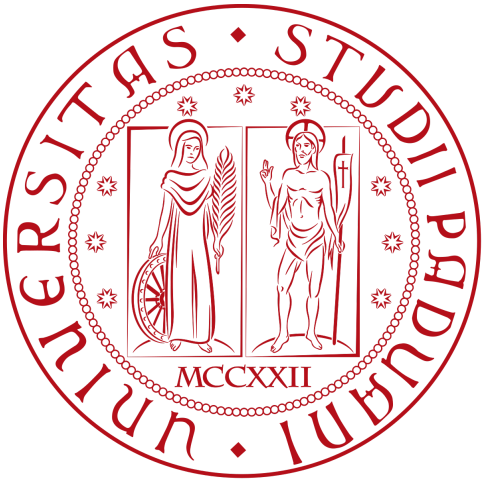
Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale
in
Statistica per le Tecnologie e le Scienze



ANALISI DEL DATASET ***acath***

Bardellone Lenny matr. 2035053
Lovato Edoardo matr. 2001319

1 Introduzione

Il muscolo cardiaco ha bisogno di un costante apporto di sangue ricco di ossigeno. Le arterie coronarie svolgono proprio la funzione di assicurarne la giusta quantità, ramificandosi dall'aorta alla sua uscita dal cuore. La malattia coronarica (coronaropatia) è una patologia che tende a restringere una o più di queste arterie coronarie e può ostruire e ridurre il flusso sanguigno, provocando così dolore toracico ed aumentando le possibilità che si formi un coagulo che, bloccando l'arteria, può causare un infarto. La sua forma più grave è la malattia dell'arteria coronaria principale sinistra o dei tre vasi (coronaropatia trivasale), ovvero la condizione quando si presentano restringimenti su tutti e tre i rami coronarici.

Il dataset considerato è ***acath***, proviene dalla Banca Dati sulle Malattie Cardiovascolari dell'Università di Duke (Durham - Carolina del Nord) e comprende 3504 pazienti e 5 variabili.
I pazienti sono stati indirizzati al Duke University Medical Center per dolore toracico.
L'obiettivo dello studio è modellare la previsione della probabilità di malattia coronarica significativa (>= 75% di restringimento del diametro in almeno un'importante arteria coronaria) e la previsione della probabilità di malattia coronarica grave dato che è stata "confermata" una qualche malattia significativa.
Nel primo modello verrà utilizzata ***sigdz*** come variabile risposta, mentre nel secondo ***tvdlm,*** prendendo in considerazione però i soli pazienti con presenza di malattia significativa (***sigdz***=1).
Per ciascun paziente, sono state rilevate le seguenti variabili:

- ***sex***: variabile dicotomica che rappresenta il sesso, assume valore 0 per i pazienti "maschi" e 1 per le "femmine".
- ***age***: variabile quantitativa che rappresenta l'età del paziente in anni.
- ***cad.dur***: variabile quantitativa che rappresenta la durata dei sintomi in giorni.
- ***sigdz***: variabile dicotomica che rappresenta la presenza o assenza della malattia coronarica significativa (assume 0 per l'assenza e 1 per la presenza).
- ***tvdlm***: variabile dicotomica che rappresenta la presenza o assenza della malattia dell'arteria coronaria principale sinistra o dei tre vasi riscontrata a livello cardiaco (assume 0 per l'assenza e 1 per la presenza).

Per una maggiore precisione nelle analisi e per non compromettere i dati, sono stati rimossi dal dataset tre pazienti perché contenevano valori di "***tvdlm***" mancanti. Si fa quindi riferimento ad uno studio con una numerosità campionaria pari a **3501 pazienti.**
Le analisi sono state eseguite con il "software R" (https://www.r-project.org), assumendo un livello di significatività pari a 0.05.
Per approfondimenti sui modelli e test utilizzati nell'analisi si rimanda a "Biostatistica, Casi di studio in R" di Ventura e Racugno, Egea (2017).

2 Analisi esplorative

L'analisi esplorativa permette di ottenere una descrizione preliminare delle informazioni date dalle variabili osservate. Inoltre, consente di osservare se sono presenti relazioni significative tra coppie di variabili attraverso opportuni test statistici.

2.1 Analisi univariate

In seguito, verranno esaminate singolarmente le variabili del dataset al fine di ottenere una sintesi dettagliata per ciascuna di esse.

Per le tre variabili qualitative ***sex***, ***sigdz*** e ***tvdlm*** sono state misurate le distribuzioni di frequenze assolute e quelle relative percentuali per ciascuna modalità.

	N	%
sex		
0 = maschi	2404	68,7
1 = femmine	1097	31,3
Totale	3501	100

Tab. 2.1.1: Distribuzione di frequenze assolute e relative percentuali per la variabile sesso (***sex***)

Dalla tabella 2.1.1 è possibile osservare che all'interno del campione i pazienti maschi sono 2404 e le femmine 1097, rispettivamente il 68.7% e il 31.3%. Eseguendo il test sulle proporzioni è risultato un X-squared pari a 487.19 ed un p-value<0.0001, rifiutando quindi l'ipotesi nulla si conclude che le due proporzioni sono significativamente diverse. Il rapporto "maschi/femmine" è pari a 2.19 e da questo si deduce che i maschi sono più del doppio rispetto alle femmine.

La variabile risposta dello studio è ***sigdz*** che rileva nei pazienti la presenza o meno della malattia coronarica significativa.

Tab. 2.1.2: Distribuzione di frequenze assolute e relative percentuali per la malattia coronarica significativa (***sigdz***)

	N	%
sigdz		
0 = assenza	1169	33,4
1 = presenza	2332	66,6
Totale	3501	100

Dalla tabella 2.1.2 è possibile osservare che all'interno del campione i pazienti con malattia coronarica significativa (***sigdz***) sono 2332 e quelli che ne sono esclusi sono 1169 , rispettivamente il 66.6% e il 33.4%. Anche in questo caso è stato eseguito il test sulle proporzioni ed è risultato un X-squared pari a 385.67 con un p-value<0.0001. Si rifiuta l'ipotesi nulla e si conclude che le due proporzioni sono significativamente diverse.

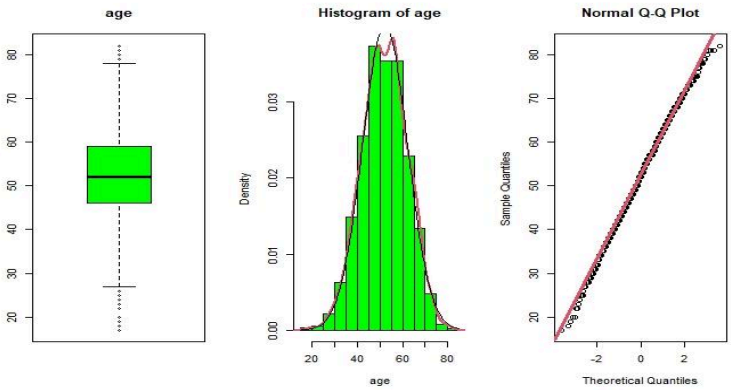
Infine, per l'ultima variabile qualitativa ***tvdlm*** che rappresenta la presenza o meno della malattia coronarica grave, tramite l'analisi esplorativa è possibile visualizzare le distribuzioni di frequenze assolute e relative percentuali per le due categorie.

Tab. 2.1.3: Distribuzione di frequenze assolute e relative percentuali per la malattia coronarica più grave (***tvdlm***)

	N	%
tvdlm		
0 = assenza	2372	67,8
1 = presenza	1129	32,2
Totale	3501	100

Dalla tabella 2.1.3 è possibile osservare che all'interno del campione i pazienti con malattia coronarica grave (***tvdlm***) sono 1129 e quelli che ne sono esclusi sono 2372 , rispettivamente il 32.2% e il 67.8%.
Anche in questo caso è stato eseguito il test sulle proporzioni ed è risultato un X-squared pari a 440.61 con un p-value<0.0001. Si rifiuta l'ipotesi nulla e si conclude, anche in questa circostanza, che le due proporzioni sono significativamente diverse.

Fig. 2.1.1: Boxplot a sinistra, istogramma con densità stimata (in rosso) e curva normale(in nero) al centro, diagramma quantile-quantile a destra per la variabile ***age***



Passando ora all'analisi esplorativa della prima variabile quantitativa ***age***, osservando la tabella seguente (2.1.4) si nota che l'età varia dai 17 agli 82 anni. L'età media dei pazienti è di 52.27 (con deviazione standard pari a 9.93) e l'età mediana è di 52 (con scarto interquartilico di 13). E' possibile affermare che media e mediana coincidono, confermando una distribuzione simmetrica. Osservando la figura 2.1.1 a sinistra si nota che il boxplot presenta alcuni outliers sia nella coda superiore che in quella inferiore.

Osservando i diagrammi quantile-quantile (figura 2.1.4 a destra) si nota un leggero allontanamento delle code dalla retta di riferimento. Di conseguenza, la statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità (W = 0.998, p-value<0.0001). Avendo però una numerosità campionaria molto elevata è possibile invocare il Teorema del Limite Centrale, il quale suggerisce che le medie campionarie saranno approssimativamente normali. E' possibile, così, andare ad accettare l'ipotesi di normalità delle medie.

Tab. 2.1.4: Statistiche di sintesi della variabile ***age***

Min.	1st.Qu.	Median	Mean	3Rd. Qu	Max.	S.d.	IQR
17.00	46.00	52.00	52.27	59.00	82.00	9.93	13.00

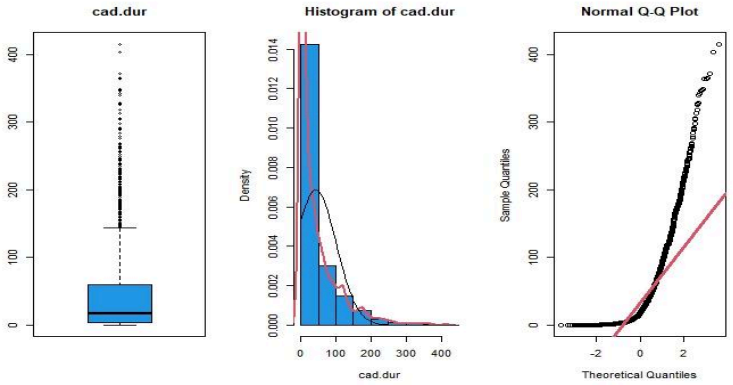


Fig. 2.1.2: Boxplot a sinistra, istogramma con densità stimata (in rosso) e curva normale(in nero) al centro, diagramma quantile-quantile a destra per la variabile ***cad.dur***

Infine, come ultima analisi esplorativa si tratta la variabile quantitativa ***cad.dur***. In particolare, dalla tabella seguente (2.1.5) si nota che la durata dei sintomi per i pazienti varia da 0 a 416 giorni. La durata media è di 42.97 (con deviazione standard pari a 58.15) e la durata mediana è di 18 (con scarto interquartilico di 56). Si osserva che la mediana è di molto inferiore alla media e questo indica una forte asimmetria nella distribuzione (figura 2.1.2 al centro). Osservando la figura 2.1.2 a sinistra, si nota che il boxplot presenta molti outliers nella coda superiore, confermando la forte asimmetria. La statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità (W = 0.728, p-value<0.0001). Anche dal grafico quantile-quantile (figura 2.1.2 a destra) si nota che il rifiuto dell'ipotesi di normalità è causato da un allontanamento dei punti dalla retta di riferimento. Vista la numerosità campionaria elevata, anche in questo contesto, può essere applicato il Teorema del Limite Centrale. Il caso specifico verrà presentato nella sezione delle analisi bivariate (2.2) in particolare nelle figure 2.2.2 e 2.2.5.

