L'R^2? No, grazie!

Giulio Palomba*

Università Politecnica delle Marche
Dipartimento di Scienze Economiche e Sociali (DISES)

Gennaio 2013

Premessa

Nell'ambito delle scienze statistiche, dell'econometria o dell'analisi empirica in generale, l'indice R^2 è noto a tutti. C'è chi lo chiama indice di determinazione, chi adattamento ai dati traducendo liberamente dall'inglese *goodness of fit* o chi semplicemente usa la dicitura compatta "Erre-Quadro". In estrema sintesi esso può essere definito come un indicatore compreso nell'intervallo $[0, 1]$, in grado di dirci se il modello stimato si adatta bene ai dati oppure no. In pratica, più è alto il valore dell'indice, meglio è; naturalmente, se l'analista (o chi per lui) limitasse la conoscenza dell'argomento alla sola definizione appena fornita, potrebbe essere indotto a pensare che la vicinanza al valore unitario rappresenti un segno inequivocabile che le cose stiano andando piuttosto bene. Non importa se lo stesso analista è uno studente, un economista/statistico applicato dell'ultima ora o chiunque si trovi a condurre un'analisi empirica sui dati. L'unica cosa rilevante è che spesso il valore dell'R^2 è utilizzato per giustificare conclusioni affrettate o del tutto arbitrarie.

Troppo spesso ho ascoltato gente mentre attribuisce una qualche facoltà divinatoria all'indice e la ragione per cui ciò avvenga è dettata da una combinazione di fattori i cui confini si mescolano e si confondono: sto parlando sicuramente di disinformazione e cattivo apprendimento a cui si vanno ad aggiungere altri elementi come la tradizione, la consuetudine e talvolta anche un briciolo di italica superstizione.

In realtà le cose sono ben diverse: con queste pagine non intendo assolutamente screditare l'uso di un indicatore "storico" e di uso comune che ha saputo ritagliarsi un posto d'onore negli output di stima presso qualsiasi software statistico-econometrico. Il mio intento è semplicemente aprire gli occhi dei lettori sul fatto che l'R^2 di certo non è il *salvatore della patria* e soprattutto non ha le caratterstiche per diventarlo. Alla luce di questa premessa, posso solo aggiungere che la sua importanza andrebbe (una volta per tutte) ridimensionata o quantomeno discussa più approfonditamente.

Dopo aver introdotto l'indice e le sue proprietà di base nella sezione 1, provvederò alla sua critica nelle sezioni 2 e 3: nella prima delle due sarò abbastanza cattivo con lui, ma poi lo riabiliterò, seppur parzialmente, nella successiva.

*Questo lavoro beneficia di alcuni preziosi commenti e suggerimenti da parte di amici che fanno il mio mestiere: sto parlando, in ordine rigorosamente alfabetico, di Francesca Di Iorio, Luca Fanelli, Jack Lucchetti e Claudia Pigini. Anche se all'apparenza si potrebbe pensare ad un complotto contro l'R^2, in realtà l'idea e la realizzazione, ivi compresi gli errori e/o le imprecisioni, sono imputabili esclusivamente al sottoscritto.

Indice

1 Definizione formale

Sfortunatamente mi tocca: devo definire formalmente cosa sia l'R^2. Non posso esimermi perché dovrò approfondire diversi aspetti tecnici del problema e qualche nozione di base ci vuole. Forza e coraggio, dunque; da qui in avanti cercherò di rendere la parte formale il meno invadente possibile.

Si consideri pertanto il modello lineare

$$y_i = X_i'\beta + \varepsilon_i \tag{1}$$

dove y_i è la variabile dipendente relativa all'i-esima osservazione campionaria, X_i' è il vettore riga contenente le k variabili esplicative ed infine ε_i è quel termine che viene comunemente definito come l'errore o il disturbo del modello (1). Se definisco il numero n come la numerosità del campione da cui l'i-esima osservazione è estratta, direi che fin è qui tutto regolare.

Chi ha studiato l'"A-B-C" dell'econometria si è sicuramente imbattuto nella formula

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2}$$

che è quella propria dello stimatore dei minimi quadrati ordinari o, meglio ancora, dello stimatore OLS. Si noti semplicemente che y_i è un'osservazione appartenente al vettore y di dimensione n, mentre X_i' rappresenta l'i-esima riga della matrice dei regressori X che ha dimensione $n \times k$.

Non voglio annoiare nessuno con le proprietà di questa funzione statistica oppure sul come e sul perché essa venga applicata. Per questo c'è un'infinità di testi che è in grado di spiegare la cosa molto meglio di me. A me è sufficiente che la formula (2) possa essere calcolata, il che equivale che una stima OLS dei k parametri contenuti nel vettore β sia effettivamente disponibile. Con estrema fantasia, chiamerò questa stima $\hat{\beta}$.

La logica dei modelli lineari come quello dell'equazione (1) ci dice sostanzialmente questo: è possibile spiegare le variazioni di y_i ponderando (o, se volete, combinando linearmente) le osservazioni contenute nell'i-esima riga della matrice dei regressori attraverso gli elementi di $\hat{\beta}$. Va da sé che la migliore previsione che ho disponibile sarà equivalente a

$$\hat{y}_i = X_i'\hat{\beta}. \tag{3}$$

Per gli amanti del rigore statistico, la quantità $\hat{y}_i$ rappresenta una stima del valore atteso condizionale $E(y_i|X_i)$ qualora effettivamente ci muovessimo all'interno di un mondo lineare, ma inserico questa considerazione più per completezza che per un effettiva utilità nell'ambito di questa trattazione.

A questo punto è chiaro che da una parte ho le vere osservazioni campionarie, mentre dall'altra ho alcune osservazioni "artificiali" che in un qualche modo dovrebbero ricalcare le prime. Chiamiamo queste ultime valori stimati di y_i, previsioni, modello stimato, oppure diamoci quell'aria da *pseudo intellectual-chic* utilizzando improbabili neologismi come ad esempio modello "fittato" (participio passato italiano del verbo anglosassone *to fit*), di fatto il discorso non cambia: se, per ogni i, $\hat{y}_i$ è vicino ad y_i abbiamo fatto un buon lavoro, altrimenti il nostro modello lineare è da rivedere o addirittura da buttare. Qui interviene il nostro amico, definito come

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}, \tag{4}$$

dove il vettore di dimensione n $\hat{\varepsilon} = y - \hat{y}$ corrisponde alla stima dell'errore del modello o, più elegantemente, al residuo della regressione. Osservando la (4) è facile notare che:

1. l'equazione si concentra sui momenti secondi ed in particolare ci dice qual è il contributo del momento secondo calcolato sui valori previsti $\hat{y}_i$ su quello calcolato sui valori effettivi y_i;
2. poiché $y = \hat{y} + \hat{\varepsilon}$, allora $R^2 \in [0, 1]$ (nell'universo OLS vale la relazione magica $\hat{y}'\hat{\varepsilon} = \hat{\varepsilon}'\hat{y} = 0$);
3. se gli errori stimati sono molto piccoli (con estremo abuso di notazione identificherei questa situazione attraverso l'approssimazione $\hat{\varepsilon} \approx 0$), significa che i veri valori di y e quelli previsti dal modello $\hat{y}$ sono pressoché identici, quindi il nostro modello econometrico è assolutamente OK (quindi, per coerenza: $R^2 \approx 1$).

