UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M. FANNO"

DIPARTIMENTO DI *SCIENZE STATISTICHE*

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

"SPIEGARE O PREVEDERE: COME UTILIZZARE I MODELLI STATISTICI"

RELATORE:

PROF.SSA LUISA BISAGLIA

LAUREANDO: LUCA CECCARELLI

MATRICOLA N. 1888280

ANNO ACCADEMICO 2021 – 2022

Dichiaro di aver preso visione del "Regolamento antiplagio" approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione 'Riferimenti bibliografici'.

I hereby declare that I have read and understood the "Anti-plagiarism rules and regulations" approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section 'References'.

Firma (signature) …Luca Coccarelli…

Indice

Capitolo 1

I modelli statistici tra spiegare e prevedere

1.1 Modelli statistici esplicativi e predittivi

I modelli statistici possono essere suddivisi tra **modelli statistici esplicativi** e **modelli statistici predittivi**. Nell'ambito della filosofia della scienza si è dibattutto a lungo su questa distinzione. Come afferma Shmueli (2010), solo recentemente si è, infatti, ammessa la necessità di considerare in maniera separata le due famiglie di modelli, rilevando come la distinzione tra il potere esplicativo e quello predittivo dei modelli statistici costituisca un elemento di analisi essenziale nella preliminare *selezione* e loro successiva applicazione in base all'obiettivo di studio. Nella costruzione di modelli statistici multivariati è, infatti, necessario chiarire se lo scopo dell'analisi sia esplicativo oppure predittivo. Da una parte l'analisi **esplicativa** rivolge la propria attenzione ad identificare i fattori che sono *causalmente* correlati ad un determinato esito, mentre dall'altra l'analisi **predittiva** si pone l'obiettivo di trovare la combinazione di fattori che meglio predicono un evento *futuro* (Sainani, 2014). Non è raro, tuttavia, che gli studiosi confondano il potere predittivo con quello esplicativo dei modelli in questione e che questo conduca a numerosi errori (Shmueli, 2010).

1.1.1 Potere esplicativo contro potere predittivo

Shmueli (2010) definisce gli elementi che differenziano l'approccio metodologico nelle analisi a scopo esplicativo rispetto quelle a scopo predittivo. "Perchè dovrebbe esserci una differenza tra spiegare e predire? La risposta risiede nel fatto

che i dati misurabili non sono rappresentazioni accurate dei costrutti sottostanti." La condensazione di costrutti e teorie in modelli statistici deve quindi essere operata in base all'obiettivo dello studio, sia questo esplicativo o predittivo.
Shmueli definisce queste differenze più formalmente: si consideri una teoria che si propone di spiegare come la variabile X determini Y attraverso la funzione F, cosìcchè $Y=F(\mathrm{X})$. L'obiettivo sarà quindi costruire un modello f a partire dalla funzione F tale che $E(Y) = f(X)$. Questo processo di costruzione dipende dalla funzione F che può essere scelta seguendo approcci differenti in base alla tipologia dei dati oggetto di analisi.
Le differenze emergono poichè nella costruzione di modelli *esplicativi* l'obiettivo è quello di adattare il più possibile la funzione f a F per la validazione delle ipotesi teoriche alla base dell'analisi. I dati X, Y costituiscono gli strumenti con i quali stimare f che verrà poi a sua volta utilizzata per l'applicazione dei test d'ipotesi sui dati. D'altro canto nella modellazione a scopo *predittivo* l'interessse ricade direttamente sulle variabili X e Y e la funzione f assume rilievo solo nella misura in cui permette di generare buone previsioni.
Shmueli (2010) e Sainani (2014) forniscono una rappresentazione del "sentiero di modellazione statistica" illustrandone le differenze in base allo scopo predittivo o esplicativo dell'analisi. In particolare Shmueli (2010) propone (Figura 1.1) il processo di modellazione statistica che si sviluppa in otto fasi:

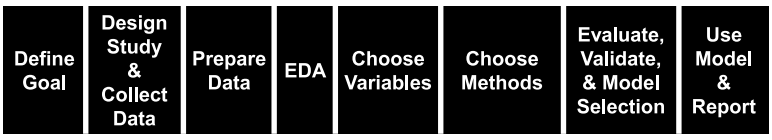


Figura 1.1: Le fasi del processo di modellazione statistica (Smueli, 2010).

(1) **Definizione dell'obiettivo**: questa fase permette di definire l'obiettivo dell'analisi statistica, sia questa esplicativa o predittiva.

(2) **Raccolta dei dati**: La scelta tra analisi predittiva ed esplicativa risulta, già nelle fasi iniziali, un elemento essenziale per portare avanti il processo di modellazione statistica. È in queste fasi che si deve, ad esempio, definire la metodologia di acquisizione dei dati e da questo punto di vista la dimensione del campione assume tra le altre cose un'importanza rilevante. Nella formulazione di modelli esplicativi infatti l'obiettivo primario è stimare in maniera più accurata possibile la funzione f con l'obiettivo di operare su di essa test d'ipotesi e procedure inferenziali. Per raggiungere questo obiettivo

è necessario ottenere un buon livello di potenza statistica e una ridotta distorsione. Questo richiede un campione sufficientemente grande. In questo tipo di modelli tuttavia tali vincoli dimensionali possono essere trascurati già dopo il superamento di una certa dimensione campionaria. Più stringenti risultano invece le caratteristiche dimensionali per modelli predittivi: il campione dev'essere sufficientemente ampio da ottenere una distorsione e una varianza ridotte.

(3) **Preparazione dei dati**: uno degli aspetti più importanti di cui tenere conto è la possibilità di dover operare con **valori mancanti**. L'obiettivo della modellazione, in entrambi i casi considerati, modifica totalmente l'approccio a questo problema. Come scrive Sarle (1998) davanti a dati mancanti l'approccio può essere duplice in base allo scopo della modellazione: predittiva o esplicativa. Se ad esempio risultano poche unità statistiche con dati mancanti, nel caso di modellazione *esplicativa* è possibile semplicemenete rinunciare a tali unità; diversamente, se l'obiettivo è fare delle *previsioni* su variabili che presentano dati mancanti questa opzione non è operabile.

(4) **Analisi esplorativa dei dati** : questa fase è comune sia alla modellazione predittiva che esplicativa, ma viene, tuttavia, utilizzata con fini differenti. Questo tipo di analisi consiste nella visualizzazione grafica o rappresentazione numerica dei dati. Se nella modellazione esplicativa questa fase è totalmente incanalata nell'impianto teorico sul quale si basa l'analisi, in ambito predittivo è utilizzata invece in maniera "libera" con l'obiettivo di visualizzare e trovare nuove relazioni tra variabili.

(5) **Scelta delle variabili** : anche la procedura di scelta delle variabili differisce in maniera piuttosto rilevante in base allo scopo del processo di modellazione statistica; nel caso di modelli esplicativi infatti la scelta della variabili e del loro ruolo è vincolata strettamente agli assunti teorici e alla "struttura causale" tra variabili. Nei modelli esplicativi la scelta delle variabili è inoltre cruciale al fine di evitare i cosìddetti errori da **variabile omessa** che si evidenziano tramite correlazione tra una delle variabili esplicative e il termine di errore. Nei modelli predittivi l'interesse è invece rivolto all'*associazione* tra variabili piuttosto che alla loro relazione *causale*. Questo fa sì che la scelta delle variabili possa avvenire in maniera meno stringente: i criteri rilevanti per operare una corretta scelta delle variabili in questo caso sono

piuttosto legati alla **qualità** dei dati e la *disponibilità* dei predittori che determinano la variabile risposta.

(6) **Scelta del modello** : la scelta del modello da utilizzare segue la definizione di modello predittivo o esplicativo sopracitata. L'obiettivo dei modelli esplicativi, come abbiamo già detto, si identifica nello stimare più accuratamente possibile la funzione f a partire da F. Per questo tipo di approccio sono particolarmente utilizzati modelli statistici come quello di regressione o il SEM (Structural Equation Modeling, che analizzeremo in particolare nel capitolo 2). Nel caso dei modelli predittivi, dove l'interesse è invece generare previsioni accurate, una famiglia di modelli ampiamente utilizzata è quella dei modelli di serie storiche (che analizzeremo in particolare nel capitolo 4). Un approccio particolarmente utile ma scarsamente utilizzato (Breiman, 2001) in ambito predittivo è inoltre la **modellazioni algoritmica**.

(7) **Validazione, Valutazione e Selezione dei modelli** : l'ultima fase del processo di modellazione statistica è identificata da Shmueli (2010) nella (1) *validazione* , (2) *valutazione* del comportamento del modello in base ai dati oggetto d'analisi e infine nella (3) *selezione* di un modello tra i vari disponibili in base alle esigenze:

   (a) per quanto riguarda i modelli esplicativi il processo di **validazione** si compone di due fasi: la prima fase (validazione del modello) guarda all'efficiacia di f di rappresentare adeguadatemente F; la seconda fase (adattamento del modello) verfica che $\hat{f}$ si adatti bene ai dati. Nel caso della modellazione predittiva l'interesse è rivolto invece alla *generalizzazione* del modello ovvero la capacità di $\hat{f}$ di prevedere nuovi dati.

   (b) la **valutazione** dei modelli statistici riguarda principalmente due aspetti: il potere esplicativo e il potere predittivo per i quali tali modelli si caratterizzano. Di conseguenza la priorità nella valutazione di modelli esplicativi sarà trovare un rilevante potere *esplicativo* ad essi associato e, al contrario, per modelli predittivi sarà prioritario ottenere buone *performance* nella predizione di nuovi dati.

   (c) Una volta *validati* e *valutati* adeguatamente, i modelli statistici subiscono l'ultima fase del processo, ovvero quella della **selezione**. Tra modelli esplicativi la selezione può avvenire identificando il modello

che possiede il maggiore potere esplicativo; da questo deriva la propensione degli studiosi all'utilizzo dei modelli "nidificati" (nested) poichè garantiscono un più semplice raffronto in termini di potere esplicativo. Per quanto concerne invece la selezione di modelli predittivi è possibile rilevare in letteratura alcuni criteri di informazione automatica come l'AIC o il BIC che permettono di valutare il potere predittivo dei modelli e di conseguenza operarne la selezione.

1.1.2 Modelli esplicativi e predittivi in pratica

1. Modelli esplicativi

Sono numerosi i campi di ricerca, soprattutto nell'ambito delle scienze sociali, nei quali i modelli statistici esplicativi vengono utilizzati con l'obiettivo di studiare relazioni *causali* tra variabili. Come abbiamo già discusso in precedenza in questi modelli si assume che una serie fattori, misurati da k variabili X_i con i = 1, 2, ..., k, determinino un **effetto** su una variabile detta Y. Un tipico esempio di modelli dal prevalente potere esplicativo sono i modelli di regressione che assumono la forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i \qquad \text{i} = 1, 2, ..., \text{n},$$

dove n è la numerosità capionaria (Shmueli e Koppius, 2009).

Dato il largo utilizzo di questa tipologia di modelli, non ci soffermeremo ulteriormente per spiegarne il funzionamento in maniera più dettagliata.

Un'altra importante famiglia di modelli largamente utilizzati in ambito esplicativo sono invece i **i modelli di equazioni strutturali**: I modelli di equazioni strutturali (**SEM**) sono modelli che combinano essenzialmente l'*analisi fattoriale confermativa*, nella quale il ricercatore opera delle ipotesi a priori sulle relazioni tra variabili (Corbetta, 2002) e la *path analysis*, volta ad analizzare invece l'intensità delle relazioni tra variabili per ottenere così informazioni sui rapporti causali sottostanti (Schumacker e Lomax, 2010). Rispetto ai modelli di regressione (che sono un caso particolare dei modelli di equazioni strutturali) i SEM risultano più *potenti* e suscitano perciò maggiore interesse nel perseguire gli obiettivi dell'analisi esplicativa. Avremo l'occasione di descrivere più nel dettaglio i modelli di equazione strutturali nel secondo capitolo.

2. Modelli predittivi

Nell'ambito dei modelli **predittivi** può essere utile fornire un'esempio che possa chiarire maggiormente il loro impianto teorico, la loro costruzione e la loro applicazione empirica.

Prendiamo in considerazione il caso di studio sviluppato da Kuhn et al. (2013). Questa applicazione si propone di prevedere il *risparmio di carburante* in base alla cilindrata del motore di un campione di veicoli. L'obiettivo dello studio era di applicare tali risultati alla successiva pianificazione produttiva. In questo modello introduttivo alla modellazione predittiva ci si limita all'utilizzo di un unico predittore, la *cilindrata dei veicoli*. Il modello in questione si propone di prevedere il risparmio di carburante (misurato in base all'indice di efficienza MPG "meters per gallon" per modelli d'auto del 2010 e del 2011) tramite l'unico predittore della cilindrata.

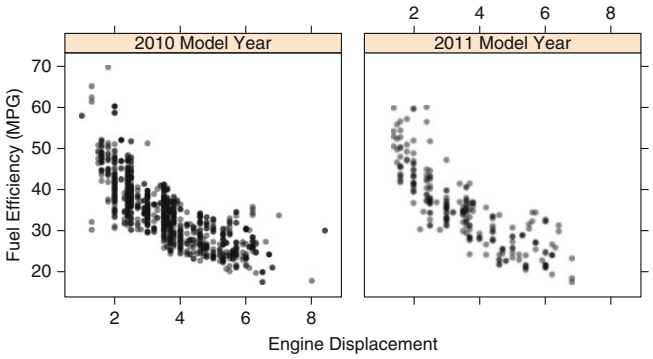


Figura 1.2: Relazione tra cilindrata e risparmi di carburante.
Applied Predictive Modeling - Kuhn M., Johnson K (2013)

(1) Il primo passo da seguire riguarda la comprensione dei dati oggetto d'analisi; per maggiore chiarezza si rappresentano tali dati graficamente con l'ausilio di uno scatter plot (vedi figura 1.2). Osservando i grafici per i modelli dell'anno 2010 e 2011 è possible notare che in entrambi i casi il riparmio di carburante diminuisce radicalmente all'aumentare della cilindrata delle automobili esaminate. La relazione tra le due variabili potrebbe essere rappresentata da una retta orientata negativamente; notiamo inoltre che la nuvola di dati presenta una curvatura per valori estremi della cilindrata.

