

Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea in

Statistica per le Tecnologie e le Scienze



Studio statistico su cani affetti da infiammazioni
al sistema nervoso centrale

Relatore: prof. Laura Ventura
Dipartimento di Scienze statistiche

Correlatore: Aurora Cocchetto
Clinica Veterinaria San Marco

Laureando: Francesco Tomaselli
Matricola n. 1220245

Anno Accademico 2021/2022

Indice

1	Introduzione	5
1.1	In ingresso	5
1.2	In fase di recidiva	6
2	Analisi esplorative	7
2.1	Analisi esplorativa univariata	7
2.1.1	In ingresso	7
2.1.2	In fase di recidiva	13
2.2	Analisi esplorativa bivariata	18
2.2.1	In ingresso	18
2.2.2	In fase di recidiva	24
3	Analisi multivariate	31
3.1	In ingresso	31
3.1.1	Analisi delle componenti principali	31
3.1.2	Analisi dei gruppi	33
3.2	In fase di recidiva	37
3.2.1	Analisi delle componenti principali	37
3.2.2	Analisi dei gruppi	39
4	Modello di regressione logistico	43
4.1	In ingresso	43
4.1.1	Guarigione contro Recidiva	43
4.1.2	Guarigione contro Decesso	44
4.1.3	Guarigione contro Recidiva e/o Decesso	45
4.2	In fase di recidiva	46
5	Metodi asintotici di ordine elevato	49
5.1	In ingresso	49
5.1.1	Guarigione contro Recidiva	49
5.1.2	Guarigione contro Decesso	50
5.1.3	Guarigione contro Decesso e/o Recidiva	51
5.2	In fase di Recidiva	52
6	Regressione robusta	55
6.1	In ingresso	55
6.1.1	Guarigione contro Recidiva	55
6.1.2	Guarigione contro Decesso	56
6.1.3	Guarigione contro Recidiva e/o Decesso	56
6.2	In fase di recidiva	57

7	Analisi di durata	61
7.1	Analisi esplorative	61
7.2	Modello di Cox	65
7.3	Bontà di adattamento	66
8	Conclusioni	69
8.1	In ingresso	69
8.2	In fase di recidiva	70

Capitolo 1

Introduzione

I dati di questo elaborato fanno riferimento a uno studio caso-controllo di tipo osservazionale (si veda Vineis *et al.*, 1988 [11]) condotto su 52 cani selezionati dalla [Clinica Veterinaria San Marco](#) di Veggiano. Tra i soggetti portati in clinica tra il 2017 e il 2021, sono stati selezionati i cani con un'età maggiore di sei mesi, sui quali sono state diagnosticate delle patologie al sistema nervoso centrale e presentano delle anomalie nei parametri sanguigni.

La patologia di interesse è lo stato dell'infiammazione al sistema nervoso centrale, che viene suddiviso in guarigione, recidiva o decesso per tale malattia. Le domande di ricerca poste dalla clinica sono suddivise in due categorie: in fase di ingresso dei cani in clinica e in fase di recidiva (esami di follow-up).

Più precisamente, gli **obiettivi** del report, richiesti dalla clinica veterinaria San Marco, sono i seguenti:

1. In ingresso:

- Differenze nei parametri sanguigni tra i cani che guariscono con i cani che muoiono per infiammazione al sistema nervoso centrale.
- Differenze nei parametri sanguigni tra i cani che guariscono da infiammazione al sistema nervoso centrale con i cani che sviluppano la recidiva.
- Differenze nei parametri sanguigni tra i cani che guariscono da infiammazione al sistema nervoso centrale con i cani che muoiono oppure che sviluppano la recidiva.
- Valutare se la terapia steroidea in atto influenza lo stato della malattia del soggetto.

2. In fase di recidiva:

- Differenze nei parametri sanguigni tra i cani che guariscono da infiammazione al sistema nervoso centrale con i cani che sviluppano la recidiva in fase di recidiva.
- Studiare quali fattori abbiano influenzato il tempo di attesa alla recidiva nei pazienti.

Vista la natura delle domande di ricerca, nel il report verranno separate le analisi e le conclusioni per i soggetti in ingresso da quelle per i soggetti in fase di recidiva.

1.1 In ingresso

Il *dataset* dei soggetti in ingresso presenta 52 unità statistiche e le informazioni raccolte su di essi sono le seguenti:

1. *Stato*: variabile qualitativa sconnessa che rappresenta lo stato di malattia del soggetto. Assume valore "Guarigione" se il cane è guarito dall'infiammazione al sistema nervoso centrale, "Recidiva" se ha sviluppato la recidiva e "Decesso" se è deceduto.
2. *Terapia*: variabile dicotomica che indica se il cane è attualmente sottoposto a una terapia steroidea.

3. *Leucociti*: variabile quantitativa che rappresenta la quantità di Leucociti presenti nel sangue del soggetto, espressa in migliaia per millilitro di sangue.
4. *Neutrofili*: variabile quantitativa che rappresenta la quantità di Neutrofili presenti nel sangue del soggetto, espressa in migliaia per millilitro di sangue.
5. *PON1*: variabile quantitativa che rappresenta la quantità di Paraoxonase 1 nel sangue del soggetto, espressa in IU (*International Units*) per litro di sangue.
6. *CRP*: variabile quantitativa che rappresenta la quantità di proteina C-Reattiva nel sangue del soggetto, espressa in milligrammi per decilitro di sangue.
7. *Aptoglobina*: variabile quantitativa che rappresenta la quantità di Aptoglobina nel sangue del soggetto, espressa in milligrammi per decilitro di sangue.
8. *Ferritina*: variabile quantitativa che rappresenta la quantità di Ferritina nel sangue del soggetto, espresso in nanogrammi per millilitro di sangue.
9. *Albumine*: variabile quantitativa che rappresenta la quantità di Albumine presente nel sangue del soggetto, espressa in grammi per decilitro di sangue.
10. *IgM*: variabile quantitativa che rappresenta la quantità di immunoglobuline IgM presente nel sangue soggetto, espressa in milligrammi per decilitro di sangue.
11. *IgA*: variabile quantitativa che rappresenta la quantità di immunoglobuline IgA presente nel sangue soggetto, espressa in milligrammi per decilitro di sangue.
12. *IgG*: variabile quantitativa che rappresenta la quantità di immunoglobuline IgG presente nel sangue soggetto, espressa in milligrammi per decilitro di sangue.
13. *Fibrinogeno*: variabile quantitativa che rappresenta la quantità di Fibrinogeno presente nel sangue del soggetto, espressa in milligrammi per decilitro di sangue.

La **variabile di interesse** risulta essere lo stato della malattia del paziente.

Il dataset originale presentava un valore mancante per il livello di *IgA* del cane con codice identificativo 50. Tale valore, è stato sostituito con la mediana della variabile, dato che la distribuzione del livello di *IgA* risulta essere asimmetrica (si veda la Sezione 2.1.1).

1.2 In fase di recidiva

Il *dataset* dei soggetti in fase di recidiva (follow-up) presenta 24 unità statistiche e le informazioni su di essi sono analoghe a quelle dei soggetti in ingresso, con le seguenti eccezioni:

1. La variabile *Stato* è una variabile dicotomica che rappresenta unicamente la guarigione oppure lo sviluppo della recidiva dell'infiammazione al sistema nervoso centrale dei pazienti in fase di recidiva.
2. Viene aggiunta una variabile che rappresenta il *Tempo*, che rappresenta il tempo di attesa alla recidiva per i soggetti in fase di recidiva, calcolata come il tempo in mesi intercorso tra gli esami in ingresso e in fase di recidiva. Questa variabile verrà utilizzata esclusivamente nel Capitolo 7.
3. La variabile che rappresenta la presenza di terapia steroidea in atto non è stata rilevata per i soggetti in fase di recidiva.
4. La variabile che rappresenta il livello di *Fibrinogeno* nel sangue dei pazienti in fase di recidiva è stata trascurata perché presenta 6 valori mancanti su 24 valori totali.

La **variabile di interesse** risulta essere lo stato della malattia del paziente.

Capitolo 2

Analisi esplorative

In questo capitolo vengono presentate le analisi esplorative univariate e bivariate coerenti con gli obiettivi del report. Per i metodi utilizzati per le analisi esplorative si veda Ventura e Racugno, 2017 [10].

2.1 Analisi esplorativa univariata

In questa sezione vengono riportate le analisi esplorative univariate per i soggetti affetti da infiammazione al sistema nervoso centrale.

2.1.1 In ingresso

Nel seguito vengono presentate le analisi esplorative univariate per i soggetti in ingresso nella clinica.

Stato della malattia e terapia steroidea

In tabella 2.1 sono riportate le distribuzioni di frequenza dello stato della malattia dei soggetti e della presenza della terapia steroidea in atto.

Stato	Frequenze Assolute	Frequenze Percentuali	Terapia	Frequenze Assolute	Frequenze Percentuali
Guarigione	24	46.15%	Sì	7	13.46%
Decesso	15	28.85%	No	45	86.54%
Recidiva	13	25.00%	Totale	52	100%
Totale	52	100%			

Tabella 2.1: Distribuzione di frequenza assoluta e percentuale dello stato di malattia e della presenza di terapia steroidea in atto nei pazienti.

Dalla tabella 2.1 emerge che la maggior parte dei soggetti presi in considerazione sono guariti dall'infiammazione al sistema nervoso centrale (46.15%) e non presentano una terapia steroidea in atto (86.54%). In figura 2.1 sono riportati i grafici a barre (con i relativi intervalli di confidenza di livello 0.95 per le proporzioni dei gruppi) per lo stato della malattia e la presenza di terapia steroidea in atto per i soggetti.

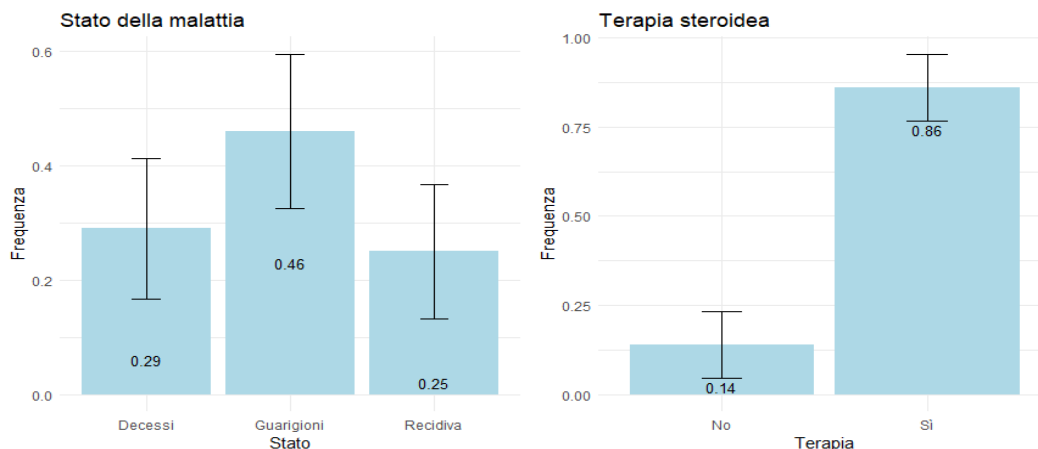


Figura 2.1: Grafico a barre per lo stato della malattia nei pazienti e la presenza della terapia steroidea in atto.

Il test per verificare l'ipotesi che la proporzione di pazienti deceduti sia uguale alla proporzione dei pazienti guariti riporta un valore osservato della statistica test pari a 5.58 e un p-value pari a 0.018. Il test per verificare l'ipotesi che la proporzione di pazienti con la recidiva sia uguale alla proporzione dei pazienti guariti riporta un valore osservato della statistica test pari a 8.52 e un p-value pari a 0.003. Il test per verificare l'ipotesi che la proporzione di pazienti deceduti sia uguale alla proporzione di pazienti con la recidiva riporta un valore osservato della statistica test pari a 0.23 e un p-value pari a 0.631. Infine, il test per verificare che la proporzione di pazienti con la terapia steroidea in atto sia pari al 25% riporta un valore osservato della statistica test pari a 3.91 e un p-value pari a 0.042. Quindi, a un livello di α del 5%, la proporzione dei pazienti deceduti e la proporzione dei pazienti con la recidiva possono considerarsi uguali tra loro e significativamente inferiori alla proporzione di pazienti guariti nel campione, e la proporzione di pazienti con la terapia steroidea in atto è inferiore al 25%.

Parametri sanguigni

Per i parametri infiammatori in ingresso, in tabella 2.2 vengono riportate le principali statistiche di sintesi e il p-value del test di Shapiro-Wilk per la verifica dell'ipotesi di normalità.

Variabile	Min	1st Qu	Median	Mean	3rd Qu	Max	Sd	Shapiro-Wilk p-value
Leucociti	4.62	8.04	9.17	10.11	11.40	22.43	3.34	<0.01
Neutrofili	2705	6020	6750	7619.9	8917.5	17260	2624.3	<0.01
PON1	2.61	3.92	4.49	4.45	5.11	6.12	0.89	0.415
CRP	0.01	0.01	0.20	0.88	1.00	12.76	2.02	<0.01
Aptoglobina	1.00	39.25	119.00	177.56	242.00	703.00	184.93	<0.01
Ferritina	92.00	149.25	185.50	219.98	256.25	676.00	118.89	<0.01
Albumine	2.50	3.27	3.50	3.50	3.80	4.50	0.43	0.524
IgG	268.00	348.75	402.50	432.56	478.75	829.00	121.80	<0.01
IgA	0.84	11.00	12.25	16.29	15.00	166.00	22.16	<0.01
IgM	53.00	100.75	113.00	114.54	131.50	163.00	23.60	0.582
Fibrinogeno	108.00	204.00	264.00	309.50	390.00	680.00	139.68	<0.01

Tabella 2.2: Statistiche di sintesi per i parametri infiammatori in ingresso dei soggetti e p-value del test di Shapiro-Wilk per la verifica della normalità.

Dalla tabella 2.2 si nota che solo per il livello di *PON1*, il livello di *Albumine* e il livello di *IgM* nel sangue dei pazienti in ingresso il test di Shapiro-Wilk per verificare l'ipotesi di normalità conduce all'accettazione a

tutti i livelli usuali di α .

Vengono riportati i principali grafici esplorativi del livello nel sangue dei pazienti in ingresso di *Leucociti* (figura 2.2), *Neutrofili* (figura 2.3), *CRP* (figura 2.4), *Aptoglobina* (figura 2.5), *Ferritina* (figura 2.6), *IgG* (figura 2.7) e *Fibrinogeno* (figura 2.8). Per questi parametri sanguigni dalle rispettive figure emerge che il livello nel sangue dei pazienti in ingresso dei parametri infiammatori in questione riportano dei valori anomali sulla coda destra della distribuzione e asimmetria a destra, come mostrato dai boxplot e dagli istogrammi; inoltre, il grafico quantile contro quantile per la verifica della normalità presenta degli allontanamenti sistematici dalla retta di riferimento in prossimità delle code della distribuzione, principalmente a causa dei valori anomali.

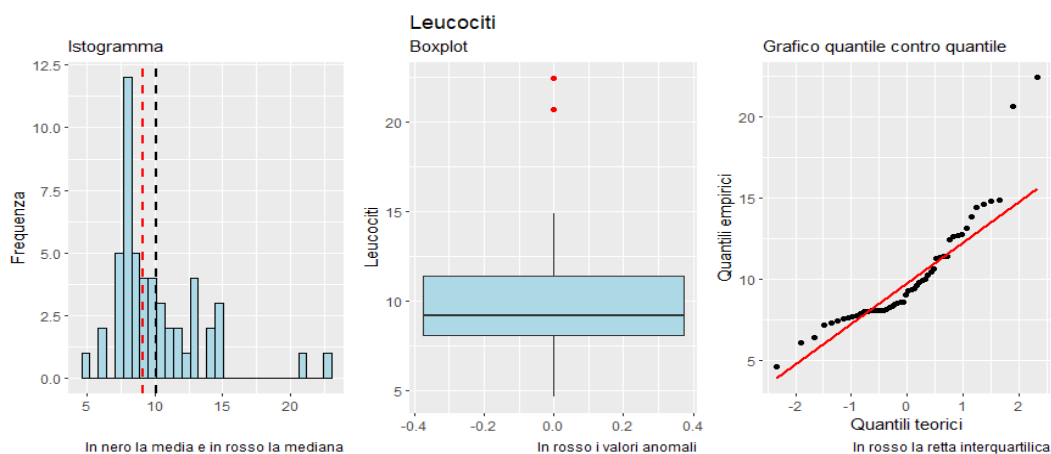


Figura 2.2: Grafici esplorativi per il livello di *Leucociti* nel sangue dei soggetti.

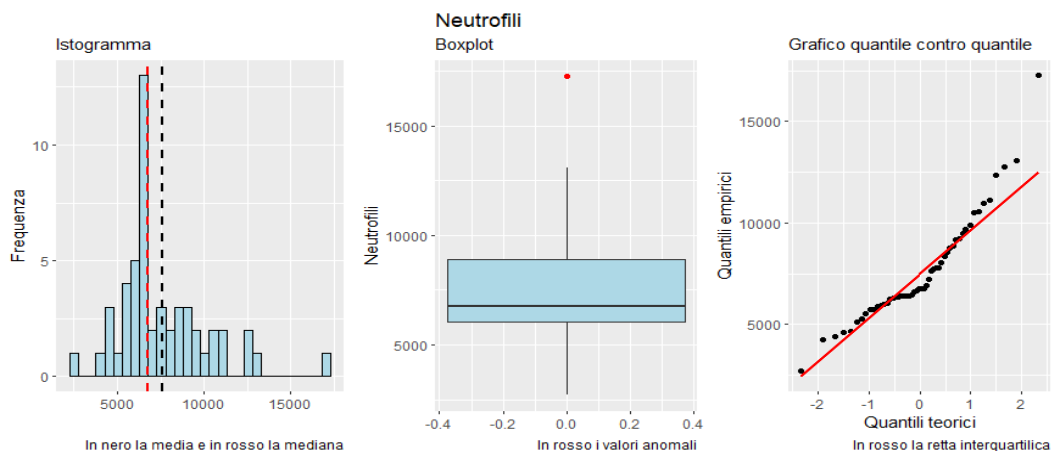


Figura 2.3: Grafici esplorativi per il livello di *Neutrofili* nel sangue dei soggetti.

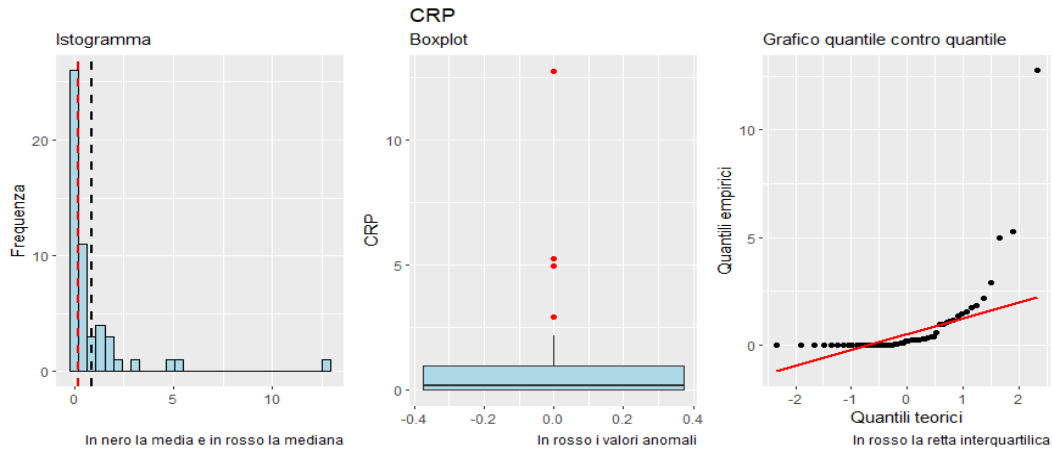


Figura 2.4: Grafici esplorativi per il livello di *CRP* nel sangue dei soggetti.

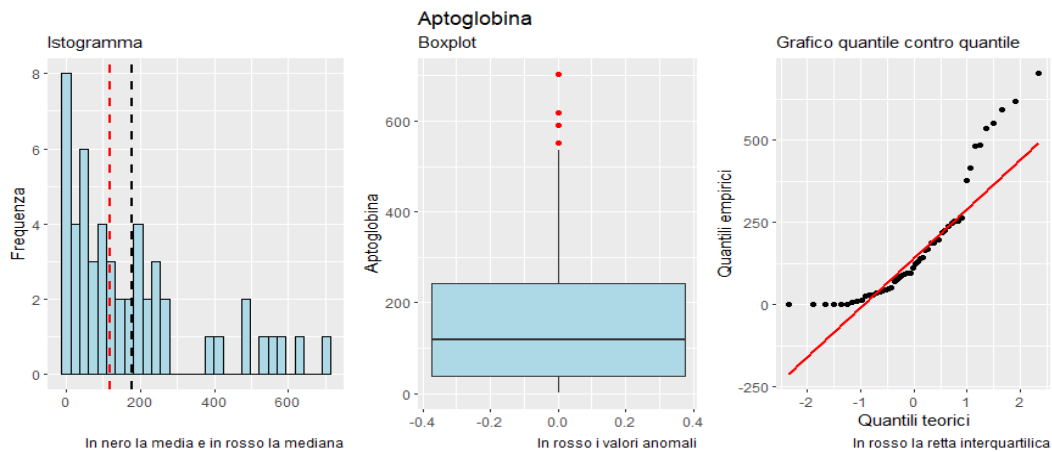


Figura 2.5: Grafici esplorativi per il livello di *Aptoglobina* nel sangue dei soggetti.

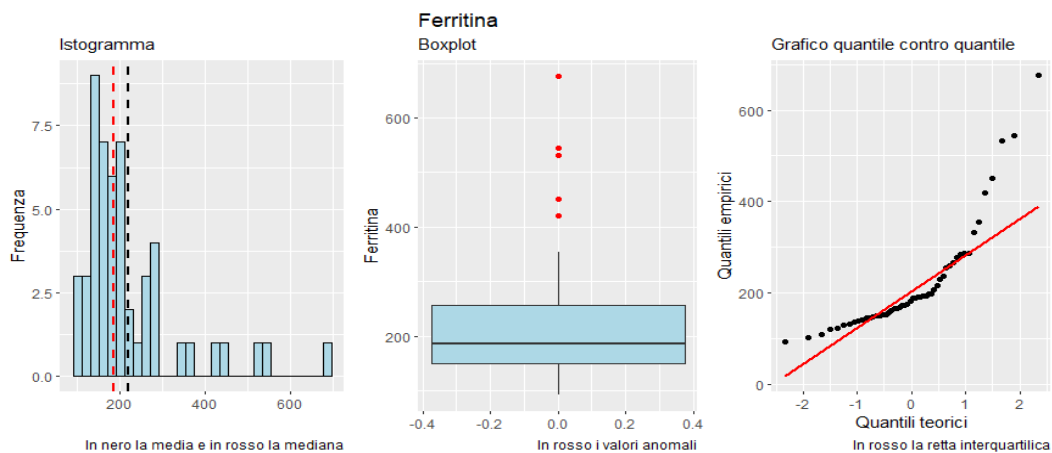


Figura 2.6: Grafici esplorativi per il livello di *Ferritina* nel sangue dei soggetti.

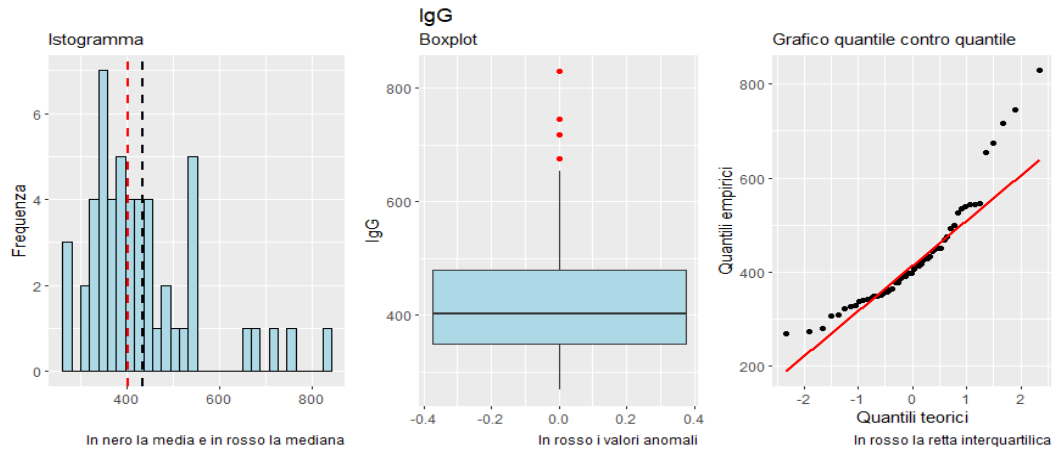


Figura 2.7: Grafici esplorativi per il livello di *IgG* nel sangue dei soggetti.

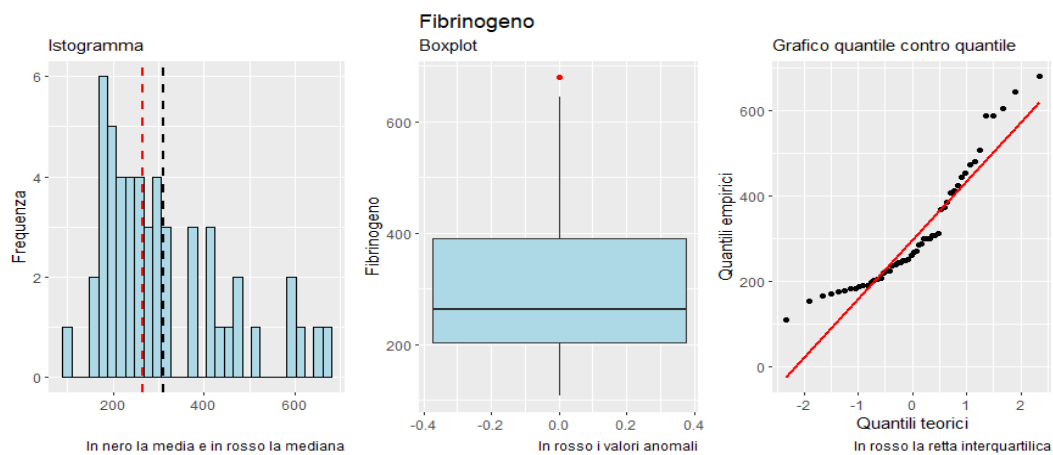


Figura 2.8: Grafici esplorativi per il livello di *Fibrinogeno* nel sangue dei soggetti.

