# Progetto – Statistica Inferenziale

Il seguente documento conterrà il progetto di Statistica Inferenziale del master professionale in Data Science di Profession AI.

## Task 1 – Dataset e import

Il dataset riguarda 2500 neonati di 3 ospedali differenti.

Il dataset è composto da dieci 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

L’import dei dati su R è stato effettuato tramite la seguente linea di codice:

**data <- read.csv("neonati.csv")**

che ha permesso di leggere l’intero dataset e inserirlo all’interno dell’ambiente di R.

## Task 2 – Descrizione del dataset e obiettivo di studio

Il dataset viene esplorato tramite la funzione base di R “head” che ci permette di visualizzare le righe e colonne del dataset.Di seguito un’analisi di ciascuna variabile del dataset:

* Età della madre: variabile quantitativa discreta
* Numero di gravidanze sostenute: variabile quantitativa discreta
* Madre fumatrice: variabile qualitativa codificata (dummy)
* N° di settimane di gestazione: variabile quantitativa discreta
* Peso in grammi del neonato: variabile quantitativa continua
* Lunghezza in mm del neonato: variabile quantitativa continua
* Diametro in mm del cranio del neonato: variabile quantitativa continua
* Tipo di parto: variabile qualitativa su scala nominale
* Ospedale: variabile qualitativa codificata (dummy)
* Sesso del neonato: variabile qualitativa su scala nominale

In definitiva, sono presenti 3 variabili quantitative discrete, 3 variabili quantitative continue, 2 variabili qualitative su scala nominale e 2 variabili qualitative codificate (dummy).

Lo studio si pone come obiettivo quello di capire se c’è una relazione tra lo stile di vita della madre e la nascita del neonato. In base a questa relazione, andare a prevedere il peso del neonato alla nascita.

## Task 3 – Analisi descrittiva del dataset

Su tutte le variabili del dataset si è proceduto con un’analisi descrittiva dei dati al fine di comprenderne i comportamenti prima di passare allo sviluppo del modello statistico.

### Variabile – Anni.madre

La media di età delle mamme è di 38 anni.  
La deviazione standard è pari a 5, quindi c’è una variabilità in media di 5 anni tra le madri censite nel dataset.  
La funzione di densità di probabilità mostra un andamento quasi come una normale.  
Gli indici di forma indicano una distribuzione asimmetrica positiva e leptocurtica.

**Istogramma e funzione di densità della variabile Anni.madre**

Immagine che contiene diagramma, Diagramma, linea, testo

Descrizione generata automaticamente

### Variabile – N.gravidanze

La media di gravidanze delle mamme è di 1.  
La deviazione standard è pari a 1, quindi c’è una variabilità in media di 1 gravidanza tra le madri censite nel dataset.  
La funzione di densità di probabilità mostra un andamento non proprio come una normale standard poiché la media si discosta dallo 0.  
Gli indici di forma indicano una distribuzione asimmetrica positiva e leptocurtica.

**Istogramma e funzione di densità della variabile N.gravidanze**

Immagine che contiene diagramma, Diagramma, linea, pendio

Descrizione generata automaticamente

### Variabile – Gestazione

La media di settimane di gestazione delle mamme è di 39.  
La deviazione standard è pari a 1.8, quindi c’è una variabilità in media di 1.8 mesi di gestazione tra le madri censite nel dataset.  
La funzione di densità di probabilità mostra al centro due cunette e questo potrebbe star ad indicare che in quel punto ci sia un valore fuori dal “normale”, quindi un probabile valore di outliers.  
Gli indici di forma indicano una distribuzione asimmetrica negativa e leptocurtica.

