REPORT di Statistica medica

Quantità di calorie bruciate in funzione di massa corporea e livello di attività fisica in soggetti esaminati sotto sforzo: due modelli di regressione a confronto

Corso: Statistica medica
Docente: Prof.ssa
Studente: Marco Lazzarini

1. MOTIVAZIONI E OBIETTIVI

I dati fanno riferimento a una serie di misurazioni della quantità di calore prodotto da 24 individui testati sotto sforzo su una cyclette, nell'ambito di uno studio presentato alla Royal Society of London nel 1913 dal Prof. J. S. Macdonald.

Successivamente, Glazebrook & Dye, in una nota allegata allo studio, elaborarono un modello di regressione non lineare multipla con il quale cercarono di spiegare in termini più generali i risultati delle analisi del Prof. J. S. Macdonald: il modello è della forma: $E(ho) = \beta_0 + \beta_1 bm + (wl / (\beta_3 + \beta_4 bm))$, in cui le calorie prodotte (*ho*) sono la variabile risposta ed il peso corporeo (*bm*) ed il livello di lavoro (*wl*), ovvero il numero di calorie orarie bruciate durante l'attività, i regressori. β_0, β_1, β_3 e β_4 sono costanti reali ricavate per via grafica[1].
Nel 1918, sulla base di tale lavoro, Major Greenwood, statistico ed epidemiologo, decise di avanzare un ulteriore modello di regressione multipla, stavolta di tipo lineare, del tipo: $E(ho) = \beta_0 + \beta_1 bm + \beta_2 wl$, convenendo che, nella sua forma, potesse semplificare la generalizzazione operata da Glazebrook & Dye.

L'obiettivo di questo report è, in seguito ad un'analisi esplorativa dei dati sui soggetti esaminati, studiare nel dettaglio i modelli proposti da Greenwood e Glazebrook & Dye, valutandoli singolarmente e, successivamente, paragonandoli tra loro in termini di validità e di bontà d'adattamento, stabilendo quale tra i due sia preferibile nel descrivere la quantità di calorie bruciate in funzione di peso e lavoro. I modelli saranno infine utilizzati per effettuare delle previsioni.

Le variabili che fanno riferimento al dataset sono:

- **bm** (Body Mass): Variabile quantitativa su scala continua: rappresenta una misura antropometrica, espressa in chilogrammi (*kg*), del peso corporeo del soggetto esaminato.
- **wl** (Work Level): Variabile quantitativa su scala continua: indica la quantità di calorie per ora (*cal/h*) prodotte dal soggetto esaminato sotto sforzo.
- **ho** (Heat Output): Variabile quantitativa su scala discreta: indica, in calorie (*cal*), la quantità di calorie bruciate dal soggetto esaminato sotto sforzo. Nello studio costituisce la variabile d'interesse.

2. ANALISI UNIVARIATE

	Min.	1° Quartile	Media	Mediana	3° Quartile	Max.	Dev. standard	Scarto Interq.	Coeff. di curtosi	Coeff. di variazione
bm	43.70	54.60	57.54	58.80	61.90	66.70	6.590	7.3	3.21	11.45
wl	13.00	19.00	34.04	38.75	43.00	56.00	16.361	24.0	1.52	48.06
ho	160.0	205.0	260.0	272.0	318.8	352.0	65.902	113.8	1.53	25.34

Tabella 1: Principali statistiche di sintesi delle variabili *bm*, *wl* e *ho*

Dalle statistiche di sintesi emerge un moderato scostamento tra media e mediana nella distribuzione della variabile *bm* (57.54 v. 58.80), iò è conseguenza di un'asimmetria negativa in distribuzione confermata da un indice di asimmetria standardizzato di Pearson $\gamma = -0.819$; la deviazione standard risulta pari a 6.590.

[1] *M. Greenwood (1918). "On the Efficiency of Muscular Work," Proceedings of the Royal Society of London, Series B, Containing Papers of a Biological Character, p.199.*

Nella variabile *wl*, la mediana risulta maggiore di circa 4.71 rispetto alla media (38.75 v. 34.04), la deviazione standard è pari a 16.361, con un indice di asimmetria standardizzato di Pearson γ = -0.278.

Analogamente, nella variabile *ho* gli indici di posizione risultano uno maggiore dell'altro (272.0 v. 260.0), con una deviazione standard pari a 65.902 ed un indice di asimmetria standardizzato di Pearson γ = -0.269.

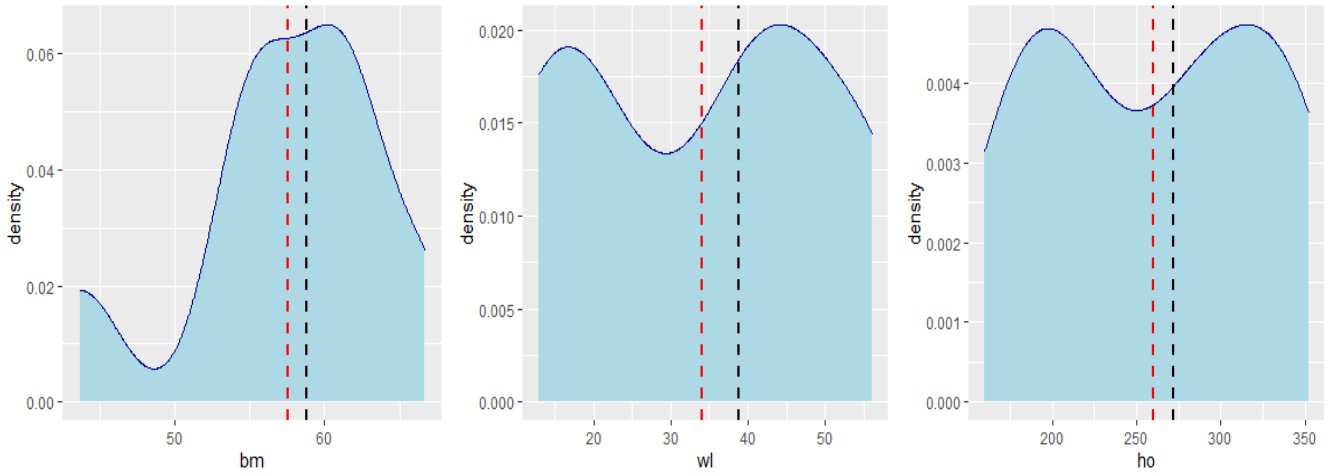


Figure 1, 2 e 3: Grafici delle densità delle distribuzioni di *bm*, *wl* e *ho* (la linea rossa indica la media, quella nera indica la mediana)

I boxplot delle distribuzioni delle tre variabili (figure 4, 5 e 6) non evidenziano la presenza di valori anomali.

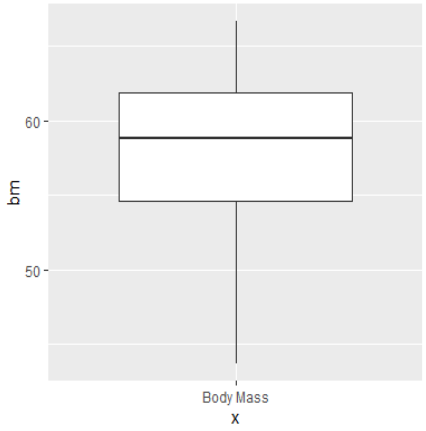


Figura 4: Boxplot (*bm*)

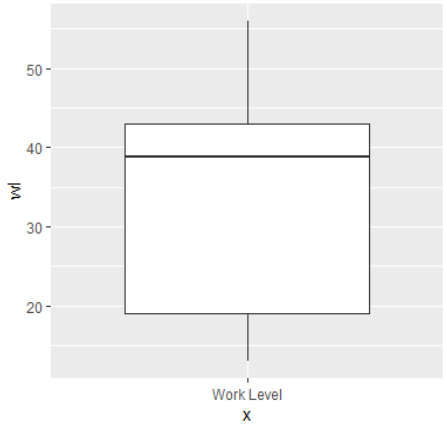


Figura 5: Boxplot (*wl*)

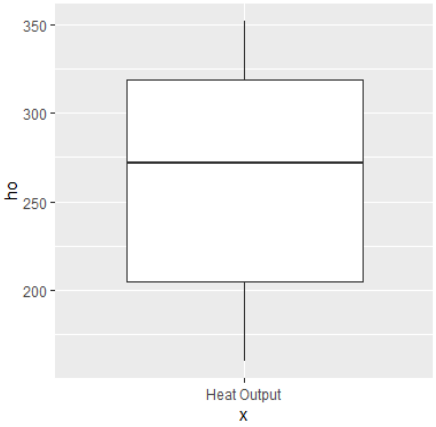


Figura 6: Boxplot (*ho*)

E' stato eseguito il diagramma quantile contro quantile della variabile *bm* (figura 7), il quale ha evidenziato per via grafica un allontanamento dalla retta di riferimento, rappresentante i quantili di una distribuzione normale standard; tale scostamento è stato confermato dal test di Shapiro-Wilk: quest'ultimo, con un valore osservato pari a t^{oss} = 0.88005 ed un **p-value p < 0.01**, risulta significativo e porta al rifiuto dell'ipotesi di normalità per *bm* a qualsiasi livello di α usuale. Similmente, è stato eseguito il diagramma quantile contro quantile per la variabile *wl* (figura 8), anche qui il test di Shapiro-Wilk è risultato significativo contro l'ipotesi di normalità a qualsiasi livello di α usuale (t^{oss} = 0.86019, **p-value p < 0.01**).

