Big Data
in
Public Health

Analisi sul Tumore alla Mammella

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Giannelli Alessio
Imbonati Lorenzo

INDICE

- ANALISI ESPLORATIVA

  Dataset, Statistiche descrittive e tipo variabili

- PREPROCESSING e DATA CLEANING

  dati ripetuti, incongruenza tra date, incongruenza età

- DOMANDE di RICERCA

ANALISI ESPLORATIVA

CANCER

idnum	Stadio	incidenza	tipotumore	geneticm
Min. : 3	: 7	Min. :1984-01-11	: 3	0:8891
1st Qu.: 2964	Stadio I :1607	1st Qu.:1984-01-13	altro :3250	1:1114
Median : 5984	Stadio II :5507	Median :1984-01-15	colon :2062	
Mean : 5975	Stadio III:1197	Mean :1984-01-15	polmone:2072	
3rd Qu.: 8965	Stadio IV :1687	3rd Qu.:1984-01-18	seno :2618	
Max. :12000		Max. :1984-01-20		
		NA's :5		

Si evince dalle statistiche descrittive come vi siano 3 tipi di variabili entro i dataset: categoriali, continue e data.

SDO

idnum	Prestazione	dataprestazione	dimissione	ospedale
Min. : 3	chemioterapica:1988	Min. :1984-01-22	Min. :1984-06-22	1 :1150
1st Qu.: 2966	chirurgica :3318	1st Qu.:1984-02-19	1st Qu.:1984-07-17	7 :1126
Median : 5986	radioterapica :4696	Median :1984-03-08	Median :1984-08-02	8 :1119
Mean : 5976		Mean :1984-03-19	Mean :1984-08-12	9 :1119
3rd Qu.: 8966		3rd Qu.:1984-04-09	3rd Qu.:1984-09-01	6 :1110
Max. :12000		Max. :1984-10-07	Max. :1985-02-12	5 :1107
		NA's :1	NA's :3	(Other):3271

DEATH

idnum	dead	enddate
Min. : 1	0:5130	Min. :1984-06-29
1st Qu.:1938	1:2618	1st Qu.:1985-11-12
Median :3874		Median :1987-05-27
Mean :3874		Mean :1987-03-31
3rd Qu.:5811		3rd Qu.:1988-11-05
Max. :7748		Max. :1988-12-31

GERMANH

idnum	smoke	sex	married	kids	work	education	age
Min. : 1	no :6146	Female:3871	no :1690	no :4260	no :7274	low :6922	Min. : 26.00
1st Qu.:1938	yes:1602	Male :3877	yes :6036	yes :3474	yes: 474	medium/high: 790	1st Qu.: 41.00
Median :3874			NA's: 22	NA's: 14		NA's : 36	Median : 46.00
Mean :3874							Mean : 48.21
3rd Qu.:5811							3rd Qu.: 54.00
Max. :7748							Max. :108.00

Su molteplici feature dovranno essere gestiti valori mancanti (NA) mentre su alcune variabili categoriali gestiremo le modalità non definite.

PRE-PROCESSING

DATI RIPETUTI

- righe duplicate solo nel dataset Cancer per i 3 pazienti 192, 363 e 1933 → righe eliminate

INCONGRUENZA DATE

- incongruenze trovate nel dataset SDO dato che per 42 obs la data di dimissione dall'ospedale è minore della data di prestazione dell'operazione → righe eliminate

INCONGRUENZA ETA'

- incongruenze trovate nel dataset GermanH dato che si hanno 8 pazienti con età superiore ai 100 anni → righe eliminate

idnum <int>	Stadio <chr>
192	Stadio IV
363	Stadio II
1933	Stadio I

idnum	Prestazione	dataprestazione	dimissione
5377	radioterapica	1984-07-22	1984-07-21
4456	chirurgica	1984-07-15	1984-07-01
2813	radioterapica	1984-09-19	1984-09-13
245	chemioterapica	1984-07-19	1984-07-17

work	education	age
no	low	108
no	low	108
no	low	104
no	low	104
no	low	102
no	low	102
no	low	100
no	low	100
no	low	98

Domanda di Ricerca 2

Effettuare il record-linkage con lo scopo di costruire l'indicatore 'Intervento chirurgico di asportazione del tumore al seno entro 60 giorni dalla data di diagnosi' su base mensile per i casi incidenti nel mese di gennaio 1984.

