Un modello statistico per prevedere il peso dei neonati

Elisa Lipari

2024-02-14

1)Importa il dataset "neonati.csv" e controlla che sia stato letto correttamente dal software

```
setwd("C:/Users/SAMSUNG/Desktop/ALL FOLDERS/MASTER Data Science/statistica inferen
ziale")

getwd()
```

```
## [1] "C:/Users/SAMSUNG/Desktop/ALL FOLDERS/MASTER Data Science/statistica infere
nziale"
```

```
dati=read.csv("neonati.csv", stringsAsFactors = T)
attach(dati)
n=nrow(dati)
```

2) Descrivi il dataset, la sua composizione, il tipo di variabili e l'obbiettivo dello studio

I dati provengono da 3 ospedali e riguardano 2500 neonati. Il dataset contiene 10 variabili riportate di seguito:

-età della madre: variabile quantitativa continua

-numero di gravidanze sostenute: variabile quantitativa discreta

-Madre fumatrice (0=NO, SI=1): variabile dummy

-N° di settimane di gestazione: variabile quantitativa continua

-peso in grammi del neonato: variabile quantitativa continua

-Lunghezza in mm del neonato: variabile quantitativa continua

-Diametro in mm del cranio del neonato: variabile quantitativa continua

-Tipo di parto: variabile qualitativa su scala nominale

-Ospedale: variabile qualitativa su scala nominale

-Sesso del neonato: variabile qualitativa su scala nominale

Lo scopo dello studio è quello di valutare se è possibile prevedere il peso del neonato alla nascita considerando soprattutto la relazione con le variabili della madre, per capire se queste hanno o meno un effetto significativo.

3) Indaga le variabili effettuando una breve analisi descrittiva, utilizzando indici e strumenti grafici che conosci

Utilizzo la funzione *summary()* per ottenere le principali info delle variabili. Inoltre per le variabili quantitative valuto gli indici di forma mentre per le variabili qualitative valuto la distribuzione di frequenza.

ETA'DELLA MADRE

```
summary(Anni.madre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   25.00   28.00   28.16   32.00   46.00
```

Dal summary mi accorgo che c'è un errore nel range min-max. Esplorando il file csv identifico 0 e 1 in questa variabile. Quindi escludo le righe della variabile Anni.madre in cui siano presenti lo 0 e l'1 dal dataset e ripeto il *summary()*

```
dati<- subset(dati, Anni.madre != 0 & Anni.madre != 1)
n=nrow(dati)
attach(dati)
```

```
## The following objects are masked from dati (pos = 3):
##
##     Anni.madre, Cranio, Fumatrici, Gestazione, Lunghezza, N.gravidanze,
##     Ospedale, Peso, Sesso, Tipo.parto
```

```
summary(Anni.madre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   25.00   28.00   28.19   32.00   46.00
```

l'età media delle madri è circa 28 anni.

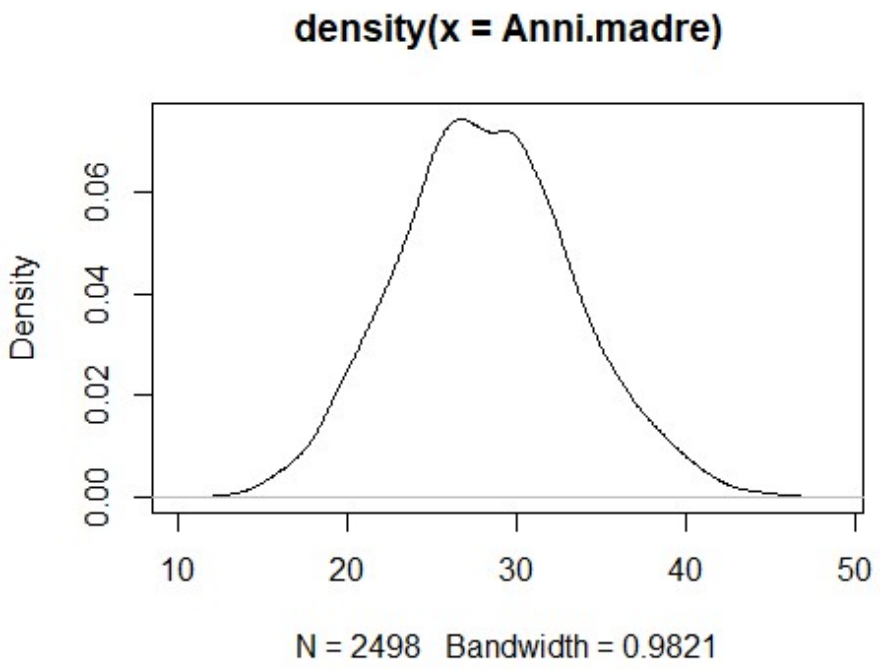
```
#indici di forma:

moments::skewness(Anni.madre)
```

```
## [1] 0.1510624
```

```
moments::kurtosis(Anni.madre)-3
```

```
## [1] -0.1056061
```

```
plot(density(Anni.madre))
```

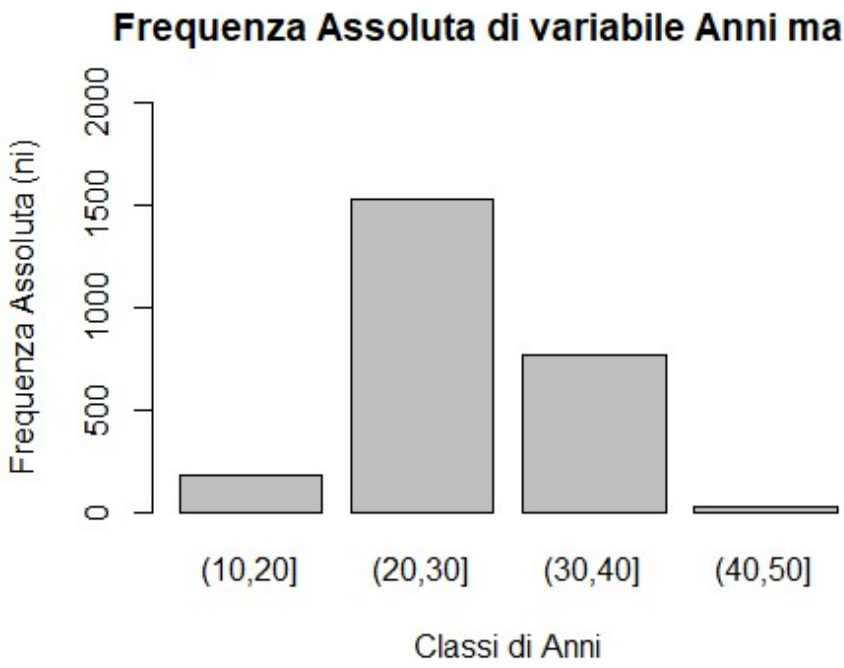


I valori di skewness e kurtosis indicano una distribuzione asimmetrica positiva e platicurtica.

Osserviamo graficamente i dati

```
#creo la classe Anni_cl con la funzione cut

anni_cl= cut (Anni.madre, breaks = c(10,20,30,40,50))


#creo la distribuzione di frequenze di anni_cl

ni=table (anni_cl)
fi=table(anni_cl)/n
Ni=cumsum(ni)
Fi=Ni/n

#la visualizzo in forma tabellare:

distr_freq=as.data.frame (cbind(ni,fi,Ni,Fi))
distr_freq

##           ni         fi   Ni         Fi
## (10,20]  175 0.07005604  175 0.07005604
## (20,30] 1527 0.61128903 1702 0.68134508
## (30,40]  767 0.30704564 2469 0.98839071
## (40,50]   29 0.01160929 2498 1.00000000
```

```r
#faccio il grafico delle frequenze assolute


Freq=barplot(distr_freq$ni,
        main = "Frequenza Assoluta di variabile Anni madre",
        xlab = "Classi di Anni",
        ylab = "Frequenza Assoluta (ni)",
        ylim = c(0, 2000),
        names.arg = rownames(distr_freq))
```



Il grafico mostra sull'asse delle x le classi dell'età delle donne prese in esame per questo studio e sull'asse delle y la frequenza osservata. Dal risultato ottenuto possiamo dire che la maggior parte delle donne prese in esame ha tra i 21 e 30 anni.

NUMERO GRAVIDANZE

```r
summary(N.gravidanze)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.9816  1.0000 12.0000
```

```r
#indici di forma:

moments::skewness(N.gravidanze)
```

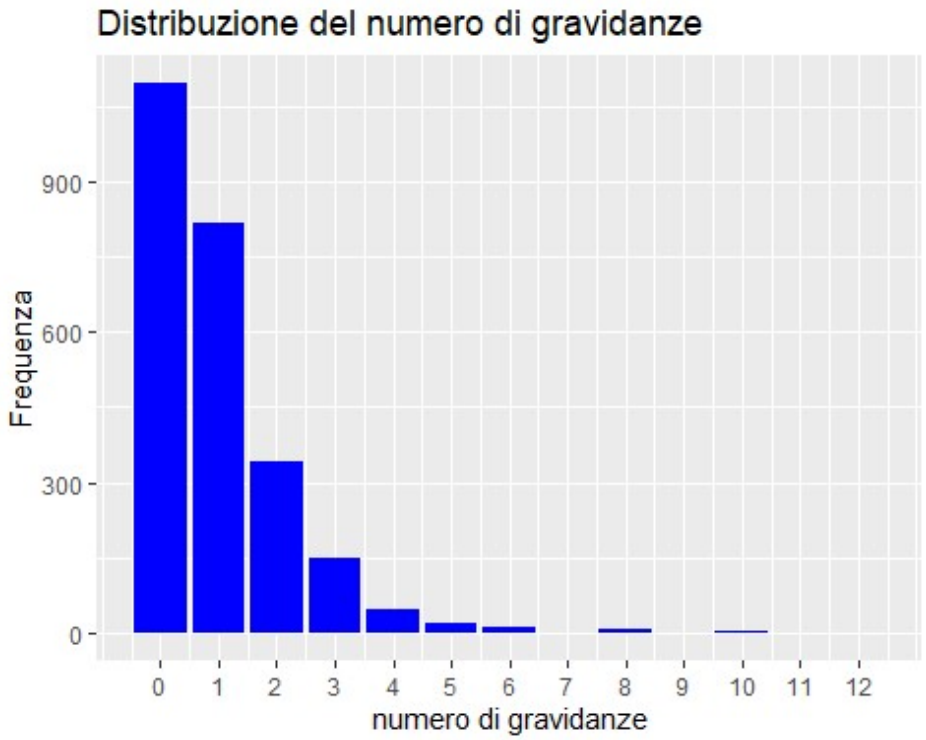
```
## [1] 2.513412
```

```r
moments::kurtosis(N.gravidanze)-3
```

```
## [1] 10.98163
```

La distribuzione di questa variabile è asimmetrica positiva e leptocurtica.

```
#visualizzo graficamente
library(ggplot2)

ggplot(dati)+
  geom_bar(aes(x = N.gravidanze), stat ="count", fill = "blue") +
  labs(title="Distribuzione del numero di gravidanze",
       x="numero di gravidanze",
       y="Frequenza")+
  scale_x_continuous(breaks=seq(0,12,1))
```



Il grafico riporta sull'asse delle x il numero di gravidanze delle donne prese in esame e sull'asse delle y la frequenza ovvero la conta delle osservazioni che ricadono in quella categoria. Il grafico mostra come la maggior parte delle donne non abbia gravidanze pregresse.

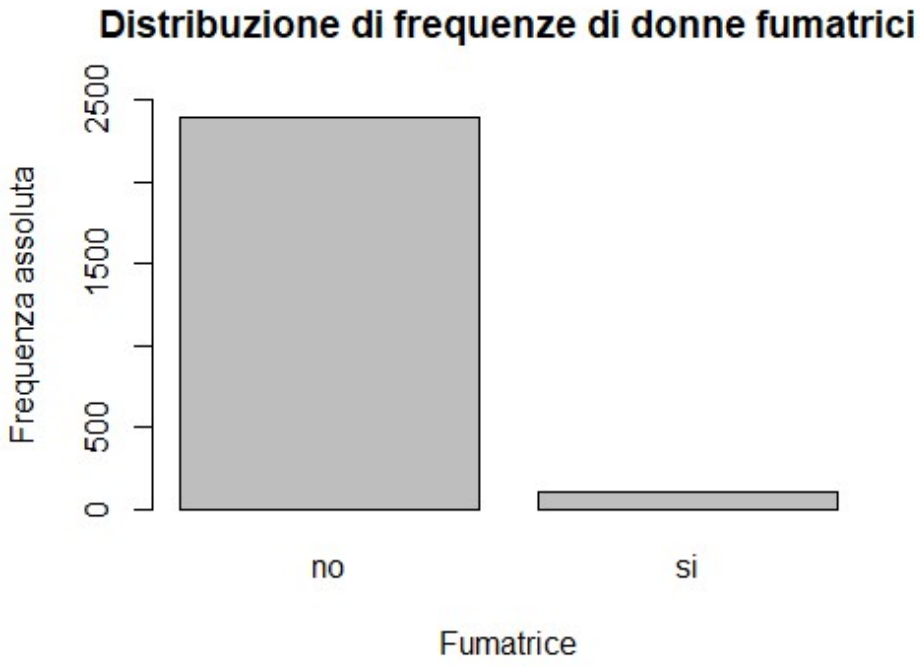
FUMATRICI

questa variabile è una variabile qualitativa in formato binario. Calcoliamo la distribuzione di frequenze.

```
Freq_ass=table(Fumatrici)
Freq_rel=table(Fumatrici)/n

distr_freq=cbind(Freq_ass,Freq_rel)
distr_freq
```

```
##   Freq_ass   Freq_rel
## 0     2394 0.95836669
## 1      104 0.04163331
```

```
barplot(Freq_ass,
        main = "Distribuzione di frequenze di donne fumatrici",
        ylab = "Frequenza assoluta",
        xlab = "Fumatrice",
        ylim = c(0, 2500),
        names.arg = c("no", "si"))
```



Su 2498 donne, il 95% delle donne prese in esame non è fumatrice. Infatti di 2498 donne solo 104 fumano.

