Metodi Statistici per l'Analisi Socio-Economica

4. L'analisi delle serie storiche con il metodo classico

Serie storiche univariate

I c.d. *metodi per serie storiche* affrontano la modellazione, generalmente a fini previsivi, dell'evoluzione di una variabile di interesse nel tempo.

Mentre nel caso della regressione lineare la variazione nella variabile obiettivo, Y, veniva "spiegata" sulla base della variazione di una o più variabili esplicative $X_1, \ldots, X_k$, qui Y viene "spiegata dal suo andamento passato".

Ci occuperemo di serie storiche relative a una sola variabile (*univariate*). L'analisi delle serie storiche *multivariate* è un parente stretto della regressione lineare, in cui la serie Y_t dipende da un'altra serie X_t *e dal proprio passato* $Y_{t-1}, \ldots, Y_{t-h}$.

Scopo dell'analisi di serie storiche

Compredere l'andamento di una serie storica può essere importante ai fini interpretativi, ma spesso è essenziale ai fini della *previsione*. In azienda sono regolarmente oggetto di previsione (*budgeting*):

- la domanda di prodotti finiti
- il fabbisogno di risorse umane
- il fabbisogno di materie prime
- le scorte
- ...

La pianificazione e il controllo delle attività produttive che consentono di bilanciare i cicli secondo cui si svolge la vita dell'azienda necessitano continuamente di previsioni dei valori futuri di queste grandezze.

Cos'è una serie storica

Una successione di dati osservati su una variabile Y nel tempo:

$$y_t, \quad t = 1, \dots T$$

I dati possono essere misurati

- in un istante (*serie di stato*)
- su un intervallo (*serie di flusso*)

In una serie storica, è comune la presenza di *dipendenza*, che prende il nome di *correlazione seriale*.
Possiamo considerare serie

- deterministiche: possono essere previste esattamente sulla base della loro storia passata
- stocastiche: sono determinate solo in maniera parziale e non possono essere previste senza errore

Obiettivi dell'analisi di serie storiche

- Descrizione (analisi grafica, individuazione degli outliers)
- Spiegazione (comprensione del fenomeno)
- Previsione (inferenza sui valori futuri del fenomeno di interesse)
- Filtraggio (estrazione/stima di componenti non osservabili della serie stessa)
- Controllo (riguarda i processi produttivi e le loro caratteristiche di qualità)

L'approccio classico

L'approccio classico all'analisi delle serie storiche postula che il processo generatore dei dati sia costituito da una parte deterministica $f(t)$ e una stocastica u_t:

$$Y_t = f(t) + u_t$$

- La componente deterministica (*legge di evoluzione temporale* del fenomeno) costituisce la parte *sistematica* e in quanto tale è perfettamente prevedibile
- la componente casuale (*errori accidentali*) rappresenta tutte le circostanze non considerate esplicitamente in $f(t)$

La componente casuale viene immaginata *stazionaria* e *incorrelata* (*white noise*). Nell'approccio moderno ci si concentra invece su quest'ultima, ammettendo che essa possa essere autocorrelata (ma in genere si richiede la stazionarietà).

Stazionarietà

Una serie storica riesce "stazionaria" se nel processo stocastico Y_t che la genera ricorrono le seguenti tre condizioni:

- il valore atteso di Y_t è costante (*stazionarietà in media*)

$$E(Y_t) = \mu \ \forall t$$

- la varianza di Y_t è costante

$$Var(Y_t) = \sigma^2 \ \forall t$$

- la covarianza tra due elementi Y_t e Y_s dipende soltanto dalla *distanza*

$$Cov(Y_t, Y_s) = Cov(Y_{t+m}, Y_{s+m}) = f(t-s) \ \forall t, s, m$$

Un simile processo si dice *stazionario in covarianza*. Una serie osservata y_t sarà "stazionaria" se generata da un processo stazionario.

Le componenti di una serie storica

Le serie storiche presentano (*possono* presentare!) tipicamente le seguenti componenti:

- Trend: movimento tendenziale di fondo dovuto all'evoluzione di lungo periodo del fenomeno
- Ciclo: oscillazione congiunturale di carattere ricorrente, spesso dovuto all'oscillare di un sistema economico attorno alle condizioni di equilibrio
- Stagionalità: regolarità empirica legata ai periodi dell'anno e dovuta a fattori climatici (alternanza delle stagioni) oppure organizzativi (ferie, festività)
- Accidentalità: componente residuale rispetto alle cause strutturali 1)-3), in genere relativa a molte influenze di piccola entità o comunque non chiaramente identificabili né suscettibili di modellazione esplicita (v. *errori* del modello OLS)

Le prime tre, se presenti, costituiscono la c.d. *parte sistematica*.

