Análise de Regressão Linear

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

Probabilidade e Estatística
Robson Fernandes - robson.fernandes@usp.br, Francisco Louzada - louzada@icmc.usp.br

Novembro de 2016

1 Introdução

Este trabalho visa 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. Propõe descrever um model linear que melhor represente o conjunto dados.

Abaixo são apresentadas apenas as variáveis tempo de corrida e taxa de consumo de oxigênio.

```
> exigenio = read.table('oxigenio.txt',header=T)
> attach(exigenio)
> exigenio

   tempo exigenio
1  11.37   44.609
2  10.07   45.313
3   8.65   54.297
4   8.17   59.571
5   9.22   49.874
6  11.63   44.811
7  11.95   45.681
8  10.85   49.091
9  13.08   39.442
10  8.63   60.055
11 10.13   50.541
12 14.03   37.388
13 11.12   44.754
14 10.60   47.273
15 10.33   51.855
16  8.95   49.156
17 10.95   40.836
```

```
18  10.00    46.672
19  10.25    46.774
20  10.08    50.388
21  12.63    39.407
22  11.17    46.080
23   9.63    45.441
24   8.92    54.625
25  11.08    45.118
26  12.88    39.203
27  10.47    45.790
28   9.93    50.545
29   9.40    48.673
30  11.50    47.920
31  10.50    47.467
>
```

2 Análise Descritiva

Na tabela 1 tem-se a análise descritiva do conjunto de dados, onde são avaliados *(Média, Mediana, Mínimo, Máximo, Variância e Desvio Padrão)* das variáveis *tempo* e *oxigênio*.

Tabela 1: Estatística descritiva

	Tempo	Oxigênio
Mínimo	8.17	37.39
Máximo	14.03	60.05
Média	10.59	47.38
Mediana	10.47	46.77
Variância	1.924918	28.37938
Desvio Padrão	1.387414	5.327231

3 Diagrama de Dispersão

Diagrama de dispersão entre as variáveis *tempo* e *oxigênio*:

```
> plot(tempo, oxigenio, xlab="Tempo", ylab="Oxigenio")
> points(mean(tempo), mean(oxigenio), col="red", lwd=5, lty=9)
>
```

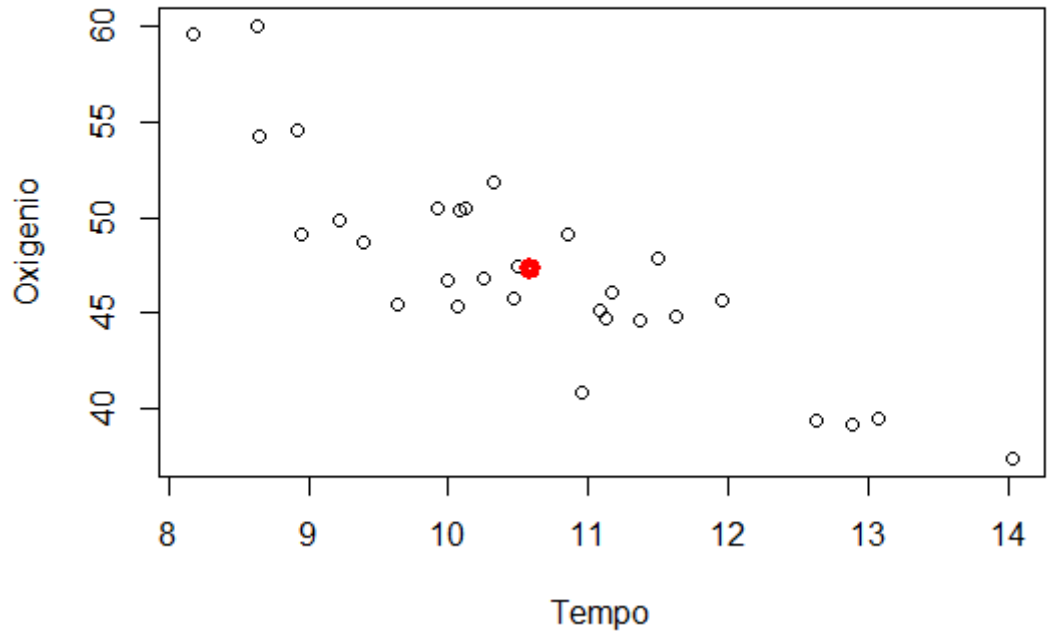

Figura 1: Gráfico de Dispersão de *tempo* versus *oxigênio*: ● ponto médio.

