Progetto Finale

Big Data in Public Health

Sanja Stanisic, matr. 800409

DATASET "GERMANH.CSV"

Numero	Attributo	Tipo di variabile	Descrizione	
1	idnum	numerica	7676 record	
2	smoke	dicotomica	no	6099
			sì	1577
3	sex	dicotomica	femminile	3835
			maschile	3841
4	married	dicotomica	no	1676
			sì	6000
5	kids	dicotomica	no	4222
			sì	3454
6	work	dicotomica	no	7206
			sì	470
7	education	dicotomica	bassa	6888
			medio/alta	788
8	age	continua	minimo	26
			massimo	108
			media	48.2
			mediana	45.5

German Health Registry

Questo dataset è un estratto del Registro sanitario tedesco, che riporta i dati relativi ad una piccola cittadina tedesca per l'anno **1984.**

Dataset originale

Sui **7748** record ci sono stati:

- → **36** dati mancati (NA) nell'attributo education
- → **22** NA nello stato civile (married)
- → **14** NA nell'attributo kids

DATASET "CANCERREGISTER.CSV"

Numero	Attributo	Tipo di variabile	Descrizione	
1	idnum	numerica	9993 record	
2	stadio	categorica	I	1605
			II	5505
			III	1197
			IV	1686
3	incidenza	data	minima	11/01/1984
			massima	20/01/1984
4	tipotumore	categorica	altro	3247
			colon	2061
			polmone	2072
			seno	2613
5	geneticm	dicotomica	0	8879
			1	1114

Registro dei tumori

Registro dei tumori tedesco relativo al mese di gennaio del **1984**.

Dataset originale

Sui **10005** record ci sono stati:

- → **7** records con stringa vuota nell'attributo **Stadio** del tumore
- → **3** records con stringa vuota dell'attributo **tipotumore**
- → **5** dati mancanti (NA) nell'**incidenza.**
- → **3** duplicati (idnum **192, 363 1933**)

DATASET "SDO.CSV"

Numero	Attributo	Tipo di variabile	Descrizione	
1	idnum	numerica	9956 record	
2	prestazione	categorica	chemioterapica	1980
			chirurgica	3301
			radioterapica	4675
3	dataprestazione	data	minima	22/01/1984
			massima	07/10/1984
4	dimissione	data	minima	23/06/1984
			massima	12/02/1985
5	ospedale	numerica	1 - 9 ospedali	

Distribuzione dei pazienti per ospedale								
1	2	3	4	5	6	7	8	9
1141	1079	1086	1094	1103	1105	1121	1114	1113

Schede di dimissione

Schede di dimissione ospedaliera dei soggetti ricoverati per trattamenti oncologici in Germania tra gennaio 1984 e ottobre 1984.

Dataset originale

Sui **10002** record ci sono stati:

- → **1** record con dato mancante (NA) nell'attributo **dataprestazione**
- → **3** NA nell'attributo **dimissione**
- → **42** records avevano date incongruenti, ovvero i soggetti sarebbero stati dimessi prima di essere sottoposti alle terapie

DATASET "DEATHREGISTER.CSV"

Numero	Attributo	Tipo di variabile	Descrizione	
1	idnum	numerica	7748 record	
2	dead	categorica	0	5130
			1	2618
3	enddate	data	minima	29/06/1984
			massima	31/12/1988

Registro di mortalità

Estratto del Registro di mortalità della cittadina tedesca che riporta la mortalità dal 1984 al 1988 e lo stato in vita alla fine del 1988 .

Dataset originale

Sui ***7748*** record non ci sono né dati mancanti né duplicati.

