NORM_COL 2 - Data set: ABALONE

INTRODUZIONE

Il data set contiene informazioni riguardanti 4177 misurazioni relative ad "abaloni" con lo scopo di prevederne l'età. Gli attributi sono i seguenti:

1. SEX: M, F
2. LENGTH: lunghezza
3. DIAMETER: diametro
4. HEIGHT: altezza
5. WHOLE_WEIGHT: peso intero
6. SHUCKED_WEIGHT: peso della carne
7. VISCERA_WEIGHT: peso delle viscere
8. SHELL_WEIGHT: peso del guscio
9. RINGS: età in anni

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare

```r
#-- R CODE

library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
panderOptions('knitr.auto.asis', FALSE)

#-- White test function
white.test <- function(lmod,data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM,"P value"=p.value)
}

#-- funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVARION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\2.Norm-Col copy(3)\\2.Norm-Col\\dati.csv")
d <- na.omit(d)
names(d) <- c("SEX","LENGTH","DIAMETER","HEIGHT","WHOLE_WEIGHT","SHUCKED_WEIGHT","VISCERA_WEIGHT","SHEL
```

```r
#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- names(d)[2:ncol(d)]

#-- print delle prime 6 righe del dataset
pander(head(d),big.mark=",")
```

Table 1: Table continues below

	SEX	LENGTH	DIAMETER	HEIGHT	WHOLE_WEIGHT	SHUCKED_WEIGHT
1	M	0.455	0.365	0.095	0.514	0.2245
3	M	0.35	0.265	0.09	0.2255	0.0995
5	F	0.53	0.42	0.135	0.677	0.2565
7	M	0.44	0.365	0.125	0.516	0.2155
9	I	0.33	0.255	0.08	0.205	0.0895
11	I	0.425	0.3	0.095	0.3515	0.141

	VISCERA_WEIGHT	SHELL_WEIGHT	RINGS
1	0.101	0.15	15
3	0.0485	0.07	7
5	0.1415	0.21	9
7	0.114	0.155	10
9	0.0395	0.055	7
11	0.0775	0.12	8

