Regressione lineare per la previsione della percentuale di bodyfat

Progetto di MMIS

Eleonora Banterle, Noemi Bongiorni, Matteo Chiesa, Luca Tagliabue

Summary del dataset bodyfat

Age	Weight	Height	Hip	Thigh	Knee
Min. :22.00	Min. : 56.70	Min. : 74.93	Min. : 85.30	Min. :49.30	Min. :33.00
1st Qu.:35.00	1st Qu.: 72.12	1st Qu.:173.36	1st Qu.: 95.50	1st Qu.:56.08	1st Qu.:36.98
Median :43.00	Median : 79.89	Median :178.12	Median : 99.30	Median :59.00	Median :38.50
Mean :44.57	Mean : 81.18	Mean :178.24	Mean : 99.94	Mean :59.47	Mean :38.59
3rd Qu.:53.25	3rd Qu.: 89.36	3rd Qu.:183.52	3rd Qu.:103.53	3rd Qu.:62.55	3rd Qu.:39.90
Max. :81.00	Max. :164.72	Max. :197.49	Max. :147.70	Max. :87.30	Max. :49.10

Neck	Chest	Abdomen	Ankle	Biceps	Forearm
Min. :31.10	Min. : 83.40	Min. : 70.40	Min. :19.10	Min. :24.8	Min. :21.00
1st Qu.:36.40	1st Qu.: 94.35	1st Qu.: 84.58	1st Qu.:22.00	1st Qu.:30.2	1st Qu.:27.30
Median :37.95	Median : 99.60	Median : 90.90	Median :22.80	Median :32.1	Median :28.70
Mean :37.99	Mean :100.79	Mean : 92.47	Mean :23.12	Mean :32.3	Mean :28.68
3rd Qu.:39.42	3rd Qu.:105.30	3rd Qu.: 99.12	3rd Qu.:24.00	3rd Qu.:34.4	3rd Qu.:30.00
Max. :51.20	Max. :136.20	Max. :148.10	Max. :33.90	Max. :45.0	Max. :34.90

Wrist	class	Gender
Min. :15.80	Min. : 0.70	Length:248
1st Qu.:17.60	1st Qu.:12.47	Class :character
Median :18.30	Median :19.20	Mode :character
Mean :18.22	Mean :19.11	
3rd Qu.:18.80	3rd Qu.:25.20	
Max. :21.40	Max. :47.50	

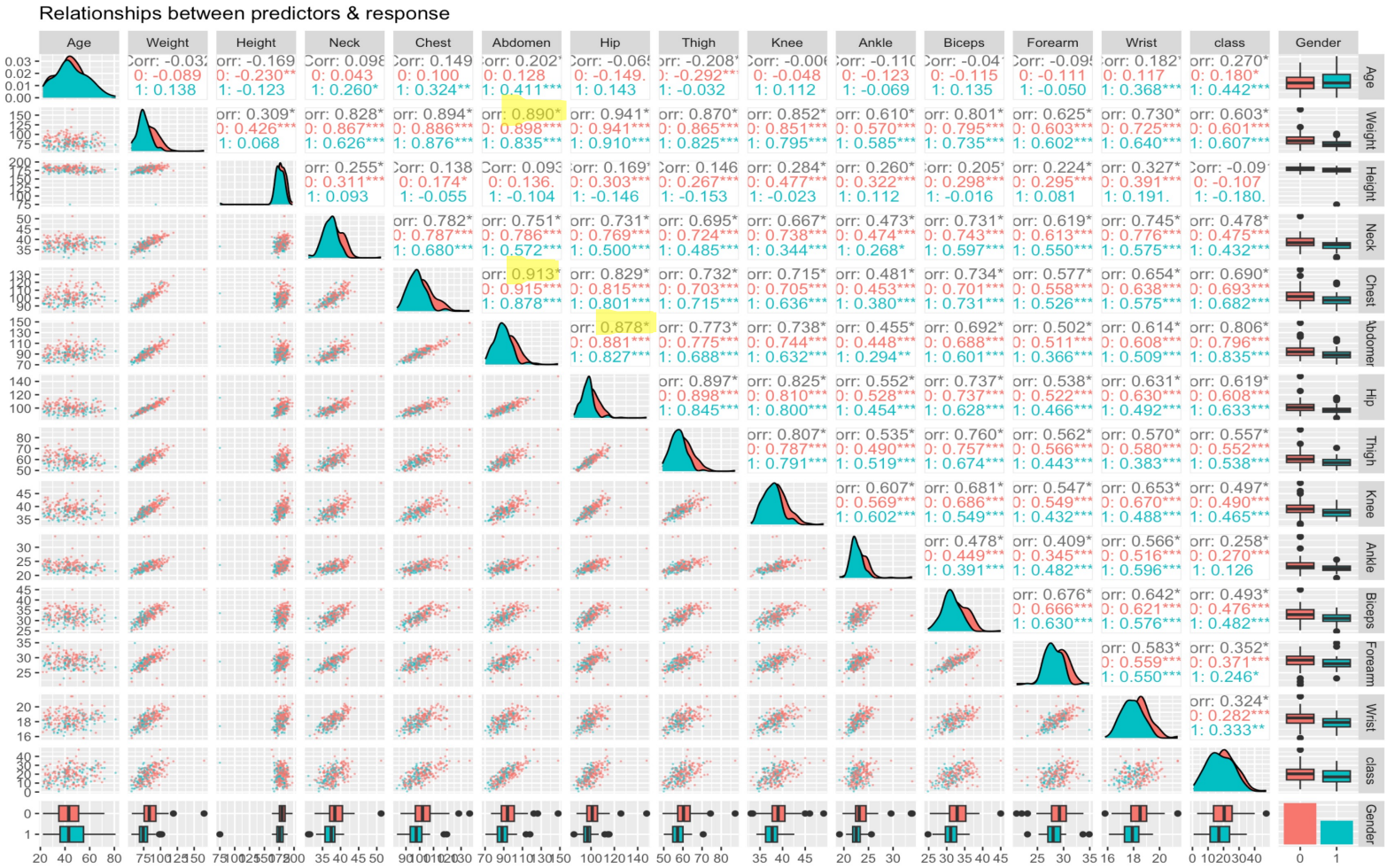
Modifiche preliminari

```
data <- data[-182, ] # rimuoviamo dato con 0 grasso
data <- data[,-1] # togliamo colonna Density
```

```
> # controllo se ci sono degli NA
> print(sapply(data,function(x) any(is.na(x))))
    Age  Weight  Height    Neck   Chest Abdomen     Hip   Thigh    Knee   Ankle  Biceps Forearm   Wrist   class
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
 Gender
  FALSE
```

```
train = sample(248,190) # circa 75% dei dati

# Dividiamo il dataframe in due per comodità
data_train_cat = data[train,]
data_test_cat = data[-train,]

# prende i nomi delle colonne con tipo numerico
data_num_columns <- unlist(lapply(data, is.numeric), use.names = FALSE)
# data_num è il dataset con solo covariate numeriche
data_num <- data[ , data_num_columns]
data_train <- data_train_cat[ , data_num_columns]
data_test <- data_test_cat[ , data_num_columns]
```

