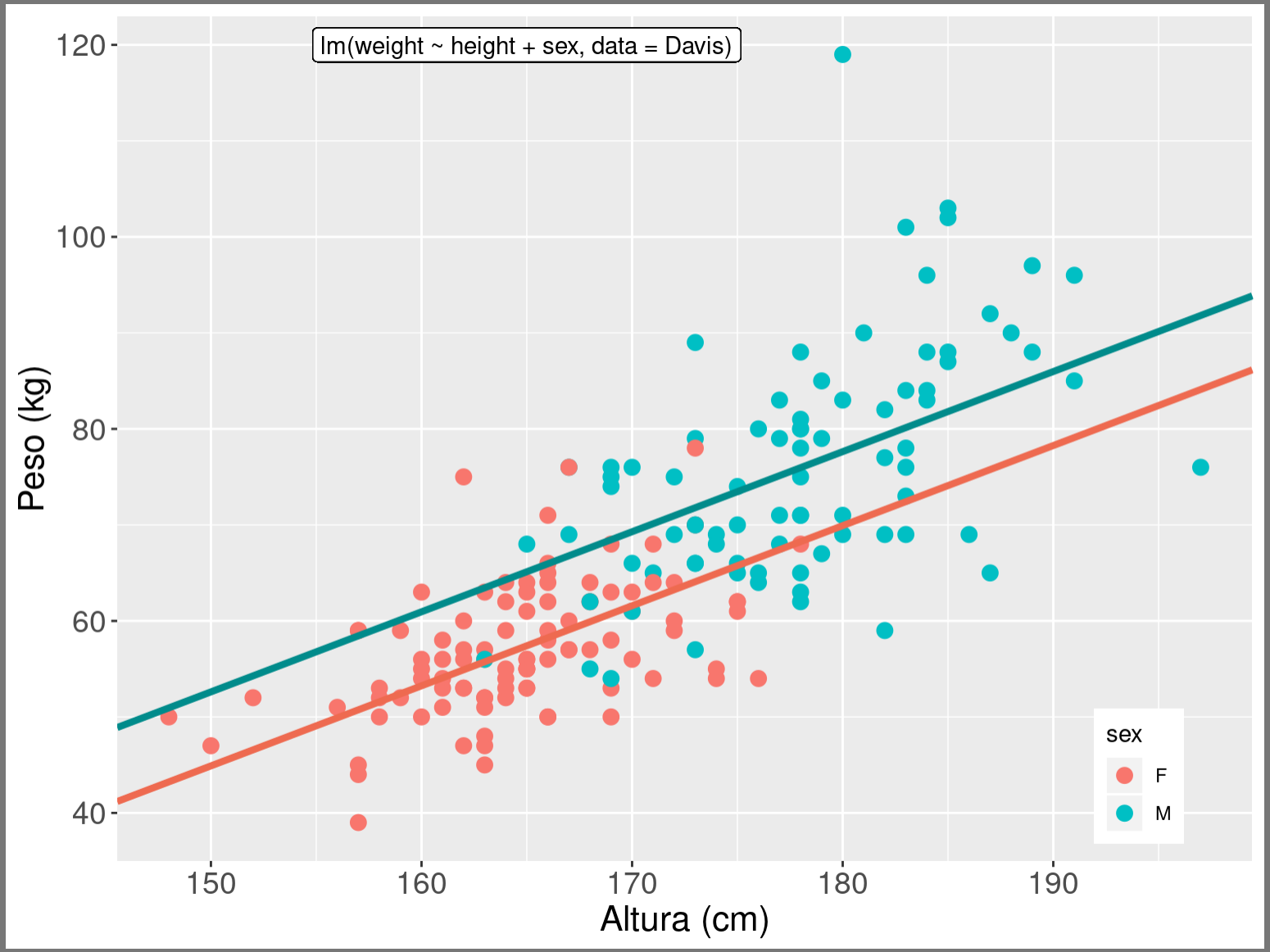
```
## Analysis of V
##
## Model 1: weig
## Model 2: weig
##   Res.Df   RS
## 1    179 3236
## 2    178 1327
## ---
## Signif. codes
```

p_{va}

p_{valo}

Modelo Linear:

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 '
##
## 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)

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

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

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

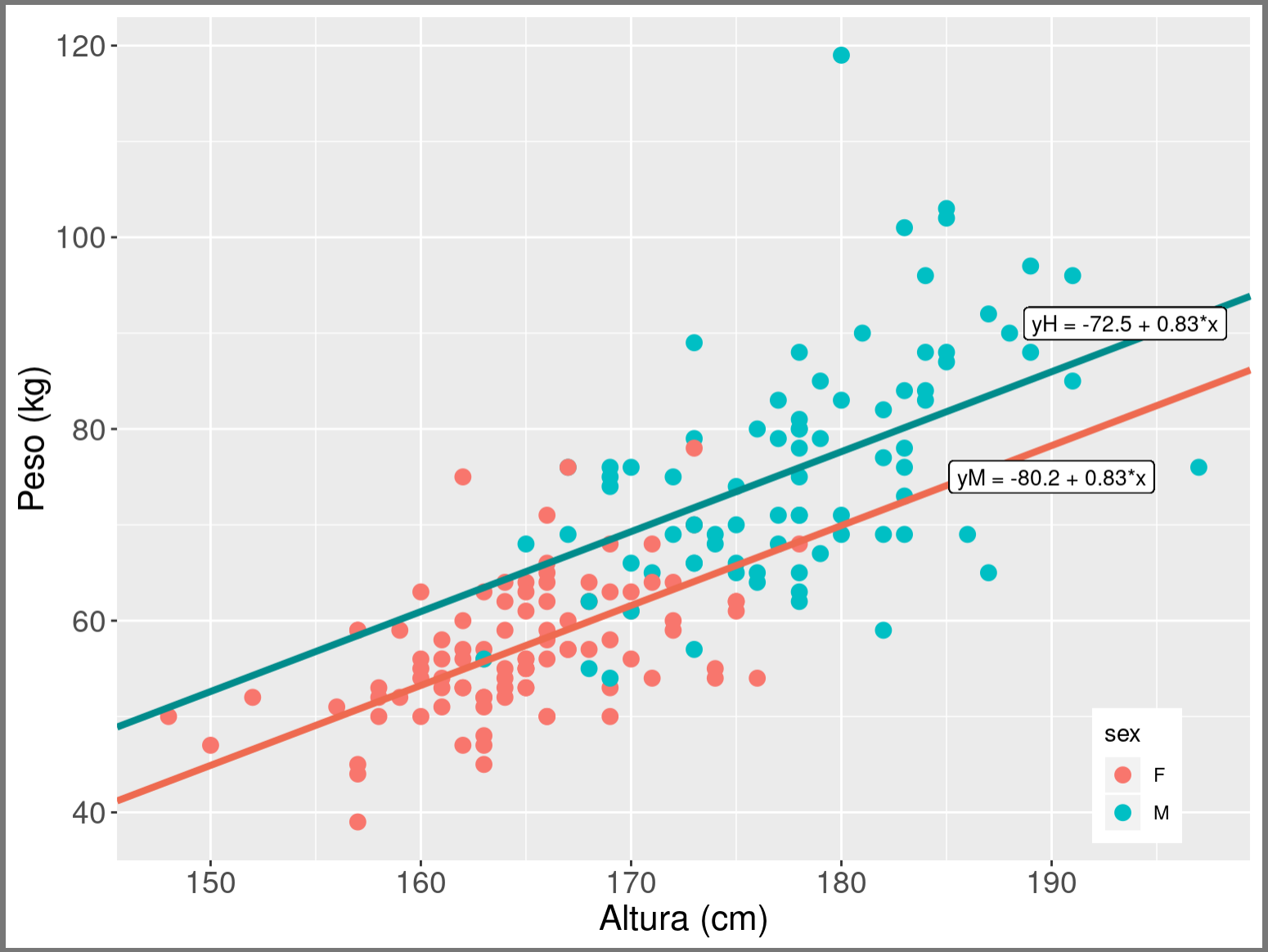
Homem (sex=1)

