PROGETTO STATISTICA INFERENZIALE

2024-02-01

```r
# PUNTO 1
setwd("C:\\Users\\Dell\\Desktop\\profession ai\\progetto Statistica
Inferenziale")
data=read.csv("neonati.csv")
nrow(data)
```

```
## [1] 2500
```

```r
head(data[1:5])
```

```
##   Anni.madre N.gravidanze Fumatrici Gestazione Peso
## 1         26            0         0         42 3380
## 2         21            2         0         39 3150
## 3         34            3         0         38 3640
## 4         28            1         0         41 3690
## 5         20            0         0         38 3700
## 6         32            0         0         40 3200
```

```r
head(data[6:10])
```

```
##   Lunghezza Cranio Tipo.parto Ospedale Sesso
## 1       490    325        Nat     osp3     M
## 2       490    345        Nat     osp1     F
## 3       500    375        Nat     osp2     M
## 4       515    365        Nat     osp2     M
## 5       480    335        Nat     osp3     F
## 6       495    340        Nat     osp2     F
```

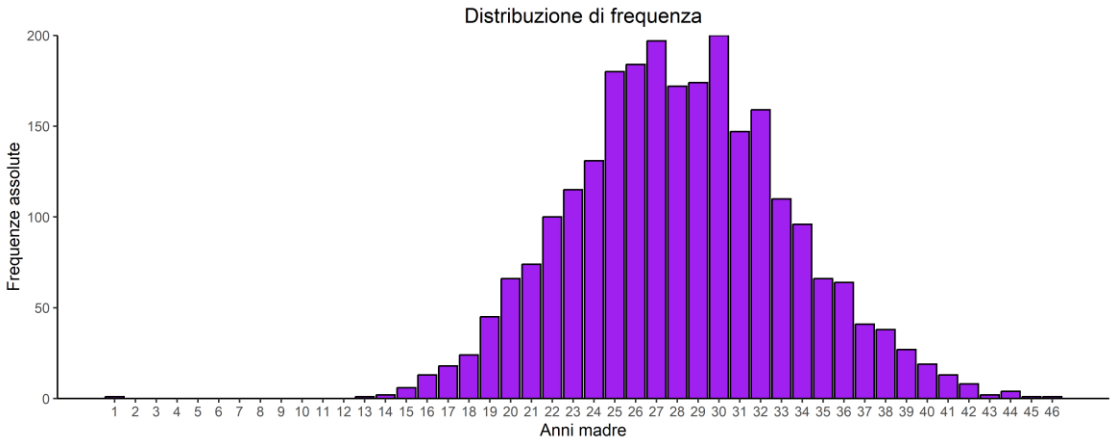
```r
data=na.omit(data) # rimuovo le osservazioni con valori mancanti
nrow(data) # nessuna osservazione presenta valori mancanti
```

```
## [1] 2500
```

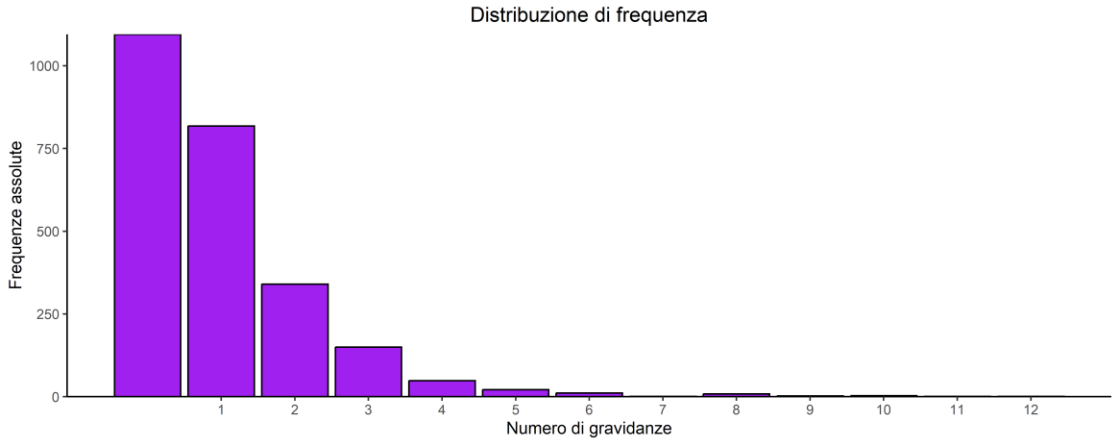
```r
# PUNTO 2
# Il dataset iniziale contiene 2500 ossevazioni di 10 variabili relative a
dei neonati.
# L'obiettivo dell'analisi statistica è trovare un modello per prevedere il
peso alla nascita dei neonati.
# Le variabili sono principalmente caratteristiche del neonato e della madre:
#
# età della madre -> quantitativa continua
# numero di gravidanze sostenute -> quantitativa discreta
# madre fumatrice (0=NO, 1=SI')-> qualitativa nominale dicotomica
# numero di settimane di gestazione -> quantitava continua
# tipo di parto (naturale o cesareo) -> qualitativa nomimale
```

```r
#
# peso in grammi del neonato -> quantitativa continua
# lunghezza in mm del neonato -> quantitativa continua
# diametro in mm del cranio del neonato -> quantitativa continua
# sesso del neonato -> qualitativa nominale dicotomica
#
# ospedale (1,2,3) -> qualitativa nominale

# PUNTO 3
attach(data)

data_quant=subset(data, select = -c(Fumatrici,Tipo.parto,Sesso,Ospedale))

attach(data_quant)

## I seguenti oggetti sono mascherati da data:
##
##     Anni.madre, Cranio, Gestazione, Lunghezza, N.gravidanze, Peso

summary(data_quant[1:3])

##    Anni.madre     N.gravidanze       Gestazione
##  Min.   : 0.00   Min.   : 0.0000   Min.   :25.00
##  1st Qu.:25.00   1st Qu.: 0.0000   1st Qu.:38.00
##  Median :28.00   Median : 1.0000   Median :39.00
##  Mean   :28.16   Mean   : 0.9812   Mean   :38.98
##  3rd Qu.:32.00   3rd Qu.: 1.0000   3rd Qu.:40.00
##  Max.   :46.00   Max.   :12.0000   Max.   :43.00

summary(data_quant[4:ncol(data_quant)])

##       Peso        Lunghezza         Cranio
##  Min.   : 830   Min.   :310.0   Min.   :235
##  1st Qu.:2990   1st Qu.:480.0   1st Qu.:330
##  Median :3300   Median :500.0   Median :340
##  Mean   :3284   Mean   :494.7   Mean   :340
##  3rd Qu.:3620   3rd Qu.:510.0   3rd Qu.:350
##  Max.   :4930   Max.   :565.0   Max.   :390

# osservo che il minimo di Anni.madre è 0, valore sicuramente errato

data[Anni.madre=="0",] # alla riga 1380 il dato Anni.madre è 0, per cui si
elimina l'osservazione

##      Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio
## 1380          0            0         0         39 3060       490    330
##      Tipo.parto Ospedale Sesso
## 1380        Nat     osp3     M

data=subset(data, Anni.madre != 0) # Costruisco la distribuzione di frequenza
per osservare i valori assunti dalla variabile
```

```r
library(ggplot2)
ggplot(data=data)+
  geom_bar(aes(x=Anni.madre),
           stat="count",
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Anni madre",
       y="Frequenze assolute")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(1,46,1))+
  scale_y_continuous(expand = c(0, 0))
```




```r
# Alcuni valori della variabile Anni.madre non sono plausibili; altri lo sono poco

ggplot(data=data)+
  geom_bar(aes(x=N.gravidanze),
           stat="count",
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Numero di gravidanze",
       y="Frequenze assolute")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(1,12,1))+
  scale_y_continuous(expand = c(0, 0))
```




```
# Alcuni valori per la variabile N.gravidanze sono poco plausibili

# https://www.istat.it/it/files/2023/10/Report-natalita-26-ottobre-2023.pdf
# Il numero medio di figli per donna in Italia nel 2022 era pari a 1,87 per
le donne straniere e 1,18 per le donne italiane: rimuoviamo le osservazioni
in cui le donne hanno più di 4 figli
# Il tasso di fecondità in Italia nel 2022 al di sotto dei 20 anni è molto
basso:
# rimuoviamo le osservazioni per cui l'età della madre è sotto 20; inoltre,
supponiamo
# che tutte le madri abbiano la prima gravidanza non prima dei 20 anni e
# non più di una gravidanza all'anno

data=subset(data, Anni.madre >= 20 & N.gravidanze<=4)
lista_indexes_to_remove=list()
for (i in 0:2){
  indexes_to_remove=which(data$Anni.madre==as.numeric(20)+i &
data$N.gravidanze>as.numeric(0)+i)
  lista_indexes_to_remove[[i+1]]=c(indexes_to_remove)
  if (length(c(indexes_to_remove))>0){
    data=data[-c(indexes_to_remove),]
  }
}
lista_indexes_to_remove

## [[1]]
##  [1]  134  364  513  529  746  876 1019 1116 1125 1269 1333 1372 1402 1407
1562
## [16] 1666 2104 2156 2290 2298 2323
##
## [[2]]
## [1]    2   12   15  791 1854
##
## [[3]]
## [1]  625 1624
```

```r
summary(data)
```

```
##    Anni.madre     N.gravidanze     Fumatrici         Gestazione
##  Min.   :20.00   Min.   :0.000   Min.   :0.00000   Min.   :25.00
##  1st Qu.:25.00   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:38.00
##  Median :28.00   Median :1.000   Median :0.00000   Median :39.00
##  Mean   :28.64   Mean   :0.898   Mean   :0.04237   Mean   :38.98
##  3rd Qu.:32.00   3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:40.00
##  Max.   :45.00   Max.   :4.000   Max.   :1.00000   Max.   :43.00
##       Peso         Lunghezza         Cranio       Tipo.parto
##  Min.   : 830   Min.   :310.0   Min.   :245.0   Length:2313
##  1st Qu.:3000   1st Qu.:480.0   1st Qu.:330.0   Class :character
##  Median :3300   Median :500.0   Median :340.0   Mode  :character
##  Mean   :3290   Mean   :494.9   Mean   :340.2
##  3rd Qu.:3620   3rd Qu.:510.0   3rd Qu.:350.0
##  Max.   :4930   Max.   :565.0   Max.   :390.0
##    Ospedale            Sesso
##  Length:2313        Length:2313
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```r
#
https://www.health.ny.gov/community/pregnancy/why_is_40_weeks_so_important.htm#:~:text=Pregnancy%20lasts%20for%20about%20280,between%2029%20and%2033%20weeks.
#
https://www.salute.gov.it/portale/donna/dettaglioContenutiDonna.jsp?area=Salute+donna&id=4478&menu=nascita
# Un neonato prematuro nasce indicativamente dopo un numero di settimanane di gestazione tra 23 e 37,
# per cui il minimo valore osservato (26) risulta plausibile.
# Un neonato post-termine nasce indicativamente dopo un numero di settimane superiore o uguale a 42,
# per cui il massimo valore osservato (43) risulta plausibile.

# https://media.tghn.org/articles/newbornsize.pdf
# L'articolo riportato mostra:
# Pesi del neonato tra circa 1 e 5,5 kg
# Lunghezze del neonato tra circa 38 e 56 cm
# Circonferenze del cranio del neonato tra circa 26 e 40 cm
# L'articolo riporta i dati di bambini da diversi stati, Italia compresa.
# Ipotizzando che anche il dataset della nostra analisi possa contenere osservazioni di bambini stranieri
# nati in Italia, consideriamo i nostri valori plausibili rispetto ai range forniti da questo altro studio:
# il nostro range non è sempre interno al range fornito dall'articolo, ma si ipotizza che i valori al di fuori siano corretti
```

```r
data_quant=subset(data, select = -c(Fumatrici,Tipo.parto,Sesso,Ospedale))
attach(data)
```

```
## I seguenti oggetti sono mascherati da data_quant:
##
##     Anni.madre, Cranio, Gestazione, Lunghezza, N.gravidanze, Peso
```

```
## I seguenti oggetti sono mascherati da data (pos = 5):
##
##     Anni.madre, Cranio, Fumatrici, Gestazione, Lunghezza, N.gravidanze,
##     Ospedale, Peso, Sesso, Tipo.parto
```

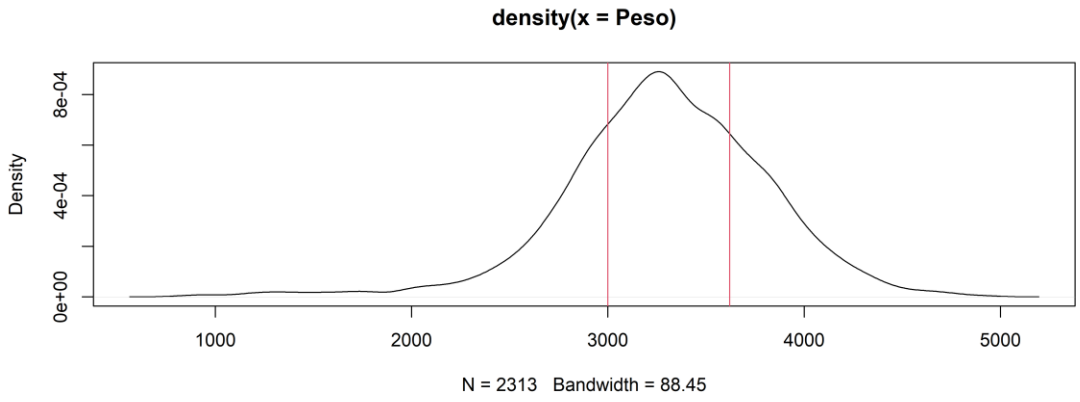
```r
summary(data_quant[1:3])
```

```
##    Anni.madre     N.gravidanze     Gestazione
##  Min.   :20.00   Min.   :0.000   Min.   :25.00
##  1st Qu.:25.00   1st Qu.:0.000   1st Qu.:38.00
##  Median :28.00   Median :1.000   Median :39.00
##  Mean   :28.64   Mean   :0.898   Mean   :38.98
##  3rd Qu.:32.00   3rd Qu.:1.000   3rd Qu.:40.00
##  Max.   :45.00   Max.   :4.000   Max.   :43.00
```

```r
summary(data_quant[4:ncol(data_quant)])
```

```
##       Peso        Lunghezza         Cranio
##  Min.   : 830   Min.   :310.0   Min.   :245.0
##  1st Qu.:3000   1st Qu.:480.0   1st Qu.:330.0
##  Median :3300   Median :500.0   Median :340.0
##  Mean   :3290   Mean   :494.9   Mean   :340.2
##  3rd Qu.:3620   3rd Qu.:510.0   3rd Qu.:350.0
##  Max.   :4930   Max.   :565.0   Max.   :390.0
```

