Microti data analysis

Introduzione

Sulla base delle considerazioni fatte nella videocall di fine Giugno, ho identificato due variabili outcome di interesse:

- stato microbiologico dei linfonodi (variabile dicotomica Pos/Neg)
- grado istologico dei granulomi (variabile ordinale con quattro livelli: g1/g2/g3/g4)

Sulla base della natura della variabile outcome ho adattato un modello di regressione specifico con lo scopo di individuare quali variabili tra quelle misurate influenzano in modo rilevante l'outcome sia come direzione (positiva-negativa) che come entità. In questo senso le variabili assumo il significato di predittori, in grado quindi di predire il valore dell'outcome con un certo grado di errore.

Stato microbiologico

Per questa analisi l'unità statistica è costituita dal singolo linfonodo. Lo stato microbiologico dei linfonodi (Micro: Positivo/Negativo) è stato modellato utilizzando un modello bayesiano di regressione logistica utilizzando i seguenti predittori:

- Dimensione del linfonodo: Diametro calcolato a partire dell'area e quindi standardizzato in z-score
- NAF presenza nei granulomi del linfonodo di batteri acido resistenti. Questa variabili è stata ottenuta calcolando per ogni singolo linfonodo la media di batteri presenti nei deversi granulomi e quindi categorizzando il linfonodo stesso come NAF = 1 quando la media risultava superiore a 0 e NAF = 0 (cioè assenza di batteri acido resistenti) nel caso di medie pari a 0.
- MNC analogamente a quanto fatto per NAF è stato fatto per la categorizzazione dei linfonodi sulla base della presenza o meno di cellule nei granulomi.
- ngr numero totale di granulomi per linfonodo. Questa variabile è usata sia per calcolare le variabili sG1 ecc... che come tale dopo trasformazione logaritmica.
- sG1, sG2, sG3, sG4: per cogliere l'influenza che il grado istologico dei granulomi ha sullo stato microbiologico di un linfonodo ho costruito quattro nuove variabili una per ogni grado istologico che identificano per ogni singolo linfonodo la proporzioni di granulomi dei diversi gradi. Per ogni linfonodo ho calcolato il numero di granulomi dei diversi gradi istologici e poi l'ho divisa per il numero complessivo di granulomi per linfonodo , ottenendo quindi un valore compreso tra 0 e 1 che corrispondente alla proporzione di granulomi dei diversi gradi. Ad esempio nel linfonodo n.8 sono stati osservati 16 granulomi, di cui 0 di Grado 1 e 2 , 13 di grado 3 e 3 di grado 4 , quindi per il linfondo 8 avremo questo profilo sG1 = 0, sG2 = 0, sG3 = 13/16 (0.81), sG4 = 3/16 (0.19). A titolo di esempio riporto nella seguente tabella 5 linfonodi e il loro profilo relativamente allo stato microbiologico alla proporzione di granulomi dei diversi gradi istologici e il numero complessivo di granulomi.

```
## # A tibble: 5 x 7
##   Idlinf Micro   sG1    sG2   sG3   sG4   ngr
##    <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1     76     1     0 0.0526 0.737 0.211    19
## 2    598     0     0 0.485  0.364 0.152    33
```

