La regressione lineare applicata a dati economici

Matteo Pelagatti

7 febbraio 2008

Indice

Sommario

In questa breve dispensa si pongono le basi per l'utilizzo della regressione su dati economici reali. Si comincia con la definizione del modello lineare e si affronta intuitivamente il problema della sua stima sotto le ipotesi classiche. Dato che tali ipotesi molto spesso vengono violate quando si lavora su dati reali, soprattutto quando in forma di serie storica, per ogni ipotesi classica "a rischio" si

- forniscono le tecniche per verificare la compatibilità delle ipotesi con i dati,
- indicano le conseguenze sulle stime e sui test forniti dai software di regressione delle violazioni di tali ipotesi,
- indicano le soluzioni per potere analizzare i dati anche in assenza di alcune delle ipotesi classiche.

Il livello matematico della discussione è molto basso e accessibile, al lettore si richiede solamente di avere almeno un'idea di che cosa sia uno stimatore e una statistica test.

1 Il modello lineare

Supponiamo che la relazione tra un fenomeno, misurato dalla variabile y e k fenomeni, misurati dalle variabili $x_1, \ldots, x_k$ sia rappresentata, o almeno approssimata, dal modello lineare

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_p + \varepsilon, \tag{1}$$

dove ε è una quantità casuale a media nulla che "sporca" la relazione, altrimenti deterministica, tra la variabile dipendente y e i regressori $x_1, \ldots, x_k$.

Si noti che il modello è lineare nei parametri $\beta_0, \ldots, \beta_k$ e non necessariamente nelle variabili, che infatti possono avere subito trasformazioni precedenti.

Per fare un esempio, si pensi alla possibile relazione tra consumo c e reddito r (che analizzeremo approfonditamente nel corso). È possibile che la relazione sia lineare nei livelli

$$c = \beta_0 + \beta_1 r + \varepsilon$$

o dopo una trasformazione logaritmica di una o entrambe le variabili:

$$\begin{aligned} c &= \beta_0 + \beta_1 \ln r + \varepsilon, \\ \ln c &= \beta_0 + \beta_1 \ln r + \varepsilon. \end{aligned}$$

In entrambi i casi la forma (1) è preservata.

Per condurre un'analisi statistica è necessario procurarsi un campione della relazione (1), collezionando le $k+1$-uple $(y_t, x_{1,t}, \ldots x_{k,t})$, per $t = 1, 2, \ldots, n$, dove n è l'ampiezza campionaria. In questo modo abbiamo n osservazioni della medesima relazione

$$y_t = \beta_0 + \beta_1 x_{1,t} + \ldots \beta_k x_{k,t} \qquad t = 1, 2, \ldots, n,$$

ed è quindi possibile fare inferenza sui coefficienti ignoti $\beta_0, \beta_1, \ldots, \beta_k$.

2 La stima dei coefficienti e le ipotesi classiche

Il metodo più frequentemente usato per stimare i coefficienti ignoti è il *metodo dei minimi quadrati*, detto anche *dei minimi quadrati ordinari* (OLS = Ordinary Least Squares). Le stime OLS dei coefficienti di regressione sono date da quei valori dei coefficienti che risolvono il seguente problema di minimizzazione

$$\min_{\beta_0,\ldots,\beta_k} = \sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1 x_{1,t} - \ldots - \beta_k x_{k,t}\right)^2,$$

e che chiameremo $\hat{\beta}_i$, $i = 1, \ldots, n$.

In figura 1 è rappresentata la retta di regressione OLS dei consumi finali pro capite (Cons) sul reddito nazionale lordo pro capite (PNL). La retta OLS è quella che tra tutte le rette nel piano minimizza la somma dei quadrati delle distanze verticali tra la retta stessa e i punti campione.

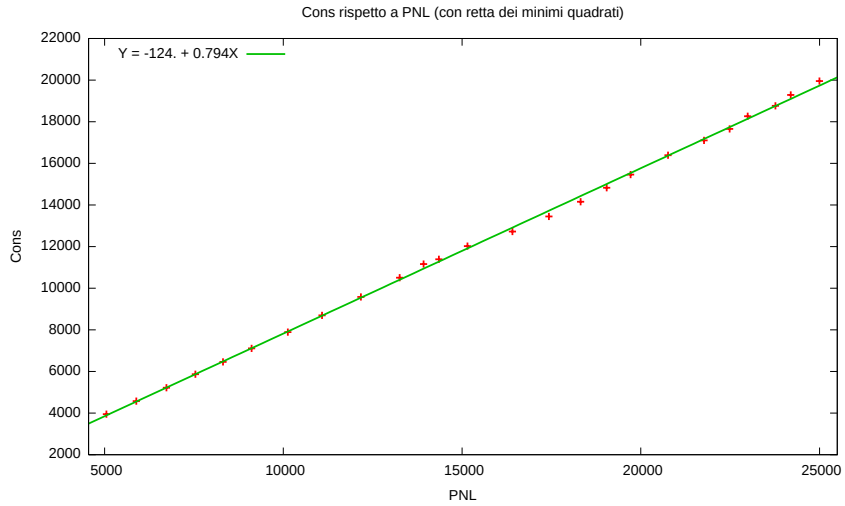

Figura 1: Prodotto nazionale lordo pro capite a prezzi correnti (ascisse), consumi finali pro capite a prezzi correnti (ordinate) e retta stimata con gli OLS.

