Un modello statistico per prevedere il peso dei neonati

```
options(warn = -1)
#install.packages("ggplot2")
#install.packages("ineq")
#install.packages("car")
#install.packages("MASS")
#install.packages("lmtest")
#install.packages("dplyr")
#tinytex::install_tinytex()
library(ggplot2)
library(ineq)
library(car)
library("MASS")
library(lmtest)
library(dplyr)
```

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

```
df = read.csv("neonati.csv")
```

Controllo se c'è qualche valore NaN

```
 any(is.na(df))
```

```
## [1] FALSE
```

Il DataFrame è stato importato correttamente.

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

Tipologia variabili:

1) Anni.madre: quantitativa discreta (anni della madre)
2) N.gravidanze: quantitativa discreta (numero di gravidanze precedenti)
3) Fumatrici: qualitativa Categoriale (SI, NO)
4) Gestazione: quantitativa discreta (numero di settimane di gestazione)
5) Peso: quantitativa continua (in grammi)
6) Lunghezza: quantitativa continua (in mm)
7) Cranio: quantitativa continua (diametro del cranio, in mm)
8) Tipo.parto: qualitativa Categoriale (Naturale o Cesareo)
9) Ospedale: qualitativa Categoriale (1, 2, 3)
10) Sesso: qualitativa Categoriale (M o F)

Summary variabili quantitative:

```r
df_selected <- select(df, -Fumatrici, -Tipo.parto, -Ospedale)
summary(df_selected)
```

```
##    Anni.madre     N.gravidanze       Gestazione         Peso
##  Min.   : 0.00   Min.   : 0.0000   Min.   :25.00   Min.   : 830
##  1st Qu.:25.00   1st Qu.: 0.0000   1st Qu.:38.00   1st Qu.:2990
##  Median :28.00   Median : 1.0000   Median :39.00   Median :3300
##  Mean   :28.16   Mean   : 0.9812   Mean   :38.98   Mean   :3284
##  3rd Qu.:32.00   3rd Qu.: 1.0000   3rd Qu.:40.00   3rd Qu.:3620
##  Max.   :46.00   Max.   :12.0000   Max.   :43.00   Max.   :4930
##    Lunghezza         Cranio       Sesso
##  Min.   :310.0   Min.   :235   Length:2500
##  1st Qu.:480.0   1st Qu.:330   Class :character
##  Median :500.0   Median :340   Mode  :character
##  Mean   :494.7   Mean   :340
##  3rd Qu.:510.0   3rd Qu.:350
##  Max.   :565.0   Max.   :390
```

La variabile Anni.madre ha dei valori anomali (esempio 0,1..). Tutti i valori inferioria 12 vengono quindi sostituiti con la media degli anni delle madri.

```r
# Calcola la media escludendo i valori inferiori a 12
media_validi <- mean(df$Anni.madre[df$Anni.madre >= 12], na.rm = TRUE)

# Sostituisci i valori inferiori a 12 con la media calcolata
df$Anni.madre[df$Anni.madre < 12] <- media_validi

attach(df)
```

Frequenza delle variabili qualitative;

```r
table(Fumatrici)
```

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

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

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

```r
table(Ospedale)
```

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

Obiettivo: Si vuole scoprire se è possibile prevedere il peso del neonato alla nascita, date tutte le altre variabili.

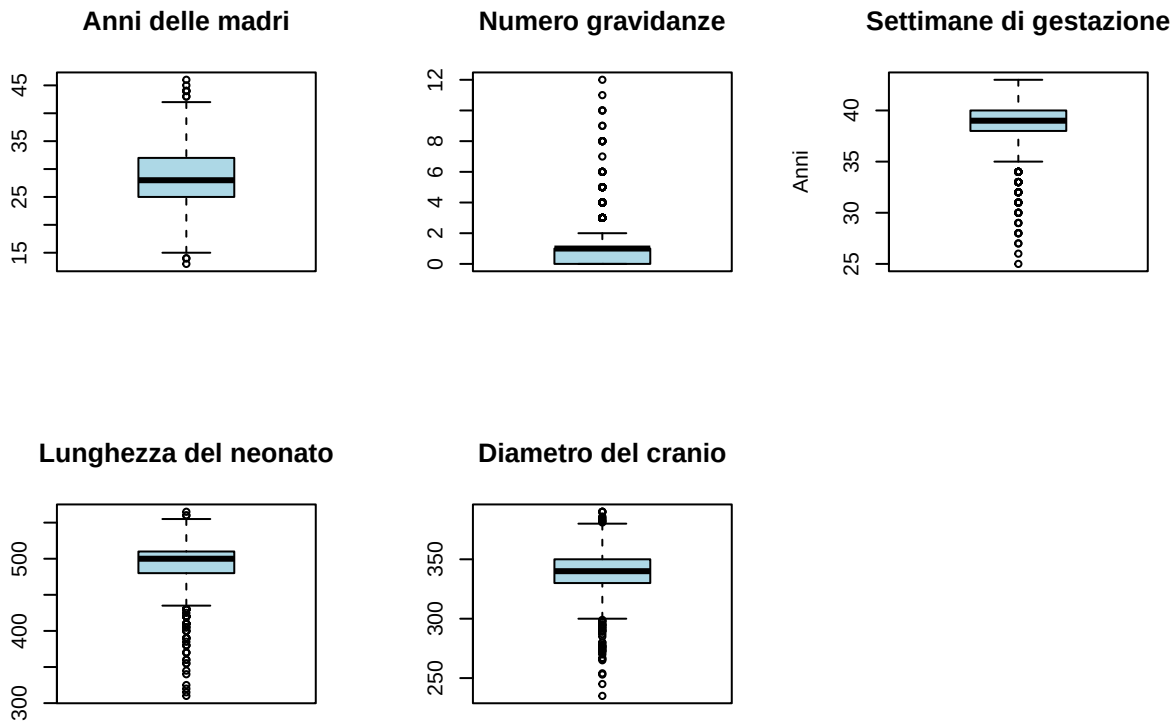
Variabile di risposta: Peso

Variabili esplicative: Anni.madre, N.gravidanze, Fumatrici, Gestazione, Lunghezza, Cranio, Tipo.parto, Ospedale, Sesso, di cui Lunghezza, Cranio, Sesso sono variabili di controllo.

3. Indaga le variabili effettuando una breve analisi descrittiva

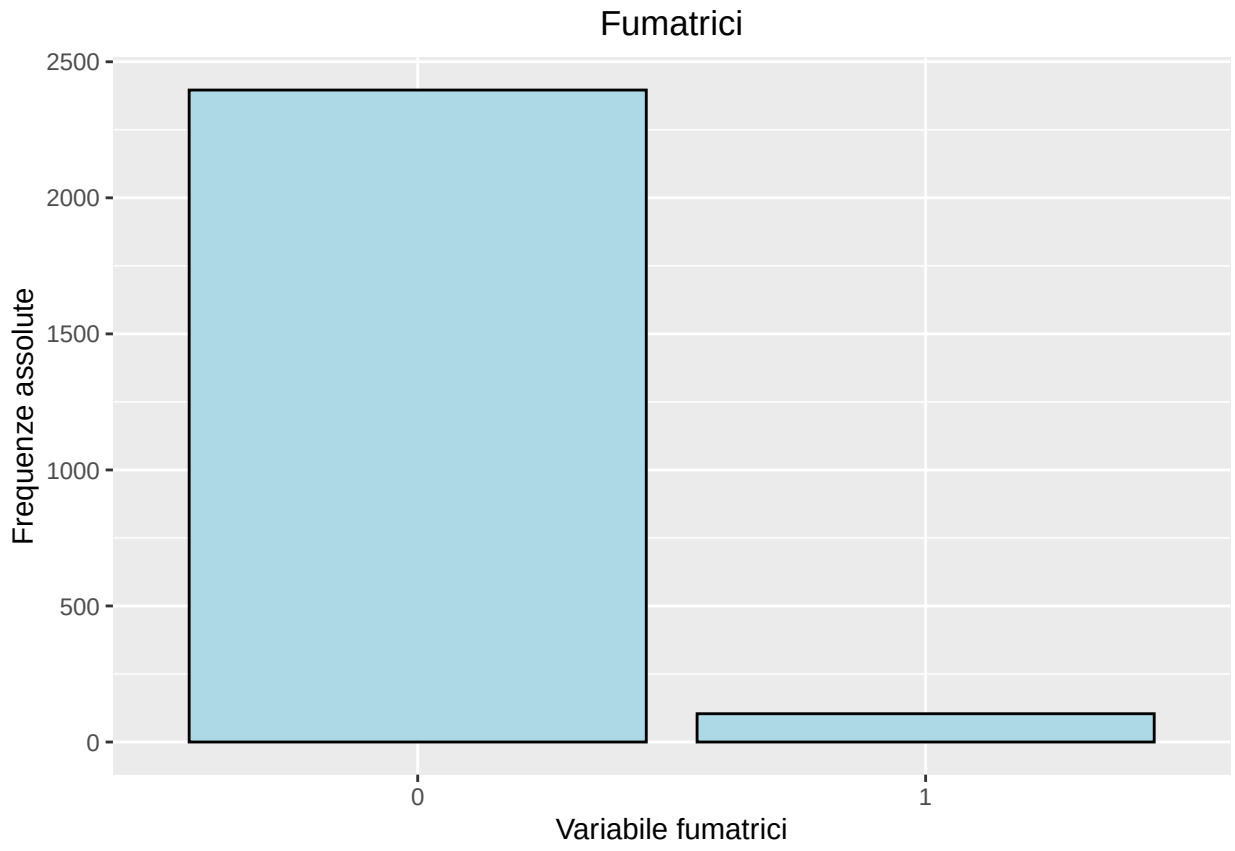
3.1 Analisi dei regressori:

```
par(mfrow = c(2,3))
boxplot(Anni.madre, main = "Anni delle madri", col="lightblue")
boxplot(N.gravidanze, main = "Numero gravidanze", col="lightblue")
boxplot(Gestazione, ylab = "Anni", main = "Settimane di gestazione", col="lightblue")
boxplot(Lunghezza, main = "Lunghezza del neonato", col="lightblue")
boxplot(Cranio, main = "Diametro del cranio", col="lightblue")
```

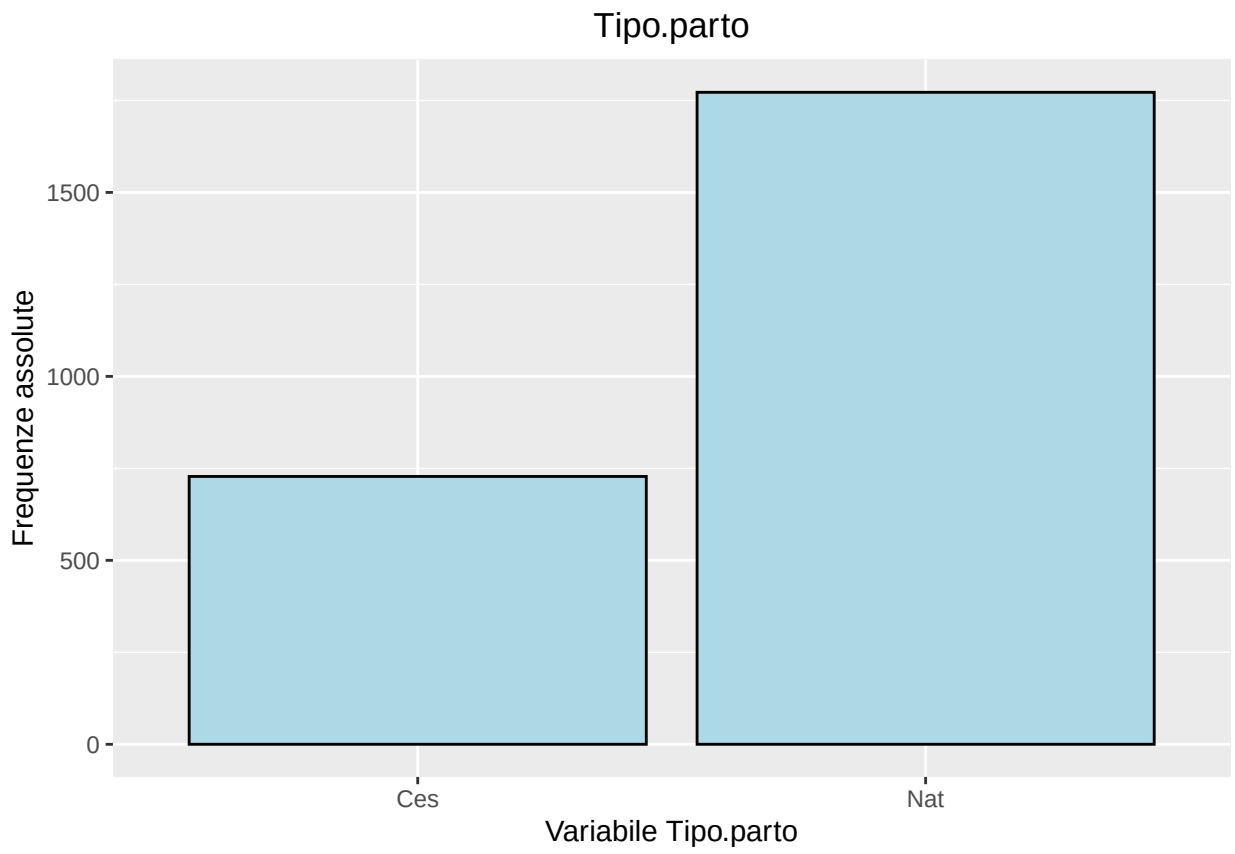



```
ggplot(data = df)+
  geom_bar(
    aes(x= as.factor(Fumatrici)),
    stat = "count",
    fill = "lightblue",
    color = "black")+
  labs(title = "Fumatrici",
```