1. Record-linkage dei dataset Cancer, GermanH e SDO
2. Filtraggio dataset → tumore seno, prestazione chirurgica, sesso femminile, tumore di stadio I e II
3. Si ottiene un dataset di 314 osservazioni
4. Creazione nuova variabile "giorni_passati" calcolando il numero di giorni tra la data di incidenza del tumore e la data di esecuzione dell'intervento
5. Creazione variabile dicotomica "entro_60_giorni" pari a 1 se "giorni_passati" ≤ 60 giorni, altrimenti pari a 0
6. Calcolo indicatore d'interesse → num = 172
   den = 314
   indicatore = 54,7%

	idnum	Stadio	incidenza	tipotumore	geneticm	Prestazione
1161	1162	Stadio II	*NA*	seno	0	chirurgica

Paziente 1162 con dato mancante

Distribuzione "entro_60_giorni"

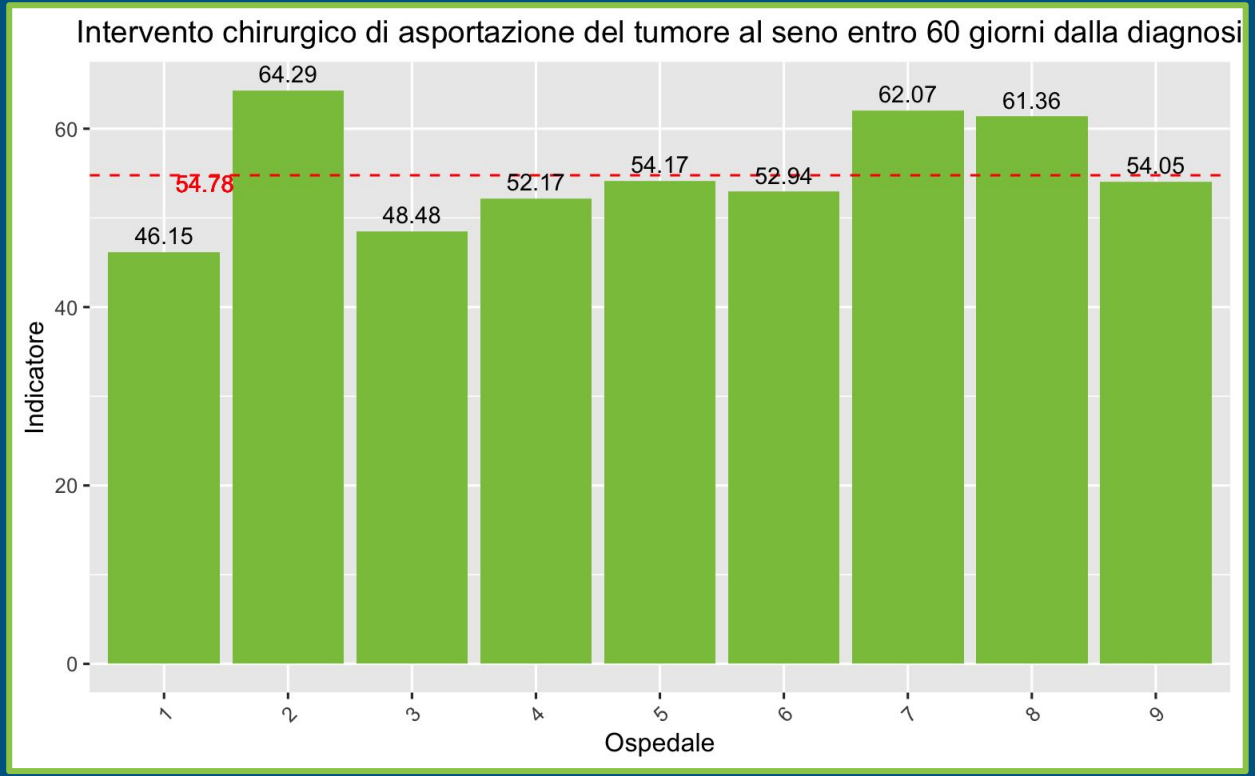
0	1
142	172

Domanda di Ricerca 3

Calcolare l'indicatore 'Intervento chirurgico di asportazione del tumore al seno entro 60 giorni dalla data di diagnosi' per ospedale e darne rappresentazione grafica, includendo come valore di riferimento nel grafico l'indicatore calcolato sull'intero dataset.

Il grafico a destra mostra la percentuale di "interventi chirurgici di asportazione del tumore al seno entro 60 giorni dalla diagnosi per ogni ospedale", rispetto all'indice generale pari a 54.78% (linea rossa tratteggiata). Il valore per ciascun ospedale è rappresentato da una barra.

Si può notare che l'ospedale con il valore più alto (64.29%) è il secondo, mentre quello con il valore più basso (46.15%) è il primo.