**Istogramma e funzione di densità della variabile Gestazione**

Immagine che contiene diagramma, Diagramma, linea, pendio

Descrizione generata automaticamente

### Variabile – Peso

La media di peso dei neonati è pari a 3284g.  
La deviazione standard è pari a 525, quindi c’è una variabilità in media di 525g di peso tra i neonati censiti nel dataset.  
La funzione di densità di probabilità non mostra particolarità che potrebbe influire particolarmente lo studio del modello.  
Gli indici di forma indicano una distribuzione asimmetrica negativa e leptocurtica.

**Istogramma e funzione di densità della variabile Peso**

Immagine che contiene diagramma, Diagramma, linea, testo

Descrizione generata automaticamente

### Variabile – Lunghezza

La media di lunghezza dei neonati è pari a 497.7mm.  
La deviazione standard è pari a 26, quindi c’è una variabilità in media di 26mm di lunghezza tra i neonati censiti nel dataset.  
La funzione di densità di probabilità mostra una cunetta molto stretta e i valori tendono ad allagarsi di molto rispetto alla media.  
Gli indici di forma indicano una distribuzione asimmetrica negativa e leptocurtica.

**Funzione di densità della variabile Lunghezza**

Immagine che contiene diagramma, Diagramma, linea, pendio

Descrizione generata automaticamente

### Variabile – Cranio

La media di diametro del cranio dei neonati è pari a 340mm.  
La deviazione standard è pari a 16, quindi c’è una variabilità in media di 16mm di diametro del cranio tra i neonati censiti nel dataset.  
La funzione di densità di probabilità mostra una doppia cunetta al centro e questo potrebbe indicare dei valori di outliers presenti nel dataset.  
Gli indici di forma indicano una distribuzione asimmetrica negativa e leptocurtica.

**Istogramma e funzione di densità della variabile Cranio**

Immagine che contiene diagramma, Diagramma, linea, testo

Descrizione generata automaticamente

### Variabile – Fumatrici

La variabile fumatrici essendo qualitativa si è proceduto con la visualizzazione delle frequenze assolute.  
Dal grafico si evince che la maggior parte delle madri censite nel dataset non sono fumatrici.

**Frequenze assolute della variabile fumatrici**

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

### Variabile – Ospedale

La variabile ospedale essendo qualitativa si è proceduto con la visualizzazione delle frequenze assolute.  
Dal grafico si evince che la variabile mostra una quasi equidistribuzione tra le sue modalità.  
*(> gini.index(Ospedale) = 0.9998683)*

**Frequenze assolute della variabile ospedale**

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

### Variabile – Sesso

La variabile sesso essendo qualitativa si è proceduto con la visualizzazione delle frequenze assolute.  
Dal grafico si evince che la variabile mostra una quasi equidistribuzione tra le sue modalità.   
*(> gini.index(Sesso) = 0.999977)*

**Frequenze assolute della variabile ospedale**

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

## Task 4 – Saggiare l’ipotesi della media di peso e lunghezza del campione rispetto alla popolazione

La media della popolazione per il peso è intorno ai 3300g, mentre per la lunghezza è intorno ai 50cm secondo l’ISTAT.

Recuperate queste informazioni si è proceduto tramite il test t-student il test per saggiare l’ipotesi che le medie di peso e altezza del campione dei neonati del dataset risultino significativamente uguali alle medie della popolazione.

Si sono impostati i test dando in input le variabili Peso e Lunghezza, si è scelto il livello di significatività alfa pari a 0.95 e si è scelto l’opzione di verificare per entrambe le code per identificare un intervallo di confidenza.

Di seguito i risultati:

* Peso
  + Il p-value risulta essere maggiore del livello di significatività scelto pari al 5% (0.05) e per questo motivo non si rifiuta l’ipotesi nulla. La media del campione risulta essere significativamente uguale a quella della popolazione ed è possibile affermarlo con una percentuale di confidenza pari al 95.

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

* Lunghezza
  + Il p-value risulta essere minore del livello di significatività scelto pari al 5% (0.05) e per questo motivo si rifiuta l’ipotesi nulla a favore di quella alternativa. La media del campione risulta essere significativamente diversa da quella della popolazione.

