Un modello statistico per prevedere il peso dei neonati

Svolgi i punti uno alla volta e produci un **documento di testo word, pdf, HTML o markdown** in cui, per ogni punto, posso visualizzarne il codice (anche a blocchi), l'output di R e il tuo commento, spiegando ciò che hai fatto e il ragionamento.
Puoi consegnare anche il file.R per sicurezza, ma non deve essere obbligatorio da leggere per me per capire cosa hai fatto.

NOTA BENE: questo non è un progetto di programmazione, ma di statistica, e mi aspetto di leggere commenti e considerazioni statistiche per i vari passaggi e risultati!

Progetto:

Per questo studio medico si analizzano i dati raccolti da 3 ospedali, riguardanti 2500 neonati.

In particolare, si sono rilevate le seguenti variabili:

- età della madre
- numero di gravidanze sostenute
- Madre fumatrice (0=NO, SI=1)
- N° di settimane di gestazione
- peso in grammi del neonato
- Lunghezza in mm del neonato
- Diametro in mm del cranio del neonato
- Tipo di parto: Naturale o Cesareo
- Ospedale: 1, 2, 3
- Sesso del neonato: M o F

Si vuole scoprire se è possibile prevedere il peso del neonato alla nascita date tutte le altre variabili.

In particolare, si vuole studiare una relazione con le variabili della madre, per capire se queste hanno o meno un effetto significativo, come ad esempio l'effetto potenzialmente dannoso del fumo (nascite premature?). Si usano anche lunghezza e diametro del cranio del neonato perché si possono stimare già dalle ecografie, ma in generale potrebbero anche fungere da variabili di controllo.

Puoi seguire i punti che ti scrivo io come traccia e svolgili uno alla volta, commentando ovviamente i risultati.

1. **Importa il dataset "neonati.csv" e controlla che sia stato letto correttamente dal software**
2. **Descrivi il dataset, la sua composizione, il tipo di variabili e l'obiettivo dello studio**

Il dataset utilizzato in questo studio medico è costituito da dati raccolti da tre ospedali e riguarda un totale di 2500 neonati. Le variabili rilevate includono:

- **Età della madre**: Variabile quantitativa continua che rappresenta l'età della madre al momento del parto;
- **Numero di gravidanze sostenute**: Variabile quantitativa discreta che indica il numero di gravidanze che la madre ha avuto in passato;
- **Madre fumatrice**: Variabile qualitativa su scala nominale (dummy) che indica se la madre è una fumatrice o meno (0 per "No", 1 per "Sì");

- **Numero di settimane di gestazione**: Variabile quantitativa continua che rappresenta la durata della gestazione in settimane;
- **Peso in grammi del neonato**: Variabile quantitativa continua che indica il peso del neonato alla nascita, che è l'obiettivo principale dello studio;
- **Lunghezza in mm del neonato**: Variabile quantitativa continua che rappresenta la lunghezza del neonato alla nascita;
- **Diametro in mm del cranio del neonato**: Variabile quantitativa continua che indica il diametro del cranio del neonato alla nascita;
- **Tipo di parto**: Variabile qualitativa su scala nominale che indica se il parto è stato naturale o cesareo;
- **Ospedale**: Variabile qualitativa su scala nominale che indica da quale ospedale proviene il dato;
- **Sesso del neonato**: Variabile qualitativa su scala nominale (dummy) che indica il sesso del neonato (Maschio o Femmina).

L'obiettivo dello studio è determinare se è possibile prevedere il peso del neonato alla nascita utilizzando le altre variabili disponibili nel dataset. In particolare, si intende studiare la relazione tra le variabili della madre (come età, numero di gravidanze sostenute e se è fumatrice) per comprendere se queste hanno un effetto significativo sulla predizione del peso alla nascita. L'inclusione di variabili come la lunghezza e il diametro del cranio del neonato potrebbe fungere da variabili di controllo per tenere conto di fattori specifici del neonato che potrebbero influenzare il peso alla nascita. Infine, l'analisi include anche il tipo di parto, l'ospedale di provenienza e il sesso del neonato come variabili potenzialmente influenti.

3. Indaga le variabili effettuando una breve analisi descrittiva, utilizzando indici e strumenti grafici che conosci

Da un'osservazione delle distribuzioni delle frequenze della variabile *Anni.Madre,* si nota la presenza di due outliers, dovuti probabilmente a degli errori durante la raccolta dei dati: non è infatti realistico che la gravidanza avvenga all'età di 0 o 1 anno. È stata perciò definita la seguente tecnica di imputazione dei dati in questione: suddividere in classi la variabile *N.Gravidanze,* calcolare la media degli anni delle madri in ogni classe e sostituire i dati errati con il valore medio ottenuto in base alla classe di appartenenza.
Dai grafici a barre in Figura 1 si osserva che le madri del campione in considerazione sono prevalentemente (>95%) non fumatrici e più del 70% hanno partorito naturalmente. Frequenze percentuali simili sono invece registrate per il sesso dei neonati (dove osserviamo per entrambi i sessi percentuali vicine al 50%) e per l'ospedale scelto per il parto, essendo le frequenze in tutti e 3 i casi oscillanti tra il 30 e il 35%.