Domanda di Ricerca 4

Utilizzare il dataset ottenuto per valutare l'associazione a livello individuale tra il livello di educazione ed il valore dell'indicatore 'Intervento chirurgico di asportazione del tumore al seno entro 60 giorni dalla data di diagnosi'.

Eliminazione delle 2 obs mancanti per "education" → Creazione variabile binaria "education_bin" → Stima modello di regressione logistica: dipendente indicatore predittore "education_bin"

Tabella di contingenza

	Outcome +	Outcome -
Exposed +	121	19
Exposed -	149	23

Odds Ratio

	OR	2.5 %	97.5 %
(Intercept)	1.2314050	0.9688018	1.565189
education_bin	0.9830449	0.5115325	1.889181

Associazione non significativa dato che IC presenta il valore 1

Domanda di Ricerca 5

Calcolate la stessa misura di effetto, questa volta aggiustata per la sola variabile ‘working’, mediante il metodo Mantel Haenszel.

Tabella di contingenza aggiustata

Work = 0	Outcome +	Outcome -
Exposed +	111	19
Exposed -	140	21

Tabella di contingenza aggiustata

Work = 1	Outcome +	Outcome -
Exposed +	10	0
Exposed -	9	2

Odds Ratio Mantel Haenszel

Odds ratio (crude)	0.98 (0.51, 1.89)
Odds ratio (M-H)	0.98 (0.51, 1.88)
Odds ratio (crude:M-H)	1.00

L'OR tra l’indicatore e "education_bin", dopo l'aggiustamento per la variabile "working" tramite il metodo Mantel-Haenszel, è pari a 0.98.

Dato il rapporto tra OR grezzo e OR M-H pari a 1 si può affermare che “working” non è una variabile confondente.

