UNIVERSIDADE DA REGIÃO DE JOINVILLE - UNIVILLE

Bacharelado em Engenharia de Software (BES)

Estatística para computação

Professora Priscila Ferraz Franczak

Engenheira Ambiental - UNIVILLE

Mestre em Ciência e Engenharia de Materiais - UDESC

Doutora em Ciência e Engenharia de Materiais - UDESC

priscila.franczak@gmail.com

Plano de Aula

Regressão

1. Polinomial Simples

1.1 De Grau Igual a 1

1.2 De Grau Maior que 1

2. Polinomial Múltipla

3. Modelos Não Polinomiais

4. Exercícios

1. Polinomial Simples

1.1 De Grau Igual a 1

- A regressão linear simples tem como objetivo estimar uma equação que relacione matematicamente duas variáveis, sendo que uma delas é explicada pela outra.

- A variável explicada geralmente é denominada **variável resposta** ou **variável dependente (Y).**

- A **variável explicativa** é denominada **variável explanatória** ou **variável independente (X).**

- A equação que representa o modelo de regressão linear simples é:

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 X$$

- O "chapéu" sobre as letras indica que foi feita uma estimativa dos parâmetros do modelo com base em dados obtidos através de uma amostra.

$\hat{Y}$ = variável dependente

$\hat{\beta}_o$ = primeiro parâmetro da equação de regressão, indica o **intercepto** no eixo Y, ou seja, o valor de Y quando X = 0.

$\hat{\beta}_1$ = segundo parâmetro da equação de regressão, chamado **coeficiente angular**, que indica a inclinação da reta de regressão.

X = variável independente.

- A análise de regressão se distingue da correlação por supor uma **relação de causalidade entre as variáveis resposta e explanatória**.

- A análise geralmente se baseia numa referência teórica, que justifique uma relação matemática de causalidade.

- A estimativa dos parâmetros β_0 e β_1 do modelo se dá a partir das seguintes fórmulas:

$$\hat{\beta}_1 = \frac{n\sum XY - \sum X * \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$\bar{Y}$ = média dos valores de Y

$\bar{X}$ = média dos valores de X

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 \mathrm{X}$$

- Quando o diagrama de dispersão apresenta os pontos agrupados em torno de uma reta imaginária, provavelmente existe urna relação linear entre as variáveis envolvidas.

Exemplo: Um engenheiro civil coleta dados em um laboratório, a fim de estudar a dilatação de um pilar de concreto segundo a temperatura ambiente no local onde o pilar se encontra:

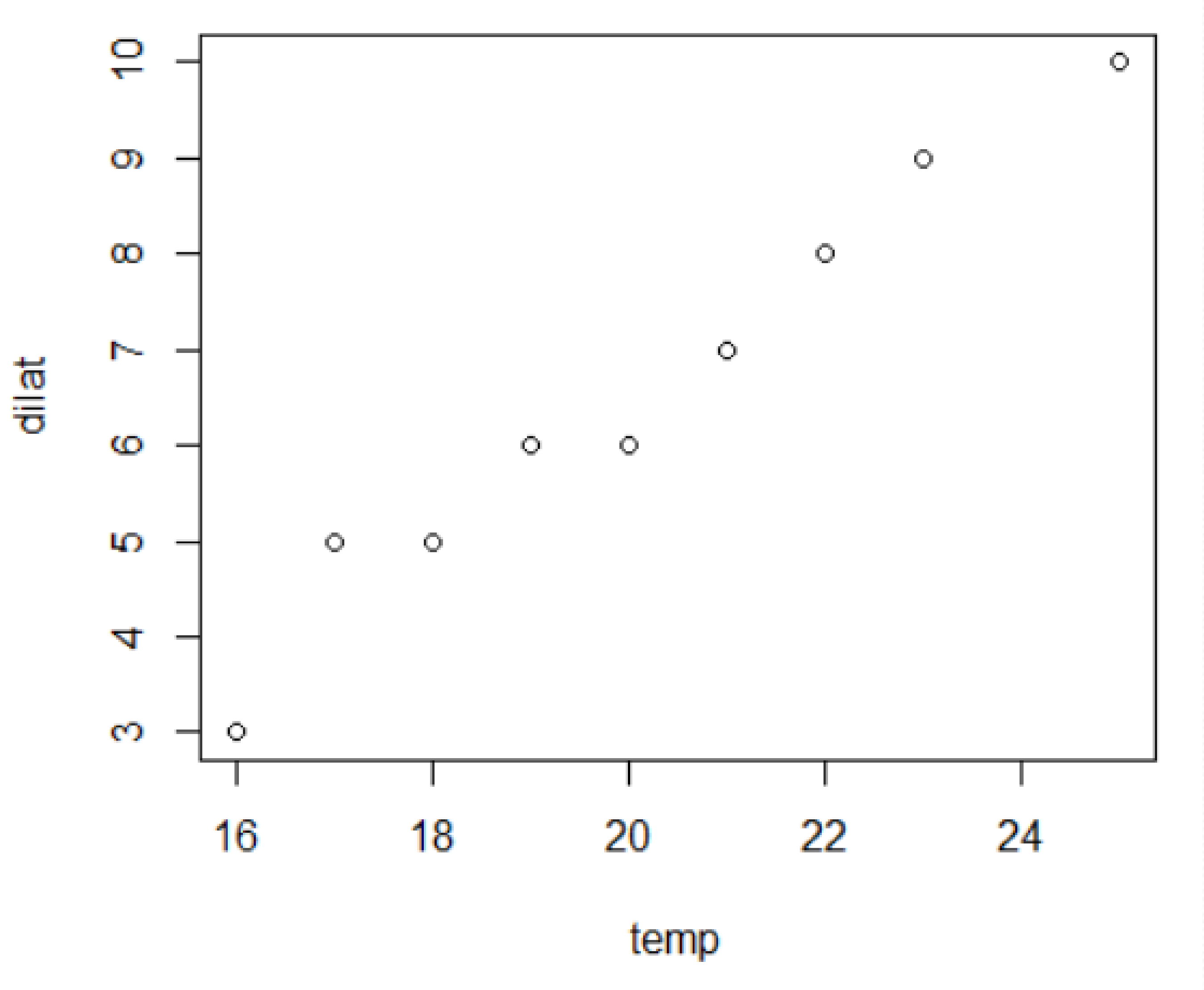
T(°C)	18	16	25	22	20	21	23	19	17
Dilat. Linear (mm)	5	3	10	8	6	7	9	6	5

