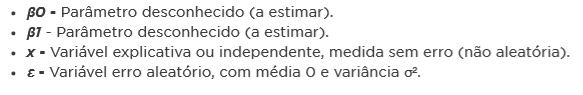
**Ciência de Dados - Análise de Regressão Linear    
Aplicado ao Consumo de Oxigênio relacionado a Distância Percorrida**

Marcio Ortolan   
Analista de Sistema

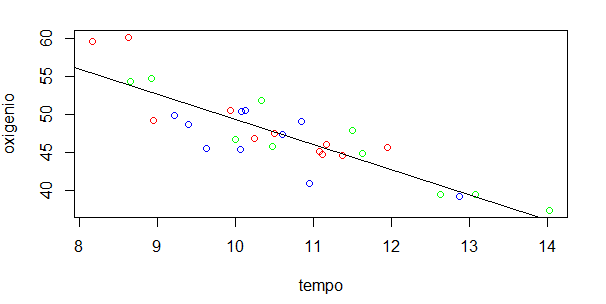
**Introdução Análise de Regressão** A análise de regressão estuda o relacionamento entre uma variável, chamada variável dependente, e outras variáveis, chamadas independentes.

Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes.

**Modelo de Regressão Linear Simples**  O modelo de regressão linear simples, define uma relação linear entre a variavel dependente e a variável independente. Possui esse nome porque considera que a relação entre as variáveis é descrita por uma função linear (equação da reta ou do plano).



**Equação da Reta – Regressão Linear Simples**

  
Como podemos observar no gráfico abaixo, o modelo de regressão linear simples é descrito por uma função linear (equação da reta ou do plano).

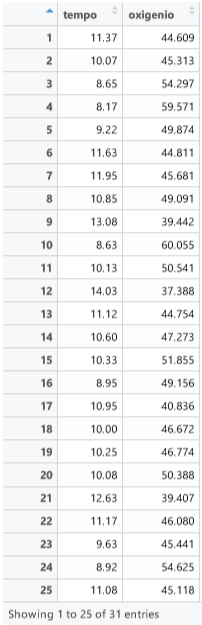
**Exemplos:**

1. Relação entre trecho percorrido e o tempo gasto.  
   (x: distância; y: tempo).
2. Relação entre os batimentos cardíacos e a velocidade atingida.  
   (x: batimentos; y: velocidade).

**Estudo de Caso**

Analisar o tempo gasto para correr 1.5 milhas (minutos) e a taxa de consumo de oxigênio (ml por minuto) em um grupo de homens envolvidos em um curso de fitness na na N.C. State University.

**Conjunto de Dados**



O conjunto de dados possui informações do tempo gasto para percorrer as 1,5 milhas e o consumo de oxigênio gasto no trecho percorrido.

O modelo de regressão linear simples deverá analisar a relação entre as variáveis tempo e oxigênio consumido (x: tempo; y: oxigênio).

**Desenvolvimento do Algoritmo (Implementação em R)**

**Leitura de Dados**

Primeiramente, vamos organizar os dados como objetos de dados no R através de um data frame (planilha). Para isso, é necessário que a tabela acima se encontre em uma estrutura tabular, na qual as colunas representam as variáveis e as linhas representam os valores. Logo, seja o arquivo de texto oxigenio.txt (download), utiliza-se a função read.table para que  o arquivo seja lido pelo R:



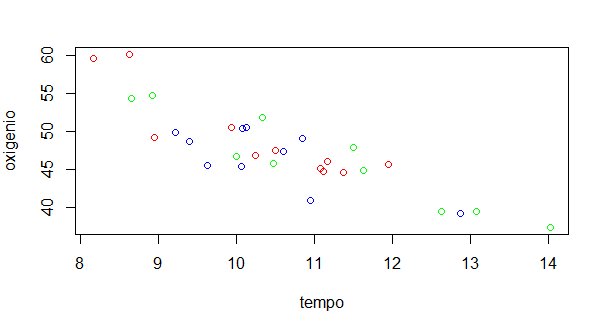
Observe que o argumento header=T indica que a primeira linha do arquivo contém os rótulos da planilha e que a função *attach* anexa o objeto consumo no ambiente de variáveis globais do R.

**Diagrama de Dispersão**

Para verificar a existência de alguma relação entre as variáveis tempo e oxigenio, deve-se construir um Diagrama de Dispersão para as duas variáveis:



A figura abaixo apresenta o Diagrama de Dispersão produzido pelo código em R acima. É possível fazer uma interpretação da existência de uma Correlação Negativa, onde é observado que quanto menor é o tempo gasto no percurso variável x:tempo, acarreta em um consumo maior de oxigênio variável y:oxigenio, assim se controlarmos x, y será também controlado.



**Coeficiente de Correlação Linear**

Para calcular o Coeficiente de Correlação Linear de Pearson entre as variáveis, utilize a função cor:



O R irá retornar o valor 0.8621949 o que evidencia uma forte correlação linear entre as variáveis. Este coeficiente, assume apenas valores entre -1 e 1.

