10. REGRESSÃO LINEAR SIMPLES

Regressão é um procedimento estatístico de estimar uma expressão algébrica (modelo) que explique a variação de uma variável dependente Y em função de p variáveis dependentes, ou explanatórias, X_1, X_2, ..., X_p.
A Regressão Linear Simples é um particular tipo de Regressão em que se tem uma variável dependente (Y) e uma única variável independente (X) e o modelo é da forma:
$Y=A+B\ X+\varepsilon$,
Onde: Y é a variável dependente e X a variável independente; ε o vetor de erros aleatórios (resíduos); A e B são os parâmetros do modelo.

As estimativas de A e B, pelo método dos mínimos quadrados, são os valores que minimizam a soma de quadrados $\sum|\varepsilon_i|^2$.

A análise de variância é esquematizada como:

F.V.	G.L.	S.Q.	Q.M.	F	*p-value*
Modelo	1	SQ(Mod.)	QM(Mod.)	QM(Mod.) / QM(Res.)	p
Resíduo	N-2	SQ(Res.)	QM(Res.)		
Total	N-1	SQ(Tot.)			

F.V. – Fontes de Variação, G.L. – Graus de Liberdade, S.Q. – Somas de Quadrados, Q.M. – Quadrados Médios, N – Número de observações

A estatística F testa a hipótese: H_0: B=0 vs H_1:B≠0.

O valor p (***p-value***) é obtido supondo que a estatística F tem uma distribuição F central com 1 e N-2 graus de liberdade. Essa pressuposição é válida se os erros forem iid - independentes e identicamente distribuídos, com distribuição normal $N(0,\sigma^2)$.

Para exemplificar a Análise de Regressão Linear Simples no R, considere o exemplo:

Exemplo(RLS_ex1): Foram observados 9 valores de uma variável X (temperatura) e os correspondentes valores Y (dilatação linear do metal) e os resultados foram:

X	18	16	25	22	20	21	23	19	17
Y	5	3	10	8	6	7	9	6	5

Entrada do dados no R:

```
> X <- c(18, 16, 25, 22, 20, 21, 23, 19, 17)
> Y <- c(5, 3, 10, 8, 6, 7, 9, 6, 5)
> X;Y
```

10.1 Gráfico auxiliar (opcional)

Representar os pontos graficamente (gráfico de dispersão) para visualizar se existe uma aparente relação linear entre X e Y.

```
> plot(X,Y)
```

10.2 Definição do modelo e estimação dos parâmetros

```
> mod <- lm(Y~X)
> mod
```

10.3 Análise de Regressão Linear Simples e obtenção do R^2.

> **anova(mod)**
> **summary.lm(mod)**

O Coeficiente de Determinação R^2, que é a porcentagem da variação que é explicada pelo modelo, pode ser obtido com o comando **summary.lm(mod)** do R.

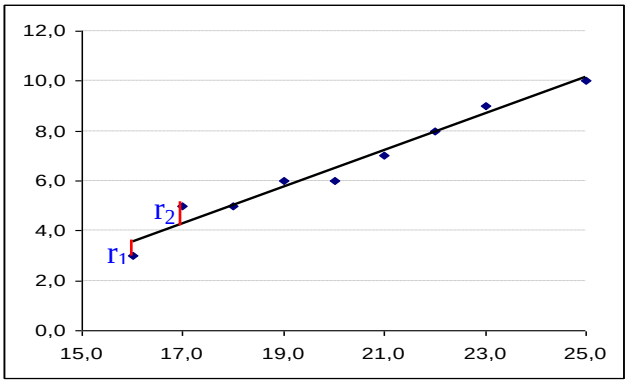
Sabe-se que esta análise considera que os erros sejam iid (independentes e identicamente distribuídos) e que tenham distribuição normal - $N(0,\sigma^2)$. Para isso faz-se necessário os diagnósticos para a análise.

10.4 Diagnósticos para a análise de modelos lineares

Encontram-se na literatura três tipos de resíduos que são: resíduo ordinário, resíduo padronizado e resíduo estudentizado.

- **Resíduo ordinário:**

$r_i = Y_i - \hat{Y}_i$ (Y_i – valor observado e $\hat{Y}_i$ - valor estimado pelo modelo ou valor predito). Os resíduos ordinários (r_1, r_2, ...) são apresentados na Figura a seguir.



No R:
> **res <- residuals(mod); res**

- **Resíduo padronizado internamente (Studentized residual)**

$rs_i = r_i / \sqrt{\hat{V}(r_i)}$, onde $\hat{V}(r_i)$ é estimativa da variância residual.

No R
> **rp <- rstandard(mod); rp**

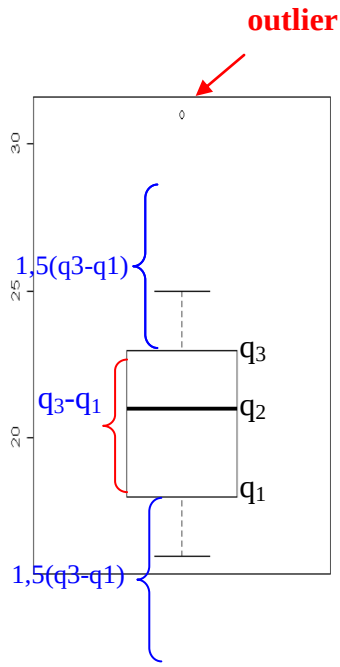
- **Resíduo padronizado externamente (Jacknife residual, Rstudent)**

$Rs_i = r_i / \sqrt{\hat{V}_{(i)}(r_i)}$, onde $\hat{V}_{(i)}(r_i)$ é estimativa da variância residual sem a observação *i*. *Rs*~t(*N-p*-1) onde N é o número de observações e p número de parâmetros.

