ME613 - Análise de Regressão

Parte 9

Benilton S Carvalho & Rafael P Maia - 2S2020

Modelo de Regressão Polinomial

Introdução

Podemos considerar funções polinomiais como um caso particular do modelo de regressão linear já visto.

Por exemplo, quando temos uma única variável preditora, podemos escrever Y como um função polinomial de X de grau q.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_2 X^q + \varepsilon$$

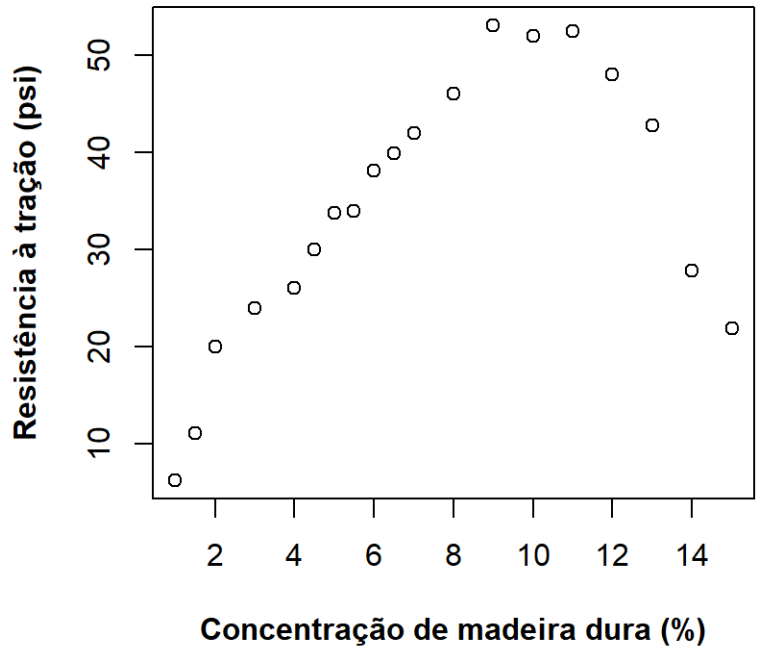
Regressão Polinomial - Multicolinearidade

Em regressão polinomial a matriz desenho do modelo $\mathbf{X}$ é formada por colunas que são potências de uma mesma variável;

Essa construção implica no mau condicionamento da matriz, ou seja, na quase dependência linear de suas colunas;

Uma medida corretiva consiste em usar os dados centrados (ou seja, substituir X por $X_* = X - \bar{X}$).

Exemplo : Concentração de madeira dura em celulose e força de tração do papel Kraft



Padronizando a variável X

Correlação entre X, X^2, X^3, X^4

```
##        x    x2    x3    x4
## x  1.000 0.970 0.921 0.874
## x2 0.970 1.000 0.987 0.961
## x3 0.921 0.987 1.000 0.993
## x4 0.874 0.961 0.993 1.000
```

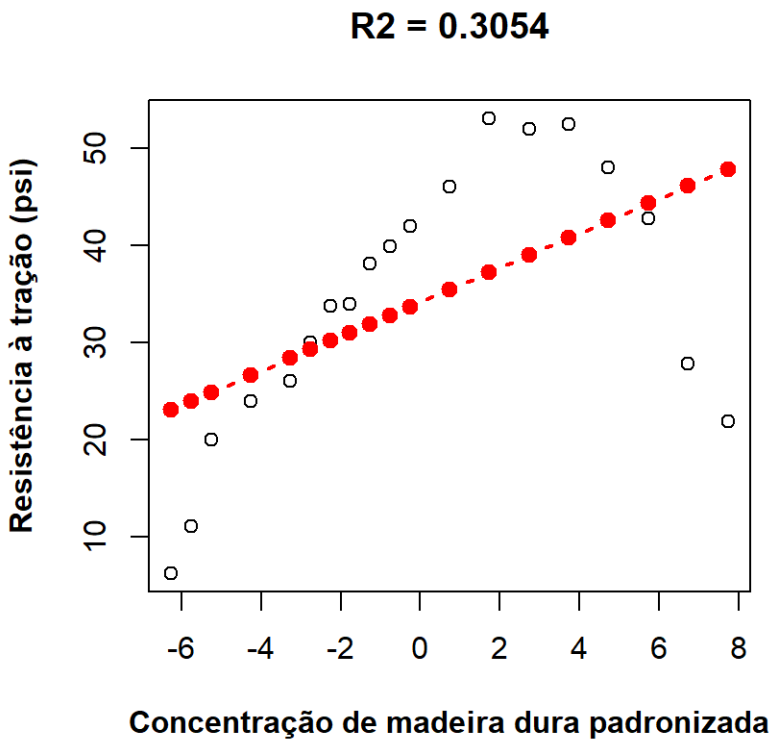
Seja $X_* = X - \bar{X}$

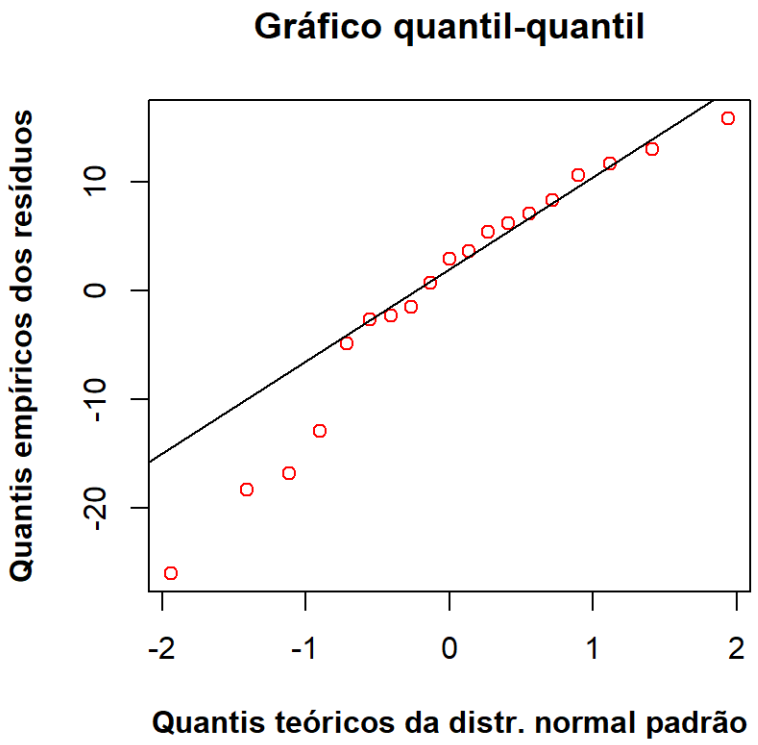
Correlação entre X_*, X_*^2, X_*^3, X_*^4

```
##        x_c  x_c2  x_c3  x_c4
## x_c  1.000 0.297 0.910 0.399
## x_c2 0.297 1.000 0.424 0.948
## x_c3 0.910 0.424 1.000 0.574
## x_c4 0.399 0.948 0.574 1.000
```

Ajuste regressão linear simples

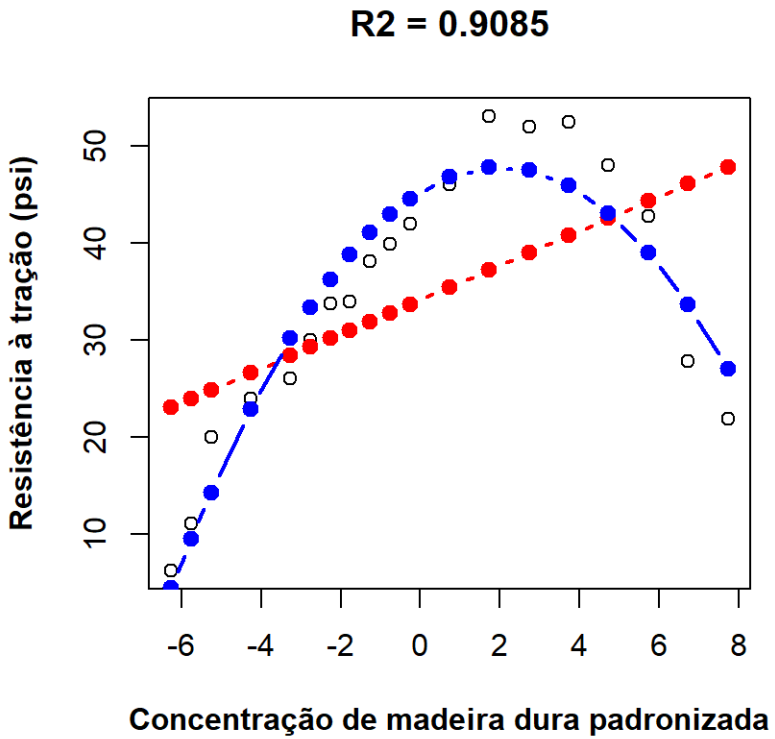
```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 34.18421    2.71075 12.61061  0.00000
## x_c          1.77099    0.64781  2.73379  0.01414
```