Domanda di Ricerca 6

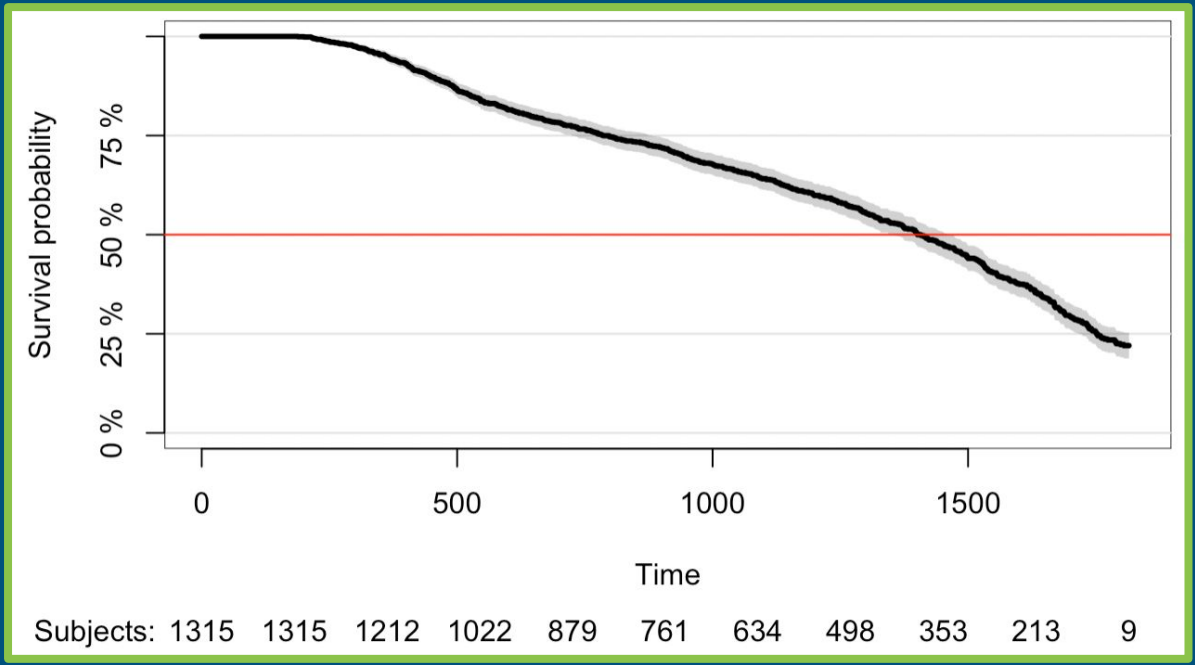
Stimate l'associazione a livello individuale tra il livello di educazione ed il valore dell'indicatore, aggiustata per tutte le variabili disponibili che ritenete opportuno inserire come potenziali confondenti, mediante un modello di regressione logistica. Su quanti soggetti avete effettuato l'analisi?
Quali variabili sono associate all'indicatore? In che modo?

	OR grezzo	OR M-H	Rapporto tra OR
geneticm	0.98	0.947	1.037
smoke	0.98	0.972	1.011
age	0.98	0.977	1.006

Il rapporto tra OR grezzo e OR M-H è, per tutte le variabili considerate, simile a 1 → si può affermare che non si tratta di variabili confondenti

Domanda di Ricerca 7

Selezionate i record relativi ai tumori al colon e stimate la sopravvivenza a 5 anni

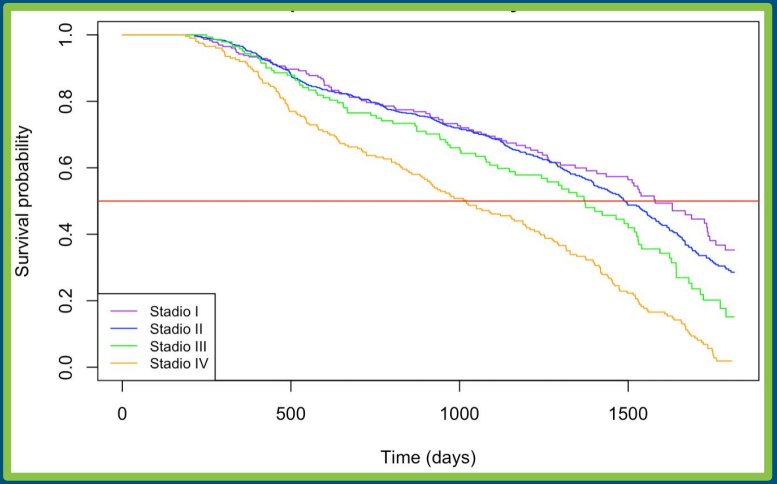


I soggetti a inizio studio sono 1315 e quelli che decedono nei 5 anni di studio sono esattamente 715.