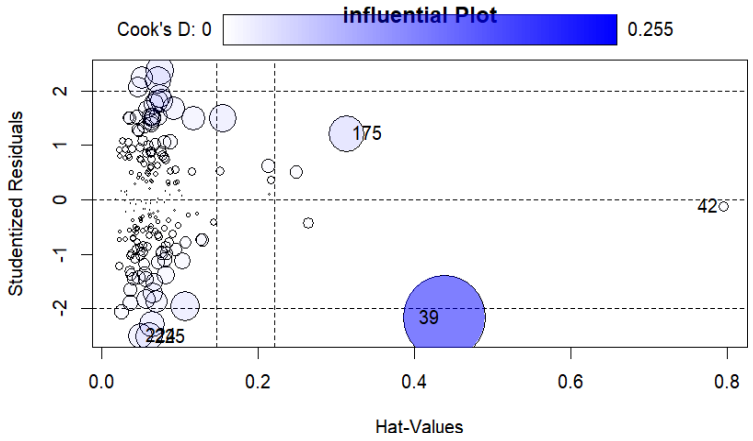
Visualizzazione dei dati

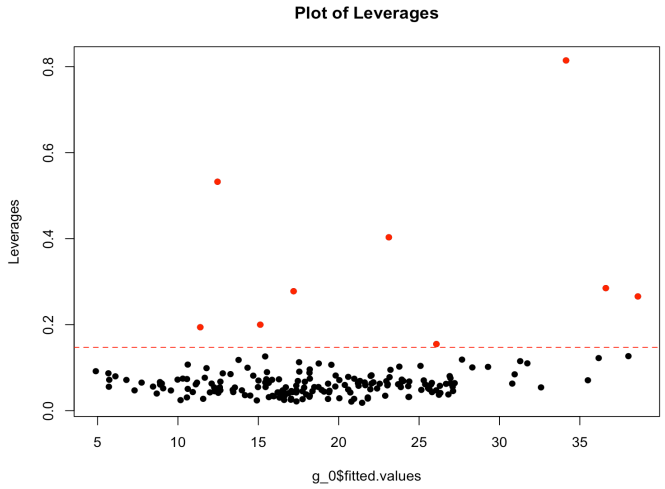


Eleonora Banterle, Noemi Bongiorni, Matteo Chiesa, Luca Tagliabue

Primo modello lineare

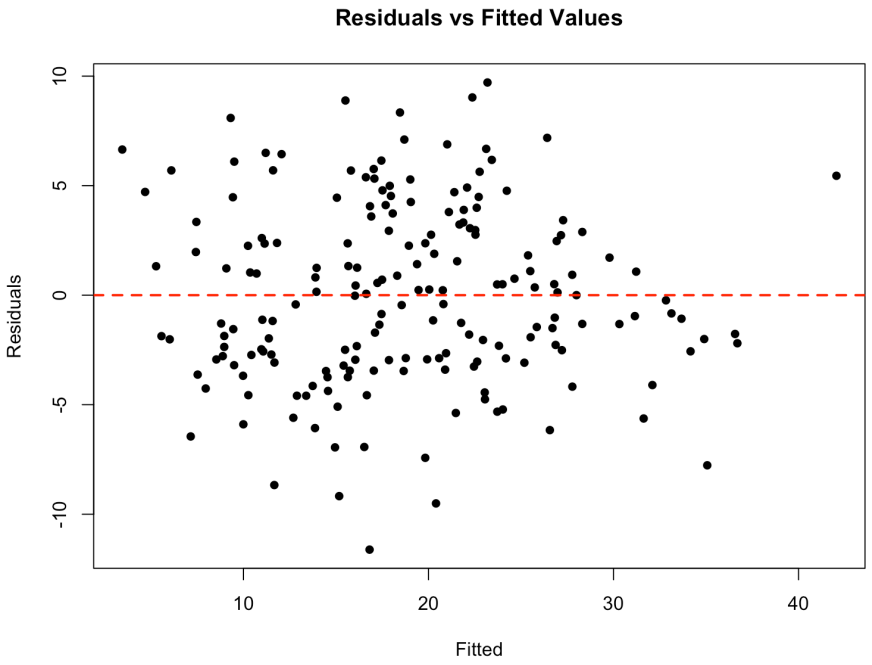
```
Call:
lm(formula = class ~ ., data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1167  -3.0494   0.0215   2.9533   9.4324

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.77981   19.04077  -0.829   0.4084
Age           0.07067    0.03548   1.992   0.0479 *
Weight       -0.14308    0.12797  -1.118   0.2650
Height       -0.02086    0.03884  -0.537   0.5919
Neck         -0.60539    0.27399  -2.210   0.0284 *
Chest         0.06711    0.10881   0.617   0.5382
Abdomen       0.84366    0.09834   8.579 4.86e-15 ***
Hip          -0.20722    0.15999  -1.295   0.1969
Thigh         0.28221    0.16157   1.747   0.0824 .
Knee         -0.07867    0.28238  -0.279   0.7809
Ankle         0.15327    0.27545   0.556   0.5786
Biceps        0.07636    0.18122   0.421   0.6740
Forearm       0.41395    0.20803   1.990   0.0482 *
Wrist        -1.41696    0.62642  -2.262   0.0249 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.221 on 176 degrees of freedom
Multiple R-squared:  0.7259,    Adjusted R-squared:  0.7056
F-statistic: 35.85 on 13 and 176 DF,  p-value: < 2.2e-16
```

