Università degli studi di Milano-Bicocca

High Dimensional Data Analysis

Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes

Autori:
Lorenzo Camaione - 850380 - l.camaione@campus.unimib.it
Martino Lorenzo Tenconi - 803154 - m.tenconi3@campus.unimib.it
Lorenzo Famiglini - 838675- l.famiglini@campus.unimib.it

Abstract

Il progetto si pone l'obiettivo di replicare le analisi affrontate nel paper [1] dove lo scopo è quello di verificare l'associazione tra nutrienti e batteri della flora intestinale. Oltre a tale studio, sono stati sviluppati dei modelli in grado di prevedere il bmi in funzione di batteri del microbioma, nutrienti ed altri fattori demografici quali età e sesso.

1 Introduzione

La dieta influenza significativamente la salute umana ed in parte anche il microbioma dell'intestino. In questo studio, da un campione di 97 persone, sono stati analizzati dati sulla dieta e sulle sequenze 16S rDNA estratte dalle feci degli individui. I batteri presenti nelle feci possono essere clusterizzati in enterotipi, che differiscono tra loro per i livelli di Bacteroides e Prevotella. Inoltre, la composizione degli enterotipi è dipendente dalla dieta di lungo periodo. In particolare, alti livelli di Bacteroides possono essere associati a diete ricche di proteine e grasso animale mentre alti livelli di Prevotella a carboidrati. Una volta studiate tali relazioni, sono stati sviluppati dei modelli che prevedono il livello di BMI di una persona in funzione di: nutrienti, microbioma e fattori demografici.

2 Datasets

I dataset a disposizione per le analisi sono forniti dal portale QIITA (study ID 1011) e sono composti da una tabella OTU, una tavola tassonomica e un questionario FFQ relativo ai nutrienti assunti dai soggetti nella dieta a lungo termine. La tabella OTU, di dimensioni 3393 x 100, contiene il numero di sequenze che sono state osservate per ognuno dei 3393 batteri presenti nel microbioma. La tavola tassonomica permette di collegare gli identificativi dei batteri presenti nella tabella OTU alle loro tassonomie. Ad essa, sono indicati, per ognuno dei 3393 batteri, i sette livelli tassonomici: regno, phylum, classe, ordine, famiglia, genere e specie. Infine, il questionario dei nutrienti contiene, per 97 soggetti, le quantità di ciascuno dei nutrienti assunti in una dieta di lungo periodo. In aggiunta a questi tre dataset, vi è una quarta tabella contenente gli id degli individui e una serie di variabili, come ad esempio il genere, l'altezza, il peso, l'età e il Body Mass Index (BMI) a loro associate. Le informazioni estratte da questo ultimo dataset, combinate con le variabili elaborate nella fase di preprocessing, saranno utilizzate per scopi di analisi predittiva del BMI.

2.1 Preprocessing

Si è effettuata una prima pulizia della tavola tassonomica rimuovendo caratteri speciali dai nomi delle tassonomie e si sono analizzate quindi le osservazioni per ciascuna colonna indicante il livello tassonomico. Poiché lo studio[1] è improntato sull'analisi dei batteri a livello di genere, ci si concentra sulle variabili indicanti i nomi delle tassonomie fino al sesto livello, ignorando la specie dei batteri osservati. La tavola tassonomica presenta un totale di 1526 batteri con genere non identificato. Perciò, ai fini dell'analisi, si considerano, per questi batteri, i soli livelli precedenti: ad esempio, per un batterio che ha classificazione fino al quarto livello ("ordine"), si considerano unicamente le prime quattro tassonomie, eliminando quindi i valori mancanti per i successivi livelli. A questo punto, vengono raggruppati tra loro i batteri nella tabella OTU che hanno la stessa classificazione fino all'ultimo livello presente. Si ottengono quindi cinque dataset, ognuno per ogni livello tassonomico da phylum a genere, contenenti le possibili aggregazioni osservate per la relativa profondità. Per ognuna di queste cinque tabelle, sono calcolate le abbondanze relative dei batteri, ovvero la presenza in termini percentuali di ciascun batterio in ognuno degli individui considerati. Unendo queste cinque tabelle, si ottengono 198 possibili

livelli di aggregazione dei batteri osservati. Sono state considerate, in seguito, solamente le tassonomie presenti in almeno il 10% dei soggetti che avessero un valore di abbondanza relativa maggiore o uguale allo 0.2% in almeno uno dei soggetti. In questo modo, sono state ottenute 76 tassonomie. Inoltre, per scopi successivi nelle analisi, si decide di creare una tabella OTU contenente i batteri aggregati al solo livello di genere o superiore ove mancante (famiglia, genere, classe, ...), ottenendo così 119 tassonomie differenti. Si mostrano le abbondanze nei 97 pazienti dei generi di batteri osservati:

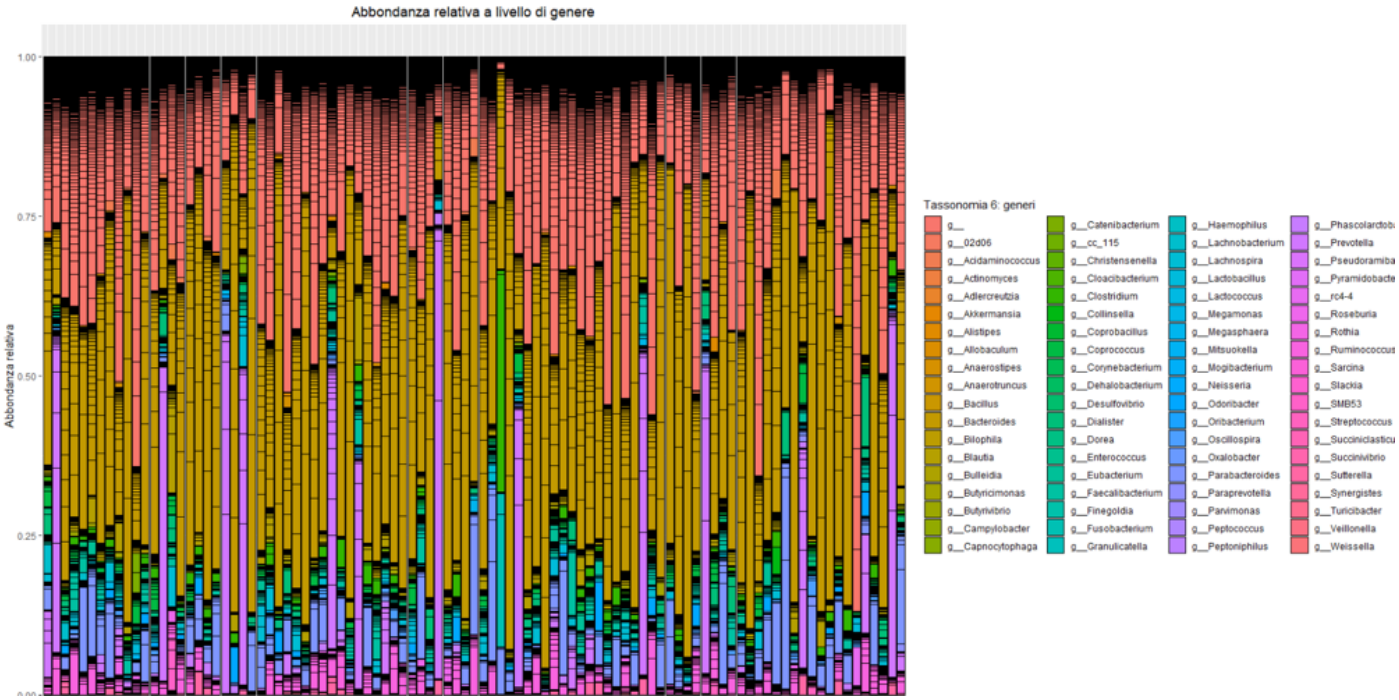


Figura 1. Tavola abbondanza per generi

Per quanto riguarda il dataset dei nutrienti, di dimensione 98 x 215, sono state selezionate soltanto le 97 righe con identificativo del soggetto nello studio presente anche nella tabella OTU citata in precedenza. Per valutare quali nutrienti fossero maggiormente significativi, a livello di impatto sulla composizione intestinale, sono stati considerati 3393 batteri, non aggregati per livello tassonomico, come indicatore di variabilità della composizione del microbioma stesso. Sulle abbondanze della tabella OTU originaria, sono state dunque calcolate le distanze tra i microbiomi intestinali dei 97 soggetti con la metrica Unifrac [3] non pesata (tiene conto solo della presenza/assenza del batterio). Essa risulta essere un metodo maggiormente discriminante per le differenze di composizione dei microbiomi rispetto alla stessa metrica Unifrac pesata, che prende in considerazione la quantità di ciascun batterio nel microbioma. Il calcolo della distanza, utilizzando la Unifrac non pesata, avviene tramite la formula:

$$\frac{sum\ of\ unshared\ branch\ lengths}{sum\ of\ all\ tree\ branch\ lengths} = fraction\ of\ total\ unshared\ branch\ lengths$$

Una volta calcolata la matrice delle distanze non pesate Unifrac, sui 97 microbiomi intestinali, è stato possibile effettuare la PERMANOVA[2] su ciascuno dei nutrienti standardizzati della tabella. A differenza della ANOVA, che si basa su assunzioni parametriche, la PERMANOVA effettua delle permutazioni casuali, nei regressori, senza dover verificare ipotesi alla base dei dati. Ognuno dei nutrienti è stato valutato significativo se il valore del False Discovery Rate FDR (percentuale di volte in cui si rifiuta l'ipotesi nulla quando invece è vera) è minore del 25%. Come descritto in [1], è stata usata una soglia, relativamente alta, per non escludere dall'analisi nutrienti associati a batteri con valori di abbondanza relativa non particolarmente elevati. I batteri e i nutrienti selezionati sono stati utilizzati

per costruire due dataset con lo scopo di effettuare l'analisi predittiva del BMI: dal momento che il numero di regressori p è superiore al numero delle osservazioni n, al fine di poter sviluppare dei modelli di regressione multivariata lineare con stimatore OLS (oltre al Lasso), sono state estratte, dai nutrienti e dai batteri, le prime componenti principali, rispettivamente per un totale di varianza catturata del 70% e dell'80%. A seconda dei modelli utilizzati,sono state prese in considerazione diverse tipologie di aggregazione per i batteri. Nel primo dataset sono presenti 119 tassonomie e nell'altro 35 generi di batteri. Una volta uniti alle le variabili dei nutrienti e demografiche sono stati suddivisi in training e test: la parte dedicata all'addestramento dei modelli è risultata essere il 90% (usata per la cross validation), lasciando il 10% dei dati per la validazione dei modelli.

