Caso di studio “Nascite Neonati”

Il seguente documento ha lo scopo di spiegare i procedimenti e i risultati dell’analisi di un dataset contenente le caratteristiche di diversi neonati e delle loro madri. Questo allo scopo di capire se le abitudini e le caratteristiche della madre influenzano le caratteristiche del nascituro, un esempio potrebbe essere “il fumo causa la nascita prematura del bambino ?”. Un altro obiettivo costruire un modello di regressione lineare per predire il peso del bambino date la altre informazioni.

Analisi dataset

Iniziamo elencando le variabili presenti nel dataset fornendo duna breve descrizione.
Per ogni riga del dataset avremo:

- **Anni madre**: Che rappresenta gli anni della madre del neonato. **QU discreta**
- **N. Gravidanze**: Che rappresenta il numero di gravidanze che la madre ha avuto. **QU discreta**
- **Fumatrici**: Che rappresenta se la madre è fumatrice o meno (0=NO, 1=SI). **QA nominale codificata come variabile QU discreta**
- **Gestazione**: Che rappresenta il numero di settimane di gestazione. **QU discreta**
- **Peso**: Che rappresenta il peso in grammi del neonato. **QU continua**
- **Lunghezza**: Che rappresenta la lunghezza in millimetri del neonato. **QU discreta**
- **Cranio**: Che rappresenta il diametro in millimetri del cranio del neonato. **QU discreta**
- **Tipo Parto**: Che rappresenta la tipologia di parto eseguita (Ces=Cesario, Nat=Naturale). **QA nominale**
- **Ospedale**: Che rappresenta l’ospedale dove stata effettuata la rilevazione, ci sono 3 ospedali (osp1=Ospedale 1, osp2=Ospedale 2, osp3=Ospedale 3). **QA nominale**
- **Sesso**: Che rappresenta il sesso del nascituro (M=maschio, F=femmina). **QA nominale**

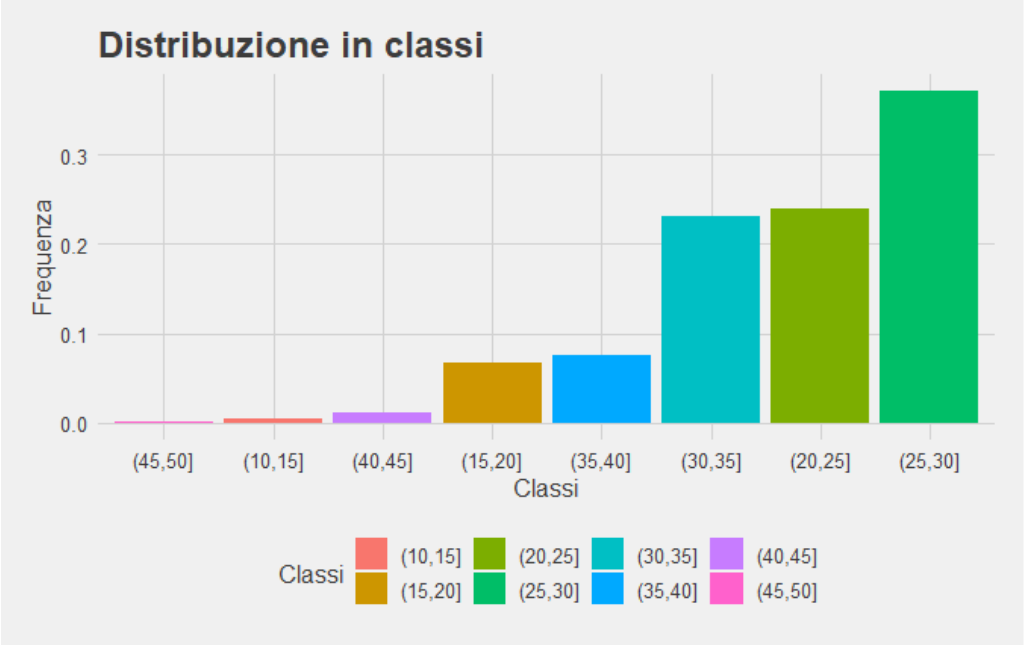
Le variabili hanno i seguenti tipi:
QA: Variabile Qualitativa
QU: Variabile Quantitativa

Andiamo ora ad analizzare singolarmente le variabili.

Anni madre

Facendo una prima analisi del dataset notiamo che ci sono delle imperfezioni in delle registrazioni, ci sono delle madri che hanno età 0 e 1. Questo è ovviamente impossibile, dovremo quindi fare un passaggio di pulizia del dataset. Elimineremo dal dataset tutte le registrazioni che hanno “Anni Madre” minore di 12 (età minima per rimanere incinte).

Osserviamo ora la distribuzione della variabile divisa in classi, ogni classe avrà un range di 5 anni di età.



Osserviamo inoltre i vari indici della variabile.

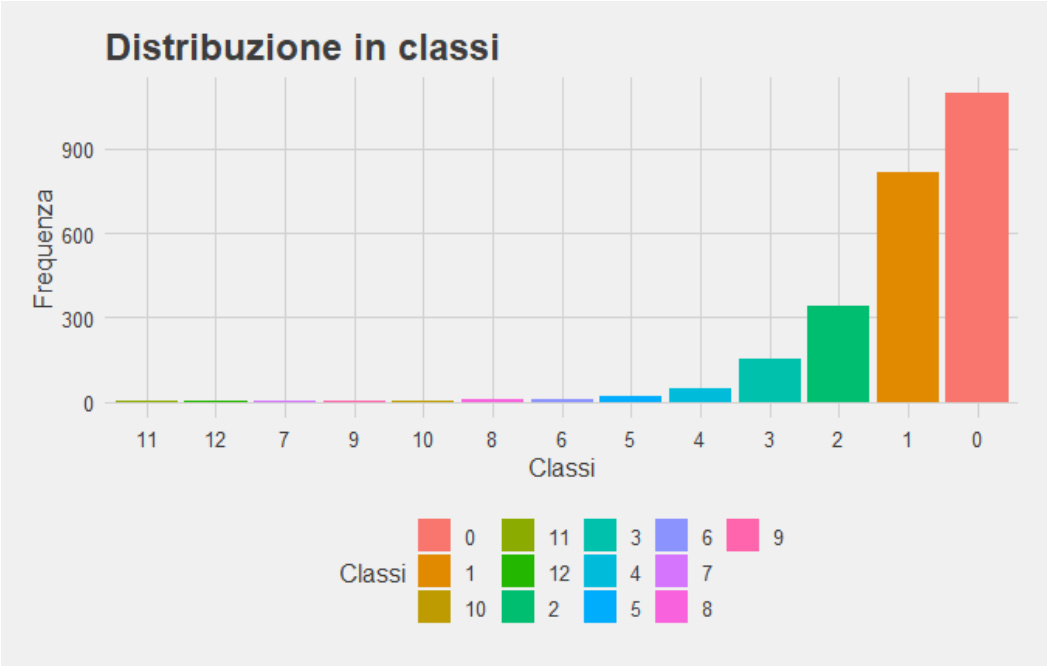
	Anni Madre
Min.	13
1st Qu.	25
Median	28
Mean	28.1861489191353
3rd Qu.	32
Max.	46
Range	33
IQR	7
Mode	30
Var	27.2192393883068
SD	5.21720609026582
CV	18.5098223430016
Asymmetry	0.151062429717029
Curtosi	-3.10560608559434

N.gravidanze

Osservando i vari indici della variabile

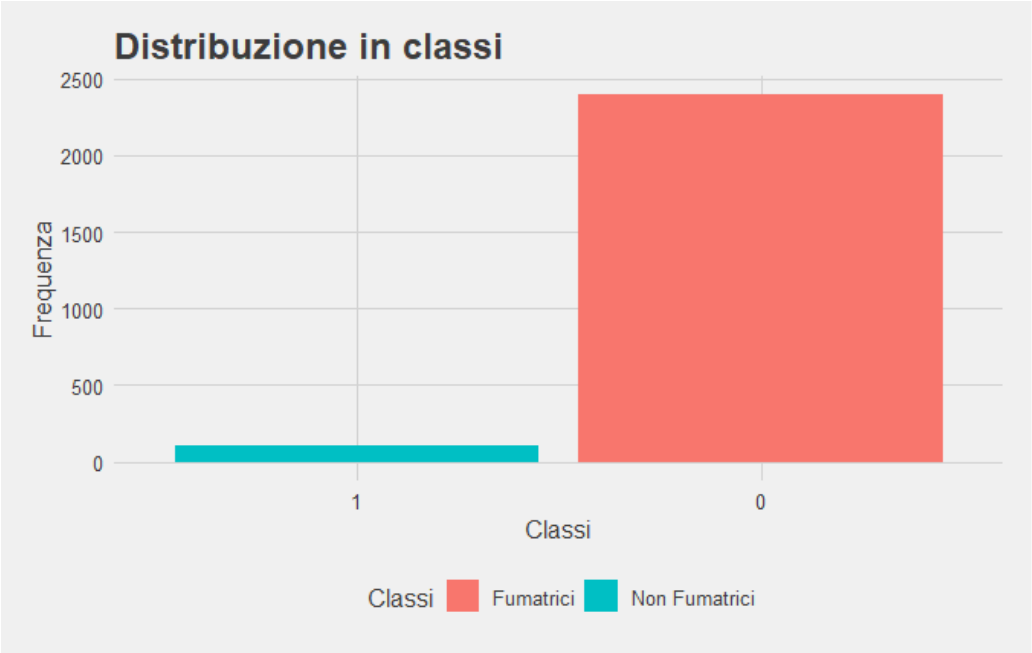
	N.Gravidanze
Min.	0
1st Qu.	0
Median	1
Mean	0.981585268214572
3rd Qu.	1
Max.	12
Range	12
IQR	1
Mode	0
Var	1.64083016513331
SD	1.28094893150871
CV	130.497978422054
Asymmetry	2.51341228629277
Curtosi	7.98162689091485

Dalla media della classe notiamo che tutte le donne hanno avuto un figlio, ma la classe modale è 0. Osserviamo il grafico delle classi per completezza:



Fumatrici

Analizziamo un grafico per capire quante pazienti fumano



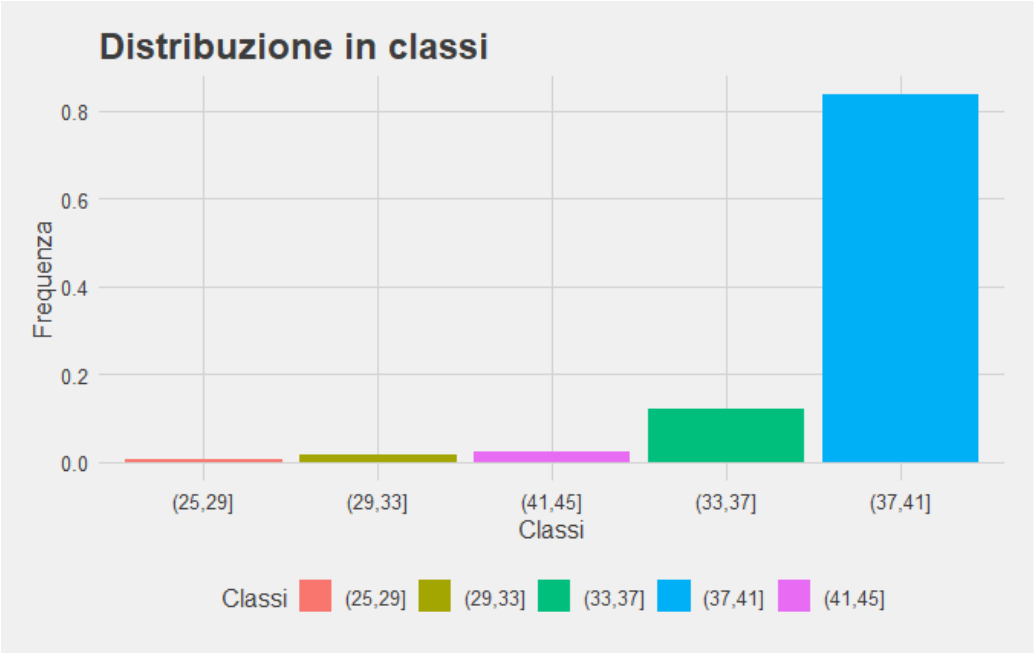
La quasi la totalità delle pazienti sono delle non fumatrici

Gestazione

Osserviamo gli indici

	Gestazione
Min.	25
1st Qu.	38
Median	39
Mean	38.9795836669336
3rd Qu.	40
Max.	43
Range	18
IQR	2
Mode	40
Var	3.49297507689772
SD	1.8689502606805
CV	4.79469015536443
Asymmetry	-2.06513085638468
Curtosi	5.25551575392918

E il grafico delle classi:



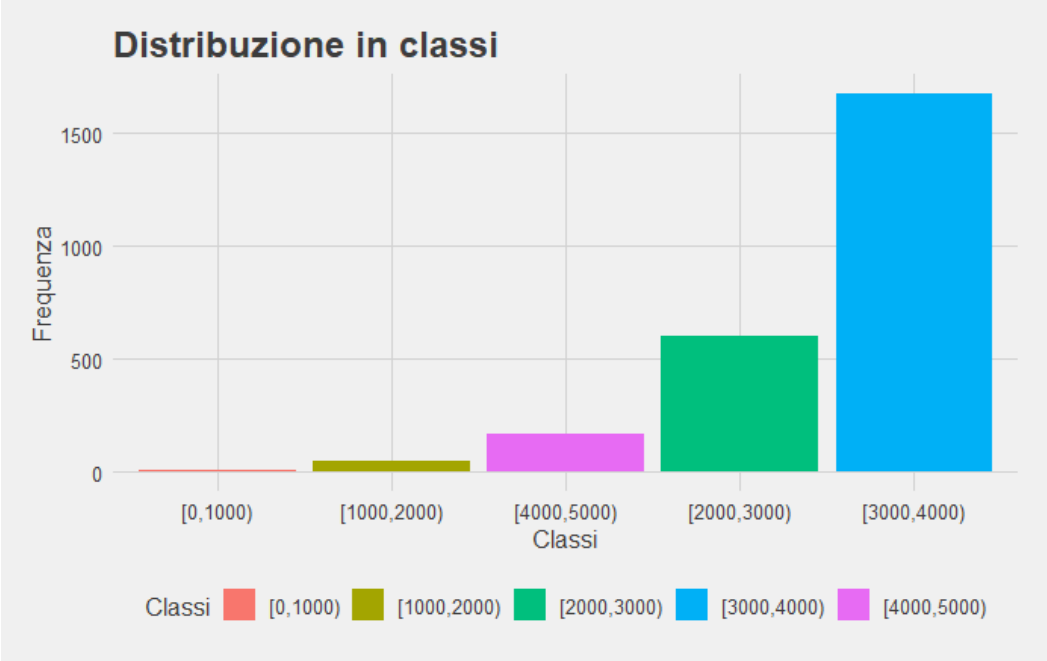
Quindi la maggior parte delle pazienti hanno dalle 37 alle 41 settimane di gestazione.

Peso

Osserviamo i vari indici

	Peso
Min.	830
1st Qu.	2990
Median	3300
Mean	3284.18414731785
3rd Qu.	3620
Max.	4930
Range	4100
IQR	630
Mode	3300
Var	275865.895591603
SD	525.229374265762
CV	15.9926895297485
Asymmetry	-0.64740356386541
Curtosi	-0.971246919174976

E il grafico delle classi per capire quale la classe di peso maggiore



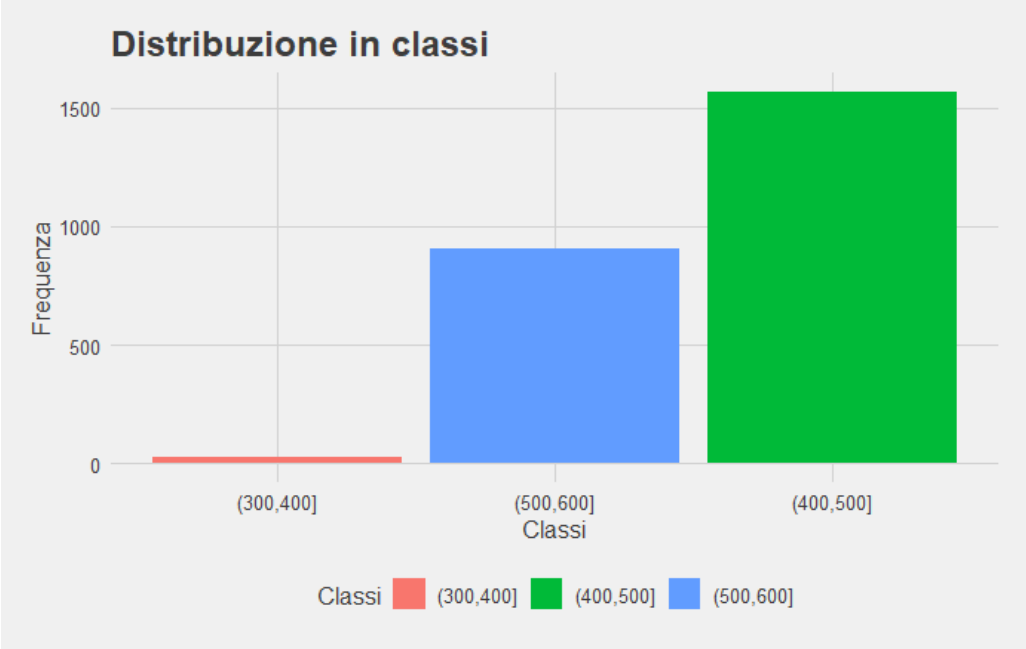
Su 2500 osservazioni la 1500 ricadono tra i 3 e i 4 kg

Lunghezza

Vediamo i vari indici

	Lunghezza
Min.	310
1st Qu.	480
Median	500
Mean	494.695756605284
3rd Qu.	510
Max.	565
Range	255
IQR	30
Mode	500
Var	693.208159799766
SD	26.3288465337881
CV	5.32223011461887
Asymmetry	-1.51457455685883
Curtosi	3.48093044337591

E la divisione in classi

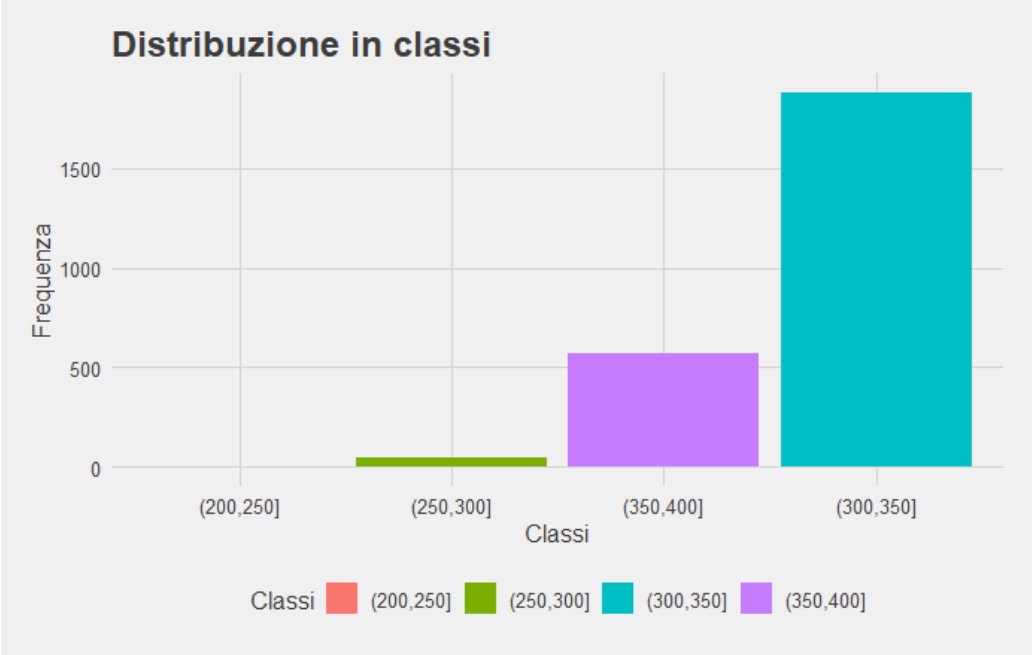


Cranio

Osserviamo i vari indici

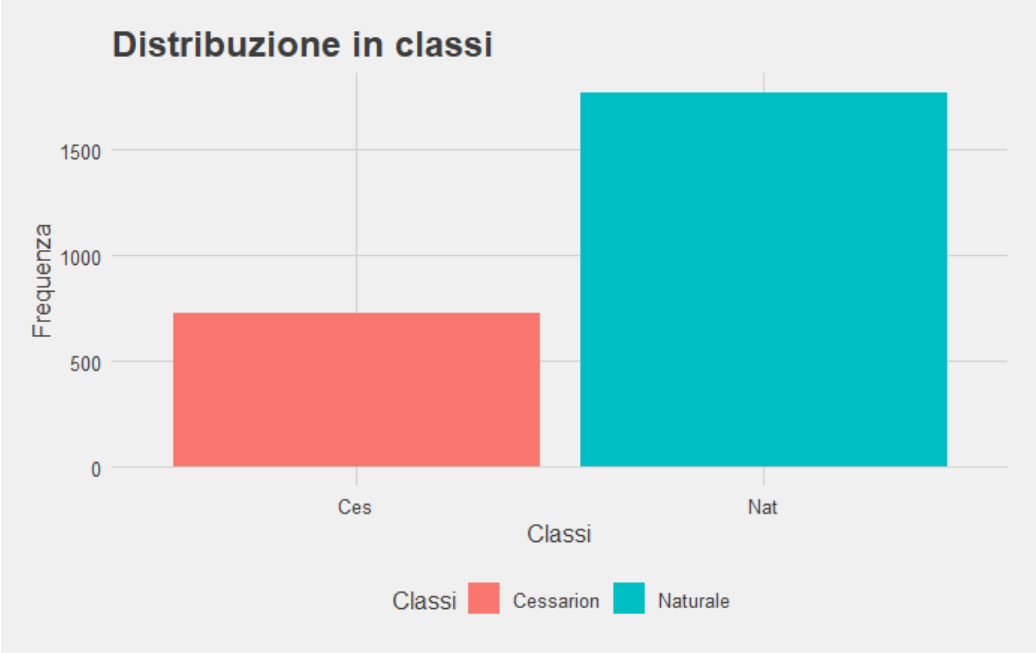
	Cranio
Min.	235
1st Qu.	330
Median	340
Mean	340.029223378703
3rd Qu.	350
Max.	390
Range	155
IQR	20
Mode	340
Var	269.927459628897
SD	16.429469243676
CV	4.8317815393702
Asymmetry	-0.785090629706357
Curtosi	-0.0551296322234931

