PROGETTO

DI

STATISTICA INFERENZIALE

Santone Carmine

c.santone@outlook.it

1. OBIETTIVI DELLO STUDIO

Con il presente studio s'intende prendere in esame un dataset contenente variabili riguardanti i parametri di neonati e rispettive madri allo scopo di fare previsioni sul peso del nascituro, un indicatore utilizzato per valutare lo stato di salute di quest'ultimo. Inoltre, si vuole saggiare l'ipotesi che il consumo di sigarette possa influenzare la salute del bambino.

2. EXPLORATORY DATA ANALYSIS

2.1. STRUTTURA DATASET

Il primo passo da compiere è indagare la composizione del dataset. Innanzitutto bisogna considerare se sono presenti valori mancanti:

```
> df <- read.csv("neonati.csv")
> sum(is.na.data.frame(df))
[1] 0
```

Non ci sono NA.

Si visualizza la struttura del dataset:

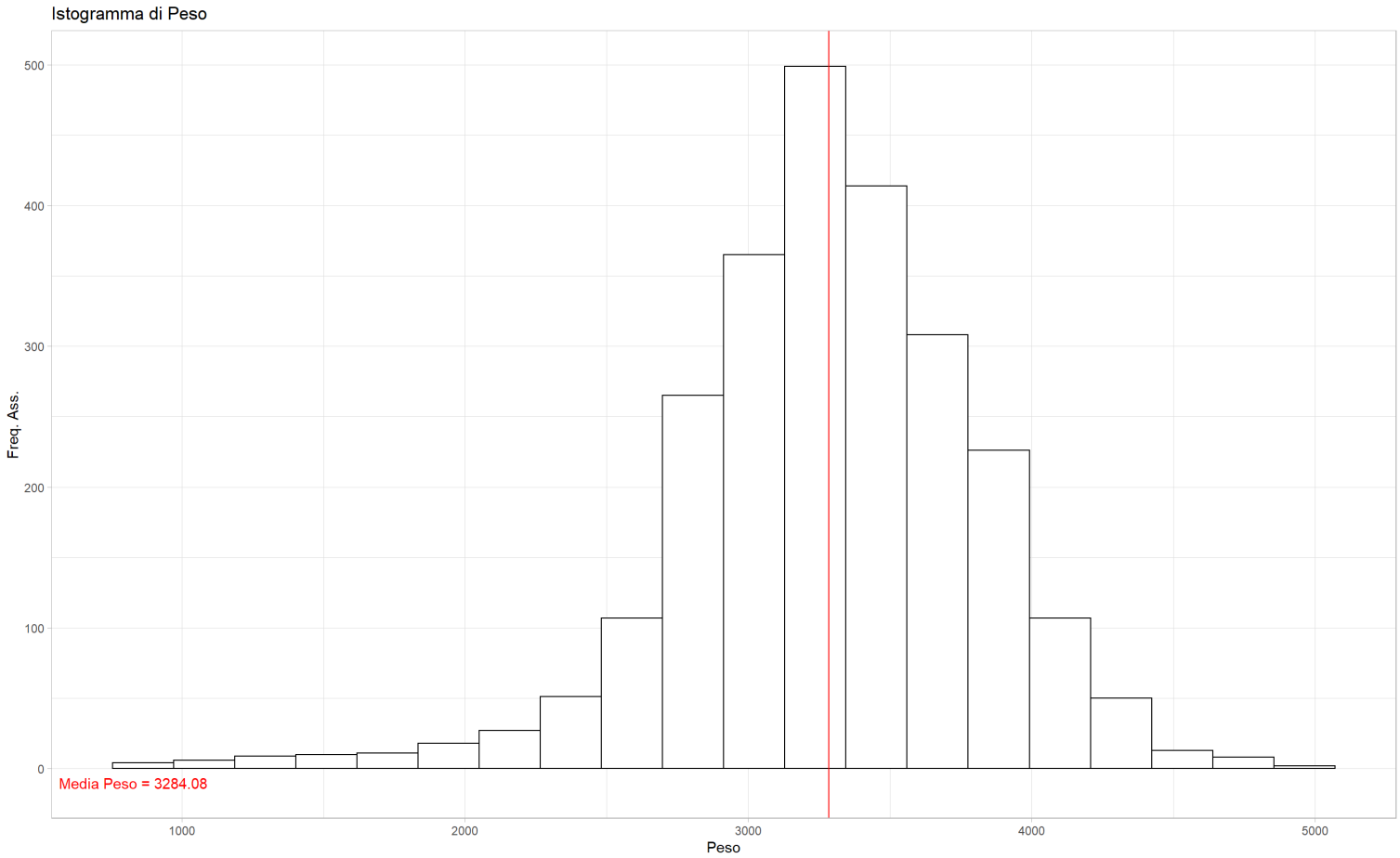
```
> str(df)


'data.frame': 2500 obs. of  10 variables:
 $ Anni.madre  : int  26 21 34 28 20 32 26 25 22 23 ...
 $ N.gravidanze: int  0 2 3 1 0 0 1 0 1 0 ...
 $ Fumatrici   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Gestazione  : int  42 39 38 41 38 40 39 40 40 41 ...
 $ Peso        : int  3380 3150 3640 3690 3700 3200 3100 3580 3670 3700 ...
 $ Lunghezza   : int  490 490 500 515 480 495 480 510 500 510 ...
 $ Cranio      : int  325 345 375 365 335 340 345 349 335 362 ...
 $ Tipo.parto  : chr  "Nat" "Nat" "Nat" "Nat" ...
 $ Ospedale    : chr  "osp3" "osp1" "osp2" "osp2" ...
 $ Sesso       : chr  "M" "F" "M" "M" ...
```

Il dataset è composto da 2500 osservazioni e 10 variabili. Ci sono 6 variabili quantitative (*Anni.madre*, *N.gravidanze*, *Gestazione*, *Peso*, *Lunghezza* e *Cranio*) e 4 variabili qualitative di cui 1 già codificata (*Fumatrici*) e altre sotto forma di stringa (*Tipo.parto*, *Ospedale*, *Sesso*).

2.2. ANALISI DESCRITTIVA

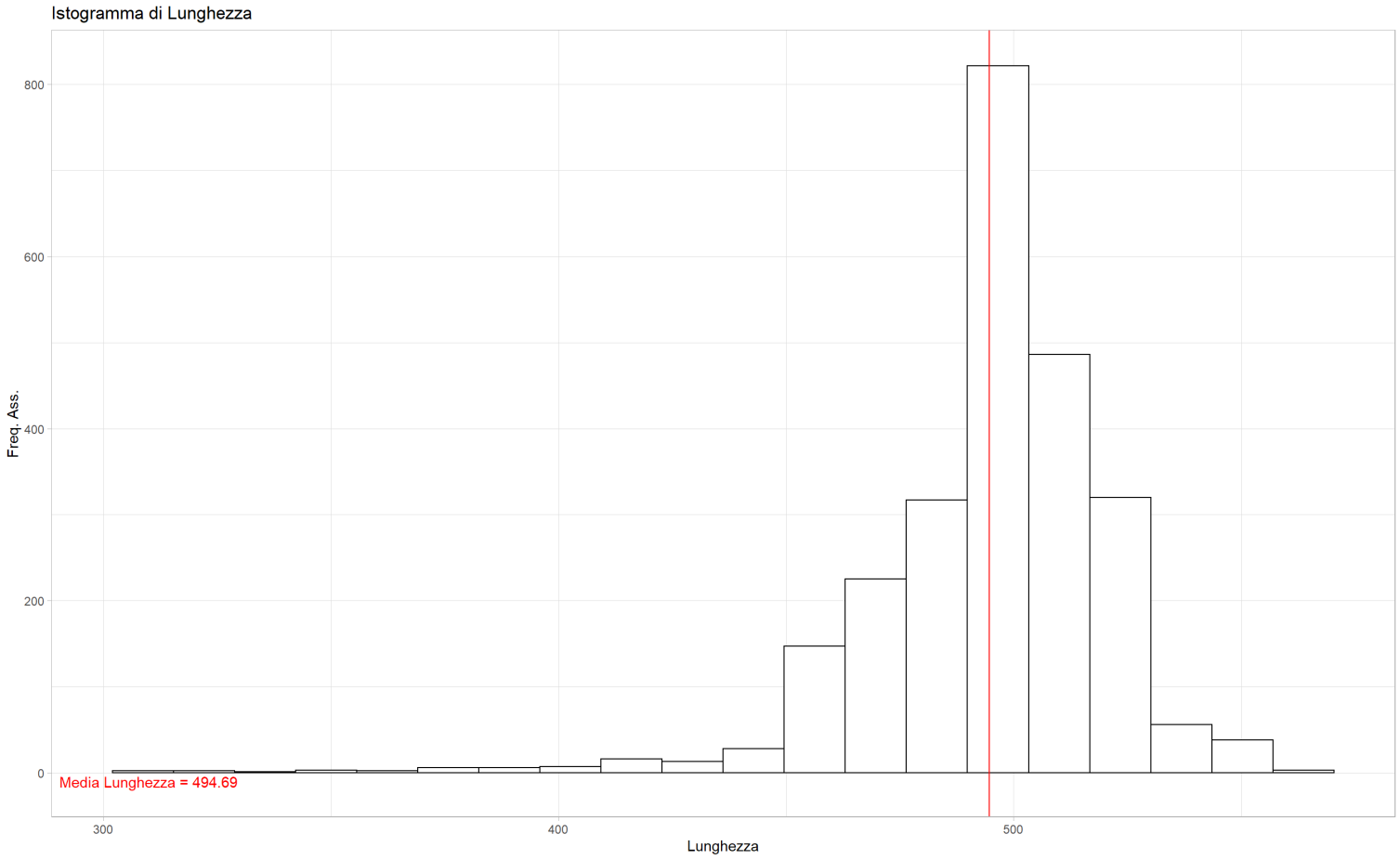
Si procede alla visualizzazione e ad una breve analisi descrittiva di ciascuna variabile presente nel dataset:

1) *Peso*



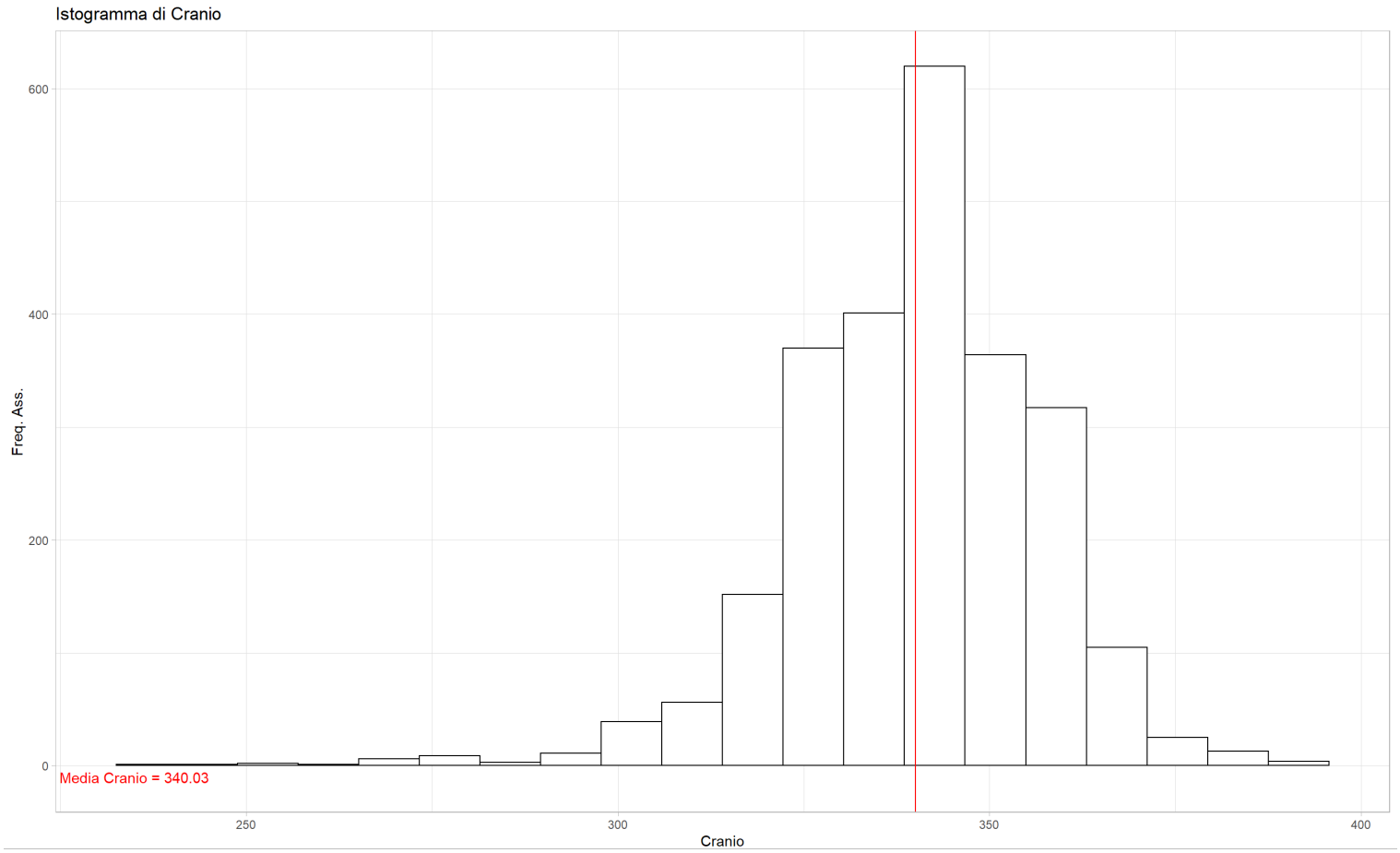
La variabile indica il peso del neonato alla nascita registrato in grammi. Si nota che gran parte delle osservazioni presenta un peso compreso tra i 3000 gr e i 4000 gr, con una media del campione pari a circa 3284 gr.

2) *Lunghezza*



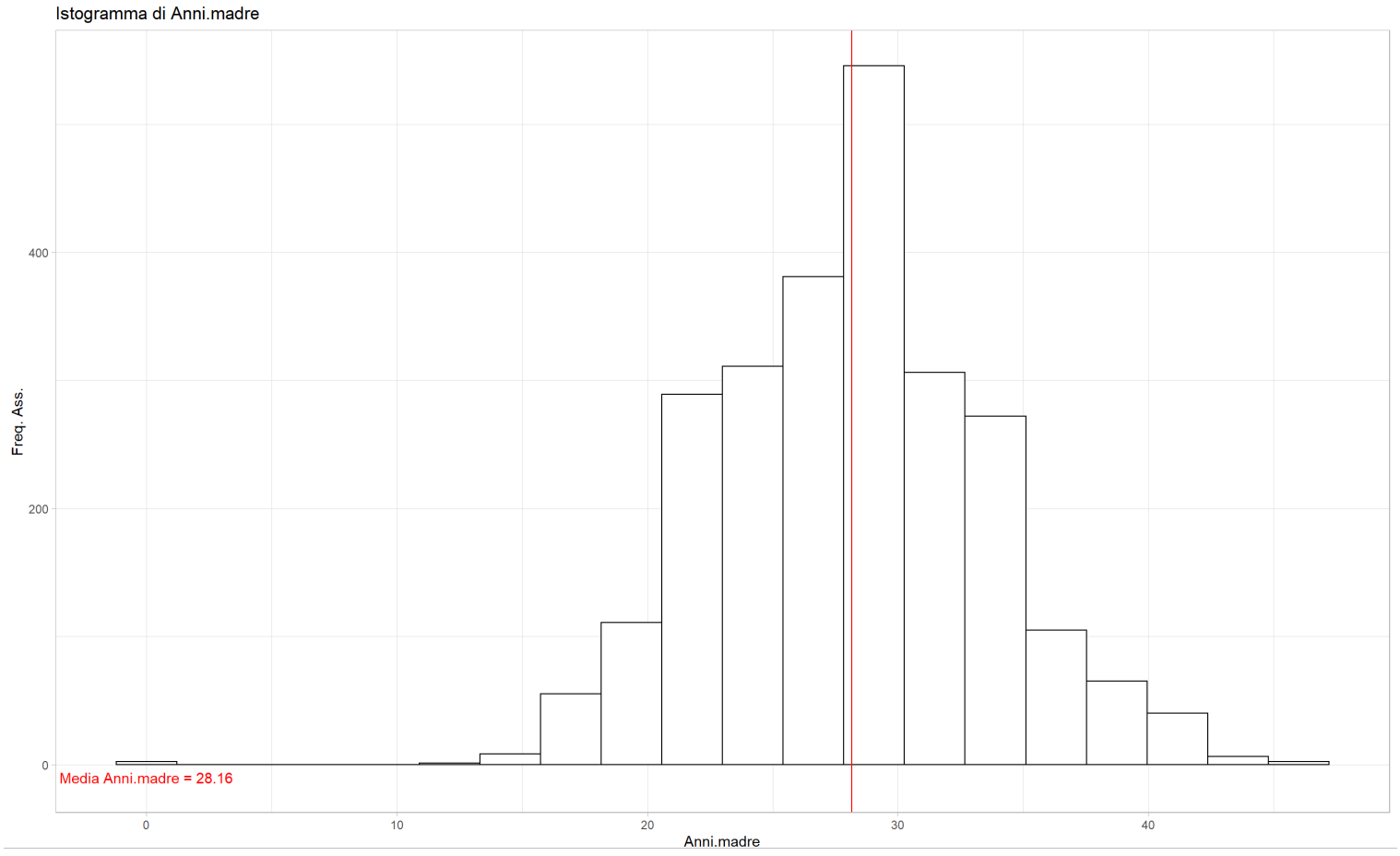
La variabile indica la lunghezza del neonato in mm. Dal grafico si vede che la maggior parte dei nascituri ha una lunghezza intorno ai 500 mm con una media di 494.69 mm.