Posso realizar um estudo de regressão nestes dados?
Qual modelo usar?
Como montar a equação que relaciona a temperatura com a dilatação neste estudo?
A temperatura realmente exerce influência na dilatação do pilar?
É possível quantificar essa relação?

- Cria-se o data.frame:

```
> temp<-c(18,16,25,22,20,21,23,19,17)
> dilat<-c(5,3,10,8,6,7,9,6,5)
> dados<-data.frame(dilat,temp)
> dados
  dilat temp
1     5   18
2     3   16
3    10   25
4     8   22
5     6   20
6     7   21
7     9   23
8     6   19
9     5   17
```

```
plot(temp,dilat)
```



- O diagrama sugere uma tendência linear dos dados.

- Faremos, portanto, um modelo de regressão linear simples (simples, pois existe apenas uma variável independente temp relacionada a variação da variável dependente dilat).

```
> reglin<-lm(dilat~temp,
+              dados)
> reglin

Call:
lm(formula = dilat ~ temp, data = dados)

Coefficients:
(Intercept)         temp
    -8.1710       0.7323
```

- Com base neste modelo ajustado, temos duas informações: o valor do intercepto (valor em que a reta de regressão intercepta o eixo das ordenadas) e o valor que representa o coeficiente de inclinação da reta, ou seja, a relação entre a dilatação e a temperatura (o quanto a dilatação varia para cada variação unitária da temperatura).

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 \mathrm{X}$$

$$dilat = -8{,}1710 + 0{,}7323.\mathrm{temp}$$

```
Coefficients:
(Intercept)         temp
    -8.1710       0.7323
```

- Com o comando predict () podemos obter os valores calculados de dilat, de acordo com o modelo ajustado, para os valores observados de temp.