STATISTICHE DESCRITTIVE

Si presentano innanzitutto le statistiche descrittive.

```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

Table 3: Table continues below

LENGTH	DIAMETER	HEIGHT	WHOLE_WEIGHT
Min. :0.075	Min. :0.0550	Min. :0.0000	Min. :0.0020
1st Qu.:0.450	1st Qu.:0.3500	1st Qu.:0.1150	1st Qu.:0.4415
Median :0.545	Median :0.4250	Median :0.1400	Median :0.7995
Mean :0.524	Mean :0.4079	Mean :0.1395	Mean :0.8287
3rd Qu.:0.615	3rd Qu.:0.4800	3rd Qu.:0.1650	3rd Qu.:1.1530
Max. :0.815	Max. :0.6500	Max. :1.1300	Max. :2.8255

SHUCKED_WEIGHT	VISCERA_WEIGHT	SHELL_WEIGHT	RINGS
Min. :0.0010	Min. :0.0005	Min. :0.0015	Min. : 1.000
1st Qu.:0.1860	1st Qu.:0.0935	1st Qu.:0.1300	1st Qu.: 8.000
Median :0.3360	Median :0.1710	Median :0.2340	Median : 9.000

SHUCKED_WEIGHT	VISCERA_WEIGHT	SHELL_WEIGHT	RINGS
Mean :0.3594	Mean :0.1806	Mean :0.2388	Mean : 9.934
3rd Qu.:0.5020	3rd Qu.:0.2530	3rd Qu.:0.3290	3rd Qu.:11.000
Max. :1.4880	Max. :0.7600	Max. :1.0050	Max. :29.000

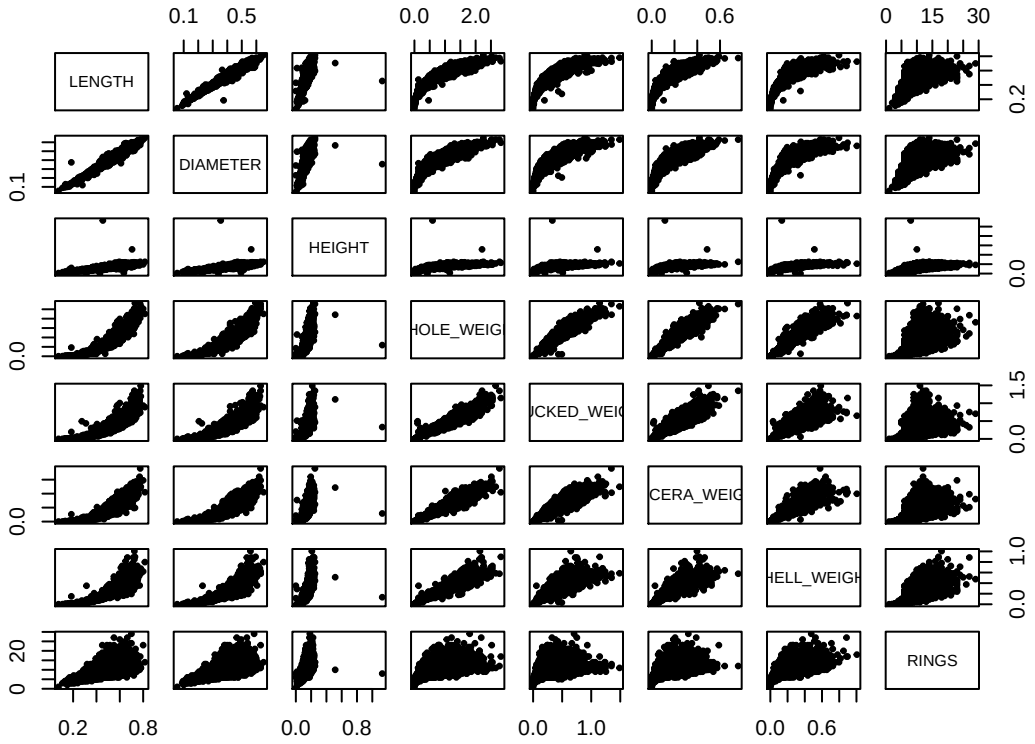
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 5: Table continues below

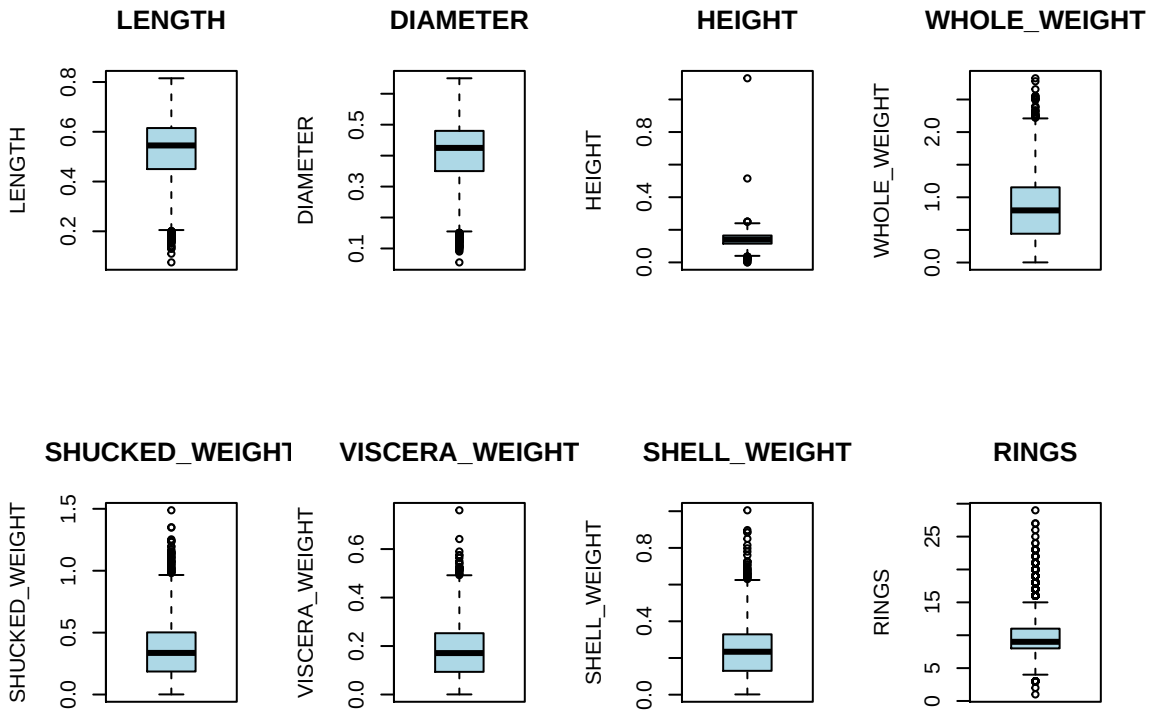
	LENGTH	DIAMETER	HEIGHT	WHOLE_WEIGHT
LENGTH	1	0.9868	0.8276	0.9253
DIAMETER	0.9868	1	0.8337	0.9255
HEIGHT	0.8276	0.8337	1	0.8192
WHOLE_WEIGHT	0.9253	0.9255	0.8192	1
SHUCKED_WEIGHT	0.8979	0.8932	0.775	0.9694
