Modelos Lineares

unificação metodológica

Alexandre Adalardo de Oliveira

PlanECO 2017

Conceitos

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

Modelos Lineares

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	$logit(\beta_1)=1$

O modelo de regressão

$$y = \hat{\alpha} + \hat{\beta}x + \epsilon$$

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

```
set.seed(2)
(x1 = seq(1,5, by=0.5))
```

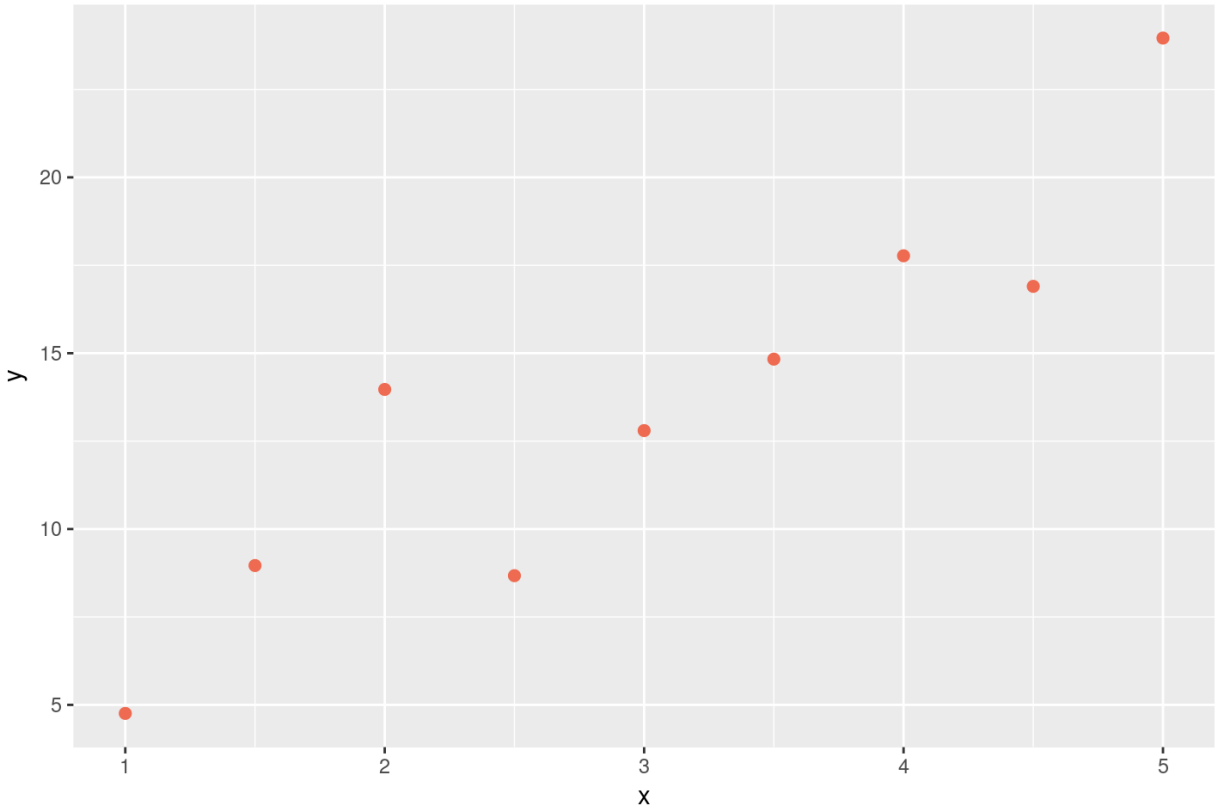
```
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