In figura 2.9 e in figura 2.10 vengono riportati i principali grafici esplorativi per il livello nel sangue di *PON1* e *Albumine* nel sangue dei pazienti in ingresso. Dalle figure emerge una distribuzione simmetrica e senza valori anomali, come mostrato dai boxplot e dagli istogrammi; inoltre, il grafico quantile contro quantile per la verifica della normalità non riporta allontanamenti sistematici dalla retta di riferimento.

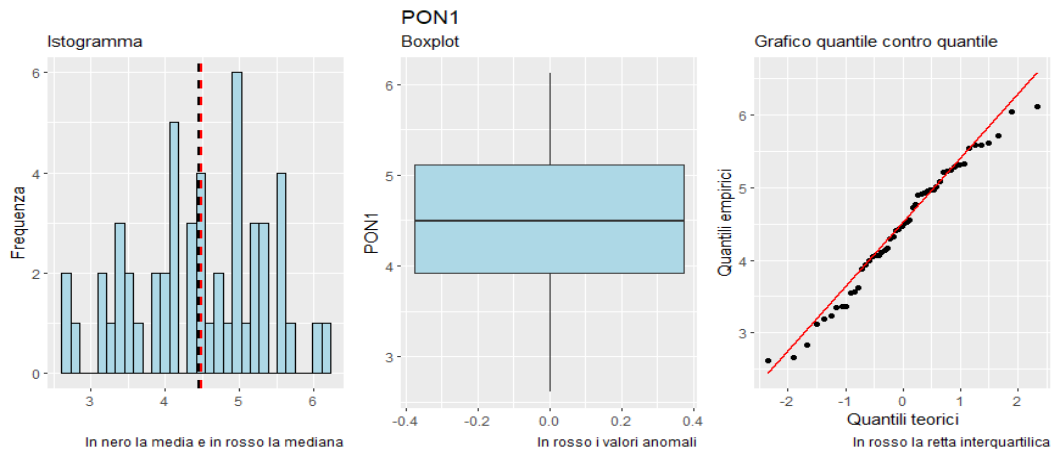


Figura 2.9: Grafici esplorativi per il livello di *PON1* nel sangue dei soggetti.

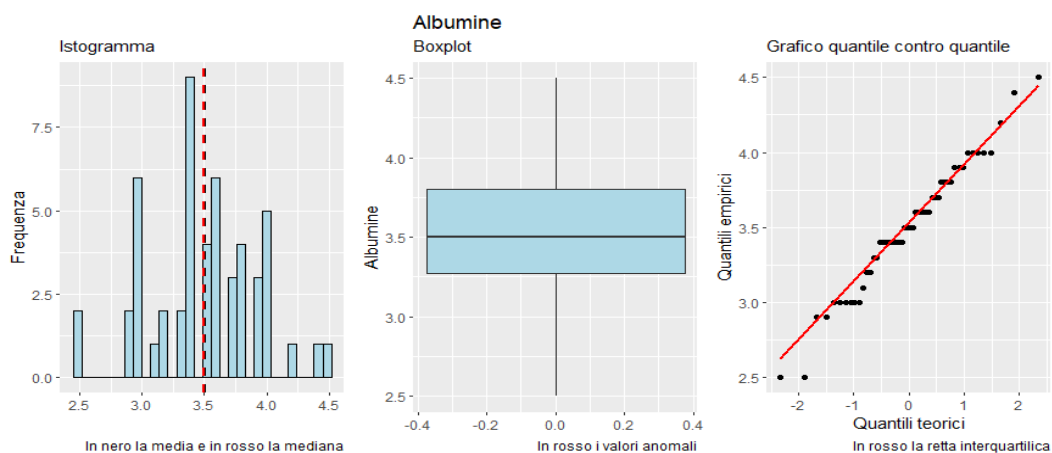


Figura 2.10: Grafici esplorativi per il livello di *Albumine* nel sangue dei soggetti.

In figura 2.11 sono riportati i principali grafici esplorativi per il livello di *IgM* nel sangue dei soggetti.

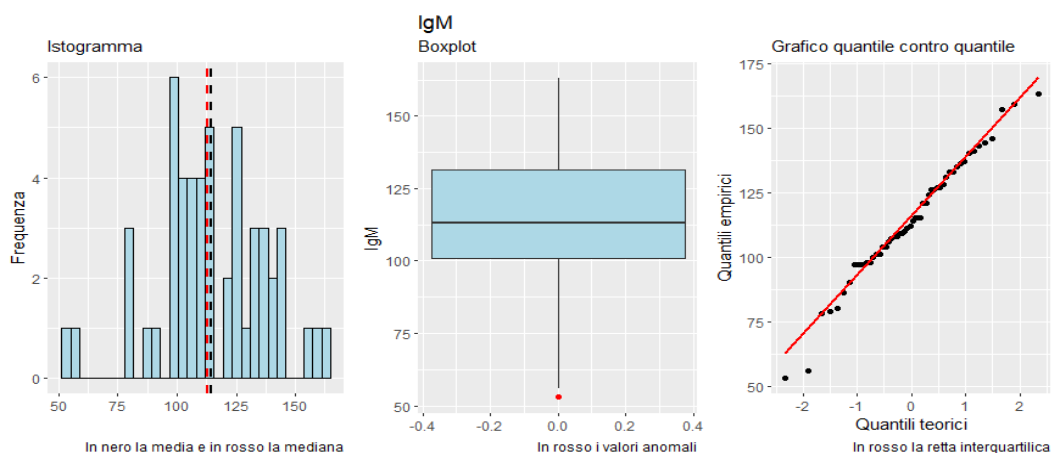


Figura 2.11: Grafici esplorativi per il livello di *IgM* nel sangue dei soggetti.

Dalla figura 2.11 emerge un valore anomalo sulla coda sinistra della distribuzione e una distribuzione simmetrica del livello di IgM nel sangue dei pazienti in ingresso, come mostrato dal boxplot e dall'istogramma. Il grafico quantile contro quantile per la verifica della normalità non mostra allontanamenti sistematici dalla retta di riferimento.

In figura 2.12 sono riportati i principali grafici esplorativi per il livello di IgA nel sangue dei soggetti.

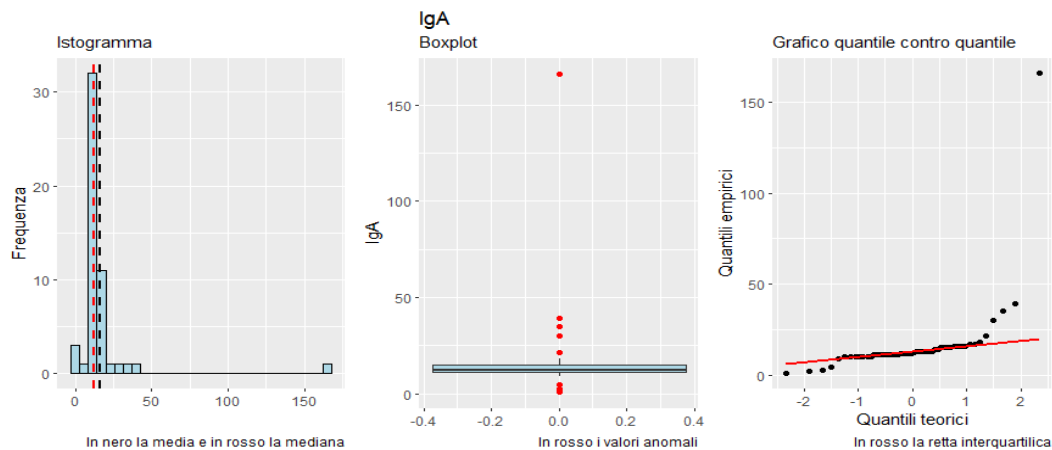


Figura 2.12: Grafici esplorativi per il livello di IgA nel sangue dei soggetti.

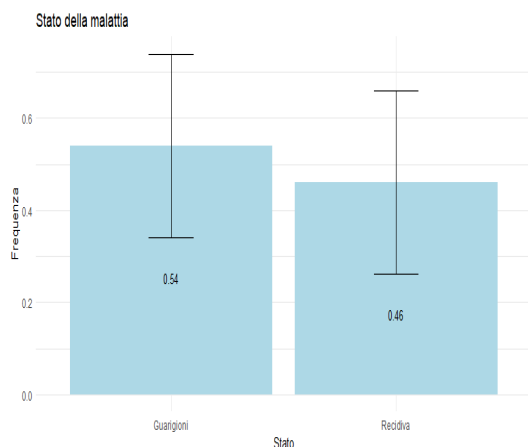
Dalla figura 2.12 emergono diversi valori anomali (in particolare un valore fortemente anomalo sulla coda destra) sulle code della distribuzione del livello di IgA nel sangue dei pazienti in ingresso e una distribuzione simmetrica, come mostrato dall'istogramma e dal boxplot. Il grafico quantile contro quantile per la verifica della normalità evidenzia un' allontanamento sistematico dalla retta di riferimento, in particolar modo sulle code della distribuzione in corrispondenza dei valori anomali.

2.1.2 In fase di recidiva

In questa sezione vengono presentate le analisi esplorative univariate per i soggetti in fase di recidiva.

Stato della malattia

In tabella 2.3 è riportata la distribuzione di frequenza dello stato della malattia dei pazienti, mentre in figura 2.13 è riportato il grafico a barre (con intervalli di confidenza al 95% per le proporzioni dei gruppi) e frequenze relative dello stato di malattia dei pazienti.



Stato	Freq. Assolute	Freq. Percentuali
Guarigione	13	54.17%
Recidiva	11	45.83%
Totale	24	100%

Figura 2.13: Grafico a barre con frequenze relative per lo stato della malattia dei pazienti.

Tabella 2.3: Distribuzione di frequenza dello stato di salute dei pazienti.

Dalla figura emerge che il campione può considerarsi bilanciato rispetto allo stato di salute dei pazienti; infatti, il test per verificare l'ipotesi che la proporzione dei pazienti guariti e la proporzione dei pazienti malati sia uguale a 0.5 viene accettata a tutti i livelli usuali di α (con valore osservato della statistica test 0.579 e un p-value 0.199).

Parametri sanguigni

In tabella 2.4 vengono riportate le principali statistiche di sintesi per i soggetti in fase di recidiva e il p-value del test di Shapiro-Wilk per la verifica della normalità.

Varibaile	Min	1st Qu	Median	Mean	3rd Qu	Max	Sd	Shapiro-Wilk p-value
Leucociti	4.090	7.303	8.795	9.994	11.972	19.250	3.949	0.161
Neutrofili	2550	5178	6620	7584	9590	17000	3599	0.045
PON1	2.800	3.955	4.485	4.786	5.548	7.370	1.156	0.227
CRP	0.010	0.010	0.115	0.198	0.328	1.000	0.256	<0.01
Aptoglobina	1.0	71.0	131.0	231.5	311.5	783.0	239.5	<0.01
Ferritina	120.0	166.8	198.5	200.4	214.8	550.0	82.6	<0.01
Albumine	2.700	3.200	3.500	3.462	3.700	4.200	0.369	0.986
IgG	253	367	396	426	439	1162	166	<0.01
IgM	74.0	93.2	108.5	108.5	122.2	148.0	21.2	0.513
IgA	0.020	10.750	13.000	15.168	17.250	36.000	8.126	<0.01

Tabella 2.4: Principali statistiche di sintesi per i parametri infiammatori dei pazienti in fase di recidiva e p-value del test di Shapiro-Wilk per la verifica della normalità.

Dalla tabella 2.4 si nota che il test di Shapiro-Wilk porta ad accettare l'ipotesi di normalità a tutti i livelli usuali di α per il livello di *Leucociti*, *PON1*, *Albumine* e *IgM*.

Dai principali grafici esplorativi del livello in fase di recidiva dei pazienti di *Leucociti* (figura 2.14), *Neutrofili* (figura 2.14), *CRP* (figura 2.14), *Aptoglobina* (figura 2.14) e *Ferritina* (figura 2.14) emergono dei valori anomali sulla coda destra della distribuzione e asimmetria a destra della distribuzione, come mostrato dagli istogrammi e dai boxplot.

I grafici esplorativi del livello nel sangue dei pazienti in fase di recidiva di *PON1* (figura 2.16), *Albumine* (figura 2.20) e *IgM* (figura 2.22) mostrano simmetria nella distribuzione dei parametri in questione. I grafici esplorativi per il livello di *IgG* (figura 2.21) e il livello di *IgA* (figura 2.23) nel sangue dei pazienti in fase di recidiva riportano valori anomali su entrambe le code della distribuzione.

Infine, il grafico quantile contro quantile per la verifica della normalità non mostra allontanamenti sistematici dalla retta di riferimento del livello nel sangue dei pazienti in fase di recidiva di *Leucociti* (figura 2.14), *PON1* (figura 2.16), *Albumine* (figura 2.20) e *IgM* (figura 2.22), per gli altri parametri infiammatori, invece, si nota un allontanamento sistematico dalla retta di riferimento. I grafici esplorativi per il livello di *IgG* (figura 2.21) e il livello di *IgA* (figura 2.23) nel sangue dei pazienti in fase di recidiva riportano valori anomali su entrambe le code della distribuzione.

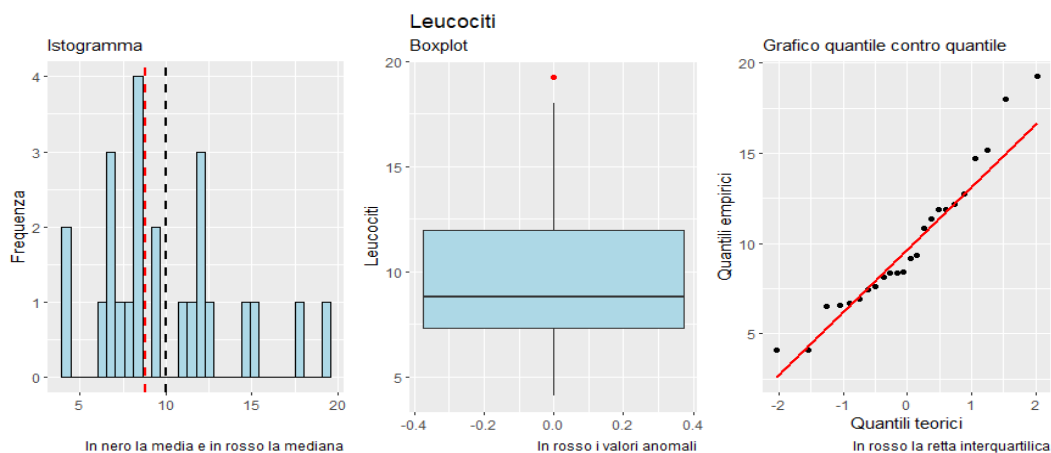


Figura 2.14: Grafici esplorativi per il livello di *Leucociti* nel sangue dei soggetti in fase di recidiva.

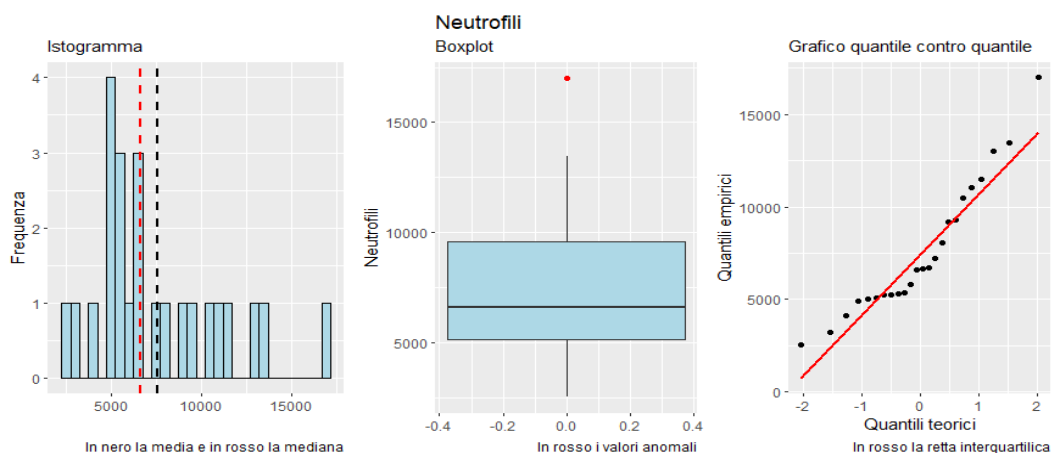


Figura 2.15: Grafici esplorativi per il livello di *Neutrofili* nel sangue dei soggetti in fase di recidiva.

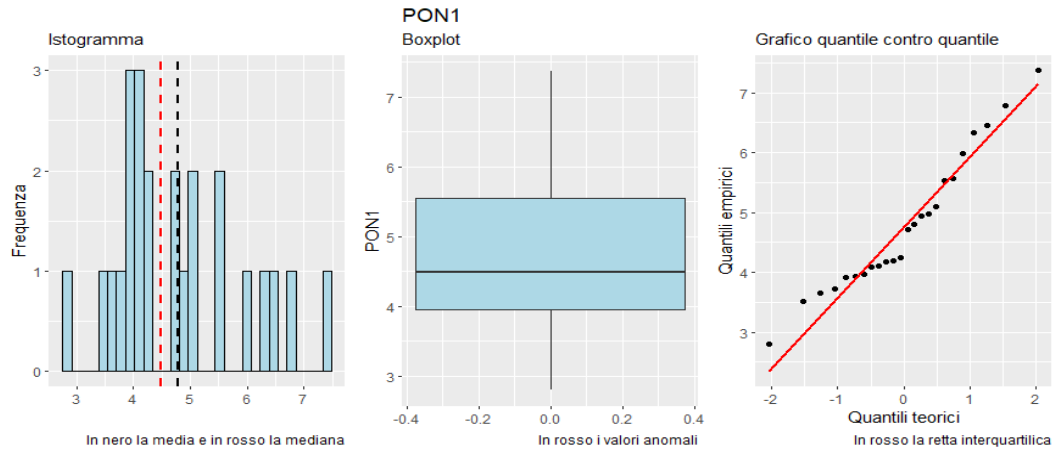


Figura 2.16: Grafici esplorativi per il livello di *PON1* nel sangue dei soggetti in fase di recidiva.

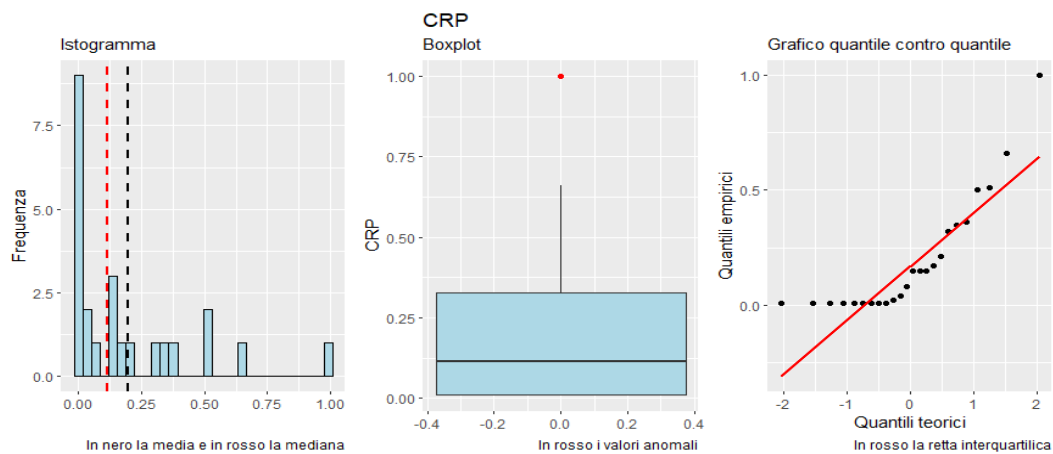


Figura 2.17: Grafici esplorativi per il livello di *CRP* nel sangue dei soggetti in fase di recidiva.

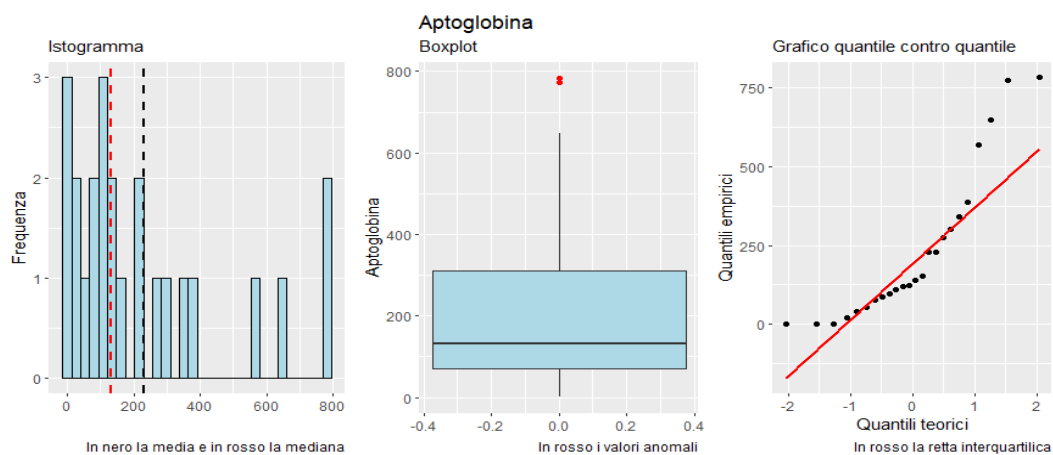


Figura 2.18: Grafici esplorativi per il livello di *Aptoglobina* nel sangue dei soggetti in fase di recidiva.

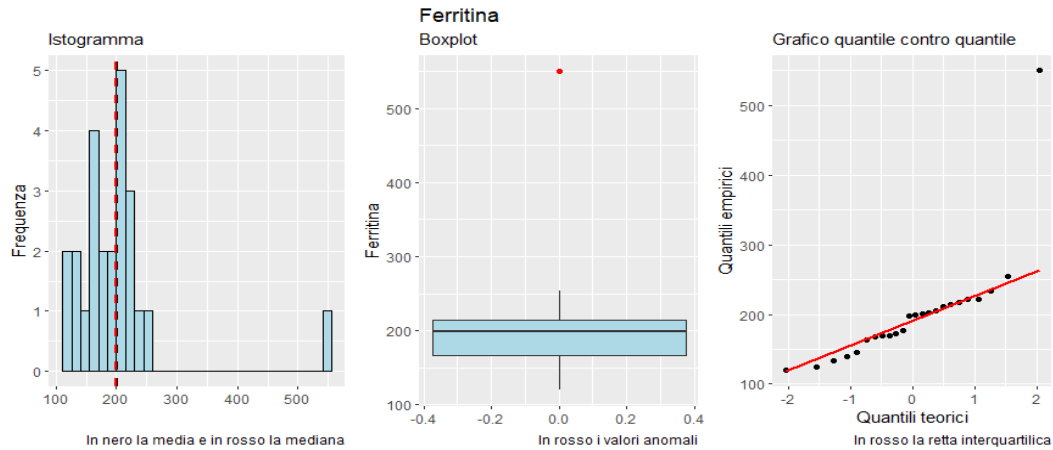


Figura 2.19: Grafici esplorativi per il livello di *Ferritina* nel sangue dei soggetti in fase di recidiva.

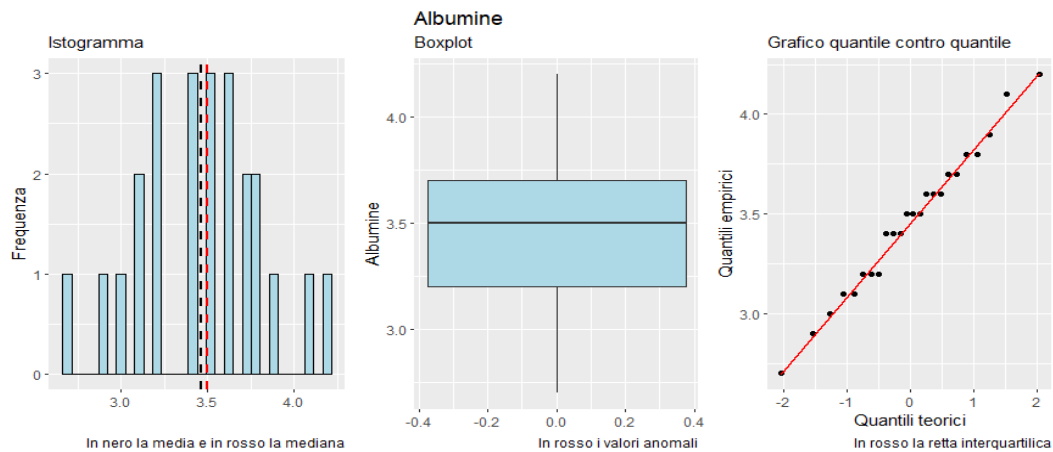


Figura 2.20: Grafici esplorativi per il livello di *Albumine* nel sangue dei soggetti in fase di recidiva.

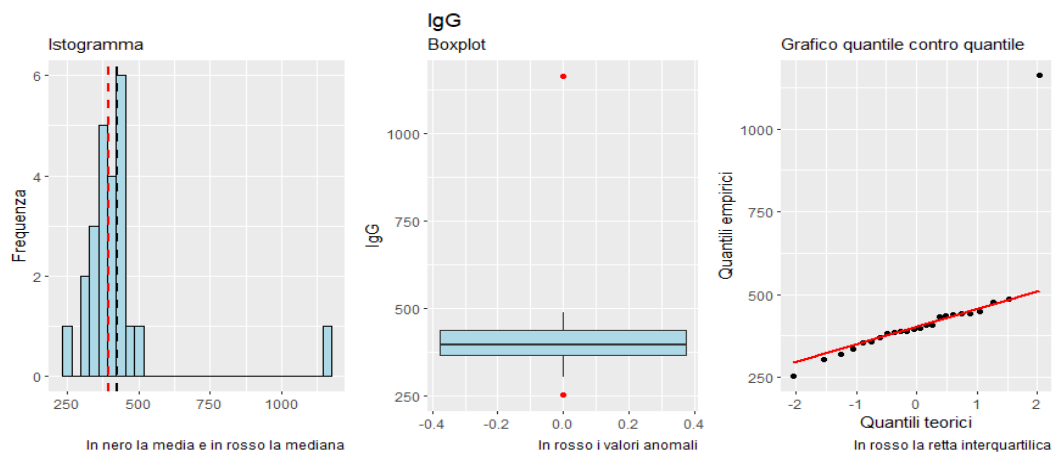


Figura 2.21: Grafici esplorativi per il livello di *IgG* nel sangue dei soggetti in fase di recidiva.