VISCERA_WEIGHT	0.903	0.8997	0.7983	0.9664
SHELL_WEIGHT	0.8977	0.9053	0.8173	0.9554
RINGS	0.5567	0.5747	0.5575	0.5404

	SHUCKED_WEIGHT	VISCERA_WEIGHT	SHELL_WEIGHT	RINGS
LENGTH	0.8979	0.903	0.8977	0.5567
DIAMETER	0.8932	0.8997	0.9053	0.5747
HEIGHT	0.775	0.7983	0.8173	0.5575
WHOLE_WEIGHT	0.9694	0.9664	0.9554	0.5404
SHUCKED_WEIGHT	1	0.932	0.8826	0.4209
VISCERA_WEIGHT	0.932	1	0.9077	0.5038
SHELL_WEIGHT	0.8826	0.9077	1	0.6276
RINGS	0.4209	0.5038	0.6276	1

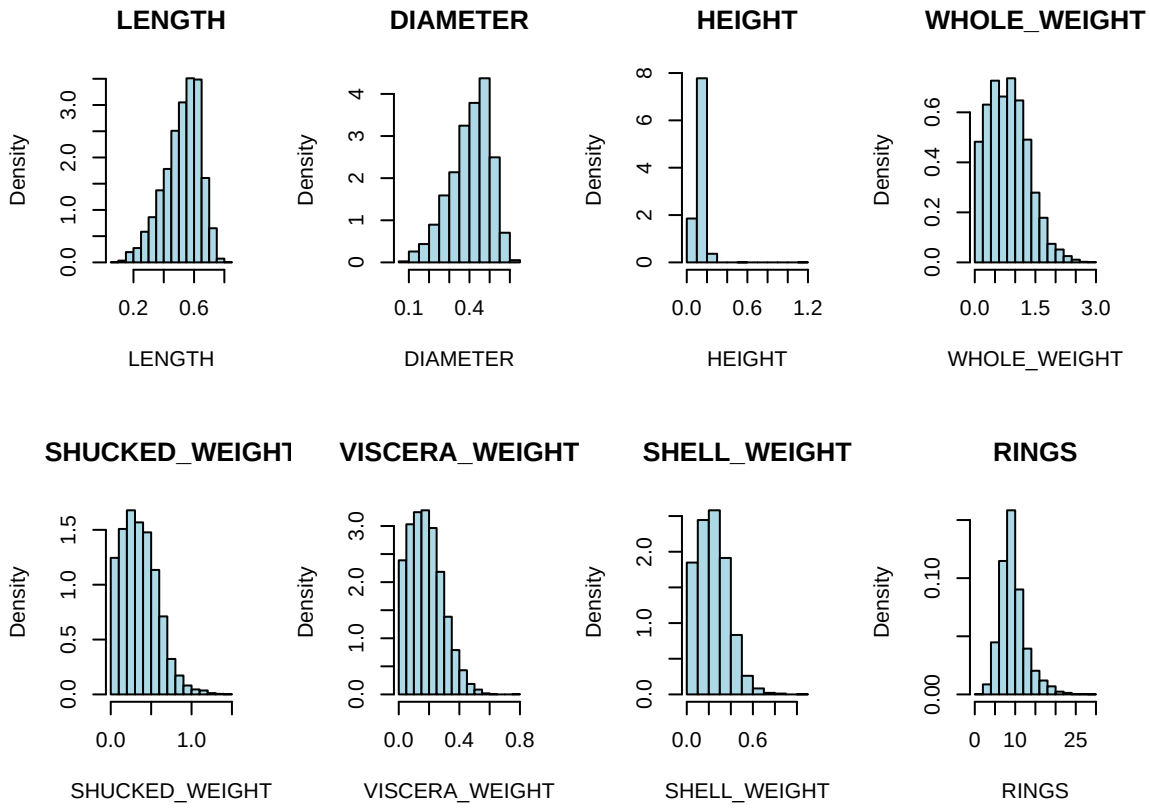
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```
par(mfrow=c(2,4))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```



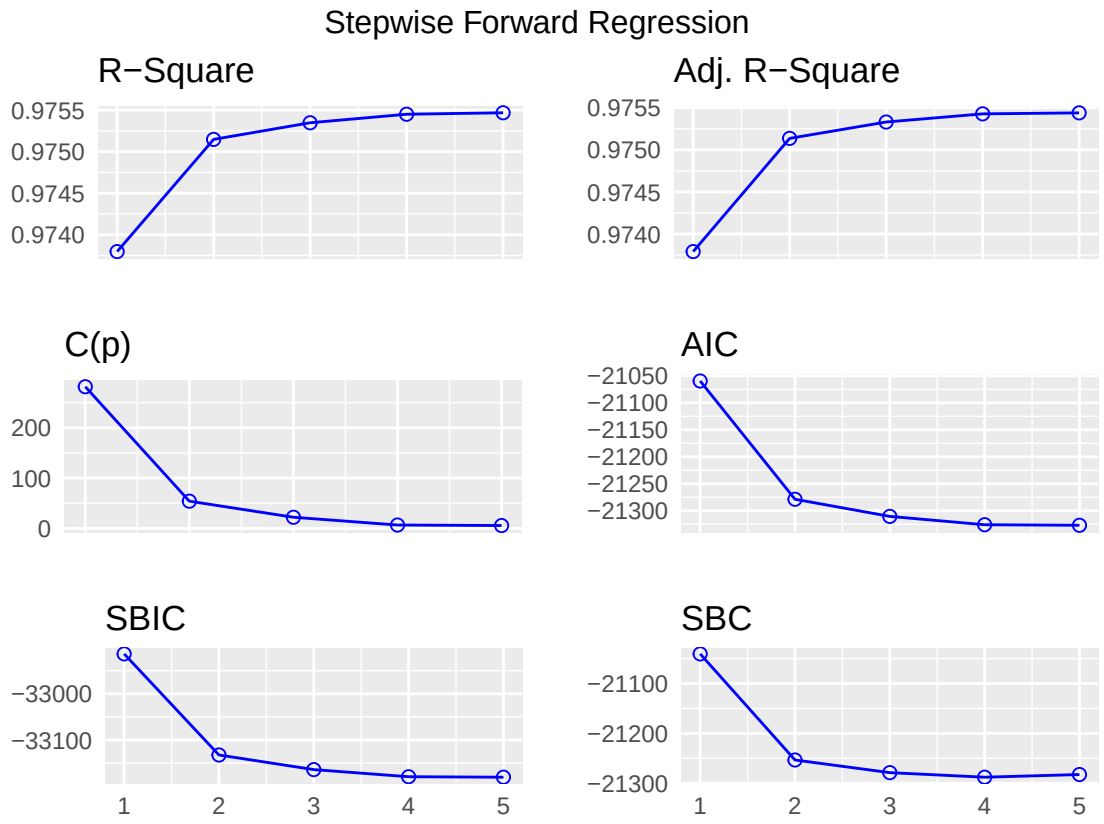