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

## Task 5 – Differenze statisticamente significative tra maschio e femmina

È possibile individuare le differenze statisticamente significative tra i due sessi, maschio e femmina, sia per le variabili lunghezza e peso, ma anche per le dimensioni del cranio. Per effettuare questo studio si è pensato di procedere con la suddivisione del campione in due gruppi: Maschi e Femmine, al fine di condurre un test t per campioni indipendenti.

Di seguito i risultati:

* Lunghezza
  + Il p-value risulta essere molto vicino allo zero e inferiore rispetto al livello di significatività scelto pari a 5% (0.05). Per questo motivo si rifiuta l’ipotesi nulla a favore di quella alternativa potendo affermare con una confidenza del 95% che mediamente la lunghezza tra neonati maschi e femmine è significativamente diversa.  
    Immagine che contiene testo, schermata, Carattere

    Descrizione generata automaticamente
* Peso
  + Il p-value risulta essere molto vicino allo zero e inferiore rispetto al livello di significatività scelto pari a 5% (0.05). Per questo motivo si rifiuta l’ipotesi nulla a favore di quella alternativa potendo affermare con una confidenza del 95% che mediamente il peso tra neonati maschi e femmine è significativamente diverso.  
    Immagine che contiene testo, schermata, Carattere

    Descrizione generata automaticamente
* Cranio
  + Il p-value risulta essere molto vicino allo zero e inferiore rispetto al livello di significatività scelto pari a 5% (0.05). Per questo motivo si rifiuta l’ipotesi nulla a favore di quella alternativa potendo affermare con una confidenza del 95% che mediamente il diametro del cranio tra neonati maschi e femmine è significativamente diverso.  
    Immagine che contiene testo, schermata, Carattere

    Descrizione generata automaticamente

## Task 6 – Verificare in quale Ospedale si eseguono più parti cesareo

Per individuare l’ospedale in cui sono stati effettuati maggiori parti di tipo cesareo, si è optato di utilizzare la libreria ggplot2 e creare un grafico a barre affiancate, al fine di individuare graficamente in quale ospedale sono stati fatti più parti cesareo.

Si è inserito all’interno delle estetiche la variabile Ospedale come ascissa e la variabile Tipo.parto come fill, ed infine l’opzione position pari a dodge per affiancare le barre.

Di seguito il risultato:

Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente

È possibile affermare che nell’ospedale 2 vengono effettuati più parti cesarei.

## Task 7 – Analisi multidimensionale: verifica relazioni a due a due tra le variabili

L'obiettivo dello studio è quello di capire se è possibile effettuare delle previsioni sul peso dei neonati alla nascita in base allo stile di vita delle madri. Prima di procedere allo sviluppo del modello atto a fare le previsioni, si è proceduto con lo studio delle correlazioni tra le variabili esplicative con la variabile risposta. In particolare, si è ritenuto opportuno non studiare la correlazione tra la variabile Ospedale con la variabile risposta Peso, poiché non ha nessun senso. Per questo si è proceduto con lo studio di correlazione tra tutte le restanti variabili esplicative e la variabile risposta.

Di seguito i risultati:

* **Anni madre** con il **Peso** ha un coefficiente di correlazione non parametrico di Spearman che presenta un p-value di 0.7648, quindi superiore al livello di significatività del 5%. Questo risultato viene ad indicare che non è possibile rigettare l’ipotesi nulla di correlazione pari a 0 ma che la correlazione non risulta significativa o quantomeno molto debole. Inoltre, andando a plottare le due variabili in uno scatterplot non si evince nessun pattern tra le due variabili. In particolare, anni madre essendo una variabile temporale e sotto certi aspetti importante, potrebbe essere inserita come variabile di controllo.