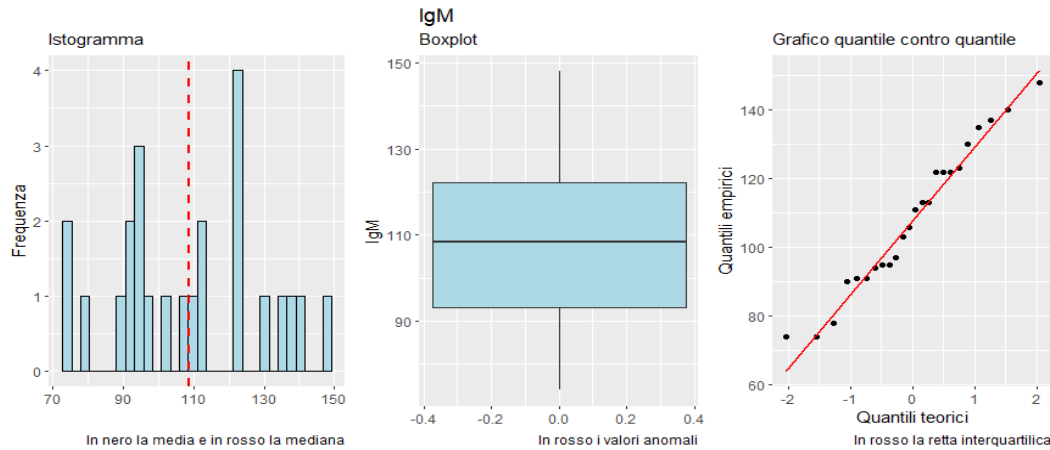


Figura 2.22: Grafici esplorativi per il livello di *IgM* nel sangue dei soggetti in fase di recidiva.

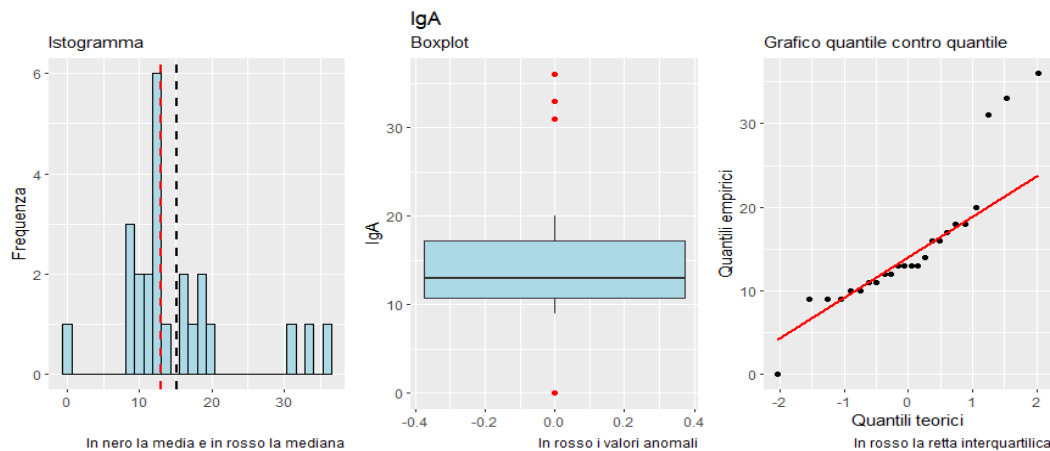


Figura 2.23: Grafici esplorativi per il livello di *IgA* nel sangue dei soggetti in fase di recidiva.

2.2 Analisi esplorativa bivariata

In questa sezione vengono presentate le analisi esplorative bivariate coerenti con gli obiettivi del report.

2.2.1 In ingresso

In questa sezione vengono svolte le analisi bivariate per i soggetti in ingresso nella clinica.

Terapia steroidea rispetto allo stato della malattia

In tabella 2.5 viene riportata la tabella di frequenza assoluta e percentuale con totale di riga per lo stato della malattia e la presenza della terapia steroidea.

Stato/ Terapia	Decesso	Guarigione	Recidiva	Totale
No	11 24%	22 49%	12 27%	45 100%
Sì	4 57%	2 29%	1 14%	7 100%
Totale	15 81%	24 78%	13 41%	

Tabella 2.5: Tabella di frequenza con totale di riga dello stato dell'inflammazione al sistema nervoso centrale nei soggetti e della terapia steroidea in atto.

Dalla tabella emerge che il 57% dei pazienti con la terapia steroidea in atto sono deceduti; mentre, il 49% dei pazienti senza terapia steroidea in atto sono guariti.

Il test esatto di Fisher per verificare l'ipotesi di indipendenza tra lo stato dell'inflammazione al sistema nervoso centrale nei soggetti e la presenza di terapia steroidea in atto riporta un p-value pari a 0.323: l'ipotesi di indipendenza tra le due variabili è accettata a tutti i livelli usuali di α .

Parametri infiammatori in ingresso rispetto allo stato della malattia

In questa sezione vengono effettuati dei test per verificare l'omogeneità delle distribuzioni dei parametri infiammatori rispetto allo stato dell'inflammazione al sistema nervoso centrale e vengono, inoltre, studiate le correlazioni tra i parametri infiammatori. Per verificare l'omoschedasticità, verrà utilizzato il test di Bartlett quando le distribuzioni dei parametri in ingresso suddivise per lo stato della malattia risultano sufficientemente simmetriche; altrimenti, verrà utilizzato il test di Levene, che risulta essere più robusto rispetto a valori anomali. Qualora l'ipotesi di omoschedasticità fosse rispettata per il parametro infiammatorio suddiviso nei tre gruppi e il test di Shapiro Wilk portasse ad accettare l'ipotesi di normalità in tutti e tre i gruppi, per verificare l'omogeneità delle distribuzioni del parametro in ingresso nei tre gruppi verrà utilizzato il test parametrico ANOVA per il confronto tra le medie; invece, se fosse accettata l'ipotesi di normalità nei gruppi e fosse rifiutata l'ipotesi di omoschedasticità per i gruppi, si procederà con un test parametrico ANOVA per gruppi eteroschedastici; infine, se fosse rifiutata l'ipotesi di normalità, indipendentemente dal test per l'omoschedasticità, verrà effettuato il test non parametrico di Kruskal-Wallis per il confronto tra mediane. Viene fissato un livello nominale di α del 5%.

In seguito, vengono riportati i grafici delle medie dei parametri sanguigni rispetto allo stato della malattia qualora le distribuzioni dei parametri infiammatori risultino simmetrici nei tre gruppi; altrimenti, verrà riportato il grafico delle mediane.

Sono riportati i boxplot e il grafico delle medie (con intervalli di confidenza di livello 0.95) per il livello nel sangue di *Neutrofili* (figura 2.24), *PON1* (figura 2.25) e *Albumine* (figura 2.26) rispetto allo stato della malattia. Nonostante la presenza di un valore anomalo sulla coda destra del livello di *Neutrofili* dei soggetti deceduti e alcuni valori anomali per il livello di *Albumine* dei pazienti guariti e con la recidiva, dalle figure emerge simmetria nella distribuzione delle variabili, come mostrato dai boxplot; il grafico delle medie, invece, mostra come le medie dei parametri sanguigni nei tre gruppi riportano valori simili e intervalli di confidenza sovrapponibili.

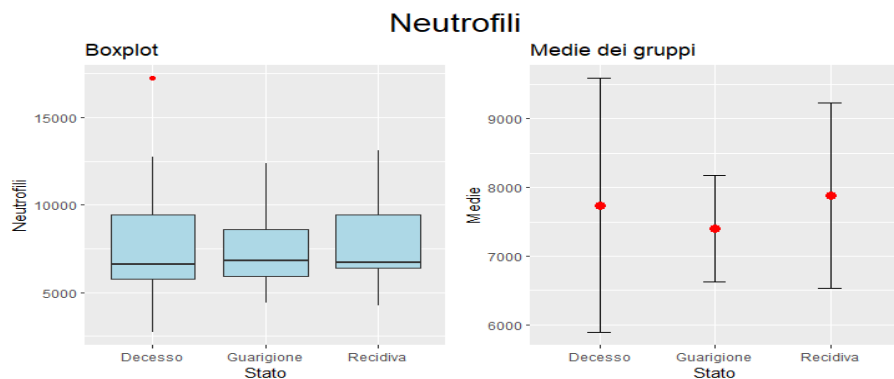


Figura 2.24: Boxplot e grafico delle medie per il livello di *Neutrofili* nel sangue dei pazienti rispetto allo stato della malattia.

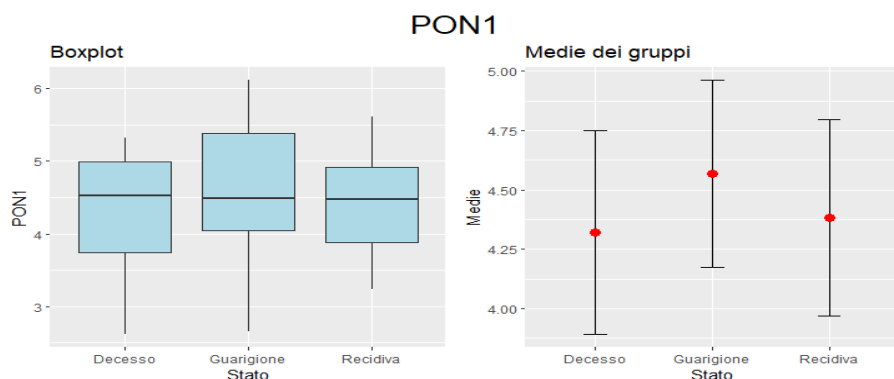


Figura 2.25: Boxplot e grafico delle medie per il livello di *PON1* nel sangue dei pazienti rispetto allo stato della malattia.

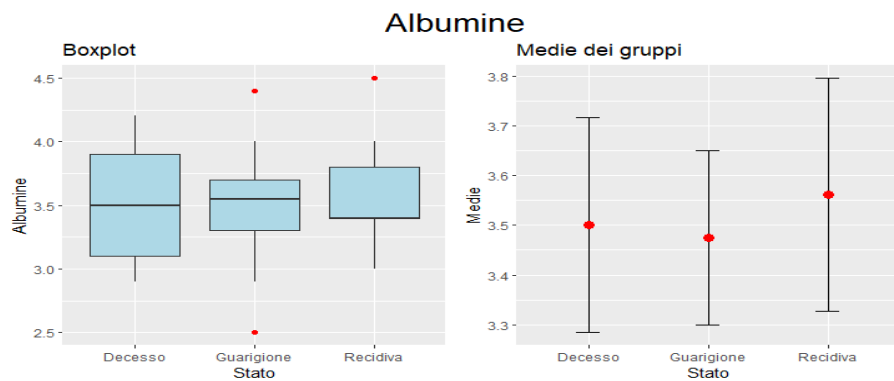


Figura 2.26: Boxplot e grafico delle medie per il livello di *Albumine* nel sangue dei pazienti rispetto allo stato della malattia.

Sono riportati i boxplot e il grafico delle mediane (con intervalli di confidenza 0.95¹) per il livello nel sangue di *Leucociti* (figura 2.27), *CRP* (figura 2.28), *Aptoglobina* (figura 2.29), *Ferritina* (figura 2.30), *IgG* (figura 2.31), *IgM* (figura 2.32), *IgA* (figura 2.33) e *Fibrinogeno* (figura 2.34) rispetto allo stato della malattia. Dalle

¹L'intervallo di confidenza per la mediana è ottenuto tramite replicazioni bootstrap, con l'ausilio del pacchetto R "DescTools" (si veda Davinson, 1997 [3]).

figure emergono diversi valori anomali nelle distribuzioni dei parametri sanguigni rispetto allo stato della malattia e tali distribuzioni risultano asimmetriche, come mostrato dai boxplot; il grafico delle mediane, invece, riporta valori simili e intervalli di confidenza sovrapponibili.

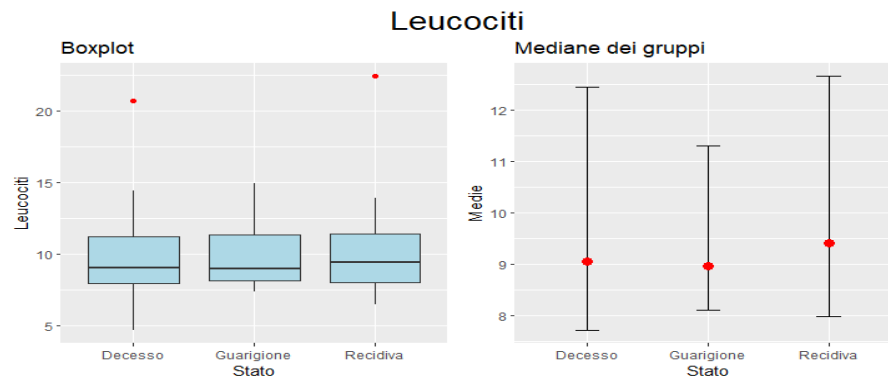


Figura 2.27: Boxplot e grafico delle mediane per il livello di *Leucociti* nel sangue dei pazienti rispetto allo stato della malattia.

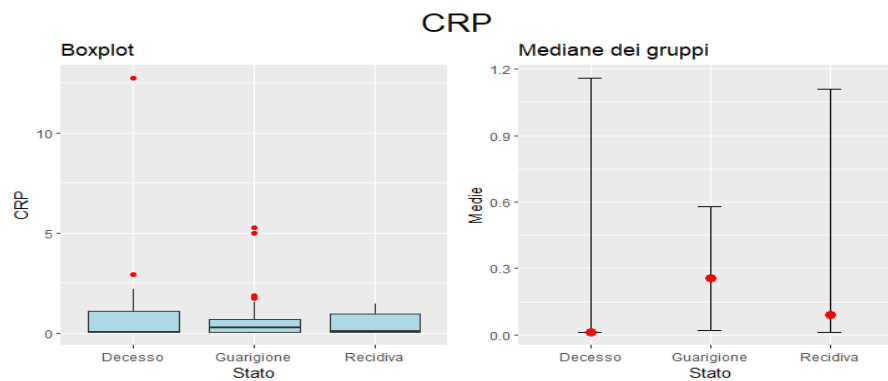


Figura 2.28: Boxplot e grafico delle mediane per il livello di *CRP* nel sangue dei pazienti rispetto allo stato della malattia.

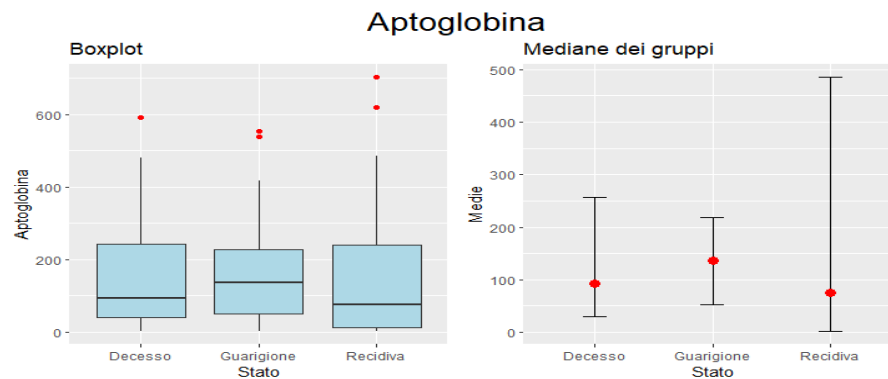


Figura 2.29: Boxplot e grafico delle mediane per il livello di *Aptoglobina* nel sangue dei pazienti rispetto allo stato della malattia.

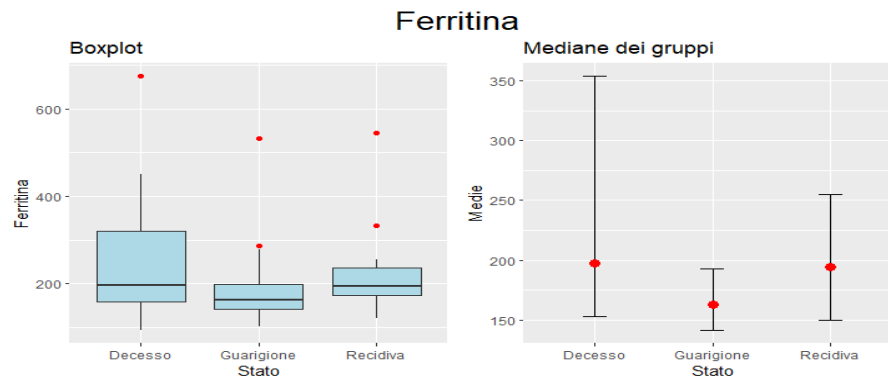


Figura 2.30: Boxplot e grafico delle mediane per il livello di *Ferritina* nel sangue dei pazienti rispetto allo stato della malattia.

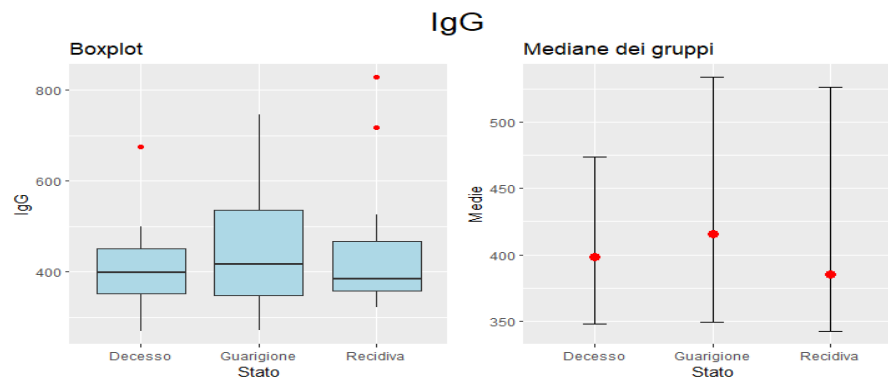


Figura 2.31: Boxplot e grafico delle mediane per il livello di *IgG* nel sangue dei pazienti rispetto allo stato della malattia.

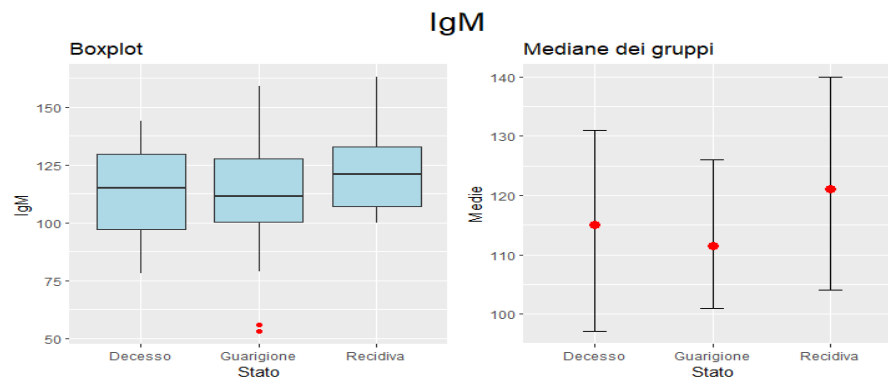


Figura 2.32: Boxplot e grafico delle mediane per il livello di *IgM* nel sangue dei pazienti rispetto allo stato della malattia.

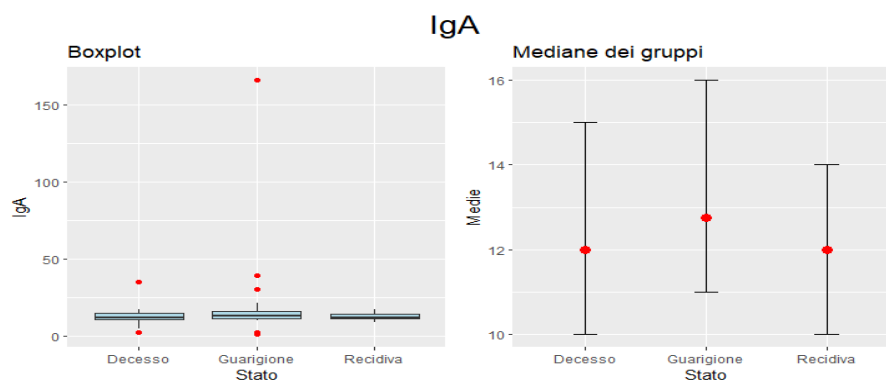


Figura 2.33: Boxplot e grafico delle mediane per il livello di *IgA* nel sangue dei pazienti rispetto allo stato della malattia.

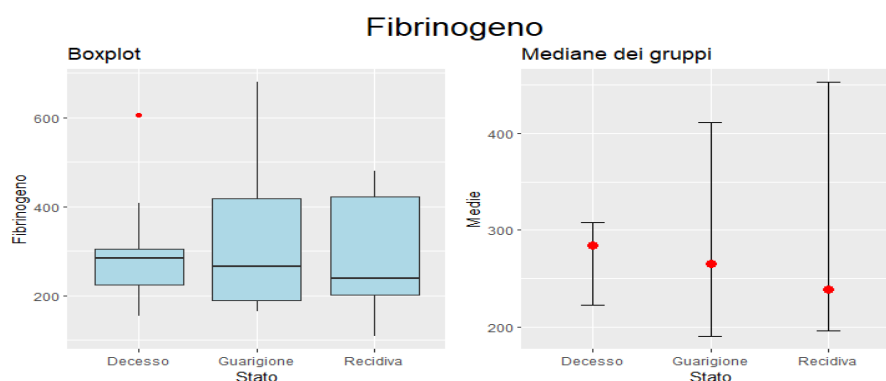


Figura 2.34: Boxplot e grafico delle mediane per il livello di *Fibrinogeno* nel sangue dei pazienti rispetto allo stato della malattia.

In tabella 2.6 vengono riportati i p-value dei test necessari per verificare l'omogeneità dei parametri infiammatori in ingresso rispetto allo stato dell'infiammazione al sistema nervoso centrale. Viene riportato in tabella solo il p-value del test di Shapiro-Wilk più basso per il parametro suddiviso nei tre gruppi.

Variabile	Bartlett test	Levene test	Shapiro-Wilk test	ANOVA	Kruskall-Wallis
Leucociti	-	0.6	<0.01	-	0.88
Neutrofili	-	0.35	0.07	0.86	-
PON1	0.59		0.17	0.67	
CRP	-	0.44	<0.01	-	0.64
Aptoglobina		0.49	<0.01	-	0.7
Ferritina	-	0.16	<0.01	-	0.14
Albumine	-	0.95	0.23	0.14	-
IgG	-	0.55	<0.01	-	0.88
IgM	-	0.9	0.11	-	0.58
IgA	-	0.42	<0.01	-	0.75
Fibrinogeno	-	0.33	<0.01	-	0.91

Tabella 2.6: P-value dei test per verificare che la distribuzione dei parametri infiammatori in ingresso rispetto allo stato dell'infiammazione al sistema nervoso centrale del soggetto sia omogenea.

Dalla tabella 2.6 emerge che, a un livello di significatività del 5%, i test portano ad accettare l'ipotesi di omogeneità in distribuzione tra i parametri infiammatori in ingresso suddivisi nei tre gruppi.

In conclusione, a seguito delle analisi esplorative bivariate e dei test, a un livello di significatività del 5%, non risultano esserci differenze in distribuzione tra i parametri sanguigni suddivisi nei tre gruppi per i pazienti in ingresso nella clinica.

Parametri infiammatori incrociati in ingresso

In figura 2.35 vengono riportati i grafici di dispersione per ogni incrocio tra i parametri sanguigni dei pazienti in ingresso, con le rispettive correlazioni di Spearman e la significatività del test: gli asterischi indicano che l'indice di correlazione di Spearman è significativamente diverso da zero a un livello di α del 10% (un asterisco), 5% (due asterischi) oppure 1% (tre asterischi).

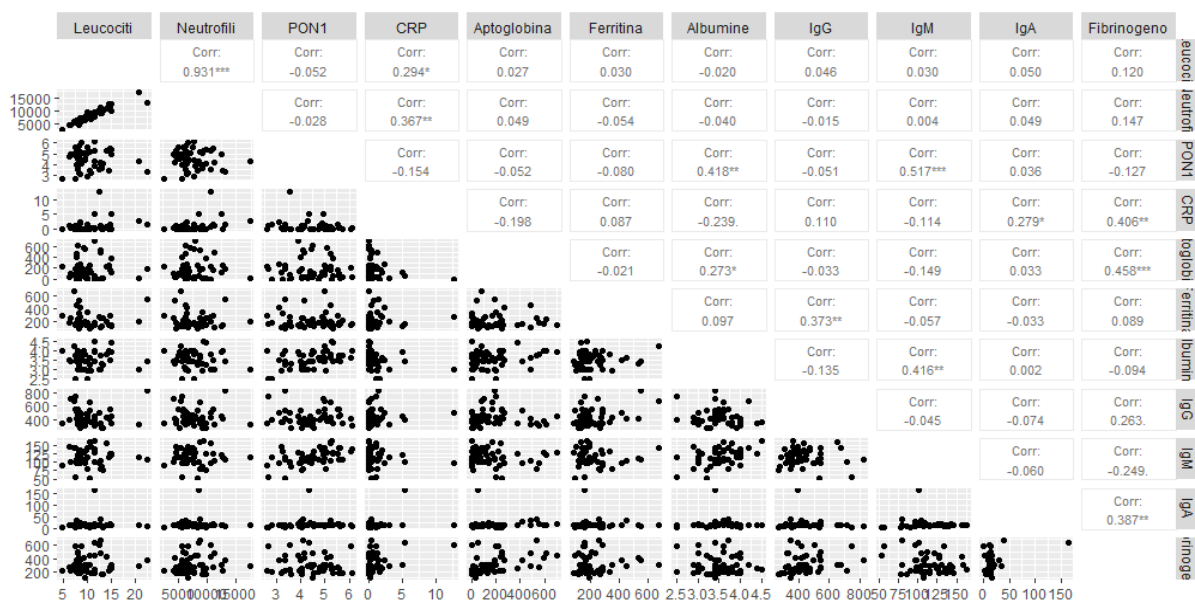


Figura 2.35: Grafici di dispersione per gli incroci tra i parametri infiammatori.

Dalla figura 2.35 si nota come tra il livello di *Leucociti* e *Neutrofili* intercorra una correlazione molto elevata, pari a 0.931 (significativa a tutti i livelli usuali di α), una correlazione elevata tra il livello di *IgM* e il livello di *PON1* pari a 0.517 (significativa a tutti i livelli usuali di α) e una correlazione elevata tra il livello di *Fibrinogeno* e il livello di *Aptoglobina* pari a 0.456 (significativa a tutti i livelli usuali di α). Inoltre, si nota la presenza di forti valori anomali per il livello di *CRP* e *IgA* rispetto alle altre variabili.

2.2.2 In fase di recidiva

In questa sezione vengono svolte le analisi bivariate per i soggetti in fase di recidiva.