3 Analisi descrittiva

Nella figura S1 (in appendice), si mostrano le correlazioni di Spearman ottenute tra i generi estratti dalle 35 tassonomie rilevanti e i nutrienti selezionati con la PERMANOVA. Dalla heatmap delle correlazioni (fig. S1) è evidente che gli amminoacidi come arginina, alanina, glicina e leucina siano correlati positivamente con il genere Bacteroides. Invece, risulta debole e negativa la correlazione di tali nutrienti con i due generi Prevotella osservati nei microbiomi dei soggetti in studio. D'altra parte, Prevotella mostra un'alta correlazione positiva con nutrienti ricchi di zuccheri, come saccarosio, fruttosio, carboidrati e indice glicemico. La quantità relativa di Bacteroides nel microbioma intestinale è invece poco correlata con i carboidrati. Per quanto riguarda le fibre e i composti d'origine vegetale, non si mostra una forte correlazione per nessuno dei due generi presi in esame. Il Ruminococcus, preso in considerazione in [1], risulta invece correlato positivamente con quest'ultima categoria di nutrienti. Tuttavia, la sua correlazione con i carboidrati e gli amminoacidi rimane debole e negativa.
Successivamente, si è utilizzata l'aggregazione dei batteri nelle 119 tassonomie a livello di genere per effettuare un raggruppamento delle composizioni dei microbiomi in cluster. Come suggerito da [1], si è effettuato un clustering PAM (Partitioning Around Medoid). Sono state provate tre differenti metriche di distanza: la distanza euclidea, la dissimilarità di Bray-Curtis, fortemente utilizzata in ambito biologico, ed infine la Jensen-Shannon divergence JSD [4]. In tutti e tre i casi il numero ottimale di cluster da considerare è stato 2, a cui sono associati i coefficienti di Silhouette maggiori.

Metrica	k ottimale	Silhouette
Dissimilarità di Bray-Curtis	2	0.41
Distanza Euclidea	2	0.51
Divergenza di Jensen-Shannon	2	0.57

Tabella 1. Risultati indice Silhoutte

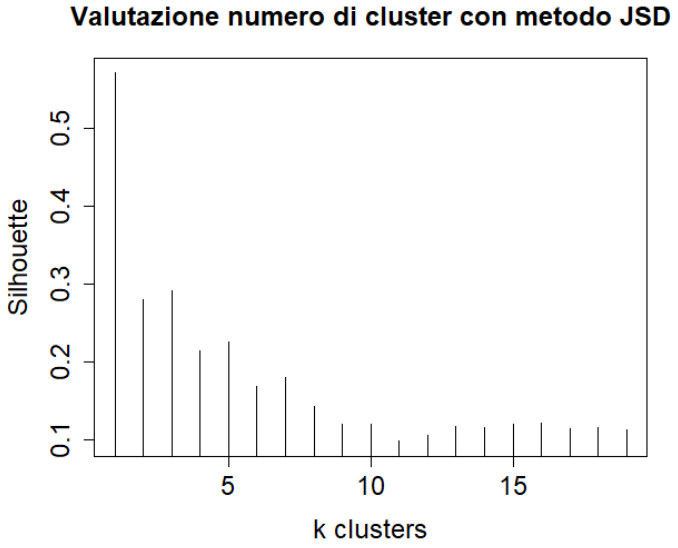


Figura 2. Indice di Silhouette in base al numero dei cluster

Il clustering calcolato mediante la matrice delle distanze con metrica JSD, è stato selezionato perchè

presenta il maggiore coefficiente di Silhouette: tale scelta massimizza la varianza tra gruppi e la coesione all'interno di essi rispetto alle altre metriche (figura 2). Sulla stessa matrice delle distanze JSD, sono state estratte le prime due componenti principali utilizzando la tecnica PCoA che, a differenza della PCA, accetta in input una matrice delle distanze invece che la matrice dei dati. In figura 3, si possono vedere le osservazioni rappresentate nel piano delle prime due componenti principali, che spiegano rispettivamente il 40.2% e il 17.3% della variabilità dei batteri considerati.

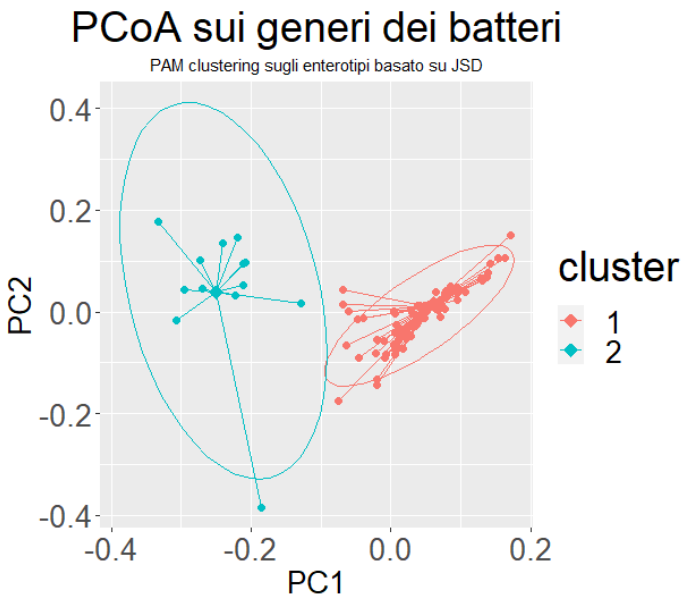


Figura 3. PCoA su matrice delle distanze e clustering

Dopo aver generato le prime due componenti principali, sono stati osservati i 5 batteri più correlati con le componenti. Per la prima componente principale si osservano i seguenti batteri:

	PC1
Bacteroides	0.80
Parabacteroides	0.38
Rikenellaceae	0.21
Enterococcus	0.20
f_Clostridium	0.18

	PC2
Clostridiales	-0.64
Ruminococcaceae	-0.60
f_S24-7	-0.55
Dehalobacterium	-0.54
RC4-4	-0.54

Tabella 2. Coefficiente di correlazione: i primi 5 batteri più correlati con le prime due componenti principali

Dalla tabella 2, vengono mostrate i 5 batteri più correlati rispetto alle due componenti principali: nel caso della PC1, sono più correlati i batteri che si sviluppano maggiormente in una dieta ad alto consumo di proteine e grassi animali, mentre la componente PC2 sembra essere negativamente associata con batteri che sono fortemente associati al consumo di zuccheri.
Osservando le composizioni dei due diversi cluster (in figura 4), si nota una netta differenziazione in termini di abbondanza dei generi Bacteroides e Prevotella. Mentre il cluster 1, rappresentato da soli 14 individui, registra un'alta abbondanza di Prevotella e un'abbondanza di Bacteroides prossima allo zero, il secondo raggruppamento ha dei valori esattamente opposti: in esso risiede una maggiore concentrazione di Bacteroides a scapito di una presenza del genere Prevotella molto ridotta. Inoltre, si è analizzata la possibile rilevanza della presenza di Ruminococcus all'interno dei due cluster ottenuti,

per verificare se la concentrazione di questo genere di batterio avesse una qualche influenza nella composizione del cluster. Tuttavia, si è osservato che Ruminococcus si distribuiva in egual modo nei due gruppi, suggerendo una scarsa capacità di discriminazione degli enterotipi ottenuti.

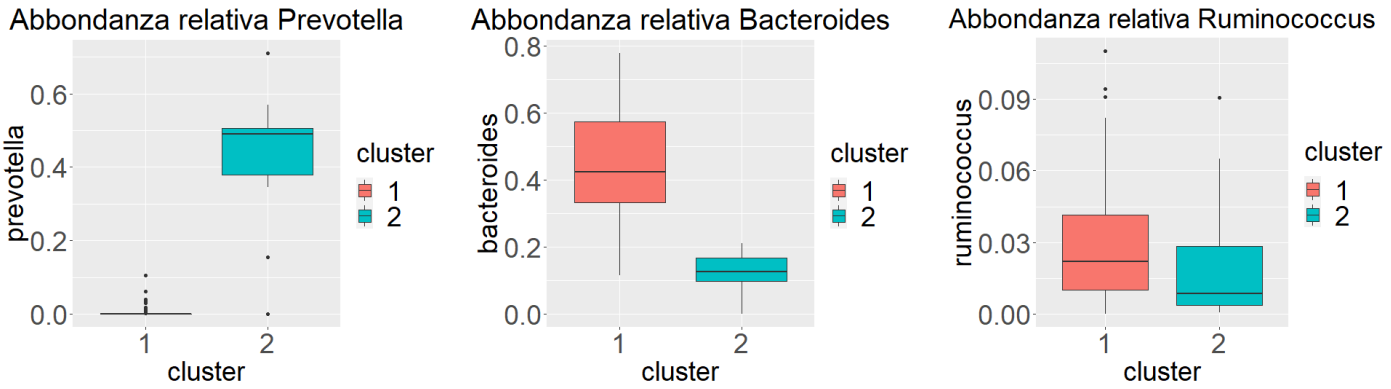


Figura 4. Abbondanze relative dei tre batteri suddivisi per cluster

Per ognuno dei due enterotipi sono state valutate le associazioni con i nutrienti selezionati in precedenza. L'associazione è stata calcolata come la media dei nutrienti standardizzati all'interno dei due differenti gruppi. In generale, come si evince dalla figura S2 (in appendice), si osserva in maniera netta un'associazione forte dell'enterotipo 2 (Prevotella) con carboidrati, zuccheri e glucosio. Questa associazione risulta invece essere opposta per diete ricche di colesterolo, grassi e proteine animali. L'enterotipo 1, caratterizzato da una maggiore presenza di Bacteroides, è invece debolmente associato, in maniera positiva, con nutrienti riconducibili ad amminoacidi, proteine e grassi animali. L'associazione diminuisce quando dal gruppo delle proteine e dei grassi animali si passa a quelli delle fibre vegetali e tutti quei nutrienti ricchi di zuccheri fino a diventare negativa.

