Alma Mater Studiorum - Università di Bologna

Dipartimento di Scienze statistiche "Paolo Fortunati"
Corso di Laurea Magistrale in Scienze statistiche, finanziarie e attuariali

Modelli GLM per il lapse risk nei fondi pensione

Tesi di laurea in Modelli statistici per le scienze attuariali

Presentata da:
Riccardo Riminucci
N. Matr.0001058963

Relatore:
Chiar.mo Prof.
Paolo Foschi

Anno Accademico 2022-2023

Ai miei genitori, per essere sempre stati al mio fianco e avermi sempre sostenuto.
Ai miei amici, per avermi dato spensieratezza quando più ne avevo bisogno, per aver condiviso tanti bei momenti, nella speranza di viverne altrettanti ancora.
Ai miei nonni, che mi chiedevano sempre degli esami all'università.
Un sentito ringraziamento va al Professor Paolo Foschi, per la disponibilità dimostrata.

Abstract

La previdenza complementare sta acquisendo una rilevanza sempre maggiore nei Paesi occidentali e non solo. Il peso specifico degli agenti privati nel settore della previdenza è in continua crescita. Questa transizione dal pubblico al privato si sviluppa contestualmente a un invecchiamento sempre più rapido della popolazione. In questo scenario è inevitabile che i fondi pensione assumano un ruolo centrale e che diventino uno dei principali temi di interesse della ricerca, accademica e non. In questo lavoro ci focalizzeremo sul fenomeno dei riscatti nei fondi pensione, indagando i possibili fattori che influenzano il fenomeno. Ricaveremo quindi dei modelli, afferenti alla famiglia dei Generalized Linear Models, con cui cercheremo di prevedere le dinamiche del fenomeno dei riscatti.

Indice

Introduzione

L'attività di assicurazione deve le sue origini a forme mutualistiche di assistenza tra cittadini e nasce dunque da delle esigenze di privati. Nel corso della sua evoluzione storica, l'attività assicurativa ha tuttavia assunto un ruolo sempre più diffuso e rilevante nella società, spesso diventando materia di pubblica utilità. La previdenza è diventata così un onere in capo (in tutto o in parte) alle finanze pubbliche in molte economie avanzate.

L'attività assicurativa nel suo complesso è quindi permeata da un'ambivalenza di interessi al contempo pubblici e privati. Si potrebbe sostenere che la massima espressione di questa ambivalenza si raggiunga nelle forme di previdenza complementare. A partire da interessi privati, i cittadini scelgono di integrare la previdenza del settore pubblico. Quest'ultimo, per motivi di pubblica utilità, incentiva i privati ad affidarsi a questa forma di previdenza parallela e accessoria. Vi è quindi un continuo rimando di interessi pubblici e privati che influisce sull'attività assicurativa in tutte le sue sfaccettature.

Il tema dei riscatti nei fondi pensione coinvolge quindi inevitabilmente interessi sia pubblici sia privati. In primo luogo, i riscatti da forme di previdenza complementare tolgono risorse da questo pilastro della previdenza nazionale. In secondo luogo, un riscatto deriva da una decisione individuale privata. Nel caso di prodotti previdenziali offerti da compagnie assicurative, la decisione di riscattare queste polizze si riverbera anche sul settore privato delle assicurazioni.

Il fenomeno dei riscatti ha quindi degli effetti rilevanti e di ampia portata, per cui diventa cruciale capire o almeno rilevare quali siano i fattori più incisivi alla base dell'esercizio di questa opzione. Solo una buona comprensione del fenomeno può gettare le basi per un successivo tentativo di previsione dei riscatti.

La ricerca delle cause e il tentativo di previsione dei riscatti costituiscono l'obiettivo di fondo di questa tesi.

La struttura del lavoro si compone delle seguenti sezioni:

- Capitolo 1: verranno brevemente illustrati i fondi pensione e con essi la normativa relativa ai riscatti degli stessi.
- Capitolo 2: verranno presentati vari paper e articoli scientifici che trattano del tema dei riscatti nel ramo vita, con approcci prevalentemente di tipo GLM.
- Capitolo 3: si darà una prospettiva tecnica sulla metodologia utilizzata in questo lavoro, ovvero sui Generalized Linear Models.
- Capitolo 4: si farà una breve panoramica sulle metriche utilizzate per valutare quanto sia preciso nelle previsioni un classificatore binario.
- Capitolo 5: si approfondirà l'analisi di un case study, inerente alla modellizzazione e previsione dei riscatti di un fondo pensione PIP.
- Conclusioni: verranno sintetizzati i risultati ottenuti nella parte di analisi, che verranno poi confrontati con quelli della letteratura esistente, evidenziando eventuali punti di discordanza.

Capitolo 1

Il riscatto nei fondi pensione

Le grandi economie occidentali stanno attraversando da ormai qualche decennio la fase conclusiva della transizione demografica, con una piramide delle età ormai invertita. In un contesto di forte sbilanciamento tra percettori e finanziatori della previdenza pubblica, il patto generazionale è stato messo in serio pericolo. Questa tendenza è stata anticipata dai mercati, portando alla nascita di numerose forme di previdenza complementare che vadano ad alleviare il peso delle prestazioni in capo alla previdenza pubblica. La combinazione di un equilibrio demografico precario e della crescente privatizzazione del settore previdenziale (non solo in Italia) ha incrementato notevolmente la diffusione e l'importanza dei fondi pensione. Nel presente capitolo illustreremo molto sinteticamente gli aspetti principali di un fondo pensione e faremo un approfondimento sulla normativa relativa al riscatto di un fondo pensione.

1.1 I fondi pensione in Italia

Il Ministero del Lavoro e delle Politiche Sociali (MinLav) illustra con un focus sintetico le caratteristiche del secondo pilastro del sistema pensionistico, ovvero la previdenza complementare, il cui scopo è appunto quello di integrare la copertura previdenziale fornita dal primo pilastro (previdenza di base obbligatoria). La previdenza complementare infatti si contraddistingue per il fatto di non essere a carattere obbligatorio, motivo per cui il legislatore negli anni ha deciso di creare delle misure di incentivo e di detassazione che potessero invogliare i cittadini a ricorrere a questa copertura previdenziale aggiuntiva. La logica di base di tutte le forme di previdenza complementare risiede nella raccolta di risparmio previdenziale che va a concorrere, assieme ai rendimenti maturati spettanti e al netto dei costi, alla posizione previdenziale individuale del contribuente. Le tipologie di fondi pensione possono tuttavia

differire anche in maniera significativa. Riportiamo da MinLav un elenco con una breve descrizione delle varie tipologie dei fondi:

- I fondi chiusi (art. 3 del D.lgs. 252/2005) di origine "negoziale", sono forme pensionistiche complementari istituite dai rappresentanti dei lavoratori e dei datori di lavoro nell'ambito della contrattazione nazionale, di settore o aziendale.

- I fondi aperti (art. 12 del D.lgs. 252/2005) sono forme pensionistiche complementari istituite da banche, imprese di assicurazioni, società di gestione del risparmio (SGR) e società di intermediazione mobiliare (SIM).

- I Piani pensionistici individuali (PIP) (art. 13 del D.Lgs. 252/2005), rappresentano i contratti di assicurazione sulla vita con finalità previdenziale. Le regole che li disciplinano non dipendono solo dalla polizza assicurativa ma anche da un regolamento basato sulle direttive della COVIP. Lo scopo è garantire all'utente gli stessi diritti e prerogative analoghi alle forme pensionistiche complementari.

- I fondi pensione preesistenti. Si tratta dei fondi pensione già esistenti al 15 novembre 1992, ovvero prima del Decreto legislativo del 21 aprile 1993, n. 124 (provvedimento abrogato dal D.lgs. 5 dicembre 2005, n. 252) che ha istituito la previdenza complementare. Questi fondi hanno caratteristiche proprie che li distinguono dai fondi istituiti successivamente. Possono, ad esempio, gestire direttamente le risorse senza ricorrere a intermediari specializzati. Si tratta di Fondi collettivi per i quali l'adesione dipende da accordi o contratti aziendali o interaziendali.

1.2 Legislazione sui riscatti nei fondi pensione

L'istituzione di riferimento per la vigilanza normativa sui fondi pensione viene detta COVIP, acronimo della Commissione di Vigilanza sui fondi Pensione. La commissione è stata istituita nel 1993 con il D. Lgs. 124/1993 come autorità preposta alle funzioni di vigilanza sulle forme pensionistiche complementari. La disciplina in materia è contenuta nel D.Lgs. 252/2005 (si veda DLgs). La normativa è sicuramente ricca e dettagliata, in questa sede tuttavia la nostra attenzione sarà focalizzata sulla parte inerente ai riscatti. Nel testo di legge del D.Lgs. 252/2005, art. 14 comma 2, si elencano le eventualità e le condizioni in cui è possibile riscattare la polizza, in caso di riscatto parziale o di riscatto totale. Riassumendone il contenuto, il COVIP (si veda il Sito COVIP FAQ) evidenzia che:

- È possibile riscattare la posizione individuale in forma parziale, nella misura del 50 per cento della posizione individuale maturata nei casi di:
  - cessazione dell'attività lavorativa che comporti inoccupazione per un periodo non inferiore a 12 mesi e non superiore a 48 mesi;
  - mobilità, licenziamento, cassa integrazione ordinaria o straordinaria;
- E' consentito il riscatto totale nei casi di:
  - invalidità permanente, da cui deriva una riduzione della capacità lavorativa a meno di un terzo;
  - inoccupazione oltre i 48 mesi;
  - perdita dei requisiti di partecipazione al fondo (ad esempio per licenziamento).

Sia riscatti totali che parziali rappresentano una circostanza generalmente negativa per l'ente che eroga la prestazione, poiché lo costringono a dover liquidare anticipatamente delle posizioni, esponendolo anche alla possibile realizzazione di minusvalenze latenti. Focalizzeremo la nostra attenzione sui Piani Individuali Pensionistici: chi eroga la prestazione è la compagnia assicuratrice, che dovrà far fronte a una richiesta di liquidità, necessaria a pagare anticipatamente le prestazioni.
Si sottolinea anche che per la compagnia un trasferimento in uscita della posizione previdenziale è di fatto assimilabile a un riscatto totale. Il contraente può infatti decidere, trascorsi due anni di iscrizione alla forma pensionistica complementare, di trasferire la propria posizione individuale verso un'altra forma pensionistica complementare iscritta all'albo COVIP. Concludiamo sottolineando che molti prodotti PIP attualmente in commercio in Italia prevedono un periodo iniziale minimo durante il quale non è possibile riscattare la polizza.

Capitolo 2

Stato dell'arte

In questa sezione verranno illustrati sinteticamente alcuni articoli scientifici che trattano con vari metodi il tema della previsione del rischio di riscatto nelle polizze assicurative.

2.1 Cerchiara et al.

Un primo paper che ha avuto un discreto numero di citazioni è senz'altro quello degli autori Cerchiara, Edwards e Gambini (Cerchiara et al. [2009]). Il paper propone nuovi orizzonti di utilizzo dei modelli GLM nell'ambito della previsione dei decrementi, nel quadro delle polizze vita. In questo lavoro gli autori elencano varie caratteristiche di una polizza vita che un assicuratore potrebbe (o dovrebbe, secondo normativa) analizzare per produrre una best estimate delle proprie obbligazioni e contestualmente dei propri cash flows. In questo scenario quindi gli autori decidono di soffermarsi sui tassi di decremento, dove con decremento intendono eventi di lapse e mortalità. I tassi di riscatto (lapse rate) assumono un ruolo centrale nella modellazione dei futuri cash flows generati da una polizza. La natura del fenomeno dei riscatti infatti sembra essere molto più volatile rispetto alla variabilità dei tassi di mortalità e per questo quindi anche molto più impattante sull'affidabilità delle stime best estimate. Un'interessante considerazione formulata in questo lavoro rimarca una distinzione tra un livello base di riscatti, detto componente irrazionale, e un livello variabile di riscatti che varia con il mercato e che viene pertanto definito come componente razionale. La reattività dei contraenti (policyholders) alle variazioni di mercato dipende anche dal tipo di policyholder, in primo luogo dal suo grado di consapevolezza e di conoscenze nel valutare la redditività di una polizza, ma anche di un investimento alternativo. Come sottolineano gli autori, il loro studio è volto non tanto al pricing di una surrender option, quanto a una precisa calibrazione del lapse risk. Nella sezione dedicata al lapse risk, gli autori propongono un modello GLM di

tipo logit con distribuzione binomiale; propongono in alternativa un modello GLM di tipo logaritmico e distribuzione Poisson, ma solo nel caso in cui l'analisi sia di tipo più qualitativo e/o rivolta a un'audience senza l'adeguata preparazione tecnica. Gli autori ipotizzano un'interessante e possibile applicazione di un modello GLM per i riscatti: un modello che tenga conto del calendar year of exposure permette di legare indirettamente i riscatti alle variabili di mercato (es. rendimento di bonds a medio-lungo termine), ottenendo così una buona stima della sensitività dei riscatti rispetto al mercato. Una tale relazione riveste grande importanza nella calibrazione di modelli interni e di Economic Scenario Generators.

2.1.1 Il case study

Per quanto riguarda l'analisi, gli autori hanno preso in considerazione i dati di riscatto di una grande compagnia italiana di bancassicurazione, rilevati nel periodo dal 1991 al 2007 per polizze vita di risparmio. In tutto il dataset sono stati riscontrati 279 mila riscatti su oltre 6 milioni di anni-polizza di esposizione. La terminologia anno-polizza tornerà utile anche in seguito poiché fa riferimento a una metodologia di analisi che verrà adottata in questo lavoro di tesi. Il concetto di anno-polizza rimarca il fatto che una stessa polizza costituisca più osservazioni nel dataset, dove ciascuna osservazione corrisponde alla determinata polizza osservata in uno specifico anno. In altri termini, la combinazione di anno di osservazione e numero polizza costituiscono, congiuntamente, la chiave univoca del dataset.

Le variabili considerate nel modello sono le seguenti: Product, Year of exposure, Duration, Year of policy inception. I risultati del modello GLM vengono illustrati con un accento sulle stime dei coefficienti di relatività e sulle rispettive bande di confidenza. Questi risultati vengono poi confrontati con un modello a un solo fattore, mostrando come quest'ultimo non riesca a catturare le correlazioni nei dati. Stando ai risultati ottenuti dagli autori, la durata della polizza (duration) influisce in modo incisivo sulla propensione al riscatto, con dei picchi in corrispondenza di 2 e 5 anni, presumibilmente legati a delle decorrenze fiscali. Sembra invece che la propensione al riscatto si riduca notevolmente oltre i 10 anni di durata della polizza: gli autori interpretano una simile dinamica nella prospettiva del processo di autoselezione dei policyholder, ovvero che oltre un certo periodo di permanenza i sottoscrittori rimasti siano molto poco propensi a riscattare la polizza. Anche l'anno di calendario di esposizione (calendar year of exposure) e l'età sono risultati significativi nell'influenzare i tassi di riscatto.

2.2 Milhaud et al.

In questo lavoro (Milhaud et al. [2011]) gli autori Milhaud, X., Loisel, S. e Maume Deschamps, V. si propongono di indagare quali siano le variabili determinanti che portano ad un innalzamento del livello dei riscatti. La motivazione scaturisce dalle conseguenze dei riscatti sul business delle compagnie di assicurazione. Gli autori individuano tre modi in cui i riscatti danneggiano le compagnie. In primo luogo, un elevato livello di riscatti potrebbe impedire alla compagnia di rientrare dei vari costi che sostiene prima della stipula o contestualmente ad essa. Un elevato livello di riscatti potrebbe inoltre tradursi in una selezione negativa di portafoglio, in cui solo i policyholder con uno stato fragile di salute non riscattano la polizza. Ultimo ma non meno importante, un elevato livello di riscatti nell'immediato ha effetti non ininfluenti sulla liquidità di una compagnia, che è soggetta al rischio di oscillazioni del tasso di interesse. Gli autori elencano poi le varie teorie riscontrabili ad oggi nella letteratura sul tema dei riscatti. Viene citata la teoria detta "Emergency Fund Hypothesis", secondo cui i policyholders decidono di riscattare la polizza per ottenere il valore di riscatto della polizza e utilizzarlo in momenti in cui hanno problemi di liquidità. Un'altra teoria proposta dalla letteratura, che chiameremo "Interest rate hypothesis", assume che i tassi di riscatto aumentino in corrispondenza di aumenti del tasso di interesse poiché questi rendono più redditizi i nuovi titoli (o le nuove polizze) emessi e i policyholder decidono di riscattare la polizza e investire il valore di riscatto in questi titoli di nuova emissione. Una simile dinamica di fatto porterebbe a un impatto asimmetrico dei tassi di interesse sulla compagnia assicurativa, che ne risente molto di più quando i tassi aumentano perché c'è un impatto aggiuntivo dato dall'aumento dei riscatti.

Gli autori in questo caso propongono due diverse metodologie. In un primo momento adottano un metodo di Classification And Regression Tree (CART). Nella presente trattazione non illustreremo questo modello, ma ci limiteremo a riportare le considerazioni degli autori in merito alla seconda metodologia utilizzata, ovvero il modello logistico, che viene utilizzato sia in un'analisi statica, sia in un'analisi dinamica. Partendo da un dataset di portafoglio delle polizze vita di una nota compagnia spagnola, gli autori cominciano illustrando l'analisi statica. La staticità risiede nell'impostazione dello studio, che prende in esame una fotografia di portafoglio scattata nel 2007, a partire da un periodo di osservazione che va dal 1999 al 2007. Questo significa che nel 2007 vengono osservati i valori che le variabili assumono nel 2007 per le polizze ancora in portafoglio, oppure al momento del riscatto se le polizze sono state riscattate prima del 2007. Questo approccio pone diversi problemi, uno su tutti l'incapacità di catturare l'effetto del new business, oltre

all'impossibilità di utilizzare la durata in portafoglio della polizza come fattore di spiegazione. Riconoscendo i limiti di un approccio statico, gli autori passano quindi a una versione dinamica del modello logistico. Con approccio dinamico si intende un approccio in cui l'unità statistica è data dalla singola polizza osservata in un singolo periodo, costruendo così un dataset in cui una polizza si ripete tante volte quanti sono i periodi in cui rimane nel portafoglio. Un'importante assunzione alla base di questo approccio è l'indipendenza temporale: la scelta, anche dello stesso policyholder, di riscattare in t+1 è indipendente dalla scelta di riscattare in t. L'ipotesi, come sottolineano gli autori, è sicuramente poco verosimile (e ancor meno in periodi di crisi economica prolungata) ma l'ampiezza della base dati ci permette di forzare questa ipotesi se si considera che nell'aggregato ci sono molte polizze che non riscattano.

Tra le variabili esplicative utilizzate (usate anche nell'analisi statica), quelle che sono risultate più significative sono:

- l'età: una maggiore età è stata associata a una maggiore propensione al riscatto,
- periodicità del premio: le polizze a premio annuale o a doppia rata mensile sono state riscattate di più,
- ricchezza: i ceti meno abbienti e quelli più ricchi riscattano di meno. Per i primi, questi potrebbero non avere le risorse per pagare le penali di riscatto, per i secondi invece le condizioni economiche di mercato non destano preoccupazioni e/o non stimolano la ricerca di rendimenti migliori.
- durata: il fattore più significativo, soprattutto in corrispondenza del termine del periodo in cui vigono restrizioni fiscali o legali, oppure delle penali contrattuali sul riscatto.

Il fattore genere non è risultato significativo.

2.3 Xong e Kang

Vediamo ora un interessante lavoro degli autori Xong and Kang [2019], che si focalizza sulla previsione dell'evento binario di riscatto mediante diversi modelli di classificazione binaria. Al centro di quest'analisi vi è una comparazione in termini di capacità previsiva di vari algoritmi, più precisamente la regressione logistica, K-Nearest Neighbour, Neural Network, Support Vector Machine. Vengono inizialmente elencati diversi studi scientifici inerenti problemi di classificazione binaria di

variabili economico-sociali (soprattutto sul default) e si riscontrano pro e contro tra i vari modelli, rendendo più difficile la scelta.

2.3.1 Case study

Il dataset preso in esame comprende un portafoglio di un'assicurazione con sede in Malesia, le cui polizze sono in mano a policyholder di età dai 21 ai 65 anni e per le quali sono disponibili le osservazioni per le seguenti variabili:

- status della polizza (se "lapsed" o "in force", al 30 giugno 2017)
- premium frequency (periodicità di pagamento del premio)
- entry age (età all'emissione)
- policy term (periodo di copertura, se a vita intera o temporanea)
- somma assicurata
- genere

I vari algoritmi di classificazione vengono confrontati in termini di capacità previsiva, misurata sia mediante l'indicatore di accuracy, sia mediante la Receiver Operating Characteristic (ROC) Curve e l'Area Under the Curve (AUC) relativa. Di seguito vengono presentate brevemente tali metriche, per una trattazione più esaustiva si rimanda al capitolo 4 di questo lavoro, dedicato alle Metriche di performance previsiva.

L'accuracy fa riferimento al rapporto di previsioni corrette sulla numerosità del dataset di testing del modello. Si possono sollevare delle obiezioni su questo metodo di valutazione poiché non tiene conto della funzione di costo di un'assicurazione, che sarà verosimilmente asimmetrica: penalizzerà molto più i falsi negativi (chi riscatta la polizza senza che la compagnia l'avesse previsto) rispetto ai falsi positivi (chi non riscatta la polizza nonostante la compagnia avesse previsto il riscatto).

La curva ROC traccia invece la relazione tra True Positive Rate e False Positive Rate in corrispondenza di diversi livelli possibili di soglia di classificazione (che varia comunque tra 0 e 1). Un classificatore casuale otterrebbe una curva ROC di tipo lineare con coefficiente angolare unitario. Proprio per questo motivo l'AUC, ovvero l'area sottesa alla curva ROC fino alla retta bisettrice del quadrante, rappresenta una misura di capacità di classificazione. Più è estesa l'AUC, maggiore sarà la capacità predittiva.

2.3.2 Risultati

La Neural Network risulta il modello con la più alta accuracy misurata nel dataset di test. La Support Vector Machine è però la metodologia che si è rilevata più precisa in termini di Area Under the Curve (AUC). Gli autori concludono che le Neural Networks e la Support Vector Machine sono gli algoritmi che hanno dimostrato la maggiore capacità previsiva.

2.4 Barucci et al.

Il paper di Barucci et al. [2020] si propone di investigare le determinanti del riscatto delle polizze vita di una grande compagnia italiana di assicurazioni. Questo proposito viene perseguito con due diverse metodologie. In un primo momento viene adottata una metodologia di studio sulla base di dati dei singoli contratti, di seguito chiamati "microdata". Il database include i dati delle polizze dal 2008 al 2017, includendo quindi il periodo della crisi dell'Eurozona. Le polizze in esame sono per il 60 per cento circa di tipo tradizionale (di Ramo I e quindi con meccanismi di rivalutazione e garanzie minime di rendimento che tutelano il policyholder) e per la restante parte sono contratti a forte contenuto finanziario (unit linked, che non hanno garanzie di rendimento), su cui però si concentra solo il 25 per cento del capitale assicurato. Una prima analisi descrittiva del dataset rivela che i tassi di riscatto sono più alti e più volatili per le polizze unit rispetto a quelle tradizionali, come è ragionevole aspettarsi dal momento che il rischio di mercato rimane interamente in capo al contraente solo per le polizze unit. Le variabili rilevate per ogni singolo contratto sono:

- genere
- età all'ingresso
- regione di residenza
- professione
- tipo di polizza (traditional oppure unit linked)
- data stipula
- capitale assicurato
- età della polizza
- periodicità del premio

- rendimento del fondo

Gli autori scelgono di considerare solo i riscatti totali (escludendo quindi le anticipazioni e i riscatti parziali) come lapse. Viene impiegato, oltre a un modello basato sulla survival analysis, un modello GLM di tipo logistico, con distribuzione binomiale per la variabile binaria del riscatto. Le variabili esplicative vengono discretizzate utilizzando dei boxplot dei lapse rates suddivisi per le varie variabili continue.

2.4.1 Risultati

Utilizzando un modello GLM Poisson, un modello GLM Binomiale e un modello basato sul Proportional Hazard ratio (PH), gli autori ottengono delle stime dei parametri che mantengono lo stesso segno su tutti e tre i modelli. Inoltre, con poche eccezioni quasi tutte le stime dei parametri sono risultate significative al livello 1 per cento.

Gli autori traggono le seguenti conclusioni:

- Product type: il lapse rate delle polizze unit è più alto di circa il 300 per cento rispetto a quanto avviene per le tradizionali.
- Gender: in linea generale le donne tendono a riscattare più spesso degli uomini.
- Age: le fasce più giovani e quelle più anziane sono quelle con i maggiori tassi di riscatto. Questa evidenza sembra rinforzare l'ipotesi Emergency Fund Hypothesis, ovvero che le polizze vengano riscattate da chi ha bisogno di liquidità.
- Calendar Year: i modelli indicano un effetto calendario che vede un trend in aumento dei riscatti dal 2008 al 2017, seppur con un'inversione di tendenza rilevata solo negli anni 2013 e 2014. Tutto ciò sembra avvalorare la tesi per cui il tasso di riscatti sia sensibile alle turbolenze di mercato.
- Region: il nord-ovest presenta i più alti tassi di riscatto, ed essendo la regione più ricca d'Italia questa evidenza sembra andare in contrasto con la Emergency Fund Hypothesis.
- Premium Frequency: lapse rate più alti in corrispondenza di contratti a premio unico rispetto a quelli a premio ricorrente.
- Capital: lapse rate più alti in corrispondenza di contratti con una somma assicurata di minore entità. A conferma della Emergency Fund Hypothesis, i policyholder con maggiori somme assicurate (e quindi verosimilmente quelli

più ricchi) sono quelli che avvertono meno il bisogno di liquidità e quindi riscattano di meno.

- Anti-duration: il più alto tasso di riscatti si riscontra nella classe di antidurata dagli anni 4 a 7, all'aumentare dell'età della polizza diminuiscono i tassi di riscatto.

Gli autori decidono poi di suddividere il dataset tra polizze tradizionali e polizze unit linked.

Per le polizze tradizionali i tassi di riscatto tendono ad avere dei picchi nelle età più giovani e in quelle più anziane (a sostegno della EFH), mentre per le polizze unit linked i tassi di riscatto aumentano con le età.

Per le polizze tradizionali i tassi di riscatto sono più alti al Sud e nelle isole, mentre per le polizze unit linked i tassi di riscatto sono più alti al Nord.

Per le polizze tradizionali i tassi di riscatto aumentano dal 2008 al 2017 ma la loro sensibilità all'anno di calendario sembra debole, mentre per le polizze unit linked i tassi di riscatto sono molto più sensibili all'anno di calendario.

Mentre nell'intero campione (tradizionali e unit linked) erano i contratti a premio unico ad essere i più interessati dai riscatti, separando i campioni il coefficiente associato cambia segno in entrambi i sottocampioni, ovvero i riscatti si concentrano maggiormente nei contratti con premio ricorrente sia per le polizze unit, sia per le tradizionali.

