Prof.ª

Paula Shinozaki

Monitora:

Macileide Oliveira

Regressão linear simples com uso do software R

UNIVASF
UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO

Proex
PRÓ-REITORIA DE EXTENSÃO

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
CODAI
COLÉGIO AGRÍCOLA DOM AGOSTINHO IKAS DA UFRPE

Validando o modelo

- Não existe “O” modelo e sim “UM DOS” modelos possíveis;
- Dado o critério escolhido, cabe ao pesquisador defender as razões dessa escolha
- Sempre haverá outros caminhos:
  - Podem levar a outros modelos melhores sob algum critério ou não
  - Sempre escolher o modelo mais plausível e parcimonioso.

Análise de diagnósticos

- Outliers
  - Pontos discrepantes, influências e alavancagem
- Normalidade
- Homocedasticidade

$Y = X\beta + \epsilon$

Pelo método dos mínimos quadrados, temos

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Logo os valores ajustados pelo modelo serão estimados por

$$\hat{Y} = X\hat{\beta}$$

Reescrevendo a equação, temos

$$\hat{Y} = X(X'X)^{-1}X'Y = HY$$

Em que

$H = (X'X)^{-1}X'$

Pontos de alavanca

São valores desproporcionais de X que perturbam a estimativa de Y. Para tal análise faz-se necessário verificar a diagonal principal h_{ii} da matriz de projeção H, dada por

$$H = X(X^T X)^{-1} X$$

Uma perturbação não necessariamente implica em alterar as estimativas dos parâmetros do modelo.

Distância de Cook

Propõe avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo ou nos dados, ao invés da avaliação pela retirada individual ou conjunta de pontos.

Um ponto é dito influência quando a ausência/presença dele no modelo **altera** as estimativas dos parâmetros.

Distância de Cook (D_i)

$$D_\delta = \frac{(\hat{\beta} - \hat{\beta}_\delta)^T X^T X (\hat{\beta} - \hat{\beta}_\delta)}{ps^2}$$

E mede quanto a perturbação $\boldsymbol{\delta} = (\delta_1, \ldots \delta_n)^T$ afasta $\hat{\beta}_\delta$ de $\hat{\beta}$ segundo a métrica $M = X^T X$.

Resíduo Padronizado (Studentized residuals)

O resíduo padronizado é igual ao valor de um resíduo, e_i, dividido por uma estimativa de seu desvio padrão. Resíduos padronizados maiores que 2 e menores que -2, $-2 < e_i < 2$, são geralmente considerados grandes.

É importante analisar não só o comportamento do resíduo em si mas também como se comportam os valores ajustados com esses resíduos.

Gráfico de envelope

Atkinson (1981) propõe a construção, por simulação de Monte Carlo, de uma banda de confiança para os resíduos da regressão normal linear, a qual denominou **envelope**, e que permite uma melhor comparação entre os resíduos e os percentis da distribuição normal padrão.

Gráfico de envelope

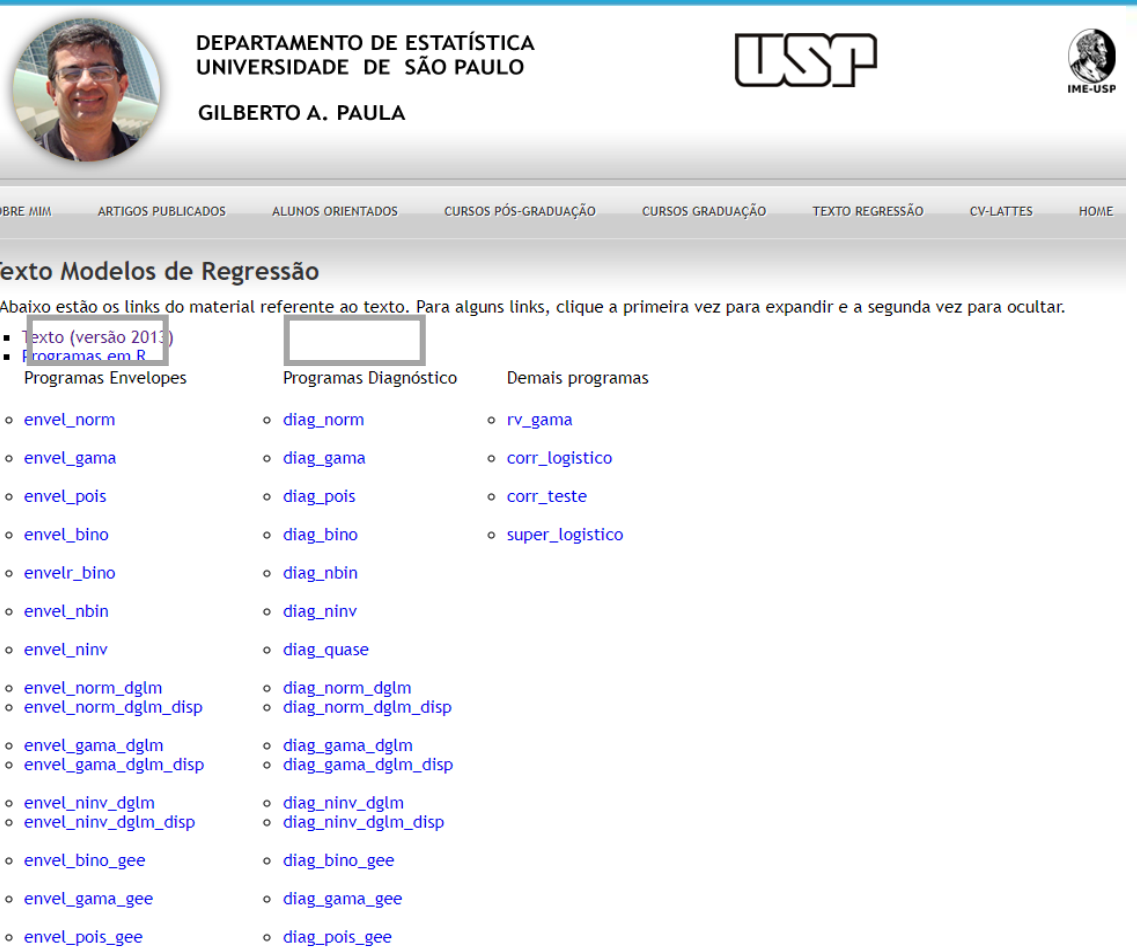
Portanto se a suposição de normalidade dos dados for atendida, espera-se que 95% dos dados encontrem-se dentro das bandas de confiança (envelope).