I possibili approcci e le fasi dell'analisi

Si distinguono due approcci all'analisi delle serie storiche a fini previsivi:

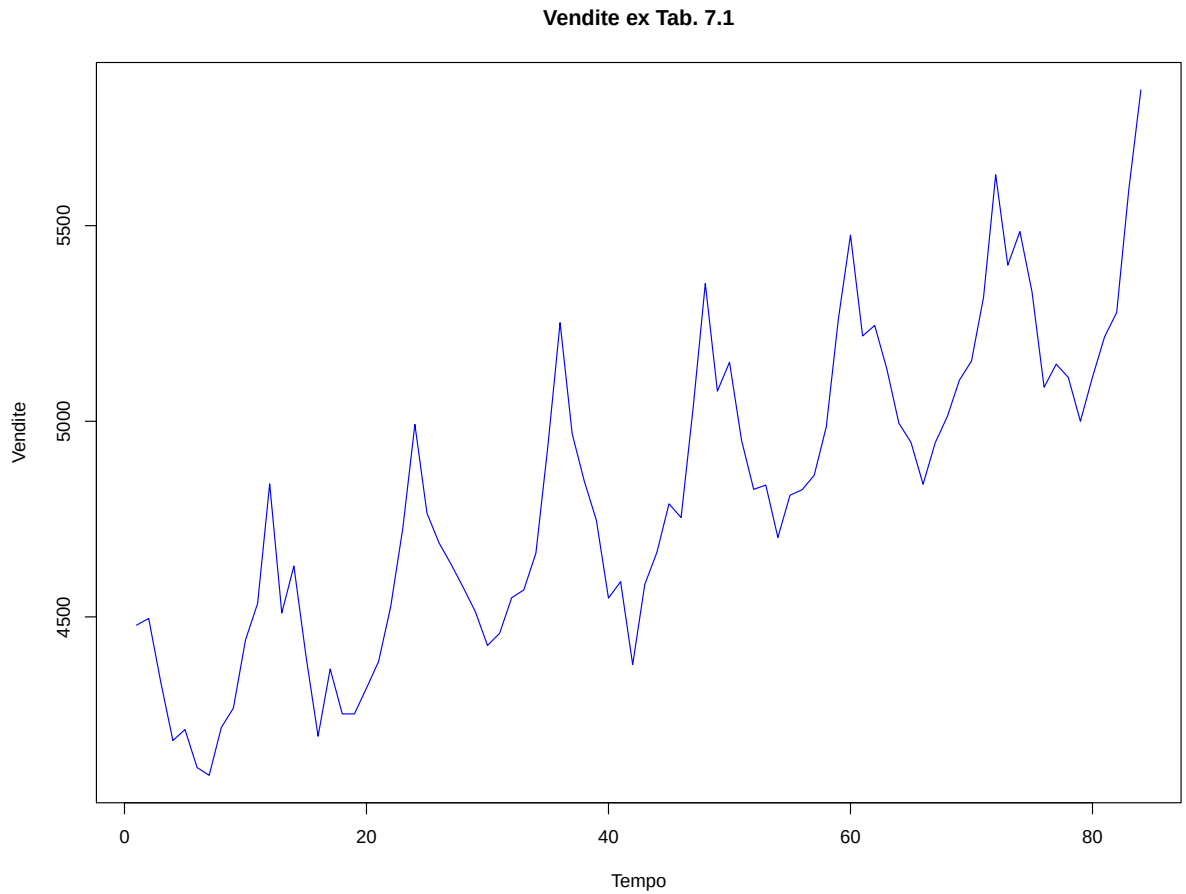
- Classico: scomposizione della serie nelle componenti sopra descritte (sola parte sistematica) e proiezione di ciascuna separatamente
- Moderno: considera il processo Y_t come un tutt'uno di carattere stocastico da modellare con tecniche probabilistiche

Le fasi di un'analisi volta alla previsione saranno:

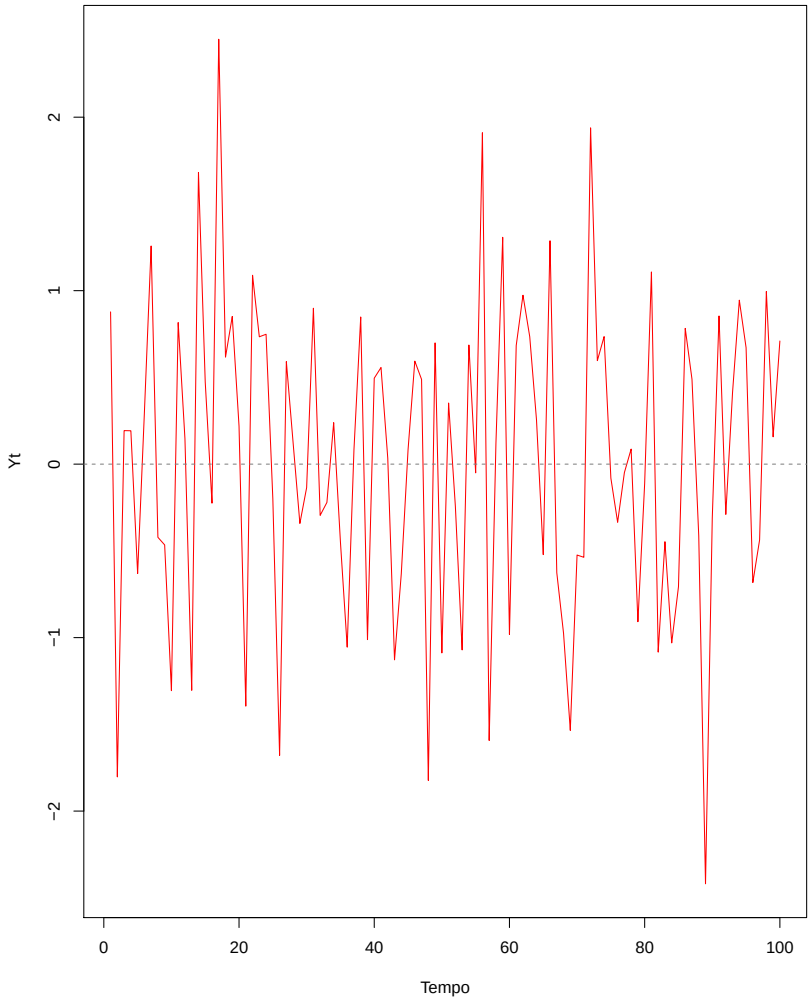
- analisi del problema
- raccolta dei dati
- analisi preliminare della struttura della serie storica
- scelta e stima del modello
- valutazione della bontà del modello a fini previsivi
- (utilizzo in pratica!)

