METODI STATISTICI IN BIOMEDICINA

STUDIO SU ANTIGENE PROSTATITICO SPECIFICO

ALBERTO AMADESSI

S4821511

OBIETTIVO DELLO STUDIO

L'obiettivo di questo studio di tipo osservativo è quello di verificare se un nuovo trattamento sia più efficace rispetto ad un trattamento già in uso nella pratica clinica per ridurre i livelli dell'Antigene Prostatico Specifico (PSA).

L'Antigene Prostatico Specifico è una proteina quantificabile nel sangue che viene sintetizzata nelle cellule della prostata ed è un noto *biomarker* per malattie prostatiche.

In questa ricerca sono stati coinvolti 1312 maschi di età compresa tra i 41 e 76 anni ai quali sono stati misurati i valori di PSA all'ingresso in studio (PSA basale) ed un anno dopo il trattamento (PSA finale) durato sei mesi.

VARIABILI DI STUDIO

Le variabili raccolte sono:

- Data di inizio osservazione
- Data di nascita
- Ipertrofia prostatica (Assente, lieve o severa)
- Indicatore del trattamento (Non trattato o trattato)
- Livelli basali di PSA
- Livelli finali di PSA

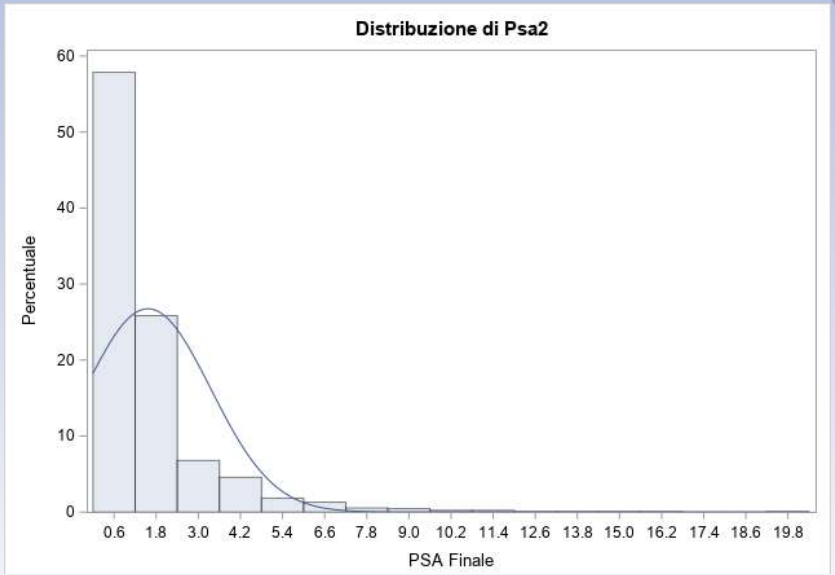
Da queste si sono ricavate ulteriori tre variabili quali l'età del paziente, l'anno di studio e la stagione di inizio osservazione in quanto si ritiene che queste caratteristiche possano incidere significativamente sullo studio in esame.