Adesso che abbiamo fatto gli onori di casa, possiamo partire: a prima vista l'R^2 sembra effettivamente costituire la soluzione ideale a tutti i nostri problemi perché sarebbe sufficiente dargli un'occhiata e sincerarsi che esso sia il più vicino possibile al suo estremo superiore. Tutto sommato basterebbe scegliere una soglia convenzionale oltre la quale possiamo affermare che il nostro modello è buono. Ebbene, in realtà questa soglia di fatto non esiste, dato che in certi contesti un $R^2 = 0.95$ potrebbe essere insufficiente, mentre in altri un valore pari a 0.2 potrebbe risultare soddisfacente. Questa caratteristica, già di per sé, potrebbe far storcere il naso di fronte a questo indicatore ma, come vedremo in seguito, questa è solo la punta di un iceberg irto di problemi.

2 Sfatiamo il mito

Già la definizione di cui all'equazione (4) nasconde un inganno per le menti più inesperte: in realtà l'R^2 non è propriamente una misura circa la bontà del modello statistico applicato ai dati. La questione è più sottile perché esso in realtà fornisce un'informazione sintetica soltanto sulla qualità dell'approssimazione lineare imposta dal modello (1).

Le cose perciò stanno più o meno così: se il meccanismo generatore dei dati (DGP) di y_i è davvero costituito da una relazione lineare condizionata dai valori dei regressori contenuti in X_i', allora il nostro indicatore ci segnalerà effettivamente in che percentuale il modello stimato $\hat{y}_i$ si adatta all'andamento della variabile dipendente. Se invece il DGP è qualcosa di diverso, allora l'R^2 è un qualcosa che misura la performance lineare di un modello che lineare non è.

Ma la cosa più incredibile è un'altra: per costruzione, l'R^2 assume il valore più alto possibile qualora il modello stimato sia effettivamente l'OLS e questa caratteristica vale sempre, indipendentemente dal "vero" DGP.

Infine, questo indice non ci dice nulla circa alcuni aspetti basilari della stima effettuata attraverso il metodo OLS, infatti:

1. non aiuta di certo a capire se una variabile esplicativa sia statisticamente significativa,

2. non ci dice nulla sul fatto che potremmo avere carenza di variabili esplicative all'interno del nostro modello,[1]

3. non dà informazioni sufficienti per stabilire se i regressori utilizzati siano in effetti quelli più appropriati per spiegare le dinamiche della variabile dipendente,

4. non fornisce alcuna indicazione circa le possibili trasformazioni sulle variabili indipendenti che potrebbero migliorare la stima,

5. è invariante alle riparametrizzazioni del modello lineare (sarò più preciso nella sezione 2.4).

2.1 Quale R^2?

Usando le sommatorie, l'equazione (4) può essere riscritta come segue

$$R^2_{nc} = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} y_i^2} \tag{5}$$

e basta moltiplicare numeratore e denominatore per il fattore $1/n$, oppure $1/(n-1)$ per gli statistici ortodossi, per accorgersi che si sta parlando del rapporto tra momenti campionari. L'equazione (5) è quella dell'R^2 **non centrato**, nel quale sia i valori della variabile dipendente y_i, sia i residui $\hat{\varepsilon}_i$ non sono espressi in deviazione dalle rispettive medie campionarie. In quest'ambito è giusto ricordare che la media dei residui del modello OLS non è necessariamente zero, infatti basta che il vettore X_i' di cui alla (1) non contenga un termine costante affinché risulti

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \hat{\beta}_1,$$

dove $\hat{\beta}_1$ è il parametro relativo alla costante stessa.

Per essere sicuri che i residui siano a media nulla, occorre perciò inserire sempre una costante nel modello! Sotto questa condizione il momento secondo calcolato sui residui, noto anche come media dei quadrati dei residui, equivale alla varianza dei residui stessi. Considerando anche la varianza della variabile dipendente, l'indice R^2 può essere ridefinito come segue:

$$R^2 = \frac{Var(\hat{y}_i)}{Var(y_i)} = 1 - \frac{Var(\hat{\varepsilon}_i)}{Var(y_i)} = 1 - \frac{\frac{1}{n-1}\sum_{i=1}^{n} \varepsilon_i^2}{\frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{6}$$

oppure, in termini matriciali,

$$R^2 = 1 - \frac{\hat{\varepsilon}' M_\iota \hat{\varepsilon}}{y' M_\iota y} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y' M_\iota y} \tag{7}$$

dove $\bar{y}$ è la media campionaria di y_i, $\iota = [\ 1 \quad 1 \quad \dots \quad 1\ ]'$ e $M_\iota = I_n - \iota(\iota'\iota)^{-1}\iota'$ è la nota matrice idempotente di proiezione che, applicata ad una variabile, ritorna i suoi scarti dalla media. Le formule (6)-(7) si riferiscono all'R^2 "ufficiale", cioè quello che i software statistico-econometrici pubblicano insieme alle stime OLS.

Una simpatica proprietà è pertanto $R^2_{nc} \geq R^2$. Naturalmente l'uguaglianza stretta è ottenuta quando $E(y_i) = 0$.

[1] Tecnicamente, mi riferisco al problema delle variabili omesse.