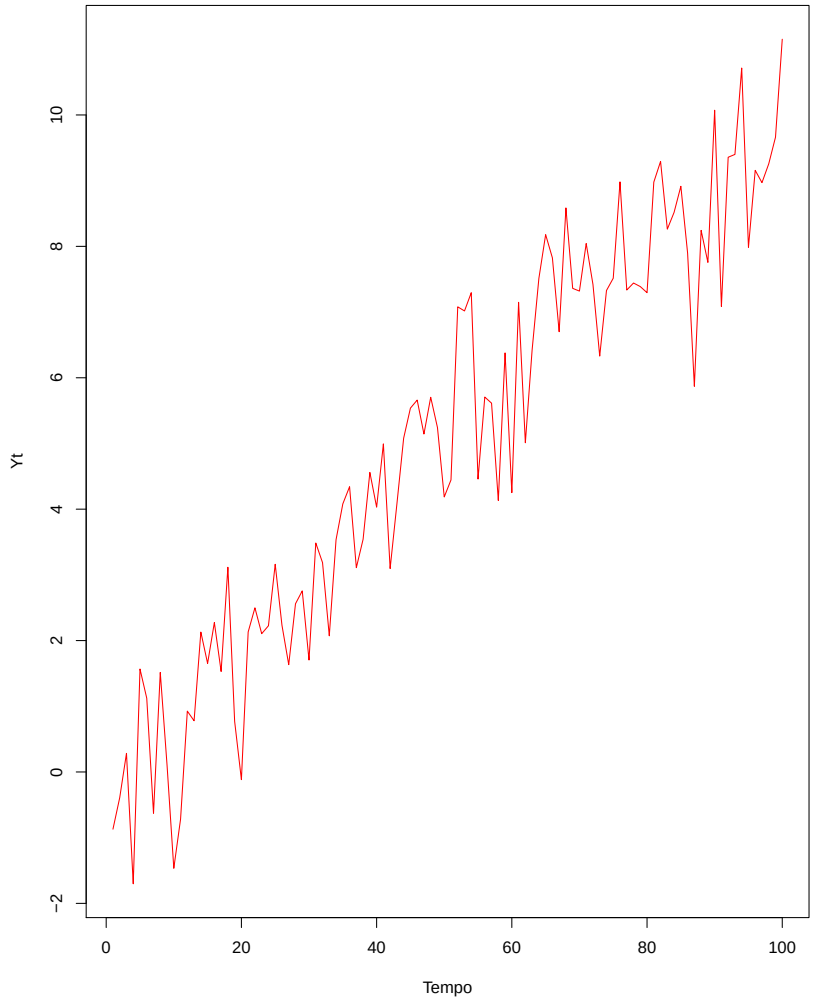
Rappresentazione grafica delle serie storiche

Iniziamo dalla rappresentazione grafica fornendo alcune intuizioni; nella prossima sezione preciseremo meglio i concetti. *(Vedi Biggeri Cap7_Tab7.1.R)*