4 Analisi del bmi e dei regressori

La seconda parte dello studio, si pone come obiettivo quello di sviluppare modelli capaci di prevedere la variabile dipendente BMI, in funzione di diversi regressori. Sono state utilizzate tecniche di riduzione della dimensionalità, in modo tale da poter sviluppare sia modelli lineari multivariati (con stimatore OLS), sia modelli che si basano sulla procedura di Stepwise e l'uso di stimatori Lasso per poter effettuare feature selection nella fase di training. Prima di entrare nel dettaglio di tali modelli, è stata analizzata la variabile dipendente. Nella tabella successiva vengono mostrate delle misure sulla distribuzione del BMI:

	Media	**Mediana**	**Deviazione Standard**	**Skewness**
Bmi	24.5	23.62	5.44	1.39

Tabella 3. Misure sulla variabile di interesse

Dai risultati, in tabella 3, emerge che la media è maggiore della mediana, e questo testimonia un'asimmetria positiva nella distribuzione, confermato anche dal valore della skewness pari a 1.39. Nella figura 5, viene rappresentata la distribuzione empirica della y a confronto con quella teorica normale.

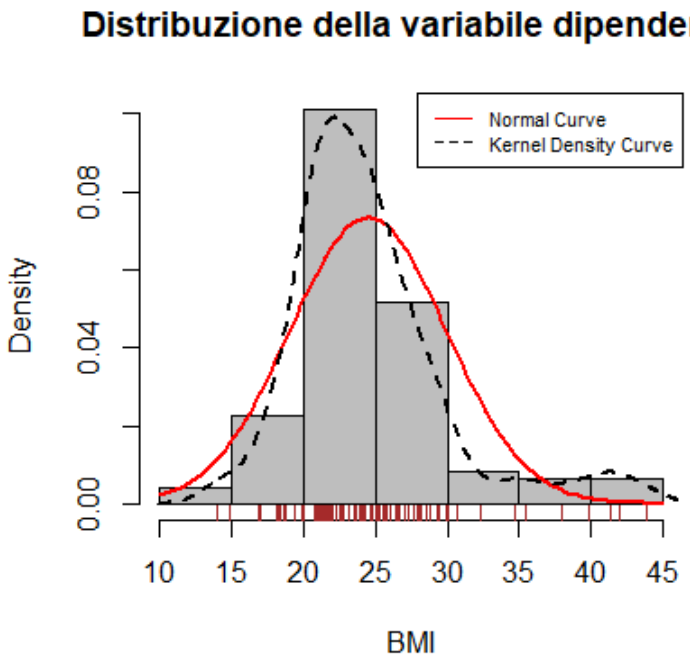


Figura 5. Distribuzione variabile y (BMI)

Dal grafico 5 emerge che il BMI presenta, nella sua distribuzione, un'asimmetria dovuta possibilmente a degli outlier. Per effettuare un'ulteriore analisi, si mostra il grafico (in figura 6) qqplot della variabile d'interesse e viene applicato il test non parametrico Shapiro-Wilk utile per verificare la normalità della distribuzione:

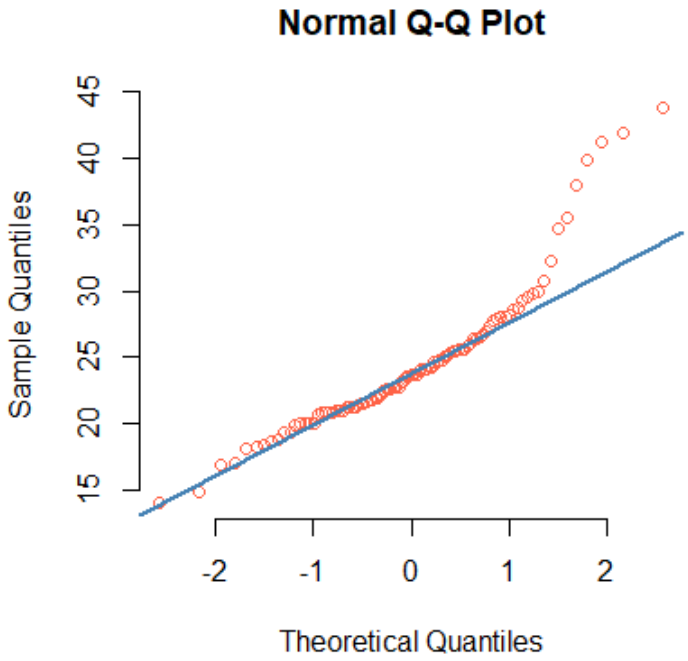


Figura 6. Normal QQ-plot della variabile bmi

Il grafico 6 assume una forma più asimmetrica che normale, e si mette in evidenza come la coda di destra influenza la distribuzione empirica. Applicando il test Shapiro-Wilk, si ottiene una statistica

W pari a 0.88 e viene rigettata l'ipotesi nulla di normalità con alfa pari a 0.05. Per porre rimedio a tale problema, si fa uno studio degli outlier nella y. Esistono diverse tecniche per identificare i valori anomali, tra cui il box plot:

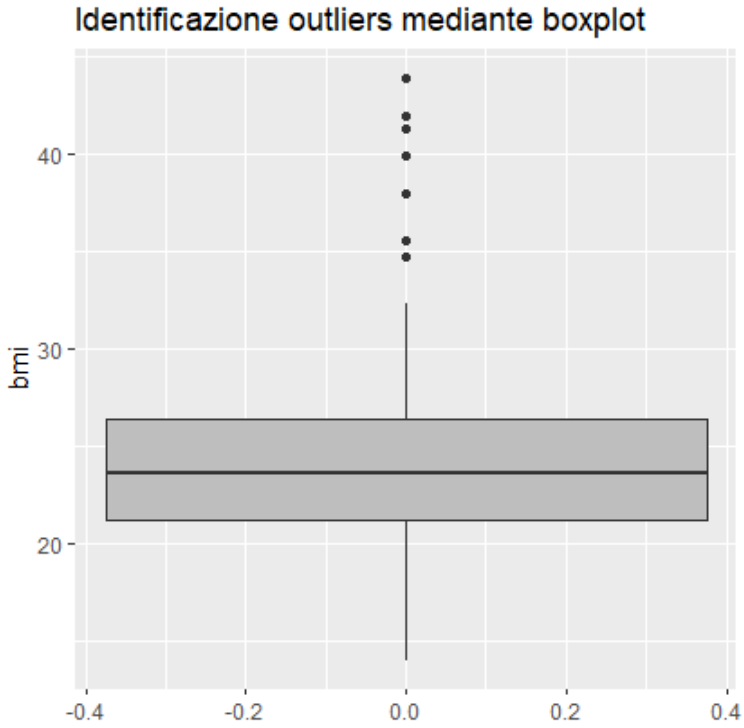


Figura 7. Box-plot BMI

Osservando la figura 7, si può concludere che esistono degli outlier che influenzano la distribuzione. Perciò una volta identificati, sono state rimosse 6 osservazioni che presentavano un bmi troppo alto. Dal grafico 8, si può osservare come tale approccio abbia corretto il problema delle code pesate:

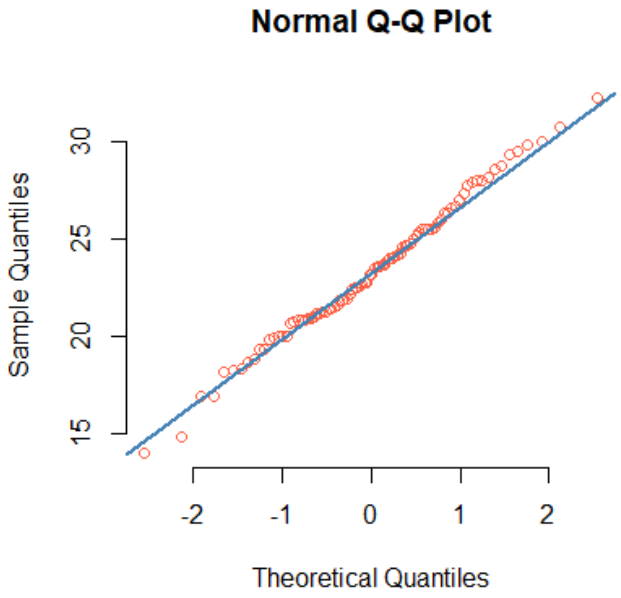


Figura 8. Normal QQ-plot della variabile bmi senza outliers

Per confermare l'efficacia della rimozione degli outlier, è stato applicato di nuovo il test non parametrico SW, accettando l'ipotesi nulla di normalità con alfa pari a 0.05 e statistica W pari a 0.99. Conseguentemente a questo, viene rispettata anche l'ipotesi di normalità degli errori che permette di ottenere risultati di grande rilievo sia per quanto riguarda la stima puntuale e intervallare dei coefficienti di regressione, sia per ciò che attiene alla verifica di ipotesi su di essi. Successivamente, è stata analizzata la distribuzione della y rispetto all'età e il sesso, in modo tale da identificare eventuali sottopopolazioni nei dati. Una volta verificata l'assunzione di normalità rispetto alla y e di varianza costante nei gruppi, abbiamo applicato il t-test a due campioni per verificare se la distribuzione della popolazione maschile differisse in modo significativo da quella femminile: il test accetta l'ipotesi nulla di uguaglianza, in media, tra le due popolazioni per alfa pari a 0.05. In figura 9, viene mostrata tale distribuzione. Un'altra variabile interessante è quella legata all'età dei pazienti: dopo aver discretizzato l'età in 2 intervalli basati sui quantili della distribuzione, si è applicato il test Chi-quadro per verificare un'eventuale dipendenza tra le varie fasce di età e il BMI.

