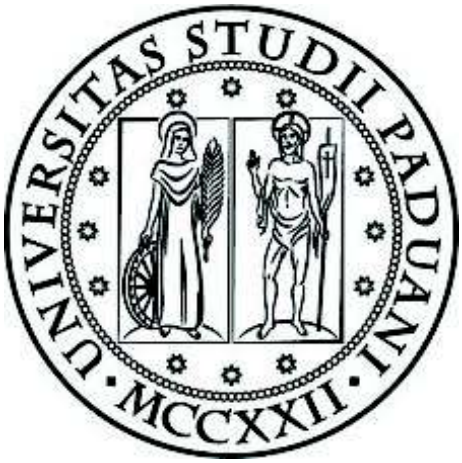
UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA
IN SCIENZE STATISTICHE, ECONOMICHE,
FINANZIARIE E AZIENDALI



TESI DI LAUREA

MODELLI GLLAMM PER LA STIMA DEL RISCHIO DI CHIUSURA DELL'ATTIVITÀ: IL CASO DELLE AZIENDE AGRICOLE VENETE

RELATORE: PROF. ADRIANO PAGGIARO

LAUREANDO: ANDREA VALENTE

ANNO ACCADEMICO 2009/2010

Indice

1. Introduzione e sommario

I criteri di sovvenzione alle imprese agricole sono stati sostanzialmente modificati dalle riforme della Politica Agricola Comunitaria (PAC) iniziate con Agenda 2000. Nel tentativo di avvicinarsi a una logica di mercato si è deciso di introdurre una maggiore selettività per le imprese da sovvenzionare valutandone la "vitalità", cercando di capire quali sono le caratteristiche di quelle che rimangono attive nel tempo consentendo un utilizzo più efficiente dei fondi della PAC.

Questa tesi analizza la sopravvivenza delle imprese agricole venete utilizzando un dataset che integra dati sulla sopravvivenza delle imprese ricavati dagli archivi REA delle Camere di Commercio, con i dati ISTAT del V Censimento Generale dell'Agricoltura.

La tesi si incentra sostanzialmente sul confronto tra il modello a effetti casuali con variabile risposta binaria stimabile con STATA attraverso la funzione *xtlogit,* e il modello a due livelli stimabile attraverso il pacchetto esterno di STATA, **gllamm**, sviluppato per il software stesso da Skrondal, Rabe-Hesketh e Pickles (2004).

Il fine principale è verificare se l'approccio Bayesiano empirico utilizzato per "inserire" l'eterogeneità non osservata nelle previsioni produca miglioramenti nella classificazione delle imprese sopravvissute rispetto all'approccio che non considera tale eterogeneità.

A differenza dell'analisi di Biasiolo (2006) che prendeva in considerazione tutte le aziende presenti nell'archivio REA, si è deciso di ridurre le analisi alla coorte delle imprese nuove iscritte nell'anno 1999. Questa scelta permette di osservare l'ingresso di tutte le unità nel dataset, eliminando possibili problemi di *length bias*. L'abbinamento di archivi REA e Censimento consente da un lato di seguire le aziende fino al 2004, dall'altro di osservare tutte le caratteristiche dell'azienda disponibili nel censimento.

La tesi è strutturata come segue: nel capitolo 2 viene presentata la Politica Agraria Comunitaria, introducendo quindi l'importanza della determinazione della probabilità di sopravvivenza delle aziende agricole in un ottica di maggiore selettività dei finanziamenti alle stesse. Nel capitolo 3 presentiamo una breve analisi

descrittiva dei dati e delle variabili utilizzate per la stima dei modelli di interesse. Nel capitolo 4 introducendo i modelli di durata a tempi discreti (ricostruendo il dataset in formato *episode-splitting*), andiamo a stimare sui dati in nostro possesso il modello ad effetti casuali per variabile risposta binaria (*xtlogit*) utilizzando le variabili UDE e OTE indicative rispettivamente dell'orientamento tecnico-economico delle aziende e della dimensione economica aziendale, e ne verifichiamo la capacità di classificazione attraverso l'approccio analitico alla *curva ROC.* Nel capitolo 5 presentiamo in maniera relativamente dettagliata un approfondimento sui modelli Gllamm (*Generalized Linear Latent and Mixed Models*), sulla stima dei parametri e sulla previsione empirico - bayesiana degli effetti casuali e delle variabili dipendenti riconducendoci poi al caso del modello a due livelli in un contesto di modelli di durata a tempi discreti. Viene trattato il metodo di previsione proposto da Biasiolo (2006) chiamato "*cluster-averaged expectation*" e viene introdotto un metodo alternativo, il "*prior prediction method",* spiegando come si ottengono tali previsioni dal pacchetto **gllamm**. Nel capitolo 6 andremo invece ad effettuare delle simulazioni di dataset simili a quello utilizzato per le analisi sulla sopravvivenza delle aziende agricole con il fine di illustrare eventuali differenze in termini di stima dei parametri e di classificazione utilizzando anche in questo caso l'approccio analitico alla *Roc Curve* e *Roc Area*: a questo scopo sono stati confrontati i diversi metodi in assenza/presenza di eterogeneità non osservata nel caso in cui le analisi fossero svolte su un panel bilanciato e su un dataset in formato *episode-splitting*. Il capitolo 7, infine, contiene essenzialmente l'applicazione sui dati reali, trattati nei primi capitoli, delle metodologie implementate dal pacchetto *gllamm* e dalla funzione interna di STATA, *xtlogit*.

Essendo principalmente lo scopo della tesi quello di indagare se vi fosse effettivamente un miglioramento delle stime delle probabilità di cessazione adottando l'approccio GLLAMM, si è giunti a verificare che in questi termini in realtà le due metodologie si equivalgono.

Dalla trattazione teorica dei modelli Gllamm nel capitolo 5 e dalle simulazioni effettuate nel capitolo 6, si è capito che non è possibile utilizzare l'approccio "*cluster-averaged expectation*" per fare previsioni su un nuovo "cluster", in quanto sono necessari i valori della variabile risposta osservati nel cluster stesso. Questa

caratteristica rende poco utile questo metodo per le analisi di interesse della tesi e smentirà sostanzialmente le conclusioni ottenute da Biasiolo (2006).

Per questo motivo si introduce una metodologia implementata in **gllamm** che sembra essere più adatta ai nostri scopi predittivi, il "*prior prediction method"* che teoricamente utilizza la distribuzione a priori dell'effetto casuale per effettuare la previsione. Tale metodologia, dalle simulazioni svolte, non sembra però fornire risultati migliori in termini classificativi dell'approccio predittivo che utilizza la funzione interna di Stata (*xtlogit*), la quale assume nella previsione assenza di eterogeneità: nella quasi totalità delle casistiche trattate in ambito di simulazione, considerare l'approccio previsivo fornito da *xtlogit* e utilizzare il "*prior prediction method"* di *gllamm* portano essenzialmente alle medesime conclusioni. Si è visto infatti che sebbene la distribuzione delle probabilità stimate dai due metodi possano essere anche diverse, in realtà nessuna delle due metodologie sfrutta in maniera migliore l'informazione, e ciò si vede dai vari test effettuati sulle *Roc Areas*.

In termini di stima dei parametri *gllamm* e *xtlogit* sembrano diversificarsi essenzialmente per i diversi metodi di quadratura utilizzata: *xtlogit* utilizza il metodo della quadratura di Gauss-Hermite sviluppato per Stata da Liu, Qing e Pierce (1994), *gllamm* utilizza il metodo sviluppato da Naylor e Smith (1982) e poi implementato da Skrondal e Rabe-Hesketh (2004). *Xtlogit* adatta la quadratura utilizzando la moda e la curvatura della moda della funzione di verosimiglianza mentre *gllamm* utilizza una stima della media e della varianza di questa. Sembra che dall'esperienza degli autori quest'ultimo tipo di quadratura funzioni molto meglio nei casi di elevata eterogeneità non osservata.

Per quanto riguarda le applicazioni sui dati reali, si è visto che è efficace utilizzare un modello ridotto che tiene in considerazione le variabili OTE e UDE ai fini di una politica di sostegno alle aziende agricole. Considerando tutte le aziende ugualmente meritevoli, la probabilità di sovvenzionare un'azienda che chiuderà entro 5 anni è pari al 23.84%; tale probabilità scende al 13.48% se si utilizzano determinate soglie con entrambe le metodologie. Tale decisione comporta però il mancato finanziamento del 59.84% delle aziende che sopravvivranno per più di 5 anni. Si è verificato inoltre, attraverso l'analisi *Roc*, che utilizzare tutte le informazioni disponibili come esplicative fornisce un miglioramento statisticamente significativo,

intorno ai 2 punti percentuali, delle performance. Si è visto, comunque, come in entrambi i modelli la dimensione economica sia fortemente determinante per la sopravvivenza delle aziende prese in esame. Rispetto all'obiettivo metodologico della tesi, le conclusioni sostanziali cui si giunge con l'approccio gllamm sono essenzialmente le medesime ottenute con i modelli di durata ad effetti casuali, in particolare se si sfrutta tutta l'informazione disponibile.

2.La Politica Agricola Comunitaria

La politica agricola comune o comunitaria (PAC) è una delle politiche comunitarie di maggiore importanza, impegnando circa il 44% del bilancio dell'Unione Europea ed è prevista dal Trattato Istitutivo delle Comunità.

L'agricoltura ha rappresentato, fin dai tempi dei negoziati del trattato di Roma, uno degli obiettivi prioritari delle istanze politiche decisionali europee. A quell'epoca era ancora vivo il ricordo delle penurie alimentari dell'immediato dopoguerra, e l'agricoltura ha costituito un elemento chiave delle politiche europee fin dagli esordi della Comunità.

La Politica Agricola Comunitaria (PAC) consiste in una serie di norme e meccanismi che regolano la produzione, gli scambi e la lavorazione dei prodotti agricoli nell'ambito dell'Unione Europea. La base giuridica della politica agraria comune è definita ad oggi agli articoli 32-38 del titolo II del Trattato CE. Il Trattato di Roma definiva gli obiettivi generali della politica agraria comune, principi fissati durante la conferenza di Stresa del luglio 1958. Nel 1960, i sei membri fondatori della Comunità europea hanno adottato i meccanismi della PAC e due anni dopo, nel 1962, la PAC è entrata in vigore.

Le finalità della PAC, secondo quanto stabilito dall'articolo 33 del Trattato CE, sono le seguenti:

- Incrementare la produttività dell'agricoltura, sviluppando il progresso tecnico, assicurando lo sviluppo razionale della produzione agricola e un impiego migliore dei fattori di produzione, in particolare della manodopera.
- Assicurare alla popolazione agricola un tenore di vita equo, intervenendo, in particolare, sul miglioramento del reddito individuale di coloro che lavorano nell'agricoltura.
- Stabilizzare i mercati.
- Garantire la sicurezza degli approvvigionamenti.
- Assicurare prezzi ragionevoli per i consumatori.

Per raggiungere tali obiettivi, l'articolo 34 del Trattato prevede la creazione di una organizzazione comune dei mercati agricoli (OCM) che, a seconda dei prodotti, andava a stabilire:

- Regole comuni in materia di concorrenza.
- Un coordinamento obbligatorio delle diverse organizzazioni nazionali del mercato.
- Un'organizzazione europea del mercato.

Tre principi fondamentali, definiti nel 1962, caratterizzano il mercato agricolo comune e quindi le OCM:

- Un mercato unificato, inteso come libera circolazione dei prodotti agricoli nell'ambito degli Stati membri.
- La preferenza comunitaria, ovvero la priorità negli scambi per i prodotti agricoli dell'Unione europea, in modo tale da renderli più vantaggiosi dal punto di vista dei prezzi rispetto ai prodotti importati dai paesi terzi e proteggerli dalle grandi fluttuazioni sul mercato mondiale.
- La solidarietà finanziaria, ovvero il sostenimento di tutte le spese e i costi inerenti l'applicazione della PAC da parte del bilancio comunitario.

Le organizzazioni comuni dei mercati sono state introdotte in maniera graduale; attualmente esse esistono per la maggior parte dei prodotti agricoli e costituiscono gli strumenti del mercato agricolo comune in quanto eliminano gli ostacoli agli scambi intracomunitari dei prodotti e mantengono barriere doganali comuni nei confronti dei paesi terzi.
Lo strumento finanziario della PAC è il Fondo europeo agricolo d'orientamento e di garanzia (FEAOG) che rappresenta una parte sostanziale del bilancio comunitario. Il FEAOG è stato istituito nel 1962 e suddiviso nel 1964 in due sezioni:

- La sezione "orientamento", che fa parte dei fondi strutturali e che contribuisce alle riforme agricole strutturali e allo sviluppo delle zone

rurali (ad esempio, tramite investimenti nelle nuove attrezzature e tecnologie).

- La sezione "garanzia", che finanzia le spese inerenti l'organizzazione comune dei mercati (ad esempio, tramite l'acquisto o lo stoccaggio delle eccedenze e la promozione delle esportazioni dei prodotti agricoli).

Nel corso degli anni, la PAC ha realizzato con successo i suoi obiettivi iniziali: è riuscita, infatti, a promuovere sia la produzione che la produttività, ha stabilizzato i mercati, assicurato l'approvvigionamento dei prodotti e protetto gli agricoltori contro le fluttuazioni dei prezzi sui mercati mondiali.

La PAC ha subito, nel corso dei quattro decenni della sua esistenza, numerose riforme. Il primo tentativo risale al 1968, con la pubblicazione da parte della Commissione di un "Memorandum sulla riforma della PAC", comunemente detto "Piano Mansholt", dal nome del suo promotore, Sicco Mansholt, all'epoca vice presidente della Commissione e responsabile della PAC. Il piano prevedeva la riduzione della popolazione attiva in agricoltura e l'incoraggiamento alla formazione di unità di produzione agricola più grandi e più efficienti.

Nel 1972 sono state introdotte misure strutturali rivolte in particolare alla modernizzazione dell'agricoltura e, nel 1983, la Commissione propose una riforma sostanziale, che fu formulata ufficialmente due anni dopo con la pubblicazione del libro verde sulle "Prospettive della politica agraria comune" (1985). Il documento aveva l'obiettivo primario di ristabilire l'equilibrio tra l'offerta e la domanda, di formulare nuove soluzioni per ridurre la produzione nei settori in difficoltà e, in genere, di proporre possibili alternative per il futuro della PAC.

Nel 1988, il Consiglio europeo ha raggiunto un'intesa su un insieme di interventi di riforma che limitavano la percentuale della spesa della PAC nel quadro del bilancio generale.

Nel 1991 la Commissione e Ray MacSharry, membro responsabile dell'agricoltura, hanno presentato due documenti di riflessione sullo sviluppo e il futuro della PAC; tali documenti costituivano la base per un'intesa politica sulla riforma raggiunta effettivamente il 21 maggio 1992. I cambiamenti più importanti riguardavano la

compensazione per le perdite di reddito subite dagli agricoltori, i meccanismi del mercato e la protezione dell'ambiente.
Nel luglio 1997, la Commissione ha proposto una riforma della PAC nel quadro di Agenda 2000, nell'ottica dell'allargamento ad est dell'Unione. Le trattative si conclusero nel marzo 1999, al consiglio di Berlino. Agenda 2000 ha rappresentato un cambiamento radicale della politica agraria comune: portando avanti il processo iniziato nel 1992, infatti, ha fornito una solida base per il futuro sviluppo dell'agricoltura nell'Unione, contemplando tutti gli ambiti di competenza della PAC (economico, ambientale e rurale) (Vieri, 2001).
La riforma ha compreso, in particolare, misure intese a :

- Rafforzare la competitività delle materie prime agricole sui mercati interni e mondiali.
- Promuovere un tenore di vita adeguato alla comunità agricola.
- Creare posti di lavoro sostitutivi e altre fonti di reddito per i lavoratori agricoli.
- Elaborare una nuova politica dello sviluppo rurale come secondo pilastro della PAC.
- Integrare maggiormente questioni ambientali e strutturali.
- Migliorare la qualità dei prodotti alimentari.
- Semplificare la legislazione in materia agraria e decentralizzarne l'applicazione, in vista di una maggiore chiarezza, trasparenza e accessibilità di norme e regolamenti.

Con tale riforma si sono create le condizioni per lo sviluppo di un'agricoltura comunitaria multifunzionale, sostenibile e concorrenziale.
A metà 2003 sono state introdotte ulteriori importanti riforme che danno attuazione ai principi dell'Agenda 2000: esse rappresentano globalmente la ristrutturazione più radicale subita dalla PAC dal 1958 ad oggi.
Non solo viene dato un taglio netto alle sovvenzioni alla produzione a vantaggio degli aiuti diretti agli agricoltori, ma la stessa concessione di tali aiuti è subordinata al rispetto delle norme vigenti in materia di ambiente, benessere degli animali, igiene

e conservazione del paesaggio rurale. Le trasformazioni hanno anche preparato la PAC ad affrontare la sfida dell'allargamento, avvenuto nel maggio 2004, e che ha comportato, con il passaggio da 15 a 25 membri, l'aumento di quasi il 70% di agricoltori nell'UE (Cioccolo et al., 2004).

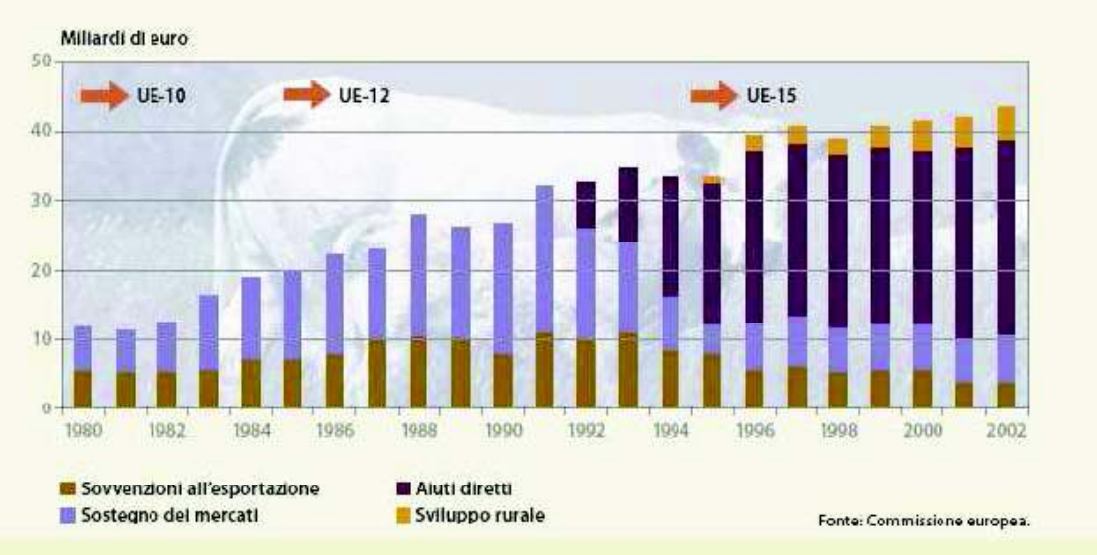


Figura 2.1*: Evoluzione della spesa per la PAC*.

Nella Figura 2.1 è rappresentata l'evoluzione della spesa per la PAC dal 1980 al 2002. Si nota in generale una continua crescita della spesa e, dall'inizio degli anni '90, un sempre maggiore incremento degli aiuti diretti a scapito della spesa dedicata al sostegno dei mercati.

2.2 La riforma Fischler (2003)

Nasce come "revisione di medio termine" (*Mid-Term Review* – MTR) di Agenda 2000, doveva infatti essere una semplice revisione di metà percorso per verificare l'effetto delle riforme introdotte nel 1999: in realtà la proposta contiene novità che vanno ben oltre la semplice verifica, la *Mid-Term Review* si è trasformata in una riforma più incisiva della stessa Agenda 2000, approfondendo il processo di riorientamento degli strumenti e la finalità stessa del sostegno.
Gli obiettivi di questa riforma senz'altro furono quelli di:

- Migliorare la competitività dell'agricoltura europea.
- Riorientare la produzione al mercato.
- Promuovere un'agricoltura sostenibile e socialmente accettabile.
- Rafforzare lo sviluppo rurale.
- Semplificare il regime di sostegno.
- Rendere la PAC più rispondente agli impegni assunti o da assumere in sede WTO (*World Trade Organization*).

La PAC post-riforma è rappresentata da un sostegno in larghissima parte disaccoppiato, legato al possesso della terra sulla quale deve svolgersi l'attività agricola e vincolato al rispetto di standard minimi ambientali, di qualità alimentare, di salubrità dei prodotti agricoli, di benessere degli animali, di gestione dei terreni agricoli. Tale riforma ritaglia un ampio ruolo per gli Stati membri chiamati ad operare una serie di scelte per adattare la PAC alle specifiche realtà territoriali.
Ruota attorno a tre cardini fondamentali:

1) Disaccoppiamento degli aiuti e istituzione del "regime di pagamento unico" (RPU).
2) Modulazione degli aiuti diretti.
3) Condizionalità degli aiuti diretti.

Il disaccoppiamento rappresenta il cuore della riforma della PAC, in cui l'aiuto non è più "accoppiato" alla quantità prodotta ma appunto "disaccoppiato", non è più legato a *cosa* si produce ma alla più generale *attività agricola*: possesso della terra e esercizio dell'attività produttiva, dove per essa si intende anche il mantenimento della terra in buone condizioni agronomiche e ambientali. Nel corso di 40 anni si è passati da un aiuto legato a "*quanto* si produce", ad un aiuto legato a "*cosa* si produce".
Si è quindi transitati da una produzione "orientata" ai sussidi a una produzione "orientata" al mercato: prima della riforma, la scelta di cosa produrre era dettata dall'entità degli aiuti che si potevano ottenere a seconda di come veniva ripartita la superficie tra le diverse colture, indipendentemente dalle richieste del mercato. Quindi anche in "assenza" di questo, la produzione realizzata poteva essere venduta a un prezzo minimo garantito. La riforma, slegando l'aiuto dal prodotto, determina un maggiore orientamento delle scelte dei produttori alle condizioni del mercato.
Gli effetti del disaccoppiamento divengono:

1) La riduzione della produzione: non è più necessario produrre per ottenere l'aiuto (il rovescio della medaglia è il rischio di abbandono) e inoltre la riduzione non è uniforme tra i prodotti e riguarda soprattutto le produzioni prima "orientate" all'ottenimento dei sussidi.
2) Il mercato si regola equilibrando domanda e offerta ed è meno influenzato dalle distorsioni dei meccanismi di formazione dei prezzi agricoli.
3) Si ha un passaggio vero e proprio da un sostegno al prodotto a un sostegno al reddito dei produttori.

L'RPU (*Regime di pagamento unico*) diventa una sorte di contenitore nel quale vengono travasati la maggior parte degli aiuti diretti prima contenuti nelle singole OCM, inquadrandoli in una cornice unica di diritti e obblighi.
Gli aiuti diretti che confluiscono nel RPU sono elencati nella Tabella 2.1:

Aiuto alla superficie per seminativi, legumi da granella, riso	
Aiuto alla produzione per le sementi	
Indennità ai produttori di patate da amido	
Premi OCM carni bovine e ovicaprine	
Aiuti alle regioni ultraperiferiche	
Pagamenti per foraggi essiccati	
Premi e pagamenti supplementari OCM latte (tra 2005 e 2007)	
Aiuti per superficie per il luppolo	
Aiuti alla produzione olio d'oliva	
Aiuti alla produzione tabacco dal 2006	dal 2006
Aiuto alla produzione cotone	
Aiuti "compensativi" zucchero	

Tabella 2.1: *Elenco degli aiuti diretti che confluiscono nel RPU*

Ogni agricoltore che accede al RPU è titolare di un certo numero di *diritti all'aiuto per ettaro,* cioè per ricevere il pagamento dell'importo fissato nel titolo, ogni diritto all'aiuto deve essere abbinato ad un ettaro di superficie aziendale.

Per ciascuno Stato membro è fissato un massimale nazionale, cioè un tetto all'ammontare di sostegno a cui ha diritto ed è calcolato in base alla media degli aiuti storici ricevuti da ciascuno Stato negli anni di riferimento.

La modulazione è senza dubbio un altro punto importante della riforma e prevede la riduzione di tutti i pagamenti diretti, allo scopo di finanziare la nuova politica di

sviluppo rurale. E' obbligatoria per gli Stati membri partendo dal 2005, con un taglio che va dal 3% al 5% in tre anni (dal 2007 fino al 2012 rimane il 5%); prevede una franchigia sui primi 5000 euro di aiuti per ciascuna azienda, mentre le risorse tagliate, al netto dell'aiuto aggiuntivo, sono destinate allo sviluppo rurale.
Le risorse passano allo sviluppo rurale con due diversi criteri: il 20% rimane allo Stato membro, mentre il restante 80% torna all'UE e viene ridistribuito secondo "criteri oggettivi" (SAU 65%, occupazione agricola 35%; il PIL pro-capite viene utilizzato come fattore di ponderazione).
L'ultimo cardine fondamentale della riforma è legato al concetto di condizionalità (che la riforma Fischler punta a rafforzare): il sostegno pubblico è *condizionato* al rispetto di standard ambientali, di sicurezza alimentare, di salute e benessere degli animali e di salute delle piante; soltanto il rispetto di queste norme garantisce il pagamento completo degli aiuti finanziari (diviene una sorta di "scambio" tra maggiori vincoli e accesso ai finanziamenti pubblici).
L'obiettivo dello strumento è di introdurre nuovi requisiti non (ancora) contemplati dalla legge e migliorare il rispetto delle norme legali vigenti; nella riforma ogni agricoltore beneficiario di pagamenti diretti è tenuto a :

- rispettare i criteri di gestione obbligatori (CGO), 18 atti comunitari in materia ambientale, di sicurezza alimentare, di salute e benessere degli animali e delle piante con un applicazione graduale dal 2005 al 2007.
- Mantenere i terreni agricoli in buone condizioni agronomiche ambientali (BCAA): hanno il compito di contrastare l'abbandono delle superfici conseguente al disaccoppiamento degli aiuti attraverso direttive relative all'erosione del suolo, ai livelli di presenza di sostanza organica minima nello stesso.

L'osservanza dei CGO e della BCAA riguarda tutta la superficie agricola dell'azienda, comprese le terre messe a riposo e quelle impiegate per attività che non comportano ottenimento di un pagamento diretto; il mancato rispetto delle norme comporta una riduzione degli aiuti diretti fino alla completa esclusione dagli aiuti stessi (D'Andrea, 2006).

2.3 La situazione italiana e le sovvenzioni alle imprese agricole

L'universo delle aziende agricole italiane è composto per il 76% da aziende con una superficie agricola di meno di 5 ettari, le quali coprono meno del 20% del totale delle terre coltivate in Italia. Mentre le aziende che dispongono di oltre 20 ettari, pur essendo solo meno del 5% , controllano quasi il 60% della superficie agricola.

Le piccole e grandi imprese non si differenziano soltanto dal punto di vista tecnologico (quindi rispetto alla possibilità per una grande impresa agricola di acquistare macchine che una piccola impresa non può permettersi), ma la differenza di *scala* si traduce senza dubbio in un vantaggio a favore delle grandi imprese: si pensi non solo ai vantaggi in termini di diminuzione dei costi unitari ma anche ad altri ambiti fondamentali per la vita dell'impresa, come il fatto di ottenere prezzi più bassi negli acquisti delle materie prime e dei semilavorati (economie di scala), vantaggi dal punto di vista dell'accesso al credito (il fatto anche di poter disporre di un capitale proprio di dimensioni rilevanti permette alle grandi imprese di poter fruire di condizioni più vantaggiose in termini di finanziamenti o anche all'autofinanziamento). Le piccole imprese dal canto loro possono raggiungere un più alto grado di flessibilità che permette loro di offrire prodotti più personalizzati, possono inoltre aderire a consorzi o a società cooperative (Cioccolo *et al.*, 2004).

In un quadro italiano in cui un' impresa vive mediamente quasi 12 anni, 1 impresa su 4 chiude entro 3 anni di vita e oltre 4 su 10 nei primi 5 anni (Infocamere, 2002), diviene di fondamentale importanza riuscire ad individuare quelle aziende agricole che, grazie agli aiuti diretti provenienti dalla PAC, possano rimanere in "vita" garantendo l'autosufficienza per i principali generi alimentari e quindi in una logica di mercato e di flussi di denaro siano le scelte più efficaci ed efficienti.

3. I dati

I dati che andremo ad utilizzare in questa ricerca provengono da un dataset che integra dati sulla sopravvivenza delle imprese agricole venete, ricavati dagli archivi REA (Repertorio delle notizie Economiche e Amministrative) del sistema delle CCIAA con i dati ISTAT del V Censimento Generale dell'Agricoltura del 2000.
La Direzione del Sistema Statistico Regionale Veneto, in possesso dei questionari del Censimento compilati dalle aziende localizzate in Regione, ha provveduto all'operazione di abbinamento con la quale si sono collegate in maniera rigorosamente anonima informazioni provenienti dal Registro REA con quelle di fonte ISTAT.
Oggetto di studio è la popolazione delle imprese che soddisfano i seguenti requisiti:

- Essere iscritte negli archivi REA a fine 1999
- Essere state rilevate nel Censimento del 2000

Di questa popolazione viene ricostruita la sopravvivenza negli archivi REA sino alla fine del 2004.
Per una completa ed esauriente trattazione di questo argomento e quindi anche dell'abbinamento dei dati del Registro Imprese e dello stesso Censimento è utile fare riferimento a Biasiolo (2006) e Bassi *et al.* (2010).
Il registro delle Imprese presso le Camere di Commercio è un'anagrafe giuridico – economica completamente informatizzata, le imprese iscritte alla Camera di Commercio sono obbligate a iscriversi anche negli archivi REA; questi si prestano bene all'analisi di sopravvivenza poiché le denuncie da effettuare al REA devono essere presentate entro 30 giorni dalla manifestazione dell'evento denunciato.
L'unità di osservazione finale è l'azienda agricola come definita dal Censimento: *una singola unità tecnico-economica costituita da terreni, anche in appezzamenti non contigui, ed eventualmente da impianti ed attrezzature varie, in cui si attua la produzione agraria ad opera di un conduttore, cioè persona fisica, società od ente*

che ne sopporta il rischio sia da solo (conduttore coltivatore e conduttore con salariati e/o compartecipanti), sia in associazione (Istat, 2002).

