Project_neonati_Jorge_Suarez

Jorge Suarez Linares

2023-12-27

Librerie utilizzate:

```
library(dplyr)
```

```
## Warning: il pacchetto 'dplyr' è stato creato con R versione 4.2.3
```

```
##
## Caricamento pacchetto: 'dplyr'
```

```
## I seguenti oggetti sono mascherati da 'package:stats':
##
##     filter, lag
```

```
## I seguenti oggetti sono mascherati da 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.2.3
```

```
library(moments)
library(knitr)
library(ineq)
library(rmarkdown)
library(tinytex)
library(e1071)
```

```
##
## Caricamento pacchetto: 'e1071'
```

```
## I seguenti oggetti sono mascherati da 'package:moments':
##
##     kurtosis, moment, skewness
```

```
library(car)
```

```
## Warning: il pacchetto 'car' è stato creato con R versione 4.2.3
```

```
## Caricamento del pacchetto richiesto: carData
```

```
##
## Caricamento pacchetto: 'car'
```

```
## Il seguente oggetto è mascherato da 'package:dplyr':
##
##     recode
```

```
library(lmtest)
```

```
## Warning: il pacchetto 'lmtest' è stato creato con R versione 4.2.3
```

```
## Caricamento del pacchetto richiesto: zoo
```

```
##
## Caricamento pacchetto: 'zoo'
```

```
## I seguenti oggetti sono mascherati da 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(stats)
library(ggcorrplot)
```

```
## Warning: il pacchetto 'ggcorrplot' è stato creato con R versione 4.2.3
```

```
library(MASS)
```

```
##
## Caricamento pacchetto: 'MASS'
```

```
## Il seguente oggetto è mascherato da 'package:dplyr':
##
##     select
```

```
library(gridExtra)
```

```
##
## Caricamento pacchetto: 'gridExtra'
```

```
## Il seguente oggetto è mascherato da 'package:dplyr':
##
##     combine
```

Dataset:

```
dati_neonati <- read.csv("neonati.csv", sep = ",", stringsAsFactors = T)
attach(dati_neonati)
```

In questo progetto andremo a fare un'analisi dei fattori che influenzano il peso dei neonati alla nascita. Nello specifico, oltre a uno studio tra le variabili raccolte, cercheremo di comprendere se le variaibli legate alla madre possono essere utilizzate per preverede il peso dei neonati.

Di seguito comincieremo con una breve analisi descrittiva delle variaibli quantitative continue coinvolte:

```
tabella_summary <- dati_neonati %>%
  summarise(
    mean_Anni_madre = mean(Anni.madre),
```

```
    sd_Anni_madre = sd(Anni.madre),
    median_Anni_madre = median(Anni.madre),
    min_Anni_madre = min(Anni.madre),
    max_Anni_madre = max(Anni.madre),
    range_Anni_madre = diff(range(Anni.madre)),

    mean_Ngravidanze = mean(N.gravidanze),
    sd_Ngravidanze = sd(N.gravidanze),
    median_Ngravidanze = median(N.gravidanze),
    min_Ngravidanze = min(N.gravidanze),
    max_Ngravidanze = max(N.gravidanze),
    range_Ngravidanze = diff(range(N.gravidanze)),

    mean_Gestazione = mean(Gestazione),
    sd_Gestazione = sd(Gestazione),
    median_Gestazione = median(Gestazione),
    min_Gestazione = min(Gestazione),
    max_Gestazione = max(Gestazione),
    range_Gestazione = diff(range(Gestazione)),

    mean_Peso = mean(Peso),
    sd_Peso = sd(Peso),
    median_Peso = median(Peso),
    min_Peso = min(Peso),
    max_Peso = max(Peso),
    range_Peso = diff(range(Peso)),

    mean_Lunghezza = mean(Lunghezza),
    sd_Lunghezza = sd(Lunghezza),
    median_Lunghezza = median(Lunghezza),
    min_Lunghezza = min(Lunghezza),
    max_Lunghezza = max(Lunghezza),
    range_Lunghezza = diff(range(Lunghezza)),

    mean_Cranio = mean(Cranio),
    sd_Cranio = sd(Cranio),
    median_Cranio = median(Cranio),
    min_Cranio = min(Cranio),
    max_Cranio = max(Cranio),
    range_Cranio = diff(range(Cranio)),
  )

print(tabella_summary)
```

```
##   mean_Anni_madre sd_Anni_madre median_Anni_madre min_Anni_madre max_Anni_madre
## 1          28.164      5.273578                28              0            46
##   range_Anni_madre mean_Ngravidanze sd_Ngravidanze median_Ngravidanze
## 1               46           0.9812       1.280587                  1
##   min_Ngravidanze max_Ngravidanze range_Ngravidanze mean_Gestazione
## 1               0              12                12         38.9804
##   sd_Gestazione median_Gestazione min_Gestazione max_Gestazione
```

```
## 1      1.868639                39             25             43
##   range_Gestazione mean_Peso  sd_Peso median_Peso min_Peso max_Peso range_Peso
## 1               18  3284.081 525.0387        3300      830     4930       4100
##   mean_Lunghezza sd_Lunghezza median_Lunghezza min_Lunghezza max_Lunghezza
## 1        494.692     26.31864              500           310           565
##   range_Lunghezza mean_Cranio sd_Cranio median_Cranio min_Cranio max_Cranio
## 1             255    340.0292  16.42533           340        235        390
##   range_Cranio
## 1          155
```

Di seguito proseguiremo con la costruzione delle tabelle assolute e relative delle variabili categoriche. Costruiremo anche i relativi grafici barre per avere uno sguardo d'insieme delle distribuzioni dei dati coinvolti.

```
#FUMATRICI

tabella_frequenza_fumatrici <- table(Fumatrici)
print(tabella_frequenza_fumatrici)
```

```
## Fumatrici
##    0    1
## 2396  104
```

```
tabella_frequenza_relativa_fumatrici <-
tabella_frequenza_fumatrici/nrow(dati_neonati)
print(tabella_frequenza_relativa_fumatrici)
```

```
## Fumatrici
##      0      1
## 0.9584 0.0416
```

```
plot_fumatrici <- ggplot()+
  geom_bar(aes(x= as.factor(Fumatrici), fill= as.factor(Fumatrici)), position=
"dodge")+

  labs(title = "Fumatrici",
       x = "S/N",
       y= "Madri",
       fill= "Fumatrici" )+
  geom_text(aes(x = as.factor(Fumatrici), y = 0, label = sprintf("%.2f",
tabella_frequenza_relativa_fumatrici[as.factor(Fumatrici)])),
            position = position_dodge(width = 0.9), vjust = -5, size=3)+
  theme_minimal()


#OSPEDALE

tabella_frequenza_ospedali <- table(Ospedale)
print(tabella_frequenza_ospedali)
```

```
## Ospedale
## osp1 osp2 osp3
##  816  849  835

tabella_frequenza_relativa_ospedali <-
tabella_frequenza_ospedali/nrow(dati_neonati)
print(tabella_frequenza_relativa_ospedali)

## Ospedale
##   osp1   osp2   osp3
## 0.3264 0.3396 0.3340

plot_ospedali <- ggplot()+
  geom_bar(aes(x= as.factor(Ospedale), fill= Ospedale), position= "dodge")+

  labs(title = "Ospedali coinvolti",
       x = "Neonati",
       y = "Ospedale")+
   scale_fill_manual(values = c("osp1" = "#f0e68c", "osp2" = "#CC0000",
"osp3"="#ff7f50"))+
  scale_x_discrete(labels = as.character(unique(dati_neonati$Ospedale))) +
   geom_text(stat = "count", aes(label = ..count..), position =
position_dodge(width = 0.9), vjust = -0.5)+
  geom_text(aes(x = as.factor(Ospedale), y = 0, label = sprintf("%.2f",
tabella_frequenza_relativa_ospedali[as.factor(Ospedale)])),
            position = position_dodge(width = 0.9), vjust = -5, size=3)+
  theme_minimal()



#TIPO DI PARTO

tabella_frequenza_parto <- table(Tipo.parto)
print(tabella_frequenza_parto)

## Tipo.parto
##  Ces  Nat
##  728 1772

tabella_frequenza_relativa_parto  <- tabella_frequenza_parto /nrow(dati_neonati)
print(tabella_frequenza_relativa_parto)

## Tipo.parto
##    Ces    Nat
## 0.2912 0.7088

plot_parto <- ggplot()+
  geom_bar(aes(x= as.factor(Tipo.parto), fill= Tipo.parto), position= "dodge")+

  labs(title = "Tipologia di parto",
       x = "Cesareo/Naturale",
       y = "Quantità madri")+
```

```
  geom_text(aes(x = as.factor(Tipo.parto), y = 0, label = sprintf("%.2f",
tabella_frequenza_relativa_parto[as.factor(Tipo.parto)])),
            position = position_dodge(width = 0.9), vjust = -5, size=3)+
  theme_minimal()


# SESSO

tabella_frequenza_Sesso <- table(Sesso)
print(tabella_frequenza_Sesso)
```

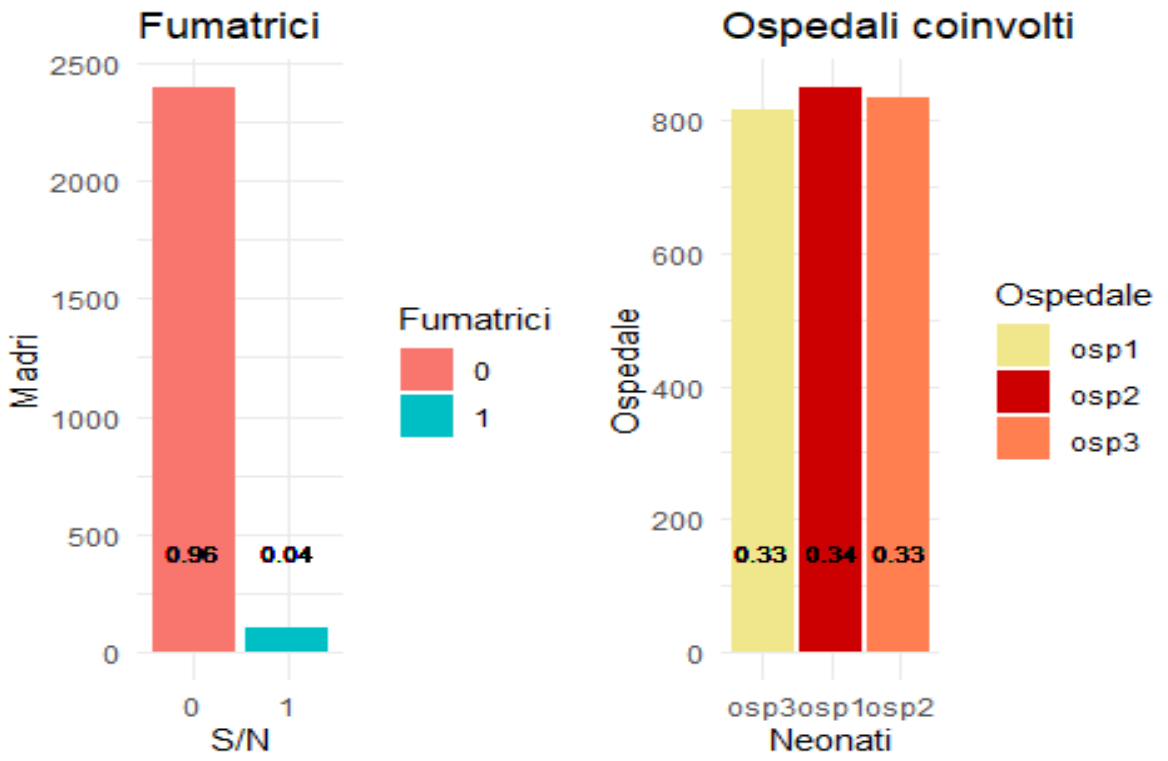
```
## Sesso
##    F    M
## 1256 1244
```

```
tabella_frequenza_relativa_Sesso  <- tabella_frequenza_Sesso /nrow(dati_neonati)
print(tabella_frequenza_relativa_Sesso)
```