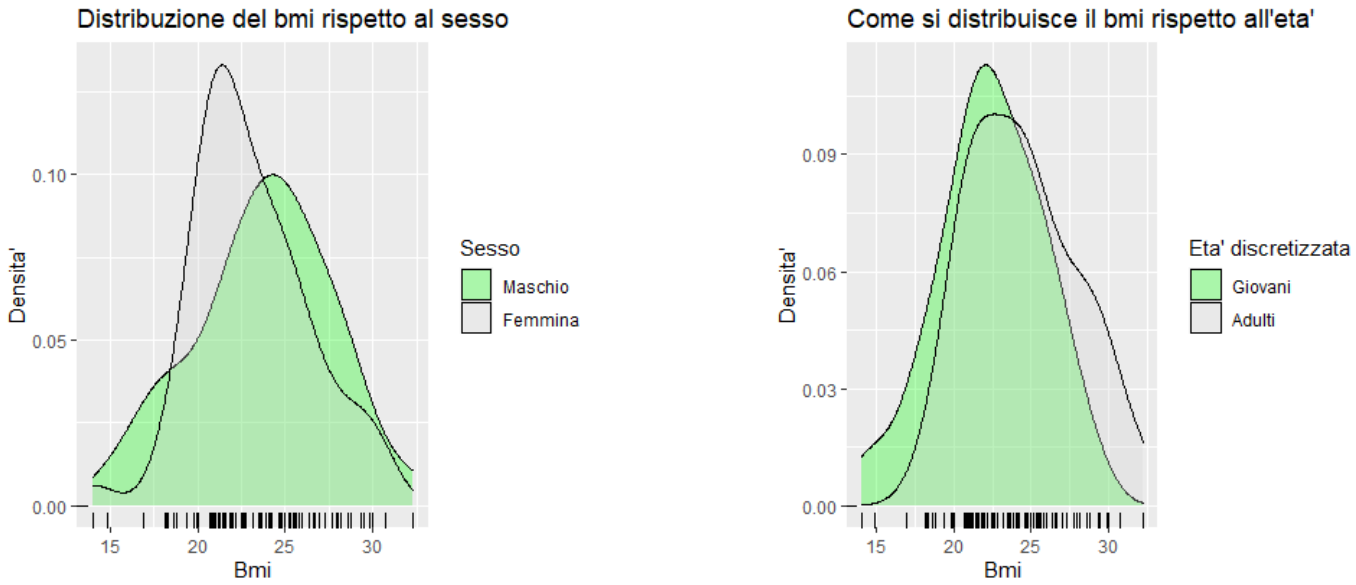


Figura 9. Distribuzione bmi rispetto alle variabili categoriali sesso ed età (discretizzata)

I risultati ottenuti evidenziano che non esiste una stretta dipendenza tra i due gruppi di età e la variabile y, e per questo si accetta l'ipotesi nulla di indipendenza con alfa pari a 0.05. Dalla figura 9, si nota come la fascia di età degli adulti tende ad avere un BMI più alto rispetto ai giovani.
Successivamente, è stata effettuata l'analisi della correlazione sulla base del coefficiente di Spearman (questo ci permette di non fare assunzioni sulle distribuzioni). Nella figura 10, viene studiata la relazione tra il bmi e un campione dei regressori, come ad esempio: le componenti principali dei batteri, dei nutrienti e alcune variabili demografiche.

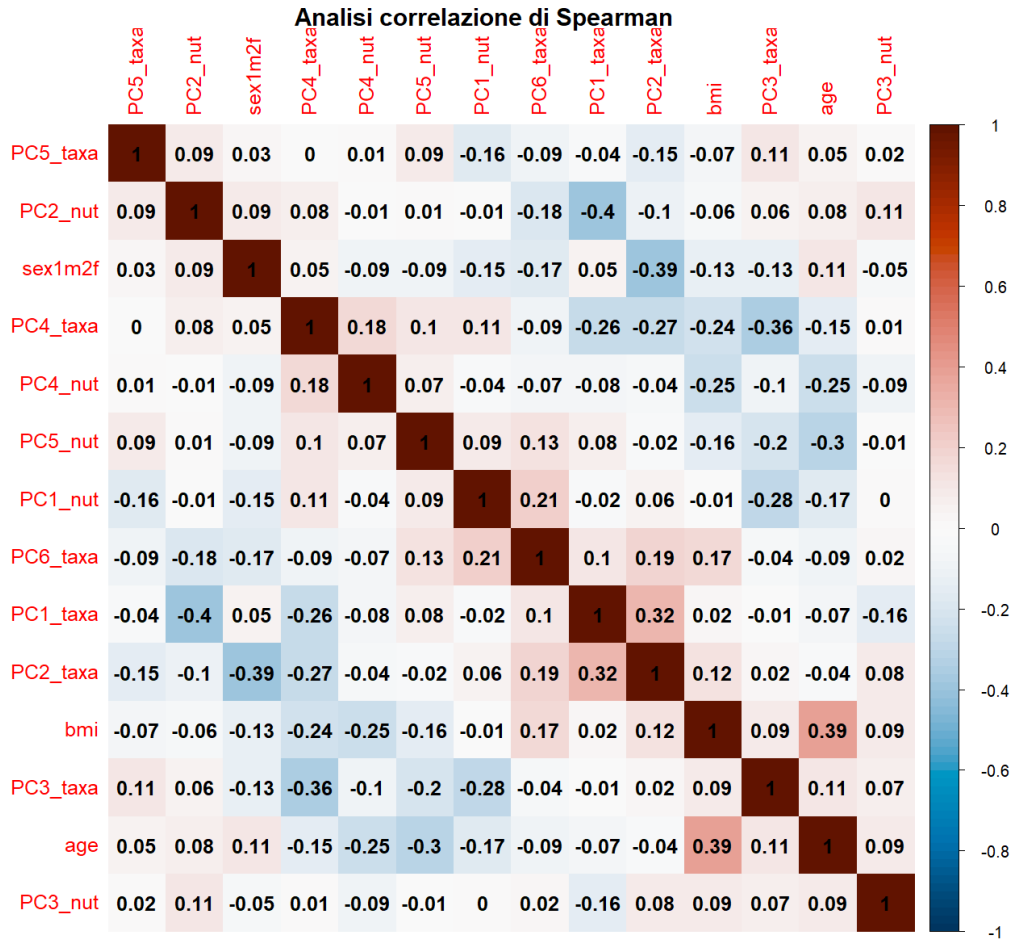


Figura 10. Matrice di correlazione di Spearman

Grazie a questo tipo di misura, si può osservare, a priori, se esiste una possibile collinearità tra i vari regressori. Da come si evince dalla figura 10, non esistono correlazioni forti da poter generare un problema di multicollinearità. Per avere la conferma di quanto detto, dopo aver sviluppato i diversi modelli, sono stati utilizzati indicatori come il VIF o il TOL.

5 Modelli

Per quanto riguarda il task di previsione del BMI, è stato svolto uno studio comparativo tra diversi modelli basandosi su due metriche: RMSE e R-squared adjusted. In totale, sono stati sviluppati 6 modelli con diversi regressori di input. Tre modelli si basano su stimatori OLS mentre gli altri 3 su stimatore LASSO: modelli con stimatore lasso si prestano ad essere usati quando n < p: è sia un metodo di Shrinkage sia un metodo di selezione delle variabili. Nel paragrafo dei modelli lasso, verrà ricercato l'iperparametro lambda nell'intervallo tra 0.1 e 2.
L'analisi è strutturata in due parti: valutazione delle metriche in 10-folds CV, in modo tale da fornire dei risultati più parsimoniosi e generare dei possibili intervalli di confidenza (rispetto al RMSE) e il confronto finale, dei diversi modelli, valutando la previsione fatta sul test set.

5.1 Modello lineare 1

Il primo modello scelto si basa sulle prime 6 componenti principali dei batteri, l'età, il sesso e il cluster identificato dagli enterotipi: i coefficienti significativi di questo modello sono: β_0,β_4,β_7, β_8 per un alfa pari a 0.05. Il test F sul modello è significativo per un alfa pari a 0.05.

$$bmi = \beta_0 + \beta_1 PC1taxa + \beta_2 PC2taxa + \beta_3 PC3taxa + \beta_4 PC4taxa + \beta_5 PC5taxa + \beta_6 PC6taxa + \beta_7 age + \beta_8 sex + \beta_9 cluster + \epsilon \quad (1)$$

5.2 Modello lineare 2

I regressori utilizzati per questo modello sono le prime 5 componenti principali dei nutrienti, l'età, il sesso e il cluster: i coefficienti significativi di questo modello sono β_0,β_6,β_7 per un alfa pari a 0.05. Il test F sul modello è significativo per un alfa pari a 0.05.

$$bmi = \beta_0 + \beta_1 PC1nutrients + \beta_2 PC2nutrients + \beta_3 PC3nutrients + \beta_4 PC4nutrients + \beta_5 PC5nutrients + \beta_6 age + \beta_7 sex + \beta_8 cluster + \epsilon \quad (2)$$

