Regressão Linear Simples

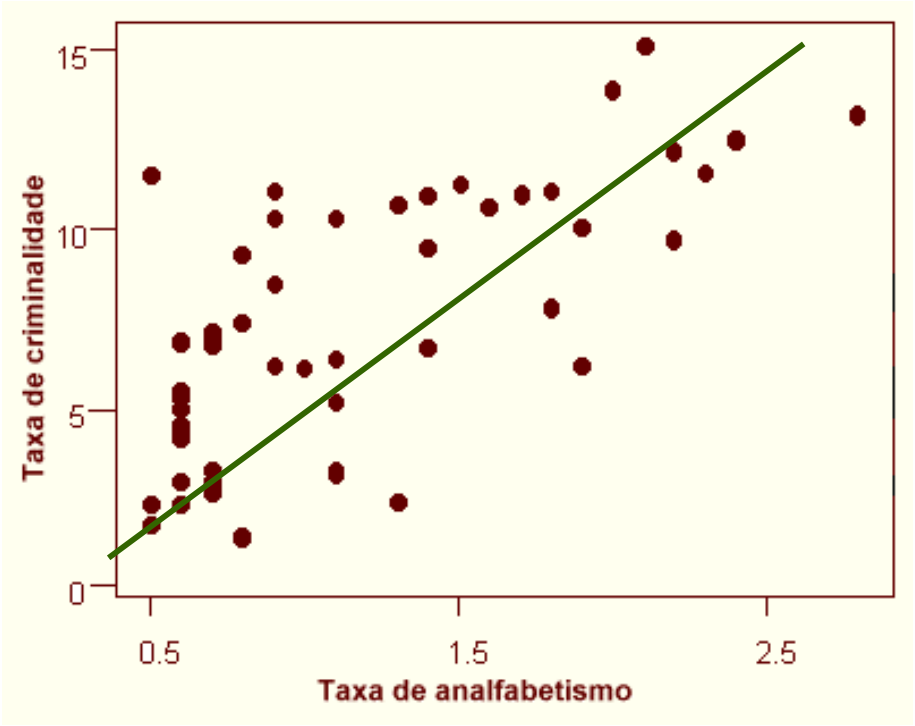
Análise de Dados e Simulação

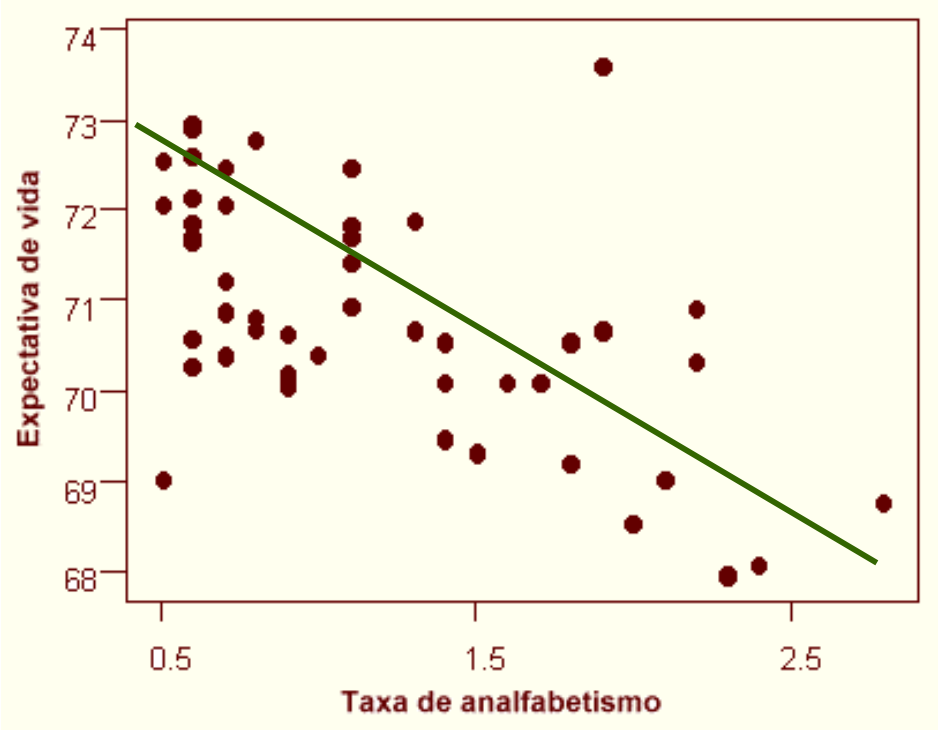
Márcia D'Elia Branco
http://www.ime.usp.br/~mbranco

Apoio: Andressa Cerqueira (Aluna do Programa PAE)

Análise de Regressão

Diagramas de Dispersão





⇒ **Explicar a forma da relação por meio de uma função matemática:** $Y = a + bX$

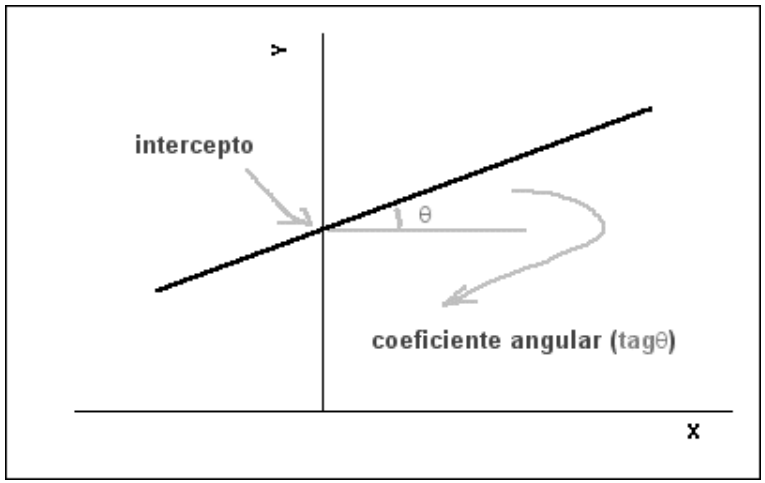
Análise de Regressão

Reta ajustada: $\hat{Y} = a + bX$

O que são *a* e *b*?