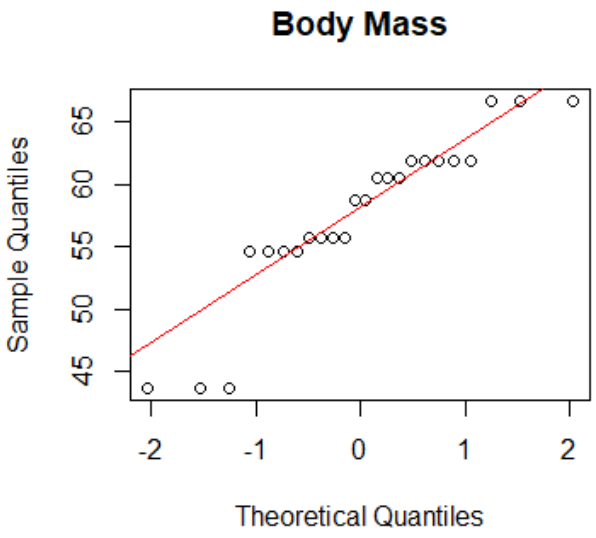


Figura 7: Diagramma quantile contro quantile (*bm*)

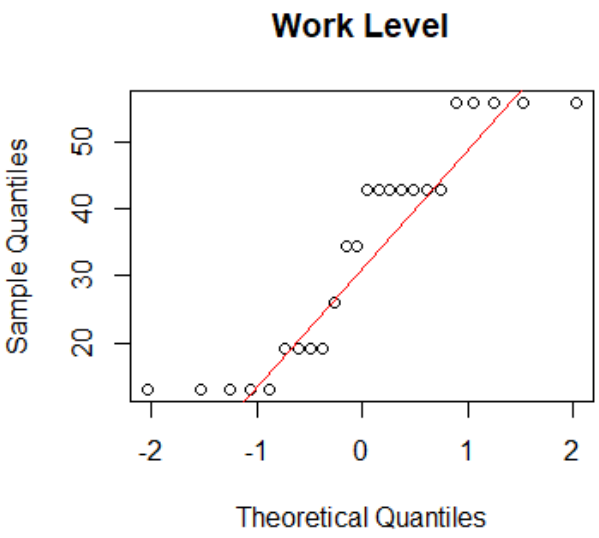


Figura 8: Diagramma quantile contro quantile (*wl*)

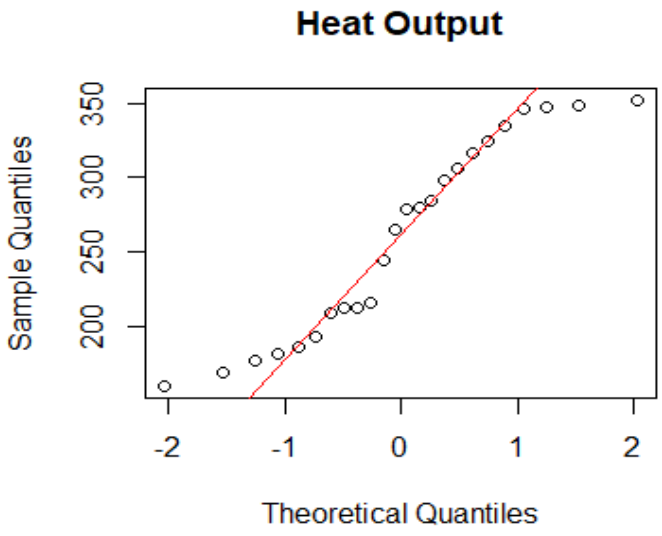


Figura 9: Diagramma quantile contro quantile (*ho*)

Per la variabile *ho* il test di Shapiro-Wilk, con t^{oss} = 0.91176 ed un **p-value p = 0.038**, porta al rifiuto dell'ipotesi di normalità ad un livello di α pari al 5%, mentre ad un'accettazione della stessa ad un livello di α pari all'1%.

Considerate le asimmetrie di tipo negativo, in particolare nella variabile *bm*, sono state considerate delle trasformate di tipo quadratico delle variabili, le quali tuttavia, non avendo portato miglioramenti ai fini dell'adattamento all'ipotesi di normalità e anzi rivelandosi addirittura controproducenti relativamente allo scopo in taluni casi, sono state scartate.

Per approfondire meglio l'ipotesi di adattamento alla normalità, considerato l'alto grado di variabilità di forma tra le variabili, è stato eseguito un test di Lilliefors: questo test rappresenta una variante del test non parametrico di Kolmogorov-Smirnov, il quale verifica l'ipotesi che una variabile segua una determinata distribuzione (in tal caso, la distribuzione gaussiana).

Il test di Lilliefors, con un valore osservato t^{oss} = 0.203 e un **p-value p = 0.0119**, porta al rifiuto dell'ipotesi di normalità per la variabile *bm* ad un livello di α pari al 5%, mentre ad un'accettazione della stessa ad un livello di α pari all'1%. Con un valore osservato t^{oss} = 0.165 e un **p-value p = 0.0913**, l'ipotesi di normalità per la variabile *ho* viene accettata ad un livello di significatività α pari

sia all'1% che al 5%, rifiutando solo per α = 10%. Con un valore osservato t^{oss} = 0.208 ed un **p-value p < 0.01**, l'ipotesi di normalità per la variabile *wl* viene rifiutata ad ogni livello di α usuale.

Il test di Anscombe-Glynn, il quale opera sotto l'ipotesi nulla che il coefficiente di curtosi sia uguale a 3 nel campione, risulta significativo per la variabile *ho* (kurt = 1.5301, t^{oss} = -3.1446, **p-value p < 0.01**).

3. ANALISI BIVARIATE

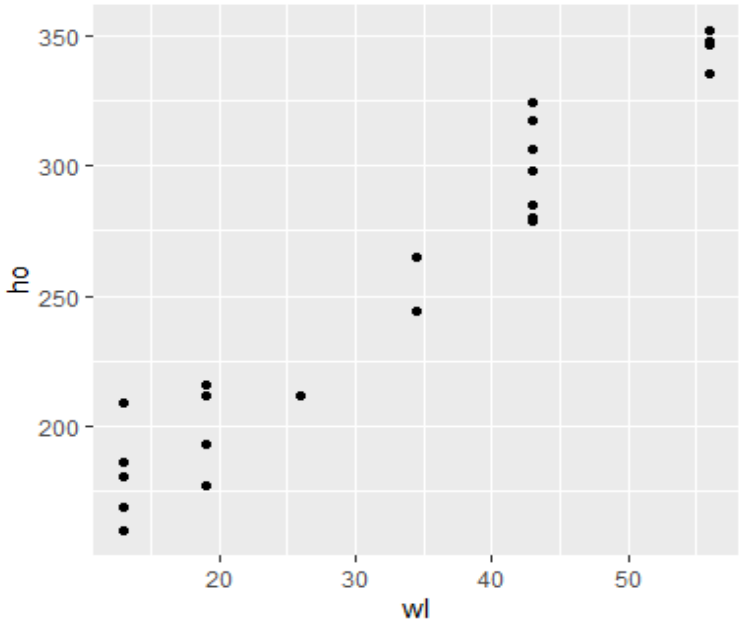


Figura 10: Diagramma di dispersione (*wl* e *ho)*

Le principali analisi bivariate hanno indagato le relazioni tra la variabile d'interesse e le altre variabili. Il diagramma di dispersione di *ho* e *wl* (figura 10) suggerisce una netta relazione di tipo lineare positiva tra quantità di calorie bruciate e livello di lavoro svolto, ciò viene confermato anche da un coefficiente di correlazione lineare tra le due variabili **ρ = 0.973**.

I diagrammi di dispersione delle variabili *bm* e *ho (figura 11)* e di *bm* e *wl* (figura 12) non hanno evidenziato a livello grafico la presenza di particolari relazioni o di andamenti sistematici.