```
y1 = 4 + 3 * x1 + rnorm(n= 9, mean= 0, sd= 2.5 )
y1
```

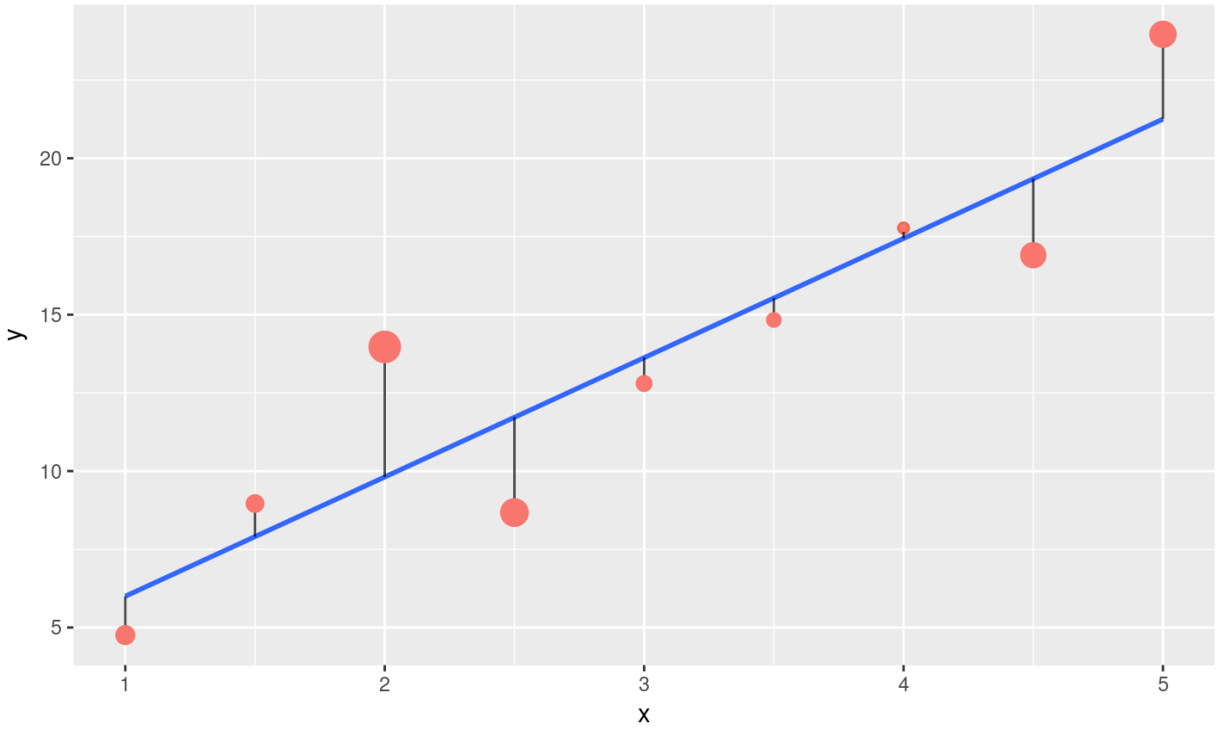
```
## [1]  4.757714  8.962123 13.969613  8.674061 12.799371 14.831051 17.769887
## [8] 16.900755 23.961185
```

4/52

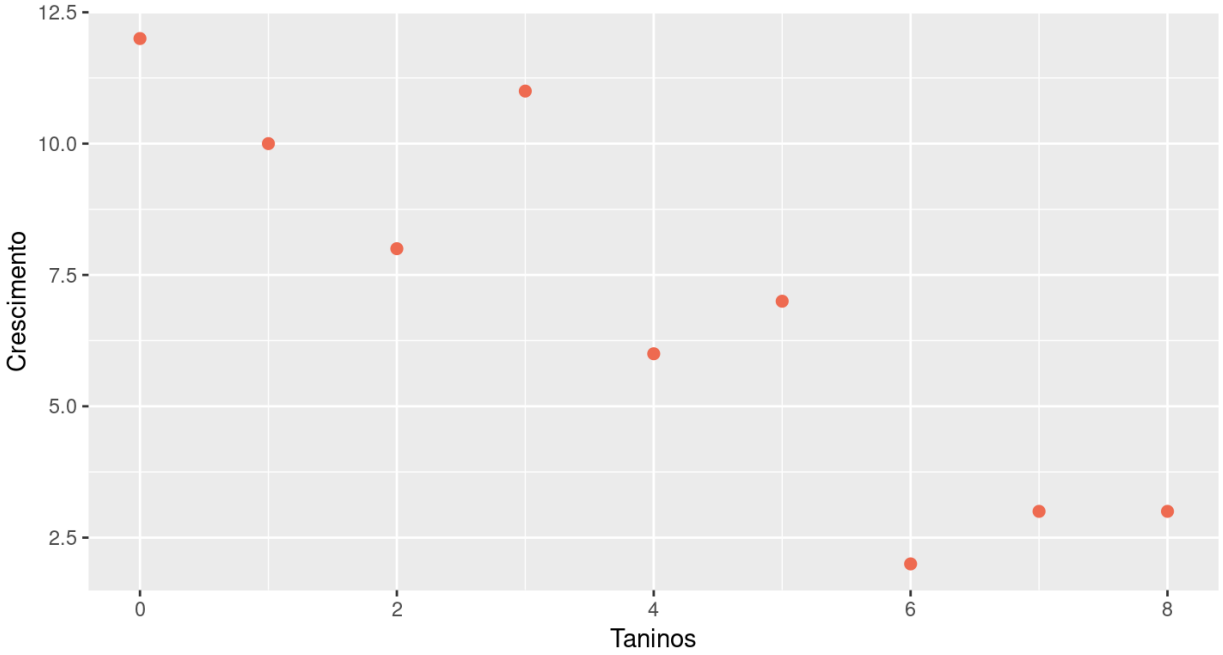
Regressão: dados simulados



Desvios quadráticos

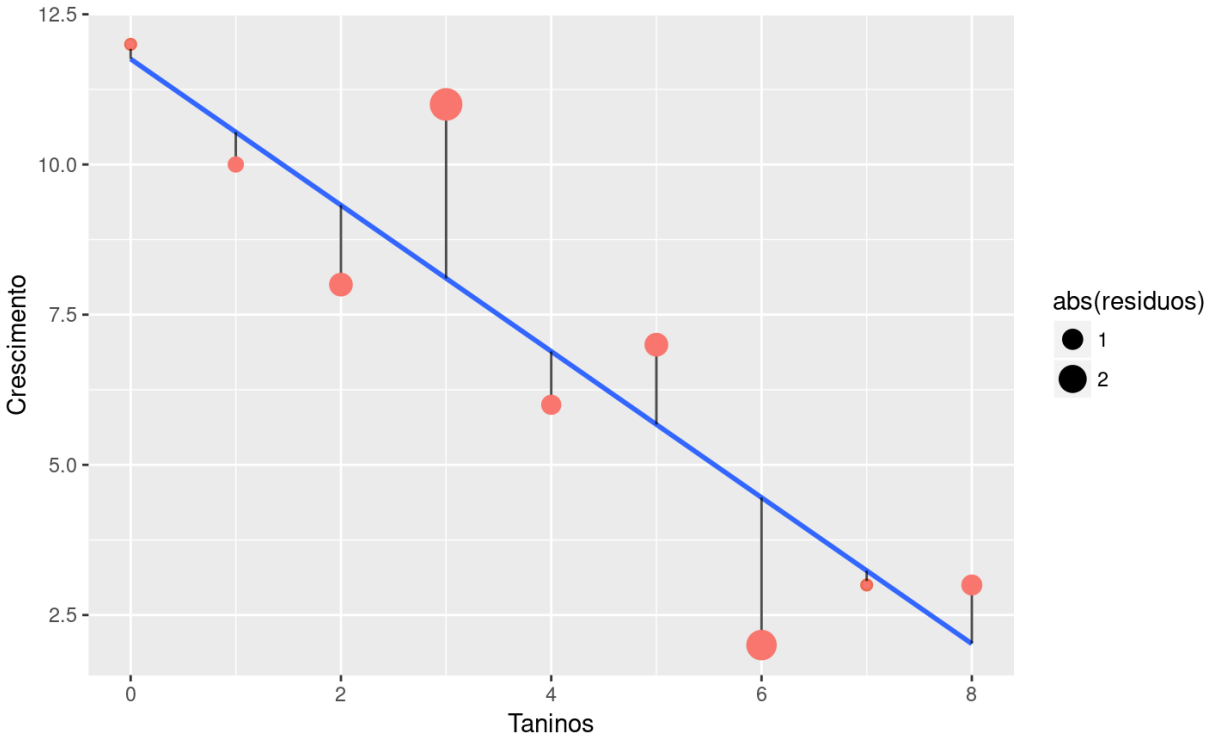


Exemplo: dieta de lagarta

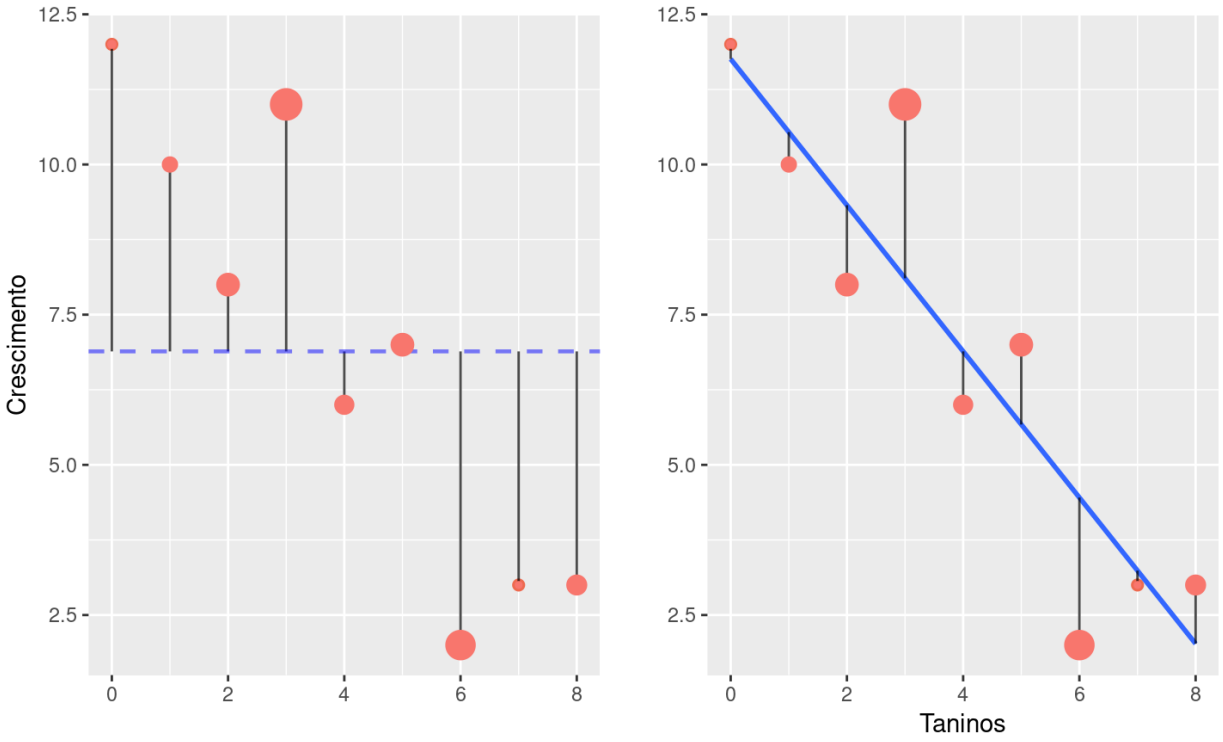


7/52

Modelo de Regressão: lagartas



Modelo Mínimo: lagarta



9/52

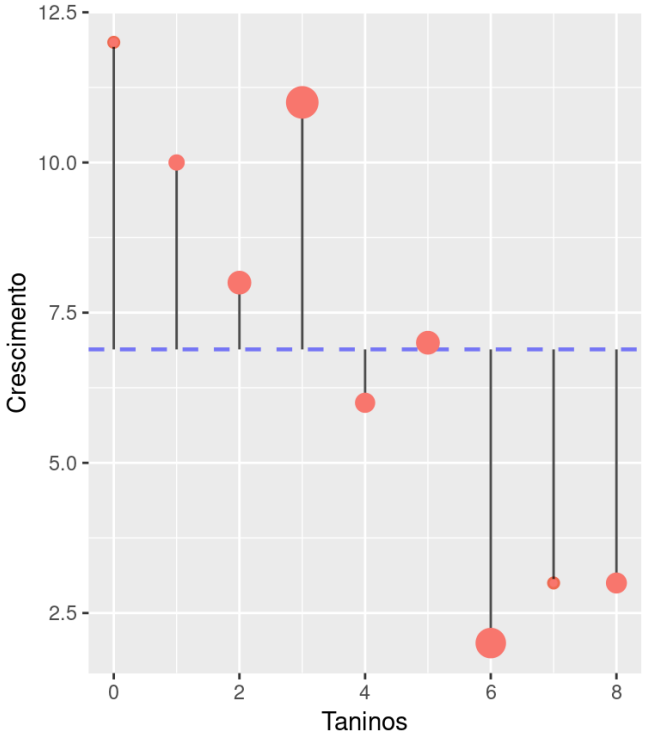
Lógica da Anova

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

Lógica da Regressão

$$SS_{total} = SS_{regr} + SS_{erro}$$

Modelo mínimo



$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

11/52

Desvios quadráticos total

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

