# Un modello statistico per prevedere il peso dei neonati

Dopo aver caricato il dataset e utilizzato un attach(), andiamo a studiare i dati nel dettaglio con la funzione head() per visualizzare le prime righe e, soprattutto, con la funzione summary() per esaminare gli indici di posizione.

```
> summary(neonati)
   Anni.madre     N.gravidanze       Fumatrici        Gestazione        Peso        Lunghezza        Cranio
 Min.   : 0.00   Min.   : 0.0000   Min.   :0.0000   Min.   :25.00   Min.   : 830   Min.   :310.0   Min.   :235
 1st Qu.:25.00   1st Qu.: 0.0000   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 :0.0000   Median :39.00   Median :3300   Median :500.0   Median :340
 Mean   :28.16   Mean   : 0.9812   Mean   :0.0416   Mean   :38.98   Mean   :3284   Mean   :494.7   Mean   :340
 3rd Qu.:32.00   3rd Qu.: 1.0000   3rd Qu.:0.0000   3rd Qu.:40.00   3rd Qu.:3620   3rd Qu.:510.0   3rd Qu.:350
 Max.   :46.00   Max.   :12.0000   Max.   :1.0000   Max.   :43.00   Max.   :4930   Max.   :565.0   Max.   :390
  Tipo.parto         Ospedale            Sesso
 Length:2500        Length:2500        Length:2500
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

Sono presenti diverse variabili quantitative continue, tra cui Anni.madre, N.gravidanze, Gestazione, Peso, Lunghezza e Cranio, oltre a variabili categoriali come Fumatrici, Sesso, Ospedale e Tipo di parto.

Andiamo ora a esaminare più da vicino la variabile Anni.madre, che è una variabile quantitativa continua. Considerando il contesto dello studio, il valore minimo della variabile, 0, non ha senso. Allo stesso modo, un'età troppo bassa risulta improbabile per una madre.

Per affrontare questa problematica, procederemo ordinando i primi valori di Anni.madre in ordine crescente, al fine di osservare meglio la presenza di eventuali valori non plausibili.

```
var_cresc <- sort(neonati$Anni.madre)
head(var_cresc, 5)
```

```
[1]  0  1 13 14 14
```

Sono presenti solo due valori problematici, 0 e 1, nella variabile Anni.madre. Trattandosi di soli due casi su 2500 (circa lo 0.08% del totale), potremo escluderli senza problemi dal dataset o, in alternativa, sostituire i valori incriminati con la media. In questo caso procederemo con la sostituzione.

Una volta eseguita la sostituzione dei due valori, controlliamo gli indici per tutte le variabili non categoriali.

| | Anni.madre | Cranio | Fumatrici | Gestazione | Lunghezza | N.gravidanze | Peso |
|---|---|---|---|---|---|---|---|
| Mean | 28.18614892 | 340.02922338 | 0.04163331 | 38.97958367 | 494.69575661 | 0.98158527 | 3284.18414732 |
| Std.Dev | 5.21720609 | 16.42946924 | 0.19978977 | 1.86895026 | 26.32884653 | 1.28094893 | 525.22937427 |
| Min | 13.00000000 | 235.00000000 | 0.00000000 | 25.00000000 | 310.00000000 | 0.00000000 | 830.00000000 |
| Q1 | 25.00000000 | 330.00000000 | 0.00000000 | 38.00000000 | 480.00000000 | 0.00000000 | 2990.00000000 |
| Median | 28.00000000 | 340.00000000 | 0.00000000 | 39.00000000 | 500.00000000 | 1.00000000 | 3300.00000000 |
| Q3 | 32.00000000 | 350.00000000 | 0.00000000 | 40.00000000 | 510.00000000 | 1.00000000 | 3620.00000000 |
| Max | 46.00000000 | 390.00000000 | 1.00000000 | 43.00000000 | 565.00000000 | 12.00000000 | 4930.00000000 |
| MAD | 4.44780000 | 14.82600000 | 0.00000000 | 1.48260000 | 22.23900000 | 1.48260000 | 459.60600000 |
| IQR | 7.00000000 | 20.00000000 | 0.00000000 | 2.00000000 | 30.00000000 | 1.00000000 | 630.00000000 |
| CV | 0.18509822 | 0.04831782 | 4.79879667 | 0.04794690 | 0.05322230 | 1.30497978 | 0.15992690 |
| Skewness | 0.15097173 | -0.78461925 | 4.58665318 | -2.06389091 | -1.51366518 | 2.51190318 | -0.64701485 |
| SE.Skewness | 0.04898001 | 0.04898001 | 0.04898001 | 0.04898001 | 0.04898001 | 0.04898001 | 0.04898001 |
| Kurtosis | -0.10792299 | 2.94011162 | 19.04501197 | 8.24650594 | 6.47334115 | 10.97043487 | 2.02472766 |
| N.Valid | 2498.00000000 | 2498.00000000 | 2498.00000000 | 2498.00000000 | 2498.00000000 | 2498.00000000 | 2498.00000000 |
| Pct.Valid | 100.00000000 | 100.00000000 | 100.00000000 | 100.00000000 | 100.00000000 | 100.00000000 | 100.00000000 |

Il minimo di Anni.Madre è ora 13, un valore sensato per la natura della variabile. Studiando Skewness e Curtosi possiamo assumere che:

- Per Anni.madre, la skewness vicina allo zero suggerisce una distribuzione piuttosto asimmetrica, mentre la curtosi leggermente inferiore allo zero indica code più leggere rispetto a una distribuzione normale. Il coefficiente di variazione è piuttosto basso, il che indica una bassa variabilità dei dati rispetto alla media.
- Per la variabile Cranio, la skewness leggermente negativa suggerisce una distribuzione asimmetrica verso la coda sinistra, mentre la curtosi positiva indica code più pesanti rispetto a una distribuzione normale. Il coefficiente di variazione è estremamente basso.
- Per Gestazione, la skewness piuttosto negativa indica una distribuzione asimmetrica con coda più lunga verso sinistra, mentre la curtosi positiva suggerisce code più pesanti rispetto a una distribuzione normale. Anche in questo caso, il coefficiente di variazione è basso, indicando una bassa variabilità dei dati rispetto alla media.
- Per Lunghezza, la skewness negativa suggerisce una distribuzione asimmetrica con coda più lunga verso sinistra, mentre la curtosi positiva indica code più pesanti.
- Per N.Gravidanze, la skewness estremamente positiva indica una distribuzione fortemente asimmetrica con una coda molto lunga verso destra, mentre la curtosi molto alta suggerisce code più pesanti rispetto a una distribuzione normale. Inoltre, il coefficiente di variazione alto indica una grande variabilità dei dati.
- Per quanto riguarda la variabile Peso, la skewness leggermente positiva suggerisce una distribuzione leggermente asimmetrica verso sinistra, mentre la curtosi positiva indica code più pesanti rispetto a una distribuzione normale. Il coefficiente di variazione indica una variabilità bassa rispetto alla media."

Passiamo alle tabelle di frequenza delle variabili Fumatrici, Tipo.parto, Ospedale e Sesso, che sono state trasformate in proporzioni percentuali.

```
table(Fumatrici)
freq_fum <- prop.table(table(Fumatrici))*100
freq_fum
table(Tipo.parto)
freq_parto <- prop.table(table(Tipo.parto))*100
freq_parto
table(Ospedale)
freq_osp <- prop.table(table(Ospedale))*100
freq_osp
table(Sesso)
freq_sesso <- prop.table(table(Sesso))*100
freq_sesso
```

Le non fumatrici (0) compongono la quasi totalità del nostro campione, con le madri fumatrici(1) che si assestano ad appena il 4% del totale.

```
Fumatrici
   0    1
2394  104
```

```
Fumatrici
        0         1
95.836669  4.163331
```

I parti naturali (quasi il 71%) sono molto più comuni dei Cesarei, il 29% delle nostre osservazioni.

```
Tipo.parto
 Ces  Nat
 728 1770
```

```
Tipo.parto
     Ces      Nat
29.14331 70.85669
```

I parti osservati nel nostro studio si dividono equivalente tra le tre strutture ospedaliere interessate.

```
Ospedale
osp1 osp2 osp3
 816  848  834
```

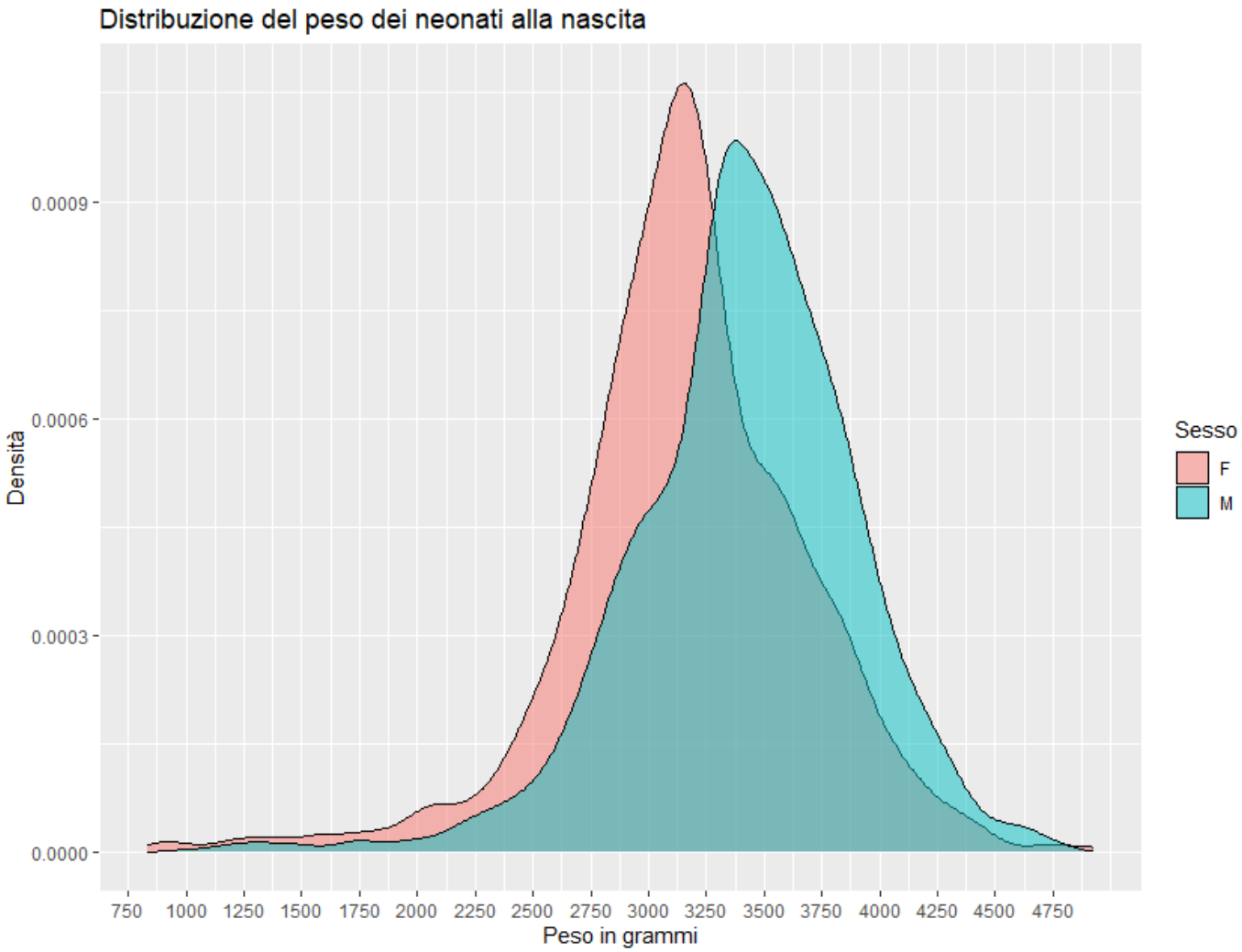
```
Ospedale
    osp1     osp2     osp3
32.66613 33.94716 33.38671
```

Scopriamo inoltre che non ci sono differenze significative nel numero di maschi e femmine.