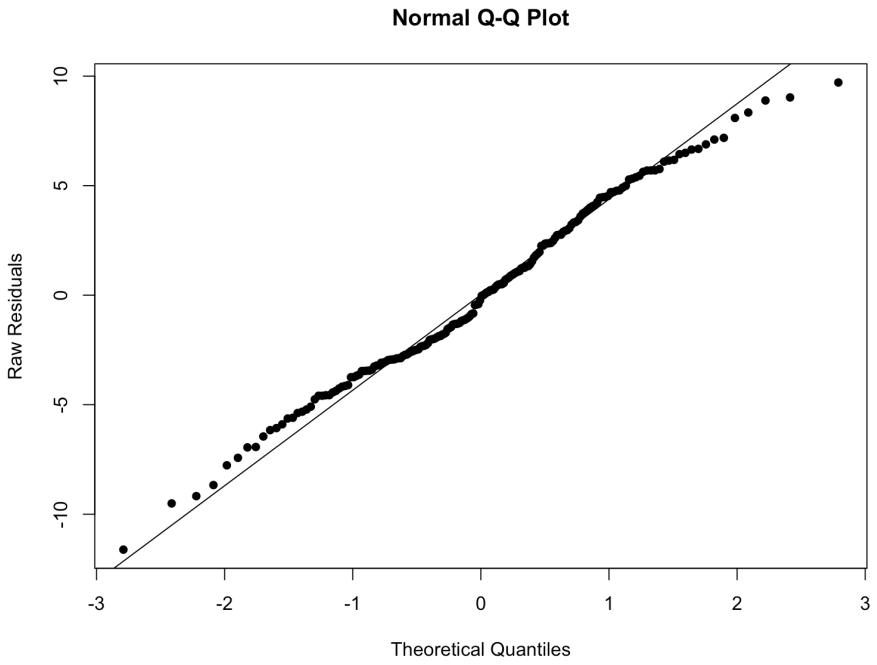




Eleonora Banterle, Noemi Bongiorni, Matteo Chiesa, Luca Tagliabue

Controllo ipotesi





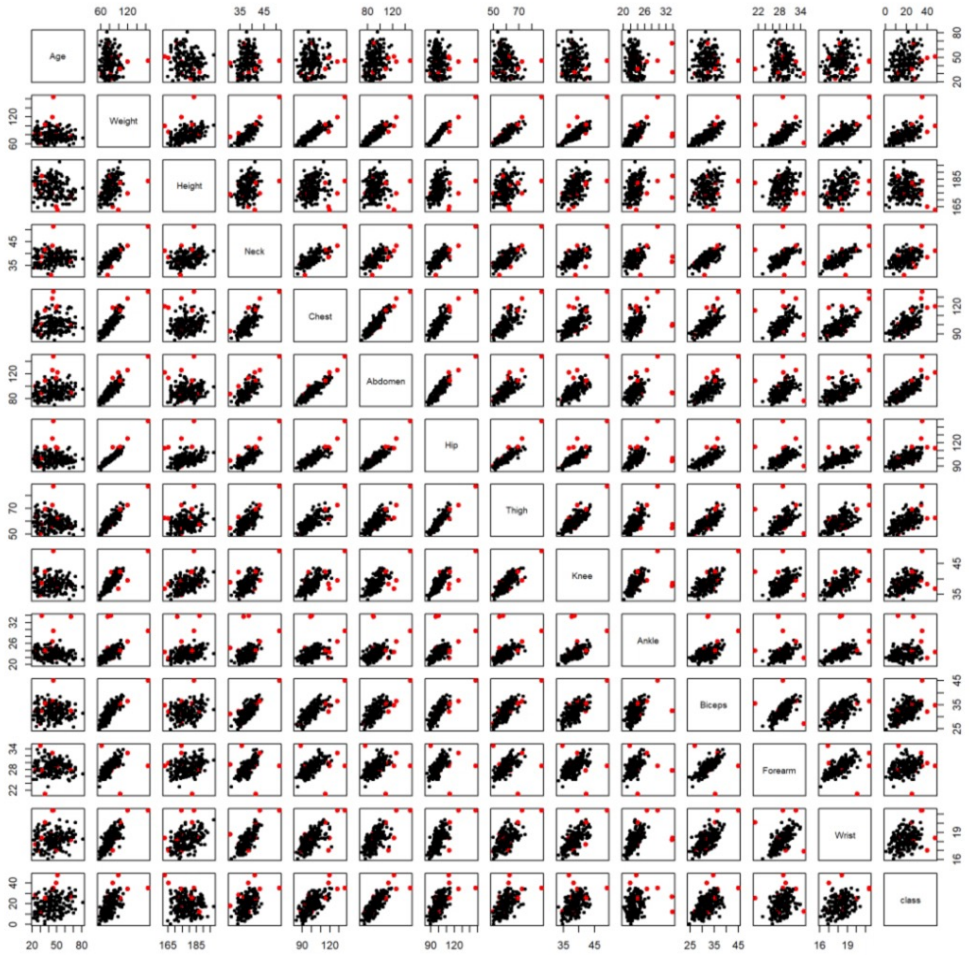

```
Shapiro-Wilk normality test

data:  g_0$res
W = 0.99063, p-value = 0.2526
```

Labor limae

```
Residual standard error: 4.16 on 167 degrees of freedom
Multiple R-squared:  0.718,     Adjusted R-squared:  0.696
F-statistic:  32.7 on 13 and 167 DF,  p-value: < 2.2e-16

> AIC( g_no_leverages )
[1] 1045.16
```

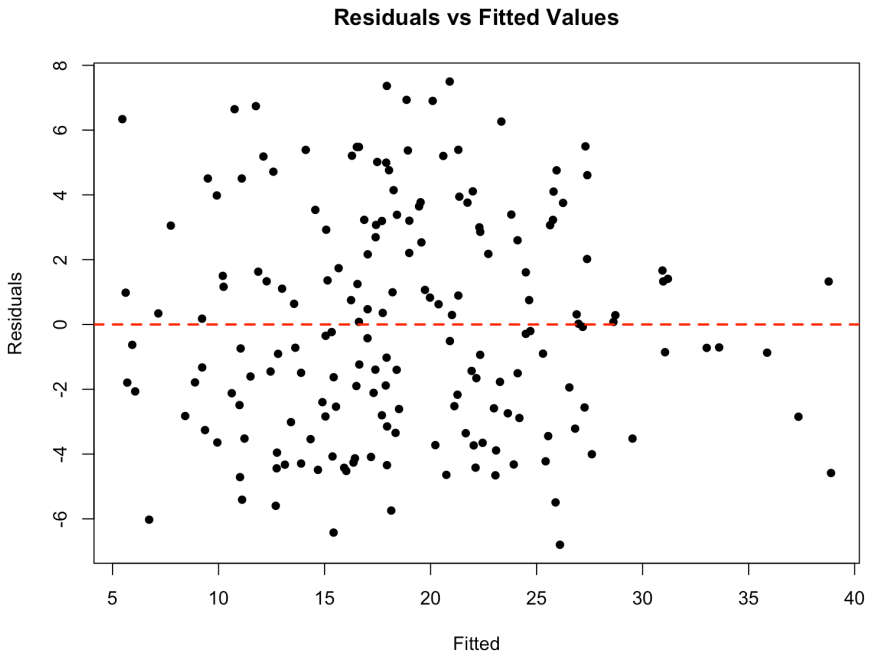
```
Residual standard error: 3.578 on 166 degrees of freedom
Multiple R-squared:  0.7928,    Adjusted R-squared:  0.7766
F-statistic: 48.85 on 13 and 166 DF,  p-value: < 2.2e-16

> AIC( g_no_stud )
[1] 985.1623
```