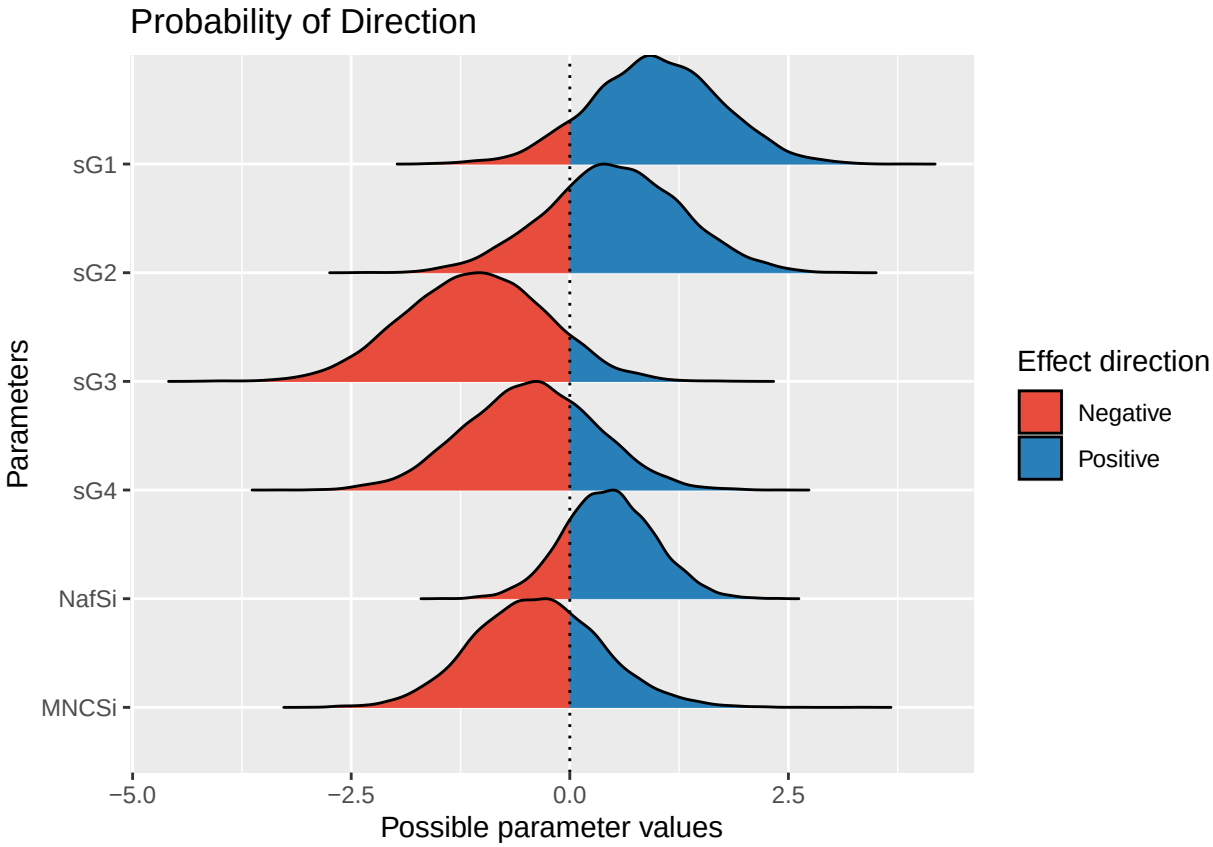
```
## 3    601     0     0 0.286  0.286 0.429     7
## 4    619     0     0 0      0.5   0.5       6
## 5    620     0     0 0      0.7   0.3      10
```

Tralascio qui i dettagli relativi alla costruzione del modello, che però metteremo nell'eventuale paper e/o tesi.

Risultati

In questo caso è stato adattato un modello bayesiano di regressione logistica. Il modello, per un singolo predittore , stima (con certo grado d'incertezza) un coefficiente che misura l'influenza che i cambiamenti del predittore hanno sullo stato microbiologico dei linfonodi (outcome), tenendo costante il valore degli altri predittori nel modello. I coefficienti assumono valori negativi se il cambiamento dei valori del predittore riduce la probabilità del linfonodo di essere positivo, al contrario valori positivi indicano un aumento della probabilità di essere positivo. Un coefficiente con valore = 0 indica che il predittore non influenza l'outcome. Quanto più il coeffiente è lontano da 0, in entrambe le direzioni, tanto maggiore è l'influenza del predittore . I coefficienti sono delle stime e quindi sono sempre accompagnati da una intervallo di valori che indica l'incertezza della stima, quanto più ampio è l'intervallo maggiore è l'incertezza. Un predittore è un predittore rilevante quanto più è lontano da 0 e con una bassa incertezza. In questi casi l'interpretazione del coefficiente non è in senso causale, ma in senso comparativo.

Il modello ha selezionato come predittori la proporzione dei granulomi dei diversi gradi istologici per linfonodo, la presenza di NAF e di MNC. I risultati del modello sono riportati nel seguente grafico. Per ogni predittore viene riportata la distribuzione della stima del coefficiente di regressione e viene indicata con diversi colori l'area corrispondente ai valori con effetto negativo e a quelli con effetto positivo. Considerando che l'area complessiva della distribuzione dei valori delle stime dei coefficienti è uguale a 1 (100%), è possibile calcolare l'area corrispondente ai valori positivi e negativi. In questo modo si ottiene una misura dell' importanza dell'effetto stimato dal modello chiamata P_direction (PD).



Di seguito in tabella sono riassunti per tutti i predittori i coefficenti stimati, l'errore standard associato e il valore di PD

```
##           Estimate Est.Error   PD.
## Intercept    -1.42      0.97 93.25
## sG1           0.98      0.74 90.21
## sG2           0.50      0.77 74.55
## sG3          -1.09      0.80 91.34
## sG4          -0.45      0.74 72.25
## NafSi         0.44      0.51 80.79
## MNCSi        -0.37      0.73 70.24
```

In sostanza all'aumentare della proporzione di granulomi di grado 1 per linfonodo c'è una probabilità del 90% che aumenti la probabilità che il linfondo sia microbiologicamente positivo (tenendo costante il valore di tutte le altre variabili). Analogo effetto c'è per la variabile sG2 anche se con una probabilità del 74%. All'aumentare della proporzione di granulomi di grado 3 e 4 tendenzialmente c'è una riduzione della probabilità che il linfonodo risulti microbiologicamente postivo con una probibilità rispettivamente del 91% e 72%. In linfondi in cui sono presenti granulomi con batteri acido resistenti c'è una maggior probabilità che siano positivi microbiologicamente nell'81% dei casi. Al contrario nel 70% dei casi la presenza di MNC in granulomi riduce la probabilità dei linfondi di essere microbiologicamente positivi.

