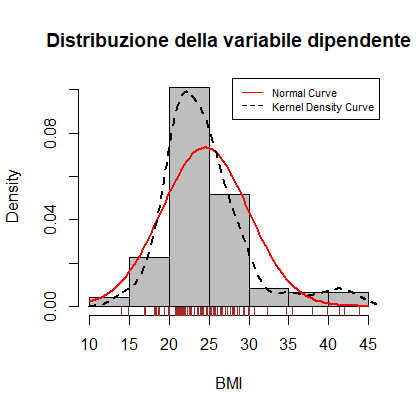
**Analisi del bmi e dei regressori:**

La seconda parte dello studio, si pone come obiettivo quello di sviluppare modelli capaci di prevedere la variabile y BMI, in funzione di diversi regressori. Verranno utilizzate tecniche di riduzione della dimensionalità in modo tale da poter sviluppare sia modelli lineari multivariati, 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, abbiamo analizzato la nostra variabile dipendente oggetto di interesse, e studiato le varie relazioni con alcuni regressori. Nella tabella successiva vengono mostrate alcune misure legate al bmi:

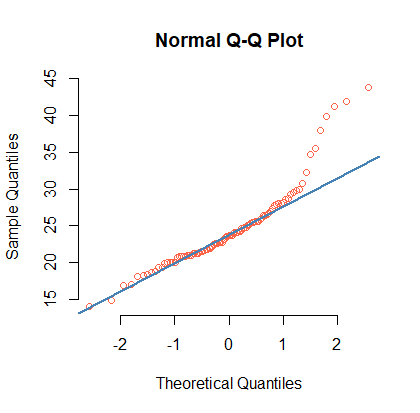
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Media | Mediana | Skewness | Stand. Dev. |  |
| B.M.I. | 24.50 | 23.62 | 1.39 | 5.44 |  |
|  |  |  |  |  |  |

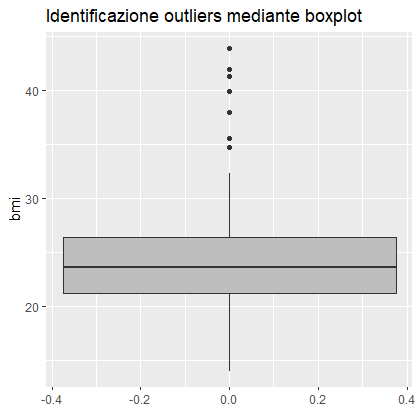
Dai i risultati emersi si evidenzia come la media è maggiore della media, e questo ci dice che sussiste un’asimmetria positiva nella distribuzione, confermato anche dalla skewness che è pari a 1.39. Nella figura successiva viene rappresentata la distribuzione empirica della y e come dovrebbe essere nel caso di normalità.



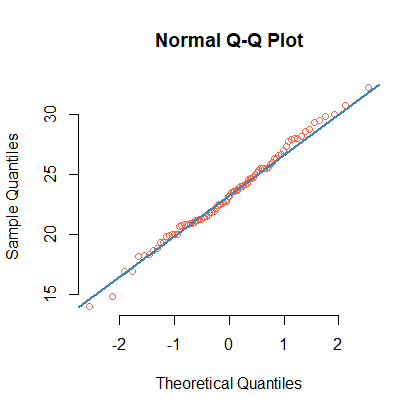
Dal grafico emerge che il bmi presenta, nella sua distribuzione un’asimmetria dovuta da possibili outliers. Per effettuare un ulteriore analisi mostriamo il qqplot della variabile di interesse e applichiamo un test non parametrico, come lo Shapiro-Wilk, per verificare la normalità della distribuzione:

Qqplot y con outliers:

  
Dal grafico superiore si vede come la forma sia più leptocurtica piuttosto che normale, e si mette in evidenza come le code influenzano la distribuzione empirica. Applicando lo Shapiro-Wilk otteniamo una statistica W pari a 0.88 e rigettiamo l’ipotesi nulla di normalità per un alfa pari a 0.05. Per porre rimedio a tale problema, analizziamo gli eventuali outliers nella y. Esistono diverse tecniche per identificare i valori anomali, tra cui il box plot:

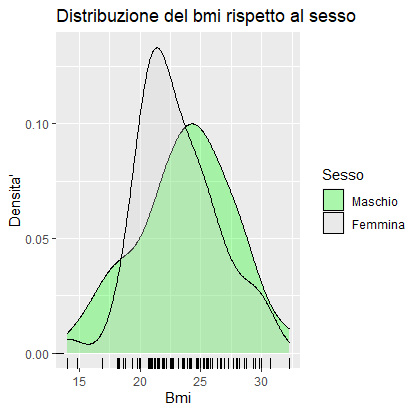


Osservando semplicemente il box plot possiamo concludere che esistono degli outliers che influenzano la distribuzione, perciò una volta identificati sono stati rimossi (6 pazienti che presentavano un bmi troppo alto e troppo basso rispetto alla nuvola dei punti del bmi). Osserviamo come tale approccio ha corretto il problema delle code pesate:

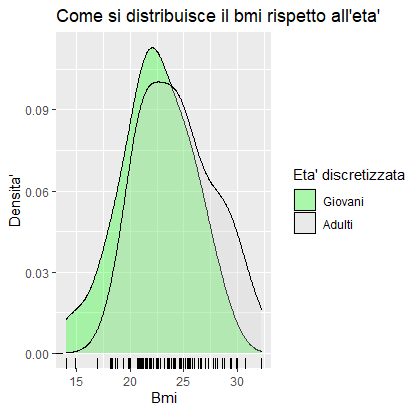


Dalla figura si evince come l’eliminazione degli outliers hanno portato ad avere una forma meno leptocurtica. Per conferma, abbiamo applicato di nuovo il test non parametrico di SW, dove si accetta l’ipotesi nulla di normalità con un alfa pari a 0.05 e una statistica W pari a 0.99. Conseguentemente a 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 sia quella intervallare. Oltre a cercare di rispettare le assunzioni alla base del modello lineare, avendo scelto di utilizzare come metrica di valutazione l’RMSE, abbiamo eliminato tali valori anomali in modo tale da renderla più robusta.

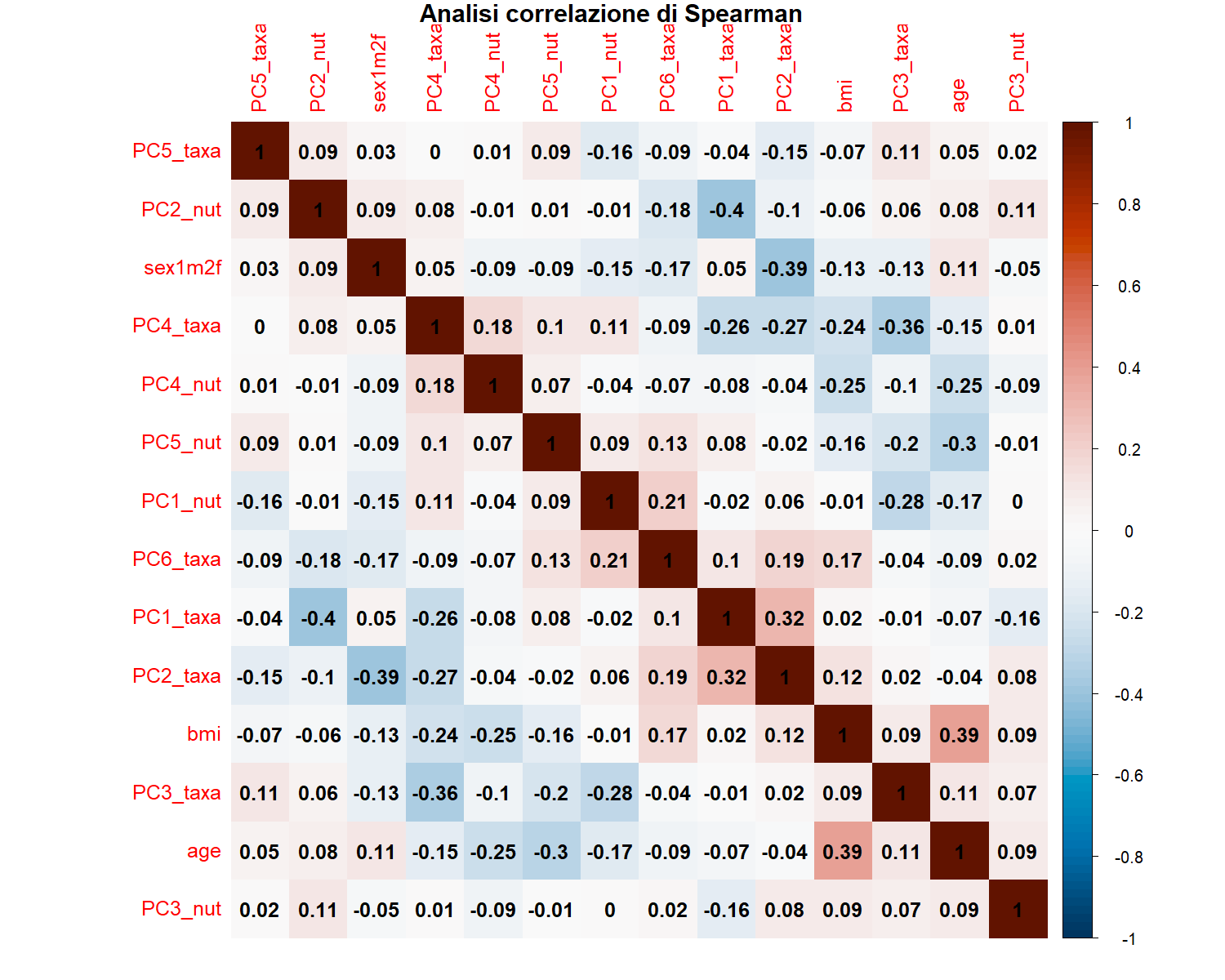
Successivamente, è stata analizzata la distribuzione della y rispetto all’età e il sesso in modo tale da identificare eventuali sottopopolazioni nei dati. Una volta verificata l’assunzione di normalità rispetto alla y e di varianza costante nei gruppi, abbiamo applicato il t-test a due campioni per verificare se la distribuzione della popolazione maschile differisse in modo significativo da quella delle femmine: il test accetta l’ipotesi nulla di uguaglianza, in media, tra le due popolazioni per un alfa pari a 0.05. Mostriamo il grafico della distribuzione della y rispetto al sesso:



Un’altra variabile interessante è quella legata all’età dei pazienti. Una volta discretizzata l’età in 2 principali intervalli e il bmi, sulla base dei quantili, abbiamo applicato il test del Chi-quadro per capire se esiste una dipendenza significativa tra le varie fasce di età e il bmi. I risultati ottenuti fanno emergere che non esiste una stretta dipendenza tra l’età e la variabile y, infatti si accetta l’ipotesi nulla di indipendenza per un alfa pari a 0.05.



Dal grafico superiore, si nota come la fascia di età degli adulti tende ad avere un bmi più alto rispetto ai giovani. Successivamente, abbiamo effettuato l’analisi della correlazione sulla base del coefficiente di Spearman, questo ci permette di non fare assunzioni sulle distribuzioni. Nella figura successiva, 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.



Grazie a questo tipo di misura, possiamo analizzare a priori se esiste una possibile collinearità tra i vari regressori. Da come si evince dalla matrice di correlazione in figura sopra, non esistono correlazioni forti da poter generare un problema di multicollinearità. Per avere la conferma di quanto detto, dopo aver sviluppato i diversi modelli, ci siamo avvalsi di indicatori come il VIF o il TOL.

**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. 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, e il confronto finale, dei diversi modelli, valutando la previsione fatta sul test set.

Modello lineare 1:

Il primo modello scelto si basa sulle prime 6 componenti principali dei batteri, l’età, il sesso e il cluster di appartenenza rispetto alla composizione batterica:

I coefficienti significativi di questo modello sono: B0,B4, B7, B8 per un alfa pari a 0.05. Il test F sul modello è significativo per un alfa pari a 0.05.

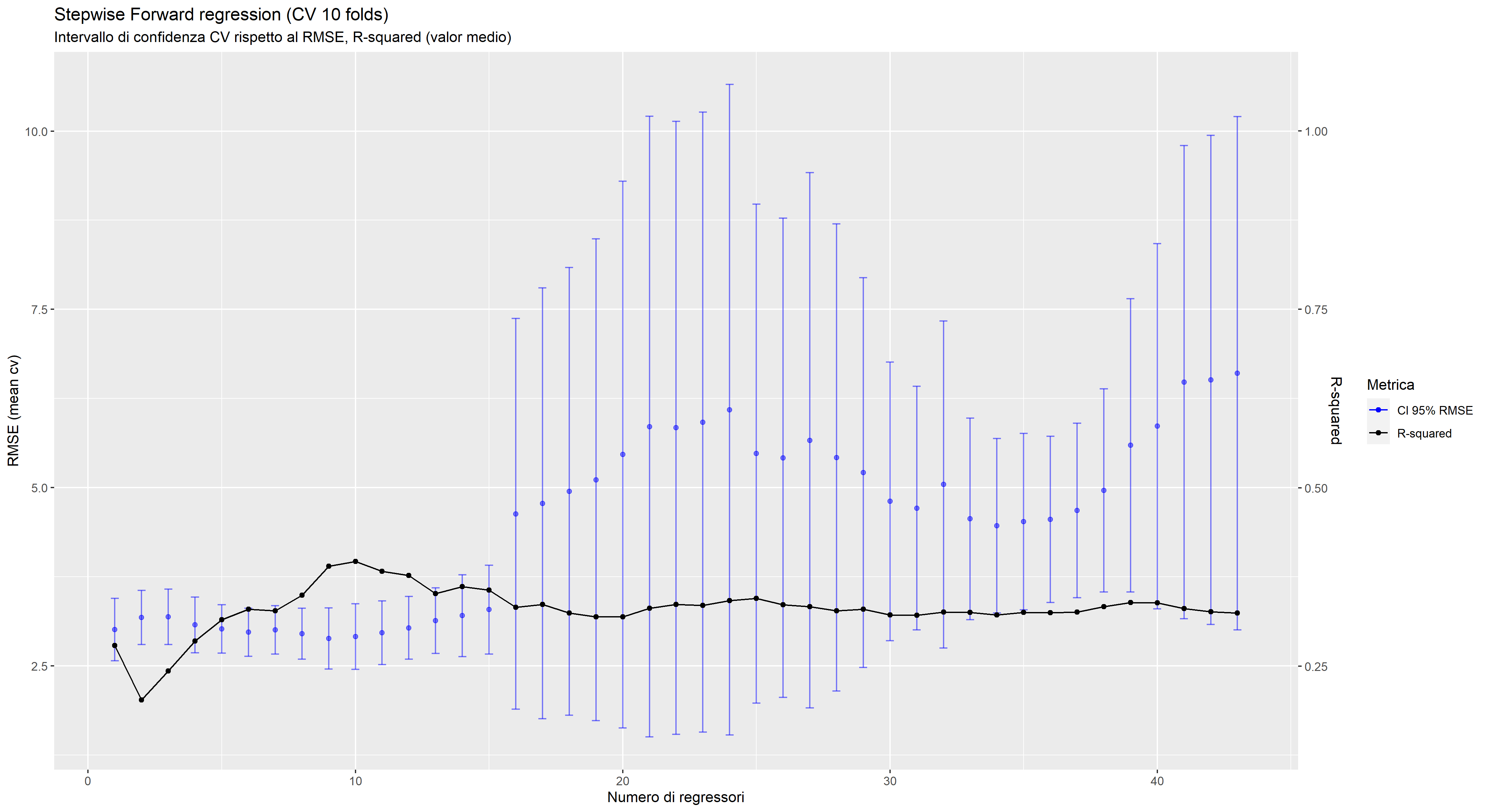
Modello lineare 2:

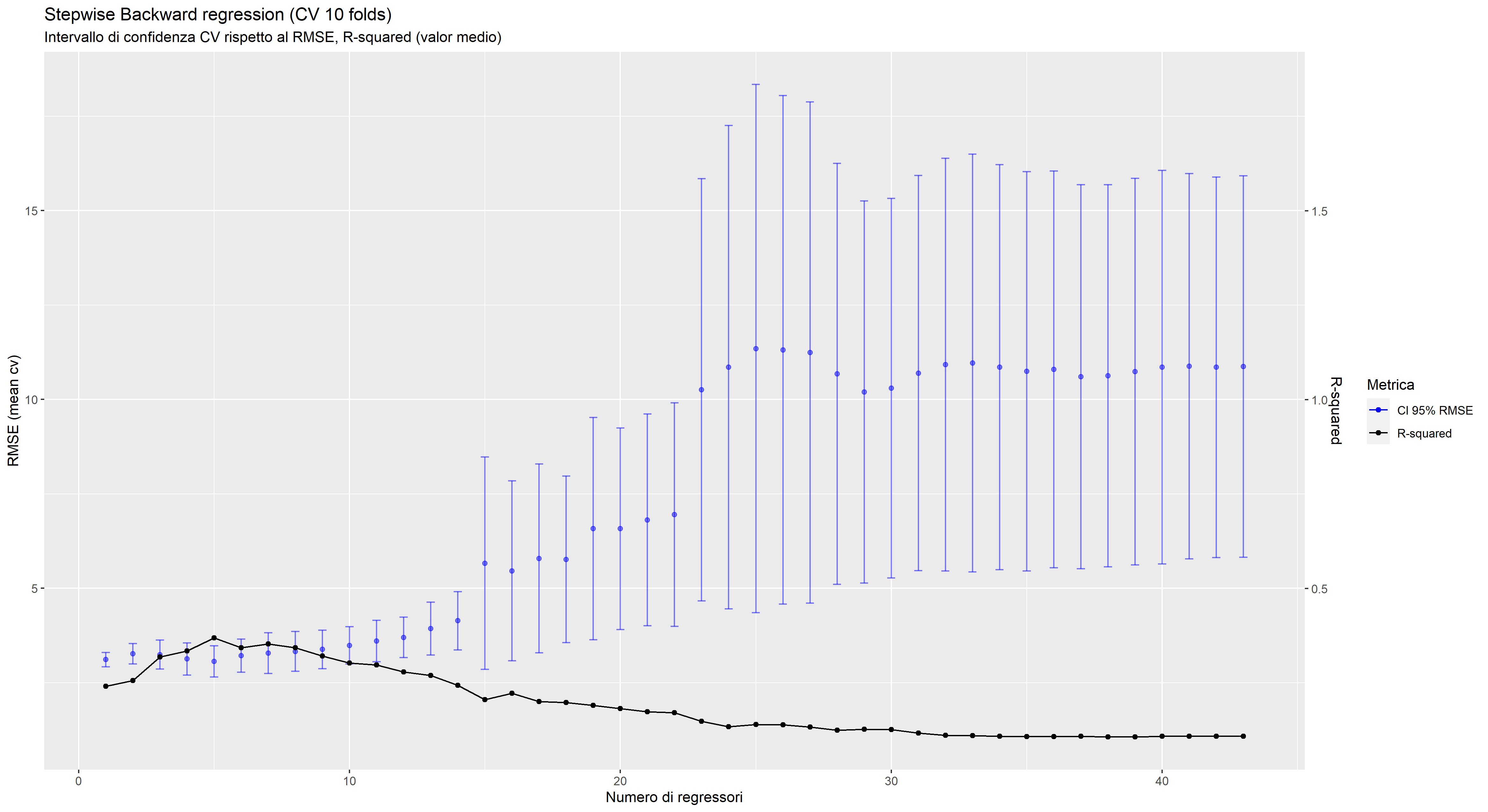
I regressori utilizzati per questo modello sono le prime 5 componenti principali dei nutrienti, l’età, il sesso e il cluster dei batteri:

I coefficienti significativi di questo modello sono: B0,B6, B7 per un alfa pari a 0.05. Il test F sul modello è significativo per un alfa pari a 0.05.

