Scoring models for country risk

Doretti, Marco

Università degli Studi di Firenze

14 February 2012

Università degli Studi di Firenze

FACOLTÀ DI ECONOMIA
Corso di Laurea Magistrale in Scienze Statistiche

Tesi in Econometria

Modelli di *scoring* per il rischio paese

Relatore:
Prof. Giorgio Calzolari

Candidato:
Marco Doretti

Correlatore:
Dott.ssa Benedetta Mazzolli

A.A. 2010/2011

Indice

Elenco delle tabelle

Elenco delle figure

Ringraziamenti

Nonostante si trovino all'inizio della tesi, queste pagine sono state scritte per ultime. Faccio notare ciò perchè mentre preparavo le altre parti ho provato più volte ad immaginare il momento della loro composizione, accompagnato dalla convinzione che esso sarebbe stato particolare. Non mi sbagliavo: quanto state leggendo è infatti la semplice copia di un foglio scritto a mano pochi minuti dopo le cinque di un sabato notte. Date queste premesse, sicuramente comprenderete l'abbandono dello stile impersonale utilizzato nella tesi in favore di un tono più colloquiale, che mi porterà a ringraziare "il babbo e la mamma" anzichè i genitori oppure a porre l'articolo determinativo davanti ai nomi di persona femminili, consuetudine tanto scorretta quanto carica di quel fascino che forse solo noi toscani percepiamo.

Il primo ringraziamento va all'Area *Risk Management* della Banca Monte dei Paschi di Siena, all'interno della quale è stato progettato e realizzato l'intero lavoro. Secondariamente voglio ringraziare tutti gli autori dei numerosi manuali LaTeX che un utente inesperto come il sottoscritto ha dovuto consultare anche durante la stesura di questi ringraziamenti (sì, è proprio così) e che meriterebbero un ruolo di primo piano nella bibliografia.

Grazie al babbo e alla mamma, alla Nonna Tina e a Luchino, sempre presenti e comprensivi, sopratutto quando, dopo dieci ore filate di algoritmi e programmazione, magari non ero proprio in splendida forma. Ringrazio inoltre tutti i componenti della mia splendida, numerosa e rumorosa famiglia, sparsi in giro per la Toscana, l'Italia, l'Europa, il mondo.

Grazie ai compagni di viaggio: ad Ale e Gabri, all'occorrenza maestri, allievi, consulenti. Il LA-PA-DO in tutti questi anni non ha fallito un colpo, tant'è che se potessi tornare indietro nel tempo li sostituirei soltanto con due

persone un po' meno fanatiche di Hattrick. Le mezz'ore passate ad aspettare che finissero di scannarsi sulla formazione o sulla gestione dello spogliatoio sono infatti l'unica cosa che posso rimproverare loro. A Pippo, mai puntuale ma sempre disposto a farsi in quattro per aiutarti e, di qualsiasi cosa tu ti occupi, sempre con il *paper* che fa al caso tuo! A Nicco: il recente terremoto non è riuscito a distrarmi quanto alcuni dei suoi comizi, ma solo lui, tra banane e OGM, sarebbe stato capace di portare una simile ventata di genio, sregolatezza e simpatia. Alla Chiara e alla Katia, uniche esponenti del gentil sesso, perchè nessuno sa come abbiano potuto sopportare questo branco di maschiacci. Insieme a loro voglio ringraziare tutti gli altri miei compagni della laurea triennale e magistrale, perchè al mitico "Beppe Parenti" io mi sono sempre sentito un po' a casa.

Grazie agli amici storici, in prima fila in questa come in altre mille occasioni: quelli che sono diventati espertissimi di rischio paese, declassamenti e triple A e quelli che «bada se ti movi a abbassare lo *spread*». È inutile fare distinzioni o elenchi: chi in cuor suo sente di essere compreso in questo gruppo ne fa automaticamente parte.

Grazie alla Giulia, perchè anche se abbiamo iniziato a condividere traguardi ed emozioni quando c'era ancora il cinema Adriano ed a malapena mettevamo insieme trent'anni in due, io non mi sono ancora stancato. Questa volta non c'è stato bisogno delle ripetizioni di letteratura italiana, ma il suo ruolo è stato, come al solito, fondamentale.

Infine voglio ringraziare te, chiunque tu sia. Se stai leggendo queste righe sicuramente avrò qualche buon motivo per farlo.

Introduzione

Negli ultimi anni i mercati finanziari hanno conosciuto un rapido sviluppo, che ha avuto come inevitabile conseguenza l'introduzione, da parte degli analisti che lavorano nel settore, di strumenti quantitativi sempre più sofisticati. In tale ottica si inserisce questo lavoro, il cui obiettivo è la stima, mediante adeguati modelli di *scoring*, del merito creditizio relativo ad alcune realtà non sempre valutate dalle agenzie di *rating*.

Il primo capitolo richiama i principali concetti, soffermandosi in particolare sulla differenza tra rischio paese e rischio sovrano, e inquadra l'operato delle agenzie illustrando gli indicatori e le metodologie da esse utilizzati. Nel secondo viene invece presentato il data-set di riferimento, elencando le variabili disponibili e svolgendo le preliminari analisi descrittive, mentre il terzo esamina alcuni aspetti della teoria dei dati mancanti. Gli elementi considerati sono quelli funzionali al duplice scopo di verificare l'ammissibilità del processo di determinazione del *rating* costruito e di occuparsi della fase di imputazione. Successivamente (capitolo 4) trova spazio la trattazione teorica dei modelli statistici impiegati. Il capitolo 5, infine, passa alla diretta applicazione sul campione unitamente all'interpretazione dei risultati e si chiude stabilendo le tematiche da approfondire o riconsiderare per eventuali sviluppi futuri.

Capitolo 1

Rischio paese e rischio sovrano

1.1 La differenza tra rischio paese e rischio sovrano

A partire dagli anni '70 la differenza tra rischio paese e rischio sovrano è stata fonte di numerose discussioni tra accademici e studiosi. Alcune correnti di pensiero tendevano infatti a scindere i due concetti, mentre altre erano convinte della loro sostanziale equivalenza. Questa confusione era imputabile alla mancanza di rigorose definizioni che spazzassero via i dubbi in merito.

Attualmente c'è la sensazione di essere giunti, pur non senza difficoltà, ad una distinzione sufficientemente condivisa: con il termine rischio sovrano si vuole fare specifico riferimento ai crediti maturati nei confronti dei governi centrali, mentre il rischio paese abbraccia una più vasta gamma di aspetti, ponendosi l'obiettivo di valutare l'affidabilità di un *generico* tipo di investimento all'estero. Naturalmente è più che ragionevole ipotizzare l'esistenza di un'associazione positiva tra il rischio paese ed il rischio sovrano, in quanto variazioni del *business environment* all'interno di un paese sono spesso accompagnate da mutamenti nella stessa direzione della solvibilità del proprio governo. Tuttavia tali relazioni non sono deterministiche, motivo per cui risulta importante tenere distinte le due nozioni.

Un forte impulso in questo senso è stato dato dall'intervento di Meldrum, che nel 2000 ha proposto un esauriente impianto definitorio individuando sei determinanti del rischio paese, una delle quali identificabile proprio con il

rischio sovrano.[1] Questa operazione ha contribuito a stabilire forse definitivamente un chiaro rapporto tra i due concetti. Meldrum ha inquadrato il rischio paese come

> "l'insieme dei rischi che non si sostengono se si effettuano delle transizioni nel mercato domestico ma che emergono nel momento in cui si effettua un investimento in un Paese estero [...]"

ed ha elencato le sue sei componenti, di seguito brevemente descritte:

- **rischi economici:** analizzano l'insorgenza di fattori che potrebbero nuocere all'investimento direttamente (microeconomici) o indirettamente, ovvero modificando la situazione congiunturale nel paese (macroeconomici);
- **rischio di trasferimento:** considera l'eventualità che le Autorità del paese in cui si effettua l'investimento applichino restrizioni ai movimenti di capitale, facendo sorgere difficoltà principalmente nelle operazioni di rimpatrio di profitti e dividendi;
- **rischio di fluttuazione del tasso di cambio:** misura la probabilità che il suddetto tasso vari in modo sfavorevole all'investitore o che ci siano cambiamenti nel regime (da tassi fissi a tassi variabili o viceversa);
- **rischio di localizzazione geografica:** prende in esame l'avvento di calamità naturali o la creazione di situazioni di tensione derivanti da rapporti difficili con paesi confinanti;
- **rischio di merito creditizio governativo (rischio sovrano):** riguarda, come già accennato, esclusivamente chi investe nelle obbligazioni governative e valuta la possibilità e la volontà del governo centrale di far fronte agli impegni finanziari assunti;
- **rischio politico:** comprende l'insieme delle eventuali decisioni del governo (espropri, confische, dazi ed altri atti unilaterali) che potrebbero danneggiare gli investitori.

Naturalmente queste componenti non formano dei compartimenti stagni ma risultano interdipendenti ed influenzate da fattori comuni.

[1] *Cfr.* [7].

1.2 Le agenzie di *rating*

Nel suddetto contesto si collocano le agenzie di *rating*, il cui compito è quello di stabilire, mediante metodologie quantitative e qualitative, delle graduatorie in cui i vari stati siano ordinati in base ai differenti livelli di rischio a loro associati. Le numerose agenzie esistenti hanno finalità simili ma vengono classificate in due gruppi secondo un criterio strettamente legato alla distinzione tra rischio paese e rischio sovrano:

- Il primo gruppo produce un **Global Country Risk Ranking**, in cui la valutazione degli stati tiene conto di *tutte* le possibili fonti di rischio paese. Ne fanno parte organismi come il *Business Environment Risk Intelligence* (*BERI*), il *Control Risk Group* o le *Export Credit Agencies* (*ECA*), ma anche alcune riviste specializzate come *Commercio Internazionale*.

- Il secondo gruppo si concentra *esclusivamente* sul **Country Credit Rating** o **Sovereign Rating**. I suoi membri cercano di quantificare il rischio che i governi non riescano a fronteggiare il proprio debito, dichiarando così *default* e venendo meno agli impegni presi con gli investitori. Le più note agenzie di questa categoria sono tre: Moody's, Fitch e Standard&Poor's.[2] Si osservi, per quanto riguarda la stima della probabilità di *default*, che ad essere valutata non è soltanto la *capacità* di onorare il debito, ma anche la *volontà* di farlo; l'autorità di cui godono gli stati sovrani consente infatti loro, con modalità precise e norme specifiche che non è opportuno analizzare in questa sede, di non rimborsare o rimborsare solo parzialmente i creditori.

1.2.1 Standard&Poor's, Moody's e Fitch

Per comprendere più approfonditamente il meccanismo che conduce alla formazione delle graduatorie, viene preso in esame l'operato delle tre più famose agenzie di *sovereign rating*. In tutti i casi la scala dei *ranking* è composta da un certo numero di sigle che vanno dal giudizio migliore (AAA per S&P e Fitch, Aaa per Moody's) a quello peggiore, rappresentato dalla

[2] D'ora in avanti per brevità ci si riferirà a questa agenzia con la sigla S&P.

lettera D che corrisponde al *default*. I debitori ritenuti più affidabili saranno assegnati alle prime categorie, quelli più a rischio nelle fasce più basse e così via, giungendo infine alle situazioni di *default*. La tabella 1.1 illustra le tre scale complete.

Si osservi il particolare ruolo della soglia BBB-, che divide i cosidetti titoli *investment grade* o di qualità bancaria da quelli di livello più basso, denominati *speculative grade* o *high yield* in quanto offrono rendimenti più elevati proprio a causa del maggiore rischio che il loro acquisto comporta. L'importanza di questa soglia è dovuta all'esistenza di norme di vigilanza che impediscono ad alcuni operatori finanziari istituzionali di acquistare i titoli appartenenti alla seconda categoria, proprio perchè non ritenuti abbastanza sicuri.

Sebbene le scale utilizzate dalle tre agenzie siano simili, i criteri di classificazione utilizzati per stilare gli effettivi giudizi sugli stati differiscono lievemente: S&P discrimina solo sulla base della probabilità di *default* (PD) in un arco temporale corrispondente alla durata dei titoli, mentre Moody's include nelle sue valutazioni anche il tasso stimato di perdite sui crediti (*loss given default* o LGD), considerando il prodotto $PD \cdot LGD$. Fitch si pone in una situazione ibrida, dal momento che i suoi *rating* coincidono con quelli di S&P *prima* che accada l'episodio di *default* e con quelli di Moody's *dopo* il verificarsi di tale evento.

A differenza di altri organismi, le tre agenzie non pubblicano i loro risultati con cadenza regolare, ma solo quando lo ritengono più opportuno; tuttavia ciò non significa che l'insorgenza di ogni nuova informazione comporti necessariamente un repentino mutamento dei giudizi. L'orientamento degli addetti ai lavori è rivolto proprio nella direzione opposta: la caratteristica considerata generalmente più auspicabile per i *rating* è infatti la *stabilità* nel tempo. In quest'ottica gli aggiustamenti che avvengono nell'arco di dodici mesi e che comportano la variazione, in entrambe le direzioni, di almeno tre *notches* (ovvero il *rating* varia di tre categorie) sono definiti *fallimenti del rating*.[3]

[3] *Cfr.* [8]. Non si considerano fallimenti tutti gli aggiustamenti da, verso, e all'interno delle categorie CCC (Caa per Moody's) o inferiori: la motivazione di tale scelta risiede nella maggiore volatilità a cui sono sottoposti i giudizi a quei livelli.

Tabella 1.1: scala dei livelli di *rating*

S&P	**Moody's**	**Fitch**	**Caratteristiche**
	Investment grade		
AAA	Aaa	AAA	Migliore qualità.
AA+	Aa1	AA+	Ampi margini di
AA	Aa2	AA	protezione di
AA-	Aa3	AA-	principale e interessi.
A+	A1	A+	Grado medio-elevato. Buoni
A	A2	A	fattori di sicurezza per
A-	A3	A-	principale e interessi.
BBB+	Baa1	BBB+	Grado medio. Margini di protezione
BBB	Baa2	BBB	adeguati nel presente; alcuni
BBB-	Baa3	BBB-	elementi di incertezza futura.
	Speculative grade		
BB+	Ba1	BB+	Bond con elementi speculativi.
BB	Ba2	BB	Il loro futuro non può
BB-	Ba3	BB-	essere considerato assicurato.
B+	B1	B+	Limitata sicurezza
B	B2	B	del pagamento di
B-	B3	B-	principale e interessi.
CCC+	Caa1	CCC+	Possibilità di *default*.
CCC	Caa2	CCC	
CCC-	Caa3	CCC-	
CC	Ca	CC	Elevato rischio *default*.
C	C	C	*Default* o
D/SD	D/SD	D/SD	*default* selettivo.

I giudizi possono rivolgersi al breve (1 anno) o al lungo periodo (3-5 anni); questi due orizzonti temporali ricalcano le tradizionali scadenze dei titoli emessi dai principali stati sovrani. Al fine di evitare revisioni troppo frequenti, nei *rating* di lungo periodo è prassi consolidata affiancare al giudizio espresso una direzione (positiva, negativa o stabile) che informi sulla tendenza ritenuta più probabile per il futuro e che prende il nome di *outlook*.