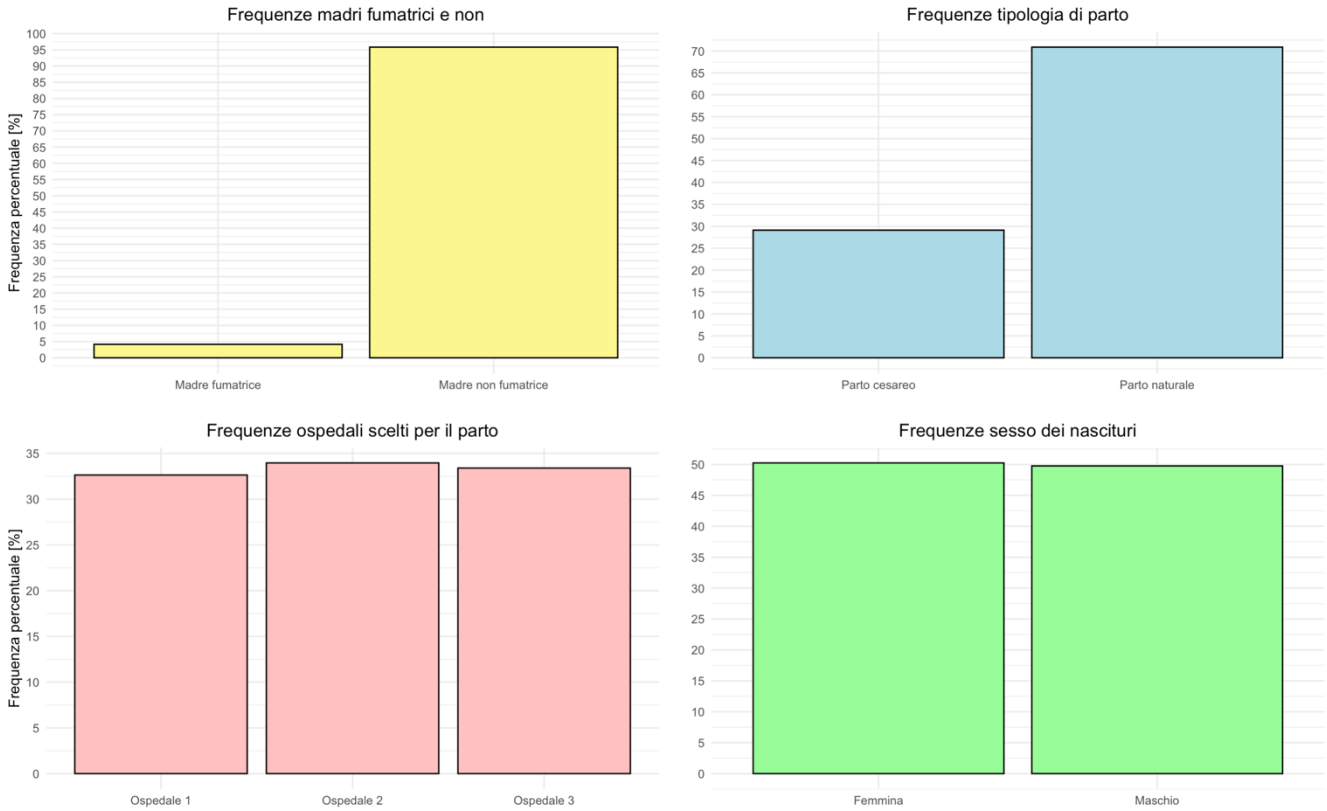

Figura 1. Frequenze percentuali delle variabili qualitative del dataset

I dati relativi alle caratteristiche dei neonati possono essere meglio descritti da dei boxplot che mettono a confronto le variabili tra i due sessi (Figura 2). In generale, vediamo come la posizione dei box sia tendenzialmente più alta per il sesso maschile per tutte e tre le variabili: *Lunghezza*, *Peso*, *Cranio*.
Per tutte e tre le variabili si osservano solo delle leggere asimmetrie nelle distribuzioni, in particolar modo per la variabile *Lunghezza* e per la variabile *Cranio* per il sesso femminile. Nelle altre variabili la mediana taglia invece a metà i box, per cui le distribuzioni possono definirsi simmetriche.
In tutti e 3 i grafici si nota la presenza di outliers, per lo più presenti nella parte bassa dei boxplot: è possibile formulare l'ipotesi che le variabili *Peso*, *Lunghezza* e *Diametro* siano tra loro correlate e che quindi i valori siano coerenti tra loro e, probabilmente, anche con la variabile *Gestazione*, indicante le settimane, appunto, di gestazione, indice di eventuale nascita prematura del neonato. Dal grafico a barre delle frequenze assolute delle settimane di gestazione (Figura 3) si vede come infatti siano presenti dei dati a favore di quanto ipotizzato, se si considera prematuro ogni parto avvenuto prima della 37ª settimana. Queste ipotesi puramente intuitive e dovute a delle prime osservazioni grafiche verranno verificate nel corso dell'elaborato.
In ultimo, il grafico a barre rappresentante la frequenza di madri all'interno del campione con delle gravidanze alle spalle, ci mostra come la maggior parte di esse risulta non averne, e solo una piccolissima percentuale di averne sostenute più di 7.

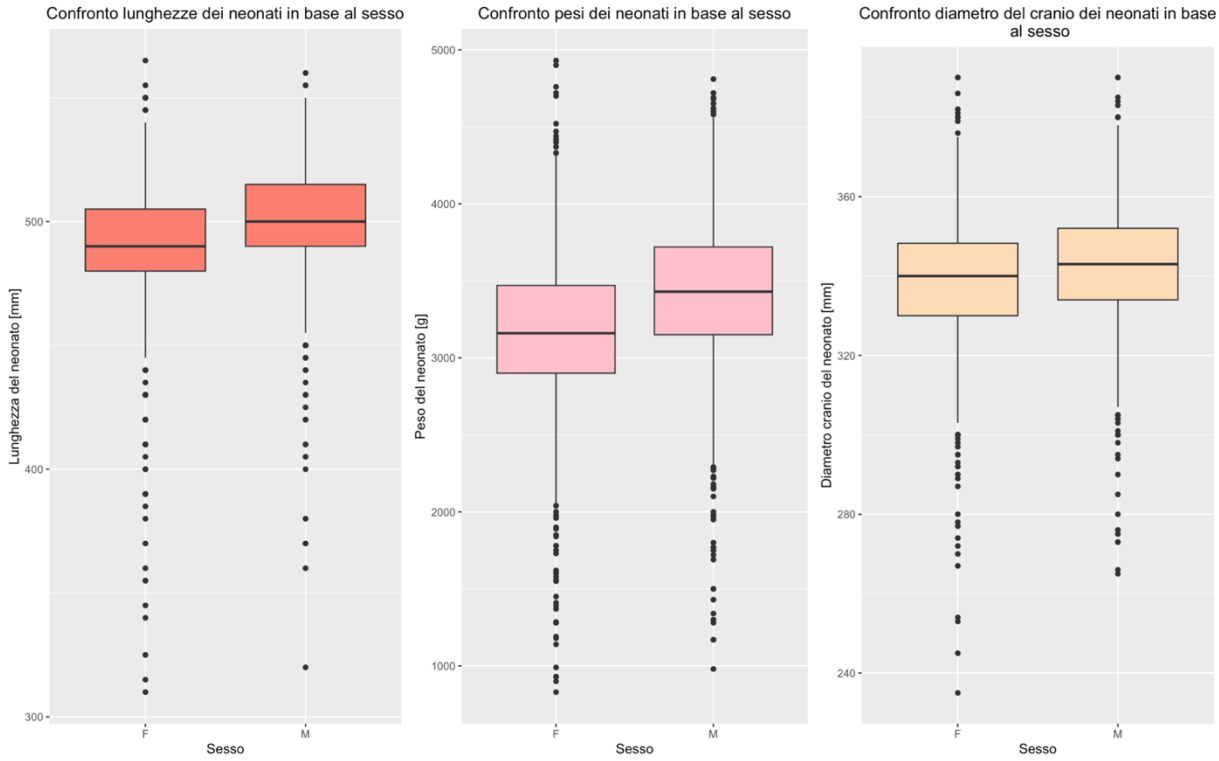

Figura 2. Boxplot di relazione tra Lunghezza, Peso e Diametro del cranio con il Sesso del neonato

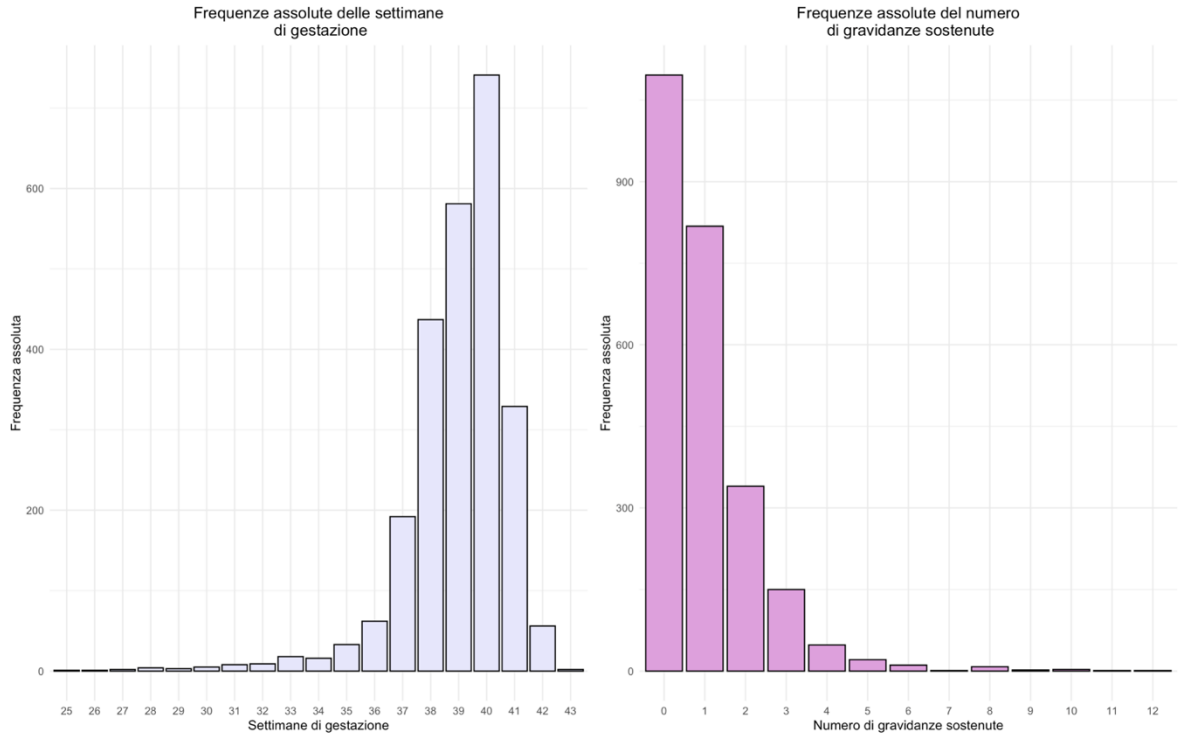

Figura 3. Frequenze assolute delle settimane di gestazione e del numero di gravidanze sostenute

4. **Saggia l'ipotesi che la media del peso e della lunghezza di questo campione di neonati siano significativamente uguali a quelle della popolazione**

Per decidere il test statistico da effettuare per poter saggiare l'ipotesi che la media del peso e della lunghezza di questo campione di neonati siano significativamente uguali a quelle della popolazione, è stata verificata la normalità della distribuzione dei dati di interesse.
Andando a rappresentare i grafici Q-Q delle due variabili (Figura 4), notiamo come in entrambi i casi essi si discostino significativamente dalla bisettrice del quadrante, suggerendo la non-normalità delle distribuzioni. Infatti, visualizzando i grafici delle distribuzioni di probabilità (Figura 5), la non-normalità è visibile soprattutto per la non simmetria rispetto alla media (rappresentata in rosso) e per la presenza di code più lunghe su uno dei due lati delle campane.