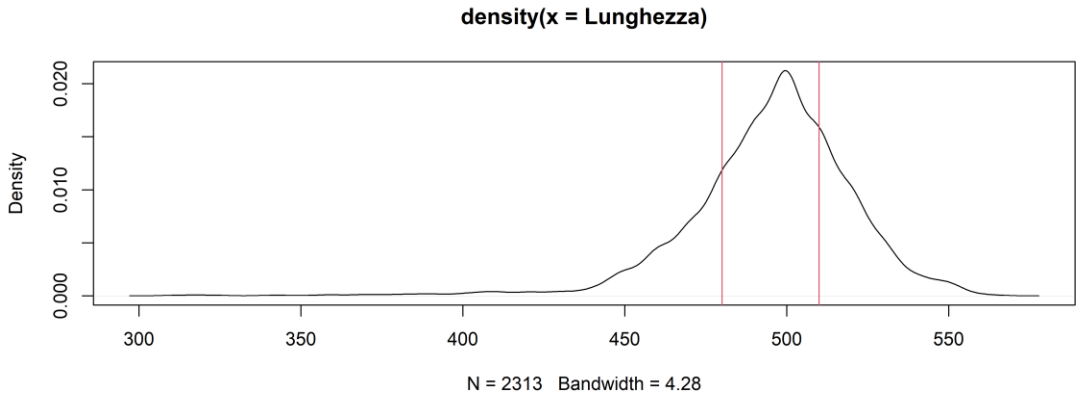
```r
plot(density(Peso)) # la variabile risposta (Peso) sembra distribuita secondo una normale con una coda sinistra più lunga
abline(v=quantile(Peso,probs=c(0.25,0.75)),
       col=2)
```



```r
quantile(Peso,probs=c(0.25,0.75)) # la metà del valori si colloca nel range 3000-3620 g
```

```
##  25%  75%
## 3000 3620
```

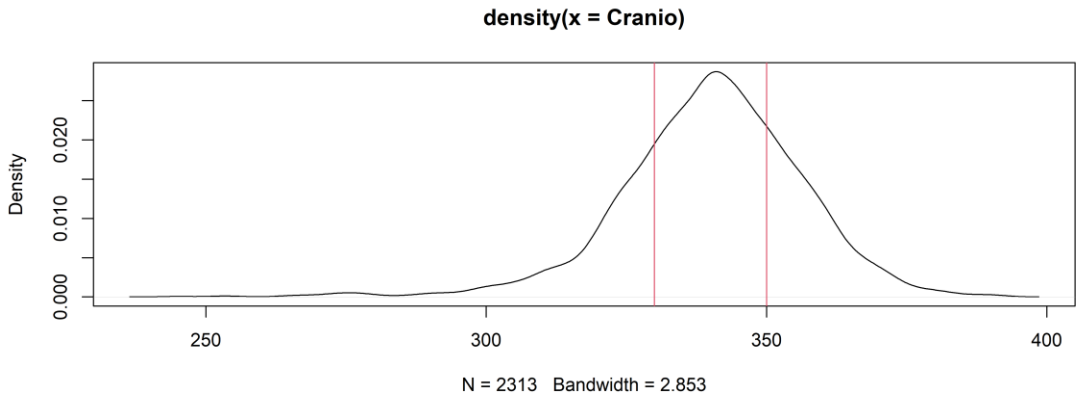
```r
plot(density(Lunghezza)) # la variabile Lunghezza sembra distribuita secondo una normale con una coda sinistra più lunga
abline(v=quantile(Lunghezza,probs=c(0.25,0.75))
       ,col=2)
```




```r
quantile(Lunghezza,probs=c(0.25,0.75)) # la metà del valori si colloca nel range 480-510 mm
```

```
## 25% 75%
## 480 510
```

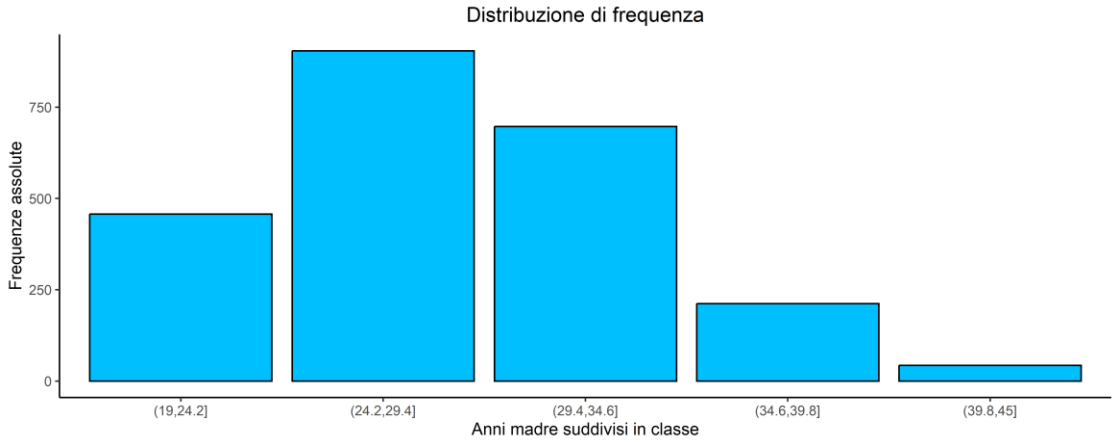
```r
plot(density(Cranio)) # la variabile Cranio sembra distribuita secondo una normale con una coda sinistra più lunga
abline(v=quantile(Cranio,probs=c(0.25,0.75)),
       col=2)
```




```r
quantile(Cranio,probs=c(0.25,0.75)) # la metà del valori si colloca nel range 330-350 mm
```

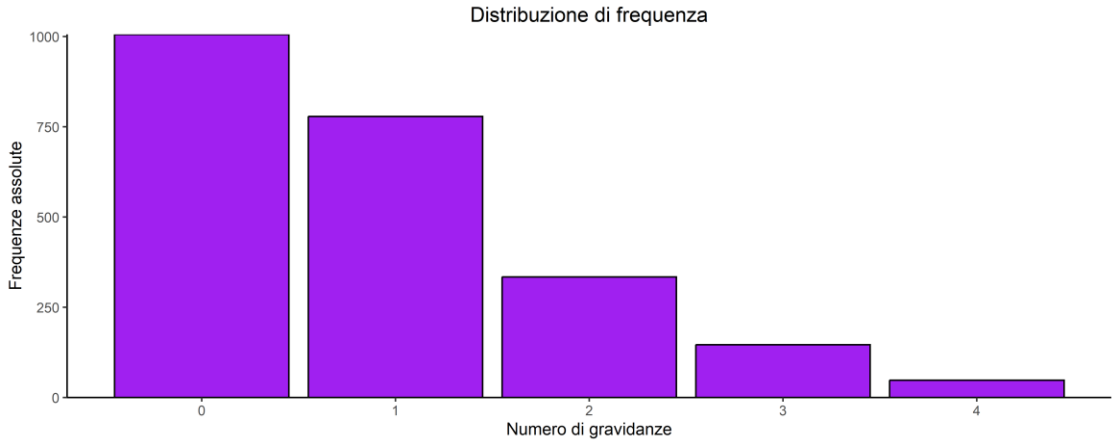
```
## 25% 75%
## 330 350
```

```
data_cl=data.frame(matrix(nrow =nrow(data), ncol = 0))
data_cl$Anni.madre.cl=cut(data$Anni.madre, breaks=seq(from=min(Anni.madre)-
1,to=max(Anni.madre),length.out=6))
ggplot(data=data_cl)+
  geom_bar(aes(x=Anni.madre.cl),
           stat="count",
           col="black",
           fill="deepskyblue")+
  labs(title="Distribuzione di frequenza",
       x="Anni madre suddivisi in classe",
       y="Frequenze assolute")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```

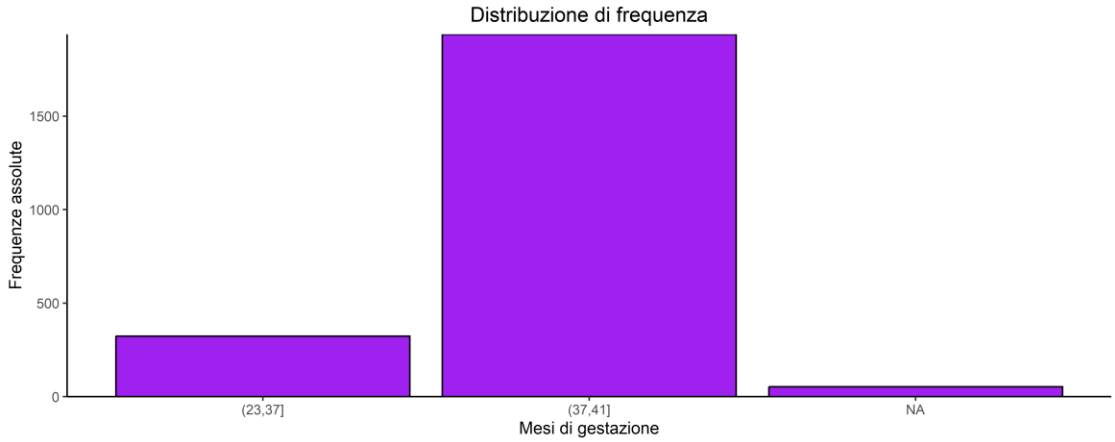



```
# La maggior parte delle madri hanno età compresa tra i 25 e i 30 anni, e,
con minor frequenza, nella fascia 30-35 anni

moments::skewness(Anni.madre)

## [1] 0.4253974

# La skewness positiva della variabile Anni.madre conferma che sono più
frequenti valori bassi

ggplot(data=data)+
  geom_bar(aes(x=N.gravidanze),
           stat="count",
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Numero di gravidanze",
       y="Frequenze assolute")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
```

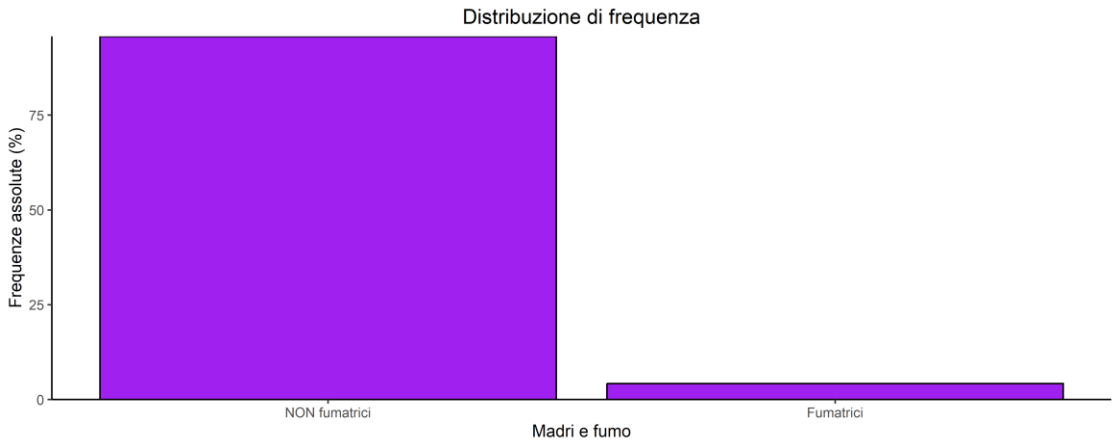
```
  scale_x_continuous(breaks=seq(min(N.gravidanze),max(N.gravidanze),1))+
  scale_y_continuous(expand = c(0, 0))
```



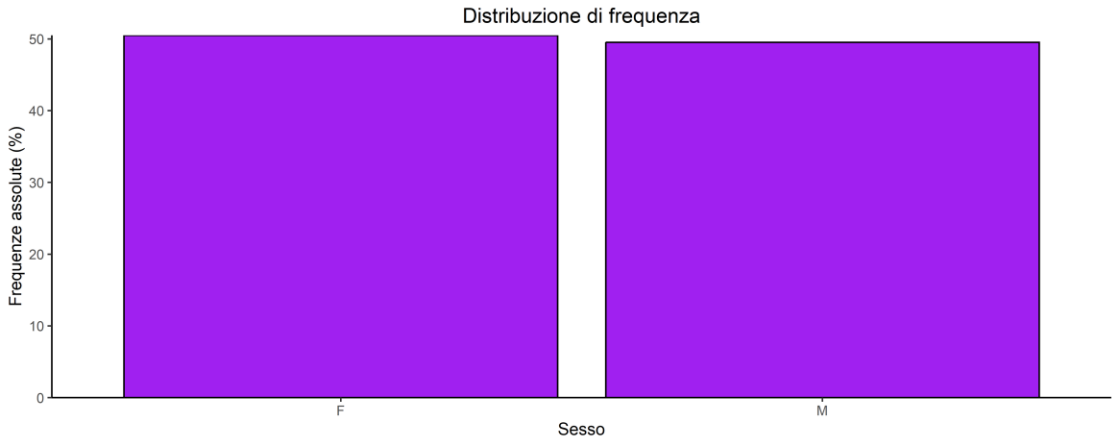

```
# All'aumentare del numero di gravidanze diminuisce il numero di ossevazioni:
cioè la maggior parte della madri hanno avuto pochi o nessun figlio

moments::skewness(N.gravidanze)

## [1] 1.071242

# La skewness positiva della variabile N.gravidanze conferma che sono più
frequenti valori bassi

data_cl$Gestazione_cl=cut(data$Gestazione, breaks=c(23,37,41))
ggplot(data=data_cl)+
  geom_bar(aes(x=Gestazione_cl),
           stat="count",
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Mesi di gestazione",
       y="Frequenze assolute")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(expand = c(0, 0))
```



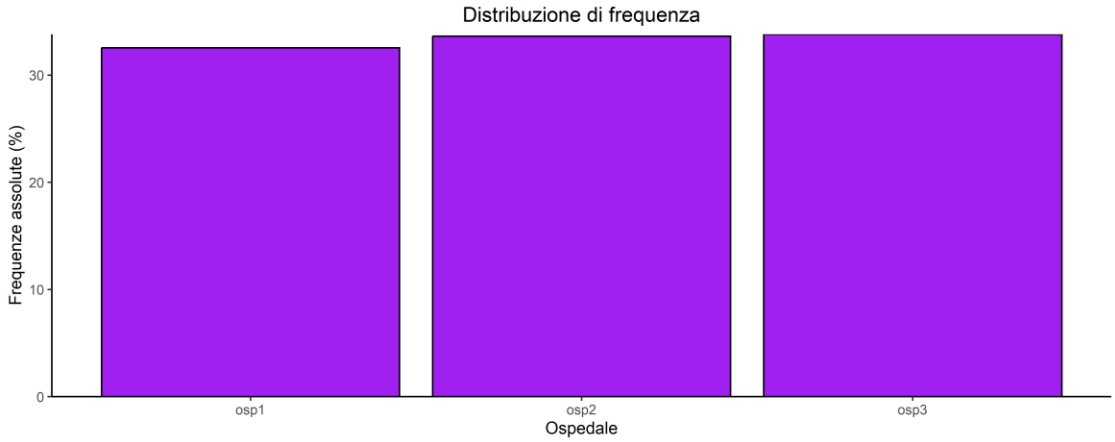