Parametri infiammatori in fase di recidiva rispetto allo stato della malattia

Vista la bassa numerosità campionaria, non si può ricorrere a teorie asintotiche per grandi campioni per valutare l'omogeneità in distribuzione dei parametri sanguigni rispetto allo stato della malattia. Di conseguenza, qualora fossero verificate le ipotesi di normalità e di omoschedasticità, verrà effettuato il test *t* di Student a due campioni indipendenti per verificare l'omogeneità delle distribuzioni dei parametri sanguigni rispetto allo stato della malattia; se fosse rifiutata l'ipotesi di omoschedasticità si ricorrerà al test *t* di Welch; infine, se fosse rifiutata la normalità, verrà utilizzato il test non parametrico basato sui ranghi di Mann-Whitney. Per verificare l'omoschedasticità tra i parametri sanguigni in fase di recidiva rispetto allo stato della malattia

dei pazienti si ricorre al test esatto di Fisher, se è accettata l'ipotesi di normalità, altrimenti, si ricorre al test non parametrico di Levene. Viene fissato un livello di significatività del 5%.

Sono riportati i boxplot e i grafici delle medie (con intervallo di confidenza 0.95 per le medie) per il livello in fase di recidiva di *Leucociti* (figura 2.36), *Neutrofili* (figura 2.37), *PON1* (figura 2.38), *Albumine* (figura 2.39) e *IgM* (figura 2.40) rispetto lo stato della malattia dei soggetti. Dalle figure si nota un valore anomalo sulla coda destra del livello di *Leucociti*, *Neutrofili* e *PON1* dei pazienti con la recidiva, come mostrato dai boxplot.

Infine, il grafico delle medie del livello di *Albumine*, *IgM* e *Neutrofili* riporta dei valori diversi per questi parametri sanguigni, mentre il livello degli altri parametri sanguigni rispetto allo stato della malattia riportano valori delle medie simili e intervalli di confidenza sovrapponibili.

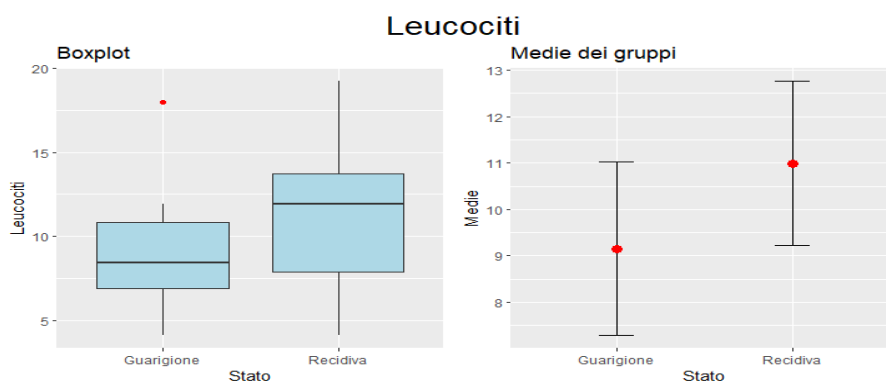


Figura 2.36: Boxplot e grafico delle medie per il livello di *Leucociti* nel sangue dei pazienti rispetto allo stato della malattia.

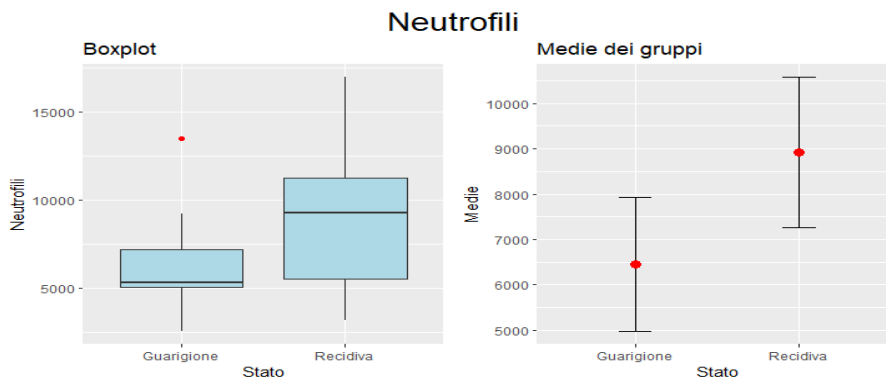


Figura 2.37: Boxplot e grafico delle medie per il livello di *Neutrofili* nel sangue dei pazienti rispetto allo stato della malattia.

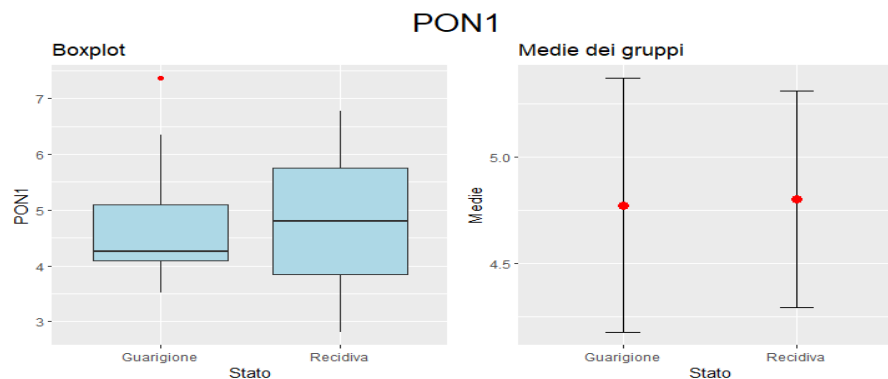


Figura 2.38: Boxplot e grafico delle medie per il livello di *PON1* nel sangue dei pazienti rispetto allo stato della malattia.

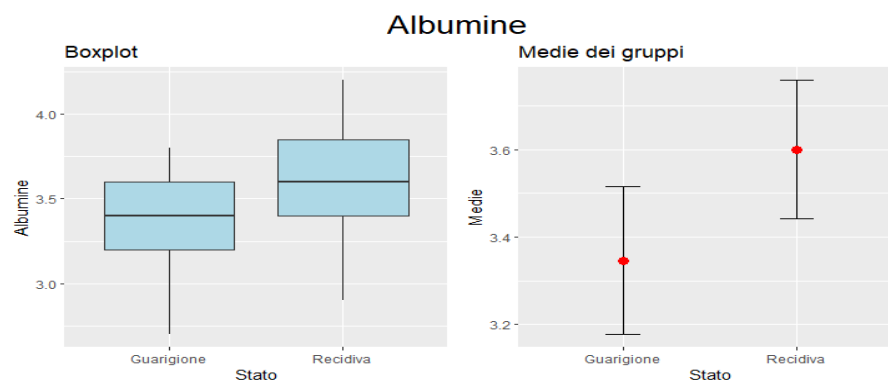


Figura 2.39: Boxplot e grafico delle medie per il livello di *Albumine* nel sangue dei pazienti rispetto allo stato della malattia.

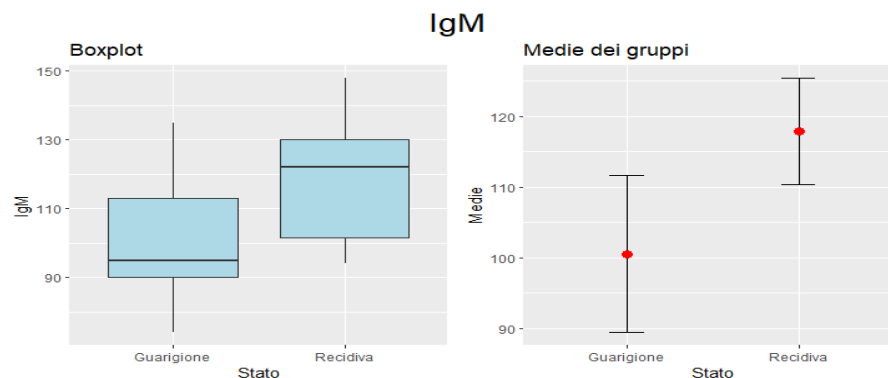


Figura 2.40: Boxplot e grafico delle medie per il livello di *IgM* nel sangue dei pazienti rispetto allo stato della malattia.

Vengono riportati i boxplot e i grafici delle mediane (con intervalli di confidenza 0.95 per la mediana²) del livello in fase di recidiva di *CRP* (figura 2.41), *Aptoglobina* (figura 2.42), *Ferritina* (2.43), *IgG* (figura 2.45) e *IgA* (figura 2.45) rispetto allo stato della malattia dei soggetti. Dalle figure emergono delle distribuzioni

²L'intervallo di confidenza per la mediana è ottenuto tramite replicazioni bootstrap, con l'ausilio del pacchetto R "DescTools" (si veda Davinson, 1997 [3]).

dei parametri sanguigni in fase di recidiva con diversi valori anomali, soprattutto sulla coda destra; inoltre, il livello dei parametri sanguigni nel sangue dei pazienti in fase di recidiva rispetto allo stato della malattia riportano variabilità e mediane simili, con intervalli di confidenza sovrapponibili.

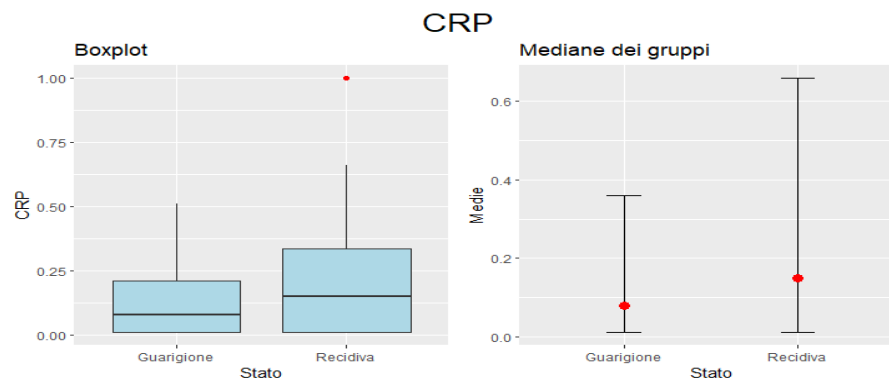


Figura 2.41: Boxplot e grafico delle mediane per il livello di *CRP* nel sangue dei pazienti rispetto allo stato della malattia.

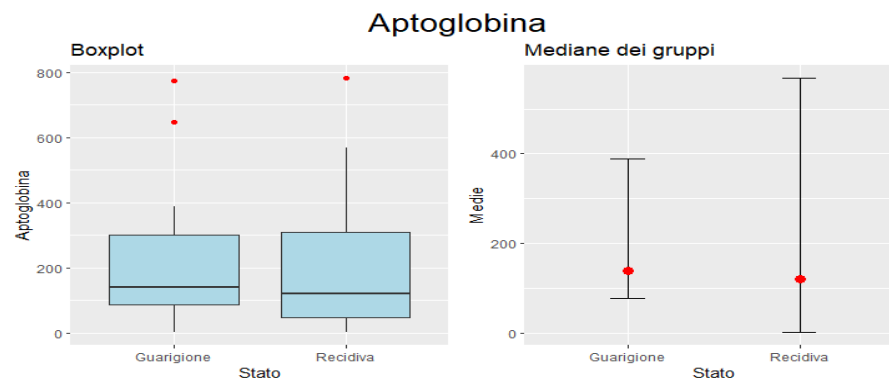


Figura 2.42: Boxplot e grafico delle mediane per il livello di *Aptoglobina* nel sangue dei pazienti rispetto allo stato della malattia.

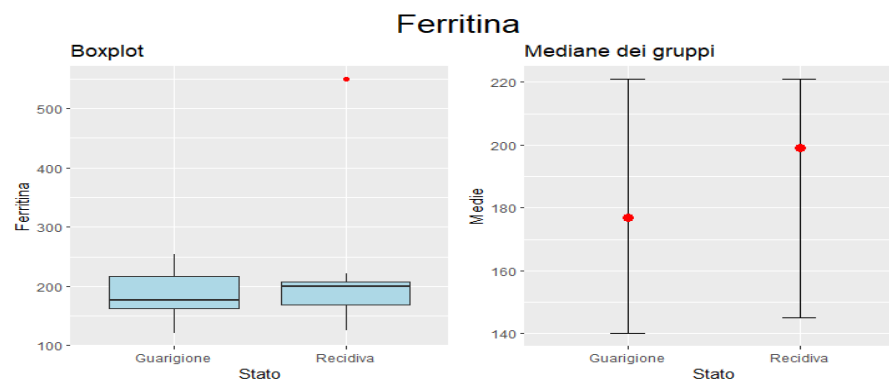


Figura 2.43: Boxplot e grafico delle mediane per il livello di *Ferritina* nel sangue dei pazienti rispetto allo stato della malattia.

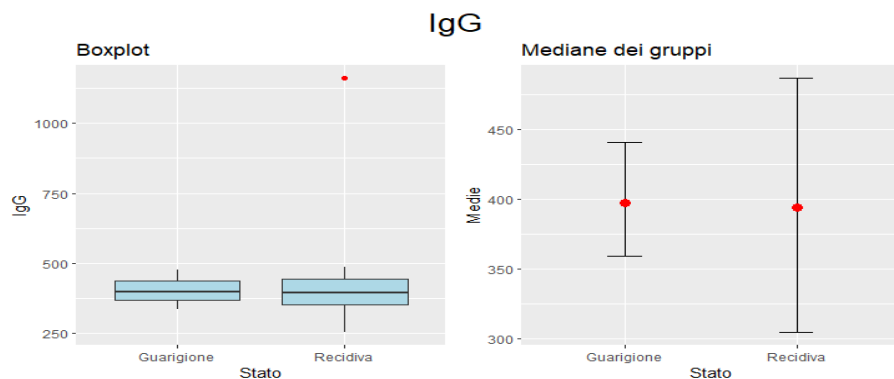


Figura 2.44: Boxplot e grafico delle mediane per il livello di *IgG* nel sangue dei pazienti rispetto allo stato della malattia.

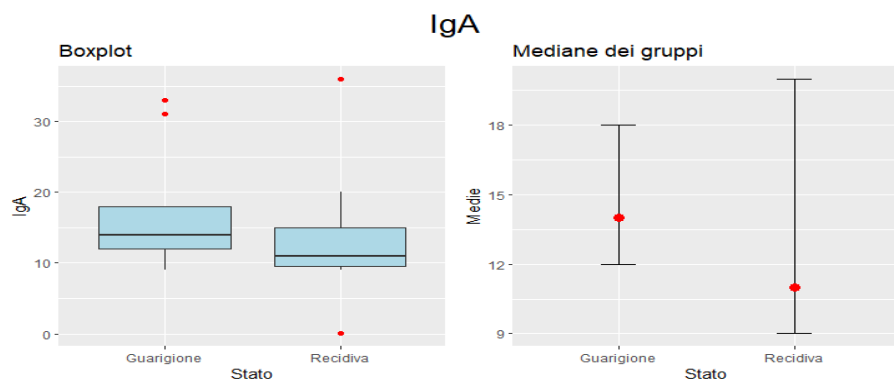


Figura 2.45: Boxplot e grafico delle mediane per il livello di *IgA* nel sangue dei pazienti rispetto allo stato della malattia.

In tabella 2.7 vengono riportati i p-value dei test per verificare l'omogeneità in distribuzione dei parametri sanguigni in fase di recidiva dei pazienti divisi per stato della malattia. In tabella viene riportato il p-value del test di Shapiro-Wilk dei parametri sanguigni rispetto allo stato della malattia che risulta minore.

Variabile	Shapiro-Wilk test	Fisher test	Levene test	Test t di Student	Mann-Whitney test
Leucociti	0.149	0.398	-	0.264	-
Neutrofili	0.093	0.171	-	0.093	-
PON1	0.057	0.622	-	0.954	-
CRP	<0.01	-	0.308	-	0.721
Aptoglobina	0.014	-	0.783	-	0.75
Ferritina	<0.01	-	0.524	-	0.954
Albumine	0.737	0.426	-	0.094	-
IgG	<0.01	-	0.185	-	0.954
IgM	0.435	0.741	-	0.043	-
IgA	<0.01	-	0.8	-	0.172

Tabella 2.7: P-value dei test per verificare che la distribuzione dei parametri infiammatori in fase di recidiva rispetto allo stato dell'infiammazione al sistema nervoso centrale del soggetto sia omogenea.

Dalla tabella 2.7 emerge che, a un livello di significatività del 5%, il test *t di Student* porta a rifiutare l'ipotesi che le medie del livello di *IgM* nel sangue dei pazienti in fase di recidiva rispetto ai due gruppi sia la stessa;

invece, per gli altri parametri sanguigni, viene accettata l'ipotesi di omogeneità in distribuzione del livello dei parametri sanguigni rispetto allo stato della malattia.

In conclusione, a seguito delle analisi esplorative bivariate e dei test, a un livello di significatività del 5%, tutte le distribuzioni dei parametri sanguigni in fase di recidiva rispetto allo stato della malattia risultano omogenee, fatta eccezione per il livello di *IgM*.

Parametri infiammatori incrociati in fase di recidiva

In figura 2.46 vengono riportati i grafici di dispersione per ogni incrocio tra i parametri sanguigni dei pazienti in ingresso, con le rispettive correlazioni di Spearman e la significatività del test: gli asterischi indicano che l'indice di correlazione di Spearman è significativamente diverso da zero a un livello di α del 10% (un asterisco), 5% (due asterischi) oppure 1% (tre asterischi).

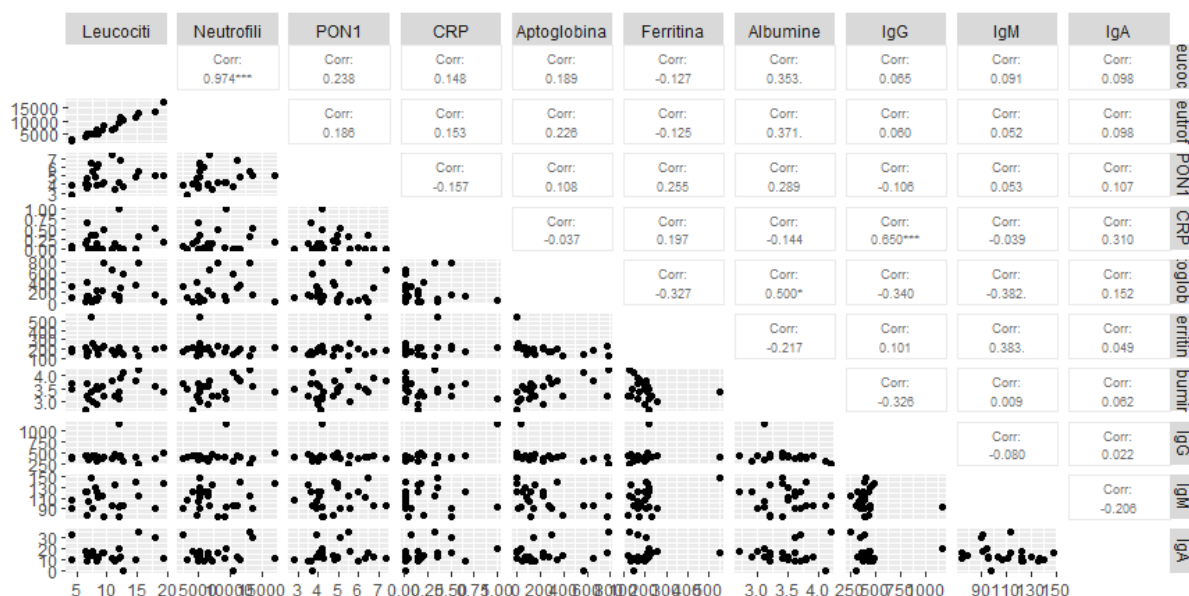


Figura 2.46: Grafici di dispersione per gli incroci tra i parametri infiammatori.

Dalla figura 2.46 emerge una correlazione molto elevata tra il livello di *Leucociti* e il livello di *Neutrofili* nel sangue dei pazienti in fase di recidiva, pari a 0.974 e una correlazione pari a 0.66 tra il livello di *CRP* e il livello di *IgG*, entrambe le correlazioni risultano significativamente diverse da 0 a tutti i livelli usuali α . Infine, si nota la presenza di un valore fortemente anomalo nel grafico di dispersione per il livello di *IgG* e di *Ferritina* con gli altri parametri sanguigni.

Capitolo 3

Analisi multivariate

In questo capitolo vengono presentati i risultati di alcune tecniche di analisi multivariata per i dati a disposizione, come l'analisi delle componenti principali e il metodo delle k-medie (si veda Johnson *et al.*, 2013 [7]). Gli obiettivi di questo capitolo sono i seguenti:

1. Ridurre la dimensionalità delle variabili tramite l'analisi delle componenti principali: valutare se pochi parametri infiammatori (solitamente 2 o 3) riescono a descrivere maggior parte dell'informazione descritta da tutti parametri (solitamente circa il 75% della variabilità).
2. Analisi dei gruppi tramite il metodo delle k-medie: quantificare l'accuratezza con cui i parametri sanguigni riescono a distinguersi nello stato della malattia o nella presenza della terapia steroidea in atto.

3.1 In ingresso

In questa sezione vengono svolte le analisi multivariate per i soggetti in ingresso nella clinica.

3.1.1 Analisi delle componenti principali

Viene svolta un'analisi delle componenti principali sui parametri infiammatori dei pazienti in ingresso per valutare quante componenti principali sono necessarie per descrivere circa il 75% della variabilità totale dei parametri infiammatori dei pazienti in ingresso.

In figura 3.1 viene riportato lo Screeplot della varianza spiegata dalle componenti principali dei parametri infiammatori in ingresso. In rosso il 75% della variabilità cumulata richiesta.

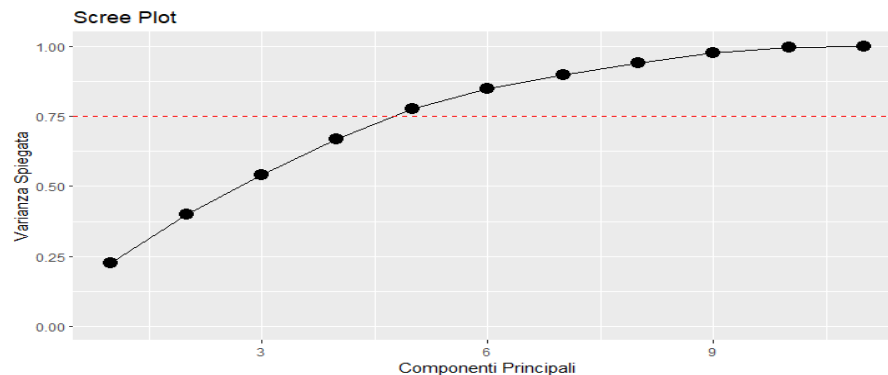


Figura 3.1: Screeplot della varianza spiegata dalle componenti principali dei parametri infiammatori dei pazienti in ingresso.

Dalla figura 3.1 non emergono gomiti particolari e si nota che circa il 75% della variabilità complessiva viene spiegata dalle prime 5 componenti principali. In tabella 3.1 vengono riportati i risultati dell'analisi delle componenti principali.

Componenti Principali	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	CP9	CP10	CP11
Deviazione standard	1.58	1.38	1.25	1.17	1.11	0.87	0.76	0.68	0.62	0.46	0.24
Varianza cumulata	0.23	0.4	0.54	0.67	0.78	0.85	0.9	0.94	0.98	0.99	1

Tabella 3.1: Risultati dell'analisi delle componenti principali.

Dalla tabella 3.1 si nota che sono necessarie 5 componenti principali per spiegare una proporzione di varianza sufficientemente elevata, perché i parametri sanguigni dei pazienti in ingresso risultano poco correlati tra loro; infatti, la varianza spiegata della quinta componente principale dei parametri infiammatori standardizzati riporta una deviazione standard di 1.11 (oltre la media che è pari ad 1) e una proporzione di varianza spiegata cumulata pari al 78%.

In tabella 3.2 vengono riportati i pesi delle prime due componenti principali componenti principali associati a ciascun parametro infiammatorio per i pazienti in ingresso.

Variabile	PC1	PC2
Leucociti	0.40280911	-0.467797588
Neutrofili	0.41615212	-0.473515760
PON1	-0.28607805	-0.409526402
CRP	0.43320927	-0.058837170
Aptoglobina	0.07454190	-0.004372444
Ferritina	0.08665275	0.104536354
Albumine	-0.26901923	-0.372463669
IgG	0.16302631	0.150960000
IgM	-0.28997382	-0.451201952
IgA	0.20819056	-0.033999168
Fibrinogeno	0.39447315	0.092576111

Tabella 3.2: Pesi delle prime due componenti principali per i pazienti in ingresso.

Dalla tabella emerge che la prima componente principale è maggiormente correlata con il livello di *Leucociti* (pesi pari a 0.403), *Neutrofili* (peso pari a 0.416) e *CRP* (peso pari a 0.433), che entrano tutte con correlazione positiva; mentre, per la seconda componente principale, le variabili più influenti risultano essere il livello di *Leucociti* (peso pari a -0.468), *Neutrofili* (peso pari a -0.473) e *IgM* (-0.451), che entrano tutte con segno negativo.

In figura 3.2 è riportato il grafico delle correlazioni per i parametri infiammatori sul piano delle prime due componenti principali.

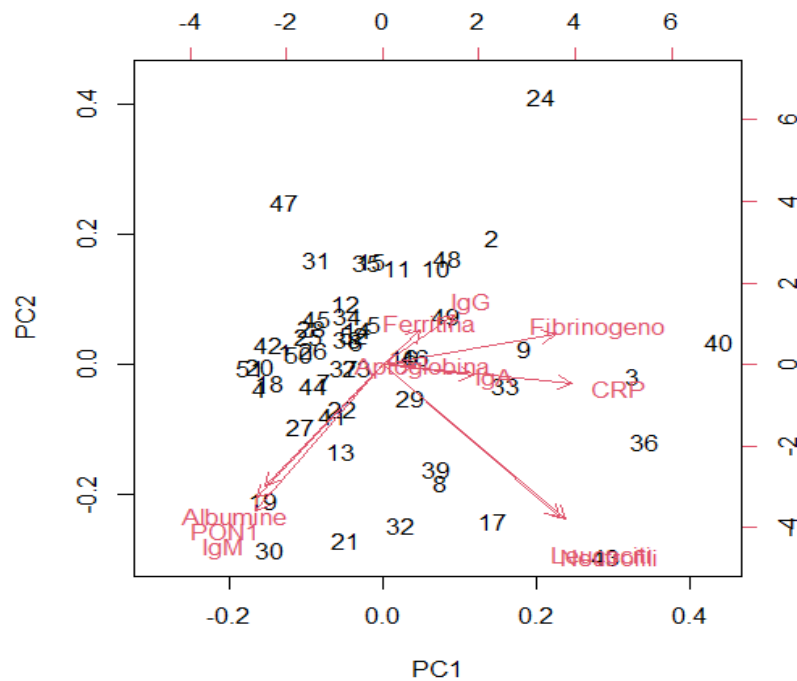


Figura 3.2: Grafico delle correlazioni sul piano delle prime due componenti principali.