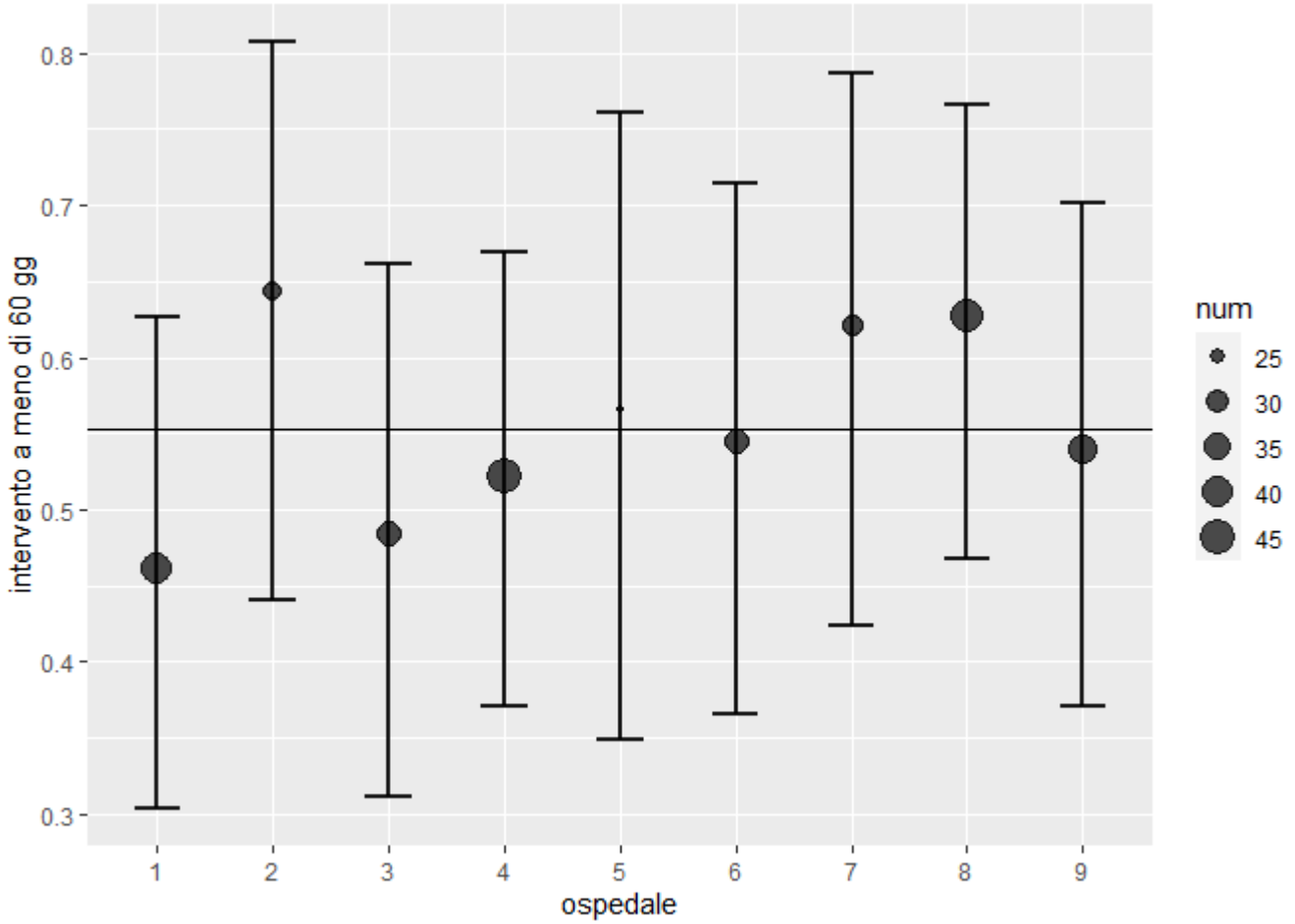
2. INDICATORE

→ 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 è costruito facendo il record-linkage tra i dataset: German Health Register, Cancer register e Schede di dimissione ospedaliera (linkage by: codice identificativo del paziente)

Descrizione	Definizione operativa numeratore	Definizione operativa denominatore
Proporzione delle pazienti, diagnosticate con il tumore al seno in stadio I-II, sottoposte all'intervento chirurgico entro 60 giorni dalla diagnosi.	Tutti i soggetti al denominatore con intervallo tra la data d'incidenza e la data dell'intervento minore o uguale a 60 giorni	Tutti i soggetti di sesso femminile con tumore al seno insorto tra 01/01/1984 e 31/01/1984, in stadio I o II, che hanno subito un intervento chirurgico.

$$I = 172/311 = 0.5530547$$

3. INDICATORE PER OSPEDALE



Ospedale	Indicatore
1	0.46154
2	0.64285
3	0.48485
4	0.52174
5	0.56522
6	0.54545
7	0.62069
8	0.62791
9	0.54054

4. ASSOCIAZIONE: EDUCAZIONE VS. INDICATORE

- La misura di effetto da stimare è l'**Odds Ratio.**

Tabella di contingenza

Educazione	Indicatore		Tot.
	0	1	
Low	120	149	269
Medium/High	19	23	42
Tot.	139	192	311