***a* : intercepto**

***b* : inclinação ou coeficiente angular**

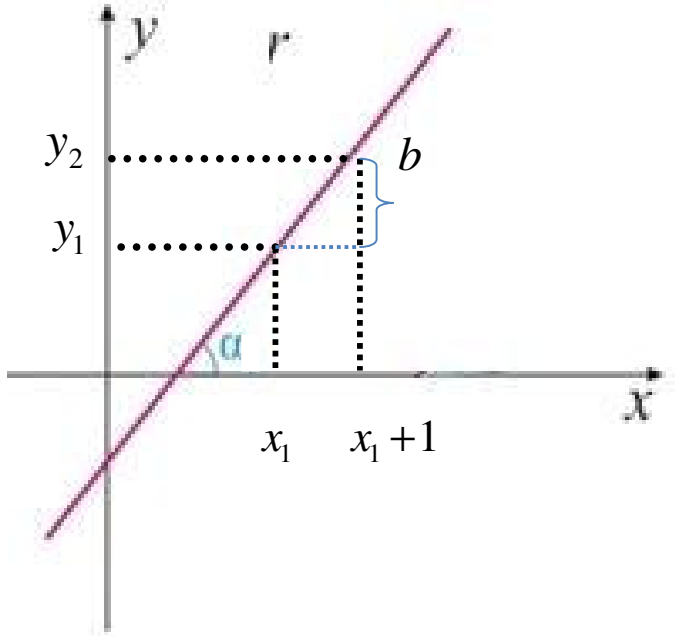


Reta ajustada: $\hat{Y} = a + bX$

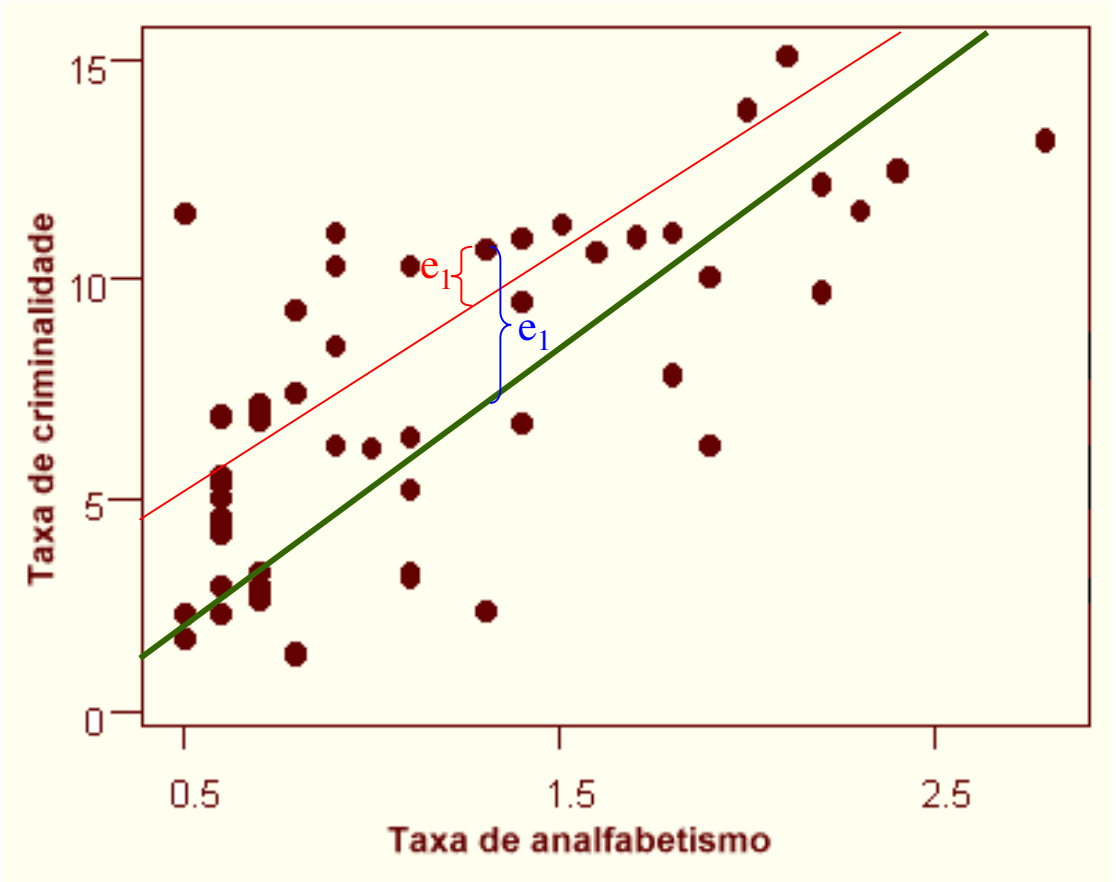
Interpretação de *b*:

Para cada aumento de uma unidade em *X*, temos um aumento médio de *b* unidades em *Y*.

$$tag(\alpha) = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_1 + 1 - x_1}$$
$$= y_2 - y_1 = b$$



Reta ajustada
(Método de mínimos quadrados)



Reta ajustada
(Método de mínimos quadrados)

Os coeficientes *a* e *b* são calculados da seguinte maneira:

$$\mathrm{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$\mathrm{a} = \overline{y} - \mathrm{b}\overline{x}$$

Exemplo 1. Progressão Continuada e Seriação: um estudo comparativo.

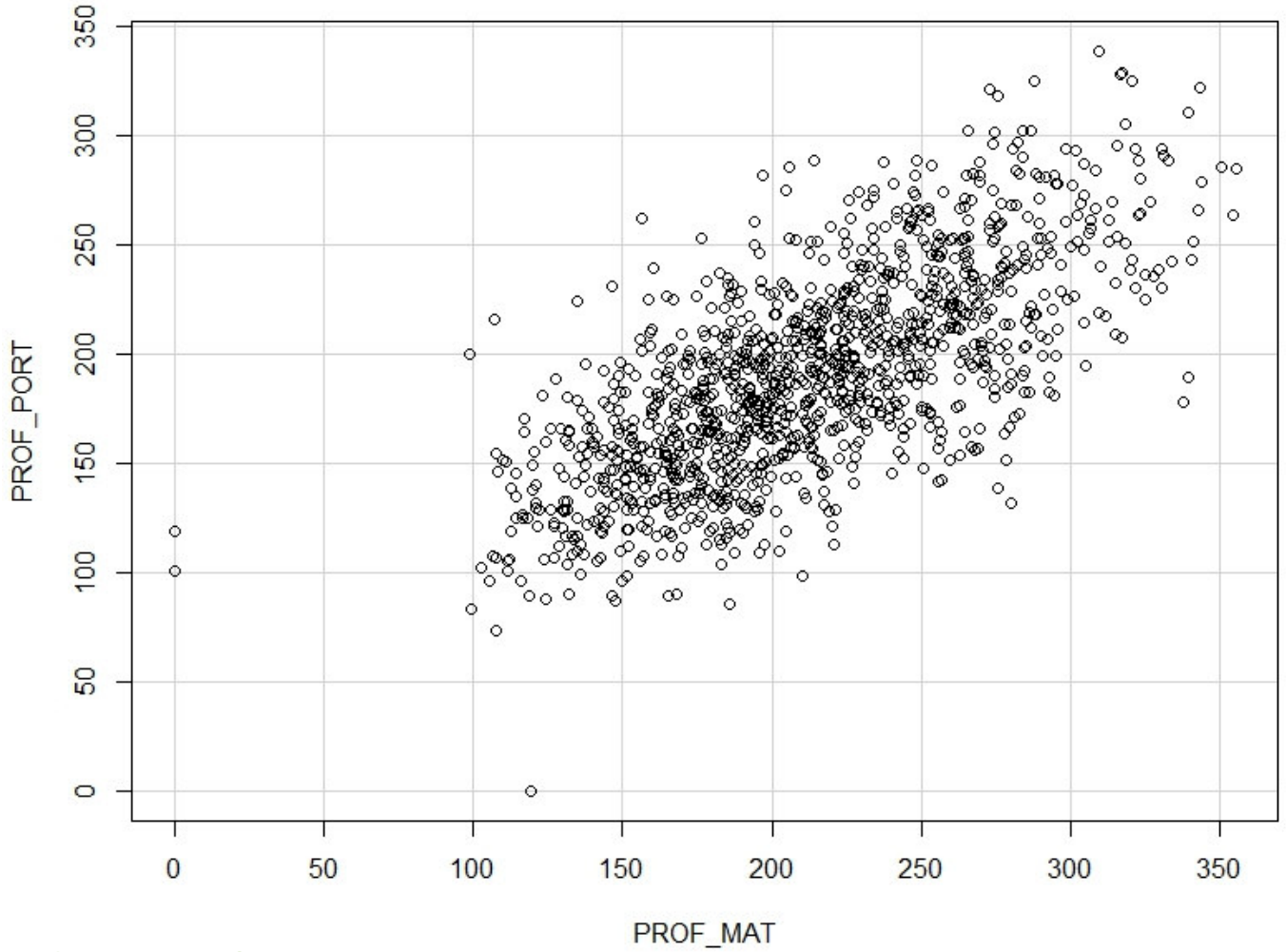
Estudar a relação entre a proficiência em português e a proficiência em matemática de acordo com os dados da Prova Brasil de 2007 e 2009.

