Statistica per l'impresa

7. L'analisi delle serie storiche

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*.
Può ben darsi, tuttavia (*al contrario di quel che dice il libro*) che le manifestazioni successive del fenomeno siano tra loro indipendenti!

- estrazioni successive del lotto, o roulette
- errori del modello di regressione lineare sub ipotesi OLS
- *rendimenti finanziari*

(Possibili) 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:

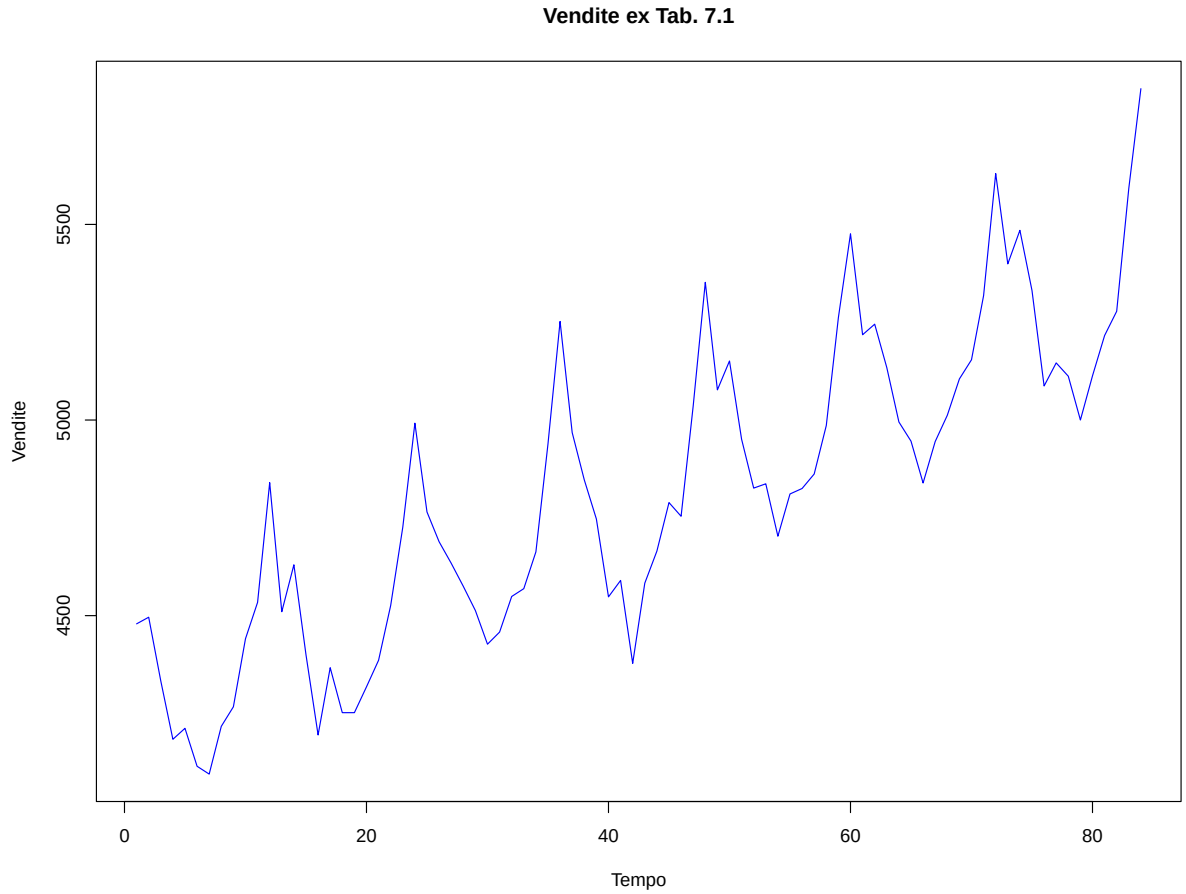
- Classico: scomposizione della serie nelle componenti sopra descritte (sola parte sistematica) e proiezione di ciascuna separatamente
- Moderno: considera il processo Y_t come un tutt'uno di carattere stocastico da modellare con tecniche probabilistiche