3) *Cranio*



La variabile sta ad indicare il diametro del cranio del neonato, misurato in mm. Si può notare che la maggior parte dei bambini nasce con un cranio il cui diametro va dai 320 ai 360 mm con una media di circa 340 mm.

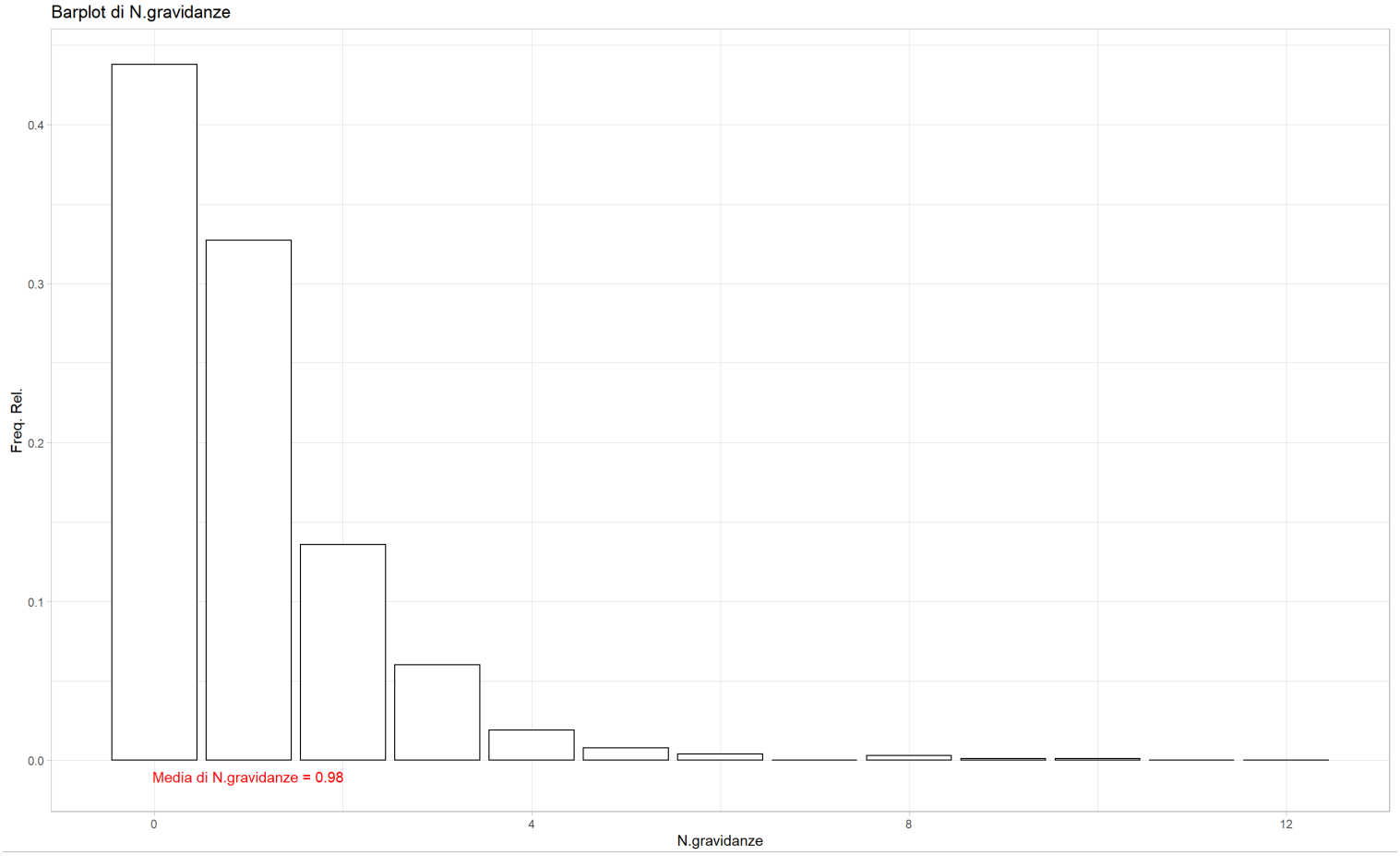
4) *Anni.madre*



Gran parte delle donne presenti all'interno del campione ha un età compresa tra i 24 e i 32 anni con una media di 28.16 anni. Visualizzando il grafico si nota la presenza di osservazioni con un'età inferiore ai 10 anni, il che porta a pensare ad errori di misurazione. In seguito si procede alla rimozione di tali osservazioni dal dataset.

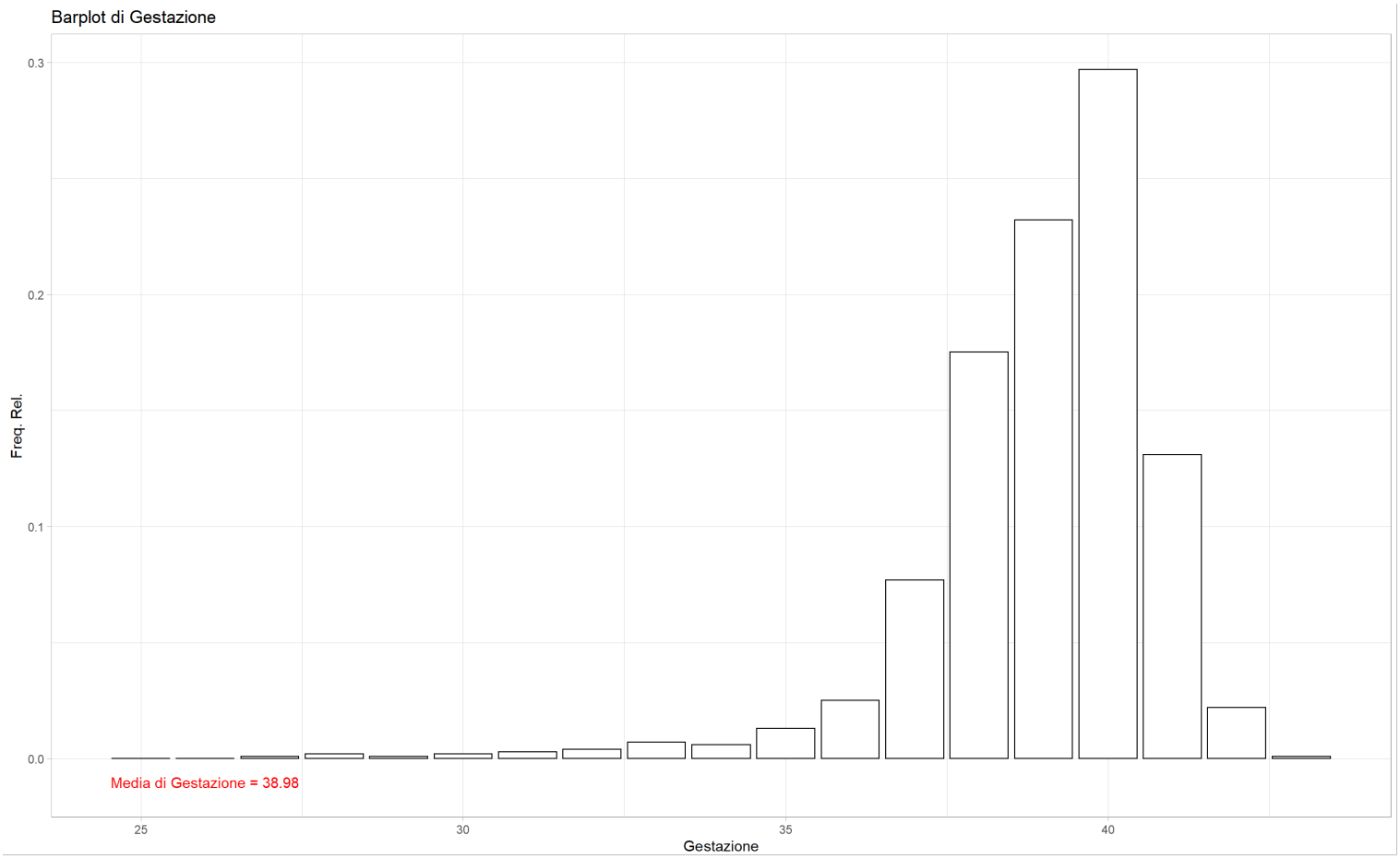
```
> del_oss <- which(df$Anni.madre<10)

> del_oss

[1] 1152 1380

> df<-df[-del_oss,]
```

5) *N.gravidanze*



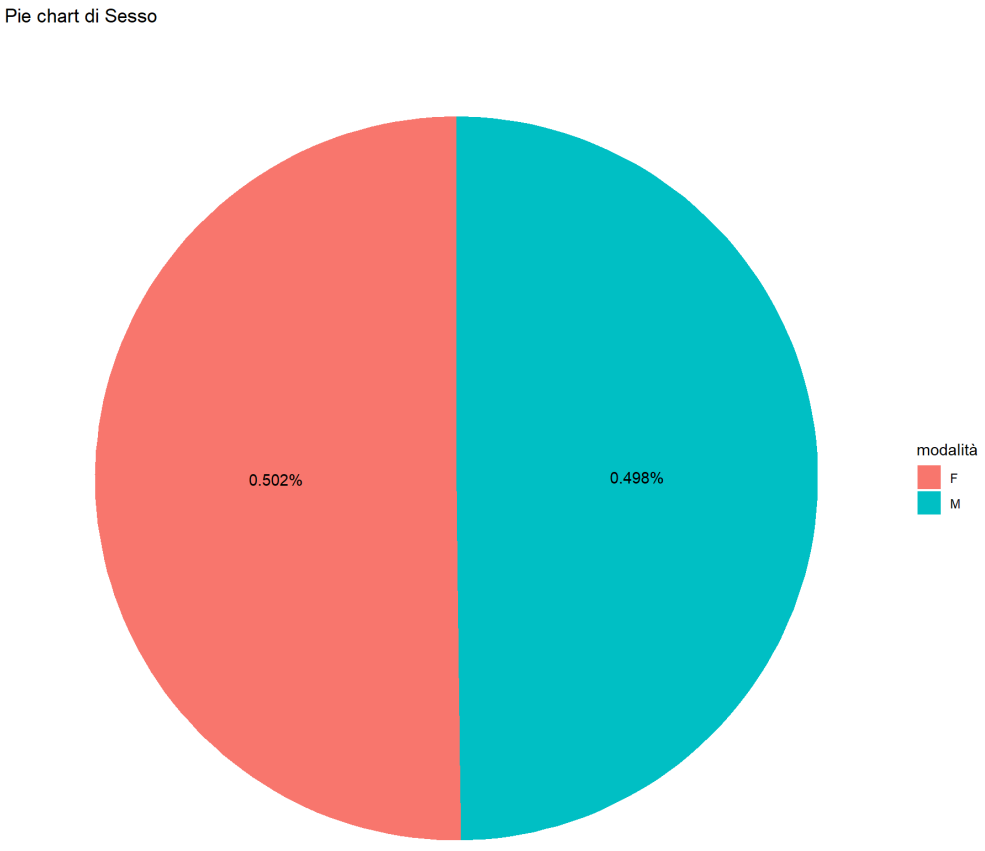
La variabile indica il numero di gravidanze sostenute precedentemente alla raccolta dei dati. Circa il 40% del campione è composto da donne alla prima gravidanza.

6) *Gestazione*



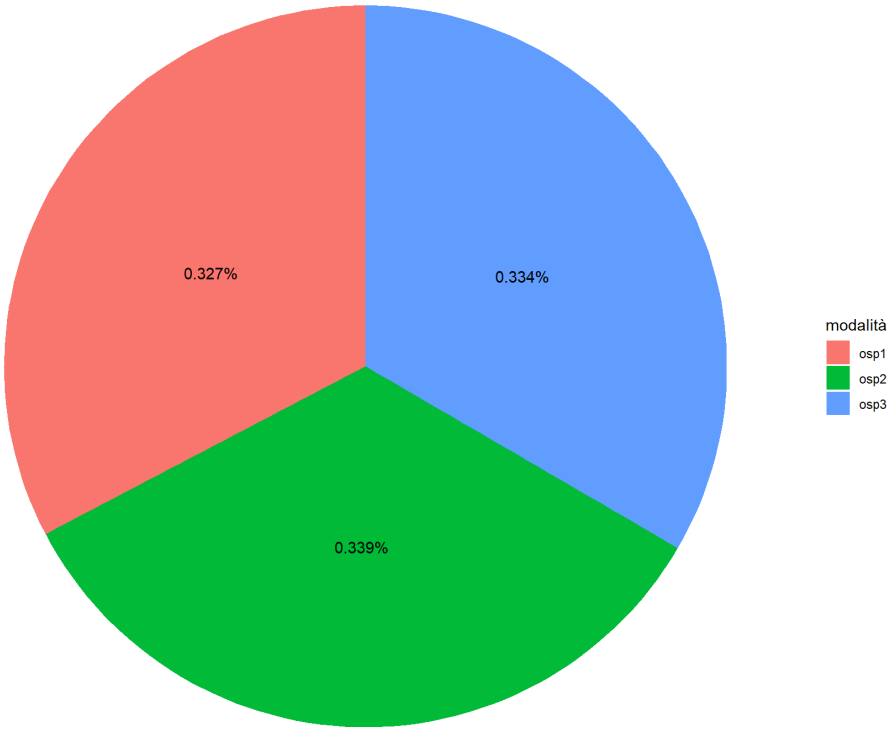
La variabile sta ad indicare il numero di settimane di gestazione per la gravidanza in atto al momento della raccolta dei dati. Per il 30% del campione è stata registrata una durata della gestazione pari a 40 settimane, ovvero circa 9.2 mesi. Inoltre si riscontra una media pari 38.98 settimane di gestazione, in linea con le tempistiche comunicate dal Ministero della Salute.