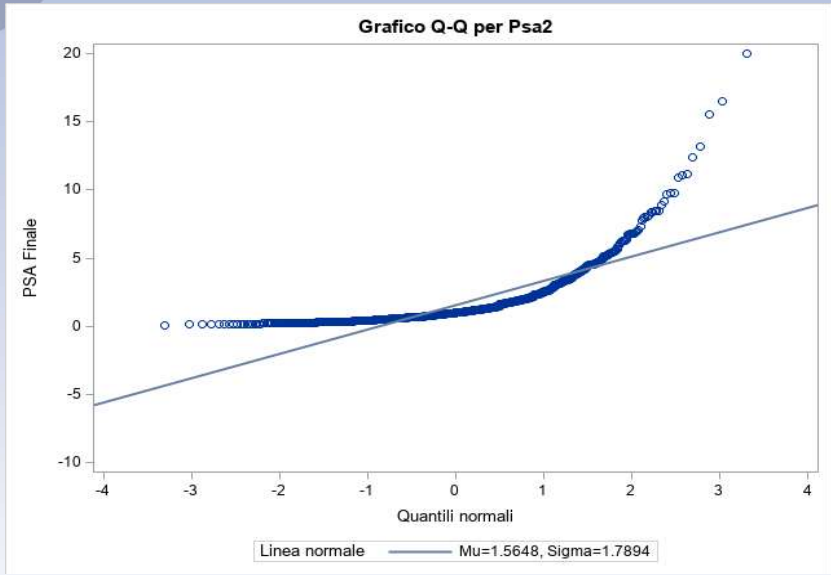
DISTRIBUZIONE PSA

Il valore finale dell'antigene prostatitico specifico viene considerato come sola variabile di interesse nell'analisi della covarianza e viene quindi trattato come unica variabile risposta del modello lineare.

Il primo passo è dunque quello di osservare la distribuzione dei valori di PSA e la relazione che presenta questa variabile con le covariate presenti nel predittore lineare.

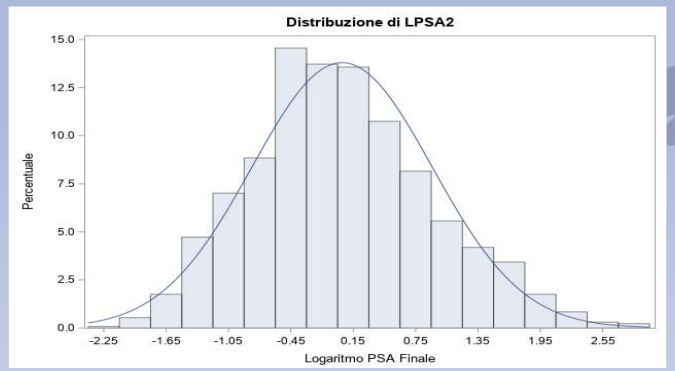
I seguenti grafici riportano la distribuzione percentuale del valore finale di PSA ed il rispettivo Q-Q plot.

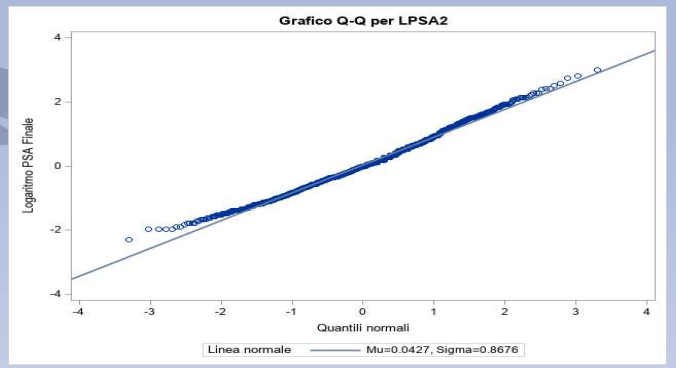




TRASFORMAZIONE PSA

Poiché l'antigene prostatico specifico sembra assumere un andamento log-normale viene effettuata una trasformazione logaritmica sulla variabile risposta al fine di rendere gaussiana la sua distribuzione.





Indicando con Y la variabile risposta (PSA) e con T la variabile trasformata con la funzione logaritmo (LPSA) si vuole verificare se valgono le seguenti relazioni:

$$\mu_Y = e^{(\mu_T + \sigma_T^2/2)}$$

$$\sigma_Y^2 = [e^{(\sigma_T^2)} - 1] * e^{(2\mu_T + \sigma_T^2)}$$

$$MG(Y) = Me(Y) = e^{(\mu_T)}$$

$$CV(Y) = \sqrt{e^{(\sigma_T^2)} - 1}$$

STATISTICA	PSA	LPSA	RAPPORTO
MA	1.565	1.520	1.029
CV	114.357	105.955	1.079
VAR	3.202	2.595	1.234
P50	0.990	1.044	0.949

ANALISI DESCRITTIVA

Dopo aver effettuato una trasformazione logaritmica sui valori basali e finali di PSA viene studiata la relazione tra la variabile risposta e le variabili esplicative.