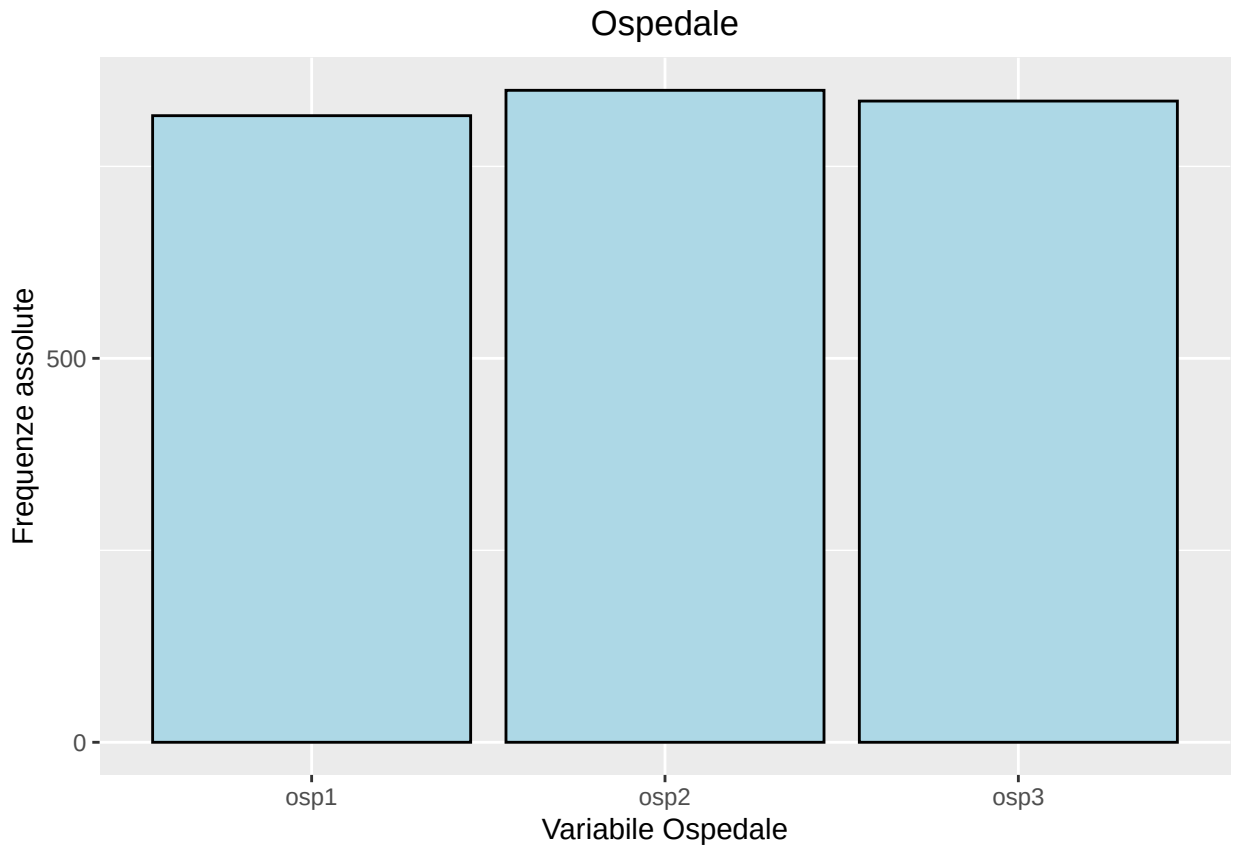
```
      x="Variabile fumatrici",
      y="Frequenze assolute")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(breaks = seq(0,2500,500))
```



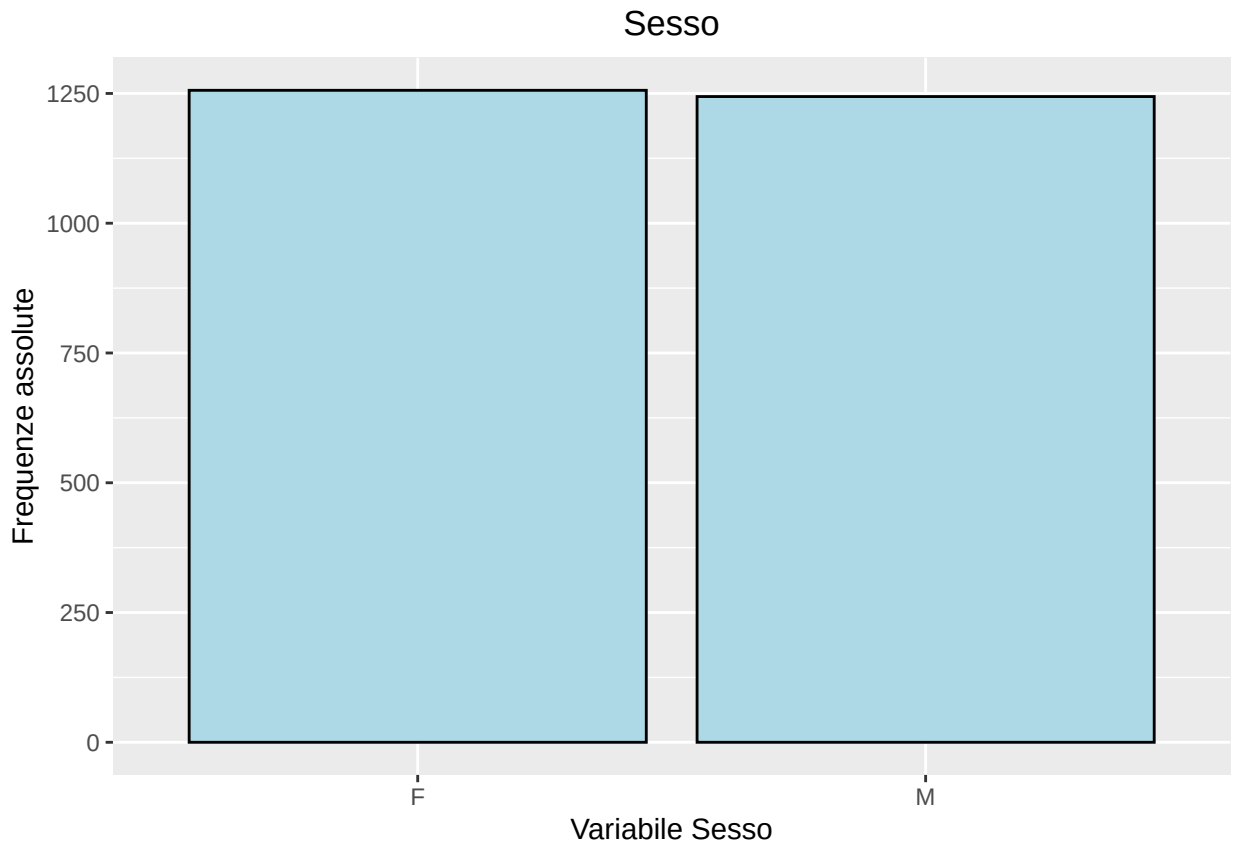

```
ggplot(data = df)+
  geom_bar(
    aes(x=Tipo.parto),
    stat = "count",
    fill = "lightblue",
    color = "black")+
  labs(title = "Tipo.parto",
       x="Variabile Tipo.parto",
       y="Frequenze assolute")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(breaks = seq(0,2500,500))
```




```
ggplot(data = df)+
  geom_bar(
    aes(x=Ospedale),
    stat = "count",
    fill = "lightblue",
    color = "black")+
  labs(title = "Ospedale",
       x="Variabile Ospedale",
       y="Frequenze assolute")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(breaks = seq(0,2500,500))
```




```
ggplot(data = df)+
geom_bar(
  aes(x=Sesso),
  stat = "count",
  fill = "lightblue",
  color = "black")+
labs(title = "Sesso",
     x="Variabile Sesso",
     y="Frequenze assolute")+
theme(plot.title = element_text(hjust = 0.5))+
scale_y_continuous(breaks = seq(0,2500,250))
```



Eccetto le variabili "Fumatrici" e "Tipo.parto" le variabili qualitative presentano una buona omogeneità, ovvero un'equidistribuzione delle classi.

Al contrario la variabile "Tipo.parto" presenta una marcata eterogeneità mentre la variabile "Fumatrici" presenta quasi la massima concentrazione, ovvero tutti i valori concentrati in un'unica classe.

Confrontando i risultati grafici con l'indice di GINI si ottengono le medesime conclusioni:

Fumatrici:

```
source("Utils.R")
indice_gini(Fumatrici)
```

```
## [1] 0.1594778
```

Tipo.parto:

```
indice_gini(Tipo.parto)
```

```
## [1] 0.8256102
```

Ospedale:

```
indice_gini(Ospedale)
```

```
## [1] 0.9998683
```

Sesso:

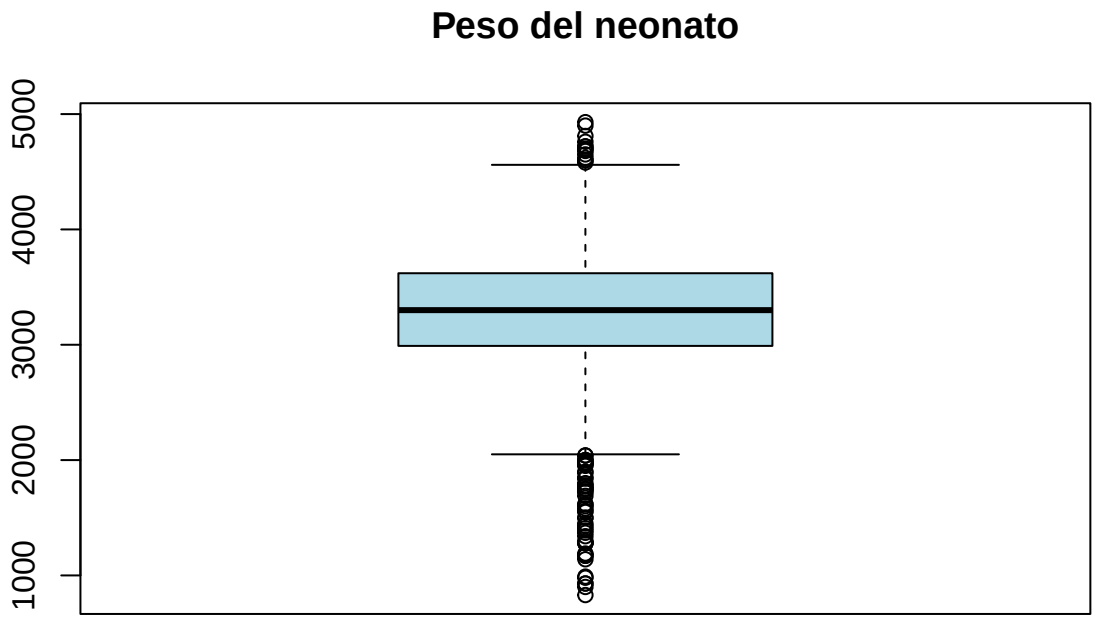
```
indice_gini(Sesso)
```

```
## [1] 0.999977
```

Dove l'indice di GINI assume valore pari a 1 quando le classi sono equidistribuite e al contrario un valore pari a zero quando le classi hanno il massimo grado di eterogeneità.

3.2 Analisi della variabile di risposta:

```
boxplot(Peso, main = "Peso del neonato", col="lightblue")
```



Dal BoxPlot la variabile di risposta presenta un buon grado di simmetria, con la presenza di outlier, soprattutto nella coda sinistra.

Verifichiamo la vicinanza alla distribuzione normale con gli indici di curtosi e simmetria:

```
moments::skewness(Peso)
```

```
## [1] -0.6470308
```

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

```
## [1] 2.031532
```

l'indice di simmetria è prossimo allo zero, ovvero indica una distribuzione simmetrica rispetto alla media.

l'indice di curtosi è lievemente positivo, ovvero i dati presentano una gobba leggermente più alta rispetto alla distribuzione normale con delle code più basse e strette.

