Un modello statistico per prevedere il peso dei neonati

Dario De Caro

2024-10-01

Carichiamo i pacchetti necessari

```
library(readr)
library(knitr)
library(ggplot2)
library(ggpubr)
library(lmtest)
library(MASS)
library(car)
library(rgl)
library(moments)
```

Importiamo il dataset

```
data <- read_csv("neonati.csv")
attach(data)
```

Verifichiamo le prime righe del dataset

```
kable(head(data))
```

Anni.madre	N.gravidanze	Fumatrici	Gestazione	Peso	Lunghezza	Cranio	Tipo.parto	Ospedale	Sesso
26	0	0	42	3380	490	325	Nat	osp3	M
21	2	0	39	3150	490	345	Nat	osp1	F
34	3	0	38	3640	500	375	Nat	osp2	M
28	1	0	41	3690	515	365	Nat	osp2	M
20	0	0	38	3700	480	335	Nat	osp3	F
32	0	0	40	3200	495	340	Nat	osp2	F

Verifichiamo la struttura del dataset

```
str(data)
```

```
## spc_tbl_ [2,500 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Anni.madre  : num [1:2500] 26 21 34 28 20 32 26 25 22 23 ...
##  $ N.gravidanze: num [1:2500] 0 2 3 1 0 0 1 0 1 0 ...
##  $ Fumatrici   : num [1:2500] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Gestazione  : num [1:2500] 42 39 38 41 38 40 39 40 40 41 ...
##  $ Peso        : num [1:2500] 3380 3150 3640 3690 3700 3200 3100 3580 3670 3700 ...
##  $ Lunghezza   : num [1:2500] 490 490 500 515 480 495 480 510 500 510 ...
##  $ Cranio      : num [1:2500] 325 345 375 365 335 340 345 349 335 362 ...
##  $ Tipo.parto  : chr [1:2500] "Nat" "Nat" "Nat" "Nat" ...
##  $ Ospedale    : chr [1:2500] "osp3" "osp1" "osp2" "osp2" ...
##  $ Sesso       : chr [1:2500] "M" "F" "M" "M" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Anni.madre = col_double(),
##   ..   N.gravidanze = col_double(),
##   ..   Fumatrici = col_double(),
##   ..   Gestazione = col_double(),
##   ..   Peso = col_double(),
##   ..   Lunghezza = col_double(),
##   ..   Cranio = col_double(),
##   ..   Tipo.parto = col_character(),
##   ..   Ospedale = col_character(),
##   ..   Sesso = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Il dataset contiene 2500 osservazioni e 10 variabili. In particolare, sono presenti 3 variabili quantitative continue (Peso, Lunghezza e Cranio), 3 quantitative discrete (Anni madre, Gestazione e N. Gravidanze), e 4 variabili qualitative (Sesso, Ospedale, Tipo di Parto e Fumatrici). L'obiettivo dello studio è scoprire se è possibile prevedere il peso del neonato alla nascita date tutte le altre variabili. In particolare, si vuole studiare una relazione con le variabili della madre, per capire se queste hanno o meno un effetto significativo.

Calcoliamo alcune statistiche descrittive

```
df <- data[, sapply(data, is.numeric)]
df <- subset(df, select = -Fumatrici)
kable(summary(df))
```

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

Skewness

```
kable(sapply(df, skewness))
```

	x
Anni.madre	0.0428115
N.gravidanze	2.5142541
Gestazione	-2.0653133
Peso	-0.6470308
Lunghezza	-1.5146991
Cranio	-0.7850527

Le variabili Lunghezza e Gestazione mostrano una asimmetria negativa, mentre N. di Gravidanze positiva. Le altre variabili sono pressoché simmetriche.

Kurtosis

```
kable(sapply(df, kurtosis))
```

	x
Anni.madre	3.380416
N.gravidanze	13.989406
Gestazione	11.258150
Peso	5.031532
Lunghezza	9.487174
Cranio	5.946206

Tutte le variabili presentano una distribuzione leptocurtica.

```
df <- data[, sapply(data, is.factor) | sapply(data, is.character)]
df$Fumatrici = Fumatrici
kable(lapply(df, table))
```

Var1	Freq
Ces	728
Nat	1772

Var1	Freq
osp1	816
osp2	849
osp3	835

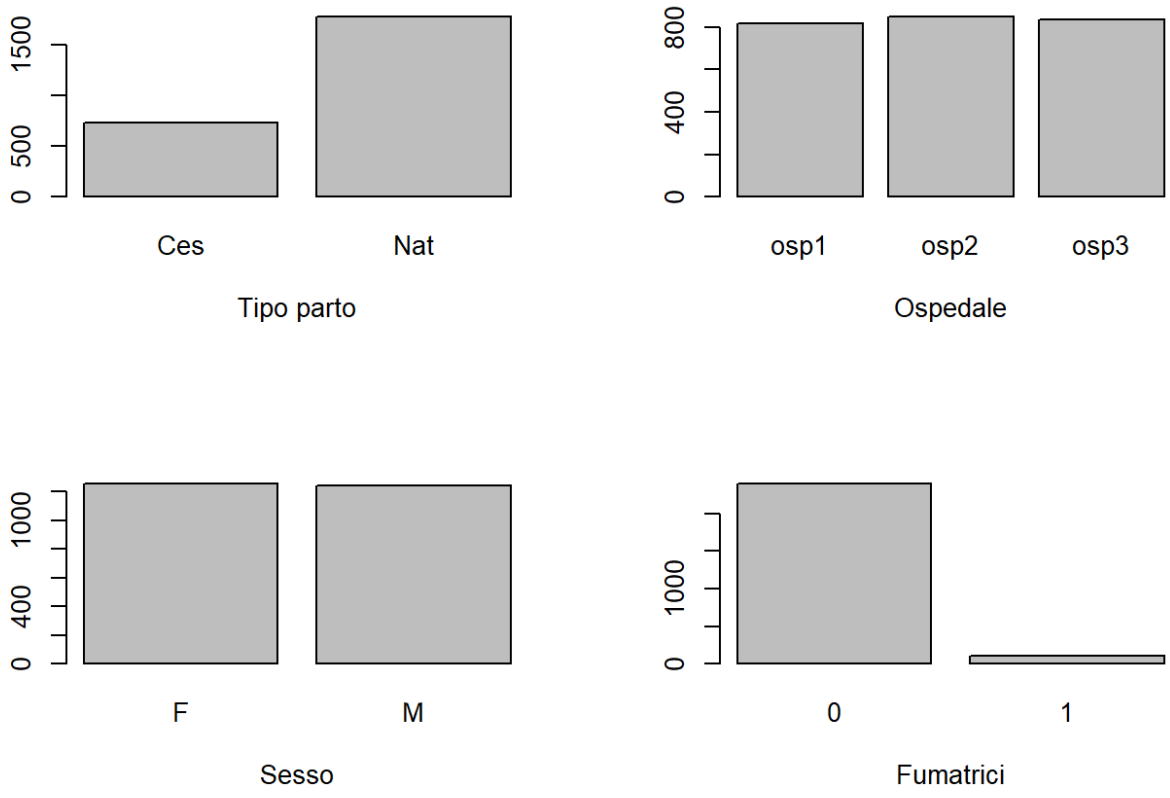
Var1	Freq
F	1256
M	1244

Var1	Freq
0	2396
1	104

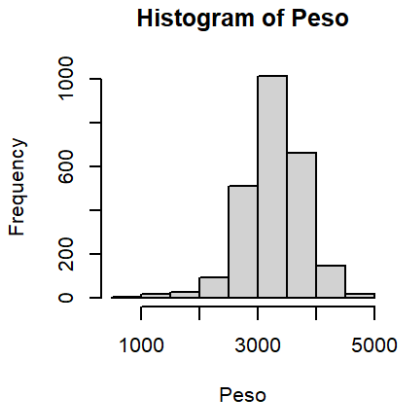
Le variabili Ospedale e Sesso mostrano una distribuzione bilanciata, mentre Tipo di parto e Fumatrici appaiono sbilanciate rispettivamente verso i parti Naturali e le madri non fumatrici.

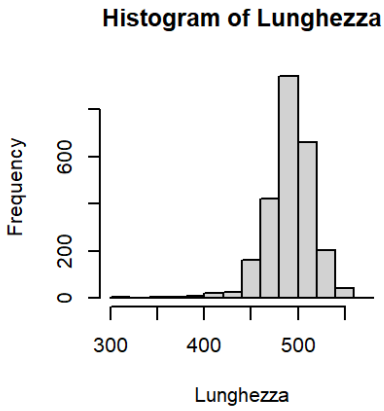
Visualizziamo graficamente

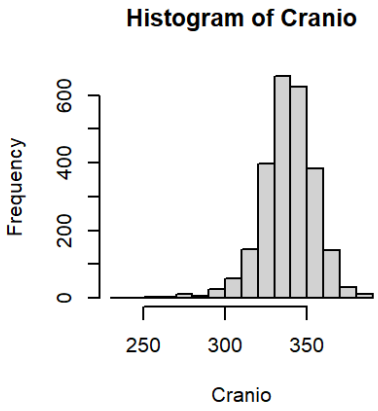
```
par(mfrow=c(2, 2))