L'ultima importante distinzione da operare riguarda i debiti contratti in valuta *locale* o *estera*. Gli episodi di *default* che storicamente occorrono con maggiore frequenza coinvolgono il debito in valuta estera, a causa della sua più difficile reperibilità da parte degli stati sovrani.[4] Per questi motivi si è giunti alla formazione di due *rating* distinti, dove quello per la valuta locale è tendenzialmente superiore di qualche categoria rispetto all'altro.

1.3 Le metodologie e le variabili di interesse

Le metodologie statistiche impiegate spaziano dall'analisi discriminante ai modelli di *scoring* e alla regressione logistica e coinvolgono numerosi indicatori considerati influenti sui livelli di rischio e quindi fondamentali nella determinazione delle graduatorie. Tuttavia, prima di elencare tali indicatori, è necessario evidenziare che tutti i produttori di *rating* sono accomunati dalla consapevolezza che lo svolgimento esclusivo di analisi quantitative non può ritenersi esaustivo in relazione ad un problema di tale entità, a causa della grande rilevanza assunta da fattori non misurabili e da elementi soggettivi (si pensi soprattutto al rischio politico). Per questo motivo nella maggior parte dei casi le conclusioni quantitative portano ad una *proposta di rating*, successivamente integrata e modificata con considerazioni specifiche di tipo qualitativo formulate dai valutatori stessi. Le dinamiche in questione sono in larga parte riscontrabili nelle procedure implementate dalle tre agenzie specializzate, di cui viene di seguito fornita una breve panoramica.

[4] In linea puramente teorica uno stato sovrano potrebbe disporre senza alcun limite della propria valuta nazionale, ad esempio semplicemente stampando moneta o alzando indiscriminatamente le imposte. In realtà tali opzioni vengono prese in considerazione molto raramente a causa delle ovvie ripercussioni sociali e politiche che comportano.

Fitch sottopone agli emissori di titoli che richiedono la sua valutazione un ampio questionario, a cui segue un'intervista condotta dagli analisti incaricati. Il questionario è composto da numerose sezioni contenenti informazioni sotto forma di variabili numeriche o descrizioni. Gli argomenti trattati spaziano dalla politica macroeconomica alla bilancia dei pagamenti, dall'analisi di produzione e commercio a quella del mercato del lavoro o ad aspetti demografici e sociali.

Standard&Poor's invece individua cinque elementi chiave, sintetizzati in altrettanti *score* numerici con valori da 1, che indica il giudizio migliore, a 6. I diversi tipi di *score* sono:

- politico;
- economico;
- estero;
- fiscale;
- monetario.

I primi due costituiscono il *profilo politico-economico*, mentre gli altri il *profilo di flessibilità e performance*. L' aggregazione dei due profili determina il *rating* in valuta estera, sulla base del quale può essere operato un ritocco verso l'alto di una o due categorie al fine di stabilire il giudizio per la valuta locale. Per ogni elemento-chiave esistono fattori *primari*, che determinano uno *score* iniziale, e *secondari*, analizzando i quali tale *score* può essere migliorato o peggiorato fino ad un massimo di due punti. La tabella 1.2 include per ogni *score* alcune delle variabili quantitative e qualitative che influenzano i suddetti fattori.

Per Moody's, infine, alla formazione del *rating* di un paese prendono parte due componenti, l'*elasticità economica* e la *robustezza finanziaria*, ciascuna a sua volta scomponibile in due aspetti. L'elasticità economica (*Economic Resiliency*) intende misurare la capacità di assorbimento di *shock* esterni da parte del paese mediante la valutazione della *forza economica* e della *qualità delle istituzioni*. La robustezza finanziaria (*Financial Robustness*) invece si pone l'obiettivo di riprodurre il grado di vulnerabilità delle finanze

Tabella 1.2: fattori chiave di S&P

Score	**Fattori primari**	**Fattori secondari**
Politico	Stabilità delle istituzioni, Efficacia dell'attività politica.	Trasparenza dei processi governativi, Qualità dei dati ufficiali.
Economico	PIL procapite.	Crescita del PIL (variazione percentuale), Grado di concentrazione.
Esterno	Status della valuta locale nelle transazioni internazionali, Bisogni Finanziari Esteri.	Posizione Estera Netta.
Fiscale	Debito/PIL.	Indice di liquidità, Indice di sviluppo umano.
Monetario	Regime nazionale dei tassi di cambio, Credibilità della politica monetaria.	Inflazione.

pubbliche e si scinde in *forza finanziaria* e *suscettibilità ad eventi esterni* dell'intero paese. Si riporta di seguito una breve descrizione dei quattro aspetti, segnalando al contempo le variabili più importanti:

Forza Economica: insieme di misure che ruotano attorno al PIL procapite. Per evitare distorsioni dovute ai naturali cicli economici e per migliorare la comparazione tra paesi, si usa il suo valore medio su un periodo di 3-5 anni corretto per le parità di potere d'acquisto. Altri indicatori considerati sono il livello di innovazione e il livello di investimenti nel capitale umano.

Qualità delle Istituzioni: quantificazione dell'affidabilità e della stabilità del governo volta a capire se sarà garantito il rispetto dei contratti stipulati. I due principali indici esaminati sono il *Rule of Law* e il *Government Effectiveness Index (GEI)*, proposti dalla Banca Mondiale.[5]

[5] *Cfr.* [24]

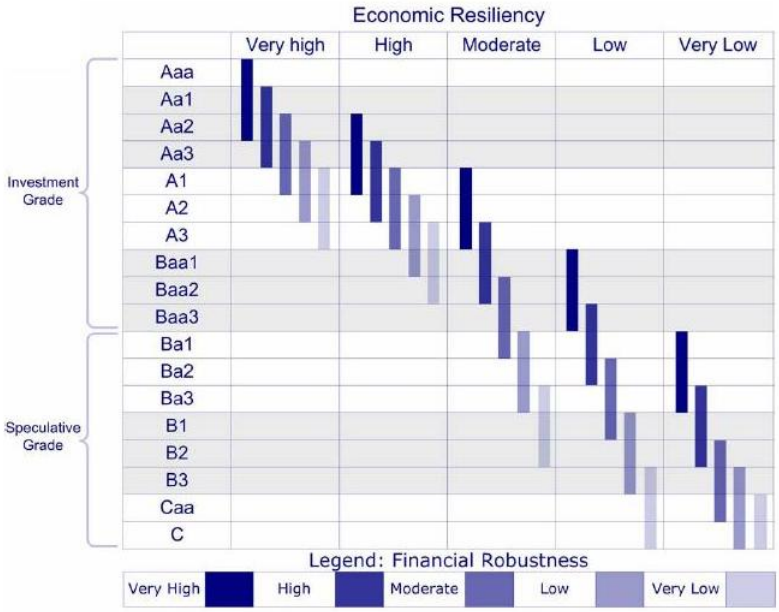

Figura 1.1: mappa del *rating* per Moody's

Forza Finanziaria: valutazione congiunta sull'entità del debito, misurato in quota al PIL, ma anche sulle strategie attuate dal governo per fronteggiarlo e gestirlo.

Suscettibilità ad eventi esterni: stima qualitativa del rischio di insorgenza di eventi che potrebbero in qualche modo minacciare la restituzione del debito. Tale classe di avvenimenti è alquanto ampia e include, ad esempio, guerre, crisi finanziarie internazionali o catastrofi naturali.

La scala ordinale utilizzata prevede cinque modalità (molto alto, alto, medio, basso e molto basso); i fattori si aggregano a coppie per ottenere i due giudizi di componente, la cui valutazione congiunta conduce ad un intervallo sulla scala di *rating*, come si può osservare dalla mappa in figura 1.1. Una volta stabilito questo intervallo, la scelta finale della categoria tiene conto di comparazioni qualitative tra tutti gli stati sovrani appartenenti all'analisi.

In sintesi, nonostante i modelli implementati e le variabili selezionate siano sostanzialmente differenti da caso a caso, è possibile individuare tre aree a cui tutti fanno riferimento: quella *macroeconomica*, quella *finanziaria* e quel-

Tabella 1.3: principali variabili per il rating

Gruppo	**Variabile**
Macroeconomiche	PIL procapite
	Variazione percentuale del PIL
	Tasso di inflazione
	Tasso di disoccupazione
	Investimenti netti/PIL
	Indice dei prezzi al consumo
Finanziarie	Debito estero/PIL
	Entrate/PIL
	Debito pubblico/PIL
	Debito estero/Esportazioni
	Interessi/Esportazioni
	Debito estero/Tot. Debito
Politiche	Indice di sviluppo umano
	Rule of Law
	GEI
	Indice di unità di governo

la *socio-politica*. Esaminando le metodologie adottate dalle tre agenzie, si nota un frequente ricorso alla scala ordinale. Questa scelta è quasi obbligata per gli indicatori del terzo gruppo, i quali coinvolgono valutazioni prettamente soggettive, ma è altrettanto diffusa per le altre misure, nonostante esse siano disponibili sotto forma di "variabili reali".

La tabella 1.3 raccoglie gli indici che ricorrono con maggiore frequenza in letteratura.

Capitolo 2

La base-dati

2.1 Presentazione del data-set

Le informazioni d'interesse sono state estratte da un archivio relativo all'anno 2010 predisposto dall'*Economist Intelligence Unit (EIU)*.[1] Si tratta in particolare di 33 indicatori economico-finanziari di natura quantitativa riguardanti 200 paesi di tutto il mondo. La tabella 2.1 include l'elenco completo con la classificazione per settore, mentre l'appendice A.1 riporta le corrispondenti etichette insieme alle unità di misura.

Le variabili sono affiancate da una misura di *rating* ufficiale, che sintetizza i giudizi delle tre agenzie andando a prendere il secondo migliore (***second best***) dei tre. In questa fase preliminare non viene considerato il segno associato alle categorie; si osservano pertanto sette modalità (da AAA a CCC), anche se il giudizio più basso è assegnato solamente a due paesi. In ben 94 casi l'informazione non è disponibile ed è stata adottata l'usuale codifica per i dati mancanti `NA`. La distribuzione di frequenza completa è riportata in tabella 2.2 a pagina 14.

Per molte variabili la presenza di valori non osservati è rilevante; le maggiori criticità sono riscontrabili in particolare per tre indicatori: il grado di concentrazione, l'andamento della borsa valori e il rapporto tra riserve internazionali e debito estero a breve termine. In questi casi la proporzione di *missing values* supera il 50 per cento.

[1] *Cfr.* [2].

Tabella 2.1: variabili estratte dall'archivio *EIU*

Settore	**Variabile**
strutturali	PIL nominale, PIL reale, PIL procapite
	PIL (Volatilità 10 anni)
	Grado di concentrazione
	Importazioni, Esportazioni
	Grado di apertura commerciale
reali e monetarie	PIL (Variazione percentuale)
	Inflazione
	Credito interno
	Credito/PIL
	Tasso di interesse sul mercato monetario
	Tasso di cambio effettivo reale
	Tasso di cambio (Volatilità 10 anni)
posizione con l'estero	Saldo di conto corrente/PIL
	Bilancia dei pagamenti
	Saldo di bilancia commerciale/PIL
	Flusso IDE in entrata/PIL
	Riserve/Uscite correnti
posizione debitoria estera	Riserve internazionali
	Debito/PIL
	Servizio del debito/Entrate valutarie
	Debito estero totale
	Debito totale di breve periodo
	Riserve internazionali/Debito estero a breve
	Variazione riserve internazionali
	Attività nette/PIL
posizione debitoria interna	*Stock* di debito pubblico/PIL
	Saldo di bilancio pubblico/PIL
	Debito estero del settore pubblico/PIL
prezzi e mercati	Andamento borsa valori
	Credit spread per *CDS*

Tabella 2.2: distribuzione di frequenza dei *rating* ufficiali

Livello	AAA	AA	A	BBB	BB	B	CCC	NA	Totale
Frequenza	15	10	11	23	20	25	2	94	200

Tabella 2.3: PIL procapite: statistiche descrittive per livello di *rating*

Livello	**Min**	**Media**	**Max**	**Dev.Std.**	**Freq.**
AAA	289.28	369.54	749.85	117.16	15
AA	66.13	284.68	578.72	130.59	10
A	117.79	171.92	247.57	48.42	11
BBB	30.13	149.23	553.87	117.38	23
BB	12.43	69.35	127.91	39.35	20
B	8.05	43.14	135.57	35.10	25
CCC	86.38	161.70	237.02	106.52	2
NA	1.44	82.77	1175.74	167.69	72
Tot.complessivo	1.44	126.19	1175.74	155.66	178

2.2 Analisi descrittive

Il primo passo da effettuare consiste nel calcolo di alcune statistiche descrittive, che si pongono l'obiettivo di individuare le variabili con maggiore capacità discriminatoria per il *rating*. Si ritiene opportuno a tal fine produrre due tabelle per ogni indicatore a disposizione; in questa sede vengono proposte a titolo esemplificativo quelle relative al PIL procapite.[2]

La prima contiene la media aritmetica, il minimo, il massimo, la deviazione standard e la frequenza per ogni livello di *rating*. È ragionevole attendersi che nelle variabili più discriminanti tali statistiche seguano un *trend* approssimativamente crescente (o decrescente) rispetto al *rating* stesso.

La seconda prevede la costruzione di sei classi ordinate sulla base dei quantili empirici, in accordo al criterio esposto dalla tabella 2.5 nella pagina successiva. In questa fase i valori mancanti sono mantenuti tali e codificati con la medesima sigla `NA`. La scelta di costituire sei classi, e non sette quante le categorie presenti nell'archivio, è dettata dalla scarsa numerosità

[2] *Cfr.* tabelle 2.3 e 2.4.

Tabella 2.4: PIL procapite: distribuzione delle classi condizionata ai giudizi

Livello	**1**	**2**	**3**	**4**	**5**	**6**	**NA**	**Tot.**
AAA	0.0	11.8	11.8	23.5	17.6	23.5	11.8	100
AA	0.0	40.0	40.0	20.0	0.0	0.0	0.0	100
A	7.7	15.4	7.7	15.4	38.5	7.7	7.7	100
BBB	17.4	13.0	17.4	17.4	13.0	4.3	17.4	100
BB	13.6	4.5	22.7	22.7	13.6	9.1	13.6	100
B	7.1	14.3	25.0	14.3	17.9	7.1	14.3	100
CCC	0.0	0.0	0.0	50.0	50.0	0.0	0.0	100
NA	9.4	12.9	24.7	27.1	8.2	9.4	8.2	100
Tot.complessivo	9.0	13.5	22.0	22.5	13.5	9.0	10.5	100

Tabella 2.5: criteri per la determinazione delle classi

Classe	**Valori**
1	minori del decimo percentile
2	tra il decimo percentile ed il primo quartile
3	tra il primo quartile e la mediana
4	tra la mediana e il terzo quartile
5	tra il terzo quartile ed il novantesimo percentile
6	maggiori del novantesimo percentile

dei giudizi di tipo CCC. Le classi così costruite vengono successivamente incrociate con i livelli di *rating*, riempiendo le celle con le frequenze relative percentuali riportate ai totali di riga, al fine di ottenere le distribuzioni condizionate ai diversi giudizi. In questo modo le tabelle associate alle misure più rappresentative dovrebbero registrare le frequenze più alte nelle celle prossime alla "pseudo-diagonale" principale o secondaria, a seconda che ci sia rispettivamente discordanza o concordanza tra la variabile in esame ed il *rating*.