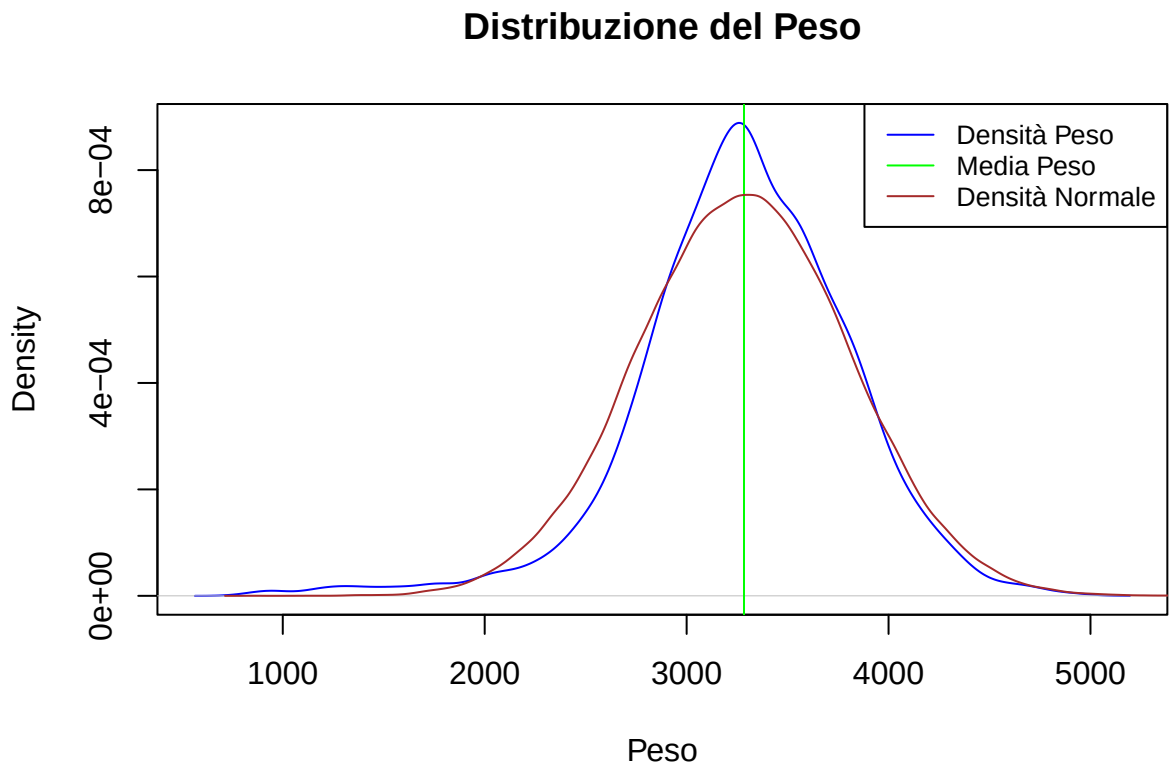
Verifichiamo l'ipotesi che i dati seguano una distribuzione Normale con il Test d'ipotesi di Shapiro-Wilk.

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

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

l' ipotesi di normalità viene rifiutata in quanto il p-value è inferiore al livello di significatività alpha=0.05.

```
plot(density(Peso), main = "Distribuzione del Peso", xlab = "Peso", col = "blue")
abline(v = mean(Peso), col = "green")
lines(density(rnorm(100000, mean = mean(Peso), sd = sd(Peso))), col = "brown")
legend("topright", legend = c("Densità Peso", "Media Peso", "Densità Normale"),
       col = c("blue", "green", "brown"), lty = 1:1, cex = 0.8)
```



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

media peso popolazione: circa 3300 gr. media lunghezza popolazione: circa 50 cm.

Essendo che le variabili in questione non seguono una distribuzione normale si opta per usare il test-t che è adatto a questo tipo di situazioni. Infatti il test-t, per verificare l'ipotesi tra media del campione e parametro sotto H0 può essere usato nei seguenti casi:

1 - campione piccolo 2 - varianza popolazione non nota 3 - incertezza sul modello che segue i dati 4 - distribuzione non normale

Test per la verifica del peso:

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

```
##
##  One Sample t-test
##
## data:  Peso
## t = -1.516, df = 2499, p-value = 0.1296
## alternative hypothesis: true mean is not equal to 3300
## 95 percent confidence interval:
##  3263.490 3304.672
## sample estimates:
## mean of x
##  3284.081
```

p-value = 0.1287. quindi non si rifiuta l'ipotesi che la media del peso dei neonati è statisticativamente uguale a quella della popolazione (3300 gr).

Test per la verifica della lunghezza:

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

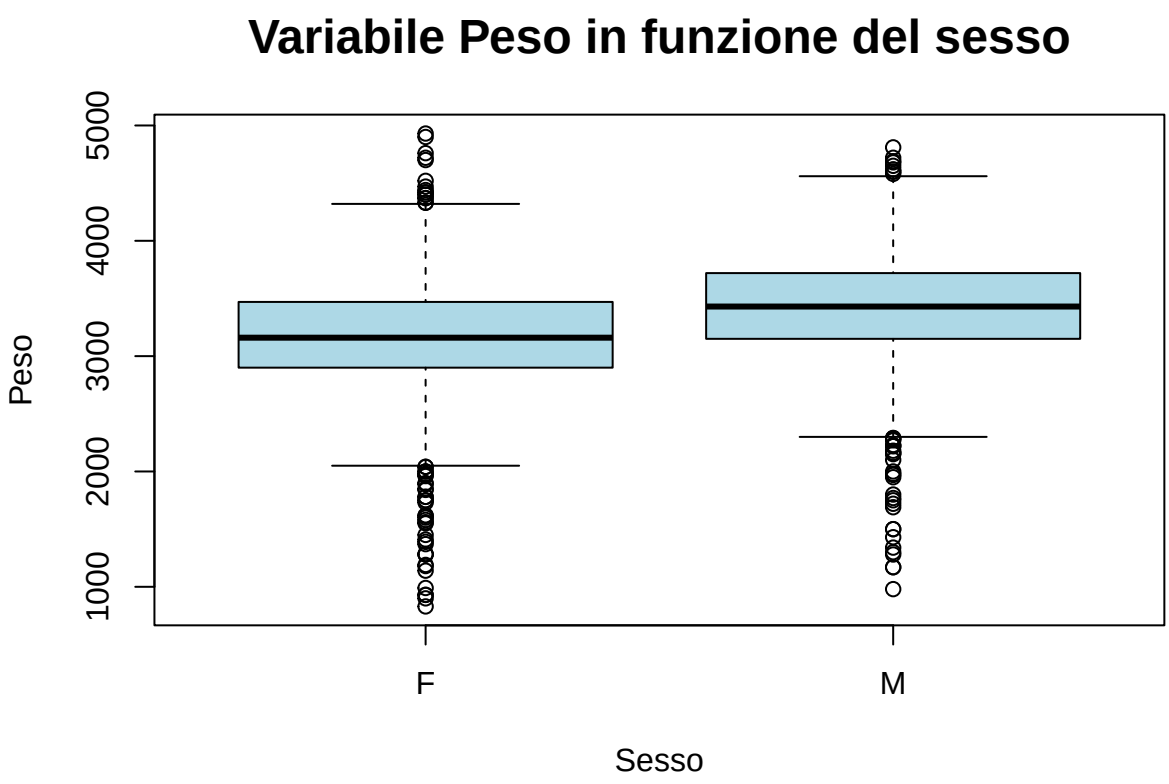
```
##
##  One Sample t-test
##
## data:  Lunghezza
## t = -10.084, df = 2499, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
##  493.6598 495.7242
## sample estimates:
## mean of x
##   494.692
```

p-value < 2.2e-16. quindi si rifiuta l'ipotesi che la media della lunghezza dei neonati è statisticativamente uguale a quella della popolazione (500 mm). I dati della popolazione sono comunque leggermente variabili quindi è possibile riscontrare incongruenze a causa di questa non univocità.

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

-> Relazione Peso-Sesso:

```r
boxplot(Peso ~ Sesso, col="lightblue")
title(main = "Variabile Peso in funzione del sesso", cex.main = 1.5)
```



verifica assunzioni per test-t d'ipotesi tra gruppi indipendenti:

```r
shapiro.test(Peso[Sesso=="M"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Peso[Sesso == "M"]
## W = 0.96647, p-value = 2.321e-16
```

```r
shapiro.test(Peso[Sesso=="F"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Peso[Sesso == "F"]
## W = 0.96285, p-value < 2.2e-16
```

Entrambe le variabili non sono distribuite normalmente quindi non si può usare il Test-t per confrontare medie di gruppi diversi.

Si utilizza allora un test NON parametrico per esempio il Wilcoxon e Mann-Whitney test.

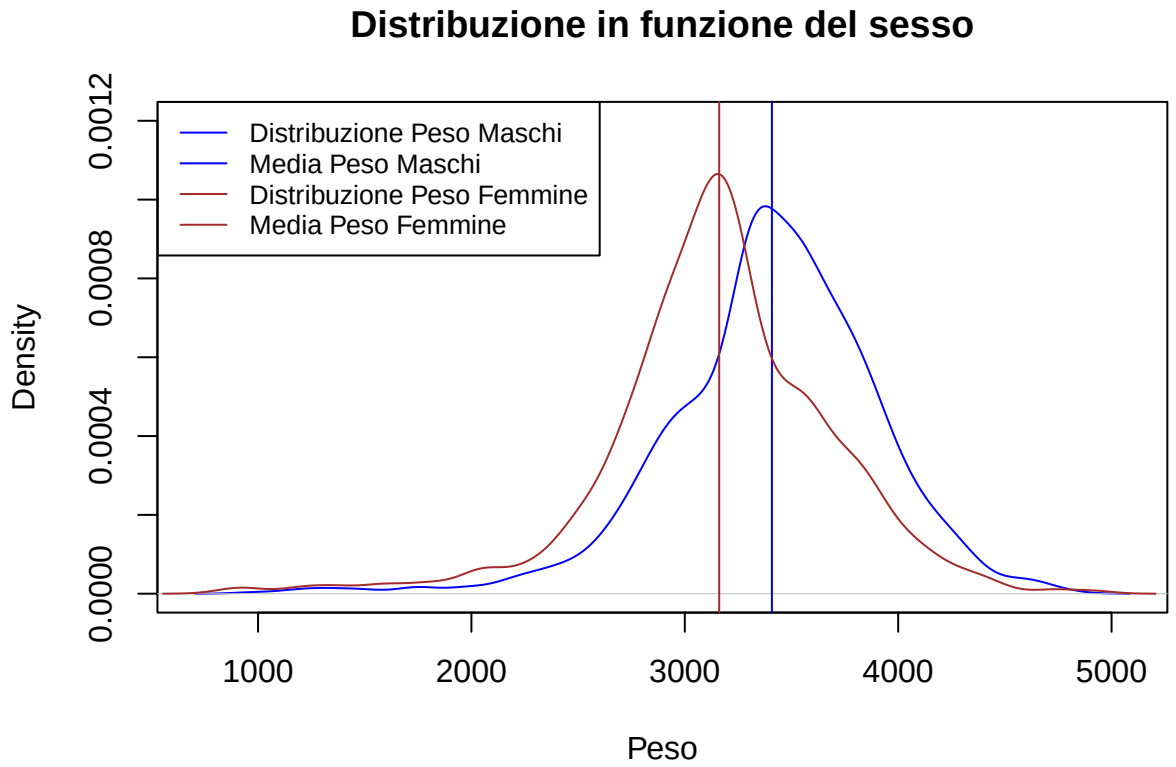
```r
wilcox.test(Peso[Sesso=="M"], Peso[Sesso=="F"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Peso[Sesso == "M"] and Peso[Sesso == "F"]
## W = 1023824, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

viene rigettata l'ipotesi nulla, quindi i due gruppi differiscono significativamente nella media.

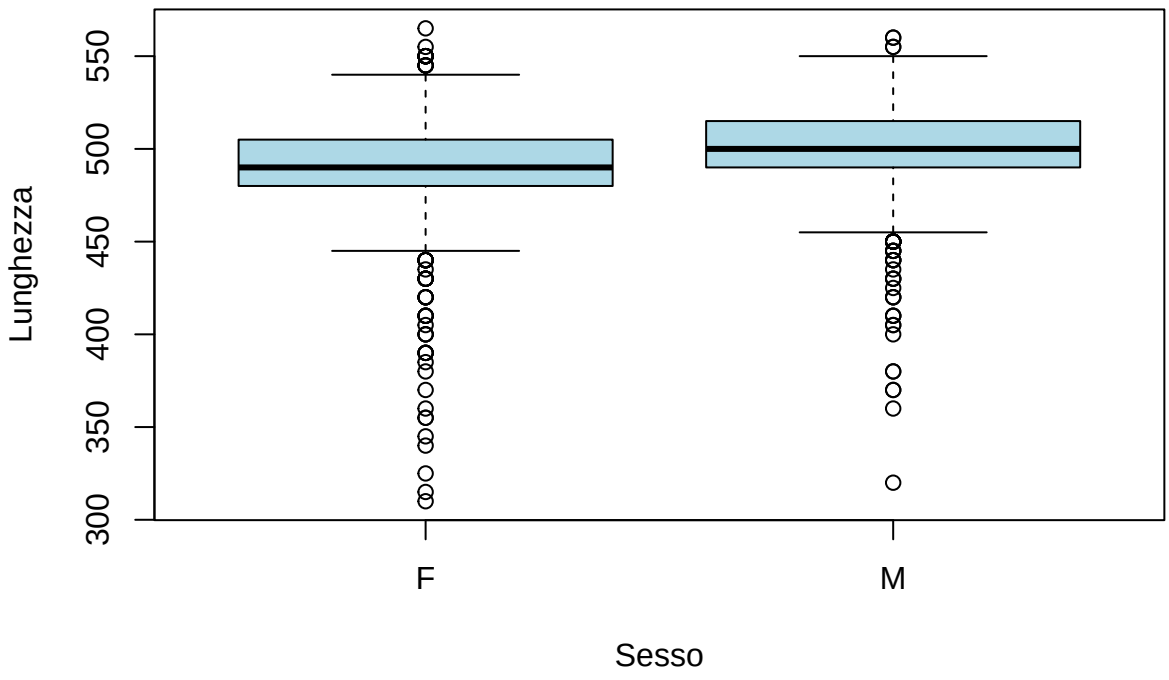
Per completezza si riporta il grafico delle due distribuzioni:

```r
plot(density(Peso[Sesso=="M"]), main = "Distribuzione in funzione del sesso", xlab = "Peso", ylim = c(0
abline(v = mean(Peso[Sesso=="M"]), col = "blue")
lines(density(Peso[Sesso=="F"]), col = "brown")
abline(v = mean(Peso[Sesso=="F"]), col = "brown")
legend("topleft", legend = c("Distribuzione Peso Maschi", "Media Peso Maschi", "Distribuzione Peso Femm
       col = c("blue", "blue", "brown", "brown"), lty = 1:1, cex = 0.8)
```



-> Relazione Lunghezza-Sesso:

```r
boxplot(Lunghezza ~ Sesso, col="lightblue")
title(main = "Variabile Lunghezza in funzione del sesso", cex.main = 1.5)
```




```
shapiro.test(Lunghezza[Sesso=="M"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Lunghezza[Sesso == "M"]
## W = 0.92028, p-value < 2.2e-16
```

```
shapiro.test(Lunghezza[Sesso=="F"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Lunghezza[Sesso == "F"]
## W = 0.89953, p-value < 2.2e-16
```

Anche in questo caso le variabili non sono distribuite normalmente quindi non si può usare il Test-t per confrontare medie di gruppi diversi.

Si userà quindi il test NON parametrico di Wilcoxon e Mann-Whitney per gruppi indipendenti.

