Progetto Statistica Inferenziale

Andrea Ricciardelli

Luglio 2023

1 Dataset e obiettivo dello studio

1.1 Dataset

Il dataset è composto da dati medici raccolti da 3 ospedali, riguardanti 2500 neonati. Per ogni neonato sono state rilevate 10 variabili.

1.2 Obiettivo dello studio

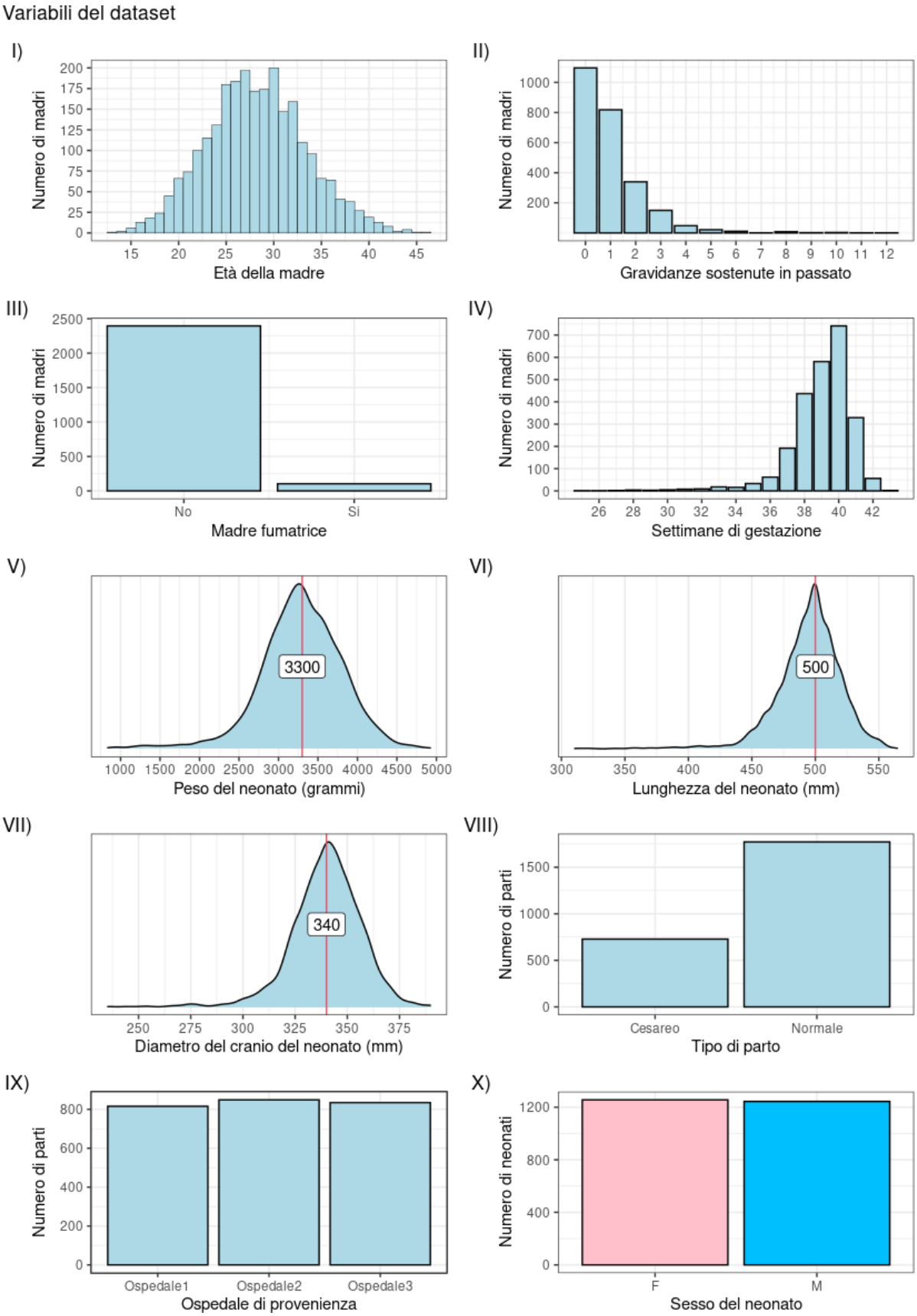
Si vuole scoprire se è possibile prevedere il peso del neonato alla nascita date tutte le altre variabili. In particolare, si vuole studiare una relazione con le variabili della madre per capire se queste hanno o meno un effetto significativo sul neonato (ad esempio, l'effetto potenzialmente dannoso del fumo potrebbe portare a nascite premature).