La rilevanza dei predittori in termini di lontanza da 0 non è però particolarmente elevata, e anche l'incertezza delle stime appare piuttosto ampia. Probabilmente l'aggregazione delle informazioni utilizzata per poter analizzare i dati a livello linfonodo non è particolarmente efficace nel cogliere i sottostanti effetti. Da questo punto di vista presenterei i risultati di questo modello come secondari commentando la suggestione tendenziale che suggerisce soprattutto se al di là degli aspetti statistici gli effetti colti sono biologicamente sensati.

Grado istologico

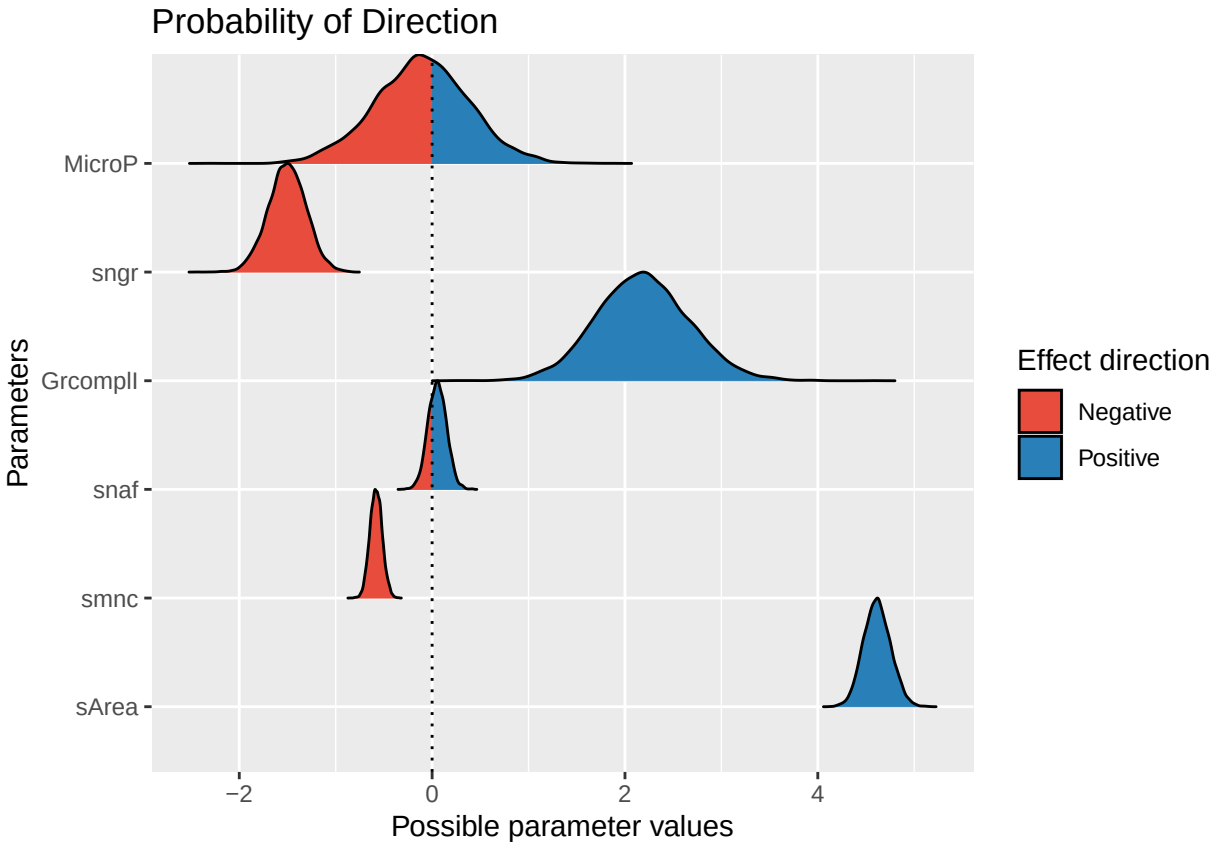
In questo caso è stato adattato un modello di regressione logistica ordinale di tipo sequenziale. Questo tipo di modello assume che le differenti classi della variabile outcome (in questo caso il grado istologico), siano espressione di un processo sequenziale per cui ad esempio un granuloma è classificato di grado 4, dopo essersi "evoluto" dai gradi precedenti. I modelli di regressione ordinale sono di diverso tipo, ho scelto questo perchè l'assunto del modello rispetto alla variabile outcome è coerente con lo sviluppo istologico dei granulomi (come confermato da Claudio. . . ..).