barplot(table(Tipo.parto), xlab = 'Tipo parto')
barplot(table(Ospedale), xlab = 'Ospedale')
barplot(table(Sesso), xlab = 'Sesso')
barplot(table(Fumatrici), xlab = 'Fumatrici')
```

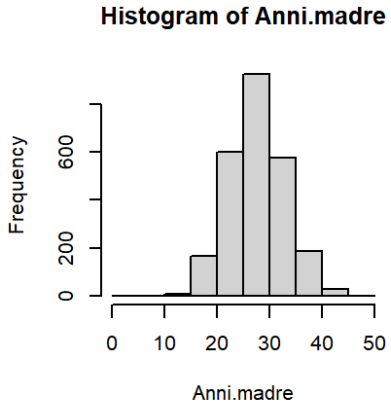


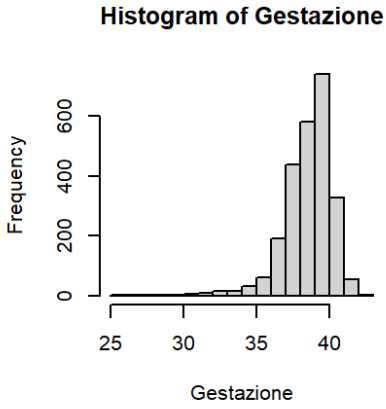

```
par(mfrow=c(2, 3))

hist(Peso)
hist(Lunghezza)
hist(Cranio)
hist(Anni.madre)
hist(Gestazione)
hist(N.gravidanze)
```

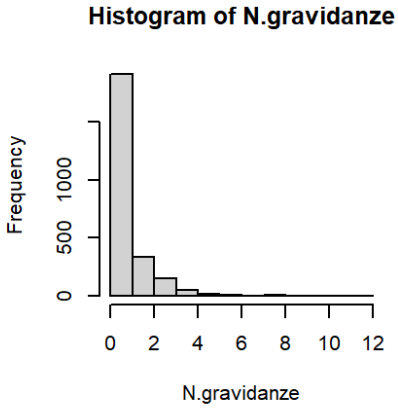












L’analisi grafica conferma i risultati ottenuti precedentemente.

Test t per la media del peso

```
shapiro.test(data$Peso)
```

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

```
t.test(data$Peso, mu=3300)
```

```
##
##  One Sample t-test
##
## data:  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
```

La media del peso di questo campione di neonati è significativamente uguale a quelle della popolazione. Tuttavia, il campione non mostra una distribuzione normale.

Test t per la media della lunghezza

```
shapiro.test(data$Lunghezza)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Lunghezza
## W = 0.90941, p-value < 2.2e-16
```

```
t.test(data$Lunghezza, mu=500)
```

```
##
##  One Sample t-test
##
## data:  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
```

La media della lunghezza di questo campione di neonati è significativamente uguale a quelle della popolazione. Tuttavia, il campione non mostra una distribuzione normale.

I valori medi della popolazione sono stati ricavati da:

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. (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.)

Test per verificare differenze significative nel peso tra maschi e femmine

Test per verificare le ipotesi

```
shapiro.test(data$Peso[data$Sesso == 'M'])
```

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

```
shapiro.test(data$Peso[data$Sesso == 'F'])
```

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

```
bptest(data$Peso  ~ data$Sesso)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  data$Peso ~ data$Sesso
## BP = 2.3503, df = 1, p-value = 0.1253
```

Le distribuzioni mostrano varianza omogenea, ma non distribuzione normale, pertanto è consigliabile esegire anche test non parametrici.

```
t.test(Peso ~ Sesso, data = data)
```

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

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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Peso by Sesso
## W = 538641, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Entrambi i test confermano una differenza significativa nel peso tra i due sessi.

Test per verificare differenze significative nella lunghezza tra maschi e femmine

Test per verificare le ipotesi

```
shapiro.test(data$Lunghezza[data$Sesso == 'M'])
```

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

```
shapiro.test(data$Lunghezza[data$Sesso == 'F'])
```

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

```
bptest(data$Lunghezza  ~ data$Sesso)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  data$Lunghezza ~ data$Sesso
## BP = 5.2544, df = 1, p-value = 0.02189
```

Le distribuzioni mostrano varianza non omogenea e distribuzione non normale, pertanto è consigliabile esegire anche test non parametrici.

```
t.test(Lunghezza ~ Sesso, data = data)
```

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

```
wilcox.test(Lunghezza ~ Sesso, data = data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Lunghezza by Sesso
## W = 594455, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Entrambi i test confermano una differenza significativa nella lunghezza tra i due sessi.

Test per verificare differenze significative nel diametro del cranio tra maschi e femmine

Test per verificare le ipotesi

```
shapiro.test(data$Cranio[data$Sesso == 'M'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Cranio[data$Sesso == "M"]
## W = 0.97046, p-value = 3.006e-15
```

```
shapiro.test(data$Cranio[data$Sesso == 'F'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Cranio[data$Sesso == "F"]
## W = 0.95543, p-value < 2.2e-16
```

```
bptest(data$Cranio  ~ data$Sesso)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  data$Cranio ~ data$Sesso
## BP = 1.8761, df = 1, p-value = 0.1708
```

Le distribuzioni mostrano varianza omogenea e distribuzione non normale, pertanto è consigliabile esegire anche test non parametrici.

```
t.test(Cranio ~ Sesso, data = data)
```

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

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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Cranio by Sesso
## W = 641638, p-value = 9.633e-15
## alternative hypothesis: true location shift is not equal to 0
```

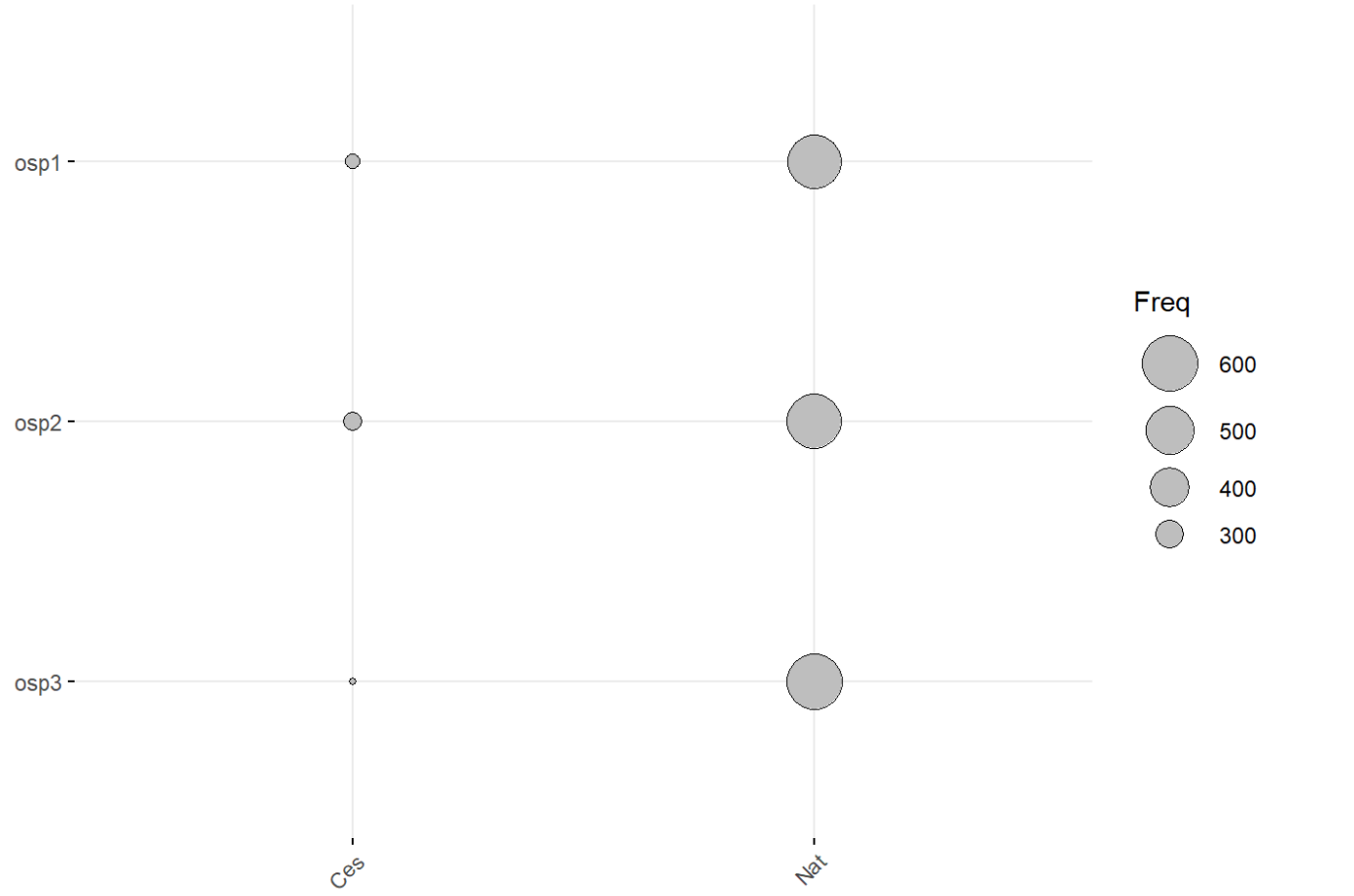
Entrambi i test confermano una differenza significativa nella lunghezza tra i due sessi.

Tabella di contingenza tra tipo di parto e ospedale

