Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in Statistica per l'Economia e l'Impresa

Relazione Finale

Diagnosi della Leishmaniosi tramite lo studio della regressione logistica penalizzata

Relatore: Professoressa Laura Ventura
Dipartimento di Scienze Statistiche

Candidato: Giorgia Morello
N.matricola: 2033961

Anno Accademico: 2023-2024

Indice

1 Capitolo 1: Introduzione

1.1 Un primo sguardo alla malattia

La Leishmaniosi canina è una delle principali malattie fatali per l'uomo e per i cani, poiché è possibile una trasmissione dall'animale all'uomo. Tale patologia è diffusa in più di 70 paesi del mondo [1]. E' presente in alcune regioni dell'Europa meridionale, Africa, Asia, America meridionale e centrale ed è stata segnalata anche negli Stati Uniti d'America (Figura 1).

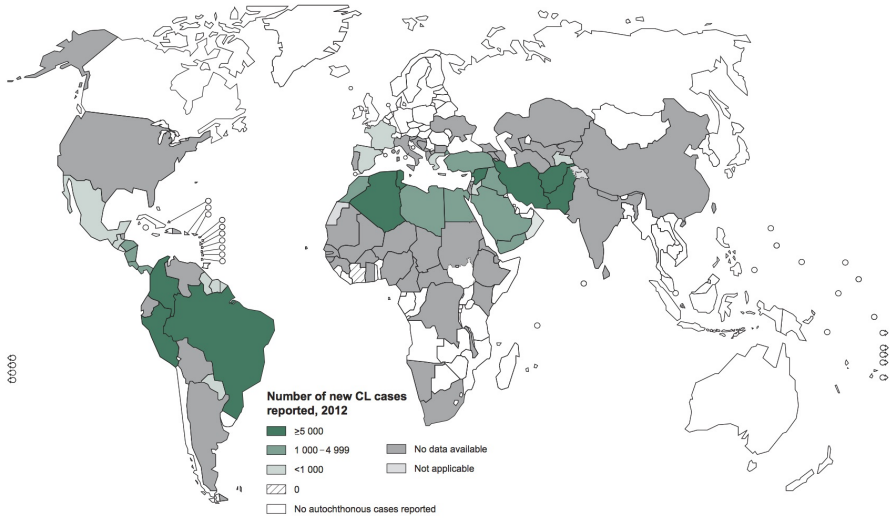

Figura 1. Distribuzione nel mondo della Leishmaniosi canina

La Leishmaniosi canina si manifesta in un ampio spettro di segnali clinici e gravità. Un gruppo di scienziati veterinari, denominato LeishVet, composto da veterinari appartenenti a istituzioni accademiche del bacino del Mediterraneo e del Nord America, si occupano di tale patologia. Il loro obiettivo è andare a fondo nello studio dell'infezione al fine di poter comprendere con una maggiore precisione quando tale malattia si verifica, la velocità del suo sviluppo e trovare eventuali prevenzioni al fine che si contragga con la minor frequenza possibile.

1.2 I sintomi

I quadri clinici della Leishmaniosi canina sono molteplici, la raccolta dei segni clinici deve essere accompagnata da esami di laboratorio per giungere d una diagnosi corretta. I sintomi possono comparire in tempi differenti e associarsi nella più varia combinazione [2]. L'andamento tende comunque a essere di tipo sub-acuto o cronico e solo raramente vengono registrate forme più acute con insorgenza di febbre; tuttavia alcuni sintomi sono più frequenti di altri,

come riportati nella Tabella 1.

Segni clinici	**proporzione positività**
linfoadenomegalia	0.57
pallore delle mucose	0.58
moderata o severa splenomegalia	0.53
perdita di peso	0.32
dermatite secca esfogliativa	0.56
ulcere	0.4
alopecia periorbitale	0.18
alopecia diffusa	0.14
onicogrifosi	0.24
segni oculari	0.16

Tabella 1: prevalenza dei principali segni clinici infetti

1.3 Il ciclo della Leishmaniosi

Il ciclo della Leismaniosi inizia con l'infezione di un vertebrato, che avviene attraverso la puntura del vettore, solitamente appartenente a varie specie di flebotomo[1]. Il vettore infetto deposita i promastigoti[2] nella cute del cane; successivamente un tipo particolare di cellule, i monociti, del sistema immunitario del cane "ingloba" questi promastigoti, che si trasformano in amastigoti e iniziano a moltiplicarsi all'interno dell'ospite vertebrato. Raggiunto un certo numero di parassiti, il monocita subisce il processo di lisi[3] e libera gli amastigoti che andranno a invadere altre cellule della stessa linea.

[1] Principale vettore della Leishmaniosi è un insetto di colore giallo pallido o sabbia di piccole dimensioni, da 2 a 4 mm, con il corpo ricoperto da una fitta peluria che interessa anche le grandi ali.

[2] Forma infettante nelle ghiandole salivari del vettore.

[3] Demolizione e dissoluzione di una cellula, causata dalla rottura della membrana cellulare.

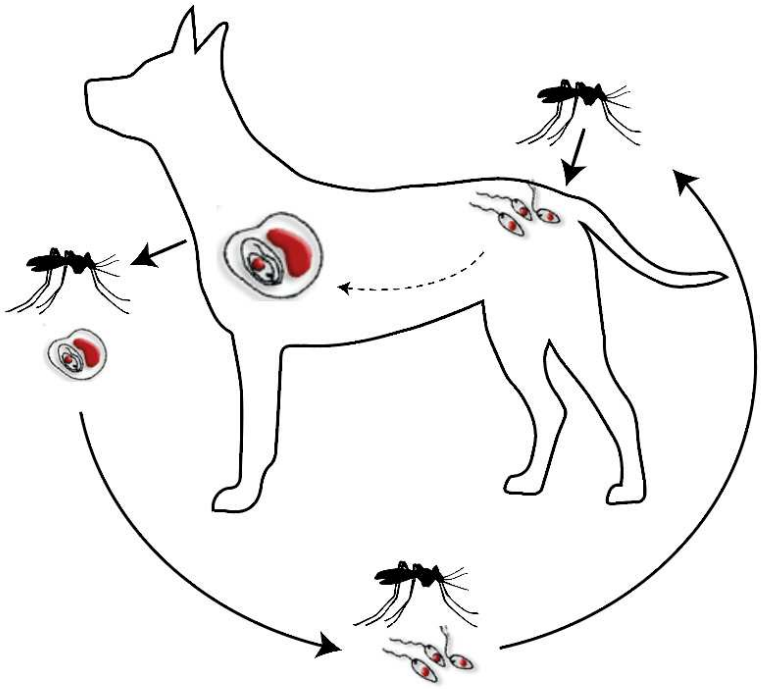

Figura 2. Ciclo Leishmaniosi canina

La malattia si realizza in quei soggetti in cui prevale un certo tipo di risposta immunitaria, che non è in grado di attivare efficacemente una serie di meccanismi che porterebbero alla produzione di cellule immunitarie, capaci di sconfiggere il parassita. Alcune cause per le quali prevale un certo tipo di risposta immunitaria sono eventuali malattie intercorrenti o stress di varia natura, quali gravidanza, malnutrizione o addirittura affaticamento eccessivo.

1.4 Test ELISA

L'ELISA è una tecnica immunoenzimatica che permette di rilevare e quantificare gli anticorpi specifici nei confronti di Leishmania Infantum [3]. Diversi sono i suoi vantaggi, tra i quali un'elevata sensibilità e specificità, che permette di discriminare anche lievi aumenti o diminuzioni del livello di anticorpi ed una lettura oggettiva poichè il test si effettua mediante un lettore spettrofotometrico[4].
L'esito può essere "Negativo", "Dubbio" o "Positivo":

1. Negativo: il cane non ha anticorpi specifici contro la Leishmania Infantum, quindi se è presente un forte sospetto che gli eventuali segni clinici, lesioni

[4]Dispositivo che legge le micropiastre.

o alterazioni laboratoristiche siano compatibili con questa infezione, si consiglia l'esecuzione di altre indagini nei tessuti più sospetti;

2. Dubbio: il comportamento dovrebbe essere lo stesso del test risultante come negativo, relativo ai cani sieronegativi. Si consiglia di ripetere il test sierologico dopo 4-6 mesi per valutare un'eventuale variazione del livello anticorporale;

3. Positivo: il test positivo testimonia la presenza di anticorpi specifici contro la Leishmania infantum. L'interpretazione deve essere correlata a livello anticorporale e a tale fine sono state definite tre categorie di positività:

   Livello basso (Positivo basso): in questo caso, se è presente un altro sospetto che i segni clinici, le lesioni o le alterazioni laboratoristiche siano compatibili con l'infezione da Leishmania Infantum, si consiglia l'esecuzione di ulteriori indagini e si consiglia di ripetere il test sierologico dopo 3-6 mesi per valutare un'eventuale variazione a livello anticorporale;

   Livello medio (Positivo medio): si consiglia anche in questo caso l'esecuzione di ulteriori indagini e si invita a ripetere il test sierologico dopo 3-6 mesi per valutare, come nel caso precedente, eventuali variazioni a livello anticorporale;

   Livello alto (Positivo alto): è presente un livello anticorporale elevato e ciò indica la disseminazione del parassita nel cane. E' probabile che i segni clinici, lesioni o alterazioni laboratoristiche siano causate da Leishmania Infantum.

Ovviamente ciò non permette di escludere altre malattie o infezioni concomitanti. Il medico veterinario può compiere altri passi diagnostici utili per il monitoraggio del cane dopo l'inizio della terapia, come una PCR Real Time[5] o la misurazione delle proteine di fase acuta, quali proteina C reattiva[6], aptoglobina[7] ed altre.

1.5 Uno sguardo al campione

1.5.1 Presentazione caso di studio

I dati raccolti fanno riferimento a uno studio clinico svolto presso una clinica veterinaria. E' stato esaminato un campione di 57 cani, sottoposti a una visita. Per ogni paziente sono state rilevate 33 variabili sia di tipo anagrafico

[5] E' un metodo che simultaneamente amplifica e quantifica il DNA

[6] E' un indice di infiammazione

[7] E' una proteina prodotta dal fegato che ha il compito di rimuovere l'emoglobina libera nel circolo sanguigno

(come l'età, il sesso, lo stato sessuale e la razza) sia variabili di tipo medico (quali valori nel sangue, registrazioni dei valori di alcune proteine) ed infine una suddivisione in alcuni gruppi per identificare vari stadi della malattia nel caso di positività. Le variabili che fanno riferimento al dataset sono:

1. Group: variabile fattoriale a due livelli, indica se i pazienti sono sani o malati (0 se sani, 1 se malati);
2. BREED: variabile fattoriale a due livelli, indica se il cane sia di razza (*Yes* se è di razza, *NO* se è di razza mista);
3. AGE: variabile quantitativa continua, che rileva l'età del paziente al momento della visita;
4. SEX: variabile fattoriale a due livelli, che rileva il sesso del paziente (M se maschio, F se femmina);
5. SEXUALSTATUS: variabile fattoriale a due livelli, che indica lo stato sessuale del paziente (C se castrato, I se intero);
6. BW: variabile quantitativa continua, rileva il peso in Kg;
7. HR: variabile quantitativa continua, indica la frequenza cardiaca (si misura in bpm);
8. RR: variabile quantitativa continua, indica la frequenza respiratoria (si misurano il numero di respiri al minuto);
9. SBP: variabile quantitativa continua, indica la pressione sistolica (si misura in mmHg);
10. Clinicalsigs: variabile fattoriale a due livelli, rileva se sono presenti dei segni clinici nel paziente (*Yes* se li presentano, *none* se sono asintomatici)[8];
11. repellents: variabile fattoriale a due livelli, indica se il cane è stato trattato con un prodotto repellente in precedenza (*YES* se ha effettuato il trattamento, *NO* se non lo ha effettuato);
12. urinaryMCP: variabile quantitativa continua, corrisponde a un parametro utile nell'esame delle urine (pg/ml);
13. urineMCP: variabile quantitativa continua, combacia con urineMCP trasformato in scala in scala e-7 (mg/dl)[9];
14. urinarycreatinine: variabile quantitativa continua, indica la quantità di creatina nelle urine (si misura in mg/dL);
15. uAmCr: variabile quantitativa continua, corrisponde all'amilasi di creatinina urinaria (si misura in unità internazionali per litro, U.I./l);

[8] Tale variabile viene esclusa dallo studio
[9] Si decide di lavorare solo con urinaryMCP

16. uMCP: variabile quantitativa continua, corrisponde ad una proteina contenuta nelle urine;

17. CVurine: variabile quantitativa discreta, corrisponde ad un catetere vescicale (cm);

18. serumMCP: variabile quantitativa continua (pg/ml);

19. CVserum: variabile quantitativa discreta;

20. UPC: variabile quantitativa continua, corrisponde al rapporto proteine urinarie-creatinina urinaria;

21. USG: variabile quantitativa continua, corrisponde al peso specifico (g/ml di urina);

22. urea: variabile quantitativa continua, è una sostanza contenuta nel plasma (mg/dL);

23. creatinine: variabile quantitativa continua, corrisponde alla creatinina nel sangue (mg/dL);

24. SDMA: variabile quantitativa continua, corrisponde alla dimetilarginina simmetrica, che è un marker della funzionalità renale (g/dL);

25. PON1: variabile quantitativa continua, corrisponde alla glicoproteina sintetizzata nel fegato, che tende a diminuire in vari stati patologici (IU/L);

26. CRP: variabile quantitativa continua, corrisponde alla proteina C reattiva (mg/dL);

27. Hp: variabile quantitativa continua, corrisponde all'aptoglobina (mg/dL);

28. Ft: variabile quantitativa continua, corrisponde alla ferritina nelle urine (mg/grCrea);

29. Iron: variabile quantitativa continua, corrisponde al ferro totale (g/dL);

30. TIBC: variabile quantitativa continua, corrisponde alla capacità delle proteine di circolo ematico di legare il ferro (g/dL);

31. LeisvetStaging: variabile fattoriale a 5 livelli, indica degli stadi corrispondenti alla gravità di sviluppo della malattia (0, Ila, llb, lll,lV);

32. IRISStaging: variabile fattoriale a 5 livelli, è un altro tipo di raggruppamento che analizza la gravità della malattia (0, I, II, III, IV);

33. IRISstagingroup: variabile fattoriale a 3 livelli, è una suddivisione meno dettagliata di IRISStaging, raggruppandola in "soli" 3 gruppi (0, 1, 2);

1.6 Suddivisione in capitoli della relazione

Questa relazione è suddivisa in 4 capitoli. Dopo una presentazione del dataset, con la spiegazione delle variabili del caso di studio, si passa al secondo capitolo contenente le analisi esplorative univariate e bivariate, con lo scopo di individuare le caratteristiche di ciascuna variabile e le relazioni tra esse. Il terzo capitolo riguarda lo studio delle correzioni di Firth [7], Kosmidis *et al.* [8] e Kenne Pagui *et al.* [10] per il modello logistico penalizzato che si andrà ad adottare, studiando le differenze e le conclusioni che ciascuno di questi portano, valutando anche le bontà di adattamento di ciascun modello. Infine il quarto capitolo è dedicato ai risultati e alle conclusioni ottenuti dall'analisi statistica effettuata. Le analisi sono state effettuate tramite il software statistico RStudio (*www.rstudio.com*), considerando il livello di significatività pari a 0.05.
Nel prossimo capitolo si effettueranno le prime analisi esplorative dei dati a disposizione sia a livello descrittivo sia a livello grafico, per avere un'idea dei dati stessi ed anche per analizzare le relazioni tra variabili.