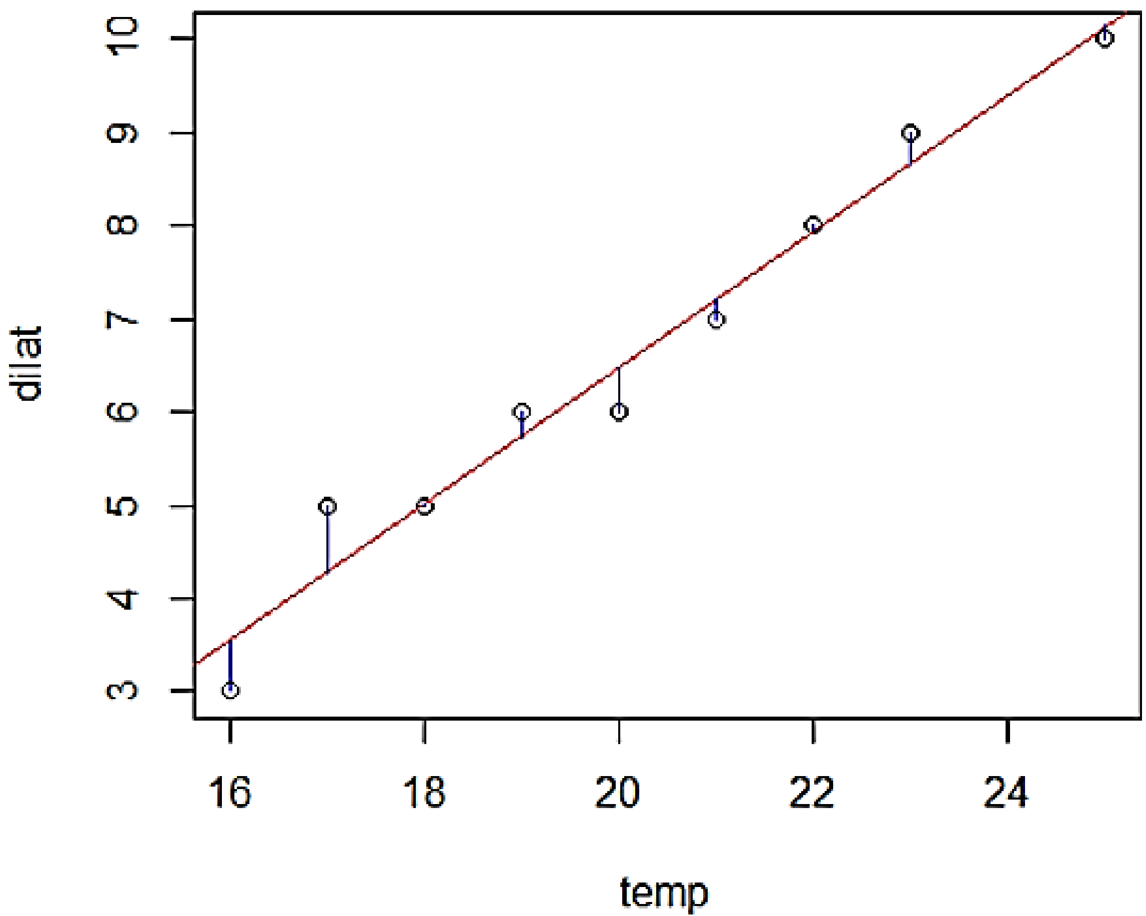
- Podemos também obter os resíduos associados a cada observação. Esses resíduos seriam simplesmente a diferença entre os valores observado e calculado correspondente a cada observação.

```
> predict(reglin)
        1         2         3         4 
 5.009677  3.545161 10.135484  7.938710 
        5         6         7         8 
 6.474194  7.206452  8.670968  5.741935 
        9 
 4.277419 
> resid(reglin)
           1            2            3 
-0.009677419 -0.545161290 -0.135483871 
           4            5            6 
 0.061290323 -0.474193548 -0.206451613 
           7            8            9 
 0.329032258  0.258064516  0.722580645 
```

```
> result<-data.frame(
+   dilat,
+   temp,
+   calculado=predict(reglin),
+   residuos=resid(reglin))
> result
  dilat temp calculado     residuos
1     5   18  5.009677 -0.009677419
2     3   16  3.545161 -0.545161290
3    10   25 10.135484 -0.135483871
4     8   22  7.938710  0.061290323
5     6   20  6.474194 -0.474193548
6     7   21  7.206452 -0.206451613
7     9   23  8.670968  0.329032258
8     6   19  5.741935  0.258064516
9     5   17  4.277419  0.722580645
```

- Agora vamos plotar novamente os dados e acrescentar ao gráfico, além da reta de regressão ajustada, segmentos de reta representando os resíduos, ou seja, segmentos de reta que vão dos valores observados (pontos) aos calculados (reta).

```
plot(temp,dilat)
abline(reglin,  #reta de regressão ajustada
       col=2)
segments(      #desenha segmentos de reta
  result$temp, #de (coord. x)
  result$dilat, #de (coord. y)
  result$temp,  #para (coord. x)
  result$calculado, #para (coord. y)
  col=4        #cor azul
)
```



- Podemos também realizar uma análise de variância da regressão da seguinte forma:

```
> anova(reglin)
Analysis of Variance Table

Response: dilat
          Df Sum Sq Mean Sq F value
temp       1 36.938  36.938   201.4
Residuals  7  1.284   0.183
             Pr(>F)
temp      2.048e-06 ***
Residuals
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1
```

- Por meio dessa análise podemos verificar que o coeficiente de x é significativo*(p-value* encontrado foi da ordem de 10^{-6}), ou seja, a temperatura influencia significativamente a dilatação.

- Com o comando summary ( ) podemos obter muitas outras informações:

```
> summary(reglin)

Call:
lm(formula = dilat ~ temp, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54516 -0.20645 -0.00968  0.25806  0.72258 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -8.1710     1.0475  -7.801 0.000107 ***
temp          0.7323     0.0516  14.191 2.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4283 on 7 degrees of freedom
Multiple R-squared:  0.9664,	Adjusted R-squared:  0.9616 
F-statistic: 201.4 on 1 and 7 DF,  p-value: 2.048e-06
```

- O valor do coeficiente de determinação (R^2) é apresentado em:

- **Multiple R-Squared**: 0,9664 e representa o quanto da variação da dilatação pode ser explicada pela variação da temperatura neste experimento.