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 Cap7_Tab7.1.R)*



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

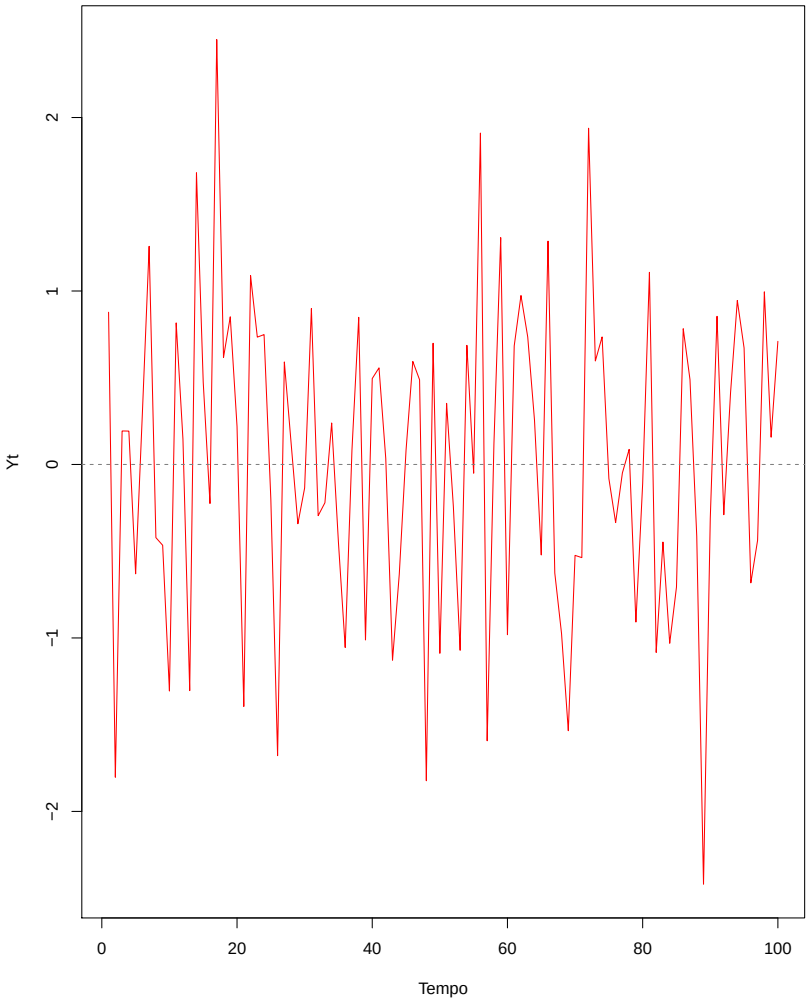
$$Var(Y_t) = \sigma^2 \ \forall t$$

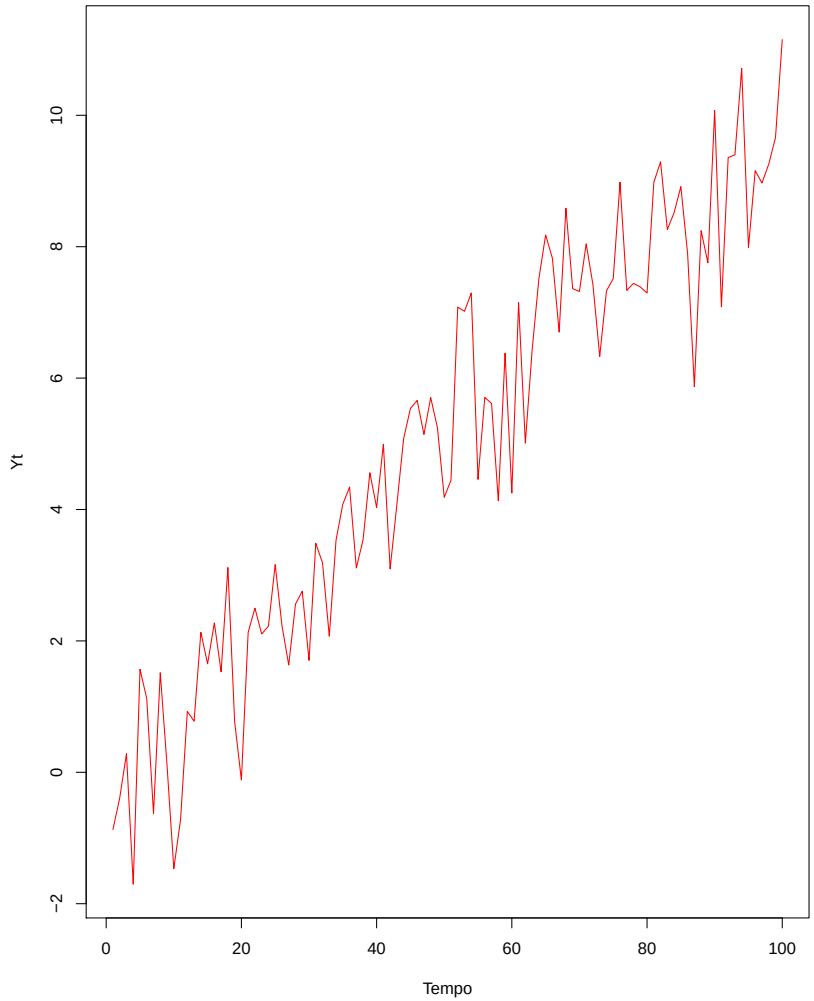
- la covarianza tra due elementi Y_t e Y_s dipende soltanto dalla *distanza*