Serie stazionarie (in media) vs. evolutive



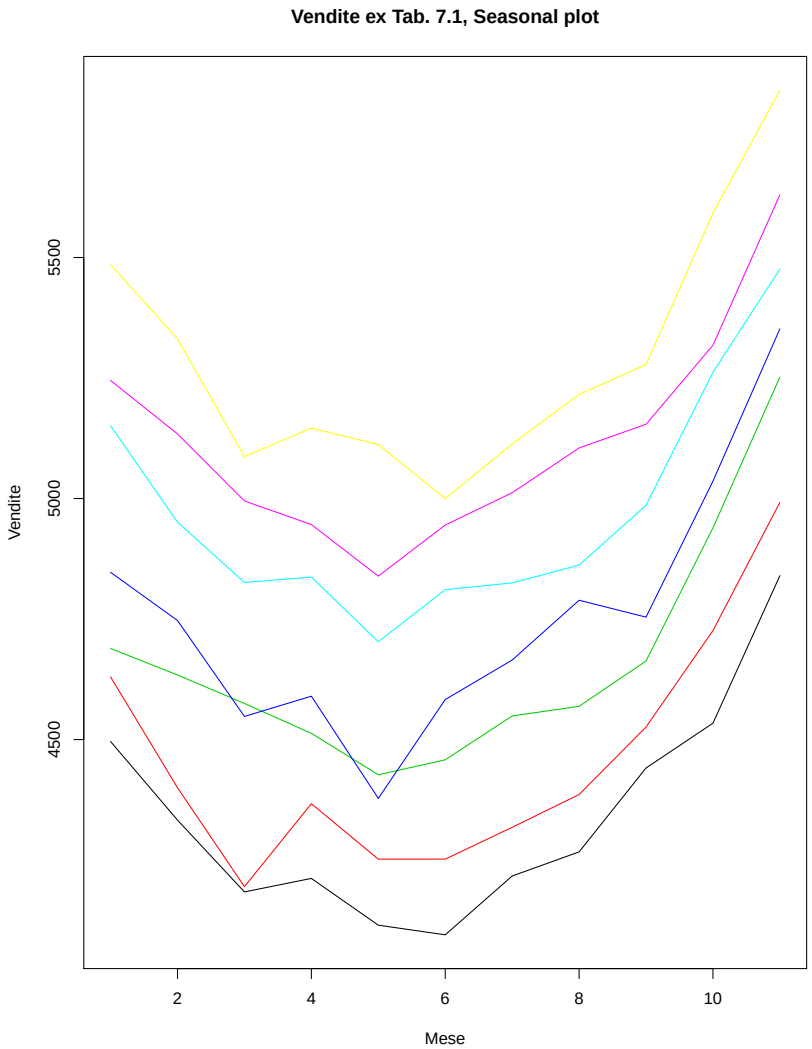


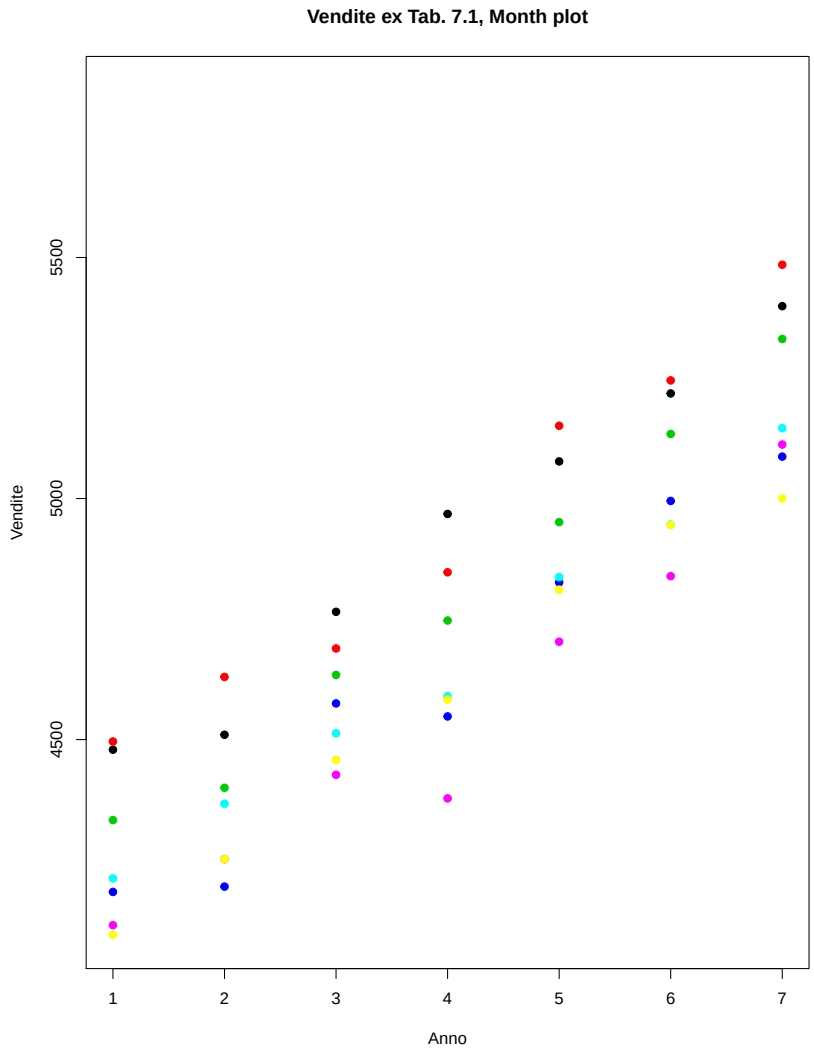
Stagionalità

Se una serie ha frequenza infra-annuale, possono presentarsi regolarità legate alle stagioni.

- Per evidenziare a livello descrittivo la *stagionalità* può essere utile visualizzare l'andamento della serie attraverso i periodi dell'anno, p. es. i mesi, anno per anno, con grafici sovrapposti: questo è il c.d. *seasonal plot*
- oppure si possono visualizzare i mesi di ogni anno, mese per mese (*monthplot*)

Seasonal plot e Month plot





Correlazione

L'*indice di autocorrelazione* è definito come la covarianza standardizzata (=il coeff. di correlazione) tra la stessa variabile in due istanti diversi:

$$\rho(h) = Cov(Y_t, Y_{t+h})/Var(Y_t)$$

- Al variare di h tra 0 e $T - h$ si ottiene la *funzione di autocorrelazione*
- Il *correlogramma* è il diagramma degli indici di autocorrelazione in funzione di h

Coefficiente di autocorrelazione

Per valutare l'autocorrelazione di Y_t è utile il concetto di *ritardo* (*lag*): in ogni istante t il ritardo h-esimo di Y_t è Y_{t-h}.

