Università degli Studi di Padova

Facoltà di Scienze Statistiche

Corso di Laurea Specialistica
in Scienze Statistiche Economiche,
Finanziarie e Aziendali

TESI DI LAUREA

Stima della probabilità di chiusura dell'attività per le aziende agricole venete

Relatore: Prof. FRANCESCA BASSI
Correlatore: Dott. ADRIANO PAGGIARO

Laureando: BIASIOLO SAMUELE

ANNO ACCADEMICO 2006/2007

Alla memoria di mio nonno Esaù.

"Fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza".

Dante Alighieri, Divina Commedia, Inferno, Canto XXVI.

Indice

Introduzione

Le recenti riforme della Politica Agricola Comunitaria (PAC), modificano sostanzialmente i criteri di sovvenzione alle imprese agricole: viene introdotta una maggiore selettività, nella scelta delle imprese da sovvenzionare, nel tentativo di avvicinarsi ad una logica di mercato.
A questo fine, è interessante valutare la "vitalità" delle imprese agricole, per comprendere le caratteristiche di quelle rimangono attive nel tempo, per consentire un utilizzo più efficiente dei fondi della PAC.

Questa tesi prende spunto un lavoro di Bassi et al. (2006), i quali hanno analizzato la sopravvivenza delle imprese agricole venete, utilizzando un nuovo dataset. Questo integra dati sulla sopravvivenza delle imprese, ricavati dagli archivi REA (Repertorio delle Notizie Economiche e Amministrative) del sistema delle CCIAA con i dati ISTAT del V Censimento Generale dell'Agricoltura.
Più precisamente, la popolazione oggetto di studio è data dalle imprese che soddisfano due requisiti:

- essere iscritte negli archivi REA a fine 1999,

- essere state rilevate nel Censimento.

Di tale popolazione viene ricostruita la sopravvivenza negli archivi REA sino alla fine del 2004.

Rispetto al contributo di Bassi et al. (2006), che prendeva in considerazione la coorte delle imprese nuove iscritte nell'anno 1999, questa ricerca analizza tutte le imprese presenti nel dataset.

Viene prima stimata, mediante modelli logit e di sopravvivenza, la probabilità di cancellazione dal registro REA in funzione delle caratteristiche

socio-economiche dell'impresa. Successivamente, si procede ad un'estensione dei modelli di sopravvivenza tenendo in considerazione l'eterogeneità non osservata.

La tesi è strutturata come segue: nel Capitolo 1 viene brevemente analizzata la riforma della PAC, nel Capitolo 2 si presentano: la popolazione oggetto di studio, la tabella dei regressori e alcune statistiche descrittive. Il Capitolo 3 include le stime della sopravvivenza nell'intero periodo di riferimento 1999-2004, ottenute con un modello logit. Nel Capitolo 4 viene stimato un modello di sopravvivenza a tempi discreti per il periodo dal 2000 al 2004. Nei capitoli 5 e 6 si procede ad un'estensione delle analisi: viene verificata la presenza di eterogeneità non osservata, e successivamente, utilizzando un approccio bayesiano empirico, viene stimata la probabilità di chiusura dell'attività, nell'intero periodo di riferimento, includendo l'eterogeneità non osservata.

Capitolo 1

La Politica Agricola Comunitaria

1.1 La Politica Agricola Comunitaria

La Politica Agricola Comunitaria (PAC) ha avuto origine nell'Europa occidentale degli anni cinquanta, dopo anni di guerra che avevano danneggiato il tessuto sociale e paralizzato l'agricoltura rendendo incerto l'approvvigionamento di viveri. [UE, 2005a] Originariamente, la PAC mirava a favorire l'incremento della produttività nella catena alimentare affinché i consumatori potessero contare su approvvigionamenti stabili di alimenti a prezzi accessibili, ma anche per garantire la redditività del settore agricolo comunitario. La PAC offriva agli agricoltori sovvenzioni e prezzi garantiti, incentivandoli così a produrre e forniva aiuti finanziari per la ristrutturazione del settore, ad esempio sostenendo gli investimenti nelle aziende agricole per garantirne lo sviluppo, sia dal punto di vista delle dimensioni, sia sotto il profilo delle capacità gestionali e tecnologiche, per adeguarsi al clima sociale ed economico dei tempi.

Malgrado la grande efficacia della PAC nel conseguire l'obiettivo dell'autosufficienza, negli anni ottanta la Comunità Europea si trovò a fare i conti con eccedenze quasi continue dei principali prodotti agricoli, alcuni dei quali erano esportati (con l'aiuto di sovvenzioni), mentre altri dovevano essere immagazzinati o eliminati all'interno della Comunità. Queste misure

avevano un costo di bilancio elevato, causavano distorsioni in alcuni mercati mondiali, non sempre erano nel pieno interesse degli agricoltori e divennero impopolari agli occhi dei consumatori e dei contribuenti. Nello stesso periodo andava crescendo, nella società, la preoccupazione per la sostenibilità ambientale dell'agricoltura: all'inizio degli anni novanta il vertice di Rio sulla Terra[1] rappresentò una tappa estremamente significativa sotto questo aspetto.

Nei primi anni di vita della Comunità, la PAC rappresentava una quota notevole delle spese di bilancio, superando in alcuni casi i due terzi del totale. La maggiore severità della disciplina di bilancio, la crescita delle attività comunitarie in altri settori e una serie di riforme della PAC hanno determinato una diminuzione della percentuale di risorse destinate alla politica agricola comune. La PAC, al giorno d'oggi, costa circa 50 miliardi di euro l'anno: meno del 50 % del bilancio comunitario. Meno dell'1% del PIL è speso per il 5,5 % della popolazione dedita all'agricoltura (dati relativi ai paesi dell'UE-15 prima dell'allargamento del 2004) [UE, 2005a].

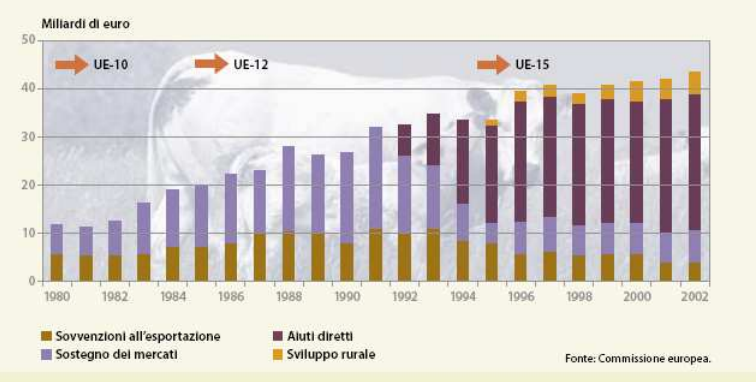


Figura 1.1: Evoluzione della spesa per la PAC.

[1] Conferenza delle Nazioni Unite sull'ambiente e lo sviluppo (United Nations Conference on Environment and Development −UNCED), svoltasi a Rio de Janeiro dal 3 al 14 giugno 1992.

1.2 La riforma della PAC

La PAC è stata più volte riformata negli ultimi anni per adeguare le norme comunitarie relative al settore agricolo ai nuovi equilibri del mercato comunitario ed internazionale, alle nuove esigenze finanziarie del bilancio dell' Unione Europea e alle nuove aspettative dei cittadini e dei consumatori.

Ben tre riforme sono state emanate negli ultimi 11 anni: la riforma MacSharry del 1992, la riforma di Agenda 2000 nel 1999 e quella più recente, denominata ormai comunemente riforma Fischler, che è diventata operativa dal primo Giugno 2005 [Frascarelli, 2004]. Quest'ultima, approvata con il compromesso di Lussemburgo del 25 Giugno 2003 e successivamente promulgata con i regolamenti del Consiglio e della Commissione, riveste una particolare importanza in quanto cambia radicalmente il modo in cui l'Unione Europea sostiene il settore agricolo, in una prospettiva di lungo periodo, visto che le previsioni finanziarie sono state fissate fino al 2013.

A partire dall'autunno 2004, gli agricoltori devono confrontarsi con regole nuove e con cambiamenti che andranno a influenzare in maniera sostanziale le scelte produttive aziendali.

La riforma si articola in sei punti essenziali.

1. Disaccoppiamento: un pagamento unico per azienda agli agricoltori UE, indipendente dalla produzione; sostituisce la maggior parte dei pagamenti diretti della PAC. Gli agricoltori, in linea di principio, riceveranno il pagamento unico per azienda sulla base delle somme percepite nel periodo di riferimento 2000-2002.

2. Condizionalità: il pagamento unico per azienda sarà condizionato al rispetto delle norme in materia di salvaguardia ambientale, sicurezza alimentare, sanità animale e vegetale, protezione degli animali e all'obbligo di mantenere la terra in buone condizioni agronomiche ed ecologiche.

3. Modulazione: riduzione di tutti i pagamenti diretti allo scopo di finanziare la nuova politica di sviluppo rurale. La modulazione si applica

alle aziende che ricevono più di 5000 euro/anno di pagamenti diretti, nelle seguenti percentuali -3% nel 2005, -4% nel 2006, -5% dal 2007 in poi.

4. Aumento delle risorse per lo sviluppo rurale: potenziamento della politica di sviluppo rurale, nuove misure a favore dell'ambiente, della qualità e del benessere animale, aiuto agli agricoltori per interventi che favoriscano l'adeguamento alle norme di produzione in vigore nell'UE.

5. Disciplina finanziaria: meccanismo finanziario atto a garantire, sino al 2013, il rispetto delle previsioni finanziarie della PAC.

6. Revisione di alcune organizzazioni comuni di mercato (OCM)[2]

7. Riforme di alcuni importanti OCM quali: settore lattiero caseario, riso, foraggi essiccati, olio di oliva e tabacco.

1.2.1 Il disaccoppiamento

Il disaccoppiamento rappresenta il cuore della nuova PAC. Uno dei privilegi più rilevanti del disaccoppiamento è quello di orientare l'agricoltura al mercato e ridurre le molteplici distorsioni indotte dal precedente regime di pagamento. Dal punto di vista della teoria economica, il disaccoppiamento viene visto come una misura auspicabile soprattutto per la sua capacità di restituire al mercato la sua funzione di determinare i prezzi, di rendere più trasparente il sostegno e, quindi, di orientare le scelte dei produttori in direzioni più rispondenti agli interessi della collettività. In quest'ottica, il regime unico di pagamento disaccoppiato conseguirà una maggiore rispondenza dell'offerta alla domanda dei consumatori e potrà portare un beneficio

[2]Le organizzazioni comuni di mercati (OCM) rappresentano il primo pilastro della PAC. Costituiscono lo strumento fondamentale di regolazione dei mercati nella misura in cui disciplinano la produzione e il commercio dei prodotti agricoli di tutti gli Stati membri dell'Unione Europea:

- eliminando gli ostacoli agli scambi intracomunitari di prodotti agricoli;
- mantenendo una barriera doganale comune nei confronti dei paesi terzi.

ai produttori, che potranno trarre pienamente vantaggio dalle opportunità di mercato.

Gli agricoltori temono che il disaccoppiamento possa costituire il primo passo verso il progressivo smantellamento del sostegno agricolo; invece, secondo la Commissione Europea, con il disaccoppiamento gli agricoltori beneficeranno di una PAC più semplice, senza pregiudizio per l'ammontare di aiuti che essi ricevono. .

Contestualmente, non si possono e non si devono nascondere i rischi del disaccoppiamento, che sono altrettanto importanti: tra questi vanno rilevati i rischi di abbandono dell'attività produttiva agricola da parte delle aziende meno competitive, soprattutto nelle zone montane e svantaggiate, dove gli agricoltori potrebbero "incassare" il pagamento unico disaccoppiato e disattivare la produzione, portandola al livello minimo richiesto dalla normativa (ad esempio, convertendo la produzione in prati o pascoli).
Tra gli svantaggi del disaccoppiamento vanno incluse anche le molteplici carenze di equità distributiva; effettivamente le modalità di attribuzione dei diritti all'aiuto cristallizzano il sostegno in base al comportamento degli agricoltori nel periodo 2000-2002, creando situazioni di disparità, che penalizzano gli agricoltori che in passato hanno adottato una buona pratica agricola, mediante rotazioni agrarie, mentre nel futuro penalizzano i giovani e le imprese che effettuano investimenti.

1.3 Le sovvenzioni alle imprese agricole

Con la nuova PAC le sovvenzioni alle aziende agricole dovrebbero rispondere maggiormente ad una logica di mercato, per cui i flussi di denaro saranno instradati verso soggetti che possono ripagare con certezza il loro "debito". Trattandosi di sovvenzioni, gli agricoltori non sono tenuti a restituire quanto ricevono, ma viene chiesto loro di mantenere "in vita" la propria azienda. La PAC è nata con questo obbiettivo: mantenere in vita le imprese agricole per garantire l'autosufficienza per i principali generi alimentari al fine di scongiurare nuove situazioni di penuria alimentare come quelle del dopoguerra. [UE, 2005a]

In Italia il 76,8% delle aziende agricole ha una dimensione inferiore ai 5

ettari [UE, 2005b], e il 79,4% utilizza una forza lavoro inferiore ad 1 unità (ULA). [Istat, 2003] Questo dato è direttamente collegato con la capacità dell'impresa di rimanere attiva nel tempo. Le piccole e grandi imprese non si differenziano soltanto, come parrebbe a prima vista per un elemento tecnologico (si pensi alla possibilità per una grande impresa agricola di acquistare macchine che una piccola impresa non può permettersi): le differenze di scala si traducono in un vantaggio *generale* a favore delle grandi imprese. Tra tutti i fattori che esercitano un'influenza sulle "fortune relative delle imprese di differente dimensione" la diminuzione dei costi unitari è un fattore importante, ma non l'unico: la scala, infatti agisce in modo pervasivo in tutti gli ambiti della vita dell'impresa. Le grandi imprese possono ottenere prezzi più bassi negli acquisti delle materie prime e dei semilavorati e conseguire economie nella gestione delle riserve accumulate. Considerazioni analoghe valgono per la distribuzione. Ad essere coinvolto, da ultimo, è lo stesso mercato del credito. Il fatto di poter disporre di un capitale proprio di dimensioni rilevanti permette alle grandi imprese di aver meno bisogno del credito e di ottenerlo a condizioni più vantaggiose rispetto alle piccole imprese [Solinas, 2005]. Le piccole imprese, d'altro canto, possono produrre beni maggiormente personalizzati e talora possono essere più flessibili. Per superare i loro limiti, possono associarsi tra di loro, aderire a consorzi, o a società cooperative. Per un nuova impresa, i primi anni di vita, di norma, sono tra i più critici: il recupero dei costi di avviamento e la scarsità dello stock di conoscenze possono annoverarsi tra le principali fonti di difficoltà. Le imprese italiane vivono in media quasi 12 anni, 1 impresa su 4 chiude entro 3 anni di vita e oltre 4 su 10 nei primi 5 anni. [Infocamere, 2002]
In questa analisi si considera *vitale* un'azienda con un'aspettativa di vita non inferiore ai 5 anni.

Capitolo 2

I dati

In questa ricerca viene utilizzato 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. 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.

2.1 Il Registro REA

Il Registro delle Imprese presso le Camere di Commercio è un'anagrafe giuridico-economica, completamente informatizzata, che assicura un sistema organico di pubblicità e informazione per le imprese. Sono tenuti all'iscrizione tutti i soggetti che svolgono attività economica. Tuttavia, in base all'art. 2, comma 3, della legge n. 77 del 1997 "i produttori agricoli che nell'anno solare precedente hanno realizzato un volume d'affari non superiore ai 2.500 €, sono esonerati da tutti gli obblighi documentali e contabili, e quindi la loro iscrizione al registro delle imprese non è obbligatoria".

Il comma 31 della legge 24 Novembre 2006 n.286, che converte il Decreto Legge 3 rOttobre 2006 n.262, modifica l'art. 34 del d.p.r. 633/1972

(Decreto IVA) e stabilisce che i produttori agricoli che nell'anno solare precedente hanno realizzato, o in caso di inizio attività, prevedono di realizzare un volume d'affari non superiore a 7.000 €, sono esonerati dal versamento dell'imposta. Poiché l'art. 2 della legge 25 Marzo 1997 n.77 prevede che i produttori agricoli in regime di esonero non sono obbligati all'iscrizione al registro delle imprese, ne deriva che non sono obbligati all'iscrizione al registro delle imprese gli imprenditori agricoli che hanno realizzato, o prevedono di realizzare, un volume d'affari non superiore a 7.000 €. Questa modifica non va ad intaccare la qualità dei dati utilizzati poiché è avvenuta dopo il 2004. A fine 1999, il Veneto contava 114.901 sedi in totale.

Le imprese iscritte alla Camera di Commercio sono obbligate ad iscriversi anche agli archivi REA; questi si prestano bene alle analisi di sopravvivenza poiché le denunce da effettuare al REA devono essere presentante entro trenta giorni dalla manifestazione dell'evento denunciato. La cancellazione dall'archivio dovrebbe coincidere con la cessazione dell'impresa, dato che l'iscrizione comporta una spesa annua non trascurabile.

2.2 Il Censimento Generale dell'Agricoltura 2000

E' utile fare un'osservazione: il Censimento ha rilevato le aziende agricole, un'entità che non coincide necessariamente con le sedi di impresa dell'archivio REA. Nell'archivio REA vengono distinte le unità locali, definite come "l'impianto funzionalmente autonomo e fisicamente distinto dala sede dell'impresa dove si esercitano attività relative o connesse a quella esercitata dall'impresa stessa." Per l'ISTAT invece l'unità di rilevazione è definita come "l'unità tecnico-economica costituita da terreni anche non contigui, ed eventualmente da impianti ed attrezzature varie in cui si attua la produzione agraria, forestale e zootecnica ad opera di un conduttore, e cioè persona fisica, società o ente, che ne sopporta il rischio sia da solo, sia in forma associata" [Istat, 2000a]. Il Censimento ha permesso di raccogliere informazioni su vari aspetti dell'azienda agricola, sia economici che sociali. Le singole sezioni del questionario riguardavano notizie:

- di carattere generale sull'azienda (sistema di conduzione, forma giuridica, svolgimento di attività di vendita dei prodotti, ecc.);

- sull'utilizzazione dei terreni (nell'annata agraria 1 novembre 1999-31 ottobre 2000) per le coltivazioni principali e la secondaria successiva (seminativi, coltivazioni legnose agrarie, ecc.), 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, ecc.);

- sull'utilizzazione di mezzi meccanici e sulle sue 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, ecc.;

- sulle modalità di acquisto dei mezzi tecnici, sullo svolgimento di attività connesse all'agricoltura, sulla commercializzazione dei prodotti, sull'utilizzo di attrezzature informatiche.

2.3 Abbinamento dei dati del Registro Imprese e del Censimento

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, le informazioni provenienti dal Registro REA con quelle di fonte ISTAT. L'unità di osservazione finale è l'azienda agricola così come definita dal Censimento.

La tabella 2.1 presenta sinteticamente i risultati dell'abbinamento dei dati. Degli originali 114.901 record contenenti i dati sulle sedi di impresa

registrate presso gli archivi REA, 19 sono stati eliminati perché privi della chiave identificativa (Codice Fiscale o Partita IVA), 28 sono stati eliminati perché contenevano una chiave identificativa ripetuta. Dei rimanenti 114.854 record, 88.891 sono stati abbinati ai dati censuari tramite Codice Fiscale, 2.343 tramite Partita IVA. Il tasso di abbinamento, a partire dai record utilizzabili contenuti nei registri REA, è del 79%.

Tabella 2.1: Risultato dell'abbinamento: archivio REA Veneto e V Censimento Agricoltura.

	Sedi dati CCIAA	Unità locali dati Censimento	Sedi + unità locali dati Censimento
Accoppiamento tramite Codice Fiscale	88.561	330	88.891
Accoppiamento tramite Partita IVA	2.340	3	2.343
Accoppiamento totale	90.901	333	91.234
Numero sedi (al netto di duplicazione e dati mancanti)	114.854		
Mancato accoppiamento con dati CCIAA	23.958 (21%)		

2.4 Le variabili del dataset

La tabella 2.2 descrive sinteticamente le variabili considerate quali possibili regressori nei modelli di sopravvivenza.[1]

Tabella 2.2: Regressori per il modello di sopravvivenza.

Nome 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_meno10m	Valore dei prodotti venduti	Binaria
val_prod_vend_tra_10_50m	Valore dei prodotti venduti	Binaria
val_prod_piu50m	Valore dei prodotti venduti	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 di produttori	Binaria
parchi	Rientra in parchi o aree protette	Binaria
perc_altra	% sulla superficie totale dell'azienda	
perc_arbo	% sulla superficie totale dell'azienda	
perc_barb	% superficie sulla sau	
perc_boschi	% sulla superficie totale dell'azienda	
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	

[1] Come si può rilevare dal contenuto della tabella 2.2, non sono state prese in considerazione tutte le variabili rilevate dal Censimento, alcune sono state scartate perchè ritenute non rilevanti per lo studio della sopravvivenza, altre perché affette da elevate percentuali di dati mancanti.