Immagine che contiene testo, diagramma, schermata, linea

Descrizione generata automaticamente

* **N.gravidanze** con il **Peso** ha un coefficiente di correlazione non parametrico di Spearman che presenta un p-value di 0.4037, quindi superiore al livello di significatività del 5%. Questo risultato viene ad indicare che non è possibile rigettare l’ipotesi nulla di correlazione pari a 0 ma che la correlazione non risulta significativa o quantomeno molto debole. Inoltre, andando a plottare le due variabili in uno scatterplot non si evince nessun pattern tra le due variabili.

Immagine che contiene testo, schermata, diagramma, Parallelo

Descrizione generata automaticamente

* **Gestazione** con il **Peso** ha un coefficiente di correlazione non parametrico di Spearman che presenta un p-value di 2 alla -16, quindi molto inferiore al livello di significatività del 5%. Questo risultato viene ad indicare che è possibile rigettare l’ipotesi nulla di correlazione pari a 0 e che la correlazione risulta significativa e forte tra le due variabili. Inoltre, andando a plottare le due variabili in uno scatterplot non si evince nessun pattern tra le due variabili.

Immagine che contiene testo, schermata, diagramma, Parallelo

Descrizione generata automaticamente

* **Lunghezza** con il **Peso** ha un coefficiente di correlazione lineare di Pearson pari a 0.7960. Questo risultato indica una forte correlazione lineare positiva tra le due variabili. Inoltre, andando a plottare le due variabili in uno scatterplot si evince possibile pattern tra le due variabili.

Immagine che contiene testo, diagramma, linea

Descrizione generata automaticamente

* **Cranio** con il **Peso** ha un coefficiente di correlazione lineare di Pearson pari a 0.7048. Questo risultato indica una moderata correlazione lineare positiva tra le due variabili. Inoltre, andando a plottare le due variabili in uno scatterplot si evince possibile pattern tra le due variabili.

Immagine che contiene testo, diagramma, schermata

Descrizione generata automaticamente

* **Sesso** con il **Peso** risulta, da un test di indipendenza di chi-quadro Pearson, un p-value pari a 2 alla -16. Questo risultato indica che l’ipotesi nulla di indipendenza viene rigettata a favore di quella alternativa di dipendenza tra le due variabili.
* **Fumatrici** con il **Peso** risulta, da un test di indipendenza di chi-quadro Pearson, un p-value pari a 0.0995. Questo risultato indica che l’ipotesi nulla di indipendenza non viene rigettata a favore di quella alternativa di dipendenza tra le due variabili.

## Task 8 – Creazione modello di regressione lineare multipla

Prima di procedere con la creazione del modello di regressione lineare multipla, si è ritenuto opportuno studiare se la variabile risposta si comporti normalmente o almeno simile ad una normale. Per questo si sono studiati i due indici di forma: Fisher e Curtosi.

L’indice di Fisher è pari a -0.6470 e questo indica una distribuzione asimmetrica negativa. L’indice di Curtosi è pari a 2.031 e questo indica una distribuzione leptocurtica. Infine, per verificare che la variabile risposta si comporti normalmente è stato eseguito il test di normalità di Shapiro-Wilk che ha presentato come risultato un p-value di 2 alla -16. Per questo motivo, non si rifiuta l’ipotesi nulla di normalità e la variabile risposta si comporta normalmente.

A questo punto, si è proceduto con la creazione del primo modello, includendo tutte le variabili come richiesto dal task.