```
par(mfrow=c(2,4))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```



REGRESSIONE

Si effettua ora una regressione stepwise per scegliere quali variabili regredire rispetto alla variabile dipendente "lenght".

```
#-- R CODE
mod1 <- lm(LENGTH ~ DIAMETER +HEIGHT +WHOLE_WEIGHT+SHUCKED_WEIGHT+VISCERA_WEIGHT+SHELL_WEIGHT+RINGS, d)

step <- ols_step_forward(mod1)
plot(step)
```




```
mod2 <- lm(LENGTH ~ DIAMETER + SHUCKED_WEIGHT + VISCERA_WEIGHT + SHELL_WEIGHT, d)

pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.05778	0.002068	27.93	1.511e-157
DIAMETER	1.103	0.007981	138.2	0
SHUCKED_WEIGHT	0.03098	0.003873	7.998	1.621e-15
VISCERA_WEIGHT	0.05976	0.008616	6.936	4.646e-12
SHELL_WEIGHT	-0.02392	0.005722	-4.18	2.973e-05

Table 8: Fitting linear model: LENGTH ~ DIAMETER + SHUCKED_WEIGHT + VISCERA_WEIGHT + SHELL_WEIGHT

Observations	Residual Std. Error	R^2	Adjusted R^2
4177	0.01883	0.9755	0.9754

```
pander(anova(mod2),big.mark=",")
```

Table 9: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DIAMETER	1	58.65	58.65	165,492	0
SHUCKED_WEIGHT	1	0.08137	0.08137	229.6	1.577e-50
VISCERA_WEIGHT	1	0.01204	0.01204	33.97	6.024e-09
SHELL_WEIGHT	1	0.006192	0.006192	17.47	2.973e-05
Residuals	4,172	1.479	0.0003544	NA	NA

```
pander(white.test(mod2),big.mark=",")
```

Test.statistic	P.value
10.93	0.004225

```
pander(dwtest(mod2),big.mark=",")
```

Table 11: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.777	2.294e-13 * * *	true autocorrelation is greater than 0

```
pander(ols_vif_tol(mod2),big.mark=",")
```

Variables	Tolerance	VIF
DIAMETER	0.1353	7.392
SHUCKED_WEIGHT	0.1148	8.709
VISCERA_WEIGHT	0.09515	10.51
SHELL_WEIGHT	0.1338	7.475