Dimostrazione:

$$\begin{aligned} 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} &\geq 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'M_\iota y} \\ \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} &\leq \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'M_\iota y} \\ y'y &\geq y'M_\iota y \end{aligned}$$

Questa disuguaglianza può essere risolta in due modi. La soluzione statistica ci suggerisce di moltiplicare entrambi i membri per $1/n$ e notare che a sinistra abbiamo il momento secondo di y_i, mentre a destra c'è la varianza; poiché $E(y_i^2) \geq Var(y_i)$ per definizione, il gioco è fatto. La soluzione algebrica invece ci direbbe di osservare l'equazione

$$\begin{aligned} &y'y - y'M_\iota y \geq 0 \\ &y'[I_n - M_\iota]'y \geq 0 \\ &y'[\iota(\iota'\iota)^{-1}\iota']'y \geq 0 \end{aligned}$$

per poi notare che essa è una forma quadratica all'interno della quale la matrice $P_\iota = \iota(\iota'\iota)^{-1}\iota'$ è una matrice di proiezione che trasforma una variabile in un vettore costante i cui elementi corrispondono alla media campionaaria. Ciò che rileva qui è che questa matrice, oltre ad essere quadrata ed idempotente, è soprattutto definita positiva: anche in questo caso perciò, la relazione è dimostrata.

L'indice definito dalle equazioni (6)-(7) però soffre di un problema rilevante: se nel modello lineare che si sta stimando si aggiungono via via sempre più regressori, finisce che l'R^2 aumenta o, quantomeno, di sicuro non diminuisce: in pratica, se voglio aumentare l'R^2, mi basta utilizzare più variabili esplicative, statisticamente rilevanti oppure no. Estremizzando ed in linea del tutto teorica, tecnicamente potrei sempre raggiungere il livello massimo $R^2 = 1$ se il numero k di tali variabili esplicative fosse elevato o al limite infinito.

Questa proprietà genera due conseguenze spiacevoli per l'analista: da un lato, bisogna prendere atto che questo indicatore non rappresenta assolutamente uno strumento utile per la specificazione del modello, specialmente quando si tratta di aggiungere/togliere variabili.[2] Dall'altro, l'affollamento di variabili esplicative effettuato con lo scopo di aumentare il valore dell'R^2 non fa il paio con la strategia di rendere il modello parsimonioso, ovvero quella di spiegare la variabilità di y_i con il minimo indispensabile di variabili all'interno di X_i'. Come ne usciamo?

Naturalmente è stato pensato un altro R^2 (e sono 3!). Questo indice ha la praticissima caratteristica di correggere le varianze per i gradi di libertà, quindi incrementa il proprio valore numerico solo nel caso in cui l'aumento di k nel modello lineare è determinato dall'aggiunta di variabili statisticamente significative. Questo indice è noto come R^2 **corretto** la cui definizione formale è

$$\bar{R}^2 = 1 - \frac{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}}{\frac{y'M_\iota y}{n-1}} = 1 - \frac{\frac{1}{n-k}\sum_{i=1}^{n}\varepsilon_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{8}$$

Con un po' di algebra è facile ottenere la relazione

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2) \tag{9}$$

dalla quale emerge il *trade off* (termine caro agli economisti) in base al quale l'aumento del numero k dei parametri comporta:

- da un lato, l'aumento del rapporto $\dfrac{n-1}{n-k}$, quindi anche quello di $\bar{R}^2$;

[2] In questo caso si parla di modelli *nested*.

- dall'altro, l'aumento di R^2 che però ha un incidenza al ribasso per $\bar{R}^2$.

Infine, non ci vuole molto a capire che:

1. in presenza di un solo regressore nel modello lineare la versione corretta dell'indice coincide con quella ufficiale (in formule: $\bar{R}^2 = R^2 \Leftrightarrow k = 1$),
2. l'estremo inferiore di $\bar{R}^2$ in pratica non è mai esattamente zero (salvo ipotetici scenari in cui $n \to \infty$) perché dipende da k, infatti risulta

$$\bar{R}^2 \in \left[1 - \frac{n-1}{n-k}, 1\right].$$

3. dalla precedente proprietà segue che l'indice $\bar{R}^2$ può assumere **valori negativi** per $k > 1$ e valori di R^2 molto vicini allo zero;
4. quando $k > 1$, risulta $\bar{R}^2 < R^2$ per $\forall\ n$. Generalizzando, vale perciò la disuguaglianza

$$\bar{R}^2 \leq R^2 \leq R^2_{nc}. \tag{10}$$

Dimostrazione:

In breve:

$$1 - \frac{n-1}{n-k}(1 - R^2) < R^2$$
$$1 - \frac{n-1}{n-k} < \left(1 - \frac{n-1}{n-k}\right) R^2$$

Se $k = 1$ oppure per $R^2 = 1$ si ottiene l'uguaglianza stretta tra i due indici; se invece $k > 1$ la quantità a sinistra del segno di disuguaglianza e quella tra parentesi sono negative, quindi la relazione è garantita per ogni $R^2 \in [0, 1)$.

In definitiva, resta la domanda fondamentale: quale indice bisogna scegliere? Una risposta esatta non c'è, anche perché non esiste alcun vincolo di esclusione su questo o quest'altro indice, dato che tutti i diversi tipi sono calcolabili (o al limite programmabili) con poco sforzo. Il problema fondamentale è l'informazione che uno cerca a cui si potrebbe aggiungere un'eventuale omissione del software in uso.

Dal punto di vista informativo R^2_{nc} appare senza dubbio come il meno affidabile dato che non è interpretabile come un contributo del modello stimato sulla varianza totale di y_i. Questo indice poi condivide con l'R^2 "ufficiale" la cattiva proprietà di essere funzione non decrescente del numero di regressori (k). Alla luce di queste affermazioni, un bravo analista dovrebbe perciò affidarsi sempre e solo alla versione corretta $\bar{R}^2$. In un ipotetico mondo ristretto ai tre indici forse sì, ma sappiamo tutti che la scienza statistica si è evoluta e non mancano di certo strumenti accessori oppure alternativi di valutazione.

Infine, c'è il problema del software. Non tutti pubblicano i tre indici e spesso accade che solo R^2 sia visibile negli output di stima. Credo che la ragione per la quale ciò accada sia semplicemente tradizione, anche se ciò non aiuta granché l'analista al quale non resta altro da fare che dare una rapida occhiata per poi studiare il problema più approfonditamente.