2 Capitolo 2: Analisi esplorative

L'analisi esplorativa consiste una preliminare indagine delle informazioni relative alle variabili considerate. Nel seguito, si effettuano inoltre opportuni test statistici[10] per individuare relazioni tra coppie di tali variabili, riportando solo i risultati significativi. I commenti si limiteranno a questi ultimi. Per tutti gli approfondimenti sui test utilizzati in questa analisi si rimanda a [4].

2.1 Analisi esplorative univariate

Nel campione in esame ci sono 19 pazienti sani e 38 malati (Tabella 1).

	N	%
0	19	0.333
1	38	0.667
Totale	57	100

Tabella 1: Frequenze assolute (N) e percentuali (%) della variabile group.

Il campione risulta essere per il 61.5% di razza e 38.5% di razza mista. Inoltre, è equamente distribuito per stato sessuale del cane, poiché il 50.8% è castrato, mentre il 49.2% non lo è. Il 43.8% è di sesso femminile e 56.2% è di sesso maschile. I cani che presentano dei segnali clinici sono in maggioranza (66.6%), mentre i soggetti asintomatici sono pari al 33.4%. Le variabili e i vari stadi della gravità della leishmaniosi sono descritti nella Tabella 2.

[10] Shapiro-Wilk per valutare la normalità, X2 di Pearson per le associazioni tra fattori, test t di Student e di Mann-Whitney (analogo non parametrico test t-Student) per variabili quantitative in due gruppi e ANOVA parametrica e test di Kruskal-Wallis (analogo non parametrico test ANOVA parametrica) per variabili quantitative in più gruppi per verificare rispettivamente omogeneità in medie o in distribuzioni (mediane).

	N	%
0	19	0.333
IIa	12	0.211
IIb	7	0.123
III	10	0.175
IV	9	0.158
Totale	57	100

Tabella 2: Frequenze assolute (N) e percentuali (%) della variabile LeisvetStaging.

Oltre alla classificazione Leisvet è stata effettuata anche la classificazione IRIS, la prima sempre in 5 gruppi, come nel caso della classificazione Lesivet, mentre la seconda raggruppando i gruppi dei vari stati di malattia in 2 macrogruppi, ottenendo una variabile qualitativa a 3 fattori, che si vedrà essere migliore per studiare i vari gruppi degli stati della malattia rispetto a IRISStaging, a causa della bassa numerosità all'interno di determinate classi.

	N	%
0	19	0.333
I	31	0.544
II	4	0.071
III	1	0.017
IV	2	0.035
Totale	57	100

Tabella 3

	N	%
0	19	0.333
I	31	0.544
II	7	0.123
Totale	57	100

Tabella 4

Tabella 3 e Tabella 4: Frequenze assolute (N) e percentuali (%) delle variabili IRISStaging e IRISStagingroup.

Per quanto riguarda la variabile repellents, risulta che al 55% del campione è stato applicato un repellente, mentre al 45% no.

Passando allo studio relativo alle variabili quantitative, l'età media riulta essere pari a 71.2 mesi (sd = 32.87 mesi). La statistica test di Shapiro-Wilk porta ad accettare l'ipotesi nulla di normalità della variabile AGE ($SW^{oss} = 0.97$, p-value = 0.51).

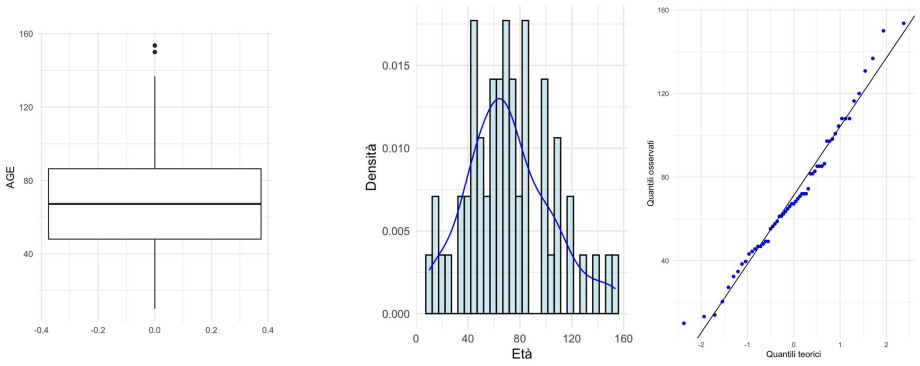

Figura 1. Boxplot, istogramma con densità lisciata e grafico q-q della variabile AGE

La statistica test di Shapiro-Wilk conferma la normalità anche per le variabili BW, urinarycreatinine, USG e Iron. Il peso medio dei cani (BW) è di 23.5 Kg (sd = 11.48). Il campo di variazione del peso va da un minimo di 4.8 Kg a un massimo di 60.0 Kg (Figura 2).

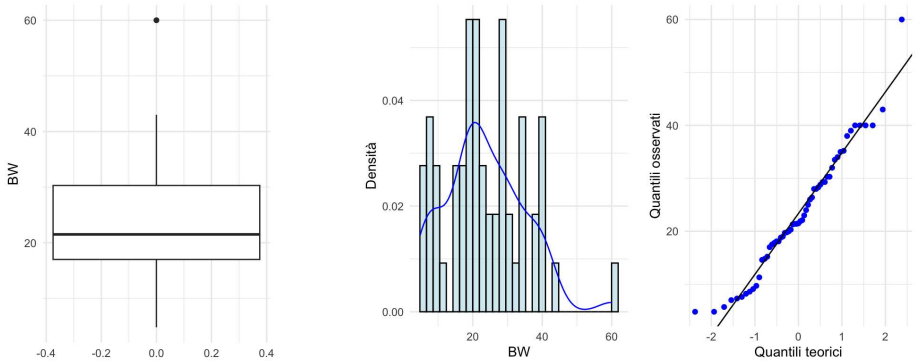

Figura 2. Boxplot, istogramma con densità lisciata e grafico q-q della variabile BW

La quantità di creatinina nelle urine ha media di 204 mg/dL (sd = 109.2), il valore minimo è di 28 mg/dL e il valore massimo è di 545 mg/dL (Figura 3).

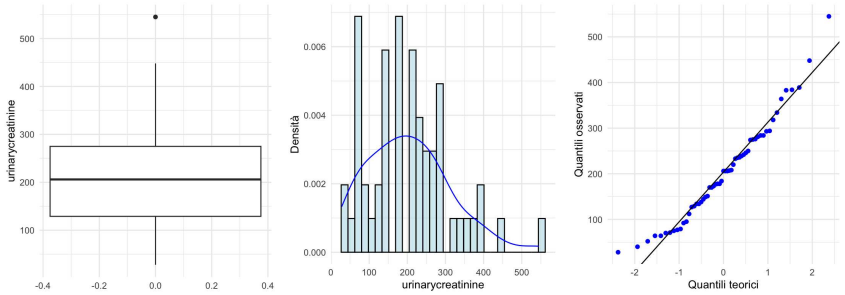

Figura 3. Boxplot, istogramma con densità lisciata e grafico q-q della variabile urinarycreatinine

Il peso specifico (USG, Figura 4) si estende da 1011 g/ml a 1070 g/ml, la media vale 1038 g/ml (sd = 14.06). Il valore mediano è di 1039 g/ml (IQR = 21).

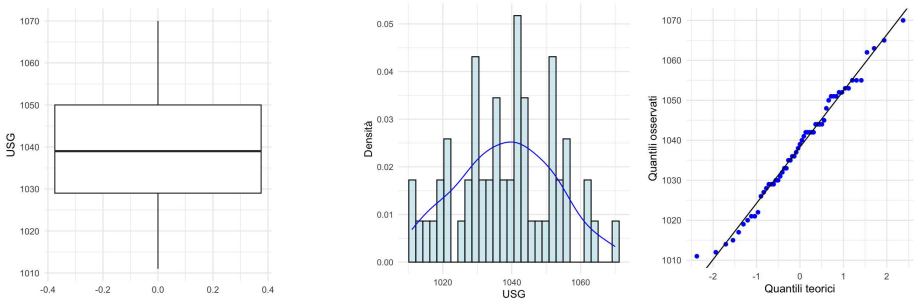

Figura 4. Boxplot, istogramma con densità lisciata e grafico q-q della variabile USG

La quantità di ferro totale (Iron, Figura 5) varia da 22 g/dL a 253 g/dL, la media vale 115 g/dL (sd = 52.1).

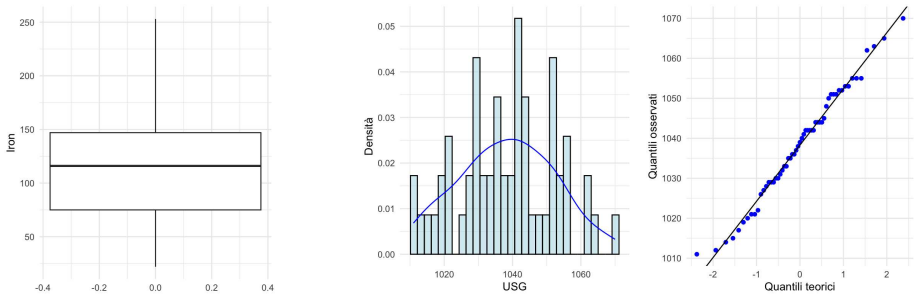

Figura 5. Boxplot, istogramma con funzione di densità e grafico q-q della variabile Iron

Altre variabili presentano un p-value di poco inferiore a 0.05 nel test di Shapiro-Wilk per testare la normalità: queste sono HR, SDMA, PON-1 e TIBC. Iniziando con lo studio della frequenza cardiaca (HR, Figura 6), essa varia da 60 bpm a 190 bpm, la media vale 120 bpm (sd = 27.438). La statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità della variabile HR (SW^{oss} = 0.957, p-value = 0.042).

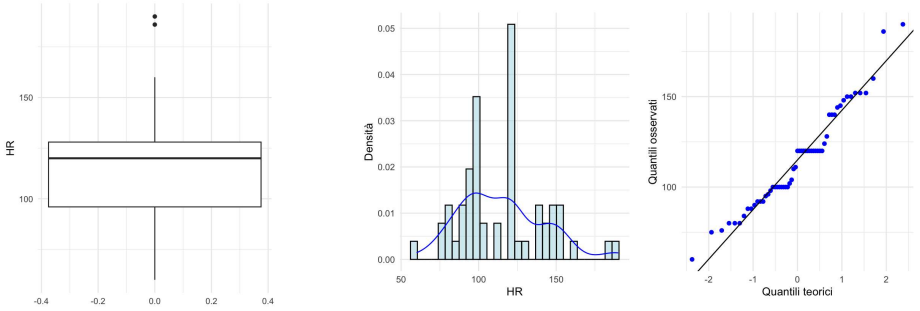

Figura 6. Boxplot, istogramma con densità lisciata e grafico q-q della variabile HR

La variabile SDMA (Figura 7) varia da un valore minimo di 4.6 a un valore massimo di 31.0. Il valore medio è di 13.8 g/dL (sd = 5.73).

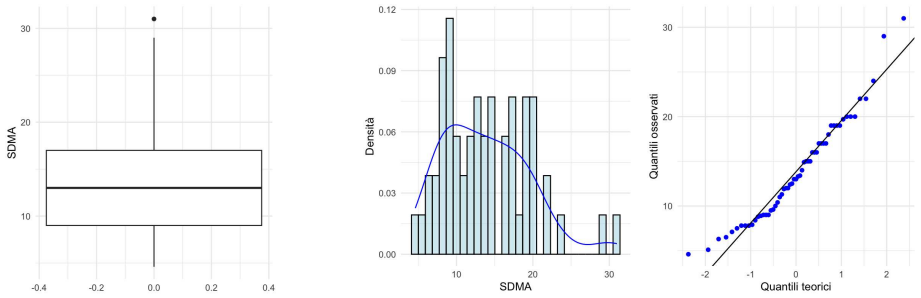

Figura 7. Boxplot, istogramma con densità lisciata e grafico q-q della variabile SDMA

La glicoproteina sintetizzata nel fegato (PON-1, Figura 8) ha un valore minimo di 1.9 IU/L e un valore massimo di 6.21 IU/L. Il valore medio corrisponde a 3.43 IU/L (sd = 0.87) e la mediana è di 3.32 IU/L (IQR = 1.12). La statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità per la variabile PON-1 ($SW^{oss} = 0.953$, p-value = 0.027).

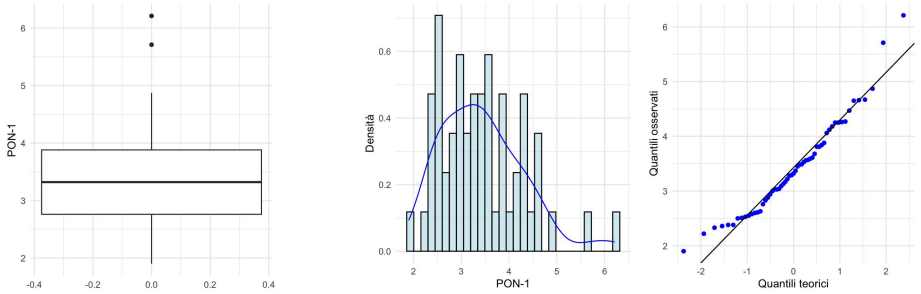

Figura 8. Boxplot, istogramma con densità lisciata e grafico q-q della variabile PON-1

La variabile TIBC (Figura 9) si estende da un minimo di 137 g/dL a un massimo di 253 g/dL, il valore medio corrisponde a 115 (sd = 52.1) e la mediana è pari a 116 (IQR = 72). La statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità per la variabile TIBC ($SW^{oss} = 0.956$, p-value = 0.036).