4 Ajuste do Modelo Linear

Seja a variável explicativa *oxigênio* (taxa de consumo de oxigênio) e a variável resposta *tempo* (tempo de corrida em minutos). Propõe-se um modelo de regressão linear para explicar a variável *tempo*, dado pela equação:

$$Y = \beta_0 + \beta_1 X$$

onde:

Y *é o valor a ser predito*
β_0 *é o intercepto (valor quando* $x = 0$*)*
β_1 *é a inclinação da reta de regressão*
X *é o valor da variável preditora*

```
> modelo = lm(tempo ~ oxigenio)
> modelo
>
```

5 Teste de Correlação - Coeficiente de Pearson

O teste indidica uma forte correlação entre as variáveis *tempo* e *oxigênio*, sendo R = -0.8621949, bem próximo de -1, indicando que há correlação negativa perfeita entre as duas variáveis, isto é,se uma aumenta, a outra sempre diminui.

```
> cor.test(dataOxigenio$tempo, dataOxigenio$oxigenio)
>
```

6 Testes de Significância do Modelo

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

```
> modelo = lm(tempo ~ oxigenio)
> summary(modelo)
>
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2243     1.1678  18.175  < 2e-16
oxigenio     -0.2245     0.0245  -9.166 4.59e-10

Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7345
```

Ao analisar o teste, o *p-valor* da variável *oxigenio* está bem próxima de 0, isso demonstra que esta é uma variável significante ao modelo.
O R^2 ajustado do modelo é 0,7345, isto significa que 73,45% da variável dependente consegue ser explicada pelos regressores presentes no modelo.

7 Teste de Normalidade

A normalidade da amostra é confirmada pelo Teste de normalidade de *Shapiro-Wilk*, cujo *P-valor* 0.4295 $\geq$ 0,05.

```
> shapiro.test(residuals(modelo))
>
```

8 Análise de Resíduos

Considera-se os seguintes gráficos para realizar a Análise dos Resíduos:

```
> plot(fitted(modelo),residuals(modelo),
                          xlab="Valores Ajustados",
                          ylab="Residuos")
> abline(h=0)

> plot(tempo,residuals(modelo),xlab="tempo",ylab="Residuos")
> abline(h=0)

> qqnorm(residuals(modelo), ylab="Residuos")
> qqline(residuals(modelo))
>
```