2.2 Il quadrato della correlazione lineare

Si supponga per un momento che il modello lineare (1) contenga solo due regressori di cui il primo è la costante, cioè risulti $X_i' = [\ 1 \quad x_i\ ]$. In questo caso si ha $k = 2$ e lo stimatore ha la nota soluzione

$$\begin{bmatrix} \hat{\beta}_1 \\ \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_2 \bar{x} \\ \\ \dfrac{Cov(x,y)}{Var(x)} \end{bmatrix}, \tag{11}$$

dove $\bar{y}$ e $\bar{x}$ sono le medie del campione rispettivamente di y_i e x_i. Con un po' di algebra si arriva all'altrettanto noto risultato

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 x_i - \hat{\beta}_1 - \hat{\beta}_2 \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\hat{\beta}_2^2 (x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \\ &= \hat{\beta}_2^2 \frac{Var(x)}{Var(y)} = \left[\frac{Cov(x,y)}{Var(x)}\right]^2 \frac{Var(x)}{Var(y)} = \frac{Cov(x,y)^2}{Var(x)Var(y)} = \end{aligned}$$

$$R^2 = \rho_{xy}^2. \tag{12}$$

Il senso della relazione (12) è semplice: l'R^2 è pari al quadrato del coefficente di correlazione ρ_{xy}. L'interpretazione statistica di questo risultato ci suggerisce che se la variabile dipendente è correlata con il regressore x_i, allora l'R^2 è grande ed il nostro modello interpreta bene i suoi movimenti. Il tutto non fa una piega.

Adesso però facciamo le seguenti considerazioni:

1. la correlazione indica solo una relazione **lineare** tra le variabili, ma di questo ho già parlato in precedenza;
2. il segno della correlazione indica se tale relazione lineare è diretta o inversa. Poiché l'R^2 equivale al quadrato della correlazione, l'informazione sulla direzione della correlazione va persa;
3. ci vuole un secondo a calcolare un coefficiente di correlazione, quindi, in un secondo, possiamo attingere ad un'informazione più completa rispetto all'R^2.

La relazione (12) può essere facilmente generalizzata per $k > 2$ come segue

$$R^2 = \sum_{j=2}^{n} \frac{Cov(x_j, y)^2}{Var(x_j)Var(y)} = \sum_{j=2}^{n} \rho_{x_j y}^2 \tag{13}$$

dove x_j è la j-esima colonna della matrice X. In questo caso l'R^2 rappresenta un aggregato, quindi conserva il vantaggio di essere un indice sintetico. Resta comunque la perdita di informazione sui segni delle singole correlazioni, ma questo è un male assolutamente sopportabile, dato che tali segni appaiono all'interno del vettore stimato $\hat{\beta}$.

2.3 Trasformazioni di variabili

Consideriamo l'esempio più ovvio: dopo aver stimato il modello (1), effettuiamo la stessa stima sul modello

$$\dot{y}_i = \dot{X}_i' b + u_i, \tag{14}$$

dove $\dot{y}_i = \ln y_i$, $\dot{x}_i = \ln x_i$, u_i è il termine d'errore, mentre i coefficienti contenuti nel vettore b sono interpretabili come le elasticità della variabile dipendente y_i rispetto ai singoli regressori contenuti in X_i. Come si comporta l'R^2?

Innanzi tutto, fornisco l'equazione:

$$R_l^2 = \frac{Var(\dot{X}\hat{b})}{Var(\dot{y})} = \frac{Var\left(\sum_{j=1}^{k} \ln X_j^{\hat{b}_j}\right)}{Var(\ln y)}, \tag{15}$$

dove il pedice l si riferisce al modello lineare nei logaritmi e X_j e $\hat{b}_j$ indicano rispettivamente la j-esima colonna della matrice X ed il j-esimo elemento (scalare) del vettore stimato $\hat{b}$. Si nota immediatamente che l'equazione (15) è diversa dalla (6), quindi $R_l^2 \neq R^2$ ed il legame tra i due indici dipende esclusivamente dalla trasformazione logaritmica applicata. Aggiungo inoltre che questa proprietà è estendibile a qualsiasi trasformazione continua di variabili, ma quello che più interessa in quest'ambito è che un valore elevato di R^2 non implica necessariamente un valore elevato dell'omologo indice calcolato sul modello trasformato. Tutto dipende dalla funzione imposta. Più in dettaglio potrei dire che, se la trasformazione "migliora" il legame lineare tra variabile dipendente e regressori, allora $R_l^2 > R^2$. Ovviamente vale la relazione opposta se le variabili originali mostrano un legame lineare più forte rispetto a quello ottenibilie dopo la trasformazione.

In conclusione, posso solo dire che l'indice non ci è di nessun aiuto quando abbiamo intenzione di applicare trasformazioni alle variabili. E qui passo e chiudo.

2.4 Cambiare tutto per non cambiare niente

Ho già affermato in precedenza che, se vogliamo riparametrizzare il modello lineare (1), l'R^2 non cambia. Non ci credete? Bene, prima di tutto mettiamo in chiaro cosa vuol dire riparametrizzare. La riparametrizzazione non consiste in un cambio di modello, poiché questo resta sempre lo stesso; in pratica, la struttura lineare è mantenuta e le variabili esplicative utilizzate all'interno di essa cambiano semplicemente veste. Per chiarezza inserisco nella Tabella 1 un esempio piuttosto immediato relativo al potere esplicativo dell'istruzione dei genitori sul livello di istruzione dei figli (variabile dipendente `STUD`).

Tabella 1: Riparametrizzazione del modello lineare

```
Model 1:
OLS, using observations 1-7951 (n = 6571)
Missing or incomplete observations dropped: 1380
Dependent variable: STUD

          coeff.    std.err.   t-stat   p-value
-------------------------------------------------------
const    3.81654   0.104305    36.59   0.0000 ***
ETA     -0.02339   0.001172   -19.95   0.0000 ***
SEX     -0.25535   0.034109   -7.486   0.0000 ***
STUPCF   0.04672   0.043463    1.075   0.2824
PARAVG   0.76576   0.050742    15.09   0.0000 ***

Mean dep. var    3.866991  S.D. dep. var   1.675581
Sum sq. resid    12381.75  S.E. of regr.   1.373221
R-squared        0.328748  Adj. R-squared  0.328339
F(4, 6566)       803.9299  P-value(F)      0.000000
Log-likelihood  -11405.40  Akaike crit.    22820.80
Schwarz crit.    22854.75  Hannan-Quinn    22832.53
```

```
Model 2:
OLS, using observations 1-7951 (n = 6571)
Missing or incomplete observations dropped: 1380
Dependent variable: STUD

          coeff.   std.err.   t-stat   p-value
-------------------------------------------------------
const    3.81654   0.104305    36.59   0.0000 ***
AGE     -0.02339   0.001172   -19.95   0.0000 ***
SEX     -0.25535   0.034109   -7.486   0.0000 ***
FATHER   0.42960   0.022043    19.49   0.0000 ***
MOTHER   0.38288   0.025371    15.09   0.0000 ***

Mean dep. var    3.866991  S.D. dep. var   1.675581
Sum sq. resid    12381.75  S.E. of regr.   1.373221
R-squared        0.328748  Adj. R-squared  0.328339
F(4, 6566)       803.9299  P-value(F)      0.000000
Log-likelihood  -11405.40  Akaike crit.    22820.80
Schwarz crit.    22854.75  Hannan-Quinn    22832.53
```

Le variabili esplicative utilizzate nella stima sono l'età (AGE), il sesso (SEX, il valore 1 è asssegnato ai maschi), il titolo di studio rispettivamente del padre e della madre (FATHER e MOTHER) ed infine il titolo di studio medio dei genitori dato dalla variabile PARAVG=0.5(FATHER+MOTHER).