La mediana di sopravvivenza corrisponde a circa 1400 giorni.

Domanda di Ricerca 8

Stimare la sopravvivenza nei primi 5 anni dalla diagnosi per Stadio e effettuare un test d'ipotesi per verificare se l'azzardo di morte sia diverso per stadio di malattia alla diagnosi.

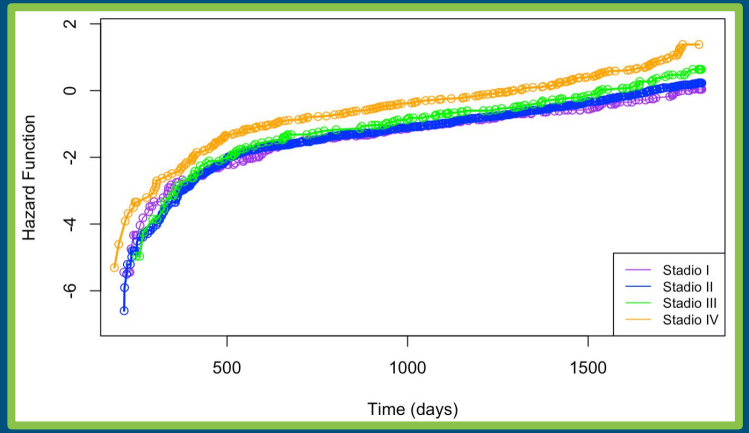



```
        Pairwise comparisons using Pairwise comparison of proportions

data:  observed

           Stadio I Stadio II Stadio III
Stadio II  0.4737   -         -
Stadio III 0.0073   0.1338    -
Stadio IV  < 2e-16  < 2e-16   3.7e-11

P value adjustment method: bonferroni
```

Solamente tra Stadio I e Stadio II non si notano differenze significative per quanto riguarda l'azzardo di morte

```
        Pearson's Chi-squared test

data:  observed
X-squared = 137.45, df = 3, p-value < 2.2e-16
```

Ci sono differenze significative tra i gruppi, si analizza tra quali sussistono queste differenze



Domanda di Ricerca 9

Applicare un modello per valutare l'associazione tra sesso e mortalità e interpretare la misura di effetto stimata.

```
Call:
glm(formula = dead ~ sex, family = binomial, data = df_colon)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.350  -1.164   1.014   1.191   1.191

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.03279    0.07722  -0.425 0.671105
sexMale      0.42921    0.11145   3.851 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1812.9  on 1314  degrees of freedom
Residual deviance: 1798.0  on 1313  degrees of freedom
AIC: 1802

Number of Fisher Scoring iterations: 4
```

	OR	2.5 %	97.5 %
(Intercept)	0.9677419	0.8318216	1.125872
sexMale	1.5360360	1.2346231	1.911034

Odds Ratio del modello che ha come target la morte e covariata il sesso è pari a 1.54.
Inoltre IC non presenta al suo interno il valore 1 quindi associazione tra sesso e mortalità statisticamente significativa.

I pazienti di sesso maschile hanno una probabiltà maggiore di morire per tumore al colon rispetto ai pazienti di sesso femminile, pari ail 54%

Domanda di Ricerca 10

Quali variabili sono associate alla mortalità? Riportare le relative stime di effetto con gli intervalli di confidenza.

```
Call:
glm(formula = dead ~ smoke + married + kids + work + education +
    age, family = binomial(), data = df_colon)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3747  -1.0706   0.4744   1.0618   1.7405

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.650109   0.357310 -10.216  < 2e-16 ***
smokeyes              0.385317   0.150848   2.554  0.01064 *
marriedyes           -0.070219   0.149617  -0.469  0.63884
kidsyes               0.127328   0.119996   1.061  0.28864
workyes               0.146981   0.257539   0.571  0.56819
educationmedium/high  0.689696   0.216233   3.190  0.00142 **
age                   0.077129   0.006907  11.167  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1798.4  on 1303  degrees of freedom
Residual deviance: 1620.0  on 1297  degrees of freedom
  (11 observations deleted due to missingness)
AIC: 1634

Number of Fisher Scoring iterations: 4
```