Vengono riportate in seguito le tabelle contenenti le principali statistiche descrittive delle variabili quantitative e la rispettiva tabella di correlazione.

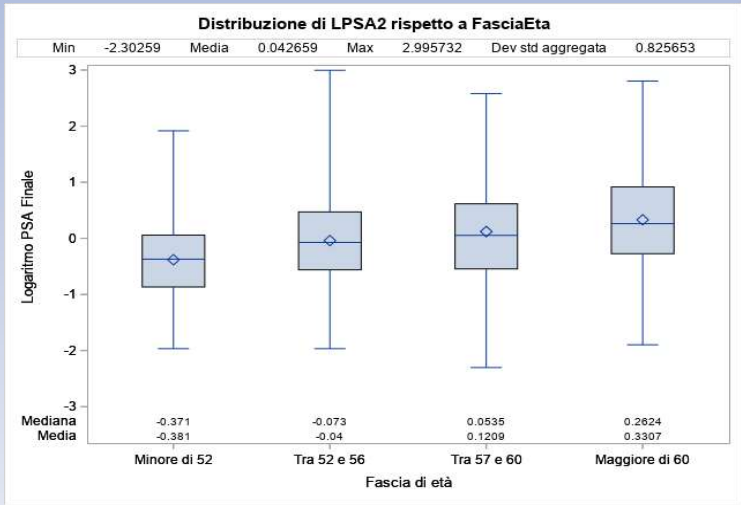
Statistiche semplici						
Variabile	N	Media	Dev std	Somma	Minimo	Massimo
Eta	1312	56.62271	5.80001	74289	41.00000	76.00000
LPSA1	1312	-0.06651	0.84085	-87.25912	-2.30259	2.72785
LPSA2	1312	0.04266	0.86756	55.96830	-2.30259	2.99573

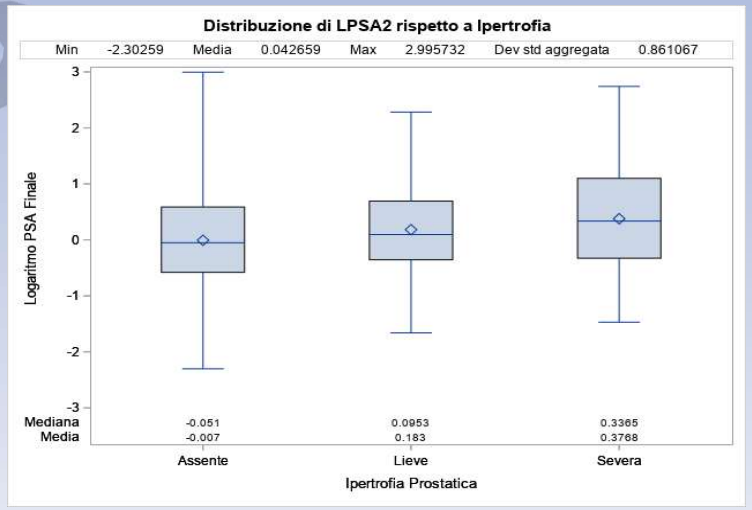
Coefficienti di correlazione di Pearson, N = 1312			
	Eta	LPSA1	LPSA2
Eta	1.00000	0.29804	0.32058
LPSA1	0.29804	1.00000	0.81012
LPSA2	0.32058	0.81012	1.00000

ANALISI DESCRITTIVA

Al fine di verificare la relazione tra le variabili qualitative e la variabile risposta sono stati analizzati diversi grafici con le variabili ritenute di interesse.

Vengono riportati in seguito i box-plot che mostrano come varia la distribuzione del logaritmo del PSA finale in relazione alle caratteristiche fisiologiche degli individui.

