

# INTRODUÇÃO AO MODELO DE REGRESSÃO LINEAR GERAL



# Introdução

- **Análise de Regressão:** obter uma equação que explique satisfatoriamente a relação entre uma variável resposta e uma ou mais variáveis explicativas.



Prever valores da variável de interesse

Uma análise de regressão linear pode ser:

- Simples

Uma única variável explicativa

└→ Temperatura

- Múltipla

Mais de uma variável explicativa

└→ Temperatura e concentração





# Modelo de Regressão Linear Simples (MRLS)

- Relação linear entre a variável dependente (Y) e a variável independente (X).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

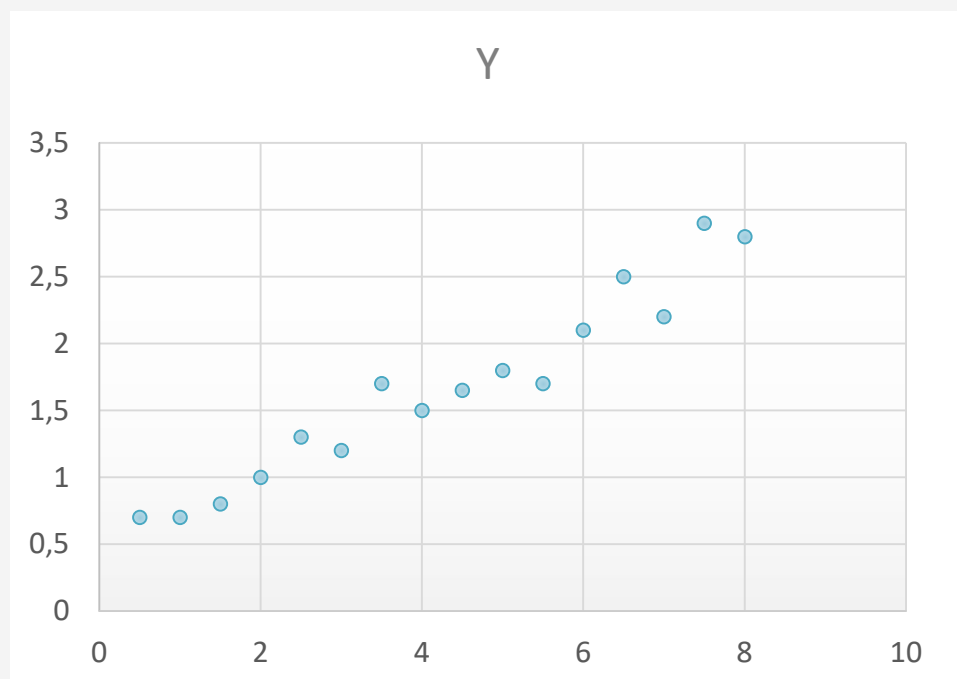


# Diagrama de Dispersão

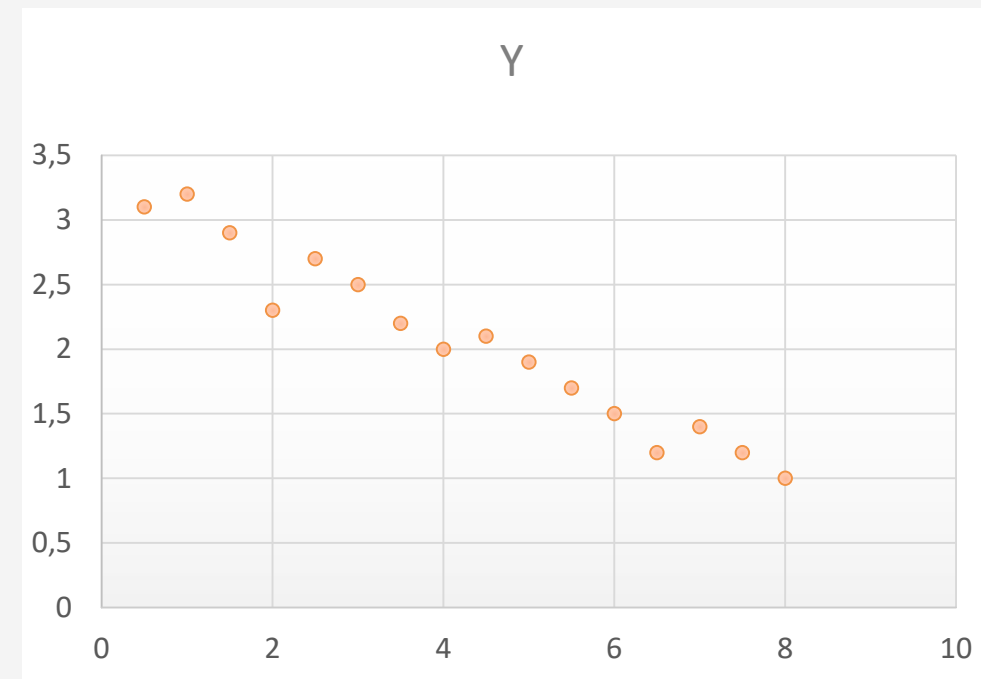
Dado um conjunto de dados com duas variáveis quantitativas  $X$  e  $Y$ , pode-se construir um diagrama de dispersão.

