Mattia Pennati 07/07/2017

Analisi di sopravvivenza:

applicazione ad un campione utilizzato precedentemente da PiCnIc.

Sommario

- Descrizione
- Dati
- Obiettivo
- Considerazioni iniziali
- Modelli utilizzabili
- Considerazioni finali
- Conclusioni e scelte (utilizzo di modelli)
- Risultati
- Nuove Considerazioni
- Nuovi raggruppamenti
- Risultati dei nuovi raggruppamenti
- Conclusioni finali

Descrizione

Lo studio consiste nella creazione di un *survival model* per un campione precedentemente analizzato da PiCnIc, enfatizzando i risultati forniti dall'algoritmo di inferenza per le progressioni tumorali.

Dati

I dati da considerare per questa esperienza sono :

- un campione su cui è stato precedentemente applicato PiCnIc,
- modelli (MSI e MSS) e relativi grafi (DAG) ottenuti da PiCnIc,
- dati clinici del campione.

I dati clinici saranno utilizzati come *survival data* per l'analisi di sopravvivenza, per questo motivo si terrà conto solamente di :

- codice di identificazione del paziente
- stato (LIVING / DECEASED)
- ultimo tempo registrato (tempo del decesso / tempo di *censoring*)
- presenza di determinate mutazioni su geni.
    - si terrà conto della presenza delle mutazioni risultate come d'interesse dopo l'applicazione dell'algoritmo PiCnIc.

Obiettivo

Gli obiettivi dello studio sono i seguenti:

- Applicare analisi di sopravvivenza per ottenere un modello che stimi il più accuratamente possibile *survival function* e *hazard function* del campione considerato.
- Evidenziare la possibile influenza delle diverse progressioni tumorali stimate da PiCnIc sulla sopravvivenza stimata del campione.

Considerazioni iniziali

Le considerazioni che portano alla scelta di un determinato modello piuttosto che un altro sono:

- Si escludono a priori i *parametric survival models* perché basati su una distribuzione parametrica di probabilità non facilmente ipotizzabile.

- La maggior parte dei dati sono *right-censored* (ovvero è terminato il periodo di osservazione prima che il soggetto potesse riscontrare un'occorrenza dell'evento), serve quindi un modello che ne tenga considerazione nelle stime, limitando la scelta a :
    - modello non parametrico utilizzando gli stimatori di Kaplan-Meier per la *survival function* (ricavando l'hazard function) e di Nelson-Aalen per la *cumulative hazard function*.
    - modello semi-parametrico, considerando gli hazard (i fattori che influenzano l'*hazard function*) tra loro proporzionali; questo porta all'utilizzo di un *Cox PH(proportional hazards) Model*.

- Si vuole tenere conto delle considerazioni riguardanti gli ordini di precedenza nelle progressioni tumorali, basate sulle mutazioni su geni, ottenute da PiCnIc. Per questo motivo saranno necessarie una o più delle seguenti azioni sul campione :
    - "raggruppare" il campione in accordo con i risultati dell'algoritmo di inferenza (in entrambi i possibili modelli descritti precedentemente), per ottenere gruppi che considerino le progressioni stimate grazie ai quali effettuare stime di sopravvivenza più precise.
    - stratificare il campione (*sse* si utilizza un *Cox Model*), supponendo che ogni possibile gruppo abbia una *baseline hazard function* differente.

Modelli utilizzabili

La scelta del modello dipende da:

- accuratezza desiderata,
- fattori che si preferisce mettere in risalto,
- raggruppamento e/o stratificazione utilizzati.

Il raggruppamento è di fondamentale importanza e viene effettuato sul grafo corrispondente al modello ottenuto tramite PiCnIc, basandosi sulle discendenze date dagli archi. Si è scelto di considerare principalmente due casi :

- ***raggruppamento1*** :
    - Visitare il grafo partendo dalla/e radice/i e scendendo verso le foglie, eseguendo praticamente una BFS, per ottenere raggruppamenti del tipo (radice, discendenti).
    - In questo modo si considerano tutti i cammini che partono da una determinata radice come appartenenti allo stesso gruppo e verrebbero ignorate le relazioni logiche fra i nodi del grafo.
    - Ogni gruppo risulta composto da pazienti che presentano almeno una mutazione di quel gruppo.

- ***raggruppamento2*** :
    - Visitare il grafo partendo da ogni foglia e risalendo verso la/le radice/i , ottenendo dei cammini impropri (essendo il grafo un DAG), ovvero cammini radice-foglia che comprendono tutti i possibili percorsi alternativi che conducono dalla radice alla foglia.
    - In questo modo si considera ogni possibile cammino improprio come un gruppo distinto dagli altri.
      Inoltre vengono considerate le relazioni logiche fra i nodi del grafo, esse sono infatti il motivo della presenza di cammini alternativi radice-foglia in un gruppo.
    - Ogni gruppo risulta composto da soggetti che presentano le mutazioni di almeno uno dei cammini alternativi del gruppo.

In base a questi fattori si potrebbero fare principalmente **6** scelte:

- *Modelli non parametrici* :
  Questi modelli non consentono di considerare le mutazioni come possibili parametri di influenza per il modello e presuppongono che l’intero campione abbia la stessa hazard function.

Nonostante ciò permetterebbero comunque di ottenere una stima differente per ogni gruppo se per ognuno si creasse un differente modello.

I primi due possibili modelli sono quindi :

1) **Modello non parametrico applicato a raggruppamento1.**
2) **Modello non parametrico applicato a raggruppamento2.**

- *Cox Models* :
  In questi modelli si presuppone che gli unici fattori ad influenzare le stime siano gli *effect parameters*, ovvero parametri che corrispondono all' effetto di diverse covariate (in questo caso la presenza o meno di determinate mutazioni andrebbe a costituire un vettore di covariate per ogni soggetto).
  Si presupporrebbe però che ogni covariata abbia la stessa baseline hazard function, ovvero verrebbe considerata alla pari di ogni altra.
  Questo porterebbe ad un modello caotico influenzato da troppi parametri, ma si aggirerebbe il problema sfruttando i raggruppamenti e utilizzando un diverso modello per ogni gruppo, cosa che porterebbe anche ad ottenere una differente baseline hazard function per ogni gruppo.

  Si potrebbero usare quindi due ulteriori modelli :

  3) **Cox Model applicato a *raggruppamento1*.**
  4) **Cox Model applicato a *raggruppamento2*.**

- *Stratified Cox Models* :
  In questi modelli si agisce come in qualsiasi altro Cox Model, ma si "stratificano" le covariate di influenza.
  In questo modo si otterrebbe un unico modello dove ogni *strato* ha però una *baseline hazard function* diversa da quella degli altri strati, ottenendo quindi una curva di sopravvivenza (e ogni altra quantità derivabile dal modello) distinta per ogni strato.
  Gli *effect parameters* sarebbero comunque le influenze delle diverse covariate, ma sarebbero differenti rispetto a quelle dei modelli precedenti, essendo dipendenti dallo strato di appartenenza.
  Per costituire gli strati si sfrutterebbe comunque un raggruppamento, considerando le mutazioni di ogni gruppo come covariate di ogni strato.

  Gli ultimi modelli possibili sono quindi :

  5) **Stratifed Cox Model applicato a *raggruppamento1*.**
  6) **Stratifed Cox Model applicato a *raggruppamento2*.**

Considerazioni finali

Data la numerosità non elevata del campione in considerazione (TCGA-COADREAD), e di campioni simili, l'utilizzo di **raggruppamento2** rischierebbe di produrre gruppi poco consistenti.
Il grafo in output da PiCnIc è un DAG probabilistico contenente anche mutazioni rare (*es*. presenti in 1 solo paziente), rendendo quindi difficile la suddivisione del campione in gruppi. Il problema principale consiste nel fatto che un paziente rientra in un determinato gruppo *sse* possiede tutte le mutazioni presenti in uno dei cammini alternativi che caratterizzano il gruppo.
Risulta quindi improbabile ottenere gruppi composti da un numero di soggetti abbastanza elevato da poterci fare una stima di sopravvivenza affidabile.
Questa considerazione frena quindi l'utilizzo dei modelli basati su **raggruppamento2** (modelli 2, 4 e 6), che potrebbero essere utilizzati solo permettendo l'appartenenza ad un determinato gruppo anche a pazienti che hanno almeno una certa probabilità delle mutazioni di un certo cammino.

Sfruttando il **raggruppamento1** si presupporrebbe invece di avere a disposizione gruppi abbastanza ampi da poter essere analizzati con risultati accettabili.
Dall'altro canto però questo raggruppamento non avrebbe una correlazione con l'output di PiCnIc così elevata come quella dell'altro possibile raggruppamento.

I modelli non parametrici (1 e 2) non riscontrano particolari problemi, ma non permettono di evidenziare l'influenza della presenza di determinate mutazioni su geni, essendo basati solo su stime non parametriche ottenute considerando:

- numero di decessi ad ogni istante di tempo
- numero di soggetti considerati a rischio ad ogni istante di tempo

In questo modo si otterrebbero comunque stime di sopravvivenza attendibili e differenti per i gruppi, ma enfatizzerebbero la differenza fra un gruppo e l'altro senza mettere in risalto l'effetto di ogni mutazione presente in un gruppo.

Nei modelli di Cox (3, 4, 5, 6), stratificati o meno, l'effetto di ogni singola mutazione sarebbe invece presente singolarmente; ad ogni covariata del modello infatti corrisponde un parametro rappresentante l'influenza della covariata sulla sopravvivenza degli individui del campione.
In questo modo si potrebbero enfatizzare sia le differenze fra i vari gruppi che l'effetto delle singole mutazioni all'interno dei gruppi.

Conclusioni e scelte (utilizzo di modelli)

In accordo con le considerazioni finali:

- Si esclude l'utilizzo di modelli non parametrici (1 e 2) perchè si desidera evidenziare l'effetto di ogni mutazione.

- Viene momentaneamente messo in stallo, ma non escluso, l'utilizzo di modelli basati sul **raggruppamento2** (modelli 4 e 6), che garantirebbe la maggior corrispondenza con i risultati di PiCnIc, in attesa di considerazioni che ne rendano possibile l'utilizzo.

- La scelta iniziale ricade quindi sull'utilizzo di un *Cox Models* (o *Stratified Cox Models*) basato sul **raggruppamento1**, limitando la scelta ai modelli 3 e 5. Verranno portati parallelamente avanti entrambi perchè:

    - si ritiene interessante confrontarne le possibili differenze

    - esprimono la stessa considerazione, ovvero l'effetto dell'appartenenza ad un determinato gruppo e l'effetto di ogni mutazione di quel gruppo.

    - permettono incroci "ibridi", potendo stratificare anche all'interno di un determinato gruppo, supponendo magari che alcune mutazioni interne al gruppo abbiano un'influenza maggiore rispetto ad altre.

Risultati

Una volta eseguita un'analisi di sopravvivenza sfruttando i modelli 3 e 5 si ottengono dei risultati poco apprezzabili che non riescono ad evidenziare ciò che ci si poteva aspettare.

OSS: Si considerano i risultati solo per la parte del campione con MSI_status = 'MSS'. Questo perché l'altra parte, quella con MSI_status = 'MSI-HIGH', è numericamente irrilevante.

- Per quanto riguarda l'utilizzo del **modello 3** :

  Raggruppando secondo il **raggruppamento1** si ottengono 10 gruppi distinti. Per ognuno di questi gruppi si ottiene un modello di **Cox** che utilizza solo le

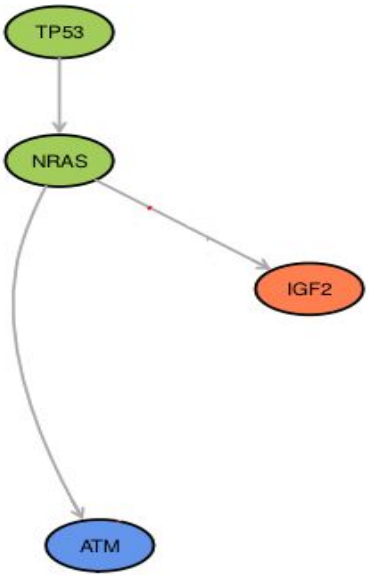
mutazioni di quel gruppo come fattori di influenza.
Da ogni modello si ottengono poi :

- **survival function**
- **baseline cumulative hazard function**
- **cox snell residuals** : metodo utilizzato per controllare l'**overall fit** del modello
- **scaled schoenfeld residuals** : metodo utilizzato per controllare l'ipotesi di proporzionalità di ogni singola covariata.