(2) Una volta osservati i dati si passa alla costruzione del modello: un tipico approccio a tale problematica consiste nell'estrarre un **campione casuale** a partire dai dati oggetto di analisi per costruire il modello e utilizzare il resto dei dati

disponibili per analizzarne le prestazioni. Questo aspetto è molto importante nel processo di modellazione in generale: utilizzare lo stesso campione di dati per la costruzione e per la validazione di un modello potrebbe portare ad una a risultati **troppo ottimistici** in fase valutativa. L'obiettivo della modellazione è ottenere informazioni per la costruzione di una nuova linea di auto perciò possiamo utilizzare il campione proveniente dai modelli di auto del 2010 per costruire il modello e poi valutarne la "performance" sul campione dell'anno successivo. In altri termini usiamo i dati del 2010 come **training set** e quelli del 2011 come **validation set**.

(3) Il passo successivo è quello della modellazione vera e propria; come primo approccio potremmo cercare di ottenere una previsione del *risparmio di carburante* utilizzando un modello di regressione semplice. Utilizzando i dati provenienti dal "training set" si giunge alla stima di una retta di regressione OLS con intercetta 50.6 e pendenza -4.5 (figura 1.3). Il grafico seguente paragona i valori osservati e quelli previsti tramite il modello di regressione OLS. Questo modello non mostra rilevante potere predittivo. Come si può notare questo infatti sottostima il valore predetto del risparmio di carburante per valori di cilindrata inferiori a 2 MPG o superiori a 6 MPG.

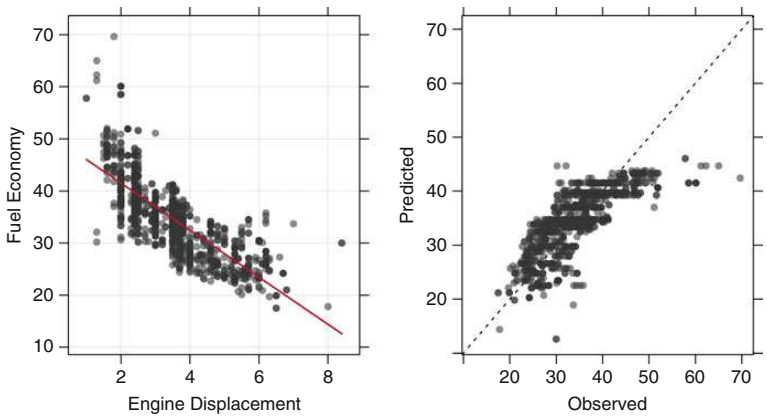


Figura 1.3: Il modello di regressione nell'analisi predittiva.
Applied Predictive Modeling - Kuhn M., Johnson K (2013).

(4) Come avevamo anticipato, la nuvola dei dati presenta una notevola curvatura; questo suggerisce una relazione non tanto lineare ma piuttosto quadratica tra le variabili *cilindrata* e *risparmio di carburante*. Questo ci permette di costruire un nuovo modello in cui stavolta il predittore "*cilindrata*" entri sia in forma lineare

che quadratica:

$$Risparmio\ di\ carburante = 63.2 - 11.9 * Cilindrata + 0.94 * Cilindrata^2$$

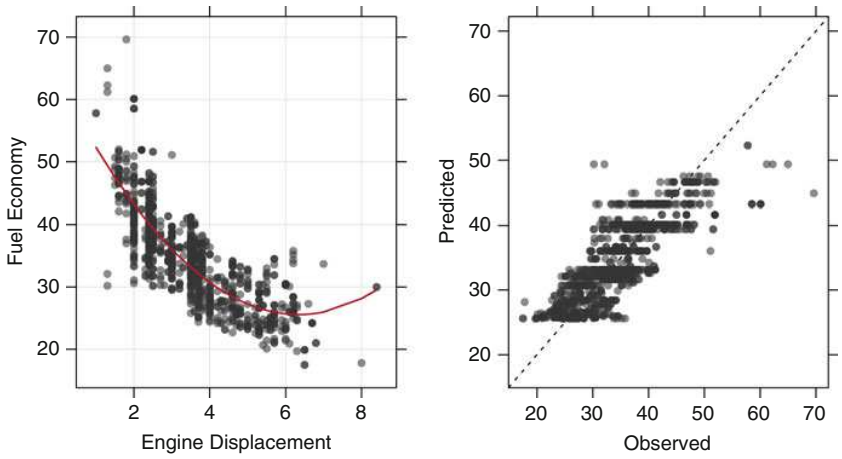


Figura 1.4: Bontà di adattamento del modello di regressione quadratico.
Applied Predictive Modeling - Kuhn M., Johnson K. (2013).

Notiamo subito come l'introduzione del predittore in forma quadratica nel modello migliori notevolmente il suo potere predittivo (figura 1.4).

Un ulteriore miglioramento al modello

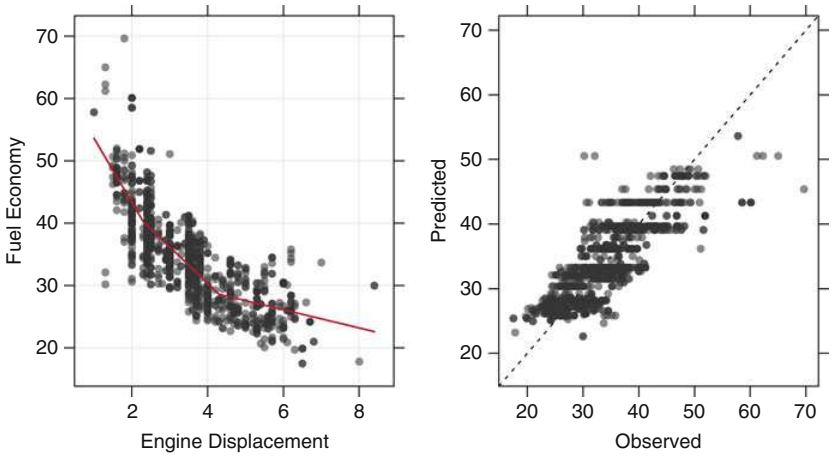


Figura 1.5: Bontà di adattamento del modello MARS sui dati del training set.
Apllied predictive Modeling - Kuhn M., Johnson K. (2013).

Per quanto l'obiettivo della modellazione a scopo predittivo sui dati forniti possa dirsi raggiunto, un ulteriore sviluppo della modellazione predittiva può essere portato avanti guardando ad un altro modello, stavolta più *sofisticato*, introdotto

da Friedman (1991): il modello MARS. Il modello MARS (Multivarate Adaptive Regression Spline) permette di costruire diverse rette di regressione, ciascuna per un diverso range di valori del predittore. É inoltre interessante sapere che il modello in questione contiene un *algoritmo* che permette di valutare il numero di rette di regressione che è necessario stimare in base ai dati oggetto di analisi.
La figura 1.5 mostra la migliore bontà di adattamento del modello MARS ai dati.

1.1.3 L'approccio *prevalentemente* esplicativo nella ricerca

Una volta definite le varie famiglie di modelli e compresa la differenza tra lo scopo *esplicativo* e *predittivo* nella modellazione statistica si rileva interessante sollevare la questione, sostenuta da numerosi studiosi (per esempio Shmueli e Koppius (2009), Ehrenberg e Bound (1993)) secondo cui in ambito accademico la modellazione con fini esplicativi continua ad essere sostanzialmente preferita a quella predittiva. Come afferma in merito Shmueli e Koppius (2009), in letteratura si registra la tendenza a trattare modelli prevalentemente esplicativi e molto più raramente a porre l'attenzione verso obiettivi di carattere predittivo. In merito a questo Shmueli sostiene che: (1) la modellazione predittiva permette una maggiore comprensione degli assunti teorici sottostanti. Ne è un esempio l'analisi del "comportamento d' intensificazione dell'impegno nello sviluppo di software" portata avanti da Keil et al. (2000). L'articolo mostra come l'approccio predittivo possa gettare le basi per una maggiore comprensione e una più lungimirante analisi teorica di quanto sia possibile fare secondo un approccio esclusivamente esplicativo. (2) La reticenza dei ricercatori verso l'analisi predittiva riguarda il riduttivo assunto per cui la scienza si occupa di spiegare i fenomeni mentre prevedere sia un'attività piuttosto pragmatica e non accademica. Infine, (3) talvolta è la confusione tra prevedere e spiegare che porta alla scarsa considerazione attribuita alla modellazione predittiva: il termine *predittore* (inteso come il ruolo della variabile che predice un certo effetto in un modello predittivo) è spesso utilizzato in maniera *allargata* per intendere invece ad esempio "variabile dipendente" o ancora il "potere predittivo" è spesso confuso per il "potere *esplicativo" di un modello.*

1.2 I modelli statistici

Una volta definite le differenze sostanziali che distinguono la modellazione statistica con scopo predittivo dalla modellazione con scopo esplicativo, è utile adden-

trarsi nella teoria statistica elencando e descrivendo alcuni dei modelli principalmente utilizzati nell'ambito della ricerca economica e non solo. A questo merito si è scelto di trattare una serie di importanti modelli statistici che possono essere intesi come lo sviluppo del modello di regressione lineare che, già ampiamente trattato ed utilizzato in letteratura, ci limiteremo a citare. Ciascuno dei modelli che definiremo di seguito può essere più o meno adatto allo studio di uno o l'altro fenomeno: questo dipende sostanzialmente dalle carattistiche che un certo fenomeno assume e, soprattutto, dalla tipologia di dati utilizzati. I modelli che analizzeremo nei prossimi paragrafi si mostrano inoltre adatti ad applicazioni con scopo sia esplicativo che predittivo pur mostrando ciascuno uno spiccato (ma non esclusivo) potere esplicativo.

Di seguito rivolgeremo l'attenzione a definire in particolare i:

1. Modelli **Binomiali e Multinomiali**
2. Modelli **Probit e Logit**
3. Modelli di **per dati Panel**

1.3 Modelli e applicazioni

Prima di focalizzare l'attenzione su ciascuna delle classi di modelli che intediamo analizzare, è utile fornire una definizione di quello che è, più in generale, un *modello statistico*.

La definizione di *Modello statistico* si poggia sull'idealizzazione fondamentale nell'inferenza statistica per cui i dati osservati y^{oss} costituiscono una realizzazione di un vettore casuale Y. In particolare y^{oss}=(y_1^{oss}, ..., y_n^{oss}) è un campione causale semplice di Y=(Y_1, ..., Y_n) dove le n componenti di y sono variabili casuali indipendenti ed identicamente distribuite (i.i.d)
Dal momento che la distribuzione di probabilità ,P^0, è il **modello probabilistico** generatore dei dati, il primo passo è quello di delimitare le forme ritenute possibili per P^0: in concreto si specifica una famiglia F di distribuzioni di probabilità almeno qualitativamente compatibili con y^{oss} che prende il nome di **modello statistico**.
Un modello statistico è, inoltre, pensato come un'*approssimazione della realtà*: la funzione F non è infatti definita per spiegare perfettamente i dati ma piuttosto

per darne una corretta rappresentazione ai fini di ricerca (Pace e Salvan , 2001).

Lo schema concettuale che porta alla definizione di un *Modello Statistico* è rappresentato nella figura 1.6: il punto di partenza consiste nell'osservazione della realtà che si concretizza con la raccolta dei **dati**. Sulla base di tali dati è, nella seconda fase, che si definisce un modello **probabilistico** che dia una corretta rappresentazione di quanto conosciamo della realtà e del modo in cui i dati sono stati ottenuti. Tramite analisi di carattere statistico è possibile infine trarre conlusioni sul modello e, in secondo luogo, sui dati stessi (Kroese e Chan, 2014).

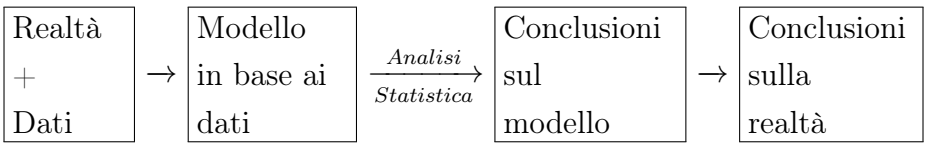


Figura 1.6: **Costruzione di un modello statistico** (Kroese e Chan, 2014)

1.3.1 Modelli Binomiali

Molte applicazioni della Statistica, anche in campo economico, riguardano la costruzione di modelli per variabili che posso assumere solo due valori. Tipicamente si assume che tali variabili, dette variabili **dicotomiche**, possano assumere solo i valori 0 o 1, associati rispettivamente al verificarsi (successo) o meno (insuccesso) di un particolare fenomeno.
L'obiettivo è quindi stimare la proporzione del *numero di successi y* sul *numero di prove n* che indica la probabilità π di successo nella singola prova. Il valore π, con $0 < \pi < 1$ rappresenta appunto la propensione di lungo termine dell'esperimento a realizzare successi nelle circostanze date.

Il **modello statistico** di base per dati binari $y_1, y_2, ..., y_n$ è che le osservazioni siano realizzazioni indipendenti di una variabile bernoulliana, Bi(1,π). L'osservazione y è quindi osservazione di una Bi(n,π), con funzione di probabilità:

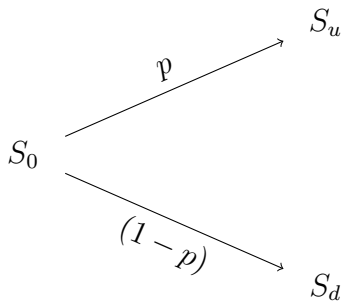
$$p_y(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}. \tag{1.1}$$

(Pace e Salvan , 2001).

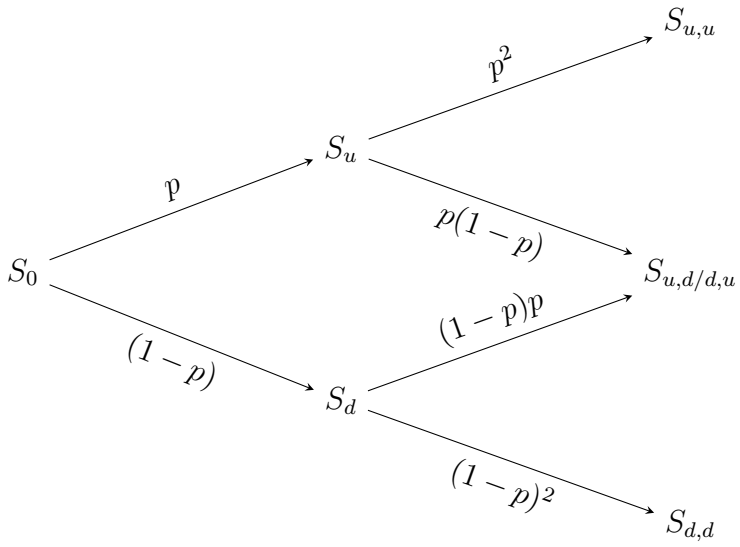
I modelli binomiali per l'analisi della volatilità

Un'interessante applicazione in ambito economico del modello binomiale riguarda l'analisi della volatilità degli strumenti finanziari sviluppata in un articolo di Erzegovesi (1999). Nell'analisi di volatilità degli strumenti finanziari si considera come "blocchetto elementare" la distribuzione binomiale a un periodo; per dare rappresentazione grafica di questo modello, costruiamo un albero binomiale che mostra l'evoluzione del prezzo di uno strumento finanziario durante un unico periodo. Si suppone che lo strumento abbia un certo prezzo al tempo zero e che dopo il periodo considerato (che chiamiamo step) questo possa subire un rialzo oppure, alternativamente, un ribasso.