5.3 Stepwise regression: Forward vs Backward

Poichè il numero di regressori p è maggiore del numero di osservazioni n, si deve ridurre la dimensionalità dei regressori. Procedure come best subset selection potrebbero individuare un sottoinsieme ottimale di modelli, a discapito del tempo computazionale. In questo caso, p è talmente elevato che non permette di provare tutte le combinazioni. Per tale motivo, è stata utilizzata la procedura stepwise su un dataset aggregato sulla tassonomia 6, componenti principali dei nutrienti e le variabili demografiche. In questo caso, sono stati valutati 46 regressori differenti. Mostriamo i risultati ottenuti dalla procedura forward e backward in 10 fold cross validation. Gli intervalli di confidenza al 95%, per la metrica RMSE, sono stati generati sulla base della seguente formula (3):

$$RMSE_{meancv} \pm 1.96 \times ((1/\sqrt{k}) sd(Err^{-1}, ..., Err^{-k})) \quad (3)$$

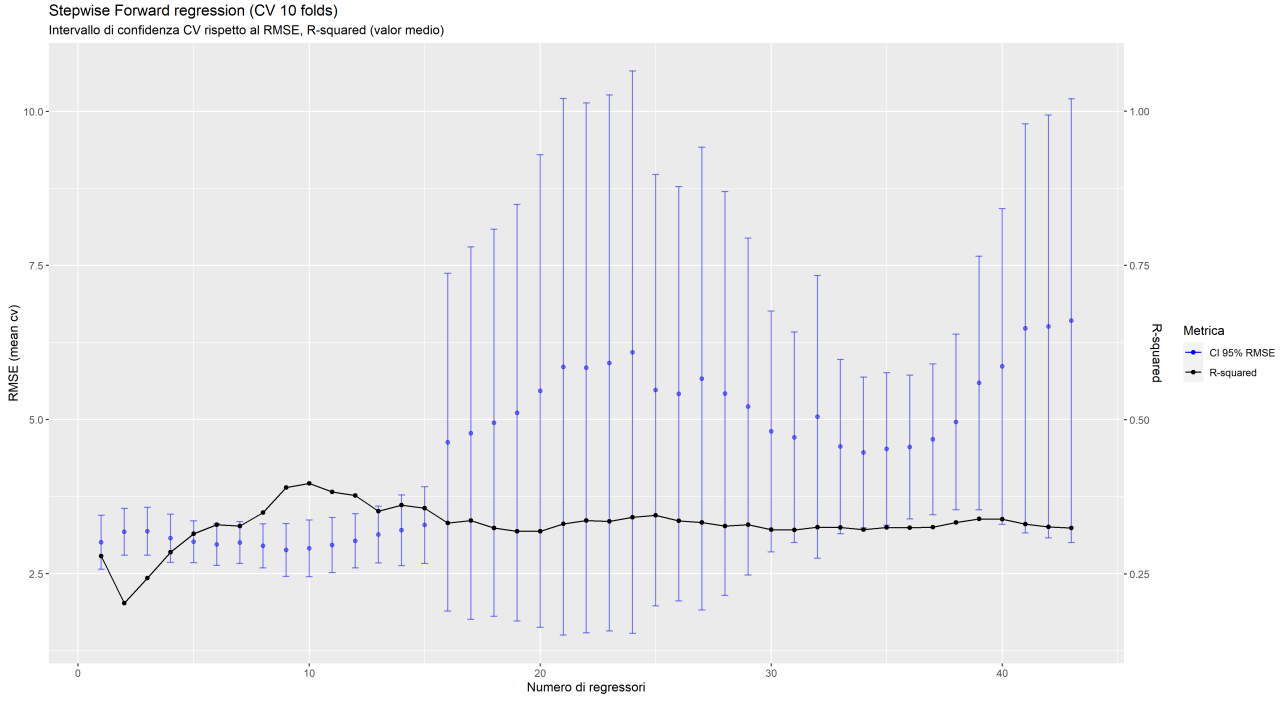


Figura 11. Procedura Stepwise Forward in 10 folds cross validation

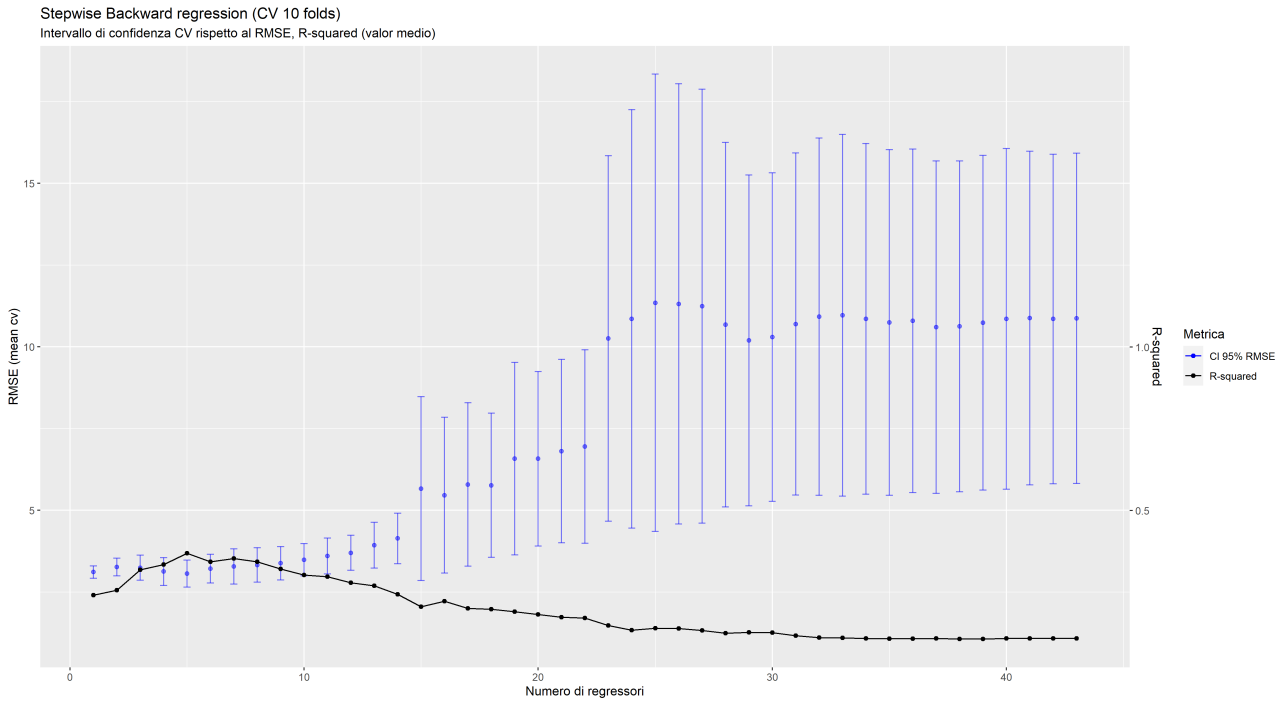


Figura 12. Procedura Stepwise Backward in 10 folds cross validation

Da come si evince dalle figure 11 e 12, la procedura stepwise forward produce dei risultati migliori rispetto alla backward. Infatti, riesce a produrre un modello con valori migliori di R-squared e RMSE. Tuttavia, osservando gli intervalli di confidenza, del RMSE, per modelli con un numero di regressori inferiore (fino a 10), non vi è una differenza significativa. Pertanto, è stato scelto il modello con 9 regressori generato dalla procedura stepwise forward. Segue l'equazione del modello.

$$bmi = \beta_0 + \beta_1 Parabacteroides + \beta_2 Blautia + \beta_3 Lachnobacterium + \beta_4 Megasphaera + \beta_5 Phascolarctobacterium + \beta_6 Veilonella + \beta_7 Catenibacterium + \beta_8 age + \beta_9 sex + \epsilon \quad (4)$$

Tutti i coefficienti risultano significativi per un alfa pari a 0.05 e per il test F sul modello si rigetta l'ipotesi nulla per un alfa pari a 0.05.

5.4 Modello lasso 1

Per quanto riguarda il lasso, i regressori usati sono i seguenti: tutti i nutrienti assunti dalle persone, sesso, età e cluster dell'enterotipo. Tra tutti questi, quelli risultati significativi per spiegare il BMI sono i seguenti: vitamina k1 , acido stearico, idrossiprolina, flavonoidi, pelargonidina antocianidina, sesso ed età. Il regressore risultato più importante è proprio l'età: segno che essa influenzi molto il BMI. Nella figura 13, si può osservare l'andamento del RMSE con relativo intervallo di confidenza al 95% di significatività ed R^2 adjusted al variare dell'iperparametro Lambda. Il valore scelto è 0.67 dato che esso massimizza l'R^2 adjusted e minimizza l'RMSE, sebbene esso non sia significativamente migliore di tutti gli altri lambda.