```
Sesso
   F    M
1255 1243
```

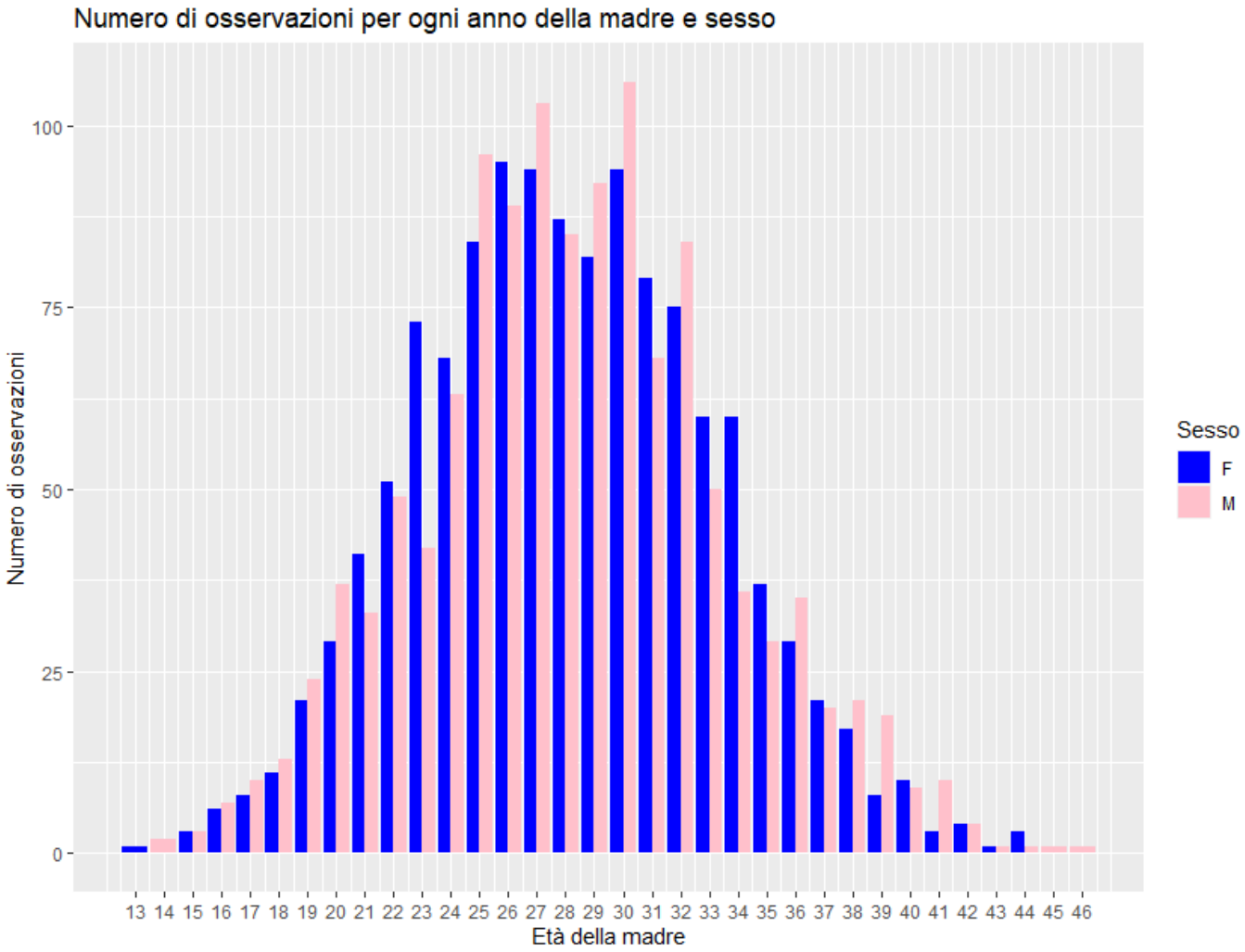
```
Sesso
       F        M
50.24019 49.75981
```

# Studio dei grafici




Nel grafico di densità del peso dei neonati per sesso, è evidente che il campione femminile presenta un valore di picco più elevato rispetto a quello maschile. Inoltre, la distribuzione del peso per le femmine appare più uniforme rispetto a quella dei maschi, che mostra una coda più pronunciata a destra. Questo suggerisce una maggiore concentrazione di osservazioni vicino al peso massimo nei neonati maschi.

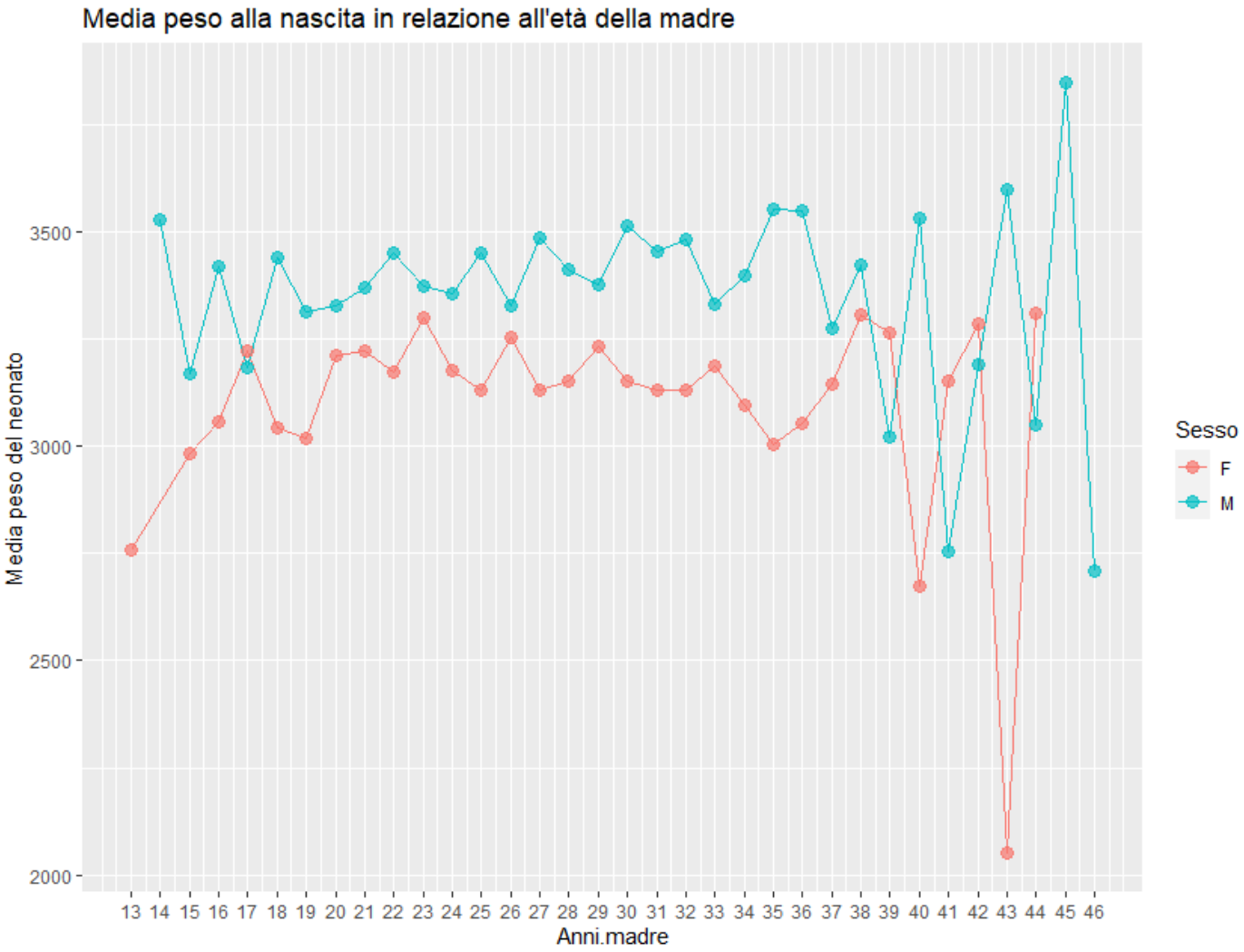
Inoltre, notiamo che il campione femminile mostra una distribuzione più snella nella parte centrale, il che indica una variabilità leggermente inferiore attorno al picco.




L'istogramma evidenzia un 'range' comune per l'età delle madri durante il parto, compreso tra i 25 ei 32 anni. Notiamo che a 33 anni si registra un picco negativo seguito da una diminuzione costante. Questo grafico fornisce indicazioni utili per comprendere i periodi di fertilità più comuni.

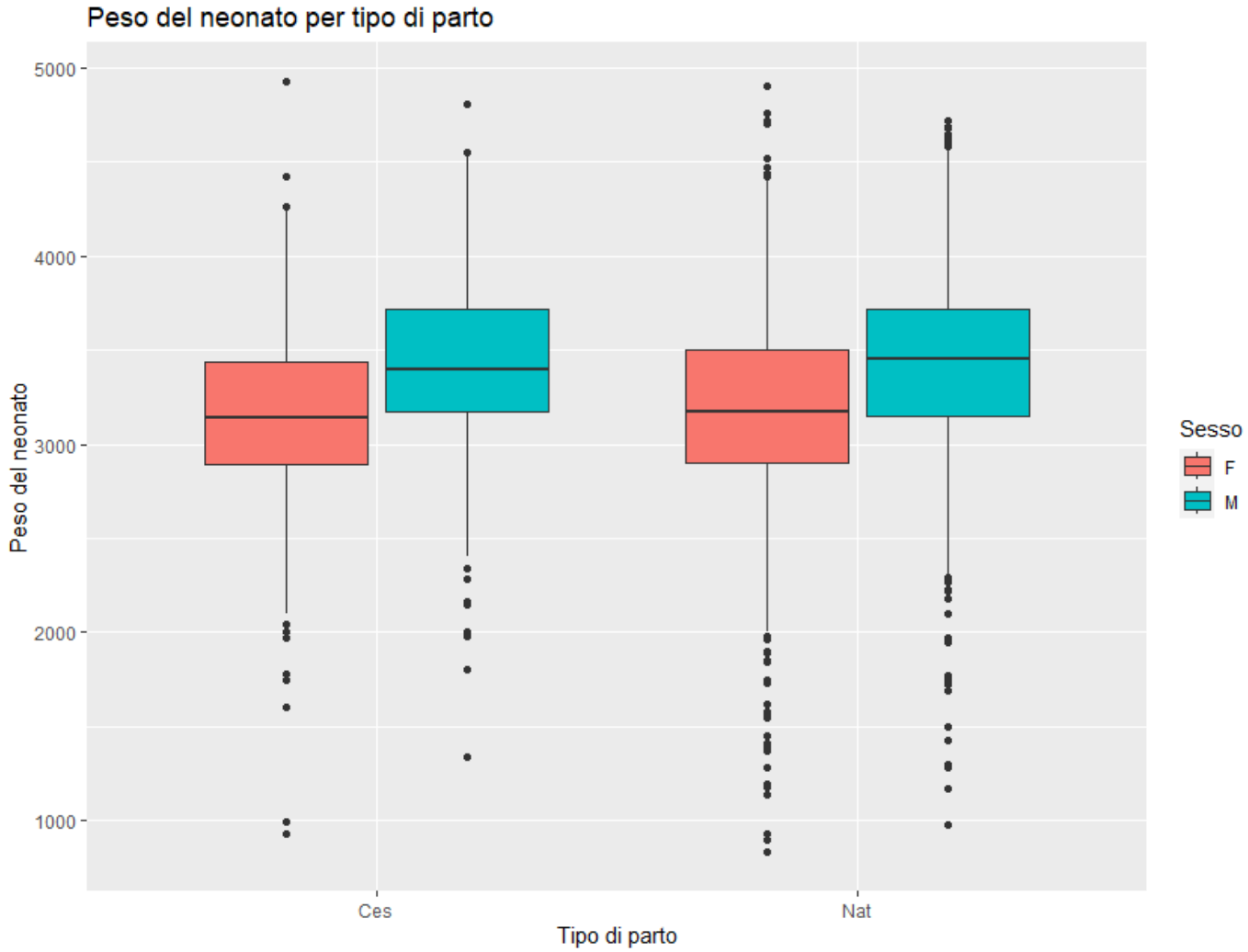
Non sembrano emergere differenze significative riguardo al sesso dei neonati. Sebbene siano presenti alcuni picchi isolati in entrambi i casi, come una lieve concentrazione di neonati femmine nelle madri trentenni e un picco di neonati maschi nelle madri di 23 anni, non si osservano trend particolari al di là della normalità.

È importante considerare anche il possibile impatto delle condizioni socioeconomiche della madre sul 'range' comune di età riscontrato. Sebbene al di fuori del nostro ambito di studio medico, questo fattore potrebbe influenzare significativamente i risultati e merita quindi una riflessione.




Nel grafico a linee che esamina il peso dei neonati in relazione all'età della madre e al sesso del bambino, abbiamo notato che il numero di neonati nati da madri sopra i 40 anni o sotto i 20 è limitato, il che influisce sulla visualizzazione del peso medio. Pertanto, per ottenere una visione più accurata, ci concentriamo sui valori con un numero maggiore di osservazioni.

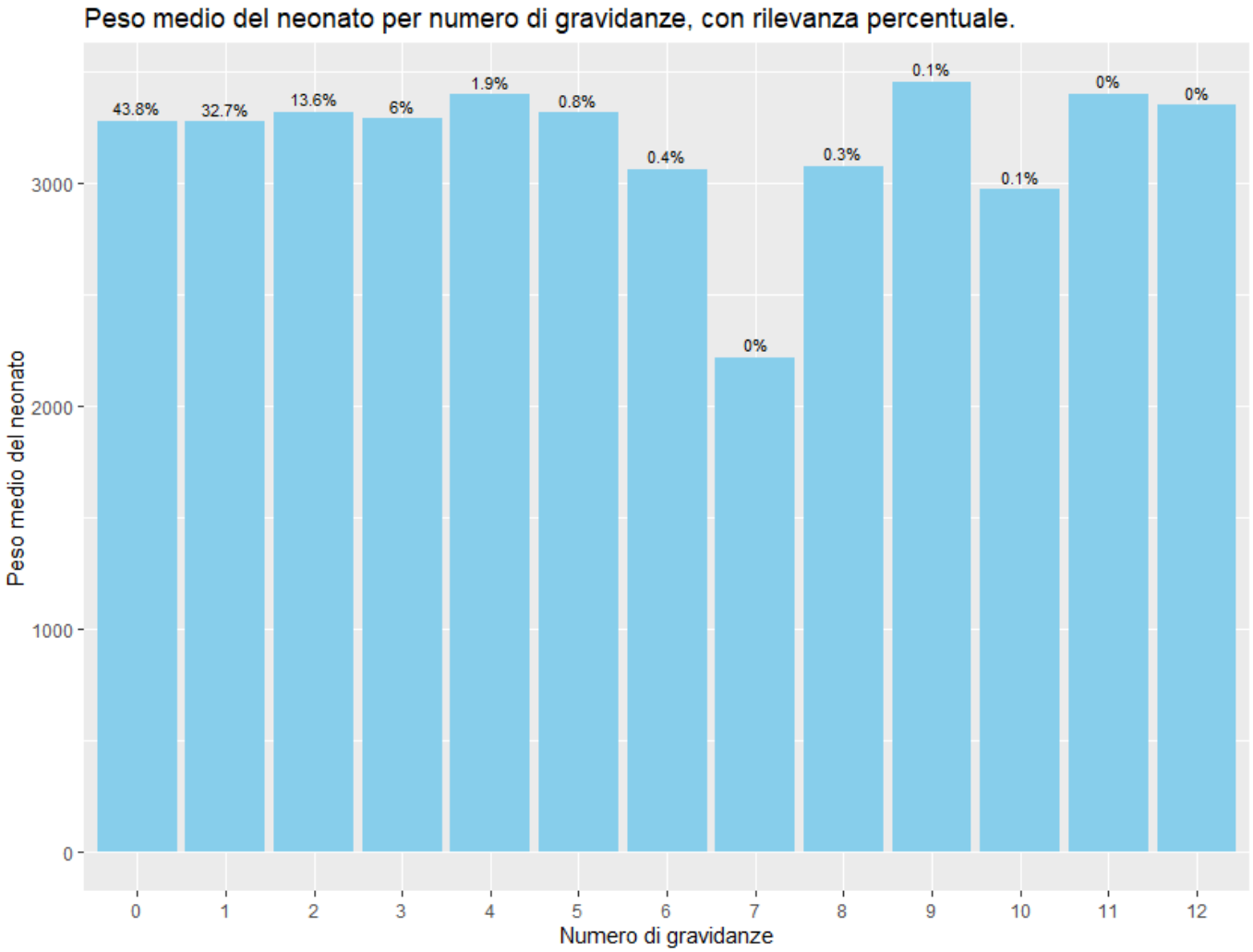
È evidente che i neonati maschi hanno un peso medio maggiore rispetto alle femmine, con una differenza di circa 500 grammi nel punto con la maggiore disparità tra i due sessi. Tuttavia, escludendo questo punto estremo, le variazioni nel peso tra i due sessi e tra le diverse fasce d'età sembrano avvicinarsi. Nella parte centrale del grafico, che contiene il maggior numero di osservazioni significative, notiamo che le variazioni di peso tra i maschi e le femmine si collocano entro un range di circa 250 grammi. In particolare, i maschi tendono ad avere un peso medio compreso tra i 3250 e i 3500 grammi, mentre per le femmine il peso medio oscilla tra i 3000 e i 3250 grammi.



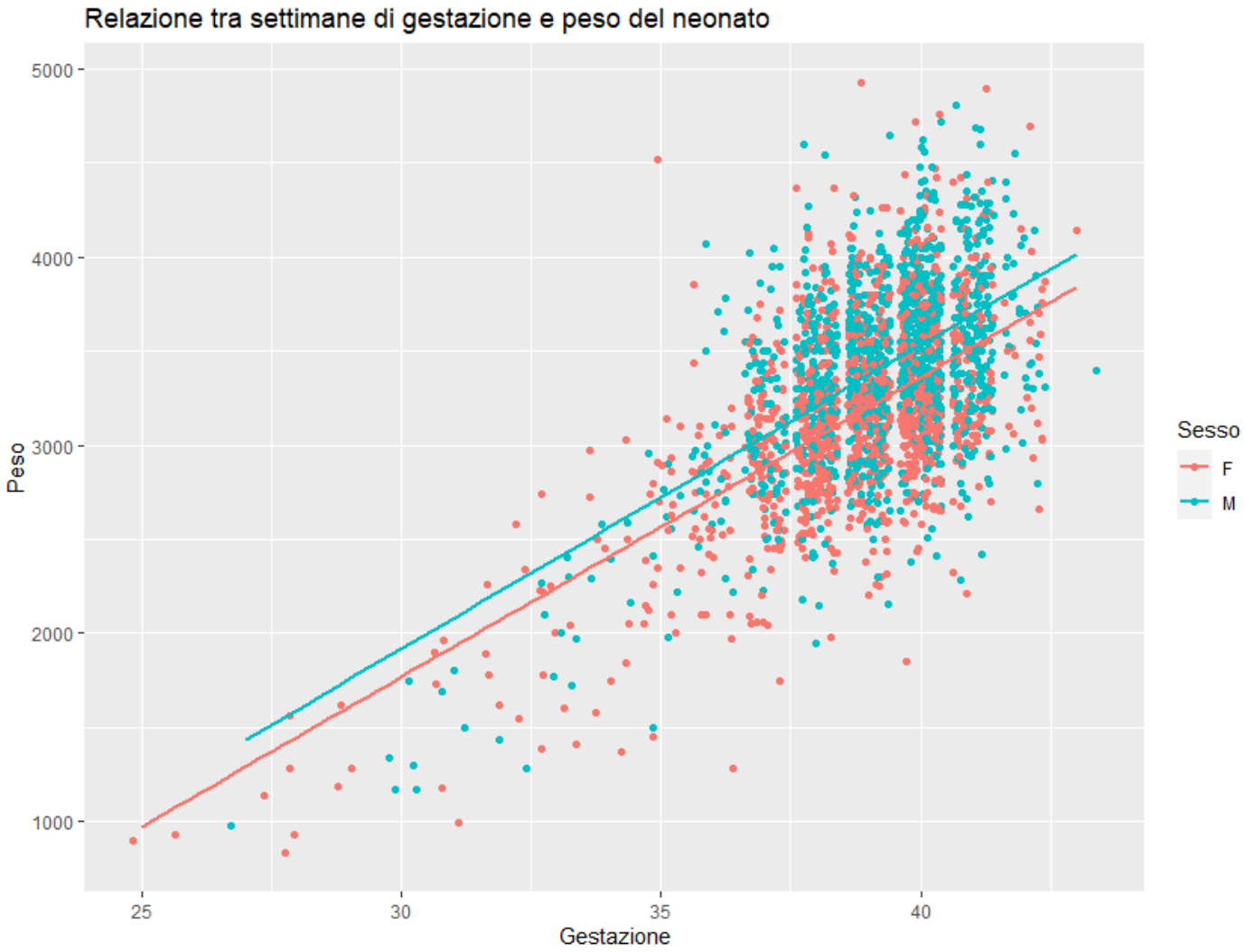

Il boxplot non mostra differenze significative nel peso in relazione al tipo di parto, suggerendo quindi l'assenza di correlazione tra la modalità di parto e il peso del neonato, così come per il sesso del nascituro. Le distribuzioni dei pesi sembrano essere bilanciate tra le varie modalità di parto.

Tuttavia, è interessante notare un dettaglio già osservato nel grafico precedente a linee: il range di differenza di peso in base al sesso, qui espresso in termini di range interquartile anziché di peso medio. Si può osservare che il 50% dei neonati di sesso femminile ha un peso compreso tra i 2900 e i 3400 grammi, mentre per i neonati maschi la variazione è compresa tra i 3200 e i 3700 grammi.

Un'ulteriore osservazione riguarda i valori outlier per il parto naturale, che sembrano essere più numerosi rispetto ai valori outlier dei parti cesarei.

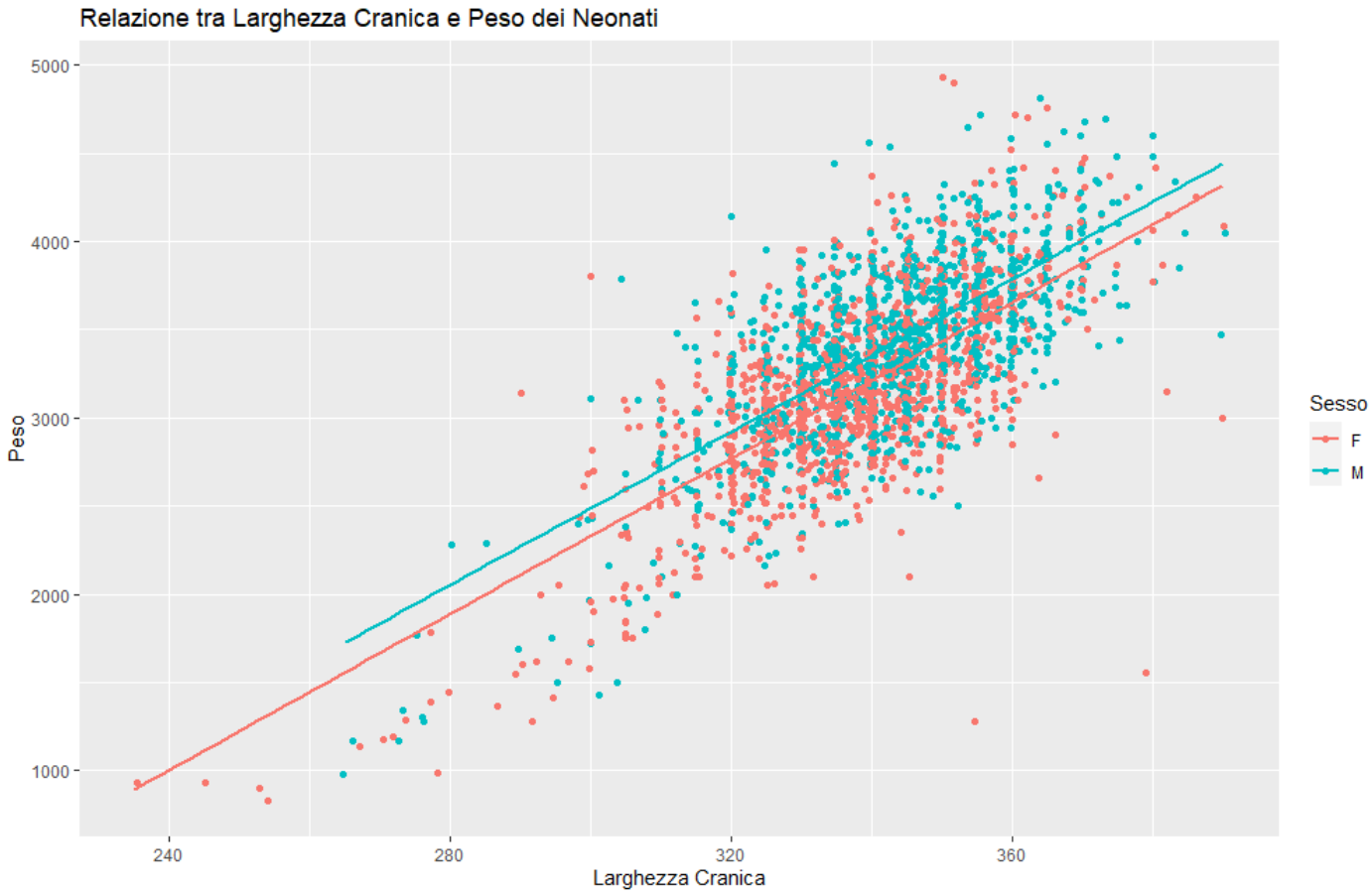



Nel grafico, stiamo esaminando se vi è una correlazione tra il numero di gravidanze e il peso medio del neonato. I valori sono generalmente simili, ad eccezione di quelli relativi a chi ha avuto sette gravidanze. Questa disparità può essere attribuita al numero limitato di osservazioni per questo gruppo, il che non consente un'analisi approfondita. Abbiamo incluso la percentuale di rilevanza di ciascun numero di gravidanze, indicando quanto frequentemente appare nel nostro studio e la sua importanza statistica. Notiamo immediatamente che le gravidanze 0, 1, 2 e 3 costituiscono circa il 96% del totale, e che tra di loro, le variazioni nel peso medio del neonato sono minime.



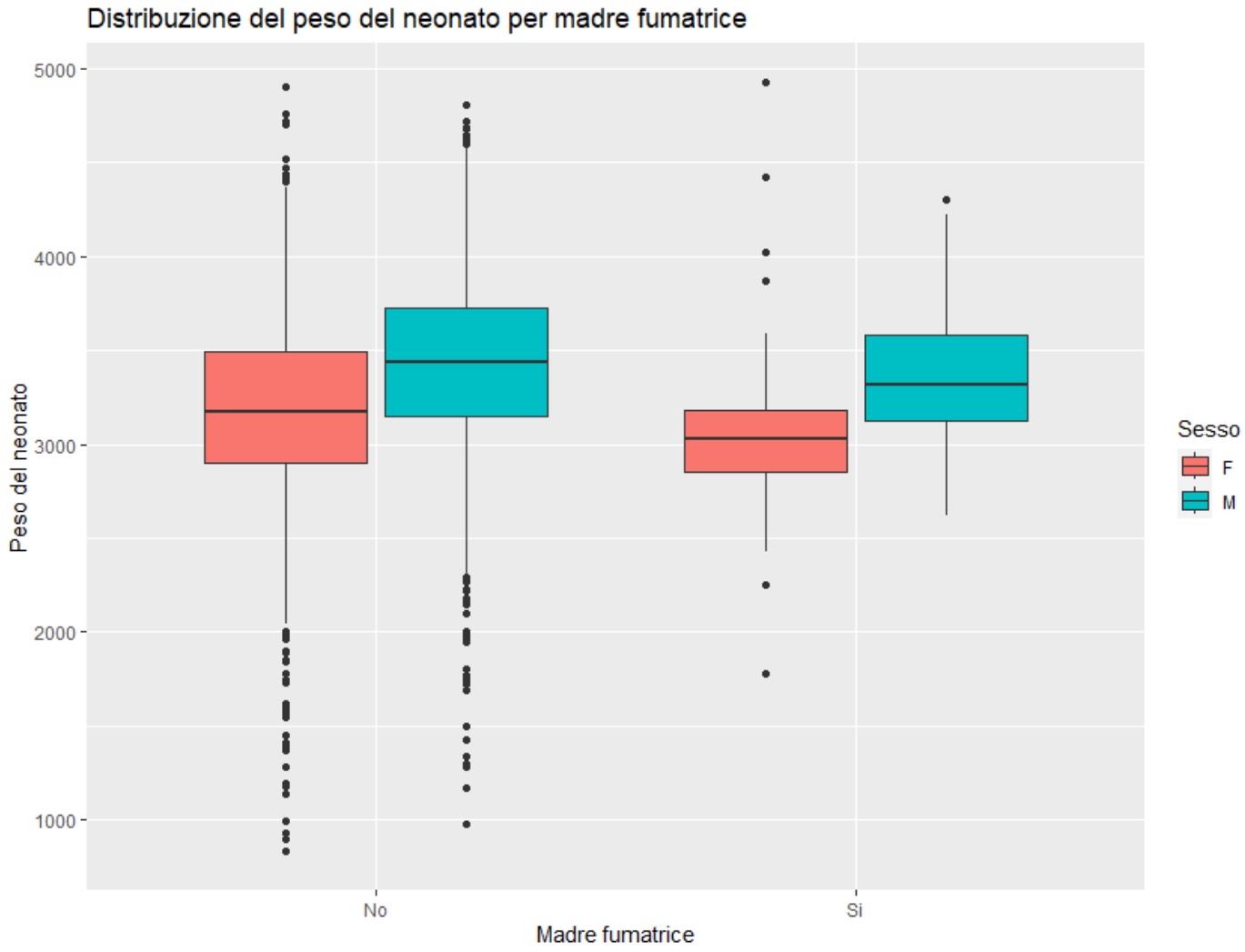

Nel grafico a dispersione che analizza l'influenza delle settimane di gestazione sul peso del neonato, diviso per sesso, emerge una correlazione positiva. Si osserva che il peso del neonato tende ad aumentare all'aumentare del periodo di gestazione, come evidenziato dalla concentrazione di molte osservazioni tra la 35ª e la 40ª settimana. Non sembrano emergere pattern particolari rispetto al sesso, con una distribuzione equa tra i due gruppi. L'unica particolarità, dovuta però a due singole osservazioni, è data dai valori minimi di gestazione in due casistiche femminili, che allungano la parte iniziale della retta F.