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

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

Serie stazionarie (in media) vs. evolutive



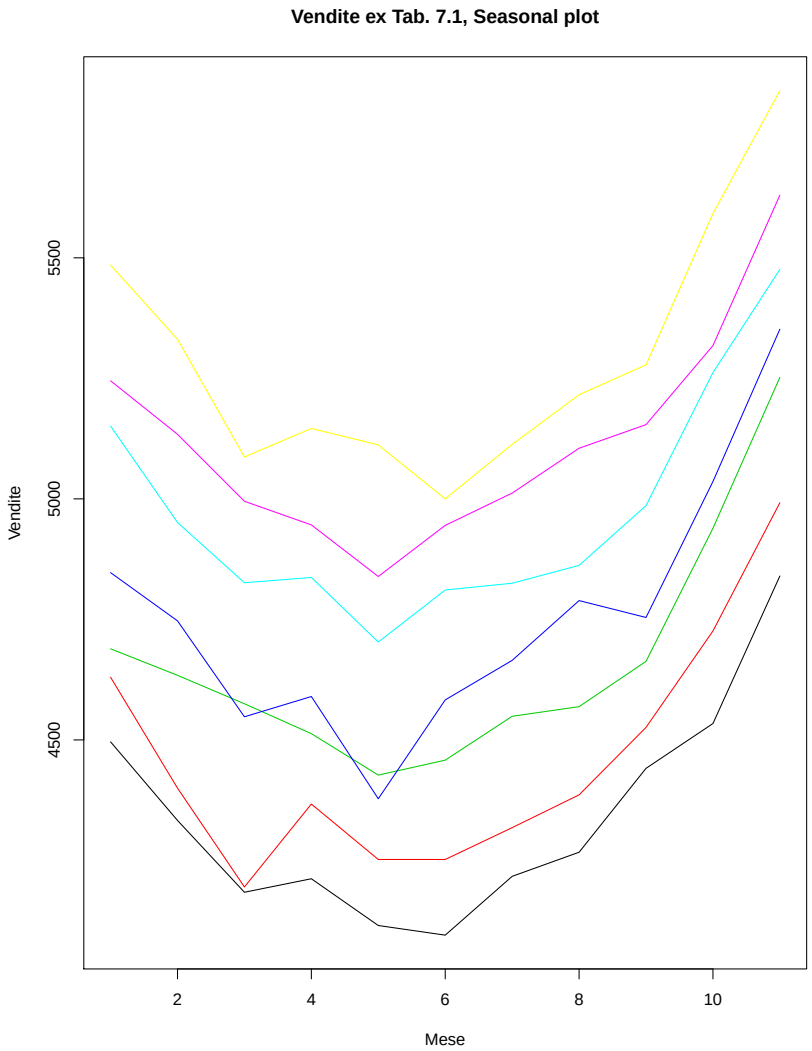


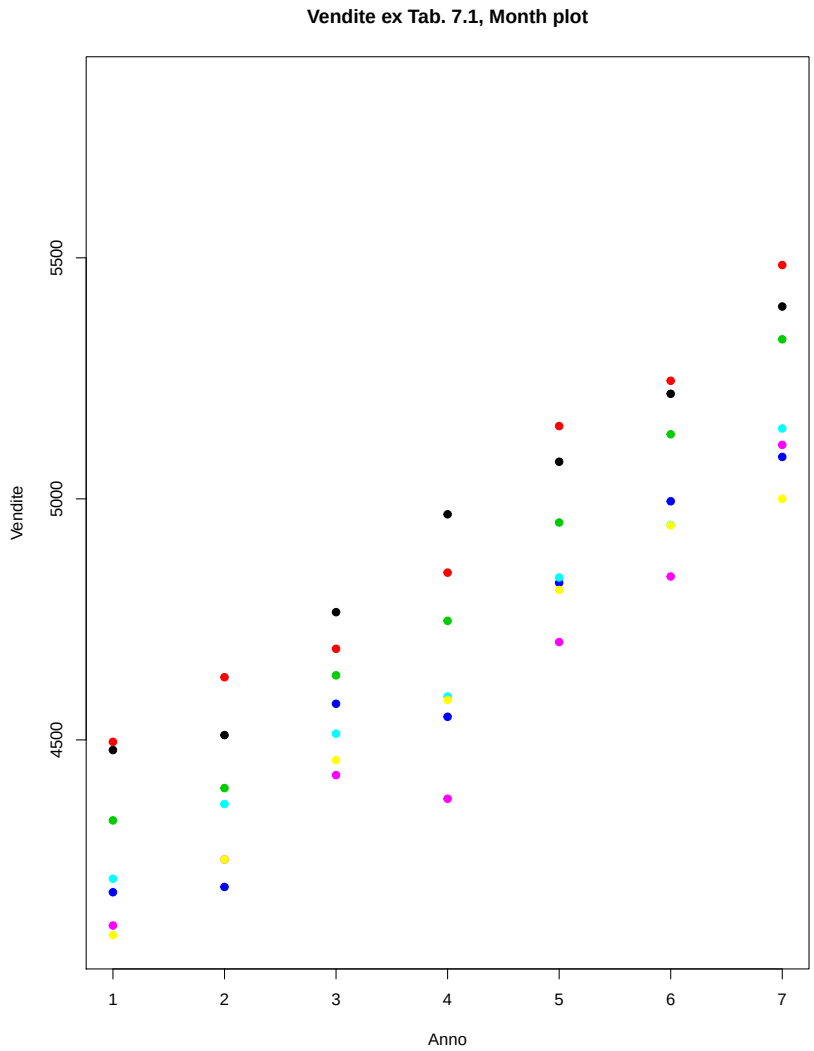