No R:
> **rs <- rstudent(mod); rs**

Os diagnósticos utilizados em Análise de Regressão Linear envolvem:

- Interpretação subjetiva de gráficos apropriados (histograma, boxplot, normalidade e análise do resíduo).
- Estudo da presença de pontos discrepantes (fora do padrão), que podem ser
  - *Outlier* - valores menores que q1-1,5(q3-q1) ou maiores que q3+1,5(q3-q1) – Representados na Figura a seguir, ou
  - *Valores influentes*.



Apresenta-se a seguir os comandos R e detalhes sobre os passos para os diagnósticos de um modelo de Regressão Linear Simples e os testes de normalidade dos erros, aplicando para os dados do Exemplo(**RLS_ex1**).

a) Gráficos de diagnósticos (histograma, boxplot e normalidade)

No R

```
> par(mfrow=c(1,3))
> hist(rs); boxplot(rs); qqnorm(rs); qqline(rs)
```

b) Análise de resíduos

A análise de resíduos é feita usando o gráfico "*valor predito*" x "Resíduo", que permite diagnosticar a condição dos erros serem iid (independentes e identicamente distribuídos). Ver Figuta 1.

Para ilustrar qual resíduo utilizar na análise de resíduos, representar graficamente este gráfico usando os três resíduos, numa mesma figura e observar o que acontece.

Comando R:

```
> mod <- lm(Y~X)
> res <- residuals(mod); res
> rp <- rstandard(mod); rp
> rs <- rstudent(mod); rs
> yp <- predict.lm(mod)
> par(mfrow=c(1,3))
> plot(yp,res); plot(yp,rp); plot(yp,rs)
```

Observe que, sem considerar as escalas, a distribuição dos pontos é a mesma.

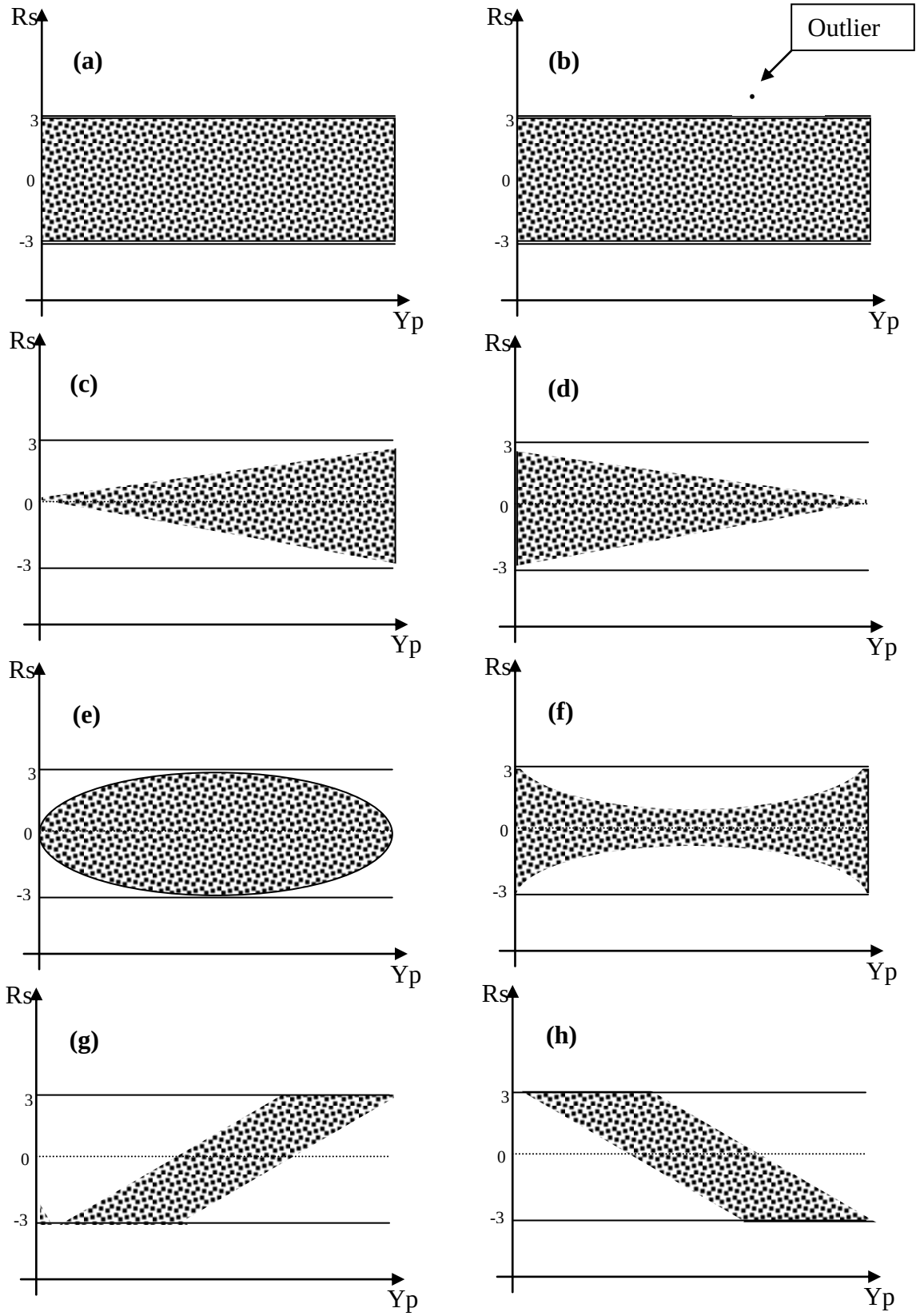


Figura 1. Gráficos dos resíduos estudentizados (Rs) por Valores preditos (Yp). **(a)** Satisfaz as exigências do modelo, **(b)** Presença de outliers, **(c)** variâncias aumentam com Yp, **(d)** variâncias diminuem com Yp , **(e)** variâncias maiores paraYp próximos da média, **(f)** variâncias menores paraY_p próximos da média, **(g)** erros dependentes, **(h)** erros dependentes.

c) Diagnóstico de valores discrepantes (outliers ou influentes)

c.1) Outliers

Um valor é considerado *outlier* se o Resíduo Padronizado Externamente (Rs) satisfizer a condição: Rs_i < -2,33 ou Rs_i > 2.33. Isso justifica utilizar na análise de resíduos o (Rs), que além das informações se os erros são iid (independentes e identicamente distribuídos), permite visualizar outiliers. Encontram-se referências que usam como limites valores entre -2 e 2 ou -3 e 3. Neste material usaremos os limites entre -3 e 3.