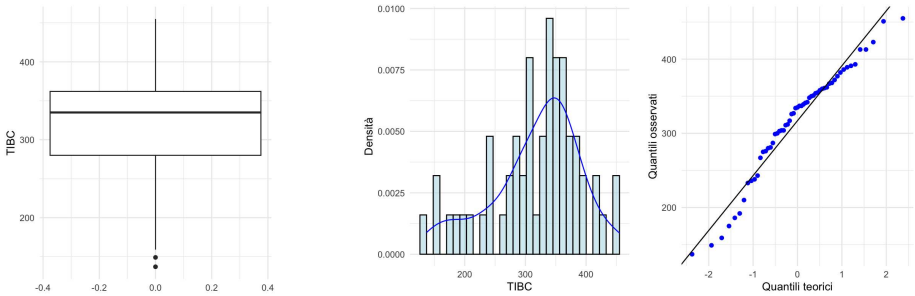

Figura 9. Boxplot, istogramma con densità lisciata e grafico q-q della variabile TIBC

Si riporta infine una sintesi di tutte le variabili quantitative nella Tabella 5.

Variabili	Min.	1 st. Qu.	Median	IQR	Mean	S.D.	3 rd. Qu.	Max.
AGE	10	48	67.2	38.4	71.2	32.87	86.4	153.6
BW	4.8	17	21.5	13.3	23.5	11.48	30.3	60
HR	60	96	120	32	115	27.44	128	190
RR	14	24	30	16	37.8	23.69	40	120
SBP	120	136	152	27	145	24.87	163	230
Antibody	1.48	6.25	23.9	36.82	24.98	18.92	43.07	60.28
urinaryMCP	131.7	357.4	748.5	1420.76	998.9	699.35	1778.2	2066.7
urinarycreatinine	28	129	206	146	204	1092	275	545
uAmCr	0	0.6	3.9	318.2	389.9	769.3	318.8	3700
uMCP	3.87	15.84	35.91	105.7	92.64	131.95	121.6	734.9
CVurine	0	0	2	4	2.77	3.13	4	12
serumMCP	42.3	139.9	280.1	335.9	401.5	396.5	475.9	1626.4
CVserum	0	1	3	5	4.25	4.72	3	28
UPC	0.1	0.2	0.3	1.4	2.5	5.09	1.6	27.7
USG	1011	1029	1039	21	1038	14.06	1050	1070
urea	11	24	33	15	46.3	51.73	39	331
creatinine	0.34	0.83	1.07	0.42	1.24	0.94	1.25	5.77
SDMA	4.6	9	13	8	13.8	5.73	17	31
PON-1	1.9	2.76	3.32	1.12	3.43	0.87	3.88	6.21
CRP	0.01	0.01	0.76	2.49	2.62	4.84	2.5	24.84
Hp	1	30	104	184	150	154	214	640

Variabili	Min.	1 st. Qu.	Median	IQR	Mean	S.D.	3 rd. Qu.	Max.
Ft	133	230	480	751	771	893.3	981	4860
Iron	22	75	116	72	115	52.1	147	253
TIBC	137	280	335	82	317	74	362	455

Tabella 5: Statistiche di sintesi per variabili quantitative: minimo e massimo, primo e terzo quartile, indici di centralità (media e mediana) con i corrispettivi indici di variabilità (deviazione standard e scarto interquartilico).

Si osserva asimmetria nelle distribuzioni per le variabili che non hanno verificata la normalità. Si è deciso di approfondire in particolare quelle con asimmetria più pronunciata.

La variabile amilasi di creatinina urinaria (uAmCr, Figura 10) presenta una forte asimmetria positiva, con una media di 389.9 (sd = 769.3) circa 100 volte maggiore della mediana pari a 3.9 U.I./L (IQR = 318.2). Il suo valore minimo è pari a 0 U.I./L e il suo valore massimo è 3700 U.I./L.

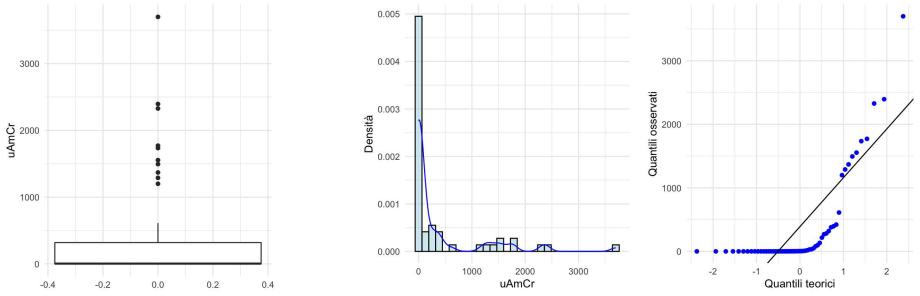

Figura 10. Boxplot, istogramma con densità lisciata e grafico q-q della variabile uAmCr

La variabile uMCP (Figura 11) varia da un minimo di 3.87 a un massimo di 734.9. La media è pari a 92.64 (sd = 131.95), circa 2.6 volte la mediana, che vale 35.91 (IQR = 105.7).

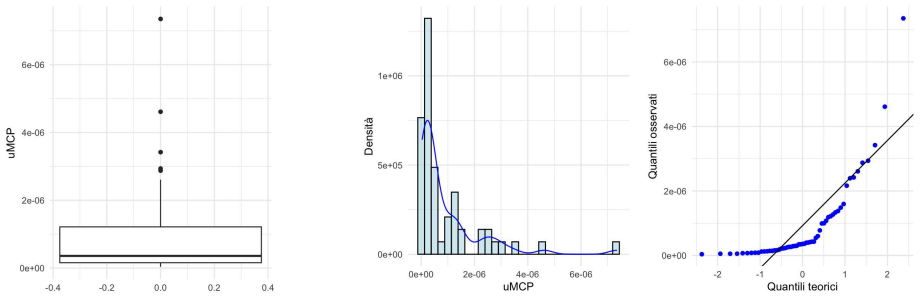

Figura 11. Boxplot, istogramma con densità lisciata e grafico q-q della variabile uMCP

La variabile serumMCP (Figura 12) ha un valore minimo di 42.3 e un valore massimo di 1626.4. La media è 401.5 (sd = 396.5) che è circa la metà della mediana che corrisponde essere circa la metà della mediana di 280.1 (IQR = 335.9).

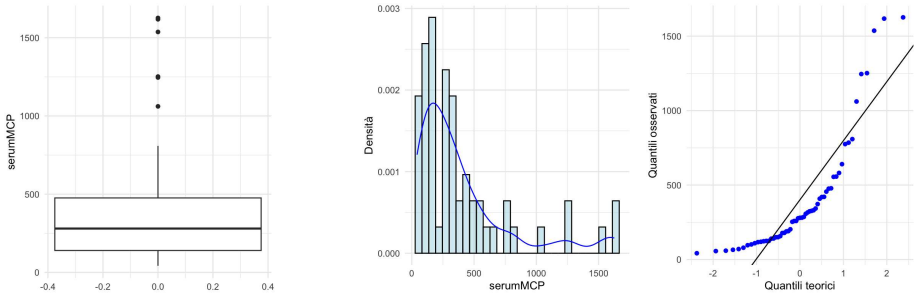

Figura 12. Boxplot, istogramma con densità lisciata e grafico q-q della variabile serumMCP

La variabile che indica il rapporto proteine urinarie-creatinina urinaria (UPC, Figura 13) varia da un valore minimo di 0.1 a un valore massimo di 27.7. Il valore medio vale 2.5 (sd = 5.095) e la mediana è 0.3 (IQR = 1.4).

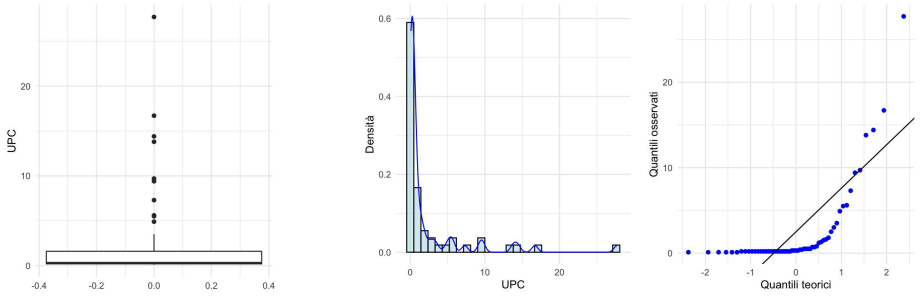

Figura 13. Boxplot, istogramma con densità lisciata e grafico q-q della variabile UPC

La variabile corrispondente alla proteina C reattiva (CRP, Figura 14) ha un valore minimo di 0.01 mg/dL e un valore massimo di 24.84. Il valore medio è di 2.62 (sd = 4.84) e la mediana è 0.76 (IQR = 2.49).

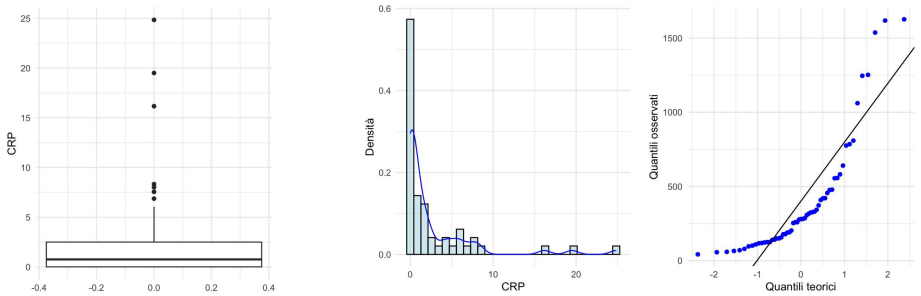

Figura 14. Boxplot, istogramma con densità lisciata e grafico q-q della variabile CRP

La variabile che indica la quantità di ferro totale (Ft, Figura 15) ha un valore minimo di 133 di mg/grCrea e un valore massimo di 981 mg/grCrea. Il valore medio è di 771 mg/grCrea (sd = 893.3) e la mediană è 771 (IQR = 751).

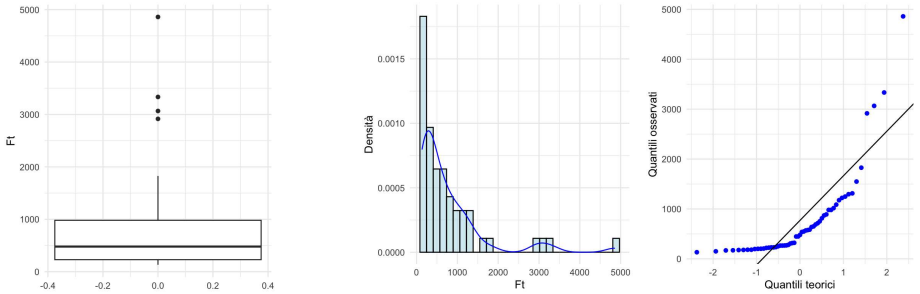

Figura 15. Boxplot, istogramma con densità lisciata e grafico q-q della variabile Ft

Si riporta infine una tabella riassuntiva con i risultati del test di Shapiro-Wilk e i rispettivi p-value per le variabili quantitative (Tabella 6).

	SW^{oss}	oss
AGE	0.97	0.32
BW	0.96	0.09
HR	0.95	0.04
RR	0.67	<0.01
SBP	0.88	<0.01
Antibody	0.91	<0.01
urinaryMCP	0.87	<0.01
urinarycreatinine	0.96	0.07
uAmCr	0.58	<0.01
uMCP	0.66	<0.01
CVurine	0.83	<0.01
serumMCP	0.76	<0.01
CVserum	0.71	<0.01
UPC	0.53	<0.01
USG	0.98	0.71
urea	0.52	<0.01
creatinine	0.58	<0.01
SDMA	0.95	0.027
PON-1	0.95	0.027
CRP	0.59	<0.01
Hp	0.85	<0.01
Ft	0.66	<0.01
Iron	0.976	0.31
TIBC	0.95	0.04

Tabella 6: Valore osservato della statistica test di Shapiro-Wilk e relativo p-value.

2.2 Analisi esplorative bivariate

In questo paragrafo vengono studiate prima di tutto le relazioni a coppie tra le variabili qualitative[11], successivamente per le variabili quantitative si è testata la normalità[12] e per le variabili quantitative dove è verificata la normalità anche l'omoschedasticità[13], per poi passare con la verifica dell'uguaglianza delle medie nel caso di normalità confermata all'interno dei gruppi[14]. Per le variabili non normali si è studiata l'uguaglianza in distribuzione[15]. Infine si è calcolata la correlazione di Spearman tra variabili quantitative.

Come risulta dalle Tabelle 7,8 e 9 le tre variabili di classificazione in più stadi LeisvetStaging ($\chi^{2oss} = 57$, p-value = <0.01), IRISStaging ($\chi^{2oss} = 57$, p-value = <0.01) e IRISStagingroup ($\chi^{2oss} = 57$, p-value = <0.01) sono ovviamente associate con la variabile di classificazione per sani/malati (group).

group/LeisvetStaging	0	IIa
IIb	III	IV
0	19	0
0	0	0
1	0	12
7	10	9

Tabella 7

group/IRISStaging	0	I	II	III	IV
0	19	0	0	0	0
1	0	31	4	1	2

Tabella 8

group/IRISstagingroup	0	1	2
0	19	0	0
1	0	31	7

Tabella 9

Tabella 7-9: Tabelle di contingenza delle variabili group/Clinicalsigs, group/LeisvetStaging, group/IRISStaging e group/IRISstaginggroup.

Anche altre variabili risultano essere significativamente associate, come il sesso

[11] Applicazione del test di indipendenza χ^2 di Pearson.

[12] Applicazione del Shapiro-test per ciascun gruppo delle variabili quantitative.

[13] Applicazione del test F tra due popolazioni normali.

[14] Applicazione del test t di Student se si confrontano due soli gruppi omoschedastici, test t di Welch se non lo sono, nel caso le variabili avessero più gruppi, sempre sotto condizioni di normalità, si applica il test ANOVA.

[15] Test di Mann-Whitney nel caso di variabile con due categorie di classificazione o Kruskal-Wallis nel caso della suddivisione della variabile in più gruppi utilizzando la correzione di Holm per valutare le differenze significative tra le coppie di gruppi nell'analisi post-hoc

con lo stato sessuale del cane ($\chi^{2oss} = 6.5$, p-value = 0.01, Tabella 10). Il 43.8% è rappresentato da cani di sesso femminile: il 72% di questi è castrato, mentre il restante 18% è intero, mentre per quanto riguarda i cani di sesso maschile (56.2%) tra questi il 34.4% è castrato, mentre il restante 65.6% è intero.