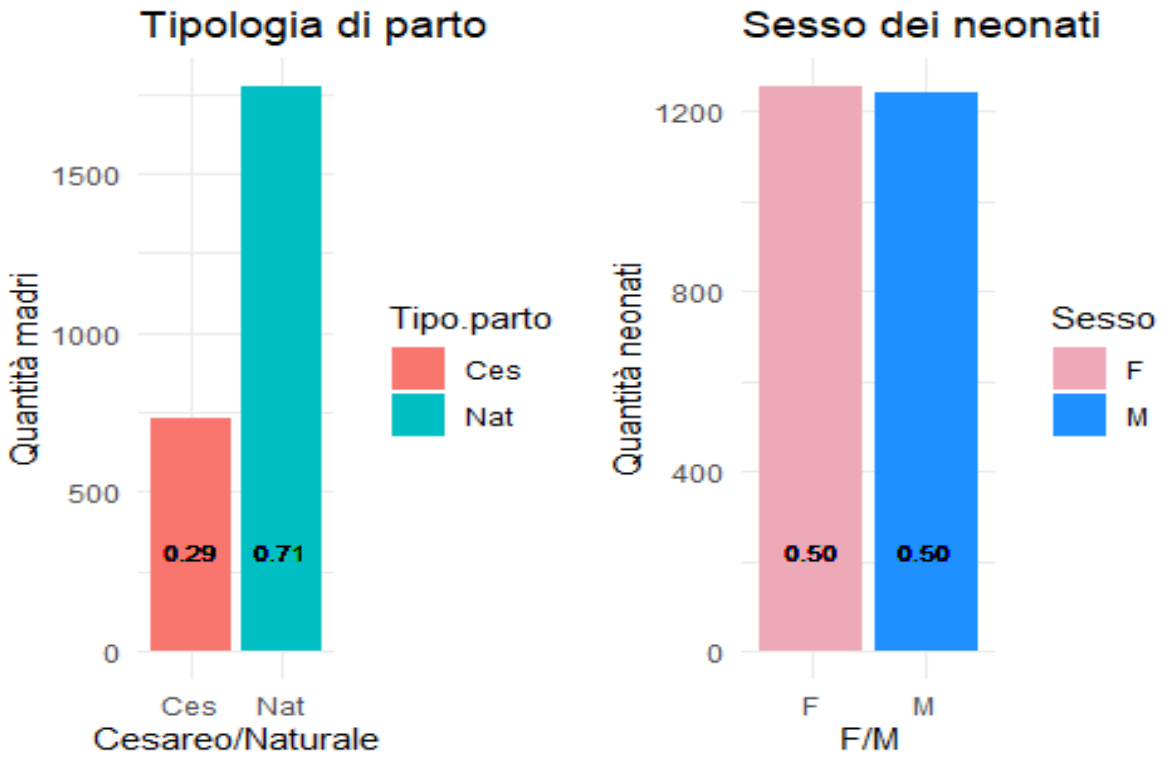
```
## Sesso
##      F      M
## 0.5024 0.4976
```

```
plot_sesso <- ggplot()+
  geom_bar(aes(x= as.factor(Sesso), fill= Sesso), position= "dodge")+

  labs(title = "Sesso dei neonati",
       x = "F/M",
       y = "Quantità neonati")+
  scale_fill_manual(values = c("F" = "pink2", "M" = "#1E90FF"))+
  geom_text(aes(x = as.factor(Sesso), y = 0, label = sprintf("%.2f",
tabella_frequenza_relativa_Sesso[as.factor(Sesso)])),
            position = position_dodge(width = 0.9), vjust = -5, size=3)+
  theme_minimal()


grid.arrange(plot_fumatrici, plot_ospedali, ncol = 2)
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



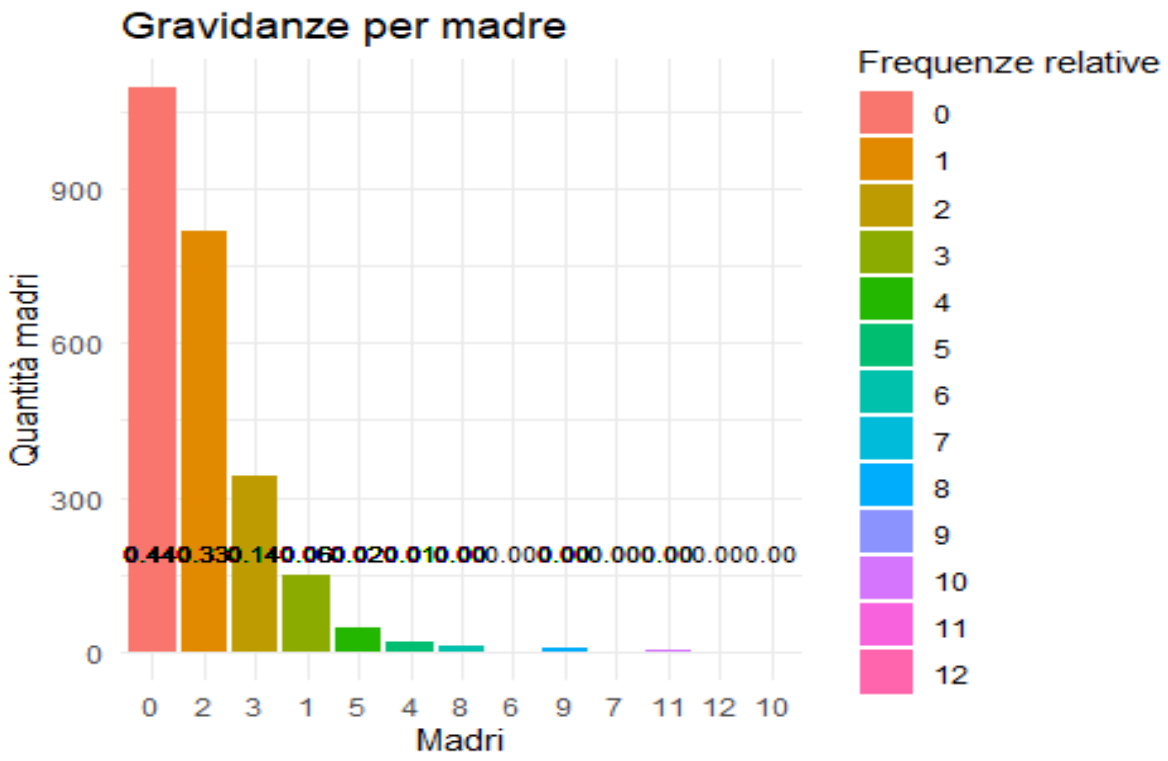

```
grid.arrange( plot_parto, plot_sesso, ncol = 2)
```



Fumatrici: Il 95,84% delle madri in esame non fuma Ospedali: I neonati sono distribuiti equamente, ~ 33% in ognuno dei tre ospedali Tipologia di parto: Il 70,88% sono parti naturali Sesso: Distribuiti equamente. Il 50,24% sono femmine mentre il 49,76% sono maschi.

Variabile "GRAVIDANZE":

```
#GRAVIDANZE
tabella_frequenza_gravidanze <- table(N.gravidanze)
print(tabella_frequenza_gravidanze)

## N.gravidanze
##    0    1    2    3    4    5    6    7    8    9   10   11   12
## 1096  818  340  150   48   21   11    1    8    2    3    1    1

tabella_frequenza_relativa_gravidanze <-
tabella_frequenza_gravidanze/nrow(dati_neonati)
print(tabella_frequenza_relativa_gravidanze)

## N.gravidanze
##      0      1      2      3      4      5      6      7      8      9     10
## 0.4384 0.3272 0.1360 0.0600 0.0192 0.0084 0.0044 0.0004 0.0032 0.0008 0.0012
##     11     12
## 0.0004 0.0004

ggplot()+
  geom_bar(aes(x= as.factor(N.gravidanze), fill= as.factor(N.gravidanze)),
position= "dodge")+

  labs(title = "Gravidanze per madre",
       x = "Madri",
       y = "Quantità madri",
       fill= "Frequenze relative")+
  scale_x_discrete(labels = as.character(unique(dati_neonati$N.gravidanze))) +
  geom_text(aes(x = as.factor(N.gravidanze), y = 0, label = sprintf("%.2f",
tabella_frequenza_relativa_gravidanze[as.factor(N.gravidanze)])),
            position = position_dodge(width = 0.9), vjust = -5, size=3)+
  theme_minimal()
```



ANALISI DESCRITTIVA DEI DATI DELL'ECOGRAFIA

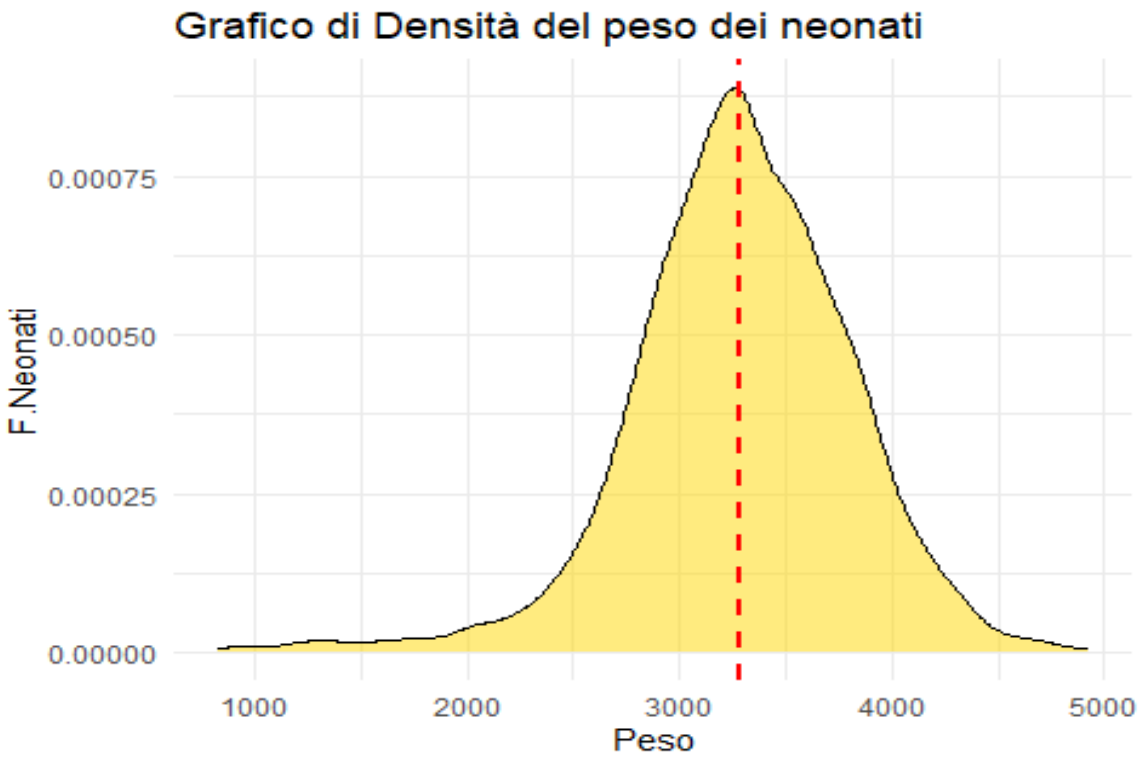
PESO

```
dati_neonati %>%
  summarise(
    peso_skew = skewness(Peso),
    peso_kurt = kurtosis(Peso)-3)
```

```
##    peso_skew  peso_kurt
## 1 -0.6466427 -0.9724926
```

```
ggplot()+
  geom_density(aes(x= Peso), fill= "#FFD700", alpha= 0.5)+
  geom_vline(xintercept = mean(Peso), color = "red", linetype= "dashed", size= 1)+
  labs(title = "Grafico di Densità del peso dei neonati",
       x = "Peso",
       y = "F.Neonati")+
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



- La skewness è negativa, indicando che la coda della distribuzione è più lunga a sinistra della media.
- La curtosi è negativa, ossia ha una distribuzione platicurtica, stando a indicare code più leggere rispetto alla gaussiana e una quantita minori di valori nei pressi della media.

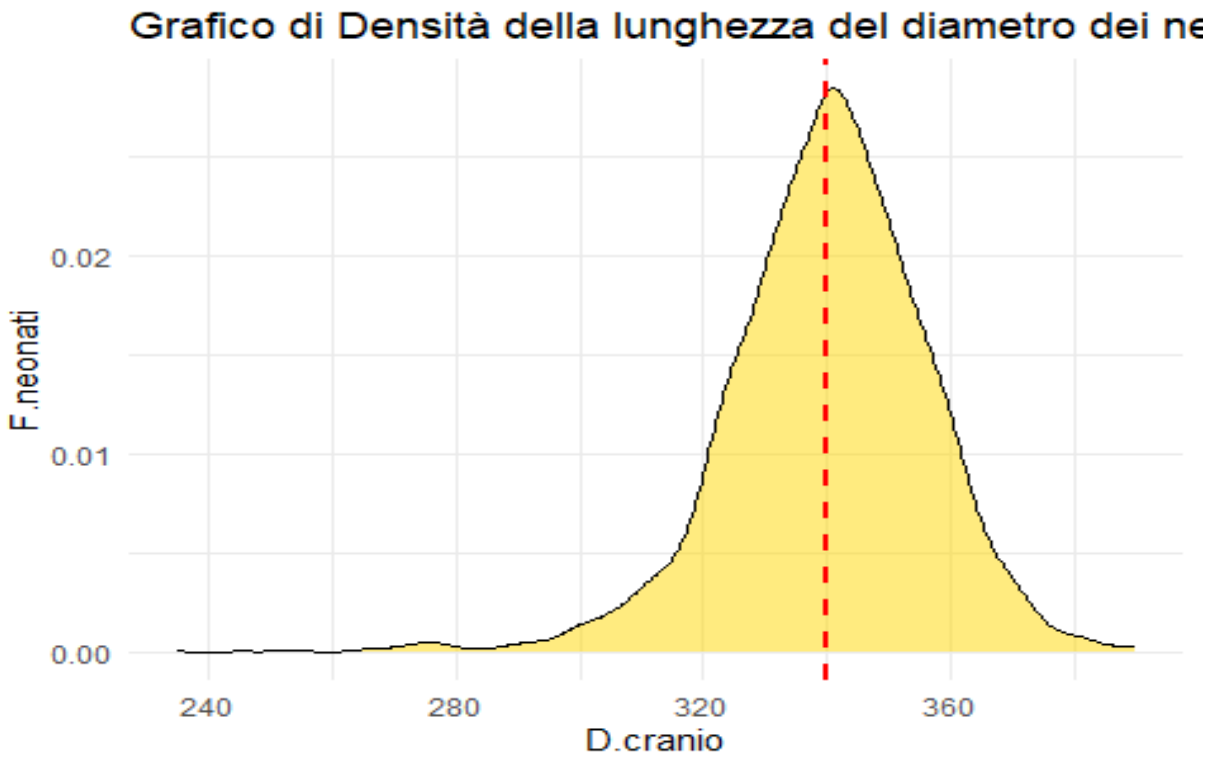
LUNGHEZZA

```
dati_neonati %>%
  summarise(
    lunghezza_skew = skewness(Lunghezza),
    lunghezza_kurt = kurtosis(Lunghezza)-3)

##   lunghezza_skew lunghezza_kurt
## 1       -1.51379       3.479586

ggplot()+
  geom_density(aes(x= Lunghezza), fill= "#FFD700", alpha= 0.5)+
  geom_vline(xintercept = mean(Lunghezza), color = "red", linetype= "dashed",
size= 1)+
  labs(title = "Grafico di Densità della lunghezza dei neonati",
       x = "Lunghezza",
       y = "F.neonati")+
  theme_minimal()
```

- La skewness è negativa, indicando che la coda della distribuzione è più lunga a sinistra della media.
- La curtosi è positiva, ossia ha una distribuzione leptocurtica, stando a indicare code più appuntite rispetto alla gaussiana e una quantita maggiore di valori nei pressi della media.