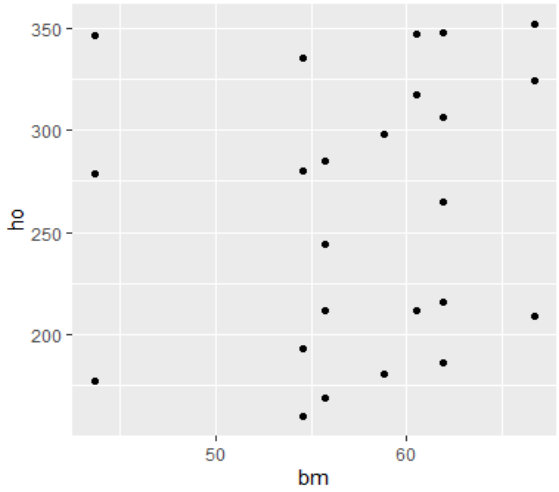


Figura 11: Diagramma di dispersione (*bm* e *ho*)

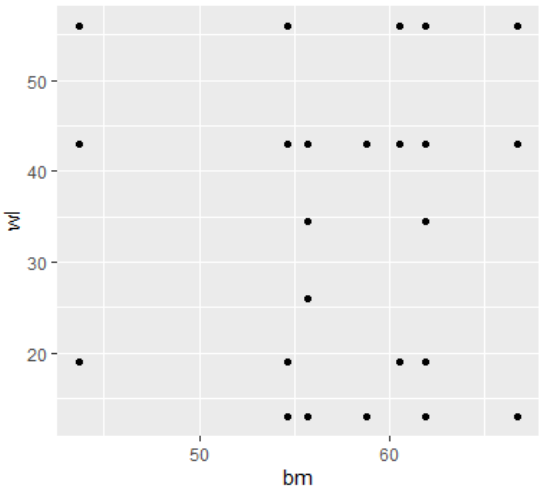


Figura 12: Diagramma di dispersione (*bm* e *wl*)

Si è dunque proceduto calcolando il grado della correlazione delle variabili sia per via parametrica mediante metodo di Pearson (tabella 2) che per via robusta mediante metodo di Spearman (tabella 3), considerato che per *bm* e *wl* l'ipotesi di normalità viene rifiutata ad ogni livello di α usuale mentre

il p-value del test di Shapiro-Wilk per la variabile *ho* risulta molto vicino al livello di significatività fissato (**p = 0.04**, α = 0.05).

	bm	wl	ho
bm	ρ = 1.000	ρ = -0.027	ρ = 0.143
wl	ρ = -0.027	ρ = 1.000	ρ = 0.973
ho	ρ = 0.143	ρ = 0.973	ρ = 1.000

Tabella 2: Coefficienti di correlazione di Pearson

	bm	wl	ho
bm	ρ_s = 1.000	ρ_s = -0.004	ρ_s = 0.261
wl	ρ_s = -0.004	ρ_s = 1.000	ρ_s = 0.957
ho	ρ_s = 0.261	ρ_s = 0.957	ρ_s = 1.000

Tabella 3: Coefficienti di correlazione per ranghi di Spearman

Il coefficiente di correlazione lineare di Pearson e quello per ranghi di Spearman calcolati per le variabili *wl* e *ho* indicano una quasi perfetta associazione positiva (**ρ = 0.973** e **ρ_s = 0.957** rispettivamente) tra queste, il test parametrico di correlazione basato sul metodo di Pearson eseguito sulle due variabili, con un valore osservato t^{oss} = 19.95 ed un **p-value p < 0.01**, rifiuta l'ipotesi che le due variabili siano linearmente incorrelate ad ogni livello di α usuale. Analogamente, sempre per le variabili *wl* e *ho*, il test robusto di correlazione basato sul metodo di Spearman rifiuta l'ipotesi che le due variabili siano incorrelate ad ogni livello di α usuale (t^{oss} = 98.207, **p-value p < 0.01**).

4. MODELLO DI REGRESSIONE LINEARE MULTIPLA

Il primo modello che si è cercato di adattare è stato quello proposto da Greenwood, ovvero un modello di regressione lineare multipla della forma: E(*ho)* = β0 + β1*bm*+β2*wl*, dove *ho* rappresenta la variabile risposta, β0 è l'intercetta del modello, β1 e β2 sono rispettivamente i coefficienti di regressione relativi alle variabili *bm* e *wl*.

Il modello stimato risulta essere: $ho_{est.}$ = 28.31 + 1.70*bm* + 3.94*wl* (il piano di regressione del modello stimato è riportato in figura 13*)*, il che significa che per ogni aumento unitario di consumo calorico sotto sforzo, la massa corporea individuata aumenta di 1.70 unità ed il livello di lavoro aumenta di 3.94 unità. Gli I.C. di livello 1 – α = 0.95 sono [0.9989 - 2.3942] per *bm* e [3.6585 - 4.2205] per *wl*.

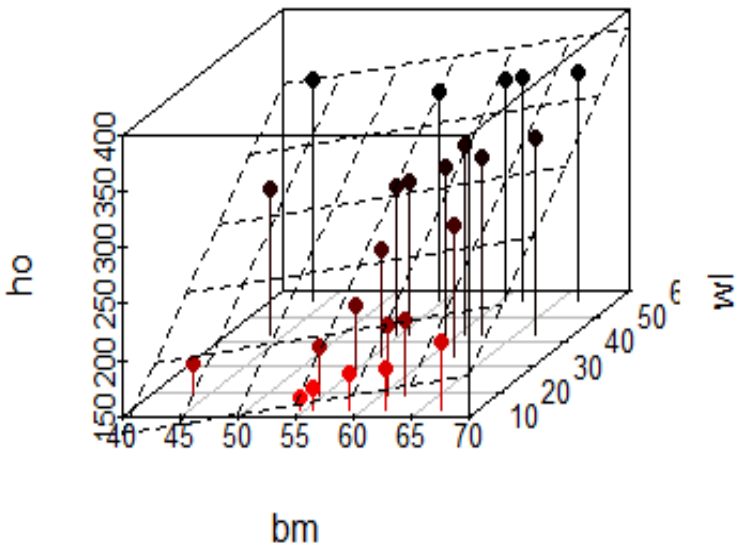


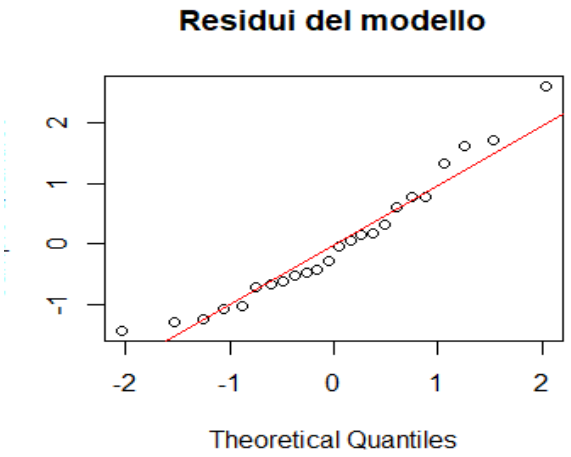
Figura 13: Scatterplot tridimensionale con piano di regressione del modello stimato

Con un valore t^{oss} della statistica F pari a 434.1 con 2 e 21 gradi di libertà ed un corrispondente **p-value p < 0.01**, il modello nel suo complesso risulta essere significativo. I regressori *bm* e *ho* risultano entrambi significativi ad ogni livello di α usuale. I valori assunti dal coefficiente di determinazione R^2, con un valore pari a **0.976**, e dall'R^2 **corretto**, con un valore pari a **0.974**, risultano essere molto elevati, indicando che il modello si adatta molto bene ai dati.

	Estimate	Std. Error	t-value	Pr(>|t|)	IC 0.95	Coefficienti beta
β_0	28.312	20.0806	1.410	0.173	[-13.447, 70.072]	-
bm	1.70	0.3355	5.057	0.035	[0.999, 2.394]	0.169
wl	3.94	0.1351	29.153	< 0.01	[3.658, 4.221]	0.978

Tabella 4: Summary del modello lineare stimato

Il **VIF** (fattore di inflazione della varianza) assume valore prossimo a **1** sia per *bm* che per *wl*, suggerendo quindi assenza di collinearità tra i regressori. Con un **coefficiente beta standardizzato** pari a **0.978**, si è ulteriormente appurato il maggior impatto sulla variabile risposta del regressore *wl* rispetto a *bm*.



A livello grafico (figura 14), il diagramma quantile contro quantile eseguito per valutare la normalità dei residui del modello appare soddisfacente; ciò viene confermato da un test di Shapiro-Wilk (t^{oss} = 0.945, **p-value p = 0.21**) per il quale l'ipotesi di normalità dei residui viene accettata a ogni livello di α usuale.

Figura 14: Diagramma quantile contro quantile (residui del modello)

Per verificare l'assunzione di omoschedasticità, sono stati considerati il grafico dei residui rispetto ai valori stimati dal modello (figura 15), il quale non ha evidenziato nessun andamento di tipo sistematico, e il grafico dei valori stimati rispetto ai valori osservati (figura 16).