Tab. 2.1.5: Statistiche di sintesi della variabile ***cad.dur***

Min.	1st.Qu.	Median	Mean	3Rd. Qu	Max.	S.d.	IQR
0.00	4.00	18.00	42.97	60.00	416.00	58.15	56.00

2.2 Analisi bivariate

Fig.2.2.1: Tabelle di sintesi per l'analisi bivariata del modello

		Test Chi-quadrato		Test di Fisher		
		X-squared	p-value	p-value	OR	IDC OR
a)	sigdz/sex	432.58	<0.0001	<0.0001	0.21	(0.18 , 0.24)

a.1)

sigdz	*sex* 0=maschi	1=femmine
0=assenza	533	636
1=presenza	1871	461

		TEST PARAMETRICO										TEST NP	
		Test di Shapiro-Wilk				Test sulle varianze		Test-t		Test di Welch		Test di Mann-Whitney	
		W0	p-value0	W1	p-value1	F	p-value	t	p-value	t	p-value	W	p-value
b)	sigdz/age	0.997	0.0557	0.996	<0.0001	1.069	0.1869	-15.207	<0.0001	-	-	968097	<0.0001
c)	sigdz/cad.dur	0.670	<0.0001	0.753	<0.0001	0.808	<0.0001	-	-	-4.908	<0.0001	1237914	<0.0001
d)	sex/age	0.998	0.0011	0.997	0.0268	0.969	0.5306	-5.973	<0.0001	-	-	1154953	<0.0001
e)	sex/cad.dur	0.735	<0.0001	0.710	<0.0001	0.996	0.9384	0.031	0.9756	-	-	1270993	0.0860

		Test di correlazione di Pearson			Test di correlazione di Spearman		
		t	p-value	corr	S	p-value	corr-S
f)	age/cad.dur (M)	16.137	<0.0001	0.313	1713077356	<0.0001	0.260
g)	age/cad.dur (F)	7.976	<0.0001	0.234	183156772	<0.0001	0.168

a) La prima analisi bivariata riguarda la variabile ***sigdz*** e la variabile ***sex***. Svolgendo il test del Chi-quadro risulta una forte associazione significativa tra il sesso e la presenza della malattia coronarica significativa. Anche il test di Fisher va a confermare tale conclusione e, tramite l'Odds Ratio di 0.21, suggerisce per le femmine una quota inferiore del 79% rispetto ai maschi.

a.1) - proporzione di maschi con malattia significativa = 0.78
- proporzione di femmine con malattia significativa = 0.42

Queste due proporzioni sono state ottenute anche dal test sulle proporzioni che con p-value <0.0001 indica una differenza significativa tra i due gruppi di pazienti.

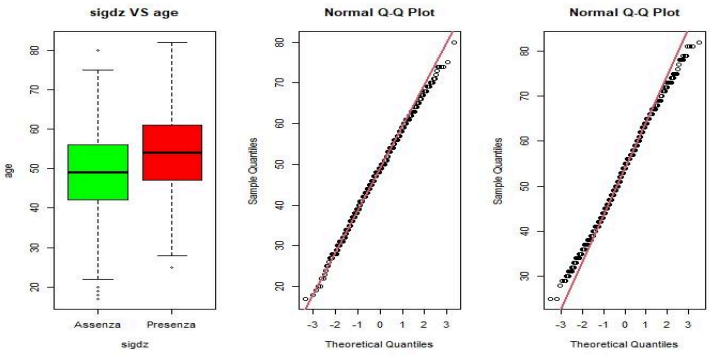


Fig. 2.2.1: Boxplot ***sigdz*** per ***age*** a sinistra, diagramma quantile-quantile sani al centro, diagramma quantile-quantile malati a destra

b) Si studia ora l'effetto dell'età (***age***) sulla presenza/assenza della malattia coronarica significativa (***sigdz***). In particolare, si nota che nei due boxplot vi sono alcuni outliers soprattutto quando la malattia non è presente e la mediana è maggiore nel gruppo dei malati (Figura 2.2.1 a sinistra). Osservando i diagrammi quantile-quantile (figura 2.2.1 al centro e a destra) si nota un leggero allontanamento delle code dalla retta di riferimento. La statistica test di Shapiro-Wilk porta ad accettare l'ipotesi di normalità nel gruppo dei sani (W=0.997, p-value=0.0557), mentre si rifiuta nel gruppo dei malati (W=0.996, p-value<0.0001). Avendo però una numerosità campionaria molto elevata è possibile, attraverso il Teorema del Limite Centrale, accettare comunque l'ipotesi di normalità delle medie.
Le distribuzioni nei due gruppi risultano omoschedastiche dato che si ottiene un valore della statistica osservata F = 1.069 con p-value = 0.1869 che porta, dunque, ad accettare l'ipotesi di uguaglianza tra le due varianze. Infine, andando a controllare l'uguaglianza tra le medie dei due gruppi attraverso il test-t a due campioni indipendenti (t = -15.207, p-value<0.0001) si va a rifiutare l'ipotesi di uguaglianza, concludendo così, che vi è una differenza significativa tra le medie nei due gruppi.
Volendo rifiutare l'ipotesi di normalità è possibile passare direttamente al test non-parametrico di Mann-Whitney che porta comunque alla stessa conclusione di rifiuto (W = 968097,p-value <0.0001). Dunque, è presente un effetto significativo dell'età sulla malattia.

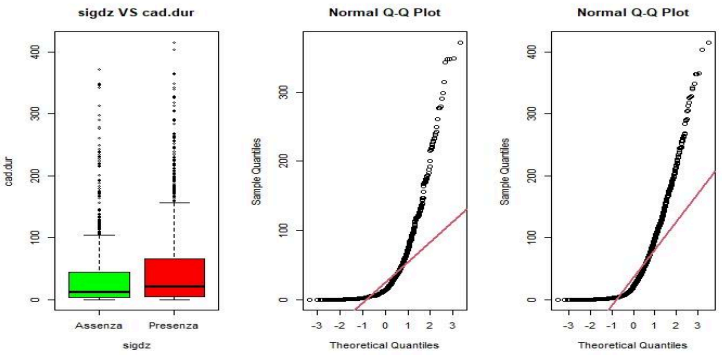


Fig 2.2.2: Boxplot ***sigdz*** per ***cad.dur*** a sinistra, diagramma quantile-quantile sani al centro, diagramma quantile-quantile malati a destra

c) Successivamente, si studia l'effetto della durata dei sintomi (***cad.dur***) sulla malattia (***sigdz***). Come osservato in precedenza dall'analisi esplorativa di ***cad.dur*** (Figura 2.1.2 sezione univariate), anche in questo caso, dai boxplot emergono molti outliers nelle code superiori in entrambi i gruppi (Figura 2.2.2 a sinistra). La statistica test di Shapiro-Wilk porta a rifiutare in entrambi l'ipotesi di normalità (W0 = 0.67046, p-value0 <0.0001 per il gruppo dei sani e W1 = 0.75339, p-value1 <0.0001 per i malati). Infatti, osservando i due diagrammi quantile-quantile (rispettivamente figura 2.2.2 al centro e a destra), è possibile notare un allontanamento consistente dei punti dalla retta di riferimento per entrambi.

Poichè la condizione principale di normalità non è soddisfatta è possibile ricorrere al test non-parametrico di Mann-Whitney che con W = 1237914 e p-value <0.0001, porta a rifiutare l'ipotesi di uguaglianza in mediana per i due gruppi, concludendo così che le mediane delle due distribuzioni per la variabile ***cad.dur*** sono significativamente diverse. Volendo passare comunque attraverso un metodo parametrico, vista la numerosità campionaria elevata, si usufruisce come in precedenza del Teorema del Limite Centrale, che porta alla medesima conclusione tramite il test di Welch con t = -4.908 e p-value<0.0001 (avendo rifiutato l'omoschedasticità). Si conclude che anche la variabile ***cad.dur*** presenta un effetto significativo sulla malattia.

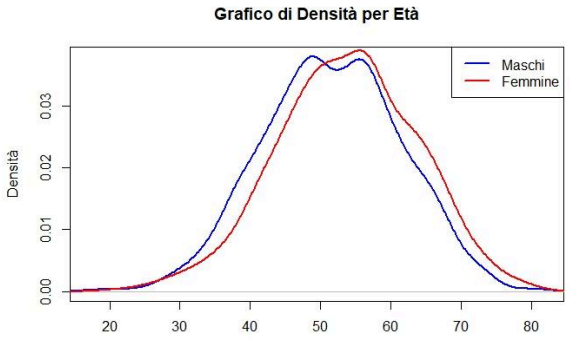


Fig. 2.2.3: Grafico di densità dell'età divisi per maschi e femmine

d) Si verifica ora la relazione tra la variabile età (***age***) e la variabile sesso (***sex***). Attraverso la visualizzazione del grafico di densità (Fig. 2.2.3) si nota che la distribuzione delle femmine è leggermente traslata a destra rispetto a quella dei maschi. I boxplot, (in figura 2.2.4 in basso a sinistra) infatti, confermano che l'età mediana è superiore per le femmine.