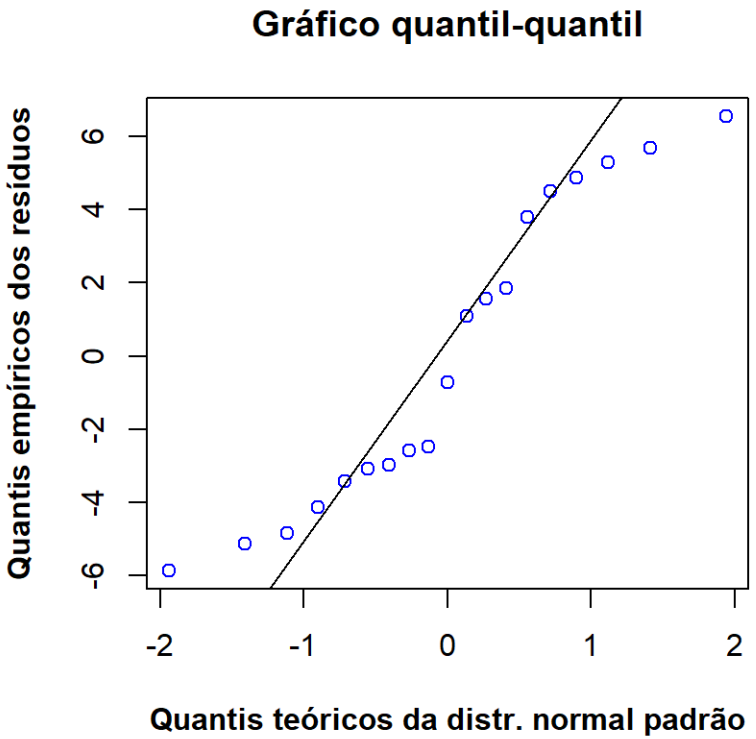




Ajuste polinômio grau 2

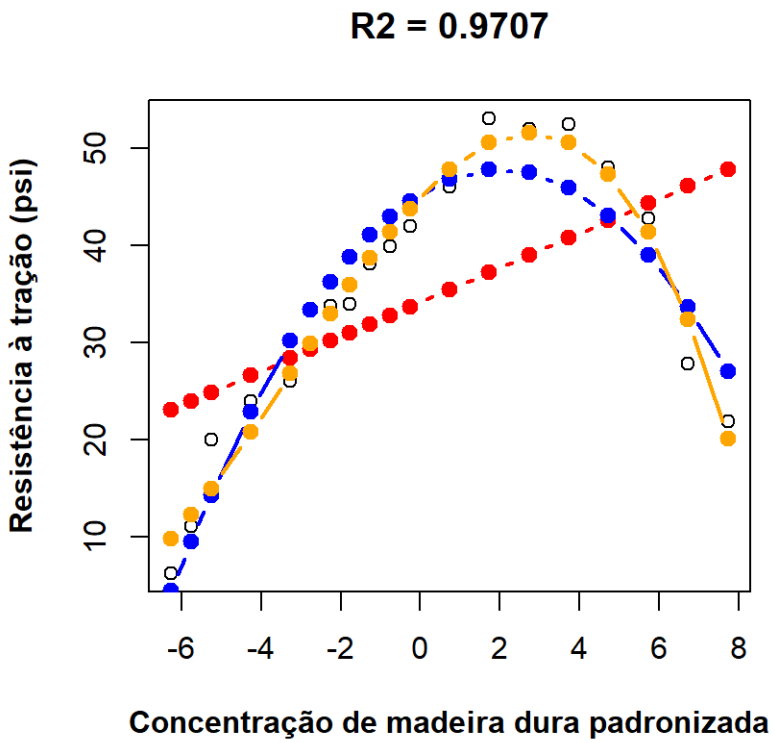
```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 45.294973   1.482873  30.54542        0
## x_c          2.546344   0.253839  10.03134        0
## x_c2        -0.634549   0.061788 -10.26973        0
```

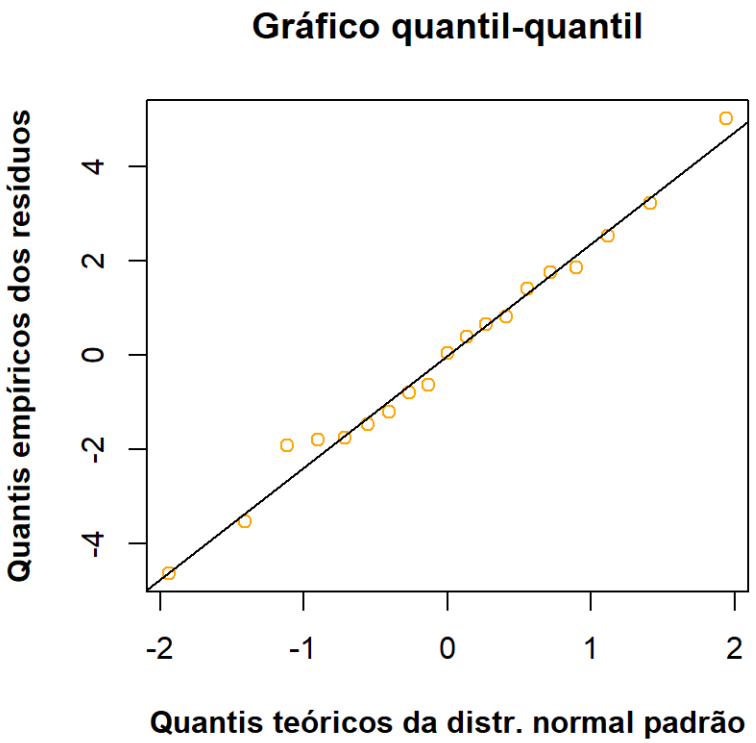




Ajuste polinômio grau 3

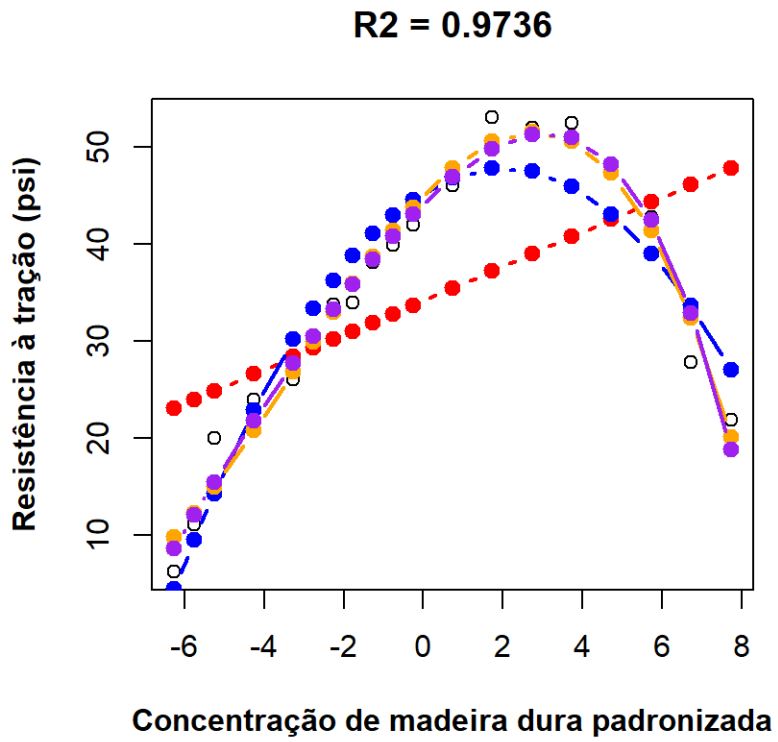
```
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 44.97556    0.86903  51.75362    0e+00
## x_c          4.33939    0.35098  12.36373    0e+00
## x_c2        -0.54887    0.03920 -14.00210    0e+00
## x_c3        -0.05519    0.00979  -5.63781    5e-05
```

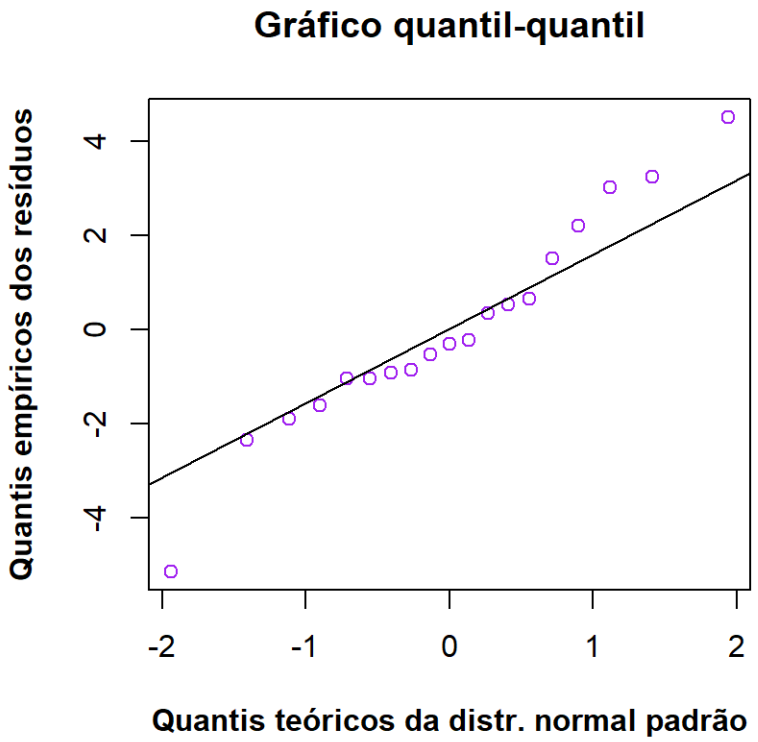




Ajuste polinômio grau 4

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 44.16512    1.07282 41.16746  0.00000
## x_c          4.11527    0.38874 10.58616  0.00000
## x_c2        -0.39417    0.12993 -3.03364  0.00894
## x_c3        -0.04527    0.01248 -3.62759  0.00274
## x_c4        -0.00351    0.00281 -1.24665  0.23298
```





Soma extra de quadrados e teste F