La tabella 3.1 presenta il risultato finale dell'abbinamento dell'archivio REA Veneto e del V Censimento dell'agricoltura, avvenuto tramite utilizzazione come chiave identificativa di Codice Fiscale o Partita IVA. Ne risulta che il tasso di abbinamento globale è pari al 79.03%, con più di 90000 aziende identificate in entrambi gli archivi ed utilizzabili quindi per le analisi.

	Numero Aziende Agricole Venete	Percentuale sul totale
Abbinamento tramite Codice Fiscale	88561	77.1%
Abbinamento tramite Partita IVA	2340	2.03%
Abbinamento totale	90901	79.03%
Mancato abbinamento con dati CCIAA	23958	20.86%
Totale	114859	100%

Tabella 3.1: *Abbinamento fra Censimento del 2000 e archivio Rea*

3.1 Le variabili

Il Censimento ha permesso di raccogliere informazioni su vari aspetti dell'azienda agricola, le singole sezioni del questionario riguardano notizie (Istat, 2002):

- di carattere generale sull'azienda (sistema di conduzione, forma giuridica, svolgimento di attività di vendita dei prodotti, etc…);
- sull'utilizzazione dei terreni (nell'annata agraria 1 novembre 1999 – 31 ottobre 2000) per le coltivazioni principali e la secondaria successiva (seminati, coltivazioni legnose agrarie, etc..), in particolare sulla vite, in ottemperanza all'apposito regolamento comunitario;

- sugli impianti di irrigazione, sui fabbricati rurali, sugli altri impianti e sulle abitazioni situate nell'azienda;
- sugli allevamenti (consistenza, tipologia, ricoveri per animali, produzione di latte, etc..);
- sull'utilizzazione di mezzi meccanici e sulle loro modalità di utilizzo (come il contoterzismo);
- sulle caratteristiche della forza lavoro impiegata in azienda;
- sull'adozione di pratiche di agricoltura biologica, sulle produzioni di qualità, sugli effetti ambientali dell'attività aziendale, etc…;
- sulle modalità di acquisto dei mezzi tecnici, sullo svolgimento di attività connesse all'agricoltura, sulla commercializzazione dei prodotti, sull'utilizzo di attrezzature informatiche;
- sull'estensione delle superfici utilizzate per le diverse coltivazioni e sulla consistenza degli allevamenti di alcune tipologie di bestiame (ciò ha permesso di derivare una mappa estremamente dettagliata dell'utilizzo del territorio, fornendo preziose indicazioni sulla destinazione dei singoli terreni e, quindi, sull'effettiva dislocazione delle diverse attività agricole).

La tabella 3.2 descrive le variabili considerate come possibili regressori nelle analisi svolte successivamente e contiene sostanzialmente la maggior parte delle informazioni presenti nel questionario.

Variabile	**Descrizione**	Modalità
cond_dir	Forma di conduzione	1 se diretta 0 altrimenti
sup_sau_azienda	Superficie totale dell'azienda in are	
sup_sau_tot2	(Superficie totale dell'azienda)^2	
senza_sup		1 se sup_sau_azienda=0
affitto	Titolo di possesso dei terreni	Binaria
uso-grat	Titolo di possesso dei terreni	Binaria
proprietà	Titolo di possesso dei terreni	Categoria di riferimento
azienda_individuale	Forma giuridica	1 se azienda individuale 0 altrimenti
val_prod_vend_meno10	Valore dei prodotti venduti in	Binaria

m	migliaia di euro	
val_prod_vend_tra_10_50m	Valore dei prodotti venduti in migliaia di euro	Binaria
val_prod_piu50m	Valore dei prodotti venduti in migliaia di euro	Categoria di riferimento
ades_consorzi	Adesione a consorzio agrario o di imprese	Binaria
adesione_soc_coop	Adesione a società cooperativa	Binaria
adesione_ass_prod	Adesione a associazioni produttori	Binaria
parchi	Rientra nei parchi o aree protette	Binaria
perc_altra	% superficie sulla sau	
perc_arbo	% superficie sulla sau	
perc_barb	% superficie sulla sau	
perc_boschi	% superficie sulla sau	
perc_cer	% superficie sulla sau	
perc_fiori	% superficie sulla sau	
perc_foraggi	% superficie sulla sau	
perc_frutta	% superficie sulla sau	
perc_legno	% superficie sulla sau	
perc_legumi	% superficie sulla sau	
perc_olivo	% superficie sulla sau	
perc_orti	% superficie sulla sau	
perc_ortive	% superficie sulla sau	
perc_patata	% superficie sulla sau	
perc_piante	% superficie sulla sau	
perc_prati	% superficie sulla sau	
perc_sanu	% superficie sulla sau	
perc_vite	% superficie sulla sau	
perc_vivai	% superficie sulla sau	
serre	Superficie in are	
bovini	Numero capi	
ovicapri	Numero capi ovini e caprini	
equini	Numero capi	
suini	Numero capi	
allev_avicoli	Numero capi	
conigli	Numero capi	
sesso	Sesso del conduttore	

eta	Età del conduttore	
eta2	(Età del conduttore)^2	
cond_prof_condu	Condizione professionale del condottore	1 se occupato, 0 altrimenti
lav_0_30	Num. di giornate di lavoro all'anno del conduttore	1 se fino a 30 gg, 0 altrimenti
lav_30_90	Num. di giornate di lavoro all'anno del conduttore	1 se tra 30 e 90 gg, 0 altrimenti
lav_90_270	Num. di giornate di lavoro all'anno del conduttore	1 se tra 90 e 270 gg, 0 altrimenti
lav_270	Num. di giornate di lavoro all'anno del conduttore	Categoria di riferimento
att_rem_extraz	Attività remunerativa extraziendale del conduttore	1 se ne svolge 0 altrimenti
lav_fam	Se è impiegata prevalentemente manodopera familiare	1 se si 0 altrimenti
capo_azienda	Se il conduttore è anche capo azienda	Binaria
lic_scu_elem	Titolo di studio del capo-azienda	Categoria di riferimento
lic_scu_inf	Titolo di studio del capo-azienda	Binaria
diploma	Titolo di studio del capo-azienda	Binaria
laurea	Titolo di studio del capo-azienda	Binaria
prod_bio	Agricoltura biologica vegetale e zootecnica	Binaria
lavoraz_prod_agric	Lavorazione di prodotti agricoli	Binaria
bel	Provincia in cui ha sede l'azienda	Binaria
pad	Provincia in cui ha sede l'azienda	Binaria
ven	Provincia in cui ha sede l'azienda	Binaria
vic	Provincia in cui ha sede l'azienda	Binaria
ver	Provincia in cui ha sede l'azienda	Binaria
rov	Provincia in cui ha sede l'azienda	Binaria
tv	Categoria di riferimento	
cod_att101	ATECO=1.1 coltivazioni agricole,orticultura,floricoltura	1 se 1.1 0 altrimenti
cod_att102	ATECO=1.2 allevamento di animali	1 se 1.2 0 altrimenti
cod_att103	ATECO=1.3 coltivazioni agricole associate all'allevamento di animali	1 se 1.3 0 altrimenti
cod_att104	ATECO=1.4 servizi connessi	1 se 1.4

	all'agricoltura e alla zootecnia	0 altrimenti
cod_att111	ATECO=1.11 coltivazioni di cereali e altri seminativi	1 se 1.11 0 altrimenti
cod_att112	ATECO=1.12 coltivazione di ortaggi	1 se 1.12 0 altrimenti
Ote1	Aziende specializzate nei seminativi	1 se cod.Ote I cifra=1 0 altrimenti
Ote2	Aziende specializzate in ortofloricoltura	1 se cod.Ote I cifra=2 0 altrimenti
Ote3	Aziende specializzate in viticoltura	1 se cod.Ote I cifra=3 0 altrimenti
Ote311	Aziende specializzate in viticoltura da vino D.O.C.	1 se cod. Ote I,II,III cifra=311, 0 altrimenti
Ote312	Aziende specializzate in viticoltura da vino comune	1 se cod. Ote I,II,III cifra=312, 0 altrimenti
Ote4	Aziende specializzate in erbivori	1 se cod. Ote I cifra=4 0 altrimenti
Ote5	Aziende specializzate in granivori	1 se cod. Ote I cifra=5 0 altrimenti
Ote6	Aziende specializzate in policoltura	1 se cod. Ote I cifra=6 0 altrimenti
Ote7	Aziende con poliallevamento	1 se cod. Ote I cifra=7 0 altrimenti
Ote8	Aziende miste coltivazioni-allevamento	1 se cod. Ote I cifra=8 0 altrimenti
Ote9	Aziende non classificabili	Categoria di riferimento
ude_2	Codice ude	Categoria di riferimento
ude2_4	Codice ude	1 se 2≤ ude≤ 4
ude4_6	Codice ude	1 se 4≤ ude≤ 6
ude6_8	Codice ude	1 se 6≤ ude≤ 18
ude8_12	Codice ude	1 se 8≤ ude≤ 12
ude12_16	Codice ude	1 se 12≤ ude≤ 16
ude16_40	Codice ude	1 se 16≤ ude≤ 40
ude40_100	Codice ude	1 se 40≤ ude≤ 100
ude100	Codice ude	1 se ude≤100

Tabella 3.2 *: Regressori utilizzati per i modelli*

3.2 Alcune analisi descrittive

In questa sezione andiamo a descrivere alcune delle variabili utilizzate nel dataset. Nella Figura 3.1, che mostra la distribuzione dell'età delle aziende agricole venete sopravvissute al censimento del 2000, è evidente che la maggior parte di queste ha un'età sostanzialmente non superiore ai 30 anni. Si nota un picco in corrispondenza dei 27 anni, questo perché circa il 18% delle aziende hanno la stessa data di inizio attività (1/1/1973) a causa dell'entrata in vigore del d.p.r.633/1972 che disciplina l'Imposta sul Valore Aggiunto (IVA).

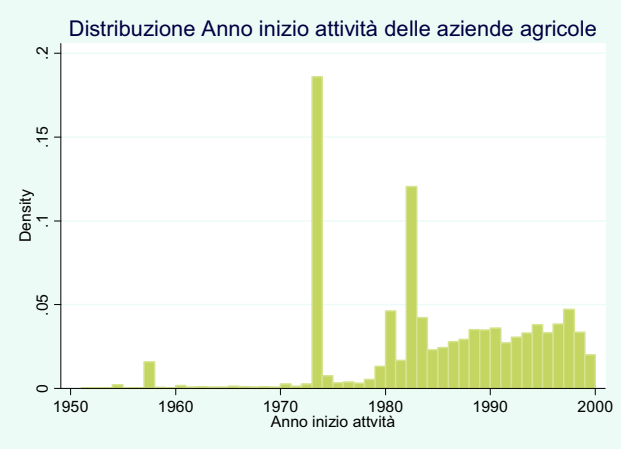


Figura 3.1: *Distribuzione dell'anno di inizio attività delle aziende agricole Venete al censimento del 2000*

Al fine di semplificare il contesto per le analisi empiriche successive, si è deciso di andare tuttavia a prendere in considerazione non tutte le aziende agricole Venete, ma solo la "coorte" di quelle aziende recenti, cioè nate nel 1999 e quindi le ultime aziende censite considerando l'anno di nascita. Per un trattazione descrittiva esaustiva di tutte le aziende del dataset è utile fare riferimento a Biasiolo (2006) e Bassi *et al.* (2010).

Le aziende agricole venete nate nell'ultimo anno possibile (1999) all'epoca del Censimento del 2000 sono 1804, la cui distribuzione geografica è rappresentata in Figura 3.2.

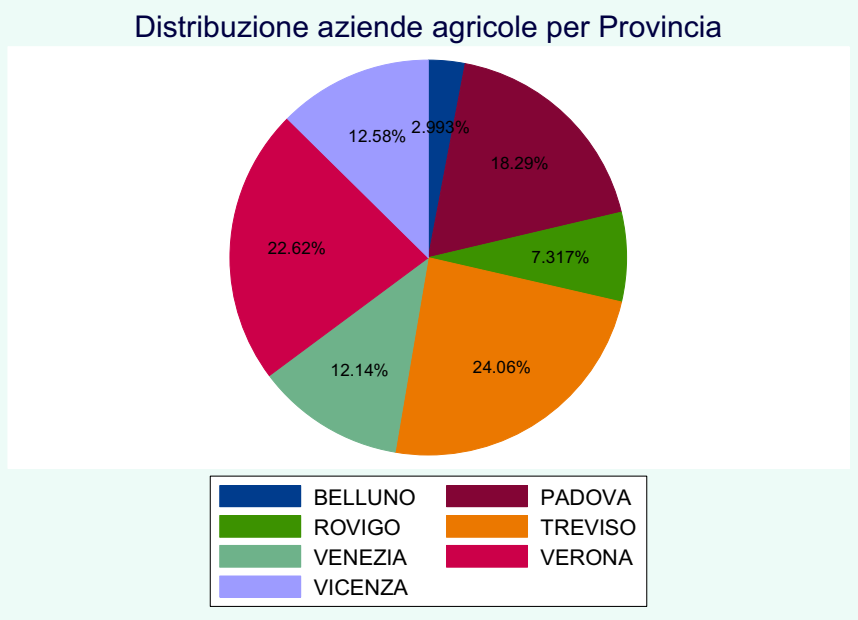


Figura 3.2: *Distribuzione Aziende Agricole Venete nate nel 1999 suddivise per Provincia*

Come si era visto in precedenza, grazie alla connessione con l'archivio Rea, possiamo anche sapere la distribuzione in termini di sopravvivenza di queste stesse aziende in un arco di tempo che va dal 2000 al 2004. Dalla Figura 3.3 si nota che nell'arco dei primi 5 anni di vita, la mortalità aziendale è pari al 23.84%.

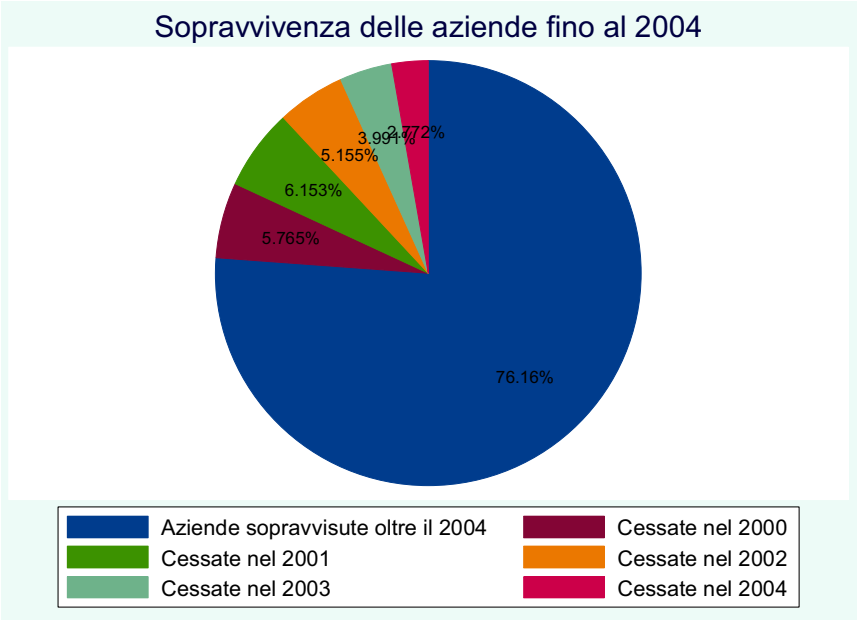


Figura 3.3: *Distribuzione in termini di sopravvivenza delle aziende agricole nate nel 1999 nel periodo 2000-2004*

Una variabile rilevante nel panorama delle aziende agricole Venete senza dubbio è l'età dei conduttori delle stesse; sostanzialmente nell'intero dataset l'età media è pari a 59 anni. L'età dei conduttori invece nel dataset ridotto è sensibilmente più bassa e si attesta intorno ai 49 anni, la distribuzione è descritta in Figura 3.4.

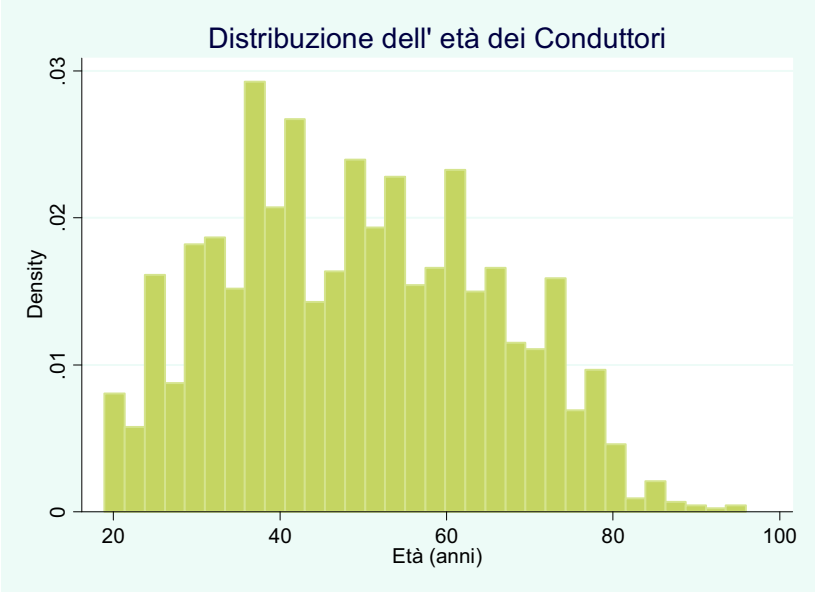


Figura 3.4*: Distribuzione dell'età dei conduttori delle aziende agricole venete nate nel 1999*

E' interessante verificare anche il livello di istruzione del conduttore rappresentato in Figura 3.5: la letteratura economica assegna al capitale umano un ruolo prioritario e di motore della crescita e di sviluppo economico. Il livello di istruzione produce due effetti: uno interno, di aumento della produttività del lavoro, derivante dall' accresciuta abilità ed efficienza del lavoratore più istruito; il secondo è un'esternalità che consiste nel miglioramento della produttività media di tutti i lavoratori coinvolti nell'attività produttiva. Verifiche empiriche dell'effetto del capitale umano sulla produttività delle risorse impiegate nell'agricoltura italiana sono state già effettuate e da queste è noto come il livello di capitale umano nell'agricoltura italiana sia basso rispetto ad altri settori (si pensi che dal censimento del 2000 risulta che solo 1% delle aziende agricole possiede attrezzature informatiche) (De Devitiis *et al.,* 2009).

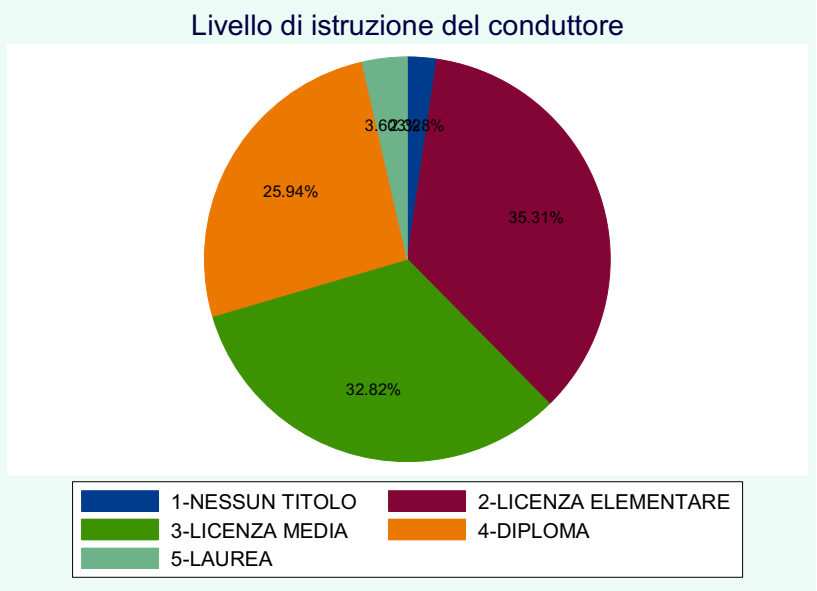


Figura 3.5: *Distribuzione del livello di istruzione del conduttore nelle aziende agricole venete nate nel 1999.*

Dalla Figura 3.5 si desume un sostanziale aumento del livello di istruzione medio del conduttore se confrontato con quello delle aziende venete nate tra gli anni ’70 e ’80, in cui ben più del 90% dei proprietari presentava al massimo una licenza media (circa il 70% possedeva la licenza elementare). L’aumento del livello di istruzione nel mondo agricolo è comunque implicito nell’aumento dell’istruzione media della popolazione italiana.

Tra le variabili presenti nel dataset di sicuro rilievo è quella che misura la SAU (*Superficie agricola utilizzata*) di ogni azienda presa in esame: l’unità di misura utilizzata storicamente è l’*ara* (100 *are*=1 *ettaro*). Lo sviluppo economico della regione Veneto ha comportato negli ultimi 30 anni una profonda trasformazione dell’assetto territoriale, con la sottrazione alla SAU di suoli destinati a processi di urbanizzazione e industrializzazione a carattere diffuso. Tale fenomeno di consumo del territorio ha profondamente inciso sul settore produttivo agricolo, sottraendo campagne sempre più vaste alla loro storica funzione e provocando profonde ed irreversibili mutazioni di paesaggi e contesti territoriali: dai dati del censimento risulta infatti che negli ultimi 30 anni il territorio veneto ha visto ridursi di 138520 ettari la SAU.

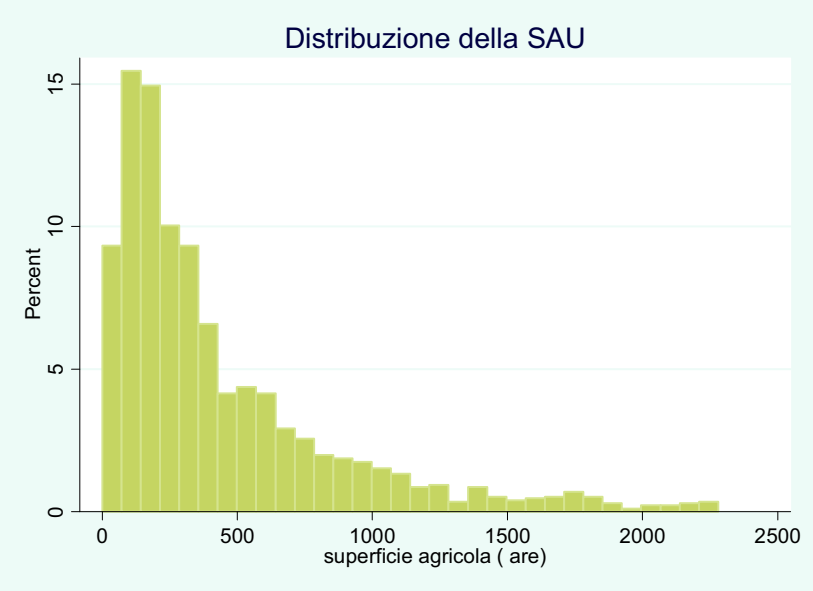


Figura 3.6 : *Distribuzione della superficie agricola utilizzata dalle aziende agricole venete nate nel 1999*

Dalla Figura 3.6 si nota come il 90% delle neonate aziende presenti una superficie dedicata allo sfruttamento agricolo inferiore ai 13 ettari, e il 75% inferiore ai 6 ettari. La variabile presenta una forte componente di variabilità legata alla presenza di outlier (il valore più grande è rappresentato da un'azienda la cui estensione della SAU è pari a 390 ettari), la media ne risente e si attesta intorno ai 6,5 ettari.

Di indubbia importanza nella descrizione economica di queste aziende è la variabile UDE *(Unità di Dimensione Economica)*: la dimensione economica infatti è altamente correlata con il livello di redditività delle aziende agricole e misura la capacità dell'azienda di impiegare efficientemente le risorse. Questo parametro rappresenta anche un pre-requisito per l'accesso agli aiuti per gli investimenti aziendali. Infatti il Regolamento della Comunità europea per il sostegno allo sviluppo rurale impone che gli aiuti possano essere concessi unicamente ad aziende che superano determinate soglie di redditività.

L'UDE rappresenta la base per il calcolo della dimensione economica aziendale: una UDE corrisponde ad un *reddito lordo standard* aziendale di 1200 euro l'anno. In quest'ottica la Commissione delle Comunità Europee ha creato delle classi di dimensione economica: la soglia di 16 UDE viene generalmente utilizzata per distinguere le piccole aziende (UDE<16) da quelle di medie e grandi dimensioni.

Il *reddito lordo standard* (RLS) esprime, in termini monetari, la differenza fra il valore della produzione lorda e l'importo dei costi specifici sostenuti per ottenere tale produzione; questa differenza viene determinata per ogni singola specie vegetale o animale. I redditi lordi standard esprimono un valore medio applicabile a tutte le aziende ricadenti in un determinato territorio che per l'Italia è identificato nella Regione. In questo modo è possibile determinare *l'orientamento tecnico- economico* (OTE) delle aziende agricole, in base all'incidenza percentuale delle varie attività produttive rispetto al *reddito lordo standard complessivo.* La classificazione delle aziende agricole per OTE si basa quindi sulla determinazione del peso economico delle varie attività produttive presenti in azienda e sulla loro classificazione. A tal fine, utilizzando i RLS della zona in cui ricade l'azienda, si moltiplicano gli ettari coltivati o il numero dei capi allevati per il corrispondente RLS: la combinazione ottenuta si confronta con uno schema tipologico che serve ad individuare gli OTE secondo criteri a livello comunitario e validi per tutte le statistiche ufficiali. Un'azienda viene definita specializzata quando il RLS di una o più attività produttive affini supera i 2/3 del RLS totale dell'azienda (Istat, 2002).

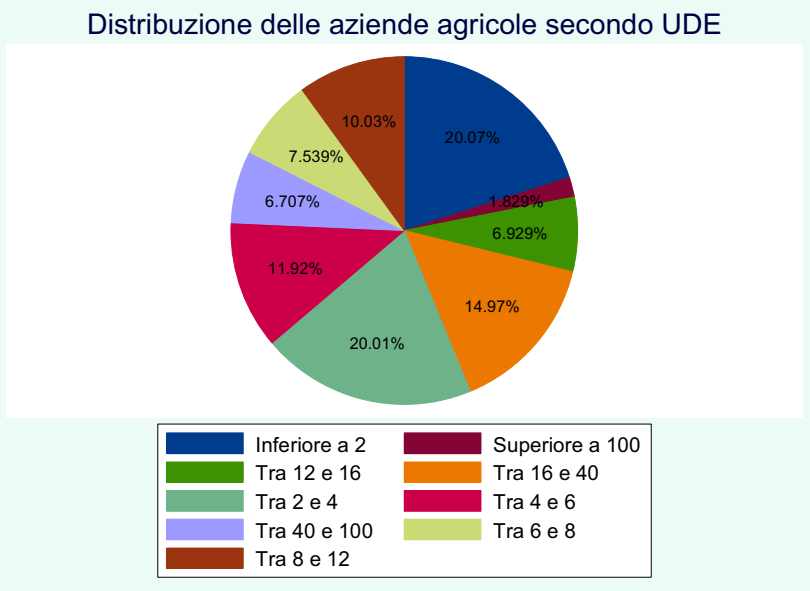


Figura 3.7: *Distribuzione delle aziende agricole secondo classi di UDE (unità di dimensione economica)*

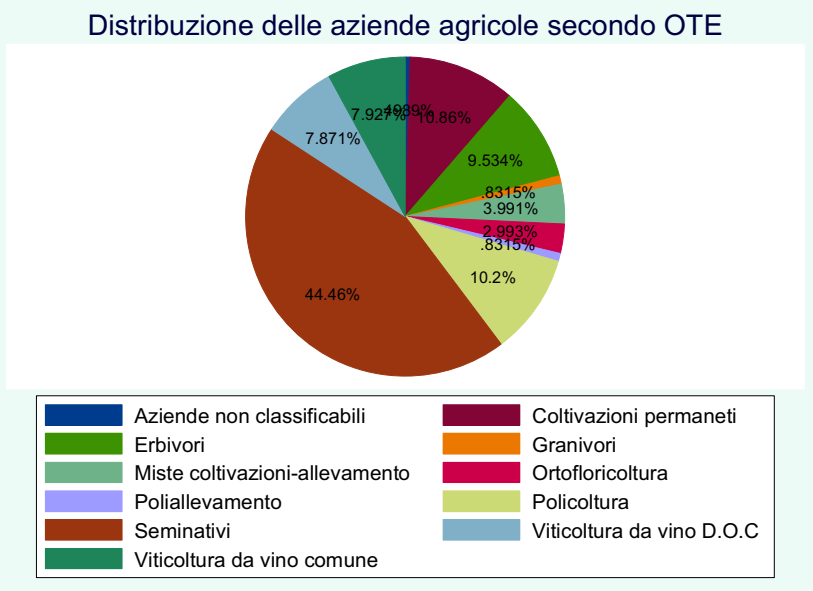


Figura 3.8: *Distribuzione della aziende agricole secondo la variabile OTE (Orientamento Tecnico-Economico)*

Come si vede dalle Figure 3.7 e 3.8, in Veneto al censimento del 2000 le aziende agricole nate lungo il 1999 per circa il 23.5% presentano una dimensione economica superiore o uguale alle 16 UDE: ci troviamo di fronte, quindi, a un panorama regionale che mostra una forte e radicata presenza di piccole-medie imprese come tradizionalmente avviene nel Nord Est in ambito imprenditoriale. Per quanto riguarda la specializzazione (e quindi facendo riferimento al codice OTE), si nota una forte presenza di aziende specializzate in Seminativi (circa il 45%), seguite da aziende specializzate in Policoltura e Coltivazioni permanenti. Le aziende specializzate nella viticoltura si suddividono equamente tra viticoltura specializzata in vini D.O.C. e viticoltura specializzata in vini comuni, e complessivamente catturano circa il 15% delle neonate aziende.