In ottemperanza a questi criteri si individuano, per ogni gruppo, le variabili che risultano più discriminanti. Per le strutturali e monetarie si osservano:

- il PIL procapite (YPCP);
- la volatilità del PIL (DGDPvol10);
- il grado di apertura commerciale (GRAPCOM);
- l'inflazione (DCPI);
- il credito interno (SODD);
- il tasso di interesse sul mercato monetario (RAT3).

Per quanto riguarda la posizione debitoria sono da segnalare:

- il rapporto tra riserve internazionali e debito estero a breve (RISINT);
- il debito estero del settore pubblico sul PIL (DEBEST);
- il saldo di bilancio pubblico sul PIL (PSBR);
- lo *stock* di debito pubblico sul PIL (PUDP).

Tra le rimanenti invece si distinguono:

- il saldo di bilancia commerciale sul PIL (TDRA);
- il rapporto tra Flusso IDE in entrata e PIL (FlussoIDE);
- il *credit spread* (spreadCDS).

In tutti i casi la direzione della relazione empirica è quella auspicata: non si riscontrano contrasti con le teorie macroeconomiche sottostanti.[3]

2.3 Obiettivi e metodologie

Svolte le analisi preliminari, è possibile iniziare la costruzione di adeguati modelli statistici per la definizione del *rating*. L'idea di base consiste nel dividere il campione osservato separando le unità per cui è disponibile il giudizio ufficiale, utilizzate per la stima dei parametri, dalle altre, su cui

[3] Ci si attende, ad esempio, che i giudizi aumentino al crescere del PIL procapite e diminuiscano al crescere del debito pubblico.

sarà effettuata la *prediction*. Per giustificare l'uso di questa tecnica è necessario prima verificare che non esista una differenza sistematica tra i due gruppi di unità, che introdurrebbe una distorsione in fase di previsione, ed implementare opportuni algoritmi di imputazione per i valori mancanti dei rimanenti indicatori. Vista l'ampia diffusione dei *missing values*, il ricorso a tecniche imputative risulta inevitabile: solamente 27 paesi presentano dei *record* completi. Questa doppia problematica è affrontata nel capitolo successivo.

Un'altra importante questione concerne la tipologia di modelli da adottare: in linea di massima la scelta classica è la regressione logistica lineare per la stima della probabilità di *default*, con variabile di risposta binaria a segnalare l'occorrenza del fallimento. Tuttavia per applicare il suddetto modello si avrebbe bisogno di una serie storica molto profonda, in cui è possibile rilevare eventi di *default*. Nel caso in esame, invece, vengono trattati dati relativi ad un solo anno. Per questo motivo lo scopo è quello di utilizzare le informazioni quantitative e qualitative dei paesi per riuscire a predire le probabilità di *default*, espressione del loro merito creditizio. Alle unità che dispongono del *rating* ufficiale è infatti possibile assegnare, tramite una procedura di *mapping*, un valore di PD.

Al fine di ottenere una classificazione più dettagliata, è opportuno in questa fase introdurre anche i segni; si passa pertanto da 7 a 19 categorie.[4]

Utilizzando questa conversione è possibile stimare un modello lineare del tipo

$$logit(PD_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

ed ottenere, a partire dai *fitted values*, i valori di PD predetti dal modello, da cui ricavare poi una misura di *rating*. Il ricorso alla trasformazione *logit* garantisce gli stessi vantaggi presenti nella regressione logistica, il principale dei quali, com'è noto, è l'estensione del supporto della variabile di risposta, che passa dall'intervallo $[\,0\,;\,1\,]$ all'intero asse reale.

Il modello appena illustrato viene in aggiunta affiancato da una regressione logistica ordinale; in questo caso, per lavorare con un sufficiente numero di osservazioni in ogni classe e per ottenere risultati più facilmente leggi-

[4] *Cfr.* tabella 2.6.

Tabella 2.6: *mapping* per la *PD*

Livello	**PD (%)**	**intervallo (%)**	
AAA	0.01	0.00	0.024
AA+	0.03	0.024	0.029
AA	0.03	0.029	0.034
AA-	0.03	0.034	0.039
A+	0.05	0.039	0.07
A	0.09	0.07	0.11
A-	0.13	0.11	0.16
BBB+	0.20	0.16	0.24
BBB	0.30	0.24	0.37
BBB-	0.46	0.37	0.56
BB+	0.69	0.56	0.85
BB	1.05	0.85	1.29
BB-	1.59	1.29	2.98
B+	3.99	2.98	4.53
B	6.31	4.53	10.43
B-	16.03	10.43	17.74
CCC+	22.12	17.74	27.32
CCC	31.63	27.32	42.07
CCC-	45.00	42.07	100

bili, è stato fatto nuovamente riferimento all'iniziale classificazione basata sulle sette categorie. La trattazione teorica della regressione logistica ordinale, inquadrata nel contesto dei modelli lineari generalizzati, è oggetto del capitolo 4.

Prima di intraprendere il percorso tracciato in questo paragrafo, sono state scartate le tre variabili in cui è assente almeno la metà delle osservazioni. Si reputa infatti eccessiva questa proporzione di dati mancanti e si ritiene che l'inserimento di tali indicatori nei modelli dopo la fase di imputazione possa risultare fuorviante.

Capitolo 3

Dati mancanti

3.1 Il meccanismo generatore dei dati mancanti

Il problema dei dati mancanti affligge la quasi totalità delle ricerche empiriche condotte in ogni campo, senza esclusione di quello preso in esame in questo lavoro. Il loro trattamento è stato pertanto teorizzato e studiato nel corso degli anni. L'approccio più seguito, le cui linee guida vengono di seguito brevemente descritte, venne portato avanti a partire dalla metà degli anni '70 da Rubin.[1] La ragionevole idea alla base afferma che se esiste una differenza sistematica tra le unità che presentano dei valori mancanti e le altre, è molto probabile che le conclusioni inferenziali tratte dal campione dei soli *record* completi siano affette da distorsione.

In termini formali è necessario definire un adeguato modello probabilistico: sia $\mathbf{Y}$ una matrice di dimensioni $(n \times p)$ corrispondente ad un data-set con n osservazioni di p variabili con dati mancanti, e sia $\mathbf{R}$ la matrice, con medesime dimensioni, associata a $\mathbf{Y}$ e contenente in ogni cella una variabile indicatrice R_{ij}, con $R_{ij} = 0$ se il valore della $j-$esima variabile non è osservato per l'unità $i-$esima e $R_{ij} = 1$ altrimenti. Si supponga inoltre che $\mathbf{Y}$ venga partizionata nell'insieme degli elementi osservati $\mathbf{Y}_{\text{obs}}$ e nell'insieme degli elementi mancanti $\mathbf{Y}_{\text{miss}}$ in modo che $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}})$.

Il meccanismo dei dati mancanti è detto ***MAR*** (***Missing At Random***) se la probabilità di occorrenza di un *missing* non dipende dal suo valore,

[1] *Cfr.* [21].

ma solo da quelli delle variabili completamente osservate nell'analisi. La formalizzazione del meccanismo *MAR* è la seguente:

$$\mathbb{P}(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}, \boldsymbol{\xi}) = \mathbb{P}(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \boldsymbol{\xi}) \tag{3.1}$$

dove $\boldsymbol{\xi}$ è l'insieme degli altri parametri che governano il modello probabilitstico per $\mathbf{R}$. Se l'assunzione di dati *MAR* è realistica, è possibile catturare la differenza tra le unità complete ed incomplete semplicemente controllando per gli elementi di $\mathbf{Y}_{\text{obs}}$. I dati mancanti costituiscono un campione casuale all'interno degli strati definiti dalle variabili osservate.

Un caso particolare si verifica quando il meccanismo dei *missing data* è indipendente anche dai valori osservati $\mathbf{Y}_{\text{obs}}$, ovvero quando

$$\mathbb{P}(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}, \boldsymbol{\xi}) = \mathbb{P}(\mathbf{R}|\boldsymbol{\xi}). \tag{3.2}$$

La situazione appena descritta viene invece definita dall'acronimo ***MCAR*** (***Missing Completely At Random***). Ovviamente l'assunzione *MCAR* è più forte dell'altra ed implica che i dati mancanti rappresentino un campione casuale non più all'interno dei livelli definiti da $\mathbf{Y}_{\text{obs}}$, bensì dell'intero archivio.

Indicando con $\boldsymbol{\theta}$ il vettore dei parametri del modello probabilistico per $\mathbf{Y}$, Rubin ha mostrato che, in caso di meccanismo *MAR* con $\boldsymbol{\theta}$ e $\boldsymbol{\xi}$ distinti, l'inferenza per $\boldsymbol{\theta}$ può basarsi su una funzione di verosimiglianza che non include al suo interno il termine relativo a $\mathbf{Y}_{\text{miss}}$. In sostanza il meccanismo dei dati mancanti, sotto queste condizioni, non incide sull'inferenza per $\boldsymbol{\theta}$ ed è perciò definito **ignorabile**.[2]

3.1.1 Il test di Little per l'ipotesi di *MCAR*

Nel 1988 Roderick Little ha proposto un test per verificare l'ipotesi di *MCAR* in un data-set $\mathbf{Y}$ composto sempre da n osservazioni di p variabili.[3] Il principale vantaggio derivante dal suo contributo consiste nella possibilità di verificare mediante un singolo test statistico l'ipotesi di distribuzione *MCAR* per un intero insieme di dati.

[2] *Cfr.* [10] e [20].
[3] *Cfr.* [20].

I metodi alternativi considerano separatamente ciascuna variabile Y con valori mancanti, dividendo le unità campionarie in due sottogruppi a seconda che il valore di Y sia osservato o meno. Successivamente gli altri indicatori vengono sottoposti a dei t-test per la differenza tra medie nei due sottogruppi. Scostamenti statisticamente significativi tra le due medie forniscono evidenza contro l'ipotesi nulla di $MCAR$.

Il maggiore difetto di tali procedure emerge al crescere della dimensione del data-set analizzato: la struttura di correlazione delle numerose statistiche t infatti è complessa e rende proibitiva la loro valutazione congiunta a livello inferenziale.

Per illustrare il test proposto da Little risulta opportuno introdurre un po' di notazione preliminare. A partire dalla matrice dei dati $\mathbf{Y}$ e dalla corrispondente matrice indicatrice $\mathbf{R}$ vengono definiti i relativi vettori riga di dimensione $(1 \times p)$: $\boldsymbol{y}_i$, i-esimo record del data-set e $\boldsymbol{r}_i$, che prende il nome di ***missing data pattern***. J è il numero di *missing data pattern* distinti (tenendo conto che un vettore di osservazioni complete conta come *pattern*), S_j è l'insieme delle unità che condividono il j-esimo *pattern* e m_j è la cardinalità di S_j, con $j = 1, \ldots, J$ e ovviamente $\sum_{j=1}^{J} m_j = n$.

Ponendo pari a p_j il numero delle variabili osservate per il j-esimo *pattern*, sia $\boldsymbol{y}_{\text{obs},i}$ il vettore $(1 \times p_j)$ contenente i valori delle variabili osservate nell'i-esima unità e sia $\bar{\boldsymbol{y}}_{\text{obs},j} = \frac{1}{m_j} \sum_{i \in S_j} \boldsymbol{y}_{\text{obs},i}$ il vettore di identiche dimensioni che contiene le medie delle variabili osservate calcolate sulle unità che condividono il j-esimo *pattern*. Viene inoltre indicata con $\mathbf{D}_j$ una matrice di dimensione $(p \times p_j)$. Ciascuna colonna di $\mathbf{D}_j$ corrisponde ad una delle variabili osservate nel j-esimo *pattern* ed assume sempre valore 0, tranne nella riga associata alla variabile in questione, in cui si assegna valore 1. A titolo esemplificativo si immagini una situazione con 5 variabili e si consideri il *pattern*

$$\boldsymbol{r}_i = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

in cui sono osservate la prima, la terza e la quinta variabile: la matrice $\mathbf{D}_j$

corrispondente sarà pertanto

$$\mathbf{D}_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Si considerino i parametri della popolazione $\boldsymbol{\mu}$ (vettore delle medie) e $\boldsymbol{\Sigma}$ (matrice di varianza e covarianza) e le relative stime di massima verosimiglianza $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\Sigma}}$ calcolate assumendo che il meccanismo generatore dei dati mancanti sia ignorabile, e si ponga infine

$$\begin{cases} \boldsymbol{\mu}_{\text{obs},j} = \boldsymbol{\mu}\mathbf{D}_j & \boldsymbol{\Sigma}_{\text{obs},j} = \mathbf{D}_j^{\text{T}}\boldsymbol{\Sigma}\mathbf{D}_j \\ \hat{\boldsymbol{\mu}}_{\text{obs},j} = \hat{\boldsymbol{\mu}}\mathbf{D}_j & \hat{\boldsymbol{\Sigma}}_{\text{obs},j} = \mathbf{D}_j^{\text{T}}\hat{\boldsymbol{\Sigma}}\mathbf{D}_j. \end{cases}$$

Assumendo questa notazione è possibile dimostrare che, sotto l'ipotesi che i valori mancanti siano generati da un meccanismo di tipo *MCAR*, la statistica test

$$d^2 = \sum_{j=1}^{J} m_j(\bar{\boldsymbol{y}}_{\text{obs},j} - \hat{\boldsymbol{\mu}}_{\text{obs},j})\hat{\boldsymbol{\Sigma}}_{\text{obs},j}^{-1}(\bar{\boldsymbol{y}}_{\text{obs},j} - \hat{\boldsymbol{\mu}}_{\text{obs},j})^{\text{T}} \tag{3.3}$$

segue una distribuzione campionaria asintotica χ^2 con f gradi di libertà, dove

$$f = \sum_{j=1}^{J} p_j - p.$$

L'ipotesi nulla sarà rifiutata per valori della statistica test superiori al valore critico associato al livello di significatività prescelto.

L'implementazione del test per l'archivio a disposizione, composto a questo punto da 30 indicatori unitamente alla probabilità di *default*, produce un valore della statistica d^2 pari a 1628.444, da confrontare con una distribuzione χ^2 con 1607 gradi di libertà per un *p-value* di 0.35. Basandosi sui dati campionari dunque non c'è evidenza per rifiutare l'ipotesi nulla di meccanismo *MCAR*. Risulta pertanto giustificato il ricorso a tecniche imputative al fine di ottenere, a partire dal data-set contenente i soli paesi che riportano la misura di *rating*, un archivio completo su cui effettuare le analisi alle quali è stato fatto riferimento.