```
anova(m4)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x_c        1 1043.43 1043.43 161.9124 4.392e-09 ***
## x_c2       1 2060.82 2060.82 319.7849 4.874e-11 ***
## x_c3       1  212.40  212.40  32.9591 5.104e-05 ***
## x_c4       1   10.02   10.02   1.5541     0.233
## Residuals 14   90.22    6.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
round(summary(m3)$coefficient,5)
```

```
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 44.97556    0.86903  51.75362    0e+00
## x_c          4.33939    0.35098  12.36373    0e+00
## x_c2        -0.54887    0.03920 -14.00210    0e+00
## x_c3        -0.05519    0.00979  -5.63781    5e-05
```

Extrapolação

A extrapolação de modelos polinomiais para valores de x fora do intervalo observado nos dados pode ser extremamente perigosa.

Modelo com dois preditores - segunda ordem

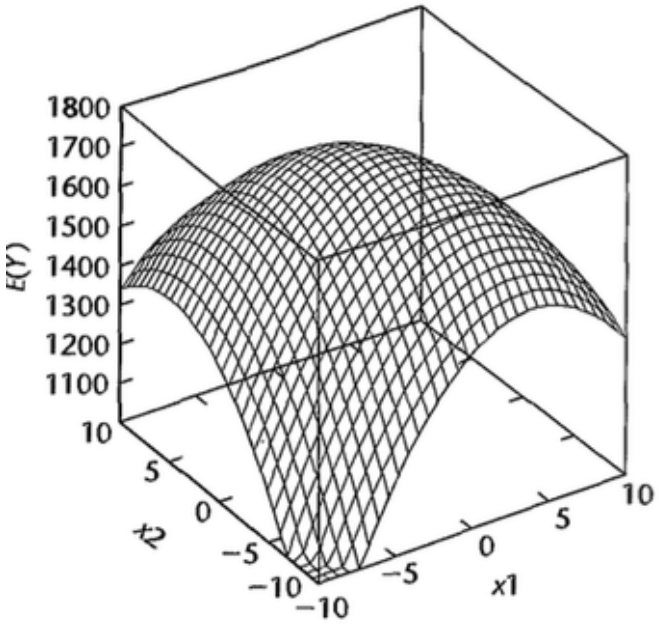
Vamos considerar um modelo de segunda ordem com duas variáveis preditoras X_1 e X_2

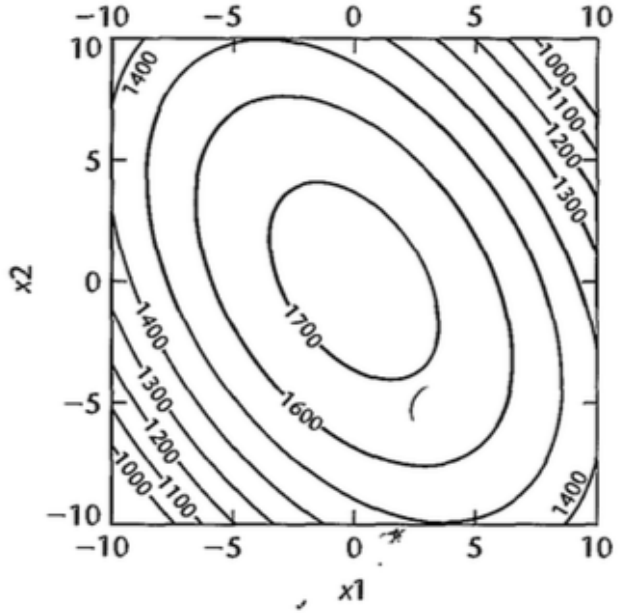
$$Y = \beta_0 + \beta_1 X_{1*} + \beta_2 X_{1*}^2 + \beta_3 X_{2*} + \beta_4 X_{2*}^2 + \beta_5 X_{1*} X_{2*} + \varepsilon$$

em que $X_{1*} = X_1 - \bar{X}_1$ e $X_{2*} = X_2 - \bar{X}_2$.

Exemplo

$$E(Y) = \beta_0 + \beta_1 X_{1*} + \beta_2 X_{1*}^2 + \beta_3 X_{2*} + \beta_4 X_{2*}^2 + \beta_5 X_{1*} X_{2*}$$





Método hierárquico de ajuste de modelo

Pode-se começar com um modelo de segunda ou terceira ordem e ir testando se os coeficientes de ordem maiores são significativos.

Por exemplo:

$$Y = \beta_0 + \beta_1 X_* + \beta_2 X_*^2 + \beta_3 X_*^3 + \varepsilon$$

Para testar se $\beta_3 = 0$ podemos utilizar $SQReg(X_*^3 \mid X_*, X_*^2)$. Se quisermos testar se $\beta_2 = \beta_3 = 0$:

$$SQReg(X_*^2, X_*^3 \mid X_*) = SQReg(X_*^2 \mid X_*) + SQReg(X_*^3 \mid X_*, X_*^2)$$

Se um termo de ordem mais alta é mantido no modelo, os de ordem mais baixa devem obrigatoriamente ser mantidos também.

Exemplo

Dados de um experimento realizado para estudar o efeito de duas variáveis, temperatura de reação (T) e concentração de reagente (C), na conversão percentual de um processo químico (Y)

```
##           T     C  Y       T_p       C_p
##  [1,] 200.00 15.00 43 -1.172692 -1.172692
##  [2,] 250.00 15.00 78  1.172692 -1.172692
##  [3,] 200.00 25.00 69 -1.172692  1.172692
##  [4,] 250.00 25.00 73  1.172692  1.172692
##  [5,] 189.65 20.00 48 -1.658187  0.000000
##  [6,] 260.35 20.00 76  1.658187  0.000000
##  [7,] 225.00 12.93 65  0.000000 -1.658187
##  [8,] 225.00 27.07 74  0.000000  1.658187
##  [9,] 225.00 20.00 76  0.000000  0.000000
## [10,] 225.00 20.00 79  0.000000  0.000000
## [11,] 225.00 20.00 83  0.000000  0.000000
## [12,] 225.00 20.00 81  0.000000  0.000000
```

Exemplo

$$Y = \beta_0 + \beta_1 X_{1*} + \beta_2 X_{2*} + \beta_3 X_{1*}^2 + \beta_4 X_{2*}^2 + \beta_5 X_{1*} X_{2*} + \varepsilon$$

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  79.749999  1.2135308 65.717328 8.349138e-10
## C_p           3.595475  0.7317866  4.913284 2.674827e-03
## T_p           8.378568  0.7317866 11.449469 2.664055e-05
## I(C_p^2)     -3.727066  0.6977733 -5.341372 1.758987e-03
## I(T_p^2)     -6.454751  0.6977733 -9.250499 9.015512e-05
## I(C_p * T_p) -5.635513  0.8824346 -6.386323 6.935235e-04
```

```
## Analysis of Variance Table
##
## Response: Y
##              Df Sum Sq Mean Sq F value    Pr(>F)
## C_p           1 142.20  142.20  24.140 0.0026748 **
## T_p           1 772.20  772.20 131.090 2.664e-05 ***
## I(C_p^2)      1  74.85   74.85  12.706 0.0118622 *
## I(T_p^2)      1 504.07  504.07  85.572 9.016e-05 ***
## I(C_p * T_p)  1 240.25  240.25  40.785 0.0006935 ***
## Residuals     6  35.34    5.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Será que um modelo de primeira ordem já seria suficiente?

- H_0: $\beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$.
- H_1: pelo menos um $\beta_q, \ldots, \beta_{p-1}$ não é zero.

(por conveniência, a notação assume que os últimos $p - q$ coeficientes do modelo serão testados)

Estatística do teste:

$$F^* \underset{\text{sob } H_0}{\sim} \frac{SQReg(X_q, \ldots, X_{p-1} \mid X_1, \ldots, X_{q-1})}{p-q} \div \frac{SQE(X_1, \ldots, X_{p-1})}{n-p} \sim F_{p-q,n-p}$$

Exemplo

$p = 6$; $n = 12$ e $q = 3$

$$F^* = \frac{SQReg(X_{1*}^2, X_{2*}^2, X_{1*}X_{2*} \mid X_{1*}, X_{2*})/3}{SQE(X_{1*}, X_{2*}, X_{1*}^2, X_{2*}^2, X_{1*}X_{2*})/13} \underset{\sim}{\text{sob}} H_0 \; F_{3,13}$$

$$\begin{aligned} SQReg(X_{1*}^2, X_{2*}^2, X_{1*}X_{2*} \mid X_{1*}, X_{2*}) &= SQReg(X_{1*}^2 \mid X_{1*}, X_{2*}) \\ &+ SQReg(X_{2*}^2 \mid X_{1*}, X_{2*}, X_{1*}^2) \\ &+ SQReg(X_{1*}X_{2*} \mid X_{1*}, X_{2*}, X_{1*}^2, X_{2*}^2) \\ &= 74.8 + 504.1 + 240.25 \\ &= 819.15 \end{aligned}$$

$$F_{obs} = \frac{819.15/3}{5.9} = 46.279661$$

Comparando com $F(0.95; 3, 13) = 3.41$. Portanto, encontramos evidências contra a hipótese nula.

Exemplo