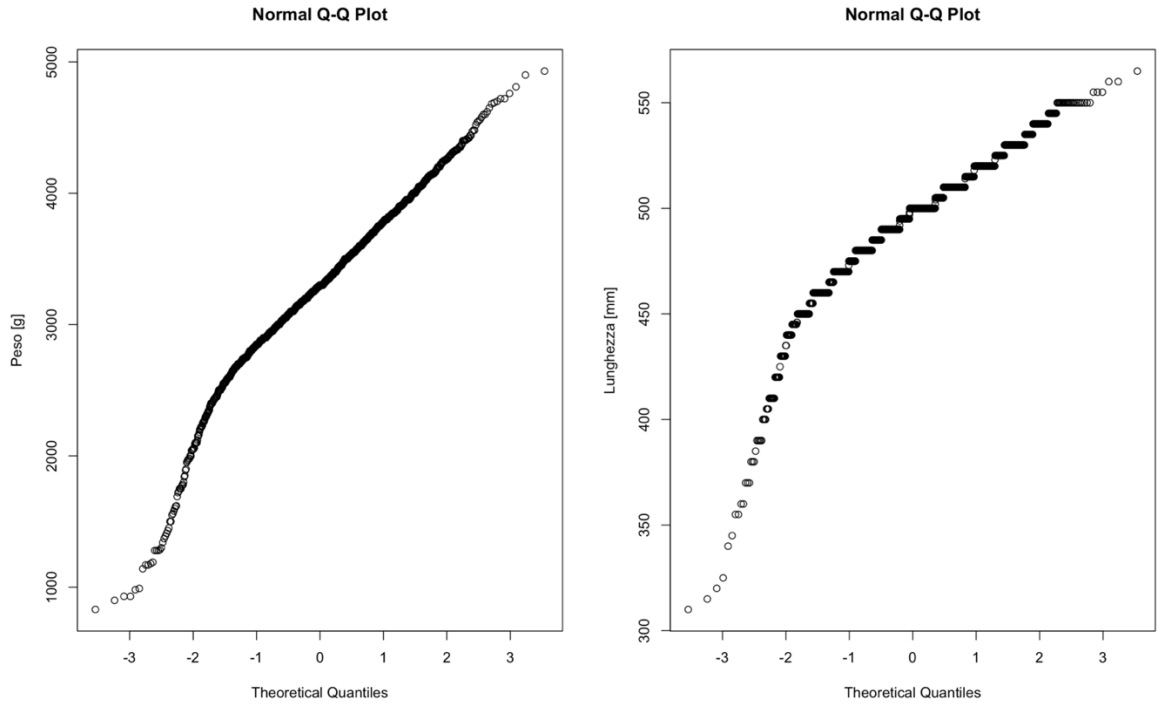

Figura 4. Q-Q plots di Peso e Lunghezza

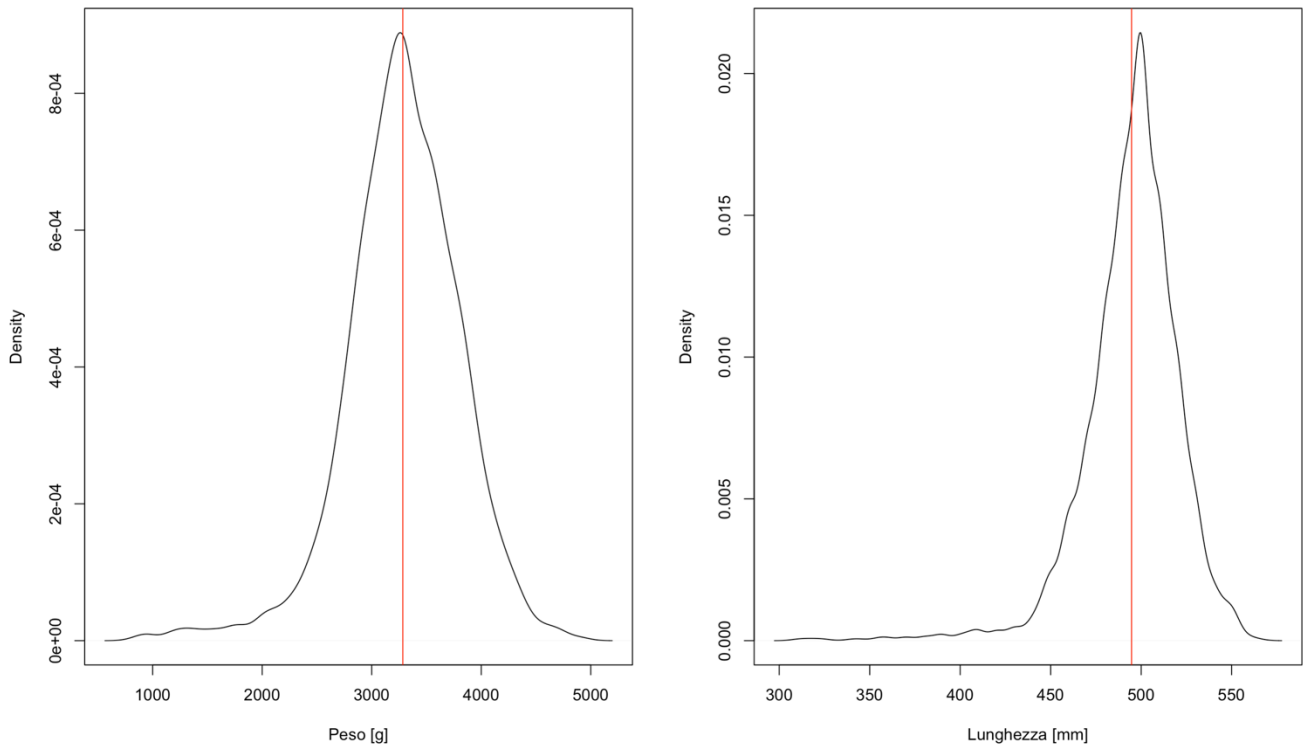

Figura 5. Densità di probabilità di Peso e Lunghezza

Di conseguenza, si rende maggiormente opportuno l'utilizzo del test dei ranghi di Wilcoxon-Mann-Whitney, per cui la normalità non risulta essere una condizione necessaria.

Dalla letteratura si apprende che il peso medio di un neonato è di 3300 g [1], mentre la lunghezza media è di 500 mm [2]. Di conseguenze, le ipotesi da saggiare diventano della forma:

- $\begin{cases} H_0 = 3300 \\ H_0 \neq 3300 \end{cases}$, per la variabile *Peso*;

- $\begin{cases} H_0 = 500 \\ H_0 \neq 500 \end{cases}$, per la variabile *Lunghezza*.

Il test svolto sulla variabile *Peso* restituisce un valore p di 0.9612, valore maggiore di 0.05, che non permette il rifiuto dell'ipotesi nulla: ciò suggerisce che vi è una probabilità molto alta di osservare un risultato simile a quello osservato anche se non ci fosse alcuna differenza significativa tra il campione e la media della popolazione.

Ciò non avviene invece per la variabile *Lunghezza*: in questo caso il p-value risulta essere tendente allo zero, quindi molto minore di 0.05, suggerendo che se l'ipotesi nulla fosse vera (cioè se non ci fossero differenze reali tra la lunghezza e il valore di riferimento di 500), la probabilità di ottenere i dati osservati sarebbe praticamente zero. Il risultato è statisticamente significativo e si può rigettare l'ipotesi nulla in favore dell'ipotesi alternativa, che suggerisce che le lunghezze osservate sono significativamente diverse da 500 mm.

[1] https://www.ospedalebambinogesu.it/da-0-a-30-giorni-come-si-presenta-e-come-cresce-80012/#:~:text=In%20media%20il%20peso%20nascita,pari%20mediamente%20a%2050%20centimetri.
[2] https://www.ospedalebambinogesu.it/da-0-a-30-giorni-come-si-presenta-e-come-cresce-80012/#:~:text=In%20media%20il%20peso%20nascita,pari%20mediamente%20a%2050%20centimetri.