	Min.	1st.Qu.	Median	Mean	3Rd. Qu	Max.	S.d.	IQR
0 = Maschi	17.00	45.00	52.00	51.59	58.00	82.00	9.83	13.00
1 = Femmine	20.00	47.00	54.00	53.74	61.00	81.00	9.99	14.00

Tab.2.2.2: Statistiche di sintesi ***age*** per ***sex***

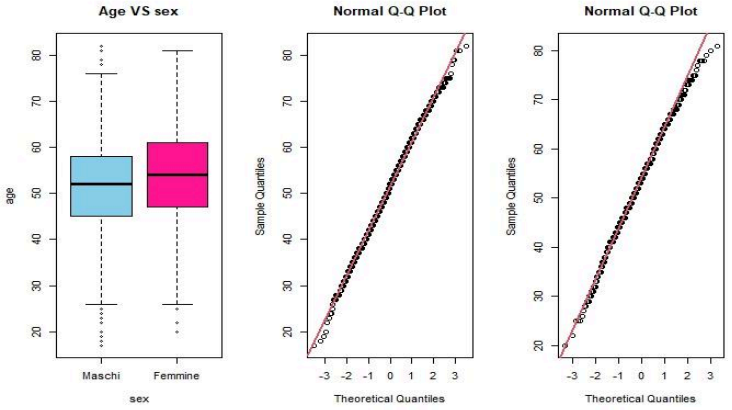


Fig.2.2.4: Boxplot ***age*** per ***sex*** a sinistra, diagramma quantile-quantile sani al centro, diagramma quantile-quantile malati a destra

La statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità sia per i maschi (W0 = 0.998, p-value = 0.0011) che per le femmine (W1 = 0.997, p-value = 0.0268). Osservando i diagrammi quantile-quantile (figura 2.2.4 al centro e a destra), infatti, si nota un leggero allontanamento delle code dalla retta di riferimento. Avendo però una numerosità campionaria molto elevata è possibile, attraverso il Teorema del Limite Centrale, accettare comunque l'ipotesi di normalità delle medie. L'omoschedasticità nei due gruppi viene accettata, ottenendo un valore della statistica osservata F = 0.969 con p-value = 0.5306. In seguito, attraverso il test-t a due campioni indipendenti (t = -5.973, p-value <0.0001) si rifiuta l'ipotesi di uguaglianza delle medie dei due gruppi, concludendo così che vi è una differenza significativa. Volendo, invece, non accettare l'ipotesi di normalità è possibile passare direttamente al test non-parametrico di Mann-Whitney che porta comunque alla stessa conclusione di rifiuto (W = 1154953, p-value<0.0001). Quindi, esiste una relazione significativa tra l'età e il sesso.

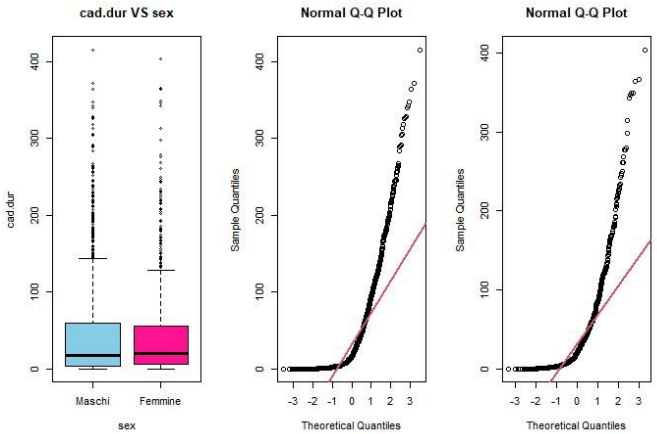


Fig. 2.2.5. : Boxplot ***cad.dur*** per ***sex*** a sinistra, diagramma quantile-quantile "maschi" al centro, diagramma quantile-quantile "femmine" a destra

e) A questo punto, si studia la relazione tra la durata dei sintomi (***cad.dur***) e il sesso (***sex***).
Come osservato in precedenza dall'analisi esplorativa di ***cad.dur*** (Figura 2.1.2 sezione univariate), anche in questo

caso, dai boxplot dei due gruppi risultano molti outliers nelle code superiori (Figura 2.2.5 a sinistra). Osservando i diagrammi quantile-quantile (figura 2.2.5 al centro e a destra) si nota un leggero allontanamento delle code dalla retta di riferimento. Infatti, verificando la normalità, anche questa volta la statistica test di Shapiro-Wilk porta al rifiuto in entrambi i gruppi (W0 = 0.735, p-value0 <0.0001 per il gruppo dei maschi e W1 = 0.710, p-value1 <0.0001 per le femmine). Ancora una volta, quindi, è utile fare ricorso al test non-parametrico di Mann-Whitney che con W = 1270993 e p-value = 0.0860 porta ad accettare l'ipotesi di uguaglianza in mediana per i due gruppi. Si conclude, così, che in mediana non vi è un tempo di durata dei sintomi significativamente diverso tra pazienti maschi e femmine. Anche in questo caso volendo passare comunque attraverso un metodo parametrico, vista la numerosità campionaria elevata, si fa uso del Teorema del Limite Centrale, che porta alla stessa conclusione di accettazione tramite il test t a due campioni indipendenti con t = 0.031 e p-value = 0.9756 (dopo aver accettato l'omoschedasticità con F = 0.996 e p-value = 0.9384).

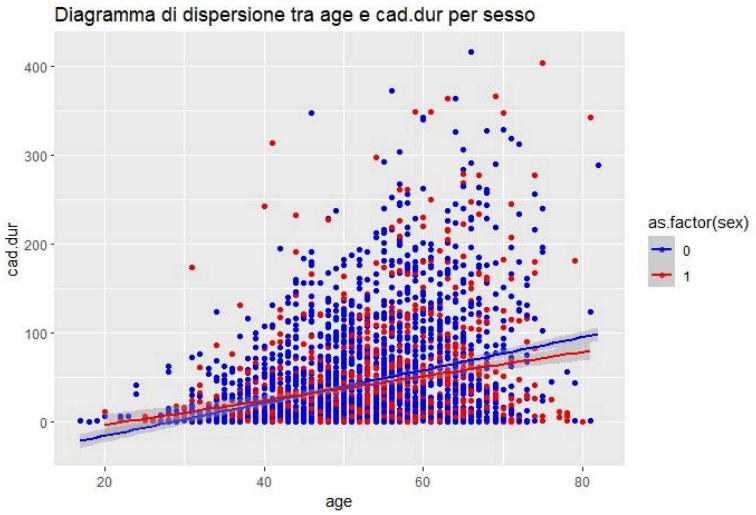


Fig. 2.2.6: Diagramma di dispersione tra ***age*** e ***cad.dur*** diviso per sesso (***sex***)

f) - g) Come ultima analisi bivariata è possibile osservare la relazione tra le due variabili quantitative del dataset, ovvero tra ***age*** e ***cad.dur.*** In particolare, osservando il diagramma di dispersione a lato è possibile notare una lieve relazione tra le due variabili. Infatti, con l'aumentare dell'età del paziente per entrambi i sessi si nota anche un aumento della durata dei sintomi. Nel dettaglio applicando il test di correlazione di Pearson tra ***age*** e ***cad.dur***, per i maschi risulta un indice pari a 0.313 con t = 16.137 ed un p-value <0.0001, mentre per le femmine un indice di 0.234 con t = 7.976 e p-value <0.0001. Quindi è possibile concludere che entrambe le correlazioni sono positive e significativamente diverse da zero. Vista la presenza di outliers, si applica anche il test più robusto di correlazione di Spearman che, con un indice pari a 0.260 con S = 1713077356 e un p-value <0.0001 per i maschi, mentre per le femmine un indice di 0.168 con S = 183156772, porta a concludere, anche in questo caso, che entrambe le correlazioni sono positive e significativamente diverse da zero. Esiste, quindi, una debole relazione tra ***age*** e ***cad.dur*** in entrambi i sessi.

3 Stima del modello

Per modellare la presenza della malattia coronarica significativa (***sigdz***) in funzione del sesso (***sex***), dell'età (***age***) e della durata dei sintomi dei pazienti (***cad.dur***), si adatta un modello di regressione logistica sulla base degli studi precedentemente svolti.

3.1 Modello di regressione logistica

Il modello di regressione logistica è un modello usualmente impiegato quando la variabile dipendente Y (variabile risposta ***sigdz***) è dicotomica e si vuole modellizzare la media di tale variabile, cioè la probabilità di successo, in funzione di *p* variabili esplicative *x1,...,xp.* La variabile risposta Y può, dunque, assumere i valori {0,1} e segue la distribuzione di Bernoulli. (Biostatistica - Ventura,Racugno,2017)

Essendo la variabile risposta (***sigdz***) una variabile qualitativa dicotomica (0=assenza, 1=presenza) si è deciso, quindi, di adattare un modello di regressione logistica di Bernoulli con link function canonica (logit).

Il modello teorico iniziale è:

- $Y_i \sim Ber(\pi_i)$ indipendenti i=1,...,3501
- $\eta_i = \beta 0 + \beta 1 * x_{i1} + \beta 2 * x_{i2} + \beta 3 * x_{i3}$
- funzione di legame canonica (logit): $\text{logit}(\pi i) = \log(\frac{\pi i}{1-\pi i}) \Rightarrow \pi i = (\frac{e^{\beta^T xi}}{1+e^{\beta^T xi}})$

Le variabili sono:

- Y_i = variabile risposta ***sigdz*** per modellare la probabilità di malattia coronarica significativa.
- x_{i1} = variabile dicotomica ***sex*** che individua il sesso del paziente e assume valore 0 se è maschio e 1 se è femmina.
- x_{i2} = variabile quantitativa ***age*** che indica l'età del paziente.
- x_{i3} = variabile quantitativa ***cad.dur*** che indica la durata dei sintomi in giorni.

Dall'analisi bivariata emerge, infatti, che tutte le covariate hanno un effetto significativo sulla variabile risposta, perciò sono tutte incluse nel modello di partenza. Partendo dal modello iniziale di adattamento del modello si nota che la variabile ***cad.dur***, riferita al coefficiente β3, risulta non essere significativa. Infatti, tramite un'analisi all'indietro (*backward*), utilizzando il test Chi-quadrato, è stata rimossa la variabile ***cad.dur.*** Si arriva, così, al modello finale (tabella 3.1.1), in cui tutti i coefficienti risultano significativi (p-value < 0.05).

Tab. 3.1.1: Tabella di adattamento del modello finale

	Stime	Intervallo di Confidenza	Std. Error	z-value	p-value
β0 = Int.	-2.575	(-3.013 , -2.143)	0.222	-11.60	<0.0001
β1 = sex(1)	-1.937	(-2.112 , -1.765)	0.088	-21.92	<0.0001
β2 = age	0.077	(0.068 , 0.086)	0.005	17.16	<0.0001
Devianza nulla	4459.6 on 3500 df				
Devianza residua	3694.9 on 3498 df				
AIC:	3700.9				

E' possibile notare che la stima del coefficiente ***sex*** è negativa e da una prima analisi si può dedurre una probabilità di malattia significativa più alta nei maschi (***sex***=0) rispetto alle femmine (***sex***=1). Allo stesso per la variabile ***age*** il coefficiente positivo indica che la probabilità di malattia significativa cresce all'aumentare dell'età. (Si analizzerà più nello specifico nella tabella 3.1.3)

A seguito dell'adattamento ne risulta il seguente modello finale stimato:

$\text{logit}(\widehat{\pi i}) = -2.575 - 1.937*\text{sex} + 0.077*\text{age}$

$$\widehat{\pi i} = \frac{exp(-2.575 - 1.937*sex + 0.077*age)}{1 + exp(-2.575 - 1.937*sex + 0.077*age)}$$

Modello 0: sigdz ~ Int.
Modello 1: sigdz ~ sex + age

	Resid. df	Resid. Dev	Df	Deviance	p-value
M0 (mod. sola int.)	3500	4459.6			
M1 (mod. corrente)	3498	3694.9	2	764.7	<0.0001

Tab. 3.1.2: Tabella di confronto con modello sola intercetta

Inoltre, come mostrato nella tabella 3.1.2 è possibile anche confrontare il modello corrente con il modello sola intercetta, verificando l'ipotesi di omogeneità contro il modello corrente tramite il test Chi-quadrato. Dalle analisi si nota che con un

W = (4459.6 - 3694.9) = 764.7 e un p-value <0.0001 si rifiuta l'ipotesi del modello con sola intercetta (Modello 0), preferendo così il modello corrente (stimato nella tabella 3.1.1).