↳ Permite verificar se pode-se assumir um relacionamento linear entre as variáveis  $X$  e  $Y$ .

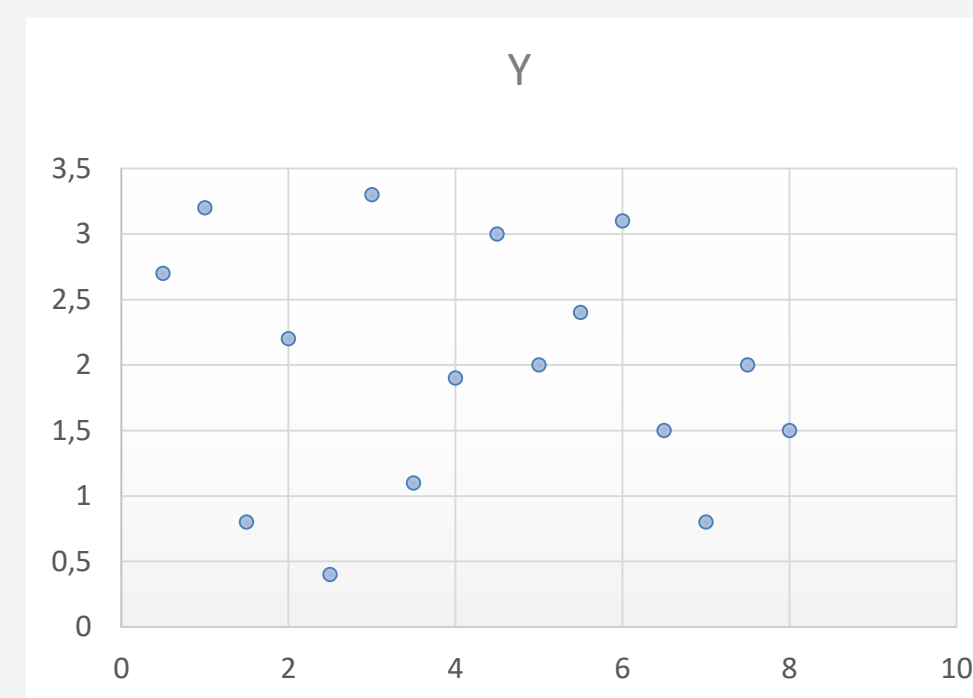
**Figura 1:** Correlação linear positiva



**Figura 2:** Correlação linear negativa



**Figura 3:** Correlação nula



- Caso se observe uma correlação linear entre as variáveis  $X$  e  $Y$ , podemos quantificar essa associação, para confirmar o que foi observado. Isso é feito pelo calculo do **coeficiente de correlação linear de Pearson ( $r$ )**.

O coeficiente de correlação linear é de tal forma que  $-1 \leq r \leq 1$ .

Dessa forma:

$$r < 0$$

Correlação linear negativa

---

$$r > 0$$

Correlação linear positiva

---

$|r|$  próximo de 1

Correlação linear forte

# Regressão Linear Simples

- Ao se confirmar uma relação linear forte, pode-se considerar uma **regressão linear simples** entre as variáveis, de forma a encontrar uma equação que explique de forma satisfatória essa relação. Essa equação terá a seguinte forma:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$




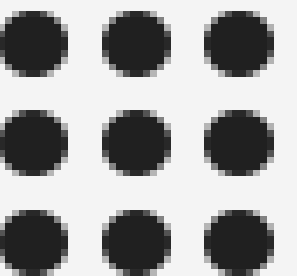
# Regressão Linear Simples

Em que:

$$\hat{\beta}_1 = \frac{S_{XX}}{S_{XY}} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- $\bar{x}$  e  $\bar{y}$  são as médias de  $x$  e  $y$ .
- $S_{XX} = \sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}$
- $S_{XY} = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}$

- 
- $\hat{\beta}_0$  representa o valor de Y quando  $X = 0$ ;
  - $\hat{\beta}_1$  representa o aumento de Y quando X é incrementado em uma unidade.





# Teste de significância da regressão

- Para testar a significância da regressão, deve-se testar as hipóteses:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

# Teste de significância da regressão

- O teste estatístico t é utilizado para testar essas hipóteses.
- Se a **hipótese nula** ( $H_0$ ) for **aceita**, a **regressão não é significativa**, e não há relação linear entre as variáveis.
- Ao contrário, se a **hipótese nula** for **rejeitada**, a **regressão é significativa**, e existe relação linear entre as variáveis.