5. **Per le stesse variabili, o per altre per le quali ha senso farlo, verifica differenze significative tra i due sessi**

Analogamente a quanto osservato per la variabile *Lunghezza* al punto 4, andando ad effettuare il test di Wilcoxon-Mann-Whitney per *Peso* e *Lunghezza* per verificare la presenza di differenze significative tra i due sessi, il p-value molto inferiore a 0.05 suggerisce il rifiuto delle ipotesi nulle di uguaglianza dei valori medi tra i due sessi.
Nella casistica descritta, poiché si parla di confronto di valori medi tra gruppi, è possibile anche l'utilizzo del t-test anche se le distribuzioni dei dati non hanno un andamento normale. Anche il t-test, però, ci porta allo stesso risultato indicato dal Wilcoxon-Mann-Whitney test sia per la variabile *Peso* che per la variabile *Lunghezza:* in entrambi i casi i p-value sono valori tendenti allo zero, indicando la presenza di effettiva differenza significativa tra le medie dei valori tra i due sessi.

6. **Si vocifera che in alcuni ospedali si facciano più parti cesarei, sai verificare questa ipotesi?**

Per verificare se in alcuni ospedali sia praticato più spesso il cesareo rispetto ad altri, è possibile effettuare un pairwise t-test, passando in input alla funzione la frequenza di ogni tipologia di parto in ogni ospedale e tutte le tipologie di ospedali. Dall'output otteniamo i p-value per ogni confronto: notiamo come in ogni combinazione il p-value sia sempre 1 (Tabella 2), quindi un valore che ci porta ad accettare l'ipotesi nulla dell'uguaglianza tra medie dei numeri di parti cesarei. Di conseguenza, non risulta essere vera l'affermazione che in alcuni ospedali si facciano più parti cesarei. Il risultato del test suggerisce quindi che quanto osservabile dal campione in esame può essere generalizzabile: calcolando infatti le frequenze di ogni tipo di parto notiamo come le percentuali siano molto simili (Tabella 1).

Tabella 1. Percentuale di parti cesarei in ogni ospedale del dataset

Ospedale	% Cesarei
Osp1	29.66 %
Osp2	29.91 %
Osp3	27.78%

Tabella 2. P-value ottenuti dal pairwise t-test

P-value	Osp1	Osp2
Osp2	1	-
Osp3	1	1

Analisi multidimensionale:

1. **Ricordati qual è l'obiettivo dello studio e indaga le relazioni a due a due, soprattutto con la variabile risposta**

Lo studio si pone l'obiettivo di prevedere il peso di un neonato alla nascita date tutte le altre variabili, in particolar modo quelle relative alla madre**.** Andando a rappresentare graficamente le relazioni a due a due della variabile

risposta, quindi il peso del neonato, con tutte le altre variabili, notiamo come coefficienti di Pearson elevati si ottengano per *Peso-Lunghezza* (0.80) e *Peso-Cranio* (0.70), mentre una correlazione media per *Peso-Gestazione* (0.59), e una correlazione quasi nulla per tutte le altre variabili. Infatti, osservando i relativi scatterplot del *Peso* e delle variabili precedentemente citate, la forte correlazione positiva è facilmente identificabile, essendo l'andamento della nuvola di punti crescente. Per ciò che concerne invece la relazione del *Peso* con la variabile *Gestazione*, graficamente si vede come la correlazione tra le due, seppur positiva e maggiore di 0.5, non è lineare, ma è presente una relazione quadratica, la cui eventuale influenza nel modello che verrà creato per la predizione del peso verrà indagata nei successivi paragrafi.

Inoltre, come ipotizzato al punto 3 dell'elaborato, esiste una certa correlazione positiva anche tra le variabili *Lunghezza* e *Cranio* stesse, o di queste ultime con la variabile *Gestazione*: di conseguenza, la presenza di valori esterni ai box, come visibile dai boxplot di figura 2, potrebbe essere spiegata da questo effetto e non trattarsi quindi di veri e propri outliers.

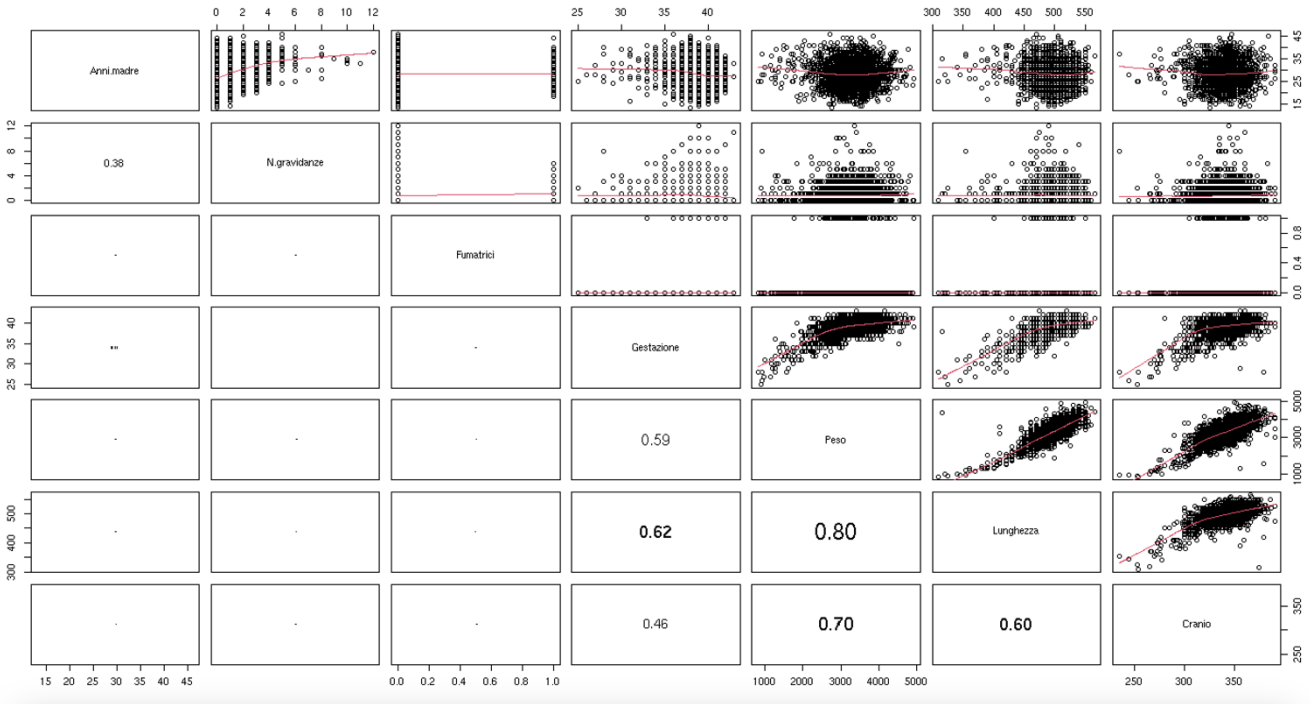

Figura 6. Scatterplot e coefficienti di Pearson tra le variabili del dataset

2. Crea un modello di regressione lineare multipla con tutte le variabili e commenta i coefficienti e il risultato ottenuto

Per permettere la predizione del peso di un neonato è stato predisposto un primo modello di regressione lineare tenente in considerazione il contributo di tutte le variabili contenute nel dataset, che ha prodotto i risultati visibili in tabella 3. Particolare attenzione deve essere posta verso la colonna 2 e 5, contenenti rispettivamente le informazioni relative ai valori marginali di ogni singola variabile sulla variabile risposta (il peso del neonato) e il relativo p-value, indice di significatività statistica.
Fissando la soglia del p-value a 0.05, si possono escludere le righe il cui contributo risulta essere non significativo, ovvero i valori di beta per i quali il p-value risulta essere maggiore del valore fissato, vale a dire gli anni della madre, l'indicatore di madre fumatrice e l'indicatore del parto avvenuto nell'ospedale numero 2 rispetto all'1.

Contributi molto significativi (quindi con p-value molto minore di 0.05, tendente a zero) sono invece dati da:

- La variabile indicante le settimane di gestazione: beta assume un valore positivo, a significare che la variabile in considerazione ha un effetto di incremento sulla variabile risposta: per ogni settimana di gestazione in più si causa infatti un aumento del peso del neonato di circa 32.6 g;
- Le variabili indicanti la lunghezza del neonato e il diametro del cranio: anche in questo caso il segno dei valori di beta è positivo, per cui per ogni cm in più di lunghezza del bambino o di diametro del cranio, si ha un incremento del peso di quest'ultimo di circa 10 g;
- Il sesso del neonato. Considerando come variabile di riferimento il sesso "Femmina", il sesso "Maschio" risulta avere causare un aumento del peso del neonato di circa 77.5 g rispetto all'altro.