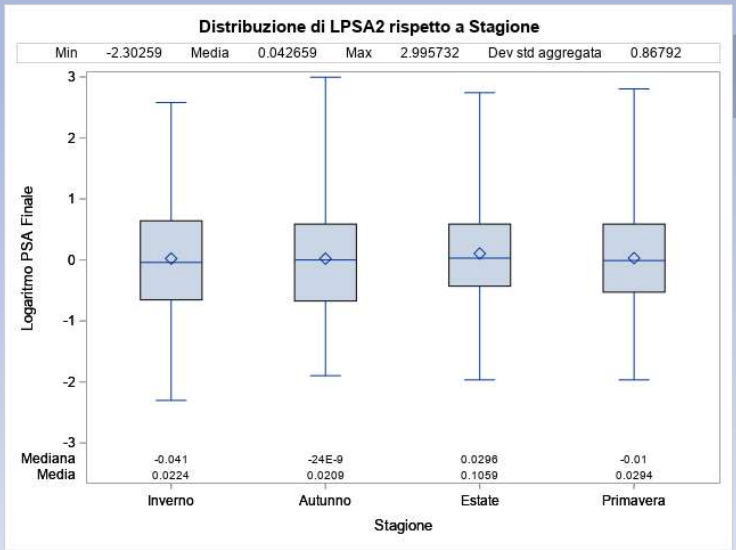


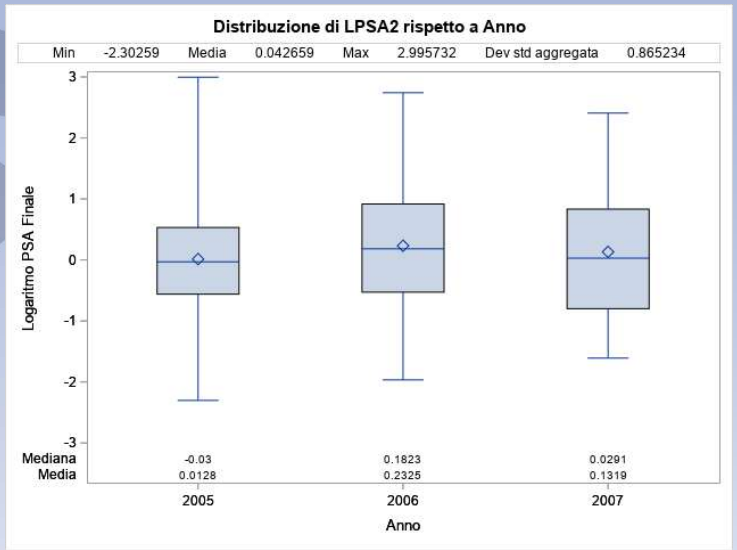


La fascia di età e l'ipertrofia prostatica sembrano incidere sul valore di PSA finale in quanto la distribuzione della variabile risposta varia nei diversi livelli delle covariate.

ANALISI DESCRITTIVA

Si riportano i box-plot relativi alla distribuzione del logaritmo di PSA finale nei livelli delle variabili temporali.



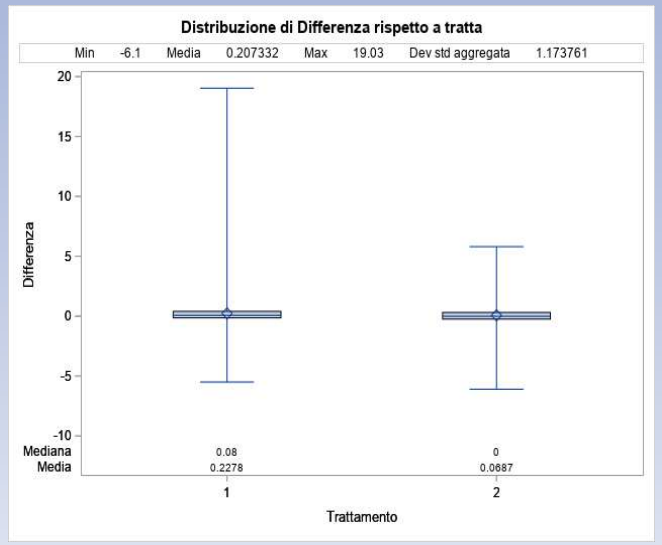


Il periodo in cui viene effettuata la rilevazione sembra incidere in modo meno significativo sui valori dell'antigene prostatico specifico in quanto si osserva come il valore di questa variabile non cambi eccessivamente nei diversi livelli.

