1. MISURE STATISTICHE DI SINTESI

Dato T = numero di osservazioni disponibili nel campione di dati, è possibile calcolare per la generica variabile x:

- **Media (campionaria);** $\mu_x = \frac{1}{T}\sum_{i=1}^{T} x_i$
- **Varianza (campionaria);** $Var_x = \frac{1}{T}\sum_{i=1}^{T}(x_i - \mu_x)^2$.
    - Quest'ultima misura la "dispersione" ossia la variabilità di un fenomeno intorno al suo valore medio
    - E' funzione crescente della dimensionalità del fenomeno (utilità delle normalizzazioni)
    - In generale le serie nominali sono più "volatili" di quelle reali
- **Covarianza (campionaria);**
  $Cov_{x,y} = \frac{1}{T}\sum_{i=1}^{T}(x_i - \mu_x)(y_i - \mu_y)$
    - Descrive il segno e l'intensità dei co-movimenti tra due variabili economiche
    - Limiti: non ammette una scala finita e limitata di misurazione
- **Correlazione (campionaria);** $Corr_{x,y} = \frac{Cov_{x,y}}{\sqrt{Var_x}\sqrt{Var_y}}$
    - Stessa interpretazione della Covarianza, ma normalizzata in modo da essere compresa tra –1 e +1.
    - Molto usata negli studi di statistica applicata dall'economia e di facile interpretazione.
    - Valori vicini a zero indicano assenza di comovimenti (i movimenti delle due variabili sono scarsamente legati tra loro), valori vicini a |1| indicano l'esistenza di una relazione "importante" (di segno positivo o negativo) tra le due grandezze
    - **Attenzione a variabili fortemente trendizzate**
      (GDP e C USA; RGDP e M2 Usa; TFT e WRN Europa)

2. PRINCIPI DI ANALISI DEI DATI: LA REGRESSIONE LINEARE

Si supponga l'esistenza di una relazione di tipo lineare tra due variabili z e x, descritta dalla seguente generica equazione:

$z_t = \alpha + \beta x_t$

nella quale

- **Intercetta:** valore di α
- **Pendenza:** valore di β

Questo tipo di equazioni è alquanto ricorrente nei modelli economici teorici, (pensate alla relazione tra Consumi e Reddito: $C_t = \overline{C} + cY_t$), nei quali si concepisce il mondo come deterministico, ossia privo di fenomeni del tutto casuali e imprevedibili. Per questo motivo tale tipo di equazione può essere soddisfatta in modo esatto (e utilizzata per stime e previsioni) una volta che intercetta e pendenza siano note.

Nel concreto si pongono però due problemi:

1. Al di là della teoria, nella realtà ci troviamo in un ***Contesto stocastico***: esistenza di una componente di errore del tutto imprevedibile e non sistematica (il cosiddetto ***rumore***), ossia priva di significato economico spiegabile in base a qualche modello teorico, e non interpretabile alla luce della teoria economica. Essa è legata a:
   a. Errori di misurazione dei dati
   b. Non conoscenza del modello vero dell'economia
   c. Esistenza dell'imponderabile (Es. Twin Towers)

2. i parametri di intercetta e pendenza non sono noti

Alla luce di ciò si consideri una tipica equazione di comportamento o di equilibrio di un modello economico, ad es. la funzione dei consumi in un banale modello reddito/spesa di ispirazione keynesiana; essa diventa:

Es. $C_t = cY_t \Rightarrow C_t = cY_t + \varepsilon_t$ [1]

Nella [1] la variabile a *dx* del segno di uguale (Y) è la variabile indipendente[1], o variabile esplicativa, o ancora *regressore;* tramite essa si cerca di spiegare il comportamento della variabile che sta a *sx* del segno di uguale (C), ossia la variabile dipendente.

La [1] comprende ovviamente una componente di errore, rappresentata dal termine ε_t. Naturalmente, se guardiamo alle cause dell'esistenza di questa componente di errore (fattori I-III) e alle loro caratteristiche possiamo renderci conto che sia nel corso del tempo, sia ripetendo più volte la stessa regressione con campioni di dati diversi, gli "errori" sono destinati a compensarsi, cosicché il termine ε_t risulta **in media** uguale a zero.

Usualmente gli analisti economici dispongono di dati reali relativi a C e Y (e più in generale relativi alle variabili economiche), ma tali dati sono il risultato sia di meccanismi di tipo economico, sia anche dell'agire dei fattori a-c. Poiché gli analisti non conoscono il modello vero dell'economia (ossia non conoscono né il processo che genera i dati veri su C e Y, nè il vero ε_t e di conseguenza nemmeno il valore reale dei parametri dell'equazione [1]), devono accontentarsi di "stimare" la realtà, "subendo" il fastidio della componente non nota di errore dovuta ai fattori a-c.