Per una trattazione più dettagliata dal punto di vista economico, possiamo andare a distinguerle per classi di fatturato relativamente al valore dei prodotti venduti. Dalla Tabella 3.3 è evidente come sostanzialmente la maggior parte delle aziende si suddivida per valore della produzione tra la classe più alta e quella più bassa, infatti ben il 43% dichiara un valore inferiore ai 5mila euro e circa il 44% dichiara un valore superiore ai 25mila euro.

Valore della produzione		*Ripartizione*
Inferiore ai 5mila euro		43.01%
Tra i 5mila e i 25mila euro		12.64%
Superiore ai 25mila euro		44.35%

Tabella 3.3: *Valore della produzione venduta suddivisa per classi di fatturato*

4. Modelli di durata a tempi discreti

I modelli utilizzabili con i dati in nostro possesso rientrano nella classe dei modelli di durata a tempi discreti. Non avendo informazioni precise sulla durata di un'azienda agricola che riguardano intervalli di tempo più brevi come settimane o mesi, ma solamente l'anno di inizio attività e cessazione della stessa, questa tipologia di modelli risulta particolarmente adatta alla stima delle probabilità di sopravvivenza.
Al fine di poter realizzare questa tipologia di analisi risulta necessario attuare una trasformazione della struttura dei dati stessa in formato *episode-splitting*.
Nella Tabella 4.1 la variabile *id_azienda* rappresenta il codice identificativo dell'azienda presa in considerazione sia nella vecchia struttura come nella nuova. Nella nuova struttura andiamo sostanzialmente ad aggiungere per ogni identificativo di azienda tante righe quanti sono gli anni in cui effettivamente l'azienda è sopravvissuta nell'arco dei 5 anni presi in considerazione dalla propria nascita.

Struttura originale				**Struttura in formato episode-splitting**			
id_azienda	y_i	x_i	anni_osservazione	id_azienda	y_{ij}	x_i	anno j
5	0	x_5	5	5	0	x_5	1
				5	0	x_5	2
				5	0	x_5	3
				5	0	x_5	4
				5	0	x_5	5
6	1	x_6	3	6	0	x_6	1
				6	0	x_6	2
				6	1	x_6	3
7	1	x_7	4	7	0	x_7	1
				7	0	x_7	2
				7	0	x_7	3
				7	1	x_7	4

Tabella 4.1: *Dataset in formato originale e dataset modificato in formato episode splitting*

Ad esempio nella vecchia struttura l'azienda 6 è stata osservata per 3 anni in quanto al terzo anno ha cessato l'attività; y_i è una variabile risposta binaria che indica semplicemente se l'azienda ha cessato l'attività (y_i=1) o se è ancora attiva (y_i=0). Nella nuova struttura l'azienda 6 essendo stata osservata per 3 anni compare con 3 record: y_{ij} in questo caso assume valore 0 per i primi due anni di osservazione e 1 al terzo in quanto ha cessato l'attività. Nel nostro caso le caratteristiche aziendali misurate dalle covariate risultano costanti nel tempo, in quanto non si dispone di informazioni su eventuali variazioni, è evidente però che il metodo proposto consente l'introduzione di variabili *time-varying*.

Nel caso in cui i tempi siano discreti, il tempo di sopravvivenza T è una variabile casuale discreta con funzione di probabilità (Jenkins, 2004):

$$f(j) \equiv f_j = \Pr(T = j)$$

dove *j* varia nel set di possibili valori dell'anno di sopravvivenza (ex: 1,2,3…..).

Si ricorda che il fatto di limitarsi alla coorte di aziende del 1999 consente di osservare l'ingresso di tutte le unità, eliminando possibili problemi di *length bias.*

In generale la funzione di sopravvivenza e quella di rischio per l'istante *j* possono essere scritte rispettivamente come:

$$S(j) = \Pr(T \geq j) = \sum_{k=j}^{\infty} f_k$$

$$h(j) = \Pr(T = j | T \geq j) = \frac{f(j)}{S(j-1)}$$

La probabilità di sopravvivere fino all'intervallo *j* è il prodotto delle probabilità di sopravvivere in tutti gli intervalli precedenti incluso quello corrente.

$$S_j = S(j) = (1-h_1)(1-h_2)\ldots\ldots(1-h_{j-1})(1-h_j)$$
$$= \prod_{k=1}^{j}(1-h_k)$$

Nel nostro specifico caso andremo quindi a stimare delle probabilità p_{ij} di sopravvivenza anno per anno *(j)* di ogni azienda *(i)*; nella nuova struttura sopra descritta possiamo anche andare a introdurre delle variabili *dummy* che definiscono i diversi anni, con l'obiettivo di verificare la variazione del rischio di cessazione durante il periodo di osservazione, o dipendenza dalla durata.

Dopo aver ottenuto le stime dei coefficienti, è possibile calcolare la probabilità di cessazione per l'intero periodo di riferimento partendo dalle stime delle probabilità di cessazione annue calcolate per tutti gli anni in cui un'azienda è stata a rischio di cessazione (se un'azienda è sopravvissuta per l'intero periodo si avranno 5 diverse probabilità, 4 se è sopravvissuta per 4 anni etc…).

Se un'azienda *i* è sopravvissuta dal 2000 al 2004, per *t* anni, con $t \leq 5$ nel dataset in formato *episode splitting* si avranno solo *t* record in corrispondenza dell'*i*-esima azienda, con *t* che varia da 1 a 5.

Questo è corretto per ottenere le stime dei coefficienti, ma se si vuole calcolare la probabilità di sopravvivenza per 5 anni è necessario avere 5 record per ogni azienda: verranno a questo scopo creati 5-*t* record con gli stessi valori che assumono le variabili per l'azienda *i*-esima (Biasiolo, 2006), andando a ottenere le stime di probabilità annue per gli eventuali anni mancanti.

Se l'obiettivo quindi diventa quello di stimare la probabilità di cessazione dell'attività per l'intero periodo di 5 anni, questa sarà:

$$\Pr(cessazione_i = 1) = 1 - \prod_{j=1}^{5}\left[1 - Pr\left(y_{ij} = 1\right)\right]$$

4.1 L'eterogeneità non osservata e il modello ad effetti casuali

La classe dei modelli ammette che le differenze tra le aziende agricole non siano catturate completamente dal vettore delle esplicative X, ma si voglia provare a prendere in considerazione la componente di eterogeneità non osservata tra i soggetti. Si ammette la presenza quindi di variabili latenti (effetti casuali) che non

sono osservate direttamente ma hanno influenza sulla sopravvivenza o meno dell'azienda agricola.

L'approccio classico prevede l'uso in questi caso di *modelli a effetti casuali con risposta binaria* per dati di panel (ad esempio, tramite il software STATA, possiamo stimare tali modelli utilizzando la funzione *xtlogit*).

Per tenere conto di questi aspetti è quindi necessario considerare un modello generale del tipo:

$$\frac{h(j,X|\zeta)}{1-h(j,X|\zeta)} = \left[\frac{h_0(j)}{1-h_0(j)}\right] exp(\beta' X + \zeta)$$

$$logit[h(j,X|\zeta)] = D(j) + \beta' X + \zeta$$

dove $D(j)$ è un termine che caratterizza la funzione di rischio di base, $\beta' X$ cattura gli effetti delle variabili osservate, mentre ζ è il termine d'errore inserito per considerare l'eterogeneità non osservata, con distribuzione Normale $(0, \sigma_\zeta^2)$ (Jenkins, 2004).

Nello specifico, un modello per dati di panel con variabile esplicativa dicotomica con effetti casuali è descritto da Wooldridge (2002) :

$$Pr(y_{it} \neq 1|x_i, \zeta_i) = Pr(y_{it} \neq 1|x_{it}, \zeta_i) = \pi(x_{it}\beta + \zeta_i) = \frac{exp(x_{it}\beta+\zeta_i)}{1+exp(x_{it}\beta+\zeta_i)} \quad , t=1,\ldots\ldots,ni$$

dove ζ_i rappresenta l'effetto non osservato e $\boldsymbol{x_i}$ contiene $\boldsymbol{x_{it}}$ per tutti i valori di *t*.

Si assuma inoltre :

(1) $y_{i1}, y_{i2}, \ldots\ldots\ldots\ldots y_{it}$,indipendenti condizionatamente a (x_i, ζ_i)

(2) $f(\zeta_i|\boldsymbol{x}_i) \sim N(0, \sigma_\zeta^2)$

Sotto queste assunzioni si può derivare la densità di $(y_{i1}, \dots\dots, y_{in_i})$ condizionatamente ai valori di (x_i, ζ_i):

$$f(y_1, \dots\dots, y_{n_i}|x_i, \zeta_i; \beta) = \prod_{t=1}^{n_i} f(y_t|x_{it}, \zeta, \beta)$$

$$f(y_t|x_{it}, \zeta, \beta) = \pi(x_t\beta + \zeta)^{y_t}[1 - \pi(x_t\beta + \zeta)]^{1-y_t}$$

Sfruttando queste equazioni, si possono stimare i parametri β e σ_ζ^2 utilizzando il metodo della verosimiglianza condizionata. Poiché i valori di ζ_i non vengono osservati, non possono comparire nella funzione di verosimiglianza. Si ottiene la distribuzione congiunta di $(y_{i1}, \dots\dots, y_{in_i})$ condizionatamente a $\boldsymbol{x_i}$ integrando la funzione in ζ_i. Utilizzando queste assunzioni si ha:

$$Pr(y_{i1}, \dots\dots y_{in_i}|x_{i1}, \dots\dots x_{in_i}) = \int_{-\infty}^{\infty} \frac{e^{-\zeta_i^2/2\sigma_\zeta^2}}{\sqrt{2\pi}\sigma_\zeta} \left\{ \prod_{t=1}^{n_i} (y_{it}, x_{it}\beta + \zeta_i) \right\} d\zeta_i$$

dove

$$f(y_{it}|x_{it}, \zeta_i, \beta) = \pi(y_{it}, x_{it}\beta + \zeta_i)^{y_{it}}[1 - \pi(y_{it}, x_{it}\beta + \zeta_i)]^{1-y_{it}}$$

Tramite la funzione *xtlogit* di Stata andiamo proprio a stimare questo specifico modello; la versione 9.2 del programma inoltre calcola una approssimazione dell'integrale sopra descritto utilizzando il metodo della quadratura adattiva di Gauss-Hermite (o Quadratura Gaussiana) che permette di approssimare integrali del tipo :

$$\int_{-\infty}^{\infty} e^{-x^2} g(x)dx \approx \sum_{m=1}^{M} \varpi_m^* g(a_m^*)$$

dove i ω_m^* rappresentano i pesi della quadratura mentre a_m^* rappresentano le ascisse (Stata, 2005). Il metodo sviluppato da Liu, Qing e Pierce (1994) in *xtlogit* adatta la quadratura attraverso la moda e la curvatura della moda della funzione di verosimiglianza.

Il metodo generalmente sembra lavorare molto bene, e spesso meglio di altri metodi alternativi quali MQL (Marginal Quasi-Likelihood) e PQL (Penalized Quasi-Likelihood), ma ci sono casi in cui questo tipo di quadratura ha performance più scarse (Skrondal e Rabe-Hesketh , 2002).

Per quanto riguarda la previsione, essa viene calcolata in ogni caso assumendo assenza di eterogeneità non osservata: alla variabile ζ_i viene assegnato il valore della sua media, cioè $E(\zeta_i) = 0$.

La probabilità stimata per un' unità con valori delle covariate $\boldsymbol{x_{it}}=\boldsymbol{x^0}$ in un ipotetico "cluster" con effetti casuali $\zeta_i = \zeta_i^0$ è data quindi da:

$$\hat{\mu}(x^0, \zeta_i^0) \equiv E_y\left(y_i|\zeta_i^0, x^0; \hat{\beta}\right) = h(x^{0\prime}\hat{\beta} + \zeta_i^0)$$

La funzione di predizione per *xtlogit* essenzialmente pone tale valore dell'effetto casuale (ζ_i^0) = 0.

4.2 Stima e previsione sui dati reali

Sfruttando la funzione *xtlogit* andiamo a stimare sui dati reali un *modello ad effetti casuali per dati di panel* con variabile esplicativa dicotomica, e ne riportiamo i principali risultati.

A fini descrittivi si è deciso di proporre un modello utile alla stima del rischio di cessazione servendosi delle variabili OTE e UDE che, come visto nel paragrafo 3.2, descrivono rispettivamente l'orientamento tecnico-economico delle aziende (e quindi in un certo senso tengono conto di tutti gli aspetti relativi alle tipologie di colture praticate che possono essere discriminanti per la cessazione o meno delle aziende), e la dimensione economica (che praticamente va a inglobare tutti quei parametri relativi alla forza e solidità economica delle aziende stesse che sono sicuramente

determinanti per la sopravvivenza). Tra le variabili esplicative andiamo anche a considerare le variabili dummy relative agli anni di osservazione. Andremo quindi a stimare un modello del tipo:

$$Pr(y_{it} \neq 1|x_{it}, \zeta_i) = \pi(x_{it}\beta + \zeta_i) = \frac{exp(UDE\_2_{it}\beta_1 + UDE\_4\_6_{it}\beta_2 \ldots\ldots\ldots + OTE1_{it}\beta_k + \cdots + \zeta_i)}{1 + exp(UDE\_2_{it}\beta_1 + UDE\_4\_6_{it}\beta_2 \ldots\ldots\ldots + OTE1_{it}\beta_k + \cdots + \zeta_i)}, \qquad t = 1 \ldots\ldots\ldots n_j$$

dove ζ_i rappresenta la variabile latente e quindi l'effetto non osservato. Si assume inoltre che le $y_{i1}, y_{i2} \ldots\ldots\ldots y_{it}$ siano indipendenti condizionatamente a (x_i, ζ_i) e che $f(\zeta_i|x_i) \sim N(0, \delta_\zeta^2)$.

La Tabella 4.2 riporta le stime dei coefficienti e i relativi *Odds Ratio* e *Standard Error*. In queste tipologie di analisi le stime dei parametri possono essere interpretate come effetti dell'incremento unitario della variabile sul *logit*, a parità di altre condizioni. Tuttavia, con fini di ricerca, sembra più utile e facilmente interpretabile valutare i corrispondenti *Odd Ratio (exp(β_j)),* interpretabili come *OR* tra due aziende che differiscono per un incremento unitario della corrispondente covariata. L'*Odd Ratio* che si ottiene valuta quindi l'aumento del rischio di cessazione in corrispondenza di un aumento unitario/presenza di una determinata variabile a parità di altre condizioni. Si può inoltre utilizzare il segno del coefficiente stimato per capire se tale variabile ha un impatto positivo o negativo sul rischio di cessazione. Chiaramente per impatto negativo si intende una diminuzione di tale rischio, per impatto positivo viceversa.

Per una trattazione più esaustiva dei risultati della regressione si rimanda al capitolo 7. In relazione alle stime dei parametri il modello non reputa significative le stime dei parametri OTE, l'orientamento tecnico economico delle aziende agricole sembra quindi non essere rilevante nella determinazione del rischio di cessazione delle imprese nel quinquennio considerato. Le variabili UDE, inerenti invece alla dimensione economica, hanno un effetto significativo sul rischio di cessazione: si nota come questo tenda a diminuire monotonicamente con la crescita della dimensione.

Variabile	*Coefficienti*	*Odds Ratio*	*Standard Error*
Base_Ote1	*0,589147*	*1,80245*	*0,652662*
Ote2	*1,426634**	*4,16465*	*0,719878*
Ote3	*0,077123*	*1,08017*	*0,679006*
Ote311	*-0,70541*	*0,49390*	*0,711546*
Ote312	*0,151713*	*1,16382*	*0,673657*
Ote4	*0,121327*	*1,12899*	*0,671331*
Ote5	*0,184872*	*1,20306*	*1,009592*
Ote6	*0,222773*	*1,24953*	*0,673329*
Ote7	*-0,48726*	*0,61430*	*0,990392*
Ote8	*-0,17981*	*0,83542*	*0,723212*
Ude2_4	*-0,36754***	*0,69243*	*0,139121*
Ude4_6	*-1,20448***	*0,29984*	*0,19435*
Ude6_8	*-1,22975***	*0,29236*	*0,236308*
Ude8_12	*-1,44118***	*0,23664*	*0,226315*
Ude12_16	*-1,71247***	*0,18041*	*0,289442*
Ude16_40	*-2,03393***	*0,13082*	*0,235676*
Ude40_100	*-2,64286***	*0,07115*	*0,433504*
Ude100	*-2,7917***	*0,06131*	*0,745702*
Costante	*-2,45801***	*0,08560*	*0,650684*
Dummy2	*0,220024*$^{\tau}$	*1,24610*	*0,146309*
Dummy3	*0,176803*	*1,19339*	*0,154111*
Dummy4	*0,016379*	*1,01651*	*0,165812*
Dummy5	*-0,27532*$^{\tau}$	*0,75932*	*0,18482*
Numerosità		*8013*	
Log -Likelihood		*-1540.0233*	
$\widehat{\sigma}_{\zeta}^{2}$		*0.297847*	

Tabella 4.2: *Modello di durata con effetti casuali*

Il risultato più interessante che si ottiene utilizzando questa metodologia, è senza dubbio la stima del parametro $\boldsymbol{\rho}$ definito come segue:

$$\rho = \frac{\sigma_{\zeta}^{2}}{\sigma_{\zeta}^{2} + \sigma_{\varepsilon}^{2}}$$

Se ρ=0 c'e' assenza di eterogeneità non osservata, perche questo implica σ_ζ^2 =0.
Stata ci fornisce un test, chiamato Test Lr (Likelihood-ratio test of rho=0), che nel nostro specifico caso fornisce questi risultati:

$\chi^2_{bar}(01)$ = 1.73 $\qquad\qquad$ $Prob >= \chi^2_{bar}(01) = 0.094.$

La distribuzione utilizzata per il test $\chi^2_{bar}(01)$ è una mistura 50:50 di due di distribuzioni: $\chi^2(0)$ e $\chi^2(1)$. Questo perché la distribuzione asintotica di ρ è una normale "troncata" a 0 (Gutierrez et.al, 2001).
Nel nostro specifico caso, a un livello di significatività del 5%, il test accetta l'ipotesi nulla di assenza di eterogeneità non osservata, a un livello di significatività del 10% rifiuta tale ipotesi.
Dopo avere ottenuto le stime dei parametri dalla regressione, queste si utilizzano per prevedere la probabilità di cessazione nei 5 anni come visto all'inizio del capitolo 4.
Dal paragrafo 4.1 si deduce che la previsione con *xtlogit* viene eseguita in Stata 9.2 assumendo che l'effetto casuale abbia un valore pari alla sua media, $E(\zeta_i) = 0$.
Successivamente si valuta la capacità classificatoria del modello comparando la distribuzione di probabilità stimata con quella osservata.

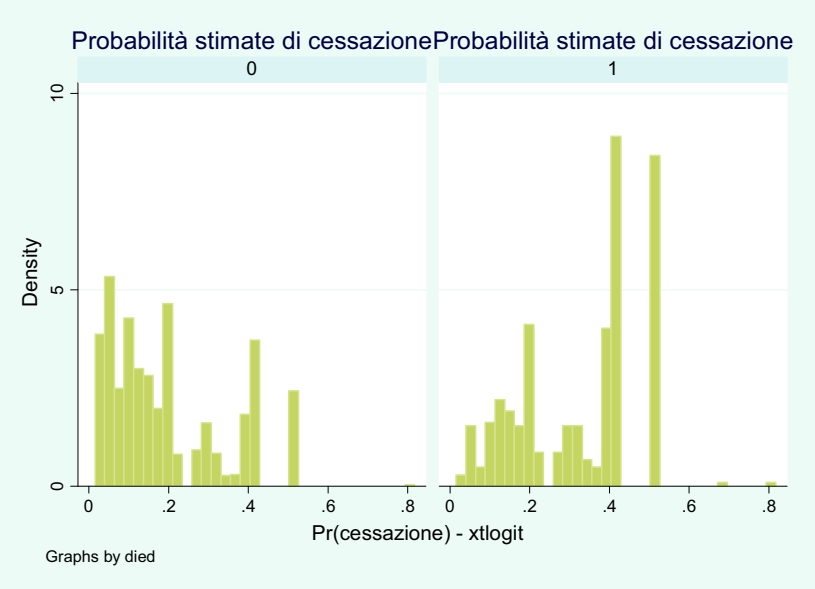


Figura 4.1: *Distribuzione delle probabilità stimate di cessazione per il periodo 2000-2004, aziende sopravvissute (0) e cessate (1)*

La Figura 4.1 riporta le distribuzioni delle probabilità stimate di cessazione per le aziende sopravvissute e per quelle cessate (0= sopravvissuta, 1= cessata) nel periodo di 5 anni (2000-2004).

La capacità classificatoria del modello si può comprendere meglio analizzando i risultati della Tabella 4.3 dove per *proporzione di falsi positivi* intendiamo : Pr(condizione effettiva = *cessata* | condizione prevista = *ancora attiva*), mentre per *proporzione di falsi negativi* intendiamo : Pr(condizione effettiva = *ancora attiva* | condizione prevista = *cessata*).

Soglie	*Corretta classificazione*	*Proporzione Falsi Positivi*	*Proporzione Falsi Negativi*
0.2	62.69%	11.44%	63.57%
0.3	72.23%	14.13%	56.04%
0.4	74.78%	17.48%	53.20%

Tabella 4.3*: Capacità classificatoria con xtlogit per tre livelli di soglia*

Come si nota dalla Tabella 4.3, la capacità classificatoria è influenzata dalla soglia scelta: aumentando la soglia di *cut off* il modello tende a classificare meglio le aziende che cessano l'attività. Per costruzione si ha anche un aumento della proporzione di falsi positivi e una diminuzione di falsi negativi.

Quindi andando ad utilizzare una soglia compresa tra 0.3 e 0.4 (cioè se la probabilità stimata di cessazione è superiore a questa, il modello classifica l'azienda come "cessata"), *xtlogit* fornisce queste prestazioni: poco più del 70% delle aziende è classificato esattamente e la probabilità di finanziare aziende che poi chiuderanno varia tra il 14 e il 18% (*Falsi positivi*). La probabilità di non finanziare aziende che poi sopravivranno invece varia tra circa il 56% e il 53% (*Falsi negativi*).

Tali risultati possono essere utilizzati ai fini di una politica di sostegno alle aziende agricole, finanziando effettivamente solo quelle che presentano una probabilità di sopravvivenza nei 5 anni superiore a una determinata soglia, che nel nostro caso può essere compresa tra 0.2 e 0.4. Si ha, in questo modo, un generale miglioramento in termini di stima della probabilità di sovvenzionare un'azienda che chiuderà entro 5 anni: passando da un 23.84% nel caso in cui si considerassero tutte le aziende

ugualmente meritevoli dell'aiuto, a un 11.44% utilizzando una soglia pari a 0.2 sui risultati del modello ridotto.

Le figure seguenti illustrano sinteticamente la variabilità di alcuni risultati presentati nella Tabella 4.3 al variare della soglia utilizzata.

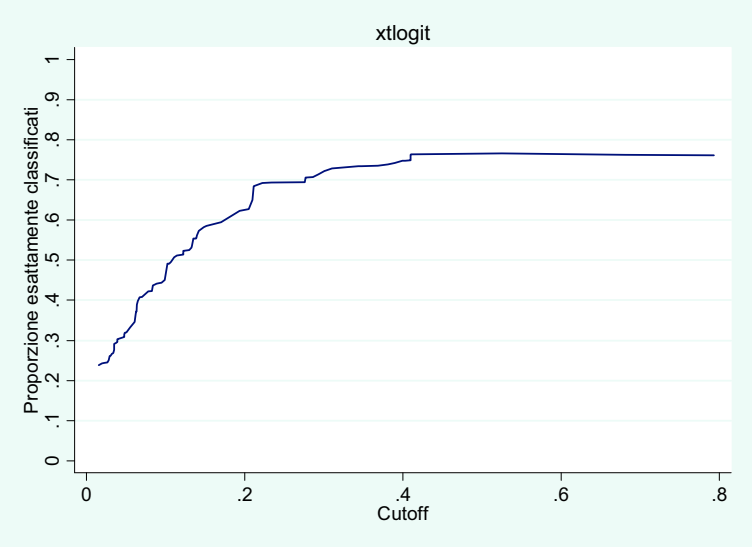


Figura 4.2 *Sensibilità della classificazione rispetto alle soglie*

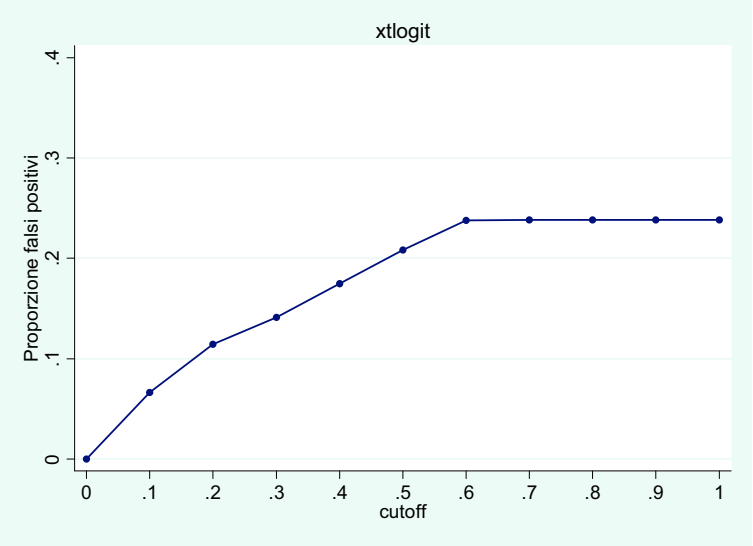


Figura 4.3:*Sensibilità della proporzione di falsi positivi alle soglie*

Nella Figura 4.2 si è andati a valutare la sensibilità della classificazione rispetto a tutte le possibili soglie di *cut-off* individuabili. Assumendo una soglia pari circa a 0.6

si può notare come la proporzione di esattamente classificati si stabilizzi. E' utile in queste analisi confrontare la Figura 4.2 con la Figura 4.3 in cui viene rappresentata la sensibilità della proporzione di falsi postivi alle varie soglie. Anche in questo caso è evidente come la proporzione si stabilizzi alla stessa soglia.

Per testare inoltre effettivamente la capacità previsiva del modello si è deciso di stimare il modello sul 50% della popolazione in oggetto, selezionata casualmente, e di analizzare le previsioni estendendole all'altra metà del campione non utilizzato per la stima del modello.

Nella Figura 4.1 è rappresentata la distribuzione delle probabilità di cessazione ottenute stimando il modello su tutto il campione e prevedendo all'interno dello stesso; nella Figura 4.4 invece è rappresentata la distribuzione delle probabilità di cessazione stimando il modello sul 50% del campione ed estendendo le previsioni all'altra metà della popolazione non utilizzata per la stima.

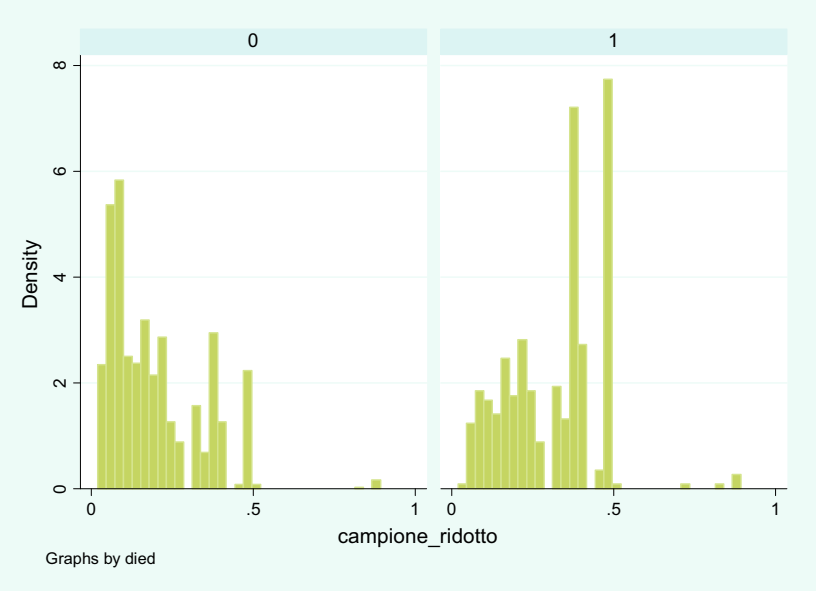


Figura 4.4: *Distribuzione delle probabilità di cessazione stimando il modello sul campione ridotto.*

La capacità classificatoria del modello che sfrutta metà della popolazione è molto simile a quella del modello che sfrutta l'intera popolazione, non si sono notati evidenti scostamenti dei risultati esposti nella Tabella 4.3. Confrontando in effetti le distribuzioni, non sembrano discostarsi in maniera evidente.

4.3 L' analisi ROC (Receiver Operating Characteristic)

L'analisi *ROC* è una metodologia sviluppata per la prima volta durante la II Guerra mondiale per l'analisi delle immagini radar e lo studio del rapporto segnale/disturbo. Essa venne ben presto applicata in altri campi della tecnica e a partire dagli anni '70, anche in campo medico, inizialmente allo scopo di quantificare l'attendibilità dei responsi di immagini radiografiche interpretate da operatori diversi.

In tempi più recenti l'utilizzo delle curve ROC si è fatto relativamente comune per la valutazione non solo delle immagini, ma anche dei più svariati test sia nel settore medico (con particolare riguardo alla valutazione dei test clinici di laboratorio) che, in minor misura, in quello veterinario.