SEX/SEXUALSTATUS	C	I
F	18	7
M	11	21

Tabella 10: Tabella di contingenza delle variabili SEX/SEXUALSTATUS.

Nella Tabella 11 si riporta il Test di Pearson con rispettivo p-value delle coppie di variabili qualitative.

	χ^{2oss}	α_{oss}
group/BREED	0.23	0.6
group/SEX	0.22	0.6
group/SEXUALSTATUS	0.0088	0.9
group/LeisvetStaging	57	<0.01
group/IRISStaging	57	<0.01
group/IRISStagingroup	57	<0.01
BREED/SEX	0.0067	0.9
BREED/SEXUALSTATUS	1.6	0.2
BREED/LeisvetStaging	1.2	0.9
BREED/IRISStaging	3.8	0.4
BREED/IRISStagingroup	3.4	0.2
SEX/SEXUALSTATUS	6.5	0.01
SEX/LeisvetStaging	8.9	0.06
SEX/IRISStaging	2.6	0.3
SEX/IRISStagingroup	4.3	0.4
SEXUALSTATUS/LeisvetStaging	6.7	0.2
SEXUALSTATUS/IRISStaging	0.21	0.9
SEXUALSTATUS/IRISStagingroup	1.1	0.9

Tabella 11. Test χ^2 di Pearson per coppie di variabili qualitative e p-value.

Successivamente si è studiata l'uguaglianza delle medie o distribuzioni delle va-

riabili quantitative all'interno dei vari gruppi. Iniziando con le variabili dove la normalità e l'omoschedasticità sono verificate, alcune di queste hanno una media all'interno dei gruppi significativamente differente. Dalla Figura 16 si evince che la variabile BW presenta differenza in media rispetto alla variabile BREED (t^{oss}=4.273, p-value <0.01).

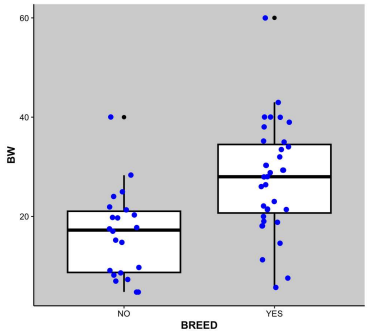

Figura 16. Boxplot della variabile BW in base alla suddivisione di BREED

La variabile urinarycreatinine (Figura 17) invece presenta differenza in media rispetto alle variabili group (t^{oss} = 5.7085, p-value <0.01), LeisvetStaging (F^{oss}=11.78, p-value <0.01), IRISStaging (F^{oss} = 9.521, p-value <0.01) e IRISstaginggroup (F^{oss} = 19.39, p-value <0.01). Presenta differenza in media tra il gruppo 0 con i gruppi III, III e IV per la variabile LeisvetStaging, tra il gruppo 0 ed il gruppo I per la variabile IRISStaging ed infine si presenta differenza in media tra il gruppo 0 con i gruppi 1 e 2 per la variabile IRISstaginggroup (Tabella 12).

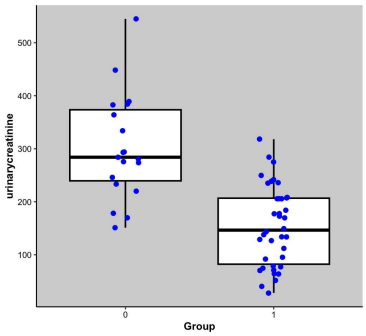

Figura 17. Boxplot della variabile urinarycreatinine in base alla suddivisione di group.

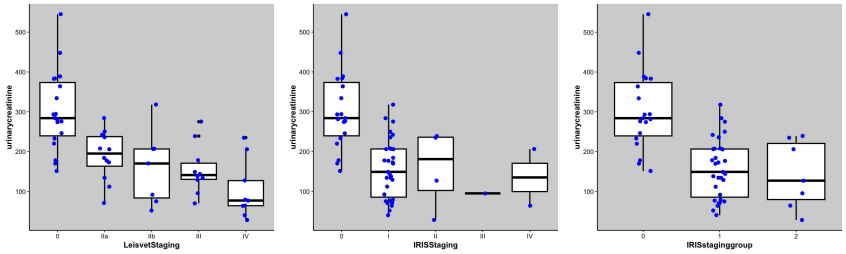

Figura 18. Boxplot della variabile urinarycreatinine in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstaginggroup

variabili	gruppi	p-value
LeisvetStaging	0-IIa	0.018
	0-III	<0.01
	0-IV	<0.01
IRISStaging	0-I	<0.01
IRISstaginggroup	0-1	<0.01
	0-2	<0.01

Tabella 12. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile urinarycreatinine all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

Le stesse differenze in media per gli stessi gruppi le presenta anche la variabile USG, quindi per la variabile group ($t^{oss} = 4.1339$, p-value <0.01, Figura 19), LeisvetStaging ($F^{oss} = 5.277$, p-value <0.01), IRISStaging ($F^{oss} = 5.261$, p-value <0.01) e IRISstaginggroup ($F^{oss} = 9.963$, p-value <0.01), come emerge dalla Figura 20. Presenta differenza in media tra il gruppo 0 e il gruppo IV per la variabile LeisvetStaging e tra il gruppo 0 con i gruppi 1 e 2 per la variabile

IRISStagingroup (Tabella 13).

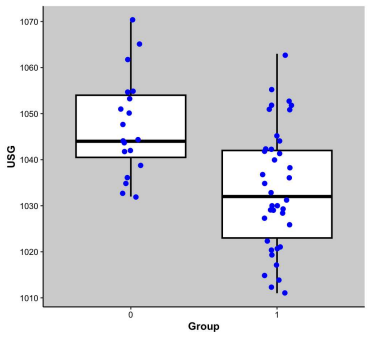

Figura 19. Boxplot della variabile USG in base alla suddivisione di group

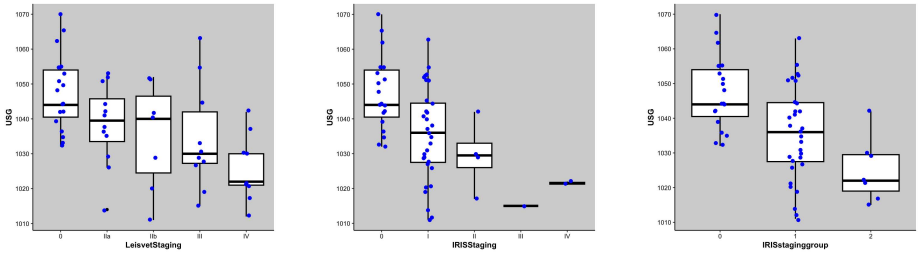

Figura 20. Boxplot della variabile USG in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstaginggroup

variabili	gruppi	p-value
LeisvetStaging	0-IV	<0.01
IRISstaginggroup	0-2	<0.01
	1-2	<0.01

Tabella 13. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile USG all'interno dei gruppi di LeisvetStaging e IRISstaginggroup

Infine, la variabile Iron presenta differenze in media per le variabili group (t^{oss} = 4.4465, p-value <0.01), LeisvetStaging (F_{oss} = 6.59, p-value <0.01, Figura 21), IRISStaging (F^{oss} = 4.25, p-value <0.01) e IRISstaginggroup (F^{oss} = 7.551, p-value <0.01), come si conferma dalla Figura 22. Presenta differenze in media tra il gruppo 0 con i gruppi IIb, III e IV per la variabile LeisvetStaging, tra il gruppo 0 e I della variabile IRISStaging e tra il gruppo 0 con i gruppi 1 e 2 per la variabile IRISstaginggroup (Tabella 14).

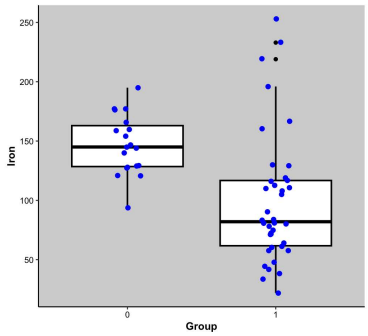

Figura 21. Boxplot della variabile Iron in base alla suddivisione di group

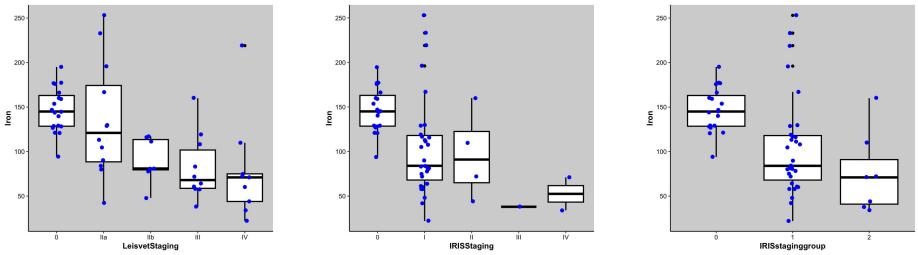

Figura 22. Boxplot della variabile Iron in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstaginggroup

variabili	gruppi	p-value
LeisvetStaging	0-IIb	<0.01
	0-III	<0.01
	0-IV	0.011
IRISStaging	0-I	<0.01
IRISstaginggroup	0-1	<0.01
	0-2	<0.01

Tabella 14. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile Iron all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

Per quanto riguarda le restanti variabili quantitative dove non è accettata l'assunzione di normalità o nel caso in cui è accettata, ma non lo è l'omoschedasticità, è stata studiata l'uguaglianza in distribuzione con il test di Mann-Whitney nel caso di variabile con due categorie di classificazione e Kruskal-Wallis nel caso della suddivisione della variabile in più gruppi (con la correzione di Holm).

La variabile RR non rispetta l'uguaglianza in distribuzione per i gruppi group (MW^{oss} = 237.5, p-value = 0.035), come si evince dalla Figura 23.

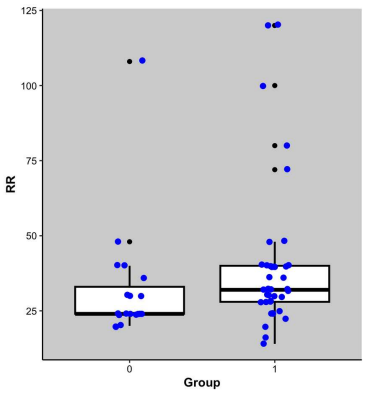

Figura 23. Boxplot della variabile RR in base alla suddivisione di group

La variabile urinaryMCP presenta differenze in distribuzione per le variabili group (MW^{oss} = 149, p-value <0.01, Figura 24), LeisvetStaging (KW^{oss} = 30.423, p-value = <0.01), IRISStaging (KW^{oss} = 16.188, p-value= <0.01)

e IRISstaginggroup (KW^{oss} = 15.644, p-value = <0.01), come si evince dalla Figura 25. La variabile presenta differenze in distribuzione tra il gruppo 0 con i gruppi III e IV per la variabile LeisvetStaging, tra il gruppo 0 e I per la variabile IRISStaging e con il gruppo 0 con i gruppi 1 e 2 della variabile IRISstaginggroup (Tabella 15).

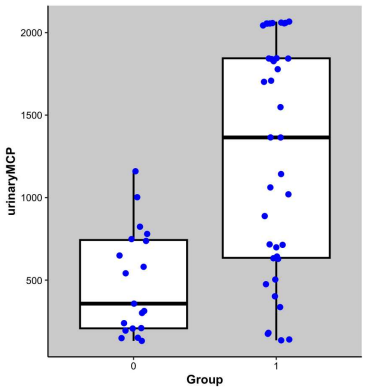

Figura 24. Boxplot della variabile urinaryMCP in base alla suddivisione di group

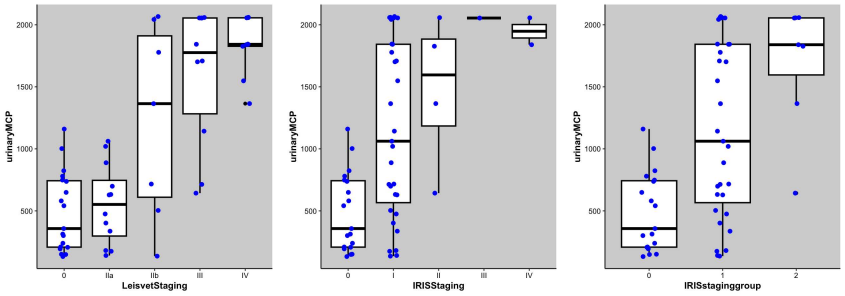

Figura 25. Boxplot della variabile urinaryMCP in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstaginggroup

Le stesse differenze in distribuzione sono state osservate anche per le variabili Ft (Figure 26-27), uAmCr (Figure 28-29), uMCP (Figure 30-31), serumMCP

variabili	gruppi	p-value
LeisvetStaging	0-III	<0.01
	0-IV	<0.01
	IIa-III	<0.01
	IIa-IV	<0.01
IRISStaging	0-I	0.019
IRISstaginggroup	0-1	0.037
	0-2	<0.01

Tabella 15. P-value significativi dell' analisi post-hoc con la correlazione Holm per la variabile urinaryMCP all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

(Figure 32-33), SDMA (Figure 34-35), CRP (Figure 36-37), Hp (Figure 38-39) e TIBC (Figure 40-41).

La variabile Ft presenta differenze in distribuzione per la variabile group (MW^{oss} = 65.5, p-value <0.01, Figura 26) e per le variabili LeisvetStaging (KW^{oss} = 29.234, p-value <0.01), IRISStaging (KW^{oss} = 25.54, p-value <0.01) e IRISstaginggroup (KW^{oss} = 25.2, p-value <0.01), come si evince dalla Figura 27. La variabile Ft presenta differenza in distribuzione tra il gruppo 0 con i gruppi IIb, III, IV della variabile LeisvetStaging, tra il gruppo 0 con i gruppi I e II della variabile IRISStaging e tra il gruppo 0 con i gruppi 1 e 2 della variabile IRISstaginggroup (Tabella 16).