GESTAZIONE

```
summary (Gestazione)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.00   38.00   39.00   38.98   40.00   43.00
```

```
#indici di forma:

moments::skewness(Gestazione)
```

```
## [1] -2.065131
```

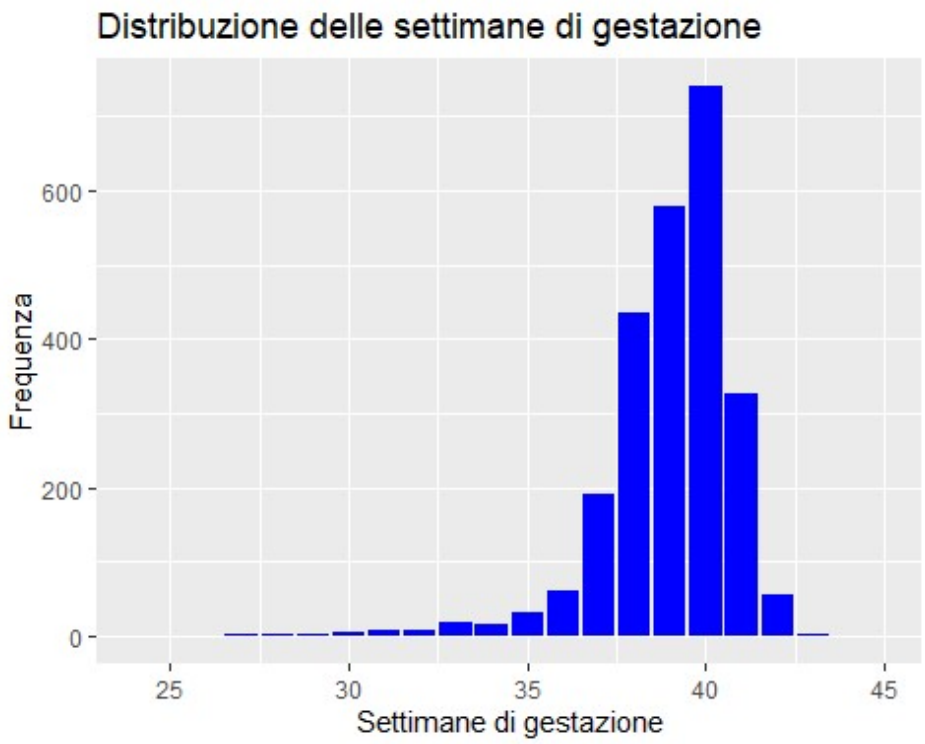
```
moments::kurtosis(Gestazione)-3
```

```
## [1] 8.255516
```

il numero di settimane di gestazione va da 25 a 43 con una media di circa 39 settimane. Dagli indici di forma possiamo dire che la distribuzione di questa variabile è asimmetrica negativa e leptocurtica

visualizziamo i dati graficamente:

```
library(ggplot2)

grafico=ggplot(dati)+
  geom_bar(aes(x = Gestazione), stat ="count", fill = "blue") +
  labs(title="Distribuzione delle settimane di gestazione",
       x="Settimane di gestazione",
       y="Frequenza")+
  scale_x_continuous(limits = c(24, 45))


grafico
```



Il grafico rappresenta sull'asse delle x le settimane di gestazione e sull'asse delle y la frequenza delle osservazioni. La maggior parte delle donne in esame in questo studio presenta 40 settimane di gestazione.

PESO in grammi del neonato

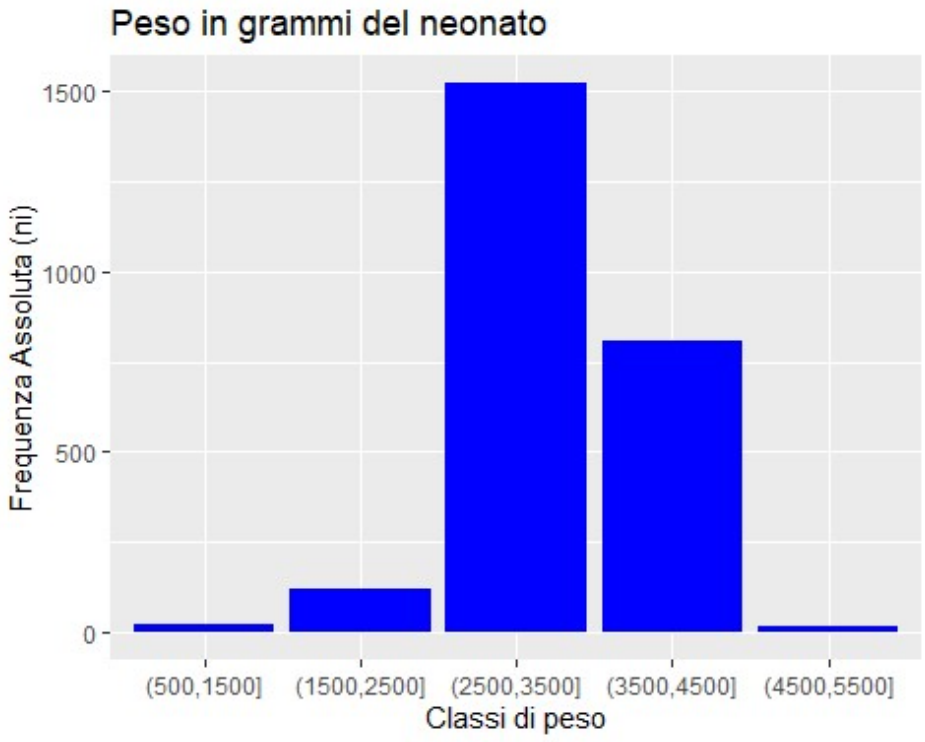
```
summary (Peso)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     830    2990    3300    3284    3620    4930
```

```r
#indici di forma:

moments::skewness(Peso)

## [1] -0.6474036

moments::kurtosis(Peso)-3

## [1] 2.028753
```

I bambini nati hanno un peso medio di circa 3,3kg con un massimo di 4,9kg e un minimo di 0,830kg. Dagli indici di forma possiamo dire che la distribuzione è leggermente asimmetrica negativa e leptocurtica.

Vediamo la frequenza di osservazioni graficamente:

```r
#creo la classe Peso_cl con la funzione cut

peso_cl= cut (Peso, breaks = c(500, 1500, 2500, 3500, 4500,5500))


#creo la distribuzione di frequenze di peso

ni=table (peso_cl)
fi=table(peso_cl)/n
Ni=cumsum(ni)
Fi=Ni/n

#la visualizzo in forma tabellare:

distr_freq=as.data.frame (cbind(ni,fi,Ni,Fi))
distr_freq

##                      ni          fi   Ni          Fi
## (500,1.5e+03]        24 0.009607686   24 0.009607686
## (1.5e+03,2.5e+03]   123 0.049239392  147 0.058847078
## (2.5e+03,3.5e+03] 1524 0.610088070 1671 0.668935148
## (3.5e+03,4.5e+03]  809 0.323859087 2480 0.992794235
## (4.5e+03,5.5e+03]   18 0.007205765 2498 1.000000000

#faccio il grafico delle frequenze assolute

library(ggplot2)
Freq=ggplot(dati)+
    geom_bar(aes(x = peso_cl), stat ="count", fill = "blue") +
    labs(title="Peso in grammi del neonato",
        x="Classi di peso",
        y="Frequenza Assoluta (ni)")+
  scale_x_discrete(labels = c("(500,1500]", "(1500,2500]", "(2500,3500]", "(3500,4
500]", "(4500,5500]"))

Freq
```



Il grafico mette in relazione il peso in grammi del neonato(asse x) e la frequenza assoluta delle osservazioni (asse y). Il peso più frequentemente osservato si aggira tra 2501 e 3500 grammi.

LUNGHEZZA(in mm del neonato)

```
summary (Lunghezza)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   310.0   480.0   500.0   494.7   510.0   565.0
```

```
#indici di forma:


moments::skewness(Lunghezza)
```

```
## [1] -1.514575
```

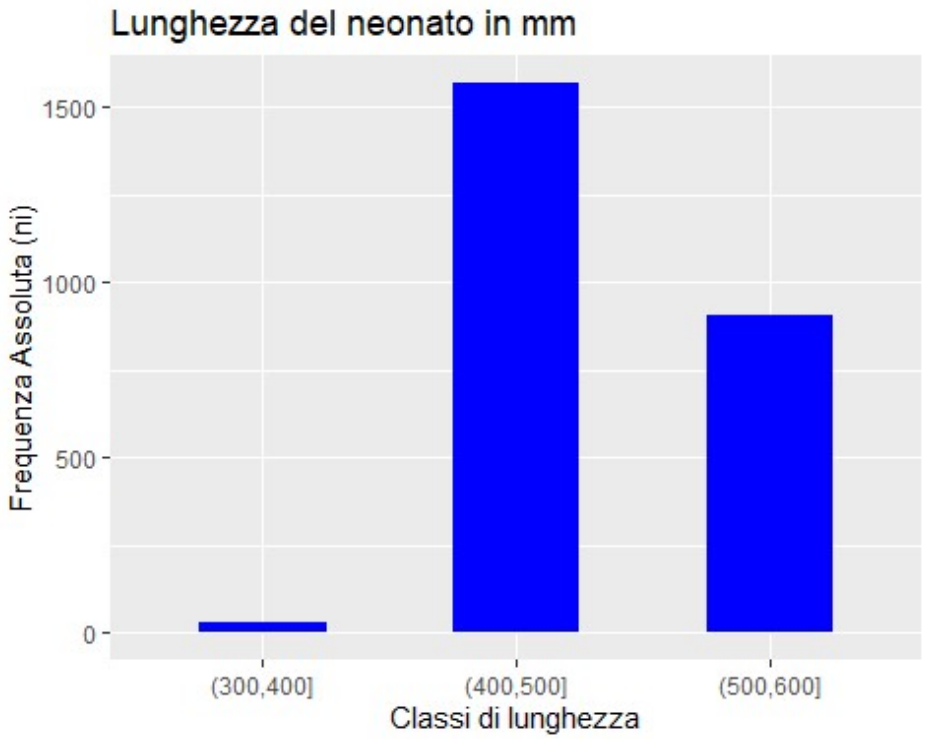
```
moments::kurtosis(Lunghezza)-3
```

```
## [1] 6.48093
```

La lunghezza del neonato in mm va da 310 a 565 con una media di circa 495mm. Dagli indici di forma possiamo dire che la distribuzione di questa variabile è asimmetrica negativa e leptocurtica.

Visualizziamo i dati graficamente:

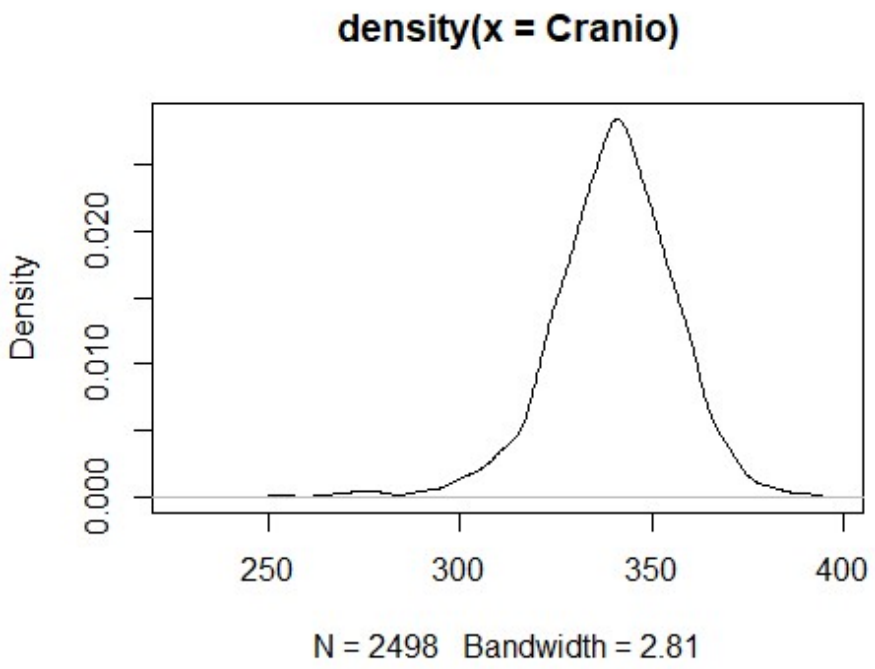
```
#creo la classe lunghezza_cl con la funzione cut

lunghezza_cl= cut (Lunghezza, breaks = c(300,400,500,600))


#creo la distribuzione di frequenze di sales

ni=table (lunghezza_cl)
fi=table(lunghezza_cl)/n
Ni=cumsum(ni)
Fi=Ni/n

#la visualizzo in forma tabellare:

distr_freq=as.data.frame (cbind(ni,fi,Ni,Fi))
distr_freq

##             ni         fi   Ni         Fi
## (300,400]   26 0.01040833   26 0.01040833
## (400,500] 1568 0.62770216 1594 0.63811049
## (500,600]  904 0.36188951 2498 1.00000000

#faccio il grafico delle frequenze assolute

library(ggplot2)
Freq=ggplot(dati)+
    geom_bar(aes(x = lunghezza_cl), stat ="count", fill = "blue",width = 0.5) +
    labs(title="Lunghezza del neonato in mm",
        x="Classi di lunghezza",
        y="Frequenza Assoluta (ni)")

Freq
```



Il grafico mostra la lunghezza in mm dei neonati presenti nel dataset(asse x) e la frequenza assoluta delle osservazioni per ogni lunghezza (asse y). La maggior parte dei bambini presi in esame in questo studio sono alti tra i 401 e 500mm.