Dalla figura si nota che le variabili più influenti sono quelle citate sopra e si nota anche la bassa correlazione tra le variabili. Poiché sono necessarie 5 componenti principali per descrivere una sufficiente proporzione di variabilità totale dei parametri infiammatori dei pazienti in ingresso, l'analisi delle componenti principali non risulta adeguata per ridurre la dimensionalità delle variabili.

3.1.2 Analisi dei gruppi

Si vuole studiare se i parametri infiammatori possono caratterizzare i tre gruppi che rappresentano lo stato della malattia dei soggetti. In figura 3.3 vengono riportati i grafici di dispersione per ogni incrocio tra i parametri infiammatori, divisi per lo stato della malattia.



Figura 3.3: Grafici di dispersione per gli incroci tra i parametri infiammatori rispetto allo stato della malattia.

Dalla figura 3.3 non sembra che i parametri infiammatori si dividano in tre gruppi ben distinti. Tramite il metodo delle k-medie, i soggetti vengono classificati in tre gruppi distinti sulla base della somiglianza dei parametri infiammatori in ingresso. In tabella 3.3 viene riportata la tabella di corretta classificazione, dove vengono posti sulle colonne gli stati della malattia osservati per i pazienti e sulle righe i gruppi trovati empiricamente con il metodo delle k-medie.

Stato	Decesso	Guarigione	Recidiva
Gruppo 1	3	10	5
Gruppo 2	10	13	7
Gruppo 3	2	1	1

Tabella 3.3: Tabella di corretta classificazione tra i gruppi ottenuti con il metodo delle k-medie e lo stato della malattia.

Dalla tabella 3.3 emerge che il metodo delle k-medie con tre centroidi classifica correttamente il 40% delle osservazioni, quindi meno della metà dei pazienti viene classificato nel gruppo che rappresenta il suo stato di malattia.

In figura 3.4 vengono riportate la Silhouette e il grafico dei gruppi sul piano delle prime due componenti principali per verificare la capacità di classificazione del metodo delle k-medie con tre centroidi nello stato di malattia dei pazienti sulla base dei parametri infiammatori in ingresso.

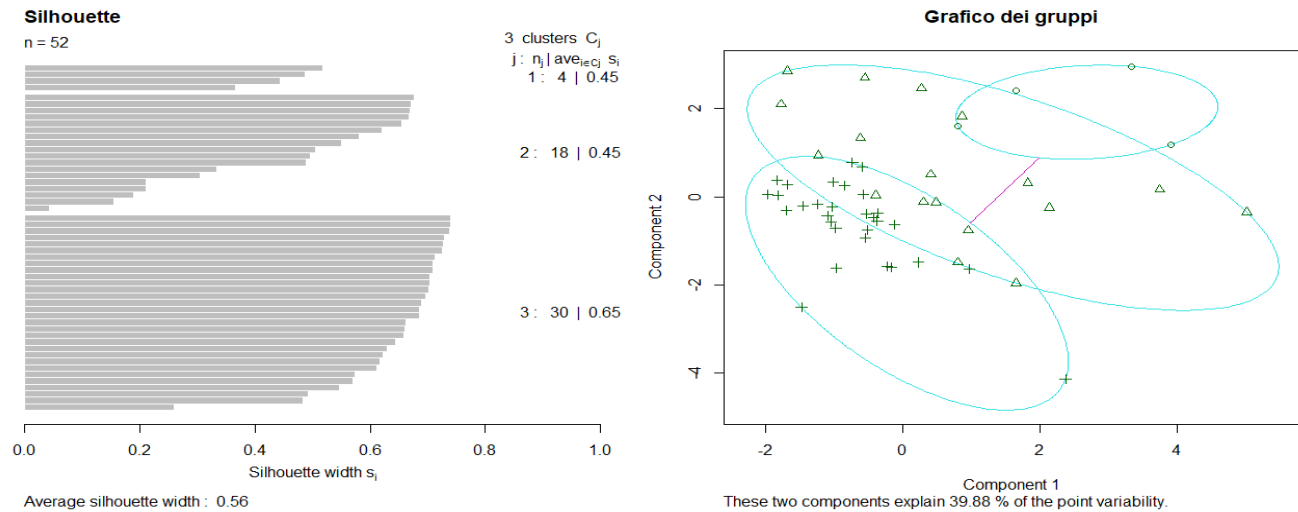


Figura 3.4: Silhouette e grafico dei gruppi sul piano delle prime due componenti principali per la valutazione della classificazione col metodo delle k-medie rispetto allo stato della malattia in ingresso.

Dalla figura 3.4 emerge che secondo la Silhouette ogni osservazione si trova meglio nel suo gruppo piuttosto che in un altro, mentre dal grafico dei gruppi sul piano delle prime due componenti principali emerge che le osservazioni non si distinguono nettamente in tre gruppi ma tendono a confondersi.

Dunque, secondo la tabella di corretta classificazione e il grafico dei gruppi, il metodo delle k-medie non è adeguato per distinguere i parametri infiammatori in ingresso dei pazienti e classificare correttamente i pazienti nello stato di malattia corrispondente. Si noti che i risultati ottenuti con le tecniche di analisi multivariate sono coerenti con le analisi bivariate.

Una seconda applicazione del metodo delle k-medie viene svolta per verificare che i parametri sanguigni si possano differenziare sulla base della guarigione contro la recidiva e/o il decesso del soggetto. Tuttavia, anche in questo caso, risulta che solo il 50% delle unità viene correttamente classificata; dunque, i parametri sanguigni non sembrano differenziarsi sulla base della guarigione e del decesso/recidiva.

Viene svolta un'analisi dei gruppi sui dati con il metodo delle k-medie per verificare che i parametri sanguigni si possano distinguere sulla base della presenza della terapia steroidea in atto. In figura 3.5 vengono riportati i grafici di dispersione per ogni incrocio tra i parametri infiammatori, divisi per la presenza della terapia steroidea in atto.



Figura 3.5: Grafici di dispersione per gli incroci tra i parametri infiammatori rispetto all'esito della terapia steroidea.

Dalla figura 3.5 i parametri infiammatori non sembrano dividersi nei soggetti sulla base della terapia steroidea. Tramite il metodo delle k-medie, vengono identificati due gruppi sulla base delle somiglianze tra i parametri infiammatori. In tabella viene riportata la tabella di corretta classificazione, dove sulle colonne è posta la presenza della terapia steroidea e sulle righe i gruppi calcolati empiricamente sulla base della somiglianza dei parametri infiammatori in ingresso.

Terapia	No	Sì
Gruppo 1	14	2
Gruppo 2	31	5

Tabella 3.4: Tabella di corretta classificazione tra i gruppi ottenuti con il metodo delle k-medie e la presenza della terapia steroidea in atto.

Dalla tabella 3.4 emerge che il metodo delle k-medie con due centroidi classifica correttamente il 63% delle osservazioni.

In figura 3.6 vengono riportati la Silhouette e il grafico dei gruppi sul piano delle prime due componenti principali per valutare se il metodo delle k-medie con due centroidi classifica le unità nella corrispondente presenza di terapia steroidea in atto sulla base dei parametri sanguigni in ingresso.

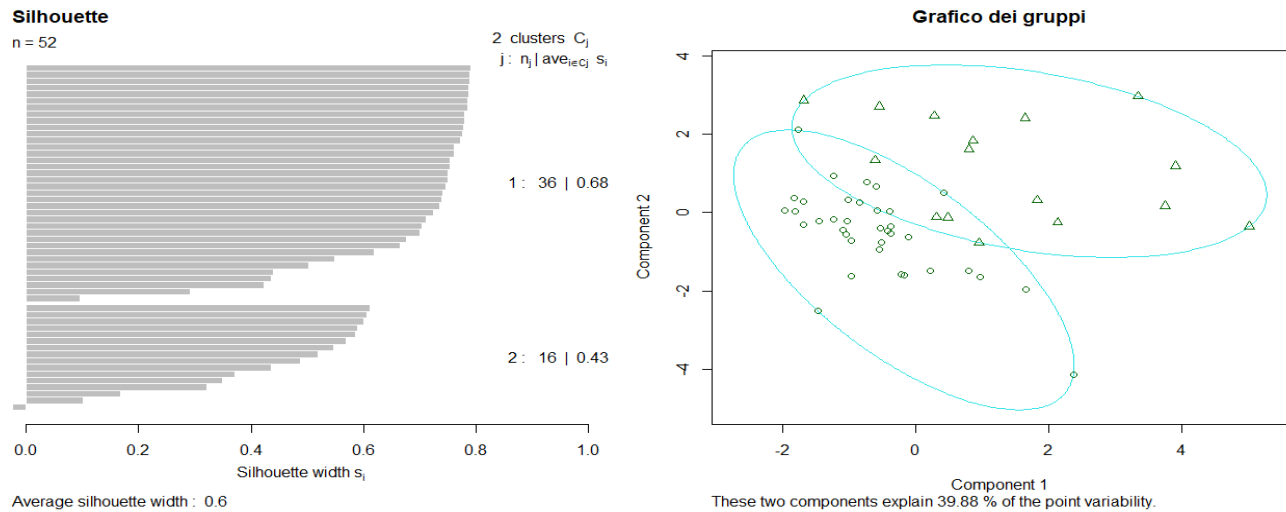


Figura 3.6: Silhouette e grafico dei gruppi sul piano delle prime due componenti principali per la valutazione della classificazione col metodo delle k-medie rispetto alla terapia steroidea in ingresso.

Dalla figura 3.6 emerge che un'unità secondo la Silhouette è stata inserita nel gruppo sbagliato, perché riporta un valore negativo (ultima osservazione), mentre, dal grafico dei gruppi sul piano delle prime due componenti principali, emerge che i due gruppi si sovrappongono e non vi è una distinzione netta delle unità. La tabella di corretta classificazione e il grafico dei gruppi portano a concludere che il metodo delle k-medie non è adeguato per classificare le unità statistiche sulla base dei parametri sanguigni nella rispettiva presenza della terapia steroidea in atto.

3.2 In fase di recidiva

In questa sezione vengono riportati risultati delle analisi multivariate per i soggetti in fase di recidiva.

3.2.1 Analisi delle componenti principali

Viene svolta un'analisi delle componenti principali per i parametri infiammatori dei pazienti in fase di recidiva per valutare quante componenti principali sono necessarie per descrivere circa il 75% della variabilità totale dei parametri infiammatori in fase di recidiva. In rosso il 75% della variabilità cumulata richiesta.

In figura 3.7 è riportato lo Screeplot della varianza spiegata dalle componenti principali dei parametri infiammatori dei pazienti in fase di recidiva.

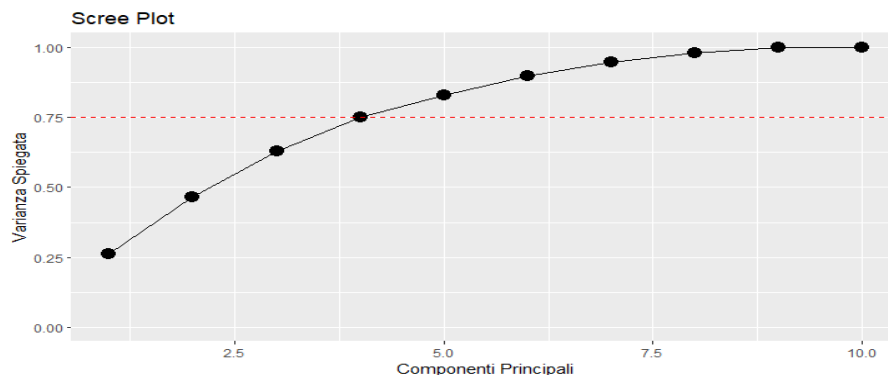


Figura 3.7: Screeplot della varianza spiegata dalle componenti principali dei parametri infiammatori dei pazienti in fase di recidiva.

Dalla figura 3.7 non emergono gomiti particolari e si nota che circa il 75% della varianza spiegata cumulata viene spiegata dalle prime quattro componenti principali.

In tabella vengono riportati i risultati dell'analisi delle componenti principali per i parametri infiammatori in fase di recidiva dei pazienti.

Componenti Principali	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Deviazione Standard	1.62	1.42	1.28	1.10	0.88	0.83	0.71	0.57	0.43	0.15
Varianza Cumulata	0.26	0.46	0.63	0.75	0.83	0.9	0.95	0.98	0.99	1

Tabella 3.5: Analisi delle componenti principali per i parametri sanguigni dei soggetti in fase di recidiva.

Dalla tabella 3.5 emerge che le prime quattro componenti principali descrivono il 75% della variabilità totale, perché i parametri infiammatori dei pazienti in fase di recidiva risultano poco correlati tra loro, e la quarta componente principale ha deviazione standard pari a 1.1, quindi al di sopra della media. Dunque, 4 componenti principali risultano essere sufficienti.

In tabella 3.6 vengono riportati i pesi delle prime due componenti principali per i pazienti in fase di recidiva.

Variabile	PC1	PC2
Leucociti	0.48051139	-0.33970527
Neutrofili	0.48667161	-0.33173364
PON1	0.22347254	-0.01722014
CRP	-0.05707822	-0.55985403
Aptoglobina	0.40172546	0.22143574
Ferritina	-0.22153643	-0.23570486
Albumine	0.45808078	0.15068249
IgG	-0.18451075	-0.53219812
IgM	-0.09979127	-0.11551464
IgA	0.12148829	-0.19221290

Tabella 3.6: Pesi delle componenti principali per i pazienti in fase di recidiva.

Dalla tabella emerge che la prima componente principale è maggiormente descritta dal livello di *Leucociti* (peso pari a 0.48), *Neutrofili* (peso pari a 0.487) e *Albumine* (peso pari a 0.458), che entrano con segno positivo; mentre, per la seconda componente principale, i parametri più influenti risultano essere il livello di

IgG (peso pari a -0.532) e *CRP* (peso pari a -0.56), che entrano con segno negativo.

In figura 3.8 è riportato il grafico delle correlazioni sul piano delle prime due componenti principali.

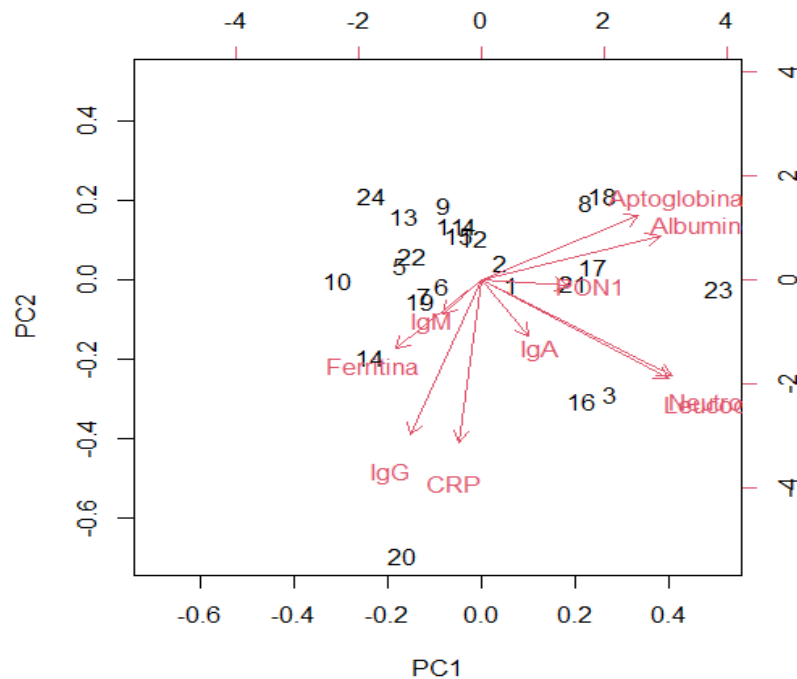


Figura 3.8: Grafico delle correlazioni per le prime due componenti principali.

Dalla figura emerge che i parametri più influenti risultano essere quelli citati sopra e si nota che le variabili risultano essere poco correlate tra loro.

Poiché sono necessarie 4 componenti principali per descrivere una sufficiente proporzione di variabilità totale dei parametri infiammatori dei pazienti in fase di recidiva, l'analisi delle componenti principali non risulta adeguata per ridurre la dimensionalità del campione.

3.2.2 Analisi dei gruppi

Si vuole studiare se i parametri infiammatori dei soggetti in fase di recidiva possano caratterizzare lo stato della malattia dei soggetti in fase di recidiva. In figura 3.9 vengono riportati i grafici di dispersione per ogni incrocio tra i parametri infiammatori in fase di recidiva, divisi per lo stato della malattia.



Figura 3.9: Grafici di dispersione per gli incroci tra i parametri infiammatori in fase di recidiva rispetto allo stato della malattia.

Dalla figura 3.9 non si notano distinzioni nette tra i parametri infiammatori dei pazienti guariti e i pazienti che hanno sviluppato la recidiva in fase di recidiva.

Tramite il metodo delle k-medie, i soggetti vengono classificati in due gruppi distinti sulla base della somiglianza dei parametri infiammatori in fase di recidiva. In tabella 3.7 viene riportata la tabella di corretta classificazione, dove viene posto sulle colonne lo stato della malattia osservata per i pazienti e sulle righe i gruppi trovati empiricamente con il metodo delle k-medie.

Stato	Guarigione	Recidiva
Gruppo 1	11	5
Gruppo 2	2	6

Tabella 3.7: Tabella di corretta classificazione tra i gruppi ottenuti con il metodo delle k-medie e lo stato della malattia in fase di recidiva.

Dalla tabella 3.7 emerge che il metodo delle k-medie con due centroidi classifica correttamente il 70.8% delle osservazioni.

In figura 3.10 vengono riportati la Silhouette e il grafico dei gruppi sul piano delle prime due componenti principali per la valutazione del processo di classificazione del metodo delle k-medie in due gruppi.

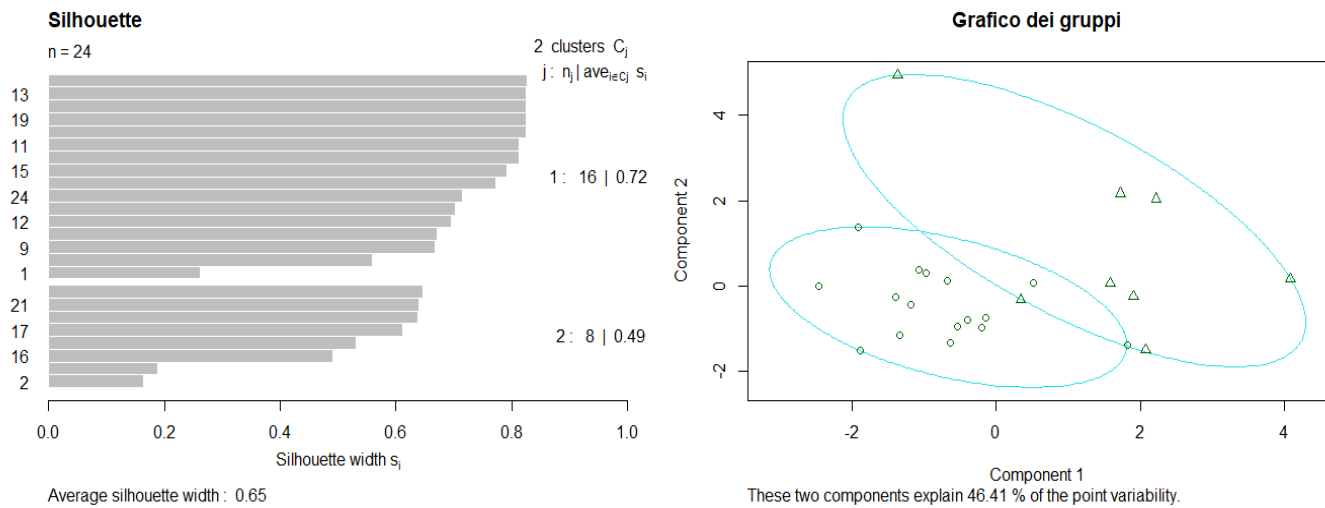


Figura 3.10: Silhouette e grafico dei gruppi sul piano delle prime due componenti principali per la valutazione della classificazione col metodo delle k-medie.

dalla figura 3.10 emerge che la Silhouette riporta solo valori positivi per i due gruppi; quindi, ogni osservazione si trova meglio nel gruppo a cui è stata assegnata piuttosto che in un altro gruppo. Il grafico dei gruppi sul piano delle prime due componenti principali riporta che i gruppi sembrano distinguersi, ma c'è una parte dei due gruppi, rappresentati dai cerchi in azzurro, che si sovrappongono.

La tabella di corretta classificazione e il grafico dei gruppi portano a concludere che il metodo delle k-medie non è adeguato per classificare le unità statistiche sulla base dei parametri sanguigni nel rispettivo stato di malattia.

Capitolo 4

Modello di regressione logistico

In questo capitolo vengono presentati i modelli di regressione utili per rispondere agli obiettivi a cui precedentemente è stata data una risposta parziale. Vista la natura dei dati e gli obiettivi del report, verranno stimati dei modelli di regressione logistica con funzione di legame *logit*¹ (si vedano Ventura *et al.*, 2017 [5], e Salvan *et al.*, 2020 [9]). La variabile risposta che rappresenta lo stato della malattia verrà dicotomizzata nelle varie sezioni a seconda dell'analisi. Durante la procedura di selezione delle variabili è stato scelto un livello di significatività del 10% per essere più protettivi nell'eliminazione delle variabili.

4.1 In ingresso

Per rispondere alle quattro domande di ricerca per i soggetti in fase di ingresso, vengono applicati ai dati dei modelli di regressione logistici. La forma del modello, inserendo tutte le covariate, presenta la seguente forma:

$$\begin{aligned} \text{logit}(\mu) = & \beta_0 + \beta_1 * \text{Terapia} + \beta_2 * \text{Leucociti} + \beta_3 * \text{Neutrofili} + \beta_4 * \text{PON1} + \beta_5 * \text{CRP} + \\ & + \beta_6 * \text{Aptoglobina} + \beta_7 * \text{Ferritina} + \beta_8 * \text{Albumine} + \beta_9 * \text{IgG} + \beta_{10} * \text{IgM} + \beta_{11} * \text{IgA} + \beta_{12} * \text{Fibrinogeno}, \end{aligned}$$

dove Y è una variabile dicotomica che rappresenta la guarigione dei pazienti, μ è la media di Y e rappresenta la probabilità di guarigione e $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ è la funzione di legame.

Nelle sezioni dedicate, verrà studiato quali covariate risultano significative per discriminare lo stato di malattia dei soggetti.

4.1.1 Guarigione contro Recidiva

In questa sezione, vengono selezionati i pazienti che sono guariti o che hanno sviluppato la recidiva per infiammazione al sistema nervoso centrale. La variabile *Stato* viene dicotomizzata nel seguente modo:

1. Vale 1 se il soggetto è guarito.
2. Vale 0 se il soggetto ha sviluppato la recidiva.

Le unità statistiche del sotto dataset risultano essere 37. Vista la natura dei dati, viene stimato inizialmente un modello di regressione logistico con funzione di legame *logit* con la sola intercetta.

A seguito di una selezione in avanti delle variabili, basata sul test log-rapporto di verosimiglianza con un livello di significatività del 10%, tutte le variabili risultano non significative: i parametri infiammatori e la presenza di terapia steroidea in atto non sono significativi per descrivere lo stato della guarigione o della recidiva nei pazienti. Il modello nullo stimato risulta dunque essere:

$$\text{logit}(\hat{\mu}) = \hat{\beta}_0$$

¹Nelle analisi è stata applicata anche una funzione di legame *probit* per verificare se i modelli si adattassero più adeguatamente ai dati, ma è risultata preferibile la funzione *logit*, anche per permettere un'interpretazione diretta dei parametri.

In tabella 4.1 viene riportato il coefficiente stimato per il modello nullo

Variabile	$\hat{\beta}$	Std. Error	z value	p-value
Intercetta	0.6131	0.3444	1.78	0.075

Tabella 4.1: Coefficienti stimati del modello logit nullo stimato.

Dalla tabella 4.1 emerge che i valori stimati $\text{logit}(\hat{\mu})$ dal modello saranno sempre pari a 0.6131, ovvero l'intercetta. Pertanto, lo stato della malattia risulta indipendente dalle covariate: le covariate non risultano utili per descrivere la guarigione oppure lo sviluppo della recidiva dei pazienti in ingresso.

4.1.2 Guarigione contro Decesso

Vengono selezionati i soggetti che sono guariti dall'infiammazione al sistema nervoso centrale oppure che sono deceduti. La variabile che rappresenta lo *Stato* di malattia del paziente viene dicotomizzata nel seguente modo:

1. Vale 1 se il soggetto è guarito.
2. Vale 0 se il soggetto è deceduto.

La numerosità del sotto dataset risulta essere pari a 39. Viene stimato un modello di regressione logistico con funzione di legame *logit* che comprende tutte le variabili iniziali.

Tramite una selezione in avanti delle variabili, basate sul test log-rapporto di verosimiglianza con un livello di significatività del 10%, solo il livello di *Ferritina* nel sangue dei pazienti risulta significativa per spiegare lo stato dei pazienti selezionati. Il modello finale risulta, dunque, essere:

$$\text{logit}(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 * \text{Ferritina}.$$

In tabella 4.2 vengono riportate le stime dei coefficienti del modello selezionato.

Variabile	$\hat{\beta}_i$	Std. Error	z value	p-value
Intercetta	1.711	0.782	2.186	0.0288
Ferritina	-0.0057	0.00328	-1.727	0.0842

Tabella 4.2: Coefficienti stimati del modello logit per i pazienti guariti e i pazienti che hanno sviluppato la recidiva.

Dalla tabella 4.2 emerge che, per ogni aumento unitario del livello di *Ferritina*, l'*odds* (quota) di pazienti guariti è 0.57% ($1 - \exp(\beta_{\text{Ferritina}})$) volte inferiore rispetto all'*odds* dei pazienti guariti con nessun aumento del livello di *Ferritina*. Alti livelli di *Ferritina* risultano essere un fattore di rischio per lo sviluppo della recidiva nei pazienti.

Viene riportato in tabella 4.3 la tabella di corretta classificazione per i valori stimati dal modello selezionato e lo stato della malattia dei soggetti.

	Decesso	Guarigione
0	4	1
1	11	23

Tabella 4.3: Tabella di corretta classificazione per i valori stimati dal modello e lo stato di malattia dei pazienti.

Dalla tabella 4.3 emerge che il ridotto stimato classifica correttamente il 69.2% dei valori.

In figura 4.1 è riportata la curva ROC per la valutazione della bontà di adattamento del modello stimato ridotto ai dati.