Le proprietà statistiche degli stimatori OLS sono determinate dalle assunzioni sulla parte stocastica (casuale) del modello, cioè di ε. In realtà le ipotesi classiche, nate e pensate in situazioni in cui i regressori erano controllati dal ricercatore (per es. quantità di un farmaco somministrato su cavie), suppongono che le $x_1, \ldots, x_k$ siano non stocastiche. Questa ipotesi, tuttavia, non ha senso quando si lavora con dati economici in cui tutta la $k+1$-upla è estratta casualmente (per es. si estrae un campione di n residenti in Italia e si chiede loro a quanto ammonta il loro reddito e quanto ne spendono in consumi finali). Quando anche i regressori sono stocastici le proprietà degli stimatori dipendono anche dalla distribuzione dei regressori, tuttavia se, come si suole fare, ci si limita ad una analisi condizionale ai valori dei regressori osservati, l'unica fonte di variabilità torna a essere l'errore di regressione ε. Le ipotesi classiche sono:

Media nulla $\mathbb{E}[\varepsilon_t] = 0$, per ogni t,

Omoschedasticità (= varianza costante) $\mathrm{Var}[\varepsilon_t] = \sigma^2$,

Indipendenza seriale ε_t e ε_t indipendenti per ogni t e s con $s \neq t$,

Indipendenza con i regressori ε_t e $x_{i,s}$ indipendenti per ogni i, s, t (questa assunzione è ovviamente vera quando le x_i non sono stocastiche),

Normalità ε_t è normalmente distribuito per ogni t.

Quando tutte queste assunzioni sono vere, lo stimatore OLS dei coefficienti di regressione gode delle seguenti proprietà:

Consistenza la probabilità che la stima OLS $\hat{\beta}_i$ disti dal valore vero β_i per più di una costante positiva arbitraria κ converge a zero al crescere della numerosità del campione, in formule $\lim_{n\to\infty} \Pr\{|\hat{\beta}_i - \beta_i| > \kappa\} = 0$ per ogni $\kappa > 0$,

Correttezza $\mathbb{E}[\hat{\beta}_i] = 0$ (lo stimatore in media "ci prende" e non tende a sovrastimare o sottostimare),

Efficienza non esiste alcuno stimatore corretto $\tilde{\beta}_i$ che abbia varianza più piccola dello stimatore OLS, in formule $\mathrm{Var}(\hat{\beta}_i) \leq \mathrm{Var}(\tilde{\beta})$,

Normalità condizionatamente alle $x_1, \ldots, x_k$ estratte, la distribuzione dello stimatore OLS $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ è normale multivariata.

Si noti che le stime e le statistiche che forniscono i software di statistica (almeno di default) sono basate sulle assunzioni classiche. In Tabella 1 è riportato l'output della procedura OLS di Gretl, che può essere considerato tipico per quanto concerne i pacchetti statistici che implementano il modello di regressione lineare. Nei software di regressione le statistiche test, le statistiche t, sono costruite per l'ipotesi $\beta_i = 0$, cioè che il coefficiente i-esimo sia nullo e quindi il relativo regressore sia da escludere dalla regressione (regressore non statisticamente significativo). Le statistiche t e il relativo p-value[1] sono riportate in Tabella 1. Se le ipotesi classiche fossero rispettate per i dati a cui si riferiscono le stime in tabella, allora il regressore PNL sarebbe da tenere ($0.0000 < 0.05$), mentre la costante potrebbe essere esclusa dal modello ($0.1038 > 0.05$).

L'output consiste di molte altre statistiche, alcune della quali ci aiutano a capire se le ipotesi classiche valogono per i nostri dati oppure no. Alcune di queste saranno discusse più avanti.

3 Le conseguenze del venir meno di alcune ipotesi classiche

Quando si lavora su dati reali, ed in particolare su serie storiche, alcune delle ipotesi classiche possono non essere più valide. Dopo avere stimato una regressione, è quindi importante verificare quali delle ipotesi classiche sembrano venir meno e ricordarsi le conseguenze di questo sugli stimatori e gli eventuali rimedi.

[1] Si ricorda che una statistica test è una funzione dei dati che misura la distanza di questi dall'ipotesi nulla e di cui si conosce la distribuzione (almeno per grandi campioni). Per esempio, nel caso in cui si voglia testare se un certo coefficiente di regressione sia nullo, la statistica è $t = \hat{\beta}_i / \hat{\sigma}_{\beta_i}$, che ovviamente sarà tanto più distante da zero quanto β_i sarà grande in valore assoluto. La divisione per la stima dell'errore standard di $\hat{\beta}_i$, cioè una misura della variabilità dello stimatore, è necessaria per conoscere la distribuzione della statistica test, che in questo caso è una t di Student con $n - k - 1$ gradi di libertà.

Il *p-value* è la probabilità di ottenere un valore della statistica test almeno così estremo quanto quello ottenuto sui dati, quando è vera l'ipotesi nulla. Se il p-value è piccolo (tipicamente più piccolo di 0.05), allora is rifiuta l'ipotesi nulla dato che sarebbe molto raro (meno del 5% dei casi) ottenere una statistica test come quella ottenuta se fosse vera l'ipotesi nulla. In questo caso, se l'ipotesi nulla è "l'i-esimo coefficiente di regressione è nullo", allora per un p-value della statistica test maggiore di 0.05 (risp. 0.01) si dice che il regressore x_i *è significativo* al 5% (risp. 1%).

Modello 1: Stime OLS usando le 25 osservazioni 1982–2006
Variabile dipendente: Cons

Variabile	Coefficiente	Errore Std.	statistica t	p-value
const	-123.78	73.0721	-1.6940	0.1038
PNL	0.794485	0.00440894	180.1989	0.0000

Media della variabile dipendente	12097.9
D.S. della variabile dipendente	5003.26
Somma dei quadrati dei residui	425239.
Errore standard dei residui ($\hat{\sigma}$)	135.973
R^2	0.999292
$\bar{R}^2$ corretto	0.999261
Gradi di libertà	23
Statistica Durbin-Watson	0.608739
Coefficiente di autocorrelazione del prim'ordine	0.715800
Log-verosimiglianza	-157.24
Criterio di informazione di Akaike	318.485
Criterio bayesiano di Schwarz	320.923
Criterio di Hannan-Quinn	319.161

Tabella 1: Output di Gretl per la procedura OLS.