1.3 Variabili

Variabile	Descrizione	Tipologia
Anni.madre	Età della madre	var. quantitativa continua su scala di rapporti
N.gravidanze	Numero di gravidanze già sostenute	var. quantitativa discreta su scala di rapporti
Fumatrici	Se la madre è fumatrice o no	var qualitativa su scala nominale (codificata: 0=NO, 1=SI)
Gestazione	Numero di settimane di gestazione	var. quantitativa continua su scala di rapporti
Peso	Peso del neonato (in grammi)	var. quantitativa continua su scala di rapporti
Lunghezza	Lunghezza del neonato (in mm)	var. quantitativa continua su scala di rapporti
Cranio	Diametro del cranio del neonato (in mm)	var. quantitativa continua su scala di rapporti
Tipo.parto	Parto naturale o cesareo?	var. qualitativa su scala nominale (Nat/Ces)
Ospedale	Ospedale di provenienza	var. qualitativa su scala nominale (osp1/osp2/osp3)
Sesso	Sesso del neonato	var. qualitativa su scala nominale (M/F)

2 Analisi descrittiva

Qui vengono mostrati i grafici relativi a ciascuna variabile. Inoltre, per le variabili quantitative vengono riportati gli indici di posizione, dispersione, simmetria - e per le variabili qualitative, vengono riportate le tabelle di frequenza.



Variabile	Anni.madre	N.gravidanze	Gestazione	Peso	Lunghezza	Cranio
Min	0	0	25	830	310	235
Q1	25	0	38	2990	480	330
Mediana	28	1	39	3300	500	340
Media	28.1	0.98	38.98	3284.0	494.6	340.0
Q3	32	1	40	3620	510	350
Max	46	12	43	4930	565	390
Intervallo	46	12	18	4100	255	155
IQR	7	1	2	630	30	20
dev.st	5.27	1.28	1.87	525.0	26.3	16.4
var	27.8	1.64	3.49	275665.6	692.6	269.7
CV	18.7	130.5	4.79	15.9	5.32	4.83
Asimmetria	0.04	2.51	-2.07	-0.65	-1.51	-0.79
Curtosi	0.38	10.99	8.26	2.03	6.49	2.95

Indici statistici delle variabili quantititative del dataset

Fumatrici	frequenza
0 (No)	2396
1 (Sì)	104

Tipo.parto	frequenza
Ces	728
Nat	1772

Ospedale	frequenza
osp1	816
osp2	849
osp3	835

Sesso	frequenza
F	1256
M	1244

Tabelle di frequenza delle variabili qualitative del dataset

3 Saggiare un'ipotesi

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

3.1 Dati della popolazione

I dati relativi alla popolazione sono stati presi dall'Ospedale Pediatrico Bambino Gesù (importante centro di ricerca pediatrico italiano). Risulta che nella popolazione la media del peso dei neonati sia 3300 grammi, e la media della lunghezza dei neonati sia 500 millimetri.

3.2 Test t con campione singolo

```
t.test(Peso, mu = 3300, conf.level = 0.95, alternative = "two.sided")
t.test(Lunghezza, mu = 500, conf.level = 0.95, alternative = "two.sided")
```

Nel primo caso (peso del neonato) risulta un p-valore di 0.13, quindi per un livello di significatività $\alpha = 0.05$ non rifiuto l'ipotesi nulla. Concludo che la lunghezza di questo campione di neonati non è significativamente diversa da quella della popolazione.
Nel secondo caso (lunghezza del neonato) risulta un p-valore minuscolo (dell'ordine di 10^{-16}), quindi rifiuto l'ipotesi nulla. Concludo che il peso di questo campione di neonati è significativamente diverso da quello della popolazione.