Il modello è ulteriormente complicato dalla struttura gerarchica dei dati, per cui le osservazioni effettuate a livello di granuloma sono nidificate sotto i linfonodi.

Quindi il grado istologico (outcome) è stato adattato utilizzando le seguenti variabili come predittori di interesse:

- lnGrArea : log dell'area del granuloma
- lnaf : log del numero di batteri acido resistenti per granuloma
- lmnc : log del numero di MNC per granuloma
- micro : stato microbiologico del linfondo (Pos / Neg)
- Grcompl : stato del granuloma completo / incompleto
- lngr : numero di granulomi per linfondo
- IdLinf : identificativo linfondo (utilizzata come variabile "random" nel modello)

I risultati del modello sono riportati nel seguente grafico.



In tabella sono riassunti per tutti i predittori i coefficenti stimati, l'errore standard associato e i valori di PD

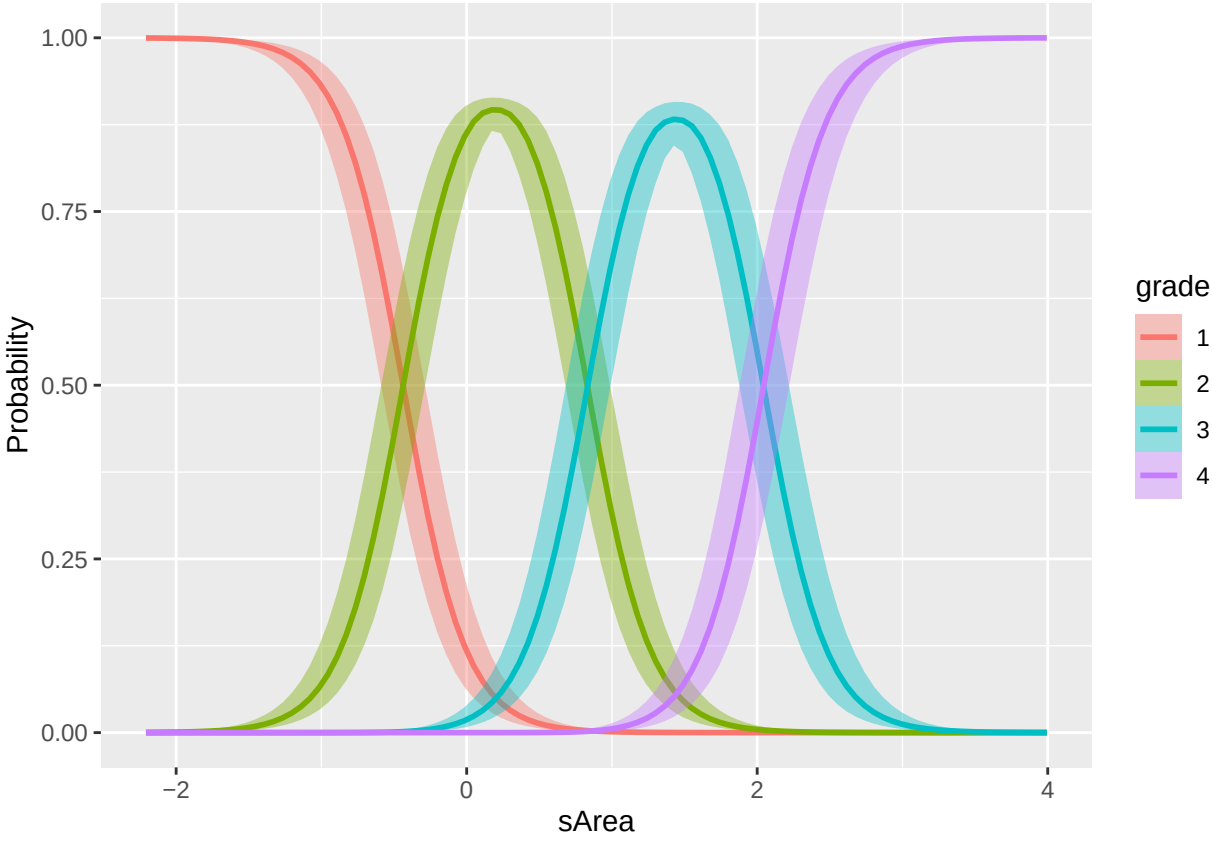
```
##              Estimate Est.Error    PD.
```