```
(dt <- lag$growth - mean(lag$growth))
```

```
## [1]  5.1111111  3.1111111  1.1111111  4.1111111 -0.8888889  0.1111111
## [7] -4.8888889 -3.8888889 -3.8888889
```

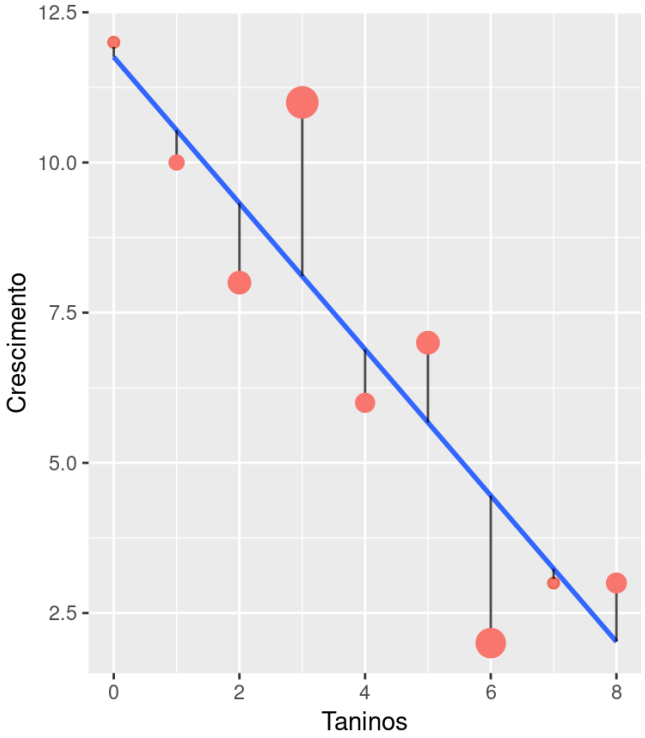
```
dt^2
```

```
## [1] 26.12345679  9.67901235  1.23456790 16.90123457  0.79012346  0.01234568
## [7] 23.90123457 15.12345679 15.12345679
```

```
(ss_total <- sum(dt^2))
```

```
## [1] 108.8889
```

Desvios quadráticos do ERRO



$$SS_{error} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

13/52

Desvios quadráticos do ERRO

$$SS_{error} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

```
(coeflag <- coef(lmlag))
```

```
## (Intercept)      tannin
##   11.755556   -1.216667
```

```
(predlag <- coeflag[1] + coeflag[2] * lag$tannin)
```

```
## [1] 11.755556 10.538889  9.322222  8.105556  6.888889  5.672222  4.455556
## [8]  3.238889  2.022222
```