L'operatore ritardo, per esempio di ordine $h = 2$, applicato al processo

$$Y = Y_1, Y_2, Y_3, Y_4, \ldots, Y_{T-2}, Y_{T-1}, Y_T$$

dà luogo a un'altro processo stocastico

$$Y_{-2} = NA, NA, Y_1, Y_2, \ldots, Y_{T-2}$$

Lo stesso vale per la serie osservata:

$$y = y_1, y_2, \ldots, y_T$$

$$y_{-2} = NA, NA, y_1, y_2, \ldots, y_{T-2}$$

Stima della funzione di autocorrelazione

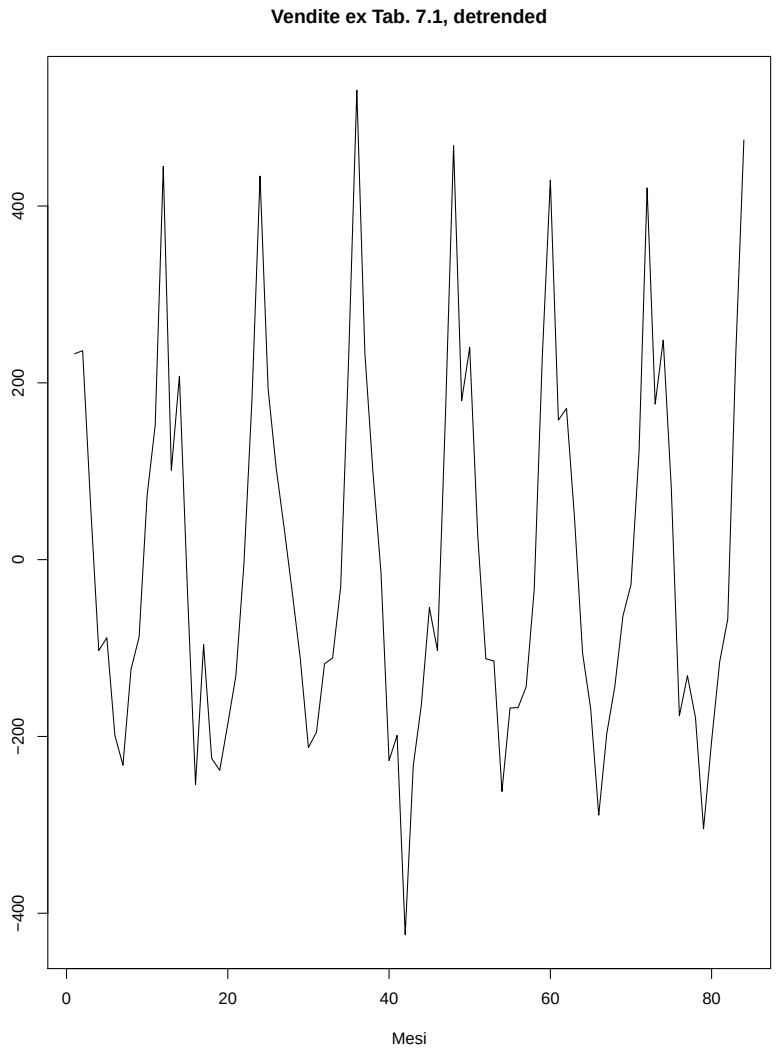
Il coefficiente di autocorrelazione di Y_t a ogni "ritardo" h viene stimato come la correlazione campionaria di Y_t e Y_{t-h}:

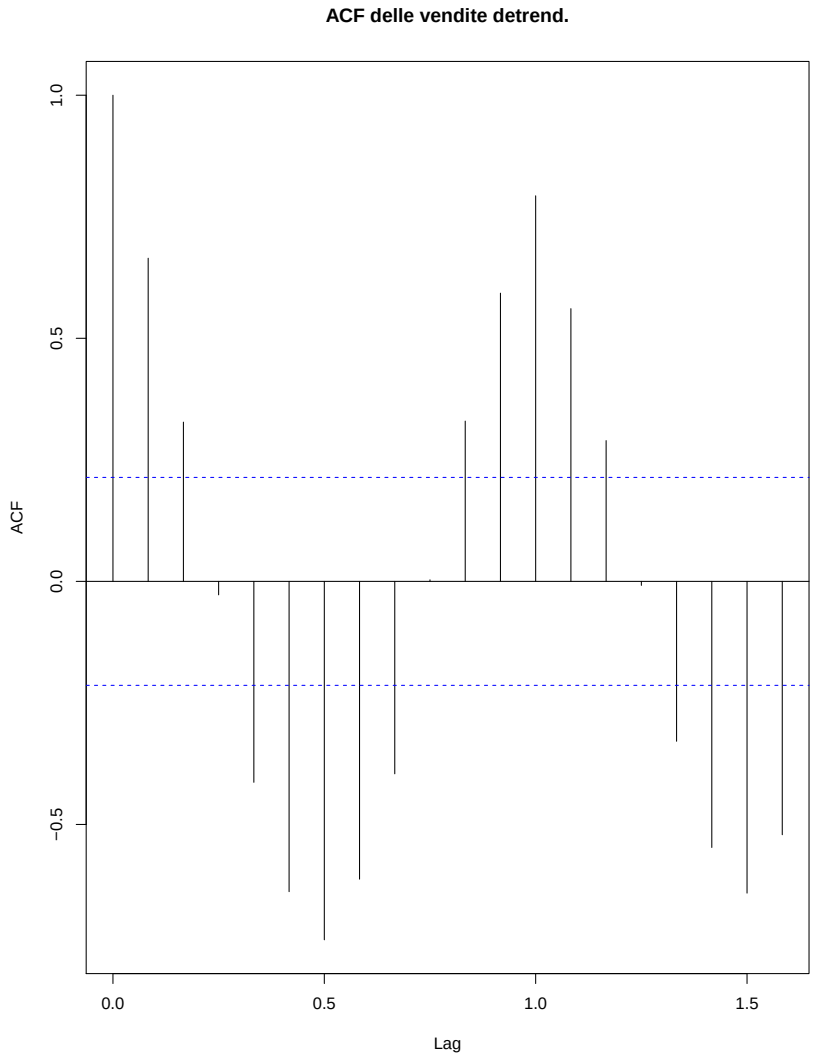
$$\rho_h = \frac{\sum_{t=h+1}^{T}(y_t - \bar{y})(y_{t-h} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