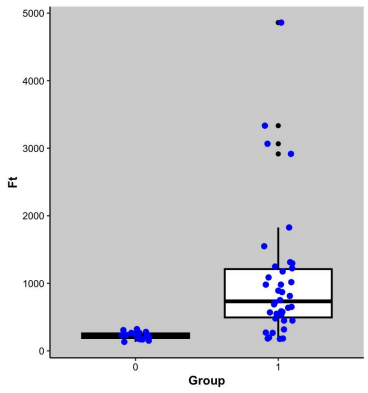

Figura 26. Boxplot della variabile Ft in base alla suddivisione di group

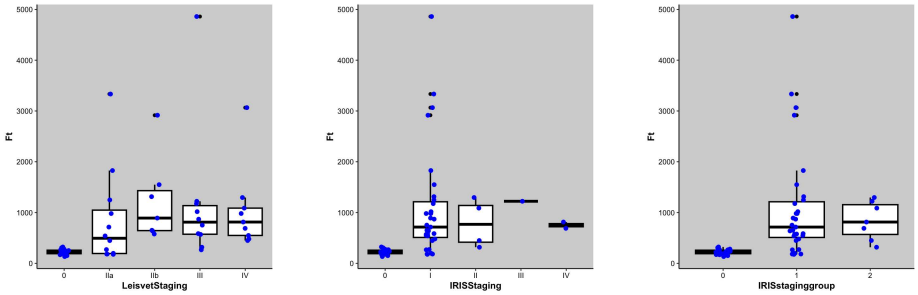

Figura 27. Boxplot della variabile Ft in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-IIb	<0.01
	0-III	<0.01
	0-IV	<0.01
IRISStaging	0-I	<0.01
	0-II	0.028
IRISstaginggroup	0-1	<0.01
	0-2	<0.01

Tabella 16. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile Ft all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile uAmCr presenta differenze in distribuzione per la variabile group ($MW^{oss} = 65.5$, p-value <0.01, Figura 28) e per le variabili LeisvetStaging ($KW^{oss} = 29.234$, p-value <0.01), IRISStaging ($KW^{oss} = 25.54$, p-value <0.01) e IRISstaginggroup ($KW^{oss} = 25.2$, p-value <0.01), come si evince dalla Figura 29. Presenta differenza in distribuzione tra il gruppo 0 e IIa, tra il gruppo III con i gruppi 0, IIa, IIb e tra il gruppo IV con i gruppi 0, IIa, IIb e III per la variabile LeisvetStaging, tra il gruppo 0 e I per la variabile IRISStaging e tra il gruppo 0 con i gruppi 1 e 2 per la variabile IRISstaginggroup (Tabella 17).

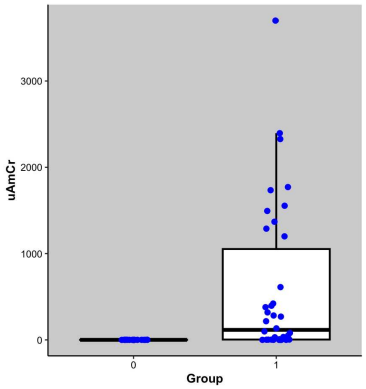

Figura 28. Boxplot della variabile uAmCr in base alla suddivisione di group

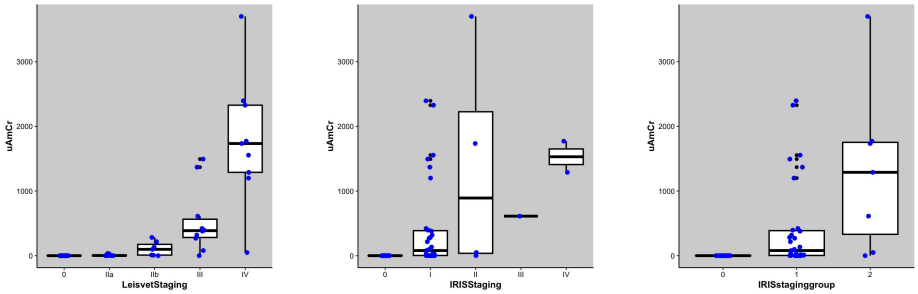

Figura 29. Boxplot della variabile uAmCr in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-IIa	<0.01
	0-III	<0.01
	IIa-III	0.037
	IIb-III	<0.01
	0-IV	<0.01
	IIa-IV	0.011
IRISStaging	0-I	<0.01
IRISstaginggroup	0-1	<0.01
	0-2	<0.01

Tabella 17. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile uAmCr all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile uMCP presenta differenze in distribuzione per la variabile group ($MW^{oss} = 90$, p-value <0.01, Figura 30) e per le variabili LeisvetStaging ($KW^{oss} = 34.912$, p-value <0.01), IRISStaging ($KW^{oss} = 22.888$, p-value <0.01) e IRISstaginggroup ($KW^{oss} = 22.515$, p-value <0.01), come si evince dalla Figura 31. Presenta differenze in distribuzione tra il gruppo 0 con i gruppi IIb, III e IV, tra il gruppo III con i gruppi III e IV per la variabile LeisvetStaging, tra il gruppo 0 con i gruppi I e II per la variabile IRISSTaging ed infine tra il gruppo 0 con i gruppi 1 e 2 per la variabile IRISstaginggroup (Tabella 18).

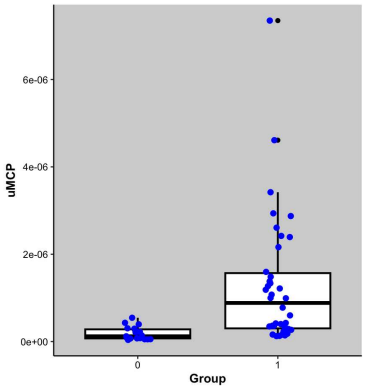

Figura 30. Boxplot della variabile uMCP in base alla suddivisione di group

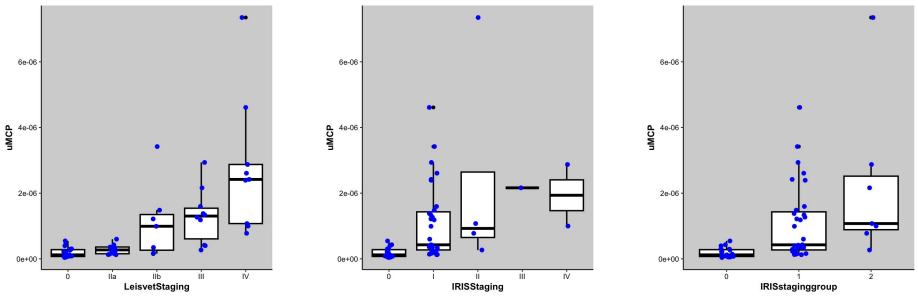

Figura 31. Boxplot della variabile uMCP in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-IIb	0.027
	0-III	<0.01
	0-IV	<0.01
	IIa-III	<0.01
	IIa-IV	<0.01
IRISStaging	0-I	<0.01
	0-II	0.037
IRISstaginggroup	0-1	<0.01
	0-2	<0.01

Tabella 18. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile uMCP all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile serumMCP presenta differenze in distribuzione per la variabile group (MW^{oss} = 62, p-value <0.01, Figura 32) e per le variabili LeisvetStaging (KW^{oss} = 38.568, p-value <0.01), IRISStaging (KW^{oss} = 29.608, p-value <0.01) e IRISstaginggroup (KW^{oss} = 28.551, p-value <0.01), come si evince dalla Figura 33. Presenta differenza in distribuzione tra il gruppo 0 con i gruppi IIb, III e IV, tra il gruppo IIa e il gruppo IV per la variabile LeisvetStaging, tra il gruppo 0 con i gruppi I e II per la variabile IRISStaging e tra il gruppo 0 con i gruppi 1 e 2 e tra il gruppo 1 e 2 per la variabile IRISstaginggroup (Tabella 19).

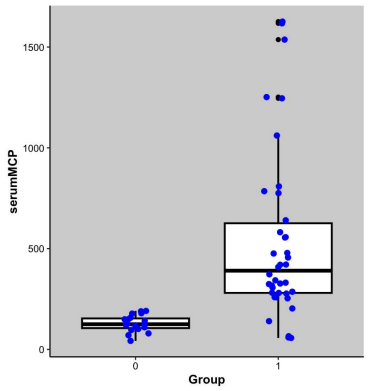

Figura 32. Boxplot della variabile serumMCP in base alla suddivisione di group

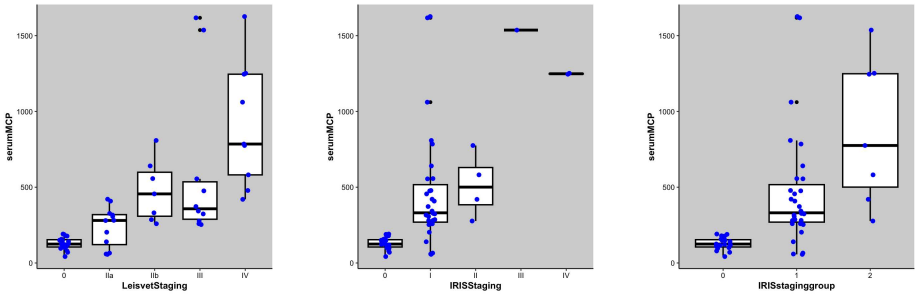

Figura 33. Boxplot della variabile serumMCP in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-IIb	<0.01
	0-III	<0.01
	0-IV	<0.01
	IIa-IV	<0.01
IRISStaging	0-I	<0.01
	0-II	<0.01
IRISstaginggroup	0-1	<0.01
	0-2	<0.01
	1-2	<0.01

Tabella 19. P-value significativi dell' analisi post-hoc con la correzione Holm per la variabile serumMCP all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile SDMA presenta differenze in distribuzione per la variabile group ($MW^{oss} = 62$, p-value <0.01, Figura 34) e per le variabili LeisvetStaging ($KW^{oss} = 38.568$, p-value <0.01), IRISStaging ($KW^{oss} = 29.608$, p-value <0.01) e IRISstaginggroup ($KW^{oss} = 28.551$, p-value <0.01), come si evince dalla Figura 35. Presenta differenze in distribuzione tra il gruppo IIb e 0 ed il gruppo IV con i gruppi 0 e IIb per la variabile LeisvetStaging, tra il gruppo 0 con i gruppi I e II per la variabile IRISStaging ed infine tra tutti i gruppi per la variabile IRISstaginggroup (Tabella 20).

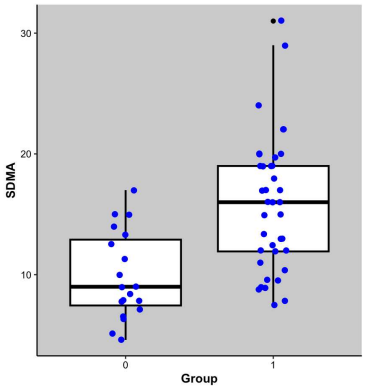

Figura 34. Boxplot della variabile SDMA in base alla suddivisione di group

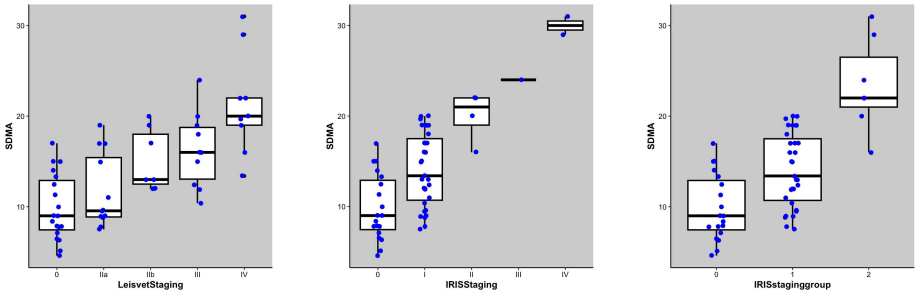

Figura 35. Boxplot della variabile SDMA in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-III	<0.01
	0-IV	<0.01
	IIa-IV	<0.01
IRISStaging	0-I	0.012
IRISstaginggroup	0-1	<0.01
	0-2	<0.01
	1-2	<0.01

Tabella 20. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile SDMA all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile CRP presenta differenze in distribuzione per la variabile group (MW^{oss} = 57.5, p-value <0.01, Figura 36) e per le variabili LeisvetStaging (KW^{oss} = 33.489, p-value <0.01), IRISStaging (KW^{oss} = 29.054, p-value <0.01) e IRISstaginggroup (KW^{oss} = 28.317, p-value <0.01), come si evince dalla Figura 37. Presenta differenze in distribuzione tra il gruppo 0 con i gruppi IIa, IIb, III e IV e tra il gruppo IIa e IV per la variabile LeisvetStaging, tra il gruppo 0 e I per la variabile IRISStaging ed infine tra il gruppo 0 con i gruppi 1 e 2 per la variabile IRISstaginggroup (Tabella 21).

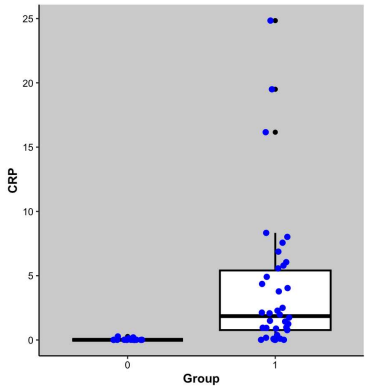

Figura 36. Boxplot della variabile CRP in base alla suddivisione di group

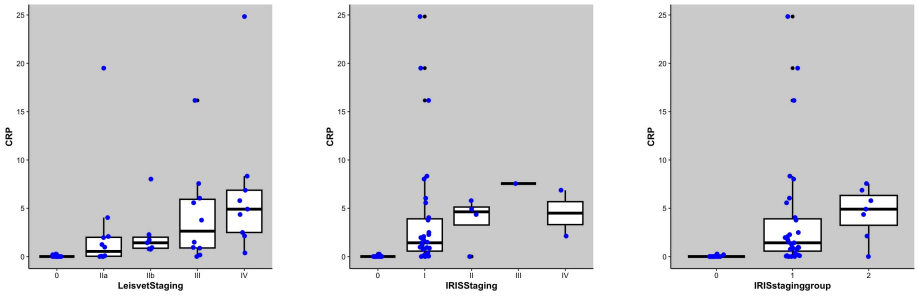

Figura 37. Boxplot della variabile CRP in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-IIa	0.012
	0-IIb	<0.01
	0-III	<0.01
	0-IV	<0.01
	IIa-IV	<0.01
IRISStaging	0-I	<0.01
IRISstaginggroup	0-1	<0.01
	0-2	<0.01

Tabella 21. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile CRP all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile Hp presenta differenze in distribuzione per la variabile group (MW^{oss} = 92, p-value <0.01, Figura 38) e per le variabili LeisvetStaging (KW^{oss} = 21.607, p-value <0.01), IRISStaging (KW^{oss} = 23.063 p-value <0.01) e IRISstaginggroup (KW^{oss} = 22.713, p-value <0.01), come si evince dalla Figura 39. Presenta differenze in distribuzione tra il gruppo 0 con i gruppi IIa, IIb, III e IV per la variabile LeisvetStaging, tra il gruppo 0 e I per la variabile IRISStaging ed infine tra il gruppo 0 con il gruppo 1 per la variabile IRISstaginggroup (Tabella 22).