Per passare dal Modello 1 a quello riparametrizzato (Modello 2) è sufficiente un po' di algebra da scuola media, infatti

$$\begin{aligned}
\texttt{STUD}_i &= \beta_1 + \beta_2\texttt{AGE}_i + \beta_3\texttt{SEX}_i + \beta_4\texttt{FATHER}_i + \beta_5\texttt{PARAVG}_i + \varepsilon_i \\
&= \beta_1 + \beta_2\texttt{AGE}_i + \beta_3\texttt{SEX}_i + \beta_4\texttt{FATHER}_i + 0.5\beta_5(\texttt{FATHER}_i + \texttt{MOTHER}_i) + \varepsilon_i \\
&= \beta_1 + \beta_2\texttt{AGE}_i + \beta_3\texttt{SEX}_i + (\beta_4 + 0.5\beta_5)\texttt{FATHER}_i + 0.5\beta_5\texttt{MOTHER}_i + \varepsilon_i \\
&= \beta_1 + \beta_2\texttt{AGE}_i + \beta_3\texttt{SEX}_i + \beta_4^*\texttt{FATHER}_i + \beta_5^*\texttt{MOTHER}_i + \varepsilon_i. \qquad (16)
\end{aligned}$$

Osservando le due stime in Tabella 1 si nota immediatamente che i parametri stimati $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$ restano invariati, così come i rispettivi standard error, mentre $\hat{\beta}_4^* = \hat{\beta}_4 + \hat{\beta}_5^*$ e $\hat{\beta}_5^* = 0.5\hat{\beta}_5$; la mini-dimostrazione di cui alla (16) risulta perciò valida. Quello che cambia è la veste nella quale il modello viene presentato, ma le variabili esplicative utilizzate sono le stesse. Ciò è confermato inoltre dal valore della logverosimiglianza e di tutti i criteri informativi. In questa situazione non esiste alcun motivo serio per cui l'R^2 dovrebbe modificare il suo valore.

Tutto ciò mi sembra coerente: se il modello è lo stesso, anche l'indice deve essere lo stesso. La ragione di questo risultato risiede fondamentalmente nel potere informativo contenuto all'interno delle variabili esplicative che, ribadisco, rimangono sempre le stesse nei modelli proposti sopra e mantengono la relazione lineare con la variabile dipendente. In situazioni come queste l'R^2 non può perciò fornire alcun aiuto: quando siamo chiamati a scegliere tra diverse parametrizzazioni, dobbiamo muoverci in base di ciò che vorremmo evidenziare con la nostra stima ed affidarci quindi all'esperienza, all'intuito o addirittura al semplice colpo d'occhio.

A questo proposito direi che frasi celebri di *Gattopardiana* memoria calzano proprio a pennello.

2.5 L'R^2 non normalizzato

Ammettiamolo: già il titolo di questa sezione mette un po' di inquietudine a tutti quelli che, dopo aver affrontato e digerito un bel corso di statistica, credevano che l'indice di determinazione godesse sempre, comunque e dovunque della nota proprietà $R^2 \in [0, 1]$. In realtà, come ho già spesso ricordato, la proprietà di normalizzazione dell'indice è sempre garantita nel caso di cui alle equazioni (6)-(7), a patto che la dinamica della variabile dipendente venga spiegata attraverso un modello lineare. In questo caso direi: "niente paura".

Esistono tuttavia diversi casi in cui possiamo avere valori dell'indice al di fuori dell'intervallo $[0, 1]$. Fondamentalmente dividerei queste violazioni del principio generale in due macro-categorie: indice R^2 relativi a modelli di regressione non lineare ed indici R^2 *ad hoc*.

Nel primo caso la dichiarata assenza di linearità[3] fa decadere la legge secondo la quale occorre dedicare attenzione alla minimizzazione della somma dei quadrati dei residui della regressione; spesso tali residui derivano da più o meno complesse trasformazioni di variabili e perciò potrebbero non avere le usuali caratteristiche di valore atteso nullo e di varianza costante oppure addirittura l'applicazione delle difinizioni (6)-(7) potrebbe non essere possibile o non avere senso.

Nel secondo caso, che chiaramente costituisce una logica conseguenza del primo, l'elaborazione di particolari indici R^2 si è curata poco del fatto che la normalizzazione degli stessi sia l'aspetto più importante da conservare. Qui potrei sbizzarirmi con gli esempi, ma questo costringerebbe il lettore a studiarsi tanta, troppa letteratura. Il mio obiettivo adesso è solamente quello di stimolare la curiosità, quindi mi limito a suggerire alcuni nomi suggestivi forniti per lo più dai modelli logit/probit:

[3]Qui si potrebbe fare un *excursus* sui modelli non lineari di regressione, ma voglio limitarmi a segnalarne qualcuno: si pensi pertanto a modelli del tipo $y_i = \alpha \exp\{X_i'\beta\}$, alle famigerate funzioni Cobb-Douglas $y_i = A\prod_{j=1}^{n} x_{j,i}^{\beta_j}$ oppure ai vari modelli logit, probit o regressioni di Poisson.

mi riferisco perciò allo Pseudo-R^2, sempre strettamente minore di 1, oppure all'R^2 di previsione.[4] Naturalmente, esistono anche indicatori *ad hoc* normalizzati come ad esempio l'R^2 di McFadden, ma questa è un'altra storia...

2.6 Modelli non lineari (R^2 basso): un esempio

Dato che ormai ho introdotto i modelli non lineari, volevo concentrare l'attenzione sul fatto che molti di essi sono reputati come largamente migliori dei modelli lineari, anche se mostrano valori dell'indice R^2 decisamente inferiori.