[1] Potrebbero, ovviamente, esserci diverse variabili indipendenti

Obiettivo degli analisti (nel nostro esempio)**:** date serie storiche di dati su C e Y si vuole ottenere una stima della regressione lineare [1], ossia una stima del parametro c (propensione marginale al consumo) che non è noto, lasciando tutto ciò che non si può spiegare sulla base dei dati disponibili e dell'equazione [1] nel ***residuo di stima***. Quest'ultimo, chiamato $\hat{C}$ il valore della variabile dipendente ricostruito ***(fittato)*** sulla base di Y noto e di c stimato, è pari a: $(C_t - \hat{C}_t) = (cY_t - \hat{c}Y_t) = \hat{\varepsilon}_t$. Insieme alla stima dei parametri del modello (c in questo caso), si ottiene, pertanto, anche una stima della componente di errore.

Stabilito che l'obiettivo è **stimare i valori dei parametri dei modelli economici,** tra le diverse stime possibili si cercano quelle che godono di precise proprietà:

- **stime non distorte**: in media sono corrette, ossia, nel caso della [1], vale che se la stima viene ripetuta N volte, utilizzando N diversi campioni di dati, in media il c stimato coincide con quello vero. Se così è, e se è vero (come abbiamo concluso poco sopra) che $\frac{1}{N}\sum_{j=1}^{N}\varepsilon_j = 0$, allora

  $$\frac{1}{N}\sum_{j=1}^{N}\hat{\varepsilon}_j = \frac{1}{N}\sum_{j=1}^{N}(C_j - \hat{C}_j) = (cY_j - \hat{c}Y_j) + \varepsilon_j = (c - \hat{c})Y_j + \varepsilon_j = 0, \forall t = 1,..., T.$$

  L'obiettivo è quello di ottenere una stima della equazione [1] tale per cui **in media** l'errore commesso nella stima (*residuo di stima*), ossia $\hat{\varepsilon}_t$ è uguale a zero, così come dovrebbe valere per la componente vera di errore (ε_t,) indotta dai fattori I-III.
- **stime consistenti**: se avessimo un numero infinito di dati le stime (di c nel nostro caso) coinciderebbero col valore vero (in tal caso conosceremmo il modello vero dell'economia e il processo che genera i dati economici).
- **stime efficienti:** minimizzano la varianza dei residui di stima

In generale si vuole cioè che la stima di c (e dei parametri in generale), $\hat{c}$, sia tale che 1) $\frac{1}{T}\sum_{j=1}^{T}\hat{\varepsilon}_t = 0$ e 2) MIN [Varianza($\hat{\varepsilon}_t$)]

Tra le diverse tecniche di stima disponibili, lo stimatore a minimi quadrati ordinari è caratterizzato da tutte le proprietà appena elencate a patto che (insieme ad altre ipotesi) sia vero che i termini di errore hanno media nulla e contengono solo fenomeni non sistematici e privi di significato economico, insomma puro e semplice rumore.

Quando si ottengono stime che godono delle proprietà precedenti, ciò significa in generale che l'equazione è stata correttamente definita (***specificata***), ossia include a *dx* del segno di uguale tutto ciò che è utile a spiegare il comportamento della dipendente, cosicché **nel termine non spiegato ($\hat{\varepsilon}$) rimane solamente la**

componente di errore che è imputabile ai fattori I-III e che è impossibile spiegare per mezzo di strumenti economici.
Se, al contrario, abbiamo dimenticato di inserire tra i regressori qualcosa che è rilevante per spiegare il comportamento della dipendente, rimarrà nel termine $\hat{\varepsilon}$ qualcosa di sistematico, che ha un significato economico (non sono state incluse tra i regressori variabili economiche rilevanti per spiegare la dipendente), o ammette comunque una interpretazione statistica (es non sono state aggiunte *dummies* indispensabili per eliminare problemi "tecnici" come stagionalità o *break* strutturali).

Un indizio dell'esistenza di un problema di *cattiva specificazione* può essere tratto dall'analisi della media campionaria dei residui stimati, che dovrebbe essere molto vicina a zero e dal loro grafico che dovrebbe oscillare ripetutamente intorno allo zero, senza evidenziare alcun fenomeno di trend, né comportamenti sistematici e chiaramente interpretabili, né rotture o break di tipo statistico.

