

Análise de Regressão

Modelos de Regressão são utilizados para descrever o relacionamento de uma variável y com outra (ou outras) variável x , por meio de uma relação matemática da forma

$$y = f(x; \beta) + \text{erro}.$$

Quando a função f é do tipo

$$f(x; \beta) = \beta_0 + \beta_1 x,$$

$\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$, tem-se um modelo de regressão linear simples. A variável x é a variável independente do modelo, enquanto y depende das variações de x , e é chamada de variável resposta. Assim, o modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis, x e y . O modelo de regressão é múltiplo quando envolve uma relação causal entre mais de duas variáveis. Ou seja, quando a variação da resposta y pode ser explicada por mais de uma variável independente, x_1, \dots, x_p , que são também denominadas variáveis explicativas ou covariáveis.

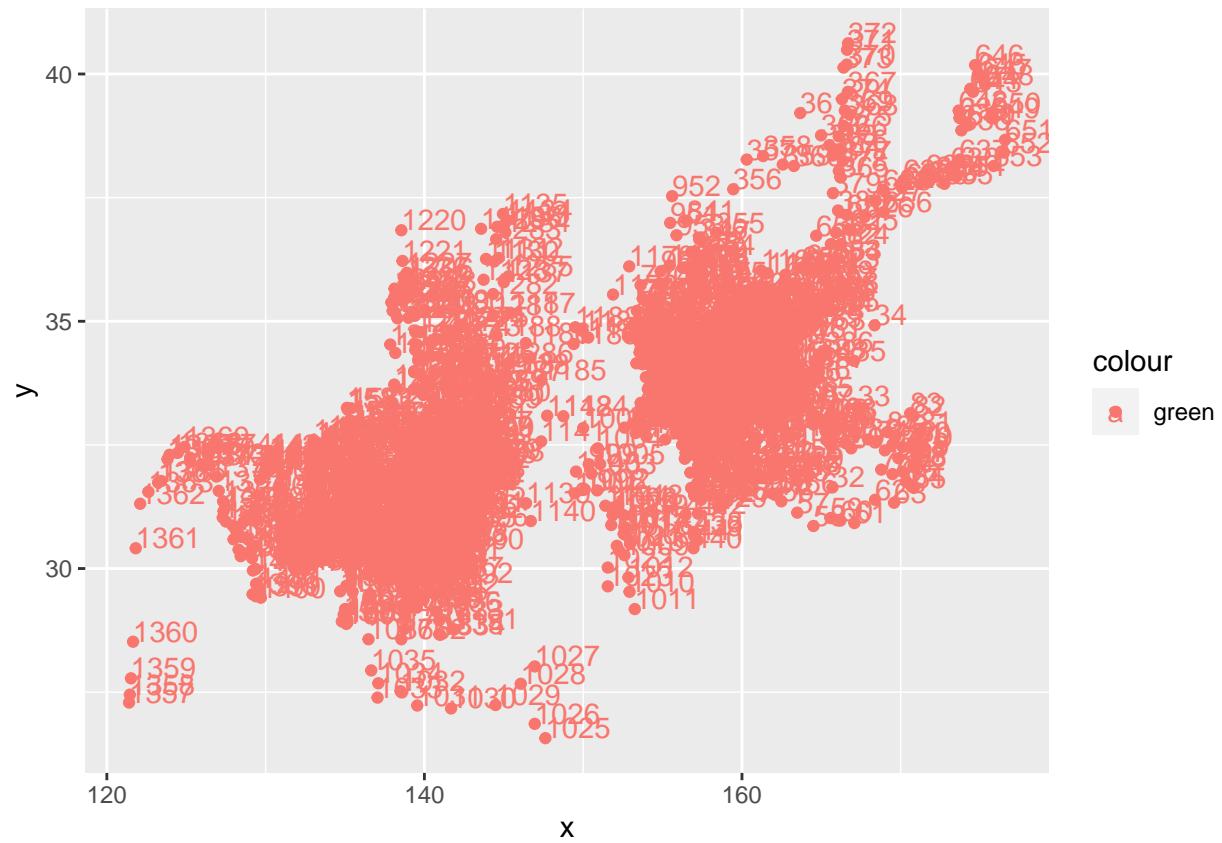
Modelos de regressão podem ser aplicados em vários tipos de probelmas.

- 1 - Em problemas em que se deseja realizar previsões sobre o comportamento futuro de algum fenômeno, extrapolando-se para o futuro as relações de causa e efeito observados no passado.
- 2 - Quando é desejado observar efeitos causados por uma variável x sobre outra variável (sobre a variável resposta) em decorrência de alterações introduzidas em seus valores.

