

TUTORIAL SOBRE ANÁLISE DE REGRESSÃO

I. – No R:

```
### Exemplo dados dos alunos: altura x peso
### x = altura e y = peso
#####
```

Entrada de Dados

```
x <- c(1.61,1.57,1.77,1.75,1.83,1.69,1.91,1.75,1.72,1.68,1.92,1.58,1.77)
y <- c(54.5,43.0,72.0,77.0,76.0,83.0,78.0,90.0,59.0,75.4,120.0,43.0,88.0)
```

a) Calculando as expressões na “raça”:

```
plot(x,y, pch=19, col="red3")          # gráfico peso x altura
n <- length(x)                          # tamanho da amostra
xbar <- mean(x)                          # média das alturas
ybar <- mean(y)                          # média dos pesos
s2x <- var(x)                            # variância das alturas
s2y <- var(y)                            # variância dos pesos
sxy <- (sum(x*y)-n*xbar*ybar)/(n-1)      # covariância entre peso e altura
rxy <- cor(x,y)                          # correlação entre peso e altura
```

Ajuste de mínimos quadrados ## estimativas dos coeficientes

```
betal <- sxy/s2x
beta0 <- ybar-xbar*betal
c(betal,betal)
```

Coeficiente de determinação

```
R2 <- (sxy^2)/(s2y*s2x)
R2
(rxy)^2                                # relação de R2 com rxy
betal*sxy/s2y                           # relação de R2 com betal
```

Valores ajustados e resíduos

```
yeh <- beta0 + betal*x                  # valores ajustados
ei <- y - yeh                           # resíduos
```

plotando a reta ajustada

```
plot(x,y, pch=19, col="red3", main="Gráfico de pontos", xlab="Altura",
      ylab="peso")
lines(x[order(x)], yeh[order(x)],col="blue3", lwd=2)
```

Análise gráfica para verificação das suposições ## do modelo (normalidade e igualdade de variâncias)

histograma

```
hist(ei, col="bisque", main="Histograma dos resíduos", xlab="Resíduos",
      ylab="Frequência", cex.main=1.25)
```

normal qq-plot dos resíduos

```
qqnorm(ei)
qqline(ei)
```

resíduos x valores ajustados

```
plot(yeh,ei, main="Resíduos x Ajustados", pch=19, col="red3")
```

resíduos x ordem (timeplot)

```
plot(ei, main="Resíduos x ordem", pch=19, col="red3")
lines(ei, lwd=2, col="blue3")
```

b) Explorando os recursos do R:

Usando o comando - lm

```
reta <- lm(y~x)          # guardando os resultados do ajuste
reta
```

```
summary(reta)           # sumário do ajuste
## O commando summary fornece:
## i) descritivas dos resíduos;
## ii) coeficientes estimados com respectivos testes de siginificância;
## iii) erro padrão dos resíduos (estimativa do desvio padrão sigma);
## iv) coef. de determinação R2 e R2-ajustado;
## v) teste F da tabela ANOVA, para testar Ho:beta1=0.
```

```
aov(reta)               # somas de quadrado dos resíduos e do modelo
res <- reta$residuals    # resíduos
sdres <- rstandard(reta) # resíduos padronizados
ajus <- reta$fitted      # valores ajustados
```

traçando a reta de regressão

```
plot(x,y, pch=19, col="red3", main="Gráfico de pontos", xlab="Altura",
      ylab="peso")
abline(coef(reta), col="blue3", lwd=2)
```

verificação das suposições do modelo:

teste de normalidade de Kolmogorov-Smirnov

```
ks.test(sdres,pnorm,0,1)
```

Análise gráfica para verificação das suposições

do modelo (normalidade e igualdade de variâncias)

gráfico 4x1

```
par(mfrow = c(2,2)) # esse comando divide a janela gráfica numa matriz 2x2
```

histograma dos resíduos

```
hist(sdres, col = "bisque", xlab="Resíduos padronizados", ylab=" ",
      main="Histograma dos Resíduos", cex.main=1)
```

gráfico normal qq-plot dos resíduos

```
qqnorm(sdres, pch=19, col="red", xlab="Quantis teóricos", ylab="Quantis
      amostrais", main="Normal Q-Q plot dos Resíduos", cex.main=1)
qqline(sdres, col="blue3")
```

gráfico resíduos x valores ajustados

```
plot(ajus,sdres,pch=19,col="red", xlab="Valores ajustados", ylab="Resíduos
      padronizados", main="Resíduos x Ajustados", cex.main=1)
lines(c(0,max(ajus)),c(0,0),lty=2)
```