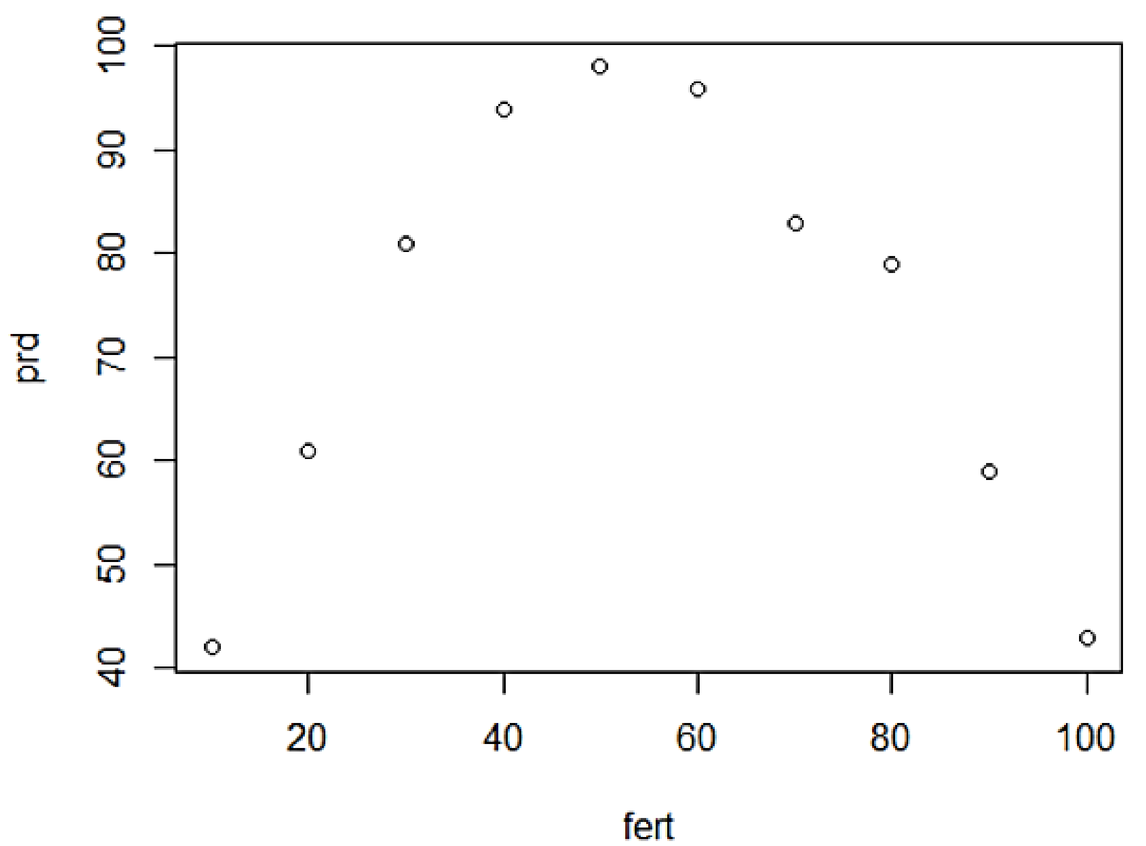
- Uma vez que o valor encontrado foi quase 97%, há indicação de que o modelo escolhido (linear) se ajusta bem aos dados.

1.2 De Grau Maior que 1

- Da mesma forma que o modelo linear ajustado, qualquer modelo de regressão polinomial pode ser obtido com o comando lm( ),que vem do inglês *linear models.*

- **Exemplo**: Os dados a seguir referem-se a produção de certa variedade de grãos (prd) em relação a quantidade de fertilizante aplicado na lavoura (fert):

```
> fert<-c(10,20,30,40,50,60,70,80,90,100) #var. indepen.
> prd<-c(42,61,81,94,98,96,83,79,59,43) #var. depend.
> plot(fert,prd)
```



- Pelo diagrama de dispersão, observa-se uma tendência quadrática nos dados.

- Dentro do comando lm ( ), observamos a necessidade de usarmos, como parte do modelo, o comando I( ).

- Esse comando permite inserirmos diretamente, no modelo, termos do tipo x ^ 2.