Per le polizze tradizionali i riscatti sono associati a livelli bassi di insured capital; viceversa per le polizze unit si riscontra un'associazione dei riscatti con i livelli più alti di insured capital.

Si notano discrepanze anche rispetto all'effetto dell'antidurata (ovvero dell'età della polizza): i riscatti sono associati alle polizze più vecchie per quanto riguarda le unit e alle polizze più recenti per quanto riguarda le tradizionali.

Restringendo poi il campione solo ai contraenti di cui si conosceva lo status lavorativo, gli autori hanno riscontrato i livelli più bassi di riscatti nella categoria degli "employed". Questo risultato va ad avvalorare ulteriormente il filone teorico della Emergency Fund Hypotesis.

Lo studio riporta anche un'analisi a livello regionale; tale studio non verrà illustrato in questa sede poiché si basa su modelli di tipo panel che sono al di fuori del perimetro di analisi della presente tesi.

2.5 Azzone et al.

Il paper di Azzone et al. [2022] riprende in parte l'analisi di Barucci et al. [2020], confermando la conclusione secondo cui i modelli di machine learning (quali ad esempio Random Forest) siano in grado di catturare l'eterogeneità delle decisioni di riscatto in misura maggiore rispetto a quanto avviene per i modelli lineari di tipo GLM. Anche il dataset utilizzato ha caratteristiche analoghe a quello utilizzato nel paper di Barucci et al. [2020]. In aggiunta al suddetto dataset vengono però considerate anche delle variabili macroeconomiche che si pensa possano influire sul lapse rate, ovvero:

- il disposable income, o meglio il tasso di crescita del reddito disponibile delle famiglie italiane,
- il tasso annuale di inflazione in Italia,
- il tasso di variazione annua dell'indice Eurostoxx,
- il tasso di interesse a 12 mesi di un BOT italiano (assimilabile a uno Zero Coupon Bond).

Anche in questo studio il dataset contiene molteplici osservazioni delle stesse polizze, ripetute finché queste rimangono all'interno del portafoglio.

Su questo dataset viene svolta un'analisi sia mediante un modello di regressione logistica, sia mediante un algoritmo di Random Forest. Le performance dei due modelli vengono comparate considerando come metriche di valutazione l'accuracy, l'Area Under the Curve (AUC) e l'Area Under Precision Recall Curve (AUPR). Quest'ultima metrica si riferisce all'area sottesa alla curva che si ottiene in un grafico, in cui negli assi troviamo vari valori di Precision e di Recall associati a diversi possibili valori della soglia. Precision e Recall sono definiti come

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN},$$

dove TP sta per True Positive (numero di osservazioni effettivamente positive e correttamente classificate come tali), FP indica i False Positive (numero di osservazioni effettivamente negative ma erroneamente classificate come positive), FN indica i False Negative (numero di osservazioni effettivamente positive ma erroneamente classificate come negative). Un'ulteriore misura utilizzata dagli autori per comparare le performance previsive di un modello logistico è la cosiddetta LogLoss. Tale metrica si basa sulla cross-entropy, non ha un lower bound ed è utilizzata solamente per confronti tra modelli (la sua valenza va dunque intesa in termini relativi).

Ne riportiamo la formula analitica e dimostriamo che è strettamente collegata alla verosimiglianza calcolata nel suo punto di massimo, ovvero quando i parametri assumono il valore delle stime di massima verosimiglianza.

$$LogLoss = -\sum_{i=1}^{n} y_i log(p_i) + (1-y_i) log(1-p_i) = -\sum_{i=1}^{n} log(p_i^{y_i}) + log((1-p_i)^{1-y_i})$$
$$= -\sum_{i=1}^{n} log(p_i^{y_i}(1-p_i)^{1-y_i}) = -log\Big(\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\Big)$$

Si noti che in questo caso le p_i sono da considerarsi come stime di massima verosimiglianza, ovvero con

$$\hat{p}_i = g^{-1}(x_i^T \hat{\beta})$$

secondo la notazione tipica dei modelli di regressione logistica che verranno illustrati nel capitolo 3 dedicato ai Generalized Linear Models. Gli autori non si limitano a valutare le performance previsive, ma si concentrano anche sulla "explainability" (interpretabilità) dei vari modelli. Il modello logistico da questo punto di vista risulta molto più facilmente interpretabile.

2.5.1 Risultati

Gli autori riscontrano i seguenti driver principali dei riscatti:

- tempo trascorso dalla stipula (durata della polizza),
- tempo residuo alla scadenza,
- contract size,
- periodicità del premio,
- tipo di campagna pubblicitaria delle polizze.

Diversamente dallo studio di Barucci et al. [2020], le variabili genere, età e regione di residenza non sono risultate particolarmente influenti sui riscatti.

2.6 Eling, Kiesenbauer

Nel paper di Eling and Kiesenbauer [2013] gli autori studiano l'impatto che hanno le caratteristiche di prodotto e di policyholder nel determinare eventi di riscatto. Anche in questo studio viene sottolineata l'importanza cruciale di prevedere il fenomeno dei riscatti, poiché quest'ultimo porta con sé molteplici rischi per la compagnia,

anche di varia natura: rischio economico (perdite economiche dovute al mancato ricavo dalle polizze), rischio di selezione avversa rispetto a mortalità e morbidità, rischio di liquidità e/o di realizzazione di minusvalenze latenti. Gli autori segnalano che storicamente non è stato sempre considerato il diritto di recesso o riscatto nella determinazione dei premi di una polizza vita (o nel pricing della polizza in generale). La più recente letteratura ha poi individuato in tale diritto di recesso/riscatto un'opzione implicita nel contratto assicurativo e ha visto nascere degli interessanti spunti di applicazione della teoria dell'option pricing al mondo assicurativo, più precisamente nell'ambito dei contratti con Guaranteed Minimum Withdrawal Benefit (GMWB).

Gli autori sottolineano anche il fatto che il dataset a loro disposizione è quello più ricco di cui si sia a conoscenza nel filone di letteratura nel campo dei modelli GLM per il lapse risk. Il dataset in questione si riferisce ai dati di portafoglio di una grande compagnia tedesca di assicurazioni nel ramo vita. I dati sono riferiti al periodo dal 2000 al 2010, comprendendo quindi anche una fase molto interessante ovvero quella tra la crisi economica del 2001 e quella del 2008. I dati contengono le osservazioni di variabili quali tipologia di prodotto, anno di calendario, antidurata (o età della polizza), età del policyholder, canale di distribuzione, eventuale presenza di coperture aggiuntive (quali ad esempio coperture sulla vita con scadenza, coperture su disabilità e incidenti, ecc...), sesso, periodicità del premio (unico o periodico). Gli autori aggregano i contratti in dei cluster, denominati "model point", sui quali calcolano i tassi di riscatto in termini di percentuali di somme assicurate che vengono riscattate. In ciascun cluster viene anche definita l'esposizione, che corrisponde alla somma del tempo che i contratti hanno trascorso appartenendo alla cella.

La metodologia adottata dagli autori si basa sui modelli GLM e più in particolare con distribuzioni di tipo Poisson e Binomiale.

Si riassumono i risultati ottenuti dagli autori sulla significatività delle variabili:

- Product type: questa variabile sembra non influire sui tassi di riscatto. Gli autori segnalano che la variabile è stata considerata anche in altri studi, ma anche che il suo significato può differire tra i diversi studi per via della specificità dei contesti nazionali.

- Calendar year: c'è un significativo effetto dell'anno di calendario, con tassi di riscatto che aumentano in corrispondenza delle crisi economiche (in accordo con la teoria dell'Emergency Fund Hypothesis).

- Antidurata (età della polizza): i tassi di riscatto sembrano essere maggiori nei primi anni della polizza, per poi decrescere stabilmente. Gli autori spiegano questa dinamica supponendo che i riscatti nei primi anni di polizza siano dovuti

al riscontro immediato che il policyholder ha in merito alla sua capacità di sostenere i pagamenti della polizza.

- Età del policyholder: secondo gli autori questo fattore non sembra influire sui tassi di riscatto in maniera importante. I più giovani (fino a 25 anni) presentano i tassi più bassi, poi in aumento dai 25 anni in su, probabilmente in corrispondenza di nuovi bisogni di liquidità dei giovani contraenti. Per le fasce più anziane i tassi aumentano fino ai 60 anni per poi scendere, presumibilmente perché ci sono porzioni di popolazione in età adulta ma non ancora pensionabile che si trovano in emergenza di liquidità.

- Distribution Channel: rispetto ai tassi di riscatto di polizze vendute da agenti, i tassi di riscatto di polizze vendute dalla banca o da broker sono rispettivamente maggiori del 25 per cento e minori del 5 per cento. Per quanto riguarda la banca, una possibile spiegazione risiede nel fatto che quest'ultima potrebbe fornire servizi più standardizzati e quindi avere una minore retention di clienti.

- Copertura accessoria: sorprendentemente gli autori osservano un aumento dei tassi di riscatto per le polizze che hanno coperture aggiuntive (quali ad esempio contro l'invalidità). Si spiegano questa evidenza come segue: il premio è più alto (quindi più oneroso sulla liquidità) in presenza di coperture aggiuntive e/o tali coperture vengono più facilmente vendute a dei contraenti privi di un'adeguata conoscenza dei prodotti assicurativi, che scoprono solo dopo la stipula di non desiderare tali coperture.

- Genere: gli autori riscontrano dei tassi di riscatto minori del 9 per cento per le contraenti femminili, in accordo con la letteratura che tendenzialmente riscontra una maggiore avversione al rischio delle donne.

- Frazionamento del premio: il tasso di riscatto di polizze a premio unico è minore del 90 per cento rispetto alle polizze con premio frazionato.

Capitolo 3

Generalized Linear Models

In questo capitolo verranno illustrati i punti principali della teoria dei modelli GLM. Per illustrare i concetti principali di tale teoria si fa riferimento ai testi di Ohlsson and Johansson [2010], McCullagh and Nelder [1989] e Dobson [2001]. La trattazione avrà un taglio incentrato su metodi e modelli di più comune utilizzo nella tariffazione del ramo danni. Come vedremo, questo approccio può essere facilmente esteso all'ambito della modellazione statistica dei riscatti nel ramo vita.

3.1 Framework

3.1.1 Ipotesi alla base

In questa sezione verranno illustrate e descritte le ipotesi alla base del framework che verrà utilizzato in questa analisi. Queste ipotesi derivano dall'impostazione utilizzata nel testo di Ohlsson and Johansson [2010].

H1: Indipendenza tra polizze

La prima ipotesi riguarda appunto la struttura di dipendenza stocastica delle variabili di interesse. In questo caso l'ipotesi di indipendenza si riferisce alla concezione più restrittiva di indipendenza, ovvero di indipendenza a due a due (e quindi poi anche congiunta). Supponendo quindi di disporre di n osservazioni di polizze (di cui ad esempio prendiamo la i-esima e la j-esima), **H1** equivale a richiedere che:

$$X_i \text{ e } X_j \text{ sono indipendenti } \forall i \neq j, \quad i, j = 1, ..., n$$

Nel nostro caso specifico, assumere indipendenza della variabile di interesse tra polizze si traduce nel richiedere che la decisione di riscatto di una polizza sia indipendente dalla decisione di riscatto delle altre polizze. Risulta evidente come questa ipotesi possa rivelarsi abbastanza distante dalla realtà dei fondi pensione: si pensi a titolo

di esempio ad un intervento normativo e/o al lancio di un nuovo prodotto concorrente, circostanze in cui si potrebbe verificare un aumento di riscatti che coinvolga l'intero portafoglio clienti.

H2: Indipendenza temporale
La seconda ipotesi riguarda l'indipendenza della variabile di interesse se questa viene osservata (anche per la stessa polizza) su intervalli di tempo disgiunti, ovvero

$$\begin{gathered}\text{dati intervalli di tempo disgiunti } I_1, \ldots, I_n, \\ \text{le variabili } X_1, \ldots, X_n \text{ sono indipendenti.}\end{gathered}$$

Nel caso dei riscatti stiamo assumendo che la decisione di riscatto presa in un trimestre non influisca sulla decisione di riscatto presa nei trimestri successivi. Si tratta ovviamente di un'assunzione surreale se si ragiona sulla singola polizza; occorre però considerare che un portafoglio molto ampio di polizze può attenuare questa problematica. In tal senso, infatti, si può pensare che la decisione di riscatto di una polizza A in un trimestre sia indipendente dalla decisione di riscatto di una polizza B durante il trimestre successivo. Ci troviamo anche qui nella situazione in cui dei fattori esterni macroeconomici potrebbero rendere improbabili le assunzioni proposte, si consideri ad esempio il caso di una tendenza positiva o negativa dei tassi di rendimento di titoli di stato o comunque di forme di investimento alternative.

H3: Omogeneità tra polizze della stessa cella tariffaria
Questa ipotesi implica che all'interno di una stessa cella tariffaria le variabili di interesse abbiano identica distribuzione. In formule:

$$\begin{gathered}\text{se } X_i \text{ e } X_j \text{ appartengono alla stessa cella,} \\ \text{allora } X_i \text{ e } X_j \text{ sono identicamente distribuite.}\end{gathered}$$

L'idea alla base è quella di creare delle celle che siano sufficientemente omogenee al loro interno, minimizzando la varianza entro i gruppi per massimizzare la varianza tra gruppi.
Si può notare che l'ipotesi **H1** implica anche l'indipendenza tra polizze appartenenti a una stessa cella tariffaria. L'ipotesi **H3** richiede l'identica distribuzione delle celle tariffarie al loro interno (le polizze che appartengono alla cella tariffaria hanno identica distribuzione). In sintesi, le ipotesi appena viste implicano che all'interno della stessa cella tariffaria le variabili di interesse siano tutte IID.

3.1.2 Esposizione e Key ratios

L'approccio seguito ricalca i metodi di tariffazione più comunemente utilizzati nel ramo danni. Abbiamo una variabile di interesse X (nel nostro caso l'esito binario della decisione di riscatto), osservata con riferimento a un'esposizione ω (nel

nostro caso ogni polizza ha la stessa esposizione unitaria, trattandosi di un unico contraente).

Nel nostro campione di n osservazioni avremo quindi le osservazioni $X_1, \dots X_n$ e le esposizioni $\omega_1, \dots, \omega_n, \quad$ con $\omega_i = 1, \forall i = 1, ..., n$.

Consideriamo ora n variabili di interesse che appartengano alla stessa cella tariffaria e studiamo le proprietà dello stimatore media campionaria, che chiameremo Y:

$$Y = \frac{\sum_{i=1}^{n} X_i}{n}.$$

H3 ci permette di assumere che all'interno della stessa cella tariffaria le variabili siano identicamente distribuite, questo implica che

- $\mathbb{E}[X_i] = \mu$ costante $\forall i$ appartenente alla cella
- $Var[X_i] = \sigma^2$ costante $\forall i$ appartenente alla cella

In una notazione più compatta indicheremo che una variabile aleatoria ha un certo valore atteso e una certa varianza nel seguente modo:

$$X_i \sim (\omega_i \mu,\ \omega_i \sigma^2),$$

ricordando che nel nostro caso ogni ω_i vale 1 perché si riferisce al fatto che ci sia un unico contraente per ogni polizza. Risulta immediato assumere una distribuzione bernoulliana per il fenomeno dei riscatti (per ciascun X_i), per cui nel nostro caso avremo

$$\mu = p \quad e \quad \sigma^2 = p(1-p).$$

A questo punto possiamo occuparci della distribuzione della statistica media campionaria. Di seguito si riportano i passaggi per ottenere media e varianza dello stimatore media campionaria.

$$\mathbb{E}[Y] = \frac{\sum_{i=1}^{n} \mathbb{E}[X_i]}{n} = \frac{\sum_{i=1}^{n} p}{n} = \frac{np}{n} = p$$

$$\begin{aligned} Var[Y] &= \frac{\sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov[X_i, X_j]}{n^2} \\ &= \frac{\sum_{i=1}^{n} Var[X_i]}{n^2} = \frac{p(1-p)n}{n^2} = \frac{p(1-p)}{n} \end{aligned}$$

Abbiamo utilizzato le ipotesi **H1** e **H2** per imporre che le covarianze siano nulle. In linea generale otteniamo uno stimatore $Y \sim (\mu; \sigma^2/\omega)$ ovvero uno stimatore non distorto e consistente. Nel nostro caso (variabili bernoulliane indipendenti e $\omega_i = 1$)

abbiamo anche che

$$nY \sim (np,\ np(1-p)) \quad e \quad nY \sim Bin,$$

ovvero che

$$Y \sim (p,\ p(1-p)/n) \quad e \quad Y \sim RelativeBinomial.$$

Quella che abbiamo chiamato media campionaria è anche interpretabile di fatto come un Key ratio, ovvero il rapporto tra la somma delle variabili di interesse e la somma delle esposizioni (infatti $\sum_{i=1}^{n} \omega_i = \sum_{i=1}^{n} 1 = n$), il tutto riferito alle sole osservazioni afferenti alla stessa cella tariffaria. Nel pricing del ramo danni abbiamo spesso dei Key ratios di interesse quali Frequency e Severity. In questo lavoro il Key ratio di interesse è la proporzione di polizze (all'interno della stessa cella) che vengono riscattate. In accordo con la teoria frequentista della probabilità, queste proporzioni rappresentano una stima delle probabilità.
Indicizzando le celle tariffarie in $j = 1, \ldots, M$, questo Key ratio Y_j rilevato nella cella j-esima rappresenta lo stimatore ottimale della probabilità di riscatto p_j riferita alla cella j-esima. Implicitamente stiamo ipotizzando che in ogni cella esista una probabilità di riscatto p_j che sia costante all'interno della cella e diversa da quella delle altre celle, cioè che $p_j \neq p_k \quad \forall j \neq k$. Il nostro obiettivo sarà quindi modellare tali p_j per spiegare e capire quali fattori tariffari (e i loro livelli) risultano più significativi nella decisione di riscatto. La scelta della dimensione di M, cioè quante celle costruire, presenta un trade-off che rende difficile la scelta del livello di aggregazione in celle tariffarie. Avere poche celle con molta esposizione al loro interno permette di ridurre la varianza del Key ratio (dunque della stima della p_j), ma potrebbe comportare un allontanamento dall'ipotesi **H3** di omogeneità dentro la cella. Avere invece molte celle comporterebbe un aumento della varianza del Key ratio, ma di contro permetterebbe di raggiungere una maggiore omogeneità all'interno della cella.

3.2 Modelli additivi e moltiplicativi

La modellazione delle varie p_j si presta a molteplici possibili approcci, che differiscono anche e soprattutto per numero di parametri e complessità del modello. Anche in questo caso faremo riferimento all'approccio più comunemente utilizzato nel Non life pricing con modelli GLM. Una considerazione preliminare va riservata alla dimensione di M. Il numero di celle, infatti, risulta pari al prodotto del numero di livelli di tutti i fattori tariffari. A titolo esemplificativo, se si scelgono come fattori tariffari l'area geografica (classificata in 5 macroregioni) e la fascia di età (supponiamo di individuare 4 fasce d'età), in assenza di altri fattori tariffari otterremo $M = 20 = 5 * 4$

celle. È immediato quindi rilevare come M possa crescere anche molto rapidamente all'aumentare del numero di fattori e di livelli tariffari. Focalizziamoci ora sul modello con più parametri possibili, ovvero un modello in cui ciascuna p_j viene stimata con un suo parametro. Avremmo in questo caso un modello con tanti parametri da stimare quante sono le celle tariffarie (M). Un simile modello presenterebbe evidenti problemi di "overfitting" per via di un'eccessiva aderenza ai dati. Proprio per questo motivo nella pratica non si modella ciascuna p_j singolarmente, ma si ricorre a dei modelli additivi o moltiplicativi, che associano un parametro a ciascun livello di ciascun fattore tariffario. Nell'esempio di cui sopra, si avrebbero $9 = 5 + 4$ parametri. Prima di illustrare i modelli additivo e moltiplicativo, introduciamo una nuova notazione più agevole. Finora abbiamo indicizzato le p_j secondo il contatore j che di fatto elenca tutte le celle possibili. Questi M parametri verranno indicati in questa sezione con la notazione $p_{h,k}$ per sottolineare che una determinata probabilità di riscatto si riferisce alla cella individuata dal livello h per il primo fattore tariffario e dal livello k per il secondo. Per una maggiore agevolezza di esposizione, i modelli additivo e moltiplicativo verranno presentati su due fattori tariffari ma si possono generalizzare a un numero arbitrario di fattori tariffari, semplicemente aggiungendo altrettanti indici a pedice.

Il modello additivo scompone il parametro di interesse di una cella in una costante (una sorta di valore base) alla quale si aggiungono dei coefficienti associati al livello tariffario che quella cella presenta per ciascun fattore tariffario. Il valore dei coefficienti indica di quanto aumenta o diminuisce il parametro di interesse se si passa dalla cella base di riferimento ad un'altra cella che presenta un diverso livello tariffario. In formule:

$$\mu_{h,k} = \beta_0 + \beta_h + \beta_k$$

Va considerato però che è possibile ottenere gli stessi valori di $\mu_{h,k}$ apportando delle piccole modifiche a questa specificazione del modello. In termini più tecnici si può dire che un modello così specificato presenta dei problemi di mancata identificazione, ovvero che esistono set di parametri diversi che possono però restituire gli stessi risultati a partire dagli stessi dati. A titolo di esempio si pensi ad un modello di questo tipo:

$$\tilde{\mu}_{h,k} = \tilde{\beta}_0 + \tilde{\beta}_h + \beta_k$$

con $\tilde{\beta}_0 = \beta_0 + c$ e con $\tilde{\beta}_h = \beta_h - c,\ \forall h$ (h generico livello tariffario del primo fattore). Si riscontra immediatamente che i due distinti set di parametri β e $\tilde{\beta}$ generano le stesse stime:

$$\tilde{\mu}_{h,k} = \tilde{\beta}_0 + \tilde{\beta}_h + \beta_k = \beta_0 + \beta_h + \beta_k = \mu_{h,k}$$

Per ovviare a questo problema si impone solitamente una restrizione su un sottoin-

sieme dei parametri in β. Questo sottoinsieme del set dei parametri deve contenere uno e un solo coefficiente per ciascun fattore tariffario (a titolo esemplificativo, il coefficiente relativo alla prima fascia di età, alla prima zona geografica, e così via per i vari altri fattori eventualmente inclusi nel modello). La restrizione in questione richiede che siano tutti nulli i valori di tale sottoinsieme del set di parametri. Questo equivale di fatto a ipotizzare un "livello base" del valore $\mu_{h,k}$ che si riscontra in una determinata cella. Di qui in avanti questa cella verrà denominata come "cella di riferimento", e ogni spostamento da questa cella comporta una variazione di $\mu_{h,k}$ dal suo livello di riferimento. La variazione è data dal coefficiente stesso β_h e/o β_k. Si supponga infatti di scegliere $\mu_{1,1}$ come valore di riferimento (ovvero la cella di riferimento presenta il livello 1 per il primo fattore e il livello 1 anche per il secondo fattore). A questo punto, imponendo le restrizioni menzionate, cioè

$$\beta_{h=1} = 0 \; e \; \beta_{k=1} = 0$$

avremo che

$$\mu_{1,1} = \beta_0 + \beta_{h=1} + \beta_{k=1} = \beta_0$$

e quindi anche che:

$$\mu_{h,k} - \mu_{1,1} = \beta_0 + \beta_h + \beta_k - \beta_0 = \beta_h + \beta_k$$

per generici livelli h del primo fattore e k del secondo fattore.

Un'alternativa al modello additivo si ritrova nel modello moltiplicativo, che scompone il parametro di interesse di una cella in una costante (una sorta di valore base) al quale vengono moltiplicati dei coefficienti (le cosiddette relatività) associati al livello che quella cella presenta per ciascun fattore tariffario. Il valore dei coefficienti indica per quale valore va moltiplicato il parametro di interesse se si passa dalla cella base di riferimento ad un'altra cella che presenta un diverso livello tariffario. Riportiamo in formule il modello moltiplicativo:

$$\mu_{h,k} = \gamma_0 * \gamma_h * \gamma_k$$

Associando un coefficiente ad ogni livello (ma non ad ogni cella), il modello moltiplicativo (così come quello additivo) non riesce a catturare bene le interazioni tra variabili, poiché in corrispondenza di ogni livello tariffario applica un coefficiente di scala in maniera "trasversale" su tutti i livelli degli altri fattori. Il testo Ohlsson and Johansson [2010] riporta come esempio, nell'ambito del motor insurance pricing, il fatto che tra i giovani guidatori ci sia una forte prevalenza di incidenti per gli uomini (interazione tra ETA=GIOVANE e GENERE=M), mentre non ci sono distinzioni

significative di genere quando si guarda al numero di incidenti provocati da persone di mezza età.

Applicando al modello moltiplicativo la funzione logaritmo, otteniamo un modello additivo per il logaritmo del parametro.

$$\log(\mu_{h,k}) = \log(\gamma_0) + \log(\gamma_h) + \log(\gamma_k) = \beta_0 + \beta_h + \beta_k$$

Questa trasformazione permette di constatare facilmente che, come per il modello additivo, anche in quello moltiplicativo ci sono dei problemi di identificazione. A titolo esemplificativo i set di parametri $(\gamma_0, \gamma_h, \gamma_k)$ e $(\gamma_0 c\ , \gamma_h/c\ , \gamma_k)$ forniscono la stessa stima di $\mu_{h,k}$. Anche in questo caso si può ovviare al problema scegliendo una cella di riferimento, uno spostamento da questa comporta una variazione determinata da un coefficiente di scala che è esattamente il coefficiente associato al livello non di riferimento. Ipotizzando che la cella di riferimento abbia il primo livello in entrambi i fattori:

$$\mu_{1,1} = \gamma_0$$

e

$$\mu_{h,k}/\mu_{1,1} = \gamma_0\gamma_h\gamma_k/\gamma_0 = \gamma_h\gamma_k$$

per generici livelli h del primo fattore e k del secondo fattore.

Si conclude segnalando che la trasformazione del modello moltiplicativo mediante il logaritmo è molto utilizzata nel pricing Non life perché si presta agevolmente alla scelta di una funzione link logaritmica nell'ambito di un modello GLM.

