UNIVERSITÀ DEGLI STUDI DI PERUGIA

FACOLTÀ DI ECONOMIA

CORSO DI LAUREA IN ECONOMIA E COMMERCIO

TESI DI LAUREA

METODI STATISTICI MULTIVARIATI APPLICATI ALL'ANALISI DEL COMPORTAMENTO DEI TITOLARI DI CARTA DI CREDITO DI TIPO REVOLVING

LAUREANDA

Fulvia Pennoni

RELATRICE

Prof.ssa Elena Stanghellini

ANNO ACCADEMICO 1998/99

«*È dalla terra, dalla solidità, che deriva necessariamente un parto pieno di gioia e il sentimento paziente di un'opera che cresce, di tappe che si susseguono, aspettate con calma, con sicurezza. Occorre soffrire perché la verità non si cristallizzi in dottrina ma nasca dalla carne*».

(Emmanuel Mounier)

Ai miei amici.

INTRODUZIONE

Una valutazione più accurata e tempestiva del rischio di credito sta avvenendo negli ultimi anni grazie all'ausilio di metodologie statistiche dette *credit scoring* che si sono rivelate notevolmente efficaci per il miglioramento che apportano alla gestione complessiva del credito.

Il processo di erogazione del credito viene suddito in varie fasi: quella istruttoria che mira ad analizzare le capacità di credito del richiedente e a valutare la compatibilità della richiesta con le strategie operative dell'istituto, quella decisionale in base alla quale il prestito viene accordato o rifiutato e la fase di monitoraggio che mira all'analisi continuativa dei crediti in essere.

Le tecniche di *credit scoring* sono sistemi di valutazione automatica che si applicano a tutte le fasi della gestione del credito. Questi non sostituiscono l'esperienza e la professionalità della persona ma si è dimostrato che consentono di integrarne in modo molto efficace le capacità valutazionali.

Nel presente lavoro si considerano le operazioni di credito al consumo per le quali i sistemi di *score* si sono rivelati particolarmente validi, in quanto garantiscono una gestione più snella del credito oltre che la correttezza metodologica e la tempestività operativa. Si considera, in modo specifico, la particolare forma tecnica con cui il credito al consumo viene erogato rappresentata dalle carte di credito cosiddette *revolving*.

Le carte di credito si vanno sempre più diffondendo, in Italia come negli altri paesi industrializzati, in quanto oltre ad essere uno strumento di credito, consentono di pagare gli acquisti effettuati, prelevare contanti sia in Italia che all'estero, se collegate al circuito internazionale. Inoltre, la forma *revolving* consente di accedere ad un finanziamento a scadenza determinata, fissato inizialmente per un determinato importo, non finalizzato all'acquisto di beni

specifici. Il prestito è rotativo nel senso che dà la possibilità di rimborsare, in genere mensilmente, l'importo speso, ricostituendo vai via il plafond iniziale a disposizione. La rateizzazione del rimborso lo rende strumento molto gradito agli utenti.

Tale lavoro vuole essere un'applicazione dello *scoring* comportamentale o *behavioural scoring* che Thomas (1999) definisce come *'the systems and models that allow lenders to make better decisions in managing existing clients by forcasting their future performance'*. Questo è costituito da quelle tecniche statistiche che mirano ad essere d'ausilio nella fase di monitoraggio del credito.

Lo *scoring* comportamentale è particolarmente utile se si conduce un portafoglio *revolving* in quanto diventa un supporto molto efficace per le decisioni che in tale ambito devono essere prese. E' utile per fare delle analisi sulla performance delle carte di credito, definire il profilo dei possessori della carta, comprendere ed analizzare i fattori chiave del rapporto carta-cliente. Tutto ciò consente alla società di suddividere il portafoglio dei clienti ed attuare azioni di marketing specifiche, oltre che avere previsioni circa lo stato finale del conto per le azioni cosiddette di sviluppo, nell'obbiettivo di fidelizzare la clientela.

L'applicazione presentata in questo lavoro si basa sull'utilizzo di strumenti statistici multivariati, che consentono una migliore comprensione delle relazioni tra le variabili che descrivono il comportamento dei titolati della carta. I modelli grafici a catena, basati su un espansione log-lineare della funzione di densità delle variabili, consentono di rappresentare associazioni orientate, inerenti sottoinsiemi di variabili. Consentono, inoltre, di individuare la struttura che rappresenti in modo più parsimonioso possibile tali relazioni e modellare simultaneamente più di una variabile risposta. Sono utili quando esiste un ordinamento anche parziale tra le variabili che permette di suddividerle in meramente esogene, gruppi d'intermedie tra loro concatenati e di risposta. Nei modelli grafici la struttura d'indipendenza delle variabili viene

rappresentata visivamente attraverso un grafo. Nel grafo le variabili sono rappresentate da nodi legati da archi i quali mostrano le dipendenze in probabilità tra le variabili. La mancanza di un arco sta a significare che due nodi sono indipendenti dati gli altri nodi.

Tali modelli sono molto potenti per la teoria che li accomuna con i sistemi esperti, per cui una volta selezionato il modello è possibile interrogare il sistema esperto per modellare la distribuzione di probabilità congiunta e marginale delle variabili.

Il primo capitolo fornisce un'introduzione alle tecniche di *scoring* ponendo l'accento sul processo di evoluzione del credito al consumo, considerando in particolare le carte di credito. Vengono presentati i metodi statistici adottati nel *credit scoring* così come si evidenzino alcune problematiche riguardanti i dati.

Il secondo capitolo prende in considerazione le variabili categoriche. Le informazioni sui titolari di carta di credito sono, infatti, compendiate in tabelle di contingenza. Si introducono le nozioni d'indipendenza tra due variabili e di indipendenza condizionata tra più di due variabili. Si studiano alcune misure d'associazione tra variabili, in particolare, si introducono i rapporti di *odds* che costituiscono la base per la costruzione dei modelli multivariati utilizzati.

Nel terzo capitolo vengono illustrati i modelli log-lineari e logistici che appartengono alla famiglia dei modelli lineari generalizzati. Essendo metodi multivariati consentono di studiare l'associazione tra le variabili considerandole simultaneamente. In particolare viene descritta una speciale parametrizzazione log-lineare che permette di tener conto della scala ordinale con cui sono misurate alcune delle variabili categoriche utilizzate. Questa è utile per trovare la migliore categorizzazione delle variabili continue. I modelli logistici, a differenza dei precedenti, assumono che ci sia una variabile risposta. In particolare si considera un modello logistico cumulativo che consente d'avere una semplice interpretazione degli effetti che le variabili esplicative esercitano su una variabile risposta ordinale. Si richiamano,

inoltre, i risultati relativi alla stima di massima verosimiglianza dei parametri dei modelli, accennando anche agli algoritmi numerici iterativi necessari per la risoluzione delle equazioni di verosimiglianza rispetto ai parametri incogniti. Si fa anche riferimento al test del rapporto di verosimiglianza per valutare la bontà di adattamento del modello ai dati.

Il capitolo quarto introduce alla teoria dei grafi, esponendone i concetti principali ed evidenziando alcune proprietà che consentono la rappresentazione visiva del modello mediante il grafo, mettendone in luce i vantaggi interpretativi. In tale capitolo si accenna anche al problema derivante dalla sparsità della tabella di contingenza, quando le dimensioni sono elevate. Vengono pertanto descritti alcuni metodi adottati per far fronte a tale problema ponendo l'accento sulle definizioni di collassabilità.

Il quinto capitolo procede ad un'applicazione dei metodi descritti su un campione composto da circa sessantamila titolari di carta di credito *revolving*, rilasciata da una delle maggiori società finanziarie italiane operanti nel settore. Le variabili prese in esame sono quelle descriventi le caratteristiche socioeconomiche del titolare della carta, desumibili dal modulo che il cliente compila alla richiesta di finanziamento e lo stato del conto del cliente in due periodi successivi. Ogni mese, infatti, i clienti vengono classificati dalla società in: 'attivi', 'inattivi' o 'dormienti' a seconda di come si presenta il saldo del conto. Lo scopo del lavoro è stato quello di ricercare indipendenze condizionate tra le variabili in particolare rispetto alle due variabili obbiettivo e definire il profilo di coloro che utilizzano maggiormente la carta.

Le conclusioni riguardanti le analisi effettuate al capitolo quinto sono riportate nell'ultima sezione. L'appendice descrive alcuni dei principali programmi relativi ai software statistici utilizzati per le elaborazioni.

CAPITOLO 1

Statistica applicata in finanza

1.1 Cambiamenti nelle metodologie di gestione del rischio di credito

L'erogazione di credito per l'azienda concedente rappresenta un'operazione rischiosa, innanzitutto perché c'è una separazione temporale tra il momento dell'erogazione e quello del rimborso del capitale e del pagamento degli interessi. Occorre quindi valutare il merito di credito del richiedente ed in base agli obbiettivi aziendali perseguiti: grado di avversione al rischio e volume dei finanziamenti erogabili, decidere se accordare o rifiutare il prestito.

Negli ultimi anni la gestione del rischio di credito sta subendo un radicale ripensamento. La microeconomia creditizia è in profonda rivisitazione. E' in corso, infatti, un più incisivo orientamento al mercato che comprende le seguenti strategie:[1]

- divisione delle attività;
- specializzazione per segmenti di mercato;
- ricerca di più profonde relazioni di clientela;
- nuove opportunità di ricavo non finanziario.

Tenendo conto di tutto ciò, il contesto decisionale del rischio di credito oggi è tale da far sì che il fabbisogno informativo dell'analista sia

maggiormente articolato e non comprenda soltanto i motivi che portano a classificare le aziende o i clienti come 'sane' o 'insolventi', 'buoni' o 'cattivi' rispettivamente ma anche l'accertamento di quali siano i segnali deboli di un eventuale crisi in atto, quando un impresa è stata valutata come 'sana' o il cliente 'buono'[2].

Si rimettono in discussione i criteri con i quali si decide se accordare o rifiutare un presto che il *pricing* dei rischi di credito, reimpostando in un'ottica di portafoglio di medio periodo la definizione di una politica generale degli impieghi.[3] In particolare rispetto al passato si pone la necessità di migliorare le capacità di valutazione di tale rischio e di affiancare alle valutazioni tradizionali misurazioni 'oggettive', che possano trovare anche riscontro nelle condizioni di prezzo delle singole operazioni, ovvero che il prezzo del prestito sia coerente con il rischio che ciascuna operazione comporta[4].

Le ragioni di una più accorta gestione dei crediti sono varie, sia economiche che finanziarie, sia ragioni di mercato che gestionali in quanto si avverte la necessità di attuare una più razionale politica dei prestiti sia nei casi di stabili relazioni con la clientela, sia nei casi di relazioni che si desidera rinsaldare. Oltre a questo occorre anche considerare che la rischiosità dei prestiti appare oggi particolarmente accentuata a causa dell'elevato numero di partite anomale (sofferenze ed incagli) e delle conseguenti perdite sui crediti. Si tenta, pertanto, di giungere a quella consapevolezza delle condizioni di credito (più opportuna combinazione di rischio e rendimento) che può assicurare alla banca o alla società finanziaria di realizzare l'efficienza allocativa[4]. In tale modo si apre la possibilità di valutare il credito in maniera

[1] Maino R. (1998), 'Nuove metodologie di gestione del rischio di credito e vantaggi competitivi per le banche', *Economia&Management*, 6, pp.73-92.
[2] Chilanti M. (1999), 'Un sistema adattivo per il supporto alle decisioni nelle analisi del rischio di credito', *Economia&Management*, 2, pp.113-126.
[3] Szegö G., Varetto F. (1999), 'Il rischio creditizio', UTET, Torino.

più oggettiva e quantificabile, al momento della decisione d'affidamento, e soprattutto della fissazione delle condizioni che assistono l'operazione. Lo scopo è avere una piena cognizione e quantificazione della rischiosità delle operazioni effettuate o effettuabili, di garantirsi al contempo che il loro svolgimento non darà luogo a sorprese, sotto il profilo economico e finanziario per l'ente finanziatore.

Alcune difficoltà oggettive che impediscono la quantificazione della rischiosità dei prestiti si stanno gradatamente superando, sia per quanto riguarda l'utilizzo di *data base* che consentono di sfruttare tutta l'informazione disponibile, tra cui anche le informazioni delle Centrali dei Bilanci e dei Rischi[5], sia per le tecniche statistiche applicate in quest'area che si stanno ampiamente sviluppando e perfezionando, in modo da permettere una misurazione e percezione più accurata di questa rischiosità, sia infine per l'affievolimento delle resistenze culturali che ostacolano l'adozione di nuove metodologie di valutazione dei fidi.

1.2 Analisi del rischio di credito

La decisione per garantire o meno il credito, fino a poco tempo fa, era presa sulla base della conoscenza personale. Fino circa vent'anni fa la maggioranza delle istituzioni finanziarie si basavano esclusivamente su un'analisi soggettiva, ovvero usavano informazioni circa le diverse caratteristiche del richiedente: il carattere (reputazione); il capitale (indebitamento); capacità di servire il debito (reddito o volatilità dei profitti) e

[4] Corigliano R.(1998), 'Il rischio di credito e pricing dei prestiti bancari', Bancaria Editrice, Roma.

[5] Le centrali di rischi sono agenzie specializzate in valutazioni dei rischi per i prestiti personali e al consumo e dotate di archivi storici nazionali dei protesti, di archivi sulle informazioni pregiudizievoli e delle visure ipotecarie e catastali e di archivi dei dati storici effettivi sul comportamento finanziario dei clienti forniti dagli stessi operatori del settore.

le garanzie collaterali; le cosiddette '4 C' del credito, per raggiungere un giudizio largamente soggettivo riguardo alla concessione o meno del prestito.

Una debolezza particolarmente avvertita nel tradizionale sistema di valutazione dei fidi era insita nella formulazione stessa del giudizio finale che era prettamente qualitativo, non in grado di 'quantificare' il rischio di credito che l'operazione comporta né giungere all'individuazione della probabilità d'insolvenza o della percentuale di recupero del credito. Proprio l'allentamento del rigore nella selezione del credito negli ultimi anni, dovuto anche all'aumento della concorrenza, è stato considerato una delle cause principali del peggioramento degli attivi delle sofferenze bancarie. Tali modalità di valutazione non consentono di prevedere con congruo anticipo il peggioramento delle condizioni economico-finanziarie del cliente o più in generale una modifica delle condizioni d'affidabilità. Queste ultime emergono in particolare dall'analisi della gestione corrente del rapporto, emergono cioè dalla fase di monitoraggio e controllo dei fidi quando la crisi o l'insolvenza dell'impresa o del cliente è ormai avviata.

Limitare, inoltre, la gestione del credito ad una semplice distinzione tra clienti adempienti o non adempienti è piuttosto riduttivo. L'obbiettivo della società concedente credito è massimizzare la redditività e la riduzione del rischio d'insolvenza è solo un aspetto di quest'obbiettivo più generale che riguarda anche la remuneravità del conto e dunque il tasso d'interesse applicato, il volume di prestito erogato, la cadenza temporale per il rimborso delle rate ed anche il momento a partire dal quale il soggetto diventa inadempiente.

Tutto ciò ha favorito negli ultimi anni da parte di banche e società finanziarie l'utilizzo di sistemi automatici per la valutazione del merito creditizio: i cosiddetti sistemi di *credit scoring*. Questi sono basati su tecniche statistiche e matematiche volte a valutare l'affidabilità creditizia dei clienti e prevedere i rischi d'insolvenza associati, assegnando ogni richiedente credito ad una determinata classe di rischio.

I sistemi di *scoring* supportano tutte le fasi della gestione del credito e non solo la fase di selezione iniziale. Essi trovano inoltre applicazione nell'analisi di fido delle imprese, nel *rating* dei titoli, nella valutazione del rischio-paese. Un'area d'applicazione in cui essi stanno avendo molta diffusione si riscontra con riferimento all'analisi della rischiosità dei fidi erogati agli individui, ovvero nell'ambito del credito al consumo, che presenta particolari specificità rispetto ad altri comparti.

Nel presente lavoro si intende analizzare questa forma di credito che nel nostro paese è stato interessato, dalla metà degli anni ottanta, da un intenso processo di crescita all'interno del sistema finanziario. Esso è il credito finalizzato all'acquisto dei beni di consumo ed è un tipico prodotto destinato al settore delle famiglie. E' concesso in Italia da società specializzate ma anche da banche e aziende commerciali che sono in forte concorrenza per i prodotti similari, quali ad esempio strumenti di pagamento come le carte di credito.

1.3 Credito al consumo

Per credito al consumo si intende la concessione di credito a favore di una persona fisica per finanziare le spese di consumo corrente o durevole, con l'esclusione di forme di sovvenzione collegate all'attività produttiva del debitore (leasing o prestiti ai professionisti). Gli importi erogati oscillano tra le 300 mila lire e i 60 milioni.

La ripresa dei consumi di beni durevoli da parte delle famiglie in Italia, che è iniziata nella seconda metà del 1996 e proseguita per tutti gli anni successivi, ha determinato un forte impulso all'espansione del credito al consumo. Invece di prelevare capitali dai propri progetti di risparmio, le famiglie italiane sono sempre più propense a ricorrere al pagamento rateale

per acquisti che in passato effettuavano soltanto quando erano in possesso dell'intera somma da pagare. Questo evidenzia la tendenza a pianificare le esigenze finanziarie modulando le spese.

Lo stock di credito erogato è cresciuto nel 1997 nel nostro paese di oltre il 30% rispetto all'anno precedente e nel 1998 di un ulteriore 20%. Nel primo semestre del 1999 i finanziamenti erogati alle famiglie sono cresciuti del 30.3%, rispetto allo stesso periodo del 1998, per un totale di prestiti pari a circa 19400 miliardi di lire (oltre 10 miliardi di euro) (Fonte Assofin[6]). La tabella 1.1 mostra le variazioni percentuali per tipologia di prestito, nel primo semestre 1999.

Esso è concesso in Italia da società finanziarie specializzate ma anche da banche e aziende commerciali. Le banche stanno riconoscendo sempre di più l'importanza di questo mercato, per le quali una funzione più evoluta di quella attualmente svolta prevede, a fianco di compiti di finanziamento alla

Tipologia di finanziamento	**Valore**	
	Miliardi	Variazione
Autoveicoli e motocicli	10746,8	25,1%
Prestiti diretti	1763,7	38,8%
Altri prestiti finalizzati	3267,0	27,1%
Carte di credito	1063,8	35,2%

Tabella 1.1 Finanziamenti erogati nel primo semestre 1999 relativi alle società finanziarie, variazione rispetto allo stesso periodo del 1998, come da tabella pubblicata su 'Sole 24 ore' 11 ottobre 1999.

[6] Associazione italiana del credito al consumo e immobiliare

produzione anche un compito di finanziamento alla domanda dei beni di consumo. Nel 1997 si è avuto un significativo aumento dello stock di crediti al consumo erogati dalle stesse (+8%) e la creazione di nuove finanziarie specializzate da parte di alcuni grandi gruppi bancari; questo perché l'operatività del comparto richiede che si sviluppino soluzioni organizzative ed istituzionali specifiche. Tant'è vero che spesso le società specializzate arrivano ad una diversificazione degli impieghi, come ha recentemente fatto Findomestic, la società leader in Italia nel credito al consumo che si è trasformata in banca ed ha distinto il settore dell'auto, le carte di credito e i prestiti personali.[7]

Il prestito è offerto con le seguenti forme tecniche:

-*i prestiti personali*: sono mutui senza vincolo di destinazione, a scadenza determinata il cui importo mediamente si aggira sui 10-12 milioni. Rientrano nella categoria i prestiti-vacanze, prestiti-nozze;

- *finanziamenti alle vendite rateali*: prestiti concessi da società finanziarie o da banche, tecnicamente collegati all'acquisto di determinati beni o servizi, per lo più a carattere durevole. Gli importi vanno dai 2 a 10 milioni e vengono erogati direttamente presso l'esercizio commerciale dove avviene la spesa. La parte più cospicua di tali finanziamenti riguarda l'acquisto dell'auto, seguono poi gli acquisti di oggetti di arredamento, di elettrodomestici o elettronica;

- *le carte di credito*: si tratta di una tessera che funge da documento di identificazione e consente al titolare di effettuare acquisti presso esercizi convenzionati di particolari beni, soprattutto non durevoli e di servizi e il prelevamento di contante agli sportelli e/o le agenzie dell'intermediario. In tale ambito recentemente si è distinto il *revolving credit*, un'operazione con la quale al cliente viene concesso un fido, utilizzabile attraverso una carta di credito, il cui ammontare si ricostituisce ogni volta che vengono effettuati i rimborsi parziali.

[7] Da 'Il Sole 24 ore' , 12 Marzo 1999, Inserto: ' Finanza e Mercati'.

1.4 Le carte di credito

Il credito al consumo rappresenta il banco di prova tipico dei prodotti finanziari innovativi quali le carte di credito e la moneta elettronica, la cui diffusione è costantemente in crescita nel nostro paese. Le caratteristiche di questi prodotti, infatti, che rendono possibile trasferire nel tempo e nello spazio il potere d'acquisto, insieme con le modifiche intervenute nelle abitudini di spesa delle famiglie, consentono di individuare ampie potenzialità di sviluppo del mercato interno ed internazionale.

Una classificazione possibile delle carte attualmente presenti sul mercato può essere fatta in relazione all'ente emittente:

- carte commerciali;
- carte Travel e Entertainment (T&E);
- carte bancarie.

Le funzioni fondamentali offerte dalla carta di credito sono:

- regolamento immediato degli scambi (strumento di pagamento attivabile tramite i terminali presso i punti vendita, o POS);
- regolamento differito degli scambi (strumento di credito);
- anticipo di contante (strumento di credito attivabile presso gli sportelli automatici, o ATM).

Inoltre accanto a queste funzioni le carte possono consentire l'accesso a servizi di svariata natura come: coperture assicurative sulla persona o sui beni acquistati, condizioni di favore sull'acquisto di beni o servizi, assistenza sanitaria di emergenza, consulenza finanziaria[8].

Attualmente il mercato delle carte di debito (cioè i PagoBancomat) e di credito, evidenzia la fine del ciclo 'familiarità e distribuzione' ovvero le carte in circolazione rappresentano un prodotto ormai maturo, tanto che alcune di esse si pongono come prodotti di massa, soprattutto grazie al fatto che la carta

è ormai uno strumento multifunzionale, per i servizi anche non finanziari che essa include.

Emerge da un convegno sulle *plastic cards* promosso da Europay International (ottobre 1999), che in Italia ci sono oltre 27 milioni di carte di pagamento in circolazione; il 63% delle famiglie italiane possiede una *plastic card*: il livello di diffusione delle carte di pagamento è pari a due carte per famiglia (fonte Databank). Inoltre la carta viene ormai accettata dalla maggior parte degli esercizi commerciali.

La tabella 1.3 illustra le carte di credito e di debito in Italia nell'anno 1998 e le variazioni rispetto agli anni 1997 e 1996.

Carte di credito in Italia	***Variazione percentuale***		
	1998	1998/1997	1997/1996
Carte di credito in circolazione (migliaia)	10150	15.0	5.5
Di cui attive:	5622	20.4	26.0
Transazioni (migliaia)	175.1	23.8	22.0
Di cui all'estero:	22.4	13.4	10.3
Carte Bancomat in circolazione(migliaia)	17989	3.9	13.4
Di cui abilitate ai POS	17000	8.3	5.3

Tabella 1.2 Carte di credito in Italia - Fonte: Banca d'Italia, *Relazione annuale*.

[8] Di Antonio M. (1994), ' Il credito al consumo', EGEA, Milano.

Tuttavia rispetto agli altri paesi europei la diffusione in Italia può considerarsi limitata; la seguente tabella illustra il numero di carte possedute per mille abitanti nei principali stati europei nel 1997.

Carte di credito e di debito	***Italia***	***Francia***	***Germania***	***Spagna***	***Regno Unito***
n. per 1000 persone	426	473	1038	897	1271
Pagamenti per carta	11	83	6	10	35

Tabella 1.3 Mercato italiano ed altri mercati europei nel 1997 -Fonte: Banca d'Italia, *Relazione annuale*

Nel corso del 1999 si è assistito ad un forte aumento del *revolving credit* con carte di credito. Come già detto si tratta di carte di pagamento collegate ad una linea di credito non garantita che consente di rateizzare l'importo speso ogni mese e rimborsarlo gradualmente proprio con la carta stessa. Pagando con questa carta non si attinge l'importo speso dal conto corrente, ma si rimborsa mensilmente pagando gli interessi passivi ed il plafond a disposizione si ricostituisce a seguito dei rimborsi mensili. Attualmente, secondo un'indagine condotta da Europay International[9] (che gestisce i marchi Mastercard e Maestro), solo il 2% delle carte di credito operanti in Italia sono *revolving*. Tuttavia si prevede un grande sviluppo per questa linea di credito che presenta una grande flessibilità perché permette di effettuare acquisti a credito senza dover chiedere finanziamento apposito in banca o presso l'esercente e di rateizzare le spese effettuate. I tassi di crescita in Italia sono più elevati di quelli registrati negli altri paesi Europei, questa tipologia di

[9] '*provider*' europeo nel campo delle carte di pagamento bancarie.

pagamento è stata utilizzata infatti in ben 3,5 milioni di operazioni, anche se l'importo medio dei finanziamenti si aggira intorno alle 300 mila lire.

Ampie potenzialità di sviluppo si intravedono poi per le carte di debito e di credito grazie al commercio elettronico; già sono effettivi i pagamenti via Internet con la carta di credito mentre si stanno studiando soluzioni per quelli con carta di debito, si stima che nei prossimi anni le transazioni on-line cresceranno considerevolmente. Inoltre soluzioni ad hoc stanno per essere perfezionate al fine di garantire non solo la sicurezza delle transazioni sul web, ma anche la certezza delle controparti[10]. Inoltre l'evoluzione delle caratteristiche tecnologiche consentirà alle carte di essere più difficilmente oggetto di contraffazione e di utilizzo fraudolento. Si prevede, infatti, che le carte a microprocessore (dette anche carte intelligenti o *smart cards*) andranno a sostituire le attuali carte a banda magnetica, poiché dotate di maggiori capacità di elaborazione e di memorizzazione, tali da consentire una identificazione immediata del titolare. Tutto ciò lascia intravedere che presto l'Italia si porterà al livello degli altri paesi europei per quanto riguarda la loro diffusione e il loro utilizzo.