Tabella 2.2: Regressori per il modello di sopravvivenza.

Nome variabile	Descrizione	Modalità
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	% sulla superficie totale dell'azienda	
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	Conduzione professionale del conduttore	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 sì 0 altrimenti
capo_azienda	Se il conduttore è anche capo-azienda	binaria
lic_scu_elem	Titolo di studio del capo-azienda	binaria
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

Tabella 2.2: Regressori per il modello di sopravvivenza.

Nome variabile	Descrizione	Modalità
nessuno	Categoria di riferimento	
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, orticoltura, 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 all'agricoltura e alla zootecnia	1 se 1.4 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 in ortofloricoltura	1 Se cod. Ote I cifra=2 0 altrimenti
Ote3	Aziende specializzate nelle coltivazioni permanenti	Se cod. Ote I cifra=3 e codice non 311 o 312, 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=311, 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

Tabella 2.2: Regressori per il modello di sopravvivenza.

Nome variabile	Descrizione	Modalità
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
Ote5	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
ude4_6	Codice ude	1 se 2$\leq$ ude $\geq$4
ude6_8	Codice ude	1 se 4$\leq$ ude $\geq$6
ude8_12	Codice ude	1 se 6$\leq$ ude $\geq$8
ude12_16	Codice ude	1 se 8$\leq$ ude $\geq$12
ude16_40	Codice ude	1 se 12$\leq$ ude $\geq$16
ude40_100	Codice ude	1 se 40$\leq$ ude $\geq$100
ude100	Codice ude	1 se ude$\leq$100

2.5 Qualità dei dati

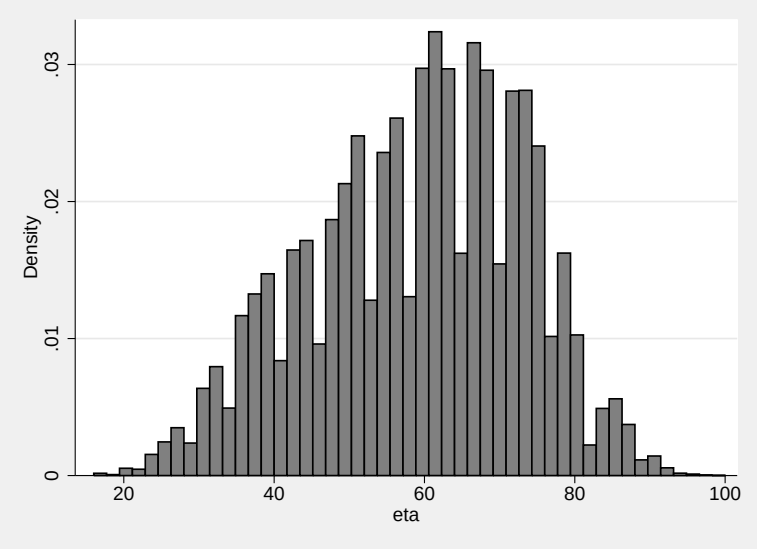
Nel dataset sono presenti dati incongruenti o mancanti in alcune variabili.

- sono presenti 16 aziende con anno di inizio attività superiore al 2000, questi record sono stati eliminati ;
- la variabile `totale_superficie_irrigata` presenta 41776 valori mancanti, per cui non potrà essere utilizzata nell'analisi;
- la variabile `tot_gg_altra_manodopera` presenta 86332 valori mancanti, per cui non potrà essere utilizzata nell'analisi;
- la variabile `serre` presenta 5 valori mancanti;
- 542 record presentano dati mancanti in corrispondenza delle variabili: `sesso`, `lav_fam`, `eta`.

2.5.1 Distribuzioni

In questa sezione vengono riportate le distribuzioni di alcune variabili continue presenti nel dataset. Si è scelto di utilizzare in alcuni casi il diagramma *box_plot* per una più semplice interpretazione dei valori che si discostano maggiormente dalla media. Per fare entrare la variabile età

Figura 2.1: Distribuzione dell'età del conduttore dell'azienda.

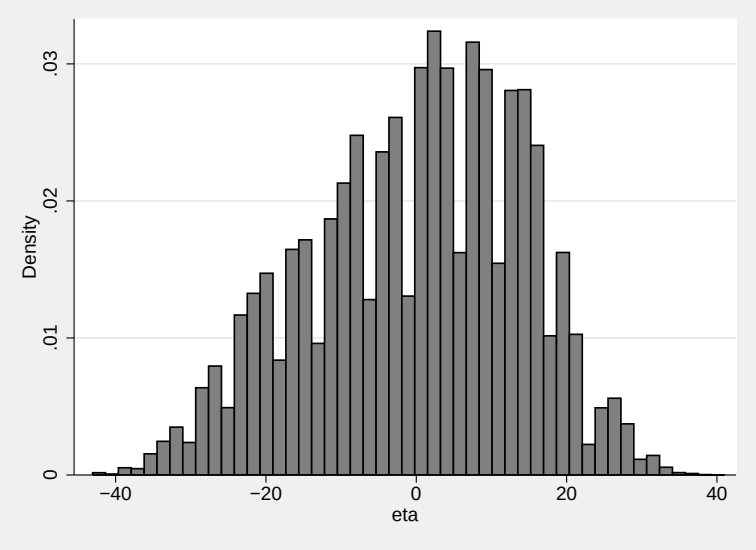


linearmente nel modello, si è provveduto a standardizzarla attraverso la trasformazione :

$$eta\_st = eta - \frac{\sum_{i=1}^{n} x_i}{n}$$

L'età media dei conduttori di aziende agricole è pari a 59 anni. La distribuzione della variabile età, dopo l'operazione di standardizzazione, è riportata nella figura 2.2.

Figura 2.2: Distribuzione standardizzata dell'età del conduttore dell'azienda.



La maggior parte delle aziende agricole venete, sopravvissute fino all'anno 2000, è nata negli anni '70. Solo lo 0.4% delle aziende ha iniziato l'attività prima del 1970, come si può notare nella figura 2.4. Circa il 18% delle aziende ha la stessa data di inizio attività, 1/1/1973: questo è dovuto dall'entrata in vigore, in quella data, del d.p.r 633/1972 che disciplina l'Imposta sul Valore Aggiunto (IVA). Dall'istogramma si evidenzia inoltre un'ulteriore concentrazione di imprese nate nel 1982, ma in questo caso non risultano interventi del legislatore; si presume quindi una modifica nelle gestione degli archivi REA.

Figura 2.3: Distribuzione dell'anno di inizio attività dell'azienda.

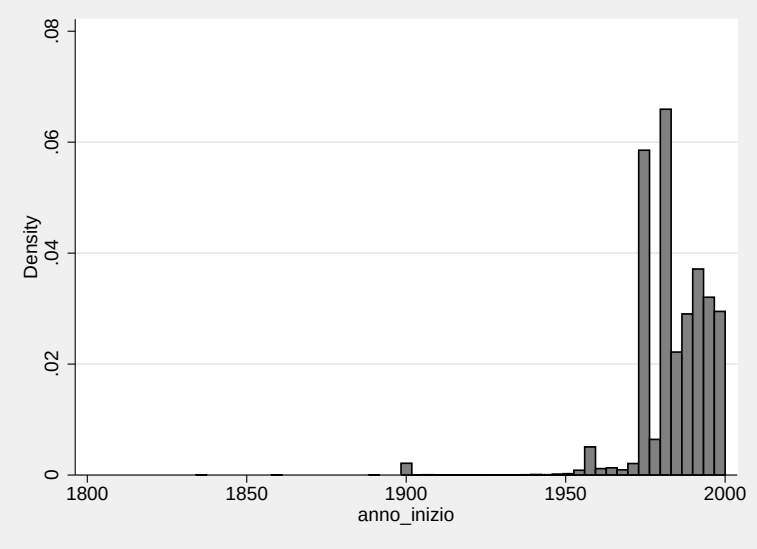
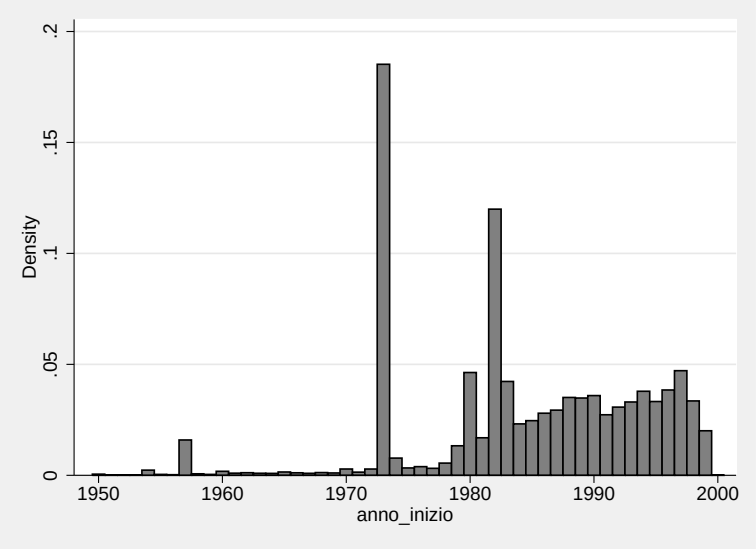


Figura 2.4: Dettaglio per anno di inizio attività dopo il 1950



La variabile che descrive l'estensione della superficie agricola utilizzata (SAU) presenta una forte variabilità. Il 98,5 % delle aziende nel dataset ha una dimensione inferiore ai 5 ettari. E l'azienda con estensione maggiore ha una dimensione di 123 ettari. L'unità di misura della variabile SAU è l'*ara*, la conversione avviene ponendo 100 are=1 ettaro. Le distribuzioni sono riportate nelle figure 2.5, 2.6

Figura 2.5: Distribuzione della SAU per aziende con estensione inferiore a 3 ettari.

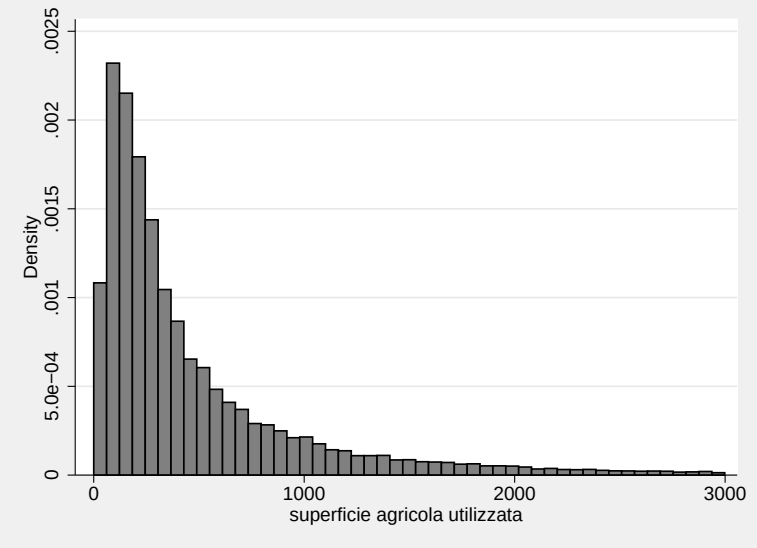
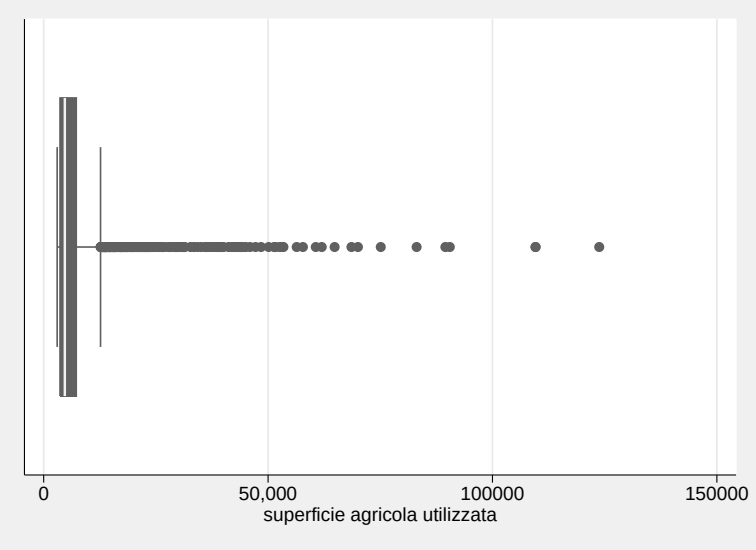


Figura 2.6: Distribuzione della SAU per aziende con estensione superiore ai 3 ettari.



Nelle figure 2.7-2.23 sono riportati i diagrammi *box_plot* relativi alle variabili che rappresentano l'utilizzo della SAU in percentuale. Le variabili: `perc_sanu, perc_boschi, perc_altra, perc_arbo` rappresentano invece l'utilizzo della Superficie Totale Aziendale in percentuale. Questa distinzione si rende necessaria poiché nel questionario del V Censimento dell'Agricoltura, la superficie destinata a boschi e arboricoltura da legno, la superficie non utilizzata (sanu), e quella destinata ad altri usi[2], non vengono sommate per calcolare la SAU. Per facilitare l'interpretazione dei diagrammi sono stati considerati solo i valori delle percentuali maggiori di 0. La tabella 2.3 presenta sinteticamente, per ogni variabile, quanti sono i valori considerati per rappresentare i diagrammi.

Tabella 2.3: Valori considerati per rappresentare i diagrammi.

variabile	numerosità considerata
cereali	63366
fiori	1015
foraggi	11358
frutta	10984
legno	246
legumi	485
olivo	3186
orti	32627
ortive	7983
patata	1634
piante	12409
prati	18695
barbabietola	6814
vite	44342
vivai	983
altra	88124
boschi	40496
sanu	24370
arbo	20963

[2]Per altri usi si intendono: aree occupate da fabbricati, cortili, strade poderali, superficie a funghi ecc.

Figura 2.7: Diagramma della percentuale di SAU coltivata a cereali.

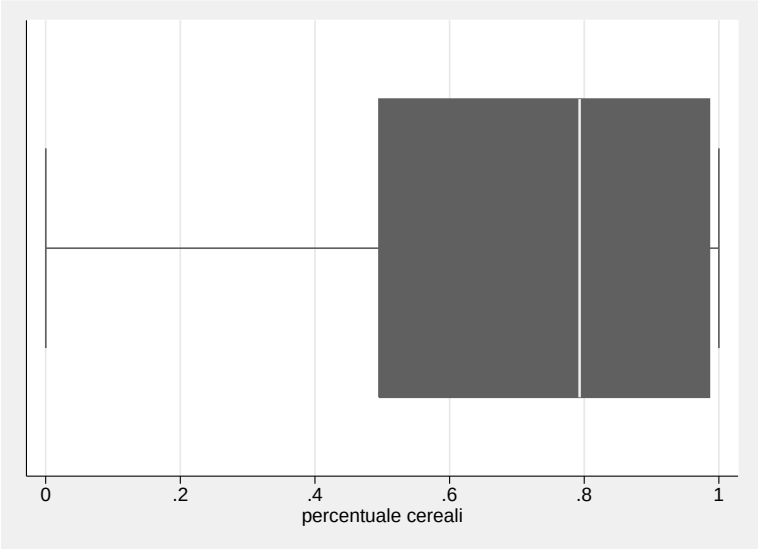


Figura 2.8: Diagramma della percentuale di SAU coltivata a fiori.

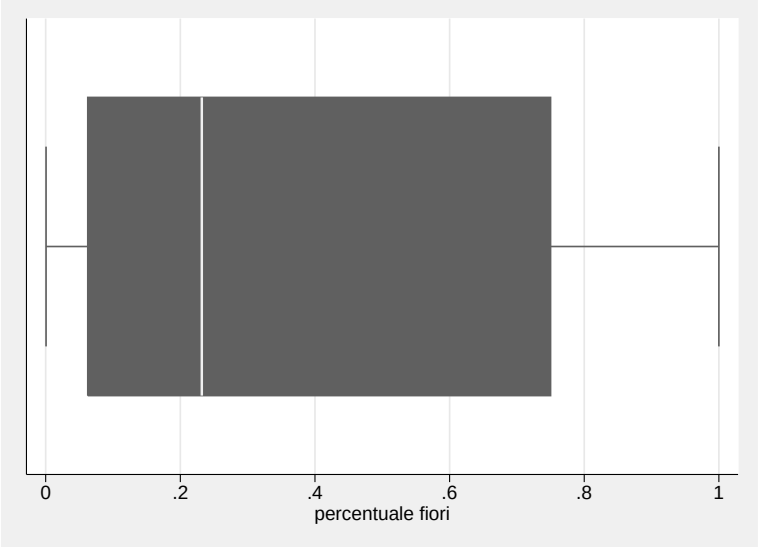


Figura 2.9: Diagramma della percentuale di SAU coltivata a foraggi.

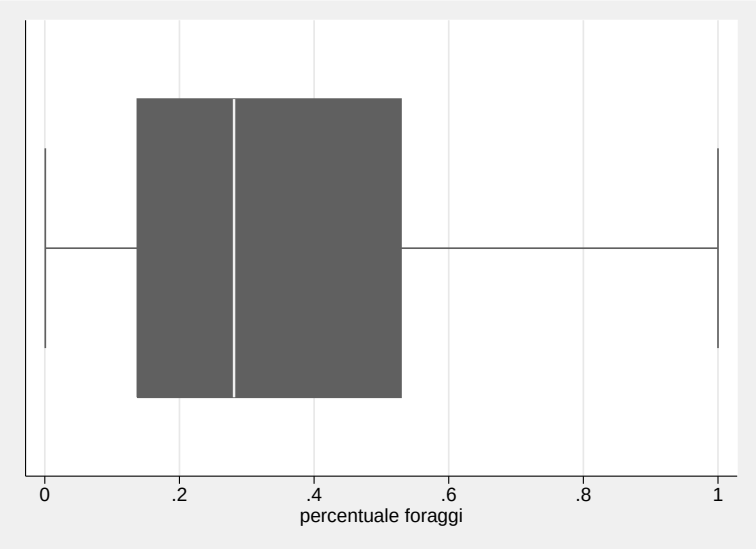


Figura 2.10: Diagramma della percentuale di SAU destinata a frutteti.

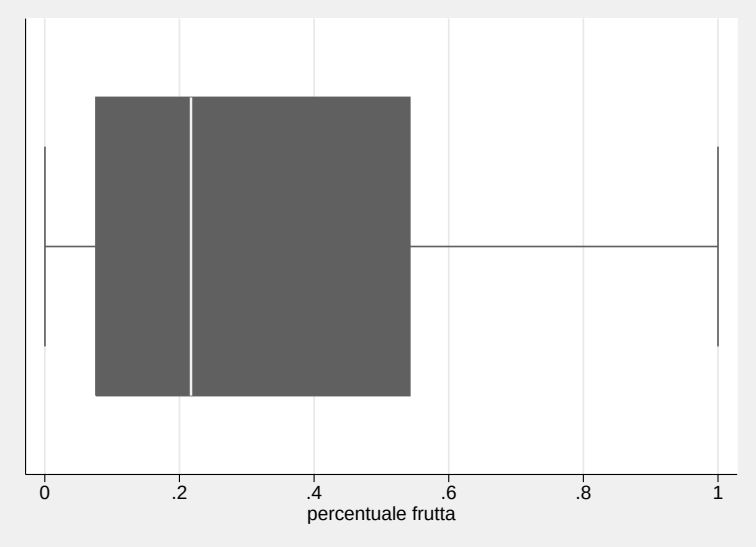


Figura 2.11: Diagramma della percentuale di SAU utilizzata per coltivazioni legnose.

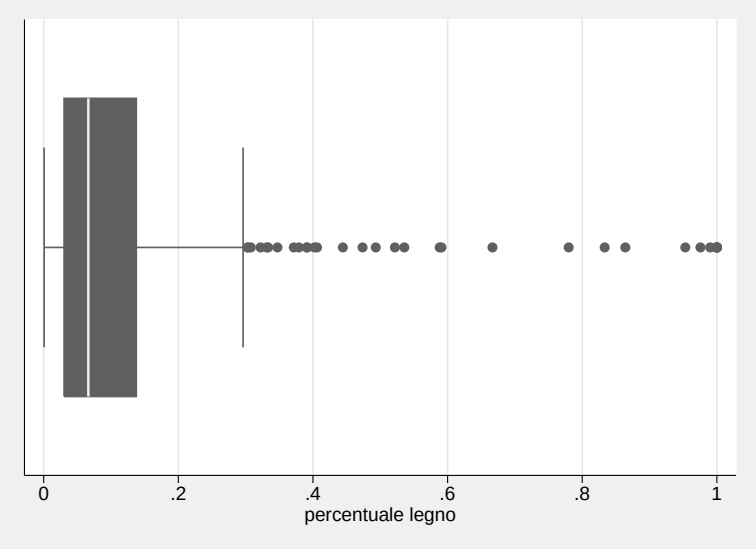


Figura 2.12: Diagramma della percentuale di SAU coltivata a legumi.