ANALISI DESCRITTIVA

Al fine di verificare l'efficacia della cura si analizzano i box-plot relativi alla differenza dei valori di PSA nei due livelli di trattamento per i valori originali e per quelli trasformati.

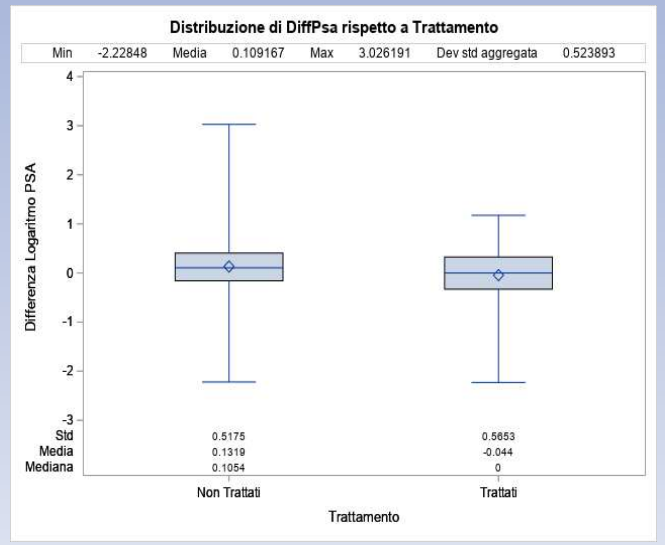
$$\text{Log}(PSA_{Finale} / PSA_{Iniziale}) = \text{Log}(PSA_{Finale}) - \text{Log}(PSA_{Iniziale})$$



Trattamento	Frequenza
Non Trattati	1143
Trattati	169

$$Log(PSA_{Finale}^{Trattati} / PSA_{Iniziale}^{Trattati}) \sim N(-0.044\,, 0.752)$$

$$Log(PSA_{Finale}^{Non\ Trattati} / PSA_{Iniziale}^{Non\ Trattati}) \sim N(0.132\,, 0.719)$$



Si osserva come la variabile non sia bilanciata e poiché i valori di media e mediana sono prossimi allo zero in entrambi i livelli si deduce come i trattamenti non siano particolarmente efficaci nell'abbassare i valori dell'antigene prostatico specifico.

MODELLO DI REGRESSIONE

Al fine di verificare se il trattamento ha ridotto significativamente i livelli finali di PSA viene sviluppato il seguente modello di regressione multiplo ad errore normale:

$$Log(PSA_2) = \beta_0 + \beta_1 * Log(PSA_1) + \beta_2 * Trattamento_2 + \beta_3 * Età_{40} + \beta_4 * Stagione_{Autunno} + \beta_5 * Stagione_{Estate} + \beta_6 * Stagione_{Primavera} + \beta_7 * Anno_{2006} + \beta_8 * Anno_{2007}.$$

Questo modello di tipo strutturato risulta in grado di controllare il possibile effetto di confondimento operato da tutte le altre variabili a disposizione e vede scalate le variabili Età ed Anno ai valori di minimo.

L'Ipertrofia Prostatica non è stata considerata nel modello in quanto ritenuta non statisticamente significativa mentre vengono riportate le stime esponenziali degli altri parametri con i relativi intervalli di confidenza.