Di seguito i risultati:

* La variabile **Anni.madre**, come anche constatato nello studio della correlazione con la variabile Peso, presenta un coefficiente beta positivo pari a circa 0.9, questo significherebbe un effetto marginale in media di 0.9 del peso. Il p-value è di 0.430, il che indica che la variabile Anni.madre non ha un effetto significativo sulla variabile risposta.
* La variabile **N.gravidanze** presenta un coefficiente beta positivo pari a 11.2, questo significherebbe che ad ogni variazione unitaria della variabile N.gravidanze, si avrebbe una variazione di circa 11.2g di Peso. Il p-value è di 0.0157, quindi poco sotto la soglia di significatività del 5%, per questo motivo ha senso includere la variabile all’interno del modello.
* La variabile **Fumatrici**, come anche constatato nello studio della correlazione con la variabile Peso, presenta un coefficiente beta negativo pari a -30, questo significherebbe un effetto marginale in media di -30 del peso. Il p-value è di 0.27, il che indica che la variabile Fumatrici non ha un effetto significativo sulla variabile risposta.
* La variabile **Gestazione** presenta un coefficiente beta positivo pari a 32.56, questo significherebbe che ad ogni variazione unitaria della variabile Gestazione, si avrebbe una variazione di circa 32.56g di Peso. Il p-value è di 2 alla -16, quindi molto inferiore al livello di significatività del 5%, per questo motivo ha senso includere la variabile all’interno del modello.
* La variabile **Lunghezza** presenta un coefficiente beta positivo pari a 10.29, questo significherebbe che ad ogni variazione unitaria della variabile Lunghezza, si avrebbe una variazione di circa 10.29g di Peso. Il p-value è di 2 alla -16, quindi molto inferiore al livello di significatività del 5%, per questo motivo ha senso includere la variabile all’interno del modello.
* La variabile **Cranio** presenta un coefficiente beta positivo pari a 10.47, questo significherebbe che ad ogni variazione unitaria della variabile Cranio, si avrebbe una variazione di circa 10.47g di Peso. Il p-value è di 2 alla -16, quindi molto inferiore al livello di significatività del 5%, per questo motivo ha senso includere la variabile all’interno del modello.
* La variabile **Tipo.parto** presenta un coefficiente beta positivo pari a 29.52, questo significherebbe che rispetto alla baseline scelta, cioè parto naturale, il peso dei neonati nati con il parto cesareo, risultano avere in media 29.52g in più. Il p- value è di 0.0146, quindi poco sotto la soglia di significatività del 5%, per questo motivo ha senso includere la variabile all’interno del modello.
* La variabile **Ospedale**, nonostante presenta dei risultati che indicherebbero la possibilità di spiegare ulteriormente la variabile risposta, non risulta avere un senso includerla nel modello poiché, in termini reale, la nascita in ospedale piuttosto che un altro, non può influenzare il peso del neonato.
* La variabile **Sesso**, presenta un coefficiente beta positivo pari a 77.54, questo significherebbe che rispetto alla baseline scelta, cioè parto maschio, il peso dei neonati maschi risultano avere in media 77.54g in più rispetto alle femmine. Il p- value è di 5.008 alla -12, quindi molto inferiore al livello di significatività del 5%, per questo motivo ha senso includere la variabile all’interno del modello.

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

Per quanto riguarda la bontà del modello e quindi di quanto quest’ultimo si adatti ai dati, il coefficiente R quadro è pari a 0.72, quindi un modello discreto per fare previsioni sul peso dei neonati.

## Task 9 – Selezione del modello migliore

La scelta del modello migliore è stata eseguita utilizzando la procedura Step-Wise. In primis, è stata utilizzata la libreria car per eseguire la procedura Step-Wise in maniera automatica con la metodologia mista. Successivamente, si è proceduto sempre con la Step-Wise, ma in maniera backward e manualmente, andando ad eliminare quelle variabile non ritenute significative e utili al modello.

La procedura di Step-Wise automatica mista ha riportato come output un modello che comprende tutte le variabili al di fuori della variabile Fumatrici, includendo così anche Ospedale e Anni.madre. Analizzando il risultato, la variabile Ospedale, pur avendo un p-value poco sotto la soglia del 5% per la modalità osp3 come baseline rispetto alle altre, non ha senso tenerla poiché nella realtà non può influire sul peso dei neonati. Per quanto riguarda Anni.madre, anch’essa ha un p-value alto e quindi non ha un effetto significativo sul peso dei neonati.