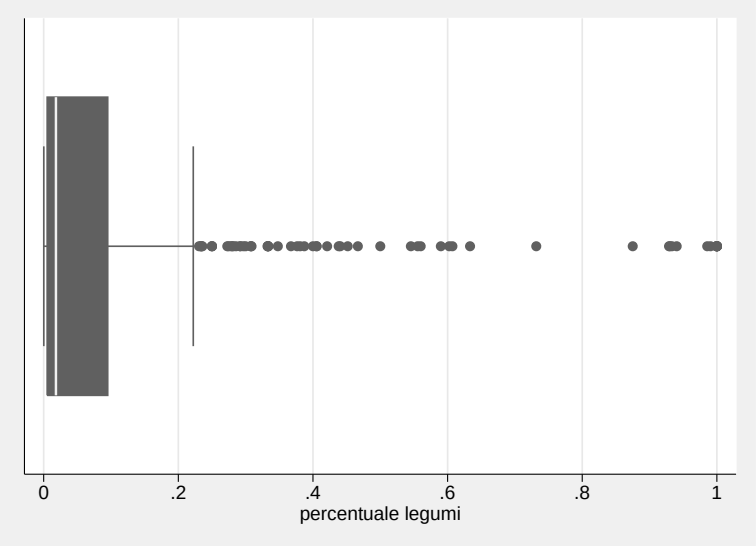


Figura 2.13: Diagramma della percentuale di SAU coltivata a olivi.

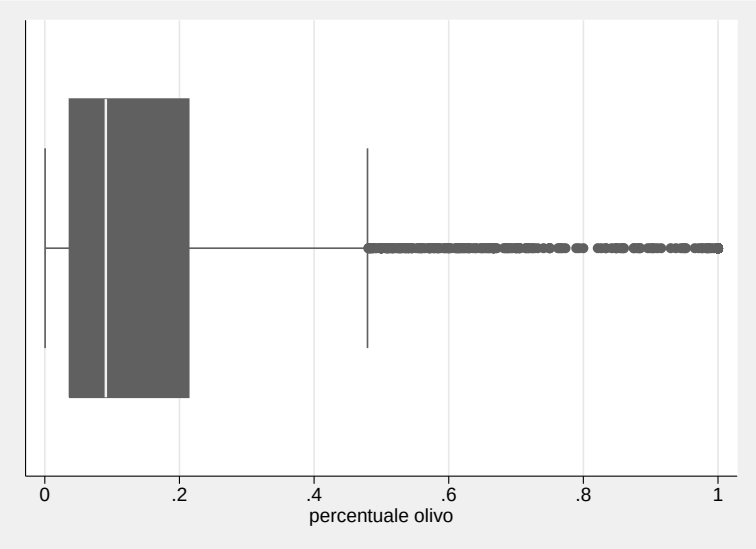


Figura 2.14: Diagramma della percentuale di SAU destinata agli orti.

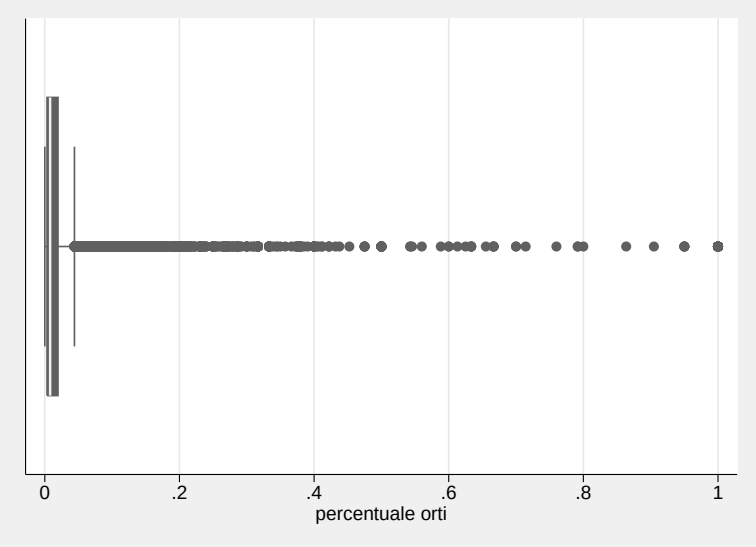


Figura 2.15: Diagramma della percentuale di SAU destinata a coltivazioni ortive.

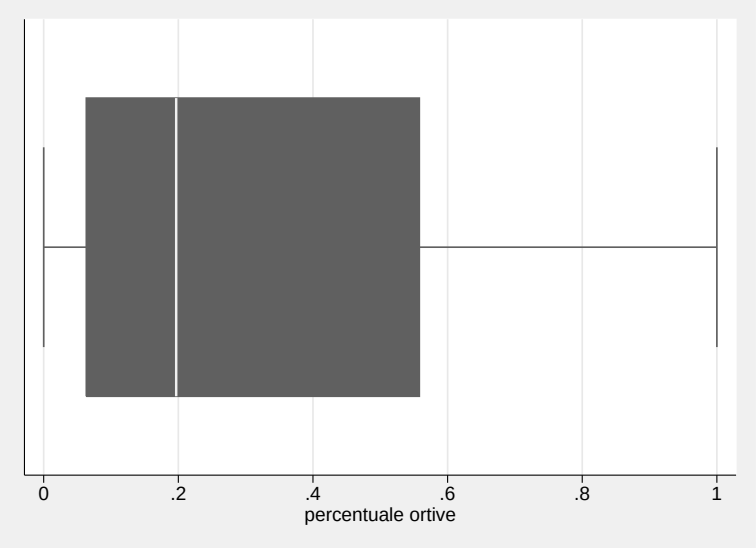


Figura 2.16: Diagramma della percentuale di SAU coltivata a patate.

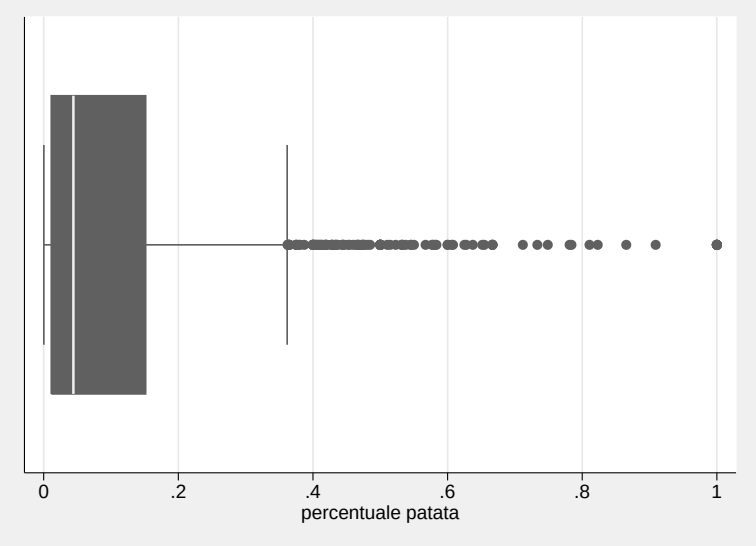


Figura 2.17: Diagramma della percentuale di SAU destinata alla coltivazione di piante.

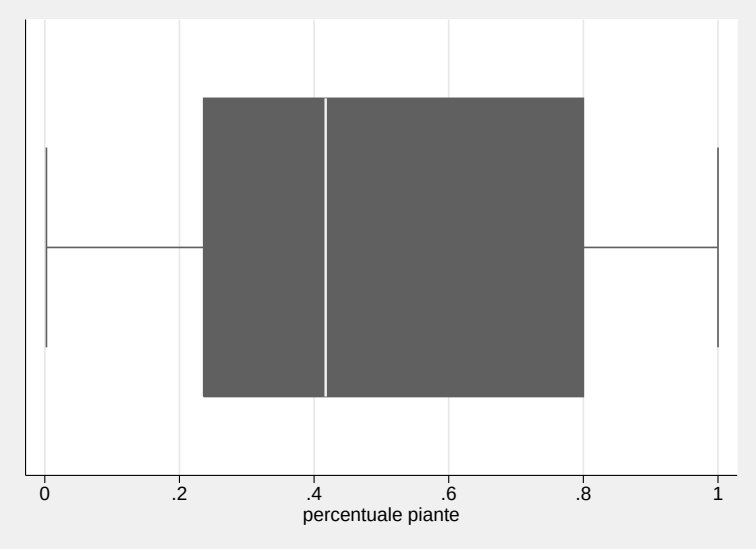


Figura 2.18: Diagramma della percentuale di SAU destinata a prati e pascoli.

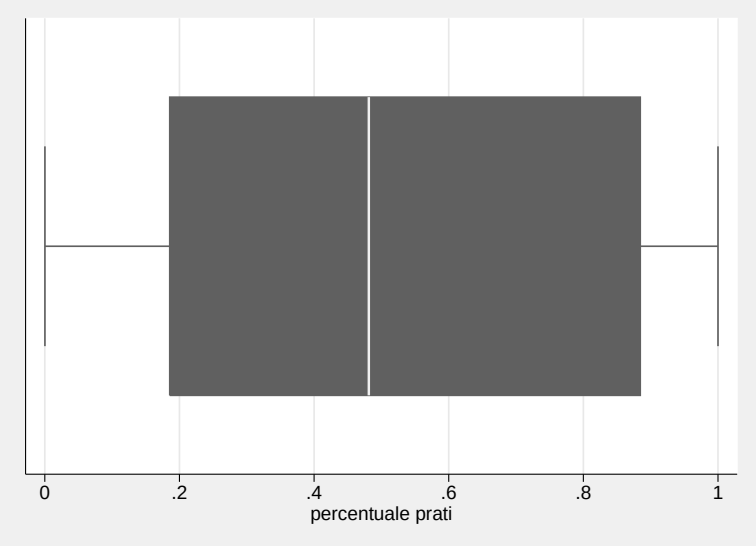


Figura 2.19: Diagramma della percentuale di SAU coltivata a barbabietola.

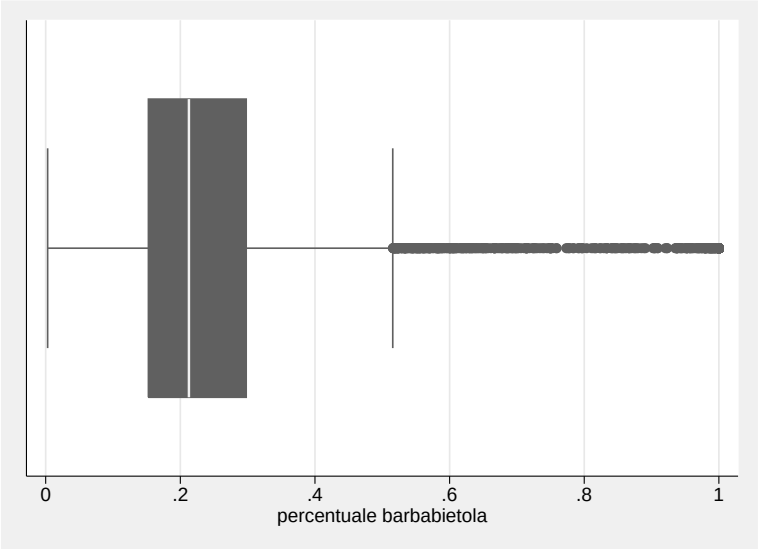


Figura 2.20: Diagramma della percentuale di SAU destinata a vigneti.

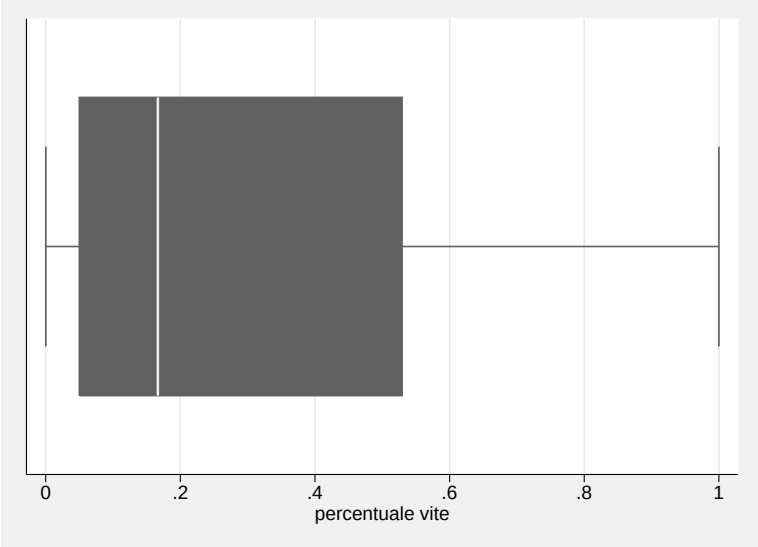


Figura 2.21: Diagramma della percentuale di SAU destinata a vivai.

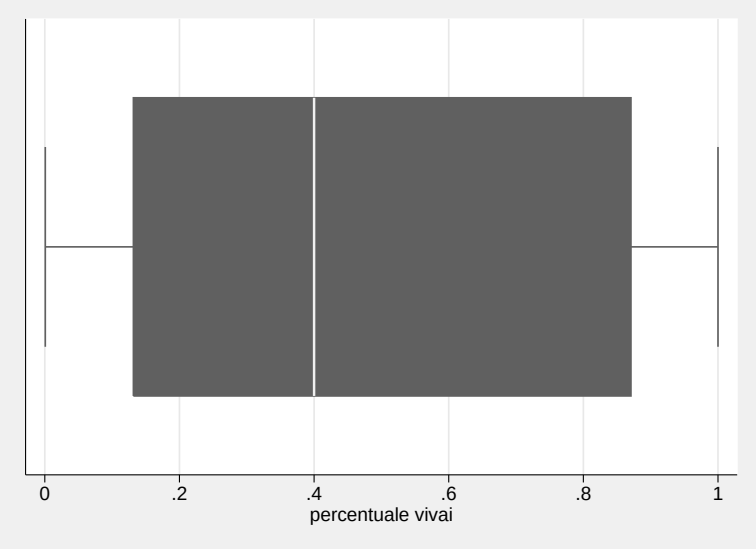


Figura 2.22: Diagramma della percentuale di superficie totale destinata ad altri usi.

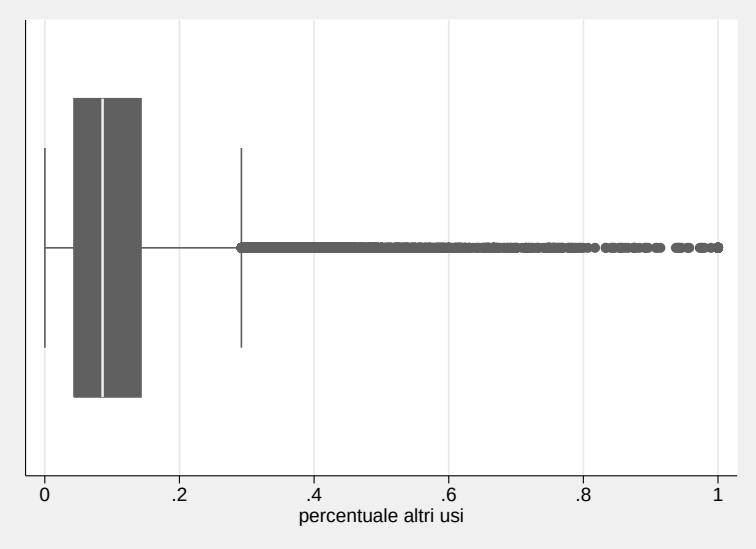
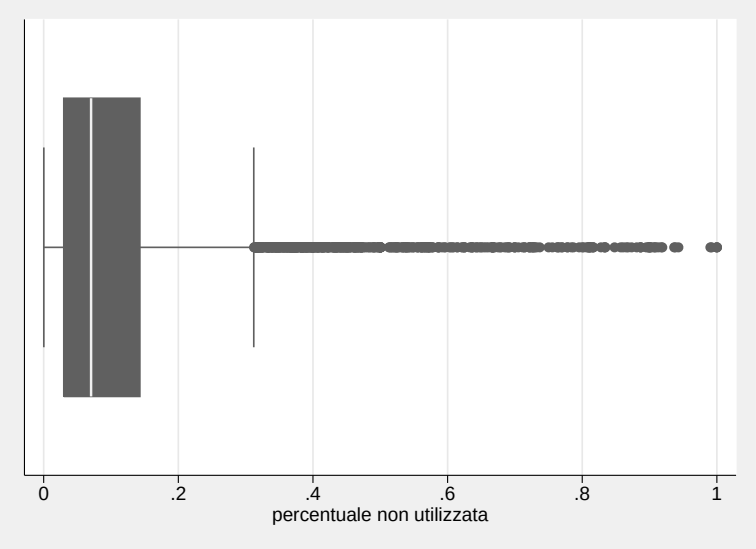


Figura 2.23: Diagramma della percentuale di superficie totale non utilizzata.



2.5.2 Tabelle di frequenza

Di seguito vengono riportate le tabelle di frequenza di alcune variabili utilizzate nell'analisi.

Tabella 2.4: Provincia dove ha sede l'azienda agricola.

PV	Freq.	Percent	Cum.
BL	1801	1.97	1.97
PD	21325	23.37	25.34
RO	6900	7.56	32.90
TV	20113	22.04	54.94
VE	13069	14.32	69.26
VI	11661	12.78	82.04
VR	16391	17.96	100.00
Total	91260	100.00	

Tabella 2.5: Anno di cessazione dell'attività.

anno cessazione	Freq.	Percent.	Cum.
2000	5976	21.85	21.85
2001	6994	25.57	47.42
2002	6157	22.51	69.92
2003	4691	17.15	87.07
2004	3536	12.93	100.00
Totale	27354	100.00	

Dalla tabella 2.5 si evidenzia il numero di aziende cessate nei 5 anni considerati.
La percentuale complessiva di aziende cessate nel periodo corrisponde al 29,97%.
La tabella 2.6 presenta i valori, divisi in classi, della variabile UDE. L' *Unità di Dimensione Economica Europea* (UDE) corrisponde ad un *Reddito Lordo Standard* (RLS) aziendale riferito a "condizioni di produzione ed a prezzi di

un prefissato periodo". Per reddito lordo si intende la differenza tra il valore della produzione lorda, proveniente dall'unità di superficie (ettaro) investita nelle singole coltivazioni e/o dal singolo capo di bestiame allevato, ed i costi specifici sostenuti per ottenerla. Poiché il calcolo dei redditi lordi non può essere effettuato a livello di singola azienda, non disponendo per ciascuna di esse dei dati contabili, per la classificazione tipologica, si è reso necessario ricorrere ai Redditi Lordi Standard, corrispondenti ad una situazione media per ogni singolo prodotto considerato nell'ambito di un dato livello territoriale. I Redditi Lordi Standard esprimono, pertanto, un valore medio applicabile a tutte le aziende ricadenti in un determinato territorio che, per l'Italia, è stato identificato con la regione [Istat, 2000b]. Nel V Censimento dell'Agricoltura ogni UDE è posto uguale ad un RLS aziendale di 1.200 ECU.

Tabella 2.6: Valori di UDE.

classe UDE	Freq.	Percent
0-2	19283	21.12
2-4	19454	21.32
4-6	10483	11.49
6-8	6713	7.36
8-12	8557	9.38
12-16	5444	5.97
16-40	13338	14.62
40-100	6063	6.64
>100	1925	2.11
Totale	91260	100.00

La tabella 2.7 si riferisce alla variabile OTE. La classificazione delle aziende agricole secondo l'Orientamento Tecnico Economico (OTE) è definita da appositi regolamenti predisposti dalla Comunità Europea [Istat, 2000b]. Ciascuna azienda è classificata in uno degli OTE in base all'incidenza percentuale del RLS delle varie attività produttive aziendali sul RLS complessivo dell'azienda.

Tabella 2.7: Valori di OTE.