Tabella 3.1: codifica numerica per il *rating*

Livello	AAA	AA	A	BBB	BB	B	CCC
Punteggio	1	2	3	4	5	6	7

3.2 Algoritmi di imputazione

I due algoritmi di imputazione costruiti sono riconducibili al noto metodo del vicino più vicino (*nearest neighbour*) poichè sfruttano le informazioni a disposizione per definire delle distanze tra le unità del data-set. Supponendo che l'osservazione A per la variabile B sia mancante, l'obiettivo è quello di determinare un certo numero di unità che siano le più *vicine* ad A tra quelle con un valore osservato per B.[4] Il dato imputato sarà costituito dalla media aritmetica dei valori assunti da queste unità. Il numero di unità che si intende selezionare per il calcolo di tale media rappresenta il primo dei tre parametri in ingresso comuni ad entrambi gli algoritmi. Esso è indicato con il simbolo `p` e il suo valore di *default* è pari a 1.

Il secondo è costituito dall'oggetto `elenco`, in cui è possibile specificare, quando necessario, il sottoinsieme di unità a cui limitare il processo di imputazione. In questo modo al termine della procedura le unità escluse conserveranno i loro eventuali valori mancanti. L'opzione di *default* ordina agli algoritmi di coinvolgere tutte le unità.[5]

L'ultimo parametro in comune è il vettore `punteggio`, che corrisponde semplicemente alla trasposizione numerica del *rating* (senza segno).[6] La sua funzione è di tipo diagnostico e verrà illustrata insieme alla descrizione dei due algoritmi presente nei prossimi paragrafi. Pur condividendo la stessa idea di base, essi presentano alcune differenze sostanziali, pertanto saranno introdotti separatamente.

L'appendice B riporta i codici del linguaggio di programmazione statistico `R` per la loro implementazione.

[4] Naturalmente il concetto di vicinanza è definito dalla distanza impiegata.

[5] L'utilità di questa opzione è limitata alla fase sperimentale degli algoritmi, in cui essi erano applicati soltanto ad un piccolo insieme di paesi, al fine di controllare più dettagliatamente il loro corretto funzionamento.

[6] *Cfr.* tabella 3.1.

3.2.1 L'algoritmo `nearneigh`

Il primo algoritmo è indicato con la funzione `nearneigh` e riceve in ingresso, oltre ai già citati parametri comuni, le variabili del data-set separando quelle completamente osservate (racchiuse nell'oggetto `nomiss`), da quelle con almeno un valore mancante (`miss`). Viene quindi definito l'oggetto `distanze`, matrice delle distanze euclidee tra le unità calcolate a partire dagli indicatori contenuti in `nomiss`. Si dichiara inoltre la matrice `diagn`, delle medesime dimensioni di `miss`, senza però istanziarne le celle.

A questo punto si procede implementando due cicli `for` annidati sulle unità (ciclo `i` per le righe) e sulle variabili (ciclo `j` per le colonne) di `miss`. Per ogni *missing value* incontrato si ordina la i-esima riga di `distanze` e si estraggono, mediante la funzione `match`, gli indici delle unità ad essa più vicine. I loro valori per la variabile in esame sono sistemati sequenzialmente nel vettore `z` dopo aver controllato che non siano mancanti. Il processo continua finchè `z` non raggiunge la lunghezza `p` desiderata; il valore imputato sarà costituito, come detto, dalla media aritmetica dei suoi elementi.

Sempre all'interno del doppio ciclo `for` trova spazio la realizzazione di misure diagnostiche: ad ogni iterazione sulle colonne si computano due vettori, composti da sette elementi e denominati `massimicond` e `minimicond`, contenenti rispettivamente i massimi e i minimi dell'indicatore di turno condizionati alle sette modalità di `punteggio`. In questi casi sono state utilizzate opportune funzioni che ignorano i dati mancanti. Per ogni passo del ciclo `i` invece si estraggono dai due vettori gli elementi di indice pari al punteggio dell'unità interessata, definendo gli scalari `maximum` e `minimum` sulla base dei quali viene determinato l'intervallo [`minimum`; `maximum`]. Le celle di `diagn` corrispondenti ai valori non osservati di `miss` assumeranno valore 0 se il dato imputato è compreso in tale intervallo e 1 altrimenti, mentre le altre non sono di interesse.

Gli oggetti restituiti dalla funzione sono il data-set privo di valori mancanti (`imputato`) e la matrice `diagn`.

3.2.2 L'algoritmo `imputaznew`

Il più grave limite della funzione `nearneigh` consiste nel basare il calcolo delle distanze esclusivamente sull'insieme delle variabili complete, scartando così il rilevante patrimonio informativo rappresentato dai rimanenti indici. Superare tale difetto è il principale obiettivo del secondo algoritmo proposto, al quale è possibile passare le variabili in un unico blocco tramite l'argomento `dati`.

All'interno del ciclo `j` viene adesso definita la matrice `daticompl` contenente tutte le colonne ma non tutte le righe: sono infatti inserite unicamente le unità con un valore osservato per la j-esima variabile. Soltanto per esse saranno calcolate, nel ciclo `i`, le distanze con l'unità in esame (che naturalmente non può comparire tra le righe di `daticompl`) come media aritmetica della differenza tra *record* in valore assoluto. Si osservi che in questo modo potrebbero essere presenti dei *missing* in entrambi i *record*: qualora vi fossero, sarebbero ignorati nel calcolo della distanza. I passi successivi ricalcano concettualmente quelli del precedente algoritmo: vengono ordinate le unità sulla base delle distanze, andando ad imputare la media aritmetica dei primi `p` valori.

Le misure diagnostiche rimangono inalterate al fine di comparare le prestazioni dei due algoritmi: la loro analisi mostra che la funzione `nearneigh` ottiene risultati meno apprezzabili dell'algoritmo `imputaznew`, che sarà pertanto adottato come scelta finale. Si è inoltre verificato empiricamente che il valore migliore da assegnare a `p` è pari a 10.

La tabella 3.2 riporta le medie e le deviazioni standard delle 19 variabili con valori mancanti prima e dopo il processo di imputazione, accompagnate dalle variazioni relative. Gli scostamenti maggiori riguardano il debito estero totale e il debito estero di breve periodo, anche se in definitiva non si ravvisano eccessive criticità.

Tabella 3.2: medie e deviazioni standard prima e dopo l'imputazione

Variabile	**Media**			**Dev.Std.**		
	Prima	**Dopo**	**Var.%**	**Prima**	**Dopo**	**Var.%**
SODD	12.96	13.04	0.58	14.15	14.10	-0.33
CREDsuPIL	87.40	86.87	-0.61	74.57	74.42	-0.21
MEXP	123.78	122.66	-0.91	255.10	254.14	-0.37
MIMP	-122.33	-121.24	-0.89	270.00	268.94	-0.39
GRAPCOM	70.93	71.00	0.09	38.61	38.43	-0.46
XRREvol10	6.55	6.56	0.14	5.39	5.34	-0.94
Tcambio	1.85	1.80	-2.91	7.93	7.87	-0.82
FlussoIDE	2.11	2.14	1.78	3.69	3.64	-1.14
ILMA	86.98	84.19	-3.21	310.90	303.81	-2.28
VARISINT	9.28	9.09	-2.02	18.14	17.68	-2.51
MCOV	6.52	6.45	-1.09	5.26	5.09	-3.21
RAT3	4.19	4.22	0.66	3.45	3.33	-3.30
PUDP	48.00	47.23	-1.60	32.92	31.61	-3.99
ATTNETTE	6.19	5.93	-4.19	52.80	49.12	-6.96
DEBEST	16.81	16.17	-3.77	12.75	11.40	-10.65
TDBT	77.45	89.24	15.22	164.73	157.25	-4.54
TSTD	17.75	20.06	13.00	36.66	34.51	-5.85
TDPY	40.51	40.50	-0.01	29.71	26.19	-11.84
TSPD	12.47	12.85	3.02	13.63	12.06	-11.54

Capitolo 4

Modelli lineari generalizzati

Questo capitolo intende presentare brevemente il modello di regressione logistica ordinale all'interno della vasta famiglia dei modelli lineari generalizzati, il cui impianto teorico è stato costruito agli inizi degli anni '70 da Nelder e Wedderburn.

4.1 Famiglia esponenziale e modelli lineari generalizzati

Definizione 1 (**Famiglia Esponenziale**)**.** Si consideri una variabile casuale Y la cui distribuzione di probabilità f dipenda da un solo parametro di interesse θ. La distribuzione f appartiene alla **famiglia esponenziale** se assume la seguente forma:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{4.1}$$

dove a, b, s and t sono funzioni note.

L'equazione (4.1) può essere facilmente riscritta come:

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)} \tag{4.2}$$

con $s(y) = e^{d(y)}$ e $t(\theta) = e^{c(\theta)}$.

Se $a(y) = y$ si parla di distribuzione **in forma canonica** e $b(\theta)$ prende il nome di **parametro naturale** di tale distribuzione. Nel caso in cui f contenga altri parametri oltre a θ, essi verranno detti **parametri di disturbo** e trattati come se fossero noti.

Definizione 2 (**Modello Lineare Generalizzato**). Si consideri un insieme di variabili casuali indipendenti $Y_1, \ldots, Y_n$, ciascuna con distribuzione appartenente alla famiglia esponenziale in forma canonica. Si supponga che tali distribuzioni siano della stessa forma e che ciascuna dipenda da un solo parametro θ_i, $i = 1, \ldots, n$ (non necessariamente di uguale valore per ogni variabile). Si ha quindi che

$$f(y_i; \theta_i) = e^{y_i b(\theta_i) + c(\theta_i) + d(y_i)}$$

con la distribuzione congiunta delle Y_i pari a

$$\begin{aligned} f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) &= \prod_{i=1}^{n} e^{y_i b(\theta_i) + c(\theta_i) + d(y_i)} \\ &= e^{\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)}. \end{aligned} \tag{4.3}$$

Un **modello lineare generalizzato** è costituito dalle seguenti tre componenti:

1. un insieme di **variabili di risposta** $Y_1, \ldots, Y_n$ con le caratteristiche sopra descritte;

2. un vettore di p **parametri** $\boldsymbol{\beta}$ ed altrettante **variabili esplicative** racchiuse nella matrice del disegno $\mathbf{X}$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

   con $p < n$;

3. una **funzione di link** monotona g tale che $g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$, dove $\mu_i = \mathbb{E}(Y_i)$ è una certa funzione di θ_i.

Dal terzo punto della definizione precedente si intuisce il legame tra il vettore dei parametri $\boldsymbol{\beta}$ e la distribuzione di probabilità $f(y_i; \theta_i)$: definendo infatti la funzione inversa $\eta = g^{-1}$ si ha che $\mu_i = \eta(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$. A sua volta μ_i è legato a θ_i da relazioni che dipendono dalla specifica distribuzione con cui si sta lavorando: ecco come, con le opportune sostituzioni matematiche, il vettore $\boldsymbol{\beta}$ entra nella determinazione di f. Questo ragionamento costituisce il punto di partenza per ricavare dal campione le stime numeriche dei parametri incogniti contenuti in $\boldsymbol{\beta}$.

4.2 Principali distribuzioni appartenenti alla famiglia esponenziale

A questo punto è necessario mostrare che le principali distribuzioni di probabilità appartengono alla famiglia esponenziale in forma canonica, e che sono pertanto utilizzabili per la costruzione di modelli lineari generalizzati.

4.2.1 Il modello binomiale

Verificare che la distribuzione binomiale è un membro della famiglia esponenziale comporta lo svolgimento di banali passaggi algebrici a partire dalla sua ben nota funzione di massa di probabilità

$$\begin{aligned} f(y;\theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\ &= e^{\ln\left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right]} \\ &= e^{\left[y\ln(\theta)-y\ln(1-\theta)+n\ln(1-\theta)+\ln\binom{n}{y}\right]} \\ &= e^{\left[y\ln(\frac{\theta}{1-\theta})+n\ln(1-\theta)+\ln\binom{n}{y}\right]}. \end{aligned} \tag{4.4}$$

Basta ora confrontare la (4.4) con la (4.2) per evincere che:

$$\begin{cases} a(y) &= y & \textbf{forma canonica} \\ b(\theta) &= \ln\left(\frac{\theta}{1-\theta}\right) & \textbf{parametro naturale} \\ c(\theta) &= n\ln(1-\theta) & \\ d(y) &= \ln\binom{n}{y}. & \end{cases} \tag{4.5}$$

Si osservi che in questo contesto il parametro di interesse è la probabilità di successo θ, mentre il numero delle prove n è quello di distrubo. Un caso particolare è costituito dalla variabile casuale binaria o di **Bernoulli**, che probabilizza il verificarsi di un successo o di un insuccesso su un'unica prova $(n = 1)$.

4.2.2 Il modello di Poisson

Ragionamenti analoghi possono essere svolti per la variabile casuale di Poisson, in cui

$$\begin{aligned} f(y;\theta) &= \frac{\theta^y e^{-\theta}}{y!} \\ &= e^{\ln\left[\frac{\theta^y e^{-\theta}}{y!}\right]} \\ &= e^{\left[y\ln(\theta)+\ln(e^{-\theta})-\ln y!\right]} \\ &= e^{[y\ln(\theta)-\theta-\ln(y!)]}. \end{aligned} \tag{4.6}$$

In questo caso:

$$\begin{cases} a(y) &= y & \textbf{forma canonica} \\ b(\theta) &= \ln(\theta) & \textbf{parametro naturale} \\ c(\theta) &= -\theta \\ d(y) &= -\ln(y!). \end{cases} \tag{4.7}$$

4.2.3 Il modello normale

Per quanto riguarda la distribuzione normale, è sufficiente osservare che

$$\begin{aligned} f(y;\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{1}{2\sigma^2}(y-\theta)^2\right]} \\ &= e^{\ln\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{1}{2\sigma^2}(y-\theta)^2\right]}\right\}} \\ &= e^{\left[-\ln(\sqrt{2\pi\sigma^2})-\frac{1}{2\sigma^2}(y^2-2y\theta+\theta^2)\right]} \\ &= e^{\left[y\frac{\theta}{\sigma^2}-\frac{\theta^2}{2\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}\right]} \end{aligned} \tag{4.8}$$

e conseguentemente

$$\begin{cases} a(y) &= y & \textbf{forma canonica} \\ b(\theta) &= \frac{\theta}{\sigma^2} & \textbf{parametro naturale} \\ c(\theta) &= -\frac{\theta^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \\ d(y) &= -\frac{y^2}{2\sigma^2}. \end{cases} \tag{4.9}$$

Si noti che il termine $-\frac{1}{2}\ln(2\pi\sigma^2)$ presente in $c(\theta)$ avrebbe potuto allo stesso modo essere inserito in $d(y)$, essendo unicamente funzione del parametro di disturbo σ^2.