```
pander(ols_eigen_cindex(mod2),big.mark=",")
```

Table 13: Table continues below

Eigenvalue	Condition Index	intercept	DIAMETER	SHUCKED_WEIGHT
4.719	1	0.0007768	0.000331	0.001357
0.2268	4.562	0.04862	0.003115	0.01899
0.03161	12.22	0.003479	0.0002698	0.343
0.01734	16.5	0.001168	0.0006091	0.5628
0.005294	29.85	0.946	0.9957	0.07379

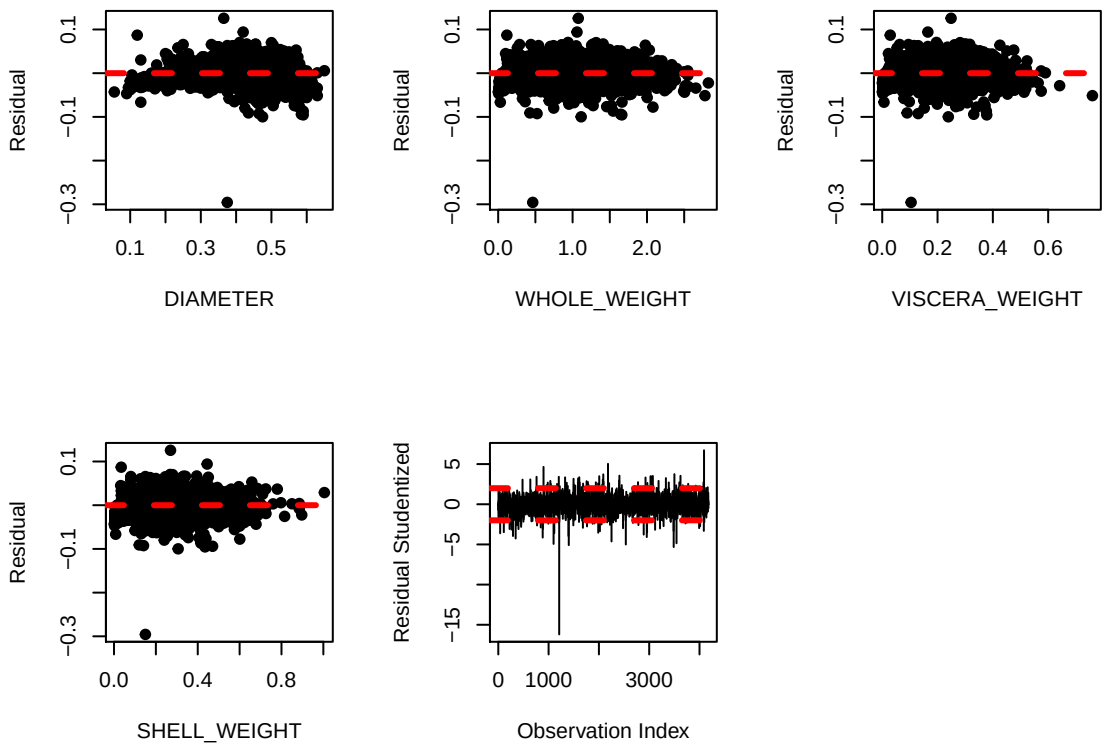
VISCERA_WEIGHT	SHELL_WEIGHT
0.001103	0.001462
0.01456	0.01282
0.02322	0.6681
0.9364	0.1087

VISCERA_WEIGHT	SHELL_WEIGHT
0.02469	0.2089

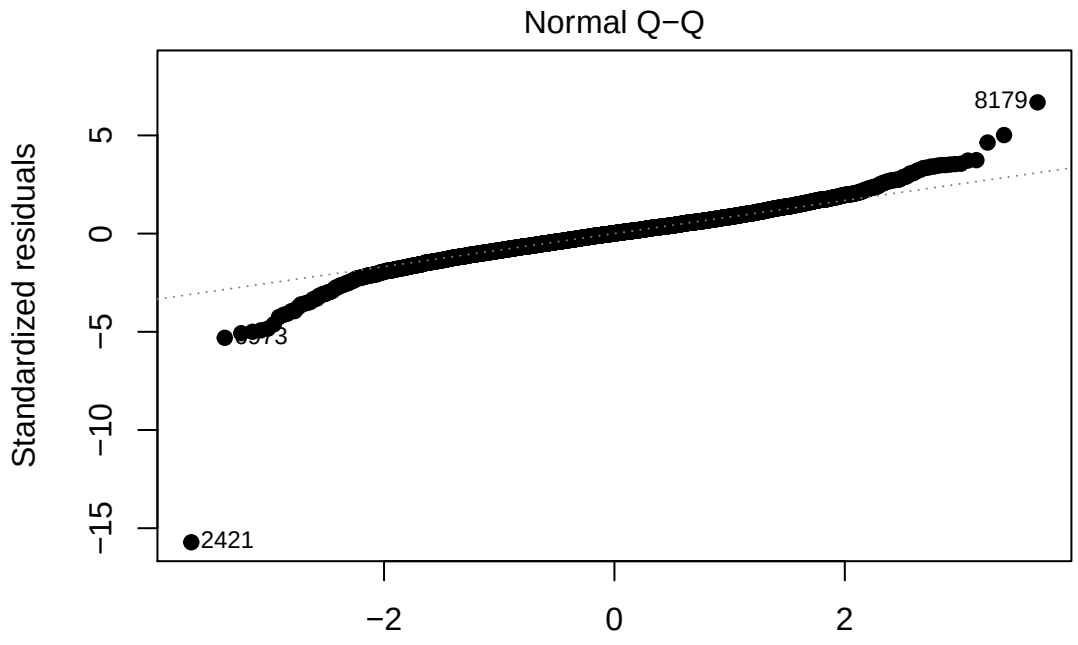
Il test F sulla bontà globale del modello porta a respingere l'ipotesi nulla di non significatività globale. L'R^2 ha un valore superiore a 0.97 quindi le variabili esplicative spiegano molto bene la varianza della variabile esplicativa.

Si passa ora all'analisi della collinearità per verificare se fra le variabili esplicative non vi siano variabili troppo correlate che potrebbero inficiare la validità dei risultati. La variance inflation e il condition index sono sempre inferiori alle rispettive soglie (rispettivamente 20 e 30). Non c'è quindi collinearità. Si può quindi passare alla analisi della normalità dei residui cominciando dall'analisi degli scatter plot e della loro densità.