7) *Sesso*



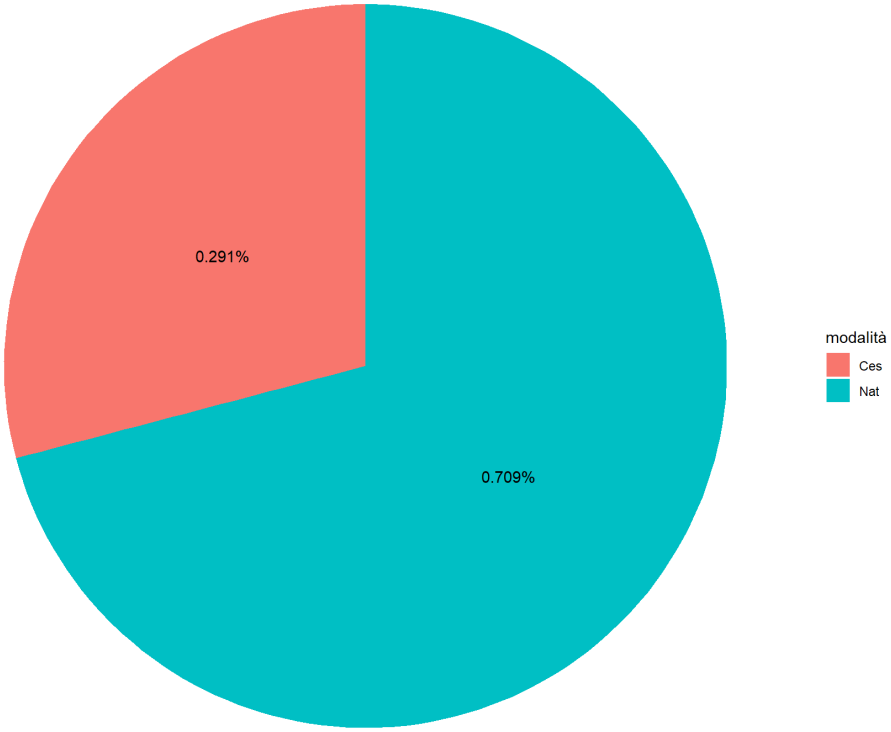
8) *Ospedale*

Pie chart di Ospedale

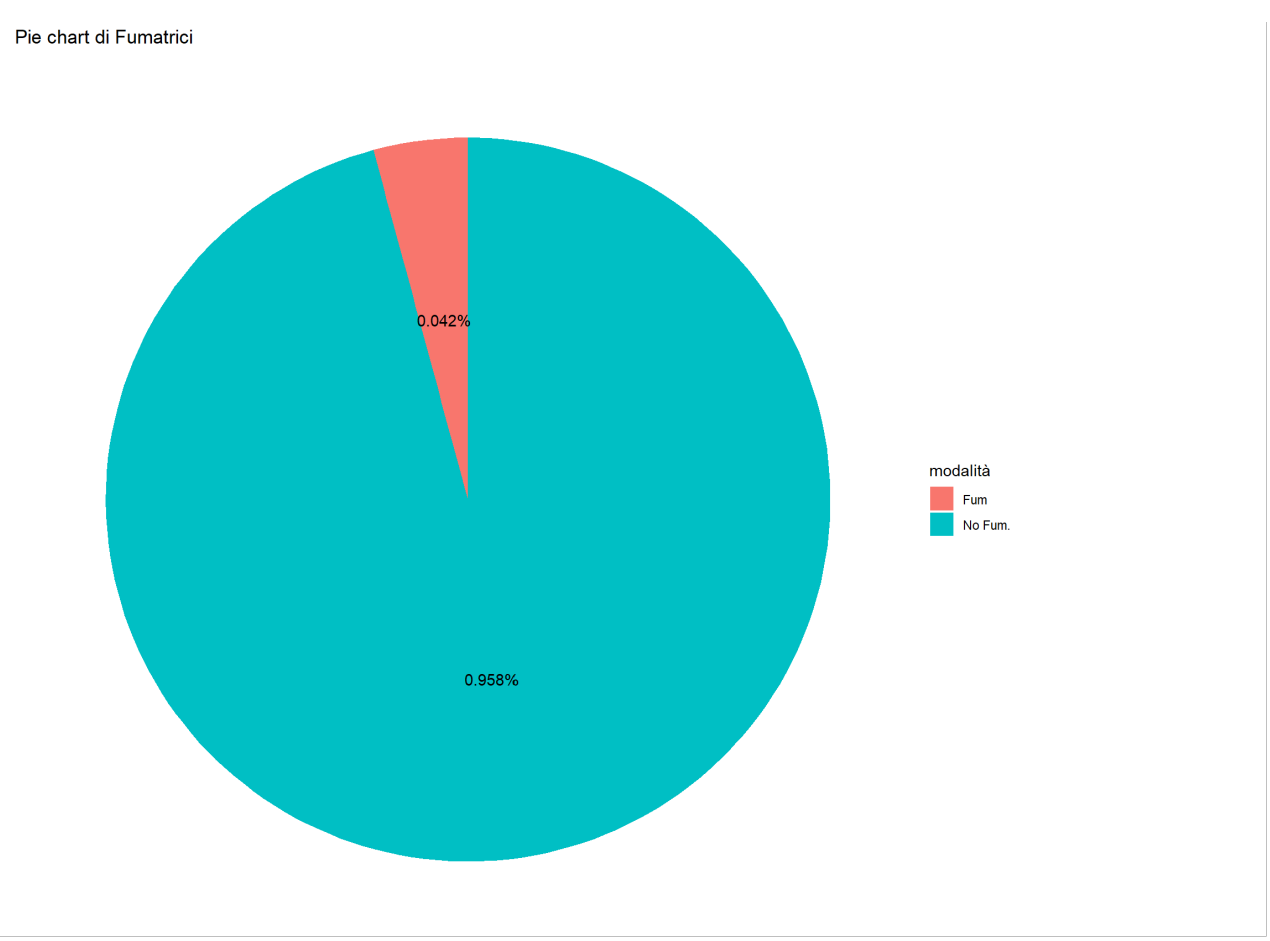


9) *Tipo parto*

Pie chart di Tipo.parto



10) *Fumatrici*



2.3. RELAZIONI TRA VARIABILI

Ecco le ipotesi che si andranno ad indagare di seguito:

1) Ci sono ospedali in cui si pratica maggiormente un tipo di parto rispetto agli altri presenti nel dataset?

2) Il tipo di parto è legato alle dimensioni del neonato, alle condizioni di salute della madre e alla gestazione. (Fonte: SISMeR)

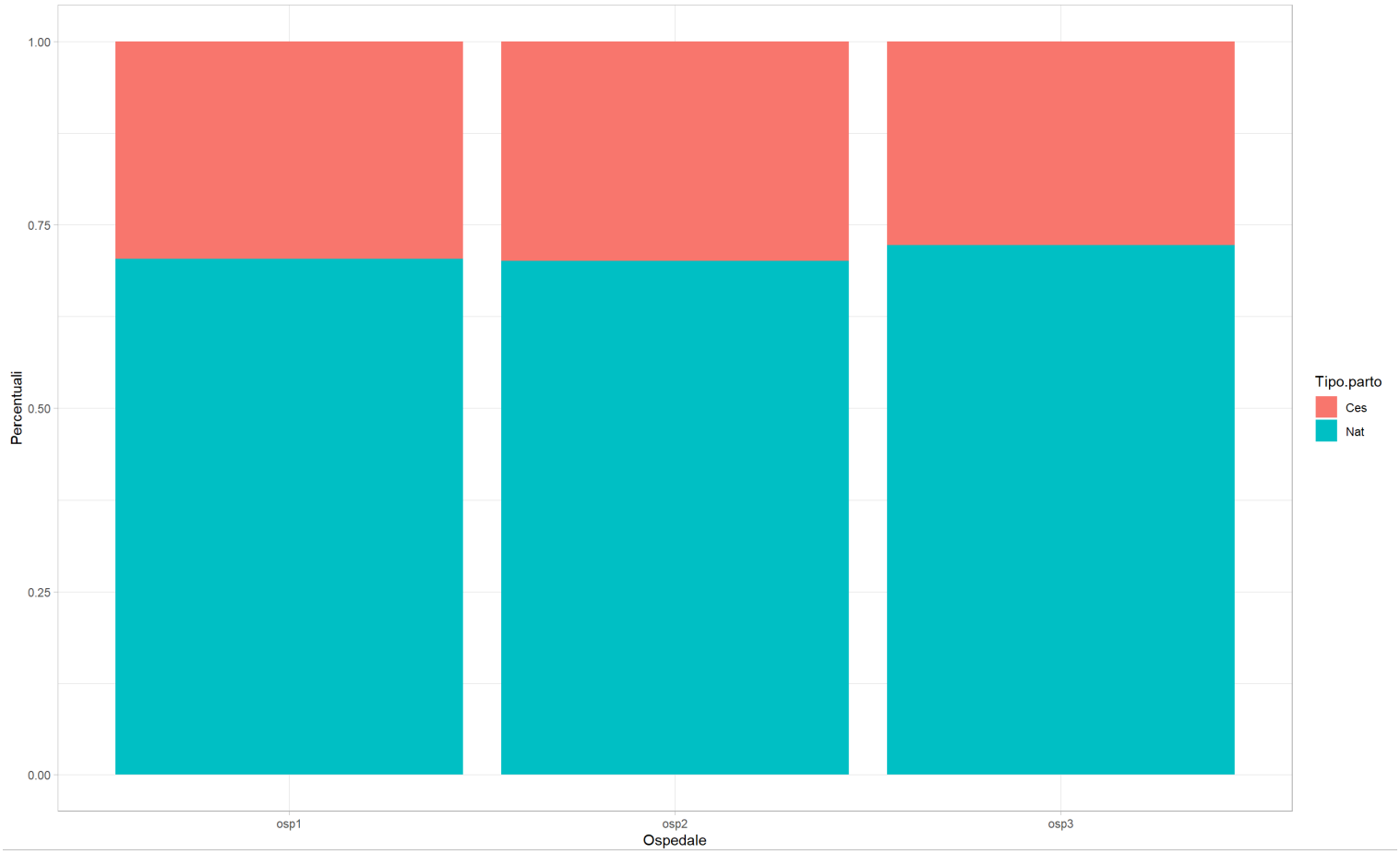
3) Il periodo di gestazione è legato all'età della madre ed alle sue condizioni di salute. (Fonte: PARWELB)

4) Il peso del neonato ha una relazione con il sesso, le condizioni di salute della madre e dal numero di gravidanze precedenti. (Fonte: myPersonalTrainer)

5) Lunghezza e diametro del cranio variano tra maschi e femmine.

Si procede all'analisi delle relazioni:

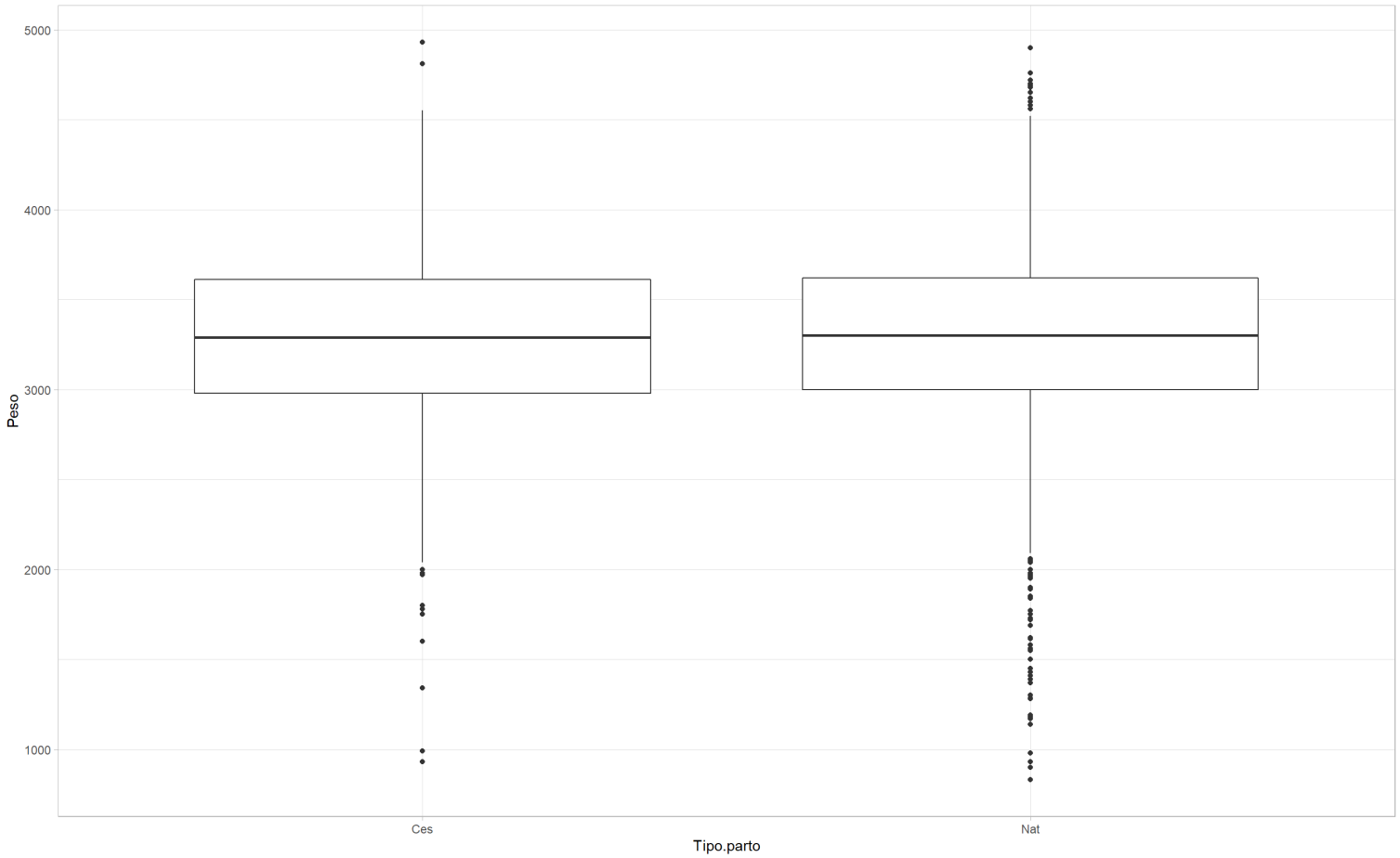
1) *Ospedale* & *Tipo parto*




```
> cont_tabOT <-table(df$Tipo.parto, df$Ospedale)

> chisq.test(cont_tabOT)

Pearson's Chi-squared test

data:  cont_tabOT
X-squared = 1.083, df = 2, p-value = 0.5819
```

Sulla base delle evidenze empiriche si esclude una relazione di dipendenza tra le due variabili. Si potrebbe concludere che non ci sono ospedali "specializzati" in un certo tipo di parto.