In un test di tipo quantitativo, occorre individuare sulla scala di lettura un valore-soglia *("cut-point"o "cut-off"*) che discrimini i risultati da dichiarare "positivi" da quelli "negativi", consentendo in questo modo di categorizzare in "positivi" e "negativi" la gamma di tutti i possibili risultati e di equiparare l'interpretazione di un test quantitativo a quella di un test qualitativo. Solitamente nella pratica è quasi impossibile individuare un valore di *cut-off* che consenta una classificazione perfetta, tale da azzerare sia falsi positivi che falsi negativi.

La *performance* di un modello in termini di classificazione per un determinato valore di *cutoff* può essere valutata attraverso una semplice tabella di contingenza che confronta l'*output* del modello con l'effettivo valore realizzato.

	Risultato effettivo		
Risultato del modello		1	0
1		*a*	*b*
0		*c*	*d*

Tabella 4.4: *Esempio di tabella di contingenza*

Il confronto fra i risultati previsti dal modello e l'effettivo stato realizzato consente di stimare due importanti parametri:

- la *sensibilità* (Se) ossia la probabilità che, riferendoci al dataset delle aziende agricole, un'azienda "cessata" sia effettivamente stimata dal modello come cessata:

$$Se = \frac{a}{(a+c)}$$

- la *specificità* (Sp) ossia la probabilità che un'azienda "sopravvissuta" risulti effettivamente stimata come sopravvissuta dal modello:

$$Sp = \frac{d}{(d+b)}$$

Se e *Sp* sono fra loro inversamente correlate in rapporto alla scelta del valore di *cut-off*. L'adozione di una soglia che offre un'elevata *Se* comporta una perdita di *Sp* e viceversa. La scelta della soglia quindi non deve essere dettata solo da questioni di ordine probabilistico volte a minimizzare la proporzione di classificazioni errate ma è necessario basarsi anche sull' impatto, nel nostro caso di tipo economico, relativamente alla generazione di Falsi positivi ($\frac{b}{a+b}$) e Falsi negativi ($\frac{c}{c+d}$).

L'analisi ROC viene effettuata attraverso lo studio della funzione che – in un test quantitativo – lega la probabilità di ottenere un risultato vero positivo (*a*) nella classe dei risultati positivi effettivamente verificatisi (*a*+*c*) ossia la *sensibilità*, alla probabilità di ottenere un risultato falso-positivo (*b)* nella classe dei risultati negativi effettivamente verificatisi *(b+d)* ossia 1-*specificità*. In un gergo medico vengono studiati i rapporti fra allarmi veri (*hit rate)* e falsi allarmi (Bottarelli e Parodi, 2003).

La relazione tra i suddetti parametri può venire raffigurata attraverso una linea che si ottiene riportando, in un sistema di assi cartesiani e per ogni possibile valore di *cut-off*, la proporzione di veri positivi in ordinata e la proporzione di falsi positivi in ascissa (un esempio è in Figura 4.5). Se il risultato è riportato su scala continua, si possono calcolare i valori di *sensibilità* e *1-specificità* per ogni valore registrato. L'unione dei punti ottenuti riportando nel piano cartesiano ciascuna coppia *(Se) e (1-Sp)* genera una curva spezzata con andamento a scaletta ("*ROC plot*"). Per interpolazione, è possibile eliminare la scalettatura (*smoothing)* ed ottenere una curva ("ROC curve").

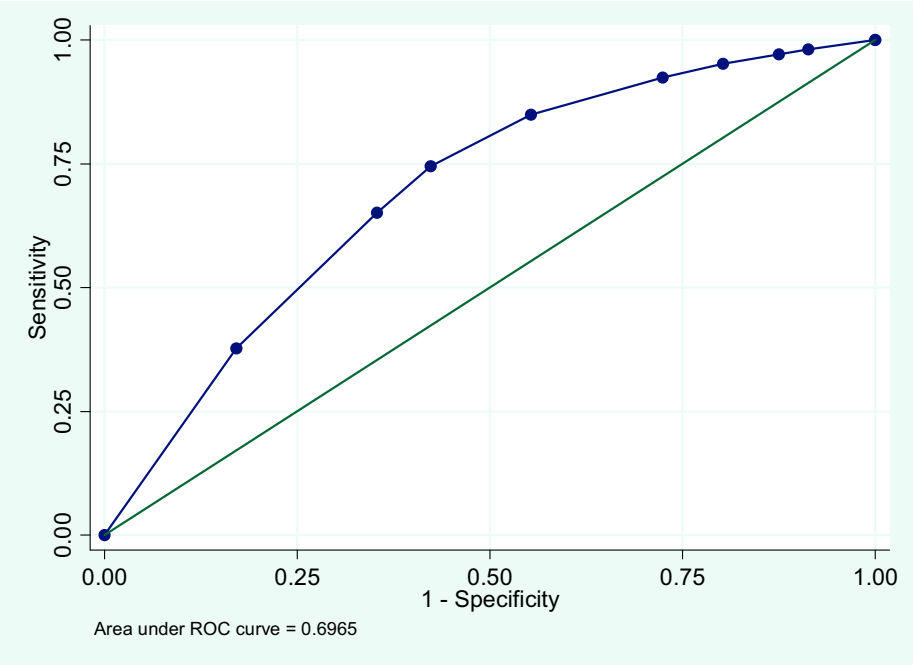


Figura 4.5: *Esempio di Curva ROC*

La capacità discriminante di un modello, ossia la sua attitudine a separare propriamente la popolazione di studio, nel nostro caso separare le aziende "cessate" da quelle "sopravvissute", è proporzionale all'estensione dell'area sottesa alla curva ROC (*Area Under Curve, AUC)* ed equivale alla probabilità che il risultato fornito da un determinato modello su un'unità estratta a caso dal gruppo delle aziende "cessate" sia superiore a quello di uno estratto a caso dal gruppo delle aziende "sopravvissute". Bamber (1975) ha dimostrato l'equivalenza tra l'area AUC sottesa ad una curva ROC, costruita per dati su scala continua, e la statistica U di Wilcoxon e Mann-Whitney.

Nel caso di un modello perfetto, ossia che non restituisce alcun falso positivo né falso negativo (capacità discriminante = 100%), la AUC passa attraverso le coordinate [0;1] ed il suo valore corrisponde all'area dell'intero quadrato delimitato dai punti di coordinate (0,0), (0,1), (1,0), (1,1), che assume valore 1 corrispondendo a una probabilità del 100% di una corretta classificazione.

Per un modello privo di valore informativo la curva ROC è rappresentata dalla diagonale *("chance line")* che passa per l'origine, con AUC=0.5.

Lo studio della curva ROC che tiene in considerazione la relazione che lega *sensibilità* e *specificità* da la possibilità di utilizzare un criterio "flessibile" per l'individuazione del *cut-off* ottimale.

Come regola generale empirica, si può affermare che il punto sulla curva ROC più vicino all'angolo superiore sinistro rappresenta il migliore compromesso tra *sensibilità e specificità* (Bottarelli e Parodi, 2003).

L'area sottesa ad una curva ROC rappresenta un parametro fondamentale per la valutazione della *performance* di un modello, in quanto costituisce una misura di accuratezza non dipendente dalla prevalenza ("*pure accurancy*").

Dalle proprietà della statistica U, *AUC* può essere considerata una variabile normale, per cui si può costruire un test ***z*** nel modo seguente (Bamber , 1975):

$$\boldsymbol{z} = \frac{\mathrm{AUC} - 0.5}{\sqrt{\sigma^2_{\mathrm{AUC}}}}$$

ove σ^2 rappresenta la varianza di AUC.

Se, ad esempio, il valore di **z** eccede il valore critico di 1.96, si può affermare che il modello presenta una performance significativamente superiore a quella di un test non discriminante.

I risultati di due modelli diversi possono inoltre essere confrontati tra di loro comparandone le *accuracy* stimate mediante l'area sottesa alle corrispondenti curve ROC (si veda esempio in Figura 4.6). Un semplice test **z** (basato sulla distribuzione normale standardizzata) può essere eseguito rapportando la differenza delle due aree all'errore standard di tale differenza.

$$z = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}}$$

r rappresenta la correlazione esistente tra le due AUC (può accadere che ci sia correlazione se applichiamo i modelli agli stessi soggetti). Il pacchetto STATA offre la possibilità di tenerne conto stimando gli *Standard Error* con il metodo di Hanley (Hanley, 1982).

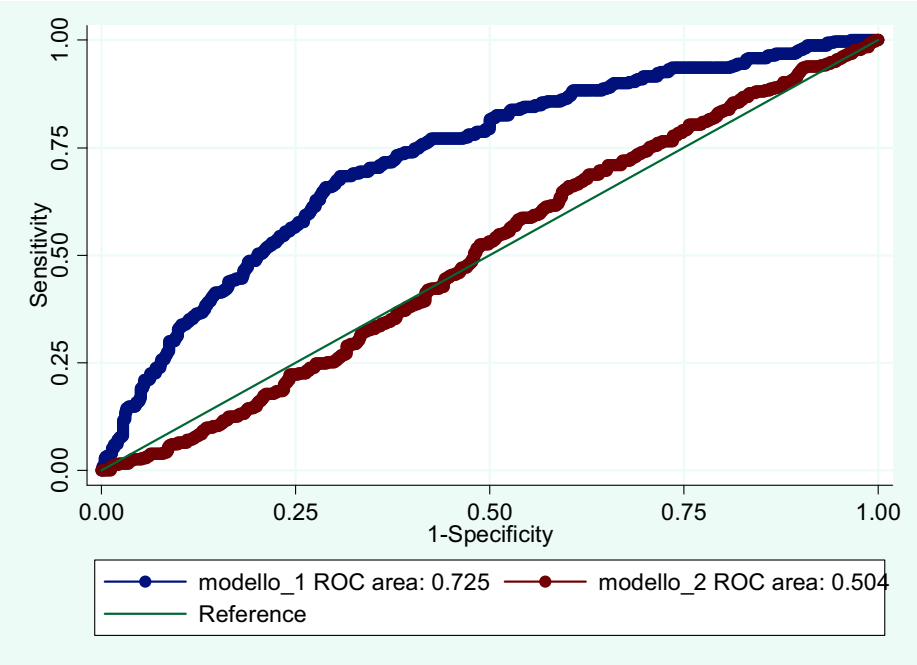


Figura 4.6: *Esempio di confronto tra due Curve Roc*

Per quanto riguarda l'interpretazione del valore di AUC, Swets (1988) propone una classificazione della capacità discriminante del modello utilizzata soprattutto in campo medico/epidemiologico. E'basata su criteri largamente soggettivi ed avviene secondo lo schema seguente:

- AUC = 0.5 il modello utilizzato non è informativo
- $0.5 < AUC \leq 0.7$ il modello è poco accurato
- $0.7 < AUC \leq 0.9$ il modello fornisce risultati moderatamente accurati
- $0.9 < AUC < 1.0$ il modello fornisce risultati altamente accurati
- AUC = 1.0 il modello è perfetto

4.3.1 Applicazione ai dati reali

Nella Figura 4.7 si riporta la *Curva Roc* e il valore della *Roc Area* relativamente alle stime di probabilità di cessazione ottenute applicando il modello ridotto ai dati tramite la funzione *xtlogit*.

La versione 9.2 di STATA offre la possibilità di calcolare l'*AUC* e relativi *standard error* con il metodo non parametrico di De-Long e Clarke-Pearson, il metodo di Bamber e quello di Hanley. La Tabella 4.5 riporta i risultati.

	ROC AREA	*Std.Error*	*Intervallo di confidenza al 95%*	
Metodo di Hanley	*0.7390*	*0.0131*	*0.71329*	*0.76480*
Metodo di Bamber	*0.7390*	*0.0131*	*0.71334*	*0.76474*
Metodo di De Long	*0.7390*	*0.0131*	*0.71334*	*0.76475*

Tabella 4.5: *Stime Roc Area e relativi Standard Error*

In queste analisi e nelle successive, i tre metodi non presentano evidenti differenze nel calcolo delle *AUC* e dei relativi *Se:* le differenze sono minimali e si riscontrano a livelli inferiori del millesimo.

Per quanto riguarda l'interpretazione del valore di *AUC*, Swets ci suggerisce che siamo di fronte a un modello che in termini di prestazioni classificative fornisce risultati moderatamente accurati.

Lo studio della curva ROC ci fornisce la possibilità di utilizzare un criterio "flessibile" per l'individuazione del *cut-off* ottimale, individuando nel punto più vicino all'angolo superiore sinistro il migliore compromesso tra *sensibilità e specificità* e quindi un modo utile per scegliere il *cut-off.*

A questo scopo nella Tabella 4.6 viene riportato un estratto della variazione in termini di specificità e sensitività e corretta classificazione, al variare delle soglie empiriche, per quelle combinazioni di *sensibilità e specificità* ritenute più plausibili.

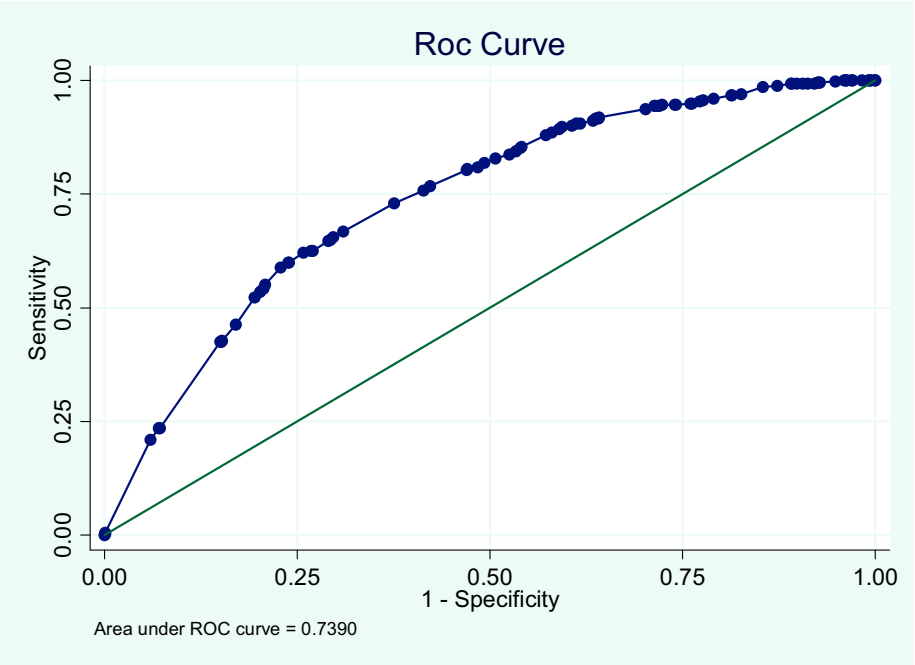


Figura 4.7: *Roc Area e Roc Curve per le probabilità di cessazione ricavate dall'applicazione del modello ridotto (Metodo di De Long)*

Xtlogit			
soglia empirica	*Sensibilità*	*Specificità*	*Class. Corrette*
.....	.....	.....	.....
0,211295	*66,74%*	*69,00%*	*68,46%*
0,222486	*65,58%*	*70,31%*	*69,18%*
0,234191	*64,88%*	*70,67%*	*69,29%*
0,276015	*64,65%*	*70,96%*	*69,46%*
0,277054	*62,56%*	*73,00%*	*70,51%*
0,286155	*62,56%*	*73,22%*	*70,68%*
0,293285	*62,09%*	*74,24%*	*71,34%*
0,301214	*60,00%*	*76,06%*	*72,23%*
.....	.....	.....	.....

Tabella 4.6: *Soglie di cut-off e relative Sensibilità, Specificità e Corretta classificazione*

Le possibili soglie che sembrano dare performance migliori in termini di classificazione sono individuabili in un *range* che va da 0.25 a 0.30, in cui l'esatta classificazione si attesta tra il 70 e il 72%.

Nel nostro particolare caso, seguendo quindi l'approccio descritto nel paragrafo 4.3, la Tabella 4.7 presenta le performance del modello per le principali soglie plausibili individuate.

Risultati utilizzando soglia 0.25 *xtlogit*			
Condizione prevista	*Condizione effettiva*		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+)	278	399	677
0=ancora attiva (-)	152	975	1127
	430	1374	
Falsi positivi	*Pr(D|-)*		13.48%
Falsi negativi	*Pr(≠ D|+)*		58.94%
Sensibilità	*Pr(+|D)*		64.65%
Specificità	*Pr(-|≠ D)*		70.96%
Classificazioni corrette			69.46%

Risultati utilizzando soglia 0.30 *xtlogit*			
Condizione prevista	*Condizione effettiva*		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+)	258	329	587
0=ancora attiva (-)	172	1045	1217
	430	1374	
Falsi positivi	*Pr(D|-)*		14.13%
Falsi negativi	*Pr(≠ D|+)*		56.04%
Sensibilità	*Pr(+|D)*		60%
Specificità	*Pr(-|≠ D)*		76.05%
Classificazioni corrette			72.23%

Tabella 4.7: *Tabelle di contingenza per le soglie 0.25 e 0.30*

5. I modelli GLLAMM

I GLLAMM (*Generalized Linear Latent and Mixed Models*) sono una classe di modelli per variabili latenti multilivello utilizzabili con vari tipi di variabili risposta: continue, conteggi, dati di durata, dicotomiche e dati categoriali. Le variabili latenti, o effetti casuali, possono avere una distribuzione discreta o normale multivariata.
Esempi di modelli appartenenti a questa classe sono: i modelli lineari generalizzati multilivello, i modelli fattoriali multilivello, i modelli a classi latenti e i modelli a equazioni strutturali multilivello.
Le stime dei modelli GLLAMM si possono ottenere utilizzando la funzione esterna **gllamm**, sviluppata per il software STATA da Skrondal, Rabe-Hesketh e Pickles, (2004) e scaricabile dal sito www.gllamm.org.
I GLLAMM possono essere definiti specificando:

1. Il valore atteso condizionato delle variabili risposta date le variabili latenti e le osservate esplicative.

2. La distribuzione condizionale delle variabili risposta date le variabili latenti osservate.

3. Equazioni strutturali per le variabili latenti che possono includere regressioni di variabili latenti sulle variabili esplicative e regressioni di variabili latenti su altre variabili latenti.

4. Le distribuzioni delle variabili latenti.

Le osservazioni contenute nel dataset in formato *episode – splitting*, come descritte nel capitolo 4 e rappresentate schematicamente in Tabella 5.1, possono essere viste attraverso l'ottica di un modello a due livelli.

Livello 1	Livello 2
Azienda 1	Anno 1 Anno 2 Anno 3 Anno 4
Azienda 2	Anno 1 Anno 2

Tabella 5.1: *Struttura dei dati per un modello a due livelli*

In questo tipo di analisi le differenze tra i soggetti non si suppongono catturate esclusivamente dal vettore delle esplicative X. Si ammette, invece, la presenza di una o più variabili, che non vengono osservate ma hanno effetto sulla sopravvivenza o meno di un'azienda agricola.

I modelli lineari generalizzati multilivello sono modelli lineari generalizzati che contengono effetti casuali normali multivariati nel predittore lineare.

Tali modelli sono anche conosciuti come modelli lineari generalizzati gerarchici o modelli lineari generalizzati a effetti misti. Un caso molto comune tra questo tipo di modelli è senza dubbio quello relativo ai modelli multilivello generalizzati per prevedere probabilità.

Gli effetti casuali rappresentano l'eterogeneità non osservata e inducono dipendenza tra le unità annidate nei "cluster".

In questa tesi cercheremo di indagare e discutere sul funzionamento delle previsioni (nel nostro caso le probabilità), per i modelli multilivello generalizzati con variabile risposta dicotomica utilizzando l'applicazione **gllamm** di Stata e i diversi comandi utili alla previsione su dati simulati e reali.

5.1 Modelli lineari multilivello e modelli lineari generalizzati

Ci limiteremo nella trattazione a considerare modelli a due livelli, perché per modelli di ordine di livello superiore la notazione diverrebbe troppo complicata. L'idea è comunque che ciò che viene qui presentato possa essere esteso a modelli con più di due livelli.

Per la variabile risposta y_{ij} di un unità i in un cluster j (N.B.: si mantiene qui la notazione utilizzata in letteratura diversa da quella vista per il modello di durata descritto al capitolo 4), il modello lineare a due livelli può essere espresso come:

$$y_{ij} = x_{ij}^{\intercal}\beta + z_{ij}^{\intercal}\zeta_j + \varepsilon_{ij}$$

dove x_{ij} sono le covariate con coefficienti fissati a β, z_{ij} sono le covariate con gli effetti casuali ζ_j e ε_{ij} sono errori di primo livello.

E' anche possibile scrivere il modello per le unità nel cluster come segue:

$$y_j = X_j\beta + Z_j\zeta_j + \varepsilon_j$$

Si assume che le covariate X_j e Z_j siano casuali e strettamente esogene (Skrondal e Rabe-Hesketh, 2004) nel senso che $E(\varepsilon_{ij}|\zeta_j, X_j, Z_j) = E\left(\varepsilon_{ij}\middle|\zeta_j, x_{ij}, z_{ij}\right) = E\left(\varepsilon_{ij}\right) = 0$ e $E\left(\zeta_j\middle|X_j, Z_j\right) = E\left(\zeta_j\right) = 0$. Gli effetti casuali e gli errori di livello 1 si assume abbiano una distribuzione normale multivariata: $\zeta_j|X_j, Z_j \sim N(0, \Psi)$ e $\varepsilon_j|\zeta_j, X_j, Z_j \sim N(0, \Theta_j)$.

Un modello lineare generalizzato a due livelli può invece essere scritto come:

$$h^{-1}\{E(y_{ij}|\zeta_j, x_{ij}, z_{ij})\} = x_{ij}^{\intercal}\beta + z_{ij}^{\intercal}\zeta_j \equiv \eta_{ij}$$

dove $h^{-1}(\cdot)$ è una funzione di legame e η_{ij} è il predittore lineare. In altre parole, il valore atteso condizionato della variabile risposta, date le covariate e gli effetti casuali, è:

$$\mu_{ij} \equiv E\left(y_{ij}\middle|\zeta_j, x_{ij}, z_{ij}\right) = h\left(x_{ij}^{\intercal}\beta + z_{ij}^{\intercal}\zeta_j\right) = h(\eta_{ij})$$

Come per i modelli lineari si assume che gli effetti casuali si distribuiscano come normali multivariate e le covariate siano strettamente esogene. Si assume che le variabili risposta inoltre siano condizionatamente indipendenti date le covariate e gli effetti casuali e che abbiano distribuzioni condizionali derivanti dalla famiglia esponenziale.

Per questa famiglia di distribuzioni, la varianza condizionale è data da:

$$var\left(y_{ij}\middle|\mu_{ij}\right) = \phi_{ij}V(\mu_{ij})$$

dove $_{ij}$ è un parametro di dispersione e $V(\mu_{ij})$ è una funzione della varianza che mette in evidenza la relazione tra la varianza condizionale e il valore atteso condizionale.

Il modello lineare multilivello risulta nel momento in cui viene specificato il link di identità, cioè $\mu_{ij} = \eta_{ij}$ in combinazione alla distribuzione condizionale normale per la variabile risposta $y_{ij}|\mu_{ij} \sim N(\mu_{ij}, \theta)$. In questo caso, la funzione di varianza è pari a 1 e il parametro di dispersione è un parametro libero $\phi_{ij} = \theta$.

Un altro caso speciale è il modello di regressione logistica per variabili risposta dicotomiche che combina una funzione di legame logit, $\text{logit}(\mu_{ij}) \equiv \log\{\mu_{ij}/(1-\mu_{ij})\} = \eta_{ij}$, con una distribuzione condizionale di Bernoulli per la variabile risposta, $y_{ij}|\mu_{ij} \sim Bernoulli\left(\mu_{ij}\right)$. La funzione di varianza è ora $V\left(\mu_{ij}\right) = \mu_{ij}(1-\mu_{ij})$ e il parametro di dispersione è 1 (Skrondal e Rabe-Hesketh, 2008).

Considerando θ come vettore dei parametri del modello da stimare, il contributo di verosimiglianza per il *j*-esimo cluster, $l_j(\theta) \equiv g\left(y_j|X_j, Z_j; \theta\right)$, diventa

$$l_j(\theta) = \int \varphi\left(\zeta_j, \Psi\right) f\left(y_j\middle|\zeta_j, X_j, Z_j; \theta^f\right) d\zeta_j$$

$$= \int \varphi(\zeta_j; \Psi) \prod_{i=1}^{n_j} f\left(y_{ij}|\zeta_j, x_{ij}, z_{ij}; \theta^f\right) d\zeta_j$$

Il primo termine dell'integrale è la densità relativa agli effetti casuali (Normali Multivariate con medie pari a zero e matrice di covarianza Ψ) e il secondo termine è la densità condizionale (o probabilità) della variabile risposta dati gli effetti casuali e le covariate.

Si utilizza la notazione θ^f per evidenziare il vettore dei parametri che appare nella distribuzione condizionale della variabile risposta, in questo modo θ consiste di θ^f e degli elementi di Ψ.

Dal momento che i cluster si assumono indipendenti, la verosimiglianza per il campione diventa $l(\theta) = \prod_{j=1}^{J} l_j(\theta)$.

Ad eccezione del caso dei modelli lineari multilivello, gli integrali di solito non hanno soluzioni analitiche ma devono essere risolti numericamente, tipicamente con metodi di quadratura adattativa (Skrondal e Rabe-Hesketh, 2009).

5.2 Previsione degli effetti casuali nell'ottica Bayesiana empirica

In questa sezione si discute come vengono stimati i valori degli effetti casuali $\zeta_j=(\zeta_{1j}\ldots\ldots\ldots,\zeta_{Qj})$ per i cluster $j=1,\ldots\ldots J$ nell'ottica Bayesiana empirica. Questa tipologia di assegnazione di solito inizia dopo che i parametri del modello sono stati stimati, con le stime dei parametri $\hat{\theta}$ trattate come parametri noti.

Quando i parametri del modello sono trattati come noti, il problema di assegnare dei valori agli effetti casuali può essere risolto con diverse metodologie e approcci, tra i quali un approccio Bayesiano empirico.

Nell'approccio Bayesiano, l'inferenza riguardante ζ_j per il cluster j è basata sulla distribuzione a posteriori di ζ_j condizionatamente ai dati noti del cluster che vengono trattati come valori osservati di variabili casuali.

Trattando i parametri del modello come noti e uguali alle stime ottenute con il metodo della massima verosimiglianza, abbiamo due fonti di informazione relative agli effetti casuali.

La prima fonte di informazione è la distribuzione a priori degli effetti casuali $g(\zeta_i; \widehat{\Psi})$, che rappresenta la nostra conoscenza *a priori* sugli effetti casuali prima di

"vedere" i dati per il cluster *j*. La seconda parte di informazione sono i dati y_j, X_j, *e* Z_j per il cluster j.

Un modo naturale di combinare le fonti di informazione relative agli effetti casuali è attraverso la distribuzione a posteriori $\varpi(\zeta_j|y_j, X_j, Z_j, \hat{\theta})$ di ζ_j :

$$\frac{g(\zeta_j; \hat{\Psi}) f(y_j|\zeta_j; X_j; Z_j; \widehat{\theta^f})}{g(y_j|X_j, Z_j, \hat{\theta})}$$

Il denominatore rappresenta il contributo di verosimiglianza $l_j(\hat{\theta})$ del *j*-esimo cluster e di solito non ha una forma finita ma è necessario valutarlo in maniera numerica. Qui i parametri sono trattati come noti e uguali alle loro stime, in questo modo la distribuzione è "empirica"o "stimata" (Skrondal e Rabe-Hesketh, 2009).

Per modelli lineari segue dai risultati standard sulle densità condizionate di normali multivariate che la densità a posteriori è normale multivariata.

Per altri tipi di variabili risposta, segue dal teorema del limite centrale Bayesiano (Carlin e Louis, 2000) che la densità a posteriori tende a una normale multivariata a mano a mano che le unità nel cluster aumentano.

La previsione empirica Bayesiana è il metodo più utilizzato per assegnare valori agli effetti casuali.

I predittori Empirici Bayesiani degli effetti casuali ζ_j sono i valori attesi della distribuzione a posteriori empirica (ottenuta inserendo le stime dei parametri $\hat{\theta}$).

$$\widetilde{\zeta_j^{EB}} = E(\zeta_j|y_j, X_j, Z_j; \hat{\theta}) = \int \zeta_j \varpi(\zeta_j|y_j, X_j, Z_j; \hat{\theta}) d\zeta_l$$

La ragione per cui si utilizza il termine "Empirico Bayesiano", che venne coniato da Robbins (1955), è che sono adottati principi Bayesiani su un approccio frequentista, inserendo nel modello dei parametri stimati. I puri Bayesiani andrebbero a ottenere la distribuzione a posteriori degli effetti casuali, assumendo una distribuzione a priori per θ, invece di implementare semplicemente le stime $\hat{\theta}$ per θ (Skrondal e Rabe-Hesketh, 2009).