E il grafico delle classi



Tipo parto

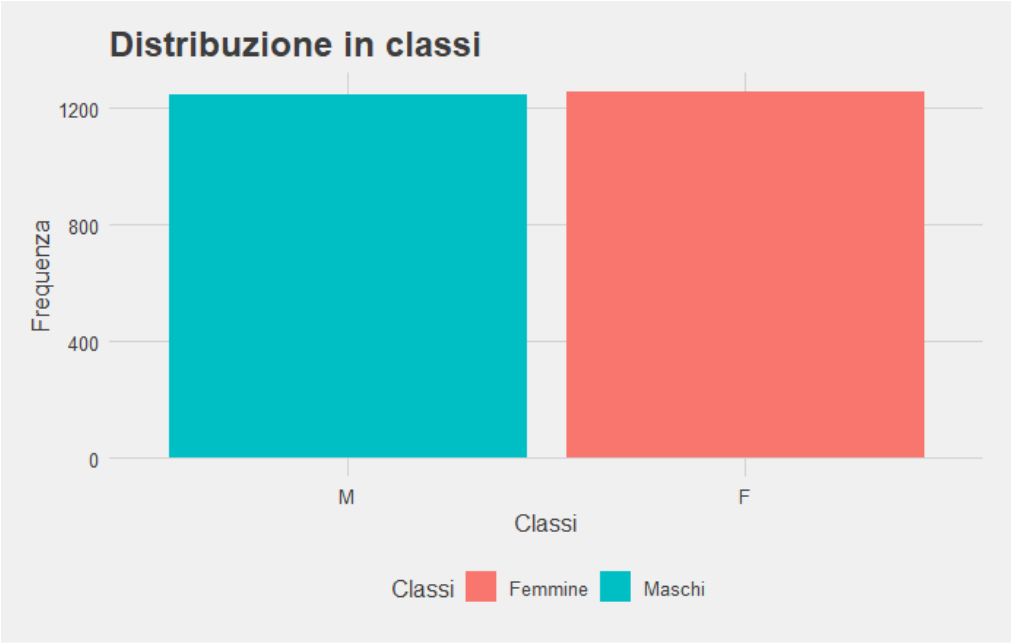
Vediamo la distribuzione delle classi



Scopriamo che la maggior parte delle donne decidono di avere un parto naturale

Sesso

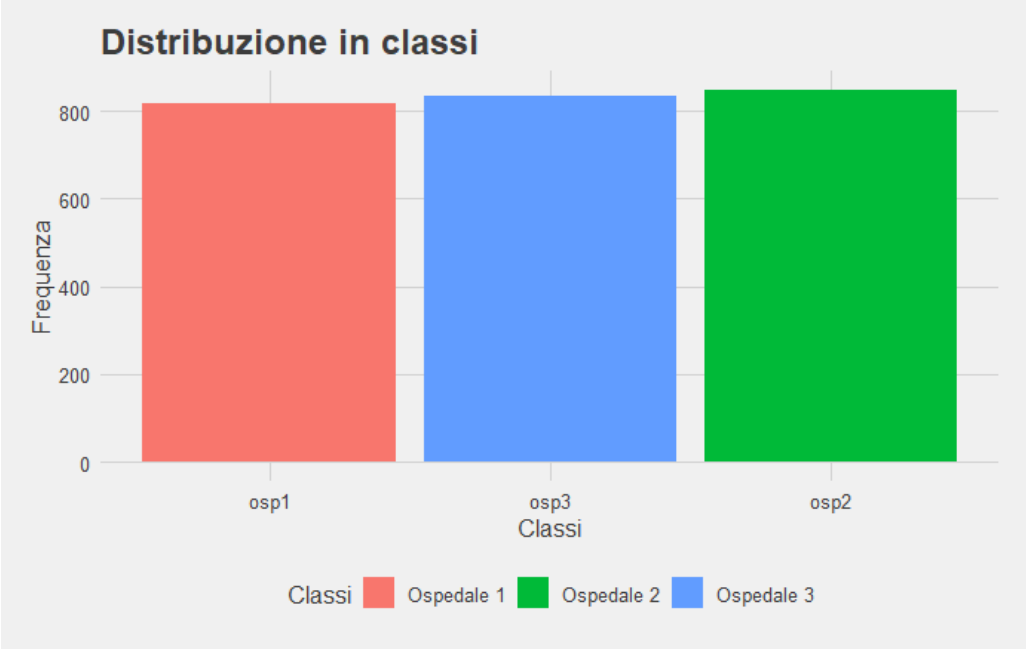
Osserviamo la distribuzione dei bambini nel nostro dataset



Come possiamo vedere sono praticamente alla pari

Ospedale

Infine vediamo la distribuzione delle nascite nei tre ospedali

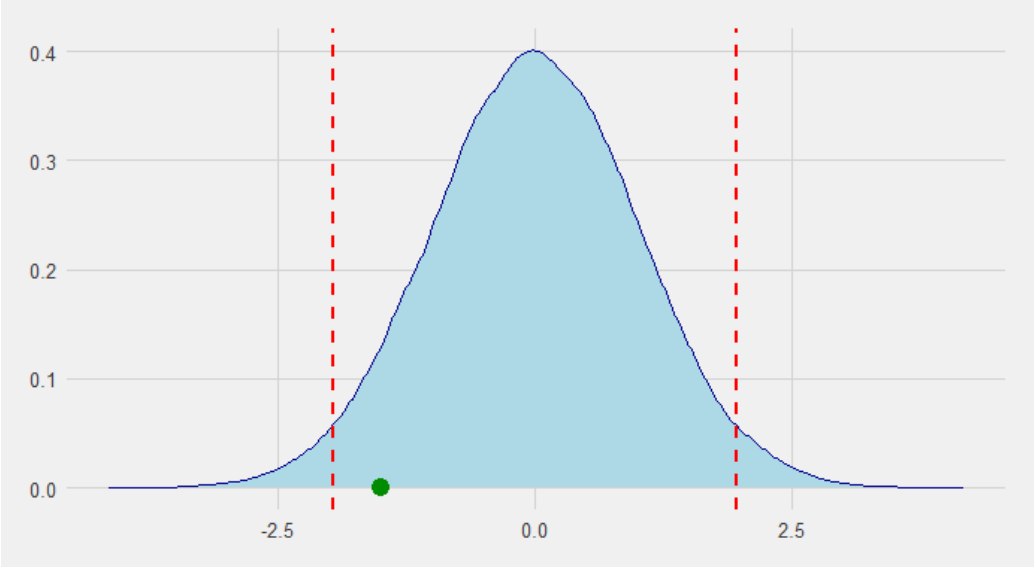


Ipotesi

Facciamo ora diversi test di ipotesi per verificare che il nostro campione di riferimento abbia le stesse informazioni della popolazione che intendiamo studiare.

Peso

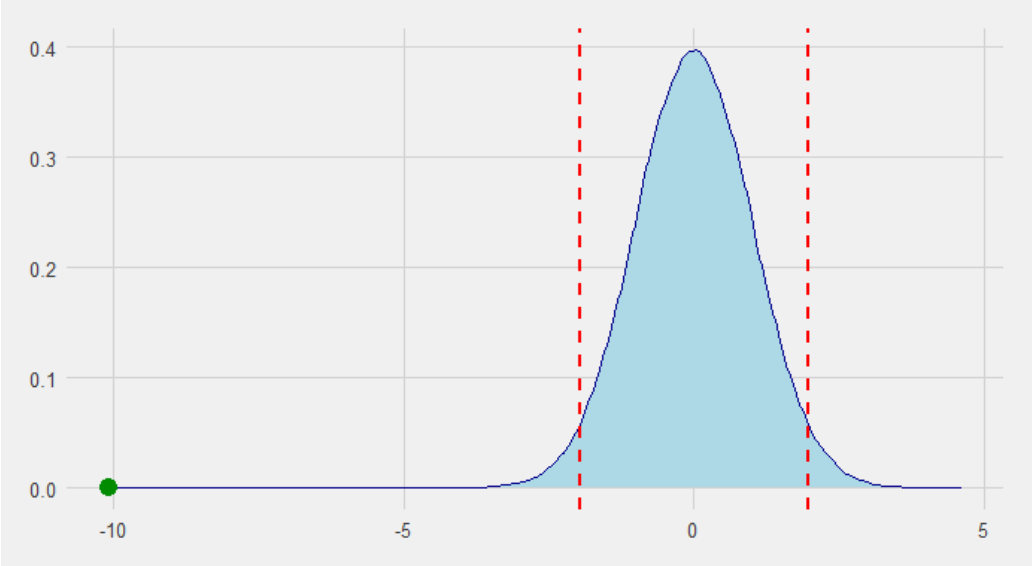
Vogliamo saggiare l'ipotesi che la media del peso del nostro campione sia diversa dalla media della popolazione.
Eseguiamo quindi un t test, il valore del p.value è di 0.1324 il che ci permette di accettare l'ipotesi nulla di uguaglianza, capiamo quindi che le due medie coincidono.
Vediamo il tutto a livello grafico:



Lunghezza

Come per il peso vogliamo saggiare l'ipotesi di diversità fra la media del campione e della popolazione originale
Eseguiamo quindi un t test, il valore del p.value in questo caso si avvina incredibilmente allo 0. Rifiutiamo quindi l'ipotesi nulla e accentiamo l'ipotesi di diversità tra le due medie. Dovremo quindi fare attenzione alle informazioni che dedurremo da questo dataset.