```
modeloreduz <- lm(Y ~ C_p + T_p)
anova(modeloreduz, modelo)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ C_p + T_p
## Model 2: Y ~ C_p + T_p + I(C_p^2) + I(T_p^2) + I(C_p * T_p)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1      9 854.51
## 2      6  35.34  3    819.17 46.354 0.0001524 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo de Regressão com Interação

Efeitos de interação

Um modelo de regressão com $p-1$ variáveis preditoras com efeitos aditivos tem função de regressão da forma:

$$E(Y) = f_1(X_1) + f_2(X_2) + \ldots + f_{p-1}(X_{p-1})$$

em que $f_1, f_2, \ldots, f_{p-1}$ podem ser quaisquer funções.

Por exemplo:

$$E(Y) = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f_1(X_1)} + \underbrace{\beta_3 X_2}_{f_2(X_2)}$$

O efeito de X_1 e X_2 em Y é **aditivo**.

Efeitos de interação

Já no exemplo a seguir, o efeito não é aditivo, há efeito de interação:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_1 X_2$$

Outro exemplo:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$$

O efeito de uma variável sobre Y irá depender do nível da variável com a qual ela interage.

Interpretação: interação e efeitos lineares

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Suponha que $X_1 = a$:

$$E(Y) = \beta_0 + \beta_1 a + \beta_2 X_2 + \beta_3 a X_2$$

Suponha que $X_1 = a + 1$:

$$E(Y) = \beta_0 + \beta_1 (a+1) + \beta_2 X_2 + \beta_3 (a+1) X_2$$

Diferença no valor esperado de Y quando aumentamos X_1 em 1 unidade:

$$\beta_0 + \beta_1 (a+1) + \beta_2 X_2 + \beta_3 (a+1) X_2 - (\beta_0 + \beta_1 a + \beta_2 X_2 + \beta_3 a X_2)$$

$$= \beta_1 + \beta_3 X_2$$

Interpretação: interação e efeitos lineares

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Suponha que $X_2 = a$:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 a + \beta_3 X_1 a$$

Suponha que $X_2 = a + 1$:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 (a + 1) + \beta_3 X_1 (a + 1)$$

Diferença no valor esperado de Y quando aumentamos X_2 em 1 unidade:

$$\beta_0 + \beta_1 X_1 + \beta_2 (a + 1) + \beta_3 X_1 (a + 1) - (\beta_0 + \beta_1 X_1 + \beta_2 a + \beta_3 X_1 a)$$

$$= \beta_2 + \beta_3 X_1$$

Interpretação: interação e efeitos lineares

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Diferença no valor esperado de Y quando aumentamos X_1 em 1 unidade:

$$\frac{\partial E(Y)}{\partial X_1} = \beta_1 + \beta_3 X_2$$

Diferença no valor esperado de Y quando aumentamos X_2 em 1 unidade:

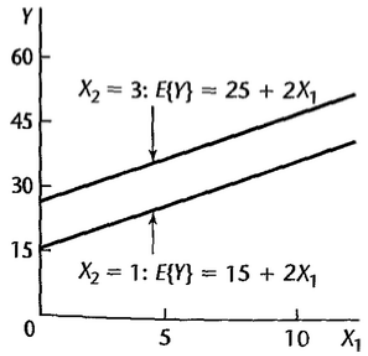
$$\frac{\partial E(Y)}{\partial X_2} = \beta_2 + \beta_3 X_1$$

Interpretação: interação e efeitos lineares

Modelo aditivo:

$$E(Y) = 10 + 2X_1 + 5X_2$$

β_1: mudança no valor esperado de Y quando X_1 aumenta em 1 unidade, mantendo X_2 constante.



Mantendo X_2 constante: não importa se $X_2 = 1$ ou $X_2 = 3$ o efeito é sempre β_1 no valor esperado quando X_1 aumenta em 1 unidade (retas paralelas).

Interpretação: interação e efeitos lineares

Modelo com interação:

$$E(Y) = 10 + 2X_1 + 5X_2 + 0.5X_1X_2$$

Se $X_2 = 1$:

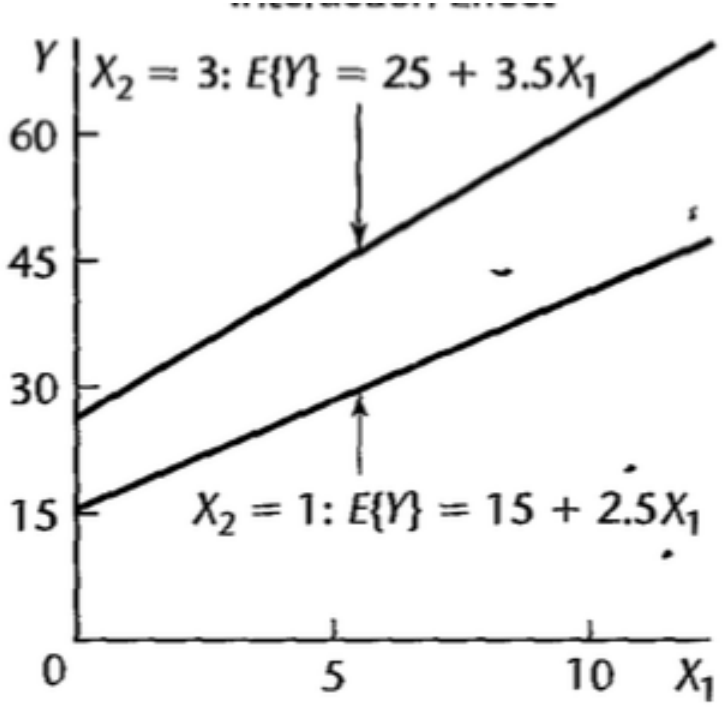
$$E(Y) = 10 + 2X_1 + 5 \times 1 + 0.5X_1 \times 1 = 15 + 2.5X_1$$

Se $X_2 = 3$:

$$E(Y) = 10 + 2X_1 + 5 \times 3 + 0.5X_1 \times 3 = 25 + 3.5X_1$$

Interpretação: interação e efeitos lineares

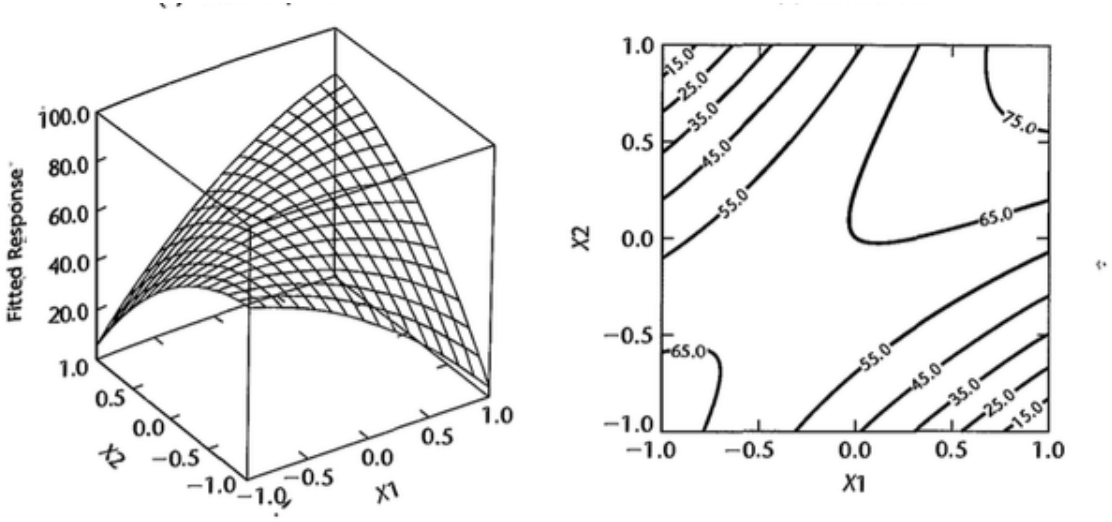
Para avaliarmos o efeito de 1 unidade de aumento em X_1, devemos considerar o valor de X_2 (retas não paralelas).