Si ottengono dei modelli validi solamente per 6 gruppi su 10, dei restanti gruppi infatti 3 non contengono eventi mentre per 1 una covariata (mutazione) non è posseduta da nessun paziente, rendendo quindi impossibile ottenere un modello di Cox.
Inoltre dei 6 gruppi per i quali si ottiene un modello di Cox si ottengono dei risultati sensati per ogni oggetto ottenuto (funzioni di sopravvivenza/hazard e residui) solamente per 2, per i quali i risultati ottenuti sono i seguenti :

Gruppo 2.

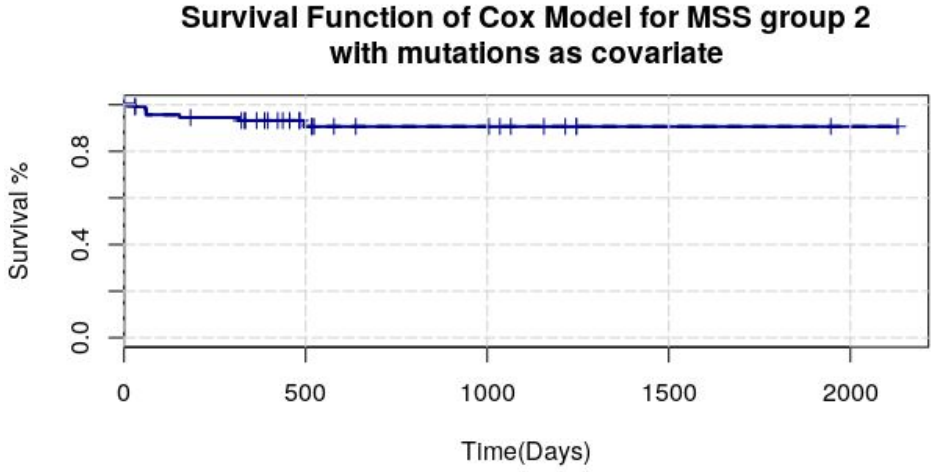
Il gruppo rispecchia la seguente porzione del grafo relativo a MSI_status = 'MSS' ottenuto da PiCnIc.

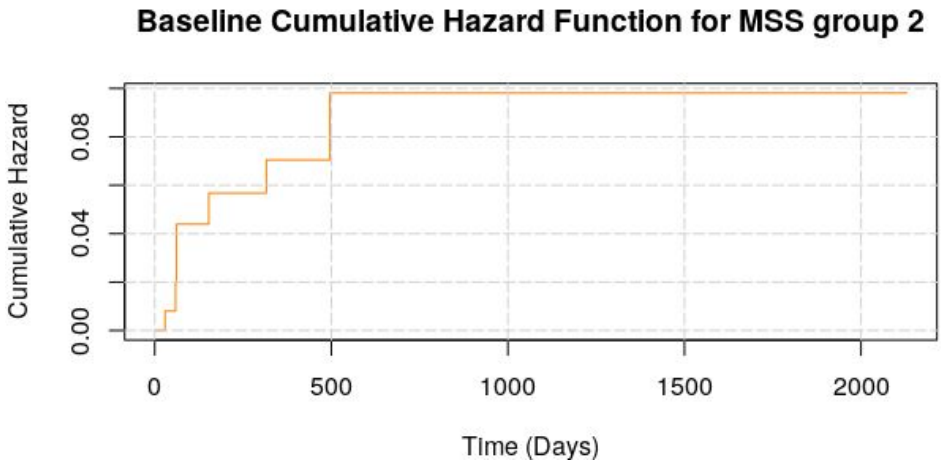


Si ottiene il seguente modello di Cox, dal quale si evincono quali covariate hanno un'influenza negativa sulla sopravvivenza con la relativa entità (Quelle che hanno un parametro di efficienza, rappresentato da '*coef*' positivo, tanto è maggiore, tanto è più influente la mutazione) e il relativo p-value.

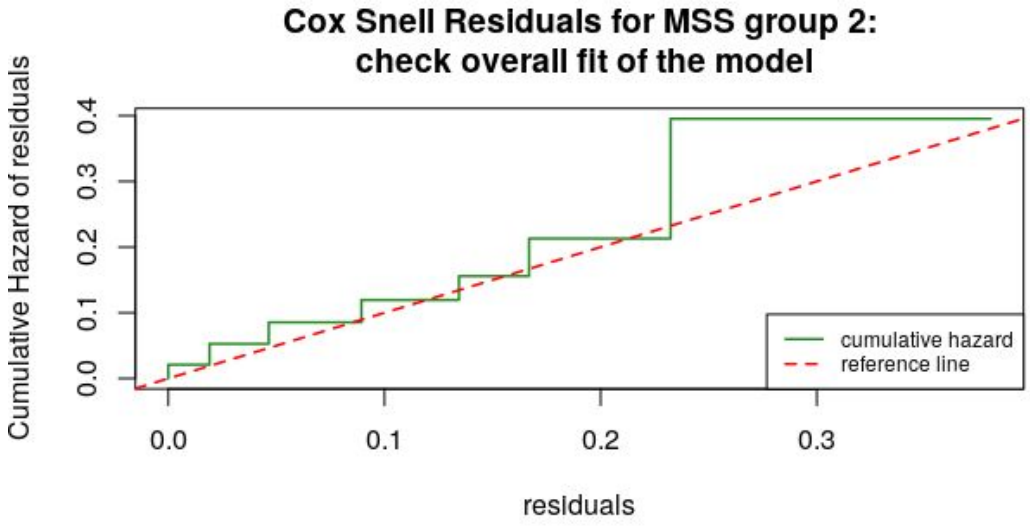
```
Call:
coxph(formula = formula, data = data)

  n= 94, number of events= 7

                       coef exp(coef)  se(coef)     z Pr(>|z|)
TP53_Mutation      1.752e+01 4.081e+07 3.616e+04 0.000    1.000
NRAS_Mutation      4.930e-01 1.637e+00 8.407e-01 0.586    0.558
ATM_Deletion              NA        NA 0.000e+00    NA       NA
IGF2_Amplification 4.205e-02 1.043e+00 3.745e+04 0.000    1.000

                   exp(coef) exp(-coef) lower .95 upper .95
TP53_Mutation      4.081e+07  2.451e-08    0.0000       Inf
NRAS_Mutation      1.637e+00  6.108e-01    0.3151     8.506
ATM_Deletion              NA         NA        NA        NA
IGF2_Amplification 1.043e+00  9.588e-01    0.0000       Inf

Concordance= 0.559  (se = 0.087 )
Rsquare= 0.01   (max possible= 0.397 )
Likelihood ratio test= 0.91  on 3 df,   p=0.8232
Wald test            = 0.34  on 3 df,   p=0.9516
Score (logrank) test = 0.67  on 3 df,   p=0.8809
```

Il modello produce le seguenti survival function e baseline hazard function:

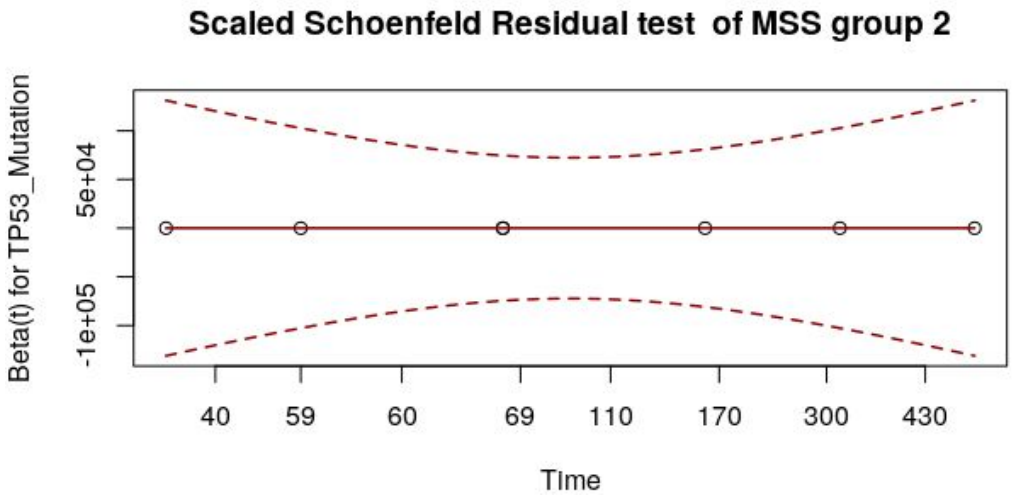


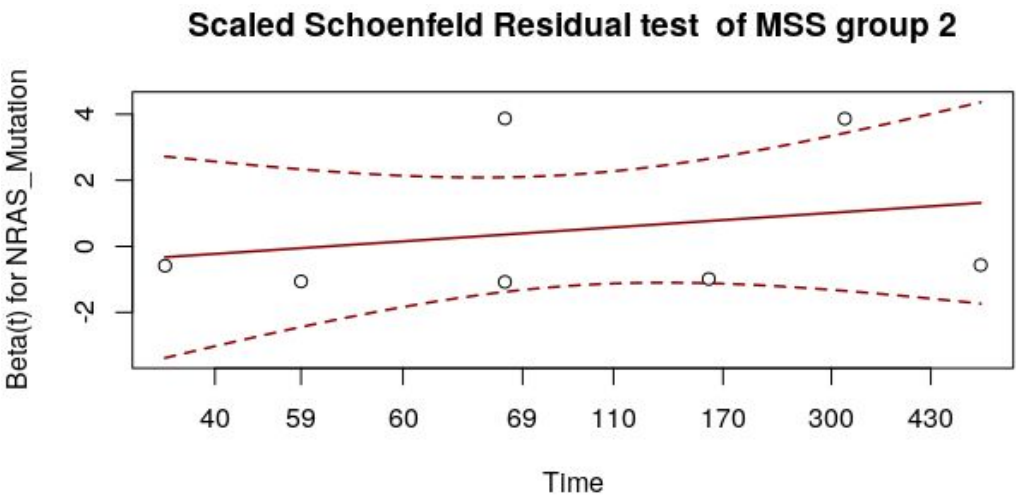


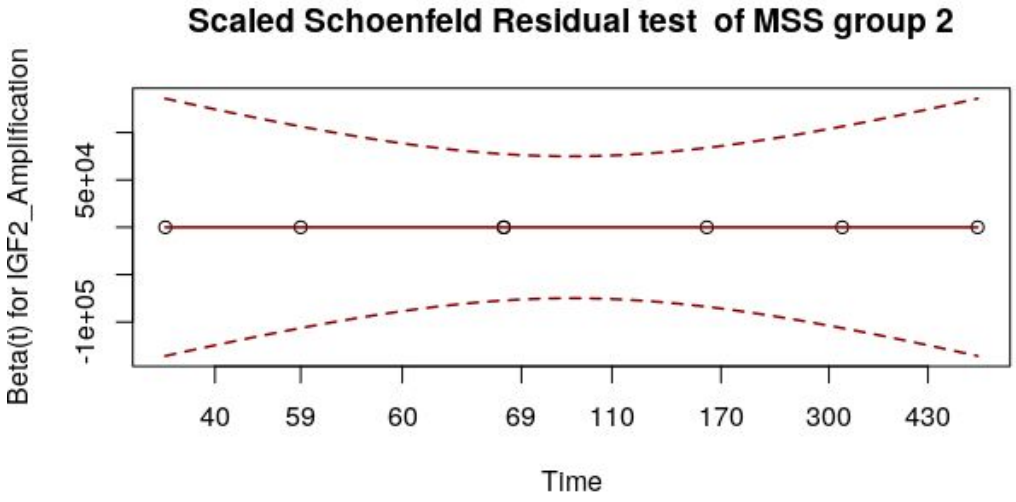
Il modello risulta attendibile, il risultato fornito tramite cox snell residuals è infatti una funzione a gradino che segue abbastanza fedelmente la retta di riferimento *x=y* come si può notare dal seguente grafico :



Per quanto riguarda le covariate il modello non riesce a stimare un parametro di influenza per ‘ATM_Deletion’, ma le restanti 3 covariate rispettano l’ipotesi di proporzionalità (‘TP53_Mutation’ e ‘IGF2_Amplification’ a pieno, mentre per ‘NRAS_Mutation’ risulta un po’ forzata) come si evince dai seguenti grafici :

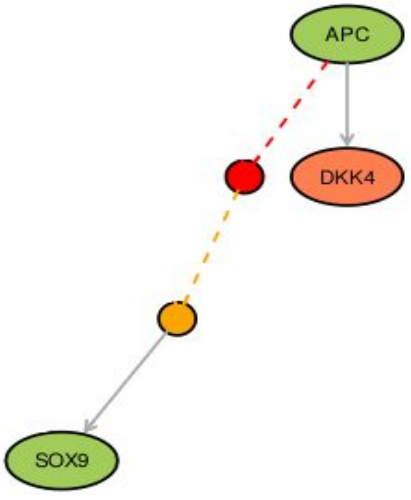






Gruppo 4.