CRANIO

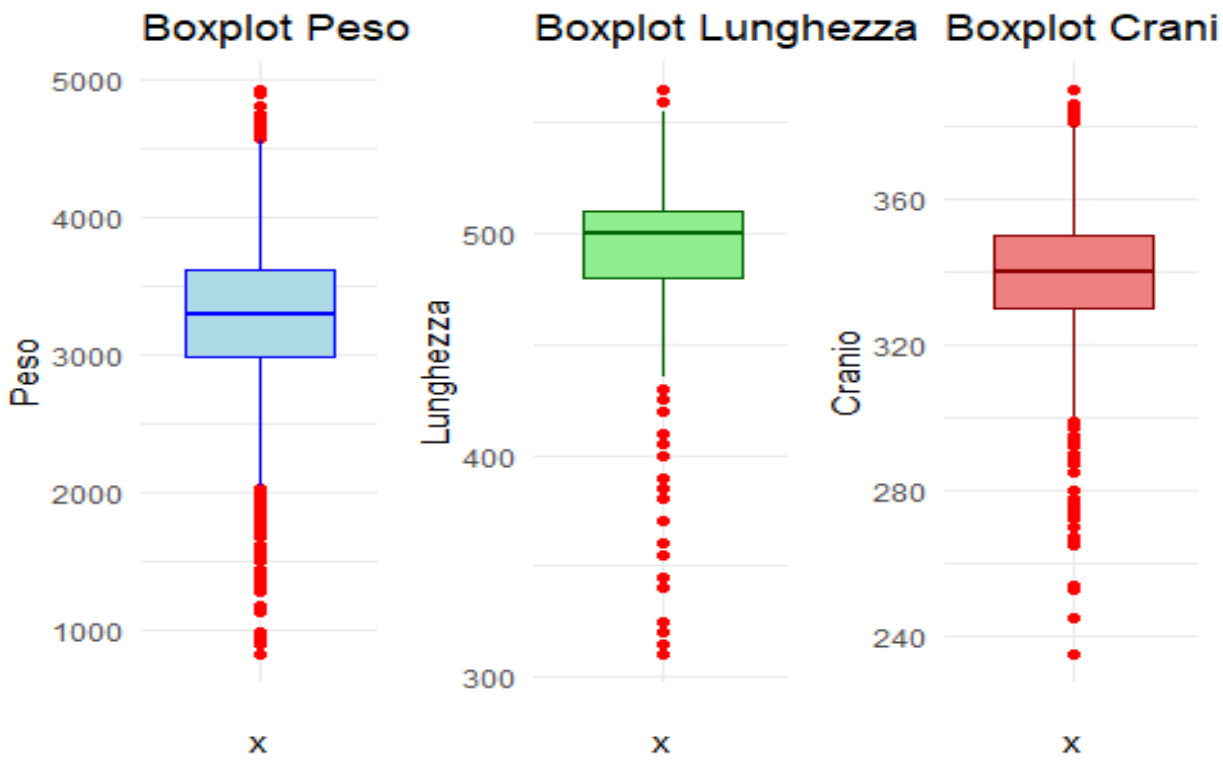
```
dati_neonati %>%
  summarise(
    cranio_skew = skewness(Cranio),
    cranio_kurt = kurtosis(Cranio)-3)

##   cranio_skew cranio_kurt
## 1  -0.7845817 -0.05854976

ggplot()+
  geom_density(aes(x= Cranio), fill= "#FFD700", alpha= 0.5)+
  geom_vline(xintercept = mean(Cranio), color = "red", linetype= "dashed", size=
1)+
  labs(title = "Grafico di Densità della lunghezza del diametro dei neonati in
mm",
       x = "D.cranio",
       y = "F.neonati")+
  theme_minimal()
```



- La skewness è negativa, indicando che la coda della distribuzione è più lunga a sinistra della media.
- La curtosi è mesocurtica, stando a indicare code simili alla gaussiana.

Adesso costruiremo alcuni boxplot per vedere se riusciamo a identificare outliers precocemente.

```
box_peso <- ggplot(data = dati_neonati, aes(x = "", y = Peso)) +
  geom_boxplot(fill = "lightblue", color = "blue", outlier.colour = "red",
outlier.shape = 16) +
  labs(title = "Boxplot Peso", y = "Peso")+
  theme_minimal()

box_lunghezza <- ggplot(data= dati_neonati, aes(x = "", y = Lunghezza)) +
  geom_boxplot(fill = "lightgreen", color = "darkgreen", outlier.colour = "red",
outlier.shape = 16) +
  labs(title = "Boxplot Lunghezza", y = "Lunghezza")+
  theme_minimal()


box_cranio <- ggplot(data=dati_neonati, aes(x = "", y = Cranio)) +
  geom_boxplot(fill = "lightcoral", color = "darkred", outlier.colour = "red",
outlier.shape = 16) +
  labs(title = "Boxplot Cranio", y = "Cranio")+
  theme_minimal()

grid.arrange( box_peso, box_lunghezza, box_cranio, ncol = 3)
```



In questi dati dell'ecografia possiamo notare una certa quantità di outliers, che potranno in un secondo moomento creare dei problemi di leverage in fase di modellizzazione.

Ora condurremo alcune ipotesi sulla media per le variabili chiave. H0 –> la media del peso nei due sessi è uguale
H1–> la media del peso nei due sessi non è uguale

```
t_test_peso_sesso <- t.test(Peso ~ Sesso, data = dati_neonati)
t_test_peso_sesso

##
##  Welch Two Sample t-test
##
## data:  Peso by Sesso
## t = -12.106, df = 2490.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is
not equal to 0
## 95 percent confidence interval:
##  -287.1051 -207.0615
## sample estimates:
## mean in group F mean in group M
##        3161.132        3408.215
```

Ci sono prove sufficienti per rifiutare l'ipotesi nulla di uguaglianza delle medie.

H0 –> la media della lunghezza nei due sessi è uguale
H1–> la media della lunghezza nei due sessi non è uguale

```r
t_test_lunghezza_sesso <- t.test(Lunghezza ~ Sesso, data = dati_neonati)
t_test_lunghezza_sesso
```

```
##
##  Welch Two Sample t-test
##
## data:  Lunghezza by Sesso
## t = -9.582, df = 2459.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is
not equal to 0
## 95 percent confidence interval:
##  -11.929470  -7.876273
## sample estimates:
## mean in group F mean in group M
##        489.7643        499.6672
```

Ci sono prove sufficienti per rifiutare l'ipotesi nulla di uguaglianza delle medie.

H0 –> la media della lunghezza nei due sessi è uguale
H1–> la media della lunghezza nei due sessi non è uguale

```r
t_test_cranio_sesso <- t.test(Cranio ~ Sesso, data = dati_neonati)
t_test_cranio_sesso
```

```
##
##  Welch Two Sample t-test
##
## data:  Cranio by Sesso
## t = -7.4102, df = 2491.4, p-value = 1.718e-13
## alternative hypothesis: true difference in means between group F and group M is
not equal to 0
## 95 percent confidence interval:
##  -6.089912 -3.541270
## sample estimates:
## mean in group F mean in group M
##        337.6330        342.4486
```

Ci sono prove sufficienti per rifiutare l'ipotesi nulla di uguaglianza delle medie.

H0 –> la media del tipo di parto e l'ospedale è uguale
H1–> la media del tipo di parto e l'osedale non è uguale

```r
tabella_chi_parto_ospedale <- table(dati_neonati$Tipo.parto,
dati_neonati$Ospedale)

chi_parto_ospedale <- chisq.test(tabella_chi_parto_ospedale)
chi_parto_ospedale
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabella_chi_parto_ospedale
## X-squared = 1.0972, df = 2, p-value = 0.5778
```

Non ci sono prove sufficienti per rifiutare l'ipotesi nulla di uguaglianza della tipologia di parto nei tre diversi ospedali

H0 –> la media del numero di settimane di gestazione è uguale tra fumatrici e non fumatrici
H1–> la media del numero di settimane di gestazione non è uguale tra fumatrici e non fumatrici

```
t.test(Gestazione~Fumatrici, data = dati_neonati)
```

```
##
##  Welch Two Sample t-test
##
## data:  Gestazione by Fumatrici
## t = -2.0824, df = 119.26, p-value = 0.03944
## alternative hypothesis: true difference in means between group 0 and group 1 is
not equal to 0
## 95 percent confidence interval:
##  -0.58791720 -0.01481813
## sample estimates:
## mean in group 0 mean in group 1
##        38.96786        39.26923
```

Ci sono prove sufficienti per rifiutare l'ipotesi nulla di uguaglianza delle medie. La qual cosa a mio parere è molto interessante, tenendo in conto per esempio che il numero di settime di gestazione influisce anche sulle variabili dell'ecografia.

Di seguito andremo a costriure un primo modello di regressione lineare multipla per prevedere il peso del neonato. Come già visto in precedenza la distribuzione della variabile dipendente ha una skewness negativa e una curtosi leptocurtica, ma senza presentare valori estremi. Ciononostante proveremo il testo di Shapiro per esplorarne la normalità:

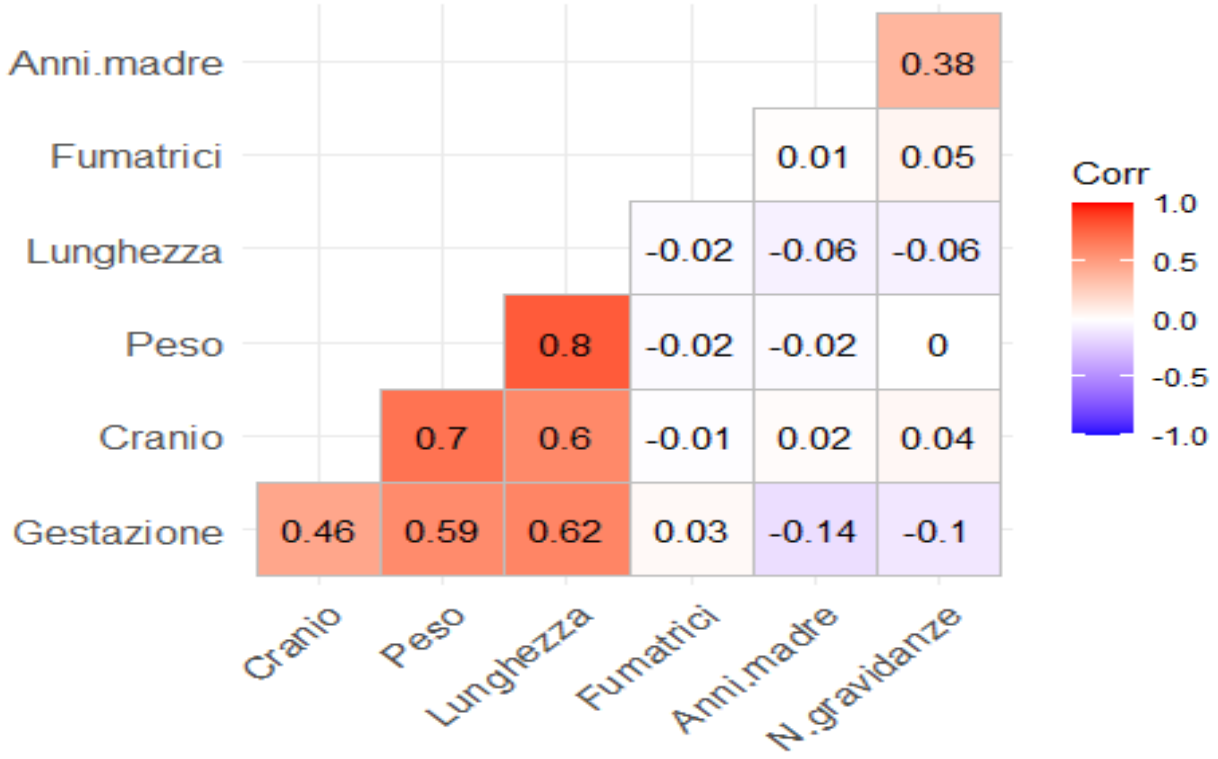
```
shapiro.test(Peso)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Peso
## W = 0.97066, p-value < 2.2e-16
```

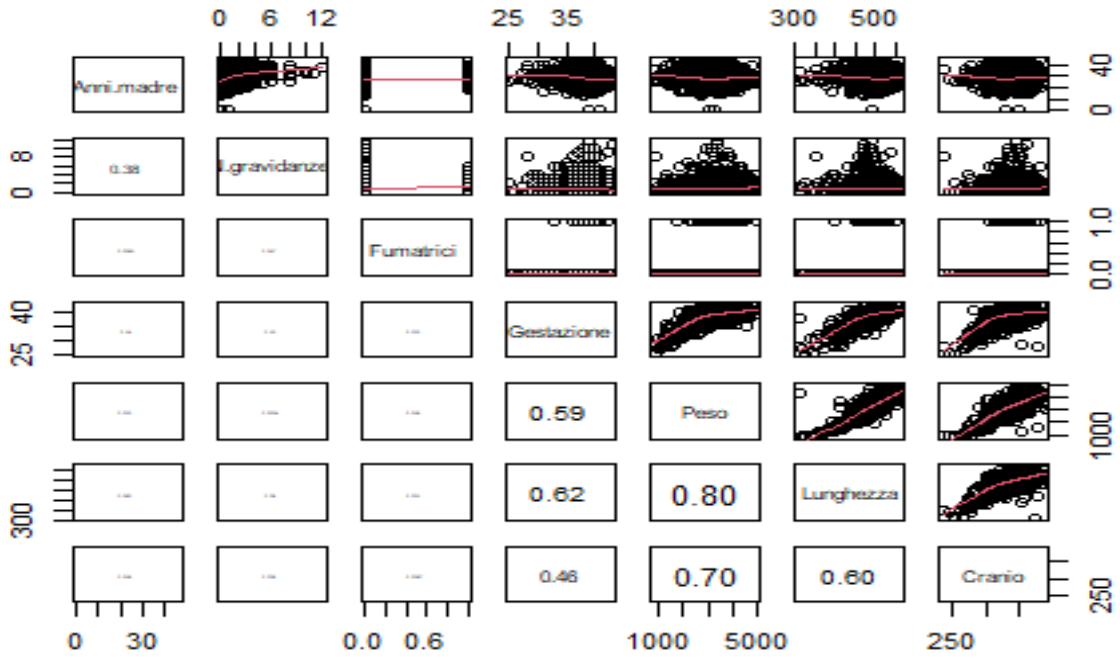
Viene respinta l'ipotesi nulla di normalità della distribuzione, ma essendo il campione abbastanza ampio, la statistica W vicina a 1 e la forma del grafico vicino alla normale, possiamo, anche per ragioni pratiche, definirla una distribuzione normale.

