Analisi sulla crescita dei ragazzi olandesi

Veronica Milazzo

INTRODUZIONE

I dati utilizzati nella seguente analisi provengono dal quarto studio olandese sulla crescita. È uno studio trasversale che misura la crescita e lo sviluppo di 7482 maschi olandesi di età compresa tra 0 e 21 anni. Il dataset è composto da 7294 unità su cui sono state rilevate due variabili quantitative:

- *bmi:* (Body Mass Index) dato biometrico calcolato come rapporto tra peso (espresso in kg) e quadrato dell'altezza (espressa in metri) ed è utilizzato come un indicatore dello stato di peso forma;
- *age:* età dei soggetti espressa in anni.

Dopo un'analisi esplorativa, l'obiettivo è studiare la relazione tra le due variabili considerando l'età come variabile esplicativa e il BMI come variabile dipendente. Inoltre, si vuole prevedere l'andamento del BMI per i ragazzi tra 21 e 30 anni. Il dataset è stato suddiviso in due gruppi di numerosità uguale pari a 3647. Il training set è stato utilizzato per stimare i modelli e il test set è stato utilizzato per calcolare il valore RMSE (Root Mean Square Error).

Analisi esplorativa

Il primo aspetto da considerare è la distribuzione della variabile dipendente. Nella tabella 1 sono presentati i valori di minimo, mediana, media, massimo, asimmetria e curtosi. Sia dal valore dell'indice di asimmetria, che si discosta molto dal valore di riferimento 0, sia dall'istogramma del BMI (grafico 1) notiamo una marcata asimmetria positiva dovuta alla presenza di valori anomali molto alti. Inoltre, il valore dell'indice di curtosi si discosta molto dal valore di riferimento 3, quindi, in questo caso non avrebbe senso ipotizzare una distribuzione normale e adattare un modello di regressione lineare semplice.

Minimo	11.17
Mediana	17.45
Media	18.03
Massimo	35.42
Asimmetria	1.04
Curtosi	4.55

Tabella 1

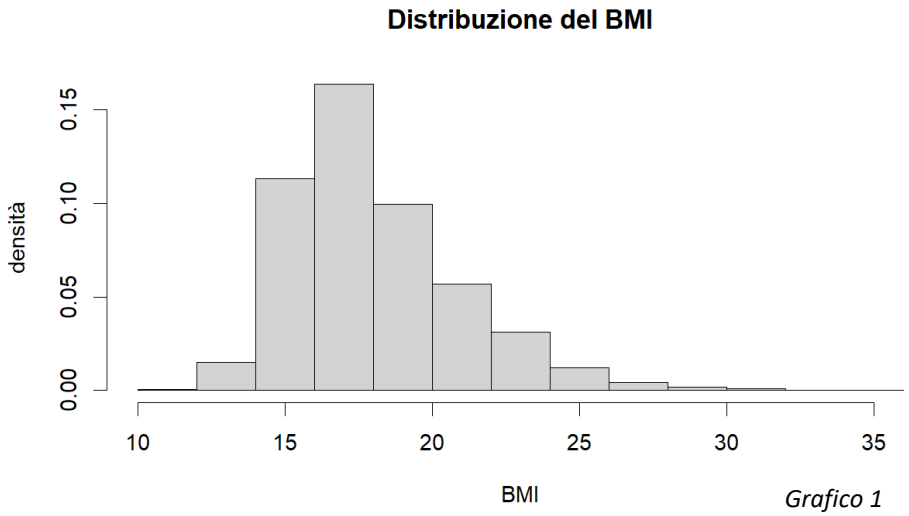

Grafico 1

Per quanto riguarda la distribuzione dell'età, i valori di sintesi sono presentati nella tabella 2. L'indice di asimmetria è molto vicino a 0. Questo non vuol dire che la distribuzione è simmetrica perché dall'istogramma dell'età (grafico 2) si evince che la distribuzione è bimodale con un picco tra 0 e 2 anni e un altro picco tra 12 e 16 anni.