- OR = 0.97 CI (0.51, 1.87)

- I soggetti con la formazione medio/alta hanno lo stesso odds di avere l'intervento chirurgico al seno entro 60 giorni dalla diagnosi come i soggetti con la formazione bassa

NON C'È ASSOCIAZIONE a livello individuale tra il livello di educazione e il valore dell'indicatore

5. ASSOCIAZIONE AGGIUSTATA PER "WORK"

Non c'è associazione

Dato che non c'è associazione a livello individuale tra il livello di educazione ed il valore dell'indicatore, non ha senso aggiustare l'OR del punto 4. per qualsiasi altra variabile

Aggiustare l'OR per la variabile «work»

In particolare, in questo caso, anche se ci fosse stata l'associazione, sarebbe impossibile aggiustare per la variabile **'work'** con il metodo **Mantel Haenszel**, dato che in una delle due tabelle di contingenza stratificate per la variabile 'work', si ha uno 0 che poi nei calcoli produce NaNs.

Status = 0		
Education	Work	
	no	yes
low	110	10
medium/high	19	0

Status = 1		
Education	Work	
	no	yes
low	140	9
medium/high	21	2

6. ASSOCIAZIONE: EDUCAZIONE VS. INDICATORE REGRESSIONE LOGISTICA (GLM)

- Le analisi sono state fatte su **311** soggetti.
- Applicando il modello di regressione logistica, veniamo alla stessa conclusione ottenuta al punto 4: a livello individuale non c'è l'associazione tra il livello di educazione e il valore dell'indicatore, non ha senso cercare i potenziali confondenti.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2165	0.1227	1.765	0.0776
Education medium/high	-0.0254	0.3334	-0.076	0.9393

	OR	2.5 %	97.5%
(Intercept)	1.2416667	0.9763374	1.579102
Education medium/high	0.9749205	0.5072023	1.873947

6. ASSOCIAZIONE: STADIO VS. INDICATORE

- L'unica variabile associata al valore dell'indicatore e la variabile **'Stadio'**, ed è un'associazione positiva.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5486	0.2463	-2.227	0.02597
StadioStadioII	0.9889	0.2796	3.537	0.0000405

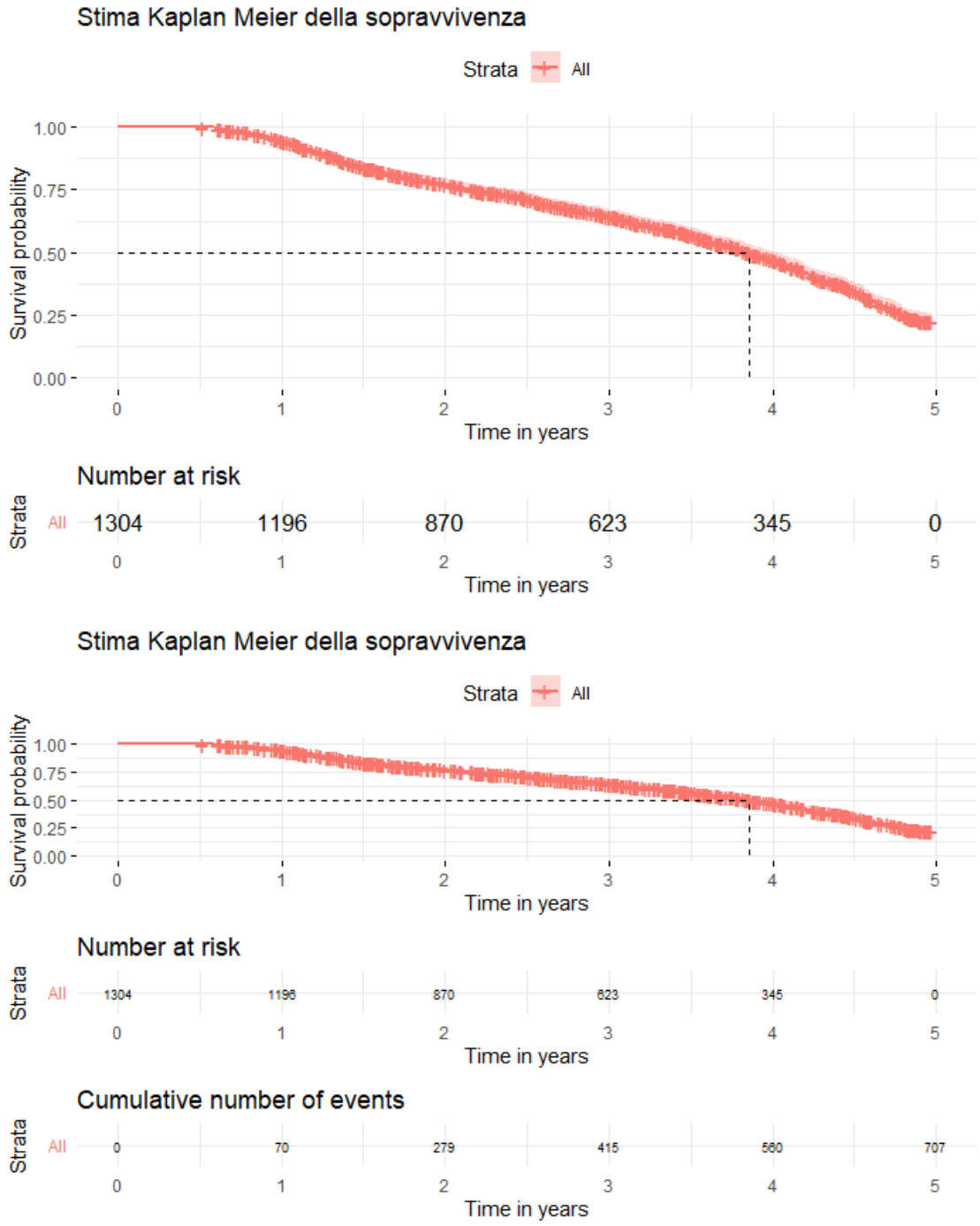
	OR	2.5 %	97.5%
(Intercept)	0.5777778	0.3565129	0.9363677
Education medium/high	2.6882160	1.5540877	4.6499986