Un esempio, alternativo a quello ormai abusato dei logit e dei probit, potrebbe essere costituito dai modelli di serie storiche per la volatilità condizionale o semplicemente modelli GARCH (Generalised Auto Regressive Conditional Heteroskedasticity, si veda ad esempio: Engle, 1982; Bollerslev, 1986), utilizzati a tal punto nei lavori di finanza operativa, tanto da garantire a colui che li ha inventati, Robert F. Engle, il Premio Nobel nel 2003. Senza addentrarmi particolarmente nei tecnicismi, potrei sintetizzare i modelli di questo tipo come un sensibile miglioramento apportato ai modelli lineari del tipo[5]

$$y_t = x_t'\beta + \varepsilon_t \tag{17}$$

perché specificano una legge di moto per la varianza condizionale $\sigma_t = E(\varepsilon_t^2|I_{t-1})$, dove il set informativo I_{t-1} contiene al proprio interno tutti i valori precedenti al tempo t della variabile dipendente e di tutte le (eventuali) variabili esplicative. Credo sia ovvio che, se la varianza condizionale è indicizzata rispetto al tempo, essa non rispetti affatto la condizione di omoschedasticità $E(\varepsilon_t^2|I_{t-1}) = \sigma^2$, quindi l'informazione che deriva dallo stimare i valori di σ_t per ogni t sia nettamente maggiore rispetto a quella fornita da un parametro assunto (talvolta erroneamente) come costante nel tempo. Dal punto di vista statistico una migliore informazione corrisponde ad una maggiore efficienza dello stimatore in questione; se aggiungo inoltre che i modelli GARCH sono stimati attraverso il metodo della massima verosimiglianza[6] non è difficile immaginare che questo stimatore goda anche delle fondamentali proprietà della consistenza e della distribuzione asintotica normale, nonostante viva in un mondo ostile nel quale il concetto di linearità non si applica.

Traduco in parole povere: se applico un modello GARCH invece di un semplice OLS ottengo sempre uno stimatore consistente, asintoticamente normale, ma più efficiente. Poi osservo l'R^2 che in questi casi è notoriamente molto basso. Cosa faccio? Butto via tutto?

A queste domande rispondo con un'altra domanda: preferireste una stima più efficiente con R^2 basso oppure una inefficente con R^2 elevato?

2.7 Maledette regressioni spurie!

Rimaniamo nell'ambito delle serie storiche. Stavolta voglio illustrare il problema con un esercizio applicato attraverso il quale effettuo una regressione OLS dell'indice della produzione industriale messicana (`mexip`) su due variabili esplicative date dal PIL della Corea del Sud (`korgdp`) ed il tasso di disoccupazione in Svizzera (`swiur`). Per completezza aggiungo che i dati sono trimestrali e vanno dal primo trimestre del 1984 al secondo trimestre del 2008.

Mentre vi state chiedendo che razza di economista io sia oppure in quale stato confusionale sia intrappolata la mia mente, vi propongo l'output di stima in Tabella 2.

L'intestazione della tabella parla chiaramente di regressione spuria e sono sicuro che il lettore attento o istruito su questo problema abbia già capito dove voglio concentrarmi. Adesso mi rivolgo perciò solo a coloro che, essendo meno esperti, si limitano a notare che

[4] Per chi volesse saperne di più suggerisco caldamente il libro di Manera e Galeotti (2005).

[5] Si noti che in questo contesto si parla di osservazioni rilevate nel tempo quindi, osservando l'equazione (1), si può notare che formalmente sono cambiati solo i pedici.

[6] Miliardi di libri di statistica o di econometria parlano di questo stimatore, quindi la scelta è amplissima; timidamente suggerisco perciò Palomba (2010).

Tabella 2: Regressione spuria

```
OLS, using observations 1984:1-2008:2 (T = 98)
Dependent variable: mexip

            coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      33.5099        1.33760       25.05     0.0000 ***
  korgdp      0.0004        1.7e-05       23.71     0.0000 ***
  swiur      -1.0007        0.58276       -1.72     0.0892 *

Mean dependent var   78.26531   S.D. dependent var   18.27667
Sum squared resid    2219.030   S.E. of regression    4.83303
R-squared             0.93151   Adjusted R-squared    0.93007
F(2, 95)            646.0796   p-value (F)           0.00000
Log-likelihood     -291.9290   Akaike IC            589.8580
Schwarz IC          597.6130   Hannan-Quinn IC      592.9947
rho                   0.95553   Durbin-Watson (DW)    0.09828
```

- i coefficienti della regressione sono tutti statisticamente significativi,
- $R^2 = 0.93151$ e $\bar{R}^2 = 0.93007$.

Senza divagare sulle proprietà caratteristiche delle regressioni spurie, vorrei far ragionare queste persone su alcuni aspetti piuttosto evidenti. In primo luogo esaminerei la significatività dei coefficienti: se ammettiamo la validità dei risultati in Tabella 2, dovremmo altresì ammettere che la produzione industriale in Messico dipenda fortemente dal PIL sudcoreano e in misura più lieve[7] dal tasso di disoccupazione in Svizzera. Adesso sfido chiunque a trovare una giustificazione economica sia per il modello stimato, sia per il risultato ottenuto.

A questo punto potreste ribattere facendomi notare che, seppure il quadro economico delineato dal modello risulti piuttosto bizzarro, effettivamente il modello interpola bene i dati perché addirittura due indici di determimazione si attestano su un incoraggiante 93%. Molto bene, ma io alzerei la voce affermando che nelle regressioni spurie tutto ciò è all'ordine del giorno perché ho a che fare con serie storiche che non hanno motivo di "muoversi insieme", ma di fatto mostrano una struttura di correlazione imponente.

Come è possibile tutto ciò? Studiando adeguatamente il problema scoprirete che state assistendo ad un effetto collaterale generato dal fatto che le serie non sono stazionarie e nell'analisi di serie storiche è appurato che la non stazionarietà manda all'aria tutti i teoremi limite relativi all'OLS.[8] Questa caratteristica spesso dà luogo a tre situazioni totalmente inaspettate quando si mettono in relazione lineare variabili che nella realtà non dovrebbero avere alcuna relazione l'una con l'altra. In estrema sintesi nelle regressioni spurie avviene che:

1. l'R^2 non assume valori "bassi",
2. i coefficienti stimati non assumono valori vicini allo zero,
3. i test t condotti per l'azzeramento di tali parametri rifiutano l'ipotesi nulla, specialmente in grandi campioni.

A questo punto è facile rendersi conto che coefficienti significativi e valori elevati di R^2 sono numeri che non hanno un senso, non solo dal punto di visto economico, ma anche da quello statistico.

[7] È evidente che questa affermazione dipende dal valore critico scelto quando si effettua il test t per la variabile relativa *swiur*.