Statistiche LR per analisi di tipo 3			
Origine	DF	Chi-quadrato	Pr > ChiQuadr
LPSA1	1	1344.47	<.0001
Trattamento	1	14.73	0.0001
ipeben	2	0.01	0.9940
Eta40	1	17.17	<.0001
Stagione	3	50.79	<.0001
Anno2005	2	42.20	<.0001

Parameter	Level1	EXP	INF_EXP	SUP_EXP
Intercept		1.12532	1.01531	1.24726
LPSA1		2.33875	2.26020	2.42002
Trattamento	2	0.85528	0.78985	0.92614
Eta40		1.01035	1.00548	1.01524
Stagione	Autunno	0.79469	0.72809	0.86739
Stagione	Estate	0.90690	0.83916	0.98012
Stagione	Primave	0.79550	0.74275	0.85199
Anno2005	1	0.86916	0.79716	0.94766
Anno2005	2	0.65015	0.56614	0.74661

MODELLO DI REGRESSIONE ALTERNATIVO

Viene sviluppato un secondo modello di regressione multipla in cui si considera come variabile risposta il cambiamento assoluto dei valori di PSA dopo la trasformazione logaritmica:

$$Log(PSA_2/PSA_1) = \beta_0 + \beta_1 * Trattamento_2 + \beta_2 * Stagione_{Autunno} + \beta_3 * Stagione_{Estate} + \beta_4 * Stagione_{Primavera} + \beta_5 * Anno_{200} + \beta_6 * Anno_{2007}.$$

Anche in questo caso la variabile risposta assume una distribuzione gaussiana in quanto si tratta della differenza tra due variabili con distribuzione normale.

L'Ipertrofia Prostatica e l'Età non sono state considerate nel modello in quanto ritenute non statisticamente significative mentre vengono riportate le stime esponenziali degli altri parametri con i relativi intervalli di confidenza.

Statistiche LR per analisi di tipo 3			
Origine	DF	Chi-quadrato	Pr > ChiQuadr
Trattamento	1	20.62	<.0001
ipeben	2	1.16	0.5600
Eta40	1	2.59	0.1073
Stagione	3	67.97	<.0001
Anno2005	2	71.84	<.0001

Parameter	Level1	EXP	INF_EXP	SUP_EXP
Intercept		1.41694	1.33900	1.49941
Trattamento	2	0.82130	0.75739	0.89060
Stagione	Autunno	0.76871	0.70279	0.84081
Stagione	Estate	0.86940	0.80314	0.94113
Stagione	Primave	0.75749	0.70651	0.81214
Anno2005	1	0.80809	0.74049	0.88187
Anno2005	2	0.58500	0.50847	0.67305

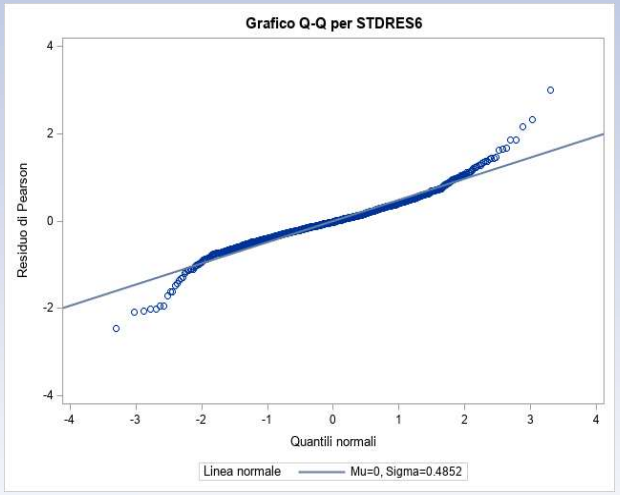
MODELLO DI REGRESSIONE OTTIMALE

Al fine di sfruttare in modo ottimale le variabili a disposizione si sono testati diversi modelli di tipo polinomiale e/o con iterazioni tra le variabili.