Dados: Prova Brasil de 2007 e 2009.

Amostra: 1.128 alunos de 6 escolas

- 2 escolas municipais (regime seriado)
- 4 escolas estaduais (regime continuado)

Diagrama de Dispersão

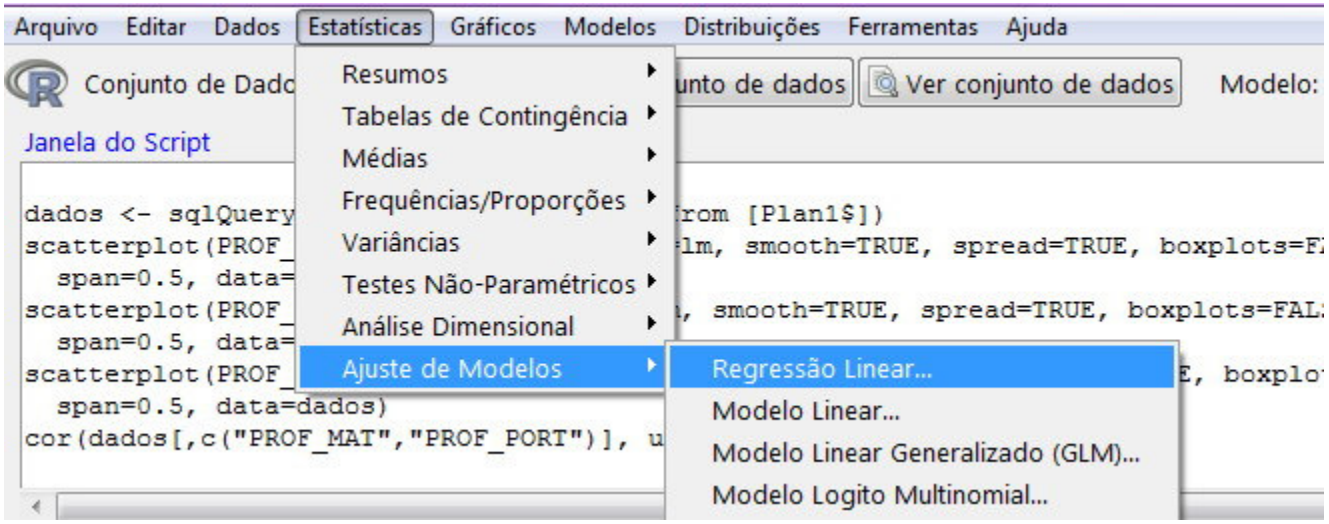


Matriz de Correlação

	PROF_MAT	PROF_PORT
PROF_MAT	1.0000000	0.7072603
PROF_PORT	0.7072603	1.0000000

$$r = 0.7072603$$

No R:



Saída do R:

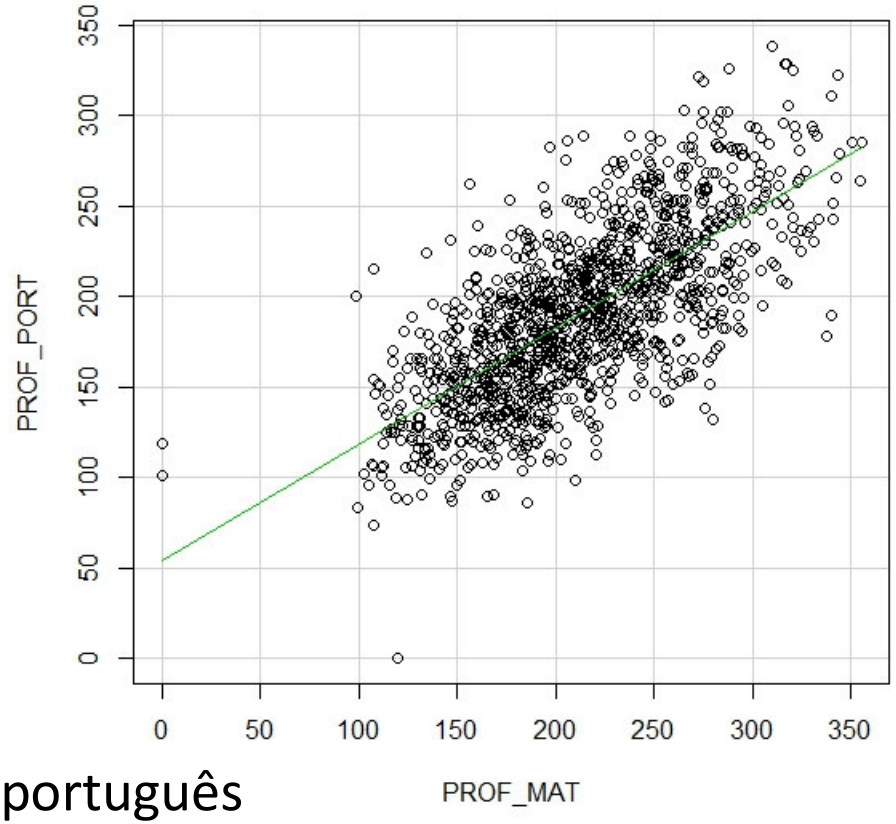
Janela de Resultados

```
> summary(RegModel.1)

Call:
lm(formula = PROF_PORT ~ PROF_MAT, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-131.020  -22.229   -0.835   22.151  107.457

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.35827    4.00317   13.58   <2e-16 ***
PROF_MAT     0.64127    0.01831   35.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.25 on 1226 degrees of freedom
Multiple R-squared: 0.5002, Adjusted R-squared: 0.4998
F-statistic:  1227 on 1 and 1226 DF,  p-value: < 2.2e-16
```

Exemplo 1.