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

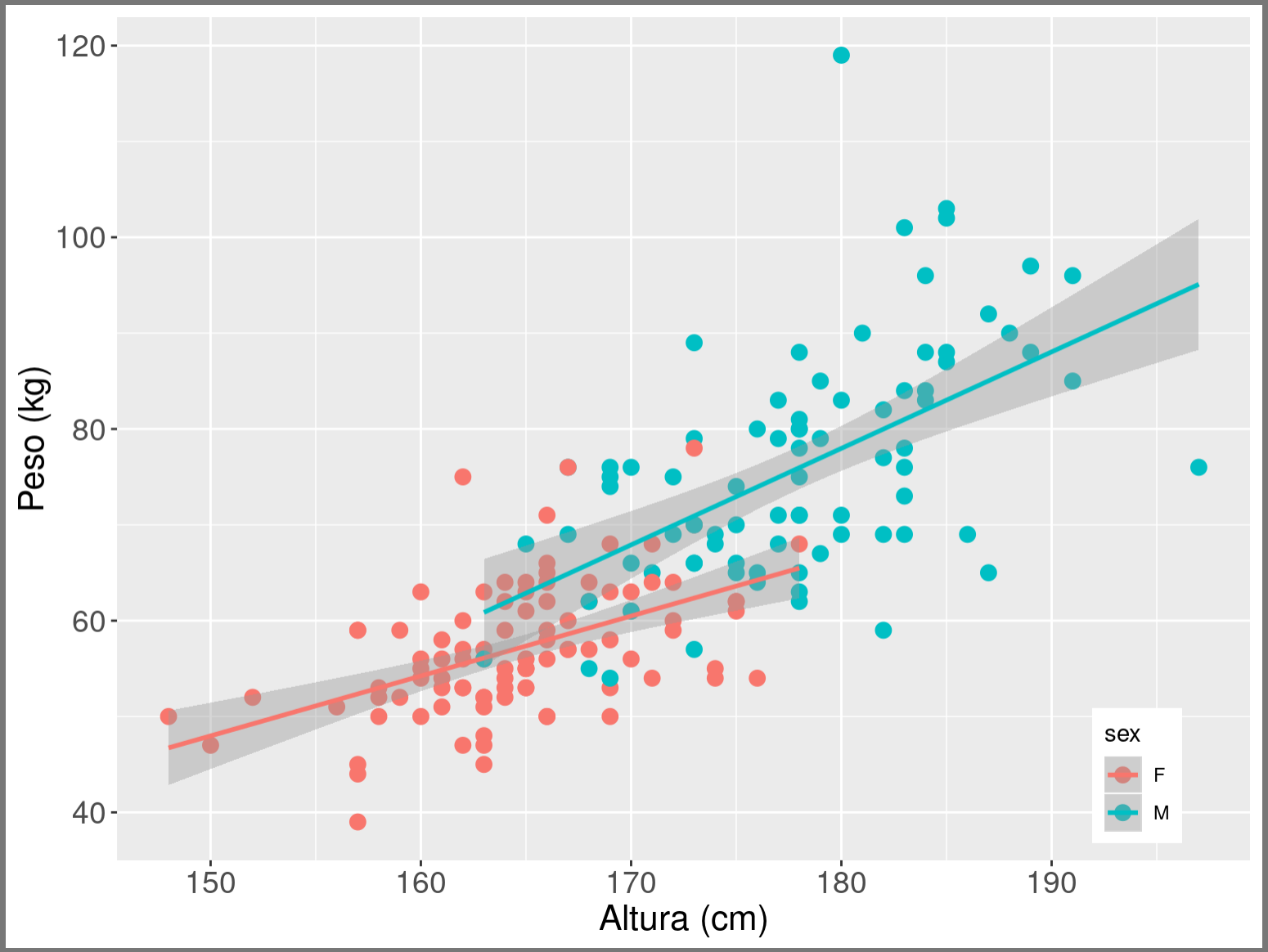
$$w_m = \hat{\alpha} + \hat{\beta}_s + \hat{\beta}_h * height$$

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

weight ~ height + sex



Interação: height:sex



weight ~ height + sex + height:sex

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

```
##
## Call:
## lm(formula = weight ~ height + sex + sex:height, data = Dav
##
## 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 '
##
## 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
```

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
```

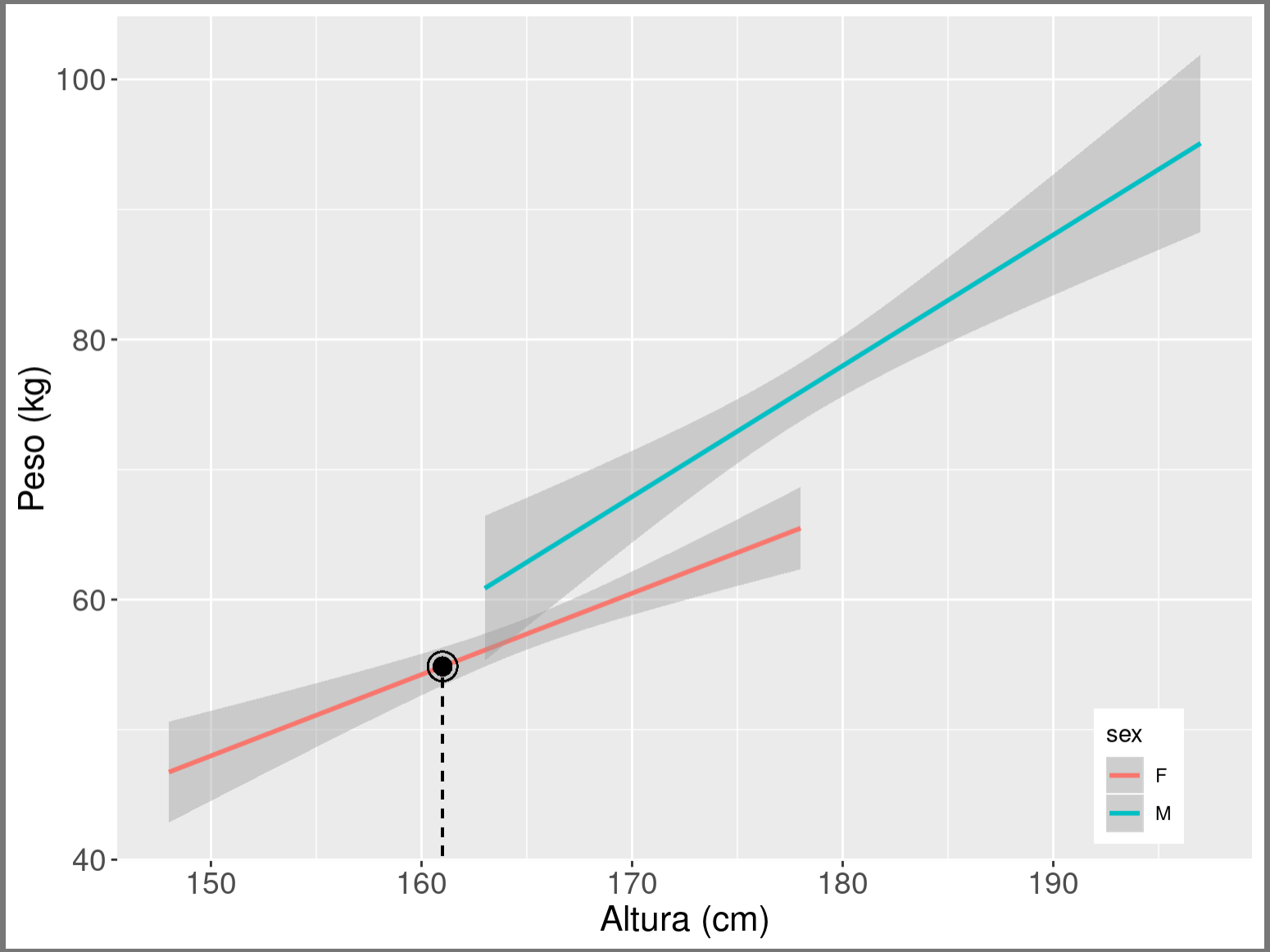
$$w = \hat{\alpha} + \hat{\beta}_h height$$