Per queste ragioni, si è proceduto con la rimozione, one-by-one, delle variabili e valutati i corrispondenti modelli. I modelli presi in esame hanno tutti lo stesso R quadro, in particolare, qualcuno varia del 0.01 rispetto agli altri, quindi una variazione irrisoria e non significativa. La scelta, quindi, del modello migliore è stata effettuata seguendo il principio del rasoio di Occam, il quale porta a selezionare ciò che è più “semplice”.

Il modello più “semplice” e che abbia un buon compromesso tra variabili esplicative e R2 discreto, è stato il modello che include le seguenti variabili:

* N.gravidanze
* Gestazione
* Sesso
* Cranio
* Lunghezza

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

## Task 10 – Studio degli effetti di interazione e non linearità

Analizzando gli scatterplot delle variabili Cranio e Lunghezza in relazione al Peso, si può sospettare ed ipotizzare un eventuale correlazione non lineare poiché le nuvola di punti presentano un inclinazione positiva e in più sembrano avere nella parte iniziale una sorta di “curva”. Per quanto riguarda gli effetti di interazione si sospetta un effetto significativo tra le dimensioni del Cranio e la Lunghezza del neonato.

Per verificare entrambe le ipotesi si è proceduto come descritto di seguito:

* L’effetto di interazione tra Lunghezza e Cranio è stata studiata sulla base del fatto che neonati che hanno un lunghezza e un cranio di dimensioni moderate, risulta avere, di conseguenza, un peso maggiore. Per questo motivo, è stata inclusa all’interno del modello la coppia di variabili moltiplicate tra loro. Il risultato, considerando il modello senza interazione, un aumento irrisorio dell’R quadro aggiustato di 0.002.
* La non linearità per la variabile Lunghezza è stata studiata provando a creare un modello lineare tra solo Peso e Lunghezza, verificare che i residui non tendono a disperdersi e creare un modello che tenga conto dell’effetto non lineare della variabile lunghezza. Infine, è stato eseguito un anova test per verificare quale dei due modelli tende a spiegare meglio la variabile Peso. Il risultato è stato che il modello non lineare presenta un p-value molto piccolo per il test F, questo indica una differenza statisticamente significativa tra i due modelli in termini di variabile risposta spiegata, per questo motivo all’interno del modello è stato inclusa la variabile Lunghezza al quadrato. L’R quadro aggiustato, con l’inserimento della Lunghezza al quadrato, ha subito un incremento discreto del 0.01.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

* La non linearità per la variabile Cranio è stata studiata allo stesso modo della variabile Lunghezza, ma riportando un risultato negativo in termini di variabile risposta spiegata e per questo motivo, la variabile Cranio è stata inserita all’interno del modello in forma lineare.

Immagine che contiene testo, schermata, diagramma

Descrizione generata automaticamente

Immagine che contiene testo, Carattere, schermata

Descrizione generata automaticamente

In conclusione, considerando il principio del rasoio di Occam e l’R quadro aggiustato, includendo gli effetti di interazione e non linearità, varia di pochissimo la bontà del modello, per questo motivo si è ritenuto più appropriato inserire la variabile Lunghezza con il suo effetto non lineare e tenere il modello con le stesse variabili individuate nel Task 9.

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

## Task 11 – Diagnostica dei residui

La diagnostica sui residui è stata effettuata sul modello finale scelto e descritto all’interno del task precedente, il quale include le variabili: N. Gravidanze, Gestazione, Lunghezza al quadrato, Cranio e Sesso. In primis, si è proceduto con un’analisi preliminare grafica e poi successivamente all’utilizzo di test statistici specifici.

Di seguito i risultati della diagnostica:

* **Shapiro-Wilk** test per verificare la normalità dei residui
  + Dal grafico del “**Normal Q-Q plot**” i punti giacciono pressappoco su tutta la bisettrice del grafico; quindi, i residui sembrano comportarsi normalmente.