4 Differenze significative tra i due sessi

Effettuo dei test t per campioni indipendenti per verificare differenze significative tra i due sessi, accompagnando i test numerici con dei boxplot per avere un riscontro grafico.

Le differenze in peso, lunghezza, diametro del cranio risultano tutte significative tra i due sessi (sempre considerando un livello di significatività di 0.05).

```
t.test(Peso ~ Sesso, paired = F)          # p-value < 2.2e-16
t.test(Lunghezza ~ Sesso, paired = F)     # p-value < 2.2e-16
t.test(Cranio ~ Sesso, paired = F)        # p-value = 1.7e-13
```

Invece, il tipo di parto (naturale o cesareo) risulta indipendente dal sesso del neonato.
In questo caso, siccome la variabile `Tipo.parto` è qualitativa, calcolo la tabella delle frequenze tra le variabili interessate ed effettuo un test chi-quadrato per saggiare l'ipotesi di indipendenza.

```
chisq.test(table(Tipo.parto, Sesso))    # p-value = 0.84
```

5 Più cesarei in certi ospedali?

Saggio un'altra ipotesi: si vocifera che in alcuni ospedali si facciano più parti cesarei, procedo a verificarla.
Siccome devo saggiare una proporzione, uso `prop.test()` sulla tabella delle frequenze tra le variabili interessate (faccio però presente che il test chi-quadrato dà lo stesso identico risultato).

```
prop.test(table(Ospedale, Tipo.parto))    # p-value = 0.57
```

Il p-valore è molto alto, quindi non rifiuto l'ipotesi nulla. Concludo che non ci sono differenze significative nel tipo di parto a seconda dell'ospedale.