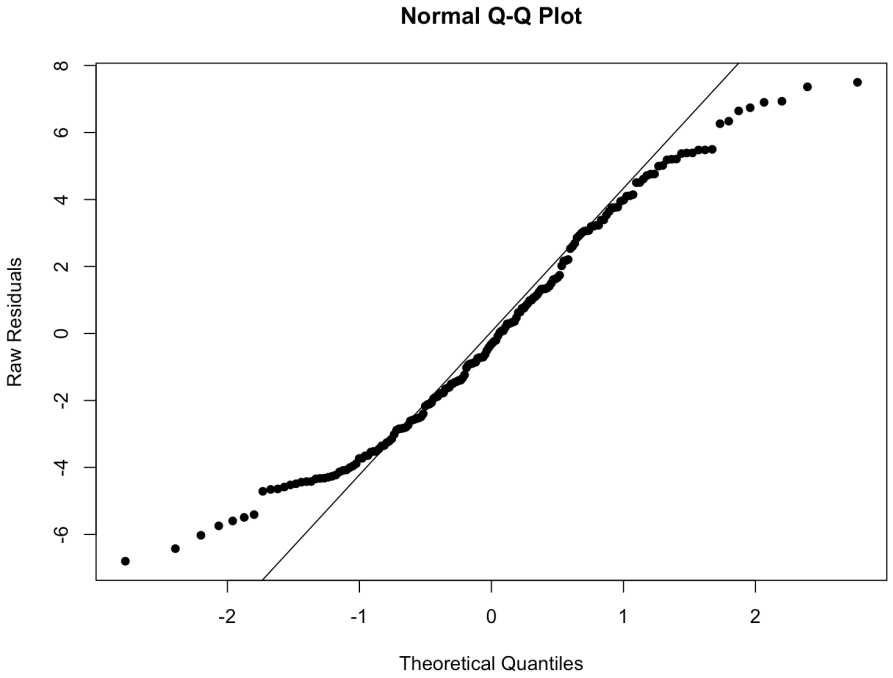
```
Residual standard error: 3.831 on 167 degrees of freedom
Multiple R-squared:  0.7663,    Adjusted R-squared:  0.7481
F-statistic: 42.11 on 13 and 167 DF,  p-value: < 2.2e-16

> AIC( g_cook )
[1] 1015.261
```