In base a quanto appena descritto immaginiamo l'evoluzione del prezzo *spot* di un'attività infruttifera. Si supponga che al termine del periodo gli stati possibili siano solo due: un rialzo e un ribasso rispetto al prezzo al tempo zero di entità misurata rispettivamente dai fattori di variazione **u** (up) e **d** (down). Attribuiamo poi a ciascuno stato, rispettivamente al rialzo ed al ribasso la probabilità **p** e **(1-p)**. Indicando ora con S_0 il prezzo al tempo zero, S_u ed S_d i prezzi in caso di rialzo o ribasso, possiamo descrivere la distribuzione tramite una semplice rappresentazione grafica Erzegovesi (1999):



Possiamo compiere un passo avanti guardando all'orizzonte bi-temporale: in questo caso associamo ancora rispettivamente al rialzo **u** ed al ribasso **d** la probabilità **p** e **(1-p)**: il modello, su due periodi, avrà la seguente rappresentazione grafica:



Per quanto semplice e poco realistico rispetto alla complessità dei mercati in cui trovano reale formazione i prezzi delle attività finanziarie questo modello permette comunque di comprendere le plausibili variazioni dei prezzi su intervalli molto brevi ed è perciò la base sulla quale costruire modelli più complessi adatti a rappresentare più realisticamente le dinamiche di prezzo di medio termine.

1.3.2 Modelli Probit e Logit

I modelli di regressione **Probit** e **Logit** sono modelli di regressione *non lineari* specificatamente disegnati per variabili dipendenti binarie. Poichè una regressione con una variabile dipendente binaria Y modella la probabilità che Y=1, è ragionevole adottare una formulazione non lineare che costringa i valori previsti ad assumere valori compresi fra zero e uno.
Dato che la variabile dipendente Y può assumere solo i valori 0 o 1, nelle regressioni Probit e Logit si utilizzano rispettivamente, al posto della funzione lineare, la *funzione di ripartizione normale standard* e la *funzione di ripartizione "logistica"* in quanto producono valori di probabilità compresi proprio fra zero e uno (Stock e Watson, 2009).

Il modello Probit

Il modello di regressione **Probit** con un singolo regressore è

$$Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X), \tag{1.2}$$

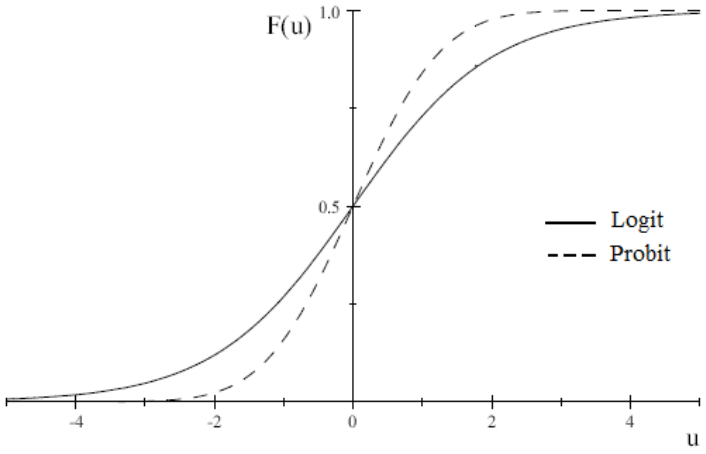


Figura 1.6: funzioni di ripartizione normale e logistica]**funzioni di ripartizione normale e logistica**
V. Verardi, Applied Microeconometrics Course, FUNDP.(2008)

dove Φ è la funzione di ripartizione normale standard (Stock e Watson, 2009).

Forniamo di seguito un'applicazione del modello Probit ad un campione di 127 osservazioni (Boston HMDA) presente nel volume di (Stock e Watson, 2009) per comprenderne il funzionamento: supponiamo di voler misurare la probabilità che un mutuo venga accettato o rifiutato conoscendo il reddito dei richiedenti e l'importo della rata di restituzione dello stesso.

Per fare questo utilizziamo appunto il modello di regressione lineare Probit nel quale Y è la variabile binaria del rifiuto del prestito *deny*, X il rapporto *rata/reddito* e la stima dei coefficienti $\widehat{\beta}_0$=-2 e $\widehat{\beta}_1$=3. Se consideriamo ad esempio rapporto rata/reddito pari a 0.4, la probabilità di rifiuto del prestito sarà data da $\Phi(\widehat{\beta}_0+\widehat{\beta}_1$*Rapporto rata/reddito) ovvero Φ(-2+3*0.4); secondo la tabella della funzione di ripartizione normale, Φ(-0.8) = Pr(Z$\leq$-0.8)=21,2%

La figura **1.2** appresenta una stima della funzione di regressione fornita dalla regressione Probit di *deny* su *P/I ratio* sulle 127 osservazioni effettuate; è possibile notare come per bassi valori del P/I (bassa rata di restituzione o alto reddito) ratio la funzione sia prossima a zero indicando una bassa probabilità di rifiuto del prestito mentre per elevati valori dell'P/I ratio (alta rata di restituzione o basso reddito) risulti prossima a uno suggerendo un'alta probabilità di rifiuto.

Il modello Probit appena descritto tiene conto di un unico regressore: questo oltre che poter indurre ad errori di *distorsione da variabile omessa* (se Y è correlata

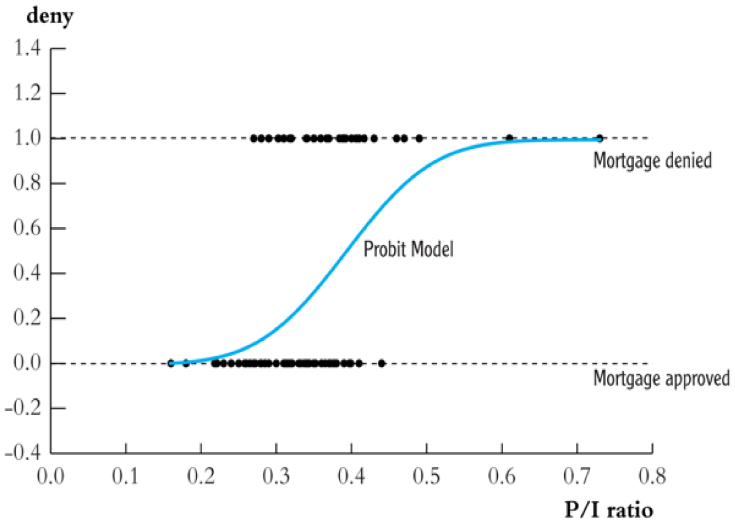


Figura 1.7: Modello probit della probabilità di rifiuto, dato P/I ratio. Stock J, Watson M - Introduzione all'econometria (2009) IV edizione

con l'unico regressore X) non rappresnta il caso piu generale che sia possibile definire:
Il modello probit con **regressori multipli** generalizza il modello probit con un unico regressore aggiungendo altri regressori nel calcolo del valore di z; così il modello probit con regressori multipli X_1,X_2,...,X_k è:

$$Pr(Y = 1|X_1, X_2, ..., X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k) \tag{1.3}$$

(Stock e Watson, 2009).

Il modello Logit

Il modello di regressione **Logit** con regressori multipli è dato da:

$$Pr(Y = 1|X_1, X_2, ..., X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k), \tag{1.4}$$

dove F indica la funzione di ripartizione **logistica** invece che la funzione di ripartizione normale standardizzata usata nel modello probit. Conoscendo la funzione di ripartizione logistica $F(x) = \dfrac{1}{1 + e^{-x}}$ (Kroese e Chan, 2014) possiamo riformulare il modello come:

$$Pr(Y = 1|X_1, X_2, ..., X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}} \tag{1.5}$$

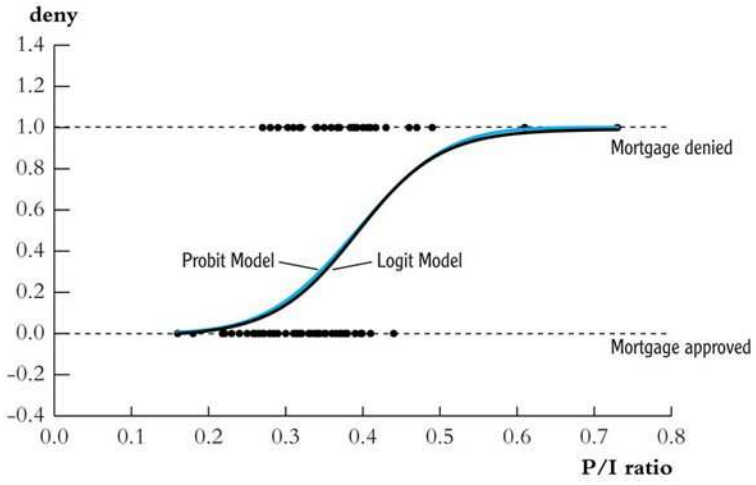


Figura 1.8: Modelli probit e logit della probabilità di rifiuto, dato P/I ratio.
Stock J, Watson W. introduzione all'econometria, (2009) IV edizione

Le funzioni di regressione probit e logit sono tra loro simili. Questo è illustrato nella figura **1.3**, che mostra le funzioni di regressione probit e logit stimate con il metodo della massima verosimiglianza sul campione di 127 osservazioni analizzato in precedenza; inoltre se storicamente la regressione logit risultava preferibile per la maggior rapidità di calcolo della funzione di ripartizione logistica rispetto a quella normale standard utilizzata nell regressione probit, ad oggi grazie a strumenti di calcolo più avanzati queste differenze si assottigliano ulteriormente (Stock e Watson, 2009).

Stima dei minimi quadrati non lineari

Nei modelli di regressione probit e logit i coefficienti entrano nella funzione di regressione in modo **non lineare**. Quello dei *minimi quadrati non lineari* è un metodo generale per stimare i parametri ignoti di una funzione di regressione in questi casi. Lo stimatore dei minimi quadrati non lineari estende lo stimatore OLS alle funzioni che non sono lineari nei parametri.
Consideriamo lo stimatore dei minimi quadrati del modello probit. Il valore atteso condizionato di Y date le variabili esplicative X è $E(Y \mid X_1,...,X_k) = Pr(Y = 1 \mid X_1,...,X_k) = \Phi(\beta_0+\beta_1X_1+...+\beta_kX_k)$. Lo stimatore dei minimi quadrati non lineari dei coefficienti è dato dai valori b_0, ..., b_k che minimizzano la somma dei quadrati degli errori di previsione:

$$\sum_{i=1}^{n}[Y_1 - \Phi(b_o + b_1X_1 + ... + b_kX_{ki})]^2 \tag{1.6}$$

Lo stimatore dei minimi quadrati non lineari ha due proprietà in comune con lo stimatore OLS utilizzato nella regressione lineare: è **consistente** e si distribuisce **normalmente** in grandi campioni (Stock e Watson, 2009).

1.3.3 Modelli per dati di Panel

I **dati di panel**, anche detti **dati longitudinali**, sono dati relativi a n unità diverse osservate in T periodi temporali diversi. Se i dati contengono osservazioni sulle variabili X e Y, potremmo indicarli come:

$$(X_{it},\, Y_{it}),\ \text{i} = 1,...,\text{n e t} = 1,...,\text{T}$$

dove il primo pedice i indica l'unità oggetto di osservazione, mentre t si riferisce al momento in cui quasta viene osservata (Stock e Watson, 2009). Se ognuna delle unità statistiche è osservata in ciascuno degli istanti temporali considerati, si parlerà allora di panel **bilanciato**; se per qualche istante temporale alcune delle osservazioni risultano assenti, parleremo invece di panel **non bilanciato** (Gujarati, 2003).