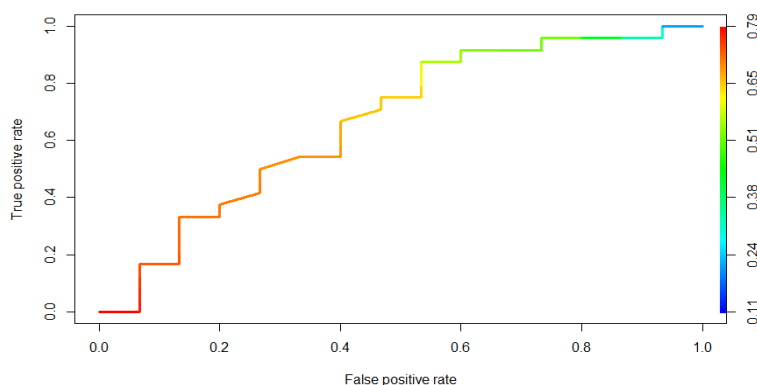


Figura 4.1: Curva ROC per la valutazione della bontà di adattamento del modello ridotto ai dati.

L'area sottesa alla curva ROC riportata in figura 4.1 risulta essere pari a 66.5% (p-value pari a 0.044); dunque, vi è un mediocre adattamento del modello ai dati. Infine, il test di Hoslem-Lemeshow per verificare l'adattamento tra frequenze attese e stimate dal modello sia soddisfacente, riporta un valore osservato della statistica test pari a 4.92 e un p-value pari a 0.77: i dati supportano l'ipotesi che le frequenze attese stimate dal modello e le frequenze osservate si adattino sufficientemente bene a tutti i livelli usuali di α .

4.1.3 Guarigione contro Recidiva e/o Decesso

Viene stimato per tutti i soggetti un modello logistico con funzione legame *logit* per verificare quali fattori influenzino la guarigione oppure la recidiva e/o il decesso dei soggetti per infiammazione al sistema nervoso centrale. La variabile *Stato* viene dicotomizzata nel seguente modo:

1. Vale 1 se il soggetto è guarito.
2. Vale 0 se il soggetto è deceduto oppure ha sviluppato la recidiva.

A seguito di una selezione in avanti delle variabili, tramite il test log-rapporto di verosimiglianza con un livello di significatività del 10%, il livello nel sangue di *Ferritina* risulta essere l'unica covariata significativa per descrivere lo stato di salute del paziente. Il modello stimato risulta essere:

$$\text{logit}(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 * \text{Ferritina}.$$

In tabella 4.4 viene riportato l'adattamento del modello.

Variabile	$\hat{\beta}_i$	Std. Error	z value	p-value
Intercetta	0.937414	0.696400	1.346	0.178
Ferritina	-0.005138	0.003129	-1.642	0.101

Tabella 4.4: Modello logit stimato per i soggetti che guariscono contro i soggetti che muoiono e/o sviluppano la recidiva.

Dalla tabella 4.4 emerge che, per ogni aumento unitario del livello di *Ferritina*, l'*odds* (quota) di pazienti guariti è 0.51% ($1 - \exp(\beta_{\text{Ferritina}})$) volte inferiore rispetto all'*odds* dei pazienti guariti con nessun aumento del livello di *Ferritina*. Alti livelli di *Ferritina* risultano essere un fattore di rischio per lo sviluppo della recidiva nei pazienti. Alti livelli di *Ferritina* risultano essere un fattore di rischio per lo sviluppo della recidiva oppure per il decesso dei pazienti.

In tabella 4.5 viene riportata la tabella di corretta classificazione per valutare la bontà di adattamento del modello ai dati.

	Decesso e/o Recidiva	Guarigione
0	17	11
1	9	15

Tabella 4.5: Tabella di corretta classificazione per verificare la bontà di adattamento del modello ridotto ai dati.

Dalla tabella 4.5 emerge che il modello classifica correttamente il 61.5% delle osservazioni. Dunque, vi è un mediocre adattamento del modello ai dati. In figura viene riportata la curva ROC per valutare la bontà di adattamento del modello ai dati.

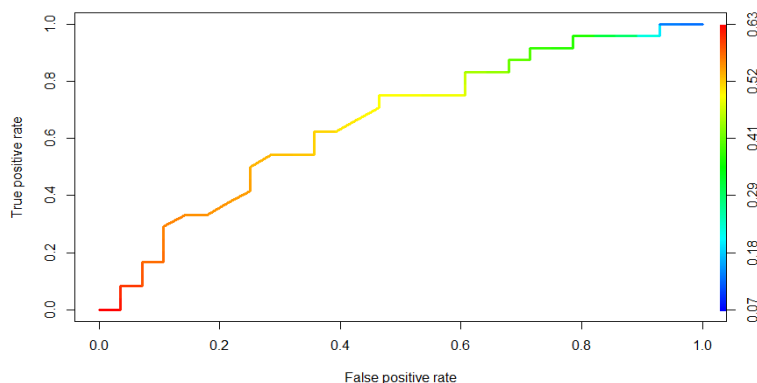


Figura 4.2: Curva ROC per la valutazione della bontà di adattamento del modello ridotto ai dati.

Dalla figura 4.5 emerge un cattivo adattamento del modello ai dati; infatti, l'area sottesa alla curva ROC risulta essere pari a 65.8% (p-value pari a 0.026). Infine, il test di Hosmer-Lemeshow per verificare la concordanza tra valori attesi stimati e valori osservati riporta un valore osservato della statistica test pari a 3.66 e un p-value pari a 0.886: i dati supportano l'ipotesi che le frequenze attese stimate dal modello ridotto e le frequenze osservate siano coerenti a tutti i livelli usuali di α .

4.2 In fase di recidiva

Per valutare l'influenza dei parametri sanguigni tra i soggetti guariti e che hanno sviluppato la recidiva nel momento della recidiva, viene stimato un modello di regressione logistico con funzione di legame *logit*, che consideri inizialmente tutte le variabili. Il modello presenta la seguente forma:

$$\begin{aligned} \text{logit}(\mu) = & \beta_0 + \beta_1 * \text{Leucociti} + \beta_2 * \text{Neutrofili} + \beta_3 * \text{PON1} + \beta_4 * \text{CRP} + \beta_5 * \text{Aptoglobina} + \beta_6 * \text{Ferritina} + \\ & + \beta_7 * \text{Albumine} + \beta_8 * \text{IgG} + \beta_9 * \text{IgM} + \beta_{10} * \text{IgA}, \end{aligned}$$

dove Y è una variabile dicotomica che rappresenta la guarigione dei pazienti, μ è la media di Y e rappresenta la probabilità di guarigione e $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ è la funzione di legame.

La variabile che rappresenta lo stato della malattia dei pazienti in fase di recidiva è stata dicotomizzata nel seguente modo:

1. Vale 1 se il paziente è guarito.
2. Vale 0 se il paziente ha sviluppato la recidiva.

Tramite un procedura di selezione delle variabili in avanti, basate sul test log-rapporto di verosimiglianza, con un livello di significatività del 10%, risultano essere significative per lo stato della malattia dei pazienti il livello di *IgM* e il livello di *Neutrofili* nel sangue dei pazienti in fase di recidiva. Il modello stimato assume la forma:

$$\text{logit}(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 * \text{Neutrofili} + \hat{\beta}_2 * \text{IgM}.$$

In tabella 4.6 viene riportata la tabella di adattamento del modello.

Variabile	$\hat{\beta}_i$	Std. Error	z value	p-value
Intercetta	8.45	3.89	2.174	0.0297
IgM	-0.055401	0.029	-1.924	0.0544
Neutrofili	-0.0003	0.0002	-1.661	0.0967

Tabella 4.6: Modello logistico stimato per lo stato dei pazienti in fase di recidiva.

Dalla tabella 4.6 emerge che, per ogni aumento unitario del livello di *IgM*, l'*odds* (quota) di pazienti guariti è 5.3% ($1 - \exp(\beta_{IgM})$) volte inferiore rispetto all'*odds* dei pazienti guariti con nessun aumento del livello di *IgM*, a parità del livello di *Neutrofili*; mentre, per ogni aumento unitario del livello di *Neutrofili*, l'*odds* (quota) di pazienti guariti è 0.03% ($1 - \exp(\beta_{Neutrofili})$) volte inferiore rispetto all'*odds* dei pazienti guariti con nessun aumento del livello di *Neutrofili*, a parità del livello di *IgM*. Dunque, alti livelli di *Neutrofili* e *IgM* risultano essere un fattore di rischio per lo sviluppo della recidiva dei pazienti.

In tabella 4.7 viene riportata la tabella di corretta classificazione per valutare la bontà di adattamento del modello ai dati.

	Recidiva	Guarigione
0	8	3
1	3	10

Tabella 4.7: Tabella di corretta classificazione del modello logistico ridotto per lo stato della malattia dei pazienti in fase di recidiva.

Dalla tabella emerge che il modello classifica correttamente il 75% delle osservazioni; dunque, vi è un buon adattamento del modello ai dati. In figura 4.3 viene riportata la curva ROC per la valutazione della bontà di adattamento del modello ai dati.

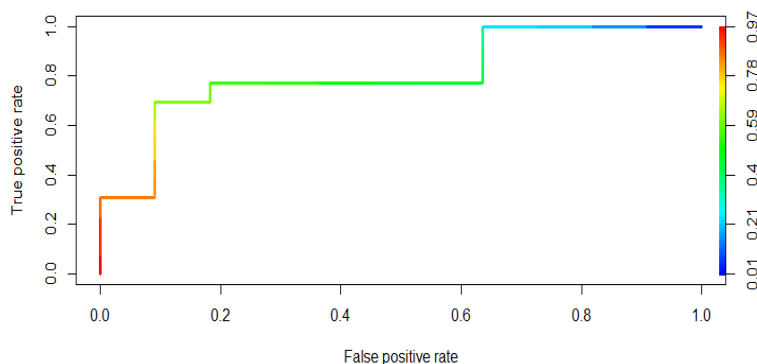


Figura 4.3: Curva ROC per la valutazione della bontà di adattamento del modello ridotto ai dati.

Dalla figura emerge un discreto adattamento del modello ai dati; infatti la curva ROC riporta un AUC (area sottesa alla curva ROC) pari a 0.81 (p-value pari a 0.012). Infine, il test Hosmer-Lemeshow riporta un valore osservato della statistica test pari a 13.01 e un p-value pari a 0.11: i dati non sono contrari all'ipotesi di un adattamento soddisfacente tra le frequenze attese stimate dal modello ridotto e le frequenze osservate.

Capitolo 5

Metodi asintotici di ordine elevato

In questo capitolo vengono utilizzati dei metodi asintotici di ordine elevato (*HOA*, *Higer-Order Asyntotics*) per stimare un modello di regressione logistico quando la numerosità campionaria è bassa (si veda Brazzale *et al.*, 2007 [2] e Brazzale, 1998 [1])¹.

Sia $r_p = \text{sgn}(\hat{\psi} - \psi)[2(l(\hat{\psi}, \hat{\lambda}) - l(\psi, \hat{\lambda}_{\psi}))]^{1/2}$ la statistica *radice con segno* profilo per un parametro scalare di interesse ψ con un errore relativo pari a $O_p(n^{-1/2})$. Si definisce la statistica *radice con segno* parziale modificata la seguente quantità:

$$r_p^* = r_p + \frac{1}{r_p} \log\left(\rho \frac{W_p}{r_p}\right),$$

dove $W_p = (\hat{\psi} - \psi_0)j_p^{1/2}(\hat{\psi})$ è la statistica alla Wald parziale e ρ un fattore di correzione. La statistica r_p^* ha una distribuzione nulla $N(0, 1)$ al terzo ordine e consente di ottenere degli intervalli di confidenza e p-value per il parametro di interesse più precisi in presenza di piccoli campioni.

Viene fissato un livello di significatività di α del 10%.

5.1 In ingresso

5.1.1 Guarigione contro Recidiva

Nella Sezione 4.1.1 è stato stimato un modello per determinare quanto i parametri infiammatori dei pazienti in ingresso e la terapia steroidea in atto influissero sulla guarigione oppure sullo sviluppo della recidiva dei pazienti; il modello in questione, ha portato a concludere che lo stato della malattia dei soggetti non dipendesse dalle covariate.

In questa sezione, viene stimato un modello logistico a partire dal modello nullo con la sola intercetta e, in seguito ad una selezione in avanti delle variabili, basata sulla statistica r_p^* (statistica *radice con segno* modificata), nessun parametro risulta significativo. In tabella 5.1 vengono riportate le stime di r_p^* con i relativi intervalli di confidenza di livello 0.9 dei parametri infiammatori in ingresso per la guarigione o lo sviluppo della recidiva dei pazienti.

Variabile	Stima	Estremo Sup	Estremo Inf
Leucociti	-0.05609	-0.2463	0.1326
Neutrofili	-0.0001035	-0.0003722	0.0001630
PON1	0.2253	-0.4735	0.9447
CRP	0.3298	-0.2759	1.2220
Aptoglobina	-0.0003988	-0.003331	0.002622
Ferritina	-0.003747	-0.010120	0.002100
Albumine	-0.4703	-2.206	1.1410
IgG	-0.001115	-0.005413	0.003228

¹Per condurre le analisi, verrà utilizzata la libreria R "hoa".

Variabile	Stima	Estremo Sup	Estremo Inf
IgM	-0.01934	-0.04633	0.005481
IgA	0.04185	-0.014190	0.1579
Fibrinogeno	0.001563	-0.002298	0.005697

Tabella 5.1: Stima degli r_p^* e dei relativi intervalli di confidenza di livello 0.9 per i parametri infiammatori dei pazienti in ingresso.

Dalla tabella 5.1 emerge che, a un livello di confidenza del 90%, tutti gli intervalli di confidenza contengono lo 0; dunque, l'inferenza basata su r_p^* conferma il risultato di indipendenza tra lo stato di malattia dei pazienti e le covariate (si veda la Sezione 4.1.1).

5.1.2 Guarigione contro Decesso

Nella sezione 4.1.2 è stato stimato un modello logistico per valutare se i parametri infiammatori dei pazienti in ingresso e la terapia steroidea in atto influenzino la guarigione oppure il decesso dei pazienti; il modello in questione, ha portato a concludere che solo il livello di *Ferritina* fosse significativo per descrivere lo stato della malattia dei pazienti.

In questa sezione, viene stimato un modello logistico a partire dal modello nullo con la sola intercetta e, in seguito, vengono valutati singolarmente i coefficienti di regressione tramite la statistica r_p^* . A seguito di una procedura manuale di selezione in avanti delle variabili, basata sulla statistica r_p^* , il livello di *Ferritina* risulta l'unico parametro sanguigno significativo, con un livello di significatività del 10%, per descrivere la guarigione oppure il decesso dei pazienti.

In figura 5.1 viene riportato un grafico con gli intervalli di confidenza di livello 0.9 per il livello di *Ferritina*, basati sulla statistica alla Wald profilo (W_p), la statistica *radice con segno* (r_p) e la statistica *radice con segno modificata* (r_p^*).

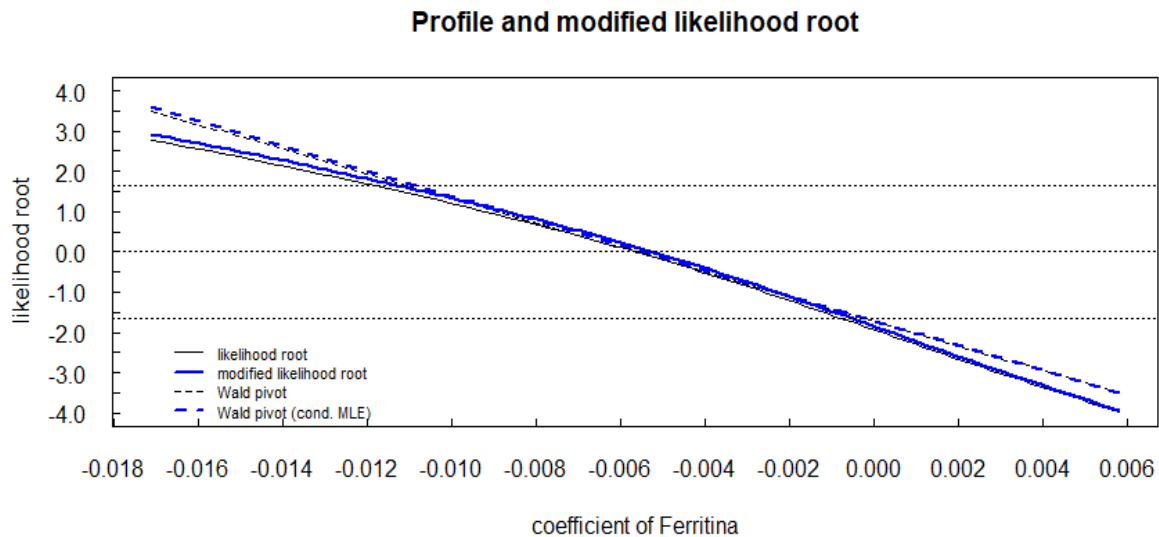


Figura 5.1: Confronto tra gli intervalli di confidenza di livello 0.9 per il livello di *Ferritina* ottenuti con le statistiche W_p , r_p e r_p^* .

Dalla figura 5.1 emerge che l'intervallo di confidenza basato sulla quantità r_p (linea nera continua) riporta un intervallo di confidenza leggermente più ristretto dell'intervallo di confidenza basato sulla quantità r_p^* (linea blu continua), soprattutto per l'estremo inferiore.

In particolare, vengono riportati in tabella 5.2 le stime dei coefficienti angolari del livello di *Ferritina* e gli intervalli di confidenza di livello 0.9 ottenuti tramite W_p , r_p e r_p^* .

Statistica	Stima	Estremo Inf	Estremo Sup
W_p	-0.005657	-0.01104	-0.0002692
r_p	-0.005657	-0.01183	-0.0007926
r_p^*	-0.005508	-0.01127	-0.0005427

Tabella 5.2: Confronto tra le stime e l'intervallo di confidenza per il livello di *Ferritina* ottenuto con W_p , r_p e r_p^* .

Dalla tabella 5.2 si nota che l'intervallo di confidenza per il coefficiente di regressione del livello di *Ferritina* ottenuto con r_p^* è più preciso di quello ottenuto con r_p .

In conclusione, sulla base della selezione del modello e delle stime ottenute con r_p^* , il modello logistico conferma i risultati ottenuti nella Sezione 4.1.2.

5.1.3 Guarigione contro Decesso e/o Recidiva

Nella sezione 4.1.3 viene stimato un modello logistico per valutare se i parametri infiammatori dei pazienti in ingresso e la terapia steroidea in atto influenzino la guarigione oppure il decesso e/o lo sviluppo della recidiva; il modello in questione ha portato a concludere che solo il livello di *Ferritina* fosse significativo per descrivere lo stato della malattia dei pazienti.

In questa sezione, viene stimato un modello logistico a partire dal modello nullo con la sola intercetta e, in seguito, vengono valutati singolarmente i coefficienti di regressione tramite la statistica r_p^* . A seguito di una procedura manuale di selezione in avanti delle variabile, basata sulla statistica r_p^* , il livello di *Ferritina* risulta l'unico parametro sanguigno significativo, con un livello di significatività del 10%, per descrivere la guarigione oppure il decesso dei pazienti.

In figura 5.2 viene riportato un grafico con gli intervalli di confidenza di livello 0.9 per il livello di *Ferritina*, basati su W_p , r_p e r_p^* .

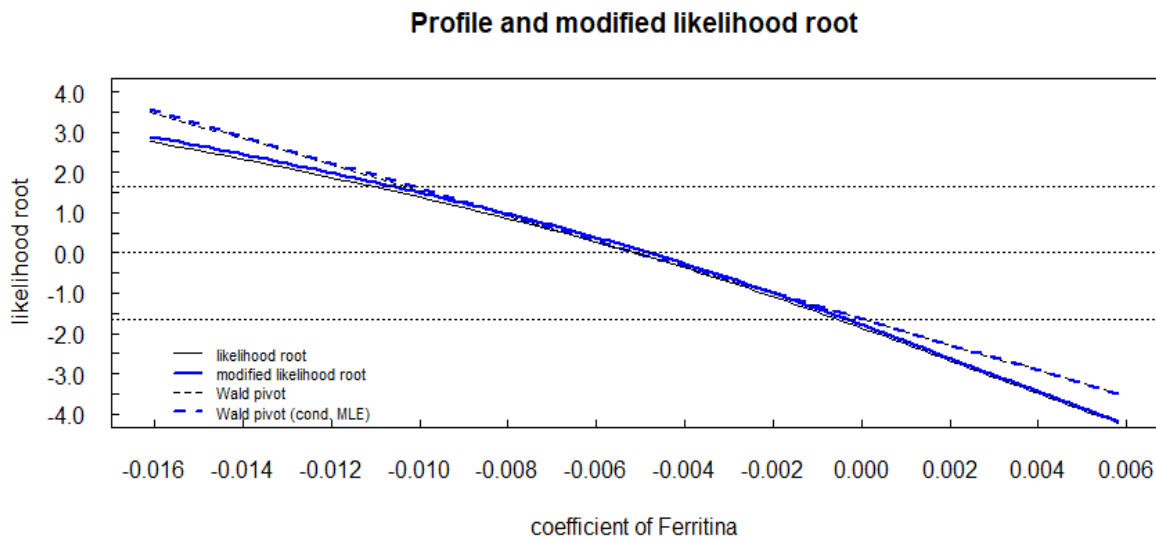


Figura 5.2: Confronto tra gli intervalli di confidenza di livello 0.9 per il livello di *Ferritina* ottenuti con W_p , r_p e r_p^*

Dalla figura 5.2 emerge che l'intervallo di confidenza basato sulla quantità r_p (linea nera continua) risulta leggermente più ristretto dell'intervallo di confidenza basato sulla quantità r_p^* (linea blu continua), soprattutto per l'estremo inferiore.

In particolare, vengono riportati in tabella 5.3 le stime dei coefficienti angolari del livello di *Ferritina* e gli intervalli di confidenza di livello 0.9 ottenuti tramite W_p , r_p e r_p^* .

Statistica	Stima	Estremo Inf	Estremo Sup
W_p	-0.005138	-0.01029	-0.0000087
r_p	-0.005138	-0.01106	-0.0005553
r_p^*	-0.005049	-0.01059	-0.0003300

Tabella 5.3: Confronto tra le stime e l'intervallo di confidenza per il livello di *Ferritina* ottenuto con W_p , r_p e r_p^* .

Dalla tabella 5.3 si nota che l'intervallo di confidenza per il coefficiente di regressione del livello di *Ferritina* ottenuto con r_p^* è più preciso di quello ottenuto con r_p .

In conclusione, sulla base della selezione del modello e delle stime ottenute con r_p^* , il modello logistico conferma i risultati ottenuti con il modello logistico nella sezione 4.1.3.

5.2 In fase di Recidiva

Nella sezione 4.2 è stato stimato un modello di regressione logistico per valutare quali parametri infiammatori influenzassero la guarigione oppure lo sviluppo della recidiva dei pazienti in fase di recidiva. Il modello porta a concludere che il livello di *Neutrofili* e il livello di *IgM* sono gli unici parametri infiammatori significativi per descrivere lo stato di malattia dei pazienti in fase di recidiva.

In questa sezione, viene stimato un modello di regressione logistico, partendo dal modello nullo con sola intercetta, tramite una selezione in avanti delle variabili basata sul test r_p^* . Tramite una selezione delle variabili in avanti, vengono inserite nel modello il livello di *Neutrofili* e il livello di *IgM*.

In figura 5.3 vengono riportati gli intervalli di confidenza di livello 0.9 basati su W_p , r_p e r_p^* per il livello di *Neutrofili*.

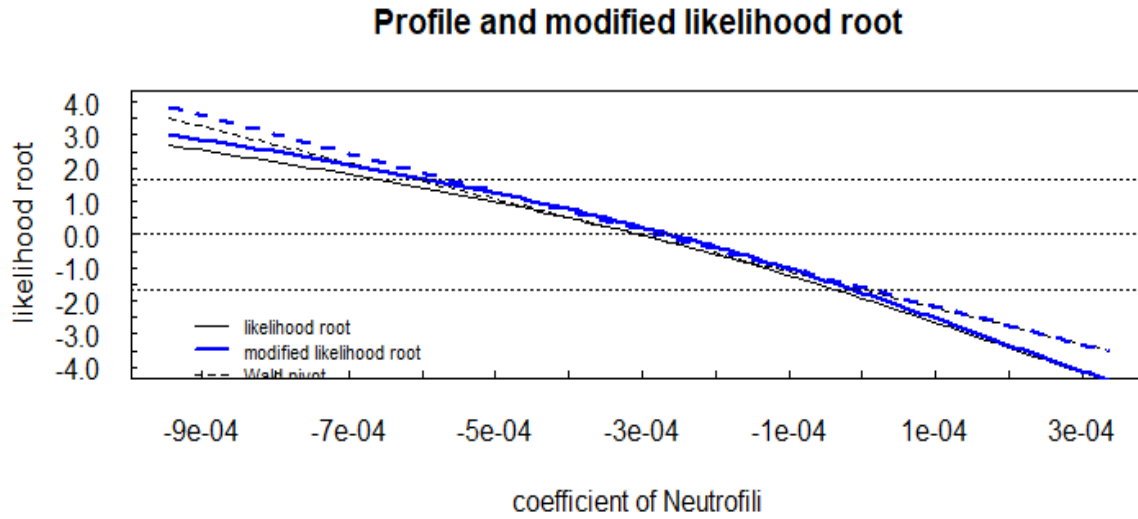


Figura 5.3: Confronto tra gli intervalli di confidenza di livello 0.9 per il livello di *Neutrofili* ottenuti con W_p , r_p e r_p^* .

Dalla figura 5.3 emerge che l'intervallo di confidenza basato su r_p^* (linea blu continua) per il livello di *Neutrofili* è più preciso di quello ottenuto con r_p (linea nera continua), soprattutto per l'estremo inferiore. In tabella 5.4 vengono riportate le stime del coefficiente di regressione del livello di *Neutrofili* con gli intervalli di confidenza di livello 0.9 basati su W_p , r_p e r_p^* .

Statistica	Stima	Estremo Inf	Estremo Sup
W_p	-0.0003040	-0.0006050	-0.00000298
r_p	-0.0003040	-0.0006559	-0.0003936
r_p^*	-0.0002766	-0.0005861	-0.0001453

Tabella 5.4: Confronto tra le stime e l'intervallo di confidenza per il livello di *Neutrofili* ottenuto con W_p , r_p e r_p^* .

Dalla tabella 5.4 emerge che l'intervallo di confidenza basato su r_p^* per il livello di *Neutrofili* è più preciso di quello basato su r_p .