Vediamo il tutto a livello grafico:

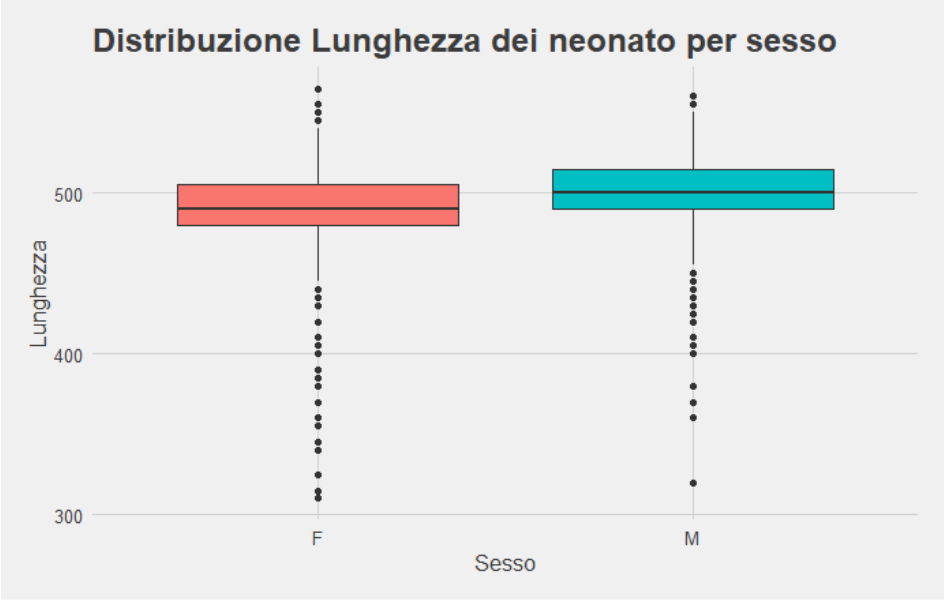


Differenza fra Sesso

Altre ipotesi che vale la pena saggiare, sono quelle che ipotizzano un valore medio diverso per i sessi. Andiamo a verificare queste ipotesi:

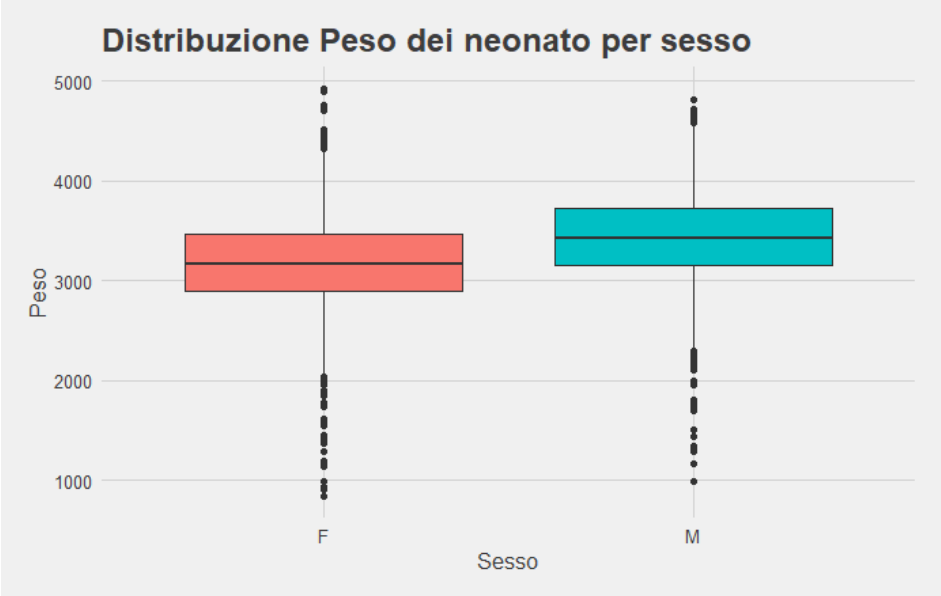
Differenza di lunghezza media

Eseguendo un test con correzione di bonferroni notiamo che il p-value e molto vicino a 0, questo ci fa rifiutare l'ipotesi di uguaglianza e ci dice che c'è una differenza significativa fra la lunghezza media per i due sessi



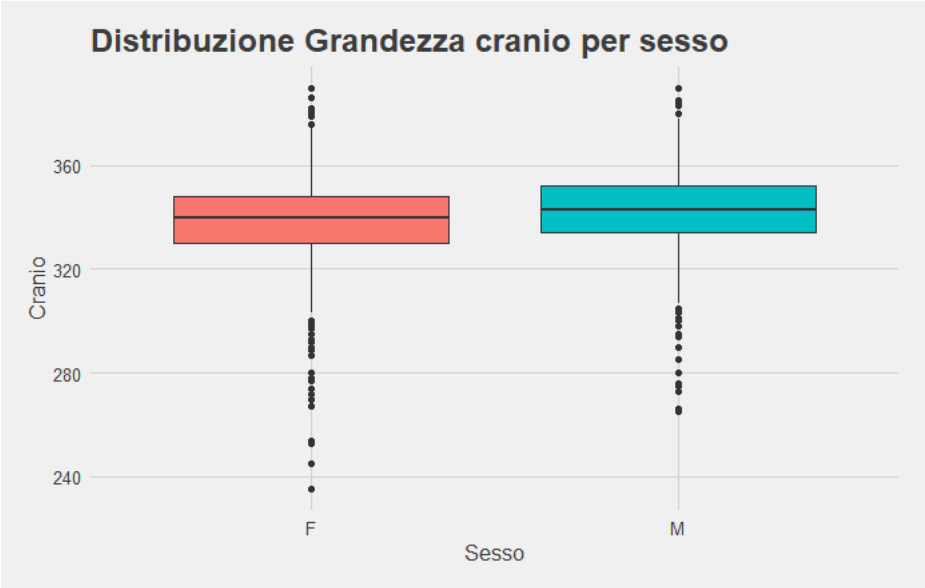
Differenza Peso medio

Eseguendo un test con correzione di bonferroni notiamo che il p-value e molto vicino a 0, questo ci fa rifiutare l'ipotesi di uguaglianza e ci dice che c'è una differenza significativa tra il peso medio per i due sessi.



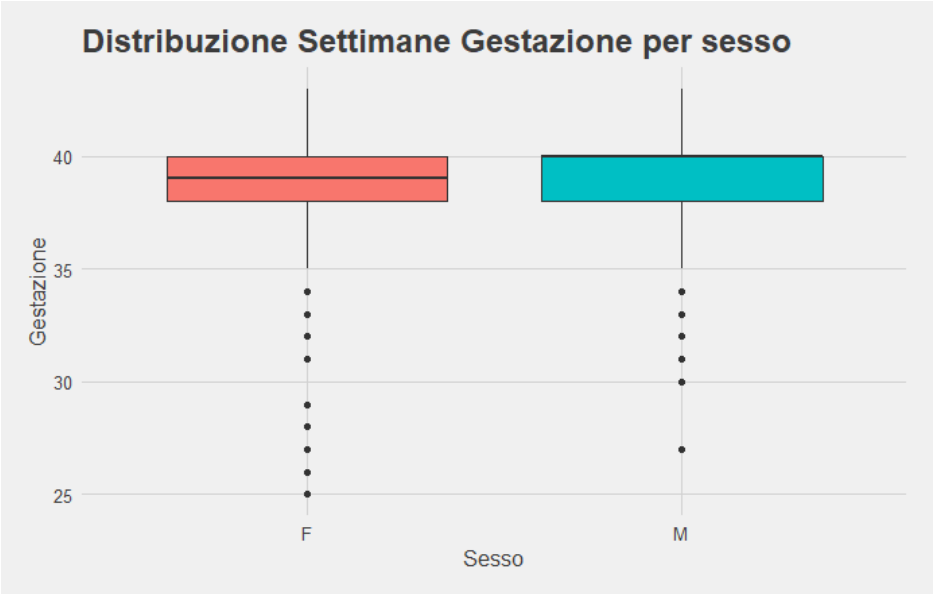
Differenza Cranio medio

Eseguendo un test con correzione di bonferroni notiamo che il p-value e molto vicino a 0, questo ci fa rifiutare l'ipotesi di uguaglianza e ci dice che c'è una differenza significativa tra la grandezza media del cranio per i due sessi.



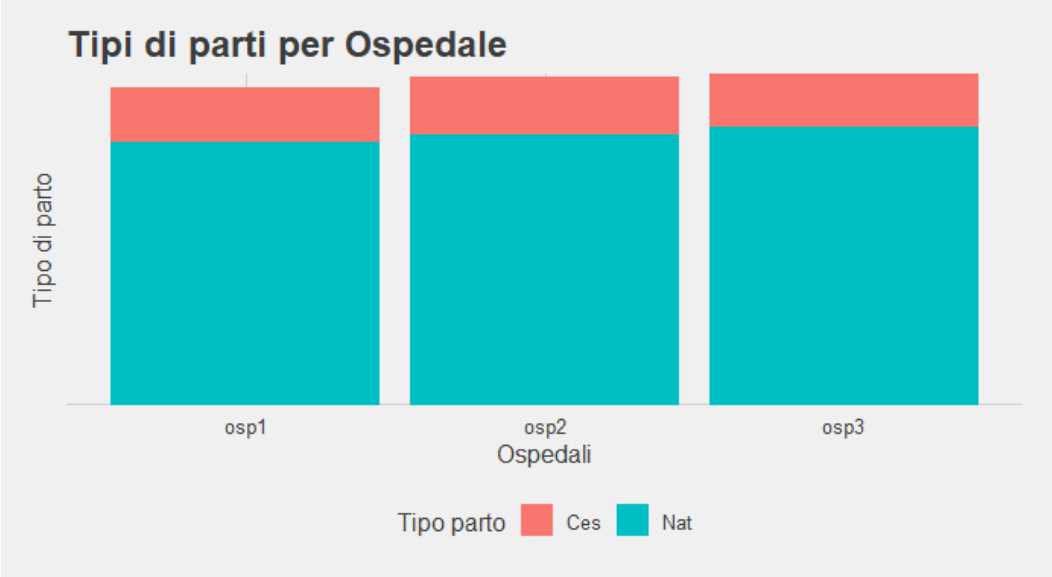
Differenza gestazione media

Eseguendo un test con correzione di bonferroni notiamo che il p-value e molto vicino a 0, questo ci fa rifiutare l'ipotesi di uguaglianza e ci dice che c'è una differenza significativa tra il tempo di gestazione medio per i due sessi.



Tipi di parto

Si dice che in alcuni ospedali si facciano più parti caseari di altri, possiamo verificare questa ipotesi basta osservare il plot delle distribuzioni.