4.2.4 Il modello multinomiale

La variabile casuale multinomiale consiste in un'estensione del caso binomiale per un generico numero J di categorie. Non si prende dunque più in considerazione solamente il numero di successi y in n prove indipendenti, bensì il vettore $\boldsymbol{y} = [y_1, y_2, \ldots, y_J]^{\mathrm{T}}$ con $\sum_{j=1}^{J} y_j = n$. La funzione di massa di probabilità dipende dal vettore $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_J]^{\mathrm{T}}$, dove $\sum_{j=1}^{J} \theta_j = 1$, e assume la seguente forma:

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{n!}{y_1!\, y_2! \cdots y_J!}\, \theta_1^{y_1}\, \theta_2^{y_2} \ \ldots\ \theta_J^{y_J}. \tag{4.10}$$

Quando $J = 2$ è facile ricondursi al caso binomiale. Nonostante ciò, non è possibile riscrivere l'equazione (4.10) come membro della famiglia esponenziale. L'uso dei modelli lineari generalizzati è giustificato quindi da una relazione esistente tra la distribuzione multinomiale e la variabile casuale di Poisson, la quale invece, come verificato in precedenza, ne fa parte.

Per comprendere tale relazione si considerino J variabili aleatorie indipendenti $Y_1, Y_2, \ldots Y_J$, ciascuna con distribuzione Poisson di parametro λ_j. La congiunta, posto $\boldsymbol{y} = [y_1, y_2, \ldots, y_J]^{\mathrm{T}}$, è pari al prodotto delle marginali:

$$f(\boldsymbol{y}) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}.$$

Un noto risultato di calcolo delle probabilità afferma che la variabile casuale somma

$$n = \sum_{j=1}^{J} Y_j$$

si distribuisce anch'essa come una Poisson con parametro $\sum_{j=1}^{J} \lambda_j$. A questo punto la distribuzione di $\boldsymbol{y}$ condizionata ad n diventa

$$\begin{aligned} f(\boldsymbol{y}/n) &= \frac{\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}}{\frac{\left(\sum_{j=1}^{J} \lambda_j\right)^n e^{-\sum_{j=1}^{J} \lambda_j}}{n!}} \\ &= \frac{n!}{y_1!\, y_2! \cdots y_J!} \left(\frac{\lambda_1}{\sum_{j=1}^{J} \lambda_j}\right)^{y_1} \cdots \left(\frac{\lambda_J}{\sum_{j=1}^{J} \lambda_j}\right)^{y_J}. \end{aligned} \tag{4.11}$$

Ponendo $\pi_j = \frac{\lambda_j}{\sum_{j=1}^{J} \lambda_j}$ si ha che $\sum_{j=1}^{J} \pi_j = 1$ e si osserva che la (4.11) coincide con la (4.10). La multinomiale corrisponde pertanto alla distribuzione

congiunta di J variabili di Poisson, condizionatamente alla conoscenza del loro totale n. Ciò non sorprende se si pensa che, noto il totale delle replicazioni indipendenti dell'esperimento n, il numero di volte in cui si osserva ciascuna delle J categorie è nient'altro che un conteggio.

4.3 Stimatori di massima verosimiglianza

Definizione 3 (**Verosimiglianza**). Sia $\boldsymbol{y} = [Y_1, \ldots, Y_n]^{\mathrm{T}}$ un vettore casuale con densità di probabilità congiunta $f(\boldsymbol{y}; \theta)$ dipendente da un vettore di parametri $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^{\mathrm{T}}$. La **funzione di verosimiglianza** $L(\boldsymbol{\theta}; \boldsymbol{y})$ corrisponde algebricamente alla funzione di densità congiunta $f(\boldsymbol{y}; \theta)$.

Il cambio di notazione sposta l'interesse dal vettore casuale $\boldsymbol{y}$, con $\boldsymbol{\theta}$ fissato, al vettore dei parametri $\boldsymbol{\theta}$ con $\boldsymbol{y}$ fissato (ossia dopo l'estrazione del campione). Lo **stimatore di massima verosimiglianza** per $\boldsymbol{\theta}$ è quel valore $\hat{\boldsymbol{\theta}}$ per cui

$$L(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) \geq L(\boldsymbol{\theta}; \boldsymbol{y}) \qquad \forall\, \boldsymbol{\theta} \in \Omega$$

dove Ω rappresenta lo spazio parametrico. Spesso per motivi algebrici si preferisce non lavorare direttamente con la verosimiglianza, ma con il suo logaritmo: è utile quindi definire la funzione di log-verosimiglianza $l(\boldsymbol{\theta}; \boldsymbol{y})$ semplicemente come

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \ln L(\boldsymbol{\theta}; \boldsymbol{y}).$$

Poichè il logaritmo è una funzione monotona, la massimizzazione rispetto all'argomento $\boldsymbol{\theta}$ di $l(\boldsymbol{\theta}; \boldsymbol{y})$ o di $L(\boldsymbol{\theta}; \boldsymbol{y})$ conduce allo stesso risultato.

4.3.1 L'algoritmo di Newton-Raphson

Nel contesto dei modelli lineari generalizzati solo in rari casi è possibile pervenire alla soluzione di questo problema di massimo per via analitica, cioè mediante il calcolo delle derivate; la maggior parte delle volte è necessario ricorrere all'impiego di algoritmi numerici di tipo iterativo, il più noto dei quali è quello di Newton-Raphson. Esso individua gli zeri di una qualsiasi funzione $t(x)$ a partire dalla sua derivata nel punto $x^{(m-1)}$, espressa come

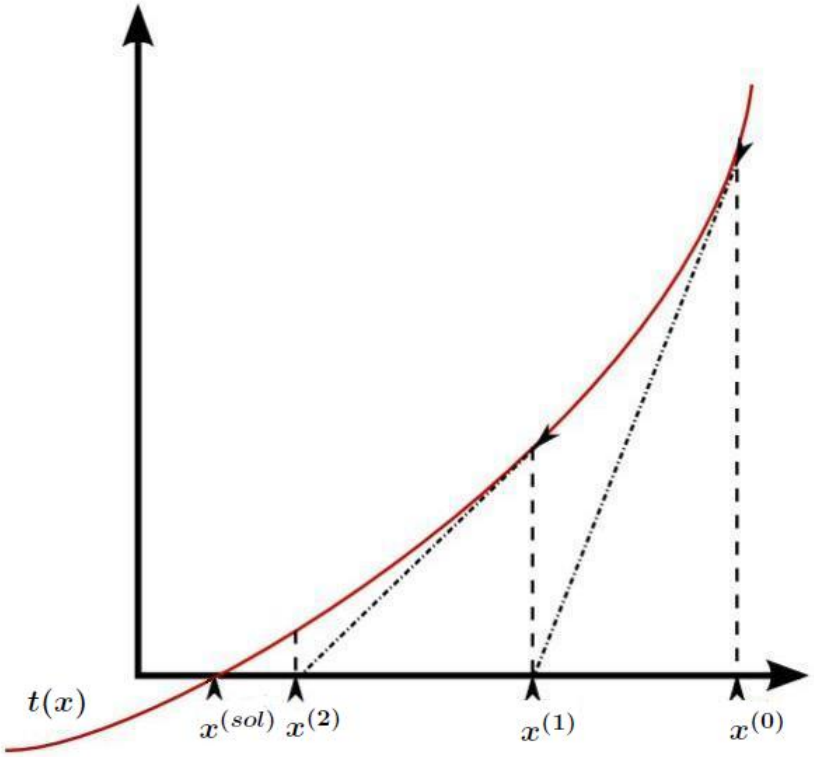

Figura 4.1: funzionamento dell'algoritmo di Newton-Raphson

rapporto incrementale

$$t'\left(x^{(m-1)}\right) = \frac{t\left(x^{(m)}\right) - t\left(x^{(m-1)}\right)}{x^{(m)} - x^{(m-1)}} \tag{4.12}$$

a patto che l'incremento $x^{(m)} - x^{(m-1)}$ sia sufficiemente piccolo. Se $x^{(m)}$ fosse la soluzione cercata, cioè tale che $t\left(x^{(m)}\right) = 0$, si potrebbe riscrivere la (4.12) come

$$x^{(m)} = x^{(m-1)} - \frac{t\left(x^{(m-1)}\right)}{t'\left(x^{(m-1)}\right)}. \tag{4.13}$$

Per trovare la soluzione è possibile calcolare iterativamente l'equazione (4.13) a partire da un valore iniziale $x^{(0)}$ finchè non viene raggiunta la convergenza, ovvero finchè il valore di $x^{(m)}$ non muta, da un iterazione all'altra, meno di un certo margine di tolleranza arbitrariamente piccolo.

Per la massimizzazione di $l(\boldsymbol{\theta}; \boldsymbol{y})$ sarà pertanto sufficiente applicare l'algoritmo alla sua derivata prima. Nel caso uniparametrico si avrà quindi

$$U = \frac{\partial l(\theta; \boldsymbol{y})}{\partial \theta}$$

da cui

$$\theta^{(m)} = \theta^{(m-1)} - \frac{U^{(m-1)}}{U'^{(m-1)}}$$

anche se spesso si è soliti sostituire la quantità U' con il suo valore atteso $\mathbb{E}(U') = -\mathcal{I}$, dove $\mathcal{I}$ rappresenta l'**informazione di Fisher**, ottenendo

$$\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{\mathcal{I}^{(m-1)}}.$$

L'algoritmo con questa variante prende il nome di **Fisher Scoring**.

L'estensione al caso multiparametrico per la stima del vettore $\boldsymbol{\beta}$ è immediata: anche in questo caso si parte da un vettore iniziale $\boldsymbol{b}^{(0)}$ da aggiornare secondo la formula iterativa

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} \boldsymbol{U}^{(m-1)} \tag{4.14}$$

in cui $\mathcal{I}$ è la **matrice di informazione attesa** e $\boldsymbol{U}$ è il vettore delle derivate parziali della log-verosimiglianza rispetto a $\boldsymbol{\beta}$, che entra nella sua determinazione in maniera indiretta come illustrato nel paragrafo 4.1 a pagina 28. Importanti proprietà della famiglia esponenziale garantiscono, sotto condizioni generali, l'esistenza e l'unicità delle stime di massima verosimiglianza.[1]

4.4 Regressione logistica ordinale

Quando la variabile di risposta Y è di tipo categorico con un naturale ordinamento tra le J categorie presenti si è soliti adottare un'estensione della regressione multinomiale logistica. Concettualmente può essere utile immaginare l'esistenza di una variabile latente continua Z su cui stabilire $J-1$ soglie ordinate $(C_1, \ldots, C_{J-1})$.

[1] *Cfr.* [3].

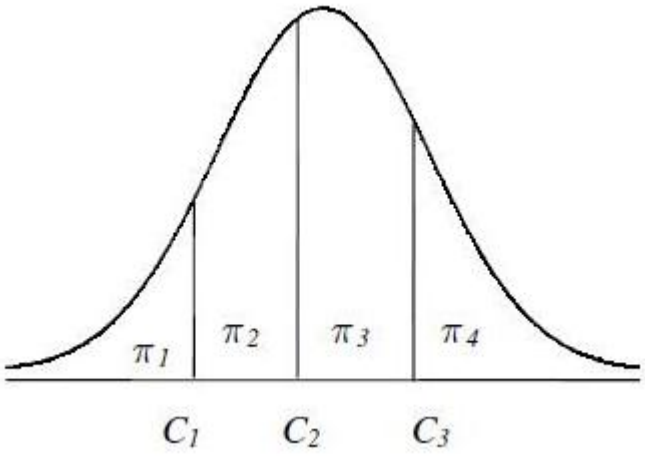

Figura 4.2: esempio di variabile latente Z

La loro determinazione definisce mediante l'equazione

$$\mathbb{P}(Y \leq j) = \mathbb{P}(Z \leq C_j) = \pi_1 + \pi_2 + \cdots + \pi_j \qquad j = 1, \ldots, J-1$$

le relative probabilità $\pi_1, \ldots, \pi_J$, dove π_J è calcolata per differenza in virtù del vincolo $\sum_{j=1}^{J} \pi_j = 1$. La figura 4.2 riporta un esempio con tre soglie, che danno origine a quattro categorie.

4.4.1 *Cumulative logit* e *proportional odds model*

In letteratura sono stati proposti numerosi modelli per variabili di risposta ordinali, il più comune dei quali è rappresentato dal ***cumulative logit model***. In esso vengono definite $J-1$ equazioni con la seguente forma:

$$logit(\mathbb{P}(Y_i \leq j)) = \ln\left(\frac{\mathbb{P}(Z_i \leq C_j)}{\mathbb{P}(Z_i > C_j)}\right) = C_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j. \tag{4.15}$$

Il segno negativo tra la soglia C_j ed il predittore lineare $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j$ serve per un'interpretazione più intuitiva dei parametri: in questo modo infatti per un coefficiente positivo un incremento del valore numerico della corrispondente covariata è associato ad un accrescimento delle probabilità relative alle categorie più alte.

Spesso si opta per un modello più parsimonioso in cui i parametri del predittore lineare non variano con le categorie e che prende il nome di ***pro-***

portional odds model. L'origine di tale denominazione deriva da un'interessante proprietà che viene a crearsi: facendo riferimento ad un generico regressore X e al suo coefficiente β_X è infatti facile verificare che la quantità

$$e^{\beta_X} = \frac{odds(\mathbb{P}(Y_i > j \mid X_i = x+1))}{odds(\mathbb{P}(Y_i > j \mid X_i = x))} \tag{4.16}$$

corrisponde, come si può evincere dalla (4.16), al rapporto tra *odds* per due unità che differiscono per un incremento unitario di X. Assumere che i parametri in $\boldsymbol{\beta}$ rimangano invariati passando da un'equazione all'altra implica che tale rapporto rimanga costante e indipendente dalla categoria considerata, originando così la proporzionalità tra gli *odds*. L'equazione (4.15) prende la forma

$$\ln\left(\frac{\mathbb{P}(Z_i \leq C_j)}{\mathbb{P}(Z_i > C_j)}\right) = C_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}. \tag{4.17}$$

Si ha così a che fare con un unico vettore di parametri affiancato dalle stime delle soglie, le quali assumono il ruolo di intercetta per ciascuna delle $J-1$ equazioni.

4.4.2 Modelli per la probabilità condizionata

In determinati contesti può essere interessante modellare delle probabilità condizionate: le formulazioni alternative della regressione logistica ordinale sono volte proprio a soddisfare questa esigenza. Le due versioni principali sono l'***adjacent categories logit model***, che contempla la possibilità di trovarsi in una categoria oppure al massimo in quella successiva

$$\mathbb{P}(Y_i = j \mid Y_i \in \{j, j+1\}) = \frac{\pi_j}{\pi_j + \pi_{j+1}}$$

e il ***continuation ratio logit model***, che invece estende il campo a tutte le categorie superiori:

$$\mathbb{P}(Y_i = j \mid Y_i \geq j) = \frac{\pi_j}{\pi_j + \cdots + \pi_J}.$$

Le equazioni dei modelli appena introdotti sono espresse rispettivamente da

$$logit\left(\frac{\pi_j}{\pi_j + \pi_{j+1}}\right) = \ln\left(\frac{\pi_j}{\pi_{j+1}}\right) = C_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j \tag{4.18}$$

per l'*adjacent categories* e

$$logit\left(\frac{\pi_j}{\pi_j + \cdots + \pi_J}\right) = \ln\left(\frac{\pi_j}{\pi_{j+1} + \cdots + \pi_J}\right) = C_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j. \qquad (4.19)$$

per l'altro.[2] Anche in questi casi esistono le varianti con pendenze comuni a tutte le categorie, in cui il predittore lineare assume la generica forma $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

In questo lavoro è stato preso in considerazione unicamente il *proportional odds model*, la cui implementazione è disponibile nei principali pacchetti statistici.