```
lag$growth
```

```
## [1] 12 10  8 11  6  7  2  3  3
```

Desvios quadráticos do ERRO

$$SS_{error} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

```
(ss_erro <- sum((lag$growth - predlag)^2))
```

```
## [1] 20.07222
```

Lógica da Regressão

$$SS_{total} = SS_{regr} + SS_{erro}$$

```
(ss_reg <- ss_total - ss_erro)
```

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

Tabela de Anova

Fonte	SumSquare	GL	MeanSquare
Regressão	88.82	1	88.82
Erro	20.07	7	2.87
Total	108.89	8	

16/52

Teste de hipótese: F e r^2

```
(r2 <- ss_reg/ss_total)
```

```
## [1] 0.8156633
```

```
(flag <- ss_reg/(ss_erro/7))
```

```
## [1] 30.97398
```

```
1- pf(flag, 1, 7)
```

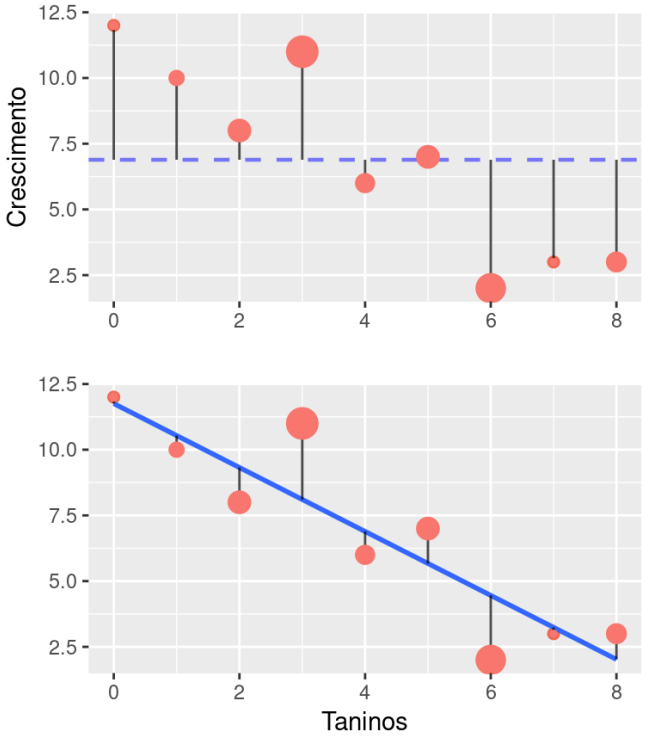
```
## [1] 0.0008460738
```

Regressão no R: lagarta

```
laglm <- lm(growth ~ tannin, data=lag)
anova(laglm)
```

```
## 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 '*' 0.05 '.' 0.1 ' ' 1
```

Comparando Modelos no R: lagarta




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

```
## Analysis of Variance Table
##
## Model 1: growth ~ 1
## Model 2: growth ~ tannin
##   Res.Df     RSS Df Sum of Sq      F    Pr(>
## 1      8 108.889
## 2      7  20.072  1    88.817 30.974 0.00084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
```

19/52

Comparando Modelos no R: lagarta

Anova do modelo: anova(laglm)

	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, laglm)

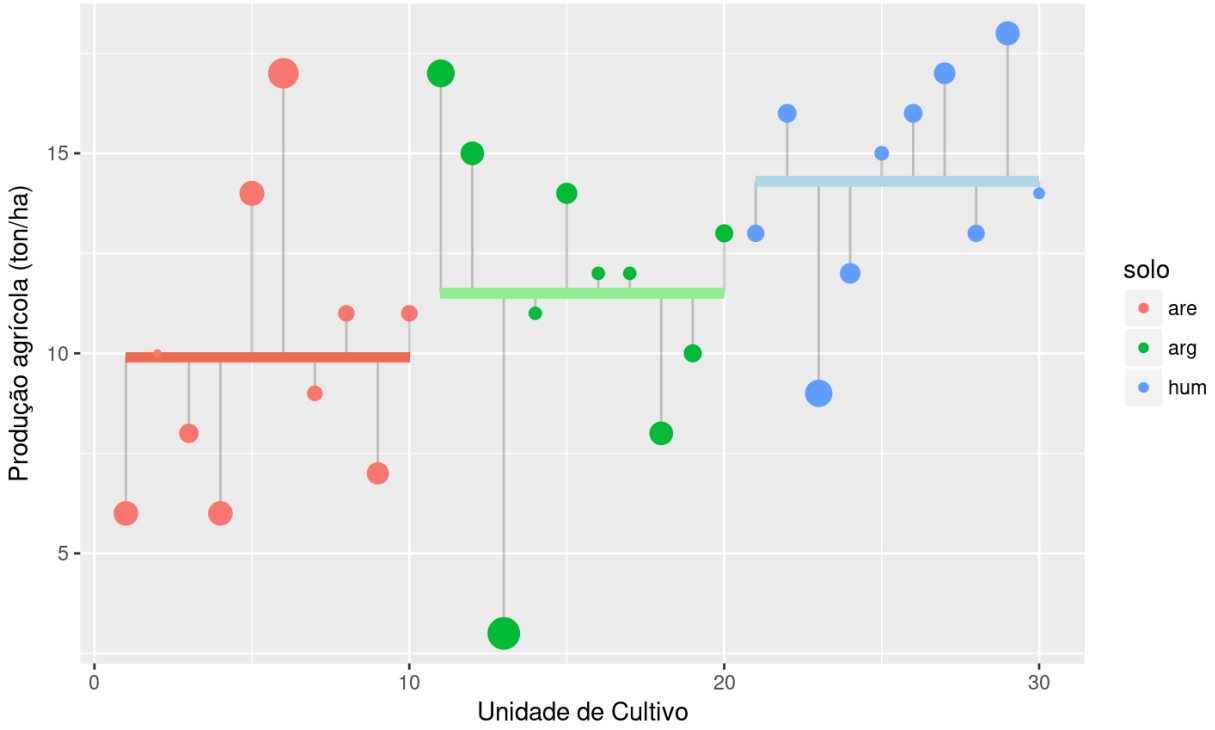
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
8	108.88889				
7	20.07222	1	88.81667	30.97398	0.0008461

NÃO DESESPERE, ESPERE! KEEP CALM!!

Atividade

Regressão de variável categórica?!