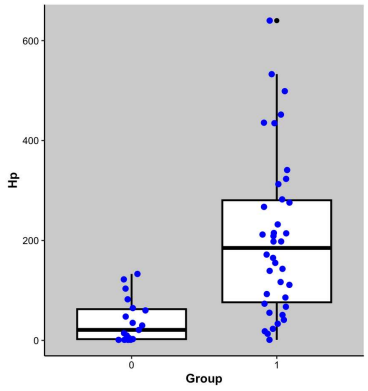

Figura 38. Boxplot della variabile Hp in base alla suddivisione di group

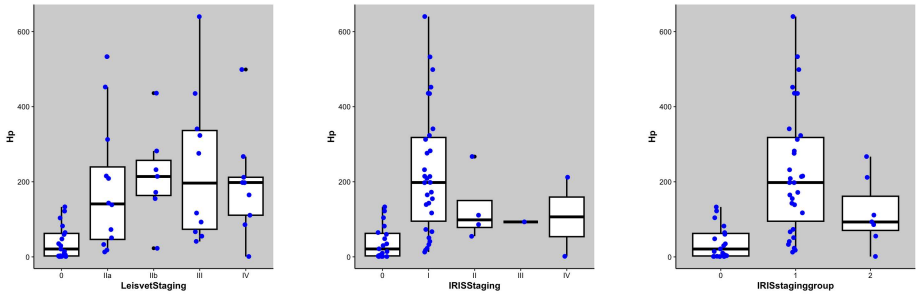

Figura 39. Boxplot della variabile Hp in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-IIa	0.012
	0-IIb	<0.01
	0-III	<0.01
	0-IV	0.015
IRISStaging	0-I	<0.01
IRISstaginggroup	0-1	<0.01

Tabella 22. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile Hp all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

La variabile TIBC presenta differenze in distribuzione per la variabile group ($MW^{oss} = 561.5$, p-value <0.01, Figura 40) e per le variabili LeisvetStaging ($KW^{oss} = 21.607$, p-value <0.01), IRISStaging ($KW^{oss} = 16.11$, p-value <0.01) e IRISstaginggroup ($KW^{oss} = 14.553$, p-value <0.01), come si evince dalla Figura 41. Presenta differenze in distribuzione tra il gruppo III e 0 e tra il gruppo IV con i gruppi 0, IIa e IIb per la variabile LeisvetStaging, tra il gruppo 0 e I per la variabile IRISStaging e tra tutti i gruppi della variabile IRISstaginggroup (Tabella 23).

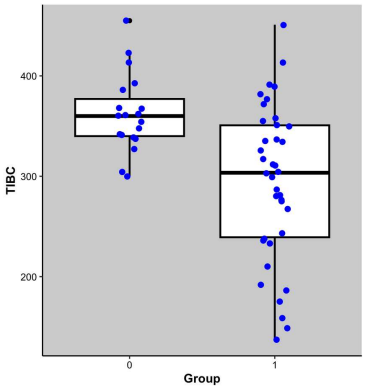

Figura 40. Boxplot della variabile TIBC in base alla suddivisione di group

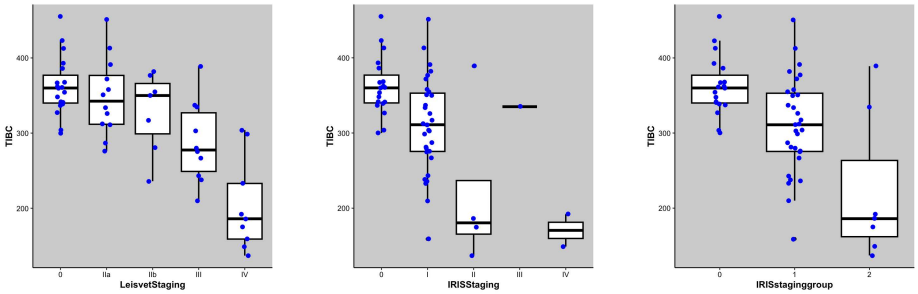

Figura 41. Boxplot della variabile TIBC in base alla suddivisione di LeisvetStaging, IRISStaging e IRISstagingroup

variabili	gruppi	p-value
LeisvetStaging	0-III	<0.01
	0-IV	<0.01
	IIa-IV	0.011
	IIb-IV	<0.01
IRISStaging	0-I	<0.01
IRISstagginggroup	0-1	<0.01
	0-2	<0.01
	1-2	0.029

Tabella 23. P-value significativi dell' analisi post-hoc con la correzione Holm per la variabile TIBC all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstagginggroup

La variabile CVurine presenta differenza in distribuzione per i gruppi della variabile LeisvetStaging ($KW^{oss} = 19.317$, p-value <0.01, Figura 42), la

variabile urea per i gruppi LeisvetStaging (KW^{oss} = 27.261, p-value <0.01), IRISStaging (KW^{oss} = 13.331, p-value <0.01), IRISstaginggroup (KW^{oss} = 12.539, p-value <0.01) e SEX (MW^{oss} = 607, p-value <0.01) come si evince dalla Figura 43 e anche la variabile creatinine per LeisvetStaging (KW^{oss} = 16.457, p-value <0.01), IRISStaging (KW^{oss} = 32.926, p-value <0.01), IRISstaginggroup (KW^{oss} = 32.833, p-value <0.01) e SEX (MW^{oss} = 528, p-value = 0.0403), come si evince dalla Figura 44.

Si riportano di seguito le Figure 42-44 per i boxplot e le Tabelle 24-26 per i rispettivi p-value .
La variabile CVurine presenta differenze in distribuzione tra il gruppo IV con i gruppi IIa e IIb della variabile LeisvetStaging (Tabella 24).

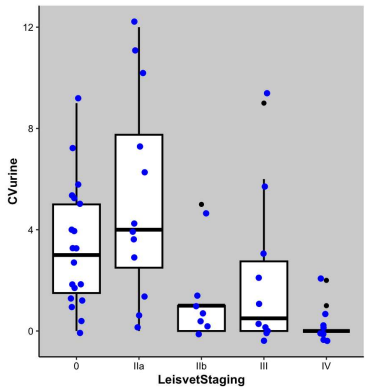

Figura 42 . Boxplot della variabile CVurine in base alla suddivisione di LeisvetStaging

variabili	gruppi	p-value
LeisvetStaging	IIa-IV	<0.01
	IIb-IV	<0.01

Tabella 24. P-value significativi dell' analisi post-hoc con la correzione Holm per la variabile CVurine all'interno dei gruppi di LeisvetStaging.

La variabile urea presenta differenze in distribuzione tra il gruppo IV con i gruppi 0, IIa e IIb per la variabile LeisvetStaging e tra il gruppo 2 con i gruppi 0 e 1 per la variabile IRISstaginggroup (Tabella 25).

variabili	gruppi	p-value
LeisvetStaging	0-IV	0.045
	IIa-IV	0.044
	IIb-IV	0.045
IRISstaginggroup	0-2	<0.01
	1-2	<0.01

Tabella 25. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile urea all'interno dei gruppi di LeisvetStaging e IRISstaginggroup

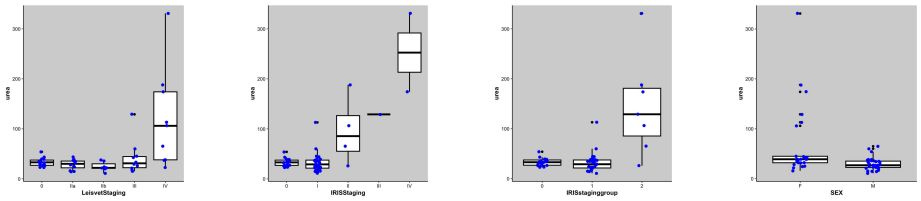

Figura 43. Boxplot della variabile urea in base alla suddivisione di LeisvetStaging, IRISStaging, IRISstagingroup e SEX

La variabile creatinine presenta differenze in distribuzione tra il gruppo 0 e IIa per la variabile LeisvetStaging, tra il gruppo 0 e I e tra il gruppo II con i gruppi 0 e I per la variabile IRISStaging ed infine tra tutti i gruppi per la variabile IRISstaginggroup (Tabella 26).

variabili	gruppi	p-value
LeisvetStaging	0-IIa	0.022
IRISStaging	0-I	<0.01
	0-II	0.024
	I-II	0.013
IRISstaginggroup	0-1	<0.01
	0-2	<0.01
	1-2	<0.01

Tabella 26. P-value significativi dell'analisi post-hoc con la correzione Holm per la variabile creatinine all'interno dei gruppi di LeisvetStaging, IRISStaging e IRISstaginggroup

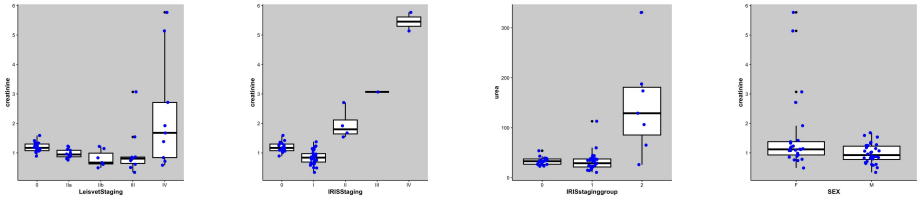

Figura 44. Boxplot della variabile creatinine in base alla suddivisione di LeisvetStaging, IRISStaging, IRISstagingroup e SEX

Si è studiata la correlazione di Spearman tra le variabili quantitative. Si osserva che l'unica variabile a non avere una correlazione significativa con nessuna delle altre variabili è la variabile AGE. Dalla Figura 45 si può osservare come si correlano le variabili tra di loro.

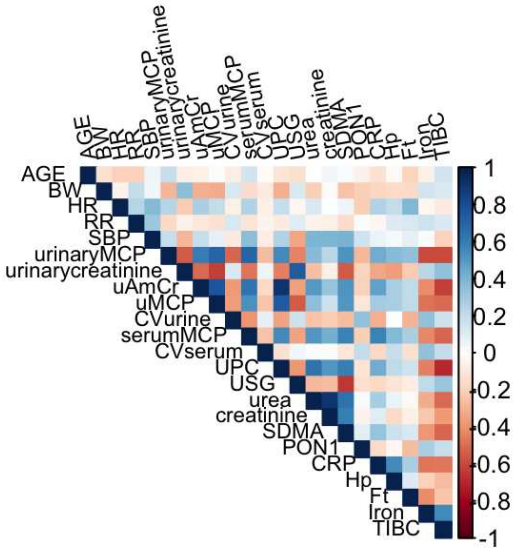

Figura 45. Correlogramma delle variabili quantitative

3 Capitolo 3: Modello logistico penalizzato

3.1 I modelli lineari generalizzati

In molte applicazioni è di interesse studiare la relazione tra una variabile risposta e altre variabili definendo un modello di regressione.

La classe dei modelli lineari generalizzati (GLM) [5], è una estensione del modello di regressione lineare normale per trattare risposte con distribuzione diversa dalla normale [6].

Sia $y = (y_1,...,y_n)$ l'osservazione della variabile risposta relativa a n unità statistiche. Si assume che y_i sia realizzazione di una variabile aleatoria (Y_i) con distribuzione appartenente alla famiglia di dispersione esponenziale univariata, con funzione di densità

$$p(y_i;\theta_i,\phi) = exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}, \tag{1}$$

con $y_i \in S \subseteq \Re, \theta_i \in \Theta \subseteq \Re, a_i(\phi) > 0$, $i = 1,...,n$. Il parametro θ_i è detto parametro naturale, mentre ϕ è detto parametro di dispersione, $b(\theta_i)$ e $c(\theta_i)$ sono funzioni note la cui scelta individua una particolare distribuzione e il dominio di Y_i non dipende da θ_i e ϕ_i.
I primi due momenti di Y_i sono rispettivamente

$$E(Y_i) = \mu_i = b'(\theta_i) = \mu(\theta_i) \tag{2}$$

e

$$Var(Y_i) = a_i(\phi)b''(\theta_i)|_{\theta_i=\theta(\mu_i)} = a_i(\phi)\upsilon(\mu_i), \tag{3}$$

dove $b'(\theta_i)$ e $b''(\phi_i)$ sono le prime due derivate di $b(\theta_i)$, $\upsilon(\mu_i) = b''(\theta_i)|_{\theta_i=\theta(\mu_i)}$ è la funzione di varianza e $\theta(\mu_i)$ è l'inversa di $\mu(\theta_i)$.
La distribuzione di Y_i è indicata con

$$Y_i = DE_1(\mu_i, a_i(\phi)\upsilon(\mu_i)), \tag{4}$$

con $\mu_i \in M$, lo spazio delle medie.
Per ogni unità si ha a disposizione il valore di p variabili esplicative $x_i = (x_{i1},x_{i2},...,x_{ip})$, $i = 1,...,n$. Ogni modello lineare generalizzato è definito da tre componenti:

- Distribuzione della risposta: le variabili casuali indipendenti Y_i si distribuiscono come descritto nella formula (4);
- Predittore lineare: per un vettore $\beta = (\beta_1,...,\beta p)^T$ di coefficienti di regressione e una matrice di covariate X, di dimensione $n \times p$, il predittore lineare è $\eta = \mathrm{X}\ \beta$ con componenti $\eta_i = \mathrm{x}_i\beta$;

- Funzione di legame: è la funzione $g(\mu_i)$ che collega μ_i al predittore lineare η_i, assunta di forma nota, derivabile con continuità e invertibile, $g(\mu_i) = x_i\beta$.

3.2 Modello di regressione logistica

Con risposte binarie non raggruppate, il modello statistico per l'i-esima osservazione è quello bernoulliano, $Ber(\mu_i)$.
La modellazione più semplice assume che le risposte siano indipendenti.
Poiché la media di Y_i assume valori nell'intervallo (0,1), come funzione di legame $g(\mu_i)$ si assume una funzione g: $[0, 1] \longrightarrow \Re$ monotona. Appartengono a questa classe tutte le funzioni inverse di funzioni di ripartizione di variabili casuali continue con supporto $\Re$. Se infatti $F(\mu_i)$ è la funzione di ripartizione di una variabile casuale continua con supporto $\Re$, la sua funzione inversa $F^{-1}(u)$ ha come dominio [0, 1], come condominio $\Re$ ed è monotona crescente. Il modello assume allora che per una tale $F(\mu_i)$,

$$g(\mu_i) = F^{-1}(\mu_i) = x_i\beta, \tag{5}$$

ossia

$$\mu_i = \pi_i = F(x_i\beta). \tag{6}$$

Le principali funzioni di legame sono:

- Funzione di legame logistica o logit (legame canonico):

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i\beta; \tag{7}$$

- Funzione di legame probit:

$$g(\mu_i) = \Phi^{-1}(\mu_i) = x_i\beta, \tag{8}$$

  dove $\Phi(.)$ rappresenta la funzione di ripartizione della distribuzione normale standard.