Stagionalità

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

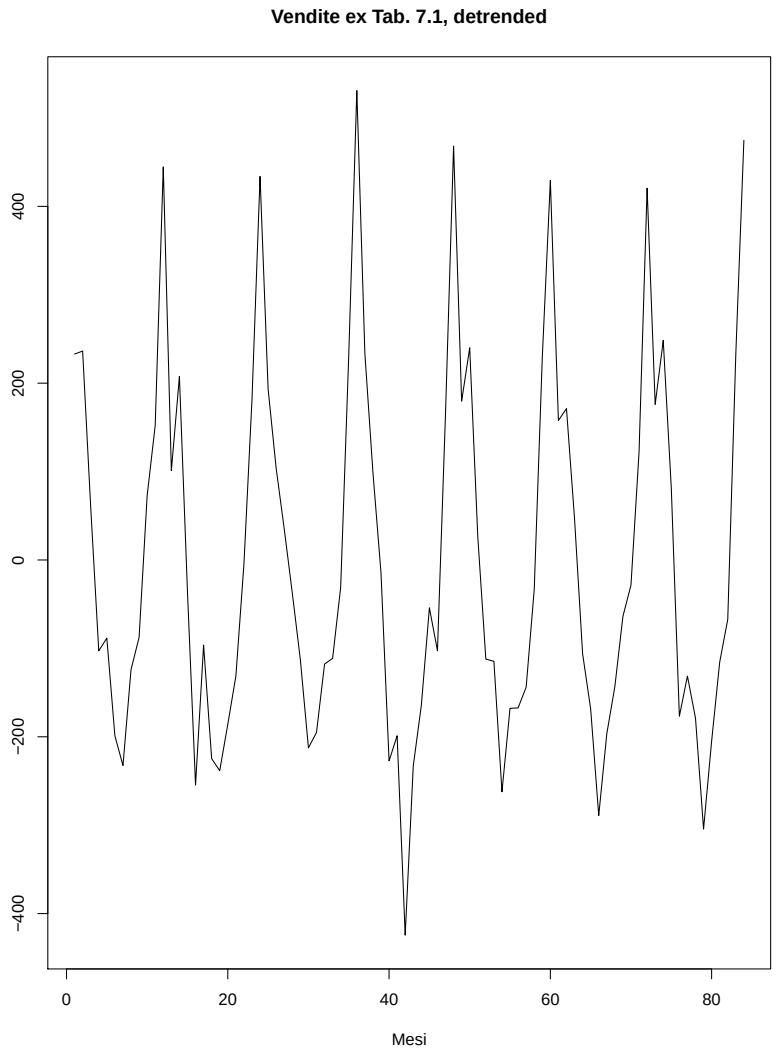
Il coefficiente di autocorrelazione di Y_t a ogni "ritardo" h viene stimato come la correlazione campionaria di Y_t e Y_{t-h}:

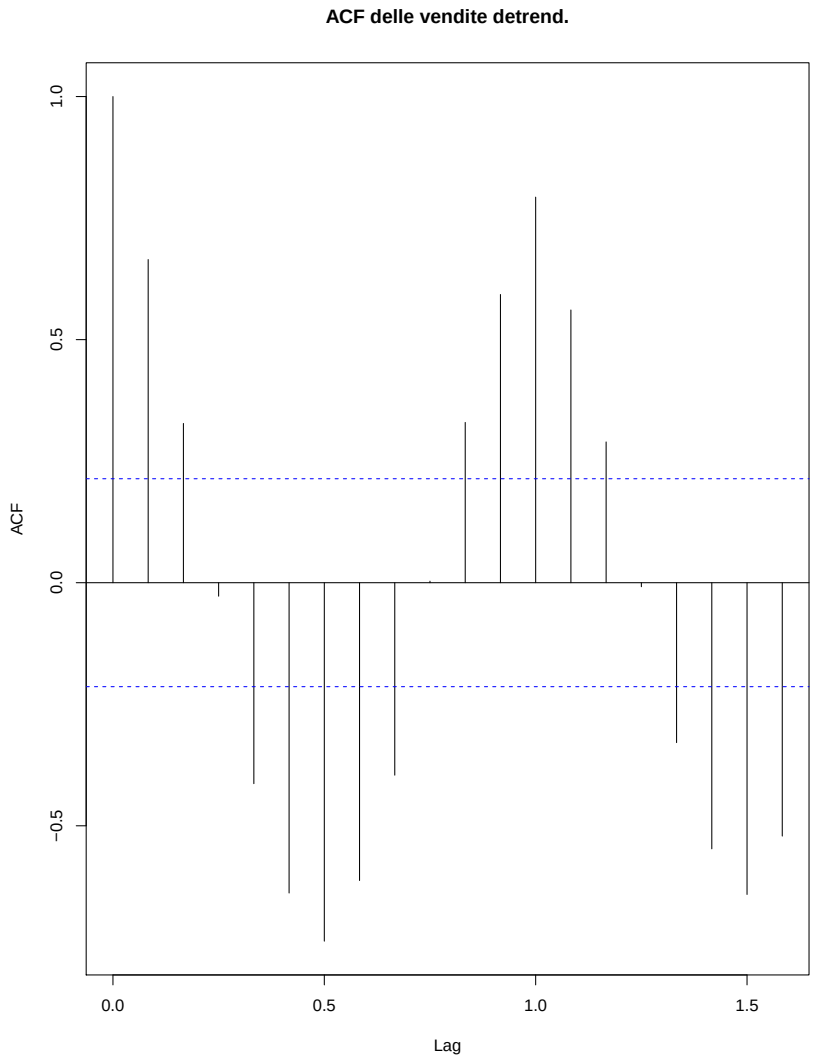
$$\rho_h = \frac{\sum_{t=h+1}^{T}(y_t - \bar{y})(y_{t-h} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

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

NB Il 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.
1.0
0.5
0.0
−0.5
ACF
0.0
0.5
1.0
1.5
Lag

La valutazione della capacità previsiva

Supponendo di aver stimato un modello univariato

$$Y_t = f(Y_1., \ldots, Y_{t-1})$$

l'analisi di bontà di adattamento del modello confronta i valori stimati (previsti) $\hat{y}_t$ con quelli effettivamente osservati y_t.
Due aspetti distinti della bontà di adattamento:

- *goodness of fit*: la capacità del modello di riprodurre i dati storici
- *goodness of forecast*: la capacità del modello di prevedere i dati futuri

Sia $y_1, y_2, \ldots, y_t, y_{t+1}, \ldots, y_T$ la serie in questione. Poniamo di stimare il modello sulla base del sottoinsieme $y_1, \ldots, y_s$ con $s < T$.

- Per valutare la *goodness of fit* si confronteranno le stime $\hat{y}_1, \ldots, \hat{y}_s$ con i dati osservati $y_1, \ldots, y_s$.
- Per valutare invece la capacità previsiva si confronteranno i valori previsti $\hat{y}_{s+1}, \ldots, \hat{y}_T$ con i dati osservati $y_{s+1}, \ldots, y_T$.