```
# La maggior parte delle donne oggetto di studio partorisce dopo un numero di
mesi di gestazione normale (37-41).
# La minoranza delle donne ha partotito neonati prematuri (23-37 mesi di
gestazione)
# Nessuna delle madri osservate ha partorito post-termine


ggplot(data=data)+
  geom_bar(aes(x=Fumatrici,
               y=after_stat(count/sum(count)*100)),
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Madri e fumo",
       y="Frequenze assolute (%)")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=c(0,1),
                     labels=c("NON fumatrici","Fumatrici"))+
  scale_y_continuous(expand = c(0, 0))
```



```r
# La maggior parte delle madri sono non fumatrici

ggplot(data=data)+
  geom_bar(aes(x=Sesso,y=after_stat(count/sum(count)*100)),
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Sesso",
       y="Frequenze assolute (%)")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(expand = c(0, 0))
```

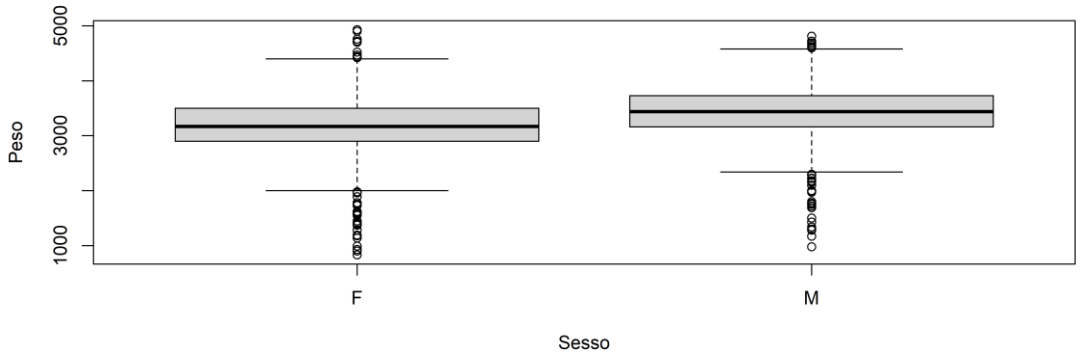



```r
# I due sessi sono quasi equamente distribuiti nel dataset

ggplot(data=data)+
  geom_bar(aes(x=Ospedale,y=after_stat(count/sum(count)*100)),
           col="black",
           fill="purple")+
  labs(title="Distribuzione di frequenza",
       x="Ospedale",
       y="Frequenze assolute (%)")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(expand = c(0, 0))
```




```r
# I tre ospedali sono quasi equamente distribuiti nel dataset

# PUNTO 4

# https://www.ospedalebambinogesu.it/da-0-a-30-giorni-come-si-presenta-e-come-cresce-80012/#:~:text=In%20media%20il%20peso%20nascita,pari%20mediamente%20a%2050%20centimetri.
# Secondo dati dell'ospedale Bambino Gesù:
# Peso medio neonati -> 3300 kg (i maschi pesano circa 150 grammi in più)
# Lunghezza media neonati -> 50 cm
# Non ci sono particolari differenze per quanto riguarda la lunghezza tra maschi e femmine.
# Lunghezza e peso possono essere diversi non solo in base al sesso,
# ma anche a fattori ereditari o ambientali. Ad esempio, se la mamma fuma
# in gravidanza, c'è un rischio aumentato di avere un neonato di basso peso (inferiore a 2500 grammi).

# Poichè voglio saggiare l'ipotesi che la media del campione non differisca da un determinato valore e non conosco la deviazione standard della popolazione utilizzo un t-test
# H0: mu_cap-mu=0
t.test(x=Peso,
       mu=3300,
       conf.level=0.95,
       alternative="two.sided")

##
##  One Sample t-test
##
## data:  Peso
## t = -0.94645, df = 2312, p-value = 0.344
## alternative hypothesis: true mean is not equal to 3300
## 95 percent confidence interval:
##  3268.194 3311.099
## sample estimates:
```

```
## mean of x
##  3289.646

# Accetto l'ipotesi nulla in quanto i criteri sotto sono verificati:
# p-value>0.025
# mu all'interno dell'intervallo di confidenza

# Poichè voglio saggiare l'ipotesi che la media del campione non differisca
da un determinato valore e non conosco la deviazione standard della
popolazione utilizzo un t-test
# H0: mu_cap-mu=0
t.test(x=Lunghezza,
       mu=500,
       conf.level=0.95,
       alternative="two.sided")

##
##  One Sample t-test
##
## data:  Lunghezza
## t = -9.2555, df = 2312, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
##  493.8500 496.0004
## sample estimates:
## mean of x
##  494.9252

# Rifiuto l'ipotesi nulla in quanto i criteri sotto NON sono verificati:
# p-value>0.025
# mu all'interno dell'intervallo di confidenza

# PUNTO 5
boxplot(Peso~Sesso,data=data)
```




```
summary(Peso[Sesso=="F"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     830    2900    3170    3170    3500    4930
```

```
summary(Peso[Sesso=="M"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     980    3160    3440    3412    3730    4810
```

```
mean(Peso[Sesso=="M"])-mean(Peso[Sesso=="F"])
```

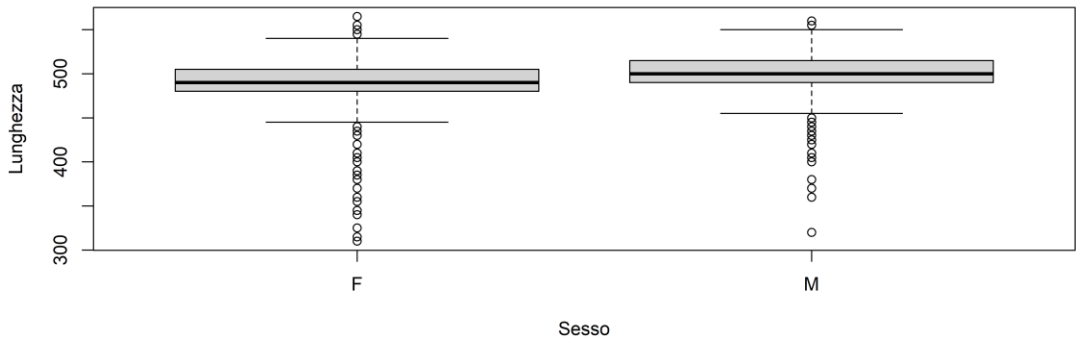
```
## [1] 242.4017
```

```
# la media del Peso è più alta per i maschi che per le femmine del campione,
ma non sappiamo se in modo significativo
# H0:mu_peso_maschi-mu_peso_femmine=0
t.test(data=data,
       Peso~Sesso,
       paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Sesso
## t = -11.389, df = 2308.4, 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:
##  -284.1409 -200.6626
## sample estimates:
## mean in group F mean in group M
##        3169.651        3412.052
```

```
# Rifiuto l'ipotesi nulla in quanto i criteri sotto NON sono verificati:
# p-value>0.025
# 0 all'interno dell'intervallo di confidenza
# Come suggerisce il test sul campione e le statistiche trovate in rete, i
neonati maschi pesano mediamente più delle neonate femmine
```

```
boxplot(Lunghezza~Sesso,data=data)
```




```r
summary(Lunghezza[Sesso=="F"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   310.0   480.0   490.0   489.9   505.0   565.0
```

```r
summary(Lunghezza[Sesso=="M"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     320     490     500     500     515     560
```

```r
mean(Lunghezza[Sesso=="M"])-mean(Lunghezza[Sesso=="F"])
```

```
## [1] 10.06868
```

```r
# la media della Lunghezza è più alta per i maschi che per le femmine del
campione, ma non sappiamo se in modo significativo
# H0:mu_lunghezza_maschi-mu_lunghezza_femmine=0
t.test(data=data,
       Lunghezza~Sesso,
       paired=FALSE)
```

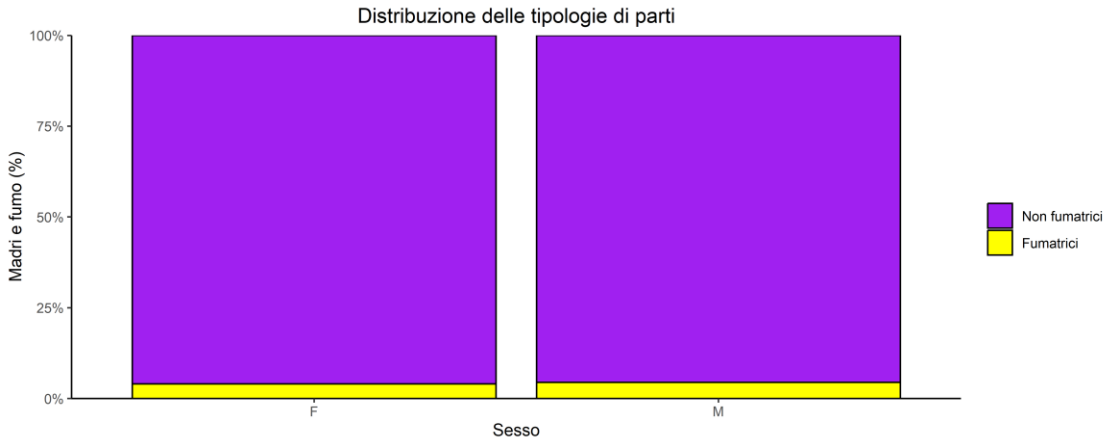
```
##
##  Welch Two Sample t-test
##
## data:  Lunghezza by Sesso
## t = -9.3624, df = 2288.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group
M is not equal to 0
## 95 percent confidence interval:
##  -12.177617  -7.959748
## sample estimates:
## mean in group F mean in group M
##        489.9409        500.0096
```

```r
# Rifiuto l'ipotesi nulla in quanto i criteri sotto NON sono verificati:
# p-value>0.025
# 0 all'interno dell'intervallo di confidenza
# Il test sul campione suggerisce che esista una differenza nella lunghezza
```

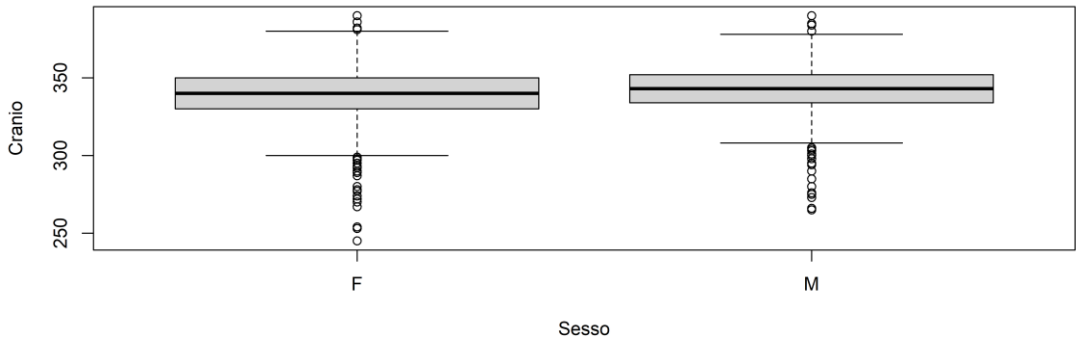
```
media dei neonati tra maschi e femmine,
# nonostante le statistiche trovate in rete dicano che il sesso non determini
una differenza.
# Tale differenza potrebbe dipendere dal caso o dal non aver effettuato un
corretto campionamento.
# In caso di non corretto campionamento, potremmo avere tante neonate da
madri fumatrici
# (e che ipotizziamo abbiano fumato in gravidanza), per cui le neonate
potrebbero avere un peso
# ancora minore rispetto a quello che avrebbero con madri non fumatrici e
questo potrebbe tradursi in una
# minore lunghezza: da come riportato sotto, vediamo che il numero di neonate
da madri fumatrici,
# così come il numero di neonati di madri fumatrici è in realtà simile.
table(Fumatrici,Sesso)
```

```
##          Sesso
## Fumatrici    F    M
##         0 1121 1094
##         1   47   51
```

```
ggplot(data=data,
       aes(x=Sesso,
           fill=factor(Fumatrici)))+
  geom_bar(position ="fill",
           color="black")+
  labs(title="Distribuzione delle tipologie di parti",
       x="Sesso",
       y="Madri e fumo (%)")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(expand=c(0,0),
                     labels = scales::percent_format())+
  scale_fill_manual(name="",
                    values = c("purple","yellow"),
                    labels=c("Non fumatrici","Fumatrici"))
```



```r
# https://www.healthychildren.org/English/ages-stages/baby/Pages/First-Month-Physical-Appearance-and-Growth.aspx
# Il diametro medio del cranio di un neonato è di circa 35 cm.
# Rispetto alle femmine, i maschi hanno in media un diametro maggiore 1 cm.
boxplot(Cranio~Sesso,data=data)
```




```r
summary(Cranio[Sesso=="F"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     245     330     340     338     350     390
```

```r
summary(Cranio[Sesso=="M"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   265.0   334.0   343.0   342.4   352.0   390.0
```

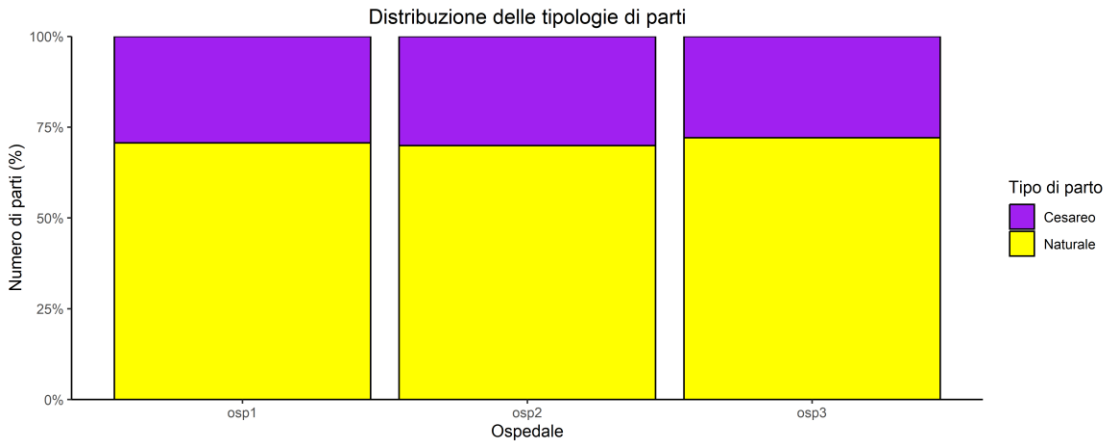
```r
mean(Cranio[Sesso=="M"])-mean(Cranio[Sesso=="F"])
```

```
## [1] 4.402397
```

```r
# la media del diametro del Cranio è più alta per i maschi che per le femmine del campione, , ma non sappiamo se in modo significativo
# H0:mu_cranio_maschi-mu_cranio_femmine=0
t.test(data=data,
       Cranio~Sesso,
       paired=FALSE)
```

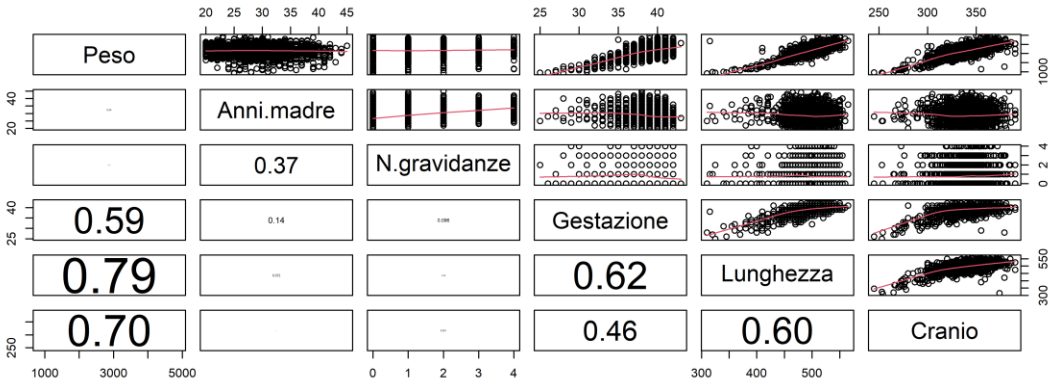
```
##
##  Welch Two Sample t-test
##
## data:  Cranio by Sesso
## t = -6.5252, df = 2309.6, p-value = 8.32e-11
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -5.725435 -3.079358
## sample estimates:
```

```
## mean in group F mean in group M
##        337.9889        342.3913

# Rifiuto l'ipotesi nulla in quanto i criteri sotto NON sono verificati:
# p-value>0.025
# 0 all'interno dell'intervallo di confidenza
# Come da info trovate online, anche in questo campione il diametro del
cranio dei
# neonati maschi risulta leggermente maggiore del diametro del cranio delle
neonate femmine,
# con una differenza minore di 1 cm.