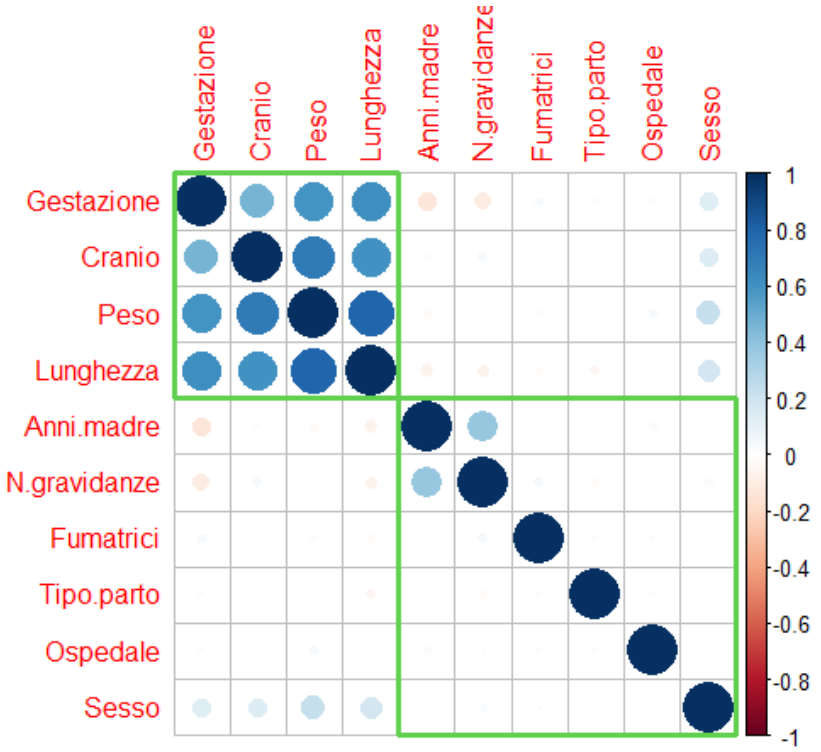


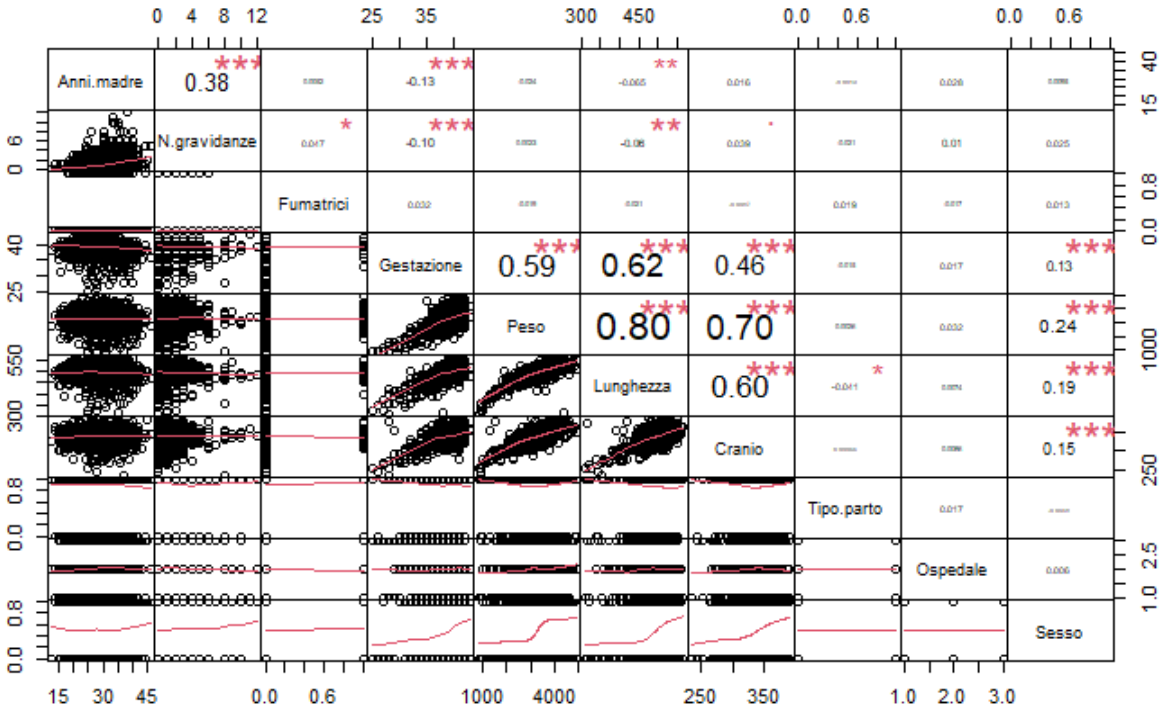
Possiamo notare che non ci sono grandi differenze nel numero di parti cesari per i diversi ospedali.

Correlazione fra variabili

Matrice di correlazione

Iniziamo ora ad analizzare le relazioni tra le variabili. Mostriamo il grafico di diverse matrici di correlazione.



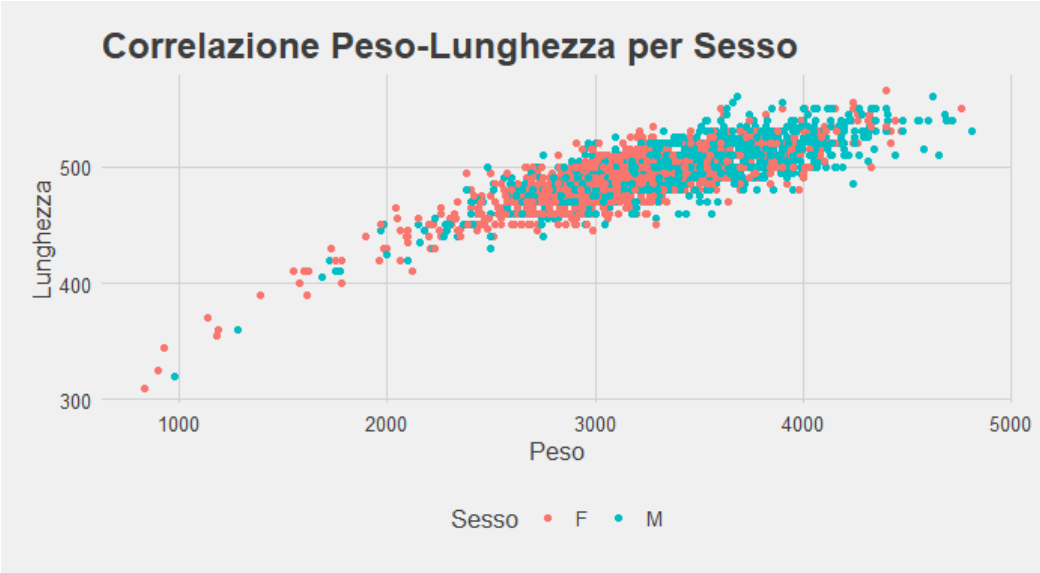


Da queste matrici possiamo subito notare che Peso è altamente correlata con Lunghezza, Cranio, Gestazione e Sesso. Questo ci fa capire che il Peso del bambino è particolarmente influenzato dalle sue caratteristiche fisiche. Considerando che abbiamo variabili molto correlate dovremo fare quindi attenzione per l'eventuale presenta di multi multicollinearità del modello.

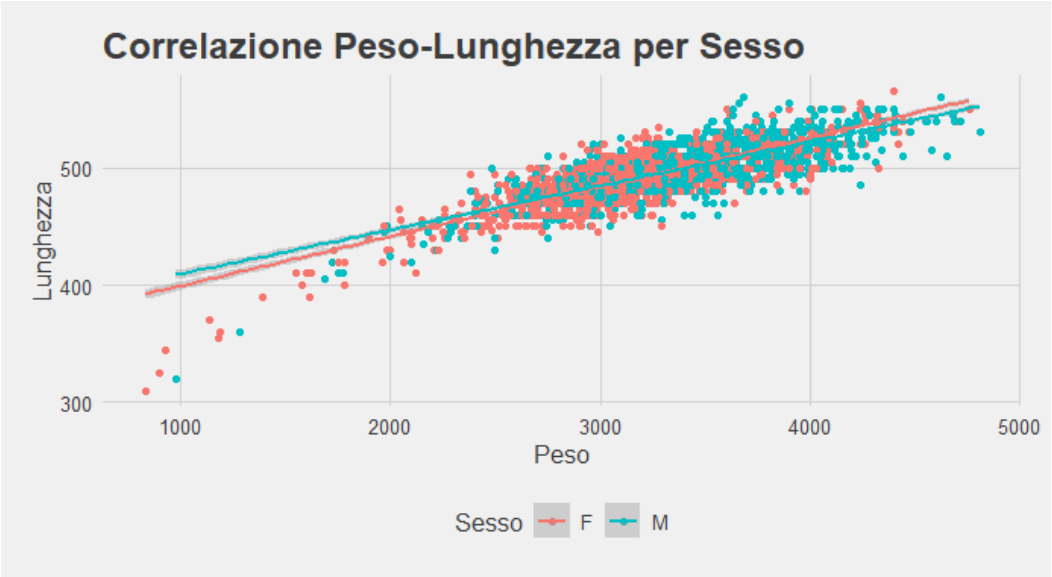
Presenza di relazioni non lineari

Un altra cosa da analizzare prima di costruire il modello è osservare se ci sono delle crescite non lineari tra i regressori, Noi ci concentreremo su quelli molto correlati alla variabile target poiché sono anche i variabili quantitative.
Facciamo qualche plot per vedere le relazioni.