1.5 Credit scoring

La valutazione del rischio nell'ambito del credito al consumo avviene principalmente attraverso tecniche di *scoring*. Il ruolo delle procedure di *scoring* è fondamentale in tale contesto per le seguenti ragioni:

- consente una metodologia uniforme di valutazione dei richiedenti che vengono valutati in base a regole predefinite e uguali per tutti. Si rinuncia così a valutazioni fatte caso per caso, garantendo l'omogeneità di trattamento nel tempo e nello spazio. Un dato che si riscontra

[10] Da 'Il Sole 24 ore' , 14 Febbraio 2000.

nell'esperienza del settore è che le valutazioni, fatte caso per caso dai funzionari, producono livelli di perdite superiori per l'impossibilità di valutare a distanza un individuo ignoto sino al momento della richiesta.[11] Da stime elaborate in vari paesi risulta inoltre che l'introduzione dello *scoring* fa calare in misura oscillante dal 15% al 45% i prestiti erogati a clienti non meritevoli;

- favorisce una notevole riduzione dei tempi una volta che il sistema di *scoring* si è sviluppato. L'attività di concessione, infatti, avviene a distanza dalla sede dell'operatore e deve essere sintonizzata sui tempi di svolgimento dell'operazione commerciale, per cui occorre che il finanziamento venga accordato in tempo reale;
- poiché il credito al consumo consiste in un numero elevatissimo di piccole operazioni, occorre disporre di sistemi di valutazione efficienti ed in grado di sfruttare le economie di scala; lo *scoring* assicura l'uniformità nella metodologia e la riduzione dei tempi che a loro volta favoriscono la riduzione dei costi unitari di ciascun prestito erogato;
- lo *scoring* consente un controllo costante del rischio permettendo di governare e prevedere il livello di perdite, dando la possibilità di pianificare l'attività e la redditività delle operazioni.

I sistemi di *scoring* supportano tutte le fasi della gestione del credito:

- *scoring di selezione,* seleziona i migliori candidati a cui proporre un prodotto di credito con azioni promozionali mirate;
- *scoring di accettazione,* consente di valutare la probabilità di insolvenza di un soggetto richiedente e guida le scelte di concessione del credito;
- *scoring comportamentale,* è utilizzato nella gestione del rapporto di credito e nel monitoraggio delle scadenze;

[11] Filotto U. Giannasca C. (1996), 'Credito al consumo: qualità del credito e gestione del rischio', *Banche e Banchieri*, 3, pp. 241-250.

- *scoring di recupero,* è in grado di stabilire qual'è la probabilità che un ritardo nei pagamenti si trasformi in un problema serio della pratica. In base ad essa consente di calibrare gli interventi da adottare per l'azione di recupero.

Dal punto di vista statistico numerose tipologie di modelli sono stati investigati e applicati soprattutto per lo *scoring* di accettazione. Metodi statistici sono stati applicati nell'area della finanza da lungo tempo, un ambito di utilizzo è certamente quello del credito ed in particolare del credito al consumo. Come ciascun altro dominio statistico esso adotta tecniche da tutto il corpo della statistica, ma la specificità delle applicazioni, le domande uniche ed i problemi distinti della finanza, richiedono il costante sviluppo di nuove classi di strumenti statistici, metodi e modelli.

Il campo principale della ricerca statistica nell'ambito del *credit scoring* è stato quello di determinare la probabilità che i nuovi richiedenti credito, una volta perfezionato il finanziamento, presentino dei ritardi di pagamento. Questo consiste nell'esame delle informazioni che si hanno sul cliente provenienti dai moduli di richiesta di finanziamento compilati dai consumatori, (*application form*), dai *credit report* forniti da agenzie specializzate e dall'archivio dei clienti della società concedente credito. Si tratta di informazioni descrittive di natura socio-demografica ed economico finanziaria, così come informazioni circa altri prodotti finanziari posseduti. Per ogni caratteristica del cliente vengono assegnati dei punteggi, che misurano il livello di rischio relativo al soggetto in esame, ovvero offrono una quantificazione numerica della probabilità che il creditore possa essere inadempiente agli impegni assunti; se la somma di tali punteggi supera una determinata soglia si decide di accordare il prestito.

Una panoramica esaustiva sui metodi di classificazione nel credito al consumo può essere trovata in Hand (1998)[12]. La tabella 1.2 fornisce una

[12] Hand D.J. e Jacka S. (1998), 'Statistics in finance', Edward Arnold, London.

rassegna su alcuni dei principali metodi statistici utilizzati nell'ambito del *credit scoring* e relative recensioni.

Metodi statistici nel Credit Scoring	
Analisi discriminante	Durand(1941), Eisenbeis (1977,1978) Moses e Liao, Myers e Forgy(1963) Grablowsky e Talley (1981)
Regressione	Orgrel (1970, 1971), Fitzpatrick (1976) Lucas (1962)
Regressione logistica	Henley (1995), Wiginton (1980)
Programmazione matematica	Hand (1981), Showres e Chakrin (1981)
Partizione ricorsiva	Breiman et al (1984), Markowski(1985), Coffman (1990)
Sistemi esperti	Zocco (1985), Davis (1987), Leonard (1993)
Reti neurali	Ripley (1984), Davis et all (1992) Rosenberg e Gleit (1994)
Metodi non parametrici	Chatterjee e Barcun (1970), Hand (1986)
Algoritmi genetici	Varetto (1998 a)
Modelli grafici	Hand, McConway, Stanghellini (1997) Sewart, Whittaker (1998) Stanghellini, McConway, Hand (1999)

Tabella 1.4 Metodi statistici utilizzati nel *credit scoring* con alcune recensioni relative.

In particolare il contributo dell'intelligenza artificiale, ha portato a sperimentare metodi come le reti neurali o i sistemi esperti che si sono rivelati molto utili nel contesto della valutazione del credito.

In generale non esiste il miglior metodo in assoluto. La validità di ciascun metodo dipende dalla situazione specifica in cui si opera: struttura dei dati, caratteristiche usate, politica di credito adottata. Inoltre la precisione nel classificare tra 'buoni' e 'cattivi' rischi è solo un aspetto per valutare il rendimento di una *score-cards* ovvero di un modello di *scoring*. Si può anche essere interessati alla velocità di classificazione, alla facilità di comprensione del metodo, alla possibilità di giustificare le decisioni raggiunte. Ogni metodo ha sue proprie peculiarità che lo rendono efficace e limitato allo stesso tempo.

Nuovi campi di ricerca si sono inoltre aperti per far fronte alle molte domande che in tale ambito sono sorte, ad esempio:

- la determinazione della profittabilità del cliente. Il rischio di insolvenza è infatti solo uno degli aspetti che interessano, l'obbiettivo della società concedente credito è quello di massimizzare la redditività. Per cui una situazione di insolvenza non rappresenta solamente una situazione di rischio ma anche una possibile fonte di guadagno. L'obbiettivo cui guardare diventa la massimizzazione del profitto atteso; questo pone la necessità di trovare soluzioni che consentano di massimizzare il profitto senza incorrere in perdite al disopra di un certo livello, di minimizzare il rischio di esposizione a perdite elevate dato un certo livello di redditività, di determinare quale sia il rischio di esposizione atteso dato un predeterminato volume del portafoglio;
- il fatto che processi di selezione implicati nella gestione comportano il mutamento della popolazione iniziale. Hand *et al*. (1997),[13] descrive bene come una popolazione iniziale attraversa vari processi di selezione, ad esempio la banca decide a chi inviare i moduli di richiesta di carta di credito o di finanziamento, solo alcuni vengono presi in considerazione e ritornano alla banca; la banca attraverso un processo di *scoring* decide a

[13] Hand D.J, McConway, M.J. e Stanghellini E. (1997),'Graphical models for applicants for credit', IMA Journal Mathematics Applied in Business. Industry, 8, pp. 143-155.

chi offrire il prestito, solo una parte dei clienti a cui il credito è stato offerto accettano il finanziamento; successivamente alcuni risultano insolventi ed altri ripagano prontamente il debito; in seguito la banca decide a chi accordare un ulteriore prestito;

- il campione usato per costruire i modelli di *score* raramente è un campione casuale dell'intera popolazione, perché esso è costruito con i dati che si hanno di coloro la cui domanda di credito è stata accettata ma serve poi a valutare l'intera popolazione dei richiedenti credito. Questo ovviamente può portare a risultati distorti. I tentativi per affrontare il problema consistono nel fare inferenza circa la restante parte della popolazione e vanno sotto il nome di *reject inference*.[14] Diviene, infatti, necessario ricostruire l'esito delle operazioni creditizie non accettate, e quindi non osservate nel seguito, sulla base del comportamento degli affidamenti concessi di cui evidentemente si conosce l'esito. Il processo d'inferenza sulla popolazione non accettata comporta la cosiddetta *augmentation* del campione di riferimento, al fine di ottenere la massima approssimazione dell'universo di studio;
- la tendenza della popolazione di riferimento ad evolvere nel tempo detta anche *population drift*. In realtà l'assunzione fondamentale che viene fatta è che la popolazione da cui un campione successivo verrà estratto, ha la stessa distribuzione di quello da cui è stato estratto il campione in esame. Molto spesso, invece, questa assunzione non è valida, soprattutto per il cambiamento delle condizioni commerciali ed economiche, dettato dalla forte concorrenza in atto. Un tentativo per risolvere il problema è stato quello di incorporare le condizioni economiche nel processo di valutazione, si veda Crook *et al* (1992), Thomas (1999), Zandi (1998).

[14] Hand D.J. e Henley, W.E. (1993), 'Can Reject inference even work?', IMA Journal Mathematics Applied in Business. Industry, 5, pp. 45-55.

Il presente lavoro costituisce un'applicazione dello *scoring* comportamentale illustrato nel seguente paragrafo.

1.6 Scoring comportamentale

Lo *scoring* comportamentale o *behavioural scoring* è un tipo di strumento di *scoring* applicabile alla fase della gestione del portafoglio crediti, in quanto consente di monitorare il comportamento del cliente al quale il credito già è stato garantito. Thomas (1999) lo definisce: ' *as the systems and models that allow lenders to make better decisions in managing existing clients by forcasting their future performance'*. Un'applicazione dello *scoring* comportamentale riguarda una stima della probabilità che un cliente passi ad uno stato d'insolvenza nel ciclo del rimborso corrente, ovvero che rimanga in situazione regolare. Questo permette al gestore di innescare azioni appropriate al livello di rischio atteso piuttosto che in base al solo stato corrente.

Rispetto allo *scoring* d'accettazione le definizioni dei clienti 'buoni' o 'cattivi' e l'applicazione della metodologia di definizione dei punteggi, si riferiscono allo stato di un cliente in essere. Uno stesso campione di conti viene analizzato in due momenti temporali consecutivi, ad esempio a distanza di sei mesi, al fine di osservare l'eventuale cambiamento di stato. Le tecniche adottate permettono di identificare le variabili che contribuiscono maggiormente a discriminare tra i diversi stati del cliente e a prevedere il gruppo di appartenenza delle nuove unità.

Lo *scoring* comportamentale è utile in molte delle decisioni che riguardano l'operazione di credito. Ad esempio se si conduce un portafoglio *revolving* il sistema è applicabile alla gestione dei limiti di credito, all'autorizzazione degli sconfinamenti, alla remissione della carta a scadenza. Inoltre può servire per prevenire le attività fraudolente che si verificano soprattutto per il fatto che vengono fornite informazioni false riguardo alle generalità del richiedente

credito, così come per promuovere azioni mirate di marketing per i prodotti innovativi a parte della clientela. Le sollecitazioni per l'acquisto di nuovi prodotti sono in genere attività piuttosto costose, per cui attraverso lo *scoring* si possono selezionare i clienti che hanno una maggiore probabilità di risposta positiva alla sollecitazione. Il limite di credito può avere un significato rilevante sulla performance dell'operazione di credito: tanto più può essere innalzato senza violare il rischio accettabile per la società tanto più l'operazione può divenire redditizia. Una strategia ragionevole sarebbe quella d'offrire ad ogni cliente un limite che lo incoraggi ad utilizzare il credito, rimanendo in una posizione di rischio accettabile per l'azienda.

Il *behavioural scoring* è uno strumento utile per se stesso, ma può diventarlo ancora di più se è inserito in un sistema di gestione che consente di esaminare strategie alternative di controllo dei crediti, valutandone l'efficacia, al fine di favorire il progressivo adattamento al cambiamento delle condizioni. In tale ottica è stato sviluppato un sistema noto come controllo adattivo, basato su una strategia di *verifica, apprendimento, adattamento*[15]. Esso consiste nell'implementare una gestione sistematica delle politiche di portafoglio, in vista non solo dei risultati ottenibili alle condizioni correnti ma anche allo scopo di ottenere informazioni utili al controllo e miglioramento dei risultati futuri. Lo scopo è di facilitare la suddivisione del portafoglio di crediti in numerosi livelli, in relazione alla loro propensione verso alcuni eventi. Per ulteriori approfondimenti si rimanda a Lewis (1994)

Solo attraverso l'utilizzo di tecnologie sofisticate, come le metodologie di *scoring*, si può assumere un comportamento che favorisca la selezione del portafoglio e consenta di assumere un comportamento proattivo e nello stesso tempo attento al rischio.[16]

[15] Lewis E. (1994), 'An Introduction to Credit scoring' , The Athena Press, California.
[16] Filotto U. Giannasca C. (1996), op. cit. pp. 241-250.

1.7 Modelli grafici per lo scoring comportamentale

Il contesto dello *scoring* comportamentale si presta molto bene ad essere analizzato attraverso una classe di strumenti statistici che prende il nome di modelli grafici a catena. Essi rientrano nella classe dei modelli grafici o modelli di indipendenza condizionata (Whittaker 1990, Lauritzen 1996) che sono modelli multivariati per studiare l'associazione e la dipendenza tra le variabili. Tali modelli sono particolarmente utili in quanto aiutano la comprensione delle relazioni tra le variabili che descrivono il comportamento dei debitori ed inoltre sono strumenti molto flessibili, in grado di mostrare le relazioni tra un elevato numero di variabili. Applicazioni dei modelli grafici per il *credit scoring* si trovano in Hand *et al*. (1997), Sewart e Whittaker (1998), Stanghellini *et al*. (1999), Ciavarella (1999).

Nel presente lavoro si utilizzano tali modelli allo scopo di definire il profilo economico e commerciale dei titolari di carta di credito di tipo *revolving* e monitorare nel tempo le loro performance, individuando quale parte della clientela ha maggiore probabilità di utilizzo della carta, oppure ha maggiore rischio di abbandono.

Nello *scoring* comportamentale rispetto allo *scoring* di applicazione sono disponibili ulteriori informazioni, come ad esempio il comportamento di spesa del cliente nel periodo di validità della carta. I modelli grafici a catena costituiscono un metodo più efficiente per tener conto di tali informazioni rispetto ai metodi tradizionalmente utilizzati come la regressione logistica e l'analisi discriminante e sono in grado di focalizzare l'attenzione su più di un aspetto del comportamento del titolare della carta diversamente dai metodi univariati. Inoltre, tali modelli sono interessanti per la teoria che li accomuna ai sistemi esperti, così che possono essere usati per derivare, a partire da un sottoinsieme di relazioni, il comportamento più probabile di un cliente in relazione ad un insieme di variabili d'interesse.

In particolare i modelli grafici a catena sono utili quando, come nel caso considerato, c'è una struttura a blocchi ricorsiva che può essere dedotta da informazioni *a priori* che si hanno sulle variabili oggetto di studio, si può cioè stabilire una sequenza temporale-causale fra di esse. Le variabili in considerazione hanno un ordine aciclico parziale, ovvero possono essere divise in gruppi. Le variabili del primo gruppo sono prevalentemente variabili esplicative, quelle dell'ultimo gruppo sono essenzialmente variabili risposta, quelle dei blocchi intermedi sono variabili risposta per quelle che le precedono mentre sono esplicative per quelle che le seguono. L'aciclicità sta nel fatto che una variabile non può essere a sua volta, anche indirettamente, esplicativa di se stessa, perciò nei modelli grafici a catena le variabili possono essere suddivise in meramente esogene, intermedie tra loro concatenanti e risposta.

Un'applicazione dei modelli grafici a catena per il *behavioural scoring* può essere trovata in A. Neri (1999).

CAPITOLO 2

Misure di associazione tra variabili categoriche nominali e ordinali

2.1 Definizione delle variabili

Le variabili categoriali o categoriche sono quelle variabili la cui scala di misura consiste in un set di categorie o livelli, si dicono variabili dicotomiche quelle con solo due categorie, mentre politomiche quelle con più di due categorie. Sono dette *nominali* quelli variabili categoriche per le quali i livelli o categorie sono interscambiabili e non presentano un ordine naturale. Sono dette *ordinali* quelle variabili categoriche che hanno livelli ordinati in modo simile ai numeri ordinali (primo, secondo,…). In tale caso non ha senso parlare di distanza o spazio tra primo e secondo né confrontare gli spazi tra coppie di categorie di risposte dato che le categorie sono interscambiabili.

I dati esaminati provengono da un campione in cui ogni unità statistica viene classificata simultaneamente in base a numerose variabili sia continue che categoriche.

In particolare, in questo lavoro, le variabili continue vengono categorigazzate così da concentrarci su modelli per variabili categoriche. Inoltre si tenta di tener conto dell'ordinalità di alcune variabili. Le variabili ordinali si presentano di frequente in numerosi campi soprattutto nelle scienze sociali.

Spesso sono ottenute trattando le variabili continue come categoriche, ad esempio l'età e il reddito. L'ordine all'interno delle categorie costituisce un elemento conoscitivo ulteriore che agevola la comprensione delle relazioni esistenti tra le variabili:[17]

- le analisi per variabili ordinali consentono di usare un'ampia gamma di modelli in grado di cogliere alcuni trend che i modelli per variabili nominali non sono in grado di cogliere. Inoltre i parametri dei modelli hanno un'interpretazione più semplice;
- per le analisi ordinali esistono modelli non saturi in situazioni in cui i modelli nominali possono essere saturi;
- i test basati sui modelli ordinali consentono di verificare numerose ipotesi alternative all'ipotesi nulla di indipendenza.

Se le relazioni tra variabili hanno un'interpretazione direzionale, come ad esempio, l'età, il numero di figli e il comportamento con la carta di credito è opportuno suddividerle in variabile risposta e variabili esplicative.

Si introducono pertanto alcune misure di associazione tra le variabili categoriche, utili anche ad individuare delle relazioni di indipendenza condizionata. Questa è uno strumento molto potente sia nella teoria probabilistica che nell'inferenza statistica. In particolare i rapporti di *odds* hanno molta rilevanza in quanto utili per la costruzione dei modelli considerati nel capitolo seguente. Questi ultimi sono modelli di regressione che descrivono come la distribuzione della variabile risposta, ad esempio il comportamento con la carta di credito, cambia a seconda dei livelli delle altre variabili ovvero i valori che una certa grandezza assume in relazione alla presenza o assenza di altre circostanze. L'interesse per tali modelli multivariati consiste nel fatto che non si studiano le relazioni tra due variabili tenendone altre sotto controllo ma si formulano modelli complessivi della realtà e delle interazioni tra variabili in essa esistenti, simulandole nella sua

[17] Agresti A. (1990), 'Categorical Data Analysis', Wiley, New York, p.262.

totalità, così da mettere a confronto i dati teorici prodotti con quelli osservati.[18] Tale approccio consente di non avere risultati fuorvianti circa la struttura delle relazioni come potrebbe accadere usando metodi univariati.

2.2 Analisi delle tabelle di contingenza

Il punto di partenza per l'analisi delle relazioni tra variabili categoriche è costituito dalle *tabelle di contingenza* o tabella a doppia entrata, che si ottiene classificando un insieme di unità statistiche, rispetto a due o più caratteri che determinano la dimensione della tabella: *bidimensionale* o *multidimensionale* rispettivamente. La tabella mostra le frequenze osservate associate ad ogni possibile combinazione delle categorie delle variabili. Le variabili categoriche vengono indicate con le lettere X, Y, Z e così di seguito mentre le relative modalità sono indicate con le lettere minuscole; ad esempio, la generica *i*-esima modalità di X è indicata con x_i. Talvolta e dove non sussistono ambiguità il valore x_i verrà indicato semplicemente con i.

L'ordine di una tabella doppia per la X e la Y è indicato come $I \times J$, dove I è il numero delle categorie della variabile che definisce le righe e J è il numero di categorie della variabile che definisce le colonne. La frequenza di una generica cella è indicata come: n_{ij} numero delle osservazioni in cui $X = x_i$, $Y = y_i$. Per ogni cella è definita la probabilità corrispondente $\pi_{ij} = pr(X = i, Y = j)$ che un'unità scelta a caso si collochi in una particolare casella.

[18] Corbetta P.(1992), 'Metodi di analisi multivariata per le scienze sociali', Bologna, Il Mulino.

Per indicare la distribuzione campionaria si usa la notazione p invece di π, ad esempio $\{p_{ij}\}$ indica la distribuzione congiunta campionaria in una tabella di contingenza doppia. Si può scrivere:

$$p_{ij} = \frac{n_{ij}}{n},$$

dove $n = \sum_i \sum_j n_{ij}$, corrisponde alla dimensione del campione, così che $\sum_i \sum_j p_{ij} = 1$.

L'insieme di probabilità $\{\pi_{ij}\}$ di una tabella doppia definisce la distribuzione congiunta delle due variabili in esame. Le due distribuzioni marginali si ottengono sommando rispetto ad ogni indice. Queste sono indicate con $\{\pi_{i.}\}$ per la variabile riga e $\{\pi_{.j}\}$ per la variabile colonna, dove

$$\pi_{i.} = \sum_j \pi_{ij} \qquad \pi_{.j} = \sum_i \pi_{ij}.$$

Nel caso di una tabella tripla $I \times J \times K$ le notazioni analoghe sono le seguenti: n_{ijk} numero di casi in cui: $X = x_i$, $Y = y_j$, $Z = z_k$. Dalla distribuzione congiunta $\pi_{ijk} = pr(X = i, Y = j, Z = k)$ è possibile ricavare le sei tabelle marginali, le probabilità marginali sono denotate da $\{\pi_{i..}\}$, $\{\pi_{.j.}\}$, $\{\pi_{..k}\}$ ed anche $\{\pi_{i.k}\}, \{\pi_{ij.}\} e \{\pi_{.jk}\}$ dove:

$$\pi_{i..} = \sum_j \sum_k \pi_{ijk} = \sum_j \pi_{ij.} = \sum_k \pi_{i.k} \qquad \pi_{.j.} = \sum_i \sum_k \pi_{ijk} \qquad \pi_{..k} = \sum_i \sum_j \pi_{ijk}$$

$$\pi_{.jk} = \sum_i \pi_{ijk} \qquad \pi_{i.k} = \sum_j \pi_{ijk} \qquad \pi_{ij.} = \sum_k \pi_{ijk}.$$

Nel trattare le frequenze osservate come variabili casuali, ciascun n_{ij} ha una distribuzione con valore atteso $m_{ij} = E(n_{ij})$. Gli $\{m_{ij}\}$ sono chiamate *frequenze attese*. E' il numero medio di osservazioni che ci si attende nella i-esima riga e j-esima colonna.

Se un soggetto è classificato nella riga i di X, si definisce $\pi_{j/i}$ la probabilità di classificazione nella colonna j di Y per $j = 1,2,.....,J$, ovvero è la probabilità condizionata di Y al livello i di X, dove $\sum_j \pi_{j/i} = 1$. Le probabilità $\{\pi_{1/i}, \pi_{2/i}, ..., \pi_{j/i}\}$ costituiscono la distribuzione condizionata di Y al livello *i* X. Quest'ultima è legata alla distribuzione congiunta da:

$$\pi_{j/i} = \frac{\pi_{ij}}{\pi_{i.}} \qquad \text{per ogni } i \text{ e } j \ .$$

Nel caso in cui Y sia una variabile risposta è utile poter confrontare la distribuzione condizionata di Y ai vari livelli delle variabili esplicative. Ad esempio se Y rappresenta lo stato del conto del cliente in un determinato periodo, X il reddito e Z l'età è interessante confrontare l'associazione tra Y e X per ogni classe d'età.

Quando le variabili sono ordinali**,** in particolare se la variabile risposta è ordinale, è maggiormente informativo far questo utilizzando la funzione di densità cumulata:

$$F_{j/i} = \sum_{b \le j} \pi_{b/j} \qquad j = 1, 2, \ldots, J.$$

$F_{j/i}$ è la probabilità di classificazione nelle prime j colonne di Y , data la classificazione nella riga i di X. Se si suppone che per due righe *h* ed *i*,

$$F_{j/h} \le F_{j/i} i \qquad j = 1, 2, \ldots, J$$

significa che è maggiormente probabile che la riga *h* abbia osservazioni al livello più alto della scala ordinale che non la riga *i*. Questo si indica dicendo che la distribuzione condizionata nella riga *h* è stocasticamente più alta di quella nella riga *i*. Ad esempio è possibile che coloro che hanno un numero di figli maggiore abbiano maggiore probabilità di avere il conto attivo o dormiente piuttosto che passivo.

La conoscenza del metodo di campionamento è il primo passo per l'analisi dei dati. Nel caso di dati categorici i risultati dell'analisi statistica sono, entro certi limiti, invarianti rispetto allo schema di campionamento con cui è stata ottenuta la tabella osservata: si può supporre che le unità statistiche facciano parte di un campione casuale di dimensione prefissata, oppure che i casi considerati siano tutti gli eventi che si verificano in un certo intervallo e quindi il loro numero complessivo non è fissato in anticipo o ancora, che il campionamento sia stratificato in cui è fissata la dimensione del campione ed anche le frequenze marginali rispetto ad uno o più caratteri. Nel primo caso la distribuzione campionaria è una distribuzione multinomiale di cui la binomiale è un caso particolare. Nel secondo tipo di schema di campionamento si assume che tutte le osservazioni delle celle siano realizzazioni di variabili casuali indipendenti aventi una distribuzione di Poisson, in essa il limite superiore dei casi possibili è infinito Nel terzo caso lo schema di campionamento è multinomiale stratificato.

I metodi d'inferenza per i dati categorici che consentono d'individuare una struttura semplificata per le probabilità di cella e una seguente stima per esse, assumono uno schema di campionamento multinomiale o di Poisson. Si dimostra che partendo da una distribuzione di Poisson e condizionandosi al totale, si ottiene una distribuzione multinomiale. I valori del campione sono utilizzati per fare inferenza circa la sottostante struttura della tabella.

2.3 Indipendenza condizionata

Dalla definizione di indipendenza si ha che due variabili casuali categoriche sono indipendenti se la funzione di probabilità congiunta è data dal prodotto delle funzioni di probabilità marginali:

$$\pi_{ij} = \pi_{i.}\pi_{j.} .$$

Quando X e Y sono indipendenti la distribuzione condizionata di Y è identica alla sua distribuzione marginale:

$$\pi_{j/i} = \frac{\pi_{ij}}{\pi_{i.}} = \frac{\pi_{i.}\pi_{j.}}{\pi_{i.}} = \pi_{.j} \qquad \text{per} \quad i = 1,\ldots..,I.$$

Ad esempio se tra X, lo stato civile e Y, la condizione lavorativa (impiegato, operaio,…) ci fosse indipendenza, la conoscenza dello stato civile non darebbe alcuna informazione su quale tipo di lavoro si svolge. Allora due variabili sono indipendenti quando la probabilità di classificazione nella colonna j è la stessa per ogni riga, per j = 1, 2,…., J. Per cui la definizione d'indipendenza può esprimersi sulla base della distribuzione condizionata o congiunta.