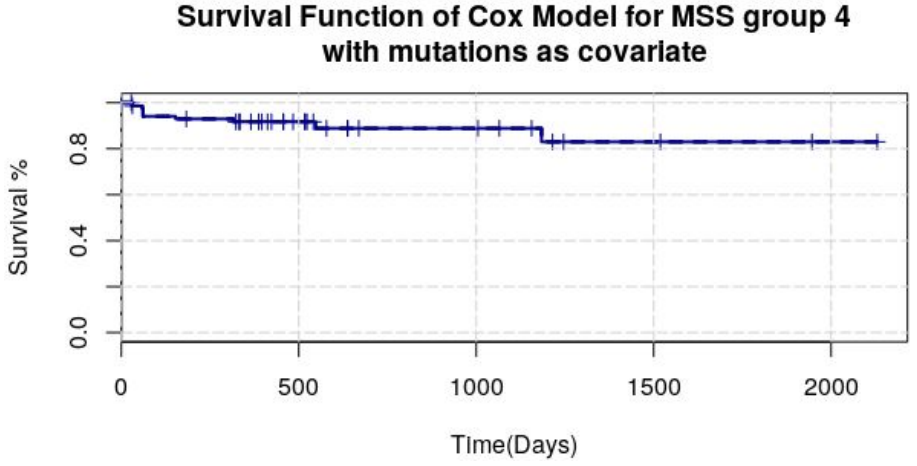
Il gruppo rispecchia la seguente porzione del grafo relativo a MSI_status = 'MSS' ottenuto da PiCnIc.

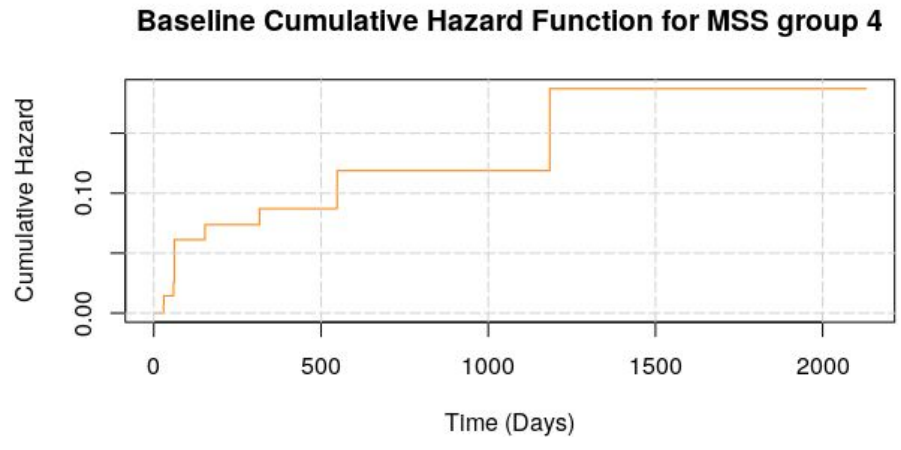


Si ottiene il seguente modello di Cox, dal quale si evincono quali covariate hanno un'influenza negativa sulla sopravvivenza con la relativa entità (Quelle che hanno un parametro di efficienza, rappresentato da '*coef*' positivo, tanto è maggiore, tanto è più influente la mutazione) e il relativo p-value.

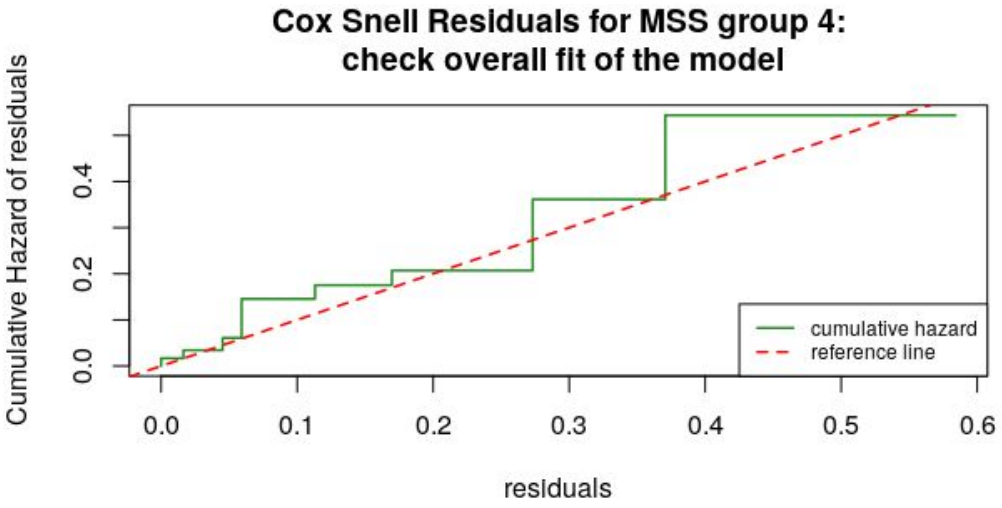
```
Call:
coxph(formula = formula, data = data)

  n= 120, number of events= 10

                        coef  exp(coef)   se(coef)      z Pr(>|z|)
APC_Mutation              NA         NA  0.000e+00     NA       NA
DKK4_Amplification -1.707e+01  3.871e-08  7.387e+03 -0.002    0.998
SOX9_Mutation       3.052e-01  1.357e+00  1.071e+00  0.285    0.776

                   exp(coef) exp(-coef) lower .95 upper .95
APC_Mutation              NA         NA        NA        NA
DKK4_Amplification 3.871e-08  2.583e+07    0.0000       Inf
SOX9_Mutation      1.357e+00  7.370e-01    0.1663     11.07

Concordance= 0.513  (se = 0.053 )
Rsquare= 0.008   (max possible= 0.436 )
Likelihood ratio test= 1.02  on 2 df,   p=0.6006
Wald test            = 0.08  on 2 df,   p=0.9602
Score (logrank) test = 0.57  on 2 df,   p=0.7517
```

Il modello produce le seguenti survival function e baseline hazard function:

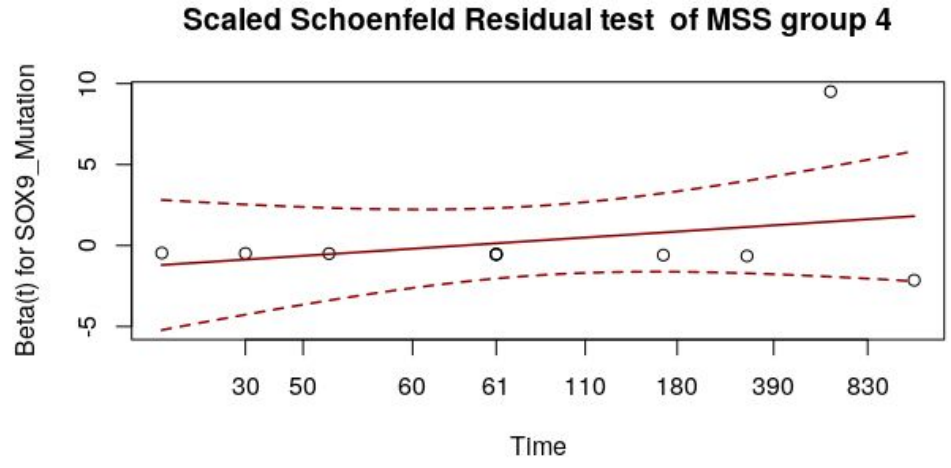


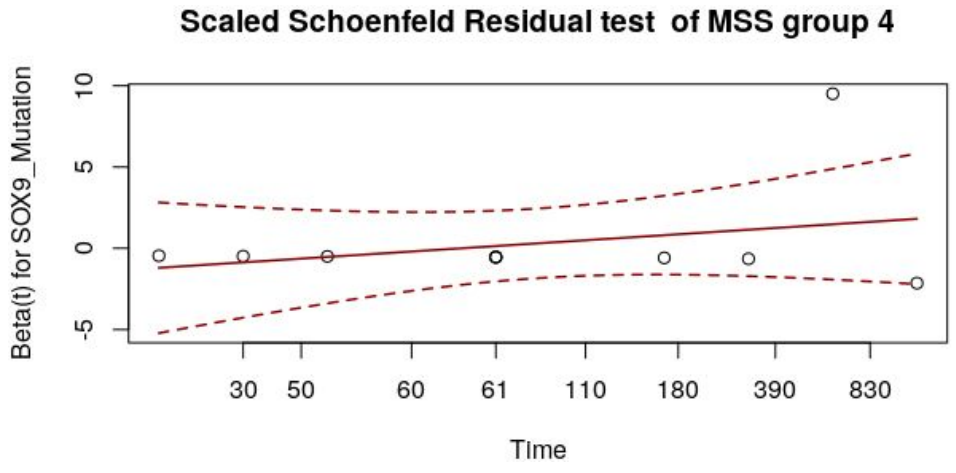


Il modello risulta attendibile, il risultato fornito tramite cox snell residuals è infatti una funzione a gradino che segue abbastanza fedelmente la retta di riferimento *x*=*y* come si può notare dal seguente grafico :



Per quanto riguarda le covariate il modello non riesce a stimare un parametro di influenza per 'APC_Mutation', ma le restanti 2 covariate rispettano l'ipotesi di proporzionalità ('DKK4_Amplification' a pieno, mentre 'SOX9_Mutation' risulta un po' forzata, ma decisamente accettabile) come si evince dai seguenti grafici :





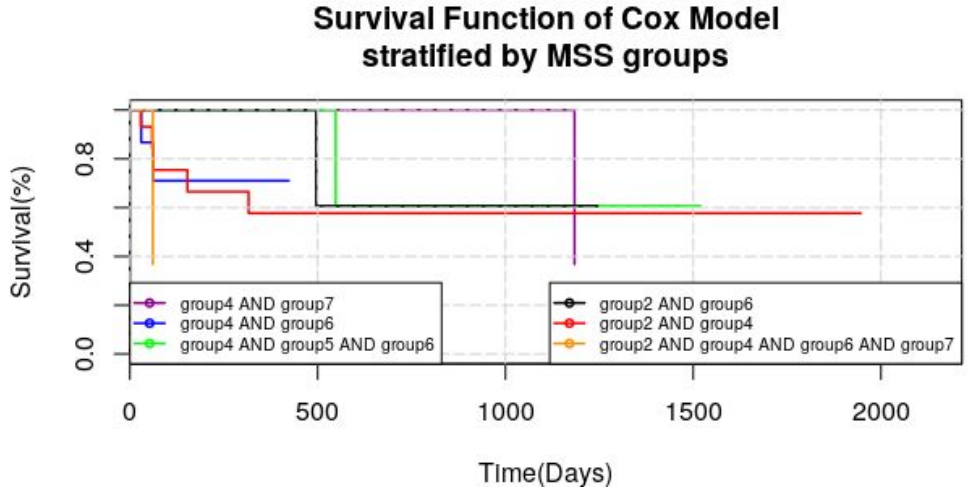
- Per quanto riguarda l'utilizzo del modello **5** :

  Sono stati considerati solo i gruppi che nel tentativo precedente avevano un modello di Cox valido, ottenendo un modello di cox stratificato per appartenenza ad uno o più di quei determinati gruppo (visto che i gruppi non sono esclusivi ed un paziente può quindi appartenere a più di un gruppo), dove le covariate sono tutte le mutazioni possibili fra tutti i gruppi.
  Il risultato è deludente e non affidabile, in quanto la maggior parte delle covariate hanno un parametro di influenza nullo; inoltre un paziente potendo appartenere a più gruppi rende minore la numerosità dei campioni relativi ai singoli gruppi.
  Per questi motivi le funzioni di sopravvivenza di tutti i singoli gruppi sono una linea retta a y=1 (survival 100%), mentre la sopravvivenza risulta più affidabile nelle stratificazioni relative a pazienti che appartengono a più gruppi.