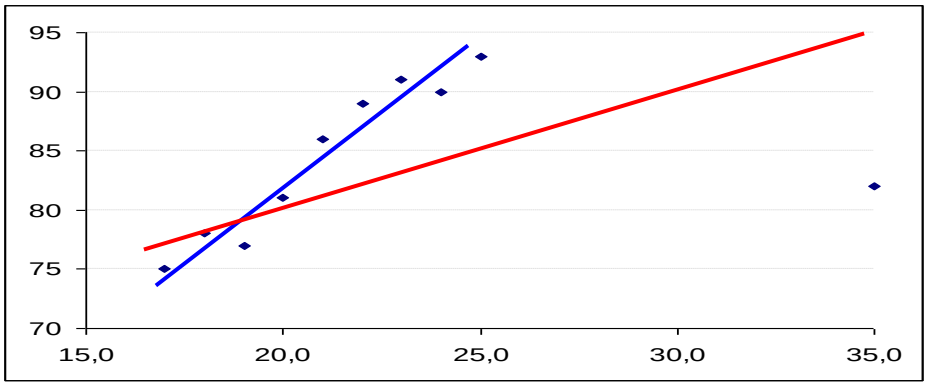
c.2) Valores influentes

Valores influentes são valores que, como o próprio nome diz, influenciam no modelo de regressão estimado.

Exemplo(RLS_ex2): Para avaliar o efeito da idade na massa muscular de indivíduos, foram tomadas 10 amostras, apresentadas a seguir:

ID	17	18	19	20	21	22	23	24	25	35
MM	75	78	77	81	86	89	91	90	93	82

O gráfico a seguir ilustra o efeito de pontos influentes.



- reta sem o indiv. de 35 anos
- reta com o indiv. de 35 anos

Observe que a observação (35,82) influenciou no modelo

Encontram-se na literatura algumas estatísticas usadas para diagnosticar pontos influentes, apresentadas na Tabela a seguir:

Tabela 1. Algumas Estatísticas utilizadas para diagnósticos de pontos influentes.

Estatística	**Denominação no R**	**Limite Crítico**
DFBetas (um por parâmetro)	dfb.1. , dfb.X, ...	$2/\sqrt{N}$
FDFitS	dffit	$2\sqrt{p/N}$
|1-COVRATIO|	cov.r	3p/N
Distância de Cook	cook.d	F(0,50;p,N-p)
Leverage – H	hat	3p/N

Essas estatísticas podem ser obtidas no R por comandos apropriados.

Os comandos do R para obter o vetor (Laverage - H) e o limite crítico são:

```
> X <- c(17:25,35)
> Y <- c(75,78,77,81,86,89,91,90,93,82); X;Y
> mod <- lm(Y~X)
> h <- hatvalues(mod);h
> P=length(mod$coefficient); N=length(Y); P; N; hc<-3*P/N;hc
```

Observe que, pela estatística hat, a observação (35,82) é um valor influente.
O comando do R para todas as estatísticas para detectar pontos influentes é:

```
> influence.measures(mod)
```

Com este comando obtem-se:

	DFBetas						
	dfb.1_	**dfb.x**	**dffit**	**cov.r**	**cook.d**	**hat**	**inf**
1	-0.5700	0.4879	-0.657	1.1409	0.20287	0.223	
2	-0.2655	0.2191	-0.327	1.3984	0.05702	0.182	
3	-0.2916	0.2277	-0.397	1.2041	0.07992	0.149	
4	-0.0767	0.0545	-0.123	1.4471	0.00854	0.124	
5	0.0647	-0.0378	0.136	1.4026	0.01041	0.108	
6	0.0795	-0.0224	0.273	1.2113	0.03884	0.101	
7	0.0336	0.0454	0.370	1.0557	0.06680	0.102	
8	-0.0304	0.0914	0.292	1.2185	0.04449	0.111	
9	-0.1413	0.2434	0.516	0.9471	0.12106	0.129	
10	**14.7123**	**-16.4311**	**-17.612**	**0.0287**	**12.55414**	**0.772**	*

Obtenção dos limites das estatísticas:

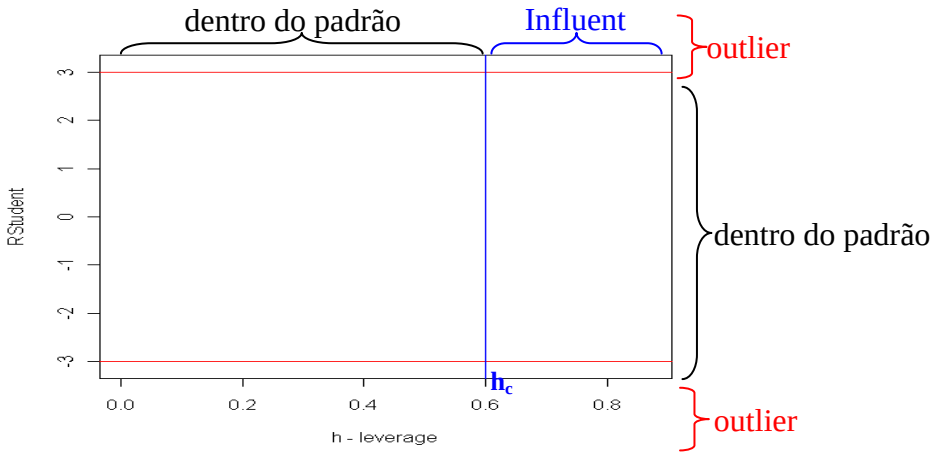
```
> limt <- list(c(DFB=2/sqrt(N),FDFits=2*sqrt(P/N),cov.r=3*P/N,
        Cook=qf(0.5,2,8, lower.tail = TRUE, log.p = FALSE),hat=3*P/N));limt
```