Nel caso di una tabella $I \times J \times K$ con X, Y e Z variabili rispettivamente, si denotano le probabilità di cella come $\{\pi_{ijk}\}$ con $\sum_i \sum_j \sum_k \pi_{ijk} = 1$. Le tre variabili sono ***mutuamente indipendenti*** quando

$$\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k} \qquad \text{per ogni} \quad i, j \text{ e } k.$$

La variabile Y è ***congiuntamente indipendente*** da X e Z quando

$$\pi_{ijk} = \pi_{i.k}\pi_{.j.} \qquad \text{per ogni} \quad i, j \text{ e } k.$$

Questa è l'indipendenza ordinaria in una tabella doppia, data dalla variabile Y e una nuova variabile composta dalle $I \times K$ combinazioni dei livelli di X e Z. In simboli questo si indica come $Y \perp (X,Z)$. In modo analogo X può essere congiuntamente indipendente da Y e Z, così come Z può essere congiuntamente indipendente da X e Y. Non è difficile dimostrare che la mutua indipendenza implica l'indipendenza congiunta.

Considerando la relazione tra X e Y controllando Z, si considerano le osservazioni delle celle di X e Y ai diversi livelli della variabile Z. Le sezioni incrociate della tabella a tre dimensioni che vengono utilizzate sono dette

tabelle parziali. In queste il valore della variabile Z è tenuto costante. La tabella ottenuta sommando ogni tabella parziale è definita *tabella marginale*. In essa Z non è controllato ma ignorato.

Se X e Y sono indipendenti nella tabella parziale per la *k*-esima categoria di Z, X e Y sono dette essere ***condizionatamente indipendenti*** al livello *k* di Z. Sia

$$\pi_{ij/k} = \frac{\pi_{ijk}}{\pi_{..k}} \qquad \text{per ogni } i, j.$$

la distribuzione congiunta di X e Y condizionata al livello *k* di Z. L'indipendenza condizionata al livello *k* di Z implica che si possa scrivere:

$$\pi_{ij/k} = \frac{\pi_{i./k}}{\pi_{.j/k}} \qquad \text{per ogni } i, j.$$

Più in generale X e Y sono ***condizionatamente indipendenti*** dato Z ($X \perp Y|Z$), se essi sono condizionatamente indipendenti per ogni livello di Z, così da poter scrivere:

$$\pi_{ijk} = \frac{\pi_{i.k}\pi_{.jk}}{\pi_{..k}} \qquad \text{per ogni } i,\ j \text{ e } k.$$

Si dimostra che l'indipendenza congiunta di Y da X e Z implica anche l'indipendenza condizionata di X e Y. Infatti, se Y è congiuntamente indipendente da X e Z

$$\pi_{ijk} = \pi_{i.k}\pi_{.j.};$$

per la definizione d' indipendenza condizionata si ha che:

$$\pi_{ij/k} = \frac{\pi_{ijk}}{\pi_{..k}} = \frac{\pi_{i.k}\pi_{.j.}}{\pi_{..k}};$$

sommando entrambi i lati per *i*, si ottiene:

$$\pi_{.j/k} = \pi_{.j.}$$

allora

$$\pi_{ij/k} = \left(\frac{\pi_{i.k}}{\pi_{..k}}\right)\pi_{.j.} = \pi_{i./k}\pi_{.j/k}$$

così X e Y sono anche condizionatamente indipendenti. Inoltre sempre dalla definizione d'indipendenza congiunta di Y da X e Z

$$\pi_{ijk} = \pi_{i.k}\pi_{.j.},$$

sommando per *k* si ottiene:

$$\pi_{ij.} = \pi_{.i..}\pi_{.j.},$$

così che X e Y sono anche marginalmente indipendenti.

Poiché la mutua indipendenza tra X, Y e Z implica l'indipendenza congiunta di Y da X e Z ,essa implica anche che X e Y sono entrambi indipendenti sia marginalmente che condizionatamente a Z. Tuttavia sapere soltanto che Y e X sono condizionatamente indipendenti non implica che essi siano anche marginalmente indipendenti. Infatti, se X e Y sono condizionatamente indipendenti si può scrivere:

$$\pi_{ijk} = \frac{\pi_{i.k}\pi_{.jk}}{\pi_{..k}},$$

sommando entrambi i lati per *k* si ottiene:

$$\pi_{ij.} = \sum_k \left(\pi_{i.k}\pi_{.jk} / \pi_{..k}\right);$$

tutti i termini dentro la parentesi comprendono *k* e in generale ciò non permette di semplificare per l'indipendenza $\pi_{i..}\pi_{.j.}$ marginale. Ad esempio il

sesso ed il reddito potrebbero essere condizionatamente indipendenti, dato il titolo di studio, però l'*odds ratio* calcolato per la tabella marginale tra il sesso e il reddito potrebbe essere diverso da uno, ovvero se si ignorasse il titolo di studio, l'*odds* di avere un reddito basso per le donne potrebbe risultare ad esempio due volte quello degli uomini.

L'indipendenza marginale dunque può essere molto diversa da quella condizionata, nel senso che vi può essere indipendenza condizionata e tuttavia una forte relazione marginale o viceversa. Tale fenomeno è noto come *paradosso di Simpson*, si veda l'esempio sulla pena di morte in Florida Agresti (1990, p.136). E' pertanto preferibile studiare l'associazione condizionata e da questa dedurre le relazioni sulle distribuzioni marginali. Soltanto in alcuni casi come si mostrerà successivamente è possibile effettuare analisi su distribuzioni marginali senza incorrere in errori.

2.4 Differenza tra le proporzioni

Un metodo per descrivere l'associazione tra due variabili categoriche in cui una è risposta dell'altra, come ad esempio, l'atteggiamento con la carta di credito e in numero di figli, consiste nel valutare la differenza tra le proporzioni. E' particolarmente utile quando si ha una variabile risposta binaria e si vuole investigare se due o più gruppi hanno differenti proporzioni di risposta in una certa categoria. In una tabella $I \times 2$ dove le colonne rappresentano i livelli della variabile risposta Y, si indica con $\pi_{1/i}$ la probabilità della risposta 1 per il soggetto nella riga i, per $i = 1, 2, \ldots, I$; $(\pi_{1/i}, \pi_{2/i}) = (\pi_{1/i}, 1 - \pi_{1/i})$ è la distribuzione condizionata della variabile risposta binaria.

Si possono confrontare due righe, per esempio *h* ed *i*, usando la differenza nelle proporzioni $(\pi_{1/h} - \pi_{1/i})$. Il confronto con la risposta 2 è identico a quello con la 1, dato che se la variabile è binaria si può scrivere: $(\pi_{2/h} - \pi_{2/i}) = (1 - \pi_{1/h}) - (1 - \pi_{1/i}) = \pi_{1/i} - \pi_{1/h}$.

Questa differenza di proporzioni cade nell'intervallo ±1. Essa è uguale a zero quando le righe *h* ed *i* hanno la stessa distribuzione condizionata. Allora si può affermare che la variabile risposta è statisticamente indipendente dalla variabile di riga quando $(\pi_{1/h} - \pi_{1/i}) = 0$ per tutte le coppie di righe *i* ed *h*.

Trattando ogni riga come un campione casuale binomiale indipendente si può ottenere una stima delle differenze. Nella riga *i*, n_{i1} ha una distribuzione binomiale con grandezza campionaria $n_{i.}$. Le proporzioni campionarie $p_{1/i} = \frac{n_{i1}}{n_{i.}}$ hanno valore atteso $\pi_{1/i}$ e varianza $\frac{\pi_{1/i}(1-\pi_{1/i})}{n_{i.}}$. Poiché queste sono indipendenti, la loro differenza ha valore atteso $E(p_{1/1} - p_{1/2}) = \pi_{1/1} - \pi_{1/2}$ ed errore standard:

$$\sigma(p_{1/1} - p_{1/2}) = \left[\frac{\pi_{1/1} - (1 - \pi_{1/1})}{n_{1+}} + \frac{\pi_{1/2} - (1 - \pi_{1/2})}{n_{2+}} \right]^{1/2},$$

sostituendo ai valori teorici $\pi_{1/i}$ le proporzioni campionarie $p_{1/i}$ si può ottenere σuna stima di σ.

Applicando il teorema del limite centrale, se le frequenze campionarie sono grandi, si può usare la distribuzione normale come approssimazione di quella delle differenze $(p_{1/1} - p_{1/2})$. Pertanto si può ottenere un intervallo di confidenza di livello α:

$$(p_{1/1} - p_{1/2}) \pm z_{\alpha/2}\hat{\sigma}(p_{1/1} - p_{1/2}).$$

Se tale intervallo contiene lo zero l'ipotesi $\pi_{1/h} = \pi_{1/i}$ si accetta. La probabilità di errore di primo tipo di questo test è pari ad α.

2.5 Odds e odds ratio

Tra le misure di associazione particolarmente utili per variabili categoriche ci sono gli *odds* (o rapporti di scommessa). E' un termine familiare molto usato nei giuochi di scommesse come nelle corse dei cavalli, dove esso sta ad indicare il rapporto tra l'eventuale vincita e la somma scommessa, (se un cavallo è dato 3 ad 1 significa che in caso di vincita la cifra scommessa viene resa triplicata allo scommettitore). Essi non vanno confusi con le probabilità .

Supponendo che un evento abbia probabilità p, l'*odds* per tale evento è definito come:

$$O = \frac{p}{1-p} = \frac{pr(che\ l'evento\ si\ verifichi)}{pr(che\ l'evento\ non\ si\ verifichi)}$$

Il rapporto considerato assume un valore tra 0 e $+\infty$, in relazione alla probabilità di verificarsi che ha l'evento; tanto più grande è l'*odds* tanto maggiore è la probabilità dell'evento. Se l'*odds* è O allora la probabilità p è data da

$$p = \frac{O}{O+1}.$$

Considerando la prima riga della tabella 2 × 2 illustrata sotto, l'*odds* che la risposta sia nella colonna 1 piuttosto che nella 2 è definito come:

$$\Omega_1 = \frac{\pi_{1/1}}{\pi_{2/1}} = \frac{\pi_{11}}{\pi_{12}}.$$

Per la seconda riga l'*odds* corrispondente è:

$$\Omega_2 = \frac{\pi_{1/2}}{\pi_{2/2}} = \frac{\pi_{21}}{\pi_{22}}.$$

Ciascun *odds* è non negativo con valore tanto più grande quanto la probabilità della risposta 1 è maggiore della 2. Ad esempio un $\Omega_1 = 4$ indica che nella prima riga la risposta 1 è quattro volte più probabile della risposta 2. Attraverso di essi è facile verificare la relazione di indipendenza tra due variabili: infatti, quando le distribuzioni condizionate all'interno delle righe sono identiche gli *odds* $\Omega_1 = \Omega_2$ devono essere uguali.

Il rapporto tra gli *odds* è definito *odds ratio* o *cross product ratio*:

$$\theta = \frac{\Omega_1}{\Omega_2};$$

	Colonne		
	1	***2***	***Totali***
Righe			
1	π_{11} $(\pi_{1/1})$	π_{12} $(\pi_{2/1})$	$\pi_{1.}$
2	π_{21} $(\pi_{1/2})$	π_{22} $(\pi_{2/2})$	$\pi_{2.}$
Totali	$\pi_{.1}$	$\pi_{.2}$	1

Tabella 2.1 Probabilità congiunte, condizionate e marginali per una tabella 2 × 2.

esso è uguale al rapporto dei prodotti delle probabilità di celle diagonalmente opposte:

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Se vale la relazione di indipendenza allora $\theta = 1$. Infatti, vale

$$\pi_{ij} = \pi_{i.}\pi_{j.},$$

che nel caso di una tabella 2 × 2 diventa:

$$\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \left(\frac{\pi_{1.}\pi_{.1}\pi_{2.}\pi_{.2}}{\pi_{1.}\pi_{.2}\pi_{2.}\pi_{.1}}\right) = 1.$$

Il valore dell'*odds ratio* varia tra (0, +∞) . Quando $0 < \theta < 1$ la prima risposta è meno probabile nella riga 1 che nella riga 2 ovvero $\pi_{1/1} < \pi_{1/2}$. Una proprietà importante degli *odds* è che essi non dipendono dalla distribuzione marginale, nel senso che se, ad esempio, le righe o le colonne sono moltiplicate per una costante, l'*odds ratio* non cambia valore. Invertendo l'ordine delle righe o delle colonne, si ottiene un valore che è l'inverso del valore originario. Se invece si cambia l'orientamento della tavola, il valore dell'*odds* resta immutato. Valori lontani da uno rappresentano una forte associazione tra le variabili. Per ovviare all'asimmetria tra i valori è più conveniente usare il logaritmo naturale $\log(\theta)$; la sua misura varia tra $(-\infty,+\infty)$. In questo caso l'indipendenza corrisponde al $\log(\theta) = 0$ così che l'*odds ratio* è simmetrico intorno allo zero.

Nel caso di una tabella *I*×*J*, gli *odds* possono essere formati utilizzando ciascuna delle $\binom{I}{2} = I(I-1)/2$ coppie i righe in combinazione con le $\binom{J}{2}$ coppie di colonne. Così per le righe a e b e le colonne c e d ci sarebbero

$\binom{I}{2}\binom{J}{2}$ *odds ratio* del tipo $\frac{\pi_{ac}\pi_{bd}}{\pi_{bc}\pi_{ad}}$. Tuttavia il set di *odds ratio* considerato contiene delle informazioni ridondanti. Quando le probabilità di cella sono positive, si dimostra che il sottogruppo di *odds ratio* $(I\text{-}1)(J\text{-}1)$ è sufficiente per descrivere tutti gli *odds ratio* che possono essere formati utilizzando tutte le coppie di righe e colonne. Ovvero il sottogruppo di *odds ratio* $(I\text{-}1)(J\text{-}1)$ descrive l'associazione in una tabella $I \times J$ senza perdita di informazione. Il modo di costruire il set miminale di *odds ratio* non è unico. Un metodo per ottenere questo set di *odds* è il seguente:

$$\alpha_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}} \qquad i = 1,2,\ \ldots,I\text{-}1, \quad j = 1,2,\ldots,J\text{-}1.$$

dove si considera l'ultima categoria della variabile Y come base.

Se si lavora con variabili ordinali si possono costruire diverse famiglie di *odds ratio*. Una di queste si ottiene utilizzando celle in righe o colonne adiacenti:

$$\theta_{ij} = \frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \qquad i = 1,2,\ \ldots,I\text{-}1, \quad j = 1,2,\ldots,J\text{-}1.$$

Questi vengono definiti *odds ratio locali*.

Un'altra famiglia di *odds ratio* che invece distingue tra le righe e le colonne è data da *odds* che sono locali per la variabile riga e globali per la variabile colonna:

$$\theta'_{ij} = \frac{\left(\sum_{k \le j} \pi_{ik}\right)\left(\sum_{b > j} \pi_{i+1,k}\right)}{\left(\sum_{k > j} \pi_{ik}\right)\left(\sum_{k \ge j} \pi_{i+1,k}\right)} \qquad i = 1,2,\ldots,I\text{-}1, \quad j = 1,2,\ldots,J\text{-}1.$$

Questi *odds* vengono definiti *odds ratio locali-globali*.

Una terza famiglia di *odds ratio* per variabili ordinali, definiti *odds ratio globali*, si ottiene utilizzando *odds* che sono globali sia per le variabili riga sia per le variabili colonna:

$$\theta''_{ij} = \frac{\left(\sum_{k\le i}\sum_{l\le j}\pi_{kl}\right)\left(\sum_{k>i}\sum_{l>j}\pi_{kl}\right)}{\left(\sum_{k\le i}\sum_{l>j}\pi_{kl}\right)\left(\sum_{k>i}\sum_{l\le j}\pi_{kl}\right)} \quad i=1,2,\ \ ,I\text{-}1, \quad j=1,2,\ldots,J\text{-}1.$$

Per ciascun set di *odds ratio*, l'indipendenza tra le variabili si ha quando tutti i logaritmi degli *odds ratio* sono uguali a zero. Convertire le probabilità di cella in uno dei set di *odds ratio* precedenti, non comporta perdita di informazione, infatti, date le probabilità marginali $\{\pi_{i.}\}$e $\{\pi_{.j}\}$, le probabilità di cella sono completamente determinate da questo set di *odds*.[19]

Nel caso di una tabella $I\times J\times K$ l'associazione marginale e parziale può essere descritta usando gli *odds ratio* relativi alla tabella marginale o condizionale. La tabella marginale $\{\pi_{ij.}\}$ di X-Y, è descritta da $(I\text{-}1)(J\text{-}1)$ *odds ratio* del tipo:

$$\theta_{ij} = \frac{\pi_{ij.}\pi_{i+1,j+1,.}}{\pi_{i,j+1,.}\pi_{i+1,j,.}}, \quad 1\le i\le I-1 \quad 1\le j\le J-1.$$

Per la tabella parziale si possono calcolare gli *odds ratio* condizionati alle varie categorie delle variabili di cui non si calcola l'associazione, ovvero si possono calcolare gli *odds ratio* tra X e Y condizionandoci alle varie categorie di Z:

$$\theta_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}}, \quad 1\le \mathrm{i}\le I-1 \quad 1\le \mathrm{j}\le J-1.$$

[19] Packett R.L. (1981), 'The Analysis of Categorical Data', Griffin, London.

L'associazione condizionata tra X e Y è descritta da (I-1)(J-1) rapporti di *odds* $\{\theta_{ij(k)}\}$ per ogni livello k di Z. Similmente, l'associazione condizionata tra X e Z è descritta da (I-1)(K-1) *odds ratio* $\{\theta_{i(j)k}\}$ per ciascuno dei J livelli di Y e l'associazione tra Y e Z da $\{\theta_{(i)jk}\}$ *odds ratio* per ogni livello di X.
Quando X e Y sono condizionatamente indipendenti dato Z vale:

$$\theta_{ij(k)} = 1, \quad i = 1,2,...,I-1, \quad j = 1,2,...,J-1 \quad k = 1,2...,K-1 \ .$$

Il rapporto tra i rapporti di associazione condizionati $\theta_{ij(k)}$ fornisce una misura della differenza tra i rapporti di associazione condizionati tra le variabili X e Y al variare della variabile Z e rappresenta una misura di interazione tra le variabili. Ad esempio se Z ha due categorie il rapporto

$$\frac{\theta_{ij(1)}}{\theta_{ij(2)}} = 1,$$

indica che non c'è interazione tripla tra le variabili: l'associazione tra due variabili non varia al variare della terza variabile. Ad esempio la relazione tra il reddito X e il comportamento con la carta di credito Y è lo stesso sia che non si abbiano figli o che si abbiano Z. E si dimostra che questo equivale a dire che l'associazione tra X e Z non varia al variare della variabile Y e che la relazione tra Y e Z non varia al variare delle categorie della X. Da cui il reddito ed il numero dei figli hanno un *effetto additivo* sul comportamento con la carta di credito.Valori superiori ad uno del rapporto tra *odds ratio* indicano invece, che la relazione tra X e Y è superiore tra le persone di centro destra, (categoria 1 di Z) che non in quelle di sinistra. Per cui le due influenze X e Z si cumulano con *effetto moltiplicativo*.

Il valore del rapporto considerato varia tra $(0,+\infty)$, anche in questo caso conviene applicare la trasformazione logaritmica, per cui se il logaritmo del rapporto è zero non c'è *interazione tripla* tra le variabili.

CAPITOLO 3

Modelli per variabili categoriche

3.1 Modelli per variabili categoriche

Introducendo una trasformazione delle probabilità di una tabella di contingenza i modelli d'interesse possono essere definiti in modo analogo ai modelli di regressione lineare. Una classe di modelli che consente di tenere conto della struttura multivariata delle variabili è costituita dai modelli *log-lineari* e *logistici*, che fanno parte di una famiglia più generale di modelli: i ***Modelli Lineari Generalizzati*** (GLM), introdotti da Nelder e Wedderburn (1972). Questi modelli sono descritti da tre componenti:

- una componente casuale, che identifica la distribuzione di probabilità della variabile risposta;
- una componente sistematica, che specifica la funzione lineare delle variabili esplicative;
- una funzione di legame (*link*) che descrive la relazione funzionale tra la componente sistematica e il valore atteso della componente casuale.

Nei modelli di regressione lineare classici un vettore di $\mathbf{y}$ osservazioni con N componenti è assunto essere la realizzazione di una variabile casuale $\mathbf{Y}$ le cui componenti sono indipendentemente distribuite con media $\boldsymbol{\mu}$. La parte sistematica del modello consiste in una specificazione del vettore $\boldsymbol{\mu}$ in termini

di un ristretto numero di parametri $\beta_1 \ldots\ldots \beta_p$. Sia i l'indice per i-esima osservazione, la parte sistematica assume la forma:

$$E(Y_i) = \mu_i = \sum_j a_{ij}\beta_j \qquad i = 1,\ldots..,N .$$

dove a_{ij} è il valore della j-esima covariata per l'osservazione i. In forma matriciale si può scrivere:

$$\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\beta}$$

dove $\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu}$, $\mathbf{A}$ è una matrice detta matrice del disegno, $\boldsymbol{\beta}$ è il vettore dei parametri del modello. Maggiori dettagli su come è costruita la matrice **A,** nel caso in cui le variabili siano categoriche, saranno dati nei paragrafi successivi.

Nei modelli lineari generalizzati si introduce una funzione *link* tra la componente sistematica e quella casuale costituito dalla funzione differenziabile monotona $g(\mu_i)$ tale che $\eta_i = g(\mu_i)$. Il vettore $\boldsymbol{\eta}$ è detto predittore lineare (*linear predictor*). Il modello lega i valori attesi delle osservazioni alle variabili esplicative attraverso la formula:

$$g(\mu_i) = \sum_j a_{ij}\beta_j \qquad i = 1,\ldots..,N .$$

Se si assume che ciascun componente di **Y** ha una ha una funzione di densità appartenente alla famiglia esponenziale, ovvero della forma:

$$f(y_i;\theta_i) = h(\theta_i)b(y_i)e^{[y_i Q(\theta_i)]},$$

dove il termine θ_i per $i = 1,\ldots..,N$ dipende dai valori delle variabili esplicative; mentre il termine $Q(\theta_i)$ è definito *parametro naturale* della distribuzione. Tale famiglia include come casi particolari la distribuzione normale, di Poisson e la binomiale. Infatti se le frequenze osservate in una tabella di contingenza sono trattate come variabili casuali di Poisson

indipendenti sia n_i l'osservazione nella *i*-esima cella e sia $m_i = E(n_i)$ il suo valore atteso per $i = 1, 2,\ldots,N$. La funzione di densità di Poisson:

$$f(n_i;m_i) = \frac{e^{-m_i}(m_i)^{n_i}}{n_i!} = e^{(-m_i)}\left(\frac{1}{n_i!}\right)e^{[n_i \log(m_i)]},$$

per valori interi non negativi di n_i è una forma esponenziale con $y_i = n_i$, $\theta_i = m_i$, $h(\theta_i) = e^{(-m_i)}$, $b(m_i) = \frac{1}{n_i!}$ e $Q(m_i) = \log(m_i)$. Poiché il parametro naturale è il $\log(m_i)$, la funzione di *link* canonica è il logaritmo del *link*: $n_i = \log(m_i)$. Il modello che utilizza questo tipo di *link* è definito ***modello log-lineare***:

$$\log(m_i) = \sum_j a_{ij}\beta_j, \qquad \text{per } i = 1,2,\ldots..,N.$$

Ovvero si applica un modello lineare al logaritmo del valore atteso delle frequenze nelle celle della tabella di contingenza. Si è interessati a vedere come all'interno di ogni cella della tabella cambia il valore del $\log(m_i)$.

Se una delle variabili nel modello è una variabile dipendente Y_i con soltanto due categorie segue una distribuzione di Bernoulli con $\pi_i = \mathrm{E}(Y_i)$, la sua funzione di densità è:

$$\begin{aligned} f(y_i;\pi_i) &= \pi_i^{y_i}(1-\pi_i^{y_i}) = \\ &= (1-\pi_i)\left[\frac{\pi_i}{1-\pi_i}\right]^{y_i} = \\ &= (1-\pi_i)e^{\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right]} \end{aligned}$$

per $y_i = 0$ e 1. Tale distribuzione appartiene alla famiglia esponenziale, il parametro naturale $Q(\pi) = \log\left[\frac{\pi}{1-\pi}\right]$ è definito il *logit* di π. Il modello con tale tipo di *link* è definito ***modello logistico***.

3.2 Modelli log-lineari

L'esposizione che segue ha lo scopo di illustrare i possibili modelli log-lineari nel caso di due e tre variabili classificatorie. Utilizzando inoltre una speciale parametrizzazione per variabili ordinali si mostra l'estensione di tali modelli al caso di più variabili categoriche. Implicitamente nei seguenti sottoparagrafi si introduce il teorema di fattorizzazione per l'indipendenza condizionata in quanto consente di esprimere la probabilità congiunta delle celle in probabilità che non dipendono simultaneamente dalle variabili indipendenti.

Nel capitolo quattro si mostrerà che i modelli esaminati possono essere rappresentati visivamente attraverso dei grafi.

La presente trattazione segue lo schema di Agresti (1990).

3.2.1 Modelli log-lineari per due variabili categoriche

Per un campione multinomiale di dimensione *n,* in una tabella $I \times J$ il set di probabilità $\{\pi_{ij}\}$ forma la distribuzione congiunta di due variabili categoriche; le frequenze attese si denotano: $\{m_{ij} = \pi_{ij} n\}$ In caso di indipendenza, come illustrato al capitolo precedente, vale: $\pi_{ij} = \pi_{i.}\pi_{j.}$ per ogni *i* e *j*. Su scala logaritmica l'indipendenza ha *forma additiva*

$\log m_{ij} = \log n + \log \pi_{i.} + \log \pi_{j.}$, ovvero il logaritmo delle frequenze attese è una funzione additiva dei parametri dell'*i*-esimo effetto riga e del *j*-esimo effetto colonna.

Chiamando la variabile riga X e la variabile colonna Y si può anche scrivere:

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y .$$

Dove

$$\lambda_i^X = \log \pi_{i.} - \left(\frac{\sum_h \log \pi_{h.}}{I} \right)$$

$$\lambda_j^Y = \log \pi_{.j} - \left(\frac{\sum_h \log \pi_{.h.}}{J} \right)$$

$$\mu = \log n + \left(\frac{\sum_h \log \pi_{h.}}{I} \right) + \left(\frac{\sum_h \log \pi_{.h}}{J} \right)$$

I parametri $\{\lambda_i^X\}$ e $\{\lambda_j^Y\}$ soddisfano il vincolo di sommare a zero:

$$\sum \lambda_i^X = \sum \lambda_j^Y = 0 .$$

I vincoli sono necessari per l'identificazione dei parametri e quelli a somma zero costituiscono una delle possibili parametrizzazioni (Goodman). Dei vincoli alternativi (Plakett 1974) sono dati da:

$$\lambda_i^X = \log \pi_{i.} - \log \pi_{1.}$$

$$\lambda_j^Y = \log \pi_{.j} - \log \pi_{.1}$$

così che $\lambda_1^X = \lambda_1^Y = 0$.