# PUNTO 6
ggplot(data=data,
       aes(x=Ospedale,
           fill=Tipo.parto))+
  geom_bar(position = "fill",
           stat = "count",
           color="black")+
  labs(title="Distribuzione delle tipologie di parti",
       x="Ospedale",
       y="Numero di parti (%)")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(expand=c(0,0),labels = scales::percent_format())+
  scale_fill_manual(name="Tipo di parto",
                    labels=c("Cesareo","Naturale"),
                    values = c("purple","yellow"))
```




```
# Le differenze tra le percentuali di parti cesarei nei 3 ospedali non
risultano marcate nel campione

#TEST CHI-QUADRATO PER CONFRONTO DI PROPORZIONI TRA GRUPPI
# H0: le probabilità di avere parto CESAREO in un determinato ospedale è la
stessa per i 3 ospedali
data$Tipo.parto_d=ifelse(Tipo.parto=="Ces",1,0)
case.vector=tapply(data$Tipo.parto_d,Ospedale,sum) # numero di casi
```

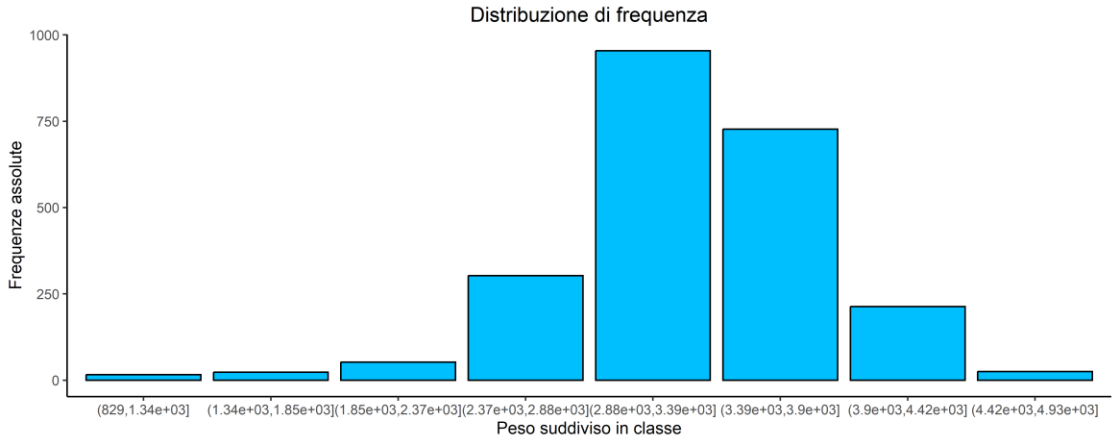
```r
favorevoli (parto cesareo) per ogni ospedale
total.vector=tapply(data$Tipo.parto_d,Ospedale,length) # numero di casi
totali per ogni ospedale
prop.test(x=case.vector,
          n=total.vector,
          conf.level = 0.95)

##
##  3-sample test for equality of proportions without continuity correction
##
## data:  case.vector out of total.vector
## X-squared = 0.94946, df = 2, p-value = 0.6221
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3
## 0.2934927 0.3007712 0.2787724

# Non si rifiuta l'ipotesi nulla, ovvero la probabilità di avere un parto
cesareo è la stessa nei 3 ospedali,
# in quanto il p-value è maggiore di 0.05.
data=subset(data, select = -Tipo.parto_d)

# ANALISI MULTIDIMENSIONALE

# PUNTO 1
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(data_quant[,c(4,1:3,5,6)],upper.panel = panel.smooth,lower.panel =
panel.cor) # matrice degli scatterplots e dei coefficienti di correlazione
tra le variabili quantitative
```



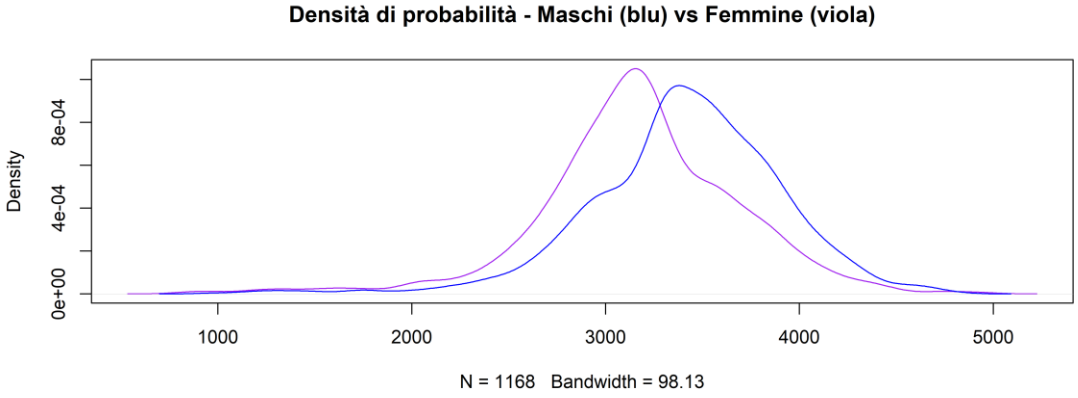
```r
# https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/#:~:text=High%20degree%3A%20If%20the%20coefficient,to%20be%20a%20small%20correlation.
# Degree of correlation:
# Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
# High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
# Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
# Low degree: When the value lies below + .29, then it is said to be a small correlation.
# No correlation: When the value is zero.
# Correlazioni medio-alte:
# Lunghezza-Peso -> 0.79 (tendenza lineare)
# Cranio-Peso -> 0.70 (tendenza lineare; sembra ci siano due pendenze)
# Gestazione-Peso -> 0.59 (tendenza quasi lineare; pendenza variabile)
# Cranio-Gestazione -> 0.46 (retta in parte quasi orizzontale o legame quadratico)
# Lunghezza-Gestazione -> 0.62 (retta in parte quasi orizzontale o legame quadratico)
# Cranio-Lunghezza -> 0.60 (tendenza a legame quadratico)


range(Peso)
```

```
## [1]  830 4930
```

```r
range_peso=max(Peso)-min(Peso)
data_cl$Peso.cl=cut(Peso, breaks=seq(from=min(Peso)-1,to=max(Peso),length.out=9)) # divido il peso in classi con un range di circa 500 (g)
ggplot(data=data_cl)+
  geom_bar(aes(x=Peso.cl),
           stat="count",
           col="black",
           fill="deepskyblue")+
  labs(title="Distribuzione di frequenza",
       x="Peso suddiviso in classe",
       y="Frequenze assolute")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```




```r
tabella_sesso_peso=table(Sesso,data_cl$Peso.cl)
test.indipendenza.1=chisq.test(tabella_sesso_peso)
test.indipendenza.1

##
##  Pearson's Chi-squared test
##
## data:  tabella_sesso_peso
## X-squared = 153.06, df = 7, p-value < 2.2e-16

# poichè il p-value è minore di 0.05 si rifiuta l'ipotesi nulla di
indipendenza delle variabili Sesso e Peso.cl
plot(density(Peso[Sesso=="F"]),col="purple",main="")
lines(density(Peso[Sesso=="M"]),col="blue")
title(main="Densità di probabilità - Maschi (blu) vs Femmine (viola)")
```



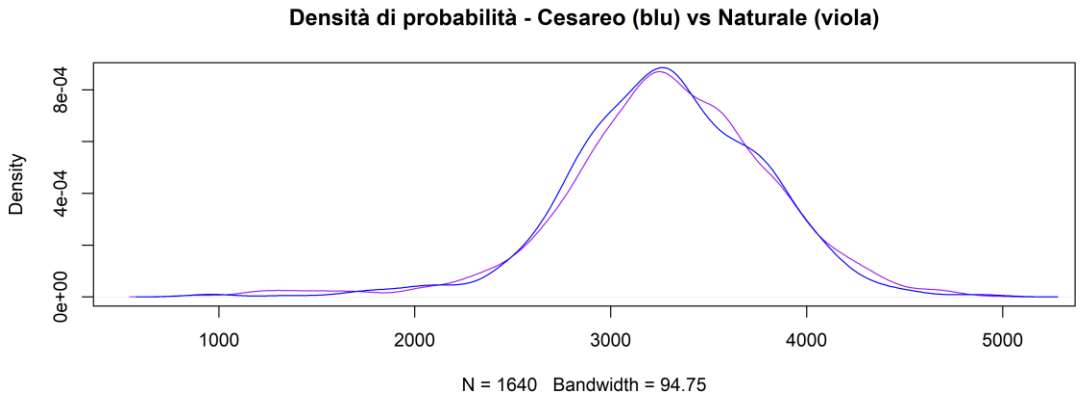

```r
# si osserva che per valori più alti del peso, le osservazioni più probabili
sono di neonati maschi


tabella_parto_peso=table(Tipo.parto,data_cl$Peso.cl)
test.indipendenza.2=chisq.test(tabella_parto_peso)
```

```
## Warning in chisq.test(tabella_parto_peso): L'approssimazione al Chi-
quadrato
## potrebbe essere inesatta
```

```
test.indipendenza.2
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabella_parto_peso
## X-squared = 5.2261, df = 7, p-value = 0.6324
```

```
# poichè il p-value è maggiore di 0.05 si accetta l'ipotesi nulla di
indipendenza delle variabili Tipo.parto e Peso.cl
plot(density(Peso[Tipo.parto=="Nat"]),col="purple",main="")
lines(density(Peso[Tipo.parto=="Ces"]),col="blue")
title(main="Densità di probabilità - Cesareo (blu) vs Naturale (viola)")
```



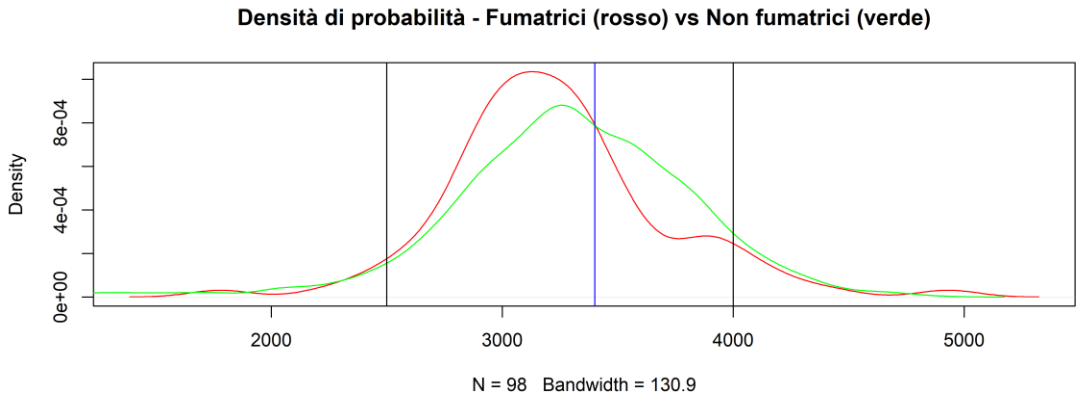

```
# non si osserva un legame tra il peso e il tipo di parto, in quanto le curve
quasi si sovrappongono


# https://academic.oup.com/aje/article/165/8/849/184757?login=false
# Peso normale alla nascita: 2500-4000 g
# https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8791242/
# Smoking during pregnancy is associated with a considerable reduction in
birth weight
# in different geographic areas, with the range of weight reduction ranging
from
# 77.7 to 232.7 g.
tabella_fumatrice_peso=table(Fumatrici,data_cl$Peso.cl)
test.indipendenza.3=chisq.test(tabella_fumatrice_peso)
```

```
## Warning in chisq.test(tabella_fumatrice_peso): L'approssimazione al
## Chi-quadrato potrebbe essere inesatta
```

```
test.indipendenza.3
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabella_fumatrice_peso
## X-squared = 11.75, df = 7, p-value = 0.1091
```

```
# poichè il p-value è maggiore di 0.05 si accetta l'ipotesi nulla di
indipendenza delle variabili Fumatrici e Peso.cl
# Questo risultato potrebbe essere dovuto al caso, o a un non corretto
campionamento o ancora potrebbe significare
# che alcune o tutte le madri fumatrici NON abbiano fumato in gravidanza
plot(density(Peso[Fumatrici=="1"]),col="red",main="")
lines(density(Peso[Fumatrici=="0"]),col="green")
title(main="Densità di probabilità - Fumatrici (rosso) vs Non fumatrici
(verde)")
abline(v=c(2500,4000))
abline(v=3400,col="blue")
```




```
# Le densità di probabilità per pesi bassi o alti sono quasi sovrapponibili
(possibile assenza di relazione tra peso basso o alto e madre fumatrice).
# Tuttavia, nella zona normopeso, il picco delle madri fumatrici è spostato
più a sinistra, indicando che valori più bassi del peso sono più probabili
per madri fumatrici
nonfumatrici_normopeso=sum(table(subset(data,Peso>=2500 & Peso <=4000 &
Fumatrici==0)$Peso))
fumatrici_normopeso=sum(table(subset(data,Peso>=2500 & Peso <=4000 &
Fumatrici==1)$Peso))
n_nonfumatrici=nrow(subset(data,Fumatrici==0))
n_fumatrici=nrow(subset(data,Fumatrici==1))
p_nonfumatrici_normopeso=nonfumatrici_normopeso/n_nonfumatrici
p_fumatrici_normopeso=fumatrici_normopeso/n_fumatrici
p_nonfumatrici_normopeso
```