4.4.3 Diagnostica

Come accade per la regressione logistica multinomiale, le principali misure diagnostiche sono basate sui confronti con il modello nullo e il modello saturo. Le comparazioni con quest'ultimo sono effettuabili quando le covariate formano dei gruppi ed è quindi possibile raggruppare i dati in *pattern* di una certa numerosità. In tal caso si può procedere al calcolo delle frequenze attese, ottenute moltiplicando le probabilità stimate dal modello per la numerosità di ciascun *pattern*. Indicando esse con e_i, i **residui di Pearson** vengono definiti come

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}} \qquad i = 1, \ldots, N$$

dove o_i rappresentano le frequenze osservate ed N è pari a J volte il numero di *pattern* distinti delle covariate. La valutazione di tali residui avviene tramite la statistica

$$X^2 = \sum_{i=1}^{N} r_i^2.$$

Un'altra quantità di interesse è rappresentata dalla **devianza**

$$D = 2[l(\boldsymbol{\beta}_{\max}) - l(\boldsymbol{\beta})]$$

dove $l(\boldsymbol{\beta}_{\max})$ è la log-verosimiglianza del modello saturo e $l(\boldsymbol{\beta})$ quella del modello sotto esame. Tenendo presente che nel caso di p variabili vengono stimati $k = p + J - 1$ parametri, sotto l'ipotesi di buon adattamento del

[2] Per verificare l'uguaglianza tra i primi due membri della (4.18) e della (4.19) è sufficiente considerare il fatto che $logit(x) = \ln(x/(1-x))$.

modello sia D che X^2 seguono una distribuzione asintotica di tipo χ^2 con $N - k$ gradi di libertà; l'ipotesi nulla sarà rifiutata per valori alti di tali statistiche.

Quando nell'analisi sono presenti dei regressori continui o la numerosità di qualche *pattern* non è elevata è preferibile operare confronti con il modello nullo mediante la statistica

$$C = 2[l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_{\min})]$$

con $l(\boldsymbol{\beta}_{\min})$ a rappresentare la log-verosimiglianza del modello nullo, e la quantità

$$Pseudo\ R^2 = \frac{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_{\min})}{l(\boldsymbol{\beta}_{\min})}.$$

Facendo sempre riferimento ad una situazione con p covariate, la statistica C si distribuisce asintoticamente come un χ^2 con p gradi di libertà. Il modello nullo infatti presenta solo i $J - 1$ parametri relativi alle soglie, esattamente p in meno di quello corrente. In questo caso valori alti della statistica C in relazione alla distribuzione $\chi^2(p)$ denotano la significatività globale dei regressori inseriti. Si osservi che la statistica C può essere impiegata anche per comparare due modelli qualsiasi, purché innestati l'uno nell'altro. I gradi di libertà della distribuzione χ^2 in tal caso saranno pari al numero di parametri di differenza tra i due modelli.

Capitolo 5

Risultati e sviluppi futuri

Dopo aver terminato le procedure di imputazione ed aver presentato dal punto di vista teorico le metodologie statistiche impiegate, è il momento di procedere alla diretta applicazione delle medesime sull'archivio a disposizione. Nasce in questa fase la necessità di individuare nell'elenco completo dei 30 indicatori quelli con maggiore capacità esplicativa, al fine di stimare modelli più parsimoniosi e meglio interpretabili.

Per selezionare le variabili si ricorre inizialmente alla *stepwise regression*, il cui impiego è stato ampiamente dibatutto in letteratura. Per tale motivo il paragrafo 5.1.1 ne descrive il funzionamento illustrando le principali critiche ad essa mosse e, senza entrare nel dettaglio, uno degli approcci alternativi.

In questo lavoro si è deciso di affiancare alla procedura *stepwise*, come ulteriore riscontro, un'analisi delle componenti principali. È però noto come la semplice applicazione delle due metodologie non possa condurre direttamente alla formulazione del modello definitivo. In virtù di tali considerazioni, ad esse è stato affidato il compito di costruire di un gruppo di *variabili-base*, partendo dal quale sono stati stimati molti modelli diversi includendo anche i rimanenti indicatori.[1]

Dopo numerose prove empiriche, accompagnate dalle relative valutazioni diagnostiche, hanno preso forma le equazioni finali per il modello lineare nella

[1] Se la teoria macroeconomica ed il giudizio degli esperti suggeriscono la rilevanza di una certa variabile non è opportuno scartarla *a priori* basandosi unicamente sulle procedure di selezione automatica del modello, ma è doveroso tentare il suo inserimento e verificarne gli effetti.

trasformata *logit* e per la regressione ordinale *proportional odds*, riportate nella sezione 5.2.

Il capitolo si chiude descrivendo due dei possibili sviluppi futuri, volti a migliorare alcuni aspetti dell'analisi.

5.1 Selezione delle variabili

5.1.1 *Stepwise regression*

La *stepwise regression* è un algoritmo per la selezione del modello con lo scopo di individuare, partendo da un archivio specificato, la combinazione di variabili che minimizzi alcune misure preposte, denominate *criteri di informazione*. Tra essi i più ricorrenti sono il **criterio di Akaike** (***AIC***), il ***Bayesian Information Criterion*** (***BIC***) e l'**indice di Hannan-Quinn** (***HQ***), definiti rispettivamente come

$$AIC = -2\,\frac{l(\hat{\boldsymbol{\theta}};\boldsymbol{y})}{n} + 2\,\frac{p}{n}$$
$$BIC = -2\,\frac{l(\hat{\boldsymbol{\theta}};\boldsymbol{y})}{n} + \ln(n)\,\frac{p}{n}$$
$$HQ = -2\,\frac{l(\hat{\boldsymbol{\theta}};\boldsymbol{y})}{n} + 2\ln(\ln(n))\,\frac{p}{n}$$

dove $l(\hat{\boldsymbol{\theta}};\boldsymbol{y})$ è la log-verosimiglianza massimizzata, p è il numero di parametri del modello ed n è la numerosità campionaria. Le differenze tra i tre indici sono determinate univocamente dal termine moltiplicato a p/n, che rappresenta la penalizzazione inflitta ai modelli meno parsimoniosi e che contrasta la scelta di stimare un eccessivo numero di parametri. L'inserimento di una covariata infatti, anche quando concettualmente scorretto o ingiustificato, comporta in ogni caso un aumento del valore numerico di $l(\hat{\boldsymbol{\theta}};\boldsymbol{y})$.

L'algoritmo considera in primo luogo il modello comprendente tutti i regressori a disposizione, escludendone uno per volta in accordo al criterio di informazione prescelto. Sarà definitivamente accantonata la variabile dalla cui eliminazione si ricava l'indice minore. Si ottiene così un nuovo modello, su cui ripetere la procedura fino a quando non è più possibile, mediante eliminazioni sequenziali, diminuire ulteriormente il valore degli indici.

Oltre alla riduzione del numero di covariate appena descritta (ottica ***backward***) è possibile anche operare nella direzione opposta, partendo cioè dal modello nullo ed aggiungendo un regressore per volta (ottica ***forward***).

In letteratura il problema della selezione del modello viene trattato a fondo: uno degli approcci più interessanti dal punto di vista econometrico è quello di Hendry, il quale propone una propria formalizzazione nota come **teoria della riduzione**.[2] L'obiettivo di base consiste nel tentare di ridurre il numero di covariate del modello di partenza, detto ***general unrestricted model (GUM)***. Si notano dunque diverse analogie con l'ottica *backward*, a cui viene però imputata la mancata realizzazione di test diagnostici ad ogni passo e l'esplorazione di un unico *tracciato* o *cammino*. Nella *stepwise regression* infatti, poichè le eliminazioni (così come le inclusioni) sono definitive, gli effetti delle operazioni nelle fasi iniziali si ripercuotono fino al modello finale e ad ogni passo lo scenario è condizionato da quanto accaduto in precedenza.

Per Hendry invece è necessaria, oltre alla validazione diagnostica del GUM e dei modelli intermedi per escludere eventuali cattive specificazioni, l'analisi di più tracciati.[3] Ciascun cammino conduce al proprio modello finale: l'autore suggerisce di definire un nuovo GUM costituito dall'unione dei regressori di tali modelli. Il processo va ripetuto finchè due *step* successivi non conducono al medesimo GUM, punto in cui non sono possibili ulteriori riduzioni ed è necessario scegliere basandosi sui criteri di informazione.

In questa sede ci si limita ad applicare l'algoritmo *stepwise* al modello lineare nella trasformata *logit*

$$logit(PD_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

riportando l'*output* per entrambe le direzioni (*backward* e *forward*). Per associare le variabili alle corrispondenti etichette è possibile fare riferimento all'appendice A.1. Il criterio di informazione utilizzato è l'AIC.

[2] *Cfr.* [18].

[3] Si parla per questo motivo di *multiple search paths*.

```
Backward direction:
Residuals:
     Min       1Q   Median       3Q      Max
-1.73708 -0.33720 -0.07863  0.33380  2.62444

            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -6.5995     0.3075 -21.4650   0.0000
DGDPvol10     0.1294     0.0315   4.1070   0.0001
spreadCDS     0.0036     0.0003  14.3304   0.0000
DGDP          0.0862     0.0228   3.7882   0.0003
DCPI         -0.0648     0.0213  -3.0494   0.0030
TDRA         -0.0095     0.0061  -1.5694   0.1201
AGDP          0.0002     0.0001   3.2723   0.0015
YPCP         -0.0065     0.0014  -4.6459   0.0000
PSBR          0.0346     0.0140   2.4687   0.0154
MIMP          0.0013     0.0006   2.2981   0.0239
GRAPCOM      -0.0033     0.0017  -1.9844   0.0503
RAT3          0.1384     0.0308   4.4991   0.0000
ATTNETTE      0.0028     0.0015   1.8308   0.0704
DEBEST        0.0182     0.0072   2.5115   0.0138
TSTD         -0.0116     0.0021  -5.5912   0.0000
TSPD          0.0127     0.0068   1.8684   0.0650

Residual standard error: 0.7206 on 90 DF. Multiple R-squared: 0.9133.
Adjusted R-squared: 0.8989. F-statistic: 63.23 on 15 and 90 DF,  p-value: < 2.2e-16.

Forward direction:
Residuals:
     Min       1Q   Median       3Q      Max
-1.79823 -0.42376 -0.08195  0.33188  2.73550

            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -6.5335     0.3368 -19.4003   0.0000
spreadCDS     0.0036     0.0002  14.7293   0.0000
YPCP         -0.0070     0.0012  -5.9334   0.0000
RAT3          0.1488     0.0324   4.5874   0.0000
TSTD         -0.0096     0.0019  -4.9469   0.0000
DEBEST        0.0202     0.0064   3.1702   0.0021
DGDP          0.0827     0.0221   3.7438   0.0003
GRAPCOM      -0.0054     0.0017  -3.1699   0.0021
TDPY          0.0057     0.0026   2.1522   0.0340
DGDPvol10     0.1159     0.0311   3.7320   0.0003
DCPI         -0.0666     0.0213  -3.1323   0.0023
ATTNETTE      0.0035     0.0011   3.2949   0.0014
PSBR          0.0246     0.0152   1.6208   0.1084

Residual standard error: 0.7271 on 93 DF. Multiple R-squared: 0.9088.
Adjusted R-squared: 0.897. F-statistic: 77.23 on 12 and 93 DF,  p-value: < 2.2e-16.
```

Le variabili più rilevanti ai fini dell'analisi sono quelle che compaiono in entrambi gli *output*. Analizzando i risultati si osservano:

- il PIL procapite (YPCP);
- il *credit spread* (spreadCDS);
- il grado di apertura commerciale (GRAPCOM);
- il tasso di interesse sul mercato monetario (RAT3);
- il debito totale di breve periodo (TSTD);
- la volatilità del PIL (DGDPvol10).

Il debito totale di breve perodo (TSTD) ha in ambedue i casi un coefficiente negativo, in disaccordo con la teoria macroeconomica: ciò conferma che i modelli ottenuti con la metodologia *stepwise* non sono definitivi, ma spesso è necessario investigare ulteriormente per eliminare problemi di questo tipo.

5.1.2 Analisi delle componenti principali

Un ulteriore metodo per individuare le variabili più rilevanti è, come anticipato, l'analisi delle componenti principali (ACP). Nel data-set esaminato gli indicatori hanno unità di misura e ordini di grandezza differenti, quindi è preferibile operare sulla matrice di correlazione anzichè su quella di covarianza.

Il *barplot* in figura 5.1 riporta le varianze delle componenti, evidenziando il ruolo di primo piano rivestito dalle prime tre: esse infatti spiegano da sole quasi il 50% della variabilità totale. Una delle regole empiriche più utilizzate consiglia di limitare il campo alle componenti con deviazione standard maggiore di 1: in questo caso, come riportato in tabella 5.1 a pagina 45, sono le prime nove a presentare questa caratteristica. La proporzione di varianza da esse spiegata, di poco inferiore all'80%, è ritenuta soddisfacente.

La determinazione delle variabili più influenti passa attraverso il calcolo delle loro correlazioni con le componenti prese in esame. In appendice A.2 è presente la matrice con tutti i coefficienti.