No entanto, a presença de vários pontos fora do envelope, além de indicar que a suposição de normalidade não foi atendida, pode sugerir a ausência de alguma variável importante no modelo.

Pacotes de diagnósticos

https://www.ime.usp.br/~giapaula/textoregressao.htm

Pacotes de diagnósticos



DEPARTAMENTO DE ESTATÍSTICA
UNIVERSIDADE DE SÃO PAULO

GILBERTO A. PAULA

SOBRE MIM | ARTIGOS PUBLICADOS | ALUNOS ORIENTADOS | CURSOS PÓS-GRADUAÇÃO | CURSOS GRADUAÇÃO | TEXTO REGRESSÃO | CV-LATTES | HOME

Texto Modelos de Regressão

Abaixo estão os links do material referente ao texto. Para alguns links, clique a primeira vez para expandir e a segunda vez para ocultar.

- Texto (versão 2013)
- Programas em R

Programas Envelopes	Programas Diagnóstico	Demais programas
envel_norm	diag_norm	rv_gama
envel_gama	diag_gama	corr_logistico
envel_pois	diag_pois	corr_teste
envel_bino	diag_bino	super_logistico
envelr_bino	diag_nbin	
envel_nbin	diag_ninv	
envel_ninv	diag_quase	
envel_norm_dglm	diag_norm_dglm	
envel_norm_dglm_disp	diag_norm_dglm_disp	
envel_gama_dglm	diag_gama_dglm	
envel_gama_dglm_disp	diag_gama_dglm_disp	
envel_ninv_dglm	diag_ninv_dglm	
envel_ninv_dglm_disp	diag_ninv_dglm_disp	
envel_bino_gee	diag_bino_gee	
envel_gama_gee	diag_gama_gee	
envel_pois_gee	diag_pois_gee	

diag_norm.txt - Bloco de Notas

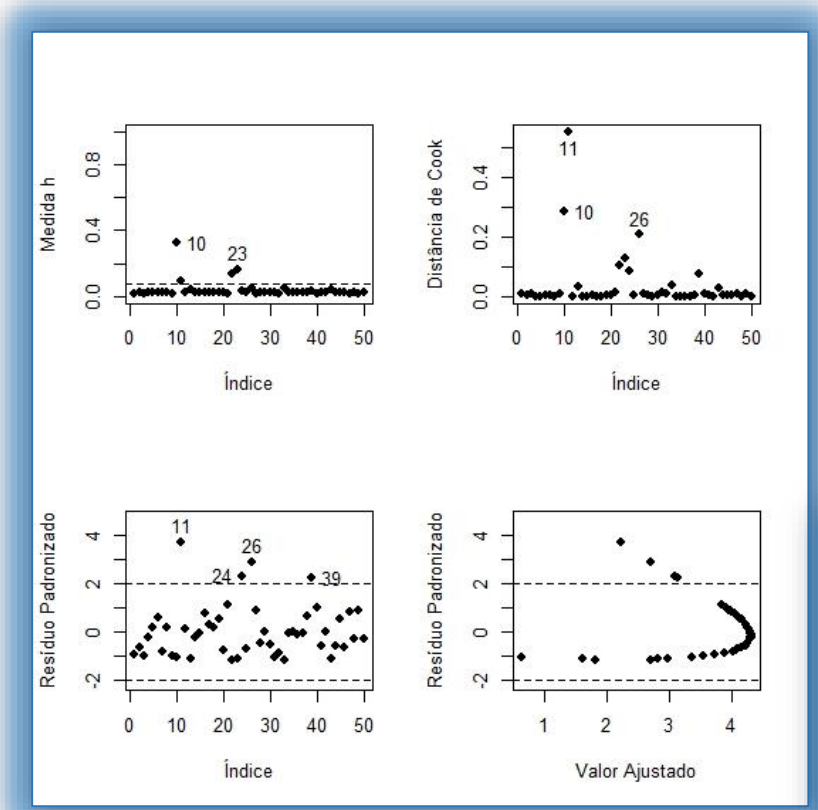
Arquivo Editar Formatar Exibir Ajuda

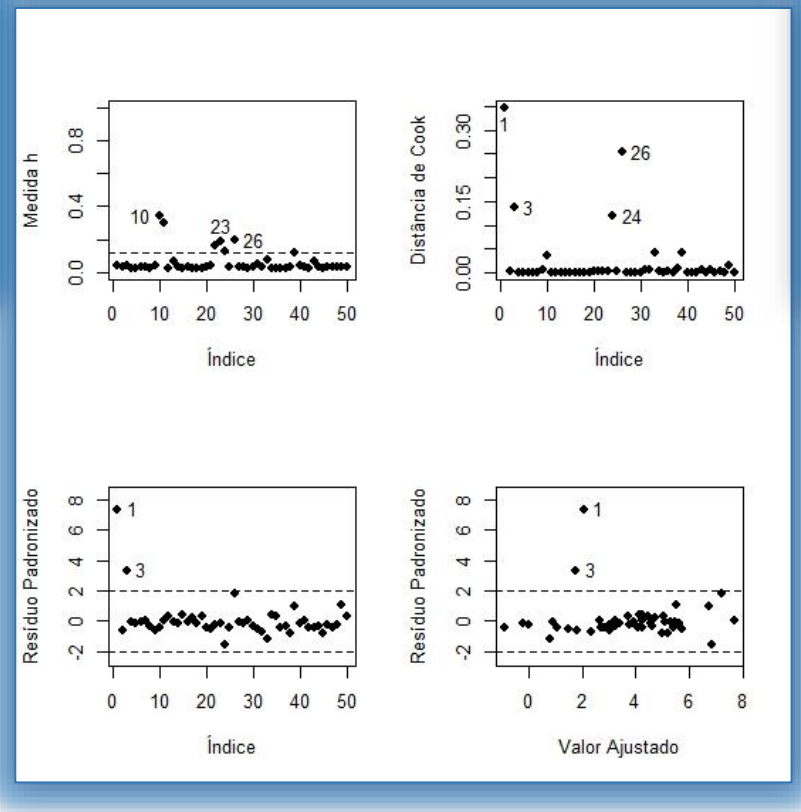
```
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
H <- X%*%solve(t(X)%*%X)%*%t(X)
h <- diag(H)
lms <- summary(fit.model)
s <- lms$sigma
r <- resid(lms)
ts <- r/(s*sqrt(1-h))
di <- (1/p)*(h/(1-h))*(ts^2)
si <- lm.influence(fit.model)$sigma
tsi <- r/(si*sqrt(1-h))
a <- max(tsi)
b <- min(tsi)
par(mfrow=c(2,2))
plot(h,xlab="Índice", ylab="Medida h", pch=16, ylim=c(0,1))
cut <- 2*p/n
abline(cut,0,lty=2)
identify(h, n=4)
#title(sub="(a)")
#
plot(di,xlab="Índice", ylab="Distância de Cook", pch=16)
identify(di, n=4)
#
plot(tsi,xlab="Índice", ylab="Resíduo Padronizado",
ylim=c(b-1,a+1), pch=16)
abline(2,0,lty=2)
abline(-2,0,lty=2)
identify(tsi, n=4)
#
plot(fitted(fit.model),tsi,xlab="Valor Ajustado",
ylab="Resíduo Padronizado", ylim=c(b-1,a+1), pch=16)
abline(2,0,lty=2)
abline(-2,0,lty=2)
identify(fitted(fit.model),tsi, n=4)
par(mfrow=c(1,1))
```