È importante notare che le gestazioni precoci mostrano un evidente calo del peso medio del neonato; la maggior parte delle osservazioni comprese tra la 25ª e la 35ª settimana si posiziona al di sotto della linea di regressione, indicando un peso inferiore rispetto al valore previsto.




Nel grafico a dispersione che analizza la correlazione tra la Larghezza Cranica e il peso del neonato, si osserva una chiara correlazione positiva. La larghezza del cranio tende ad aumentare all'aumentare del peso del neonato, come evidenziato dalla linea di regressione che si estende in un range compatto da circa 320 a 360 per la larghezza cranica e da circa 2500 a 4000 grammi per il peso.

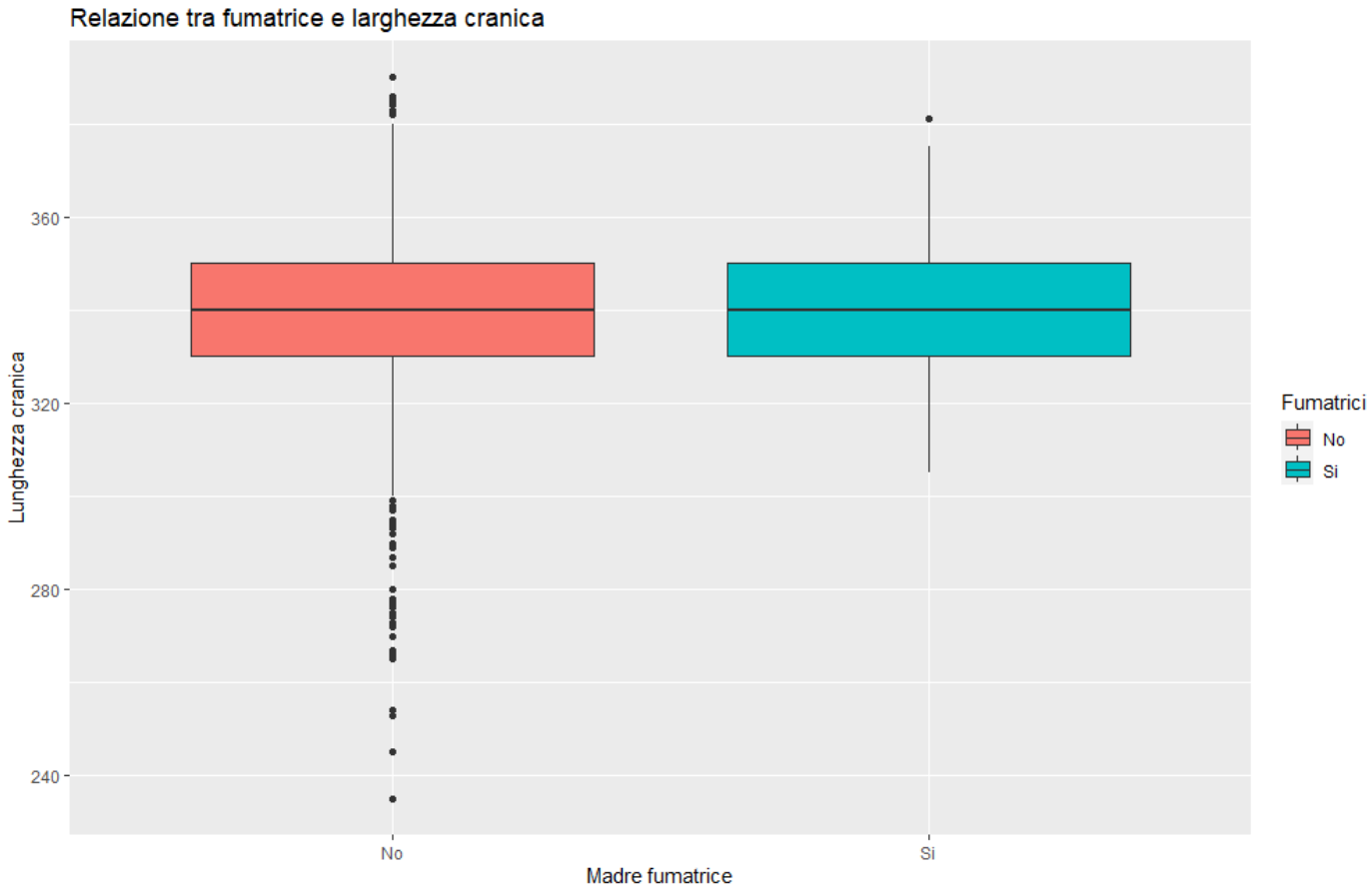
Nella parte sinistra del grafico, sotto le linee di regressione, sono presenti alcune osservazioni, indicando la presenza di neonati con peso inferiore alla media rispetto alla loro larghezza cranica. Inoltre, si nota che la linea di regressione per le femmine è più lunga sulla sinistra rispetto a quella dei maschi, suggerendo una maggiore variazione nella larghezza cranica per le femmine nei neonati con peso inferiore.



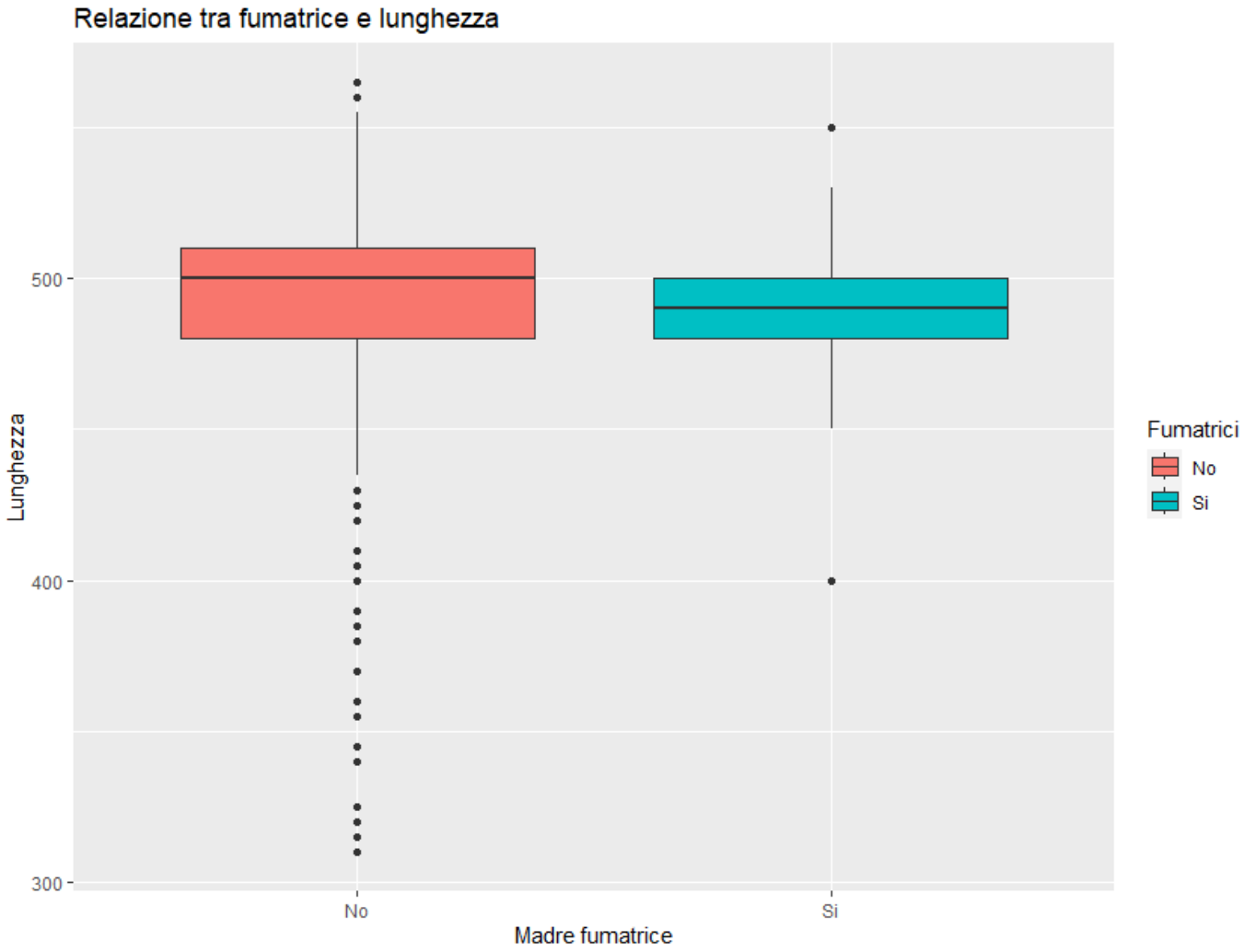

In questo boxplot, esaminiamo la relazione tra il peso del neonato, il sesso del neonato e lo status di fumatrice della madre. Le scatole, che rappresentano i dati nel range interquartile, sono significativamente più ampie per le madri non fumatrici rispetto a quelle delle madri fumatrici, indicando pesi mediamente più alti. Questa differenza è particolarmente evidente nel campione femminile.

Notiamo un gran numero di outlier nelle madri fumatrici, il che potrebbe essere attribuito alla loro minor frequenza nel campione di studio. Infatti, le madri non fumatrici costituiscono il 96% del campione, mentre le fumatrici solo il 4%. Di conseguenza, vi sono meno outlier nelle madri fumatrici.

Infine, possiamo dedurre che, data la lunghezza dei baffi nei box per le madri non fumatrici, i dati seguono una distribuzione normale. Sebbene vi sia una leggera differenza per le madri fumatrici, essa non è così significativa.

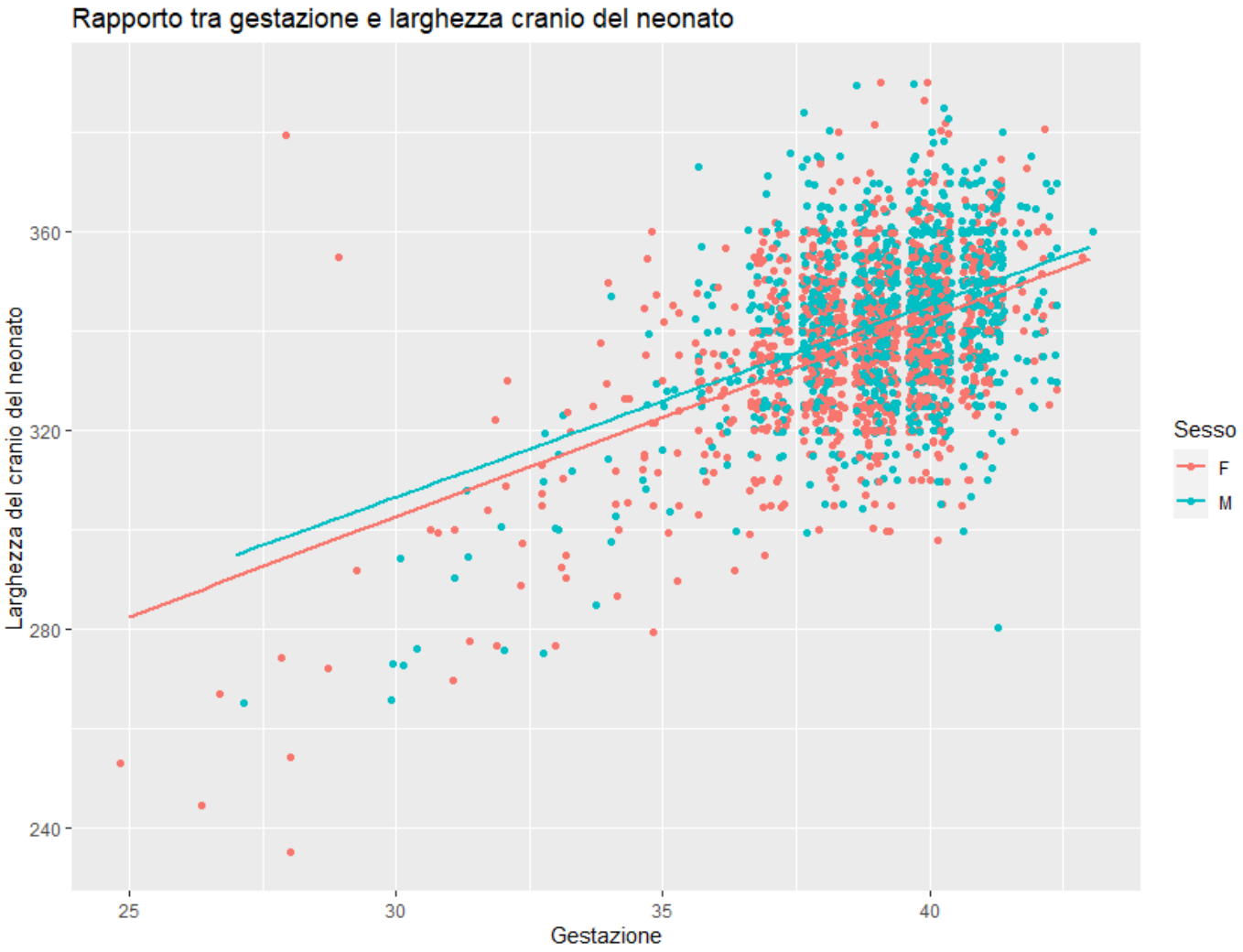



Non sembrano emergere differenze sostanziali tra le larghezze craniche dei neonati con madre fumatrice e non fumatrice. L'unica particolarità degna di nota è la presenza di numerosi outlier per le non fumatrici. Questo fenomeno potrebbe essere attribuibile alla varietà e alla quantità dei dati esaminati.



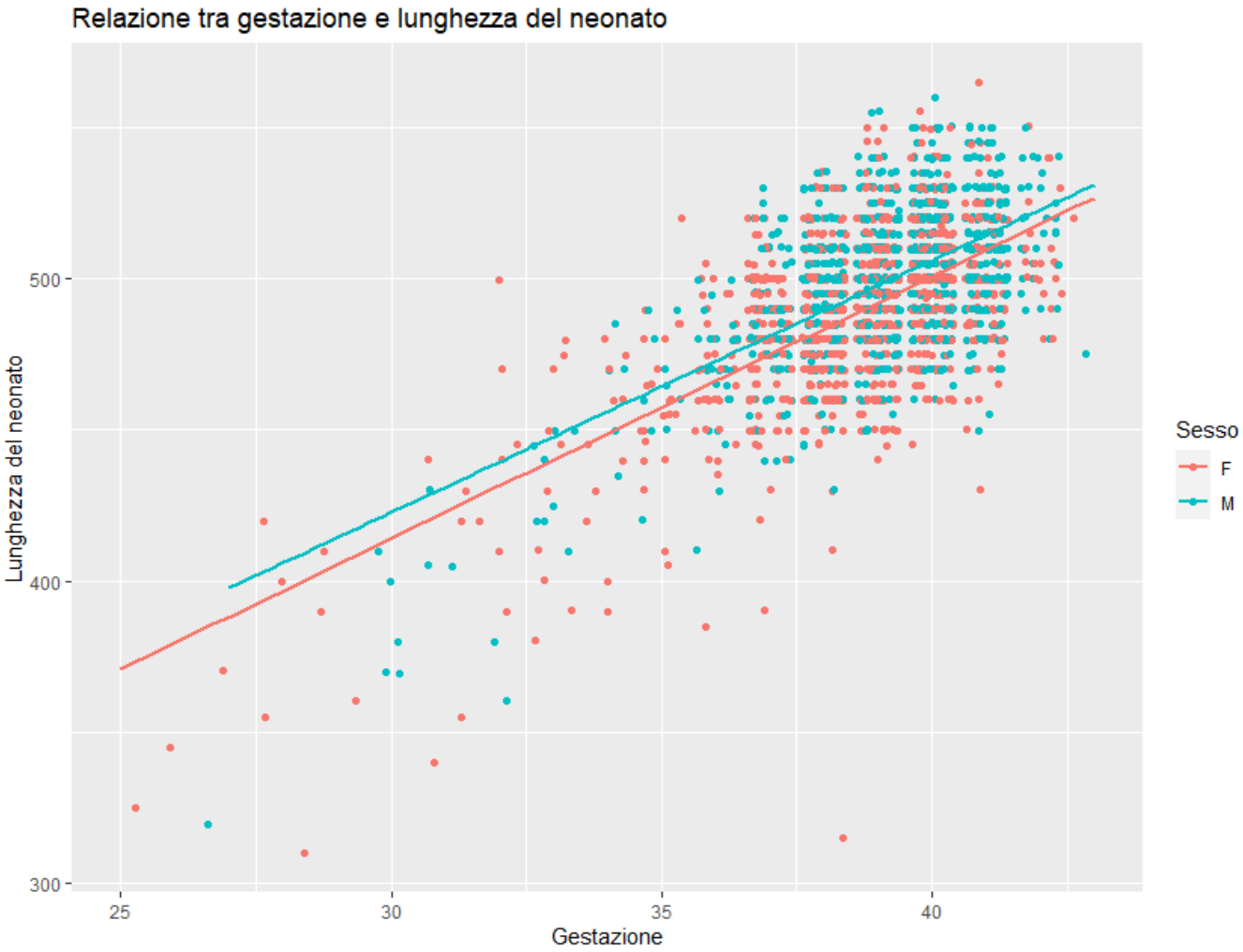

Nel boxplot relativo alle madri non fumatrici, la scatola risulta ampia quasi il doppio rispetto a quella delle fumatrici, con una mediana discretamente più alta. Tuttavia, la mediana appare leggermente sbilanciata all'interno del range interquartile, indicando una distribuzione asimmetrica negativa. Ciò suggerisce che la maggior parte delle osservazioni si concentra verso valori più alti. Al contrario, nel boxplot delle madri fumatrici, la scatola sembra indicare una distribuzione più simmetrica, sebbene con valori mediamente inferiori.

Come osservato in precedenza in altri boxplot, gli outlier sono presenti principalmente nelle osservazioni delle madri non fumatrici, in particolare concentrati nella parte bassa del grafico.




Notiamo immediatamente una correlazione positiva tra le settimane di gestazione e la larghezza del cranio del neonato. La maggior parte delle osservazioni si concentra nel range tra le 35 e 40 settimane di gestazione, con larghezza del cranio compresa tra i 320 e i 360 mm.

Nella parte sinistra del grafico, sotto le linee di regressione, sono presenti alcune importanti osservazioni. Questi dati evidenziano come, con il diminuire delle settimane di gestazione, diminuisca anche la correlazione con la larghezza del cranio. Questo suggerisce che nei neonati nati prematuramente potrebbe essere presente una variazione nella relazione tra le settimane di gestazione e le dimensioni del cranio, il che potrebbe avere implicazioni significative per la loro salute e sviluppo.




Anche in questa casistica, osserviamo una correlazione positiva tra le settimane di gestazione e la lunghezza del neonato alla nascita. La maggior parte delle osservazioni si concentra nel range tra le 35 e le 40 settimane di gestazione, con una lunghezza del neonato compresa tra i 450 e i 550 mm.

È importante soffermarci su una particolarità già osservata in precedenza: con il diminuire delle settimane di gestazione, diminuisce anche la correlazione con la lunghezza del neonato. Questo fenomeno potrebbe avere implicazioni significative per la salute e lo sviluppo dei neonati nati prematuramente, sottolineando l'importanza di considerare attentamente la prematurità nelle valutazioni della crescita neonatale.

# Ipotesi varie

Andiamo a saggiare alcuni ipotesi, ad esempio:

- La media del peso e della lunghezza di questo campione di neonati è significativamente uguale a quelle della popolazione?

Iniziamo scegliendo il test adatto, usando uno Shapiro-test per saggiare la normalità delle distribuzioni di peso.