6 Analisi multidimensionale

6.1 Normalità della variabile risposta

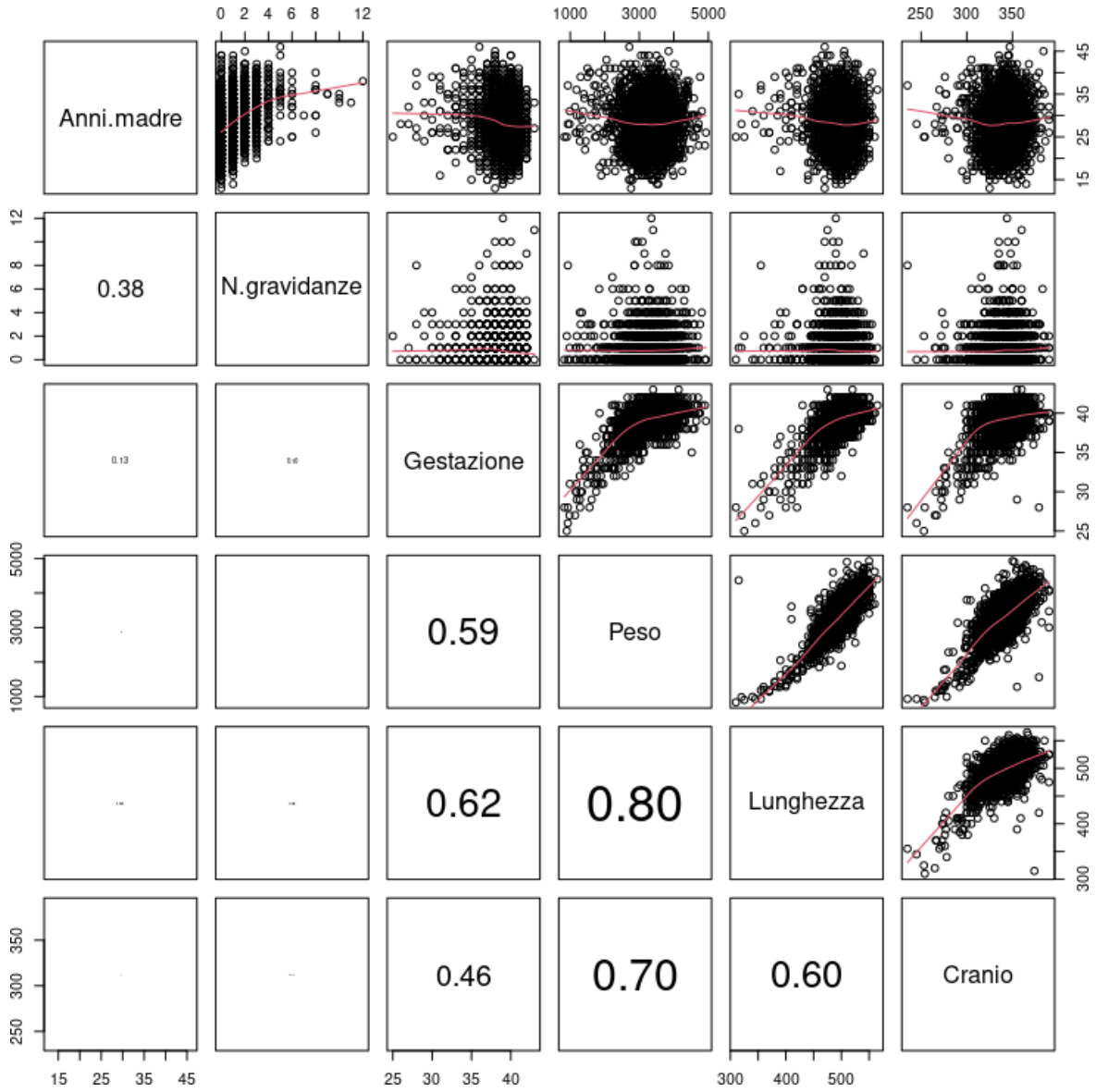
Prima di tutto, verifico che la variabile risposta (Peso) sia approssimativamente normale, andando a vedere gli indici di forma ed effettuando un test di Shapiro-Wilk. Lo verifico in anticipo perché eventuali allontanamenti dalla normalità della variabile risposta, spesso ricadono anche sui residui.

```
moments::skewness(Peso)         # -0.65
moments::kurtosis(Peso) - 3     # 2.03
shapiro.test(Peso)              # p-value < 2.2e-16
```

Noto che la variabile risposta (Peso) non segue una distribuzione normale...
Infatti, il test di Shapiro-Wilk rifiuta nettamente l'ipotesi nulla di normalità, e possiamo notare che la distribuzione è particolarmente leptocurtica (appuntita), con un valore di curtosi di 2.03. So che in questo caso sarebbe opportuno usare un GLM, ma provo comunque con un modello di regressione lineare.

6.2 Analisi

Indago le relazioni a due a due tra le variabili quantitative, sia numericamente (matrice di correlazione) che graficamente. A prima vista appaiono significative lunghezza, cranio, gestazione.



Poi indago le relazioni tra le variabili qualitative e la variabile risposta. Risulta significativa solo Sesso (invece la variabile Fumatrici - che indica se una madre sia fumatrice o meno - **non** risulta significativa, a differenza di quello che ci si poteva aspettare, con un p-valore di 0.30).

```
t.test(Peso ~ Sesso)          # p-value < 2.2e-16
t.test(Peso ~ Fumatrici)      # p-value = 0.30
t.test(Peso ~ Tipo.parto)     # p-value = 0.89
pairwise.t.test(Peso, Ospedale, paired = F, pool.sd = T, p.adjust.method =
"bonferroni")    # p-values = 1.00/0.33/0.33
```

7 Regressione con tutte le variabili

Creo il modello di regressione lineare multipla con tutte le variabili:

```
mod1 <- lm(Peso ~ .)
```