Classi di OTE	Freq.	Percent
ote1	45874	50.28
ote2	1935	2.12
ote3	7446	8.16
ote311	4858	5.32
ote312	5674	6.22
ote4	9026	9.89
ote5	999	1.09
ote6	10154	11.13
ote7	1165	1.28
ote8	3878	4.25
ote9	235	0.26
Totale	91244	100.00

Tabella 2.8: Sesso del conduttore dell'azienda.

sesso	Freq.	Percent	Cum.
M	71483	78.81	78.81
F	19219	21.19	100.00
Totale	90,702	100.00	

La tabella 2.10 evidenzia una discreta propensione delle aziende a varie forme di associazionismo.
Alcune aziende sono associate in più di una forma: 330 sono associate sia a consorzi che associazioni di produttori e società cooperative.

Nella tabella 2.13 si può notare la predominanza della società individuale, rispetto ad altre forme di società.

Tabella 2.9: Titolo di studio del conduttore dell'azienda.

Titolo di studio	Freq.	Percent
nessuno	3696	4.05
licenza elementare	52056	57.05
licenza media	21280	23.32
diploma	12513	13.71
laurea	1699	1.86
Totale	91244	100.00

Tabella 2.10: Associazioni tra imprese.

Tipo di associazione	Freq.	Percent
nessuna	48336	52.8
assoc. di produttori	10611	11.63
soc. cooperative	26469	29.01
consorzi	5828	6.39
Totale	91244	100.00

Tabella 2.11: Classificazione delle aziende agricole in base all'attività.

Codici ATECO	Freq.	Percent
nessuna	13686	15.00
cod_att101	5479	6.00
cod_att102	6666	7.31
cod_att103	11065	12.13
cod_att104	992	1.09
cod_att111	49308	54.04
cod_att112	4048	4.44
Totale	91244	100.00

Tabella 2.12: Giornate di lavoro del conduttore in un anno.

numero di giornate	Freq.	Percent
tra 0 e 30	28532	31.27
tra 30 e 90	18923	20.74
tra 90 e 270	24659	27.03
più di 270	19130	20.96
Totale	91244	100

Tabella 2.13: Forme di società.

Tipo di società	Freq.	Percent
società individuale	86923	95.26
altre forme	4321	4.74
Totale	91244	100.00

Tabella 2.14: Valore della produzione venduta.

Valore della produzione	Freq.	Percent
meno di 10 milioni di lire	39152	42.91
tra 10 e 50 milioni di lire	40228	44.09
più di 50 milioni di lire	11864	13.00
Totale	91244	100.00

Capitolo 3

Un primo approccio: il modello Logit

3.1 L'analisi della regressione logistica

L'analisi di regressione logistica è un metodo per la stima della funzione di regressione che meglio collega la probabilità del possesso di un attributo dicotomico con un insieme di variabili esplicative. In questo caso l'attributo dicotomico è rappresentato dalla cessazione, o meno, dell'azienda agricola nel periodo 1999-2004: l'analisi di regressione logistica consente di individuare le determinanti della *probabilità*, o *rischio*, della cessazione dell'azienda. Quello di regressione logistica è dunque un caso speciale dell'analisi di regressione, che trova applicazione quando la variabili dipendente è dicotomica, mentre l'analisi di regressione lineare si applica se la variabile dipendente è continua.

Oltre che per la scala di misura della variabile dipendente, l'analisi della regressione logistica si distingue da quella lineare perché per questa si ipotizza una distribuzione normale di Y, mentre se Y è dicotomica la sua distribuzione è, ovviamente, binomiale.

Analogamente, nell'analisi della regressione lineare anche la stima di Y ottenuta dalla regressione varia da $-\infty$ a $+\infty$, mentre nell'analisi della regressione logistica la stima di Y varia tra 0 e 1. La stima di Y assume allora il significato di probabilità che Y si uguale a 1 : $P(Y = 1|x) = \pi(x)$.

La funzione di regressione logistica si presenta come segue:

$$logit(\pi(x)) = \beta_0 + \sum_{i}^{q} \beta_i x_i = \boldsymbol{X\beta} \tag{3.1}$$

$$logit(\pi(x)) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \tag{3.2}$$

La scelta del *logit* per descrivere la funzione che lega la probabilità di Y alla combinazione delle variabili predittive è determinata dalla constatazione che la probabilità si avvicina ai limiti zero e uno gradualmente e descrive una figura detta "sigmoide" che assomiglia alla cumulata della distribuzione casuale degli errori detta "funzione logistica". [Fabbris, 1997]
La probabilità di Y si può, infatti, scrivere come funzione logistica :

$$\pi(x) = \frac{e^{\boldsymbol{X\beta}}}{1 + e^{\boldsymbol{X\beta}}} \tag{3.3}$$

Pur non essendo il *logit* l'unica funzione che consente di modellare la probabilità di un fenomeno, essa è privilegiata, dato che è una trasformata del rapporto tra due probabilità complementari, ovvero tra il numero di successi per ogni insuccesso del fenomeno in esame, come si può verificare nell'equazione (3.2). In inglese questa quantità è detta *odds*.

3.1.1 Stima e verifica di ipotesi

L'analisi della regressione logistica è un'estensione della regressione lineare. L'una e l'altra fanno parte di una classe di modelli, detti *lineari generalizzati*, la cui trattazione è indipendente dalla natura di Y e dal tipo di funzione che lega le variabili esplicative alla dipendente.

Nei Modelli Lineari Generalizzati la variabile dipendente e le variabili esplicative sono legate da una funzione g, monotòna e differenziabile, detta *funzione legame*:

$$E(Y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} \tag{3.4}$$

La distribuzione dell'errore delle stime ottenibili con l'applicazione di questi modelli dipende dalla natura di Y:

- se la funzione legame è l'identità ($g = 1$) e la distribuzione dell'errore è normale, ci si riconduce al modello di regressione lineare;
- se la funzione legame è $g = logit$ e la distribuzione dell'errore è binomiale, si ha il modello di regressione logistica.

Stima

La stima dei parametri ignoti β è effettuata con il metodo della massima verosimiglianza, che si basa sulla massimizzazione della probabilità di osservare l'insieme di dati osservato, in funzione di β.
Date n osservazioni indipendenti, il modello relativo alla generica unità i $(i = 1, \ldots, n)$ è :

$$\begin{aligned} y_i &= E(Y_i|\boldsymbol{x}_i) + \epsilon_i \\ &= \frac{exp\,(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + exp\,(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})} + \epsilon_i \\ &= \pi(\boldsymbol{x}_i) + \epsilon_i \end{aligned} \tag{3.5}$$

Poiché Y è dicotomica, la sua distribuzione è binomiale, con media $E(Y_i|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$, assumendo $E(\epsilon|X) = 0$, e funzione di probabilità per l'i-esima unità:

$$f(y_i|\boldsymbol{x}_i; \beta) = \pi(\boldsymbol{x}_i)^{y_i}[1 - \pi(\boldsymbol{x}_i)]^{1-y_i} \tag{3.6}$$

Data l'indipendenza delle osservazioni, la verosimiglianza del campione di n unità è il prodotto delle verosimiglianze delle unità che lo compongono:

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i; \beta) \tag{3.7}$$

per derivare la "stima di massima verosimiglianza" dei parametri, si determina il vettore β che massimizza il logaritmo di $L(\beta)$:

$$\begin{aligned} l(\beta) &= \log[L(\beta)] \\ &= \sum_{i=1}^{n} [y_i \log \pi(\boldsymbol{x}_i) + (1 - y_i) \log(1 - \pi(\boldsymbol{x}_i))] \\ &= \sum_{i=1}^{n} y_i \log \frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)} + \sum_{i=1}^{n} \log(1 - \pi(\boldsymbol{x}_i) \end{aligned} \tag{3.8}$$

Ponendo uguali a 0 le derivate parziali fatte rispetto ai $p+1$ parametri $(\beta_0, \beta_1, \ldots, \beta_p)$ da stimare si ottengono le cosiddette *equazioni di verosimiglianza*. Tali equazioni, in quanto non lineari nei parametri, richiedono l'applicazione di metodi iterativi di stima.

3.1.2 Interpretazione dei parametri

Sia X una variabile esplicativa nominale a più modalità (ad esempio una variabile risposta, anch'essa dicotomica). Per esempio siano Y = azienda cessata [0: non cessata; 1: cessata] e X = adesione_consorzi [0: azienda che non aderisce a consorzi; 1: azienda che aderisce a consorzi]. In generale, si ha che il coefficiente di regressione di una variabile misura la variazione nel *logit* di Y corrispondente al possesso dell'attributo X:

$$logit\left\{Pr\left(Y=1\,|X=1\right)\right\} - logit\left\{Pr\left(Y=1\,|X=0\right\}\right) = \quad (3.9)$$

$$= (\beta_0 + \beta_1 1) - (\beta_0 + \beta_1 0) = \beta_1 \quad (3.10)$$

Poichè X assume valori 0 o 1, il coefficiente rappresenta il logaritmo dell'*odds ratio*:

$$\beta_1 = \log \frac{Pr(Y=1|X=1) \cdot Pr(Y=0|X=0)}{Pr(Y=1|X=0) \cdot Pr(Y=0|X=1)} \quad (3.11)$$

Si consideri una variabile esplicativa nominale a cinque modalità, come il titolo di studio del conduttore dell'azienda, per la quale sono state costruite quattro variabili *dummy* D_1, D_2 D_3 , D_4 che significano rispettivamente licenza elementare ($D_1 = 1$), licenza media ($D_2 = 1$), diploma ($D_3 = 1$), laurea ($D_4 = 1$). I parametri $\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}$ rappresentano il logaritmo dell'*odds ratio* licenza elementare vs. nessun titolo di studio, licenza media vs. nessun titolo di studio e via di seguito per le altre modalità:

$$\beta_{11} = \log \frac{Pr(Y=1|D_1=1) \cdot Pr(Y=0|X=\text{ nessun titolo})}{Pr(Y=1|X=\text{ nessun titolo}) \cdot Pr(Y=0|D_1=1)} \quad (3.12)$$

Nel caso di variabili quantitative, l'interpretazione del parametro è analoga a quella per variabili dicotomiche, tuttavia è utile concentrare l'attenzione

sulla seguente domanda: "E' sensato considerare un incremento unitario di X?". L'incremento di un'unità per una variabile continua può non essere interessante ai fini della ricerca, mentre sarebbe più plausibile considera un'aumento di c unità. In generale, considerando un aumento di c unità della variabile indipendente, si avrà :

$$c\beta_1 = logit[Pr(Y = 1|X = x + c)] - logit[Pr(Y = 1|X = x)] \qquad (3.13)$$

e l'*odds ratio* che si ottiene, $\psi(c) = \exp(c\beta_1)$, valuta l'aumento di rischio corrispondente ad un aumento di c unità della variabile esplicativa.

3.2 Applicazione del modello LOGIT al dataset

Nel dataset in esame è presente una variabile dicotomica che assume valore 1 se l'azienda è cessata nel periodo 1999-2004 e 0 altrimenti. Quindi la regressione logistica ben si presta a spiegare il comportamento di questa variabile. Dopo aver eliminato i record presentanti dati mancanti rimangono utilizzabili 90.697 unità.
Inizialmente, si stima un modello considerando come covariate le variabili `UDE`,11 classi, e `OTE`, 9 classi: misure sintetiche di specializzazione colturale e redditività standardizzate, proposte a livello di Unione Europea per classificare le aziende agricole. Inoltre, vengono inserite come covariate le variabili *dummy*: `t0_8_16`, `t0_16_23`, `t0_16_23`, `t0_23`. Queste variabili sono utilizzate per controllare l'effetto "dell'anzianità" dell'azienda sulla probabilità di cessazione. I valori: 8,16,23 rappresentano rispettivamente il 25^o,50^o,75^o percentile della variabile `t0` definita come: 2000[1]`-anno_inizio`.

Quindi, a determinati valori assunti dalle variabili dummy, corrispondono degli intervalli temporali in cui è nata un'azienda :
`t0_8_16`=1 $\Longrightarrow$ 1992-1986
`t0_16_23`=1 $\Longrightarrow$ 1985-1977
`t0_23`=1 $\Longrightarrow$ 1975-1834
I risultati delle stime sono riportanti nella tabella 3.1

[1]Si ricorda che le informazioni per il V Censimento Generale dell'Agricoltura sono state raccolte nei mesi di Ottobre e Novembre 2000

Tabella 3.1: Stima del modello logit: `OTE`, `UDE`, `t0` come regressori

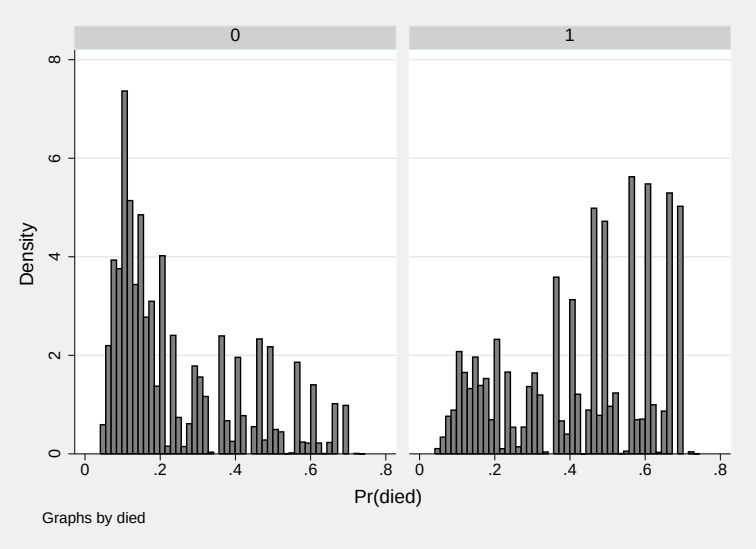
Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
ote1	0.760**	2.138	(0.136)
ote2	1.003**	2.728	(0.150)
ote4	0.329*	1.389	(0.139)
ote5	0.878**	2.405	(0.163)
ote6	0.566**	1.762	(0.138)
ote7	0.11	1.116	(0.161)
ote8	0.356*	1.427	(0.142)
ote311	0.009	1.009	(0.142)
ote312	0.402**	1.495	(0.139)
ote3	0.548**	1.731	(0.139)
ude2_4	-0.827**	0.437	(0.021)
ude4_6	-1.615**	0.199	(0.028)
ude6_8	-2.021**	0.133	(0.036)
ude8_12	-2.143**	0.117	(0.034)
ude12_16	-2.380**	0.093	(0.045)
ude16_40	-2.426**	0.088	(0.032)
ude40_100	-2.463**	0.085	(0.047)
ude100	-2.675**	0.069	(0.085)
t0_8_16	0.181**	1.198	(0.023)
t0_16_23	0.437**	1.548	(0.023)
t0_23	0.554**	1.740	(0.023)
costante	-0.492**		(0.135)

N	90697
Log-likelihood	-46952.787
$\chi^2(21)$	17036.609
$Pr. > \chi^2(21)$	0.0000

Livelli di significatività : † : 10% ∗ : 5% ∗∗ : 1%

Quasi tutti i coefficienti delle variabili risultano significativi, tranne `OTE7` e `OTE 311`. E' interessante notare come, al crescere della dimensione economica UDE, diminuisce monotonicamente e in maniera statisticamente significativa, la probabilità di cessazione, che aumenta invece con l'incremento degli anni di vita dell'impresa. Dopo aver ottenuto i risultati delle regressione, questi si utilizzano per prevedere la probabilità di cessazione delle imprese presenti nel dataset.
Successivamente, si va valutare la capacità classificatoria del modello comparando la distribuzione di probabilità stimata con quella osservata. La figura 3.1 riporta le distribuzioni delle probabilità stimate di cessazione per le aziende sopravvissute e per quelle cessate (0=azienda cessata ,1=sopravvissuta).

Figura 3.1: Probabilità stimate di cessazione.



La capacità classificatoria del modello si può comprendere meglio analizzando i risultati della tabella 3.2 che presenta l'analisi di classificazione per 3 soglie di probabilità di cessazione: 0.2, 0.3 e 0.4; oltre queste soglie l'azienda viene classificata come cessata.

Tabella 3.2: Capacità classificatoria del modello.

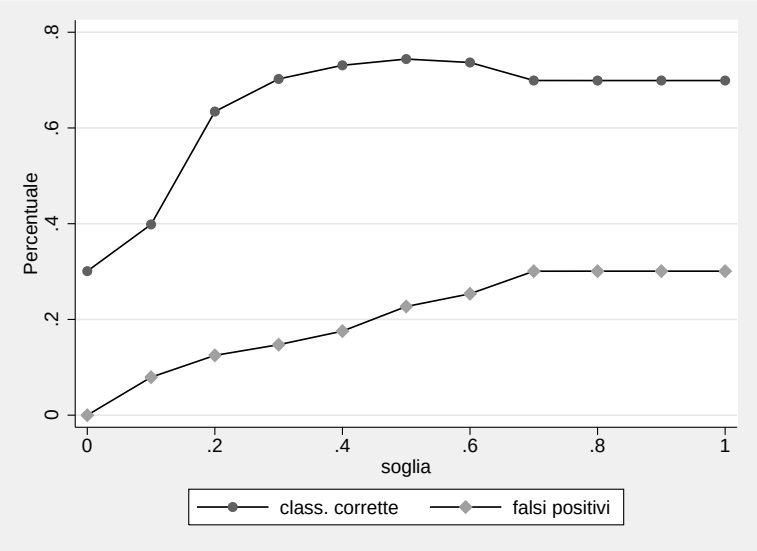
Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	22257	28137	50394
0=ancora attiva (−)	5032	35271	40303
Falsi Positivi	$Pr(D\|-)$		12.43%
Falsi Negativi	$Pr(\neq D\|+)$		55.83%
Classificazioni corrette			63.43%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	19686	19390	39076
0=ancora attiva (−)	7603	44018	51621
Falsi Positivi	$Pr(D\|-)$		14.73%
Falsi Negativi	$Pr(\neq D\|+)$		49.62%
Classificazioni corrette			70.24%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	16737	13858	30595
0=ancora attiva (−)	10522	49550	60102
Falsi Positivi	$Pr(D\|-)$		17.56%
Falsi Negativi	$Pr(\neq D\|+)$		45.29%
Classificazioni corrette			73.09%

Come si può notare dalla tabella 3.2 , la capacità classificatoria è influenzata dalla soglia scelta. La figura 3.2 illustra sinteticamente la variabilità di alcuni risultati presentati nella tabella 3.2 al variare della soglia utilizzata. Si può notare come la percentuale di aziende correttamente classificate, e la percentuale di falsi positivi, si stabilizzino se si assume una soglia superiore a 0.7.

Figura 3.2: Sensibilità della classificazione al variare della soglia.



Ai fini di una politica di finanziamento alle aziende agricole, i risultati ottenuti possono essere utilizzati in diversi modi. Il principale consiste nel finanziare solo le imprese che hanno una probabilità di sopravvivere, per almeno 5 anni, superiore ad una determinata soglia. Se si considerano tutte le aziende ugualmente meritevoli, la probabilità di sovvenzionare un'azienda che chiuderà entro 5 anni è pari al 27,28%; tale probabilità scende a 12,43% se si finanziano solamente le imprese con probabilità stimata di cessazione inferiore a 0.2. Questa decisione però comporta il mancato finaziamento del 55,83% delle aziende che sopravviveranno per più di 5 anni. Aumentando il valore della soglia, aumenta la probabilità di finanziare aziende che chiuderanno entro 5 anni, ma diminuisce la probailità di non finanziare le aziende che continueranno a rimanere attive. Successivamente si prova a stimare un modello LOGIT utilizzando tutti i regressori. I risultati della stima sono riportati nella tabella 3.3 .

Tabella 3.3: Stima modello logit : tutti i regressori.

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
adesione_ass_prod	-0.0809972**	0.922	(0.0296409)
adesione_soc_coop	-0.1766148**	0.838	(0.0217818)
allev_avicoli	-3	1.000	(1.5000000)
azienda_individuale	-0.0935330†	0.911	(0.0554470)
bovini	-0.0005186	0.999	(0.0003349)
cond_prof_condu	0.1214660**	1.129	(0.0274339)
conigli	-0.0000145	1.000	(0.0000208)
equini	-0.0147181	0.985	(0.0107480)
parchi	-0.0502224	0.951	(0.0461882)
lavoraz_prod_agric	-0.0237796	0.977	(0.0273876)
lic_scu_elem	-0.0351143	0.965	(0.0391672)
lic_scu_inf	-0.1330359**	0.875	(0.0447944)
serre	-9.7	1.000	(6.0000000)
sesso	-0.1893475**	0.827	(0.0219648)
suini	0.0000109	1.000	(0.0000602)
sup_sau_azienda	-0.0000744**	1.000	(0.0000161)
sup_sau_tot2	0.0000000**	1.000	(0.0000000)
lav_fam	0.0004568**	1.000	(0.0000753)
cond_dir	-0.1269625**	0.881	(0.0249621)
val_prod_vend_meno10m	-0.1121330*	0.894	(0.0447597)
val_prod_vend_tra_10_50m	0.0030645	1.003	(0.0429827)
prod_bio	0.0447666	1.046	(0.0886311)
nr_abit	-0.0026664	0.997	(0.0124556)
ovicapri	0.000736	1.001	(0.0009535)
laurea	-0.2501025**	0.779	(0.0834637)
diploma	-0.1753180**	0.839	(0.0468813)
bel	-0.1530594*	0.858	(0.0675855)
rov	-0.2341744**	0.791	(0.0405591)
pad	-0.3582567**	0.699	(0.0251800)

Continua nella prossima pagina...