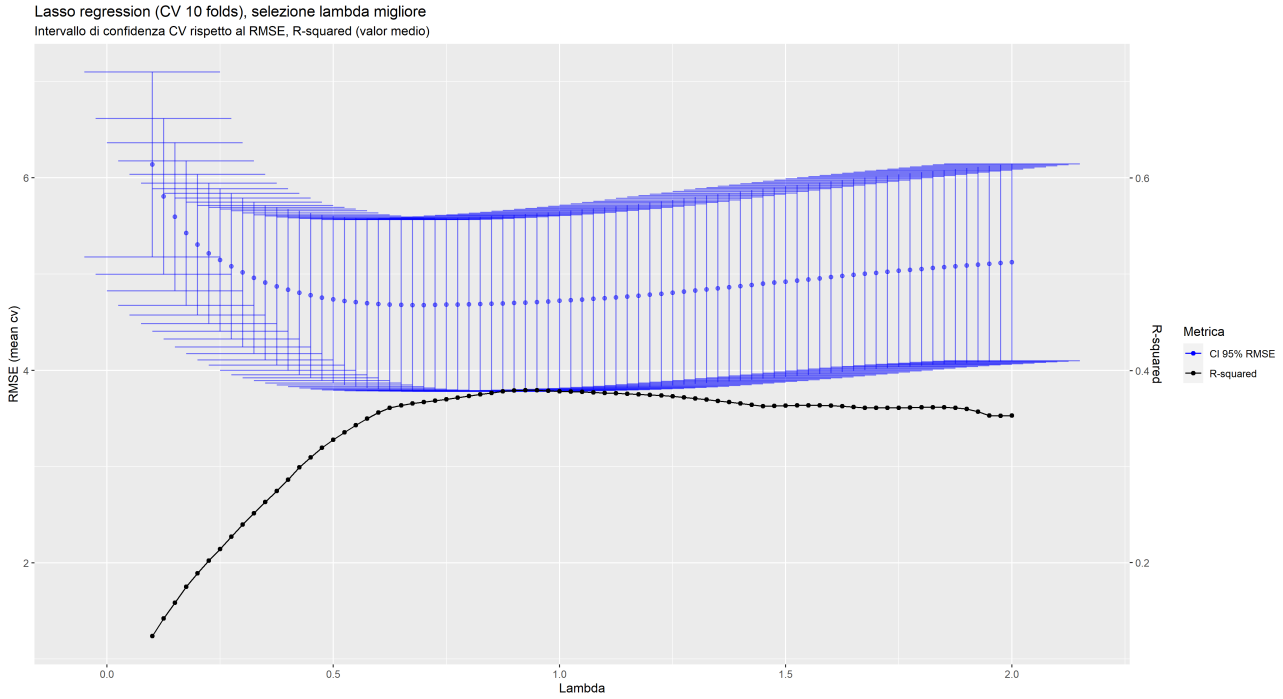


Figura 13. Selezione del λ ottimale in 10 folds cross validation, per il lasso 1

5.5 Modello lasso 2

Per il modello 2, i regressori usati sono i seguenti: batteri, sesso, età e cluster dell'enterotipo delle persone. Tra tutti questi, quelli risultati significativi per spiegare il BMI sono i seguenti: Megasphaera, sesso ed età. Nella figura 14, si può osservare l'andamento del RMSE con relativo intervallo di confidenza al 95% di significatività ed R^2 adjusted al variare dell'iperparametro Lambda. Il valore scelto è 0.70 dato che esso massimizza l'R^2 adjusted e minimizza l'RMSE sebbene esso non sia significativamente migliore di tutti gli altri lambda.

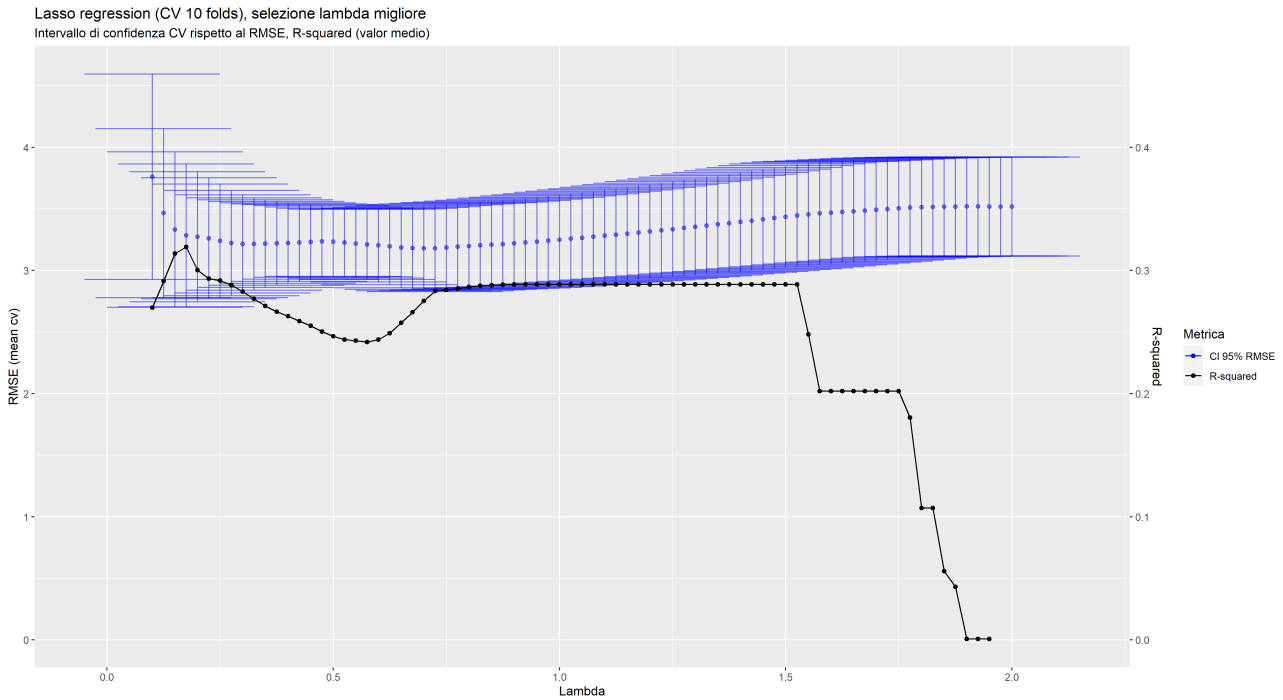


Figura 14. Selezione del λ ottimale in 10 folds cross validation, per il lasso 2

5.6 Modello lasso 3

Per il modello 3, sono stati usati tutti i regressori a disposizione. Tra tutti questi, quelli risultati significativi per spiegare il BMI sono i seguenti: età, sesso e Megasphaera. Il valore di lambda scelto è 0.725 dato che esso massimizza l'R^2 adjusted e minimizza l'RMSE.

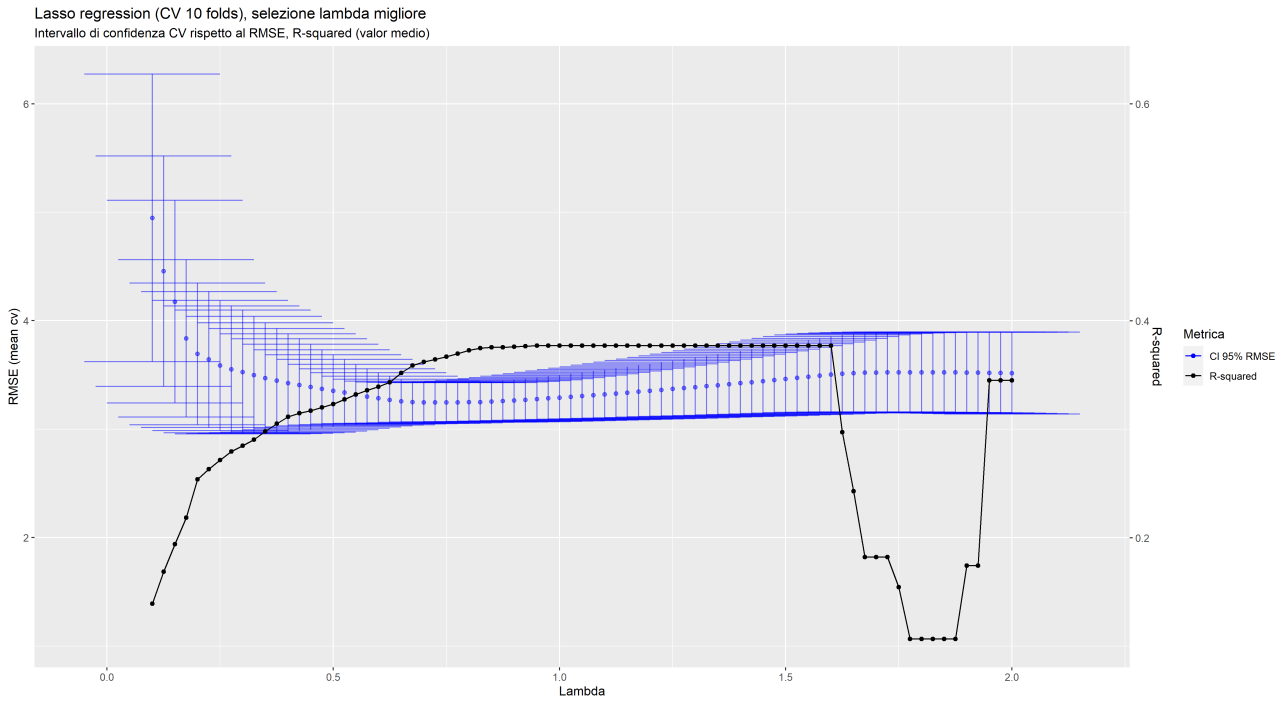


Figura 15. Selezione del λ ottimale in 10 folds cross validation, per il lasso 3

5.7 Risultati dei modelli

Una volta identificati i diversi modelli "ottimali", sono stati analizzati i risultati ottenuti mediante la cross-validation:

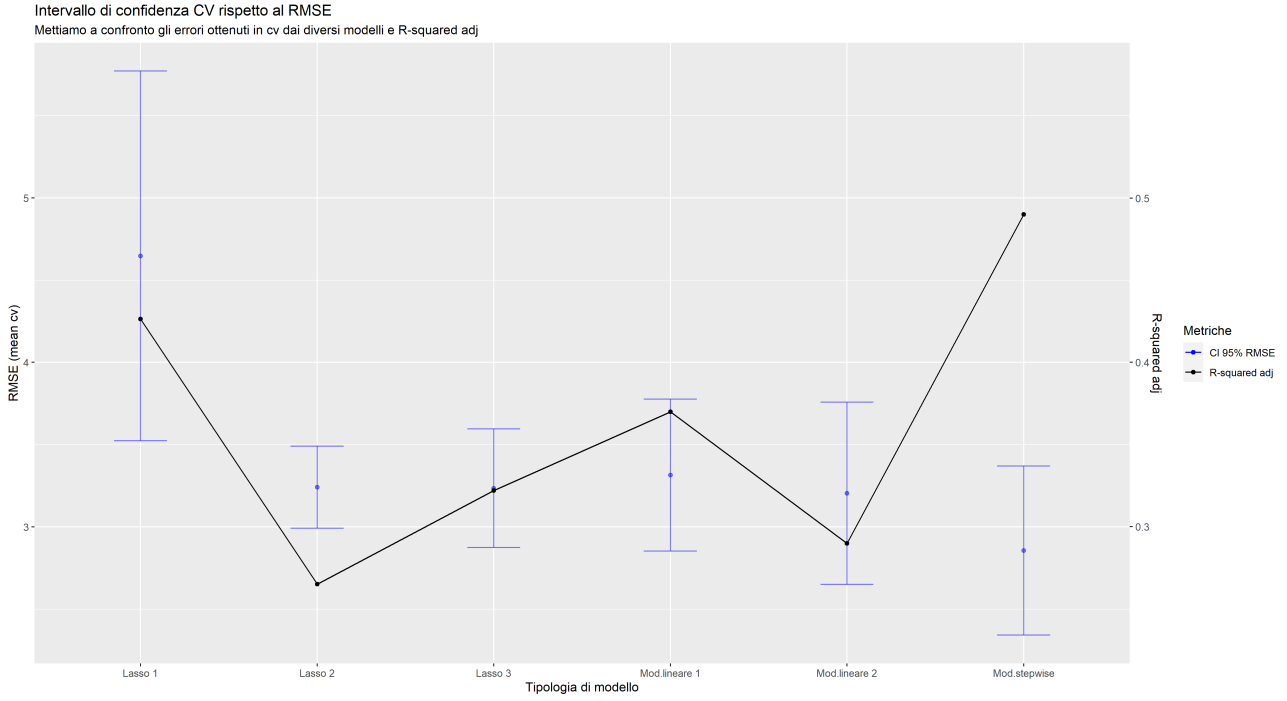


Figura 16. Confronto, in 10 folds cross validation, tra i vari modelli sviluppati

Nella figura 16, si confrontano i modelli di regressione con stimatore OLS con quelli Lasso opportunamente calcolati con i migliori iper-parametri. Sfruttando gli intervalli di confidenza, generati in fase di cross validation, sempre sulla base della formula (3), si identificano i modelli migliori. Dalla figura 16, è emerso che non esiste una differenza significativa tra i vari modelli, in termini di RMSE. Le uniche differenze significative sono tra il modello stepwise rispetto a lasso 1 e lasso 2 rispetto al lasso 1. Inoltre, il modello che riesce a catturare più varianza, è sempre quello selezionato dalla procedura forward. Una volta stimati i modelli, sono state effettuate le previsioni sul test set:

Modello	**RMSE test set**
Modello lineare 1	3.22
Modello lineare 2	3.31
Modello Stepwise	2.73
Lasso 1	3.52
Lasso 2	3.35
Lasso 3	2.92

Tabella 4. Risultati RMSE dei vari modelli rispetto al test set

Dalla tabella 4, emerge che il modello stepwise è quello più affidabile in termini di risultati tra la cross validation e il test set. Il metodo stepwise forward ha permesso di individuare un modello che in termini di trade off tra varianza e distorsione riesce ad essere quello più parsimonioso, raggiungendo dei risultati affidabili in fase di inferenza.

6 Conclusioni

Dall'analisi è emersa una forte associazione tra dieta ed enterotipo, caratterizzato dalla presenza/assenza dei batteri Bacteroides e Prevotella entrambi appartenenti al phylum Bacteroidetes. L'enterotipo con maggiore presenza di Prevotella risulta fortemente associato con diete ricche di fibre e carboidrati mentre la presenza di Bacteroides è associata con diete ricche di proteine e grassi animali. Poichè, il BMI è calcolato tramite peso ed altezza è interessante valutare se le associazioni con i nutrienti ed i batteri lo influenzano. I modelli stimati evidenziano che l'età è un fattore molto influente per il BMI mentre l'enterotipo non lo è. I batteri significativi per spiegare il BMI sono Catenibacterium, Megasphaera, Parabacteroides, Blautia, Lachnobacterium, Phascolarctobacterium, Veillonella, Catenibacterium . Tra questi solo Parabacteroides appartiene al phylum Bacteroidetes mentre gli altri sono del phylum Firmicutes. I batteri del phylum Firmicutes, ad eccezione di Lacnobacteria, influenzano positivamente il BMI mentre Parabacteroides negativamente. Tuttavia,i modelli sviluppati, dati questi regressori, in termini di R-squared non riescono a raggiungere livelli soddisfacenti affinchè si ritengano dei buoni modelli. In conclusione, dal nostro studio si evince che i firmicutes sono i batteri che potrebbero influenzare la crescita del BMI insieme all'età.

7 Sviluppi Futuri

Dal momento che il BMI è calcolato sulla base del rapporto tra peso e altezza, sarebbe interessante analizzare, come variabile dipendente, il peso. Tale analisi metterebbe in relazione il peso con i seguenti regressori: età, nutrienti, batteri e altezza. Questo permetterebbe di risalire direttamente ai fattori che influenzano il peso e a sua volta il BMI.
Come è noto, il peso è influenzato anche dallo stile di vita di una persona, ed in particolare, dall'attività fisica che essa svolge. Perciò avere informazioni aggiuntive sotto questo aspetto ritornebbe utile per la costruzione di un modello capace di essere più discriminativo.
Inoltre, i limiti che sono stati riscontrati, sono dovuti anche dalla numerosità campionaria: sarebbe ideale raccogliere altri campioni in modo tale da poter creare modelli più robusti.

References

[1] Wu, Gary D., et al. "Linking long-term dietary patterns with gut microbial enterotypes." Science 334.6052 (2011): 105-108.

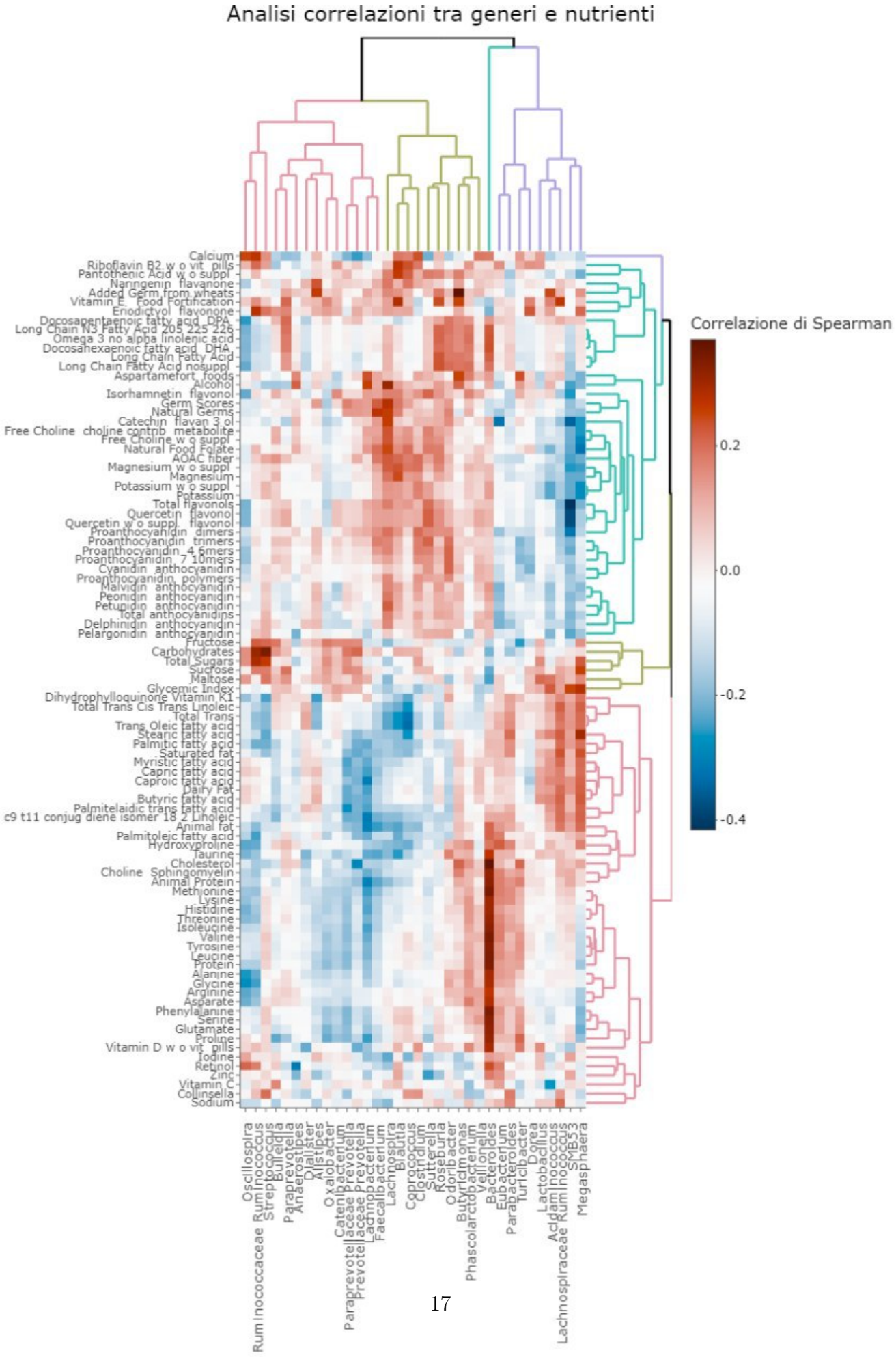
[2] McArdle, Brian H., and Marti J. Anderson. "Fitting multivariate models to community data: a comment on distance based redundancy analysis." Ecology 82.1 (2001): 290 297.