Nel modello attuale, si osserva che la devianza residua (3694.9) è maggiore dei suoi gradi di libertà (3498). Questa specifica condizione non permette di affermare il perfetto adattamento del modello corrente ai dati osservati, anche se potrebbe aver catturato la complessità degli stessi. Nei prossimi passi dell'analisi, si esaminerà più approfonditamente le performance del modello attraverso l'utilizzo di ulteriori metodi di valutazione (come l'accuratezza, la curva ROC con conseguente AUC e la valutazione dei residui) che consentiranno di valutare la validità delle previsioni e ottenere una visione più completa e dettagliata delle caratteristiche del modello.

Le stime dei parametri di regressione sono interpretabili come il logaritmo del rapporto delle quote. Gli *odds* con i relativi intervalli di confidenza per la previsione della presenza della malattia significativa (***sigdz***) sono riassunti nella tabella seguente.

	exp	IDC
sex(1)	exp(-1.937) = 0.144	(0.122 , 0.171)
age	exp(0.077) = 1.08	(1.07 , 1.09)

Tab.3.1.3: Tabella con quote per le relative variabili (***sex*** e ***age***)

Dalla tabella 3.1.3 emergono le seguenti conclusioni:

- Il valore per ***sex*** pari a 0.144 indica che la quota di probabilità della presenza di malattia significativa (***sigdz***) per le femmine è inferiore del 85.6% rispetto ai maschi (mantenendo costanti le altre variabili).
- Il valore per ***age*** pari a 1.08 sta ad indicare che, per ogni incremento unitario dell'età, la quota di probabilità della presenza della malattia significativa aumenta dell'8% (sempre tenendo costanti le altre variabili).

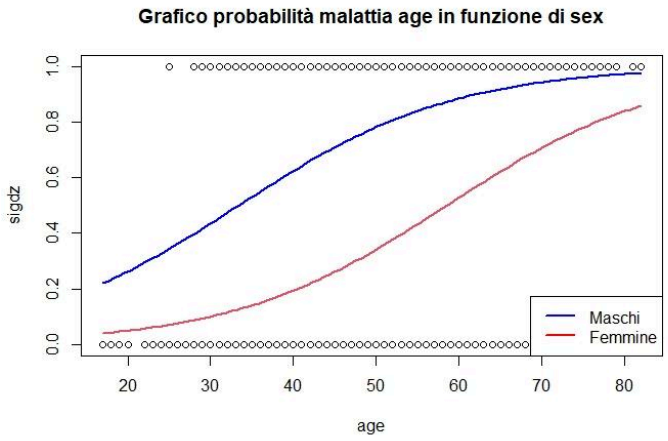


Fig. 3.1.1: Grafico della probabilità di malattia significativa (***sigdz***) in funzione di ***age*** per entrambi i sessi (***sex***)

Le precedenti interpretazioni delle quote possono essere visibili nella figura a lato (3.1.1) che mostra la probabilità di malattia significativa (***sigdz***) in funzione dell'età (***age***) divisa per sesso (***sex***). Come si può notare la probabilità di malattia per i maschi (curva blu) è maggiore rispetto alle femmine e aumentando l'età di entrambi i sessi la probabilità di malattia sale.

Di seguito viene riportata la tabella di corretta classificazione (3.1.4) che mostra come il modello adottato classifica correttamente i pazienti sani e i pazienti malati.

sigdz	FALSE	TRUE
0	559	610
1	268	2064

Tab.3.1.4: Tabella di corretta classificazione

- L'*accuratezza* del test è pari a $\frac{2064+559}{559+610+268+2064}$ = 0.749 = 74.9 %
- La *sensibilità* del test è pari a $\frac{2064}{2064+268}$ = 0.885 = 88.5 % , questo vuol dire che si sbaglia dell'11.5 % (0.115) a classificare i malati.

- La *specificità* del test è pari a $\frac{559}{610+559}$ = 0.478 = 47.8 %, questo vuol dire che si sbaglia del 52.2 % (0.521) a classificare i sani.
- La *prevalenza* della malattia è pari a $\frac{2064+268}{559+610+268+2064}$ = 0.666 = 66.6 %

E' possibile affermare che l'accuratezza e la sensibilità sono buoni, mentre la specificità ha un'alta percentuale di errore nel classificare correttamente i sani.

Inoltre, è possibile stabilire la capacità predittiva del modello anche utilizzando la curva ROC ("Receiver Operating Characteristic"), che consente il confronto tra i valori osservati della risposta e i valori stimati del modello.

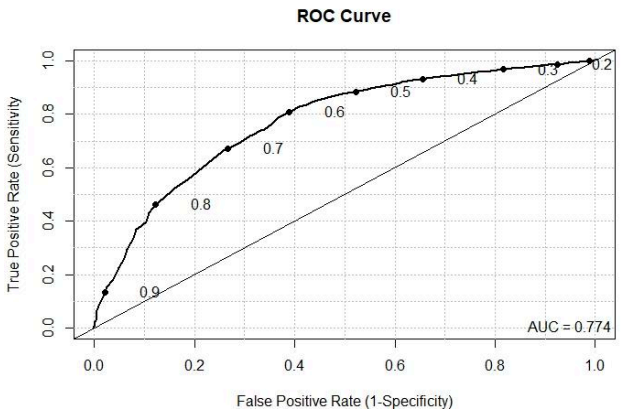


Fig. 3.1.2: Curva ROC per l'adattamento del modello e area sotto la curva (AUC)

In Figura 3.1.2 è riportato il grafico della curva ROC per i valori stimati contro la variabile risposta; la diagonale rappresenta la bisettrice. La curva appare lontana da essa, indice del fatto che i valori previsti stimano abbastanza bene i valori osservati. L'AUC in questo contesto valuta la bontà dell'adattamento del modello: un valore di AUC prossimo a 1 suggerisce che il modello ha un buon adattamento ai dati, mentre un valore vicino a 0.5 indica che il modello non si adatta meglio di un modello casuale. Questo concetto si allinea con l'ipotesi nulla nel test di Mann-Whitney, in cui l'uguaglianza di distribuzione tra gruppi è l'equivalente della mancanza di adattamento del modello alla relazione tra predittori e risposta nella regressione logistica. In questo specifico caso il p-value <0.0001 rifiuta tale ipotesi e l'area sottesa alla curva ROC (AUC, "Area Under Curve") è pari a 0.774, quindi il test risulta moderatamente accurato tra valori stimati e osservati.

Successivamente, una specifica misura di bontà di adattamento utilizzabile nella regressione logistica è la statistica di Hosmer-Lemeshow. In questo caso, un $\chi 2$ HL = 18.526 con p-value= 0.0176 indica una scarsa adattabilità del modello ai dati (a un livello di significatività 0.01 risulterebbe migliore).

L'ultimo elemento che viene riportato per la validazione del modello finale stimato è l'analisi dei residui.

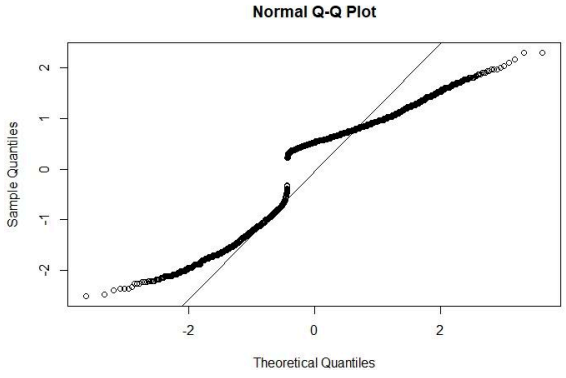


Figura 3.1.3: Diagramma quantile-quantile dei residui del modello finale stimato.

Dal diagramma quantile-quantile dei residui del modello (Fig. 3.1.3) emerge la presenza di outliers in entrambe le code. Inoltre, si evidenzia un allontanamento di molti punti dalla retta di riferimento per l'adattamento alla normalità. Infatti, tramite il test di Shapiro-Wilk che fornisce un W = 0.880 con p-value <0.0001, si rifiuta l'ipotesi di normalità. Anche il distacco in corrispondenza del valore zero nelle ordinate potrebbe indicare una deviazione dalla distribuzione normale dei residui. Nel contesto della regressione logistica ciò è abbastanza comune, poiché la regressione logistica produce residui che possono non essere normalmente distribuiti. Questo è dovuto alla natura logistica della trasformazione di probabilità. Quando la risposta è dicotomica, infatti, l'esame dei residui risulta complicato.

4 Ulteriori considerazioni

Un'altra analisi importante è la previsione della probabilità di malattia coronarica grave (***tvdlm***) nei pazienti in cui è stata "confermata" una qualche malattia significativa (***sigdz***).
E' proprio in questo caso, dunque, che entra in gioco la variabile ***tvdlm:*** una variabile dicotomica che rappresenta la presenza o assenza della malattia dell'arteria coronaria principale sinistra o dei tre vasi riscontrata a livello cardiaco (assume valore 0 per l'assenza e 1 per la presenza). Si andrà, quindi, a lavorare con ***tvdlm*** come variabile risposta, in relazione però ai soli pazienti a cui è stata riscontrata la presenza di malattia significativa ***sigdz***, usufruendo quindi della probabilità condizionata P(***tvdlm*|*sigdz*** = 1). Di conseguenza si lavora con una numerosità campionaria di 2332 pazienti. Anche in questa circostanza, si cercherà di comprendere come le variabili ***sex***, ***age*** e ***cad.dur*** influenzano la variabile risposta. A seguito dell'introduzione, per comodità nella spiegazione, da ora in poi il condizionamento (***sigdz***=1) sarà sottinteso.

4.1 Analisi univariate

Come in precedenza, verranno esaminate le variabili del dataset, però questa volta, con numerosità campionaria pari a 2332.