3.2.1 Funzione logit: modello moltiplicativo per gli odds

Come vedremo in seguito, la presente analisi utilizzerà di fatto una regressione logistica, ovvero un modello GLM in cui la funzione link sarà di tipo logit. In formule:

$$\log(\frac{\mu_{h,k}}{1-\mu_{h,k}}) = \beta_0 + \beta_h + \beta_k$$

Applicando la funzione esponenziale ad entrambi i lati possiamo notare come, nella regressione logistica che useremo, si ha di fatto un modello moltiplicativo per gli odds.

$$Odds(p_{h,k}) = Odds(\mu_{h,k}) = \frac{\mu_{h,k}}{1-\mu_{h,k}} = \exp(\beta_0 + \beta_h + \beta_k) = \gamma_0\gamma_h\gamma_k$$

Gli odds rappresentano il rapporto tra la probabilità di un evento e il suo complemento; pertanto, sono una trasformazione monotona crescente della probabilità dell'evento al numeratore, che viene mappata nel semiasse reale positivo. La proprietà di monotonia crescente ci permette di trattare gli odds alla stregua di una proxy per le probabilità di riscatto. Più in generale avremo degli odds che variano sulle diverse celle partendo da un livello di riferimento e modificandosi con un meccanismo di relatività moltiplicative. Ipotizzando che la cella di riferimento abbia il primo livello in entrambi i fattori, ovvero che:

$$\log(\frac{\mu_{1,1}}{1-\mu_{1,1}}) = \beta_0 + \beta_{h=1} + \beta_{k=1} = \beta_0 \iff \frac{\mu_{1,1}}{1-\mu_{1,1}} = \exp(\beta_0) = \gamma_0 = Odds(\mu_{1,1}).$$

Avremo che

$$OddsRatio(h,k) = \frac{Odds(\mu_{h,k})}{Odds(\mu_{1,1})} = \gamma_h\gamma_k \iff Odds(\mu_{h,k}) = Odds(\mu_{1,1})\gamma_h\gamma_k.$$

Notiamo quindi che valori maggiori di 1 di γ_h, γ_k (e quindi valori positivi di β_h, β_k) aumentano la probabilità di riscatto nella cella (h,k) rispetto alla probabilità di riscatto nella cella di riferimento (1,1).

Nel seguito si farà riferimento a distribuzioni e modelli che si prestano meglio a una rappresentazione in lista dei dati (un unico indice che corrisponde all'osservazione), mentre l'impostazione appena presentata (con dei coefficienti associati a vari livelli di rischio) si presta a una visione tabulare dei dati (dati raggruppati in celle). A livello di matrice dei dati, rappresentati in forma di lista, i modelli additivo e moltiplicativo corrispondono a impostare una matrice $\mathbf{X}$ composta da variabili dummy, associate ciascuna a un livello di ogni fattore tarrifario. Data un'osservazione i-esima e dato il livello tariffario j-esimo, il generico elemento della matrice $\mathbf{X}$, ovvero $x_{i,j}$, avrà valore 1 se quella osservazione presenta quel livello tariffario e avrà valore 0 altrimenti.

3.3 Exponential Dispersion Models (EDM)

3.3.1 Definizione ed esempi

Se nei modelli lineari classici (Ordinary Least Squares) la distribuzione di riferimento è la gaussiana, nei modelli lineari generalizzati la classe di distribuzioni di riferimento è appunto la classe delle distribuzioni EDM. Per una maggiore agevolezza nella trattazione, ritorniamo alla notazione delle p_i con singolo indice, un indice per ogni osservazione. Un aspetto centrale delle distribuzioni EDM è senz'altro la

loro funzione di probabilità, che assume una forma funzionale ben precisa, di seguito riportata.

$$f_{Y_i}(y_i, \theta_i, \omega_i, \phi) = \exp(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \omega_i, \phi)) \tag{3.1}$$

Dalla formula notiamo innanzitutto che il parametro di interesse θ_i varia appunto con l'osservazione. Richiediamo che lo spazio parametrico associato ai vari θ sia un intervallo aperto (ovvero che θ non possa assumere valori ai confini dello spazio parametrico). Questa restrizione trova fondamento in delle derivazioni che verranno descritte più avanti. Notiamo inoltre che il parametro di dispersione ($\phi > 0$ costante positiva) risulta costante su tutte le osservazioni. Sia θ sia ϕ sono parametri incogniti, ma il primo è l'oggetto della nostra indagine, mentre il secondo assume il ruolo di un parametro di disturbo. Abbiamo anche un parametro altrettanto importante che è noto, quello dell'esposizione, $\omega_i > 0$ (costante positiva). In riferimento alle singole osservazioni, per noi tutti gli ω_i saranno pari a 1; in riferimento alle celle tariffarie gli ω_j rispecchieranno quanti contraenti ricadono nella stessa cella tariffaria, risultando così variabili al variare delle celle. Nella funzione di distribuzione delle EDM rientrano anche le funzioni b( ) e c( ). La prima è detta funzione cumulante ed è quella che occorre specificare per ottenere univocamente un modello distributivo specifico all'interno della classe delle EDM. Eventuali trasformazioni lineari della funzione b( ) infatti non influiscono sul modello distributivo della variabile casuale appartenente alle EDM. La funzione cumulante ha come dominio lo spazio parametrico di θ e mappa i suoi possibili valori sull'asse dei numeri reali. Richiediamo che la funzione cumulante sia derivabile due volte rispetto a θ (quindi deve anche essere continua) e che abbia derivata prima invertibile (quindi tale funzione sarà strettamente monotona in θ). Anche in questo caso la ragion d'essere di queste restrizioni sarà discussa più avanti. Infine, la funzione c( ) rappresenta una costante che non dipende da θ: per questo motivo non entra in alcun modo nel processo di stima del parametro θ di massima verosimiglianza. Per avere conferma di questa affermazione è sufficiente osservare che nella log-verosimiglianza la funzione c( ) è soltanto una costante additiva, che si annulla nel momento in cui si deriva la log-verosimiglianza rispetto a θ.

Si riporta di seguito una tabella che riassume le varie parametrizzazioni in forma EDM di alcuni modelli distributivi tra i più conosciuti.

	θ	$b(\theta)$	ϕ	$\nu(\mu)$
$N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2	1
Poisson(λ)	$log(\lambda)$	$exp(\theta)$	1	λ
RelativeBinomial(n,μ)	logit(μ)	$log(1 + exp^{\theta})$	1	$\mu(1-\mu)$
Gamma($\omega\alpha, \omega\beta$)	$-\beta/\alpha$	-log(-θ)	$1/\alpha$	μ^2

Come già illustrato nella sezione delle ipotesi, nella nostra trattazione stiamo assumendo che le variabili osservate siano indipendenti.

3.3.2 Funzione generatrice dei cumulanti

La funzione generatrice dei momenti è una particolare funzione, associata a una variabile aleatoria, che coincide con il valore atteso dell'esponenziale di tale variabile aleatoria, moltiplicata per una variabile ausiliaria. In formule:

$$M_Y(t) = \mathbb{E}\left[e^{tY}\right].$$

Si può dimostrare che la funzione generatrice dei momenti (che chiameremo MGF da qui in avanti) restituisce il momento k-esimo della variabile aleatoria Y se si calcola la derivata k-esima della MGF in $t = 0$.

Analogamente la funzione generatrice dei cumulanti (CGF) è una funzione, associata a una variabile aleatoria, che ne restituisce il momento k-esimo centrato se si calcola la derivata k-esima della CGF in $t = 0$. Si può ricavare la CGF semplicemente applicando il logaritmo alla MGF.

$$\Psi_Y(t) = \log(M_Y(t)).$$

Nell'ambito delle distribuzioni EDM esiste un'interessante proprietà che lega la funzione dei cumulanti alla funzione generatrice dei cumulanti. D'ora in poi faremo riferimento a questa proprietà chiamandola semplicemente 'Lemma'. Riportiamo di seguito il lemma.

Lemma. Se $Y \sim EDM(\theta, b(), \phi, \omega)$, $\theta \in \Theta$, intervallo aperto, e $|t| \approx 0$, allora

$$\Psi_Y(t) = \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega}$$

Richiediamo che t sia in un intorno di 0 affinché $(\theta + t\phi/\omega)$ rientri nello spazio parametrico. Per lo stesso motivo abbiamo richiesto che lo spazio parametrico di θ fosse un intervallo aperto, altrimenti θ avrebbe potuto assumere valore all'estremo dell'intervallo e in quel caso $(\theta + t\phi/\omega)$ potrebbe non appartenere più all'intervallo. Per ricavare il primo cumulante (la media) di Y sarà opportuno calcolare la derivata

prima della funzione generatrice dei cumulanti in $t = 0$, cioé

$$\mathbb{E}[Y] = \left.\frac{d\Psi_Y(t)}{dt}\right|_{t=0}$$

Il rapporto incrementale ottenuto dal Lemma, che abbiamo visto coincidere con $\Psi_Y(t)$, una volta derivato in dt e ponendo $t = 0$, ci restituisce quindi il valore atteso della EDM, che coincide con la derivata prima di b( ) rispetto a theta. In formule:

$$\mathbb{E}[Y] = \left.\frac{d\Psi_Y(t)}{dt}\right|_{t=0} = \frac{d}{dt}\left(\frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega}\right)\Bigg|_{t=0} = b'(\theta).$$

Ricavando la derivata seconda della funzione generatrice dei cumulanti e calcolandola in 0 otteniamo poi la varianza di una EDM, che diventa la derivata seconda di $b()$ due volte rispetto a θ, scalata per ϕ/ω (dispersione fratto esposizione). In formule:

$$Var[Y] = \left.\frac{d^2\Psi_Y(t)}{dt^2}\right|_{t=0} = \frac{\phi}{\omega}b''(\theta).$$

I risultati appena ottenuti sono in linea con le conseguenze delle ipotesi assunte all'inizio della nostra analisi. Infatti notiamo che il valore atteso della EDM (del Key ratio, utilizzando la terminologia delle prime sezioni) non dipende da ω, mentre la varianza scala con $1/\omega$ (o equivalentemente, la precisione è direttamente proporzionale a ω). Nell'appendice si riportano la dimostrazione del Lemma (A.1) e un'illustrazione di due proprietà delle funzioni MGF e CGF A.2.

3.3.3 Riproducibilità

Una proprietà molto interessante delle distribuzioni EDM riguarda la possibilità di comprimere i dati di più distribuzioni. In altri termini, combinando linearmente delle distribuzioni EDM indipendenti è possibile ottenere nuovamente una distribuzione appartenente a questa famiglia. Questa proprietà risulta estremamente utile nella nostra analisi poiché ci permette di raggruppare i dati in celle tariffarie. Inizialmente calcoliamo il Key ratio per ciascuna singola osservazione, aggregando poi le osservazioni in celle andremo a calcolare (in ciascuna cella) una media, ponderata sulle esposizioni, del valore del Key ratio calcolato sulle singole osservazioni che afferiscono a quella cella. La proprietà di riproducibilità ci permette di affermare che questo Key ratio della cella è distribuito a sua volta secondo un modello EDM. A titolo esemplificativo supponiamo che solo le osservazioni $i = 1$ e $i = 2$ rientrino nella prima cella tariffaria (ma il ragionamento vale per qualsiasi numero di osservazioni).

Sappiamo che:

$$Y_1 \sim EDM(\theta, b(), \phi, \omega_1) \quad \text{e} \quad Y_2 \sim EDM(\theta, b(), \phi, \omega_2)$$

Si noti che non abbiamo indicizzato θ solamente perché le due osservazioni appartengono alla stessa cella e pertanto il loro valore di θ coincide. Infatti

$$\theta_i = \log(\frac{p_i}{1-p_i}) = x_i^T \beta,$$

e le osservazioni nella stessa cella hanno lo stesso vettore x_i^T.

A questo punto definiamo il Key ratio della cella con Y e l'esposizione totale della cella con ω.

$$Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2} = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega} = \frac{\omega_1}{\omega} Y_1 + \frac{\omega_2}{\omega} Y_2.$$

Nel nostro caso avremo X_1 e X_2 variabili bernoulliane con $\omega_1 = \omega_2 = 1$, dunque $Y_i = X_i/\omega_i = X_i$ con $i = 1, 2$. Avremo quindi che:

$$Y = \frac{1Y_1 + 1Y_2}{1+1} = \frac{X_1 + X_2}{2}$$

e più in generale

$$Y = \frac{\sum_{i=1}^{n} X_i}{n},$$

e otteniamo una media di variabili bernoulliane, che per definizione corrisponde a una Relative Binomial con $\omega = n$. Nell'appendice A.3 si riporta la dimostrazione del fatto che, se Y_1 e Y_2 sono indipendenti, vale che

$$Y \sim EDM(\theta, b(), \phi, \omega), \qquad \omega = \sum_{i \in cella} \omega_i.$$

3.3.4 Funzione varianza

Come si è già constatato, media e varianza di una EDM dipendono dalle derivate (rispettivamente prima e seconda) della funzione $b()$ rispetto a θ. Si può riscontrare una particolare ed interessante conseguenza di questa caratteristica esaminando la funzione varianza. A questo punto della trattazione possiamo dare almeno una prima giustificazione sulla restrizione che avevamo imposto richiedendo che la funzione $b()$ avesse derivata prima invertibile. L'invertibilità, infatti, permette di avere uno e un solo valore di θ come immagine (mediante la funzione inversa) di ogni singolo valore di μ. A questo punto calcolando la derivata seconda di $b()$ rispetto a θ si arriva alla varianza della EDM, a meno di una costante di proporzionalità (ϕ/ω). La funzione

composta che ci porta da μ alla varianza non scalata è detta appunto funzione varianza, e permette di capire una fondamentale relazione tra media e varianza. Nel caso della normale, la funzione varianza è una funzione costante unitaria: questo significa che traslare una normale modificandone la media non ha impatto sulla sua varianza, che rimane immutata. Diverso è il caso della Poisson, in cui la funzione varianza è la funzione identità: in questo caso quindi all'aumentare della media aumenterà anche la varianza. Riportiamo di seguito l'espressione analitica della funzione varianza.

$$\nu(\mu) = b''(b'^{-1}(\mu))$$

che discende da

$$Var[Y] = \frac{\phi}{\omega} b''(\theta) = \frac{\phi}{\omega} b''(b'^{-1}(\mu)) = \frac{\phi}{\omega}\nu(\mu)$$

La distribuzione di interesse per questo lavoro è la Relative Binomial, pertanto ci concentreremo solamente sulla sua funzione varianza. Abbiamo già visto che:

$$\begin{aligned} \mathbb{E}[Y] &= p = \mu \\ Var[Y] &= \frac{p(1-p)}{n} = \frac{\mu(1-\mu)}{n} \qquad \Longrightarrow \qquad Var[Y] = \frac{\phi}{\omega}\,\mu(1-\mu), \end{aligned}$$

poiché $\phi = 1$ e $\omega = n$. In questo caso risulta quindi triviale individuare la funzione varianza: $\nu(\mu) = \mu(1-\mu)$

Si può facilmente dimostrare che calcolare $\nu(\mu)$ come $b''(b'^{-1}(\mu))$ porterebbe allo stesso risultato. Si consideri inoltre che la funzione $\nu(\mu) = \mu(1-\mu)$ raggiunge il suo massimo in corrispondenza di $\mu = 1/2$. Questo costrutto matematico sembra anche ragionevole: in un esperimento binomiale si raggiunge la massima varianza quando i successi e i fallimenti sono equiprobabili.

3.3.5 La Binomiale e la Relative Binomial

In questa sezione esploriamo in maggior dettaglio le due distribuzioni che risultano fondamentali per la presente analisi.

La distribuzione binomiale descrive il numero di successi in un esperimento composto da n prove bernoulliane indipendenti e con identica distribuzione (dunque stessa probabilità di successo p).
La binomiale relativa (o Relative Binomial) corrisponde ad una binomiale scalata per $1/n$, ovvero alla percentuale di successi ottenuti su n prove bernoulliane indipendenti. Equivalentemente, la Relative Binomial è la distribuzione campionaria della media campionaria in un esperimento di bernoulliane IID. Questa prospettiva ci ri-

porta alle considerazioni già illustrate sui Key ratios. Le distribuzioni di bernoulli e le Relative Binomial appartengono entrambe alla famiglia delle distribuzioni EDM.

Nell'appendice A.4 si riporta la dimostrazione del fatto che la Relative Binomial appartenga alla famiglia EDM.

3.4 Generalized Linear Models (GLM)

3.4.1 Descrizione

Con la terminologia di Generalized Linear Models ci si riferisce a dei particolari modelli lineari che permettono di rilassare alcune ipotesi che vengono assunte nel quadro teorico dei modelli lineari classici. Viene rilassata innanzitutto un'ipotesi centrale nella teoria dei modelli lineari, ovvero la normalità della variabile di risposta. I modelli GLM consentono quindi di analizzare fenomeni non gaussiani, con variabili dipendenti che potrebbero essere ad esempio discrete, asimmetriche, a supporto limitato. Alcuni esempi di distribuzioni non gaussiane comunemente utilizzate nei modelli GLM sono la Poisson e la Gamma. In questa trattazione le distribuzioni di interesse sono quelle appartenenti alla famiglia delle EDM, ma va specificato che i modelli GLM consentono di adottare una qualsiasi distribuzione appartenente alla famiglia esponenziale.
Viene rilassata anche l'ipotesi OLS secondo cui la media della variabile dipendente può essere prevista come combinazione lineare delle variabili esplicative. In questo senso nei modelli GLM è possibile applicare una funzione link g( ) alla media della variabile dipendente, l'immagine di questa funzione viene poi legata alla combinazione lineare delle variabili esplicative. Equivalentemente, nei modelli GLM una qualche trasformazione monotona della media è una funzione lineare delle variabili esplicative. In formule, dato un linear predictor $\eta_i = x_i^T \beta$, con la funzione link possiamo avere ad esempio (se il link è il log):

$$\log(\mu_i) = \eta_i = x_i^T \beta.$$

Di seguito riportiamo uno schema riepilogativo della completa specificazione del modello:

- y_i osservazioni, $i = 1, ..., n$
- $x_i = (x_{i1}, x_{i2}, ...x_{iK})^T \in \mathbb{R}^K$ vettore delle K covariate per la i-esima osservazione
- $\omega_i \in \mathbb{R}$ esposizione della singola osservazione, parametro noto

- $Y_i \sim EDM(\theta_i, b(), \phi, \omega_i)$ modello distributivo per $Y_i \quad i = 1, ...n$
- Y_i stocasticamente indipendente da $Y_j \quad \forall i \neq j \quad i = 1, ...n \quad j = 1, ...n$
- $\beta \in \mathbb{R}^K$ vettore (K x 1) di parametri incogniti, ciascuno riferito a una covariata
- $g(\mathbb{E}[Y_i]) = g(\mu_i) = x_i^T \beta \quad \forall i = 1, ...n$ link che definisce la struttura di dipendenza tra le covariate e la media della risposta.

3.4.2 Funzione link

La funzione link di un modello GLM deve fondamentalmente rispettare due importanti restrizioni. Il primo requisito è quello della monotonia stretta, che permette poi l'invertibilità della funzione link stessa. In questo senso non si ha perdita di informazione perché possiamo passare da un certo valore di μ alla sua trasformata e viceversa, dal momento che una funzione invertibile è anche biunivoca. Un secondo requisito sulla funzione link prevede che il dominio di tale funzione corrisponda allo spazio parametrico di μ. A livello intuitivo si nota una certa assonanza tra il ruolo svolto dalla funzione link in un modello GLM e la funzione utilizzata nei modelli additivo e moltiplicativo. Approfondiamo questo parallelismo partendo dalla formula di un modello moltiplicativo quando i dati sono organizzati in forma tabulare.

$$\log(\mu_{h,k}) = \beta_0 + \beta_h + \beta_k.$$

Per riportare i dati in forma di lista (un unico indice cui corrisponde un'osservazione) ci serviremo di variabili dummy X_j.

Data un'osservazione i, la variabile $x_{i,j}$ assume valore 1 se β_j è inclusa nel determinare μ_i, assume valore 0 altrimenti.

In termini pratici, la variabile dummy associata a un livello tariffario (es ETA = PRIMA FASCIA) assegna valore 1 se l'osservazione appartiene alla prima fascia d'età, assegna 0 altrimenti.

Per l'analisi della probabilità di riscatto chiameremo le μ_i con la nomenclatura p_i semplicemente per ricordare che sono delle proporzioni/probabilità, il cui valore varia tra 0 e 1 per costruzione. Il link sarà per noi la funzione logit, per una serie di motivazioni che la rendono la funzione più comunemente utilizzata nei problemi di regressione logistica (vedremo infatti nel seguito che la funzione logit costituisce il link canonico di una variabile aleatoria Relative Binomial). Un'alternativa alla funzione logit è detta funzione probit, che consiste nella funzione quantile (l'inversa della funzione di ripartizione) di una normale standard. Sia la funzione logit sia la probit mappano l'intero asse reale $\mathbb{R}$ nell'intervallo aperto $(0;\ 1)$ e consentono così

di modellare proporzioni/probabilità. Imporre un link di tipo logit equivale anche ad ottenere, come ci aspettiamo, un valore previsto della p_i che sia tra 0 e 1.

Riportiamo di seguito alcune derivazioni sulla funzione inversa del logit, che dimostra quanto appena affermato. Abbiamo appena visto il link di tipo logit:

$$g(\mu_i) = g(p_i) = \log(\frac{p_i}{1-p_i}) = x_i^T \beta.$$

Abbiamo già visto anche che nella Relative Binomial:

$$\theta_i = \log(\frac{p_i}{1-p_i}).$$

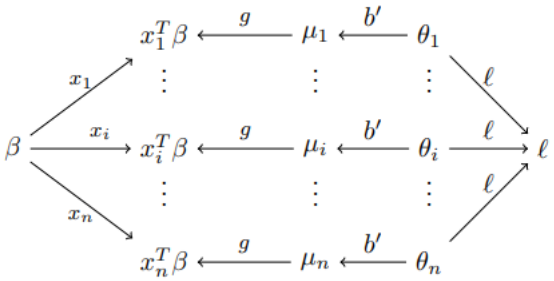
Mettendo insieme le due equazioni sopra otteniamo

$$\theta_i = \log(\frac{p_i}{1-p_i}) = g(p_i) \quad \Longleftrightarrow \quad p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}.$$

3.4.3 Equazioni di verosimiglianza

Il tema del metodo di stima è strettamente connesso alla scelta della forma distributiva EDM, poiché la formulazione di un'ipotesi distributiva delle variabili consente poi di adottare il metodo di stima di massima verosimiglianza. Ci sono molteplici ragioni per cui questo metodo di stima è quello più largamente utilizzato in ambito frequentista. Tali ragioni possono essere riassunte dalle proprietà degli stimatori di massima verosimiglianza. Gli stimatori di massima verosimiglianza infatti sono asintoticamente corretti, consistenti, efficienti e con distribuzione campionaria di tipo gaussiano. Questo significa che, in grandi campioni, lo stimatore del vettore beta sarà una normale multivariata e convergerà sul vero valore del vettore beta. Inoltre, con questo metodo di stima sarà possibile condurre un likelihood ratio test, con distribuzione Chi-quadrato, per confrontare l'adattamento ai dati di due diversi modelli. Questo metodo di stima è adottato anche nella stima dei parametri nei modelli OLS, e sotto le ipotesi di normalità della variabile endogena, lo stimatore di massima verosimiglianza nei modelli OLS coincide con lo stimatore dei minimi quadrati. Nei modelli GLM l'ipotesi di normalità non è necessariamente verificata e quindi un approccio di stima dei minimi quadrati non riconduce poi allo stimatore di massima verosimiglianza. Ci concentreremo dunque sulla derivazione degli stimatori di massima verosimiglianza, trattando il caso di una generica distribuzione EDM e poi calandolo nel caso specifico della distribuzione Relative Binomial. In appendice A.5 vengono riportate le derivazioni appena citate.

Di seguito si riporta uno schema esemplificativo che descrive le relazioni tra i vari parametri.



3.4.4 Link canonico

Nell'ambito delle distribuzioni EDM, lo score può essere trasformato, tramite alcuni passaggi algebrici, in una sorta di somma pesata degli scarti tra le osservazioni y_i dei Key ratios e il vero valore dei parametri che tali Key ratios intendono stimare. In questo senso quindi le equazioni di verosimiglianza (che si ottengono imponendo che il vettore score sia nullo) portano a richiedere che queste somme pesate degli scarti siano nulle. Questo sistema di equazioni può essere riportato anche in forma matriciale. I pesi in questione possono essere semplificati notevolmente scegliendo un link canonico. Data una determinata funzione cumulante b( ), la funzione link si dice canonica se la derivata della funzione cumulante è l'inversa della funzione link.

In appendice (A.6)si riportano i passaggi matematici, sia nel caso generale EDM sia nel caso della Relative Binomial.

Notiamo che nella Relative Binomial la funzione link canonica è rappresentata dalla funzione logit, rafforzando così la nostra scelta metodologica di adottare un link logit nel modello in esame.

3.4.5 Metodi numerici

Il primo step nella procedura di ottimizzazione della log-verosimiglianza consiste nell'imporre che la sua derivata prima (lo score) sia nulla, e risolvere per il punto critico (nel nostro caso si tratta del vettore β).

In linea generale per le distribuzioni EDM le equazioni di verosimiglianza non sono lineari in β perché questo compare anche al denominatore all'interno della funzione varianza (infatti $\nu(\mu) = \nu(g^{-1}(x_i^T\beta))$) e della funzione link (infatti $g'(\mu) = g'(g^{-1}(x_i^T\beta))$. Si configura così una relazione di dipendenza complicata in cui β non è ricavabile agevolmente in maniera analitica. Gli stimatori di verosimiglianza vengono quindi ricavati dai principali software statistici mediante degli algoritmi di risoluzione di sistemi non lineari. Tra questi metodi numerici, quelli più rinomati sono certamente l'algoritmo di bisezione e il metodo di Newton-Raphson.

Entrambi gli algoritmi partono da un vettore di coordinate iniziali β_0. Queste coordinate vengono fissate in maniera più o meno "arbitraria" e indicano un valore iniziale candidato ad essere la soluzione del sistema. In maniera iterativa la soluzione viene aggiornata per avvicinarsi alla radice della funzione fin quando non si raggiunge uno specificato parametro di tolleranza. L'algoritmo di bisezione sfrutta il teorema degli zeri (anche detto teorema di Bolzano) per aggiornare ad ogni iterazione la soluzione proposta, non richiedendo quindi di conoscere le derivate della funzione di cui si cerca la radice. Il metodo di bisezione è certamente potente ma non è applicabile al nostro caso poiché richiede che l'incognita sia monovariata, mentre noi cerchiamo come soluzione un vettore di più parametri, β.

In un certo senso il metodo di Newton risulta più esigente ma anche più sofisticato ed efficiente, poiché nell'aggiornare iterativamente la soluzione proposta tiene conto del gradiente e dell'Hessiano della funzione (nel nostro caso quindi si cerca la radice, intesa come valore in cui la funzione vale zero, della log-verosimiglianza sfruttando le informazioni fornite dal vettore score e dalla matrice delle derivate seconde, ottenuta derivando lo score rispetto a ciascun coefficiente β_j). Il metodo di Newton ha il forte vantaggio di avere convergenza quadratica (diversamente dalla bisezione), ma tale convergenza è solo locale (se il vettore di inizializzazione è distante dalla soluzione, il metodo potrebbe fallire). Inoltre, non è sempre applicabile senza restrizioni (per una trattazione più approfondita vedere Brandimarte [2013]).

La risoluzione delle equazioni di verosimiglianza, nel caso delle distribuzioni EDM, rientra tra le casistiche in cui è opportuno modificare alcune specificazioni per poter applicare l'algoritmo. Nello specifico, è abbastanza complicato esplicitare la forma analitica di un generico elemento della matrice delle derivate seconde della log-verosimiglianza. Per ovviare a questa difficoltà si tende a sostituire la matrice delle derivate seconde con il suo valore atteso (ovvero sostituendo il valore atteso μ_i al Key ratio osservato y_i). Sotto opportune condizioni di regolarità, questa matrice dei valori attesi delle derivate seconde corrisponde alla matrice di informazione di Fisher, col segno cambiato. Si definisce con il nome di "Scoring di Fisher" il metodo che prevede di utilizzare questa matrice di valori attesi nelle iterazioni che aggiornano la stima di β. Si noti che ad ogni iterazione questa matrice va ricalcolata poiché dipende da dei pesi che dipendono a loro volta da β.