Il modello ha $\mathbf{R^2_{adj} = 0.72}$, che è un valore ragionevole. Riporto i coefficienti di seguito:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6738.4762   141.3087 -47.686  < 2e-16 ***
Anni.madre        0.8921     1.1323   0.788   0.4308
N.gravidanze     11.2665     4.6608   2.417   0.0157 *
Fumatrici       -30.1631    27.5386  -1.095   0.2735
Gestazione       32.5696     3.8187   8.529  < 2e-16 ***
Lunghezza        10.2945     0.3007  34.236  < 2e-16 ***
Cranio           10.4707     0.4260  24.578  < 2e-16 ***
Tipo.partoNat    29.5254    12.0844   2.443   0.0146 *
Ospedaleosp2    -11.2095    13.4379  -0.834   0.4043
Ospedaleosp3     28.0958    13.4957   2.082   0.0375 *
SessoM           77.5409    11.1776   6.937 5.08e-12 ***
```

Questo primo modello include 10 variabili (9 in realtà, ma per il fattore Ospedale ne vengono create due perché ha 3 modalità). Alcune hanno p-valori molto alti, quindi molto probabilmente si possono togliere dal modello per snellirlo (rasoio di Occam) senza perdere una quantità significativa di varianza spiegata.

8 Ricerca del modello "migliore"

8.1 Procedura stepwise automatica

Prima di fare test manuali, guardo cosa ottengo dalla funzione che implementa la procedura stepwise automatica - usando il criterio BIC (preferibile all'AIC in quanto non sovrastima modelli sovraparametrizzati).

```
stepwise_mod <- MASS::stepAIC(mod1,
                              direction = "both",
                              k = log(n))    # k = log(n) => uso BIC
```

Noto che quel modello include le variabili `N.gravidanze, Gestazione, Lunghezza, Cranio, Sesso`. Ora provo a procedere per passi manualmente.

8.2 Ricerca manuale del modello

Nel modello con tutte le variabili noto che `Anni.madre` ha un p-valore di 0.43. Rimuovendola R^2_{adj} resta invariato, il BIC scende e il test ANOVA non indica una differenza significativa di varianza spiegata (p-value di 0.43). Quindi la rimuovo senza dubbi.

La variabile `Ospedale` ha due p-valori (perché è qualitativa con 3 modalità) di 0.40 e 0.03. Rimuovendola R^2_{adj} scende di nemmeno un punto percentuale, il BIC scende e il test ANOVA indica una differenza significativa (p-value di 0.009). Ma per una diminuzione così bassa di R^2_{adj} a fronte di una rimozione di due variabili, scelgo di non tenerla.

Ora la variabile `Fumatrici` ha un p-valore di 0.25. Rimuovendola R^2_{adj} scende solo di 0.1%, il BIC scende e l'ANOVA riporta un p-valore di 0.25. Quindi la rimuovo senza dubbi.

A questo punto, tutte le variabili hanno p-valore < 0.05. Provo a togliere la meno significativa, `Tipo.parto`, che ha p-valore di 0.01. Rimuovendola, R^2_{adj} scende di mezzo punto percentuale, il BIC scende un pochino e il test ANOVA indica una differenza significativa (p-value di 0.01). Ma per una diminuzione così bassa di R^2_{adj} a fronte della rimozione di una variabile, scelgo di non tenerla.

Ora la variabile del modello con p-valore maggiore è `N.gravidanze` con p-valore di 0.004, provo a rimuoverla. Risulta che R^2_{adj} scende di nemmeno un punto percentuale, ma il BIC per la prima volta sale. Il test ANOVA indica una differenza significativa dal modello precedente (p-value di 0.004). A fronte di ciò, scelgo di tenere la variabile nel modello e fermarmi qua, perché tutte le altre variabili hanno p-valori dell'ordine di 10^{-12} o minori.

8.3 Modello risultante

Il modello risultante è il seguente (tra l'altro coincide con il modello trovato precedentemente, con la procedura stepwise automatica per minimizzare il BIC):

```
lm(Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso)

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 ***
---
Adjusted R-squared:  0.7265
```

Verifico che non ci siano problemi di multicollinearità (tutti i VIF sono inferiori a 5 quindi nessun problema).