- I soggetti che hanno il tumore al seno allo **Stadio II** hanno l'odds di avere l'intervento chirurgico entro 60 giorni dalla diagnosi **quasi 3** volte maggiore rispetto alle pazienti aventi il tumore allo **Stadio I**.

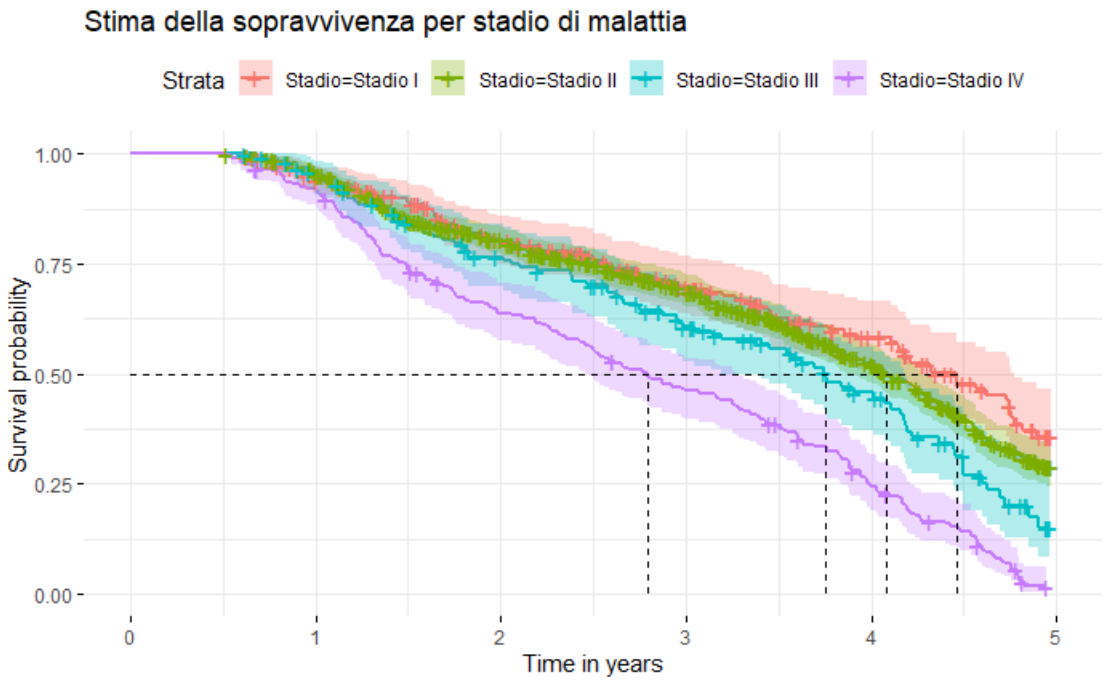
7.TUMORE AL COLON

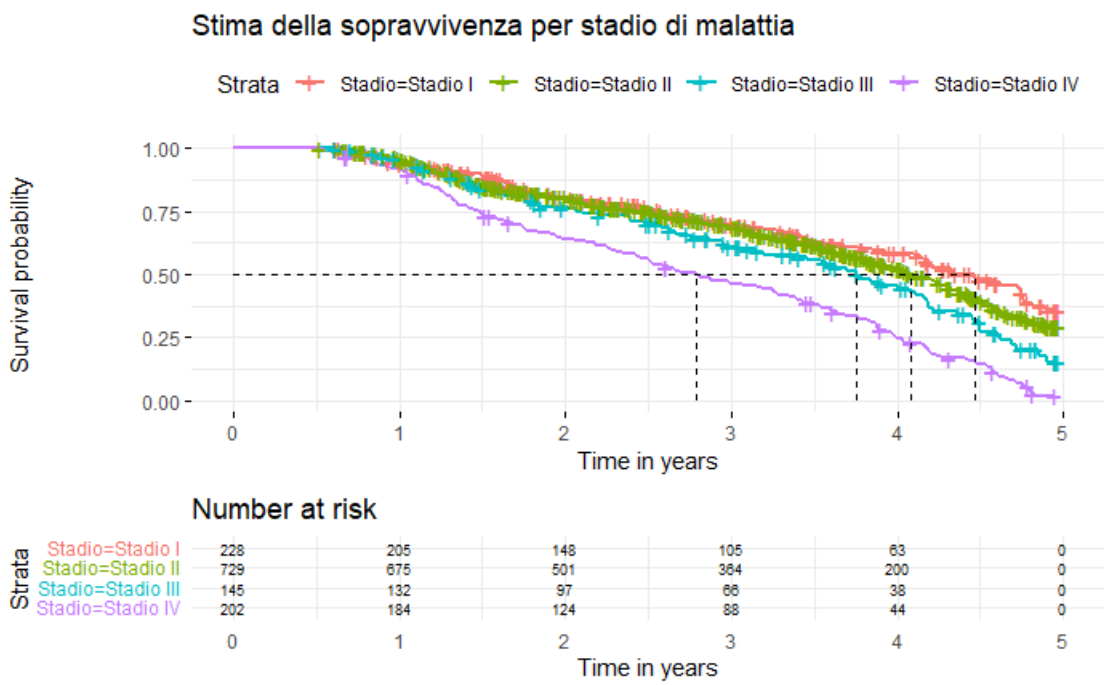
Sopravvivenza a 5 anni

- Datasets: Registro tumori, Registro di morte e Registro sanitario tedesco (record linkage by: idnum)
- **1304** soggetti sono stati inclusi In questa analisi
- Ci sono stati **707** eventi (morti)
- La stima approssimativa della mediana è **3.86** anni.



8.STIMA DI SOPRAVVIVENZA PER STADIO DI TUMORE





8. STIMA DI SOPRAVVIVENZA PER STADIO DI MALATTIA

Stima di sopravvivenza - Modello di Cox

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Stadio II	0.1613	1.1750	0.1173	1.375	0.16924
Stadio III	0.4314	1.5393	0.1506	2.865	0.00417
Stadio IV	0.9614	2.6153	0.1282	7.498	6.49e-14

	exp(coef)	exp(-coef)	lower .95	upper .95
Stadio II	1.175	0.8511	0.9336	1.4791.539
Stadio III	1.539	0.6496	1.1459	2.068
Stadio IV	2.615	0.3824	2.0341	3.362