DIAMETRO in mm del cranio del neonato

```
summary (Cranio)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     235     330     340     340     350     390

#indici di forma:

plot(density(Cranio))
```




```
moments::skewness(Cranio)
```

```
## [1] -0.7850906
```

```
moments::kurtosis(Cranio)-3
```

```
## [1] 2.94487
```

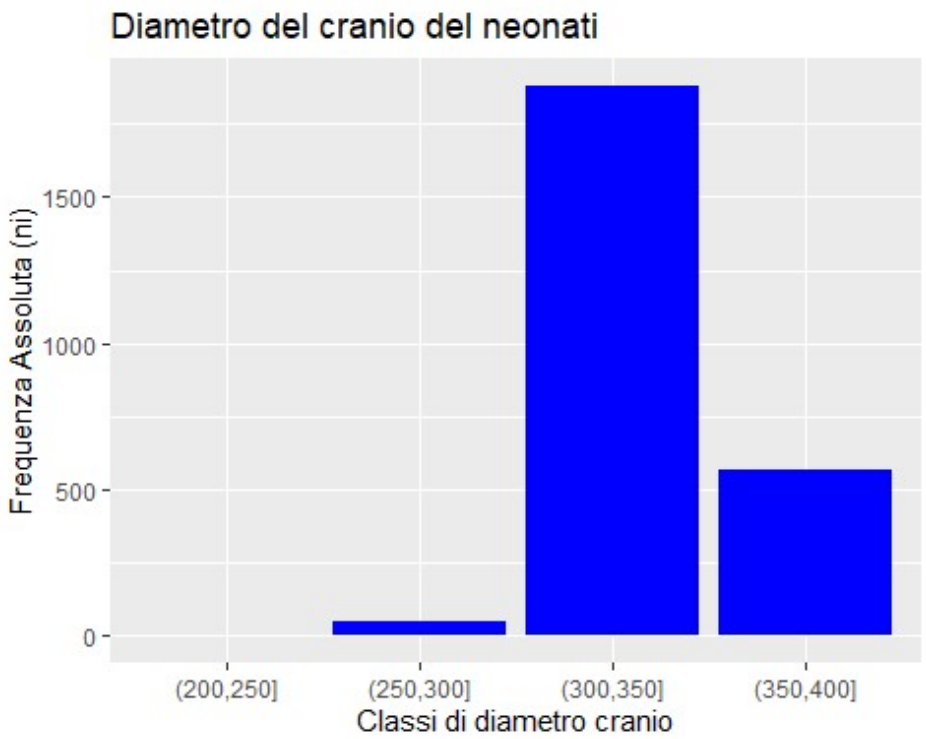
Il diametro del cranio dei bambini in questo studio va da 235 a 390mm con una media di circa 340. La distribuzione della variabile Cranio è leggermente asimmetrica negativa e leptocurtica.

visualizziamo i dati graficamente:

```
#creo la classe cranio_cl con la funzione cut

cranio_cl= cut (Cranio, breaks = c(200,250,300,350,400))

#creo la distribuzione di frequenze di sales

ni=table (cranio_cl)
fi=table(cranio_cl)/n
Ni=cumsum(ni)
Fi=Ni/n

#la visualizzo in forma tabellare:

distr_freq=as.data.frame (cbind(ni,fi,Ni,Fi))
distr_freq
```

```
##             ni           fi   Ni           Fi
## (200,250]    2 0.0008006405    2 0.0008006405
## (250,300]   49 0.0196156926   51 0.0204163331
## (300,350] 1880 0.7526020817 1931 0.7730184147
## (350,400]  567 0.2269815853 2498 1.0000000000
```

```
#faccio il grafico delle frequenze assolute

library(ggplot2)
Freq=ggplot(dati)+
    geom_bar(aes(x = cranio_cl), stat ="count", fill = "blue") +
    labs(title="Diametro del cranio del neonati",
        x="Classi di diametro cranio",
        y="Frequenza Assoluta (ni)")

Freq
```



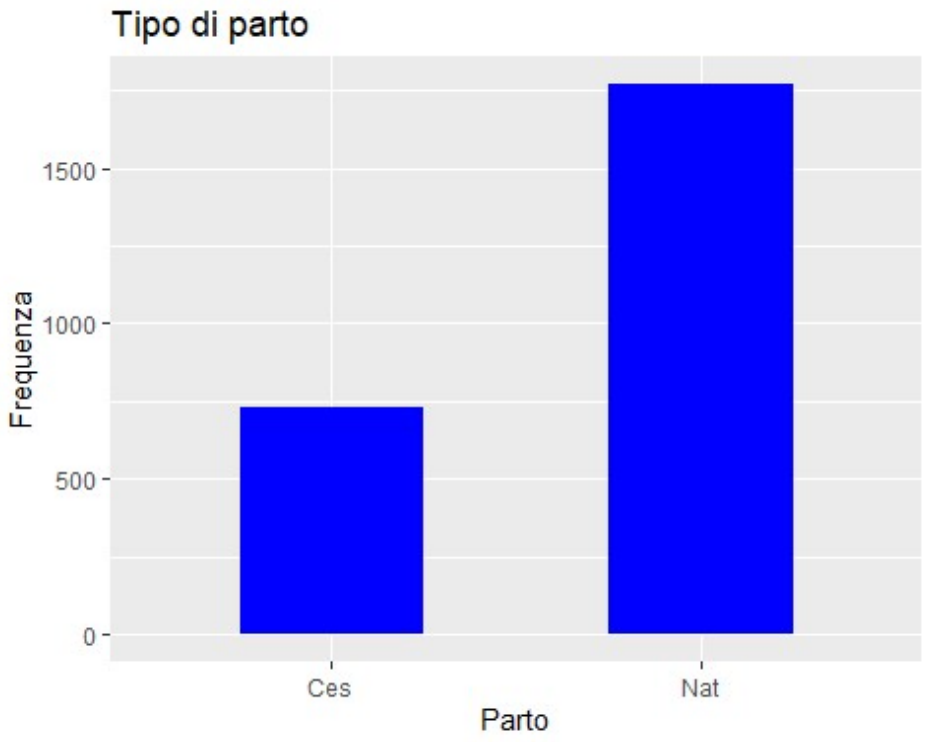
Il grafico riporta il diametro in mm del cranio dei neonati sulla asse delle x e la frequenza delle osservazioni sull'asse delle y. La maggior parte dei neonati presenta un diametro del cranio tra i 301 e 350mm.

TIPO DI PARTO

```
table(Tipo.parto)
```

```
## Tipo.parto
##  Ces  Nat
##  728 1770
```

```r
library(ggplot2)

Freq=ggplot(dati)+
  geom_bar(aes(x = Tipo.parto), stat ="count", fill = "blue",width = 0.5) +
  labs(title="Tipo di parto",
       x="Parto",
       y="Frequenza")

Freq
```

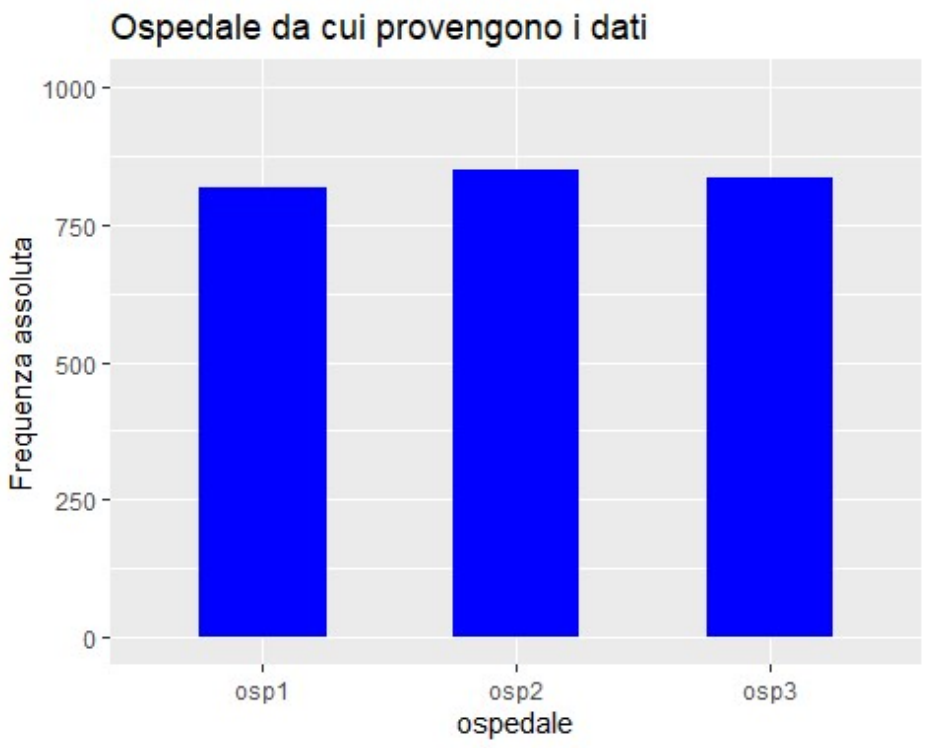


Il grafico mostra sull'asse delle x il tipo di parto, cesareo o naturale, e sull'asse delle y la frequenza assoluta. La maggior parte delle donne in questo studio ha affrontato un parto naturale.

OSPEDALE

```r
#calcolo la frequenza assoluta

table(Ospedale)

## Ospedale
## osp1 osp2 osp3
##  816  848  834

#grafico

library(ggplot2)
```

```r
Freq=ggplot(dati)+
  geom_bar(aes(x = Ospedale), stat ="count", fill = "blue",width = 0.5) +
  labs(title="Ospedale da cui provengono i dati",
       x="ospedale",
       y="Frequenza assoluta")+
  scale_y_continuous(limits = c(0,1000))

Freq
```

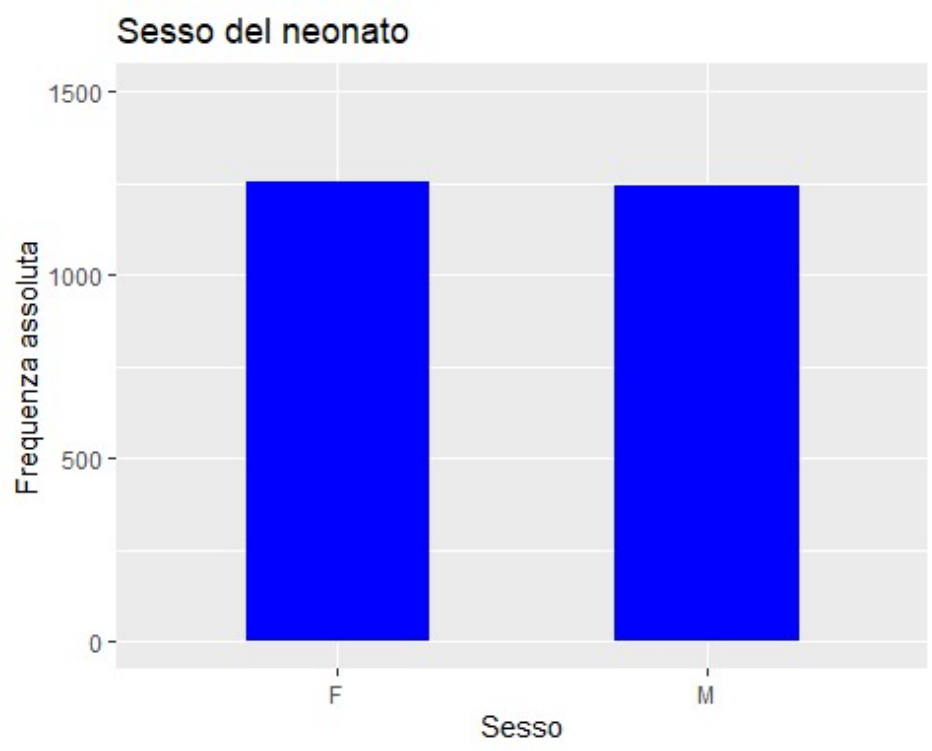


Il grafico riporta sull'asse delle x gli ospedali da cui provengono i dati raccolti e sull'asse delle y la frequenza assoluta. I dati provengono in egual modo da tutti e tre gli ospedali.

SESSO del neonato

```r
#calcolo la frequenza assoluta

table(Sesso)

## Sesso
##    F    M
## 1255 1243

#grafico

library(ggplot2)

Freq=ggplot(dati)+
  geom_bar(aes(x = Sesso), stat ="count", fill = "blue",width = 0.5) +
```

```
  labs(title="Sesso del neonato",
       x="Sesso",
       y="Frequenza assoluta")+
  scale_y_continuous(limits = c(0,1500))

Freq
```



Il grafico riporta la frequenza assoluta (asse y) della variabile sesso (asse x). La distribuzione dei neonati maschi e femmine all'interno del dataset è molto simile.

4) Saggia l'ipotesi che la media del peso e della lunghezza di questo campione di neonati siano significativamente uguali a quelle della popolazione

Mediamente il peso e l'altezza nella popolazione di neonati in italia è di 3,3kg e 50cm rispettivamente (fonte: https://www.ospedalebambinogesu.it)

Non conoscendo la varianza della popolazione, effettuo un t test.

VARIABILE PESO:

Ipotesi Nulla (H0): La media del peso dei neonati nel campione è uguale alla media della popolazione.

Ipotesi Alternativa (H1): La media del peso dei neonati nel campione è significativamente diversa dalla media della popolazione.