A reta ajustada é:

$$\widehat{Y} = 54,36 + 0,64X$$



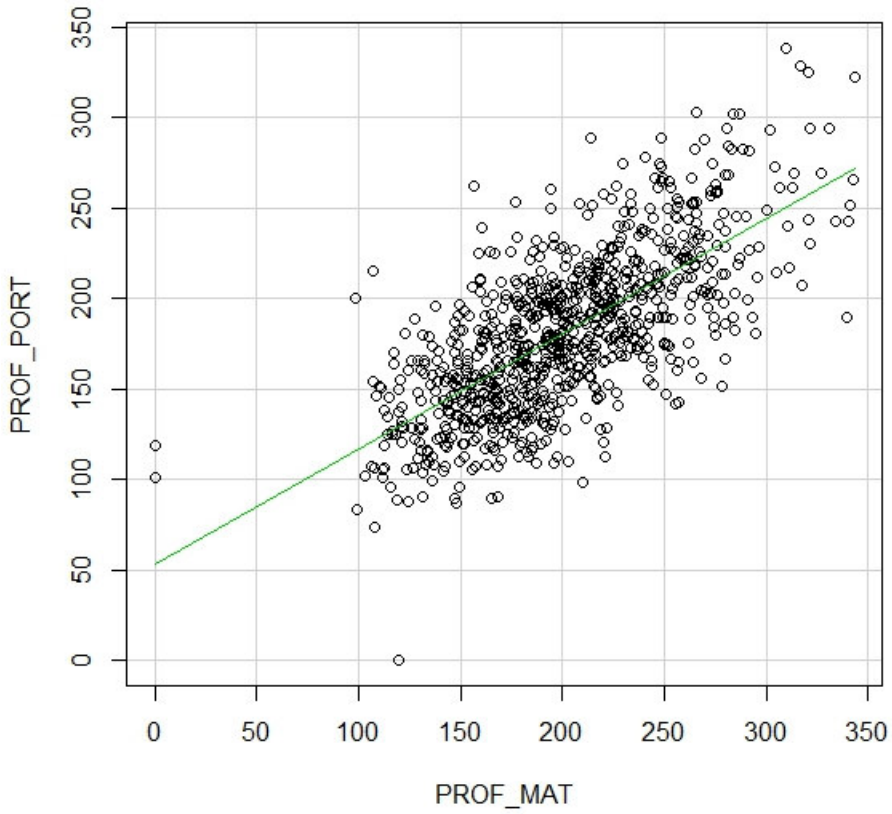
$\hat{Y}$: valor predito para proficiência em português

X : proficiência em matemática

Interpretação de *b*:

Para um aumento de uma unidade na proficiência em matemática (*X*), a proficiência em português (*Y*) aumenta, em média, 0,64 unidades.

Escola Estadual

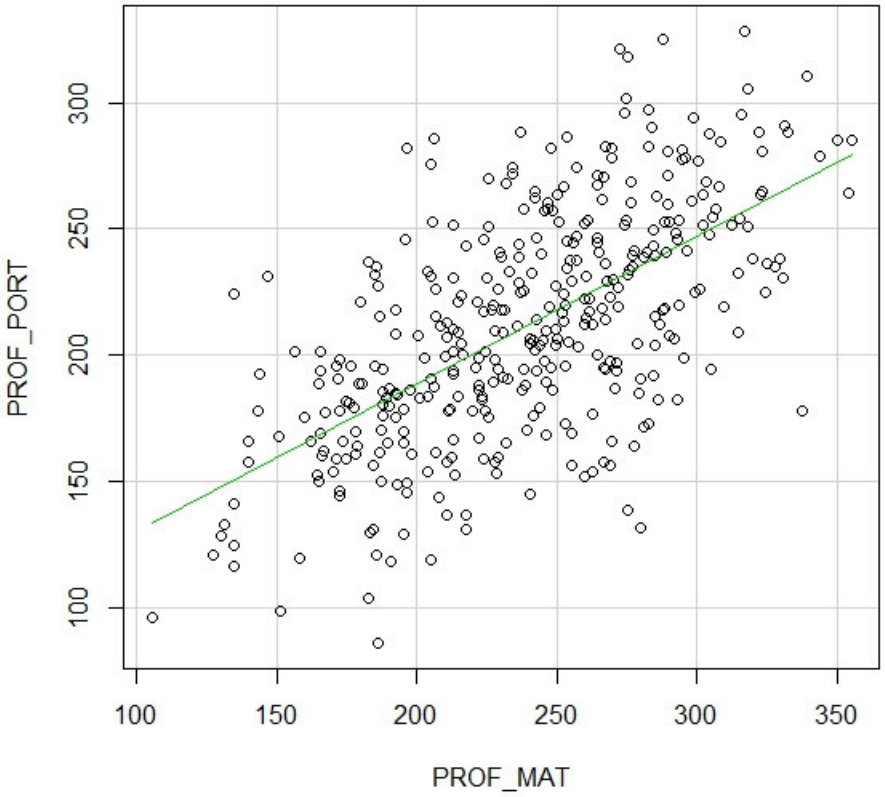


A reta ajustada é:

$$\widehat{Y} = 53,45 + 0,64X$$

$$r = 0.6990$$

Escola Municipal



A reta ajustada é:

$$\widehat{Y} = 71,38 + 0,04X$$

$$r = 0.6237$$

Exemplo 2: Relação entre o tempo que um indivíduo leva para reagir a um estímulo visual e a idade do indivíduo

Y	X	X.Y	X_2
96	20	1920	400
92	20	1840	400
106	20	2120	400
100	20	2000	400
98	25	2450	625
104	25	2600	625
110	25	2750	625
101	25	2525	625
116	30	3480	900
106	30	3180	900
109	30	3270	900
100	30	3000	900
112	35	3920	1225
105	35	3675	1225
118	35	4130	1225
108	35	3780	1225
113	40	4520	1600
112	40	4480	1600
127	40	5080	1600
117	40	4680	1600
2150	600	65400	19000

Y: tempo de reação ao estímulo

X: idade do indivíduo

$$n = 20 \quad \bar{x} = 30 \quad \bar{y} = 107{,}50$$

$$r = 0.7681$$

$$\mathrm{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

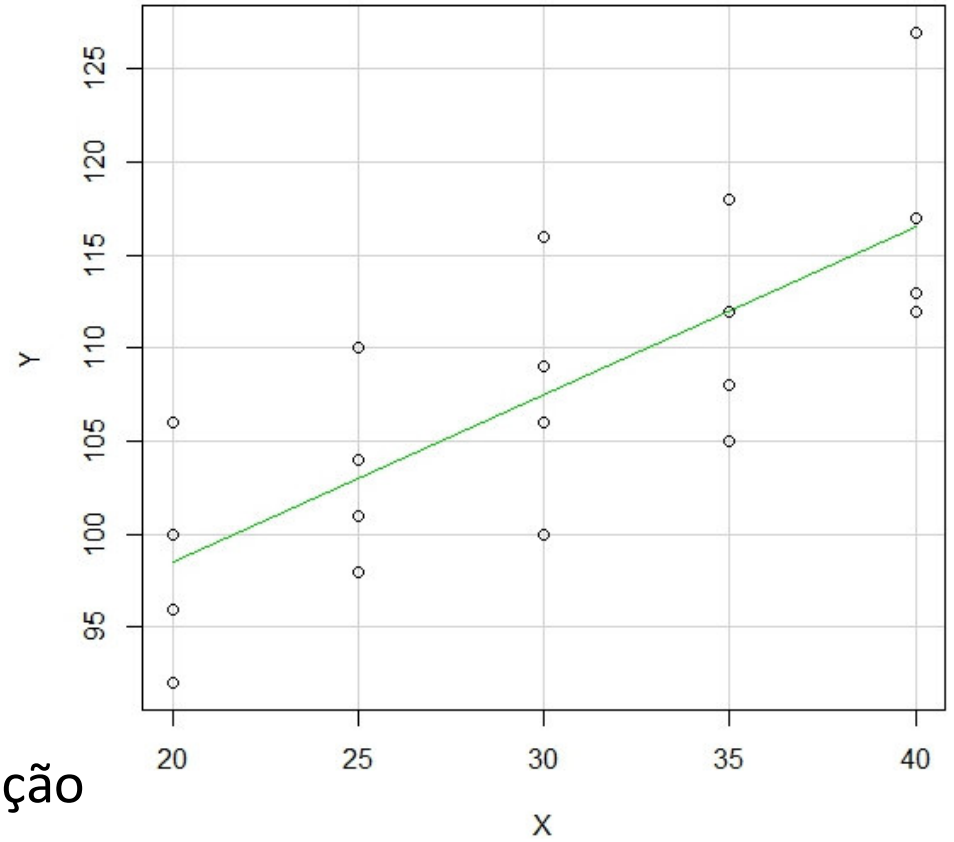
$$= \frac{65.400 - 20 \cdot 30 \cdot 107,50}{19.000 - 20 \cdot 30^2} = 0,90$$