Valori maggiori del p-value, ma pur sempre significativi, sono invece risultati dal numero di gravidanze precedenti della madre, dalla variabile indicante il tipo di parto effettuato e dalla variabile della scelta dell'ospedale 3 rispetto all'1.
Anche in questi casi, la correlazione tra le variabili citate e la variabile risposta è positiva: ogni gravidanza in più avuta in precedenza può essere causa di un aumento di circa 11.3g in più del peso del neonato, mentre la scelta del parto naturale un aumento del peso di 29.5g rispetto alla scelta del cesareo.

Tabella 3. Statistiche relative al modello 1

Mod1	Estimate (beta) [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6738.47	141.30	-47.68	0	***	
Anni.madre	0.89	1.132	0.787	0.430		
N.gravidanze	11.26	4.66	2.417	0.015	*	
Fumatrici	-30.163	27.53	-1.095	0.273		
Gestazione	32.56	3.818	8.529	0	***	
Lunghezza	10.29	0.300	34.23	0	***	0.7278
Cranio	10.47	0.426	24.57	0	***	
Tipo.partoNat	29.52	12.08	2.443	0.0146	*	
Ospedaleosp2	-11.20	13.43	-0.834	0.404		
Ospedaleosp3	28.09	13.49	2.081	0.0374	*	
SessoM	77.54	11.17	6.93	0	***	

3. Cerca il modello "migliore", utilizzando tutti i criteri di selezione che conosci e spiegali.

Per decidere il tipo di modello da utilizzare, sono stati calcolati asimmetria e indice di curtosi della variabile risposta (Peso), che risultano essere rispettivamente -0.6470 e 2.031: ci si aspetta quindi una distribuzione asimmetrica negativa e leptocurtica. Lo Shapiro test per la verifica dell'ipotesi nulla di normalità produce un p-value tendente allo zero, andando a confermare la non normalità della distribuzione, come si può anche osservare dalla figura 7:

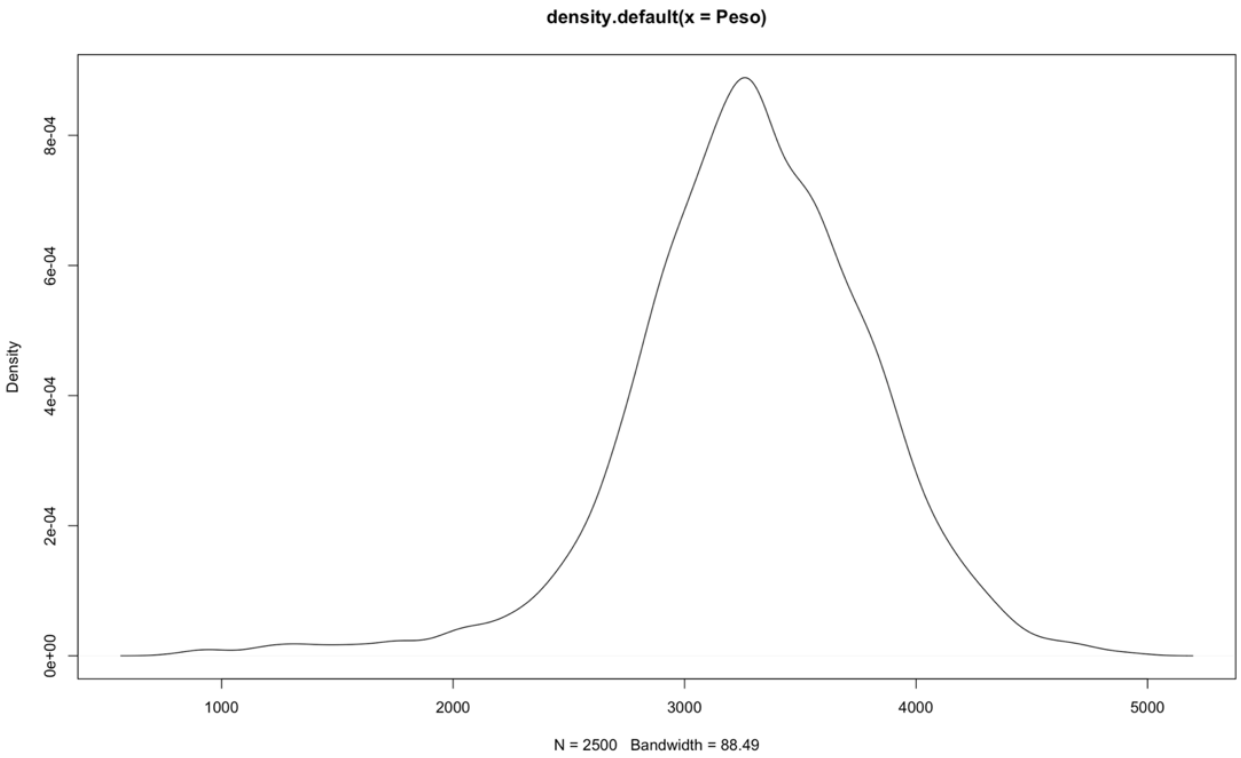

Figura 7. Densità di probabilità della variabile risposta (Peso)

Le distribuzioni note per essere utilizzate per dati con distribuzioni asimmetriche non risultano essere adatte in questo caso, in quanto, ad esempio, la gaussiana inversa o la distribuzione gamma approssimano bene distribuzioni asimmetriche positive. Pertanto, nonostante i risultati precedentemente mostrati, la gaussiana risulta essere il miglior compromesso per costruire un modello che rappresenti al meglio i dati a disposizione. Non si ricorre quindi a modelli lineari generalizzati.

Essendo *Anni.Madre* una delle variabili non significative, il modello 2 è stato costruito non considerandola. Si nota come non risultino esserci variazioni in termini di significatività delle altre variabili, nonché uno spostamento nullo del valore R^2_{adj} . Si può dunque concludere che la variabile *Anni.Madre* non è importante in termini di prestazioni del modello, ma, al contrario, superflua, dunque è possibile non considerarla.

Tabella 4 Statistiche relative al modello 2

Mod2	Estimate [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6708.107	135.939	-49.346	0	***	
N.gravidanze	12.608	4.338	2.906	0.0037	**	
Fumatrici	-30.309	27.536	-1.101	0.2711		
Gestazione	32.25	3.797	8.494	0	***	0.7278
Lunghezza	10.294	301	34.239	0	***	
Cranio	10.488	425	24.651	0	***	
Tipo.partoNat	29.535	12.083	2.444	0.0146	*	

Ospedaleosp2	-11.082	13.436	-825	0.4096	
Ospedaleosp3	28.366	13.49	2.103	0.0356	*
SessoM	77.621	11.176	6.945	0	***

Osservando le variabili non significative, si procede eliminando dal modello la variabile *Fumatrici*: ancora una volta $\boldsymbol{R^2_{adj}}$ non subisce variazioni. I valori delle stime variano poco percettibilmente, con incrementi o decrementi dell'ordine dei decimali, ma mantenendo sempre la stessa significatività dei casi precedenti. Tutto ciò indica che un modello privo della variabile *Fumatrici* risulta essere buono al pari dei modelli precedenti.

Tabella 5. Statistiche relative al modello 3

Mod3	Estimate [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6707.429	135.944	-49.34	0	***	
N.gravidanze	12.362	4.333	2.853	0.0044	**	
Gestazione	31.991	3.79	8.442	0	***	
Lunghezza	10.309	0.3	34.316	0	***	
Cranio	10.492	425	24.661	0	***	0.7278
Tipo.partoNat	29.28	12.082	2.424	0.0154	*	
Ospedaleosp2	-11.023	13.436	-0.82	0.4121		
Ospedaleosp3	28.641	13.489	2.123	0.0338	*	
SessoM	77.441	11.176	6.93	0	**	

Osservando il modello 3, la variabile con il contributo meno significativo risulta essere *Tipo.parto,* per cui il modello 4 è stato costruito omettendola. Di nuovo, ciò non causa grandi variazioni in termini di valori stimati e della rispettiva significatività, ma si nota una lieve diminuzione nel valore di $\boldsymbol{R^2_{adj}}$, che da 0.7278 (Tabella 5) diventa 0.7273 (Tabella 6). Sebbene il verificarsi del decremento dovrebbe essere un'indicazione di non procedere con il modello ottenuto, essendo la variazione avvenuta per valori minori dell'ordine di grandezza dei millesimi, è possibile non considerarla determinante ma continuare momentaneamente a considerare valido il modello privo della variabile *Tipo.parto* per procedere con l'analisi necessaria per la selezione del modello migliore.