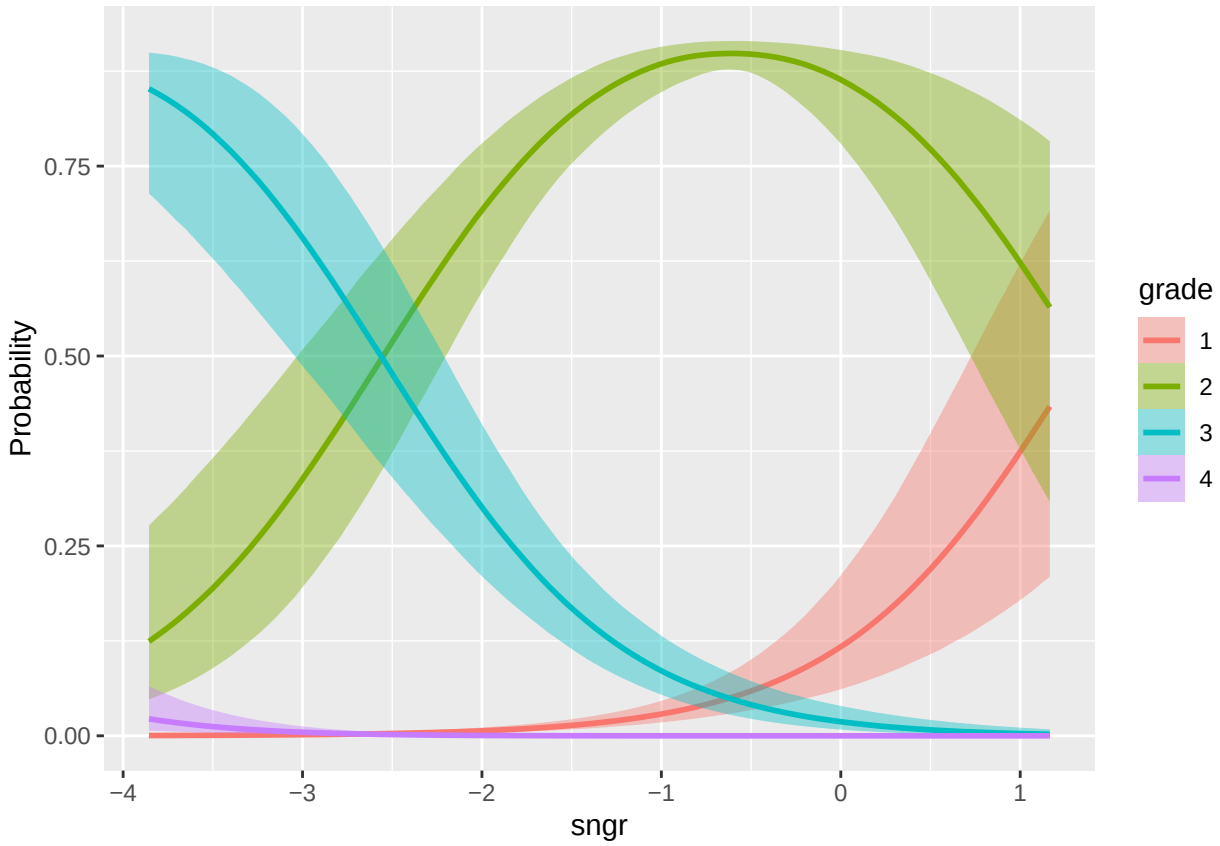
```
## Intercept[1]    -2.02      0.36 100.00
## Intercept[2]     3.84      0.36 100.00
## Intercept[3]     9.44      0.45 100.00
## sArea            4.62      0.14 100.00
## GrcomplI         2.18      0.48 100.00
## snaf             0.05      0.10  70.76
## smnc            -0.58      0.06 100.00
## MicroP          -0.11      0.48  58.86
## sngr            -1.50      0.20 100.00
```