Per indicare l'insieme delle variabili che generano il modello e la loro struttura d'interazione si utilizza una *formula*, che si compone di sottoinsiemi d'interazioni massimali. I vari sottoinsiemi sono definiti *generatori* e costituiscono la classe generatrice del modello associato. Il modello precedente viene indicato come **(X,Y)** dato che i coefficienti massimi sono λ_i^X e λ_j^Y.

Il modello **(X,Y)** è un *modello log-lineare di indipendenza* per una tabella di contingenza doppia ed è vero anche il viceversa, ovvero per una tabella $I \times J$ il modello log-lineare **(X,Y)** vale solo se c'è indipendenza tra le variabili. Si può, infatti, scrivere: $m_{ij} = e^{\mu+\lambda_i^X+\lambda_j^Y}$, sia $a = e^{\mu}, a_{X(i)} = e^{\lambda_i^X}, a_{Y(j)} = e^{\lambda_j^Y}$, con $a_{X(.)} = \sum_i a_{X(i)}$, similmente per $a_{Y(.)}$. Notando che:

$$
\begin{aligned}
p_{ij} &= n_{ij}/n = aa_{X(i)}a_{Y(j)}/n \\
p_{i.} &= aa_{X(i)}a_{Y(.)}/n \\
p_{.j} &= aa_{X(.)}a_{Y(j)}/n
\end{aligned}
$$

ed anche

$$1 = p_{..} = aa_{X(.)}a_{Y(.)}/n$$

sostituendo si ha:

$$
\begin{aligned}
p_{i.}p_{.j} &= aa_{X(i)}a_{Y(.)}aa_{X(.)}a_{Y(j)}/n^2 \\
&= \left(aa_{X(i)}a_{Y(j)}/n\right)\left(aa_{X(.)}a_{Y(.)}/n\right) \\
&= aa_{X(i)}a_{Y(j)}/n \\
&= p_{ij}.
\end{aligned}
$$

Così il modello log-lineare **(X,Y)** implica l'indipendenza.

Nel caso vi sia dipendenza tra le variabili invece il modello diventa **(XY)**:

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

Il modello **(XY)** è definito *modello saturo* per una tabella di contingenza doppia e riproduce perfettamente ciascun insieme di frequenze osservate. L'equazione ha forma assai simile a quella dell'analisi della varianza dove le deviazioni dai valori medi delle frequenze di cella possono essere interpretati in termini di 'effetti marginali e d'interazione'.

Con la prima parametrizzazione adottata i vincoli simmetrici sono:

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$$

da cui risulta che ci sono (I-1) e (J-1) parametri $\left\{\lambda_i^X\right\}$, $\left\{\lambda_j^Y\right\}$ linearmente indipendenti riga e colonna rispettivamente. I termini degli 'effetti marginali' hanno l'interpretazione:

$$\lambda_i^X = \frac{\sum_j \log m_{ij}}{J} - \frac{\sum_i \sum_j \log m_{ij}}{IJ} \quad \text{e} \quad \lambda_j^Y = \frac{\sum_i \log m_{ij}}{I} - \frac{\sum_i \sum_j \log m_{ij}}{IJ}$$

mentre il termine di interazione può scriversi come:

$$\lambda_{ij}^{XY} = \log m_{ij} - \frac{\sum_j \log m_{ij}}{J} - \frac{\sum_i \log m_{ij}}{I} + \frac{\sum_i \sum_j \log m_{ij}}{IJ}.$$

Ci sono (I-1)(J-1) parametri λ_{ij}^{XY} linearmente indipendenti che corrispondono ai gradi di libertà del *modello di indipendenza*. Nel *modello saturo* i parametri linearmente indipendenti sono: 1 + (I-1) + (J-1) + (I-1)(J-1) =IJ, ovvero il numero dei parametri nel modello saturo uguaglia il numero delle celle della tabella corrispondente.

L'utilità di tali modelli risiede nel fatto che i parametri sono interpretabili in termini di *odds ratio*. Si dimostra, ad esempio che per una tabella 2× 2 il parametro

$$\lambda_{11}^{XY} = \frac{\log\left[\left(m_{11}m_{22}\right)/\left(m_{12}m_{21}\right)\right]}{4}$$

è un quarto del logaritmo dell'*odds ratio*[20].

Se si adotta la seconda parametrizzazione, utilizzando il livello uno di ogni variabile come base, i vincoli diventano: $\lambda_1^X = \lambda_1^Y = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$. Il modello saturo è lo stesso e si perviene alle stesse conclusioni sulle relazioni tra le variabili, ma l'interpretazione dei parametri è diversa, ad esempio il parametro d'interazione corrisponde direttamente all'*odds ratio*:

$$\lambda_{ij}^{XY} = \log \frac{m_{11} m_{ij}}{m_{1j} m_{i1}} .$$

3.2.2 Modelli log-lineari per tre variabili categoriche

Il *modello saturo* log-lineare per una *tabella tripla* con X, Y e Z variabili con $i = 1,2,\ldots..,I$, $j = 1,2,\ldots..,J$, e $k = 1,2,\ldots\ldots,K$ categorie rispettivamente è indicato come:

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} .$$

La notazione utilizzata per tale modello è **(XYX)**. Ponendo alcuni di questi parametri uguale a zero si ottengono dei modelli che permettono di testare varie ipotesi sulla struttura delle relazioni tra le variabili.

Un possibile modello per la distribuzione condizionata delle variabili X, Y, Z, è il modello **(XY,YZ,XZ)**:

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} ,$$

chiamato modello con *assenza del fattore di interazione tripla*. I termini d'associazione parziale sono presenti per ogni coppia di variabili e ciò

[20] Per la dimostrazione si rimanda ad Agresti (1990), op. cit. pp133-134.

comporta che nessuna coppia è condizionatamente indipendente. In tale modello gli *odds ratio* condizionati tra due variabili sono identici per ogni livello della terza variabile. E' assente, infatti, il parametro d'interazione tripla dato dalla differenza dei logaritmi dei rapporti degli *odds* tra due variabili al variare della terza variabile. Sostituendo l'espressione del modello in quella per l'*odds ratio* condizionato $\theta_{ij(k)}$ come definito nel secondo capitolo, si ottiene:

$$\log\theta_{ij(k)} = \lambda_{ij}^{XY} + \lambda_{i+1,j+i}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}.$$

Dato che il lato destro dell'espressione è lo stesso per ogni k, l'assenza del fattore d'interazione tripla equivale a:

$$\theta_{ij(1)} = \theta_{ij(2)} = ... = \theta_{ij(K)} \qquad \text{per ogni } i, j.$$

Considerazioni analoghe per gli *odds ratio* parziali implicano che:

$$\theta_{i(1)k} = \theta_{i(2)k} = \mathbf{.....} = \theta_{i(J)k} \qquad \text{per ogni } i,k$$

ed anche:

$$\theta_{(1)jk} = \theta_{(2)jk} = ... = \theta_{(I)jk} \qquad \text{per ogni } j,k\ .$$

Pertanto nel modello **(XY,YZ,XZ)** l'associazione parziale tra due coppe di variabili è identica per ogni livello della terza variabile. Tuttavia tale associazione può essere diversa dall'associazione marginale corrispondente, ad esempio, l'associazione parziale tra X-Y, dato Z, può essere diversa dall'associazione marginale tra X-Y visto che Z è condizionatamente dipendente sia da X che da Y.[21]

Un modello d'interesse che ha una coppia di indipendenze condizionate è il seguente:

[21] Agresti A. (1984), op. cit., p.41.

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{YZ},$$

in cui X è indipendente da Z condizionatamente Y, infatti è assente il parametro d'interazione tripla, per cui il rapporto degli *odds* tra X e Z non cambia al variare di Y e manca il parametro d'interazione tra X e Z che specifica l'indipendenza condizionata tra queste due. Questo modello viene indicato come **(XY,YZ)**, i parametri $\{\lambda_{ij}^{XY}\}$ e $\{\lambda_{ij}^{YZ}\}$ corrispondono all'associazione parziale tra X-Y e Y-Z. Tuttavia benché X e Z siano condizionatamente indipendenti per ogni livello di Y, possono essere marginalmente dipendenti.

Nel caso in cui vi sia indipendenza congiunta di Z da X e Y il modello diventa **(XY,Z)**:

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}.$$

I parametri $\{\lambda_{ij}^{XY}\}$ si riferiscono all'associazione parziale tra X e Y dato Z. Il modello analogo per le probabilità di cella è il seguente: $\pi_{ijk} = \pi_{i.k}\pi_{.j.}$, ovvero X e Z sono condizionatamente indipendenti dato Y, allo stesso modo Y e Z sono condizionatamente indipendenti dato X. Ci sono tre tipi di modelli di questo tipo che corrispondono alle tre coppie di variabili che possono essere condizionatamente indipendenti: (**XY,Z**), **(YZ,X)** oppure **(XZ,Y)**.

Se nel modello (**XY,Z**) vale che tutti i $\lambda_{ij}^{XY} = 0$, anche le variabili X e Y sono condizionatamente indipendenti e il modello diventa di mutua indipendenza tra le variabili **(X,Y,Z)**:

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

In termini di probabilità di cella il modello può essere riformulato come: $\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$. Ciascuna coppia di variabili è marginalmente e condizionatamente indipendente date le restanti.

I modelli con indipendenze condizionate presentati in precedenza rientrano nella classe dei **modelli gerarchici.** Il termine gerarchico indica che se un parametro relativo all'interazione tra un certo insieme di variabili è incluso nel modello, devono esservi inclusi anche i parametri relativi a interazioni di livello inferiore delle stesse variabili. Ad esempio se si pone $\lambda_{ij}^{XY} = 0$ ma il termine di inerazione tripla λ_{ijk}^{XYZ} non è posto uguale a zero il modello risultante è un *modello log-lineare non gerarchico*. In pratica i modelli gerarchici sono quei modelli che si possono ordinare a partire dal cosiddetto modello saturo, ovvero senza alcun vincolo, fino al modello di indipendenza di ciascun carattere da tutti gli altri. Una sottoclasse dei modelli gerarchici definita modelli grafici può essere rappresentata univocamente attraverso i grafi, che mostrano la struttura di dipendenza tra le variabili come descritto al quarto capitolo.

3.2.3 Speciale parametrizzazione per variabili ordinali

Utilizzando la notazione matriciale i modelli log-lineari possono essere riformulati nel seguente modo. Considerando una tabella di contingenza di qualsiasi dimensione avente q celle, nel caso di una tabella $I \times J \times K$, $q = IJK$, le frequenze attese delle celle sono definite da un vettore $\boldsymbol{m} = (m_1, ..., m_q)'$. Come illustrato in precedenza un modello log-lineare è un modello del tipo:

$$\log(\boldsymbol{m}) = \mathbf{A}\,\ddot{e}\ .$$

Dove log(**m**) è il vettore $q \times 1$ delle frequenze, **A** è una matrice $p \times q$ con rango(**A**) = p e $\boldsymbol{\lambda}$ è un vettore $p \times 1$ di parametri incogniti. Un modo equivalente per definire tali modelli è quello di specificare una matrice **C** tale che $\mathbf{C}\boldsymbol{\mu} = \boldsymbol{\lambda}$ con $\boldsymbol{\mu} = \log(\boldsymbol{m})$. Da questo si deriva la matrice $\mathbf{A} = \mathbf{C}^{-1}$ infatti $\boldsymbol{\mu} = \mathbf{C}^{-1}\boldsymbol{\lambda}$. Nel caso in cui la dimensione del campione n sia fissata, essendo $\boldsymbol{m} = n\,\boldsymbol{\pi}$ si può ugualmente modellare il vettore log $\boldsymbol{\pi}$ oppure il vettore log $\boldsymbol{m}$**.** La matrice **C** è definita matrice dei contrasti perché specifica i vincoli che devono essere soddisfatti dal modello, Le matrici dei contrasti sono diverse a seconda dei pesi usati per definire i contrasti e così che anche i parametri del modello hanno una diversa interpretazione. Esse vengono costruite con il prodotto di Kronecker tra matrici più piccole che contengono dei codici per l'effetto generale e gli effetti principali del modello[22].

Una speciale parametrizzazione per variabili ordinali che permette di formulare ipotesi d'indipendenza di vario tipo è quella proposta da Wermuth e Cox (1998).[23] Adottando tale paramertizzazione le matrici **C** vengono costruite nel modo seguente. Sia ad esempio X una variabile categorica con

[22] Per un confronto tra codifiche alternative si veda: Wermuth (1992), 'On the Relation Between Interactions Obtained with Alternative Codings of Discrete Variables' , Methodika 6, pp. 76-85.

[23] Wermuth N. e Cox D.R. (1998), 'On the Application of Conditional Indipendence to Ordinal Data', International Statistical Review, 66, pp. 181-199.

tre modalità e Y una variabile categorica con due modalità $\underset{(6\times6)}{\mathbf{C}_{32}} = \underset{(2\times2)}{C_2} \otimes \underset{(3\times3)}{C_3}$

ovvero:

$$\underset{(6\times6)}{\mathbf{C}_{32}} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & -1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}$$

I parametri del modello log-lineare possono essere ottenuti da:

$$\hat{\boldsymbol{\eta}}^{XY} = \mathbf{C}_{32} \log \boldsymbol{\delta}^{XY},$$

ovvero:

$$\begin{pmatrix} \lambda_{-}^{XY} \\ \lambda_{12}^{X} \\ \lambda_{23}^{XY} \\ \lambda_{12}^{Y} \\ \lambda_{12.12}^{XY} \\ \lambda_{23.12}^{XY} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & -1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \log\pi_{11} \\ \log\pi_{21} \\ \log\pi_{31} \\ \log\pi_{12} \\ \log\pi_{22} \\ \log\pi_{32} \end{pmatrix}$$

da cui il parametro di interazione $\lambda_{23.12}^{XY}$ è definito come:

$$\lambda_{23.12}^{XY} = \log\pi_{21} - \log\pi_{31} - \log\pi_{22} + \log\pi_{32} = \log\frac{\pi_{21}\pi_{32}}{\pi_{31}\pi_{22}}$$

ovvero è il logaritmo dell'*odds ratio locale* nella sottotabella 2×2 che considera i livelli 2,3 di X e i livelli 1,2 di Y.

Attraverso tale parametrizzazione numerose ipotesi d'indipendenza possono essere sottoposte a verifica ponendo uguale a zero i logaritmi degli *odds ratio* nelle sottotabelle 2×2. Nell'esempio precedente:

- c'è indipendenza nelle sottotabella 2×2 tra i livelli 2,3 da X e i livelli 1,2 di Y se il parametro $\lambda_{23.12} = 0$;
- c'è indipendenza nella sottotabella 3×2 tra tutti i livelli di X e i livelli 1,2 di Y quando $\lambda_{12.12} = \lambda_{23.12} = 0$;
- c'è indipendenza come sopra con in più indipendenza nella sottotabella 2×3 tra tutti i livelli di Y e i livelli 1,2 di X se vale che $\lambda_{12.12} = \lambda_{23.12} = \lambda_{12.23} = 0$;
- la variabile X è indipendente dalla variabile Y cioè: $\lambda_{12.12} = \lambda_{23.12} = \lambda_{12.23} = \lambda_{23.23} = 0$.

Per le variabili con più di due categorie e considerando il vettore delle probabilità con i livelli della variabile X che si alternano più rapidamente cioè:

$$\pi^{XY} = \left(\pi_{11},\pi_{21},...,\pi_{I1},\pi_{12},...,\pi_{I2},\pi_{13},...,\pi_{IJ}\right)',$$

i parametri del modello log-lineare sono generati secondo il seguente ordine lessicografico:

$$\lambda_{-}^{XY},\lambda_{12}^{X},\lambda_{23}^{X},....,\lambda_{I-1,I}^{X},\lambda_{12}^{Y},\lambda_{12,12}^{XY},....\lambda_{I-1,I.12}^{XY},\lambda_{13}^{Y},\lambda_{12.13}^{XY},....\lambda_{I-1,I.J-1,J}^{XY}.$$

In generale $\lambda_{ii'.jj'}^{XY}$ indica il logaritmo dell'*odds ratio* nella tabella 2×2 dei livelli i,i' di X e j,j' di Y.

Per più di due variabili c'è un'estensione diretta. Per il modello log-lineare saturo in cui l'unico vincolo è dato dal fatto che le probabilità di cella devono sommare ad uno si può scrivere nel caso di tre variabili X, Y e Z:

$$\hat{\mathbf{e}}^{XYZ} = \mathbf{C}_{IJK}\log\mathbf{d}^{XZY};$$

dove la matrice dei contrasti è ottenuta nel seguente modo:

$$\mathbf{C}_{IJK} = C_K \otimes C_J \otimes C_I (= C_K \otimes C_{IJ} = C_{JK} \otimes C_I) .$$

Il parametro di interazione $\lambda^{XY}_{ii'.jj'}$ può essere interpretato come una media degli *odds ratio* locali condizionati ottenuti dalla sottotabella 2×2 dei livelli i,i' di X e j, j' di Y calcolati per ogni livello di X. Invece il parametro $\lambda^{XYZ}_{ii',jj',kk'}$ misura la differenza tra gli *odds ratio* locali condizionati in relazione alla tabella 2×2 dei livelli i,i' di X e j, j' di Y per i livelli di Z k e k' .

I vari modelli con indipendenze condizionate possono essere stimati ponendo alcuni dei parametri $\hat{\lambda}$ uguali a zero e quindi eliminando le colonne corrispondenti della matrice **C**.

Tale parametrizzazione è particolarmente utile quando si vuole unire due o più categorie adiacenti di una variabile ordinale. Siano X e Y due variabili ordinali con I e J livelli rispettivamente. Alcune delle possibili relazioni di indipendenza che possono essere formulate sono le seguenti:

- ♣ $X_i \perp Y_j$ per $i = i'$, i'' e j = 1,2,…..,J. C'è indipendenza in ciascuna delle (J-1) sottotabelle 2×2 relative alle categorie i', i'' di X e 1,2,….,J di Y. Di conseguenza la riga i, i' possono essere unite quando si considera B;
- ♣ $X_i \perp Y_j$ per i = 1, 2,…..I e $j = j', j''$. In modo analogo al precedente c'è indipendenza in ogni (I-1) sottotabella 2×2 delle categorie 1,2,….,I di X e j', j'' di Y. Per cui la relazione tra le due variabili può essere studiata nella tabella ottenuta unendo le colonne j' e j'';
- ♣ $X_i \perp Y_j$ per i = 1, 2,….,I e j = 1,2,…..,J. C'è indipendenza in ciascuna delle (J-1)(I-1) sottotabelle 2×2 da cui risulta che X ⊥ Y.

Se si considerano più di due variabili ad esempio X, Y e Z variabili ordinali con I, J e K livelli rispettivamente e sia Y la variabili risposta, alcune delle relazioni di indipendenza condizionata sono:

- ♣ $X \perp Z|Y$ insieme a $X_i \perp Y_k$ per $i = i', i''$ e $k = 1,2,\ldots K$. per ogni livello di Y c'è indipendenza nelle (K-1) sottotabelle 2×2 tra le categorie i', i'' di X e 1,2,….,K di Z. In pratica distinguere tra i livelli i', i'' di X non fornisce informazioni ulteriori quando si studia Z;
- ♣ $X \perp Z|Y_j$ per $j = j', j''$ insieme a $X_i \perp Z_k$ per $i = i'$, i'' e $k = 1,2,\ldots K$. La relazione precedente di indipendenza condizionata ora si applica soltanto per i livelli j', j'' di Y.

3.3 Modello logistico

Quando la tabella di contingenza viene formata incrociando una variabile dipendente (o risposta) con una o più variabili esplicative è interessante studiare come queste ultime influiscono sulla prima più che le relazioni tra le variabili esplicative stesse. Piuttosto che modellare il logaritmo delle frequenze attese delle celle, o delle probabilità come nei modelli log-lineari, si costruisce pertanto un modello per il logaritmo degli *odds* (il *logit*) della variabile risposta. Nel caso in cui la variabile risposta sia binaria c'è un unico modo di formare l'*odds*. Se si esprime il logaritmo di tali *odds* in funzione delle variabili esplicative e delle loro interazioni, il modello risultante è un ***modello logistico*** come introdotto nel paragrafo 3.1.

Sia Y la variabile risposta binaria con valori 0 e 1 ed X una variabile casuale continua, si indica E(Y) = P(Y=1) come $\pi(x)$ che riflette la dipendenza di Y della variabile esplicativa. Se vale la seguente relazione tra x e $\pi(x)$:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

per $\beta < 0$ quando $x \to \infty$, $\pi(x) \downarrow 0$, invece se $\beta > 0$ $\pi(x) \uparrow 1$, se il modello per $\beta = 0$ è accettabile, la variabile Y è indipendente dalla variabile esplicativa. L'*odds* di avere la categoria 1 della variabile Y è:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}\left(e^{\beta}\right)^{x};$$

l'*odds* cresce di e^{β} unità per ogni incremento unitario in x. Il modello diventa lineare attraverso il logaritmo dell'*odds*:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x .$$

L'equazione del modello esplicita e quantifica attraverso i parametro β, l'influenza delle variabile indipendente sulla dipendente.

Nel caso in cui la variabile esplicativa sia categorica con I categorie, nella riga i della tabella $I \times 2$, le due probabilità di risposta sono $\pi_{1/i}$ e $\pi_{2/i}$, con $\left(\pi_{1/i} + \pi_{2/i}\right) = 1$. Il modello logistico è il seguente:

$$\log\left(\frac{\pi_{1/i}}{\pi_{2/i}}\right) = \alpha + \beta_i ,$$

dove $\{\beta_i\}$ descrive gli effetti del fattore esplicativo sulla variabile risposta.

Anche in questo caso come per il modello log-lineare sono necessari dei vincoli per identificare i parametri, questi possono essere $\sum \beta_i = 0$ oppure $\beta_1 = 0$, così che i parametri β_i linearmente indipendenti sono (I-1). Tanto più è grande β_i, tanto maggiore è il *logit* nella riga i e tanto più è grande $\pi_{1/i}$, ovvero i soggetti nella categoria i hanno maggiore probabilità di avere la 1 anziché la due come categoria di risposta.

Quando un fattore non ha influenza sulla variabile risposta, il modello logistico diventa il modello di indipendenza statistica tra le variabili:

$$\log\left(\frac{\pi_{1/i}}{\pi_{2/i}}\right) = \alpha \ ,$$

dove $\beta_1 = \beta_2 = ... = \beta_I = 0$.

Nel caso in cui le variabili esplicative siano più di una, ad esempio X e Z con *I* e *K* categorie rispettivamente, per la tabella $I \times J \times 2$ sia $\pi_{1/ik}$ la probabilità delle risposta 1 per il livello *i* di X e *k* di Z così che $\pi_{1/ik} + \pi_{2/ik} = 1$, il modello logistico è il seguente:

$$\log\left(\frac{\pi_{1/ik}}{\pi_{2/ik}}\right) = \alpha + \beta_i + \beta_k \ ,$$

in cui gli effetti di X sono rappresentati dai parametri $\{\beta_i\}$ e gli effetti di Z da $\{\beta_k\}$. Il modello assume che gli effetti di ogni fattore siano gli stessi per ogni livello dell'altro fattore ovvero assume l'assenza d'interazione tripla tra le variabili, corrisponde quindi al modello log-lineare **(XZ,XY,ZY)**[24]. Questo modello tratta le osservazioni $\{n_{ik.}\}$ come fisse, mentre $\{n_{ik1}\}$ come variabili casuali binomiali indipendenti con parametri $\{\pi_{1/ik}\}$.

Il modello logistico corrispondente al modello log-lineare **(ZY,XZ)** in cui Z ha effetti su Y ma X è indipendente da Y condizionatamente a Z è il seguente:

$$\log\left(\frac{\pi_{1/ij}}{\pi_{2/ij}}\right) = \alpha + \beta_i \ .$$

I risultati si estendono facilmente per la tabella incrociata di Y con le variabili esplicative X, Z e W aventi *I*, *K* e *H* categorie rispettivamente. Il modello logistico

[24] Per la dimostrazione si veda Agresti A. (1990), op.cit. pp. 152-153.

$$\log\left(\frac{\pi_{1/ikh}}{\pi_{2/ikh}}\right) = \alpha + \beta_i + \beta_k + \beta_h + \beta_{hi}$$

corrisponde al modello log-lineare **(XZW,XZY,WY)** che contiene tutti i termini d'interazione tra le variabili esplicative e quelli d'associazione di ogni variabile con la risposta Y, così come l'effetto congiunto di X e Z su Y dato dal parametro β_{hi}. Se quest'ultimo parametro è assente ed anche il parametro β_h non è nel modello, si dimostra che il modello log-lineare corrispondente è **(XZW,XY,ZY)** in cui Y è indipendente da W condizionatamente a X e Z.

In sintesi, quando c'è una variabile risposta binaria, vi è una corrispondenza biunivoca fra i modelli logistici e i modelli log-lineari. Per cui i metodi di stima dei modelli sono gli stessi dei modelli log-lineari così come i valori stimati, il rapporto di verosimiglianza ed i gradi di libertà[25].

3.3.1 Modello logistico per un fattore con risposta multinomiale

Quando la variabile risposta ha più di due categorie i modelli log-lineari corrispondono ai ***modelli logistici generalizzati***. Se la variabile risposta Y ha *J* categorie ci sono $\binom{J}{2}$ coppie di risposte con le quali si possono costruire i *logit*. In tal caso si cerca di fare in modo che due cose alla volta siano confrontate. Un criterio per farlo è identificare le coppie di livelli da confrontare, così che si avranno tante equazioni logistiche quante il numero delle categorie della variabile risposta meno uno. Per una tabella di contingenza tripla con X e W variabili esplicative si denota $\pi_{j/ih}$ la probabilità della risposta *j* al livello *i* di X e *h* di W. Un modello logistico equivalente al

[25] Agresti A. (1990), op. cit. p.178.

modello log-lineare **(XW,XY,WY)** con assenza del fattore di interazione tripla può essere formulato considerando una categoria della variabile risposta come riferimento. Considerando, ad esempio, l'ultima, il *j*-esimo *logit* per il modello è il seguente:

$$\log\left(\frac{\pi_{j/ih}}{\pi_{J/ih}}\right) = \log\left(\frac{m_{ihj}}{m_{ihJ}}\right) = \alpha_j + \beta_{ij} + \beta_{hj}$$

per $j = 1,\ldots,J\text{-}1$, dove sia l'intercetta che i parametri β dipendono da *j* e gli effetti di A e di B sono additivi. I parametri del modello determinano anche quelli dei modelli che usano altre coppie di categorie della variabile risposta. Ad esempio:

$$\log\left(\frac{\pi_{a/ih}}{\pi_{b/ih}}\right) = \log\left(\frac{\pi_{a/ih}}{\pi_{J/ih}}\right) - \log\left(\frac{\pi_{b/ih}}{\pi_{J/ih}}\right)$$

allora l'effetto della categoria *i* di A sul logaritmo dell'*odds* per la classificazione nella categoria *a* piuttosto che *b* di Y è dato da dalla differenza dei beta $\beta_{ia} - \beta_{ib}$. Per l'identificabilità dei parametri è necessario porre dei vincoli lineari come visto in precedenza.