Il modello che presenta i valori di entropia minori è quello composto da tutte le variabili a disposizione ad eccezione dell'ipertrofia prostatica e con l'iterazione tra l'anno dell'inizio dell'osservazione ed il valore logaritmo del PSA basale:

$$Log(PSA_2) = \beta_0 + \beta_1 * Log(PSA_1) + \beta_2 * Trattamento_2 + \beta_3 * Età_{40} + \beta_4 * Stagione_{Autunno} + \beta_5 * Stagione_{Estate} + \beta_6 * Stagione_{Primavera} + \beta_7 * Anno_{2006} + \beta_8 * Anno_{2007} + \beta_9 * Anno_{2006} * Log(PSA_1) + \beta_{10} * Anno_{2007} * Log(PSA_1).$$

Statistiche di Wald per test congiunti			
Origine	DF	Chi-quadrato	Pr > ChiQuadr
Trattamento	1	15.11	0.0001
LPSA1	1	1158.56	<.0001
Eta40	1	17.03	<.0001
Stagione	3	52.84	<.0001
Anno2005	2	47.99	<.0001
LPSA1*Anno2005	2	8.74	0.0127



ANALISI DEI COEFFICIENTI

Al fine di fornire un significato statistico alle variabili trattate si riportano le stime dei coefficienti di regressione del modello ritenuto ottimale con la rispettiva con trasformazione esponenziale.

Parameter	Level1	BETA	INF	SUP	EXP	INF_EXP	SUP_EXP
Intercept		0.12072	0.01814	0.22330	1.12831	1.01831	1.25019
Trattamento	2	-0.15776	-0.23710	-0.07843	0.85405	0.78892	0.92456
LPSA1		0.82580	0.78828	0.86332	2.28370	2.19960	2.37102
Eta40		0.01018	0.00536	0.01500	1.01023	1.00538	1.01512
Stagione	Autunno	-0.23858	-0.32602	-0.15114	0.78775	0.72179	0.85973
Stagione	Estate	-0.10341	-0.18088	-0.02593	0.90176	0.83453	0.97440
Stagione	Primave	-0.23228	-0.30082	-0.16373	0.79273	0.74021	0.84897
Anno2005	1	-0.16256	-0.25099	-0.07413	0.84997	0.77803	0.92855
Anno2005	2	-0.47044	-0.61662	-0.32427	0.62473	0.53977	0.72305
LPSA1*Anno2005	1	0.11843	0.02437	0.21250	1.12573	1.02467	1.23676
LPSA1*Anno2005	2	0.12151	-0.00461	0.24764	1.12920	0.99540	1.28099

Poiché il coefficiente relativo al trattamento risulta negativo si deduce che la nuova cura sembra maggiormente efficace rispetto alla precedente mentre il coefficiente relativo all'età scalata mostra come l'aumento di un anno di età comporti in mediana un aumento dell'antigene prostatico specifico dell'1% a parità delle altre covariate.

EFFETTO CONFONDIMENTO

Al fine di quantificare l'effetto di confondimento delle variabili sulla relazione tra il valore logaritmico dell'antigene prostatico specifico ed il trattamento si considera la seguente tabella che riporta la differenza percentuale dei valori del coefficiente del trattamento rispetto al modello senza i confondenti.

Modello	Devianza	AIC	EXP	INF_EXP	SUP_EXP	Differenza
Singolo	1.0015	3354.78	0.941	0.818	1.082	0.000
Qualitative	1.0054	3352.36	0.944	0.821	1.086	0.319
Quantitative	1.0031	1921.85	0.876	0.808	0.951	-6.908
Qualitative con LPSA	1.0061	1868.85	0.836	0.772	0.905	-11.158
Completo	1.0069	1853.52	0.855	0.790	0.926	-9.139
Completo con iterazione	1.0085	1848.78	0.854	0.789	0.925	-9.245

Dalla tabella si nota come il trattamento non sia statisticamente significativo al fine della regressione se considerato singolarmente e di come la differenza assoluta percentuale del coefficiente sembri aumentare con il numero di covariate considerate.

CONCLUSIONI

Nonostante i due trattamenti non sembrino abbassare drasticamente i valori dell'antigene prostatico specifico in quanto i livelli della seconda misurazione risultano mediamente maggiori rispetto ai livelli della prima per entrambi i trattamenti si deduce che la nuova cura sia più efficace della prima in quanto nei modelli di regressione costruiti il relativo coefficiente di regressione risulta di valore negativo e statisticamente significativo.