Tabella 6. Statistiche relative al modello 4

Mod4	Estimate [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6682.009	135.671	-49.252	0	***	0.7273
N.gravidanze	12.082	4.335	2.787	0.0054	**	
Gestazione	32.046	3.793	8.448	0	***	
Lunghezza	10.272	0.3	34.204	0	***	
Cranio	10.526	426	24.729	0	***	
Ospedaleosp2	-11.074	13.449	-823	0.4104		
Ospedaleosp3	29.199	13.5	2.163	0.0306	*	
SessoM	77.497	11.186	6.928	0	***	

Infatti, mettendo a confronto i 4 modelli creati fino ad ora, ognuno contenente un numero diverso di variabili, il criterio di informazione Bayesiano suggerisce il modello 4 come il modello con il BIC minore (35226.09, come osservabile in tabella 7), che lo rende, dunque, preferibile rispetto agli altri. Come già evidenziato in precedenza, il fatto che la rimozione della variabile *Tipo.parto* causasse una diminuzione del valore di R^2_{adj} non è condizione sufficiente per decidere di scartare il modello. Una visione d'insieme suggerisce infatti che una diminuzione così minima è accettabile in quanto conseguenza di una semplificazione del modello, fattore che contribuisce a renderlo più generalizzabile. Altri indici, infatti, come ad esempio il BIC, per l'appunto, lo indicano come preferibile.

Tabella 7. BIC dei modelli 1,2 3, e 4

	df	BIC
mod1	12	35241.84
mod2	11	35234.64
mod3	10	35228.03
mod4	**9**	**35226.09**

4. Si potrebbero considerare interazioni o effetti non lineari?

Appurato che il modello 4 risulta essere il migliore nel momento in cui si considerano solo effetti di tipo lineare, è possibile fare dei tentativi di modifica al modello per renderlo più preciso e considerare la presenza di particolari effetti dovuti ad alcune variabili.
Dagli scatterplot rappresentanti le relazioni della variabile risposta (*Peso)* con tutte le altre variabili del dataset (Figura 6), si nota principalmente una relazione quadratica con la variabile *Gestazione.*

Il modello 5 è quindi ottenuto tenendo in considerazione questo tipo di effetto:

Tabella 8. Statistiche relative al modello 5

Mod5	Estimate [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6106.745	123.543	-49.43	0	***	0.7276
N.gravidanze	12.162	4.333	2.807	0.005	**	
Lunghezza	10.285	298	34.472	0	***	
Cranio	10.546	425	24.817	0	***	
Ospedaleosp2	-10.835	13.443	-806	0.4203		
Ospedaleosp3	29.069	13.493	2.154	0.0313	*	
SessoM	76.846	11.183	6.872	0	***	
I(Gestazione^2)	0.434	0.05	8.594	0	***	

Considerando l'effetto quadratico della variabile *Gestazione*, si mantiene la significatività degli effetti lineari delle altre variabili e si percepisce un aumento, seppur minimo, del valore dell'R^2_{adj}. Il valore del coefficiente beta relativo alla variabile *Gestazione* indica che il cambia il tasso di crescita del peso del neonato per unità di cambiamento nelle settimane di gestazione, considerando anche l'effetto quadratico, è di 0.434, valore molto significativo. La relazione quadratica può quindi essere tenuta all'interno del modello.
Rispetto al modello 4, il modello 5 ha anche un indice BIC minore:

Tabella 9. BIC dei modelli 4 e 5

	df	BIC
Mod4	9	35226.09
Mod5	**9**	**35223.67**

Provando ad introdurre eventuali effetti di interazioni tra variabili, il modello 5 si modifica come segue.
La prima interazione considerata è quella tra la variabile *Fumatrici* e la variabile *Anni.madre*, precedentemente scartate in quanto non considerate aventi effetti significativi singolarmente. Poiché effettuando un test chi-quadro risultano essere indipendenti, è possibile che abbiano sulla variabile risposta un effetto congiunto da non ignorare.

Tabella 10. Statistiche relative al modello 6

Mod6	Estimate [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6128.273	127.638	-48.013	0	***	0.7274
N.gravidanze	10.991	4.662	2.357	0.0185	*	
Lunghezza	10.273	299	34.371	0	***	
Cranio	10.524	426	24.724	0	***	
Ospedaleosp2	-11.007	13.45	-0.818	0.4132		
Ospedaleosp3	28.527	13.503	2.113	0.0347	*	

SessoM	76.926	11.187	6.876	0	***
I(Gestazione^2)	0.441	51	8.676	0	***
Fumatrici	-36.652	146.58	-0.25	0.8026	
Anni.madre	0.921	1.155	798	0.4251	
Fumatrici:Anni.madre	0.28	5.087	0.055	0.956	

Si nota come anche considerando l'effetto di interazione tra le due variabili, l'effetto è ancora non significativo, con nessuna variazione positiva anche per $\boldsymbol{R^2_{adj}}$. Una situazione analoga si ottiene andando a considerare l'effetto congiunto di *Fumatrici* e *Sesso*, anch'esse indipendenti.
Un'ulteriore modifica al modello viene fatta rimuovendo la variabile *Ospedale*: essendo una variabile qualitativa, le stime vengono fatte prendendo come riferimento uno degli ospedali (il numero 1), ma considerarla all'interno del modello non lo renderebbe generalizzabile ad una popolazione più estesa del campione di riferimento, in quanto le analisi e predizioni sarebbero accurate solo per studi effettuati su neonati nati in uno dei tre ospedali. Pertanto, si può analizzare la bontà del nuovo modello 8:

Tabella 11. Statistiche relative al modello 8

Mod8	Estimate [g]	Std_Error	t_value	Pr_gt_t	Significance	R^2_{adj}
(Intercept)	-6100.641	123.545	-49.38	0	***	0.7267
N.gravidanze	12.554	4.338	2.894	0.0038	**	
Lunghezza	10.261	299	34.358	0	***	
Cranio	10.56	426	24.816	0	***	
SessoM	77.329	11.198	6.906	0	***	
I(Gestazione^2)	0.438	0.051	8.667	0	***	

Si nota come l'effetto di tutte le variabili risulta ora essere molto significativo, con una lieve diminuzione dell'$\boldsymbol{R^2_{adj}}$, a favore però di un modello più snello e generalizzabile. Anche il BIC lo conferma come il modello migliore tra quelli costruiti, con un valore di 35217.52:

Tabella 12. BIC dei modelli 5 e 8

	df	BIC
Mod5	9	35223.67
Mod8	**7**	**35217.52**

5. **Effettua una diagnostica approfondita dei residui del modello e di potenziali valori influenti. Se ne trovi prova a verificare la loro effettiva influenza**
6. **Quanto ti sembra buono il modello per fare previsioni?**

Effettuando i test di verifica delle ipotesi sui residui, avendo scelto il modello 8 come modello migliore, si nota come solo il test relativo alla non correlazione dei residui viene superato, con un p-value di 0.1191 che ci consente di accettare l'ipotesi nulla. Al contrario, le ipotesi nulle di omoschedasticità e normalità dei residui devono essere necessariamente rifiutate, in quanto i relativi test producono un p-value molto vicino allo zero.