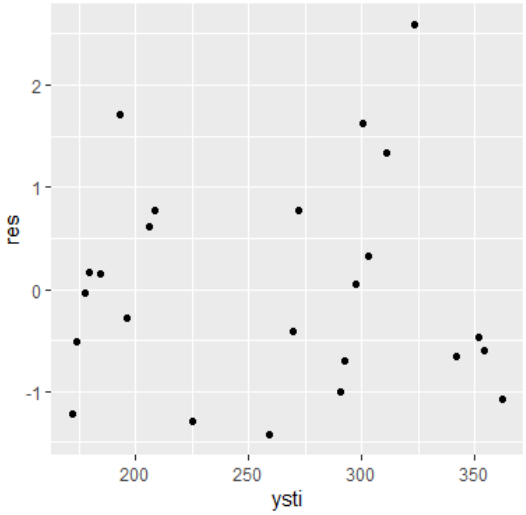


Figura 15: Residui v. valori stimati dal modello

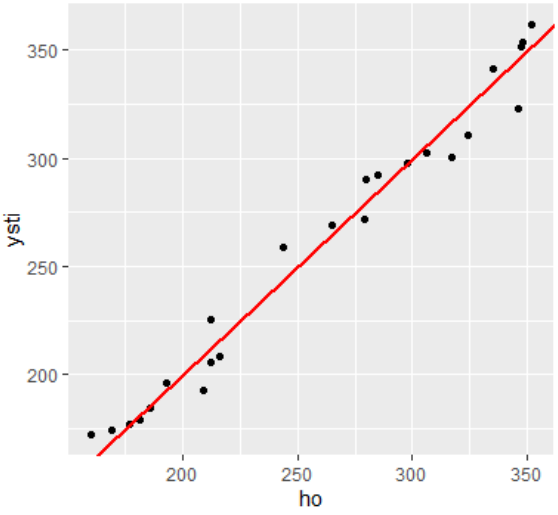


Figura 16: Valori stimati dal modello v. valori osservati

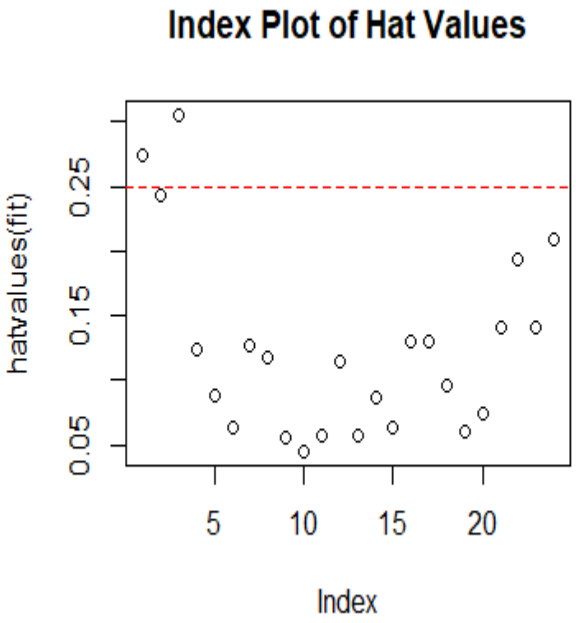
Complessivamente, si può concludere che la normalità dei residui appare soddisfacente considerata la numerosità contenuta del campione.

L'analisi dei punti leva, anomali e influenti ha evidenziato alcune criticità relative all'osservazione n° 3: il p-value aggiustato con correzione di Bonferroni per il più grande residuo studentizzato, relativo a questa unità, riporta un **p-value p = 0.143** (ricavato mediante Outlier-Test[2]), risultando dunque non significativo ad ogni livello di α usuale.

L'osservazione n° 3 si configura tuttavia come punto leva; i punti leva vengono definiti tali se si discostano sensibilmente dalla massa dei dati sul piano dei regressori, presentando valori inusuali di questi ultimi rispetto agli altri dati. Tale punto tende ad avere un residuo relativamente piccolo in quanto forza il modello a passargli vicino. La quantificazione del grado di "leva" di un punto è data dal coefficiente di leva, ovvero dal quadrato della distanza di Mahalanobis dell'osservazione dalla media dei regressori.[3] In letteratura[4] sono suggerite come soglie di attenzione il doppio o il triplo del valore medio *p/n (hat value)*, dove p è il numero di parametri del modello, compresa l'intercetta, e n il numero di osservazioni.

La linea orizzontale nel grafico in figura 17 evidenzia che le osservazioni n° 3 e n° 22 superano di circa 2.5 volte l'*hat value* medio; tuttavia, è dal grafico in figura 18 che è possibile notare come sia l'osservazione n° 3 a rappresentare, oltre che un punto leva, un punto notevolmente influente, ovvero tale da avere un impatto sui parametri del modello non proporzionale rispetto a quello delle altre osservazioni. L'osservazione in questione supera largamente la distanza di Cook, intesa come distanza standardizzata tra la previsione ottenuta col modello inclusivo di tutte le variabili e quella in cui verrebbe omessa.

[2] *Kabacoff (2011). "R in Action" Manning publications, p.200-201.*
[3] *Grigoletto, Pauli, Ventura (2017). Modello Lineare, Teoria e applicazioni con R, p.222.*
[4] *Belsey, D.A. ,Kuh, E., Welsch, R.E., Wiley (1980), Regression Diagnostics*



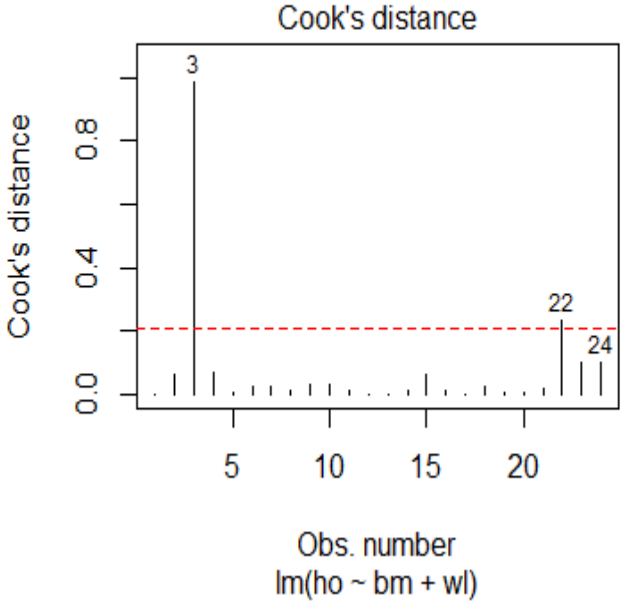
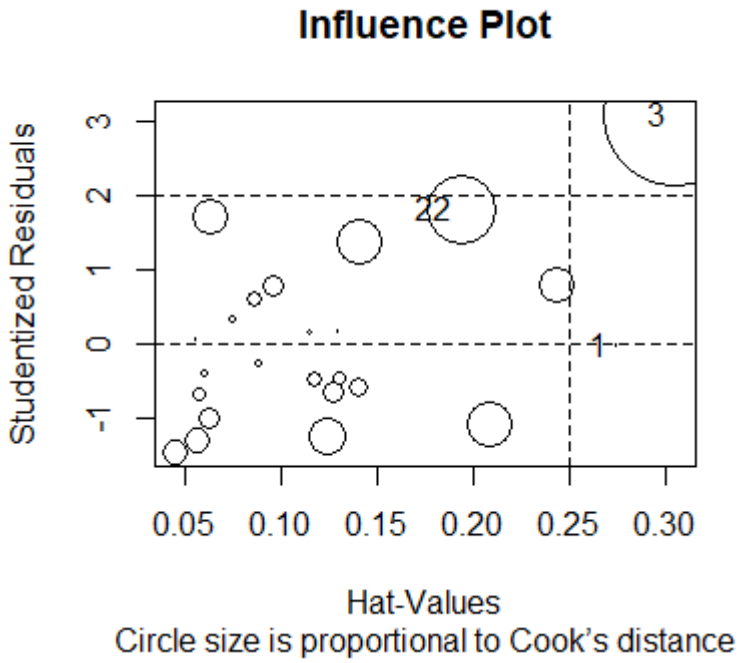


Figura 17: Grafico degli hat values per identificare punti leva. Figura 18: Grafico della distanza di Cook per identificare punti influenti



Nel grafico in figura 19 viene reso ulteriormente apprezzabile il largo impatto sul modello dell'osservazione n° 3 in quanto punto influente. Ciò ha portato a considerare un ri-adattamento del modello con un'esclusione del valore anomalo.

Figura 19: Grafico per l'individuazione dei punti influenti del modello; il diametro delle circonferenze è proporzionale alla distanza di Cook, ovvero rappresenta l'influenza dell'osservazione sui parametri del modello.