... tabella 3.3

Variabile	Coefficiente	Odds ratio	(Std. Err.)
vic	-0.0403736	0.960	(0.0299735)
ven	0.0247159	1.025	(0.0285884)
ver	-0.2861954**	0.751	(0.0314881)
senza_sup	-0.1469857	0.863	(0.2165117)
adesione_consorzi	-0.0959880**	0.908	(0.0370817)
eta	0.0316707**	1.032	(0.0008994)
eta2	0.0004326**	1.000	(0.0000405)
lav_0_30	0.8984192**	2.456	(0.0473899)
lav_30_90	0.6522317**	1.920	(0.0447936)
lav_90_270	0.3327155**	1.395	(0.0373977)
capo_azienda	-0.1963234**	0.822	(0.0452985)
cod_att101	-0.0204696	0.980	(0.0400831)
cod_att102	0.0649470†	1.067	(0.0371437)
cod_att103	0.0135055	1.014	(0.0318580)
cod_att104	-0.0080128	0.992	(0.0826625)
cod_att111	0.0055037	1.006	(0.0241335)
cod_att112	-0.0168008	0.983	(0.0447253)
affitto	0.1595937**	1.173	(0.0330727)
uso_grat	0.0500331	1.051	(0.0517890)
perc_cer	0.3895914*	1.476	(0.1579311)
perc_fiori	0.6680098**	1.950	(0.2397664)
perc_foraggi	-0.0261081	0.974	(0.1666428)
perc_frutta	0.9020286**	2.465	(0.1776108)
perc_legno	-1.3934486†	0.248	(0.7759555)
perc_legumi	0.9101408†	2.485	(0.5076891)
perc_olivo	0.2577802	1.294	(0.2262434)
perc_orti	0.7712912**	2.163	(0.2446033)
perc_ortive	0.6765320**	1.967	(0.1727163)
perc_patata	0.3217639	1.380	(0.4103431)
perc_piante	0.3806993*	1.463	(0.1632380)
perc_prati	0.3236432*	1.382	(0.1617901)

Continua nella prossima pagina...

... tabella 3.3

Variabile	Coefficiente	Odds ratio	(Std. Err.)
perc_barb	0.0128061	1.013	(0.1875178)
perc_vite	0.2396274	1.271	(0.1664845)
perc_vivai	0.9932080**	2.700	(0.2331223)
perc_altra	0.0358561	1.037	(0.0863438)
perc_boschi	0.0626138	1.065	(0.0755287)
perc_sanu	0.0556802	1.057	(0.1794674)
perc_arbo	-0.0922245	0.912	(0.2070820)
ude2_4	-0.8013507**	0.449	(0.0227921)
ude4_6	-1.5151806**	0.220	(0.0304249)
ude6_8	-1.8581397**	0.156	(0.0393362)
ude8_12	-1.8967427**	0.150	(0.0389501)
ude12_16	-2.0331956**	0.131	(0.0507062)
ude16_40	-1.9420102**	0.143	(0.0461371)
ude40_100	-1.8917138**	0.151	(0.0740587)
ude100	-1.9579277**	0.141	(0.1355729)
ote1	0.2813854	1.325	(0.2048632)
ote2	0.6013509**	1.825	(0.2197160)
ote4	0.237514	1.268	(0.2079137)
ote5	0.6966702**	2.007	(0.2143572)
ote6	0.2285484	1.257	(0.2053153)
ote7	-0.0888779	0.915	(0.2223545)
ote8	0.1452475	1.156	(0.2079931)
ote311	-0.105271	0.900	(0.2139713)
ote312	0.1479217	1.159	(0.2100968)
ote3	0.1723246	1.188	(0.2110368)

Continua nella prossima pagina...

... tabella 3.3

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
t0_8_16	0.0953965**	1.100	(0.0246480)
t0_16_23	0.1005777**	1.106	(0.0253193)
t0_23	0.1195132**	1.127	(0.0259126)
costante	-0.2292685		(0.2444636)

N	90697
Log-likelihood	-44643.1162
$\chi^2(88)$	21656
$Pr. > \chi^2(88)$	0.000

Livelli di significatività : † : 10% ∗ : 5% ∗∗ : 1%

Le variabili `UDE` continuano a rimanere significative, anche se la probabilità di cessazione non diminuisce più monotonicamente; ma questo risultato è dovuto solamente alle classi `UDE12_16` e `UDE16_40`.
La specializzazione colturale, descritta dall'indice OTE, fa aumentare significativamente la probabilità di cessazione solamente per le classi OTE_2 e OTE_5 rispettivamente: ortofloricoltura e granivori. Altri indicatori di specializzazione colturale come, ad esempio, la percentuale di superficie agraria destinata ad una particolare coltivazione risultano significative. Con l'aumento della percentuale di superficie coltivata a: fiori, frutta, orti, ortive e vivai aumenta la probabilità di cessazione dell'impresa.

Anche in questo caso, con l'aumentare dell'anzianità dell'azienda, aumenta la probabilità di cessazione, tuttavia questo aumento è debole, lo stesso vale per l'età del conduttore dell'azienda. Con l'aumentare del numero di giornate che il conduttore lavora in azienda, la probabilità di cessazione diminuisce. Anche il titolo di studio influenza l'uscita dal mercato dell'azienda: le aziende condotte da laureati hanno più probabilità di rimanere attive. Inoltre, le aziende con terreno in affitto hanno maggior probabilità di cessare rispetto alle aziende di proprietà.

Le aziende che hanno sede nella provincia di Treviso, hanno maggior

probabilità di cessare rispetto alla aziende con sedi nelle altre province del Veneto. Le aziende con minore probabilità di cessazione hanno sede a Padova.

All'aumentare della superficie aziendale, la probabilità di cessazione diminuisce significativamente, nonostante il valore molto piccolo assunto dal coefficiente, questo perchè l'intervallo dei valori di questa variabile è: 0-123788. Si rimanda alla figura 2.6 per maggiori dettagli.
Le aziende condotte da donne, *ceteris paribus*, hanno una minor probabilità di cancellazione rispetto a quelle condotte da uomini.

Un'ultima considerazione va fatta sulla scala dei coefficienti: tra i significativi, solo quelli relativi alla dimensione economica hanno valori maggiori all'unità. Questo sta a significare che gli altri coefficienti, seppur significativamente diversi da 0, hanno una minima influenza sull'aumento o sulla diminuzione del rischio di cessazione per l'impresa. Anche per questo modello viene analizzata la capacità classificatoria, come per il precedente.

Figura 3.3: Probabilità stimate di cessazione.

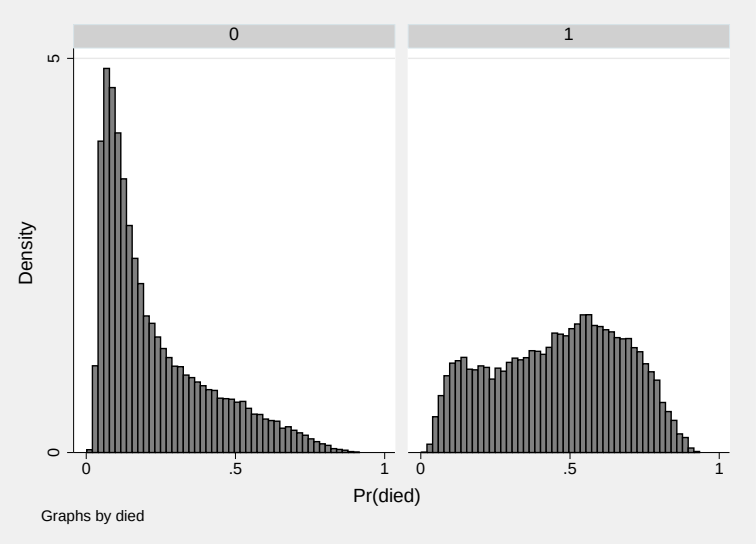


Tabella 3.4: Capacità classificatoria del modello.

Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	22921	26731	49652
0=ancora attiva (−)	4368	36667	41045
Falsi Positivi	$Pr(D\|-)$		10.64%
Falsi Negativi	$Pr(\neq D\|+)$		53.83%
Classificazioni corrette			65.71%

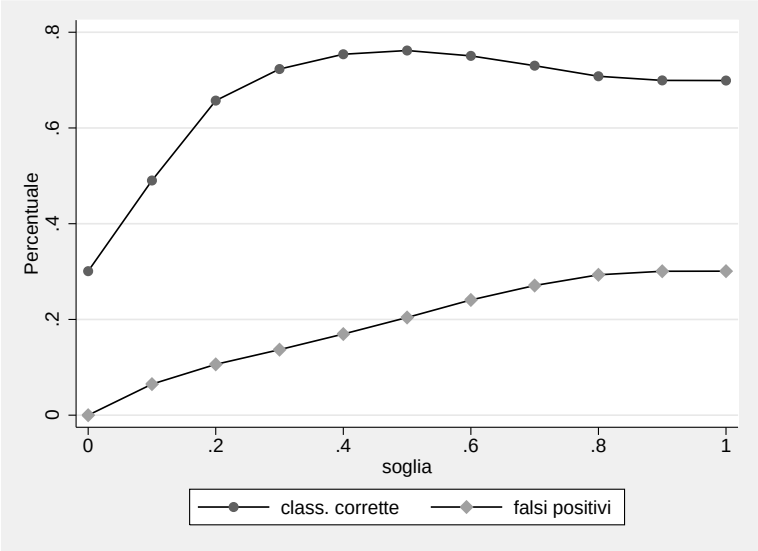
Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	20078	17896	37974
0=ancora attiva (−)	7211	45512	52723
Falsi Positivi	$Pr(D\|-)$		13.67%
Falsi Negativi	$Pr(\neq D\|+)$		47.12%
Classificazioni corrette			72.31%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	16753	11789	28534
0=ancora attiva (−)	10536	51619	62155
Falsi Positivi	$Pr(D\|-)$		16.95%
Falsi Negativi	$Pr(\neq D\|+)$		41.30%
Classificazioni corrette			75.40%

Selezionando una soglia corrispondente alla media delle probabilità di cessazione, 0.3, si finanzieranno 7211 aziende che saranno destinate a chiudere (pari al 13,67% delle aziende sovvenzionate). Inoltre non si finanzieranno 17896 aziende che invece sopravviveranno (pari al 47,12% delle aziende non finaziate).

Il modello tende ad esagerare nel classificare le imprese tra le cessate, ma

Figura 3.4: Sensibilità della classificazione al variare della soglia.



può essere efficacemente utilizzato ai fini di una politica che ha come scopo principale quello di sussidiare imprese destinate a sopravvivere.
Rispetto al modello che utilizza solo alcuni regressori, quello con tutte le variabili produce dei tenui miglioramenti nella capacità di classificazione, al massimo del 2%.

3.2.1 Capacità previsiva del modello

Per testare la capacità previsiva dei modelli, si è proceduto come segue:

1 stima del modello utilizzando un campione selezionato casualmente, pari al 50% della popolazone,

2 analisi delle previsioni estese all'altra metà della popolazione non utilizzata per la stima.

Utilizzando il modello con le covariate `UTE`, `ODE`, `t0_8_16`, `t0_16_23`, `t0_16_23`, `t0_23`, si ottengono risultati simili al modello che utilizza l'intera popolazione per quanto riguarda la probabilità di finanziare un'impresa che cesserà l'attività entro 5 anni. Si notano dei peggioramenti, nell'ordine di dieci punti

percentuali, per la probabilità di non finanziare aziende che rimangono attive per 5 anni. Risultati simili si ottengono per il modello che utilizza tutti i regressori.
I risultati utilizzati per valutare la capacità previsiva del modello con covariate `UTE`, `ODE`, `t0_8_16`, `t0_16_23`, `t0_16_23`, `t0_23`, sono riportati in figura 3.5 e in tabella 3.5

Figura 3.5: Probabilità stimate di cessazione.

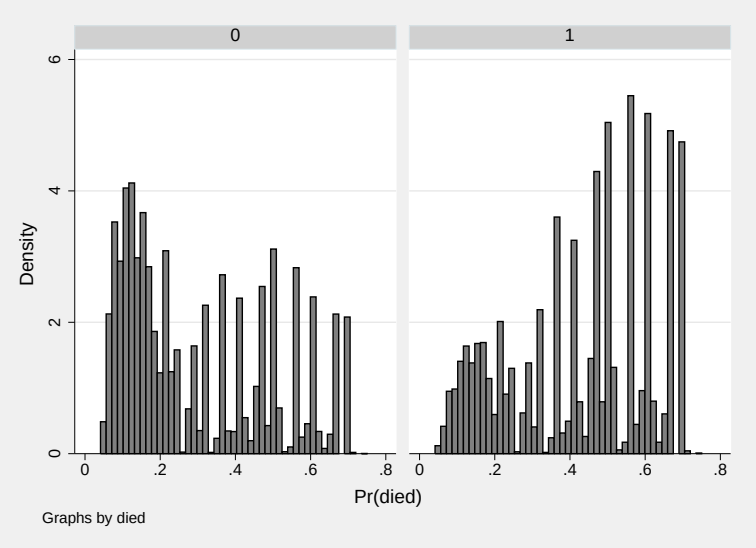


Tabella 3.5: Capacità classificatoria del modello.

Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	11232	24390	35622
0=ancora attiva (−)	2440	19490	21930
Falsi Positivi	$Pr(D\|-)$		11.13%
Falsi Negativi	$Pr(\neq D\|+)$		68.48%
Classificazioni corrette			53.83%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	9906	18665	28571
0=ancora attiva (−)	3766	25215	28981
Falsi Positivi	$Pr(D\|-)$		12.99%
Falsi Negativi	$Pr(\neq D\|+)$		65.33%
Classificazioni corrette			61.02%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	8406	14513	22919
0=ancora attiva (−)	5266	29367	34633
Falsi Positivi	$Pr(D\|-)$		15.21%
Falsi Negativi	$Pr(\neq D\|+)$		63.32%
Classificazioni corrette			65.36%

Capitolo 4

La probabilità di sopravvivenza annua

Con i dati in nostro possesso, non è possibile avere un'informazione precisa sulla durata di un'azienda agricola, poiché si conosce solamente l'anno in cui eventualmente questa è cessata. Non si è possesso di informazioni più dettagliate, che riguardano ad esempio intervalli di tempo più brevi: mesi o settimane. Questo comporta delle scelte, in termini di modelli utilizzabili, poiché non è lecito utilizzare modelli di durata ipotizzati per distribuzioni continue, ma bisogna necessariamente ricorrere a modelli per periodi discreti.

4.1 Analisi di sopravvivenza a tempi discreti

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

$$f(j) = Pr(T = j) \tag{4.1}$$

La funzione di sopravvivenza per l'istante j è data da:

$$S(j) = Pr(T \geq j) = \sum_{k=j}^{\infty} f_k \tag{4.2}$$

La funzione di rischio si può scrivere come :

$$\begin{aligned} h(j) &= \frac{f(j)}{S(j-1)} \\ &= Pr(T = j \,|T \geq j) \end{aligned} \tag{4.3}$$

$S(j)$ può ora essere scritta in una forma che evidenzia maggiormente il significato di sopravvivenza:

$$\begin{aligned} S(j) &= (1-h_1)(1-h_2)\ldots(1-h_{j-1})(1-h_j) \\ &= \prod_{k=1}^{j}(1-h_k) \end{aligned} \tag{4.4}$$

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

4.2 La verosimiglianza

Siano :

$$h_{ij} = Pr(T_i = j \,|T_j \geq j)$$

il rischio per il soggetto i di cessare all'istante j, condizionato alla sopravvivenza fino a j; c_i una variabile indicatrice che assume valore 1 se l'episodio è completo, 0 se l'episodio è censurato (cioè l'impresa è ancora attiva a fine periodo).
La verosimiglianza per un'osservazione censurata è data da:

$$\begin{aligned} L_i &= Pr(T_i > j) = S_i(j) \\ &= \prod_{k=1}^{j}(1-h_{ik}) \end{aligned} \tag{4.5}$$

per un'osservazione non censurata la verosimiglianza è:

$$\begin{aligned} L_i &= Pr(T_i = j) = f_i(j) \\ &= h_{ij}S_i(j-1) \\ &= \frac{h_{ij}}{1-h_{ij}}\prod_{k=1}^{j}(1-h_{ik}) \end{aligned} \tag{4.6}$$

La verosimiglianza per l'intero campione vale:

$$\begin{aligned} L &= \prod_{i=1}^{n} [Pr(T_i = j)]^{c_i} [Pr(T_i > j)]^{1-c_i} \\ &= \prod_{i=1}^{n} \left[\left(\frac{h_{ij}}{1-h_{ij}} \right) \prod_{k=1}^{j} (1-h_{ik}) \right]^{c_i} \left[\prod_{k=1}^{j} (1-h_{ik}) \right]^{1-c_i} \\ &= \prod_{i=1}^{n} \left[\left(\frac{h_{ij}}{1-h_{ij}} \right)^{c_i} \prod_{k=1}^{j} (1-h_{ik}) \right] \end{aligned} \tag{4.7}$$

La log-verosimglianza è data dall'equazione:

$$\log L = \sum_{i=1}^{n} c_i \log \left(\frac{h_{ij}}{1-h_{ij}} \right) + \sum_{i=1}^{n} \sum_{k=1}^{j} \log (1-h_{ik}) \tag{4.8}$$

Si definisca una nuova variabile indicatrice binaria y_{ik}; questa assumerà valore 1 se il soggetto i effettua una cambiamento di stato nell'intervallo k (l'azienda cessa), e 0 altrimenti.
Schematicamente:

$$\begin{aligned} c_i &= 1 \quad \Longrightarrow y_{ik} = 1 \text{ per } k = T_i,\ y_{ik} = 0 \text{ altrimenti} \\ c_i &= 0 \quad \Longrightarrow y_{ik} = \text{per tutti i valori di } k \end{aligned}$$

Con l'introduzione della variabile y_{ik} la log-verosimiglianza diventa :

$$\begin{aligned} \log L &= \sum_{i=1}^{n} \sum_{k=1}^{j} y_{ik} \log \left(\frac{h_{ik}}{1-h_{ik}} \right) + \sum_{i=1}^{n} \sum_{k=1}^{j} \log (1-h_{ik}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{j} [y_{ik} \log h_{ik} + (1-y_{ik}) \log(1-h_{ik})] \end{aligned} \tag{4.9}$$

Dall'equazione 4.9 si può riconoscere l'espressione della logverosimiglianza per un modello di regressione binaria dove y_{ik} è la variabile dipendente e la matrice dei dati viene trasformata secondo quanto illustra la tabella 4.2 . Con la nuova struttura dei dati ogni soggetto (azienda agricola) avrà tanti record quanti sono gli intervalli di tempo (anni) in cui è stato a rischio di cessazione; questa tecnica è conosciuta con il nome di *episode-splitting*.
Quindi la stima si ottiene applicando il modello LOGIT al dataset trasformato e utilizzando come variabile dipendente y_{ik}. Uno svantaggio di questo metodo consiste nel considerevole aumento del numero di record del dataset: per l'analisi in questione si passa da 90697 a 391730 record.

Tabella 4.1: Le due strutture dati a confronto.

struttura originale			struttura modificata				
soggetto i	c_i	T_i	soggetto i	c_i	T_i	y_{ik}	anno k
1	0	3	1	0	3	0	1
			1	0	3	0	2
			1	0	3	0	3
2	1	4	2	1	4	0	1
⋮	⋮	⋮	2	1	4	0	2
			2	1	4	0	3
			2	1	4	1	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

4.3 Stima della probabilità di sopravvivenza annua

Per la stima della probabilità di sopravvivenza anno per anno si è utilizzato il metodo proposto da Jenkins (1995) descritto precedentemente.
Un'estratto del dataset in formato *episode-splitting* è riportato nella tabella 4.3 .
La variabile *anno* k a differenza di quanto illustrato nella tabella 4.2 ha valore di "partenza" diverso per ogni soggetto poiché bisogna tener conto dell'anzianità pregressa, all'anno 2000, di ogni azienda.
Le variabili *dummy* vengono introdotte per controllare la variazione del rischio durante il periodo di riferimento [Allison, 1984]. Inizialmente si stima un modello utilizzando come regressori le variabili `OTE`, `UDE`, `t0`, come proposto precedentemente.