```
tabella_parti <- table(data$Tipo.parto, data$Ospedale)
kable(tabella_parti)
```

	osp1	osp2	osp3
Ces	242	254	232
Nat	574	595	603

```
ggballoonplot(as.data.frame(tabella_parti))
```



Sia la tabella, che il grafico non mostrano nessuna tendenza di maggiori parti cesarei tra ospedali.

Test chi-quadrato per verificare differenze significative

```
chisq.test(tabella_parti)
```

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

Il test conferma l'ipotesi iniziale.

Correlazione con variabili qualitative

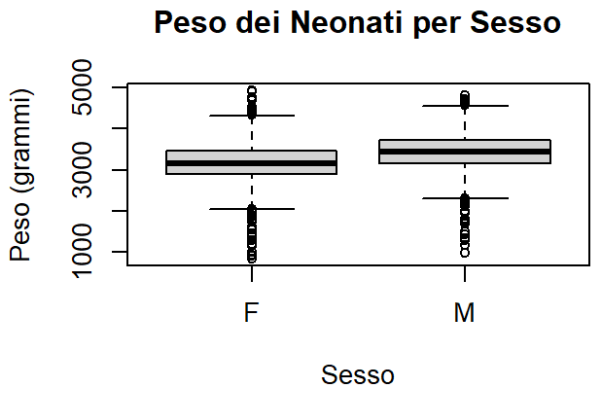
```
par(mfrow=c(2, 2))

boxplot(data$Peso ~ data$Sesso, main="Peso dei Neonati per Sesso", xlab='Sesso', ylab="Peso (gra
mmi)")

boxplot(data$Peso ~ data$Ospedale, main="Peso dei Neonati per Ospedale",xlab = 'Ospedale', ylab
="Peso (grammi)")
pairwise.t.test(data$Peso, data$Ospedale)
```

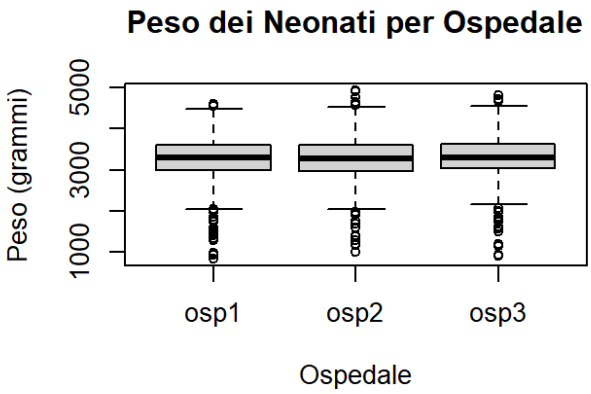
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$Peso and data$Ospedale
##
##      osp1 osp2
## osp2 0.99 -
## osp3 0.33 0.33
##
## P value adjustment method: holm
```

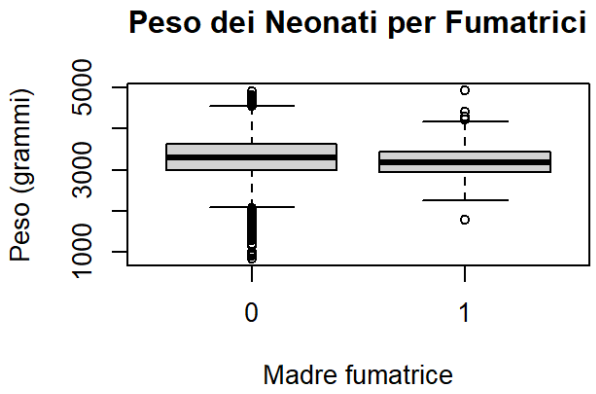
```
boxplot(data$Peso ~ data$Fumatrici, main="Peso dei Neonati per Fumatrici", xlab='Madre fumatric
e', ylab="Peso (grammi)")
t.test(Peso ~ Fumatrici, data = data)
```

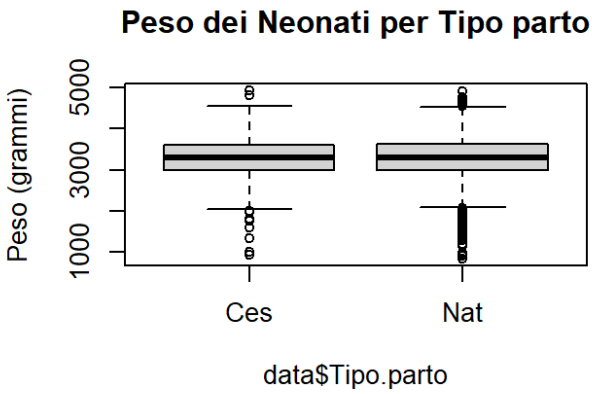
```
##
##  Welch Two Sample t-test
##
## data:  Peso by Fumatrici
## t = 1.034, df = 114.1, p-value = 0.3033
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
0
## 95 percent confidence interval:
##  -45.61354 145.22674
## sample estimates:
## mean in group 0 mean in group 1
##        3286.153        3236.346
```

```
boxplot(data$Peso ~ data$Tipo.parto, main="Peso dei Neonati per Tipo parto", ylab="Peso (gramm
i)")
```










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

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Tipo.parto
## t = -0.12968, df = 1493, p-value = 0.8968
## alternative hypothesis: true difference in means between group Ces and group Nat is not equal
to 0
## 95 percent confidence interval:
##  -46.27992  40.54037
## sample estimates:
## mean in group Ces mean in group Nat
##          3282.047          3284.916
```

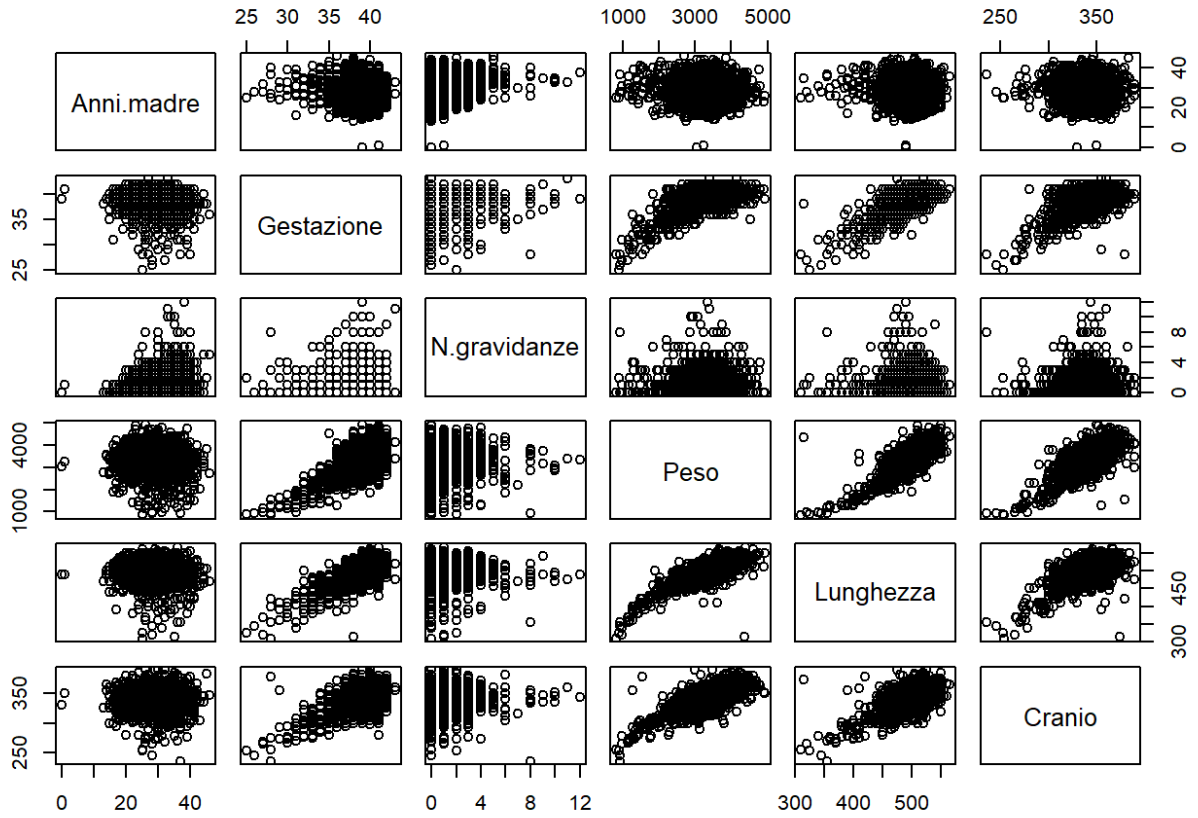
Il sesso sembra essere la variabile qualitativa che più influisce sul peso.

Matrice di correlazione per le variabili numeriche