```
-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
```

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

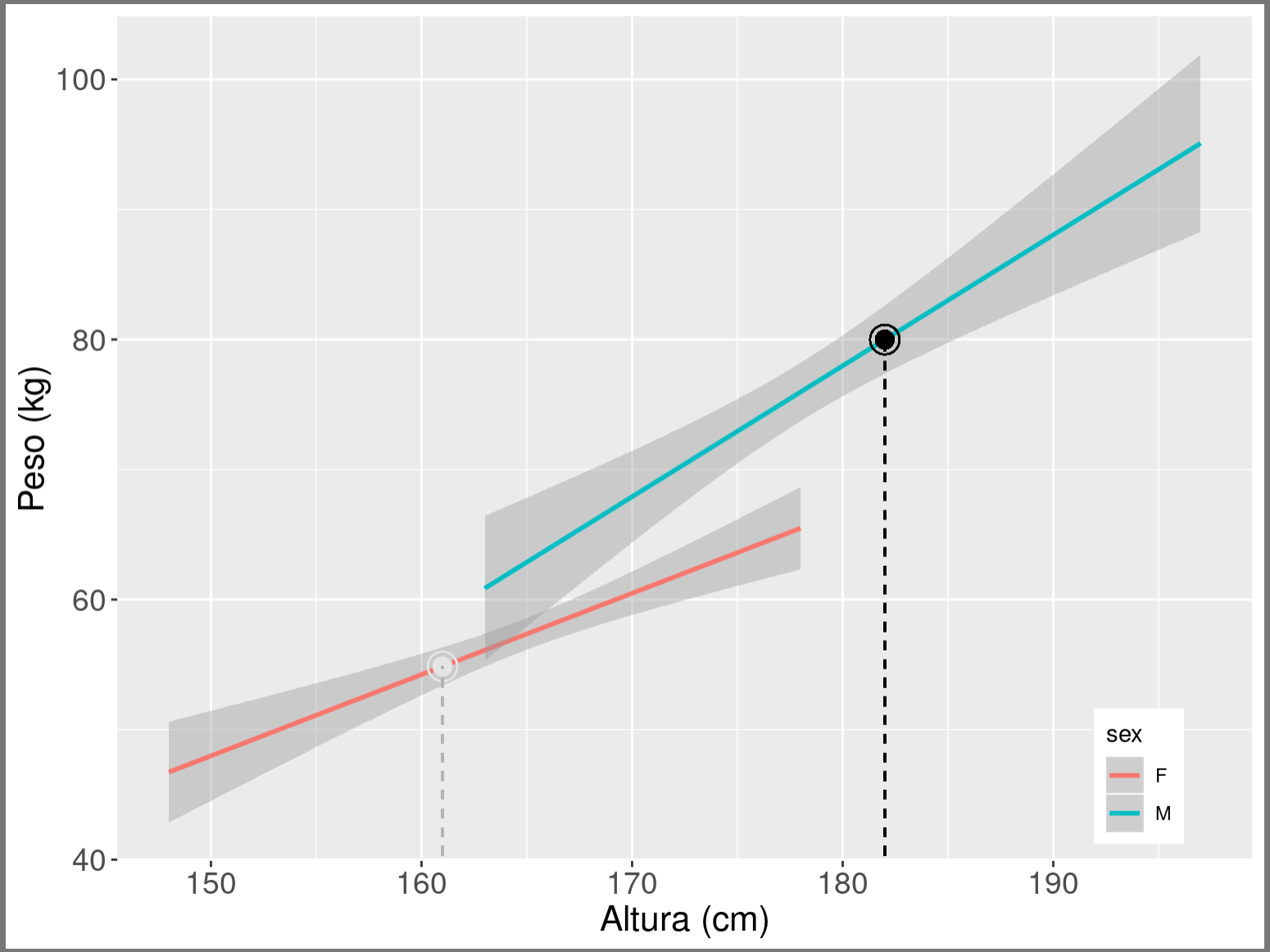
$$w = \hat{\alpha} + \hat{\beta}_s + (\hat{\beta}_h + \hat{\beta}_{s:h}) * height$$

```
-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.



Matrix do Modelo

Primeiros registros nos dados

```
##   sex weight height
## 1   M     77    182
## 2   F     58    161
```

Matrix do Modelo (linhas 1 e 2)

```
##   (Intercept) height sexM height:sexM
## 1           1    182    1         182
## 2           1    161    0           0
```

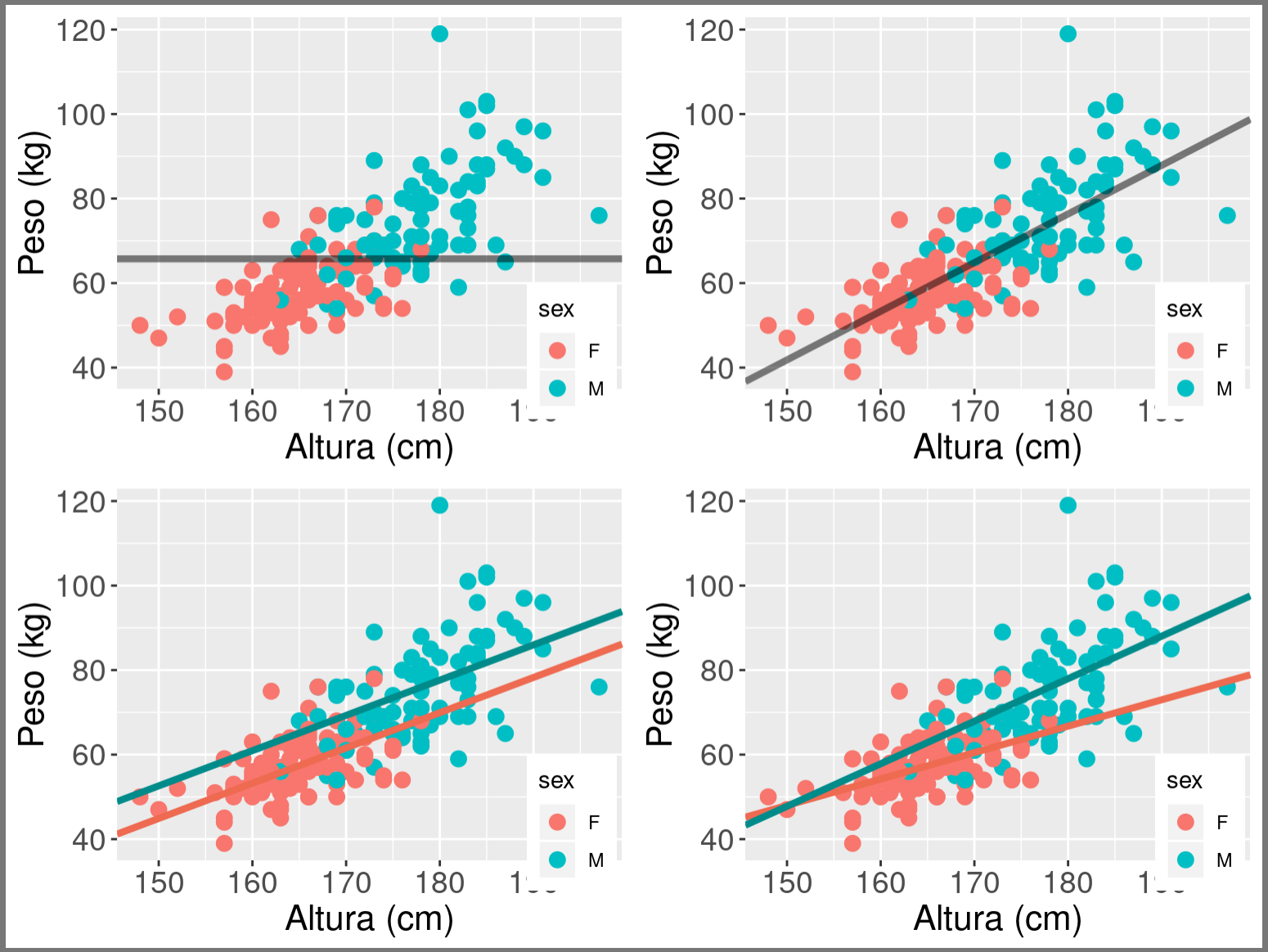
Coeficientes do Modelo

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

Multiplicação Matricial

```
##       [,1]
## 1 79.99018
## 2 54.85893
```

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

Deviance (Generalização):

Distância ao modelo saturado.

$$D = 2 * (LL_1 - LL_0)$$

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

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

Princípio da parcimônia (Navalha de Occam)

- mínimo número de parâmetros
- linear é melhor que não-linear
- reter menos pressupostos
- simplificado ao 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
3. ordem de remoção de termos:
    - interação não significativos (maior ordem)
    - termos quadráticos ou não lineares
    - variáveis explicativas não significativas
    - agrupe níveis de fatores sem diferença
    - ANCOVA: intercepto não significativo -> 0

Simplificação do modelo:

Critério para a tomada de decisão (Variância)

Compare o modelo anterior com o simplificado

A diferença não é significativa:

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

A difereça é significativa

```
* retenha o modelo complexo
* modelo MINÍMO ADEQUADO
```

Simplificando Modelo: exemplo

Modelo cheio

```
##
## Call:
## lm(formula = weight ~ height + sex + sex:height, data = Dav
##
## 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 '
##
## 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: exemplo

weight ~ height + sex + sex:height

weight ~ height + sex

```
anova(lmdavisfull, lmdavis01)
```

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

Simplificando Modelo: exemplo

weight ~ height + sex

```
##
## 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 '
##
## 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
```

Simplificando Modelo: exemplo

weight ~ height + sex

weight ~ height

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

Modelo Mínimo Adequado