Ora faremo uno studio delle correlazioni tenendo a mente che una correlazione troppo elevata dei regressori con la variaible dipendente, ~.80, può provocare dei problemi di multicollearità.

```
dati_numerici_neonati <- dati_neonati[, sapply(dati_neonati, is.numeric)]

corr_matrix_mod2 = round(cor(dati_numerici_neonati), 2)


ggcorrplot(corr_matrix_mod2, hc.order = TRUE, type = "lower",
           lab = TRUE)
```




```r
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(dati_numerici_neonati, upper.panel = panel.smooth, lower.panel = panel.cor)
```




```
round(cor(dati_numerici_neonati),2)
```

```
##              Anni.madre N.gravidanze Fumatrici Gestazione  Peso Lunghezza
## Anni.madre         1.00         0.38      0.01      -0.14 -0.02     -0.06
## N.gravidanze       0.38         1.00      0.05      -0.10  0.00     -0.06
## Fumatrici          0.01         0.05      1.00       0.03 -0.02     -0.02
## Gestazione        -0.14        -0.10      0.03       1.00  0.59      0.62
## Peso              -0.02         0.00     -0.02       0.59  1.00      0.80
## Lunghezza         -0.06        -0.06     -0.02       0.62  0.80      1.00
## Cranio             0.02         0.04     -0.01       0.46  0.70      0.60
##              Cranio
## Anni.madre     0.02
## N.gravidanze   0.04
## Fumatrici     -0.01
## Gestazione     0.46
## Peso           0.70
## Lunghezza      0.60
## Cranio         1.00
```

- Correlazione: possiamo notare una correlazione preccoupante della variabile risposta con il regressore “Lunghezza”. Si dovrebbe scartare dal futuro modello, ma trovandosi nel limite soglia, preferisco mantenerlo poichè si tratta di dati dell’ecografia e funge da varaibile di controllo. Per il resto, non ci sono altri problemi di correlazione, tranne che nel caso di “Cranio”, ma si tratta di un caso simile a “Lunghezza”.
- Relazione lineare: a colpo d’occhio possiamo notare delle mancate relazioni lineare tra la variaible risposta e i regressori “Anni. Madre”, “N. gravidanze”, “Fumatrici”, e una relazione da approfondire con “Gestazione”.

Proseguiamo adesso con la creazione di un primo modello. Con “Peso” come variaible risposta e tutte le altre variaibli del dataset come regressori.

```
mod_peso1 <- lm(Peso ~.,data = dati_neonati )
summary(mod_peso1)
```

```
##
## Call:
## lm(formula = Peso ~ ., data = dati_neonati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1124.40  -181.66   -14.42   160.91  2611.89
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6738.4762   141.3087 -47.686  < 2e-16 ***
## Anni.madre        0.8921     1.1323   0.788   0.4308
## N.gravidanze     11.2665     4.6608   2.417   0.0157 *
## Fumatrici       -30.1631    27.5386  -1.095   0.2735
## Gestazione       32.5696     3.8187   8.529  < 2e-16 ***
## Lunghezza        10.2945     0.3007  34.236  < 2e-16 ***
## Cranio           10.4707     0.4260  24.578  < 2e-16 ***
## Tipo.partoNat    29.5254    12.0844   2.443   0.0146 *
## Ospedaleosp2    -11.2095    13.4379  -0.834   0.4043
## Ospedaleosp3     28.0958    13.4957   2.082   0.0375 *
## SessoM           77.5409    11.1776   6.937 5.08e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.9 on 2489 degrees of freedom
## Multiple R-squared:  0.7289, Adjusted R-squared:  0.7278
## F-statistic: 669.2 on 10 and 2489 DF,  p-value: < 2.2e-16
```

Regressori con p-value significativo: N.gravidanze, Gestazione, Lunghezza, Cranio, Parto Naturale , 3° ospedale, Sesso M L’R quadro aggiustato indica che la variaiblità del modello viene spegata per un 72,78%

Procedura stepwise: Tolgo come primo regressore “Anni.Madre”, perchè presenta un p-value non significativo.

```
mod_peso2 <- update(mod_peso1, ~.-Anni.madre)
summary(mod_peso2)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Ospedale + Sesso, data = dati_neonati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1113.93  -180.11   -16.36   160.58  2616.96
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6708.1065   135.9394 -49.346  < 2e-16 ***
## N.gravidanze     12.6085     4.3381   2.906  0.00369 **
## Fumatrici       -30.3092    27.5359  -1.101  0.27113
## Gestazione       32.2501     3.7968   8.494  < 2e-16 ***
## Lunghezza        10.2944     0.3007  34.239  < 2e-16 ***
## Cranio           10.4876     0.4255  24.651  < 2e-16 ***
## Tipo.partoNat    29.5351    12.0834   2.444  0.01458 *
## Ospedaleosp2    -11.0816    13.4359  -0.825  0.40957
## Ospedaleosp3     28.3660    13.4903   2.103  0.03559 *
## SessoM           77.6205    11.1763   6.945 4.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.9 on 2490 degrees of freedom
## Multiple R-squared:  0.7288, Adjusted R-squared:  0.7278
## F-statistic: 743.6 on 9 and 2490 DF,  p-value: < 2.2e-16
```

```
anova(mod_peso1,mod_peso2)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza
+
##     Cranio + Tipo.parto + Ospedale + Sesso
## Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   2489 186762521
## 2   2490 186809099 -1    -46578 0.6207 0.4308
```

Si nota a livello di significativà un miglioramento nel regressore N.gravidazne. L'R quadro aggiustato è rimasto invariato. Il test di Anova registra un aumento non significativo di varianza spiegata.

Tolgo il regresso "Fumatrici" dal primo modello perchè presenta un p-value non significativo.

```
mod_peso3 <- update(mod_peso2, ~.-Fumatrici)
summary(mod_peso3)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso, data = dati_neonati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1113.18  -181.16   -16.58   161.01  2620.19
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6707.4293   135.9438 -49.340  < 2e-16 ***
## N.gravidanze     12.3619     4.3325   2.853  0.00436 **
## Gestazione       31.9909     3.7896   8.442  < 2e-16 ***
## Lunghezza        10.3086     0.3004  34.316  < 2e-16 ***
## Cranio           10.4922     0.4254  24.661  < 2e-16 ***
## Tipo.partoNat    29.2803    12.0817   2.424  0.01544 *
## Ospedaleosp2    -11.0227    13.4363  -0.820  0.41209
## Ospedaleosp3     28.6408    13.4886   2.123  0.03382 *
## SessoM           77.4412    11.1756   6.930 5.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.9 on 2491 degrees of freedom
## Multiple R-squared:  0.7287, Adjusted R-squared:  0.7278
## F-statistic: 836.3 on 8 and 2491 DF,  p-value: < 2.2e-16
```

```
anova(mod_peso2,mod_peso3)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
## Model 2: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Ospedale + Sesso
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   2490 186809099
## 2   2491 186899996 -1    -90897 1.2116 0.2711
```

Non si nota alcun miglioramento nei regressori né nell'R quadro aggiustato che rimane al 72,78%. Inoltre il test di Anova registra un aumento non significativo di varianza spiegata.

Tolgo la variabile ospedale da questa prima regressione

```
mod_peso3 <- update(mod_peso3, ~.- Ospedale)
summary(mod_peso3)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Sesso, data = dati_neonati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1129.31  -181.70   -16.31   161.07  2638.85
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6707.2971   135.9911 -49.322  < 2e-16 ***
## N.gravidanze     12.7558     4.3366   2.941   0.0033 **
## Gestazione       32.2713     3.7941   8.506  < 2e-16 ***
## Lunghezza        10.2864     0.3007  34.207  < 2e-16 ***
```

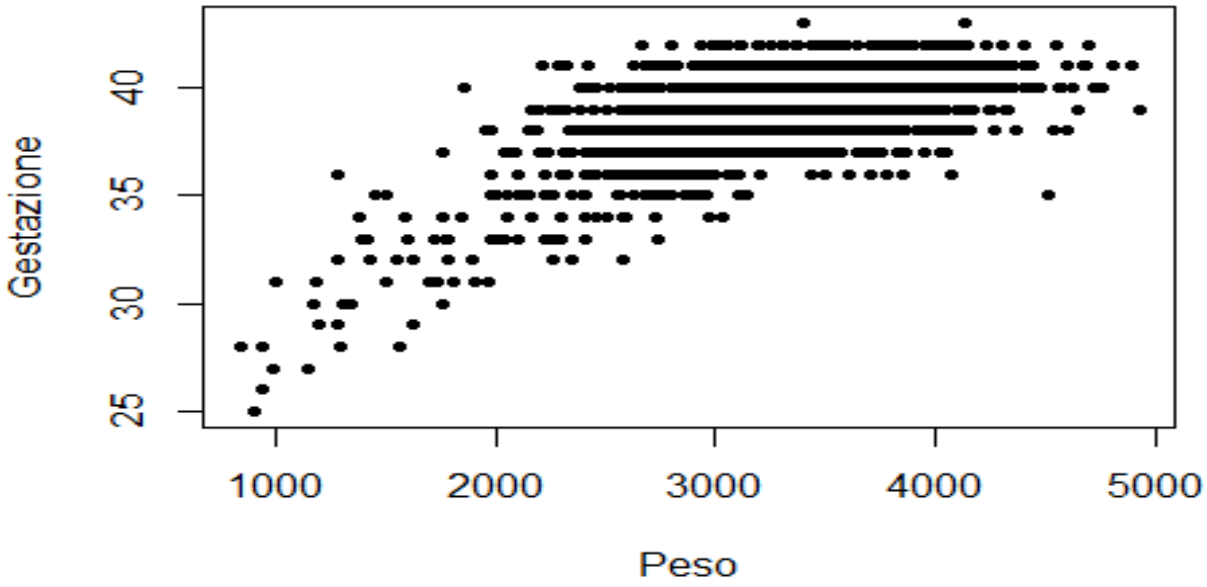
```
## Cranio           10.5057     0.4260  24.659  < 2e-16 ***
## Tipo.partoNat    30.0342    12.0969   2.483   0.0131 *
## SessoM           77.9285    11.1905   6.964 4.22e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.3 on 2493 degrees of freedom
## Multiple R-squared:  0.7277, Adjusted R-squared:  0.727
## F-statistic:  1110 on 6 and 2493 DF,  p-value: < 2.2e-16
```

Ora studieremo più approfonditamente la relazione lineare tra la variaibile risposta e Gestazione.

```
mod_peso4 <- update(mod_peso3, ~.+I(Gestazione^2))
summary(mod_peso4)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Sesso + I(Gestazione^2), data = dati_neonati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1127.83  -180.56   -16.62   162.80  2661.28
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4713.3849   897.6462  -5.251 1.64e-07 ***
## N.gravidanze       12.8408     4.3333   2.963  0.00307 **
## Gestazione        -79.0197    49.6700  -1.591  0.11176
## Lunghezza          10.3888     0.3039  34.185  < 2e-16 ***
## Cranio             10.6005     0.4278  24.780  < 2e-16 ***
## Tipo.partoNat      29.5657    12.0889   2.446  0.01453 *
## SessoM             75.7674    11.2228   6.751 1.82e-11 ***
## I(Gestazione^2)     1.4857     0.6612   2.247  0.02472 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.1 on 2492 degrees of freedom
## Multiple R-squared:  0.7282, Adjusted R-squared:  0.7275
## F-statistic: 953.9 on 7 and 2492 DF,  p-value: < 2.2e-16
```

```
plot(Peso,Gestazione, pch= 20)
```




```
anova(mod_peso3,mod_peso4)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Sesso
## Model 2: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Sesso + I(Gestazione^2)
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1   2493 187601677
## 2   2492 187222293  1    379384 5.0497 0.02472 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il nuovo regressore ha un p-value sotto la soglia di significatività, indicando che la relazione con la variaible risposta potrebbe avere un andamento diverso con l'aggiunta di ulteriori dati. l'R quadro aggiustato non presenta miglioramenti ma il test di Anova registra un aumento significativo di varianza spiegata. Per questo motivo, ritengo di continuare con questo modello.

AIC E BIC

```
AIC(mod_peso1,mod_peso2,mod_peso3,mod_peso4)
```

```
##           df      AIC
## mod_peso1 12 35171.95
## mod_peso2 11 35170.57
## mod_peso3  8 35175.16
## mod_peso4  9 35172.10
```

```
BIC(mod_peso1,mod_peso2,mod_peso3,mod_peso4)
```

```
##           df      BIC
## mod_peso1 12 35241.84
## mod_peso2 11 35234.64
## mod_peso3  8 35221.75
## mod_peso4  9 35224.51
```

AIC: mod_peso2 BIC: mod_peso2

Multicollinearità

```
vif(mod_peso2)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## N.gravidanze 1.027985  1        1.013896
## Fumatrici    1.007346  1        1.003666
## Gestazione   1.676688  1        1.294870
## Lunghezza    2.085755  1        1.444214
## Cranio       1.626661  1        1.275406
## Tipo.parto   1.004240  1        1.002118
## Ospedale     1.003421  2        1.000854
## Sesso        1.040558  1        1.020077
```

Nessun regressore presenta problemi di multicollinearità.

Faccio un'ulteriore verifica con la funzione stepwise per vedere se mi ritorna lo stesso modello:

```
mod_peso_stepwise <-MASS:: stepAIC(mod_peso1,
               direction = "both",
               k=2)
```

```
## Start:  AIC=28075.26
## Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Anni.madre    1     46578 186809099 28074
## - Fumatrici     1     90019 186852540 28075
## <none>                      186762521 28075
## - N.gravidanze  1    438452 187200974 28079
## - Tipo.parto    1    447929 187210450 28079
## - Ospedale      2    685979 187448501 28080
## - Sesso         1   3611021 190373542 28121
## - Gestazione    1   5458403 192220925 28145
## - Cranio        1  45326172 232088693 28617
## - Lunghezza     1  87951062 274713583 29038
##
## Step:  AIC=28073.88
## Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
```