Notate che la regressione lineare può essere utilizzata non solo per stimare parametri economicamente interpretabili, ma anche semplicemente per "ripulire" l'andamento della variabile dipendente da fenomeni statistico-tecnici privi di contenuto e significato economico. Questi generano esclusivamente "rumore" e costituiscono elementi di disturbo nell'analisi economica. Al fine di effettuare tale operazione di "ripulitura", occorre: a) definire opportune variabili di comodo (le *dummies*) che sintetizzano il fenomeno statistico in questione, b) specificare una regressione nella quale vengono usate come regressori solo tali variabili *dummies*. **Come risultato della regressione si ha (in generale) che ciò che rimane nei residui di stima, ossia quella parte dell'andamento della dipendente che i regressori non riescono a spiegare, è in realtà il suo comportamento effettivo ed economicamente rilevante, depurato dai fenomeni di disturbo.**

1. Cosa osservare al termine di una regressione lineare

Dopo aver effettuato la stima è utile in primo luogo valutare la qualità dei risultati ottenuti e successivamente fornirne una interpretazione economica

a) Valutare la qualità di una regressione

Un utile indicatore di "bontà" delle stime ottenute è dato dal ***coefficiente di determinazione R^2***. Stilizzando il concetto, si può dire che Esso misura la percentuale della varianza della variabile dipendente che si è riusciti a spiegare per mezzo dei regressori inclusi nell'equazione stimata. Stilizzando il concetto, si può dire (con qualche approssimazione) che il coefficiente di determinazione ci dice quanto il set di regressori prescelto è capace di spiegare il comportamento della variabile dipendente. Ovviamente valori di R^2 prossimi a 1 indicano che la specificazione dell'equazione è valida e che buona parte del comportamento della dipendente viene spiegata. Al

contrario, valori di R^2 bassi e vicini a zero devono essere intesi come un campanello di allarme circa la qualità delle stime ottenute.
Un secondo strumento utile per valutare la qualità di una regressione è ***l'analisi dei residui di stima***: essi dovrebbero avere media nulla[2] e il loro andamento non dovrebbe evidenziare alcun comportamento sistematico o tendenziale, nessun effetto di trascinamento e nessuna "rottura" anomala. In caso contrario, la qualità della specificazione è debole e l'analista deve modificare il set di regressori incluso nel modello.

b) Interpretare le stime ottenute dal punto di vista economico: il problema della significatività statistica

L'interpretazione economica dei risultati si basa ovviamente sull'analisi della dimensione e del segno dei coefficienti stimati. Tuttavia, prima di procedere a tale operazione, occorre verificare che i coefficienti stessi risultino statisticamente significativi, ossia che il loro valore sia significativamente diverso da zero. Se così non è, nulla di sensato può essere detto a loro proposito. Il problema della ***significatività statistica*** nasce dal fatto che la stima dei coefficienti è soggetta a errore: se, stimando la [1] si ottenesse per la propensione marginale al consumo un valore pari a 0.2, ma la stima fosse afflitta da un margine di errore del tipo ±0.3, tutti i valori di c compresi tra –0.1 e 0.5 (compreso lo 0.2 ottenuto con la stima) sarebbero egualmente probabili e ragionevoli e non sarebbe possibile operare distinzioni tra essi. In particolare risulterebbe ragionevole anche il valore c=0.0, che starebbe a dire che il reddito non ha nessuna capacità di influenzare i consumi. In tal caso l'equazione [1] sarebbe del tutto inutile e non fornirebbe alcuna informazione economica rilevante. Ecco perché, prima di ragionare sul messaggio economico che viene dai valori stimati dei parametri, occorre escludere che tra i valori che essi possono ragionevolmente assumere sia compreso il valore 0. A questo proposito si usa un test statistico, basato su una statistica test che ha distribuzione del tipo t di Student; essa è ottenuta come rapporto tra il valore stimato del parametro e l'errore standard[3] ad esso associato. Se il valore assoluto di tale rapporto (***Statistica t***) è > di 1.645[4] allora il coefficiente è significativamente diverso da zero, il regressore ad esso associato è rilevante ed è sensato produrre ragionamenti di tipo economico in proposito[5].

- **Europa: Fare regressione di PCR su YER e costante (equazione dei consumi)**
- **Europa: Fare regressione di PCR su YER+TIN e costante (equazione dei consumi)**
- **Italia: Fare regressione di DK su GBYL e costante (funzione degli investimenti)**

[2] oltre ad altre proprietà che qui non possiamo commentare.
[3] Che è la radice quadrata della varianza dei valori che il parametro c può assumere.
[4] Per comodità si è deciso di utilizzare sempre il valore critico asintotico al 5%.
[5] Naturalmente per uno studio rigoroso del test t (ipotesi nulla, distribuzione del test, test a due e una coda, proprietà asintotiche) si rimandano gli studenti al corso di statistica.

- **Italia: Fare regressione di DK su GBYL, GDP, trend e costante (funzione degli investimenti estesa)**
- **Italia: Fare regressione di EXP su Lira/Marco e lira/dollaro con e senza dummy**

2. L'ANALISI DEL CICLO ECONOMICO: PREDISPORRE I DATI

- **Eliminare i break (Accidentalità e cambiamenti di regime)**

I break (punti di rottura) nelle serie possono essere: a GRADINO o a IMPULSO.