Indici sintetici di bontà di adattamento

Per valutare l'adattamento *in-sample*:

- Mean Error (ME): $ME = \frac{1}{s}\sum_{t=1}^{s} y_t - \hat{y}_t$
- Mean Square Error (MSE): $MSE = \frac{1}{s}\sum_{t=1}^{s}(y_t - \hat{y}_t)^2$
- Mean Absolute Error (MAE): $MAE = \frac{1}{s}\sum_{t=1}^{s}|y_t - \hat{y}_t|$
- Mean Absolute Percentage Error (MAPE): $ME = \frac{1}{s}\sum_{t=1}^{s}\frac{|y_t - \hat{y}_t|}{y_t}$

Per valutare la bontà di previsione *out-of-sample*:

- Mean Error (ME): $ME = \frac{1}{T-s}\sum_{t=s+1}^{T} y_t - \hat{y}_t$
- Mean Square Error (MSE): $MSE = \frac{1}{T-s}\sum_{t=s+1}^{T}(y_t - \hat{y}_t)^2$
- Mean Absolute Error (MAE): $MAE = \frac{1}{T-s}\sum_{t=s+1}^{T}|y_t - \hat{y}_t|$
- Mean Absolute Percentage Error (MAPE): $ME = \frac{1}{T-s}\sum_{t=s+1}^{T}\frac{|y_t - \hat{y}_t|}{y_t}$

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 in genere di considerarla assieme al trend, al che questa componente (T) viene detta *trend-ciclo*.

La componente deterministica f può assumere diverse forme funzionali:

- additiva: $Y_t = T_t + S_t + e_t$
- moltiplicativa: $Y_t = T_t \cdot S_t \cdot e_t$
- mista

Metodi di stima delle componenti

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

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

Le componenti T ed S possono essere stimate con metodi

- empirici (perequativi): consistono in un *lisciamento* che si adatta ai dati del campione permettendo di isolare una componente *in-sample* ma non permette di estrapolarla/prevederla
- analitici (interpolativi): consistono nella scelta di una *funzione analitica* di cui stimare i parametri, la quale si può poi usare per prevedere le singole componenti

Nell'approccio moderno tali procedure vengono sostituite dalla stima di un vero e proprio modello statistico parametrico.

Medie mobili

Le *medie mobili* sono un attrezzo (device) statistico utile per "lisciare" le oscillazioni casuali e mettere in evidenza le componenti sistematiche.
La media mobile di k termini è definita, *per k dispari*, come:

$$MM_k(y_t) = \frac{\sum_{s=t-(k-1)/2}^{t+(k-1)/2}}{k}$$

è insomma una media di k termini centrati su y_t. Per esempio, se $k = 5$,

$$MM_5(y_t) = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5}$$

Se k è pari, per esempio $k = 4$, si fa una media di $k + 1$ termini assegnando agli estremi un peso di 0.5:

$$MM_4(y_t) = \frac{0.5 \cdot y_{t-2} + y_{t-1} + y_t + y_{t+1} + 0.5 \cdot y_{t+2}}{4}$$

Destagionalizzazione usando le MM

Le medie mobili (MM) eliminano o riducono le oscillazioni di periodo pari all'ampiezza della media mobile. Es. una MM_{12} su dati mensili "filtra" le oscillazioni stagionali. Per ottenere una stima *in-sample* di T ed S si può procedere come segue (modello additivo):

- calcolo di $MM_{12}(y_t) = T_t^{(1)}$ come prima approssimazione a T_t (si perdono 6+6 termini all'inizio e alla fine)
- la serie detrendizzata $y_t - T_t^{(1)}$ è una stima di $S_t + e_t$ (*stagionalità grezza*)
- ipotizzando stagionalità costante (es. $S_t = S_{t-12} = S_{t+12} \ldots$),
  a) si calcola la media delle stagionalità grezze dei vari anni per ogni mese ottenendo 12 coefficienti stagionali $\hat{S}_j, j = 1, \ldots, 12$
  b) si verifica che la media $\bar{\hat{S}}$ degli $\hat{S}_j$ sia zero (principio di conservazione delle aree) altrimenti li si centrano $\hat{\hat{S}}_t = \hat{S}_j - \bar{\hat{S}}$

  ottenendo così i *coefficienti netti di stagionalità*

(continua:) Stima del trend

Per ottenere la stima "definitiva" del *trend-ciclo* $\hat{T}_t$

- si deriva la serie destagionalizzata $D_t = y_t - \hat{\hat{S}}_t$
- si stima il trend eliminando le oscillazioni casuali con un'ulteriore media mobile di ampiezza "opportuna"