$$\mathrm{a} = 107,50 - 0,90 \cdot 30 = 80,50$$

Exemplo 2.

A reta ajustada é:

$$\hat{Y} = 80,50 + 0,90X$$



$\hat{Y}$: valor predito para o tempo de reação

X : idade do indivíduo

Interpretação de *b*:

Para um aumento de um ano na idade do indivíduo (*X*), o tempo de reação (*Y*) aumenta, em média, 0,90 unidades.

Exemplo 3. Oito indivíduos foram submetidos a um teste sobre conhecimento de língua extrangeira e, em seguida, mediu-se o tempo gasto para cada um aprender a operar uma determinada máquina. As variáveis medidas foram:

Y : tempo, em minutos, necessário para operar a máquina

X : resultado no teste (máximo = 100 pontos)

X	Y
45	343
52	368
61	355
70	334
74	337
76	381
80	345
90	375

Matriz de Correlação

	X	Y
X	1.0000000	0.2381005
Y	0.2381005	1.0000000

Saída do R:

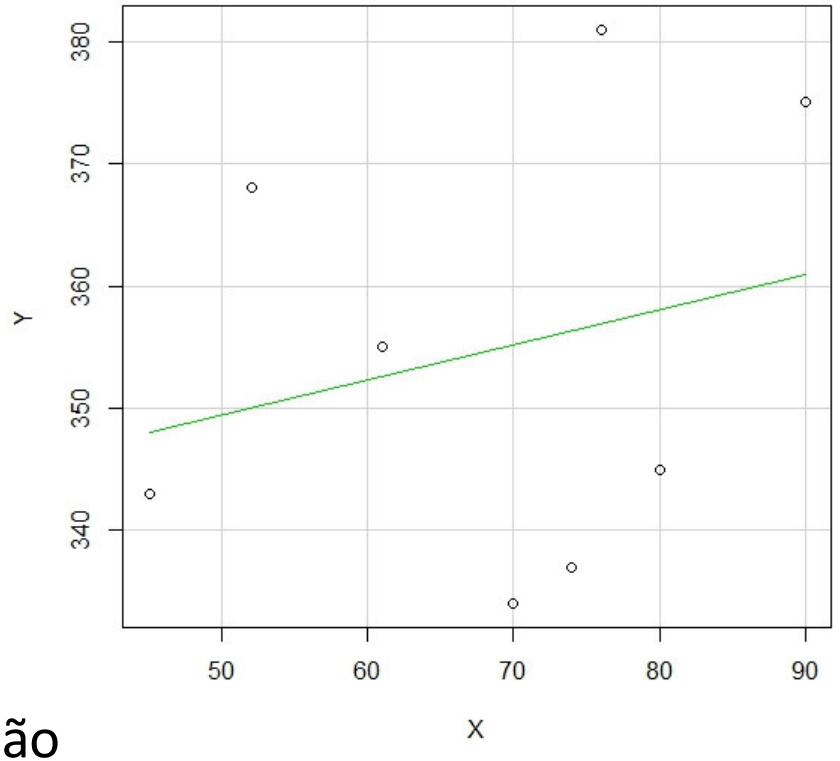
Janela de Resultados

```
Call:
lm(formula = Y ~ X, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
-21.18 -14.61  -1.32  15.07  24.11

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 335.1723    33.2748   10.07 5.56e-05 ***
X             0.2858     0.4760    0.60     0.57
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.82 on 6 degrees of freedom
Multiple R-squared: 0.05669, Adjusted R-squared: -0.1005
F-statistic: 0.3606 on 1 and 6 DF,  p-value: 0.5701
```

Exemplo 3.

A reta ajustada é:

$$\widehat{Y} = 335,17 + 0,29X$$



$\hat{Y}$: valor predito para o tempo de operação

X : resultado no teste

Interpretação de *b*:

Para um aumento de uma unidade no resultado do teste (*X*), o tempo de operação da máquina (*Y*) aumenta, em média, 0,29 unidades.

Exemplo 4. **Expectativa de vida e analfabetismo**

Considere as duas variáveis observadas em 50 estados norte-americanos.

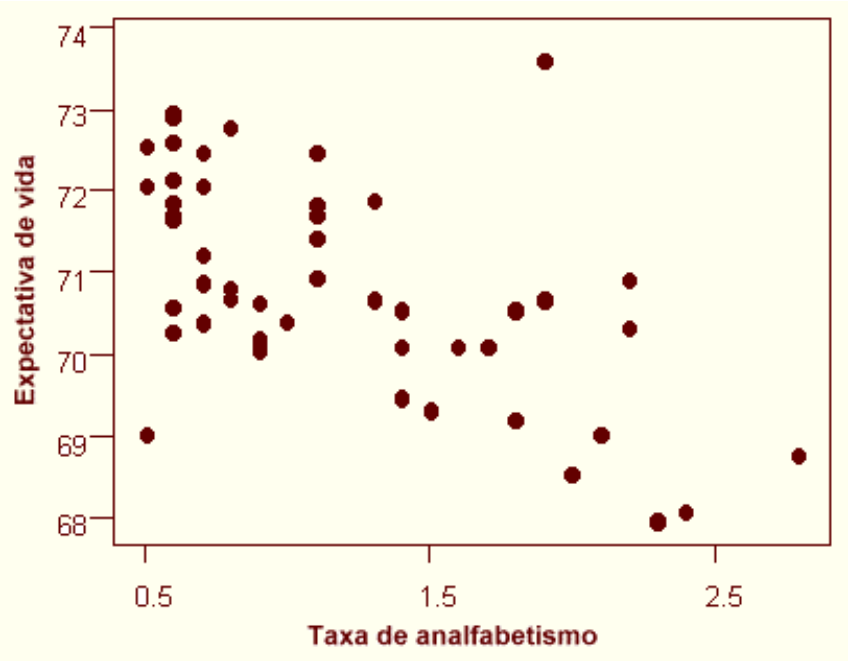
Y: expectativa de vida

X: taxa de analfabetismo

Exemplo 4.

