Progetto Statistica Inferenziale

Eugenio Pasqua

2023-12-09

Un modello Statistico per la previsione del peso dei neonati

Questo progetto si propone di esplorare il dataset riguardante diverse variabili legate alla nascita di neonati e di determinare quali di queste potrebbero essere correlate con il peso del neonato alla nascita. L'obiettivo principale è comprendere e stabilire le relazioni esistenti tra le variabili come Gestazione, sesso, misure antropometriche come Cranio e Lunghezza, peso del neonato e caratterisiche della madre. La variabile peso del neonato sarà la nostra variabile target su cui effettuare l'analisi e le previsioni.

1 - Importiamo il dataset "neonati.csv"

Procediamo con l'importazione del dataset "neonati.csv":

```
dati_sorgenti <- read.csv("neonati.csv")
dati_sorgenti <- dati_sorgenti[!is.na(dati_sorgenti$Anni.madre)
                               & dati_sorgenti$Anni.madre != 0, ]
head(dati_sorgenti,5)
```

```
##   Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio Tipo.parto
## 1         26            0         0         42 3380       490    325        Nat
## 2         21            2         0         39 3150       490    345        Nat
## 3         34            3         0         38 3640       500    375        Nat
## 4         28            1         0         41 3690       515    365        Nat
## 5         20            0         0         38 3700       480    335        Nat
##   Ospedale Sesso
## 1     osp3     M
## 2     osp1     F
## 3     osp2     M
## 4     osp2     M
## 5     osp3     F
```

2 - Descrizione del dataset, sua composizione, tipo di variabili e obiettivo dello studio:

Il dataset raccoglie dati relativi alle madri e alle caratteristiche dei neonati al momento del parto. Di seguito sono elencate le variabili associate alle madri:

- Età della madre
- Numero di gravidanze precedenti
- Fumo durante la gravidanza della madre
- Durata della gestazione in settimane Le seguenti variabili sono invece riferite alle misurazioni antropometriche dei neonati e ad altri dettagli relativi al parto:
- Peso del neonato (espresso in grammi)
- Lunghezza del neonato (espressa in millimetri)
- Diametro del cranio del neonato (espresso in millimetri)
- Tipo di parto (naturale o cesareo)

- Ospedale in cui si è svolto il parto
- Sesso del neonato

3 - Indagine sulle variabili effettuando una breve analisi descrittiva, utilizzando indici e strumenti grafici:

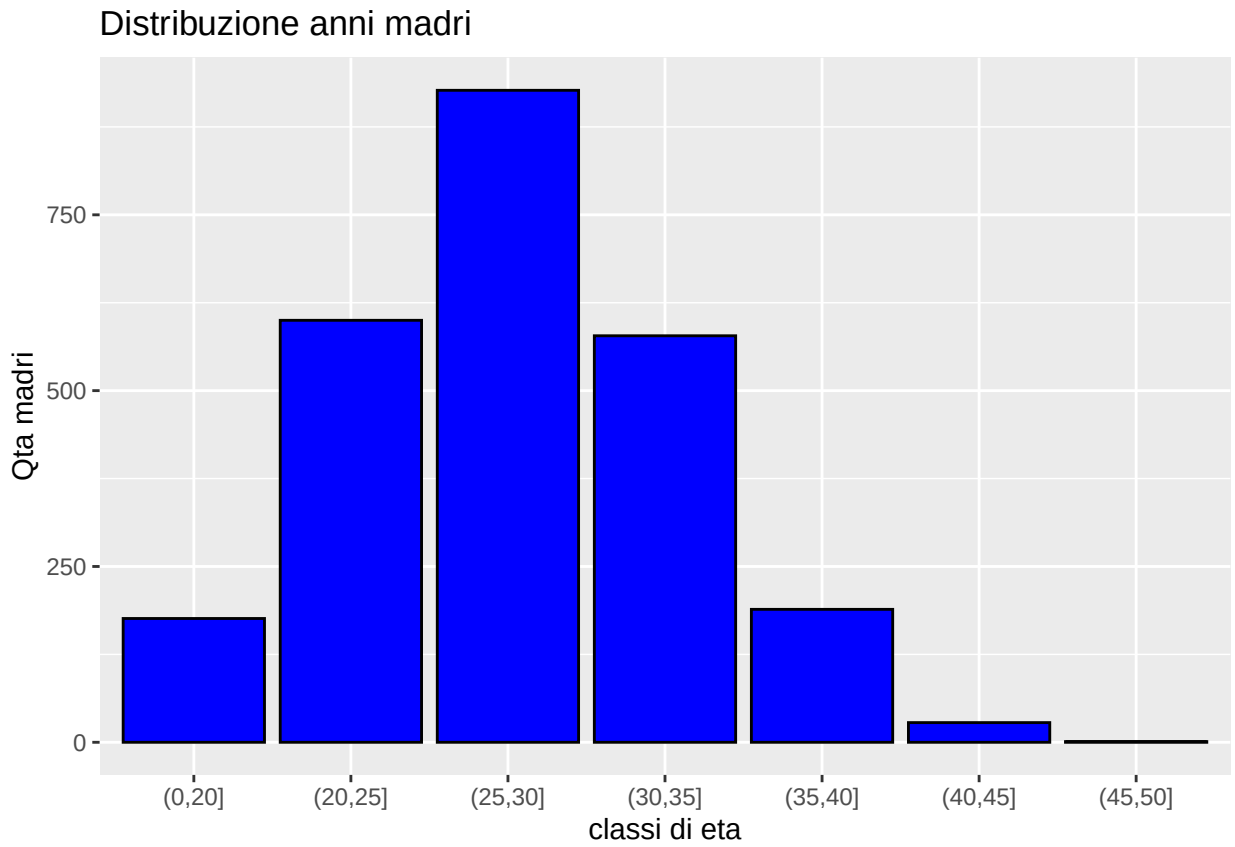
Effettuiamo un riepilogo delle variabili per esaminare le misure di posizione del dataset, offrendo una descrizione statistica di ogni variabile presente nel dataset appena importato:

```
summary(dati_sorgenti)
```

```
##    Anni.madre     N.gravidanze       Fumatrici         Gestazione
##  Min.   : 1.00   Min.   : 0.0000   Min.   :0.00000   Min.   :25.00
##  1st Qu.:25.00   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:38.00
##  Median :28.00   Median : 1.0000   Median :0.00000   Median :39.00
##  Mean   :28.18   Mean   : 0.9816   Mean   :0.04162   Mean   :38.98
##  3rd Qu.:32.00   3rd Qu.: 1.0000   3rd Qu.:0.00000   3rd Qu.:40.00
##  Max.   :46.00   Max.   :12.0000   Max.   :1.00000   Max.   :43.00
##       Peso         Lunghezza         Cranio     Tipo.parto
##  Min.   : 830   Min.   :310.0   Min.   :235   Length:2499
##  1st Qu.:2990   1st Qu.:480.0   1st Qu.:330   Class :character
##  Median :3300   Median :500.0   Median :340   Mode  :character
##  Mean   :3284   Mean   :494.7   Mean   :340
##  3rd Qu.:3620   3rd Qu.:510.0   3rd Qu.:350
##  Max.   :4930   Max.   :565.0   Max.   :390
##    Ospedale            Sesso
##  Length:2499        Length:2499
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

L'età media delle madri si aggira intorno ai 28 anni, con la maggior parte delle osservazioni concentrata tra i 25 e i 32 anni. Ora procediamo a esaminare la distribuzione di frequenza delle osservazioni suddividendo il campione in cluster e contando il numero di madri in ciascun cluster di età:

```
dati_sorgenti$Anni.madre_classi <- cut(dati_sorgenti$Anni.madre,
                                       breaks = c(0,20,25,30,35,40,45,50,60))
library(ggplot2)
ggplot(data = dati_sorgenti)+
  geom_bar(aes(x=Anni.madre_classi),
           stat="count",
           col="black",
           fill="blue")+
  labs(title="Distribuzione anni madri",x="classi di eta",y="Qta madri")
```



Esaminiamo ora la variabile relativa al numero di gravidanze. In media, ogni madre ha avuto circa una gravidanza, con un periodo di gestazione mediamente compreso tra le 38 e le 40 settimane. Passando al campione di madri, esaminiamo la distribuzione della variabile "Fumatrici" e come questa si distribuisce nel totale delle osservazioni presenti nel dataset.