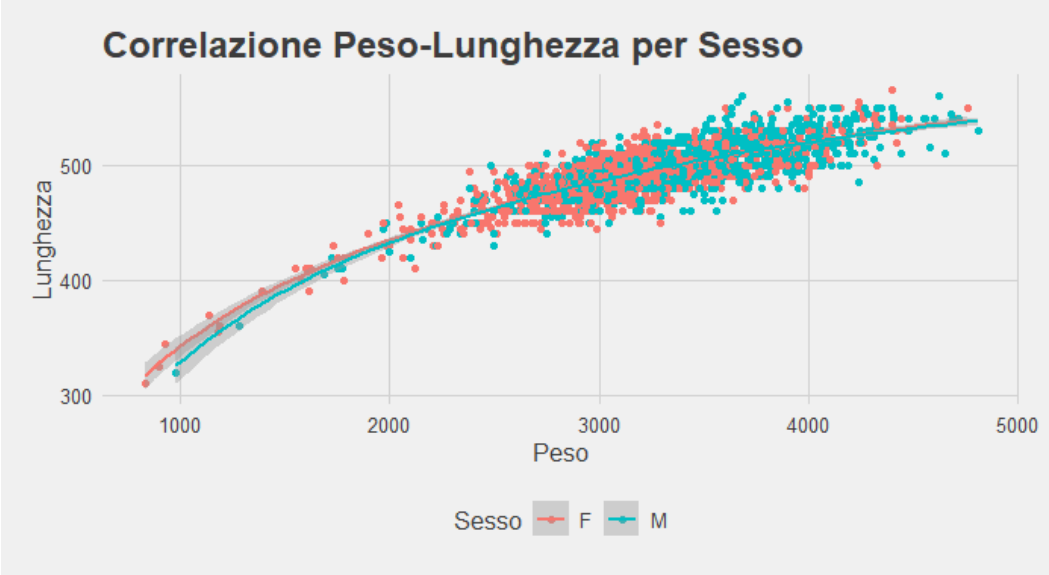
Iniziamo dalla Lunghezza



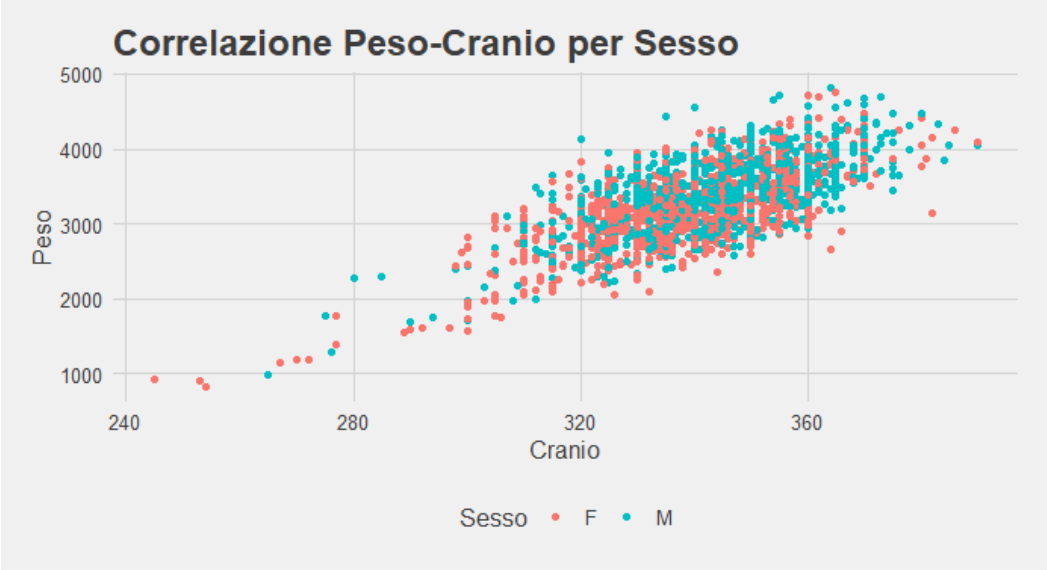
Osserviamo come sarebbe un possibile retta di regressione



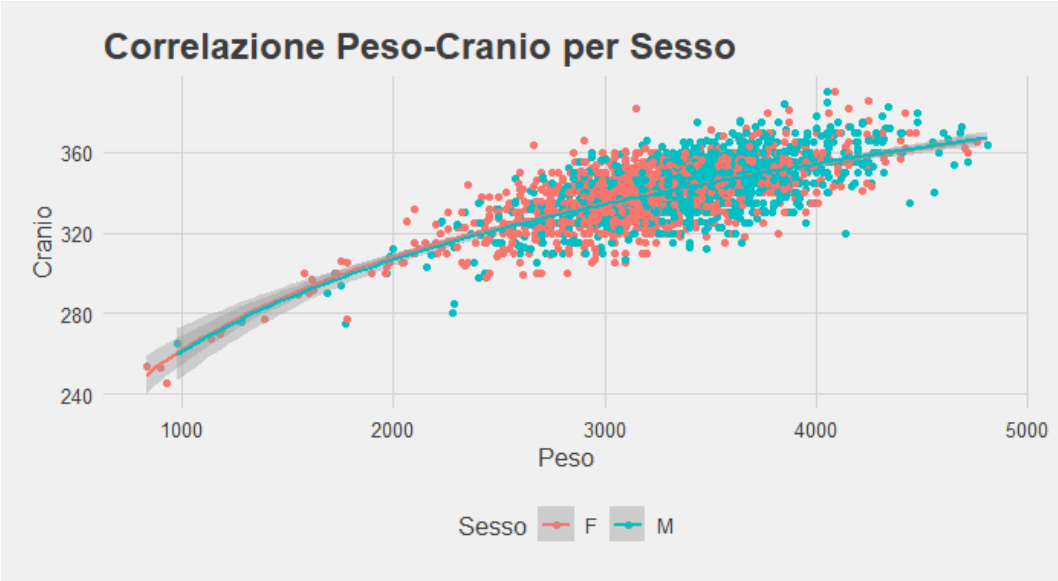
Come notiamo la crescita non si adatta perfettamente a una retta lineare, facendo qualche test notiamo che la crescita che più si adatta ai dati è quella logaritmica.



Osserviamo poi la grandezza del Cranio

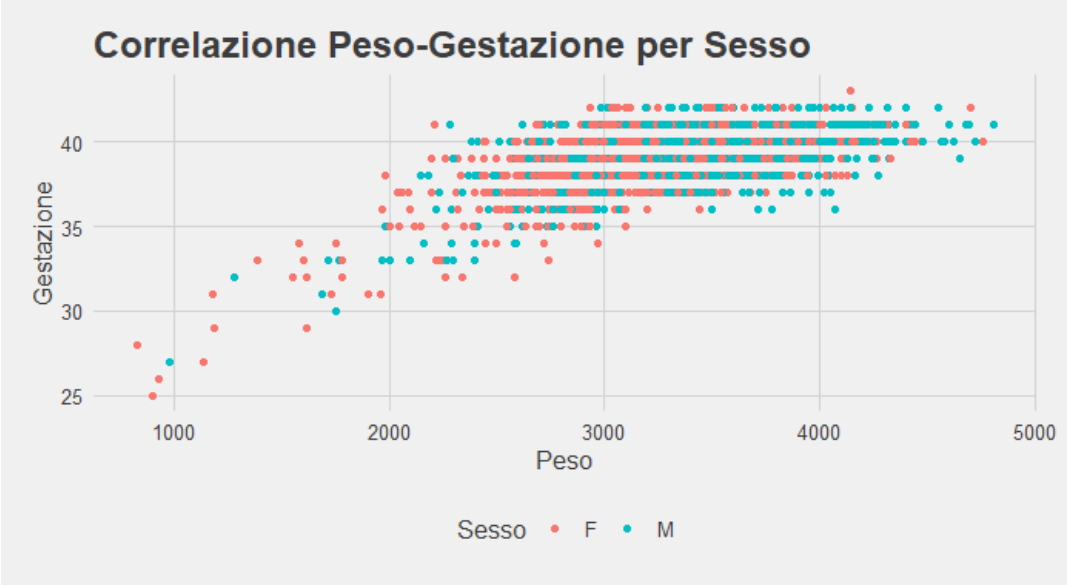


Anche in questo caso non si tratta di una correlazione lineare, facendo qualche test la soluzione migliore è quella logaritmica

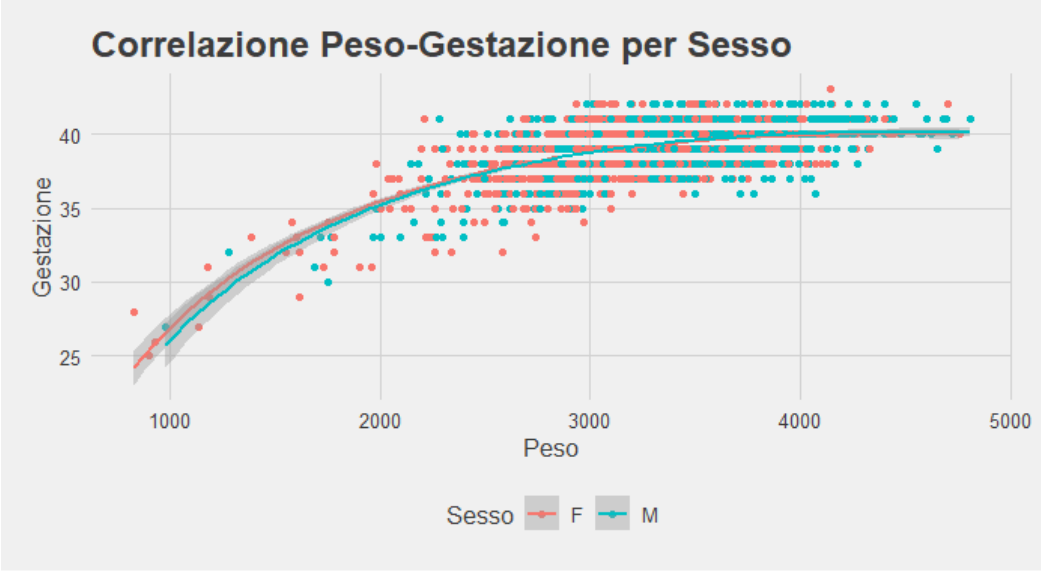


Nota: un altra crescita rilevante è quella 1/x

Osserviamo infine le settimane di Gestazione



Anche in questo caso la funzione di crescita più rilevante è quella logaritmica


Correlazione Peso-Gestazione per Sesso
Gestazione
40
35
30
25
1000
2000
3000
4000
5000
Peso
Sesso
F
M

Modelli

Prima di iniziare a trovare i modelli migliori dividiamo il dataset in training e un test set, cosi da poter testare le performance del nostro modello una volta addestrato.

Per scegliere il modello migliore utilizzeremo la procedura backward stepwise.
Partiamo quindi con un modello che contiene tutti i regressori e togliamo le variabili non rilevanti.
Il modello di partenza sarà il seguente:

```
Call:
lm(formula = Peso ~ Anni.madre + N.gravidanze + Gestazione + 
    Lunghezza + Cranio + Fumatrici + Tipo.parto + Ospedale + 
    Sesso, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-726.41 -175.75  -10.92  155.36 1026.74 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -6726.3248   145.7560 -46.148  < 2e-16 ***
Anni.madre        0.7737     1.0992   0.704 0.481546    
N.gravidanze     17.3595     4.9500   3.507 0.000462 ***
Gestazione       27.0645     3.8302   7.066 2.11e-12 ***
Lunghezza        11.1236     0.3065  36.297  < 2e-16 ***
Cranio            9.8510     0.4292  22.950  < 2e-16 ***
Fumatrici       -31.4392    26.3417  -1.194 0.232792    
Tipo.partoNat    28.9151    11.5085   2.512 0.012056 *  
Ospedaleosp2     -9.9850    12.7871  -0.781 0.434961    
Ospedaleosp3     24.1231    12.8430   1.878 0.060467 .  
SessoM           76.6249    10.6572   7.190 8.73e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 250.5 on 2294 degrees of freedom
Multiple R-squared:  0.7398,    Adjusted R-squared:  0.7386 
F-statistic: 652.1 on 10 and 2294 DF,  p-value: < 2.2e-16
```

Possiamo subito notare che il regressore Ospedale ha poca significatività statistica, verrà quindi rimosso.
Una scelta da compiere è quella se tenere o rimuovere i regressori Fumatrice e Anni.madre poiché hanno poca significativa statistica ma potrebbero essere ottime variabili di controllo.
Effettuando i test AIC e BIC sui modelli con e senza variabili notiamo che non abbiamo grandi differenze, inoltre anche l’R quadro non ne risulta significamene migliorato. Decido quindi per scelta personale di tenere il regressore Fumatrici come variabili di controllo e di rimuovere quello Anni.madre.
Il principale motivo è che quando abbiamo una fumatrice possiamo notare una riduzione del peso di 300 grammi, questo può senza dubbio aiutarci a effettuare previsioni più accurate.