```
round(cor(data[c("Anni.madre", "Gestazione", "N.gravidanze",
           "Peso", "Lunghezza", "Cranio")]), digits = 2)
```

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

```
pairs(data[c("Anni.madre", "Gestazione", "N.gravidanze",
           "Peso", "Lunghezza", "Cranio")])
```



Il peso appare linearmente correlato positivamente con la Lunghezza, il diametro del cranio e le settimane di gestazione.

Modello di regressione lineare multipla

```
data$Fumatrici <- ifelse(data$Fumatrici == 1, "Y", "N")
modello_intero = lm(Peso ~ Gestazione + Lunghezza + Ospedale + Cranio + Sesso + N.gravidanze + T
ipo.parto + Fumatrici + Anni.madre, data=data)

summary(modello_intero)
```

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

Per quanto riguarda le variabili quantitative, Gestazione, Lunghezza, Cranio e N. Gravidanze mostrano un coefficiente positivo e un p-value fortemente significativo, mentre l'età della madre non sembra avere influenza sul peso.

Per quanto riguarda le variabili qualitative, il sesso Maschile e il parto di tipo naturale, mostrano un coefficiente positivo e un p-value fortemente significativo, mentre la madre fumatrice e l'ospedale non sembrano avere effetti significativi sul Peso.

Selezione del modello tramite stepwise

Tramite procedura stepwise, vengono selezionate solo le variabili che minimizzano il BIC del modello, eliminando quelle non significative.

```
modello_step <- stepAIC(modello_intero, direction = "both", k=log(nrow(data)))
```

```
## Start:  AIC=28139.32
## Peso ~ Gestazione + Lunghezza + Ospedale + Cranio + Sesso + N.gravidanze +
##     Tipo.parto + Fumatrici + Anni.madre
##
##                Df Sum of Sq       RSS   AIC
## - Anni.madre    1     46578 186809099 28132
## - Fumatrici     1     90019 186852540 28133
## - Ospedale      2    685979 187448501 28133
## - N.gravidanze  1    438452 187200974 28137
## - Tipo.parto    1    447929 187210450 28138
## <none>                      186762521 28139
## - Sesso         1   3611021 190373542 28179
## - Gestazione    1   5458403 192220925 28204
## - Cranio        1  45326172 232088693 28675
## - Lunghezza     1  87951062 274713583 29096
##
## Step:  AIC=28132.12
## Peso ~ Gestazione + Lunghezza + Ospedale + Cranio + Sesso + N.gravidanze +
##     Tipo.parto + Fumatrici
##
##                Df Sum of Sq       RSS   AIC
## - Fumatrici     1     90897 186899996 28126
## - Ospedale      2    692738 187501837 28126
## - Tipo.parto    1    448222 187257321 28130
## <none>                      186809099 28132
## - N.gravidanze  1    633756 187442855 28133
## + Anni.madre    1     46578 186762521 28139
## - Sesso         1   3618736 190427835 28172
## - Gestazione    1   5412879 192221978 28196
## - Cranio        1  45588236 232397335 28670
## - Lunghezza     1  87950050 274759149 29089
##
## Step:  AIC=28125.51
## Peso ~ Gestazione + Lunghezza + Ospedale + Cranio + Sesso + N.gravidanze +
##     Tipo.parto
##
##                Df Sum of Sq       RSS   AIC
## - Ospedale      2    701680 187601677 28119
## - Tipo.parto    1    440684 187340680 28124
## <none>                      186899996 28126
## - N.gravidanze  1    610840 187510837 28126
## + Fumatrici     1     90897 186809099 28132
## + Anni.madre    1     47456 186852540 28133
## - Sesso         1   3602797 190502794 28165
## - Gestazione    1   5346781 192246777 28188
## - Cranio        1  45632149 232532146 28664
## - Lunghezza     1  88355030 275255027 29086
##
## Step:  AIC=28119.23
## Peso ~ Gestazione + Lunghezza + Cranio + Sesso + N.gravidanze +
##     Tipo.parto
##
##                Df Sum of Sq       RSS   AIC
## - Tipo.parto    1    463870 188065546 28118
## <none>                      187601677 28119
## - N.gravidanze  1    651066 188252743 28120
## + Ospedale      2    701680 186899996 28126
```

```
## + Fumatrici     1     99840 187501837 28126
## + Anni.madre    1     54392 187547285 28126
## - Sesso         1   3649259 191250936 28160
## - Gestazione    1   5444109 193045786 28183
## - Cranio        1  45758101 233359778 28657
## - Lunghezza     1  88054432 275656108 29074
##
## Step:  AIC=28117.58
## Peso ~ Gestazione + Lunghezza + Cranio + Sesso + N.gravidanze
##
##                Df Sum of Sq       RSS   AIC
## <none>                      188065546 28118
## - N.gravidanze  1    623141 188688687 28118
## + Tipo.parto    1    463870 187601677 28119
## + Ospedale      2    724866 187340680 28124
## + Fumatrici     1     91892 187973654 28124
## + Anni.madre    1     54816 188010731 28125
## - Sesso         1   3655292 191720838 28158
## - Gestazione    1   5464853 193530399 28181
## - Cranio        1  46108583 234174130 28658
## - Lunghezza     1  87632762 275698308 29066
```