- Funzione di legame log-log complementare (complementare log-log)

$$g(\mu_i) = log\left[-log(1-\mu_i)\right] = x_i\beta; \tag{9}$$

- Funzione di legame log-log

$$g(\mu_i) = -log[-log(\mu_i)]; \tag{10}$$

- Funzione di legame di Cauchy:

$$g(\mu_i) = tan(\pi(\mu_i - 0.5)). \tag{11}$$

L'andamento delle varie funzioni di legame è rappresentato nella Figura 90.

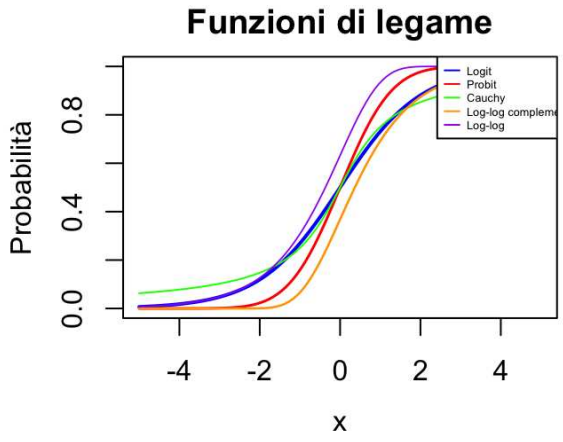

Figura 46. Funzioni di legame per dati binari.

Nel caso si utilizzi la funzione di legame logistica, l'espressione che lega le quantità è

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} = x_i\beta = \eta_i. \tag{12}$$

Approfondendo la regressione logistica, che utilizza la funzione di legame della quale ci serviremo in questo caso di studio, essa permette di interpretare il predittore lineare η_i in termini di logaritmo della quota (*log-odds*). Il generico coefficiente β_r esprime l'effetto sul logaritmo della quota di un incremento unitario di x_{ir}, fermo restando il valore delle altre variabili esplicative presenti nel modello.
Con una sola variabile esplicativa quantitativa x e il modello con $\eta_i = \beta_1 + \beta_2 x_i$, se si incrementa la variabile esplicativa di un'unità, passando da x_i a x_i+1, la variazione del predittore lineare è uguale a β_2. La quota corrispondente risulta quindi moltiplicata per $\exp(\beta_2)$. Il modello logit si basa sull'assunzione

$$Pr(Y_i = 1|x_i) = \frac{e^{\beta_1+\beta_2 x_i}}{1 + e^{\beta_1+\beta_2 x_i}}. \tag{13}$$

Il log-rapporto delle quote è

$$log\left(\frac{Pr(Y_i = 1|x_i = 1)/Pr(Y_i = 0|x_i = 1)}{Pr(Y_i = 1|x_i = 0)/Pr(Y_i = 0|x_i = 0)}\right) = \beta_2, \tag{14}$$

e si può leggere come il rapporto tra la probabilità con la quale un evento si verifica in un gruppo in esame e la probabilità con la quale lo stesso evento si verifica in un gruppo di controllo.
Un problema nella stima dei modelli di regressione per dati binari è quello della classificazione scorretta, ovvero l'assegnazione dell'oggetto di studio ad una categoria diversa da quella alla quale dovrebbe essere attribuito. Ignorare l'errata classificazione può portare a distorsioni asintotiche non nulle degli stimato di massima verosimiglianza. Questa relazione finale, tramite l'applicazione del dataset in considerazione si occupa proprio della riduzione della distorsione attraverso l'utilizzo e il confronto di diversi metodi di correzione.

3.3 Riduzione della distorsione: due metodi tradizionali

Se si dispone di un modello regolare con parametro $\beta = (\beta_1, \beta_2,...,\beta p)$ p-dimensionale, la distorsione asintotica dello stimatore di massima verosimiglianza $\hat{\beta}$ può essere scritta come

$$b(\beta) = \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n^2} + ..., \tag{15}$$

dove n è la numerosità campionaria [7].
Sono stati studiati e discussi in primis due approcci tradizionali per la riduzione della distorsione, che hanno in comune il fatto di essere "correttivi" piuttosto che preventivi (ovvero la stima di massima verosimiglianza $\hat{\beta}$ viene prima calcolata e poi corretta). Il primo metodo è chiamato *jackknife* e consiste nel ricalcolare

più volte la grandezza statistica stimata lasciando fuori dal campione un'osservazione alla volta. Tale metodo richiede solo che il termine principale nella distorsione sia di ordine n^{-1}. Il secondo approccio è chiamato *riduzione della distorsione tramite sviluppo in serie*. Anche questo ha successo nella rimozione del termine

$$\frac{b_1(\beta)}{n} \tag{16}$$

dalla distorsione asintotica. Un requisito fondamentale per l'applicazione di questi due metodi è l'esistenza di $\hat{\beta}$ finito per il campione con cui si lavora. Nel caso in cui $\hat{\beta}$ sia infinito, come avviene ad esempio nei modelli di regressione logistica, i due approcci tradizionali non sono applicabili, ma si considerano altri metodi.

3.4 Il metodo di Firth & Kosmidis

Si inizia con la correzione per la distorsione proposta da Firth [7] per la stima dei parametri non vincolato dalla finitezza di $\hat{\beta}$.
In un problema regolare di stima, le stime di massima verosimiglianza vengono ottenute come soluzione delle equazioni di verosimiglianza. Indicando con $L(\beta)$ la funzione di verosimiglianza e con $l(\beta)$ la funzione di log-verosimiglianza, le equazioni per la stima dei parametri che portano a ottenere $\hat{\beta}$ sono date da

$$\frac{\partial l(\beta)}{\partial \beta} = U(\beta) = 0. \tag{17}$$

Il metodo proposto da Firth per ridurre la distorsione dello stimatore di $\hat{\beta}$ si ottiene introducendo una correzione sistematica del meccanismo che produce la stima di massima verosimiglianza, cioè dell'equazione di verosimiglianza basata sulla funzione punteggio $\mu(\beta)$ piuttosto che della stima stessa della funzione di verosimiglianza; se si ha che $\hat{\beta}$ ha una distorsione di $b(\beta)$, la funzione punteggio viene spostata verso il basso in ogni punto β di $i(\beta)b(\beta)$, con $i(\beta) = E\left[-\frac{\partial U(\beta)}{\partial \beta}\right]$ (che corrisponde alla matrice nota come informazione di Fisher).
Viene così definita una nuova funzione punteggio

$$U^*(\beta) = U(\beta) - i(\beta)b(\beta), \tag{18}$$

la quale porta a calcolare $\tilde{\beta}$, la cui distorsione sarà minore rispetto a quella delle stime ottenute usualmente, come soluzione di

$$U^*(\beta) = 0, \tag{19}$$

e presenta distorsione asintotica $\mathrm{O}(n^{-2})$, inferiore rispetto alla distorsione di $\hat{\beta}$.

Si confrontano graficamente due modelli teorici logistico e logistico penalizzato

generati con una sequenza di 100 valori all'interno del range [-10,10] e loro rispettive funzioni punteggio (Figure 47).

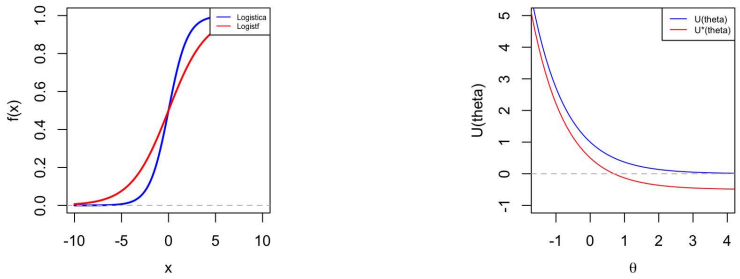

Figura 47. Confronto tra due modelli teorici logistico e logistico penalizzato generati e funzioni punteggio dei due modelli

La distorsione di $\hat{\beta}$ deriva dalla combinazione di due fattori, ovvero la non distorsione della funzione punteggio, $E\left[U(\beta)\right] = 0$ e la curvatura della funzione punteggio.
Se $U(\beta)$ è lineare in θ, allora $E(\hat{\beta}) = \beta$, ma la curvatura e la non distorsione della funzione di punteggio si combinano provocando una distorsione nello stimatore di massima verosimiglianza $\hat{\beta}$; dunque la distorsione di $\hat{\beta}$ può essere ridotta attraverso l'introduzione di una piccola distorsione nella funzione di punteggio.
Se $\hat{\beta}$ è soggetto ad una distorsione positiva (cioè se $E(\hat{\beta}) > 0$), la funzione punteggio deve essere spostata verso il basso in ogni punto β di una quantità pari a $i(\beta)b(\beta)$, dove $b(\beta)$ indica la distorsione.

Il metodo di Firth è stato reso più efficiente specialmente dal punto di vista computazionale da Kosmidis [8].

3.5 Il metodo di Kenne Pagui

Kenne Pagui *et al.* (2017) hanno sviluppato un metodo di riduzione della distorsione basato sulla mediana e non sulla media come indice di centratura.

Anche questo metodo non richiede il calcolo della stima di massima verosimiglianza e oltre a ridurre la distorsione dello stimatore garantisce la finitezza delle stime in situazioni in cui le stime di massima verosimiglianza sono sulla frontiera dello spazio parametrico. La loro proposta si basa, come per il metodo di Firth , sulla derivazione di una nuova equazione di stima [9]. Nel caso del parametro scalare di interesse, viene modificata la funzione di punteggio e, adottando la mediana come indice di centratura di essa, si ottiene la funzione di punteggio modificata sottraendo alla funzione di punteggio la sua mediana approssimata. La soluzione delle risultati equazioni di stima, purché aventi soluzione unica, restituiscono uno stimatore con distorsione ridotta in mediana.
Con il metodo di Kenne Pagui *et al.* la funzione di punteggio modificata risulta essere

$$U^*(\beta) = U(\beta) + i(\beta)M_1(\beta), \tag{20}$$

dove $M_1(\beta)$ è un termine opportuno di correzione.
Semplificazioni algebriche per il metodo proposto [9] sono state ottenute dagli stessi autori e la nuova formulazione ha diversi vantaggi, tra i quali quello di facilitare l'implementazione per modelli parametrici generali e di ottenere notevoli vantaggi computazionali.

3.6 Formulazione del modello teorico e applicazione delle correzioni

Si intende indagare la relazione fra il gruppo di sani e malati e un'insieme di variabili esplicative, al fine di confrontare la presenza di malattia o meno tra gruppi di soggetti con caratteristiche diverse e individuare i fattori e valori più importanti. Si assume che i dati di classificazione della malattia o meno siano realizzazioni della variabile casuale group, con $\text{group}_i = 0$ se il soggetto è sano e $\text{group}_i = 1$ se il soggetto è malato, *i = 1,...,n.* Viene adattato il modello logistico specificato dalle seguenti ipotesi:

1. il legame logistico è:

$$g(\mu_i) = logit(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) \tag{21}$$

2. il modello inziale è

$$logit(\mu_i) = \eta_i = \beta_0, \tag{22}$$

dal quale poi si applica la procedura *forward* al fine stimare il modello con le variabili significative.

In questo capitolo ci si focalizza nella stima del modello attraverso l'applicazione delle correzioni sopra presentate, dal momento che a causa del fenomeno

della separazione è complicato anche per l'algoritmo trovare una soluzione univoca entro il numero massimo di passaggi o iterazioni consentiti; tale fenomeno si verifica quando una o più variabili esplicative possono predire perfettamente l'esito della variabile dipendente, causando però problemi durante l'adattamento del modello poiché l'algoritmo di massima verosimiglianza non riesce a stimare correttamente i coefficienti del modello a causa della non singolarità.
Per la scelta del modello si utilizza un approccio di tipo forward partendo dal modello con solo l'intercetta, mentre la significatività dei coefficienti viene testata utilizzando la statistica test log-rapporto di verosimiglianza penalizzata per la correzione di Firth e la statistica test log-rapporto di verosimiglianza con le rispettive correzioni di Kosmidis e Kenne Pagui.

Si inizia con la stima del modello applicando la correzione di Firth (1993). Il processo di selezione porta a includere la variabile urinaryMCP, con un valore osservato del test pari a X^{2oss}= 6.198 e un p-value <0.01, la variabile TIBC con un valore osservato del test pari a X^{2oss}= 16.486 e un p-value <0.01, la variabile SDMA con un valore osservato del test pari a X^{2oss}= 12.531 e un p-value <0.01, la variabile creatinine con un valore osservato del test pari a X^{2oss}= 7.305 e un p-value <0.01 e infine la variabile RR con un valore del test osservato pari a $X2^{2oss}$= 5.298 e un p-value <0.01. Come riportato anche in Tabella 27 il modello finale stimato è:

$\hat{logit}(\mu_i)$=$\hat{\eta}_i$= 0.002 urinaryMCP$_i$ -0.011 TIBC$_i$ +0.313 SDMA$_i$ - 2.127 creatinine$_i$ + 0.030 RR$_I$.

	Coefficients	Std.Error	I.C	Chisq	p-value
urinaryMCP	0.002	0.001	(0.0004,0.005)	6.198	<0.01
TIBC	-0.011	0.006	(-0.019,-0.005)	16.486	<0.01
SDMA	0.313	0.104	(0.125,0.5580)	12.531	<0.01
creatinine	-2.127	0.863	(-4.111,-0.729)	7.305	<0.01
RR	0.030	0.016	(0.0041,0.066)	5.298	<0.01

Tabella 27. Tabella di adattamento del modello, si riportano le stime dei coefficienti e i relativi standard error, gli intervalli di confidenza al 95%, valore osservato del test log-rapporto della massima verosimiglianza penalizzata e p-value.

La capacità predittiva del modello è buona, con AUC=0.943 (si veda Figura 48).

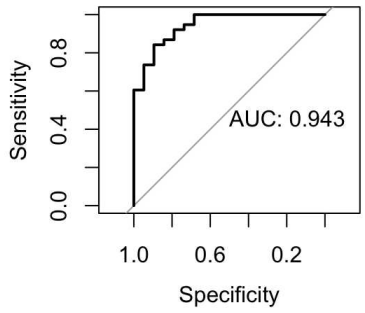

Figura 48. Curva ROC e AUC del modello logistico penalizzato senza intercetta per valutare la bontà del modello.