```
madri_fumatrici <- table(dati_sorgenti$Fumatrici)
madri_fumatrici
```

```
##
##    0    1
## 2395  104
```

Nel dataset esaminato, la presenza di madri fumatrici è estremamente rara, con la maggior parte delle madri che non fuma. Per quanto riguarda le caratteristiche dei neonati, il peso medio è di circa 3,2 kg, con la maggior parte dei neonati che si aggira tra 2,9 kg e 3,6 kg. La lunghezza media dei neonati varia da 480 mm a 510 mm, mentre le misurazioni del cranio oscillano principalmente tra 330 e 350, con una media di circa 340. Analizziamo adesso la distribuzione dei parti osservati tra parto naturale e parto cesareo nel dataset:

```
tipologie_parto <- table(dati_sorgenti$Tipo.parto)
tipologie_parto
```

```
##
##  Ces  Nat
##  728 1771
```

Come possiamo vedere piu della metà delle osservazioni sono parti "naturali". Vediamo ora il sesso dei neonati come si distribuisce sul campione descritto dal dataset:

```
sesso_neonati <- table(dati_sorgenti$Sesso)
sesso_neonati
```

```
##
##    F    M
## 1256 1243
```

Dai dati, sembra che la distribuzione sia abbastanza uniforme in questo caso. Adesso infine diamo un'occhiata alla variabile "Ospedale" per comprendere come sono distribuiti i dati all'interno del dataset:

```
tipo_osp <- table(dati_sorgenti$Ospedale)
tipo_osp
```

```
##
## osp1 osp2 osp3
##  816  849  834
```

Da quanto riportato sembra che su tutti i parti registrati, i 3 ospedali abbiano una distribuzione equilibrata.

4 - Saggiamo l'ipotesi che la media del "peso" e della "lunghezza" del campione di neonati del dataset siano significativamente uguali a quello della popolazione:

Dopo aver effettuato una ricerca online, ho trovato una media approssimativa di 3,3 kg e 50 cm per neonati al momento della nascita.

Adesso, procederemo con il test t per valutare se il peso registrato nel nostro campione è significativamente simile a quello della popolazione generale. Questo ci aiuterà a determinare se il nostro campione è rappresentativo per l'analisi dell'intera popolazione:

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

```
##
##  One Sample t-test
##
## data:  dati_sorgenti$Peso
## t = -1.5069, df = 2498, p-value = 0.132
## alternative hypothesis: true mean is not equal to 3300
## 95 percent confidence interval:
##  3263.572 3304.769
## sample estimates:
## mean of x
##   3284.17
```

Dalle evidenze sopra riportate si evince che il p-value non è inferiore a livello di significatività prescelto pertanto non possiamo rifiutare l'ipotesi nulla quindi sul peso il campione risulterebbe rappresentativo per la popolazione.

Andiamo ora a saggiare l'ipotesi della lunghezza : abbiamo detto che dovremmo avere una media di circa 50cm vediamo se è vero:

```
t.test(dati_sorgenti$Lunghezza,mu=500,conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  dati_sorgenti$Lunghezza
## t = -10.077, df = 2498, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
```

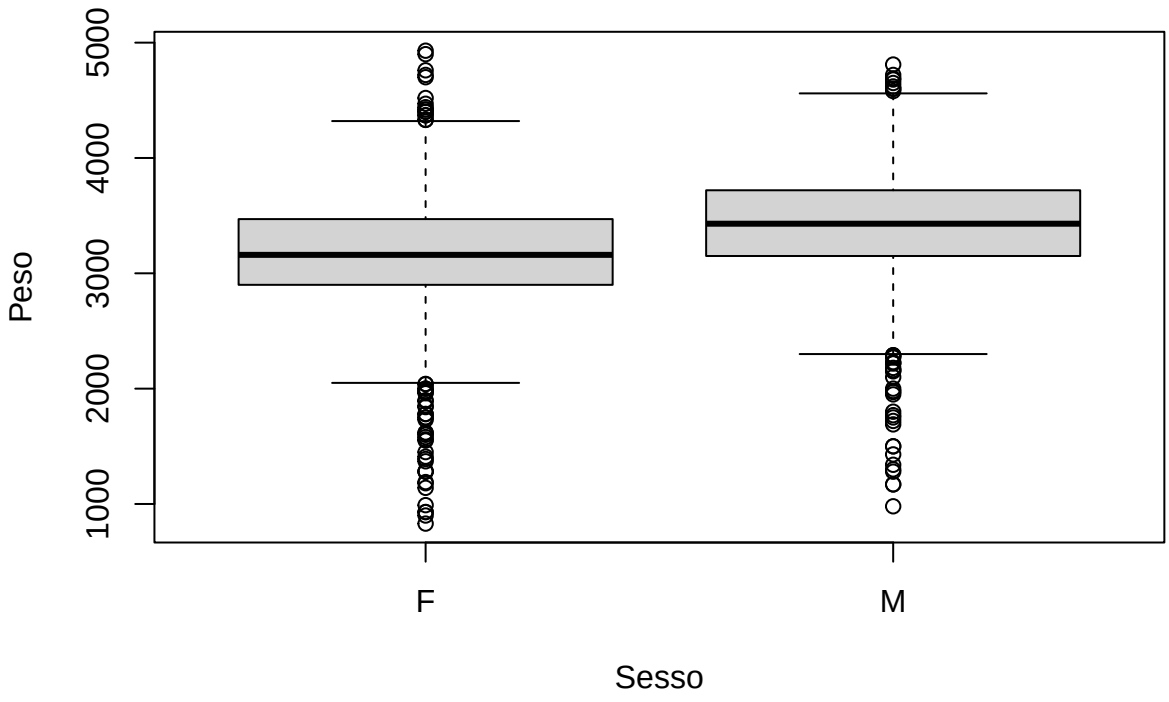
```
##  493.6613 495.7265
## sample estimates:
## mean of x
##  494.6939
```

Il test t ha restituito un p-value molto basso, indicando che la lunghezza potrebbe differire dalla media della popolazione, poiché il p-value è inferiore al livello di significatività scelto. Tuttavia, considerando la mediana, che si attesta intorno al valore 500, potremmo accettare questa misura come un indicatore sufficiente.

5 - Sempre per il "peso" e "lunghezza" o eventualmente per altre per cui ha senso farlo procedo a verificare le differenze significative tra i due sessi:

Procediamo ora a verificare se i parametri di peso e lunghezza in riferimento al sesso dei neonati rilevano differenze significative. Per far questo utilizziamo l'approccio dei boxplot condizionati:

```
attach(dati_sorgenti)
boxplot(Peso~Sesso)
```



Come mostrato dal grafico sopra riportato, i neonati maschi tendono ad assumere un peso leggermente superiore a quello delle femmine.

Riesaminiamolo utilizzando il t-test per determinare se esiste una differenza statistica significativa. Verificheremo quindi l'ipotesi nulla che il peso medio dei neonati maschi sia uguale a quello delle femmine:

```
dati_maschi <- subset(dati_sorgenti, Sesso=='M')
dati_femmine <- subset(dati_sorgenti, Sesso=='F')

# Mi calcolo il peso medio delle femmine
```

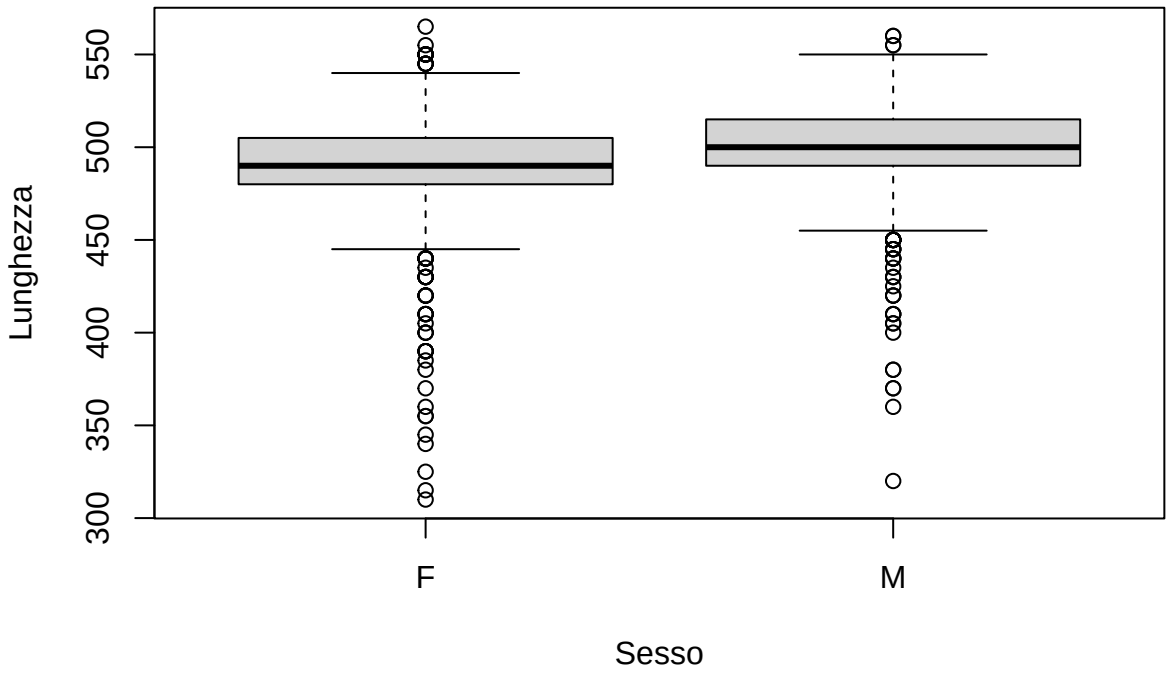
```r
media_femmine <- mean(dati_femmine$Peso)
t.test(dati_maschi$Peso,mu=media_femmine,conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  dati_maschi$Peso
## t = 17.657, df = 1242, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 3161.132
## 95 percent confidence interval:
##  3381.012 3435.979
## sample estimates:
## mean of x
##  3408.496
```

Il p-value ottenuto dal t-test è inferiore al livello di significatività, quindi respingiamo l'ipotesi nulla. Questo fornisce una forte evidenza a sostegno dell'ipotesi che i neonati maschi e femmine abbiano pesi differenti.

Vediamo in termini di lunghezza e rifacciamo la medesima analisi applicata per tale variabile:

```r
boxplot(Lunghezza~Sesso)
```



Anche la lunghezza nei maschi sembra essere leggermente superiore a quella delle femmine infatti il boxplot "M" si pone verticalmente piu in alto rispetto a quello "F". Proviamolo con il t-test :

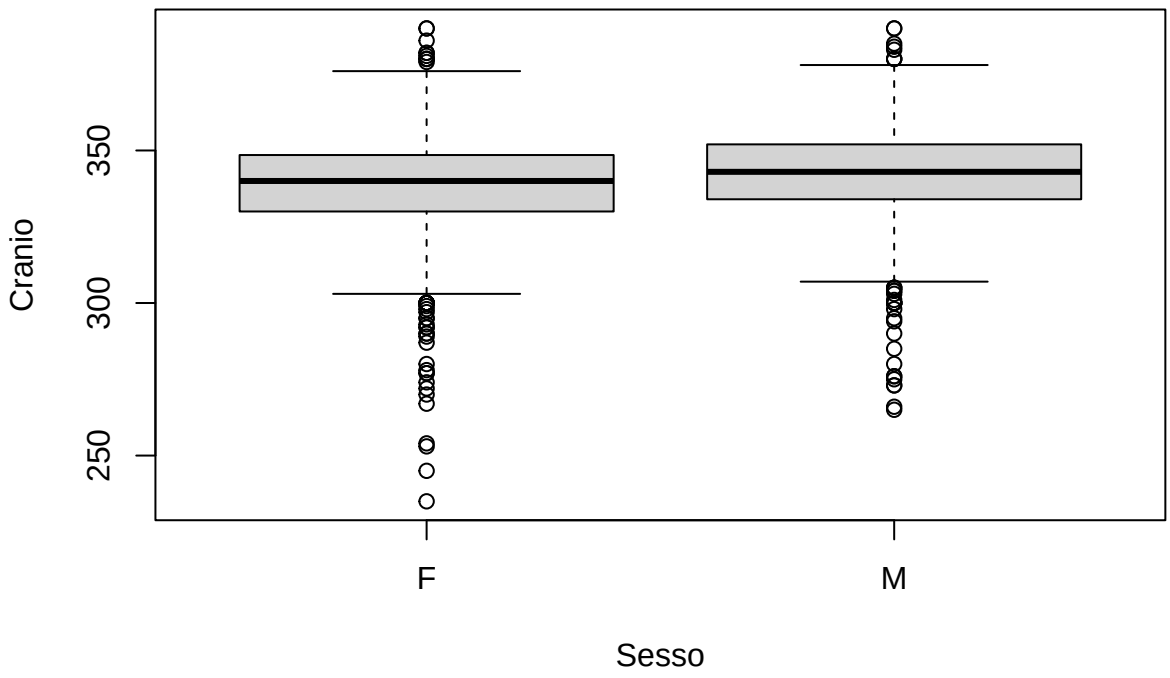
```r
dati_maschi <- subset(dati_sorgenti, Sesso=='M')
dati_femmine <- subset(dati_sorgenti, Sesso=='F')

media_femmine <- mean(dati_femmine$Lunghezza)
```

```
t.test(dati_maschi$Lunghezza,mu=media_femmine,conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  dati_maschi$Lunghezza
## t = 14.531, df = 1242, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 489.7643
## 95 percent confidence interval:
##  498.3369 501.0131
## sample estimates:
## mean of x
##   499.675
```

Nel contesto del t-test, il valore ottenuto per il p-value conferma il rifiuto dell'ipotesi di uguaglianza delle lunghezze tra neonati maschi e femmine. Proviamo a vedere anche la rispettiva dimensione del cranio:

```
boxplot(Cranio~Sesso)
```



In questo caso siamo piu vicini tra i due sessi anche se nei maschi si nota anche in tale variabile una dimensione leggermente superiore.

Riproviamo anche qui l'evidenza con il t-test:

```
dati_maschi <- subset(dati_sorgenti, Sesso=='M')
dati_femmine <- subset(dati_sorgenti, Sesso=='F')

media_femmine <- mean(dati_femmine$Cranio)
t.test(dati_maschi$Cranio,mu=media_femmine,conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  dati_maschi$Cranio
## t = 10.804, df = 1242, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 337.633
## 95 percent confidence interval:
##  341.5823 343.3348
## sample estimates:
## mean of x
##  342.4586
```

Si conferma anche con il t-test la differenza in media delle dimensioni del cranio tra neonati maschi e neonati femmine.

6 - Verifichiamo se in alcuni ospedali vengono eseguiti in maggioranza parti cesarei rispetto ad altri:

Iniziamo filtrando il dataset per i parti cesarei e procediamo creando un istogramma che illustra la distribuzione. Utilizzeremo gli ospedali sull'asse delle ascisse (x) e il numero di parti cesarei sull'asse delle ordinate (y):

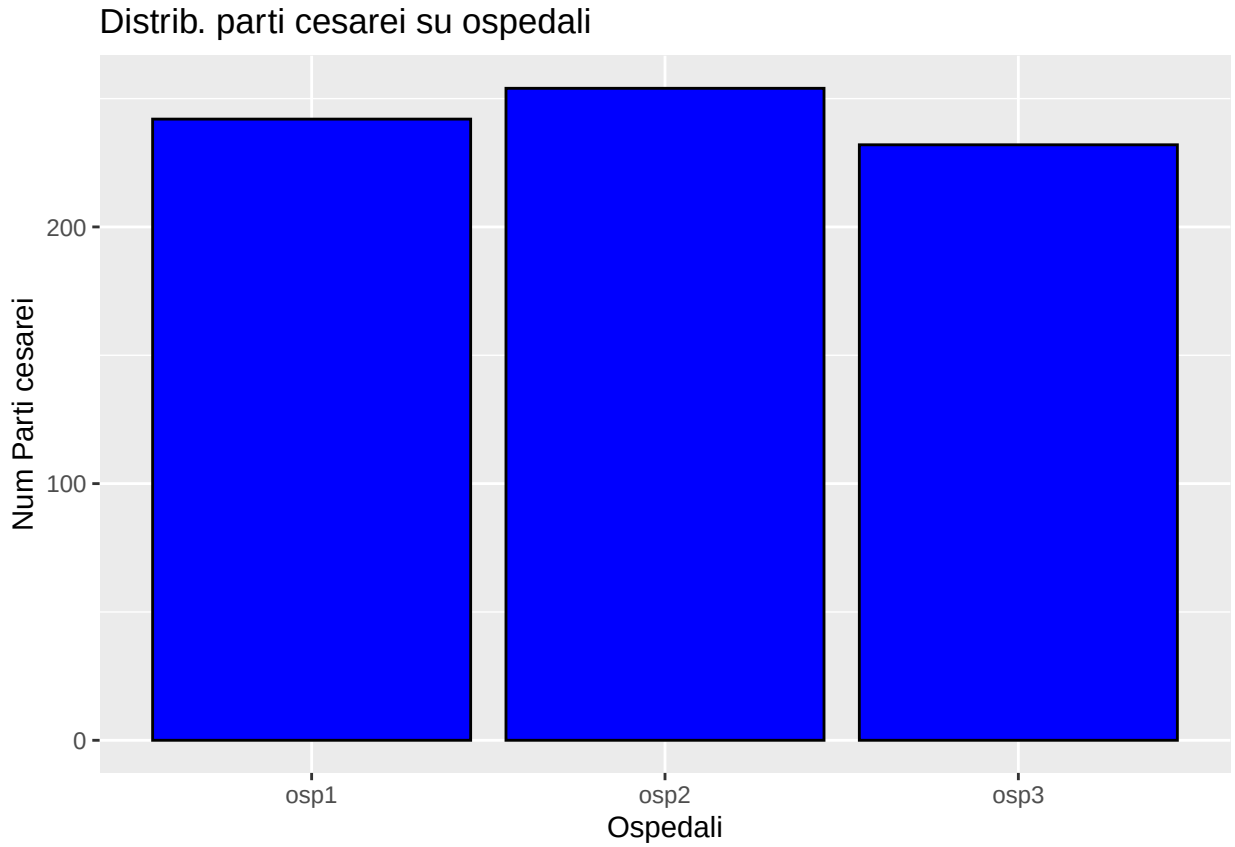
```
library(dplyr)
```

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

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

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

```
df_cesarei <- dati_sorgenti %>% filter(Tipo.parto=="Ces") %>%select(Tipo.parto,Ospedale)

library(ggplot2)
ggplot(data = df_cesarei)+
  geom_bar(aes(x=Ospedale),
           stat="count",
           col="black",
           fill="blue")+
  labs(title="Distrib. parti cesarei su ospedali",x="Ospedali",y="Num Parti cesarei")
```



Come si può notare dal grafico, nel secondo ospedale sembra essere registrato un numero maggiore di casi rispetto agli ospedali 1 e 3. Andiamo adesso a provare con il test chi-quadro se sussistono prove statisticamente significative che determinati tipi di parto vengono eseguiti in determinati ospedali:

```
# Creazione della tabella di contingenza
tabella_contingenza <- table(df_cesarei$Ospedale, df_cesarei$Tipo.parto)
chisq.test(tabella_contingenza)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tabella_contingenza
## X-squared = 1, df = 2, p-value = 0.6065
```

Dall'evidenza mostrata non ci sono prove statisticamente significative che suggeriscano un'associazione tra il tipo di parto e l'ospedale, considerando i dati nel dataset. Un p-value alto come 0.6 suggerisce che le differenze osservate nelle frequenze dei tipi di parto tra gli ospedali potrebbero essere attribuite al caso o alla variabilità casuale piuttosto che a una reale associazione tra il tipo di parto e l'ospedale.

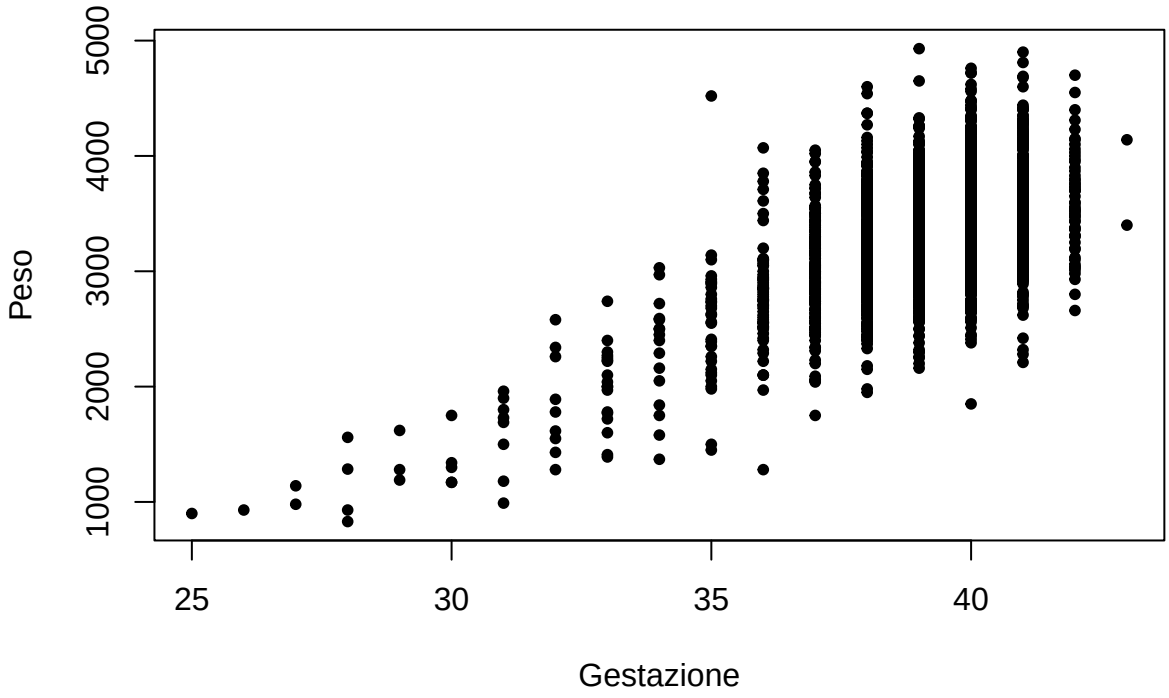
Analisi Multidimensionale e implementazione del modello di previsione

Cominciamo la creazione del nostro modello previsionale analizzando coppie di variabili una alla volta. Procederemo passo dopo passo, valutando la validità del modello appena costruito per ottenere previsioni significative sulla variabile "peso", che è al centro del nostro studio.

1 - Analisi di indagine sulle relazioni tra le variabili a coppie in riferimento alla variabile risposta del "peso"

Iniziamo a vedere la correlazione tra il peso ed il periodo di Gestazione:

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



Dall'andamento della nuvola dei punti sicuramente la variabile Gestazione risulta correlata al peso del neonato; la tendenza risulta quindi lineare positiva ovvero all'aumentare del numero di settimane di Gestazione aumenta il peso del neonato stesso. Possiamo dimostrarlo matematicamente secondo quanto di seguito riportato dalla funzione cor() in R:

```
cor(Gestazione,Peso)
```

```
## [1] 0.5917921
```

L'indice indica che esiste una correlazione positiva. Proviamo adesso a vedere la variabile riferita agli "Anni della madre":

```
cor(Anni.madre,Peso)
```

```
## [1] -0.02351812
```

Come possiamo vedere il valore è tendente allo zero pertanto la funzione ci indirizza verso un assenza di correlazione alla variabile target "peso" del neonato. Verifichiamo adesso se esiste correlazione con il numero di gravidanze:

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

```
## [1] 0.002276741
```

Anche in questo caso l'indice conferma assenza di correlazione. Passiamo allo step di verifica di correlazione

tra il “Peso” e lo status di madri Fumatrici:

```
cor(Fumatrici,Peso)
```

```
## [1] -0.01898179
```

L’indice conferma assenza di correlazione anche tra questa coppia di variabili. Ad integrazione verifichiamolo eseguendo anche un t-test per determinare se esistono differenze statistiche significative nei pesi tra neonati di madri fumatrici e non fumatrici :

```
# Esempio di test t per confrontare le medie dei due gruppi
t.test(Peso ~ Fumatrici, data = dati_sorgenti)
```

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Fumatrici
## t = 1.036, df = 114.11, p-value = 0.3024
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -45.52084 145.32289
## sample estimates:
## mean in group 0 mean in group 1
##        3286.247        3236.346
```

Come possiamo vedere un p-value pari a 0.30 suggerisce che, statisticamente, non ci sono prove significative per affermare che ci siano differenze sostanziali nei pesi dei neonati tra i due gruppi di madri (fumatrici e non) nel campione analizzato confermando l’assenza di correlazione tra madre fumatrice e non e relativo peso del neonato. Vediamo una correlazione adesso con la variabile Tipo.parto:

```
Tipo.parto_num <- ifelse(Tipo.parto == "Nat",0,1)
cor(Tipo.parto_num,Peso)
```

```
## [1] -0.002593506
```

Proviamo anche qui ad analizzare l’associazione con la variabile Tipo.parto anche tramite t-test:

```
t.test(Peso ~ Tipo.parto, data = dati_sorgenti)
```

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Tipo.parto
## t = -0.13539, df = 1493.6, p-value = 0.8923
## alternative hypothesis: true difference in means between group Ces and group Nat is not equal to 0
## 95 percent confidence interval:
##  -46.41444  40.42089
## sample estimates:
## mean in group Ces mean in group Nat
##          3282.047          3285.043
```

Anche in questo caso il t-test ci dice che non ci sono prove statisticamente significative per poter affermare che ci siano differenze sostanziali nei pesi dei neonati tra i due gruppi (cesareo e naturale) nel campione considerato.

Se invece prendiamo in considerazione le variabili di osservazione sul neonato, possiamo notare una correlazione dimostrabile con le variabili antropometriche come il cranio e la lunghezza del neonato.:

```
cor(Cranio,Peso)
```

```
## [1] 0.7047755
```

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

```
## [1] 0.7960404
```

Per valutare le correlazioni tra le variabili di un dataframe in R, è possibile eseguire un'analisi tramite una singola chiamata diretta sul dataframe stesso della funzione "cor", filtrando opportunamente i campi non rappresentativi per lo studio. :

```
df_filter <- dati_sorgenti %>% select(Anni.madre,
                                      N.gravidanze,
                                      Fumatrici,
                                      Gestazione,
                                      Lunghezza,
                                      Cranio,
                                      Peso)
cor(df_filter)
```

```
##                Anni.madre N.gravidanze    Fumatrici  Gestazione   Lunghezza
## Anni.madre    1.000000000  0.381220884  0.005644889 -0.13642580 -0.06424092
## N.gravidanze  0.381220884  1.000000000  0.046811482 -0.10150066 -0.06046588
## Fumatrici     0.005644889  0.046811482  1.000000000  0.03221022 -0.02079659
## Gestazione   -0.136425803 -0.101500657  0.032210221  1.00000000  0.61892515
## Lunghezza    -0.064240924 -0.060465881 -0.020796590  0.61892515  1.00000000
## Cranio        0.014857454  0.038827247 -0.008717190  0.46086591  0.60334626
## Peso         -0.023518120  0.002276741 -0.018981789  0.59179210  0.79604039
##                   Cranio         Peso
## Anni.madre    0.01485745 -0.023518120
## N.gravidanze  0.03882725  0.002276741
## Fumatrici    -0.00871719 -0.018981789
## Gestazione    0.46086591  0.591792099
## Lunghezza     0.60334626  0.796040389
## Cranio        1.00000000  0.704775475
## Peso          0.70477547  1.000000000
```

Come è evidente, le variabili con un indice di correlazione significativo spiccano chiaramente dalle altre quando si osserva la colonna del peso utilizzando questa prospettiva.

2 - Creazione del modello di regressione lienare multipla con tutte le variabili:

Procediamo ora alla costruzione del nostro modello di regressione lineare multipla, includendo tutte le variabili che abbiamo individuato come correlate (anche se minimamente) con la nostra variabile target "Peso".

Le variabili che risultano significative sono:

- Gestazione
- Lunghezza
- Cranio
- Sesso (trasformato in variabile dummy 0-1 per Femmine e Maschi, rispettivamente)

Andiamo avanti costruendo il nostro modello di regressione lineare utilizzando queste variabili.