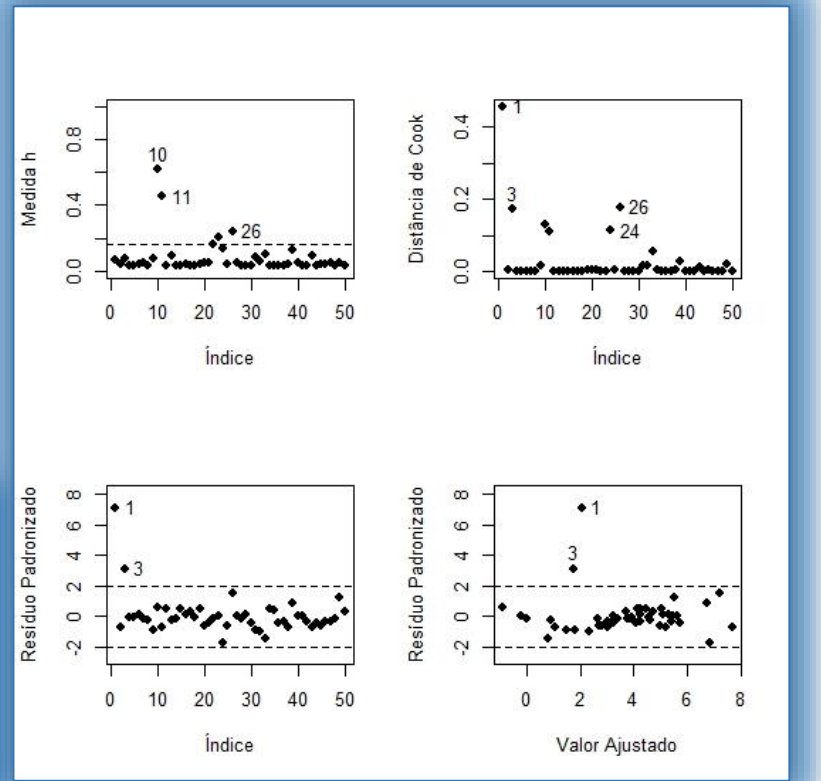
envel_norm.txt - Bloco de Notas

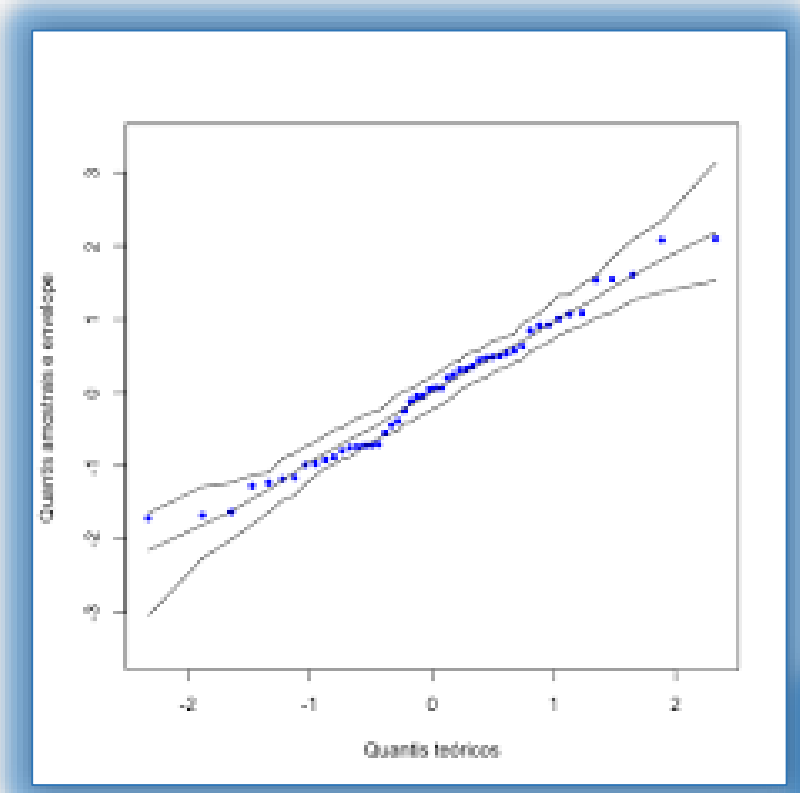
Arquivo Editar Formatar Exibir Ajuda

```
par(mfrow=c(1,1))
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
H <- X%*%solve(t(X)%*%X)%*%t(X)
h <- diag(H)
si <- lm.influence(fit.model)$sigma
r <- resid(fit.model)
tsi <- r/(si*sqrt(1-h))
#
ident <- diag(n)
epsilon <- matrix(0,n,100)
e <- matrix(0,n,100)
e1 <- numeric(n)
e2 <- numeric(n)
#
for(i in 1:100){
     epsilon[,i] <- rnorm(n,0,1)
     e[,i] <- (ident - H)%*%epsilon[,i]
     u <- diag(ident - H)
     e[,i] <- e[,i]/sqrt(u)
     e[,i] <- sort(e[,i]) }
#
for(i in 1:n){
     eo <- sort(e[i,])
     e1[i] <- (eo[2]+eo[3])/2
     e2[i] <- (eo[97]+eo[98])/2 }
#
med <- apply(e,1,mean)
faixa <- range(tsi,e1,e2)
#
par(pty="s")
qqnorm(tsi,xlab="Percentil da N(0,1)",
ylab="Residuo Studentizado", ylim=faixa, pch=16, main="")
par(new=TRUE)
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1, main="")
par(new=TRUE)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1, main="")
par(new=TRUE)
qqnorm(med,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=2, main="")
```