Riportiamo ora un'applicazione del modello di regressione per dati di panel su dati riguardanti incidenti mortali e imposte sugli alcolici nei 48 stati americani tra il 1982 e il 1988 sviluppata nel volume Stock e Watson (2009).
L'obiettivo è analizzare la relazione esistente tra la variabile dipendente Y=*tasso di mortalità* e la variabile esplicativa X=*imposta sulla birra* per studiare quanto gli interventi pubblici volti a limitare la guida in stato d'ebrezza siano efficaci nel ridurre il tasso di mortalità sulle strade. Iniziamo costruendo due modelli di regressione basati sui dati del 1982 e del 1988 separatamente.
Esaminando la relazione tra la due variabili rispettivamente nell'anno 1982 e 1988 otteniamo le seguenti rette di regressione OLS:

$$1\ \widehat{TassoDiMortalità} = 2{,}01 + 0{,}15\, ImpostaSullaBirra \qquad \text{(dati 1982)}$$

$$2\ \widehat{TassoDiMortalità} = 1{,}86 + 0{,}44\, ImpostaSullaBirra \qquad \text{(dati 1988)}$$

Possiamo notare che contrariemante a quanto atteso il coefficiente stimato con i dati del 1982 e del 1988 è *positivo*. Questo implica una relazione positiva tra imposte e tasso di mortalità: ad un aumento delle imposte sulla birra è associato un *maggiore* e non minore tasso di mortalità. Questa relazione positiva potrebbe non essere però la conclusione corretta dal momento che queste regressioni

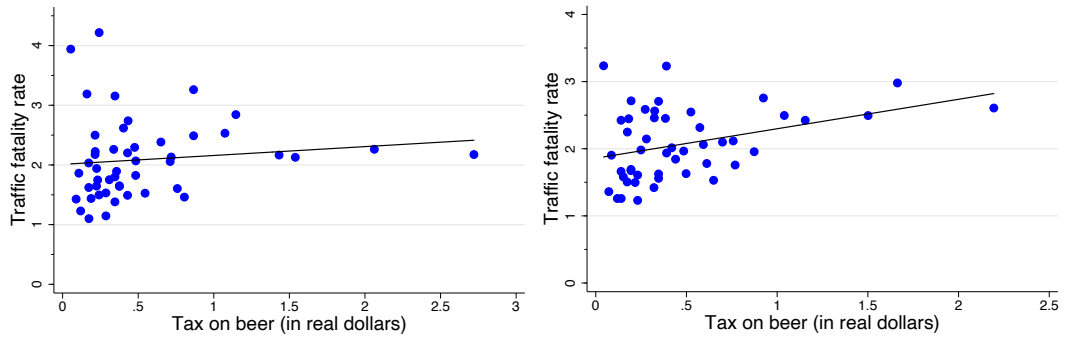


Figura 1.9: tasso di mortalità sulle strade e imposta sulla birra, 1982 - 1988.
Stock J, Watson M. - introduzione all'econometria (2009) edizione IV

potrebbero essere soggette a una sostanziale distorsione da variabili omesse. La distorsione da variabili omesse può essere ridotta introducendo altre variabili nel modello: il grado di manutenzione delle strade o l'intensità del traffico sono ad esempio alcune delle variabili che intuiamo facilmente determinare, insieme alla variabile "imposta sulla birra", il tasso di mortalità sulle strade. Talvolta però, come nel nostro caso, queste variabili possono non essere direttamente o facilmente misurabili. É possible tuttavia seguire una strada alternativa nel caso in cui tali variabili siano caratterizzate dal fatto di *non* variare nel tempo. In questi casi particolari una soluzione è infatti quella di costruire un modello di regressione a *fattori fissi* (di cui vedremo il funzionamento fra poco): questo permette di risolvere, da una parte, il problema della distorsione da variabili omesse tramite la loro esplicitazione nel modello e, dall'altra, quello della loro non misurabilità poichè vengono considerate costanti nel tempo. (Stock e Watson, 2009).
É importante sottolineare tuttavia che un modello di regressione ad effetti fissi non riesce a trarre alcuna conclusione sulle variabili omesse che *non* cambiano nel tempo poichè sono tra loro incorrelate e ciascuna esiste in maniera totalmente indipendente dalle altre (Bell et. al, 2019).

Il confronto "prima e dopo"

Se, come nel caso in esame, si dispone di dati per T = 2 periodi, è possibile confrontare i valori della variabile dipendente nel secondo periodo con quelli del primo. Chiamiamo Z_i una variabile che determina il tasso di mortalità nell'i-esimo stato, ma che non varia nel tempo (omettiamo perciò il pedice t); ad esempio Z_i può essere l'atteggiamento culturale in un certo luogo verso la guida in stato d'ebbrezza che supponiamo costante tra il 1982 e il 1988.
La regressione che mette in relazione Z_i e l'imposta sulla birra con il tasso di

mortalità sarà:

$$TassoDiMortalità_{it} = \beta_0 + \beta_1 ImpostaSullaBirra_{it} + \beta_2 Z_i + u_{it} \quad (1.7)$$

Possiamo costruire la stessa equazione per ognuno degli anni 1982 e 1988:

$$TassoDiMortalità_{i1982} = \beta_0 + \beta_1 ImpostaSullaBirra_{i1982} + \beta_2 Z_i + u_{i1982} \quad (1.8)$$

$$TassoDiMortalità_{i1988} = \beta_0 + \beta_1 ImpostaSullaBirra_{i1988} + \beta_2 Z_i + u_{i1988} \quad (1.9)$$

Se sottraiamo ora l'equazione "$TassoDiMortalità_{1988}$" dall'equazione "$TassoDi Mortalità_{1982}$" l'effetto di Z_i si annulla e otteniamo:

$$TassoDiMortalità_{1988} \text{ - } TassoDiMortalità_{1982} = \beta_1(ImpostaSullaBirra_{i1988} \text{ - } ImpostaSullaBirra_{i1982}) + u_{i1988} \text{ - } u_{i1982}$$

Utilizzando questo approccio la retta di regressione OLS è:

$$TassoDiMortalità_{1988} \text{ - } TassoDiMortalità_{1982} = \text{ - } 0{,}072 \text{ -}1{,}04^* (ImpostaSullaBirra_{i1988} \text{ - } ImpostaSullaBirra_{i1982})$$

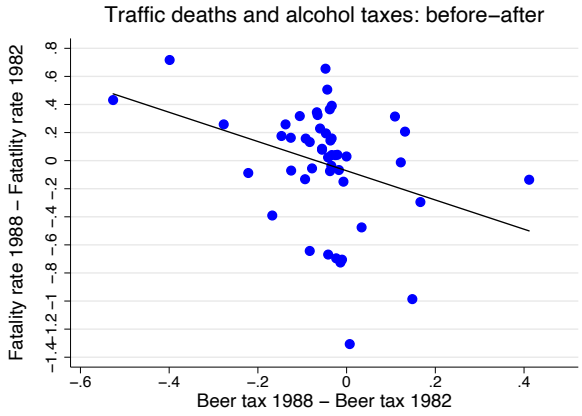


Figura 1.10: variazione dei tassi di mortalità e imposte sulla birra, 1982-1988.
Stock J, Watson M. - introduzione all'econometria (2009) edizione IV

In questo caso la pendenza negativa della retta di regressione (figura 1.10) lascia immaginare una relazione **negativa** tra l'inasprimento delle imposte sulla birra e il tasso di mortalità, in accordo con la teoria economica (Stock e Watson, 2009).

Capitolo 2

SEM: Modelli di Equazioni Strutturali

Come abbiamo avuto l'occasione di anticipare nel capitolo precedente una delle famiglie di modelli statistici maggiormente utilizzate con scopo esplicativo sono i **modelli di equazioni strutturali** o **SEM** (Structural Equation Model).

I SEM possono essere considerati un insieme di tecniche di analisi statistica tra loro *affini*: è per questo che non è possibile individuare un'unica fonte cui attribuire per intero la loro teorizzazione (Kline, 2015). I primi studi relativi ai modelli di equazioni strutturali, attribuibili a Jöreskog (1971), possono essere infatti considerati il frutto dei numerosi progressi precedenti nell'ambito della modellazione e dell'analisi statistica. Possiamo identificare tre famiglie di modelli statistici che hanno costruito i fondamenti per la formulazione successiva dei SEM. La prima famiglia è quella dei modelli di regressione lineare, nati dagli studi sulla correlazione tra variabili portata avanti da Karl Pearson già verso la fine del diciannovesimo secolo (Pearson, 1895). La seconda è quella dei modelli di analisi fattoriale, sviluppati da Charles Spearman e ripresi da Howe (1955), Andreson e Rubin (1956) e Lawley (1958) a cui si devono in particolare i modelli *confermativi* di analisi fattoriale. È infine a Sewell Wright (1918, 1921, 1934) che è riconducibile la famiglia dei modelli di *Path Analisys*, fondamentali per la costruzione dei SEM (Kline, 2015).

Queste famiglie di modelli esauriscono così gli elementi necessari per comprendere con chiarezza il percorso di sviluppo dei SEM. I modelli di equazioni strutturali *combinano* infatti l'**analisi fattoriale** e la **path analisys** nello studio delle rela-

zioni causali tra variabili: è proprio nei primi anni 70 che (con i prima citati studi di Joreskorg e le pubblicazioni di Keesling (1972) e Wiley (1973)) l'approccio *misurativo* dell'analisi fattoriali e quello *strutturale* dell'analisi di percorso vengono integrati nella formulazione dei modelli di equazioni strutturali originariamente identificati con il l'acronimo LISREL (LInear Structural RELation) (Schumacker e Lomax, 2010).

2.1 L'analisi fattoriale

Con il termine **analisi fattoriale** ci si riferisce in realtà ad un'insieme di tecniche di analisi statistica che hanno l'obiettivo di "esprimere un insieme di variabili osservate nei termini di un numero inferiore di fattori" (Corbetta, 2002). L'elemento di partenza dell'analisi fattoriale è la costruzione di una cosiddetta *matrice delle covarianze* fra le variabili osservate: nell'analisi fattoriale si assume infatti che la correlazione tra le variabili osservate possa essere giustficata dall'esistenza di fattori sottostanti. Se ad esempio si volesse condurre un analisi fattoriale tra le variabili osservate "voto in algebra lineare" e "voto in analisi", l'impianto teorico dello studio sarebbe fondato sull'esistenza di un fattore, come ad esempio "la capacità di astrazione matematica degli studenti", che possa giustificare la correlazione tra le variabili (Corbetta, 2002).
L'analisi fattoriale ha tradizionalmente due scopi fondamentali: uno prevalentemente **confermativo** ed uno invece **esplorativo**. L'approccio **confermativo** richiede, *a priori* , la rigida formulazione di un impianto torico da parte del ricercatore circa le relazioni tra le variabili oggetto d'esame.

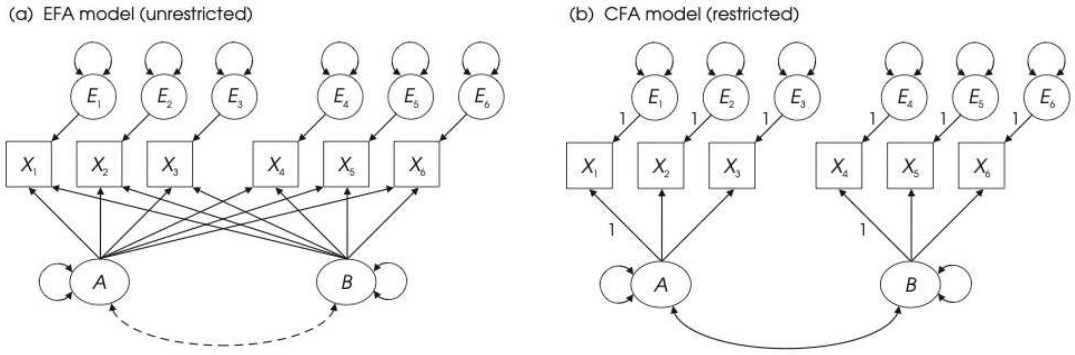


Figura 2.1: Analisi fattoriale confermativa ed esplorativa a confronto tramite path models. prnciples and practice of Structural Equation Modeling. Rex B. Kline, 2015.

In questo caso si può ad esempio assumere che i fattori non siano tra loro correlati oppure che un fattore influenzi solo alcune delle variabili del modello.

Nell'approccio esplorativo l'obiettivo è invece quello di esplorare le possibilli relazioni causali tra variabili, senza definire un modello *a-prioristico*. Questo presume che il ricercatore possa anche non fissare un numero definito di fattori latenti da ricomprendere nel modello o anche che possa non limitare l'influenza dei fattori a solo alcune delle variabili del modello stesso (questo è visibile dal maggior numero di frecce unidirezionali nel *path model* **fig. 2.1**) (Kline, 2015).
Da quanto appena accennato sull'analisi fattoriale si rileva chiaramente come lo scopo del suo utilizzo sia quindi verificare l'effetto di una serie di fattori *latenti* (e perciò non direttamente misurabili) sulle variabili oggetto di studio. Come vedremo nei paragrafi successivi questo obiettivo è proprio quello che si prepongono di raggiungere i **SEM** tramite la costruzione di modelli che incorporano variabili osservate e latenti e ne analizzano la relazione causale.

2.2 I modelli di equazioni strutturali

Agli inizi del ventesimo secolo la ricerca econometrica stava affrontando il problema delle relazioni causali tra variabili con la costruzione dei cosiddetti "*modelli di equazioni simultanee*" (Henry Shultz, 1938). In questo tipo di approccio, differentemente da quanto di interesse nei modelli di regressione, l'obiettivo era analizzare la relazione causale tra variabili, costruendo però un modello in cui queste potessero assumere *simultaneamente* il ruolo di variabili **dipendenti** ed **indipendenti** (Dragan e Topolšek, 2014). Ed ecco che, come vedremo fra poco, l'interesse dei modelli di equazioni strutturali introdotti qualche decennio dopo, sia proprio quello di analizzare, non solo le relazioni tra una variabile dipendente ed una serie di regressori, ma guardare invece *simultaneamente* alla relazione causale che intercorre tra tutte le variabili ricomprese nel modello. A questo merito, utilizzando la terminologia tipica dell'econometria ci riferiremo alle varibili di un modello di equazioni strutturali dividendole tra *endogene* ed *esogene*: consideremo quindi **esogene** quelle variabili che posso assumere *esclusivamente* il ruolo di variabile indipendente mentre chiameremo **endogene** le variabili che possono assumere il ruolo di varibili alternativamente dipendenti o indipendenti nelle equazioni strutturali che lo costituiscono (Corbetta, 2002).

2.2.1 L'intuizione dietro i modelli di equazioni strutturali

Supponiamo di voler analizzare la relazione di causalità tra le variabili X_1, X_2, X_3 e X_4 dove X_4 indica il successo scolastico dello studente, X_3 le aspirazioni

professionali degli studenti oggetto di esame, X_2 le aspirazioni professionali dei genitori degli studenti ed infine X_1 lo status socio-economico delle famiglie di provenienza. Rappresentiamo ora graficamente le relazioni tra variabili secondo l'approccio tradizionale della *path analisys* (figura **2.2** e **2.3**).

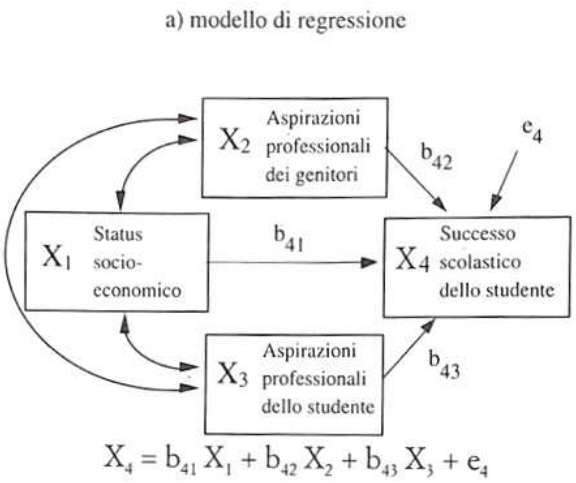


Figura 2.2: rappresentazione grafica delle relazioni tra variabili nel modello di regressione.
Metodi di analisi multivariata per le scienze sociali, Corbetta (2002).