```
sesso_num <- ifelse(Sesso == "F",0,1)
dati_sorgenti$sesso_num <- sesso_num

df_lm <- dati_sorgenti[, c("Peso", "Gestazione", "Lunghezza", "Cranio", "sesso_num")]
```

```r
mod1 <- lm(Peso ~ Gestazione + Lunghezza + Cranio + sesso_num ,data = df_lm)
summary(mod1)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + sesso_num,
##     data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1138.25  -184.45   -17.53   163.32  2627.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6650.5585   135.5473 -49.064  < 2e-16 ***
## Gestazione     31.2805     3.7862   8.262 2.30e-16 ***
## Lunghezza      10.2054     0.3007  33.934  < 2e-16 ***
## Cranio         10.6680     0.4246  25.127  < 2e-16 ***
## sesso_num      79.2016    11.2162   7.061 2.13e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275 on 2494 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7257
## F-statistic:  1653 on 4 and 2494 DF,  p-value: < 2.2e-16
```

La chiamata a `summary()` ci fornisce le stime dei "Coefficienti" associati alle variabili nel nostro modello "mod1". Ad esempio, per ogni settimana aggiuntiva di gestazione, il peso del neonato aumenta di circa 31 grammi. Lo stesso vale per la variabile "Lunghezza", dove ogni millimetro aggiuntivo si traduce in un aumento del peso di circa 10 grammi, e così via.

L'R quadro, intorno al 72,6%, rappresenta un indicatore della validità del nostro modello lineare, ed è confermato anche dall'R quadro aggiustato, che si attesta intorno al 72,5%. Questi valori indicano quanto le variabili nel modello spieghino la variazione nel peso del neonato.

3 - Ricercare il modello lineare "migliore" per la variabile Target:

Per ricercare il modello migliore utilizzeremo la procedura Stepwise che consiste nell'aggiungere o togliere le variabili una per volta e valutare passo dopo passo i risultati ottenuti. Nel nostro caso partendo dal modello che ingloba tutte le variabili che hanno rilevato un minimo di correlazione con la variabile Target vedremo quali tra esse possiamo eliminare per arrivare ad avere un modello quanto piu semplice possibile che abbia il minor numero di variabili con un R Quadro stabile o se non addirittura piu alto.

Ripartiamo con elencare le variabili utilizzate nel nostro modello che riguardano:

- Gestazione
- Lunghezza
- Cranio
- Sesso (convertito in dummy 0-1 per Femmine e Maschi rispettivamente)

Procediamo a togliere la variabile "Cranio", creandoci il modello `mod2`.

```r
mod2 <- update(mod1,~.-Cranio)
summary(mod2)
```

```
##
## Call:
```

```
## lm(formula = Peso ~ Gestazione + Lunghezza + sesso_num, data = df_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1138.2  -192.9   -22.7   186.8  3618.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5223.5515   137.7488 -37.921  < 2e-16 ***
## Gestazione     44.4900     4.1966  10.601  < 2e-16 ***
## Lunghezza      13.6013     0.3007  45.233  < 2e-16 ***
## sesso_num      90.4625    12.5434   7.212 7.28e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.8 on 2495 degrees of freedom
## Multiple R-squared:  0.6568, Adjusted R-squared:  0.6563
## F-statistic:  1591 on 3 and 2495 DF,  p-value: < 2.2e-16
```

In questo caso, se confrontiamo il modello "mod2" con il "mod1", notiamo che il "mod2" avrebbe una diminuzione nell'R^2, indicando una perdita di affidabilità. Per comprendere l'impatto della variabile "Lunghezza", procediamo eliminandola e osservando come ciò influisca sull'R^2, creando il modello "mod3":

```
mod3 <- update(mod1,~.-Lunghezza)
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Cranio + sesso_num, data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1316.22  -219.02   -10.93   210.48  1532.16
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6213.9501   163.1053 -38.098   <2e-16 ***
## Gestazione     92.6135     4.0216  23.029   <2e-16 ***
## Cranio         17.1425     0.4585  37.391   <2e-16 ***
## sesso_num     118.6295    13.4848   8.797   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 332.5 on 2495 degrees of freedom
## Multiple R-squared:  0.5996, Adjusted R-squared:  0.5991
## F-statistic:  1246 on 3 and 2495 DF,  p-value: < 2.2e-16
```

Questa conferma anche in questo caso una diminuzione dell'affidabilità del modello.

Riproviamo lo stesso per la variabile "sesso_num" che abbiamo definito in modalità dummy andandoci a creare il modello "mod4" :

```
mod4 <- update(mod1,~.-sesso_num)
summary(mod4)
```

```
##
## Call:
```

```
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio, data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1105.78  -183.32   -12.75   166.51  2623.85
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6776.8697   135.6712 -49.951   <2e-16 ***
## Gestazione     31.6946     3.8227   8.291   <2e-16 ***
## Lunghezza      10.4254     0.3020  34.516   <2e-16 ***
## Cranio         10.7878     0.4284  25.184   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.7 on 2495 degrees of freedom
## Multiple R-squared:  0.7206, Adjusted R-squared:  0.7203
## F-statistic:  2145 on 3 and 2495 DF,  p-value: < 2.2e-16
```

Nel caso in cui la variabile "sesso_num" venga rimossa dal modello, si osserva una lieve diminuzione dell'affidabilità del modello. Proviamo infine con la eliminazione dal modello della variabile "Gestazione" :

```
mod5 <- update(mod1,~.-Gestazione)
summary(mod5)
```

```
##
## Call:
## lm(formula = Peso ~ Lunghezza + Cranio + sesso_num, data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1267.86  -186.60   -17.09   167.23  2794.01
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6184.3033   124.8896 -49.518  < 2e-16 ***
## Lunghezza      11.3915     0.2678  42.537  < 2e-16 ***
## Cranio         11.1550     0.4261  26.180  < 2e-16 ***
## sesso_num      80.6369    11.3650   7.095 1.68e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 278.7 on 2495 degrees of freedom
## Multiple R-squared:  0.7186, Adjusted R-squared:  0.7183
## F-statistic:  2124 on 3 and 2495 DF,  p-value: < 2.2e-16
```

Se decidessimo di eliminare la variabile "Gestazione", anche qui seppur in misura ridotta, si verificherebbe una diminuzione dell'affidabilità del modello. Il modello pertanto piu affidabile risulterebbe essere il modello di partenza ovvero il mod1. Prossimo passo: procediamo con l'analisi utilizzando i criteri dell'AIC e del BIC per approfondire questa ipotesi. Secondo questi strumenti, il modello che presenta il valore più basso sarebbe da considerarsi preferibile in termini di adeguatezza e complessità.

```
AIC(mod1,mod2,mod3,mod4,mod5)
```

```
##      df      AIC
## mod1  6 35172.37
```

```
## mod2  5 35734.29
## mod3  5 36119.00
## mod4  5 35219.84
## mod5  5 35237.84
```

```
BIC(mod1,mod2,mod3,mod4,mod5)
```

```
##      df      BIC
## mod1  6 35207.31
## mod2  5 35763.41
## mod3  5 36148.12
## mod4  5 35248.96
## mod5  5 35266.96
```

Il modello originale (mod1) presenta un miglior coefficiente di determinazione (R^2) sia utilizzando l'AIC che il BIC confermando la nostra ipotesi che tale modello sia il migliore. Proviamo adesso ad utilizzare la funzione prevista in R che mi ricerca lei il modello migliore: la funzione stepAIC. Il suo obiettivo principale è aiutare nella costruzione del miglior modello possibile, esplorando una serie di modelli inclusi o esclusi di variabili. Il processo avviene in modo iterativo, aggiungendo o eliminando una variabile alla volta e valutando come ciò influenzi il criterio di informazione del modello (AIC o BIC); quello che farà la funzione sarebbe dunque l'automazione di quello che abbiamo appena mostrato manualmente in pratica:

```
library(MASS)
```

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

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

```
stepwise.mod <- MASS::stepAIC(mod1,direction="both",k=2)
```

```
## Start:  AIC=28078.52
## Peso ~ Gestazione + Lunghezza + Cranio + sesso_num
##
##              Df Sum of Sq       RSS   AIC
## <none>                     188676676 28078
## - sesso_num   1   3772222 192448898 28126
## - Gestazione  1   5163585 193840262 28144
## - Cranio      1  47763274 236439950 28640
## - Lunghezza   1  87112693 275789369 29025
```

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

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + sesso_num,
##     data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1138.25  -184.45   -17.53   163.32  2627.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6650.5585   135.5473 -49.064  < 2e-16 ***
## Gestazione     31.2805     3.7862   8.262 2.30e-16 ***
```

```
## Lunghezza      10.2054     0.3007  33.934  < 2e-16 ***
## Cranio         10.6680     0.4246  25.127  < 2e-16 ***
## sesso_num      79.2016    11.2162   7.061 2.13e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275 on 2494 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7257
## F-statistic:  1653 on 4 and 2494 DF,  p-value: < 2.2e-16
```

Sembra che anche secondo questa metodologia, il modello migliore sia `mod1`, poiché offre il miglior contributo all'R quadro utilizzando il minor numero di variabili.

Utilizzeremo quindi `mod1` come modello definitivo.

4 - Verifica interazioni o effetti non lineari:

In questo passaggio, esamineremo se ci sono interazioni rilevanti tra le variabili scelte per il modello "mod1". Per fare ciò, rappresenteremo le interazioni tra di esse con l'implementazione del modello "mod6" indicato di seguito.:

```
mod6 <- update(mod1,~.+Lunghezza*sesso_num
                      + Cranio*sesso_num
                      + Gestazione*sesso_num
                      + Lunghezza*Cranio
                      + Lunghezza*Gestazione
                      + Gestazione*Cranio)
summary(mod6)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + sesso_num +
##     Lunghezza:sesso_num + Cranio:sesso_num + Gestazione:sesso_num +
##     Lunghezza:Cranio + Gestazione:Lunghezza + Gestazione:Cranio,
##     data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1135.70  -182.55   -12.45   164.18  2540.71
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.186e+02  1.124e+03  -0.284  0.77676
## Gestazione           -1.750e+02  6.352e+01  -2.754  0.00593 **
## Lunghezza             1.171e+01  4.992e+00   2.346  0.01904 *
## Cranio               -6.739e+00  7.006e+00  -0.962  0.33615
## sesso_num            -3.160e+01  2.796e+02  -0.113  0.91002
## Lunghezza:sesso_num   9.677e-01  6.069e-01   1.594  0.11096
## Cranio:sesso_num     -6.550e-01  8.711e-01  -0.752  0.45220
## Gestazione:sesso_num -3.884e+00  7.692e+00  -0.505  0.61369
## Lunghezza:Cranio     -9.311e-03  1.402e-02  -0.664  0.50665
## Gestazione:Lunghezza  3.760e-02  1.072e-01   0.351  0.72580
## Gestazione:Cranio     5.822e-01  2.019e-01   2.884  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 273.4 on 2488 degrees of freedom
## Multiple R-squared:  0.7301, Adjusted R-squared:  0.729
## F-statistic:   673 on 10 and 2488 DF,  p-value: < 2.2e-16
```