```
> shapiro.test(Peso)

        Shapiro-Wilk normality test

data:  Peso
W = 0.97068, p-value < 2.2e-16

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

        Shapiro-Wilk normality test

data:  (Peso[Sesso == "F"])
W = 0.96293, p-value < 2.2e-16

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

        Shapiro-Wilk normality test

data:  (Peso[Sesso == "M"])
W = 0.96637, p-value = 2.225e-16
```

Con un valore di W molto vicino all'uno ma un p-value estremamente basso (< 2.2e-16) rigettiamo l'ipotesi nulla che i dati siano distribuiti normalmente. Pertanto, useremo un test Wilcoxon.

Iniziamo salvando come variabili le medie di nostro interesse riguardanti la Popolazione, prendendo il dato da studi clinici presi dal web. Visto che le medie di peso tra maschi e femmine differiscono leggermente, al contrario della lunghezza, faremo oltre al test unico anche dei test per sesso.

```
        Wilcoxon signed rank test with continuity correction

data:  Peso
V = 1162385, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 3400
```

In questo caso il p-value estremamente basso ci porta a rifiutare l'ipotesi nulla, e concludere che esiste una differenza statistica significativa tra il peso medio del nostro campione e quello della popolazione.

```
        Wilcoxon signed rank test with continuity correction

data:  Peso[Sesso == "F"]
V = 267680, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 3300
```

Come nel caso precedente, siamo portati a rifiutare l'ipotesi nulla, e concludere che esista una significativa differenza statistica tra le tue medie.

```
        Wilcoxon signed rank test with continuity correction

data:  Peso[Sesso == "M"]
V = 355593, p-value = 0.1086
alternative hypothesis: true location is not equal to 3450
```

Il test sul peso medio dei maschi ci porta un valore di p-value superiore al livello di significatività dello 0.05%, pertanto non abbiamo prove sufficienti per rigettare l'ipotesi nulla. Non vi è una differenza statisticamente significativa tra la media del peso per i maschi nel campione e la media della popolazione usata come confronto.

Proseguiamo con la variabile Lunghezza, e degli Shapiro-test per saggiarne la normalità.

```
> shapiro.test(Lunghezza)

        Shapiro-Wilk normality test

data:  Lunghezza
W = 0.90944, p-value < 2.2e-16

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

        Shapiro-Wilk normality test

data:  Lunghezza[Sesso == "F"]
W = 0.8996, p-value < 2.2e-16

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

        Shapiro-Wilk normality test

data:  Lunghezza[Sesso == "M"]
W = 0.92026, p-value < 2.2e-16
```

Dei p-value così bassi ci portano a rigettare l'ipotesi di normalità della distribuzione. Perciò, useremo dei Wilcoxon test.

```
        Wilcoxon signed rank test with continuity correction

data:  Lunghezza
V = 875939, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 500
```

Le due medie, Lunghezza del campione e Lunghezza dello Popolazione, presentano una differenza statisticamente significativa.

```
        Wilcoxon signed rank test with continuity correction

data:  Lunghezza[Sesso == "F"]
V = 160462, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 500
```

Anche per la media inerente alla Lunghezza dei Neonati femmine possiamo notare, visto il p-value molto basso, una differenza statisticamente significativa. Per ultimi andiamo a studiare i Neonati maschi.

```
        Wilcoxon signed rank test with continuity correction

data:  Lunghezza[Sesso == "M"]
V = 280003, p-value = 0.1414
alternative hypothesis: true location is not equal to 500
```

In questo caso non notiamo evidenze statistiche che ci portino a pensare ad una differenza significativa tra la media del campione e il valore medio della popolazione.

Terminiamo studiando la variabile Cranio, eseguendo prima degli Shapiro test per saggiarne la normalità.

```
> shapiro.test(Cranio)

        Shapiro-Wilk normality test

data:  Cranio
W = 0.96358, p-value < 2.2e-16

> shapiro.test(Cranio[Sesso=="F"])

        Shapiro-Wilk normality test

data:  Cranio[Sesso == "F"]
W = 0.95547, p-value < 2.2e-16

> shapiro.test(Cranio[Sesso=="M"])

        Shapiro-Wilk normality test

data:  Cranio[Sesso == "M"]
W = 0.97038, p-value = 2.901e-15
```

I p-value estremamente bassi ci portano a rifiutare l'ipotesi di normalità. Anche in questo caso useremo il Wilcoxon test.

```
> wilcox_cranio

        Wilcoxon signed rank test with continuity correction

data:  Cranio
V = 469692, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 350

> wilcox_cranio_femmine

        Wilcoxon signed rank test with continuity correction

data:  Cranio[Sesso == "F"]
V = 83761, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 350

> wilcox_cranio_maschi

        Wilcoxon signed rank test with continuity correction

data:  Cranio[Sesso == "M"]
V = 154592, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 350
```

In tutti e tre i casi abbiamo dei valori di p-value inferiori al livello di significatività dello 0.05%, pertanto possiamo rigettare l'ipotesi nulla. È' presente una differenza statisticamente significativa tra la media del peso per i maschi nel campione e la media della popolazione usata come confronto.

Ora verifichiamo l'ipotesi secondo cui in alcuni ospedali vengono eseguiti più parti cesarei. Creiamo una tabella e usiamo il test di Pearson del Chi-quadro.

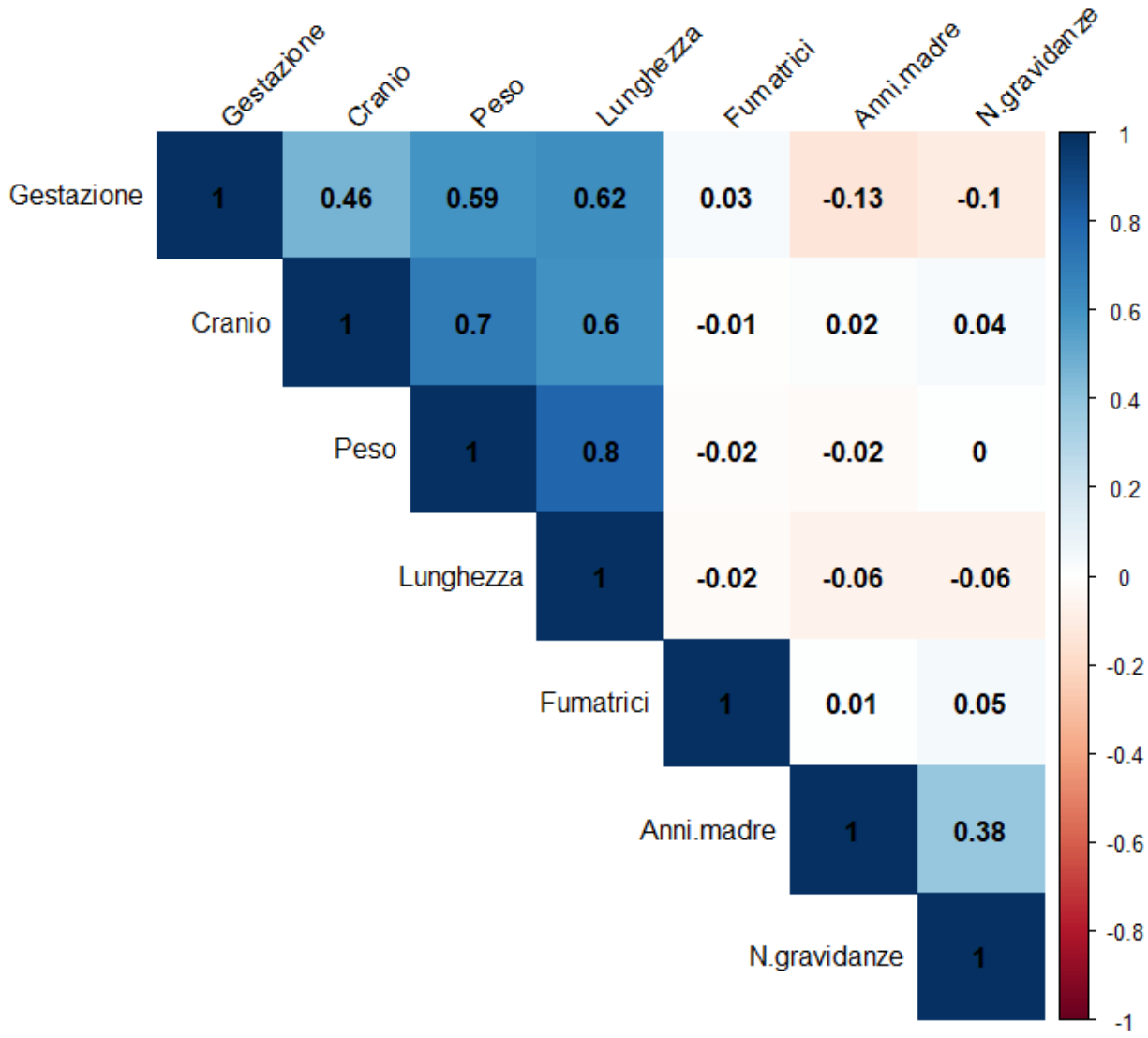
| | Tipo.parto | osp1 | osp2 | osp3 |
|---|---|---|---|---|
| 1 | Ces | 242 | 254 | 232 |
| 2 | Nat | 574 | 594 | 602 |