Si rende dunque necessaria un'analisi approfondita riguardo l'eventuale presenza di valori outliers o di leva che potrebbero aver avuto effetto sui residui. Dai plot in figura 8 notiamo che i residui sono distribuiti randomicamente intorno allo 0 come desiderato, ma dal Q-Q plot (Figura 9) si nota come alcuni valori si discostino dalla bisettrice del quadrante, ad indicare la non normalità della distribuzione, come già emerso dal t-test.
Sebbene i grafici in figura 9 rappresentanti i valori di leva e gli outliers ci mostrino come i valori che superano le rispettive soglie siano molteplici, dal grafico rappresentante la distanza di Cook che mette in relazione residui e leverages vediamo come l'unica osservazione potenzialmente critica sia la 1551, in quanto all'interno del range di allarme tra 0.5 e 1. Effettuando invece un t-test degli outliers, otteniamo che i valori che sono considerati tali effettivamente sono, di nuovo, l'osservazione 1551, la 155 e la 1306, come visibile in tabella 13

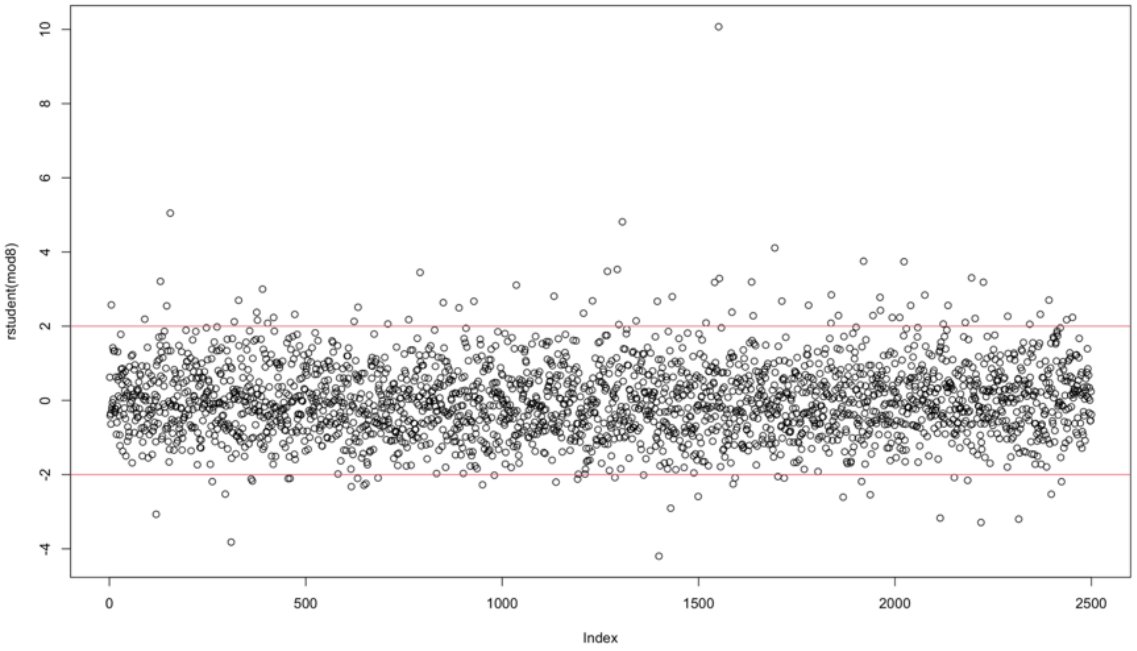

Figura 8. Residui - modello 8

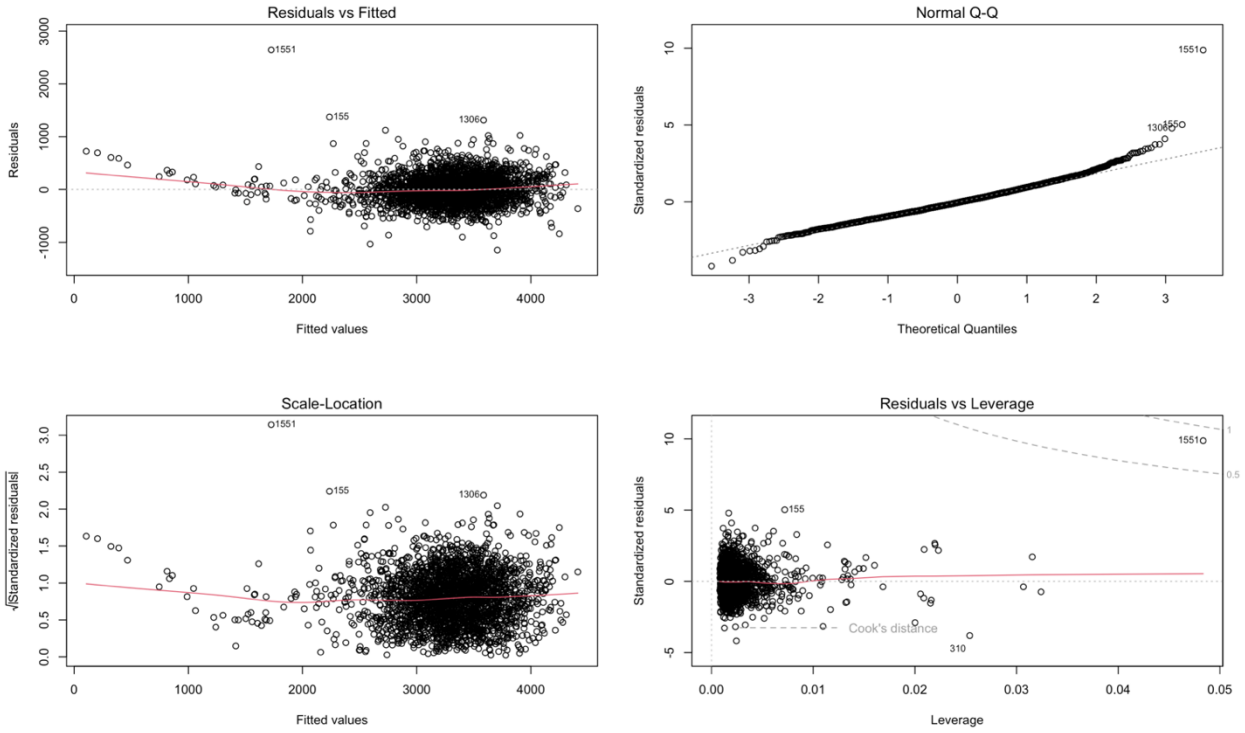

Figura 9.

Tabella 13.

Osservazione	Rstudent	p-value	Bonferroni p-value
1551	10.07	0	0
155	5.04	0	0.001
1306	4.81	0	0.003

Creando un secondo dataset privo delle osservazioni individuate come potenzialmente influenti, è stato costruito un secondo modello di regressione con le stesse caratteristiche del modello 8. Rieffettuando l'analisi dei residui, si ottengono ora verificate sia la condizione di non correlazione dei residui che la condizione di omoschedasticità, con dei p-value rispettivamente di 0.1517 e 0.1564. L'ipotesi nulla di normalità dei residui deve invece, ancora una volta, essere rifiutata, presentando un p-value tendente allo zero, risultato prevedibile in parte vista la non normalità della variabile risposta, che probabilmente si riflette sui residui.
Nonostante questa ultima condizione non ideale, 2 su 3 delle assunzioni sui residui sono verificate, per cui il modello può essere considerato abbastanza buono. Le 3 osservazioni rimosse risultavano dunque essere influenti sul modello. Tuttavia, la loro esclusione non comporta cambiamenti significativi in termini di bontà del modello, poiché tutte le variabili continuano ad essere altamente significative e l'R^2_{adj} viene leggermente incrementato a 0.74, rendendo il modello buono per fare previsioni.

7. **Fai la tua migliore previsione per il peso di una neonata, considerato che la madre è alla terza gravidanza e partorirà alla 39esima settimana. Niente misure dall'ecografia.**

Non avendo a disposizione le misure dall'ecografia, per rendere la predizione il più possibile accurata, i dati sono stati filtrati considerando solo le misure di lunghezza e diametro del cranio dei neonati di sesso femminile e nati alla 39esima settimana. Di tutti questi è stata calcolata la mediana, indice robusto e non soggetto ad oscillazioni dovute ad eventuali outliers. Dunque, i valori utilizzati per la lunghezza e per il cranio per la predizione del peso sono stati rispettivamente 495 cm e 340 cm, grazie ai quali, in aggiunta al valore di *Gestazione=39*, *N.Gravidanze=2 e Sesso="F",* il modello ha predetto un peso di 3261.11 g.

8. Cerca di creare qualche rappresentazione grafica che aiuti a visualizzare il modello. Se è il caso semplifica quest'ultimo!

Poiché il modello scelto tiene in considerazione 4 variabili, non è possibile rappresentarlo graficamente su un piano. Pertanto, per poterne visualizzare la bontà, è stata calcolata la predizione del peso per ogni punto del dataset ed è stata messa a confronto quest'ultima con i valori reali del peso a disposizione. Dunque, si ottiene lo scatterplot in figura 10. Dal grafico si nota come il modello sia maggiormente performante per pesi maggiori, mentre restituirà, con alta probabilità, risultati poco attendibili per predizioni inferiori ai 2000 g. Dal grafico si vede inoltre come la bontà del modello non dipenda dal sesso del neonato, in quanto, per valori bassi di peso, la retta di regressione si discosta dalla nuvola di punti indipendentemente dal fatto che questi rappresentino un neonato di sesso maschile o femminile.