Resulta:

DFB	**FDFits**	**cov.r**	**Cook**	**hat**
0.6324555	0.8944272	0.6000000	0.7568285	0.6000000

É considerado influente se pelo menos uma das estatísticas exceder o limite crpitico.

Na prática escolhe-se uma das estatísticas – aqui usaremos a *hat*. e plota-se o gráfico "*h*" x "rs", que permite diagnosticar: outliers, valores influentes ou ambos. O gráfico a seguir ilustra as regiões onde os valores são considerados outliers ou influentes.



Para exemplificar, considere o exemplo apresentado a seguir:

Exmplo(RLS_ex3): Para os dados apresentados a seguir, **v**erificar a presença de valores extremos (*outlier* e/ influentes) para as três situações: (X1,Y1); : (X2,Y2); : (X3,Y3). Os dados estão disponíveis em **A_RLS_ex3.txt**.

X1	Y1	X2	Y2	X3	Y3
1	8,0	1	8,0	1	8,0
2	8,2	2	8,2	2	8,2
3	6,0	3	6,0	3	6,0
4	5,5	4	5,5	4	5,5
5	3,9	5	3,9	5	3,9
6	4,0	6	**8,5**	**20**	4,0
7	3,8	7	3,8	7	3,8
8	4,2	8	4,2	8	4,2
9	3,0	9	3,0	9	3,0
10	2,7	10	2,7	10	2,7

Solução:

```
> rm(list=ls())
> ex3 <- read.table("C:/EEAR/A_RLS_ex3.txt", header=T, dec=",")
> ex3
> attach(ex3)
> X<-X1; Y<-Y1
> plot(X,Y)
```

Comandos R para detecção de valores discrepantes.

```
> mod<-lm(Y~X)
> N=length(Y); P=length(mod$coefficient)
> rs <- rstudent(mod)
> h <- lm.influence(mod)$hat; lc <- 3*P/N
> minrs=min(min(rs),-3)
> maxrs=max(max(rs),3)
> ymin=minrs-.1
> ymax=maxrs+.1
> maxh=max(max(h),lc)
> minh=min(h)
> xmin=minh-.1
> xmax=maxh+.1
> par(mfrow=c(1,1))
> plot(c(xmin,xmax),c(ymin,ymax), type="n", xlab="h - leverage", ylab="RStudent")
> abline(h=-3, col="red")
> abline(h=3,col="red"); abline(v=lc, col="blue")
> points(h,rs)
```

Repita para: X<-X2; Y<-Y2
Repita para: X<-X3; Y<-Y3

Um Script no R para diagnósticos (**S_Diag**) - usando a estatística Laverage – H (hat) e limites -3 e 3 para outliers é apresentado a seguir.

Script para diagnósticos (S_Diag)

```
#####################################################################
###                        Script  Diagnósticos                   ###
###      Depende da variável Dependente Y e do modelo - mod       ###
#####################################################################
### Gráficos de diagnósticos
rs <- rstudent(mod)
par(mfrow=c(1,3))
hist(rs, main="histograma")
boxplot(rs, main="boxplot")
qqnorm(rs, main="normalidade"); qqline(rs)

### Gráfico para Análise do Residuo
rs <- rstudent(mod); yp <- predict.lm(mod)
par(mfrow=c(1,1))
plot(yp,rs,main="Análise do Resíduo", xlab="Valores preditos",
ylab="RStudent")
abline(h=0)

### Grafico para valores discrepantes (outliers e/ou influentes)
N=length(Y)
P=length(mod$coefficient)
rs <- rstudent(mod)
h <- lm.influence(mod)$hat
lc <- 3*P/N
minrs=min(min(rs),-3)
maxrs=max(max(rs),3)
ymin=minrs-.1
ymax=maxrs+.1
maxh=max(max(h),lc)
minh=min(h)
xmin=minh-.1
xmax=maxh+.1
par(mfrow=c(1,1))
plot(c(xmin,xmax),c(ymin,ymax), type="n", xlab="h - leverage",
ylab="RStudent")
abline(h=-3, col="red"); abline(h=3,col="red"); abline(v=lc, col="blue");
points(h,rs)
```

10.5 Testes de Normalidade dos Erros.

Os Testes de Normalidade já foram apresentados no item 6 e encontram-se no Script (**S_TestNorm.R**).

Script R para a Análise de Regressão Linear Simples

A análise de Regressão Linear Simples requer certos cuidados. As condições para a análise (Diagnósticos) devem ser observadas com cuidado.

Apresenta-se a seguir um Script no R para Análise de Regressão Linear Simples (S_RLS).