```
## 'data.frame':    30 obs. of  2 variables:
##  $ solo : chr  "are" "are" "are" "are" ...
##  $ colhe: int  6 10 8 6 14 17 9 11 7 11 ...
```



22/52

Variáveis Dummy ou Indicadoras

```
croplin <- crop[,c( "colhe", "solo")]
croplin$solo
```

```
##  [1] are are are are are are are are are are arg arg arg arg arg arg arg
## [18] arg arg arg hum hum hum hum hum hum hum hum hum hum hum
## Levels: are arg hum
```

```
croplin$arg <-0
croplin$arg[crop$solo=="arg"] <- 1
croplin$hum <- 0
croplin$hum[crop$solo=="hum"] <- 1
```

Variável Dummy ou Indicadora

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

Número de níveis do fator menos 1 (intercepto)

Modelo linear: dummy

Modelo

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

Intercepto:

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

Coeficientes:

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

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

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

Regressão dummy

```
lmdum <- lm(colhe ~ arg + hum, croplin)
summary(lmdum)
```

```
##
## 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-10 ***
## arg            1.600      1.529   1.047  0.30456
## hum            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
```

Modelo Linear Normal

```
lmCrop <- lm(colhe~solo, data = crop)
summary(lmCrop)
```

```
##
## 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
```

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
```

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

Modelo

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

ercepto:

–

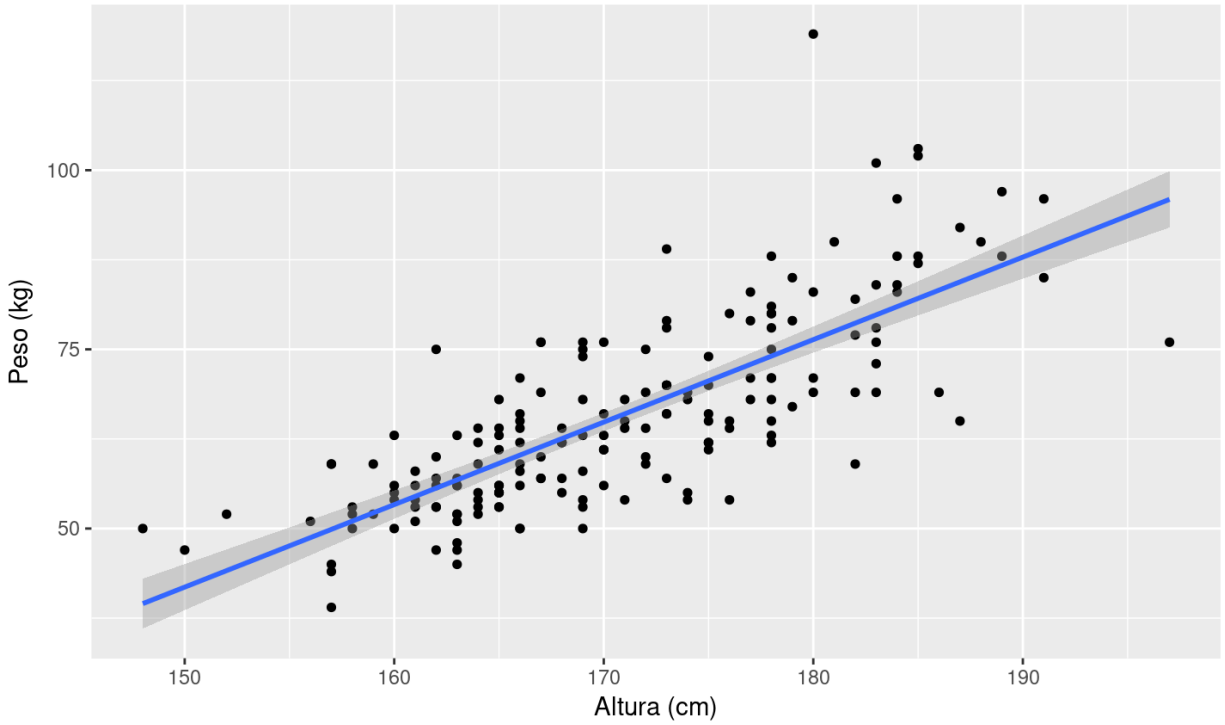
Atividade

Retomando a regressão

Davis (peso ~ altura)

```
## 'data.frame':    180 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

Gráfico da Regressão: peso ~ altura



lmdavis <- lm(weight~height, data = Davis)

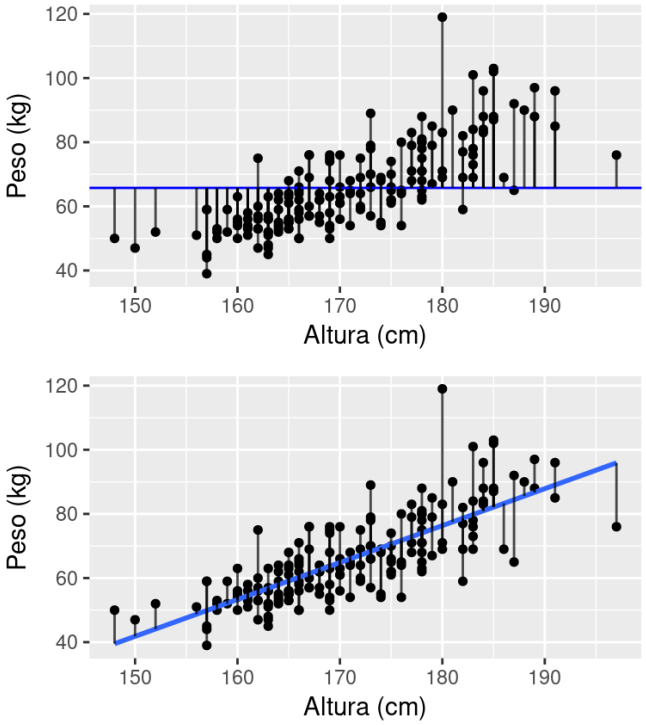
Modelo da Regressão

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
height	1	19095.04	19095.04	256.08	0
Residuals	178	13272.71	74.57		