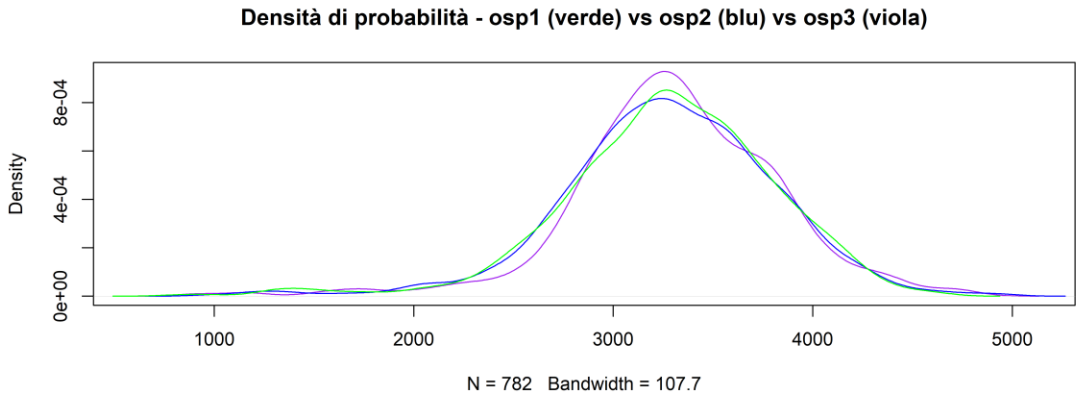
```
## [1] 0.8776524
```

```
p_fumatrici_normopeso
```

```
## [1] 0.8877551

# Le proporzioni di neonati normopeso con madri non fumatrici o fumatrici
# sono simili nel campione (possibile assenza di relazione tra normopeso e
madre fumatrice).
#TEST CHI-QUADRATO PER CONFRONTO DI PROPORZIONI TRA GRUPPI
# H0: la probabilità di avere neonati normopeso in una determinata categoria
di madre (non fumatrice/fumatrice) è la stessa per le due categorie
case.vector=c(nonfumatrici_normopeso,fumatrici_normopeso) # numero di casi
favorevoli (normopeso) per ogni categoria di madre (non fumatrice/fumatrice)
total.vector=c(n_nonfumatrici,n_fumatrici) # numero di casi totali per ogni
categoria di madre (non fumatrice/fumatrice)
prop.test(x=case.vector,
          n=total.vector,
          conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  case.vector out of total.vector
## X-squared = 0.019987, df = 1, p-value = 0.8876
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.07940085  0.05919539
## sample estimates:
##    prop 1    prop 2
## 0.8776524 0.8877551

# poichè il p-value è maggiore di 0.05 si accetta l'ipotesi nulla


nonfumatrici_normopeso=sum(table(subset(data,Peso>=2500 & Peso <=3400 &
Fumatrici==0)$Peso))
fumatrici_normopeso=sum(table(subset(data,Peso>=2500 & Peso <=3400 &
Fumatrici==1)$Peso))
n_nonfumatrici=nrow(subset(data,Fumatrici==0))
n_fumatrici=nrow(subset(data,Fumatrici==1))
p_nonfumatrici_normopeso=nonfumatrici_normopeso/n_nonfumatrici
p_fumatrici_normopeso=fumatrici_normopeso/n_fumatrici
p_nonfumatrici_normopeso

## [1] 0.5363431

p_fumatrici_normopeso

## [1] 0.7040816

# Le proporzioni di neonati con peso tra 2500 e 3400 g sono diverse
considerando madri non fumatrici o fumatrici
# (possibile relazione tra peso dato e madre fumatrice): prevale il numero di
neonati da madri fumatrici
```

```r
#TEST CHI-QUADRATO PER CONFRONTO DI PROPORZIONI TRA GRUPPI
# H0: le probabilità di avere neonati con peso tra 2500 e 3400 g per una determinata categoria di madre (non fumatrice/fumatrice) è la stessa per le due categorie
case.vector=c(nonfumatrici_normopeso,fumatrici_normopeso) # numero di casi favorevoli (peso 2500-3400) per ogni categoria di madre (non fumatrice/fumatrice)
total.vector=c(n_nonfumatrici,n_fumatrici) # numero di casi totali per ogni categoria di madre (non fumatrice/fumatrice)
prop.test(x=case.vector,
          n=total.vector,
          conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  case.vector out of total.vector
## X-squared = 9.9771, df = 1, p-value = 0.001585
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.26579347 -0.06968357
## sample estimates:
##    prop 1    prop 2
## 0.5363431 0.7040816

# poichè il p-value è minore di 0.05 si rifiuta l'ipotesi nulla



nonfumatrici_normopeso=sum(table(subset(data,Peso>=3400 & Peso <=4000 & Fumatrici==0)$Peso))
fumatrici_normopeso=sum(table(subset(data,Peso>=3400 & Peso <=4000 & Fumatrici==1)$Peso))
n_nonfumatrici=nrow(subset(data,Fumatrici==0))
n_fumatrici=nrow(subset(data,Fumatrici==1))
p_nonfumatrici_normopeso=nonfumatrici_normopeso/n_nonfumatrici
p_fumatrici_normopeso=fumatrici_normopeso/n_fumatrici
p_nonfumatrici_normopeso

## [1] 0.3544018

p_fumatrici_normopeso

## [1] 0.2142857

# Le proporzioni di neonati con peso tra 3400 e 4000 g sono diverse considerando madri non fumatrici o fumatrici
# (possibile relazione tra peso dato e madre fumatrice): prevale il numero di neonati da madri NON fumatrici
#TEST CHI-QUADRATO PER CONFRONTO DI PROPORZIONI TRA GRUPPI
# H0: le probabilità di avere neonati con peso tra 3400 e 4000 g in una
```

```r
determinata categoria di madre (non fumatrice/fumatrice) è la stessa per le due categorie
case.vector=c(nonfumatrici_normopeso,fumatrici_normopeso) # numero di casi favorevoli (peso 2500-3400) per ogni categoria di madre (non fumatrice/fumatrice)
total.vector=c(n_nonfumatrici,n_fumatrici) # numero di casi totali per ogni categoria di madre (non fumatrice/fumatrice)
prop.test(x=case.vector,
          n=total.vector,
          conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  case.vector out of total.vector
## X-squared = 7.5099, df = 1, p-value = 0.006136
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05114284 0.22908934
## sample estimates:
##    prop 1    prop 2
## 0.3544018 0.2142857

# poichè il p-value è minore di 0.05 si rifiuta l'ipotesi nulla


tabella_ospedale_peso=table(Ospedale,data_cl$Peso.cl)
test.indipendenza.4=chisq.test(tabella_ospedale_peso)
test.indipendenza.4

##
##  Pearson's Chi-squared test
##
## data:  tabella_ospedale_peso
## X-squared = 18.351, df = 14, p-value = 0.1912

# poichè il p-value è maggiore di 0.05 si accetta l'ipotesi nulla di indipendenza delle variabili Ospedale e Peso.cl
plot(density(Peso[Ospedale=="osp3"]),col="purple",main="")
lines(density(Peso[Ospedale=="osp2"]),col="blue")
lines(density(Peso[Ospedale=="osp1"]),col="green")
title(main="Densità di probabilità - osp1 (verde) vs osp2 (blu) vs osp3 (viola)")
```




```
# le 3 curve quasi si sovrappongono (anche se si osserva un picco più alto
rispetto all'ospedale 3):
# non si osserva un legame tra lo specifico ospedale e il peso del neonato


# PUNTO 2
mod1=lm(Peso~.,data=data)
summary(mod1)

##
## Call:
## lm(formula = Peso ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1116.08  -182.62   -14.63   163.52  2589.47
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6734.9938   150.4022 -44.780  < 2e-16 ***
## Anni.madre       -0.3895     1.3294  -0.293  0.76955
## N.gravidanze     18.7800     6.2244   3.017  0.00258 **
## Fumatrici       -34.8482    28.6696  -1.216  0.22430
## Gestazione       33.8684     4.0126   8.441  < 2e-16 ***
## Lunghezza        10.1630     0.3150  32.263  < 2e-16 ***
## Cranio           10.5862     0.4461  23.731  < 2e-16 ***
## Tipo.partoNat    30.5211    12.6724   2.408  0.01610 *
## Ospedaleosp2     -3.6689    14.1381  -0.260  0.79527
## Ospedaleosp3     32.2723    14.1134   2.287  0.02231 *
## SessoM           76.3894    11.7175   6.519 8.65e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276.2 on 2302 degrees of freedom
## Multiple R-squared:  0.7256, Adjusted R-squared:  0.7245
## F-statistic: 608.9 on 10 and 2302 DF,  p-value: < 2.2e-16
```

```r
# Il test di ipotesi sui coefficienti beta con ipotesi nulla beta=0 mostra, considerando un livello di
# significatività pari a 0.05, in base ai valori del p-value, che:
# N.gravidanze, Gestazione, Lunghezza, Cranio, Tipo parto, Ospedale ("osp3"), Sesso spiegano la risposta
# Adjusted R-squared:  0.7245

# PUNTO 3
# mod2: elimino le variabili non significative del modello 1
mod2=lm(Peso~N.gravidanze+Gestazione+Lunghezza+Cranio+Tipo.parto+Ospedale+Sesso)
summary(mod2)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1119.2  -182.9   -14.8   163.3  2591.4
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6748.2297   142.8971 -47.224  < 2e-16 ***
## N.gravidanze     17.6002     5.7971   3.036  0.00242 **
## Gestazione       33.6573     3.9866   8.443  < 2e-16 ***
## Lunghezza        10.1818     0.3146  32.364  < 2e-16 ***
## Cranio           10.5885     0.4457  23.758  < 2e-16 ***
## Tipo.partoNat    30.2841    12.6696   2.390  0.01691 *
## Ospedaleosp2     -3.6971    14.1317  -0.262  0.79364
## Ospedaleosp3     32.6455    14.1042   2.315  0.02072 *
## SessoM           76.1286    11.7117   6.500  9.8e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276.1 on 2304 degrees of freedom
## Multiple R-squared:  0.7255, Adjusted R-squared:  0.7245
## F-statistic:   761 on 8 and 2304 DF,  p-value: < 2.2e-16
```

```r
library(car)
```

```
## Warning: il pacchetto 'car' è stato creato con R versione 4.3.2

## Caricamento del pacchetto richiesto: carData

## Warning: il pacchetto 'carData' è stato creato con R versione 4.3.2
```

```r
# https://rpubs.com/CPEL/multlinearregression
# For each variable, GVIF (generalized VIF), DF, and GVIF normalized by DF is produced.
```

```r
# The accepted VIF cutoffs are 5 or 10, depending on who you ask.
# This is to say that if your VIF value is above 5 (or 10), you may consider
that variable
# is collinear and needs to be removed.
# If you run this without any categorical variables, the output will look
different.
# It produces VIFs (instead of GVIFs), which do not need to be normalized by
DF,
# so you can just compare them to your cutoff of 5 or 10.
vif(mod2)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## N.gravidanze 1.027677  1        1.013744
## Gestazione   1.671517  1        1.292872
## Lunghezza    2.086641  1        1.444521
## Cranio       1.614941  1        1.270803
## Tipo.parto   1.004465  1        1.002230
## Ospedale     1.004841  2        1.001208
## Sesso        1.040004  1        1.019806
```

```r
# Multicollinearità: nessuna variabile X dà problemi al modello in quanto
GVIF^(1/(2*Df)) < 5
# Adjusted R-squared:  0.7245 (non cambia, preferisco il modello 2, più
semplice)

# mod3: elimino le variabili che, in base al punto 1, sono risultate
indipendenti rispetto a Y
mod3=lm(Peso~Gestazione+Lunghezza+Cranio+Sesso)
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1135.60  -183.89   -14.58   163.49  2603.10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6686.9964   142.5710 -46.903  < 2e-16 ***
## Gestazione     32.7114     3.9789   8.221 3.32e-16 ***
## Lunghezza      10.0971     0.3153  32.026  < 2e-16 ***
## Cranio         10.7761     0.4448  24.228  < 2e-16 ***
## SessoM         77.7215    11.7531   6.613 4.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.3 on 2308 degrees of freedom
```

```
## Multiple R-squared:  0.7227, Adjusted R-squared:  0.7222
## F-statistic:  1503 on 4 and 2308 DF,  p-value: < 2.2e-16

# Adjusted R-squared:  0.7222  (non si riduce molto, preferisco il modello 3, più semplice)

# mod4: aggiungo l'effetto quadratico di Gestazione, in base allo scatterplot del punto 1
mod4=update(mod3,~.+I(Gestazione^2))
summary(mod4)

##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     I(Gestazione^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1130.58  -183.24   -15.12   164.30  2624.56
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4845.6461   940.8977  -5.150 2.83e-07 ***
## Gestazione        -70.0363    52.0489  -1.346   0.1786
## Lunghezza          10.1946     0.3189  31.967  < 2e-16 ***
## Cranio             10.8547     0.4463  24.323  < 2e-16 ***
## SessoM             75.6431    11.7925   6.415 1.71e-10 ***
## I(Gestazione^2)     1.3722     0.6931   1.980   0.0478 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.1 on 2307 degrees of freedom
## Multiple R-squared:  0.7231, Adjusted R-squared:  0.7225
## F-statistic:  1205 on 5 and 2307 DF,  p-value: < 2.2e-16

# Adjusted R-squared:  0.7225  (cresce poco, escludiamo questo modello più complesso)

# valuto se il sesso possa avere una qualche influenza sulla composizione corporea e
# quindi il contributo che la lunghezza possa dare al peso a seconda del sesso
mod5=update(mod3,~.+Sesso*Lunghezza)
summary(mod5)

##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     Lunghezza:Sesso)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1144.69  -183.53   -17.41   162.20  2554.80
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6544.1373   172.6775 -37.898  < 2e-16 ***
## Gestazione          32.8347     3.9788   8.252 2.58e-16 ***
## Lunghezza            9.8148     0.3694  26.569  < 2e-16 ***
## Cranio              10.7486     0.4451  24.151  < 2e-16 ***
## SessoM            -248.9185   223.1783  -1.115    0.265
## Lunghezza:SessoM     0.6591     0.4497   1.466    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.2 on 2307 degrees of freedom
## Multiple R-squared:  0.7229, Adjusted R-squared:  0.7223
## F-statistic:  1204 on 5 and 2307 DF,  p-value: < 2.2e-16
```

```
# l'interazione delle variabili sesso e lunghezza non è significativa
# (p-value del coefficiente relativo all'interazione Sesso-Lunghezza maggiore
di 0.05)

# valuto se il sesso possa avere una qualche influenza sulla composizione
corporea e
# quindi il contributo che il diametro della testa possa dare al peso a
seconda del sesso
mod6=update(mod3,~.+Sesso*Cranio)
summary(mod6)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     Cranio:Sesso)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1137.33  -184.63   -15.95   164.71  2612.55
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6601.5356   180.7746 -36.518  < 2e-16 ***
## Gestazione       32.8157     3.9816   8.242 2.81e-16 ***
## Lunghezza        10.0964     0.3153  32.020  < 2e-16 ***
## Cranio           10.5124     0.5617  18.716  < 2e-16 ***
## SessoM         -108.7535   242.7649  -0.448    0.654
## Cranio:SessoM     0.5479     0.7124   0.769    0.442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.3 on 2307 degrees of freedom
```

```
## Multiple R-squared:  0.7227, Adjusted R-squared:  0.7221
## F-statistic:  1203 on 5 and 2307 DF,  p-value: < 2.2e-16

# l'interazione delle variabili sesso e cranio non è significativa
# (p-value del coefficiente relativo all'interazione Sesso-Cranio maggiore di
0.05)

AIC(mod1,mod2,mod3,mod4,mod5,mod6) # il modello migliore secondo l'AIC è il
modello 2 (AIC minimo)

##      df      AIC
## mod1 12 32579.86
## mod2 10 32577.41
## mod3  6 32592.92
## mod4  7 32590.99
## mod5  7 32592.77
## mod6  7 32594.33

BIC(mod1,mod2,mod3,mod4,mod5,mod6) # il modello migliore secondo l'AIC è il
modello 3 (BIC minimo)

##      df      BIC
## mod1 12 32648.81
## mod2 10 32634.87
## mod3  6 32627.40
## mod4  7 32631.22
## mod5  7 32632.99
## mod6  7 32634.55

# si sceglie tra i due modelli il modello più semplice, mod3

n=nrow(data)
stepwise.mod=MASS::stepAIC(mod1, # tutta la procedura di ricerca del modello
migliore si può effetturare con la funzione stepAIC
                           direction="both", # si parte dal modello 1, con
procedura mixed e usando il criterio BIC
                           k=log(n))

## Start:  AIC=26077.06
## Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Anni.madre    1      6548 175583634 26069
## - Ospedale      2    605957 176183043 26070
## - Fumatrici     1    112689 175689775 26071
## - Tipo.parto    1    442433 176019519 26075
## <none>                      175577086 26077
## - N.gravidanze  1    694322 176271408 26078
## - Sesso         1   3241614 178818700 26112
## - Gestazione    1   5433872 181010958 26140
```

```
## - Cranio        1  42951870 218528956 26576
## - Lunghezza     1  79391732 254968819 26932
##
## Step:  AIC=26069.4
## Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Ospedale      2    605299 176188933 26062
## - Fumatrici     1    111346 175694980 26063
## - Tipo.parto    1    441779 176025413 26068
## <none>                      175583634 26069
## - N.gravidanze  1    741536 176325170 26071
## + Anni.madre    1      6548 175577086 26077
## - Sesso         1   3236327 178819961 26104
## - Gestazione    1   5514756 181098390 26133
## - Cranio        1  42974152 218557786 26568
## - Lunghezza     1  79439307 255022941 26925
##
## Step:  AIC=26061.86
## Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Fumatrici     1    126295 176315228 26056
## - Tipo.parto    1    463806 176652739 26060
## <none>                      176188933 26062
## - N.gravidanze  1    806690 176995623 26065
## + Ospedale      2    605299 175583634 26069
## + Anni.madre    1      5890 176183043 26070
## - Sesso         1   3253651 179442584 26096
## - Gestazione    1   5621219 181810152 26127
## - Cranio        1  43036116 219225049 26560
## - Lunghezza     1  79176920 255365853 26913
##
## Step:  AIC=26055.78
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Tipo.parto    1    457409 176772637 26054
## <none>                      176315228 26056
## - N.gravidanze  1    764149 177079377 26058
## + Fumatrici     1    126295 176188933 26062
## + Ospedale      2    620248 175694980 26063
## + Anni.madre    1      4527 176310701 26064
## - Sesso         1   3238537 179553764 26090
## - Gestazione    1   5535259 181850487 26120
## - Cranio        1  43106881 219422109 26554
## - Lunghezza     1  79620179 255935407 26910
```