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamente

* + P-value pari a 2 alla -16, quindi praticamente pari a 0. Questo risultato indica che si rifiuta l’ipotesi nulla di normalità dei residui.

Immagine che contiene testo, Carattere, schermata

Descrizione generata automaticamente

* **Breusch-Pagan** test per verificare l’omoschedasticità dei residui (varianza costante):
  + Dal grafico del “**Residuals vs Fitted**” i punti sembrano sparsi intorno alla media di 0 e quindi, sembrerebbe non esserci eteroschedasticità.

Immagine che contiene testo, schermata, diagramma

Descrizione generata automaticamente

* + P-value pari a 6.6 alla -11, quindi praticamente pari a 0. Questo risultato indica che si rifiuta l’ipotesi nulla di omoschedasticità dei residui.

Immagine che contiene testo, Carattere, schermata

Descrizione generata automaticamente

* **Durbin-Watson** test per verificare la presenza di autocorrelazione tra i residui:
  + Il grafico “**Scale-Location**” mostra grosso modo una nuvola sparsa di punti e quindi nessun pattern particolare indicando la non-presenza di autocorrelazione.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

* + P-value pari a 0.11 quindi superiore al livello di significatività del 5%, questo indica che non si rifiuta l’ipotesi nulla di indipendenza dei residui. Non c’è autocorrelazione.

Inoltre, si è proceduto con l’analisi di eventuali osservazioni leverage e outliers del modello. Di seguito i risultati:

* **Leverage**
  + Il modello presenta molti valori leverage, quindi molti valori inusuali nello spazio dei regressori.

Immagine che contiene testo, schermata, menu

Descrizione generata automaticamente

Immagine che contiene testo, schermata

Descrizione generata automaticamente

* **Outliers**
  + Il modello presenta 4 valori outliers quindi 4 valori inusuali della variabile risposta. I 4 valori indicati presentano un p-value molto piccolo e questo indica che i valori sono significativamente anomali e influenti statisticamente per il modello. Il tutto indica una poca robustezza del modello.

Immagine che contiene testo, schermata, Rettangolo, linea

Descrizione generata automaticamente

Immagine che contiene testo, Carattere, schermata, numero

Descrizione generata automaticamente

## Task 12 – Descrizione della bontà del modello

Considerando i risultati della diagnostica sui residui, il modello non rispetta le seguenti assunzioni:

* Normalità dei residui
* Omoschedasticità dei residui

Il modello presenta molto valori di leverage che risultano significativamente influenti poiché superano di grand lunga la soglia prevista e inoltre, sono presenti anche quattro valori outliers che a loro volta influenzano significativamente.

In conclusione, nonostante l’R quadro aggiustato del modello è pari a circa 0.73, non filtra bene le informazioni riversandole all’interno dei residui. Questo rende il modello discreto per effettuare previsioni, sicuramente non ottimale.

## Task 13 – Previsione

La previsione del peso di una neonata alla 39esima settimana di gestazione e con una madre che sta alla terza gravidanza è pari a: **3627g**.

## Task 14 – Rappresentazione grafica del modello

La rappresentazione grafica del modello è stata effettuata tramite gli scatter3d e la libreria ggplot2 per mostrare le relazioni della variabile Peso con le altre variabili individuate per il modello.

Gli scatter3d per problemi di visibilità, sono visibili direttamente in R utilizzando il codice legato al progetto, mentre per le rappresentazioni grafiche delle rette di regressioni si è potuto procedere anche con l’inserimento degli screen.

**Rappresentazione del modello per Peso – Gestazione - Sesso**

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

**Rappresentazione del modello per Peso – Lunghezza - Sesso**

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

**Rappresentazione del modello per Peso – Gestazione - Sesso**

Immagine che contiene diagramma, schermata, testo, linea

Descrizione generata automaticamente