Rosso: leverages Verde: Studentizzati Arancione: Cook

Controllo ipotesi



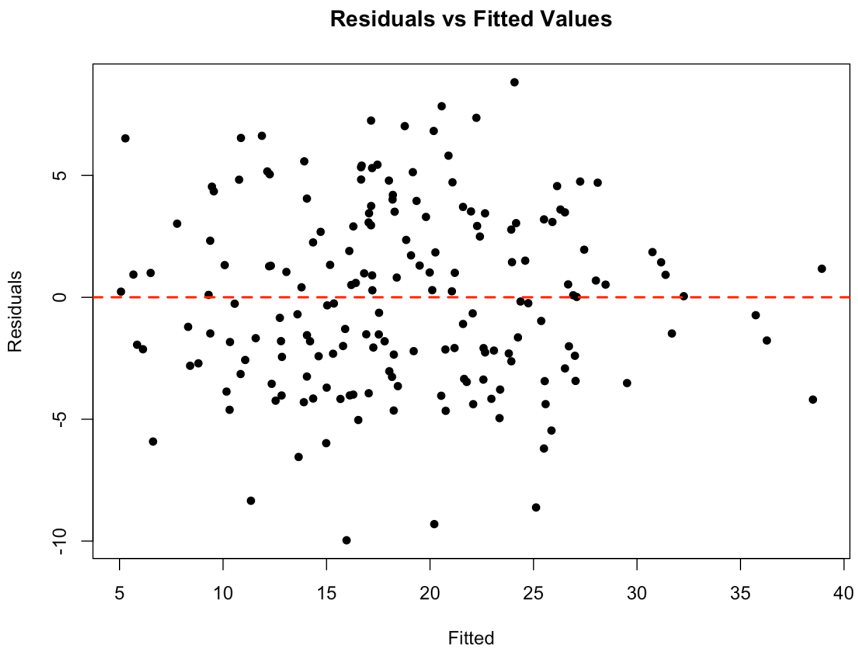


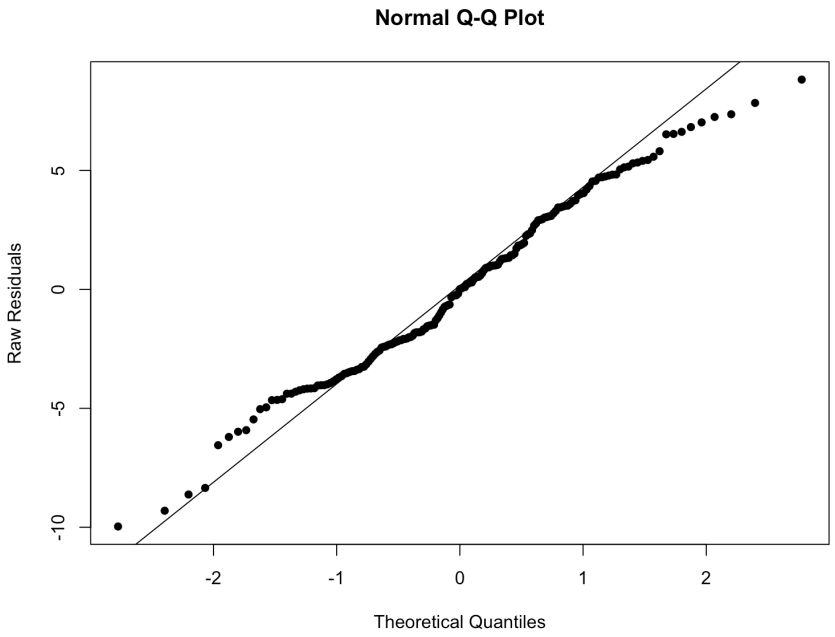

```
        Shapiro-Wilk normality test

data:  g_no_stud$res
W = 0.97096, p-value = 0.0008268
```

```
> # -------------------- BOX COX --------------------
> b = boxcox( class ~ ., data = data_train)
> #x lambda evaluated
> best_lambda_ind = which.max( b$y ); best_lambda = b$x[ best_lambda_ind ]; best_lambda
[1] 1.070707
```

Modello scelto

```
Residual standard error: 3.831 on 167 degrees of freedom
Multiple R-squared:  0.7663,    Adjusted R-squared:  0.7481
F-statistic: 42.11 on 13 and 167 DF,  p-value: < 2.2e-16
```






```
Shapiro-Wilk normality test

data:  g_cook$res
W = 0.98745, p-value = 0.1078
```

Rimozione covariate non significative

```
g_backward_removed = step( g_cook, direction = "backward" , trace = T)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.0118  -2.6260  -0.3197   3.1458  10.3445

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.87053    7.46627  -0.652  0.51504
Age          0.10862    0.03409   3.187  0.00171 **
Neck        -0.43023    0.24181  -1.779  0.07695 .