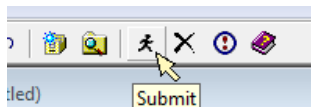
resíduos x ordem (timeplot)

```
plot(sdres, main="Resíduos x ordem", ylab="Resíduos padronizados", pch=19,
      col="red3", cex.main=1)
lines(sdres, col="blue3")
lines(c(0,max(ajus)),c(0,0),lty=2)
```

```
par(mfrow = c(1,1))
```

II. – No SAS:

- Na janela *Editor* do SAS digitar (ou colar) os comandos abaixo;
- Marcar as linhas conforme os 3 blocos de comandos: 1a. parte; 2a. parte e 3a. parte;
- Clicar no botão *submit*, (ver figura) para cada bloco de comandos marcado;
- Observar as saídas (quando houver) na janela *Output*.



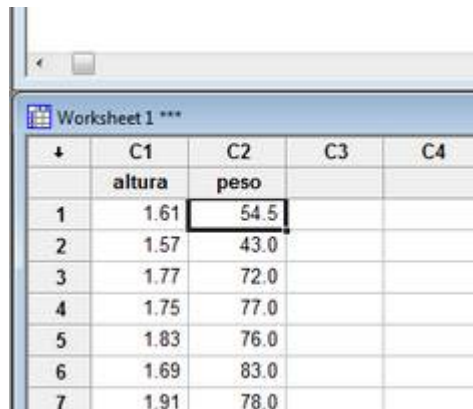
Comando	Descrição
<pre>/* 1a. parte - entrada dos dados */ data exemplo1; input altura peso; cards; 1.61 54.5 1.57 43.0 1.77 72.0 1.75 77.0 1.83 76.0 1.69 83.0 1.91 78.0 1.75 90.0 1.72 59.0 1.68 75.4 1.92 120.0 1.58 43.0 1.77 88.0 ; proc corr data=exemplo1 noprob; run;</pre>	<p>⇒ O comando data é usado para criar o conjunto de dados, que no caso, recebeu o nome de exemplo1.</p> <p>O conjunto de dados exemplo1 tem duas variáveis contínuas: altura e peso.</p> <p>⇒ O comando proc corr calcula o coeficiente de correlação linear entre as variáveis</p>
<pre>/* 2a. parte - ajuste da regressao */ title1 'Análise Grafica do Ajuste'; symbol v=dot h=1 c=green; proc reg data=exemplo1; model peso=altura; output out=saida p=ajus r=res student=sdres; title2 'Grafico de pontos'; plot peso*altura /lline=1 cline=red; run; proc print data=saida; run;</pre>	<p>⇒ As definições title1 e symbol são fixadas para os gráficos.</p> <p>⇒ O comando para o ajuste e análise da regressão é o proc reg, que utiliza os dados em exemplo1. O modelo ajustado é definido pela opção model peso=altura, que define o modelo de regressão linear simples para peso e altura. A opção output define o arquivo saída com os valores ajustados (ajus); resíduos (res) e resíduos padronizados (sdres). A opção plot constrói o gráfico de pontos peso*altura, com a reta ajustada.</p> <p>⇒ Como saídas do proc reg o SAS fornece a tabela análise de variâncias (tabela ANOVA); o coeficiente de determinação R² e os testes individuais para os parâmetros do modelo.</p>

```
/* 3a. parte - avaliação das suposições do
modelo */
title2 'Resíduos x Valores Ajustados';
plot r.*p. /nostat;
run;
title2 'Resíduos x Ordem';
plot r.*obs. /nostat;
run;
title2 'Normal QQ-plot dos Resíduos';
plot student.*nqq. /noline nostat;
run;
quit;
```

⇒ O comando `plot` é usado para a construção dos gráficos que servirão para a avaliação das suposições de normalidade e igualdade de variância. Cada gráfico é identificado pelo seu subtítulo, dado pela opção `title2`.