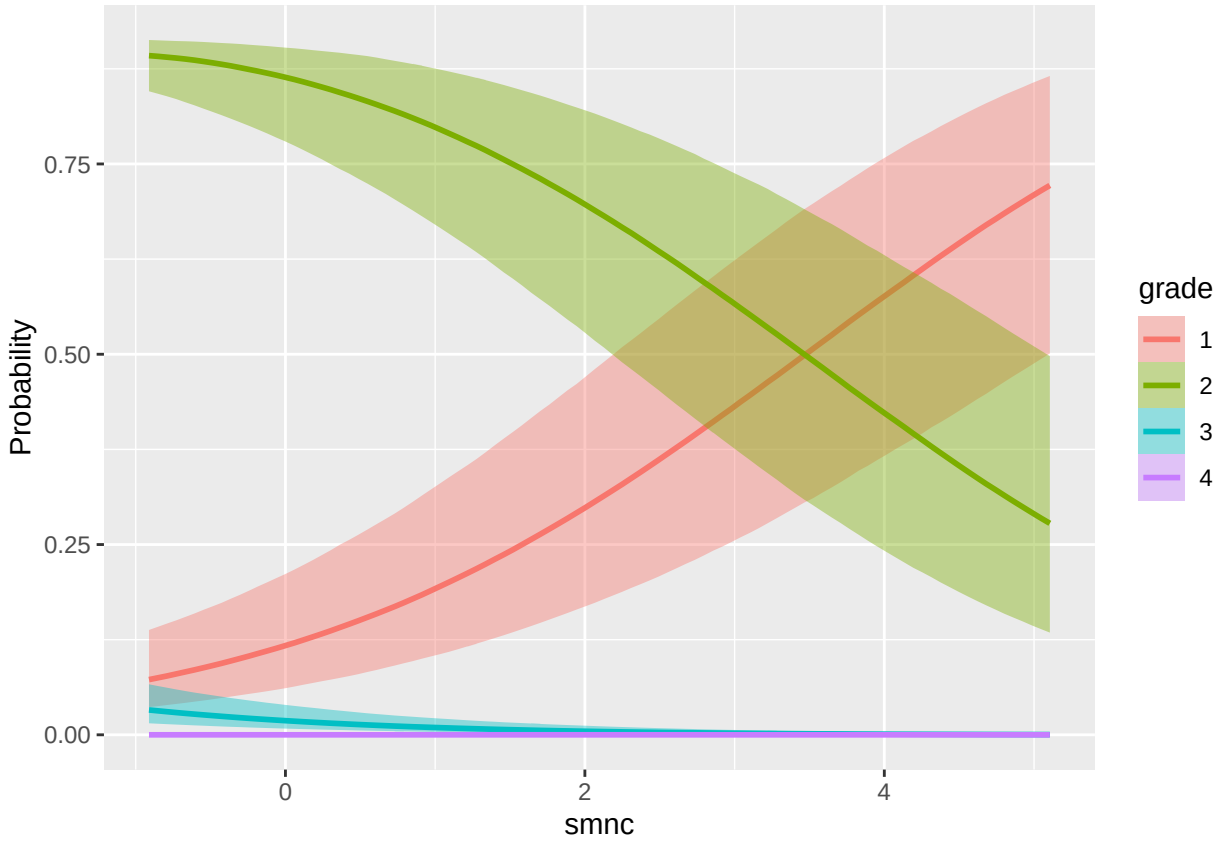
Il modello suggerisce che all'aumentare della dimensione dei granulomi aumenta in modo rilevante (il coefficiente di regressione è molto lontano da 0 !) la probabilità che il granuloma sia classificato nei gradi più alti della scala istologica. Al contrario all'aumentare del nunero di granulomi nel linfondo si riduce la probabilità che i granulomi di quel linfonodo siano di grado istologico elevato. Lo stato microbiologico non sembra essere rilevante, mentre sempre secondo il modello all'aumentare del numero di MNC si riduce la probabilità dei granulomi di essere classificato in gradi elevati. Per quello che riguarda il numero di batteri acido-resistenti il modello suggerisce che ci sia un effetto positivo, cioè all'aumentare del numero di batteri aumenta la probabilità che il granuloma sia di grado istologico alto ma questo risultato è supportato solo con una probabilità del 70%. Infine i granulomi incompleti rispetto a quelli completi hanno una maggior probabilità di essere classificati in gradi istologici elevati.