La stima dei vari $\rho_h, h = 1, \ldots, T - h$ dà luogo alla funzione di autocorrelazione (ACF) empirica.

NB II le autocorrelazioni a ogni distanza h si possono stimare solo se il DGP è stazionario in covarianza. Altrimenti per un dato h le covarianze $Cov(Y_1, Y_{1+h})$, $Cov(Y_5, Y_{5+h})$, $Cov(Y_t, Y_{t+h})$ sarebbero tutte diverse e per stimare ciascuna avrei a disposizione una sola coppia di osservazioni.

ACF plot


Vendite ex Tab. 7.1, detrended
400
200
0
−200
−400
0
20
40
60
80
Mesi


ACF delle vendite detrend.
ACF
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
1.5
Lag

Modelli di (s)composizione delle serie storiche

L'approccio classico ipotizza che la serie storica sia generata come

$$Y_t = f(T_t, C_t, S_t, e_t)$$

dove la parte deterministica può consistere di trend (T), ciclo (C) e stagionalità (S) ed e è un disturbo aleatorio.

Stimare la componente ciclica in modo separato è fuori moda. Ci si accontenta spesso di considerarla assieme al trend, al che questa componente (T) viene detta *trend-ciclo*.

Modelli di (s)composizione delle serie storiche - 2

La componente deterministica f può assumere diverse forme funzionali:

- additiva: $Y_t = T_t + C_t + S_t + e_t$
- moltiplicativa: $Y_t = T_t \cdot C_t \cdot S_t \cdot e_t$
- mista $Y_t = T_t \cdot C_t \cdot S_t + e_t$
- Nel primo caso tutte le componenti sono espresse nell'unità di misura di Y
- nel secondo C S ed e sono numeri puri (*incidenze relative* ovvero numeri indici)
- nel terzo C ed S sono numeri puri

Un modello moltiplicativo può essere *linearizzato* con una trasformazione logaritmica:

$$ln(Y_t) = ln(T_t) + ln(C_t) + ln(S_t) + v_t$$

Operazioni di pulizia della serie

Vaglio sulla continuità:

- cambiamenti di base (per i numeri indici)
- metrica della variabile (prezzi correnti o costanti, totale o pro capite)
- variazioni strutturali (es. lira/euro)

Vaglio sulla lunghezza della serie:

- numero di osservazioni
- frequenza (giornaliera, mensile, trimestrale...)
- fase ciclica (dove si "trova" il fenomeno nel periodo di osservazione)

Depurazione delle variazioni di calendario:

- i flussi possono risentire della lunghezza diseguale dei periodi. Per correggere si può:
  - aggregare i dati nel tempo (ridurre la frequenza)
  - passare a medie es. giornaliere
  - aggiustare mediante coefficienti correttivi (v. Di Fonzo e Lisi, Tab. 1.1)

Trend lineare (o linearizzabile) nei parametri

Consideriamo $f(t)$ stimabile con il metodo dei minimi quadrati.
(Ricordiamo: se $Y = \alpha + \beta X + u$, allora $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$)

Trend polinomiale:

$$f(t) = \alpha_0 + \alpha_1 t + \ldots + \alpha_q t^q$$

da cui il modello di regressione

$$y_t = \alpha_0 + \alpha_1 t + \ldots + \alpha_q t^q + \epsilon_t$$

in forma matriciale,

$$Y = P\alpha + \epsilon$$

e quindi $\hat{\alpha} = (P'P)^{-1}P'Y$, dove $p = [1, \ldots, n]'$ e $P = [p^0, p^1, p^2, \ldots, p^n]$

Determinare l'ordine del trend polinomiale

Polinomi di ordini differenti "producono" trend di forma differente. Es.:

- costante: $y_t = \alpha_0 + \epsilon_t$
- lineare: $y_t = \alpha_0 + \alpha_1 t + \epsilon_t$
- parabolico: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \epsilon_t$

Come decidere quanti termini polinomiali includere?