III. – No Minitab:

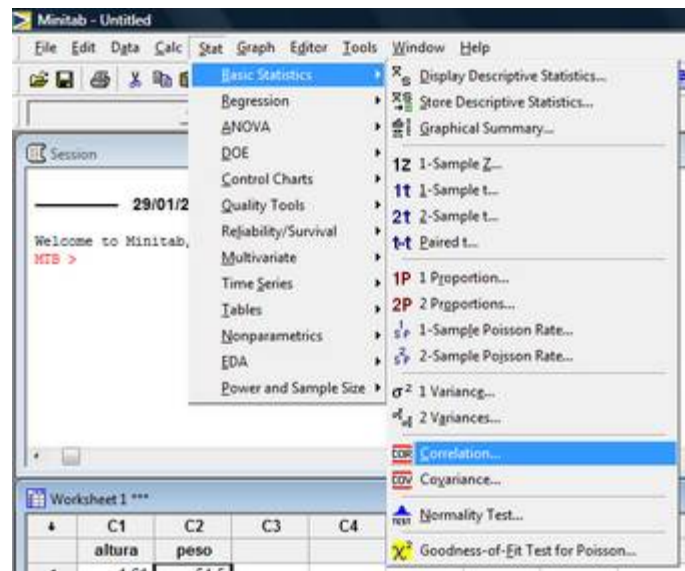
- a) Digitar os dados na planilha do Minitab (*Worksheet*) conforme figura. Pode ser copiado e colado direto do *Excel* se preferir;



The screenshot shows a Minitab worksheet titled 'Worksheet 1 ***'. It contains a table with 4 columns: C1, C2, C3, and C4. The first two columns are labeled 'altura' and 'peso' respectively. The data is as follows:

	C1	C2	C3	C4
	altura	peso		
1	1.61	54.5		
2	1.57	43.0		
3	1.77	72.0		
4	1.75	77.0		
5	1.83	76.0		
6	1.69	83.0		
7	1.91	78.0		

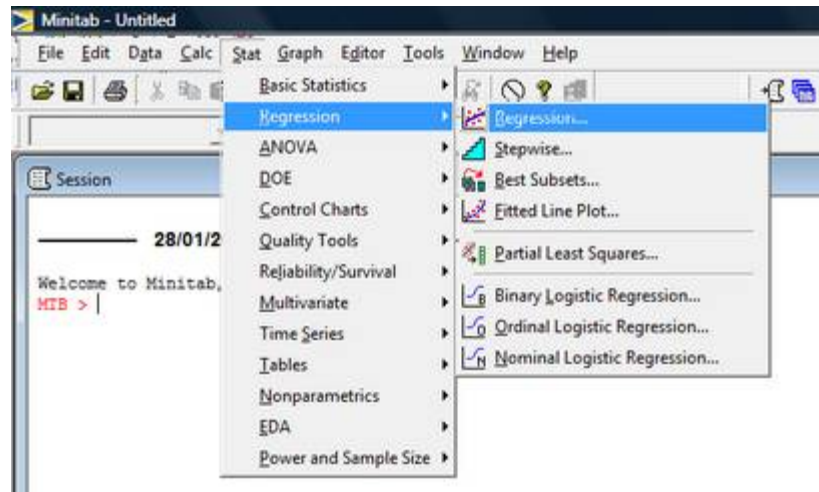
- b) No menu Stat, selecionar **Basic Statistics** e em seguida Correlation;



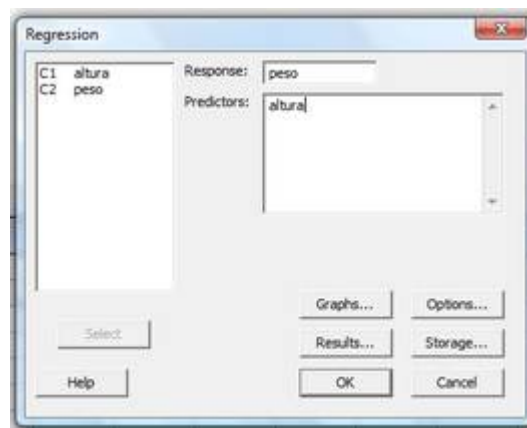
- c) Na próxima janela, selecionar as variáveis para o coeficiente de correlação linear e clicar OK;