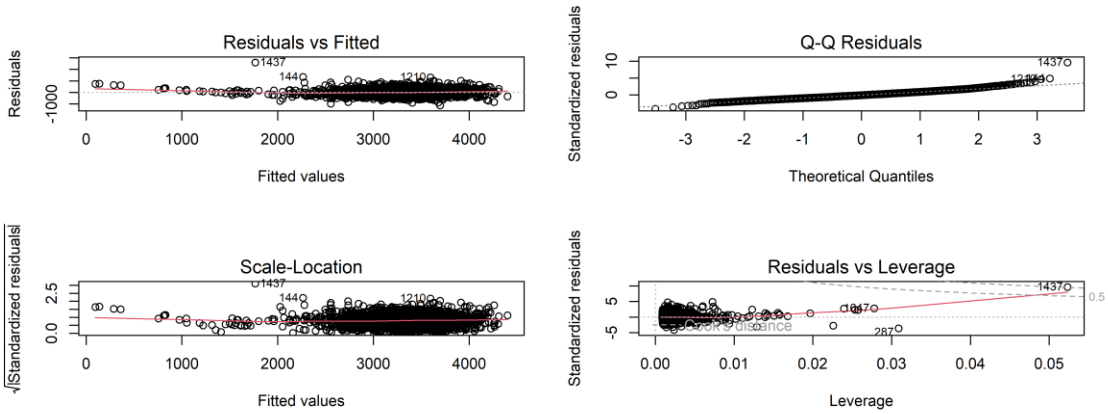
```
##
## Step:  AIC=26054.02
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
##
##                Df Sum of Sq       RSS   AIC
## <none>                      176772637 26054
## - N.gravidanze  1    717417 177490054 26056
## + Tipo.parto    1    457409 176315228 26056
## + Fumatrici     1    119898 176652739 26060
## + Ospedale      2    641967 176130670 26061
## + Anni.madre    1      4009 176768628 26062
## - Sesso         1   3247545 180020182 26088
## - Gestazione    1   5557145 182329782 26118
## - Cranio        1  43460488 220233125 26555
## - Lunghezza     1  79228176 256000813 26903
```

```r
summary(stepwise.mod)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1152.7  -181.7   -15.6   164.2  2611.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6717.5209   142.6626 -47.087  < 2e-16 ***
## N.gravidanze    17.7401     5.7977   3.060  0.00224 **
## Gestazione      34.0189     3.9946   8.516  < 2e-16 ***
## Lunghezza       10.1236     0.3148  32.156  < 2e-16 ***
## Cranio          10.6325     0.4464  23.816  < 2e-16 ***
## SessoM          76.4264    11.7395   6.510 9.18e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276.8 on 2307 degrees of freedom
## Multiple R-squared:  0.7238, Adjusted R-squared:  0.7232
## F-statistic:  1209 on 5 and 2307 DF,  p-value: < 2.2e-16
```

```r
# Adjusted R-squared:  0.7232
# questo metodo restituisce un modello che rispetto al modello 3 considera
anche la variabile
# N.Gravidanze: tuttavia l'R^2 aggiustato cresce molto poco, quindi si
seleziona ancora il modello 3

# PUNTO 5
```

```r
par(mfrow=c(2,2))
plot(mod3)
```




```r
# RESIDUALS VS FITTED: Non tutti i residui hanno media nulla (linea rossa) attorno ai valori stimati
# Q-Q RESIDUALS: I quantili delle distribuzione dei residui cadono all'incirca in corrispondenza dei
# quantili di una distribuzione normale (sulla retta)
# SCALE-LOCATION: La varianza dei residui è all'incirca costante (linea rossa quasi orizzontale) lungo i valori stimati
# RESIDUALS VS LEVERAGE: Individuato un punto levegare (osservazione 1437) che potrebbe avere influenza sul modello,
# perchè la sua distanza di Cook sembra superare la soglia di 1

shapiro.test(residuals(mod3))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod3)
## W = 0.97339, p-value < 2.2e-16

# poichè il p-value è minore di 0.05 si rifiuta l'ipotesi nulla di distribuzione normale dei residui

library(lmtest)

## Warning: il pacchetto 'lmtest' è stato creato con R versione 4.3.2

## Caricamento del pacchetto richiesto: zoo

##
## Caricamento pacchetto: 'zoo'

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

```r
bptest(mod3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 87.429, df = 4, p-value < 2.2e-16
```

```r
# poichè il p-value è minore di 0.05 si rifiuta l'ipotesi nulla di
omoschedasticità dei residui

dwtest(mod3)
```

```
##
##  Durbin-Watson test
##
## data:  mod3
## DW = 1.9454, p-value = 0.09446
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# poichè il p-value è maggiore di 0.05 non si rifiuta l'ipotesi nulla di
indipendenza dei residui

# trovo i punti ad alto leverage considerando solo quelli il cui
# leverage è 2 volte più grande del leverage medio
lev=hatvalues(mod3)
plot(lev)
p=sum(lev)
soglia=2*p/n
abline(h=soglia,col=2)
high_lev=lev[lev>soglia]
length(high_lev)
```

```
## [1] 122
```

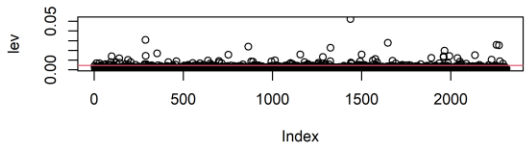
```r
which.max(lev)
```

```
## 1437
## 1437
```

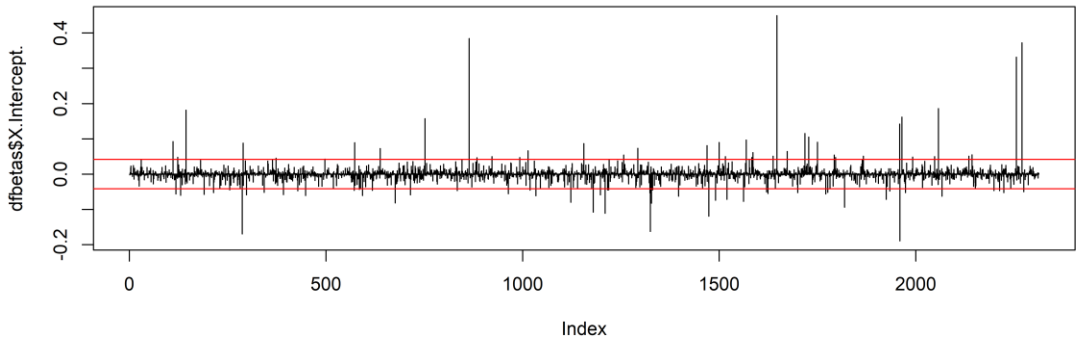
```r
# vengono identificati 122 osservazioni con alto leverage.
# L'osservazione con leverage massimo è la 1437

# test l'ipotesi nulla che le osservazioni non siano outlier
# https://quantoid.net/files/702/lecture9.pdf
outlierTest(mod3)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 1437 9.840771         2.0611e-22   4.7673e-19
## 144  4.894616         1.0531e-06   2.4358e-03
## 1210 4.724428         2.4464e-06   5.6584e-03
```

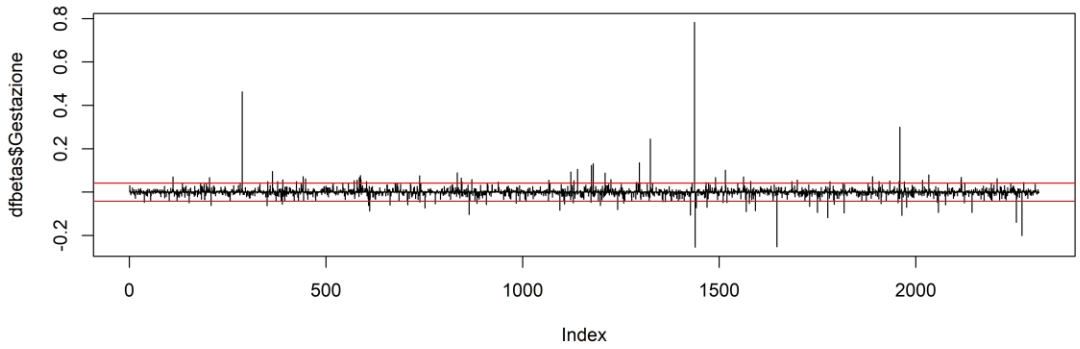
```r
# 3 outliers (p<0.05), tra cui l'osservazione 1437


# Effetto di ciascuna osservazione su ogni coefficiente beta:
# le osservazioni per cui il DFBETAS è fuori dalla soglia sono possibili valori influenti
dfbetas=data.frame(dfbetas(mod3))
thresh= 2/sqrt(n)
par(mfrow=c(1,1))
```



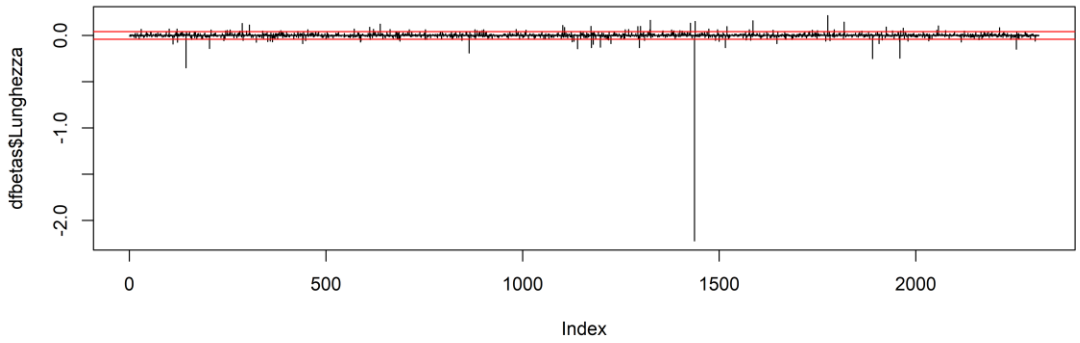

```r
plot(dfbetas$X.Intercept., type='h')
abline(h = thresh, col ="red")
abline(h = -thresh, col ="red")
```



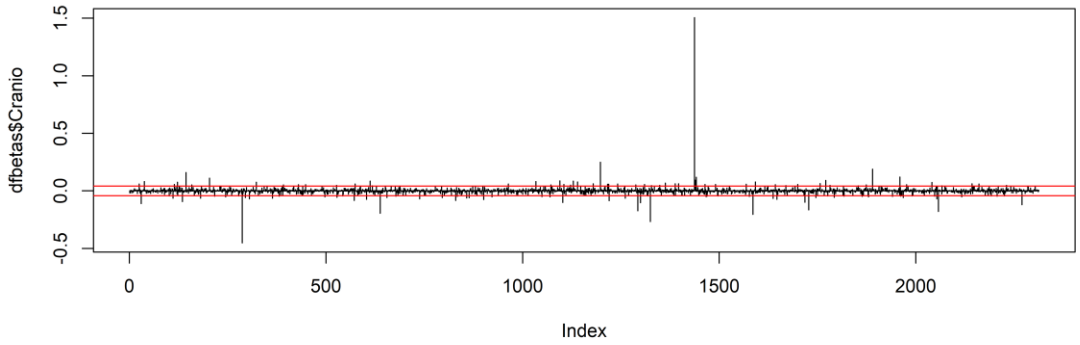