La procedura numerica di stima appena illustrata si semplifica notevolmente se si sceglie come funzione link il cosiddetto link canonico, come è stato già illustrato nella sezione precedente. Si può dimostrare che, quando si sceglie la funzione link canonica, il metodo di Newton e quello di Fisher coincidono. Scegliendo il link canonico, inoltre, si può scrivere l'Hessiano come -1 che moltiplica per un prodotto matriciale, il quale coinvolge la matrice di disegno $\mathbf{X}$ e una matrice diagonale positiva

definita. Ne consegue che l'Hessiano sia una matrice negativa definita, confermando così la concavità della log-verosimiglianza e con essa l'esistenza di un unico massimo.

3.4.6 Likelihood ratio test, devianza, modello saturato

La scelta del metodo di massima verosimiglianza per la stima dei parametri permette anche di condurre un test del rapporto di verosimiglianza. Condurre test di significatività sui singoli parametri del modello non apporta informazioni molto rilevanti poiché a ciascun parametro è associato uno e un solo livello di un fattore tariffario. Si pensi ad esempio ad un modello in cui il coefficiente associato a una fascia d'età risulta significativo, mentre il coefficiente associato a un'altra fascia di età risulta non significativo. Un simile risultato farebbe scaturire dei dubbi in merito all'opportunità di includere o meno la fascia d'età tra i fattori tariffari. La decisione di includere o meno un fattore tariffario viene solitamente presa servendosi appunto del test del rapporto di verosimiglianza. In linea generale rispetto a un modello M0, un modello M1 ha più vincoli se include un minor numero di fattori tariffari.

Partendo dal modello con un certo numero di fattori si può infatti ottenere un modello con meno fattori semplicemente imponendo la restrizione secondo cui i parametri β_j siano nulli in corrispondenza delle variabili dummy associate ai fattori tariffari che si vogliono eliminare.

Da queste considerazioni si intuisce come ogni possibile modello sia un caso particolare (restricted) di un modello con più fattori. Questa catena di inclusione culmina nel "modello saturato", ovvero un modello con il massimo numero di parametri possibile, che è pari al numero di osservazioni (la matrice $\mathbf{X}$ sarà infatti quadrata, e più specificamente sarà la matrice identità). Il modello saturato associa infatti a ciascuna osservazione il relativo parametro ($\theta_i = b'^{-1}(y_i)$ poiché fissa $\mu_i = y_i$) ed è quello che riesce ad adattarsi in maniera più fedele ai dati, raggiungendo così il massimo della funzione di verosimiglianza. Ogni altro modello con k parametri rappresenta una versione restricted (con n-k restrizioni lineari) del modello saturato e pertanto le relative verosimiglianze raggiungeranno valori pari o inferiori (essendo dei massimi vincolati) alla verosimiglianza del modello saturato. In questo senso quindi il modello saturato funge da benchmark per valutare l'adattamento degli altri modelli. All'estremità opposta troviamo il modello nullo, che associa un unico parametro ($\mu_i = \mu$ media generale) a tutte le osservazioni, e che pertanto ottiene un valore di verosimiglianza inferiore a qualsiasi altro modello.

Una misura comunemente adottata per valutare la bontà di adattamento di un modello, anche in ambiti diversi dai modelli GLM, consiste nel Likelihood Ratio

Test, in cui il modello proposto è il modello restricted e nel nostro caso il modello saturato assume il ruolo di modello unrestricted. La logica di fondo di questo tipo di test si basa sul fatto che la verosimiglianza del modello saturato funge da upper bound delle verosimiglianze di qualsiasi altro modello: ne consegue che

$$\frac{L_{prop}(\hat{\theta}_{prop}, y)}{L_{satur}(\hat{\theta}_{satur}, y)} \in [0; 1]$$

dove $\hat{\theta}$ sono le rispettive stime di massima verosimiglianza nei modelli proposto e saturato.

Spesso si ricorre a un'approssimazione asintotica per ricavare una statistica test che abbia distribuzione nota sullo spazio campionario. Tale approssimazione costituisce una trasformazione del rapporto di verosimiglianza, con la seguente forma analitica:

$$\begin{aligned} T &= -2\log(\frac{L_{prop}(\hat{\theta}_{prop}, y)}{L_{satur}(\hat{\theta}_{satur}, y)}) \\ &= -2(l_{prop}(\hat{\theta}_{prop}, y) - l_{satur}(\hat{\theta}_{satur}, y)) \\ &= 2(l_{satur}(\hat{\theta}_{satur}, y) - l_{prop}(\hat{\theta}_{prop}, y)). \end{aligned}$$

Asintoticamente $T \sim \chi^2(n-k)$ per $n \longrightarrow \infty$ e con k =numero di parametri del modello proposto. Le ipotesi del test sono:

H_0 : Mod_{satur} non è significativamente più accurato di Mod_{prop}

H_1 : Mod_{satur} è significativamente più accurato di Mod_{prop}

Rifiuteremo H_0 se osserviamo nel campione che $T(y) > \chi^2_{1-\alpha}(n-k)$.

A titolo esemplificativo, un modello proposto che abbia una verosimiglianza molto vicina a quella del modello saturato porterà il rapporto di verosimiglianza poco al di sotto di 1. Applicando il logaritmo e cambiando segno si avrà quindi un valore positivo piccolo, che molto probabilmente rientrerà nella regione di accettazione della statistica test, poiché la distribuzione Chi-quadrato ha supporto positivo e la regione di rifiuto ricade unicamente nella coda destra. Saremo quindi portati ad accettare H_0 ovvero che il modello saturato non è significativamente più accurato del modello proposto.
La formula analitica di questo test statistico è di fatto corrispondente alla devianza scalata, dove l'aggettivo "scalata" fa riferimento al fatto che la quantità in questione rappresenta la devianza del modello, moltiplicata per il reciproco del coefficiente di

dispersione. In formule:

$$\begin{aligned} D.Scaled &= 2(l_{satur}(\hat{\theta}_{satur}, y) - l_{prop}(\hat{\theta}_{prop}, y)) \\ &= 2\sum_{i=1}^{n} \frac{\omega_i}{\phi}\big((y_i\hat{\theta}_{satur} - b(\hat{\theta}_{satur})) - (y_i\hat{\theta}_{prop} - b(\hat{\theta}_{prop})\big), \end{aligned}$$

dove la seconda uguaglianza discende direttamente dalla forma analitica già vista della log-verosimiglianza di una distribuzione EDM.

$$D = (D.scaled)\phi = 2\sum_{i=1}^{n} \omega_i((y_i\hat{\theta}_{satur} - b(\hat{\theta}_{satur})) - (y_i\hat{\theta}_{prop} - b(\hat{\theta}_{prop}))$$

Una prima osservazione va rivolta alla distribuzione della devianza scalata sullo spazio campionario. Abbiamo visto che la forma analitica della devianza scalata coincide con quella dell'approssimazione del rapporto di verosimiglianza. Sotto le stesse ipotesi viste per tale approssimazione possiamo quindi affermare che

$$D.Scaled \sim \chi^2(n-k) \quad \implies \mathbb{E}\left[D.scaled\right] = (n-k),$$

per $n \longrightarrow \infty$ e con k =numero di parametri del modello proposto.

Possiamo ricavare così una stima del parametro di dispersione:

$$\mathbb{E}\left[D\right] = \mathbb{E}\left[(D.scaled)\phi\right] = (n-k)\phi \quad \implies \quad \hat{\phi} = \frac{\mathbb{E}\left[D\right]}{n-k}.$$

Notiamo inoltre che la devianza del modello è una funzione decrescente della verosimiglianza del modello proposto; pertanto, la stima con metodo di massima verosimiglianza coincide con un processo di stima che punti a minimizzare la devianza del modello. Questa devianza, infatti, misura in relazione inversa la bontà di adattamento ai dati ottenuta dal modello. Riassumendo, il modello saturato presenta una devianza che costituisce il lower bound delle devianze e una log-verosimiglianza che costituisce l'upper bound delle log-verosimiglianze. Per il modello nullo vale invece il contrario: questo modello raggiunge l'upper bound delle devianze e il lower bound delle log-verosimiglianze. Quanto appena visto rinforza il legame inverso tra verosimiglianza e devianza.

Un'ulteriore misura di bontà di adattamento è costituita dall'indice Pseudo R Squared di McFadden. Questo indice permette di estendere con maggiore continuità alcuni concetti dei modelli OLS al mondo dei modelli GLM. Nei modelli OLS infatti esiste una misura, detta appunto R Squared, che permette di valutare la bontà di adattamento di un modello semplicemente calcolando il peso della devianza spiegata

dalla regressione sulla devianza totale.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y_i})^2 - \sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} = \frac{SS(M_{null}) - SS(M_{proposed})}{SS(M_{null})}$$

dove SS sta per Sum of Squared residuals. Quest'ultima quantità coincide con le devianze del modello solo nel caso di un modello OLS con distribuzione Normale della variabile risposta. L'indicatore Pseudo R squared di McFadden riprende questa logica, ma lo fa passando per le verosimiglianze. In formule:

$$\begin{aligned} PseudoR^2 &= \frac{l(\hat{\theta}_{null}) - l(\hat{\theta}_{proposed})}{l(\hat{\theta}_{null}) - l(\hat{\theta}_{saturated})} \\ &= \frac{2(l(\hat{\theta}_{proposed}) - l(\hat{\theta}_{null}))}{2(l(\hat{\theta}_{saturated}) - l(\hat{\theta}_{null}))} \\ &= \frac{2(l(\hat{\theta}_{saturated}) - l(\hat{\theta}_{null})) - 2(l(\hat{\theta}_{saturated}) - l(\hat{\theta}_{proposed}))}{2(l(\hat{\theta}_{saturated}) - l(\hat{\theta}_{null}))} \\ &= \frac{D.scaled(M_{null}) - D.scaled(M_{prop})}{D.scaled(M_{null})} \in [0;1] \end{aligned}$$

Nella sezione di analisi utilizzeremo la funzione ANOVA di R per condurre dei test sulla significatività di interi fattori tariffari e ritroveremo indicazioni sul Likelihood Ratio Test e sulla devianza sopra illustrati.

3.4.7 Residui

Una quantità di grande interesse nell'analisi e nel confronto tra modelli è certamente rappresentata dai residui. I residui solitamente indicano infatti una distanza (con segno) tra le previsioni del modello e le osservazioni effettive. In un certo senso quindi i residui misurano la bontà di adattamento e/o la capacità previsiva di un modello. Nel caso dei modelli OLS, le distanze tra osservazioni e valori previsti sono delle semplici differenze:

$$r_i = y_i - \hat{\mu_i}$$

Nel caso dei modelli GLM non è possibile però adottare questo tipo di metrica, si utilizzano infatti i residui di Pearson.

$$r_{P_i} = \frac{y_i - \hat{\mu_i}}{\sqrt{\nu(\hat{\mu_i})/\omega_i}}.$$

Si evidenzia il fatto che i residui di Pearson di fatto scalano i residui r_i dividendoli per le rispettive deviazioni standard, che troviamo al denominatore. Tali deviazioni

standard dipendono dall'esposizione di ogni singola osservazione e dalla varianza, che a sua volta dipende dalla stima ottenuta $\hat{\mu}_i$.

In questa sede ci limitiamo solamente ad accennare al fatto che la somma dei quadrati di questi residui restituisce la statistica Chi-quadrato di Pearson, una misura di adattamento globale del modello che si distribuisce come una $\chi^2(n-k)$ dove k è il numero di parametri stimati:

$$\chi^2 = \sum_{i=1}^{n} r_{P_i}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)/\omega_i}.$$

Capitolo 4

Metriche di performance previsiva

Il presente capitolo sarà dedicato a una breve rassegna dei metodi più comunemente utilizzati per valutare la capacità di un classificatore binario di prevedere la classe cui appartiene una data osservazione. Più in particolare, queste metriche intendono misurare quanto un modello sia in grado di prevedere correttamente la classe di appartenenza di un'osservazione, pur senza aver mai visto le caratteristiche di tale osservazione in fase di apprendimento. In questo capitolo si farà spesso ricorso alla terminologia propria del machine learning e della statistica computazionale, proprio perché queste metriche di valutazione sono state ampiamente sfruttate dal filone più recente della statistica (la cosiddetta data science). Un tema di vitale importanza in questo filone è appunto quello dell'apprendimento automatico. La prospettiva è quella di sviluppare algoritmi che estrapolino degli schemi di conoscenza e delle tendenze di fondo insite nei dati empirici forniti dall'esperienza. Il paradigma di apprendimento è esattamente quello dell'essere umano: a partire dai dati formulare dei modelli o inferire dei legami che possano dare una spiegazione dei fenomeni e quindi del mondo circostante. Quando si parla quindi di learning o training di un modello, si intende in realtà la fase di stima dei parametri di un modello, che avviene a partire da un dataset apposito (il cosiddetto data-train, in questa trattazione). A partire dal data-train il modello apprende dei legami e delle tendenze, utili per poter fare previsioni. Calando questa affermazione nel caso in esame, si evidenzia che il processo di stima di una regressione logistica sul data-train culmina nell'apprendimento dei coefficienti β_j che legano i livelli tariffari alla propensione al riscatto. A livello di terminologia, un procedimento di stima di regressione logistica rientra nell'ambito del supervised statistical learning, dove l'aggettivo supervised si riferisce al fatto che il modello può utilizzare le variabili risposta (la dummy di riscatto del data-train) per inferire delle relazioni e degli schemi di conoscenza.

La prassi di analisi statistica prevede solitamente la suddivisione dei dati a disposizione in un dataset di training e uno di testing. Illustreremo quindi molto breve-

mente le più diffuse metriche di valutazione delle previsioni nell'ambito di problemi di classificazione binaria. Queste metriche si fondano su una logica di confronto tra valori effettivi nel dataset di test e previsioni date dal modello sul dataset di test.

4.1 Matrice di confusione e indicatori

La matrice di confusione o "confusion matrix" associa la situazione effettiva (status di riscatto/non riscatto disponibili nel test dataset) a quella prevista. La struttura di una matrice di confusione è la seguente.

	NegativiEffettivi	PositiviEffettivi
NegativiPrevisti	TN	FN
PositiviPrevisti	FP	TP

dove

- TP sono i True Positive, ovvero le osservazioni previste come positive che sono effettivamente positive.
- TN sono i True Negative, ovvero le osservazioni previste come negative che sono effettivamente negative.
- FP sono i False Positive, ovvero le osservazioni previste come positive che però sono effettivamente negative.
- FN sono i False Negative, ovvero le osservazioni previste come negative che però sono effettivamente positive.

La previsione come riscatto o non riscatto dipende di fatto da un valore soglia (threshold), per il quale la scelta più intuitiva e più comune è il valore di 0,5. Va però evidenziato, come si vedrà nella sezione successiva, che non sempre questa soglia rappresenta la scelta più corretta, specialmente quando il data-train di partenza presenta una variabile dummy di risposta non bilanciata (molti più casi positivi dei negativi o viceversa). Se la $\hat{p}_i$ prevista supera tale soglia, il nostro modello sta prevedendo che quella osservazione sia un riscatto, se non supera tale soglia la sta classificando come non riscattata. Ci sono numerose metriche di valutazione dell'accuratezza di previsione di un classificatore binario.

In primo luogo troviamo certamente l'accuracy, ovvero il rapporto tra previsioni corrette e totale delle osservazioni da prevedere

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN}.$$

Va notato però che, a seconda dell'ambito di applicazione, potrebbe non essere necessariamente l'accuracy generale a essere rilevante, quanto altre metriche più specifiche, quali ad esempio:

$$Sensitivity = \frac{TP}{TP + FN} = \text{True Positive Rate}(TPR)$$

Il TPR indica, sul totale delle osservazioni effettivamente positive, quante sono previste correttamente.

$$Specificity = \frac{TN}{TN + FP} = \text{True Negative Rate}(TNR)$$

Il TNR indica, sul totale delle osservazioni effettivamente negative, quante sono previste correttamente.

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

Il False Positive Rate indica, sul totale delle osservazioni effettivamente negative, quante sono previste erroneamente come positive.

Nel caso in esame dell'analisi dei riscatti, le esigenze di previsione non sono univoche. Da un lato, si potrebbe puntare a massimizzare l'accuracy nel suo complesso, in accordo con la logica di stima Best-Estimate su cui si fonda tutta la normativa Solvency II. Dall'altro, un approccio più prudenziale per la compagnia potrebbe essere quello di puntare a minimizzare i falsi negativi (e dunque a massimizzare la sensitivity), ovvero quelli che riscattano senza che la compagnia l'avesse previsto. Un modo per dare maggiore peso all'una o all'altra metrica è sicuramente quello di aggiustare la threshold di conseguenza. Vedremo nella sezione successiva come questi indicatori risentono della scelta della threshold utilizzata.

4.2 Curva ROC, AUC, Threshold tuning

Come già anticipato, variazioni della soglia di classificazione portano a variazioni del FPR e del TPR. Per visualizzare questa relazione si ricorre generalmente alla Receiver Operating Characteristics Curve (Curva ROC), la quale individua per ogni livello di soglia una coppia di coordinate cartesiane corrispondenti al TPR in ordinata e al FPR in ascissa, ottenuti utilizzando tale soglia. Ne risulta un grafico che presenta una curva che supera in altezza la bisettrice del primo quadrante, in corrispondenza della quale si avrebbe un classificatore casuale (Predicting by chance).

Un'operazione importante nello svolgere la classificazione è certamente la scelta della threshold di classificazione, soprattutto se nel dataset di training i positivi e

i negativi effettivi non sono equamente rappresentati. La calibrazione di tale soglia (threshold tuning) non ha una formula specifica o corretta in termini assoluti, presenteremo quindi l'approccio più diffuso di threshold tuning. Di fatto questo approccio prescrive di scegliere il valore di soglia che permette di ottenere l'angolo in alto a sinistra della curva ROC (si veda il capitolo di analisi per una rappresentazione grafica della curva) e più precisamente un punto che appartenga alla retta di equazione cartesiana $y = 1 - x$. Numericamente questo corrisponde a cercare una soglia t tale che:

$$TPR(t) = 1 - FPR(t).$$

Questo problema di root-finding viene spesso convertito in un problema di ottimizzazione di questo tipo:

$$\hat{t} = \arg\min_t |TPR(t) + FPR(t) - 1|.$$

La curva ROC costituisce uno strumento molto utile anche perché serve per calcolare una metrica ad essa strettamente connessa, ovvero l'area sottesa alla curva ROC, denominata Area Under the Curve (AUC). Se per la curva ROC la bisettrice costituisce una sorta di benchmark, nel caso della AUC lo stesso benchmark è appunto rappresentato da 0,5 , ovvero dall'area sottesa alla bisettrice nell'intervallo (0;1). Altri esempi di metriche utilizzabili in alternativa all'AUC per valutare l'efficacia di un modello sono il coefficiente di Gini e il test di Kolmogorov Smirnov. Il coefficiente di Gini è collegato strettamente all'AUC, infatti vale la formula

$$Gini = 2AUC - 1 \quad \iff \quad AUC = \frac{Gini + 1}{2}.$$

Nel prossimo capitolo di avremo modo di servirci di queste metriche per valutare la performance previsiva dei modelli.

Capitolo 5

Analisi

In questa sezione verrà illustrato l'intero processo che ha portato alla formalizzazione del modello, dalla fase preliminare di elaborazione dei dati fino alla derivazione dei risultati, che verranno presentati.

5.1 Preparazione del dataset

Il dataset è stato ottenuto a partire da dei database di una nota compagnia italiana di assicurazioni nel ramo vita. I database in questione costituiscono delle fotografie trimestrali della situazione di portafoglio della compagnia. Ogni database contiene quindi tutte e solo le polizze presenti nel portafoglio della compagnia in un determinato istante di osservazione. In questa tesi si è deciso di rivolgere l'attenzione alle sole polizze previdenziali. Per questo motivo sono stati utilizzati dei database della suddetta compagnia che fossero riferiti unicamente a un prodotto a scopo pensionistico. Il prodotto in questione è una polizza collocata come Piano Individuale Pensionistico di tipo assicurativo. Si tratta quindi di un fondo pensione iscritto all'Albo tenuto dalla COVIP e finalizzato all'erogazione di trattamenti pensionistici complementari. Il prodotto scade al raggiungimento dell'età pensionabile, salvo riscatti totali o decesso del contraente. Il trasferimento della posizione previdenziale verso un'altra forma pensionistica complementare può avvenire solamente una volta decorso un periodo iniziale di due anni. L'aderente, anche prima del periodo minimo di permanenza, può:

- trasferire la posizione ad un'altra forma pensionistica complementare, alla quale acceda in riferimento a una nuova situazione lavorativa;
- riscattare il 50 per cento della posizione, in caso di cessazione dell'attività lavorativa che comporti l'inoccupazione per un periodo di tempo non inferiore a dodici mesi e non superiore a quarantotto mesi, oppure in caso di ricorso da

parte del datore di lavoro a procedure di mobilità, cassa integrazione guadagni, ordinaria o straordinaria;

- riscattare l'intera posizione, in caso di invalidità permanente che comporti la riduzione della capacità di lavoro a meno di un terzo o a seguito di cessazione dell'attività lavorativa, che comporti l'inoccupazione per un periodo di tempo superiore a quarantotto mesi.

- riscattare l'intera posizione individuale maturata, oppure trasferirla ad altra forma pensionistica complementare, qualora vengano meno i requisiti di partecipazione.

In questa sede non ci occuperemo di riscatti parziali e anticipazioni, tralasciamo quindi le caratteristiche di prodotto relative all'esercizio di tali opzioni contrattuali.

A livello operativo il procedimento di integrazione dei dati ha seguito un percorso iterativo, pertanto verrà descritta la procedura seguita per un generico trimestre. Si è creata innanzitutto una chiave univoca che concatenasse il numero della polizza e il periodo di osservazione. Il numero di polizza è stato utilizzato per verificare se una determinata polizza fosse presente anche nella fotografia di portafoglio del trimestre successivo. Per le polizze non presenti, ovvero quelle che sono uscite dal portafoglio nel corso del trimestre, si è resa necessaria la ricerca del motivo dell'uscita dal portafoglio. Quest'ultima informazione è stata integrata da un ulteriore database. Va infatti precisato che una polizza può uscire dal portafoglio anche per motivi diversi dal riscatto, quali ad esempio la mortalità o la semplice scadenza (intesa come raggiungimento dell'età pensionabile).

Ai fini della presente analisi sono stati considerati tra i riscatti totali gli eventi di uscita dal portafoglio dovuti a una di queste casistiche:

- 'Comunicazione di Recesso Individuale',

- 'Comunicazione di Riscatto Totale Individuale',

- 'Insolvenza Individuale',

- 'Riscatto Totale per Perdita Status di Lavoratore Volontaria',

- 'Riscatto Totale per Perdita Status di Lavoratore non volontaria',

- 'Trasferimento Fondo'.

Il dataset completo è stato ricavato accostando in colonna le fotografie dei portafogli nei vari quarter. Tutte le fuoriuscite dai portafogli che fossero imputabili a una delle causali sopra menzionate sono state contrassegnate con valore 1 nella variabile

dummy appositamente creata e denominata nel dataset con etichetta 'riscatto2'. Si è deciso di creare anche la variabile "period" valorizzandola con il quarter di osservazione del singolo record del dataset (ovvero il momento di estrazione dati della fotografia di portafoglio da cui proviene il record).

Tra tutte le variabili presenti nel dataset sono state considerate soltanto le variabili che sono state utilizzate in letteratura nei vari studi già citati.

Il dataset è stato poi suddiviso in un dataset di allenamento del modello, il data-train, e in un dataset di verifica delle capacità previsive, il data-test. La suddivisione è avvenuta in base al periodo di osservazione: il data-test coincide con la fotografia di portafoglio all'ultimo quarter disponibile, tutti gli altri record rientrano nel data-train. Questa decisione riflette in parte anche l'intento di fondo della creazione di un modello previsivo, che riesca a prevedere il numero di riscatti in portafoglio a partire da quanto osservato in precedenza.

Si specifica inoltre che i dati sono stati anonimizzati: le variabili quantitative "riserva matematica" e "cumulo premi pagati" sono state moltiplicate per una costante di scala che per ovvie ragioni non sarà menzionata. Questo procedimento non comporta distorsioni nei dati e nelle loro relazioni con i riscatti, trattandosi semplicemente di una trasformazione di scala.

Un altro importante step ha visto la discretizzazione delle variabili quantitative di interesse. Sono state discretizzate le variabili seguenti:

- cumulo premi pagati,
- riserva matematica,
- età,
- durata totale (dalla stipula alla scadenza),
- durata residua (dalla data di valutazione alla scadenza),
- antidurata (dalla data di stipula a quella di valutazione).

La scelta della metodologia di discretizzazione delle variabili non è semplice e i possibili approcci sono molteplici e con diversi livelli di sofisticazione (si pensi ad esempio a degli algoritmi di unsupervised learning per il clustering). In questa sede si è deciso di seguire un approccio più semplificato, scegliendo sostanzialmente di dividere le variabili quantitative in 4 fasce corrispondenti ai 4 quartili delle variabili stesse. I quartili sono riferiti alla distribuzione del solo data-train. La presenza di variabili che variano con il tempo costringerà poi a dover ricalibrare le soglie (i quartili) nel tempo.

Si evidenzia più in generale una criticità: i quartili presi come criteri di classificazione sono quantili empirici e dunque soggetti a variabilità. Occorre però considerare la notevole numerosità del nostro dataset, caratteristica che ci permette di assumere che i quartili empirici siano una stima relativamente robusta dei quartili teorici. Un'altra possibile criticità risiede nel fatto che una tale suddivisione potrebbe creare degli sbilanciamenti nel momento in cui ci fossero troppo pochi eventi di riscatto in un sottocampione corrispondente a un certo quartile. A questo proposito si è deciso di osservare le proporzioni di riscatto condizionate ai vari quartili del data-train. Si riportano di seguito le suddette proporzioni, evidenziando che in nessun quartile di nessuna variabile si nota un valore eccessivamente prossimo allo 0 e nemmeno eccessivamente distante dalla proporzione di riscatti su tutto il data-train, ovvero il 2 su 1000 circa.

```
> print(cbind(lapse_by_ETA, lapse_by_CUM_PREM, lapse_by_RM))
  ETA_Q   riscatto2 CUM_PREMI_Q   riscatto2 RM_Q   riscatto2
1     1 0.002392985           1 0.001492013    1 0.001638300
2     2 0.002172537           2 0.001472754    2 0.003740648
3     3 0.001921994           3 0.005274262    3 0.002309469
4     4 0.001796339           4 0.002077693    4 0.002077387
> print(cbind(lapse_by_DUR_TOT, lapse_by_DUR_RESID, lapse_by_ANTID))
  DUR_TOT_Q    riscatto2 DUR_RESID_Q   riscatto2 ANTID_Q   riscatto2
1         1 0.0007132668           1 0.001753426       1 0.001130749
2         2 0.0018018018           2 0.001661002       2 0.002833956
3         3 0.0009341429           3 0.001646233       3 0.002797425
4         4 0.0020787075           4 0.002127280       4 0.002269126
>
```

I numeri da 1 a 4 si riferiscono alla fascia di quartile, i numeri decimali alla loro destra sono le proporzioni condizionate. Ad esempio la seconda fascia di età ha una proporzione di riscatti pari allo 0.002172537, mentre la terza fascia di riserva matematica ha una proporzione di riscatti del 0.002309469. Questa scelta di discretizzazione permette anche, come si è visto, di imporre che le 3 (4 livelli - 1 dell'intercetta) equazioni di verosimiglianza riferite al fattore discretizzato valgano per il 25% del data-train ciascuna.