8. STIMA DI SOPRAVVIVENZA PER STADIO DI MALATTIA

Stima di sopravvivenza - Modello di Cox

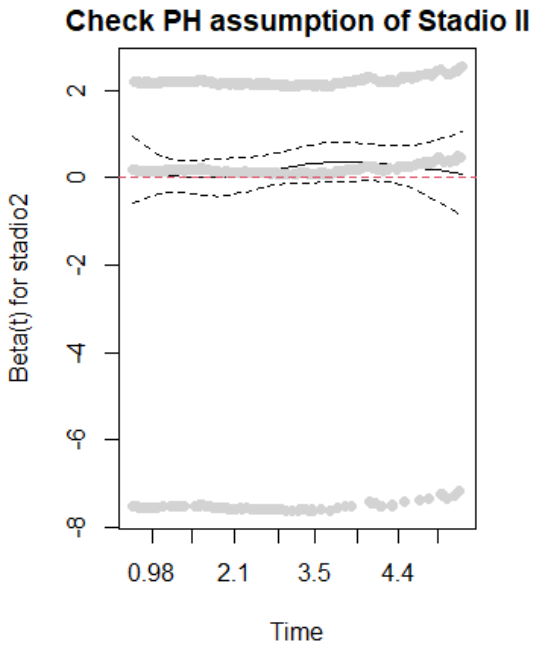
- Lo stadio di riferimento è lo **Stadio I**. Quindi, i soggetti allo **Stadio II** di malattia avrebbero **1,175** volte maggiore l'azzardo di morire rispetto ai soggetti allo **Stadio I**. Poiché l'intervallo di confidenza dell'hazard ratio, (exp(coeff)), comprende l'1 si conclude che questa differenza non è statisticamente significativa, e praticamente i soggetti allo **Stadio II** hanno lo stesso azzardo di morire come i sogetti allo **Stadio I**.

- I soggetti allo **Stadio III** hanno **1,539** volte maggiore l'azzardo di morire rispetto ai malati allo **Stadio I**. Dall'intervallo di confidenza concludiamo che nella peggiore delle ipotesi (upper .95), l'azzardo di morire dei malati allo **Stadio III** è il **doppio** dell'azzardo dei malati del gruppo di riferimento.

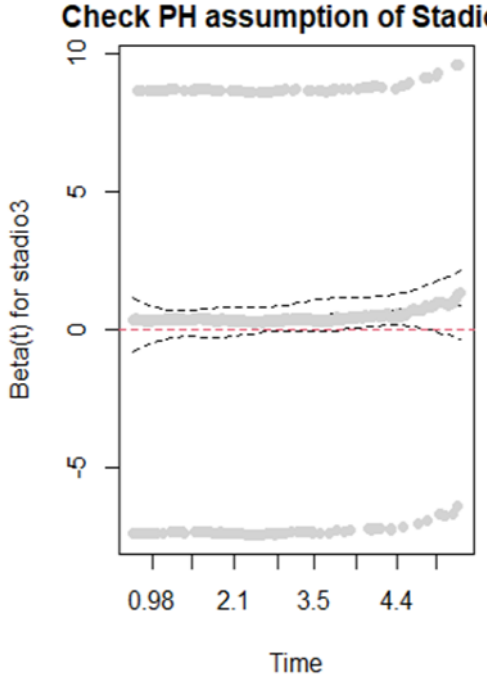
- Le persone che hanno il tumore al colon allo **Stadio IV** hanno **2,6** volte maggiore l'azzardo di morire rispetto ai malati allo **Stadio I**. Nella migliore ipotesi (limite inferiore) loro avrebbero l'azzardo raddoppiato rispetto al gruppo di riferimento, mentre nella peggiore ipotesi (limite superiore), l'azzardo di morire sarebbe **3,4** volte maggiore rispetto ai malati allo **Stadio I**.

VERIFICA: ASSUNTO DI AZZARDO PROPORZIONALO

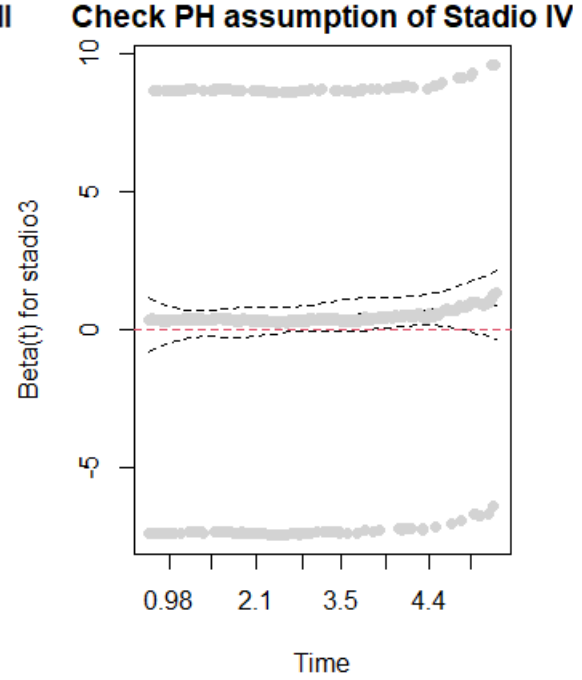
Stadio II



Stadio III



Stadio IV



L'applicazione del modello di Cox prevede l'assunto della proporzionalità di azzardi.