```
wilcox.test(Lunghezza[Sesso=="M"], Lunghezza[Sesso=="F"])
```

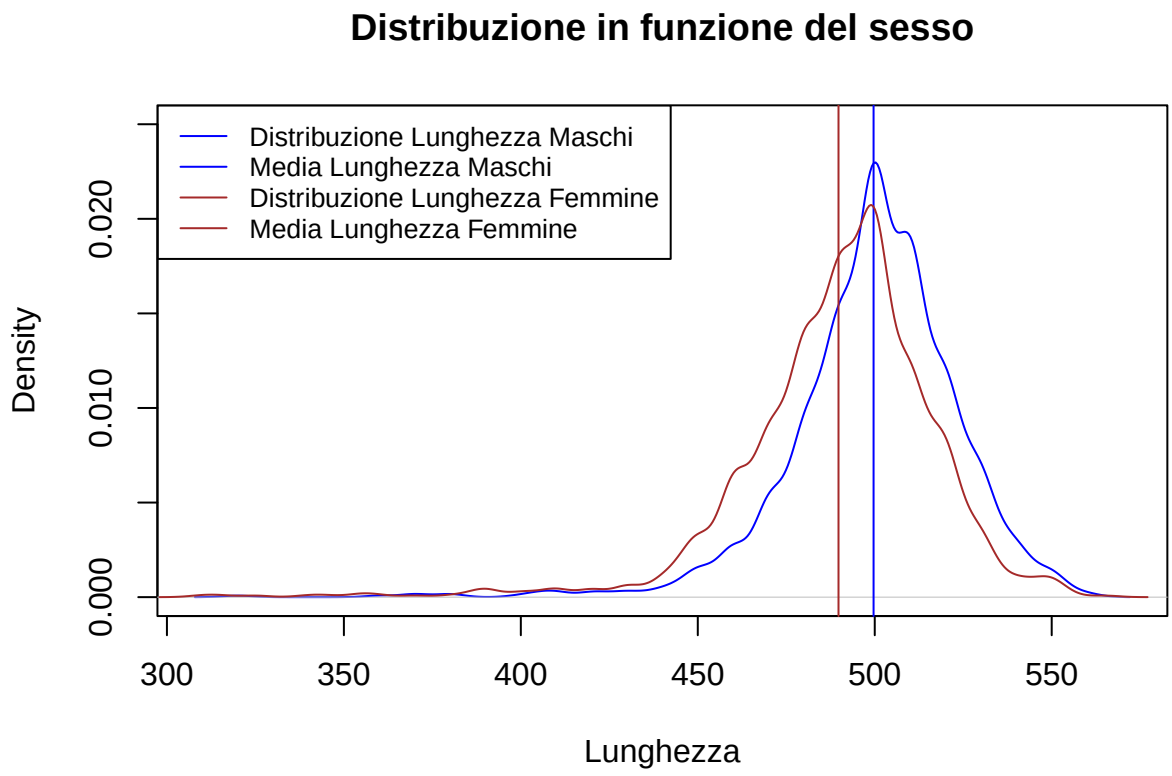
```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  Lunghezza[Sesso == "M"] and Lunghezza[Sesso == "F"]
## W = 968010, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

anche in questo caso viene rigettata l'ipotesi nulla, quindi i due gruppi differiscono significativamente nella media.

Per completezza si riporta il grafico delle due distribuzioni:

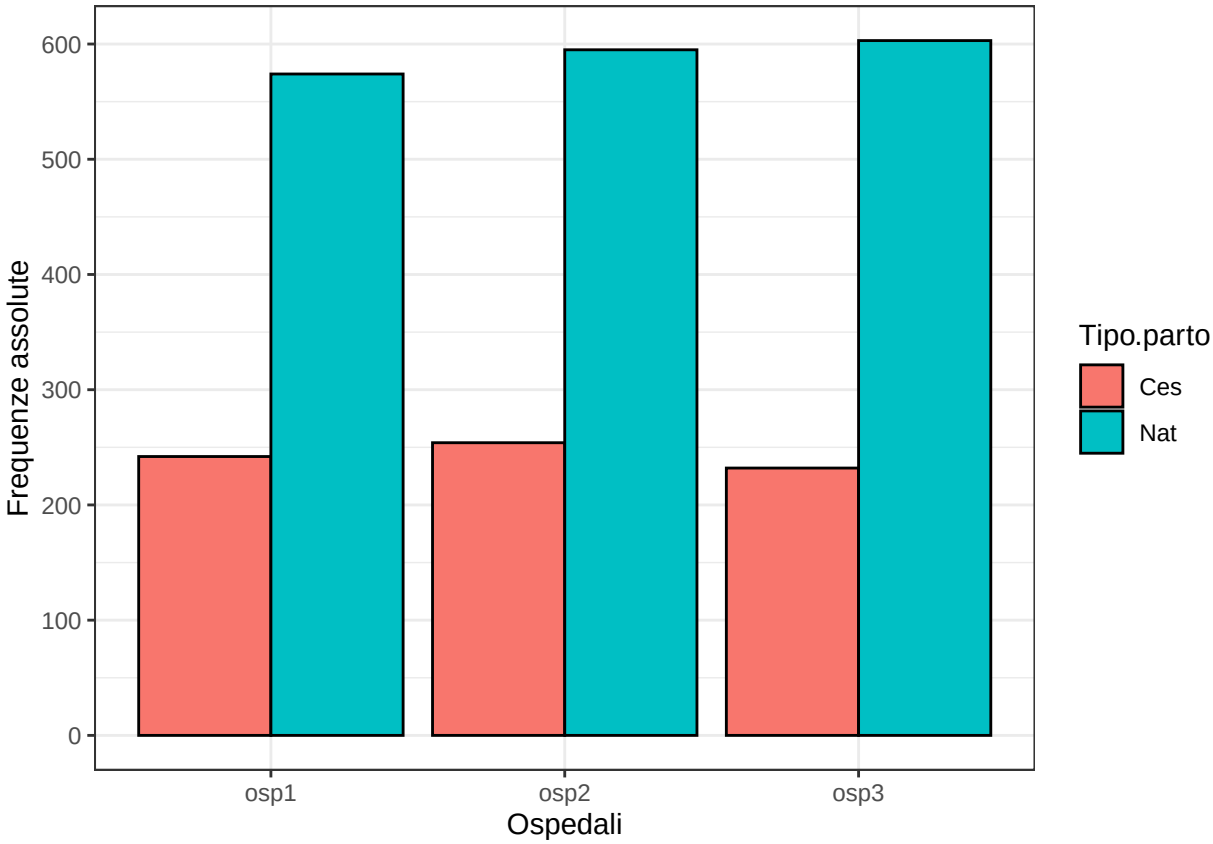
```
plot(density(Lunghezza[Sesso=="M"]), main = "Distribuzione in funzione del sesso", xlab = "Lunghezza", y
abline(v = mean(Lunghezza[Sesso=="M"]), col = "blue")
lines(density(Lunghezza[Sesso=="F"]), col = "brown")
abline(v = mean(Lunghezza[Sesso=="F"]), col = "brown")
legend("topleft", legend = c("Distribuzione Lunghezza Maschi", "Media Lunghezza Maschi", "Distribuzione
       col = c("blue", "blue", "brown", "brown"), lty = 1:1, cex = 0.8)
```



6. Si vocifera che in alcuni ospedali si facciano più parti cesarei, verifichiamo questa ipotesi.

```
ggplot(data = df)+
  geom_bar(
    aes(x=Ospedale, fill = Tipo.parto),
    position = "dodge",
    sta = "count",
```

```
    color = "black")+
  labs(x="Ospedali",
       y="Frequenze assolute")+
  theme_bw()+
  scale_y_continuous(breaks = seq(0,1500,100))
```



Si può notare una leggera differenza tra i 3 ospedali con "osp2" in testa seguito da "osp1" e "osp3", si verifica ora che queste differenze siano statisticamente significative.

Il test utilizzato è X^2 che ha una distribuzione chi-quadro con (N-1)*(M-1) gradi di libertà dove N e M sono rispettivamente righe e colonne della tabella di contingenza cosi creata.

```
tab_contingenza = table(Tipo.parto, Ospedale)["Ces", ]
tab_contingenza
```

```
## osp1 osp2 osp3
##  242  254  232
```

ovvero N=1 e M=3.

```
chisq.test(tab_contingenza)
```

```
##
##  Chi-squared test for given probabilities
##
```

```
## data:  tab_contingenza
## X-squared = 1, df = 2, p-value = 0.6065
```

quindi NON si rifiuta l'ipotesi nulla che le 3 frequenze di parti cesari provengano dalla stessa distribuzione di conseguenza le differenze osservate graficamente non sono statisticamente significative potendo concludere che le voci son false.

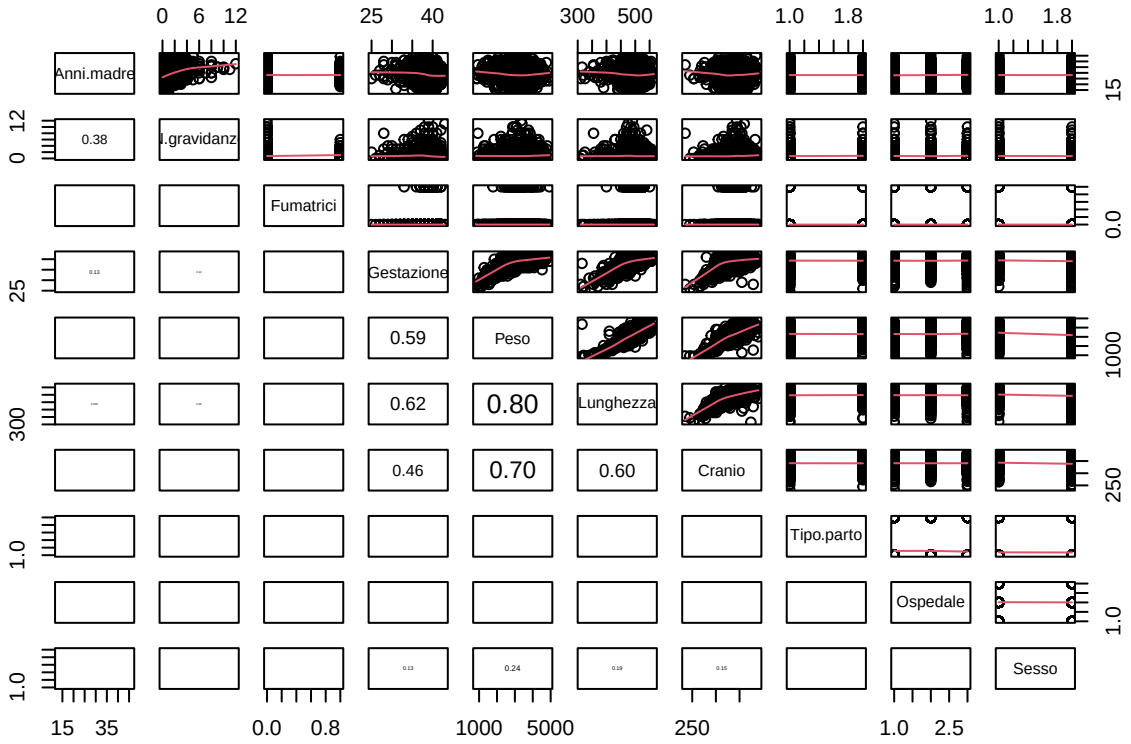
PARTE 2 - ANALISI MULTIDIMENSIONALE -

1. Indaga le relazioni a due a due, soprattutto con la variabile risposta.

```
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)
}


mapping_numerico = c("Nat" = 1, "Ces" = 2)

df$Tipo.parto = as.numeric(factor(Tipo.parto, levels = names(mapping_numerico), labels = mapping_numeri

mapping_numerico2 = c("osp1" = 1, "osp2" = 2, "osp3" = 3)

df$Ospedale = as.numeric(factor(Ospedale, levels = names(mapping_numerico2), labels = mapping_numerico2

mapping_numerico3 = c("M" = 1, "F" = 2)

df["Sesso"] = as.numeric(factor(Sesso, levels = names(mapping_numerico3), labels = mapping_numerico3))

attach(df)
```

```
## I seguenti oggetti sono mascherati da df (pos = 3):
##
##     Anni.madre, Cranio, Fumatrici, Gestazione, Lunghezza, N.gravidanze,
##     Ospedale, Peso, Sesso, Tipo.parto
```