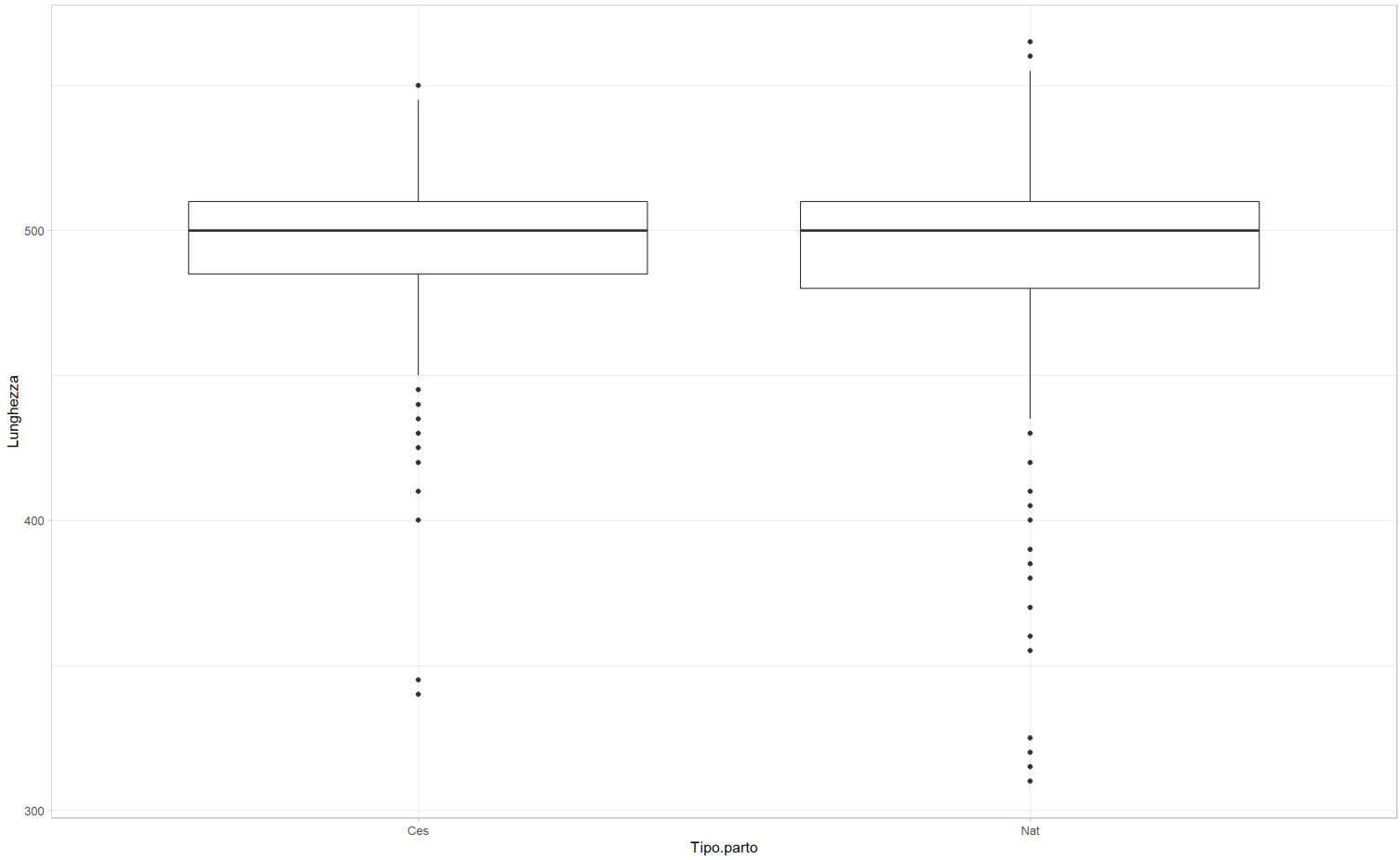
2.1) *Tipo parto & Peso*




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

Welch Two Sample t-test

data:  Peso by Tipo.parto
t = -0.13626, df = 1494.4, p-value = 0.8916
alternative hypothesis: true difference in means between group Ces and group Nat is not equal to 0
95 percent confidence interval:
-46.44246  40.40931
sample estimates:
mean in group Ces mean in group Nat
         3282.047          3285.063
```

Sulla base delle evidenze empiriche si esclude l'ipotesi che il peso abbia una relazione con la scelta del tipo di parto.

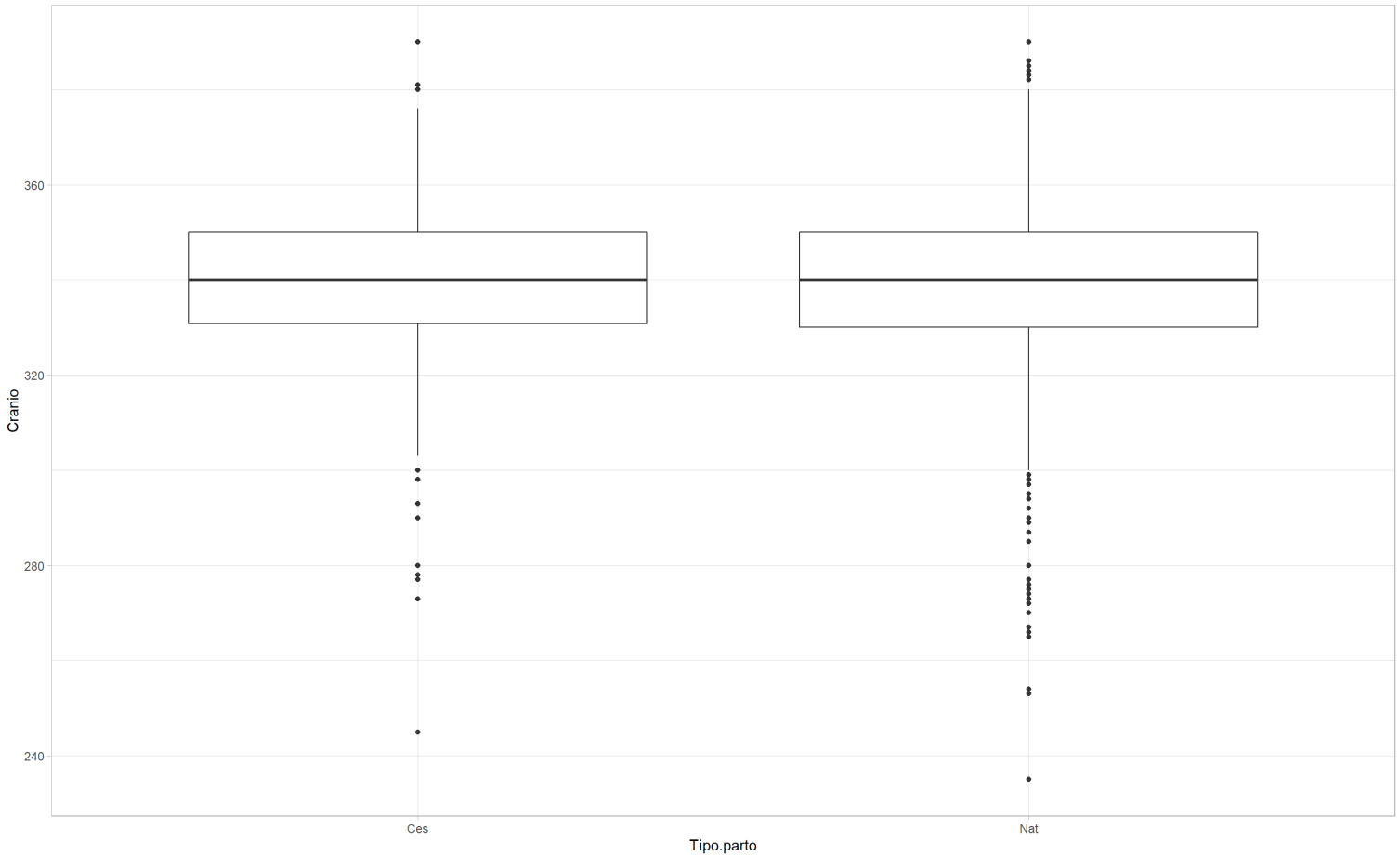
2.2) *Tipo parto & Lunghezza*




```
> t.test(Lunghezza~Tipo.parto, data=df)

Welch Two Sample t-test

data:  Lunghezza by Tipo.parto
t = 2.1389, df = 1517.3, p-value = 0.0326
alternative hypothesis: true difference in means between group Ces and group Nat is not equal to 0
95 percent confidence interval:
0.1954027 4.5172874
sample estimates:
mean in group Ces mean in group Nat
         496.3654          494.0090
```

I risultati del test indicano che, in media, a parti cesarei è associato un incremento della lunghezza del neonato pari a circa 2 mm.

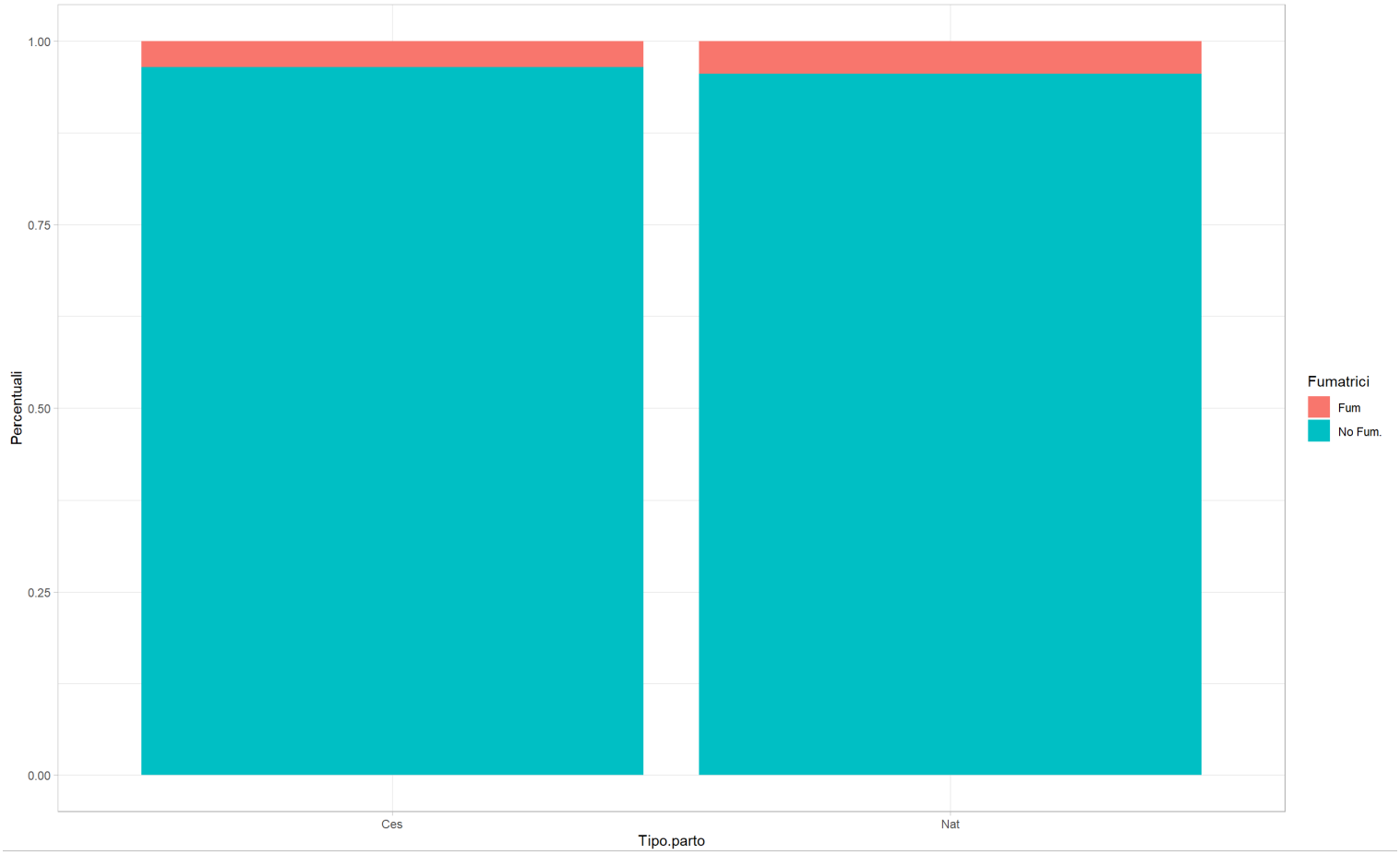
2.3) *Tipo parto* & *Diametro cranio*




```
> t.test(Cranio~Tipo.parto, data=df)

Welch Two Sample t-test

data:  Cranio by Tipo.parto
t = -0.03405, df = 1463.2, p-value = 0.9728
alternative hypothesis: true difference in means between group Ces and group Nat is not equal to 0
95 percent confidence interval:
-1.394615  1.347024
sample estimates:
mean in group Ces mean in group Nat
         340.0124          340.0362
```

In questo caso non ci sono evidenze di un legame tra il diametro del cranio e la tipologia di parto adoperata.

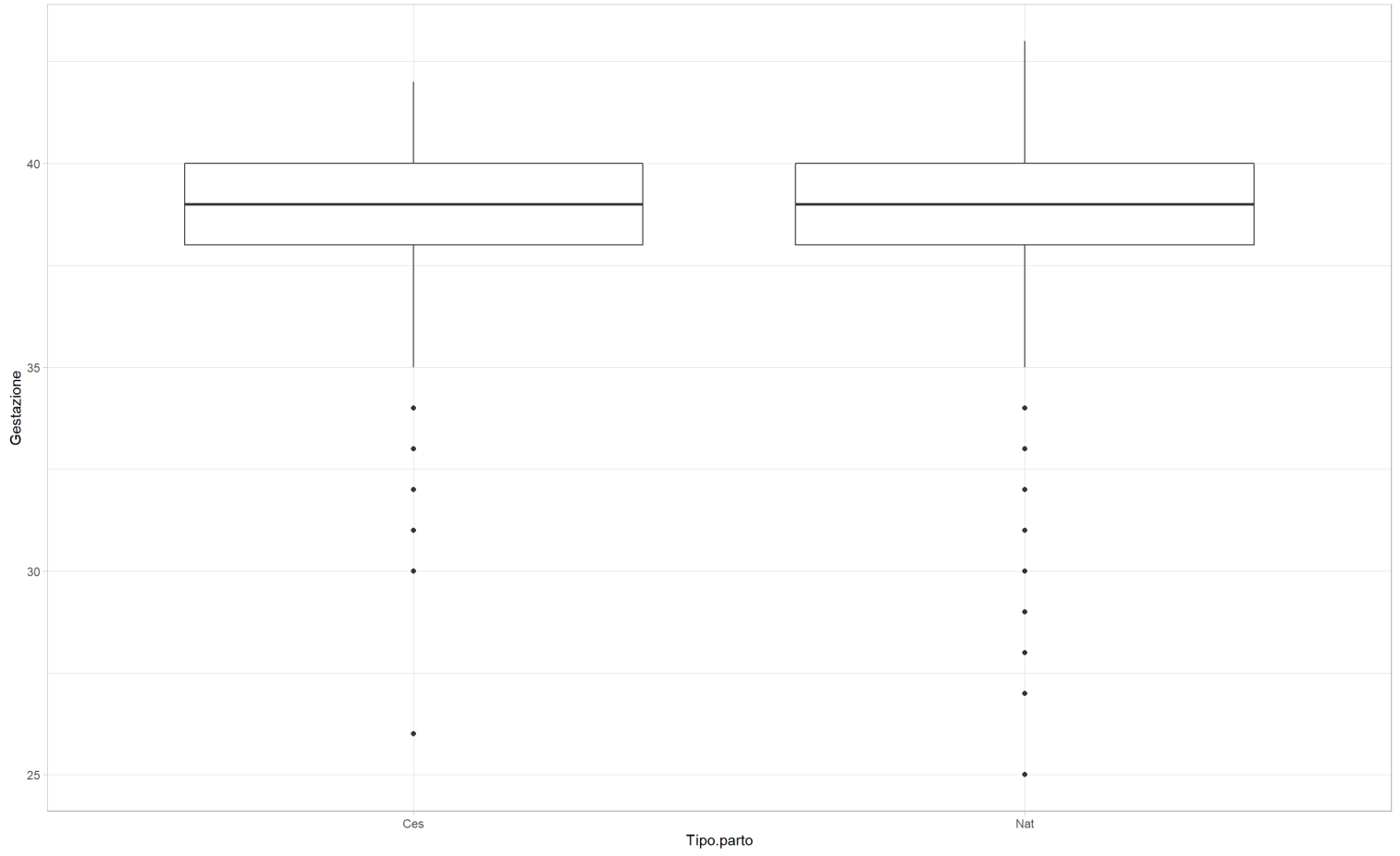
2.4) *Tipo parto* & *Fumo*




```
> chisq.test(cont_tabPF)

Pearson's Chi-squared test with Yates' continuity correction

data:  cont_tabPF
X-squared = 0.70493, df = 1, p-value = 0.4011
```

Non si evidenziano relazioni tra l'essere fumatrice e il tipo di parto scelto.

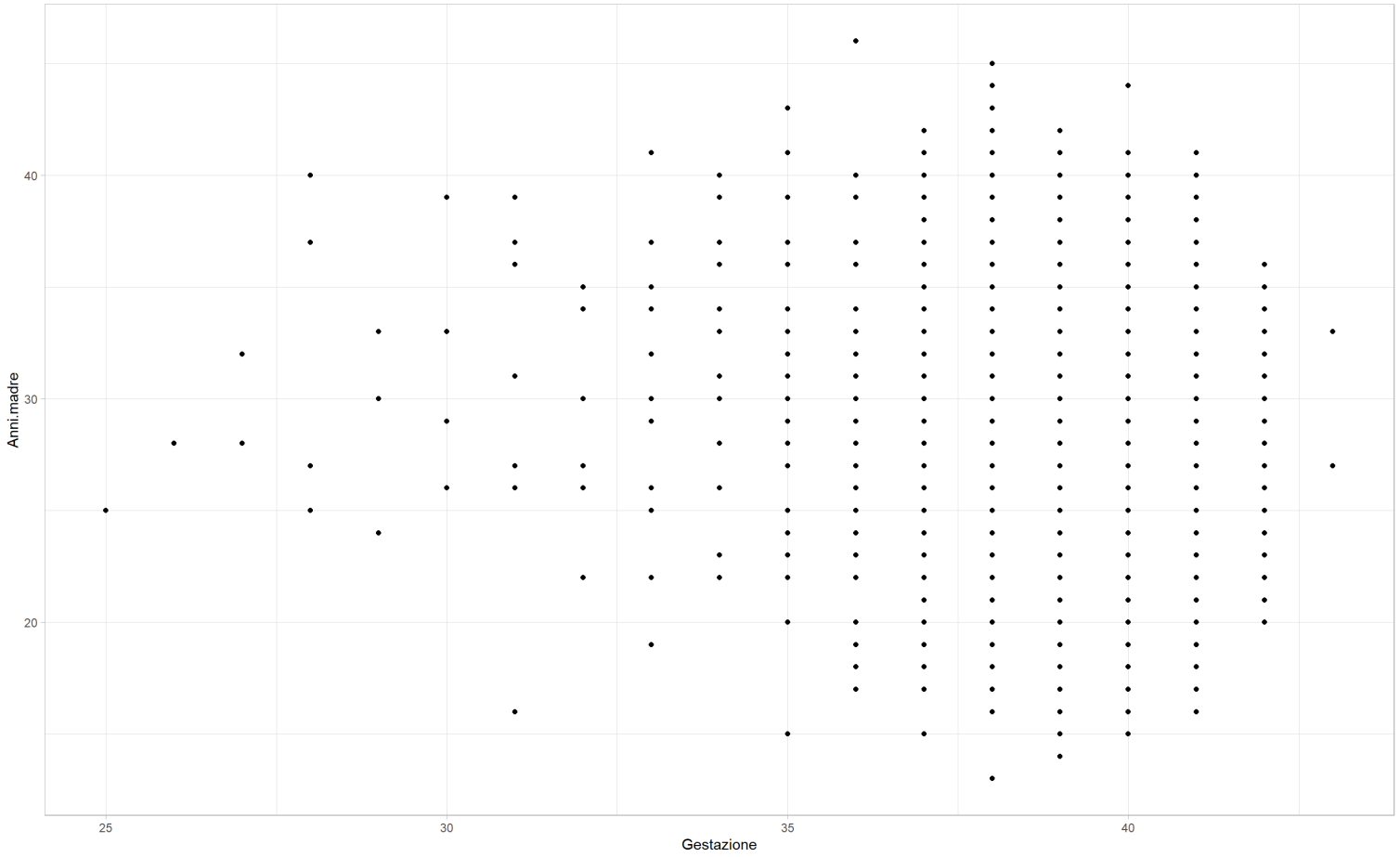
2.5) *Tipo parto & Gestazione*



```
> t.test(Gestazione~Tipo.parto, data=df)