A reta ajustada é:

$$\hat{Y} = 72{,}395 - 1{,}296\ X$$



$\hat{Y}$: valor predito para expectativa de vida

X : taxa de analfabetismo

Interpretação de *b*:

Para um aumento de uma unidade na taxa do analfabetismo (*X*), a expectativa de vida (*Y*) diminui, em média, 1,296 anos.

Graficamente, temos $\hat{Y} = 72{,}395 - 1{,}296\ X$

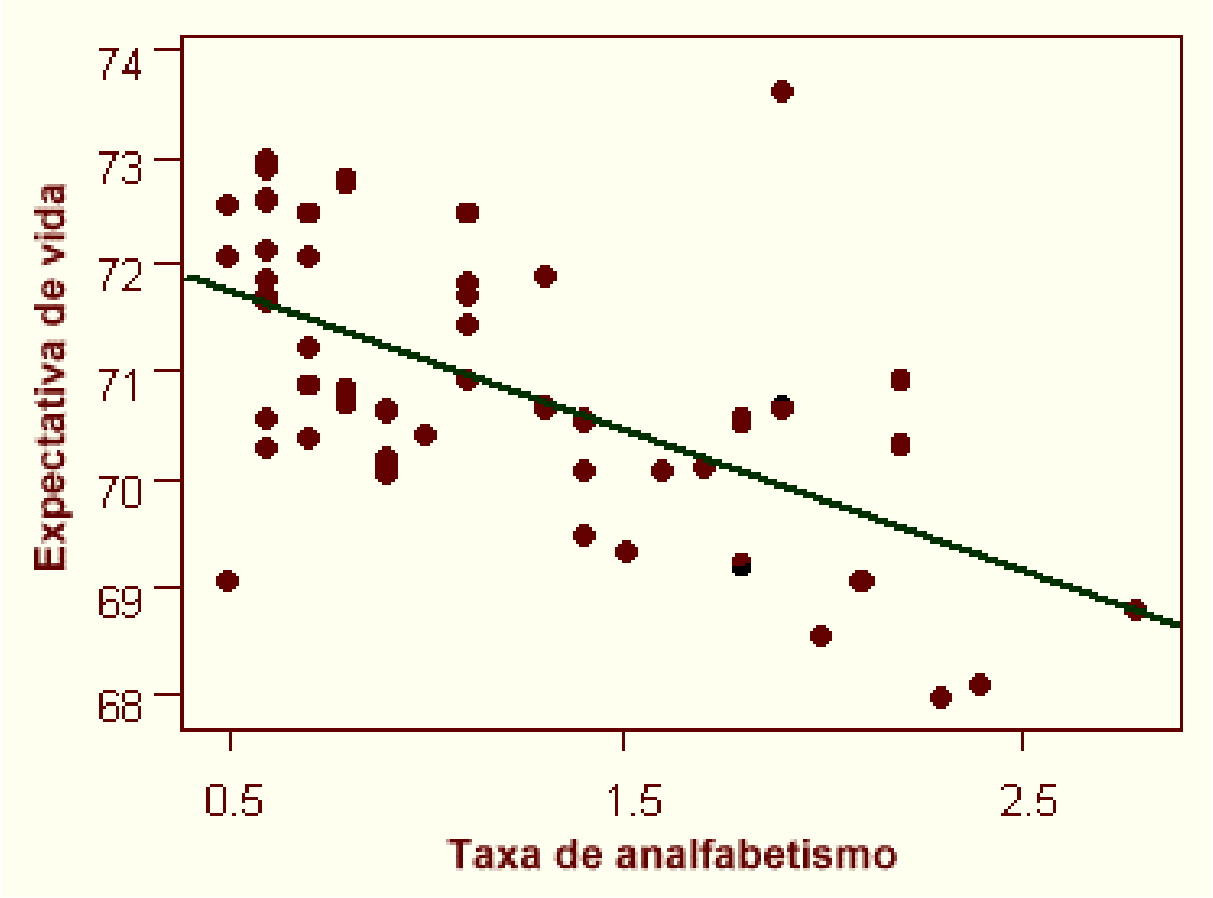


Gráfico quantis x quantis

Suponha que temos valores $x_1,..., x_n$ da variável X e valores $y_1,..., y_m$ da variável Y, todos medidos pela mesma unidade.

O gráfico *q x q* é um gráfico dos quantis de X contra os quantis de Y.

Se $m=n$ o gráfico *q x q* é um gráfico dos dados ordenados de X contra os dados ordenados de Y.

Se as distribuições dos dois conjuntos de dados fossem idênticas, os pontos estariam sobre a reta y = x.

Gráfico quantis x quantis

O gráfico de dispersão fornece uma possível relação global entre as variáveis.

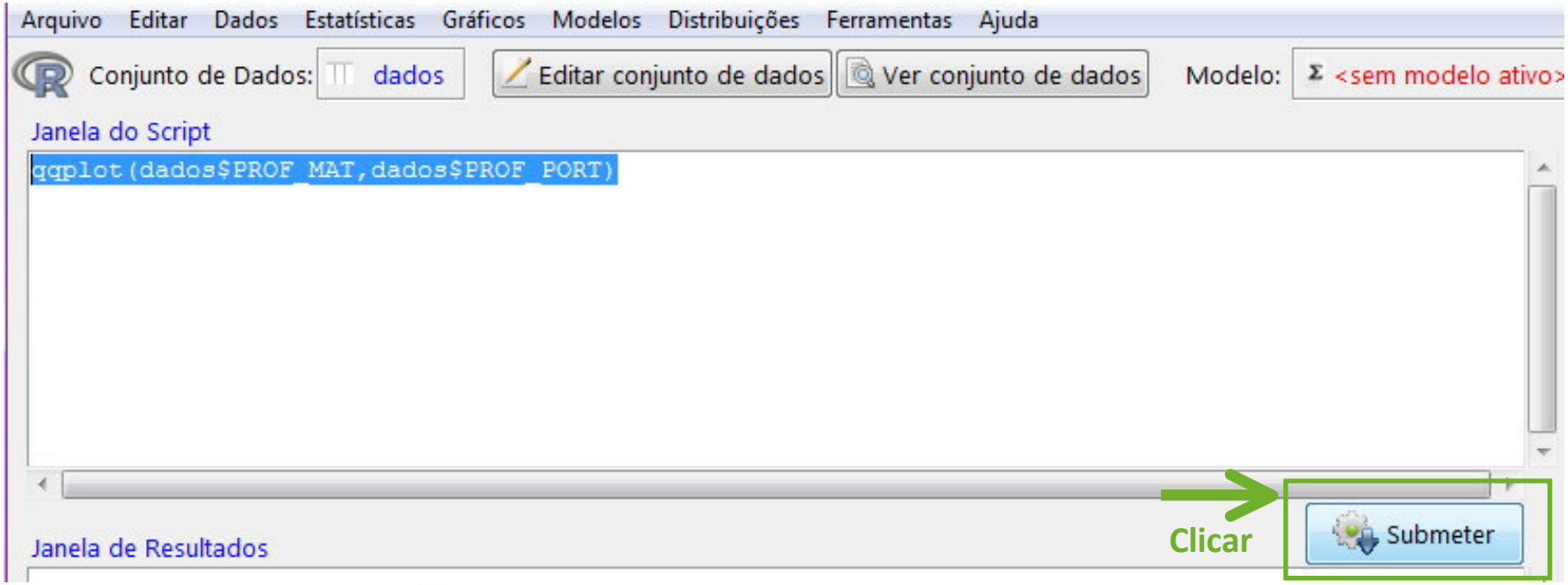
O gráfico *q x q* mostra:

• se valores pequenos de X estão relacionados com valores pequenos de Y;

• se valores intermediários de X estão relacionados com valores intermediários de Y;

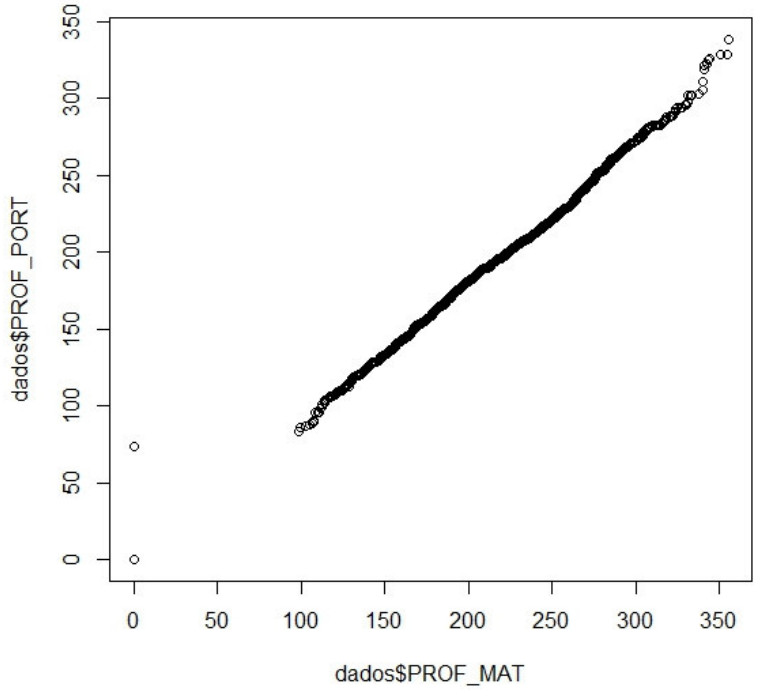
• se valores grande de X estão relacionados com valores grandes de Y.

No R: `qqplot(nome_dados$nome_variavel1,nome_dados$nome_variavel2)`



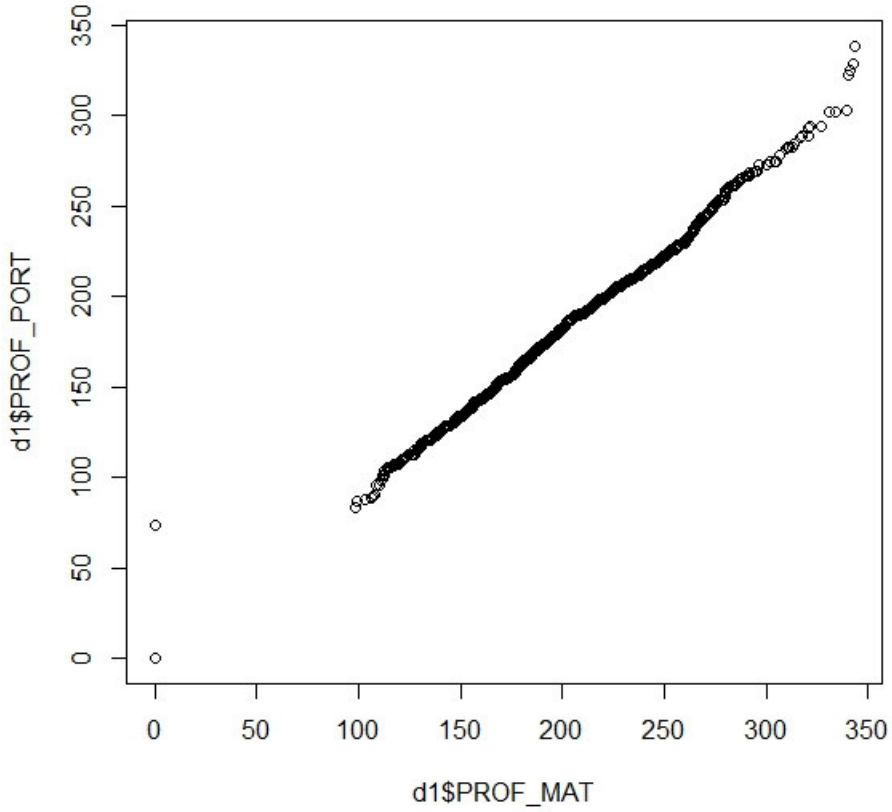
Exemplo 1.

Resultado

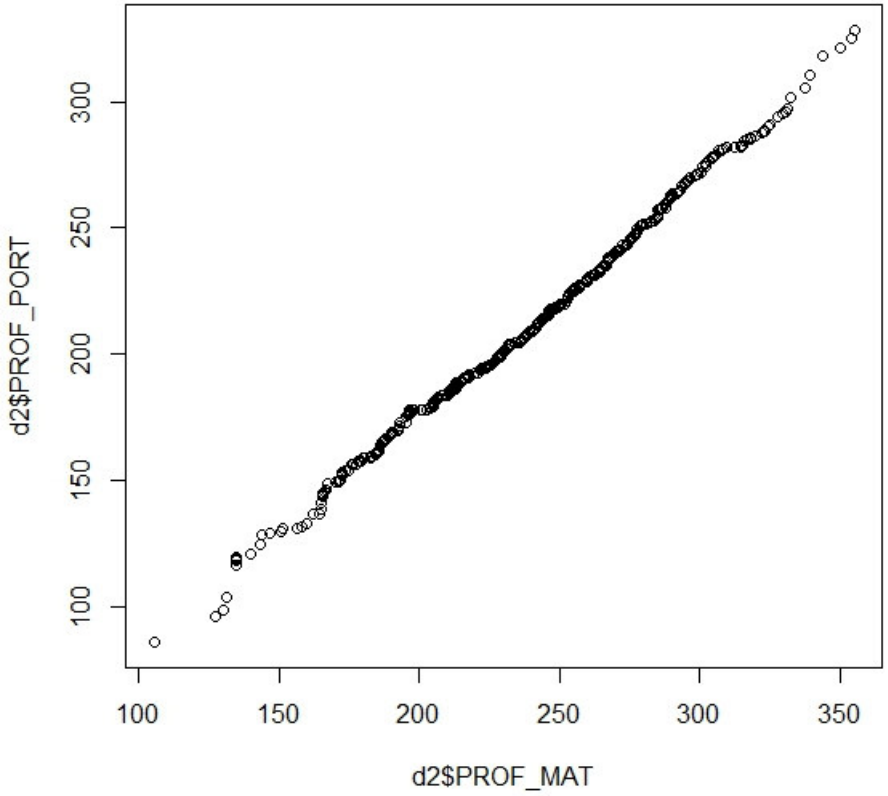


Exemplo 1.