```r
#-- R CODE
par(mfrow=c(2,3))
plot(d$DIAMETER,resid(mod1),pch=19,xlab="DIAMETER",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$WHOLE_WEIGHT,resid(mod1),pch=19,xlab="WHOLE_WEIGHT",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$VISCERA_WEIGHT,resid(mod1),pch=19,xlab="VISCERA_WEIGHT",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$SHELL_WEIGHT,resid(mod1),pch=19,xlab="SHELL_WEIGHT",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(1:nrow(d),rstudent(mod1),pch=19,xlab="Observation Index",ylab="Residual Studentized",type="h")
abline(h=2,lwd=3,lty=2,col=2)
abline(h=-2,lwd=3,lty=2,col=2)
```

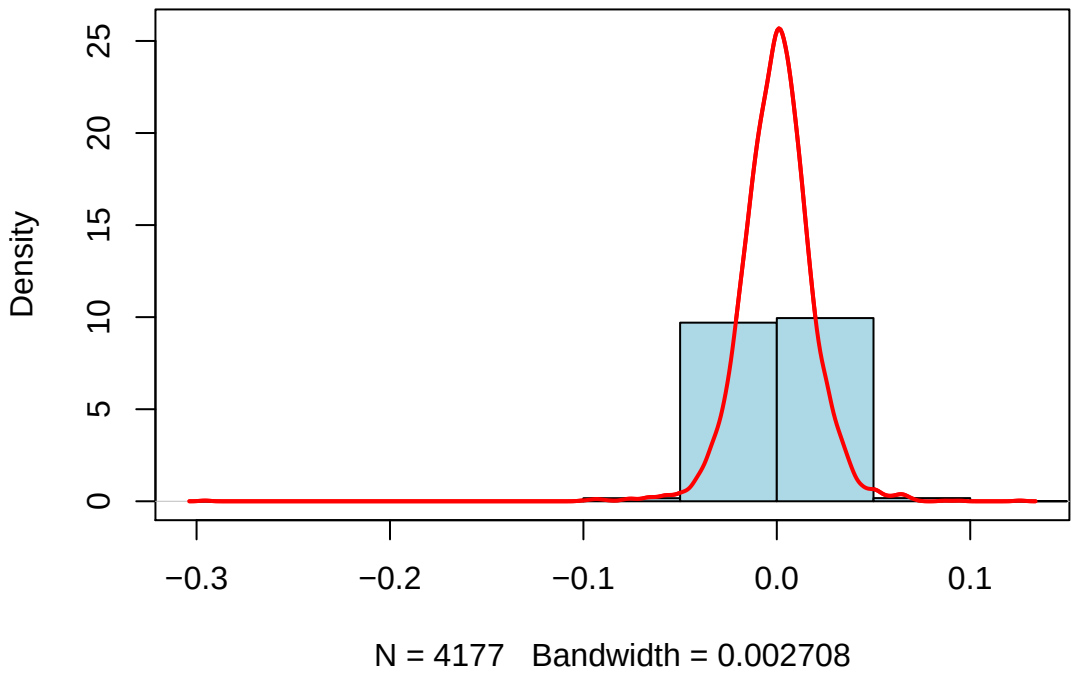



```
#-- R CODE

plot(mod1,which=2,pch=19)
```



GTH ~ DIAMETER + HEIGHT + WHOLE_WEIGHT + SHUCKED_WEIGHT + VISC