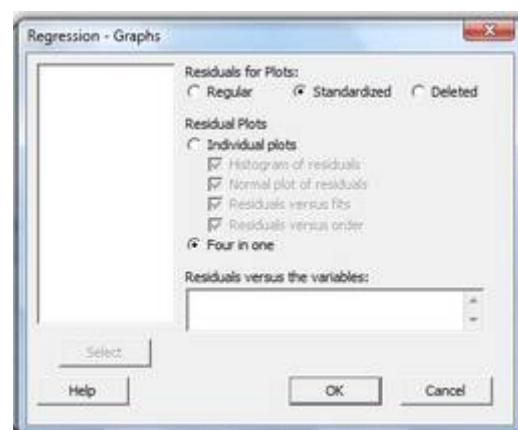
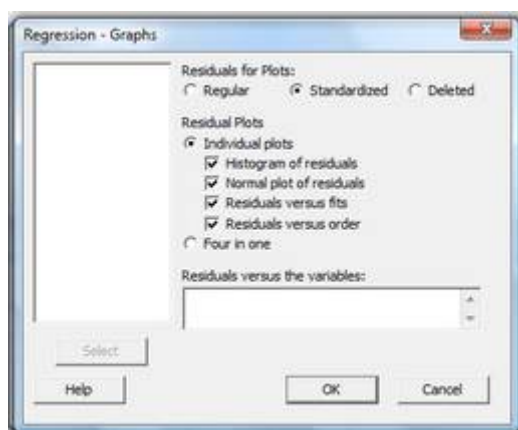
- d) No menu **Stat**, selecionar a opção **Regression** e na sequência, selecionar **Regression** novamente;



- e) Na janela que vai abrir, selecionar as variáveis resposta e preditora, conforme abaixo;



- f) Ainda na janela **Regression**, selecionar a opção **Graphs** e marcar os gráficos desejados, individualmente (Figura da esquerda) ou todos os 4 gráficos na opção quatro em um **“Four in one”** (Figura da direita). Na opção **Residuals for Plots**, marcar **Standardized**.



- g) Observar a saída no janela *Session* e o gráfico gerado na janela independente.

Correlations: altura; peso

Pearson correlation of altura and peso = 0.797

Regression Analysis: peso versus altura

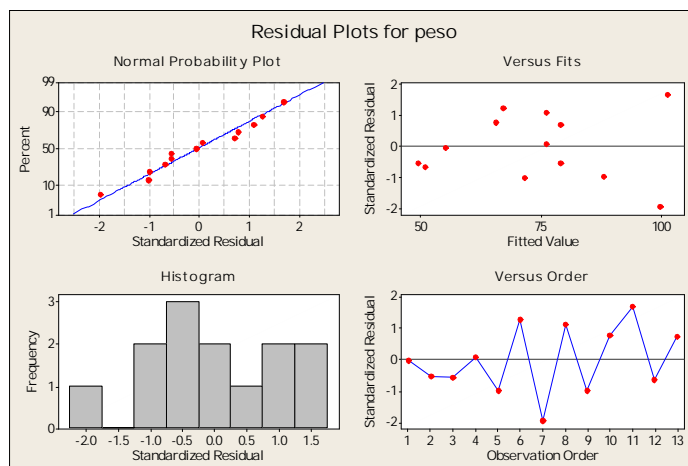
The regression equation is
peso = - 185 + 149 altura

Predictor	Coef	SE Coef	T	P
Constant	-184.87	59.18	-3.12	0.010
altura	149.10	34.05	4.38	0.001

S = 13.1420 R-Sq = 63.5% R-Sq(adj) = 60.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3310.6	3310.6	19.17	0.001
Residual Error	11	1899.8	172.7		
Total	12	5210.5			



Exemplo:

Medidas feitas em um grupo de homens envolvidos em um curso de fitness na N.C. State University. As variáveis coletadas foram: *Idade* (anos); *peso* (kg); *taxa de consumo de oxigênio* (ml por kg de peso por minuto); *tempo gasto para correr 1.5 milhas* (minutos); *frequência cardíaca em repous* (pulso); *frequência cardíaca durante a corrida* (medida ao mesmo tempo que a taxa de consumo de oxigênio) e *frequência cardíaca máxima registrada na corrida*. Abaixo são apresentadas apenas as variáveis *tempo de corrida* e *taxa de consumo de oxigênio*.

Avalie se o modelo de regressão linear simples pode ser ajustado para as variáveis *tempo de corrida* e *taxa de consumo de oxigênio*.

tempo	oxigenio
11.37	44.609
10.07	45.313
8.65	54.297
8.17	59.571
9.22	49.874
11.63	44.811
11.95	45.681
10.85	49.091
13.08	39.442
8.63	60.055
10.13	50.541
14.03	37.388
11.12	44.754
10.60	47.273
10.33	51.855
8.95	49.156
10.95	40.836
10.00	46.672
10.25	46.774
10.08	50.388
12.63	39.407
11.17	46.080
9.63	45.441
8.92	54.625
11.08	45.118
12.88	39.203
10.47	45.790
9.93	50.545
9.40	48.673
11.50	47.920
10.50	47.467