Minimo	0.03
Mediana	10.45
Media	9.29
Massimo	21.70
Asimmetria	-0.03
Curtosi	1.53

Tabella 2

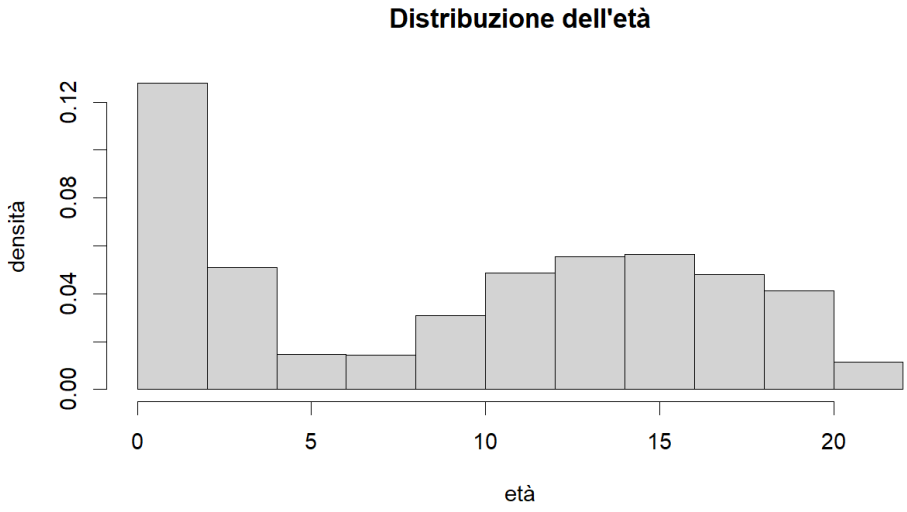

Grafico 2

Il grafico 3 rappresenta la relazione che sussiste tra BMI ed età. È una relazione non lineare, poiché il BMI cresce rapidamente nel primo anno di età per poi subire un lieve calo fino ai 7 anni e una successiva ricrescita graduale fino ai 22 anni.

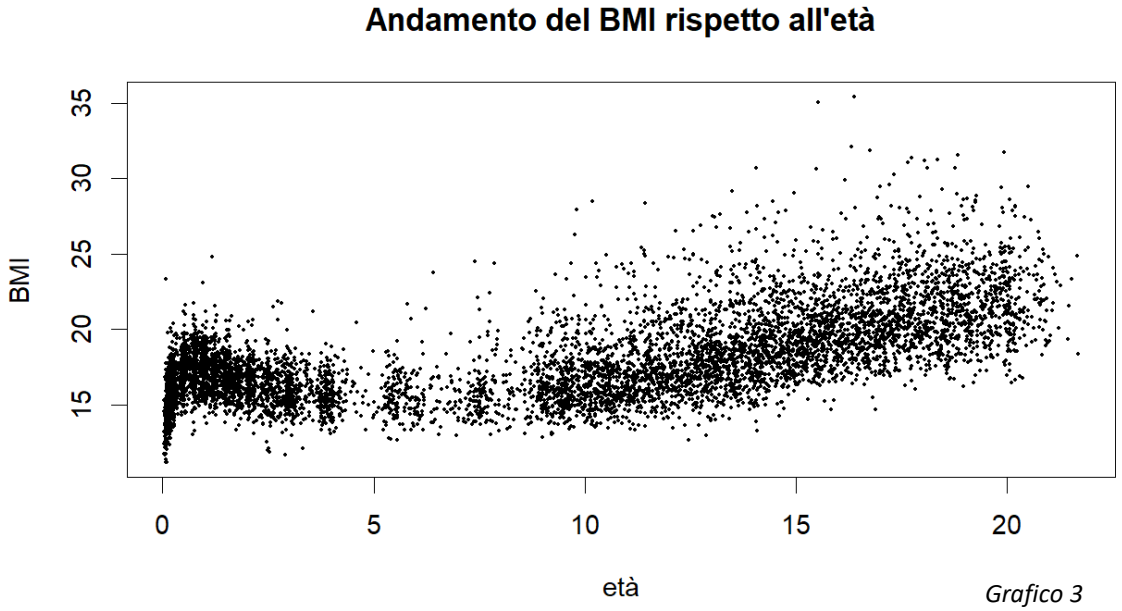

Grafico 3

METODI

Per rispondere all'obiettivo dell'analisi, i modelli non lineari da poter considerare, basandosi su quanto visto nella precedente analisi esplorativa, sono: regressione polinomiale, regressione segmentata, regressione spline e regressione quantile con termini non lineari.