La validità delle ipotesi classiche viene verificata sugli errori di regressione stimati

$$e_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1,t} - \ldots - \hat{\beta}_1 x_{k,t},$$

che sono rappresentati in Figura 2 per la regressione stimata nel precedente paragrafo.

3.1 Non-gaussianità

Un'ipotesi che spesso viene violata è quella della normalità degli errori di regressione, che tendono a essere *leptocurtici* (le osservazioni estreme sono più probabili rispetto a quanto atteso sotto l'ipotesi di normalità). I software statistici spesso hanno la possibilità di testare l'ipotesi di normalità degli errori di regressione. I test possono essere di vario tipo (Jarque-Bera, Kolmogorov-Smirnov, Cramer-Von Mises, Chi-Quadrato, ecc.), tuttavia all'utente sarà sufficiente osservare il p-value della statistica test implementata nel package, tenendo in mente che l'ipotesi nulla è che gli errori di regressione provengano da una distribuzione normale.

In Figura 3 è mostrato l'istogramma degli errori di regressione del modello esemplificativo del paragrafo precedente, sovrapposto alla densità normale con varianza stimata sui dati. Inoltre è riportato il p-value per l'ipotesi di normalità, che essendo maggiore del 5% (p-value = 0.60 > 0.05) fa concludere che l'ipotesi di normalità è compatibile con i nostri dati.

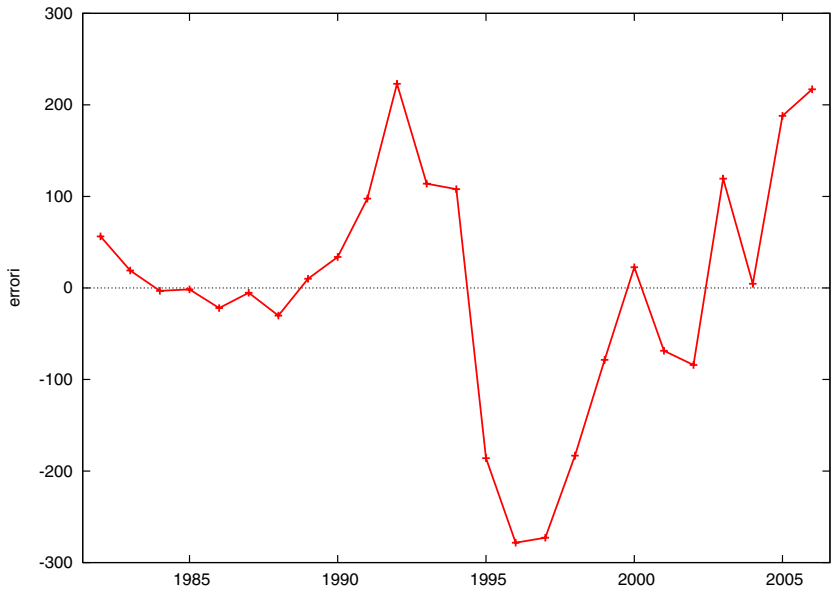

Figura 2: Errori della regressione dei consumi pro capite sui redditi pro capite.

Quando non ho più la normalità degli errori di regressione, gli OLS non sono più efficienti e la loro distribuzione non è più nota; tuttavia per campioni sufficientemente ampi (grosso modo per $n \geq 30$), la normalità rimane una buona approssimazione. La presenza di valori estremi rende la variabilità degli stimatori piuttosto alta e quindi le stime sono meno affidabili.

Per quanto detto, i test di siginificatività forniti dai software di regressione rimangono approssimativamente validi e l'approssimazione migliora con il crescere della numerosità campionaria.

3.2 Eteroschedasticità

Quando la varianza dell'errore di previsione non è costante per tutte le osservazioni, gli OLS non sono più efficienti e gli errori standard degli stimatori sono stimati in maniera impropria (non sono consistenti). Ciò comporta che anche le statistiche t e i relativi p-values siano imprecisi.

Non sempre i software forniscono test per l'ipotesi di omoschedasticità, tuttavia se si osserva il grafico degli errori di regressione stimati non è difficile rendersi conto di una eventuale tendenza nella variabilità degli errori. Gretl mette a disposizione il test di omoschedasticità di White, il cui output basato sui dati usati in precedenza è illustrato nella Tabella 2 L'ipotesi di omoschedasticità non può essere rigettata (0.067 > 0.05), benché il p-value sia piuttosto vicino al valore critico.

Quando ci si trova in presenza di eteroschedasticità, gli errori standard degli stimatori OLS possono essere stimati consistentemente. Alcuni software permet-

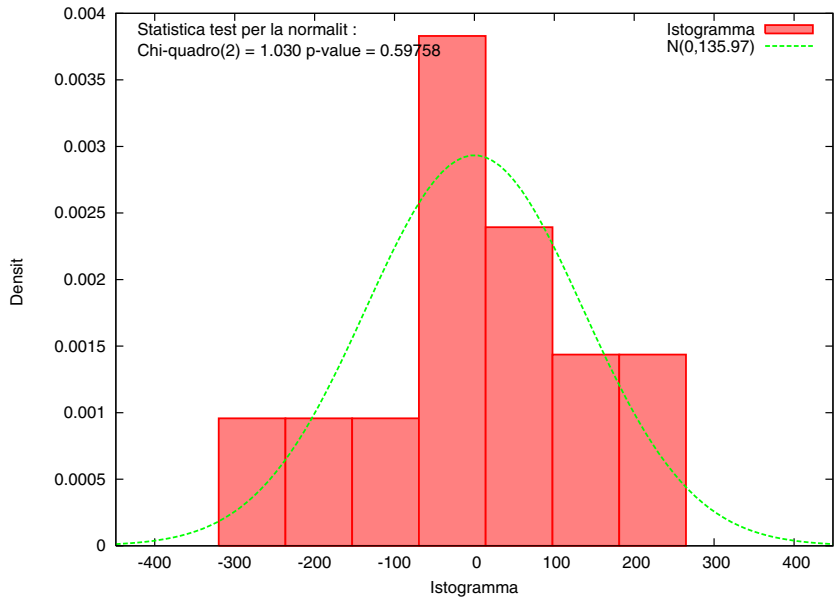

Figura 3: Istogramma, densità normale e test di normalità.

```
Test di White per l'eteroschedasticità -
  Ipotesi nulla: eteroschedasticità non presente
  Statistica test: TR^2 = 5.39406
  con p-value = P(Chi-Square(2) > 5.39406) = 0.0674055
```