Quando le categorie della variabile risposta sono ordinate il modello logistico può essere formulato costruendo dei *logit* che tengano conto di tale ordine. Per far questo non è necessario usare solo due categorie alla volta della variabile risposta, i *logit* infatti possono essere formati raggruppando le categorie contigue della scala ordinale. Ad esempio si possono formare dei *logit* di probabilità cumulate. Siano $\{\pi_1(x),\ldots,\pi_J(x)\}$ le probabilità di risposta per un set *x* di variabili esplicative. Utilizzando le probabilità cumulate per la categoria *j* della variabile risposta Y quando le variabili esplicative assumono il valore *x*

$$F_j(x) = P(Y \leq j/x) = \pi_1(x) + \ldots\ldots + \pi_j(x), \qquad j = 1,\ldots\ldots J$$

si possono costruire i *logit* in questo modo:

$$L_j(x) = \text{logit}\left[F_j(x)\right] = \log\left(\frac{F_j(x)}{1-F_j(x)}\right)$$
$$= \log\left(\frac{\pi_1(x)+.....+\pi_j(x)}{\pi_{j+1}(x)+.....+\pi_J(x)}\right), \quad j = 1,....,J-1$$

definiti come *cumulative logit* o *logit* cumulativi. Ciascun *logit* cumulativo utilizza tutte le *J* categorie della variabile risposta. Praticamente tale modello è logistico come quelli precedenti per una variabile risposta binaria in cui le categorie dalla prima alla *j*-esima formano la prima categoria e le categorie dalla *j*+1 alla *J* formano la seconda categoria.

Diversamente dagli altri modelli logistici i modelli che utilizzano tale tipo di *logit* non sono equivalenti ai modelli log-lineari.

Il modello logistico cumulativo può essere formulato nel seguente modo:

$$L_j(x) = \log\left[\frac{P(Y \le j/x)}{P(Y < j/x)}\right] = \alpha_j - \beta' x \qquad j = 1,2,.....J-1$$

α_j sono le intercette sconosciute dei parametri dette *cutpoints*, che soddisfano la condizione $\alpha_1 \le \alpha_2 \le .... \le \alpha_J$, dato che il *logit* cumulativo è funzione crescente di $F_j(x)$ che è essa stessa crescente in *j* per dato *x* e β è il vettore dei parametri che esprimono l'effetto di *x*. Occorre notare che il vettore dei coefficienti di regressione β non dipende da *j*,implicando che il modello assume che la relazione tra le variabili esplicative e Y è indipendente da *j*. Il modello soddisfa l'assunzione di *odds* proporzionali per tale ragione è definito anche *proportional odds model*. L'*odds ratio* delle probabilità cumulate è definito come:

$$\left[\frac{P(Y \le j/x_1)/P(Y > j/x_1)}{P(Y \le j/x_2)/P(Y > j/x_2)}\right],$$

odds ratio cumulativo. Il logaritmo di tale *odds* è proporzionale alla distanza tra i valori delle variabili esplicative, con la stessa costante proporzionale applicata ad ogni *cutpoint*, infatti:

$$L_j(x_1) - L_j(x_2) = \log\left[\frac{P(Y \le j/x_1)/P(Y > j/x_1)}{P(Y \le j/x_2)/P(Y > j/x_2)}\right] = \beta(x_2 - x_1).$$

Può essere interpretato dicendo che l'*odds* di avere una risposta $\le j$ è $e^{[\beta'(x_2 - x_1)]}$ volte più grande per $x = x_1$ che per $x = x_2$.

Il modello più semplice assume che la variabile risposta è simultaneamente indipendente da tutte la variabili esplicative:

$$L_j(x) = \alpha_j, \qquad j = 1,2.....,J-1.$$

Il modello logistico cumulativo è molto utilizzato nel caso di variabile risposta ordinale, mentre le variabili esplicative possono essere oltre che categoriche anche continue. Tale modello, infatti, presenta delle proprietà interessanti: resta invariato se l'ordine delle categorie di Y è invertito, risultando invertito il segno dei parametri di regressione; è inoltre invariante se le categorie della variabile risposta sono collassate o unite, in tale caso i coefficienti β restano gli stessi mentre i valori delle intercette cambiano.[26]

[26] Greenland S. (1994), 'Alternative models for ordinal logistic regression', *Stat Med*, 13, pp.1665-77.

3.4 Stime dei modelli di indipendenza condizionata

Una volta selezionato un modello log-lineare di interesse si usano i dati del campione per stimare i parametri del modello e di conseguenza le probabilità di cella e le frequenze attese. Se si assume che le osservazioni della tabella incrociata $\{n_{ijk}\}$ di tre variabili X, Y e Z siano variabili casuali di Poisson indipendenti con valore atteso $\{m_{ijk}\}$, la probabilità di osservare il campione può scriversi come:

$$\prod_i \prod_j \prod_k \frac{e^{-m_{ijk}} m_{ijk}{}^{n_{ijk}}}{n_{ijk}!}$$

dove le produttorie definiscono il prodotto fra le celle della tabella.
La funzione di verosimiglianza è la probabilità di avere $\{n_{ijk}\}$ per un determinato schema di campionamento, dove i valori osservati sono trattati come funzione di parametri incogniti. Le stime di massima verosimiglianza (ML) sono i valori dei parametri che massimizzano la probabilità di osservare il campione estratto. Poiché occorre massimizzare l'espressione precedente in relazione ad $\{m_{ijk}\}$ il termine fattoriale è costante e non ha influenza sul punto di massimo, per cui può essere trascurato e si può scrivere il logaritmo della verosimiglianza come: $L(m) = \sum_i \sum_j \sum_k n_{ijk} \log(m_{ijk}) - \sum_i \sum_j \sum_k m_{ijk}$.

Per il modello saturo della tabella incrociata $I \times J \times K$ la verosimiglianza diventa l'espressione seguente:

$$\begin{aligned} L(m) = n\mu &+ \sum_i n_{i..}\lambda_i^X + \sum_j n_{.j.}\lambda_j^Y + \sum_k n_{..k}\lambda_k^Z \\ &+ \sum_i \sum_j n_{ij.}\lambda_{ij}^{XY} + \sum_i \sum_k n_{i.k}\lambda_{ik}^{XZ} + \sum_j \sum_k n_{.jk}\lambda_{jk}^{YZ} \\ &+ \sum_i \sum_j \sum_k n_{ijk}\lambda_{ijk}^{XYZ} \\ &- \sum_i \sum_j \sum_k \exp\left[\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}\right]. \end{aligned}$$

Poiché la distribuzione di Poisson appartiene alla famiglia esponenziale si dimostra i coefficienti dei parametri del logaritmo della verosimiglianza sono statistiche sufficienti[27]. Per il modello saturo i coefficienti di $\{\lambda_{ijk}^{XYZ}\}$ sono $\{n_{ijk}\}$.

Per ottenere le stime di massima verosimiglianza di un modello log-linare vincolato occorre massimizzare l'espressione della verosimiglianza con i vincoli imposti dal modello. Per alcuni modelli si dimostra che gli stimatori di ML delle probabilità di cella sono dati dalle proporzioni campionarie delle celle. Ad esempio per il modello **(XZ,YZ)**:

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XZ} + \lambda_{ik}^{YZ},$$

in cui **X** e **Y** sono condizionatamente indipendenti dato **Z** come illustrato al paragrafo 3.2.2, il logaritmo della verosimiglianza si semplifica nel seguente modo:

$$\begin{aligned} L(m) = n\mu &+ \sum_i n_{i..}\lambda_i^X + \sum_j n_{.j.}\lambda_j^Y + \sum_k n_{..k}\lambda_k^Z \\ &+ \sum_i \sum_k n_{i.k}\lambda_{ik}^{XZ} + \sum_j \sum_k n_{.jk}\lambda_{jk}^{YZ} \\ &- \sum_i \sum_j \sum_k \exp\left[\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}\right]. \end{aligned}$$

Per ottenere le stime occorre fare la derivata di *L(m)* per il parametro μ e per i parametri λ ed uguagliarle a zero così da ottenere il punto di massimo. Ad esempio:

$$\begin{aligned} \frac{\partial L}{\partial \mu} &= n - \sum_i \sum_j \sum_k \exp\left[\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}\right] \\ &= n - \sum_i \sum_j \sum_k m_{ijk}, \end{aligned}$$

[27] Azzalini A. (1996), 'Statistical Inference', Chapman Hall, pp30-41.

uguagliando a zero l'espressione precedente si ottiene l'equazione di verosimiglianza: $\hat{m}_{...} = n$. Così che i valori stimati hanno il totale pari alle frequenze osservate.

Per il parametro λ_i^X

$$\frac{\partial L}{\partial \lambda_i^X} = n_{i..} - \sum_j \sum_k m_{ijk}$$
$$= n_{i..} - m_{i..} \quad ,$$

da cui si ottengono le equazioni $\left\{\hat{m}_{i..} = n_{i..},\ i = 1,2,.....,I\right\}$. Allo stesso modo differenziando rispetto a λ_i^Y e rispetto a λ_k^Z si ottiene l'uguaglianza a livello marginale fra le frequenze osservate e valori stimati.

$$\frac{\partial L}{\partial \lambda_{ik}^{XZ}} = n_{i.k} - m_{i.k} \quad \text{e} \quad \frac{\partial L}{\partial \lambda_{jk}^{YZ}} = n_{.jk} - m_{.jk}$$

da cui le equazioni di verosimiglianza: $\hat{m}_{i.k} = n_{i.k}$ per ogni i e k e $m_{.jk} = n_{.jk}$ per ogni j e k.

In pratica i valori stimati hanno gli stessi totali marginali dei valori osservati. Dato che le ultime due equazioni implicano le precedenti, queste ultime due determinano le stime di massima verosimiglianza. Birch[28] ha dimostrato che c'è un' unica soluzione per le equazioni di verosimiglianza e che tali stime sono identiche sia nel caso di un campione multinoniamle sia nel caso di un campionamento di Poisson. Nel modello **(XZ,YZ)** vale la formula $\pi_{ijk} = \frac{\pi_{i.k}\pi_{.jk}}{\pi_{..k}}$. Le statistiche minimali sufficienti per i parametri λ_{ik}^{XZ} e λ_{jk}^{YZ} sono $\left\{n_{i.k}\right\}$e $\left\{n_{.jk}\right\}$. I valori stimati sono dati da: $\hat{m}_{ijk} = \frac{\hat{m}_{i.k}\hat{m}_{.jk}}{\hat{m}_{..k}} = \frac{n_{i.k}n_{.jk}}{n_{..k}}$.

[28] Birch, M.W. (1963), 'Maximum likelihood in the three-way contingency table', Journal of Royal Statistic Society, B25, pp.220-223

Tuttavia vi sono modelli per i quali le equazioni di verosimiglianza non ammettono soluzioni esplicite. Di questi si parla nel seguente paragrafo.

3.4.1 Algoritmi di stima

Molto spesso i modelli log-lineari non hanno stime dirette ovvero non si hanno formule esplicite per calcolare $\{\hat{m}_{ijk}\}$. In questo caso le equazioni di verosimiglianza devono essere risolte iterativamente. Ad esempio, per il modello con assenza del fattore di interazione tripla benché le tabelle marginali doppie sono statistiche sufficienti per il modello, non si può esprimere direttamente $\{\pi_{ijk}\}$ in termini di $\{\pi_{ij.}\},\{\pi_{i.k}\}$e $\{\pi_{.jk}\}$. I metodi di stima iterativa possono essere usati oltre che per i modelli che non hanno stime dirette anche per quelli che le hanno.
Ci sono due algoritmi standard che vengono utilizzati *l'Algoritmo di Newton-Raphson* e *l'Algoritmo di Stima Proporzionale Iterativa* (*Iterative Proportional Fitting Algorithm*).

L'algoritmo di stima proporzionale iterativa proposto da Deming e Stephan (1940) consiste in una procedura di adattamento successivo dei marginali della tabella ai vincoli prescritti, modificando di volta in volta le probabilità fino a convergenza:

- le stime iniziali per $\{\hat{m}_i^{(0)}\}$ hanno una struttura di associazione meno complessa di quella del modello che occorre stimare, in genere $\{\hat{m}_i^{(0)} = 1\}$;
- moltiplicando poi per dei fattori scalari appropriati si adattano $\{\hat{m}_i^{(0)}\}$ in modo tale che uguaglino le tabelle marginali per il set delle statistiche sufficienti;
- il processo continua fintanto che la differenza tra i valori stimati ad ogni passo è sufficientemente vicina a zero.

Il motivo per cui le stime prodotte con tale metodo sono di massima verosimiglianza è che esso genera una sequenza di valori stimati che ad ogni passo fa incrementare il valore della log-verosimiglianza.

L'algoritmo di Newton-Rapson può essere illustrato introducendo la seguente notazione. Sia $f(\lambda)$ una funzione del vettore λ, tale metodo si applica per trovare $f(\lambda) = 0$. Si parte da un valore iniziale per λ, ad esempio λ_0 e si individua una sequenza di lambda $\lambda_1, \lambda_2, \ldots$ convergente al valore di $\hat{\lambda}$ che soddisfa $f(\hat{\lambda}) = 0$. La sequenza è definita ricorsivamente; si inizia con λ_0 e si definisce λ_{t+1} dato il valore di λ_t. Sia m_t la t-esima approssimazione per $\hat{m}$ ottenuta da λ_t attraverso $m = \exp(A\lambda_t)$. Il valore successivo di λ_t cioè λ_{t+1} è generato utilizzando la formula $\lambda_{t+1} = \lambda_t + [A'\text{Diag}(m_t)A]^{-1}A'(n - m_t)$,dove Diag (m_t) ha gli elementi $\{m_i^t\}$ nella diagonale principale. Questo a sua volta produce m_{t+1} e così di seguito da cui λ_{t+1} è ottenuto da λ_t attraverso la serie di regressioni pesate. Per t che cresce, m_t e λ_t convergono in modo piuttosto rapido alle stime di ML $\hat{m}$ e $\hat{\lambda}$. La matrice H= A′Diag (m_t)A converge a $\hat{H} = -A'\text{Diag}(\hat{m})A$ da cui la matrice stimata di covariaza di λ è data da $-\hat{H}^{-1}$.

3.5 Valutazione del modello

I modelli che si utilizzano per semplificare e formalizzare le relazioni tra le variabili del campione considerato possono essere valutati facendo un confronto in negativo, mediante la prova che i dati non contraddicono il modello ipotizzato. Una caratteristica dei modelli per dati categorici è che la bontà del modello stimato può essere verificata attraverso il *chi-quadrato del rapporto di verosimiglianza* o devianza che è una formulazione dello scarto tra frequenze teoriche e frequenze osservate. Per un campione multinomiale in

una tabella di contingenza doppia, la statistica nell'ipotesi che il modello formulato sia vero (ipotesi nulla H_0) è definita come:

$$G^2 = 2\sum_{ij} n_{ij} \log(n_{ij} / \hat{m}_{ij}),$$

dove $\hat{m}_{ijk} = n_{i.} n_{.j} / n_{..}$ sono le frequenze stimate relative al modello di indipendenza. Tal espressione è valida anche per altri schemi di campionamento come quello di Poisson o il multinomiale stratificato.

Per grandi campioni la statistica G^2 sotto l'ipotesi che il modello d'interesse è quello vero, ha distribuzione asintotica (al crescere della frequenza totale della tabella) di tipo chi-quadrato. Il numero dei gradi di libertà è pari al numero dei vincoli effettivi necessari per ridurre il modello saturo al modello d'interesse. Ad esempio nel caso di un campione multinomiale lo spazio dei parametri è dato da $\{\pi_{ij}\}$ con la restrizione $\sum_i \sum_j \pi_{ij} = 1$, la dimensione è data da (IJ-1). Sotto l'ipotesi nulla di indipendenza le probabilità $\{\pi_{ij}\}$ sono determinate dalle probabilità marginali $\{\pi_{i.}\}$ e $\{\pi_{.j}\}$, così le dimensioni sono (I-1) + (J-1). La differenza tra le due dimensioni (IJ-1)-[(I-1) - (J-1)] è uguale a (I-1)(J-1), da cui se il campione è sufficientemente grande G^2 ha una distribuzione chi-quadrato con (I-1)(J-1) gradi di libertà.

Nel caso di una tabella tripla la devianza diventa:

$$G^2 = 2\sum_i \sum_j \sum_k n_{ijk} \log\left(\frac{n_{ijk}}{\hat{m}_{ijk}}\right).$$

I gradi di libertà per la statistica si ottengono dalla differenza delle dimensioni tra l'ipotesi alternativa e l'ipotesi nulla. Ad esempio per un campione mutinomiale con probabilità $\{\pi_{ijk}\}$, partendo da un modello saturo, il solo vincolo che si ha è rappresentato da $\sum_i \sum_j \sum_k \pi_{ijk} = 1$ per cui i parametri

linearmente indipendenti sono (*IJK*-1). Se il modello che si vuole testare è il modello di mutua indipendenza le probabilità $\{\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}\}$ sono determinate da (*I*-1) parametri per $\{\pi_{i..}\}$ poiché $\sum_i \pi_{i..} = 1$, (*J*-1) parametri per $\{\pi_{.j.}\}$ e da (*K*-1) parametri per $\{\pi_{..k}\}$, da cui i parametri totali sono (*I*+*J*+*K*-3). Così i gradi di libertà totali sono: (*IJK*-1) – (*I*+*J*+*K*-3) = *IJK* – *I* – *J* – *K* +2. Lo stesso numero di gradi di libertà si ha nel caso in cui la distribuzione campionaria sia una distribuzione di Poisson.

Una proprietà che deriva dalla distribuzione del chi-quadrato è che la statistica in esame può essere suddivisa in componenti additive quando un modello a sua volta è suddiviso in componenti additive. Infatti se X_1 e X_2 sono variabili casuali indipendenti aventi una distribuzione del chi-quadrato con v_1 e v_2 gradi di libertà allora la variabile $X = X_1 + X_2$ ha ugualmente una distribuzione χ^2 con $df = v_1 + v_2$.

Sia $G^2(M)$ il valore di G^2 per il modello *M* basato sul confronto dei valori stimati $\hat{m}$ ai dati *n* ovvero: $G^2(M) = -2[L(\hat{m};n) - L(n;n)]$ dove $L(\hat{m},n)$ è il logaritmo della verosimiglianza sotto l'assunzione del modello e $L(n,n)$ è il logaritmo della verosimiglianza del modello saturo. Considerando due modelli M_1 e M_2 tali che M_2 è più semplice di M_1 per il fatto che ha alcuni parametri in meno, il modello M_2 è definito annidato o *nested* in M_1 ovvero M_2 è completamente contenuto in M_1 così che quando M_1 è un buon modello per i dati lo è anche M_2. Se si definiscono v_1 e v_2 i gradi di libertà relativi ai due modelli risulta che $v_1 < v_2$, poiché M_2 è più semplice di M_1. Così deve valere che $L(\hat{m}_2;n)$ sia minore di $L(\hat{m}_1;n)$, da cui:

$$G^2(M_1) \leq G^2(M_2).$$

Assumendo che il modello M_1 sia vero per testare la validità di M_2 si può usare la statistica:

$$G^2(M_2/M_1) = G^2(M_2) - G^2(M_1)$$

che ha una distribuzione chi-quadrato con gradi di libertà $df = v_2 - v_1$. Ad esempio in una tabella di contingenza doppia si può scrivere:

$$G^2(M_2/M_1) = 2\sum \hat{m}_{1i} \log\left(\frac{\hat{m}_{1i}}{\hat{m}_{2i}}\right),$$

dove $\{\hat{m}_{1i}\}$e $\{\hat{m}_{2i}\}$ sono i valori stimati per il modello M_1 e M_2.

Quando i vari modelli che si vogliono testare formano un set di modelli annidati ovvero differiscono tra loro solo per l'inclusione o meno di alcuni termini di associazione si può partizionare G^2 dal modello più semplice fino all'ultimo così da poter testare una sequenza di ipotesi. Goodman ha dimostrato che se si hanno s-1 test per confrontare s modelli annidati e tali test sono asintoticamente indipendenti, l'errore asintotico di primo tipo non può eccedere α. Per ulteriori approfondimenti si rimanda ad Agresti (1990).
Un'altra statistica utilizzata per valutare il modello è la statistica di Pearson:

$$X^2 = \sum\sum \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

che sotto l'ipotesi di validità del modello si distribuisce sempre come un chi-quadrato con (I-1)(J-1) gradi di libertà. Diversamente dalla statistica G^2 la differenza tra $X^2(M_2) - X^2(M_1)$ per i modelli annidati non ha una forma di Pearson, tale differenza non è necessariamente positiva. La statistica di Pearson che può essere usata per confrontare due modelli è:

$$X^2(M_2/M_1) = \sum \frac{(\hat{m}_{1i} - \hat{m}_{2i})^2}{\hat{m}_{2i}}.$$

Come G^2 tale statistica dipende dai dati soltanto attraverso i valori stimati ovvero le statistiche sufficienti per il modello M_1.

CAPITOLO 4

Rappresentazione dei modelli mediante grafi

4.1 Cenni sulla teoria dei grafi

Un grafo, $G = (V, E)$ è un oggetto matematico composto da un insieme finito di *vertici* V (detti anche nodi) e da un insieme finito di *archi* E (o lati) che collegano i vertici; formalmente E è composto da un insieme ordinato di coppie di vertici (α,β), $\alpha,\beta \in V$. Gli archi che collegano i vertici possono essere di due tipi: *orientati* se $(\alpha,\beta) \in E$ ed $(\beta,\alpha) \notin E$. Si tratta di archi che indicano una direzione causale che viene rappresentata da una freccia $\alpha \rightarrow \beta$, α è detto *genitore* del vertice β mentre β è definito *figlio* del vertice α; o archi *non orientati* si hanno se $(\alpha,\beta) \in E$ implica $(\beta,\alpha) \in E$, ovvero sono archi che indicano un' associazione che non ha direzione causale e sono rappresentati attraverso semplici linee tra nodi. Il grafo può essere rappresentato attraverso un diagramma:

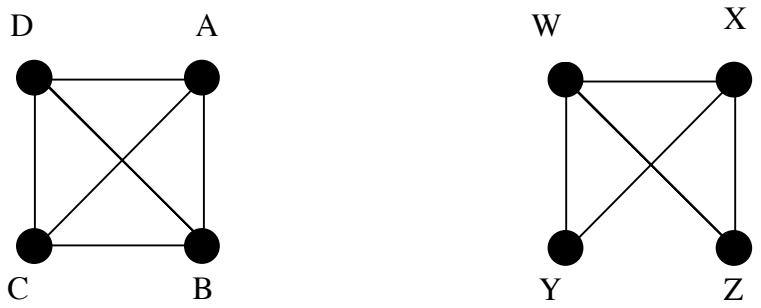

Figura 4.1 Esempi di grafi

I vertici sono disegnati come dei punti o cerchi che rappresentano le variabili categoriche pertanto indicati con le lettere maiuscole. Gli archi sono disegnati con delle linee rette che congiungono i vertici. Ovviamente è possibile disegnare un grafo in molti modi ma ciò non cambia la natura del grafo che è comunque costituito da vertici e archi.

Si dice che due vertici α, $\beta \in V$ sono *adiacenti*, si indicano con $\alpha \sim \beta$, se c'è un arco che li congiunge, ad esempio nel grafo precedente X e Y sono adiacenti mentre non lo sono Y e Z.

Ogni *sottoinsieme* di vertici $u \subseteq V$ costituisce un *sottografo* di G, si ha allora un grafo $G_u = (u,F)$ i cui F consiste in quegli archi di E dove entrambi i punti finali sono in u. Un grafo è *completo* se ogni vertice è unito agli altri vertici da un arco. Il grafo di destra disegnato è completo mentre quello di sinistra non lo è. Un sottoinsieme di vertici è completo se induce un sottografo completo.

Si definisce *clique* di un grafo quel sottoinsieme di vertici $u \subseteq V$ massimamente completo: l'aggiunta di un ulteriore vertice al sottoinsieme u comporta un sottografo non completo. Ad esempio le *clique* del grafo di destra sono $\{X,Y,W\}$ e $\{X, Z\}$.

Una sequenza di vertici distinti $X_0,\ldots,X_n$ tali che $X_{i-1} \sim X_i$ per $i= 1,\ldots,n$ si definisce *cammino* di lunghezza n in G. Graficamente un *cammino* è semplicemente una sequenza di archi che collega i due vertici. Nel grafo precedente Z,X,Y,W è un cammino di lunghezza tre tra Z e W, similmente Z,X,W è un cammino tra Z e W di lunghezza due. Se esiste un cammino tra ciascuna coppia di vertici il grafo si dice *connesso*. Un cammino $X_0, X_2,\ldots,X_n$, X_1 è detto *ciclo* di lunghezza n; ad esempio A,B,C,D,A è un ciclo di lunghezza quattro. Se gli n vertici $X_0,\ldots,X_n$ del ciclo $X_0, X_2,\ldots,X_n, X_1$ sono distinti e se vale $X_j \sim X_k$ solo se $|\, j-k \,|=1$ o $n-1$, allora si ha *una corda in un ciclo*. Un grafo è *aciclico* se non contiene alcun ciclo. Per tre sottoinsiemi a,b e s di V, si dice

che *s* separa *a* e *b* se ogni cammino da *a* a *b* passa per *s*.Per un'approfondita trattazione della tematica si rimanda a Whittaker (1990) ed Edwards (1995).

4.2 Grafi di indipendenza

Un modello di indipendenza condizionata può essere rappresentato visivamente attraverso un grafo. Il grafo, infatti, fornisce un'immagine della struttura di indipendenza tra le variabili oggetto di analisi. La base di tale schema di rappresentazione è la corrispondenza tra separazione grafica e le proprietà della relazione d'indipendenza condizionata dette *proprietà markoviane*.

Tra i grafi di indipendenza se ne possono distinguere vari tipi:

- ***grafi non orientati***, se tutti gli archi del grafo sono rappresentati da delle linee;
- ***grafi orientati***, se tutti gli archi sono costituiti da frecce;
- ***grafi a catena*** ovvero grafi costituiti sia da linee che da frecce.

- I ***grafi di indipendenza non orientati*** sono legati ad un insieme di variabili casuali nel seguente modo. Sia $V = (V_1,V_2, \ldots,V_n)$ un vettore di variabili casuali associate ai vertici del grafo, sia $f = f(V)$ la funzione di densità congiunta di V. Per un dato grafo $G(V,E)$ si dice che f soddisfa la *proprietà markoviana a coppie* (C) rispetto a G se ogni coppia di variabili non adiacenti in G è condizionatamente indipendente date le restanti, ad esempio:

$$v \not\sim w \Rightarrow v \perp w | V \setminus \{v,w\}.$$

Si dice che f soddisfa la *proprietà markoviana globale* (G) rispetto a G se per tre sottinsiemi disgiunti A, B, S di V tali che S separa A e B in G si ha che:

$$A \perp B \mid S.$$

Tale proprietà ha un'importante conseguenza: permette di individuare le relazioni di indipendenza tra le variabili condizionando rispetto alle sole variabili separanti (*minimo insieme separatore*).