Abdomen      0.84813    0.07581  11.187  < 2e-16 ***
Hip         -0.48142    0.14388  -3.346  0.00100 **
Thigh        0.45223    0.15286   2.958  0.00352 **
Forearm      0.50646    0.23731   2.134  0.03423 *
Wrist       -1.99442    0.55355  -3.603  0.00041 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.057 on 174 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7456
F-statistic: 76.79 on 7 and 174 DF,  p-value: < 2.2e-16
```

Modello finale

```
g_7=lm(class ~ Age + Neck + Abdomen + Thigh + Wrist + Hip, data_train[ id_to_keep, ])
summary(g_7)
```

#togliamo la covariata *forearm*

```
g_8=lm(class ~ Age + Neck + Abdomen + Thigh + Wrist, data_train[ id_to_keep, ])
summary(g_8)
```

#togliamo la covariata *hip*

```
g_9=lm(class ~ Age + Neck + Abdomen + Wrist, data_train[ id_to_keep, ])
summary(g_9)
```

#togliamo la covariata *thigh*

```
g_10=lm(class ~ Age + Abdomen + Wrist, data_train[ id_to_keep, ])
summary(g_10)
```

#togliamo la covariata *neck*

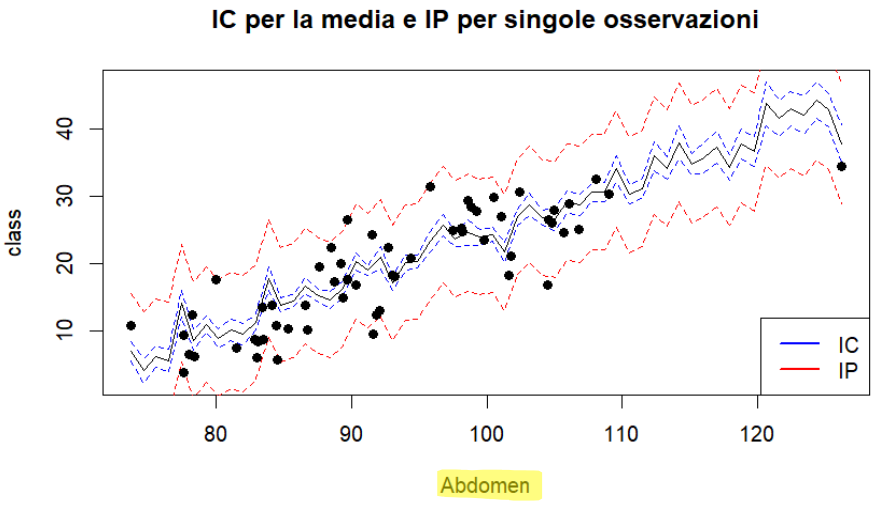
```
Residual standard error: 4.098 on 175 degrees of freedom
Multiple R-squared:  0.7491,    Adjusted R-squared:  0.7404
F-statistic: 87.06 on 6 and 175 DF,  p-value: < 2.2e-16
```