5.3 Previsione delle variabili dipendenti

La media condizionale della variabile risposta, o nel nostro caso probabilità, per un unità con valori delle covariate $\mathbf{x_{ij}=x^0}$ e $\mathbf{z_{ij}=z^0}$ in un ipotetico cluster con effetti casuali $\zeta_j = \zeta_j^0$ è data da:

$$\hat{\mu}(x^0, z^0, \zeta_j^0) \equiv E_y(y_{ij}|\zeta_j^0, x^0, z^0; \hat{\beta}) = h(x^{0\prime}\hat{\beta} + z^{0\prime}\zeta_j^0)$$

Invece di andare a considerare particolari valori di ζ_j^0 degli effetti casuali, possiamo andare a considerare la distribuzione di $\hat{\mu}(x^0, z^0, \zeta_j^0)$ nella popolazione di cluster. Un alternativa all'uso della distribuzione a priori degli effetti casuali $\varphi(\zeta_j; \widehat{\Psi})$ per ricavare la distribuzione di $\hat{\mu}(x^0, z^0, \zeta_j)$ sarebbe quella di utilizzare la distribuzione a posteriori $\varpi(\zeta_j|y_j.X_j.Z_j; \hat{\theta})$.
Sostanzialmente abbiamo quindi altri due modi per ricavare delle previsioni a seconda dei due tipi di distribuzioni utilizzate.
Utilizzando la distribuzione a priori degli effetti casuali $\varphi(\zeta_j; \widehat{\Psi})$ possiamo andare ad integrare $\hat{\mu}(x^0, z^0, \zeta_j)$ al fine di ottenere il valore $\bar{\mu}(x^0, z^0)$:

$$\bar{\mu}(x^0, z^0) \equiv E_y(y_{ij}|x^0, z^0; \widehat{\Psi}) = \int_{-\infty}^{\infty} \hat{\mu}(x^0, z^0; \zeta_j)\varphi(\zeta_j; \widehat{\Psi})d\zeta_j$$

Il valore atteso marginale può essere utilizzato per fare una previsione di un unità su un nuovo cluster, ipotizzando che sia campionato casualmente. Aftshartous e de Leeuw (2005) chiamarono questo metodo di previsione "**prior prediction method**" .
L'integrale coinvolto nella determinazione del valore atteso deve generalmente essere calcolato attraverso metodi di approssimazione numerica.
Se si vuole invece considerare il valore atteso nel cluster (anche chiamato "*cluster-averaged expectation"*), dal momento che gli effetti casuali per il singolo cluster non sono noti, non si può andare ad utilizzare la media condizionale vista sopra ma è necessario utilizzare la distribuzione a posteriori la quale rappresenta tutta la nostra conoscenza sugli effetti casuali per il "cluster".

Possiamo ottenere il "*cluster-averaged expectation*" $\tilde{\mu_j}(x^0, z^0)$ dall'integrazione di $\hat{\mu}(x^0, z^0, \zeta_j)$ sulla distribuzione a posteriori degli effetti casuali per il cluster:

$$\tilde{\mu}_j(x^0, z^0) \equiv E_\zeta\{\hat{\mu}(x^0, z^0, \zeta_j)|y_j, X_j, Z_j; \hat{\theta}\} = \int_{-\infty}^{\infty} \hat{\mu}(x^0, z^0, \zeta_j)\omega(\zeta_j|y_j, X_j, Z_j; \hat{\theta})d\zeta_j$$

Il valore atteso basato sulla distribuzione a posteriori può essere utilizzato per fare previsioni su nuove unità in un cluster *j* già esistente, sfruttando l'informazione che si ha sul cluster stesso. Non è quindi possibile utilizzare questo approccio per fare previsioni su un nuovo cluster, in quanto sono necessari i valori *y* osservati nel cluster stesso. Come si vedrà, questa caratteristica renderà poco utile questo metodo per le analisi di interesse della tesi.

5.4 Il modello a due livelli per l'analisi di sopravvivenza

Un modello a due livelli della struttura vista nella Tabella 5.1, con variabile esplicativa dicotomica, può essere descritto nella seguente equazione (Wooldridge, 2002) (N.B. In questa sezione riprendiamo la notazione utilizzata nel capitolo 4 per la descrizione del modello a effetti casuali per dati di panel):

$$Pr(y_{it} \neq 1|x_i, \zeta_i) = Pr(y_{it} \neq 1|x_{it}, \zeta_i) = \pi(x_{it}\beta + \zeta_i) = \frac{exp(x_{it}\beta+\zeta_i)}{1+exp(x_{it}\beta+\zeta_i)} \quad , \text{t=1,.....,ni}$$

dove ζ_i rappresenta l'effetto non osservato e $\boldsymbol{x_i}$ contiene $\boldsymbol{x_{it}}$ per tutti i valori di *t*.
Si assuma inoltre :

- $y_{i1}, y_{i2}, \ldots\ldots\ldots\ldots y_{it}$ indipendenti condizionatamente a (x_i, ζ_i)
- $f(\zeta_i|\boldsymbol{x}_i) \sim N(0, \sigma_\zeta^2)$

Sotto queste assunzioni la verosimiglianza è data da:

$$\prod_i \int \left\{ \prod_t f(y_{it}|x_{it}, \zeta_i) \right\} g(\zeta_i) d\zeta_i$$

Dove $f(y_{it}|x_{it}, \zeta_i)$ è la densità della variabile risposta condizionata alle variabili esplicative e alla variabile latente; $g(\zeta_i) \sim N\left(0, \sigma_\zeta^2\right)$ è la densità a priori della variabile latente.

Il contributo alla verosimiglianza per il soggetto i è:

$$L_i(\beta, \sigma_\zeta) = \underbrace{\int_{-\infty}^{\infty} \frac{e^{-\zeta_i/2\sigma_\zeta^2}}{\sqrt{2\pi}\sigma_\zeta} \left\{ \prod_t f(y_{it}|\zeta_i, \beta) \right\} d\zeta_i}_{\propto \, a \, posteriori \, di \, \zeta_i}$$

dove

$$f(y_{it}|x_{it}, \zeta_i, \beta) = \pi(x_t\beta + \zeta)^{y_t}[1 - \pi(x_t\beta + \zeta)]^{1-y_t}$$

che è la stessa che si ottiene partendo da un modello per dati di panel con variabile esplicativa dicotomica (Biasiolo, 2006).

Per quanto riguarda le stime delle probabilità, l'integrale coinvolto nel calcolo della previsione tramite il "*prior prediction method*" diventa:

$$\bar{\mu}(x^0) \equiv E_y\left(y_{it}|x^0; \widehat{\sigma_\zeta}\right) = \int_{-\infty}^{\infty} \hat{\mu}(x^0; \zeta_i) \varphi\left(\zeta_i; \widehat{\sigma_\zeta}\right) d\zeta_i \qquad (5.1)$$

Andando invece a considerare il metodo di stima delle probabilità che utilizza la distribuzione a posteriori della variabile latente, chiamato anche *cluster-averaged expectation*, l'integrale coinvolto nel nostro caso diventa:

$$\tilde{\mu}_i(x^0) \equiv E_\zeta\left\{\hat{\mu}(x^0, \zeta_i)|y_{it}, X_i, ; \hat{\theta}\right\} = \int_{-\infty}^{\infty} \hat{\mu}(x^0, \zeta_i) \omega\left(\zeta_i|y_{it}, x_i; \hat{\theta}\right) d\zeta_i \qquad (5.2)$$

5.5 Modelli GLLAMM e STATA

Come accennato all'inizio del capitolo le stime e le previsioni dei modelli GLLAMM si possono ottenere utilizzando il "pacchetto" esterno **Gllamm,** sviluppato per il software STATA (alcuni esempi che utilizzano **Gllamm** sono disponibili nel sito : http://www.gllamm.org/examples.html).
Sostanzialmente all'interno del "pacchetto" sono forniti due comandi principali, uno dedicato alla stima vera e propria dei modelli (*gllamm*) e uno dedicato esclusivamente alle operazioni "post-stima" (*gllapred).*
Per la sintassi completa con tutte le molteplici opzioni di *gllamm* si rinvia al manuale di **Gllamm** realizzato dai creatori del pacchetto stesso (Skrondal, Rabe-Hesketh e Pickles A., 2004) .
Gllamm può utilizzare un gran numero di opzioni diverse in quanto con un singolo programma possiamo stimare una grande varietà di modelli differenti; per la maggior parte dei modelli, la funzione è corredata da una serie di opzioni la cui sintassi è simile a quella utilizzata da altri comandi di stima propri di Stata.
Nel nostro caso le opzioni che è necessario specificare per stimare un modello a due livelli con variabile esplicativa dicotomica sono indicativamente tre:

- *i(variable) :* attraverso il quale si specifica la variabile che definisce la gerarchia dei cluster.

- *link :* che ci permette di specificare la funzione di link che nel nostro caso è *logit.* Ci sono altri diversi link che possono essere specificati in quanto come si è detto in precedenza la funzione *gllamm* può stimare vari modelli.

- *family(families):* che ci permette di specificare la famiglia che deve essere utilizzata per la specificazione della densità condizionale, nel nostro caso è *binomial*.

Un'evoluzione del metodo di quadratura Gaussiana descritto nel capitolo 4 è stata introdotta per superare alcuni problemi nei modelli a due livelli. Nella funzione

gllamm è implementato il metodo della Quadratura Gaussiana Adattiva, che sembra funzionare meglio nei problemi in cui la quadratura ordinaria fallisce. Inoltre la quadratura adattiva sembra essere spesso più efficiente dal punto di vista computazionale in quanto richiede un numero inferiore di punti di quadratura per raggiungere la stessa precisione della quadratura ordinaria (Skrondal e Rabe-Hesketh, 2002).

In ***gllamm*** è implementato il metodo originalmente sviluppato da Naylor e Smith (1982), che sostanzialmente utilizza per l'adattamento una stima della media e della varianza della verosimiglianza.

L'output della funzione *gllamm* è molto simile agli output di altri comandi utilizzati in Stata, nella Figura 5.1 riportiamo le principali informazioni che l'output contiene.

Dopo aver stimato i parametri del modello utilizzando *gllamm* possiamo lanciare il comando per ottenere le previsioni, sia nel caso le variabili risposta siano continue che se sono discrete: la funzione utilizzabile come già visto in precedenza è *gllapred* e a seconda dell'opzione inserita si ottengono diversi tipi di previsioni.

Tra le opzioni possibili, si può accompagnare *gllapred* all'opzione *linepred* che come dice il nome fornisce il predittore lineare. Le opzioni però sui cui ci andremo a concentrare sono sostanzialmente due:

- *Mu* : che fornisce il valore atteso della variabile risposta, per esempio nel nostro caso la probabilità prevista essendo una variabile risposta dicotomica, condizionatamente alla *distribuzione a posteriori* della variabile latente.
- *Mu Marginal* : che fornisce il valore atteso della variabile risposta condizionatamente alla *distribuzione a priori*.

Le opzioni descritte che accompagnano *gllapred* ritornano i valori attesi specificati rispettivamente nelle formule (5.1 e 5.2) descritte al paragrafo 5.4, nel prossimo capitolo andremo effettivamente a verificarlo attraverso delle simulazioni.

```
Running adaptive quadrature
Iteration 0:   log likelihood = -1002.6404
Iteration 1:   log likelihood = -958.25227
Iteration 2:   log likelihood = -935.41446
Iteration 3:   log likelihood = -921.74259
Iteration 4:   log likelihood = -921.09894

Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -921.09894
Iteration 1:   log likelihood = -921.19368
Iteration 2:   log likelihood = -921.19351
Iteration 3:   log likelihood = -921.19351

number of level 1 units = 4014
number of level 2 units = 1000

Condition Number = 19.835655

gllamm model

log likelihood = -921.19351

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    5.324308   .7460507     7.14   0.000     3.862075    6.786541
          x2 |    3.861874   .6838371     5.65   0.000     2.521578     5.20217
          x3 |   -1.914732   .4133493    -4.63   0.000    -2.724882   -1.104582
          x4 |   -1.158135    .608935    -1.90   0.057    -2.351626    .0353555
       _cons |    -7.07799   .7911379    -8.95   0.000    -8.628592   -5.527389
------------------------------------------------------------------------------

Variances and covariances of random effects
------------------------------------------------------------------------------
    var(1): 13.110433
------------------------------------------------------------------------------
```

Figura 5.1: *Esempio di output della funzione gllamm*

Vengono riportati i diversi valori assunti dalla verosimiglianza utilizzando il metodo della Quadratura Gaussiana Adattiva fino alla convergenza dell'integrale (in giallo è segnalato un particolare valore assunto). Vengono anche riportati il valore dei parametri stimati e relativi standard error, intervalli di confidenza e test di significatività sui parametri (in verde è segnalato un particolare valore del parametro). Viene riportato il condition number, definito come la radice quadrata del rapporto tra il più grande e più piccolo autovalore della matrice Hessiana: è un indicatore di identificazione del modello. Nell'output viene anche indicata (in grigio) una stima della varianza dell' *effetto casuale.*

6. Le simulazioni

In questa sezione, con un approccio del tutto empirico, si confrontano le prestazioni predittive del "pacchetto" **gllamm**, con le relative funzioni *gllamm* e *gllapred,* con la funzione classica *xtlogit* per la stima di modelli di sopravvivenza a tempi discreti. Sull'esempio del dataset utilizzato per l'analisi delle probabilità di cessazione o sopravvivenza delle aziende agricole si ricreano dei dataset casuali utilizzabili per la stima di tali modelli con variabile risposta dicotomica, analizzandone la capacità predittiva sotto diverse condizioni controllate.

Con l'introduzione inoltre della *Curva Roc* siamo andati a sottoporre ad esame l'effettiva capacità discriminatoria dei diversi metodi trattati in precedenza per la previsione delle probabilità di cessazione in presenza/assenza di eterogeneità non osservata. Si è introdotto il "*prior prediction method*" visto nei paragrafi 5.3 e 5.4 e se ne sono valutate le performance.

I dataset di volta in volta simulati presentano la struttura in tabella 6.1.

id_azienda	y_{ik}	x_i	s_i	z_i	…..	***anno k***
5	0	x_5	s_5	z_5	…..	1
5	0	x_5	s_5	z_5		2
5	0	x_5	s_5	z_5	…..	3
5	0	x_5	s_5	z_5		4
5	0	x_5	s_5	z_5	…..	5
6	0	x_6	s_6	z_6	…..	1
6	0	x_6	s_6	z_6	…..	2
6	1	x_6	s_6	z_6	….	3
7	0	x_7	s_7	z_7	…..	1
7	0	x_7	s_7	z_7	…..	2
7	0	x_7	s_7	z_7	…..	3
7	1	x_7	s_7	z_7	…..	4

Tabella 6.1: *Struttura tipica del dataset utilizzato per le simulazioni*

Tipicamente le simulazioni prevedono 1000 unità statistiche osservate in 5 occasioni. Le variabili esplicative in questo caso sono *x, s* e *z* che rappresentano delle variabili determinanti per la sopravvivenza aziendale. Nel corso delle analisi ogni qualvolta andremo a simulare un dataset, specificheremo la distribuzione delle variabili utilizzate come esplicative: negli esempi svolti e raccolti in questa sezione, le variabili sono simulate casualmente da *distribuzioni Uniformi(0,1), Normali(0,1) e Bernoulliane.*

La variabile **y** è generata dall'applicazione di una soglia, nel nostro caso derivante da una generazione casuale da una distribuzione uniforme (0,1), al risultato della funzione *logit* come segue:

$$\widehat{p_{ik}} = \frac{exp^{(\beta_0+\beta_1 x_i+\beta_2 s_i \ldots\ldots+random\ effect_i+anno_k)}}{1+exp^{(\beta_0+\beta_1 x_i+\beta_2 s_i \ldots\ldots+random\ effect_i+anno_k)}}$$

$$y_{ik} = \begin{cases} 1\ se\ \widehat{p_{ik}} > soglia \\ 0\ se\ \widehat{p_{ik}} < soglia \end{cases}$$

Anche i parametri β_j vengono di volta in volta selezionati, in modo da dare alle variabili utilizzate una maggiore o minore capacità discriminatoria; variando invece la costante essenzialmente andiamo a decidere, a soglia definita, la concentrazione maggiore o minore di aziende sopravvissute.

In alcune esemplificazioni, al fine di studiare il comportamento delle diverse metodologie trattate nel caso di eterogeneità non osservata tra i soggetti, andiamo a inserire in $\widehat{p_{ik}}$ una variabile *random effect,* che assume lo stesso valore ogni qual volta lo stesso soggetto si presenta nel dataset. Nella stima dei modelli tale variabile sarà considerata come variabile latente e assumerà il ruolo di effetto casuale. Anche questa variabile sarà generata casualmente da una distribuzione *Normale(0,*σ^2*).*

Al fine poi di rientrare in una logica piena di modello di durata a tempi discreti si è deciso di costruire alcuni dataset inserendo una variabile legata alla durata stessa: siamo andati quindi a generare casualmente una variabile che assumesse lo stesso valore per ogni anno *k* computandola in $\widehat{p_{ik}}$. In questo modo, inserendo nella stima dei modelli delle *dummy*, possiamo andare a verificare l'effettiva capacità dei

modelli di cogliere eventualmente, in una logica di determinazione della probabilità di sopravvivenza o cessazione, una variazione del rischio nel tempo.
Suddivideremo in questo senso le simulazioni in due gruppi principali. Un primo gruppo (***Simulazioni 1***) nel quale testeremo essenzialmente i metodi di stima attraverso *gllamm e xtlogit* su dati di panel bilanciati, per cercare di cogliere e verificare essenzialmente differenze legate alla stima dei parametri slegandoci dalla durata. Andremo quindi a stimare i modelli su dataset che nel nostro caso possiedono 5 record per ogni identificativo di azienda.
Nel secondo gruppo (***Simulazioni 2)*** inserendo la variabile $anno_k$ ci riconduciamo ai modelli di durata a tempi discreti caratterizzando il rischio di base, come visto nel capitolo 4, in cui:

$$logit[h(k, X|\zeta)] = D[(k)] + \beta' X + random\ effect_i$$

dove $D(k)$ è un termine che caratterizza la funzione di rischio di base, mentre $random\ effect_i$ è il termine d'errore inserito per considerare l'eterogeneità non osservata, con distribuzione Normale $(0, \sigma_\zeta^2)$ (Jenkins, 2004). Andremo quindi a stimare i modelli con le diverse metodologie su dataset della struttura vista in tabella 6.1, ricostruendo poi le probabilità di cessazione come proposto all'inizio del capitolo 4.

6.1 Il "*prior prediction method*" con *gllapred* di Gllamm e la previsione con *xtlogit* in STATA

Un'alternativa per la stima delle probabilità di sopravvivenza in modelli di durata a tempi discreti (in una logica di modello a due livelli) è l'utilizzo del **"*prior prediction method*".**
Tale metodo, come visto in precedenza, utilizza il pacchetto esterno **Gllamm** di STATA ed è in questo implementato attraverso la funzione *gllapred mu marginal.*

$$\bar{\mu}(x^0) \equiv E_y\left(y_{it}|x^0;\widehat{\sigma_\zeta}\right) = \int_{-\infty}^{\infty} \hat{\mu}(x^0;\zeta_i)\varphi\left(\zeta_i;\widehat{\sigma_\zeta}\right)d\zeta_i$$

Questa metodologia previsiva utilizza, come si vede dall'equazione sopradescritta, la distribuzione a priori della variabile latente. Con il fine di verificare che la previsione effettivamente non utilizzasse l'informazione proveniente dalla y_{it} realmente accaduta (vedi ***cluster-averaged expectation,*** paragrafi 5.3 e 6.2) si è agito come si vedrà successivamente nel paragrafo 6.2.1: a partire dalle medesime esplicative siamo andati ad effettuare le previsioni su dataset con valori *y* differenti e siamo andati a verificare che fossero le stesse. Effettivamente le distribuzioni in questo caso sono identiche, il metodo previsivo non usufruisce di informazioni provenienti dalla y_{it} realmente accaduta e di conseguenza anche le *curve Roc* sono le medesime. Questo consente di utilizzare il metodo in situazioni reali nelle quali si vogliano prevedere le *y* sulla base dei valori assunti dalle esplicative.

6.1.1 Assenza di eterogeneità non osservata

In questa sezione si è deciso di confrontare l'approccio classico alla previsione dei modelli di durata a tempi discreti e l'approccio alla previsione del tipo "**prior prediction method**" nel caso di assenza di eterogeneità non osservata, riportando le conclusioni fondamentali raggiunte esemplificate su un determinato dataset. Le analisi sono comunque state svolte su più dataset diversi in modo da testare effettivamente i risultati: sono stati simulati 40 dataset e, in termini previsivi e di stima dei parametri, le conclusioni in gran parte dei casi sono quelle esemplificate in questa sezione.

Sostanzialmente si evince che stimando i modelli in assenza di eterogeneità le stime dei parametri "praticamente" sono equivalenti: non si nota un metodo di stima "superiore", le stime puntuali sembrano essere più precise in *xtlogit,* ma gli standard-error stimati con *gllamm* risultano migliori; la differenza probabilmente è dovuta alle diverse tipologie di quadratura utilizzata dai due metodi (come visto nel capitolo 5) e a un migliore utilizzo di *gllamm* delle stime della varianza di σ_ζ^2. Le probabilità

stimate di cessazione con i due metodi (la funzione *predict* di *xtlogit* e *gllapred mu marginal)* non coincidono perfettamente ma la loro diversità è talmente irrisoria che la classificazione non ne risente assolutamente. In conclusione non sembra quindi esserci una metodologia "superiore" utilizzabile nel caso di assenza di eterogeneità per fare previsioni sulle probabilità di cessazione.

Un risultato che si è riscontrato è che *xtlogit* stima comunque $\hat{\sigma}_\zeta^2$ anche se nella simulazione non è stata inserita alcuna componente di eterogeneità non osservata (in ogni caso la componente di eterogeneità tramite il test su $\boldsymbol{\rho}$, come vedremo in seguito, non risulta statisticamente significativa nella quasi totalità dei modelli stimati). Il modello *gllamm* da questo punto di vista sembra cogliere meglio l'assenza di eterogeneità, stimando $\hat{\sigma}_\zeta^2 \cong 0$ nella totalità dei modelli utilizzati.

Le variabili utilizzate nella simulazione sono 4, x_1 e x_4 realizzazioni casuali di una Distribuzione Uniforme [0,1], x_2 realizzazione casuale di una distribuzione Normale (0,1) e x_3 realizzazione casuale di una Bernoulliana.

Simulazione 1)

Nella Tabella 6.2 si nota come entrambi i metodi tendano a comportarsi in modo abbastanza simile nella stima di ogni singolo parametro : essi sovrastimano i parametri negativi e sottostimano quelli positivi. In particolare però si nota come la metodologia *gllamm* tenda a offrire prestazioni peggiori in termini di stima puntuale, nell'ordine comunque di pochi centesimi, rispetto a *xtlogit,* anche se i valori veri sono sempre all'interno degli intervalli. In compenso gli standard-error stimati da *gllamm* risultano più bassi, generando degli intervalli di confidenza di ampiezza inferiore : *gllamm* sembra quindi essere migliore in termini di precisione.

Probabilmente questo si deve al fatto che *gllamm* riesce in questa tipologia di modelli a stimare meglio σ_ζ^2, anche se comunque la varianza stimata da *xtlogit* rimane molto piccola.

Parametri	*Valori reali*	*Stime dei parametri con xtlogit*	*Standard Error xtlogit*	*Stime dei parametri con gllamm*	*Standard Error gllamm*
β_0	*-3.0*	*-2,8112***	*0,21448*	*-2,791452***	*0,211158*
β_1	*2.2*	*1,87536***	*0,27004*	*1,865523***	*0,265896*
β_2	*1.3*	*1,24685***	*0,08584*	*1,238577***	*0,084458*
β_3	*-2.1*	*-2,3451***	*0,16236*	*-2,334299***	*0,160372*
β_4	*-1.4*	*-1,2552***	*0,25715*	*-1,246434***	*0,253054*
Numerosità		5000		5000	
Log Likelihood		*-725.27058*		*-725.00553*	
$\widehat{\sigma}_\zeta^2$		***0.046812***		***5.159e-12***	

Tabella 6.2: *Riassunto delle stime xtlogit e gllamm*

I risultati del *Test Lr* (Likelihood-ratio test of rho=0) forniti dalla procedura *xtlogit* sono :

$\chi^2_{bar}(01) = 0.53$ $\qquad$ $Prob >= \chi^2_{bar}(01) = 0.233$

La presenza di eterogeneità non osservata è statisticamente non significativa, in accordo quindi con i valori immessi nella simulazione.

Simulazione 2)

In questa seconda simulazione andiamo a inserire la variabile $anno_k$ rientrando in un contesto di modelli di durata in forma *episode-splitting* come descritto nella parte iniziale del capitolo: nella Figura 6.1 viene rappresentata la distribuzione simulata degli anni di sopravvivenza delle aziende agricole. Nella Tabella 6.3 elenchiamo i valori reali dei parametri immessi nella simulazione (*a2, a3, a4 e a5* rappresentano i valori che assume $anno_k$ in corrispondenza dei diversi istanti temporali) e i risultati delle stime ottenute con la metodologia *xtlogit*

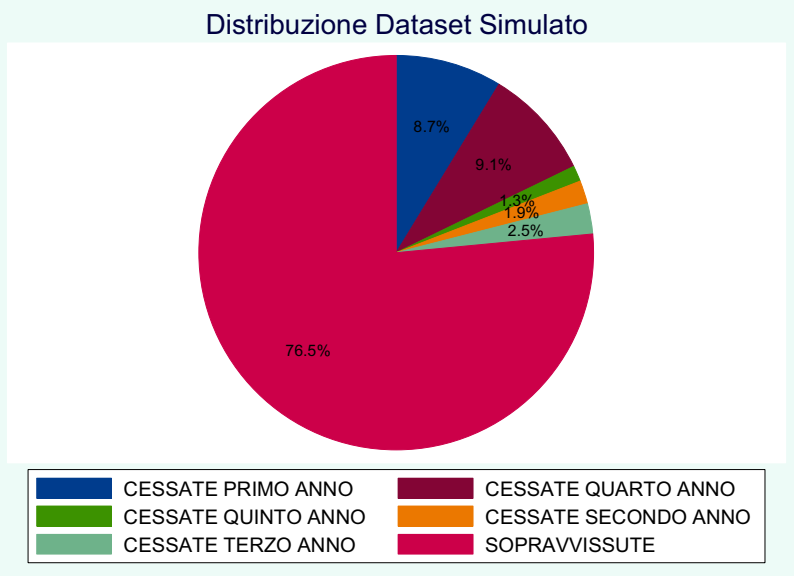


Figura 6.1: *Distribuzione per anni di sopravvivenza delle aziende agricole nel dataset simulato*

Parametri	***Valori reali***	***Stime dei parametri con xtlogit***	***Standard Error***	***[Intervallo di confidenza al 95%]***	
β_0	*-3.5*	*-3.44415**	*0,30616*	*-4,04422*	*-2.84407*
β_1	*2.2*	*1,658065**	*0,302932*	*1,06433*	*2,2518*
β_2	*1.3*	*1,255829**	*0,103585*	*1,052807*	*1,45885*
β_3	*-2.1*	*-2,32534**	*0,18302*	*-2,68405*	*-1,96663*
β_4	*-1.4*	*-1,0283**	*0,29369*	*-1,60393*	*-0,45268*
dummy2	*0.95*	*1,270098**	*0,245437*	*0,789051*	*1,751144*
dummy3	*-0.25*	*-0,14748**	*0,086887*	*-0,29687*	*0,02892*
dummy4	*1.65*	*1,555011**	*0,257213*	*1,050882*	*2,05914*
dummy5	*-0.64*	*-0,44697**	*0,21722*	*-0,8647*	*-0,02077*

Numerosità	***4563***
Log Likelihood	***-558.56576***
ρ	***.0154061***
$\hat{\sigma}_\zeta^2$	***0.05107****

Tabella 6.3: *Riassunto delle stime mediante metodologia xtlogit*

I risultati del Test Lr (Likelihood-ratio test of rho=0) sono:

$\chi^2_{bar}(01) = 0.19$ $\quad\quad$ $Prob >= \chi^2_{bar}(01) = 0.329$

Si può concludere che la presenza di eterogeneità non osservata non è statisticamente significativa.

Nella Tabella 6.4 andiamo invece ad esporre i risultati in termini di stime puntuali e intervallari ottenute con la metodologia gllamm.

Parametri	***Valori reali***	***Stime dei parametri con gllamm***	***Standard Error***	***[Intervallo di confidenza al 95%]***	
β_0	*-3.5*	*-3,41546**	*0,302547*	*-4,00844*	*-2,82248*
β_1	*2.2*	*1,640058**	*0,298852*	*1,05432*	*2,225797*
β_2	*1.3*	*1,243563**	*0,101912*	*1,043819*	*1,443307*
β_3	*-2.1*	*-2,30316**	*0,180082*	*-2,65612*	*-1,95021*
β_4	*-1.4*	*-1,01661**	*0,290166*	*-1,58532*	*-0,4479*
dummy2	*0.95*	*1,256641**	*0,244224*	*0,777972*	*1,735311*
dummy3	*-0.25*	*-0,12748**	*0,07887*	*-0,27654*	*0,03252*
dummy4	*1.65*	*1,53036**	*0,2552*	*1,030178*	*2,030542*
dummy5	*-0.64*	*-0,47281*	*0,201743*	*-0,86677*	*0,07885*

Numerosità	***4563***
Log Likelihood	***-558.46828***
$\hat{\sigma}^2_{\zeta}$	***5.931e-19***

Tabella 6.4: *Riassunto delle stime mediante metodologia gllamm*

Il pacchetto ***gllamm*** non fornisce alcun test per valutare statisticamente la presenza di eterogeneità non osservata.

Le principali indicazioni in termini di funzionamento delle stime prodotte nella prima simulazione, sono riscontrabili anche in questa seconda simulazione esemplificativa.

La lieve diversità riscontrata nelle stime dei parametri e della varianza dell'effetto casuale non sembra inficiare la distribuzione delle previsioni e la classificazione presentate in Figura 6.2.