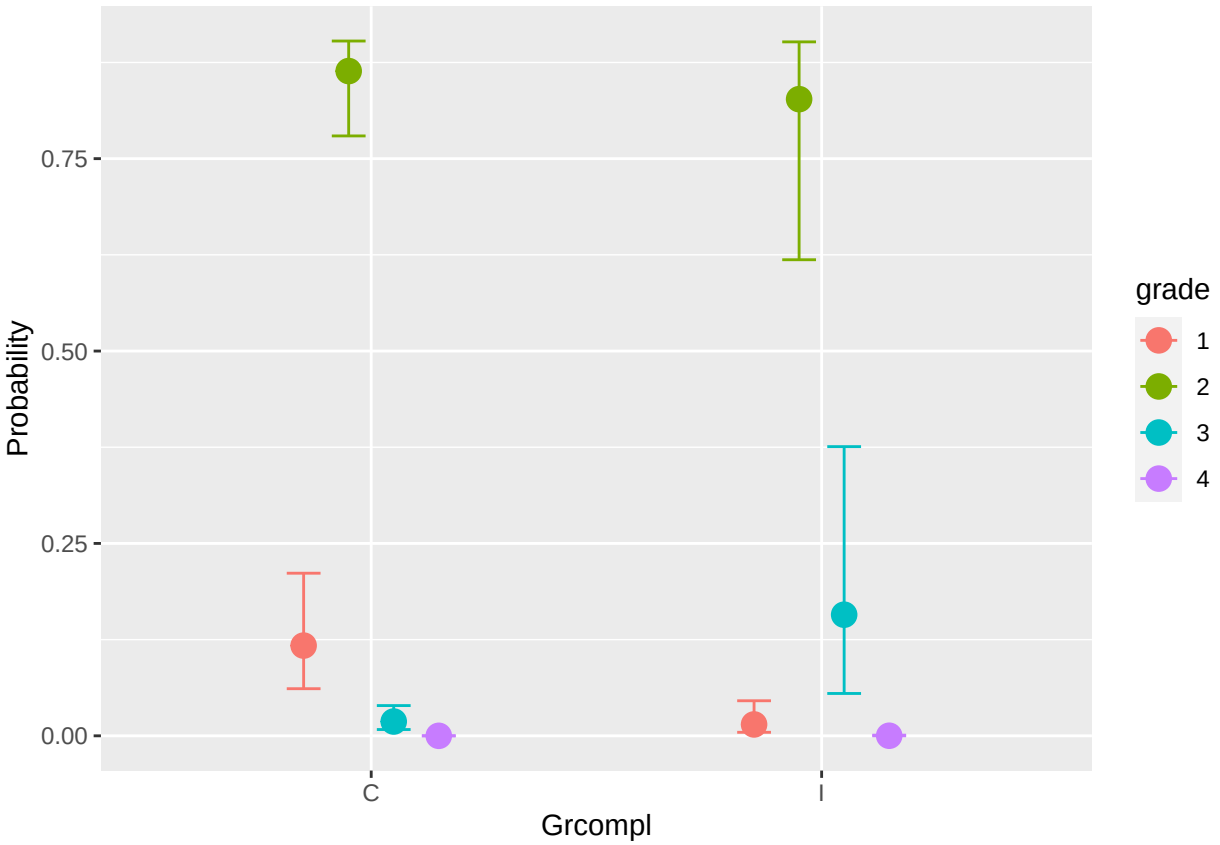
In questo tipo di modelli risulta più efficace visualizzare, almeno per i predittori più rilevanti gli effetti marginali, cioè tenendo costanti gli altri valori dei predittori osservare come varia al variare del predittore d'interesse la probabilità di un granuloma di essere classificato in uno dei quattro gradi istologici.

I grafici che seguono mostrano gli effetti dei predittori: dimensione, numero di granulomi, completezza, numero di mnc.