```
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 '
##
## 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

```
anova(lmdavis01)
```

```
## Analysis of Variance Table
##
## Response: weight
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## height       1 19095.0 19095.0  280.04 < 2.2e-16 ***
## sex          1  1203.5  1203.5   17.65 4.204e-05 ***
## Residuals 177 12069.2    68.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Modelo Mínimo Adequado

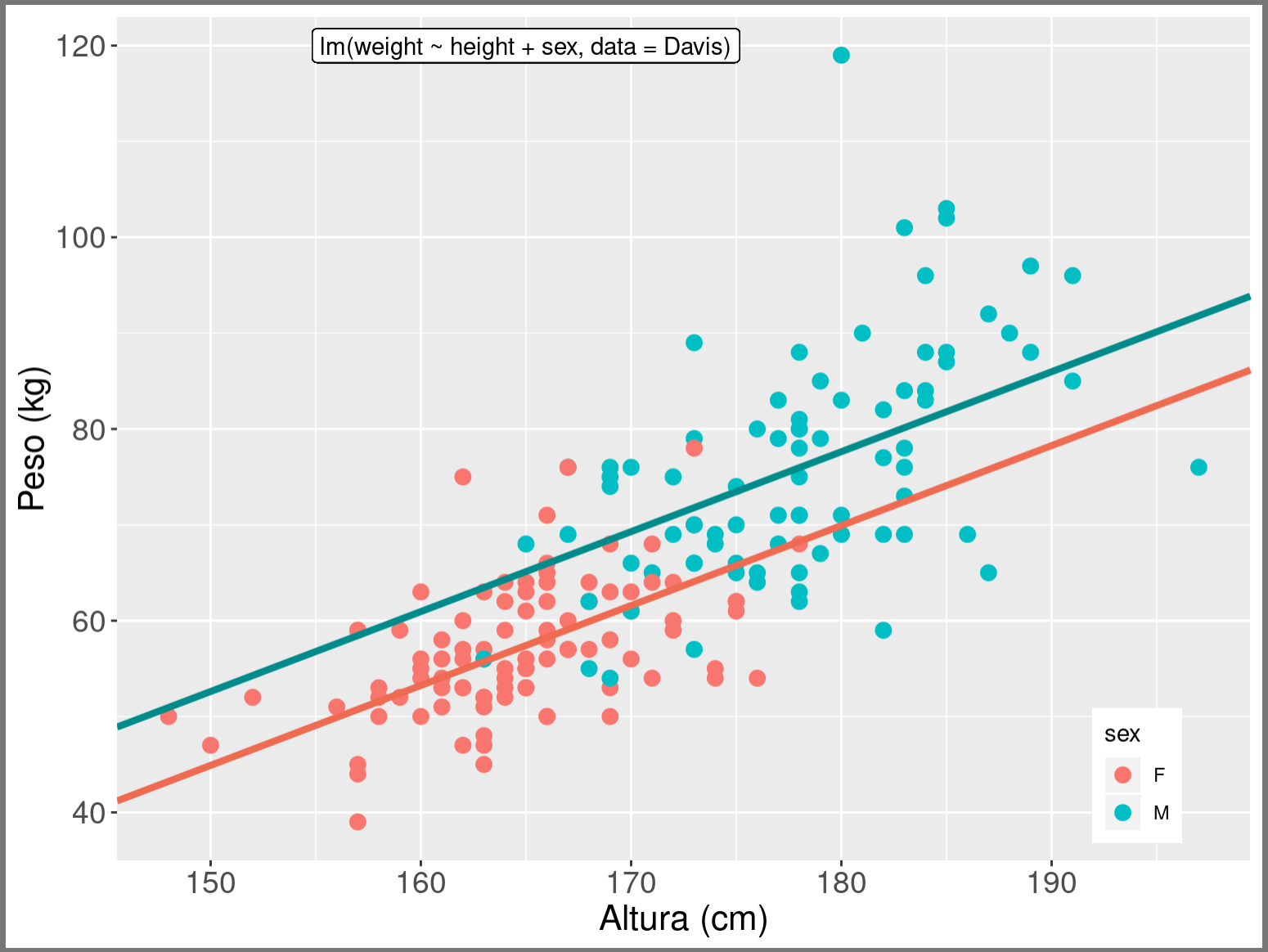
```
coef(lmdavis01)
```

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

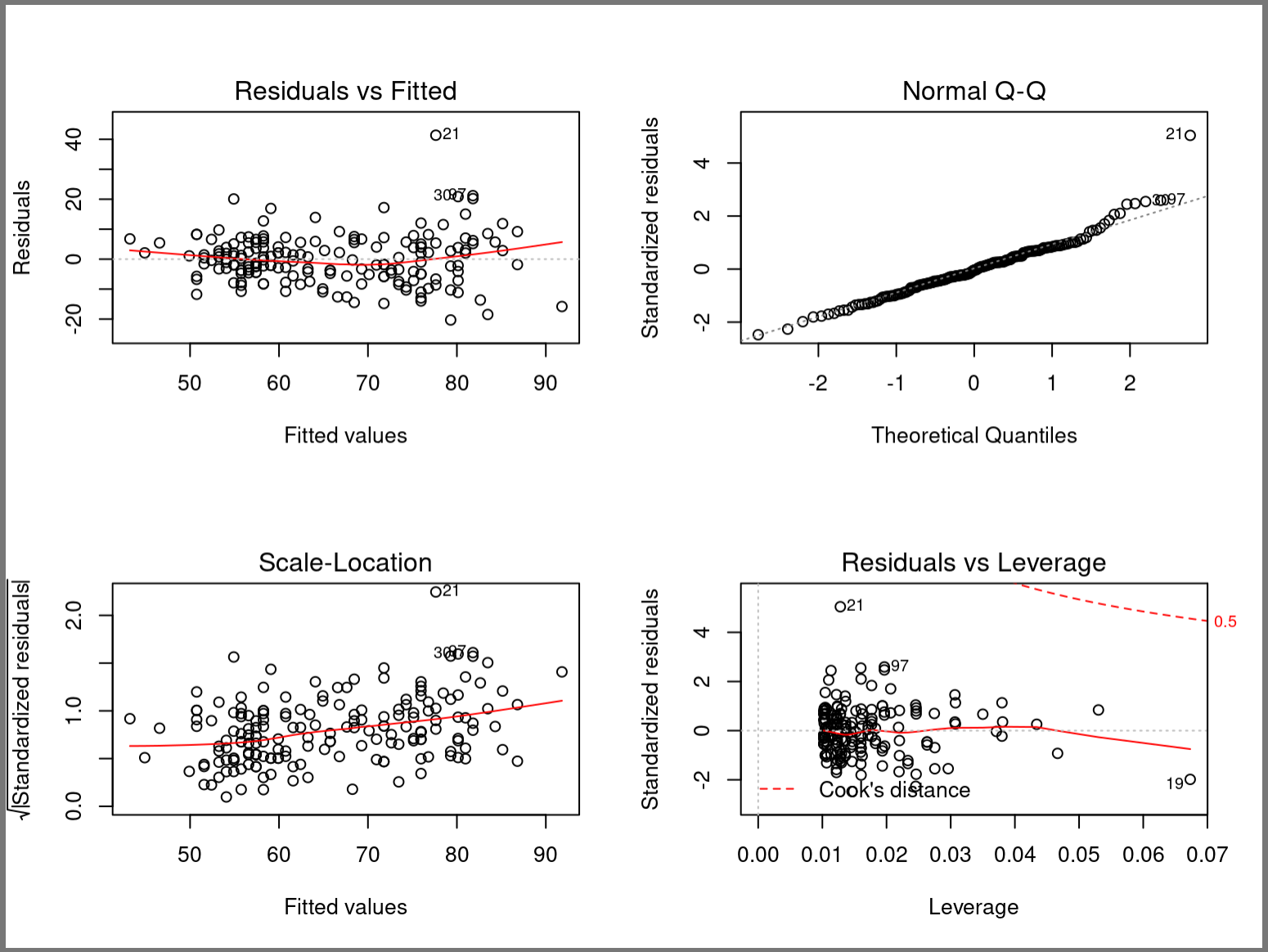
```
confint(lmdavis01)
```

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

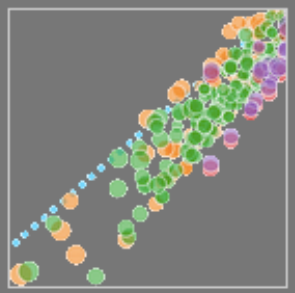
Modelo Mínimo Adequado



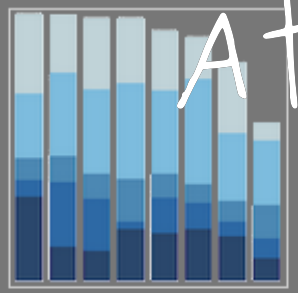
Diagnóstico do Modelo:



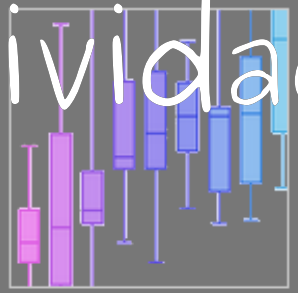
Line and Scatter Plots



Bar Charts

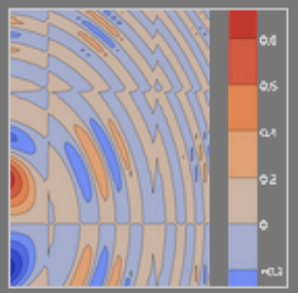


Box Plots

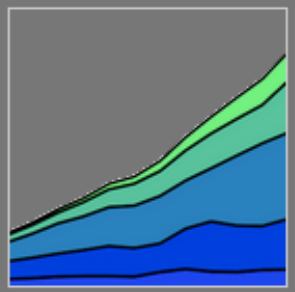


Bubble Charts