Utilizzando summary(mod6), si puo' valutare la significatività delle interazioni guardando i risultati ottenuti dai p-values associati ai coefficienti delle interazioni in quanto se il p-value associato a una particolare interazione è inferiore al livello di significatività , potrebbe indicare che quella specifica interazione è significativa nel modello. Tra le sei tipologie di interazioni considerate nel modello "mod6", sembra che l'interazione più significativa coinvolga le variabili "Gestazione e Cranio". Proviamo quindi a sviluppare un modello, denominato "mod7", che includa solamente questa interazione per valutare il suo impatto sull'affidabilità previsionale del modello.

```
mod7 <- update(mod1,~.+Gestazione*Cranio)
summary(mod7)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + sesso_num +
##     Gestazione:Cranio, data = df_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1125.54  -181.16   -13.87   163.88  2682.23
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -252.33335 1108.34750  -0.228  0.81992
## Gestazione        -139.35142   29.57929  -4.711 2.60e-06 ***
## Lunghezza           10.42183    0.30109  34.613  < 2e-16 ***
## Cranio              -9.41506    3.47883  -2.706  0.00685 **
## sesso_num           73.39419   11.18778   6.560 6.51e-11 ***
## Gestazione:Cranio    0.52588    0.09042   5.816 6.80e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.3 on 2493 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7292
## F-statistic:  1346 on 5 and 2493 DF,  p-value: < 2.2e-16
```

L'incremento minimo nell'R^2 con l'aggiunta di un'altra variabile sembra suggerire che l'ipotesi di includere l'interazione nel modello potrebbe non essere plausibile, poiché non sembra essere rilevante in modo significativo.

5 - Diagnostica approfondita dei residui del modello e di potenziali valori influenti:

A partire dal modello scelto, "mod1", procediamo con l'analisi dei residui, i quali dovrebbero mostrare le seguenti caratteristiche:

- Normalità
- Omoschedasticità
- Incorrelazione
- Media zero

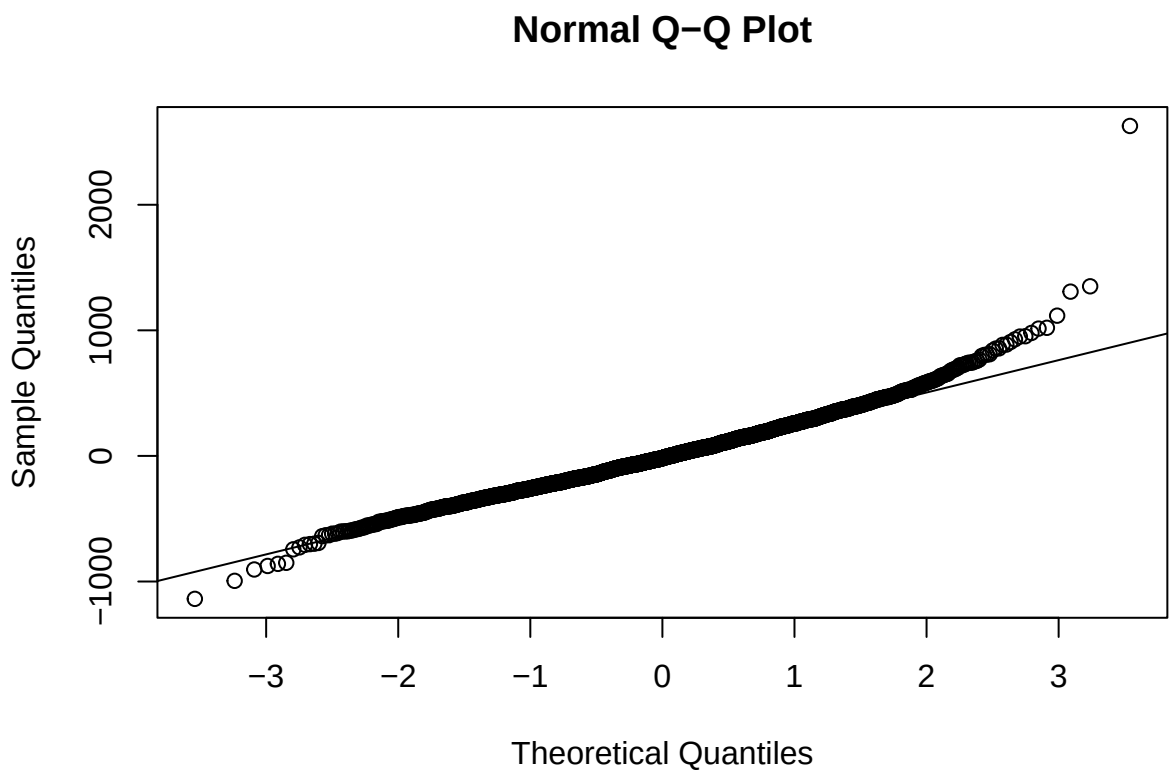
Iniziamo con il test di Shapiro-Wilk per verificare se i residui seguono una distribuzione normale.

```
residui <- residuals(mod1)
shapiro.test(residui)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residui
## W = 0.97422, p-value < 2.2e-16
```

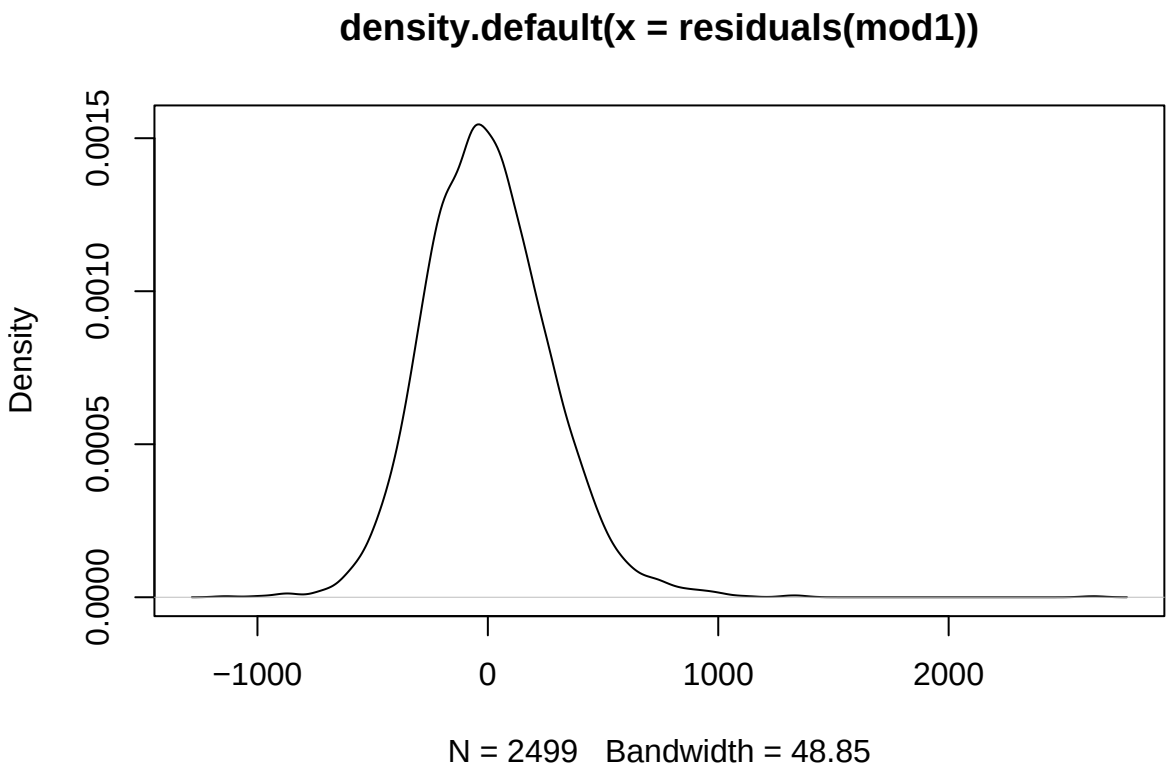
Andiamo a vedere il confronto da un punto di vista grafico :

```
qqnorm(residui)
qqline(residui)
```



Dai punti disposti lungo la linea bisettrice nel grafico, sembra che i residui seguano effettivamente una distribuzione normale. Osserviamo direttamente il grafico di densità per vedere come la curva dei residui si avvicina alla distribuzione normale.

```
plot(density(residuals(mod1)))
```



Andiamo ora a vedere il test di omoschedasticità ; per farlo utilizzeremo il test di Breusch-Pagan tramite la funzione R "btest" :

```
library(lmtest)
```