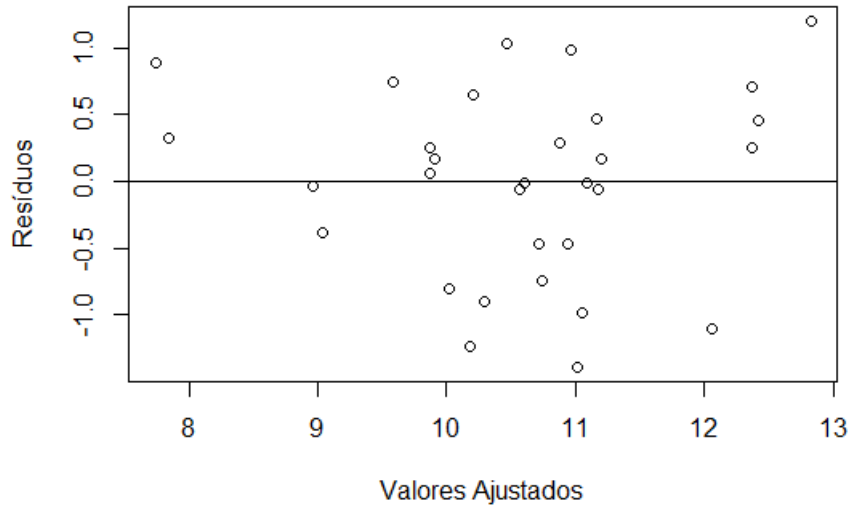

Figura 2: Gráfico de Resíduos versus Valores Ajustado

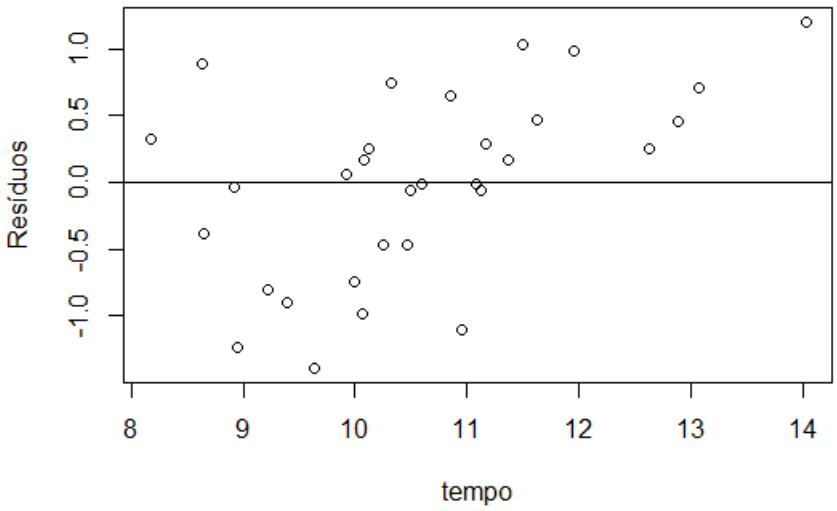

Figura 3: Gráfico de Resíduos versus Experiência

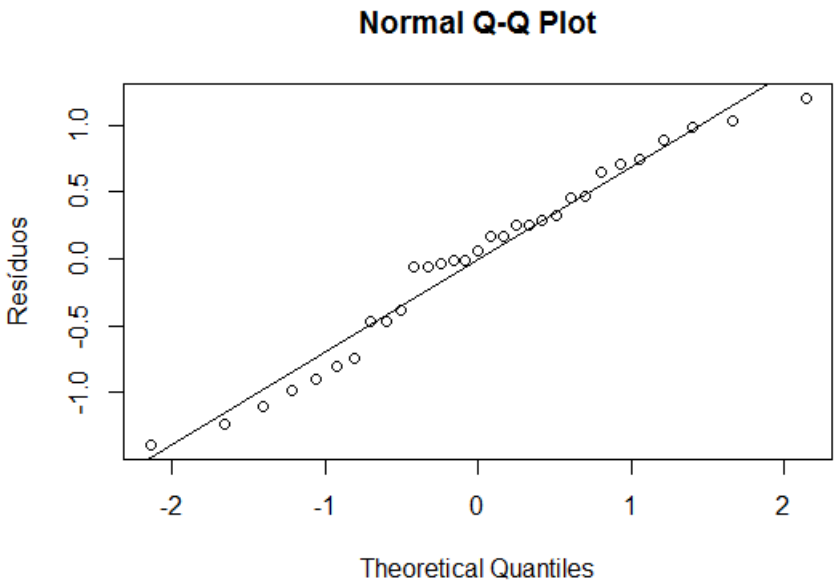

Figura 4: Normal QQ

9 Resultados e Considerações Finais

Com base na análise realizado na amostra, obteve-se um modelo de regressão linear que descreve o conjunto de dados. O R^2 ajustado do modelo é 0,7345, sendo assim, 73,45% da variável dependente consegue ser explicada pelos regressores presentes no modelo.

A equação da reta ajustada foi definida por:

$$Y = 21.2243 - 0.2245X$$

Abaixo tem-se o gráfico que descreve o modelo de regressão linear proposto:

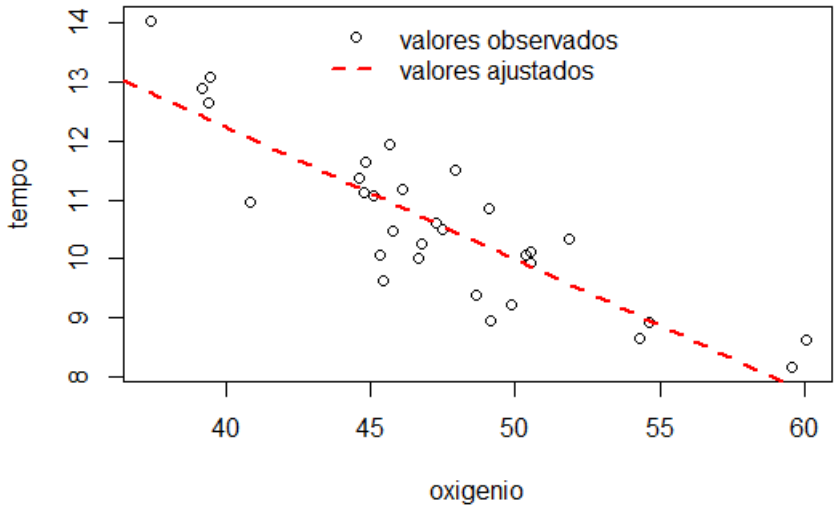

Figura 5: Gráfico de Regressão Linear