Script R (S_RLS)

```
#########################################################
###     Script para Regressão Linear Simples  - Y = A + B X     ###
#########################################################
## Entre com as variáveis X e Y   ##

#### Início do Script para Regressão Linear Simples ####
## Imprima as variaveis
X; Y

## Plote os valores observados e a reta regressora (opcional)
par(mfrow=c(1,1))
plot(Y~X); abline(lm(Y~X))

## Defina o modelo
mod <- lm(Y~X)
mod

#### Análise de Regressão Linear Simples e obtenção do R2
anova(mod)
summary.lm(mod)

#### Carregue o Script para Diagnósticos (S_Diag)
 # Gráficos: histograma, boxplot e normalidade; análise do resíduo e Valores extremos

####### Carregue o Script para testes de Normalidade (S_TestNorm)

#####################################################################
## se precisar eliminar registros, use a sintaxe:
# df <- df[-c(nr1,nr2...),]; attach(df)
# onde df=nome do data.frame
#     nr1, nr2  sao os numeros dos registros a serem eliminados
##### repita os procedimentos a partir do Início do Script #####
```

Atividade (RLS_at). Em um experimento, em cada semana (X), X=1, 2, ..., 5, obteve-se os Pesos (Y) de três animais (Repetições), cujos valores são apresentados a seguir (disponíveis em **A_RLS_at.txt**). Fazer a análise de Regressão Linear Simples desses dados.

Valores	Valores de Y		
de X	R1	R2	R3
1	8.2	6.6	9.8
2	19.7	15.7	16.0
3	28.6	25.0	31.9
4	30.8	37.8	40.2
5	40.3	42.9	32.6

a) Faça a análise de Regressão Linear Simples completa.
b) Apresentar os resultados.
c) Concluir.

11. REGRESSÃO LINEAR MÚLTIPLA

A Regressão Linear Múltipla é o caso que se tem uma variável dependente (Y) e k variáveis independentes ($X_1, X_2, ..., X_k$) e o modelo é da forma:

$Y = B_0 + B_1X_1 + B_2 X_2 + ... + B_kX_k + \varepsilon$,

Onde: Y é a variável dependente e $X_1, X_2, ..., X_k$ são as variáveis independentes; ε o vetor de erros (resíduos); $B_1, B_2, ...,B_k$ são os parâmetros do modelo.

A análise de variância é esquematizada como:

F.V.	G.L.	S.Q.	Q.M.	F	*p-value*
Modelo	k	SQ(Mod.)	QM(Mod.)	QM(Mod.) / QM(Res.)	p
Resíduo	N-k-1	SQ(Res.)	QM(Res.)		
Total	N-1	SQ(Tot.)			

F.V. – Fontes de Variação, G.L. – Graus de Liberdade, S.Q. – Somas de Quadrados, Q.M. – Quadrados Médios, N – Número de observações, k – número de variáveis independentes.

A estatística F testa a hipótese: H_0: $B_1=B_2= ... =B_k=0$ *vs* H_1: $B_i \neq B_{i'}$, para algum $i \neq i'$.

O valor p (***p-value***) é obtido supondo que a estatística F tem uma distribuição F central com K e N-k-1 graus de liberdade. Essa pressuposição é válida se os erros forem iid - independentes e identicamente distribuídos, com distribuição normal $N(0,\sigma^2)$.

Para exemplificar a Análise de Regressão Múltipla, considere o exemplo:

Exemplo(RLM_ex): Em um experimento com uma leguminosa foram observadas 17 repetições das variáveis: Y=Peso da planta, X1=Matéria Seca, X2=Diâmetro Médio, X3=Altura Média e X4=Número de folhas. Os dados estão disponíveis no arquivo Excel (**A_RLM_ex.txt**).

Y	X1	X2	X3	X4
0,25	12,51	36,50	27,00	38,00
0,35	18,12	38,00	24,50	44,50
0,19	10,67	38,50	20,50	36,50
0,25	14,24	38,50	22,50	37,50
0,29	14,95	35,00	22,00	41,50
0,17	8,83	34,50	22,00	29,00
0,18	11,51	36,50	19,50	32,00
0,23	12,13	36,25	18,50	35,00
0,19	10,11	37,25	25,50	36,00
0,27	15,67	35,25	21,00	42,50
0,17	9,76	33,50	19,00	36,00
0,11	7,04	33,00	27,00	15,00
0,19	9,25	34,00	26,00	16,00
0,25	13,43	37,50	32,50	20,50
0,34	15,49	38,50	28,00	28,50
0,25	13,00	36,50	28,50	22,00
0,15	7,91	40,25	35,00	17,50

Entrada dos dados no R.

```
> rlmex <- read.table("C:/EEAR/A_RLM_ex.txt", header=T, dec=",");rlmex
> attach(rlmex)
> Y;X1;X2;X3;X4
```

11.1 Gráficos auxiliares (opcional)

Representar graficamente a matriz de gráficos de dispersão de Y com cada um dos Xi's e entre os Xi´s.

```
> #install.packages("car")
> #require(car)
> par(mfrow=c(1,1))
> pairs(rlm)      # rlm é o nome do data.fame usado
```

11.2 Definição do modelo e estimação dos parâmetros

```
> mod <- lm(Y~X1+X2+ X3+ X4 )
> mod
```

11.3 Análise de Regressão Linear Múltipla e obtenção do R^2.

```
> anova(mod)
> summary.lm(mod)
```

11.4 Diagnósticos para a análise Regressão Linear Múltipla

Os diagnósticos utilizados em Análise de Regressão Múltipla são os mesmos utilizados para regressão linear simples.

Os passos para os Diagnósticos em uma análise de regressão múltipla são:

a) Gráficos de diagnósticos.
b) Análise de Resíduos.
c) Diagnósticos de valores discrepantes (*outliers* e influentes).

Disponíveis no Script R (**S_Diag**).

11.5. Testes de Normalidade dos Erros

Os Testes de Normalidade dos Erros já foram apresentados e utilizados para regressão linear simples.

Disponíveis no Script R (**S_TestNorm**)

Observa-se que para o exemplo(**RLM_ex**) as condições para a análise de Regressão Linear Múltipla foram satisfeitas.

11.6. Seleção do Modelo

Seleção de modelos é um procedimento estatístico para selecionar um modelo contendo apenas as variáveis que influenciam significativamente na variável dependente Y.