Regressione polinomiale

La regressione polinomiale si basa sul presupposto che molti andamenti non lineari possono essere approssimati tramite funzioni polinomiali di ordine p:

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$$

Le funzioni polinomiali sono molto flessibili e quindi aumentando l'ordine è possibile trovare funzioni che si adattano ai dati osservati in modo teoricamente perfetto, ma un modello che si adatta perfettamente ai dati non è affidabile nel caso in cui vogliamo fare delle previsioni.

Regressione segmentata

La regressione segmentata è un approccio in cui la variabile indipendente è suddivisa in intervalli e per ognuno di essi è adattata una regressione lineare. Se l'effetto della variabile X mostra un andamento che non può essere colto con un semplice modello di regressione lineare, si assume che: $E[Y|X] = \beta_0 + \beta_1 X + \delta(X - \psi)_+$ dove ψ = punto di svolta
β_0 = intercetta
β_1 = pendenza a sinistra del punto di svolta
$\beta_1 + \delta$ = pendenza a destra del punto di svolta

I punti di svolta possono essere noti oppure devono essere stimati con algoritmi appropriati.

Regressione spline

La regressione spline è un metodo di complessità maggiore rispetto alla regressione polinomiale, ma di adattamento migliore e si basa sulla regressione polinomiale a tratti. Le funzioni spline posso assumere virtualmente qualsiasi andamento:

$$\eta(E[Y|X]) = \beta_0 + s(X)$$

Un problema è che bisogna fissare sia il numero che il posizionamento dei nodi. Solitamente quattro o cinque nodi risultano appropriati e nella maggior parte dei casi possono essere posizionati in modo automatico in base ai percentili della variabile esplicativa. Le spline cubiche sono in genere instabili nelle code, cioè prima del primo nodo e dopo l'ultimo, e per ovviare a questo problema si pone solitamente un vincolo di linearità sulle code.

Regressione quantile

La regressione quantile consente di stimare l'intera distribuzione dei quantili condizionati alla variabile risposta, così da poter studiare l'influenza della variabile esplicativa sulla forma della distribuzione di Y. Quando la relazione è non lineare le curve quantili possono essere stimate tramite spline:

$$Q_Y(\tau|X) = \beta_0 + s(X) \quad \text{dove} \quad \tau \in (0,1)$$

RISULTATI

La tabella 3 mostra il valore dell'AIC e del RMSE calcolati per i vari modelli considerati. Come mostrato nel grafico 3 sussiste una relazione non lineare tra le variabili e dal confronto emerge che un modello lineare non risulta adeguato poiché entrambi i valori sono i più alti. Il modello che si adatta meglio è il modello B-spline poiché presenta l'AIC più basso e anche l'errore minore. Anche il modello segmentato si adatta bene in termini di AIC, ma presenta un valore di RMSE leggermente più alto del modello B-spline. Graficamente tutti i modelli non lineari considerati interpolano bene i dati osservati (grafico 4).

	GLM	POLINOMIALE	SEGMENTATO	B-SPLINE	N-SPLINE
AIC	16012.61	15805.24	15278.54	15272.80	15544.55
RMSE	2.216	2.030	2.033	2.027	2.076