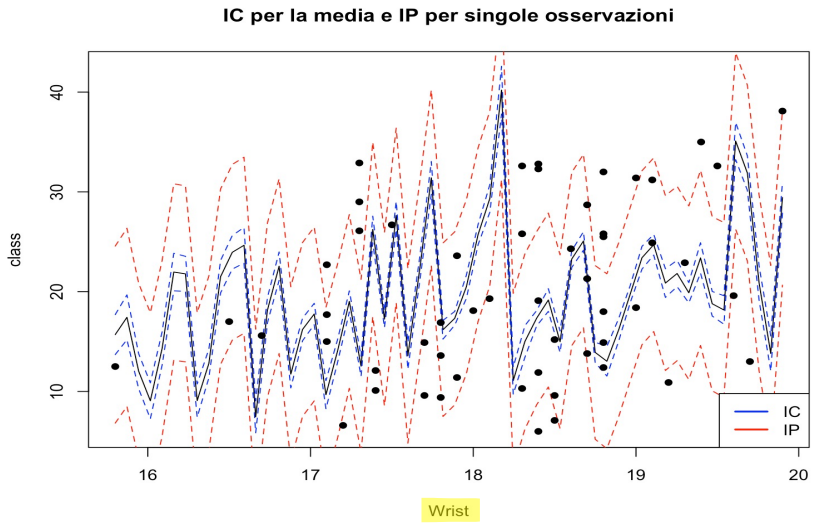
```
Residual standard error: 4.196 on 176 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.728
F-statistic: 97.88 on 5 and 176 DF,  p-value: < 2.2e-16
```

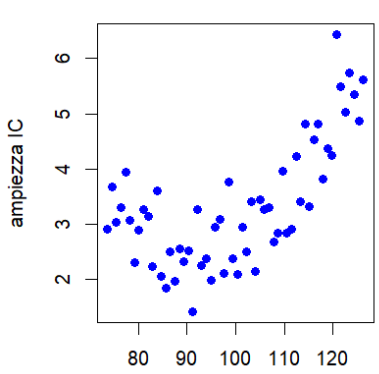
```
Residual standard error: 4.207 on 177 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.7265
F-statistic: 121.2 on 4 and 177 DF,  p-value: < 2.2e-16
```

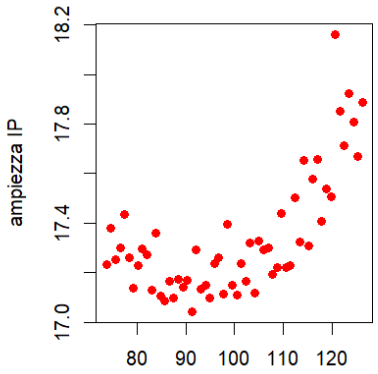
```
Residuals:
     Min       1Q   Median       3Q      Max
-10.3667  -2.9588  -0.2837   3.0893   9.5369

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.59076    6.56426  -1.918 0.056702 .
Age           0.08977    0.02542   3.531 0.000527 ***
Abdomen       0.76329    0.04058  18.810  < 2e-16 ***
Wrist        -2.35177    0.44680  -5.264 4.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.204 on 178 degrees of freedom
Multiple R-squared:  0.7315,    Adjusted R-squared:  0.7269
F-statistic: 161.6 on 3 and 178 DF,  p-value: < 2.2e-16
```

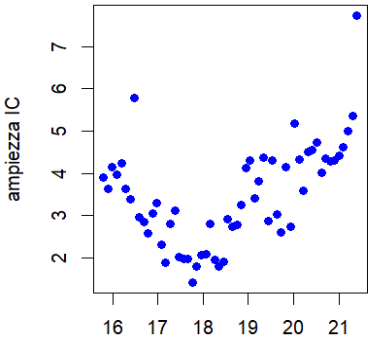
Intervalli di confidenza e di previsione

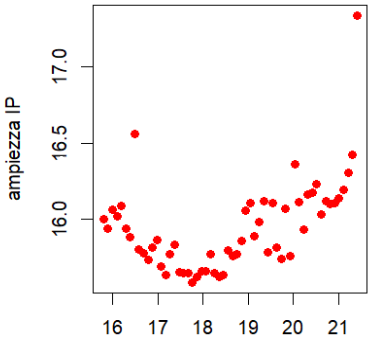












Cross-validazione

```
> #MSE sul test set
> mean( (data_test$class - predict(g_10,data_test) )^2 )
[1] 24.86077
> # Mse sul training set
> mean( (g_10$residuals )^2 )
[1] 17.28167
```

```
> cv.err = cv.glm(data_train[ id_to_keep, ],g_10, K = 10)
> cv.err$delta[1]
[1] 18.2517
```