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



Come si può notare dal grafco appena riportato la variabile di risposta Peso sembra essere correlata a 3 variabili in particolare:

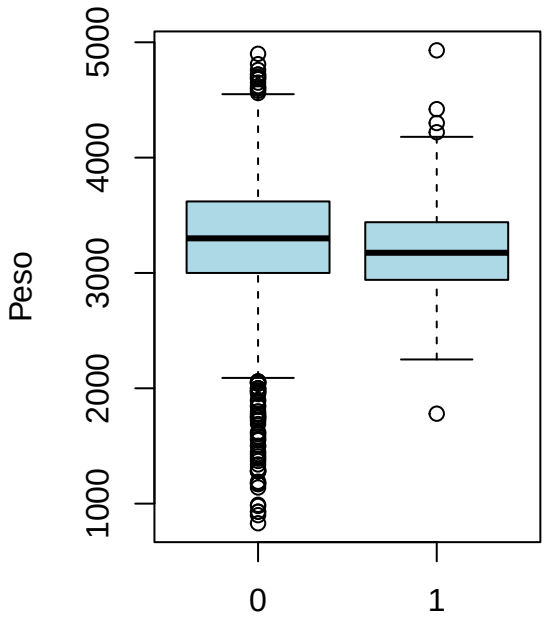
-> Lunghezza: con una correlazione pari a 0.80 -> Cranio: con una correlazione pari a 0.70 -> Gestazione: con una correlazione pari a 0.59

tutte e 3 variabili quantitative.

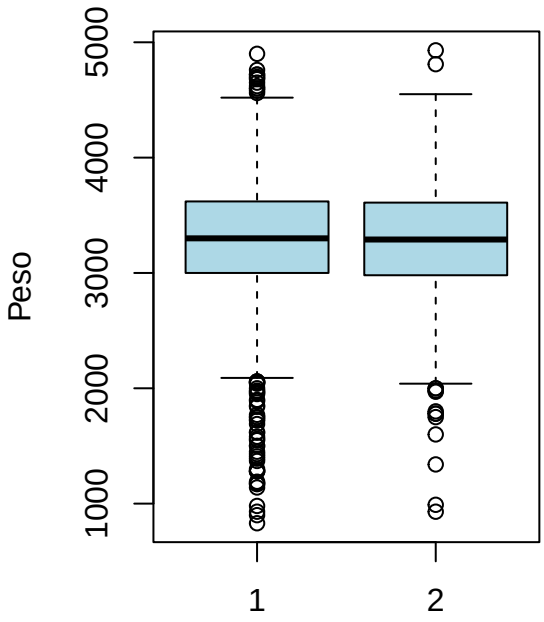
Per le variabili qualitative lo scatterplot non rappresenta un buon strumento di visualizzazione, tanto meno le indicazioni della correlazione lineare circa eventuali associazioni.

Si usa perciò il boxplot condizionato per visualizzare eventuali dipendenze.

```
par(mfrow = c(1,2))
boxplot(Peso~Fumatrici, xlab = "Relazione peso-fumatrici", col="lightblue")
boxplot(Peso~Tipo.parto, xlab = "Relazione peso-tipo parto", col="lightblue")
```



Relazione peso−fumatrici



Relazione peso−tipo parto

Dall'analisi grafica non sembrano esserci differenze significative del peso in funzione del tipo di parto ma si possono notare delle leggere fluttuazioni in funzione del fatto che la madre fumi o meno. Verifichiamo quindi con il test d'ipotesidi Wilcoxon e Mann-Whitney.

```
wilcox.test(Peso[Fumatrici=="0"], Peso[Fumatrici=="1"], mu=0)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Peso[Fumatrici == "0"] and Peso[Fumatrici == "1"]
## W = 138162, p-value = 0.05971
## alternative hypothesis: true location shift is not equal to 0
```

Non si rifiuta l'ipotesi nulla anche se siamo sulla zona di confine. L'ipotesi nulla (H0) del test di Wilcoxon-Mann-Whitney è che le due distribuzioni sono stocasticamente uguali, cioè che non ci sono differenze significative tra i gruppi. Quindi si porrà maggiore attenzione alla variabile Fumatrici in seguito.

```
wilcox.test(Peso[Tipo.parto=="1"], Peso[Tipo.parto=="2"], mu=0)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Peso[Tipo.parto == "1"] and Peso[Tipo.parto == "2"]
## W = 655268, p-value = 0.5315
## alternative hypothesis: true location shift is not equal to 0
```

Come ci si aspettava non si rifiuta l ipotesi nulla quindi la variabile Tipo.parto non influenza la variabile Peso.

Non viene valutata la correlazione Peso-Ospedale perchè non ha senso.

Per quanto riguarda le variabili esplicative invece si nota una leggera correlazione tra:

-> Gestazione e Lunghezza pari a: 0.62 -> Gestazione e Cranio pari a: 0.46 -> Lunghezza e Cranio pari a: 0.60

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

Vengono ora tolte in modo casuale 10 osservazioni dal dataset che verranno usate alla fine per la fase di testing.

```
df = read.csv("neonati.csv")
df = df[-27,]
df = df[-73,]
df = df[-516,]
df = df[-812,]
df = df[-1111,]
df = df[-1315,]
df = df[-1717,]
df = df[-1899,]
df = df[-2301,]
df = df[-2400,]
attach(df)
```

```
## I seguenti oggetti sono mascherati da df (pos = 3):
##
##     Anni.madre, Cranio, Fumatrici, Gestazione, Lunghezza, N.gravidanze,
##     Ospedale, Peso, Sesso, Tipo.parto
```

```
## I seguenti oggetti sono mascherati da df (pos = 4):
##
##     Anni.madre, Cranio, Fumatrici, Gestazione, Lunghezza, N.gravidanze,
##     Ospedale, Peso, Sesso, Tipo.parto
```

```
modello_1 = lm(Peso ~ ., data=df)
summary(modello_1)
```

```
##
## Call:
## lm(formula = Peso ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1125.21  -181.44   -14.89   161.49  2613.53
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6734.6476   141.5980 -47.562  < 2e-16 ***
```

```
## Anni.madre        0.8680     1.1347   0.765   0.4443
## N.gravidanze     11.1823     4.6675   2.396   0.0167 *
## Fumatrici       -27.7597    27.6865  -1.003   0.3161
## Gestazione       32.6044     3.8274   8.519  < 2e-16 ***
## Lunghezza        10.2937     0.3010  34.197  < 2e-16 ***
## Cranio           10.4560     0.4266  24.512  < 2e-16 ***
## Tipo.partoNat    29.1952    12.1184   2.409   0.0161 *
## Ospedaleosp2     -9.9621    13.4856  -0.739   0.4601
## Ospedaleosp3     28.3308    13.5268   2.094   0.0363 *
## SessoM           78.6919    11.2097   7.020 2.85e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.1 on 2479 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.728
## F-statistic: 667.1 on 10 and 2479 DF,  p-value: < 2.2e-16
```

Commenti:

Le variabili: N.gravidanze, Gestazione, Lunghezza, Cranio, Tipo.parto e Sesso hanno superato il test-t, ovvero hanno mostrato un livello di significatività tale da rigettare l ipotesi nulla che siano uguali a zero, di conseguenza sembrano mostrare una buona significatività nello spiegare la varianza della Risposta. In ogni caso vanno ulteriormente indagate con ulteriori test.

Questi risultati sono in linea con l'indagine fatta nel punto 1 dove le variabili Gestazione, Lunghezza e Cranio hanno mostrato una buona relazione con la risposta mentre era assente per le variabili Fumatrici e Tipo.parto.

Avendo ottenuto risultati contrastanti per la variabile Tipo.parto va ulteriormente indagata.

Siamo inoltre in linea col punto 5 della parte 1 dove si era riscontrata una differenza significativa del peso in funzione della variabile Sesso.

```
vif(modello_1)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## Anni.madre   1.186151  1        1.089106
## N.gravidanze 1.185344  1        1.088735
## Fumatrici    1.007062  1        1.003525
## Gestazione   1.694951  1        1.301903
## Lunghezza    2.084776  1        1.443876
## Cranio       1.630492  1        1.276907
## Tipo.parto   1.004158  1        1.002077
## Ospedale     1.003982  2        1.000994
## Sesso        1.040739  1        1.020166
```

Dalla statistica VIF non si rilevano particolari correlazioni tra i regressori, essendo tutte minori di 5.

Infine per quanto la variabilità spiegata dal modello si riscontra un discreto ma non ottimo risultato, come si può notare dall' R^2 aggiustato.

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

3.1 manualmente:

Dall'analisi del punto 2 il modello suggerito sarebbe:

```
modello_2 = lm(Peso ~ . -Anni.madre -Fumatrici -Ospedale, data=df)
summary(modello_2)
```

```
##
## Call:
## lm(formula = Peso ~ . - Anni.madre - Fumatrici - Ospedale, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1129.81  -182.01   -16.35   161.13  2640.18
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6703.5581   136.2451 -49.202  < 2e-16 ***
## N.gravidanze     12.6292     4.3444   2.907  0.00368 **
## Gestazione       32.3345     3.8028   8.503  < 2e-16 ***
## Lunghezza        10.2859     0.3010  34.173  < 2e-16 ***
## Cranio           10.4874     0.4266  24.586  < 2e-16 ***
## Tipo.partoNat    29.6945    12.1299   2.448  0.01443 *
## SessoM           79.1948    11.2207   7.058 2.18e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.5 on 2483 degrees of freedom
## Multiple R-squared:  0.7279, Adjusted R-squared:  0.7273
## F-statistic:  1107 on 6 and 2483 DF,  p-value: < 2.2e-16
```

Ovvero il modello 1 meno i regressori con un' influenza non significativa sulla risposta.

Il modello_2 presenta un R^2 aggiustato perossochè uguale a modello_1 con la differenza di usare 3 variabili in meno, il che rappresenta un vantaggio. Verifichiamo con gli appositi indici (dove verrà usato solo l'indice BIC in quanto tende a penalizzare di più modelli sovraparametrati rispetto al AIC e quindi in linea con il pensiero di Occam):

```
BIC(modello_1, modello_2)
```

```
##           df      BIC
## modello_1 12 35105.20
## modello_2  8 35084.48
```

Viene fatta adesso una analisi dell varianza con il test ANOVA per verificare se ci sono differenza significative della varianza spiegata dai modelli.

```
anova(modello_1, modello_2)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ Anni.madre + 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) - Anni.madre -
##     Fumatrici - Ospedale
```

```
##   Res.Df       RSS Df Sum of Sq     F  Pr(>F)
## 1   2479 186314928
## 2   2483 187106788 -4   -791860 2.634 0.03255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Secondo questo test c'è una varianza spiegata significativamente diversa nei due modelli ma viene ugualmente tenuto il modello_2 in quanto dalle indagini precedenti i regressori eliminati erano poco significativi sulla variabile di risposta, inoltre a causa dell'overfitting è facile avere una varianza spiegata maggiore quando si hanno 3 regressori in più.

Ora si prova a togliere anche la variabile Tipo.parto in quanto non risultava significativa dall' analisi del punto1.

```
modello_3 = update(modello_2, ~. -Tipo.parto)
summary(modello_3)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.74  -181.15   -15.78   163.54  2641.12
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6677.5479   135.9666 -49.112  < 2e-16 ***
## N.gravidanze    12.3451     4.3472   2.840  0.00455 **
## Gestazione      32.3968     3.8066   8.511  < 2e-16 ***
## Lunghezza       10.2487     0.3009  34.059  < 2e-16 ***
## Cranio          10.5205     0.4268  24.651  < 2e-16 ***
## SessoM          79.3038    11.2319   7.061 2.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.8 on 2484 degrees of freedom
## Multiple R-squared:  0.7273, Adjusted R-squared:  0.7267
## F-statistic:  1325 on 5 and 2484 DF,  p-value: < 2.2e-16
```

Anche in questo caso il modello ottenuto riesce a mantenere un R^2 aggiustato praticamente uguale con un parametro in meno.

test anova:

```
anova(modello_2, modello_3)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ (Anni.madre + N.gravidanze + Fumatrici + Gestazione +
##     Lunghezza + Cranio + Tipo.parto + Ospedale + Sesso) - Anni.madre -
##     Fumatrici - Ospedale
```

```
## Model 2: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1   2483 187106788
## 2   2484 187558383 -1   -451595 5.9929 0.01443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anche in questo caso il test anova conferma una perdita di varianza spiegata eliminando una variabile.