H0: μ = 3.3kg

H1: μ != 3.3kg

```
#t test

t.test(Peso, mu = 3300, conf.level = 0.95, alternative = "two.sided")

##
##  One Sample t-test
##
## data:  Peso
## t = -1.505, df = 2497, p-value = 0.1324
## alternative hypothesis: true mean is not equal to 3300
## 95 percent confidence interval:
##  3263.577 3304.791
## sample estimates:
## mean of x
##  3284.184
```

Il t test mi fornisce un p-value di 0.13, ovvero maggiore di 0.05. Questo vuol dire che non rifiuto l'ipotesi nulla H0 e che la differenza tra la media del peso del campione (3,2kg) e della popolazione (3.3kg) non è significativa.

VARIABILE LUNGHEZZA

Ipotesi Nulla (H0): La media della lunghezza in mm dei neonati nel campione è uguale a quella della popolazione.

Ipotesi Alternativa (H1): La media della lunghezza in mm dei neonati nel campione è significativamente diversa da quella della popolazione

H0: μ = 50cm

H1: μ != 50cm

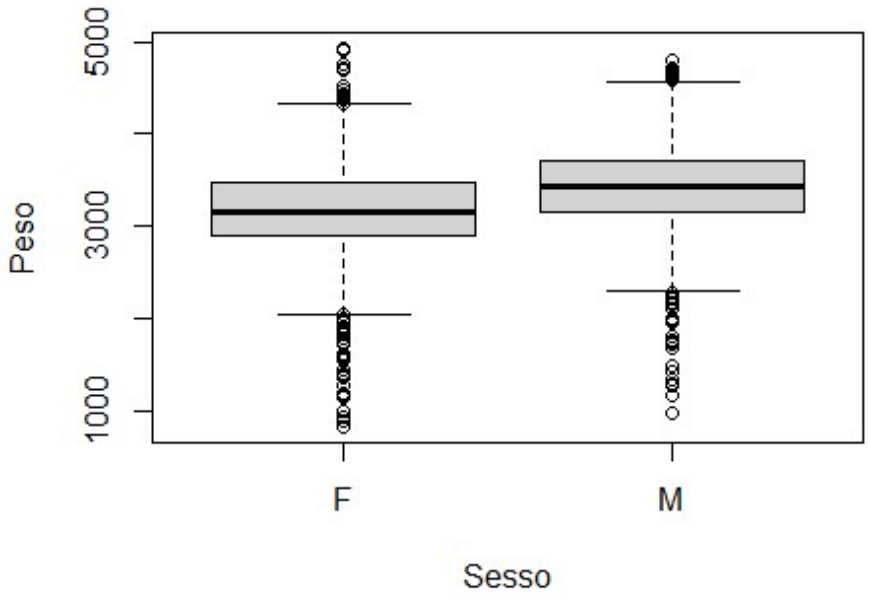
```
#t test

t.test (Lunghezza, mu = 500, conf.level = 0.95, alternative = "two.sided")

##
##  One Sample t-test
##
## data:  Lunghezza
## t = -10.069, df = 2497, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
##  493.6628 495.7287
## sample estimates:
## mean of x
##  494.6958
```

il t test ci fornisce un p-value molto piccolo portandoci a rifiutare l'ipotesi nulla di uguaglianza. Nonostante il test ci indichi una differenza significativa tra media del campione e della popolazione, ai fini dello studio questa differenza potrebbe non essere rilevante.

5) Per le stesse variabili, o per altre per le quali ha senso farlo, verifica differenze significative tra i due sessi

Variabile PESO:

```
#verifico graficamente eventuali differenze con boxplot
boxplot(Peso~Sesso)
```



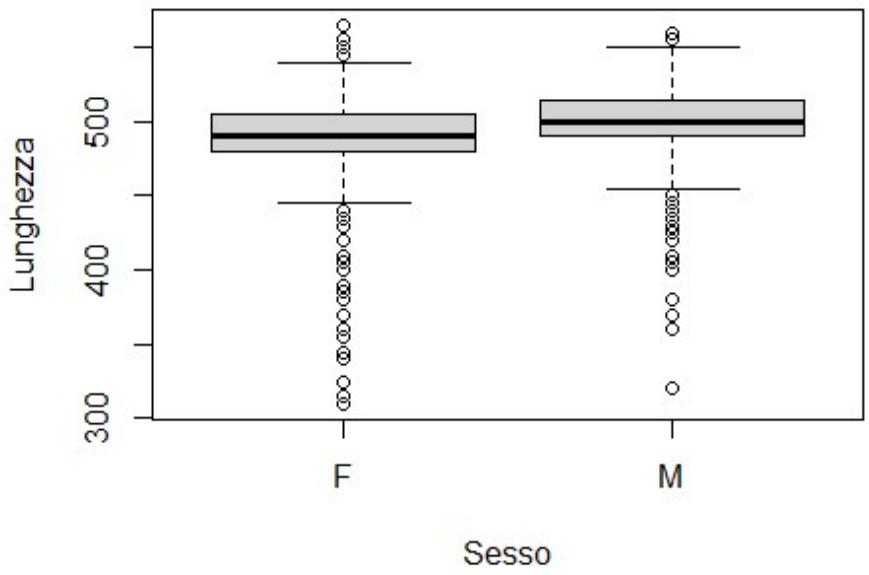

```
#verifico assunzione di normalità
shapiro.test(Peso)

##
##  Shapiro-Wilk normality test
##
## data:  Peso
## W = 0.97068, p-value < 2.2e-16

#rifiuto ipotesi di normalità ed effettuo wilcox test

wilcox.test(Peso ~ Sesso, data = dati)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Peso by Sesso
```

```
## W = 537495, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0

Peso_medio_F=mean(Peso[Sesso == "F"])
Peso_medio_M=mean(Peso[Sesso == "M"])

Peso_medio_F

## [1] 3161.061

Peso_medio_M

## [1] 3408.496
```

Il p-value ottenuto è inferiore a 0.05 indicandoci una differenza significativa nella media del peso tra femmine e maschi (3161 vs 3408 rispettivamente)

Variabile LUNGHEZZA

```
#verifico graficamente eventuali differenze con boxplot
boxplot(Lunghezza~Sesso)
```



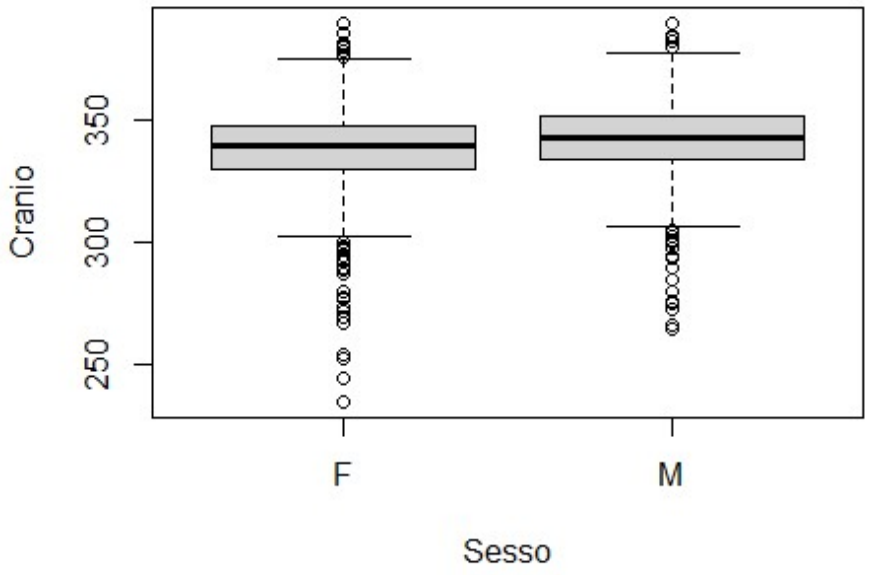

```
#verifico assunzione di normalità
shapiro.test(Lunghezza)

##
##  Shapiro-Wilk normality test
##
## data:  Lunghezza
## W = 0.90944, p-value < 2.2e-16
```

```r
#rifiuto normalità ed effettuo wilcox test

wilcox.test(Peso ~ Sesso, data = dati)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Peso by Sesso
## W = 537495, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0

Lungh_media_F=mean(Lunghezza[Sesso == "F"])
Lungh_media_M=mean(Lunghezza[Sesso == "M"])

Lungh_media_F

## [1] 489.7641

Lungh_media_M

## [1] 499.675
```

Anchein questo caso il p-value ottenuto è molto piccolo indicando una differenza significativa nella media della lunghezza tra i due gruppi (M e F).

Variabile CRANIO

```r
#verifico graficamente eventuali differenze con boxplot
boxplot(Cranio~Sesso)
```



```r
#verifico assunzione di normalità
shapiro.test(Cranio)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cranio
## W = 0.96358, p-value < 2.2e-16
```

```r
#rifiuto normalità ed effettuo wilcox test

wilcox.test(Cranio ~ Sesso, data = dati)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Cranio by Sesso
## W = 639872, p-value = 7.141e-15
## alternative hypothesis: true location shift is not equal to 0
```

```r
Cranio_F=mean(Cranio[Sesso == "F"])
Cranio_M=mean(Cranio[Sesso == "M"])

Cranio_F
```

```
## [1] 337.6231
```

```r
Cranio_M
```

```
## [1] 342.4586
```

il t test ci fornisce un p-value molto piccolo. Vi è una differenza significativa nella media del diametro del cranio tra i due gruppi (M e F).

6) Si vocifera che in alcuni ospedali si facciano più parti cesarei, sai verificare questa ipotesi?

Avendo due variabili qualitative (ospedale e tipo di parto), utilizzo il test del chi quadro per valutare se vi è un'associazione tra le due variabili.

Ho: indipendenza

H1: dipendenza

```r
#costruisco tabella di contingenza

tab_contingenza= table(Ospedale, Tipo.parto)
tab_contingenza
```

```
##         Tipo.parto
## Ospedale Ces Nat
##     osp1 242 574
##     osp2 254 594
##     osp3 232 602
```

```
# Test del chi-quadro
chisq.test(tab_contingenza)

##
##  Pearson's Chi-squared test
##
## data:  tab_contingenza
## X-squared = 1.083, df = 2, p-value = 0.5819
```

il test ci fornisce un p-value è 0.5819, che è superiore a 0.05. Quindi non rifiuto l'ipotesi nulla. In altre parole, non ci sono differenze significative nella distribuzione dei tipi di parto tra gli ospedali.

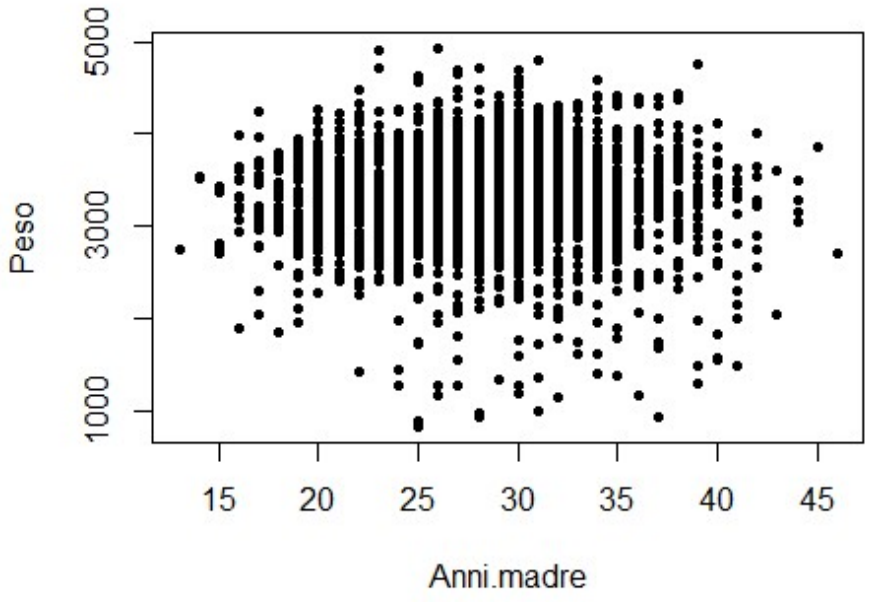
Analisi multidimensionale:

1) **Ricordati qual è l'obbiettivo dello studio e indaga le relazioni a due a due, soprattutto con la variabile risposta**

La variabile risposta dello studio è il peso del neonato. Valuto le relazioni di questa variabile con quelle relative alle madri (anni, n.gravidanze,fumatrici, gestazione) tramite il coeff di correlazione.

Peso vs Anni.madre

```
plot (Anni.madre, Peso, pch=20)
```

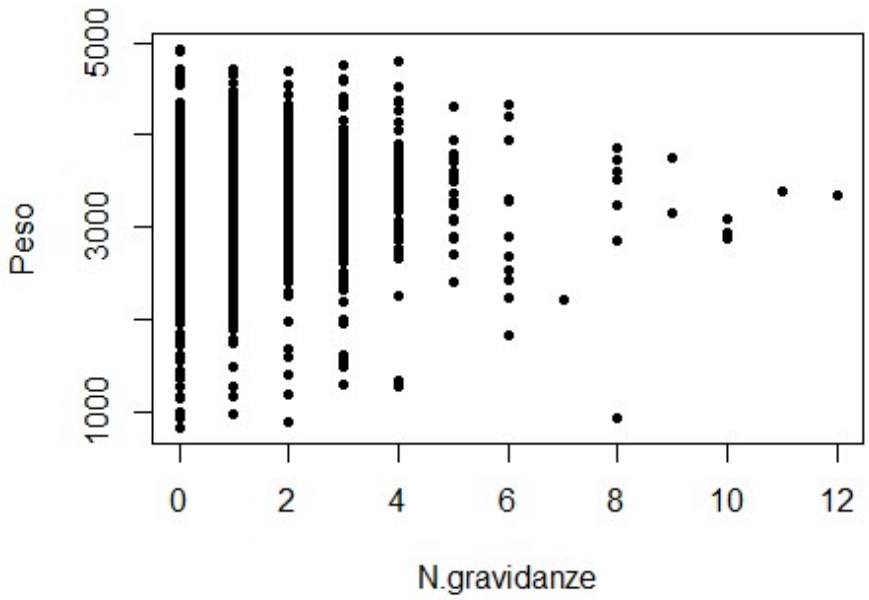



```
cor(Anni.madre, Peso)

## [1] -0.02378138
```

Il valore ottenuto è -0.024 ed essendo vicino a zero e negativo, suggerisce una correlazione lineare molto scarsa o addirittura inesistente tra le età delle madri e i pesi corrispondenti. Il grafico di dispersione mostra che non c'è una chiara relazione lineare tra le età delle madri e i pesi, e il coefficiente di correlazione conferma questa osservazione.