```
plot(density(resid(mod1)),col=2,lwd=2,main="")
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="",add=T)
lines(density(resid(mod1)),col=2,lwd=2)
```




```
pander(shapiro.test(resid(mod1)))
```

Table 15: Shapiro-Wilk normality test: `resid(mod1)`

Test statistic	P value
0.9428	1.754e-37 * * *

```
pander(ks.test(resid(mod1),"pnorm"))
```

Table 16: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

Test statistic	P value	Alternative hypothesis
0.4728	0 * * *	two-sided

Sia l'esame dgli scatter plot che l'andamento della distribuzione non perfettamente simmetrica pongono dubbi sulla reale normalità dei residui.

Si devono individuare gli outliers (ovvero gli individui che almeno in un box-plot sono individuati da pallini che si discostano in modo netto dal resto della distribuzione). Si passa quindi a ristimare il modello senza outlier.

```r
#-- R CODE

out <- c(
  which(d$LENGTH%in%boxplot(d$LENGTH,plot=F)$out),
  which(d$DIAMETER%in%boxplot(d$DIAMETER,plot=F)$out),
  which(d$SHUCKED_WEIGHT%in%boxplot(d$SHUCKED_WEIGHT,plot=F)$out),
  which(d$VISCERA_WEIGHT%in%boxplot(d$VISCERA_WEIGHT,plot=F)$out),
  which(d$SHELL_WEIGHT%in%boxplot(d$SHELL_WEIGHT,plot=F)$out)
)
out <- unique(out)
d_noout <- d[-out,]

mod2 <- lm(LENGTH ~ DIAMETER + SHUCKED_WEIGHT + VISCERA_WEIGHT + SHELL_WEIGHT, d_noout)

pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.07109	0.002367	30.03	5.436e-179
DIAMETER	1.053	0.009241	114	0
SHUCKED_WEIGHT	0.04327	0.004104	10.54	1.169e-25
VISCERA_WEIGHT	0.06785	0.008745	7.758	1.083e-14
SHELL_WEIGHT	-0.01586	0.006448	-2.46	0.01395

Table 18: Fitting linear model: LENGTH ~ DIAMETER + SHUCKED_WEIGHT + VISCERA_WEIGHT + SHELL_WEIGHT

Observations	Residual Std. Error	R^2	Adjusted R^2
4031	0.01789	0.974	0.9739

```r
pander(anova(mod2),big.mark=",")
```

Table 19: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DIAMETER	1	48.08	48.08	150,229	0
SHUCKED_WEIGHT	1	0.106	0.106	331.1	3.613e-71
VISCERA_WEIGHT	1	0.01739	0.01739	54.35	2.028e-13
SHELL_WEIGHT	1	0.001936	0.001936	6.05	0.01395
Residuals	4,026	1.288	0.00032	NA	NA

```r
pander(white.test(mod2),big.mark=",")
```

Test.statistic	P.value
66.86	2.998e-15

```
pander(dwtest(mod2),big.mark=",")
```

Table 21: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.82	4.427e-09 * * *	true autocorrelation is greater than 0

```
pander(ols_vif_tol(mod2),big.mark=",")
```

Variables	Tolerance	VIF
DIAMETER	0.1099	9.103
SHUCKED_WEIGHT	0.1157	8.645
VISCERA_WEIGHT	0.1005	9.95
SHELL_WEIGHT	0.1184	8.446

```
pander(ols_eigen_cindex(mod2),big.mark=",")
```

Table 23: Table continues below

Eigenvalue	Condition Index	intercept	DIAMETER	SHUCKED_WEIGHT
4.749	1	0.0005567	0.0002298	0.001216
0.2033	4.833	0.03866	0.00242	0.01847
0.02714	13.23	0.002667	0.0001967	0.416
0.01674	16.84	0.003573	0.001339	0.4383
0.003702	35.82	0.9545	0.9958	0.126

VISCERA_WEIGHT	SHELL_WEIGHT
0.001057	0.001149
0.01669	0.01113
0.01008	0.582
0.9599	0.14
0.0123	0.2657