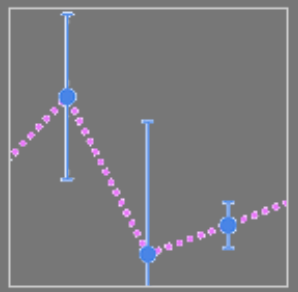
Contour Plots



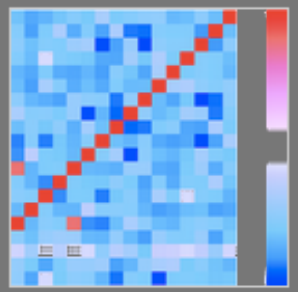
Filled Area Plots



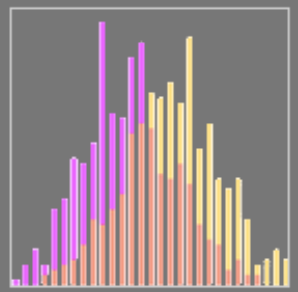
Error Bars



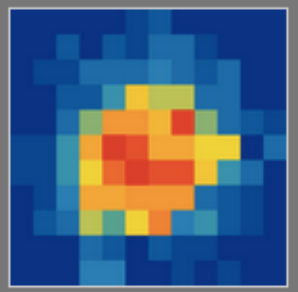
Heatmaps



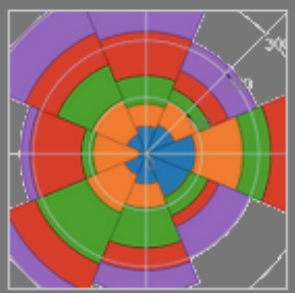
Histograms



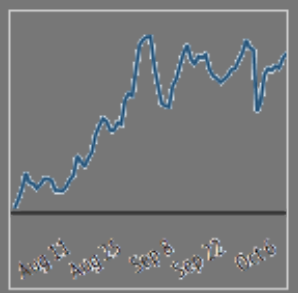
2D Histograms



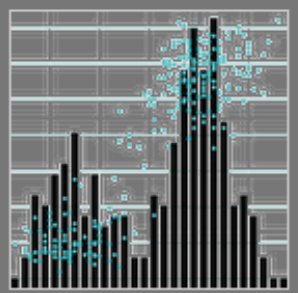
Polar Charts



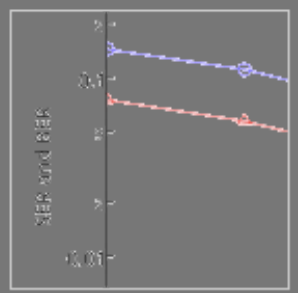
Time Series



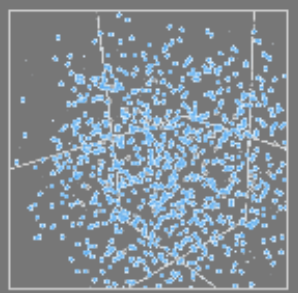
Multiple Chart Types



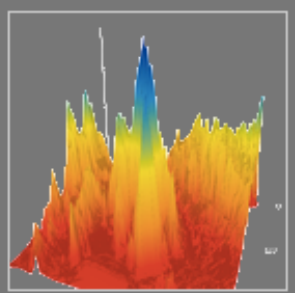
Log Plots



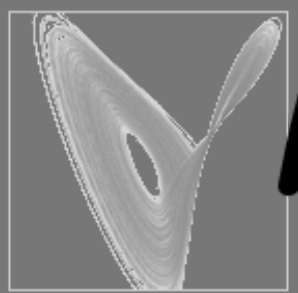
3D Scatter Plots



3D Surface Plots



3D Line Plots



Atividade

PlAnEco

Simulando dados

Modelo Linear Múltiplo:

- quais variáveis incluir
- curvatura em resposta a variável preditora
- interações entre variáveis
- correlação entre variáveis preditoras (colinearidade)
- saturação do modelo (complexidade)

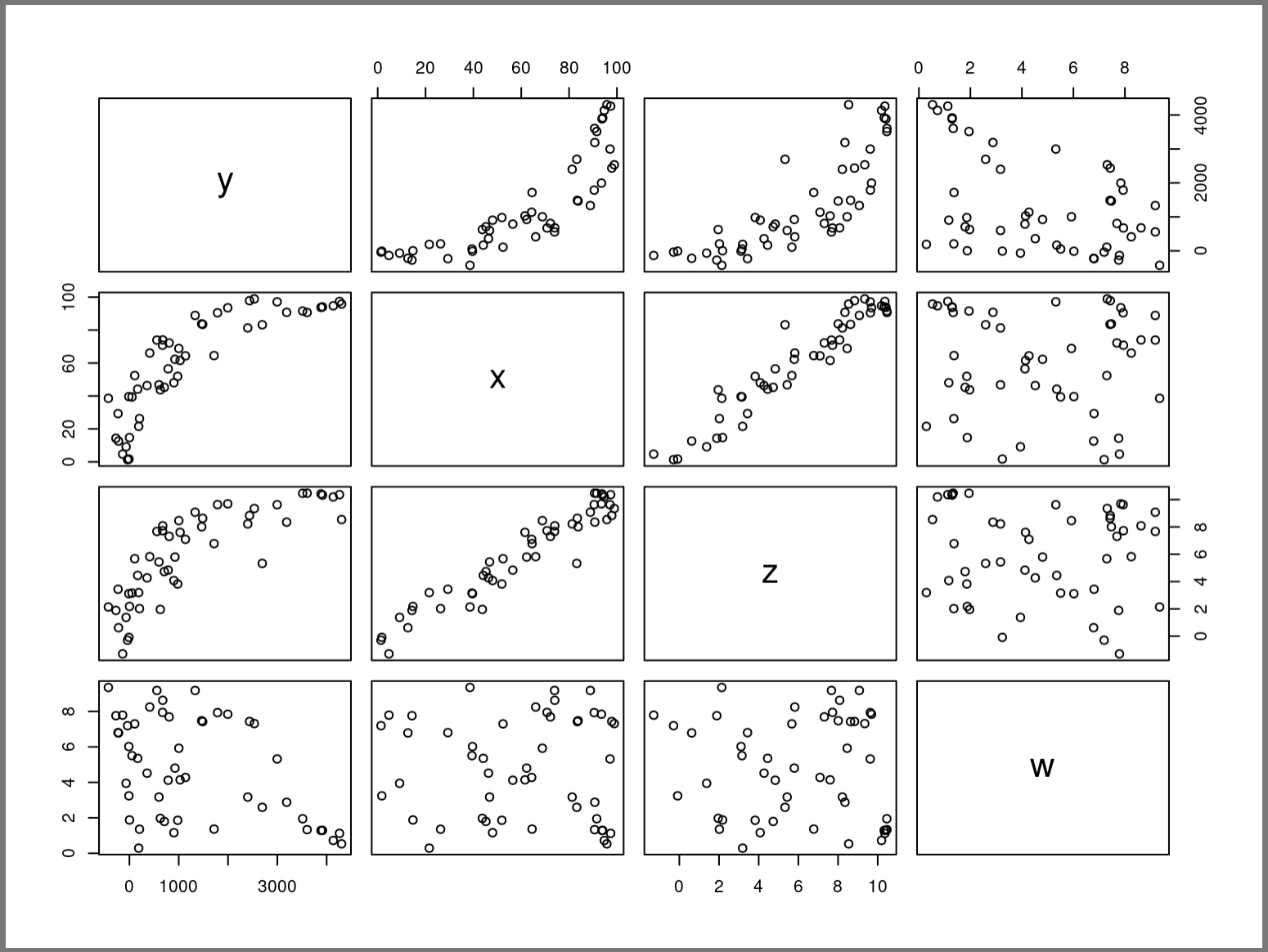
Simulando dados

y ~ x + z + w ...

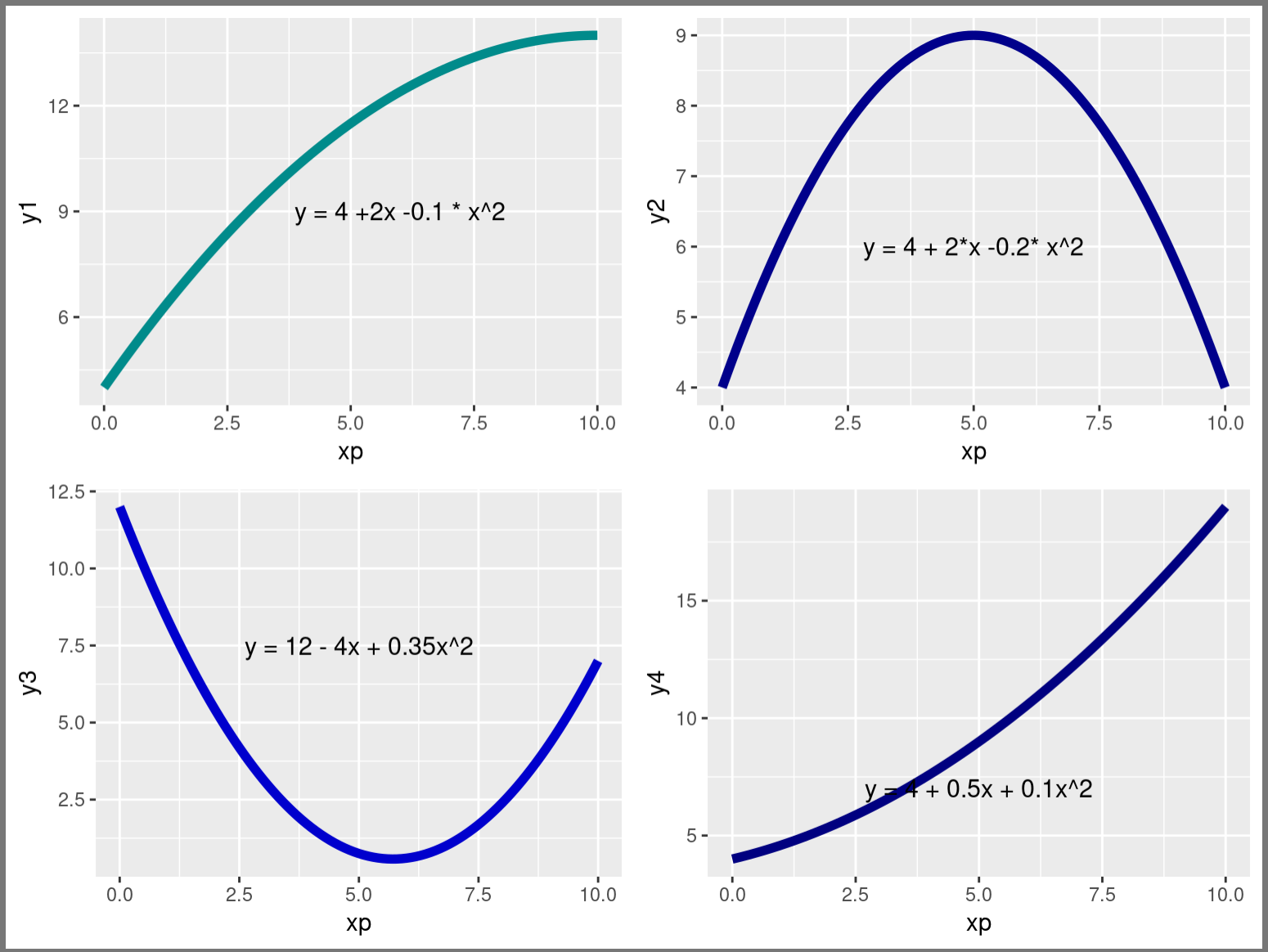