Destagionalizzazione usando le MM

Le medie mobili

- sono una procedura di adattamento *in-sample*
- fanno perdere $k/2$ periodi all'inizio e alla fine, dove k è la frequenza annua dei dati osservati

pertanto questi metodi sono inadatti alla previsione *out-of-sample*. Essi sono invece utili ai fini interpretativi.
Una procedura alternativa è la stima di un trend $\hat{T}_t$ con metodi analitici, da cui ottenere $\hat{S}_t = y_t - \hat{T}_t$. Quindi si possono calcolare i coefficienti di stagionalità come visto sopra.

Stima analitica del trend

Per estrapolare (prevedere) i valori futuri di una serie storica è necessario innanzitutto stimare analiticamente e proiettare nel futuro la componente di trend (se la componente stagionale è stata assunta costante, la si può aggiungere successivamente). Possibili forme per la componente di trend in funzione *del tempo*):

- Trend lineare (o linearizzabile) nei parametri:
  - costante: $f(t) = \beta_0$
  - lineare: $f(t) = \beta_0 + \beta_1 t$
  - polinomiale (es.: quadratica): $f(t) = \beta_0 + \beta_1 t + \beta_2 t^2$
  - esponenziale: $f(t) = \beta_0 \beta_1^t$
    quest'ultima può essere linearizzata usando i logaritmi:
    $ln(f(t)) = ln(\beta_0) + ln(\beta_1)t$
- Trend non lineare né linearizzabile:
  - esponenziale modificata: $f(t) = K + \beta_0 \beta_1^t$
  - funzione di Gompertz: $f(t) = K \beta_0^{\beta_1^t}$

Metodi di stima dei parametri del trend

Destagionalizzazione mediante medie mobili (o aggregazione) e metodi di lisciamento in generale possono essere combinati con metodi analitici basati su un modello statistico per la stima e previsione del trend:

- I trend lineari o linearizzabili possono essere stimati (*ed estrapolati*) facilmente mediante un modello di regressione OLS.
- (La linearizzazione mediante logaritmi provoca distorsioni in previsione)
- Al modello OLS del trend vanno applicate tutte le considerazioni del Cap. 6 sulla "bontà" della stima:
  - proprietà degli errori
  - appropriata forma funzionale
  - stabilità strutturale
  - efficacia previsiva

Livellamento esponenziale

Il *livellamento esponenziale* nasce nel 1957 come metodo pragmatico per la previsione delle serie storiche basato sulle medie mobili. In seguito esso è stato giustificato teoricamente anche nel quadro della teoria "moderna" delle serie storiche come caso particolare dei modelli ARMA/ARIMA.

- Si supponga di disporre, al tempo t, di una serie di osservazioni

$$y_{t-n}, y_{t-n+1}, \ldots, y_{t-3}, y_{t-2}, y_{t-1}, y_t$$

  e di voler prevedere y_{t+1}
- Si potrebbe pensare di ricorrere a una media mobile "all'indietro" di alcuni termini
- L'idea alla base del livellamento esponenziale è di modificare l'approccio delle medie mobili attribuendo più importanza alle osservazioni più recenti e in particolare all'ultima y_t

Livellamento esponenziale semplice

Il *livellamento esponenziale costante* o *semplice* parte dall'ipotesi che la serie sia stazionaria in media

- In prima approssimazione, dato che la serie è stazionaria in media, si potrebbe prendere come previsore in $t+1$ la media aritmetica delle osservazioni:

$$\hat{y}_{t+1} = \frac{\sum_{j=n}^{0} y_{t-j}}{n}$$

  ma così si darebbe lo stesso peso a ogni osservazione.

- Il livellamento esponenziale semplice generalizza quanto sopra assegnando a ogni osservazione un peso:

$$\hat{y}_{t+1} = \frac{\sum_{j=n}^{0} \omega_j y_{t-j}}{\sum_j \omega_j}$$

  (nella media aritmetica è $\omega_j = \frac{1}{n} \forall j$)

Determinazione dei pesi

Nel modello di livellamento esponenziale costante si stabilisce che i pesi ω_j decrescono esponenzialmente fino a 0 al crescere della distanza da t.