# Teste de significância da regressão

- **Regra da decisão:** a um determinado nível de significância, a decisão do teste de hipóteses se dá por:

p-valor < nível de significância  $\rightarrow$  rejeitar  $H_0$

p-valor > nível de significância  $\rightarrow$  aceitar  $H_0$

# Coeficiente de determinação ( $R^2$ )

- Fornece a proporção da variação de Y explicada pela variável X. Quando  $R^2$  é próximo de 1, isso indica que a maior parte da variação de Y é explicada pelo modelo de regressão.

$$R^2 = \frac{S_{XY}^2}{S_{XX} \times S_{YY}}$$

**EXEMPLO**





# Exemplo

- Considere o conjunto de dados a seguir:

**Tabela 1:** quantidade de procaína hidrolisada em moles/litro no plasma humano em função do tempo em minutos.

Fonte: Aven e Foldes (1951).

Tempo (minutos)	Quantidade hidrolisada
2	3,5
3	5,7
5	9,9
8	16,3
10	19,3
12	25,7
14	28,2
15	32,6





# Exemplo

- Utilizando o R, vamos construir o diagrama de dispersão dos dados, calcularemos o coeficiente de correlação linear de Pearson e, em caso de uma associação linear forte entre as variáveis  $X$  e  $Y$ , encontraremos a equação de regressão que modela a associação entre elas.

# Exemplo

- Primeiro, deve-se entrar com os dados no R:

```
x <- c(2, 3, 5, 8, 10, 12, 14, 15)
```

```
y <- c(3.5, 5.7, 9.9, 16.3, 19.3, 25.7, 28.2, 32.6)
```

- Em seguida, vamos construir um diagrama de dispersão relacionando *x* e *y* com o comando:

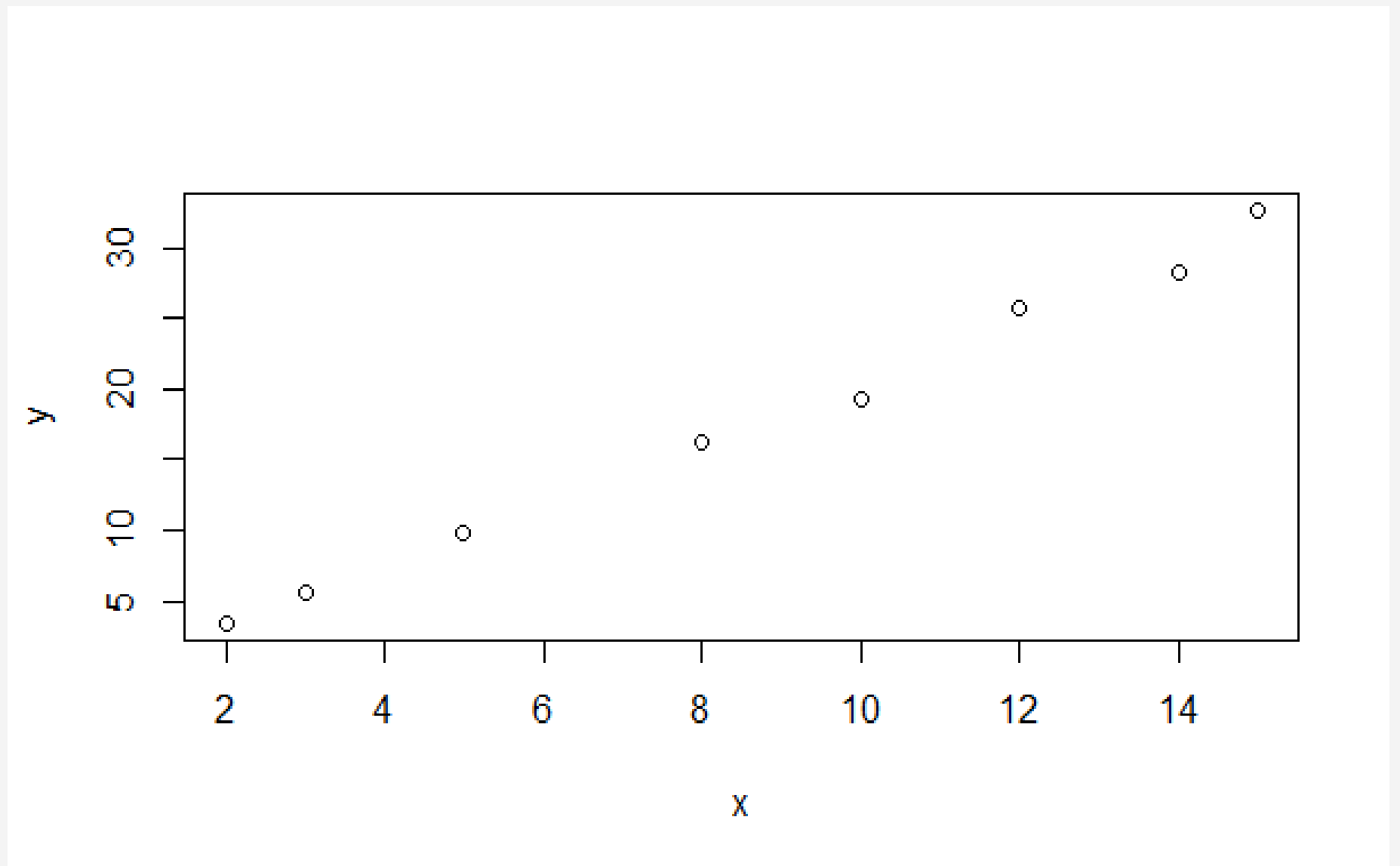
```
plot(x,y)
```

# Exemplo

- Parece haver uma associação linear positiva entre as variáveis.
- Para confirmar, vamos calcular o coeficiente de correlação linear de Pearson com o comando:

*cor(x,y)*

**Figura 4:** Diagrama de dispersão da quantidade de procaína hidrolisada em moles/litro no plasma humano (y) em função do tempo em minutos (x).



# Exemplo

- O valor encontrado para o coeficiente é de **0,9969554**, o que significa que existe de fato uma **correlação linear positiva forte** entre as variáveis  $X$  e  $Y$ , como havíamos suspeitado pelo diagrama de dispersão.

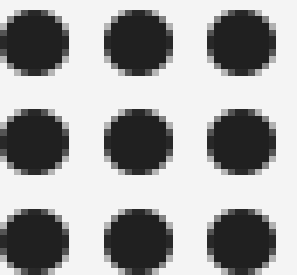




# Exemplo

- Vamos obter a equação linear estimada utilizando a função `lm()`. Para isso, precisamos construir uma tabela com os valores de `x` e `y`, utilizando a função:

```
dados <- data.frame(x,y)  
dados
```

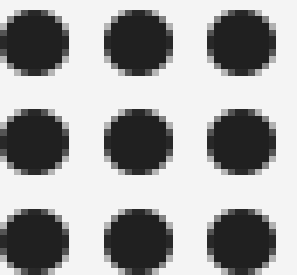




# Exemplo

> dados

	x	y
1	2	3.5
2	3	5.7
3	5	9.9
4	8	16.3
5	10	19.3
6	12	25.7
7	14	28.2
8	15	32.6



# Exemplo

- Usando a função `lm()`:

```
regressao <- lm(y ~ x, data = dados)
```

```
regressao
```



# Exemplo

> regressao

Call:

lm(formula = y ~ x, data = dados)

Coefficients:

(Intercept)	x
-0.985	2.161

No output do R são mostrados os valores dos coeficientes  $\hat{\beta}_0$  (Intercept) e  $\hat{\beta}_1$  (x).







# Exemplo

- Portanto, a equação de regressão estimada é:

$$\hat{y} = -0,985 + 2,161x$$

# Exemplo

- Faremos agora o teste de significância da regressão linear simples.
- As hipóteses testadas são da forma:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

# Exemplo

- Para encontrar o p-valor referente a este teste de hipóteses no R, utilizaremos o comando:

*summary(regressao)*

**> summary(regressao)**

**Call:**

**lm(formula = y ~ x, data = dados)**

**Residuals:**

Min	1Q	Median	3Q	Max
-1.3208	-0.2655	0.1230	0.3420	1.1763

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.98502	0.67549	-1.458	0.195
x	2.16058	0.06899	31.319	7.04e-08 ***

---

**Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 0.9044 on 6 degrees of freedom**

**Multiple R-squared: 0.9939, Adjusted R-squared: 0.9929**

**F-statistic: 980.9 on 1 and 6 DF, p-value: 7.039e-08**



# Exemplo

- Obtivemos o p-valor  $= 7,039 \times 10^{-8}$ . Considerando um nível de significância de 5%:
- $7,039 \times 10^{-8} < 0,05 \rightarrow$  Rejeita-se  $H_0$ , e aceita-se a hipótese  $H_1: \beta_1 \neq 0$ .
- Portanto, **a regressão é significativa**, e há uma relação linear entre as variáveis  $x$  e  $y$ , dada por:

$$\hat{y} = -0,985 + 2,161x$$



# Exemplo

- Uma vez ajustado o modelo, podemos verificar o valor do coeficiente de determinação  $R^2$ , de modo a encontrar qual a proporção da variabilidade de Y que é explicada por X. O  $R^2$  também é calculado pelo *summary(regressao)*.

**Multiple R-squared: 0.9939**

- Portanto, 99,39% da variabilidade de Y pode ser explicada por X, ou seja, pelo modelo de regressão ajustado.