```
## - Fumatrici     1     90897 186899996 28073
## <none>                      186809099 28074
## + Anni.madre    1     46578 186762521 28075
## - Tipo.parto    1    448222 187257321 28078
## - Ospedale      2    692738 187501837 28079
## - N.gravidanze  1    633756 187442855 28080
## - Sesso         1   3618736 190427835 28120
## - Gestazione    1   5412879 192221978 28143
## - Cranio        1  45588236 232397335 28618
## - Lunghezza     1  87950050 274759149 29036
##
## Step:  AIC=28073.1
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## <none>                      186899996 28073
## + Fumatrici     1     90897 186809099 28074
## + Anni.madre    1     47456 186852540 28075
## - Tipo.parto    1    440684 187340680 28077
## - Ospedale      2    701680 187601677 28079
## - N.gravidanze  1    610840 187510837 28079
## - Sesso         1   3602797 190502794 28119
## - Gestazione    1   5346781 192246777 28142
## - Cranio        1  45632149 232532146 28617
## - Lunghezza     1  88355030 275255027 29039
```

```
summary(mod_peso_stepwise)
```

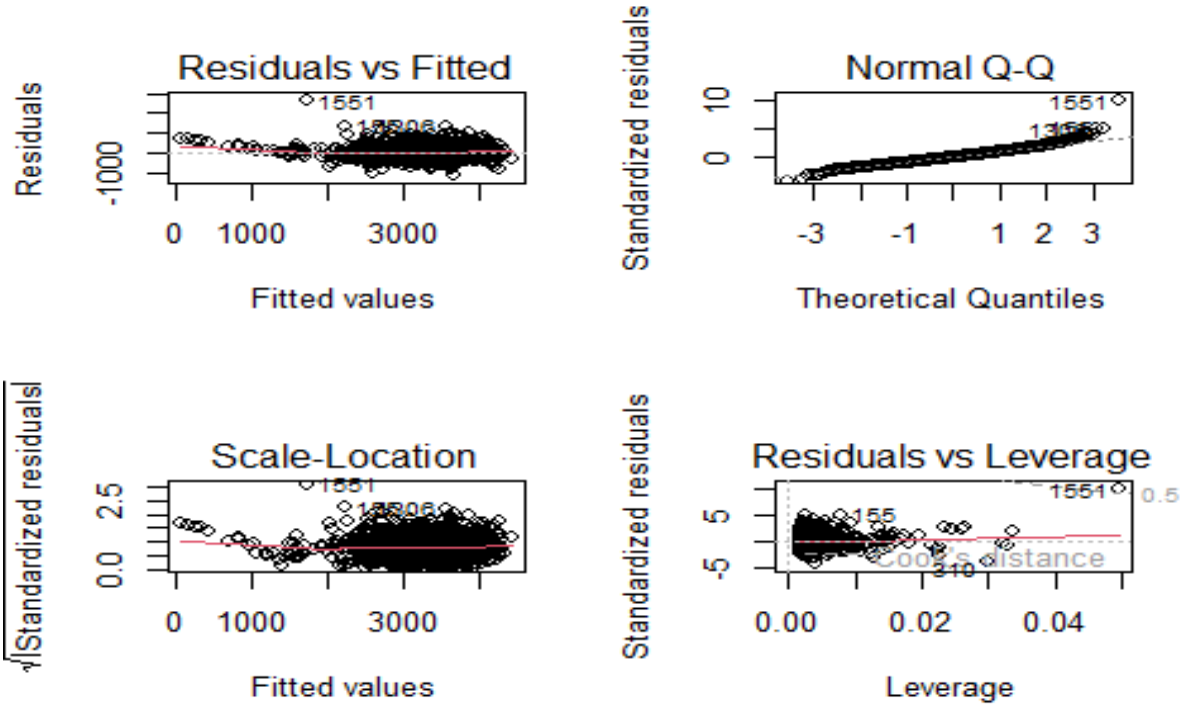
```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso, data = dati_neonati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1113.18  -181.16   -16.58   161.01  2620.19
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6707.4293   135.9438 -49.340  < 2e-16 ***
## N.gravidanze     12.3619     4.3325   2.853  0.00436 **
## Gestazione       31.9909     3.7896   8.442  < 2e-16 ***
## Lunghezza        10.3086     0.3004  34.316  < 2e-16 ***
## Cranio           10.4922     0.4254  24.661  < 2e-16 ***
## Tipo.partoNat    29.2803    12.0817   2.424  0.01544 *
## Ospedaleosp2    -11.0227    13.4363  -0.820  0.41209
## Ospedaleosp3     28.6408    13.4886   2.123  0.03382 *
## SessoM           77.4412    11.1756   6.930 5.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 273.9 on 2491 degrees of freedom
## Multiple R-squared:  0.7287, Adjusted R-squared:  0.7278
## F-statistic: 836.3 on 8 and 2491 DF,  p-value: < 2.2e-16
```

La funzione automatica del programma ha preferito il terzo modello.

Per il momento continueremo con il modello individuato dalla funzione e comincieremo con l'analisi dei residui del modello per individuare se ci sono problemi di normalità, omoschedasticità, linearità e leverage.

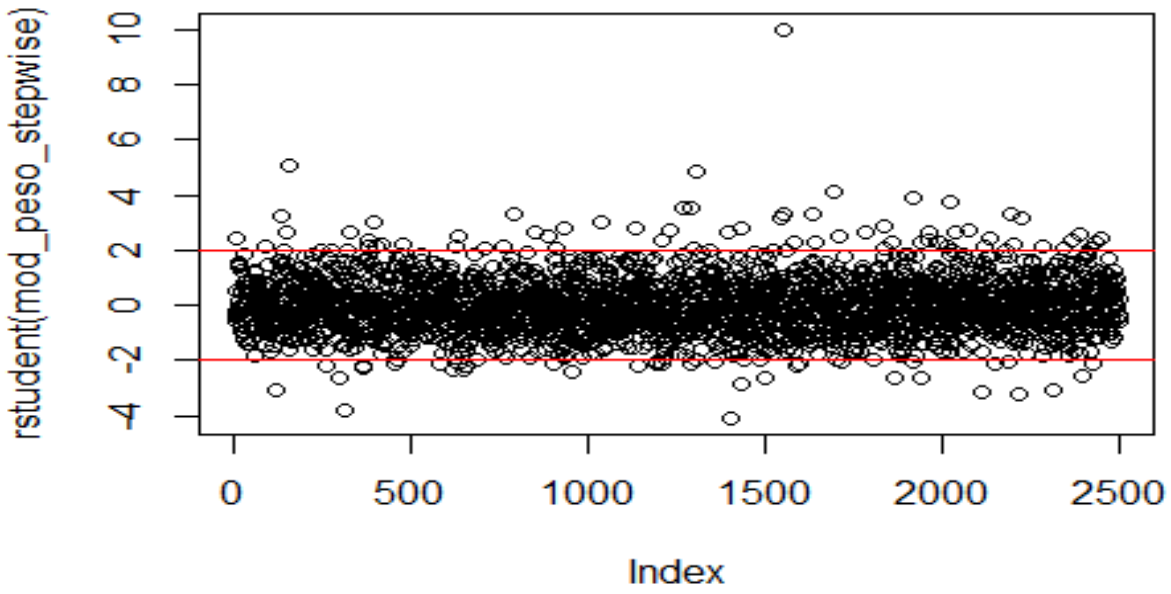
```
par(mfrow=c(2,2))
plot(mod_peso_stepwise)
```



Nel primo grafico Residuals vs Fitted: i punti presentano un pattern. E' evidente la presenza di una eteroschedasticità. Nel grafico Q-Q normal: si può indicare, nonostante una leggera deviazione agli estremi della coda, che i residui seguano una dstribuzione normale. Nel grafico Scale - Location : si nota una piccola e trascurabile curvatura. Presenta, però, un pattern piuttosto evidente. Nel grafico Residuals vs Leverage: si notano alcuni residui di osservazioni potenzialmente influenti che esplorerò di seguito

OUTLIERS

```
plot(rstudent(mod_peso_stepwise))
abline(h=c(-2,2), col="red")
```



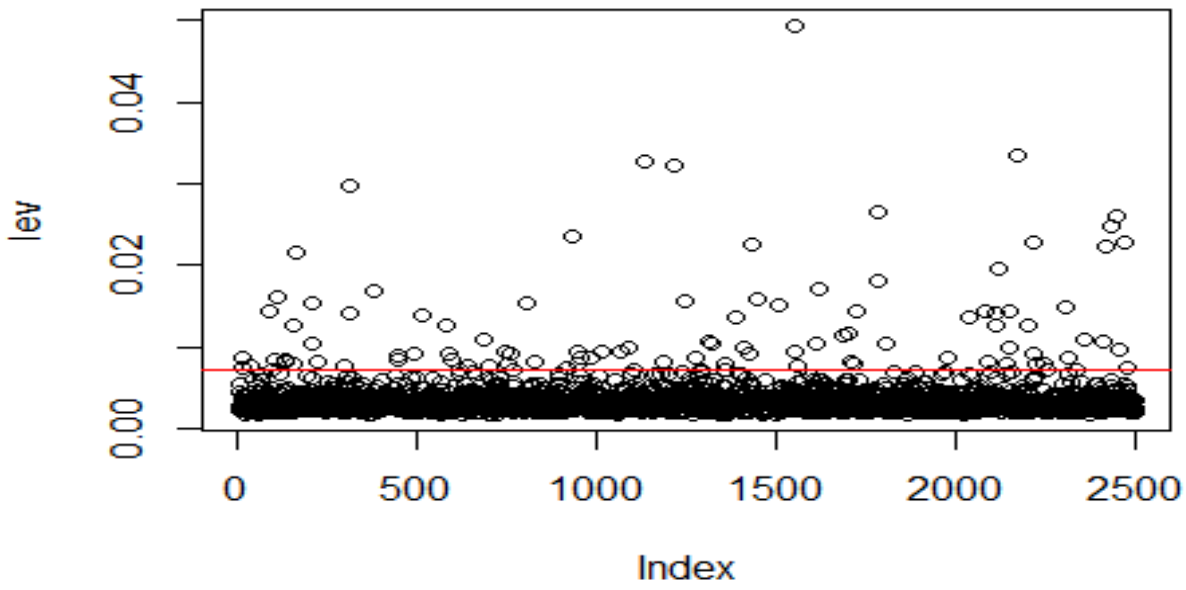

```
outlierTest(mod_peso_stepwise)
```

```
##       rstudent unadjusted p-value Bonferroni p
## 1551 10.004278         3.9642e-23   9.9104e-20
## 155   5.046640         4.8210e-07   1.2053e-03
## 1306  4.872419         1.1716e-06   2.9291e-03
```

Sono presenti tre osservazioni outliers sull'asse della variabile risposta.

LEVERAGE

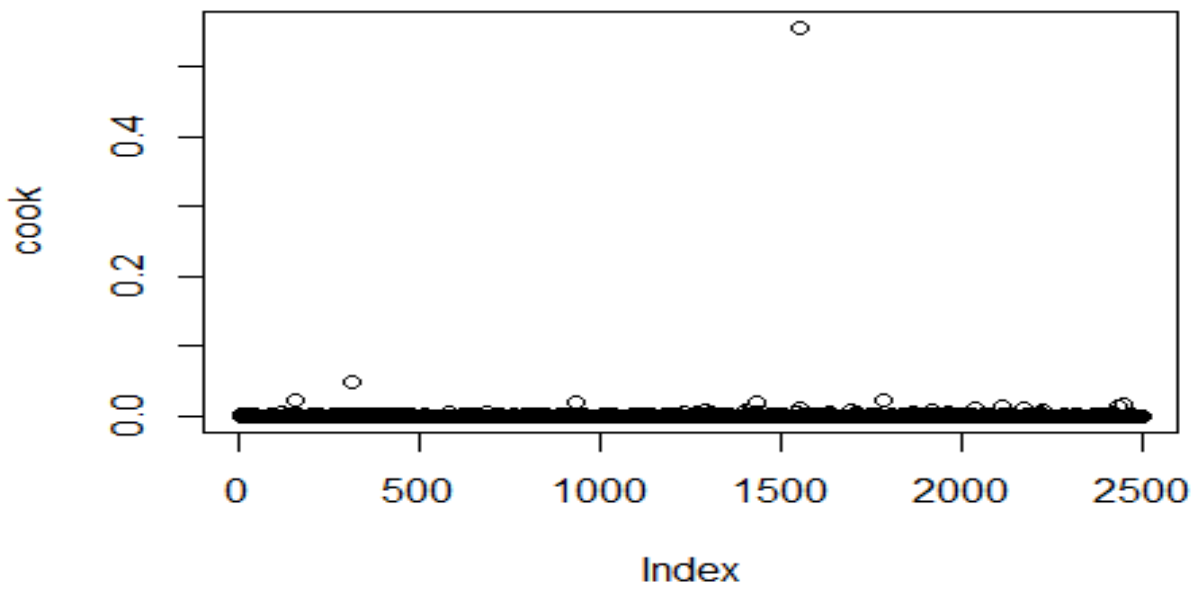
```
lev <- hatvalues(mod_peso_stepwise)
plot(lev)
p= sum(lev)
n= nrow(dati_neonati)
soglia= 2*p/n
abline(h=soglia, col="red")
```




```
lev_maggiori_soglia <- lev[lev>soglia]
n_lev <- length(lev_maggiori_soglia)
n_lev

## [1] 96
```

Sono presenti 96 osservazioni considerate come potenziali punti di leva sull'asse dei regressori. Prseguiro di seguito con lo studio della distanza di cook per vedere se qualcuna di queste osservazioni rappresenta un effettivo pericolo per il modello.