Tabella 4.2: Dataset in formato *episode-splitting*.

id_azienda	y_{ik}	anno k	dummy2	dummy3	dummy4	dummy5
4	0	4	0	0	0	0
4	0	5	1	0	0	0
4	0	6	0	1	0	0
4	0	7	0	0	1	0
4	0	8	0	0	0	1
5	0	18	0	0	0	0
5	1	19	1	0	0	0
6	0	17	0	0	0	0
6	0	18	1	0	0	0
6	1	19	0	1	0	0
7	0	13	0	0	0	0
7	0	14	1	0	0	0
7	1	15	0	1	0	0

Tabella 4.3: Stima del modello di sopravvivenza annua: `OTE` ,`UDE`, `t0` come regressori.

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
dummy2	0.3218222**	1.380	(0.0188291)
dummy3	0.3474839**	1.416	(0.0194316)
dummy4	0.1954367**	1.216	(0.0207723)
dummy5	0.0034656	1.003	(0.0224567)
ote1	0.5700261**	1.768	(0.1077140)
ote2	0.7787134**	2.179	(0.1218514)
ote4	0.1948404†	1.215	(0.1101598)
ote5	0.6387305**	1.894	(0.1325485)
ote6	0.4021583**	1.495	(0.1092397)
ote7	0.0158744	1.016	(0.1320446)
ote8	0.2409840*	1.273	(0.1132844)
ote311	-0.1105097	0.895	(0.1134245)
ote312	0.2747084*	1.316	(0.1103757)
ote3	0.3586813**	1.431	(0.1111984)
ude2_4	-0.6749518**	0.509	(0.0159284)
ude4_6	-1.3643965**	0.256	(0.0231878)
ude6_8	-1.7357062**	0.176	(0.0320768)
ude8_12	-1.8464231**	0.158	(0.0303879)
ude12_16	-2.0718619**	0.126	(0.0411322)
ude16_40	-2.1175705**	0.120	(0.0290918)
ude40_100	-2.1602009**	0.115	(0.0433110)
ude100	-2.3528448**	0.095	(0.0809696)
t0_8_16	0.1501179**	1.162	(0.0193770)
t0_16_23	0.3627300**	1.437	(0.0185256)
t0_23	0.4361293**	1.547	(0.0183161)
costante	-2.4265186**		(0.1082634)

N	391730
Log-likelihood	-89681.7788287
$\chi^2(25)$	18668.3230684
$Pr > \chi^2(25)$	0.0000

Livelli di significatività : † : 10% ∗ : 5% ∗∗ : 1%

La probabilità annua di cancellazione, a parità di caratteristiche aziendali, aumenta in modo statisticamente significativo nel 2001 e nel 2002 per poi diminuire nel 2003. Il coefficiente relativo all'anno 2004 non risulta statisticamente significativo. I coefficienti di `UDE` e `t0` continuano ad essere significativi e ad avere un andamento simile a quelli ottenuti dalle analisi proposte nel precedente capitolo.
Dopo aver ottenuto le stime dei coefficienti, è possibile calcolare la probabilità di cessazione per l'intero periodo di riferimento. Questa è ottenuta a partire dalle stime delle probabilità di cessazione annue calcolate per tutti gli anni in cui un'azienda è stata a rischio di cessazione. Quindi, se un'azienda è sopravvissuta per l'intero periodo, si avranno 5 diverse probabilità; 4 se è sopravvissuta per 4 anni e così via.
Seguendo quanto proposto da Allison (1984), si è stimato il modello senza introdurre le variabili *dummy*, ipotizzando che il rischio non vari nel tempo. Il test del rapporto di log-verosimiglianza utilizzato per confrontare i due modelli rifiuta l'ipotesi nulla di uguaglianza.

Tabella 4.4: Test del confronto tra i due modelli.

modello	log-verosimiglianza	gradi di libertà
modello senza *dummy*	-89954.98	22
modello con *dummy*	-89681.78	26
Lr test $\chi^2(4) = 546.40$	$Pr > \chi^2(4) = 0.000$	

Se un'azienda i è sopravvisuta, dal 1999 al 2004, per t anni, con $t \leq 5$, nel dataset in formato *episode-splitting* si avranno solo t record in corrispondenza dell'azienda i. Questo è corretto per ottenere le stime, ma per calcolare la probabilità di sopravvivenza per 5 anni, c'è bisogno di avere 5 record per ogni azienda. Per cui vengono creati $5 - t$ record, con gli stessi valori che assumono le variabili nell'azienda i.

Sia p_{ij} la probabilità di cessazione dell'azienda i nell'anno j.
La probabilità di cessazione in 5 anni sarà:

$$Pr(c_i = 1) = 1 - \prod_{j=1}^{5} [1 - Pr(y_{ij} = 1)] \tag{4.10}$$

Le probabilità stimate di cessazione vengono riportate in figura 4.1 .

Figura 4.1: Probabilità stimate di cessazione.

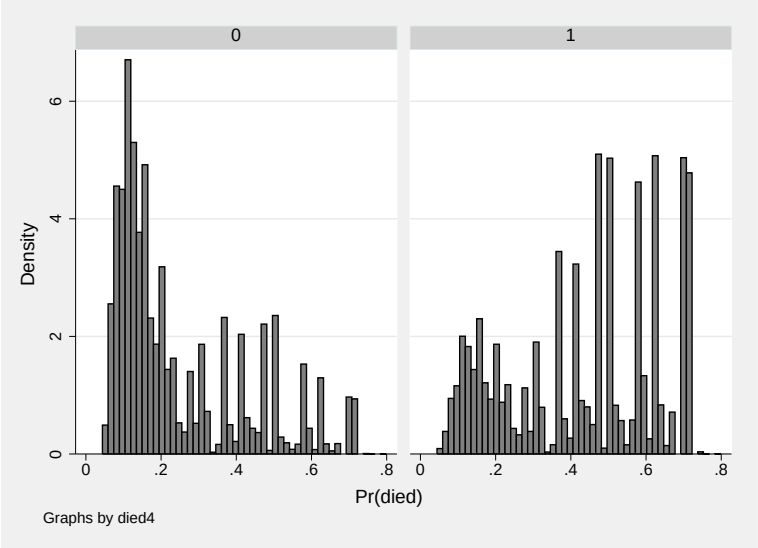


Tabella 4.5: Capacità classificatoria del modello.

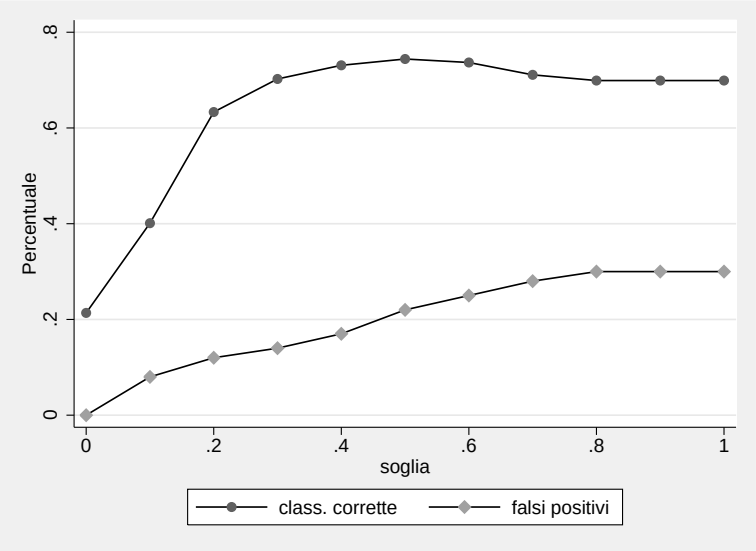
Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	22188	27885	50073
0=ancora attiva (−)	5101	35523	40624
Falsi Positivi	$Pr(D\|-)$		12.55%
Falsi Negativi	$Pr(\neq D\|+)$		55.68%
Classificazioni corrette			63.63%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	19385	19385	39069
0=ancora attiva (−)	7605	44023	51628
Falsi Positivi	$Pr(D\|-)$		14.73%
Falsi Negativi	$Pr(\neq D\|+)$		49.61%
Classificazioni corrette			70.24%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	16290	13858	30148
0=ancora attiva (−)	10999	49550	60549
Falsi Positivi	$Pr(D\|-)$		17.55%
Falsi Negativi	$Pr(\neq D\|+)$		45.29%
Classificazioni corrette			73.08%

Come già illustrato nel § 3.2, viene di seguito valutata la sensibilità della classificazione, al variare della soglia di probabilità stimata oltre cui si considera un'azienda come cessata.
In termini di capacità classificatoria del modello non si notano significativi miglioramenti rispetto al modello che fa riferimento ad un unico episodio per tutto il periodo di riferimento.

Figura 4.2: Sensibilità della classificazione al variare della soglia.



Si confronti a tal proposito la tabella 3.2 con la tabella 4.3.
Si stima ora il modello inserendo tutti i regressori utilizzabili.

Tabella 4.6: Stima del modello di sopravvivenza annua: tutti i regressori.

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
dummy2	0.3491206**	1.418	(0.0190301)
dummy3	0.3949729**	1.484	(0.0196479)
dummy4	0.2629346**	1.301	(0.0209980)
dummy5	0.0833948**	1.087	(0.0226833)
adesione_ass_prod	-0.0815393**	0.922	(0.0248147)
adesione_soc_coop	-0.1518130**	0.859	(0.0182541)
allev_avicoli	-1	0.368	(1.4000000)
azienda_individuale	-0.0685383	0.934	(0.0493990)
bovini	-0.0006909*	0.999	(0.0003264)
cond_prof_condu	0.1230817**	1.131	(0.0224323)
conigli	-0.000014	1.000	(0.0000196)

Continua nella prossima pagina...

... tabella 4.6

Variabile	Coefficiente	Odds ratio	(Std. Err.)
equini	-0.0158044	0.984	(0.0101188)
parchi	-0.003544	0.996	(0.0376976)
lavoraz_prod_agric	-0.0235245	0.977	(0.0226185)
lic_scu_elem	-0.0059021	0.994	(0.0280569)
lic_scu_inf	-0.0845168*	0.919	(0.0333623)
serre	-7.4	0.001	(5.4000000)
sesso	-0.1547784**	0.857	(0.0170474)
suini	0.0000219	1.000	(0.0000524)
sup_sau_azienda	-0.0000762**	1.000	(0.0000150)
sup_sau_tot2	0.0000000**	1.000	(0.0000000)
lav_fam	0.0003595**	1.000	(0.0000675)
cond_dir	-0.0922891**	0.912	(0.0187752)
val_prod_vend_meno10m	-0.0992717**	0.905	(0.0383038)
val_prod_vend_tra_10_50m	-0.0040011	0.996	(0.0372995)
prod_bio	-0.0078029	0.992	(0.0782892)
nr_abit	-0.0016216	0.998	(0.0099520)
ovicapri	0.0006795	1.001	(0.0008629)
laurea	-0.1942314**	0.823	(0.0699537)
diploma	-0.1146063**	0.892	(0.0351356)
bel	-0.1696944**	0.844	(0.0540043)
rov	-0.2754637**	0.759	(0.0328466)
pad	-0.3559448**	0.701	(0.0196518)
vic	-0.0847146**	0.919	(0.0234186)
ven	-0.0346275	0.966	(0.0218808)
ver	-0.3331304**	0.717	(0.0255659)
senza_sup	-0.196882	0.821	(0.1709068)
adesione_consorzi	-0.0747213*	0.928	(0.0306112)
eta	0.0238148**	1.024	(0.0007374)
eta2	0.0001823**	1.000	(0.0000320)
lav_0_30	0.8064739**	2.240	(0.0416279)
lav_30_90	0.6070247**	1.835	(0.0395865)

Continua nella prossima pagina...

... tabella 4.6

Variabile	Coefficiente	Odds ratio	(Std. Err.)
lav_90_270	0.3148912**	1.370	(0.0336452)
capo_azienda	-0.1356825**	0.873	(0.0339490)
cod_att101	-0.0109165	0.989	(0.0316171)
cod_att102	0.0659137*	1.068	(0.0292387)
cod_att103	0.0105798	1.011	(0.0251184)
cod_att104	-0.000739	0.999	(0.0654419)
cod_att111	0.0094015	1.009	(0.0190182)
cod_att112	-0.0083179	0.992	(0.0352392)
affitto	0.1137717**	1.120	(0.0281938)
uso_grat	0.052063	1.053	(0.0404029)
perc_cer	0.2985781*	1.348	(0.1218307)
perc_fiori	0.4629110*	1.589	(0.2036063)
perc_foraggi	-0.0513992	0.950	(0.1294110)
perc_frutta	0.7542761**	2.126	(0.1413859)
perc_legno	-1.2502874†	0.286	(0.6981763)
perc_legumi	0.7300734†	2.075	(0.3941981)
perc_olivo	0.240202	1.272	(0.1812743)
perc_orti	0.6082938**	1.837	(0.1671194)
perc_ortive	0.5002601**	1.649	(0.1353624)
perc_patata	0.4017749	1.494	(0.3497571)
perc_piante	0.2945223*	1.342	(0.1259382)
perc_prati	0.2680672*	1.307	(0.1249295)
perc_barb	-0.0441174	0.957	(0.1480080)
perc_vite	0.1676948	1.183	(0.1300402)
perc_vivai	0.8237799**	2.279	(0.1971637)
perc_altra	0.0486322	1.050	(0.0679432)
perc_boschi	0.0604483	1.062	(0.0595120)
perc_sanu	0.0865916	1.090	(0.1422409)
perc_arbo	-0.1039794	0.901	(0.1659993)
ude2_4	-0.6320776**	0.531	(0.0167872)
ude4_6	-1.2369284**	0.290	(0.0247365)

Continua nella prossima pagina...

... tabella 4.6

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
ude6_8	-1.5432697**	0.214	(0.0339612)
ude8_12	-1.5751270**	0.207	(0.0336810)
ude12_16	-1.7046996**	0.182	(0.0453936)
ude16_40	-1.6202265**	0.198	(0.0405437)
ude40_100	-1.5550419**	0.211	(0.0666520)
ude100	-1.5796022**	0.206	(0.1240387)
ote1	0.1308683	1.140	(0.1565570)
ote2	0.4084220*	1.504	(0.1716729)
ote4	0.0645142	1.067	(0.1589566)
ote5	0.3985669*	1.490	(0.1612905)
ote6	0.0797222	1.083	(0.1568630)
ote7	-0.1832108	0.833	(0.1734327)
ote8	0.0175109	1.018	(0.1592323)
ote311	-0.2225459	0.800	(0.1651640)
ote312	0.0165151	1.017	(0.1611780)
ote3	0.013721	1.014	(0.1626032)
t0_8_16	0.0729137**	1.076	(0.0198397)
t0_16_23	0.0909208**	1.095	(0.0200770)
t0_23	0.1010017**	1.106	(0.0204019)
costante	-2.3575548**		(0.1946044)

N	391730
Log-likelihood	-87362.9048668
$\chi^2(92)$	23306.
$Pr. > \chi^2(92)$	0.0000709921

Livelli di significatività : † : 10% ∗ : 5% ∗∗ : 1%

Anche in questo caso, non si notano differenze significative nel valore dei coefficienti, rispetto a quelli del modello stimato nel § 3.2, che utilizza gli stessi regressori, a meno delle variabili *dummy*. Anche le analisi della

capacità classificatoria del modello, che sono riportate di seguito, forniscono miglioramenti irrisori, nell'ordine di alcuni decimali.
Anche in questo caso, si è testata la capacità previsiva dei modelli come proposto nel § 3.2.1. La capacità classificatoria dei modelli che utilizzano metà della popolazione è molto simile a quella dei modelli che utilizzano l'intera popolazione: si evidenziano piccoli scostamenti, nell'ordine del punto percentuale.

Figura 4.3: Probabilità stimate di cessazione.

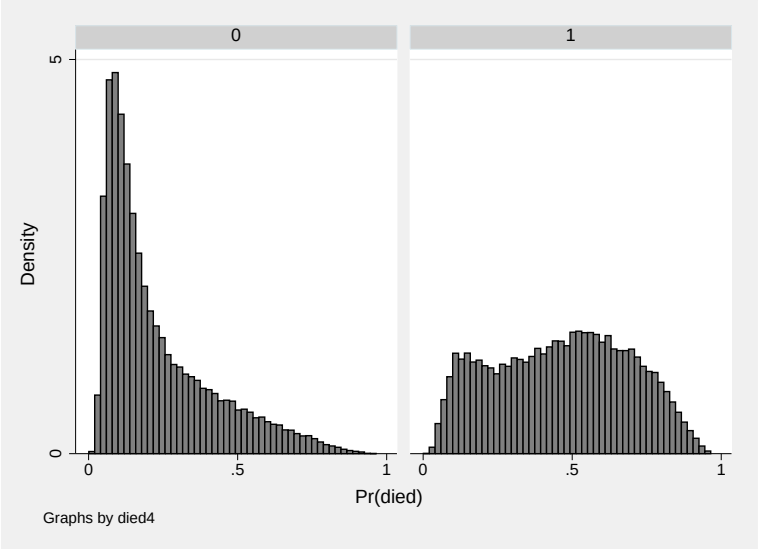
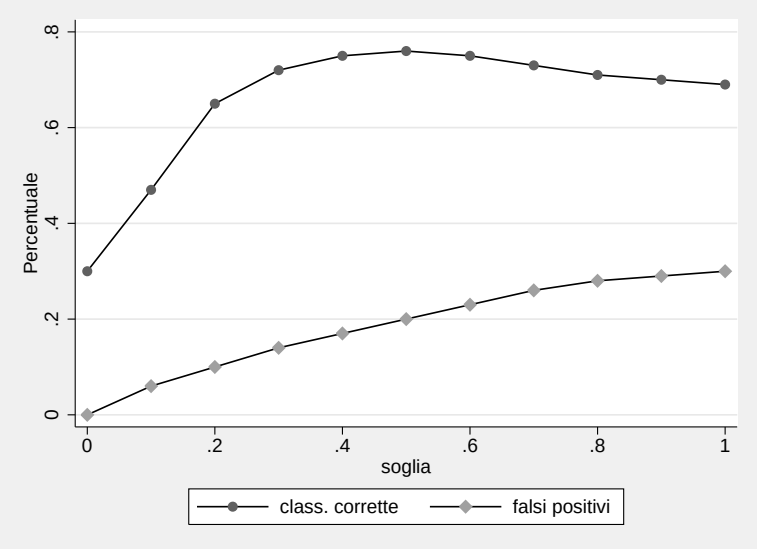


Tabella 4.7: Capacità classificatoria del modello.

Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	22801	26531	49332
0=ancora attiva (−)	4488	36877	41365
Falsi Positivi	$Pr(D\|-)$		10.85%
Falsi Negativi	$Pr(\neq D\|+)$		53.78%
Classificazioni corrette			65.79%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	19797	17473	37270
0=ancora attiva (−)	7492	45935	53427
Falsi Positivi	$Pr(D\|-)$		14.00%
Falsi Negativi	$Pr(\neq D\|+)$		46.88%
Classificazioni corrette			72.47%
Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	16445	11391	27834
0=ancora attiva (−)	10844	52017	62863
Falsi Positivi	$Pr(D\|-)$		17.24%
Falsi Negativi	$Pr(\neq D\|+)$		40.91%
Classificazioni corrette			75.49%

Figura 4.4: Sensibilità della classificazione al variare della soglia.



Capitolo 5

Eterogeneità non osservata

5.1 Il processo generatore dei dati

Il dataset che si sta analizzando contiene le caratteristiche socio-economiche di aziende agricole venete nate in dato anno t e sopravvissute almeno fino all'anno 2000, anno in cui è stato effettuato il V Censimento Generale dell'Agricoltura. Tutte le aziende cessate prima del 2000 sono automaticamente escluse dalla popolazione in analisi: vengono sistematicamente sottorappresentate le aziende che hanno avuto vita "breve" [Kiefer, 1984].
Questo aspetto provoca una distorsione nelle stime nota in letteratura come *length-bias* [Vardi, 1985], [Davidov et al., 2001].