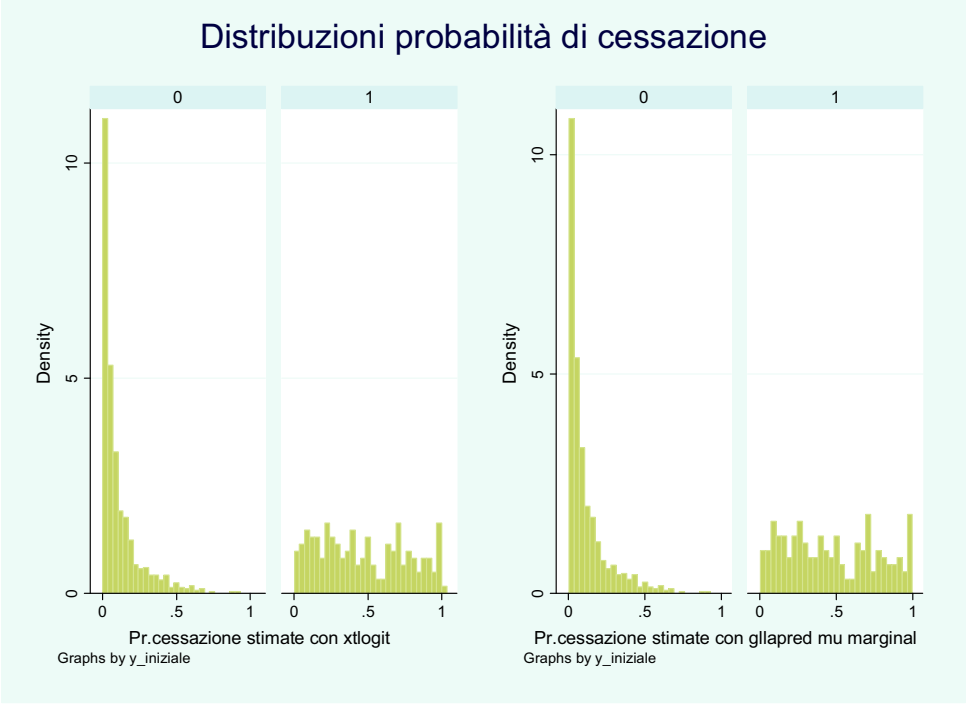


Figura 6.2: *Distribuzione delle probabilità stimate di cessazione : utilizzando xtlogit (a sinistra) e utilizzando gllapred mu marginal (a destra)*

Le distribuzioni di probabilità differiscono in maniera lieve, il "**prior prediction method**" non sembra quindi portare effettivi miglioramenti alle normali stime di probabilità che si ottengono con la previsione "classica".

In termini di classificazione inoltre non si nota alcuna capacità discriminante superiore della metodologia che sfrutta le stime *gllamm:* guardando infatti la figura 6.3, le *Roc Curve* tendono praticamente a sovrapporsi, indice che le probabilità stimate (che vanno a definire le soglie empiriche sulle quali si calcolano le varie combinazioni di *specificità* e *sensibilità* che costituiscono la curva) dall'uno e dall'altro metodo differiscono in modo irrilevante.

Il calcolo delle *Roc Area* e la relativa stima degli *Standard Error* è stato effettuato con il metodo non parametrico di De-Long e Clarke-Pearson (De Long *et al*., 1988).

Il Test utilizzato per comparare le *AUC* e implementato nella versione 9.2 di STATA è descritto in Cleves (2002). Il test, che in ipotesi nulla considera l' uguaglianza delle Aree, certifica chiaramente la sostanziale "indifferenza" in termini di capacità discriminante delle due metodologie utilizzate, e quindi che le stesse probabilità stimate sono praticamente le stesse. Nella Tabella 6.5 andiamo ad esporre i principali risultati del test implementato in Stata.

	ROC AREA	***Std.Error***	***Intervallo di confidenza al 95%***	
Stime xtlogit	*0.8787*	*0.0141*	*0.85112*	*0.90620*
Stime gllapred_mu_marginal	*0.8786*	*0.0141*	*0.85110*	*0.90619*

Tabella 6.5: *Stime Roc Area e relativi standard error*

Ho: area(p_xtlogit) = area(p_gllamm_s~l)

$\chi^2(1) =$ 0.18 Prob> $\chi^2 =$ 0.6701

Il valore del test è confrontato con un χ^2 , accetta chiaramente l'ipotesi nulla, verificando che statisticamente le *AUC* stimate non differiscono. In Figura 6.3 sono rappresentate le curve che sembrano praticamente sovrapporsi.

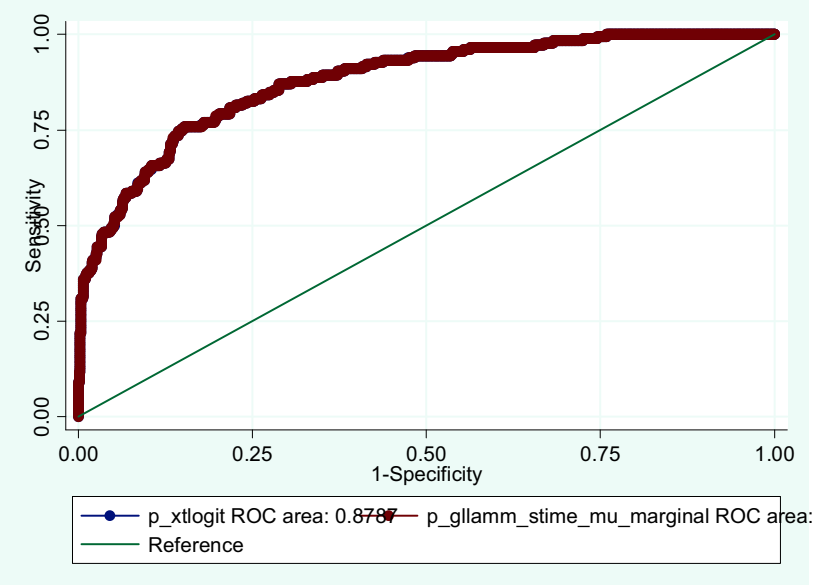


Figura 6.3: *Curve Roc e Roc Area*

6.1.2 Presenza di eterogeneità non osservata

In questa sezione si è deciso di confrontare l'approccio classico alla previsione dei modelli di durata a tempi discreti e l'approccio alla previsione del tipo "**prior prediction method**" nel caso di presenza di eterogeneità non osservata, riportando le conclusioni fondamentali raggiunte esemplificate su un determinato dataset.

Anche queste analisi sono state svolte su più dataset diversi in modo da testare effettivamente i risultati: sono stati simulati 40 dataset e, in termini previsivi e di stima dei parametri, le conclusioni in gran parte dei casi sono quelle esemplificate in questa sezione.

Dalle analisi effettuate si evince che a parità di modello stimato, aumentando la varianza del *random-effect* nella simulazione, il metodo di stima *gllamm* produce stime dei parametri β_j e σ_ζ^2 puntuali ed intervallari lievemente migliori. Come detto, il comportamento si è riscontrato nella quasi totalità dei dataset simulati e modelli *gllamm* e *xtlogit* stimati. Come in precedenza le variabili utilizzate in queste simulazioni sono 4, x_1 e x_4 realizzazioni casuali di una Distribuzione Uniforme [0,1], x_2 realizzazione casuale di una distribuzione Normale (0,1) e x_3 realizzazione casuale di una Bernoulliana.

Simulazione 1)

1) Immettiamo nella simulazione $\boldsymbol{\sigma_\zeta^2 = 4}$

Parametri	***Valori reali***	***Stime dei parametri con xtlogit***	***Standard Error xtlogit***	***Stime dei parametri con gllamm***	***Standard Error gllamm***
$\boldsymbol{\beta_0}$	*-2*	*-1,68052**	*0,258469*	*-1,68194**	*0,258828*
$\boldsymbol{\beta_1}$	*2.2*	*1,861741**	*0,316403*	*1,86085**	*0,316498*
$\boldsymbol{\beta_2}$	*1.3*	*1,189748**	*0,099283*	*1,189853**	*0,099421*
$\boldsymbol{\beta_3}$	*-2*	*-2,01344**	*0,188089*	*-2,01488**	*0,188645*
$\boldsymbol{\beta_4}$	*-1.2*	*-1,24986**	*0,306859*	*-1,24924**	*0,306908*
Numerosità		5000		5000	
Log Likelihood		*-1759.0419*		*-1759.0705*	
$\boldsymbol{\widehat{\sigma}_\zeta^2}$		***3.67923****		***3.6638505****	

Tabella 6.6: *Stime xtlogit e gllamm*

I risultati del Test Lr (Likelihood-ratio test of rho=0) sono :

$\chi^2_{bar}(01)$ = 480.06 $Prob >= \chi^2_{bar}(01) = 0.000$

Il test su ρ fornito da *xtlogit* stima significativamente la presenza di eterogeneità nei dati. Nelle analisi eseguite, non si riscontra un metodo di stima "superiore" che riesca ad effettuare stime intervallari e puntuali più precise. Di norma la metodologia che stima in maniera più precisa $\boldsymbol{\sigma^2_\zeta}$ riesce a stimare "meglio" anche gli stessi parametri chiaramente. Da questo particolare esempio, a fronte di una varianza immessa nella simulazione pari a 4, *xtlogit* riesce a stimare meglio puntualmente $\hat{\sigma}^2_\zeta$, e ciò si riflette sia nella stima dei parametri, sia nella precisione delle stime intervallari. Dal canto suo *gllamm* fornisce stime puntuali e intervallari che si discostano nell'ordine dei centesimi dalle stime *xtlogit.* La cosa interessante si è notata quando a parità di dataset simulato, inserendo progressivamente nella simulazione livelli di σ^2_ζ superiori, si vanno a stimare entrambi i modelli:

2) Immettiamo nella simulazione $\boldsymbol{\sigma^2_\zeta = 9}$

Parametri	***Valori reali***	***Stime dei parametri con xtlogit***	***Standard Error xtlogit***	***Stime dei parametri con gllamm***	***Standard Error gllamm***
$\boldsymbol{\beta_0}$	*-2*	*-1,73042**	*0,32349*	*-1,82303**	*0,366149*
$\boldsymbol{\beta_1}$	*2.2*	*2,025946**	*0,389006*	*2,122351**	*0,43115*
$\boldsymbol{\beta_2}$	*1.3*	*1,091934**	*0,116687*	*1,153697**	*0,132018*
$\boldsymbol{\beta_3}$	*-2*	*-1,81576**	*0,23326*	*-1,9243**	*0,267106*
$\boldsymbol{\beta_4}$	*-1.2*	*-1,30834**	*0,38024*	*-1,36703**	*0,424015*
Numerosità		5000		5000	
Log Likelihood		*-1883.6591*		*-1881.6966*	
$\boldsymbol{\hat{\sigma}^2_\zeta}$		***7.21373****		***7.9502174****	

Tabella 6.7: *Stime xtlogit e gllamm*

I risultati del Test Lr (Likelihood-ratio test of rho=0) sono :

$\chi^2_{bar}(01)$ = 1103.69 $Prob >= \chi^2_{bar}(01) = 0.000$

Anche in questo caso *xtlogit* si accorge perfettamente della presenza di etereogeneità non osservata nel dataset. Il modello *gllamm,* sebbene al pari di *xtlogit* sottostimi $\hat{\sigma}^2_\zeta$, si è notato che nella quasi totalità delle simulazioni eseguite riesce a stimare in maniera più accurata tale parametro e inoltre migliorano anche le stime puntuali (quelle che poi effettivamente vengono utilizzate ai fini previsivi) anche degli altri parametri. Chiaramente in entrambi i casi l'aumento della variabilità della componente non osservata va a scapito della precisione intervallare delle stime, gli *standard error* diventano sempre più grandi. Quindi in termini concreti, sembrerebbe più efficace utilizzare le stime *gllamm* in dataset di questo tipo in cui ci sia una forte componente di eterogeneità non catturata dalle variabili osservate. Ad esempio, quindi, se nel nostro dataset di aziende agricole non viene catturata dalle covariate una forte componente di variabilità (dovuta a diverse caratteristiche non osservate tra aziende) diventerebbe preferibile*,* al fine di una più efficace identificazione dei parametri "non osservati" che stanno alla base del modello, la procedura *gllamm* rispetto a *xtlogit.*

Simulazione 2)

Anche in questa seconda simulazione andiamo a inserire la variabile $anno_k$ rientrando in un contesto di modelli di durata a tempi discreti come descritto nella parte iniziale del capitolo. Nella Figura 6.4 viene rappresentata la distribuzione delle aziende per anno di sopravvivenza.

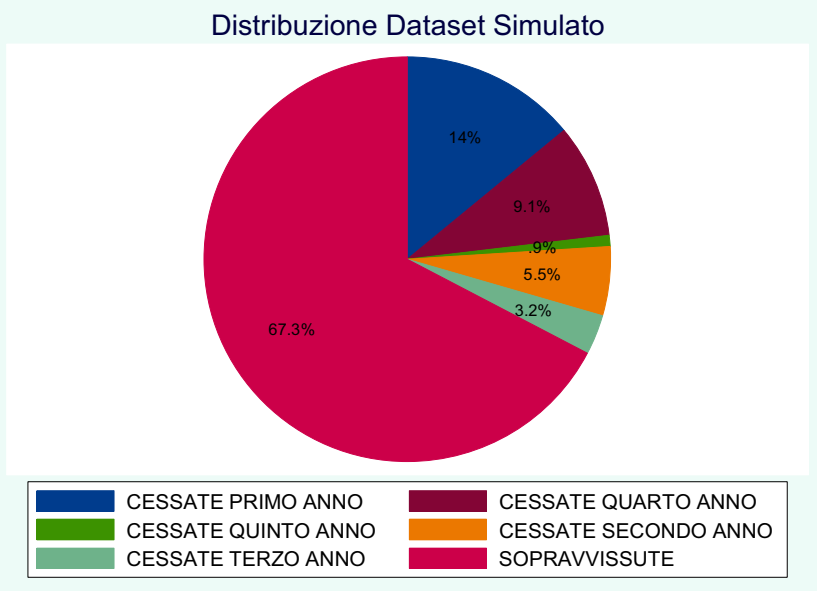


Figura 6.4: *Distribuzione per anni di sopravvivenza delle aziende agricole nel dataset simulato*

Immettiamo nella simulazione $\sigma_\zeta^2 = 4$

Parametri	***Valori reali***	***Stime dei parametri con xtlogit***	***Standard Error xtlogit***	***Stime dei parametri con gllamm***	***Standard Error gllamm***
β_0	-2.5	-2,02543*	0.210205	-2,42952*	0,511513
β_1	2.2	1,215227*	0,232637	1,553707*	0,456436
β_2	1.3	0,78934*	0,075481	1,001893*	0,242515
β_3	-2	-1,18515*	0,137005	-1,49928*	0,37556
β_4	-1.2	-0,68908*	0,224669	-0,93152*	0,378692
dummy2	1.43	0,72389*	0,153606	1,045791*	0,393589
dummy3	-1.25	-1,80304*	0,330226	-1,43744*	0,525708
dummy4	1.65	0,684246*	0,175486	1,22793*	0,613468
dummy5	-1.64	-3,18806*	0,77197	-2,62584*	0,941856
Numerosità		4120		4120	
Log Likelihood		-970.37417		-969.54129	
$\hat{\sigma}_\zeta^2$		***0.30340****		***1.7251739****	

Tabella 6.8: *Stime xtlogit e gllamm*

I risultati del Test Lr (***Likelihood-ratio test of rho=0***) sono :

$\chi^2_{bar}(01)$ = 1.38 $Prob >= \chi^2_{bar}(01) = 0.090$

Anche in questo caso *xtlogit* si accorge della presenza di etereogeneità non osservata nel dataset a un livello di significatività del test del 10% ma non al 5%.

Le principali indicazioni in termini di funzionamento delle stime prodotte nella prima simulazione non sono riscontrabili anche in questa seconda simulazione esemplificativa: le stime puntuali risultano sensibilmente diverse, probabilmente questo è in parte spiegabile dal fatto che le due metodologie utilizzino, come già visto in precedenza, metodi di quadratura diversi. Nelle simulazioni eseguite la metodologia utilizzata da *gllamm* sembra produrre risultati migliori in termini di precisione puntuale e intervallare nei casi di elevata variabilità dell'effetto casuale.

Dal punto di vista previsivo, utilizzare la metodologia che sfrutta le stime dei parametri del modello *gllamm* sembra essere più corretto. La procedura in *xtlogit* va a determinare le previsioni delle probabilità stimate di cessazione nell'ipotesi in cui il parametro ρ sia uguale a 0: quindi utilizzare questo metodo di previsione non tiene in considerazione il fatto che il test abbia verificato statisticamente la presenza di eterogeneità non osservata. Il "**prior prediction method**" da questo punto di vista ci da la possibilità di ottenere le stime di probabilità di cessazione tenendo in considerazione la "diversità" tra aziende agricole non catturata dalle covariate.

Tuttavia andando a verificare i risultati in termini classificativi attraverso l'analisi della *ROC Area* e della *ROC Curve* (Figura 6.6 e Tabella 6.9) si evince che il risultato fornito dalla previsione con *xtlogit* (che come si è visto non tiene conto dell'eterogeneità non osservata) non assume una differenza statisticamente significativa se si considera la previsione con *gllamm,* che tiene conto dell'eterogeneità non osservata.

In definitiva, quindi, l'approccio analitico alla *Curva Roc* ci aiuta a stabilire che in questa tipologia di dataset, tenendo in considerazione l'eterogeneità non osservata, lo stimare dei modelli a due livelli o stimare dei modelli di durata a tempi discreti ci fornisce le stesse indicazioni in termini classificativi.

Tali affermazioni sono state verificate anche in questo caso su più dataset diversi, riscontrando nella quasi totalità dei casi questo comportamento.

	ROC AREA	***Std.Error***	***Intervallo di confidenza al 95%***	
Stime xtlogit	*0.7690*	*0.0153*	*0.73902*	*0.79895*
Stime gllapred_mu_marginal	*0.7689*	*0.0153*	*0.73893*	*0.79883*

Tabella 6.9: *Stime Roc Area e relativi standard error*

Ho: area(p_xtlogit) = area(p_gllamm_s~l)

$\chi^2(1) =$ 0.13 Prob> $\chi^2 =$ 0.7147

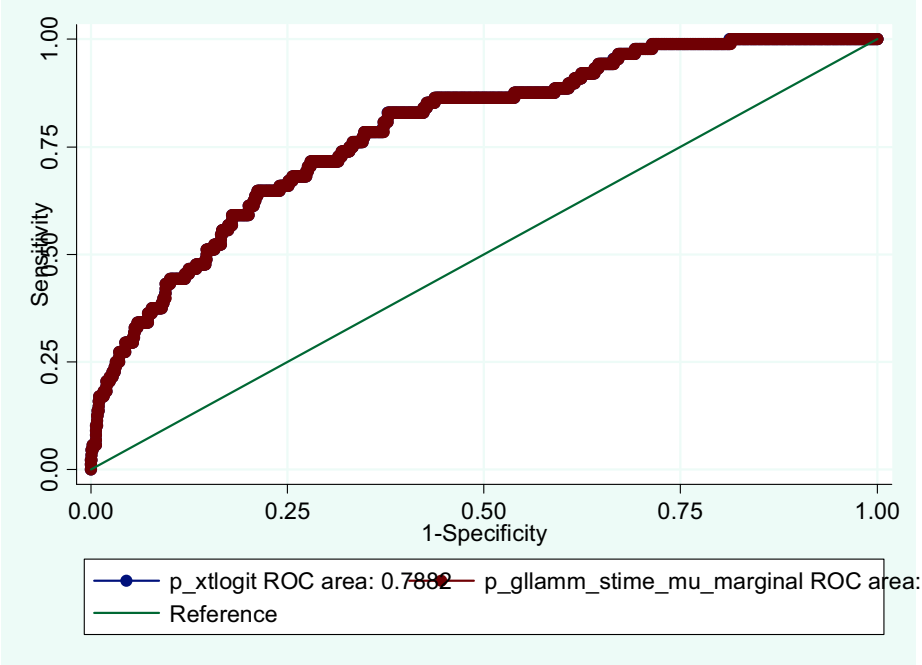


Figura 6.6: *Rappresentazione delle due Curve Roc*

Le due *Curve Roc* interpolate vanno praticamente a sovrapporsi, le *Roc Area* sono sostanzialmente le stesse, la diversità degli intervalli di confidenza ci fa percepire che in realtà non si tratta della stessa curva. Dalla stima delle *Roc Area* si deduce che secondo la classificazione empirica di Swets, entrambi i modelli forniscono delle

previsioni in termini classificativi "moderatamente accurate". Il test sull'uguaglianza delle *AUC* prova statisticamente che l'approccio predittivo del tipo "**prior prediction method**" e l'approccio predittivo "classico" essenzialmente forniscono gli stessi risultati in termini classificativi.
In conclusione inoltre, si è riscontrato che ai fini della stima delle probabilità di cessazione/sopravvivenza e successiva classificazione, in questa tipologia di dataset, il fatto di considerare o meno eterogeneità non sembra influenzare la classificazione stessa.

6.2 La "*cluster-averaged expectation"* e *gllapred mu*

In questa sezione si metterà invece in evidenza come la "**cluster-averaged expectation**" non sia un approccio utile in un contesto di stima di probabilità di sopravvivenza in modelli di durata a tempi discreti.
Da numerose simulazioni ci si è resi conto sostanzialmente di due cose fondamentali:

- Le stime delle probabilità di "cessazione" dipendono "fortemente" dal valore effettivamente realizzatosi.

- La distribuzione delle stime delle probabilità di "cessazione" migliora in termini classificativi all'aumentare della varianza stimata degli effetti casuali. A parità di condizioni infatti, più la stima della varianza dell'effetto casuale cresce più la distribuzione delle probabilità stimate diventa effettivamente più discriminante.

6.2.1 Le stime "*cluster-averaged expectation"* e i valori effettivi

Dal paragrafo 5.4 si evince che nel calcolo della previsione *cluster-averaged expectation*, utilizzando la *distribuzione a posteriori* degli effetti casuali, tra le variabili condizionanti viene considerata anche la stessa variabile dipendente:

$$\varpi\left(\zeta_j|y_j, X_j, \hat{\theta}\right) = \frac{g\left(\zeta_j; \hat{\Psi}\right) f\left(y_j|\zeta_j; X_j; \widehat{\theta^f}\right)}{g\left(y_j|X_j, \hat{\theta}\right)}$$

Si è deciso quindi di verificare empiricamente se effettivamente utilizzando questa metodologia di previsione, ci sia del condizionamento "forzato" della variabile dipendente, in questo modo si spiegherebbe in parte la capacità di tale metodo predittivo di dare risultati ottimali, osservata da Biasiolo (2006).

Per questo motivo si è deciso di stimare un modello *gllamm* su un dataset del tipo visto in Tabella 6.1; le previsioni con *gllapred* sono state calcolate sul dataset stesso e, per verifica, su uno identico in cui pero la variabile dipendente y_{it} fosse diversa, ricreata come visto nella parte iniziale del capitolo, modificando la soglia casuale.

Osservando l'integrale coinvolto nel calcolo di questa previsione (vedi formula seguente), con il metodo appena descritto, stiamo andando a verificare che questa non dipenda *ceteris paribus* dalla *y* effettivamente realizzata.

$$\tilde{\mu}_i(x^0) \equiv E_\zeta\left\{\hat{\mu}(x^0, \zeta_i)|y_{it}, X_i, ; \hat{\theta}\right\} = \int_{-\infty}^{\infty} \hat{\mu}(x^0, \zeta_i)\omega\left(\zeta_i|y_{it}, x_i; \hat{\theta}\right)d\zeta_i$$

Secondo le nostre ipotesi se la previsione non considerasse come parte integrante la realizzazione effettiva di y_{it} nei "cluster" su cui vogliamo fare previsione, dovremmo ottenere risultati previsivi identici sia nel primo come nel secondo dataset: la distribuzione delle probabilità stimate di cessazione e le performance di classificazione dovrebbero essere identiche.

Il secondo dataset è quindi un dataset di controllo che serve a verificare, come dovrebbe essere in un contesto di analisi previsive, che il metodo utilizzato non sfrutti l'informazione proveniente dalla realizzazione a posteriori della *y* (nella realtà infatti, se si vuole fare "previsione", a priori non si dovrebbe conoscere l'effettivo valore realizzato, cosa possibile solo in un contesto di verifica empirica dei modelli).

Riportiamo in questa sezione delle analisi esemplificative che utilizzano un dataset di 1000 unità statistiche osservate per 5 occasioni (abbiamo quindi complessivamente 5000 record) con 4 covariate, realizzazioni casuali di variabili Normali (0,1) *(x_2)*,

Uniformi *(x_1 e x_4)* e Bernoulli *(x_3)*, e una variabile utilizzata come *random effect* anch'essa Normale (0,1). Le variabili dipendenti sono generate come visto in precedenza.

Parametri	**Valori dei parametri utilizzati per la simulazione**	**Stime dei parametri con *gllamm***	**[Intervallo di confidenza al 95%]**	
$\boldsymbol{\beta_0}$	*-3.5*	*-3.408569**	*-3.81810*	*-2.9990*
$\boldsymbol{\beta_1}$	*2*	*2.15031**	*1.72213*	*2.57848*
$\boldsymbol{\beta_2}$	*-2*	*-2.050768**	*-2.22208*	*-1.8794*
$\boldsymbol{\beta_3}$	*2*	*2.020224**	*1.71215*	*2.3282*
$\boldsymbol{\beta_4}$	*-2.1*	*-2.464345**	*-2.89585*	*-2.0328*
$\boldsymbol{\sigma^2_\zeta}$	*1*	*1.0224804**		

Tabella 6.10: *Parametri veri e stimati da gllamm*

Dalla Tabella 6.10 notiamo che le stime sono tutte significative, la costante è sovrastimata come il parametro β_1 *e* β_3. La funzione *gllamm* inoltre sembra stimare in maniera praticamente esatta la varianza del *random effect.*

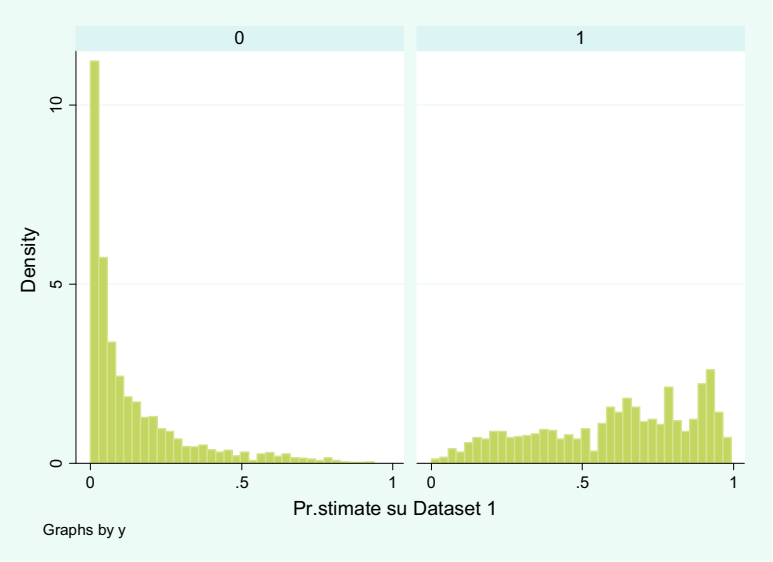


Figura 6.7: *Distribuzione delle probabilità stimate di cessazione sul Dataset1*

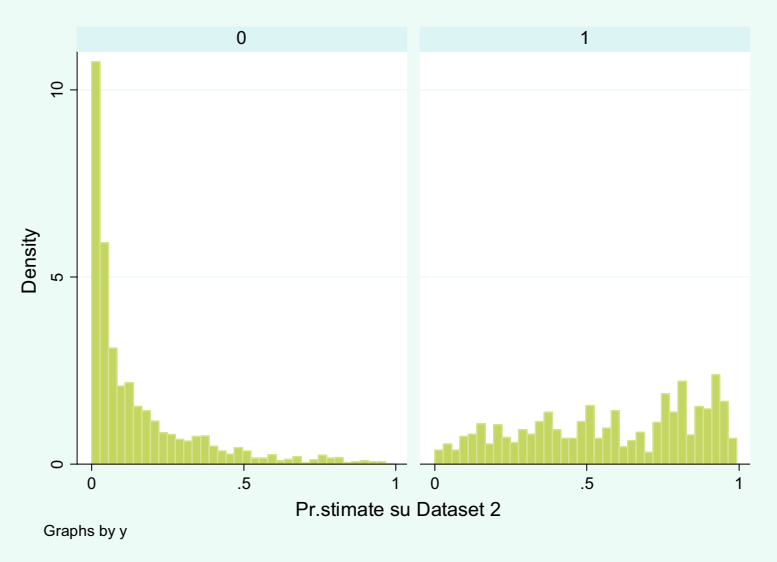


Figura 6.8: *Distribuzione delle probabilità stimate di cessazione sul Dataset2 di controllo*

Nelle Figure 6.7 e 6.8 andiamo a visualizzare la distribuzione rispettivamente delle probabilità stimate di cessazione nel Dataset1 e nel Dataset2; se il metodo di previsione non utilizzasse l'informazione a posteriori data dalla dipendente diversa presente nel Dataset2 le distribuzioni stimate dovrebbero essere le stesse. In questo caso anche la classificazione dovrebbe essere identica, ma come si evince dalla Figura 6.9 nemmeno le *curve Roc* stimate combaciano.

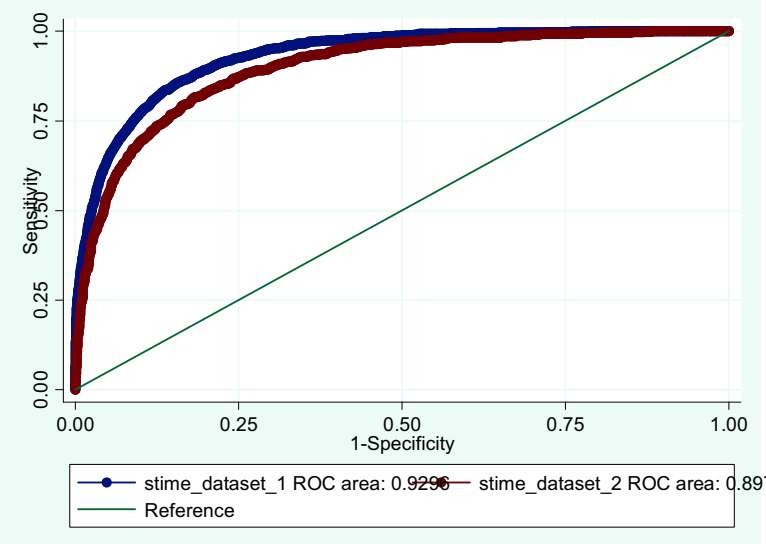


Figura 6.9: *Curve Roc e corrispondenti Roc-Area per le stime sul Dataset1 e sul Dataset2 di controllo.*

L'esempio riportato è esemplificativo del comportamento generale della previsione, ed è stato testato riproducendo diversi dataset: il comportamento registrato alla fine delle analisi è stato sempre lo stesso.