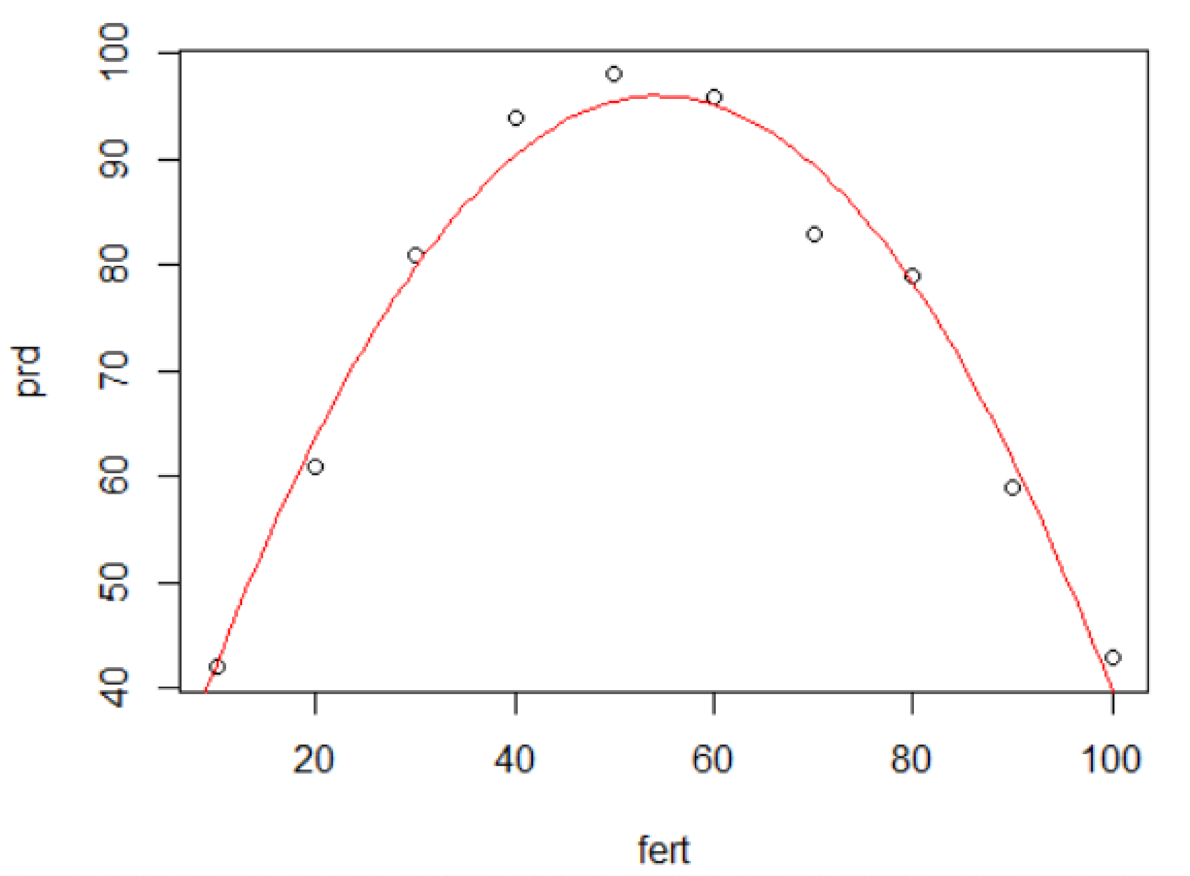
```
> reg<-lm(                    #ajusta uma regressão
+   prd~fert+I(fert^2)) #modelo quadrático
> reg                         #exibe o modelo ajustado

Call:
lm(formula = prd ~ fert + I(fert^2))

Coefficients:
(Intercept)         fert    I(fert^2)
   15.51667      2.95720     -0.02716
```