Un risultato interessante è stato dimostrato da Pearl e Paz (1986)[29]. Se f soddisfa la condizione di positività $f > 0$ allora (C) $\Leftrightarrow$(G). Di conseguenza un grafo costruito in accordo con una di tali proprietà soddisfa anche l'altra.

La semplice separazione grafica permette allora di ricavare informazione sulla struttura d'associazione tra le variabili del grafo. Ad esempio per un modello con quattro variabili W, X, Y e Z per la quali si ha che $W \perp Z \mid (X,Y)$ e che $Y \perp W \mid (X,Z)$ si può formare il grafico illustrato in figura 4.2

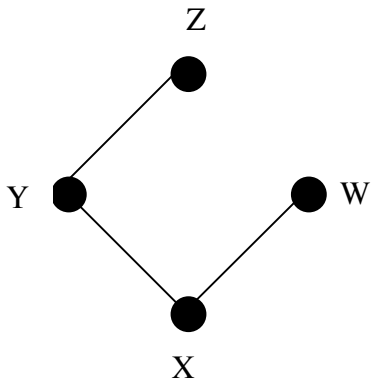

Figura 4.2 Esempio di grafo.

dal quale si deduce che $W \perp Z \mid Y$ e che $X \perp Z \mid Y$. La proprietà globale markoviana consente pertanto di legare la proprietà di separazione della teoria dei grafi alla proprietà statistica d'indipendenza condizionata. Pearl e Paz (1986) hanno anche dimostrato che l'equivalenza delle proprietà markoviane può essere

[29] Pearl J. e Paz A. (1986), 'Graphoids. A graph-based for reasoning about relevancy relations.' *Proceedings of the 7th European conference on Artificial Intelligence*, Bringhton.

ottenuta anche attraverso delle induzioni formali, utilizzando solo alcune semplici proprietà d'indipendenza condizionata come assiomi.

- Nei ***grafi orientati*** le frecce che congiungono i vertici stanno ad indicare una direzione causale. Tali grafi sono stati inizialmente studiati in relazione alla *path analysis* per variabili gaussiane (Wright, 1921) e categoriche (Goodman,1973). In sintesi, questi assumono che la distribuzione in esame è costituita da un ordine completo tra le variabili (V_1, ……, V_n) tale che V_i è anteriore a V_{i+1} per $i = 1, 2,\ldots n\text{-}1$ nell'ordinamento. La fattorizzazione della funzione congiunta d'interesse statistico diventa pertanto:

$$f(v_1)f(v_2|v_1)\ldots..f(v_n|v_{n-1}\, v_{n-2}\ldots v_1).$$

Nel grafico per $i < j$, viene disegnata una freccia da V_i a V_j a meno che $f(v_j|v_i\ldots.v_1)$ non dipenda da v_i, ovvero eccetto che si abbia

$$V_i \perp V_j|\ \{V_1\ldots V_j\}\setminus\{V_i,V_j\}.$$

Questa rappresenta la versione diretta della proprietà markoviana a coppie.

I grafi generati in questo modo sono grafi *aciclici orientati*; *orientati* perché tutte gli archi sono delle frecce e *aciclici* perché la sequenza causale di $\{V_i\}$ impedisce il formarsi di *cicli*.

Dato che la funzione di densità condizionata $f(v_j|\ v_{j-1}\ldots.v_1)$ può essere liberamente specificata, ogni modello univariato può essere applicato. Tuttavia tale tipologia di grafi non è stata molto utilizzata in statistica; una spiegazione possibile potrebbe essere che raramente un ordine causale può essere specificato in modo così completo.

- I ***grafi a catena*** combinano i grafi orientati e non orientati in un unico contesto. Questi sono basati su una partizione dell'insieme di vertici V in sottoinsiemi numerati tali da formare una cosiddetta *catena di dipendenze* $V=S_1\cup\ S_2\ \cup\ldots\cup S_k$ e la corrispondente fattorizzazione della funzione di

densità congiunta $f\ (V_1,\ldots..,V_n)$ di interesse statistico nel rispetto dell'ordinamento è la seguente:

$$f\,(S_1)\,f\,(S_2|\,S_1)\ldots f\,(S_k|\,S_{k\text{-}1}\cup S_{k\text{-}2}\cup\ldots..\cup S_1).$$

Vale a dire che è possibile scrivere la funzione di densità come prodotto di funzioni di densità condizionali rispetto all'ordinamento tra le variabili.

I sottoinsiemi S_i sono chiamati *componenti* della catena di dipendenza (o *struttura ricorsiva a blocchi*); in pratica si tratta degli insiemi i cui elementi sono connessi. Le variabili che si trovano nello stesso componente sono covariate, ovvero la loro struttura di associazione è assunta simmetrica, senza un ordine causale. I componenti della catena sono ordinati in modo che S_i è anteriore a S_{i+1} per $i = 1,\ldots..,k\text{-}1$. Così tutti gli archi tra vertici appartenenti allo stesso sottoinsieme sono non orientati, mentre tutti gli archi tra vertici appartenenti a sottoinsiemi differenti sono archi orientati con direzione determinata dall'ordinamento dei sottoinsiemi. La figura mostra un esempio di grafo a catena.

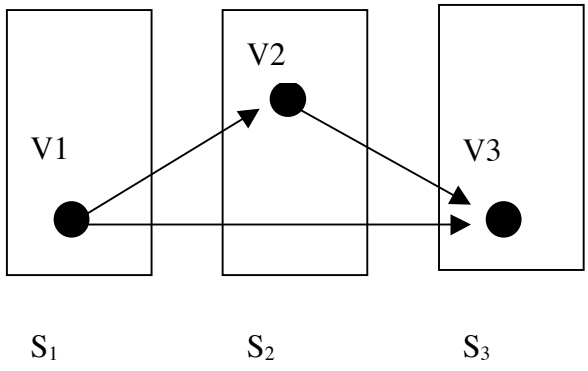

Figura 4.3. Esempio di grafo a catena con tre blocchi.

Se manca una linea tra due vertici dello stesso componente S_i, o una freccia da $v \in S_j$ a $w \in S_i$, per $j < i$ questo sta a significare che

$$v \perp w \mid S_1 \cup S_2 \cup \ldots \cup S_i \,. \qquad (4.1)$$

La (4.1) fornisce la versione della *proprietà markoviana a coppie per il grafo a catena*[30]. Per dimostrazioni e ulteriori risultati si rimanda a Frydenberg (1989)[30].

Tali grafi non contengono cicli orientati, ovvero se ci si muove dal vertice v rispettando la direzione degli archi, allora non si può tornare a v una volta superato un arco orientato. In tal senso un *ciclo* deve essere composto interamente da linee. Dato il grafo, i componenti della catena possono essere facilmente identificati guardando i vertici che rimangono adiacenti quando tutti gli archi del grafo sono stati rimossi. I componenti della catena sono interpretabili soprattutto in relazione alle strategie per l'analisi dei dati: per le variabili in S_1 verrà scelto un modello senza direzione causale (a meno che S_1 non sia considerato esogeno), poi verrà modellata la distribuzione la distribuzione condizionata di S_2 dato S_1, poi ancora la distribuzione condizionata di S_3 dato S_1 e S_2 e così di seguito fino all'ultimo componente.

I modelli statistici adatti in questo caso sono modelli di risposta multivariati per S_i dati $S_1 \cup S_2 \cup \ldots \cup S_{i-1}$, per i quali un insieme arbitrario di relazioni di indipendenza condizionata del tipo (4.1) può essere specificato.

---

[30] Frydenberg M. (1989), 'The chain graph Markov property', *Research Report, 186*. Department of Theoretical Statistic, University of Aarhus.

4.3 Modelli grafici log-lineari

I grafi di indipendenza consentono di integrare la metodologia dei modelli log-lineari in quanto permettono una visualizzazione grafica del modello. Questo è particolarmente utile per modelli di dimensioni elevate per i quali il grafo consente di chiarire in modo considerevole la struttura e l'interpretazione del modello stesso.

Si definiscono *modelli grafici* quei modelli tali che ad ogni *clique* del grafo corrisponde il medesimo ordine d'interazione tra le variabili. Ovvero nei modelli grafici c'è una corrispondenza biunivoca tra grafo di indipendenza e modello.

Considerando, ad esempio, il modello log-lineare di mutua indipendenza **(X,Y,Z),** per la tabella tridimensionale corrispondente alle variabili discrete X, Y e Z, il grafo associato con il modello è il seguente:

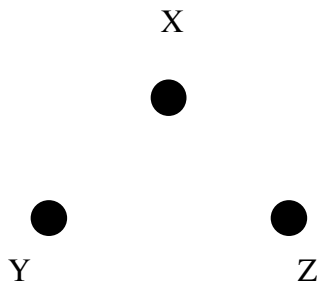

Figura 4.4 Grafo relativo al modello log-lineare (X,Y,Z).

Invece per il modello log-lineare indicato come **(XY,XZ)**

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$$

dato che tale espressione implica

$$m_{ijk} = \exp(\mu + \lambda_i^X + \lambda_k^Z + \lambda_{ik}^{XZ})\exp(\lambda_i^X + \lambda_{ij}^{XY})$$

si può vedere che le probabilità di cella si fattorizzano in termini che non coinvolgono simultaneamente Y e Z; che come evidenziato al terzo capitolo è una delle caratteristiche dell'indipendenza condizionata tra due variabili. Il *criterio di fattorizzazione per l'indipendenza condizionata* afferma, infatti, che le variabili casuali X e Y sono condizionatamente indipendenti dato Z se e solo se è possibile fattorizzare la funzione di densità congiunta in funzioni che non dipendono contemporaneamente da X e Y. Tale criterio è molto utilizzato nella verifica della relazione di indipendenza condizionata[31]. In tal caso il grafo di indipendenza è il seguente:

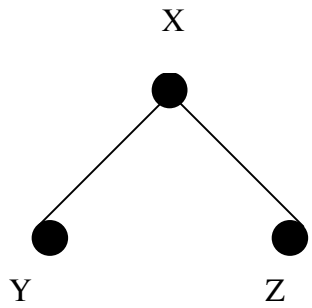

Figura 4.5 Grafo relativo al modello log-lineare (XY,XZ).

Dalla definizione data in precedenza di modello grafico si deduce che il grafo in figura 4.6 rappresenta il modello **(XYZ,YZW)** dato che è interpretabile interamente in termini di relazioni di indipendenza condizionata in quanto X e W sono condizionatamente indipendenti date le restanti variabili

[31] Dawid A. (1979), 'Conditional indipendence in statistical theory',Jouornal of the Royal Statistical Society,B, 39, pp.1-38.

e il modello non contiene il termine d'interazione $\lambda_{il}^{XW} = 0$ così come tutti gli altri termini di ordine maggiore che lo comprendono.

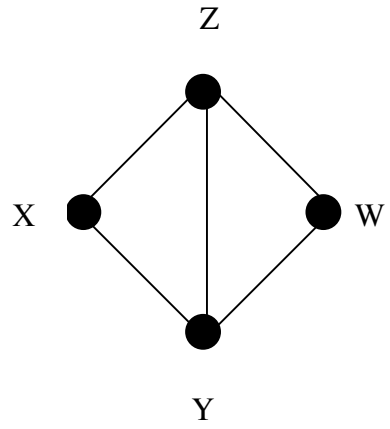

Figura 4.6 Grafo relativo al modello (XYZ,YZW)

In pratica un modello log-lineare gerarchico è grafico se contiene tutti i coefficienti a due fattori generati da un coefficiente di ordine più elevato nonché il coefficiente di interazione di ordine più elevato.[32] Per una definizione in termini probabilistici si rimanda a Whittaker (1990)[33]. Si dimostra che nella stima dei modelli grafici log-lineari, le equazioni di verosimiglianza sono caratterizzate dall'uguaglianza delle funzioni di densità marginale osservate e stimate per tutti i margini corrispondenti le *clique*.

Una *sottofamiglia* di modelli grafici sono i *modelli scomponibili* ovvero quei modelli log-lineari che presentano espressioni dirette per le stime dei parametri, per cui non è necessario nessun procedimento di adattamento ricorsivo proporzionale. Questi sono stati inzialmente studiati da Goodman (1970) e caratterizzati come modelli grafici i cui grafi sono *triangolari* (*o*

[32] Christensen R. (1990), 'Log-linear Models', Sprinter-Verlag, New York, p.103.
[33] Whittaker J. (1990), op. cit., p.207.

cordali), ossia si tratta di grafi che non contengono nessun *ciclo* di lunghezza maggiore di tre.

Nel seguito si introduce il concetto di collassabilità date le forti implicazioni con i modelli grafici a catena che sono formulati interamente in termini di distribuzioni condizionali e marginali.

4.4 Collassabilità

Un problema che si riscontra nella valutazione del modello è dovuto alla presenza dei cosiddetti *random zeros* ovvero delle celle che hanno una probabilità positiva di verificarsi ma non si verificano nel campione considerato. Diverso è invece il caso dei *fixed zeros* o zeri strutturali che si presentano quando la combinazione delle categorie delle variabili è impossibile. Le frequenze interne a tali celle devono essere sempre pari a zero e le tabelle di contingenza sono dette *tabelle incomplete*. Queste ultime tuttavia non costituiscono un problema dato che la probabilità che un osservazione sia in una determinata cella è pari a zero, per cui anche il valore atteso m_i deve esserlo. Nella stima del modello log-lineare è sufficiente non considerare gli zeri strutturali.[34]

Nel caso, invece, dei *random zeros* il problema è noto come *sparsità della tabella di contingenza*, questo fatto comporta innanzitutto che i risultati asintotici non sono più validi. Una delle assunzioni fatte per il modello è, infatti, che le probabilità di cella siano tutte positive; se così non è la distribuzione asintotica del rapporto di verosimiglianza non è più un chi-quadrato e questo porta a risultati distorti nella selezione del modello. Alcuni

[34] Christensen R. (1990), op. cit. p.333.

esempi di come la sparsità della tabella può provocare l'aumento della devianza quando si aggiungono ulteriori variabili al modello sono presentati in Sweart e Whittaker[35]. In secondo luogo non è certo che le stime di massima verosimiglianza esistano. Come visto in precedenza, ad esempio per un modello log-lineare saturo con tre variabili A, B e C, i valori osservati devono uguagliare quelli stimati: $m_{ijk} = \hat{m}_{ijk}$. Dato che il modello è log-lineare deve essere definito $\log m_{ijk}$ ma se per alcuni i, j o k $m_{ijk} = 0$, le stime di massima verosimiglianza non esistono.

Per far fronte a questa situazione un approccio suggerito è quello di identificare le celle per le quali i vincoli della verosimiglianza implicherebbero che $\hat{m} = 0$ e trattarle come zeri strutturali, cioè eliminarle dal modello. In tale modo i gradi di libertà del modello sarebbero dati da quelli del modello meno il numero degli zeri strutturali, che costituiscono dei gradi di libertà persi, dato che non forniscono 'informazione'.[36] Tuttavia tale metodo è abbastanza laborioso dal punto di vista computazionale.

Ancora un altro metodo utilizzato quando i risultati asintotici sembrano non essere più validi, è quello di valutare la distribuzione campionaria esatta delle osservazioni delle celle. In tale modo si ottengono *test esatti condizionati* ovvero basati sulle probabilità della distribuzione condizionata.

Tuttavia molte strategie sono state proposte per analizzare tabelle di contingenza con un elevato numero di celle. Un approccio particolarmente interessante è legato al concetto di *collassabilità*. Ci sono due principali modi di intenderla, uno si riferisce alle condizioni che garantiscono l'assenza di fenomeni come il *Paradosso di Simpson*, ovvero, quando è possibile studiare le relazioni tra alcune variabili esaminando semplicemente la tabella marginale e i parametri d'interazione sono gli stessi nella tabella completa e in quella

[35] Sweart P. e Whittaker J. (1998), 'Graphical models in credit scoring', IMA, Journal of Mathematics Applied in Business & Industry, 9, pp.224-266.

collassata. Questa è definita anche *collassabilità parametrica*.[37] Per ulteriori approfondimenti si rimanda a Bishop (1971), Whittemore (1978), Wermuth(1987).

L'altro concetto di collassabiltà introdotto da Asmussen e Edwards (1983) può essere definito *model collapsibility* in quanto è legato all'idea di invarianza dei modelli quando alcune variabili non vengono osservate. Sia p una tabella di contingenza multidimensionale per un insieme di variabili $K=\{1,2,\ldots,k\}$. Sia $p(x)$ la probabilità che una generica osservazione di un campione di dimensione fissa n cada nella cella x, con $\sum p(x) = 1$. Si indichi con p_a la tabella marginale ovvero quella ottenuta sommando $p(x)$ rispetto a tutti i livelli delle variabili nell'insieme K/a. Si indichi inoltre con L il modello log-lineare tale che ($p \in$ L) con L_a il modello marginale su p_a. Il modello L è definito collassabile in $a \subseteq k$ se vale una delle seguenti proprietà:

- per ogni $p(x) \in$ L si ha che $p_a(x_a) \in L_a$;
- per tutti $\hat{p}(x_a) = \hat{p}_a(x_a)$,

dove $p(x_a)$ sono le stime di massima verosimiglianza delle probabilità delle celle in a sotto il modello L, mentre $p_a(x_a)$ sono le probabilità stimate di ML delle celle in a per L_a. Questa condizione stabilisce che le interazioni marginali tra le variabili in a possono essere studiate sia nella tabella completa che nella tabella collassata, così come deve essere uguale il test del rapporto di verosimiglianza cioè le relazioni di indipendenza tra le variabili in a restano inalterate dopo la marginalizzazione.

In pratica L è collassabile in a se e solo se a è contenuto in un generatore. In altre parole questa condizione stabilisce che la stima delle relazioni per a può essere ottenuta nella tabella marginale di a. Tale concetto è molto utile per

[36] Christensen R. (1990), op.cit., pp.341.
[37] Lauritzen S. (1996), op. cit., p.121.

ridurre la dimensione della tabella e utilizzare le tabelle marginali che hanno meno probabilità di essere sparse.

Attraverso il modello grafico è possibile individuare visivamente se il modello è collassabile rispetto ad un certo sottoinsieme di variabili b, con $b \subseteq K$. Un grafico G è collassabile se e solo se la frontiera di ciascun componente collegato in b è completa in G. Per ulteriori profondimenti ed esempi si rimanda a Whittaker (1990).

CAPITOLO 5

Caratteristiche del comportamento con la carta di credito

5.1 Introduzione

Nel presente capitolo si procederà ad un'applicazione dei modelli grafici log-lineari e logistici nell'ambito dello *scoring* comportamentale.

Lo studio viene condotto per ottenere informazioni sul comportamento del cliente una volta che gli è stata concessa la carta di credito. Lo scopo principale della società finanziaria è quello di 'capire' il comportamento dei clienti per avere un'idea più chiara circa la parte della clientela alla quale offrire condizioni più vantaggiose (*marketing segmentation*), ad esempio tassi d'interesse più bassi, servizi di consulenza a prezzi ridotti, nell'ottica della fidelizzazione della clientela stessa, ed avere, inoltre, delle previsioni su quale parte dei clienti ha maggiore probabilità di avere un saldo del conto 'inattivo' per un lungo periodo. Allo stesso tempo è anche interessante capire quali clienti ristabiliranno una relazione con la società una volta che essa si è conclusa. Lo studio delle relazioni tra le variabili che determinano il comportamento del cliente è utile per investigare tale contesto.

Dapprima si procede a descrivere il campione analizzato ed alcuni problemi riguardanti i dati. Dopo aver illustrato la categorizzazione delle variabili continue si riportano i risultati della stima del modello grafico a

catena e di quello logistico cumulativo. Lo scopo dell'approccio seguito è di spiegare le relazioni tra l'insieme delle variabili, cercando un modello che catturi le informazioni nel campione considerato.

Alcune analisi sono poi state condotte utilizzando il sistema esperto di Hugin (Olesen *et al*. 1992) per il calcolo delle probabilità marginali e condizionali data la stretta connessine con i modelli presentati, queste sono riportate nell'ultimo paragrafo.

5.2 Il campione

L'analisi per la costruzione di un modello grafico a catena è condotta su un campione composto di circa 60000 osservazioni estratte con campionamento casuale semplice da una popolazione di titolari di una carta di credito *revolving*, rilasciata da una delle maggiori società finanziarie italiane operanti nel settore. Esso riguarda tutti coloro che hanno iniziato un business con la società nei primi sei mesi del 1997.

Le variabili prese in esame sono costituite da:

- un insieme di variabili dette *application variables,* che sono quelle desumibili dal modulo che il cliente compila al momento della richiesta della carta di credito;
- lo stato del conto del cliente in due periodi successivi: dicembre 1997 e 1998. Il titolare della carta ogni mese viene classificato dalla società come '*inattivo*' se il saldo del conto non è attivo da più di sei mesi, '*dormiente*', se il saldo del conto non è attivo da meno di sei mesi ed '*attivo*' altrimenti. L'osservazione consecutiva di tale variabile permette di sapere se un determinato soggetto ha movimentato il conto subito dopo l'apertura, il periodo in cui non ha utilizzato la carta e le sue successive utilizzazioni.

Alcune delucidazioni sono utili sul campione utilizzato e sui problemi statistici ad esso associati. Il campione è interessato da processi di selezione che potrebbero essere fonte di distorsione. Una prima selezione da cui è affetto influenza il *credit scoring* ma non il *behavioural scoring*, è dovuta al fatto che i dati provengono dal normale processo di accettazione dei clienti seguito dalla banca, invece il metodo di *score* che si costruisce con il *credit scoring* che consente di discriminare tra 'buoni' e 'cattivi' clienti andrà applicato a tutti i richiedenti crediti. I risultati resterebbero validi fintanto che il processo di selezione della società non muta, tuttavia si può far inferenza sulla popolazione non accettata utilizzando il processo conosciuto come *reject inference* trattato al primo capitolo.

Nello *scoring* comportamentale, invece, l'interesse risiede proprio in coloro la cui richiesta è stata accettata, in pratica in coloro che hanno uno *score* al sopra di una determinata soglia. La forma di distorsione che è presente in questo caso sta nel fatto che le persone possono rifiutarsi di accettare la carta che gli viene concessa. Per cui la popolazione che possiede una carta si è auto-selezionata in base alla decisione di accettarla o rifiutarla. Il campione preso in esame è costituito solo da coloro ai quali la carta è stata effettivamente attivata. Per cui i risultati dei modelli considerati risulterebbero validi solo per i possessori di carta di credito e non anche per coloro ai quali viene proposto l'utilizzo di una carta da parte della banca. Per poter estendere i risultati del modello anche a coloro che hanno rifiutato la carta di credito, occorre accettare l'assunzione che, per un dato insieme di variabili esplicative considerate, se il cliente accetta o meno la carta questo non comporta informazioni sul futuro comportamento con la carta stessa. Se vale questa relazione di indipendenza tra la decisione di accettare la carta e il comportamento con la carta, condizionatamente alle variabili esplicative, si

dimostra che i parametri di interazione del modello stimato sono gli stessi del modello costruito per l'intera popolazione[38].

Occorre inoltre tener presente che il campione potrebbe essere affetto da fattori stagionali. Ad esempio, per l'acquisto di determinati beni la società offre la possibilità di attivare il conto senza pagare interessi. I clienti che aprono le relazioni con la società in questa forma tendono a non richiedere ulteriori prestiti, ovvero a disattivarsi una volta esaurito il plafond iniziale. Essendo l'acquisto di questi beni soggetto a variazioni stagionali, (in particolare aumentano nella stagione primaverile) è possibile che tali clienti siano sovrarappresentati nel campione in studio. Per maggiore precisione sarebbe necessario conoscere quale parte della popolazione è interessata da tali politiche, tuttavia questa informazione non è stata resa disponibile.

5.3 Descrizione delle variabili

Per ogni cliente che possiede la carta vengono registrate un insieme di variabili descriventi il comportamento demografico e finanziario. In questo lavoro si è tenuto in considerazione soltanto un sottoinsieme di tali variabili:

- l'età del cliente;
- il reddito mensile del cliente e del coniuge;
- il numero dei figli;
- lo status residenziale.

Come già detto Si sono considerate le due variabili risposta di maggior interesse rappresentate dallo stato del conto in dicembre 1997 e in dicembre 1998, in modo tale da studiare il comportamento del cliente nel lungo periodo, ed in particolare il comportamento di spesa durante il periodo natalizio.

[38] Stanghellini E.*et al* (1999), op. cit. pp 239-251.
Una dimostrazione teorica di come il grafo sia collassabile sulle variabili esplicative e le variabili risposta può essere trovata in Edwards D. (1995), op.cit. pp.90-102.

Sia il reddito del cliente che quello del coniuge presentano dei valori mancanti, che potrebbero essere non dichiarati perché il coniuge è inesistente o magari perché non si è a conoscenza della cifra o per altre cause. In alcune circostanze il dato mancante potrebbe fornire comunque delle indicazioni ma in questo caso non se ne è tenuto conto data l'esiguità delle mancate risposte rispetto alla numerosità del campione.

Le variabili quali l'età e il reddito sono variabili continue mentre le altre sono categoriche. Poiché in tale contesto vengono utilizzati modelli per dati categorici, le variabili continue sono state categorizzate, mentre per le altre si è proceduto a ridurre il numero delle categorie per avere un modello più parsimonioso e maggiormente interpretabile.

Una prima analisi delle tabelle dello stato del conto del cliente nei due periodi considerati è utile al fine di avere una visione più chiara dello studio in esame.

Nel dicembre 1997, circa un anno dal rilascio della carta, il 75% dei clienti risultano avere un conto attivo mentre soltanto il 6% hanno un conto inattivo, i dormienti sono il 18%. La percentuale elevata degli attivi potrebbe anche essere sovrastimata per via dei clienti interessati dalle politiche stagionali della società alle quali si fatto riferimento in precedenza. La seguente tabella riporta le frequenze assolute.

Stato al 12/97	*Frequenze*	*Percentuali %*
Inattivo	3924	6.2
Dormiente	11635	18.4
Attivo	47782	75.4
Totale	63341	100

Tabella 5.1 Stato del conto del cliente nel Dicembre 1997.

Se si considera, infatti, lo stato del conto al 12/98 si nota che la percentuale di attivi si aggira intorno al 54% mentre è cresciuta considerevolmente quella degli inattivi passando al 37%, i dormienti sono soltanto il 6%. Si nota perciò che dopo due anni dalla attivazione della carta l'utilizzo della stessa o avviene con frequenza o non avviene più. La tabella è illustrata in tabella 5.2.

Stato al 12/98	*Frequenze*	*Percentuali %*
Inattivo	23530	37.1
Dormiente	5443	8.6
Attivo	34368	54.3
Totale	63341	100

Tabella 5.2 Stato del conto del cliente nel Dicembre 1998.

Considerando la tabella incrociata si nota che circa il 71% di coloro che sono inattivi al 12/97 lo sono anche un anno dopo, mentre dei dormienti al primo anno il 73% diventano inattivi al 12/98. Tra gli attivi al 12/97 il 48% lo sono anche al 12/98, mentre il 19% sono inattivi.