Tab.4.1.1: Distribuzione di frequenze assolute e relative percentuali per la variabile ***tvldm*** a sinistra e per la variabile ***sex*** a destra

	N	%
tvdlm		
0 = assenza	1203	51.6
1= presenza	1129	48.4
Totale	2332	100

	N	%
sex		
0 = maschio	1871	80.2
1= femmina	461	19.8
Totale	2332	100

Dalla tabella 4.1.1 a sinistra è possibile osservare che all'interno del campione i pazienti con malattia coronarica grave (***tvdlm***) sono 1129, mentre quelli senza sono 1203, rispettivamente il 48.4% e il 51.6%. Eseguendo il test sulle proporzioni è risultato un X-squared pari a 2.285 ed un p-value = 0.1306. Si accetta, quindi, l'ipotesi nulla concludendo che le due proporzioni sono significativamente uguali.

Dalla tabella 4.1.1 a destra analizzando la variabile ***sex*** si nota che all'interno del campione i pazienti maschi sono 1871 e le femmine 461, rispettivamente l'80.2% e il 19.8%.
In questa circostanza svolgendo il test sulle proporzioni è risultato un X-squared = 851.32 e un p-value <0.0001. Si rifiuta l'ipotesi nulla concludendo questa volta che le due proporzioni sono significativamente diverse. Il rapporto "maschi/femmine" è pari a 4.06, deducendo che i maschi sono presenti 4 volte di più rispetto alle femmine.

	Min.	1st.Qu.	Median	Mean	3Rd. Qu.	Max.	S.d.	IQR
age	25.00	47.00	54.00	54.02	61.00	82.00	9.51	14.00
cad.dur	0.00	5.00	22.00	46.26	66.00	416.00	59.93	61.00

Tab.4.1.2:Statistiche di sintesi delle variabili ***age*** e ***cad.dur***

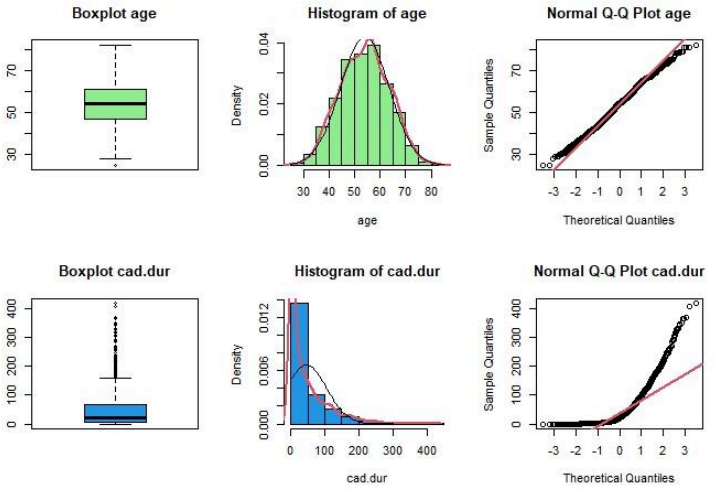


Fig.4.1.1: Boxplot, istogramma e diagramma quantile-quantile per la variabile ***age*** in alto e per la variabile ***cad.dur*** in basso.

Passando ora alla variabile ***age*** dalla figura 4.1.1 in alto, si nota che il boxplot presenta un outlier nella coda inferiore e la distribuzione nell'istogramma appare simmetrica. Dalla visualizzazione del diagramma quantile-quantile emerge un leggero allontanamento di entrambe le code (Figura 4.1.1 in alto a destra). Infatti, la statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità (W = 0.996, p-value < 0.0001), ma avendo un

campione con numerosità campionaria elevata è possibile invocare il Teorema del Limite Centrale che permette di proseguire accettando l'ipotesi di normalità nelle medie.
Per quanto riguarda la variabile ***cad.dur***, invece, osservando la figura 4.1.1 in basso, emerge che il boxplot presenta molti outliers nella coda superiore, confermando la forte asimmetria anche nell'istogramma a fianco. Dal diagramma quantile-quantile si nota un forte scostamento dei punti dalla retta di riferimento (Figura 4.1.1 in basso a destra). Infatti, la statistica test di Shapiro-Wilk porta a rifiutare l'ipotesi di normalità (W = 0.753, p-value<0.0001). Il caso specifico inerente all'applicazione del Teorema del Limite Centrale verrà presentato nella prossima sezione delle analisi bivariate (4.2) in particolare nelle figure 4.4.3 e 4.4.5.

4.2 Analisi bivariate

Come accennato in precedenza, in questa analisi tutte le variabili prese in considerazione sono condizionate ai soli pazienti con presenza di malattia significativa (***sigdz***=1).

Fig.4.2.1: Tabelle di sintesi per l'analisi bivariata del secondo modello

		Test Chi-quadrato		Test di Fisher		
		X-squared	p-value	p-value	OR	IDC OR
a)	tvdlm/sex	13.024	0.0003	0.0003	0.68	(0.55 , 0.84)

a.1)

tvdlm	*sex*	
	0=maschi	1=femmine
0=assenza	930	273
1=presenza	941	188

		TEST PARAMETRICO										TEST NP	
		Test di Shapiro-Wilk				Test sulle varianze		Test-t		Test di Welch		Test di Mann-Whitney	
		W0	p-value0	W1	p-value1	F	p-value	t	p-value	t	p-value	W	p-value
b)	tvdlm/age	0.996	0.0016	0.996	0.0057	1.043	0.473	-8.884	<0.0001	-	-	538352	<0.0001
c)	tvdlm/cad.dur	0.712	<0.0001	0.802	<0.0001	0.520	<0.001	-	-	-9.836	<0.0001	503767	<0.0001
d)	sex/age	0.996	<0.0001	0.990	0.0038	0.870	0.0524	-8.467	<0.0001	-	-	323211	<0.0001
e)	sex/cad.dur	0.760	<0.0001	0.719	<0.0001	0.986	0.8407	0.316	0.7522	-	-	4323288	0.5377

		Test di correlazione di Pearson			Test di correlazione di Spearman		
		t	p-value	corr	S	p-value	corr-S
f)	age/cad.dur (M)	14.047	<0.0001	0.309	527109739	<0.0001	0.236
g)	age/cad.dur (F)	5.277	<0.0001	0.239	49582457	<0.0001	0.223

a) La prima analisi bivariata riguarda la variabile ***tvdlm*** e la variabile ***sex***, svolgendo il test del chi-quadro risulta una forte associazione significativa tra il sesso e la presenza della malattia coronarica grave (***tvdlm***). Anche il test di Fisher va a confermare tale conclusione, e l'*Odds Ratio* pari a 0.68 suggerisce per le femmine una quota inferiore del 32% rispetto ai maschi.

a.1) - proporzione di maschi con malattia grave = 0.50
- proporzione di femmine con malattia grave = 0.41

Queste due proporzioni sono state ottenute anche dal test sulle proporzioni che con p-value 0.0003 indica una differenza significativa tra i due gruppi di pazienti per sesso.

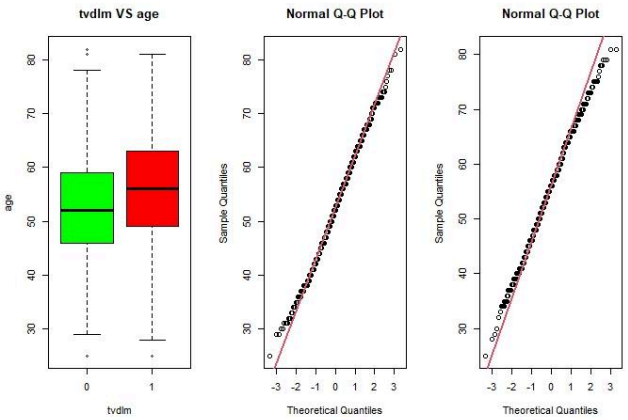


Fig. 4.4.2: Boxplot ***tvdlm*** per ***age*** a sinistra, diagramma quantile-quantile sani al centro, diagramma quantile-quantile malati a destra

b) Si studia ora l'effetto dell'età (***age***) sulla presenza/assenza della malattia coronarica grave (***tvdlm***). Si nota dalla figura 4.4.2 a sinistra, che in entrambi i boxplot vi sono alcuni outliers e la mediana è maggiore nel gruppo dei malati. L'ipotesi di normalità attraverso il test di Shapiro-Wilk viene rifiutata in entrambi i gruppi, infatti, osservando i diagrammi quantile-quantile (Figura 4.4.2 al centro e a destra) si nota un leggero allontanamento delle code dalla retta di riferimento. Avendo però una numerosità campionaria molto elevata è possibile, attraverso il Teorema del Limite Centrale,

proseguire comunque con l'ipotesi di normalità delle medie. Successivamente, il test sulle varianze accetta l'ipotesi di omoschedasticità nei due gruppi. Infine, con il test-t a due campioni indipendenti si va a rifiutare l'ipotesi di uguaglianza, concludendo così che vi è una differenza significativa tra le medie nei due gruppi. Per un ulteriore controllo, volendo non accettare l'ipotesi di normalità, è possibile passare direttamente al test non-parametrico di Mann-Whitney che porta comunque alla stessa conclusione di rifiuto di uguaglianza tra le due distribuzioni. E' presente, quindi, un effetto significativo dell'età sulla malattia grave.

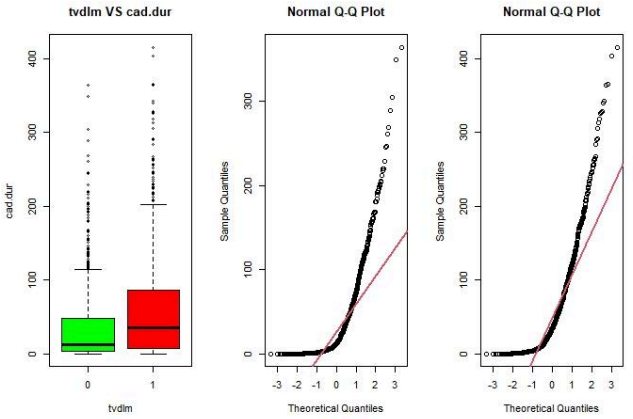


Fig. 4.4.3: Boxplot ***tvdlm*** per ***cad.dur*** a sinistra, diagramma quantile-quantile sani al centro, diagramma quantile-quantile malati a destra

c) Si studia ora l'effetto della durata (***cad.dur***) sulla presenza/assenza della malattia coronarica grave (***tvdlm***). Dai boxplot (Fig. 4.4.3 a sinistra) si nota che vi sono molti outliers in entrambe le code superiori e la mediana è maggiore nel gruppo dei malati. L'ipotesi di normalità viene rifiutata in entrambi i gruppi e osservando i diagrammi quantile-quantile (Figura 4.4.3 al centro e a destra), infatti, si nota un forte allontanamento dalla retta di riferimento. Per far fronte a questo, è possibile ricorrere al test non-parametrico di Mann-Whitney che porta a rifiutare l'ipotesi di uguaglianza in mediana per i due gruppi, concludendo che le mediane delle due distribuzioni per la variabile ***cad.dur*** sono significativamente diverse. Volendo passare comunque attraverso un metodo parametrico, vista la numerosità campionaria elevata, si usufruisce come in precedenza del Teorema del Limite Centrale che porta alla medesima conclusione tramite il test di Welch (non avendo l'ipotesi di omoschedasticità).