Questo assunto che è stato verificato usando il Schoenfeld test (di residui) che ha confermato la proporzionalità di azzardi tra diversi stadi di malattia.

9. ASSOCIAZIONE: SESSO vs. MORTALITÀ

L'associazione è stata valutata tramite regressione logistica (Generalized Linear Models)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.03593	0.07739	-0.464	0.642455
SexMale	0.42439	0.11189	3.793	0.000149

	OR	2.5 %	97.5%
(Intercept)	0.9647059	0.8289271	1.122725
Education medium/high	1.5286609	1.2276360	1.903499

- L'essere maschio è un fattore di rischio, perché i maschi hanno l'odds di morire **1.5** volte maggiore delle donne.

10. ASSOCIAZIONE: MORTALITÀ vs. ALTRE VARIABILI

→ Le variabili : **stadio**, **geneticm**, **smoke**, **education** ed **age** sono associate alla mortalità

		OR	CI
Stadio	Stadio II	1.34	(0.9929221 , 1.8160142)
	Stadio III	2.09	(1.3713977, 3.1979680)
	Stadio IV	14.24	(8.285601, 24,4665630)
geneticm		6.29	(3.7246681, 10.609802)
smoke (yes)		1.49	(1.1287853, 1.964621)
education (medium/high)		1.82	(1.217759, 2.732643)
age		1.08	(1.06470092, 1.09363976)

11. CONFONDIMENTO

→ L'unica variabile confondente nella valutazione dell'associazione tra sesso e mortalità è la variabile **age**.

	OR	2.5 %	97.5%
(Intercept)	0.023167	0.01194212	0.04494259
sex (Male)	1.648330	1.30452531	2.08274350
age	1.080599	1.06606383	1.09533229

→ Aggiustamento per la variabile age, modifica l'**OR** dal **1.53** al **1.65**, cioè più del 5%.

11. INTERAZIONE

Interazione tra variabile sex ed education

Sex: Female

education	Mortalità		Tot.
	0	1	
Low	325	289	614
Medium/High	15	39	54
Tot.	340	328	668

OR = 2.92 CI = (1.58, 5.41)

Sex: Male

education	Mortalità		Tot.
	0	1	
Low	234	340	574
Medium/High	23	39	62
Tot.	257	379	636

OR = 1.17 CI = (0.68, 2.01)

L'OR nei maschi non è significativamente diverso da 1, ovvero la mortalità di maschi non dipende dalla loro educazione.
Le donne con un'educazione medio/alta hanno, invece, l'odds di morire quasi **3** volte maggiore delle donne con un educazione bassa

MODELLO FINALE

1. Tutte e 6 variabili associate alla mortalità sono candidati da considerare fattori di rischio

2. La variabile age agisce da confondente nell'associazione tra la mortalità e sesso, e quindi dobbiamo includerle tutte e due, sesso e age.

3. C'è interazione tra l'educazione e sesso. Nonostante l'educazione interferisce diversamente sui maschi e sulle donne, viene inclusa nel modello per il suo contributo nella riduzione della devianza residua.

MODELLO FINALE

Analysis of Deviance Table					
Model: binomial, link: logit					
	Df	Deviance resid.	Df resid.	Deviance	Pr (>Chi)
NULL			1303	1798.4	
Sex	1	14.474	1302	1784.0	0.0001421
Stadio	3	159.749	1299	1624.2	< 2.2e-16
geneticm	1	64.917	1298	1559.3	7.81e-16
age	1	223.788	1297	1335.5	< 2.2e-16
smoke	1	9.510	1296	1326.0	0.0020439
education	1	10.059	1295	1315.9	0.0015157

- Tutte e sei le variabili sono state incluse nel modello finale poiché tutte contribuiscono in maniera significativa (in particolare age e Stadio) alla riduzione della varianza residua (Anova, test Chi-quadro).
- Sex, Stadio, geneticm, age, smoke ed education le possiamo considerare fattori di rischio della mortalità dopo la diagnosi del tumore al colon.