5.4 Categorizzazione delle variabili continue

La categorizzazione delle variabili continue e la riduzione del numero di categorie per quelle categoriche è stata condotta attraverso l'utilizzo di un modello log-lineare adottando la parametrizzazione per variabili ordinali di Wermuth e Cox introdotta al capitolo 3. Tale parametrizzazione consente di costruire degli *odds ratio* locali relativi alle categorie adiacenti delle variabili.

Per le variabili continue è stata scelta inizialmente una categorizzazione molto raffinata e ciascuna variabile è stata incrociata in una tabella di contingenza con la variabile risposta lo 'stato del conto del cliente al 12/98'. Poi sulla base del modello log-lineare stimato con la parametrizzazione adottata si sono considerati i logaritmi degli *odds ratio* di livelli adiacenti divisi per la loro deviazione standard. Per grandi campioni e sotto l'ipotesi che il logaritmo dell'*odds ratio* sia zero, possono essere trattati come valori di una distribuzione normale con media zero, da cui i valori in modulo maggiori di 1.96 sono molto improbabili, ovvero evidenziano la dipendenza nella tabella 2 × 2 corrispondente. Il modello ridotto stimato pone i parametri non significativi pari a zero. Se il modello risultante è un buon modello, si procede ad aggregare le classi corrispondenti. Tale metodo consente, infatti, di tener conto della scala ordinale che presentano le variabili continue categorizzate.

Considerando la variabile 'numero di figli' (**A**), dapprima le classi considerate sono state quattro: uno, due, tre o più di tre figli e si è formata la tabella 4 × 3 ottenuta dalla classificazione incrociata di tale variabile con lo 'stato al 12/98' (**B**). Le entrate delle celle sono riportate nella seguente tabella.

N. Figli **(A)**	***Stato al 12/98*** **(B)**			
	Inattivo	Dormiente	Attivo	Totali
0	11165	2515	15579	29199
1	5107	1255	7832	14194
2	5564	1319	8077	14960
≥ 3	1694	354	2940	4988
Totali	**23530**	**5443**	**34368**	**63341**

Tabella 5.3 Valori osservati per la classificazione incrociata del numero di figli contro lo stato del cliente al 12/98

Il logaritmo delle frequenze della tabella viene disposto in un unico vettore in cui i valori delle celle vengono posti in modo ordinato procedendo per colonna, ovvero con i valori della variabile (**A**) che si alternano più velocemente. Utilizzando il software statistico *S-plus* questo è stato fatto costruendo la funzione che consente di vettorizzare i valori della tabella per colonna, riportata in appendice in A.1.

Si è poi costruita la matrice dei contrasti fra coppie di termini con la parametrizzazione per variabili ordinali di Wermuth e Cox. La matrice deriva dal prodotto di Kronecker delle matrici definite nel seguente modo:

$$C_3 = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \quad \text{e} \quad C_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Tali matrici sono state ottenute in *S-plus* costruendo una funzione che permette di averle di qualsiasi dimensione, la funzione è riportata in appendice in A.3.

Stimando il modello log-lineare con tale parametrizzazione si ottengono i rapporti degli *odds* che, ad esempio, per l'ultima colonna della matrice del disegno sono quelli relativi alla sotto tabella 2 × 2 per i livelli 3 e 4 di (**A**) e i livelli 2 e 3 di (**B**).

Il procedimento che consente di stimare il modello in *S-Plus* è riportato in appendice in A.4.

La tabella seguente riporta i parametri d'interazione, del modello log-lineare stimato saturo e ridotto, divisi per la loro deviazione standard, che corrispondono al logaritmo degli *odds ratio* di livelli adiacenti.

Termini d'interazione studentizzati				***Termini d'interazione studentizzati***			
Riga	*Colonna*	*Modelllo saturo*	*Modello ridotto*	*Riga*	*Colonna*	*Modelllo saturo*	*Modello ridotto*
12	*12*	2.265	2.30	*12*	*23*	0.303	0.05
23	*12*	-0.819	0	*23*	*23*	-0.44	0
34	*12*	-1.911	-2.30	*34*	*23*	4.79	4.91

Tabella 5.4 Valori stimati per il modello saturo e per quello ridotto.

Le stime di massima verosimiglianza del modello sono state ottenute attraverso l'algoritmo di stima proporzionale iterativa (*Newton Raphson algorithm*). Il modello ridotto ottenuto uguagliando a zero i parametri d'interazione non significativi, ovvero eliminando le colonne corrispondenti della matrice del disegno, comporta la relazione d'indipendenza nella sottotabella relativa ai livelli 2,3 della variabile (**A)** e i livelli 1,2,3 della variabile (**B**). La devianza è pari a $G^2 = 2.81$ con $df = 2$, *p-value* = (0.2454), dato che tale valore non è significativo neppure al 5%, è ragionevole accettare il modello. Questo permette di unire i livelli 2,3 della variabile (**A)**, ovvero l'associazione marginale tra **(A)** e **(B)** può essere indifferentemente studiata in una tabella di contingenza collassata 3×3.

Da cui le categorie prese in esame per tale variabile sono le seguenti:

Categorie della variabile n.figli
0
1-2
> 3

Il reddito considerato è espresso in euro ed è stato ottenuto sommando il reddito mensile dichiarato, del cliente e del coniuge se esistente. Questo è stato dapprima suddiviso in sette categorie. La tabella di contingenza per tale variabile incrociata con la variabile risposta è riportata nella tabella 5.5. I valori mancanti riscontrati per la prima categoria della variabile risposta sono 27, per la seconda categoria 10 e per la terza 119. Come detto in precedenza tali valori non sono stati tenuti in considerazione.

Reddito	***Stato al 12/98***			
	Inattivo	Dormiente	Attivo	Totali
1-600	199	73	596	868
601-1230	504	144	1030	1678
1231-1450	2089	525	3655	6269
1451-1870	3908	887	5906	10701
1871-2120	3246	806	4743	8795
2121-3210	6900	1549	9504	17953
≥ 3120	6657	1449	8815	16921
Totali	**23503**	**5433**	**34249**	**63185**

Tabella 5.5 Valori osservati per la classificazione incrociata del reddito contro lo stato del cliente al 12/98.

Come visto in precedenza il logaritmo delle frequenze delle celle è stato vettorizzato procedendo per colonna. Utilizzando il metodo sopra descritto con le matrici del contrasto C_3, C_7 si è stimato il modello log-lineare.

Si riportano di seguito i valori stimati dei parametri del modello, la varianza e i rapporti studentizzati, relativi soltanto ai coefficienti d'interazione.

Righe	*Colonne*	*Stime dei parametri d'interazione*	*Errore standard*	*t-value*
12	*12*	-0.2499	0.5600	-4.4035
23	*12*	-0.1283	0.1121	-1.4479
34	*12*	-0.1019	0.0996	-1.0231
45	*12*	0.0911	0.0465	2.2775
56	*12*	-0.1028	0.0671	-1.5309
67	*12*	0.0388	0.1202	0.3171
12	*23*	0.1013	0.0696	1.4540
23	*23*	-0.0271	0.1277	-0.2118
34	*23*	-0.0445	0.0407	-1.0943
45	*23*	-0.1236	0.2584	-0.4780
56	*23*	0.0418	0.0659	0.6339
67	*23*	0.0895	0.4610	-2.2375

Tabella 5.6 Valori stimati dei parametri d'interazione del modello, errore standard e *t*-value.

Il modello ridotto stimato è stato ottenuto ponendo uguali a zero *odds ratio* corrispondenti alle righe 2,3 3,4 e 5,6,7. Il modello include, pertanto, le ipotesi di indipendenza nella sottotabella 2×3 tra i livelli 2,3 e 3,4 del reddito e i livelli 1,2,3 dello 'stato del conto al 12/98' ($A_i \perp B_j$ per i=3,4 e j=1,2,3) ed anche nella sottotabella 3×3 tra i livelli 5,6,7 del reddito e i livelli 1,2,3 dello 'stato del conto al 12/98' ($A_i \perp B_j$ per i= 5,6,7 e j=1,2,3). Dato che tale modello si dimostra adeguato con G^2=14.3 con df=6, *p-value* = (0.0257) è sufficiente studiare l'associazione marginale in tra le due variabili, anziché nella tabella completa 7×3, nella tabella collassata 4×3, dove i livelli della variabile reddito sono i seguenti:

Categorie della variabile reddito mensile •
0 – 600
601 - 1450
1451 – 2120
>2121

Anche per la variabile età del cliente si è proceduto in modo analogo al precedente, dapprima suddividendola in nove categorie. Stimando il modello è risultato che soltanto tre livelli sono sufficienti per descrivere l'associazione marginale con la variabile 'stato del conto al 12/98'. Le classi risultanti sono le seguenti.

Categorie della variabile età del cliente
18-44
45-54
55 e oltre

5.5 Stima della differenza tra le proporzioni

In tale paragrafo si riportano i risultati della stima tra le proporzioni di coloro che presentano un conto 'attivo', 'inattivo o dormiente', tra i proprietari con mutuo e tra i locatari. La variabile 'codice di residenza' presenta sei categorie elencate come in tabella 5.7 ed ha 15 valori mancati:

Categorie del codice di residenza	***Frequenze assolute***
Proprietario con mutuo	3327
Presso datori di lavoro	1333
Presso parenti	15258
Locatario	18031
Proprietario	25366
Altro	11
Valori mancanti	15
Totale	**63341**

Tabella 5.7 Frequenze assolute per le categorie della variabile codice di residenza.

La variabile risposta 'stato del conto al 12/97' è stata resa binaria aggregando le categorie di inattivo e dormiente, ovvero ponendo l'interesse su coloro che presentano un saldo attivo. La tabellina marginale binaria è la seguente.

Status di residenza	***Stato del conto al 12/97***		
	Inattivo/dormiente al 12/97	Attivo al 12/97	**Totali**
Proprietari con mutuo	697	2630	**3327**
Locatario	3827	14204	**18031**
Totali	**4524**	**16834**	**21358**

Tabella 5.8 Tabella doppia per due categorie dello status di residenza e lo stato del conto al 12/97.

La proporzione campionaria di attivi al 12/97 tra i proprietari con mutuo è pari a (2630/3327) = 0.79052, mentre tra i locatari essa ammonta a (14204/18031) = 0.78775. La differenza tra le proporzioni è pari a 0.00275 con errore standard 0.007684, da cui l'intervallo di confidenza è (0,01785204, -0,01231064). Poiché tale intervallo contiene lo zero si può concludere che la probabilità di avere il conto attivo al 12/97 è la stessa tra i proprietari con mutuo e locatari.

Considerando invece la seconda variabile risposta lo 'stato del conto al 12/98' si è stimato un modellino log-lineare, che consente di tener conto anche delle altre categorie, per vedere se anche in questo caso l'associazione del 'codice di residenza' con la variabile risposta poteva essere studiata non distinguendo tra proprietari con mutuo e locatari. La tabella incrociata per tale variabile con la variabile risposta stato del conto al 12/98 è riportata in tabella 5.9.

Codice di residenza	*Stato 12/98*			
	Inattivo	Dormiente	Attivo	Totale
Proprietario	10417	2304	12645	4720
Presso datori di lavoro	477	110	746	1333
Presso parenti	5866	1039	8083	14988
Locatario	5749	1451	10831	18031
Proprietario con mutuo	264	1012	2051	3327
Altro	4	2	5	11
Totali	**23525**	**5440**	**34361**	**63326**

Tabella 5.9 Valori osservati per la classificazione incrociata del reddito contro lo stato del cliente al 12/98.

In tal caso poiché la variabile in esame è nominale la matrice del contrasto utilizzata considera la prima categoria della variabile come riferimento:

$$C_6 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Il programma creato in *S-plus* per ottenere tali matrici di qualsiasi dimensione è riportato in appendice in A.2

Dal prodotto di Kronecker tra la matrice sopra e C_3 come definita prima, stimando il modello saturo e il modello ridotto in cui le categorie proprietario con mutuo e locatario sono state unite, si ottiene una devianza pari a $G^2 = 6.73$ con $df = 2$, $p\text{-}value = (0.0346)$ da cui si conclude che c'è indipendenza nella sottotabella 2×3 tra i livelli 1,2 del 'codice di residenza' e i livelli 1,2,3 della variabile 'stato del conto al 12/98'. Anche per l'associazione marginale con la seconda variabile risposta è irrilevante distinguere tra proprietari con mutuo e locatari.

5.6 Modello grafico a catena

L'ordinamento appropriato per le variabili considerate è rappresentato dal grafo illustrato in figura 5.1

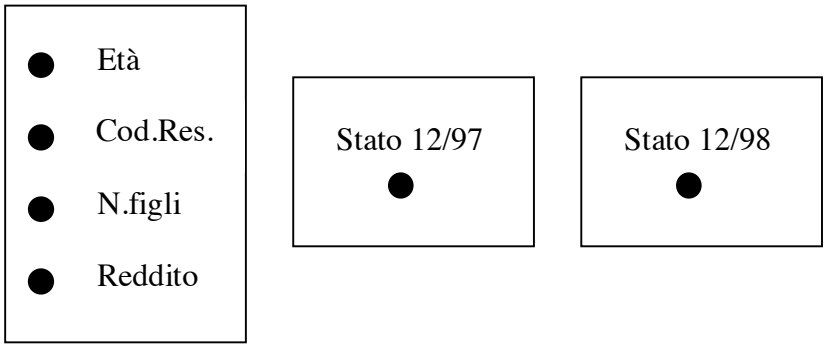

Figura 5.1 Grafo rappresentante l'ordinamento ipotizzato per le variabili.

Le variabili sono state raggruppate in tre blocchi distinti: nel primo riquadro sono riportate le variabili che descrivono la condizione sociale del cliente, queste sono potenziali cause delle variabili nei riquadri successivi quali lo 'stato del conto al 12/97' e al '12/98'. Il secondo blocco è la variabile risposta per il blocco precedente mentre esplicativa per il blocco seguente giacché permette di predire lo stato finale del conto 'al 12/98' per ogni stato del '12/97'. Ad esempio è possibile avere una prima idea di quale parte del portafoglio ha maggiore probabilità di chiudere il conto se al 12/97' lo stato del conto risulta attivo.

Inizialmente viene stimato un modello grafico non orientato per le variabili nel primo blocco, poi si seleziona il modello per la distribuzione congiunta della variabile del secondo blocco, condizionatamente a quelle del primo blocco, si aggiunge poi il terzo blocco e si stima la distribuzione della variabile di questo, condizionatamente alle variabili nei primi due.

Le variabili del primo blocco formano una tabella di contingenza non molto ampia grazie anche alla categorizzazione in variabile binaria del codice di residenza. Si è preferito, infatti, aggregare le categorie di tale variabile in 'proprietari della casa' e 'non proprietari', aggregando in un'unica classe i locatari, i proprietari con mutuo, coloro che abitano presso i datori di lavoro e altri. Ciò ha consentito di avere un modello più parsimonioso e maggiormente interpretabile ed ovviare ai problemi di sparsità della tabella.

N.FIGLI (B)	REDDITO (A)	COD.RES. (D)					
		Proprietario			Non proprietario		
		ETA (C)					
		17-44	45-54	≥55	17-44	45-54	≥55
0	1-600	ab	ab	abc	ab	ab	ab
	600-1450	abc	abc	abcd	abcd	abc	abc
	1451-2120	abc	abc	abcd	abcd	abc	abcd
	>2121	abc	abc	abcd	abc	abc	abcd
1-2	1-600	ab	ab	abc	ab	ab	ab
	600-1450	abc	abc	abc	abc	abc	abc
	1451-2120	abc	abcd	abcd	abc	abcd	abcd
	>2121	abc	abcd	abcd	abcd	abcd	abcd
≥3	1-600	a	a	ab	a	ab	ab
	600-1450	ab	ab	ab	abc	abc	ab
	1451-2120	abc	abc	abc	abc	abc	abc
	>2121	abc	abc	abc	abc	abc	abc

Tabella 5.10 Tabella di contingenza per le variabili del primo blocco dove il numero delle lettere è indicativo del numero delle cifre decimali presenti nella cella corrispondente.

La tabella considerata è rappresentata in tabella 5.10 dove i valori di ogni cella sono stati sostituiti con delle lettere per rispettare la riservatezza dei dati del fornitore commerciale.

Le celle della tabella sono 72 e non ci sono celle vuote, inoltre le frequenze all'interno di ogni cella sono abbastanza elevate.

Dato che tutte le variabili sono ordinali si è utilizzata di nuovo la parametrizzazione di Wermuth e Cox e si è stimato il modello log-lineare saturo, una volta ottenuta una stima dei parametri d'interazione standardizzati, si è proceduto ad individuare ipotesi d'indipendenza da verificare.

Per valutare l'ipotesi d'indipendenza tra due variabili condizionatamente alle altre, ad esempio tra N.figli (B) e il Reddito (A) per ogni classe di età e del codice di residenza, si è proceduto a vincolare a zero i logaritmi degli *odds ratio* corrispondenti ai coefficienti d'interazione tra A e B così come tutti i coefficienti d'interazione di ordine tre ADB e ACB e le interazioni del quarto ordine ABCD. Nell'ipotesi che il modello d'indipendenza sia un buon modello si potrà rappresentare il grafo senza l'arco che collega le variabili A e B.

Dal modello risulta però una forte associazione tra tutte le variabili oggetto di studio. Il modello con assenza del fattore d'interazione d'ordine quattro presenta devianza pari a 201.34 con $df = 12$, *p-value* = (0.0000), sono infatti pari a (4-1)(3-1)(3-1)(2-1)= 12 i parametri che vanno vincolati a zero.

L'unica semplificazione trovata tra le variabili è stata l'indipendenza tra le prime due classi del reddito e la seconda e terza dei figli, condizionatamente ad ogni classe d'età e del codice di residenza. Il modello ridotto, infatti, in cui sono stati posti uguali a zero gli *odds ratio* tra i livelli 1,2 del reddito e dei figli 2,3 in ogni sottotabella relativa all'età e al codice di residenza, presenta devianza pari a 4.9731 con $df = 6$, *p-value* = (0.5472) ma tale informazione non è significativa ai fini del presente studio dato che non coinvolge le due variabili risposta.

Da cui il modello grafico risultante per le variabili del primo blocco è riportato in figura 5.2

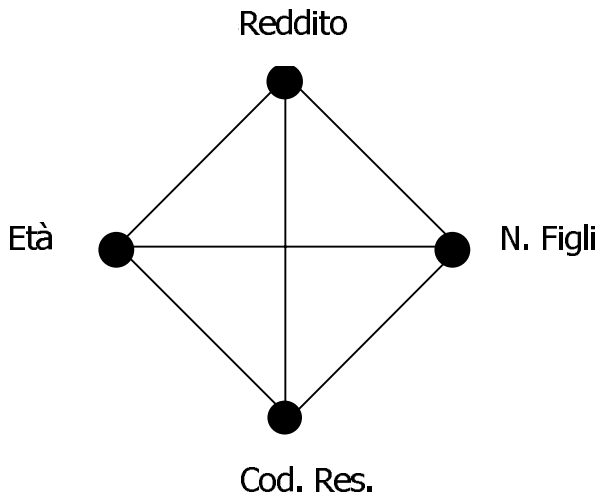

Figura 5.2 Grafo relativo al modello selezionato per le variabili del primo blocco.

Per le variabili dei blocchi seguenti che sono le due variabili risposta si è verificato, utilizzando il software CoCo (Badsberg 1995), se alcuni archi orientati del modello grafico potevano essere rimossi, ovvero se vi era indipendenza tra la variabile risposta 'stato del conto al 12/97' e una o più variabili del primo blocco così come tra la variabile risposta 'stato del conto al 12/98' e una o più variabili dei blocchi precedenti condizionatamente alle altre variabili.

Per le variabili dei primi due blocchi la tabella di contingenza è formata da 216 celle con alcune celle vuote, è stato pertanto utilizzato il test esatto del rapporto di verosimiglianza condizionato. Tuttavia tutte le variabili del primo blocco sono molto associate con lo 'stato del conto al 12/97', infatti, nessun arco orientato può essere rimosso.

Inoltre anche per le variabili del primo blocco così come per la variabile del secondo si ha un'associazione molto stretta con la variabile 'stato del conto al 12/98'. La tabella di contingenza presenta 648 celle di cui circa il

15% vuote, anche in questo caso si è utilizzato il test esatto, ma nessun arco può essere rimosso. Tale risultato può essere spiegato con il fatto che tutte le variabili esplicative ad eccezione dello 'stato del conto al 12/97' vertono sul medesimo oggetto: la condizione socio-economica, pertanto è possibile che ci sia una forte correlazione tra di esse.

Nella tabella seguente vengono riportati i risultati del test effettuato in CoCo per la rimozione degli archi relativi al terzo blocco che come si può vedere risultano tutti significativi. In appendice in A.6 vengono descritti i comandi utilizzati per ottenere tali stime.

-drop edge **CF**
Test of [[ABCDE] [ABDEF]]
Against [ABCDEF]
G^2 417.7278 P= 0.0000 P-Exact = 0.0000
X^2 384.3117
DF 252

-drop edge **AF**
Test of [[ABCDE] [BCDEF]]
Against [ABCDEF]
G^2 478.383 P= 0.0000 P-Exact = 0.0000
X^2 459.8617
DF 298

-drop edge **BF**
Test of [[ABCDE] [ACDEF]]
Against [ABCDEF]
G^2 461.8326 P= 0.0000 P-Exact = 0.0000
X^2 434.5070
DF 261

-drop edge **DF**
Test of [[ABCDE] [ABCEF]]
Against [ABCDEF]
G^2 635.2265 P= 0.0000 P-Exact = 0.0000
X^2 459.8617
DF 187

-drop edge **EF**
Test of [[ABCDE] [ABCDF]]
Against [ABCDEF]
G^2 1.08E+04 P= 0.0000 P-Exact = 0.0000
X^2 1.09E+04
DF 262

Tabella 5.11 Risultati del test per la rimozione degli archi orientati del primo e secondo blocco. E ed F rappresentano rispettivamente la prima e la seconda variabile risposta, mentre le altre lettere si riferiscono alle restanti variabili come in tabella 5.9.

Dai risultati precedenti si può notare che l'arco meno significativo sembra essere quello tra l'età (C) e lo 'stato del conto del conto al 12/98'(F). Pertanto l'età non sembrerebbe avere particolare influenza come predditore dello stato del conto, date le informazioni relative alle altre variabili. Vi è, tuttavia, solo una debole evidenza verso tale ipotesi ed il test esatto non consente di eliminare l'arco che collega le due variabili.

L'unica semplificazione trovata è che il modello non contiene le interazioni al disopra del quarto ordine, ovvero la relazione tra tre variabili scelte a caso non varia al variare della quarta. La tabella 5.12 riporta i risultati della stima del modello.

Test of [[ABCD] [BCEF] [BDEF] [ADEF] [ACEF] [ABEF] [ABCF] [ABCE] [ABDF] [ABDE] [ACDF] [ACDE] [BCDF] [BCDE]
Against [ABCDEF]
G^2 219.5931 P-Exact = 0.0207
X^2 201.0069 P-Exact = 0.1245
DF 179

Tabella 5.12 Risultati della stima del modello con assenza del fattore d'interazione d'ordine quattro.

Tenuto conto di tutto ciò, il modello selezionato per le variabili in esame è può essere rappresentato visivamente attraverso il grafo seguente, in cui nessun arco è assente.

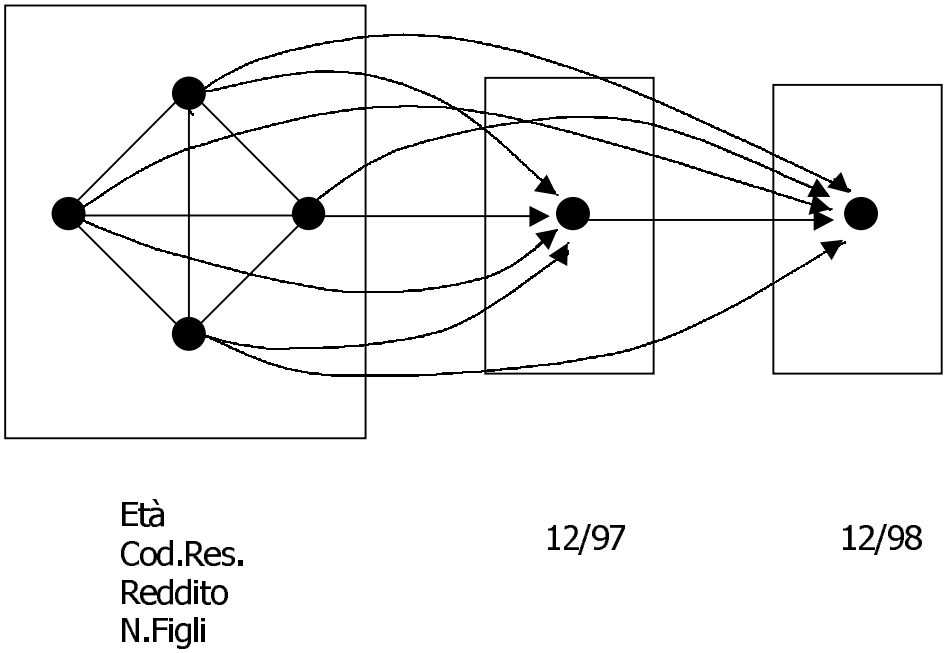

Figura 5.3 Grafo relativo al modello grafico a catena risultante dallo studio condotto.

5.7 Modello logistico cumulativo

Un modello logistico cumulativo con l'assunzione di *odds* proporzionali è stato stimato per la prima variabile risposta 'stato del conto del cliente al 12/97', contro le variabili del primo blocco, ed un secondo modello logistico per la seconda variabile risposta 'stato del conto del cliente al 12/98' contro le variabili del primo e secondo blocco.

Per la stima si è ricorso software statistico SAS; in A.5 dell'appendice è riportato il programma utilizzato.

Considerando lo 'stato al 12/97' il modello con *odds* proporzionali, con soltanto gli effetti principali delle variabile esplicative, si dimostra un modello adeguato. La tabella 5.13 riporta i risultati della stima del modello logistico cumulativo.

Score test for proportional odds assumption

Chi-Square = 7.2044 with 4 DF (p=0.1255)

Model fitting information and testing the global null hypothesis Beta=0

Criterion	Intercept Only	Intercept and covariates	Chi-Square for covariates
AIC	88022.213	87622.762	.
SC	88040.321	87677.085	.
-2 log L	88018.213	87610.762	407.451 with 4 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
Intercept1	1	-3.3677	0.0570	3490.0830	0.0001	.
Intercept2	1	-1.7645	0.0551	1024.6347	0.0001	.
Età	1	0.1454	0.0129	126.2370	0.0001	0.061281
Figli	1	-0.1585	0.0160	97.9508	0.001	-0.054819
Reddito	1	0.1377	0.1250	121.1515	0.001	0.059646
Cod.res	1	0.2408	0.0200	145.1677	0.001	0.065059

Tabella 5.13 Risultati della stima del modello logistico cumulativo per la variabile 'stato del conto al 12/97'

Lo *score test* per l'assunzione di *odds* proporzionali è pari a 7.2044, che rispetto ad una distribuzione del chi-quadrato con 4 gradi di libertà, risulta non significativo (*p-value* = 0.1255). Da cui il modello di *odds* proporzionali è appropriato per i dati analizzati.