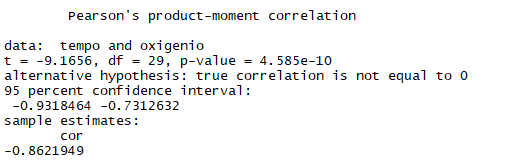
**Interpretação dos valores do Coeficiente de Correlação**

* 1.0 positivo ou negativo. Indica uma correlação perfeita
* 0.9 positivo ou negativo. Indica uma correlação muito forte.
* 0.7 a 0.9 positivo ou negativo. Indica uma correlação forte.
* 0.5 a 0.7 positivo ou negativo. Indica uma correlação moderada.
* 0.3 a 0.5 positivo ou negativo. Indica uma correlação fraca.
* 0 a 0.3 positivo ou negativo. Indica uma correlação desprezível.

**Teste de Hipóteses para o Coeficiente de Correlação**

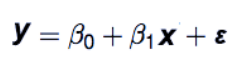
Para que possamos avaliar se esse resultado é significativo, vamos realizar um Teste de Hipóteses para o Coeficiente de Correlação.



Ao executar o Teste de Hipóteses, teremos o seguinte resultado:  
  


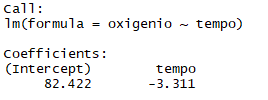
Observem que o **p-value** do teste **(p-value = 4.585e-10)** é bem pequeno, conclui-se que o valor do Coeficiente de Correlação Linear de Pearson tem significância Estatística.   
  
 O teste de **Correlação de Pearson**, também indica que a hipótese deve ser aceita, visto o intervalo de confiança entre 0,9318464 e  0,7312632.   
  
**Regressão Linear Simples**

Ajuste do Modelo Linear Sejam x e y, respectivamente, as variáveis tempo (explicativa) e oxigênio (resposta). Propõe-se um modelo de regressão linear, dado pela equação abaixo.



Onde β0 e β1são parâmetros desconhecidos e ε é o erro aleatório. Para ajustar um modelo de regressão linear no R utiliza-se a função lm, conforme o código abaixo.



  
Ao executar o modelo ajustado irá retornar as seguintes informações:

Beta 0 Beta 1

Note que função lm() é chamada com o formato lm(y ˜ x), ou seja, a variável resposta é y e a explicativa é x , sempre nessa ordem.

O R retorna o valor dos coeficientes de β0 e β1 estimados via Método de Mínimos Quadrados. Logo, a equação da reta ajustada é dada por:

**y = 82,422 + (-3,311) \* x**

Onde poderíamos interpretar que, para estimar ou prever o valor do consumo de oxigenio basta informar o tempo na equação e realizar o cálculo.

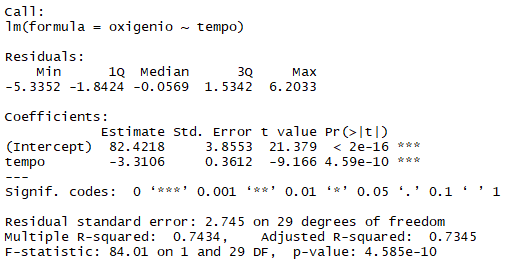
**consumo = 82,422 + (-3,311) \* tempo**

**Teste de Significância do Modelo**

Realizando o teste de significância, verifica-se o P-valor das variáveis através da saída da função **summary**:



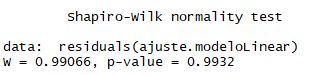
Ao analisar o teste, o P-valor da variável tempo está bem próxima de 0, isso demonstra que esta é uma variável significante ao modelo. Conforme a saída da função **summary** abaixo.



**Teste de Normalidade**

A normalidade da amostra é confirmada pelo Teste de Normalidade de Shapiro-Wilk, cujo P-valor 0,9932 > 0,05. Conforme instrução em R abaixo





**Estimativa do Consumo Oxigênio**

Tendo uma função matemática significante que representa nosso conjunto de dados, vamos realizar a estimativa de consumo de oxigênio com um tempo de (12,5) minutos do conjunto de dados utilizando a função predict no R.



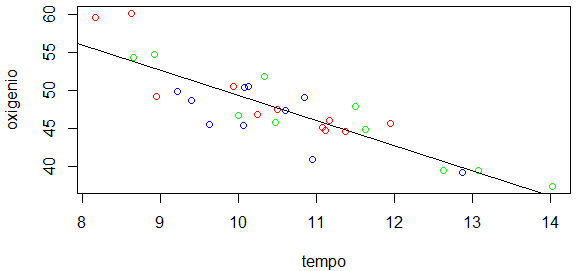
Ao executar teremos uma estimativa de consumo de Oxigênio de 41.03983.

**Reta Ajustada**

Para esboçar a reta ajustada no diagrama de dispersão, vamos utilizar a função **abline:**



Ao executar o trecho em R, teremos o Diagrama de Dispersão com a reta ajustada, conforme figura abaixo.



**Avaliação da Qualidade do Modelo**

**Coeficiente de Determinação**

O coeficiente de determinação R², reflete na qualidade do ajuste do modelo, ou seja, o quanto o modelo consegue explicar o conjunto de dados. Vamos utilizar a função summary do R para extrair o valor do coeficiente de determinação R²



Sendo o R² = 0.7433801, logo

O modelo consegue explicar cerca de 74% dos dados observados.