Impostando il problema secondo un modello di regressione e indicando X_4 come variabile *dipendente* otteniamo l'equazione *strutturale*:

$$X_4 = b_{41}X_1 + b_{42}X_2 + b_{43}X_3 + e_4$$

Come notiamo dalla rappresentazione grafica del modello ciascuna delle variabili X_1,X_2 e X_3 determina X_4: questa relazione è sottolineata dalle frecce unidirezionali tra le variabili. Un modello simile non garantisce però informazioni aggiuntive sulle relazioni causali tra le variabili indipendenti: l'unico aspetto che traspare è la correlazione non nulla (evidenziata dalle frecce bidirezionali) tra le variabili indipendenti del modello, senza però che ne vengano specificate le relazioni causali.

Differentemente da quanto catturato dal modello di regressione, un **modello di equazioni strutturali** si propone di analizzare non solo la relazione tra una variabile dipendente del modello e le restanti variabili indipendenti, ma di ampliare lo studio anche alle relazioni causali tra le variabili indipendenti. Seguire la logica dei modelli di equazioni strutturali ci permette perciò di tenere conto delle relazioni che è facile immaginare esistere tra le variabili indipendenti: lo *status socio-economico* (X_1) potrebbe ad esempio influenzare le *aspirazioni professionali degli studenti* (X_3) che a loro volta potrebbero essere influenzate dalle

aspirazioni professionali dei genitori (X_2). Questa relazione è evidenziata in una delle quattro equazioni (la seconda) che costituiscono il **modello di equazioni strutturali**, riportato nella figura 2.3.

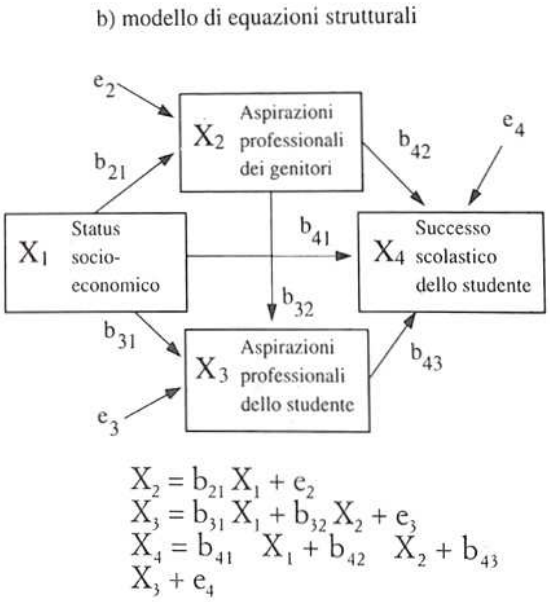


Figura 2.3: rappresentazione grafica delle relazioni tra variabili in un modello ad equazioni strutturali. Metodi di analisi multivariata per le scienze sociali, Corbetta (2002).

Passare dal modello di regressione ad un modello di equazioni strutturali permette allora di cogliere con maggior chiarezza le effettive relazioni tra le variabili: così come avviene nel caso della regressione lineare i coefficienti b vengono stimati a partire dai dati (con tecniche differenti dal *metodo dei minimi quadrati*) e informano circa gli effetti che, in ciascuna delle n-equazioni del modello, la variazione unitaria di una variabile indipendente ha sulla variabile dipendente.
Nel caso della regressione lineare i coefficienti b permettono di stimare l'effetto solamente *parziale* della variazione unitaria di una variabile indipendente sulla variabile dipendente: se si vuole stimare l'effetto di una variazione unitaria di X_2 su X_4 il coefficiente b_{42} informerà solo sull'effetto *parziale* di tale variazione. Questo è dovuto al fatto che, in questo modo, non teniamo conto dell'influenza che a sua volta X_1 ha su X_2: l'effetto *totale* di una variazione unitaria di X_2 su X_1 può essere quindi valutato solo guardando congiuntamente al coefficiente b_{42} e al coefficiente b_{21}.

Ecco che, in questo senso costruire un modello di equazioni strutturali può essere molto più informativo sulle relazioni causali tra variabili di quanto lo sia un modello di regressione lineare (Corbetta, 2002).

2.2.2 Formulazione teorica dei SEM

La formulazione originale dei modelli di equazioni strutturali è dovuta a Karl Jöreskog e D. Sörbom agli inizi degli anni '70 (Jöreskog, 1971). I modelli di equazioni strutturali rivolgono la loro attenzione all'analisi della relazione di causalità non solo tra le variabili osservate ma anche con le cosiddette variabili **latenti**. Quest'ultime sono variabili non direttamente rilevabili e che quindi vengono "misurate tramite due o più indicatori rilevabili" (Faraci e Musso, 2013).

Consideriamo di seguito la specificazione dei modelli di equazioni strutturali fornita da Karl Jöreskog e che segue la notazione LISREL che definiamo nella tabella di seguito.

tabella 2.1.2. La notazione LISREL (Corbetta, 2002).

η	variabili latenti endogene	ξ	variabili latenti esogene
Y	variabili osservate endogene	X	variabili osservate esogene
ζ	errori stocastici delle variabili η	ϵ	errori stocastici delle variabili Y
δ	errori stocastici delle variabili X	λ_y	coefficienti strutturali tra η e Y
λ_x	coefficienti strutturali tra ξ e X	β	coefficienti strutturali tra η e η
γ	coefficienti strutturali tra ξ e η	ϕ	varianze-covarianze tra le variabili ξ
ψ	varianze-covarianze tra gli errori ζ	θ_ϵ	varianze-covarianze tra gli errori ϵ
θ_δ	varianze-covarianze tra gli errori δ		

Il modello LISREL può essere descritto come segue:

consideriamo il vettore η' = (η_1, η_2, ..., η_m) delle osservazioni della variabile latente η e il vettore ξ' = (ξ_1, ξ_2, ..., ξ_n) delle realizzazioni della variabile osservata ξ e consideriamo l'equazione strutturale che ne definisce la relazione lineare come:

$$\eta = \boldsymbol{B}\eta + \boldsymbol{\Gamma}\epsilon + \zeta$$

dove **B** e **Γ** sono le matrici dei coefficienti trutturali del modello, di dimensione rispettivamente (m × m) e (m × n) mentre ζ_1 = (ζ_1, ζ_2, ..., ζ_m) è il vettore (m × 1) degli errori stocastici.

Guardando ai coefficienti strutturali, ovvero le m × m componenti della matrice **B** e le m × n componenti della matrice **Γ** possiamo notare come il loro significato non differisce da quello assunto dai coefficienti di un modello di regressione lineare. La componente β_{ij} della matrice B dei coefficienti strutturali identificherà, infatti, l'effetto che, la variazione unitaria della j-esima componente di η, η_j ha sulla i-esima componente di η, η_i (con j = 1, 2, ..., m, e i = 1, 2, ..., n); medesimamente la componente γ_{ij} indicherà l'effetto della variazione unitaria della j-esima componente di ϵ, ϵ_j sulla i-esima componente di η, η_i.
Nel modello in questione, come ben rappresentato nella tabella della notazione LISREL, le variabili η e ξ sono variabili cosiddette **latenti** e, perciò, non sono direttamente osservate. Supponiamo invece che ad essere osservate siano le variabili x e y. Rspettivamente le loro osservazioni sono x = (x_1, x_2, ..., x_q) e y = (y_1, y_2, ..., y_p) tali che:

$$\begin{aligned} \text{y} &= \boldsymbol{\Lambda_y}\eta + \epsilon, \\ \text{x} &= \boldsymbol{\Lambda_x}\xi + \delta \end{aligned}$$

dove ϵ e δ indicano i vettori dell'errore stocastico, mentre $\boldsymbol{\Lambda_x}$ e $\boldsymbol{\Lambda_y}$ sono le matrici dei coefficienti strutturali dei modelli di misurazione. Le funzioni di regressione di y su η e di x su ξ sono appunto anche definite *modello di misurazione* per la variabile y e per la variabile x. Il modello LISREL può essere così riassunto tramite le tre equazioni sopracitate:

$$\begin{aligned} \text{Modello di equazioni strutturali} &: \eta = \beta\eta + \gamma\epsilon + \zeta \\ \text{Modello di misurazione per y} &: \text{y} = \boldsymbol{\Lambda_y}\eta + \epsilon \\ \text{Modello di misurazione per x} &: \text{x} = \boldsymbol{\Lambda_x}\xi + \delta \end{aligned}$$

(Jöreskog, 1971).

Il modello LISREL così formulato necessita però di una serie di restrizioni che elenchiamo di seguito:

1. Le variabili sono misurate in termini dello scarto dalle loro medie, ovvero:

$$\begin{aligned} \text{E}(\eta) = \text{E}(\xi) &= 0 \\ \text{E}\ (\zeta) &= 0 \\ \text{E}(\text{Y}) = \text{E}(\epsilon) &= 0 \\ \text{E}(\text{X}) = \text{E}(\delta) &= 0 \end{aligned}$$

2. Le variabili indipendenti e gli errori sono fra loro incorrelati nella stessa equazione:

$$\begin{aligned} \mathrm{E}(\xi\ \zeta') &= 0 \\ \mathrm{E}(\eta\ \epsilon') &= 0 \\ \mathrm{E}(\xi\ \delta') &= 0 \end{aligned}$$

mentre tra equazioni diverse del modello:

$$\begin{aligned} \mathrm{E}(\eta\ \delta') &= 0 \\ \mathrm{E}(\xi\ \epsilon') &= 0 \end{aligned}$$

3. Gli errori delle diverse equazioni sono incorrelati l'un l'altro:

$$\begin{aligned} \mathrm{E}(\zeta\ \epsilon') &= 0 \\ \mathrm{E}(\zeta\ \delta') &= 0 \\ \mathrm{E}(\epsilon\ \delta') &= 0 \end{aligned}$$

4. La matrice $\boldsymbol{I}$ **-** $\boldsymbol{B}$ dev'essere non singolare. Essendo $\mathbf{B}$ una matrice quadrata $m \times m$ la condizione di non singolarità è soddisfatta se il det$(\mathbf{I} - \mathbf{B}) \neq 0$. Questa condizione è rilevante in quanto garantisce che il modello sia costituito da equazioni *non* ridondanti, ovvero che ciascuna equazione del modello sia indipendente e che perciò non possa essere espressa come una combinazione delle altre equazioni del modello.

(Corbetta, 2002).

La corretta specificazione del modello teorico richiede inoltre la derivazione di altre quattro matrici di *covarianza*, una delle quali ($\mathbf{\Phi}$) tra le variabili esogene e le altre tre ($\mathbf{\Psi}$, $\mathbf{\Theta_\delta}$ e $\mathbf{\Theta_\epsilon}$) tra gli errori stocastici del modello. Un modello di equazioni strutturali è quindi composto da un totale di otto matrici: quattro di *coefficienti strutturali* $\mathbf{\Gamma}$, $\mathbf{B}$, $\mathbf{\Lambda_y}$ e $\mathbf{\Lambda_x}$ e quattro di covarianze $\mathbf{\Phi}$, $\mathbf{\Theta_\delta}$, $\mathbf{\Theta_\epsilon}$ e $\mathbf{\Psi}$ (Jöreskog, 1971).

Tramite le 8 matrici del modello è possibile derivare un'ulteriore importante matrice, ovvero la matrice Σ, che contiene le covarianze tra le variabili X, le covarianze tra le variabili Y ed infine quelle tra le variabili X e Y. La matrice Σ assume un importante rilievo ai fini della nostra analisi in quanto è essenziale per

definire la bontà di adattamento di un modello di equazioni strutturali rispetto ai dati. Qualunque sia infatti il metodo di stima dei parametri strutturali (2SLS (minimi quadrati a due fasi), GLS (minimi quadrati generalizzati), ...) l'obiettivo sarà comunque di stimare i parametri al fine di minimizzare la discrepanza tra la matrice *teorica* delle covarianze Σ e la matrice delle covarianze *osservate* (Corbetta, 2002; Jöreskog, 1971).

2.2.3 Ricerca econometrica e SEM

Economie sommerse e *corruzione* (Buehn e Schneider, 2009)

Il primo esempio di modello di equazioni strutturali che prendiamo in considerazione riguarda l'analisi della relazione tra **economia sommersa** e **corruzione**. Questo studio (Buehn e Schneider, 2009) è stato portato avanti costruendo un modello di equazioni strutturali che utilizza una serie di cause ed indicatori per spiegare i fenomeni della corruzione e dell'economia sommersa. Questo approccio è necessario dal momento che entrambe le variabili, essendo *latenti*, non sono osservabili direttamente. Per comprendere meglio gli obiettivi dello studio è ne-

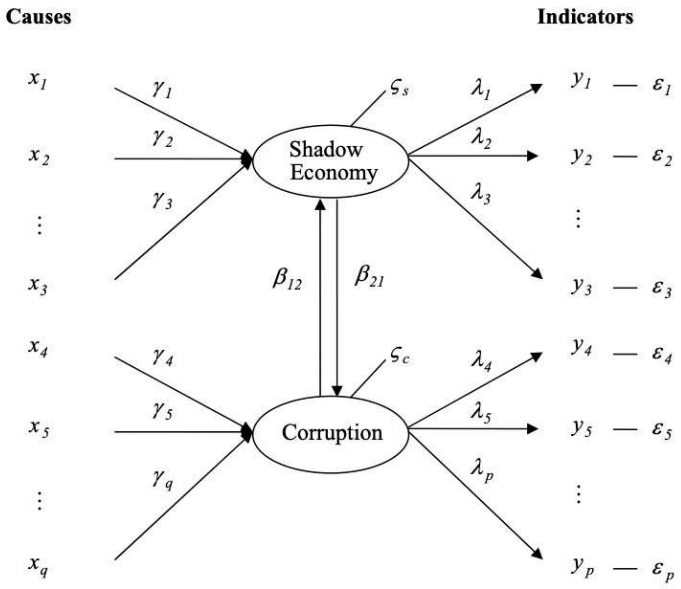


Figura 2.4: Il modello di equazioni strutturali tra corruzione ed economia sommersa.
Buenhn, Schneider (2009)

cessario definire il concetto di *economia sommersa* e *corruzione*. Per economia sommersa s'intende, secondo la definizione di Smith (1994) qualunque attività di produzione di beni o servizi, non necessariamente di stampo illegale, che sfugge

alla stima del PIL. Il concetto di "corruzione" è invece definito dalla Banca Mondiale come "l'impiego distorto delle leggi che mina le fondamenta di una nazione e in più colpisce i più poveri, già largamente svantaggiati". Ciò premesso, lo studio volge l'attenzione alla costruzione di un modello di equazioni strutturali che, per l'impianto teorico descritto nei paragrafi precedenti, è definito da una componente strutturale ed un modello di misurazione. Per fare questo guardiamo innanzitutto alle relazioni tra variabili e introduciamo le cause (x_1, x_2, ..., x_q) e gli indicatori del modello (y_1, y_2, ..., y_p) (**figura 2.4**) .