Anche in questo caso si ha interesse a rimuovere l'effetto "rottura" dalla serie, e ottenere serie "depurate" dai break

Un metodo per rimuovere i break dalla serie:

a) individuare esistenza e collocazione temporale dei break della serie attraverso un'analisi grafica e puntuale dei dati.
b) costruire una (o più) variabile dummy (D^b_t) che assume valore 1 in corrispondenza dei punti di rottura e 0 altrove.
c) eseguire la seguente regressione: $y_t=\alpha D^b_t + \varepsilon_t$.
d) definire i residui stimati della regressione: $\hat{\varepsilon}_t = \hat{y}_t - \alpha D^b_t$ che costituiscono la serie originale "depurata" dal break.
e) Fare un grafico dei residui stimati e della serie originale e confrontarli.

(Lit/DM per ITA, ITALABF per ITA, EER e EEN per Euro area).

- **Omogeneizzare i dati**

- **Destagionalizzare i dati**

Il problema della stagionalità: si trova in quelle serie che mostrano periodicamente (ogni stesso periodo dell'anno, tipicamente), un comportamento diverso in una particolare osservazione (ad es: nell'ottavo mese....) rispetto alle restanti osservazioni (...dell'anno). Non essendo un fenomeno che trasmette informazioni di natura economica rilevante si è interessati a rimuoverlo, "depurando" le serie storiche originali dai suoi effetti.

Alcuni metodi per rimuovere la stagionalità (***DESTAGIONALIZZAZIONE***): vengono presentati due metodi molto semplici per rimuovere la stagionalità. Il primo è basato sulle medie mobili. Il secondo è basato sulla regressione della variabile affetta da stagionalità su un insieme si variabili dummies stagionali.

- **Il metodo delle medie mobili** (supponiamo di usare dati trimestrali)

a) calcolare la media mobile a 5 termini della serie originale:
b) $x_t = (0{,}5y_{t-2} + y_{t-1} + y_t + y_{t+1} + 0{,}5y_{t+2})/4$
c) definire la variabile $_t = (y_t - x_t)$.
d) calcolare gli indici stagionali i_h come media, lungo il campione di osservazioni disponibili, dei $_t$ relativi a quel trimestre (con dati trimestrali si hanno 4 indici; ad es. per il 4° trimestre $i_h=i_4$).
e) aggiustare gli indici in modo che la loro somma sia uguale a 0:

f) $s_j = i_j - i^*$ dove $i^* = \frac{1}{4}\sum_{k=1}^{4} i_k$

g) ricavare il dato de-stagionalizzato sottraendo al dato originale del trimestre j-esimo l'indice s_j:

h) il tutto può essere iterato

(es:ITAIP per ITA e M1, M2, M3 per US)

- **Il metodo delle variabili *dummies*.** (supponiamo di usare dati trimestrali)

a) selezionare la variabile da de-stagionalizzare (y_t) e individuare il trimestre "q" in cui si riscontra il fenomeno stagionale.

b) eseguire la seguente regressione: $y_t = D^q_t + \ _t$.

c) dove D^q_t è una variabile dummy che assume valore 1 nei trimestri in cui si ha stagionalità e zero altrove.

d) La serie costituita dai residui stimati della regressione: $\hat{\varepsilon}_t = y_t - \hat{\beta} D^q_t$ rappresenta la serie originale destagionalizzata

e) Fare un grafico dei residui stimati (serie destagionalizzata) e della serie originale per confrontarli.

(es:ITAIP per ITA e M1, M2, M3 per US).

- **<u>Detrendizzare i dati</u>**
- Richiamo alla distinzione tra ciclo economico (oscillazioni di breve periodo) e tendenza di lungo periodo (crescita; equilibrio strutturale)
- La detrendizzazione serve ad eliminare da una serie economica la sua componente di lungo periodo solitamente rappresentata da un trend di tipo lineare.
- Ciò che resta a seguito della detrendizzazione è l'andamento ciclico della variabile.
- Schema di detrendizzazione:

  Regressione lineare: $y_t = \alpha + \beta T + \varepsilon_t$

  La serie detrendizzata è data da ε_t

Detrendizzare RGDP, UR e PRIND per l'Italia nonché RGDP per USA

- **Il Filtro HP**

- **Fare regressione di ITAIP su SEAS3 e costante (destagionalizzazione) e altro metodo**
- **Fare regressione di ITALABF su DUMLABF e costante (per eliminare break)**
- **Fare regressione di RGDP su TREND e costante (detrendizzazione) e altro metodo**