```
##
## 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
##
```

32/52

Modelo da Regressão: peso ~ altura



anova(davisNull,lmdavis)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
179	32367.75				
178	13272.71	1	19095.04	256.0832	0

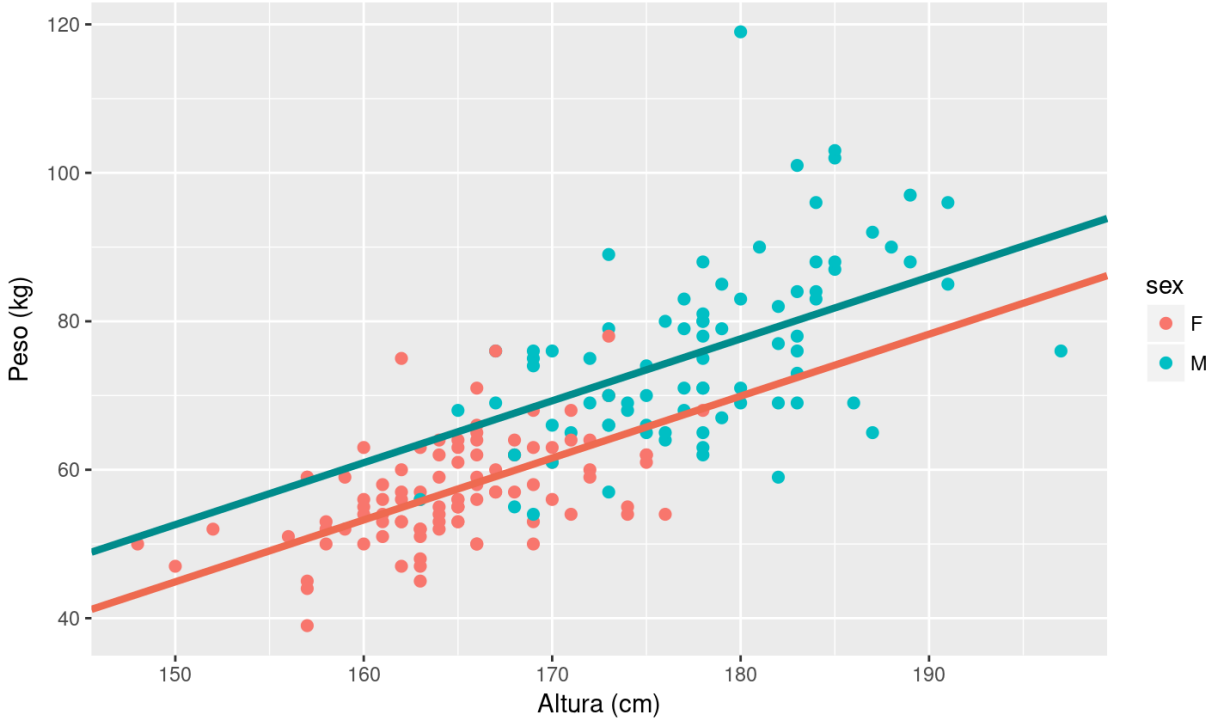
$p_{valor} = 2.2e-16$

$p_{valor} = 2.2 * 10^{-16}$

$r^2 = 0.587$

Modelo de Regressão:

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



34/52

sexo: variável dummy com dois níveis (mulher = 0, homem = 1)

```
lmdavis01 <- lm(weight~ height + sex, data = Davis)
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
```

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

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

Mulher ($sex = 0$)

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

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

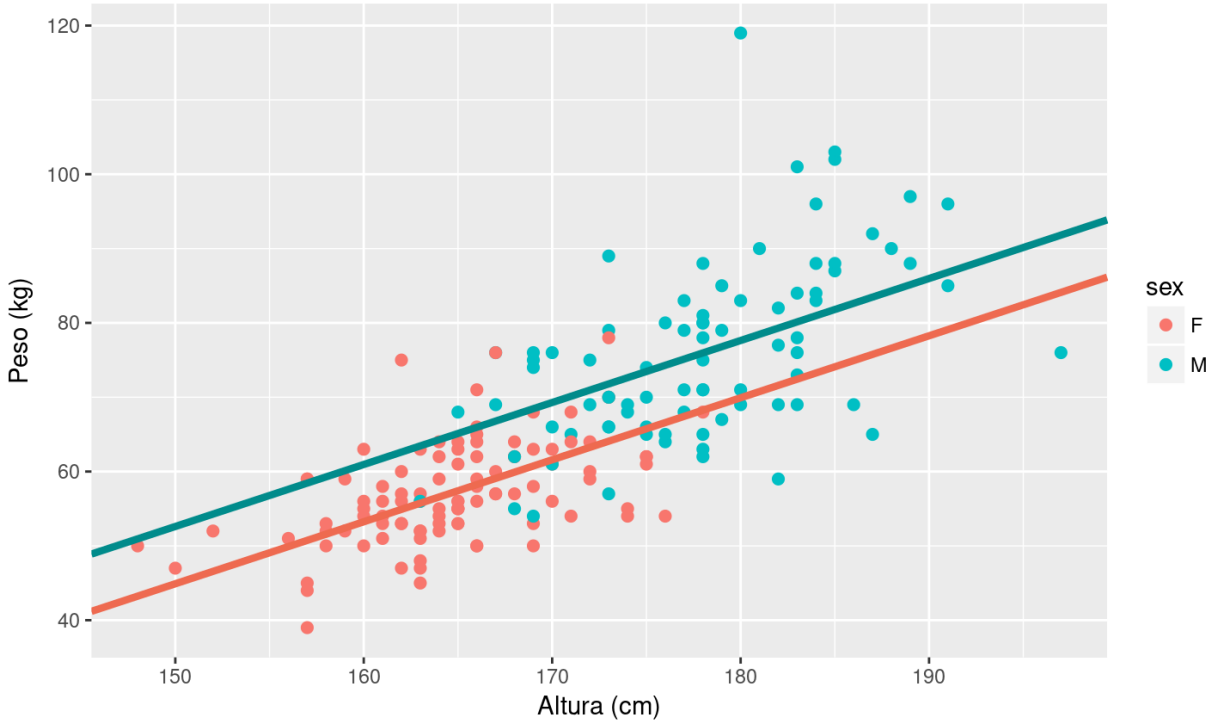
Homem ($sex = 1$)

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

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

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

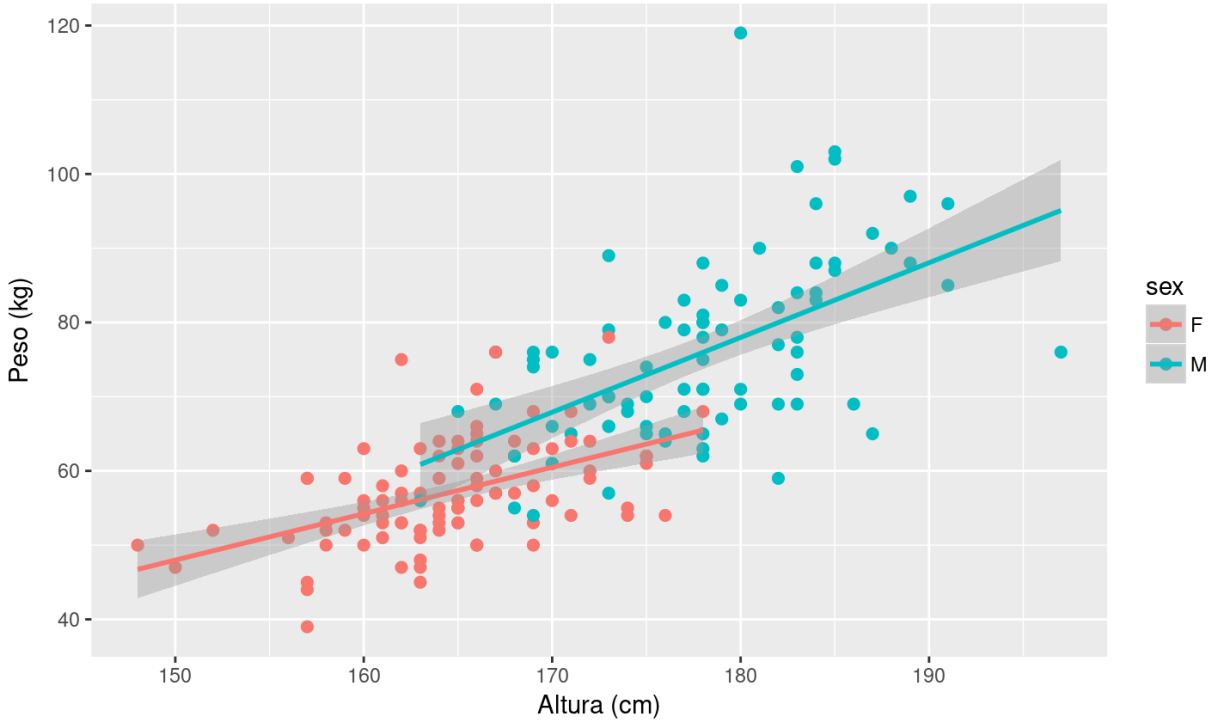
```
gg06 +
  geom_abline(slope= coeflm01[2], intercept = coeflm01[1], col="coral2", lwd=1.5 )+
  geom_abline(slope= coeflm01[2], intercept = coeflm01[1]+ coeflm01[3], col="cyan4", lwd=1.
```



37/52

Interação

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



38/52

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