Existem vários métodos estatísticos para a seleção de modelos, dentre eles pode-se citar: Forward, Backward, Stepwaise, Mallows, Máximo R^2 etc.

Na literatura sugere-se:

Se K<=15, usar o método Stepwise ou o Backward (K=número de variáveis regressoras).

Se K <<< N, sugere-se o Método Cp (N=número de observações).

Sintaxes dos métodos de seleção de modelos no R.

a) Selecao de Modelo - método Cp (Mallows´Cp) - (Menor Cp - melhor o modelo).

Restrições: Não aceita observações perdidas, se for o caso, tirar todo o registro

Sintaxe:

```
> #install.packages("leaps")
> #require(leaps)
### Complete o cbind com todas as variáveis dependentes ###
> selcp <- leaps(x=cbind(X1,X2,...), y=Y,method=c("Cp"))
> nparcp <- selcp$size
> Cp <- selcp$Cp
> dfr <- data.frame(cbind(nparcp,Cp,selcp$which))
> sdfr <- dfr[order(dfr$Cp),]
> sdfr
```

b) Selecao de Modelo - método R2 Ajustado (Maior R2 - melhor o modelo)

Sintaxe:

```
> #install.packages("leaps")
> require(leaps)
### Complete o cbind com todas as variáveis dependentes ###
> selr2 <- leaps(x=cbind(X1,X2,....), y=Y,method=c("adjr2"))
> nparr2 <- selr2$size
> r2.aj <- selr2$adjr2
> dfr <- data.frame(cbind(nparr2,r2.aj,selr2$which))
> sdfr <- dfr[order(dfr$r2.aj,decreasing="T"),]
> sdfr
```

c) Selecao de Modelo - método backward ou forward ou stepwise.

Seleção é feita pelo valor de AIC (Critério Akaike).

Sintaxe:

```
> step(mod, direction="backward")     # método  (backward).
> step(mod, direction="forward")      # método  (forward).
> step(mod, direction="both")         # método  (stepwise).
```

Um Script no R para a seleção de modelo (**S_SelMod**), disponibilizando os métodos: Mallows, Máximo R^2, Forward, Backward, Stepwaise é apresentado a seguir.

Script R (S_SelMod)

```
#########################################################################
####                    Seleção De Modelos                           ####
#########################################################################
##          Método Cp (Mallows´Cp) - (Menor Cp - melhor o modelo)      ##
##     Não aceita observações perdidas, se for o caso, tirar todo o registro  ##

#install.packages("leaps")
#require(leaps)
# COMPLETE O cbind COM O TOTAL DE VARIAVEIS ESPLANATÓRIAS
selcp <- leaps(x=cbind(X1,X2,...), y=Y,method=c("Cp"))
```

```
nparcp <- selcp$size
Cp <- selcp$Cp
dfr <- data.frame(cbind(nparcp,Cp,selcp$which))
sdfr <- dfr[order(dfr$Cp),]
sdfr

#####          Método R2 Ajustado (Maior R2 - melhor o modelo)       ###
#install.packages("leaps")
require(leaps)
# COMPLETE O cbind COM O TOTAL DE VARIAVEIS ESPLANATÓRIAS
selr2 <- leaps(x=cbind(X1,X2,....), y=Y,method=c("adjr2"))
nparr2 <- selr2$size
r2.aj <- selr2$adjr2
dfr <- data.frame(cbind(nparr2,r2.aj,selr2$which))
sdfr <- dfr[order(dfr$r2.aj,decreasing="T"),]
sdfr

#####          Método backward   ou   forward ou  stepwise            ###
# a seleção é feita pelo valor de AIC (Critério Akaike)
step(mod, direction="backward")            # método (backward).
#step(mod, direction="forward")            # método (forward).
#step(mod, direction="both")               # método (stepwise).
```

Apresenta-se a seguir um Script no R para Análise de Regressão Linear Múltipla (S_RLM).

Script R (S_RLM)

```
#########################################################################
###     Script para Regressão Linear Múltipla - Y = A + B X1 + CX2 + ...     ###
#########################################################################
## Entre com as variáveis Y, X1, X2, ....

#### Inicio do Script para Regressão Linear Múltipla ####
### Imprima as variáveis (complete)
Y; X1; X2; ...

## Plote a matriz de dispersão de Y com cada Xi e entre os Xis(Opcional)
#install.packages("car");
#require(car)
par(mfrow=c(1,1))
pairs(df)    # df=nome do data.fame

## Defina o modelo (complete)
mod <- lm(Y~X1+X2+ ... )
mod

## Análise de Regressão Linear Múltipla e obtenção do R2
anova(mod)
summary.lm(mod)
```

Carregue o Script para Diagnósticos (S_Diag)
 # Gráficos: histograma, boxplot e normalidade; análise do resíduo e Valores extremos

Carregue o Script para testes de Normalidade (S_TestNorm)

```
#####################################################################
## se precisar eliminar registros, use a sintaxe:
# df <- df[-c(nr1,nr2...),]; attach(df)
# onde df=nome do data.frame
#      nr1, nr2  sao os numeros dos registros a serem eliminados
##### repita os procedimentos a partir do Início do Script #####
#####################################################################
```

Carregue o Script para seleção de modelos

Faça a Análise de Regressão do modelo selecionado

```
mod <- lm(Y~X1+X2+ ... )
summary(mod)
```

Atividade(RLM_at). Em um experimento para avaliar quais dos nutrientes: X_1=**N**, X_2=**P**, X_3=**K**, X_4=**Ca**, X_5=**Mg**, X6=**S** influenciam na granulometria do solo. Obteve-se os dados apresentados a seguir, onde **Y =** % terra retida na peneira 18.
Dados disponíveis no site no arquivo **A_RLM_at.txt**.

a) Faça a análise de Regressão Linear Múltipla completa.
b) Apresentar os resultados.
c) Concluir.