Un altro metodo utilizzato per testare la bontà del modello è il test Hosmer-Lemeshow, il quale riporta un valore del test pari a $X^2 = 4.247$ con rispettivo p-value pari a 0.8342.

Si riporta infine la matrice di confusione (Tabella 28) o tabella di errata classificazione:

previsti/osservati	0	1
0	15	4
1	4	34

Tabella 28. Matrice di confusione modello logistico penalizzato con correzione di Firth.

Il modello stimato ha una buona sensibilità, specificità e accuratezza: la sensibilità è pari a 0.895, la specificità è pari a 0.789 e l'accuratezza è pari a 0.859.

Dal modello stimato si evince che:

1. Il coefficiente stimato relativo a urinaryMCP pari a 0.002 evidenzia un effetto significativo positivo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.002}$ =1.002 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di urinaryMCP aumenta dello 0.2% a parità delle altre variabili.

2. Il coefficiente stimato relativo a TIBC pari a -0.011 evidenzia un effetto significativo negativo tra pazienti malati e pazienti sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{-0.011}$ =0.98 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di TIBC diminuisce del 2% a parità delle altre variabili.

3. Il coefficiente stimato relativo a SDMA pari a 0.313 evidenzia un effetto significativo positivo tra pazienti malati e pazienti sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.313}$ =1.367 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di SDMA aumenta del 36.7% a parità delle altre variabili.

4. Il coefficiente stimato relativo a creatinine è pari a -2.127 evidenzia un effetto significativo negativo tra pazienti malati e pazienti sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{-2.127}$ =0.119 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di creatinina diminuisce dell'88% circa a parità delle altre variabili.

5. Il coefficiente stimato relativo a RR pari a 0.030 evidenzia un effetto significativo positivo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.030}$ =1.03 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di RR aumenta del 3% a parità delle altre variabili.

Utilizzando lo stesso modello teorico applicando la correzione di distorsione in media di Kosmidis e Firth (riadattata nel 2009 sotto un punto di vista computazionale più efficiente), sempre con l'approccio di tipo forward, il processo di selezione porta a includere nel modello le variabili urinaryMCP, con un valore del test W^{oss} pari a 2.131 e un p-value di 0.033, la variabile TIBC, con un valore del test W^{oss} pari a -2.856 e un p-value <0.01, la variabile SDMA con un valore W^{oss} pari a 2.877 e un p-value <0.01 e infine la variabile creatinine con un valore W^{oss} di -1.885 e un p-value <0.01. Come riportato anche in Tabella 2 il modello finale stimato è:

$$\text{logit}(\hat{\mu}_i)=\hat{\eta}_i= 0.002\ \text{urinaryMCP}_i - 0.006\ \text{TIBC}_i + 0.285\ \text{SDMA}_i - 1.885\ \text{creatinine}_i$$

	Coefficients	Std.Error	z value	p-value
urinaryMCP	0.002	0.001	2.131	0.033
TIBC	-0.006	0.002	-2.846	<0.01
SDMA	0.285	0.098	2.877	<0.01
creatinine	-1.885	0.711	-2.653	<0.01

Tabella 29. Tabella di adattamento del modello, si riportano le stime dei coefficienti e i relativi standard error, valore osservato del test di Wald e p-value.

La capacità predittiva del modello è molto buona, con AUC=0.913 (si veda Figura 49).

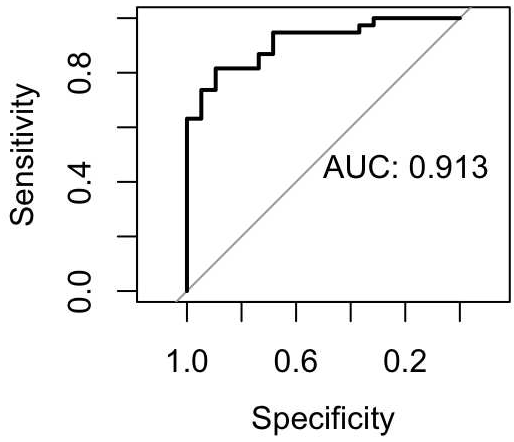

Figura 49. Curva ROC e AUC del modello logistico con correzione distorsione di Firth e Kosmidis in media, per valutare la bontà del modello.

Il test Hosmer-Lemeshow riporta un valore del test pari a $X^2 = 4.247$ con rispettivo p-value pari a 0.8342.

La matrice di confusione (Tabella 30) presenta una lieve differenza rispetto a quella relativa al modello stimato con la correzione di Firth, poiché il modello ha un vero negativo in più e ha un vero positivo in meno.

previsti/osservati	0	1
0	14	5
1	5	33

Tabella 30. Matrice di confusione modello logistico penalizzato con correzione in media di Kosmidis.

Il modello stimato ha una buona sensibilità e specificità: la sensibilità è pari a 0.868 e la specificità è pari a 0.737 e l'accuratezza è pari a 0.825.

Dal modello stimato si evince che:

1. Il coefficiente stimato relativo a urinaryMCP pari a 0.002 evidenzia un effetto significativo positivo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.002} = 1.002$ indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di urinaryMCP aumenta dello 0.2% a parità delle altre variabili.

2. Il coefficiente stimato relativo a TIBC pari a -0.006 evidenzia un effetto significativo negativo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{-0.006} = 0.994$ indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di TIBC diminuisce del 0.6% a parità delle altre variabili.

3. Il coefficiente stimato relativo a SDMA pari a 0.287 evidenzia un effetto significativo positivo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.287} = 1.332$ indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di SDMA aumenta del 33.2% a parità delle altre variabili.

4. Il coefficiente stimato relativo a creatinine è pari a -1.885 evidenzia un effetto significativo negativo tra l'essere malati e l'essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{-1.885} = 0.152$ indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di creatinina diminuisce dell'85% circa a parità delle altre variabili.

Servendosi invece sempre dello stesso modello teorico descritto nel paragrafo 3.5, con l'approccio di tipo forward, però con la correzione della distorsione in mediana, il processo di selezione porta a includere nel modello le variabili urinaryMCP, con un valore del test W_{oss} pari a 2.131 e un p-value di 0.033, la variabile TIBC, con un valore del test W^{oss} pari a -2.846 e un p-value <0.01, la variabile SDMA con un valore W^{oss} pari a 2.877 e un p-value <0.01 e infine la variabile creatinine con un valore W^{oss} di -2.653 e un p-value <0.01. Come riportato anche in Tabella 2 il modello finale stimato è:

$$\text{logit}(\hat{\mu}_i) = \hat{\eta}_i = 0.002\ \text{urinaryMCP}_i - 0.006\ \text{TIBC}_i + 0.285\ \text{SDMA}_i - 1.885\ \text{creatinine}_i$$

	Coefficients	Std.Error	z value	p-value
urinaryMCP	0.002	0.001	2.181	0.029
TIBC	-0.007	0.002	-2.846	<0.01
SDMA	0.299	0.103	2.907	<0.01
creatinine	-2.004	0.761	-2.653	<0.01

Tabella 31. Tabella di adattamento del modello, si riportano le stime dei coefficienti e i relativi standard error e valore osservato del test di Wald e p-value.

La capacità predittiva del modello è buona, con AUC=0.914 (si veda Figura 50).

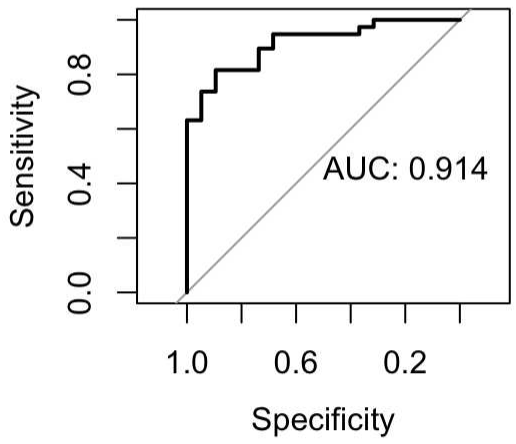

Figura 50. Curva ROC del modello logistico di Kenne Pagui con correzione distorsione in mediana, per valutare la bontà del modello.

Il test Hosmer-Lemeshow riporta un valore del test pari a $X^2 = 2.244$ con rispettivo p-value pari a 0.973.

La matrice di confusione (Tabella 32) presenta una differenza rispetto alle due precedenti, poiché sono ora presenti 34 veri positivi e 14 veri negativi.

previsti/attuali	0	1
0	14	4
1	5	34

Tabella 32. Matrice di confusione modello logistico penalizzato con correzione in mediana di Kenne Pagui.

Il modello stimato ha una buona sensibilità e specificità: la sensibilità è pari a 0.894 e la specificità è pari a 0.737 e l'accuratezza è pari a 0.842.
Dal modello stimato si evince che:

1. Il coefficiente stimato relativo a urinaryMCP pari a 0.002 evidenzia un effetto significativo positivo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.002} = 1.002$ indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di urinaryMCP aumenta dello 0.2% a parità delle altre variabili.

2. Il coefficiente stimato relativo a TIBC pari a -0.007 evidenzia un effetto significativo negativo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{-0.007}$ =0.993 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di TIBC diminuisce dello 0.7% a parità delle altre variabili.

3. Il coefficiente stimato relativo a SDMA pari a 0.299 evidenzia un effetto significativo positivo tra l'essere malati e essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{0.299}$ =1.348 indica che la quota per la probabilità che un paziente sia malato rispetto che sia sano all'aumentare di un'unità del valore di SDMA aumenta del 35% circa a parità delle altre variabili.

4. Il coefficiente stimato relativo a creatinine è pari a -2.004 evidenzia un effetto significativo negativo tra l'essere malati e l'essere sani all'aumentare di un'unità della variabile. In termini di *odds ratio*, il valore $e^{-2.004}$ =0.135 indica che la quota per la probabilità che un paziente sia malato rispetto

che sia sano all'aumentare di un'unità del valore di creatinina diminuisce dell' 86% circa a parità delle altre variabili.

Può essere d'interesse, oltre ai vari metodi di bontà di stima dei modelli utilizzati precedentemente, calcolare il criterio di informazione di Akaike (AIC) (si veda Pace Salvan, 1996, cap. 4.5): il modello logistico penalizzato con la correzione di Firth ha un AIC di 33.294, il modello stimato con la correzione in media di Kosmidis ha un valore AIC di 46.672 ed infine il modello stimato con la correzione in mediana di Kenne Pagui ha un valore AIC pari a 46.412.
Visti i valori di AIC, della curva ROC e anche dell'accuratezza, specificità e sensibilità sembrerebbe che i risultati migliori si tengano utilizzando il modello logistico con correzione di Firth, con rispettive variabili esplicative urinaryMCP, TIBC, creatinine e RR.

4 Capitolo 4: Conclusioni

La diagnosi della Leishmaniosi prevede di norma la determinazione nei pazienti di diversi valori. Grazie ai vari metodi di bontà per i modelli stimati, i quali confermano che tutti i tre modelli siano ben stimati, si possono considerare tutti e tre simultaneamente viste le stime dei parametri simili e tutte significative. La variabile urinaryMCP nel caso di studio di questa relazione ha valori che stanno all'interno del range 131.7-2066.6 pg/ml; all'aumentare di un pg/ml, la quota della probabilità che il paziente sia malato rispetto che il paziente sia sano aumenta dello 0.2%. La variabile TIBC nel caso di studio di questa relazione ha valori che stanno all'interno del range 137-455 g/dL; all'aumentare di un g/dL la quota per la probabilità che il paziente sia malato rispetto che il paziente sia sano diminuisce di circa lo 0.7%. La variabile SDMA nel caso di studio di questa relazione ha valori che si estendono nel range 4.60-31 g/dL; all'aumentare di un g/dL la quota per la probabilità che il paziente sia malato rispetto che il paziente sia sano aumenta di circa il 34%. La variabile creatinine ha valori che stanno all'interno del range 0.34-5.77 mg/dL; all'aumentare di un mg/dL la quota per la probabilità che il paziente sia malato rispetto che il paziente sia sano diminuisce di circa l'85%. Infine, considerando anche la variabile RR relativa alla frequenza respiratoria, nel caso di studio di questa relazione i valori si estendono all'interno del range 14-120 respiri al minuto; all'aumentare di un respiro al minuto la quota per la probabilità che il paziente sia malato rispetto che il paziente sia sano aumenta del 3%.
Si conclude che in questo caso di studio è fondamentale per la diagnosi della Leishmaniosi analizzare 5 variabili: urinaryMCP, ovvero una quantità di proteine nelle urine, TIBC, ovvero la capacità delle proteine di circolo ematico di legare con il ferro, SDMA, ovvero la dimetilarginina simmetrica, creatinine, ovvero la quantità di creatina nel sangue ed infine RR, ovvero a frequenza respiratoria. Di ciascuna di queste si è analizzato che la probabilità di contrarre la malattia aumenta all'aumentare dei valori delle variabili urinaryMCP, SDMA e RR, mentre la probabilità di contrarre la malattia diminuisce all'aumentare dei valori delle variabili TIBC e creatinine.

Riferimenti bibliografici

[1] Xavier Roura e Laura Ordeix, *Caratteristiche dermatologiche della leismaniosi canina*, n. 28.1, Veterinary FOCUS, 2020.

[2] Istituto Zooprofilattico Sperimentale delle Venezie. *Leishmaniosi canina, come evitarla, come difendersi*, presentazione n.2, III edizione, 2020.

[3] Gruppo Leishmania, *Leishmaniosi canina: Approccio diagnostico e classificazione del paziente leishmaniotico e gestione del paziente proteinurico*, Parte I, Giugno 2007.

[4] Laura Ventura, Walter Racugno, *Biostatistica. Casi di studio in r.*, Ed. Egea, 2017.

[5] Nelder J.A. and Weddenburn R.W.M. (1972), *Generalized Linear Models*, Jorunal of the Royal Statistic Society.Series A, **135**, No.3, 370-384.

[6] Alessandra Salvan, Nicola Sartori, Luigi Pace (2020), *Modelli lineari generalizzati*, Ed. Springer Verlag.

[7] Firth, D. (1993), *Bias reduction of maximum likelihood estimates*, Biometrika, **80**, 1993, 27-38.

[8] Ioannis Kosmidis and David Firth (2009), *Bias reduction in exponential family nonlinear models*, Biometrika, **96**, No.4, 793-804.

[9] Kenne Pagui, Ioannis Kosmidis, Nicola Sartori (2019), *Mean and median bias reduction in generalized linear models*, Statistics and Computing, 43-59.

[10] Kenne Pagui, E.C., Alessandra Salvan, Nicola Sartori, (in fase di elaborazione, 2019), *Efficient implementation of median bias reduction.*