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

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

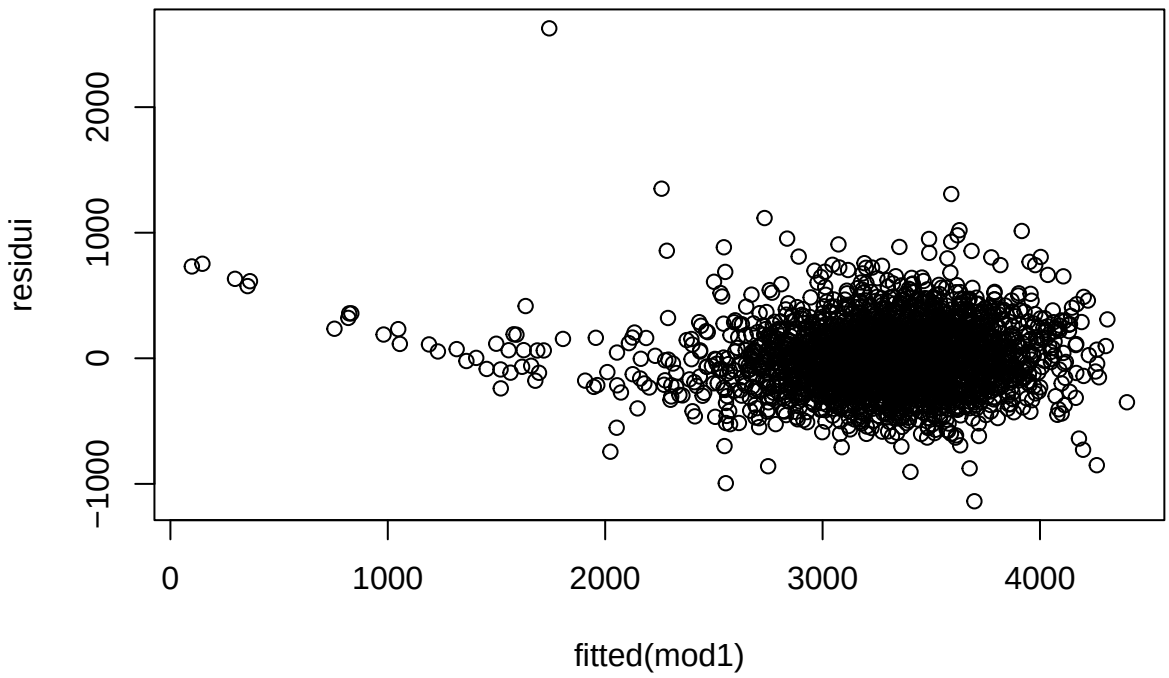
```
## I seguenti oggetti sono mascherati da 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(mod1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 89.082, df = 4, p-value < 2.2e-16
```

Questo test produce un valore p che, se inferiore al livello di significatività, suggerisce la presenza di eteroschedasticità nei residui. Al contrario, se il valore p risulta essere maggiore, indicherebbe una natura omoschedastica dei residui. Esploriamo questa caratteristica anche graficamente per una comprensione più approfondita. :

```
plot(fitted(mod1),residui)
```

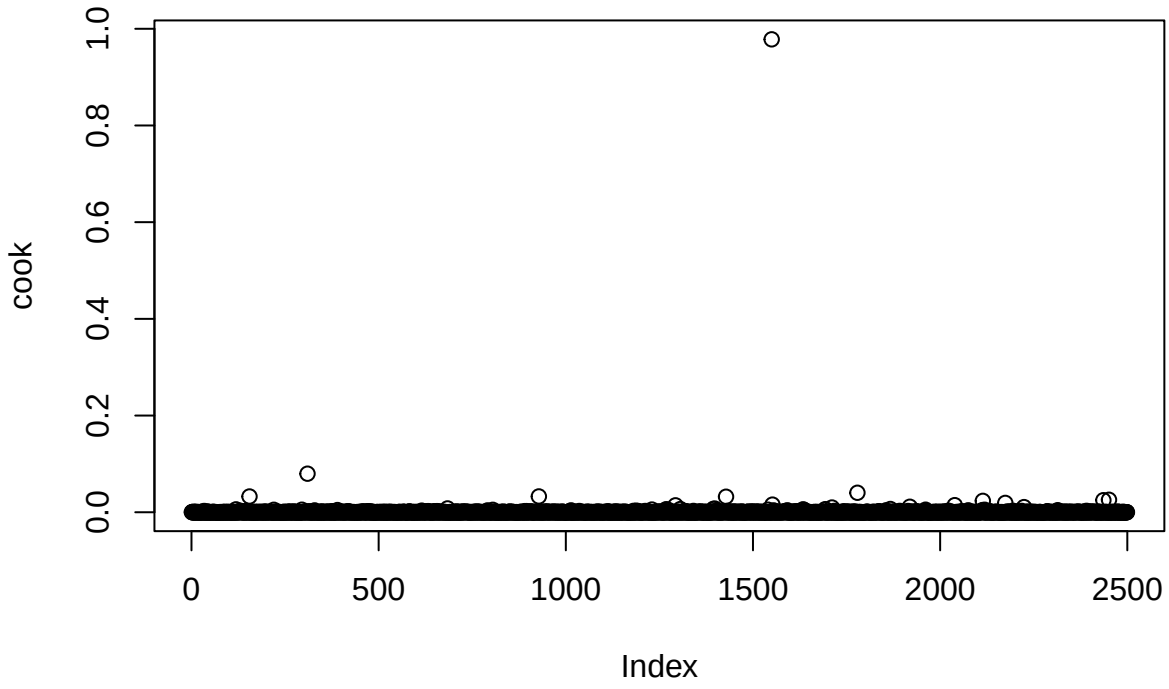


In questo contesto, i numeri dovrebbero distribuirsi intorno allo zero.

L'osservazione di una nuvola di punti concentrati attorno allo zero nel grafico dei residui rispetto ai valori predetti potrebbe suggerire che, in generale, i residui non mostrano una chiara eteroschedasticità. Tuttavia, l'esito negativo del test di Breusch-Pagan potrebbe essere causato dalla presenza di alcuni outlier che influenzano la varianza dei residui.

Per indagare ulteriormente, possiamo analizzare la presenza di tali outlier utilizzando la "Distanza di Cook". Essa ci permette di individuare specifici casi che influenzano in modo significativo il modello.

```
cook <- cooks.distance(mod1)
plot(cook)
```



Come possiamo vedere abbiamo individuato un outlier ; proviamo a definire quanto è la sua distanza di cook :

```
max(cook)
```

```
## [1] 0.9781323
```

Cerchiamo di individuare adesso l'outlier :

```
cook_d <- cooks.distance(mod1)
indice_outlier <- which.max(cook_d)
print(indice_outlier)
```

```
## 1551
## 1550
```

Proviamo adesso ad eliminarlo, procedere dunque al ricaricamento del dataset bonificato e infine al ricalcolo del modello aggiornato secondo quanto sotto riportato :

```
dati_senza_outlier <- dati_sorgenti[-indice_outlier,] #rimuovo il valore anomalo
attach(dati_senza_outlier)
```

```
## Il seguente oggetto è mascherato _da_ .GlobalEnv:
##
##     sesso_num
```

```
## I seguenti oggetti sono mascherati da dati_sorgenti:
##
##     Anni.madre, Anni.madre_classi, Cranio, Fumatrici, Gestazione,
##     Lunghezza, N.gravidanze, Ospedale, Peso, Sesso, Tipo.parto
```

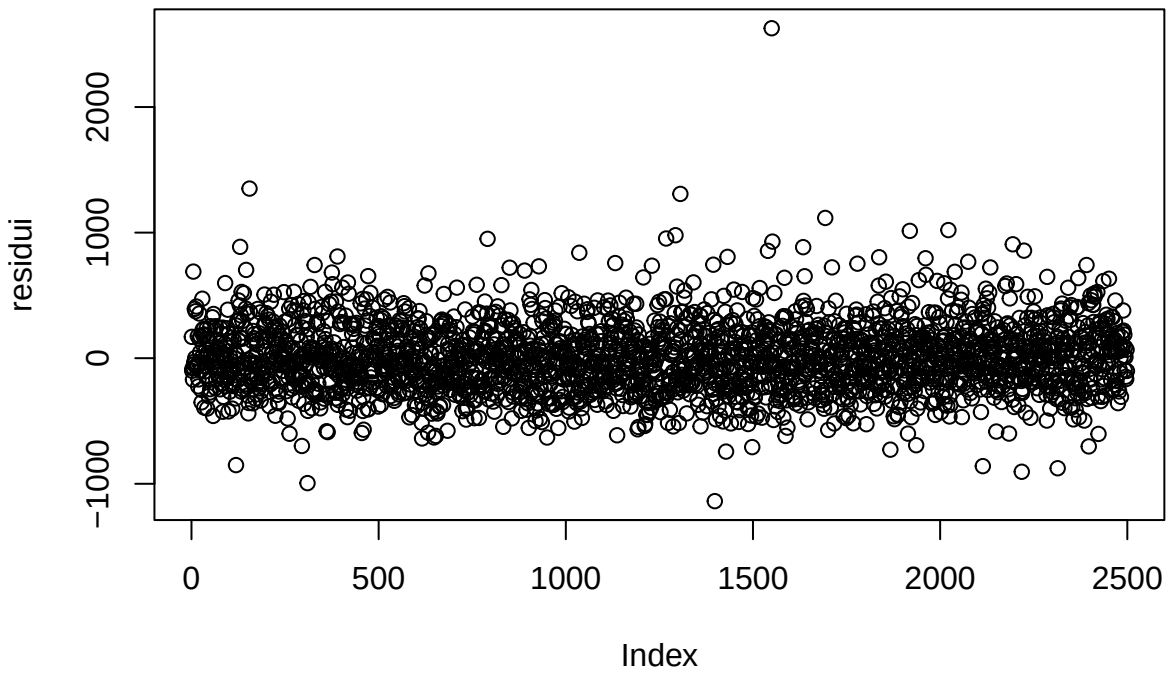
```r
sesso_num <- ifelse(Sesso == "F",0,1)
dati_senza_outlier$sesso_num <- sesso_num
mod1 <- lm(Peso ~ Gestazione
           + Lunghezza
           + Cranio
           + sesso_num ,data = dati_senza_outlier)
library(lmtest)
bptest(mod1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 9.3326, df = 4, p-value = 0.0533
```

Ora il pvalue risulta leggermente superiore al livello di significatività quindi siamo in omoschedasticità.