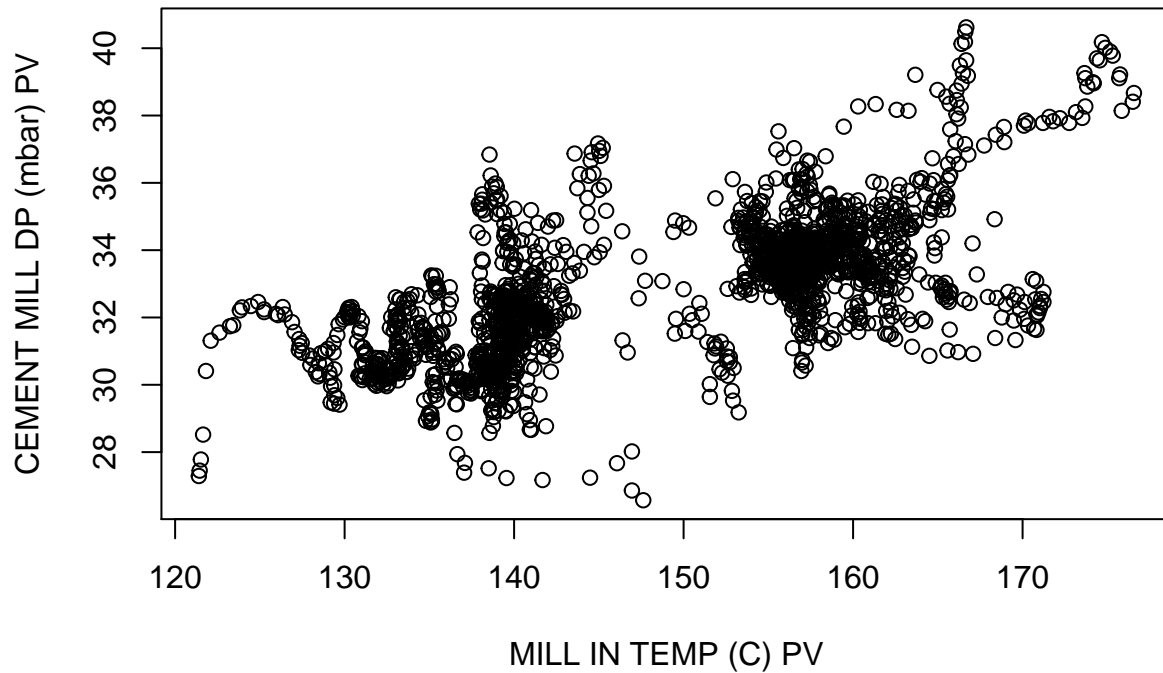
Modelos de Regressão Linear Simples

Considere a variável Diferencial de Pressão e a variável Temperatura Interna dos dados do moinho de cimento. Vamos assumir que y represente a variável Diferencial de Pressão (variável resposta) e x represente a variável Temperatura Interna do moinho (variável independente). Vamos considerar os valores das variáveis observados em um período de um dia (28/01/2019), considerando intervalos entre as coletas de três horas (3h). Essas dados são coletados a cada 30min, no entanto esses valores podem conter uma dependência ao longo do tempo, que pode ser amenizada considerando um intervalo maior entre as coletas.

```
lab=which(x==x)
ggplot(df, aes(x= x, y= y, colour="green", label=lab))+
  geom_point() +geom_text(aes(label=lab),hjust=0, vjust=0)
```



```
plot(x,y, xlab="MILL IN TEMP (C) PV", ylab="CEMENT MILL DP (mbar) PV")
```



Note que parece existir uma associação entre as a variáveis, pois a medida que x aumenta, a variável y parece tender a também aumentar. Se a relação existe, em geral, é desejado saber qual é a função que pode descrever o relacionamento. Neste caso, pode-se fazer a suposição inicial de que esta função seja uma reta, ou seja, pode-se supor que um modelo de regressão linear seja apropriado. Assim, o interesse é aproximar os dados a melhor reta possível, para tentar prever o comportamento da variável x em função de y.

Para mostrar a função da reta que melhor decreve os dados deve-se fazer a aproximação dos betas, que seguem as seguintes funções:

$$\hat{\beta}_0 = \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_0 \cdot \bar{x}$$

Para aproximação do β_1 será necessário o valor das medias amostrais, que serão expressa pelas seguintes fórmulas:

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\bar{y} = \frac{\sum y_i}{n}$$

Modelos de Regressão Linear Múltiplo

O modelo de regressão linear múltiplo é expresso pela função linear:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

em que:

y é a variável resposta (variável dependente no modelo)

x_1, \dots, x_p são as covariáveis (variáveis independentes ou explicativas), supostamente independentes entre si;
 β_0, \dots, β_p são os coeficientes da regressão;
 ε é o erro aleatório.