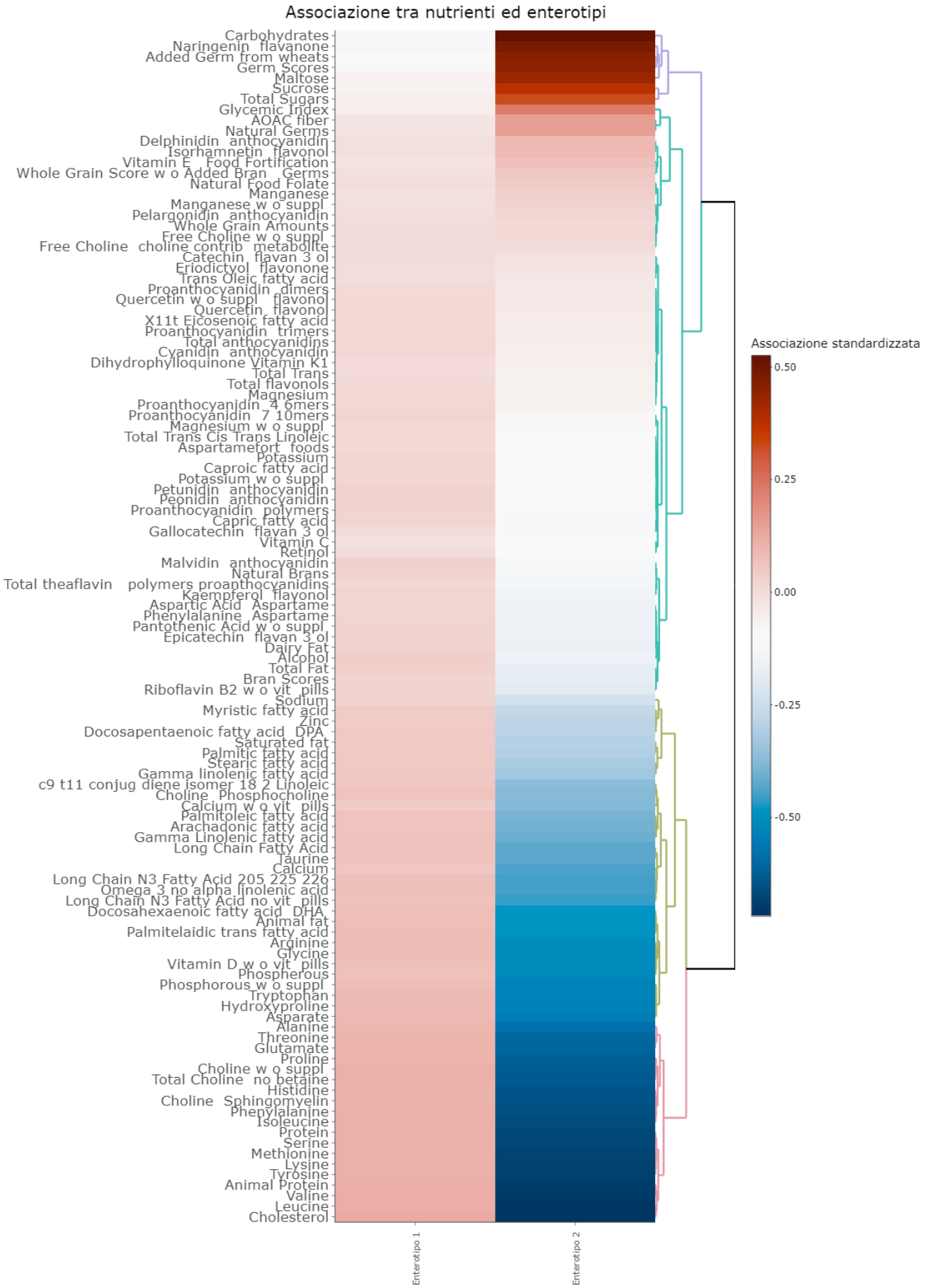
[3] Lozupone, Catherine, et al. "UniFrac: an effective distance metric for microbial community comparison." The ISME journal 5.2 (2011): 169 172.

[4] https://bit.ly/2zs5YFv

8 Appendice

Nella la figura S1 rappresenta la correlazione di Spearman tra nutrienti e batteri, mentre la S2 mostra le associazioni standardizzate tra enterotipi e nutrienti


Analisi correlazioni tra generi e nutrienti
Correlazione di Spearman
0.2
0.0
-0.2
-0.4
Calcium
Riboflavin B2 w o vit pills
Pantothenic Acid w o suppl
Naringenin flavanone
Added Germ from wheats
Vitamin E Food Fortification
Eriodictyol flavonone
Docosapentaenoic fatty acid DPA
Long Chain N3 Fatty Acid 205 225 226
Omega 3 no alpha linolenic acid
Docosahexaenoic fatty acid DHA
Long Chain Fatty Acid
Long Chain Fatty Acid nosuppl
Aspartamefort foods
Alcohol
Isorhamnetin flavonol
Germ Scores
Natural Germs
Catechin flavan 3 ol
Free Choline choline contrib metabolite
Free Choline w o suppl
Natural Food Folate
AOAC fiber
Magnesium w o suppl
Magnesium
Potassium w o suppl
Potassium
Total flavonols
Quercetin flavonol
Quercetin w o suppl flavonol
Proanthocyanidin dimers
Proanthocyanidin trimers
Proanthocyanidin 4 6mers
Proanthocyanidin 7 10mers
Cyanidin anthocyanidin
Proanthocyanidin polymers
Malvidin anthocyanidin
Peonidin anthocyanidin
Petunidin anthocyanidin
Total anthocyanidins
Delphinidin anthocyanidin
Pelargonidin anthocyanidin
Fructose
Carbohydrates
Total Sugars
Sucrose
Maltose
Glycemic Index
Dihydrophylloquinone Vitamin K1
Total Trans Cis Trans Linoleic
Total Trans
Trans Oleic fatty acid
Stearic fatty acid
Palmitic fatty acid
Saturated fat
Myristic fatty acid
Capric fatty acid
Caproic fatty acid
Dairy Fat
Butyric fatty acid
Palmitelaidic trans fatty acid
c9 t11 conjug diene isomer 18 2 Linoleic
Animal fat
Palmitoleic fatty acid
Hydroxyproline
Taurine
Cholesterol
Choline Sphingomyelin
Animal Protein
Methionine
Lysine
Histidine
Threonine
Isoleucine
Valine
Tyrosine
Leucine
Protein
Alanine
Glycine
Arginine
Asparate
Phenylalanine
Serine
Glutamate
Proline
Vitamin D w o vit pills
Iodine
Retinol
Zinc
Vitamin C
Collinsella
Sodium
Oscillospira
Ruminococcus
Streptococcus
Bulleidia
Paraprevotella
Anaerostipes
Dialister
Alistipes
Oxalobacter
Catenibacterium
Paraprevotellaceae Prevotella
Prevotellaceae Prevotella
Lachnobacterium
Faecalibacterium
Lachnospira
Blautia
Coprococcus
Clostridium
Sutterella
Roseburia
Odoribacter
Butyricimonas
Phascolarctobacterium
Veillonella
Bacteroides
Eubacterium
Parabacteroides
Turicibacter
Dorea
Lactobacillus
Adlercreutzia
Ruminococcaceae Ruminococcus
Lachnospiraceae Ruminococcus
SMB53
Megasphaera


Associazione tra nutrienti ed enterotipi
Carbohydrates
Naringenin flavanone
Added Germ from wheats
Germ Scores
Maltose
Sucrose
Total Sugars
Glycemic Index
AOAC fiber
Natural Germs
Delphinidin anthocyanidin
Isorhamnetin flavonol
Vitamin E Food Fortification
Whole Grain Score w o Added Bran Germs
Natural Food Folate
Manganese
Manganese w o suppl
Pelargonidin anthocyanidin
Whole Grain Amounts
Free Choline w o suppl
Free Choline choline contrib metabolite
Catechin flavan 3 ol
Eriodictyol flavonone
Trans Oleic fatty acid
Proanthocyanidin dimers
Quercetin w o suppl flavonol
Quercetin flavonol
X11t Eicosenoic fatty acid
Proanthocyanidin trimers
Total anthocyanidins
Cyanidin anthocyanidin
Dihydrophylloquinone Vitamin K1
Total Trans
Total flavonols
Magnesium
Proanthocyanidin 4 6mers
Proanthocyanidin 7 10mers
Magnesium w o suppl
Total Trans Cis Trans Linolejc
Aspartamefort foods
Potassium
Caproic fatty acid
Potassium w o suppl
Petunidin anthocyanidin
Peonidin anthocyanidin
Proanthocyanidin polymers
Capric fatty acid
Gallocatechin flavan 3 ol
Vitamin C
Retinol
Malvidin anthocyanidin
Natural Brans
Total theaflavin polymers proanthocyanidins
Kaempferol flavonol
Aspartic Acid Aspartame
Phenylalanine Aspartame
Pantothenic Acid w o suppl
Epicatechin flavan 3 ol
Dairy Fat
Alcohol
Total Fat
Bran Scores
Riboflavin B2 w o vit pills
Sodium
Myristic fatty acid
Zinc
Docosapentaenoic fatty acid DPA
Saturated fat
Palmitic fatty acid
Stearic fatty acid
Gamma linolenic fatty acid
c9 t11 conjug diene isomer 18 2 Linoleic
Choline Phosphocholine
Calcium w o vit pills
Palmitoleic fatty acid
Arachadonic fatty acid
Gamma Linolenic fatty acid
Long Chain Fatty Acid
Taurine
Calcium
Long Chain N3 Fatty Acid 205 225 226
Omega 3 no alpha linolenic acid
Long Chain N3 Fatty Acid no vit pills
Docosahexaenoic fatty acid DHA
Animal fat
Palmitelaidic trans fatty acid
Arginine
Glycine
Vitamin D w o vit pills
Phospherous
Phosphorous w o suppl
Tryptophan
Hydroxyproline
Asparate
Alanine
Threonine
Glutamate
Proline
Choline w o suppl
Total Choline no betaine
Histidine
Choline Sphingomyelin
Phenylalanine
Isoleucine
Protein
Serine
Methionine
Lysine
Tyrosine
Animal Protein
Valine
Leucine
Cholesterol
Enterotipo 1
Enterotipo 2
Associazione standardizzata
0.50
0.25
0.00
-0.25
-0.50