Tabella 3

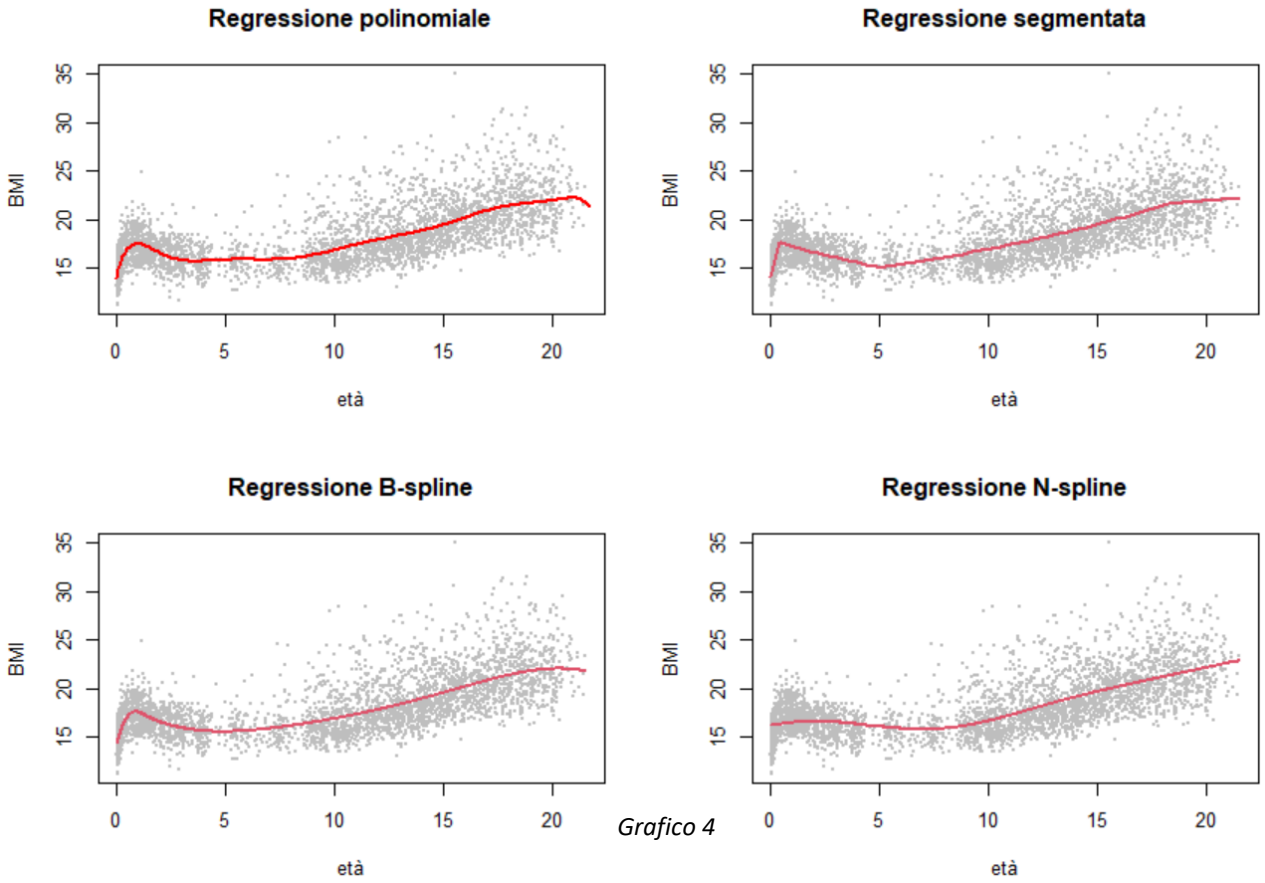

Grafico 4

In termini di capacità predittiva, guardando le previsioni fino a 30 anni (grafico 5), si preferisce il modello segmentato che mostra un aumento contenuto del BMI all'aumentare dell'età. Inoltre, il modello polinomiale e il modello B-spline sono da escludere poiché non è plausibile che il BMI diminuisca all'aumentare dell'età fino a raggiungere il valore zero o addirittura assuma valori inferiori a zero.