Medida h
Índice
Distância de Cook
Índice
Resíduo Padronizado
Índice
Resíduo Padronizado
Valor Ajustado


Medida h
Índice
Distância de Cook
Índice
Resíduo Padronizado
Índice
Resíduo Padronizado
Valor Ajustado


Medida h
Índice
Distância de Cook
Índice
Resíduo Padronizado
Índice
Resíduo Padronizado
Valor Ajustado


Quantis amostrais e envelope
Quantis teóricos

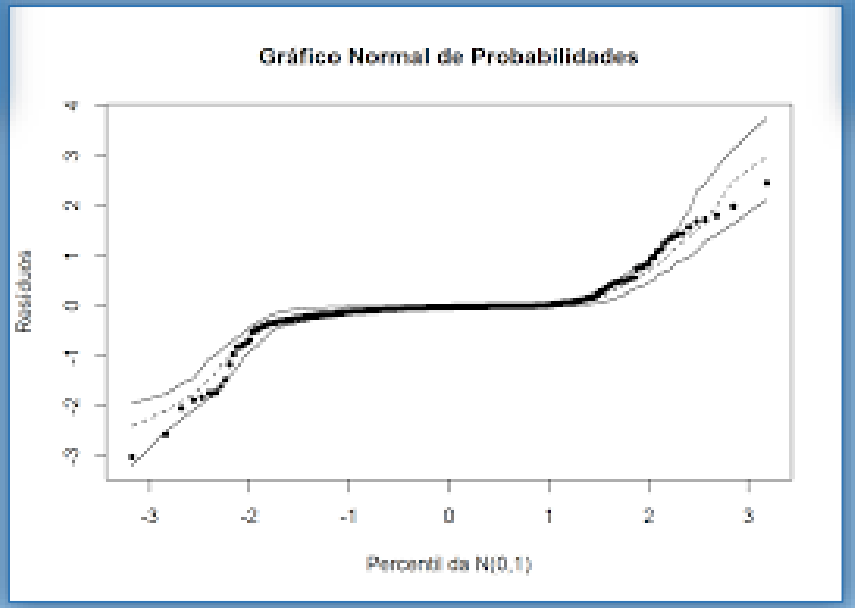
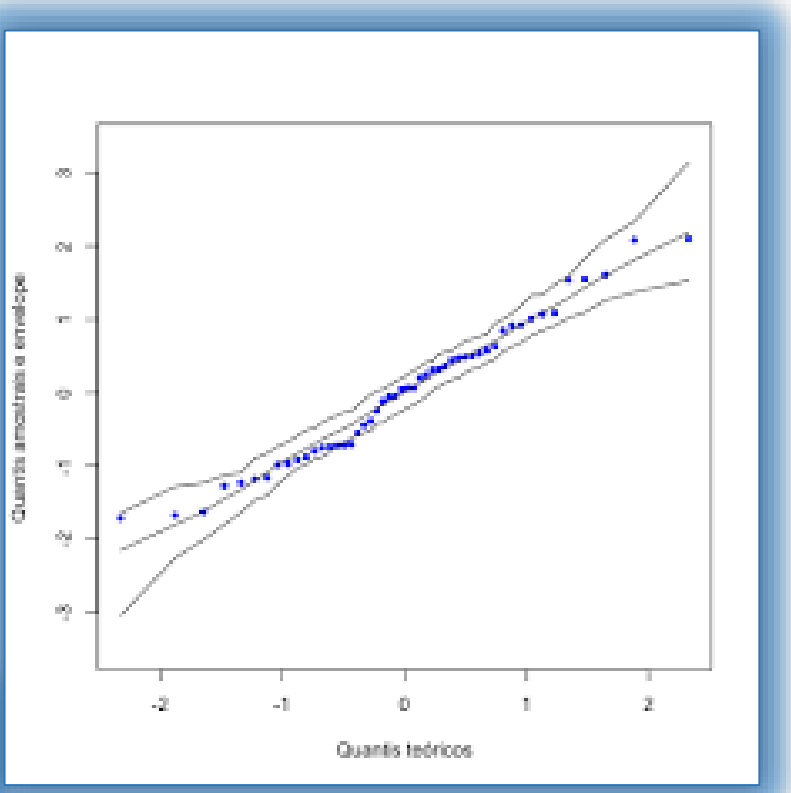
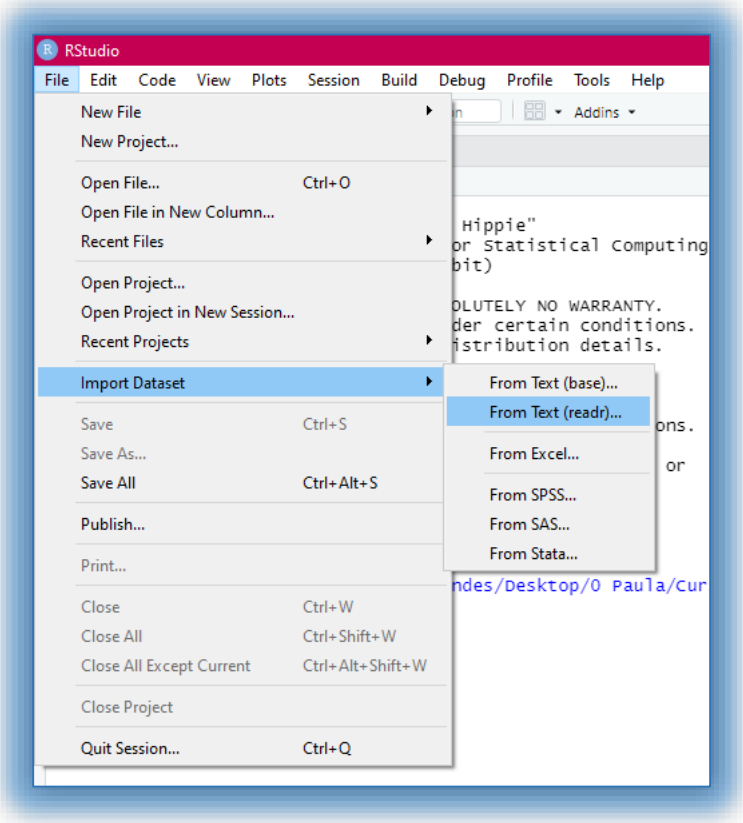
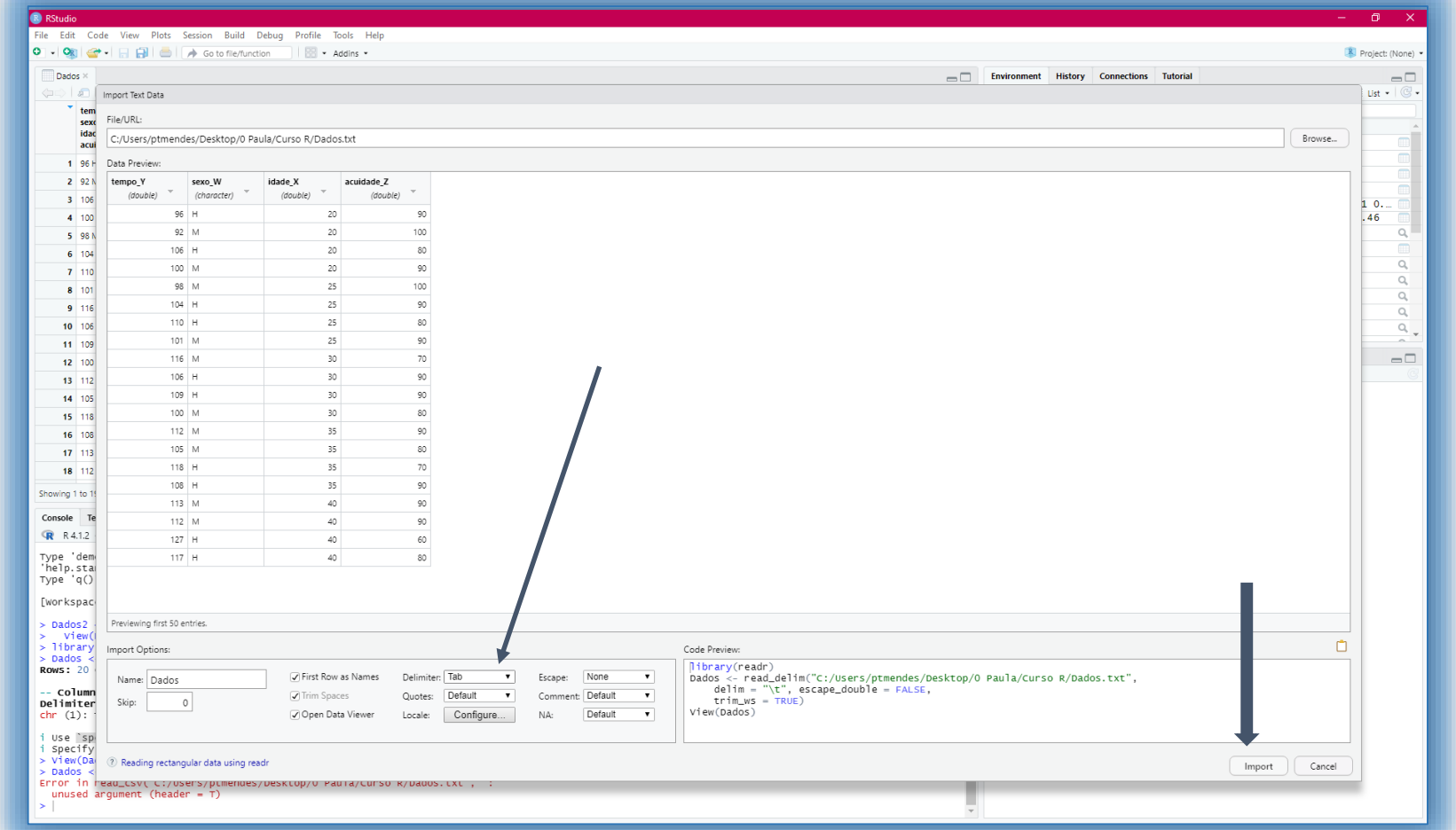


Gráfico Normal de Probabilidades
Resíduos
Percentil da N(0,1)