Quais variáveis estão relacionadas à resposta?

y	x	z	w
−37.479581	1.390885	−0.2913806	7.193786
−9.218105	1.726080	−0.0846613	3.240860
−137.144153	4.705672	−1.2925959	7.788095
−67.182923	9.161318	1.3762292	3.944410
−220.748670	12.631249	0.6231300	6.785929

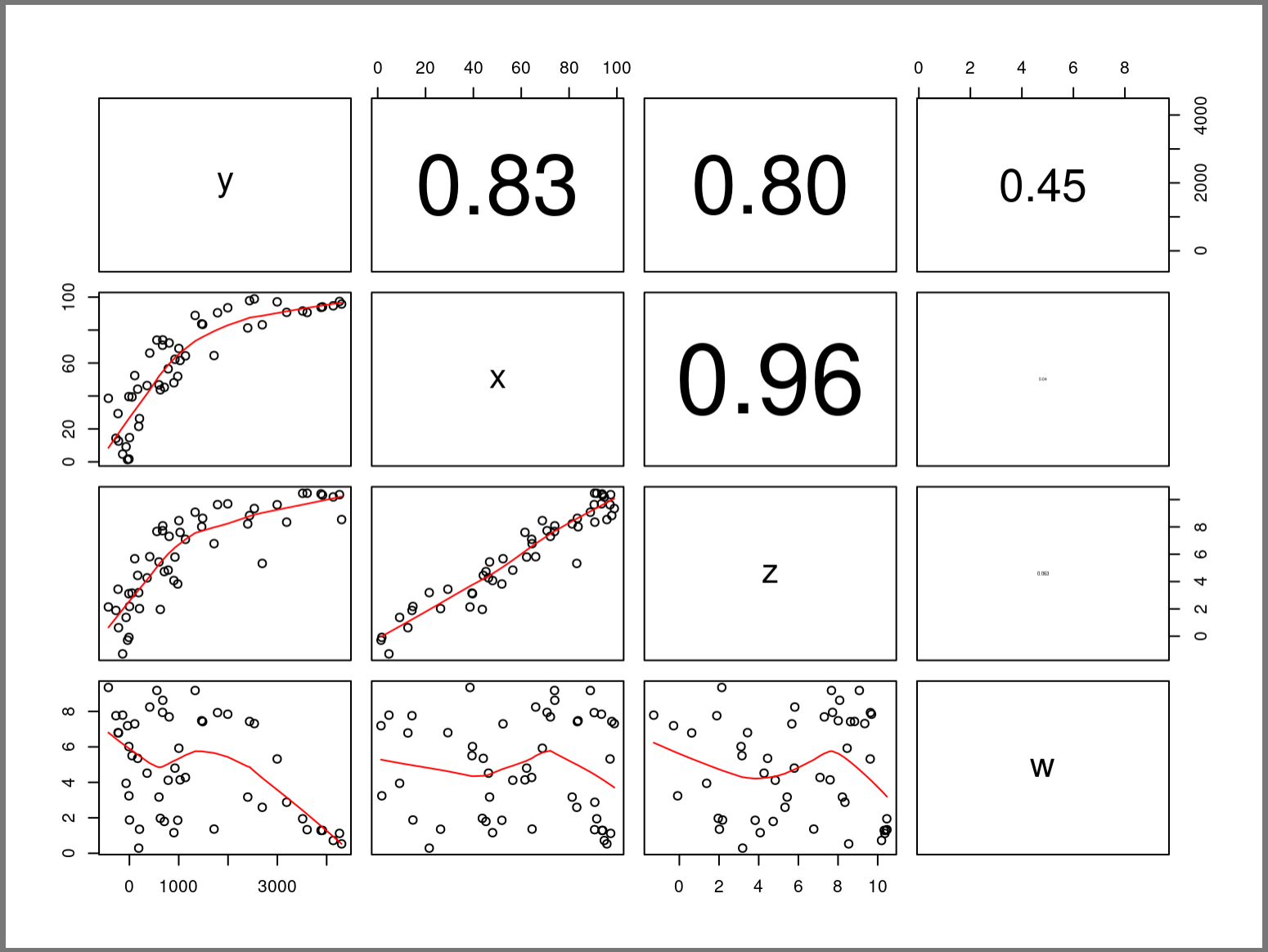
Análise Exploratória



Curvatura da relação: polinômios


y1
12
9
6
y = 4 +2x -0.1 * x^2
0.0
2.5
5.0
7.5
10.0
xp
y2
9
8
7
6
5
4
y = 4 + 2*x -0.2* x^2
0.0
2.5
5.0
7.5
10.0
xp
y3
12.5
10.0
7.5
5.0
2.5
y = 12 - 4x + 0.35x^2
0.0
2.5
5.0
7.5
10.0
xp
y4
15
10
5
y = 4 + 0.5x + 0.1x^2
0.0
2.5
5.0
7.5
10.0
xp

Correlação entre preditoras



Correlação entre preditoras

Indice de colinearidade (confirmar)

VIF: Variance Inflation Factor

Proporcional a variação compartilhada com outras preditoras

$$VIF = \frac{1}{1 - R_k^2}$$

R_k^2 : coeficiente de determinação da preditora (k) em relação a outras preditoras do modelo

- $VIF = 1$: não há variação compartilhada;
- $VIF = 4$: 75% de variação explicada ;
- $VIF = 10$: 90% de variação explicada;

VIF