5. MODELLO DI REGRESSIONE LINEARE CON ESCLUSIONE DELL'OSSERVAZIONE N° 3

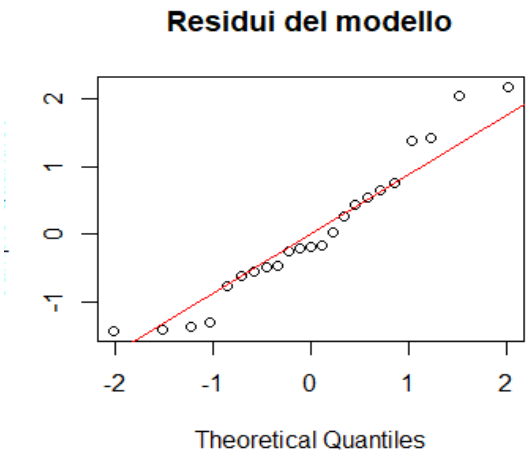
Il modello stimato risulta in questo caso essere: $ho_{est.} = 4.89 + 2.15bm + 3.83wl$ (la rappresentazione grafica del piano di regressione è stata in questo caso omessa in quanto ampiamente simile a quella precedente), il che significa che per ogni aumento unitario di consumo calorico sotto sforzo, la massa corporea individuata aumenta di 2.15 unità ed il livello di lavoro aumenta di 3.83 unità. Gli

I.C. di livello 1 – α = 0.95 sono [1.4814 – 2.8115] per *bm* e [3.5765 - 4.0766] per *wl*. Con un valore t^{oss} della statistica F pari a 565.5 con 2 e 20 gradi di libertà ed un corrispondente **p-value p < 0.01**, il modello nel suo complesso risulta essere significativo. I regressori *bm* e *ho* risultano entrambi significativi ad ogni livello di α usuale. I valori assunti dal coefficiente di determinazione **R^2**, con un valore pari a **0.982**, e dall'**R^2 corretto**, con un valore pari a **0.989**, prossimi a 1, indicano una bontà d'adattamento del modello superiore a quella del modello precedente. Il **VIF**, con valore pari a **1.014** sia per *bm* che *wl*, esclude presenza di collinearità tra i regressori.

	Estimate	Std. Error	t-value	Pr(>|t|)	IC 0.95	Coefficienti beta
β_0	4.89	18.587	0.263	0.795	[-33.885, 43.662]	-
bm	2.15	0.319	6.733	< 0.01	[1.481, 2.811]	0.120
wl	3.83	0.120	31.925	< 0.01	[3.576, 4.077]	0.948

Tabella 5: Summary del modello lineare stimato con omissione dell'osservazione n° 3

Il coefficiente beta standardizzato di *wl*, sebbene risulti inferiore a quello del modello comprensivo dell'osservazione n° 3, conferma il maggiore impatto del regressore wl sul modello rispetto a *bm*.



A livello grafico (figura 20), il diagramma quantile contro quantile eseguito per valutare la normalità dei residui del modello senza l'osservazione n° 3 evidenzia un lieve scostamento dalla retta di riferimento nella parte destra. Il test di Shapiro-Wilk (t^{oss} = 0.944, **p-value p = 0.23**) non porta al rifiuto dell'ipotesi di normalità per i residui a ogni livello di α usuale.

Figura 20: Diagramma quantile contro quantile (residui del modello)

Per verificare l'assunzione di omoschedasticità, sono stati considerati il grafico dei residui rispetto ai valori stimati dal modello (figura 21), il quale non ha evidenziato nessun andamento di tipo sistematico, e il grafico dei valori stimati rispetto ai valori osservati (figura 22).

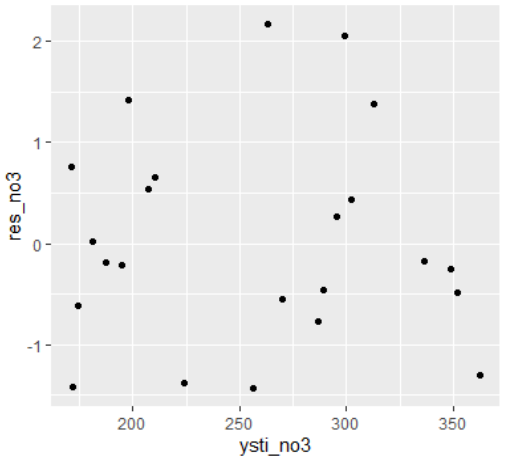


Figura 21: Residui v. valori stimati dal modello

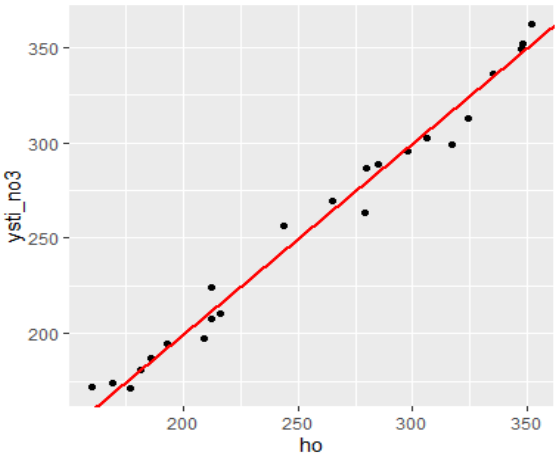


Figura 22: Valori stimati dal modello v. valori osservati

Anche qui, si può concludere che la normalità dei residui appare soddisfacente considerata la numerosità ridotta del campione.

L'analisi del leverage, degli outlier e dei punti influenti non ha evidenziato criticità tali da giustificare ulteriori interventi di natura correttiva sul modello.

6. MODELLO DI REGRESSIONE NON LINEARE MULTIPLA

Il secondo modello che si è proceduto ad analizzare è il modello di Glazebrook & Dye: questo è un modello di regressione non lineare multipla così definito: $E(ho) = \beta0 + \beta1 bm + (wl / (\beta3 + \beta4\, bm))$. Questo genere di modello assume che la media della variabile risposta sia specificata mediante una funzione parametrica non lineare dei regressori, nel caso specifico tale funzione è: $\mu(x; \underline{\beta}) = \beta0 + \beta1 bm + (wl / (\beta3 + \beta4 bm))$.

Prima di discutere dell'adattamento del modello, è opportuno premettere che i modelli di regressione non lineari vengono stimati mediante algoritmi numerici iterativi i quali richiedono l'inserimento, all'interno dell'input, dei valori di partenza dei parametri di regressione; ciò è necessario al fine di risolvere le equazioni di verosimiglianza soggiacenti. L'utilizzo di diversi valori di partenza per i parametri di regressione può portare l'algoritmo a convergere in diversi punti di massimo locali, questo a sua volta può portare a stime (e dunque a conclusioni) relative al modello anche sensibilmente differenti tra loro.

Si è dunque cercato di adattare il modello non lineare usando due linee d'approccio differenti: avvalendosi delle risorse di letteratura a disposizione sul caso, la prima ha consistito nell'utilizzare i valori delle costanti riportate da Glazebrook & Dye ricavate per via grafica (la cui natura è tuttavia non meglio specificata), la seconda ha sfruttato un algoritmo iterativo di tipo NLS (*Nonlinear Least Squares*) a partire da valori di β0, β1, β3 e β4 inizialmente settati a 1.

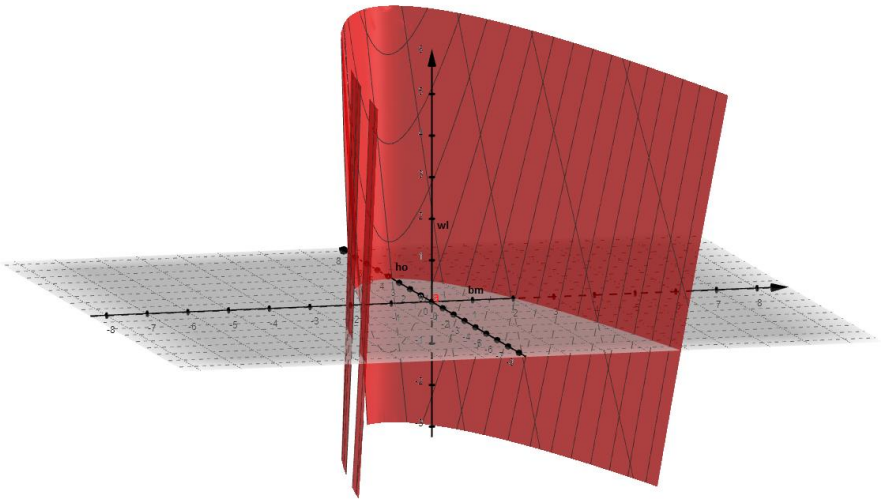


Figura 23: Approssimazione grafica della superficie di regressione nel piano tridimensionale relativa al modello non lineare

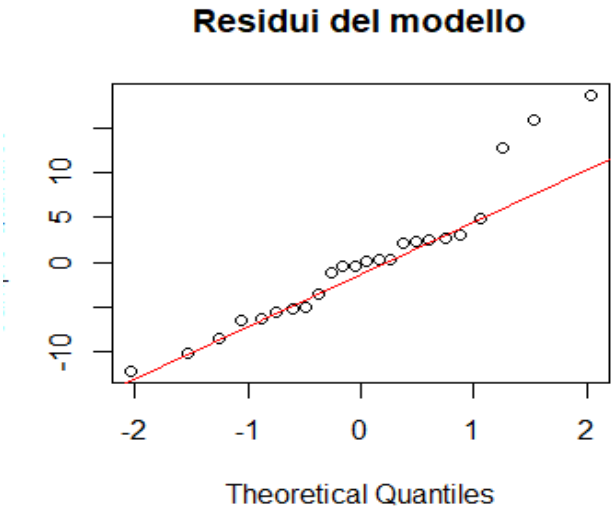
Le due procedure hanno condotto a risultati identici in termini di stime, standard error, t-value, significatività dei parametri, residual standard error ed AIC (174.3675 per entrambi), differendo unicamente nel numero di iterazioni necessarie a raggiungere la convergenza da parte dell'algoritmo (6 nel primo caso, 4 col secondo). In virtù di ciò, non verrà nel seguito fatta distinzione

tra il modello non lineare stimato mediante costanti di Glazebrook & Dye e quello ottenuto via NLS e si farà invece riferimento allo stesso unicamente in quanto "modello di regressione non lineare".