Welch Two Sample t-test

data:  Gestazione by Tipo.parto
t = 0.78196, df = 1484.4, p-value = 0.4344
alternative hypothesis: true difference in means between group Ces and group Nat is not equal to 0
95 percent confidence interval:
-0.09318183  0.21672129
sample estimates:
mean in group Ces mean in group Nat
         39.02335          38.96158
```

Non ci sono differenze in media in termini di durata della gestazione tra i due tipi di parto. Ciò significa che la tipologia di parto eseguita non è legata ad una gravidanza prematura o protratta.

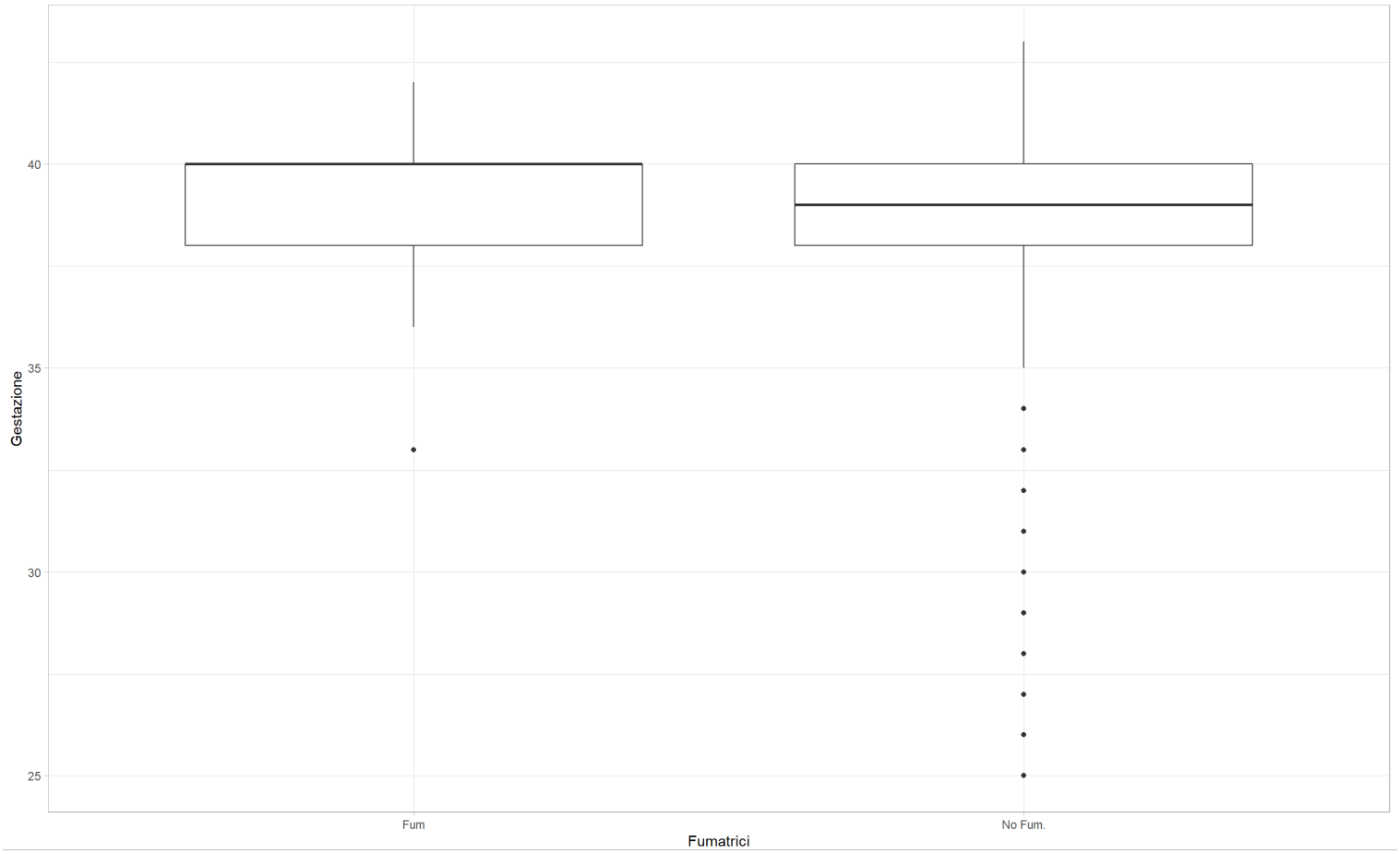
3.1) *Gestazione* & *Età madre*




```
> cor(df$Anni.madre, df$Gestazione)

[1] -0.134942
```

Dal grafico e dal coefficiente di correlazione si denota che non c'è alcuna relazione tra gli anni della madre e le settimane di gestazione.

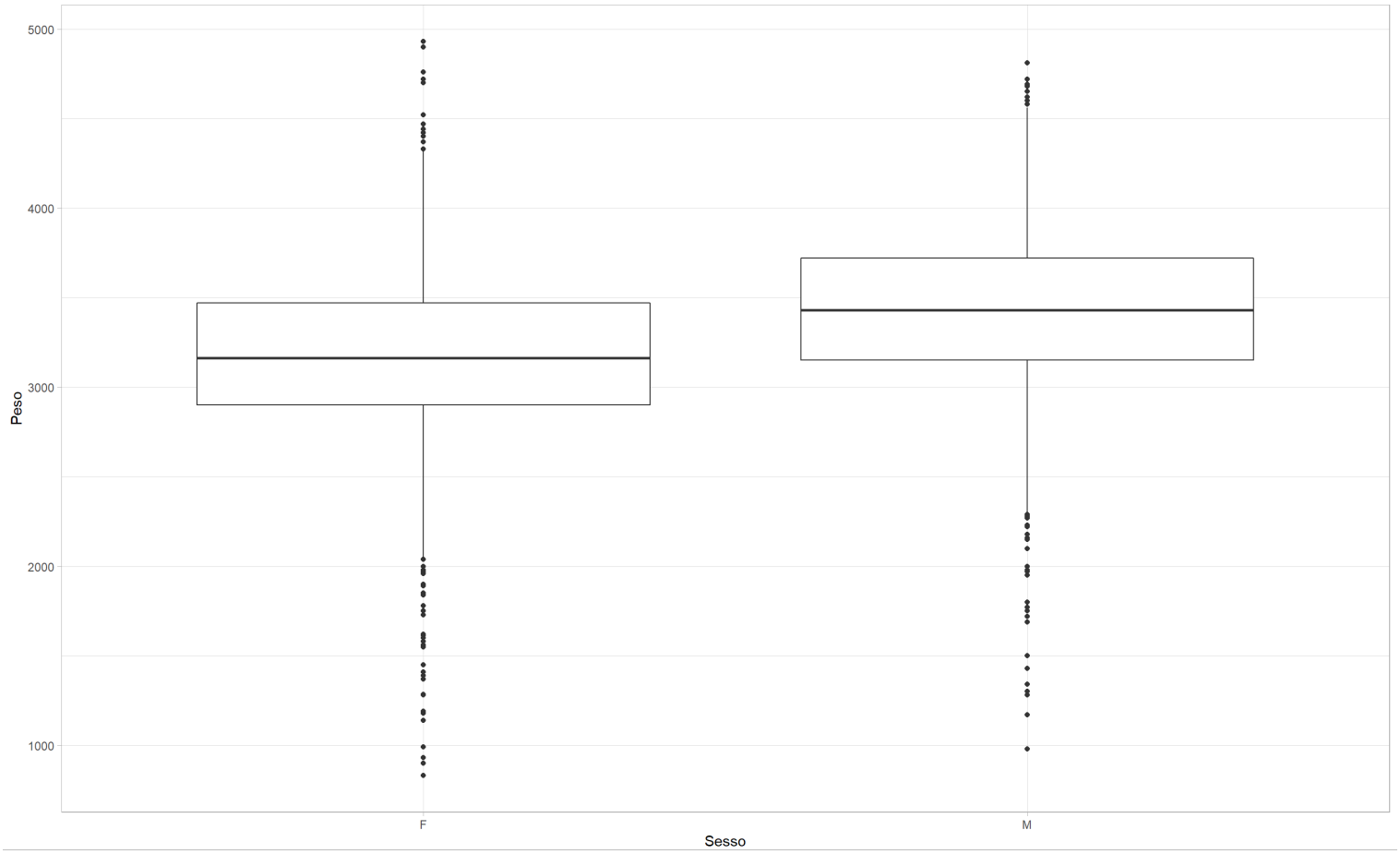
3.2) *Gestazione* & *Fumo*




```
> t.test(Gestazione~Fumatrici, data=df)

Welch Two Sample t-test

data:  Gestazione by Fumatrici
t = 2.0883, df = 119.28, p-value = 0.0389
alternative hypothesis: true difference in means between group Fum and group No Fum. is not equal to 0
95 percent confidence interval:
0.01566887 0.58879100
sample estimates:
    mean in group Fum mean in group No Fum.
              39.26923              38.96700
```

Ad un livello di significatività del 5% si può concludere che, in media, la gestazione delle madri non fumatrici duri qualche giorno in rispetto a quella delle donne fumatrici. Siccome per entrambe le categorie siamo nel range di settimane (37-41) previsto dal Ministero della Salute il fatto di fumare o meno non ha un impatto sulla salute del bambino.

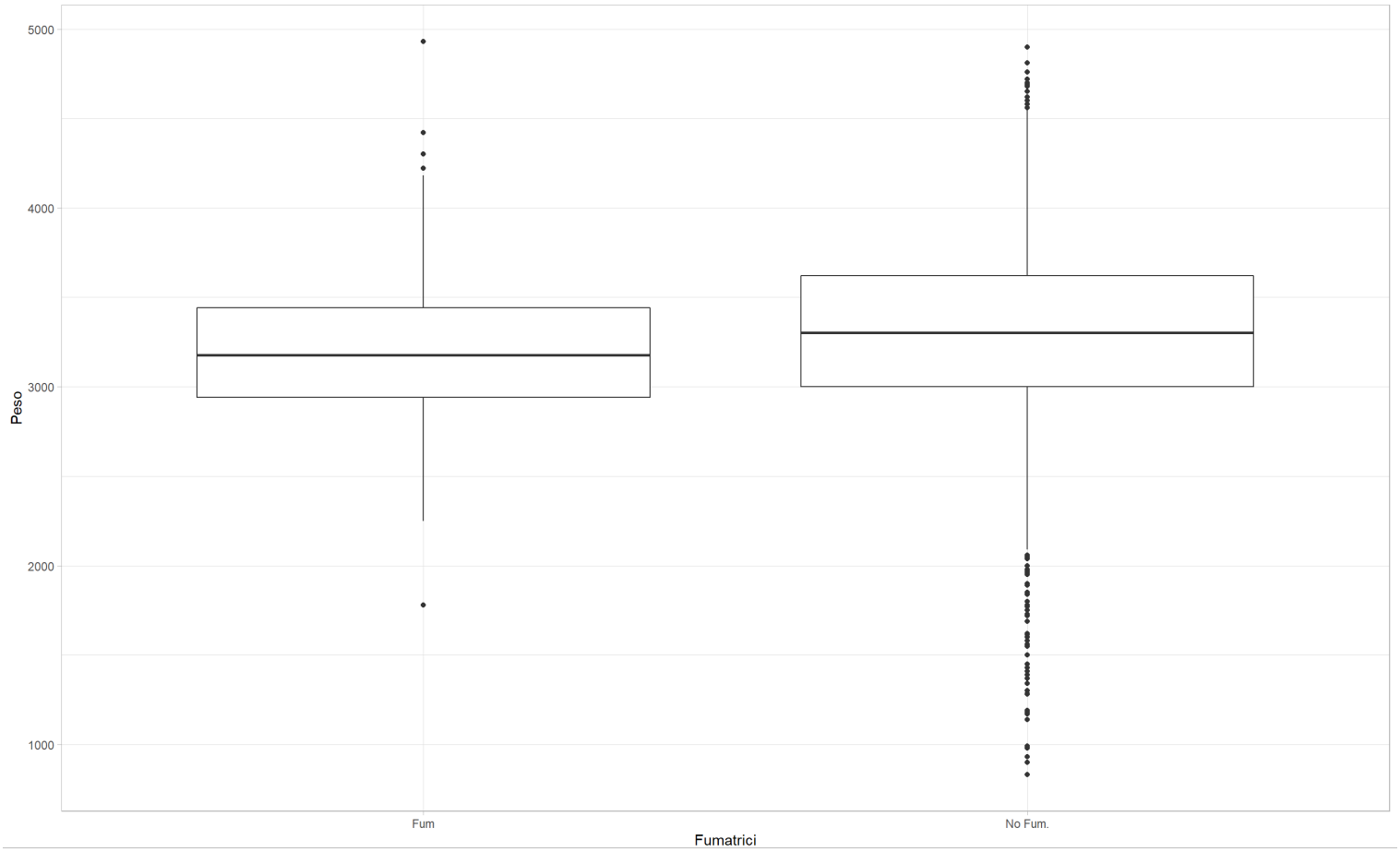
4.1) *Peso & Sesso*




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

Welch Two Sample t-test

data:  Peso by Sesso
t = -12.115, df = 2488.7, 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:
-287.4841 -207.3844
sample estimates:
mean in group F mean in group M
       3161.061        3408.496
```

I risultati del test confermano l'ipotesi di differenza in media tra il peso alla nascita tra maschi e femmine, in particolare i maschi pesano, in media, 300 gr in più.

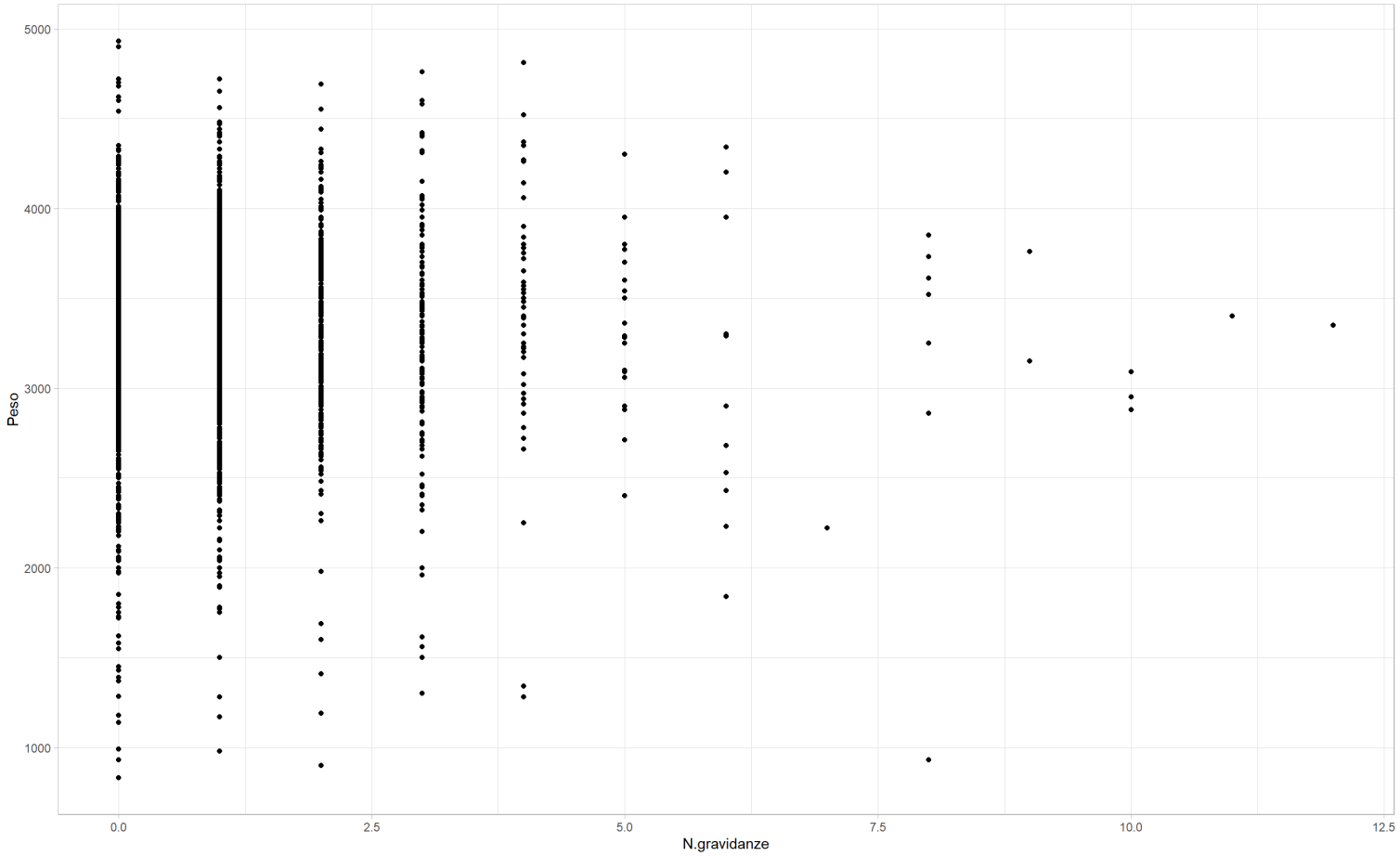
4.2) *Peso* & *Fumo*




```
> t.test(Peso~Fumatrici, data=df)

Welch Two Sample t-test

data:  Peso by Fumatrici
t = -1.0362, df = 114.12, p-value = 0.3023
alternative hypothesis: true difference in means between group Fum and group No Fum. is not equal to 0
95 percent confidence interval:
-145.3399   45.5076
sample estimates:
    mean in group Fum mean in group No Fum.
             3236.346              3286.262
```

Il risultato test ci fornisce un'ulteriore conferma del fatto che l'essere fumatrice o meno non ha una relazione diretta significativo sulla salute del neonato.