Stepwise regression: Forward vs Backward

Data la numerosità elevata dei regressori p, dove p > n, bisogna avvalersi di diverse tecniche per ridurre la dimensionalità di p. Procedure come best subset selection potrebbero produrre risultati migliori, a discapito del tempo computazionale. In questo caso di regressione, p è talmente elevato che non permette di poter provare tutte le combinazioni. Per tale motivo, è stata utilizzata la procedura stepwise su un dataset aggregato per la 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 con 10 folds CV. Gli intervalli di confidenza al 95%, per la metrica RMSE, sono stati generati sulla base della seguente formula:





Da come si evince dalle due figure, la procedura stepwise forward produce dei risultati migliori rispetto alla backward. Infatti, riesce ad individuare un numero di regressori il cui errore è più basso e un R-squared più alto. Tuttavia, osservando gli intervalli di confidenza per un numero di regressori ridotto, fino a 10, non esiste una differenza significativa. Perciò, è stato scelto il modello con 9 regressori generato dalla procedura stepwise forward avendo un R-squared più elevato:

Modello Lasso1:

Dihydrophylloquinone.Vitamin.K1 0.73598134

Stearic.fatty.acid 0.10482176

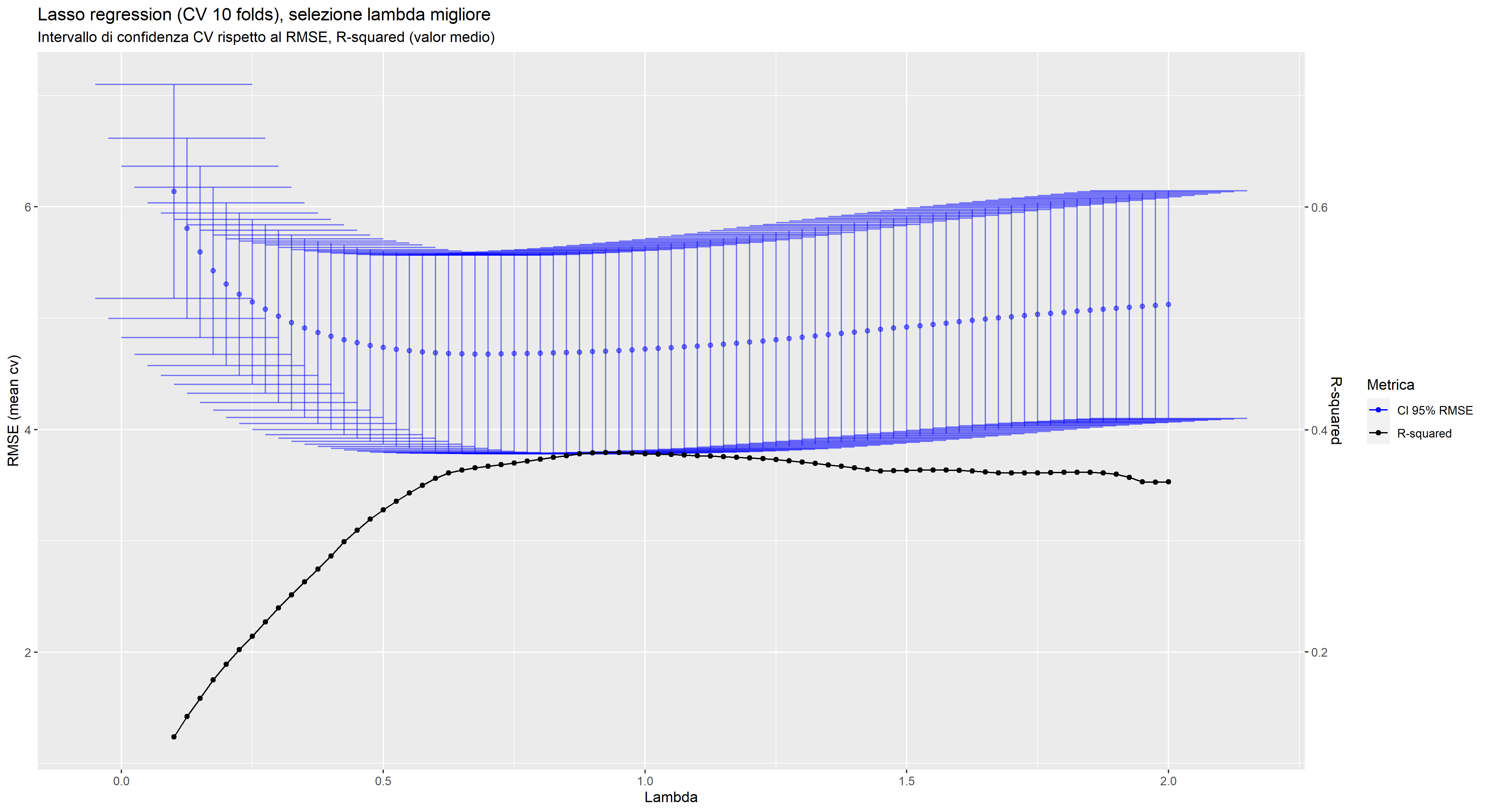
Hydroxyproline 0.62255868

Total.flavonols -0.04738795

Pelargonidin..anthocyanidin -0.08737270

sex1m2f -0.02064929

age 1.97828512

Lambda: 0.675

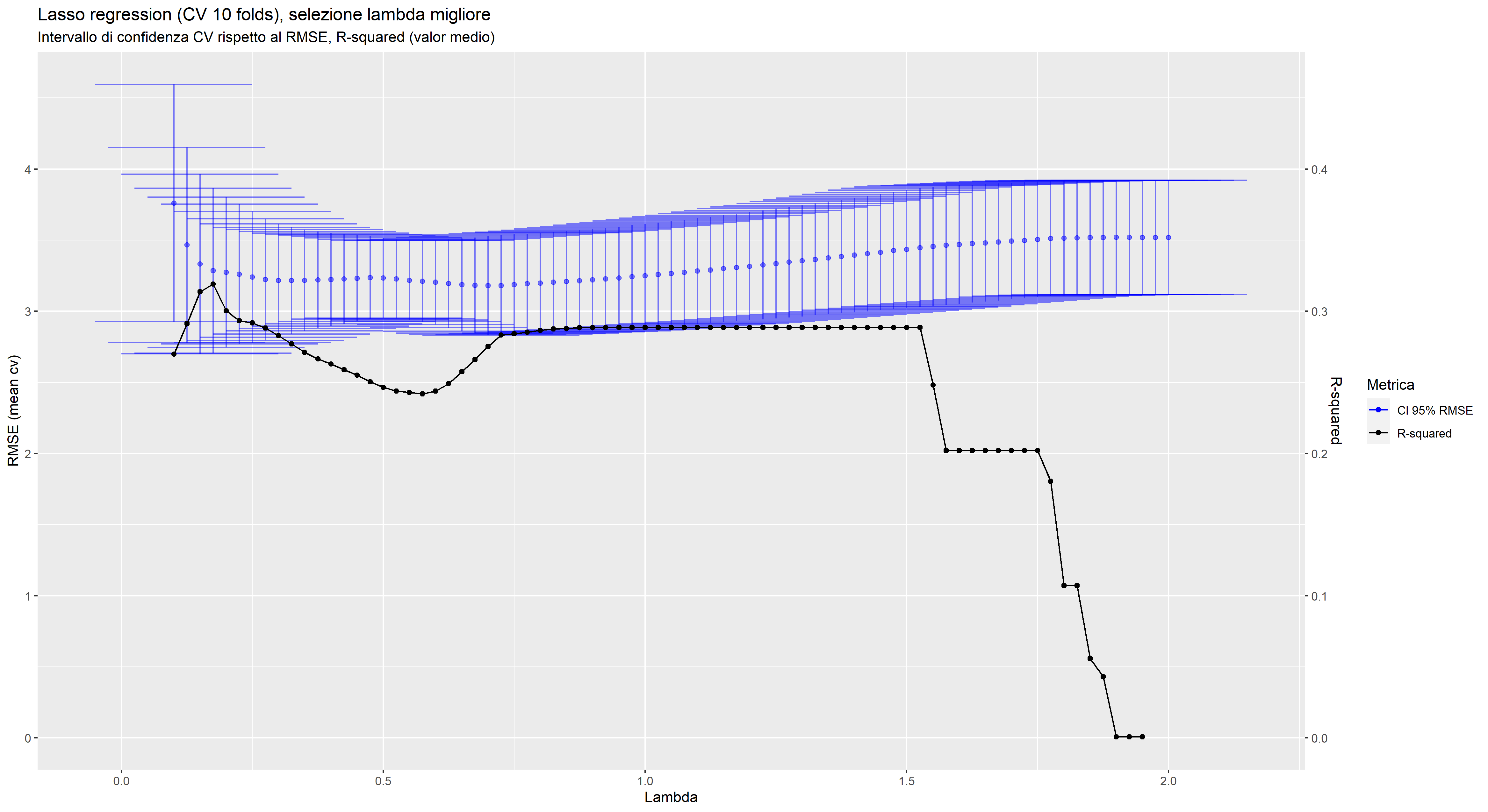
Modello Lasso2:

sex1m2f -0.07527547

age 1.10304431

Veillonellaceae.g\_\_Megasphaera 0.07498628

Lambda: 0.70



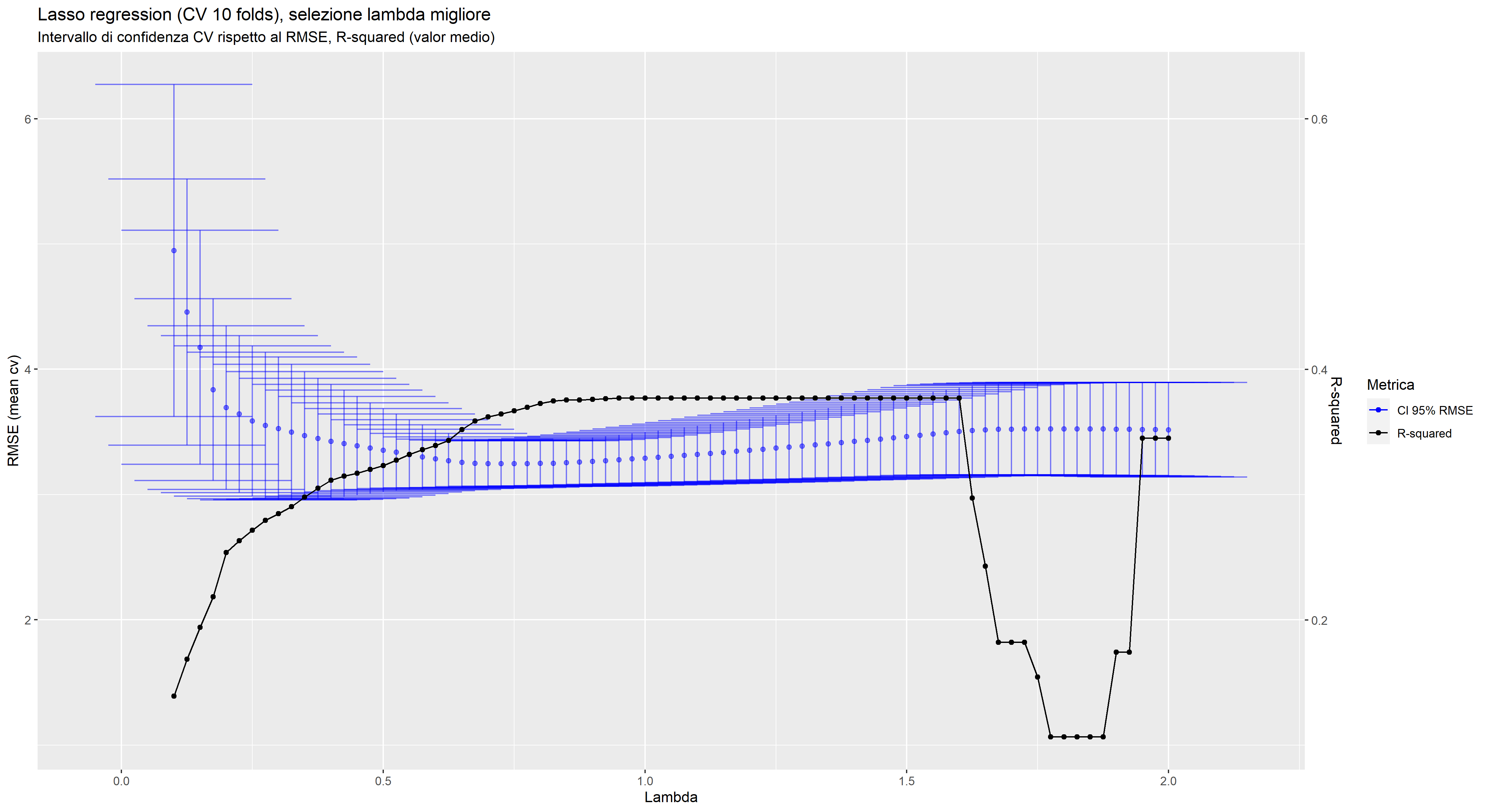
Modello Lasso3:

sex1m2f -0.04623762

age 1.07378549

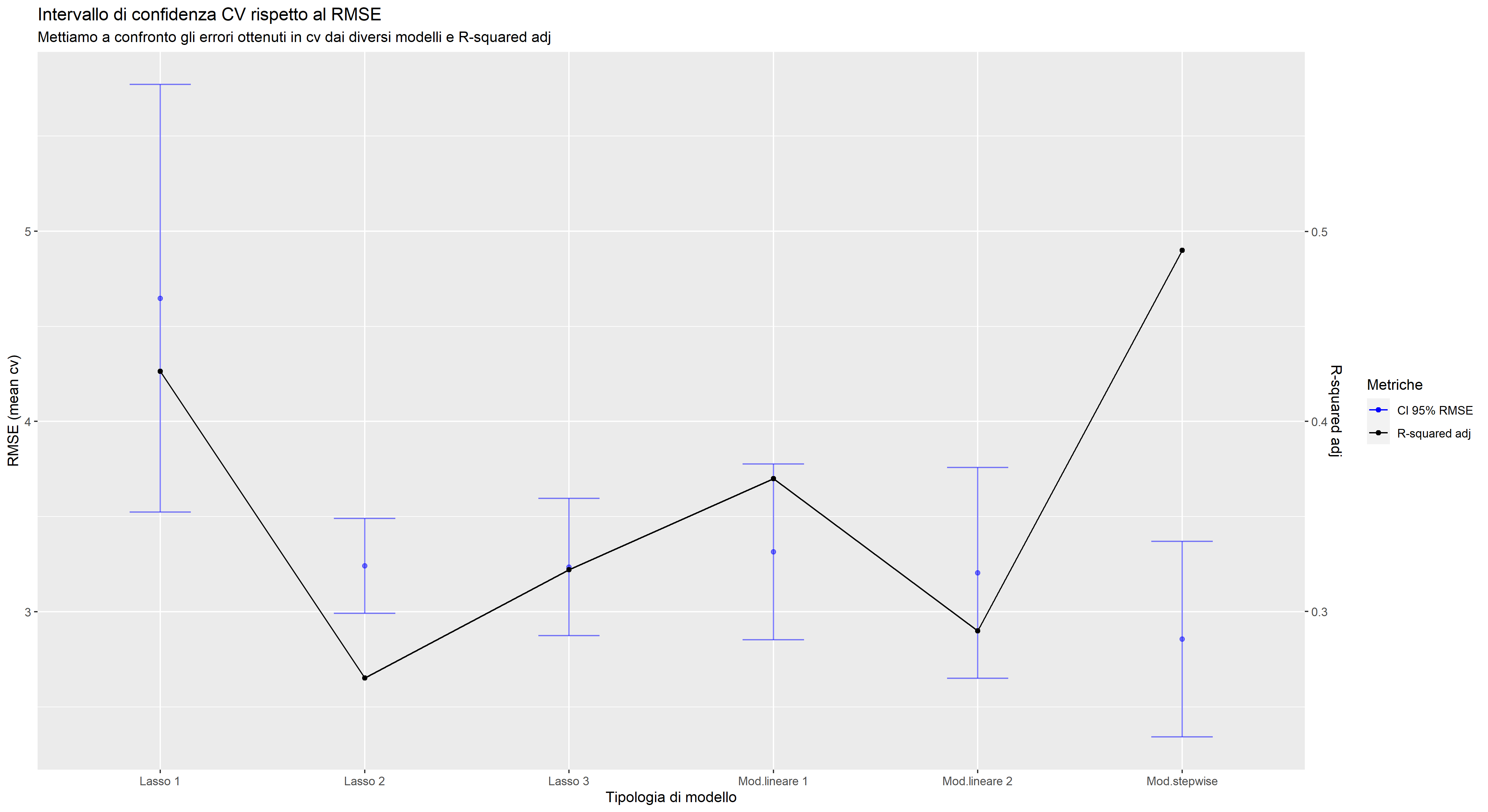
Veillonellaceae.g\_\_Megasphaera 0.04803838

Lambda: 0.725



Risultati dei modelli:

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



Dalla figura superiore, si mettono a confronto i modelli di regressione con stima OLS con quelli Lasso dopo aver identificato gli iper-parametri migliori: nel metodo di stepwise corrisponde al numero di regressori e per i modelli lasso è la scelta di lambda ottimale. Successivamente, per poter confrontare tutti i modelli è stato preso il training set e suddiviso in diversi fold per la CV, questo ci ha permesso di generare degli intervalli di confidenza, sempre sulla base della formula (x), e di identificare il modello migliore. Dal grafico è emerso che non esiste una differenza significativa tra i vari modelli, in termini di RMSE, tranne per il modello stepwise 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 valutati i modelli sulla CV, sono state effettuate le previsioni sul test set.

|  |  |
| --- | --- |
| RISULTATI TEST (RMSE): | RISULTATI TEST (R-squared): |
| LASSO 1: 3.524 | LASSO 1: 0.530 |
| LASSO 2: 3.351 | LASSO 2: 0.493 |
| LASSO 3: 2.921 | LASSO 3: 0.551 |
| Modello lineare 1: 3.221 | Modello lineare 1: 0.534 |
| Modello lineare 2: 3.313 | Modello lineare 2: 0.520 |
| Stepwise regression: 2.734 | Stepwise regression: 0.501 |

Dalla tabella, emerge che il modello stepwise sembrerebbe quello più affidabile in termini di risultati tra la cross validation e il test set, anche se esso è composto da sole 10 osservazioni, e quindi questo può influenzare i risultati in termini di R^2. Infatti, nei restanti modelli, rispetto alla Cross-validation, il coefficiente di determinazione raggiunge valori superiori, facendo emergere una presenza di underfitting nei modelli. Questo significa che tali modelli risultano essere più flessibili rispetto al test set ma più rigidi nel training set. Per questo principio sceglieremo il modello stepwise, essendo quello più parsimonioso in termini di trade off di varianza e distorsione sia per il training set sia per il test set. Questa valutazione è stata fatta rispetto al valore puntuale stimato, sarebbe interessante stimarsi anche l’intervallo di confidenza del coefficiente di determinazione, in modo tale da poter osservare la stima intervallare. Ovviamente, avendo dei valori del R-squared inferiori a 0.6, tutti i modelli non riescono a prevedere, in modo accurato, il BMI. Questo è anche dovuto dal fatto che tra i regressori e la variabile risposta, non esistono forti dipendenze tale per cui si riesce a spiegare bene l’indice di massa corporea.