c b a

$$y = ax^2+bx+c$$

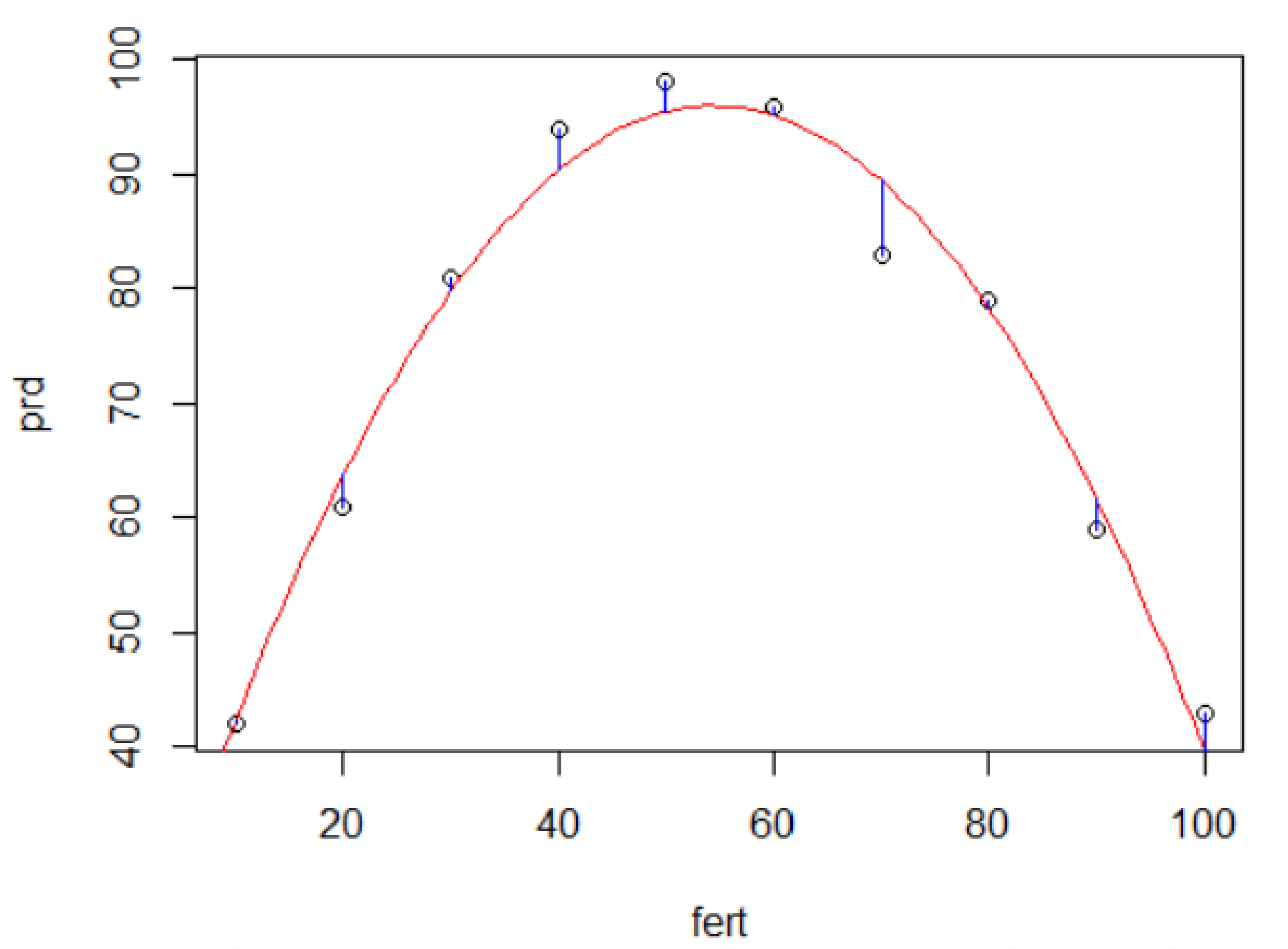
- Para acrescentar a curva ajustada no gráfico anterior...

```
curve(15.51667+2.95720*x-0.02716*x*x, #equação
      0,100,  #limites do eixo x
      col=2,  #curva em vermelho
      add = T)  #adicionar ao gráfico existente)
```



- E acrescentar os segmentos de retas que representam os resíduos:

```
segments(           #desenha segmentos de reta
  fert,             #de (coord. x)
  prd,              #de (coord. y)
  fert,             #para (coord. x)
  predict(reg),     #para (coord. y)
  col=4
)
```



prd
100
90
80
70
60
50
40
20
40
60
80
100
fert

2. Polinomial Múltipla

- Na regressão múltipla, uma variável resposta se relaciona a duas ou mais variáveis explanatórias.

- O objetivo também é predizer os valores de Y com base nas variáveis explanatórias.

Exemplos:

Para prever o preço de revenda de um automóvel, o analista de dados pode utilizar diversas variáveis, como:

- idade,
- número de quilômetros rodados,
- presença de vidros elétricos,
- presença de ar condicionado,
- consumo de combustível na estrada,
- consumo de combustível na cidade,
- estado de conservação dos pneus,
- estado de conservação da pintura, etc.

- Modelos de regressão ajudam na decisão dos bancos sobre conceder ou não um empréstimo para determinado candidato.

- Para isso, o banco geralmente levanta diversas variáveis para estimar a probabilidade de o cliente ser ou não um bom pagador.

- Na maioria das vezes, uma variável resposta se relaciona a mais de uma variável explanatória.

- Nessa situação, também podemos utilizar o método dos mínimos quadrados para obter uma equação que relacione as variáveis.

- Nesse caso, temos uma regressão múltipla.

- **Exemplo**: Consideramos que se queira ajustar uma superfície de resposta, ou de tendência - uma equação de regressão polinomial de grau 2, que descreva o comportamento das coordenadas de pontos que representam o relevo de determinado local.

- As coordenadas são dadas nos eixos cartesianos (x, y, z), em que z é a cota do ponto.

- Estamos supondo que z seja função de x e y.

- Um modelo polinomial de 2º grau tem a forma:

$$z = \beta_o + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2 + \varepsilon$$

```
x<-rep(1:10,10)
y<-as.numeric(gl(10,10))
set.seed(1234)
z<-rnorm(100,30,15)
coord<-data.frame(x,y,z)
```

```
> coord
     x  y         z
1    1  1 11.894014
2    2  1 34.161439
3    3  1 46.266618
4    4  1 -5.185466
5    5  1 36.436870
6    6  1 37.590838
7    7  1 21.378901
8    8  1 21.800522
9    9  1 21.533220
10  10  1 16.649433
```

…

- Agora vamos montar o modelo:

```
> modelo<-z~x+y+I(x^2)+I(x*y)+I(y^2)
> ajuste<-lm(     #ajusta a regressão
+   modelo,       #modelo utilizado
+    coord)       #conjunto de dados
> ajuste

Call:
lm(formula = modelo, data = coord)

Coefficients:
(Intercept)            x            y       I(x^2)
   31.60876     -1.93079     -2.11543      0.14627
   I(x * y)       I(y^2)
    0.06096      0.28101
```

3. Modelos Não Polinomiais

- Apesar de os modelos polinomiais serem úteis em muitas situações, há casos em que a disposição dos pontos no diagrama de dispersão, ou mesmo o problema do qual os dados foram obtidos, indique a necessidade ou exigência de modelos mais específicos.