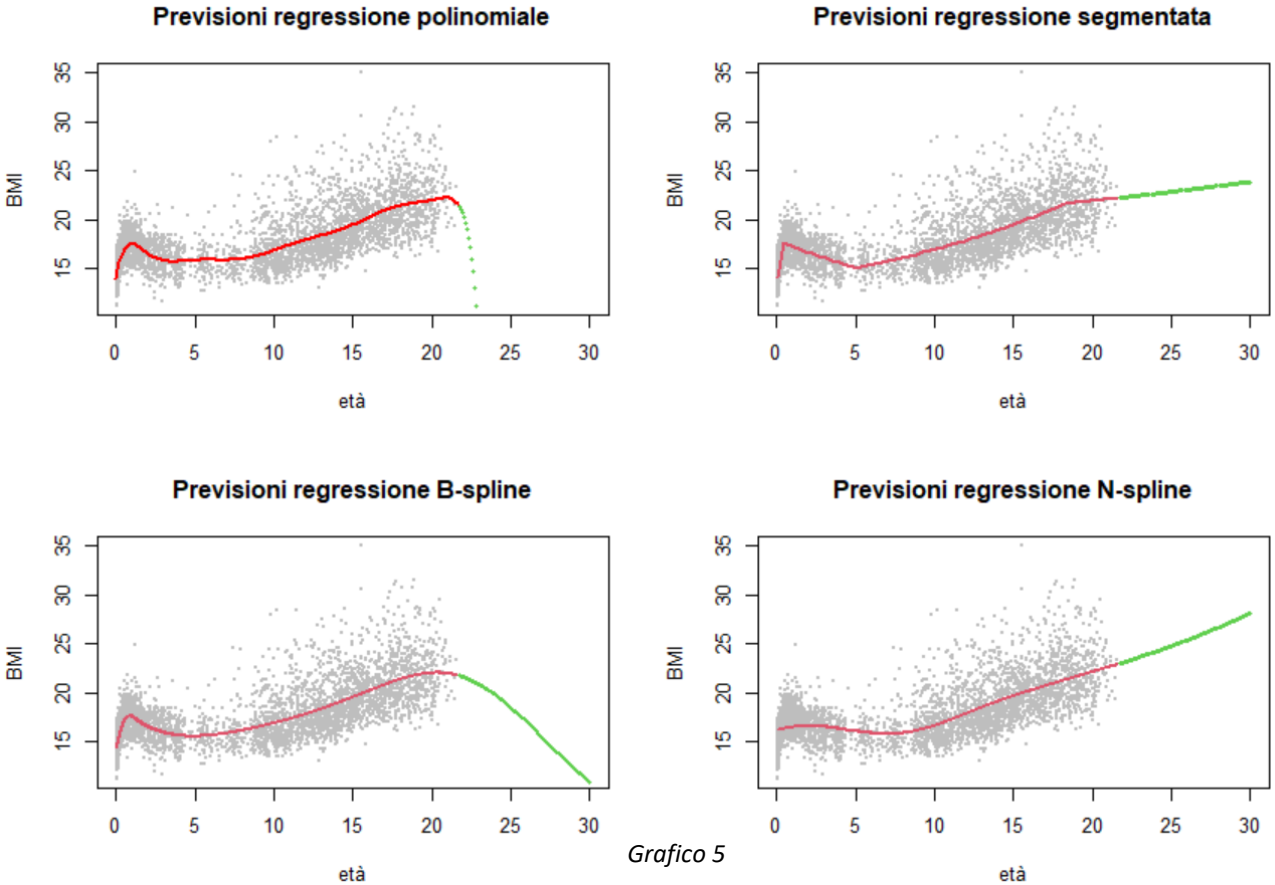

Grafico 5

Tra tutti, il modello da preferire è quello segmentato in cui otteniamo tre punti di svolta in corrispondenza di 0.404, 5.052, 18.210 anni e le rispettive pendenze sono mostrate nella tabella 4.

	STIMA	IC(95%).INF	IC(95%).SUP
SLOPE 1	-0.03689	-0.04621	-0.02756
SLOPE 2	0.00209	0.00169	0.002482
SLOPE 3	-0.00153	-0.00162	-0.001447
SLOPE 4	-0.00037	-0.00111	0.00036

Tabella 4

Si ipotizza che la variabile risposta si distribuisce come una Gamma e si considera il link canonico che è la funzione inversa. L'interpretazione dei parametri è la seguente:

- la prima pendenza è la più elevata in valore assoluto e il segno negativo suggerisce che al crescere dell'età si riduce l'inverso del livello medio del BMI; quindi, tra 0 e 5 mesi circa il livello medio del BMI aumenta al crescere dell'età;
- la seconda pendenza suggerisce che, tra 5 mesi e 5 anni circa, al crescere dell'età il livello medio del BMI si riduce;
- la terza pendenza suggerisce che, tra 5 e 18 anni circa, al crescere dell'età il livello medio del BMI aumenta;
- la quarta pendenza non risulta significativamente diversa da zero; quindi, dopo i 18 anni non c'è nessun effetto dell'età sul livello medio del BMI.

CONCLUSIONI

La diagnosi di obesità si basa sull'indice di massa corporea (BMI), ma comunque non è possibile fissare dei range di valori rigidi dal momento che sensibili variazioni del peso durante la crescita possono essere fisiologiche. In particolare, tra 0 e 2 anni il calcolo del BMI non è indicato mentre tra i 2 e i 18 anni vengono fornite delle tabelle di riferimento dall'Organizzazione Mondiale della Sanità (OMS). Più precisamente, vengono fornite due tabelle: una relativa alla fascia di età 2-5 anni e un'altra relativa alla fascia di età 5-18 anni. Per individui con età superiore ai 18 anni i valori di riferimento sono quelli degli adulti:

- grave magrezza: BMI < 16;
- sottopeso: BMI compreso tra 16 e 18.5;
- normopeso: BMI compreso tra 18.5 e 25;
- sovrappeso: BMI compreso tra 25 e 30;
- obesità: BMI compreso tra 30 e 40;
- grave obesità: BMI > 40.