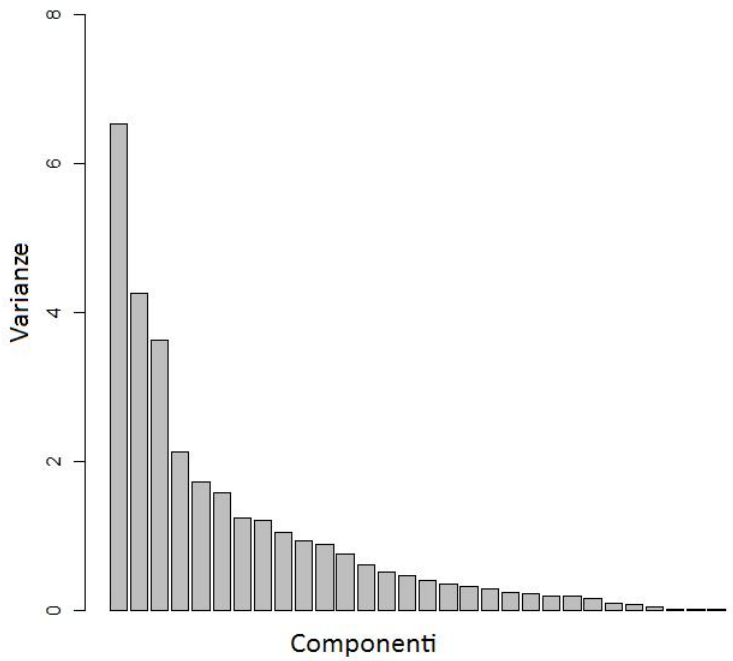

Figura 5.1: *barplot* per le componenti principali

5.2 Modelli finali

L'analisi delle componenti principali conferma l'importanza delle variabili individuate con la regressione *stepwise* e ne segnala altre. Vengono inoltre inseriti nel gruppo-base anche gli indicatori con un importante significato economico. L'elenco completo include:

- il *credit spread* (spreadCDS);
- il PIL procapite (YPCP);
- la volatilità a 10 anni del PIL (DGDPvol10);
- il grado di apertura commerciale (GRAPCOM);
- il tasso di interesse sul mercato monetario (RAT3);
- il debito totale di breve periodo (TSTD);
- il saldo di bilancia commerciale sul PIL (TDRA);

Tabella 5.1: varianza spiegata dalle prime nove componenti

Componente	**Dev.Std.**	**Prop.varianza**	**Cum.Prop.**
1	2.555	0.218	0.218
2	2.062	0.142	0.359
3	1.902	0.121	**0.480**
4	1.458	0.071	0.551
5	1.314	0.058	0.608
6	1.257	0.053	0.661
7	1.115	0.041	0.703
8	1.097	0.040	0.743
9	1.020	0.035	**0.777**

- il saldo di bilancio pubblico sul PIL (PSBR);
- il tasso di cambio effettivo reale (Tcambio);
- lo *stock* di debito pubblico sul PIL (PUDP).

A partire da questo gruppo sono state elaborate varie soluzioni, garantendo la coerenza tra i segni dei coefficienti e le teorie macroeconomiche e mantenendo in ogni caso le variabili con un rilevante significato economico o finanziario, anche quando non significative. Si è giunti così alle formulazioni definitive riportate di seguito.

5.2.1 Regressione lineare per la trasformata *logit*

Per tale modello l'equazione finale contiene sette indicatori. Il grado di apertura commerciale e il credito interno sono mantenuti nonstante presentino coefficienti statisticamente non significativi (con *p-value* rispettivamente pari a 0.37 e 0.49) a causa dell'importanza concettuale che rivestono. Il segno di tutti i coefficienti è quello atteso. L'adattamento, misurato tramite l'indice R^2, sfiora l'85% e il test-F per la significatività globale dei regressori fornisce ampia evidenza contro il modello nullo. L'*output* del modello è il seguente:

Tabella 5.2: test di normalità dei residui

Test	**Statistica-test**	**Valore**	**P-value**
Shapiro-Wilk	W	0.989	0.554
Jarque-Bera	X-squared	1.954	0.377
D'Agostino	Chi2 (Omnibus)	2.561	0.278
	Z3 (Skewness)	1.193	0.233
	Z4 (Kurtosis)	1.067	0.286

```
Residuals:
      Min        1Q    Median        3Q       Max
-2.228548 -0.641477  0.004026  0.483668  2.516137

            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -6.8639     0.3498 -19.6240   0.0000
YPCP         -0.0063     0.0011  -5.9784   0.0000
DGDPvol10     0.1011     0.0425   2.3771   0.0194
TDRA         -0.0155     0.0067  -2.3274   0.0220
RAT3          0.1473     0.0354   4.1576   0.0001
spreadCDS     0.0036     0.0003  10.6418   0.0000
SODD          0.0066     0.0074   0.8908   0.3752
GRAPCOM      -0.0011     0.0016  -0.6876   0.4933

Residual standard error: 0.8975 on 98 DF. Multiple R-squared: 0.8536.
Adjusted R-squared: 0.8431. F-statistic: 81.62 on 7 and 98 DF,  p-value: < 2.2e-16.
```

Il *qq-plot* dei residui, riportato in figura 5.2 a pagina 47, suggerisce una discreta aderenza alla distribuzione normale. L'impressione grafica è confermata da specifici test per la normalità, i cui risultati vengono riassunti dalla tabella 5.2. In tutti i casi i *p-value* elevati indicano che non c'è evidenza campionaria per rifiutare l'ipotesi nulla di provenienza da una distribuzione normale.

Calcolando la PD predetta mediante la formula

$$PD_i = \frac{e^{logit(PD_i)}}{1 + e^{logit(PD_i)}}$$

e sfruttando le relazioni presentate in tabella 2.6 a pagina 18, si ottengono le classi di *rating* stimate dal modello per ogni paese. Il confronto con i *rating* ufficiali denota che in più dell'80% dei casi le differenze si mantengono

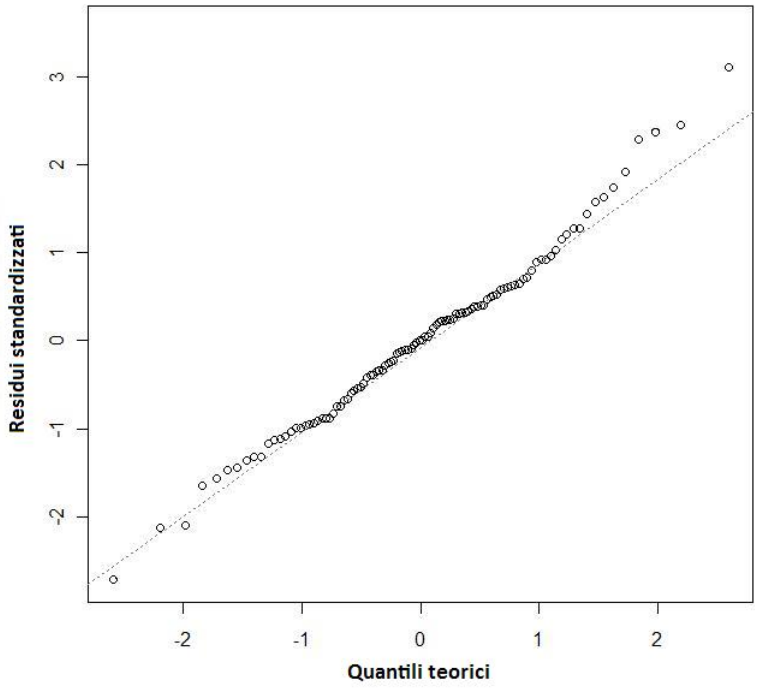

Figura 5.2: *qqplot* dei residui

entro due *notches* e possono pertanto essere imputate alla naturale variabilità del modello piuttosto che ad una sua cattiva specificazione: il *fitting* nel complesso è giudicato soddisfacente.

5.2.2 Regressione logistica ordinale

Per il modello ordinale l'insieme dei regressori è analogo: in questo caso, oltre alle due variabili citate in precedenza, anche il rapporto tra il saldo di bilancia commerciale e il PIL (TDRA) presenta un coefficiente non significativo (*p-value* 0.19); tuttavia, per i soliti motivi, si opta per mantenerlo nell'equazione. Il test del rapporto di verosimiglianza per il confronto con il modello nullo denota la globale capacità esplicativa dei regressori: il valore della statistica test, da confrontare con un $\chi^2(7)$, è infatti 199.28. Anche lo *Pseudo-*R^2 (0.523) è abbastanza elevato. L'*output* seguente riporta, oltre ai coefficienti con i relativi errori standard e alle statistiche appena commentate, anche i valori per le sei soglie.

```
Ordered logistic regression                       Number of obs   =        106
                                                  LR chi2(7)      =     199.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -90.771273                       Pseudo R2       =     0.5233

------------------------------------------------------------------------------
   punteggio |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  YPCP       |  -.0165879   .0033727    -4.92   0.000    -.0231982   -.0099776
  DGDPvol10  |   .2799031    .120182     2.33   0.020     .0443506    .5154556
  TDRA       |   -.019363   .0146904    -1.32   0.187    -.0481557    .0094297
  RAT3       |   .2146213   .0960251     2.24   0.025     .0264156     .402827
  spreadCDS  |   .0104256   .0016725     6.23   0.000     .0071475    .0137036
  SODD       |   .0213128   .0208653     1.02   0.307    -.0195823     .062208
  GRAPCOM    |  -.0046628   .0060925    -0.77   0.444     -.016604    .0072783
-------------+----------------------------------------------------------------
       /cut1 |  -3.467187   1.175874                     -5.771858   -1.162516
       /cut2 |  -1.385627   1.047547                     -3.438781     .667527
       /cut3 |   .4764592   1.004403                     -1.492135    2.445053
       /cut4 |   3.839993    1.08253                      1.718273    5.961712
       /cut5 |   7.154111   1.336793                      4.534045    9.774177
       /cut6 |   13.21562   2.093867                       9.11172    17.31953
------------------------------------------------------------------------------
```

Per ogni paese è possibile ottenere il vettore con le probabilità di appartenenza alle sette categorie stimate dal modello: il calcolo si fa per differenza a partire dalle cumulate, ricavabili con la formula

$$\mathbb{P}(Y_i \leq j) = \frac{e^{C_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{C_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}.$$

Un'ulteriore misura diagnostica è rappresentata dal confronto tra il *rating* ufficiale e quello più probabile secondo la previsione del modello: solo in meno del 5% dei casi la differenza è maggiore di una categoria.[4]

I parametri stimati serviranno per creare uno *score* quantitativo per i paesi senza *rating*: l'attribuzione specifica alle classi dovrà tenere in considerazione anche fattori qualitativi, la cui importanza è stata sottolineata nel capitolo 1.

[4] In questo caso si considerano gli scostamenti di una anzichè due categorie perchè la classificazione è meno fine rispetto al modello precedente.

5.3 Sviluppi futuri

5.3.1 Imputazione multipla

La tecnica *nearest neighbour* utilizzata nel capitolo 3 rientra, insieme ad altre più o meno sofisticate, nei **metodi di imputazione deterministica**, così chiamati perchè la loro ripetuta applicazione al medesimo insieme di dati produce sempre risultati identici. Un ulteriore esempio è il metodo della *media generale*, nel quale si sostituiscono i valori mancanti con la media aritmetica calcolata sulle sole unità per cui la misura in esame è disponibile. In altri casi vengono sfruttate delle informazioni ausiliarie per dividere l'intero insieme di unità in classi (imputando la media di classe anzichè quella generale), o per stimare dei parametri di regressione impiegati per sostituire i *missing* con una *prediction* (*imputazione con regressione*).

Per i metodi appena descritti esiste una corrispondente versione **stocastica**, caratterizzata dall'introduzione di una componente aleatoria. Nell'imputazione con media o con regressione solitamente viene aggiunto un termine residuale ϵ_i di cui va specificata la distribuzione, mentre l'equivalente della tecnica del vicino più vicino prende il nome di *metodo del donatore casuale* e prevede, all'interno delle già citate classi, la selezione casuale dei donatori ogniqualvolta occorra un dato mancante.

A differenza di quanto accade nei metodi deterministici, l'applicazione ripetuta di tecniche *random* origina risultati ogni volta differenti, ma come negli altri casi viene imputato *un unico* dato, che prende subito il posto del valore mancante: per tale ragione si parla di **imputazione stocastica singola**.

L'idea di effettuare un maggior numero di imputazioni per ogni *missing* prende il nome di **imputazione multipla** ed è stata proposta da Rubin parallelamente agli altri suoi contributi in materia.[5] Il principale vantaggio dell'imputazione multipla consiste nel poter trarre conclusioni inferenziali a partire da $m > 1$ data-set completi piuttosto che da uno solo, avendo così la possibilità di considerare l'incertezza generata dal processo di imputazione stesso.

[5] *Cfr.* [10].

Un altro vantaggio è rappresentato dall'impiego di procedure iterative che in alcuni casi garantiscono la *stabilità* dei valori imputati: a tal proposito viene di seguito presentato l'algorimo ***SRMI*** (***Sequential Regression Imputation Method***).[6] Esso prevede due fasi ed implementa una sequenza di regressioni valutando di volta in volta la tipologia di variabile dipendente.[7]

La prima fase termina con la determinazione di un archivio privo di valori mancanti e il suo funzionamento è descritto dai passi riportati nell'elenco seguente. Indicando con $\mathbf{X}$ la matrice delle variabili complete (in cui introdurre anche una colonna per l'intercetta) e con $\mathbf{Y}$ la matrice dei k indicatori $Y_1, \ldots, Y_k$ affetti dalla presenza di *missing* (ordinati considerando prima quelli con un maggior numero di valori osservati), tali *step* si ripetono ciclicamente sulle k colonne di $\mathbf{Y}$. Con riferimento ad un generico Y_j è necessario, per
$j = 1, \ldots, k$:

1. definire $\mathbf{U} = \mathbf{X}$;
2. definire $Y = Y_j$;
3. stimare, partendo dal modello $Y = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, il vettore dei parametri $\boldsymbol{B} = (\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}Y$ sulla base delle osservazioni complete;
4. porre in RSS la somma dei residui al quadrato, in df il numero di gradi di libertà del modello e in $\mathbf{T}$ la decomposizione di Cholesky di $(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}$, per cui $\mathbf{T}^{\mathrm{T}}\mathbf{T} = (\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}$;
5. estrarre un numero casuale u da una distribuzione χ^2 con df gradi di libertà e definire $\sigma_*^2 = RSS/u$;
6. generare un vettore casuale $\boldsymbol{z}$ della stessa lunghezza di $\boldsymbol{B}$ campionando per ogni elemento da una distribuzione $N(0,1)$ ed impiegarlo per definire $\boldsymbol{\beta_*} = \boldsymbol{B} + \sigma_*\mathbf{T}\boldsymbol{z}$;

[6] *Cfr.* [19].

[7] Il data-set esaminato comprende unicamente indicatori continui, quindi non c'è ragione di abbandonare il modello normale, ma è l'algoritmo contempla anche l'utilizzo di regressioni logistiche o di Poisson in caso di risposte binarie o di conteggio *etc.*

7. indicare con $\mathbf{U}_{\text{miss}}$ la sottomatrice contenente solo le righe associate ad unità per cui Y è mancante;

8. calcolare i valori imputati $Y_* = \mathbf{U}_{\text{miss}}\boldsymbol{\beta}_* + \sigma_*\boldsymbol{v}$, dove $\boldsymbol{v}$ è di nuovo un vettore di numeri casuali normali con lunghezza pari al numero di righe di $\mathbf{U}_{\text{miss}}$, e sistemarli nelle opportune posizioni di Y;

9. aggiornare $\mathbf{U} = (\mathbf{U}, Y)$;

10. incrementare $j = j + 1$.

La seconda fase applica iterativamente questo schema ma, poichè è già disponibile un archivio completo, nelle regressioni vengono sempre coinvolte tutte le variabili e non solo quelle contenute in $\mathbf{U}$. La procedura si ripete un numero prestabilito di volte, oppure fino alla convergenza dei valori imputati.