	Min.	1st.Qu.	Median	Mean	3Rd. Qu.	Max.	S.d.	IQR
0 = Maschi	28.00	47.00	53.00	53.20	60.00	82.00	9.24	13.00
1 = Femmine	25.00	51.00	58.00	57.33	65.00	81.00	9.90	14.00

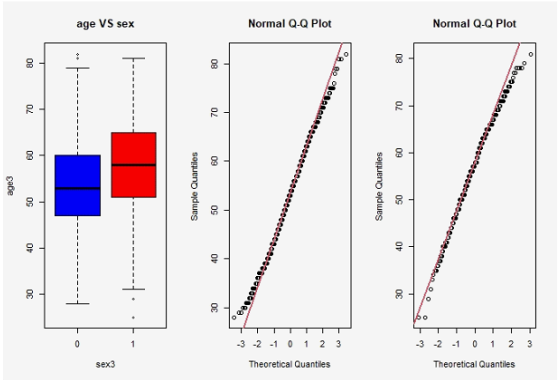


Fig. 4.4.4: In alto tabella per le statistiche di sintesi della variabile ***age***, in basso Boxplot ***age*** per ***sex*** (a sinistra), diagramma quantile-quantile maschi (al centro), diagramma quantile-quantile femmine (a destra)

d) Si studia ora la relazione tra l'età (***age***) e il sesso (***sex***) dei pazienti. Dai boxplot (Fig. 4.4.4 in basso a sinistra) si nota che vi sono alcuni outliers in entrambi i sessi e la mediana per le femmine è maggiore. L'ipotesi di normalità viene rifiutata in entrambi i gruppi e osservando i diagrammi quantile-quantile (Figura 4.4.4 in basso al centro e a destra), infatti, si nota un leggero allontanamento delle code dalla retta di riferimento. Anche questa volta, è possibile proseguire comunque con l'ipotesi di normalità delle medie tramite l'utilizzo del Teorema del Limite Centrale.
In seguito, il test sulle varianze accetta l'ipotesi di omoschedasticità nei due gruppi e, tramite il test-t a due campioni indipendenti, si va a rifiutare l'ipotesi di uguaglianza concludendo così che vi è una differenza significativa tra le medie dell'età nei due gruppi di pazienti.
Volendo rifiutare l'ipotesi di normalità è possibile passare direttamente al test non-parametrico di Mann-Whitney che porta comunque alla stessa conclusione di rifiuto.

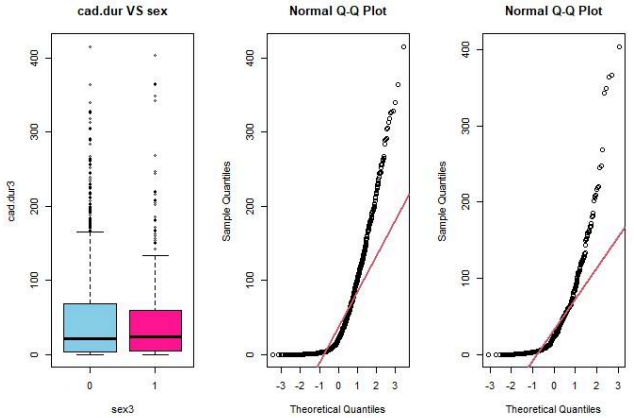


Fig. 4.4.5: Boxplot ***cad.dur*** per ***sex*** a sinistra, diagramma quantile-quantile maschi al centro, diagramma quantile-quantile femmine a destra

e) A questo punto, si esamina la relazione tra la durata dei sintomi (***cad.dur***) e il sesso (***sex***) dei pazienti. Dai boxplot (Figura 4.4.5 a sinistra) si nota che vi sono molti outliers in entrambe le code superiori e la mediana è circa uguale per maschi e femmine. Osservando i diagrammi quantile-quantile (Figura 4.4.5 al centro e a destra) si nota un forte allontanamento dei punti dalla retta di riferimento, infatti, l'ipotesi di normalità viene rifiutata in entrambi i gruppi. Come nel caso (*c*) è possibile ricorrere al test non-parametrico di Mann-Whitney, che porta questa volta ad accettare l'ipotesi di uguaglianza in mediana per i due gruppi. Si conclude così, che le mediane per la variabile ***cad.dur*** nelle distribuzioni di maschi e femmine, non sono significativamente diverse.
Nel caso si volesse comunque proseguire accettando la normalità delle medie, tramite il Teorema del Limite Centrale, si nota che si arriva alla stessa conclusione di uguaglianza per le due distribuzioni omoschedastiche.

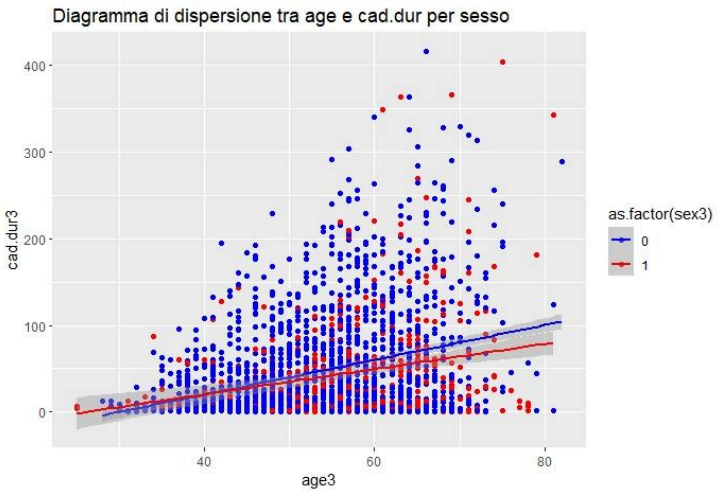


Fig. 4.4.6: Diagramma di dispersione tra ***age*** e ***cad.dur*** diviso per sesso (***sex***)

f) - g) Come ultima analisi bivariata della tabella in figura 4.2.1, è possibile osservare la relazione tra le due variabili quantitative del dataset: ***age*** e ***cad.dur***. Osservando quindi il diagramma di dispersione a lato, è possibile notare una lieve relazione tra le due variabili. Infatti, con l'aumentare dell'età dei pazienti per entrambi i sessi si nota anche un aumento della durata dei sintomi. Osservando la tabella f)-g) in figura 4.2.1 è possibile concludere che il test di correlazione di Pearson mostra correlazioni positive e significativamente diverse da zero sia per i maschi che per le femmine. Vista la presenza di outliers si applica anche il test di correlazione di Spearman (più robusto) che, anche in questo caso, porta a concludere che entrambe le correlazioni sono positive e significativamente diverse da zero. Esiste, quindi, una debole relazione tra ***age*** e ***cad.dur*** in entrambi i sessi.

4.3 Stima del modello

Per modellare la presenza della malattia coronarica più grave (***tvdlm***) in funzione del sesso (***sex***), dell'età (***age***) e della durata dei sintomi dei pazienti (***cad.dur***), si adatta un modello di regressione logistica sulla base degli studi precedentemente svolti.

4.4 Modello di regressione logistica

Poiché la variabile risposta (***tvdlm***) è una variabile qualitativa dicotomica (0=assenza,1=presenza) si è deciso, anche in questo caso, di adattare un modello di regressione logistica di Bernoulli con link function canonica (logit).

Il modello teorico iniziale è:

- $Y_i \sim Ber(\pi_i)$ indipendenti i=1,...,2332
- $\eta_i = \beta 0 + \beta 1 * x_{i1} + \beta 2 * x_{i2} + \beta 3 * x_{i3}$
- funzione di legame canonica (logit): $\text{logit}(\pi i) = \log(\frac{\pi i}{1-\pi i}) \Rightarrow \pi i = (\frac{e^{\beta^T xi}}{1+e^{\beta^T xi}})$

Le variabili sono:

- Y_i = variabile risposta ***tvdlm*** per modellare la probabilità di malattia coronarica grave.
- x_{i1} = variabile dicotomica ***sex*** che individua il sesso del paziente e assume valore 0 se è maschio e 1 se è femmina.
- x_{i2} = variabile quantitativa ***age*** che indica l'età del paziente.
- x_{i3} = variabile quantitativa ***cad.dur*** che indica la durata dei sintomi in giorni.

Dall'analisi bivariata emerge, infatti, che tutte le covariate hanno un effetto significativo sulla variabile risposta, perciò sono tutte incluse nel modello di partenza.

Tab. 4.4.1: Tabella di adattamento del modello finale

	Stime	Intervallo di Confidenza	Std. Error	z-value	p-value
β0 = Int.	-2.079	(-2.588 , -1.578)	0.257	-8.076	<0.0001
β1 = sex(1)	-0.546	(-0.766 , -0.329)	0.111	-4.899	<0.0001
β2 = age	0.034	(0.025 , 0.044)	0.005	7.042	<0.0001
β3 = cad.dur	0.006	(0.004 , 0.007)	0.001	7.220	<0.0001
Devianza nulla	3230.5 on 2331 df				
Devianza residua	3066.8 on 2328 df				
AIC:	3074.8				

Dalla tabella 4.4.1 di adattamento del modello si nota che i parametri, riferiti alle corrispondenti variabili, sono tutti significativi (p-value < 0.05). L'analisi all'indietro (*backward*) effettuata conferma che il modello finale corrisponde a quello iniziale. In questo caso, a differenza del primo modello per la malattia coronarica significativa (***sigdz***), la variabile ***cad.dur*** rimane nel modello.

Da una prima analisi è possibile notare che la stima negativa del coefficiente ***sex***, indica una probabilità di malattia significativa più alta nei maschi (***sex***=0) rispetto alle femmine (***sex***=1). Per la variabile ***age*** il coefficiente positivo evidenzia che la probabilità di malattia significativa cresce all'aumentare dell'età e, analogamente, per la variabile ***cad.dur*** si può applicare lo stesso ragionamento, ma lo si nota in modo meno accentuato. (Il tutto sarà analizzato più nello specifico nella tabella 4.4.3).