12. REGRESSÃO POLINOMIAL

A Regressão Polinomial é o caso que se tem uma variável dependente (Y), uma variável independentes (X). A equação do modelo é um polinômio de grau *k* em X, ou seja:
$Y= B_0 + B_1X + B_2 X^2 + ... + B_kX^k + \varepsilon$,
Onde: *Y* é o vetor das observações; X^1, X^2, ..., X^k as potências de X; $\boldsymbol{\varepsilon}$ o vetor de erros aleatórios (resíduos); B_1, B_2, ...,B_k são os parâmetros do modelo.

A análise de variância é esquematizada como:

F.V.	G.L.	S.Q.	Q.M.	F	*p-value*
Modelo	k	SQ(Mod.)	QM(Mod.)	QM(Mod.) / QM(Res.)	p
Resíduo	N-k-1	SQ(Res.)	QM(Res.)		
Total	N-1	SQ(Tot.)			

F.V. – Fontes de Variação, G.L. – Graus de Liberdade, S.Q. – Somas de Quadrados, Q.M. – Quadrados Médios, N – Número de observações, k – número de variáveis independentes.

A estatística F testa a hipótese: H_0: $B_1=B_2= ... =B_k=0$ vs H_1: $B_i \neq B_{i'}$, para algum $i \neq i'$.

O valor p (***p-value***) é obtido supondo que a estatística F tem uma distribuição F central com K e N-k-1 graus de liberdade. Essa pressuposição é válida se os erros forem iid - independentes e identicamente distribuídos, com distribuição normal $N(0,\sigma^2)$.

Para exemplificar a Análise de Regressão Múltipla, considere o exemplo:

Exemplo(RPOL_ex). Fazer a regressão Polinomial para os dados da quantidade do produto (X) e do tempo que o líquido demora para congelar (Y), apresentados a seguir. Disponíveis no arquivo **A_RPol_ex.txt**.

X	Tempo para Gelar (Y)		X	Tempo para Gelar (Y)	
	Rep.1	Rep.2		Rep.1	Rep.2
2.50	7.39	7.30	2.80	5.97	5.90
2.55	7.00	7.03	2.85	5.90	5.82
2.60	6.90	6.95	2.90	5.80	5.80
2.65	6.85	6.80	2.95	6.15	6.00
2.70	6.70	6.30	3.00	6.30	6.15
2.75	6.33	6.20			

Entrada do dados no R:

```
> rp <- read.table("C:/EEAR/A_RPOL_ex.txt", header=T, dec=","); rp
> attach(rp);
> X;Y
```

12.1 Gráfico auxiliar (opcional)

Representar os pontos graficamente (gráfico de dispersão).

```
> plot(X,Y)
```

12.2 Determinar o grau do polinômio

O grau do polinômio deve ser determinado por critérios estatísticos. O critério mais usado consiste em desdobrar os graus de liberdade do modelo em partes.

Passo 1. Verificar se o grau do polinômio é 1 (primeiro grau). Neste caso o esquema da análise é:

F.V.	G.L.
Modelo	10
Resíduo	11
Total	21

F.V.	G.L.
Linear – X1	1
Desvio de Regressão	9
Resíduo	11
Total	21

Se o Desvio de Regressão for significativo, é que o polinômio tem grau superior a 1. Assim sendo verifica-se se o grau é 2.

Passo 2: Verificar se o grau do polinômio é 2 (segundo grau). Neste caso o esquema da análise é:

F.V.	G.L.
Linear – X1	1
Quadrático – X2	1
Desvio de Regressão	8
Resíduo	11
Total	21

Se o Desvio de Regressão for significativo, é que o polinômio tem grau superior a 2. Assim sendo verifica-se se o grau é 3. Repete-se o processo até obter o Desvio de Regressão não significativo.

Determinar o grau do polinômio com os dados do exemplo(**RPOL_ex**).

```
FX <- as.factor(X)

# Grau 1
X1 <- X
ar1 <- aov(lm(Y~X1+FX))
summary(ar1)
```

Observa-se que o desvio de regressão é significativo p<0,05)

```
# Grau 2
X2 <- X^2
ar2 <- aov(lm(Y~X1+X2+FX))
summary(ar2)
```

Observa-se que o desvio de regressão é significativo (p<0,05)

```
# Grau 3
X3 <- X^3
ar3 <- aov(lm(Y~X1+X2+X3+FX))
summary(ar3)
```

Observa-se que o desvio de regressão é não significativo (p>0,05), logo o polinômio é de grau 3.

Apresenta-se a seguir um Script no R para Análise Determinação do grau do Polonômio (**S_Grau_Pol**).

Script R (S_Grau_Pol)

```
############################################################
###         Script para determinar o Grau do Polinômio     ###
###                Depende das variáveis X e Y             ###
############################################################
FX <- as.factor(X)

# Grau 1
X1 <- X
ar1 <- aov(lm(Y~X1+FX))
summary(ar1)
## Se Desvio de Regressão (FX) for não significativo o Grau é 1, senão verifique o Grau 2.

# Grau 2
X2 <- X^2
ar2 <- aov(lm(Y~X1+X2+FX))
summary(ar2)
## Se Desvio de Regressão (FX) for não significativo o Grau é 2, senão verifique o Grau 3.

# Grau 3
X3 <- X^3
ar3 <- aov(lm(Y~X1+X2+X3+FX))
summary(ar3)
## Se Desvio de Regressão (FX) for não significativo o Grau é 3, senão verifique o Grau 4.


## Repita o processo até obter o Desvio de Regressão Não Significativo.
```

Determinado o grau do polinômio, define-se o modelo.