```r
plot(dfbetas$Gestazione, type='h')
abline(h = thresh, col ="red")
abline(h = -thresh, col ="red")
```



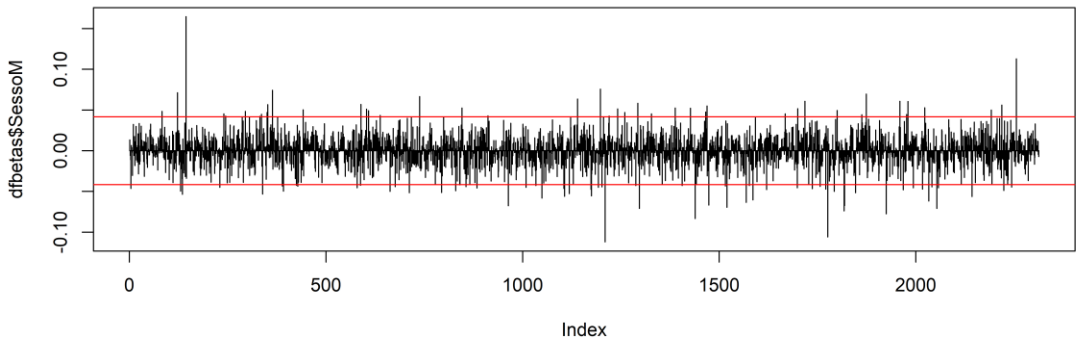

```r
plot(dfbetas$Lunghezza, type='h')
abline(h = thresh, col ="red")
abline(h = -thresh, col ="red")
```



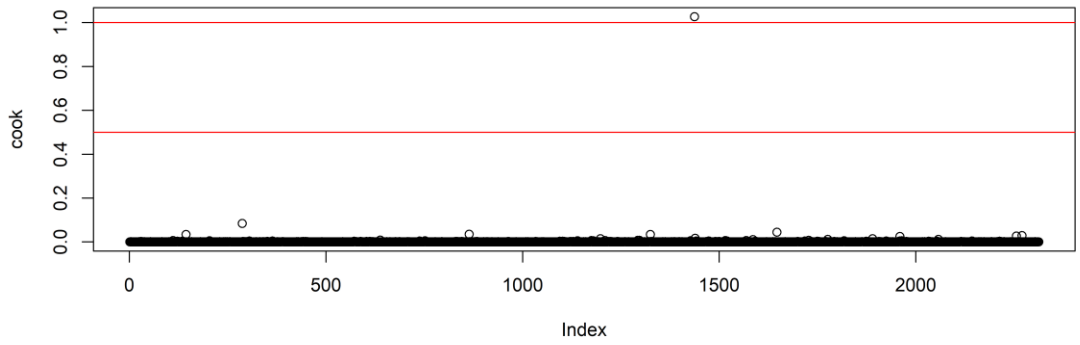

```r
plot(dfbetas$Cranio, type='h')
abline(h = thresh, col ="red")
abline(h = -thresh, col ="red")
```



```r
plot(dfbetas$SessoM, type='h')
abline(h = thresh, col ="red")
abline(h = -thresh, col ="red")
```




```r
# possibili osservazioni influenti sono presenti per ogni coefficiente

# distanza di cook
cook=cooks.distance(mod3)
plot(cook)
soglia1=0.5
soglia2=1
abline(h=soglia1, col="red")
abline(h=soglia2, col="red")
```




```r
# solamente un'osservazione supera la distanza di cook
max_cook=max(cook)
influencials=cook[cook>1 & cook==max(cook)] # considero come dato influente quello con massima distanza di Cook (maggiore di 1)
influencials
```

```
##     1437
## 1.026884
```

```r
# l'osservazione 1437 ha distanza di cook maggiore di 1
rows_of_influencials=names(influencials)

# https://rpubs.com/christianthieme/769935
# rimuovo le osservazioni che hanno distanza di cook maggiore di 1 fino a
quando ce ne sono e
# non creando più di 10 modelli di regressione
loops=0
list_of_influencials=list(rows_of_influencials)
while (length(influencials)!=0){
  loops=loops+1
  if (loops==1){
    data_reduction=subset(data,row.names(data)!=c(rows_of_influencials))
  } else {

data_reduction=subset(data_reduction,row.names(data_reduction)!=c(rows_of_inf
luencials))
  }
  mod_x=lm(Peso~Gestazione+Lunghezza+Cranio+Sesso,data=data_reduction)
  cook=cooks.distance(mod_x)
  max_cook=max(cook)
  influencials=cook[cook>1 & cook==max(cook)]
  rows_of_influencials=names(influencials)
  list_of_influencials=append(list_of_influencials,rows_of_influencials)
  if (loops>10) {
    break
  }
}
list_of_influencials
```

```
## [[1]]
## [1] "1437"
##
## [[2]]
## [1] "1551"
```

```r
nrow(data)-nrow(data_reduction)
```

```
## [1] 2
```

```r
# 2 osservazioni influenti sono state rimosse dal dataset
mod3.1=mod_x
summary(mod3.1)
```

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

```
## -1152.18  -184.38   -13.92   166.12  1389.10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6686.6885   139.7130 -47.860  < 2e-16 ***
## Gestazione     29.7138     3.9119   7.596 4.42e-14 ***
## Lunghezza      10.7780     0.3168  34.022  < 2e-16 ***
## Cranio         10.1248     0.4409  22.962  < 2e-16 ***
## SessoM         77.7065    11.5200   6.745 1.92e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.7 on 2306 degrees of freedom
## Multiple R-squared:  0.7334, Adjusted R-squared:  0.733
## F-statistic:  1586 on 4 and 2306 DF,  p-value: < 2.2e-16
```

```
# Adjusted R-squared:  0.7334 (aumentato rispetto al modello 3, quindi è un
possibile modello migliore)


# https://stats.oarc.ucla.edu/r/dae/robust-regression/
# https://rpubs.com/DragonflyStats/Robust-Regression-Weighting
# https://datascienceplus.com/robust-regressions-dealing-with-outliers-in-r/
# https://en.wikipedia.org/wiki/Robust_regression
# Gli approcci robusti sono metodi di regressione alternativi al metodo dei
minimi quadrati,
# usati quando le assunzioni richieste da questo metodo non sono rispettate
(per esempio ci sono outliers)
# Aggiusto il modello pesando ogni osservazione in base a quanto è un
outlier:
# più è outlier, minore sarà il suo peso
library(MASS)
```

```
## Warning: il pacchetto 'MASS' è stato creato con R versione 4.3.2
```

```
mod3.2=rlm(Peso~Gestazione+Lunghezza+Cranio+Sesso) #robust fitting of linear
model using an M estimator
summary(mod3.2)
```

```
##
## Call: rlm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1149.691  -175.285    -7.223   174.815  2765.949
##
## Coefficients:
##             Value      Std. Error t value
## (Intercept) -6786.1671   136.3084   -49.7854
## Gestazione     29.6185     3.8041     7.7859
## Lunghezza      10.8673     0.3014    36.0519
## Cranio         10.2715     0.4252    24.1545
```

```
## SessoM         80.3030    11.2368     7.1464
##
## Residual standard error: 259.6 on 2308 degrees of freedom
```

```
hweights=data.frame(Peso=data$Peso,  #Observation
                    resid = mod3.2$resid, #Residual
                    weight = mod3.2$w) #Weight
hweights=hweights[order(mod3.2$w), ]
head(hweights)
```

```
##      Peso     resid    weight
## 1437 4370  2765.949 0.1262227
## 144  3610  1404.438 0.2485830
## 1210 4900  1313.951 0.2656901
## 1297 2560 -1149.691 0.3036525
## 1569 3850  1140.293 0.3061595
## 1874 4650  1022.342 0.3414752
```

```
hweights["1437",]
```

```
##      Peso    resid    weight
## 1437 4370 2765.949 0.1262227
```

```
hweights["1551",]
```

```
##      Peso    resid weight
## 1551 2550 26.96245      1
```

```
# notare che il modello mod3.1 ha il peso più basso in corrispondenza
# dell'osservazione 1437, ma peso 1 in corrispondenza dell'osservazione 1551:
# la mia interpretazione è che l'osservazione 1551 diventa un outlier solo
quando l'osservazione
# 1437 viene rimossa

# PUNTO 6
AIC(mod3,mod3.2)
```

```
##        df      AIC
## mod3    6 32592.92
## mod3.2  6 32601.88
```

```
BIC(mod3,mod3.2)
```

```
##        df      BIC
## mod3    6 32627.40
## mod3.2  6 32636.35
```

```
# secondo i criteri AIC e BIC, il modello 3.2 è peggiore rispetto al modello
3,
# ma abbiamo già osservato che il modello 3.1 è migliore rispetto al modello
3
```

```
mod3$coefficients
```

```
## (Intercept)  Gestazione   Lunghezza      Cranio      SessoM
## -6686.99643    32.71141    10.09709    10.77613    77.72154
```

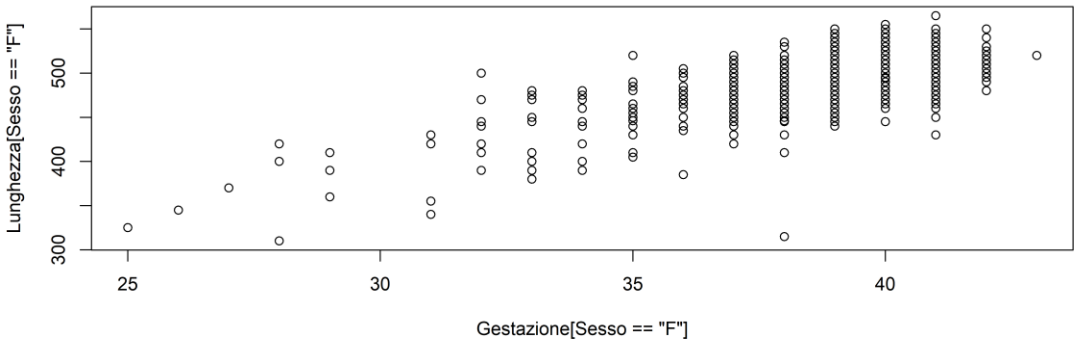
```
mod3.1$coefficients
```

```
## (Intercept)  Gestazione   Lunghezza      Cranio      SessoM
## -6686.68851    29.71380    10.77803    10.12484    77.70652
```

```
mod3.2$coefficients
```

```
## (Intercept)  Gestazione   Lunghezza      Cranio      SessoM
## -6786.16709    29.61854    10.86725    10.27147    80.30299
```

```
# analizzo i coefficienti per verificare quale eventuale modello stima il
valore più basso/alto
# di peso dei neonati, poichè, a seconda delle esigenze dello studio, può
essere più utile stimare un valore
# più alto piuttosto che più basso.
# I coefficienti non crescono/decrescono tutti allo stesso modo da un modello
a un altro
# (per esempio, il coefficiente per la variabile Gestazione decresce dal
modello 3 al modello 3.2,
# mentre il coefficiente per la variabile Lunghezza cresce), per cui in
generale
# il modello che fornisce il peso più alto/basso può variare a seconda dei
dati specifici.


# PUNTO 7
plot(x=Gestazione[Sesso=="F"],
     y=Lunghezza[Sesso=="F"])
```




```
new_born_len_med =median(Lunghezza[Sesso=="F" & Gestazione==39]) # 495 mm
new_born_len_med
```

```
## [1] 495
```

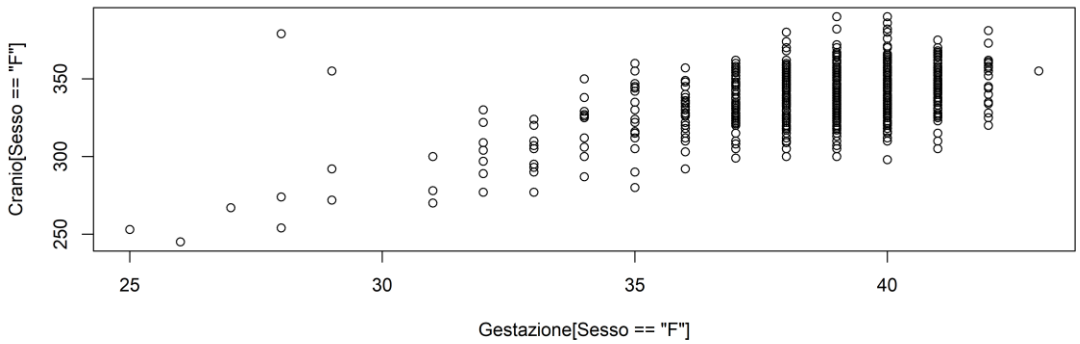
```
new_born_len_min =min(Lunghezza[Sesso=="F" & Gestazione==39]) # 440 mm
new_born_len_min
```

```
## [1] 440
```

```
new_born_len_max =max(Lunghezza[Sesso=="F" & Gestazione==39]) # 550 mm
new_born_len_max
```

```
## [1] 550
```

```
# nel nostro dataset, la lunghezza mediana di una neonata nata alla 39esima
settimana è di 495 mm
plot(x=Gestazione[Sesso=="F"],
     y=Cranio[Sesso=="F"])
```




```
new_born_head_med=median(Cranio[Sesso=="F" & Gestazione==39]) # 340 mm
new_born_head_med
```

```
## [1] 340
```

```
new_born_head_min=min(Cranio[Sesso=="F" & Gestazione==39]) # 300 mm
new_born_head_min
```

```
## [1] 300
```

```
new_born_head_max=max(Cranio[Sesso=="F" & Gestazione==39]) # 390 mm
new_born_head_max
```

```
## [1] 390
```

```
# nel nostro dataset, il diametro mediano del cranio di una neonata nata alla
39esima settimana è di 340 mm
new_born_med=data.frame(Gestazione=39,Sesso="F",Lunghezza=new_born_len_med,Cr
anio=new_born_head_med)
new_born_estim_weigth_med=predict(mod3.1, newdata = new_born_med) # 3250 g
print(new_born_estim_weigth_med)
```

```
##        1
## 3249.719

new_born_min=data.frame(Gestazione=39,Sesso="F",Lunghezza=new_born_len_min,Cr
anio=new_born_head_min)
new_born_estim_weigth_min=predict(mod3.1, newdata = new_born_min) # 2252 g
print(new_born_estim_weigth_min)

##        1
## 2251.934

new_born_max=data.frame(Gestazione=39,Sesso="F",Lunghezza=new_born_len_max,Cr
anio=new_born_head_max)
new_born_estim_weigth_max=predict(mod3.1, newdata = new_born_max) # 4349 g
print(new_born_estim_weigth_max)

##        1
## 4348.753

lett_median_weigth=3000
(new_born_estim_weigth_med-lett_median_weigth)/(max(Peso)-min(Peso))*100 #
differenza tra le due stime pari a circa il 6% del range

##        1
## 6.090714

(new_born_estim_weigth_min-new_born_estim_weigth_med)/(max(Peso)-
min(Peso))*100 # differenza tra le due stime pari a circa il 24% del range

##         1
## -24.33622

(new_born_estim_weigth_max-new_born_estim_weigth_med)/(max(Peso)-
min(Peso))*100 # differenza tra le due stime pari a circa il 27% del range

##       1
## 26.8057

# https://media.tghn.org/articles/newbornsize.pdf
# I dati reali ricavati su neonati ed esposti nell'articolo sopra, mostrano
che il
# 50esimo percentile (mediana) delle neonate nate alla 39esima settimana di
gestazione
# hanno una lunghezza attorno a 49 cm, circonferenza della testa attorno a
33/34 cm e peso
# attorno a 3 kg. Dunque, i valori di lunghezza e circonferenza della testa
IPOTIZZATI
# al fine di stimare il peso, sono coerenti con i dati della letteratura.