BIC:

```
BIC(modello_2, modello_3)
```

```
##           df      BIC
## modello_2  8 35084.48
## modello_3  7 35082.67
```

bic del modello_3 leggermente inferiore del modello_2.

Per ora non si prendono decisioni sulla scelta del modello_2 o modello_3 ma si indaga ulteriolmente.

3.2 selezione del modello tramite la funzione stepAIC:

```
n = nrow(df)
step_wiseAIC = MASS::stepAIC(modello_1, direction = "both", k=log(n))
```

```
## Start:  AIC=28031.07
## Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Anni.madre    1     43987 186358915 28024
## - Ospedale      2    658251 186973179 28024
## - Fumatrici     1     75555 186390483 28024
## - N.gravidanze  1    431392 186746320 28029
## - Tipo.parto    1    436220 186751148 28029
## <none>                      186314928 28031
## - Sesso         1   3703734 190018663 28072
## - Gestazione    1   5454056 191768985 28095
## - Cranio        1  45156723 231471652 28564
## - Lunghezza     1  87889299 274204227 28986
##
## Step:  AIC=28023.84
## Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Fumatrici     1     76066 186434981 28017
## - Ospedale      2    664431 187023346 28017
## - Tipo.parto    1    436190 186795105 28022
## <none>                      186358915 28024
## - N.gravidanze  1    619982 186978897 28024
## + Anni.madre    1     43987 186314928 28031
## - Sesso         1   3712892 190071807 28065
```

```
## - Gestazione    1   5411508 191770423 28087
## - Cranio        1  45406726 231765641 28559
## - Lunghezza     1  87891414 274250330 28978
##
## Step:  AIC=28017.03
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Ospedale      2    671807 187106788 28010
## - Tipo.parto    1    429842 186864823 28015
## <none>                      186434981 28017
## - N.gravidanze  1    599563 187034544 28017
## + Fumatrici     1     76066 186358915 28024
## + Anni.madre    1     44498 186390483 28024
## - Sesso         1   3699630 190134611 28058
## - Gestazione    1   5353972 191788952 28080
## - Cranio        1  45442184 231877165 28552
## - Lunghezza     1  88277759 274712740 28974
##
## Step:  AIC=28010.35
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Tipo.parto    1    451595 187558383 28009
## <none>                      187106788 28010
## - N.gravidanze  1    636812 187743601 28011
## + Ospedale      2    671807 186434981 28017
## + Fumatrici     1     83442 187023346 28017
## + Anni.madre    1     50798 187055990 28018
## - Sesso         1   3753730 190860518 28052
## - Gestazione    1   5447923 192554712 28074
## - Cranio        1  45548801 232655590 28545
## - Lunghezza     1  88000470 275107258 28962
##
## Step:  AIC=28008.53
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
##
##                Df Sum of Sq       RSS   AIC
## <none>                      187558383 28009
## - N.gravidanze  1    608919 188167302 28009
## + Tipo.parto    1    451595 187106788 28010
## + Ospedale      2    693560 186864823 28015
## + Fumatrici     1     76758 187481625 28015
## + Anni.madre    1     50869 187507514 28016
## - Sesso         1   3764137 191322520 28050
## - Gestazione    1   5469179 193027562 28072
## - Cranio        1  45883404 233441787 28546
## - Lunghezza     1  87588001 275146384 28955
```

```
summary(step_wiseAIC)
```

```
##
```

```
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.74  -181.15   -15.78   163.54  2641.12
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6677.5479   135.9666 -49.112  < 2e-16 ***
## N.gravidanze    12.3451     4.3472   2.840  0.00455 **
## Gestazione      32.3968     3.8066   8.511  < 2e-16 ***
## Lunghezza       10.2487     0.3009  34.059  < 2e-16 ***
## Cranio          10.5205     0.4268  24.651  < 2e-16 ***
## SessoM          79.3038    11.2319   7.061 2.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.8 on 2484 degrees of freedom
## Multiple R-squared:  0.7273, Adjusted R-squared:  0.7267
## F-statistic:  1325 on 5 and 2484 DF,  p-value: < 2.2e-16
```

```
BIC(step_wiseAIC, modello_3)
```

```
##              df      BIC
## step_wiseAIC  7 35082.67
## modello_3     7 35082.67
```

La funzione stepAIC sembra confermare il modello_3 precedentemente scelto.

4. Si potrebbero considerare interazioni o effetti non lineari?

relazioni quadratiche:

```
modello_4_1 = update(modello_3, ~. + I(Gestazione^2))
modello_4_2 = update(modello_3, ~. + I(Lunghezza^2))
modello_4_3 = update(modello_3, ~. + I(Cranio^2))
modello_4_4 = update(modello_3, ~. + I(N.gravidanze^2))
BIC(modello_4_1)
```

```
## [1] 35085.28
```

```
BIC(modello_4_2)
```

```
## [1] 34997.96
```

```
BIC(modello_4_3)
```

```
## [1] 35055.64
```

```
BIC(modello_4_4)
```

```
## [1] 35086.9
```

relazioni tra variabili:

```
modello_5_1 = update(modello_3, ~. + Gestazione*Cranio )
modello_5_2 = update(modello_3, ~. + Lunghezza*Gestazione )
modello_5_3 = update(modello_3, ~. + Lunghezza*Cranio )
BIC(modello_5_1)
```

```
## [1] 35055.78
```

```
BIC(modello_5_2)
```

```
## [1] 35064.05
```

```
BIC(modello_5_3)
```

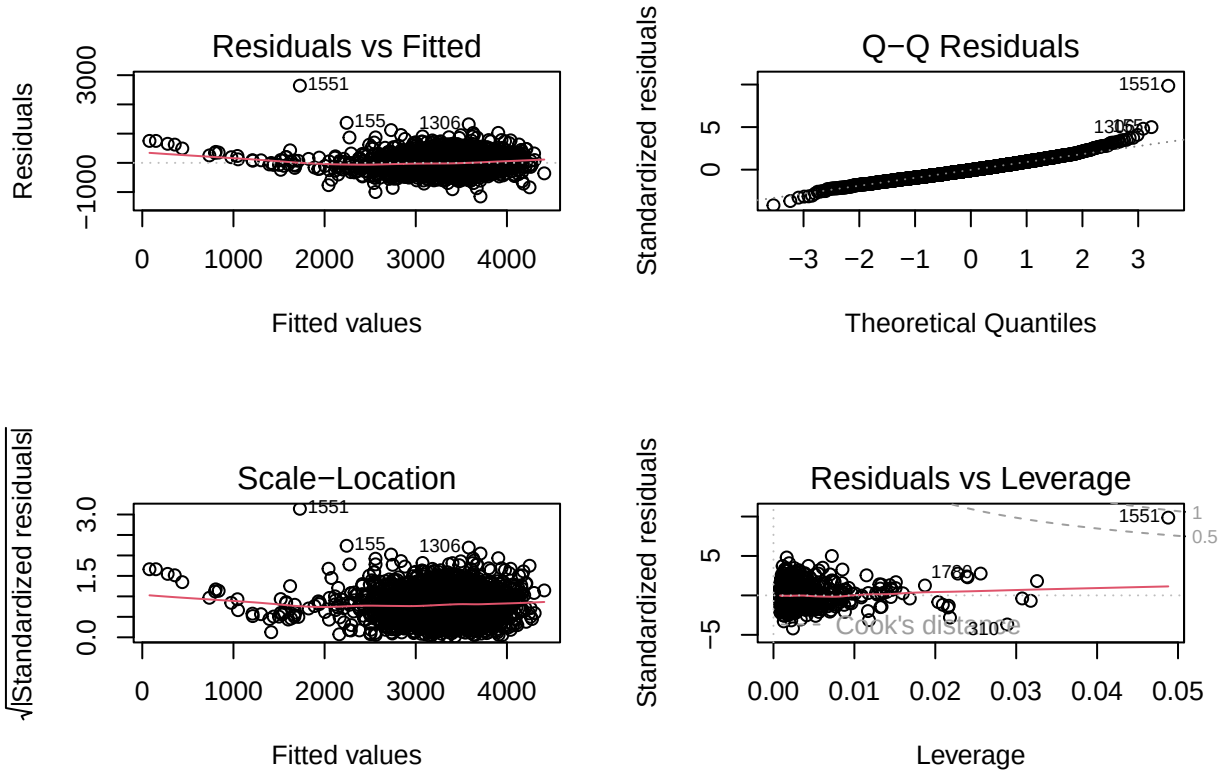
```
## [1] 35067.21
```

Il modello_4_2 è quello con il BIC minore, tra i modelli non lineari. In conclusione i modelli per ora in competizione sono il modello_3 e modello_4_2.

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

-> modello_3:

```
par(mfrow=c(2,2))
plot(modello_3)
```



dai grafici 1 e 3 i residui sembrano avere una leggera maggiore varianza nella zona centrale che tende poi a chiudersi alle estremità.

dal graifco 2 i residui sembrano seguire una buona approssimazione della normale, eccetto nelle code che tendono leggermente a distaccarsi dalla bisettrice.

dal grafico 4 si nota l il residuo 1551 nella zona di attenzione, ovvero superiore a 0.5.

Verifichiamo la presenza di outliers:

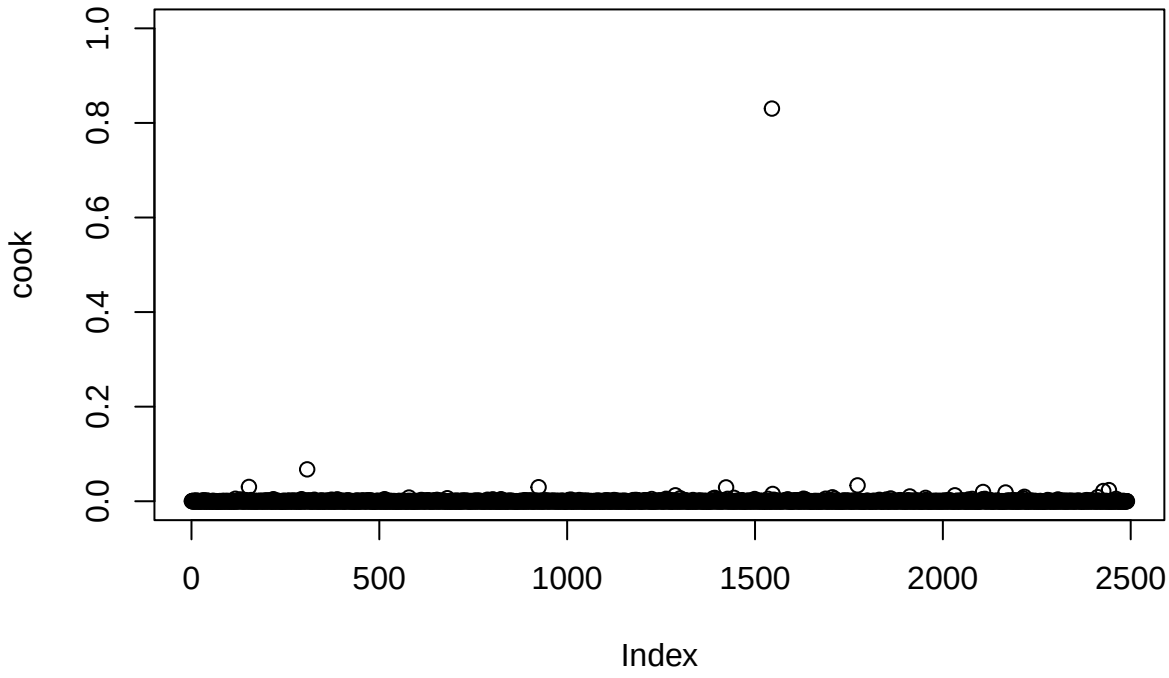
```
#distanza di cook
cook<-cooks.distance(modello_3)
max(cook)
```

```
## [1] 0.8303653
```

```
which.max(cook)
```

```
## 1551
## 1545
```

```
plot(cook,ylim = c(0,1))
```



Dall'analisi degli outliers con la distanza di cook risulta che l'osservazione 1551 supera la soglia di avvertimento.

```
df_no_out = df[-1551,]

modello_3_no_out = lm(Peso ~ . -Anni.madre -Fumatrici -Ospedale -Tipo.parto, data=df_no_out)
```

Viene quindi eliminata dal dataset (anche se tipicamente si dovrebbe calcolare la media ma in questo caso viene fatto così in quanto i dati sono tanti).