```
vif(lm(y ~ x + z + w ))
```

```
##          x          z          w
## 12.155388 12.183386  1.008662
```

```
vif(lm(y ~ x + w))
```

```
##       x       w
## 1.00163 1.00163
```

```
vif(lm(y ~ z + w))
```

```
##       z       w
## 1.003937 1.003937
```

Colinearidade: soluções

- reter apenas uma das variáveis colineares
- reduzir as dimensões das variáveis colineares (PCA)

Definir os termos do modelo cheio

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

Modelo Cheio

```
lmfull <- lm(y ~ x + w + z +
                       I(x^2) + I(z^2) +
                       x:w + x:z + z:w + z:w:x, data = yxzw)
```

```r
summary(lmfull)
```

```
##
## 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
## -16.8495  -5.7956  -0.3322   4.2633  29.6627
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.193915  11.442878   1.852   0.0714 .
## x           -2.572884   0.536259  -4.798 2.25e-05 ***
## w           -4.332059   1.963077  -2.207   0.0331 *
## z            5.068630   5.895282   0.860   0.3950
## I(x^2)       0.518748   0.012522  41.426  < 2e-16 ***
## I(z^2)       0.516183   1.069157   0.483   0.6319
## x:w         -3.022575   0.070387 -42.942  < 2e-16 ***
## x:z         -0.199390   0.224934  -0.886   0.3807
## w:z          0.279907   0.726247   0.385   0.7020
## x:w:z       -0.000639   0.006779  -0.094   0.9254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.262 on 40 degrees of freedom
```

Simplificando o modelo

```
lm01 <-  lm(y ~ x + w + z +
                    I( x^2) + I(z^2) +
                    x:w + x:z + z:w, data = yxzw)

anova(lmfull, lm01)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + z:w + z:
## Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + z:w
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     40 3431.1
## 2     41 3431.9 -1  -0.76203 0.0089 0.9254
```

```r
summary(lm01)
```

```
##
## Call:
## lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + z:w,
##     data = yxzw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8821  -5.8719  -0.4094   4.2821  29.5060
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.39839    7.63268   2.673  0.01075 *
## x           -2.56245    0.51833  -4.944 1.34e-05 ***
## w           -4.19093    1.25405  -3.342  0.00178 **
## z            5.34953    5.02464   1.065  0.29326
## I(x^2)       0.51873    0.01237  41.938  < 2e-16 ***
## I(z^2)       0.51146    1.05499   0.485  0.63040
## x:w         -3.02560    0.06186 -48.911  < 2e-16 ***
## x:z         -0.20205    0.22044  -0.917  0.36471
## w:z          0.23783    0.56583   0.420  0.67645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.149 on 41 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
```

Simplificando o modelo

```
lm02 <- lm(y ~ x + w + z +
                     I( x^2) + I(z^2) +
                     x:w + z:w, data = yxzw)
anova(lm01, lm02)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + x:z + z:w
## Model 2: y ~ x + w + z + I(x^2) + I(z^2) + x:w + z:w
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     41 3431.9
## 2     42 3502.2 -1   -70.327 0.8402 0.3647
```

```r
summary(lm02)
```

```
##
## Call:
## lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w + z:w, data =
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4363  -6.1137  -0.4808   4.3176  29.9606
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.818074   7.604429   2.738  0.00904 **
## x           -2.324930   0.448053  -5.189 5.75e-06 ***
## w           -4.240332   1.250502  -3.391  0.00153 **
## z            3.209428   4.440771   0.723  0.47386
## I(x^2)       0.507757   0.003094 164.089  < 2e-16 ***
## I(z^2)      -0.422661   0.272336  -1.552  0.12817
## x:w         -3.035963   0.060703 -50.013  < 2e-16 ***
## w:z          0.329621   0.555841   0.593  0.55635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.132 on 42 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.644e+05 on 7 and 42 DF,  p-value: < 2.2e-16
```

Simplificando o modelo

```
lm03 <- lm(y ~ x + w + z +
                    I( x^2) + I(z^2) +
                    x:w, data = yxzw)
anova(lm02, lm03)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x + w + z + I(x^2) + I(z^2) + x:w + z:w
## 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 3502.2
## 2     43 3531.6 -1   -29.324 0.3517 0.5564
```

```r
summary(lm03)
```

```
##
## Call:
## lm(formula = y ~ x + w + z + I(x^2) + I(z^2) + x:w, data = yxzw
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6707  -6.3689  -0.2563   4.6686  29.5533
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 20.717359   7.545000    2.746 0.008776 **
## x           -2.474234   0.367821   -6.727  3.2e-08 ***
## w           -4.355709   1.225927   -3.553 0.000939 ***
## z            4.906277   3.370301    1.456 0.152728
## I(x^2)       0.507920   0.003059  166.052  < 2e-16 ***
## I(z^2)      -0.461620   0.262293   -1.760 0.085531 .
## x:w         -3.001515   0.017489 -171.624  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.063 on 43 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.948e+05 on 6 and 43 DF,  p-value: < 2.2e-16
```

Simplificando o modelo

```
lm04 <- lm(y ~ x + w + z +
                     I( x^2) + x:w, data = yxzw)
anova(lm03, lm04)
```

```
## 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 3531.6
## 2     44 3785.9 -1   -254.38 3.0974 0.08553 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lm04)
```

```
##
## Call:
## lm(formula = y ~ x + w + z + I(x^2) + x:w, data = yxzw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0896  -5.7077  -0.5103   4.5774  30.7566
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 23.531249   7.547344    3.118  0.00321 **
## x           -1.992187   0.251285   -7.928 5.12e-10 ***
## w           -5.205080   1.153480   -4.513 4.73e-05 ***
## z           -0.515414   1.399219   -0.368  0.71437
## I(x^2)       0.503364   0.001668  301.786  < 2e-16 ***
## x:w         -2.988520   0.016227 -184.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.276 on 44 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.231e+05 on 5 and 44 DF,  p-value: < 2.2e-16
```

Simplificando o modelo

```
 lm05 <- lm(y ~ x + w + I( x^2) + x:w, data = yxzw)
anova(lm04, lm05)
```

```
## 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 3785.9
## 2     45 3797.6 -1   -11.675 0.1357 0.7144
```

```r
summary(lm05)
```

```
##
## Call:
## lm(formula = y ~ x + w + I(x^2) + x:w, data = yxzw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9424  -5.5396  -0.7859   4.6774  30.3060
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 23.437267   7.470240    3.137    0.003 **
## x           -2.040476   0.212311   -9.611 1.78e-12 ***
## w           -5.156007   1.134704   -4.544 4.13e-05 ***
## I(x^2)       0.503330   0.001649  305.172  < 2e-16 ***
## x:w         -2.989057   0.016005 -186.753  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.186 on 45 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.843e+05 on 4 and 45 DF,  p-value: < 2.2e-16
```

Simplificando o modelo

```
lm06 <- lm(y ~ x + w + I( x^2), data = yxzw)
anova(lm05, lm06)
```

```
## 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    3798
## 2     46 2947081 -1  -2943284 34877 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

```
lm07 <- lm(y ~ x + w + x:w, data = yxzw)
anova(lm05, lm07)
```

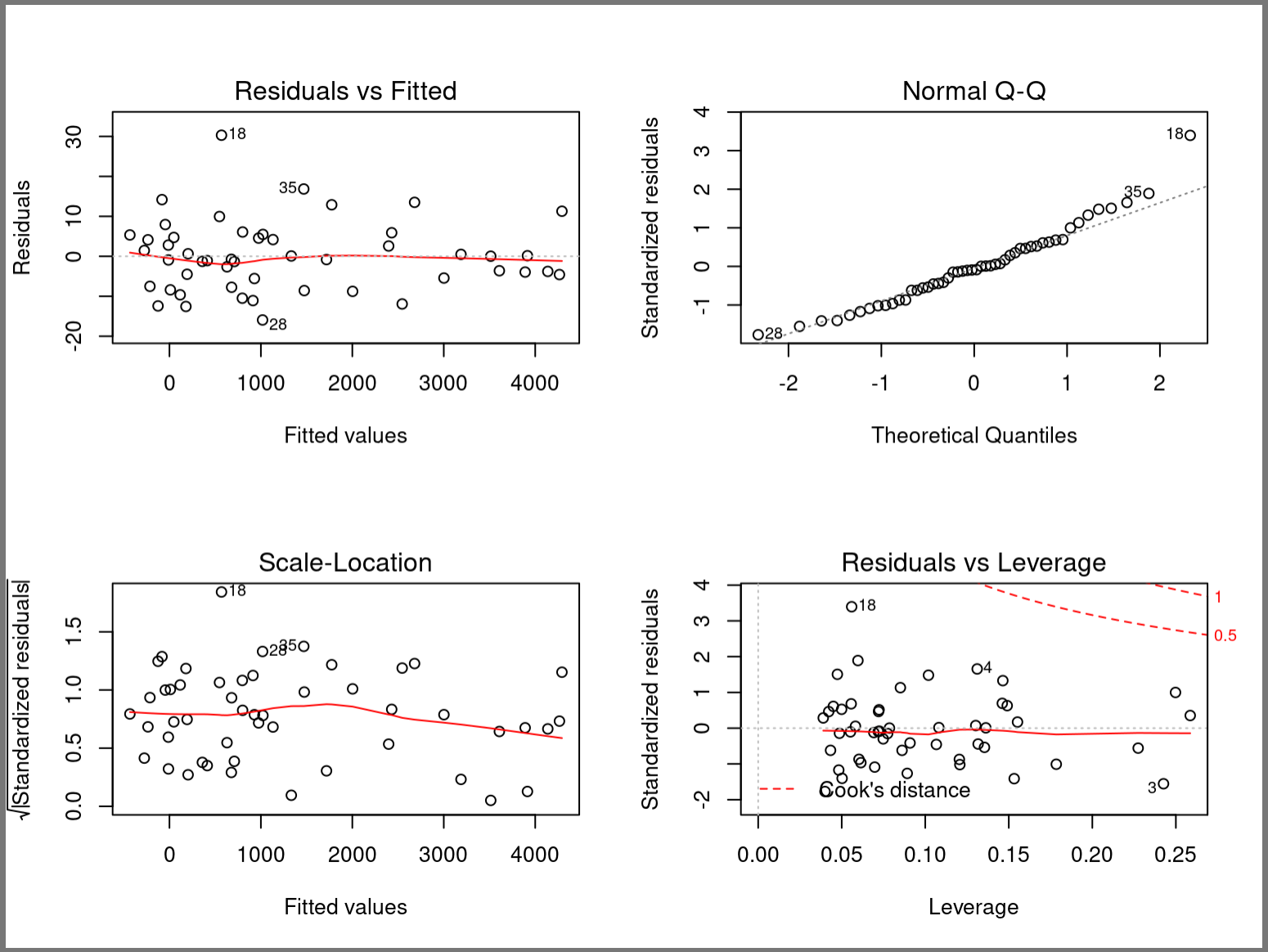
```
## Analysis of Variance Table
##
## Model 1: y ~ x + w + I(x^2) + x:w
## Model 2: y ~ x + w + x:w
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1     45    3798
## 2     46 7863171 -1  -7859373 93130 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Modelo Mínimo Adequado

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