5.2 Aggregazione in celle

Una volta discretizzate le variabili quantitative continue si è resa possibile l'aggregazione in celle del nostro data-train. Seguendo gli approcci visti in letteratura si è deciso di utilizzare i seguenti fattori di aggregazione:

- period: trimestre di calendario,

- SESSO del policyholder,
- FRAZ: periodicità dei versamenti,
- LAV AUTO: variabile dummy, vale 1 se il contraente è un lavoratore autonomo,
- ETA Q: fasce di età in quartili,
- RM Q: fasce di riserva matematica in quartili,
- CUM PREMI Q: fasce di cumulo premi in quartili,
- DUR RESID Q: fasce di durata residua in quartili,
- DUR TOT Q: fasce di durata totale in quartili,
- ANTID Q: fasce di antidurata (età della polizza) in quartili.

In ciascuna cella si ottiene il valore di esposizione (pari al numero di polizze-trimestre presenti nella cella) e il numero di riscatti presenti nella cella. Si ricava il valore empirico del Key ratio Y nelle varie celle semplicemente rapportando il numero di riscatti nella cella al rispettivo valore di esposizione. Si riporta un estratto del dataframe costituito dalle celle appena descritte.

```
> head(celle)
  period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q riscatto2 esposiz Key_ratio
1   21q4     F    1       NO     3    1           1           1         1       1         0       3         0
2   22q1     F    1       NO     3    1           1           1         1       1         0       4         0
3   22q2     F    1       NO     3    1           1           1         1       1         0       4         0
4   22q3     F    1       NO     3    1           1           1         1       1         0       2         0
5   22q4     F    1       NO     3    1           1           1         1       1         0       2         0
6   21q3     M    1       NO     3    1           1           1         1       1         0       1         0
> tail(celle)
      period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q riscatto2 esposiz   Key_ratio
10485   21q3     M   12       SI     4    4           4           4         4       4         6    1453 0.004129387
10486   21q4     M   12       SI     4    4           4           4         4       4         2    1465 0.001365188
10487   22q1     M   12       SI     4    4           4           4         4       4         3    1472 0.002038043
10488   22q2     M   12       SI     4    4           4           4         4       4         2    1518 0.001317523
10489   22q3     M   12       SI     4    4           4           4         4       4         0    1607 0.000000000
10490   22q4     M   12       SI     4    4           4           4         4       4         2    1633 0.001224740
>
```

5.3 Modello GLM

Proseguiamo quindi con la stima dei modelli GLM. Per la stima dei parametri è stato utilizzato il comando "glm" con parametro family = binomial(link = "logit"), il modello è stato stimato sul data-train. Il parametro "weights" del comando glm non è stato specificato: in questo modo il modello stimato sul data-train considera ogni riga come un'osservazione di una polizza trimestre, mantenendo così $\omega_i = 1 \ \forall i$ come desiderato.

5.3.1 Modello 1

Inizialmente si è deciso di partire da un modello (mod1) che includesse tutti i fattori tariffari individuati per la creazione delle celle, ad eccezione del periodo di osservazione. Quest'ultimo infatti non può essere incluso nel modello se si vogliono effettuare delle previsioni sul data-test; ciò è dovuto al modo in cui è stato ricavato il data-test stesso (filtrando cioè le sole osservazioni riferite all'ultimo period).

La stima di tale modello avviene mediante delle apposite funzioni del software R e restituisce i seguenti coefficienti, ciascuno associato a un livello tariffario.

```
Call:
glm(formula = cbind(riscatto2, esposiz - riscatto2) ~ SESSO +
    FRAZ + LAV_AUTO + ETA_Q + RM_Q + CUM_PREMI_Q + DUR_RESID_Q +
    ANTID_Q + DUR_TOT_Q, family = binomial(link = "logit"), data = data_train,
    x = TRUE)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.412289   0.748121  -9.908  < 2e-16 ***
SESSOM       -0.018557   0.028520  -0.651 0.515259
FRAZ2        -0.090923   0.084254  -1.079 0.280519
FRAZ3         0.260764   0.119483   2.182 0.029077 *
FRAZ4        -0.158195   0.043976  -3.597 0.000322 ***
FRAZ6        -0.114476   0.058651  -1.952 0.050961 .
FRAZ12       -0.205007   0.035538  -5.769 7.99e-09 ***
LAV_AUTOSI    0.082718   0.040012   2.067 0.038704 *
ETA_Q2       -0.130923   0.037599  -3.482 0.000497 ***
ETA_Q3       -0.219029   0.038829  -5.641 1.69e-08 ***
ETA_Q4       -0.177661   0.045969  -3.865 0.000111 ***
RM_Q2         0.021067   0.787496   0.027 0.978657
RM_Q3        -1.170075   0.966404  -1.211 0.225991
RM_Q4        -1.440623   0.697283  -2.066 0.038824 *
CUM_PREMI_Q2  0.609532   1.156439   0.527 0.598141
CUM_PREMI_Q3  2.574420   0.779777   3.301 0.000962 ***
CUM_PREMI_Q4  1.845356   0.718550   2.568 0.010224 *
DUR_RESID_Q2  0.007228   0.133844   0.054 0.956933
DUR_RESID_Q3  0.017763   0.125370   0.142 0.887332
DUR_RESID_Q4  0.244832   0.058367   4.195 2.73e-05 ***
ANTID_Q2      0.938825   0.074363  12.625  < 2e-16 ***
ANTID_Q3      0.929262   0.072761  12.771  < 2e-16 ***
ANTID_Q4      0.752974   0.045879  16.412  < 2e-16 ***
DUR_TOT_Q2    0.730842   1.000706   0.730 0.465191
DUR_TOT_Q3   -0.185881   1.000853  -0.186 0.852662
DUR_TOT_Q4    0.221824   0.710320   0.312 0.754822
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La prima linea di analisi del modello mod1 passa per la valutazione di questi coefficienti stimati. Nella notazione vista finora, le voci "Estimate" corrispondono ai β_j con j generico livello tariffario. Gli "Std. Error" sono le deviazioni standard degli stimatori dei vari β_j. Iniziamo osservando che la cella di riferimento avrà le seguenti caratteristiche: sesso femminile, con frazionamento annuale del premio, lavoratori

NON autonomi, nella prima fascia di quartile per tutte le variabili discretizzate in quartili.

Notiamo innanzitutto che il sesso presenta una stima negativa di β_j ma poco significativa. Possiamo concludere (limitatamente al nostro campione) che in questo modello il genere del policyholder non risulta significativo. La periodicità del premio, ovvero il suo frazionamento, presenta tutti coefficienti negativi (ad eccezione della classe di frazionamento in quadrimestri, FRAZ3), anche se solo i coefficienti negativi associati ai livelli FRAZ4 e FRAZ12 sono fortemente significativi. Si potrebbe forse ipotizzare (al netto di qualche irregolarità) una tendenza generale per cui i riscatti sono tanto meno probabili quanto più è frazionato il premio: un premio più frazionato rappresenta per il policyholder un'uscita periodica di entità minore rispetto a un premio con pagamenti meno frequenti ma di maggiore entità. Questi ultimi potrebbero generare delle crisi di liquidità, portando ad un aumento dei riscatti, in accordo con la Emergency Fund Hypothesis (EFH). Si sottolinea però la debolezza dell'evidenza a disposizione in tal senso, motivo per cui per il momento non si possono fare altro che congetture sulla dinamica del fenomeno.

Lo status di lavoratore autonomo sembra avere un effetto di incremento sulla probabilità di riscatto, con un coefficiente che è comunque significativo per un livello di confidenza del 5 per cento. Anche qui l'evidenza non è particolarmente forte o comunque non è indicato trarre conclusioni sul primo modello stimato, ma si può certamente ritrovare una (seppur debole) argomentazione a favore della EFH se si considera che i lavoratori autonomi possono essere più facilmente esposti a fluttuazioni di reddito e quindi a crisi di liquidità.

Si riscontra una notevole importanza del fattore tariffario ETA, per il quale possiamo osservare che tutti i livelli sono molto significativi. Quanto al segno dei coefficienti, si evidenzia che siano tutti valori negativi: abbiamo quindi indicazione del fatto che la fascia più giovane di policyholder tende a riscattare più spesso. Stando a queste evidenze, infatti, le prime due fasce di età sembrano essere quelle più propense al riscatto, seguite poi dall'ultima fascia di età. La terza fascia di età sembra essere quella che riscatta meno di tutte le altre. Anche qui in letteratura sono state riscontrate delle dinamiche simili, spesso ascritte alla teoria di EFH ma anche a quella di Interest Rate Hypothesis.

Possiamo invece osservare come la riserva matematica non sembri essere particolarmente significativa in nessun livello.

Per il fattore cumulo premi abbiamo evidenze contrastanti, dal momento che due livelli su tre hanno coefficienti significativi. Serviranno quindi ulteriori analisi per determinare l'apporto di questo fattore nella spiegazione della propensione al riscatto, per il momento ci limitiamo ad osservare che le due classi con il cumulo

premi più alto sembrano più propense al riscatto.

Anche per quanto riguarda la durata residua valgono le considerazioni fatte sopra sulla dubbia significatività del fattore.

Diversa è la situazione per l'antidurata (età della polizza): tutti i coefficienti sono molto significativi. Quanto ai segni, vediamo che le fasce più propense al riscatto sono le due centrali: questo rispecchia quanto avevamo già visto esaminando le proporzioni dei riscatti condizionate alle fasce di antidurata ("lapse-by-antid").

La durata totale non sembra invece influenzare la propensione al riscatto per nessun livello.

Le variabili di durata totale, durata residua e antidurata sono perfettamente correlate ma in questo frangente non dobbiamo necessariamente allarmarci dal momento che stiamo utilizzando le loro versioni discretizzate secondo i rispettivi quartili. Va comunque sottolineata l'evidenza contrastante che deriva da questi fattori, sul piano della significatività ancor prima che sul piano dei segni dei coefficienti ad essi associati. Torneremo in seguito su questa problematica.

Si riporta in seguito un'indicazione sulle devianze relative al modello.

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74403  on 2497025  degrees of freedom
Residual deviance: 73923  on 2497000  degrees of freedom
AIC: 73975

Number of Fisher Scoring iterations: 9

>
```

Notiamo l'indicazione delle devianze rispettivamente nulla e residua, due quantità che sono coinvolte nella formula analitica dell'indicatore di adattamento Pseudo R Squared di McFadden.

Si conduce anche una verifica empirica sulla soluzione numerica dell'algoritmo di massima verosimiglianza: la soluzione raggiunta dall'algoritmo sembra essere in linea con i risultati teorici delle equazioni di verosimiglianza $\mathbf{X}^T(y-p)=0$, a meno di errori numerici con un ordine di grandezza molto modesto. L'output di seguito riportato conferma quanto illustrato in appendice in merito alla matrice $\mathbf{X}$ trasposta: di fatto somma per ogni riga gli scarti (y_i-p_i) di tutte le osservazioni che presentano un determinato livello.

```
> t(X1) %*% (data_train$riscatto2 - p_i_hat) # e-07 praticamente vettore di zeri
                    [,1]
(Intercept)  -4.262040e-06
SESSOM       -2.547961e-06
FRAZ2        -1.286682e-07
FRAZ3        -6.408195e-08
FRAZ4        -3.980878e-07
FRAZ6        -2.379509e-07
FRAZ12       -1.191860e-06
LAV_AUTOSI   -5.518366e-07
ETA_Q2       -4.461382e-07
ETA_Q3       -1.145398e-06
ETA_Q4       -2.001730e-06
RM_Q2        -1.383011e-08
RM_Q3        -7.425051e-09
RM_Q4        -4.193239e-06
CUM_PREMI_Q2 -3.090628e-09
CUM_PREMI_Q3 -1.422372e-08
CUM_PREMI_Q4 -4.198867e-06
DUR_RESID_Q2 -4.585142e-08
DUR_RESID_Q3 -4.698380e-08
DUR_RESID_Q4 -2.262736e-06
ANTID_Q2     -8.048560e-08
ANTID_Q3     -6.180351e-08
ANTID_Q4     -2.469528e-07
DUR_TOT_Q2   -2.364842e-09
DUR_TOT_Q3   -2.181626e-07
DUR_TOT_Q4   -2.341661e-06
>
```

Contestualmente a questi bilanciamenti andiamo a verificare anche che il numero di riscatti previsto dal modello mod1 sul dataset "celle" (costruito aggregando il data-train secondo i fattori tariffari già visti) sia di fatto corrispondente al numero di riscatti effettivamente osservati. In linea con quanto visto per i bilanciamenti, il nostro modello mod1 verifica il bilanciamento (tra effettivi e previsti) dei riscatti sulle celle. Il numero previsto di riscatti nella cella (pred-n-risc) è stato ottenuto moltiplicando l'esposizione della cella per la probabilità di riscatto prevista dal modello sulla cella.

```
> round(sum(celle$pred_n_risc1),0) == sum(celle$riscatto2) # = 5184 tot riscatti data_train
[1] TRUE
>
```

A questo punto occorre però valutare la significatività di ciascun fattore tariffario nel suo complesso. Questa operazione viene svolta mediante l'apposito comando ANOVA, specificando tra i suoi parametri che desideriamo condurre un Likelihood Ratio Test sulla significatività dei vari fattori. Possiamo notare dall'output di seguito riportato che vi sono svariati fattori che apparentemente non forniscono un apporto significativo nella spiegazione del fenomeno dei riscatti. Si tratta di fattori quali: Sesso, Status di lavoratore autonomo, Riserva matematica, Durata residua e Durata totale. Va però ricordato che un diverso ordine delle variabili nella formula del modello potrebbe portare a risultati diversi del test ANOVA. A livello numerico, infatti, questo comando conduce dei Likelihood Ratio Test considerando i vari

fattori nell'ordine in cui sono specificati nella formula, di fatto imponendo quindi un ordine di inclusione in una collezione di modelli annidati. Si segnala comunque che questo comando è stato lanciato con varie possibili disposizioni dei regressori e i risultati ottenuti dall'ANOVA sono stati quasi sempre analoghi a quello qui riportato. Ad ogni modo, il fatto che il fattore tariffario Sesso non risulti significativo è abbastanza verosimile ed è anche in linea con una parte della letteratura. Tralasceremo quindi questo fattore dai successivi modelli. Non è invece altrettanto scontato che variabili quali Lavoratore autonomo e Riserva matematica siano poco significative: sono entrambe legate abbastanza strettamente con lo status socio-economico del policyholder e dunque di riflesso con la sua capacità di fronteggiare emergenze di liquidità. Almeno in un primo momento, lasceremo queste variabili nel modello. Per quanto riguarda la Riserva matematica abbiamo anche un'argomentazione a favore di questa scelta, che risiede nella suddivisione Lapse-by-RM-Q. In tale suddivisione si vede una certa oscillazione dei tassi di riscatto in funzione della fascia di RM di appartenenza. Venendo poi alle variabili legate alle durate, si decide in prima battuta di escludere dal modello solamente la variabile della Durata totale, poiché le voci di Durata residua e Antidurata di fatto apportano congiuntamente informazioni equivalenti.

```
> anova(mod1, test='LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(riscatto2, esposiz - riscatto2)

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     2497025      74403
SESSO       1     0.00   2497024      74403 0.9754815
FRAZ        5    21.77   2497019      74381 0.0005801 ***
LAV_AUTO    1     0.27   2497018      74381 0.6015819
ETA_Q       3    72.04   2497015      74309 1.559e-15 ***
RM_Q        3     3.13   2497012      74306 0.3721991
CUM_PREMI_Q 3     8.19   2497009      74298 0.0422163 *
DUR_RESID_Q 3     3.19   2497006      74295 0.3639701
ANTID_Q     3   371.05   2497003      73924 < 2.2e-16 ***
DUR_TOT_Q   3     0.93   2497000      73923 0.8191081
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Ci concentriamo ora sull'analisi dei residui. Viene in seguito riportato un grafico che mostra in ascissa la probabilità di riscatto prevista dal modello per ciascuna osservazione, mentre in ordinata si rilevano i valori dei residui di Pearson per ciascuna osservazione. In linea generale quindi tutti i punti che giacciono approssimativa-

mente sulle ascisse rappresentano delle osservazioni per le quali i residui di Pearson hanno valore molto ridotto e prossimo allo zero. Si ricorda inoltre che i residui di Pearson hanno la seguente forma:

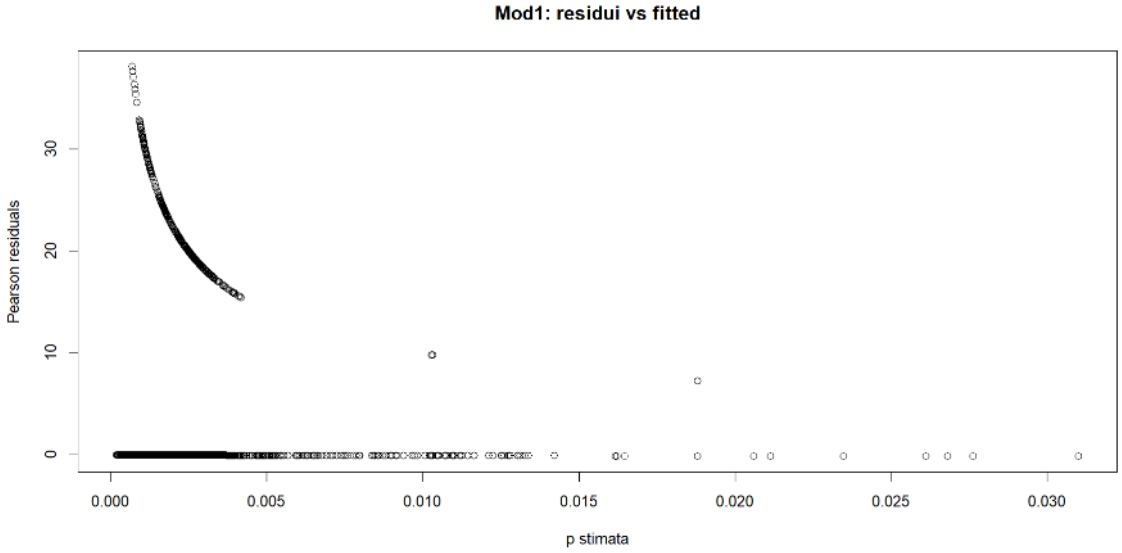
$$\frac{y_i - \hat{\mu}_i}{\sqrt{\nu(\hat{\mu}_i)/\omega_i}}.$$

Nella nostra situazione avremo dei residui di questo tipo

$$\frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)/1}},$$

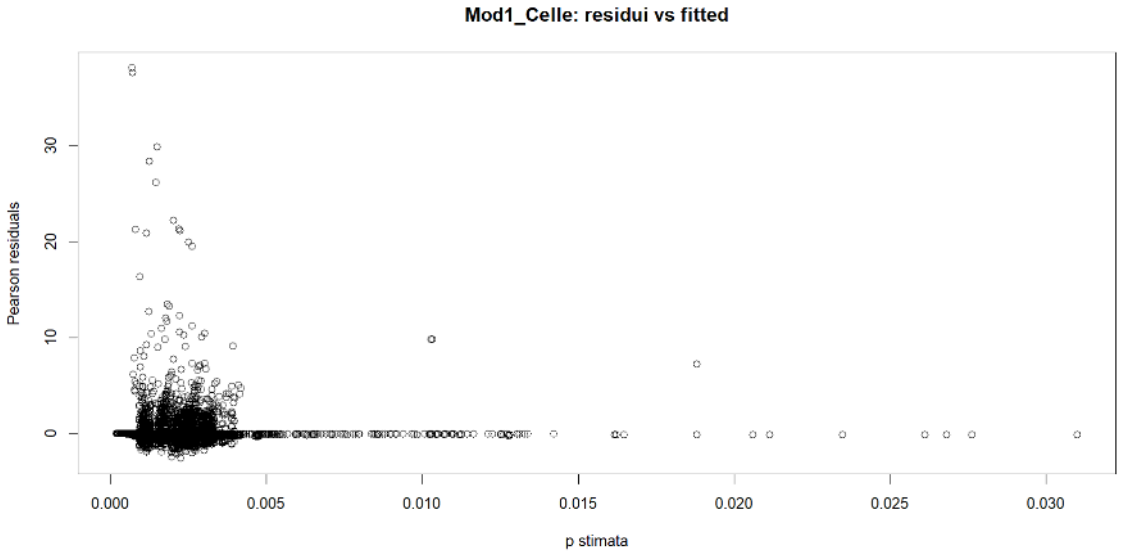
con $y_i = 0$ oppure $y_i = 1$ poiché $\omega_i = 1 \quad \forall i$.

Va considerato che tutte le $\hat{p}_i$ previste dal modello mod1 hanno valori molto ridotti (il loro massimo è circa 0.03). Si potrebbe quindi supporre che, al netto di eccezioni dovute alla scala data dal denominatore, molte osservazioni sull'asse delle ascisse hanno il valore di dummy riscatto fissato a 0. Con un ragionamento analogo notiamo che tutte le osservazioni della curva in alto nel grafico verosimilmente avranno la dummy riscatto $y_i = 1$, inoltre all'aumentare di $\hat{p}_i$ i residui di Pearson diminuiscono (con una velocità che non sembra lineare, per via del denominatore che contiene le $\hat{p}_i$). Un grafico di questo tipo non sembra quindi molto esplicativo perché risente del fatto che $\hat{p}_i \in (0;1)$ (intervallo continuo) mentre $y_i \in \{0;1\}$ (insieme di due elementi) poiché $\omega_i = 1 \quad \forall i$.



Abbiamo visto però che esiste una certa corrispondenza tra le variabili EDM sulle singole osservazioni e sulle osservazioni radunate in celle. Si riscontra, stimando lo stesso modello sul data-train e poi sul data-train aggregato in celle, che i coefficienti di massima verosimiglianza coincidono. Nel data-train aggregato in celle troviamo il

valore di y_i espresso come Key ratio (numero riscatti su esposizioni). In presenza di esposizioni maggiori di 1 (come avviene nella maggior parte delle celle), il Key ratio della cella j sarà un valore $y_j \in [0; 1]$ (va inteso come insieme dei numeri razionali tra 0 e 1, estremi inclusi). Questo permette di confrontare in maniera più agevole le stime ottenute con le osservazioni empiriche. Riportiamo quindi di seguito il grafico dei valori previsti e dei residui di Pearson, entrambi riferiti a un modello identico al mod1 ma stimato sulle celle.



Possiamo notare anche in questo caso una serie di punti allineati in corrispondenza dell'asse delle ascisse: tali punti corrispondono a delle celle nelle quali il Key ratio (proporzione di riscatti osservata) è molto vicino alla probabilità di riscatto prevista dal modello per la suddetta cella. In corrispondenza della nuvola di punti in basso a sinistra abbiamo le celle per le quali la probabilità prevista non supera lo 0.005. In queste celle i residui raggiungono anche valori importanti, con un massimo oltre i 30. Questi ampi scostamenti potrebbero essere dovuti a un'incapacità del modello di prevedere bene le probabilità in tali celle. Dei residui così importanti potrebbero essere dovuti anche a un "effetto scala" di $\hat{p}_i$: tanto più tale stima dista da 0.5, tanto minore sarà la $\nu(\hat{\mu}_i) = \hat{p}_i(1 - \hat{p}_i)$ al denominatore e quindi tanto maggiori saranno i residui di Pearson in valore assoluto. Questo meccanismo potrebbe anche spiegare perché i residui diminuiscano andando verso destra nel grafico. I residui con i valori più grandi potrebbero essere dovuti anche al fatto che le celle cui sono riferiti hanno un'esposizione prossima a 1. Di fatto torneremmo alla situazione vista in precedenza nel grafico dei residui del modello mod1 (stimato cioè sulle singole osservazioni). Tramite software è stato verificato che i valori più alti (in valore assoluto) dei residui sono riferiti prevalentemente a celle con esposizione unitaria e con livelli tariffari non

troppo omogenei. Di seguito riportiamo alcuni dati delle celle che presentano i 10 residui più grandi in valore assoluto.

```
> celle[match(sort(abs(residuals.glm(mod1_prova, type = 'pearson')), decreasing = TRUE)[1:10],
residuals.glm(mod1_prova, type = 'pearson')),1:15]
     period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q
779    22q3     F    1       SI     4    4           4           2         3       1
1813   21q3     F    2       NO     3    1           1           4         4       1
2626   22q2     M    1       NO     3    3           4           4         4       1
2163   22q4     F    4       NO     3    2           2           4         4       1
8870   21q3     F   12       NO     2    1           1           4         4       4
3684   21q4     M    6       NO     3    4           4           1         4       2
3832   20q4     M    4       SI     4    4           4           1         4       2
1187   21q2     M    4       NO     4    4           4           2         4       1
2544   21q4     F    6       NO     4    4           3           4         4       1
6833   21q4     F    6       NO     3    1           1           1         4       4
     riscatto2 esposiz Key_ratio   pred_prob1 pred_n_risc1
779          1       1 1.0000000 0.0006879191 0.0006879191
1813         1       1 1.0000000 0.0007057208 0.0007057208
2626         2       3 0.6666667 0.0014893292 0.0044679877
2163         1       1 1.0000000 0.0012389415 0.0012389415
8870         1       1 1.0000000 0.0014589345 0.0014589345
3684         1       1 1.0000000 0.0020274050 0.0020274050
3832         1       1 1.0000000 0.0021966980 0.0021966980
1187         2      11 0.1818182 0.0007977508 0.0087752583
2544         1       1 1.0000000 0.0022289980 0.0022289980
6833         1       2 0.5000000 0.0011452386 0.0022904772
> |
```

Si nota in linea generale una certa prevalenza dei livelli più alti per fattori come Età, Riserva, Cumulo premi, Durata residua. Per l'Antidurata invece prevale il primo livello. Va sottolineata però la scarsa rilevanza di questi risultati, poiché si tratta per la maggior parte di celle a esposizione unitaria con Key ratio pari a 1. Si riscontra quindi una certa difficoltà a interpretare il grafico dei residui, anche sfruttando l'aggregazione in celle.

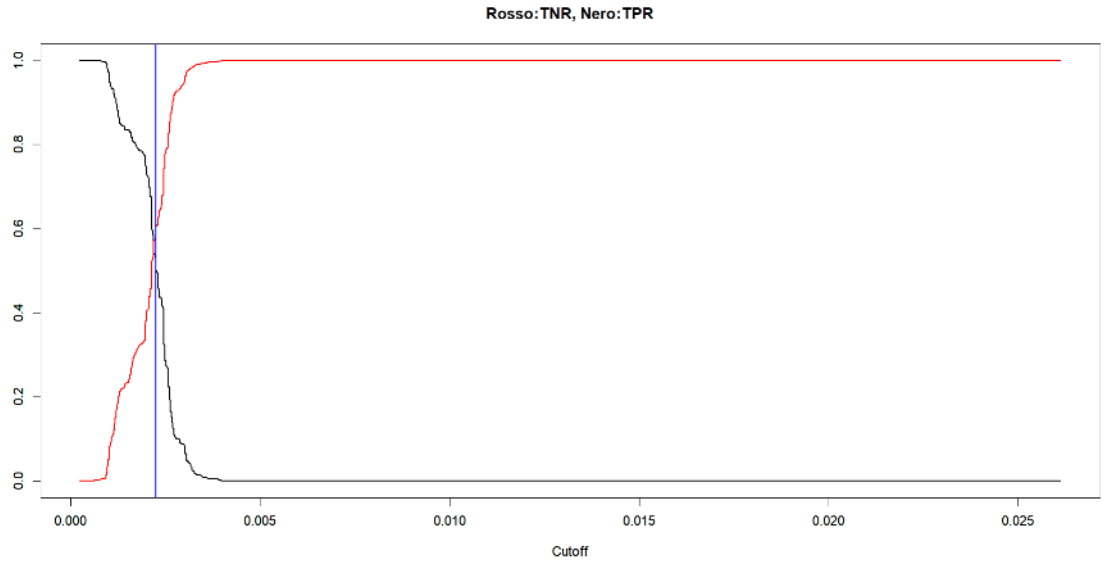
Selezioniamo quindi le 50 celle cui corrispondono i residui di Pearson maggiori in valore assoluto, da queste filtriamo solo quelle che hanno un'esposizione sufficientemente grande o comunque maggiore di una certa soglia (la soglia di esposizione è stata qui individuata arbitrariamente nel valore di 10).