- E' possibile "abbondare" e non si sbaglia (ma si possono perdere gradi di libertà)
- Altrimenti si può usare il criterio:
  - delle differenze successive
  - del confronto dell'R^2

Operatori ritardo e differenza prima

Definiamo i seguenti operatori:

- Operatore *ritardo*: $B : By_t = y_{t-1}$
- Operatore *identita'*: $I = B^0$, $Iy_t = y_t$
- Operatore *differenza prima all'indietro*: $(I - B) : (I - B)y_t = y_t - y_{t-1}$

NB si noti che applicare più volte, p. es. due, l'operatore differenza prima è ben diverso dall'applicare la differenza seconda $(y_t - y_{t-2}$. Infatti, mentre

$$B^2 y_t = y_{t-2}$$

risulta

$$(I - B)^2 y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$$

$$= (I - B)(I - B)y_t = [I^2 - 2BI + B^2]y_t = y_t - 2y_{t-1} + y_{t-2}$$

Ordine del polinomio: criterio delle differenze

La differenza prima riduce il grado di un polinomio di 1. In generale, se $f(t)$ e' di grado q,

$$(I - B)^q f(q) = k$$

$$(I - B)^r f(q) = 0, r > q$$

Quindi si possono calcolare le successive differenze e fermarsi quando la serie diventa approssimativamente costante.
Bisogna tener presente pero' che la componente stocastica e_t viene differenziata a sua volta, e questo ne incrementa la varianza.
Esempio: se $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$ ma $Var(y_t) = Var(e_t)$ allora

$$Var[(I - B)^2 y_t] = Var[e_t - 2e_{t-1} + e_{t-2}] = \sigma^2 + 4\sigma^2 + \sigma^2 = 6\sigma^2$$

Ordine del polinomio: criterio dell'R2

Il *coefficiente di determinazione* e' un indice di adattamento statistico del modello ai dati dato dal rapporto fra la varianza dei valori stimati $Var(\hat{y})$ e quella della variabile dipendente $Var(y)$

- Il coefficiente di determinazione R^2 cresce all'aggiunta di ogni regressore, nel caso della polinomiale e' sempre

$$R_r^2 < R_{r+1}^2$$

  pertanto non e' utilizzabile come criterio di specificazione
- La versione *corretta* $\bar{R}^2$ invece, penalizzando per il numero di regressori, permette di scegliere la "miglior" combinazione di capacita' esplicativa e parsimonia

Per ogni valore di r si ottiene un $\bar{R}^2$: il maggiore indica la migliore approssimazione.

Trend esponenziale

La funzione esponenziale

$$f(t) = \alpha_0 e^{\alpha 1 t}, \alpha_0 > 0$$

e' adatta a rappresentare fenomeni di tipo "esplosivo" (incrementi in progressione geometrica)
Se $\alpha_1 > 0$ allora la curva e' crescente.

- La derivata rispetto al tempo della funzione esponenziale rappresenta il tasso di crescita della curva al tempo t. Esso e' proporzionale al livello della funzione.
- Il *tasso di crescita relativo* $\frac{df(t)}{dt}$ e' costante e pari ad α_1

Trend esponenziale - trasformata logaritmica

La trasformata logaritmica del trend esponenziale e' una retta:

$$logf(t) = log\alpha_0 + \alpha_1 t$$

pertanto

$$(I - B)logf(t)$$

$$(I - B)^r logf(t) = 0, r > 1$$

I parametri della funzione esponenziale possono quindi essere stimati mediante gli OLS sfruttando la trasformazione logaritmica, a patto che la funzione originale sia *moltiplicativa*, anche nel termine di errore:

$$y_t = f(t)e_t = \alpha_0 e^{\alpha_1 t}$$

da cui

$$logy_t = log\alpha_0 + \alpha_1 t + loge_t = \alpha_0^* + \alpha_1 t + e_t^*$$

Trend esponenziale - trasformata logaritmica - 2

Attenzione però alle proprietà dell'errore: per una stima OLS ottimale, l'errore *trasformato* e_t^* deve avere media nulla.
Se e_t^* è normale, allora e è log-normale con media $e^{\sigma^2/2}$, pertanto la previsione della relazione originale basata sulla stima OLS della relazione trasformata risulta *distorta*.

- Si potrebbe adottare il modello

$$y_t = f(t) + e_t = \alpha_0 e^{\alpha_1 t} + e_t$$

  (errore additivo) ma questa forma non è più linearizzabile
- Va stimata mediante procedimenti iterativi (es. *minimi quadrati non lineari*)

Stima della componente stagionale

Il modello di regressione puo' essere usato anche per stimare la componente stagionale, espressa come una funzione *periodica* $g(t)$

- il cui valore in t si riproduce a intervalli costanti
- la lunghezza s dell'intervallo e' detta *periodo*

$$g(t) = g(t+s) = g(t+2s) + \ldots$$

Assumiamo che il processo generatore della serie storica sia

$$Y_t = S_t + e_t$$

dove $S_t = g(t)$

Stima della componente stagionale con le dummies

Supponiamo che la funzione periodica $g(t)$ sia rappresentabile come

$$g(t) = \sum_{j=1}^{S} \gamma_j d_{jt}$$

con $t = 1, 2, \ldots, n$, dove

- $d_{jt} = 1$ nel periodo $j-$esimo dell'anno cui appartiene t
- $d_{jt} = 0$ altrimenti

Il modello di regressione associato a questa rappresentazione e':

$$y = D\gamma + e$$

con, nel caso di serie trimestrale ($S = 4$), $D = [d_1, d_2, d_3, d_4]$ e $d_1 = [1, 0, 0, 0, 1, 0, 0, \ldots]'$, $d_2 = [0, 1, 0, 0, 0, 1, 0, , \ldots]'$ ecc.