6.2.2 La varianza dell'effetto casuale e le stime "*cluster-averaged expectation"*

Nelle analisi effettuate è emersa un'altra problematica di tale metodo previsivo, relativa all'effetto della varianza stimata del *random effect* sulle previsioni e di conseguenza sulla capacità classificatoria del modello.

Nelle analisi di simulazione, infatti, si è notato che stimando i modelli con *gllamm* ed effettuandone le previsioni con il metodo del ***"cluster-averaged expectation"***, aumentando la varianza dell'effetto casuale sullo stesso dataset la capacità classificatoria del modello tende ad aumentare di precisione.

Anche in questo caso esemplifichiamo le problematiche sorte attraverso la simulazione; si è ricreato anche in questa occasione un dataset della struttura utilizzata per le ***Simulazioni 1*** con le metodologie già spiegate: abbiamo 1000 unità statistiche osservate per 5 anni, il dataset ha quindi 5000 records.

In questa sezione siamo andati a confrontare le prestazioni in termini di classificazione stimando lo stesso modello con le stesse esplicative ma di volta in volta aumentando la varianza del *random effect.* Anche in questo caso utilizziamo come variabili esplicative delle realizzazioni casuali di variabili Normali (0,1) *(x_2)*, Uniformi *(x_1 e x_4)* e Bernoulli *(x_3)*, e una variabile utilizzata come *random effect* anch'essa Normale $(0,\sigma_\zeta^2)$. Le variabili dipendenti sono generate come visto nella formula all'inizio del capitolo.

Nelle Tabella 6.11 riportiamo le stime dei parametri ottenute con *gllamm* relativamente al caso in cui σ_ζ^2 imputato nella simulazione sia uguale a 1, 4 e 9.

A) $\sigma_\zeta^2 = 1; \hat{\sigma}_\zeta^2$ =0.950638

Parametri	***Valori reali***	***Stime dei parametri con gllamm***	***[Intervallo di confidenza al 95%]***	
β_0	*-2*	*-2,31859**	*-2,6373*	*-1,9998*
β_1	*2*	*2,168799**	*1,79252*	*2,54508*
β_2	*1*	*1,060915**	*0,93962*	*1,18221*
β_3	*-2*	*-1,94934**	*-2,1769*	*-1,7218*
β_4	*2*	*2,382458**	*2,01015*	*2,75476*

B) $\sigma_\zeta^2 = 4; \hat{\sigma}_\zeta^2$ =3.46237

Parametri	***Valori reali***	***Stime dei parametri con gllamm***	***[Intervallo di confidenza al 95%]***	
β_0	*-2*	*-2,52253**	*-2,99156*	*-2,0535*
β_1	*2*	*2,235943**	*1,688085*	*2,78380*
β_2	*1*	*1,08973**	*0,919003*	*1,26045*
β_3	*-2*	*-1,86762**	*-2,20245*	*-1,5327*
β_4	*2*	*2,541933**	*1,998192*	*3,08567*

C) $\sigma_\zeta^2 = 9; \hat{\sigma}_\zeta^2$ =7.629386

Parametri	***Valori reali***	***Stime dei parametri con gllamm***	***[Intervallo di confidenza al 95%]***	
β_0	*-2*	*-2,56472*	*-3,24686*	*-1,8825*
β_1	*2*	*2,206138*	*1,409012*	*3,00326*
β_2	*1*	*1,108026*	*0,869235*	*1,34681*
β_3	*-2*	*-1,66367*	*-2,15154*	*-1,1758*
β_4	*2*	*2,465309*	*1,676697*	*3,25392*

Tabella 6.11: *Sintesi delle stime gllamm a livelli di varianza 1, 4 e 9*

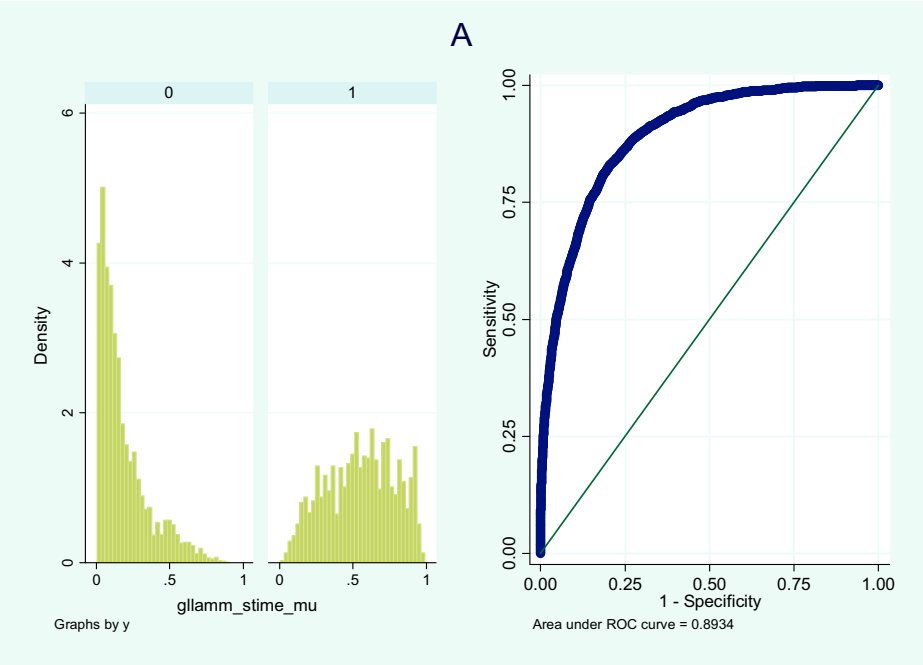


Figura 6.10: *Distribuzione delle probabilità stimate e relativa Roc Area, varianza bassa*

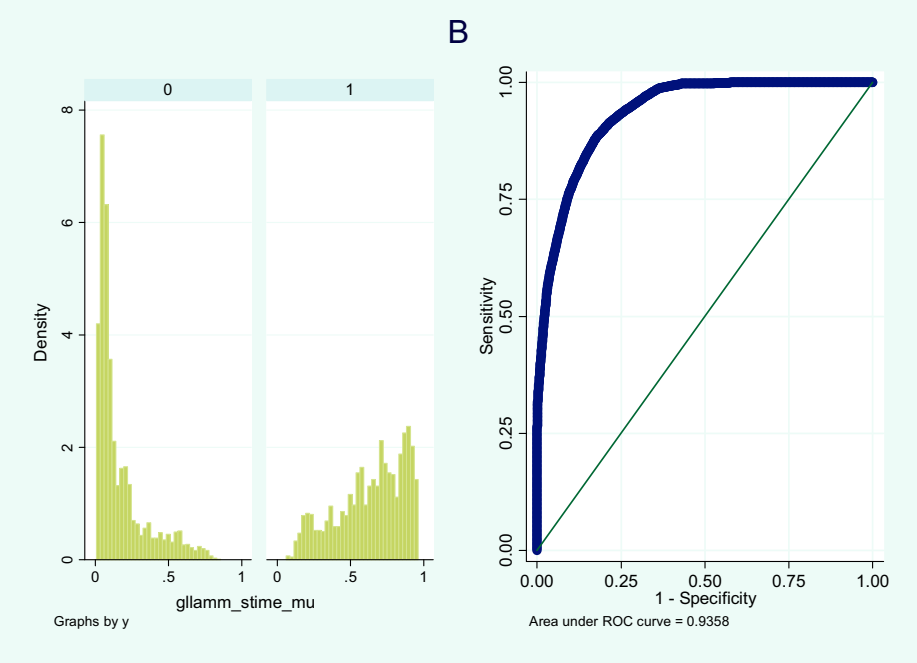


Figura 6.11: *Distribuzione delle probabilità stimate e relativa Roc Area, varianza media*

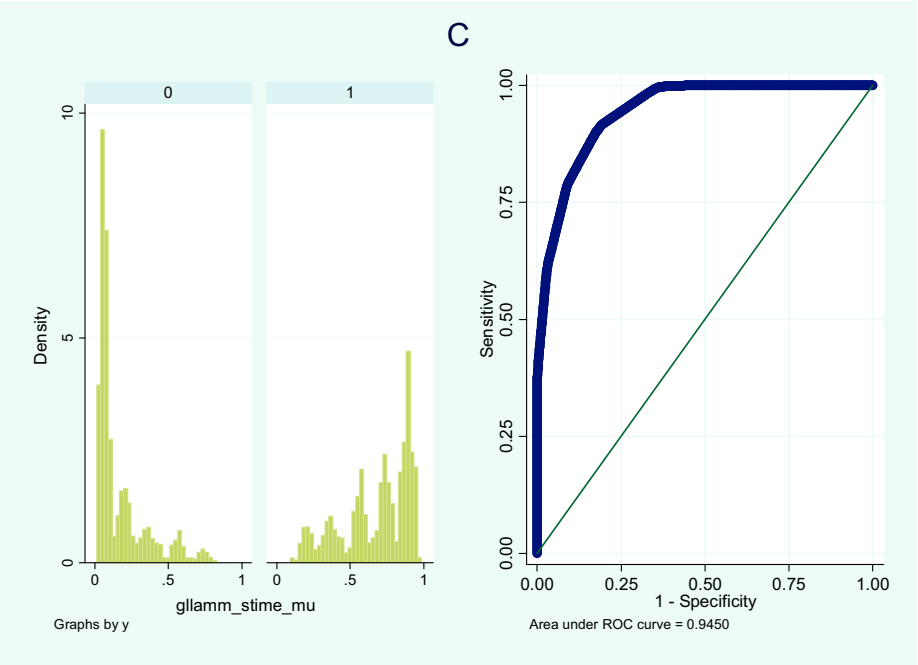


Figura 6.12: *Distribuzione delle probabilità stimate e relativa Roc Area, varianza alta*

Dalle Figure 6.10, 6.11 e 6.12 si evince che a parità di informazione contenuta nelle covariate e aumentando gradualmente la componente di eterogeneità non osservata nel dataset e quindi l'informazione non catturata dal modello stesso, le distribuzioni delle probabilità stimate diventano gradualmente più discriminanti.

La stessa *Roc Area,* che come abbiamo visto in precedenza è un indicatore della capacità classificatoria dei modelli, tende gradualmente ad aumentare.

I risultati che ritroviamo in questo esempio sintetizzano il comportamento generale della previsione: è stato testato effettuando le stesse analisi su altri 20 dataset simulati nella stessa maniera. Si è inoltre verificata la capacità classificatoria e il comportamento stesso delle previsioni stimando i modelli sul 50% del campione ed estendendo le previsioni all'altra metà, e le conclusioni sono state le medesime.

Da queste analisi e dalle precedenti è chiaro come il metodo previsivo ***"cluster-averaged expectation",*** che nel pacchetto **Gllamm** di STATA è implementato attraverso il comando *gllapred* corredato dell'opzione *mu*, non sia la metodologia più adatta utilizzabile per stimare probabilità di cessazione in modelli di durata a tempi discreti. Quindi le conclusioni ottenute da Biasiolo (2006) sono sostanzialmente state smentite.

L’applicazione di questo metodo previsivo probabilmente diventa più efficace per ottenere delle “post-diction” (predizioni dopo il fatto) in altre tipologie di analisi: ad esempio, per dati longitudinali binari, possono essere utilizzate per generare grafici relativi alle traiettorie individuali di crescita per meglio comprendere aspetti del modello e dei dati stessi (Skrondal e Rabe-Hesketh, 2008).

7. Applicazioni ai dati reali

In quest' ultimo capitolo andiamo a riprendere i risultati ottenuti nel capitolo 4 con *xtlogit* sul dataset originale e andiamo a confrontarli con i risultati che si ottengono tramite il pacchetto *gllamm.* Inoltre andiamo a verificare per entrambi le stime delle probabilità di cessazione, valutandole poi attraverso l'approccio della *curva Roc.* Allo stesso modo andremo ad applicare le stesse metodologie prendendo in considerazione tutte le variabili (stimando così il modello completo). Sarà anche interessante utilizzare l'approccio di tipo *Roc Analisys* per confrontare i risultati in termini classificativi del modello ridotto e del modello completo.

7.1 Modello ridotto

In questo paragrafo riprendiamo prima di tutto il modello già stimato nel paragrafo 4.2 riportando le stime effettive dei parametri ottenute con l'approccio "classico" e andando a confrontarle con le stime *gllamm*. Inoltre proponiamo un raffronto delle stime di probabilità previste con il metodo *Roc.* Si è deciso di utilizzare inizialmente le variabili OTE e UDE, che come si è visto in precedenza sono indicative rispettivamente dell'orientamento tecnico-economico delle aziende e della dimensione economica aziendale.

La Tabella 7.1 presenta il confronto fra i due modelli. In relazione alle stime dei parametri, i modelli sembrano comportarsi in maniera sostanzialmente diversa anche se le indicazioni generali che se ne traggono sono relativamente simili: entrambi i metodi non reputano significative le stime dei parametri relativi all'insieme delle variabili OTE, l'orientamento tecnico economico delle aziende agricole sembra quindi non essere rilevante nella determinazione del rischio di cessazione delle imprese nel quinquennio considerato. Ambedue i metodi rilevano significatività statistica solamente nella stima del parametro *Ote_2*, il rischio di cessazione tende quindi ad aumentare per quelle aziende che sono specializzate nell'ortofloricoltura.

Entrambi i metodi invece ritengono statisticamente rilevanti le stime dei parametri delle variabili UDE, inerenti quindi alla dimensione economica aziendale: si nota come al crescere della dimensione economica aziendale, il rischio di cessazione, a parità di condizione, tenda a diminuire monotonicamente.

Una differenza rilevante sta nella stima delle *dummies* legate agli anni di studio, la metodologia *gllamm* da questo punto di vista stima dei parametri significativi, cogliendo come il rischio di cessazione tenda ad aumentare nei primi 4 anni di vita delle imprese e a diminuire al quinto, fermo restando comunque che nei 5 anni analizzati sia pressoché elevato. Da questo punto di vista la metodologia che sfrutta la funzione *xtlogit* non coglie alcuna significatività legata ai medesimi parametri, non evidenziando quindi alcuna sostanziale differenza nella variazione del rischio di cessazione nel quinquennio considerato. Un particolare inoltre molto rilevante è il fatto che la metodologia *gllamm* stima una componente di varianza dell'effetto casuale molto più corposa rispetto alla funzione *xtlogit*. Inoltre i risultati del Test Lr (Likelihood-ratio test of rho=0) sono :

$\chi^2_{bar}(01)$ = 1.73 $\qquad\qquad$ $Prob >= \chi^2_{bar}(01) = 0.094$

Il test su ρ mette in evidenza quindi che, considerando un livello di significatività del 5%, si accetta l'ipotesi nulla di assenza di eterogeneità non osservata, si rifiuta invece con un livello di significatività del test al 10%.

Variabile	*Coeff. gllamm*	*Odds Ratio*	*Standard Error*	*Coeff. xtlogit*	*Odds Ratio*	*Standard Error*
Base_Ote1	*0,779303*	*2,179953*	*1,07062*	*0,589147*	*1,80245*	*0,652662*
Ote2	*1,96359*$^{\tau}$	*7,124859*	*1,19068*	*1,426634**	*4,16465*	*0,719878*
Ote3	*-0,00445*	*0,995557*	*1,10333*	*0,077123*	*1,08017*	*0,679006*
Ote311	*-1,18134*	*0,306868*	*1,14550*	*-0,70541*	*0,49390*	*0,711546*
Ote312	*0,109788*	*1,116042*	*1,09882*	*0,151713*	*1,16382*	*0,673657*
Ote4	*-0,00622*	*0,993802*	*1,09573*	*0,121327*	*1,12899*	*0,671331*
Ote5	*0,174023*	*1,190083*	*1,59540*	*0,184872*	*1,20306*	*1,009592*
Ote6	*0,191135*	*1,210623*	*1,09765*	*0,222773*	*1,24953*	*0,673329*
Ote7	*-0,91112*	*0,402074*	*1,54452*	*-0,48726*	*0,61430*	*0,990392*
Ote8	*-0,44867*	*0,638477*	*1,16952*	*-0,17981*	*0,83542*	*0,723212*
Ude2_4	*-0,63803**	*0,528333*	*0,25512*	*-0,36754***	*0,69243*	*0,139121*
Ude4_6	*-2,01432***	*0,133411*	*0,41345*	*-1,20448***	*0,29984*	*0,19435*
Ude6_8	*-2,04698***	*0,129124*	*0,46081*	*-1,22975***	*0,29236*	*0,236308*
Ude8_12	*-2,41278***	*0,089566*	*0,47667*	*-1,44118***	*0,23664*	*0,226315*
Ude12_16	*-2,73869***	*0,064655*	*0,54773*	*-1,71247***	*0,18041*	*0,289442*
Ude16_40	*-3,15729***	*0,042541*	*0,51087*	*-2,03393***	*0,13082*	*0,235676*
Ude40_100	*-4,10272***	*0,016528*	*0,77257*	*-2,64286***	*0,07115*	*0,433504*
Ude100	*-4,4856***	*0,01127*	*1,20457*	*-2,7917***	*0,06131*	*0,745702*
Costante	*-3,51272***	*0,029816*	*1,12725*	*-2,45801***	*0,08560*	*0,650684*
Dummy2	*0,739559***	*2,095012*	*0,25704*	*0,220024*$^{\tau}$	*1,24610*	*0,146309*
Dummy3	*1,021484***	*2,777313*	*0,34506*	*0,176803*	*1,19339*	*0,154111*
Dummy4	*1,076307***	*2,933825*	*0,40743*	*0,016379*	*1,01651*	*0,165812*
Dummy5	*0,929811**	*2,53403*	*0,45140*	*-0,27532*$^{\tau}$	*0,75932*	*0,18482*
Numerosità		*8013*			*8013*	
Log - Likelihood		*-1536.1264*			*-1540.0233*	
$\widehat{\sigma}_{\zeta}^{2}$		*4.5766821*			*0.297847*	
	Livelli di significatività	***τ: 15%***	****: 5%***		*****: 1%***	

Tabella 7.1: *Stime dei coefficienti Gllamm e Xtlogit*

Le differenze si attenuano se andiamo invece a confrontare, attraverso il metodo della *Curva Roc*, le diverse tipologie di previsione viste finora: il "*prior prediction method*" e la previsione classica per un modello a effetti casuali sfruttando la

funzione *xtlogit*. Si confermano invece ovviamente le differenze con il "*cluster-averaged expectation*", come si è visto non adatto per le previsioni.

Dalla Figura 7.1 si nota come la capacità classificatoria dell' ultimo metodo sia chiaramente superiore alle altre due, l'*AUC* stimata è $\cong 1$, siamo quindi di fronte riprendendo lo schema di Swets a dei risultati "altamente accurati" ma come si è visto non utilizzabili a fini "puramente" previsivi (vedi paragrafo 6.2). Anche per quanto riguarda gli altri due metodi, i risultati sembrano rispecchiare le conclusioni ottenute in ambito "controllato", per cui le capacità di previsione di *xtlogit* e *gllamm* sono essenzialmente le stesse, nonostante le differenze nelle stime puntuali.

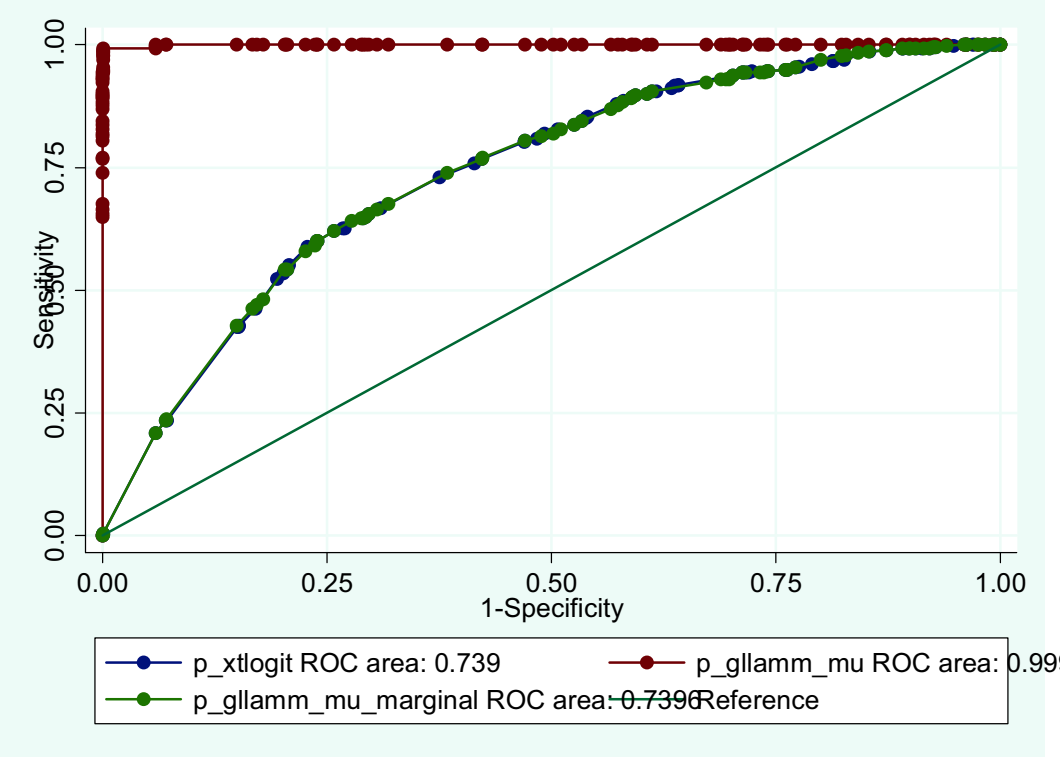


Figura 7.1: *Roc Curve e Roc Area relative ai tre diversi metodi di previsione*

I valori del test effettuato sulle *Roc Area* in Tabella 7.2 certificano la sostanziale uguaglianza delle aree: la statistica test assume un valore molto piccolo (0.23) a fronte di un p_value pari a 0.6339, si va quindi ad accettare l'ipotesi nulla.

$\chi^2(1) =$ 0.23 Prob> $\chi^2 =$ 0.6339

	ROC AREA	***Std.Error***	***Intervallo di confidenza al 95%***	
Xtlogit	*0.7390*	*0.0131*	*0.71334*	*0.76475*
Gllamm marginal	*0.7396*	*0.0131*	*0.71387*	*0.76524*

Tabella 7.2: *Stime Roc Area e relativi standard error*

La parità delle due metodologie in termini di stima delle probabilità di cessazione e successiva classificazione è ancora più evidente se andiamo a vedere la tabella 7.2 nella quale viene rappresentato un estratto della variazione in termini di specificità e sensitività e corretta classificazione, al variare delle soglie empiriche utilizzando i due diversi approcci.

Come regola generale empirica, si può affermare che il punto sulla curva *Roc* più vicino all'angolo superiore sinistro rappresenta il migliore compromesso tra *sensibilità e specificità* (Bottarelli e Parodi, 2003), e quindi un modo utile per scegliere il *cut-off*. Nell'estratto della tabella seguente si è deciso quindi di visualizzare per ciascun metodo alcune combinazioni.

Xtlogit				***Glamm marginal***			
soglia empirica	*Sensibilità*	*Specificità*	*Class. Corrette*	*soglia empirica*	*Sensibilità*	*Specificità*	*Class. Corrette*
0,047391	*99,30%*	*9,39%*	*30,82%*	**0,059091**	*99,30%*	*9,39%*	*30,82%*
0,047909	*99,30%*	*10,19%*	*31,43%*	**0,061871**	*99,30%*	*10,04%*	*31,32%*
0,04804	*99,30%*	*10,84%*	*31,93%*	**0,064107**	*99,30%*	*10,12%*	*31,37%*
.....	.....	.....	.....	.....	.....	.....	.....
0,211295	*66,74%*	*69,00%*	*68,46%*	**0,343315**	*66,51%*	*69,43%*	*68,74%*
0,222486	*65,58%*	*70,31%*	*69,18%*	**0,361176**	*65,58%*	*70,31%*	*69,18%*
0,234191	*64,88%*	*70,67%*	*69,29%*	**0,395354**	*64,88%*	*70,67%*	*69,29%*
0,276015	*64,65%*	*70,96%*	*69,46%*	**0,422645**	*64,65%*	*70,96%*	*69,46%*
.....	.....	.....	.....	.....	.....	.....	.....
0,409988	*42,56%*	*85,01%*	*74,89%*	**0,615876**	*42,79%*	*85,08%*	*75,00%*
0,41015	*23,49%*	*92,79%*	*76,27%*	**0,622796**	*23,72%*	*92,87%*	*76,39%*
0,410334	*23,49%*	*92,94%*	*76,39%*	**0,626404**	*23,49%*	*92,94%*	*76,39%*

Tabella 7.2: *Soglie di cut-off e relative Sensibilità, Specificità e Corretta Classificazione*

Dalla Tabella 7.2 si capisce come sostanzialmente per ogni soglia empirica utilizzata per generare la *Roc curve* con la funzione *xtlogit* vi sia una corrispondente soglia empirica ottenuta con la metodologia *gllamm* che mantiene i rapporti tra *Sensibilità* e *Specificità* e gli stessi complementi a 1 (*1- Sensibilità e 1- Specificità)* praticamente inalterati: questo significa che nelle corrispondenti tabelle di contingenza generate si hanno rapporti marginali di riga e di colonna identici.

Da ciò, sebbene si utilizzino soglie diverse per ogni metodologia, i risultati in termini di classificazione risultano eguali: il test sulle *Roc Area* va a dimostrare statisticamente proprio questo concetto.

Nel nostro particolare caso, seguendo quindi l'approccio sopradescritto, la Tabella 7.3 presenta le tabelle di contingenza per le principali soglie plausibili individuate :

Risultati utilizzando soglia 0.25 *xtlogit* e 0.40 per *gllamm marginal*			
Condizione prevista	*Condizione effettiva*		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+)	278	399	677
0=ancora attiva (-)	152	975	1127
Falsi positivi	*Pr(D|-)*		13.48%
Falsi negativi	*Pr(≠ D|+)*		58.94%
Classificazioni corrette			69.46%

Risultati utilizzando soglia 0.30 *xtlogit* e 0.48 per *gllamm marginal*			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+)	258	329	587
0=ancora attiva (-)	172	1045	1217
Falsi positivi	*Pr(D|-)*		14.13%
Falsi negativi	*Pr(≠ D|+)*		56.04%
Classificazioni corrette			72.23%

Tabella 7.3: *Tabelle di contingenza per le diverse soglie*

Ai fini di una politica di finanziamento alle aziende agricole, i risultati ottenuti possono essere utilizzati in modi diversi. Il principale consiste nel finanziare solo le imprese che hanno una probabilità di sopravvivere per almeno 5 anni superiore a una determinata soglia. Considerando tutte le aziende ugualmente meritevoli, la probabilità di sovvenzionare un'azienda che chiuderà entro 5 anni è pari al 23.84% ; tale probabilità scende al 13.48% se si finanziano solamente le imprese con probabilità stimata di cessazione inferiore a 0.25 se si utilizza la funzione *xtlogit* e 0.40 se si utilizza *gllamm marginal*. Tale decisione comporta però il mancato finanziamento del 59.84% delle aziende che sopravviveranno per più di 5 anni. Ovviamente, per costruzione, aumentando il valore della soglia, aumenta la probabilità di non finanziare le aziende che chiuderanno entro 5 anni, ma diminuisce la probabilità di non finanziare le aziende che continueranno a rimanere attive.

7.1.1 Analisi su un campione di prova

Si è deciso di testare la capacità previsiva, stimando il modello su un campione selezionato casualmente pari al 50% della popolazione e analizzando le previsioni estese all'altra metà della popolazione non utilizzata per la stima.
I parametri stimati da entrambe le metodologie rispecchiano in generale le conclusioni sopracitate utilizzando tutto il campione, per cui non sono qui riportati. Le Figure 7.2 e 7.3 e le Tabelle 7.4 e 7.5 riportano le analisi tramite *Roc Curve* .

	ROC AREA	***Std.Error***	***Intervallo di confidenza al 95%***	
Xtlogit	*0.7298*	*0.0190*	*0.69264*	*0.76704*
Gllamm marginal	*0.7285*	*0.0191*	*0.69106*	*0.76586*

Tabella 7.4: *Stime Roc Area e relativi standard error, campione di stima*

$\chi^2(1) =$ 0.70 Prob> $\chi^2 =$ 0.4013

	ROC AREA	***Std.Error***	***Intervallo di confidenza al 95%***	
Xtlogit	*0.7417*	*0.0182*	*0.70593*	*0.77741*
Gllamm marginal	*0.7415*	*0.0191*	*0.70559*	*0.77749*

Tabella 7.5: *Stime Roc Area e relativi standard error, campione di previsione*

$\chi^2(1) =$ 0.01 Prob> $\chi^2 =$ 0.9242

Anche in questo caso i valori del test effettuato sulle *Roc Area* certificano la sostanziale uguaglianza delle aree: la statistica test assume un valore molto piccolo a fronte di un p_value pari a 0.4013 per il campione di stima e 0.9242 per quello di previsione, per cui si va quindi ad accettare l'ipotesi nulla.

Entrambi i metodi forniscono secondo la classificazione di Swets dei risultati moderatamente accurati in termini predittivi, anche questo in linea con i risultati ottenuti effettuando le medesime analisi su tutto il campione.