```
> chi_square

        Pearson's Chi-squared test

data:  cont_table2[-1]
X-squared = 1.083, df = 2, p-value = 0.5819
```

Dato che il valore p è maggiore del livello di significatività del 5%, non possiamo rigettare l'ipotesi nulla. Non possiamo dunque asserire che vi sia un'associazione significativa tra il tipo di parto (naturale o cesareo) e l'ospedale in cui avviene il parto.

# Analisi multidimensionale

Indaghiamo le relazioni tra le variabili, concentrandoci soprattutto sulla variabile Peso. Per far ciò useremo una matrice di correlazione, una heatmap.




Una delle correlazioni più forti è quella tra Peso e Gestazione, che ci indica come, all'aumentare del periodo di gestazione, tenda ad aumentare il peso del neonato.

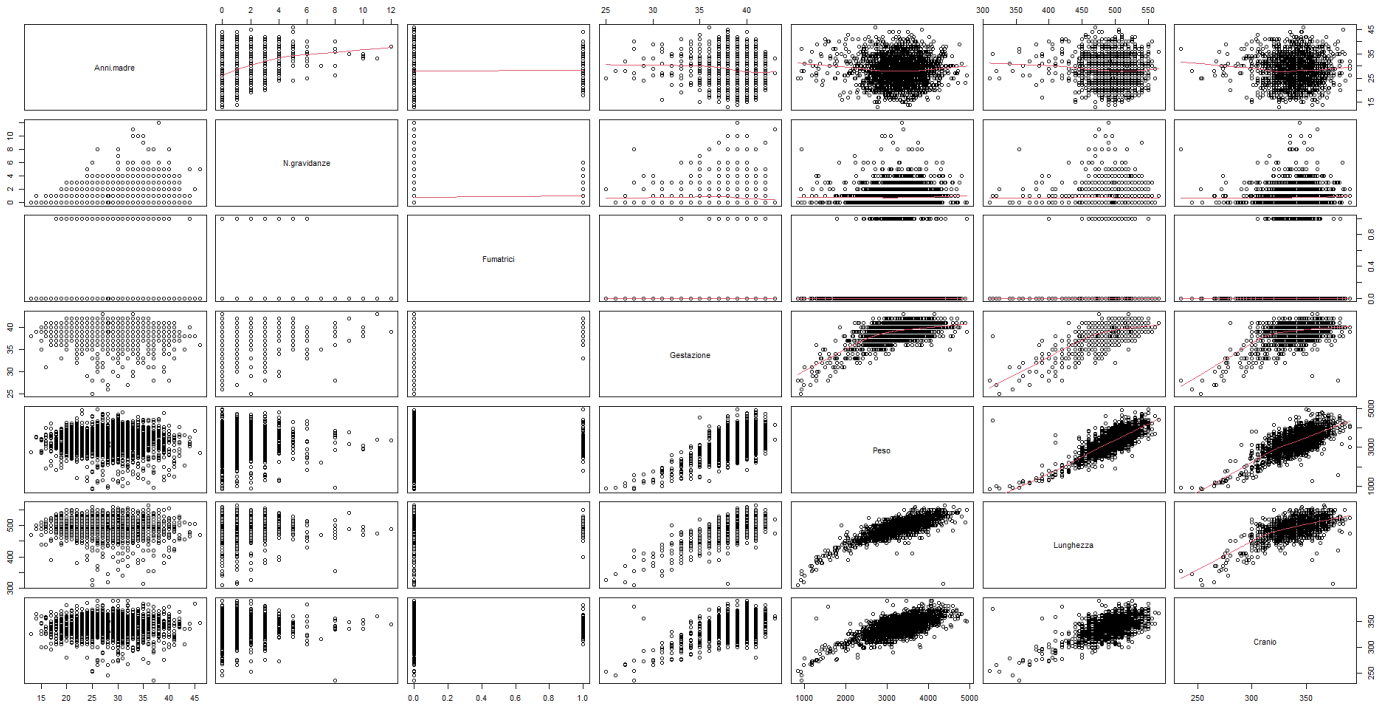
Anche tra Gestazione e Lunghezza vi è una correlazione positiva molto simile, con neonati nati dopo gestazioni più lunghe che dimostrano una lunghezza maggiore.

Peso e Lunghezza sono anch'essi fortemente correlati, dato che i neonati più pesanti tendono ad essere più lunghi. Questo assunto è presente anche nella correlazione tra Peso e Cranio, fortemente positiva: neonati più grandi tendono ovviamente ad avere crani più grandi.

Sembra che non vi sia praticamente alcuna correlazione, positiva o negativa, tra Fumatrici e le altre variabili. Questo ci suggerisce l'assenza di evidenze dirette dell'impatto del fumo sulla lunghezza, il peso, il periodo di gestazione o il diametro del cranio del neonato.

Quasi lo stesso può dirsi per la variabile Anni Madre, che mostra il coefficiente più alto con N.gravidanze, suggerendo una modesta relazione tra l'età della madre e il numero di gravidanze che ha avuto.

Sono inoltre presenti diverse correlazioni molto deboli, vicine allo zero. Una di interesse è quella tra Gestazione e Anni madre, che ci suggerisce come le madri più anziane possano avere tendenzialmente dei periodi di gestazione leggermente più brevi.




Vista la grandezza del grafico consiglio di visualizzarlo su R. Sono presenti alcuni effetti non lineari, soprattutto tra Lunghezza e Cranio, Gestazione e Lunghezza e Gestazione e Cranio.

Creiamo un modello di regressione con tutte le variabili.

```
mod1 <- lm(Peso ~ . , data=neonati)
summary(mod1)
```

```
> summary(mod1)

Call:
lm(formula = Peso ~ ., data = neonati)

Residuals:
    Min      1Q  Median      3Q     Max
-1123.3  -181.2   -14.6   160.7  2612.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6735.1400   141.3974 -47.633  < 2e-16 ***
Anni.madre      0.7975     1.1463   0.696   0.4867
N.gravidanze   11.4130     4.6665   2.446   0.0145 *
Fumatrici     -30.1567    27.5396  -1.095   0.2736
Gestazione     32.5262     3.8179   8.519  < 2e-16 ***
Lunghezza      10.2951     0.3007  34.237  < 2e-16 ***
Cranio         10.4725     0.4261  24.580  < 2e-16 ***
Tipo.partoNat  29.5025    12.0848   2.441   0.0147 *
Ospedaleosp2  -11.2217    13.4388  -0.835   0.4038
Ospedaleosp3   28.0985    13.4972   2.082   0.0375 *
SessoM         77.5473    11.1779   6.938 5.07e-12 ***
---
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: 669.1 on 10 and 2489 DF,  p-value: < 2.2e-16
```

Il modello presenta un buon R-quadro complessivo, ma ci sono diverse variabili che non contribuiscono significativamente con il modello, visto il loro p-value superiore o vicino alla soglia di 0.05. Procederemo quindi ad eliminare per step questi valori, e creare altrettanti modelli.

```
mod2 <- update(mod1,~.-Anni.madre)
summary(mod2)
anova(mod2,mod1)
BIC(mod2,mod1)
car::vif(mod2)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6708.1065   135.9394 -49.346  < 2e-16 ***
N.gravidanze     12.6085     4.3381   2.906  0.00369 **
Fumatrici       -30.3092    27.5359  -1.101  0.27113
Gestazione       32.2501     3.7968   8.494  < 2e-16 ***
Lunghezza        10.2944     0.3007  34.239  < 2e-16 ***
Cranio           10.4876     0.4255  24.651  < 2e-16 ***
Tipo.partoNat    29.5351    12.0834   2.444  0.01458 *
Ospedaleosp2    -11.0816    13.4359  -0.825  0.40957
Ospedaleosp3     28.3660    13.4903   2.103  0.03559 *
SessoM           77.6205    11.1763   6.945 4.81e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 273.9 on 2490 degrees of freedom
Multiple R-squared:  0.7288,    Adjusted R-squared:  0.7278
F-statistic: 743.6 on 9 and 2490 DF,  p-value: < 2.2e-16
```

Levando Anni.madre l’r quadro aggiustato non è variato, ma N.gravidanze sembra essere più significativo. Utilizziamo un test ANOVA per capire se la rimozione della variabile aiuta in modo significativo il modello.

```
> anova(mod2,mod1)
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Ospedale + Sesso
Model 2: Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza +
    Cranio + Tipo.parto + Ospedale + Sesso
  Res.Df       RSS Df Sum of Sq     F Pr(>F)
1   2490 186809099
2   2489 186772779  1     36321 0.484 0.4867
```

L’RSS non varia praticamente di nulla, mentre il valore F è piuttosto basso, il che indica che la variabile Anni.madre non contribuisce in modo significativo al modello. Il p-value superiore alla soglia di 0.05 ci suggerisce che l’aggiunta della variabile non migliora il modello predittivo.

Usiamo il modello BIC per mettere a confronto i due modelli.

```
> BIC(mod2,mod1)
     df      BIC
mod2 11 35234.64
mod1 12 35241.97
```

Con un BIC inferiore, il mod2 rimane preferibile al mod1 che, nonostante includa un parametro aggiuntivo, non migliora abbastanza il modello da giustificare l’aggiunta.

Infine, eseguiamo un test VIF, un indice che misura l'eventuale presenza di multicollinearità.

```
> car::vif(mod2)
                GVIF Df GVIF^(1/(2*Df))
N.gravidanze 1.027985  1        1.013896
Fumatrici    1.007346  1        1.003666
Gestazione   1.676688  1        1.294870
Lunghezza    2.085755  1        1.444214
Cranio       1.626661  1        1.275406
Tipo.parto   1.004240  1        1.002118
Ospedale     1.003421  2        1.000854
Sesso        1.040558  1        1.020077
```

Non ci sono VIF maggiori di 5, pertanto il test può dirsi superato.

Aggiorniamo il modello levando la variabile Ospedale.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6708.074    135.984 -49.330  < 2e-16 ***
N.gravidanze     13.012      4.342   2.997  0.00276 **
Fumatrici       -31.759     27.570  -1.152  0.24946
Gestazione       32.541      3.801   8.561  < 2e-16 ***
Lunghezza        10.272      0.301  34.129  < 2e-16 ***
Cranio           10.501      0.426  24.648  < 2e-16 ***
Tipo.partoNat    30.296     12.098   2.504  0.01234 *
SessoM           78.114     11.191   6.980 3.77e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274.3 on 2492 degrees of freedom
Multiple R-squared:  0.7278,    Adjusted R-squared:  0.7271
F-statistic:   952 on 7 and 2492 DF,  p-value: < 2.2e-16
```

L'r quadro è diminuito di pochissimo, e abbiamo un modello più semplice.

```
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Ospedale + Sesso
  Res.Df       RSS Df Sum of Sq      F   Pr(>F)
1   2492 187501837
2   2490 186809099  2    692738 4.6168 0.009969 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nonostante il p-value indichi che la variabile Ospedale sia significativa, abbiamo visto nella parte descrittiva del nostro progetto come il peso medio non vari in base alla struttura ospedaliera.

Creiamo un nuovo modello (mod4) estraendo la variabile Tipo.parto.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6681.6714   135.7178 -49.232  < 2e-16 ***
N.gravidanze    12.7185     4.3450   2.927  0.00345 **
Fumatrici      -30.4634    27.5948  -1.104  0.26972
Gestazione      32.5914     3.8051   8.565  < 2e-16 ***
Lunghezza       10.2341     0.3009  34.011  < 2e-16 ***
Cranio          10.5359     0.4262  24.718  < 2e-16 ***
SessoM          78.1713    11.2028   6.978 3.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274.6 on 2493 degrees of freedom
Multiple R-squared:  0.7271,    Adjusted R-squared:  0.7265
F-statistic:  1107 on 6 and 2493 DF,  p-value: < 2.2e-16
```

```
> anova(mod4,mod3)
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Sesso
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1   2493 187973654
2   2492 187501837  1    471817 6.2707 0.01234 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L’r quadro aggiustato è rimasto praticamente invariato, e il modello è più semplice. Inoltre, nella parte descrittiva del progetto, abbiamo potuto vedere graficamente come non vi fosse differenza nel peso medio dei neonati per via della tipologia di parto. Nonostante il test ANOVA indichi che la variabile Tipo.parto sia statisticamente significativa il risultato è molto più vicino alla soglia di 0.05 rispetto alle altre variabili; decido quindi di escluderla dal modello, preferendo continuare con uno più semplice e con variabili più solide dal punto di vista statistico.

Inoltre, vi è anche un leggero miglioramento nel BIC e nessuna multicollinearità tra le variabili.

```
> BIC(mod4,mod3,mod2,mod1)
     df      BIC
mod4  8 35226.70
mod3  9 35228.24
mod2 11 35234.64
mod1 12 35241.97
```

```
> car::vif(mod4)
N.gravidanze    Fumatrici   Gestazione    Lunghezza       Cranio        Sesso
    1.026120     1.006607     1.675575     2.078644     1.624603     1.040271
```

Nonostante Fumatrici sia una variabile non significativa per il momento è stata mantenuta come variabile di controllo. Proviamo ad escluderla, per capire se ciò porta ad un significativo miglioramento del modello.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6681.1445   135.7229 -49.226  < 2e-16 ***
N.gravidanze   12.4750     4.3396   2.875  0.00408 **
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 ***
---
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
```

```
> anova(mod4nofum, mod4)
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Sesso
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1   2494 188065546
2   2493 187973654  1     91892 1.2187 0.2697
> BIC(mod4nofum,mod4,mod3,mod2,mod1)
          df      BIC
mod4nofum  7 35220.10
mod4       8 35226.70
mod3       9 35228.24
mod2      11 35234.64
mod1      12 35241.97
```

L’r quadro aggiustato non è variato. Secondo il test di ANOVA la variabile Fumatrici non migliora in modo significativo la capacità del modello di predire il peso dei neonati. Inoltre, il BIC senza questa variabile è migliorato sensibilmente.

Ciononostante, decido di tenerla in quanto variabile di controllo trattandosi di uno studio clinico, e tenendo da conto anche i risultati ottenuti nei grafici boxplot di studio, nei quali era chiaro un peso medio minore per i neonati nati da madre fumatrice.

Procedo creando diversi modelli che considerino interazioni ed effetti non lineari. Per chiarezza del progetto non riporterò qui tutti i modelli, possono essere trovati nel file .R. Uno dei migliori risulta essere questo:

```
Call:
lm(formula = Peso ~ N.gravidanze + Fumatrici + Gestazione + Cranio +
    Lunghezza + I(Lunghezza^2) + Sesso, data = neonati)

Residuals:
     Min       1Q   Median       3Q      Max
-1170.35  -181.95   -11.83   162.98  1785.71

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    198.983315 723.822749   0.275 0.783411
N.gravidanze    14.285516   4.269659   3.346 0.000833 ***
Fumatrici      -23.907419  27.104975  -0.882 0.377845
Gestazione      42.685197   3.879372  11.003  < 2e-16 ***
Cranio          10.646792   0.418709  25.428  < 2e-16 ***
Lunghezza      -20.214842   3.162204  -6.393 1.94e-10 ***
I(Lunghezza^2)   0.031592   0.003267   9.671  < 2e-16 ***
SessoM          70.165729  11.031579   6.360 2.39e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 269.6 on 2492 degrees of freedom
Multiple R-squared:  0.737,     Adjusted R-squared:  0.7363
F-statistic: 997.6 on 7 and 2492 DF,  p-value: < 2.2e-16
```