```
> celle[match(sort(abs(residuals.glm(mod1_prova, type = 'pearson')), decreasing = TRUE)[1:50], residuals.g
lm(mod1_prova, type = 'pearson')),1:15] %>% filter(esposiz > 10)
   period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q riscatto2 esposiz
1    21q2     M    4       NO     4    4           4           2         4       1         2      11
2    22q3     M    4       NO     2    4           4           2         4       4         3      34
3    21q2     F    1       NO     3    1           1           4         4       4         3      44
4    21q1     M    6       SI     4    4           4           4         4       2         2      13
5    22q1     F    2       NO     3    4           4           1         4       4         5     135
6    20q4     F    1       NO     4    4           4           2         4       1         2      54
7    20q4     F    4       SI     3    4           4           4         4       1         1      14
8    21q2     M   12       NO     4    4           4           1         4       1         1      21
9    21q1     M    2       NO     2    4           4           4         4       3         2      26
10   22q1     F    1       NO     4    4           4           3         4       1         2      81
11   21q3     F    4       SI     3    4           4           4         4       4         3      79
12   22q4     M   12       NO     4    4           4           3         4       2         1      12
13   21q4     F    1       NO     3    4           4           1         1       1         1      35
14   21q2     F   12       NO     4    4           4           2         4       3         1      13
15   22q1     M   12       NO     4    4           4           2         4       3         1      14
    Key_ratio   pred_prob1 pred_n_risc1
1  0.18181818 0.0007977508  0.008775258
2  0.08823529 0.0017731827  0.060288212
3  0.06818182 0.0016395567  0.072140496
4  0.15384615 0.0029293241  0.038081214
5  0.03703704 0.0017564372  0.237119022
6  0.03703704 0.0009518373  0.051399213
7  0.07142857 0.0010738769  0.015034277
8  0.04761905 0.0007558158  0.015872131
9  0.07692308 0.0028657948  0.074510665
10 0.02469136 0.0009619079  0.077914539
11 0.03797468 0.0022774210  0.179916256
12 0.08333333 0.0019648570  0.023578284
13 0.02857143 0.0007264712  0.025426493
14 0.07692308 0.0019618394  0.025503912
15 0.07142857 0.0019258380  0.026961732
```

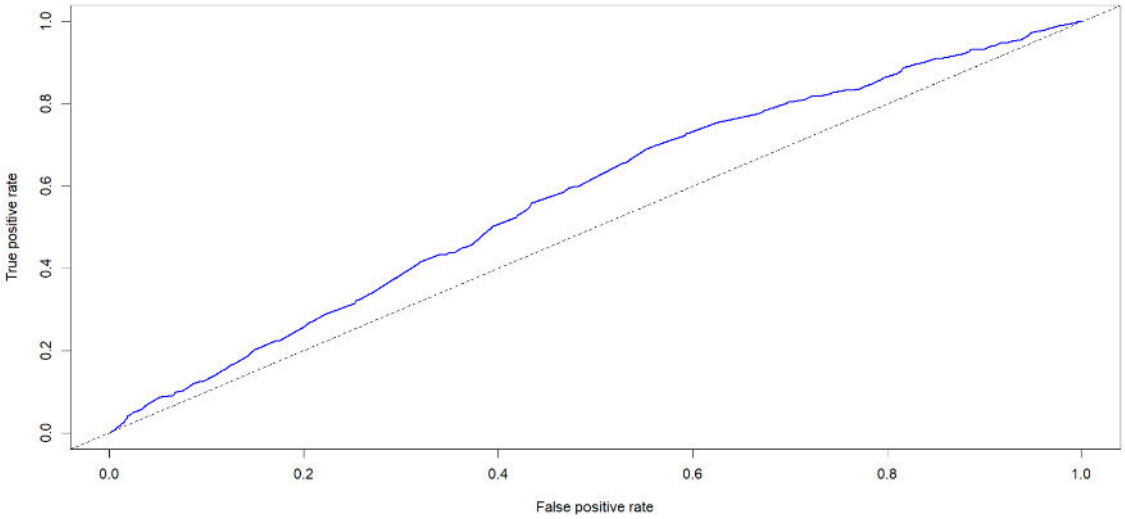
Si è verificato tramite software che i residui che risultano maggiori in valore assoluto sono tutti di segno positivo. L'output sopra riportato segnala quindi dei casi di sottostima della probabilità di riscatto. Si nota una forte prevalenza del quarto livello per la variabile Riserva matematica, per la variabile Cumulo premi e per la variabile Durata totale. Si potrebbe forse inferire che il modello mod1

stia sottostimando le probabilità di riscatto in corrispondenza di contratti a lunga durata, con alti valori di cumulo premi e di riserve. Questo rappresenta un rischio importante per la compagnia, poiché sta sottostimando le probabilità di riscatto proprio per quei contratti che, se riscattati, richiedono di liberare maggior liquidità.

Concludiamo la trattazione del modello mod1 valutandone le performance previsive sul dataset di test. Il data-train ha una distribuzione sbilanciata della variabile dummy riscatto (5184 riscatti su oltre 2 milioni di polizze-trimestre); pertanto si è reso necessario calibrare la soglia che permettesse di ottenere il punto più a nord-ovest della curva ROC. Riportiamo il grafico che mostra la sensitivity (il TPR, in nero) e la specificity (il TNR, in rosso) associate a diversi livelli di threshold. La retta verticale blu rappresenta la soglia calibrata. Si verifica visivamente che con tale soglia riusciamo a massimizzare TPR e FPR contemporaneamente, passando per la loro intersezione.



Seguono poi la curva ROC e la confusion matrix basata sulla threshold (calibrata come illustrato nel capitolo dedicato).




```
> confusionMatrix(pred_labels_mod1, factor(data_test$riscatto2), positive = '1')
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 178123    307
         1 128798    343

               Accuracy : 0.5802
                 95% CI : (0.5785, 0.582)
    No Information Rate : 0.9979
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0011

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.527692
            Specificity : 0.580355
         Pos Pred Value : 0.002656
         Neg Pred Value : 0.998279
             Prevalence : 0.002113
         Detection Rate : 0.001115
   Detection Prevalence : 0.419874
      Balanced Accuracy : 0.554023

       'Positive' Class : 1

>
```

Segnaliamo innanzitutto l'accuracy del nostro modello, che risulta maggiore di 0,5 ma comunque sicuramente non elevata. Lo stesso si può affermare per la specificity e la sensitivity. In linea generale la matrice evidenzia che il problema principale del nostro modello è un'eccessiva tendenza a prevedere dei positivi: i False Positive sono dello stesso ordine di grandezza dei True Negative.

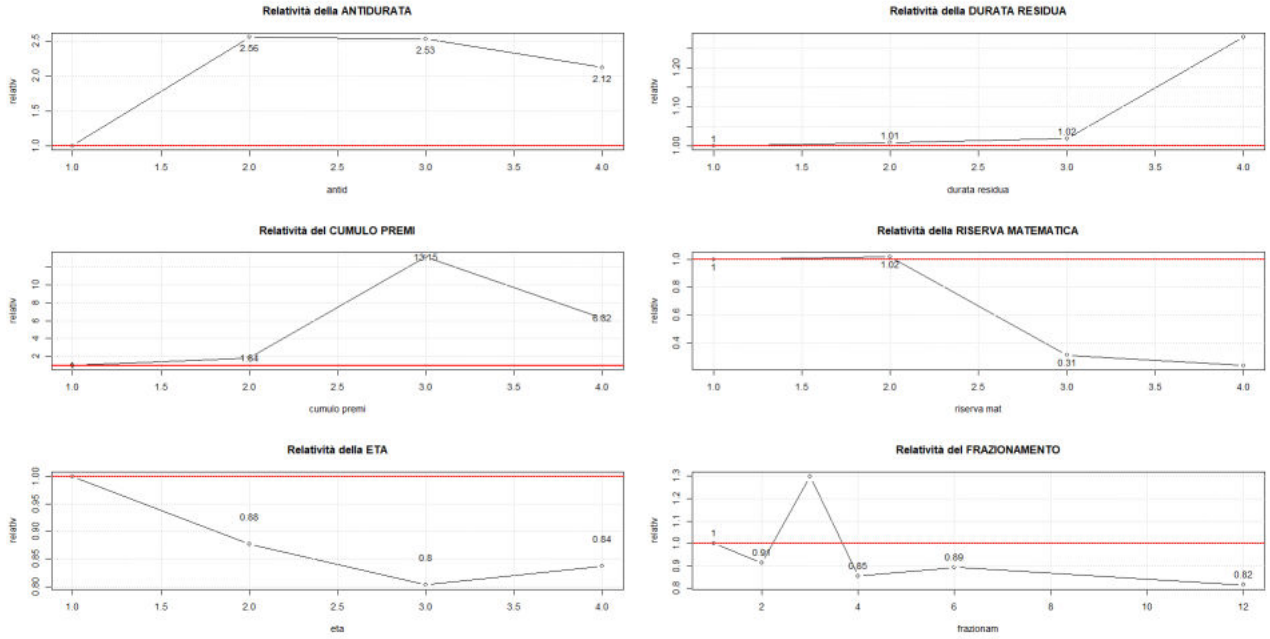
5.3.2 Modello 2

Si decide di procedere alla stima del modello mod2 basandosi sulle considerazioni formulate in sede di commento ai risultati del comando ANOVA per il modello mod1. In questa seconda versione infatti verranno tolte le variabili Sesso e Durata totale.

Il modello viene naturalmente stimato anche in questo caso sui dati del data-train. Si riportano di seguito i coefficienti stimati.

```
Call:
glm(formula = cbind(riscatto2, esposiz - riscatto2) ~ FRAZ +
    LAV_AUTO + ETA_Q + RM_Q + CUM_PREMI_Q + DUR_RESID_Q + ANTID_Q,
    family = binomial(link = "logit"), data = data_train, x = TRUE)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.203412   0.256192 -28.117  < 2e-16 ***
FRAZ2        -0.089549   0.084229  -1.063 0.287710
FRAZ3         0.261898   0.119472   2.192 0.028370 *
FRAZ4        -0.157386   0.043960  -3.580 0.000343 ***
FRAZ6        -0.112912   0.058614  -1.926 0.054059 .
FRAZ12       -0.204207   0.035519  -5.749 8.97e-09 ***
LAV_AUTOSI    0.081900   0.039977   2.049 0.040494 *
ETA_Q2       -0.130817   0.037598  -3.479 0.000503 ***
ETA_Q3       -0.219281   0.038825  -5.648 1.62e-08 ***
ETA_Q4       -0.177723   0.045962  -3.867 0.000110 ***
RM_Q2         0.018919   0.788036   0.024 0.980846
RM_Q3        -1.170177   0.966698  -1.210 0.226091
RM_Q4        -1.438784   0.697799  -2.062 0.039218 *
CUM_PREMI_Q2  0.610714   1.156755   0.528 0.597531
CUM_PREMI_Q3  2.576127   0.780148   3.302 0.000960 ***
CUM_PREMI_Q4  1.844081   0.719066   2.565 0.010331 *
DUR_RESID_Q2  0.007978   0.133772   0.060 0.952444
DUR_RESID_Q3  0.018572   0.125249   0.148 0.882119
DUR_RESID_Q4  0.245369   0.057953   4.234 2.30e-05 ***
ANTID_Q2      0.938873   0.074313  12.634  < 2e-16 ***
ANTID_Q3      0.928831   0.072689  12.778  < 2e-16 ***
ANTID_Q4      0.752946   0.045699  16.476  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Si può subito notare che le stime dei coefficienti non variano sostanzialmente e anzi rimangono pressoché ferme sugli stessi valori. Lo stesso si può dire a grandi linee delle deviazioni standard di tali stime e dunque anche della significatività dei coefficienti stessi. Per questo motivo i coefficienti di questo modello non verranno commentati, per una migliore illustrazione degli stessi si rimanda al commento dei coefficienti del modello mod1.
A titolo riassuntivo dei coefficienti si riportano i loro esponenziali (le relatività), ovvero i fattori di scala per cui moltiplicare l'odds ratio quando si passa dal livello di riferimento a un altro livello.



Riportiamo poi la devianza del modello mod2.

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74403  on 2497025  degrees of freedom
Residual deviance: 73924  on 2497004  degrees of freedom
AIC: 73968

Number of Fisher Scoring iterations: 9
```

Anche in questo caso conduciamo una verifica empirica della convergenza della soluzione numerica a quella teorica delle equazioni di verosimiglianza. Dall'output di seguito riportato possiamo concludere che, a meno di errori numerici trascurabili, la soluzione ha trovato un punto abbastanza vicino al punto di massimo della verosimiglianza.

```
> t(X2) %*% (data_train$riscatto2 - p_i_hat2) # e-07 praticamente vettore di zeri
                    [,1]
(Intercept)  -1.948391e-06
FRAZ2        -6.469890e-08
FRAZ3        -4.561061e-08
FRAZ4        -2.576583e-07
FRAZ6        -1.702751e-07
FRAZ12       -7.497644e-07
LAV_AUTOSI   -3.807638e-07
ETA_Q2       -3.227698e-07
ETA_Q3       -3.811659e-07
ETA_Q4       -6.960885e-07
RM_Q2        -3.220417e-09
RM_Q3        -2.161897e-09
RM_Q4        -1.927443e-06
CUM_PREMI_Q2 -2.382590e-09
CUM_PREMI_Q3 -4.552024e-09
CUM_PREMI_Q4 -1.926880e-06
DUR_RESID_Q2 -1.493017e-08
DUR_RESID_Q3 -2.330541e-08
DUR_RESID_Q4 -1.898414e-06
ANTID_Q2      1.072628e-08
ANTID_Q3      6.607309e-09
ANTID_Q4      9.913058e-08
```

Controlliamo inoltre che vi sia bilanciamento tra riscatti effettivi e previsti a livello di celle.

```
> round(sum(celle$pred_n_risc2),0) == sum(celle$riscatto2) # = 5184 tot riscatti data_train
[1] TRUE
> |
```

Proseguiamo poi con la valutazione della significatività di interi fattori tariffari, sempre con il comando ANOVA. Si riporta il relativo output.

```
> anova(mod2, test='LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(riscatto2, esposiz - riscatto2)

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       2497025      74403
FRAZ          5    21.76   2497020      74381 0.0005807 ***
LAV_AUTO      1     0.27   2497019      74381 0.6033273
ETA_Q         3    72.03   2497016      74309 1.572e-15 ***
RM_Q          3     3.13   2497013      74306 0.3720656
CUM_PREMI_Q   3     8.19   2497010      74298 0.0422099 *
DUR_RESID_Q   3     3.17   2497007      74295 0.3662781
ANTID_Q       3   370.66   2497004      73924 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Possiamo notare che la significatività globale dei fattori ricalca abbastanza fedelmente quella già ottenuta quando è stato lanciato il comando ANOVA sul modello

mod1. Come già anticipato, il risultato più controintuitivo è sicuramente il fatto che la riserva matematica sembri non essere significativa. Si è deciso quindi di ripetere il comando ANOVA su un modello analogo al mod2, con l'unica differenza che risiede nell'ordine dei regressori. Pertanto si è deciso di specificare la riserva matematica come primo regressore, affinché il comando ANOVA ne vada a testare la significatività rispetto al modello nullo, isolando così l'apporto informativo del fattore in questione. Riportiamo di seguito l'output, segnalando che la Riserva matematica non sembra avere un apporto significativo rispetto al modello nullo.

```
> anova(mod2_prova1, test = 'LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(riscatto2, esposiz - riscatto2)

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      2497025      74403
RM_Q         3     1.99   2497022      74401 0.5754807
FRAZ         5    22.34   2497017      74379 0.0004515 ***
LAV_AUTO     1     0.23   2497016      74379 0.6308134
ETA_Q        3    72.63   2497013      74306 1.164e-15 ***
CUM_PREMI_Q  3     8.19   2497010      74298 0.0422099 *
DUR_RESID_Q  3     3.17   2497007      74295 0.3662781
ANTID_Q      3   370.66   2497004      73924 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Analoghe considerazioni valgono per il fattore Lavoratore autonomo.

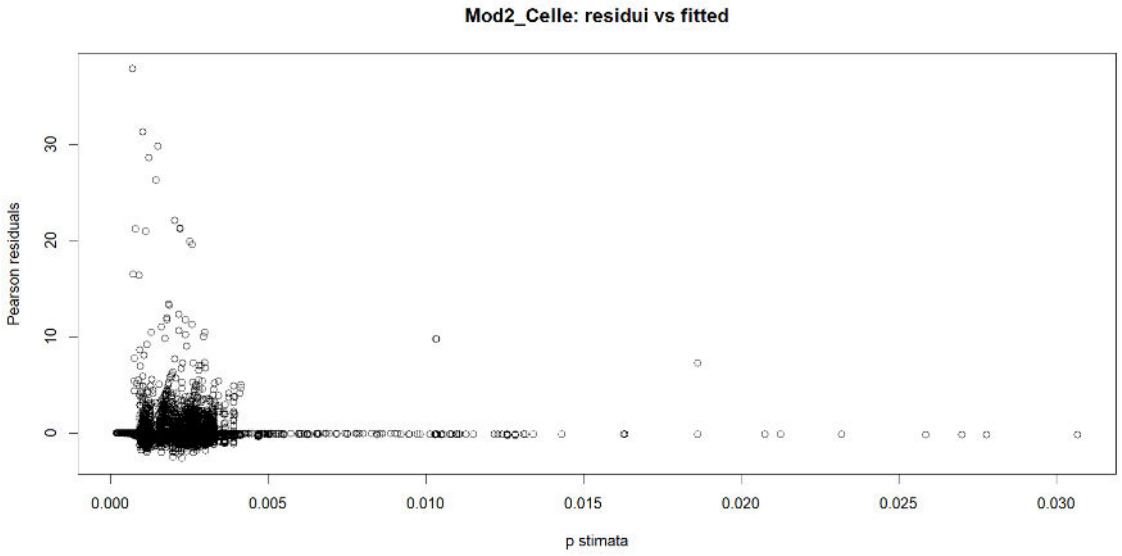
```
> anova(mod2_prova2, test = 'LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(riscatto2, esposiz - riscatto2)

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      2497025      74403
LAV_AUTO     1     0.16   2497024      74403 0.6924730
RM_Q         3     1.97   2497021      74401 0.5785639
FRAZ         5    22.43   2497016      74379 0.0004342 ***
ETA_Q        3    72.63   2497013      74306 1.164e-15 ***
CUM_PREMI_Q  3     8.19   2497010      74298 0.0422099 *
DUR_RESID_Q  3     3.17   2497007      74295 0.3662781
ANTID_Q      3   370.66   2497004      73924 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Diverso invece è il caso del fattore Durata residua (si veda l'output riportato a fine paragrafo). L'apporto di questo fattore è decisamente significativo se preso singolarmente. Con questo ordinamento dei fattori stiamo anteponendo la Durata residua ad altri fattori che contengono delle informazioni simili. Si nota ad esempio una aumento importante del p-value, e quindi una "riduzione di significatività", della variabile Età. Questa evidenza sembra anche abbastanza ragionevole: ci si aspetta una relazione inversa abbastanza robusta tra l'Età alla data e la durata residua della polizza, motivo per cui si ritiene che le due variabili possano contenere informazioni simili. Considerando una singola polizza ripetuta su più trimestri, infatti, è evidente come l'età aumenti ogni volta di un trimestre e la durata residua si riduca dello stesso ammontare.

```
> anova(mod2_prova3, test = 'LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(riscatto2, esposiz - riscatto2)

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      2497025      74403
DUR_RESID_Q  3    23.07   2497022      74380 3.900e-05 ***
LAV_AUTO     1     0.45   2497021      74380 0.5031838
RM_Q         3     1.92   2497018      74378 0.5888482
FRAZ         5    22.95   2497013      74355 0.0003444 ***
ETA_Q        3    51.99   2497010      74303 3.014e-11 ***
CUM_PREMI_Q  3     8.17   2497007      74295 0.0427177 *
ANTID_Q      3   370.66   2497004      73924 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esiste tuttavia un metodo più strutturato per selezionare le variabili, procedendo iterativamente alla stima di modelli da cui vengono aggiunti o tolti dei singoli regressori in ciascuna iterazione, scegliendo le variabili sulla base del Likelihood Ratio Test. Il modello successivo mod3 verrà pertanto stimato con questo procedimento automatizzato.

Passiamo poi alla fase di analisi dei residui del modello mod2. Viene riportato un grafico che mostra i valori previsti e i residui di Pearson di un modello analogo al mod2 ma stimato sulle celle (il motivo per cui non si riportano i valori riferiti al modello mod2 stimato sulle osservazioni è stato già discusso in sede di commento dei residui del mod1).



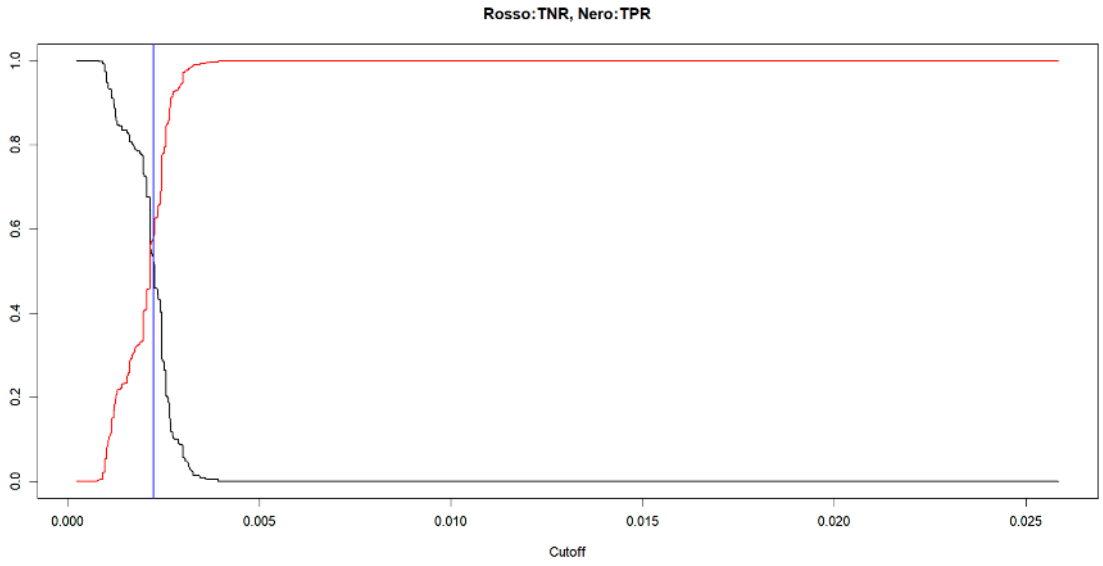
Il grafico presenta una nuvola di punti analoga a quella vista per il modello mod1. Si rimanda pertanto al commento dello stesso, segnalando però che rispetto al mod1 è cambiata la cella cui corrisponde il massimo residuo di Pearson: questa presenta dei livelli diversi su tutti i fattori tariffari, ma ha anch'essa esposizione unitaria. Riportiamo le caratteristiche delle celle che presentano i 10 residui più grandi in valore assoluto, segnalando che sono praticamente le stesse già viste per il modello mod1, seppur ordinate in maniera leggermente diversa.

```
> celle[match(sort(abs(residuals.glm(mod2_prova, type = 'pearson')), decreasing = TRUE)[1:10],
residuals.glm(mod2_prova, type = 'pearson')),c(1:13,16,17)]
     period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q
1813   21q3     F    2       NO     3    1           1           4         4       1
779    22q3     F    1       SI     4    4           4           2         3       1
2626   22q2     M    1       NO     3    3           4           4         4       1
2163   22q4     F    4       NO     3    2           2           4         4       1
8870   21q3     F   12       NO     2    1           1           4         4       4
3684   21q4     M    6       NO     3    4           4           1         4       2
3832   20q4     M    4       SI     4    4           4           1         4       2
2544   21q4     F    6       NO     4    4           3           4         4       1
1187   21q2     M    4       NO     4    4           4           2         4       1
6833   21q4     F    6       NO     3    1           1           1         4       4
     riscatto2 esposiz Key_ratio   pred_prob2 pred_n_risc2
1813         1       1 1.0000000 0.0006978050  0.000697805
779          1       1 1.0000000 0.0010209913  0.001020991
2626         2       3 0.6666667 0.0014960640  0.004488192
2163         1       1 1.0000000 0.0012231797  0.001223180
8870         1       1 1.0000000 0.0014422299  0.001442230
3684         1       1 1.0000000 0.0020427064  0.002042706
3832         1       1 1.0000000 0.0022102183  0.002210218
2544         1       1 1.0000000 0.0022126818  0.002212682
1187         2      11 0.1818182 0.0008038883  0.008842771
6833         1       2 0.5000000 0.0011319749  0.002263950
>
```

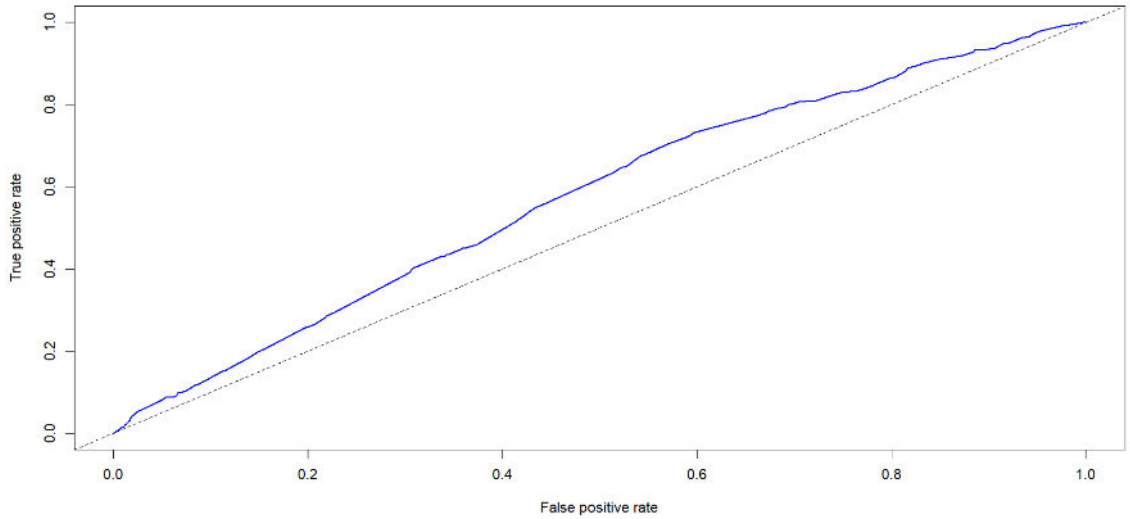
Anche in questo caso andiamo a selezionare le celle cui corrispondono i 50 residui maggiori in valore assoluto, filtrando poi quelle con esposizione maggiore di 10. I risultati dell'output sotto riportato sono analoghi a quelli del modello mod1, si rimanda pertanto al commento dello stesso.