```
cook <- cooks.distance(mod_peso_stepwise)
plot(cook)
```




```
max(cook)

## [1] 0.5557527
```

Il valore di Cook è di 0.55. Da considerarsi una influenza misurata, poichè secondo la pratica comune, la distanza di cook diventa preoccupante quando è >1.

OMOSCHEDASTICITA'

```
bptest(mod_peso_stepwise)

##
##  studentized Breusch-Pagan test
##
## data:  mod_peso_stepwise
## BP = 91.768, df = 8, p-value < 2.2e-16
```

Viene rifiutata l'ipotesi nulla quindi la varianza non viene ritenuta costante ossia c'è un problema di eteroschedastica.

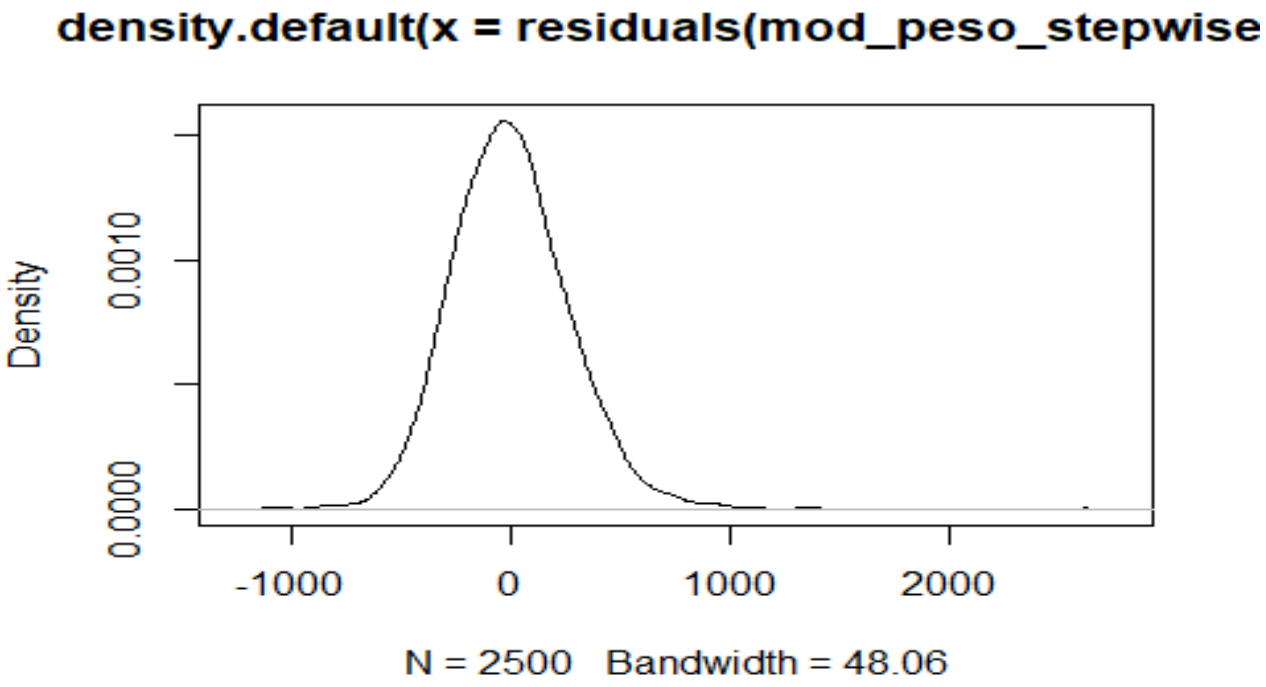
AUTOCORRELAZIONE

```
dwtest(mod_peso_stepwise)

##
##  Durbin-Watson test
##
## data:  mod_peso_stepwise
## DW = 1.9527, p-value = 0.1184
## alternative hypothesis: true autocorrelation is greater than 0
```

I residui non sono autocorrelati.

NORMALITA'

```
shapiro.test(residuals(mod_peso_stepwise))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_peso_stepwise)
## W = 0.97408, p-value < 2.2e-16

plot(density(residuals(mod_peso_stepwise)))
```



Si rifiuta l'ipotesi nulla di normalità per via del leverage, e come si può vedere dal grafico di densità, la distribuzione dei residui presenta una lunga coda a destra. Ciononostante il grafico di densità segue la forma della distribuzione normale, e visto che il test di Shapiro è estremamente sensibile, preferisco tenere come valida l'ipotesi di normalità.

Infine il nostro modello non ha superato il test di omoschedastcità. Ha presentato dei problemi di normalità e leverage, ma entrambi da ritenersi non particolarmente problematici. Proverò quindi a fare alcune trasformazioni per vedere se riesco a ottenere un modello migliore.

Il primo passo che proverò ad affrontare è quello di dare pesi diversi a ciascuna osservazione, in quanto i pesi sono in grado di riflettere l'importanza relativa delle singole osservazioni nel modello.

```
pesi <- 1/sqrt(fitted(mod_peso_stepwise))
```

```
mod_stepwise_pesato <- update(mod_peso_stepwise,~.,weights= pesi)
summary(mod_stepwise_pesato)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso, data = dati_neonati, weights = pesi)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -155.60  -24.11   -2.29   21.67  387.77
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6182.7986   115.0052 -53.761  < 2e-16 ***
## N.gravidanze     10.4910     4.3664   2.403   0.0163 *
## Gestazione       28.0771     3.6981   7.592 4.42e-14 ***
## Lunghezza         9.6773     0.2963  32.660  < 2e-16 ***
## Cranio           10.3267     0.4266  24.209  < 2e-16 ***
## Tipo.partoNat    28.8156    12.2881   2.345   0.0191 *
## Ospedaleosp2    -15.9515    13.6368  -1.170   0.2422
## Ospedaleosp3     28.0652    13.6905   2.050   0.0405 *
## SessoM           81.0427    11.3537   7.138 1.24e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.99 on 2491 degrees of freedom
## Multiple R-squared:  0.7675, Adjusted R-squared:  0.7668
## F-statistic:  1028 on 8 and 2491 DF,  p-value: < 2.2e-16
```

Da notare che con l'aggiunta dei pesi, c'è stato anche un aumento dell'R quadro al 76,68% dal 72% di prima.

OMOSCHEDASTICITA'

```
bptest(mod_stepwise_pesato)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod_stepwise_pesato
## BP = 9.2048, df = 8, p-value = 0.3253
```

E' stato risolto il problema di eteroschedasticità.

Multicollinearità

```
vif(mod_stepwise_pesato)
```

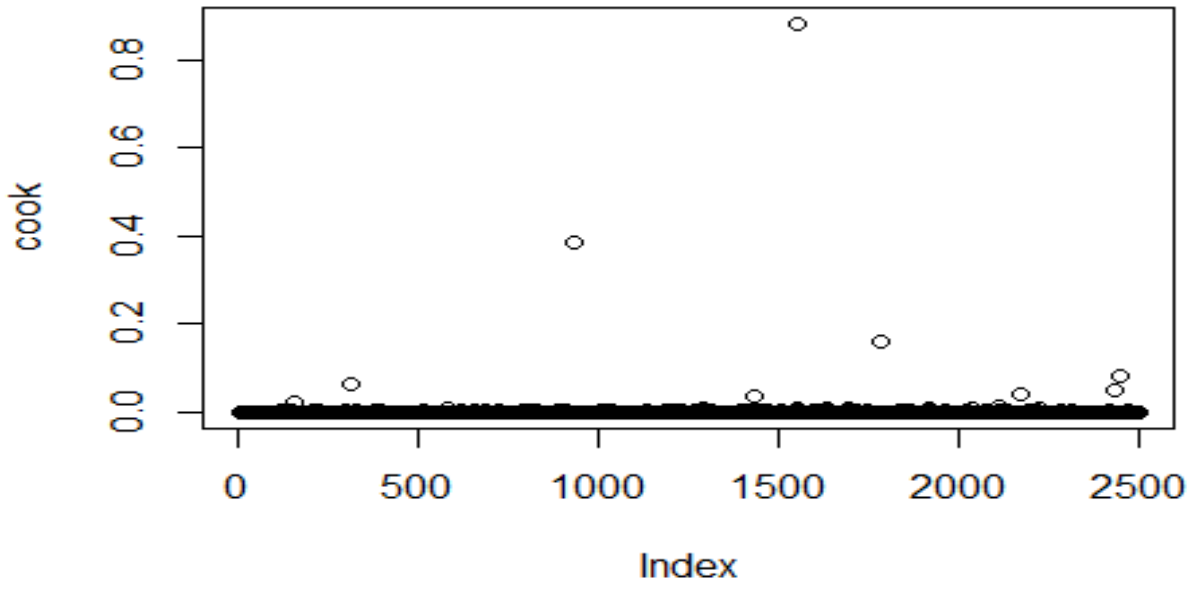
```
##                 GVIF Df GVIF^(1/(2*Df))
## N.gravidanze 1.023320  1        1.011593
## Gestazione   2.137963  1        1.462177
```

```
## Lunghezza    2.685473  1        1.638741
## Cranio       1.977764  1        1.406330
## Tipo.parto   1.004752  1        1.002373
## Ospedale     1.003349  2        1.000836
## Sesso        1.040307  1        1.019955
```

I regressori non presentano problemi di multicollinearità

DISTANZA DI COOK

```
cook <- cooks.distance(mod_stepwise_pesato)
plot(cook)
```




```
max(cook)
```

```
## [1] 0.8820776
```

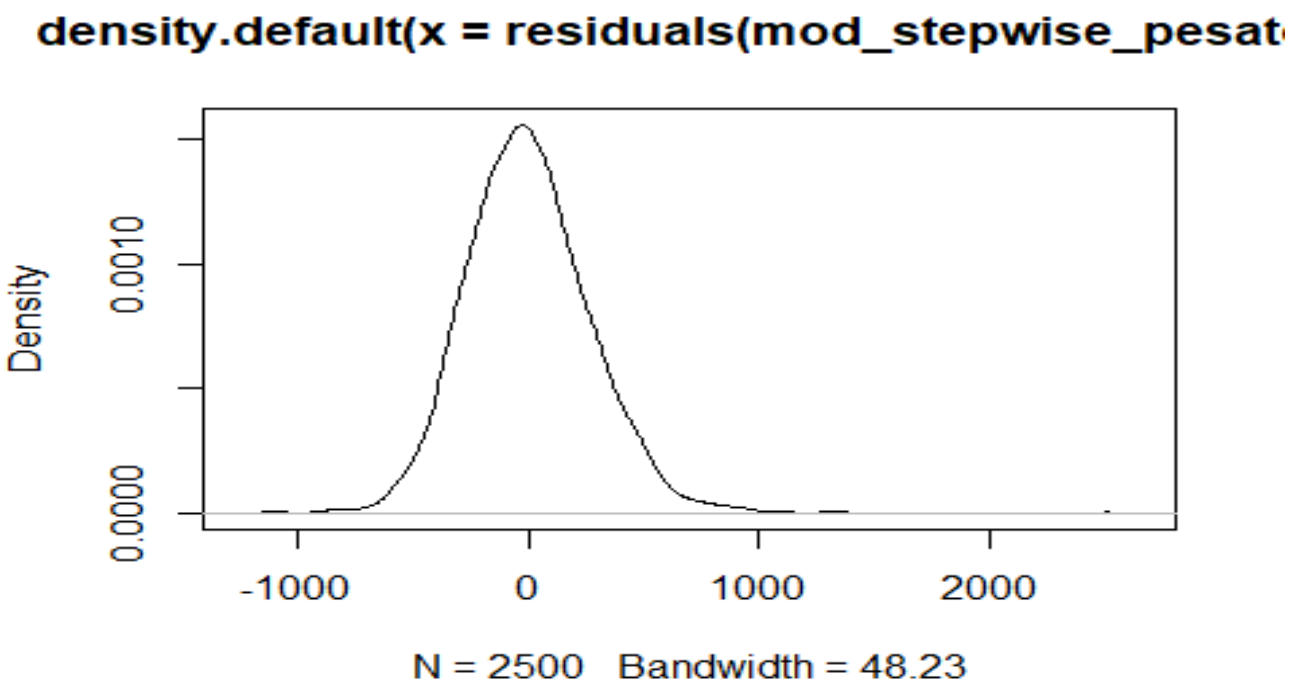
La distanza di Cook si è alzata notevolmente. Ciononostante è ancora sotto al livello soglia di 1.

NORMALITA'

```
shapiro.test(residuals(mod_stepwise_pesato))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_stepwise_pesato)
## W = 0.97654, p-value < 2.2e-16
```

```
plot(density(residuals(mod_stepwise_pesato)))
```



Si rifiuta ancora una volta l'ipotesi nulla di normalità.

Prima di procedere alla funzione di predizione, proveremo a costruire un modello robusto, meno sensibile ai fattori di leverage.

```
modello_robusto <- rlm(Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
Tipo.parto + Sesso, data = dati_neonati)
summary(modello_robusto)

##
## Call: rlm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Sesso, data = dati_neonati)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1144.647  -172.464    -9.951   167.093  2798.300
##
## Coefficients:
##               Value      Std. Error t value
## (Intercept)   -6798.6531   131.1493   -51.8390
## N.gravidanze     12.3356     4.1822     2.9495
## Gestazione       29.3922     3.6590     8.0328
## Lunghezza        11.0289     0.2900    38.0304
## Cranio            9.9993     0.4109    24.3370
## Tipo.partoNat    27.2661    11.6662     2.3372
## SessoM           81.7850    10.7921     7.5782
##
## Residual standard error: 252.1 on 2493 degrees of freedom
```

```
modello_robusto_pesato <- update(modello_robusto,~.,weights= pesi)
summary(modello_robusto_pesato)
```

```
##
## Call: rlm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Sesso, data = dati_neonati, weights = pesi)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -146.473  -23.142   -1.013   22.133  427.493
##
## Coefficients:
##               Value      Std. Error t value
## (Intercept)   -6574.5055   108.4265   -60.6356
## N.gravidanze     11.9335     4.1186     2.8975
## Gestazione       27.4815     3.4894     7.8757
## Lunghezza        10.8239     0.2795    38.7301
## Cranio            9.8591     0.4025    24.4951
## Tipo.partoNat    26.5774    11.5939     2.2924
## SessoM           83.2881    10.7123     7.7750
##
## Residual standard error: 33.79 on 2493 degrees of freedom
```

Tutti i regressori presentano un t-value significativo.

omoschedasticità modello robusto.