	OR	2.5 %	97.5 %
(Intercept)	0.0260	0.0129	0.0524
smokeyes	1.4701	1.0938	1.9758
marriedyes	0.9322	0.6953	1.2499
kidsyes	1.1358	0.8978	1.4369
workyes	1.1583	0.6992	1.9189
educationmedium/high	1.9931	1.3046	3.0450
age	1.0802	1.0657	1.0949

Si nota che le variabili 'smoke' e 'age' risultano significativamente associate con il target 'dead' sia analizzando la significatività dei coefficienti del modello che osservando gli intervalli di confidenza (non presentano il valore 1 nell'IC).
Si puo' quindi supporre che chi fuma ha una probabilità del 47% maggiore di morire di tumore al colon rispetto a chi non fuma; mentre chi ha un'età più anziana ha una probabilità dell'8% maggiore di morire di tumore al colon.
Si deve valutare se 'education' si tratta di una variabile confondente o meno.

Domanda di Ricerca 11

Valutare la presenza di confondenti e/o modificatori di effetto tra le variabili disponibili nel German health register e nel registro tumori nella valutazione dell'associazione tra sesso e mortalità. Se identificate un'interazione tra sesso e un'altra variabile riportare le stime di effetto per maschi e femmine separatamente e commentare il tipo di interazione trovato.

Dead / Sex	OR grezzo	OR M-H	Rapporto tra OR	M-H test Omogeneità (p)
Smoke	1.54	1.53	1.01	0.636
Married	1.53	1.53	1	0.313
Kids	1.53	1.53	1	0.656
Work	1.54	1.54	1	0.223
Education	1.54	1.53	1.01	**0.032**

Non ci sono variabili confondenti. La variabile "Education" risulta essere un modificatore d'effetto. Dopo aver fatto interagire le stime di effetto per maschi e femmine separatamente con la feature "Education" si è giunti alla conclusione che LE DONNE PIU' ISTRUITE affette da TUMORE AL COLON hanno una probabilità 2.93 superiore di morire rispetto alle DONNE MENO ISTRUITE.

Domanda di Ricerca 12

A seguito delle considerazioni effettuate nei punti precedenti scegliete un modello finale per valutare i fattori di rischio della mortalità dopo diagnosi di tumore al colon e commentate i risultati.

```
model_cox <- coxph(Surv(time_d, I(as.numeric(dead))) ~ relevel(sex, "Male") + geneticm  + Stadio + smoke + age + education, data=df_colon)
```

```
model_cox_2 <- coxph(Surv(time_d, I(as.numeric(dead))) ~ relevel(sex, "Male") * geneticm + relevel(sex, "Male") * Stadio + relevel(sex, "Male") * education + smoke + age, data=df_colon)
```

I due modelli sono significativamente diversi e grazie alla ANOVA si può affermare che il modello che include le interazioni risulta migliore del primo. I test di ipotesi riportano che il modello di Cox è significativo. In particolare, risultano significative le variabili "sex", "geneticm", "Stadio" per il quarto livello e "age". In merito alle interazioni inserite nel modello, risultano significative quelle tra "sex" e "Stadio II", tra "sex" e "Stadio III".

L'Hazard Ratio risulta significativamente superiore a 1 per le variabili "geneticm", "Stadio IV" e "age" e per le interazioni tra "sex" e "Stadio II" e tra "sex" e "Stadio III". Queste variabili sono positivamente associate con il rischio di mortalità. L'Hazard Ratio risulta invece significativamente inferiore a 1 per la variabile "sex". Le restanti variabili hanno un Hazard Ratio non significativamente diverso da 1.

Giannelli Alessio
Imbonati Lorenzo

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

GRAZIE della vostra ATTENZIONE