Tabella 2: Output Gretl per il test di omoschedasticità di White.

tono questa correzione che è basata su una diversa stima della matrice di covarianza degli stimatori indicata con i seguenti nomi:

- matrice di covarianza di White
- matrice di covarianza robusta (o errori standard robusti)
- matrice HC (heteroskedasticity-consistent = eteroschedasticità-consistente)

Gretl mette a disposizione diverse versioni di tale correzione (HC0, HC1, HC2, HC3), la cui discussione va ben oltre il livello di questa dispensa. La parte rilevante dell'output della regressione errori standard HC1 è riportata nella Tabella 3.

Come avrete notato, le stime dei coefficienti di regressione sono le medesime, ma i loro errori standard e, pertanto, le statistiche t con i relativi p-values sono cambiati. Ora, anche la costante è significativa. Questo può sorprendere, dato che il test di White non ci ha portato a rigettare l'ipotesi di omoschedasticità. Tuttavia,

```
Variabile dipendente: Cons
Errori standard robusti rispetto all'eteroschedasticità, variante HC1

      VARIABILE      COEFFICIENTE       ERRORE STD    STAT T   P-VALUE

  const              -123.781           43.0813       -2.873   0.00859 ***
  PNL                   0.794485         0.00357785  222.057  <0.00001 ***
```

Tabella 3: Stime OLS con errori standard consistenti anche in presenza di eteroschedasticità.

come vedremo fra poco, i residui della nostra regressione violano un'altra delle ipotesi classiche, che ha persino peggiori conseguenze sugli errori standard stimati.

3.3 Correlazione seriale

Osservando il grafico degli errori stimati in Figura 2 ci si accorge che gli errori non sembrano tra loro indipendenti. Infatti, quando sono sopra alla media (che per stime OLS è sempre zero) tendono a rimanerci, e altrettanto succede quando sono sotto alla media. Quindi, se al tempo t osservo un errore maggiore di zero, mi aspetto che al tempo $t + 1$ l'errore sarà più probabilmente positivo piuttosto che negativo. Questo è un classico esempio di correlazione seriale (o autocorrelazione) positiva. Se, invece, gli errori tendessero a cambiare segno per tempi consecutivi, allora si parlerebbe di correlazione seriale negativa. Quando vi è correlazione, che è una importante forma di dipendenza, è violata l'ipotesi di indipendenza seriale e gli stimatori OLS perdono la correttezza, l'efficienza e gli errori standard calcolati dai pacchetti software non sono consistenti.

Per testare la presenza di correlazione seriale negli errori di regressione esistono diversi test. Molti pacchetti statistici forniscono nell'output standard una statistica chiamata *Durbin-Watson* (a volte solo DW). Nella Tabella 1 è riportato un valore della statistica di Durbin-Watson di circa 0.61.

La statistica di DW è utilizzata per testare la presenza di correlazione tra due errori consecutivi (tra e_t e e_{t+1}, autocorrelazione di ordine 1). La statistica DW assume valori tra 0 e 4. In caso di assenza di correlazione la DW teorica (nella popolazione) è pari a 2. Al crescere della correlazione tra e_t e e_{t+1} la DW scende fino a raggiungere 0, quando la correlazione lineare tra gli errori consecutivi è pari a 1. Al decrescere della correlazione la DW cresce fino a raggiungere il valore di 4, quando la correlazione tra gli errori consecutivi è par a -1. Purtroppo la distribuzione della statistica DW sotto l'ipotesi di assenza di correlazione è di complicata derivazione e, data una ampiezza del test, per es. del 5%, esistono due intervalli di $[0, 4]$ in cui non è possibile determinare, se non caso per caso, la risposta del test. Durbin e Watson forniscono quindi una tabella per diversi valori di n e di k, che va letta come segue. Si prenda il valore della statistica DW calcolata sui

valori campionari e se DW è minore di 2 si ponga $dw = \text{DW}$ altrimenti si ponga $dw = 4-\text{DW}$). Si individuino i valori d_L e d_U rilevanti in Tabella 4. Se $dw < d_L$ si

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$		$k'=6$		$k'=7$		$k'=8$		$k'=9$		$k'=10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.012	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

Tabella 4: Valori critici del test DW con ampiezza 5%. Si noti che qui $k' = k + 1$ è il numero di regressori più la costante.

rifiuta l'ipotesi di incorrelazione seriale tra osservazioni consecutive, se $dw > d_U$ non si rigetta l'ipotesi nulla di incorrelazione tra osservazioni consecutive, mentre se $dw \in [d_L, d_U]$ il test non porta a decisione univoche.

Nel caso in esempio, abbiamo $n = 25$ e $k' = 2$ e dalla tabella $d_L = 1.206$ e $d_U = 1.550$. Pertanto, avendo $dw = 0.61 < 1.206$ si può concludere che i nostri errori mostrano autocorrelazione di ordine 1 significativa.

Un modo più generale per individuare autocorrelazione di qualsiasi ordine è per mezzo dell'autocorrelogramma, cioè del grafico della correlazione campionaria tra e_t ed e_{t-h} per diversi valori di h.