Interpretação: interação e efeitos curvilinear

Exemplo:

$$E(Y) = 65 + 3X_1 + 4X_2 - 10X_1^2 - 15X_2^2 + 35X_1X_2$$



Interpretação: interação e efeitos curvilinear

Se $X_1 = 1$:

$$E(Y) = 65 + 3 \times 1 + 4X_2 - 10 \times (1^2) - 15X_2^2 + 35 \times 1 \times X_2$$

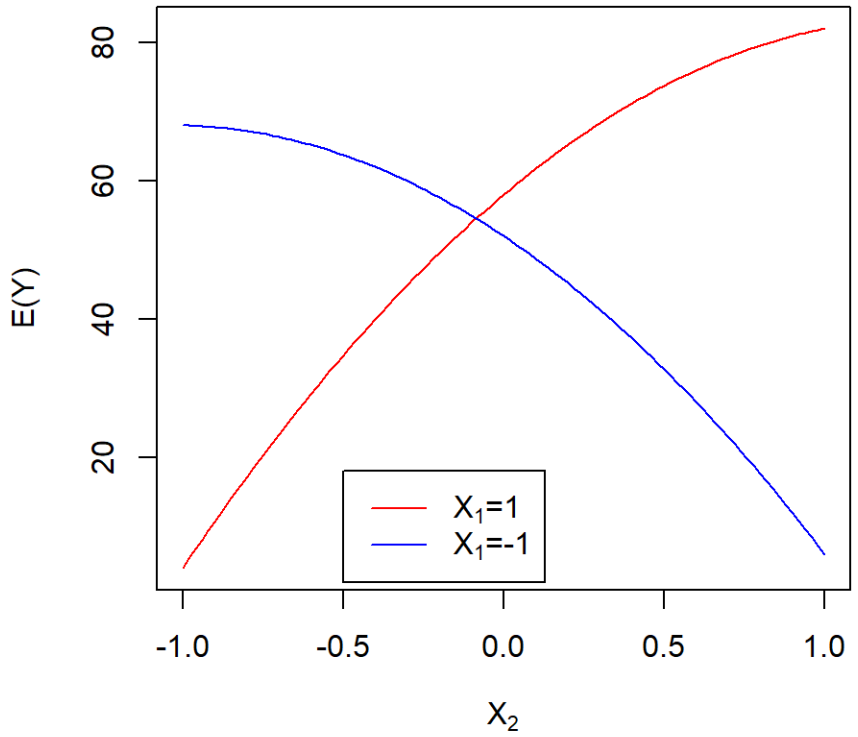
$$E(Y) = 58 + 39X_2 - 15X_2^2$$

Se $X_1 = -1$:

$$E(Y) = 65 + 3 \times (-1) + 4X_2 - 10 \times (-1^2) - 15X_2^2 + 35 \times (-1) \times X_2$$

$$E(Y) = 52 - 31X_2 - 15X_2^2$$

Interpretação: interação e efeitos curvilinear



Exemplo

X_1: tríceps, $X_{1*} = X_1 - \bar{X}_1$.

X_2: coxa, $X_{2*} = X_2 - \bar{X}_2$.

X_3: antebraço, $X_{3*} = X_1 - \bar{X}_3$.

Y: gordura corporal

Exemplo

X1	X2	X3	Y	x1	x2	x3
19.5	43.1	29.1	11.9	-5.805	-8.07	1.48
24.7	49.8	28.2	22.8	-0.605	-1.37	0.58
30.7	51.9	37.0	18.7	5.395	0.73	9.38
29.8	54.3	31.1	20.1	4.495	3.13	3.48
19.1	42.2	30.9	12.9	-6.205	-8.97	3.28
25.6	53.9	23.7	21.7	0.295	2.73	-3.92
31.4	58.5	27.6	27.1	6.095	7.33	-0.02
27.9	52.1	30.6	25.4	2.595	0.93	2.98
22.1	49.9	23.2	21.3	-3.205	-1.27	-4.42
25.5	53.5	24.8	19.3	0.195	2.33	-2.82
31.1	56.6	30.0	25.4	5.795	5.43	2.38
30.4	56.7	28.3	27.2	5.095	5.53	0.68
18.7	46.5	23.0	11.7	-6.605	-4.67	-4.62
19.7	44.2	28.6	17.8	-5.605	-6.97	0.98
14.6	42.7	21.3	12.8	-10.705	-8.47	-6.32
29.5	54.4	30.1	23.9	4.195	3.23	2.48
27.7	55.3	25.7	22.6	2.395	4.13	-1.92
30.2	58.6	24.6	25.4	4.895	7.43	-3.02
22.7	48.2	27.1	14.8	-2.605	-2.97	-0.52
25.2	51.0	27.5	21.1	-0.105	-0.17	-0.12

Exemplo

$$E(Y) = \beta_0 + \beta_1 X_{1*} + \beta_2 X_{2*} + \beta_3 X_{3*} + \beta_4 X_{1*}X_{2*} + \beta_5 X_{1*}X_{3*} + \beta_6 X_{2*}X_{3*} + \varepsilon$$

```
modelo <- lm(Y ~ x1 + x2 + x3 + I(x1*x2) + I(x1*x3) + I(x2*x3),data=dat)
summary(modelo)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 20.526893531 1.07362646 19.1192136 6.699796e-11
## x1           3.437808068 3.57866572  0.9606396 3.542612e-01
## x2          -2.094717339 3.03676957 -0.6897848 5.024579e-01
## x3          -1.616337237 1.90721068 -0.8474875 4.120550e-01
## I(x1 * x2)   0.008875562 0.03085046  0.2876963 7.781144e-01
## I(x1 * x3)  -0.084790836 0.07341774 -1.1549093 2.689155e-01
## I(x2 * x3)   0.090415385 0.09200130  0.9827621 3.436619e-01
```

Exemplo

```
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x1          1 352.27  352.27 52.2238 6.682e-06 ***
## x2          1  33.17   33.17  4.9173   0.04503 *
## x3          1  11.55   11.55  1.7117   0.21343
## I(x1 * x2)  1   1.50    1.50  0.2217   0.64552
## I(x1 * x3)  1   2.70    2.70  0.4009   0.53760
## I(x2 * x3)  1   6.51    6.51  0.9658   0.34366
## Residuals  13  87.69    6.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$

H_1: pelo menos um dentre $\beta_4, \beta_5, \beta_6$ é diferente de 0.

$p = 7$

$n = 20$

$q = 4$

$$F^* = \frac{SQReg(X_{1*}X_{2*}, X_{1*}X_{3*}, X_{2*}X_{3*} \mid X_{1*}, X_{2*}, X_{3*})/3}{SQE(X_{1*}, X_{2*}, X_{3*}, X_{1*}X_{2*}, X_{1*}X_{3*}, X_{2*}X_{3*})/13} \underset{\sim}{\text{sob}} H_0 \; F_{3,13}$$

Exemplo

$$\begin{aligned} SQReg(X_{1*}X_{2*}, X_{1*}X_{3*}, X_{2*}X_{3*} \mid X_{1*}, X_{2*}, X_{3*}) &= SQReg(X_{1*}X_{2*} \mid X_{1*}, X_{2*}, X_{3*}) \\ &+ SQReg(X_{1*}X_{3*} \mid X_{1*}, X_{2*}, X_{3*}, X_{1*}X_{2*}) \\ &+ SQReg(X_{2*}X_{3*} \mid X_{1*}, X_{2*}, X_{3*}, X_{1*}X_{2*}, X_{1*} \\ &= 1.5 + 2.7 + 6.514836 \\ &= 10.714836 \end{aligned}$$

$$F_{obs} = \frac{10.714836/3}{6.7} = 0.5330764$$

Comparando com $F(0.95; 3, 13) = 3.41$, não encontramos evidências contra a hipótese nula.

Exemplo