Modelo Exponencial

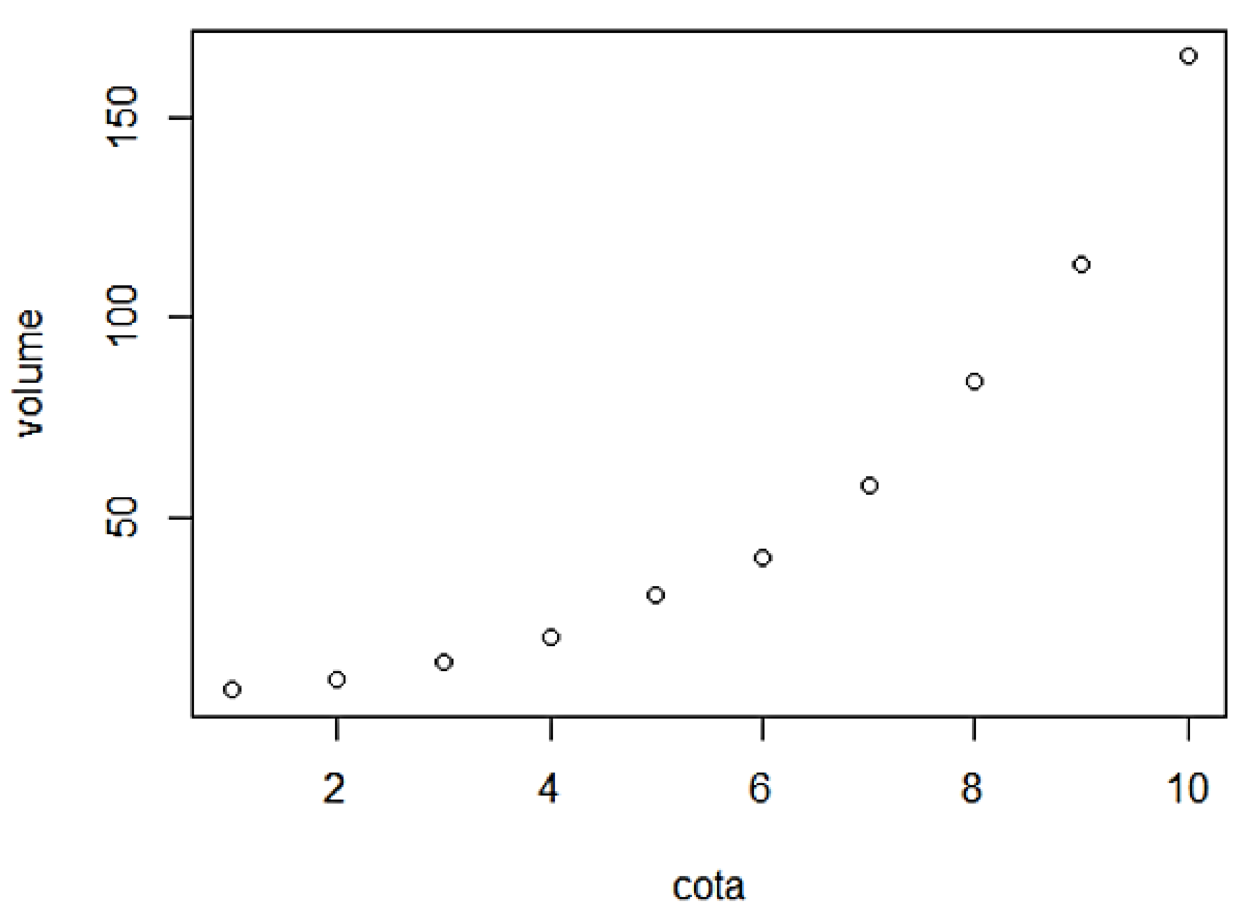
Exemplo:

- Num projeto de construção de urna barragem é de grande interesse equacionar a relação entre a cota do nível de água e o volume armazenado quando esta cota é atingida.

- Essa relação é obtida a partir de um diagrama cota-volume, estimado através do levantamento topográfico, com suas respectivas curvas de nível, da região onde será construída a barragem.