[8] Vorrei qui aggiungere che la stazionarietà di una serie storica è testabile, quindi la presenza di questa caratteristica dovrebbe sempre essere valutata prima di effettuare qualsiasi operazione di stima. Sto parlando dei test di radice unitaria dei quali attualmente esiste un'abbondante letteratura.

Ma come si fa a riconoscere una regressione spuria? Tralasciando il fatto che prima di tutto ci vuole buon senso quando si selezionano le variabili in un modello OLS, direi che un bel test preventivo di radice unitaria sulle serie già ci dovrebbe un'informazione importante. Più precisamente: se le serie sono stazionarie il mondo OLS è salvo e non ci sono problemi, neppure ad utilizzare gli indici R^2; viceversa, la presenza di non stazionarietà non indica necessariamente che la regressione è spuria, ma almeno dovrebbe generare il sospetto. E poi c'è una utile regola del pollice in base alla quale si ha una regressione spuria quando l'indice R^2 risulta maggiore del valore assunto dalla statistica di Durbin e Watson (o semplicemente DW, 1950):[9] partendo dal presupposto che, quando le cose "vanno bene", dovrebbe risultare $R^2 \approx 1$ e DW≈ 2, è chiaro che una situazione tale per cui risulta $R^2 >$DW dovrebbe insinuare nella mente dell'analista legittimi dubbi.

Adesso concentro di nuovo l'attenzione sulla Tabella 2 e, forte della regoletta appena introdotta, osservo in un attimo che la regressione è effettivamente spuria. Cosa posso fare per spiegare le relazioni tra le 3 variabili? Tecnicamente c'è poco da fare: dapprima differenzio le serie poi stimo lo stesso modello OLS sulle serie differenziate.[10] Trovate tutto nella Tabella 3 con le nuove variabili `dmexip`, `dkorgdp` e `dswiur`.

Tabella 3: Modello finale

```
OLS, using observations 1984:2-2008:2 (T = 97)
Dependent variable: dmexip

             coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       0.621143      0.194959       3.186     0.0020  ***
  dkorgdp    -2.85e-06      8.20e-05      -0.035     0.9724
  dswiur     -1.039910      0.837748      -1.241     0.2176

Mean dependent var   0.591753   S.D. dependent var   1.356748
Sum squared resid    173.8603   S.E. of regression   1.359992
R-squared            0.016145   Adjusted R-squared  -0.004788
F(2, 94)             0.771273   P-value (F)          0.465328
Log-likelihood      -165.9388   Akaike IC            337.8776
Schwarz IC           345.6017   Hannan-Quinn IC      341.0009
rho                  0.444251   Durbin-Watson (DW)   1.107156
```

Nella Tabella 3 il quadro della situazione cambia drasticamente: il modello OLS stimato impiega solo variabili stazionarie (fidatevi), quindi i suoi risultati in questo caso sono sicuramente più affidabili, infatti:

1. i coefficienti stimati non sono statisticamente significativi e questo, dal punto di vista economico, dovrebbe tranquillizzarci,
2. $R^2 = 0.016145$ e addirittura abbiamo $\bar{R}^2 < 0$: tutto è coerente,
3. $R^2 <$DW, come è normale che sia.

3 La dignità di un indice

Bisogna ammettere che, nella sezione 2, l'indice R^2 esce un po' con le ossa rotte, quindi mi sembra opportuno in questa sede salvare tutto quello che di lui è salvabile; è una questione di stile e (perché no?) di rispetto.

[9] In breve, la statistica DW è un indicatore immediato per valutare l'assenza di autocorrelazione di ordine 1 nei residui di una regressione lineare.

[10] Differenziare una serie storica y_t significa determinare la serie storica $\Delta y_t = y_t - y_{t-1}$.

3.1 Il fantastico mondo lineare

In questa sezione, come nella prossima, sarò telegrafico perché il problema della linearità del modello l'ho già introdotto e discusso in precedenza. Il mondo lineare rappresenta l'habitat naturale dell'indice R^2 nel quale può esprimere tutta la sua valenza dal punto di vista informativo. Ho già detto che quest'indice è stato creato per modelli lineari di regressione ed in tale ambito raggiunge il suo massimo splendore.

La questione perciò è un'altra: la dinamiche dell'economia sono davvero rappresentabili attraverso il fantastico mondo lineare? La risposta a tale domanda, che può apparire decisiva per le sorti dell'analisi economica e/o dell'econometria in generale, in realtà va fornita caso per caso oppure va ricercata attraverso test statistici più o meno complessi. Sta di fatto che spesso per molti fenomeni ci si accontenta di fornire un'approssimazione lineare oppure si effettuano ipotesi di linearità più o meno ardite. È un costo da pagare per avere la possibilità di applicare un modello statistico/econometrico che sia facilmente maneggiabile dal punto di vista analitico, magari anche facilmente stimabile, ma soprattutto facilmente interpretabile dal punto di vista economico. Adesso però mi rendo conto che questi problemi hanno natura più filosofica che tecnica, quindi sto andando fuori tema. Mi limito perciò alla seguente considerazione: di sicuro, non si propende per una specificazione lineare con l'obiettivo di massimizzare l'R^2.

3.2 Derivati e generalizzazioni

Facciamo il conto della serva oltre all'R^2 classico o ufficiale: esistono ad esempio l'R^2 non centrato e l'R^2 corretto per i g.d.l., come ampiamente descritto nella sezione 2.1, l'R^2 within, l'R^2 between e l'R^2 totale nei modelli panel data, l'R^2 di previsione, l'R^2 di McFadden, lo pseudo-R^2, l'R^2 di Cox e Snell o l'R^2 di Negekerke nei modelli con variabili dipendenti qualitative, ecc. Di sicuro non ne ho nominati molti altri, ma già questo elenco fa capire che, l'indice non solo ha una certa importanza o utilità, ma la filosofia su cui esso si basa è riconosciuta come universalmente valida, data la proliferazione di altri indicatori ad esso omologhi. Se è vero che "l'imitazione è la più sincera forma di ammirazione", direi proprio che siamo sulla buona strada.

3.3 Il cacciatore di collinearità

Spesso, a seguito di una stima effettuata applicando il metodo OLS, può capitare di imbattersi in una situazione tale per cui risulta che:

1. l'R^2 assume valori elevati, ma i test t di azzeramento dei parametri segnalano che le variabili esplicative non sono statisticamente significative in quanto gli errori standard stimati sono elevati. In termini più formali, ciò significa che gli intervalli di confidenza intorno ai coefficienti di regressione sono piuttosto ampi;

2. le variabili esplicative del modello sono tra loro fortemente correlate.