Escola Estadual

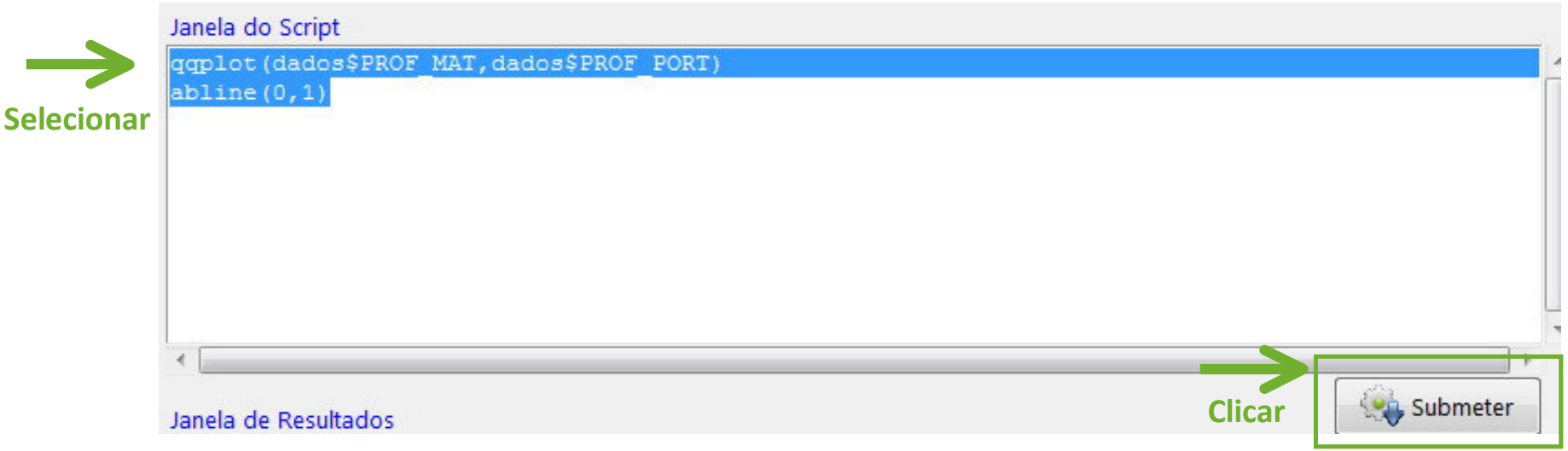


Escola Municipal



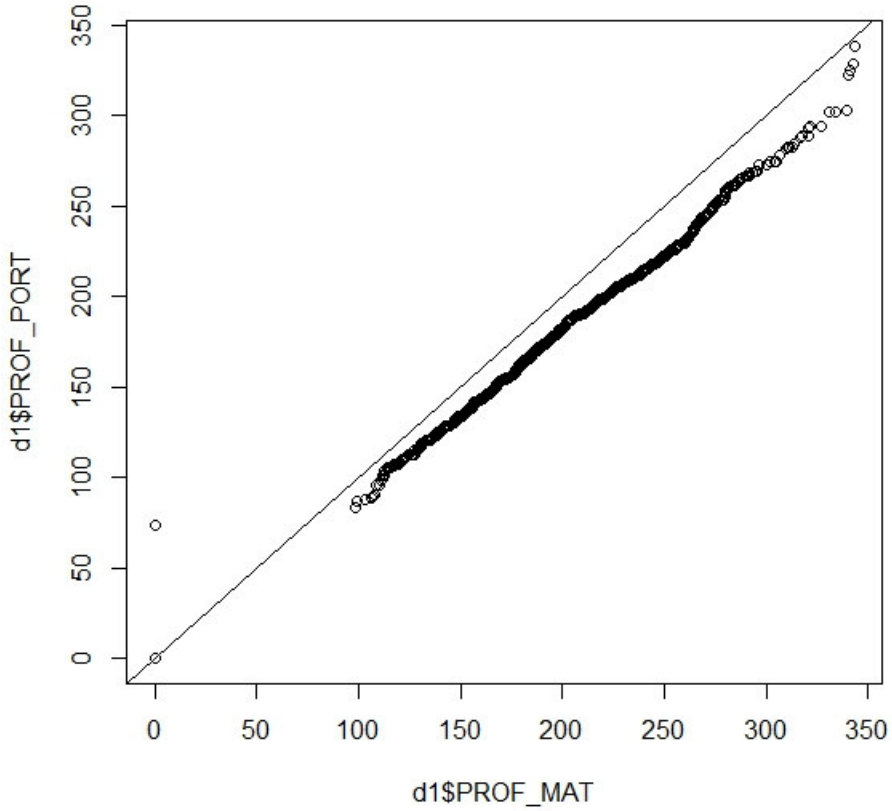
No R: Para adicionar a reta y = x no gráfico:

Comando: `abline(0,1)`

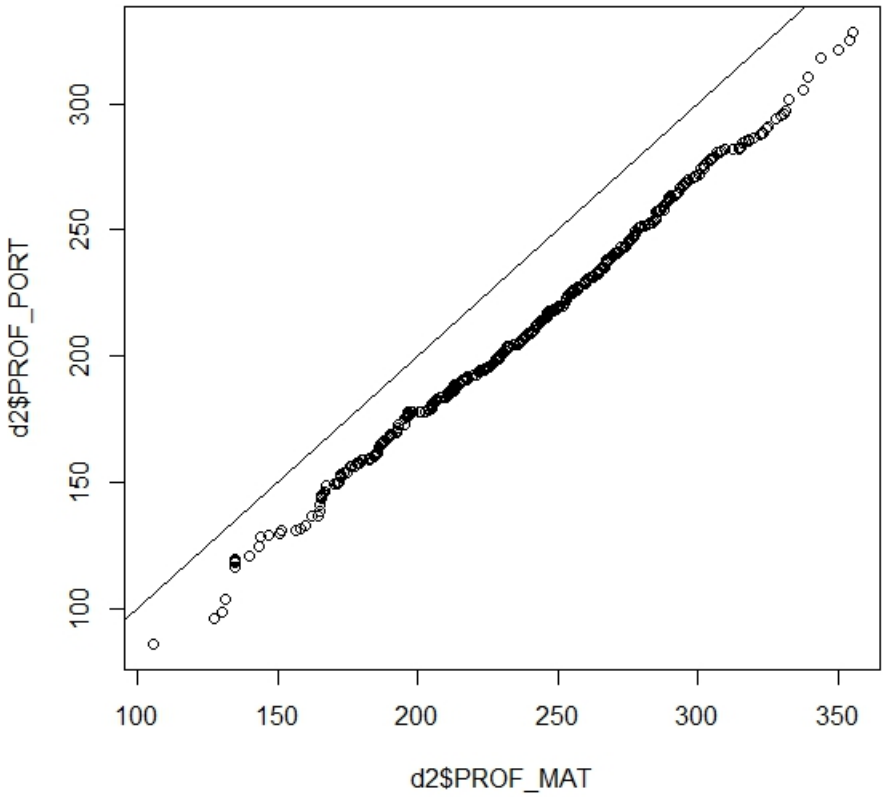


Exemplo 1.

Escola Estadual



Escola Municipal



Referências

- Slides da disciplina MAE116

- Bussab, W. e Morettin, P. (2010). Estatística Básica.

Saraiva, 6ª edição