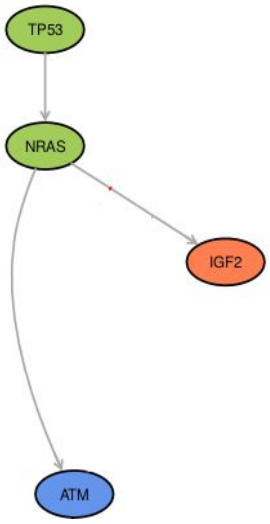
```
Call:
coxph(formula = formula, data = dataset_notnullgroups)

  n= 151, number of events= 11

                          coef  exp(coef)   se(coef)      z Pr(>|z|)
TP53_Mutation               NA         NA  0.000e+00     NA       NA
NRAS_Mutation       -2.876e-01  7.500e-01  1.127e+00 -0.255    0.799
ATM_Deletion                NA         NA  0.000e+00     NA       NA
IGF2_Amplification          NA         NA  0.000e+00     NA       NA
FBXW7_Mutation       2.120e+01  1.615e+09  9.091e+04  0.000    1.000
BRAF_Amplification          NA         NA  0.000e+00     NA       NA
CTNNB1_Mutation             NA         NA  0.000e+00     NA       NA
SOX9_Mutation               NA         NA  0.000e+00     NA       NA
SOX9_Amplification          NA         NA  0.000e+00     NA       NA
APC_Mutation                NA         NA  0.000e+00     NA       NA
DKK4_Amplification  -2.027e+01  1.570e-09  3.091e+04 -0.001    0.999
ARID1A_Mutation             NA         NA  0.000e+00     NA       NA
TP53_Deletion               NA         NA  0.000e+00     NA       NA
KRAS_Mutation       -2.221e+01  2.256e-10  4.708e+04  0.000    1.000
SMAD4_Mutation      -2.843e-07  1.000e+00  8.155e+04  0.000    1.000
ATM_Mutation                NA         NA  0.000e+00     NA       NA
DKK4_Mutation               NA         NA  0.000e+00     NA       NA
PIK3CA_Mutation      5.261e-07  1.000e+00  8.155e+04  0.000    1.000
TCF7L2_Deletion      4.559e+01  6.271e+19  1.363e+05  0.000    1.000
ERBB2_Amplification -1.009e+00  3.644e-01  7.777e+04  0.000    1.000
ERBB2_Mutation       1.957e+01  3.140e+08  4.648e+05  0.000    1.000
PTEN_Mutation               NA         NA  0.000e+00     NA       NA
PTEN_Deletion               NA         NA  0.000e+00     NA       NA
SMAD4_Deletion              NA         NA  0.000e+00     NA       NA
FAM123B_Mutation            NA         NA  0.000e+00     NA       NA
```

```
                    exp(coef) exp(-coef) lower .95 upper .95
TP53_Mutation              NA         NA        NA        NA
NRAS_Mutation        7.500e-01  1.333e+00   0.08231     6.834
ATM_Deletion               NA         NA        NA        NA
IGF2_Amplification         NA         NA        NA        NA
FBXW7_Mutation       1.615e+09  6.190e-10   0.00000       Inf
BRAF_Amplification         NA         NA        NA        NA
CTNNB1_Mutation            NA         NA        NA        NA
SOX9_Mutation              NA         NA        NA        NA
SOX9_Amplification         NA         NA        NA        NA
APC_Mutation               NA         NA        NA        NA
DKK4_Amplification   1.570e-09  6.368e+08   0.00000       Inf
ARID1A_Mutation            NA         NA        NA        NA
TP53_Deletion              NA         NA        NA        NA
KRAS_Mutation        2.256e-10  4.433e+09   0.00000       Inf
SMAD4_Mutation       1.000e+00  1.000e+00   0.00000       Inf
ATM_Mutation               NA         NA        NA        NA
DKK4_Mutation              NA         NA        NA        NA
PIK3CA_Mutation      1.000e+00  1.000e+00   0.00000       Inf
TCF7L2_Deletion      6.271e+19  1.595e-20   0.00000       Inf
ERBB2_Amplification  3.644e-01  2.744e+00   0.00000       Inf
ERBB2_Mutation       3.140e+08  3.184e-09   0.00000       Inf
PTEN_Mutation              NA         NA        NA        NA
PTEN_Deletion              NA         NA        NA        NA
SMAD4_Deletion             NA         NA        NA        NA
FAM123B_Mutation           NA         NA        NA        NA

Concordance= 0.691  (se = 0.154 )
Rsquare= 0.071   (max possible= 0.192 )
Likelihood ratio test= 11.12  on 9 df,   p=0.2674
Wald test            = 0.07  on 9 df,   p=1
Score (logrank) test = 11.03  on 9 df,   p=0.2737
```

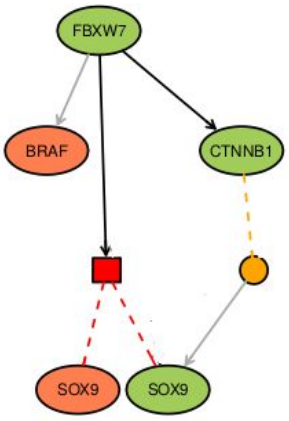
Per questo motivo si è optato per un modello di Cox stratificato che ignori le mutazioni come covariate e consideri solo i differenti strati per poterne mettere in risalto le differenze.
Il modello sarà quindi un modello che non subirà influenza da nessuna covariata, risultando quindi una semplice stima della funzione di sopravvivenza per diversi gruppi/unione di gruppi.
Le funzioni di sopravvivenza non nulle sono le seguenti :



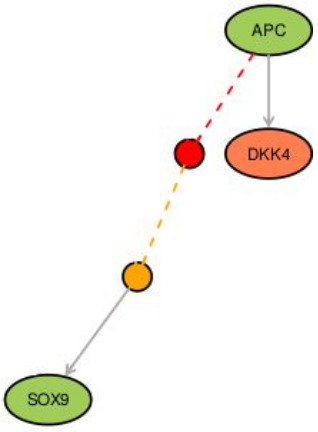
I gruppi in questione sono i seguenti :



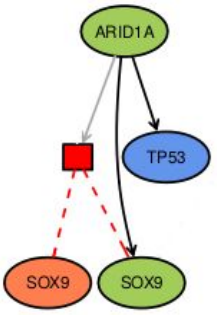
group 2



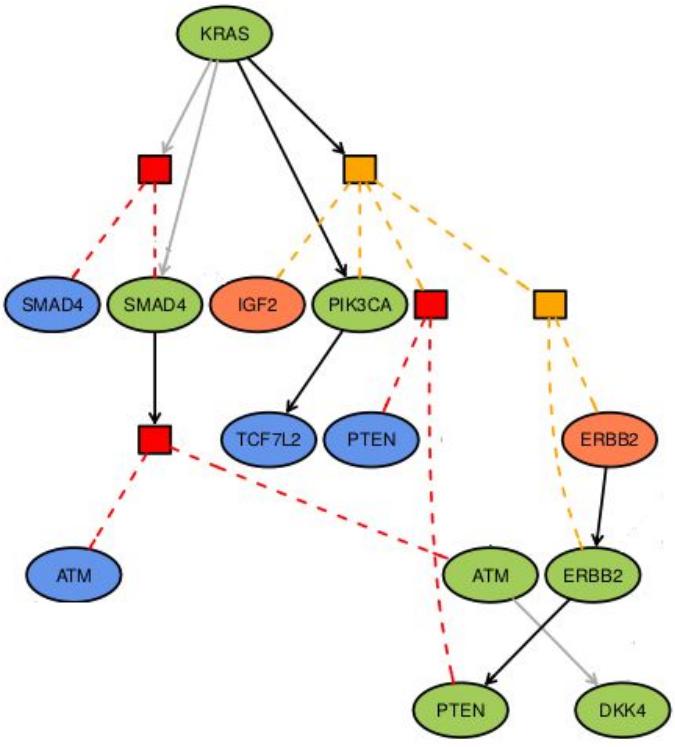
group 3



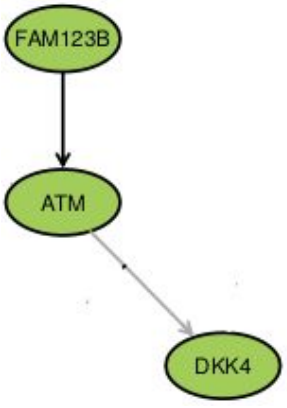
group 4



group 5



group 6



group 7

Nuove Considerazioni

I risultati dei modelli utilizzati sono piuttosto deludenti e poco affidabili.
Questo è dovuto quasi esclusivamente alla **numerosità del campione**; essa infatti è troppo bassa se paragonata al numero dei raggruppamenti ottenuti e al numero di mutazioni.

- La bassa presenza di eventi registrati durante il tempo di osservazione dei soggetti rende quindi molti gruppi inconsistenti per poter eseguire un'analisi di sopravvivenza.

- La stessa analisi con un campione 8/10 volte più numeroso avrebbe sicuramente permesso di ottenere risultati più affidabili e esaustivi.

- La parte del campione con 'MSI_status = MSI-HIGH' è talmente poco numerosa da poter essere considerata inutilizzabile per un'analisi di sopravvivenza, verrà quindi ignorata nelle successive modellazioni.

- Con un campione di queste dimensioni e con così tante mutazioni di interesse i parametri di influenza di ogni mutazione stimati tramite cox non risultano affidabili, rendendo i risultati utili solo a livello di confronto fra loro.

A scopo informativo verrà quindi eseguita un'analisi solo su una parte del dataset con MSI_status = 'MSS' considerando solamente i gruppi 2 e 4 , ovvero i gruppi che avevano dato i risultati migliori nei precedenti modelli.

Inoltre dato che l'obiettivo principale era evidenziare l'influenza sulla sopravvivenza dei cammini rappresentanti progressioni tumorali ottenuti da PiCnIc si è deciso di provare un approccio che consideri raggruppamenti differenti.
Questi nuovi raggruppamenti si concentreranno sulla **profondità** dei nodi (rappresentanti mutazioni) nel grafo.
Ogni **livello di profondità** sarà quindi visto come un gruppo differente e l'obiettivo sarà quello di provare a dimostrare che la sopravvivenza dei pazienti peggiorerà tanto più il loro gruppo di appartenenza rappresenterà un livello di profondità maggiore.
Per numerosità del campione e natura probabilistica dei grafi ottenuti da PiCnIc si è deciso di considerare solo i **primi 3 livelli** di profondità.

Caso specifico

Per sfruttare i pochi risultati interessanti ottenuti dai tentativi precedenti si considerino solamente i gruppi 2 e 4 ottenuti dal grafo MSS (quelli per i quali si ottengono modelli, sopravvivenza e residui di senso compiuto).