```
car::vif(mod5)    # Tutti i valori sono < 5, OK
```

9 Effetti non lineari e interazioni

Dai grafici creati in precedenza qua noto un possibile effetto quadratico tra Gestazione e la Y. Provando ad inserirlo nel modello, ho che R^2_{adj} sale solo dello 0.04%, il p-valore di Gestazione^2 è 0.02 (ma quello di Gestazione è passato da 0.3% a 11%), il BIC sale e l'ANOVA indica una differenza significativa tra i due modelli.
Anche se graficamente mi pare che l'effetto quadratico possa essere rilevante, i test numerici non indicano un miglioramento significativo, quindi scelgo di non aggiungerla al modello.

Scelgo di non considerare interazioni tra le variabili, perché ragionandoci mentalmente non immagino nessuna relazione che possa essere significativa tra le variabili nel modello finale (escludendo le variabili antropometriche di controllo).

10 Analisi dei residui

Verifico, sia graficamente che numericamente, se il modello rispetta tutte le assunzioni sui residui (normalità con media 0, omoschedasticità, indipendenza).

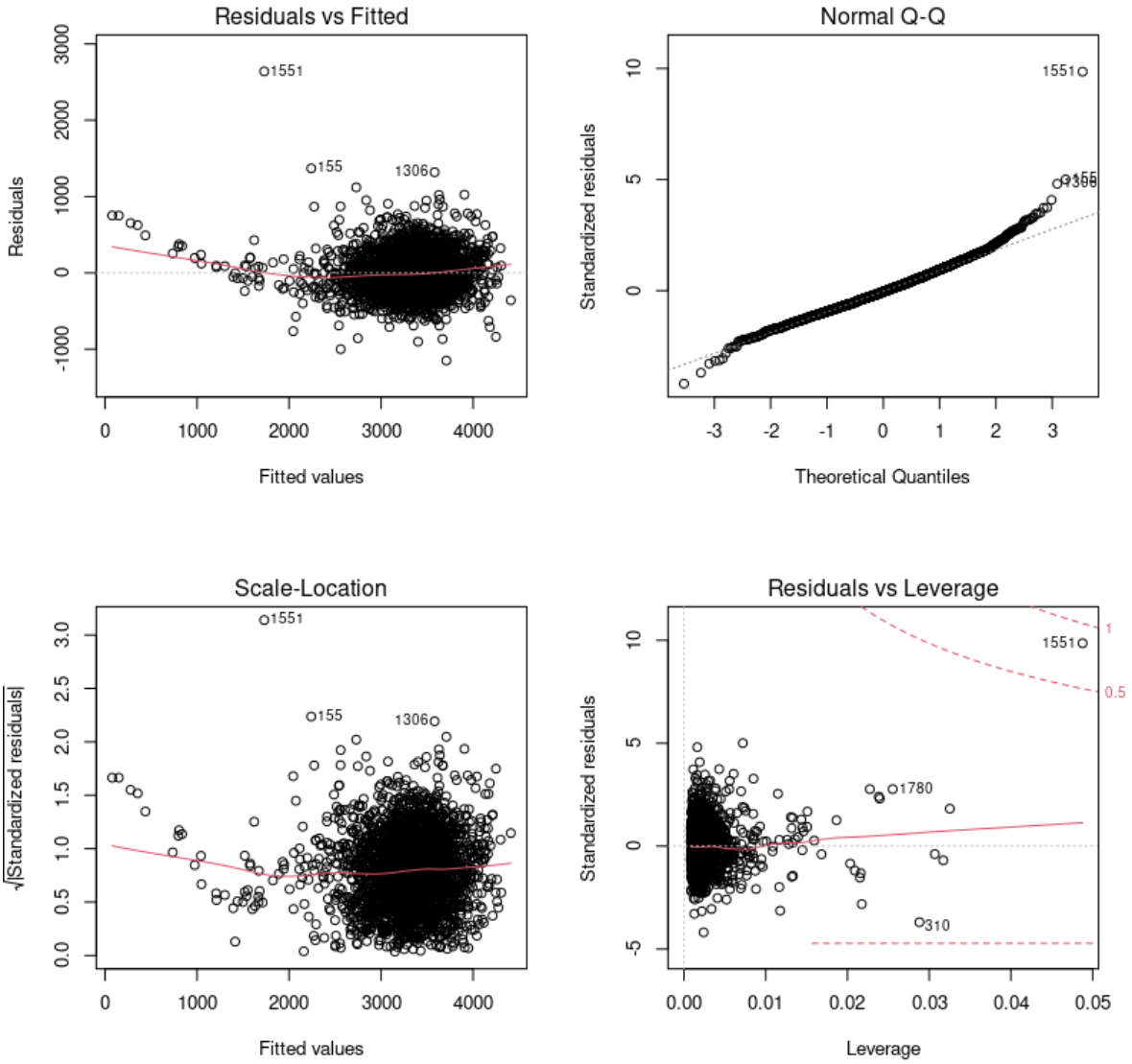


Figura 1: Analisi grafica dei residui del modello finale