Un altra cosa da fare è modificare i regressori del modello perché come abbiamo visto alcuni di questi hanno crescite logaritmica e non lineare. Modificheremo la variabile target e i regressori Gestazione, Lunghezza e Cranio trasformandoli in modo logaritmico.
Otterremo quindi il modello seguente:

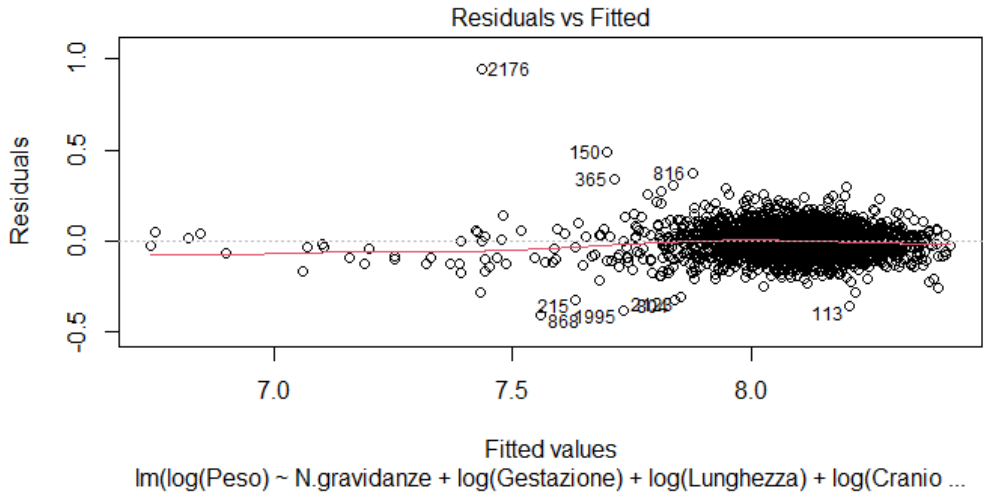
```
Call:
lm(formula = log(Peso) ~ N.gravidanze + log(Gestazione) + log(Lunghezza) +
    log(Cranio) + Tipo.parto + Sesso + Fumatrici, data = train)

Residuals:
      Min        1Q    Median        3Q       Max
-0.257669 -0.051121 -0.000371  0.050019  0.266190

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11.078341   0.219323 -50.512  < 2e-16 ***
N.gravidanze      0.006584   0.001386   4.750 2.16e-06 ***
log(Gestazione)   0.498720   0.043708  11.410  < 2e-16 ***
log(Lunghezza)    1.782339   0.045103  39.517  < 2e-16 ***
log(Cranio)       1.073771   0.043752  24.542  < 2e-16 ***
Tipo.partoNat     0.008063   0.003459   2.331   0.0198 *
SessoM            0.020525   0.003200   6.414 1.71e-10 ***
Fumatrici        -0.008406   0.007914  -1.062   0.2883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07529 on 2297 degrees of freedom
Multiple R-squared:  0.7903,    Adjusted R-squared:  0.7897
F-statistic:  1237 on 7 and 2297 DF,  p-value: < 2.2e-16
```

Come possiamo notare il punto R quadro e significativamente migliorato.

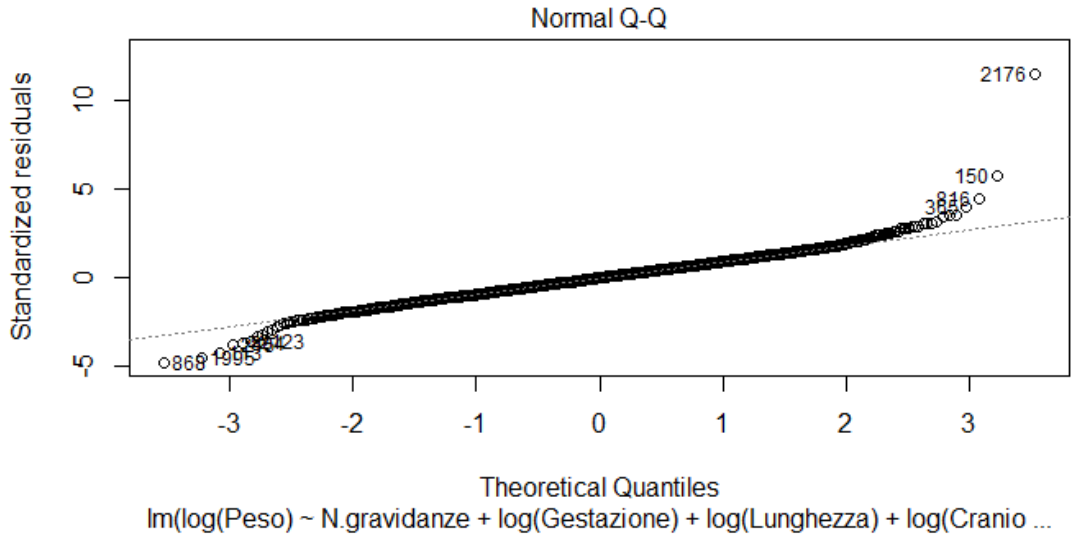
Questo risulta il nostro modello migliore fino a ora, il prossimo passo è vedere se il modello rispetta tutte le condizioni sui residui per poter essere affidabile.
Prima di tutto facciamo un test per verificare la presenza di multicollinearità, notiamo che non è presente multicollinearità.

Residui

Osserviamo i grafici dei residui:

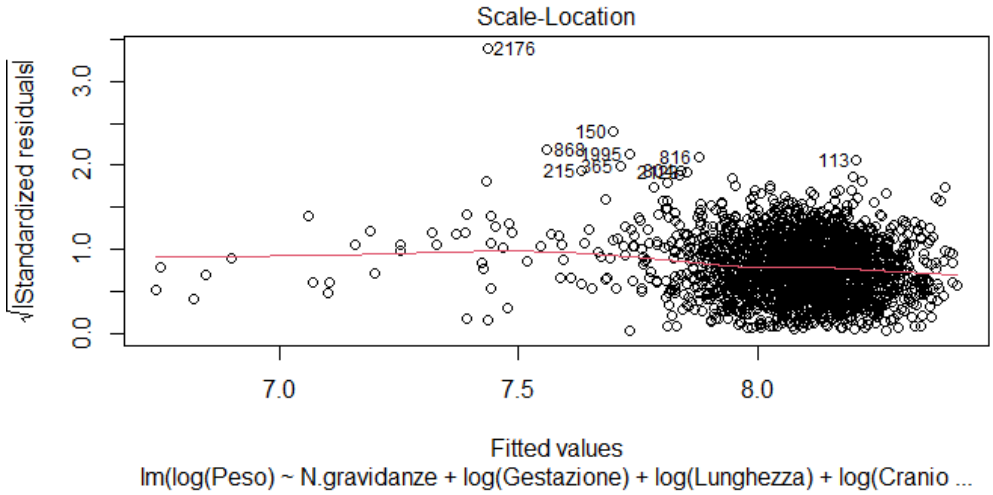


Da questo primo grafico notiamo che non abbiamo delle crescite nascoste dei dati e che i residui hanno media molto vicina allo 0. Questo ci fa be sperare per la normalità dei residui.
Vediamo il secondo grafico:

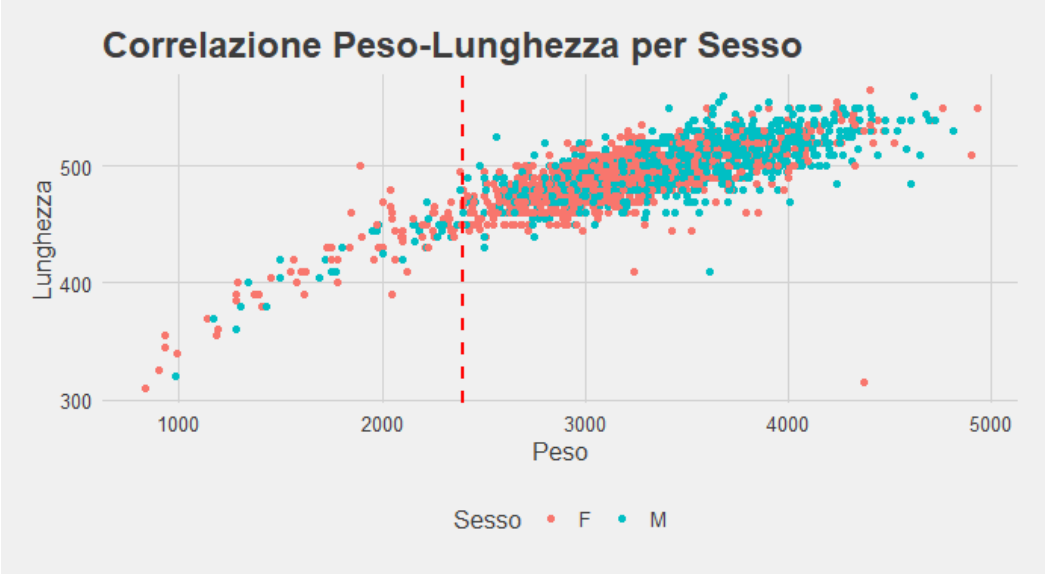


Possiamo notare che ci sono degli outlier che si discostano dalla linea teorica dei quantili, questo probabilmente riduce di molto la normalità dei residui, probabilmente aumentando l'asimmetria.
Effettuiamo uno shapiro test per verificare la normalità dei residui, purtroppo rifiutiamo l'ipotesi di normalità.