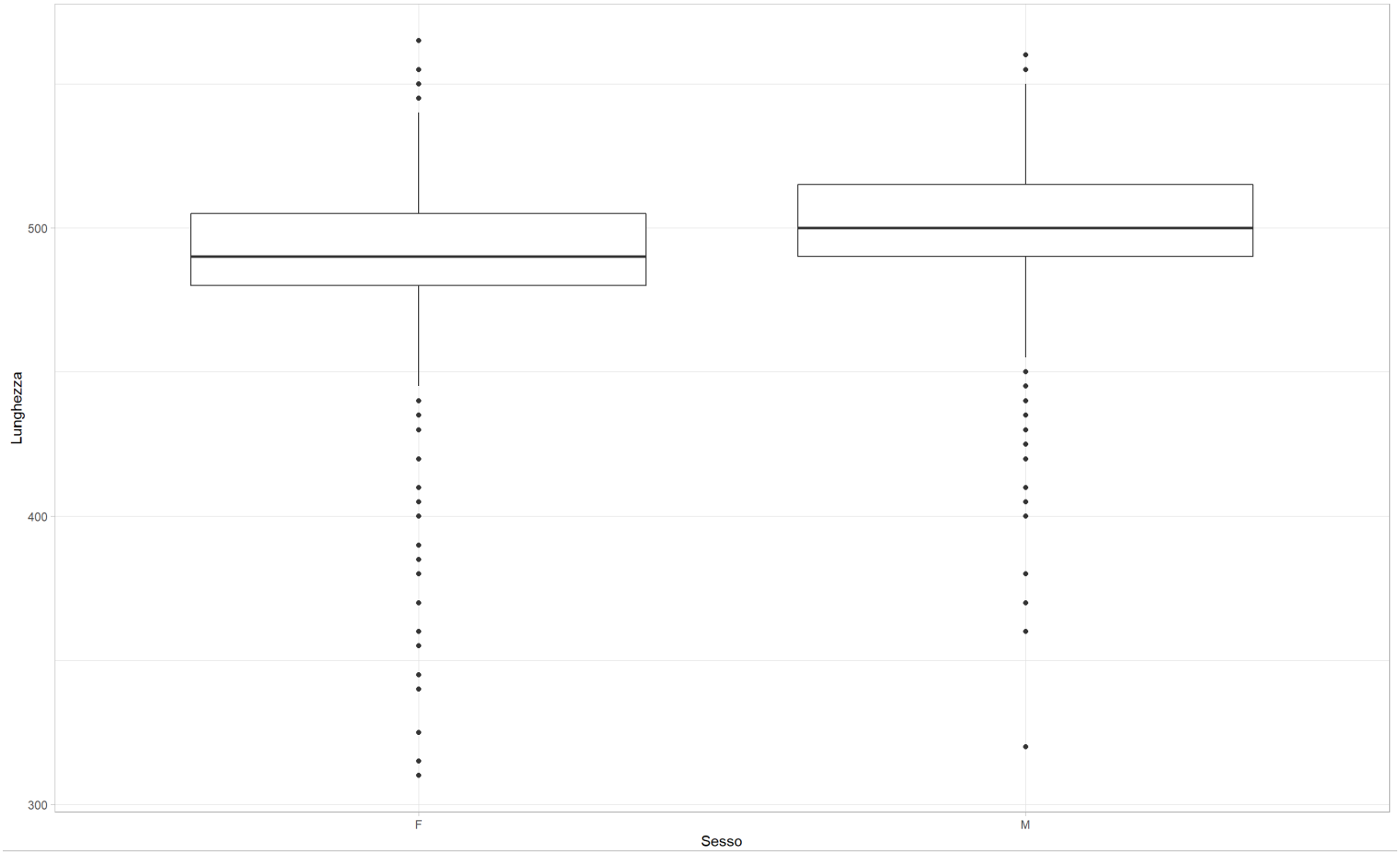
4.3) *Peso & Numero gravidanze precedenti*




```
> cor(df$Peso, df$N.gravidanze)
[1] 0.00227711
```

Si esclude ogni tipo di legame tra le variabili.

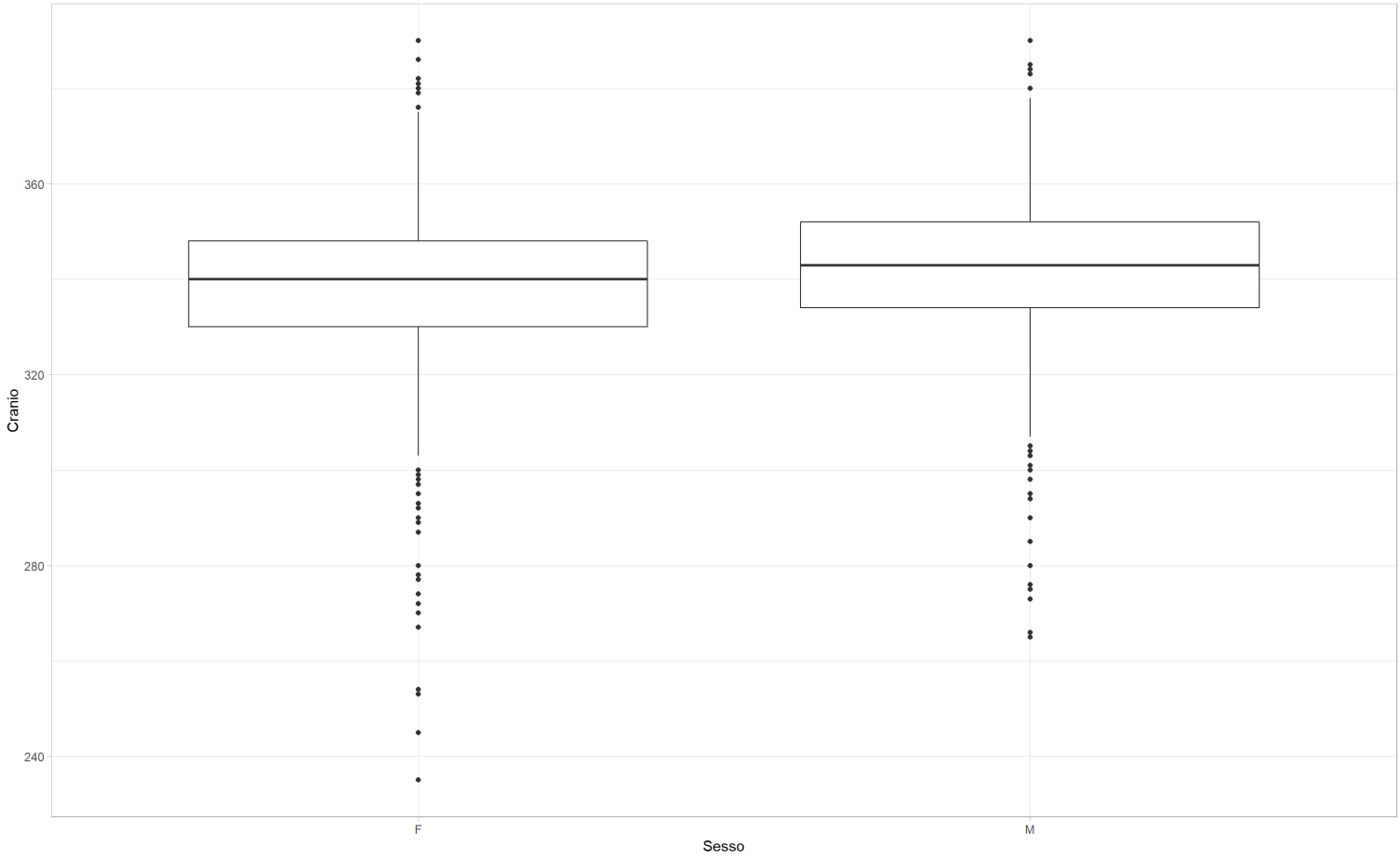
5.1) *Sesso* & *Lunghezza*



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

Welch Two Sample t-test

data:  Lunghezza by Sesso
t = -9.5823, df = 2457.3, 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:
-11.939001  -7.882672
sample estimates:
mean in group F mean in group M
       489.7641        499.6750
```

Sulla base delle evidenze empiriche si può affermare che, in media, i neonati maschi vengono al mondo con una lunghezza maggiore di 9.5 mm rispetto alle femmine.

5.2) *Sesso* & *Diametro del cranio*




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

Welch Two Sample t-test

data:  Cranio by Sesso
t = -7.4366, df = 2489.4, p-value = 1.414e-13
alternative hypothesis: true difference in means between group F and group M is not equal to 0
95 percent confidence interval:
-6.110504 -3.560417
sample estimates:
mean in group F mean in group M
       337.6231        342.4586
```

I dati del campione mostrano che in media i neonati maschi hanno una misura del cranio maggiore di circa 5 mm rispetto alle femmine.

2.4 TEST SULLA MEDIA DELLA POPOLAZIONE

Si effettuano due test di verifica di ipotesi per confrontare le medie del campione delle variabili *Peso* e *Lunghezza*. A tal proposito, la fonte utilizzata per la scelta delle medie come parametri della popolazione è l'Ospedale Pediatrico Bambino Gesù.

Media *Peso*

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

One Sample t-test

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

Non si rigetta l'ipotesi nulla di uguaglianza delle medie ad un livello di significatività del 5%.

Media *Lunghezza*

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

One Sample t-test

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

In questo caso si conclude che la media campionaria della lunghezza del neonato è statisticamente diversa dal vero parametro della popolazione.

3. ANALISI MULTIDIMENSIONALE

Nel seguente capitolo si cercherà di raggiungere due importanti risultati:

- Capire se, inserendo variabili di controllo all'interno di un modello di regressione, possa esserci un'influenza del fumo sul peso del neonato, quindi fare inferenza causale;

-Costruire il miglior modello possibile per fare predizioni sulla base delle variabili inserite nel dataset.

3.1 INFERENZA CAUSALE

Si parte dall'analizzare un modello di regressione lineare semplice con il peso come variabile di risposta e la variabile dicotomica *Fumatrici* come predittore:

```
> mod_fum <- lm(Peso~Fumatrici, data = df)

> summary(mod_fum)

.........

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3286.26      10.73 306.131   <2e-16 ***
Fumatrici     -49.92      52.61  -0.949    0.343
```

Il modello è affetto da distorsione da variabile omessa, in questo caso il coefficeinte potrebbe essere stato sottostimato (sovrastimato in valore assoluto) in quanto potrebbe includere la variabilità di altri fattori correlati con la variabile indipendente che abbiano un effetto su quella dipendente. Si vanno a verificare le condizioni che segnalano un problema di distorsione:

- Variabili correlate con *Fumatrici*;

- Variabili determinanti di *Peso*.

```
> cor(df, df$Fumatrici)

                       [,1]
Anni.madre      0.005240840
N.gravidanze    0.046813090
Fumatrici       1.000000000
Gestazione      0.032308217
Peso           -0.018987400
Lunghezza      -0.020811781
Cranio         -0.008667282
Tipo.parto_ces -0.019003895
Sesso_m         0.013026054
Ospedale1       0.012935995
Ospedale2       0.00294124
```

Non sembrano esserci correlazioni con altre variabili presenti nel dataset.

Si può pensare ad eventuali effetti interazione in quanto l'effetto del fumo sulla salute del nascituro può dipendere dall'età della madre.

```
> mod_fum2 <- lm(Peso~Fumatrici*Anni.madre, data=df)

> summary(mod_fum2)

 ..........

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3347.987     59.079  56.669   <2e-16 ***
Fumatrici             73.008    280.827   0.260    0.795
Anni.madre            -2.190      2.062  -1.062    0.288
Fumatrici:Anni.madre  -4.330      9.744  -0.444    0.657
```

Il modello non è significativo, si ricorre a S.E. robusti all'eteroschedasticità per provare a rendere significativi i coefficienti nel caso in cui il modello soffra di eteroschedasticità.

```
> library(estimatr)

> mod_fum2_rob <- lm_robust(Peso~Fumatrici*Anni.madre, data=df)

> summary(mod_fum2_rob)

 ..........

Coefficients:
                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)          3347.99     60.692 55.1638   0.0000 3228.975 3466.998 2494
Fumatrici              73.01    222.413  0.3283   0.7427 -363.124  509.141 2494
Anni.madre             -2.19      2.181 -1.0044   0.3153   -6.466    2.086 2494
Fumatrici:Anni.madre   -4.33      7.806 -0.5547   0.5791  -19.638   10.977  249
```

Non si notano miglioramenti significativi del modello, in definitva si può concludere che, date le variabili contenute del dataset, non ci sono evidenze statistiche sull'influenza del fumo sul peso del neonato. In un futuro lavoro si può pensare di raccogliere altre variabili relative al consumo di alcool e alimentazione per utilizzarle come variabili di controllo.

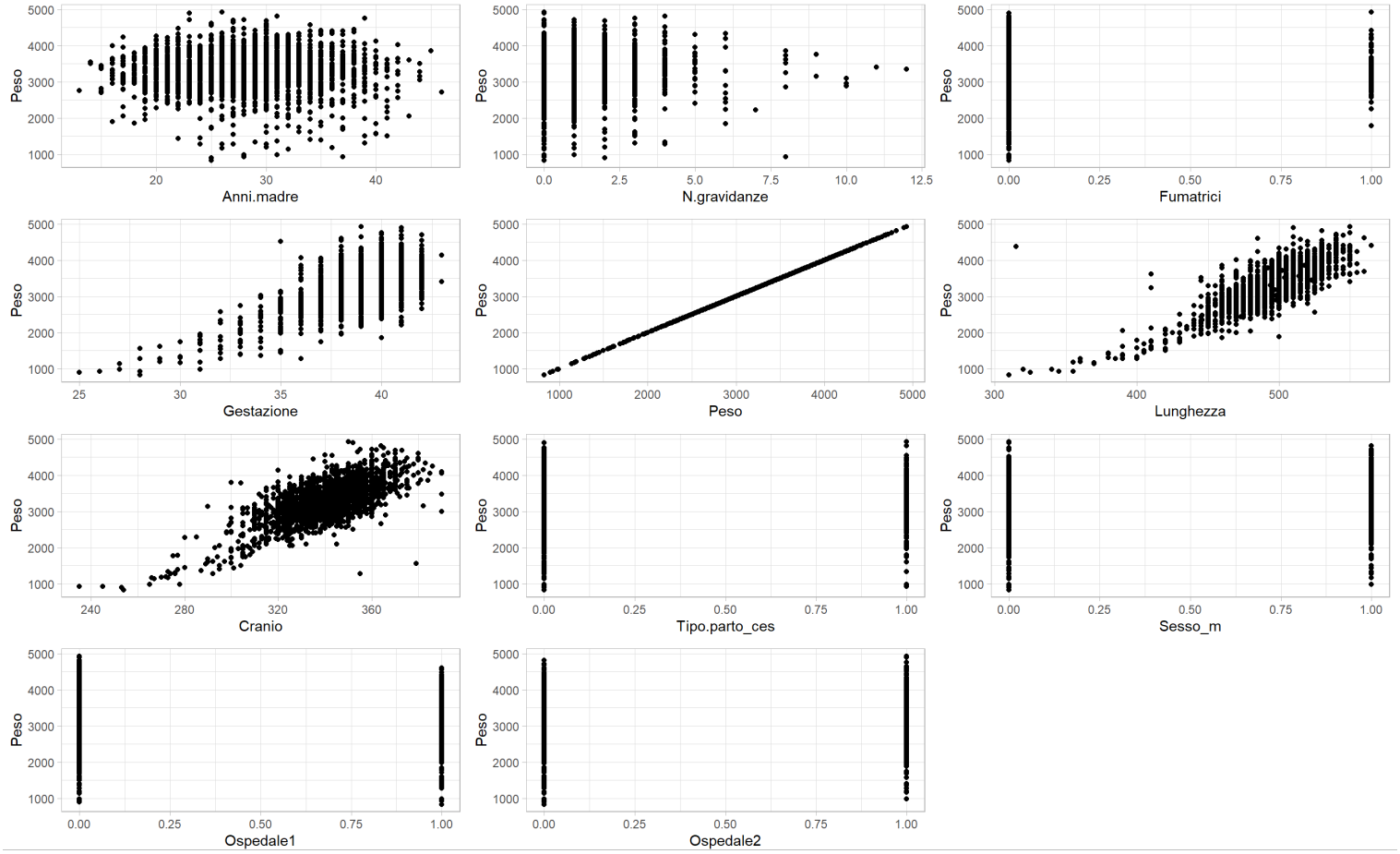
3.2 MODELLO PREDITTIVO

In questo capitolo si cercherà di selezionare il miglior modello per fare previsione del peso di un neonato, date le variabili contenute nel dataset.