```
modeloreduz <- lm(Y ~ x1 + x2 + x3,data=dat)
anova(modeloreduz,modelo)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ x1 + x2 + x3
## Model 2: Y ~ x1 + x2 + x3 + I(x1 * x2) + I(x1 * x3) + I(x2 * x3)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     16 98.405
## 2     13 87.690  3    10.715 0.5295 0.6699
```

Preditores Qualitativos

Exemplo: Seguros

Y = meses até a implementação

X_1 = tamanho da firma (em milhões de dólares)

$$X_2 = \begin{cases} 1, & \text{se a firma tem ações na bolsa} \\ 0, & \text{caso contrário} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Exemplo: Seguros

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

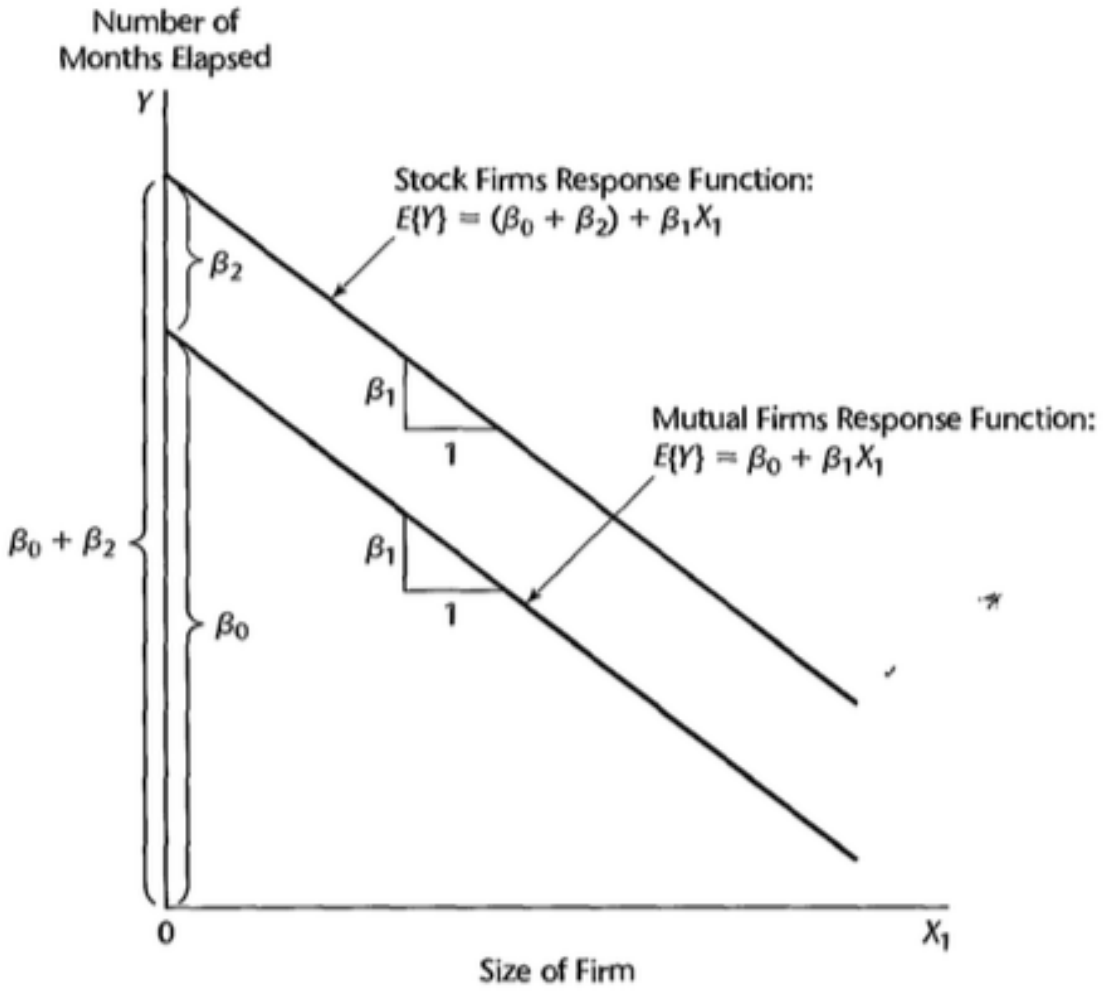
Se a firma não tem ações na bolsa, então $X_2 = 0$:

$$E(Y) = \beta_0 + \beta_1 X_1$$

Se a firma tem ações na bolsa, então $X_2 = 1$:

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$$

Exemplo: Seguros

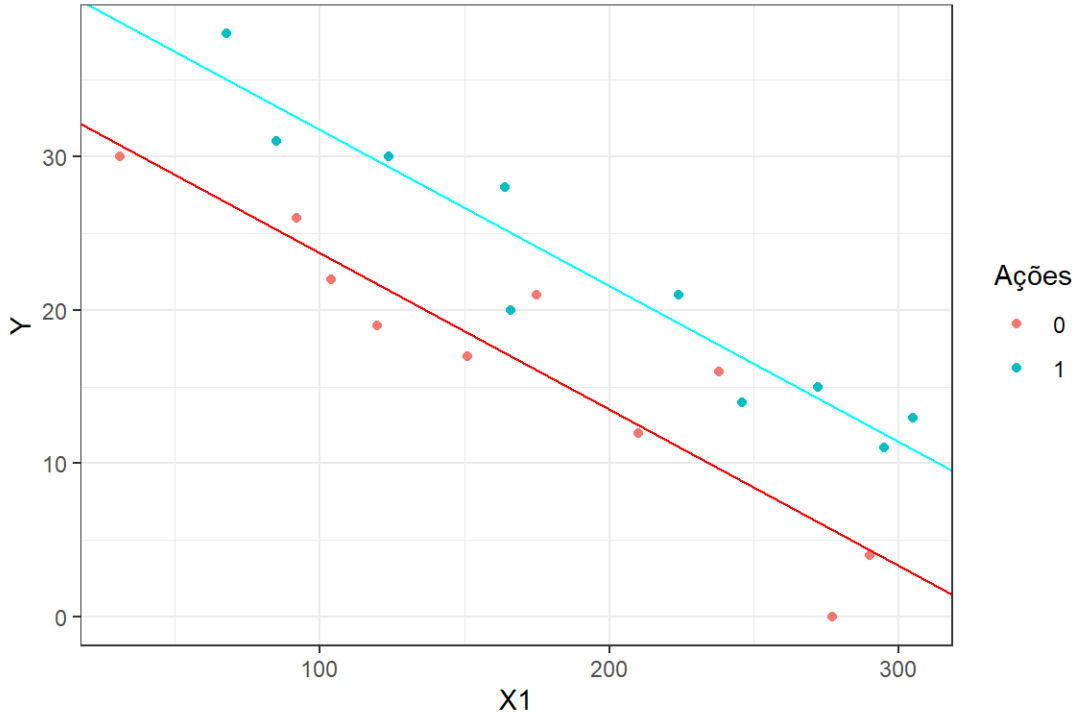


Exemplo: Seguros

```
##     Y  X1 X2
## 1  17 151  0
## 2  26  92  0
## 3  21 175  0
## 4  30  31  0
## 5  22 104  0
## 6   0 277  0
## 7  12 210  0
## 8  19 120  0
## 9   4 290  0
## 10 16 238  0
## 11 28 164  1
## 12 15 272  1
## 13 11 295  1
## 14 38  68  1
## 15 31  85  1
## 16 21 224  1
## 17 20 166  1
## 18 13 305  1
## 19 30 124  1
## 20 14 246  1
```

Exemplo: Seguros

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 33.8740690 1.813858297  18.675146 9.145269e-13
## X1          -0.1017421 0.008891218 -11.442990 2.074687e-09
## X2           8.0554692 1.459105700   5.520826 3.741874e-05
```



Exemplo: Seguros

Incluindo termo de interação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

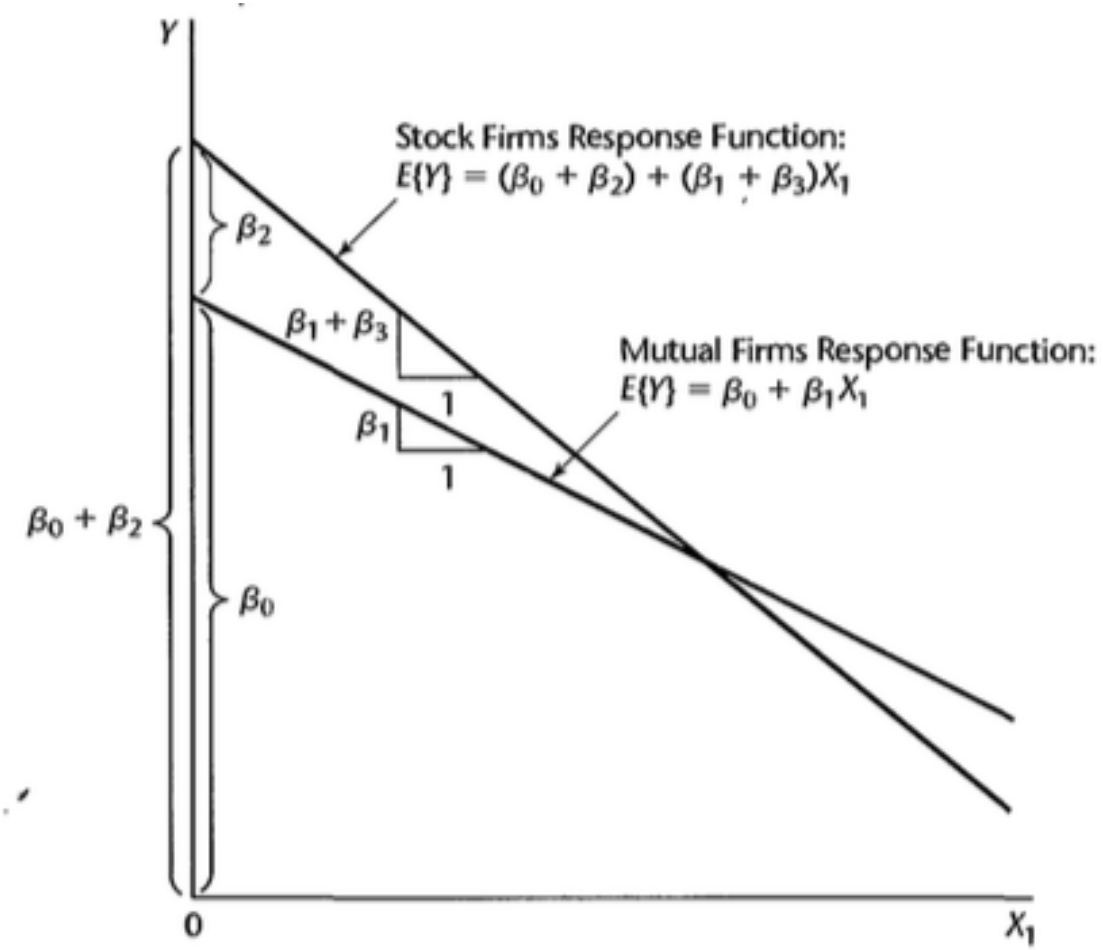
Se a firma não tem ações na bolsa, então $X_2 = 0$:

$$E(Y) = \beta_0 + \beta_1 X_1$$

Se a firma tem ações na bolsa, então $X_2 = 1$:

$$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

Exemplo: Seguros



Exemplo: Seguros

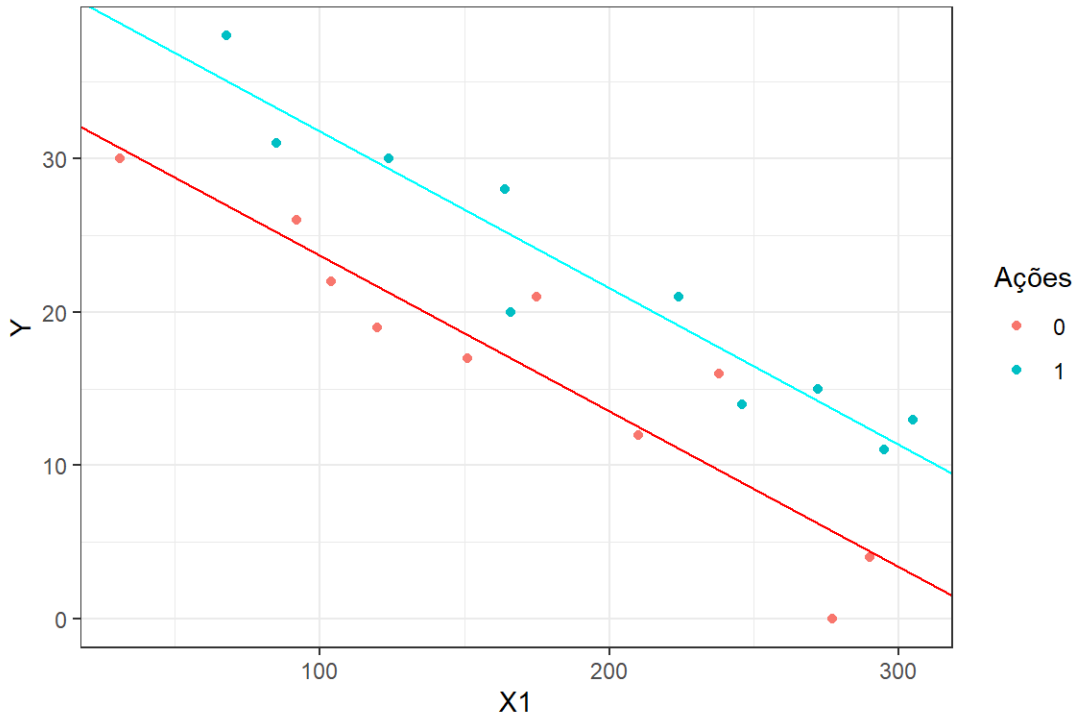
```
modelo <- lm(Y ~ X1 + X2 + X1*X2, data=dados); round(summary(modelo)$coef,5)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 33.83837    2.44065 13.86449  0.00000
## X1          -0.10153    0.01305 -7.77861  0.00000
## X2           8.13125    3.65405  2.22527  0.04079
## X1:X2       -0.00042    0.01833 -0.02276  0.98213
```

```
modelo <- lm(Y ~ X1*X2, data=dados); round(summary(modelo)$coef,5)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 33.83837    2.44065 13.86449  0.00000
## X1          -0.10153    0.01305 -7.77861  0.00000
## X2           8.13125    3.65405  2.22527  0.04079
## X1:X2       -0.00042    0.01833 -0.02276  0.98213
```

Exemplo: Seguros



Variável preditora com mais de duas classes

Exemplo: Salário (Y), anos de experiência desde a última titulação (X_1) e uma variável qualitativa $titulo$ que indica o grau de titulação do funcionário.

Existem 3 níveis de titulação: bacharel, mestre e doutor.

Definimos 2 variáveis *dummy* ou indicadoras:

$$X_2 = \begin{cases} 1, & \text{se mestre} \\ 0, & \text{caso contrário} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{se doutor} \\ 0, & \text{caso contrário} \end{cases}$$

Paramétrização por *casela de referência* a escolha de qual o nível será a referência é arbitrária.

Variável preditora com mais de duas classes

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Se título = bacharel:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \beta_3 \times 0 = \beta_0 + \beta_1 X_1$$

Se o título = mestre:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 \times 1 + \beta_3 \times 0 = (\beta_0 + \beta_2) + \beta_1 X_1$$

Se o título = doutor:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \beta_3 \times 1 = (\beta_0 + \beta_3) + \beta_1 X_1$$

O modelo de primeira ordem implica no fato de que o efeito dos anos de experiência é linear e com o mesmo coeficiente angular para todos os níveis de titularidade. Temos diferentes interceptos para cada modelo.

Variável preditora com mais de duas classes

- β_1: mudança esperada no salário médio (Y) para cada unidade de aumento nos anos de experiência (X_1), considerando o mesmo nível de titularidade.
- β_2: diferença esperada no salário médio do bacharel para o mestre, considerando a mesma velocidade.
- β_3: diferença esperada no salário médio do bacharel para o doutor, considerando a mesma velocidade.

Variável preditora com mais de duas classes

```
##        Y    X1   titulo
## 1   58.8  4.49   doutor
## 2   34.8  2.92 bacharel
## 3  163.7 29.54   doutor
## 4   70.0  9.92   doutor
## 5   55.5  0.14   doutor
## 6   85.0 15.96   mestre
## 7   34.0  2.27 bacharel
## 8   29.7  1.20 bacharel
## 9   56.1  5.33   mestre
## 10  70.6 15.74   doutor
```

Variável preditora com mais de duas classes

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  25.79587  4.0030933 6.443984 2.072072e-08
## X1            2.38670  0.2590196 9.214361 3.704340e-13
## titulomestre 13.38309  4.6068743 2.905025 5.109154e-03
## titulodoutor 28.56642  4.9012373 5.828410 2.265887e-07
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 24851.9 24851.9 116.708 8.546e-16 ***
## titulo     2  7342.7  3671.4  17.241 1.161e-06 ***
## Residuals 61 12989.4   212.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variável preditora com mais de duas classes

Qual a diferença esperada no salário médio entre o mestre o doutor?

Quando o título é mestre temos que:

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$$

Quando o título é doutor temos que:

$$E(Y) = (\beta_0 + \beta_3) + \beta_1 X_1$$

A diferença entre mestre e doutor, mantendo os anos de experiência:

$$(\beta_0 + \beta_3) + \beta_1 X_1 - [(\beta_0 + \beta_2) + \beta_1 X_1] = \beta_3 - \beta_2$$

Variável preditora com mais de duas classes

Após obtermos estimativas: $\hat{\beta}_3 - \hat{\beta}_2$ e devemos também fornecer o erro-padrão da estimativa.