Osserviamo il terzo grafico:

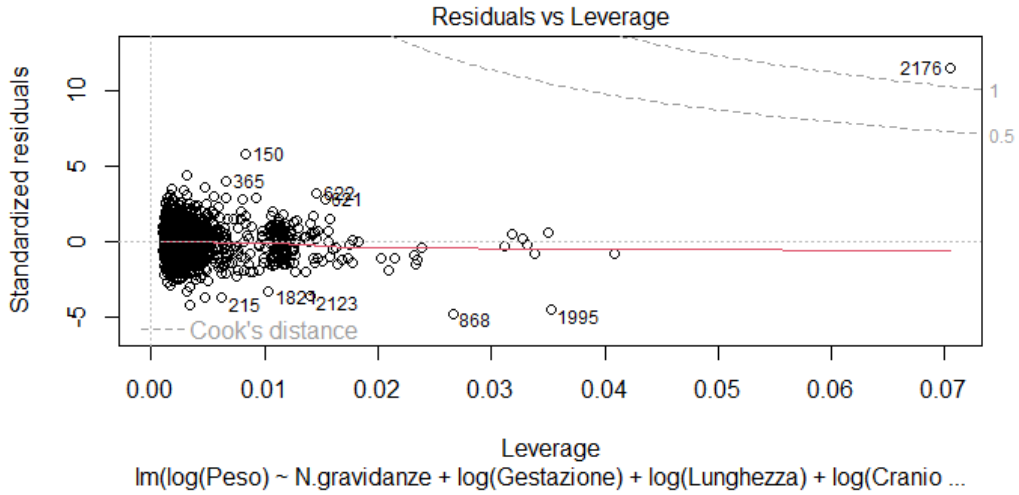


Purtroppo notiamo che la varianza non è uguale per tutti i residui, questo probabilmente causa eteroschedasticità.
Ma da cosa è data questa bassa varianza ?
Probabilmente da due fattori, il primo che le variabili Lunghezza, Cranio e Gestazione all'aumentare del peso hanno più registrazioni e quindi tendiamo ad avere una varianza squilibrata. Per fare un esempio vediamo la seguente immagine:

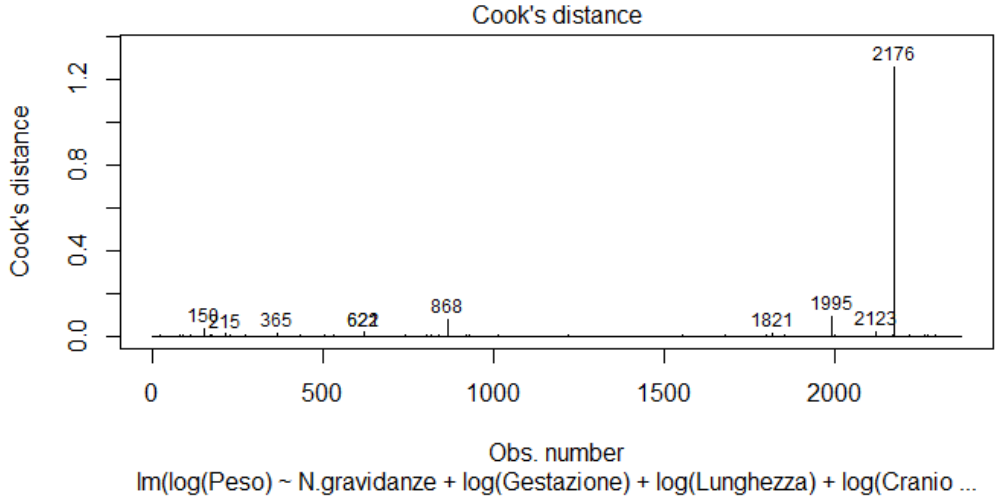


In questo caso a destra della linea rossa abbiamo molte più registrazioni che a sinistra, questo sposta la varianza. Un altra spiegazione è la presenza dei regressori Fumatrice, Tipo.Parto e Sesso che essendo variabili qualitative hanno pochissima varianza. Per verificare l'ipotesi di eteroschedasticità utilizziamo il Breusch–Pagan test. Purtroppo rifiutiamo l'ipotesi di omoschedasticità accertando cosi la presenza di eteroschedasticità.

Guardiamo ora le ultime due immagini relative agli outliers è al punteggio di cook:



In questo grafico possiamo notare i principali outliars, in particolare ce un valore che influenza molto il modello, osserviamo il secondo grafico per capire meglio.



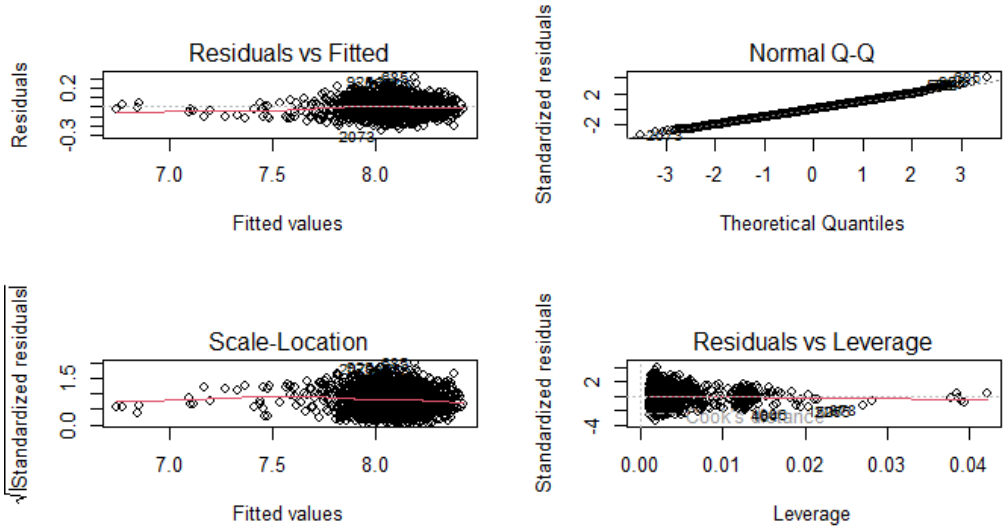
La registrazione numero 2176 influenza molto il nostro modello.
Per sicurezza effettuiamo il Durbin-Watson Test per verificare la presenza di Incorrelazione, il test conferma che il nostro modello non presenza incorrelazione.

Attualmente il modello ha diverse problematiche, per risolvere elimineremo diversi outlier cosi da ridurre il loro rumore e rendere il modello normale e omoschedastico.

Una volta rimossi punti più rumorosi e riaddestrato il modello presenta il seguente punteggio

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.25443 -0.05067  0.00027  0.05000  0.30482

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -11.368844   0.216488 -52.515  < 2e-16 ***
N.gravidanze     0.005841   0.001341   4.356 1.38e-05 ***
log(Gestazione)  0.528638   0.043495  12.154  < 2e-16 ***
log(Lunghezza)   1.765812   0.045095  39.158  < 2e-16 ***
log(Cranio)      1.122334   0.043870  25.583  < 2e-16 ***
Tipo.partoNat    0.008126   0.003488   2.330   0.0199 *
SessoM           0.021547   0.003214   6.704 2.54e-11 ***
Fumatrici       -0.007358   0.008138  -0.904   0.3660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07581 on 2304 degrees of freedom
Multiple R-squared:  0.7997,    Adjusted R-squared:  0.7991
F-statistic:  1314 on 7 and 2304 DF,  p-value: < 2.2e-16
```

Il punteggio R quadro è leggermente migliorato, inoltre ora i residui del modello seguono l’ipotesi di normalità, abbiamo anche ridotto l’eteroschedasticità senza però eliminarla del tutto. Purtroppo l’eteroschedasticità è contenuta nei dati stessi e quindi risulta particolarmente difficile rimuoverla dal modello.
Riosserviamo il grafico dei residui dopo la rimozione di diversi ouliers:



Predizioni

Testiamo ora il nostro modello.
Proviamo a dare un informazione al nostro modello e vedere la predizione che farà.

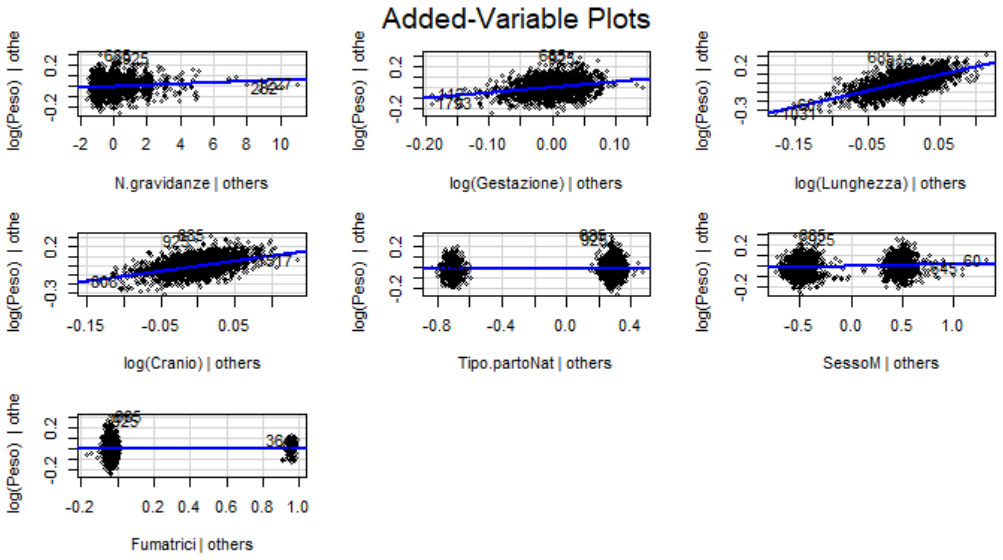
Ipotizziamo di avere *una madre alla terza gravidanza e che si stima partorirà alla 39 settimana.*
In queste informazioni parziali aggiungiamo le medie delle informazioni che ci mancano, come la lunghezza, il cranio ecc.
La predizione del nostro modello sarà uguale a 3263.714 grammi.
Ma non sapendo il peso corretto non sapremo di quanto si è sbagliato il nostro modello.
Per osservare effettivamente di quanto si sbaglia il nostro modello utilizziamo dataset di test che ci siamo messi da parte. Facciamo delle previsioni e calcoliamo delle metriche per capire come di quanto si sbaglia il nostro modello.
Come metriche utilizziamo:

- Root Mean Square Error: che misura l'errore medio eseguito dal modello nel prevedere l'esito di un'osservazione. → il nostro modello ha un RMSE di 287.4587
- Mean Absolute Error: la distanza assoluta media fra i due punti, simile al RMSE ma meno influnezato dagli outlier → il nostro modello ha un MAE di 223.2993

Facendo una media delle metriche avremo che il nostro modello si allontana dal valore coretto di circa 255.379 grammi, personalmente lo ritengo un buon punteggio.

Visualizzazione

Infine cerchiamo di dare una visualizzazione grafica del nostro modello, purtroppo essendo a più di 3 dimensioni non possiamo plottarlo. Un alternativa è quella di utilizzare i grafici di regressione parziale per vedere come la retta taglia i singoli regressori.



Questo grafico mostra come la nostra retta riesca a tagliare correttamente le variabili.