Stima della componente stagionale con dummies - 2

A questo punto la stagionalita' si puo' stimare a OLS:

$$\hat{\gamma}^* = (D'D)^{-1}D'y$$

ottenendo i *coefficienti stagionali grezzi* $\hat{\gamma}^*$ e la serie destagionalizzata "grezza" sotto forma di residui della regressione:

$$y^{*d} = y - D\hat{\gamma}^*$$

Spesso ha senso richiedere (es. *flussi*) che le somme annuali di y e y^d coincidano: per questo si richiede che

$$\sum_{j=1}^{S} \hat{\gamma}_j^* = 0$$

che puo' essere ottenuto trasformando in scarti dalla media $\hat{\gamma}_j = \hat{\gamma}_j^* - \bar{\gamma}_j$.

Stima simultanea di trend e stagionalita'

Nella maggior parte dei casi sono presenti entrambe le componenti

- tendenziale
- stagionale

In un approccio di regressione risulta immediato combinare la modellazione di trend e stagionalita':

$$y = P\alpha + D\gamma + e$$

da cui la forma

$$y = [P; D]\theta + e$$

e la stima OLS $\hat{\theta} = [\hat{\alpha}; \hat{\gamma}]$
La serie destagionalizzata si puo' sempre calcolare come

$$y^d = y - D\hat{\gamma}$$

Trend non lineari nei parametri: le curve di crescita

Molte serie storiche esibiscono dei trand non bene approssimabili con curve polinomiali in t. In particolare quando il fenomeno manifesta fasi di crescita accelerata seguite da fasi di stasi.

Andamenti come questi possono essere descritti da particolari funzioni dette *curve di crescita*.

- la curva esponenziale modificata
- la curva logistica
- la curva di Gompertz

Tutte queste curve sono non-lineari nei parametri.

La curva esponenziale modificata

Consideriamo una situazione in cui il tasso di crescita di un fenomeno al tempo t sia proporzionale alla crescita (totale) ancora da raggiungere $\alpha > 0$:

$$\frac{df(t)}{dt} = k[\alpha - f(t)]$$

con $k > 0$.
Una soluzione della precedente equazione differenziale è

$$f(t) = \alpha(1 - \beta e^{-kt})$$

con $\beta > 0$, detta *esponenziale modificata*.
La curva parte da $-\infty$ ed ha un asintoto superiore in α, che è anche un fattore di scala della funzione; k controlla la posizione sull'asse t e β determina l'intersezione con l'asse y.

La curva esponenziale modificata - 2

L'esponenziale modificata si può scrivere come

$$f(t) = \beta_0 + \beta_1 e^{\beta_2^* t})$$

con $\beta_0 = \alpha > 0$, $\beta_1 = -\alpha\beta < 0$ e $\beta_2^* = -k < 0$.

- La differenza rispetto all'esponenziale semplice è l'asintoto β_0.
- Perciò questa curva non è linearizzabile.
- Tuttavia, se conoscessimo β_0, la differenza $\beta_0 - f(t)$ sarebbe linearizzabile:

  $$log[\beta_0 - f(t)] = log(-\beta_1) + \beta_2^* t$$

  e varrebbero le considerazioni fatte a proposito dell'esponenziale semplice.

In generale, tuttavia, l'asintoto non è noto. In questo caso bisogna ricorrere a metodi numerici.

La curva logistica

Supponiamo che il tasso di crescita di un fenomeno sia proporzionale (con fattore $k > 0$) al prodotto tra il livello raggiunto e l'ammontare di crescita ancora da raggiungere (la crescita totale è $\alpha > 0$):

$$\frac{df(t)}{dt} = \frac{kf(t)[\alpha - f(t)]}{\alpha}$$

Ora è il tasso di crescita relativo a essere proporzionale al totale da raggiungere:

$$\frac{df(t)}{dt}/f(t) = \frac{k}{\alpha}[\alpha - f(t)]$$

da cui, integrando,

$$f(t) = \frac{\alpha}{1 + \beta e^{-kt}}$$

con $\beta > 0$.

La curva logistica - 2

La curva

$$f(t) = \frac{\alpha}{1 + \beta e^{-kt}}$$

è detta *curva logistica*. Essa

- ha due asintoti orizzontali: 0 per $t \to -\infty$ e α per $t \to +\infty$
- il punto di flesso cade in $t = 0$ se $\beta = 1$, "prima" se $0 < \beta < 1$ e "dopo" se $\beta > 1$

In generale,

- α determina la scala della funzione (in verticale)
- k determina la scala lungo l'asse dei tempi (orizzontale) e quindi la pendenza della curva
- β determina il punto di incontro con l'asse verticale

La curva logistica - 3

Il reciproco della curva logistica

$$\frac{1}{f(t)} = \frac{1+\beta e^{-kt}}{\alpha} = \frac{1}{\alpha} + \frac{\beta}{\alpha} e^{-kt}$$

ridefinendo $\alpha^* = \frac{1}{\alpha}$ e $\beta^* = \frac{\beta}{\alpha}$, è un'esponenziale modificata

$$f*(t) = \alpha^* + \beta^* e^{-kt}$$

che può essere stimata come visto in precedenza.

La curva di Gompertz

Se il tasso di crescita è

$$\frac{df(t)}{dt} = kf(t)log[\alpha 7 f(t)]$$

con $k, \alpha > 0$, integrando si ottiene la *curva di Gompertz*

$$f(t) = \alpha e^{-\beta e^{-kt}}$$

La curva di Gompertz

- ha forma simile alla logistica (α, k, β hanno lo stesso significato)
- ma non è simmetrica attorno al punto di flesso

Inoltre, $logf(t)$ assume la forma di un'esponenziale modificata.