Dall'analisi emerge che due dei tre punti di svolta trovati (5.052 e 18.210) si avvicinano proprio ai valori (5 e 18) che individuano le classi di età per cui è necessario cambiare i valori di riferimento. Il BMI dei bambini viene considerato con valori normali se si colloca tra il 5° e il 95° percentile, più nello specifico, si definisce:

- sottopeso: BMI < 5° percentile;
- normopeso: BMI compreso tra 5° e 85° percentile;
- sovrappeso: BMI compreso tra 85° e 95° percentile;
- obesità: BMI > 95° percentile.

Considerato quanto appena detto, per seguire la crescita dei bambini nel tempo è utile adattare un modello di regressione quantile per il 5°, 85° e 95° percentile. Le curve quantili sono mostrate nel grafico 6. Come ci aspettiamo una minima parte dei bambini sono sottopeso, la maggior parte dei bambini sono normopeso e una buona parte, soprattutto dai 10 anni in poi, sono sovrappeso o obesi.

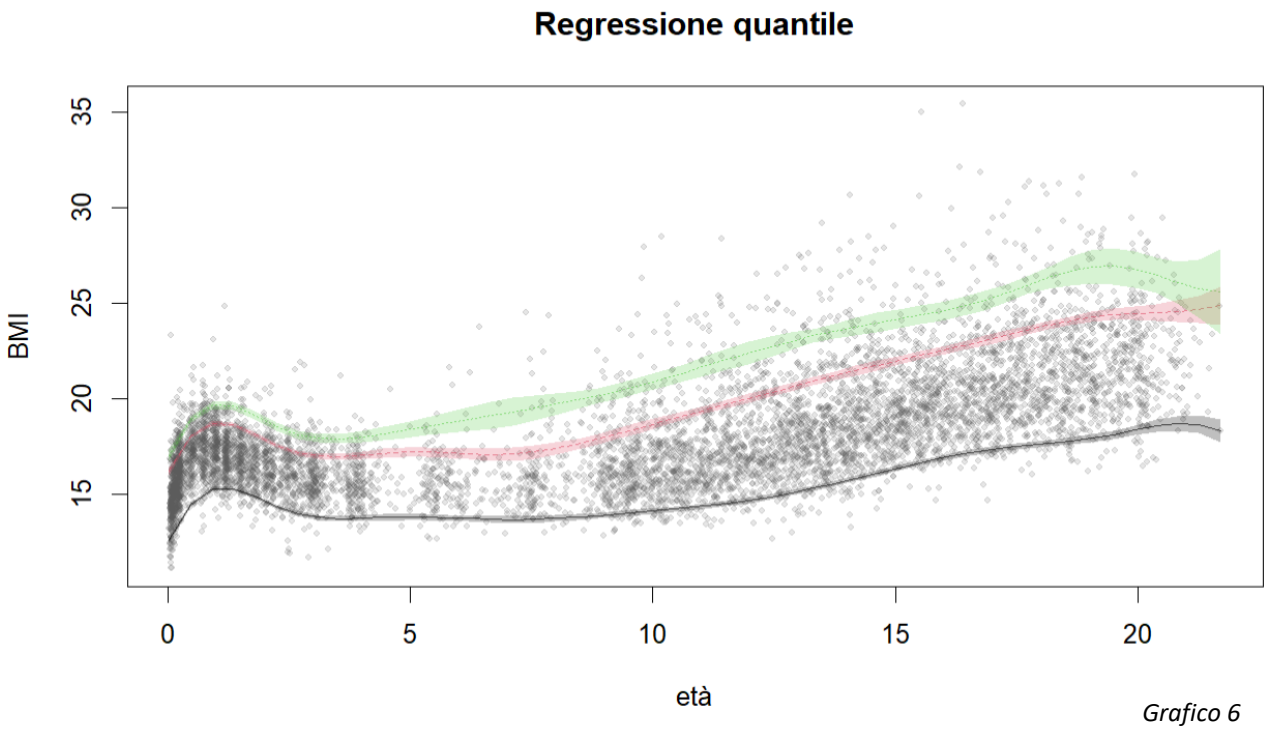

Grafico 6