Peso vs N.gravidanze

```
plot (N.gravidanze, Peso, pch=20)
```



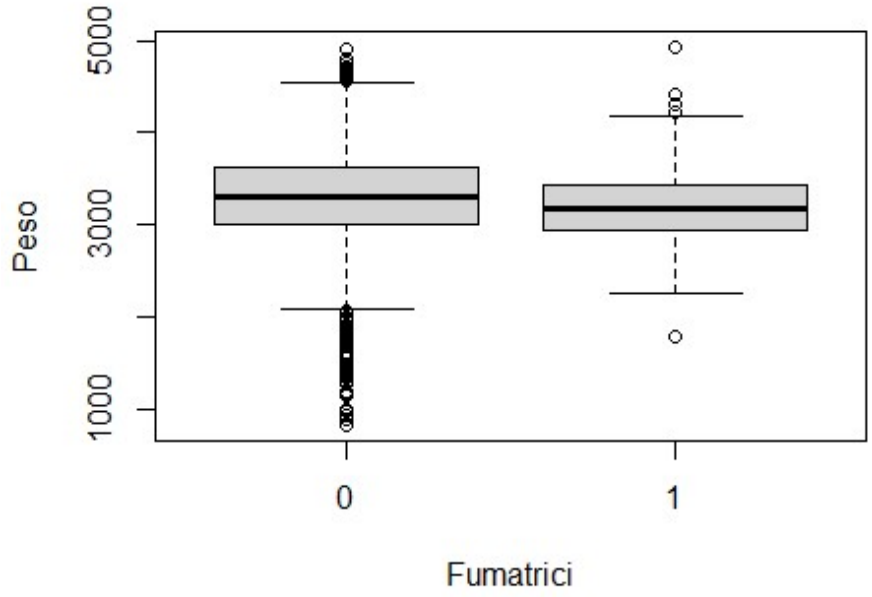

```
cor(N.gravidanze, Peso)
```

```
## [1] 0.002277118
```

Il coeff di correlazione, vicinissimo allo 0, indica che non c'è una correlazione lineare significativa tra queste due variabili. Il numero di gravidanze non sembra essere correlato al peso dei neonati nel nostro campione di dati.

Peso vs Fumatrici

```
#verifico graficamente eventuali differenze con boxplot
boxplot(Peso~Fumatrici)
```

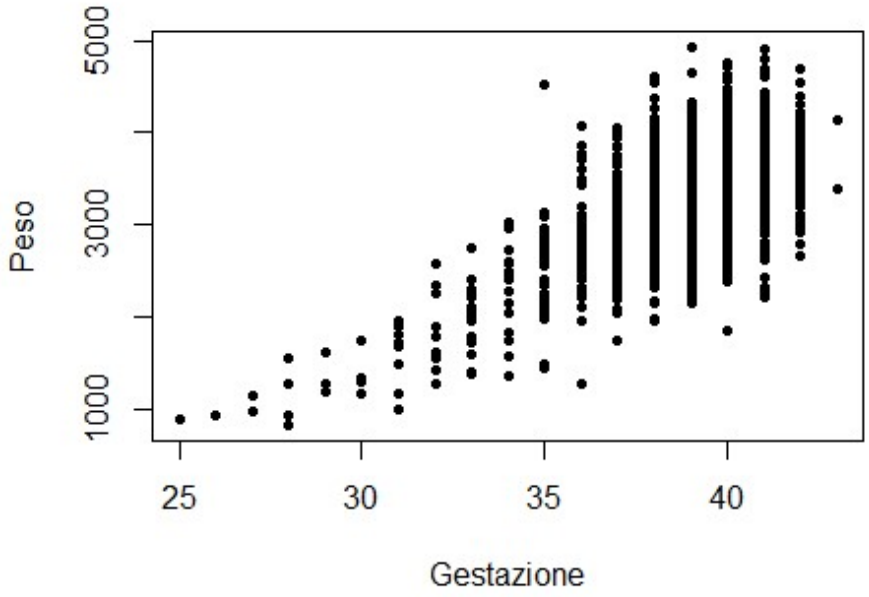



```
#correlazione
cor(Fumatrici, Peso)

## [1] -0.0189874
```

Il risultato della correlazione tra “Fumatrici” e “Peso” è -0.0189874, il che indica una correlazione molto debole e vicina a zero. Questo valore suggerisce che non c’è una relazione significativa tra il fatto di essere fumatrici e il peso del neonato

Peso vs Gestazione

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

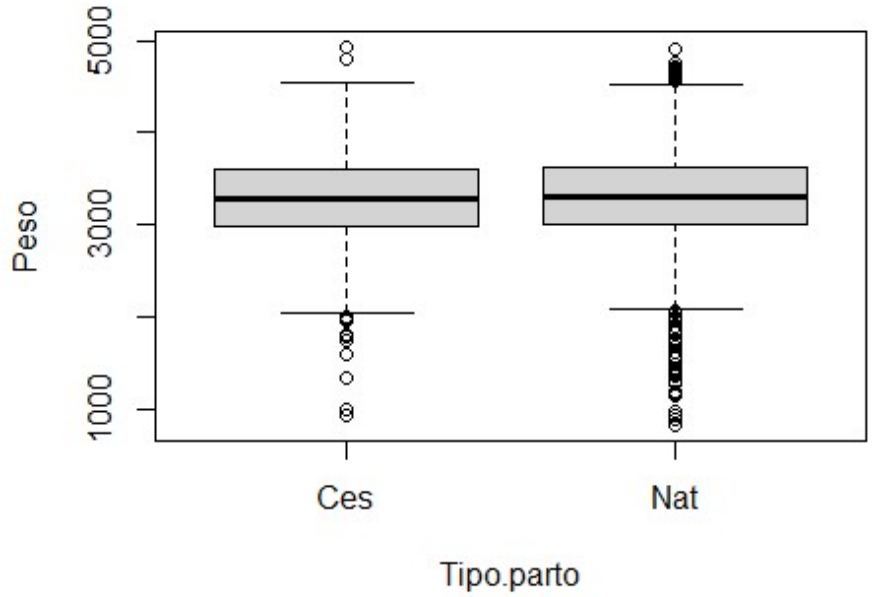



```
cor(Gestazione, Peso)

## [1] 0.5919592
```

Dal plot e dal coeff di correlazione possiamo dire che vi è una correlazione positiva tra il peso del neonato e il numero di settimane di gestazione della madre, indicando che all'aumentare delle settimane di gestazione, aumenta anche il peso del neonato.

Peso vs Tipo.parto

```
#verifico graficamente eventuali differenze con boxplot
boxplot(Peso~Tipo.parto)
```



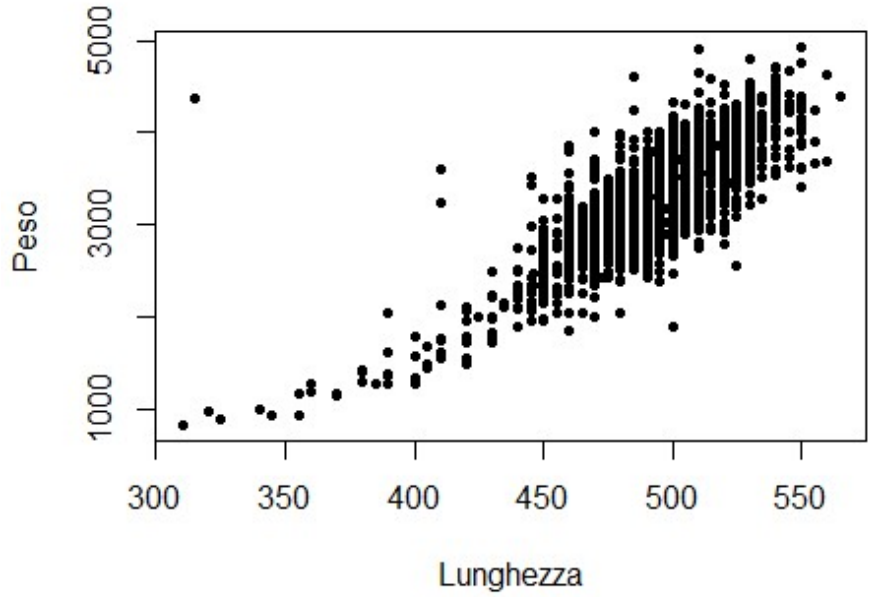

```
#correlazione
tipodiparto=ifelse(Tipo.parto=="Nat",1,0)
cor(tipodiparto, Peso)

## [1] 0.002610428
```

La correlazione tra il tipo di parto e il peso del bambino alla nascita è estremamente debole (positiva). Il tipo di parto quindi non sembra influenzare il peso del bambino.

Verifico le correlazioni con variabili come lunghezza e diametro cranio del neonato.

Peso vs Lunghezza

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



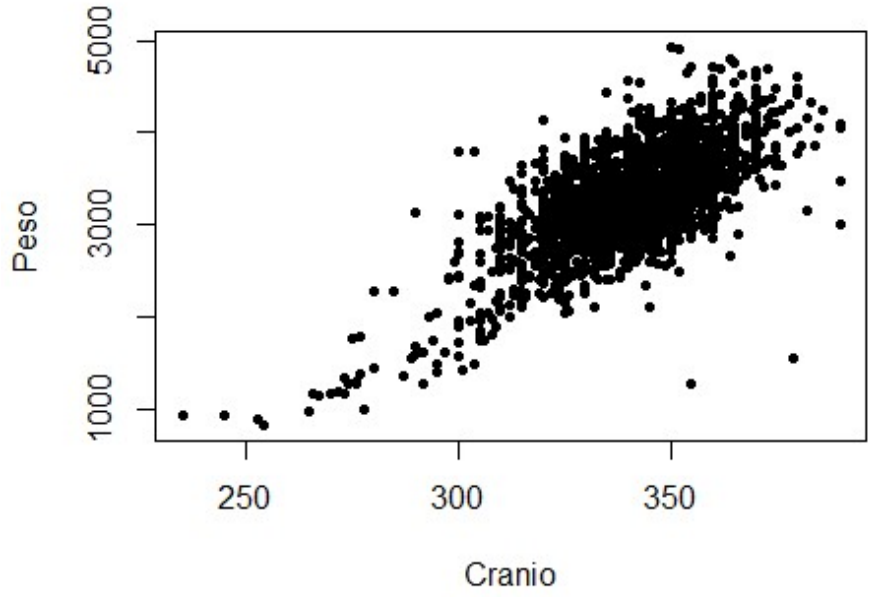

```
cor(Lunghezza, Peso)
```

```
## [1] 0.7960415
```

Come aspettato, tra il peso dei neonati e la lunghezza vi è una forte correlazione positiva.

Peso vs Cranio

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




```
cor(Cranio, Peso)

## [1] 0.7048438
```

Tra la variabile peso e cranio vi è una forte correlazione positiva. All'aumentare del peso, aumenta anche il diametro del cranio.

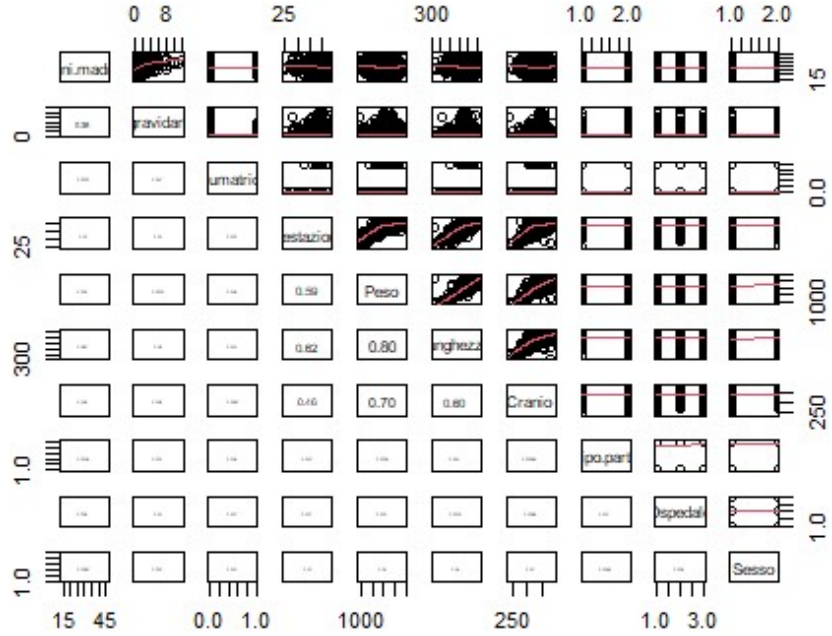
2) Crea un modello di regressione lineare multipla con tutte le variabili e commenta i coefficienti e il risultato ottenuto

Abbiamo visto in precedenza con il shapiro test che la variabile risposta Peso rifiuta l'ipotesi di normalità e che gli indici di forma indicano che la distribuzione è asimmetrica negativa e leptocurtica.