# Exemplo

- Adicionando a equação de regressão e o  $R^2$  ao gráfico:

*plot(x,y)*

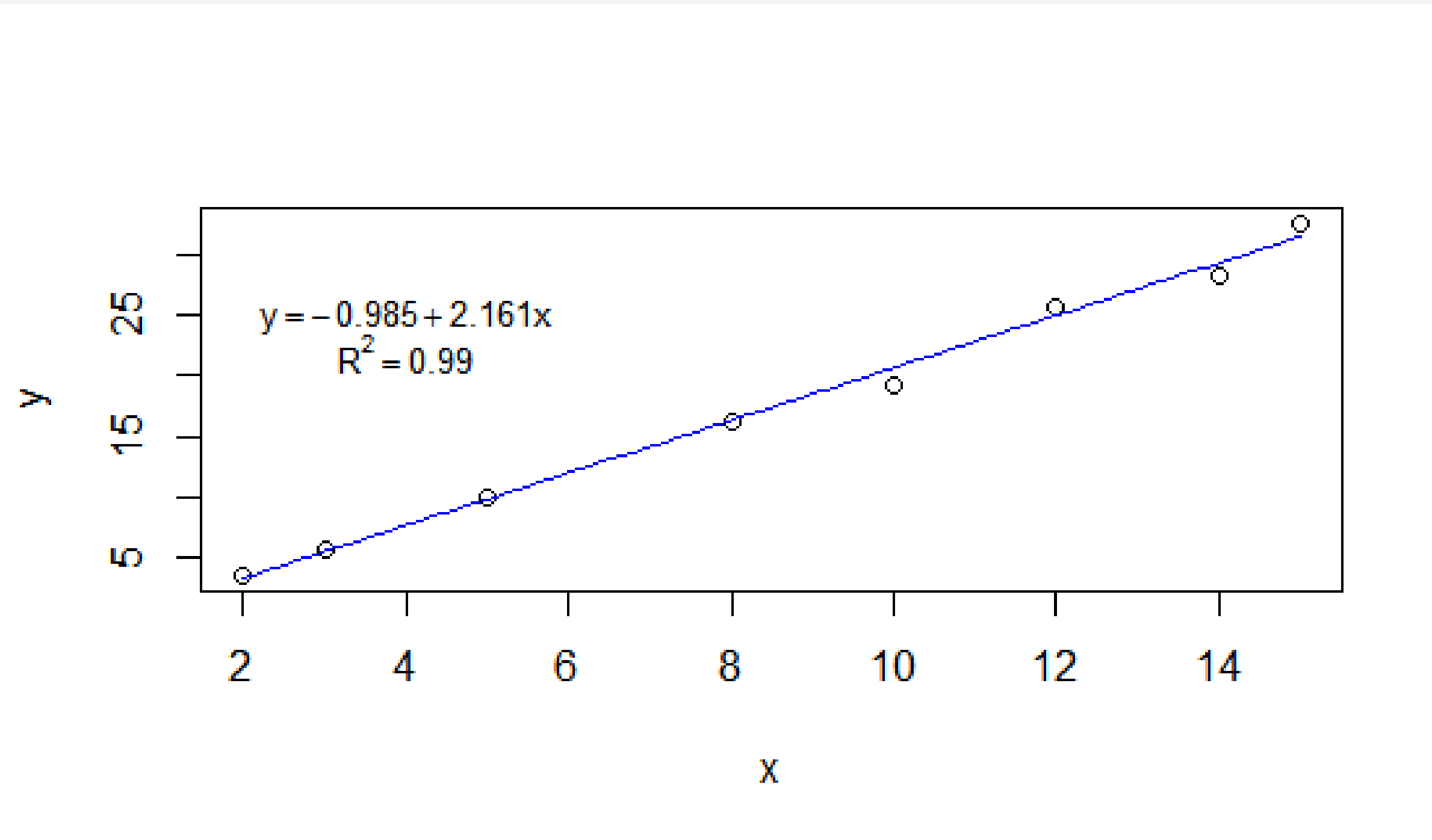
*curve(-0.985+2.161\*x, add = T, col = "blue")*

*text(4, 25, expression(y== -0.985+2.161\*x), cex = 0.8)*

*text(4, 22, expression(R^2==0.99), cex = 0.8)*

# Exemplo

**Figura 5:** Equação da regressão ajustada no diagrama de dispersão







# Dados de secagem

- O conjunto de dados que será analisado no trabalho se refere ao peso final após secagem de flores de *Viola × wittrockiana* Gams, conhecidas comumente como amor-perfeito, em relação ao tempo de secagem.



# Dados de secagem

tempo	pf
15	0.8797
15	0.8496
15	0.8696
30	0.799
30	0.7301
30	0.7659
45	0.7186
45	0.6554
45	0.6763
60	0.6404
60	0.5669
60	0.594
75	0.5043
75	0.4262
75	0.4104
90	0.4271
90	0.3626

Essas são as 17 primeiras linhas dos dados de secagem. Esse conjunto de dados possui no total 48 linhas.