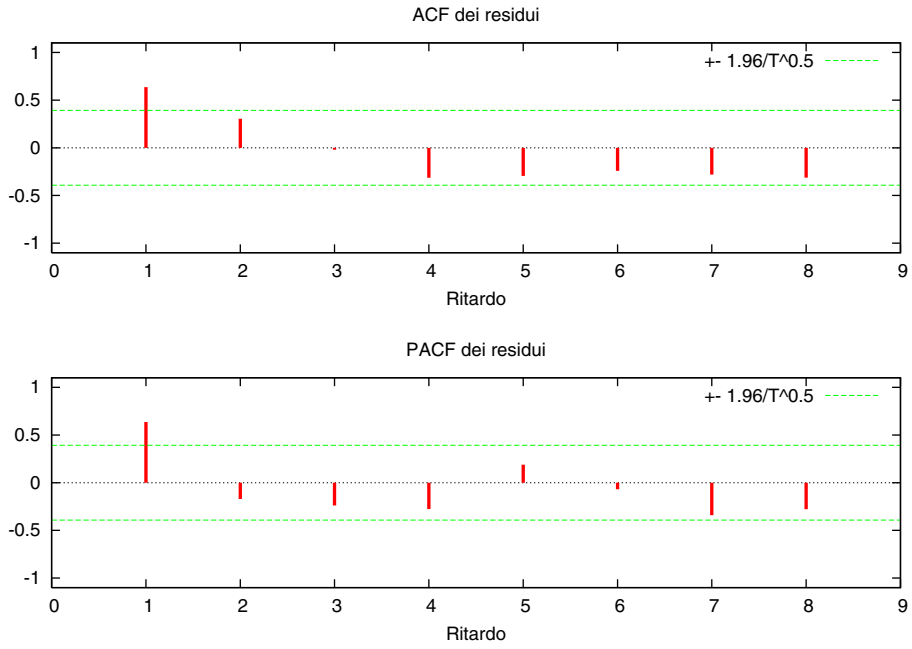

Figura 4: Funzione di autocorrelazione (ACF) campionaria per $h = 1, \ldots, 8$.

Come avrete notato dalla Figura 4, il correlogramma contempla una secondo grafico con la funzione di autocorrelazione parziale, ma per ora limitatevi ad ignorarlo. Le bande presenti del correlogramma consentono di testare l'ipotesi che una data correlazione sia nulla: se la barra supera le bande in una delle due direzioni, allora possiamo rifiutare l'ipotesi di assenza di correlazione a quel dato ritardo. La statistica di Ljung-Box che tipicamente completa il correlogramma, permette di testare l'ipotesi che tutte le correlazioni fino a quelle di un ritardo h' prefissato siano nulle.

Per i nostri dati la versione testuale del correlogramma con test di Ljung-Box è riportata in Tabella 5. Se scegliamo il ritardo $h = 8$, la statistica di Ljung-Box è pari a 26.60 ed il relativo p-value è $0.001 < 0.05$, che spinge a rigettare l'ipotesi nulla che tutte le correlazioni siano nulle: $\text{Corr}(e_t, e_{t-1}) = \text{Corr}(e_t, e_{t-2}) = \ldots = \text{Corr}(e_t, e_{t-8}) = 0$.

Quando si rileva la presenza di correlazione seriale tra gli errori di regressione sono possibili due strade: i) modellare direttamente la correlazione, ii) stimare gli errori standard degli stimatori in maniera consistente. La prima soluzione richiede

Funzione di autocorrelazione dei residui

LAG	ACF		PACF		Q-stat.	[p-value]
1	0.6365	***	0.6365	***	10.5193	[0.001]
2	0.3040		-0.1702		13.0138	[0.001]
3	-0.0197		-0.2386		13.0247	[0.005]
4	-0.3139		-0.2763		15.9172	[0.003]
5	-0.2946		0.1892		18.5804	[0.002]
6	-0.2419		-0.0689		20.4609	[0.002]
7	-0.2809		-0.3413	*	23.1246	[0.002]
8	-0.3130		-0.2779		26.6046	[0.001]

Tabella 5: Funzioni di autocorrelazione e autocorrelazione parziale campionarie e statistica di Ljung-Box (Q-stat).

competenze in *analisi delle serie storiche*, anche se un modello del tipo

$$y_t = \beta_0 + \beta_1 x_{1,t} + \ldots + \beta_k x_{k,t} + \eta_t$$
$$\eta_t = \phi_1 \eta_{t-1} + \varepsilon_t$$

spesso potrebbe essere sufficiente. In tale modello gli errori di regressioni η_t sono fatti dipendere dagli errori immediatamente precedenti η_{t-1}. Tali errori di regressione sono detti seguire un processo autoregressivo di ordine 1 o AR(1). Molti software permettono di stimare modelli di questo tipo. Dato che il parametro ϕ_1 coincide con la correlazione di η_t con η_{t-1}, esso sarà compreso tra -1 e 1.

Le stime di tale modello sui nostri dati sono mostrate in Tabella 6. Mentre il correlogramma delle stime di ε_t, e_t, sono mostrate in Figura 5.

Variabile dipendente: Cons

VARIABILE	COEFFICIENTE	ERRORE STD	STAT T	P-VALUE	
phi_1	0.738271	0.138751	5.321	<0.00001	***
PNL	0.789623	0.00408463	193.316	<0.00001	***

Tabella 6: Stime modello regressivo con errori AR(1).

Il test Ljung-Box a ritardo $h = 8$ è pari a $Q(8) = 5.55$ con un p-value di $0.697 > 0.05$, pertanto tutta l'autocorrelazione degli errori di regressione sembra essere stata modellata.