Visualizziamo graficamente la matrice di correlazione che mostra le correlazioni a due a due tra le variabili su intero dataset

```
?pairs

## starting httpd help server ... done

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)
```

```
}
#correlazioni

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



Le correlazioni maggiori che si riscontrano sono quelle tra peso vs lunghezza, peso vs cranio , peso vs gestazione.

Creo un modello di regressione lineare multipla con tutte le variabili:

```
mod1= lm(Peso~., data=dati)
summary(mod1)

##
## Call:
## lm(formula = Peso ~ ., data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1123.26  -181.53   -14.45   161.05  2611.89
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6735.7960   141.4790 -47.610  < 2e-16 ***
## Anni.madre        0.8018     1.1467   0.699   0.4845
## N.gravidanze     11.3812     4.6686   2.438   0.0148 *
## Fumatrici       -30.2741    27.5492  -1.099   0.2719
## Gestazione       32.5773     3.8208   8.526  < 2e-16 ***
```

```
## Lunghezza        10.2922     0.3009  34.207  < 2e-16 ***
## Cranio           10.4722     0.4263  24.567  < 2e-16 ***
## Tipo.partoNat    29.6335    12.0905   2.451   0.0143 *
## Ospedaleosp2    -11.0912    13.4471  -0.825   0.4096
## Ospedaleosp3     28.2495    13.5054   2.092   0.0366 *
## SessoM           77.5723    11.1865   6.934 5.18e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274 on 2487 degrees of freedom
## Multiple R-squared:  0.7289, Adjusted R-squared:  0.7278
## F-statistic: 668.7 on 10 and 2487 DF,  p-value: < 2.2e-16
```

Guardando i coefficienti, le variabili con un effetto significativo sono N.gravidanze, Gestazione,Lunghezza,Cranio, Tipo di parto, Ospedale e Sesso.

All'aumentare del numero di gravidanze, aumenta il peso del neonato di 11 gr. Ad ogni settimana di gestazione, si ha un incremento di 32,6gr del neonato. Per ogni mm di lunghezza in più. il peso del neonato aumenta di 10,3gr. per ogni mm di diametro del cranio in più, il peso aumento di 10,5 gr. Tenendo fisse le altre variabili, nel parto naturale si rileva un peso medio di 30g in più rispetto ai neonati nati da parto cesario. Nell'ospedale 3 sembra esserci un incremento di peso dei nascituri di 28g rispetto ai dati raccolti nell'osp1. Infine, nei neonati di sesso maschile si rileva un peso medio di 77g in più rispetto alle femmine.

l'R2 del modello 1 è 0,727.

3)Cerca il modello "migliore", utilizzando tutti i criteri di selezione che conosci e spiegali.

Proviamo a migliorare il modello escludendo le variabili non significative come gli anni della madre.

```
mod2=update(mod1, ~. -Anni.madre)
summary(mod2)

##
## Call:
## lm(formula = Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Ospedale + Sesso, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1113.82  -180.30   -16.22   160.66  2616.32
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6708.6189   136.0211 -49.320  < 2e-16 ***
## N.gravidanze     12.5833     4.3400   2.899  0.00377 **
## Fumatrici       -30.4268    27.5455  -1.105  0.26944
## Gestazione       32.2996     3.7997   8.501  < 2e-16 ***
## Lunghezza        10.2916     0.3008  34.209  < 2e-16 ***
## Cranio           10.4874     0.4257  24.638  < 2e-16 ***
```

```
## Tipo.partoNat    29.6654    12.0892   2.454  0.01420 *
## Ospedaleosp2    -10.9509    13.4442  -0.815  0.41541
## Ospedaleosp3     28.5171    13.4986   2.113  0.03474 *
## SessoM           77.6452    11.1849   6.942 4.91e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274 on 2488 degrees of freedom
## Multiple R-squared:  0.7288, Adjusted R-squared:  0.7279
## F-statistic: 743.1 on 9 and 2488 DF,  p-value: < 2.2e-16
```

Non riscontriamo grandi differenze rispetto al mod1 e nel pvalue . Facciamo un ulteriore verifica con Anova e BIC

```
anova(mod2,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
## Model 2: Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza
+
##     Cranio + Tipo.parto + Ospedale + Sesso
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   2488 186779904
## 2   2487 186743194  1     36710 0.4889 0.4845
```

```
BIC(mod1,mod2)
```

```
##      df      BIC
## mod1 12 35215.45
## mod2 11 35208.12
```

L'anova ci fornisce un Pr(>F) di 0,5 indicando che la differenza tra i modelli non è significativa e possiamo escludere la variabile Anni.madre dal modello. Il BIC è molto simile tra i due modello quindi procediamo con il migliorare il mod2.

Provo a creare un mod3 escludendo un'altra variabile non significativa come Fumatrici.

```
mod3=update(mod2, ~. -Fumatrici)
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1113.07  -181.71   -16.66   161.08  2619.57
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   -6707.9252   136.0257 -49.314  < 2e-16 ***
## N.gravidanze     12.3360     4.3344   2.846  0.00446 **
## Gestazione       32.0386     3.7925   8.448  < 2e-16 ***
## Lunghezza        10.3059     0.3006  34.286  < 2e-16 ***
## Cranio           10.4920     0.4257  24.648  < 2e-16 ***
## Tipo.partoNat    29.4080    12.0875   2.433  0.01505 *
## Ospedaleosp2    -10.8939    13.4447  -0.810  0.41786
## Ospedaleosp3     28.7917    13.4969   2.133  0.03301 *
## SessoM           77.4657    11.1842   6.926 5.48e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274 on 2489 degrees of freedom
## Multiple R-squared:  0.7287, Adjusted R-squared:  0.7278
## F-statistic: 835.7 on 8 and 2489 DF,  p-value: < 2.2e-16
```

Anche in questo caso non sembrano esserci grosse differenze con il modello precedente. valuto Anova, BIC

```
anova(mod3,mod2)

## Analysis of Variance Table
##
## Model 1: Peso ~ N.gravidanze + 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 186871503
## 2   2488 186779904  1     91599 1.2201 0.2694

BIC(mod1,mod2,mod3)

##      df      BIC
## mod1 12 35215.45
## mod2 11 35208.12
## mod3 10 35201.52
```

L'anova ci fornisce un Pr(>F) di 0,3 indicando che la differenza tra i modelli non è significativa e possiamo escludere la variabile Fumatrice dal modello. Il BIC nel mod3 è ancora più basso del modello 2. Procediamo con il modello 4 escludendo la variabile Ospedale.

```
mod4=update(mod3, ~. -Ospedale)
summary(mod4)

##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Sesso, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1129.14  -181.97   -16.26   160.95  2638.18
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6708.0171   136.0715 -49.298  < 2e-16 ***
## N.gravidanze     12.7356     4.3385   2.935  0.00336 **
## Gestazione       32.3253     3.7969   8.514  < 2e-16 ***
## Lunghezza        10.2833     0.3009  34.177  < 2e-16 ***
## Cranio           10.5063     0.4263  24.648  < 2e-16 ***
## Tipo.partoNat    30.1601    12.1027   2.492  0.01277 *
## SessoM           77.9171    11.1994   6.957 4.42e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.4 on 2491 degrees of freedom
## Multiple R-squared:  0.7277, Adjusted R-squared:  0.727
## F-statistic:  1109 on 6 and 2491 DF,  p-value: < 2.2e-16
```

```
BIC(mod3,mod4)
```

```
##      df      BIC
## mod3 10 35201.52
## mod4  8 35195.25
```

Sulla base del BIC, preferiamo ancora il mod4.

Infine escludiamo anche la variabile Tipo di parto e facciamo anche un test dell'Anova e Vif per valutare multicollinearità

```
mod5=update(mod4, ~. -Tipo.parto)
summary(mod5)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.37  -180.98   -15.57   163.69  2639.09
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6681.7251   135.8036 -49.201  < 2e-16 ***
## N.gravidanze    12.4554     4.3416   2.869  0.00415 **
## Gestazione      32.3827     3.8008   8.520  < 2e-16 ***
## Lunghezza       10.2455     0.3008  34.059  < 2e-16 ***
## Cranio          10.5410     0.4265  24.717  < 2e-16 ***
## SessoM          77.9807    11.2111   6.956 4.47e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.7 on 2492 degrees of freedom
```

```
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.7265
## F-statistic:  1327 on 5 and 2492 DF,  p-value: < 2.2e-16
```

```
BIC(mod4,mod5)
```

```
##      df      BIC
## mod4  8 35195.25
## mod5  7 35193.65
```

```
anova(mod5,mod4)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
## Model 2: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Sesso
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1   2492 188042054
## 2   2491 187574428  1    467626 6.2101 0.01277 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::vif(mod5)
```

```
## N.gravidanze   Gestazione    Lunghezza       Cranio        Sesso
##     1.023462     1.669779     2.075747     1.624568     1.040184
```

dal summary non notiamo grosse differenze e il BIC e l'anova ci dicono che il modello 5 è il migliore. Nel vif tutte le variabili hanno valori inferiori a 5 quidni si esclude multicollinearità

Utilizzo pacchetto MASS per confermare modello migliore

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("MASS")
```

```
library (MASS)
```

```
model=MASS::stepAIC(mod1,
                    direction="both",


summary(model)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.37  -180.98   -15.57   163.69  2639.09
##
## Coefficients:
```

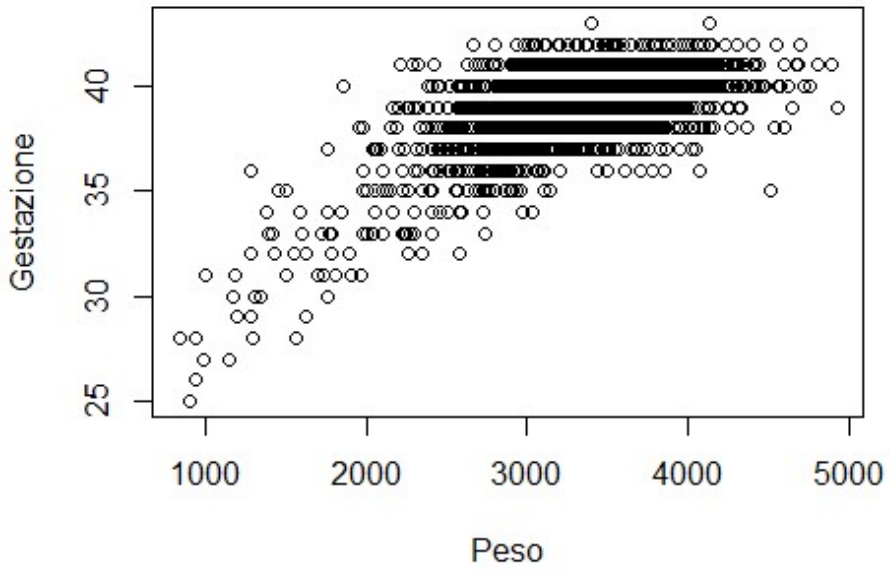
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6681.7251   135.8036 -49.201  < 2e-16 ***
## N.gravidanze    12.4554     4.3416   2.869  0.00415 **
## Gestazione      32.3827     3.8008   8.520  < 2e-16 ***
## Lunghezza       10.2455     0.3008  34.059  < 2e-16 ***
## Cranio          10.5410     0.4265  24.717  < 2e-16 ***
## SessoM          77.9807    11.2111   6.956 4.47e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.7 on 2492 degrees of freedom
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.7265
## F-statistic:  1327 on 5 and 2492 DF,  p-value: < 2.2e-16
```

Il pacchetto mass ci conferma che il modello migliore è il 5.

4) Si potrebbero considerare interazioni o effetti non lineari?

Guardando il grafico ottenuto con Pairs le relazioni non lineari sono osservabili tra Peso vs Gestazione. Valuto l'effetto quadratico di Gestazione.

```
plot(Gestazione ~Peso)
```




```
mod6=update(mod5,~. + I(Gestazione^2))
summary(mod6)

##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
```

```
##     Sesso + I(Gestazione^2), data = dati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1144.0  -181.5   -12.9   165.8  2661.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4646.7158   898.6322  -5.171 2.52e-07 ***
## N.gravidanze       12.5489     4.3381   2.893  0.00385 **
## Gestazione        -81.2309    49.7402  -1.633  0.10257
## Lunghezza          10.3502     0.3040  34.045  < 2e-16 ***
## Cranio             10.6376     0.4282  24.843  < 2e-16 ***
## SessoM             75.7563    11.2435   6.738 1.99e-11 ***
## I(Gestazione^2)     1.5168     0.6621   2.291  0.02206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.5 on 2491 degrees of freedom
## Multiple R-squared:  0.7276, Adjusted R-squared:  0.7269
## F-statistic:  1109 on 6 and 2491 DF,  p-value: < 2.2e-16
```

```
BIC(mod5,mod6)
```

```
##      df      BIC
## mod5  7 35193.65
## mod6  8 35196.21
```

Nel mod6 la variabile Gestazione perde di significatività e il suo coefficiente è molto diverso da quello del mod5 in quanto assume addirittura valore negativo indicando una relazione inversa con la variabile Peso che in questo caso non ha senso. Inoltre, la variabile quadratica è comunque meno significativa di quella non quadratica del modello 5.Pertanto, come suggerito anche dal BIC più basso, si continua a preferire il modello 5. Valutiamo anche possibili interazioni delle variabili come Gestazione e Cranio. Valuto se vi è un effetto sinergico sulla variabile risposta creando il mod7.