In figura 5.4 viene riportato il grafico degli intervalli di confidenza di livello 0.9 basati su W_p , r_p e r_p^* per il livello di *IgM*.

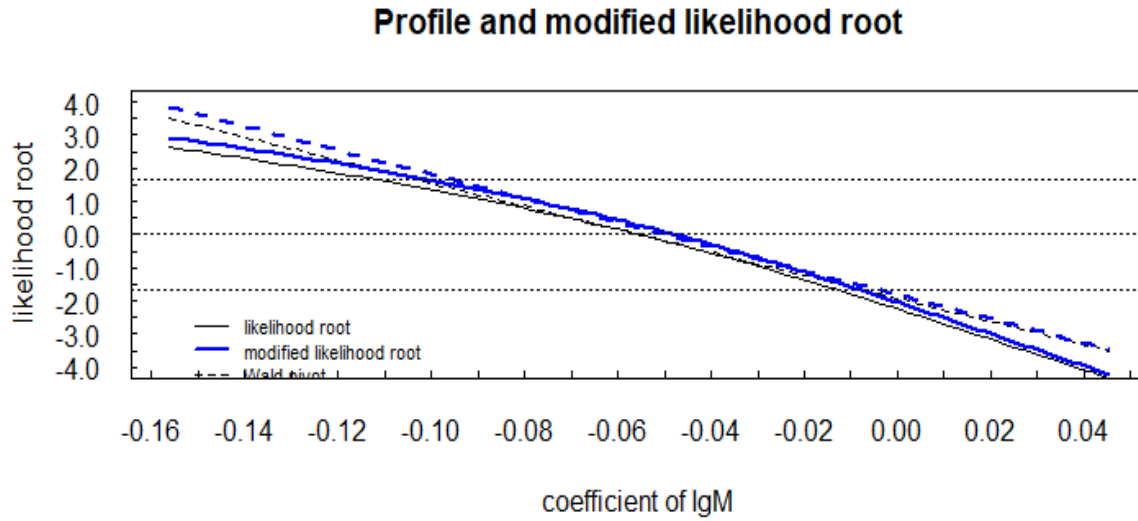


Figura 5.4: Confronto tra gli intervalli di confidenza di livello 0.9 per il livello di IgM ottenuti con W_p , r_p e r_p^* .

Dalla figura 5.4 emerge che l'intervallo di confidenza basato su r_p^* è più preciso dell'intervallo di confidenza basato su r_p , in particolar modo per l'estremo inferiore dell'intervallo. In tabella 5.5 sono riportate le stime di W_p , r_p e r_p^* con i relativi intervalli di confidenza per il livello di IgM nel sangue dei pazienti in fase di recidiva.

Statistica	Stima	Estremo Inf	Estremo Sup
W_p	-0.0003040	-0.10280	-0.008026
r_p	-0.0003040	-0.0006559	-0.0003936
r_p^*	-0.0002766	-0.0005861	-0.0001453

Tabella 5.5: Confronto tra le stime e l'intervallo di confidenza per il livello di IgM ottenuto con W_p , r_p e r_p^* .

Dalla tabella 5.5 emerge che l'intervallo di confidenza di r_p^* è più preciso di quello basato su r_p . Non risultano esserci altre covariate significative secondo la statistica r_p^* per descrivere lo stato della malattia dei pazienti in fase di recidiva.

In conclusione, il modello logistico conferma i risultati ottenuti con il modello logistico nella sezione 4.2.

Capitolo 6

Regressione robusta

Poiché dalle analisi esplorative condotte nel Capitolo 2 sono emersi diversi valori anomali nelle distribuzioni dei parametri sanguigni dei pazienti, sia in ingresso che in fase di recidiva, in questo capitolo vengono stimati dei modelli logistici robusti rispetto a valori anomali delle covariate. L'obiettivo di questo capitolo è confermare i modelli stimati nel Capitolo 4 tramite delle stime robuste dei parametri (si veda Heritier *et al.*, 2009 [6] e Farcomeni e Ventura, 2012 [4])¹.

Viene fissato un livello di significatività di α del 10%.

6.1 In ingresso

In questa sezione vengono stimati i modelli logistici robusti per i pazienti in ingresso alla clinica.

6.1.1 Guarigione contro Recidiva

Nella Sezione 4.1.1 la procedura di selezione delle variabili ha portato a concludere che lo *Stato* della malattia è indipendente dalle covariate.

Per i pazienti che sono guariti oppure che hanno sviluppato la recidiva, viene stimato un modello di regressione logistico robusto, partendo dal modello nullo e selezionando le variabili con una procedura di selezione in avanti, basata sulla significatività del test alla Wald. In tabella 6.1 sono riportati gli output del modello logistico robusto stimato per lo stato della malattia dei pazienti guariti o con la recidiva.

Variabile	Stima	Std. Error	z-value	p-value
Terapia	0.08701	1.27655	0.068	0.9457
Leucociti	-0.06023	0.10977	-0.549	0.583
Neutrofili	-0.0001089	0.0001633	-0.667	0.505
CRP	0.2656	0.3764	0.706	0.480
PON1	0.2130	0.3875	0.550	0.583
Aptoglobina	-0.0004494	0.0018067	-0.249	0.804
Ferritina	-0.003715	0.003641	-1.02	0.3076
Albumine	-0.4567	0.8262	-0.553	0.580
IgG	-0.001133	0.002620	-0.433	0.665
IgM	-0.01776	0.01577	-1.126	0.260
IgA	0.03563	0.06056	0.588	0.556
Fibrinogeno	0.001434	0.002427	0.591	0.555

Tabella 6.1: Modello logistico robusto stimato per lo stato della malattia nei pazienti guariti e con la recidiva.

¹Per condurre le analisi, verrà utilizzata la libreria R "robustbase"

Dalla tabella 6.1 emerge che nessuna variabile risulta significativa per descrivere lo stato di malattia dei pazienti guariti o con la recidiva. Pertanto, a tutti i livelli usuali di α , il modello logistico robusto conferma l'indipendenza tra lo stato della malattia e le covariate discussa nella Sezione 4.1.1.

6.1.2 Guarigione contro Decesso

Nella Sezione 4.1.2 il modello logistico ha portato a concludere che il livello di *Ferritina* fosse l'unico parametro sanguigno significativo per descrivere la guarigione oppure il decesso dei pazienti.

In questa sezione, per i pazienti guariti oppure deceduti, viene stimato un modello di regressione logistico robusto con la sola intercetta. A seguito di una selezione in avanti delle variabili, basata sulla significatività del test alla Wald, il livello di *Ferritina* risulta l'unico parametro sanguigno significativo per descrivere lo stato di malattia dei pazienti, ad un livello di significatività del 10%. In tabella 6.2 viene riportato l'adattamento del modello ai dati.

Variabile	Stima	Std. Error	z-value	p-value
Intercetta	1.817210	0.822150	2.21	0.0271
Ferritina	-0.006213	0.003530	-1.76	0.0784

Tabella 6.2: Modello logistico robusto stimato per la guarigione oppure per il decesso dei pazienti in ingresso.

Dalla tabella emerge che, per ogni aumento unitario del livello di *Ferritina*, la probabilità di pazienti guariti diminuisce 0.61% volte rispetto alla quota di pazienti guariti senza nessun aumento unitario del livello di *Ferritina*. In figura 6.1 sono riportate le stime del coefficiente di regressione per il livello di *Ferritina* del modello logistico e del modello logistico robusto, con i relativi intervalli di confidenza 0.90.

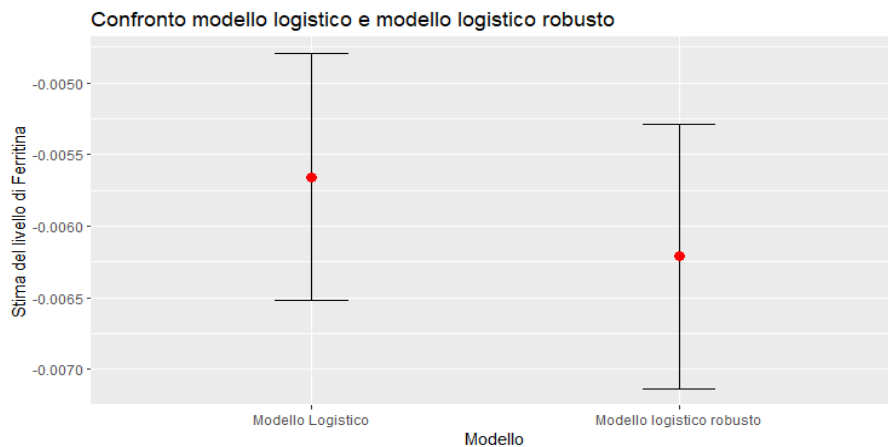


Figura 6.1: Stime dei coefficienti di regressione del modello logistico e del modello logistico robusto con intervalli di confidenza di livello 0.9.

Dalla figura emerge che il modello di regressione logistico e il modello di regressione logistico robusto riportano delle stime per il coefficiente di regressione del livello di *Ferritina* e degli intervalli di confidenza molto simili. In conclusione, il modello logistico robusto conferma i risultati ottenuti con il modello logistico nella Sezione 4.1.2.

6.1.3 Guarigione contro Recidiva e/o Decesso

Nella Sezione 4.1.3 il modello logistico ha portato a concludere che il livello di *Ferritina* fosse l'unico parametro sanguigno significativo per descrivere la guarigione e lo sviluppo della recidiva o il decesso.

In questa sezione, per lo stato della malattia dei pazienti, viene stimato un modello di regressione logistico robusto con la sola intercetta. A seguito di una selezione in avanti delle variabili, basata sulla significatività

del test alla Wald, il livello di *Ferritina* risulta l'unico parametro sanguigno significativo per descrivere lo stato di malattia dei pazienti, a un livello di significatività del 10%. In tabella 6.3 viene riportato l'adattamento del modello ai dati.

Variabile	Stima	Std. Error	z-value	p-value
Intercetta	1.058510	0.750334	1.411	0.158
Ferritina	-0.005799	0.003495	-1.659	0.097

Tabella 6.3: Modello logistico robusto stimato per la guarigione e per il decesso oppure la recidiva dei pazienti in ingresso.

Dalla tabella emerge che, per ogni aumento unitario del livello di *Ferritina*, la probabilità di pazienti guariti diminuisce 0.57% volte rispetto alla quota di pazienti guariti senza nessun aumento unitario del livello di *Ferritina*. In figura 6.2 vengono riportate le stime dei coefficienti di regressione ottenuti con il modello logistico e il modello logistico robusto per il livello di *Ferritina* nel sangue dei pazienti con il relativo intervallo di confidenza di livello 0.9.

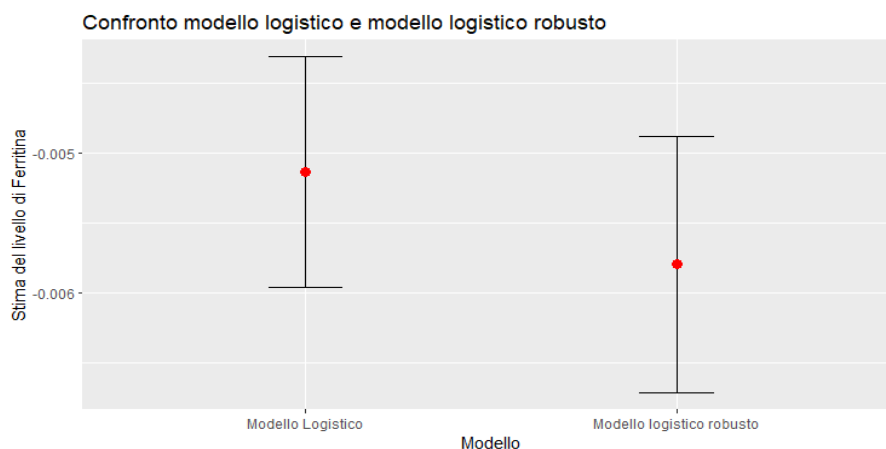


Figura 6.2: Stime dei coefficienti di regressione del modello logistico e del modello logistico robusto con intervalli di confidenza di livello 0.9.

Dalla figura emerge che tramite il modello logistico e il modello logistico robusto si ottengono delle stime e degli intervalli di confidenza del coefficiente di regressione simili per il livello di *Ferritina*.

In conclusione, il modello logistico robusto conferma i risultati ottenuti con il modello logistico nella Sezione 4.1.3

6.2 In fase di recidiva

Nella Sezione 4.2 è emerso che il livello di *IgM* e *Neutrofili* sono gli unici parametri sanguigni significativi per descrivere lo stato di malattia dei pazienti in fase di recidiva.

In questa sezione, viene stimato un modello logistico robusto con la sola intercetta e viene attuata una procedura di selezione delle variabili in avanti, basata sulla significatività del test alla Wald. Nella selezione in avanti delle variabili, il livello di *IgM* risulta essere l'unica covariata significativa, seguita dal livello di *IgM*, mentre le altre covariate non risultano significative per descrivere lo stato di malattia dei pazienti in fase di recidiva. In tabella 6.4 vengono riportate le stime dei parametri del modello logistico robusto.

Variabile	Stima	Std. Error	z-value	p-value
Intercetta	7.776	3.90	-1.99	0.046
IgM	-0.0497468	0.0285354	-1.743	0.0813

Tabella 6.4: Modello logistico robusto stimato per lo stato di malattia dei pazienti in fase di recidiva.

Dalla tabella emerge che, a parità del livello di *Neutrofili*, per ogni aumento unitario del livello di *IgM*, la quota di pazienti guariti diminuisce 4.84% volte rispetto alla quota di pazienti guariti con nessun aumento unitario del livello di *IgM*; mentre, diversamente dal risultato ottenuto nella sezione 4.2, dall'adattamento del modello logistico robusto emerge che il livello di *Neutrofili* non risulta essere significativo ad un livello di α del 10%. In figura 6.3 vengono riportate le stime dei coefficienti di regressione ottenute con il modello logistico e con il modello logistico robusto per il livello di *Neutrofili* con i relativi intervalli di confidenza 0.9.

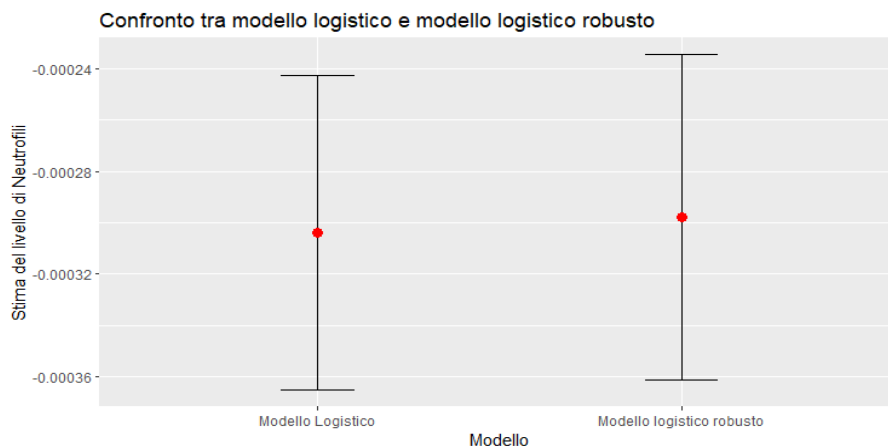


Figura 6.3: Stime dei coefficienti di regressione del modello logistico e del modello logistico robusto con intervalli di confidenza di livello 0.9 per il livello di *Neutrofili*.

Dalla figura emerge che il livello di *Neutrofili* non risulta essere significativo per descrivere lo stato della malattia dei pazienti in fase di recidiva secondo il modello logistico robusto. In figura 6.4 vengono riportate le stime dei coefficienti di regressione ottenute con il modello logistico e con il modello logistico robusto per il livello di *IgM* con i relativi intervalli di confidenza 0.9.

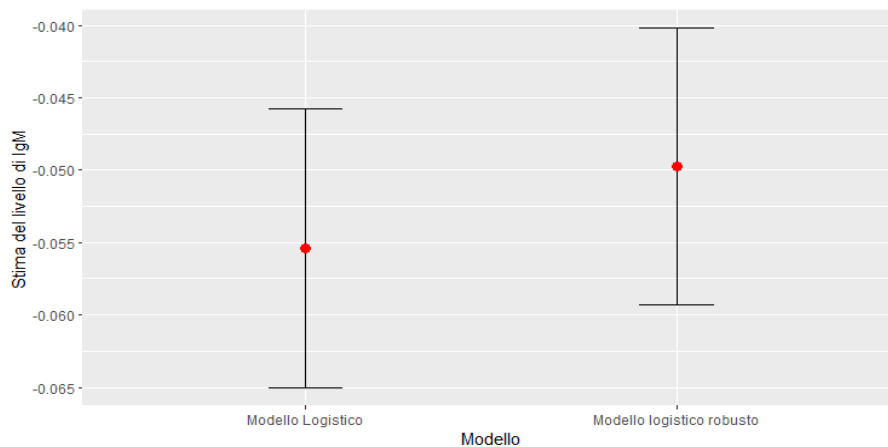


Figura 6.4: Stime dei coefficienti di regressione del modello logistico e del modello logistico robusto con intervalli di confidenza di livello 0.9 per il livello di *IgM*.

Dalla figura emerge che le stime e gli intervalli di confidenza del coefficiente di regressione del livello di IgM ottenute tramite il modello logistico e tramite il modello logistico robusto risultano essere molto simili. In conclusione, dal modello logistico robusto emerge che solo il livello di IgM risulta essere significativo per descrivere lo stato della malattia dei pazienti in fase di recidiva.

Capitolo 7

Analisi di durata

In questo capitolo viene studiato il tempo di attesa alla recidiva dei pazienti (si veda Klein *et al.*, 2013 [8]). Il 54.2% dei pazienti nel momento della visita di controllo non hanno sviluppato la recidiva (si veda la Sezione 2.1.2) e risultano, dunque, dati censurati. La censura risulta essere una censura a destra di tipo casuale e la coorte è dinamica.

7.1 Analisi esplorative

In questa sezione vengono presentati le principali analisi esplorative univariate del tempo di attesa alla recidiva dei soggetti.

In figura 7.1 viene riportata la curva di Kaplan-Meier per tutti i soggetti, con i rispettivi intervalli di confidenza di livello 0.95. I segni rossi lungo la curva rappresentano i dati censurati, ovvero i pazienti che non hanno sviluppato la recidiva.

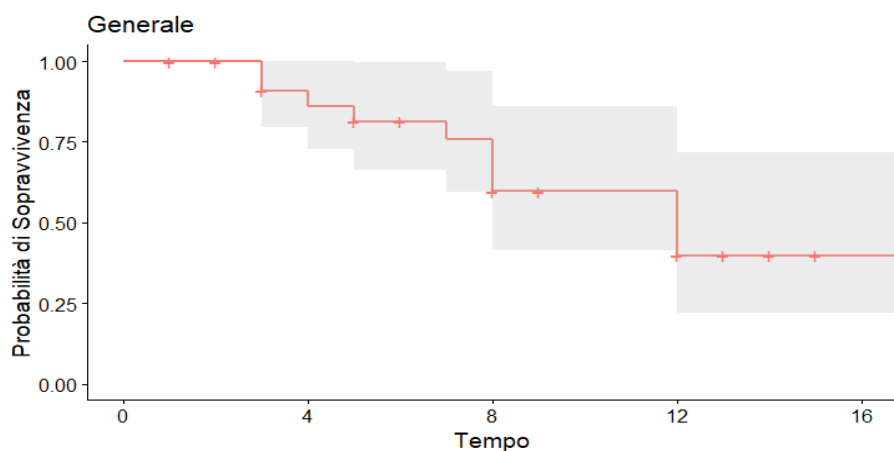


Figura 7.1: Curva di K-M generale dei soggetti al momento della visita di controllo.

Dalla figura 7.1 emerge che il tempo mediano di sopravvivenza dei pazienti è pari a 12 settimane.

Vengono riportate le funzioni di sopravvivenza (con intervalli di confidenza 0.95) dei pazienti, ottenuti tramite lo stimatore di *Kaplan-Meier*, rispetto ai quartili del livello di *Leucociti* (figura 7.2), *Neutrofili* (figura 7.3), *PON1* (figura 7.4), *CRP* (figura 7.5), *Aptoglobina* (figura 7.6), *Ferritina* (figura 7.7), *Albumine* (figura 7.8), *IgG* (figura 7.9), *IgM* (figura 7.10) e *IgA* (figura 7.11). Viene, inoltre, riportato il test di *Peto-Peto* per verificare l'uguaglianza delle curve di sopravvivenza, che risulta più potente del *log-rank test* quando le curve

nei gruppi non sono proporzionali¹. Essendo i gruppi suddivisi in quartili, le numerosità campionarie dei parametri sanguigni per ciascun gruppo sarà circa pari a 6 unità statistiche.

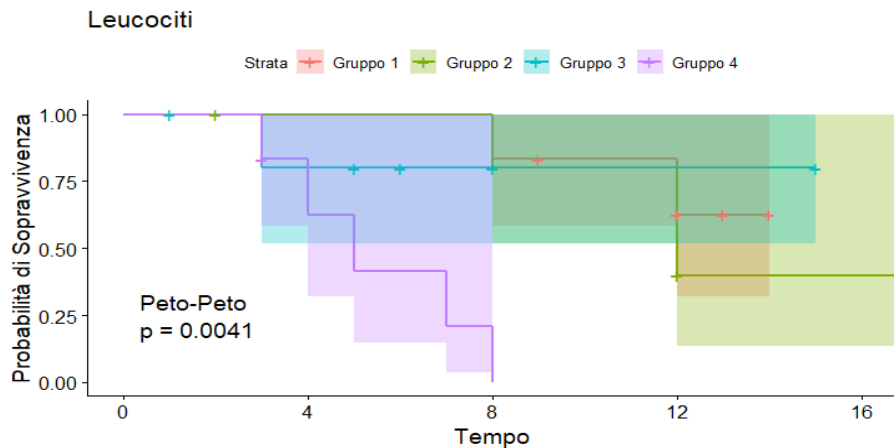


Figura 7.2: Curva di K-M dei soggetti rispetto al livello di *Leucociti* al momento della visita di controllo.

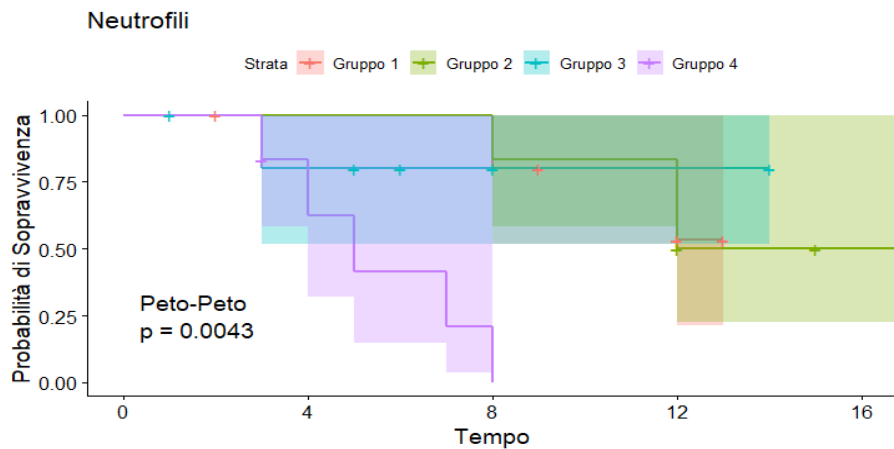
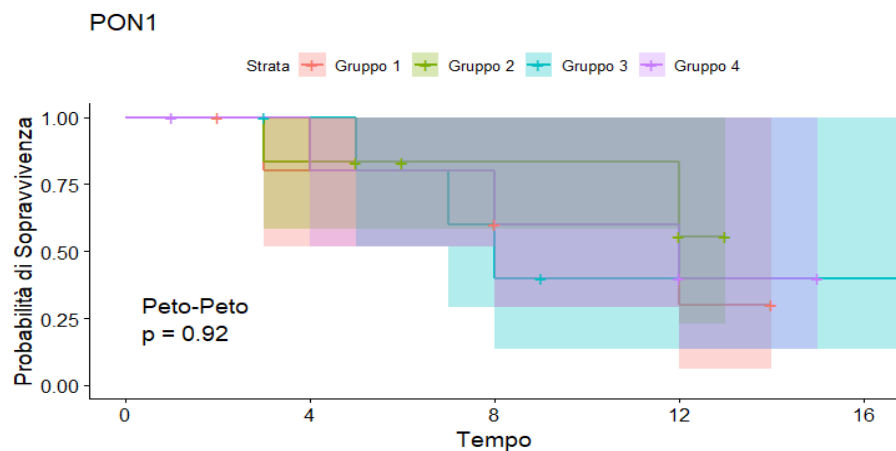
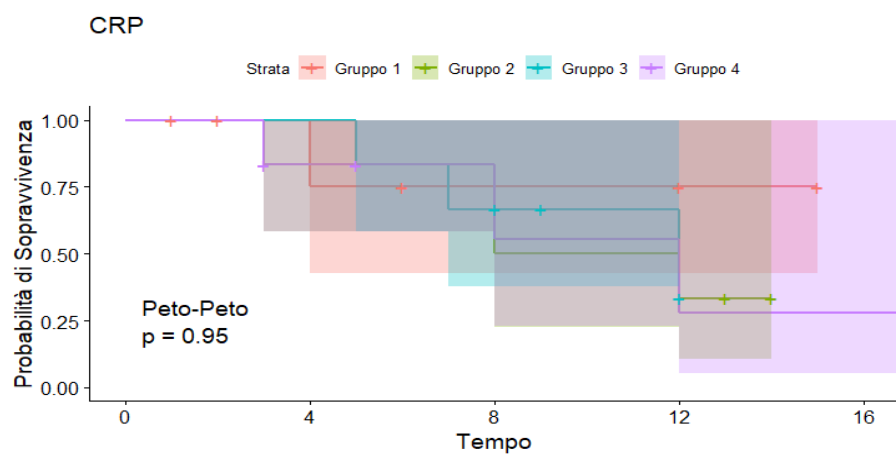
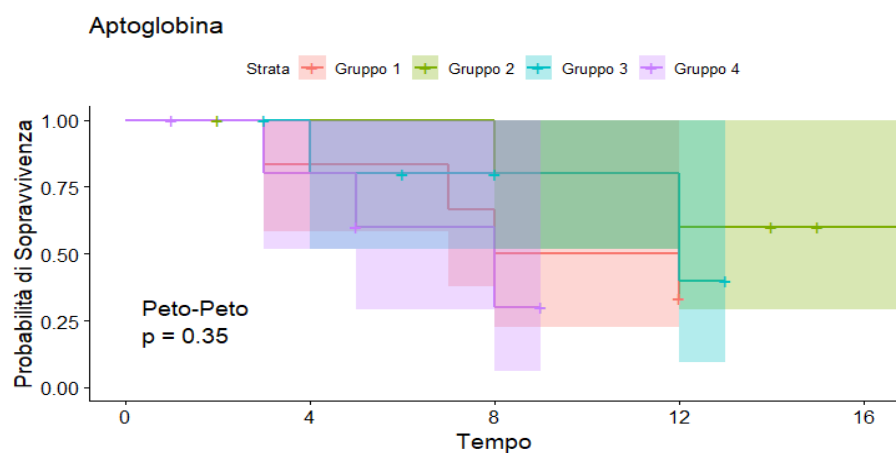


Figura 7.3: Curva di K-M dei soggetti rispetto al livello di *Neutrofili* al momento della visita di controllo.