Si saggiano le ipotesi di normalità, omoschedasticità e indipendenza rispettivamente:

```
shapiro.test(residuals(modello_3_no_out))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modello_3_no_out)
## W = 0.97401, p-value < 2.2e-16
```

```
bptest(modello_3_no_out)
```

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

```
## data:  modello_3_no_out
## BP = 89.289, df = 5, p-value < 2.2e-16
```

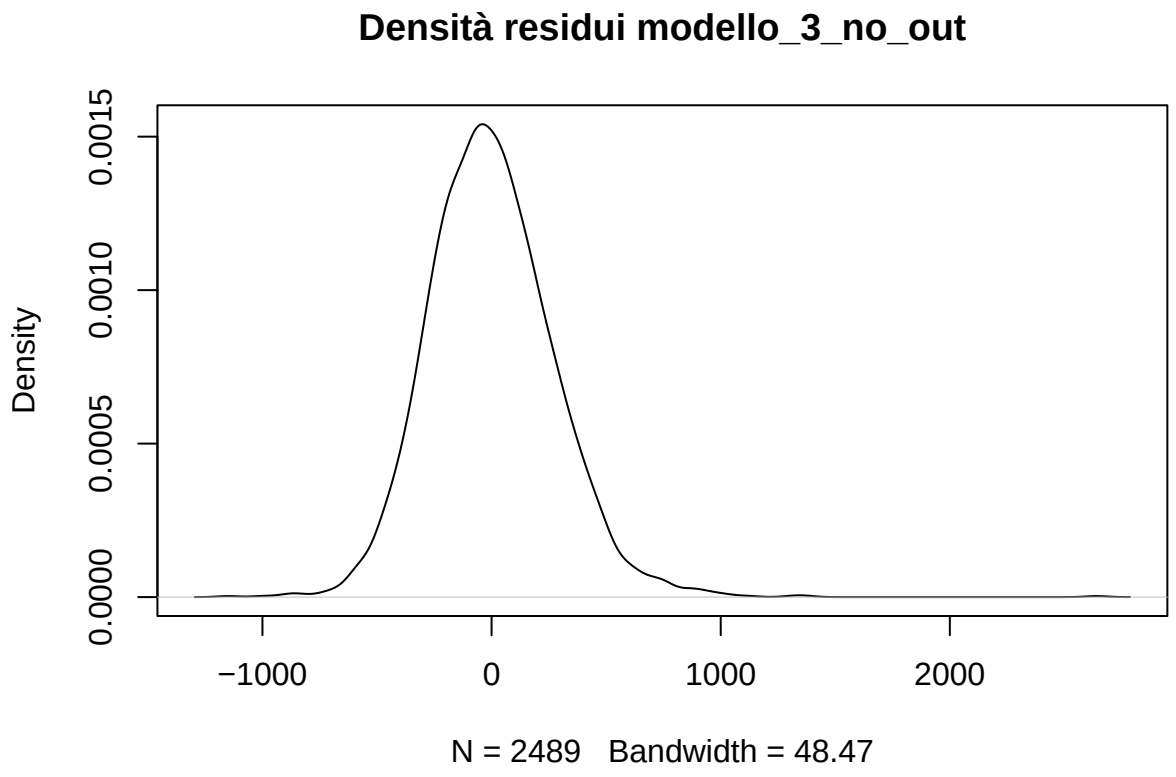
```r
dwtest(modello_3_no_out)
```

```
##
##  Durbin-Watson test
##
## data:  modello_3_no_out
## DW = 1.9591, p-value = 0.1535
## alternative hypothesis: true autocorrelation is greater than 0
```

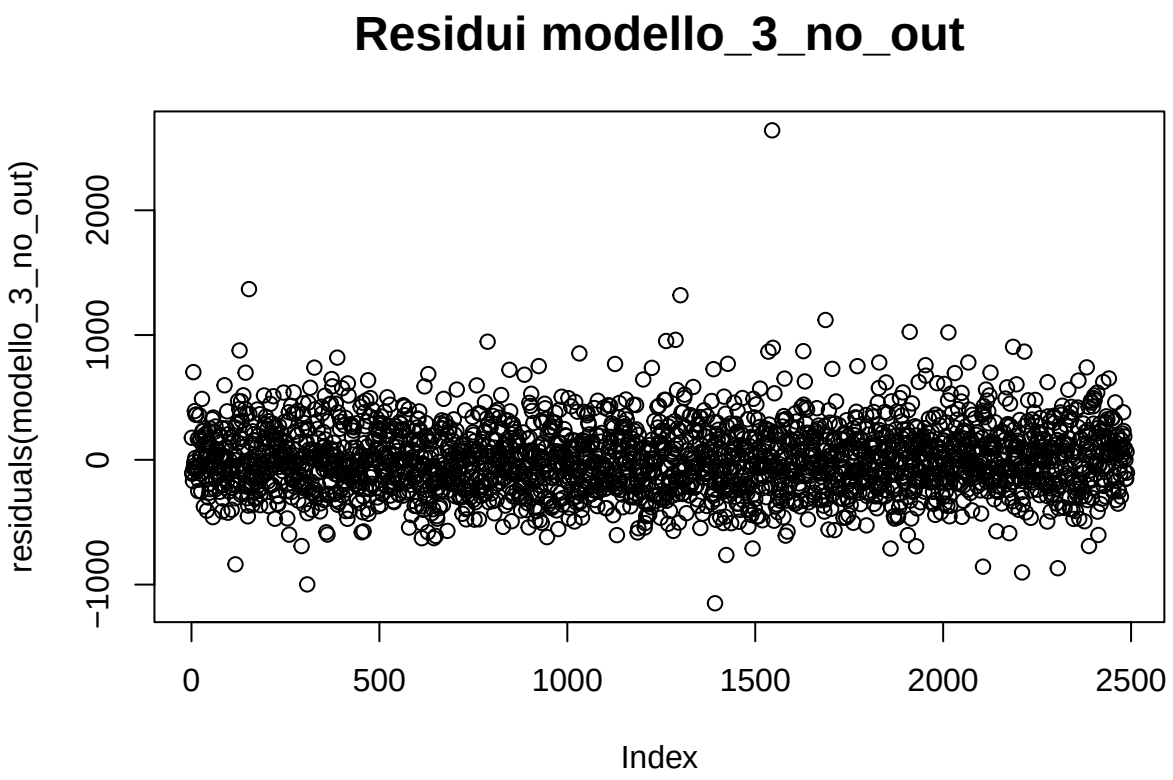
Dalla quale: -> residui con una distribuzione non normale -> residui non omoschedastici -> residui indipendenti

Infine:

```r
plot(density(residuals(modello_3_no_out)), main="Densità residui modello_3_no_out")
```




```r
plot(residuals(modello_3_no_out))
title(main = "Residui modello_3_no_out", cex.main = 1.5)
```



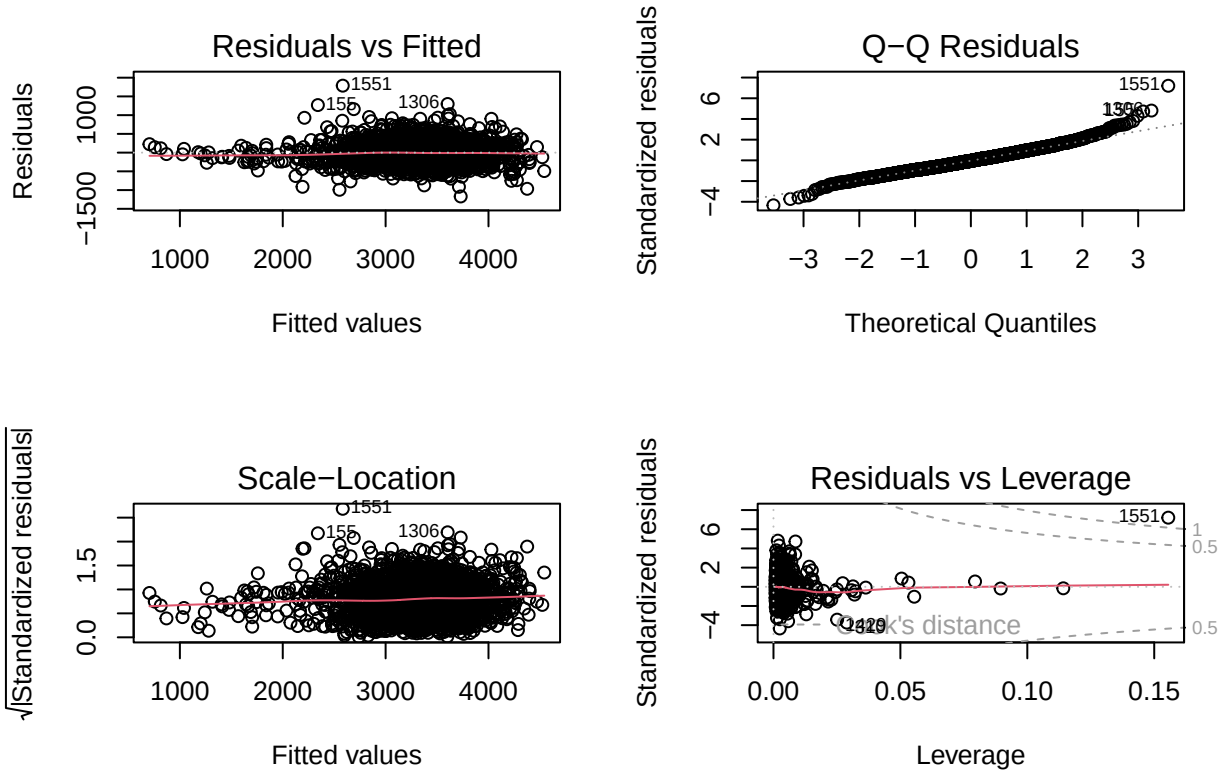
Dagli ultimi due grafici invece si possono notare:

-> residui con una buona approssimazione alla normale -> residui omoschedastici, ovvero senza particolari pattern nella varianza.

Essendo che con campioni di grandi dimensioni, il test di Shapiro-Wilk può diventare statisticamente significativo anche se le deviazioni dalla normalità sono trascurabil (questo è dovuto alla sua sensibilità elevata) si è deciso di tenere comunque in considerazione il modello 3 con esclusa l'osservazione 1551 ovvero il "modello_3_no_out".

-> modello_4_2

```
par(mfrow=c(2,2))
plot(modello_4_2)
```



Le stesse considerazioni dello stesso grafico del modello_3 possono essere fatte per il modello_4_2.

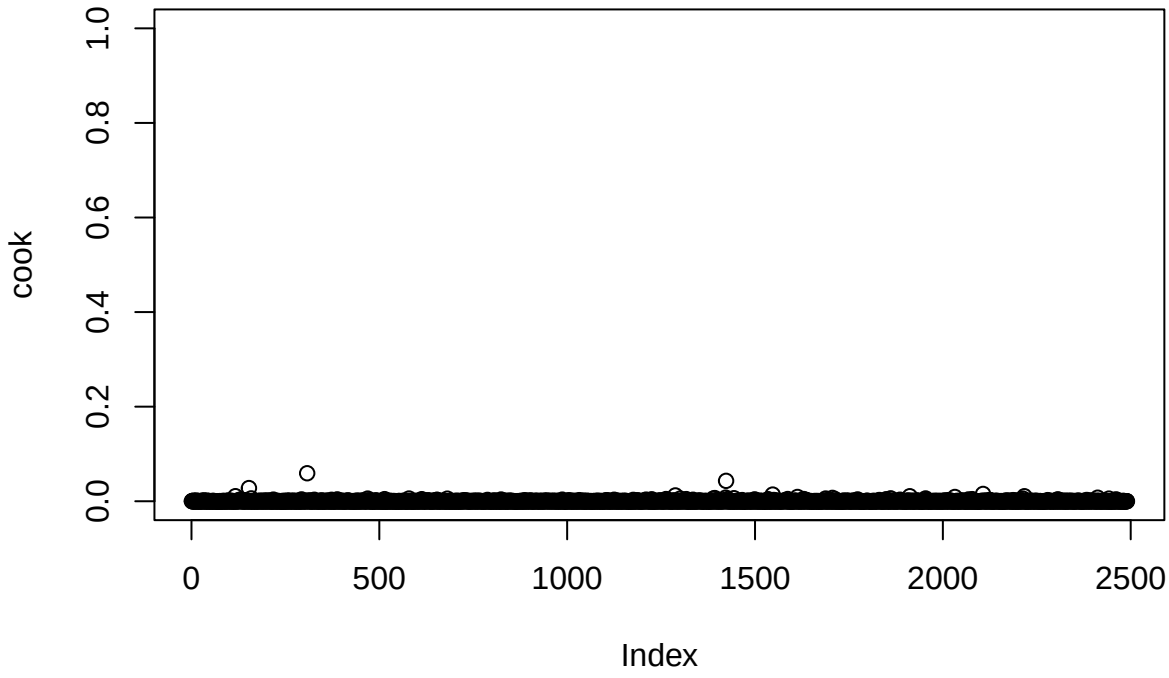
```
#distanza di cook
cook<-cooks.distance(modello_4_2)
max(cook)
```

```
## [1] 1.366412
```

```
which.max(cook)
```

```
## 1551
## 1545
```

```
plot(cook,ylim = c(0,1))
```



Nessun particolare valore oltre la distanza di cook.