```
mod7=update(mod5, ~. + (Gestazione*Cranio))
summary(mod7)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso + Gestazione:Cranio, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1137.04  -181.47   -12.19   167.45  2695.30
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -187.95215 1106.93645  -0.170  0.86519
## N.gravidanze        13.12748    4.31382   3.043  0.00237 **
```

```
## Gestazione        -140.78001   29.53978  -4.766 1.99e-06 ***
## Lunghezza           10.46687    0.30113  34.759  < 2e-16 ***
## Cranio              -9.85430    3.47659  -2.834  0.00463 **
## SessoM              72.00219   11.18136   6.439 1.43e-10 ***
## Gestazione:Cranio    0.53389    0.09033   5.910 3.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.8 on 2491 degrees of freedom
## Multiple R-squared:  0.7308, Adjusted R-squared:  0.7301
## F-statistic:  1127 on 6 and 2491 DF,  p-value: < 2.2e-16
```

L'effetto sinergico Gestazione:cranio sembra essere significativo e anche il p value sembra essere migliorato. Tuttavia le singole variabili Gestazione e cranio assumono un coeff negativo indicando una relazione inversa con la variabile Peso ma non ha alcun senso. Infatti ci aspettiamo che all'aumentare delle settimane di gestazione aumenti anche il peso del bambino, così come all'aumentare del diametro del crano. Pertando continuiamo a preferire il mod5. Proviamo a fare la stessa cosa per valutare sinergia tra Gestazione e lunghezza.

```
mod8=update(mod5, ~. + (Gestazione*Lunghezza) )
summary(mod8)
```

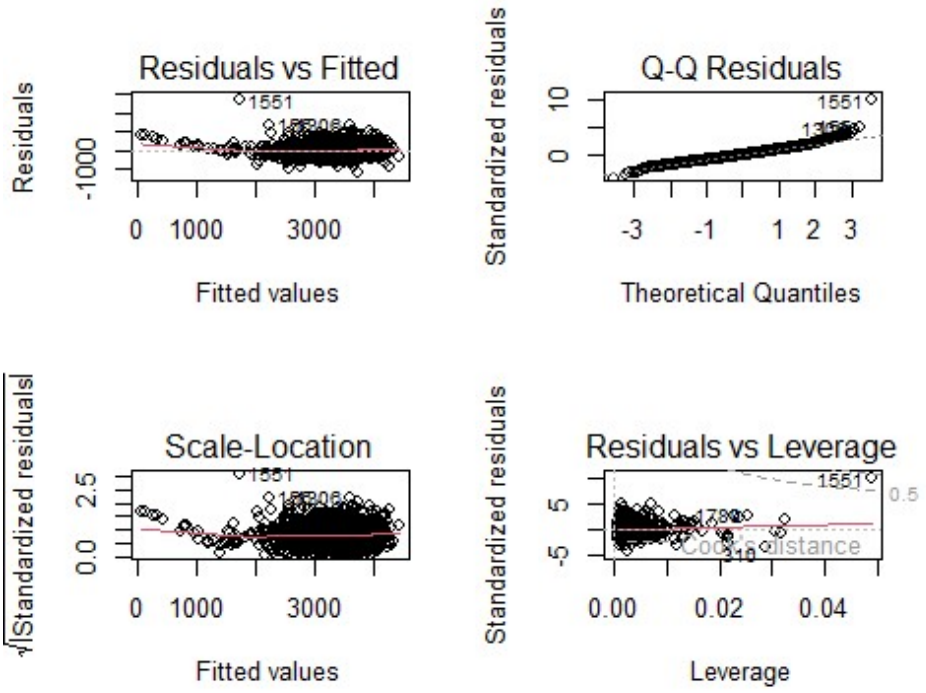
```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso + Gestazione:Lunghezza, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1133.41  -179.98   -11.52   168.93  2652.65
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.991e+03  9.206e+02  -2.163 0.030631 *
## N.gravidanze          1.303e+01  4.321e+00   3.015 0.002594 **
## Gestazione           -9.391e+01  2.481e+01  -3.785 0.000157 ***
## Lunghezza            -8.476e-02  2.028e+00  -0.042 0.966661
## Cranio                1.076e+01  4.264e-01  25.234  < 2e-16 ***
## SessoM                7.225e+01  1.121e+01   6.445 1.38e-10 ***
## Gestazione:Lunghezza  2.729e-01  5.298e-02   5.151 2.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.3 on 2491 degrees of freedom
## Multiple R-squared:  0.7299, Adjusted R-squared:  0.7292
## F-statistic:  1122 on 6 and 2491 DF,  p-value: < 2.2e-16
```

Questo modello non apporta nessuna miglioria rispetto al modello 5. la variabile Lunghezza perde la significatività e anche in questo caso le variabili assumono coeff negativi indicando una relazione inversa con la variabile Peso. Pertando lo rifiutiamo. il nostro modello definitivo è Mod5.

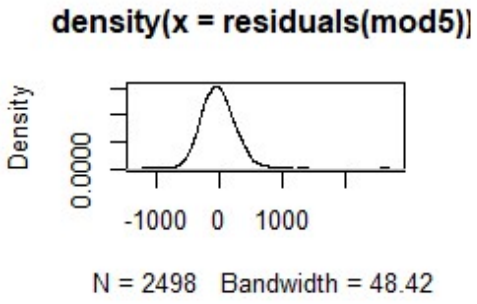
5) Effettua una diagnostica approfondita dei residui del modello e di potenziali valori influenti. Se ne trovi prova a verificare la loro effettiva influenza

divido la finestra grafica in 4 parti per visualizzare il modello.

```
par(mfrow=c(2,2))
plot(mod5)
```




```
#shapiro test per saggiare ipotesi di normalità

shapiro.test(residuals(mod5))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod5)
## W = 0.97414, p-value < 2.2e-16

#vediamo graficamente la distribuzione dei residui
plot(density(residuals(mod5)))
```



Otteniamo un valore molto piccolo quindi rifiutiamo l'ipotesi di normalità. Ad ogni modo graficamente la distribuzione somiglia a quella di una normale con una coda un po' più lunga a destra.

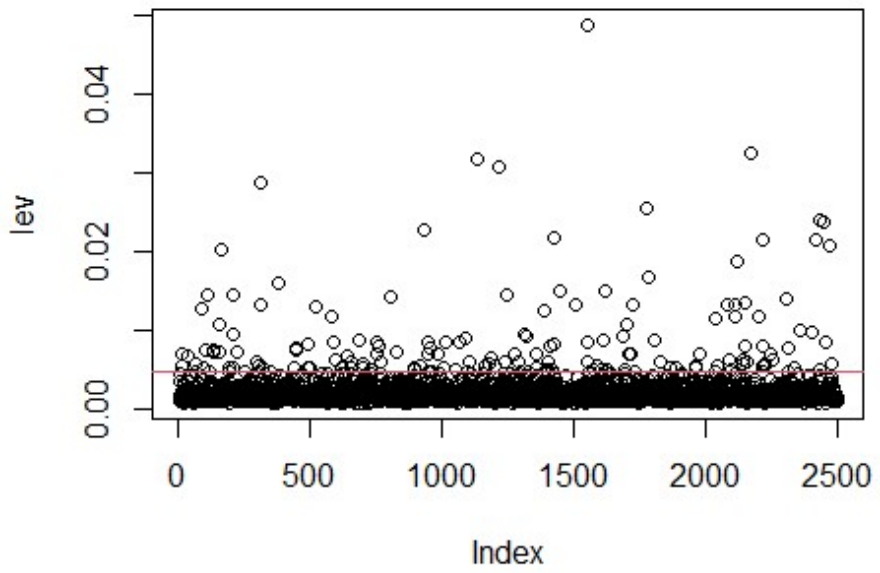
Saggiamo l'ipotesi di omoschedasticità ovvero di varianza costante e l'ipotesi di non correlazione dei residui utilizzando il test di Breusch-Pagan e Durbin-Watson rispettivamente

```
install.packages("lmtest")

library(lmtest)

bptest(mod5)

##
##  studentized Breusch-Pagan test
##
## data:  mod5
## BP = 90.297, df = 5, p-value < 2.2e-16

dwtest(mod5)

##
##  Durbin-Watson test
##
## data:  mod5
## DW = 1.9532, p-value = 0.1209
## alternative hypothesis: true autocorrelation is greater than 0
```

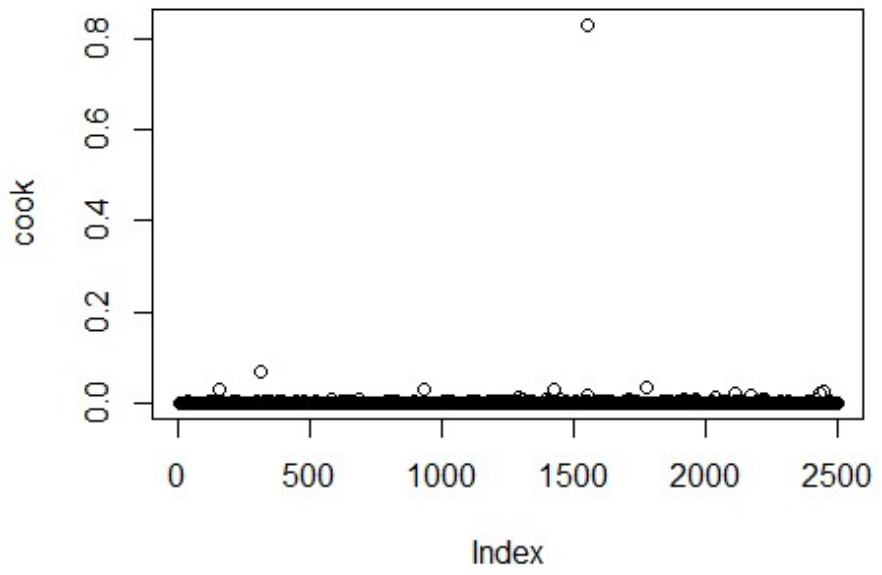
Rifiuto l'ipotesi di omoschedasticità, mentre non rifiuto l'ipotesi nulla per il test di Durbin-Watson. Il mio modello non rispetta già due assuzioni (normalità e omoschedasticità).

Valuto i leverage e gli outliers

```
#leverage

lev=hatvalues(mod5)
plot(lev)
p=sum(lev)
soglia=2*p/n
abline(h=soglia, col=2)
```




```
#outliers

car::outlierTest(mod5)

##       rstudent unadjusted p-value Bonferroni p
## 1551 10.046230         2.6345e-23   6.5810e-20
## 155   5.025345         5.3818e-07   1.3444e-03
## 1306  4.824963         1.4848e-06   3.7092e-03

#distanza di cook
cook=cooks.distance(mod5)
plot(cook)
```

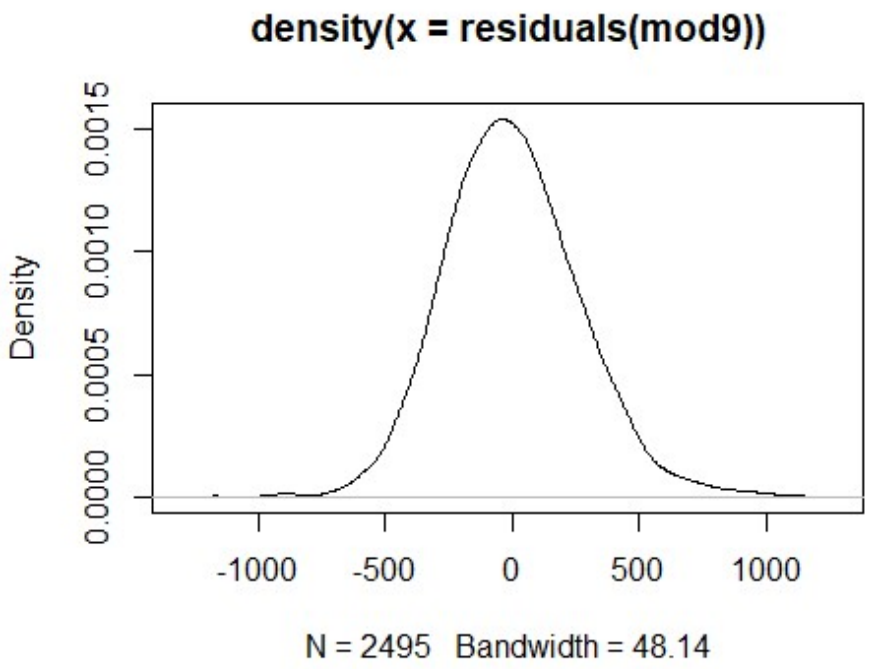



```
max_cook=max(cook)
max_cook

## [1] 0.8297645

osservazioni_influenti <- which(cook==max_cook)
osservazioni_influenti

## 1551
## 1549
```

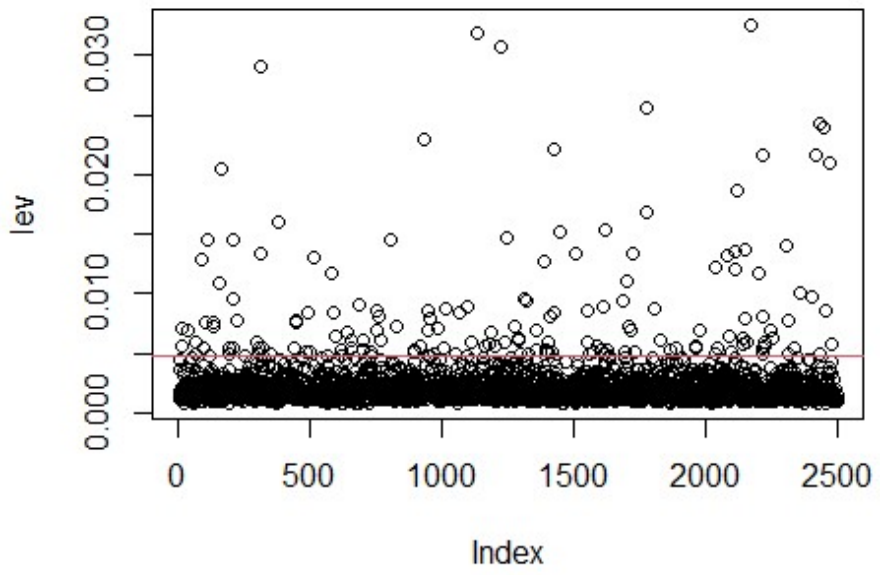
Abbiamo un bel po di leverage e 3 outliers. l'osservazione massima di cook supera la soglia di allarme quindi potrebbe in qualche modo avere una influenza significativa sui risultati del modello. Guardando i dati relativi agli outliers e alle osservazioni influenti di cook, vediamo che nella riga 1551 vi è un peso del neonati che supera alla 38esina settimana i 4kg ma la sua lunghezza è molto piccola. Anche osservando l'outlier alla riga 1306 vediamo che ci sono valori come il peso che è anomalo (quasi 5kg). Anche il campione 155 sembra avere un peso elevato rispetto alla lunghezza considerando la 36esima settimana di gestazione. Pertanto si decide di escludere questi valori dal dataset per vedere se il modello migliora.