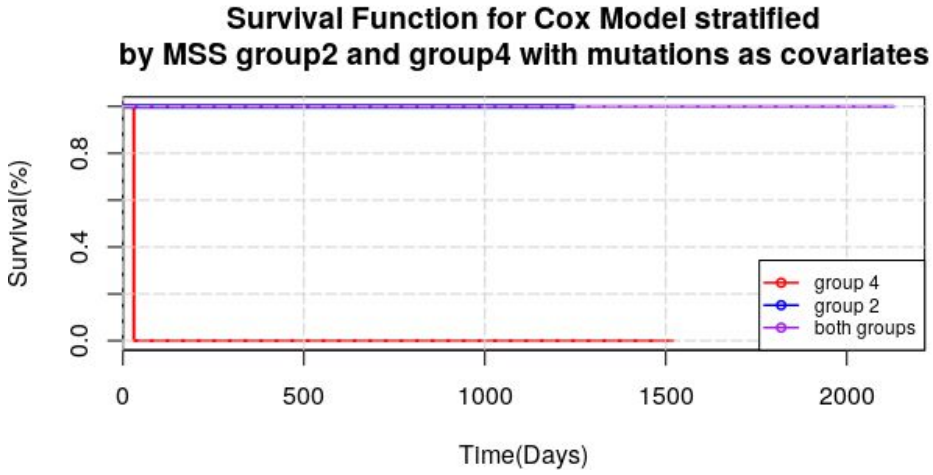
Considerando come dataset un sottoinsieme di quello iniziale comprendente solo i pazienti appartenenti al gruppo 2 e/o al gruppo 4 si sono eseguiti i seguenti tentativi:

- **Cox model stratificato** per il gruppo gruppo di appartenenza e considerano le mutazione solo dei gruppi 2 e 4 come covariate.
  I risultati seguono più o meno quelli ottenuti in precedenza, avendo quindi ancora

troppe covariate per cui non si riesce a stimare l'influenza e continuando ad avere il problema delle sovrapposizioni fra gruppi.

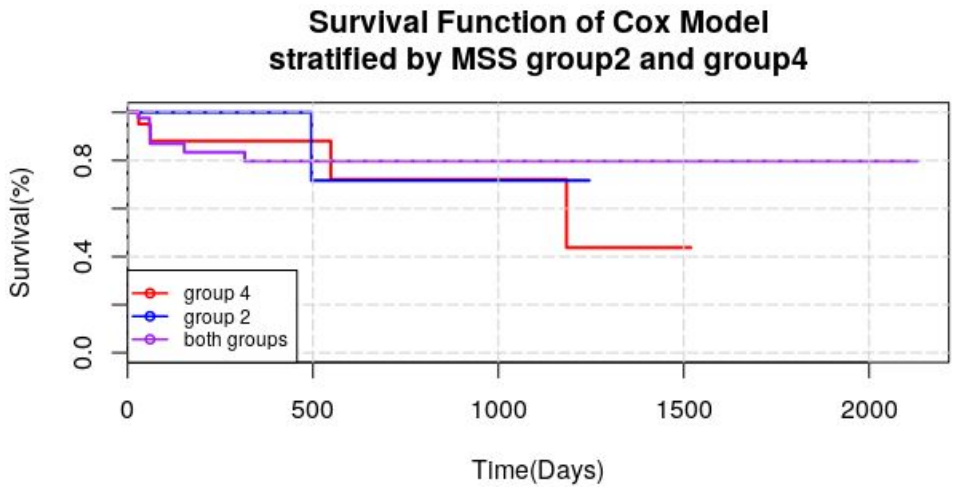
```
Call:
coxph(formula = formula_g2g4, data = dataset_g2g4)

  n= 138, number of events= 11

                        coef  exp(coef)   se(coef)      z Pr(>|z|)
TP53_Mutation       1.854e+01  1.128e+08  5.393e+04  0.000    1.000
NRAS_Mutation       4.312e-01  1.539e+00  8.462e-01  0.510    0.610
ATM_Deletion               NA         NA  0.000e+00     NA       NA
IGF2_Amplification  4.129e-02  1.042e+00  5.648e+04  0.000    1.000
APC_Mutation               NA         NA  0.000e+00     NA       NA
DKK4_Amplification -1.810e+01  1.383e-08  1.404e+04 -0.001    0.999
SOX9_Mutation      -3.105e-01  7.331e-01  1.211e+00 -0.256    0.798

                   exp(coef) exp(-coef) lower .95 upper .95
TP53_Mutation      1.128e+08  8.867e-09   0.00000       Inf
NRAS_Mutation      1.539e+00  6.497e-01   0.29309     8.082
ATM_Deletion              NA         NA        NA        NA
IGF2_Amplification 1.042e+00  9.596e-01   0.00000       Inf
APC_Mutation              NA         NA        NA        NA
DKK4_Amplification 1.383e-08  7.231e+07   0.00000       Inf
SOX9_Mutation      7.331e-01  1.364e+00   0.06834     7.864

Concordance= 0.587  (se = 0.107 )
Rsquare= 0.012   (max possible= 0.34 )
Likelihood ratio test= 1.63  on 5 df,   p=0.8982
Wald test            = 0.33  on 5 df,   p=0.9971
Score (logrank) test = 1.03  on 5 df,   p=0.9598
```

Si possono comunque notare differenze nei risultati di sopravvivenza, ma le funzioni sono assolutamente non attendibili e di scarsa accuratezza.



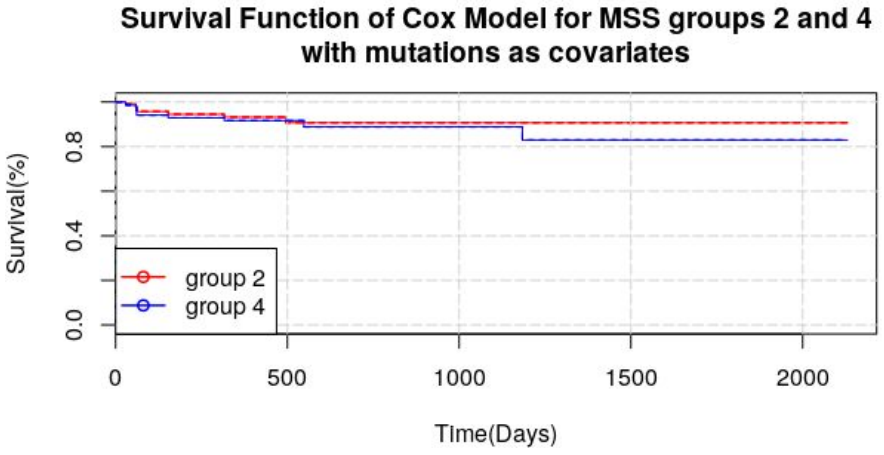
- **Cox model stratificato** per il gruppo di appartenenza e senza l'ausilio di covariate.

Si ottiene un modello che non ha parametri di influenza e utilizza la stratificazione solo per stimare la sopravvivenza dei diversi strati.
Le funzioni di sopravvivenza risultanti sono :



- Gruppi 2 e 4 considerati **separatamente.**
  Si sfruttano i modelli e le funzioni di sopravvivenza ottenute nel tentativo precedente tramite il modello 5 in modo da poterli confrontare e in modo da poterci effettuare dei test statistici per controllarne la differenza.

  Il confronto delle funzioni di sopravvivenza considerando separatamente i due modelli ottenuti in precedenza mostra una differenza, anche se non troppo evidente.



  Per controllare effettivamente la possibile differenza fra i due gruppi si applica quindi un **log rank test.**
  Si assume quindi come ipotesi **H0** che i due differenti gruppi abbiano la stessa distribuzione di sopravvivenza, mentre come ipotesi alternativa che seguano distribuzioni differenti.
  Eseguendo il test non si riesce a negare l'ipotesi nulla, non potendo quindi

confermare la differenza fre le due funzioni di sopravvivenza, come si può evincere dai risultati:

```
Call:
survdiff(formula = SurvObj ~ group, data = dataset_g2g4)

             N Observed Expected (O-E)^2/E (O-E)^2/V
group=2     18        1     1.51     0.172    0.2017
group=4     44        4     3.23     0.184    0.2642
group=both 76        6     6.26     0.011    0.0258

 Chisq= 0.4  on 2 degrees of freedom, p= 0.83
```

Nuovi raggruppamenti

Come considerato successivamente ai risultati dei modelli scelti inizialmente si è deciso di provare a basarsi sulla profondità, ottenendo quindi due nuovi raggruppamenti:

- **levelGrouping**
  Si creano 3 livelli:
  - level 1 : solo mutazioni che sono radici nel grafo (profondità = 1)
  - level 2 : solo mutazioni che sono figlie di radici nel grafo (profondità = 2)
  - level 3 : solo mutazioni che sono figlie di figlie di radici nel grafo (profondità = 3)

  Un paziente appartiene ad un solo determinato livello e, nello specifico, apparterrà a :
  - level 1 : se ha solo mutazioni radice
  - level 2 : se ha sia mutazioni radice che mutazioni a profondità 2
  - level 3 : se ha mutazioni radice, mutazioni di profondità 2 e mutazioni di profondità 3.

  Un paziente che appartiene ad un gruppo con profondità maggiore non appartiene ai gruppi con profondità minore.

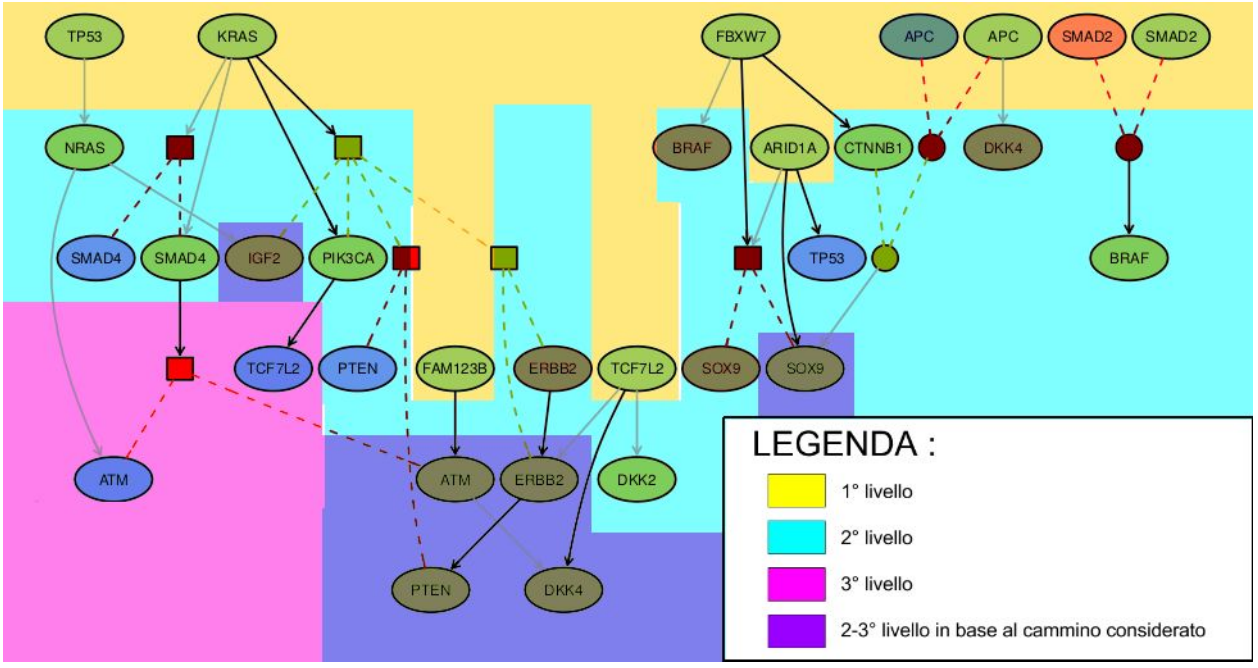
- **pathLevelGrouping**
  Ogni gruppo è un sottoinsieme dei gruppi ottenuti con con la considerazione precedente.
  La limitazione che viene aggiunta è che un paziente appartiene al livello 2 o al livello 3 **sse** le mutazioni che possiede nei vari livelli sono lungo un **cammino del grafo.**

Un paziente che appartiene ad un gruppo ottenuto con levelGrouping non apparterrà per forza anche al relativo gruppo ottenuto con pathLevelGrouping, mentre il contrario è assicurato.
Questo perchè con il primo raggruppamento si considera un paziente appartenente ad

un determinato gruppo se ha una qualsiasi delle mutazioni del livello relativo al gruppo, mentre nel secondo caso apparterrebbe al gruppo solo se avesse almeno una mutazione che è discendente di una qualsiasi mutazione del livello precedente al proprio.