Interazione tra le covariate

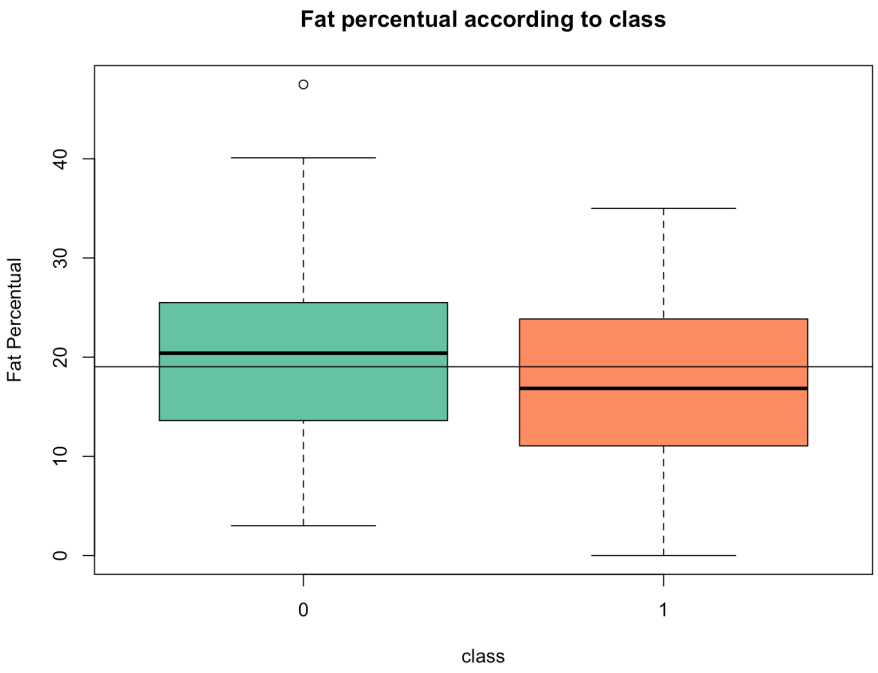
```
g_11=lm(class ~ Age + Abdomen + Wrist + Age:Gender + Abdomen:Gender + Wrist:Gender, data_train_cat[ id_to_keep, ])
summary(g_11)

g_12=lm(class ~ Age + Abdomen + Wrist + Age:Gender + Wrist:Gender, data_train_cat[ id_to_keep, ])
summary(g_12)

g_13=lm(class ~ Age + Abdomen + Wrist + Age:Gender, data_train_cat[ id_to_keep, ])
summary(g_13)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -13.53807    6.96189  -1.945   0.0534 .
Age              0.05999    0.03299   1.818   0.0707 .
Abdomen          0.79188    0.04948  16.004  < 2e-16 ***
Wrist           -2.38922    0.47878  -4.990 1.45e-06 ***
Age:Gender1      0.07434    0.05359   1.387   0.1672
Abdomen:Gender1 -0.08039    0.08878  -0.905   0.3665
Wrist:Gender1    0.25837    0.42526   0.608   0.5443
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.196 on 175 degrees of freedom
Multiple R-squared:  0.7369,    Adjusted R-squared:  0.7279
F-statistic: 81.69 on 6 and 175 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -13.90746    6.94637  -2.002   0.0468 *
Age              0.06193    0.03291   1.882   0.0615 .
Abdomen          0.76704    0.04116  18.638  < 2e-16 ***
Wrist           -2.24738    0.45219  -4.970 1.58e-06 ***
Age:Gender1      0.06089    0.05147   1.183   0.2384
Wrist:Gender1   -0.10706    0.13389  -0.800   0.4250
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.194 on 176 degrees of freedom
Multiple R-squared:  0.7357,    Adjusted R-squared:  0.7281
F-statistic: 97.96 on 5 and 176 DF,  p-value: < 2.2e-16
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.13282    6.76831  -2.236  0.02661 *
Age           0.07723    0.02674   2.888  0.00436 **
Abdomen       0.77112    0.04080  18.901  < 2e-16 ***
Wrist        -2.24132    0.45166  -4.962 1.63e-06 ***
Age:Gender1   0.02143    0.01460   1.468  0.14386
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.19 on 177 degrees of freedom
Multiple R-squared:  0.7347,    Adjusted R-squared:  0.7287
F-statistic: 122.5 on 4 and 177 DF,  p-value: < 2.2e-16
```

ANOVA




```
> P_shap = tapply( data$class, data$Gender, function( x ) ( shapiro.test( x )$p ) )
> P_shap
        0         1
0.1889364 0.4837790
> Var = tapply( data$class,data$Gender , var )
> Var
       0        1
69.59572 65.76328
```

```
> leveneTest(data$class, data$Gender)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.0032 0.9547
      247
> bartlett.test(data$class, data$Gender)

        Bartlett test of homogeneity of variances

data:  data$class and data$Gender
Bartlett's K-squared = 0.091321, df = 1, p-value = 0.7625
```

```
> mod_aov = aov( data$class ~ data$Gender )
> summary(mod_aov)
             Df Sum Sq Mean Sq F value Pr(>F)
data$Gender   1    422   421.6   6.183 0.0136 *
Residuals   247  16841    68.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```