Come si è già accennato, la seconda soluzione è quella di ottenere stime consistenti degli errori standard nonostante la presenza di autocorrelazione. Tali stime

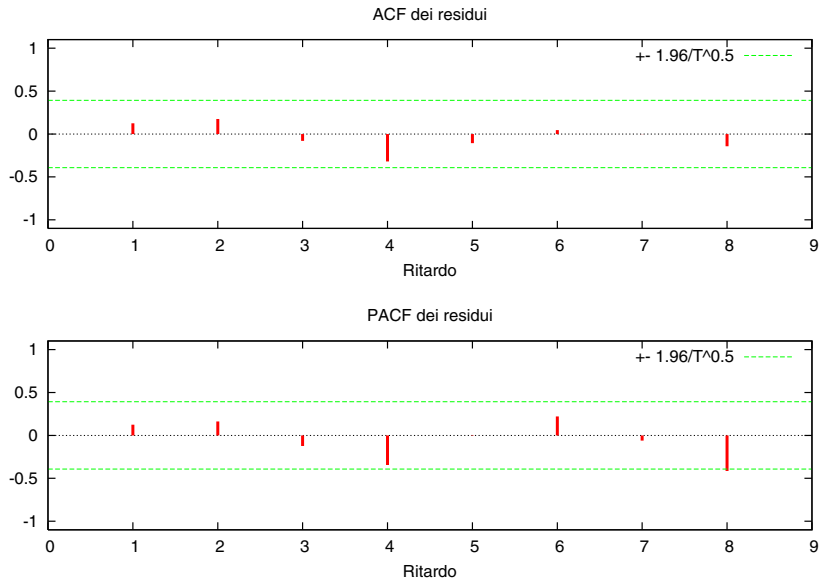

Figura 5: Autocorrelogramma degli errori del modello di regressione con errori AR(1).

della matrice di covarianza degli stimatori prendono solitamente il nome *HAC*, che sta per *Heteroskedasticity Autocorrelation Consistent*, e sono consistenti anche in caso di eteroschedasticità oltre che di errori autocorrelati.

Un estratto della stima con errori standard HAC è riportato in Tabella 7

```
Variabile dipendente: Cons
Errori standard robusti rispetto alla correlazione seriale, ordine di ritardo 2

      VARIABILE      COEFFICIENTE       ERRORE STD    STAT T   P-VALUE

  const              -123.781           53.6956       -2.305   0.03052 **
  PNL                   0.794485         0.00450698  176.279  <0.00001 ***
```

Tabella 7: Stime OLS con errori standard HAC.

Come è facile notare, il valore della stima del coefficiente β_1 relativo alla variabile PNL è pressoché il medesimo nei due modelli e metodi di stima.

4 Processi integrati e cointegrazione

4.1 Processi stazionari e processi integrati

Quando si fanno regressioni su serie storiche, vi è una cosa a cui prestare molta attenzione. In genere, una serie storica è vista come una *traiettoria* di un *processo stocastico*. La disamina, anche solo superficiale, del concetto di processo stocastico va molto oltre le intenzioni di questa dispensa. Per quanti ci riguarda, si può pensare a un processo stocastico come a una "macchinetta genera dati"in maniera non esattamente prevedibile.

Particolare importanza giocano i *processi stocastici stazionari*, cioè sequenze di variabili casuali $x_1, x_2, \ldots, x_n$ con

- media costante: $\mathbb{E}[x_t] = \mu$
- varianza costante $\mathrm{Var}[x_t] = \sigma^2$
- covarianza seriale costante $\mathrm{Cov}(x_t, x_{t-h}) = \gamma(h)$ (la covarianza dipende dalla distanza temporale h tra le due osservazioni e non dal tempo t in cui la si calcola).

Una ulteriore classe di processi molto importanti è quella dei *processi integrati*. Un processo stocastico $\{x_t\}$ si dice *integrato di ordine* 1, se x_t non è stazionario, mentre la sua differenza $y_t = x_t - x_{t-1}$ è stazionaria.

Il processo integrato più semplice è il processo *passeggiata aleatoria con deriva* (random walk with drift), che dato un valore iniziale x_0 è generato da

$$x_t = x_{t-1} + \delta + \varepsilon_t,$$

con ε_t sequenza di variabili indipendenti e identicamente distribuite (i.i.d.) a media varianza finita e δ è una costante, detta deriva (drift). Se tale costante è nulla, allora il processo è detto solo passeggiata aleatoria, se $\delta > 0$ il processo ha una tendenza a crescere e se $\delta < 0$ il processo tende a declinare. Tale processo è integrato di ordine 1, infatti la sua differenza è la variabile casuale ε_t, che essendo una sequenza i.i.d. è processo stazionario. Se ε_t ha media nulla, allora il valore atteso di x_t è il primo valore della sequenza x_0, mentre la varianza è pari a t volte la varianza di ε_t. In Figura 6 sono illustrate traiettorie di $\varepsilon_t \sim N(0, 1)$, $x_t = x_{t-1} + \varepsilon_t$ e $z_t = z_{t-1} + 0.1 + \varepsilon_t$.

4.2 Test di radice unitaria e di stazionarietà

Spesso si riesce a capire se una serie storica è stata generata da un processo integrato semplicemente guardandone il grafico, tuttavia esistono diversi test statistici, detti test di radice unitaria (unit root tests), quando l'ipotesi nulla è l'integrazione, e test di stazionarietà, quando l'ipotesi nulla è la stazionarietà.