Risultati dei nuovi raggruppamenti

Considerano quindi i nodi fino a profondità 3 (che casualmente nel grafo utilizzato per questo studio è anche la profondità massima) questa è la ripartizione per livelli del grafo:

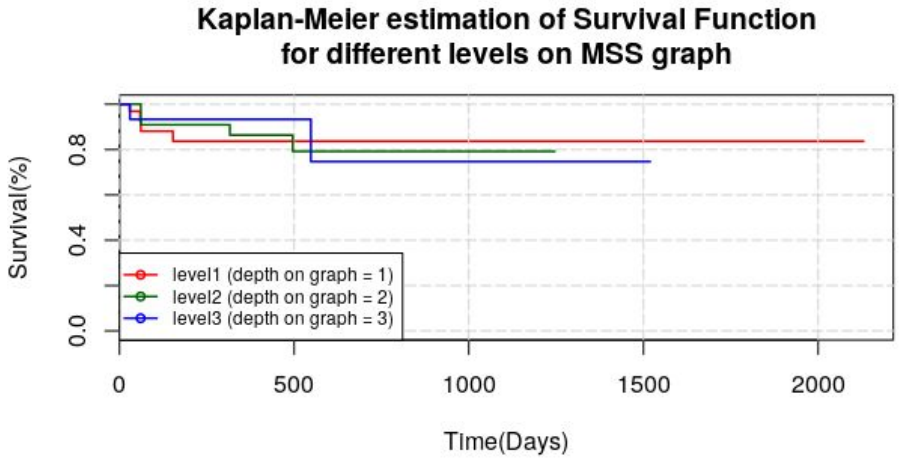


Si è quindi eseguita l'**analisi di sopravvivenza** per entrambi i nuovi raggruppamenti (sia per **levelGrouping** che per **pathLevelGrouping**) tramite diversi tentativi in modo da poter avere un confronto più attendibile.
I tentativi sono stati :

- **Kaplan-Meier** per ogni livello e **log-rank test** fra i vari modelli

  Esso è il metodo più semplice per confrontare gruppi di uno stesso campione e permette di poter valutare le differenze fra le diverse funzioni di sopravvivenza. Le funzioni risultanti dall'applicazione di questo modello sono effettivamente tra loro diverse, ma non in maniera rilevante.

  Esse sono le seguenti nel caso del raggruppamento **levelGrouping**:



Notando quindi una differenza nelle funzioni e avendo a disposizione sia l'intero campione che i diversi livelli si esegue facilmente un log-rank test fra le varie combinazioni di gruppi ottenendo i seguenti risultati :

- test su **tutti e 3 i livelli**:

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_levels_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=0  4        1    0.391    0.9479    1.0157
level=1 64        4    4.295    0.0203    0.0342
level=2 50        4    4.590    0.0759    0.1327
level=3 34        2    1.723    0.0444    0.0546

 Chisq= 1.1  on 3 degrees of freedom, p= 0.768
```

  Non riesce comunque a rifiutare l'ipotesi che le 3 curve seguano la stessa distribuzione, non potendo quindi definirle tra loro differenti.

- test fra i **livelli 1 e 2**

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_levels_1_2_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=1 64        4     3.87   0.00422   0.00838
level=2 50        4     4.13   0.00396   0.00838

 Chisq= 0  on 1 degrees of freedom, p= 0.927
```

  Non si può rifiutare l'ipotesi nulla di uguaglianza fra le distribuzioni di sopravvivenza, non potendo aggiungere quindi altro.

- test fra i **livelli 2 e 3**

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_levels_2_3_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=2 50        4     4.37    0.0313     0.121
level=3 34        2     1.63    0.0840     0.121

 Chisq= 0.1  on 1 degrees of freedom, p= 0.728
```

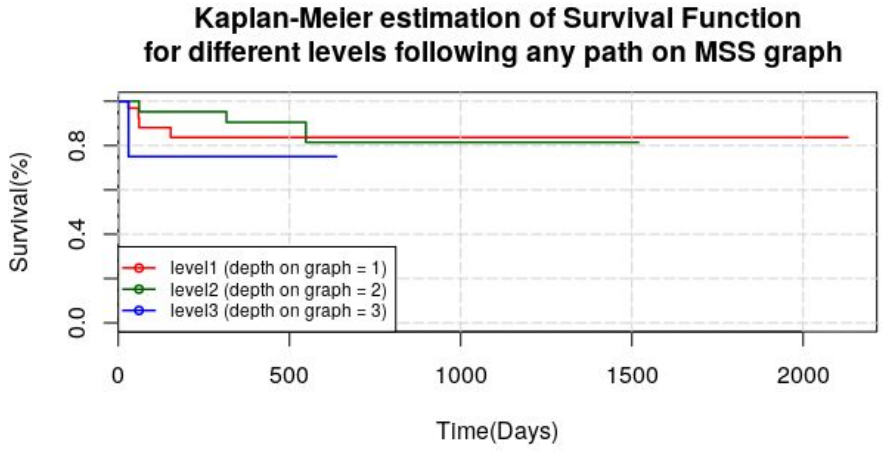
Anche in questo caso non si può rifiutare l'ipotesi nulla.

- test fra i **livelli 1 e 3**

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_levels_1_3_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=1 64        4     4.35    0.0276     0.104
level=3 34        2     1.65    0.0726     0.104

 Chisq= 0.1  on 1 degrees of freedom, p= 0.747
```

Nemmeno in quest'ultimo caso si può rifiutare l'ipotesi nulla.

Esse sono invece le funzioni risultanti ed i loro confronti nel caso del raggruppamento **pathLevelGrouping**:



Notando quindi una differenza nelle funzioni e avendo a disposizione sia l'intero campione che i diversi livelli si esegue facilmente un log-rank test fra le varie combinazioni di gruppi ottenendo i seguenti risultati :

- test su **tutti e 3 i livelli**:

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_pathlevels_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=0 29        3    1.478    1.5688    1.8456
level=1 64        4    4.295    0.0203    0.0342
level=2 48        3    4.909    0.7424    1.3819
level=3 11        1    0.318    1.4621    1.5320

 Chisq= 3.9  on 3 degrees of freedom, p= 0.273
```

  Non riesce a rifiutare l'ipotesi che le 3 curve seguano la stessa distribuzione, non potendo quindi definirle tra loro differenti, anche se dati i risultati si sarebbe comunque più propensi a considerarle tali.

- test fra i **livelli 1 e 2**

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_pathlevels_1_2_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=1 64        4     3.43    0.0946     0.188
level=2 48        3     3.57    0.0909     0.188

 Chisq= 0.2  on 1 degrees of freedom, p= 0.665
```

  Non si può rifiutare l'ipotesi nulla di uguaglianza fra le distribuzioni di sopravvivenza, non potendo aggiungere quindi ulteriori informazioni.

- test fra i **livelli 2 e 3**

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_pathlevels_2_3_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=2 48        3    3.691     0.129       1.7
level=3 11        1    0.309     1.545       1.7

 Chisq= 1.7  on 1 degrees of freedom, p= 0.192
```

  Anche in questo caso non si può rifiutare l'ipotesi nulla.

- test fra i **livelli 1 e 3**

```
Call:
survdiff(formula = SurvObj ~ level, data = MSSsurvival_pathlevels_1_3_dataset)

         N Observed Expected (O-E)^2/E (O-E)^2/V
level=1 64        4    4.638    0.0878      1.23
level=3 11        1    0.362    1.1248      1.23

 Chisq= 1.2  on 1 degrees of freedom, p= 0.267
```

  Nemmeno in quest'ultimo caso si può rifiutare l'ipotesi nulla, ma come nel caso più generico si sarebbe più propensi a considerare le distribuzioni di sopravvivenza dei livelli 1 e 3 come differenti piuttosto che come uguali.

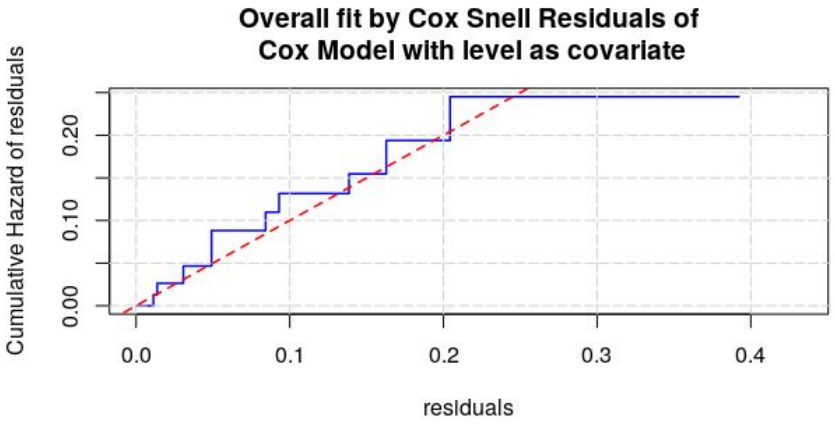
- **Cox model** con l'appartenenza ad un determinato livello come covariata

  Si è scelto di ottenere anche un semplice modello di Cox con un'unica covariata, ovvero l'appartenenza ad un determinato livello, per vedere se e quanto essa può influenzare la sopravvivenza.

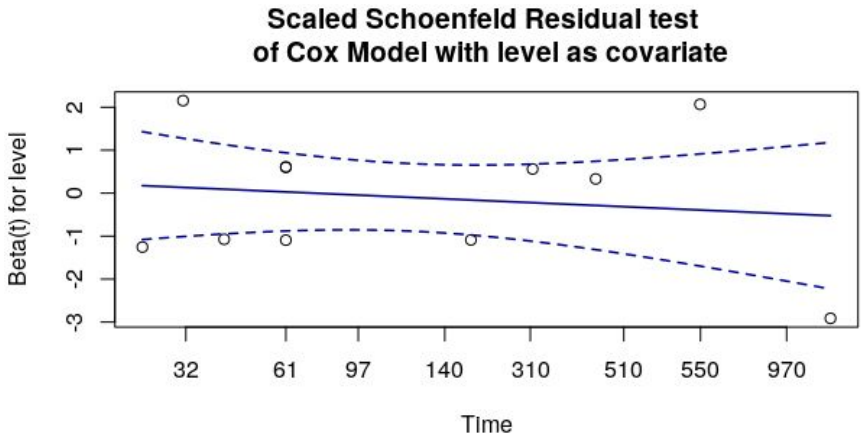
  Di seguito i risultati per **levelGrouping**:

```
Call:
coxph(formula = SurvObj ~ level, data = MSSsurvival_levels_dataset)

  n= 152, number of events= 11

         coef exp(coef) se(coef)      z Pr(>|z|)
level -0.0990    0.9057   0.3931 -0.252    0.801

      exp(coef) exp(-coef) lower .95 upper .95
level    0.9057      1.104    0.4192     1.957

Concordance= 0.497  (se = 0.091 )
Rsquare= 0   (max possible= 0.41 )
Likelihood ratio test= 0.06  on 1 df,   p=0.8009
Wald test            = 0.06  on 1 df,   p=0.8012
Score (logrank) test = 0.06  on 1 df,   p=0.8011
```

  Il parametro di influenza ottenuto stima un effetto positivo del livello sulla sopravvivenza. Questo risultato può essere considerato sensato se si nota che la numerosità dei livelli è tanto maggiore quanto è alto il livello, fatto che porta a leggere il parametro come "la sopravvivenza è tanto migliore quanto più è alto il livello".
  Questa considerazione è comunque forzata, dato anche il basso valore del parametro stimato.
  Per aumentare la validità dell'osservazione si osservano i risultati dei test sul modello.



  Il modello risulta abbastanza affidabile, mentre per quanto riguarda la correttezza dell'assunzione di proporzionalità sull'unica covariata del modello:

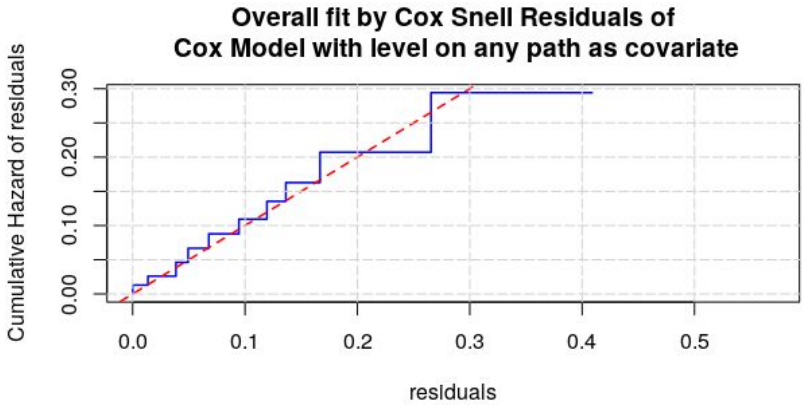


L'assunzione di proporzionalità per la covariata 'level' risulta corretta, anche se molti residui sono all'esterno dell'intervallo di confidenza, non garantendo la correttezza del risultato.

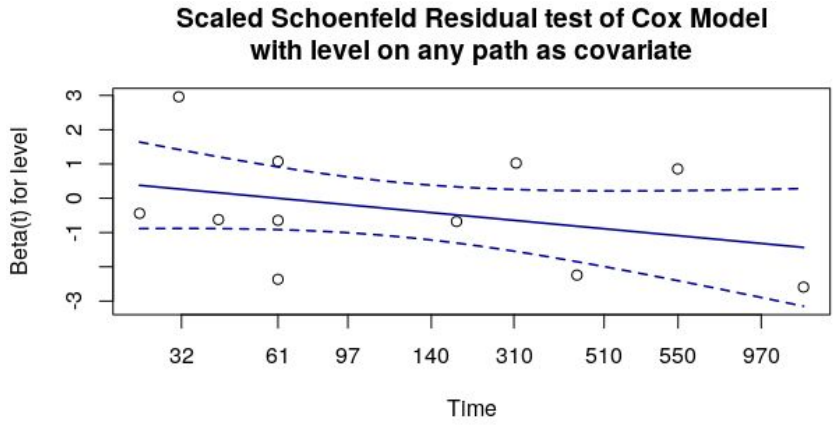
Di seguito i risultati per **pathLevelGrouping**:

```
Call:
coxph(formula = SurvObj ~ level, data = MSSsurvival_pathlevels_dataset)

  n= 152, number of events= 11

         coef exp(coef) se(coef)      z Pr(>|z|)
level -0.3334    0.7165   0.3953 -0.843    0.399

      exp(coef) exp(-coef) lower .95 upper .95
level    0.7165      1.396    0.3301     1.555

Concordance= 0.509  (se = 0.091 )
Rsquare= 0.005   (max possible= 0.41 )
Likelihood ratio test= 0.7  on 1 df,   p=0.4019
Wald test            = 0.71  on 1 df,   p=0.399
Score (logrank) test = 0.72  on 1 df,   p=0.397
```

Il modello ottenuto conferma le considerazioni del caso precedente, assegnando però un'influenza maggiore per l'appartenenza ad un determinato livello.
Prima di fornire osservazioni aggiuntive si osservano i risultati dei test sul modello :



Il modello risulta affidabile, mentre per quando riguarda l'assunzione di proporzionalità per la covariata 'level' :



La covariata non rispetta l'assunzione di proporzionalità rendendo di fatto inutile ogni ulteriore considerazione.

- **Cox model** con ogni livello come una covariate differente

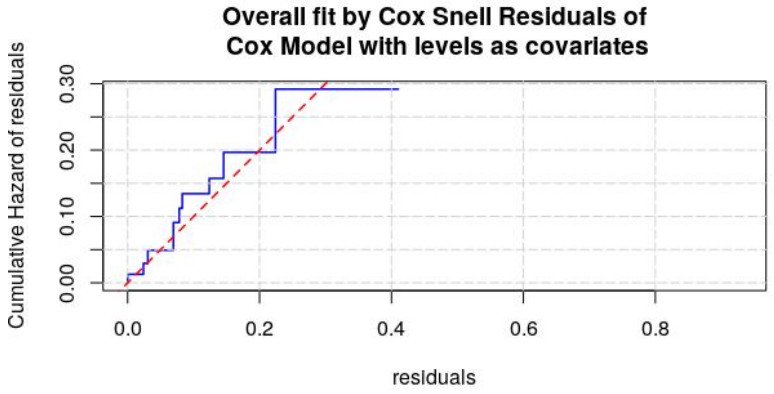
  Sì è ottenuto anche un modello di Cox dove ogni livello viene considerato come una differente covariata, per aumentare le informazioni a disposizione per eventuali considerazioni sull'influenza di un livello.

  Di seguito i risultati per **levelGrouping**:

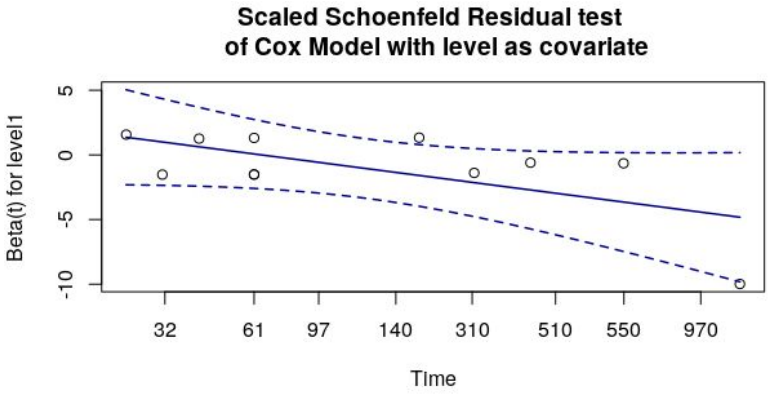
```
Call:
coxph(formula = SurvObj ~ level1 + level2 + level3, data = MSSsurvival_levels_dataset)

  n= 152, number of events= 11

          coef exp(coef) se(coef)      z Pr(>|z|)
level1 -1.0632    0.3454   1.1515 -0.923    0.356
level2 -1.1206    0.3261   1.1459 -0.978    0.328
level3 -0.8189    0.4409   1.2375 -0.662    0.508

       exp(coef) exp(-coef) lower .95 upper .95
level1    0.3454      2.896   0.03615     3.299
level2    0.3261      3.067   0.03451     3.081
level3    0.4409      2.268   0.03899     4.986

Concordance= 0.543  (se = 0.091 )
Rsquare= 0.005   (max possible= 0.41 )
Likelihood ratio test= 0.83  on 3 df,   p=0.8417
Wald test            = 1.03  on 3 df,   p=0.7929
Score (logrank) test = 1.12  on 3 df,   p=0.7729
```

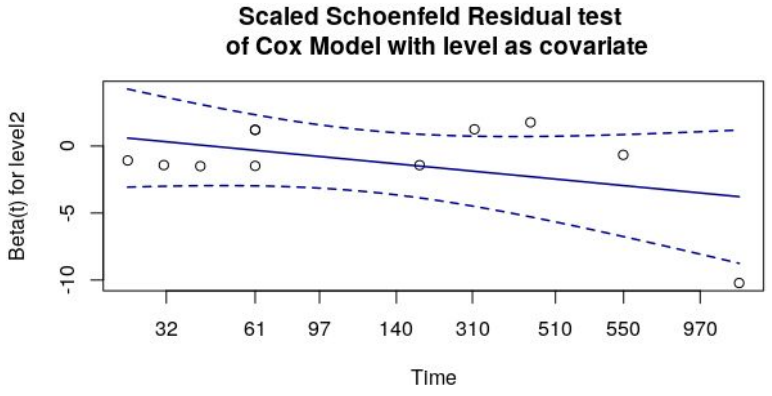
Dai valori dei parametri si nota che l'appartenenza ad un qualsiasi livello porta ad un'influenza positiva sulla sopravvivenza. Questo fatto è particolare, ma potrebbe essere dovuto al basso numero di morti registrate nel campione di partenza.

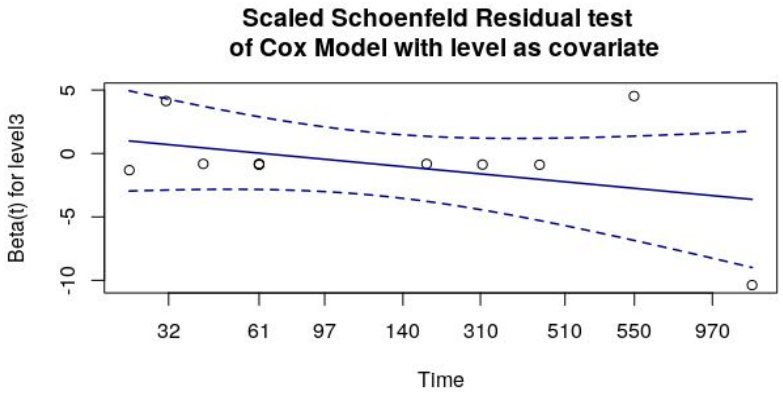
Si può comunque notare come sia minore l'influenza positiva del terzo livello (il più profondo sul grafo) rispetto a quelli superiori.
Prima di aggiungere ulteriori considerazioni si osservano i risultati dei test sul modello e sulle covariate.



Il modello risulta quindi affidabile, per quanto riguarda le covariate invece :





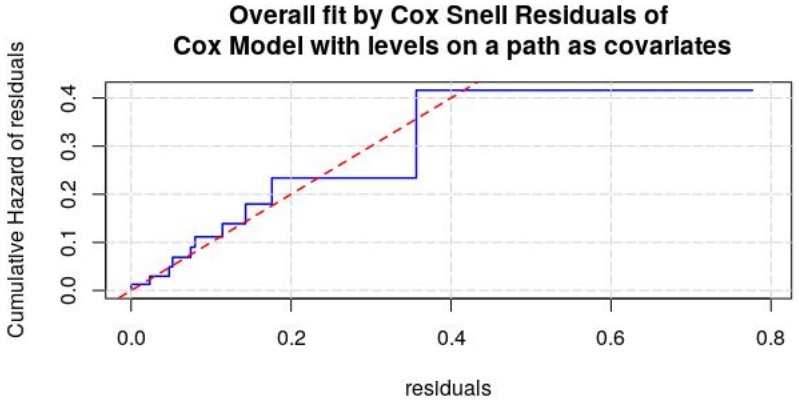


Nessuna delle covariate rispetta a pieno l'ipotesi di proporzionalità, rendendo ogni ulteriore considerazione inutile ed ogni osservazione precedente priva di fondamento.

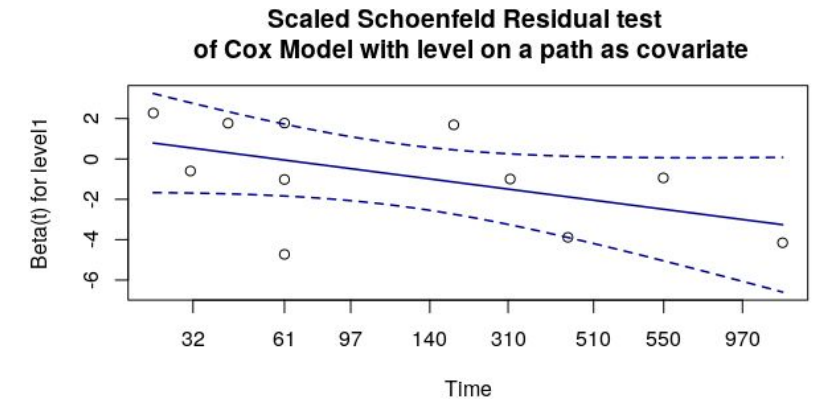
Di seguito i risultati per **pathLevelGrouping**:

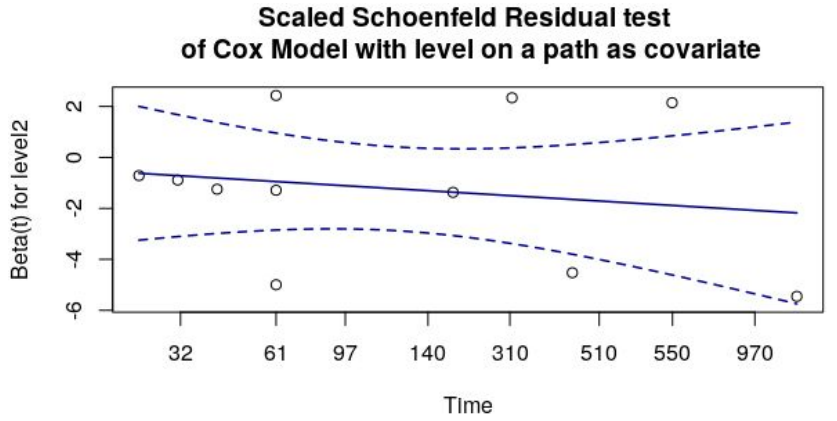
```
Call:
coxph(formula = SurvObj ~ level1 + level2 + level3, data = MSSsurvival_pathlevels_dataset)

  n= 152, number of events= 11

          coef exp(coef) se(coef)      z Pr(>|z|)
level1 -0.8004    0.4491   0.7691 -1.041    0.298
level2 -1.2329    0.2914   0.8218 -1.500    0.134
level3  0.4486    1.5661   1.1650  0.385    0.700

       exp(coef) exp(-coef) lower .95 upper .95
level1    0.4491     2.2265   0.09948     2.028
level2    0.2914     3.4313   0.05822     1.459
level3    1.5661     0.6385   0.15965    15.362