Il modello viene definito in forma matriciale come segue:

$$(1)\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}=\begin{bmatrix}0 & \beta_{12}\\ \beta_{21} & 0\end{bmatrix}\cdot\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}+\begin{bmatrix}\gamma_1 & \gamma_2 & \gamma_3 & 0 & 0 & ... & 0\\ 0 & 0 & 0 & \gamma_3 & \gamma_5 & ... & \gamma_q\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ ...\\ x_q\end{bmatrix}+\begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix},$$

$$(2)\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ ...\\ y_q\end{bmatrix}=\begin{bmatrix}1 & 0\\ \lambda_2 & 0\\ \lambda_3 & 0\\ 0 & 1\\ 0 & \lambda_5\\ ... & ...\\ 0 & \lambda_p\end{bmatrix}\cdot\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}+\begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\\ \epsilon_5\\ ...\\ \epsilon_q\end{bmatrix},$$

dove, la (1) indica la componente strutturale mentre la (2) il modello di misurazione. η_1 e η_2 sono le variabili latenti ovvero rispettivamente "corruzione" e "economia sommersa" mentre i coefficienti β indicano l'effetto della variazione di una delle variabili latenti sull'altra.

Poichè, come abbiamo anticipato, le variabili η_1 e η_2 non sono direttamente osservabili il modello è stato costruito utilizzando una serie di *cause* ed *indicatori* che specifichiamo di seguito.

Cause e indicatori dell'economia sommersa

Le *cause*:

1. **la pressione fiscale**. La pressione fiscale influenza molto il mercato del lavoro: una maggiore pressione fiscale genera maggiore costo del lavoro. Questo può rendere maggiormente conveniente lavorare per l'"economia sommersa" che, non essendo sottoposta a tassazione, offre salari superiori.

2. **l'intensità della regolazione**. Le politiche di regolazione del mercato del lavoro (che si manifestano ad esempio con l'introduzione del salario minimo, limitazioni ai licenziamenti, ecc...) provocano un aumento dei costi del lavoro favorendo una maggiore offerta di lavoro verso l'economia sommersa.

3. **La disoccupazione**. La disoccupazione ha un effetto ambiguo sull'economia sommersa. Da un lato infatti offrendo beni e servizi a basso prezzo, maggiormente accessibili ai soggetti disoccupati, l'economia sommersa potrebbe essere favorita dall'aumento della disoccupazione per effetto dell'aumento di domanda. D'altro canto però i nuovi disoccupati potrebbero vedere le loro entrate diminuire a tal punto da non poter accedere neppure ai beni offerti a minor prezzo dall'economia sommersa. L'effetto finale è quindi definito dalla prevalenza dell'uno sull'altro.

Gli *indicatori*:

1. **L'impiego di moneta**. Un rilevante indicatore circa l'ampiezza e lo sviluppo delle economie sommerse è dato dall'impiego degli aggregati monetari M_0 (base monetaria) e M_1 (moneta più depositi): come è facile immaginare infatti nelle economie sommerse le transazioni sono prevalentemente sostenute tramite moneta M_0 o M_1.

2. **La crescita del PIL**. Lo sviluppo dell'economia sommersa può avere effetti ambigui sul tasso di crescita del PIL di un'economia; da un lato infatti le sostenute attività dell'economia sommersa potrebbero comunque trainare l'economia "ufficiale" promuovendo la crescita. Dall'altro però poichè le attività sommerse non sono tipicamente soggette a tassazione, questo potrebbe generare la necessità di un aumento delle imposte nell'economia emersa generando un effetto di rallentamento. Stante questo ambiguo effetto totale, guardando alle analisi di (Loayza, 1996) è possibile os-

servare una relazione negativa tra sviluppo dell'economia sommersa e tasso di crescita del PIL.

3. **Il tasso di partecipazione**. Guardare alla variazione del tasso di partecipazione può essere utile in questo modello in quanto può indicare lo spostament o della forza lavoro dall'economia "ufficiale" all'economia sommersa e viceversa.

Cause e indicatori della corruzione
Le *cause*:

1. **Il sistema politico**. Un sistema politico e giudiziario poco efficiente creano condizioni che limitano il rispetto delle norme e perciò, favoriscono la corruzione.

2. **Il fattore culturale**. Si rileva come i soggetti scarsamente scolarizzati considerano la corruzione come un fattore intrinseco della società in cui vivono (Pasuk e Sungsidh, 1994). In questo modello assumiamo perciò, per semplicità, che minore scolarizzazione *primaria* implichi una maggior corruzione.

3. **Il fattore economico**. Nelle economie in cui il governo inteferisce con politiche regolatorie particolarmente stringenti, gli attori economici potrebbero favorire la corruzione con l'obiettivo di ritrovare la "libertà economica", eludendo le norme ufficiali.

Gli *indicatori*:

1. **La crescita economica**. Sono numerosi gli studi (tra cui ad esempio Mauro (1995)) che collegano la corruzione allo scarso sviluppo economico. Assumeremo perciò che un basso livello del PIL possa indicare la presenza del fenomeno della corruzione.

Una volta definite la cause e gli indicatori delle variabili del modello, i ricercatori hanno proceduto alla specificazione del modello. Le stime dei parametri strutturali con i relativi errori standard (riportate integralmente nell'appendice) sono frutto dei cinque test di specificazione che sono stati effettuati. La maggior parte dei coefficienti strutturali del modello sono risultati *statisticamente significativi* e hanno mostrato un segno concorde alle ipotesi sulle relazioni causali definite in origine sulla base della teoria economica. (1) Per quanto concerne le *cause e indicatori dell'economia sommersa* i coefficienti associati alle variabili "intensità della regolazione" e "disoccupazione" mostrano una chiara rilevanza nello sviluppo delle economie sommerse; allo stesso modo ha mostrato conferma l'ipotesi che lega l'aumento della pressione fiscale allo sviluppo del fenonemo delle economie

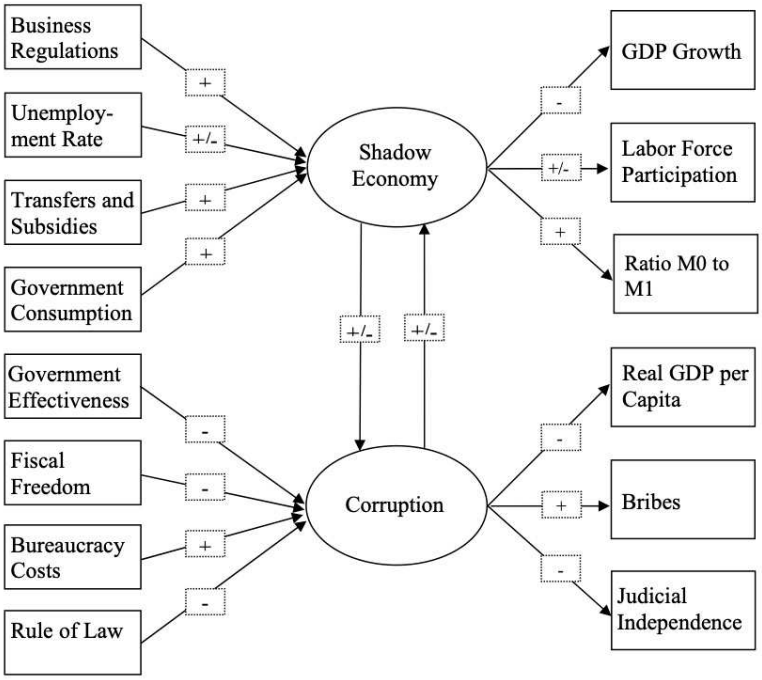


Figura 2.5: Path diagram delle relazioni tra variabili latenti, cause e indicatori. Buenhn, Schneider (2009)

sommerse. Gli aggregati monetari M_0 e M_1, la diminuzione del tasso di crescita del PIL e il tasso di partecipazione mostrano tutti una relazione negativa con lo sviluppo dell'economia sommersa come ipotizzato in precedenza.
(2) Riguardo le *cause e gli indicatori della corruzione* la specificazione del modello conferma la relazione positiva tra le politiche di intervento del governo nell'attività economica e la presenza di fenomeni di corruzione. Non trova giustificazione invece la relazione ipotizzata teoricamente tra *bassa scolarizzazione* e *corruzione* (Pasuk e Sungsidh, 1994). Allo stesso modo anche il coefficiente legato al sistema giudiziario non è statisticamente significativo. È invece statisticamente significativo il coefficiente del sistema *politico*: un efficiente sistema politico è perciò negativamente correlato con il fenomeno della corruzione.

Anche i coeffcienti β_{12} e β_{21} risultano statisticamente significativi: ciò giustifica la relazione causale, alla base dell'intero modello, tra l'esistenza delle *economie sommerse* e la *corruzione*. In ognuna delle (5) specificazioni del modello il coefficiente β_{21} ha sempre assunto valore maggiore del coefficente β_{12}. Questo singifica che l'effetto dell'economia sommersa sulla corruzione risulta tendenzialmente *maggiore* dell'effetto della corruzione sull'economia sommersa (Buehn e Schneider, 2009).

Il modello I di Klein (1950)

Un' altra interessante applicazione empirica dei modelli SEM in ambito economico, è fornito dall'analisi macroeconomica portata avanti da Klein (1950). Il modello in questione tiene conto delle seguenti variabili *esogene* ed *endogene*:

ENDOGENE	ESOGENE
C_t, consumo al tempo t	K_t, capitale al tempo t
I_t, investimento al tempo t	G_t, spesa pubblica (no salari) al tempo t
W_{pt}, salario *privato* al tempo t	T_t tassazione su attività ed esportazioni
X_t, domanda di equilibrio al tempo t	W_{gt} salario minimo
P_t, profitti priviati al tempo t	A_t, trend temporale dal 1931

Klein costruisce il modello di equazioni simultanee seguendo la struttura che segue:

1. $C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 \ (W_{pt} + W_{gt}) + \epsilon_{1t}$ (consumo),
2. $I_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 \ (K_{t-1} + \epsilon_{2t}$ (investimento),
3. $W_{tp} = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 \ (A_t + \epsilon_{1t}$ (salari privati),
4. $X_t = C_t + I_t + G_t$ (domanda di equilibrio),
5. $P_t = X_t$ - T_t - W_{pt} (profitti privati),
6. $K_t = K_{t-1} + I_t$ (stock di capitale).

Nelle sei equazioni del modello è possibile identificare una serie di variabili *esogene* quali: (1) la spesa pubblica non destinata al pagamento dei salari G_t, (2) la tassazione indiretta più le esportazioni nette T_t, (3) il salario minimo W_{gt}, (4) il trend misurato dal 1931 A_t e (5) il termine costante.
Il modello è composto da tre *equazioni di comportamento* (1, 2, 3) che spiegano come ciascuna delle variabili indipendenti determini le variabile dipendente posta nel membro di sinistra, una condizione di equilibrio macroeconomico della domanda (4) e infine due equazioni (5, 6) che impongono condizioni di *identità economica*, come ad esempio (6) che lo stock di capitale al tempo t dev'essere dato dallo stock di capitale al tempo (t-1) + l'investimento al tempo t (Greene, 2008).
Come possiamo notare l'assunto di base del modello è il raggiungimento della condizione di equilibrio macroeconomico attraverso due condizioni: la massimizzazione dell'utilità e dei profitti delle imprese e la prospettiva di massimizzazione

dell'utilità dei privati. Nel modello di Kleine queste due condizioni giustifiche-rebbero, tramite le interazioni di questi due gruppi di attori, le determinazioni dei prezzi e del livello salariale (Klein, 1950).

A differenza dei modelli che abbiamo considerato finora, l'analisi di Klein non comprende variabili latenti. Nonostante questo però, lo studio in esame apre ad ulteriori considerazioni che saranno oggetto di analisi nel prossimo capitolo. Come possiamo notare alcune delle variabili incluse nel modello sono variabili ritardate. La domanda "X", I profitti "P" e il capitale "K" sono infatti presenti nel modello sia in forma *ritardata* (con pedice t-1) che in forma non ritardata (con pedice t). Questo tipo di variabili sono oggetto di analisi in quelle che vengono definite *serie storiche*.

Capitolo 3

Modelli per Serie Storiche

In ambito statistico, con il termine "*serie*" ci si riferisce ad un insieme di dati che risultano ordinati secondo un criterio *qualitativo*. Una **Serie *Storica*** è una successione di dati numerici nella quale il criterio qualitativo di ordinamento è quello *temporale*. I dati numerici a cui ci si riferisce possono essere frutto di osservazioni di una o più variabili contemporaneamente: possiamo in questo senso distinguere tra serie storiche *univariate* e *multivariate*: la prima categoria fa riferimento a serie che associano a ciascun istante temporale una sola variabile osservata, differentemente si parlerà di serie storiche *multivariate* o anche *multiple* se a ciascun instante è associata la realizzazione di due o più variabili (Di Fonzo e Lisi, 2000).