```
BIC(modello_step)
```

```
## [1] 35220.1
```

```
summary(modello_step)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     N.gravidanze, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.44  -180.81   -15.58   163.64  2639.72
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6681.1445   135.7229 -49.226  < 2e-16 ***
## Gestazione      32.3321     3.7980   8.513  < 2e-16 ***
## Lunghezza       10.2486     0.3006  34.090  < 2e-16 ***
## Cranio          10.5402     0.4262  24.728  < 2e-16 ***
## SessoM          77.9927    11.2021   6.962 4.26e-12 ***
## N.gravidanze    12.4750     4.3396   2.875  0.00408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.6 on 2494 degrees of freedom
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.7265
## F-statistic:  1328 on 5 and 2494 DF,  p-value: < 2.2e-16
```

Il modello finale ottenuto comprende solo le variabili quantitative Gestazione, Lunghezza, Cranio e N. di Gravidanze (tutte con coefficiente positivo) e la variabile qualitativa Sesso. Un valore di R^2 di 0.72 indica che le variabili sono in grado di spiegare il 72% della variabilità del campione.

#Verifica di multicollinearità

Il Variance Inflation Factor (VIF) è una misura utilizzata per rilevare la multicollinearità nelle regressioni multiple. La multicollinearità si verifica quando due o più variabili indipendenti nel modello sono altamente correlate.

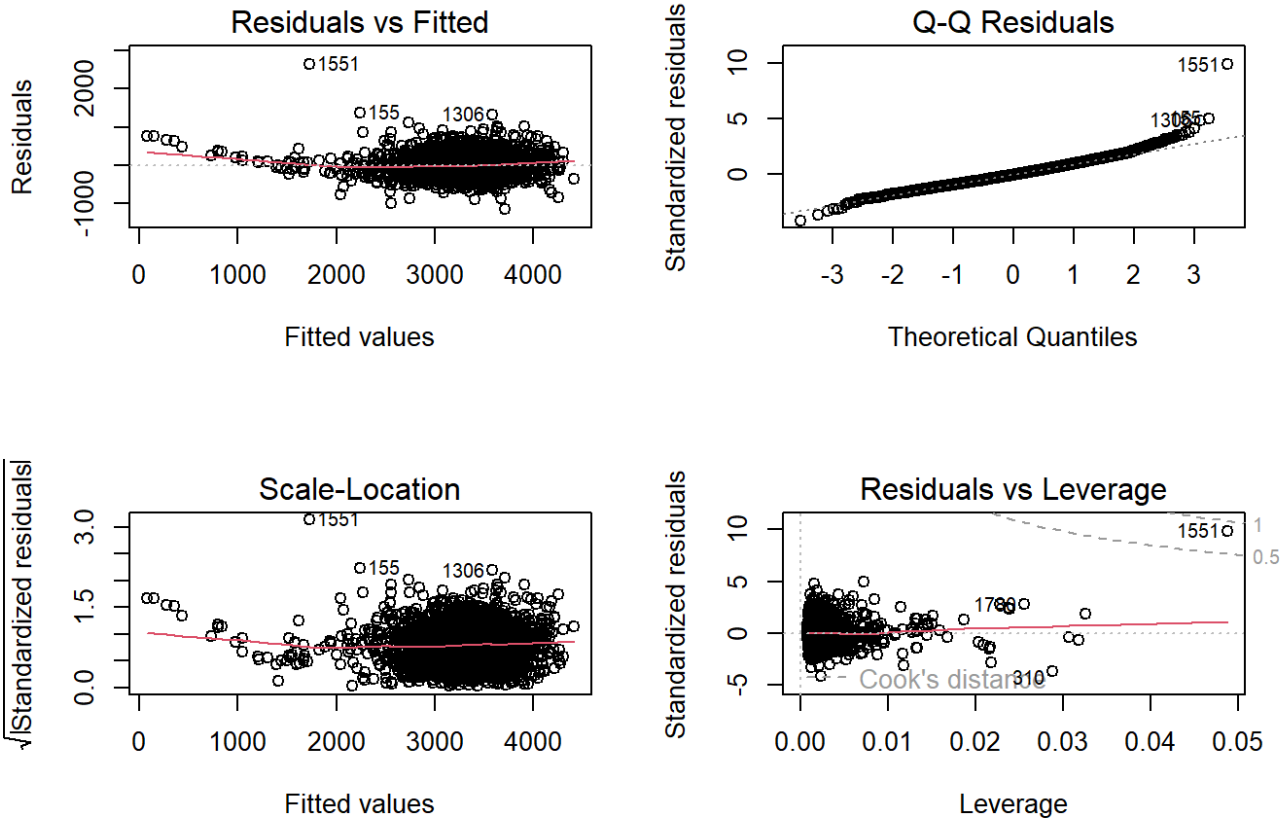
```
vif(modello_step)
```

```
##   Gestazione    Lunghezza       Cranio        Sesso N.gravidanze
##     1.669189     2.074689     1.624465     1.040054     1.023475
```

Tutte le variabili mostrano un VIF < 5, pertanto non è presente multicollinearità.

Diagnostica del modello: grafici dei residui

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

plot(modello_step)
```



1. Grafico dei residui vs. valori predetti: I residui sono distribuiti casualmente intorno allo zero.
2. Grafico Q-Q (Quantile-Quantile): I punti seguono approssimativamente la linea diagonale, i residui sono normalmente distribuiti.
3. Il dato 1551 appare come unico valore influente del modello.

Test sui residui

```
shapiro.test(residuals(modello_step))
```

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

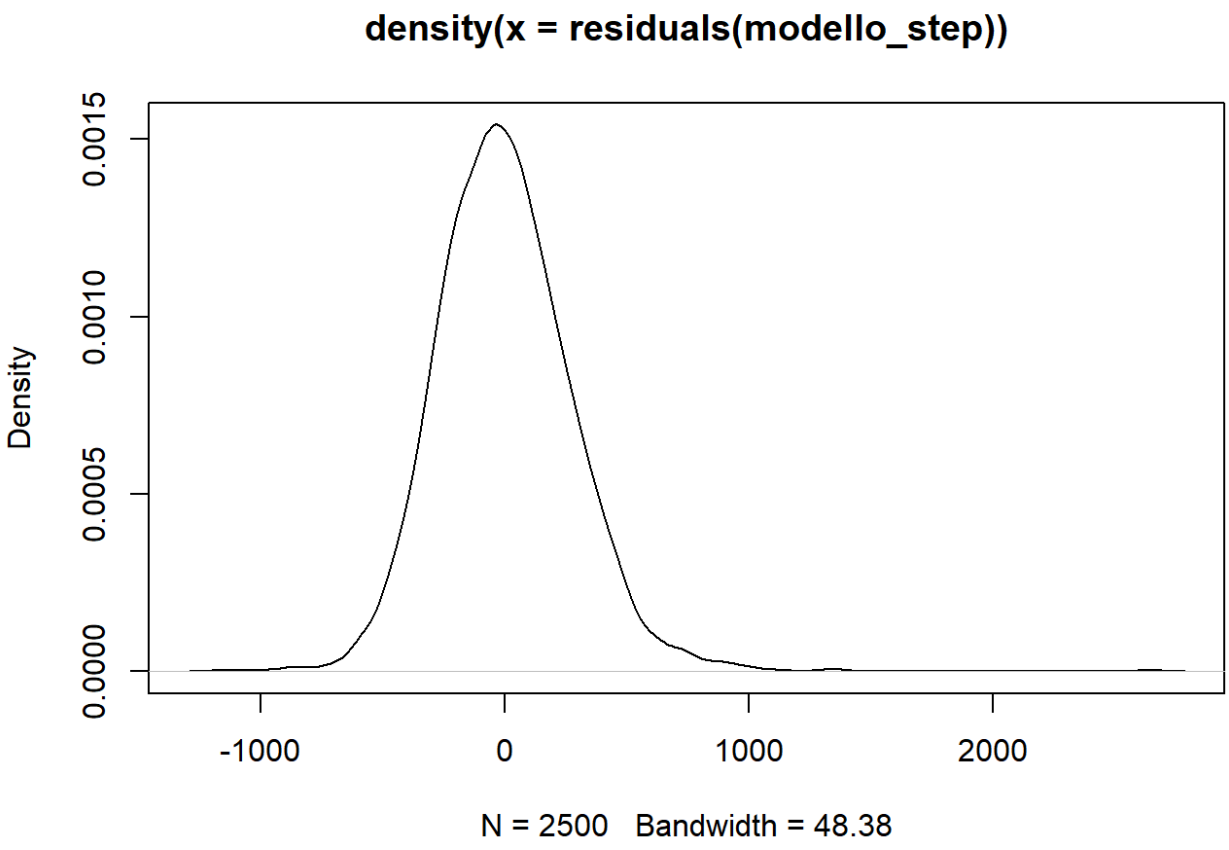
```
bptest(modello_step)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modello_step
## BP = 90.253, df = 5, p-value < 2.2e-16
```

```
dwtest(modello_step)
```

```
##
##  Durbin-Watson test
##
## data:  modello_step
## DW = 1.9535, p-value = 0.1224
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(density(residuals(modello_step)))
```