Dal punto di vista algebrico, una correlazione molto elevata tra due o più variabili esplicative si potrebbe ripercuotere nel fatto che le colonne della matrice X non siano tutte linearmente indipendenti e qui gli amanti del genere affermerebbero che il rango-colonna di X non è pieno. Un tale scenario preclude l'utilizzo dello stimatore OLS definito dall'equazione (2) poiché il prodotto $X'X$ configura una matrice con rango non pieno o, meglio ancora, singolare (non invertibile). Un vero incubo.

D'accordo, sto parlando del famoso problema della collinearità in presenza del quale il valore dell'R^2 evidentemente non individua alcuna bontà di adattamento del modello ai dati. In questa situazione l'indice trova invece la sua esaltazione; si consideri pertanto un modello di regressione del tipo

$$X_i = \sum_{j \neq i} \gamma_j X_j + u_i, \tag{18}$$

dove la variabile dipendente e tutti i regressori rappresentano le colonne della matrice X di cui parlavo poc'anzi, mentre i γ_j sono semplicemente coefficienti. In pratica sto cercando di capire se effettivamente le colonne di tale matrice sono effettivamente correlate, quindi posso calcolarmi un indice R_i^2 per ogni $i = 1, 2, \ldots, k$. Poiché il concetto di correlazione identifica per definizione l'esistenza del fantastico mondo lineare, il potere informativo dei vari R_i^2 è massimo per costruzione. Traduco: se l'i-esimo indice è prossimo al valore 1, allora l'i-esima variabile contenuta in X è collineare, il gioco è semplice. Non ci credete?

Allora vi informo che esiste un indicatore immediato di collinearità che si chiama *Variance Inflation factor* (per gli amici VIF) il quale è così definito

$$\text{VIF}_i = (1 - R_i^2)^{-1}. \tag{19}$$

Si nota immediatamente che l'indice di determinazione è decisivo ed è in relazione diretta con il VIF: nella prassi, solitamente si associa la presenza di collinearità ad un VIF> 10. Questo accade quando $R_i^2 > 0.9$. Credo che ulteriori spiegazioni non servano.

3.4 Test di specificazione e diagnostica

In questa sezione mostrerò finalmente che l'indice R^2 è davvero uno strumento indispensabile in certi contesti. In particolare mi riferisco alla moltitudine di test di specificazione e/o diagnostica che solitamente vengono inflitti agli studenti in un qualsiasi corso di econometria.

La lista è lunga e contiene numerosi contributi illustri: si va dal test RESET al test ai *Conditional Moment Test*, dal test di sovraidentificazione di Sargan a quello di Hausman, dai test di variabili omesse a quelli di autocorrelazione tra i quali mi piace citare il test di Breusch-Godfrey tanto per rendere l'idea. Infine ci sono parecchie procedure di valutazione per la presenza di eteroschedasticità nei dat e qui non posso evitare il riferimento al test di White o al test ARCH.

La letteratura in quest'ambito è pressoché sconfinata e sicuramente non ho fatto menzione a diversi contributi illustri (*mea culpa*) poiché tutti questi test hanno in comune il fatto di consistere fondamentalmente in un classicissimo test dei moltiplicatori di Lagrange (test LM) nel quale la statistica test è data da

$$\text{LM} = nR^2, \tag{20}$$

dove l'R^2 è quello relativo ad una specifica regressione ausiliaria di interesse. Poiché, per definizione, la statistica test (20) si distribuisce asintoticamente come una tranquillizante chi-quadro, è facile comprendere che l'utilizzo di tali procedure è assolutamente di pubblico dominio.

In questo contesto (o dovrei dire in questi contesti?) l'indice R^2, benché moltiplicato per una costante n che generalmente equivale alla numerosità del campione, conquista la sua vera gloria in quanto permette all'analista di fare inferenza statistica identificandosi di fatto in una statistica test.

4 Conclusioni

Cos'è davvero l'R^2? Spero vivamente che queste pagine che ho scritto servano quantomeno da avvertimento per l'analista incauto che crede di aver effettuato un buon lavoro solo perché tale indice è vicino al valore unitario. Se da un lato è vero che lo stesso indice si è creato una reputazione, dall'altra gli "addetti ai lavori" diffidano spesso dei risultati che esso fornisce. Il fatto che goda di un esagerato credito presso troppi utilizzatori della scienza statistico-econometrica deriva più da ragioni legate alla semplicità del suo approccio, dalla notorietà presso il grande pubblico (chi non ha studiato l'R^2 come mezzo di valutazione *ex post* di una stima OLS?) e anche perché oggi praticamente non esiste alcun sofware che non ne pubblichi il valore in coda ad una stima appena effettuata. La diffidenza verso i risultati che esso fornisce a mio avviso già dovrebbe nascere osservandone il larghissimo utilizzo, ma mi rendo conto che questa non è assolutamente una ragione valida per boicottare niente e nessuno. Per questa ragione ho deciso di discutere nel dettaglio i diversi aspetti di quest'indicatore, sia positivi che negativi, e la conclusione a cui giungo è sostanzialmente un monito: FATE ATTENZIONE!

L'R^2 è solo un indicatore, nulla di più e questo occorre tenerlo sempre a mente. Nonostante sotto ipotesi molto stringenti, mi riferisco soprattutto alla linearità del meccanismo generatore dei dati, addirittura può diventare sopraffino. La sua valenza informativa va quindi sempre valutata in riferimento al contesto in cui esso è calcolato: così facendo vi renderete subito conto che l'R^2 deve essere considerato semplicemente come un'utile informazione accessoria da affiancare ad analisi più accurate. Fidarsi troppo dei suoi valori è un po' come pretendere di giudicare il sapore di un piatto osservandone solo l'aspetto o i colori: è chiaro che siamo in grado di attingere alcune informazioni, ma rischiamo che queste siano superficiali o quantomeno distorte rispetto alla varietà di sapori che scopriremmo dopo l'assaggio.

Riferimenti bibliografici

Bollerslev, T. (1986). *Generalized autoregressive conditional heteroskedasticity.* Journal of Econometrics, 31: 307–327.

Durbin, J. e Watson, G. (1950). *Testing for serial correlation in least squares regression.* Biometrika, (37): 409–428.

Engle, R. F. (1982). *Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation.* Econometrica, 50(4): 987–1007.

Manera, M. e Galeotti, M. (2005). *Microeconometria.* Carocci.

Palomba, G. (2010). *Elementi di statistica per l'econometria.* CLUA libri, Ancona. 2^a edizione.