Nel primo grafico (che mette in relazione le stime ottenute dal modello e i rispettivi residui), per la maggior parte si osserva una nuvola casuale di punti intorno ad una media di 0, che significherebbe tutto nella norma. L'unico problema visibile è che la coda di sinistra devia verso l'alto (cioè, il modello sottostima le previsioni del peso di neonati leggeri).

Nel secondo grafico (che confronta i quantili di Peso con i quantili della normale) in buona parte i punti del grafico si allineano lungo la retta $y = x$, indicando che i residui si allineano lungo una normale. Anche stavolta si nota che le code presentano un pattern diverso dalla maggior parte dei punti.

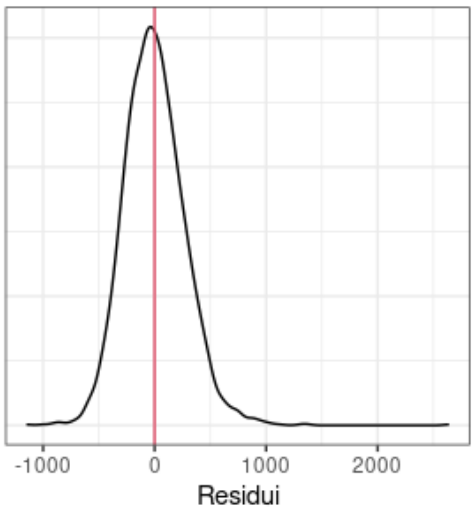
Nel terzo grafico notiamo, similmente al primo, principalmente una nuvola casuale di punti grosso modo orizzontale intorno ad un valore di y (ciò starebbe ad indicare una varianza costante, rispettando l'ipotesi di omoschedasticità). Anche qua però la coda di sinistra presenta il problema di deviare un po'verso l'alto.

Nel quarto grafico (valori anomali) solo un valore supera la soglia di avvertimento (distanza di Cook > 0.5) e nessuno supera la soglia di allarme di 1, quindi graficamente non vengono mostrati problemi con i valori anomali. Ora procedo con i test numerici.

10.1 Analisi numerica dei residui