Innanzitutto si considera un modello che ha come predittori tutte le variabili a disposizione:

```
> mod_gen <- lm(Peso~., data=df)

> summary(mod_gen)

..........


Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6677.9129   141.3984 -47.228  < 2e-16 ***
Anni.madre         0.8018     1.1467   0.699  0.48449
N.gravidanze      11.3812     4.6686   2.438  0.01485 *
Fumatrici        -30.2741    27.5492  -1.099  0.27191
Gestazione        32.5773     3.8208   8.526  < 2e-16 ***
Lunghezza         10.2922     0.3009  34.207  < 2e-16 ***
Cranio            10.4722     0.4263  24.567  < 2e-16 ***
Tipo.parto_ces   -29.6335    12.0905  -2.451  0.01432 *
Sesso_m           77.5723    11.1865   6.934 5.18e-12 ***
Ospedale1        -28.2495    13.5054  -2.092  0.03657 *
Ospedale2        -39.3408    13.3838  -2.939  0.00332 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274 on 2487 degrees of freedom
Multiple R-squared:  0.7289, Adjusted R-squared:  0.7278
```

Si considerino eventuali effetti non lineari:



Si osserva che la relazione tra *Peso* e le variabili *Gestazione*, *Lunghezza* e *Cranio* può essere meglio approssimata da un polinomio di secondo grado.

Inoltre sarà valutato nuovamente l'effetto interazione tra il fumo e l'età della madre.

```
> mod_gen2 <- update(mod_gen,~.+Fumatrici:Anni.madre+I(Gestazione^2)+I(Lunghezza^2)+I(Cranio^2))


> mod_gen3 <- update(mod_gen,~.+Fumatrici:Anni.madre+poly(Gestazione,2)+poly(Lunghezza,2)+poly(Cranio,2))
```

Sono stati due modi per costruire i polinomi di secondo grado:

- I(x^2) permette di costruire un polinomio elevando semplicemente la modalità della variabile al quadrato;

- poly() realizza dei polinomi ortogonali, andando a limitare il naturale incremento di multicollinearità associato all'utilizzo di trasformazioni polinomiali.

In seguito si utilizzerà il criterio del BIC ai tre modelli per la feature selection, ovvero selezione delle variabili che forniscono informazione.

```
> mod_gen_upd <- stepAIC(mod_gen, direction = "both", k=log(nrow(df)))

> summary(mod_gen_upd)

.........
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6681.7251   135.8036 -49.201  < 2e-16 ***
N.gravidanze    12.4554     4.3416   2.869  0.00415 **
Gestazione      32.3827     3.8008   8.520  < 2e-16 ***
Lunghezza       10.2455     0.3008  34.059  < 2e-16 ***
Cranio          10.5410     0.4265  24.717  < 2e-16 ***
Sesso_m         77.9807    11.2111   6.956 4.47e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274.7 on 2492 degrees of freedom
Multiple R-squared:  0.727, Adjusted R-squared:  0.7265



> mod_gen_upd2 <- stepAIC(mod_gen2, direction = "both", k=log(nrow(df)))



> summary(mod_gen_upd2)



..........

Coefficients:

                  Estimate Std. Error t value Pr(>|t|)

(Intercept)     -1.584e+03  9.066e+02  -1.747 0.080769 .

N.gravidanze     1.441e+01  4.247e+00   3.392 0.000704 ***

Gestazione       3.624e+02  6.260e+01   5.789 7.96e-09 ***

Lunghezza       -3.016e+01  4.039e+00  -7.467 1.13e-13 ***

Sesso_m          7.239e+01  1.100e+01   6.580 5.73e-11 ***

I(Gestazione^2) -4.206e+00  8.233e-01  -5.109 3.49e-07 ***

I(Lunghezza^2)   4.167e-02  4.145e-03  10.055  < 2e-16 ***

I(Cranio^2)      1.542e-02  6.179e-04  24.965  < 2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 268.5 on 2490 degrees of freedom

Multiple R-squared:  0.7395, Adjusted R-squared:  0.7387

> mod_gen_upd3 <- stepAIC(mod_gen3, direction = "both", k=log(nrow(df)))

> summary(mod_gen_upd3)

..........

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3234.073      8.635 374.512  < 2e-16 ***
N.gravidanze           14.395      4.248   3.389 0.000713 ***
Sesso_m                72.309     11.005   6.570 6.09e-11 ***
poly(Gestazione, 2)1 4795.271    401.107  11.955  < 2e-16 ***
poly(Gestazione, 2)2 -1865.951   377.571  -4.942 8.25e-07 ***
poly(Lunghezza, 2)1  12427.454   418.443  29.699  < 2e-16 ***
poly(Lunghezza, 2)2   3508.406   391.643   8.958  < 2e-16 ***
poly(Cranio, 2)1      8376.968   366.984  22.827  < 2e-16 ***
poly(Cranio, 2)2       637.758   405.579   1.572 0.115970
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 268.5 on 2489 degrees of freedom
Multiple R-squared:  0.7395, Adjusted R-squared:  0.7387
```

Si ricorre nuovamente al criterio del BIC, questa volta per la model selection, ovvero selezionare il modello che gestisce meglio il trade-off tra complessità e informazione.

```
> BIC(mod_gen_upd, mod_gen2_upd, mod_gen3_upd)


             df      BIC

mod_gen_upd   7 35193.65

mod_gen2_upd  9 35092.68

mod_gen3_upd 10 35100.25
```

Si sceglie il modello con il valore BIC, quindi si dovrebbe propendere per il secondo modello. Però, dando una rapida occhiata ai predittori selezionati ci si accorge che sono stati inclusi polinomi di secondo grado ma non le variabili originali. Sarà utilizzato quindi mod_gen3_upd come modello di riferimento.

```
> mod_gen_def <- mod_gen3_upd

> summary(mod_gen_def)

Call:
lm(formula = Peso ~ N.gravidanze + Sesso_m + poly(Gestazione,
    2) + poly(Lunghezza, 2) + poly(Cranio, 2), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1187.0  -182.9   -11.8   162.7  1469.8

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3234.073      8.635 374.512  < 2e-16 ***
N.gravidanze           14.395      4.248   3.389 0.000713 ***
Sesso_m                72.309     11.005   6.570 6.09e-11 ***
poly(Gestazione, 2)1 4795.271    401.107  11.955  < 2e-16 ***
poly(Gestazione, 2)2 -1865.951    377.571  -4.942 8.25e-07 ***
poly(Lunghezza, 2)1  12427.454    418.443  29.699  < 2e-16 ***
poly(Lunghezza, 2)2   3508.406    391.643   8.958  < 2e-16 ***
poly(Cranio, 2)1      8376.968    366.984  22.827  < 2e-16 ***
poly(Cranio, 2)2       637.758    405.579   1.572 0.115970
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 268.5 on 2489 degrees of freedom
Multiple R-squared:  0.7395, Adjusted R-squared:  0.7387
F-statistic: 883.2 on 8 and 2489 DF,  p-value: < 2.2e-16
```

Si passa, ora, all'analisi della multicollinearità.

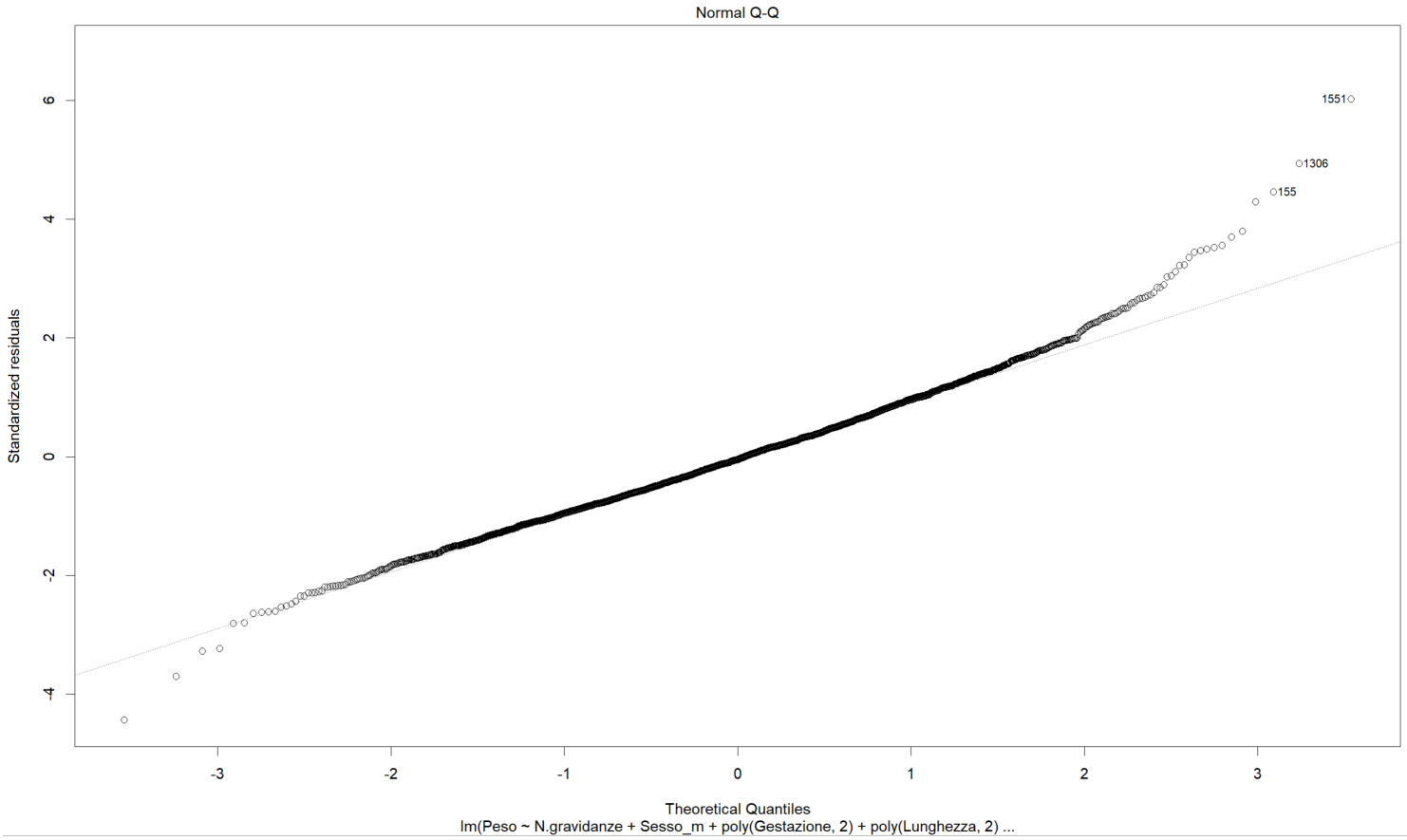
```
> vif(mod_gen_def, type="predictor")

                GVIF Df GVIF^(1/(2*Df)) Interacts With
N.gravidanze  1.025546  1        1.012692           --
Sesso_m       1.049072  1        1.024242           --
Gestazione   11.823068  0             Inf           --
Lunghezza    11.823068  0             Inf           --
Cranio       11.823068  0             Inf           --
                                          Other Predictors
N.gravidanze              Sesso_m, Gestazione, Lunghezza, Cranio
Sesso_m              N.gravidanze, Gestazione, Lunghezza, Cranio
Gestazione   N.gravidanze, Sesso_m, Gestazione, Lunghezza, Cranio
Lunghezza    N.gravidanze, Sesso_m, Gestazione, Lunghezza, Cranio
Cranio       N.gravidanze, Sesso_m, Gestazione, Lunghezza, Cranio
```

Utilizzo l'argomento type="predictor" per annullare l'effetto dei polinomi (ovviamente correlati con la variabile di grado 1) sul calcolo dei VIF. VIF superiore a 10 per tre predittori: *Gestazione*, *Lunghezza* e *Cranio*.

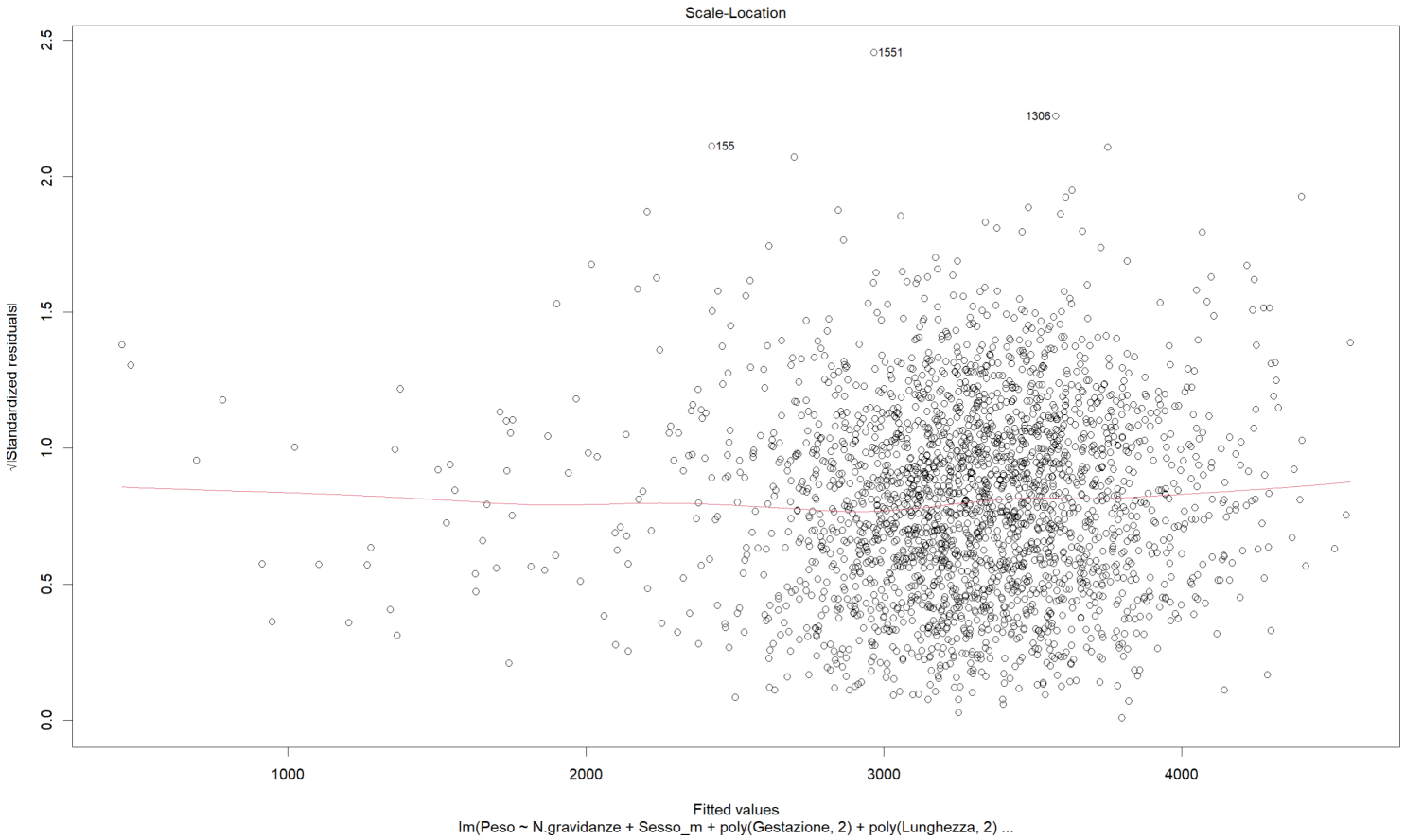
Quindi, al fine di risolvere il problema di multicollinearità, ho dapprima escluso il secondo grado della variabile *Cranio* e poi ho calcolato nuovamente i VIF. Non si evidenziano problemi.

Infine si va ad effettuare la fase di diagnostica sui residui.