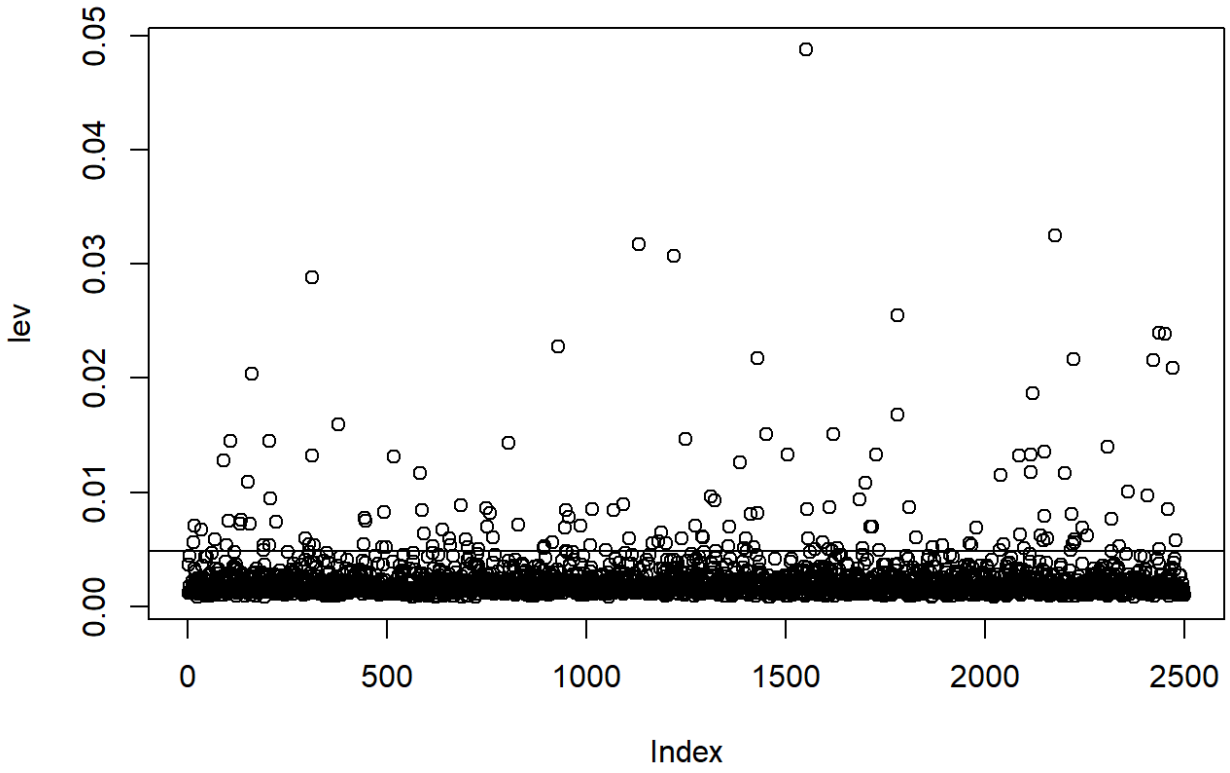


1. Shapiro-Wilk Test ($p < 0.05$): Rifiutiamo l'ipotesi nulla che i residui siano normalmente distribuiti.

2. Breusch-Pagan Test (p<0.05): Rifiutiamo l'ipotesi nulla di omoscedasticità. Ciò suggerisce che la varianza degli errori non è costante
3. Durbin-Watson Test (p> 0.05): Non possiamo rifiutare l'ipotesi nulla che non ci sia autocorrelazione nei residui. Questo suggerisce che i residui sono indipendenti l'uno dall'altro.

Verifica dei valori influenti con la statistica Cook's distance

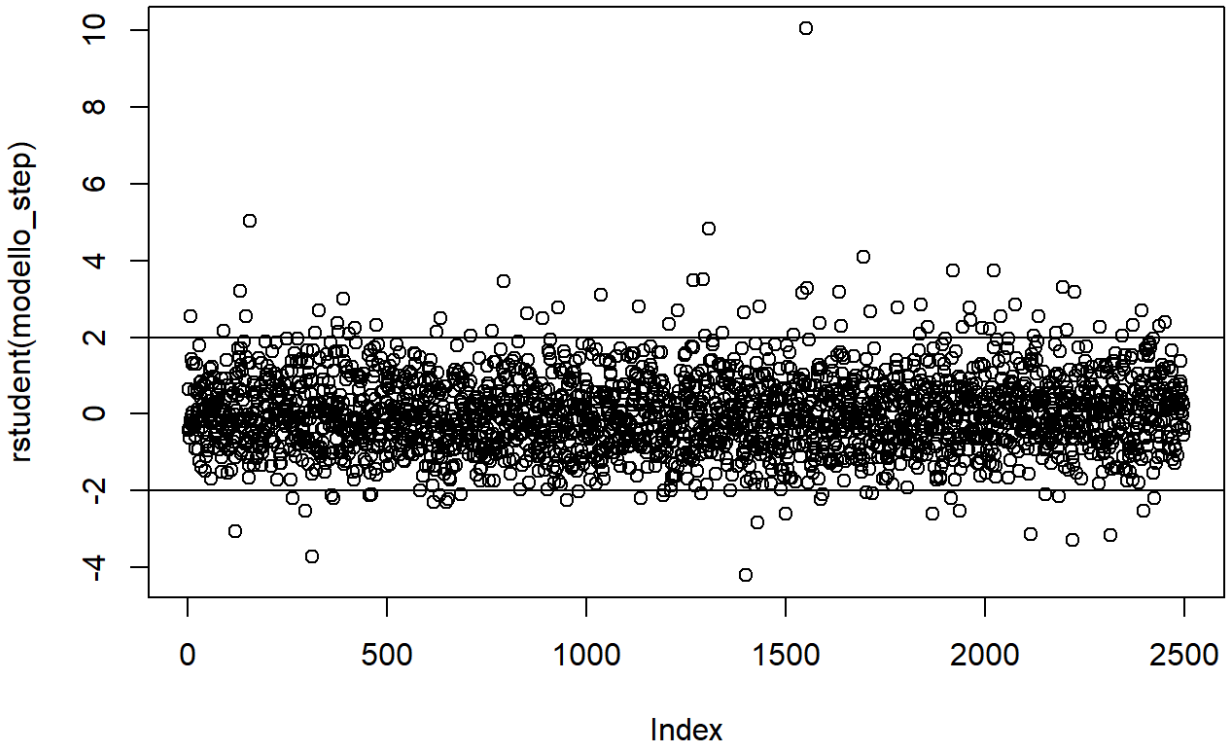
```
lev = hatvalues(modello_step)
p=sum(lev)
soglia = 2*p/nrow(data)
plot(lev)
abline(h=soglia)
```



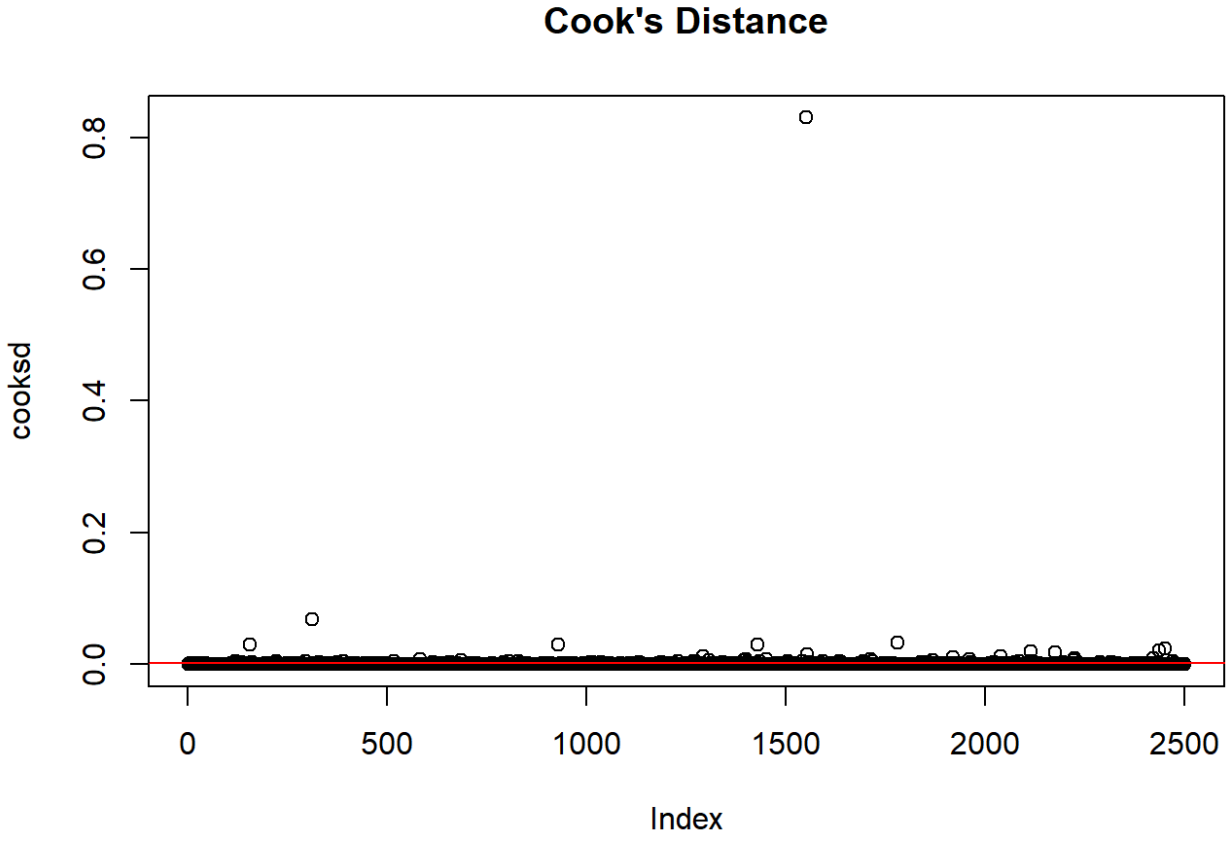

```
outlierTest(modello_step)
```

```
##       rstudent unadjusted p-value Bonferroni p
## 1551 10.051908         2.4906e-23   6.2265e-20
## 155   5.027798         5.3138e-07   1.3285e-03
## 1306  4.827238         1.4681e-06   3.6702e-03
```

```
plot(rstudent(modello_step))
abline(h=c(-2,2))
```