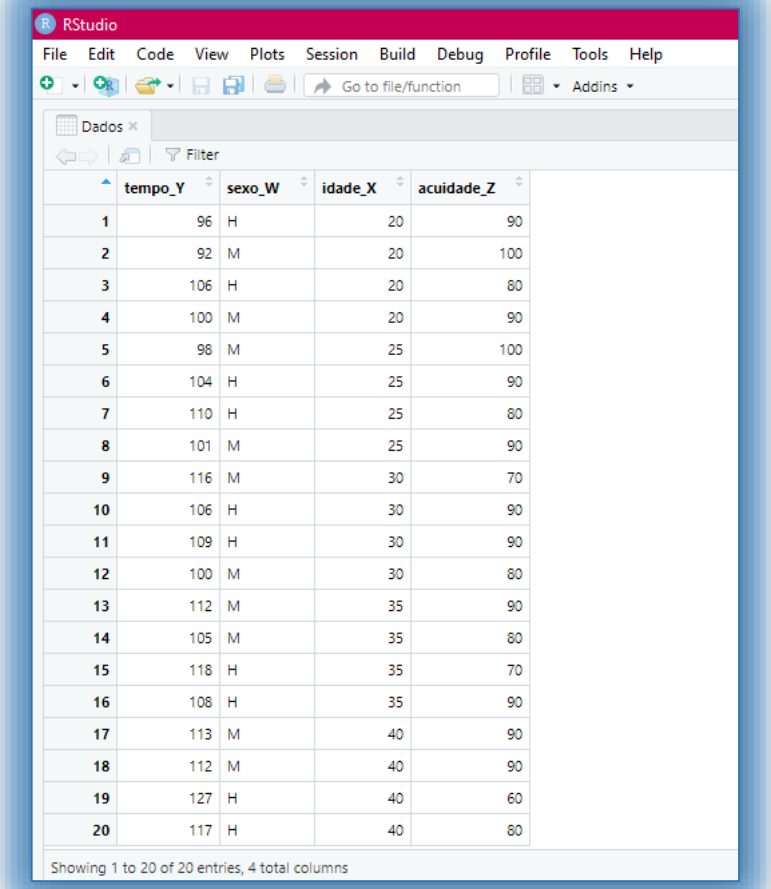

Quantis amostrais e envelope
Quantis teóricos

Carregando os dados





Carregando os dados



	tempo_Y	sexo_W	idade_X	acuidade_Z
1	96	H	20	90
2	92	M	20	100
3	106	H	20	80
4	100	M	20	90
5	98	M	25	100
6	104	H	25	90
7	110	H	25	80
8	101	M	25	90
9	116	M	30	70
10	106	H	30	90
11	109	H	30	90
12	100	M	30	80
13	112	M	35	90
14	105	M	35	80
15	118	H	35	70
16	108	H	35	90
17	113	M	40	90
18	112	M	40	90
19	127	H	40	60
20	117	H	40	80

Showing 1 to 20 of 20 entries, 4 total columns

- Função attach: Indexa os dados

$$attach(Dados)$$

```
> summary(Dados)
    tempo_Y         sexo_W              idade_X     acuidade_Z
 Min.   : 92.0   Length:20          Min.   :20   Min.   : 60
 1st Qu.:100.8   Class :character   1st Qu.:25   1st Qu.: 80
 Median :107.0   Mode  :character   Median :30   Median : 90
 Mean   :107.5                      Mean   :30   Mean   : 85
 3rd Qu.:112.2                      3rd Qu.:35   3rd Qu.: 90
 Max.   :127.0                      Max.   :40   Max.   :100
> 
```

Modelo linear

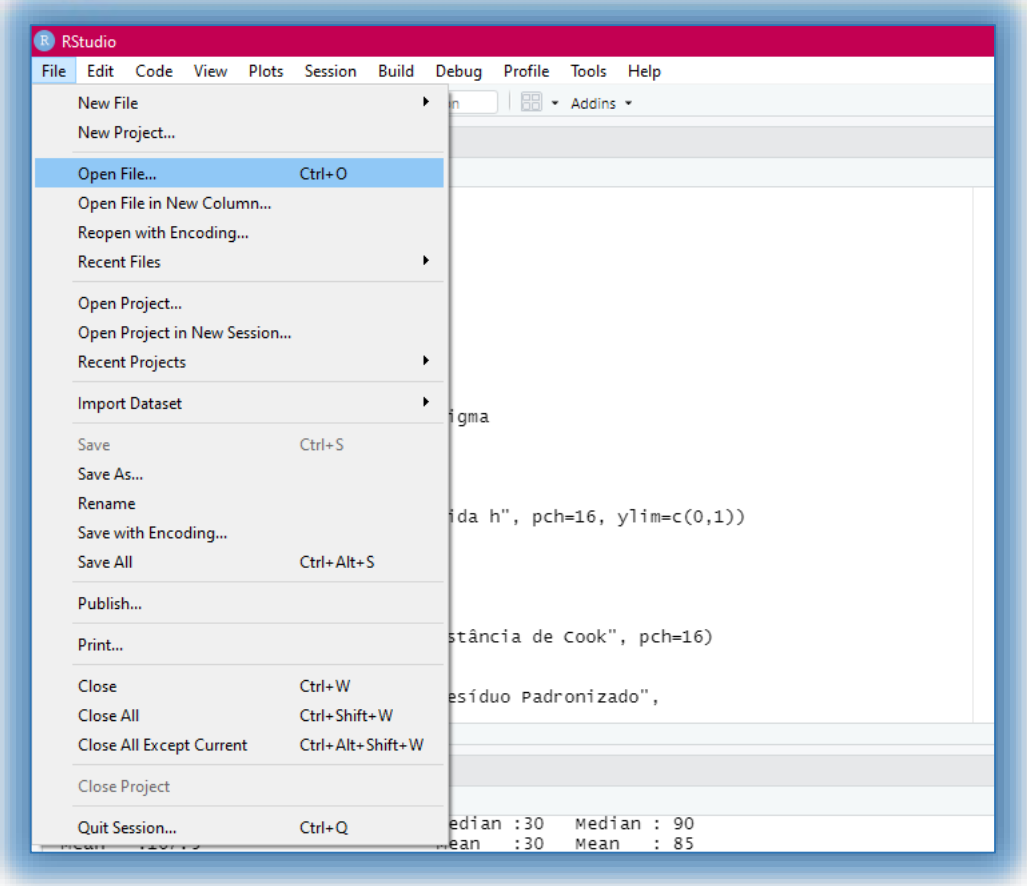
```
fit.model<- lm(acuidade_Z~tempo_Y)
fit.model
```

```
> fit.model<-lm(acuidade_Z~tempo_Y)
> fit.model

Call:
lm(formula = acuidade_Z ~ tempo_Y)

Coefficients:
(Intercept)      tempo_Y
   180.5208       -0.8886
```