Il modello stimato risulta essere: $ho_{est.} = -8.652 + 4.222bm + (wl / (0.422+0.195bm))$. I parametri di regressione risultano essere tutti significativi eccetto per β3 (come si può vedere in tabella 6), si è scelto tuttavia di non adottare un approccio di tipo *backward elimination*[5], eliminando il parametro di regressione non significativo, in quanto si sarebbe corso il rischio di stimare un modello diverso i cui presupposti teorici non sono noti.

	Estimate	Std. Error	t-value	Pr(>|t|)	IC 0.95
β_0	-8.652	9.071	-3.509	< 0.01	[-13.929, -38.782]
β_1	4.222	2.117	7.336	<0.01	[2.867, 5.456]
β_3	0.422	0.555	0.760	0.4561	[-0.746, 0.136]
β_4	0.195	0.139	5.178	0.0307	[0.102, 0.005]

Tabella 6: Summary del modello non lineare stimato



Si è proceduto con un'analisi dei residui del modello stimato: il diagramma quantile contro quantile (figura 24) evidenzia un accentuato allontanamento sistematico dalla retta di riferimento, verosimilmente imputabile ad un'asimmetria di tipo positivo.

Figura 24: Diagramma quantile contro quantile (residui del modello)

Il test di Shapiro-Wilk, con un valore osservato $t^{oss} = 0.9244$ e un **p-value p = 0.073** non ha fornito evidenza circa la violazione dell'ipotesi di normalità per i residui del modello non lineare ad un livello di significatività α pari all'1% e al 5%.

Per verificare l'assunto di omoschedasticità, sono stati considerati il grafico dei residui rispetto ai valori stimati dal modello (figura 25) e il grafico dei valori stimati rispetto ai valori osservati (figura 26). Anche per questo modello la normalità dei residui appare complessivamente soddisfacente limitatamente alla ridotta numerosità campionaria; è altresì interessante notare come i grafici di residui v. valori stimati e di valori osservati v. valori stimati del modello risultino molto simili alle controparti esaminate per il modello lineare di Greenwood con e senza osservazione anomala (*par.4 e 5*).

[5] *Ventura, Racugno (2017). "Biostatistica: casi di studio in R," Egea s.p.a, p.188.*

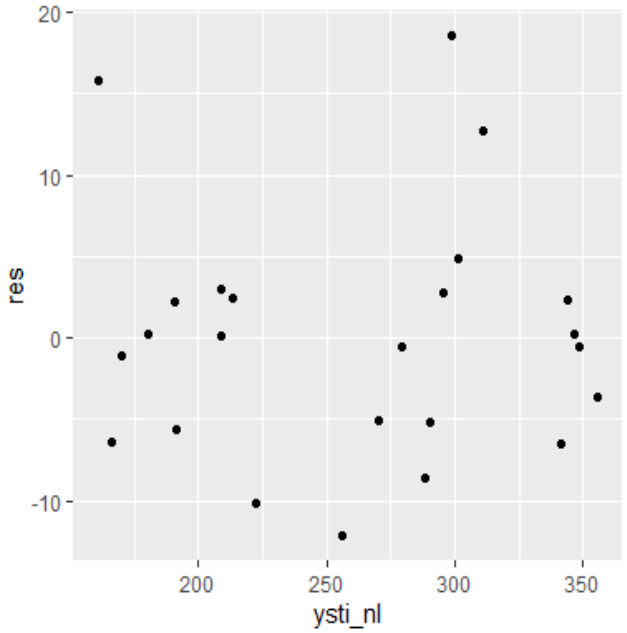


Figura 25: Residui v. valori stimati dal modello

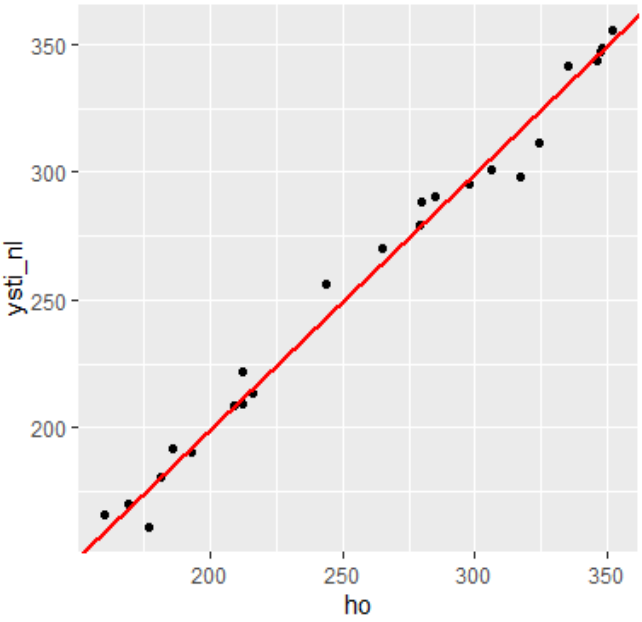


Figura 26: Valori stimati dal modello v. valori osservati

Per questo modello non sono emersi punti influenti, anomali o di leva rappresentanti criticità tali da escludere determinate osservazioni dall'input e operare un riadattamento.

7. CONFRONTO TRA MODELLI, PREVISIONI E CRITICITÀ DEI VALORI CALORICI NEGATIVI

I grafici dei residui rispetto ai valori stimati e dei valori osservati rispetto ai valori stimati dei tre modelli, come già detto nel *par. 6*, hanno evidenziato andamenti e caratteristiche di dispersione simili tra loro, mentre nel diagramma quantile contro quantile indagato per valutare l'adattamento all'ipotesi di normalità dei residui del modello non lineare e di quello lineare senza osservazione anomala sono presenti allontanamenti dalla retta di riferimento.

Considerato che i modelli lineari e quello non lineare sono tra loro non annidati, un indice che è parso sensato ricavare ai fini di un loro paragone è stato l'AIC; l'AIC (criterio d'informazione di Akaike) fornisce una misura della qualità della stima del modello tenendo conto sia della complessità dello stesso che della sua bontà di adattamento. La regola è di preferire il modello con AIC più basso.

Per il modello di Greenwood comprensivo di osservazione n° 3 l'AIC è pari a 186.221, per quello senza osservazione anomala è pari a 170.870 e per quello non lineare di Glazebrook & Dye è pari a 174.367; dunque il modello non lineare risulta, in questi termini, preferibile a quello lineare adattato con inclusione dell'osservazione n° 3, ma non a quello in cui quest'ultima viene esclusa.

Un'altra misura utile in termini di comparazione tra modelli è il BIC (criterio di informazione bayesiano, noto anche come criterio di Schwarz), un criterio per la selezione di un modello tra una classe di modelli parametrici con diverso numero di parametri. Il BIC assume valore 190.933 per il modello lineare con osservazione anomala, 175.411 per quello privo di quest'ultima e 180.257 per quello non lineare. Analogamente all'AIC, la regola è scegliere il modello con BIC inferiore; dunque,

anche in questo caso, il modello non lineare risulta preferibile solo rispetto al modello lineare comprensivo dell'osservazione n° 3.

	AIC	BIC
Modello lineare con osservazione n° 3	-8.652	9.071
Modello lineare senza osservazione n° 3	4.222	2.117
Modello non lineare	0.422	0.555

Tabella 7 : AIC e BIC dei modelli adattati.

Dunque, complessivamente, il modello non lineare di Glazebrook & Dye risulta preferebile in termini di paragone rispetto al modello di Greenwood con osservazione n° 3 ma non a quello in cui quest'ultima viene omessa. Il modello lineare di Greenwood con esclusione del valore anomalo gode, per altro, di una bontà d'adattamento estremamente alta considerato il valore del coefficiente di determinazione R^2 prossimo al 99% e un migliore adattamento dei residui all'ipotesi di normalità, sebbene questa non venga violata a livelli di significatività α pari all'1% e al 5% nel modello non lineare.

Previsioni fatte con il modello lineare senza omissione dell'osservazione n° 3 per un individuo di 75kg ad un livello di lavoro di 20cal/h riportano una produzione stimata di **158cal**, I.C. di livello 1 – α = 0.95: [137.77,178,23], bande di previsione inferiore e superiore: [128.08, 187.91].