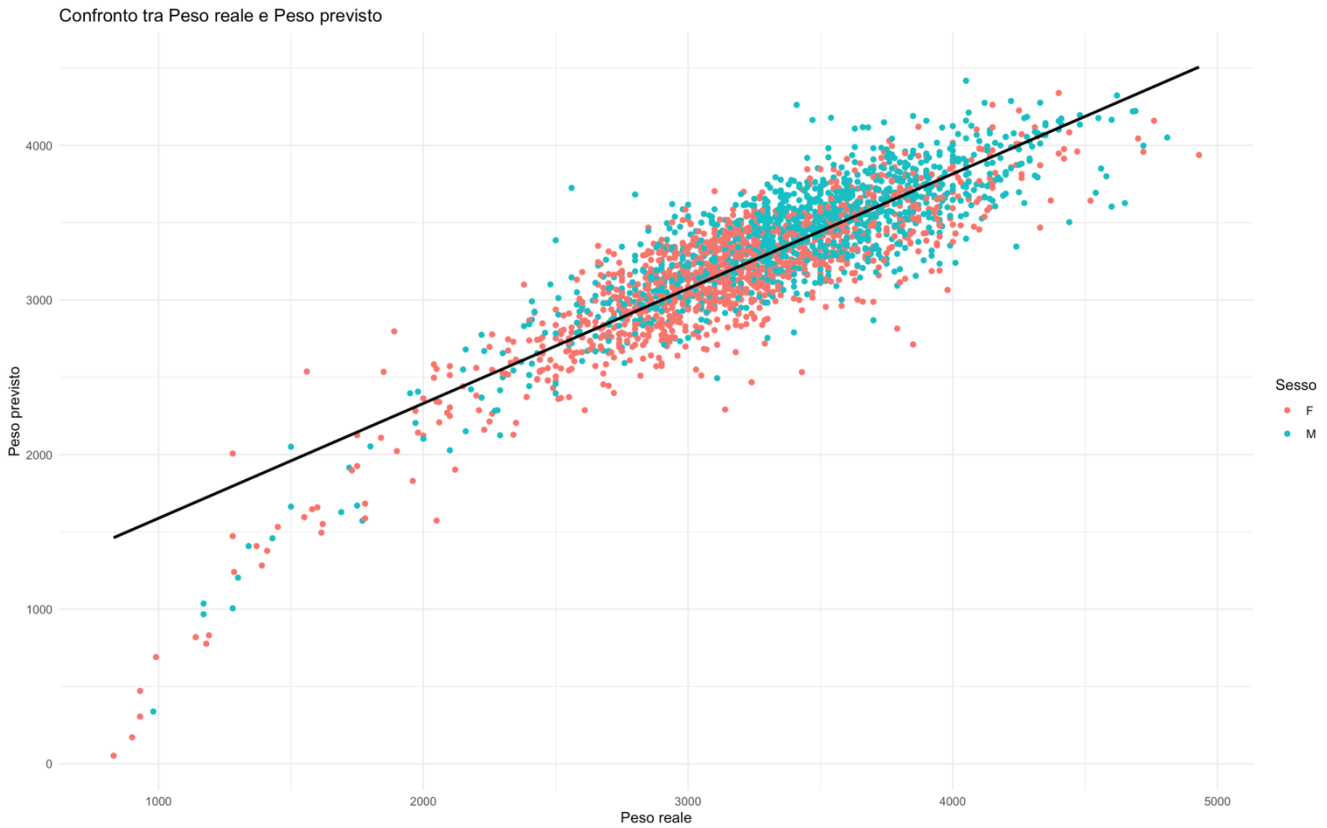

Figura 10. Rappresentazione del peso reale e del peso previsto dal modello in base al sesso del neonato

Osservando invece il grafico di Figura 11, rappresentante analogamente la relazione tra peso reale e peso predetto dal modello, ma senza distinzione per sesso, bensì per settimane di gestazione della madre, è evidente

come valori più accurati si ottengano per un numero maggiore di settimane di gestazione. Dunque, il modello non risulta essere adatto a predire il peso di neonati nati prematuramente (prima di 37 settimane), mentre approssima sufficientemente bene il peso di nascite avvenute in tempistiche normali. È possibile concludere che, sebbene il modello scelto non copra con sufficiente bontà tutte le possibili casistiche, è bene tenere a mente che la nascita prematura è un evento non ordinario, ma può dipendere da possibili complicazioni della gravidanza, dovute a loro volta da fattori vari, spesso differenti di caso in caso: risulta perciò difficile fare predizioni precise, in quanto sarebbero necessarie numerose informazioni (al momento non disponibili nel dataset corrente) per rendere il modello il più possibile generalizzabile.

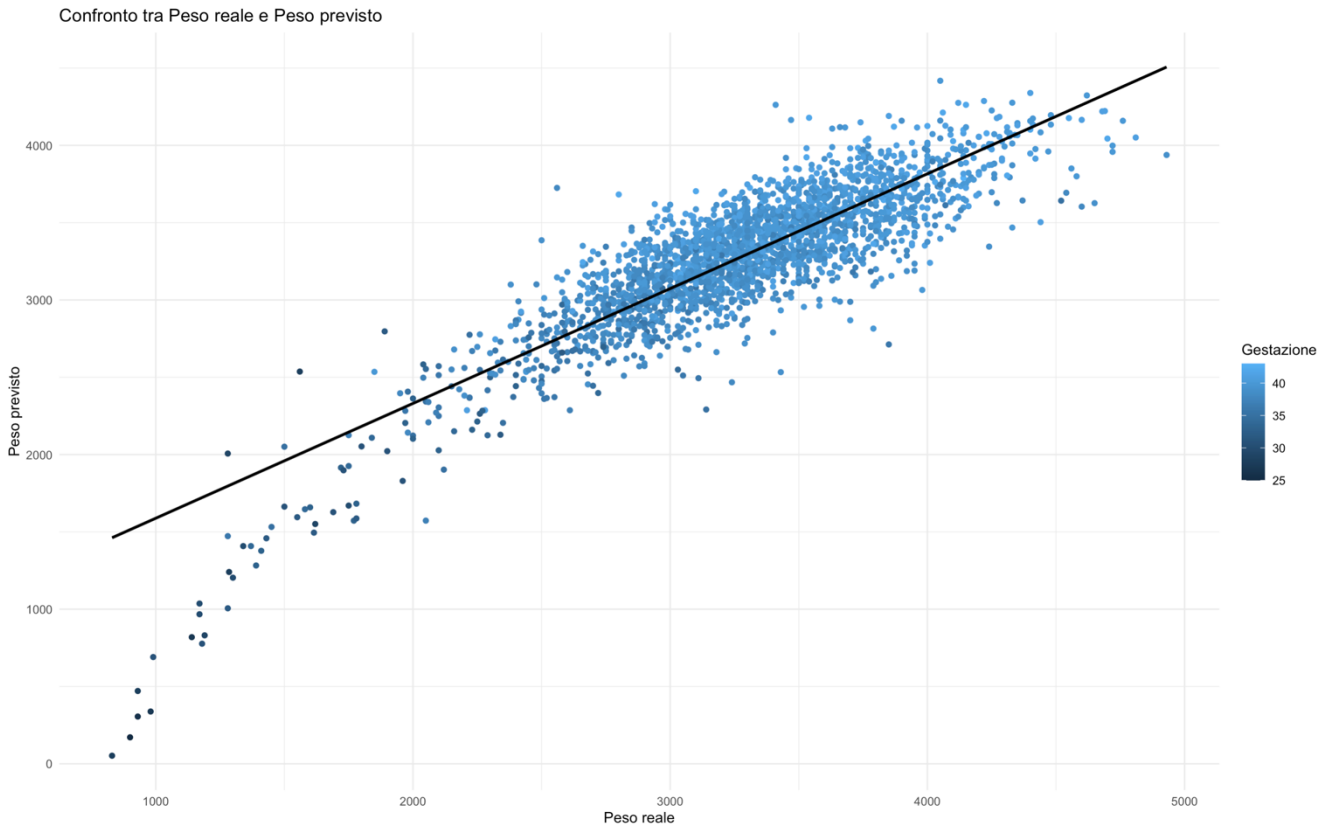

Figura 11. Rappresentazione del peso reale e del peso previsto dal modello in base al numero di settimane di gestazione della madre