Una serie storica $\{y_t$, con t = 1, 2, ..., $n\}$ può essere definita allora come una successione ordinata di numeri reali che misura un certo fenomeno Y, analizzato in funzione della sua evoluzione temporale (e quindi rispetto alla variabile t). Dal punto di vista grafico una tale successione può essere rappresentata per mezzo di diagrammi cartesiani che mostrano, sull'asse x, i vari istanti temporali nei quali si analizza il fenomeno, mentre, sull'asse delle ordinate, le singole osservazioni della

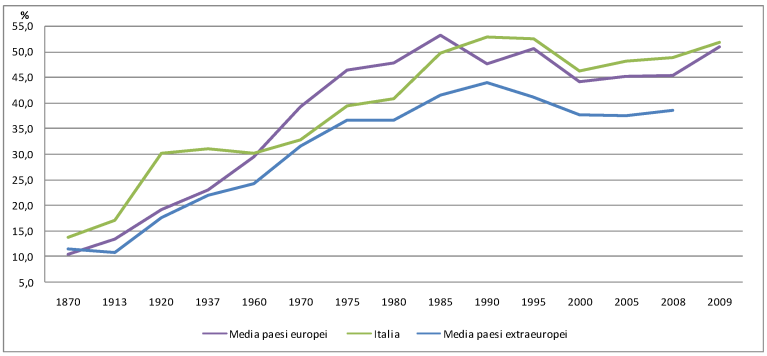


Figura 3.1: La serie storica della variabile "spesa pubblica in percentuale al PIL" Ministero dell'Economia e delle finanze, Ragioneria Generale dello Stato (2009)

variabile Y (Piccolo, 1990).
Rappresentare graficamente le osservazioni di una serie storica costituisce il primo passo per portare avanti un'analisi predittiva: questo permette infatti di poter visualizzare alcuni aspetti importanti come ad esempio la sua tendenza di fondo (detta *trend*), i valori estremi, la presenza di una componente stagionale e/o eventuali shock imporovvisi che posso essere ampiamente informativi sul comportamento futuro della serie (Chatfield, 1995). Una particolarità delle serie storiche è il rapporto di dipendenza che è possibile individuare tra le sue osservazioni. Tramite l'impiego di modelli statistici idonei (alcuni dei quali verrano definiti nei paragrafi succesivi), le osservazioni passate di una serie storica possono essere utilizzate per ottenere previsoni più o meno accurate sulla sua evoluzione futura. In questo senso è bene operare la distinzione fondamentale tra serie storiche *deterministiche* e serie storiche *stocastiche*. Una serie storica $\{y_t\}_{t=1}^n$ si dice *deterministica* se esiste una funzione

$$\varphi_t = \varphi(t, y_{t-1}, y_{t-2}, ...)$$

tale che:

$$E(y_t - \varphi_t)^2 = 0, \text{ (con t} \in \tau \text{ (spazio campionario))}.$$

In altre parole una serie storica si dice *deterministica* se è prevedibile esattamente a partire dalle sue osservazioni precedenti (Di Fonzo e Lisi, 2000).
Diremo invece che una serie storica è *stocastica* se non è possibile prevederla esattamente a partire dalle osservazioni precedenti, a meno di un errore che chiameremo u_t. A questo proposito definiamo come segue il modello stocastico che *descrive il processo generatore dei dati* per una serie storica $\{y_t\}_{t=1}^n$:

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

dove f(t) indica la componente deterministica del modello mentre il termine d'errore u_t ne rappresenta la componente stocastica (Di Fonzo e Lisi, 2000).

Approccio classico vs approccio moderno

Una distinzione degna di nota per quanto riguarda le serie storiche concerne i diversi approcci rilevabili in letteratura circa il *trattamento del modello*. A questo proposito si possono individuare due diversi approcci nello studio delle serie storiche: un approccio *classico* ed uno *moderno* (vedi per esempio Di Fonzo

e Lisi (2000)). Per considerare le differenze sostanziali tra l'uno e l'altro punto di vista guardiamo al *processo generatore dei dati* che abbiamo già definito nel paragrafo precedente come:

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

(1) **L'approccio classico** all'analisi delle serie storiche considera la funzione $f(\mathrm{t})$ come la *legge di evoluzione temporale* del fenomeno. Il termine u_t viene invece inteso come la componente stocastica, ovvero l'insieme di tutti i fattori che non possono essere definiti esplicitamente attraverso la funzione $f(\mathrm{t})$. Per questo la componente stocastica del modello è, secondo l'approccio classico, definibile come un processo di *white noise* (WN) in cui le variabili casuali sono tutte incorrelate, con media zero e varianza costante.
(2) In maniera molto differente da quello classico, **l'approccio moderno** focalizza l'attenzione, non tanto sulla funzione $f(\mathrm{t})$, quanto piuttosto sulla componente u_t del modello. In questo senso la componente stocastica si presume essere determinata da una funzione g delle variabili ritardate di Y $(Y_{t-1}, Y_{t-2}, ...)$ e del processo *white noise* (ϵ_{t-1}, ϵ_{t-2}, ...) tale che:

$$u_t = g(Y_{t-1}, Y_{t-2}, ..., \epsilon_{t-1}, \epsilon_{t-2}, ...) + \epsilon_t, \tag{3.1}$$

$$\text{con } \epsilon_t \sim \mathrm{WN}(0,\sigma^2)$$

3.1 Le componenti del modello di serie storiche

L'approccio **tradizionale** all'analisi delle serie storiche rivolge particolare interesse alla scomposizione delle *variazioni* in quelle che vengono definite **componenti** della serie (Di Fonzo e Lisi, 2000). Anche quando si assume che una serie storica sia *stazionaria*, ovvero che la *legge di probabilità* che ne definisce il comportamento non cambi nel tempo (Cryer e Chan, 2008), è possibile identificare alcune sue variazioni rispetto alla *tendenza di fondo* che è utile analizzare separatamente.

Tradizionalmente tali variazioni possono essere scomposte in:

1. **Ciclo**
   Alcune serie storiche mostrano una tendenza ciclica: le osservazioni si succedono alternando fasi ascendenti e discendenti. Un esempio può essere la

variazione della temperatura durante il giorno oppure, in ambito economico, le fluttuazioni collegate alle fasi di espansione e contrazione dell'intero sistema economico.

2. **Trend**
   Per *trend* si intende la variazione del valor medio della serie storica, visibile nel lungo-termine. In questo senso è importante considerare come, in alcuni fenomeni, l'ampiezza temporale delle oscillazione renda difficile distinguere una variazione ciclica da un trend. Nei cambiamenti climatici ad esempio, in cui un ciclo di oscillazione può durare oltre 50 anni, considerare un orizzonte temporale inferiore potrebbe nascondere la natura ciclica di tali variazioni (Chatfield, 1995).

3. **Stagionalità**

   Molte serie storiche presentano, nel corso dell'anno, variazioni che sono dovute a fattori sociali, economici o climatici. Una serie storica che descrive ad esempio la produttività delle imprese italiane nel tempo potrebbe subire uno scostamento dalla sua tendenza di fondo in corrispondenza dei mesi estivi, meno produttivi.

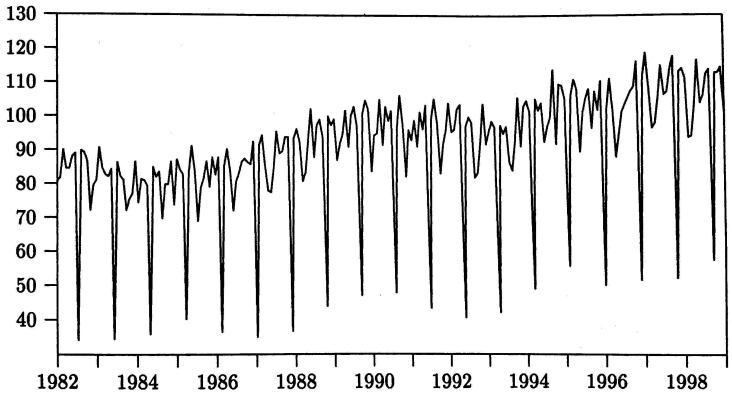


Figura 3.2: Indice generale della produzione industriale.

Le componenti che abbiamo appena elencato necessitano ora di essere analizzate con maggiore dettaglio guardando soprattutto ai più frequenti modelli di decomposizione delle serie storiche. In generale è possibile identificare due modalità di decomposizione delle serie storiche: il primo, secondo quello che viene chiamato *modello additivo*, e il secondo che viene invece definito *modello moltiplicativo*. Secondo l'approccio moltiplicativo, una serie storica Y_t può essere scomposta come *prodotto* delle sue componenti, ovvero:

$$Y_t = T_t \times C_t \times S_t \times \epsilon_t \tag{3.2}$$

Secondo l'approccio additivo, invece, una serie storica Y_t può essere scomposta come *addizione* delle sue componenti:

$$Y_t = T_t + C_t + S_t + \epsilon_t \tag{3.3}$$

In alcuni casi, in base alla tipologa dei dati oggetto di analisi, il modello additivo può essere derivato, a partire del modello moltiplicativo, tramite una trasformazione logaritmica e ne rappresenta perciò un caso particolare. Infatti:

$$\log Y_t = \log T_t + \log C_t + \log S_t + \log \epsilon_t \tag{3.4}$$

Oltre questi due modelli principali rileviamo anche la possibilità di seguire un approccio *misto* alla scomposizione di una serie storica. come ad esempio:

$$Y_t = T_t \times C_t + S_t + \epsilon_t, \tag{3.5}$$

dove la serie storica Y_t è determinata sia dalla somma che dal prodotto delle sua componenti (Piccolo, 1990).

In ognuno dei tre approcci, abbiamo, rispetto alla t-esima componente della serie storica Y_y:

1. il t-esimo elemento della componente tendenziale T_t
2. il t-esimo elemento della componente ciclica C_t
3. il t-esimo elemento della componente stagionale S_t
4. il t-esimo elemento dell'errore ϵ_t

Una volta riconosciute le componenti di una serie storica si può passare alla fase di analisi delle stesse. In particolare l'obiettivo di questo processo è di "filtrare" la serie storica per eliminarne l'una o l'altra componente. Questo può essere eseguito per motivi diversi: si può ad esempio voler depurare la serie da ogni componente, oppure si può voler filtrare solamente alcune di esse per evidenziarne meglio altre. Dal punto di vista metodologico l'eliminazione di una o più componenti da una serie storica è possibile tramite diversi approcci, che possono essere però più o meno adatti in base ai dati a disposizione. Per quanto concerne la componente

stagionale, è ad esempio possibile isolarla (e quindi **destagionalizzare** una serie) tramite tre approcci fondamentali: un primo metodo è quello basato sulle *medie mobili*, un altro si basa sui *modelli regressivi* e infine l'ultimo sui *modelli stocastici*.

3.2 Processi Stocastici e Modelli di Serie Storiche

Un *processo stocastico* può essere definito come una sequenza di v.c. Y_t (con t = 0, ±1, ±2, ±3, ...) che costituisce il modello di riferimento per le osservazioni di una serie storica (Cryer e Chan, 2008). Prenderemo in considerazione di seguito alcuni dei processi stocastici più importanti per la costruzione dei modelli di serie storiche: a partire dai *processi stocastici* autoregressivi AR(p), a media mobile MA(q) e ARMA(p, q) è infatti possibile costruire alcuni dei modelli di serie storiche più utilizzati in ambito di ricerca. L'assunto fondamentale alla base di questo tipo di modelli è la *correlazione* non nulla tra la realizzazione al tempo t di una serie storica e le sue osservazioni precedenti. In questo senso i modelli che analizzaremo di seguito forniscono utili strumenti per poter *prevedere*, a partire dalle sue osservazioni precedenti, le evoluzioni future di una serie storica.

3.2.1 il modello AR(p)

I modelli *Autoregressivi* si basano sull'assunto per cui i *valore al tempo t* di una serie storica può essere determinato in funzione dei suoi p valori precedenti x_{t-1}, x_{t-2}, ..., x_{t-p}. Il valore attuale x_t della serie potrà allora essere stimato tramite il modello:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + u_t,$$

dove ϕ_1, ϕ_2, ..., ϕ_p sono i coefficeinti del modello e $u_t \sim \text{WN}(0,\sigma^2)$ è l'errore stocastico. Il modello appena descritto prende il nome di *Modello Autoregressivo di ordine p*, anche definito secondo l'acronimo inglese **AR(p)** (Shumway e Stoffer, 2011).

Per costruire un modello autoregressivo è necessario definire l'ordine p. A questo merito, costruire un modello AR con ordine particolarmente basso, permettrebbe di compiere solo pochi "passi indietro" nella serie storica: per l'obiettivo predittivo che abbiamo definito in precedenza questo potrebbe non esssere sufficientemente informativo. Allo stesso modo però un modello di ordine p troppo elevato restituisce una funzione di regressione con molti regressori (pari al numero p) che può essere invece difficile da maneggiare.

3.2.2 il modello MA(q)

Un altro importante esempio di modello per serie storiche è quello a **medie mobili**. Di Fonzo definisce un processo stocastico a Media Mobile di seguito definito come:

$$Y_t = \epsilon_t - \theta_1\epsilon_{t-1} - ... - \theta_q\epsilon_{t-q}, \tag{3.6}$$

dove θ_j (con j = 1, 2, ..., q) sono i coeffcienti del modello mentre ϵ_t, ϵ_{t-1}, ...,ϵ_{t-q} sono osservazioni di un processo White Noise. In altre parole quindi "*la variabile Y_t può essere considerata il risultato di una somma pesata di impulsi casuali presenti (ϵ_t) e passati (ϵ_{t-1}, ..., ϵ_{t-q})* (Di Fonzo e Lisi, 2000).

Un aspetto essenziale di questo modello è che il termine ϵ_t dev'essere un processo *White Noise*. I processi stocastici White Noise si caratterizzano per essere *puramente casuali*. Questi sono composti da una serie di osservazioni con media nulla e varianza costante che sono tra loro incorrelate.

3.2.3 Il modello ARMA

Dalla combinazione dei modelli a media mobile MA(q) e dei modelli autoregressivi AR(p) sono scaturiti invece i cosiddetti *modelli autoregressvi a media mobile ARMA*. Il modello ARMA (p, q) può essere definito come segue:

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t - \theta_1\epsilon_{t-1} - ... - \theta_q\epsilon_{t-q} \\ &= \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \end{aligned} \tag{3.7}$$

dove $\epsilon_t \sim$ WN(0, σ^2).

Come notiamo dall'equazione (3.3) il modello ARMA è proprio la combinazione dei due modelli, quello autoregressivo e quello a media mobile, analizzati in precedenza: notiamo infatti che la prima componente, denotata dai parametri ϕ, deriva dai modelli autoregressivi AR(p) mentre la seconda componente con paramteri θ è data dal modello a media mobili MA(q). Sia per quanto riguarda il parametro p del modello AR(p) che per il parametro q del modello AM(q) esistono una serie di criteri (come ad esempio il criterio di AIC o la funzione

di autocorrelazione parziale PACF) che ne permettono la determinazione e che possono essere utilizzati anche nel modello ARMA(p, q).