Andiamo ora a vedere che non esista alcuna condizione di correlazione tra i residui:

```r
plot(residui)
```



Analogamente andiamo ad effettuare il test di Durbin-Watson tramite la funzione in R dwtest():

```r
library(lmtest)
dwtest(mod1)
```

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

```
## data:  mod1
## DW = 1.956, p-value = 0.1353
## alternative hypothesis: true autocorrelation is greater than 0
```

In questo caso il test di Durbin-Watson indica che il valore p è superiore al livello di significatività quindi gli stessi non sono autocorrelati pertanto il modello supera anche la condizione di autocorrelazione.

Infine vediamo il test se la media dei residui è vicino allo 0:

```
mean(residui)
```

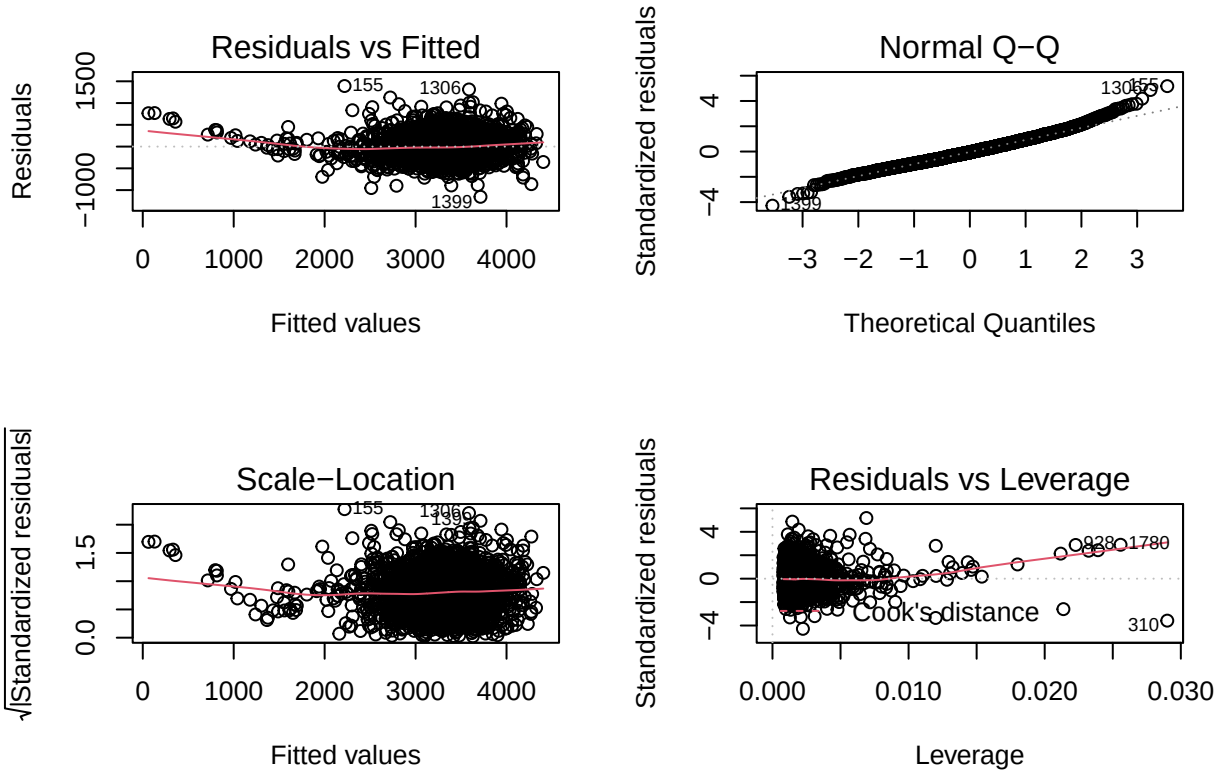
```
## [1] 3.886947e-16
```

Anche in questo caso il test viene superato nella diagnostica. A questo punto possiamo affermare che la diagnosi risulta superata nei punti indicati per il modello mod1 prescelto.

6 - Considerazioni sulla bontà del modello:

Il modello che abbiamo sviluppato mostra un R quadro del 72%, il che indica una misura ragionevole di affidabilità. È rassicurante notare che supera con successo la diagnosi dei residui, confermando la sua capacità predittiva complessiva. Possiamo esaminare i controlli principali relativi alla diagnosi dei residui appena eseguita utilizzando la seguente chiamata:

```
par(mfrow=c(2,2))
plot(mod1)
```



7 - Previsione del peso di una neonata, in considerazione della madre alla terza gravidanza, e gestazione alla 39esima settimana:

Come abbiamo visto il N.gravidanze non è variabile correlata al peso ma le altre due ne hanno confermato la correlazione. Utilizzeremo dunque la funzione predict in R per utilizzare il nostro modello "mod1" a fare inferenza per la previsione richiesta :

```
predict(mod1, newdata = data.frame(Gestazione=39, sesso_num=0, Cranio=340, Lunghezza=494))
```

```
##        1
## 3236.269
```

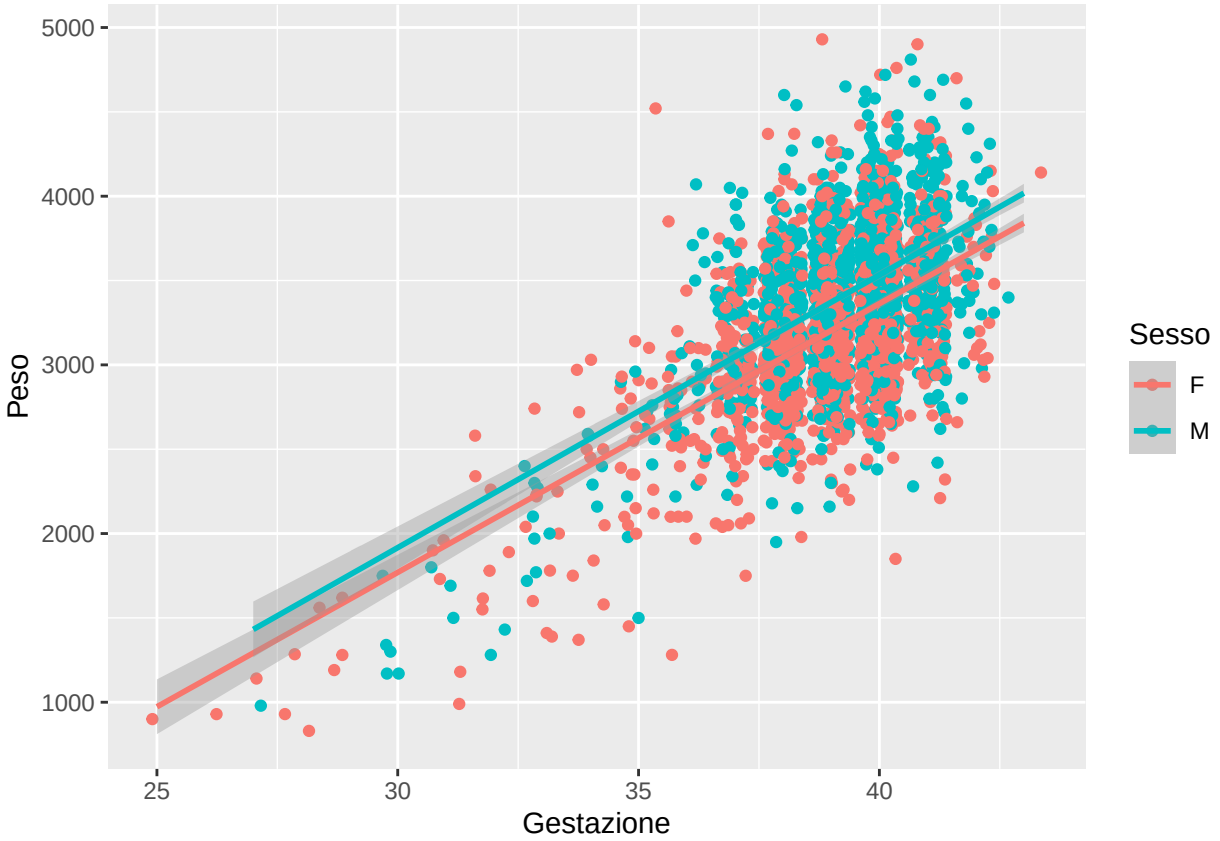
Abbiamo utilizzato il valore medio del dataset per "Cranio" e "Lunghezza" nel nostro modello per non influenzare le previsioni. Secondo il modello costruito, il peso previsto in grammi per una neonata nata da madre alla 39ª settimana è di 3237g.

8 - Rappresentazione grafica del modello realizzato:

Proviamo a rappresentare il modello nelle variabili principali da un punto di vista grafico :

```
library(ggplot2)
ggplot(data=dati_sorgenti)+
  geom_point(aes(x=Gestazione,
                 y=Peso,
                 col=Sesso), position="jitter")+
  geom_smooth(aes(x=Gestazione,
                 y=Peso,
                 col=Sesso) , method="lm")
```

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



Come si può notare dal grafico, c'è una differenza nel peso dei neonati in base al sesso, con un aumento tendenziale del peso all'aumentare della "Gestazione". Come osservato, il numero più frequente di settimane di gestazione è 39, come evidenziato nel grafico. La "Gestazione" mostra una correlazione lineare con il peso, mentre il sesso aggiunge un ulteriore aspetto alla correlazione con il peso del neonato.