```
##
## Call:
## lm(formula = weight ~ height + 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
```

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

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

Mulher ($sex = 0$)

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

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

Homem ($sex = 1$)

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

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

Predição do modelo

Uma mulher de 160cm de altura

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

$$sex = 0$$

```
(coefull <- coef(lmdavisfull))
```

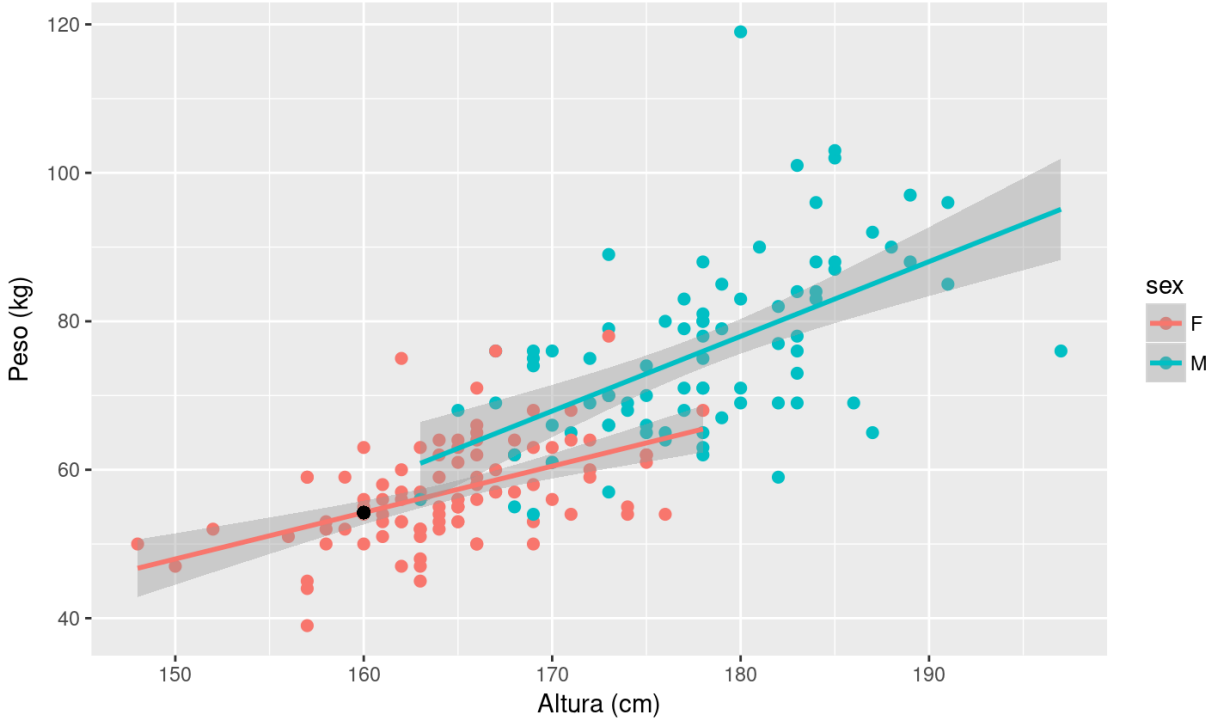
```
## (Intercept)      height        sexM height:sexM
## -45.7988220   0.6252035 -57.4326307   0.3815088
```

```
predMulher <- coefull[1] + coefull[2] * 160
(predMulher <- as.numeric(predMulher))
```

```
## [1] 54.23373
```

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

- Uma mulher com 160cm de altura tem peso 54.23 kg.



Predito do Modelo

Homem com 180cm

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