12.3 Definição do modelo e estimação dos parâmetros

> **mod <- lm(Y~X1+X2+ X3)**
> **mod**

12.4 Análise de Regressão Polinomial e obtenção do R^2.

> **anova(mod)**
> **summary.lm(mod)**

12.5 Diagnósticos para a análise de Regressão Polinomial

Os diagnósticos utilizados em Análise de Regressão Polinomial são os mesmos utilizados para regressão linear simples.
Disponíveis no Script R (**S_Diag**).

12.6. Testes de Normalidade dos Erros

Os Testes de Normalidade dos Erros já foram apresentados e utilizados para regressão linear simples.

Disponíveis no Script R (**S_TestNorm**)

Apresenta-se a seguir um Script no R para Análise de Regressão Polinomial (**S_RPol**).

Script R (S_RPol)

```
#####################################################################
###     Script para Regressão Polinomial - Y = B0 + B1 X1 + B2 X2 + ...    ###
#####################################################################
## Entre com as variáveis X e Y

## Imprima as variàveis
X;Y

## Plote os pontos  (opcional)
par(mfrow=c(1,1))
plot(X,Y)

## Carregue o Script para determinar o grau do Polinomio

## Defina o modelo (complete), Análise de Regressão Polinomial e obtenção do R2
mod <- lm(Y~X1+X2+ ... )
mod; anova(mod); summary.lm(mod)

### Carregue o Script para Diagnósticos

####### Carregue o Script para testes de Normalidade (S_TestNorm)

#####################################################################
## se precisar eliminar registros, use a sintaxe:
```

```
# df <- df[-c(nr1,nr2...),]; attach(df)
# onde df=nome do data.frame
#      nr1, nr2  sao os numeros dos registros a serem eliminados
##### repita os procedimentos a partir do Início do Script #####
#####################################################################

### Represente a curva graficamente
cf <- summary.lm(mod)
plot(X,Y)
## complete de acordo com o grau do polinomio
B0 <-cf$coefficients[1]; B1 <-cf$coefficients[2]; B2 <-cf$coefficients[3];  # ……
curve(B0+B1*x+B2*x^2 +B3*x^3+ …, col="red",add=T)
```

Atividade(RPOL_at). Em um experimento para avaliar consumo (Y) em função do tempo em meses (X) obteve-se os resultados apresentados a seguir. Disponíveis em **A_RPOL_at.txt**:

X	Y	X	Y	X	Y	X	Y
4,0	13,3	4,2	10,8	4,6	9,3	5,0	11,8
4,0	12,8	4,4	10,3	4,6	9,8	5,0	12,3
4,0	12,3	4,4	9,8	4,8	10,3	5,0	12,8
4,2	11,8	4,4	9,3	4,8	10,8	5,2	13,3
4,2	11,3	4,6	8,8	4,8	11,3	5,2	13,8

a) Faça a análise de Regressão Polinomial completa (α=10%).
b) Apresentar os resultados.
c) Concluir.

13. REGRESSÃO NÃO LINEAR

A Regressão Não Linear é o caso que se tem uma variável dependente (Y), uma variável independentes (X). A equação do modelo é do tipo não-linear.

Para exemplificar a Análise de Regressão Não Linear, considere o exemplo:
Exemplo. Os dados referem-se a um experimento de digestibilidade da matéria seca de capim coast-cross, em função do tempo de Incubação. Os dados estão disponíveis em **A_RNL_ex.txt**).

Tempo	Repetições					
de Incub.	1	2	3	4	5	6
0	14,85	14,80	15,07	14,27	20,46	20,07
3	17,39	18,11	18,50	16,03	22,83	21,75
6	22,32	21,04	23,43	19,53	27,94	23,98
12	29,21	30,20	29,78	32,52	36,27	31,78
24	40,73	40,96	41,36	43,88	46,00	39,13
48	47,34	51,34	51,61	53,35	57,07	49,49
72	50,41	54,62	55,72	56,21	59,93	55,12

Deseja-se Estimar os parâmetros do modelo de ORSKOV & McDONALD (1979): $Deg(t)=a+b(1-e^{-ct})$.

Entrada dos dados no R

```
##Entrada dos dados no R
rnl <- read.table("c:/eear/a_rnl_ex.txt", header=T, dec=",");rnl
attach(rnl)
X;Y

##Gráfico auxiliar (opcional) - gráfico de dispersão
plot(Y~X)

##Definição do modelo e estimação dos parâmetros
func <- Y~a+b*(1-exp(-c*X))
mod <- nls(func, start=c(a=15,b=45,c=0.05))
mod

##Diagnósticos para a análise de Regressão Não Linear
# Obtenção dos resíduos e valores preditos
cf <- summary(mod)
ae <- cf$coefficients[1]; be <- cf$coefficients[2]; ce <- cf$coefficients[3]
rp<-residuals(mod)/sd(residuals(mod))
yp<- ae+be*(1-exp(-ce*X)); yp

## Gráfico para Análise do Residuo
par(mfrow=c(1,1))
plot(yp,rp,main="Análise do Resíduo", xlab="Valores preditos", ylab="Studentized residual")
abline(h=0)
```

```
## Testes de Normalidade
rs<-rp

# Teste de Shapiro-Wilk
shapiro.test(rs)

# Teste de Kolmogorov-Smirnov
#install.packages("nortest")
#require(nortest)
lillie.test(rs)

#    Teste Cramer-von Mises
#install.packages("nortest")
#require(nortest)
cvm.test(rs)

#  Teste de Anderson-Darlin
#install.packages("nortest")
#require(nortest)
ad.test(rs)

##Representação Gráfica dos valores observados e modelo ajustado
xmi <- min(X)-.1; xma <- max(X)+.1
ymi <- min(Y)-.1; yma <- max(Y)+.1
plot(c(xmi,xma),c(ymi,yma), "n", main="Gráfico - Regressão Não Linear",
  col.main="blue",  xlab="X", ylab="Y", col.lab="blue")
plot(X,Y)
curve(ae+be*(1-exp(-ce*x)),add=T,col="red")
```