3.3 Serie Storiche Finanziarie

In ambito econometrico, i modelli di serie storiche sono largamente utililizzati soprattutto con l'obiettivo di studiare la *volatilità* dei rendimenti associati agli strumenti finanziari. Il celebre modello di Black-Scholes, utilizzato per la determinazione del prezzo c_t di un'opzione call europea, ne è l'esempio. Questo modello definisce infatti il prezzo di un opzione call europea come:

$$\begin{aligned} c_t = P_t\phi(x) - Kr^{-l}\phi(x - \sigma_t\sqrt{l}), con \\ x = \frac{\log(P_t/Kr^{-l})}{\sigma_t\sqrt{l}} + \frac{1}{2}\sigma_t\sqrt{l} \end{aligned} \tag{3.8}$$

dove ϕ(x) è la distribuzione normale standradizzata calcolata in x, l è il periodo che intercorre tra l'acquisto e la scadenza dell'opzione e infine r è il tasso di interesse *risk-free*.
Al di là delle considerazioni sul modello nella sua componente strettamente finanziaria, quello che interessa in questa sede è la sua componente σ_t. La deviazione standard in questione rappresenta proprio la *volatilità* dello strumento finanziario sottostante all'opzione call in esame. Un altro esempio di analisi della volatilità è l'indice VIX: questo indicatore, introdotto dalla Chicago Board Option Exchange (CBOE) nel 2004 fornisce infatti informazioni oggiornate circa la volatilità attesa nei mercati finanziari (Tsay, 2005).

L'interesse nei confronti dell'analisi della volatilità giustifica il sempre più frequente utilizzo degli strumenti della statistica inferenziale in ambito finanziario. I modelli ARCH (AutoRegressive Conditional Heteroschedastic) e GARCH (Generalised AutoRegressive Conditional Heteroschedastic) sono ad esempio alcuni dei modelli di serie storiche che vengono utilizzati in ambito finanziaro con l'obiettivo di analizzare la volatilità dei rendimenti, guardando alla loro *varianza condizionale* (Cryer e Chan, 2008). Come abbiamo visto nei paragrafi precedenti, i modelli ARMA assumono infatti che la serie storica oggetto di studio abbia varianza costante. L'analisi statistica in ambito finanziario si pone invece l'obiettivo di analizzare la varianza di una serie storica proprio in quanto indicatore della

volatilità di un certo strumento finanziario. Da qui deriva la necessità di costruire nuovi modelli in cui la varianza *non* sia presunta costante: in questo senso è risultato essenziale l'apporto di Engle (1982) grazie alla sua prima formulazione dei suddetti modelli ARCH e GARCH (Shumway e Stoffer, 2011).

3.3.1 il modello ARCH(1)

Il modello ARCH fu introdotto da Engle (1982) con l'intendo di analizzare uno dei fenomeni tipici delle serie storiche di tipo finanziario ovvero l'alternarsi di periodi caratterizzati da bassa volatilità ad altri caratterizzati da alta volatilità nei rendimenti. Questo suggerisce che la varianza condizionata ai rendimenti passati non sia costante. In questo modello ci riferiremo alla varianza condizionale (detta anche *volatilità* condizionale) di un rendimento r_t come $\sigma^2_{t|t-1}$, dove il pedice (t-1) indica appunto che la varianza è condizionata al rendimento al tempo (t-1). Ora, se r_t è noto, il suo quadrato, r_t^2, è uno stimatore corretto della varianza condizonale $\sigma^2_{t|t-1}$. In questo senso allora una serie di ritorni quadratici piccoli potrebbero prevedere un perido poco volatile, mentre, al contrario, una serie di ritorni quadratici più grandi potrebbero prevedere un periodo di maggiore volatilità.

In generale il modello ARCH(1) assume che la serie storica dei rendimenti $\{r_t\}$ sia data da:

$$r_t = \sigma^2_{t|t-1}\epsilon_t \tag{3.9}$$

$$\sigma^2_{t|t-1} = \omega + \alpha r^2_{t\,t-1} \tag{3.10}$$

dove α e ω sono i parametri ignoti del modello e $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$. Distribuendosi secondo un processo white noise, ϵ_t ha varianza unitaria e perciò (dall'equazione 3.9) $r_t = \sigma^2_{t|t-1}$. Per questo possiamo dire che:

$$\begin{aligned} E(r_t^2|r_{t-j}, j = 1, 2, ...) &= E(\sigma^2_{t|t-1}\epsilon_t^2|r_{t-j}, j = 1, 2, ...) \\ &= \sigma^2_{t|t-1}E(\epsilon_t^2|r_{t-j}, j = 1, 2, ...) \\ &= \sigma^2_{t|t-1}E(\epsilon_t^2) \\ &= \sigma^2_{t|t-1} \end{aligned} \tag{3.11}$$

La seconda equazione è data dal fatto che $\sigma_{t|t-1}$ è nota a partire dai rendimenti passati. La terza equazione è invece giustificata dal fatto che ϵ_t è indipendente dai rendimenti passati. La quarta equazione è invece data dal fatto che ϵ_t ha

varianza unitaria poichè si distribuisce secondo un processo white noise (Cryer e Chan, 2008).

3.3.2 Il modello ARCH(q) e GARCH(p, q)

La principale problematica legata al modello ARCH(1) è quella di prevedere le varianze condizionali future sulla base delle più recenti osservazioni quadratiche del rendimento. In questo senso è ragionevole immaginare che il modello in questione possa essere migliorato includendo tutte le osservazioni quadratiche dei rendimenti, attribuendo pesi minori alle varianze più lontane nel tempo. Possiamo quindi introdurre a questo merito il modello ARCH(q), che può essere inteso come la generalizzazione del modello ARCH(1), per cui:

$$\sigma^2_{t|t-1} = \omega + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + ... + \alpha_q r^2_{t-q} \tag{3.12}$$

dove il termine q indica proprio l'ordine del modello ARCH.
Un altro approccio al problema è stato sviluppato da Bollerslev (1986) e Taylor (1986) con la formulazione del modello generalizzato GARCH(p, q). La differenza fondamentale tra il modello ARCH(q) e il modello GARCH(p, q) è che quest'ultimo include un numero p di ritardi della varianza condizionale e può essere perciò definito come:

$$\begin{aligned}\sigma^2_{t|t-1} = \omega + \beta_1 \sigma^2_{t|t-1} + ... + \beta_p \sigma^2_{t-p|t-p} + \alpha_1 r^2_{t-1} \\ + \alpha_2 r^2_{t-2} + ... + \alpha_q r^2_{t-q}\end{aligned} \tag{3.13}$$

(Cryer e Chan, 2008)

Appendice A

Stime dei parametri da Buehn e Schneider (2009)

Specification	**(1)**		**(2)**		**(3)**		**(4)**		**(5)**	
Latent Variables	SE	C	SE	C	SE	C	SE	C	SE	C
Causes										
Business Regulations	0.18**		0.13*		0.21**		0.18**		0.18**	
	(2.00)		(1.84)		(2.18)		(2.02)		(1.98)	
Unemployment	0.19**				0.16*		0.17*		0.20**	
	(1.98)				(1.78)		(1.93)		(2.02)	
Transfers and Subsidies	0.09		0.05		0.11		0.09		0.09	
	(1.16)		(1.09)		(1.35)		(1.22)		(1.15)	
Government Consumption	0.16**		0.11*				0.15*		0.17**	
	(1.98)		(1.76)				(1.91)		(2.05)	
Labor Market Regulations			0.22**							
			(2.05)							
Size of Government					0.14*					
					(1.66)					
Government Effectiveness		-0.22***		-0.15**		-0.20***		-0.23***		-0.21***
		(3.13)		(2.25)		(2.66)		(3.36)		(3.01)
Fiscal Freedom		-0.15***		-0.09*		-0.15***		-0.14**		-0.17***
		(2.48)		(1.81)		(2.27)		(2.37)		(2.68)
Bureaucracy Costs		0.42***		0.34***		0.41***		0.40***		0.45***
		(5.15)		(2.95)		(4.29)		(4.79)		(5.52)
Rule of Law		-0.01		0.01		-0.01				-0.02
		(0.10)		(0.19)		(0.09)				(0.38)
School Enrollment								0.06		
								(1.01)		
Indicators										
GDP Growth	-0.51		-0.47		-0.46		-0.50		-0.51	
Labor Force Participation	-0.41***		-0.44***		-0.43***		-0.41***		-0.40***	
	(4.15)		(4.02)		(4.04)		(4.13)		(4.15)	
Ratio M0 to M1	0.31***		0.34***		0.35***		0.32***		0.30***	
	(3.33)		(3.33)		(3.52)		(3.36)		(3.32)	
Real GDP per Capita		-0.78		-0.75		-0.74		-0.78		-0.77
Bribes		0.15*		0.16**		0.16*		0.15**		0.14*
		(1.73)		(1.99)		(1.95)		(1.74)		(1.71)
Judicial Independence		-0.06		-0.08		-0.07		-0.06		
		(0.73)		(0.99)		(0.80)		(0.71)		
Freedom from Corruption										0.12
										(1.46)
Latent variables										
Shadow Economy → Corruption	0.68***		1.07***		0.81***		0.69***		0.67***	
	(4.23)		(4.34)		(3.98)		(4.19)		(4.23)	
Corruption → Shadow Economy	0.42***		0.43***		0.37***		0.47***		0.39***	
	(2.64)		(2.70)		(2.27)		(2.95)		(2.50)	

Absolute z-statistics appear in parenthesis. * = significance at 10% level, ** significance at 5 % level, *** = significance at 1% level.
Note: SE = shadow economy; C = corruption.

Figura A.1: Stima dei parametri del modello. Buehn e Schneider (2009)

Bibliografia

Anderson, T.W. e Rubin, H. (1956). Statistical Inference in Factor Analysis. *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*. 5, 111-150.

Bell, A., Fairbrother, M. e Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality and Quantity*. 53, 1051–1074.

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

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistist. Sci.*. 16, 199-215

Buehn, A. e Schneider, F. (2009). Estimating the Size of the Shadow Economy: Methods, Problems and Open Questions. *IZA Discussion Paper*. 9820, 1–30.

Chatfield, C. (1995). *Time series forecasting*. 1^a edizione. Londra: Chapman Hall.

Corbetta, P. (2002). *Metodi di analisi multivariata per le scienze sociali. I modelli di equazioni strutturali*. 1^a edizione. Bologna: Il Mulino.

Cryer, J. D. e Chan, K. (2008). *Time Series Analysis. With applications in R*. 2^a edizione. New York: Springer.

Lisi, F. e Di Fonzo, T. (2000). *Complementi di Statistica Economica: analisi delle Serie Storiche univariate*. 1^a edizione. Padova: Cleup editore.

Dragan, D. e Topolšek, D. (2014). Introduction to structural equation modeling: review, methodology and practical applications. *The 11th International Conference on Logistics Sustainable Transport 2014*, pages 1-27.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*. 5, 987–1007.

Ehrenberg, A. S. C. e Bound, J. A. (1993). Predictability and Prediction. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*. 156, 167-206.

Erzegovesi, L. (1999). Capire la volatilità con il modello binomiale. *Alea Tech Reports*. 4, 1-4.

Faraci, P. e Musso, P. (2013). *La valutazione dei modelli di equazioni strutturali: temi e prospettive*. 1^a edizione. Milano: LED. pp. 111-150.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*. 19, 1-67.

Greene, W. H. (2008). *Econometric Analysis*. 6^a edizione. Upper Saddle River: Pearson Prentice Hall.

Gujarati, D. N. (2003). *Basic econometrics*. 4^a edizione. New York: McGraw Hill.

Howe, W. G. (1955). Some contributions to factor analysis. Report No. ORNL-1919, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*. 36, 409–426.

Keesing, J. W. (1972). *Maximum likelihood approaches to causal flow analysis. Unpublished doctoral dissertation*. Chicago: Università di Chicago

Keil, M., Tan, B. C. Y., Wei, K., Saarinen, T., Tuunainen, V. e Wassenaar, A. (2000). A cross-cultural study on escalation of commitment behavior in software projects. *MIS Quarterly*. 24, 299-325.

Klein, L. R. (1950). *Economic Fluctuations in the United States. 1921-1941*. 1^a edizione. New York: John Wiley & Sons.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. 4^a edizione. New York: Guilford pubblications.

Kroese, D. P. e Chan, J. C. C. (2014). *Statistical Modeling and Computation*. 4^a edizione. New York: Springer.

Kuhn, M., Johnson, K. et al. (2013). *Applied predictive modeling*, volume 26. 1^a edizione. New York: Springer.

Lawley, D. N. (1958). Estimation in factor analysis under various initial assumptions. *British Journal of Statistical Psychology*. 11, 1–12.

Loayza, N. (1996). The economics of the informal sector: a simple model and some empirical evidence from latin america. *Carnegie-Rochester Conference Series on Public Policy*. 45, 129-162.

Mauro P. (1996). Corruption and growth. *The quarterly journal of economics*. 3, 681-712.

Pace, L. e Salvan, A. (2001). *Introduzione alla statistica: vol.2*. 1^a edizione. Padova: Cedam.

Pasuk, P. e Sungsidh, P. (1994). *Corruption and democracy in thailand.* 1^a edizione. Chiang Mai: Silkworm Books

Pearson, K. (1895). Correlation coefficient. *Royal Society Proceedings 1895*. 58, page 214.

Piccolo, D. (1990). *Introduzione all'analisi delle Serie Storiche*. 1^a edizione. Roma: Edizioni Carocci.

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R*. 6, 841-844.

Sarle, W. S. (1998). Prediction with missing inputs. *JCIS 98 Proceedings (P.Wang, ed.)*. 2, 399-402.

Schumacker, R. E. e Lomax, R. G. (2010). *A beginner's guide to structural equation modeling*. 3^a edizione. New York: Routledge.

Shmueli, G. e Koppius, O: (2009). The challege of prediction in information systems research. *Robert H. Smith School Research Paper*. 06-152, 43-58.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*. 25, 289-310.

Shumway, R. H. e Stoffer, D. S. (2000). *Time series analysis and its applications*. 3^a edizione. New York: Springer.

Stock, J. H. e Watson, M. (2009). *Introduzione all'econometria*. 3^a edizione. Milano: Pearson Italia.

Taylor, S. J. (1986). *Modelling financial time series*. 1^a edizione. Chichester: John Wiley & Sons.

Tsay, R. S. (2005). *Analysis of financial time series*. 1^a edizione. New York: John Wiley & Sons.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. *A. S. Goldberger & O. D. Duncan (Eds.), Structural equation models in the social sciences* (pp. 69-83) New York: Seminar Press.

Wright, S. (1918). On the nature of size factors. *The Annals of Mathematical Statistics*. 3, 367–374.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea–pigs. *Proceedings of the National Academy of Sciences*. 6, 320–332.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*. 5, 161–215.