Probability
0.75
0.50
0.25
0.00
−4
−3
−2
−1
0
1
sngr
grade
1
2
3
4

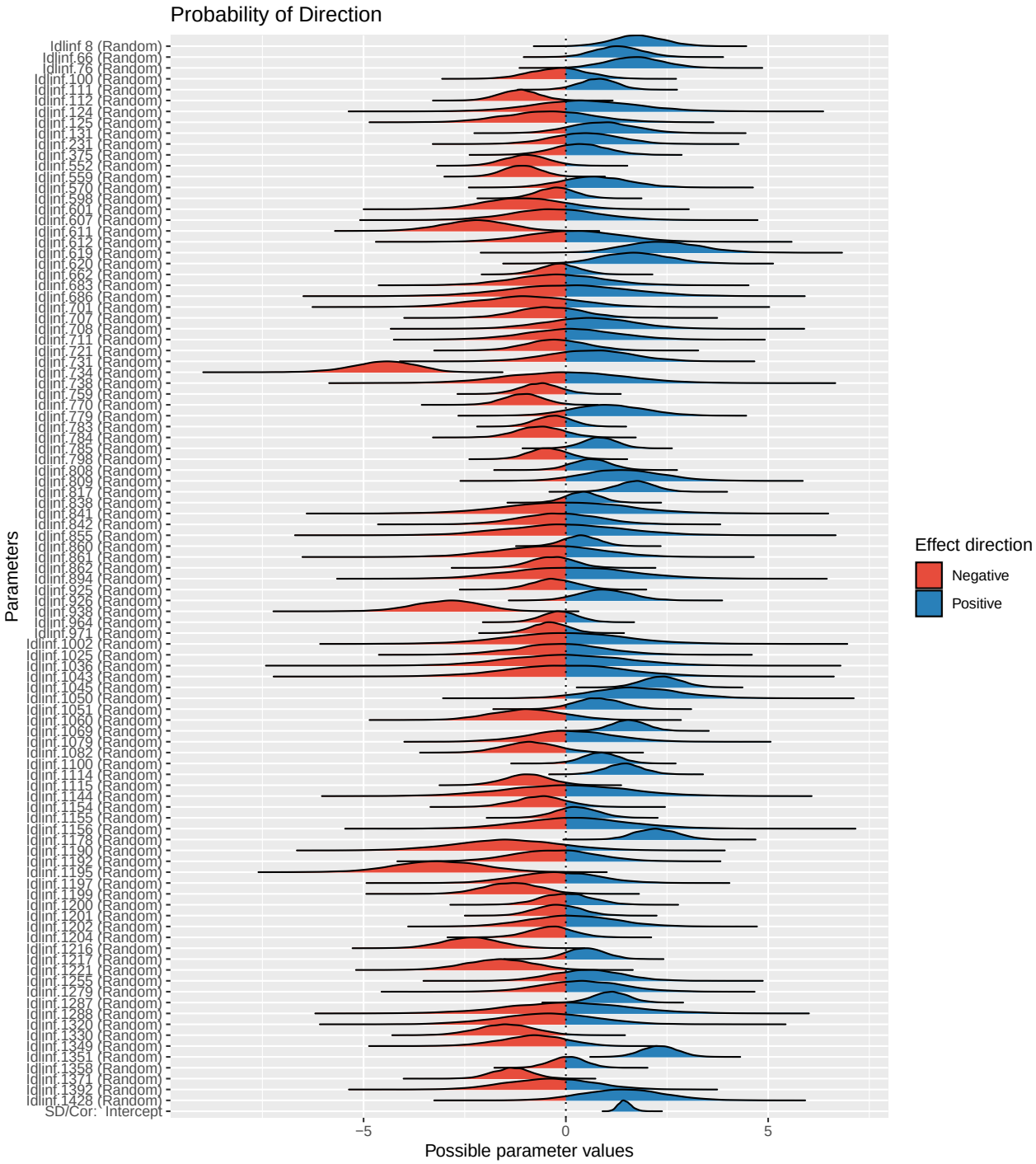

Probability
0.75
0.50
0.25
0.00
0
2
4
smnc
grade
1
2
3
4



Per l'interpretazione dei grafici che si riferiscono alle variabili su scala continua, dovete considerare che per necessità di costruzione del modello , tutte le variabili vengono standardizzate e quindi i valori sono scalati rispetto alla media e alla deviazione standard. Quindi il valore zero corrisponde alla media e i valori + o - sono le deviazioni standard dalla media... Ma al di là di questa complicazione è sufficiente osservare la direzione del grafico verso destra aumenta il valore della variabile verso sinistra diminuisce... se prendete il primo grafico quello che riporta l'effetto della dimensione dei granulomi vedete 4 curve ognuna fa riferimento a un grado istologico come riportato nella legenda a fianco. Il modello ragiona in questi termini, assegna al singolo granuloma sulla base dei valori dell'area e tenendo costante tutte le altre variabili 4 valori di probabilità per grado istologico. Quindi se provate a tirare idealmente una linea corrispoendente a -1 cioè area di dimensioni inferiori di una deviazione standard dalla media dei granulomi vedete che il modello assegna 0 probabilità ai gradi 3 e 4, una bassa probabilità al grado due e infine una probabilità elevata al grado 1. Se fate la stessa cosa con il valore di area uguale a +2 deviazioni standard noterete che il modello attribuisce una probabilità pari a 0 e poco più per i gradi 1 e 2 mentre attribuisce una probabilitò del 50% sia al grado 3 che al grado 4. La stessa modalità di lettura si applica anche alle altre variabili. L'area risulta un predittore molto efficace infatti questo grafico mostra una netta separazione tra i diversi gradi istologici.

Effetti random

Infine il grafico seguente riporta per tutti i linfondodi la variabilità del grado istologico sotto forma di stima. La linea verticale corrispondente a 0 indica il valore medio della variabile continua non osservabile direttamente del grado istologico. I linfonodi che sono spostati a destra di 0 cioè che hanno mediamente valori positivi, sono linfonodi che hanno mediamente un profilo di granulomi con grado istologico elevato, mentre quelli che si collocano tendenzialmente a sinistra sono linfonodi con profilo di basso grado istologico. Quello che suggerisce in particolar modo questo grafico è l'ampia variabilità dei linfonodi dal punto di vista del grado istologico dei granulomi in essi presenti...



Ad esempio il linfonodo identificato come IdLinf8 ha 16 grnulomi di cui 3 di grado 4 e 13 di grado 3; nel grafico sopra il linfondo 8 si colloca a destra del valore medio indicato dalla linea verticale 0. Al contrario il linfonodo 734 è costitutito da 237 granulomi tutti di grado 1, e infatti nel grafico è rappresentato da una curva a sinistra della linea verticale.