```
shapiro.test(residuals(mod5))    # p-value < 2.2e-16
```

Il test di Shapiro-Wilk rifiuta nettamente la normalità dei residui, ma nel Q-Q plot avevamo previsto una ragionevole possibilità di normalità. Approfondiamo osservando direttamente la distribuzione dei residui:



In questo grafico si vede che la distribuzione dei residui assomiglia ad una normale; è prevalentemente simmetrica (con la coda di destra è particolarmente lunga), però sembra decisamente leptocurtica. Infatti la curtosi, calcolata con `moments::kurtosis(residuals(mod5)) - 3`, risulta 4.16, piuttosto alta.

Ora procedo con i test di Durbin-Watson e di Breusch-Pagan, che hanno come ipotesi nulle rispettivamente la non correlazione tra i residui e l'omoschedasticità.

```
lmtest::dwtest(mod5)    # p-value = 0.11 => residui non autocorrelati (bene)
lmtest::bptest(mod5)    # p-value < 2.2e-16 => !!! ho eteroschedasticita' (ahia)
```

10.1.1 Valori anomali

L'unico valore precedentemente individuato con distanza di Cook > 0.5 è l'osservazione 1551. Esaminandola, noto che Lunghezza=315 e Peso=4370. Confrontando il peso con quello di neonati di lunghezza simile, ho numeri estremamente diversi (intorno a 1000) quindi probabilmente sarà stato un errore di battitura[1]. Siccome nessun altro valore supera la soglia di avvertimento, e nessun valore arriva alla soglia di allarme, non ci sono problemi con valori anomali.

Ad ogni modo, per completezza esamino separatamente valori estremi nella variabile risposta, e nello spazio dei regressori (valori outlier e leverage):
La funzione `outlierTest` del pacchetto `car` restituisce 3 outlier (osservazioni 1551, 155 e 1306). Invece, il calcolo manuale dei valori di leva restituisce 152 valori (il 6% sul totale di 2500), che corrispondono a tutte le osservazioni che si trovano lontane dalle altre nello spazio dei regressori.

```
# Leverage
lev <- hatvalues(mod4)
p <- sum(lev)
soglia <- 2 * p/n
length(lev[lev > soglia])    # 152
```

11 Previsioni con il modello

Nel complesso, R^2_{adj} del modello è 0.72, cioè il 72% della variabilità del Peso è spiegato dal modello. Non ci sono valori anomali allarmanti. I residui presentano media 0 e non sono autocorrelati. A fronte di ciò, pure se c'è eteroschedasticità e il test di Shapiro-Wilk rifiuta la normalità, il modello sembra abbastanza buono per spiegare la variabilità della variabile risposta.

Faccio una previsione per il peso di una neonata: la madre è alla terza gravidanza (quindi N.gravidanze sarà 2, ossia le precedenti) e partorirà alla 39esima settimana. Non sono disponibili misure dall'ecografia.

Per fare una previsione in R devo passare un data frame con tutte le variabili, anche quelle che non ho. Per quelle quantitative (Lunghezza, Cranio) posso semplicemente usare la media di esse, mentre per Tipo.parto scelgo di considerare entrambe le modalità e farne la media pesata. Risulta un peso di 3261g, perfettamente ragionevole.

```
newdata_nat <- data.frame(N.gravidanze=2, Gestazione=39, Lunghezza=mean(Lunghezza),
               Cranio=mean(Cranio), Tipo.parto=factor("Nat"), Sesso=factor("F"))
newdata_ces <- data.frame(N.gravidanze=2, Gestazione=39, Lunghezza=mean(Lunghezza),
               Cranio=mean(Cranio), Tipo.parto=factor("Ces"), Sesso=factor("F"))
pred1 <- predict(mod5, newdata = newdata_nat)
pred2 <- predict(mod5, newdata = newdata_ces)
prediction = 1 / n * (pred1 * dim(data[Tipo.parto == "Nat",])[1]
                      + pred2 * dim(data[Tipo.parto == "Ces",])[1])
```

12 Rappresentazioni grafiche del modello

Riporto nuovamente la formula del modello finale:

```
Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
```

[1] Se dovessi indovinare, considerando che anche Cranio e Gestazione hanno valori non in linea con quelli di neonati di Lunghezza simile, la Lunghezza sarebbe potuta essere 515 invece che 315.

Per le variabili `N.gestazione`, `Lunghezza`, `Cranio` mostro grafici a dispersione (con relative linee di fit del modello, distinte per maschi e femmine) . Per `N.gravidanze` scelgo un boxplot perché permette una visualizzazione migliore. In tutti i grafici la variabile risposta `Peso` è sull'asse y.