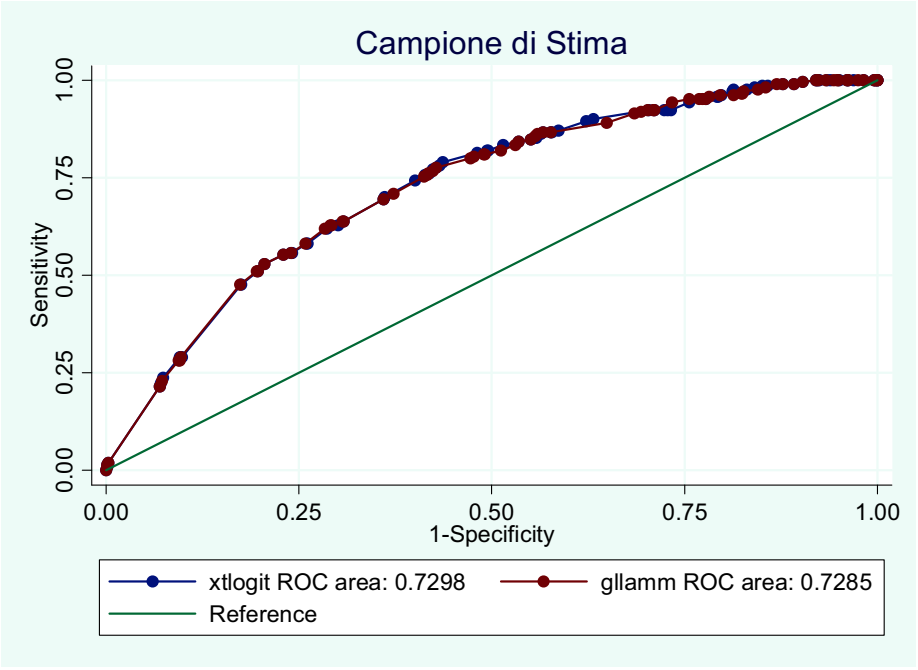


Figura 7.2: *Roc Curve e Roc Area relative alle previsioni ottenute con xtlogit e gllamm nel campione di stima*

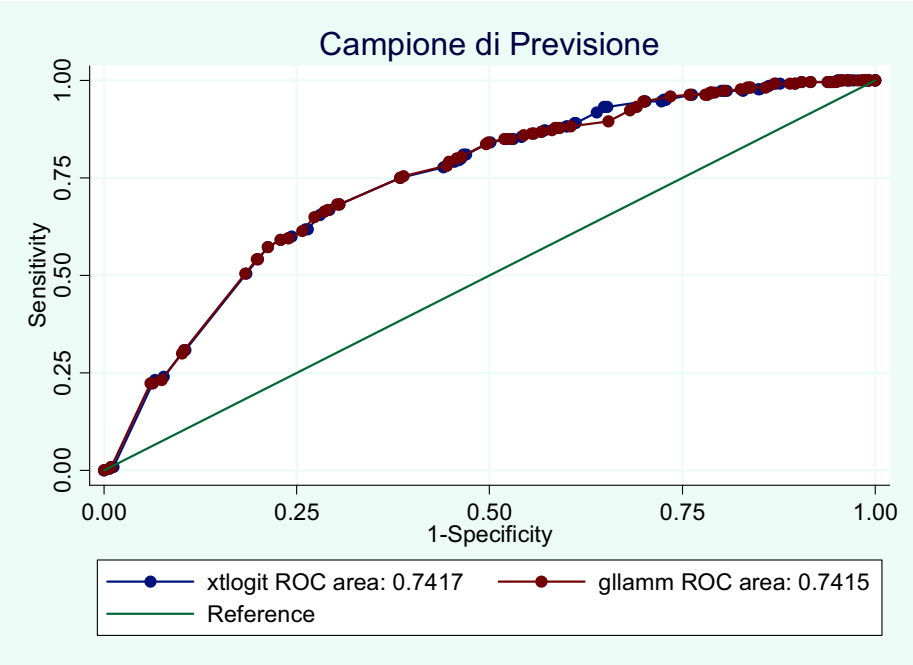


Figura 7.3: *Roc Curve e Roc Area relative alle previsioni ottenute con xtlogit e gllamm sul campione di previsione*

Seguendo il procedimento sopradescritto per la determinazione della soglia ottimale di *cutoff* , le soglie ottimali sembrano essere praticamente le stesse sia utilizzando tutto il campione, sia utilizzando il solo campione di stima o il solo campione di previsione. Di seguito riportiamo le tabelle di contingenza (Tabella 7.6) relative al campione di previsione utilizzando le medesime soglie ottimali viste precedentemente.

In questo caso particolare, utilizzando le soglie 0.25 e 0.40, rispettivamente per *xtlogit* e *gllamm*, i risultati non sono identici ma si differenziano per circa mezzo punto percentuale: *xtlogit* fornisce una probabilità di finanziare aziende che in realtà cessano la loro attività, superiore ai risultati offerti da *gllamm*, ma allo stesso tempo fornisce un risultato migliore se si valuta la probabilità di non finanziare aziende che in realtà sopravvivono (la probabilità stimata con *xtlogit* è infatti inferiore). In generale i risultati trovati sono in linea con le conclusioni generali ricavate sfruttando tutta l'informazione campionaria.

Risultati utilizzando soglia 0.25 *xtlogit* e 0.40 per ***gllamm marginal***			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+) 0=ancora attiva (-)	147/***150*** 73/***70***	199/***208*** 483/***474***	346/***358*** 556/***544***
Falsi positivi	*Pr(D|-)*		13.13% ***12.87%***
Falsi negativi	*Pr(≠ D|+)*		57.51% ***58.10%***
Classificazioni corrette			69.84% ***69.18%***

P.S. in grassetto i risultati per gllamm

Risultati utilizzando soglia 0.30 *xtlogit* e 0.48 per *gllamm marginal*			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+) 0=ancora attiva (-)	131 89	163 519	294 608
Falsi positivi	*Pr(D|-)*		14.64%
Falsi negativi	*Pr(≠ D|+)*		55.42%
Classificazioni corrette			72.06%

Tabella 7.6: *Tabelle di contingenza relative al campione di previsione*

7.2 Modello completo

In questa sezione proviamo a utilizzare tutte le variabili presenti nel nostro dataset, cercando di valutare se effettivamente ci siano effettivi miglioramenti in termini di determinazione della probabilità di cessazione. Utilizzando tutte le variabili a nostra disposizione, in Tabella 7.7 si nota che si coglie parte delle differenze tra le aziende non catturate in precedenza dal modello ridotto, riducendo teoricamente in maniera significativa la componente di eterogeneità non osservata

Tabella 7.7: *Stime xtlogit e gllamm (modello completo)*

Variabile	*Coeff. gllamm*	*Odds Ratio*	*Coeff. xtlogit*	*Odds Ratio*
adesione_a~d	*0,045456*	*1,046504*	*0,046053*	*1,04713*
adesione_s~p	*-0,21874*^{T}	*0,803529*	*-0,21858*^{T}	*0,803659*
adesione_c~i	*-0,23296*	*0,792185*	*-0,23477*	*0,790753*
allev_avic~i	*-3,6E-05*	*0,999964*	*-3,6E-05*	*0,999964*
bovini	*-0,00194*	*0,998062*	*-0,00196*	*0,998047*
cond_prof_~u	*0,018619*	*1,018794*	*0,0182*	*1,018366*
conigli	*-5,1E-05*	*0,999949*	*-5,2E-05*	*0,999948*
equini	*-0,2269*^{T}	*0,796998*	*-0,22731*^{T}	*0,796677*
parchi	*0,065359*	*1,067542*	*0,068247*	*1,07063*
lavoraz_pr~c	*-0,43381**	*0,648035*	*-0,43526**	*0,647094*
serre	*-5,21E-06*	*0,999995*	*-5,35E-06*	*0,999995*
sesso	*-0,20458**	*0,81499*	*-0,20584**	*0,81396*
suini	*0,000322*	*1,000322*	*0,000325*	*1,000325*
sup_sau_az~a	*-0,00045*	*0,99955*	*-0,00044*	*0,999557*
sup_sau_tot2	*-4,26E-08*	*1*	*-4,35E-08*	*1*
lav_fam	*0,000179*	*1,000179*	*0,00018*	*1,00018*
cond_dir	*0,105448*	*1,111209*	*0,105846*	*1,111651*
val_prod~10m	*0,010473*	*1,010528*	*0,008935*	*1,008975*
val_prod~50m	*0,10834*	*1,114426*	*0,107215*	*1,113173*
prod_bio	*0,432429*	*1,540997*	*0,433967*	*1,543368*
nr_abit	*-0,13641**	*0,872482*	*-0,13755**	*0,871491*
ovicapri	*-0,06105*	*0,940779*	*-0,06135*	*0,940498*
laurea	*-0,31209*	*0,731914*	*-0,31246*	*0,731644*
diploma	*-0,26643*	*0,766113*	*-0,26651*	*0,766045*
lic_scu_inf	*-0,26952*	*0,763746*	*-0,27078*	*0,762782*
bel	*-0,51896*^{T}	*0,59514*	*-0,52254*^{T}	*0,593015*
rov	*0,158427*	*1,171666*	*0,15968*	*1,173136*
pad	*-0,21808*^{T}	*0,804062*	*-0,22006*^{T}	*0,802472*
vic	*-0,10315*	*0,901992*	*-0,10215*	*0,902896*
ven	*0,217678*^{T}	*1,243187*	*0,218118*^{T}	*1,243733*
ver	*-0,45474**	*0,634615*	*-0,45632**	*0,63361*
senza_sup	*-0,24483*	*0,78284*	*-0,24121*	*0,785679*
eta	*0,012443***	*1,01252*	*0,012539***	*1,012617*
eta2	*0,000162*	*1,000162*	*0,000164*	*1,000164*

lav_30_90	*0,291763*	*1,338786*	*0,292466*	*1,339728*
lav_0_30	*0,510871*^{T}	*1,666742*	*0,512519*^{T}	*1,669491*
lav_90_270	*0,403344*^{T}	*1,496821*	*0,405064*^{T}	*1,499399*
capo_azienda	*0,080438*	*1,083761*	*0,079354*	*1,082587*
cod_att101	*-0,09438*	*0,909934*	*-0,09508*	*0,909299*
cod_att102	*0,057709*	*1,059407*	*0,057056*	*1,058716*
cod_att103	*-0,10522*	*0,900122*	*-0,10554*	*0,899839*
cod_att104	*0,576189*^{T}	*1,779244*	*0,577736*^{T}	*1,782*
cod_att111	*-0,02383*	*0,976451*	*-0,02423*	*0,976061*
cod_att112	*-0,21647*	*0,805358*	*-0,21764*	*0,804416*
affitto	*-0,26701*^{T}	*0,765663*	*-0,26829*^{T}	*0,764689*
uso_grat	*0,239383*	*1,270465*	*0,23972*	*1,270894*
perc_altra	*-0,21156*	*0,809317*	*-0,21229*	*0,808733*
perc_boschi	*-0,0206*	*0,97961*	*-0,02049*	*0,979716*
perc_cer	*-1,01606*^{T}	*0,362019*	*-1,01577*^{T}	*0,362123*
perc_fiori	*-0,42684*	*0,652566*	*-0,42393*	*0,654472*
perc_foraggi	*-1,70118**	*0,182468*	*-1,70378**	*0,181994*
perc_frutta	*-0,64906*	*0,522538*	*-0,64571*	*0,524288*
perc_legno	*0,309732*	*1,363059*	*0,315155*	*1,370472*
perc_legumi	*-48,4822*	*8,8E-22*	*-48,7582*	*6,68E-22*
perc_olivo	*-0,50056*	*0,606193*	*-0,49862*	*0,607369*
perc_orti	*-0,82808*	*0,436885*	*-0,83993*	*0,43174*
perc_ortive	*-0,6114*	*0,542592*	*-0,60895*	*0,543922*
perc_patata	*0,59864*	*1,819642*	*0,600679*	*1,823356*
perc_piante	*-0,82577*^{T}	*0,437898*	*-0,82437*^{T}	*0,438513*
perc_prati	*-0,81858*^{T}	*0,441056*	*-0,82031*^{T}	*0,440297*
perc_sanu	*0,044401*	*1,045402*	*0,044582*	*1,04559*
perc_barb	*-2,06525**	*0,126787*	*-2,06924***	*0,126282*
perc_arbo	*0,050116*	*1,051393*	*0,050492*	*1,051788*
perc_vite	*-1,03028*^{T}	*0,356908*	*-1,02996*^{T}	*0,357022*
perc_vivai	*-2,35911*^{T}	*0,094504*	*-2,35743*^{T}	*0,094664*
ude2_4	*-0,24323**	*0,784089*	*-0,24533**	*0,782447*
ude4_6	*-0,75429***	*0,470344*	*-0,75806***	*0,468574*
ude6_8	*-0,66619**	*0,51366*	*-0,66974**	*0,511843*
ude8_12	*-0,69231**	*0,500418*	*-0,69763**	*0,497766*
ude12_16	*-0,73506**	*0,479475*	*-0,741**	*0,476639*
ude16_40	*-0,59861**	*0,549574*	*-0,60375**	*0,546756*
ude40_100	*-0,44664*^{T}	*0,639775*	*-0,45133*^{T}	*0,636779*
ude100	*-0,49102*^{T}	*0,612004*	*-0,49592*^{T}	*0,609008*

Base_ote1	*1,432099**	*4,187479*	*1,440858**	*4,224319*
ote2	*1,501426*^{T}	*4,488085*	*1,511964*^{T}	*4,53563*
ote4	*1,478921**	*4,388208*	*1,491398**	*4,443303*
ote5	*0,803143*	*2,232546*	*0,813279*	*2,255291*
ote6	*1,213509*^{T}	*3,365273*	*1,221333*^{T}	*3,391706*
ote7	*0,828819*	*2,290611*	*0,834439*	*2,303522*
ote8	*1,220037*^{T}	*3,387313*	*1,229892*^{T}	*3,42086*
ote311	*0,558506*	*1,748058*	*0,565094*	*1,759614*
ote312	*1,067539*	*2,908214*	*1,074913*	*2,929738*
ote3	*1,055063*^{T}	*2,872156*	*1,061874*^{T}	*2,891785*
dummy2	*0,072549*	*1,075246*	*0,072787*	*1,075501*
dummy3	*-0,12341*^{T}	*0,883904*	*-0,12378*^{T}	*0,883571*
dummy4	*-0,40141***	*0,669375*	*-0,40253***	*0,668626*
dummy5	*-0,78904***	*0,454281*	*-0,79097***	*0,453405*
_cons	*-2,44405***	*0,086809*	*-2,46302***	*0,085177*
Numerosità	*8013*		*8013*	
Log - Likelihood	*-1555,9373*		*-1558.7924*	
$\hat{\sigma}_{\zeta}^{2}$	*5,485e-16*		*0.039457*	

Livelli di significatività ***T: 15%*** ****: 5%*** *****: 1%***

Entrambi i metodi stimano una componente di varianza dell'effetto casuale ≅ 0 . Inoltre i risultati del Test Lr (Likelihood-ratio test of rho=0) sono :

$\chi^2_{bar}(01)$ = 1.98 $Prob >= \chi^2_{bar}(01) = 0.048$

Il test condotto mediante la procedura classica *xtlogit* ammette assenza di eterogeneità non osservata al limite della significatività al 5%, mentre *gllamm* la stima essenzialmente nulla.

Come si è già visto nel capitolo trattato in precedenza, nel caso in cui nella simulazione non viene inserita alcuna componente di eterogeneità non osservata, con entrambe le metodologie utilizzate si vanno a stimare parametri praticamente identici (si è desunto che la leggera differenza tra le stime fosse dovuta ai diversi metodi di quadratura utilizzata nel procedimento di massimizzazione della verosimiglianza).

Andando ad analizzare i risultati effettivi vediamo come una gran parte dei parametri stimati non siano statisticamente significativi a un livello almeno del 15%. E' interessante notare come il rischio di cessazione diminuisca in maniera significativa, secondo entrambi i metodi di stima, se la conduzione dell'azienda agricola è affidata ad una donna a parità delle altre condizioni. Si nota inoltre come le distinte locazioni geografiche delle aziende influenzino il rischio di cessazione: aziende bellunesi e veronesi risultano più favorite in tal senso, delle colleghe veneziane. Anche il parametro relativo all'età del conduttore aziendale risulta significativo, ma l'impatto che ha nella determinazione della probabilità di cessazione è relativamente piccolo.
Secondo entrambi i metodi inoltre, a parità delle altre condizioni, la superficie agricola dedicata alle differenti colture sembra avere raramente un impatto significativo sulla sopravvivenza delle aziende stesse: superfici dedicate alla coltivazione di barbabietole e dedicate alla coltura della vigna sembrano avere un significativo impatto sulla probabilità di cessazione, diminuendone il rischio, come anche la presenza di vivai. Come già visto nel modello adottato precedentemente, i parametri relativi alla dimensione economica risultano tutti fortemente rilevanti nella valutazione del rischio, ma non più in maniera inequivocabile con un andamento lineare monotono. Nel modello completo inoltre si rivelano significativi anche alcuni parametri relativi alla specializzazione colturale. La dimensione temporale sembra anch'essa avere un ruolo significativo nella determinazione della probabilità di cessazione, si nota essenzialmente una diminuzione del rischio col procedere degli anni.
Dalla Figura 7.4 è evidente che la distribuzione delle probabilità stimate di cessazione, come già visto nella sezione delle simulazioni, sia praticamente identica utilizzando le due diverse metodologie in assenza di eterogeneità non osservata. Le stime puntuali di tali probabilità differiscono nell'ordine dei centesimi. In assenza di eterogeneità non osservata i due metodi quindi forniscono essenzialmente risultati identici.

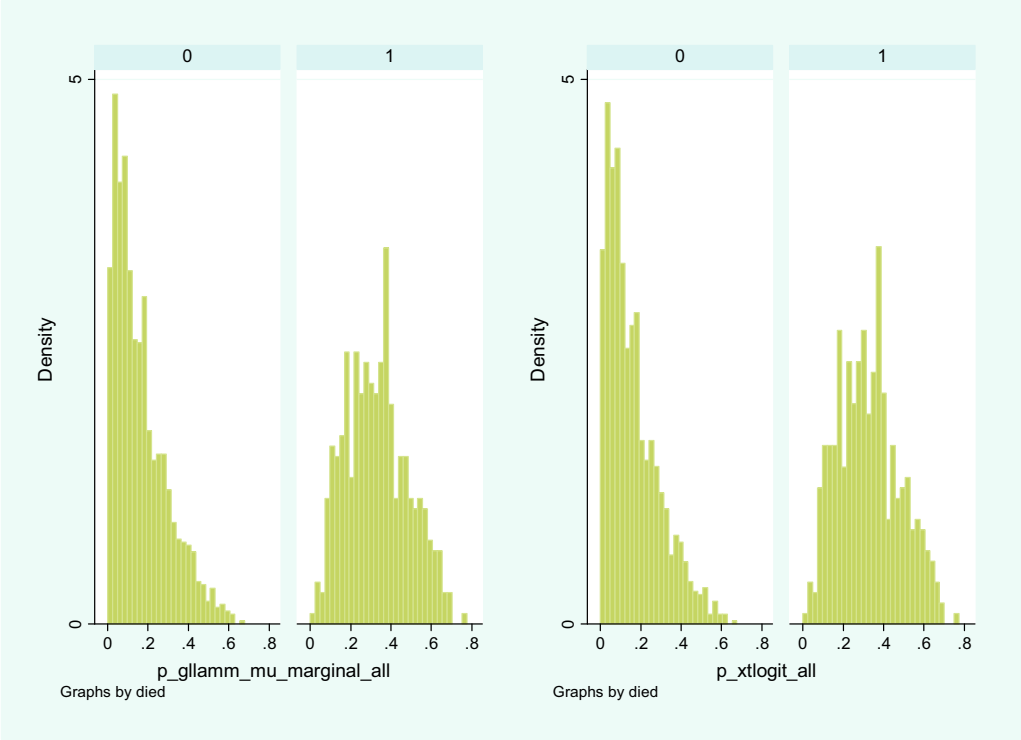


Figura 7.4: *Distribuzione delle probabilità di cessazione, a sinistra sfruttando metodologia gllamm marginal e a destra sfruttando xtlogit*

Le affermazioni sopra riportate sono supportate anche dall'analisi delle *Roc Curve* e delle *Roc Area*: le curve tendono a sovrapporsi in ogni punto quasi in maniera perfetta.

I valori del test effettuato sulle *Roc Area* certificano essenzialmente l' uguaglianza delle aree: la statistica test assume un valore pari a 2.58 e un p_value pari a 0. 1079, si va quindi ad accettare l'ipotesi nulla. Secondo la classificazione di Swets entrambe le metodologie forniscono dei risultati moderatamente accurati.

Diventa interessante andare a verificare se effettivamente il modello completo fornisca informazioni rilevanti in termini classificativi rispetto al modello ridotto. Per semplicità, visto che essenzialmente stime *gllamm* e stime *xtlogit* forniscono gli stessi risultati, utilizziamo l'analisi *Roc* sulla prima metodologia.

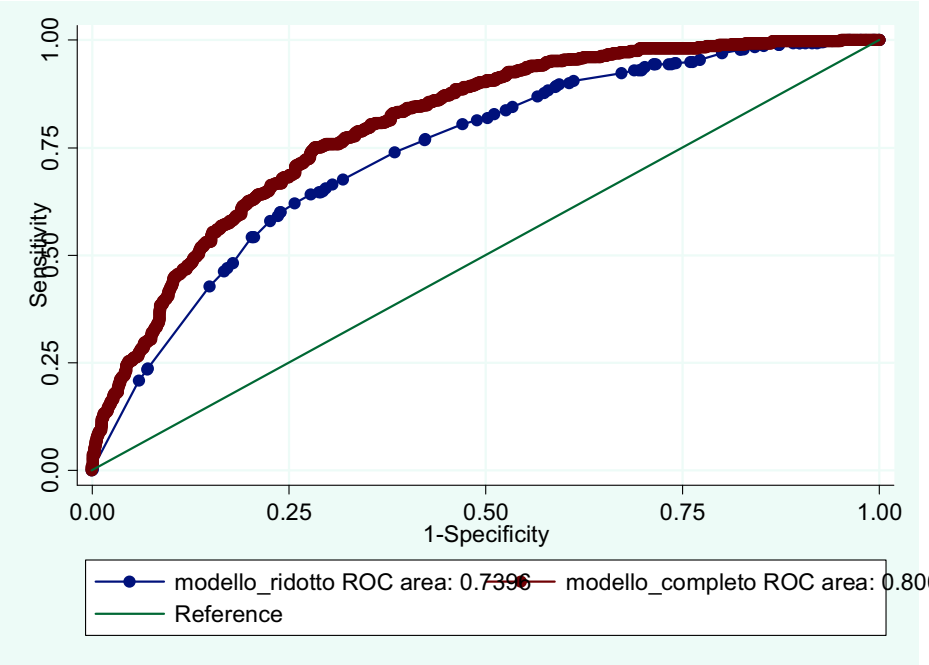


Figura 7.5: *Confronto Roc Area e Roc Curve relativamente al modello ridotto e al modello completo*

	ROC AREA	***Std.Error***	***Intervallo di confidenza al 95%***	
modello ridotto	*0.7396*	*0.0131*	*0.71387*	*0.76524*
modello completo	*0.8003*	*0.0114*	*0.77807*	*0.82261*

Tabella 7.8: *Stime Roc Area e relativi standard error*

Dai risultati del test in Tabella 7.8 si evince come effettivamente il modello completo superi in termini classificativi, e quindi di stima del rischio di cessazione, il modello ridotto.

Le *AUC* differiscono in maniera statisticamente rilevante: la statistica test assume un valore molto grande (44.85) a fronte di un p_value pari a 0, si va quindi a rifiutare l'ipotesi nulla che prevede uguaglianza delle Aree.

Dalla Figura 7.5 si riesce a carpire istantaneamente il vantaggio che si ha utilizzando il modello completo, in quanto la curva risulta superiore per ogni combinazione possibile di *Specificità* e *Sensibilità*.

Al fine di quantificare in termini classificativi il generale miglioramento, andiamo a controllare i risultati tramite tabella di contingenza, fissando ad esempio lo stesso

livello di *specificità* utilizzato per ricavare le soglie ideali nel modello ridotto: fissando ad esempio la soglia in corrispondenza di un livello di specificità del 76.06% (nel modello ridotto tale soglia ottimale era 0.48, nel modello completo passa a 0.25) si hanno i risultati in Tabella 7.9:

Risultati utilizzando soglia 0.25 per *gllamm marginal*			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva (≠ D)	Totale
1= cessata (+)	287	323	610
0=ancora attiva (-)	143	1051	1194
Falsi positivi	*Pr(D|-)*		11.98%
Falsi negativi	*Pr(≠ D|+)*		52.95%
Classificazioni corrette			74.17%

Tabella 7.9: *Tabella di contingenza per stime gllamm marginal con soglia 0.25*

Dalla Tabella 7.9 si evince come a parità di *specificità* (vedi paragrafo 7.1) il modello completo fornisca risultati generalmente migliori nell'ordine di 1.5/2 punti percentuali: utilizzando in questo caso una soglia pari a 0.25 abbiamo che il modello classifica correttamente il 74.17% delle osservazioni a fronte di un 72.23%.

Inoltre la probabilità di finanziare un azienda che nell'arco dei 5 anni andrà sicuramente a cessare l'attività è inferiore di circa due punti percentuali utilizzando il modello completo (11.98% contro i 14.13% del modello ridotto). Si ha anche un miglioramento nella stima della probabilità di non finanziare aziende che poi sopravvivranno, si passa da un 56.04% del modello ridotto al 52.95% del modello completo. Si è visto, comunque, come in entrambi i modelli la dimensione economica sia fortemente determinante per la sopravvivenza delle aziende prese in esame. In generale a soglia di specificità o sensibilità fissata, il modello completo fornisce performance migliori.

Rispetto all'obiettivo metodologico della tesi, le conclusioni sostanziali cui si giunge con l'approccio gllamm sono essenzialmente le medesime ottenute con i modelli di durata con effetti casuali, in particolare se si sfrutta tutta l'informazione disponibile.

Bibliografia

Afshartous D. e de Leeuw J. (2005) : Prediction in multilevel models, *J. Educ. Behav. Statist*., 30: 109-139.

Bamber D. (1975) : The area above the ordinal dominance graph and the area below the receiver operating characteristic graph*, Journal of Mathematical Psychology*, 12: 387-415.

Bassi F., Chillemi O. e Paggiaro A. (2010) : Survival of farms in the Veneto region of Italy 1999–2004, *Agribusiness*, 26 (in corso di pubblicazione).

Biasiolo S. (2006) : *"Stima della probabilità di chiusura dell'attività per le aziende agricole venete"*, Tesi di Laurea, Facoltà di Scienze Statistiche, Università degli studi di Padova.

Bottarelli E. e Parodi S. (2003) : Un approccio per la valutazione della validità dei test diagnostici: le curve R.O.C. (Receiver Operating Charateristic)*, Ann. Fac. Medic. Vet. di Parma*, 23: 49-68.

Carlin B. e Louis T. (2000) : *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn., Chapman and Hall, Londra.

Cioccolo V., Furfaro P., Mussolin S. e Piras P. (2004) : *IL PUNTO SU .... La Politica Agricola Comunitaria (PAC),* http://www.lavoro.gov.it/ ilpuntosupac.pdf.

Cleves M. (2002) : From the help desk: Comparing areas under receiver operating characteristic curves from two or more models*, The Stata Journal,* 2,3: 301–313.

D'Andrea M. (2006) : La riforma del 2003 di Fischler, *Agriregionieuropa*, 2, 4, Dicembre 2006 - associazione Alessandro Bartola (studi e ricerche di economia e politica agraria).

De Devitiis B. e Wanda Maietta O. (2009) : Capitale umano e produttività del lavoro agricolo nelle regioni dell'Unione Europea, *Agriregionieuropa*, 5, 16, Marzo 2009 – associazione Alessandro Bartola (studi e ricerche di economia e politica agraria).

DeLong E.R., DeLong D.M. e Clarke-Pearson D.L. (1988) : Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, 44: 837-45.

Gutierrez R., Carter S. e Drukker D. (2001) : On boundary-value likelihood-ratio tests, *Stata Technical Bulletin*, 60: 15-18.

Hanley J. A. (1982) : The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 143: 29-36.

Infocamere (2002) : *Comunicato Stampa Indagine Movimprese*, 3 trimestre 2002.

Istat (2002) : *Piano generale del V Censimento dell'Agricoltura,*
http: //censagr.istat.it

Jenkins G. (2004) : *Survival Analysis*, unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK,
http://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf.

Liu Q. e Pierce D. A. (1994) : A note on Gauss Hermite quadrature, *Biometrika,* 81:624-629.

Naylor J.C. e Smith A. F. M. (1982) : Applications of a method for the efficient computation of posterior distributions, *Applied Statistics*, 31:214-225.

Robbins H. (1955) : An empirical Bayes approach to statistics, in *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, 157-164, Berkley: University of California Press.

Skrondal A. e Rabe-Hesketh S. (2002) : Reliable estimation of generalized linear mixed models using adaptive quadrature*, The Stata Journal* 2,1:1-21.

Skrondal A. e Rabe-Hesketh S. (2004) : *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models,* Chapman Hall/CRC.

Skrondal A. e Rabe-Hesketh S. (2008) : *Multilevel and Longitudinal Modeling using Stata, 2nd edition,* College Station: Stata press.

Skrondal A. e Rabe-Hesketh S. (2009) : Prediction in multilevel generalized linear models*, Journal of the Royal Statistical Society A*, 172: 659-687.

Skrondal A. Rabe-Hesketh S. e Pickles A. (2004) : *GLLAMM Manual,* http://www.bepress.com/ucbbiostat/paper160/.

Stata (2005) : *Stata Longitudinal/Panel Data Reference Manual,* release 9, copyright *2005,* University of California, Berkeley.

Swets J.A. (1988) : Measuring the accuracy of diagnostic systems, *Science*, 240:1285-93.

Vieri S. (2001) : *Politica agraria comunitaria, nazionale e regionale*, Edagricole, Bologna, 2001.

Wooldridge J.M. (2002) : *Econometric Analysis of Cross Section and Panel Data,* MIT Press, Cambridge, MA.