A seguito dell'adattamento ne risulta il seguente modello finale stimato:

$\text{logit}(\widehat{\pi i}) = -2.079 - 0.546*\text{sex} + 0.034*\text{age} + 0.006*\text{cad.dur}$

$$\widehat{\pi i} = \frac{exp(-2.079 - 0.546*sex + 0.034*age + 0.006*cad.dur)}{1 + exp(-2.079 - 0.546*sex + 0.034*age + 0.006*cad.dur)}$$

Modello 0: tvdlm ~ Int.
Modello 1: sigdz ~ sex + age + cad.dur

	Resid. df	Resid. Dev	Df	Deviance	p-value
M0 (mod. sola int.)	2331	3230.5			
M1 (mod. corrente)	2328	3066.8	3	163.7	<0.0001

Tab. 4.4.2: Tabella di confronto con modello sola intercetta

Inoltre, come mostrato nella tabella precedente 4.4.2, è possibile confrontare il modello corrente con il modello sola intercetta, verificando l'ipotesi di omogeneità contro il modello corrente tramite il test Chi-quadrato. Si nota che con un W = (3230.5 - 3066.8) = 163.7 e un p-value <0.0001 si rifiuta l'ipotesi del modello con sola intercetta (Modello 0), preferendo così il modello corrente (stimato in precedenza nella tabella 4.4.1). L'analisi della devianza per il modello corrente, rivela che la devianza residua (3066.8) è maggiore dei suoi gradi di libertà (2328). Questa condizione non permette di affermare il perfetto adattamento del modello corrente ai dati osservati. Nei prossimi passi dell'analisi si esaminerà più approfonditamente le performance del modello attraverso l'utilizzo di ulteriori metodi di valutazione (tra cui l'accuratezza, l'AUC e la valutazione dei residui) che consentiranno di valutare la validità delle previsioni e ottenere una visione più completa e dettagliata delle caratteristiche del modello.

Inoltre, anche questa volta le stime dei parametri di regressione sono interpretabili come il logaritmo del rapporto delle quote. Gli *odds* con i relativi intervalli di confidenza per la previsione della presenza di malattia grave (***tvdlm***) sono riassunti nella seguente tabella.

	exp	IDC
sex(1)	exp(-0.546) = 0.579	(0.466 , 0.720)
age	exp(0.034) = 1.035	(1.024, 1.045)
cad.dur	exp(0.066) = 1.006	(1.004 , 1.008)

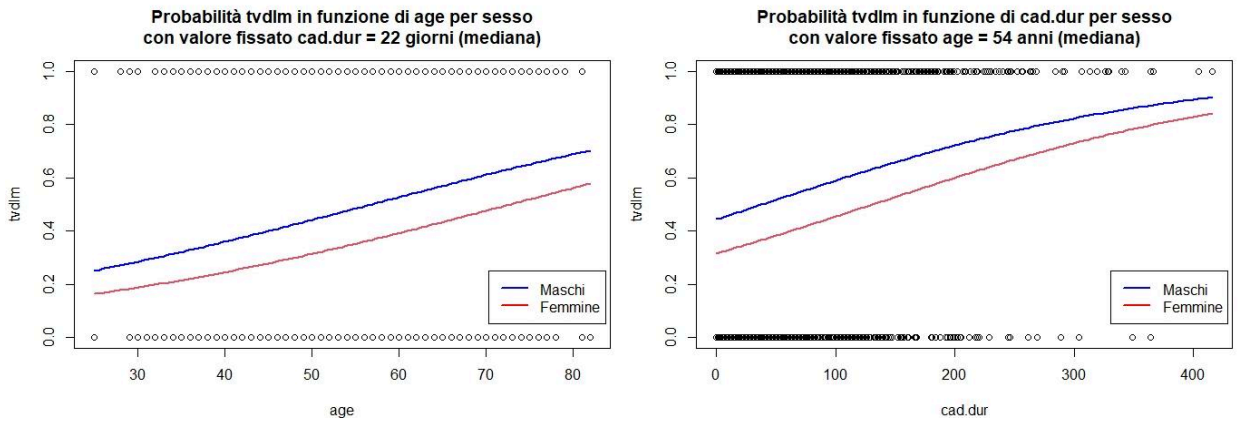
Tab.4.4.3: Tabella con quote per le relative variabili (***sex***, ***age*** e ***cad.dur***)

Dalla tabella 4.4.3 emergono le seguenti conclusioni:

- il valore per ***sex*** pari a 0.579 indica che la quota di probabilità della presenza di malattia grave (***tvdlm***) per le femmine è inferiore del 42.1% rispetto ai maschi (tenendo costanti le altre variabili).
- il valore per ***age*** pari a 1.035 sta ad indicare che, per ogni incremento unitario dell'età, la quota di probabilità della presenza della malattia grave aumenta del 3.5% (mantenendo costanti le altre variabili).
- il valore per ***cad.dur*** pari a 1.006 sta ad indicare che, per ogni incremento unitario della durata dei sintomi, la quota di probabilità della presenza della malattia grave aumenta dello 0.6%. (sempre mantenendo costanti le altre variabili).

Le precedenti interpretazioni delle quote possono essere visibili nella figura seguente (4.4.1) che mostra a sinistra la probabilità di malattia grave (***tvdlm***) in funzione dell'età (***age***) divisa per sesso (***sex***), e analogamente, in funzione della durata dei sintomi (***cad.dur***) a destra. Come si può notare la probabilità di malattia per i maschi (curva blu) è maggiore rispetto alle femmine in entrambi i grafici. Inoltre, aumentando a sinistra l'età e a destra la durata dei sintomi, si nota che in entrambi i sessi la probabilità di malattia cresce.

Fig. 4.4.1: Grafico della probabilità di malattia grave (***tvdlm***) nei due casi



Di seguito viene riportata la tabella di corretta classificazione che mostra come il modello adottato classifica correttamente i pazienti sani e i pazienti malati.

tvdlm	FALSE	TRUE
0	841	362
1	546	583

Tab.4.4.4: Tabella di corretta classificazione

- L'*accuratezza* del test è pari a $\frac{583+841}{841+362+546+583}$ = 0.61 = 61 %
- La *sensibilità* del test è pari a $\frac{583}{583+546}$ = 0.516 = 51.6 % , questo vuol dire che si sbaglia del 48.4 % (0.484) a classificare i malati.
- La *specificità* del test è pari a $\frac{841}{362+841}$ = 0.699 = 69.9 % , questo vuol dire che si sbaglia del 30.1 % (0.301) a classificare i sani.
- La *prevalenza* della malattia è pari a $\frac{583+546}{841+362+546+583}$ = 0.484 = 48.4 %

In linea generale si nota che l'accuratezza e la specificità sono discretamente accettabili, mentre la sensibilità ha un'alta percentuale di errore nel classificare correttamente i malati.

Analogamente anche in questo caso, si valuta la capacità predittiva del modello utilizzando la curva ROC ("Receiver Operating Characteristic"), che consente il confronto tra i valori osservati della risposta e i valori stimati del modello.

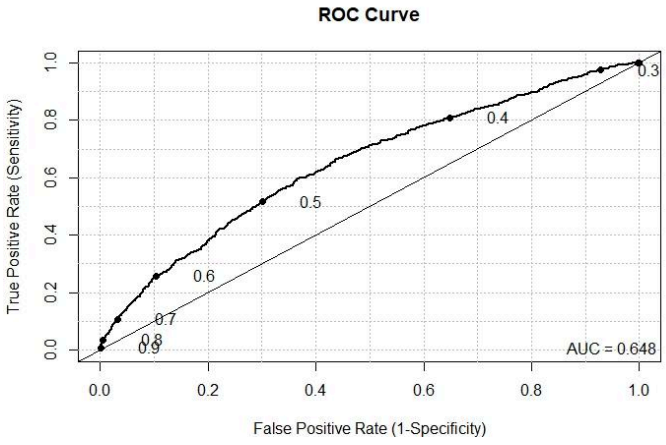


Fig. 4.4.2: Curva ROC per l'adattamento del modello e area sotto la curva (AUC)

In figura 4.4.2 a lato, è riportato il grafico della curva ROC per i valori stimati contro la variabile risposta. La curva appare sufficientemente spostata dalla bisettrice, indice del fatto che i valori previsti stimano abbastanza bene i valori osservati. In questo caso, l'area sottesa alla curva ROC (AUC, "Area Under Curve") è pari a 0.648, quindi il test risulta poco accurato tra valori stimati e osservati. Anche in questo caso specifico il p-value risulta <0.0001. Si rifiuta dunque l'ipotesi nulla del test di Mann-Whitney, in cui l'accettazione dell'uguaglianza di distribuzione tra gruppi si tradurrebbe in una mancanza di adattamento del modello alla relazione tra predittori e risposta.

La statistica di Hosmer-Lemeshow con un $\chi 2$ HL = 5.838 con p-value = 0.6654 indica una buona adattabilità del modello tra frequenze stimate e osservate.

L'ultimo elemento che viene riportato per la validazione del modello finale stimato è l'analisi dei residui.

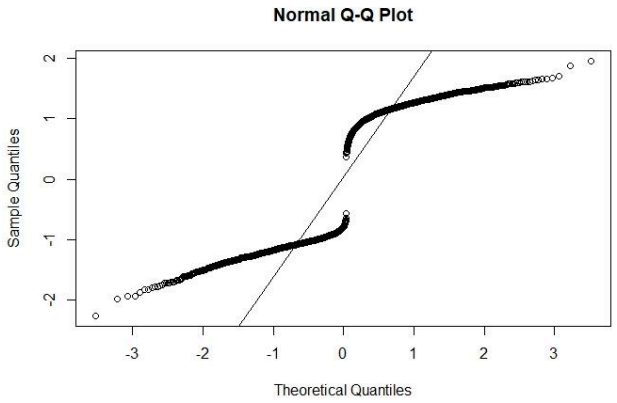


Figura 4.4.3: Diagramma quantile-quantile dei residui del modello finale stimato

Il diagramma quantile-quantile dei residui del modello (Fig. 4.4.3 a lato) evidenzia la presenza di outliers in entrambe le code. Emerge, inoltre, un allontanamento di molti punti dalla retta per l'adattamento alla normalità di riferimento. Attraverso il test di Shapiro-Wilk che fornisce un W = 0.798 con p-value

<0.0001, infatti, si rifiuta l'ipotesi di normalità. Come nel modello precedente, anche in questo caso, il distacco in corrispondenza del valore zero nelle ordinate potrebbe segnalare una deviazione dalla distribuzione normale dei residui. Quando la risposta è dicotomica, infatti, l'esame dei residui risulta complesso.