Le stime di massima verosimiglianza sono ottenute in SAS attraverso l'algoritmo di Fisher (*Fisher scoring alghoritm*) che differisce dall'algoritmo di Newton-Raphson soltanto per il fatto che nella matrice delle derivate seconde sono usati i valori attesi piuttosto che quelli osservati.

Come illustrato al capitolo 3, il modello ipotizza la seguente relazione: $logit\ [P(\ Y \leq j)] = \alpha_j - \beta(numero\ delle\ righe)$; sostituendo il segno meno al più, si ottiene l'interpretazione di $\beta > 0$ come 'effetto positivo'. Per l'età il coefficiente $\hat{\beta}$ è pari a (- 0.1454) con errore standard di 0.0129. Per cui si deduce che per la classe d'età (i+1), l'*odds ratio* stimato per lo stato del conto maggiore di j è $e^{-0.1454} = 0.86$ volte l'*odds ratio* stimato per la classe d'età i, con i = 1,2. Allora la probabilità di avere il conto attivo al 12/97 è minore per coloro che hanno un'età più elevata, ovvero alti livelli di età sono associati con una scarsa movimentazione del conto. L'intervallo di confidenza al 95% per tale *odds ratio* è pari a exp[-0.1454±1.96(0.0129)] = (1.113424, 0.8430). L'*odds ratio* di essere attivo per l'ultima classe dell'età è pari a 0.75 volte l'*odds ratio* stimato per la prima classe d'età, infatti exp[-0.1454(3-1)] = 0.75.

Per la variabile 'numero dei figli' vale invece, la relazione opposta, dato che l'*odds ratio* stimato per la classe i+1 di questa variabile è pari a 1.17175 volte quello per la classe i = 1,2 si deduce che coloro che hanno un numero di figli maggiore tendono ad usare la carta più volte. L'intervallo di confidenza è pari a (1.20908, 1.1355).

L'uso più frequente della carta, dopo un anno dalla concessione, viene fatto da coloro che hanno bassi livelli di reddito rispetto a coloro che hanno reddito più elevato, infatti, l'*odds ratio* stimato per stimato per la classe i+1 di

questa variabile è pari a 0.87 volte quello per la classe i = 1,2,3 e l'*odds ratio* per l'ultima classe contro la prima è pari a 0.76.

Coloro che non sono proprietari della casa movimentano maggiormente il conto rispetto ai proprietari. L'*odds* per la categoria due della variabile 'codice di residenza' ovvero 'proprietari' è $e^{-0.2408} = 0.79$ volte l'*odds ratio* per i 'non proprietari'.

I risultati mettono anche in luce che il modello d'indipendenza con $\beta = 0$ non è plausibile, in quanto il test del rapporto di verosimiglianza per $\beta = 0$ è pari a 407.451 con $df = 4$, da cui si deduce che tutte le variabili esplicative nel modello considerato sono importanti predittori per 'lo stato del conto al 12/97' come verificato con CoCo.

Si è notato, inoltre, che aggiungendo le interazioni tra le variabili, il modello non migliora in quanto, il valore del *AIC* (*Akaike information criterion) = -2 log L + 2(k+s)*, dove k è il numero delle categorie ordinate della variabile risposta, mentre s è il numero delle variabili esplicative, è comunque pari a 82022 per entrambi i modelli. Tale valore si è dimostrato essere il più basso rispetto a tutti i modelli stimati in cui una o più variabili non venivano incluse.

Per quanto riguarda la variabile risposta 'stato del conto al 12/98', inizialmente si nota che il modello logistico cumulativo per tale variabile contro la variabile 'stato del conto al 12/97', non consente di accettare l'assunzione di *odds ratio* proporzionali, infatti, lo *score test* ha valore pari a 224.1218 con $df = 1$. Tale risultato è plausibile dato il fatto che, come visto in precedenza dall'analisi delle tabelle, coloro che sono dormienti al 12/97 tendono ad essere inattivi al 12/98 e non attivi come postula il modello considerato.

Dalla stima del modello logistico per l'ultimo blocco contro i primi due si evidenzia che l'assunzione di *odds* proporzionali non è valida data la presenza

della variabile stato del conto al 12/97, il modello presenta, infatti, uno score test pari a 260.4356 con *df* = 5.

Il modello logistico cumulativo si dimostra adeguato se si considerano come variabili esplicative per lo stato al 12/98 solo quelle del primo blocco. I risultati del modello sono illustrati in tabella 5.14.

Score Test for the Proportional Odds Assumption

Chi-Square = 10.3612 with 4 DF (p=0.0348)

Model Fitting Information and Testing the Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and covariates	Chi-Square for covariates
AIC	115099.44	114290.09	.
SC	115117.55	114344.41	.
-2 log L	115095.44	114278.09	817.35 with 4 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
Intercept1	1	-1.1628	0.0570	3490.0830	0.0001	.
Intercept2	1	-0.8034	0.0551	1024.6347	0.0001	.
Età	1	0.1310	0.0129	126.2370	0.0001	0.061281
Figli	1	-0.1833	0.0160	97.9508	0.001	-0.054819
Reddito	1	0.1448	0.125	121.1515	0.001	0.059646
Cod.res	1	0.3673	0.0200	145.1677	0.001	0.065059

Tabella 5.14 Risultati della stima del modello logistico cumulativo per la variabile 'stato del conto al 12/98'.

Le relazioni tra le variabili vanno nella stessa direzione di quelle per lo 'stato del conto al 12/97'. Si ha, infatti, che coloro che hanno minore età

hanno maggiore probabilità di essere attivi anche al 12/98 oltre che al 12/97. In particolare sembra accentuata la relazione tra un numero di figli maggiore e la probabilità che il conto sia attivo, l'*odds ratio* per la classe i+1 di tale variabile è infatti 1,2 volte *l'odds ratio* per la classe i con i = 1,2. Per il codice di residenza l'*odds ratio* per i non proprietari è pari a 1.44 volte l'*odds ratio* per i proprietari.

Le persone più anziane tendono a movimentare di meno il conto; coloro che hanno infatti l'*odds ratio* di avere il conto 'attivo' per coloro che hanno il reddito basso è pari a 1.56 volte l'*odds ratio* per coloro con reddito più elevato.

Anche in questo caso il modello di completa indipendenza presenta una devianza molto elevata (pari a 817.36 con *df*= 4) da cui si deduce la forte associazione delle variabili esplicative con la risposta.

I modelli presi in esame consentono di avere una semplice interpretazione delle relazione tra la variabile risposta ordinale e le variabili esplicative. Per esaminare la bontà del modello e la validità dell'assunzione di *odds* proporzionali si è utilizzata un analisi 'informale' stimando dei modelli cumulativi separati, ottenuti collassando la variabile risposta in dicotomica, dapprima unendo le categorie in 'inattivo e dormiente' e poi le categorie di 'dormiente' ed 'attivo'.[39] L'assunzione di *odds* proporzionali non sembra essere violata sia nel caso in cui la variabile risposta sia lo 'stato del conto al 12/97' o 'al 12/98', dato che in entrambi i casi gli effetti delle variabili esplicative hanno stessi segni e valori non dissimili. La tabella 5.15 riporta i valori dei parametri stimati relativi al modello che ha come variabile risposta lo stato al 12/97.

[39] Tali modelli pur se hanno stime di ML indipendenti possono essere usati per diagnosticare il modello originale ed esaminarne la plausibilità delle ipotesi come suggeriscono Clogg C. e Shihadeh E. (1996), op. cit. pp.159-160.

Variabili esplicative	*Logit cumulativo*	*Modelli separati*	
	Parametri	***k*=1**	***k*=2**
α_1	-3.3677	-3.2058	-
α_2	-1.7645	-	-1.7795
Età	0.1454	0.1495	0.1014
Numero figli	-0.1585	-0.1580	-0.1645
Reddito	0.1377	0.1385	0.1889
Cod.Res.	0.2408	0.2456	0.1613

Tabella 5.15 Confronto tra i parametri stimati del modello logistico cumulativo, per la variabile dipendente 'stato del conto al 12/97' e i parametri di modelli separati ottenuti unendo la prima e la seconda categoria della variabile risposta (*k*=1) e la seconda e la terza (*k*=2).

Sarebbe tuttavia opportuno, come suggeriscono Ananth e Kleinbaum[40] comparare tali modelli con modelli alternativi ad esempio con il modello logistico per categorie adiacenti, per verificarne adeguatamente la validità delle assunzioni, così come fare delle analisi diagnostiche ulteriori quali l'analisi dei residui.

5.8 Profilo della popolazione

Per investigare diversi aspetti delle relazioni tra le variabili, è necessario manipolare la distribuzione di probabilità congiunta rappresentata dal modello grafico stimato in precedenza ed illustrato in figura 5.3. Questo può essere fatto utilizzando i sistemi esperti basati su algoritmi computazionali, per la teoria che li accomuna con i modelli grafici. Il modello grafico selezionato

[40] Ananth. C. e Kleinbaum D. (1997), 'Regression Models for Ordinal Response: A Review of Methods and Applications', International Journal of Epidemiology, p.1323.

fornisce al sistema esperto la struttura di base delle variabili e la funzione di probabilità sottostante.

In questo lavoro al fine di avere informazioni utili per l'istituto che rilascia la carta è stato utilizzato il sistema esperto Hugin (Olesen *et. al.* 1992) che tratta i grafi aciclici orientati. Esso consente di investigare il grafo individuando, attraverso condizionamenti, sottogruppi di popolazione con profili particolari.

Un'applicazione interessante è quella di prevedere sulla base dello stato al 12/97 il comportamento al 12/98 in differenti sottogruppi di clienti. Questo può essere fatto condizionandosi allo stato al 12/97 e valutando per diverse combinazioni delle variabili esplicative, la distribuzione di probabilità relativa allo stato al 12/98.

Dall'analisi risulta che la probabilità che un cliente sia inattivo al 12/98 essendo inattivo l'anno precedente è di circa il 71% come mostrano le tabelle 5.16 e 5.17 che riportano l'interfaccia di Hugin ottenuta condizionandosi sullo stato del conto inattivo al 12/97 e marginalizzando rispetto alle altre variabili.

Stato al 12/97	
* 100	Inattivo
	Dormiente
	Attivo

Stato al 12/98	
71.33	Inattivo
5.02	Dormiente
23.65	Attivo

Tabelle 5.16 e 5.17 Interfaccia di Hugin come si presenta essendosi condizionati sulla prima categoria della variabile 'stato del conto al 12/97' ed avendo propagato sullo 'stato dal conto al 12/98'.

In particolare tra coloro che sono inattivi al 12/97, coloro che hanno un numero di figli pari a tre o più hanno una probabilità ridotta al 68% di essere inattivi al 12/98.

Coloro che sono dormienti al 12/97 hanno una probabilità pari al 73% di risultare inattivi un anno dopo e tale probabilità cresce al 74% fra coloro che hanno un reddito nella classe più alta (maggiore di 2121 •) ed arriva al 76% per coloro che hanno questo reddito e sono proprietari della casa.

Se invece si considerano gli attivi al 12/97 di questi circa il 26% risultano inattivi al 12/98. Se sono proprietari della casa la probabilità di essere inattivi al 12/98 ammonta la 29% mentre cala al 23% per coloro che hanno un numero di figli maggiore di tre e un reddito nella prima classe (minore di 600 •).

Risulta, inoltre, che il profilo dei clienti che hanno maggiore probabilità di diventare attivi di nuovo una volta che essi hanno un saldo inattivo è dato da coloro che:

- hanno un età al disotto di 44 anni;
- reddito nella classe più bassa;
- numero di figli compreso tra uno e due;
- non proprietari della casa.

Per cui questa fascia della popolazione che sembra avere maggiore sensibilità nel passare da uno stato inattivo a quello attivo potrebbe essere quella a cui indirizzarsi con campagne di marketing mirate riguardanti anche l'offerta di nuovi prodotti.

La fascia della popolazione alla quale proporre tassi d'interesse ridotti potrebbe essere quella che utilizza di meno la carta ovvero quella che ha maggior probabilità di avere un conto inattivo sia al 12/97 che al 12/98. Il profilo individuato risulta essere:

- un età maggiore di 54 anni;
- nessun figlio;
- un reddito nella fascia più alta;

- proprietario della casa.

Il grafo può inoltre essere investigato in molti altri modi, i condizionamenti possibili sono, infatti, pari alle combinazioni delle variabili, per ognuno di essi è possibile esaminare tutti gli aspetti della distribuzione ritenuti interessanti. La flessibilità garantita dal sistema esperto implementato sul modello selezionato lo rende pertanto uno strumento molto utile sia come tecnica esplorativa che come strumento decisionale.

CONCLUSIONI

Il presente lavoro ha permesso di evidenziare le potenzialità offerte dalle nuove strategie e tecniche d'analisi multivariata presentate nel contesto del *behavioural scoring*.

In particolare, utilizzando modelli per variabili ordinali, è stato possibile categorizzare in modo accurato le variabili continue così da lavorare solo su modelli per dati categorici. È stato possibile, inoltre, ridurre il numero di categorie delle variabili categoriche in modo tale che i modelli siano più parsimoniosi e maggiormente interpretabili.

L'esito delle elaborazioni sulle variabili socioeconomiche dei titolari di carta di credito, attivata nei primi mesi del 1997, ha evidenziato una struttura delle variabili piuttosto complessa. Nessuna relazione d'indipendenza condizionata è stata trovata. Tuttavia considerare il modello grafico a catena ha permesso di modellare l'effetto delle variabili esplicative su entrambe le variabili risposta indicanti lo stato del conto del cliente: 'attivo', 'inattivo', 'dormiente', al dicembre 1997 e al dicembre 1998. Attraverso la rappresentazione grafica, l'analisi compiuta sulle variabili è interpretabile oltre che precisa nell'evidenziare le struttura delle dipendenze. Nello studio condotto sono risultati assenti i fattori d'interazione tra le variabili d'ordine superiore al quarto.

La stima del modello logistico cumulativo ha permesso di evidenziare che coloro che hanno un età più elevata, maggiore di cinquantacinque anni, risultano avere un conto 'inattivo' con maggiore probabilità in entrambi i periodi considerati. Il numero dei figli ha invece un effetto positivo sulla probabilità di avere il conto 'attivo' sia al 12/97 che al 12/98. Coloro che hanno livelli di reddito più bassi, minori di 600 • mensili, movimentano di più il conto rispetto a coloro che hanno alti livelli di reddito. I proprietari della

casa hanno minore probabilità di essere 'attivi' sia al 12/97 che al 12/98 rispetto ai non proprietari.

Le manipolazioni della distribuzione di probabilità congiunta del modello selezionato, realizzate attraverso il sistema esperto di Hugin, hanno permesso di mostrare le considerevoli potenzialità offerte dai modelli grafici, consentendo di evidenziare vari profili d'interesse della popolazione con comportamenti particolari. Ad esempio, è risultato che coloro che hanno un età al disotto di 44 anni, un reddito minore di 600 • mensili, un numero di figli compreso tra uno e due e non sono proprietari della casa, essendo 'attivi' al 12/97 è molto probabile che abbiano il conto di nuovo 'attivo' dopo un anno.

L'uso di tali modelli consente pertanto, una conoscenza più accurata della popolazione e dà la possibilità di instaurare relazioni attive con la clientela come proporre offerte mirate per nuovi prodotti, offrire commissioni di consulenza a prezzi inferiori, incentivare all'uso della carta. Aiutano cioè a mettere a punto azioni strategiche per la società, che risultano fondamentali, dato il contesto altamente competitivo in cui operano le istituzioni che offrono credito al consumo.

Un'analisi più accurata richiederebbe l'inserimento di altre variabili esplicative, anche al fine di aiutare la semplificazione della struttura delle relazioni tra le variabili, che è risultata alquanto complessa nel modello stimato. Un'altra variabile d'interesse può essere, ad esempio, il limite di credito concesso al cliente dalla società, ovvero la cifra massima che gli è consentito prendere in prestito. Questa è stabilita a seguito del processo di *score* e risulterebbe utile considerarla, per studiare meglio il comportamento di spesa, oltre che per verificarne le potenzialità come strumento di marketing che essa sembra avere. Un'utile variabile esplicativa per lo stato del conto al 12/98 è il numero dei prestiti erogati nel periodo precedente a tale data. Si potrebbe tener conto, inoltre, degli aspetti d'insolvenza del cliente, se egli effettua i rimborsi parziali in modo puntuale ogni mese o ha delle insolvenze

più o meno lunghe; così come la sua propensione attuale a contrarre prestiti, le preferenze circa l'ente finanziatore, o ancora la tipologia prevalente dei beni acquistati, che forniscono ulteriori indicazioni oltre quelle socio-economiche.

L'approccio descritto nel presente lavoro vuole quindi essere esplorativo in tale contesto e non pretende di esaurirne tutti gli aspetti ma soltanto introdurre ad ulteriori studi che debbono essere compiuti. Consente di evidenziare, tuttavia, che sfruttare i metodi multivariati e tenere in considerazione la scala ordinale con cui le variabili si presentano permette di giungere a conclusioni interessanti rispetto al comportamento prevalente in sottogruppi della clientela.

APPENDICE

Si presenta di seguito la descrizione dei principali programmi utilizzati per i software usati nelle analisi condotte al quinto capitolo.

Dapprima vengono descritte le funzioni create per il software S-PLUS; di seguito i comandi utilizzati in CoCo per il test circa la rimozione degli archi orientati del modello grafico e il programma per stimare il modello logistico cumulativo in SAS.

A.1 Descrizione della funzione per S-plus che consente di vettorizzare le frequenze di una tabella di contingenza.

```
vettcol <- function(matrice) {
as.vector(matrice)
}
```

A.2 Funzione che consente di creare le matrici del contrasto per caratteri nominali.

```
contrasto<- function(n){
nome<-matrix(nrow=n,ncol=n)
for(i in 1:n){
for(j in 1:n){
nome[i,j]<-0}
nome[i,i]<-1}
Icol<-c(rep(1,n))
for(i in 2:n){
Icol[i]<- -1}
for (i in 2:n){
nome[i,1]<-Icol[i]}
Icol
nome
}
```

A.3 Funzione che consente di creare le matrici del contrasto per caratteri ordinali.

```
contrasto1<-function(n){
nom<-matrix(nrow=n,ncol=n)
for(i in 1:n){
for(j in 1:n){
nom[i,j]<-0}
nom[i,i]<-1
}
Irig <- c(rep(n,1))
for(i in 1:n){
Irig[j]<-1}
for(j in 1:n) {
nom[1,j]<-Irig[j]
}
for(j in 1:n){
if(i= =j+1)
nom[i,j] <- -1}}
nom
}
```

A.4 Procedimento che consente di stimare il modello log-lineare con la parametrizzazione per variabili ordinali.

vettcol(*a*)-> *b*
kronecker (contrasto1 (*dimensione matrice*), contrasto1 (*dimensione matrice*))->*c*
data.frame (cbind (*b*, solve (*c*)))->*d*
glm (*b* ~ . –1, family = poisson, data = *d*)->*e*
coefficients (summary (*e*))->*f*
d [,-c(*parametri da togliere*)]->*g*
glm(*b* ~ . –1, family = poisson, data = *g*)->*e*
deviance (*e*)->*h*
pchisq (*e, gradi di libertà*)

A.5 Descrizione del programma che consente di stimare il modello logistico cumulativo in SAS. Le lettere si riferiscono alle variabili come in tabella 5.10 del quinto capitolo.

```
Filename  x ' c:\dati' ;

data y ;

infile x ;

input A B C D E F ;

data y ;

set x;

title 'modello logistico cumulativo';

proc logistic  ;

model E = A  B  C  D ;

run ;
```

A.6. Descrizione dei comandi utilizzati per valutare in CoCo la rimozione degli archi del grafo come descritto in tabella 5.10 del quinto capitolo.

```
set keyboard on
read factors A 4 / B 3 / C 3 / D 2 / E 3 / F  3//
set inputfile observations  (nome del file)
read table
set ordinal ABCDEF
read model ABCDEF
base
```

set exact test on

drop edge CF

BIBLIOGRAFIA

ANANTH C.V.e KLEINBAUM D. (1997), 'Regression Models for Ordinal Response: A Review of Methods and Applications', *International Journal of Epidemiology*, pp.1323-1333.

AGRESTI A. (1984), '*Analysis of ordinal categorical data*', Wiley, New York.

AGRESTI A.(1990), '*Categorical Data Analysis*', Wiley, New York.

ASMUSSEN S. e EDWARDS D. (1983), 'Collapsibility and response variables in contingency tables', *Biometrika*,70, pp.567-568.

AZZALINI A. (1996), '*Statistical Inference*', Chapman Hall.

BADSBERG J. H. (1995), '*A Guide to CoCo*', Research Report R 91-43, Insitute for Electronic Systems, University of Aalborg.

BIRCH M.W (1963), 'Maximum likelihood in the three-way contingency table', *Journal of Royal Statistic Society*', B25, pp.220-223.

CANDE V.A., DAVID G.K. (1997), 'Regression models for ordinal response: a review of methods and applications', *International Journal of Epidemiology*, Vol. 26, N.6, pp.1323-1333.

CHILANTI M. (1999), 'Un sistema adattivo per il supporto alle decisioni nelle analisi del rischio di credito', *Economia&Management*, 2, pp.113-126.

CHRISTENSEN R. (1990), '*Log-linear Models and Logistic Regression*', Sprinter-Verlag, New York,

CIAVARELLA C. (1999), '*Modelli grafici applicati al credit scoring*', Tesi di Laurea, Università degli Studi di Perugia.

CICCHITELLI G. (1990), '*Probabilità e Statistica*', Maggioli Editore, Rimini.

CLOGG C., SHIHADEH E., (1996), '*Statistical Models for Ordinal Variables*', Sage Publications, Thousand Oaks.

CORIGLIANO R. (1998), '*Il rischio di credito e pricing dei prestiti bancari*', Bancaria Editrice, Roma

CORBETTA P. (1992), '*Metodi di analisi multivariata per le scienze sociali*', Bologna, Il Mulino.

CROOK J.N., HAMILTON R., THOMAS L.C. (1992), 'The Degradation of the Scorecard over the business cycle', IMA *Journal Mathematics Applied in Business & Industry*, 4, pp.111-123.

DARROCK J.N., LAURITZEN S.L., SPEED T.P. (1980), 'Markow fields and log-linear interaction models for contingency tables', *The Annals of Mathematical Statistics*, 8, pp.522-539.

DAWID A.P. (1979), 'Conditional indipendence in statistical theory', *Journal of the Royal Statistical Society*,B, 39, pp.1-38.

DE LUCA A. (1998), '*Marketing bancario e metodi statistici applicati*', Franco Angeli Editore.

DI ANTONIO M. (1994), '*Il credito al consumo*', EGEA, Milano.

EDWARDS E.D. (1995), '*Introduction Graphical Modelling*', Springer-Verlag.

FILOTTO U GIANNASCA C (1996), ' Credito al consumo: qualità del credito e gestione del rischio', *Banche e Banchieri*,3, pp. 241-250.

FORCINA A.(1996), '*Appunti di Statistica Descrittiva*', Cafaro Editrice, Perugia.

FRYDENBERG M. (1989), 'The chain graph Markov property', *Scandinavian Journal of statistics*,17, pp.333-353.

GREENLAND S. (1994), 'Alternative models for ordinal logistic regression', *Statistics in Medicine*, 13, pp.1665-77.

HAND D.J. e HENLEY W.E. (1993), 'Can Reject inference even work?', IMA *Journal Mathematics Applied in Business & Industry*, 5, pp. 45-55.

HAND D.J. e JACKA S. (1998), '*Statistics in finance*', Edward Arnold, London.

HAND D.J, McCONWAY M.J., STANGHELLINI E. (1997), 'Graphical models for applicants for credit', IMA *Journal Mathematics Applied in Business. Industry*, 8, pp. 143-155.

LAURITZEN S. L. (1996), '*Graphical Models*', Oxford University Press, Oxford.

LEWIS E. (1994), '*An Introduction to Credit scoring*', The Athena Press, California.

MAINO R. (1998), ' Nuove metodologie di gestione del rischio di credito e vantaggi competitivi per le banche', *Economia&Management*, 6, pp.73-92.

McCULLAGH P., NELDER J.A. (1989), '*Generalized Linear Models*', Chapman and Hall, London.

OLESEN K.G., LAURITZEN S.F., JENSEN F.V. (1992), 'Hugin: a system creating adaptive causal probabilistic networks', *Uncertainty in artificial Intelligence 8*, Morgan Kaufmann, San Matteo, CA.

PACKETT R.L. (1981), '*The Analysis of Categorical Data*', Griffin, London.

PEARL J., PAZ (1986), 'Graphoids. A graph-based for reasoning about relevancy relations.' *Proceedings of the 7th European conference on Artificial Intelligence*, Bringhton U.K .

NERI A. (1999). '*I modelli grafici a catena per variabili ordinali: un'applicazione nel campo del behavioural scoring*', Tesi di dottorato, Università degli Studi di Firenze.

ROCCATO A. (1999), 'Le reti neurali nella soluzione SAS per il data mining', *Scienza&Business*, 1, pp.74-76.

SIMPSON C.H. (1951), 'The interpretation of interaction in contingency tables', *Journal of Royal Statistical Society*, B, 13, pp.238-241.

STANGHELLINI E., McCONWAY M.J., HAND D. J (1999), 'A discrete variable chain graph for applicants for credit', *Applied Statistics*, 48, 2, pp. 241-266.

STANGHELLINI E., BIGGERI A. (1998), 'Uso di modelli grafici per l'analisi della relazione tra l'inquinamento atmosferico e disturbi respiratori nell'infanzia', *Atti XXXIX riunione scientifica SIS*, Maggioli Editore, Rimini, pp.215-223.

SWEGÖ G. VARETTO (1999), '*Il Rischio Creditizio*', UTET, Torino.

SWEART P. e WHITTAKER J. (1998), 'Graphical models in credit scoring', IMA, *Journal of Mathematics Applied in Business & Industry*, 9, pp.224-266.

TAGLIAVINI G. (1988), 'Valutazione del merito creditizio nei prestiti personali e nel credito al consumo', *Il Risparmio*, 2, Marzo-Aprile.

THOMAS L.C. (1999), 'A Survey of Credit and Behavioural scoring, Forecasting financial risk of lending to customer', Proceedings of Credit Scoring and Credit Control VI.

WERMUTH N. COX D.R. (1998). 'On the application of conditional indipendence to ordinal data', *International Statistical Review*, 66, 2, pp.181-199.

WERMUTH N. (1992), 'On the Relation Between Interactions Obtained with Alternative Codings of Discrete Variables' , *Methodika* 6, pp. 76-85.

WHITTAKER J. (1990), '*Grafical Models in Applied multivariate Statistics*'. Wiley , Chichester.