- A primeira coluna indica o tempo de secagem em minutos, e a segunda o peso final da flor em gramas.



# Dados de secagem

- Para importar um arquivo do seu computador no R, como é o caso aqui, você pode fazer essa importação da seguinte maneira. Primeiro escolha o diretório, clicando em **Session** → **Set Working Directory** → **Choose Directory**, e escolhendo o diretório. Em seguida, utilizam-se os comandos:

```
dados<-read.table("nome_do_arquivo.txt", h=T)
```

```
attach(dados)
```

Com isso, os dados são armazenados no R.



# Dados de secagem

- Para facilitar na hora de escrever os comandos, vamos chamar o tempo de  $x$  e o pf de  $y$ . Para isso usamos:

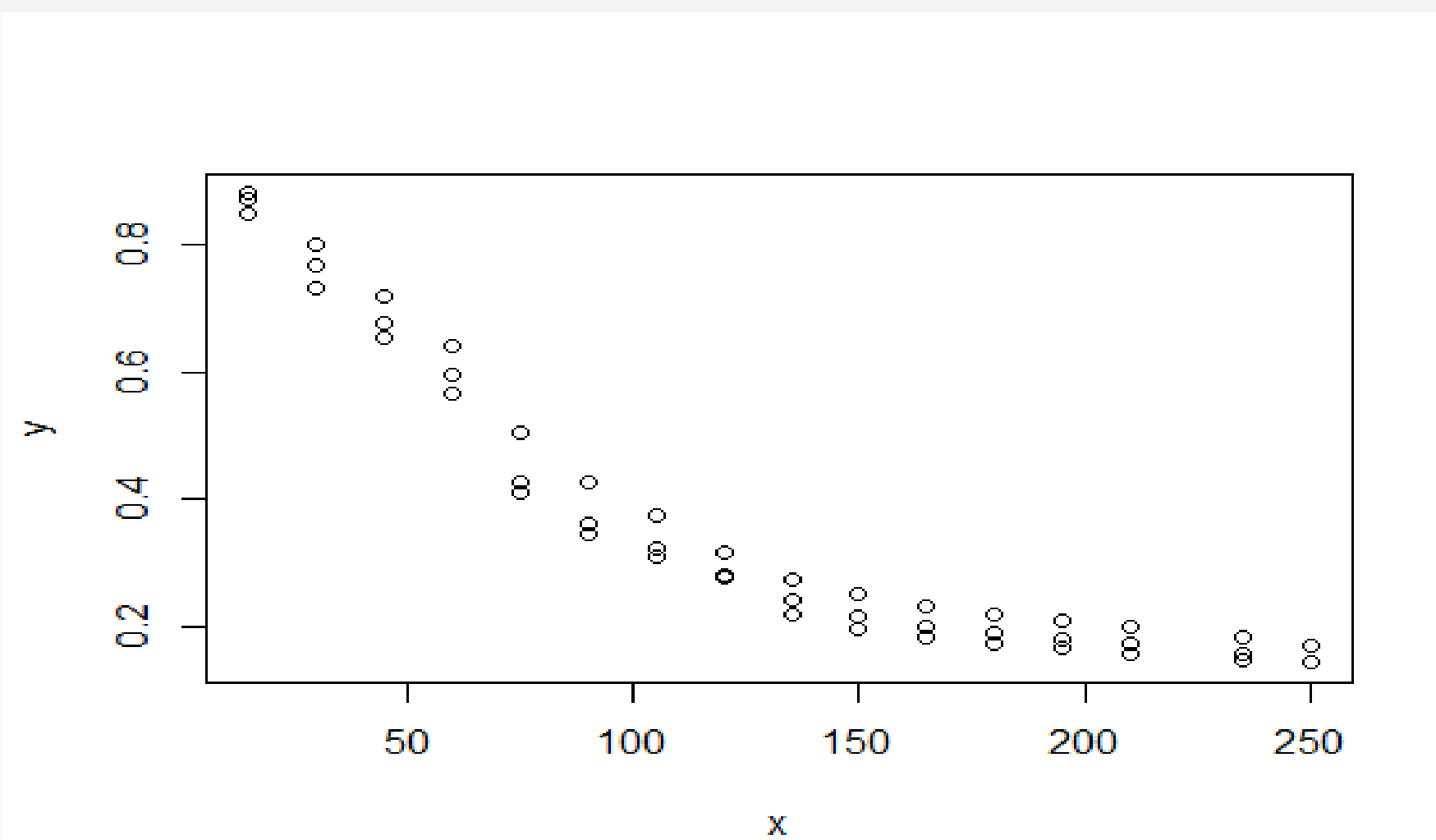
*$x \leftarrow tempo$*

*$y \leftarrow pf$*

# Dados de secagem

- Ao construir o diagrama de dispersão do tempo (x) em relação ao peso final (y), com o comando *plot(x,y)*, podemos observar que não parece haver uma relação linear entre as variáveis.