Lembre que:

$$Var(\hat{\beta}_3 - \hat{\beta}_2) = Var(\hat{\beta}_3) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_3, \hat{\beta}_2)$$

Variável preditora com mais de duas classes

$\hat{\beta}_3$ - $\hat{\beta}_2$ = 28.57 - 13.38 = 15.18

$\hat{Var}(\hat{\beta})$

```
##              (Intercept)    X1 titulomestre titulodoutor
## (Intercept)        16.02 -0.43       -13.03       -11.63
## X1                 -0.43  0.07        -0.04        -0.26
## titulomestre      -13.03 -0.04        21.22        13.48
## titulodoutor      -11.63 -0.26        13.48        24.02
```

$$Var(\hat{\beta}_3 - \hat{\beta}_2) = 24.02 + 21.22 - 2 \times 13.48 = 18.28$$

Exemplo: Fábrica de sabão

Y: resíduo de sabão

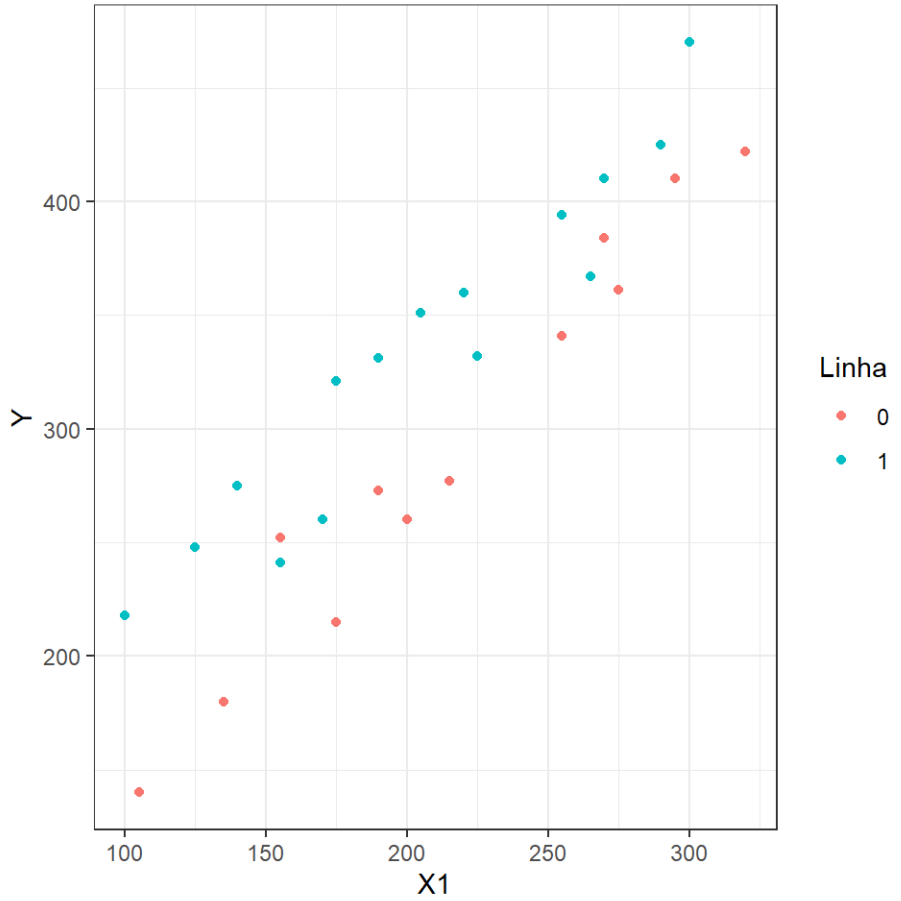
X_1: velocidade

$$X_2 = \begin{cases} 1, & \text{se produção na linha 1} \\ 0, & \text{caso contrário} \end{cases}$$

Exemplo: Fábrica de Sabão

Y X1 X2 1 218 100 1 2 248 125 1 3 360 220 1 4 351 205 1 5 470 300 1 6 394 255 1 7 332 225 1 8 321 175 1 9 410 270 1 10 260 170 1 11 241 155 1 12 331 190 1 13 275 140 1 14 425 290 1 15 367 265 1 16 140 105 0 17 277 215 0 18 384 270 0 19 341 255 0 20 215 175 0 21 180 135 0 22 260 200 0 23 361 275 0 24 252 155 0 25 422 320 0 26 273 190 0 27 410 295 0

Exemplo: Fábrica de sabão



Exemplo: Fábrica de sabão

Iremos ajustar um modelo assumindo que:

- a relação entre a quantidade de resíduo e velocidade é linear para as duas linhas de produção;
- retas diferentes para as duas linhas de produção;
- as variâncias dos termos de erros ao redor de cada reta são iguais.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Para a linha 1: $E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$.

Para a linha 2: $E(Y) = \beta_0 + \beta_1 X_1$.

Exemplo: Fábrica de sabão

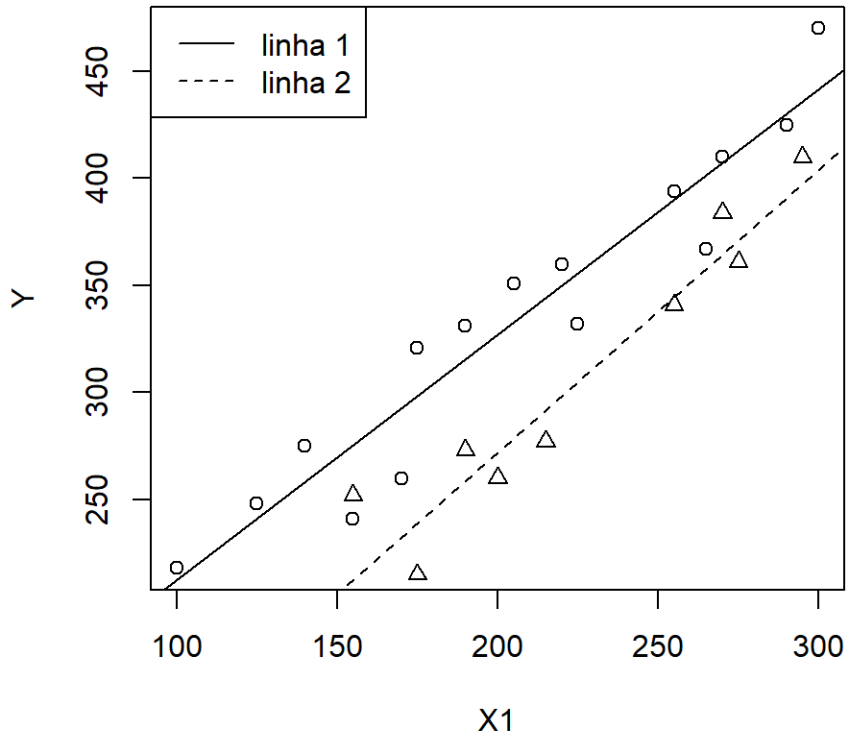
```
modelo <- lm(Y ~ X1 + X2 + I(X1*X2), data=dados)
summary(modelo)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  7.5744646 20.8696979  0.3629408 7.199633e-01
## X1           1.3220488  0.0926247 14.2731771 6.446165e-13
## X2          90.3908632 28.3457320  3.1888703 4.085851e-03
## I(X1 * X2)  -0.1766614  0.1288377 -1.3711932 1.835463e-01
```

```
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## X1          1 149661  149661 347.5548 2.224e-15 ***
## X2          1  18694   18694  43.4129 1.009e-06 ***
## I(X1 * X2)  1    810     810   1.8802    0.1835
## Residuals  23   9904     431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo: Fábrica de sabão



Exemplo: Fábrica de sabão

Se quisermos testar a hipótese nula de que temos apenas uma reta para representar as duas linhas:

H_0: $\beta_2 = \beta_3 = 0$

H_a: pelo menos um entre β_2 e β_3 é diferente de zero.

Estatística do teste:

$$F^* = \frac{SQReg(X_q, \ldots, X_{p-1} \mid X_1, \ldots, X_{q-1})}{p-q} \div \frac{SQE(X_1, \ldots, X_{p-1})}{n-p}$$

$$\text{sob} \underset{\sim}{H_0} \; F_{p-q,n-p}$$

Exemplo: Fábrica de sabão

$p = 4$

$n = 27$

$q = 2$

$$F^* = \frac{SQReg(X_2, X_1X_2 \mid X_1)/2}{SQE(X_1, X_2, X_1X_2)/23} \underset{\sim}{\text{sob}} H_0 \; F_{2,23}$$

$$\begin{aligned} SQReg(X_2, X_1X_2 \mid X_1) &= SQReg(X_2 \mid X_1) + SQReg(X_1X_2 \mid X_1, X_2 \\ &= 1.86941 \times 10^4 + 809.6 \\ &= 1.95037 \times 10^4 \end{aligned}$$

$$F_{obs} = \frac{1.95037 \times 10^4/2}{430.6} = 22.6471203$$

Comparando com $F(0.95; 2, 23) = 3.42$, encontramos evidências contra a hipótese nula.

Exemplo: Fábrica de sabão

```
modeloreduz <- lm(Y ~ X1, data=dados)
anova(modeloreduz,modelo)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2 + I(X1 * X2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     25 29407.8
## 2     23  9904.1  2     19504 22.646 3.669e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo: Fábrica de sabão

Se quisermos testar a hipótese nula de que para as duas linhas de produção o coeficiente angular é o mesmo:

$H_0: \beta_3 = 0$

$H_a: \beta_3 \neq 0.$

$p = 4$

$n = 27$

$q = 3$

$$F^* = \frac{SQReg(X_1X_2 \mid X_1, X_2)/1}{SQE(X_1, X_2, X_1X_2)/23} \underset{\sim}{\text{sob}} H_0 \; F_{1,23}$$

Exemplo: Fábrica de sabão

$$F_{obs} = \frac{809.6/1}{430.6} = 1.8801672$$

Comparando com $F(0.95; 1, 23) = 4.28$, não encontramos evidências contra a hipótese nula.

Agradecimento

- Slides criados por Samara F Kiihl / IMECC / UNICAMP
- Editado por Rafael P Maia / IMECC / UNICAMP

Leitura

- Applied Linear Statistical Models: Seções 8.1-8.3, 8.5-8.7.
- Faraway - Linear Models with R: Capítulo 14.
- Draper & Smith - Applied Regression Analysis: Capítulo 12.