- Si impone:

$$\omega_j = \alpha(1-\alpha)^j$$

  con $0 < \alpha < 1$ e $\sum^{\infty} \omega_j = 1$

- Sostituendo ricorsivamente ad ogni termine y_h la previsione fatta in $h-1$: $\hat{y}_h$, si ottiene il seguente modello:

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$$

  dove α è chiamato *parametro di livellamento* (smoothing).

Il livellamento esponenziale come correzione sequenziale degli errori di previsione

Si vede come il modello si fondi su una logica di *aggiornamento sequenziale*:

- la previsione a un passo $\hat{y}_{t+1}$ è una media dell'ultimo termine e di tutti i precedenti, sintetizzati nella previsione precedente $\hat{y}_t$.
- Inoltre, riscrivendo la formula come

$$\hat{y}_{t+1} = \alpha y_t + \hat{y}_t - \alpha \hat{y}_t = \hat{y}_t + \alpha(y_t - \hat{y}_t)$$

si nota come la previsione corrente $\hat{y}_{t+1}$ sia uguale alla precedente $\hat{y}_t$ modificata per l'errore di previsione $(y_t - \hat{y}_t)$ commesso al passo precedente, moltiplicato per il parametro di smussamento α.

Si adotta pertanto una logica di correzione sequenziale degli errori di previsione.

Come scegliere il parametro α?

A questo punto, rimane libero il parametro α:

- il criterio di *ottimalità* per la sua stima dovrà essere basato sull'impiego pratico del modello
- pertanto è naturale cercare $\hat{\alpha}$ tale da "minimizzare gli errori di previsione", per esempio sotto forma di somma dei quadrati:

$$min_{\hat{\alpha}} SS(\hat{\alpha}) = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

  la stima viene ottenuta con metodi numerici (p. es. *grid search*)
- un altro problema (minore) è come *inizializzare* la serie dei valori previsti, ovvero cosa sostituire per $\hat{y}_1$: si può usare y_1 o una media dei primi valori.

Il metodo di Holt e Winters

Consideriamo una serie storica non stazionaria in media, che ammette

- un trend
- una componente stagionale

Se la serie storica ammette una tendenza di fondo *localmente rettilinea*, un modo di adattare il livellamento esponenziale al caso è di scomporre il valore y_{t+1} in

- un livello medio in t, m_t, e
- un trend T_t tra il tempo t e $t+1$

In generale, su un intervallo di lunghezza Θ, il valore previsto della serie al tempo $t+\Theta$ sarà esprimibile come

$$\hat{y}_{t+\Theta} = \hat{m}_t + \hat{T}_t\Theta$$

Il metodo di Holt e Winters: stima - 1

Anziché stimare congiuntamente le due componenti, si scompone il procedimento utilizzando due modelli di livellamento esponenziale:

- uno per il livello medio

$$\hat{m}_t = \delta_1 y_t + (1 - \delta_1)(\hat{m}_{t-1} + \hat{T}_{t-1})$$

- e uno per il trend

$$\hat{T}_t = \delta_2(\hat{m}_t - \hat{m}_{t-1}) + (1 - \delta_2)\hat{T}_{t-1}$$

dove il primo modello ricostruisce, secondo il solito processo ricorsivo/di correzione dell'errore, il livello medio, il secondo il trend (il cui valore osservato T_t è definito come differenza tra i livelli medi).

Il metodo di Holt e Winters: stima - 2

Nel caso vi fosse una componente stagionale, ovvero

$$\hat{y}_{t+\Theta} = \hat{m}_t + \hat{T}_t\Theta + \hat{S}_{t+\Theta-s}$$

si aggiungerebbe una terza equazione:

- livello medio:

$$\hat{m}_t = \delta_1(\bar{y}_t - \hat{S}_{t-s}) + (1-\delta_1)(\hat{m}_{t-1} + \hat{T}_{t-1})$$

- trend:

$$\hat{T}_t = \delta_2(\hat{m}_t - \hat{m}_{t-1}) + (1-\delta_2)\hat{T}_{t-1}$$

- stagionalità:

$$\hat{S}_t = \delta_3(\tilde{y}_t - \hat{m}_t) + (1-\delta_3)\hat{s}_{t-s}$$

dove nell'equazione del livello medio $\bar{y}_t$ è il valore *destagionalizzato* di y_t, e nell'equazione della stagionalità $\tilde{y}$ è il valore *detrendizzato* di y_t.