Le stesse, eseguite con il modello lineare con omissione dell'osservazione n° 3, riportano una produzione stimata di **242cal**, I.C. di livello 1 – α = 0.95: [229.75, 255.04], bande di previsione inferiore e superiore: [219.85, 264.96].

Il modello non lineare, per questa previsione, ha fornito una stima di **143.84cal**, I.C. di livello 1 – α = 0.95 bootstrap: [127.79, 164.21], bande di previsione inferiore e superiore: [128.86, 168.04]. Dunque, la stima a parità di valori in input differisce di circa 15cal tra il modello lineare con osservazione n°3 e quello non lineare, mentre di ben 99cal tra il modello senza osservazione n° 3 e quello non lineare.

Un problema sollevato da Greenwood all'interno del suo studio è che sia la sua formula che quella di Glazebrook & Dye presentino la criticità di fornire valori calorici negativi per valori piccoli del peso quando il lavoro svolto è pari a zero, ciò tuttavia è spiegabile nella misura in cui entrambe le formule siano, di base, delle interpolazioni, dunque valori calorici negativi, pur non essendo coerenti con valori "sensati" del fenomeno indagato (il dispendio calorico minimo per qualsiasi tipo di attività è per definizione 0), risultano leciti considerata la natura intrinseca dei modelli. Ciò è stato confermato per via computazionale: la stima calorica di un individuo di 20kg con il modello lineare per il quale si assume lavoro esterno pari a 0cal/h è di **-32.66cal**, I.C. di livello 1 – α = 0.95: [-78.51,13.19], bande di previsione inferiore e superiore: [-78.55, 13.12].

8. CONCLUSIONI

Sono state eseguite le principali analisi univariate e bivariate delle variabili presenti all'interno del dataset. Dalle prime è emersa un'importante asimmetria all'interno della distribuzione della variabile

bm ed un rifiuto dell'ipotesi di normalità ad ogni livello di α usuale per le variabili *bm* e *wl* esaminate con test di Shapiro-Wilk e test di Lilliefors. Le analisi grafiche non hanno evidenziato la presenza di valori anomali. Il test di Anscombe-Glynn per la valutazione del grado di curtosi è risultato significativo per la variabile *ho*. Trasformazioni di scala delle variabili si sono rivelate inefficaci ai fini di una loro normalizzazione.

Le analisi bivariate si sono concentrate nell'indagare l'intensità del grado di correlazione tra le variabili: questa è stata valutata sia per via parametrica mediante metodo di Pearson che per via robusta mediante metodo di Spearman, considerato sia il numero ridotto di osservazioni, sia il fatto che per *bm* e *wl* l'ipotesi di normalità viene rifiutata ad ogni livello di α usuale mentre il p-value del test di Shapiro-Wilk per la variabile *ho* è molto vicino al livello di significatività fissato ($p = 0.04$, $\alpha = 0.05$). È emersa una forte correlazione lineare positiva tra calorie bruciate sotto sforzo e livello di lavoro ($\rho = 0.973$, $\rho_s = 0.957$).

Lo scopo principale dello studio era indagare sulla relazione tra calorie bruciate sotto sforzo e livello di lavoro (inteso come calorie bruciate all'ora) e peso corporeo: a tal fine sono stati adattati due modelli di regressione multipla, uno lineare della forma della forma $E(ho) = \beta0 + \beta1bm + \beta2wl$ e l'altro non lineare della forma $E(ho) = \beta0 + \beta1bm + (wl / (\beta3 + \beta4bm))$, quest'ultimo stimato a partire dalle costanti ottenute da Glazebrook & Dye e mediante un algoritmo iterativo di tipo NLS con valori delle costanti inizialmente settati a 1. I due modelli stimati ottenuti sono risultati identici, differendo unicamente nel numero di iterazioni necessarie per raggiungere la convergenza). Particolare attenzione è stata rivolta all'osservazione n°3, la quale, rappresentando un punto sia di leva che influente, ha portato a considerare un ulteriore adattamento del modello lineare previa esclusione della stessa dai dati in input. Quest'ultimo modello è risultato migliore del precedente in termini di bontà di adattamento, con un valore dell'R^2 prossimo al 99%. Lo stesso si è rivelato migliore a livello di AIC e BIC sia rispetto alla controparte comprensiva di osservazione anomala che rispetto al modello non lineare, basato sulla formula di Glazebrook & Dye.

In conclusione, nonostante sarebbe lecito a fronte dei risultati ottenuti assumere che il modello lineare di Greenwood sia più adatto allo scopo dell'indagine, il fatto che il modello lineare sia risultato sensibilmente migliore di quello non lineare solo in seguito all'esclusione di un'osservazione di carattere anomalo porta a concludere che una maggiore disponibilità campionaria sarebbe determinante al fine di comprendere meglio quale dei due modelli proposti sia preferibile in termini di performance e di paragone. Pro tanto ne è il fatto che le stime delle previsioni fornite con il modello non lineare e con quello lineare privo di osservazione n° 3 siano moderatamente discostate tra loro (di circa 99cal).

9. BIBLIOGRAFIA

- "On the Efficiency of Muscular Work" Proceedings of the Royal Society of London, Series B, Containing Papers of a Biological Character, M. Greenwood (1918)
- Regression Diagnostics, Belsey, D.A. ,Kuh, E., Welsch, R.E., Wiley (1980)
- Biostatistica: casi di studio in R, Ventura, Racugno, Egea s.p.a. (2017)
- Modello lineare: teoria e applicazione con R, Grigoletto, Pauli, Ventura, G. Giappichelli Editore (2017)
- R for Data Science, Wickham, Grolemund. O'Reilly (2017)
- R in Action, Kabacoff, Manning (2011)