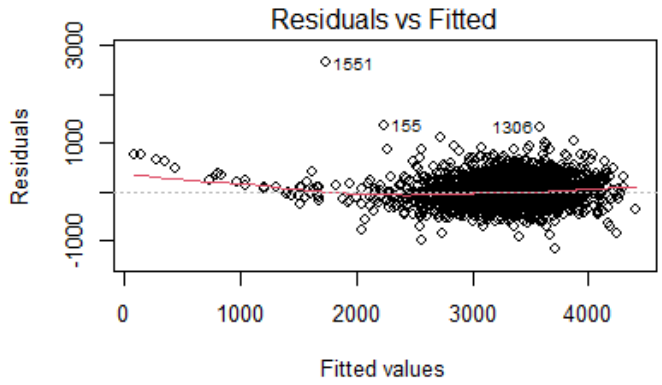
```
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Cranio + Lunghezza +
    I(Lunghezza^2) + Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Sesso
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   2492 181173497
2   2493 187973654 -1  -6800157 93.535 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

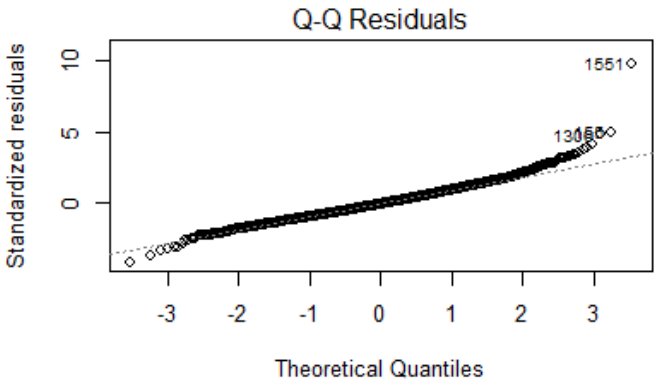
Il modello ha un r quadro aggiustato migliore del mod4, ma è anche più complesso. Con il test ANOVA vediamo come il modello più complesso fornisce una migliore adattabilità ai dati in modo statisticamente significativo.

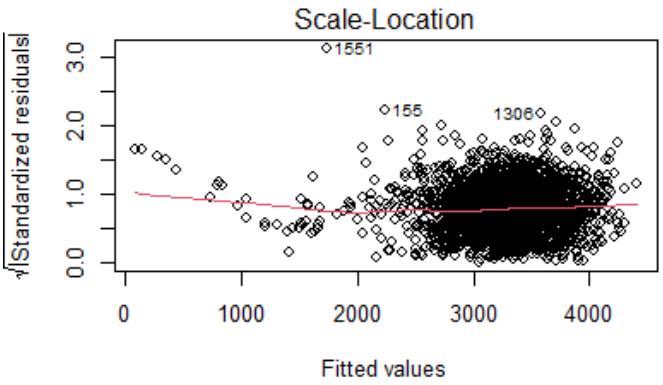
```
> BIC(mod_non_lin2,mod4)
             df      BIC
mod_non_lin2  9 35142.41
mod4          8 35226.70
> car::vif(mod_non_lin2)
  N.gravidanze      Fumatrici     Gestazione
      1.027600       1.007237       1.806315
        Cranio      Lunghezza I(Lunghezza^2)
      1.625823     238.081896     230.152216
         Sesso
      1.046162
```

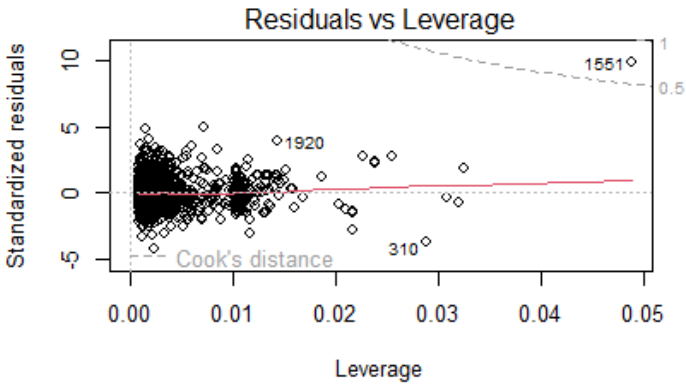
Il BIC del mod_non_lin2 è di certo migliore di quello del mod4, ma il test VIF ci mostra alcune problematiche. Lunghezza e il suo quadrato presentano una forte multicollinearità. Per questo preferiremo il mod4.

Analizziamo i residui.




E' presente una chiara eteroschedasticità, con la varianza dei residui non costante. Studiando il Q-Q notiamo una leggera coda positiva, ma per la maggior parte indica una distribuzione normale.

Anche con lo Scale-Location notiamo Eteroschedasticità. Infine per Residui vs Leverage possiamo notare l'osservazione 1551, oltre la soglia di 0.5 della distanza di cook.

```
        Shapiro-Wilk normality test

data:  residuals(mod4)
W = 0.9741, p-value < 2.2e-16
```

Secondo lo Shapiro-wilk rifiutiamo l'ipotesi nulla che i residui seguano una distribuzione normale.

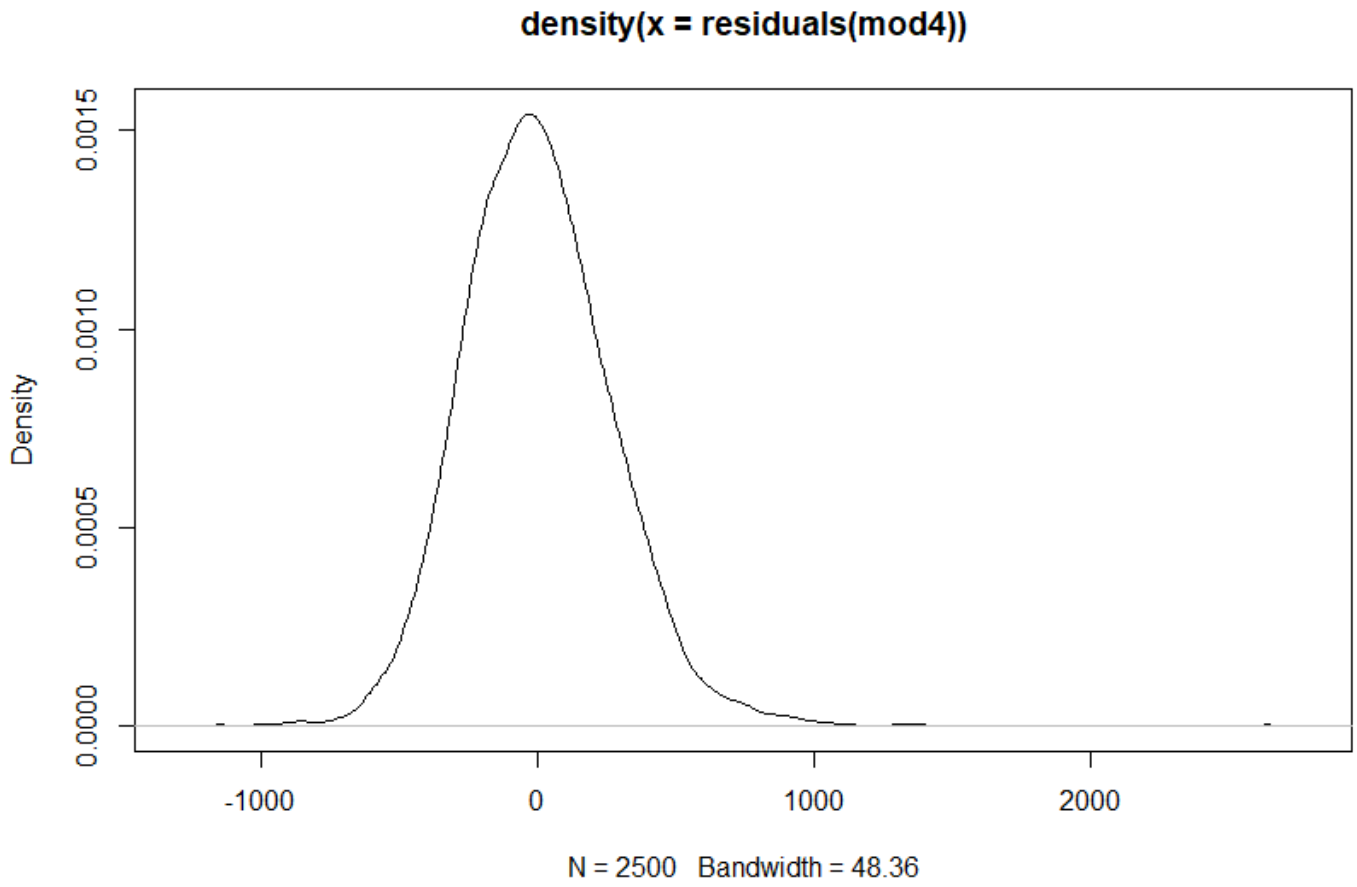
```
        studentized Breusch-Pagan test

data:  mod4
BP = 89.798, df = 6, p-value < 2.2e-16
```

Il test ci suggerisce una eteroschedasticità dei residui, per cui la varianza dei residui non è costante.

```
        Durbin-Watson test

data:  mod4
DW = 1.9542, p-value = 0.126
alternative hypothesis: true autocorrelation is greater than 0
```

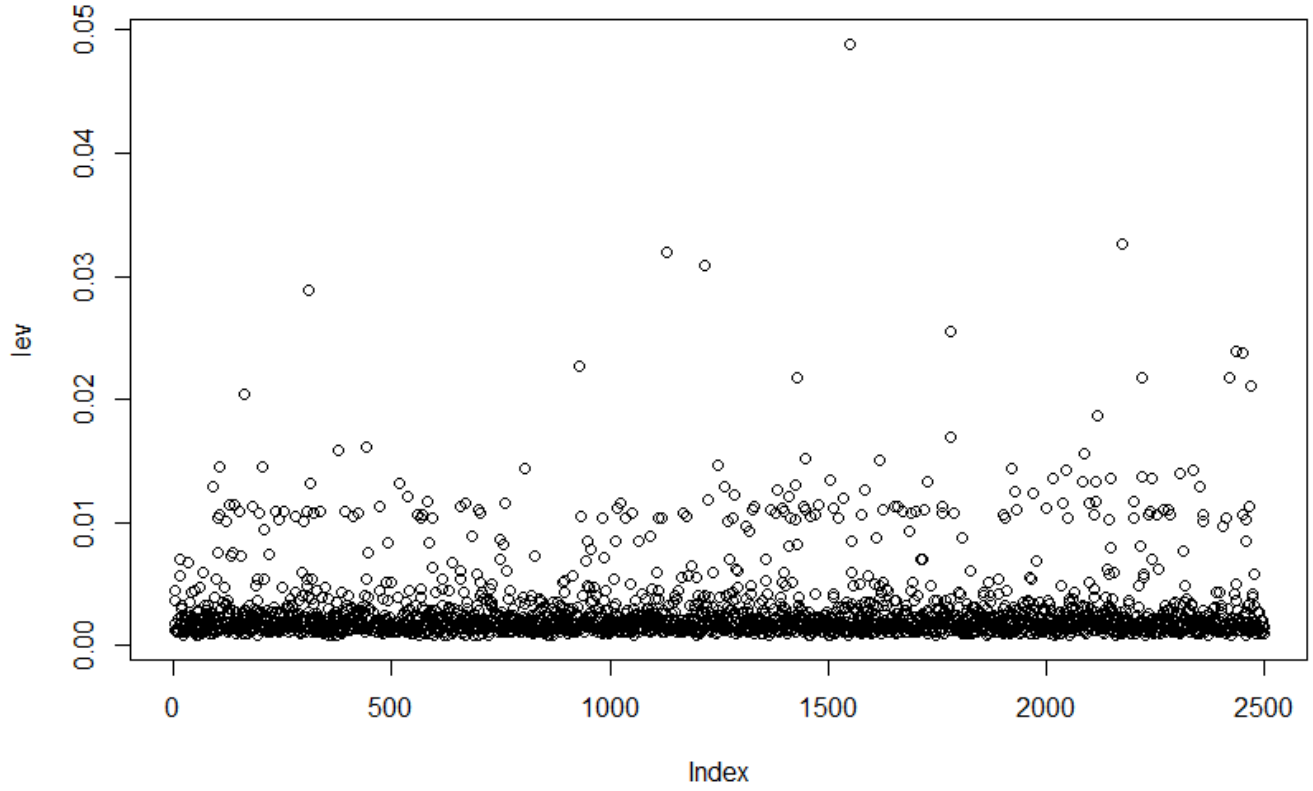
Il test di Durbin-Watson indica l'assenza di autocorrelazione tra i residui del modello.

Andiamo a studiare il density plot dei residui del modello.




Risulta in una distribuzione normale, anche se sono presenti delle code soprattutto nel lato destro del grafico di densità, allungato per una singola osservazione (la 1551).

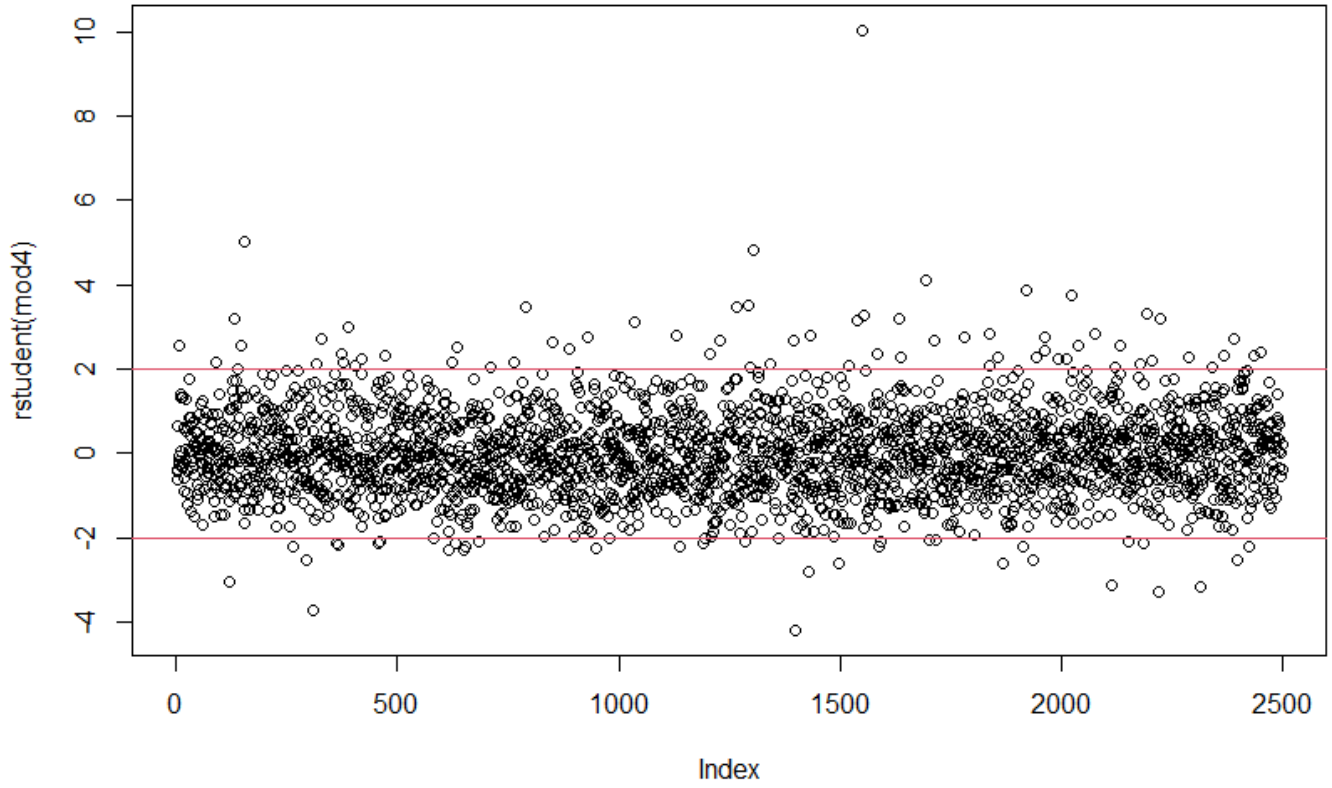
Andiamo a studiare i leverage.




Non sono presenti valori oltre la soglia, che peraltro non viene nemmeno visualizzata nel plot. Per conferma richiamiamo i leverage superiori alla soglia, che dovranno essere zero.

```
> lev[lev>soglia]
named numeric(0)
```

Passiamo agli outliers.




```
      rstudent unadjusted p-value Bonferroni p
1551 10.039719         2.8060e-23   7.0149e-20
155   5.022108         5.4723e-07   1.3681e-03
1306  4.823102         1.4986e-06   3.7465e-03
```

L'osservazione 1551 ha un residuo altissimo, significativamente diverso da zero, e un Bonferroni molto basso. È quindi un outlier.

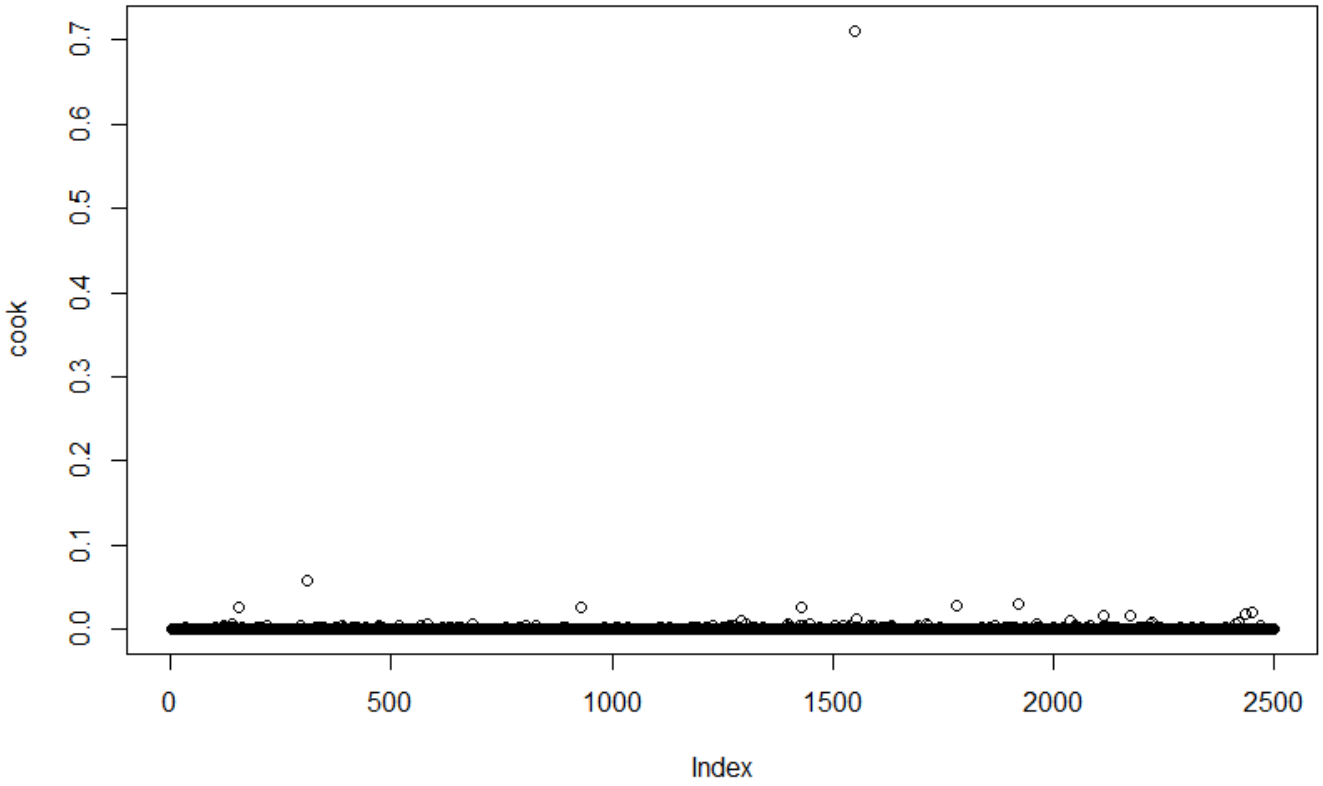
Anche per quanto riguarda le osservazioni 155 e 1306 abbiamo residui piuttosto alti, e risultano diversi da zero. Per quanto il Bonferroni non sia basso come per l'osservazione 1551 anch'essi possono essere classificati come outlier.

Studiamoli più nel dettaglio.