Concordance= 0.635  (se = 0.091 )
Rsquare= 0.02   (max possible= 0.41 )
Likelihood ratio test= 3.11  on 3 df,   p=0.3757
Wald test            = 3.45  on 3 df,   p=0.3278
Score (logrank) test = 3.89  on 3 df,   p=0.2735
```

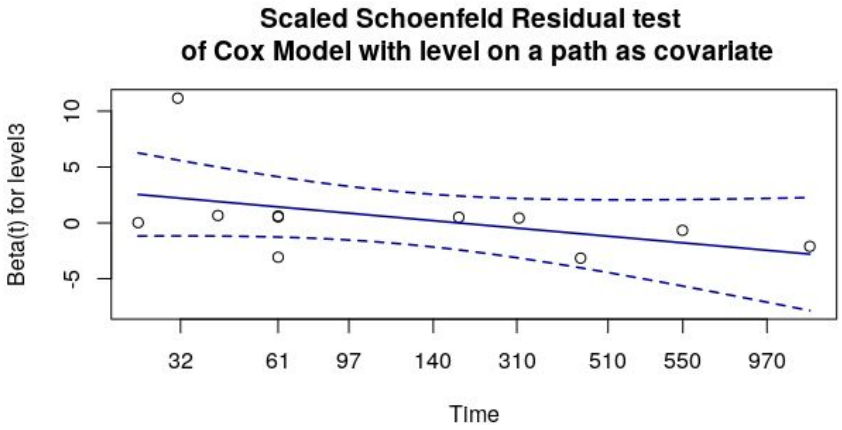
Dai parametri stimati si potrebbe osservare come la sopravvivenza è influenzata negativamente per i pazienti appartenenti al terzo livello, mentre lo è positivamente per l'appartenenza ai livelli superiori, anche se stranamente risulta più conveniente appartenere al secondo livello rispetto al primo (fatto probabilmente dovuto alla numerosità del campione iniziale).
Prima di ogni ulteriore considerazione osserviamo i risultati dei test sul modello:



Il modello risulta abbastanza affidabile, mentre per quanto riguarda le covariate :





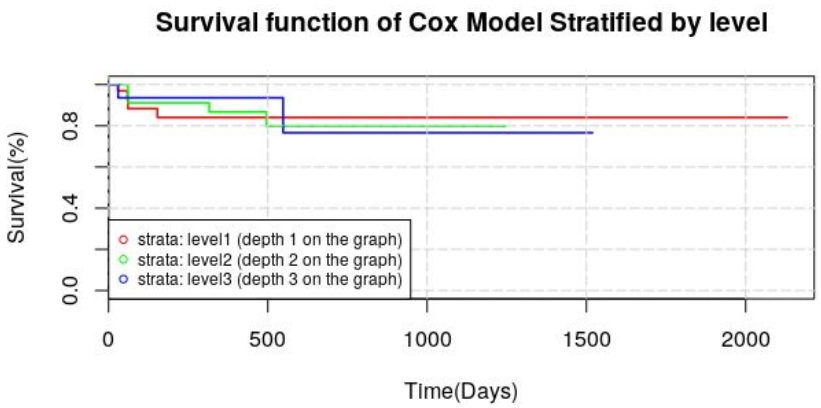


Nessuna delle covariate rispetta a pieno l'ipotesi di proporzionalità (potrebbe essere accettata per 'level2') non potendo confermare nessuna delle osservazioni precedenti e non permettendo di aggiungere nuove considerazioni.

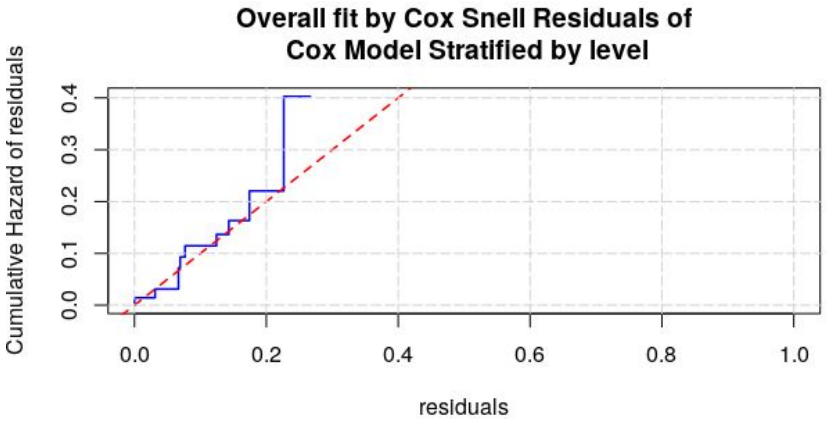
- **Cox model stratificato** per livello

  Si è ottenuto anche un semplice modello di cox senza l'influenza di nessuna covariate e stratificato per appartenenza ad un determinato livello.
  Da esso ci si aspettano gli stessi risultati ottenuti con il modello non parametrico risultante dalle stime di kaplan-meier.

  Di seguito i risultati per **levelGrouping**:

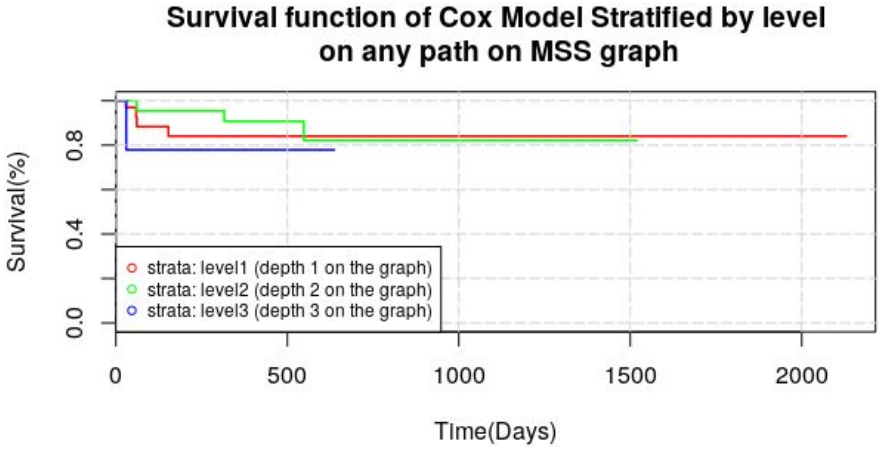


Si notano in effetti delle differenze fra le funzioni di sopravvivenza dei diversi livelli che concordano con quelle ottenute con l'utilizzo dello stimatore di kaplan-meier.
Si controlla l'affidabilità del modello eseguendo il test sull'*overall fit*, che produce il seguente risultato:



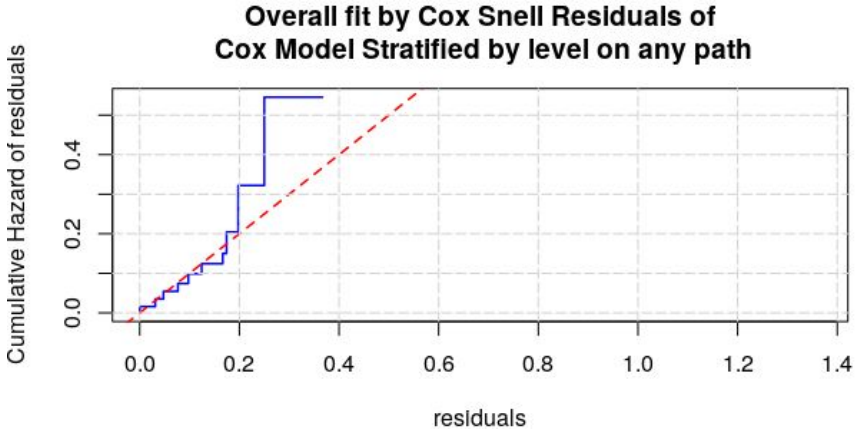
Il modello non risulta troppo affidabile, anche se non è nemmeno da considerare sbagliato.

Di seguito i risultati per **pathLevelGrouping**:



Anche in questo caso i risultati sono molto simili a quelli ottenuti non parametricamente.
La correttezza del modello viene testata come in ogni precedente caso tramite

cox snell residuals producendo il seguente risultato:



In questo caso ci si discosta troppo dalla linea di riferimento *x=y* rendendo il modello poco affidabile.

Conclusioni finali

L'analisi di sopravvivenza partiva da un campione decisamente troppo limitato, ma serviva utilizzarne uno per cui si avessero dei risultati provenienti dall'esperienza PiCnIc.
L'obiettivo era infatti un'ulteriore conferma dei grafi delle progressioni tumorali stimate dall'algoritmo di inferenza.

Per il target si è quindi deciso di provare a mostrare le possibili differenze:

- fra diversi cammini sul grafo, utilizzando dei **modelli** che confrontassero diverse porzioni del grafo,
- fra diversi livelli di profondità sul grafo, utilizzando modelli basati su **raggruppamenti** diversi a quelli del primo caso.

Per quanto riguarda il primo tentativo i risultati sono abbastanza deludenti.
Si possono infatti notare differenze di sopravvivenza fra diverse porzioni del grafo, ma la maggior parte dei modelli sono poco affidabili o nulli.
Inoltre non si è riuscito a mettere in evidenza l'influenza delle varie mutazioni all'interno dei gruppi, non potendo utilizzare i risultati per arricchire il grafo risultante da PiCnIc.
Questa seconda osservazione è quella che rende quasi inutilizzabili le informazioni ottenute, infatti l'obiettivo di questo tentativo era quello di poter aggiungere un valore percentuale che rappresentasse la sopravvivenza ai nodi del grafo.

Fallito il primo approccio si è provato quindi con un secondo tentativo, atto a evidenziare un altro aspetto utile alla conferma dei grafi, ovvero il peggioramento della sopravvivenza nel tempo scendendo di profondità in un determinato grafo.
I risultati in questo caso sono interessanti, ma non possono essere considerati affidabili.
Le funzioni di sopravvivenza differiscono infatti nei vari livelli, soprattutto peggiorano

all’aumentare della profondità relativa al livello, confermando parzialmente l’ipotesi iniziale (fatto confermato sia con un modello non parametrico, come kaplan-meier, che con uno semi-parametrico, come il modello di cox).
Ciò che rende la conferma solo parziale sono i test sulla differenza delle varie curve e i modelli di cox che stimano l’influenza dell’appartenenza in un determinato livello.
In entrambi i raggruppamenti non viene confermata infatti nessuna differenza di distribuzione di sopravvivenza; mentre per quanto riguarda l’influenza di un determinato livello stimata tramite un modello di cox i risultati rispettano l’ipotesi, ma i modelli risultano non statisticamente affidabili.
Non si ha quindi nessuna evidenza statistica che mostri la reale differenza di sopravvivenza fra diverse profondità del grafo.

Concludendo :

- si possono evidenziare differenze fra diverse porzioni del grafo, ma non abbastanza affidabili da poterlo arricchire con nuove informazioni riguardanti la sopravvivenza nel tempo.
- si possono evidenziare peggioramenti nella sopravvivenza dei pazienti corrispondenti a livelli di profondità maggiori nel grafo, ma non si hanno risultati statisticamente accettabili per poter confermare l’ipotesi.
- in entrambi i casi il problema principale è dovuto alla numerosità del campione troppo limitata a cui si aggiunge una bassissima presenza di eventi (morti) registrate, fattori importanti per un’analisi di sopravvivenza.
- i risultati sono quindi leggermente inferiori alle aspettative, ma non negano nessuna ipotesi
- ripetere l’esperienza con un campione con numerosità 8/10 volte maggiore permetterebbe di avere dei risultati considerabili affidabili.