¹I risultati risentono della bassa numerosità campionaria.

Figura 7.4: Curva di K-M dei soggetti rispetto al livello di *PON1* al momento della visita di controllo.Figura 7.5: Curva di K-M dei soggetti rispetto al livello di *CRP* al momento della visita di controllo.Figura 7.6: Curva di K-M dei soggetti rispetto al livello di *Aptoglobina* al momento della visita di controllo.

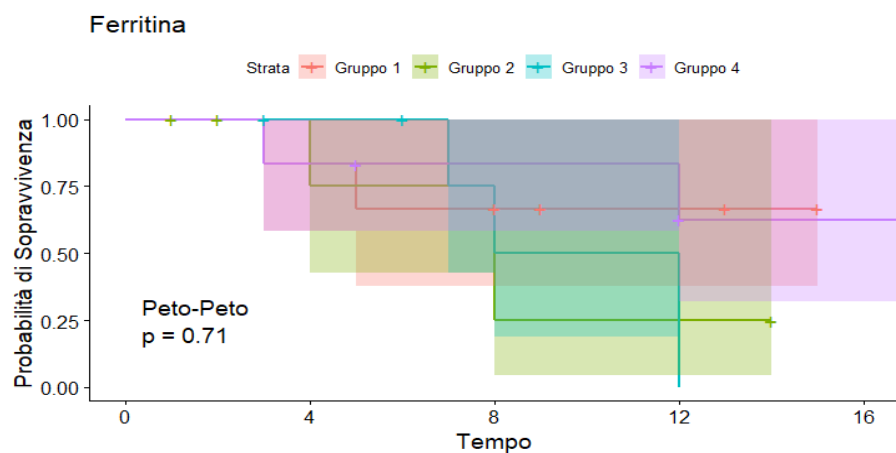


Figura 7.7: Curva di K-M dei soggetti rispetto al livello di *Ferritina* al momento della visita di controllo.

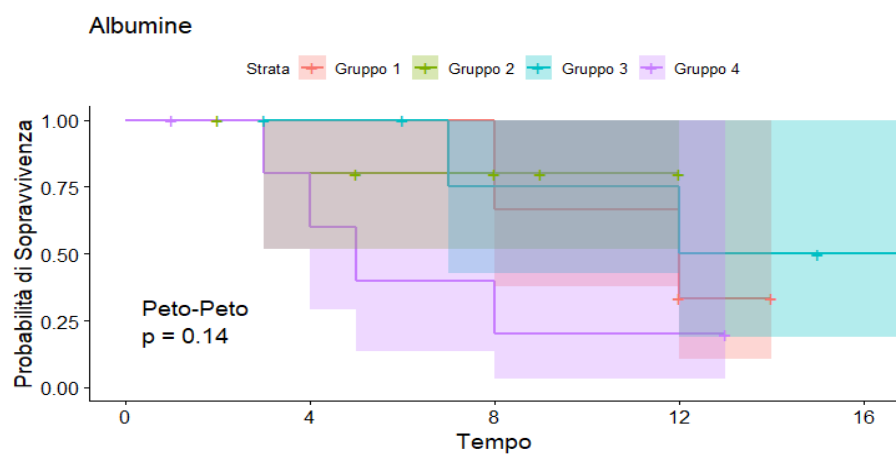


Figura 7.8: Curva di K-M dei soggetti rispetto al livello di *Albumine* al momento della visita di controllo.

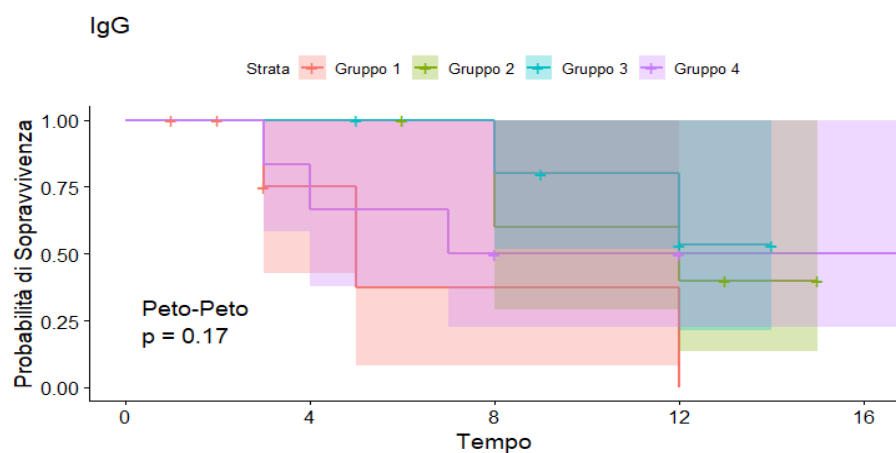


Figura 7.9: Curva di K-M dei soggetti rispetto al livello di *IgG* al momento della visita di controllo.

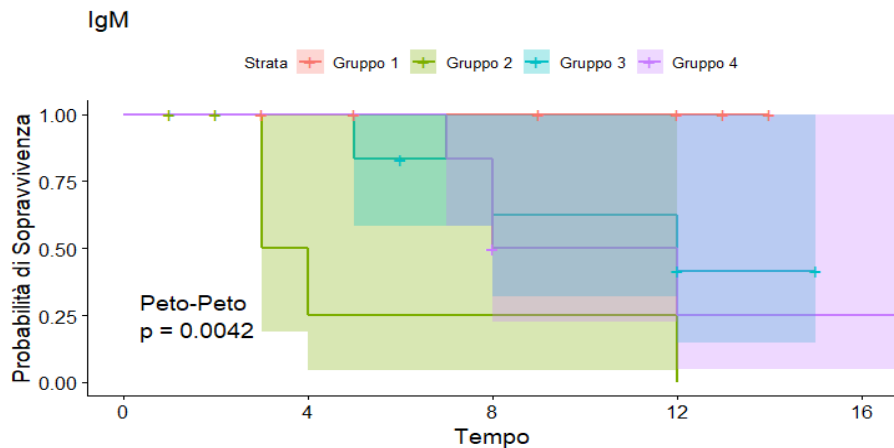


Figura 7.10: Curva di K-M dei soggetti rispetto al livello di *IgM* al momento della visita di controllo.

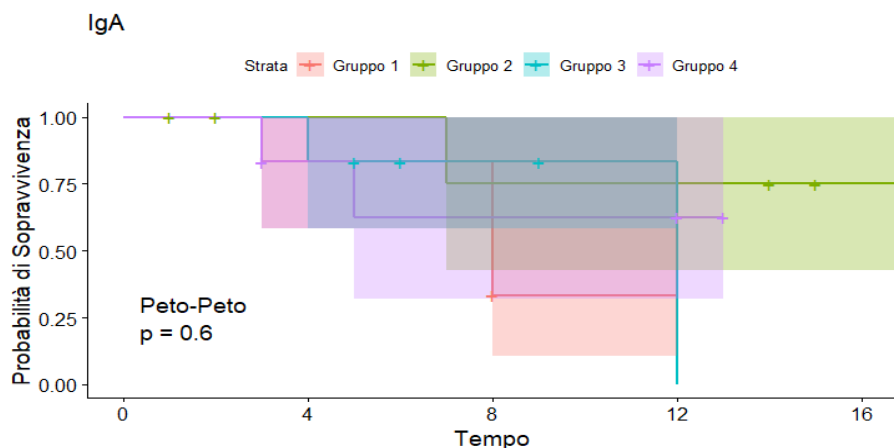


Figura 7.11: Curva di K-M dei soggetti rispetto al livello di *IgA* al momento della visita di controllo.

Dalle figure emerge che i soggetti con alti livelli di *Leucociti* e *Neutrofili*, riportano tempi minori di attesa tra la visita di ingresso e lo sviluppo della recidiva. Inoltre, il test di *Peto-Peto*, a un livello di significatività del 5%, porta a concludere che il tempo di attesa alla recidiva è significativamente diversa per i diversi quartili del livello di *Leucociti*, *Neutrofili* e *IgM*.

7.2 Modello di Cox

In questa sezione viene stimato un modello di regressione semiparametrico di Cox a rischi proporzionali per verificare quali fattori aumentino o diminuiscano il rischio di sviluppare la recidiva nei soggetti. Il modello di Cox a rischi proporzionali presenta la seguente forma:

$$h(X|Z) = h_0(X)exp(\beta^t Z),$$

dove Z rappresenta le covariate, β i coefficienti di regressione, X il *Tempo* e h_0 la funzione di rischio di base. A seguito di una procedura automatica di selezione delle variabili basata su un criterio di minimizzazione dell'AIC (si veda Salvan *et al.*, 2020), si ottiene il modello stimato finale riportato in tabella 7.1.

Variabile	β_i	$\exp(\beta_i)$	$\text{se}(\beta_i)$	test z	p-value
Neutrofili	0.0002681	1.0002681	0.0001051	2.550	0.01076
PON1	0.4261705	1.5313818	0.2798449	1.523	0.12779
Aptoglobina	0.0049668	1.0049791	0.0020262	2.451	0.01423
IgG	0.0065676	1.0065892	0.0025133	2.613	0.00897
IgA	-0.0913195	0.9127261	0.0636819	-1.434	0.15157

Tabella 7.1: Modello di Cox a rischi proporzionali stimato.

Dalla tabella 7.1 emerge che per ogni aumento unitario del livello di *Neutrofili*, il rischio di sviluppare la recidiva aumenta del 0.2%; per ogni aumento unitario del livello di *Aptoglobina*, il rischio di sviluppare la recidiva aumenta del 0.4%; mentre, per ogni aumento unitario del livello di *IgG*, il rischio di sviluppare la recidiva aumenta del 0.6%. Si nota, infine, che ad un livello di α del 10%, il livello di *PON1* e il livello di *IgA* non risultano significativi.

In conclusione, alti livelli di *Neutrofili*, *Aptoglobina* e *IgG* risultano dei fattori di rischio per lo sviluppo della recidiva.

7.3 Bontà di adattamento

Per verificare la bontà di adattamento globale del modello di Cox ai dati, in figura 7.12 viene riportato il grafico dei residui di Cox-Snell (in rosso la bisettrice).

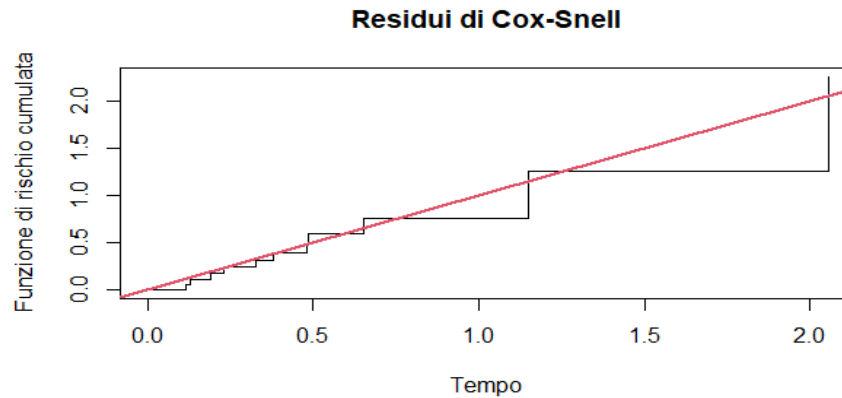


Figura 7.12: Residui di Cox-Snell per la verifica dell'adattamento globale del modello di Cox finale ai dati.

Dalla figura 7.12 emerge che la funzione di rischio cumulata (ottenuta con lo stimatore di Nelson-Aalen), calcolata nei residui di Cox-Snell, contro il tempo, giace sulla bisettrice; dunque, il modello stimato di Cox sembra adattarsi bene ai dati.

Oltre all'adeguamento globale del modello ai dati, bisogna anche verificare che le variabili non dipendano dal tempo e che sia rispettato il vincolo di proporzionalità dei rischi, sia globalmente che per ogni variabile. Per verificare queste due ipotesi e convalidare il modello, si studiano i residui scalati di Schoenfeld per il livello di *Neutrofili* (figura 7.13), *PON1* (figura 7.13), *Aptoglobina* (figura 7.14), *IgG* (figura 7.14) e *IgA* (figura 7.15).

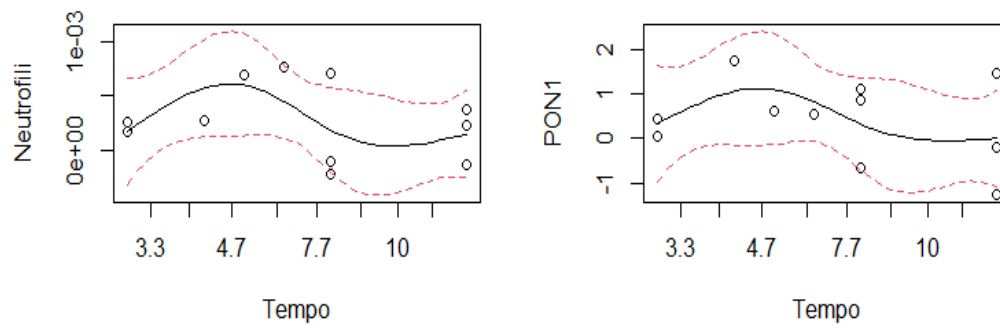


Figura 7.13: Residui di Shoenfeld per il livello di *Neutrofili* e il livello di *PON1*

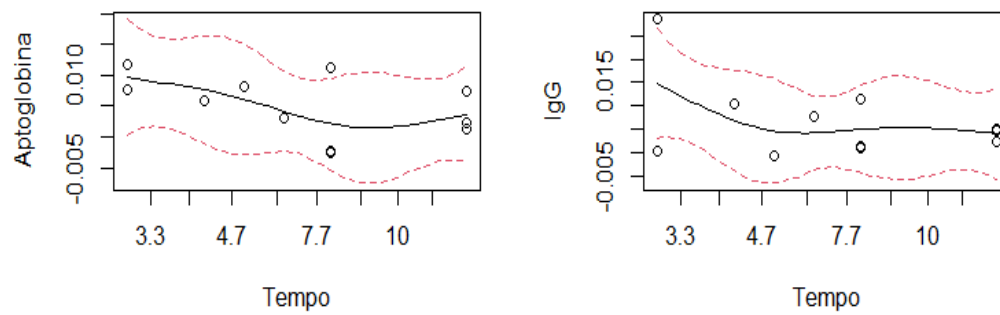


Figura 7.14: Residui di Shoenfeld per il livello di *Aptoglobina* e il livello di *IgG*

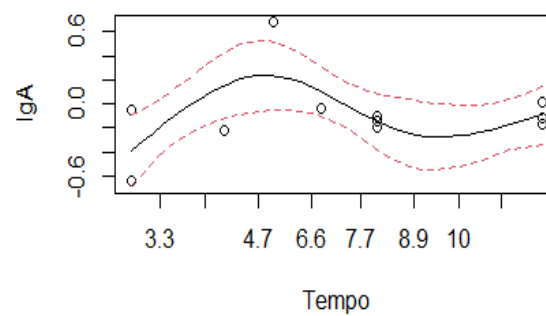


Figura 7.15: Residui di Shoenfeld per il livello di *IgA*

Dalle figure non emergono andamenti sistematici dei residui di Shoenfeld in funzione del tempo (covariate tempo indipendenti) e sembra esserci una retta orizzontale che possa interpolare i residui di Shoenfeld (verifica della proporzionalità). I p-value dei test per verificare che i coefficienti angolari delle curve che interpolano i dati siano pari a zero sono riportati in tabella 7.2.

Variabile	Test	P-value
Neutrofili	0.2707	0.6
PON1	0.1303	0.72
Aptoglobina	0.0907	0.76
IgG	2.6531	0.10
IgA	0.0168	0.9
Globale	7.961	0.16

Tabella 7.2: Test e p-value del test per verificare la proporzionalità dei rischi delle variabili del modello di Cox finale.

Dalla tabella 7.2 emerge che l'ipotesi di proporzionalità è accettata a tutti i livelli usuali di α , sia globalmente che rispetto a tutte le covariate.

In conclusione, il modello di Cox, secondo i residui di Cox-Snell si adatta bene ai dati e i residui scalati di Shoenfeld convalidano il modello perché confermano l'ipotesi di indipendenza tra le covariate e il tempo e confermano l'ipotesi di proporzionalità tra i rischi, sia globalmente che rispetto ad ogni singola covariata.

Capitolo 8

Conclusioni

In questo capitolo vengono presentati i principali risultati e le principali conclusioni del report e le risposte a tutte le domande di ricerca, secondo i risultati ottenuti.

8.1 In ingresso

I dati presentano 52 cani selezionati in ingresso nella clinica, di cui il 29% risulta essere deceduto per infiammazione al sistema nervoso centrale, il 46% è guarito e il 25% ha sviluppato la recidiva. Le analisi esplorative bivariate hanno rivelato come lo stato della malattia nei soggetti sia indipendente dalla presenza di una terapia steroidea in atto, mentre i parametri infiammatori non presentino distribuzioni diverse in media o mediana rispetto allo stato della malattia: indipendentemente dallo stato della malattia dei soggetti, i parametri sanguigni in ingresso sono statisticamente equivalenti nei tre gruppi. Inoltre, i parametri sanguigni risultano molto poco correlati tra loro, fatta eccezione per il livello di *Neutrofili* e di *Leucociti*, che riportano una correlazione di Spearman pari a 0.931. Un'analisi dei gruppi effettuata sui soggetti risulta non presentare significative differenze tra i parametri sanguigni dei soggetti in ingresso con lo stato della malattia e la terapia steroidea. Il modello di regressione logistico per verificare quali parametri infiammatori in ingresso influenzino lo stato della malattia dei soggetti riportano i seguenti risultati:

1. La terapia steroidea e i parametri infiammatori in ingresso dei soggetti non sono significativi per descrivere la guarigione oppure lo sviluppo di recidiva di infiammazione al sistema nervoso centrale nel paziente.
2. Il livello di *Ferritina* risulta essere l'unica covariata significativa per descrivere la guarigione oppure il decesso del soggetto per infiammazione al sistema nervoso centrale: ad ogni aumento unitario del livello di *Ferritina* nel sangue dei pazienti, la quota di pazienti guariti è 0.57% volte inferiore rispetto alla quota di pazienti guariti con nessun aumento del livello di *Ferritina*; quindi la *Ferritina* risulta essere un fattore di rischio per il decesso del paziente affetto da infiammazione al sistema nervoso centrale. Il modello riporta un'accuratezza di classificazione pari a 69.2% e un AUC pari a 66.5%; il test di Hosmer-Lemeshow riporta un p-value pari a 0.77, che porta ad accettare a tutti i livelli usuali di α l'ipotesi che le frequenze attese stimate dal modello e le frequenze osservate siano coerenti.
3. Il livello di *Ferritina* risulta essere l'unica covariata significativa per descrivere la guarigione oppure il decesso e la recidiva del soggetto per infiammazione al sistema nervoso centrale: ad ogni aumento unitario del livello di *Ferritina* nel sangue dei pazienti, la quota di pazienti guariti è 0.52% volte inferiore rispetto alla quota di pazienti guariti con nessun aumento del livello di *Ferritina*; quindi la *Ferritina* risulta essere un fattore di rischio per il decesso oppure per lo sviluppo della recidiva del paziente affetto da infiammazione al sistema nervoso centrale. Il modello riporta un'accuratezza di classificazione pari a 61.5% e un AUC pari a 65.8%; il test di Hosmer-Lemeshow riporta un p-value pari a 0.886, che porta ad accettare a tutti i livelli usuali di α l'ipotesi che le frequenze attese stimate dal modello e le frequenze osservate siano coerenti.

4. La terapia steroidea non risulta significativa per descrivere lo stato della malattia dei pazienti, ad un livello di significatività del 10%, in nessun modello preso in considerazione.

I metodi asintotici di ordine elevato (maggiore precisione per piccoli campioni) e la regressione robusta (robusta rispetto a valori anomali delle covariate) confermano i risultati ottenuti con la regressione logistica per i pazienti in ingresso.

8.2 In fase di recidiva

I dati presentano 24 cani selezionati in fase di recidiva, di cui il 54.17% è guarito da infiammazione al sistema nervoso centrale e il 45.83% ha sviluppato la recidiva. Le analisi esplorative bivariate hanno rivelato che non c'è differenza in distribuzione tra i parametri infiammatori in fase di recidiva tra i pazienti guariti e che hanno sviluppato la recidiva, tranne che per il livello di *IgM*, a un livello di α del 5%. I parametri infiammatori in fase di recidiva risultano essere poco correlati fra loro, fatta eccezione per il livello di *Leucociti* e il livello di *Neutrofili* nel sangue dei pazienti in fase di recidiva, tra i quali intercorre una correlazione di Spearman pari a 0.974. Le tecniche di analisi multivariata, poiché le variabili sono poche correlate tra loro, riportano che le prime 4 componenti principali spiegano il 75% della variabilità totale; mentre, i gruppi ottenuti con il metodo delle k-medie con due centroidi sono in grado di classificare correttamente il 70.8% delle osservazioni nello stato della malattia corrispondente, quindi i parametri infiammatori in fase di recidiva non distinguono in modo sufficientemente netto il gruppo dei pazienti guariti dal gruppo dei pazienti che hanno sviluppato la recidiva. Il modello di regressione logistico con la sola intercetta, con funzione di legame *logit*, a seguito di una procedura di selezione in avanti delle variabili basata sul test log-rapporto di verosimiglianza, con un livello di significatività del 10%, porta a concludere che solo il livello di *Neutrofili* e il livello di *IgM* nel sangue dei pazienti nel momento della recidiva sia significativo per descrivere lo stato di malattia dei pazienti. Alti livelli di *Neutrofili* e di *IgM* risultano essere dei fattori di rischio per la probabilità di sviluppare la recidiva. In particolare, per ogni aumento unitario del livello di *Neutrofili* nel sangue dei pazienti, la quota di pazienti guariti è 0.03% volte inferiore rispetto alla quota di pazienti guariti con nessun aumento unitario del livello di *Neutrofili*; mentre per ogni aumento unitario del livello di *IgM* nel sangue dei pazienti in fase di recidiva, la quota di pazienti guariti è 5.3% volte inferiore rispetto alla quota di pazienti guariti con nessun aumento del livello di *IgM*. La tabella di corretta classificazione riporta un'accuratezza di classificazione del modello pari a 75%, l'AUC (area sottesa alla curva ROC) riporta un valore pari a 0.81: il modello si adatta sufficientemente bene ai dati. Infine, il test di Hosmer-Lemeshow riporta un valore osservato della statistica test pari a 13.01 e un p-value pari a 0.11: i dati non sono contrari all'ipotesi che le frequenze attese stimate dal modello e le frequenze osservate siano coerenti.

I metodi asintotici di ordine elevato (maggiore precisione per piccoli campioni) confermano i risultati ottenuti con la regressione logistica per i pazienti in fase di recidiva; mentre, secondo il modello logistico robusto, il livello di *Neutrofili* non è significativo per descrivere lo stato della malattia dei pazienti in fase di recidiva.

Le analisi esplorative per i dati di durata portano a concludere che le curve di sopravvivenza dei parametri sanguigni dei pazienti in fase di recidiva suddivisi in quartili, secondo il test di *Peto-Peto*, a un livello di significatività del 5%, risultano omogenee, fatta eccezione per il livello di *Leucociti*, *Neutrofili* e *IgM*. Il modello semiparametrico di Cox a rischi proporzionali a seguito di una procedura automatica di selezione delle variabili basata sull'AIC, porta a concludere che il rischio di sviluppare la recidiva nei pazienti aumenta del 0.2% per ogni aumento unitario del livello di *Neutrofili*, del 0.4% per ogni aumento unitario del livello di *Aptoglobina* e del 0.6% per ogni aumento unitario del livello di *IgG*; mentre, il livello di *PON1* e *IgA* non risultano significativi ad un livello di significatività del 10%. Infine, il modello di Cox, secondo i residui di Cox-Snell si adatta bene ai dati e, secondo i residui scalati di Shoenfeld per la convalida del modello, le ipotesi di variabili tempo indipendenti e di proporzionalità tra i rischi (sia globalmente che rispetto ad ogni covariata) sono accettate a tutti i livelli usuali di α .

Bibliografia

- [1] A.R. Brazzale. Approximate conditional inference in logistic and loglinear models. 1998.
- [2] A.R. Brazzale, A.C. Davison, and N. Reid. *Applied Asymptotics: Case Studies in Small-sample Statistics*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, 2007.
- [3] A.C. Davison, D.V. Hinkley, R. Gill, and B.D. Ripley. *Bootstrap Methods and Their Application*. Bootstrap Methods and Their Application. Cambridge University Press, 1997.
- [4] A. Farcomeni and L. Ventura. An overview of robust methods in medical research. 1(111-133), 2012.
- [5] M. Grigoletto, L. Ventura, and F. Pauli. *Modello Lineare: Teoria e Applicazioni con R*. Giappichelli, 2017.
- [6] S. Heritier, E. Cantoni, S. Copt, and M.P. Victoria-Feser. *Robust Methods in Biostatistics*. Wiley Series in Probability and Statistics. Wiley, 2009.
- [7] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis: Pearson New International Edition*. Pearson Education, 2013.
- [8] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer New York, 2013.
- [9] A. Salvan, N. Sartori, and L. Pace. *Modelli Lineari Generalizzati*. UNITEXT. Springer Milan, 2020.
- [10] L. Ventura and W. Racugno. *Biostatistica. Casi di studio in R*. I Manuali. EGEA, 2017.
- [11] P. Vineis, P. Duca, and P. Pasquini. *Manuale di metodologia epidemiologica*. Fuori collana. Carocci, 1988.

Note metodologiche

Il metodo di campionamento delle unità statistiche e la selezione delle variabili è stata attuata da [Clinica Veterinaria San Marco](#) di Veggiano.

Il software utilizzato per le analisi è [RStudio Team \(2020\)](#). [RStudio: Integrated Development for R](#). RStudio, PBC, Boston, MA.

Il software utilizzato per la stesura del report è $\text{\LaTeX}$, tramite l'ausilio di [Overleaf](#).