```
     Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio Tipo.parto Ospedale Sesso
1551         35            1         0         38 4370       315    374        Nat     osp3     F
155          30            0         0         36 3610       410    330        Nat     osp1     M
1306         23            0         0         41 4900       510    352        Nat     osp2     F
```

La 1306 mostra un peso di 4900 grammi e una lunghezza di 510mm, valori elevati per un neonato. Anche nelle altre due osservazioni questi valori risultano diversi dalla media, il che potrebbe essere il motivo per loro riconoscimento come outlier. Cranio non sembra molto lontana dai valori mediani, così come Anni.madre e N.gravidanze.

Studiamo anche la distanza di cook tramite un grafico, e richiamiamo il valore massimo.




La distanza massima è 0.7, alquanto elevata.

Studiamo singolarmente le osservazioni problematiche.

```
> print(righe_interessate)
     Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio Tipo.parto Ospedale Sesso
1551         35            1         0         38 4370       315    374        Nat     osp3     F
155          30            0         0         36 3610       410    330        Nat     osp1     M
1306         23            0         0         41 4900       510    352        Nat     osp2     F
310          40            3         0         28 1560       420    379        Nat     osp3     F
```

Per quanto presentino dei dati oltre la media non si tratta di valori fisiologicamente errati. Più avanti studieremo un nuovo modello senza l'osservazione 1551 e l'osservazione 310, per chiarire eventuali variazioni.

# Previsione mod4.

Tentiamo ora una previsione del peso usando il modello 4, per un neonato femmina la cui madre ha già avuto tre gravidanze e si trova alla 39esima settimana di gestazione. Faremo la previsione sia per madre fumatrice che per non fumatrice.

```
nuovi_neonatifumSI <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 1,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)

  )

previsionifumSI <- predict(mod4, newdata=nuovi_neonatifumSI)
```

Il modello calcola che la neonata dovrebbe pesare 3242.302 grammi. Passiamo al modello per madre non fumatrice.

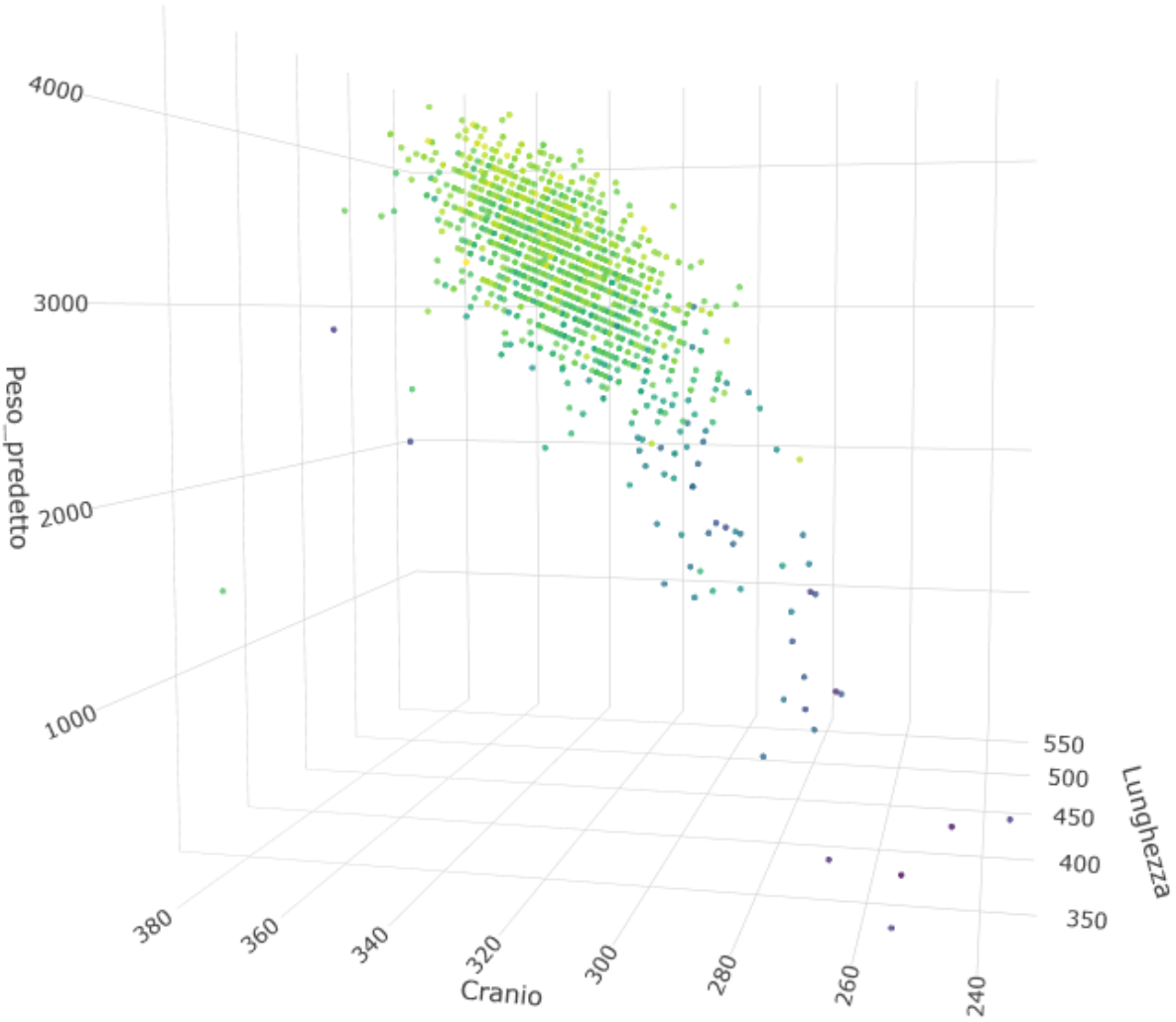
```
nuovi_neonatifumNO <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 0,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)

)

previsionifumNO <- predict(mod4, newdata=nuovi_neonatifumNO)
```

In questo caso il modello ci riporta un peso previsto di 3272.765 grammi.

Avere una rappresentazione grafica di questo modello sarebbe impossibile, viste le tante variabili utilizzate. Tentiamo di costruire un modello con tre variabili, rappresentandoli con un plot3D. Per farlo sceglieremo quelle con p-value più basso.

```
mod_simpl <- lm(Peso ~ Lunghezza + Cranio, data = neonati)
summary(mod_simpl)
library(plotly)
df_plot <- neonati
df_plot$Peso_predetto <- predict(mod_simpl, newdata = neonati)

# Creare il grafico
plot_ly(data = df_plot, x = ~Lunghezza, y = ~Cranio, z = ~Peso_predetto,
              type = 'scatter3d', mode = 'markers',
              marker = list(size = 2, color = ~Gestazione, colorscale='Viridis', opacity = 0.8))
```

A seguire uno stamp, ma trattandosi di un plot 3D consiglio di visionarlo tramite il file R.


4000
3000
2000
1000
Peso_predetto
380
360
340
320
300
280
260
240
Cranio
550
500
450
400
350
Lunghezza

# Previsione mod4plus

Precedentemente abbiamo visto come due particolari osservazioni, la 1551 e la 310, fossero oltre la soglia dello 0.5 di Cook. Proviamo a costruire un nuovo modello escludendole.

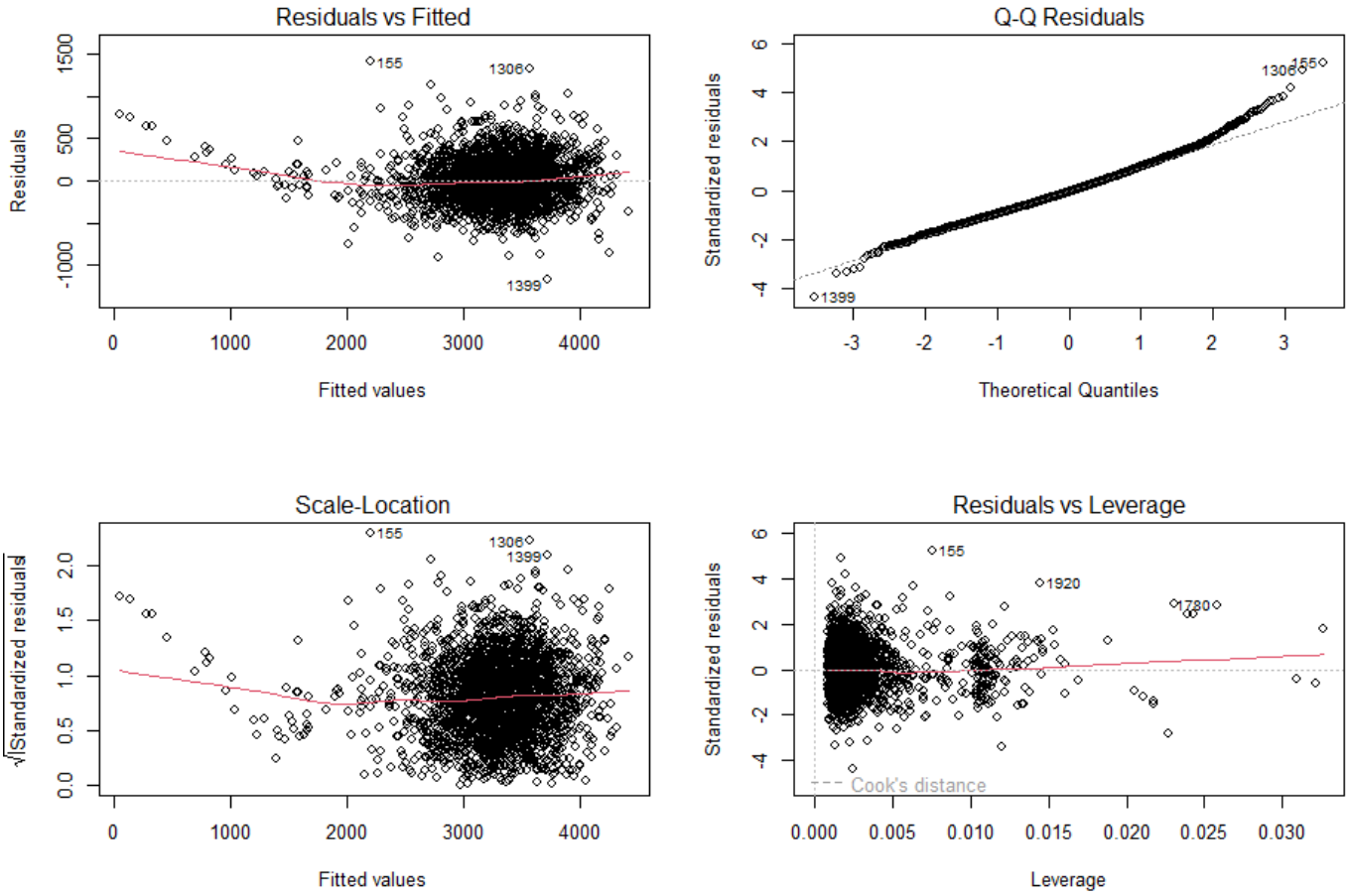
```
neonati2 <- neonati[-c(310, 1551), ]

mod4plus <- lm(formula = Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza +
                 Cranio + Sesso, data = neonati2)
summary(mod4plus)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1168.77  -179.86   -14.73   161.76  1402.96

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6662.9185   132.8843 -50.141  < 2e-16 ***
N.gravidanze   13.4937     4.2509   3.174  0.00152 **
Fumatrici     -27.2839    26.9948  -1.011  0.31225
Gestazione     28.2164     3.7585   7.507 8.36e-14 ***
Lunghezza      10.8386     0.3014  35.955  < 2e-16 ***
Cranio         10.0986     0.4245  23.789  < 2e-16 ***
SessoM         77.8064    10.9593   7.100 1.63e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 268.6 on 2491 degrees of freedom
Multiple R-squared:  0.7375,    Adjusted R-squared:  0.7369
F-statistic:  1167 on 6 and 2491 DF,  p-value: < 2.2e-16
```

In confronto al modello precedente gli indici dei residui cambiando tutti considerevolmente, a parte il terzo quartile. Anche alcuni gradi di significatività variano, senza però inficiare il modello (notiamo un aumento in Fumatrici, gestazione e SessoM).

L’adjusted R-squared beneficia di un leggero aumento, da 0.7265 a 0.7369.

Proseguiamo con l’analisi dei residui del nuovo modello.



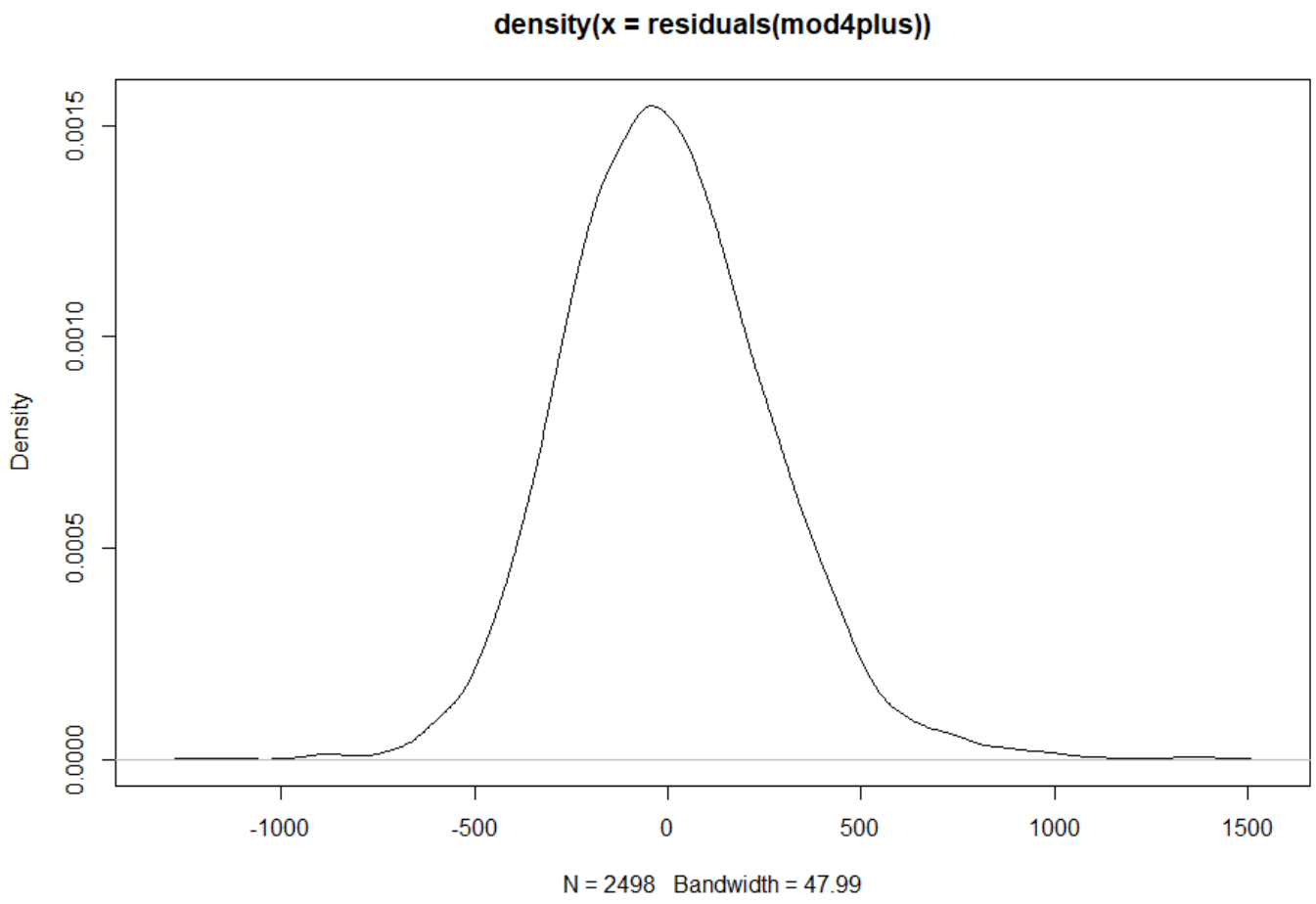

L'eteroschedasticità è ancora presente, ma con questo modello non abbiamo osservazioni che superano le soglie della distanza di Cook.