**Figura 6:** Diagrama de dispersão do peso final após secagem das flores (y) em relação ao tempo de secagem (x).





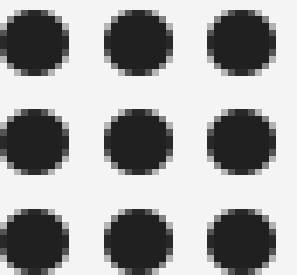
# Dados de secagem

- Para confirmar essa suspeita, vamos obter a equação de regressão linear utilizando a função `lm()`.

```
reg <- lm(y ~ x)  
summary(reg)
```

- E obtemos a equação da regressão estimada:

$$\hat{y} = 0,745 - 0,0029x$$



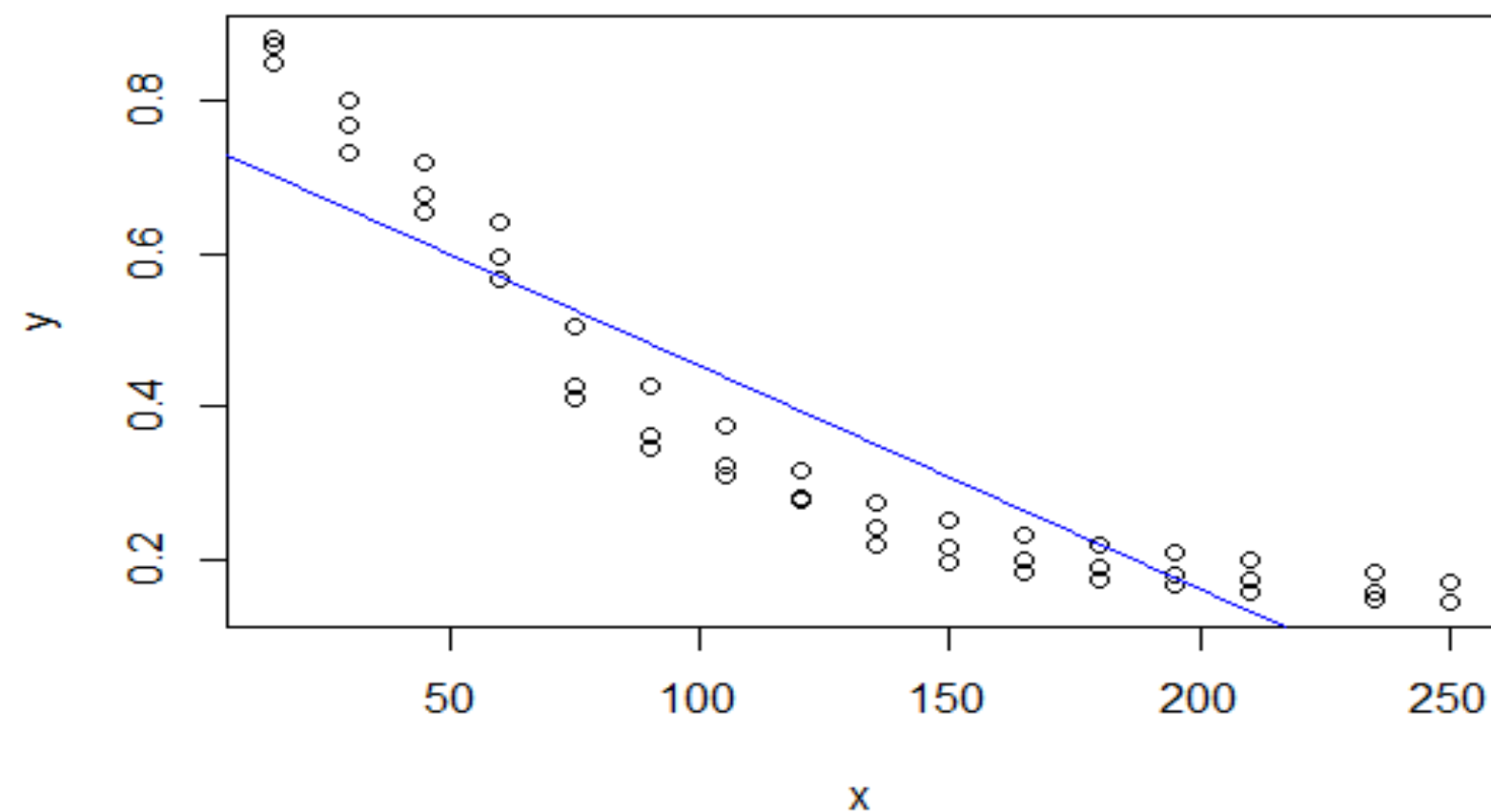
# Dados de secagem

- Adicionamos a equação de regressão obtida no gráfico de dispersão:

*`curve(0.7444486-0.0029114*x, add = T, col = "blue")`*

- Podemos perceber que o modelo não se ajusta bem ao conjunto de dados de secagem.