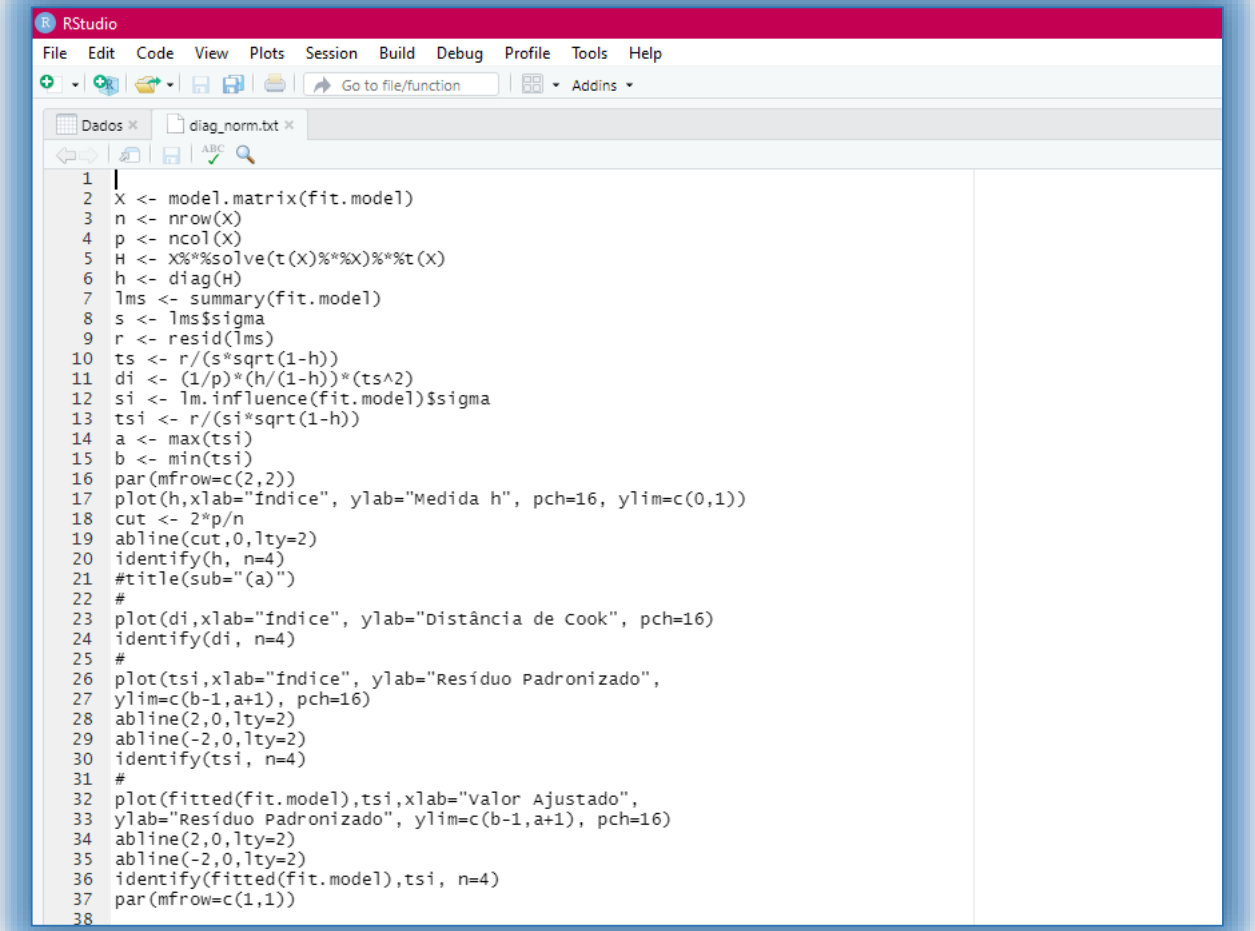
```
summary(fit.model)
```

```
> summary(fit.model)

Call:
lm(formula = acuidade_Z ~ tempo_Y)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6642  -5.8358   0.2258   5.6664   9.8871

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 180.5208    19.5909   9.215  3.1e-08 ***
tempo_Y      -0.8886     0.1817  -4.890 0.000118 ***
---
```