CODICE R

```
# REPORT STATISTICA MEDICA
# muscle1.dat
# UPLOAD E ATTACHMENT
dati =
read.table("dati_muscl.txt",
header = TRUE)
attach(dati)
dim(dati)
head(dati)
dati

# ANALISI UNIVARIATE
# STATISTICHE DI SINTESI
# bm (BODY MASS) quantità di
massa corporea espressa (kg)
summary(bm)
sd(bm)

# wl (WORK LEVEL) livello di
lavoro (calorie orarie)
summary(wl)
sd(wl)

# ho (HEAT OUTPUT) calorie
consumate (calorie)
summary(ho)
sd(ho)

# CURTOSI
install.packages("moments")
library(moments)
kurtosis(bm)
kurtosis(wl)
kurtosis(ho)
anscombe.test(ho)

# BOXPLOT
boxplot(bm)
ggplot(data=dati,aes(x="Body
Mass", y = bm))+
  geom_boxplot()
ggplot(data=dati,aes(x="Work
Level", y = wl))+
  geom_boxplot()
ggplot(data=dati,aes(x="Heat
Output", y = ho))+
  geom_boxplot()

# GRAFICI DI DENSITA'
library(ggplot2)
#bm
# Basic density
p <- ggplot(dati, aes(x=bm)) +
  geom_density(color="darkblue",
fill="lightblue")+
  geom_vline(xintercept =
median(bm),     # Add line for
median
           col =
"black",linetype="dashed",lwd =
1)
# Add mean line
p+
geom_vline(aes(xintercept=mean
(bm)),
            color="red",
linetype="dashed", size=1)
p

#wl
p2 <- ggplot(dati, aes(x=wl)) +
  geom_density(color="darkblue",
fill="lightblue")+
  geom_vline(xintercept =
median(wl),     # Add line for
median
           col =
"black",linetype="dashed",lwd =
1)
# Add mean line
p2+
geom_vline(aes(xintercept=mean
(wl)),
            color="red",
linetype="dashed", size=1)
p2

#ho
p3 <- ggplot(dati, aes(x=ho)) +
  geom_density(color="darkblue",
fill="lightblue")+
  geom_vline(xintercept =
median(ho),     # Add line for
median
           col =
"black",linetype="dashed",lwd =
1)
# Add mean line
p3+
geom_vline(aes(xintercept=mean
(ho)),
            color="red",
linetype="dashed", size=1)
p3

# COEFFICIENTI DI VARIAZIONE
# bm
(sd(bm)/mean(bm))*100
#wl
(sd(wl)/mean(wl))*100
#ho
(sd(ho)/mean(ho))*100

# INDICI DI ASIMMETRIA
skew = function(x){
n = length(x)
s3 = sqrt(var(x)*(n-1)/n)^3
mx = mean(x)
sk = sum((x-mx)^3)/s3
sk/n}
skew(bm)
skew(wl)
skew(ho)

# TEST DI NORMALITA'
# bm (BODY MASS)
qqnorm(bm, main = "Body
Mass")
qqline(bm, col="red")
shapiro.test(bm)

# wl (WORK LEVEL)
qqnorm(wl, main = "Work Level")
qqline(wl, col="red")
shapiro.test(wl)

# ho (HEAT OUTPUT)
qqnorm(ho, main = "Heat
Output")
qqline(ho, col="red")
shapiro.test(ho)

install.packages("nortest")
library(nortest)
lillie.test(bm)
lillie.test(wl)
lillie.test(ho)

# TRASFORMAZIONI
QUADRATICHE DELLE VARIABILI
shapiro.test(bm)
shapiro.test(bm^2)
shapiro.test(wl)
shapiro.test(wl^2)
shapiro.test(ho)
shapiro.test(ho^2)

# ANALISI BIVARIATE
plot(dati)

# scatterplot tra wl e ho
plot(wl,ho)
ggplot(dati, aes(x = wl, y = ho)) +
  geom_point()

# scatterplot tra bm e ho
plot(bm, ho)
ggplot(dati, aes(x = bm, y = ho)) +
  geom_point()

# scatterplot tra bm e wl
```

```
plot(bm,wl)
ggplot(dati, aes(x = bm, y = wl)) +
  geom_point()

# CORRELAZIONI (PEARSON e
SPEARMAN)
cor(dati, method = "pearson")
cor.test(wl, ho, method =
"pearson")
cor.test(bm,ho,method =
"pearson")
cor.test(bm,wl,method="pearson
")

cor(dati, method="spearman")
cor.test(wl, ho, method =
"spearman")
cor.test(bm,ho,method =
"spearman")
cor.test(bm,wl,method="spearm
an")

# MODELLO LINEARE
modello1 = lm(ho ~ bm+wl, data
= dati)
summary(modello1)
vif(modello1)
# STIME STANDARDIZZATE
C.BETA
install.packages("QuantPsyc")
library(QuantPsyc)
lm.beta(modello1)

#RAPPRESENTAZIONE GRAFICA
DEL MODELLO LINEARE
scatterplot3d(bm,wl,ho)
library(scatterplot3d)
s3d <- scatterplot3d(bm, wl, ho,
pch=16, highlight.3d = TRUE, type
= "h")
fit0 <- lm(ho ~ bm+wl)
s3d$plane3d(fit0)
?scatterplot3d

# ANALISI DELLA
MULTICOLLINEARITA'
library(car)
vif(modello1)

#INTERVALLI DI CONFIDENZA PER
I PARAMETRI DI REGRESSIONE
confint(modello1)

#ANALISI DEI RESIDUI
res = rstandard(modello1)
qqnorm(res, main = "Residui del
modello")
qqline(res, col="red")
shapiro.test(res)

#RES.V.VAL.ST. E VALORI STIMATI
VS VALORI OSSERVATI
(OMOSCHEDASTICITA')
ysti = fitted(modello1)
ggplot(dati, aes(x = ysti, y = res))
+
  geom_point()

plot(ho,ysti)
abline(0,1,col="red")

ggplot(dati,
       aes(x = ho,
           y = ysti)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              size = 0.8)

# PREVISIONI
install.packages("propagate")
library(propagate)
predict(modello1)
newdata<-data.frame(bm=30,
wl=20)
predict(modello1,newdata,interv
al="prediction")
predict(nl_fit,newdata,interval=c(
"prediction"))
newdata1<-
data.frame(bm=20,wl=0)
predict(nl_fit,newdata=newdata1
)
predictNLS(nl_fit,newdata)
predictNLS(nl_fit,newdata1)

# MODELLO NON LINEARE
GLAZEBROOK & DYE
nl_fit = nls(ho ~
b0+b1*bm+(wl/(b3+b4*bm)),dat
a = dati, start = list(b0=-
138,b1=4.5,b3=0.08,b4=0.003))
summary(nl_fit)

# MODELLO NON LINEARE CON
COSTANTI SETTATE A 1
nl_fit_one = nls(ho ~
b0+b1*bm+(wl/(b3+b4*bm)),dat
a = dati, start = list(b0=-
116.593,b1=4.2128,b3=0.03183,
b4=0.0039))
summary(nl_fit_one)

AIC(nl_fit_one)
AIC(nl_fit)

#INTERVALLI DI CONFIDENZA PER
I PARAMETRI DI REGRESSIONE
confint(nl_fit_one)

#RAPPRESENTAZIONE GRAFICA
MODELLO NON LINEARE n.1
scatterplot3d(bm,wl,ho)
s3d <- scatterplot3d(bm, wl, ho,
pch=16, highlight.3d = TRUE, type
= "h")
s3d$plane3d(mod)

#ANALISI DEI RESIDUI NON LIN.
res = residuals(nl_fit)
qqnorm(residuals(nl_fit), main =
"Residui del modello")
qqline(residuals(nl_fit),
col="red")
shapiro.test(residuals(nl_fit))

# residui v. valori stimati
ysti_nl = fitted(nl_fit)
ggplot(dati, aes(x = ysti_nl, y =
residuals(nl_fit))) +
  geom_point()

# valori stimati v. valori osservati
ggplot(dati,
       aes(x = ho,
           y = ysti_nl)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              size = 0.8)

# AIC e BIC MODELLI
AIC(modello1)
AIC(nl_fit)
BIC(modello1)
BIC(nl_fit)

# ULTERIORI GRAFICI
TRIDIMENSIONALI
# x, y, z variables
x <- bm
y <- wl
z <- ho
# Compute the linear regression
fit_nova <- lm(z ~ x + y, data =
dati)
# predict values on regular xy
grid
grid.lines = 50
x.pred <- seq(floor(min(x)),
ceiling(max(x)), length.out =
grid.lines)
```

```
y.pred <- seq(floor(min(y)),
ceiling(max(y)), length.out =
grid.lines)
xy <- expand.grid( x = x.pred, y =
y.pred)
z.pred <- matrix(predict(fit_nova,
newdata = xy),
            nrow = grid.lines, ncol =
grid.lines)
# fitted points for droplines to
surface
fitpoints <- predict(fit_nova)
# scatter plot with regression
plane
scatter3D(x, y, z, pch = 16, cex =
0.5, colvar=FALSE,
        col="black",theta = 30, phi
= 20, bty="b2",
        expand =0.7,
        xlab = "Body Mass", ylab =
"Work Level", zlab = "Heat
Output",
        surf = list(x = x.pred, y =
y.pred, z = z.pred,
                facets = NA,col
="red",fit = fitpoints ),main = " ")
# PUNTI ANOMALI
library(car)
outlierTest(modello1)

# PUNTI LEVA
hat.plot <- function(fit) {
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main="Index
Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red",
lty=2)
  identify(1:n, hatvalues(fit),
names(hatvalues(fit)))
}
hat.plot(modello1)

# PUNTI INFLUENTI
cutoff <- 4/(nrow(dati)-
length(modello1$coefficients)-2)
plot(modello1, which=4,
cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")

influencePlot(modello1,
id.method="identify",
main="Influence Plot",
          sub="Circle size is
proportional to Cook's distance")
```

```
# MODELLO DI REGRESSIONE
LINEARE (OSSERVAZIONE 3
ESCLUSA)
dati_no3 =
read.table("dati_no_oss3.txt",
header = TRUE)
modello_no3 = lm(ho ~ bm+wl,
data = dati_no3)
summary(modello_no3)
confint(modello_no3)
vif(modello_no3)
# STIME STANDARDIZZATE
C.BETA
install.packages("QuantPsyc")
library(QuantPsyc)
lm.beta(modello_no3)

#ANALISI DEI RESIDUI
res_no3 =
rstandard(modello_no3)
qqnorm(res_no3,main = "Residui
del modello")
qqline(res_no3, col="red")
shapiro.test(res_no3)

#RES.V.VAL.ST. E VALORI STIMATI
VS VALORI OSSERVATI
(OMOSCHEDASTICITA')---------
ysti_no3 = fitted(modello_no3)
ggplot(dati_no3, aes(x =
ysti_no3, y = res_no3)) +
 geom_point()
ggplot(dati_no3,
     aes(x = ho,
       y = ysti_no3)) +
 geom_point() +
 geom_abline(intercept = 0,
          slope = 1,
          color = "red",
          size = 0.8)

# PUNTI ANOMALI
library(car)
outlierTest(modello_no3)

# PUNTI LEVA
hat.plot <- function(fit) {
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main="Index
Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red",
lty=2)
  identify(1:n, hatvalues(fit),
names(hatvalues(fit)))
}
hat.plot(modello_no3)
```

```
# PUNTI INFLUENTI
cutoff <- 4/(nrow(dati_no3)-
length(modello_no3$coefficients
)-2)
plot(modello_no3, which=4,
cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")

influencePlot(modello_no3,
id.method="identify",
main="Influence Plot",
          sub="Circle size is
proportional to Cook's distance")

# ANALISI DEI PUNTI ANOMALI,
LEVA E INFLUENTI (MOD. NON
LINEARE)
influencePlot(nl_fit,
id.method="identify",
main="Influence Plot",
          sub="Circle size is
proportional to Cook's distance")

#AIC e BIC mod. lineare senza
oss.3
AIC(modello_no3)
BIC(modello_no3)
```