# Il peso stimato risulta superiore al peso mediano della letteratura, ma
solamente di una quantità pari al 6 % del range.
```

```r
# Se la lunghezza e il diametro del cranio sono pari ai valori minimi del range (per neonate alla 39esima settimana) allora il peso scende del 25 % rispetto al range (tanto)

# Se la lunghezza e il diametro del cranio sono pari ai valori massimi del range (per neonate alla 39esima settimana) allora il peso sale del 27 % rispetto al range (tanto)

# Dunque, avere queste due variabili (lunghezza e diametro del cranio) è importante per stimare il peso in maniera più accurata

# https://www.statology.org/how-to-interpret-residual-standard-error/#:~:text=Residual%20standard%20error%20%3D%20%E2%88%9A%CE%A3,total%20number%20of%20model%20parameters.
# https://library.virginia.edu/data/articles/getting-started-with-gamma-regression
# Creo un modello lineare generalizzato suppongendo una distribuzione degli errori di tipo Gamma e una link function di tipo "inverse"
mod3.3=glm(Peso~Gestazione+Lunghezza+Cranio+Sesso,family=Gamma(link="inverse"))
summary(mod3.3)
```

```
##
## Call:
## glm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso,
##     family = Gamma(link = "inverse"))
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.443e-03  1.787e-05   80.76  < 2e-16 ***
## Gestazione  -6.352e-06  4.505e-07  -14.10  < 2e-16 ***
## Lunghezza   -1.054e-06  3.386e-08  -31.12  < 2e-16 ***
## Cranio      -1.061e-06  4.697e-08  -22.58  < 2e-16 ***
## SessoM      -4.842e-06  1.223e-06   -3.96 7.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.009100302)
##
##     Null deviance: 69.553  on 2312  degrees of freedom
## Residual deviance: 21.102  on 2308  degrees of freedom
## AIC: 33110
##
## Number of Fisher Scoring iterations: 4
```

```r
AIC(mod3,mod3.3)
```

```
##        df      AIC
## mod3    6 32592.92
## mod3.3  6 33109.82
```

```r
BIC(mod3,mod3.3)
```

```
##        df     BIC
## mod3    6 32627.4
## mod3.3  6 33144.3
```

```r
# secondo i criteri AIC e BIC, il modello 3.3 è mogliore rispetto al modello
3

new_born_estim_weigth_med=predict(mod3.3, newdata = new_born_med,type =
"response") # 3194 g
new_born_estim_weigth_med
```

```
##        1
## 3194.371
```

```r
(new_born_estim_weigth_med-lett_median_weigth)/(max(Peso)-min(Peso))*100 #
differenza tra le due stime pari a circa il 5% del range
```

```
##        1
## 4.740753
```

```r
new_born_estim_weigth_min=predict(mod3.3, newdata = new_born_min,type =
"response") # 2419 g
new_born_estim_weigth_min
```

```
##       1
## 2418.74
```

```r
new_born_estim_weigth_max=predict(mod3.3, newdata = new_born_max,type =
"response") # 4949 g
new_born_estim_weigth_max
```

```
##        1
## 4949.132
```

```r
(new_born_estim_weigth_min-new_born_estim_weigth_med)/(max(Peso)-
min(Peso))*100 # differenza tra le due stime pari a circa il 19% del range
```

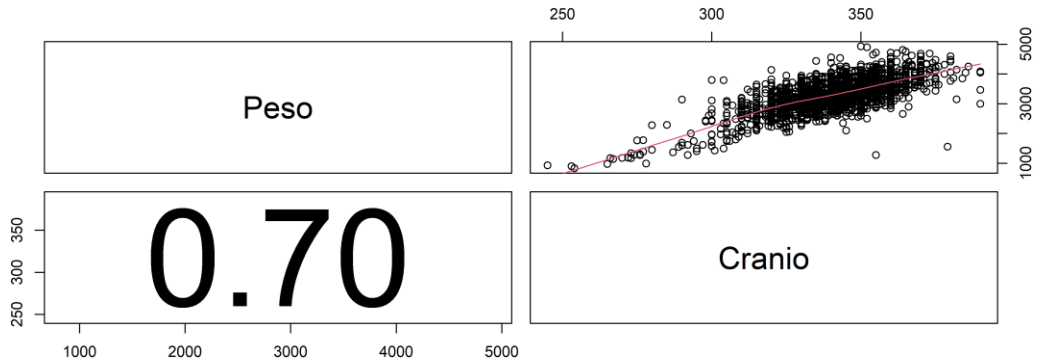
```
##         1
## -18.91781
```

```r
(new_born_estim_weigth_max-new_born_estim_weigth_med)/(max(Peso)-
min(Peso))*100 # differenza tra le due stime pari a circa il 43% del range
```

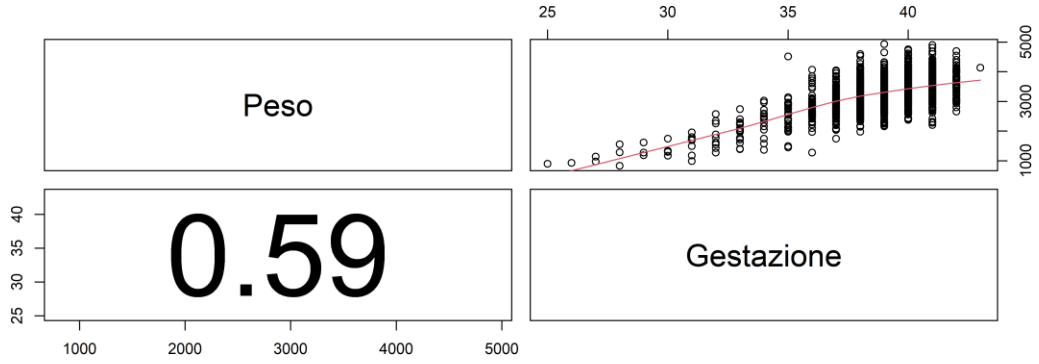
```
##        1
## 42.79904
```

```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]
```

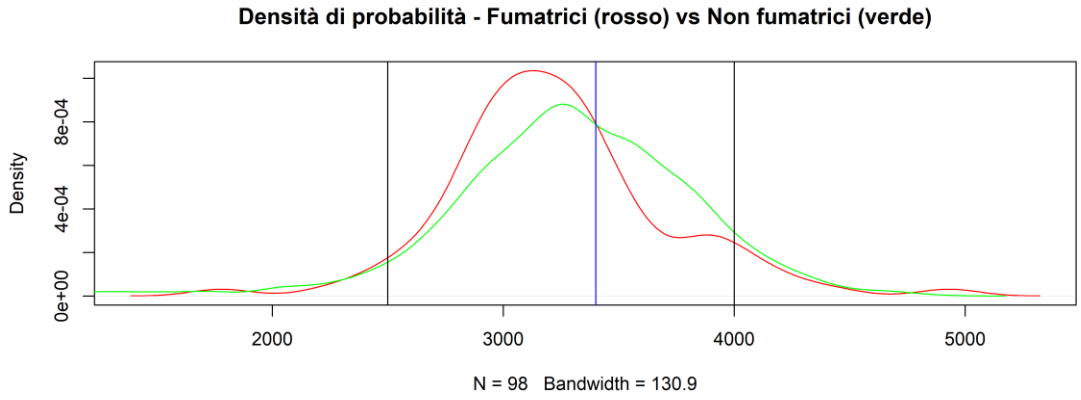
```
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(data_quant[,c(4,6)],upper.panel = panel.smooth,lower.panel = panel.cor)
```




```
# la pendenza sembra cambiare attorno al diametro del cranio di 325 mm,
quindi
# si potrebbero costruire due modelli separati per i due range di diametro
del cranio



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(data_quant[,c(4,3)],upper.panel = panel.smooth,lower.panel = panel.cor)
```

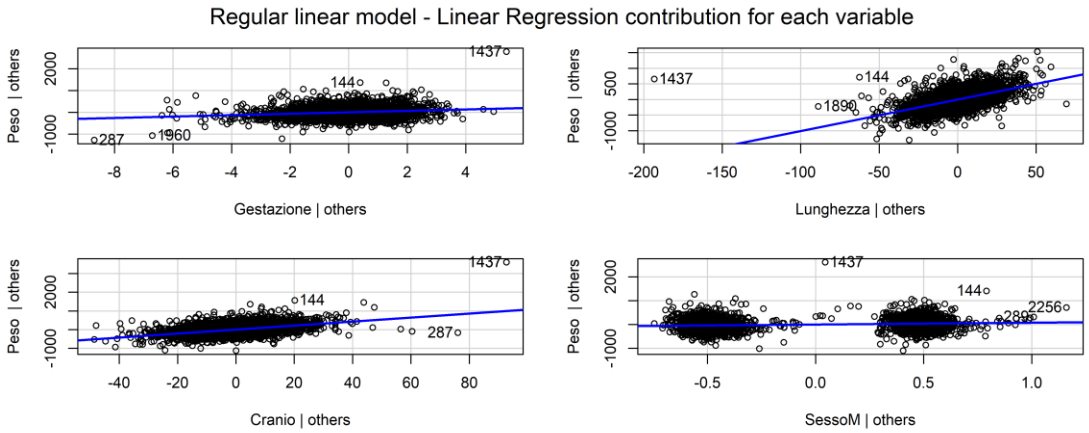



```
# la pendenza sembra cambiare alla settimana 37, quindi
# si potrebbero costruire due modelli separati per i due range di settimane
di gestazione


plot(density(Peso[Fumatrici=="1"]),col="red",main="")
lines(density(Peso[Fumatrici=="0"]),col="green")
title(main="Densità di probabilità - Fumatrici (rosso) vs Non fumatrici
(verde)")
abline(v=c(2500,4000))
abline(v=3400,col="blue")
```

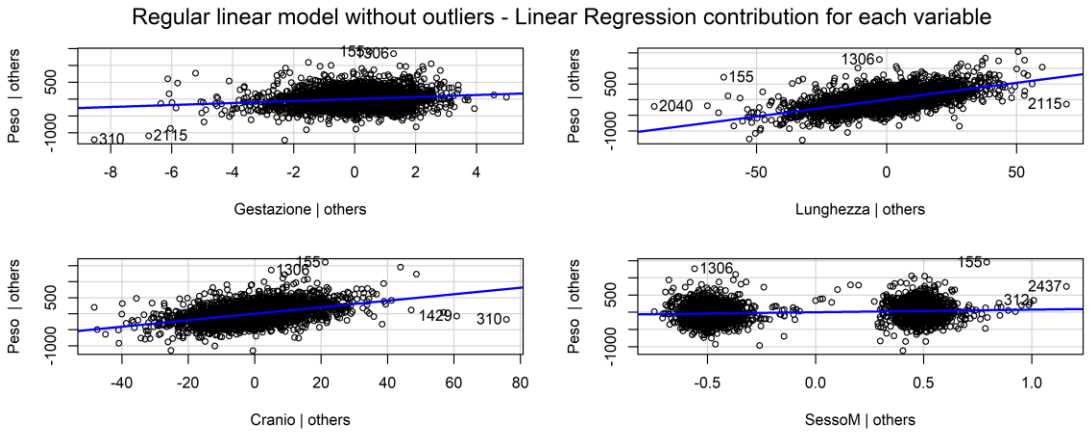



```
# come osservato in precedenza, sembra che la variabile madre fumatrice/non
fumatrice
# abbia un effetto sulla variabile peso.
# E' possibile che modelli differenti da quelli definiti, che contengono la
variabile
# Fumatrice possano tener conto di questo effetto

# PUNTO 8
# https://www.statology.org/plot-multiple-linear-regression-in-r/
avPlots(mod3,
```

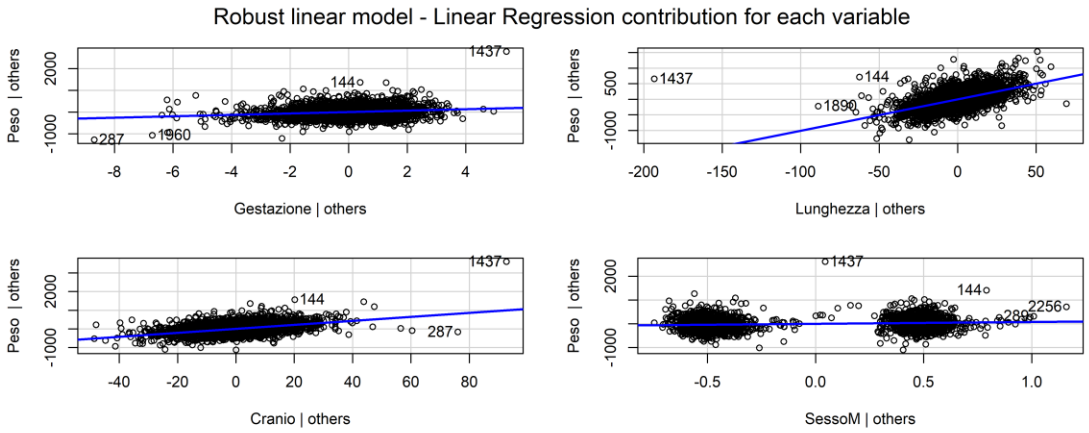
```
        main="Regular linear model - Linear Regression contribution for each
variable")
```




```
avPlots(mod3.1,
        main="Regular linear model without outliers - Linear Regression
contribution for each variable")
```




```
avPlots(mod3.2,
        main="Robust linear model - Linear Regression contribution for each
variable")
```




```r
avPlots(mod3.3,
        main="Generealized linear model without outliers - Linear Regression contribution for each variable")
```