Diagnósticos



“abrir” o identify

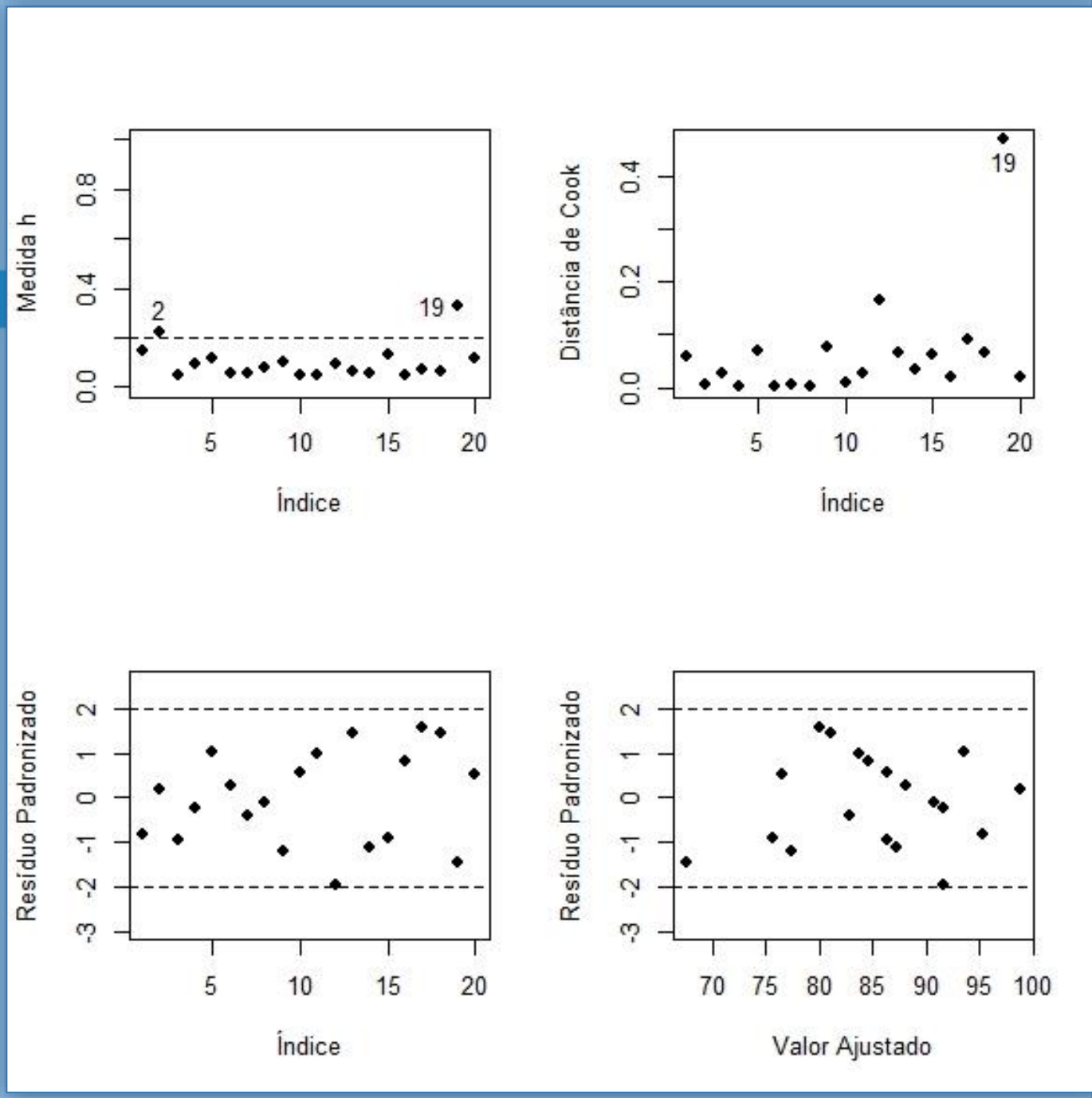


Diagnósticos

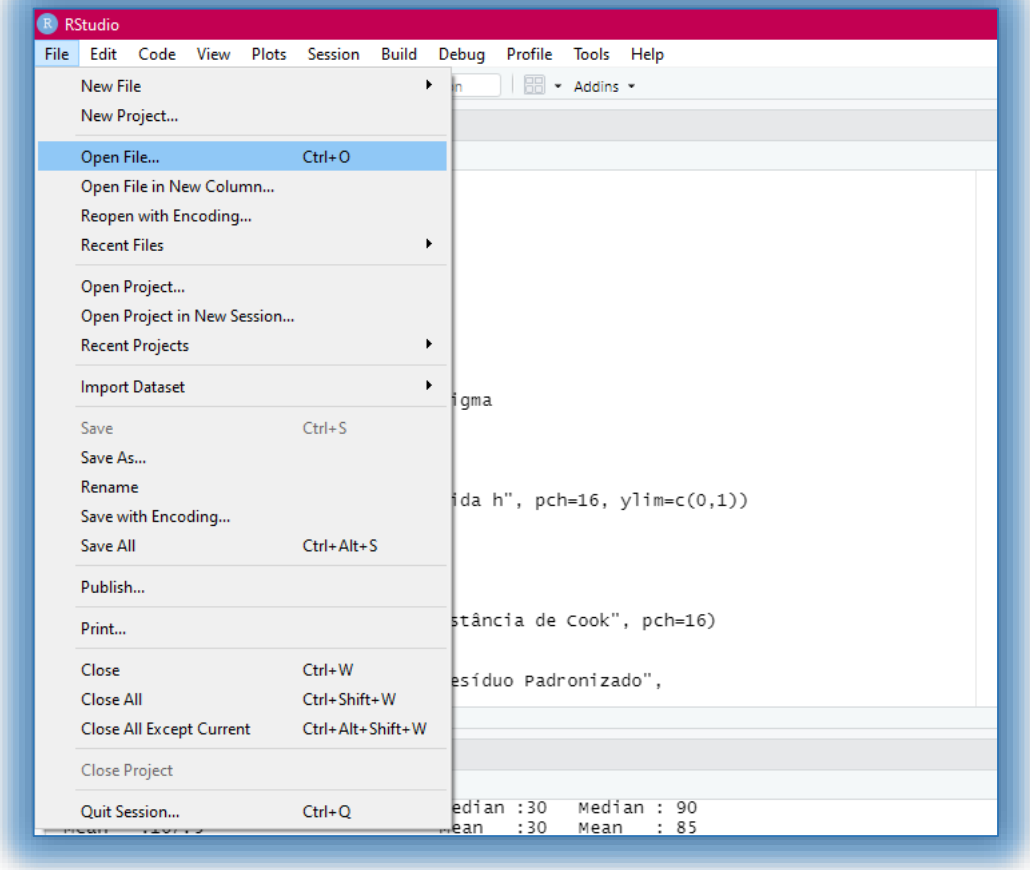
- Observações 19 e 2 são aberrantes;

- Observação 19 é influente

- Comportamento adequado dos resíduos



Envelope



envel_norm.txt - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
par(mfrow=c(1,1))
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
H <- X%*%solve(t(X)%*%X)%*%t(X)
h <- diag(H)
si <- lm.influence(fit.model)$sigma
r <- resid(fit.model)
tsi <- r/(si*sqrt(1-h))
#
ident <- diag(n)
epsilon <- matrix(0,n,100)
e <- matrix(0,n,100)
e1 <- numeric(n)
e2 <- numeric(n)
#
for(i in 1:100){
     epsilon[,i] <- rnorm(n,0,1)
     e[,i] <- (ident - H)%*%epsilon[,i]
     u <- diag(ident - H)
     e[,i] <- e[,i]/sqrt(u)
     e[,i] <- sort(e[,i]) }
#
for(i in 1:n){
     eo <- sort(e[i,])
     e1[i] <- (eo[2]+eo[3])/2
     e2[i] <- (eo[97]+eo[98])/2 }
#
med <- apply(e,1,mean)
faixa <- range(tsi,e1,e2)
#
par(pty="s")
qqnorm(tsi,xlab="Percentil da N(0,1)",
ylab="Residuo Studentizado", ylim=faixa, pch=16, main="")
par(new=TRUE)
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1, main="")
par(new=TRUE)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1, main="")
par(new=TRUE)
qqnorm(med,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=2, main="")
```

Envelope

- Todas as observações encontram-se dentro das bandas de confianças.

- Indicando bom ajuste do modelos, ou seja, o modelo captou adequadamente a variabilidade das variáveis

- As suposições de normalidade foram atendidas