5.1.1 Length-bias

Quando la probabilità, per un soggetto, di essere incluso nel campione è correlata ad una variabile misurata, si ha una selezione del campione distorta per lunghezza (*lenght-biased*).
Si supponga G l'ignota distribuzione della variabile casuale positiva Y. In una procedura di campionamento che soffre di *lenght-bias* non si osservano i valori Y_i, ma bensì i valori $Z_i \cdots Z_n$ la cui funzione di ripartizione, nel caso continuo, è [Gill et.al, 1988]:

$$F(z) = \frac{1}{\mu} \int_0^z y \, \mathrm{d}G(y) \qquad \text{per} \quad z \geq 0 \tag{5.1}$$

Tuttavia, se si assumono tassi di ingresso ed uscita costanti nel tempo, la trattazione si fa più semplice. Le aziende che sono state attive per t anni e che sono entrate nel campione forniscono lo stesso contributo alla verosimiglianza delle aziende attive per t anni che sono cessate prima dell'anno 2000 [Jenkins, 1995], considerando la necessità di modellare l'intervallo di tempo che va dagli anni 1999 al 2004.
Si ponga l'anno $t = 1834$ come il primo anno in cui si osserva la nascita di un'azienda agricola, mentre $t = \tau$ l'anno di selezione del campione (in questo caso l'anno 2000). La situazione delle imprese viene rilevata all'anno $t = \tau, t = \tau + 1, \cdots, t = \tau + 4$.

Si consideri l'azienda i nata nell'anno 1990, ancora attiva all'anno 2003. La probabilità di rimanere attiva dall'anno 1990 all'anno 2003 è:

$$(1 - h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(\cdots)(1 - h_{i,1990}) \tag{5.2}$$

dove $h_{it} = Prob(T_i = t | T_i \geq t;\ X_{it})$.

La probabilità, per l'azienda i di rimanere attiva dal 1990 al 2003, condizionatamente ad essere sopravvissuta fino all'anno 2000 risulta:

$$\frac{(1 - h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(\cdots)(1 - h_{i,1990})}{(1 - h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(1 - h_{i,2000})} =$$

$$= (1 - h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(1 - h_{i,2000}) \tag{5.3}$$

Mentre la probabilità, per l'azienda i di cessare nel 2003, condizionatamente alla sua sopravvivenza fino all'anno 2000 sarà:

$$\frac{(h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(\cdots)(1 - h_{i,1990})}{(1 - h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(1 - h_{i,2000})} =$$

$$= (h_{i,2003})(1 - h_{i,2002})(1 - h_{i,2001})(1 - h_{i,2000}) \tag{5.4}$$

La probabilità di sopravvivenza condizionata, e quindi il contributo alla verosimiglianza, dipendono solamente dai dati per gli anni in cui l'azienda è a rischio tra il 1999 e il 2004.

5.2 L'eterogeneità non osservata

Nelle analisi considerate nei precedenti capitoli, tutte le differenze tra i soggetti (le aziende agricole), si supponevano catturate dal vettore delle

esplicative, X.
Si consideri ora un'estensione dei modelli proposti che permetta di considerare eterogeneità non osservata tra i soggetti.

Considerare eterogeneità non osservata significa ammettere la presenza di una, o più variabili, che non vengono osservate, ma che hanno effetto sulla sopravvivenza di un'azienda agricola.
Per tenere conto di questo aspetto è necessario considerare un nuovo modello. [Jenkins, 2004]

$$\frac{h(j,\mathbf{X}|\nu)}{1-h(j,\mathbf{X}|\nu)} = \left[\frac{h_0(j)}{1-h_0(j)}\right] exp(\beta'\mathbf{X}+\nu) \tag{5.5}$$

$$logit[h(j,\mathbf{X}|\nu)] = D(j) + \beta'\mathbf{X} + \nu \tag{5.6}$$

dove :

- D(j) è un termine che caratterizza la funzione di rischio base.
- ν è un termine di errore, inserito per considerare l'eterogeneità non osservata, con distribuzione normale, con media 0 e varianza finita.

In un dataset in formato *episode-splitting*, il modello descritto dall'equazione 5.6 può essere stimato con un modello per dati di panel a risposta binaria, con effetti casuali[1]. Si è in presenza di un panel non bilanciato poiché non tutti i soggetti sono osservati per lo stesso periodo di tempo.

5.2.1 Modello ad effetti casuali con variabile risposta binaria

Un modello per dati di panel con variabile esplicativa dicotomica è descritto nell'equazione 5.7 [Wooldridge, 2002].

$$Pr(y_{it} \neq 1|\boldsymbol{x}_i,\nu_i) = Pr(y_{it} \neq 1|\boldsymbol{x}_{it},\nu_i) = \pi(\boldsymbol{x}_{it}\beta+\nu_i) = \frac{exp(\boldsymbol{x}_{it}\boldsymbol{\beta}+\nu_i)}{1+exp(\boldsymbol{x}_{it}\boldsymbol{\beta}+\nu_i)}, \qquad t=1,\cdots,n_i \tag{5.7}$$

[1] Il modello è conosciuto in letteratura anche come modello a componenti di varianza.

dove ν_i rappresenta l'effetto non osservato e $\boldsymbol{x}_i$ contiene $\boldsymbol{x}_{it}$ per tutti i valori di t.
Si assuma inoltre:

(1) $y_{i1}, y_{i2}, \cdots y_{it}$ indipendenti condizionatamente a $(\boldsymbol{x}_i, \nu_i)$

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

Si può evidenziare il modello 5.7 come un modello a componenti di varianza:

$$y_{it} \neq 1 \Longleftrightarrow \boldsymbol{x}_i\boldsymbol{\beta} + \nu_i + \epsilon_{it} > 0 \tag{5.8}$$

dove ϵ_{it} ha distribuzione logistica con media 0 e varianza $\sigma_\epsilon^2 = \pi^2/3$, indipendentemente da ν_i.

Sotto le assunzioni (1) e (2) e 5.7 si può derivare la densità di $(y_{i1}, \cdots, y_{in_i})$ condizionatamente ai valori di $(\boldsymbol{x}_i, \nu_i)$:

$$f(y_1, \cdots, y_{n_i}|\boldsymbol{x}_i, \nu_i; \boldsymbol{\beta}) = \prod_{t=1}^{n_i} f(y_t|\boldsymbol{x}_{it}, \nu, \boldsymbol{\beta}) \tag{5.9}$$

$$f(y_t|\boldsymbol{x}_{it}, \nu, \boldsymbol{\beta}) = \pi(\boldsymbol{x}_t\boldsymbol{\beta} + \nu)^{y_t}[1 - \pi(\boldsymbol{x}_t\boldsymbol{\beta} + \nu)]^{1-y_t} \tag{5.10}$$

Partendo dall'equazione 5.9 si possono stimare i parametri $\boldsymbol{\beta}$ e σ_ν^2 utilizzando il metodo della massima verosimiglianza condizionata.
Poiché i valori di ν_i non vengono osservati, non possono comparire nella funzione di verosimiglianza. Si può ottenere la distribuzione congiunta di $(y_{i1}, \cdots, y_{it})$ condizionatamente a $\boldsymbol{x}_i$ integrando la funzione in ν_i.
Utilizzando tutte le assunzioni fatte in precedenza,

$$Pr(y_{i1}, \cdots, y_{in_i}|\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{in_i}) = \int_{-\infty}^{\infty} \frac{e^{-\nu_i^2}/2\sigma_\nu^2}{\sqrt{2\pi}\sigma_\nu} \left\{ \prod_{t=1}^{n_i} f(y_{it}, \boldsymbol{x}_{it}\boldsymbol{\beta} + \nu_i) \right\} \mathrm{d}\nu_i \tag{5.11}$$

dove

$$f(y_{it}, \boldsymbol{x}_{it}\boldsymbol{\beta} + \nu_i) = \pi(y_{it}, \boldsymbol{x}_{it}\boldsymbol{\beta} + \nu_i)^{y_{it}}[1 - \pi(y_{it}, \boldsymbol{x}_t\boldsymbol{\beta} + \nu_i)]^{1-y_{it}} \tag{5.12}$$

Il software STATA ver 9.1 calcola un'approssimazione dell'integrale 5.11 utilizzando il metodo della quadratura di Gauss-Hermite che permette di approssimare integrali del tipo :

$$\int_{-\infty}^{\infty} e^{-x^2} g(x)\mathrm{d}x \approx \sum_{m=1}^{M} w_m^* g(a_m^*) \tag{5.13}$$

dove i w_m^* rappresentano i pesi della quadratura mentre a_m^* ne denotano le ascisse [Stata, 2005].
Le analisi sono state condotte con un valore di $M = 16$. La scelta di un valore di $M = 12$ o $M = 20$ non porta a conclusioni significativamente diverse.

5.2.2 Gli effetti sui coefficienti

Gli effetti dell'omissione di regressori ortogonali, si possono evidenziare semplicemente nel caso di modelli probit [Wooldridge, 2002].
Si supponga che il modello di interesse sia :

$$P(y = 1|x, c) = \Phi(\boldsymbol{X\beta} + \gamma c) \tag{5.14}$$

L'equazione 5.14 si può riscrivere in forma di variabile latente:

$$y^\star = \boldsymbol{X\beta} + \gamma c + e \tag{5.15}$$

dove $y = 1[y^\star > 0]$ e $e|x, c \sim N(0, 1)$.
Si supponga c indipendente rispetto a x e $c \sim N(0, \tau^2)$.
Sotto queste assunzioni, il termine di errore $\gamma c + e$ risulta indipendente da x e ha distribuzione $N(0, \gamma^2\tau^2 + 1)$.
Quindi:

$$P(y = 1|\boldsymbol{x}) = P(\gamma c + e > -\boldsymbol{x\beta}|\boldsymbol{x}) = \Phi(\boldsymbol{x\beta}/\sigma) \tag{5.16}$$

dove $\sigma^2 \equiv \gamma^2\tau^2 + 1$. Utilizzando la 5.16 si dimostra che $plim\ \hat{\beta} = \beta_j/\sigma$. Poichè $\sigma = \sqrt{\gamma^2\tau^2 + 1} > 1$ (a meno che $\gamma = 0$ o $\tau^2 = 0$), $|\beta_j/\sigma| < |\beta_j|$.
Nei modelli logit, il termine di errore ha distribuzione logistica[2], e questo aspetto rende problematica la dimostrazione analitica degli effetti dell'omissione di regressori ortogonali. Studi di simulazione [Cramer, 2007] evidenziano comunque, anche per i modelli logit, una riduzione dei restanti coefficienti verso lo 0.
Tuttavia, le distorsioni dei coefficienti non hanno un grosso impatto sulle derivate, e di conseguenza i risultati delle analisi non variano di molto.

In un modello logit, l'effetto marginale di X_j su $Pr(Y_j = 1|X_j)$ dipende

[2]La funzione di densità di probabilità di una v.c. logistica è $f(x; \mu) = \frac{e^{-(x-\mu)}}{(1+e^{-(x-\mu)})^2}$

da X_j:

$$\frac{\delta Pr(Y_j = 1|X_j)}{\delta X_j} = \beta_j \cdot \pi(\boldsymbol{X}'\boldsymbol{\beta})(1 - \pi(\boldsymbol{X}'\boldsymbol{\beta})) \tag{5.17}$$

Quindi, una diminuzione di β_j, può essere compensata da cambiamenti di segno opposto di altri termini.

Confrontando i coefficienti delle tabelle 5.3 e 4.6 si ottengono risultati coerenti con quanto illustrato precedentemente. Includendo l'eterogeneità non osservata, i coefficienti risultano quasi tutti aumentati (solo 6 su 93 diminuiscono, ma questo effetto è dovuto alle approssimazioni utilizzate per massimizzare la verosimiglianza). Il rapporto medio tra i coefficienti nei due modelli risulta pari a 2.48. Si evidenzia un solo cambiamento di segno, nella variabile relativa alla presenza di produzioni biologiche all'interno dell'azienda, comunque il coefficiente relativo a questa variabile risulta non significativo in entrambi i modelli.

Per alcuni coefficienti, il test di significatività, con un livello $\alpha = 5\%$ porta a conclusioni diverse rispetto al modello che non considera eterogeneità non osservata.

La tabella 5.1 evidenzia queste differenze.

Tabella 5.1: Variazioni nella significatività dei coefficienti.

Coefficiente	Eterogeneità non osservata $P > \|z\|$	Eterogeneità osservata $P > \|z\|$
azienda_individuale	0.04	0.165
bovini	0.091	0.034
ven	0.003	0.114
perc_legno	0.03	0.073
perc_legumi	0.047	0.064
perc_prati	0.088	0.032

Il risultato più interessante che si ottiene è la stima del parametro ρ:

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2} \tag{5.18}$$

Se $\rho = 0$, non c'è presenza di eterogeneità non osservata, perché questo implica $\sigma_\nu^2 = 0$, quindi il termine ν può non essere considerato.

Con i dati in esame si ottiene:

$$\begin{aligned} \sigma_\nu &= 1.917969 \quad \text{Std.Err.}(0.1206) \\ \rho &= 0.5279 \quad \text{Std.Err.}(0.03134) \\ \text{Lr Test } \rho = 0 &: \chi^2 bar(01) = 516.09 \\ Pr &\geq \chi^2 bar(01) = 0.000 \end{aligned}$$

Si può concludere che la presenza di eterogeneità non osservata è statisticamente significativa.
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, 2002]

Tabella 5.2: Stima del modello di sopravvivenza con eterogeneità non osservata.

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
dummy2	0.929908 **	2.534276013	(0.0657908)
dummy3	1.363645 **	3.91042084	(0.1011945)
dummy4	1.4959 **	4.463351762	(0.1247723)
dummy5	1.505744 **	4.507505968	(0.1413447)
adesione_ass_prod	-0.1187545 **	0.888025784	(0.039281)
adesione_soc_coop	-0.2647425 **	0.767403531	(0.0307919)
allev_avicoli	-0.000000436	0.999999564	(0.00000206)
azienda_individuale	-0.1490198 *	0.861552056	(0.0726733)
bovini	-0.0007185 †	0.999281758	(0.0004251)
cond_prof_condu	0.2004619 **	1.221967054	(0.0370927)
conigli	-0.0000256	0.9999744	(0.0000279)
equini	-0.0137628	0.986331474	(0.0126456)
parchi	-0.0215446	0.978685827	(0.0608008)
lavoraz_prod_agric	-0.0495553	0.95165253	(0.0362072)

Continua nella prossima pagina...

... tabella 5.2

Variabile	Coefficiente	Odds ratio	(Std. Err.)
lic_scu_elem	-0.0190675	0.981113135	(0.0495807)
lic_scu_inf	-0.1655969 **	0.847387747	(0.0577198)
serre	-0.0000106	0.9999894	(0.00000778)
sesso	-0.2220861 **	0.800846408	(0.0294394)
suini	0.0000174	1.0000174	(0.0000814)
sup_sau_azienda	-0.0000918 **	0.999908204	(0.0000195)
sup_sau_tot2	1.1E-09 **	1.000000001	(0.000000000218)
lav_fam	0.0005199 **	1.000520035	(0.0001016)
cond_dir	-0.1557331 **	0.85578757	(0.0326242)
val_prod_vend_meno10m	-0.1651286 **	0.847784672	(0.0583452)
val_prod_vend_tra_10_50m	-0.0124538	0.987623428	(0.0557298)
prod_bio	0.0249282	1.025241506	(0.116885)
nr_abit	-0.0103474	0.98970595	(0.0163326)
ovicapri	0.0007508	1.000751082	(0.0012492)
laurea	-0.2877052 **	0.749982655	(0.1093608)
diploma	-0.188826 **	0.827930554	(0.0604601)
bel	-0.2747572 **	0.75975657	(0.0896013)
rov	-0.4810915 **	0.618108358	(0.0565281)
pad	-0.6388164 **	0.527916897	(0.0422692)
vic	-0.1668724 **	0.846307593	(0.0399065)
ven	-0.1129485 **	0.893196659	(0.0378408)
ver	-0.5571158 **	0.572858923	(0.0466378)
senza_sup	-0.3099993	0.73344747	(0.2838621)
adesione_consorzi	-0.1137498 *	0.892481227	(0.0488849)
eta	0.038315 **	1.039058485	(0.0018348)
eta2	0.000416 **	1.000416087	(0.0000548)
lav_0_30	1.264146 **	3.540068227	(0.0786075)
lav_30_90	0.9117055 **	2.488563161	(0.0680067)
lav_90_270	0.4408762 **	1.554068297	(0.0515758)
capo_azienda	-0.2223733 **	0.800616438	(0.0589337)
cod_att101	-0.0113094	0.988754311	(0.0524995)

Continua nella prossima pagina...

... tabella 5.2

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
cod_att102	0.1037061 *	1.109274391	(0.0488528)
cod_att103	0.0195474	1.019739701	(0.0417712)
cod_att104	-0.009821	0.990227069	(0.1084552)
cod_att111	0.0245811	1.024885706	(0.0316428)
cod_att112	-0.0272233	0.973143914	(0.0586449)
affitto	0.1716058 **	1.187209743	(0.0437716)
uso_grat	0.1261035 †	1.134399573	(0.0679836)
perc_cer	0.4222614 *	1.525407214	(0.206186)
perc_fiori	0.7367889 *	2.08921605	(0.3182925)
perc_foraggi	-0.1179097	0.888776305	(0.21765)
perc_frutta	1.133596 **	3.10680852	(0.2358399)
perc_legno	-2.216092 *	0.109034384	(1.020859)
perc_legumi	1.338696 *	3.814066718	(0.6730014)
perc_olivo	0.1973226	1.218136948	(0.2993129)
perc_orti	1.107015	3.02531434	(0.3068713)
perc_ortive	0.7880325 **	2.199065506	(0.2271882)
perc_patata	0.6289961	1.875726592	(0.5344673)
perc_piante	0.419185 *	1.520721662	(0.2131877)
perc_prati	0.3606996 †	1.434332523	(0.2114068)
perc_barb	-0.1305883	0.877578999	(0.245674)
perc_vite	0.1782616	1.195137928	(0.2178544)
perc_vivai	1.228169 **	3.414970996	(0.3093511)
perc_altra	0.0784847	1.081646806	(0.1133708)
perc_boschi	0.1075449	1.113540858	(0.0991258)
perc_sanu	0.1060774	1.111907935	(0.2354432)
perc_arbo	-0.1480466	0.862390927	(0.2731842)
ude2_4	-1.143874 **	0.31858244	(0.0549781)
ude4_6	-2.178634 **	0.113196051	(0.0970973)
ude6_8	-2.641872 **	0.071227806	(0.1177961)
ude8_12	-2.678462 **	0.068668685	(0.1182624)
ude12_16	-2.861522 **	0.057181664	(0.131495)

Continua nella prossima pagina...

... tabella 5.2

Variabile	**Coefficiente**	**Odds ratio**	**(Std. Err.)**
ude16_40	-2.724682 **	0.06556705	(0.1235506)
ude40_100	-2.64958 **	0.070680893	(0.1432256)
ude100	-2.699217 **	0.067258155	(0.2054571)
ote1	0.345904	1.413266936	(0.2667579)
ote2	0.7297871 *	2.07463887	(0.28798)
ote4	0.2521508	1.286790071	(0.270702)
ote5	0.7657609 **	2.150630167	(0.2797121)
ote6	0.2365689	1.266894842	(0.2671827)
ote7	-0.1111469	0.894807292	(0.2893527)
ote8	0.1526155	1.164876998	(0.2706872)
ote311	-0.1469234	0.863360108	(0.2787104)
ote312	0.1688209	1.183908082	(0.2737253)
ote3	0.2065105	1.229380643	(0.2751883)
t0_8_16	0.1047426 **	1.11042475	(0.032581)
t0_16_23	0.1343226 **	1.143761738	(0.0335971)
t0_23	0.1627254 **	1.17671352	(0.0344782)
costante	-3.404128 **		(0.3406676)

N	391730
Log-likelihood	-87104.859

Livelli di significatività : † : 10% ∗ : 5% ∗∗ : 1%

Le previsioni vengono calcolate assumendo $\rho = 0$, cioè assenza di eterogeneità non osservata: alla variabile ν_i viene assegnato il valore della sua media: $E(\nu_i) = 0$, che deriva dalle ipotesi fatte nella 5.6 .

Figura 5.1: Probabilità stimate di cessazione.