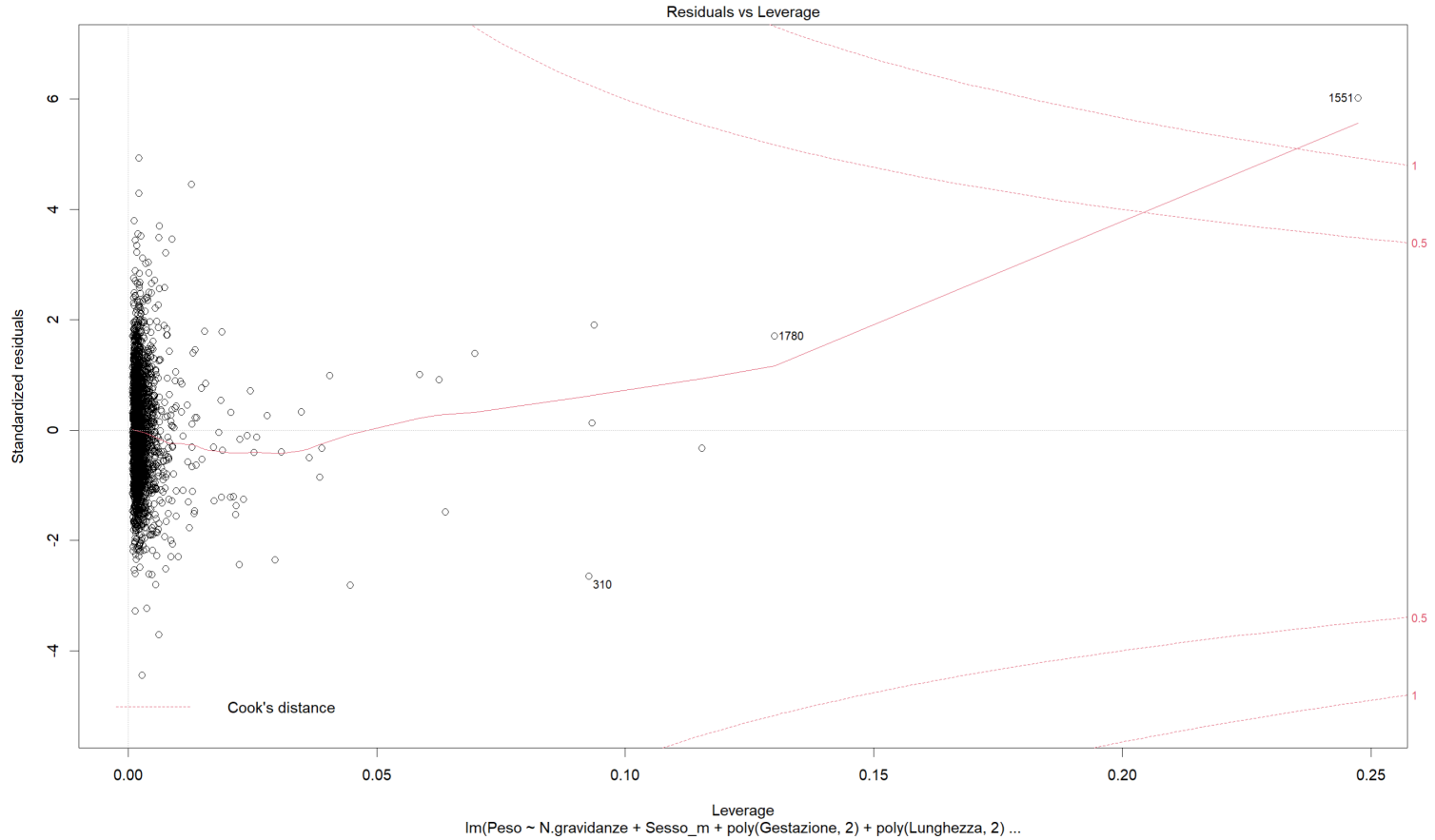
Assunzione degli errori a media 0 rispettata, non visualizzo dei pattern.



Assunzione distribuzione normale errori non perfettamente rispettata, problema nelle code che presentano osservazioni che si discostano dalla bisettrice.

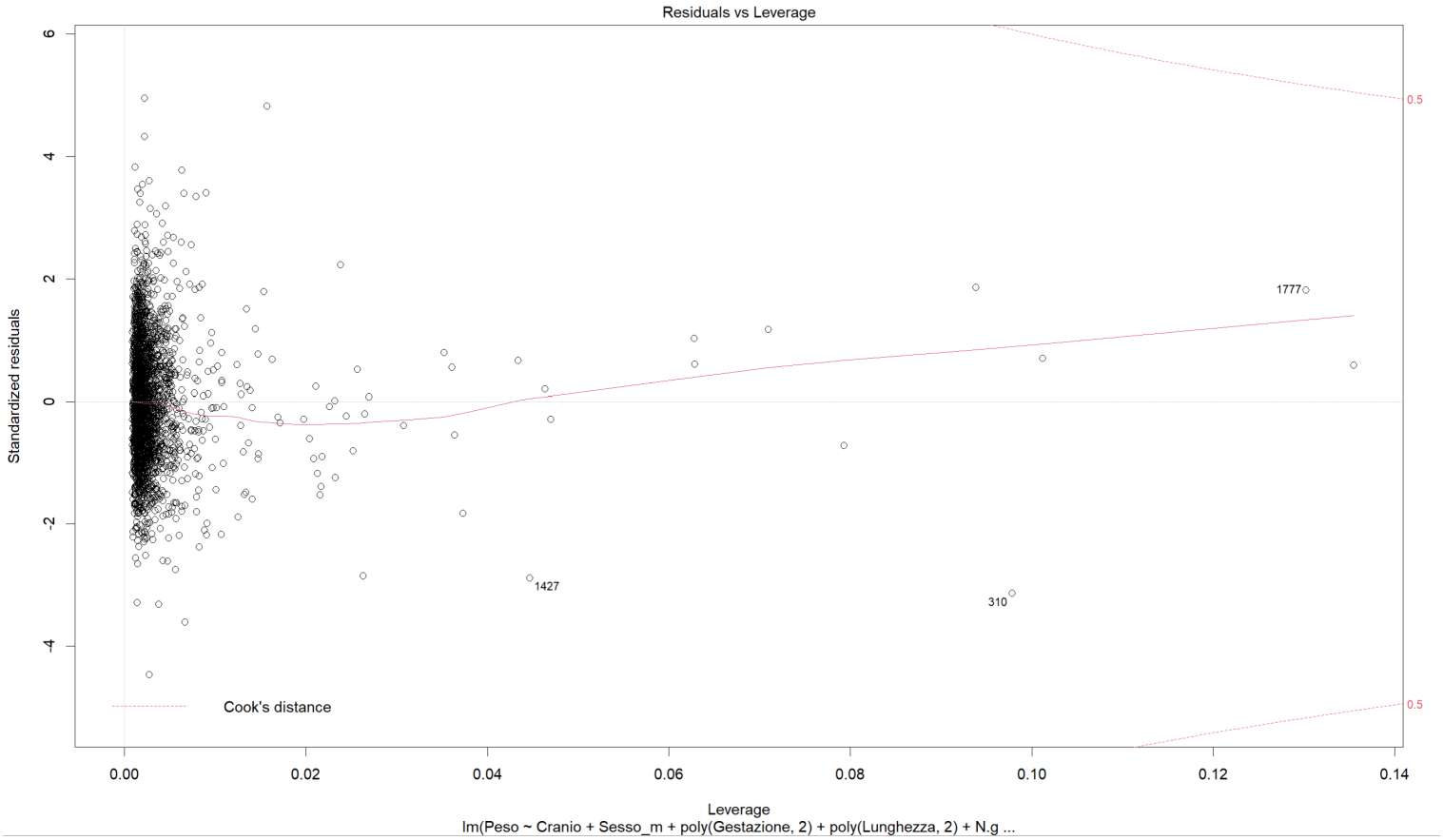


Assunzione omoschedasticità rispettata, visualizzo una nube casuale di punti.



L' osservazione 1551 è un outlier che potrebbe distorcere i risultati del modello. Quindi elimino l'osservazione dal dataset.

In seguito ho ricalcolato i BIC sulle tre tipologie di modelli presentate all'inizio del paragrafo ed i risultati vanno comunque nella direzione perseguita precedentemente. Ho constatato che non ci fossero problemi di collinearità ed effettuato nuovamente la diagnostica sui residui, i cui risultati sono invariati.

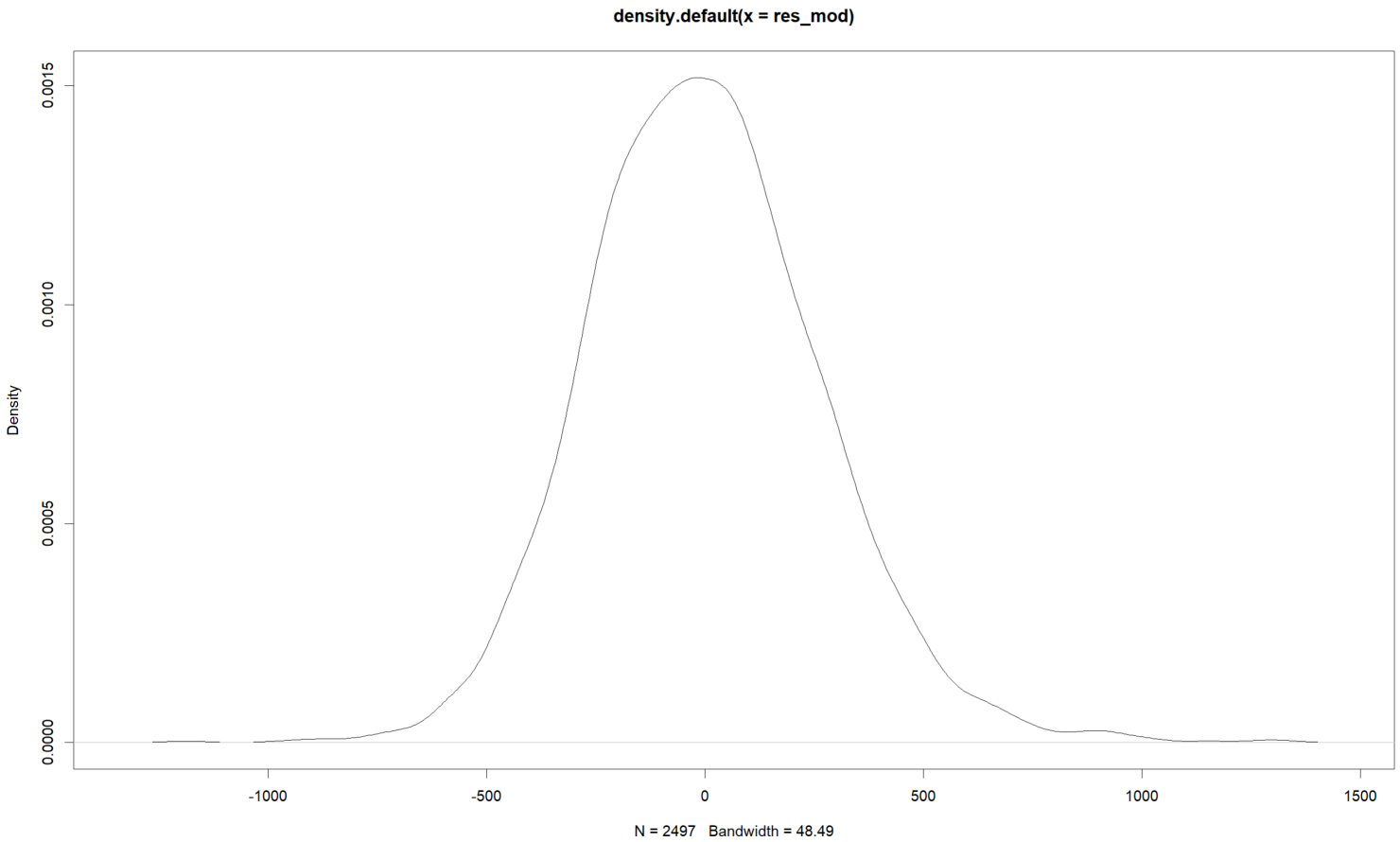


Dal grafico si osserva che non ci sono osservazioni che rientrano nella distanza di Cook.

Infine ho calcolato i vari indici per trovare conferma dalle intuizioni date dalle osservazioni grafiche.

```
> res_mod <- residuals(mod_gen_def)
> shapiro.test(res_mod)

Shapiro-Wilk normality test

data:  res_mod
W = 0.99048, p-value = 8.393e-1
```

Secondo il test i residui non seguono una distribuzione normale, da indagare ulteriormente con un grafico della densità di probabilità dei residui.



Dopo aver visualizzato il grafico ritengo che la non normalità della distribuzione dei residui, evidenziata dai risultati dello Shapiro test, non influenzi significativamente il potere predittivo del modello.

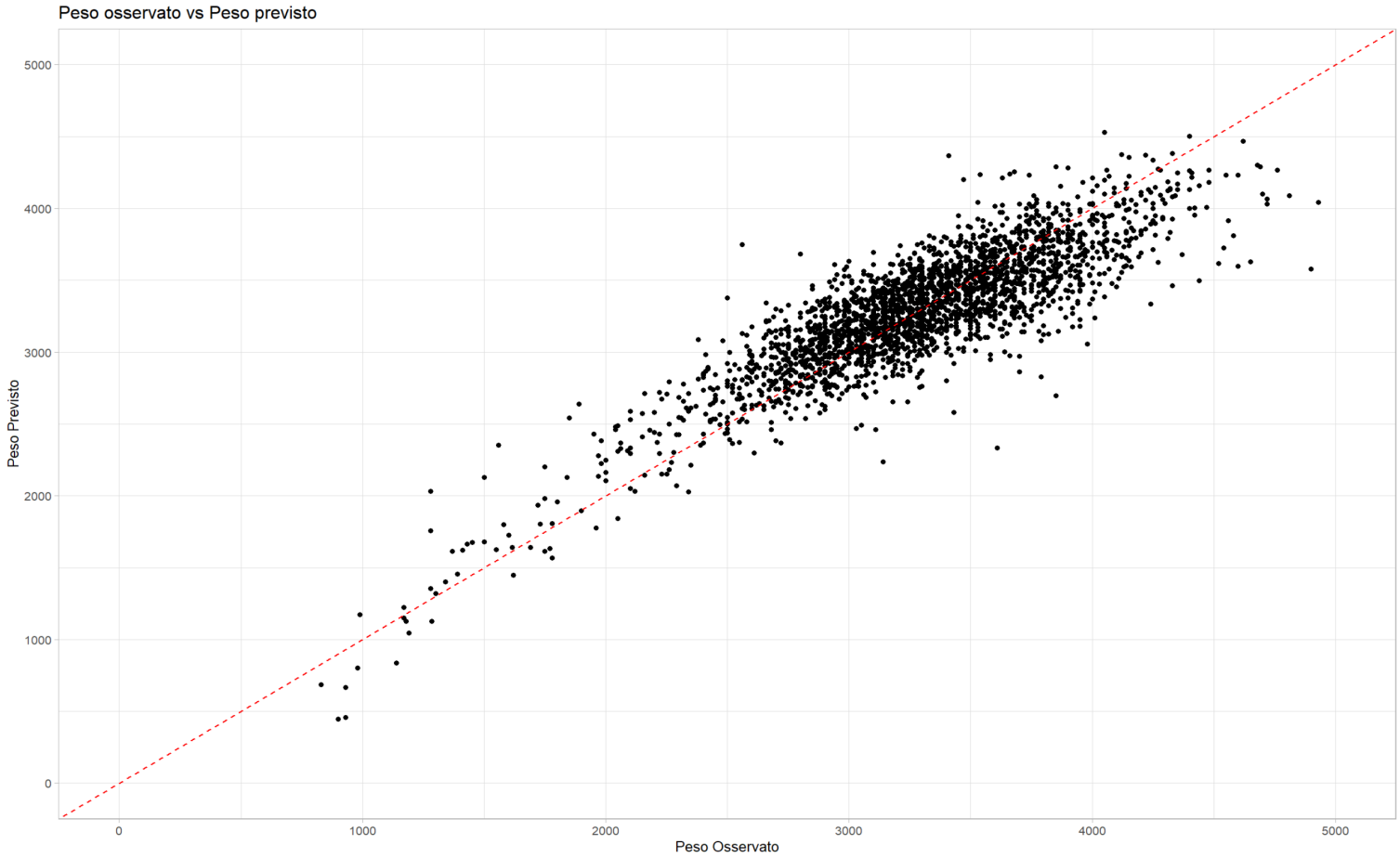
```
> bptest(mod_gen_def)

studentized Breusch-Pagan test

data:  mod_gen_def
BP = 17.125, df = 7, p-value = 0.0166
```

Si rigetta l'ipotesi nulla di omoschedasticità ad un livello di significatività del 5% ma non all'1%. Significa che c'è un leggero problema di eteroschedasticità che non credo vada ad inficiare sul potere predittivo del modello.

```
> dwtest(mod_gen_def)

Durbin-Watson test

data:  mod_gen_def
DW = 1.9496, p-value = 0.104
alternative hypothesis: true autocorrelation is greater than 0
```

Non si rigetta l'ipotesi nulla di auto-correlazione dei residui pari a 0.

In conclusione ho creato un grafico che cercasse in qualche modo di mostrare il potere predittivo del modello scelto. Siccome la rappresentazione di tante variabili è una procedura complessa sia da un punto di vista computazionale che cognitivo, ho deciso di costruire un grafico che mettesse a confronto i valori reali e quelli predetti della variabile dipendente Y.



In un modello "perfetto" tutti i punti si troverebbero sulla bisettrice in rosso. In questo possiamo comunque affermare che il modello è in grado di spiegare discretamente il fenomeno.