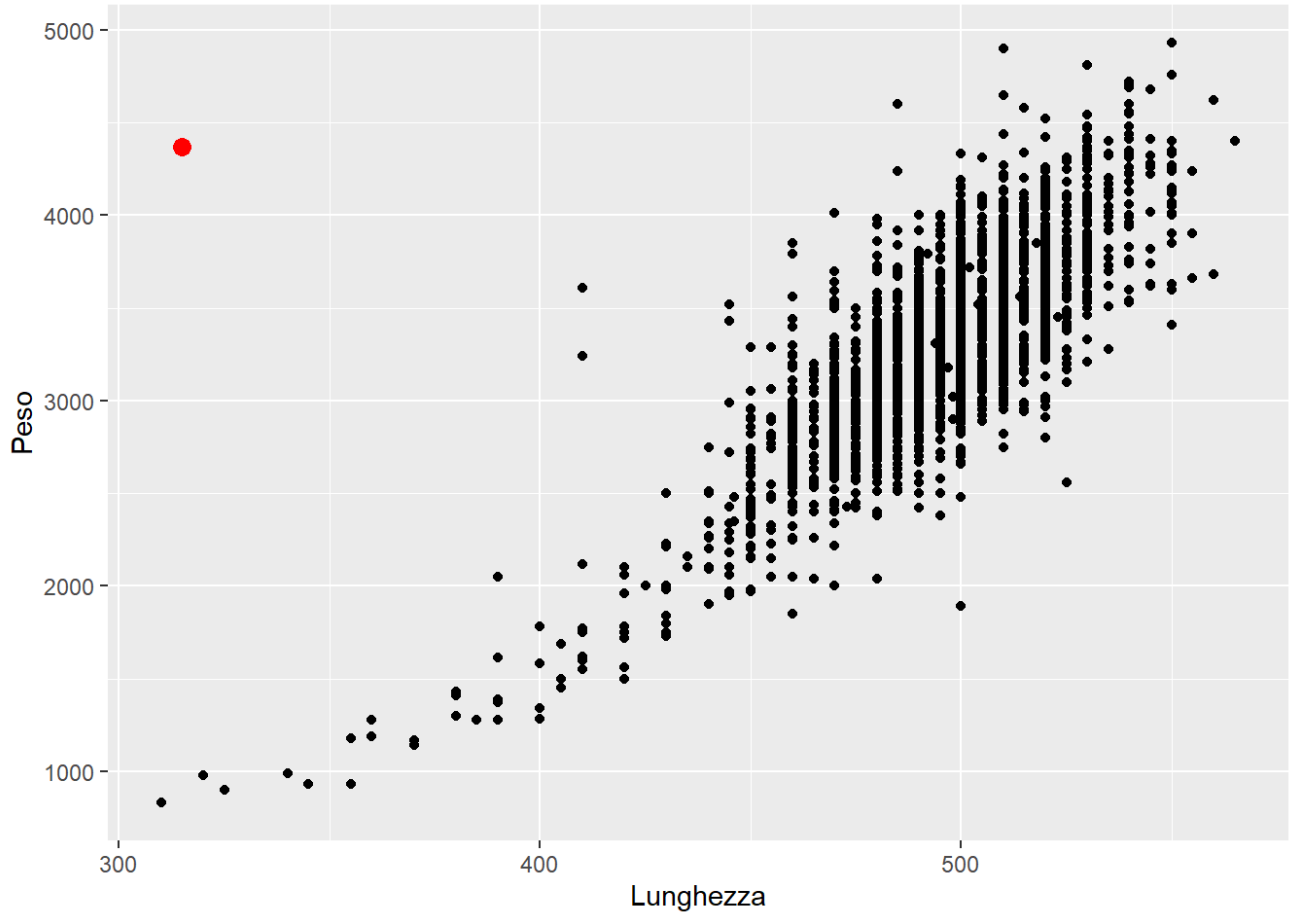



```
cooksd <- cooks.distance(modello_step)
plot(cooksd, main="Cook's Distance")
abline(h = 4/(nrow(data)-length(coef(modello_step))-1), col="red")
```



Il dato 1551 si conferma come unico valore influente del modello, con una distanza di Cook pari a circa 0.8.

```
ggplot(data=data)+
  geom_point(aes(x=Lunghezza, y=Peso))+
  geom_point(aes(x=Lunghezza[1551], y=Peso[1551]), color='red', size=3)
```



Il punto sembra essere l'unico a non seguire il trend lineare tra lunghezza e peso. Proviamo ad eliminarlo e ricostruire il modello.

Modello senza outlier

```
df_senza_outlier <- data[-1551, ]

modello_step_senza_outlier <- lm(Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
    N.gravidanze, data = df_senza_outlier)

summary(modello_step_senza_outlier)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     N.gravidanze, data = df_senza_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1165.74  -179.59   -12.74   162.89  1410.88
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6683.4142   133.0802 -50.221  < 2e-16 ***
## Gestazione      29.5891     3.7340   7.924 3.43e-15 ***
## Lunghezza       10.8927     0.3017  36.109  < 2e-16 ***
## Cranio           9.9187     0.4225  23.476  < 2e-16 ***
## SessoM          78.1348    10.9840   7.114 1.47e-12 ***
## N.gravidanze    13.1652     4.2557   3.094    0.002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.3 on 2493 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7367
## F-statistic:  1399 on 5 and 2493 DF,  p-value: < 2.2e-16
```

```
BIC(modello_step_senza_outlier)
```

```
## [1] 35107.74
```

Il valore di R^2 è aumentato di circa 1%, mentre il BIC si è abbassato di 113. Entrambi i risultati indicano un miglioramento del modello. Ripetiamo i test sui residui.

Test sui residui modello senza outlier

```
shapiro.test(residuals(modello_step_senza_outlier))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modello_step_senza_outlier)
## W = 0.98886, p-value = 4.764e-13
```

```
bptest(modello_step_senza_outlier)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modello_step_senza_outlier
## BP = 11.393, df = 5, p-value = 0.04411
```

```
dwtest(modello_step_senza_outlier)
```

```
##
##  Durbin-Watson test
##
## data:  modello_step_senza_outlier
## DW = 1.954, p-value = 0.1251
## alternative hypothesis: true autocorrelation is greater than 0
```

Il Breusch-Pagan test adesso risulta al limite della significatività, suggerendo che l'ipotesi nulla di omoscedasticità potrebbe essere rispettata. Tuttavia, il test di Shapiro-Wilk mostra ancora un'elevata significatività, pertanto non possiamo accettare l'ipotesi di normalità.

Nel complesso, il modello sembra essere affidabile, con variabili significative e buone proprietà diagnostiche, ad eccezione della normalità dei residui.

Creiamo un nuovo dataframe con le caratteristiche specifiche

```
nuovo_neonato <- data.frame(N.gravidanze = 3, Gestazione = 39, Sesso = "F")
```

Creiamo un nuovo modello senza includere lunghezza e diametro del cranio

```
modello_ridotto <- lm(Peso ~ N.gravidanze + Gestazione + Sesso,
                      data=df_senza_outlier)

summary(modello_ridotto)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Sesso, data = df_senza_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1493.24  -272.78   -14.31   266.99  1893.87
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3142.295    175.092 -17.947  < 2e-16 ***
## N.gravidanze    23.396      6.497   3.601 0.000323 ***
## Gestazione     162.138      4.491  36.105  < 2e-16 ***
## SessoM         166.144     16.701   9.948  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.4 on 2495 degrees of freedom
## Multiple R-squared:  0.3799, Adjusted R-squared:  0.3792
## F-statistic: 509.5 on 3 and 2495 DF,  p-value: < 2.2e-16
```

```
BIC(modello_ridotto)
```

```
## [1] 37237.33
```

Le variabili sono rimaste significative. Tuttavia, il p-value è sceso a 0.38, mentre il BIC si è alzato a 37237, pertando il modello è peggiorato.