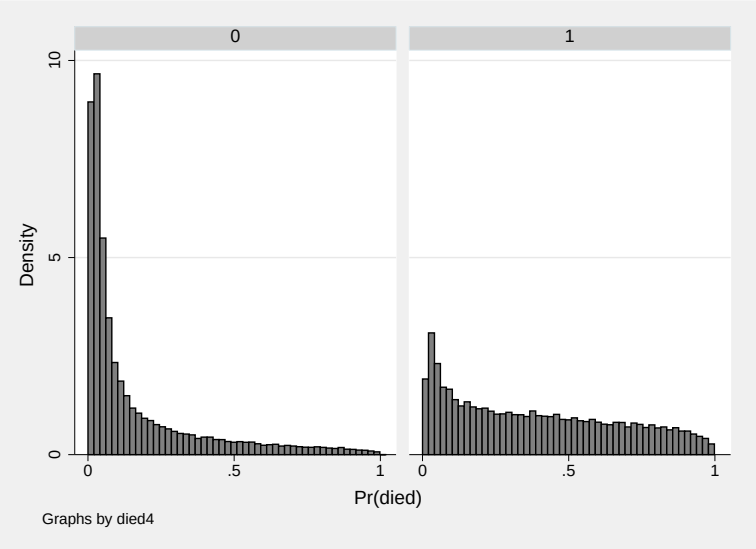


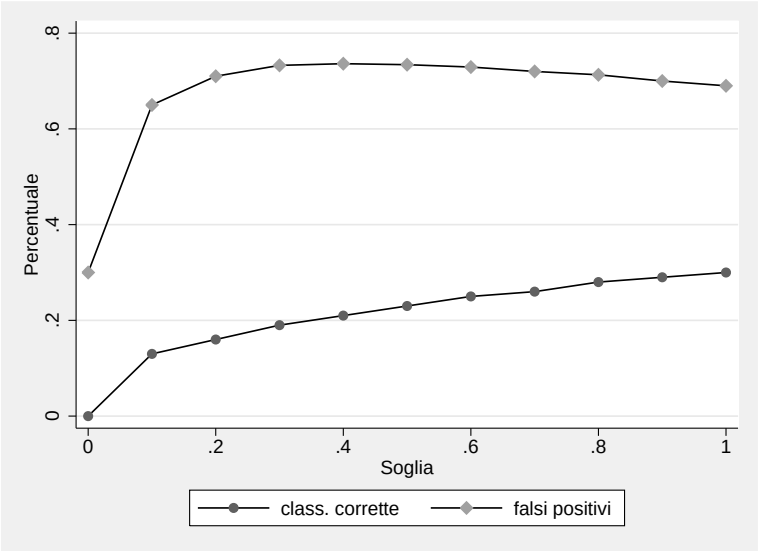
Tabella 5.3: Capacità classificatoria del modello.

Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	17947	16560	34507
0=ancora attiva (−)	9342	46848	56190
Falsi Positivi	$Pr(D\|-)$		16.62%
Falsi Negativi	$Pr(\neq D\|+)$		48.00%
Classificazioni corrette			71.44%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	14965	11919	26884
0=ancora attiva (−)	12324	51489	63813
Falsi Positivi	$Pr(D\|-)$		19.31%
Falsi Negativi	$Pr(\neq D\|+)$		44.33%
Classificazioni corrette			73.27%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	12182	8813	20995
0=ancora attiva (−)	15107	54595	69702
Falsi Positivi	$Pr(D\|-)$		21.67%
Falsi Negativi	$Pr(\neq D\|+)$		41.97%
Classificazioni corrette			73.62%

Figura 5.2: Sensibilità della classificazione al variare della soglia.



Rispetto al modello senza eterogeneità non osservata, si possono evidenziare trascurabili miglioramenti della percentuale di aziende classificate correttamente, considerando soglie pari a 0.2 e 0.3. Scegliendo la soglia 0.4 si hanno lievissimi peggioramenti delle prestazioni nel modello con eterogeneità non osservata.

Utilizzando le soglie 0.2 e 0.3, il modello con eterogeneità osservata produce una percentuale di falsi positivi rispettivamente del 16.62% e 19.31%, mentre le stesse percentuali, per il modello che non considera eterogeneità non osservata, valgono 10,85% e 14,00%. L'effetto contrario avviene per le

percentuali dei falsi negativi: in questo caso, il modello con eterogeneità osservata ha prestazioni migliori, comunque sempre nell'ordine di un punto percentuale.
Per testare la capacità previsiva, si è stimato il modello utilizzando un campione casuale semplice pari a metà della popolazione e, successivamente, si sono analizzate le previsioni calcolate utilizzando l'altra metà della popolazione: i risultati non si discostano significativamente da quanto illustrato nella tabella 5.3.

Capitolo 6

Un approccio bayesiano

6.1 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 distribuzione discreta o normale multivariata [Skrondal, A. e Rabe-Hesketh, S. ,2004].
Le osservazioni contenute nel dataset in formato *episode-splitting* possono essere viste attraverso l'ottica di un modello a due livelli:

Tabella 6.1: Struttura multilivello.

livello 1	livello 2
azienda 1	anno 1
	anno 2
	anno 3
	anno 4
	anno 5
azienda 2	anno 1
	anno 2
	anno 3

Per un modello con due livelli, assumendo quanto detto nel §5.2.1 la verosimiglianza è data dalla 6.1.

$$\prod_i \int \left\{ \prod_t f(y_{it}|\boldsymbol{x}_{it}, \nu_i) \right\} g(\nu_i) \mathrm{d}\nu_i \tag{6.1}$$

dove $f(y_{it}|\boldsymbol{x}_{it}, \nu_i)$ è la densità della variabile risposta condizionata alle variabili esplicative e alla variabile latente; $g(\nu_i) \sim N(0, \sigma^2)$ è la densità a priori della variabile latente.
Il contributo alla verosimglianza per il soggetto i è:

$$L_i(\beta, \sigma) = \int_{-\infty}^{\infty} \underbrace{\frac{e^{-\nu_i^2}/2\sigma_\nu^2}{\sqrt{2\pi}\sigma_\nu} \left\{ \prod_t f(y_{it}|\nu_i, \boldsymbol{\beta}) \right\}}_{\propto \text{ a posteriori di } \nu_i} \mathrm{d}\nu_i \tag{6.2}$$

Si può notare come la 6.2 coincida con la 5.11.
In un'ottica bayesiana empirica è possibile ottenere la media e la varianza a posteriori della variabile latente ν [Skrondal, A., Rabe-Hesketh, S., Pickles , 2004a].

$$\tilde{\nu_i} = E[\nu_i|y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\beta}}, \hat{\sigma}] = \frac{\int \nu_i \phi(\nu_i; 0, \hat{\sigma}) \prod_t f(y_{ti}|\nu_i; \hat{\boldsymbol{\beta}}) \mathrm{d}\nu_i}{L_j(\hat{\boldsymbol{\beta}}, \hat{\sigma})} \tag{6.3}$$

$$\tilde{\tau_i}^2 = var[\nu_i|y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\beta}}, \hat{\sigma}] = \frac{\int \nu_i^2 \phi(\nu_i; 0, \hat{\sigma}) \prod_t f(y_{ti}|\nu_i; \hat{\boldsymbol{\beta}}) \mathrm{d}\nu_i}{L_j(\hat{\boldsymbol{\beta}}, \hat{\sigma})} - \tilde{\nu_i}^2 \tag{6.4}$$

Per calcolare $Pr(y_{it} \neq 1)$, utilizzando metodi bayesiani empirici non si può inserire $\tilde{\nu_i}$ nell'equazione $Pr(y_{it} \neq 1|\boldsymbol{x}_i, \nu_i) = \frac{exp(\boldsymbol{x}_{it}\boldsymbol{\beta}+\nu_i)}{1+exp(\boldsymbol{x}_{it}\boldsymbol{\beta}+\nu_i)}$, ma è necessario integrare la funzione non lineare rispetto alla distribuzione a posteriori della variabile latente [Skrondal, A. e Rabe-Hesketh, S. ,2004].

6.2 Stime

Le stime dei modelli GLLAMM, si possono ottenere utilizzando la funzione esterna `gllamm`, sviluppata per il software STATA. Utilizzando questo pacchetto, specificando le ipotesi che permettono di trattare un modello logit con eterogeneità non osservata, si ottengono le medesime stime che produce il comando `xtlogit`, la routine "interna" di STATA utilizzata per trattare

lo stesso tipo di modello.
Una grossa differenza tra `gllamm` e `xtlogit` consiste nella velocità con cui viene stimato il modello: utilizzando la stessa macchina, `xtlogit` produce gli stessi risultati di `gllamm` in un tempo dieci volte inferiore[1].

Le previsioni, tenendo conto dell'eterogeneità non osservata, possono essere ottenute utilizzando la funzione esterna `gllapred`, che permette inoltre di ottenere le stime, ottenute con metodi bayesiani empirici, della media e della deviazione standard della variabile latente.
Di seguito vengono riportate le analisi delle probabilità di cessazione stimate utilizzando la funzione `gllapred`, con i coefficienti del modello con eterogeneità non osservata, descritto nel capitolo precedente.
Dalla figura 6.1 si evidenza chiaramente la capacità del modello di classificare correttamente le aziende che non sono sopravvissute. I modelli esposti nei precedenti capitoli riuscivano a classificare abbastanza correttamente le aziende sopravvissute, ma la classificazione delle aziende che cessano l'attività non era altrettanto precisa.

[1]La stima del modello con tutti i regressori, con una macchina equipaggiata con processore Pentium Celeron 1,4 Ghz e 512 Mb di RAM, richiede circa 20 ore utilizzando `xtlogit`, mentre sono necessari circa 11 giorni per stimare lo stesso modello utilizzando `gllamm`

Figura 6.1: Probabilità stimate di cessazione.

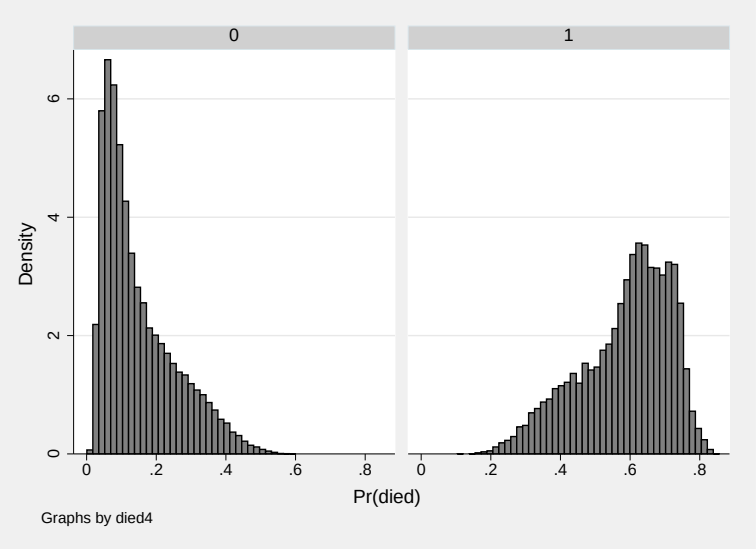


Tabella 6.2: Capacità classificatoria del modello.

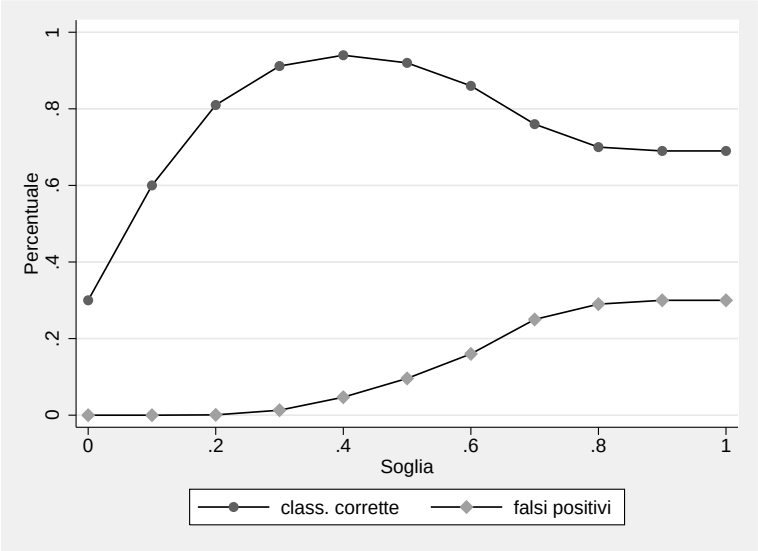
Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	27249	17123	44372
0=ancora attiva (−)	40	46285	46235
Falsi Positivi	$Pr(D\|-)$		0.086%
Falsi Negativi	$Pr(\neq D\|+)$		38.58%
Classificazioni corrette			81.00%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	26537	7255	33792
0=ancora attiva (−)	752	56153	56905
Falsi Positivi	$Pr(D\|-)$		1.32%
Falsi Negativi	$Pr(\neq D\|+)$		21.47%
Classificazioni corrette			91.17%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	24203	1792	25995
0=ancora attiva (−)	3086	61616	64702
Falsi Positivi	$Pr(D\|-)$		4.77%
Falsi Negativi	$Pr(\neq D\|+)$		6.89%
Classificazioni corrette			94.62%

Dalla figura 6.2 si evidenziano chiaramente le soglie che forniscono le migliori prestazioni del modello: 0.3 e 0.4. La soglia 0.3 può essere scelta per minimizzare la probabilità di finanziare aziende che chiuderanno l'attività entro 5 anni, pari 0.0132. La soglia 0.4 fornisce una percentuale elevata (94.62%) di aziende classificate correttamente, anche se in questo caso aumenta la probabilità di finanziare aziende che chiuderanno; ma diminuisce

Figura 6.2: Sensibilità della classificazione al variare della soglia.

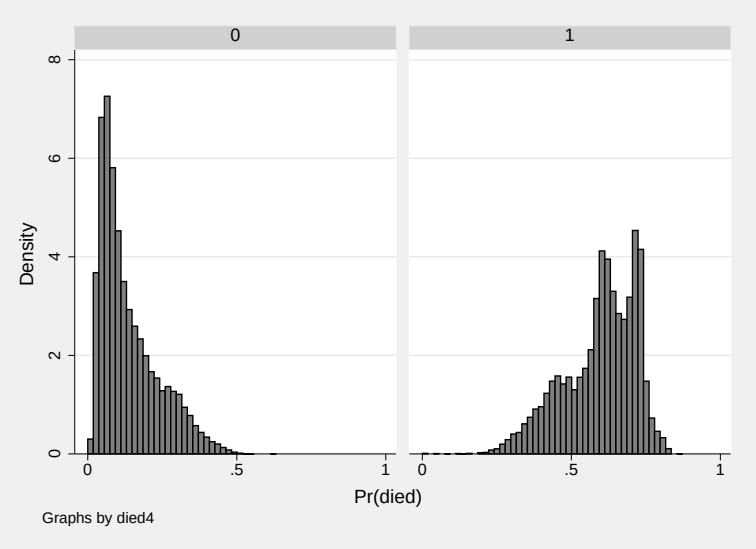


drasticamente, rispetto alla soglia 0.3, la probabilità di non finanziare aziende che continueranno l'attività.

6.3 Capacità previsiva del modello

Anche per questo modello si è testata la capacità previsiva, seguendo lo stesso schema di analisi proposto nei capitoli precedenti.
La figura 6.3 riporta le probabilità di cessazione stimate per le imprese appartenenti alla metà della popolazione non utilizzata per stimare il modello. I risultati sono confortanti poiché la distribuzione rappresentata in figura 6.3 ha un'andamento simile a quella di figura 6.1

Figura 6.3: Probabilità stimate di cessazione.



La tabella 6.3 riporta le analisi della capacità classificatoria del modello utilizzato per testare la capacità previsiva: si ottengono risultati leggermente migliori rispetto al modello che utilizza tutta l'informazione, ma questo è dovuto unicamente alla variabilità campionaria.

Tabella 6.3: Capacità classificatoria del modello.

Soglia=0.2			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	13599	7343	44372
0=ancora attiva (−)	18	24106	24124
Falsi Positivi	$Pr(D\|-)$		0.074%
Falsi Negativi	$Pr(\neq D\|+)$		16.54%
Classificazioni corrette			83.66%

Soglia=0.3			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	13406	2796	16202
0=ancora attiva (−)	211	28653	28864
Falsi Positivi	$Pr(D\|-)$		0.731%
Falsi Negativi	$Pr(\neq D\|+)$		17.25%
Classificazioni corrette			93.32%

Soglia=0.4			
Condizione prevista	Condizione effettiva		
	1=cessata (D)	0=ancora attiva ($\neq D$)	Totale
1=cessata (+)	1609	11403	13012
0=ancora attiva (−)	165	31889	32054
Falsi Positivi	$Pr(D\|-)$		3.42%
Falsi Negativi	$Pr(\neq D\|+)$		3.79%
Classificazioni corrette			96.46%

Conclusioni

I risultati, ottenuti da questa tesi, suggeriscono che l'integrazione dei dati di sopravvivenza, provenienti dagli archivi CCIIA, con le informazioni socio-economiche di provenienza ISTAT, aiutano a comprendere meglio la struttura produttiva del comparto agricolo Veneto.

In tutte le analisi che sono state condotte, è emerso l'effetto della dimensione economica sulla probabilità di sopravvivenza di un'azienda: le aziende più piccole sono molto più a rischio di cessare l'attività rispetto a quelle più grandi. Gli effetti di altre variabili, anche se significative, non risultano essere altrettanto "forti".

Rispetto al modello che utilizza come regressori solo le variabili che descrivono la dimensione economica, la specializzazione colturale e la durata pregressa dell'azienda, un modello che utilizza tutti i regressori disponibili fornisce dei tenui miglioramenti nella capacità classificatoria, al massimo del 2%.

Analizzando la probabilità di sopravvivenza annua, non si notano significativi miglioramenti, nella capacità classificatoria, rispetto al modello che fa riferimento ad un unico episodio per tutto il periodo di riferimento.

I test condotti suggeriscono la presenza di eterogeneità non osservata, e questo provoca un'aumento, in modulo, del valore dei coefficienti. Tuttavia, la capacità classificatoria del modello non subisce miglioramenti, se non si considera nella previsione il parametro che descrive l'eterogeneità non osservata.

L'approccio bayesiano empirico, utilizzato per "inserire" l'eterogeneità non osservata nelle previsioni, ha prodotto ottimi risultati: la percentuale di aziende classificate correttamente è aumentata da 73% a 94%.

Uno degli obbiettivi che ci si era posti riguardava la diminuzione della probabilità di finanziare aziende che chiuderanno entro 5 anni: si è passati dal 23%, che si ottiene se le aziende sono ritenute egualmente meritevoli, allo 0.074%, utilizzando il modello con eterogeneità non osservata con soglia di classificazione pari a 0.2.

Bibliografia

[Allison, 1984] *Event History Analysis*, Sage.

[Bassi et. al, 2006] Bassi F., Chillemi 0., Paggiaro A. (2006), *La vitalità delle aziende agricole Venete: situazione attuale e prospettive*, Atti del convegno "Le statistiche agricole verso il Censimento del 2010: valutazioni e prospettive", Cassino 26-27 Ottobre 2006.

[Cramer, 2007] Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-specified Disturbances, *Oxford Bulletin of Economics and Statistics*, OnlineEarly Articles.

[Davidov et al., 2001] Referent sampling, family history and relative risk: the role of length-biased sampling, *Biostatistics*, 2, 173-181.

[Fabbris, 1997] *Statistica Multivariata: analisi esplorativa dei dati*, McGraw-Hill.

[Frascarelli, 2004] *La riforma della Pac*, Edagricole.

[Gill et.al, 1988] Large sample theory of emprical distributions in biased sampling models, *Annals of Statistics*, 16, 1069-1112.

[Gutierrez et.al, 2002] On boundary-value likelihood-ratio tests, *Stata Technical Bulletin*, 60, 15-18.

[Istat, 2000a] Piano generale del V Censimento dell'Agricoltura.

[Istat, 2000b] Caratteristiche tipologiche delle aziende agricole.

[Istat, 2003] *Annuario Statistico Italiano 2003.*

[Infocamere, 2002] *Comunicato Stampa Indagine Movimprese*, 3 trimestre 2002.

[Jenkins, 1995] Easy ways to estimate discrete time duration models, *Oxford Bullettin of Economics and Statistics*, 57, 129-138.

[Jenkins, 2004] Survival Analysis. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester,UK.

[Kiefer, 1984] Economic duration data and hazard functions, *Journal of Economic Literature*, 26, 646-679.

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

[Skrondal, A., Rabe-Hesketh, S., Pickles , 2004a] *GLLAMM Manual. U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 160. `htpp://www.bepress.com/ubbbiostat/paper160`

[Solinas, 2005] *I processi di formazione, la crescita e la sopravvivenza delle piccole imprese*, Franco Angeli.

[Stata, 2005] *Stata Longitudinal/Panel Data Reference Manual*

[Vardi, 1985] Nonparametric estimation in the presence of length bias, *Annals Statistic*, 10, 616-620.

[UE, 2005a] *La politica agricola comune alla portata di tutti* `http://europa.eu.int/comm/agriculture/index_it.htm`.

[UE, 2005b] *L'agricoltura nell'Unione Europea - Informazioni statistiche ed economiche* `http://ec.europea.eu/agriculture/agrista/index_it.htm`.

[Wooldridge, 2002] *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge MA.