```
dati_esclusi <- dati[rownames(dati) != "1551" & rownames(dati) != "1306"& rownames
(dati) != "155", ]

mod9 = lm(Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso, data = da
ti_esclusi)
summary(mod9)
```

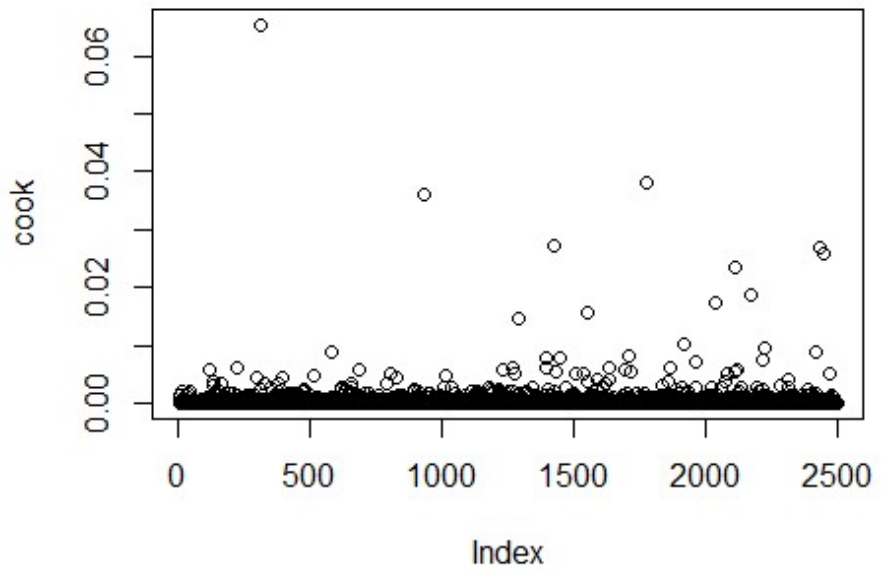
```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = dati_esclusi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1168.15  -178.63   -12.52   163.97  1134.68
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6696.0015   131.9461 -50.748  < 2e-16 ***
## N.gravidanze    13.9491     4.2164   3.308 0.000952 ***
## Gestazione      29.2575     3.6998   7.908 3.90e-15 ***
## Lunghezza       11.0084     0.2996  36.746  < 2e-16 ***
## Cranio           9.8211     0.4188  23.453  < 2e-16 ***
## SessoM          77.5066    10.8908   7.117 1.44e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.6 on 2489 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7411
## F-statistic:  1429 on 5 and 2489 DF,  p-value: < 2.2e-16

#verifico multicollinearità
car::vif(mod9)

## N.gravidanze   Gestazione    Lunghezza       Cranio        Sesso
##     1.023992     1.676572     2.134833     1.658923     1.040557

#shapiro test per saggiare ipotesi di normalità

shapiro.test(residuals(mod9))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod9)
## W = 0.9923, p-value = 3.11e-10

#vediamo graficamente la distribuzione dei residui
plot(density(residuals(mod9)))
```




```
library(lmtest)

bptest(mod9)

##
##  studentized Breusch-Pagan test
##
## data:  mod9
## BP = 11.445, df = 5, p-value = 0.04324

dwtest(mod9)

##
##  Durbin-Watson test
##
## data:  mod9
## DW = 1.9538, p-value = 0.124
## alternative hypothesis: true autocorrelation is greater than 0

#leverage

lev=hatvalues(mod9)
plot(lev)
p=sum(lev)
soglia=2*p/n
abline(h=soglia, col=2)
```




```
lev[lev>soglia]

#outliers

car::outlierTest(mod9)

##       rstudent unadjusted p-value Bonferroni p
## 1399 -4.402529         1.1151e-05     0.027823
## 1694  4.274359         1.9891e-05     0.049627

#distanza di cook
cook=cooks.distance(mod9)
plot(cook)
```




```
max_cook=max(cook)
index_max_cook <- which(cook == max_cook)
```

> index_max_cook

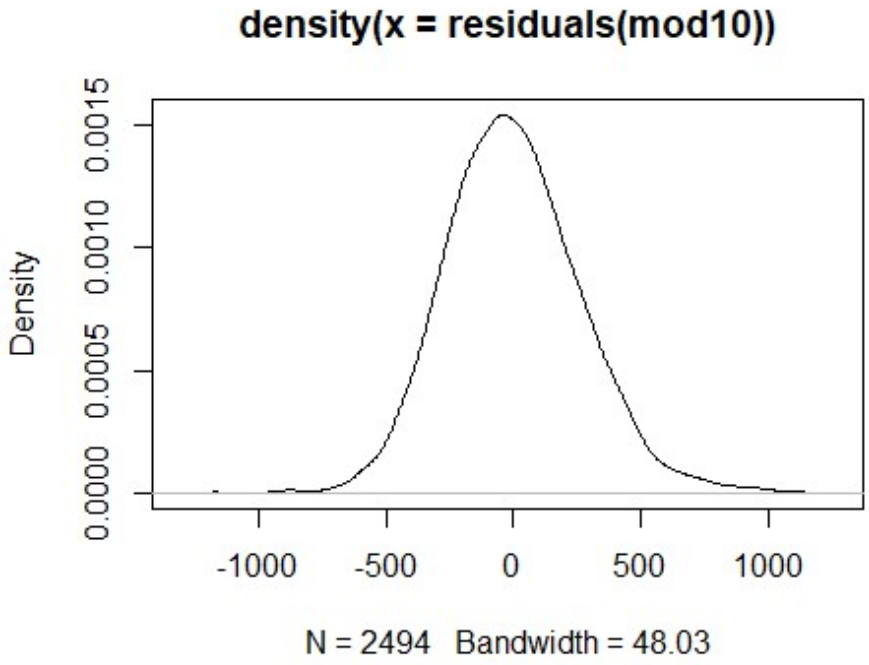
310

309

Nonostante R^2 migliori e non vi sia multicollinearità, le assunzioni non vengono ancora rispettate. Valutando la distanza di cook, il campione della riga 310 riporta un diametro del cranio sopra la media per le settimane di gestazione. Pertanto escludo anche questo campione e ripeto i test.

```
dati_esclusi2 <- dati[rownames(dati) != "1551" & rownames(dati) != "1306"& rowname
s(dati) != "155" & rownames(dati) != "310", ]

mod10 = lm(Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso, data = d
ati_esclusi2)
summary(mod10)

##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = dati_esclusi2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1170.39  -179.56   -12.89   162.23  1128.98
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6675.0772   131.7513 -50.664  < 2e-16 ***
## N.gravidanze    14.0575     4.2062   3.342 0.000844 ***
## Gestazione      27.6621     3.7169   7.442 1.36e-13 ***
## Lunghezza       10.9678     0.2991  36.675  < 2e-16 ***
## Cranio          10.0030     0.4207  23.775  < 2e-16 ***
## SessoM          77.0309    10.8651   7.090 1.74e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266 on 2488 degrees of freedom
## Multiple R-squared:  0.7419, Adjusted R-squared:  0.7414
## F-statistic:  1430 on 5 and 2488 DF,  p-value: < 2.2e-16

#verifico multicollinearità
car::vif(mod10)

## N.gravidanze   Gestazione    Lunghezza       Cranio        Sesso
##     1.023026     1.676825     2.130752     1.679041     1.040295

#shapiro test per saggiare ipotesi di normalità

shapiro.test(residuals(mod10))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod10)
## W = 0.9923, p-value = 3.105e-10

#vediamo graficamente la distribuzione dei residui
plot(density(residuals(mod10)))
```




```
library(lmtest)

bptest(mod10)

##
##  studentized Breusch-Pagan test
##
## data:  mod10
## BP = 6.5571, df = 5, p-value = 0.2557

dwtest(mod10)

##
##  Durbin-Watson test
##
## data:  mod10
## DW = 1.9576, p-value = 0.1448
## alternative hypothesis: true autocorrelation is greater than 0
```

Il mio modello rifiuta ancora le assunzioni di normalità ma non rifiuto l'ipotesi di omoschedasticità e di autocorrelazione di Durbin-Watson. Inoltre anche l'R quadro è migliorato rispetto ai modelli precedenti. Il vif per tutte le variabili è inferiore a 5 quindi non c'è rischio di multicollinearità nel modello.

6) Quanto ti sembra buono il modello per fare previsioni?

In generale, nonostante il modello non rispetti l'assunzione di normalità, tutti i test effettuati e l'R quadro mi fanno dire che sia un buon modello per le previsioni.

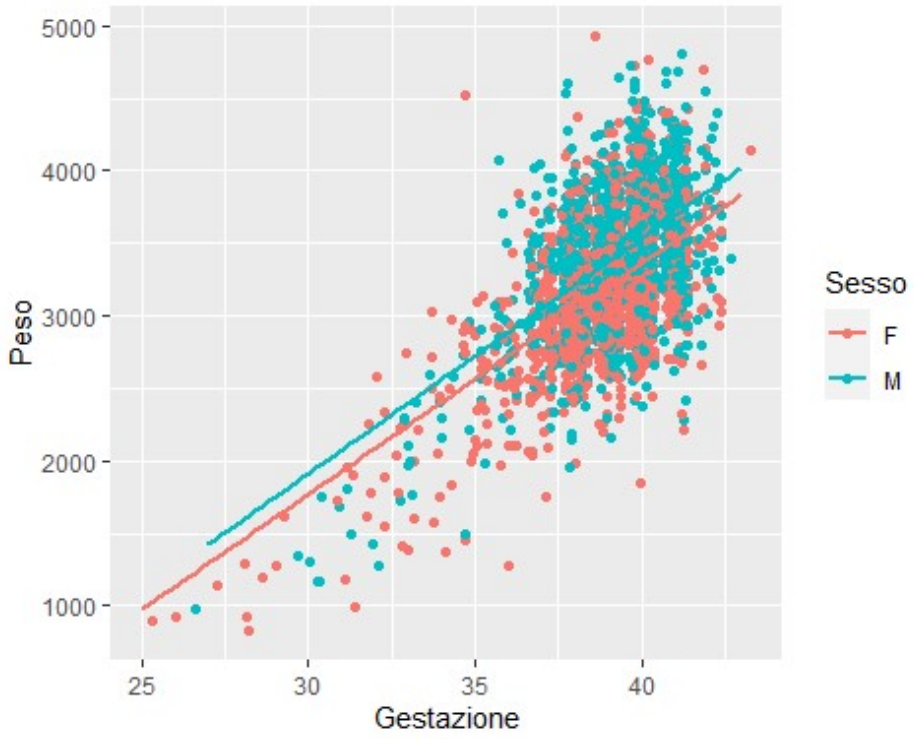
7) **Fai la tua migliore previsione per il peso di una neonata, considerato che la madre è alla terza gravidanza e partorirà alla 39esima settimana. Niente misure dall'ecografia.**

```
#utilizzando il mio modello per intero ottengo messaggio di errore: Error: variabl
es 'Lunghezza', 'Cranio' were specified with different types from the fit
:

predict(mod10, newdata=data.frame(N.gravidanze=3, Gestazione=39, Sesso = as.factor
("F"), Lunghezza = NA, Cranio = NA))



# Per risolvere, elimino dal mio modello le variabili di cui non ho info (lunghezz
a e Cranio) e faccio la previsione:

mod_prev <- lm(Peso ~ N.gravidanze + Gestazione + Sesso, data = dati_esclusi2)
Prediction=predict(mod_prev, newdata = data.frame(N.gravidanze = 3, Gestazione = 3
9, Sesso = as.factor("F")))

Prediction

##        1
## 3251.205
```

il peso previsto è di 3251 g, in linea con la media del peso alla 39esima settimana di gestazione per le neonate.

8) **Cerca di creare qualche rappresentazione grafica che aiuti a visualizzare il modello. Se è il caso semplifica quest'ultimo!**

```
library (ggplot2)

#Peso vs settimane di gestazione


grafico1=ggplot( data=dati_esclusi2)+
  geom_point(aes(x=Gestazione,
                 y=Peso,
                 col=Sesso),position="jitter")+
  geom_smooth(aes (x=Gestazione,
                   y=Peso,
                   col=Sesso), se=F, method="lm")


grafico1

## `geom_smooth()` using formula = 'y ~ x'
```



Il grafico permette di visualizzare la relazione tra Peso e settimane di gestazione nei due sessi F e M. Come vediamo le due rette hanno lo stesso andamento ma i maschi sembrano pesare leggermente di più rispetto alle femmine.

```
#visualizzo in forma 3d la relazione tra il Peso, la lunghezza e le settimane di g
estazione
library(rgl)

plot3D=with(dati_esclusi2, plot3d(Lunghezza, Gestazione, Peso, col=ifelse(Sesso ==
"M", "blue", "pink")))
legend3d("topright", legend=c("Maschio", "Femmina"), pch=20, col=c("blue", "pink")
)

plot3D
```

In questo grafico 3D possiamo visualizzare la relazione della variabile risposta Peso con altre 3 variabili: lunghezza, settimane di gestazione e Sesso. Dal grafico si nota come all'aumentare delle settimane di gestazione e della lunghezza aumenti anche il Peso del neonato.