5 Conclusioni

5.1 Conclusioni per il primo modello

Il primo modello si concentra sulla previsione della probabilità di malattia coronarica significativa (***sigdz***). La variabile ***sigdz*** è una variabile dicotomica che rappresenta la presenza o assenza della malattia coronarica significativa (0 per l'assenza e 1 per la presenza). Dopo aver rimosso dal dataset i tre pazienti con dati mancanti, l'analisi si è svolta con una numerosità campionaria di 3501 pazienti. Lo scopo dell'analisi è stato cercare di capire come le variabili ***sex***, ***age*** e ***cad.dur*** influenzano la presenza della malattia coronarica significativa. Come primo passo sono state eseguite le principali analisi esplorative per studiare la variabile risposta malattia coronarica significativa (***sigdz***). Dall'analisi bivariata, è emerso che le covariate età, sesso e durata dei sintomi hanno tutte un'effetto significativo sulla variabile risposta e, per questo motivo, sono state tutte introdotte nel modello. Per modellare la presenza della malattia si è deciso di adattare un modello di regressione logistica e, tramite l'analisi all'indietro, il modello finale è risultato con le sole covariate ***sex*** ed ***age***. E' stato adottato, dunque, un modello di regressione logistica con predittore lineare $\eta i = \beta 0 + \beta 1 * \textbf{\textit{sex}} + \beta 2 * \textbf{\textit{age}}$. E' emerso che la quota di presenza della malattia nelle femmine è dell' 85.6% inferiore rispetto ai maschi (tenendo costanti le altre variabili). In aggiunta, è risultato che la quota di presenza della malattia cresce dell'8% ad ogni aumento unitario dell'età (sempre mantenendo costanti le altre variabili). Per quanto riguarda la fase di validazione del modello, diversi indicatori sono stati utilizzati per comprendere la sua affidabilità e capacità predittiva. Per prima cosa, la devianza residua (3694.9) per il modello corrente non è risultata essere inferiore ai gradi di libertà associati (3498), pertanto questa condizione non permette di affermare il perfetto adattamento del modello corrente ai dati osservati. In seguito, l'accuratezza del modello è risultata buona, poiché il valore di 0.749 indica che il 74.9 % delle previsioni del modello è corretto rispetto alla variabile risposta. La curva ROC ha restituito un valore di AUC pari a 0.774 suggerendo che il modello fornisce una classificazione moderatamente accurata. Successivamente, la statistica di Hosmer-Lemeshow con un $\chi 2$ HL = 18.526 e p-value = 0.0176 ha indicato una scarsa adattabilità del modello tra frequenze stimate e osservate. Invece, l'osservazione del diagramma quantile-quantile dei residui ha mostrato un significativo allontanamento di diversi punti dalla retta di riferimento. Tale comportamento può essere attribuito alla natura logistica della trasformazione di probabilità in un modello di regressione logistica. Infatti l'esame dei residui risulta difficile quando si tratta di variabili dicotomiche. In conclusione, dalla combinazione di accuratezza, AUC, statistica di Hosmer-Lemeshow e analisi dei residui emerge che il modello dimostra una certa capacità predittiva e viene fornito un quadro complessivo della validità del modello, indicando aree di forza e potenziali miglioramenti. Durante la fase di individuazione del modello, si è considerata la possibilità di stimare un modello di regressione logistica, ma questa volta con tutte le covariate trasformate in variabili dicotomiche, incluse l'età (***age***) e la durata dei sintomi (***cad.dur***). Questa trasformazione è stata effettuata considerando la mediana come punto di riferimento. Per l'età si è assunto 0=età maggiore di 52 anni e 1=età minore di 52; mentre per la variabile durata 0=durata maggiore di 18 giorni e 1=durata minore di 18 giorni. Tuttavia, attraverso un'analisi all'indietro (*backward*), il modello finale ha confermato il modello di partenza, poiché tutti i parametri relativi alle covariate risultavano significativi. Nonostante ciò, l'indice di Akaike (AIC) di 3820.7 per il modello dicotomico è risultato superiore e quindi peggiore rispetto all'AIC di 3700.9 del modello finale principale, che includeva le variabili originali. Riflettendo sulla possibilità che la trasformazione in variabili dicotomiche potesse comportare la perdita di informazioni essenziali e quindi ridurre la precisione del modello, si è deciso di escludere il modello dicotomico. Si è preferito continuare con il modello finale principale che mostrava, comunque, una maggiore precisione con un AIC più basso e assicurava una conservazione di informazioni fondamentali nel modello. In conclusione, il modello migliore è risultato quello di regressione logistica con le variabili originali e

funzione di legame canonico (logit). Lo si è preferito anche allo stesso modello contenente, questa volta, la funzione di legame “probit”, poichè l’AIC pari a 3703.6 si è rivelato leggermente superiore a quello del modello scelto (3700.9) e non sono emersi miglioramenti significativi o vantaggi evidenti. In aggiunta, la scelta della funzione di legame “logit” è motivata anche dalla semplificazione nell’interpretazione delle stime di verosimiglianza e dal suo ampio utilizzo in ambito medico.

5.2 Conclusioni per il secondo modello

Si è ritenuto importante effettuare un’analisi della previsione della probabilità di malattia coronarica grave (***tvdlm***) nei pazienti in cui è stata "confermata" una qualche malattia significativa (***sigdz***). La variabile ***tvdlm*** è una variabile dicotomica che rappresenta la presenza o assenza della malattia dell'arteria coronaria principale sinistra o dei tre vasi riscontrata a livello cardiaco (assume valore 0 per l’assenza e 1 per la presenza). Oltre a ***tvdlm***, anche le altre variabili sesso (***sex***), età (***age***) e durata dei sintomi (***cad.dur***) sono state condizionate rispetto ai soli pazienti a cui è stata riscontrata la presenza di malattia significativa ***sigdz.*** Di conseguenza l’analisi si è svolta con una numerosità campionaria di 2332 pazienti. L’obiettivo è stato cercare di comprendere come le variabili ***sex***, ***age*** e ***cad.dur*** influenzano la presenza della malattia coronarica più grave nei pazienti che già soffrono della malattia coronarica significativa. Per lo studio della variabile risposta malattia coronarica grave (***tvdlm***), sono state eseguite le principali analisi esplorative. Dall’analisi bivariata, si è colto che le covariate età, sesso, durata dei sintomi hanno tutte un’effetto significativo sulla variabile malattia coronarica grave e, per questo motivo, sono state tutte introdotte nel modello. Per modellare la presenza della malattia coronarica più grave (***tvdlm***) si è deciso di adattare un modello di regressione logistica e, tramite l’analisi all’indietro, il modello finale è risultato corrispondere al modello iniziale. E’ stato adottato, dunque, un modello di regressione logistica con predittore lineare $\eta i = \beta 0 + \beta 1 * \textbf{\textit{sex}} + \beta 2 * \textbf{\textit{age}} + \beta 3 * \textbf{\textit{cad.dur}}$. E’ emerso che la quota di presenza della malattia più grave nelle femmine è il 42.1% inferiore rispetto ai maschi (con le altre variabili che si mantengono costanti). In aggiunta, è risultato che la quota di presenza della malattia grave aumenta del 3.5% ad ogni aumento unitario dell’età (sempre mantenendo le altre variabili costanti) e dello 0.6% per ogni incremento unitario della durata dei sintomi (anche in questo caso tenendo le altre variabili costanti). Per quanto riguarda la fase di validazione del modello, diversi indicatori sono stati utilizzati per comprendere la sua affidabilità e capacità predittiva. Per prima cosa, la devianza residua (3066.8) per il modello corrente non è risultata essere inferiore ai gradi di libertà associati (2328), pertanto questa condizione non permette di affermare il perfetto adattamento del modello corrente ai dati osservati. In seguito, l'accuratezza del modello è risultata accettabile, poiché il valore di 0.61 indica che il 61% delle previsioni del modello è corretto rispetto alla variabile risposta. La curva ROC ha restituito un valore di AUC pari a 0.648 suggerendo che il modello non fornisce una classificazione particolarmente accurata. Successivamente, la statistica di Hosmer-Lemeshow con un $\chi 2$ HL = 5.8381 e p-value = 0.6654 ha indicato una buona adattabilità del modello tra frequenze stimate e osservate. Invece, l’osservazione del diagramma quantile-quantile dei residui ha rivelato un significativo allontanamento di diversi punti dalla retta di riferimento. Anche in questo caso, come nel modello precedente il comportamento può essere attribuito alla natura logistica della trasformazione di probabilità in un modello di regressione logistica. Quando si tratta di variabili dicotomiche, infatti, l’esame dei residui risulta complicato. In conclusione, dalla combinazione di accuratezza, AUC, statistica di Hosmer-Lemeshow e analisi dei residui risulta che il modello dimostra una certa capacità di previsione e viene presentato un quadro complessivo della validità del modello, indicando aree di forza e potenziali miglioramenti. Anche in questo caso, si è considerata la possibilità di stimare un modello di regressione logistica, con tutte le covariate trasformate in variabili dicotomiche, comprese l'età (***age***) e la durata dei sintomi (***cad.dur***), sempre prendendo la mediana come punto di riferimento. Per l’età si è assunto 0=età maggiore di 52 anni e 1=età minore di 52; e per la durata 0=durata maggiore di 22 giorni e 1=durata minore di 22 giorni. Tuttavia, attraverso un'analisi all’indietro (*backward*), il modello finale è risultato: $\eta i = \beta 0 + \beta 1 * \textbf{\textit{sex(1)}} + \beta 2 * \textbf{\textit{age(1)}} + \beta 3 * \textbf{\textit{cad.dur(1)}}$, poiché tutti i parametri risultavano significativi. Nonostante ciò, l'indice di Akaike (AIC) di 3107.7 per il modello dicotomico è risultato superiore e quindi peggiore rispetto all'AIC di 3074.8 del modello principale con le variabili originali. Anche in questa

situazione, riflettendo sulla possibilità che la trasformazione in variabili dicotomiche potesse comportare la perdita di informazioni fondamentali e quindi ridurre la precisione del modello, si è deciso di escludere il modello dicotomico preferendo il modello principale. Con un AIC più basso assicurava, infatti, una conservazione di informazioni essenziali nel modello. In conclusione, il modello migliore è risultato quello di regressione logistica con le variabili originali e funzione di legame canonico (logit). Lo si è preferito anche allo stesso modello contenente, questa volta, la funzione di legame "probit", poichè l'AIC pari a 3075.1 si è rivelato leggermente superiore a quello del modello scelto (3074.8) e non sono emersi miglioramenti significativi o vantaggi evidenti. In aggiunta, la scelta della funzione di legame "logit" è motivata, anche in questo caso, dalla semplificazione nell'interpretazione delle stime di verosimiglianza e dal suo ampio utilizzo in ambito medico.

6 Bibliografia e Sitografia

- **Bibliografia**

- Laura Ventura, Walter Racugno. "Biostatistica: Casi di studio in R". A cura di EGEA, 2017.

- **Sitografia**

- www.agnesevardanega.eu - Ricerca sociale con r - Agnese Vardanega: Grafici con ggplot2.
- www.alleatiperlasalute.it - Cardiologia - Malattia cardiovascolare - Coronaropatia.