```
> shapiro.test(residuals(mod4plus))

        Shapiro-Wilk normality test

data:  residuals(mod4plus)
W = 0.98884, p-value = 4.641e-13

> lmtest::bptest(mod4plus)

        studentized Breusch-Pagan test

data:  mod4plus
BP = 6.3139, df = 6, p-value = 0.389

> lmtest::dwtest(mod4plus)

        Durbin-Watson test

data:  mod4plus
DW = 1.9584, p-value = 0.1493
alternative hypothesis: true autocorrelation is greater than 0
```

Lo Shapiro test ci restituisce dei dati si migliori, ma che non ci permettono di accettare l'ipotesi nulla. Pertanto, confermiamo che i residui non seguono una distribuzione normale.

Il Breusch-pagan test è variato considerevolmente; ora non abbiamo prove sufficienti per rifiutare l'ipotesi nulla di omoschedasticità. Possiamo affermare che la varianza dei residui sia ora costante.

Il Durbin-Watson test non presenta variazioni degne di nota, confermando l'assenza di autocorrelazione tra i residui del modello.

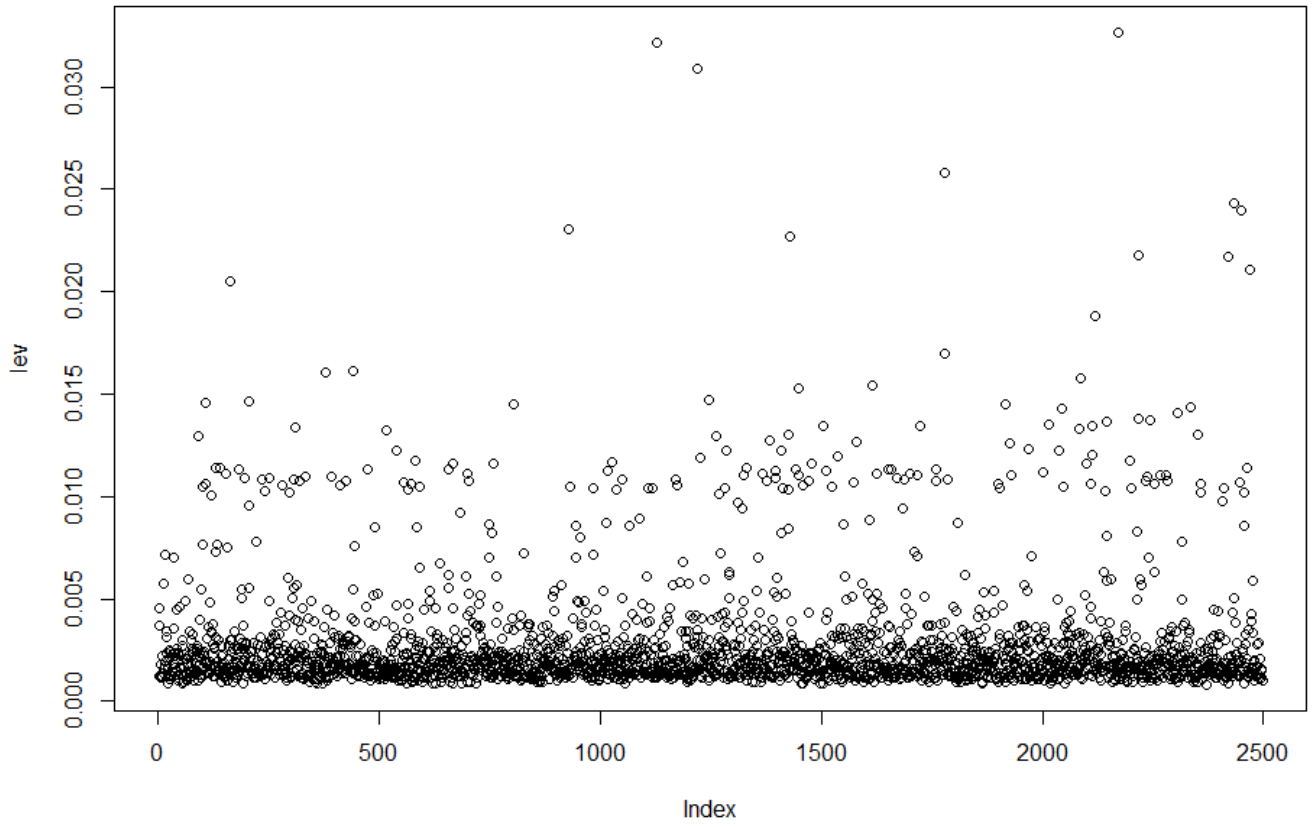



Il grafico di densità mostra ora una coda sulla destra più contenuta, data l'assenza dell'osservazione 1551. A parte ciò non si dimostrano cambiamenti di tendenza, con lunghe code a sinistra e a destra del grafico.

```
#leverage
lev <- hatvalues(mod4plus)
plot(lev)
p <- sum(lev)
soglia=2*p/n
abline(h=soglia,col=2)
lev[lev>soglia]
```

Controllando i leverage del modello pulito, ne possiamo ancora una volta constatare la totale assenza. La soglia non rientra nella visualizzazione.

```
> lev[lev>soglia]
named numeric(0)
>
```



Riguardo gli outliers ne sono presenti tre, già osservati in precedenza.

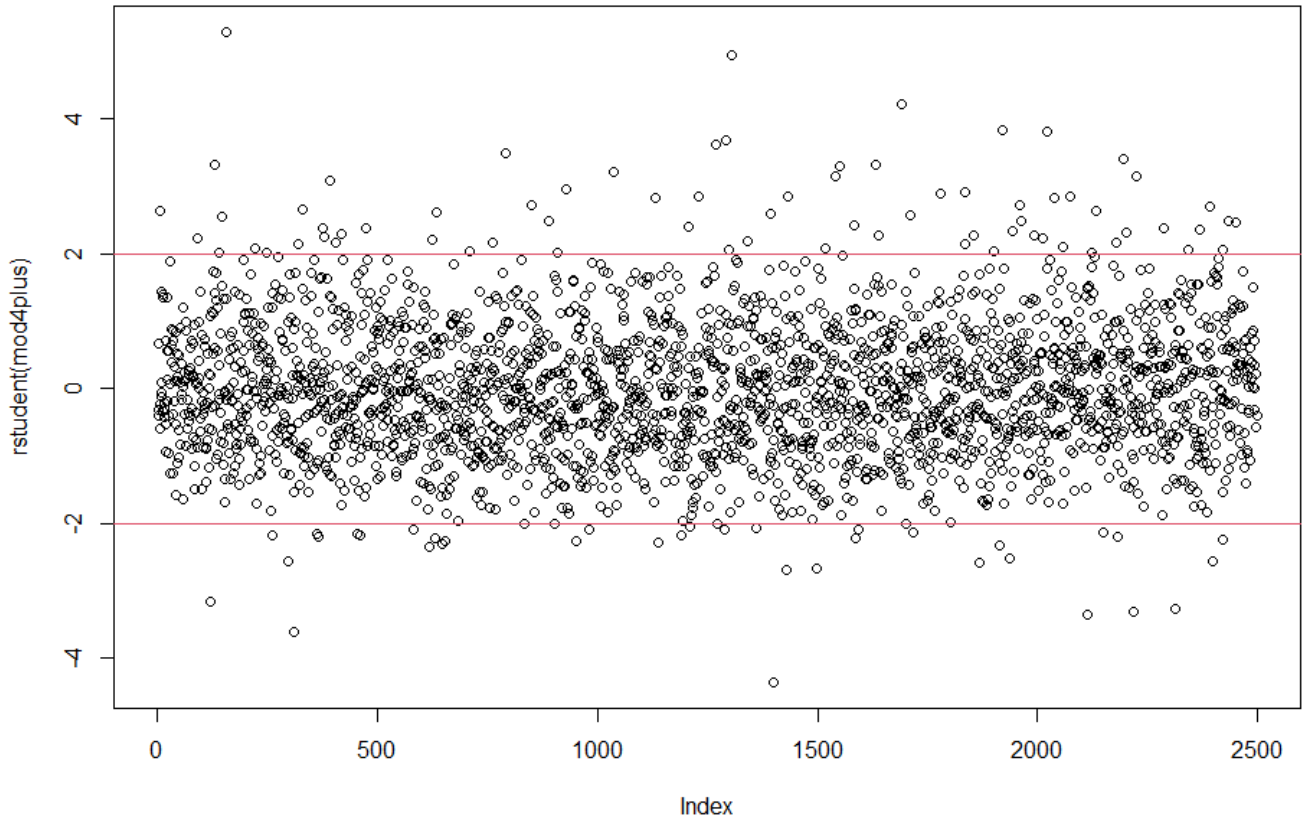
```
#outliers
plot(rstudent(mod4plus))
abline(h=c(-2,2), col=2)
car::outlierTest(mod4plus)
```

```
> car::outlierTest(mod4plus)
      rstudent unadjusted p-value Bonferroni p
155   5.282337         1.3859e-07   0.00034634
1306  4.938830         8.3797e-07   0.00209410
1399 -4.353267         1.3954e-05   0.03487100
```

L'osservazione 155 ha un residuo elevato, con un p-value piuttosto basso e un p-value di Bonferroni altrettanto basso, il che ci porta a classificarlo come outlier.

L'osservazione 1306 ha anch'essa un residuo elevato e un p-value di Bonferroni basso, suggerendoci che si tratta di un outlier.

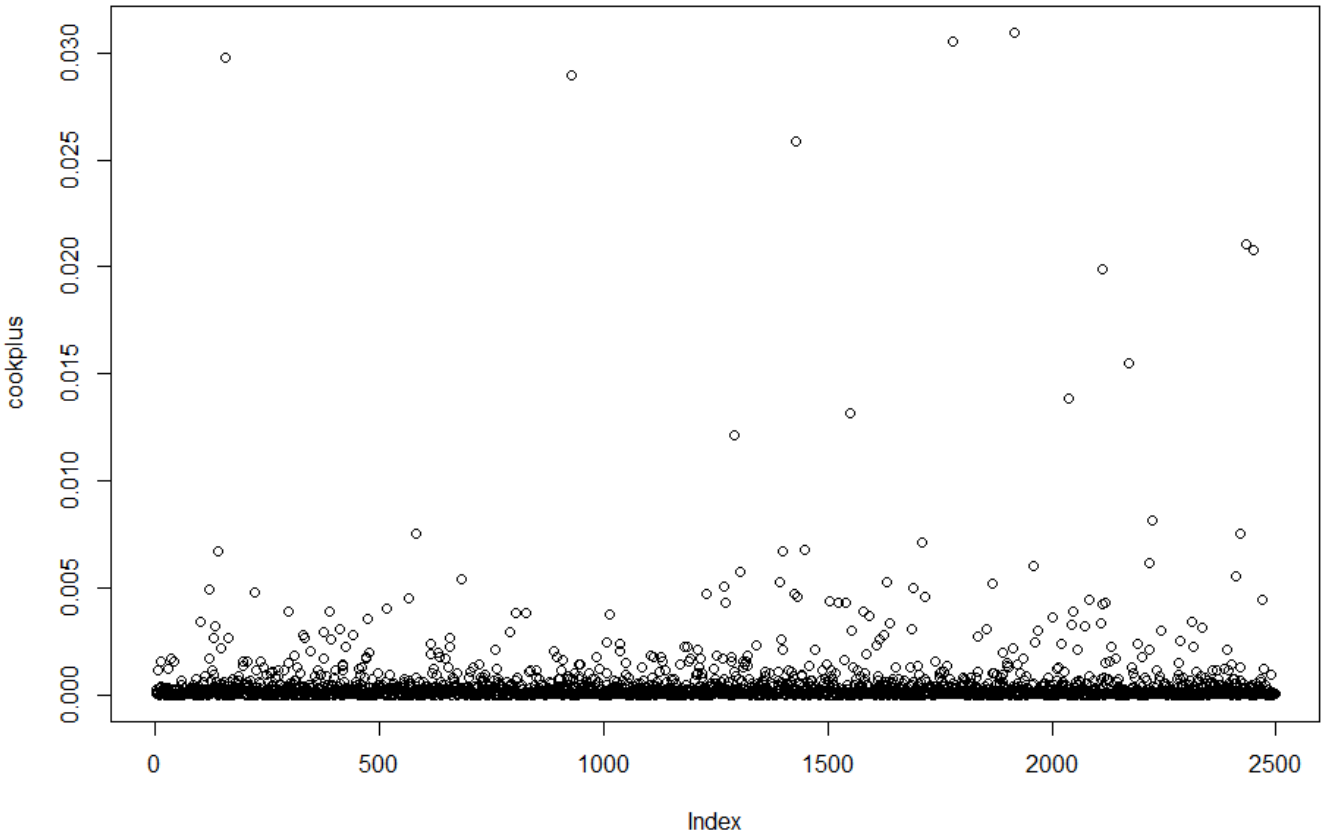
L'osservazione 1399 ha un residuo negativo, e un basso p-value di Bonferroni, più alto delle altre due osservazioni ma comunque sotto la soglia. Anche in questo caso consideriamo l'osservazione come un outlier significativo.



```r
#distanza di cook
cookplus <-cooks.distance(mod4plus)
plot(cookplus)
max(cookplus)
```

Ulteriore controllo con la distanza di cook, per capire se vi siano valori che vanno oltre la soglia. Eseguiamo il controllo sia con un plot che richiamando il valore massimo, che in questo caso sarà 0.03093833.

Il plot ci conferma che non ci sono più valori oltre la soglia



Ora che il modello è pulito tentiamo una nuova previsione, sempre considerando i valori per una neonata femmina con una madre alla trentanovesima settimana di gestazione e alla terza gravidanza, sia fumatrice che no.

```
plus_neonatifumSI <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 1,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)
  
)

plus_previsionifumSI <- predict(mod4plus, newdata=plus_neonatifumSI)
```

Il peso predetto per una neonata nata da madre fumatrice è di 3246.302 grammi, come nel precedente modello.

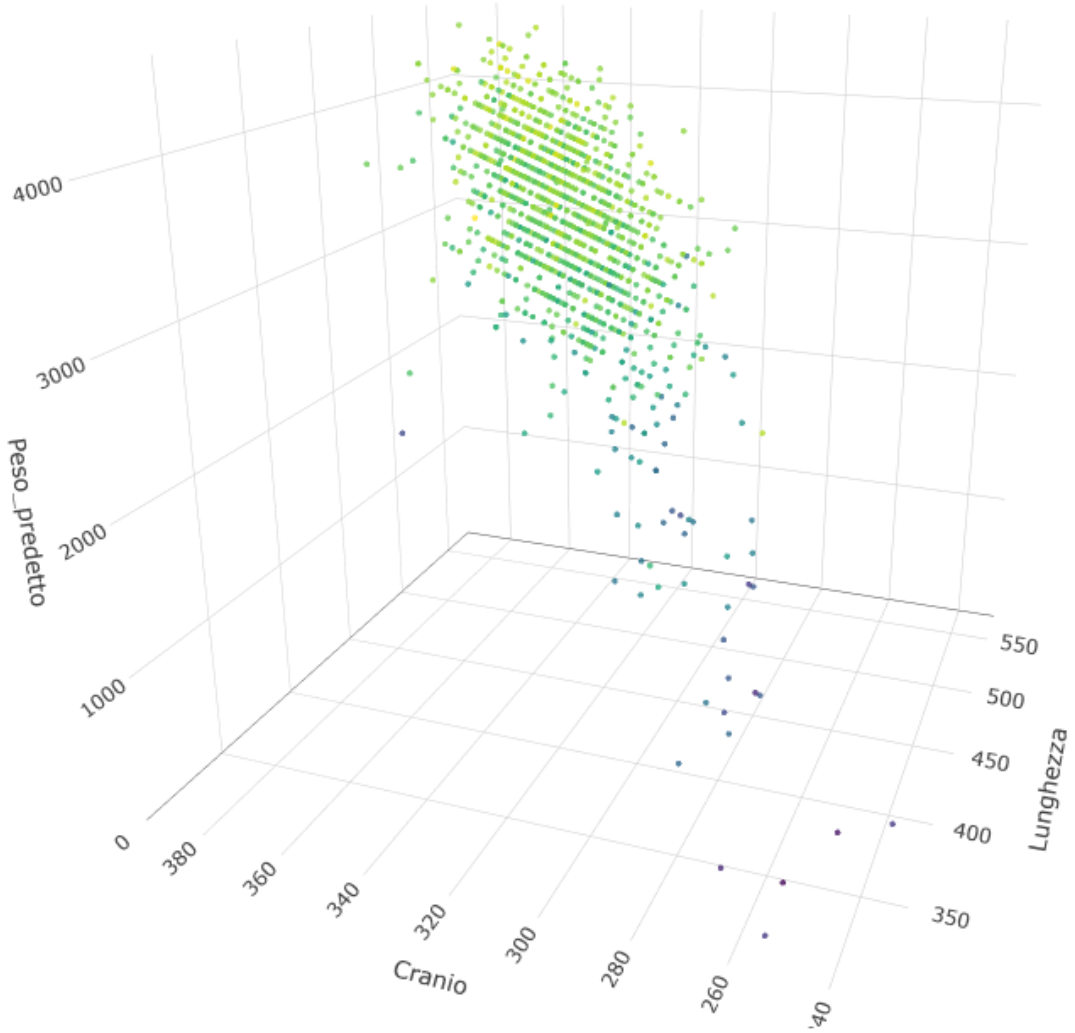
```
plus_previsionifumSI <- predict(mod4plus, newdata=plus_neonatifumSI)

plus_neonatifumNO <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 0,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)
  
)

plus_previsionifumNO <- predict(mod4plus, newdata=plus_neonatifumNO)
```

Il peso predetto per una neonata nata da madre non fumatrice è di 3273.586 grammi.

Procediamo ora alla costruzione di un modello più semplificato, rappresentabile con un plot3D.

```
mod_simpl_plus <- lm(Peso ~ Lunghezza + Cranio, data = neonati2)
summary(mod_simpl_plus)
library(plotly)
df_plot2 <- neonati2
df_plot2$Peso_predetto <- predict(mod_simpl_plus, newdata = neonati2)

# Creare il grafico
plot_ly(data = df_plot2, x = ~Lunghezza, y = ~Cranio, z = ~Peso_predetto,
        type = 'scatter3d', mode = 'markers',
        marker = list(size = 2, color = ~Gestazione, colorscale='Viridis', opacity = 0.8))
```

Anche in questo caso invito a visionare il plot dal file R, qui allegherò uno stamp per una prima visione.


4000
3000
2000
1000
0
Peso_predetto
380
360
340
320
300
280
260
Cranio
550
500
450
400
350
Lunghezza