Il test di radice unitaria più noto (anche perché il primo comparso in letteratura) è sicuramente il *test di Dikey-Fuller aumentato* (ADF). L'ipotesi nulla è che

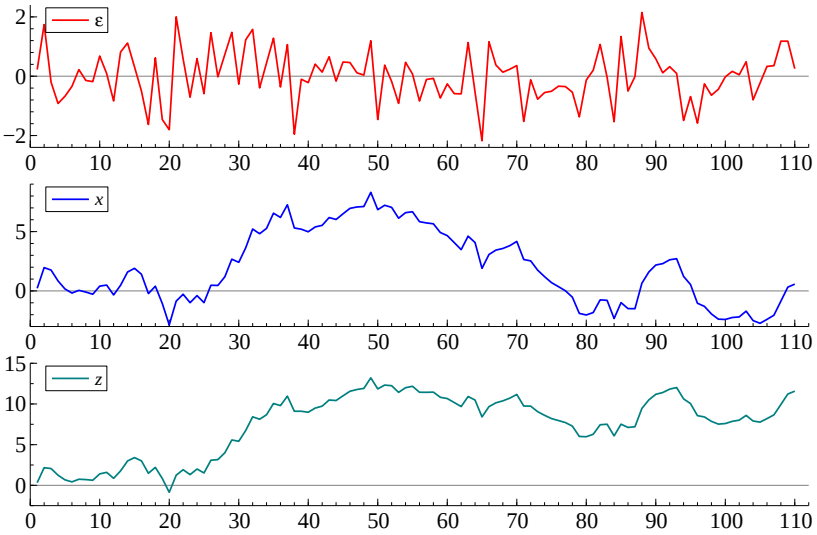

Figura 6: Traiettorie dei processi: a) ε_t i.i.d., b) x_t passeggiata aleatoria, c) z_t passeggiata aleatoria con deriva positiva.

una certa serie storica sia stata generata da un processo integrato di ordine 1. Per applicare il test ADF, che è basato sulla statistica t di una regressione ausiliaria, bisogna fissare il numero di ritardi della variabile differenziata da includere a destra dell'uguale della funzione di regressione

$$\Delta y_t = \alpha + \delta t + \gamma y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + \varepsilon_t,$$

dove Δ è l'operatore differenza, $\Delta x_t = x_t - x_{t-1}$ e l'ordine p delle variabili differenziate ritardate è da stabilire a priori. Tipicamente si stimano regressioni per diversi valori di p e si sceglie quello che minimizza un criterio di informazione. Le ultime tre righe della Tabella 1 riportatano i criteri di Akaike, Schwarz e Hannan-Quinn, gli ultimi; gli ultimi due sono i più utilizzati per modelli di serie storiche. Alcuni software (come Gretl) fanno questa operazione in automatico. L'output di Gretl per il test ADF è illustrato in Tabella 8. Dal valore del p-value ($0.41 > 0.05$) concludiamo che l'ipotesi di radice unitaria (il processo è integrato) è supportata dai dati.

Alternativamente, si può testare l'ipotesi nulla che il processo sia stazionario contro l'alternativa che esso sia integrato. Il test di stazionarietà più usato è il *KPSS* (dalle iniziali dei quattro autori). Anche in questo test esiste un parametro da specificare a priori, tuttavia i software solitamente propongono un valore, che noi accetteremo in ogni caso. Senza entrare in dettagli tecnici che oltrepassano i fini di questa dispensa, ci limitiamo a mostrare l'output tipico del test in Tabella 9. Il test ha una distribuzione non standard, e solitamente i software non sono in grado di calcolare i p-values, ma si limitano a fornire le soglie critiche per diverse

```
Test Dickey-Fuller per PNL
Ampiezza campionaria 23
Ipotesi nulla di radice unitaria: a = 1

   Con costante e trend
   Modello: (1 - L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   Ordine dei ritardi: 1
   Coefficiente di autocorrelazione del prim'ordine per e: 0.104
   Valore stimato di (a - 1): -0.453427
   Statistica test: tau_ct(1) = -2.34497
   p-value asintotico 0.4089

P-value basati su MacKinnon (JAE, 1996)
```

Tabella 8: Output test ADF per la variabile PNL.

ampiezze del test. Si ricordi che il test è "a coda destra", pertanto si rifiuta l'ipotesi nulla quando la statistica test è maggiore della soglia critica stabilita. Nel caso in

```
Test KPSS per PNL (trend incluso)

Parametro di troncamento del ritardo = 2
Statistica test = 0.14903

                  10%      5%    2.5%      1%
Valori critici: 0.119   0.146   0.176   0.216
```

Tabella 9: Output test KPSS per la variabile PNL.

esempio si ha che la statistica test (0.149) è maggiore della soglia critica al 5% (0.146), e pertanto l'ipotesi di stazionarietà è da rigettare. Il risultati del KPSS test concorda con quello dell'ADF, ma si noti che non sempre ciò accade. Quando i due test concordano possiamo ritenere i risultati piuttosto robusti.

4.3 Regressione tra serie storiche integrate

Mentre ha senso regredire serie storiche stazionarie su serie storiche stazionarie, e si è visto in precedenza come modificare il modello classico quando i residui mostrano correlazione, non ha alcun senso regredire processi stazionari su processi integrati e viceversa. Infatti non si può creare un processo stazionario (y_t) moltiplicando un processo integrato (x_t) per una costante (β_1) e sommandogli un processo stazionario (ε_t). Analogamente è possibile ottenere un processo integrato (y_t) moltiplicando un processo stazionario (x_t) per una costante (β) e sommandogli un altro processo stazionario (ε_t).

Più interessante e pericoloso il caso in cui sia la variabile dipendente, sia i regressori sono integrati. Infatti, in assenza di una particolare condizione detta

cointegrazione, che discuteremo fra qualche riga, la regressione non ha senso (è detta spuria), ma i test t sui coefficienti di regressione tenderanno a indicarci che la relazione tra regressori e variabile dipendente è statisticamente significativa. Ciò è dovuto al fatto che se la y_t e le $x_{i,t}$ non sono *cointegrate*, l'errore di regressione ε_t sarà a sua volta integrato e avrà varianza che cresce con t. Questo fa sì che i classici test t divergano, portando a rifiutare l'ipotesi nulla di assenza di relazione lineare sempre più frequentemente, con il crescere della numerosità campionaria n.