```
> celle[match(sort(abs(residuals.glm(mod2_prova, type = 'pearson')), decreasing = TRUE)[1:50], residuals.g
lm(mod2_prova, type = 'pearson')),1:15] %>% filter(esposiz > 10)
   period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q riscatto2 esposiz
1    21q2     M    4       NO     4    4           4           2         4       1         2      11
2    22q3     M    4       NO     2    4           4           2         4       4         3      34
3    21q2     F    1       NO     3    1           1           4         4       4         3      44
4    21q1     M    6       SI     4    4           4           4         4       2         2      13
5    22q1     F    2       NO     3    4           4           1         4       4         5     135
6    20q4     F    1       NO     4    4           4           2         4       1         2      54
7    20q4     F    4       SI     3    4           4           4         4       1         1      14
8    21q2     M   12       NO     4    4           4           1         4       1         1      21
9    21q1     M    2       NO     2    4           4           4         4       3         2      26
10   22q1     F    1       NO     4    4           4           3         4       1         2      81
11   21q3     F    4       SI     3    4           4           4         4       4         3      79
12   22q4     M   12       NO     4    4           4           3         4       2         1      12
13   21q2     F   12       NO     4    4           4           2         4       3         1      13
14   22q1     M   12       NO     4    4           4           2         4       3         1      14
15   22q3     F    1       SI     4    4           4           2         4       1         1      27
    Key_ratio   pred_prob1 pred_n_risc1
1  0.18181818 0.0007977508  0.008775258
2  0.08823529 0.0017731827  0.060288212
3  0.06818182 0.0016395567  0.072140496
4  0.15384615 0.0029293241  0.038081214
5  0.03703704 0.0017564372  0.237119022
6  0.03703704 0.0009518373  0.051399213
7  0.07142857 0.0010738769  0.015034277
8  0.04761905 0.0007558158  0.015872131
9  0.07692308 0.0028657948  0.074510665
10 0.02469136 0.0009619079  0.077914539
11 0.03797468 0.0022774210  0.179916256
12 0.08333333 0.0019648570  0.023578284
13 0.07692308 0.0019618394  0.025503912
14 0.07142857 0.0019258380  0.026961732
15 0.03703704 0.0010338349  0.027913541
>
```

Proseguiamo il commento al modello mod2 descrivendone le metriche di previsione. Riportiamo il grafico di TPR e TNR in relazione alle varie soglie e notiamo che la soglia calibrata è molto vicina a quella calibrata per il modello 1, ovvero circa il 2 per mille.



Seguono poi la curva ROC e la confusion matrix basata sulla threshold (calibrata come illustrato nel capitolo dedicato).




```
> confusionMatrix(pred_labels_mod2, factor(data_test$riscatto2), positive = '1')
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 102871    146
         1 204050    504

               Accuracy : 0.3361
                 95% CI : (0.3344, 0.3378)
    No Information Rate : 0.9979
    P-Value [Acc > NIR] : 1

                  Kappa : 7e-04

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.775385
            Specificity : 0.335171
         Pos Pred Value : 0.002464
         Neg Pred Value : 0.998583
             Prevalence : 0.002113
         Detection Rate : 0.001639
   Detection Prevalence : 0.665063
      Balanced Accuracy : 0.555278

       'Positive' Class : 1

>
```

L'accuracy è decisamente troppo bassa per poter ipotizzare un impiego del modello mod2 in ambito previsionale. L'errore del nostro modello anche in questo caso è una eccessiva tendenza a prevedere come positive le osservazioni. Si noti che i falsi positivi superano i veri negativi e che la specificity (TNR) arriva poco sopra il 30 per cento (dunque il FPR arriva poco sotto il 70 per cento).

5.3.3 Modello 3

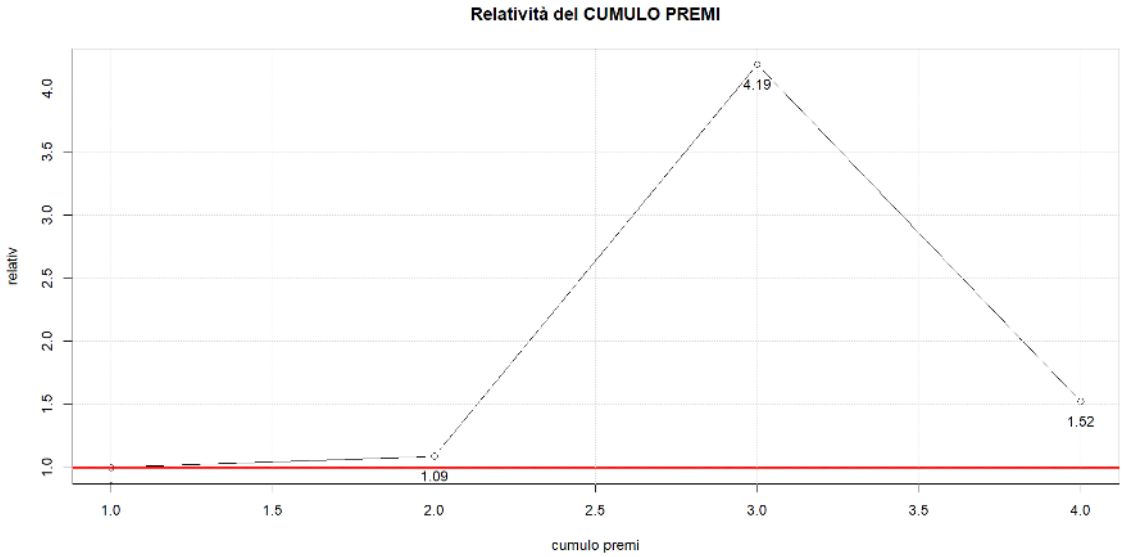
Veniamo infine al modello ricavato selezionando le variabili mediante degli appositi algoritmi di tipo stepwise. Si è deciso di partire dal modello mod1 con tutte le variabili, e su questo applicare un algoritmo stepwise con direzione "backward/forward".

Questo significa che l'algoritmo inizialmente elimina una variabile dal modello con tutte le variabili. La scelta della variabile da eliminare viene operata sulla base del criterio di informazione di Akaike. Sarebbe possibile specificare in alternativa il criterio BIC, ma tramite software si è notato che tale criterio porterebbe alla selezione delle sole variabili di età e antidurata (età della polizza), ottenendo così un modello troppo parsimonioso. Il primo passo di eliminazione della variabile è il cosiddetto passo backward, cui segue un passo forward in cui l'algoritmo aggiunge una variabile, sempre sulla base del criterio di informazione di Akaike. L'algoritmo procede quindi iterativamente in questo modo fin quando non può più ridurre il valore del criterio di informazione. Il modello mod3 così ricavato contiene tutte le variabili del modello mod2, ad eccezione della variabile Riserva matematica RM-Q.
Riportiamo di seguito i coefficienti stimati.

```
glm(formula = cbind(riscatto2, esposiz - riscatto2) ~ FRAZ +
    LAV_AUTO + ETA_Q + CUM_PREMI_Q + DUR_RESID_Q + ANTID_Q, family = binomial(link = "logit"),
    data = data_train, x = TRUE)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.216996   0.254162 -28.395  < 2e-16 ***
FRAZ2       -0.089831   0.084229  -1.067 0.286192
FRAZ3        0.263246   0.119466   2.204 0.027558 *
FRAZ4       -0.158171   0.043959  -3.598 0.000320 ***
FRAZ6       -0.113364   0.058617  -1.934 0.053113 .
FRAZ12      -0.204803   0.035519  -5.766 8.12e-09 ***
LAV_AUTOSI   0.081581   0.039976   2.041 0.041278 *
ETA_Q2      -0.131183   0.037597  -3.489 0.000484 ***
ETA_Q3      -0.219802   0.038823  -5.662 1.50e-08 ***
ETA_Q4      -0.178501   0.045958  -3.884 0.000103 ***
CUM_PREMI_Q2  0.085307   1.029906   0.083 0.933987
CUM_PREMI_Q3  1.432843   0.510189   2.808 0.004978 **
CUM_PREMI_Q4  0.420713   0.244276   1.722 0.085018 .
DUR_RESID_Q2  0.007816   0.133772   0.058 0.953406
DUR_RESID_Q3  0.018506   0.125250   0.148 0.882540
DUR_RESID_Q4  0.245307   0.057952   4.233 2.31e-05 ***
ANTID_Q2      0.938115   0.074307  12.625  < 2e-16 ***
ANTID_Q3      0.928226   0.072685  12.770  < 2e-16 ***
ANTID_Q4      0.752316   0.045691  16.465  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si nota subito che solo i coefficienti legati al fattore Cumulo premi sono cambiati in maniera significativa. Tutti gli altri coefficienti infatti sono rimasti quasi invariati, a meno di piccole variazioni nell'ordine di grandezza di 10^{-4} che non modificano però il segno o il livello di significatività dei coefficienti stessi. Per il fattore Cumulo premi, invece, vi sono importanti differenze nei coefficienti tra mod2 e mod3. Come si evince anche dal grafico delle relatività riportato in precedenza, per il mod2 la seconda fascia di cumulo premi aveva una relatività (riferita agli odds di riscatto) di poco superiore a 1, mentre tale relatività per la terza e quarta fascia raggiungeva valori in un intorno di 13 e di 6 rispettivamente. Quindi, generalizzando questi risultati, secondo

il modello mod2 i contraenti con i cumuli premi più alti sono anche quelli che riscattano di più. Nel modello mod3 non cambia la struttura delle relatività (intesa come forma della curva delle relatività), ma si riduce significativamente la magnitudo delle relatività associate alle due fasce più alte di cumulo premi. Si evidenzia che nel modello mod3 è stata rimossa la variabile della Riserva matematica, che verosimilmente conteneva delle informazioni in stretta relazione con il cumulo premi (tanti più premi ha conferito il contraente, tanto maggiore sarà la riserva matematica a parità di rendimento del fondo). Le relatività del modello mod2 associate alla terza e quarta fascia di Riserva matematica erano inferiori a 1 (quindi i coefficienti erano negativi): togliendo il fattore Riserva matematica dal modello è verosimile che il fattore Cumulo premi abbia incorporato questo effetto, sotto la ragionevole ipotesi che ci sia una buona corrispondenza tra fascia di Riserva matematica e fascia di Cumulo premi. Si riporta di seguito il grafico delle relatività del modello 3 per il fattore Cumulo premi. Si conclude quindi sottolineando che l'impatto della rimozione del fattore Riserva matematica sia stato quasi interamente assorbito dal fattore Cumulo premi.



Proseguiamo il commento del modello mod3 riportandone la devianza e i bilanciamenti delle equazioni di massima verosimiglianza.

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74403  on 2497025  degrees of freedom
Residual deviance: 73929  on 2497007  degrees of freedom
AIC: 73967

Number of Fisher Scoring iterations: 9
```

```
> t(X3) %*% (data_train$riscatto2 - p_i_hat3) # e-07 praticamente vettore di zeri
                        [,1]
(Intercept)  -1.776455e-06
FRAZ2        -5.942531e-08
FRAZ3        -4.302080e-08
FRAZ4        -2.283843e-07
FRAZ6        -1.570564e-07
FRAZ12       -6.767314e-07
LAV_AUTOSI   -3.538231e-07
ETA_Q2       -2.814356e-07
ETA_Q3       -3.390153e-07
ETA_Q4       -6.605393e-07
CUM_PREMI_Q2 -1.471272e-09
CUM_PREMI_Q3 -4.567405e-09
CUM_PREMI_Q4 -1.756960e-06
DUR_RESID_Q2 -1.279172e-08
DUR_RESID_Q3 -2.093880e-08
DUR_RESID_Q4 -1.743529e-06
ANTID_Q2      1.347047e-08
ANTID_Q3      1.499566e-08
ANTID_Q4      2.362780e-07
>
```

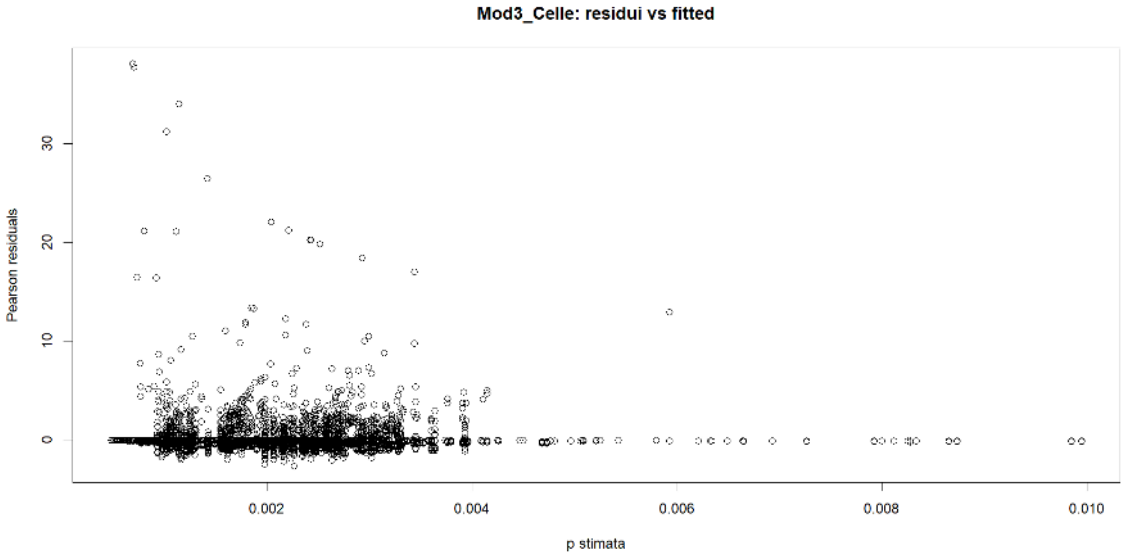
Si può dunque affermare che anche il modello mod3 abbia avuto una convergenza verso una soluzione delle equazioni di verosimiglianza. Controlliamo anche in questo caso che i riscatti effettivi e quelli previsti risultino bilanciati anche nell'aggregato quando si stima il modello a livello di celle.

```
> round(sum(celle$pred_n_risc3),0) == sum(celle$riscatto2) # = 5184 tot riscatti data_train
[1] TRUE
```

Arriviamo poi alla valutazione della significatività dei vari fattori tariffari. Riportiamo a fine paragrafo l'output del comando ANOVA. Notiamo che la significatività globale dei fattori rimane pressoché la stessa rispetto a quanto visto nella prima impostazione del comando ANOVA per il modello mod2: sembrano essere significativi i fattori Frazionamento, Età e Antidurata (età della polizza). Si è invece ridotta la significatività del fattore cumulo premi. Questo potrebbe essere dovuto allo stretto legame che verosimilmente intercorre tra Riserva matematica e Cumulo premi. Per quanto riguarda il fattore Durata residua, si rimanda al commento al mod2 per quanto riguarda la spiegazione sul perché tale fattore sia in realtà significativo. Per quanto riguarda il fattore Lavoratore autonomo, questo sembra non essere significativo ma viene comunque lasciato nel modello per due motivi: in primo luogo un'indicazione simile dovrebbe essere strettamente correlata alla situazione di liquidità dell'individuo (la quale riveste un ruolo decisivo secondo la Emergency Fund Hypothesis), in secondo luogo questo fattore è stato selezionato dalla procedura stepwise e quindi l'averlo incluso ha portato a una minore penalizzazione in termini di criterio informativo di Akaike.

```
> anova(mod3, test='LRT')
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(riscatto2, esposiz - riscatto2)

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     2497025      74403
FRAZ        5    21.76   2497020      74381 0.0005807 ***
LAV_AUTO    1     0.27   2497019      74381 0.6033273
ETA_Q       3    72.03   2497016      74309 1.572e-15 ***
CUM_PREMI_Q 3     6.82   2497013      74302 0.0780189 .
DUR_RESID_Q 3     3.17   2497010      74299 0.3655461
ANTID_Q     3   370.16   2497007      73929 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Ci concentriamo ora sull'analisi dei residui del modello mod3. Anche in questo caso facciamo riferimento a un modello analogo al modello mod3, ma stimato sulle celle. Otteniamo il seguente grafico.



Come prima cosa notiamo che rispetto al modello mod2 le probabilità di riscatto stimate sulle varie celle sono generalmente più piccole. Si potrebbe ipotizzare che la rimozione del fattore di Riserva matematica abbia portato a una riduzione generalizzata dei valori stimati. Si è già discusso infatti di come la rimozione di questo fattore abbia fatto diminuire la magnitudo dei coefficienti associati al fattore Cumulo premi, che pur rimanendo positivi sono diminuiti. Rimane comunque pressoché invariata la conformazione della nuvola di punti: all'aumentare della probabilità stimata diminuiscono i residui; inoltre in corrispondenza di probabilità stimate entro lo 0.004 si riscontrano i residui più significativi, soprattutto di segno positivo. Anche in questo

caso si è verificato tramite software che le celle che presentano i residui più grandi in valore assoluto hanno quasi tutte esposizione e Key ratio unitari: questi residui più ampi si devono a celle a ridotta esposizione. Riportiamo a seguire l'estrazione dei dati delle celle con i 10 residui più ampi in valore assoluto. Coincidono con le celle già viste per i modelli mod1 e mod2, ad eccezione della cella "9586".

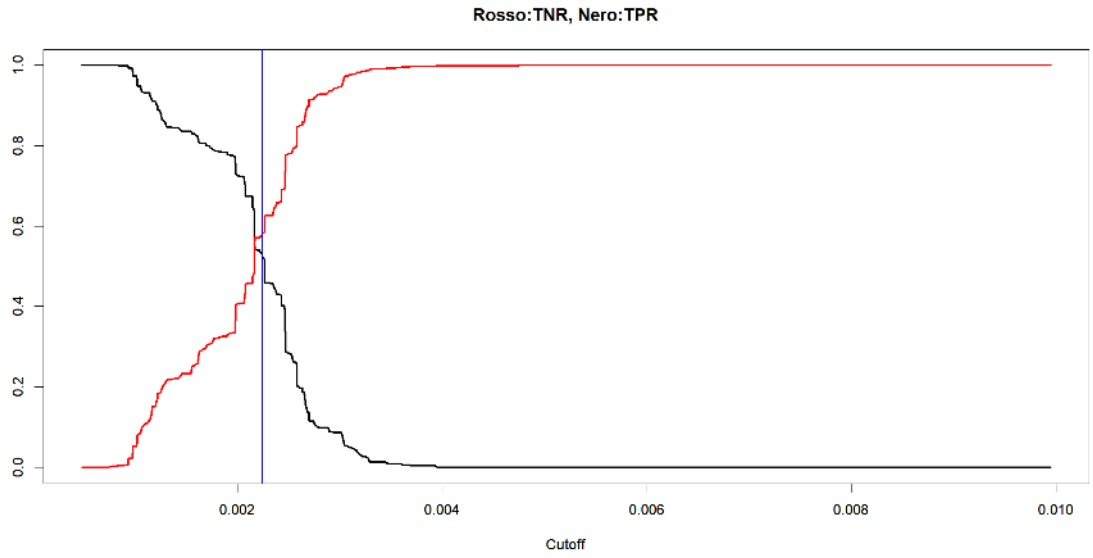
```
> celle[match(sort(abs(residuals.glm(mod3_prova, type = 'pearson')), decreasing = TRUE)[1:10],
residuals.glm(mod3_prova, type = 'pearson')), c(1:13, 18, 19)]
     period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q
1813   21q3     F    2       NO     3    1           1           4         4       1
2163   22q4     F    4       NO     3    2           2           4         4       1
2626   22q2     M    1       NO     3    3           4           4         4       1
779    22q3     F    1       SI     4    4           4           2         3       1
8870   21q3     F   12       NO     2    1           1           4         4       4
3684   21q4     M    6       NO     3    4           4           1         4       2
3832   20q4     M    4       SI     4    4           4           1         4       2
1187   21q2     M    4       NO     4    4           4           2         4       1
6833   21q4     F    6       NO     3    1           1           1         4       4
9586   22q4     M    2       NO     2    2           4           4         4       4
     riscatto2 esposiz Key_ratio   pred_prob3 pred_n_risc3
1813         1       1 1.0000000 0.0006878020 0.0006878020
2163         1       1 1.0000000 0.0006995634 0.0006995634
2626         2       3 0.6666667 0.0011454893 0.0034364680
779          1       1 1.0000000 0.0010215765 0.0010215765
8870         1       1 1.0000000 0.0014204520 0.0014204520
3684         1       1 1.0000000 0.0020429130 0.0020429130
3832         1       1 1.0000000 0.0022084396 0.0022084396
1187         2      11 0.1818182 0.0008039753 0.0088437279
6833         1       2 0.5000000 0.0011149321 0.0022298642
9586         1       1 1.0000000 0.0024245711 0.0024245711
>
```

Riportiamo poi una visione, filtrata scegliendo solo le celle con esposizione maggiore di 10, delle celle cui corrispondono i 50 residui maggiori in valore assoluto.

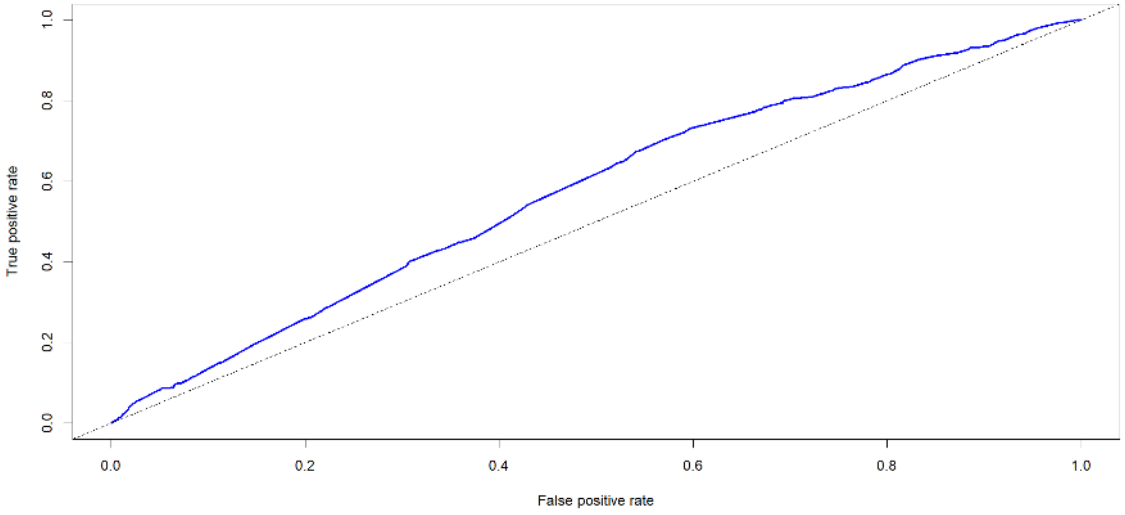
```
> celle[match(sort(abs(residuals.glm(mod3_prova, type = 'pearson')), decreasing = TRUE)[1:50], residuals.g
lm(mod3_prova, type = 'pearson')),1:15] %>% filter(esposiz > 10)
   period SESSO FRAZ LAV_AUTO ETA_Q RM_Q CUM_PREMI_Q DUR_RESID_Q DUR_TOT_Q ANTID_Q riscatto2 esposiz
1   21q2     M    4       NO     4    4           4           2         4       1         2      11
2   22q3     M    4       NO     2    4           4           2         4       4         3      34
3   21q2     F    1       NO     3    1           1           4         4       4         3      44
4   21q1     M    6       SI     4    4           4           4         4       2         2      13
5   22q1     F    2       NO     3    4           4           1         4       4         5     135
6   20q4     F    1       NO     4    4           4           2         4       1         2      54
7   20q4     F    4       SI     3    4           4           4         4       1         1      14
8   21q2     M   12       NO     4    4           4           1         4       1         1      21
9   21q1     M    2       NO     2    4           4           4         4       3         2      26
10  22q1     F    1       NO     4    4           4           3         4       1         2      81
11  21q3     F    4       SI     3    4           4           4         4       4         3      79
12  22q4     M   12       NO     4    4           4           3         4       2         1      12
13  21q2     F   12       NO     4    4           4           2         4       3         1      13
14  22q1     M   12       NO     4    4           4           2         4       3         1      14
15  22q3     F    1       SI     4    4           4           2         4       1         1      27
    Key_ratio   pred_prob1 pred_n_risc1
1  0.18181818 0.0007977508  0.008775258
2  0.08823529 0.0017731827  0.060288212
3  0.06818182 0.0016395567  0.072140496
4  0.15384615 0.0029293241  0.038081214
5  0.03703704 0.0017564372  0.237119022
6  0.03703704 0.0009518373  0.051399213
7  0.07142857 0.0010738769  0.015034277
8  0.04761905 0.0007558158  0.015872131
9  0.07692308 0.0028657948  0.074510665
10 0.02469136 0.0009619079  0.077914539
11 0.03797468 0.0022774210  0.179916256
12 0.08333333 0.0019648570  0.023578284
13 0.07692308 0.0019618394  0.025503912
14 0.07142857 0.0019258380  0.026961732
15 0.03703704 0.0010338349  0.027913541
>
```

Per il commento di questi risultati, vedere il commento ai risultati analoghi riscontrati nel modello mod1.

Concludiamo il commento al modello mod3 focalizzando l'attenzione sulle sue performance di previsione. Segue il grafico di TPR e TNR riferiti alle varie soglie, con la soglia calibrata illustrata come al solito dalla retta verticale blu. La soglia calibrata rimane sempre vicina al 2 per mille (come già accaduto per i modelli mod1 e mod2), che approssimativamente si avvicina alla probabilità di riscatto osservata sia su tutto il data-train sia su tutto il data-test.