```
##
## Call:
## lm(formula = y ~ x + w + I(x^2) + x:w, data = yxzw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9424  -5.5396  -0.7859   4.6774  30.3060
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 23.437267   7.470240    3.137    0.003 **
## x           -2.040476   0.212311   -9.611 1.78e-12 ***
## w           -5.156007   1.134704   -4.544 4.13e-05 ***
## I(x^2)       0.503330   0.001649  305.172  < 2e-16 ***
## x:w         -2.989057   0.016005 -186.753  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.186 on 45 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.843e+05 on 4 and 45 DF,  p-value: < 2.2e-16
```

DIAGNÓSTICO DO MODELO



Estimativa do Modelo

```
coef(lm05)
```

```
## (Intercept)           x           w      I(x^2)         x:w
##  23.4372672  -2.0404761  -5.1560065   0.5033303  -2.9890572
```

```
confint(lm05)
```

```
##                  2.5 %     97.5 %
## (Intercept)  8.3914314 38.4831031
## x           -2.4680920 -1.6128602
## w           -7.4414185 -2.8705945
## I(x^2)       0.5000084  0.5066522
## x:w         -3.0212937 -2.9568206
```

$$y = 23.44 - 2.04x + 0.50x^2 - 5.16w - 2.99xw$$

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)
```

```
coef(lm05)
```

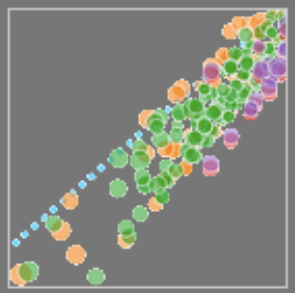
```
## (Intercept)           x           w      I(x^2)         x:w
##  23.4372672  -2.0404761  -5.1560065   0.5033303  -2.9890572
```

Problema da colinearidade

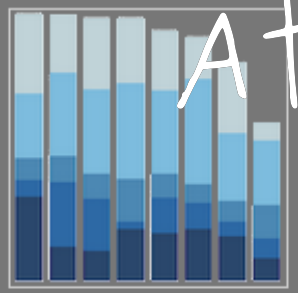
```
lm10 <- lm(y ~  w + z +  I(z^2) + z:w, data = yxzw)
summary(lm10)
```

```
##
## Call:
## lm(formula = y ~ w + z + I(z^2) + z:w, data = yxzw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -784.03 -252.24  -22.12  103.84 1594.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  169.306    367.407   0.461   0.6472
## w            -60.680     54.945  -1.104   0.2753
## z             91.107    102.293   0.891   0.3779
## I(z^2)        31.160      7.215   4.319 8.53e-05 ***
## w:z          -20.429      7.769  -2.630   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 455.7 on 45 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.894
## F-statistic: 104.3 on 4 and 45 DF,  p-value: < 2.2e-16
```

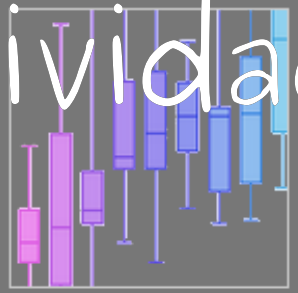
Line and Scatter Plots



Bar Charts

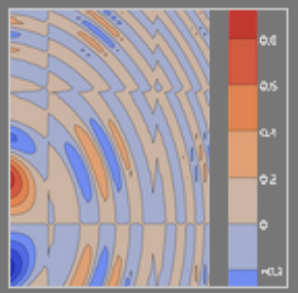


Box Plots



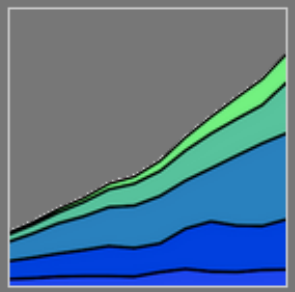
Bubble Charts

Contour Plots

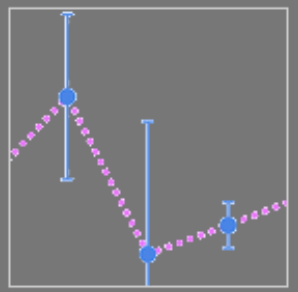


Atividade

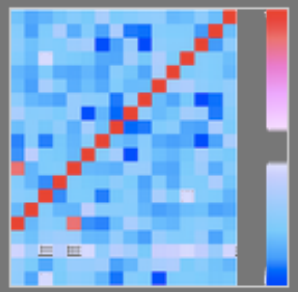
Filled Area Plots



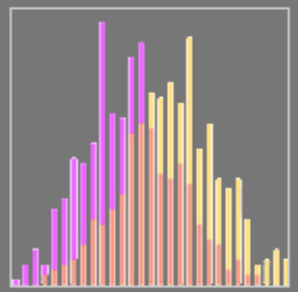
Error Bars



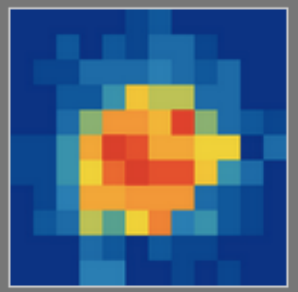
Heatmaps



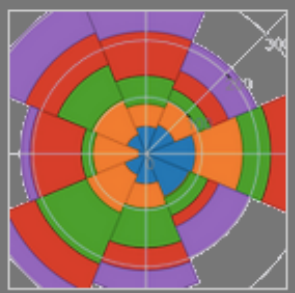
Histograms



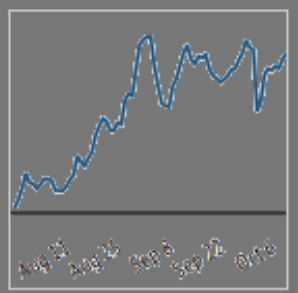
2D Histograms



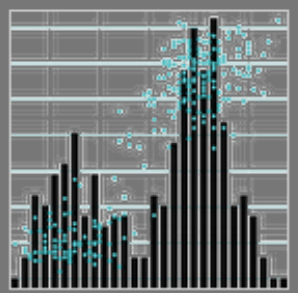
Polar Charts



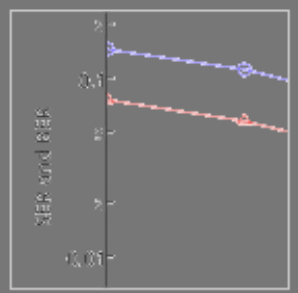
Time Series



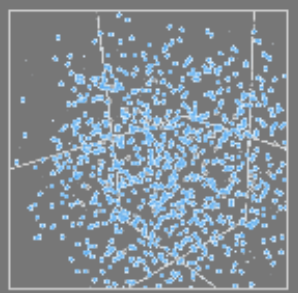
Multiple Chart Types



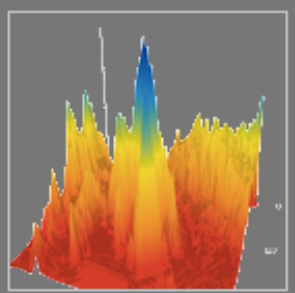
Log Plots



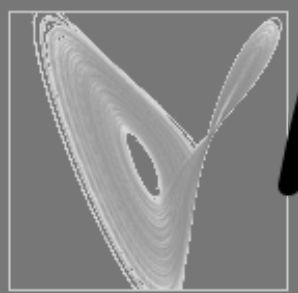
3D Scatter Plots



3D Surface Plots



3D Line Plots



PlAnEco