Agora, considere uma amostra y_1, \dots, y_n de y em que cada y_i está associado às p variáveis explicativas, $x_i, x_{i1}, \dots, x_{ip}$, $i = 1, \dots, n$, assim pelo modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i,$$

$i = 1, \dots, n$ e $n > p$, que pode ser visto como um modelo de regressão linear amostral.

No modelo de regressão usual, os ε_i 's são variáveis aleatórias sujeitas as seguintes condições:

- $E[\varepsilon_i] = 0$;
- $Var(\varepsilon_i) = \sigma^2$;
- $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j, j = 1, \dots, n$.

Note que, x_{i1}, \dots, x_{ip} são variáveis numéricas, não são variáveis aleatórias. No entanto, cada y_i depende da quantidade aleatória ε_i e portanto é uma variável aleatória. Assim, a média de y_i é dada por:

$$E[y_i] = E[\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

e a variância é dada por:

$$Var(y_i) = Var(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i) = \sigma^2, \quad \forall i.$$

Usando a notação matricial, o modelo é dado por:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \Rightarrow \mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

A matriz X de dimensão $n \times (p+1)$ e chamada de matriz de regressão quando $rank[X] = p+1$, e é chamada de matriz de delineamento quando $rank[X] = r < p+1$. A coluna de 1's na matriz refere-se ao intercepto, β_0 . O vetor Y de dimensão $n \times 1$ contém as variáveis y_1, \dots, y_n , β é o vetor de dimensão $(p+1) \times 1$ dos coeficientes de regressão e ε é o vetor de erros aleatórios de dimensão $n \times 1$.

Conseqüentemente,

$$E[\mathbf{Y}] = E[\mathbf{X}\beta + \varepsilon]$$

e

$$Var(\mathbf{Y}) = Var(\mathbf{X}\beta + \varepsilon) = \sigma^2 I_n,$$

em que I_n é a matriz identidade de dimensão $n \times n$.

Estimação por mínimos quadrados

O método de mínimos quadrados (*MMQ*) aplica-se somente aos parâmetros $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, e é frequentemente aplicado em situações em que não se dispõe de mais especificações, além das que já foram feitas, sobre os erros. Este método consiste em estimar $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ por $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ de modo que o vetor de valor esperado $E[\mathbf{Y}] = \mathbf{X}\beta$ esteja tão perto quanto possível do vetor de observações \mathbf{y} de \mathbf{Y} . Ou seja, os estimadores de mínimos quadrados de $\beta_0, \beta_1, \dots, \beta_p$ devem minimizar a soma dos quadrados dos erros, dada por:

$$U(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 = \varepsilon^T \varepsilon$$

Nota que $U(\beta)$ pode ser expresso por:

$$U(\beta) = \mathbf{Y}^T \mathbf{Y} - \beta^T X^T \mathbf{Y} - \mathbf{Y}^T X \beta + \beta^T X^T X \beta = \mathbf{Y}^T \mathbf{Y} - 2\beta^T X^T \mathbf{Y} + \beta^T X^T X \beta$$

Derivando $U(\beta)$ com respeito a β e igualando a zero, temos:

$$\frac{\partial U(\beta)}{\partial \beta} = -2X^T \mathbf{Y} + 2X^T X \hat{\beta} = 0,$$

que resulta em:

$$X^T X \beta = X^T \mathbf{Y},$$

denominadas equações normais.

Se $\text{rank}[X] = p + 1$, $X^T X$ é positiva definida e, portanto, inversível (não-singular). Assim, as equações normais possuem uma solução única dada por:

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y},$$

em que $\hat{\beta}$ é o estimador de mínimos quadrados (*EMQ*) de β .

O modelo de regressão ajustado, correspondente ao vetor, \mathbf{Y} , é dado por:

$$\hat{\mathbf{Y}} = X \hat{\beta} = X(X^T X)^{-1} X^T \mathbf{Y} = H \mathbf{Y}.$$

A matriz $H = X(X^T X)^{-1} X^T$ de dimensão $n \times n$ é geralmente chamada de matriz chapéu. Essa matriz possui algumas propriedades importantes que são enunciadas no teorema a seguir.

Teorema 3.1: Suponha que X é uma matriz $n \times (p + 1)$ de *rank* completo $p + 1$. Então,

- H e $(I_n - H)$ são simétricas e idempotente;
- $\text{rank}[I_n - H] = \text{Tr}[I_n - H] = n - (p + 1) = n - p - 1$;
- $HX = X$.