Definiamo ora il concetto di cointegrazione e vediamo come capire quando questa relazione è presente tra variabili che compongono una regressione.

Cointegrazione Due o più serie storiche integrate $x_{1,t}, \ldots, x_{k,t}$ si dicono cointegrate se esiste almeno una loro combinazione lineare non banale[2] $\beta_1 x_{1,t} + \ldots + \beta_k x_{k,t}$ che è stazionaria. Il vettore $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ che raccoglie i coefficienti della combinazione lineare è detto *vettore di cointegrazione*.

Intuitivamente, si può pensare alla relazione di cointegrazione come alla presenza di trend stocastici (tipo passeggiata aleatoria) comuni alle serie storiche. Essendo il trend la componente che nel lungo periodo prevale, ovvero è responsabile della maggior parte della variabilità di una serie storica, la cointegrazione è relazione molto forte. Due serie storiche cointegrate non possono divergere da una relazione di equilibrio esistente tra loro se non per brevi periodi.

Se le variabili $y_t, x_{1,t}, \ldots, x_{k,t}$ sono cointegrate, allora il metodo OLS mi fornisce stime consistenti (in realtà super-consistenti) di un vettore di cointegrazione. Pertanto, il modo più semplice per stabilire se le variabili integrate di una regressione sono cointegrate è

1. stimare la regressione con gli OLS,
2. verificare se gli errori di regressione stimati sono stazionari,
3. se gli errori sono stazionari, ma correlati, aggiustare la regressione nel modo visto nella sezione 3.3.

Si noti che il primo test di cointegrazione presentato in letteratura da Engle e Granger era propio basato su questa procedura. Engle e Granger proponevano di applicare il test ADF sui residui della regressione. Tuttavia, ricordatevi che se applicate la statistica ADF sui residui di una regressione, la sua distribuzione non è diversa da quella standard. Alcuni software (come Gretl) implementano la procedura di Engle-Granger in automatico, e vi forniscono i valori critici corretti. In Tabella 10 è riportato l'output del test di Engle-Granger di Gretl. Dato che l'ipotesi nulla del test ADF sugli errori di regressione non può essere rifiutato (p-value$= 0.61$), concludiamo che le due serie non sono cointegrate. Tuttavia, tenete conto del fatto che i test di cointegrazione tendono a essere poco potenti se i dati

[2]Per banale si intende la combinazione con coefficienti tutti nulli.

non sono numerosi. Osservando la Figura 2, l'allontanarsi dalla media (nulla) degli errori di regressione sembra dovuta a fattori congiunturali (come la profonda crisi economica del 1991-1995) piuttosto che permanenti (strutturali).

```
Passo 1: regressione di cointegrazione

Regressione di cointegrazione -
Stime OLS usando le 25 osservazioni 1982-2006
Variabile dipendente: Cons

      VARIABILE      COEFFICIENTE       ERRORE STD    STAT T   P-VALUE

  const              -123.781           73.0721       -1.694   0.10377
  PNL                   0.794485         0.00440894  180.199  <0.00001 ***

  R-quadro = 0.999292
  R-quadro corretto = 0.999261
  Statistica Durbin-Watson = 0.608739
  Coefficiente di autocorrelazione del prim'ordine = 0.7158
  Criterio di informazione di Akaike (AIC) = 318.485
  Criterio bayesiano di Schwarz (BIC) = 320.923
  Criterio di Hannan-Quinn (HQC) = 319.161

Passo 2: test Dickey-Fuller sui residui

ordine dei ritardi 1
Ampiezza campionaria 23
Ipotesi nulla di radice unitaria: a = 1

   Valore stimato di (a - 1): -0.350078
   Statistica test: tau_c(2) = -1.84656
   p-value asintotico 0.6071

P-value basati su MacKinnon (JAE, 1996)

Ci sono sintomi di una relazione di cointegrazione se:
(a) L'ipotesi di radice unitaria non è rifiutata per le singole variabili.
(b) L'ipotesi di radice unitaria è rifiutata per i residui (uhat) della
    regressione di cointegrazione.
```

Tabella 10: Output del test Engle-Granger per le variabili Cons e PNL.

Dato che in molti software il test do Engle-Granger non è implementato, mentre l'ADF è presente pressoché ovunque. Si riportano in Tabella 11 i valori critici del test ADF applicato sugli errori di regressione stimati. Si rammenti che il test ADF da applicare agli errori di regressione è quello *senza costante*, infatti gli errori OLS hanno per costruzione media nulla, e che il test ADF è a coda sinistra (si rifiuta se la statistica test è minore del valore critico).

Quando alcune serie storiche sono integrate, ma non cointegrate, è comunque possibile studiare la relazione tra loro intercorrente per mezzo della regressione, tuttavia è prima necessario differenziale per renderle stazionarie.

Si noti, inoltre, che quando tra alcune serie storiche sussiste cointegrazione, non è più necessario imporre l'indipendenza tra i regressori e l'errore di re-

Numero regressori	Ampiezza del test		
(costante esclusa)	0.01	0.05	0.10
Regressione con costante			
1	-3.96	-3.37	-3.07
2	-4.31	-3.77	-3.45
3	-4.73	-4.11	-3.83
4	-5.07	-4.45	-4.16
Regressione con costante e trend lineare			
1	-4.36	-3.80	-3.52
2	-4.65	-4.16	-3.84
3	-5.04	-4.49	-4.20
4	-5.58	-5.03	-4.73

Tabella 11: Tavola dei valori critici per il test ADF applicato a errori di regressione.

gressione, infatti, essendo quest'ultimo stazionario, la relazione tra l'errore di regressione e i regressori sarà necessariamente più debole rispetto alla relazione di cointegrazione presente tra le serie storiche.