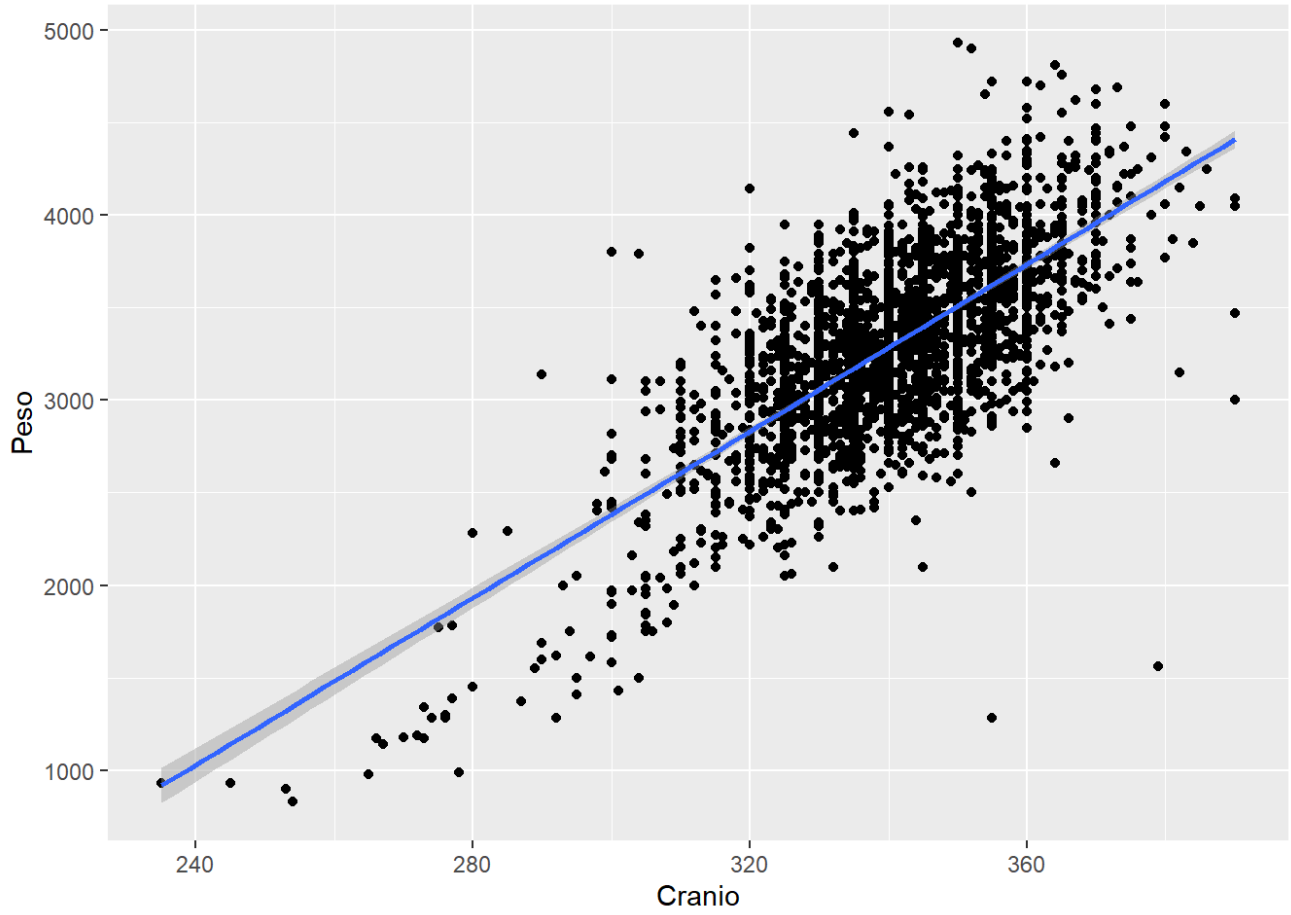
Predizione con il modello ridotto

```
predizione <- predict(modello_ridotto, newdata = nuovo_neonato)
predizione
```

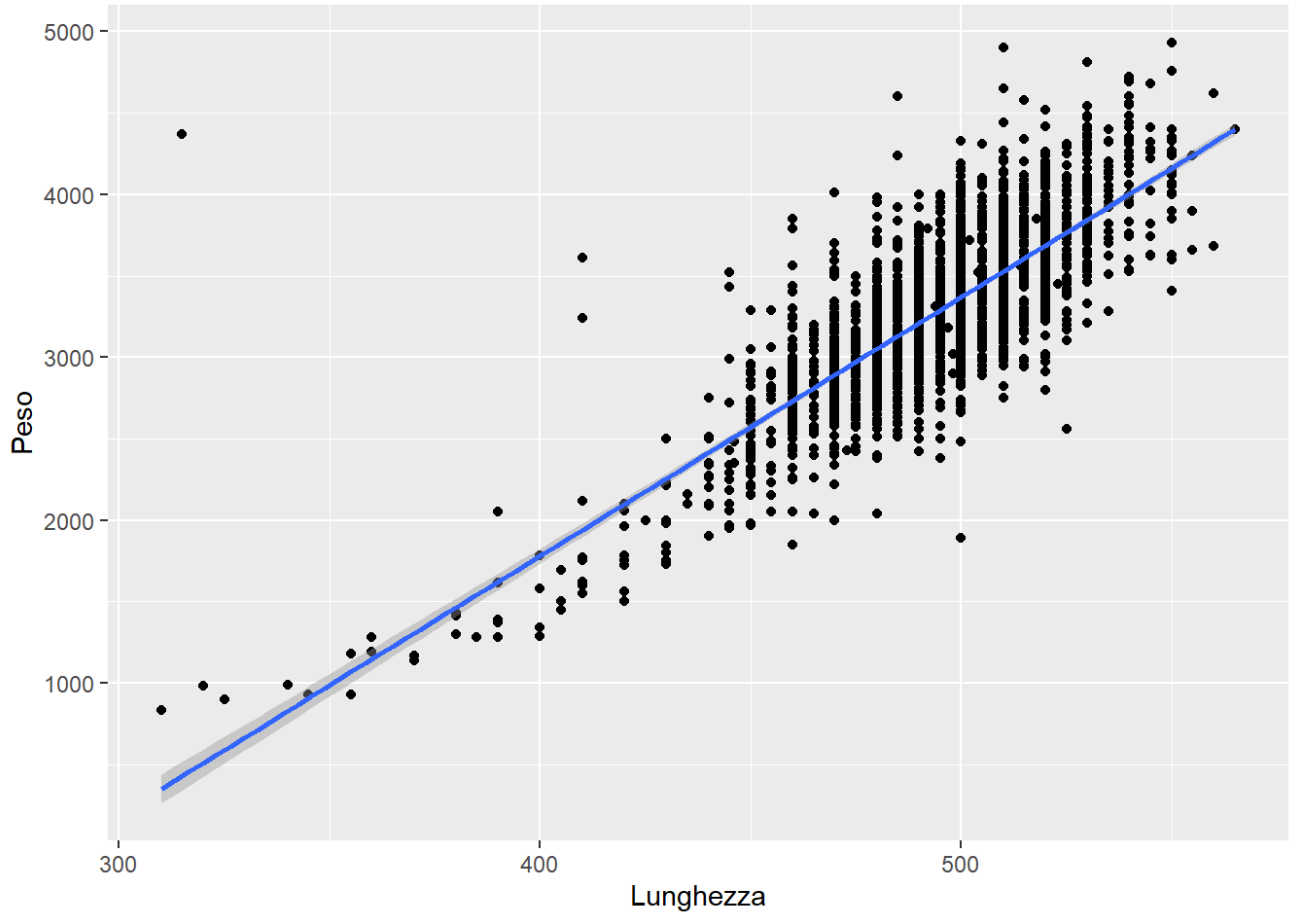
```
##        1
## 3251.287
```

#Visualizziamo il modello graficamente

```
ggplot(data=data)+
  geom_point(aes(x=Cranio, y=Peso)) +
  geom_smooth(aes(x=Cranio, y=Peso), method='lm')
```




```
ggplot(data=data)+
  geom_point(aes(x=Lunghezza, y=Peso)) +
  geom_smooth(aes(x=Lunghezza, y=Peso), method='lm')
```




```
ggplot(data=data)+
  geom_point(aes(x=Gestazione, y=Peso)) +
  geom_smooth(aes(x=Gestazione, y=Peso), method='lm')
```