L'algoritmo ha molte caratteristiche interessanti ma presenta anche alcuni difetti: in presenza di numerose variabili è possibile incappare, ad esempio, in problemi di multicollinearità.

5.3.2 Non-linearità

Nei modelli presentati si è scelto di non investigare su relazioni non lineari, introducendo unicamente termini di primo grado. In altri contesti, sempre legati alla stima della PD, l'inserimento di effetti quadratici o con esponente superiore può invece rivelarsi vantaggioso, sopratutto quando:

- la teoria non riesce a fornire precise indicazioni circa il segno atteso di un coefficiente;

- l'effetto marginale di un regressore sulla variabile di risposta non è costante, in valore assoluto e/o in direzione, sull'intero supporto.

Nelle situazioni appena descritte limitarsi allo studio delle sole relazioni lineari può talvolta condurre a stime di coefficienti con segno opposto a quello atteso, dai quali vengono giocoforza tratte conclusioni controintuitive.[8]

[8] Il problema si è verificato durante la fase di selezione del modello con la variabile TSTD (pagina 43), il cui coefficiente aveva segno negativo.

Il metodo classico per il trattamento delle non-linearità nei modelli, ovvero l'aggiunta dei termini polinomiali $\{x^2, x^3, \dots x^s\}$, è stato riconsiderato dai più recenti articoli in letteratura, i quali evidenziano come piccole perturbazioni delle covariate possono causare un significativo peggioramento del *fitting* delle funzioni polinomiali.[9] Una delle alternative suggerite prevede di affiancare all'insieme $\{x^2, x^3, \dots x^s\}$ gli M termini

$$\{(x - k_1)^s_+, \dots, (x - k_M)^s_+\}$$

dove

$$(x - k_i)^s_+ = \begin{cases} 0 & \text{se } x \leq k_i \\ (x - k_i)^s & \text{se } x > k_i \end{cases} \qquad i = 1, \dots, M.$$

Gli scalari $k_1, \dots, k_M$ prendono il nome di ***knots*** ed è compito del ricercatore decidere il loro numero e la loro locazione. Alcune delle soluzioni proposte in letteratura suggeriscono di trattarli come parametri incogniti da stimare insieme ai coefficienti di regressione oppure di seguire approcci basati sulla minimizzazione del BIC.

Si osservi che l'informazione necessaria per l'adozione della tecnica rimane invariata; inoltre il modello resta lineare nei parametri quindi la fase di stima può essere implementata con le usuali procedure. In sostanza, l'utilizzo di questo strumento non comporta maggiori "oneri informativi" e garantisce, a detta degli autori, maggiore accuratezza in fase previsiva oltre ad un aumento del *fitting* globale.

[9] *Cfr.* [16].

Appendice A

Tabelle riassuntive per le variabili

A.1 Etichette e unità di misura

Tabella A.1: etichette e unità di misura: variabili strutturali e monetarie

Variabile	**Etichetta**	**Udm**
PIL nominale	GDPD	bil. $
PIL reale	AGDP	bil. $
PIL procapite	YPCP	bil. $
PIL (Volatilità 10 anni)	DGDPvol10	%
Grado di concentrazione	XPP1	%
Importazioni	MIMP	bil. $
Esportazioni	MEXP	bil. $
Grado di apertura commerciale	GRAPCOM	%
PIL (Variazione percentuale)	DGDP	%
Inflazione	DCPI	%
Credito interno	SODD	%
Credito/PIL	CREDITO	%
Tasso di interesse sul mercato monetario	RAT3	%
Tasso di cambio effettivo reale	Tcambio	%
Tasso di cambio (Volatilità 10 anni)	XRREvol10	%

Tabella A.2: etichette e unità di misura: posizione debitoria

Variabile	**Etichetta**	**Udm**
Riserve internazionali	ILMA	bil. $
Debito/PIL	TDPY	%
Servizio del debito/Entrate valutarie	TSPD	%
Debito estero totale	TDBT	bil. $
Debito totale di breve periodo	TSTD	bil. $
Riserve internazionali/Debito estero a breve	RISINT	%
Variazione riserve internazionali	VARISINT	%
Attività nette/PIL	ATTNETTE	%
Stock di debito pubblico/PIL	PUDP	%
Saldo di bilancio pubblico/PIL	PSBR	%
Debito estero del settore pubblico/PIL	DEBEST	%

Tabella A.3: etichette e unità di misura: altre variabili

Variabile	**Etichetta**	**Udm**
Saldo di conto corrente/PIL	CARA	%
Bilancia dei pagamenti	BILpagam	%
Saldo di bilancia commerciale/PIL	TDRA	%
Flusso IDE in entrata/PIL	FlussoIDE	%
Riserve/Uscite correnti	MCOV	mesi
Andamento borsa valori	BORSAvalori	%
Credit spread per *CDS*	spreadCDS	bps

A.2 Correlazioni con le componenti principali

Tabella A.4: correlazioni tra le variabili e le prime nove componenti

Variabile	**Componenti**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
DGDPvol10	-0.17	0.3	0.2	**-0.58**	0.42	0.14	0.15	0.02	0.04
spreadCDS	**-0.5**	-0.18	-0.29	-0.08	0.34	-0.18	0.09	-0.18	0.44
DGDP	-0.09	**0.52**	-0.45	0.36	-0.2	0.11	-0.08	0.14	0.01
DCPI	-0.45	0.26	-0.47	-0.35	0.2	-0.08	0.04	0.09	0.16
GDPD	**0.77**	-0.14	-0.38	0.01	0.24	-0.06	-0.08	-0.09	-0.03
CARA	0.38	**0.72**	0.45	-0.12	0.07	-0.07	0.1	-0.01	0.07
TDRA	0.3	**0.71**	0.28	-0.21	0.07	-0.01	0.12	-0.15	0.14
BILpagam	0.21	**0.78**	0.36	-0.08	0.14	0.24	0.15	0	0.05
AGDP	**0.78**	-0.06	-0.47	0.03	0.23	-0.01	-0.1	-0.02	-0.03
YPCP	**0.54**	-0.07	**0.59**	0.02	0.02	-0.27	0.06	0.28	0.09
PSBR	0.09	**0.58**	0.28	0.06	-0.21	0.07	-0.39	0.17	0.14
SODD	-0.36	0.32	-0.49	-0.17	0.19	-0.18	-0.1	0.27	-0.01
CREDITO	0.43	-0.42	0.34	0.03	0.05	-0.09	0.42	0.19	0
MEXP	**0.88**	-0.01	-0.3	0.07	0.22	0.06	-0.02	-0.01	-0.02
MIMP	**-0.86**	0.1	0.32	-0.07	-0.24	0	0.07	0.1	0.05
GRAPCOM	-0.09	0.13	0.41	0.23	0.32	0.41	-0.06	0	-0.37
XRREvol10	-0.23	0.34	-0.38	-0.48	0.03	-0.22	0.33	0.06	-0.25
Tcambio	0.06	0.19	-0.22	-0.07	**-0.67**	-0.06	0.38	-0.06	-0.39
FlussoIDE	-0.43	-0.06	-0.28	0.12	0.12	**0.62**	0.05	0.04	-0.07
ILMA	**0.62**	0.17	-0.4	0.1	0.11	0.23	0.05	0.28	0.03
VARISINT	0.12	0.14	-0.05	0.35	-0.38	0.23	0.2	-0.21	**0.58**
MCOV	0.15	0.47	-0.23	0.03	-0.12	0.34	0.22	0.42	0.09
RAT3	**-0.52**	0.03	**-0.57**	-0.1	-0.17	-0.04	0.01	0.14	0.09
PUDP	0.18	**-0.55**	0.01	0.16	0.14	0.11	**0.61**	0.04	0.12
ATTNETTE	0.02	-0.02	0.21	0.28	0.02	**-0.5**	-0.03	**0.59**	0.07
DEBEST	-0.46	-0.5	-0.07	0.06	0.07	0.26	0.01	0.37	0.05
TDBT	**0.67**	-0.17	-0.19	-0.39	-0.34	0.03	-0.09	0.01	0.07
TSTD	**0.83**	-0.03	-0.3	-0.19	-0.15	0.09	-0.07	0.11	0.03
TDPY	-0.06	**-0.51**	0.39	-0.46	-0.05	0.36	-0.17	0.26	0.08
TSPD	0.28	-0.44	0.15	**-0.68**	-0.31	0.16	-0.12	0	0.11

Appendice B

Codici R per gli algoritmi di imputazione

B.1 Algoritmo nearneigh

```
source("functions.R")
nearneigh <- function (miss, nomiss, p=1, elenco=row.names(nomiss), punteggio) {
  imputato <- diagn <- miss
  distanze <- as.matrix(dist(nomiss))
  if (length(distanze[distanze == 0]) != dim(imputato)[1]) {
    print ("Warning: other null distances in addition to diagonal ones")
  }
  for (j in 1:dim(miss)[2]) {
    minimicond <- tapply(miss[,j],punteggio,minn)
    massimicond <- tapply(miss[,j],punteggio,maxn)
    for (i in 1:dim(miss)[1]) {
      diagn[i,j] <- NA
      if (is.na(miss[i,j]) & row.names(miss)[i] %in% elenco) {
        vett <- sort(distanze[i,])
        k <- 0
        z <- NA
        while (length(z) < p+1) {
          minimo <- vett[2+k]
          valore <- miss[match(minimo,distanze[i,]),j]
          if (!is.na(valore)){
            z <- c(z,valore)
          }
          k <- k+1
        }
        imputato[i,j] <- meann(z)
        minimum <- minimicond[punteggio[i]]
        maximum <- massimicond[punteggio[i]]
        diagn[i,j] <- ifelse(imputato[i,j]<=maximum & imputato[i,j]>=minimum,0,1)
      }
      i <- i+1
    }
    j <- j+1
  }
  list(imputato=imputato, diagn=diagn)
}
```

B.2 Algoritmo imputaznew

```
source("functions.R")
imputaznew <- function(dati, p=1, elenco=row.names(dati), punteggio){
        daticopia <- diagn <- dati
  for (j in 1:dim(daticopia)[2]){
    minimicond <- tapply(daticopia[,j],punteggio,minn)
    massimicond <- tapply(daticopia[,j],punteggio,maxn)
    daticompl <- data.matrix(daticopia[which(!is.na(daticopia[,j])),])
    for (i in 1:dim(daticopia)[1]){
      diagn[i,j] <- sc[i,j] <- NA
      if (is.na(daticopia[i,j]) & row.names(daticopia)[i] %in% elenco){
        distmedie <- rep(NA,nrow(daticompl))
        rec <- data.matrix(daticopia)[i,]
        for (k in 1:length(distmedie)) {
          unita <- daticompl[k,]
          distmedie[k] <- meann(t(abs(rec-unita)))
          k <- k+1
        }
        if(length(distmedie[is.nan(distmedie)])>=1 | length(distmedie[is.na(distmedie)])>=1){
          print("Warning: there are invalid distances")
        }
        z <- cbind(distmedie,daticompl[,j])
        z <- z[order(distmedie,daticompl[,j]),]
        daticopia[i,j] <- imp <- mean(z[1:p,2])
        minimo <- minimicond[punteggio[i]]
        massimo <- massimicond[punteggio[i]]
        diagn[i,j] <- ifelse(imp <= massimo & imp >= minimo,0,1)
      }
      i <- i+1
    }
  j <- j+1
  }
  list(imputato=daticopia,diagn=diagn)
}
```

Bibliografia

[1] **ABI** (2008) - *Modello di Scoring per le Crisi*, ABI Country Risk Forum, documento interno.

[2] **ABI** (2008) - *ABI Country Risk Compass - Early Warning delle crisi bancarie e di liquidità nei Paesi emergenti*, ABI Country Risk Forum.

[3] **Agresti, A.** (1984) - *Analysis of Ordinal Categorical Data*, John Wiley & Sons.

[4] **Agresti, A.** (2002) - *Categorical Data Analysis*, John Wiley & Sons.

[5] **Agresti, A.** (2007) - *An Introduction to Categorical Data Analysis*, John Wiley & Sons.

[6] **Baldacci, E. & Chiampo, L.** (2007) - *L'analisi del rischio-paese: L'approccio di SACE*, SACE Working paper.

[7] **Bernè, F. & Burello, E. & Ciprian, M. & Gasparet, G. & Pediroda, V. & Robba, C.** (2004) - *Il Rischio Paese, determinazione, rilievo, applicazioni*, Gruppo di ricerca 'Rischio-Paese' - Università degli Studi di Trieste.

[8] **Bhatia, A. V.** (2002) - *Sovereign Credit Ratings Metodology: An Evaluation*, IMF Working Paper.

[9] **Borio, C. & Packer, F.** (2004) - *Analisi dei nuovi orientamenti in materia di rischio-paese*, Rassegna trimestrale BRI.

[10] **Calzolari G. & Neri L.** (2002) - *A Method of Simulated Scores for Imputation of Continuous Variables Missing at Random*, Quaderni del Dipartimento di Statistica 'Giuseppe Parenti', Firenze.

[11] **Dobson, A. J.** (2002) - *An Introduction To Generalized Linear Models*, Chapman&Hall/CRC.

[12] **Fitch Ratings** (2002) - *Fitch Sovereign Ratings - Rating Methodology*, 2002.

[13] **Fitch Ratings** (2008) - *Guide to Sovereign Credit Report.*

[14] **Fitch Ratings** (2010) - *Fitch Ratings Sovereign 2009 Transition and Default Study.*

[15] **Fitch Ratings** (2011) - *Sovereign Rating Methodology.*

[16] **Giordani, P. & Jacobson, T. & Von Schedvin, E. & Villani, M.** (2011) - *Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial Ratios*, Sveriges Riskbank Working Paper Series.

[17] **Hendry, D. F.** (1995) - *Dynamic Econometrics*, Oxford University Press.

[18] **Hendry, D. F. & Krolzig, H. M.** (2002) - *New Developments in Automatic General-to-specific Modelling*, Econometrics and the Philosophy of Economics, Princeton University Press.

[19] **Lepkowski, J. & Raghunathan, T. & Solenberger, P. & Van Hoewyk, J.** (2001) - *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*, Statistics Canada, Survey Methodology.

[20] **Little, R.** (1988) - *A Test of Missing Completely at Random for Multivariate Data With Missing Values*, Journal of the American Statistical Association.

[21] **Little, R. & Rubin, D.** (1987) - *Statistical Analysis with Missing Data*, John Wiley & Sons.

[22] **Marotta, G.** (2009) - *I rating sul rischio sovrano*, `http://www.economia.unimore.it`.

[23] **Moody's Investors Service** (2004) - *A Quantitative Model for Foreign Currency Government Bond Ratings*, Special Comment.

[24] **Moody's Investors Service** (2008) - *Sovereign Bond Ratings.*

[25] **Standard&Poor's** (2011) - *Sovereign Government Rating Methodology And Assumptions*, Standard&Poor's Global Credit Portal.

[26] **Zaninelli, M.** (2007) *Il processo di costruzione del country-rating: una breve rassegna*, `http://www.assbb.it/contenuti/file/2007_02_04.pdf`.