**Figura 7:** equação do modelo linear aplicada ao diagrama de dispersão, para os dados de secagem





# Dados de secagem

- Da mesma forma, vamos obter a equação do modelo de regressão quadrático. Para isso, vamos utilizar os comandos:

```
reg2 <- lm(y ~ x + I(x^2), data = dados)  
summary(reg2)
```

- E obtemos a equação estimada:

$$\hat{y} = 0,979 - 0,0079x + 0,00002x^2$$



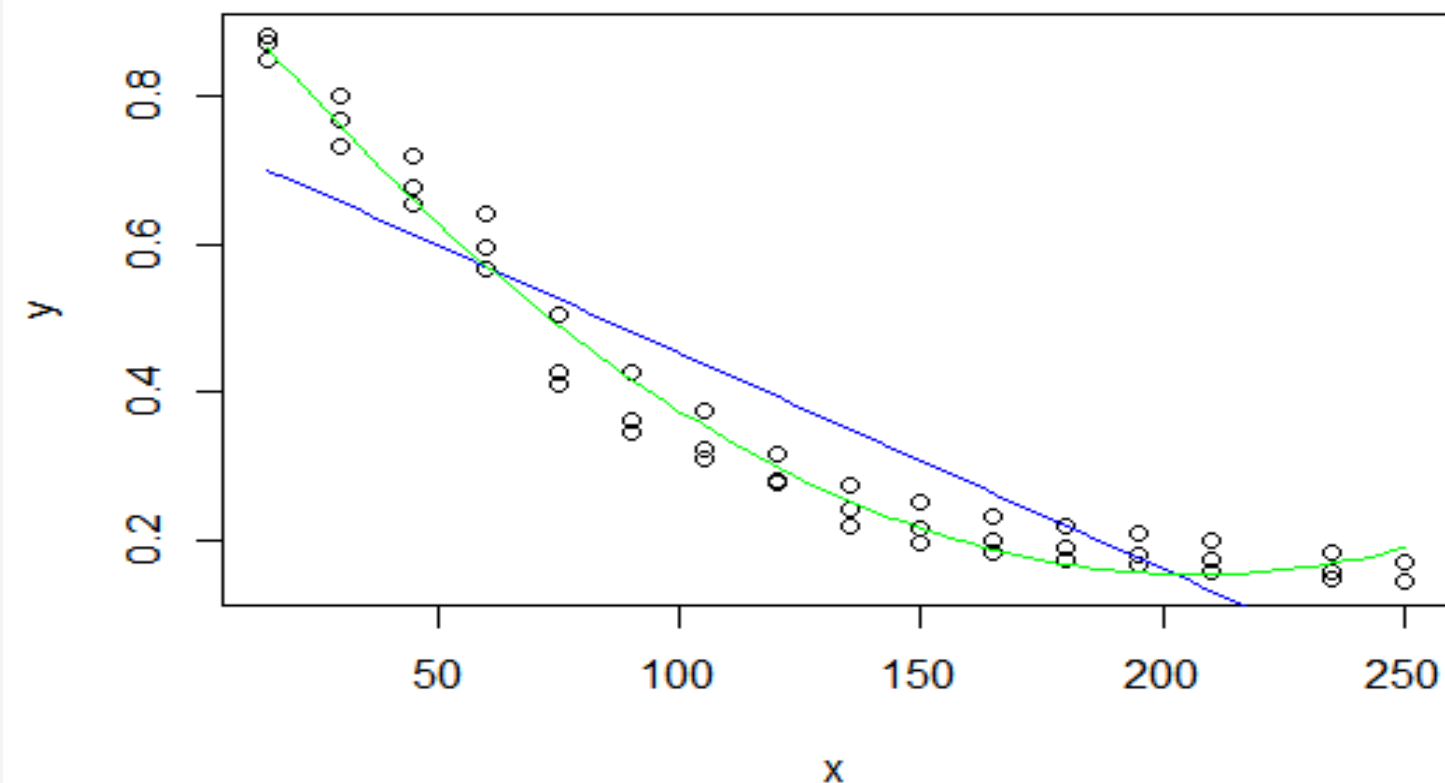

# Dados de secagem

- Adicionamos a equação de regressão obtida no gráfico com o comando:

*`curve(0.9786-0.007955*x+0.00001919*x^2, add = T, col = "green")`*

Esse modelo se ajusta melhor ao conjunto de dados. Porém, por ser uma parábola, a partir de certo ponto, os valores de y vão voltar a aumentar, e, portanto, não é adequado.

**Figura 8:** equação de modelo de regressão quadrático (verde) aplicado ao diagrama de dispersão dos dados de secagem



# Exemplo

- Portanto, podemos concluir que para o conjunto de dados de secagem apresentado, deve ser utilizado um **modelo de regressão não-linear**.