```
modello_robusto_pesato <- update(modello_robusto,~.,weights= pesi)
summary(modello_robusto_pesato)
```

```
##
## Call: rlm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Sesso, data = dati_neonati, weights = pesi)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -146.473  -23.142   -1.013   22.133  427.493
##
## Coefficients:
##               Value      Std. Error t value
## (Intercept)   -6574.5055   108.4265   -60.6356
## N.gravidanze     11.9335     4.1186     2.8975
## Gestazione       27.4815     3.4894     7.8757
## Lunghezza        10.8239     0.2795    38.7301
## Cranio            9.8591     0.4025    24.4951
## Tipo.partoNat    26.5774    11.5939     2.2924
## SessoM           83.2881    10.7123     7.7750
##
## Residual standard error: 33.79 on 2493 degrees of freedom
```

```
bptest(modello_robusto_pesato)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  modello_robusto_pesato
## BP = 9.17, df = 6, p-value = 0.1642
```

Il modello non presenta problemi di eteroschedasticità.

Multicollinearità

```
vif(modello_robusto_pesato)
```

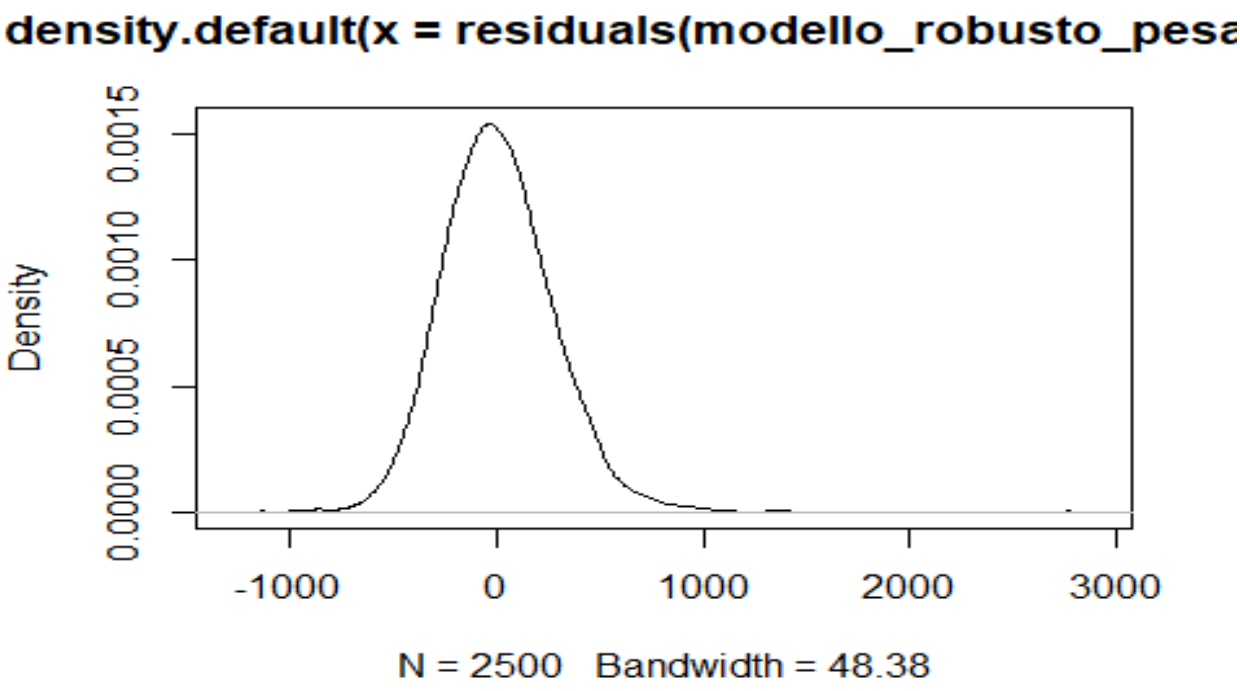
```
## N.gravidanze    Gestazione     Lunghezza        Cranio    Tipo.parto         Sesso
##     1.022440      2.137540      2.682782      1.977464      1.004446      1.040005
```

Il modello non presenta problemi di multicollinearità.

```
shapiro.test(residuals(modello_robusto_pesato))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modello_robusto_pesato)
## W = 0.97163, p-value < 2.2e-16
```

```
plot(density(residuals(modello_robusto_pesato)))
```



Il modello presenta problemi di normalità, ancora una volta dovuti ai punti di leva. Ma questo modello dovrebbe essere meno sensibile a questi fattori.

Ora costruiremo un modello con nessuna misura dell'ecografia.

```r
mod_peso_senza_misure <- lm(Peso ~ Anni.madre + Fumatrici + N.gravidanze +
Gestazione + Tipo.parto + Ospedale + Sesso, data = dati_neonati)
summary(mod_peso_senza_misure)
```

```
##
## Call:
## lm(formula = Peso ~ Anni.madre + Fumatrici + N.gravidanze + Gestazione +
##     Tipo.parto + Ospedale + Sesso, data = dati_neonati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1488.2  -270.1   -13.6   262.7  1905.9
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3306.932    187.373 -17.649  < 2e-16 ***
## Anni.madre        3.742      1.706   2.193  0.02838 *
## Fumatrici      -110.330     41.498  -2.659  0.00790 **
## N.gravidanze     18.437      7.010   2.630  0.00859 **
## Gestazione      163.461      4.520  36.163  < 2e-16 ***
## Tipo.partoNat    15.329     18.213   0.842  0.40007
## Ospedaleosp2     -1.426     20.273  -0.070  0.94395
## Ospedaleosp3     23.949     20.366   1.176  0.23974
## SessoM          164.712     16.699   9.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.4 on 2491 degrees of freedom
## Multiple R-squared:  0.3821, Adjusted R-squared:  0.3801
## F-statistic: 192.6 on 8 and 2491 DF,  p-value: < 2.2e-16
```

I regressori presentano una buona significatività, tranne i regressori “Ospedale” e ” Tipo. Parto”. Inoltre, e cosa più importante, l’R quadro è molto basso, appena il 38%, ovvero il modello può spiegare il 38% della variaibilità del modello. Di conseguenza ogni previsione possibile non è lontanamente affidabile.

Proveremo con una procedura stepwise, per vedere se riusciamo a migliorare l’R quadro.

```r
mod_peso_senza_misure_stepwise <-MASS:: stepAIC(mod_peso_senza_misure,
               direction = "both",
               k=2)
```

```
## Start:  AIC=30130.78
## Peso ~ Anni.madre + Fumatrici + N.gravidanze + Gestazione + Tipo.parto +
##     Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Ospedale      2    338648 426000786 30129
## - Tipo.parto    1    121043 425783181 30130
## <none>                      425662138 30131
## - Anni.madre    1    821976 426484114 30134
## - N.gravidanze  1   1182030 426844167 30136
```

```
## - Fumatrici     1   1207859 426869997 30136
## - Sesso         1  16624132 442286270 30225
## - Gestazione    1 223470325 649132463 31184
##
## Step:  AIC=30128.76
## Peso ~ Anni.madre + Fumatrici + N.gravidanze + Gestazione + Tipo.parto +
##     Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Tipo.parto    1    130235 426131021 30128
## <none>                      426000786 30129
## + Ospedale      2    338648 425662138 30131
## - Anni.madre    1    845501 426846287 30132
## - N.gravidanze  1   1206111 427206897 30134
## - Fumatrici     1   1231187 427231973 30134
## - Sesso         1  16674239 442675025 30223
## - Gestazione    1 223933725 649934511 31183
##
## Step:  AIC=30127.53
## Peso ~ Anni.madre + Fumatrici + N.gravidanze + Gestazione + Sesso
##
##                Df Sum of Sq       RSS   AIC
## <none>                      426131021 30128
## + Tipo.parto    1    130235 426000786 30129
## + Ospedale      2    347841 425783181 30130
## - Anni.madre    1    846924 426977945 30131
## - N.gravidanze  1   1188774 427319795 30133
## - Fumatrici     1   1215268 427346289 30133
## - Sesso         1  16667830 442798852 30222
## - Gestazione    1 223815017 649946038 31181
```

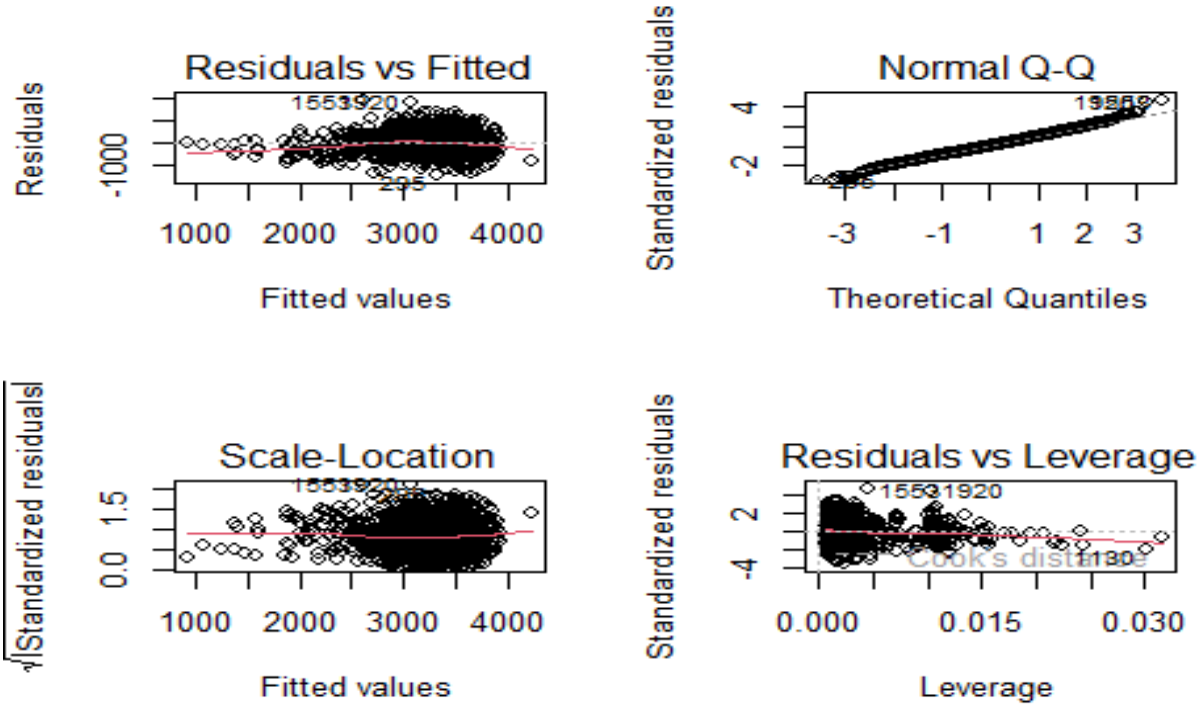
```
summary(mod_peso_senza_misure_stepwise)
```

```
##
## Call:
## lm(formula = Peso ~ Anni.madre + Fumatrici + N.gravidanze + Gestazione +
##     Sesso, data = dati_neonati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1466.6  -271.3   -12.0   261.1  1901.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3292.719    186.469 -17.658   <2e-16 ***
## Anni.madre       3.797      1.705   2.226   0.0261 *
## Fumatrici     -110.625     41.480  -2.667   0.0077 **
## N.gravidanze    18.477      7.005   2.638   0.0084 **
## Gestazione     163.525      4.518  36.193   <2e-16 ***
## SessoM         164.913     16.697   9.877   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 413.4 on 2494 degrees of freedom
## Multiple R-squared:  0.3814, Adjusted R-squared:  0.3802
## F-statistic: 307.6 on 5 and 2494 DF,  p-value: < 2.2e-16
```

La funzione, ha tolto i due regressori non significativi ma l'R quadro è rimasto invariato. A mio avviso, non avrebbe senso continuare con il modello poichè il risultato della previsione non sarebbe attendibile. Ciononostante proveremo a proseguire per vedere secondo questo modello quanto peserebbe il neonato di una madre alla 39sima settimana di gestazione e alla sua terza gravidanza.

```
par(mfrow=c(2,2))
plot(mod_peso_senza_misure_stepwise)
```



Nel primo grafico Residuals vs Fitted: i punti presentano un pattern. E' evidente la presenza di una eteroschedasticità. Nel grafico Q-Q normal: si può indicare, nonostante una leggera deviazione agli estremi della coda, che i residui seguano una dstribuzione normale. Nel grafico Scale - Location : si nota una piccola e trascurabile curvatura. Presenta, però, un pattern piuttosto evidente. Nel grafico Residuals vs Leverage: si notano alcuni residui di osservazioni potenzialmente influenti che esplorerò di seguito

OMOSCHEDATICITA'

```
bptest(mod_peso_senza_misure_stepwise)

##
##  studentized Breusch-Pagan test
##
```

```
## data:  mod_peso_senza_misure_stepwise
## BP = 8.8749, df = 5, p-value = 0.1142
```

I residui mostrano non avere problemi di eteroschedastcità

AUTOCORRELAZIONE

```
dwtest(mod_peso_senza_misure_stepwise)
```

```
##
##  Durbin-Watson test
##
## data:  mod_peso_senza_misure_stepwise
## DW = 1.8934, p-value = 0.00383
## alternative hypothesis: true autocorrelation is greater than 0
```

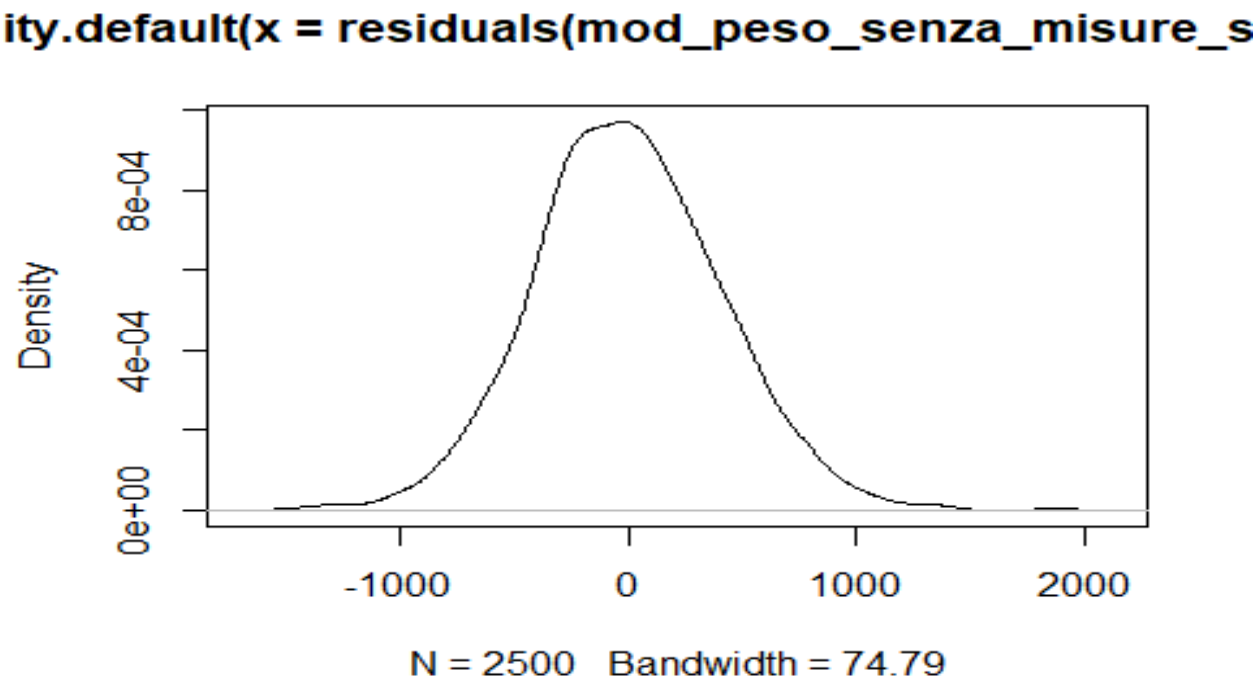
I residui mostrano dei problemi di autocorrelazione, ossia le stime dei coefficienti possono non essere accurate così come la previsione può non essere precisa.

NORMALITA'