Una volta calibrata la soglia per il modello mod3 classifichiamo le osservazioni del data-test e otteniamo la curva ROC e la confusion matrix seguenti.



```
> confusionMatrix(pred_labels_mod3, factor(data_test$riscatto2), positive = '1')
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 192466    353
         1 114455    297

               Accuracy : 0.6267
                 95% CI : (0.625, 0.6284)
    No Information Rate : 0.9979
    P-Value [Acc > NIR] : 1

                  Kappa : 9e-04

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.4569231
            Specificity : 0.6270864
         Pos Pred Value : 0.0025882
         Neg Pred Value : 0.9981693
             Prevalence : 0.0021133
         Detection Rate : 0.0009656
   Detection Prevalence : 0.3730911
      Balanced Accuracy : 0.5420048

       'Positive' Class : 1

> |
```

Notiamo innanzitutto che l'accuracy è nettamente migliorata rispetto al modello mod2, superando il valore di 60 per cento e con esso l'accuracy del modello mod1. Rimane il problema della sovrastima dei positivi: i falsi positivi hanno lo stesso ordine di grandezza dei veri negativi. Inoltre nel modello mod2 avevamo una sensitivity molto alta e una specificity molto bassa, mentre per il modello mod3 è la specificity che supera la sensitivity, seppur risultando più bilanciato il raffronto (specificity e sensitivity hanno valori meno estremi). Più in generale è cambiata la tendenza che assume il modello nel fare previsioni: il modello mod3 tende a prevedere meno positivi rispetto al mod2, il che migliora da un lato la specificity poiché riduce i falsi positivi in favore dei veri negativi, ma dall'altro riduce la sensitivity perché riduce anche i veri positivi in favore dei falsi negativi. In maniera più sintetica si potrebbe dire che il modello mod3 è più "preciso" nel suo complesso ma anche meno "prudente" (sono aumentati i falsi negativi, ovvero i contraenti che riscattano senza che la compagnia lo avesse previsto). Preferendo l'accuratezza generale della previsione, saremo dunque portati a scegliere il modello mod3 per prevedere i riscatti.

Capitolo 6

Conclusioni

In questo lavoro di tesi si è cercato di analizzare il fenomeno dei riscatti in un Piano Individuale Pensionistico, con l'obiettivo di indagare quali siano le variabili più strettamente connesse con la propensione al riscatto totale della posizione nel fondo pensione. A tal proposito è stata proposta la metodologia tipica di alcuni rami del Non life pricing, ovvero i Generalized Linear Models con covariate discretizzate. Più precisamente è stato impiegato un modello GLM con distribuzione bernoulliana delle osservazioni (e Relative binomial delle celle) e con funzione link di tipo logit: di fatto si configura la scelta di un modello di regressione logistica.
I risultati empirici ottenuti con questi modelli verranno qui ricondotti, quando possibile, alle teorie elaborate in letteratura per la modellazione dei riscatti nei prodotti del ramo vita. Le teorie in questione legano il fenomeno dei riscatti alternativamente a bisogni di liquidità o a reazioni dei contraenti a una variazione dei tassi di interesse.

Risultati

In tutti i modelli sono stati considerati sia fattori anagrafici del contraente, sia caratteristiche della polizza in questione. Sono state considerate tutte le variabili disponibili nei dati grezzi che fossero state utilizzate anche in letteratura. Tra queste si segnalano innanzitutto le variabili che **non** sono risultate significative:

- Sesso: in accordo con Milhaud et al. [2011], il sesso del contraente non sembra influire in maniera significativa sulla propensione al riscatto.

- Riserva matematica: la mancanza di significatività di questo fattore, per quanto controintuitiva, viene rafforzata sia da dei test ANOVA di significatività del fattore, sia da processi stepwise di selezione delle variabili. Bisogna tuttavia specificare che i dati della Riserva matematica sono valorizzati con il controvalore delle quote nel fondo, che può avere delle oscillazioni anche intense. Il carattere previdenziale della polizza PIP potrebbe spiegare perché i riscatti

non reagiscano particolarmente a queste oscillazioni: il contraente tipo accantona delle somme e non monitora costantemente il controvalore delle sue quote per capire quando riscattare, poiché non si tratta di un investitore che ha acquistato una polizza di investimento puro.

- Lavoratore autonomo: è un altro fattore che non risulta significativo nel suo complesso, pur essendo strettamente connesso alla situazione di liquidità dell'individuo. Va però notato che questo fattore viene incluso nel modello quando si utilizzano metodi di selezione delle variabili. Si potrebbe migliorare il modello sostituendo questa variabile con un indicatore della ricchezza o del reddito del contraente: è ragionevole ipotizzare che queste informazioni possano risultare significative nel prevedere i riscatti. A titolo esemplificativo, un lavoratore autonomo particolarmente facoltoso sarà meno soggetto a crisi di liquidità rispetto a un lavoratore dipendente a basso reddito.

- Durata totale: questa informazione è già contenuta nelle variabili Durata residua e Antidurata (età della polizza), al punto da risultare ridondante. Sembra inoltre anche più intuitivo parlare di queste due variabili anziché di Durata totale.

Si illustrano poi le variabili che sono risultate significative e le relatività ad esse associate:

- Frazionamento: ad eccezione delle polizze con pagamento quadrimestrale, si nota una tendenza generale che vede una progressiva diminuzione della propensione al riscatto all'aumentare del frazionamento. Un premio meno frazionato avrà un importo maggiore e potrebbe essere percepito dal contraente come una spesa "una tantum", per la quale non vengono preventivamente accantonati fondi a sufficienza. Questo risultato quindi va a sostegno della Emergency Fund Hypothesis ed è in accordo con Milhaud et al. [2011] e Barucci et al. [2020].

- Età: questo fattore è risultato estremamente significativo e le relatività ad esso associate indicano che si ha la massima propensione al riscatto nella prima fascia di età, seguita poi dalla seconda fascia. La terza fascia di età è quella che tende a riscattare di meno dal fondo pensione. Questo risultato è in linea con Barucci et al. [2020], dove gli autori forniscono una spiegazione della curva delle relatività che può essere facilmente ricondotta alla Emergency Fund Hypothesis. Si ritiene che i più giovani e i più anziani siano i più esposti a crisi di liquidità, poiché costituiscono la popolazione che sta entrando o che è in fase di uscita dal mercato del lavoro. Le fasce giovani verosimilmente

avranno redditi più bassi mentre le fasce più anziane potrebbero ritrovarsi in situazioni di disoccupazione prima di arrivare all'età pensionabile e in tal caso potrebbero attingere alla somma riscattata dal fondo pensione per traghettarsi verso l'età pensionabile.

- Cumulo premi: questo fattore è risultato abbastanza significativo e i suoi coefficienti hanno risentito (per magnitudo ma non a livello di curva di relatività) dell'esclusione della Riserva matematica dal modello, segnalando quindi la possibile presenza di associazioni con tale variabile. I contraenti con cumulo premi nelle due fasce più basse hanno le minori probabilità di riscatto, le quali raggiungono un massimo in corrispondenza della terza fascia di cumulo premi. Anche per la fascia più alta di cumulo premi la probabilità di riscatto è significativamente più alta rispetto alle polizze con i cumuli premi più bassi. Questa dinamica non ha una spiegazione univoca: un maggior cumulo premi potrebbe essere dovuto a una maggiore età della polizza ma anche a una maggiore disponibilità economica. Va però considerato che ci sono dei periodi iniziali in cui la polizza non può essere riscattata (o comunque in cui ci sono maggiori limitazioni al riscatto). Durante questi periodi il cumulo premi sarà tendenzialmente più basso e rientrerà nelle prime due fasce. Questo meccanismo potrebbe spiegare almeno in parte la forma della curva delle relatività riferite al cumulo premi.

- Durata residua: in questo caso la curva delle relatività indica che la propensione al riscatto non cambia nelle tre fasce di durata residua più bassa, mentre nella fascia con la maggiore durata residua si rileva una propensione decisamente maggiore. Questo può essere dovuto al fatto che le polizze più vicine alla scadenza (ovvero all'età pensionabile) vengano comprensibilmente riscattate di meno, poiché i contraenti ormai prossimi alla pensione saranno poco incentivati a trasferire la propria posizione ad altre forme pensionistiche complementari. Viceversa le polizze con la maggior durata residua saranno stipulate da contraenti mediamente più giovani, che saranno maggiormente portati a riscattare la polizza per bisogno di liquidità o per maggior propensione al cambiamento.

- Antidurata: l'età della polizza risulta estremamente significativa. Questa evidenza è forse quella che ci si aspetta maggiormente: i vincoli e i limiti al riscatto vigenti nei primi anni della polizza condizionano il comportamento di riscatto dei contraenti che necessariamente riscatteranno di meno durante il periodo iniziale. La curva delle relatività cattura questa dinamica, con una propensione di riscatto che ha il suo minimo in corrispondenza della prima fascia (che corrisponde abbastanza fedelmente al periodo iniziale durante il

quale vigono restrizioni sul riscatto). Le relatività delle altre tre fasce indicano che al di fuori del periodo iniziale la propensione al riscatto è decisamente maggiore. Questi risultati sono in linea con Milhaud et al. [2011] e Barucci et al. [2020].

I fattori sopra menzionati sono stati utilizzati anche per formulare delle previsioni sul dataset di test del modello. Si è deciso di dare maggior peso all'accuracy nella valutazione della performance di previsione del modello. Pertanto, il modello con le migliori capacità previsive è il mod3, che considera le variabili Frazionamento, Lavoratore autonomo, Età, Cumulo premi, Durata residua, Antidurata (età della polizza).

La metodologia utilizzata è sicuramente migliorabile e questo lavoro può servire come punto di partenza per l'implementazione di modelli o algoritmi più sofisticati. In primo luogo si potrebbe migliorare il modello includendo delle variabili legate al reddito del contraente: questo fattore potrebbe migliorare significativamente le capacità previsive del modello. In secondo luogo si potrebbero adottare metodi più sofisticati per discretizzare le variabili, anziché basarsi semplicemente sui quartili. Una migliore discretizzazione, o anche semplicemente un maggior numero di livelli individuati dai quantili, potrebbe migliorare la precisione delle previsioni del modello. Infine per la parte di previsione si potrebbero utilizzare degli algoritmi come le Neural Networks e Support Vector Machine, che generalmente riescono a prevedere meglio le osservazioni rispetto al modello logistico.

La dinamica dei riscatti nei fondi pensione costituisce un tema di grande rilevanza, al quale verranno molto probabilmente dedicati ulteriori studi. Si auspica che questo lavoro possa servire a supporto di eventuali future analisi.

Appendice A

Dimostrazioni

A.1 Lemma sulla CGF delle EDM

Dimostrazione per una Y variabile aleatoria EDM continua

$$M_Y(t) = \mathbb{E}\left[\exp^{tY}\right] = \int_{\mathbb{R}} e^{ty} f_Y(y)\mathrm{d}y = \int_{\mathbb{R}} \exp^{ty} \exp\Big(\frac{y\theta - b(\theta)}{\phi/\omega} + c(y,\omega,\phi)\Big)\mathrm{d}y$$
$$= \int_{\mathbb{R}} \exp\Big(\frac{y(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega} + c(y,\omega,\phi)\Big)\mathrm{d}y.$$

(definiamo $\tilde{\theta} = \theta + t\phi/\omega$)

$$= \int_{\mathbb{R}} \exp\Big(\frac{b(\tilde{\theta}) - b(\theta)}{\phi/\omega} + \frac{y\tilde{\theta} - b(\tilde{\theta})}{\phi/\omega} + c(y,\omega,\phi)\Big)\mathrm{d}y$$
$$= \exp\Big(\frac{b(\tilde{\theta}) - b(\theta)}{\phi/\omega}\Big) \int_{\mathbb{R}} \exp\Big(\frac{y\tilde{\theta} - b(\tilde{\theta})}{\phi/\omega} + c(y,\omega,\phi)\Big)\mathrm{d}y.$$

Se $|t| \approx 0$ e Θ intervallo aperto, la funzione integranda è una densità EDM e l'integrale vale 1. Quindi $M_Y(t) = e^{\frac{b(\tilde{\theta}) - b(\theta)}{\phi/\omega}}$ e

$$\Psi_Y(t) = \log(M_Y(t)) = \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega}.$$

A.2 Proprietà di MGF e CGF

In questa sezione enunciamo delle proprietà delle funzioni generatrici dei momenti (MGF) e dei cumulanti (CGF). Queste serviranno poi per dimostrare la proprietà di riproducibilità delle EDM.

A.2.1 Trasformazioni di scala

Data una costante $c \in \mathbb{R}$ e una variabile aleatoria $Y = cX$ vale che:

$$M_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[e^{tcX}\right] = M_X(tc) \implies \Psi_Y(t) = \Psi_{cX}(t) = \Psi_X(tc)$$

A.2.2 Somma di variabili indipendenti

Date le variabili aleatorie indipendenti X e Y e la loro somma $Z = X + Y$ vale che:

$$M_Z(t) = \mathbb{E}\left[e^{tZ}\right] = \mathbb{E}\left[e^{tX+tY}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right]$$

(per ipotesi di indipendenza)

$$= \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_Y(t).$$

Quindi, $\Psi_Z(t) = \Psi_{X+Y}(t) = \Psi_X(t) + \Psi_Y(t)$.

A.3 Riproducibilità delle EDM

Dimostriamo di seguito la proprietà di riproducibilità delle EDM per una coppia di variabili aleatorie EDM indipendenti, rimarcando il fatto che tale dimostrazione e dunque tale proprietà può essere estesa facilmente alla combinazione lineare di un numero arbitrario di variabili aleatorie.

Date le variabili indipendenti $Y_1 \sim EDM(\theta, b(), \phi, \omega_1)$ e $Y_2 \sim EDM(\theta, b(), \phi, \omega_2)$ e il Key ratio sul loro aggregato

$$Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2} = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega} = \frac{\omega_1}{\omega}Y_1 + \frac{\omega_2}{\omega}Y_2 = Z_1 + Z_2$$

Studiamo la distribuzione di Y passando per la derivazione della sua CGF $\Psi_Y(t)$. Notiamo che, per la proprietà della CGF della somma di variabili indipendenti, $\Psi_Y(t) = \Psi_{Z_1+Z_2}(t) = \Psi_{Z_1}(t) + \Psi_{Z_2}(t)$, e che, per la proprietà della CGF della trasformata in scala, $\Psi_{Z_i}(t) = \Psi_{Y_i}(t\omega_i/\omega)$. Segue che

$$\begin{aligned}\Psi_{Z_i}(t) = \Psi_{Y_i}(t\omega_i/\omega) &= \frac{b(\theta + t\frac{\phi}{\omega_i}\frac{\omega_i}{\omega}) - b(\theta)}{\phi/\omega_i} \\ &= \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega_i} = \frac{\omega_i}{\phi}(b(\theta + t\phi/\omega) - b(\theta))\end{aligned}$$

Abbiamo quindi che

$$\begin{aligned}\Psi_Y(t) &= \frac{\omega_1}{\phi}(b(\theta + t\phi/\omega) - b(\theta)) + \frac{\omega_2}{\phi}(b(\theta + t\phi/\omega) - b(\theta)) \\ &= \frac{\omega_1 + \omega_2}{\phi}(b(\theta + t\phi/\omega) - b(\theta)) = \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega}.\end{aligned}$$

Abbiamo così ottenuto la CGF di una EDM, più precisamente:

$$Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2} \sim EDM(\theta, b(), \phi, \omega_1 + \omega_2).$$

Concludiamo questa sezione sottolineando che in questa tesi la variabile aleatoria osservata su ogni singola osservazione, ovvero Y_i, è distribuita secondo un modello di Bernoulli, mentre l'esposizione di ciascuna singola osservazione sarà pari a 1 ($\omega_i = 1\ \forall i$). In questo frangente quindi il Key ratio Y della cella, ottenuto come media ponderata delle Y_i, diventa di fatto una media semplice e più precisamente una proporzione di successi nella cella. La proporzione di successi nella cella, ovvero Y, corrisponde a una variabile aleatoria binomiale divisa per il rispettivo parametro del numero di prove. Questo modello di variabile aleatoria viene detto "Relative Binomial" e appartiene alle EDM (si veda la sezione seguente per una dimostrazione), in linea con quanto affermato dalla proprietà di riproducibilità delle EDM.

A.4 Appartenenza alle EDM

In questa sezione dimostriamo che la Relative Binomial appartiene alle EDM. Segnaliamo che una variabile aleatoria bernoulliana può essere considerata anche come una Relative Binomial con $\omega = 1$, quindi quanto segue può essere facilmente esteso alle variabili aleatorie bernoulliane.

A.4.1 Relative Binomial

Se $Y \sim$ Relative Binomial$(\mu = p, \omega = n)$

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(X = ny) = \binom{n}{ny} p^{ny}(1-p)^{(n-ny)}$$

e quindi

$$\begin{aligned}\log(\mathbb{P}(Y=y)) &= \log(\binom{n}{ny}) + ny\log(p) + n(1-y)\log(1-p)\\ &= \log(\binom{n}{ny}) + ny(\log(p)-\log(1-p)) + n\log(1-p)\\ &= \frac{y\ logit(p)+\log(1-p)}{1/n} + \log(\binom{n}{ny})\end{aligned}$$

Definiamo $\theta = logit(p)$. Ovvero, $p = \frac{e^\theta}{1+e^\theta}$ e $\log(1-p) = -\log(1+e^\theta) = -b(\theta)$. Quindi, $\log(\mathbb{P}(Y=y)) = \frac{y\theta-\log(1+e^\theta)}{1/n} + \log(\binom{n}{ny})$. Possiamo quindi concludere che Y appartiene alla classe EDM.

A.5 Equazioni di verosimiglianza, derivazione max likelihood

A.5.1 EDM Generica

La verosimiglianza del vettore y in caso di indipendenza è data da $L(\beta, y) = \Pi_{i=1}^n f_{Y_i}(y_i \mid \beta)$, per cui la Log-verosimiglianza sarà $l(\beta, y) = \sum_{i=1}^n \log(f_{Y_i}(y_i \mid \beta)) = \sum_{i=1}^n l_i$, con

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c().$$

Vogliamo trovare β tale che $\frac{dl}{d\beta_j} = \sum_{i=1}^n \frac{dl_i}{d\beta_j} = 0 \quad \forall j$ livello tariffario. A questo scopo, consideriamo la catena:

$$\frac{dl_i}{d\beta_j} = \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_j}$$

ed andiamo ad analizzare ogni fattore.

Il primo fattore si può riscrivere come

$$\frac{dl_i}{d\theta_i} = \frac{d}{d\theta_i}(\omega_i(y_i\theta_i - b(\theta_i)/\phi) = (y_i-\mu_i)\omega_i/\phi.$$

Il secondo fattore è dato da

$$\frac{d\theta_i}{d\mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{\nu(\mu_i)}$$

poiché $\frac{d\mu_i}{d\theta_i} = \frac{db'(\theta_i)}{d\theta_i} = b''(\theta_i)$ e $b'(\theta_i)$ invertibile (sfruttiamo il teorema delle funzioni

implicite, richiediamo che $b''(\theta_i) \neq 0$).

Il terzo fattore è

$$\frac{d\mu_i}{d\eta_i} = \frac{1}{g'(\mu_i)}$$

poiché $\dfrac{d\eta_i}{d\mu_i} = \dfrac{dg(\mu_i)}{d\mu_i} = g'(\mu_i)$ e $g(\mu_i)$ invertibile (sfruttiamo il teorema delle funzioni implicite, richiediamo che $g'(\mu_i) \neq 0$). Infine,

$$\frac{d\eta_i}{d\beta_j} = \frac{d(x_i^T\beta)}{\beta_j} = x_{i,j}.$$

Moltiplicando questi fattori e otteniamo

$$\frac{dl_i}{d\beta_j} = \frac{(y_i - \mu_i)\omega_i}{\phi} \; \frac{1}{\nu(\mu_i)} \; \frac{1}{g'(\mu_i)} \; x_{i,j}$$

Assemblando il tutto:

$$\begin{aligned}\frac{dl}{d\beta_j} &= \sum_{i=1}^{n} \frac{dl_i}{d\beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\omega_i}{\phi} \; \frac{1}{b''(\theta_i)} \; \frac{1}{g'(\mu_i)} \quad x_{i,j} \\ &= \sum_{i=1}^{n} \frac{1}{\phi} \; x_{i,j} \; \frac{\omega_i}{\nu(\mu_i)g'(\mu_i)} \; (y_i - \mu_i) = \frac{1}{\phi} \sum_{i=1}^{n} x_{i,j} \, \tilde{\omega}_i \; (y_i - \mu_i).\end{aligned}$$

In questa equazione la dipendenza da β è insita nella dipendenza da $\mu_i = g^{-1}(x_i^T\beta)$.

L'equazione è non lineare in β, ad eccezione del caso gaussiano con link identità (Weighted least Squares).

Raduniamo le equazioni dei vari livelli in un sistema in forma matriciale con tante righe quanti sono i livelli, otteniamo il vettore score (gradiente):

$$\frac{dl}{d\beta} = \frac{1}{\phi} \, \mathbf{X}^T \, \tilde{\mathbf{W}} \, (y - \mu), \qquad \text{con } \tilde{\mathbf{W}} \text{ matrice diagonale di } \tilde{\omega}_i.$$

Le equazioni di verosimiglianza si possono riassumere con: $\mathbf{X}^T \; \tilde{\mathbf{W}} \; (y - \mu) = 0$. Si noti che $\mathbf{X}^T$ è una matrice le cui righe sono le variabili dummy e una riga di 1 associata all'intercetta. Questo significa che in ogni equazione del sistema la somma sugli $x_{i,j}$ di fatto consiste in una aggregazione sulle osservazioni associate a ciascun livello.

A.5.2 Relative Binomial

Il procedimento per ricavare le equazioni di verosimiglianza si semplifica notevolmente se si considera che la funzione logit costituisce il link canonico. Nella sezione

successiva si dimostra che, utilizzando il link canonico, vale che:

$$\tilde{\omega}_i = \omega_i \quad \Longrightarrow \quad \frac{dl}{d\beta} = \frac{1}{\phi}\,\mathbf{X}^T\,\tilde{\mathbf{W}}\,(y-p) = \frac{1}{\phi}\,\mathbf{X}^T\,\mathbf{W}\,(y-p) = \frac{1}{\phi}\,\mathbf{X}^T\,(y-p)$$

dove l'ultima uguaglianza discende dal fatto che la matrice diagonale dei pesi $\mathbf{W} = \mathbf{I}$ matrice identità nel nostro caso di Relative Binomial. La struttura in dummy della matrice $\mathbf{X}$ di fatto impone che, aggregando per ciascun livello tariffario, le probabilità stimate di riscatto p coincidano con la frequenza empiricamente osservata di riscatto. La struttura di $\mathbf{X}^T$ sarà infatti la seguente (n colonne quante le osservazioni, K righe quanti i livelli) :

$$\begin{bmatrix} DummyLivello_1 \\ ... \\ DummyLivello_j \\ ... \\ DummyLivello_K \end{bmatrix}$$

Nel prodotto matriciale con il vettore (y-p), in ogni riga si sommeranno i valori di $(y_i - p_i)$ per tutte e sole le osservazioni che presentano quel livello tariffario.

A.6 Link canonico, EDM e Relative Binomial

A.6.1 EDM Generica

Si è già visto che nelle equazioni di verosimiglianza entra la quantità $\tilde{\omega_i} = \dfrac{\omega_i}{\nu(\mu_i)g'(\mu_i)}$. Riscriviamo il denominatore $g'(\mu_i)\nu(\mu_i) = g'(b'(\theta_i))b''(\theta_i) = \dfrac{d}{d\theta_i}g(b'(\theta_i))$ e imponiamo che valga $\dfrac{d}{d\theta_i}g(b'(\theta_i)) = 1$. Integriamo in $d\theta_i$ a entrambi i lati e otteniamo $g(b'(\theta_i)) = \theta_i \implies b'() = g^{-1}()$.

Vediamo quindi che il link canonico $g(\mu_i)$ assume forma funzionale inversa rispetto alla derivata prima rispetto a θ della funzione cumulante.

A.6.2 Relative Binomial

Per la relative binomial $b'(\theta_i) = \dfrac{e^{\theta_i}}{1+e^{\theta_i}}$. Per avere il link naturale, imponiamo $g(b'(\theta_i)) = g(\dfrac{e^{\theta_i}}{1+e^{\theta_i}}) = \theta_i$. Ipotizziamo un link di tipo logit: $g(\mu_i) = g(b'(\theta_i)) =$

$\log(b'(\theta_i)) - \log(1 - b'(\theta_i)) = \log(\frac{e^{\theta_i}}{1+e^{\theta_i}}) - \log(1 - \frac{e^{\theta_i}}{1+e^{\theta_i}}) = \theta_i - \log(1+e^{\theta_i}) - \log(\frac{1}{1+e^{\theta_i}}) = \theta_i - \log(1+e^{\theta_i}) + \log(1+e^{\theta_i}) = \theta_i.$

Bibliografia

Michele Azzone, Emilio Barucci, Giancarlo Giuffra Moncayo, and Daniele Marazzina. A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, 2022. URL doi.org/10.1016/j.eswa.2021.116261.

E. Barucci, T. Colozza, D. Marazzina, and E.Rroji. Lapse risk in life insurance contracts. 2020. URL re.public.polimi.it/bitstream/11311/1134359/4/PrePrint.pdf.

Paolo Brandimarte. *Numerical Methods in Finance and Economics*. Wiley, 2013.

R. Cerchiara, M. Edwards, and A. Gambini. Generalized linear models in life insurance: Decrements and risk factor analysis under solvency ii. 2009. URL hiips://wwwactuariesorg/AFIR/Colloquia/Rome2/Cerchiara_Edwards_Gambini.pdf.

DLgs. Decreto legislativo 252 del 2005. 2005. hiips://wwwcovipit/sites/default/files/notizie/A016Decreto-252.pdf.

Annette J. Dobson. *Introduction to Generalized Linear Models*. Chapman Hall/CRC Press, 2001.

Martin Eling and Dieter Kiesenbauer. What policy features determine life insurance lapse? an analysis of the german market. *Journal of Risk and Insurance*, 2013. URL hiips://doi.org/10.1111/j.1539-6975.2012.01504.x.

Peter J. McCullagh and John A. Nelder. *Generalized Linear Models (second edition)*. Routledge, 1989.

X. Milhaud, S. Loisel, and V. Maume-Deschamps. Surrender triggers in life insurance: what main features affect the surrender behavior in a classical economic context? 2011. URL hiips://hal.science/hal-00450003/document.

MinLav. Previdenza complementare. 2023. hiips://wwwlavorogov.it/temi-e-priorita/previdenza/focus-on/previdenza-complementare/pagine/default.

E. Ohlsson and Bj. Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. 2010.

Sito COVIP FAQ. 2023.

Lim Jin Xong and Ho Ming Kang. A comparison of classification models for life insurance lapse risk. 2019. URL wwwijrteorg/wp-content/uploads/papers/v7i5s/ES2151017519.pdf.