Si saggiano le ipotesi di normalità, omoschedasticità e indipendenza rispettivamente:

```
shapiro.test(residuals(modello_4_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modello_4_2)
## W = 0.98564, p-value = 3.641e-15
```

```
bptest(modello_4_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modello_4_2
## BP = 127.03, df = 6, p-value < 2.2e-16
```

```
dwtest(modello_4_2)
```

```
##
##  Durbin-Watson test
##
```

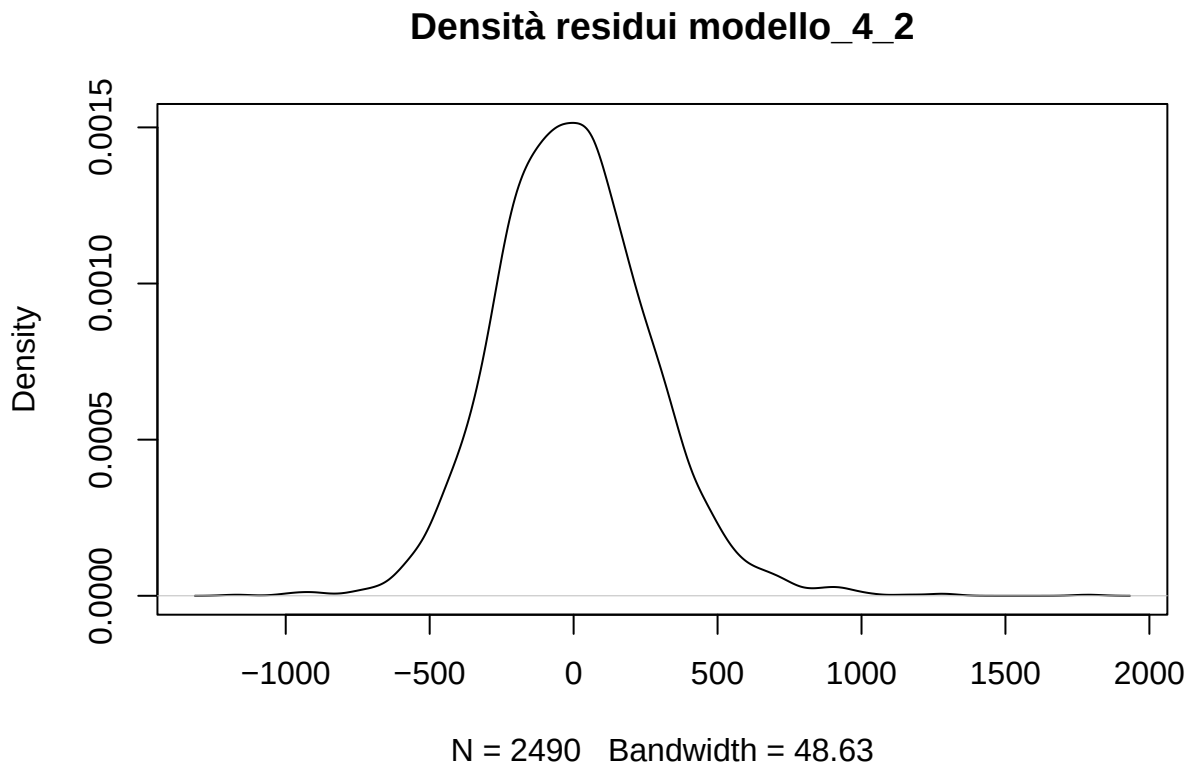
```
## data:  modello_4_2
## DW = 1.9519, p-value = 0.1149
## alternative hypothesis: true autocorrelation is greater than 0
```

Dalla quale:

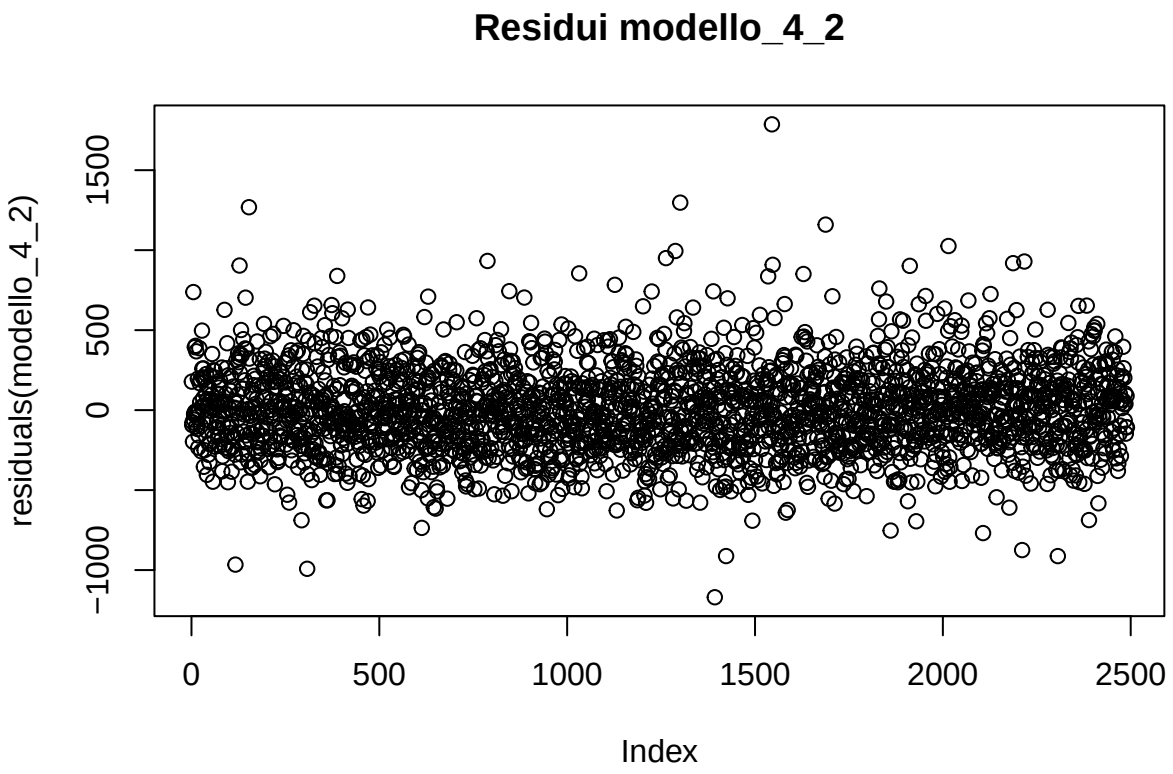
-> residui con una distribuzione non normale -> residui non omoschedastici -> residui indipendenti

Infine:

```
plot(density(residuals(modello_4_2)), main="Densità residui modello_4_2")
```




```
plot(residuals(modello_4_2))
title(main="Residui modello_4_2")
```



Anche in questo caso dagli ultimi due grafici invece si possono notare:

-> residui con una buona approssimazione alla normale -> residui omoschedastici, ovvero senza particolari pattern nella varianza.

In conclusione si trova:

1) differenza non significativa nel parametro R^2 tra i due modelli.

```
summary(modello_3_no_out)$r.squared
```

```
## [1] 0.7271496
```

```
summary(modello_4_2)$r.squared
```

```
## [1] 0.7372122
```

2) BIC inferire del modello_4_2.

```
BIC(modello_3_no_out)
```

```
## [1] 35069.56
```

```
BIC(modello_4_2)
```

```
## [1] 34997.96
```

3) numero inferiore di regressori del modello_3.

4) una leggera migliore approssimazione alla normale dei residui del modello_4_2.

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

uso il mean square error:

```
df_test = read.csv("neonati.csv")

p31 = predict(modello_3_no_out, df_test[27,])
p41 = predict(modello_4_2, df_test[27,])

p32 = predict(modello_3_no_out, df_test[73,])
p42 = predict(modello_4_2, df_test[73,])

p33 = predict(modello_3_no_out, df_test[516,])
p43 = predict(modello_4_2, df_test[516,])

p34 = predict(modello_3_no_out, df_test[812,])
p44 = predict(modello_4_2, df_test[812,])

p35 = predict(modello_3_no_out, df_test[1111,])
p45 = predict(modello_4_2, df_test[1111,])

p36 = predict(modello_3_no_out, df_test[1315,])
p46 = predict(modello_4_2, df_test[1315,])

p37 = predict(modello_3_no_out, df_test[1717,])
p47 = predict(modello_4_2, df_test[1717,])

p38 = predict(modello_3_no_out, df_test[1899,])
p48 = predict(modello_4_2, df_test[1899,])

p39 = predict(modello_3_no_out, df_test[2301,])
p49 = predict(modello_4_2, df_test[2301,])

p310 = predict(modello_3_no_out, df_test[2400,])
p410 = predict(modello_4_2, df_test[2400,])

v1 = c(p31, p32, p33, p34, p35, p36, p37, p38, p39, p310)
v2 = c(p41, p42, p43, p44, p45, p46, p47, p48, p49, p410)

mu = c(df_test[27,]$Peso, df_test[73,]$Peso, df_test[516,]$Peso, df_test[812,]$Peso, df_test[1111,]$Pes


mse_3_no_out = MSE(v1,mu)
mse_4_2 = MSE(v2,mu)

mse_3_no_out
```

```
## [1] 44824.36
```

```
mse_4_2
```

```
## [1] 47789.78
```

In conclusione si può dire che entrambi i modelli siano dei buoni adattamenti per questo Dataset ma viene preso in considerazione il modello_3/modello_3_no_out in quanto a parità di R^2 è quello che ha meno parametri ed è anche quello che sembra predire meglio dati nuovi (mse_3_no_out < mse_4_2).

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.

Essendo che non si hanno valori per le variabili di controllo lunghezza e cranio prima di effettuare la predizione le rimuovo dal modello per non ottenere risultati sballati.

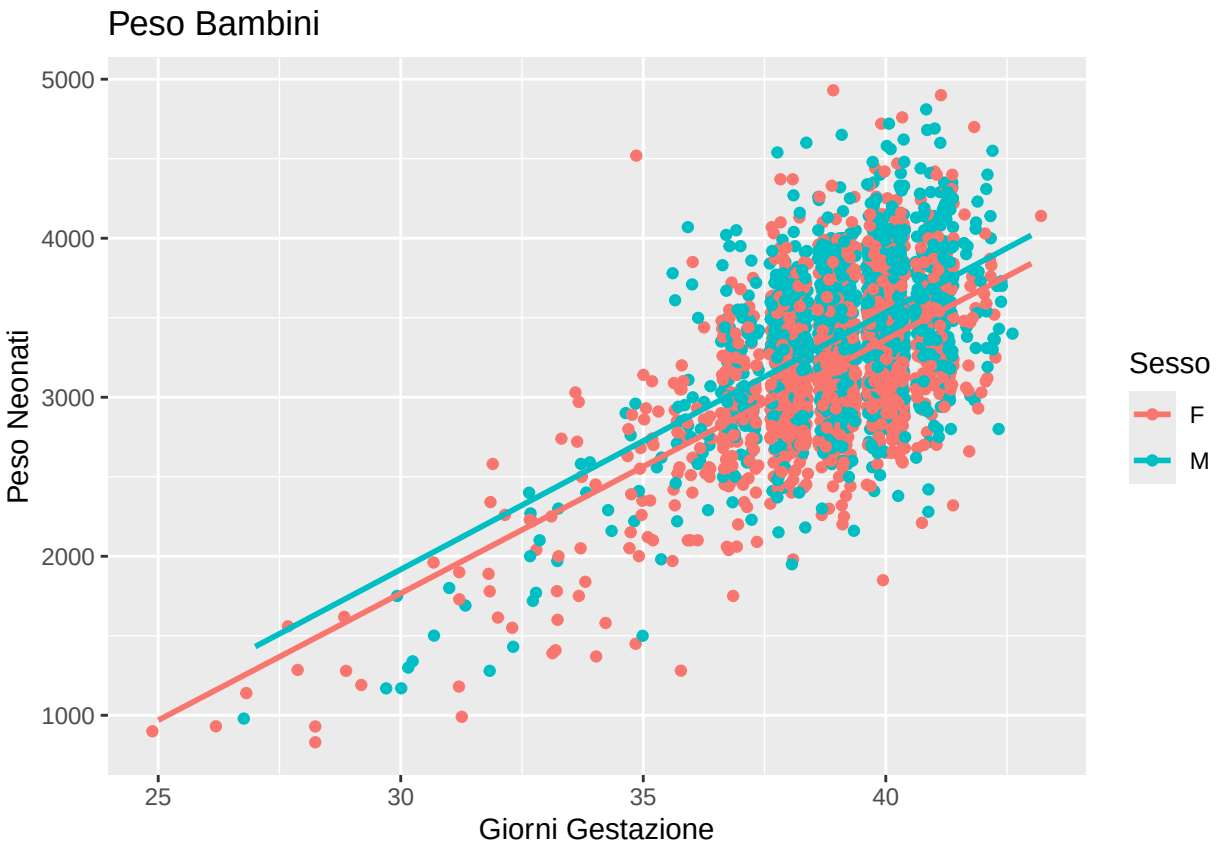
```
modello_6 = update(modello_3_no_out, ~. -Lunghezza -Cranio)
osservazione = data.frame(N.gravidanze=3,Gestazione=39, Sesso="F")

predict(modello_6, osservazione)
```

```
##        1
## 3251.248
```

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

```
ggplot(data = df)+
  geom_point(aes(x = Gestazione,
                 y = Peso,
                 col = Sesso),position = "jitter")+
  geom_smooth(aes(x = Gestazione,
                  y = Peso,
                  col = Sesso),se=F,method = "lm")+
  labs(title = "Peso Bambini",
       x="Giorni Gestazione",
       y="Peso Neonati")
```

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



Il peso dei bambini cresce al crescere delle settimane di gestazione. La maggior parte dei bambini nasce dalla 35° settimana di gestazione. Non ci sono differenze particolari nell'andamento della crescita del peso al crescere delle settimane di gestazione al variare del sesso del bambino (le rette hanno diversa intercetta ma simile pendenza).