```
shapiro.test(residuals(mod_peso_senza_misure_stepwise))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_peso_senza_misure_stepwise)
## W = 0.99693, p-value = 6.3e-05
```

```
plot(density(residuals(mod_peso_senza_misure_stepwise)))
```



Multicollinearità

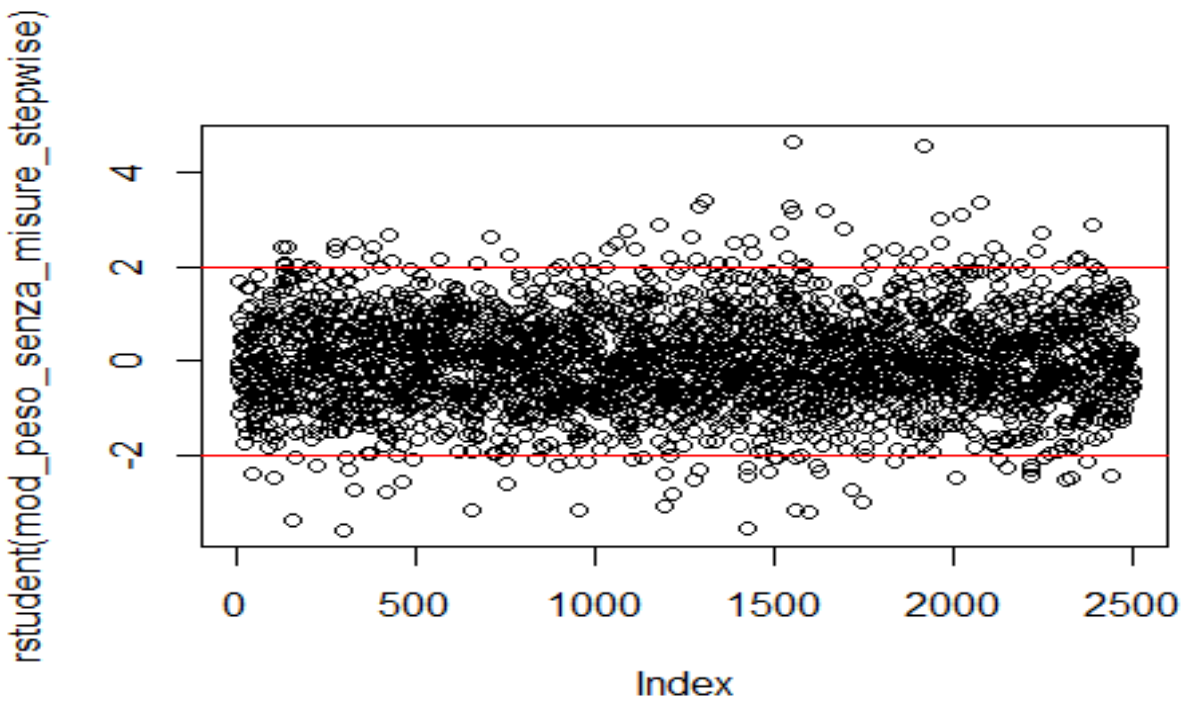
```
vif(mod_peso_senza_misure_stepwise)
```

```
##   Anni.madre    Fumatrici N.gravidanze   Gestazione        Sesso
##     1.182927     1.003719     1.176920     1.042552     1.019768
```

Non ci sono problemi di multicollinearità

OUTLIERS

```
plot(rstudent(mod_peso_senza_misure_stepwise))
abline(h=c(-2,2), col="red")
```



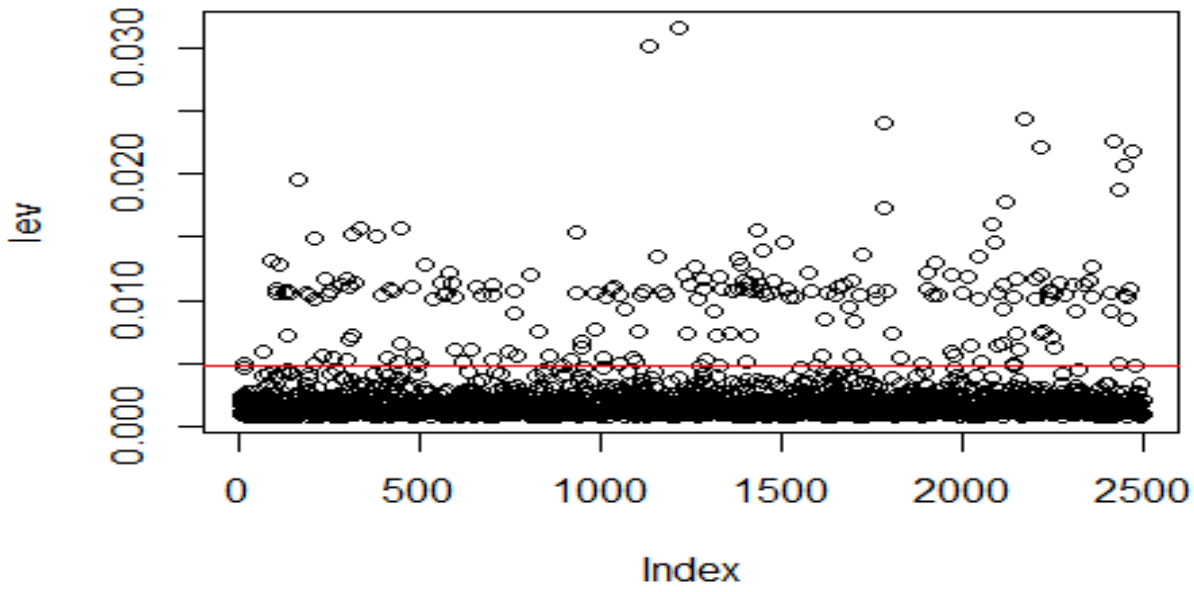

```
outlierTest(mod_peso_senza_misure_stepwise)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 1553 4.629687         3.8500e-06     0.009625
## 1920 4.534175         6.0563e-06     0.015141
```

Sono presenti 2 osservazioni outliers sull'asse della variabile risposta.

LEVERAGE

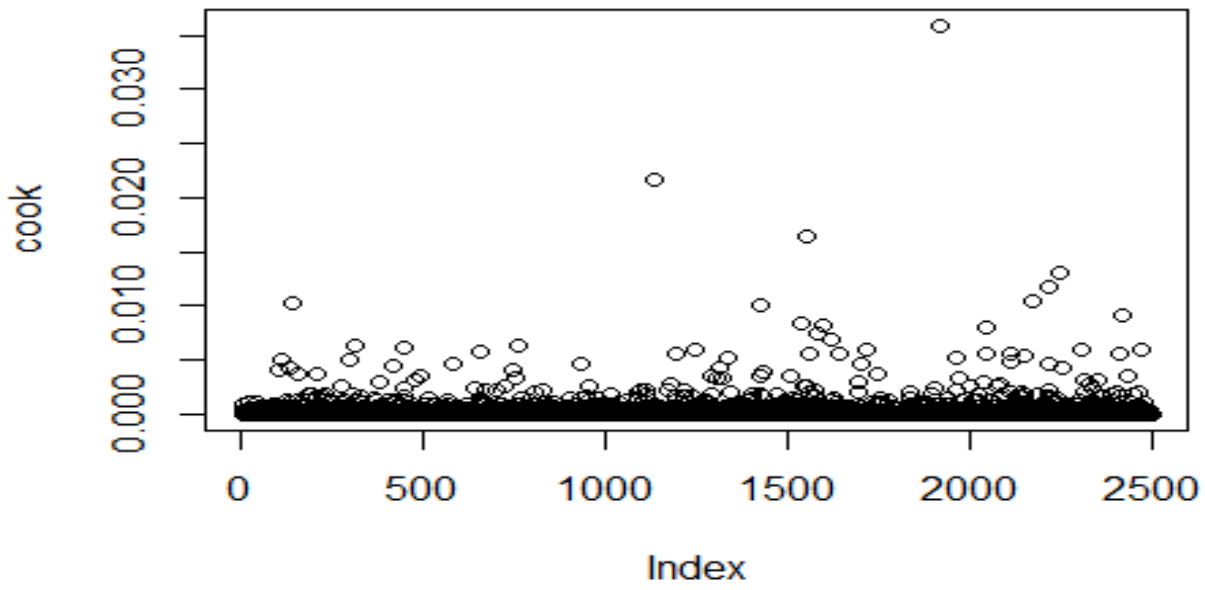
```
lev <- hatvalues(mod_peso_senza_misure_stepwise)
plot(lev)
p= sum(lev)
n= nrow(dati_neonati)
soglia= 2*p/n
abline(h=soglia, col="red")
```




```
lev_maggiori_soglia <- lev[lev>soglia]
n_lev <- length(lev_maggiori_soglia)
n_lev

## [1] 211
```

Sono presenti 211 osservazioni considerate come potenziali punti di leva sull'asse dei regressori. Prseguiro di seguito con lo studio della distanza di cook per vedere se qualcuna di queste osservazioni rappresenta un effettivo pericolo per il modello.

```
cook <- cooks.distance(mod_peso_senza_misure_stepwise)
plot(cook)
```




```
max(cook)
```

```
## [1] 0.03588354
```

Il valore di Cook è molto basso perciò i punti di leva del modello sono accettabili.

Di seguito procederemo con le funzioni di predizione in tale ordine: - PREDIZIONE CON FUNZIONE LM - STEPWISE E PESI - PREDIZIONE CON MODELLO ROBUSTO - PREDIZIONE CON MODELLO SENZA MSIURE DELL'ECOGRAFIA (prima con madre fumatrice e poi non fumatrice)

PREDIZIONE CON FUNZIONE LM - STEPWISE E PESI

```
nuovi_dati_peso <- data.frame(

  N.gravidanze = 3,
  Tipo.parto = "Nat",
  Sesso = "M",
  Lunghezza = mean(dati_neonati$Lunghezza),
  Ospedale = "osp2",

  Cranio =mean(dati_neonati$Cranio),
  Gestazione = 39
)



previsione_peso <- predict(mod_stepwise_pesato, newdata = nuovi_dati_peso)
previsione_peso
```

```
##        1
## 3336.238
```

Abbiamo creato una previsione del peso del futuro neonato con dati presi dalla media delle misure dell'ecografia e dalle variabili più significative delle variabili categoriche.

PREDIZIONE CON MODELLO ROBUSTO

```
dati_modello_robusto <- data.frame(

  N.gravidanze = 3,
  Gestazione = 39,
  Lunghezza = mean(dati_neonati$Lunghezza),
  Cranio = mean(dati_neonati$Cranio),
  Tipo.parto = "Nat",
  Sesso = "M"
)



previsione_modello_robusto <- predict(modello_robusto_pesato, newdata =
dati_modello_robusto)
previsione_modello_robusto
```

```
##        1
## 3349.799
```

Il peso stimato dal modello robusto è di 3349.799 gr.

PREDIZIONE CON MODELLO SENZA MSIURE DELL'ECOGRAFIA

```
dati_senza_eco <- data.frame(

  N.gravidanze = 3,
  Anni.madre = mean(dati_neonati$Anni.madre),
  Sesso = "M",
  Ospedale = "Osp2",
  Fumatrici = "0",
  Gestazione = 39
)

dati_senza_eco$Anni.Madre <- as.numeric(NA, levels =
levels(dati_neonati$Anni.Madre))
dati_senza_eco$Fumatrici <- as.numeric(dati_senza_eco$Fumatrici)



previsione_peso_senza_eco <- predict(mod_peso_senza_misure_stepwise, newdata =
dati_senza_eco)
previsione_peso_senza_eco
```

```
##        1
## 3412.042
```

La stima è di 3412 grammi. Di seguito proveremo a vedere se c'è un cambio di peso con la madre fumatrice.

```
dati_senza_eco <- data.frame(

  N.gravidanze = 3,
  Anni.madre = mean(dati_neonati$Anni.madre),
  Sesso = "M",
  Ospedale = "Osp2",
  Fumatrici = "1",
  Gestazione = 39
)

dati_senza_eco$Anni.Madre <- as.numeric(NA, levels =
levels(dati_neonati$Anni.Madre))
dati_senza_eco$Fumatrici <- as.numeric(dati_senza_eco$Fumatrici)



previsione_peso_senza_eco_fumatrice <- predict(mod_peso_senza_misure_stepwise,
newdata = dati_senza_eco)
previsione_peso_senza_eco_fumatrice
```

```
##        1
## 3301.417
```

Effettivamente c'è un leggero cambiamento di peso nel neonato. Di fatto, pesa di meno quando la madre è fumatrice. Tuttavia il risultato della predizione è da prendere con cautela in quanto l'R quadro del modello è di appena il 38%.

In conclusione abbiamo avuto modo di creare 3 modelli diversi di regressione multipla, fare alcune considerazioni di statistica descrittiva e formulare alcune ipotesi sulla media delle variaibili chiave. Grazie a questi modelli, ci è possibile stimare il futuro il peso di un neonato a partire dai dati fisiologici e dalle ecografie della madre. Ma abbiamo anche visto che fattori comportamentali, come l'essere fumatrice, può incidere sul peso del futuro neonato così come incidere sulla media del numero di settimane di gestazione tra una madre fumatrice e non fumatrice.