```
coefull
```

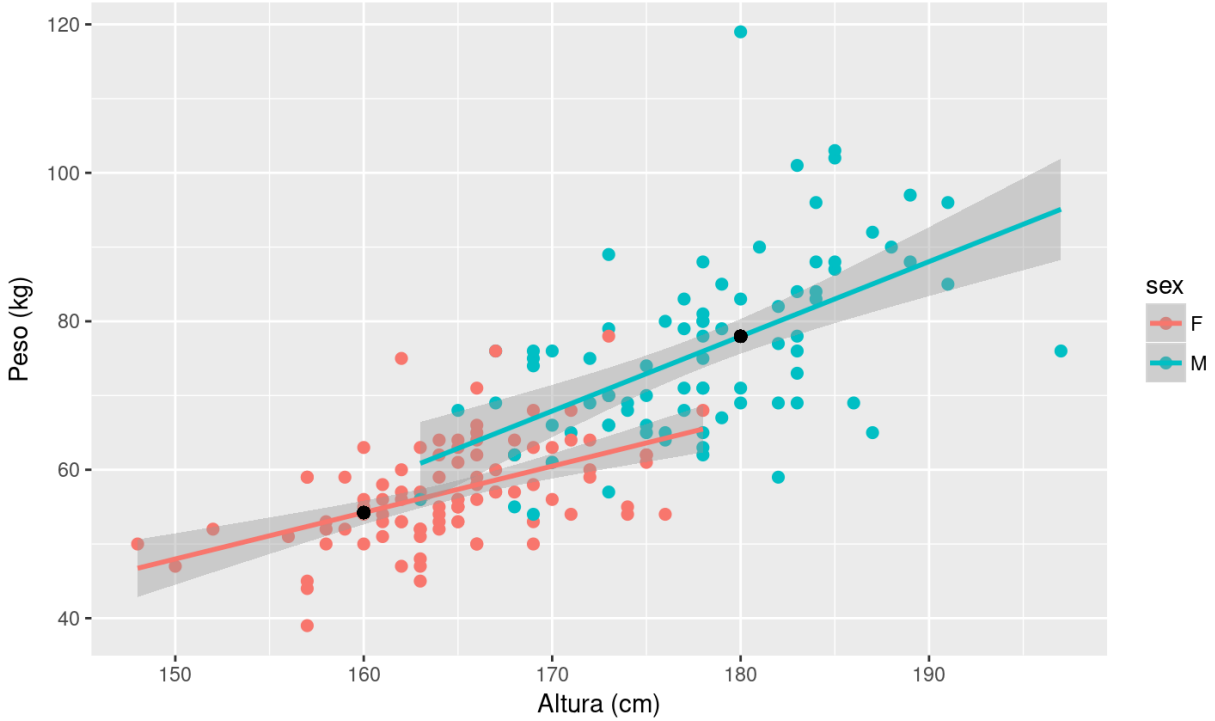
```
## (Intercept)      height        sexM height:sexM
## -45.7988220   0.6252035 -57.4326307   0.3815088
```

```
predHomem <- (coefull[1]+ coefull[3]) + (coefull[2]+ coefull[4]) * 180
(predHomem <- as.numeric(predHomem))
```

```
## [1] 77.97676
```

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

- Um homem com 180cm de altura tem peso 77.98 kg.



Qual o melhor modelo

Princípio da parcimônia (Navalha de Occam)

- devem ter menos parâmetros possível
- linear é melhor que não-linear
- reter menos pressupostos
- simplificado ao mínimo adequado
- explicações mais simples são preferíveis

Simplificação do modelo

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 significativoa -> 0

Simplificação do modelo: continuação

- **Compare o modelo anterior com o simplificado**
- A diferença não é significativa:
  - retenha o modelo mais simples
  - continue simplificando
- A difereça é significativa
  - retanha o modelo complexo
  - este é o modelo **MINÍMO ADEQUADO**

Simplificando Modelo: exemplo

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

```
## Analysis of Variance Table
##
## Model 1: weight ~ height + 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 ' ' 1
```

Simplificando Modelo: exemplo

```
anova(lmdavis01, lmdavis)
```

```
## 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 ' ' 1
```

Modelo Mínimo Adequado

```
coeflm01
```

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

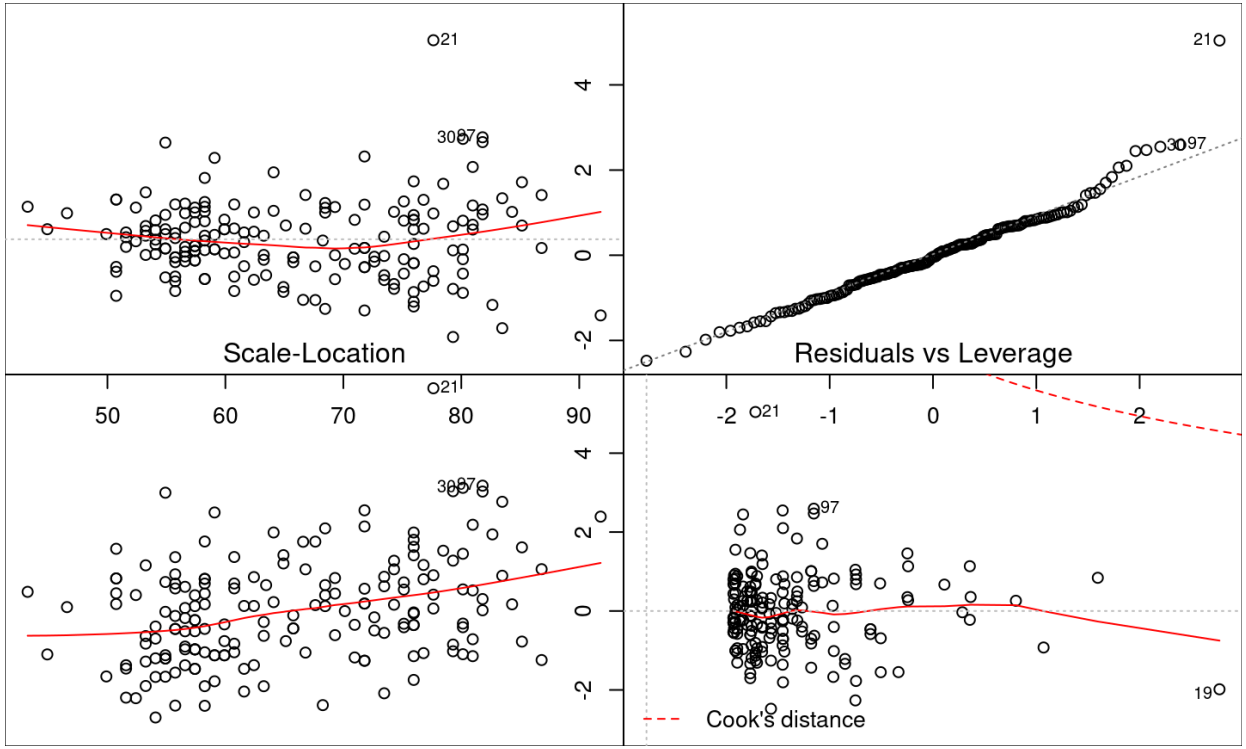
```
confint(lmdavis01)
```

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

50/52

Diagnóstico do Modelo: plot(modelo)

```
oldpar <- par(mfrow=c(2,2), mar=c(0,0,0,0))
plot(lmdavis01)
```



51/52

```
par(oldpar)
```

Atividade