Suponha os dados a seguir, com a, cota dada em metros e o volume em quilômetros cúbicos:

```
cota<-c(1,2,3,4,5,6,7,8,9,10)
volume<-c(7,10,14,20,31,40,58,84,113,165)
dados<-data.frame(cota,volume)
plot(dados)
```



- Apesar de ser possível ajustar um modelo polinomial para os dados em questão, um modelo mais apropriado seria baseado na função $y = a.e^{b.x}$

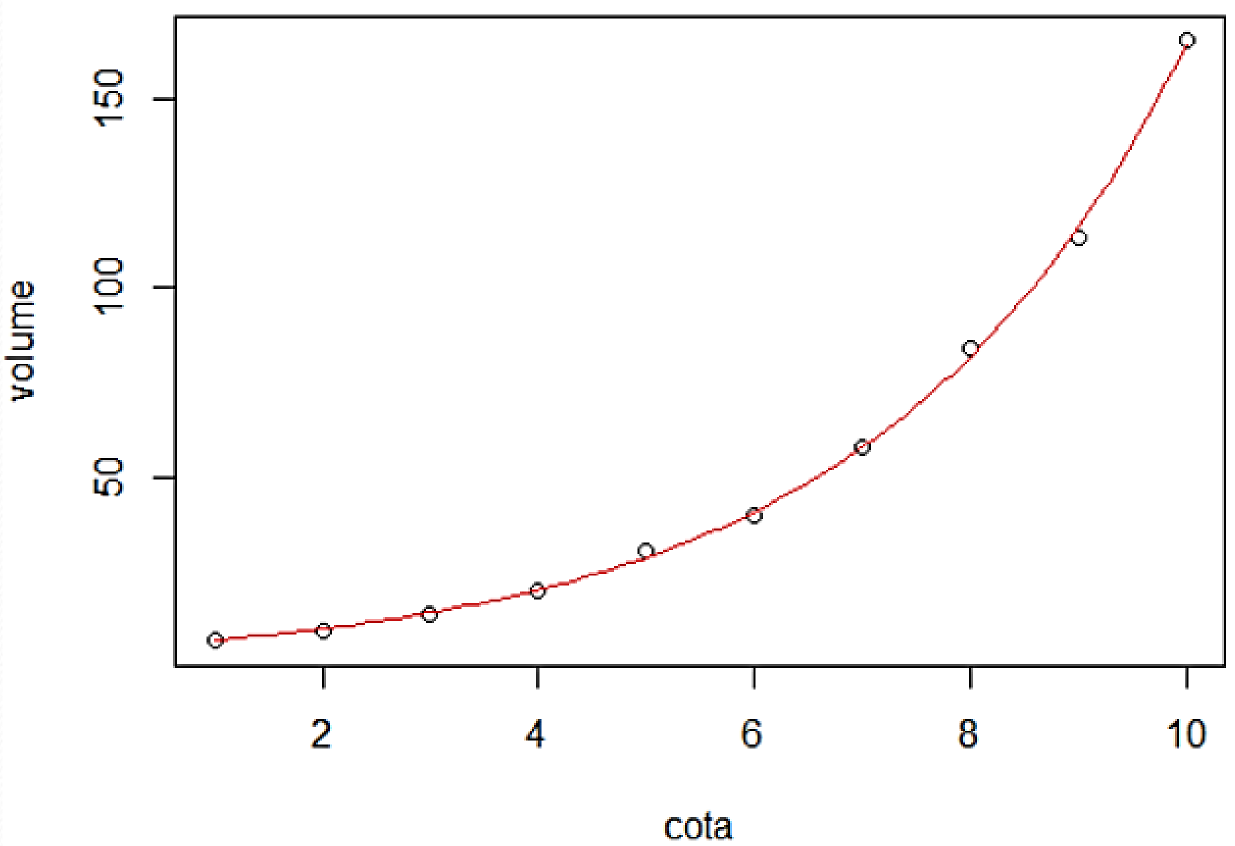
```
funcao<-volume~a*exp(b*cota) #modelo
exponencial<-nls(             #ajust. modelos não lineares
  funcao,                     #modelo a ajustar
  dados,                      #conjunto de dados
  start = c(a=1,b=1))         #valores iniciais dos estimadores
exponencial
```

```
Nonlinear regression model
  model: volume ~ a * exp(b * cota)
   data: dados
     a      b
5.1164 0.3467
 residual sum-of-squares: 19.44

Number of iterations to convergence: 14
Achieved convergence tolerance: 2.694e-07
```

- A obtenção de estimativa dos mínimos quadrados dos coeficientes se dá através do comando nls ( ), do inglês *Nonlinear Least Squares.*
- Comando ??regression no console abre mais opções de modelo de regressão.

- Desenhando a curva ajustada:

```
curve(                       #desenha a curva
  5.1164*exp(0.3467*x),      #equação ajustada
  1,                         #lim. inf. do eixo das abcissas
  10,                        #limite superior
  add = T,                   #acrescentar no gráfico anterior
  col=2)                     #cor vermelha
```



4. Exercícios