UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA

IL POSSESSO DELLE ATTIVITÀ FINANZIARIE DEGLI ITALIANI. ALCUNE ANALISI

Relatore: Prof. OMAR PACCAGNELLA

Laureando: ALESSANDRO DA BOIT

Matricola: 599052-GEI

INDICE

1- INTRODUZIONE

La Banca d'Italia effettua ogni due anni un'indagine sui bilanci delle famiglie italiane, nella quale si raccolgono molteplici informazioni di carattere socio-demografico, economico e culturale. In questa trattazione l'attenzione è rivolta alle attività finanziarie possedute dagli Italiani e all'ammontare degli investimenti destinati ad esse. L'obiettivo che si prefigge questo lavoro è ricercare delle relazioni significative tra le variabili oggetto d'esame, cercando di indagare come, definendo determinate caratteristiche possedute dal capofamiglia o dalla famiglia stessa, queste possano aiutarci nell'individuare potenziali investitori per ciascuna attività finanziaria. In un secondo momento, condizionandoci ai risultati positivi, in termini di possesso di un determinato investimento, studiare anche come l'ammontare relativo di queste risorse sia influenzato dalle medesime variabili.

Le aziende, lo Stato, le banche e altri enti che trattano le attività finanziarie degli investitori avrebbero così a disposizione uno strumento che potrebbe essere in grado di orientarne le scelte strategiche, definendone possibili applicazioni in situazioni reali. Le conclusioni ottenute dunque, dovranno dunque essere supportate dai possibili utilizzi del modello da parte dei diversi enti, per rendere questa analisi efficace e di valenza strategica. Ciò permetterebbe di effettuare interventi mirati a specifiche categorie di clienti attuali e potenziali, per esempio in termini di fidelizzazione, corretta informazione e offerte diversificate per clientele effettivamente differenti (*Grandinetti, 2008*).

Di seguito verranno proposte delle analisi descrittive del fenomeno oggetto d'indagine, allo scopo di sintetizzare alcune informazioni contenute nel data set ed evidenziare le relazioni che sussistono tra le variabili prese in considerazione. Le analisi descrittive costituiscono comunque solo una base per avere un'idea generale su cui lavorare, ma sono utili in quanto forniscono delle linee guida e aggiungono informazioni rilevanti (per esempio riguardo alla distribuzione dei redditi e alla conseguente capacità di spesa).

Le relazioni tra le variabili verranno poi sintetizzate in fase di stima di modelli di regressione multivariati, i quali permettono l'analisi congiunta di tutti i fattori in

gioco; i modelli stimati dovranno poi essere validati in riferimento alla loro bontà con ulteriori verifiche delle ipotesi sottostanti i modelli tramite test statistici e analisi grafiche. Tutte le elaborazioni numeriche sono state eseguite con il software R.

2- DEFINIZIONE DEL DATASET

Nel periodo compreso tra marzo e ottobre del 2007 sono stati somministrati i questionari relativi all'indagine campionaria sui bilanci delle famiglie italiane del 2006. Lo schema di campionamento è lo stesso delle precedenti indagini, e prevede una procedura di selezione a due stadi (*Report dell'indagine, 2006*): le unità di primo stadio sono i comuni e le unità di secondo stadio le famiglie. Le unità di primo stadio sono inoltre stratificate in base alla regione di residenza e alla classe di ampiezza demografica.

Il data set di riferimento è relativo all'indagine sui bilanci delle famiglie italiane effettuata dalla Banca d'Italia nel 2006: il campione è composto da 7768 famiglie, ma è stato ridotto a 5018 in funzione degli scopi di questa trattazione. I dati infatti, riguardano esclusivamente gli ultracinquantenni, e la ricerca si propone di scovare delle relazioni significative tra le varie forme di investimento di quest'ultimi e una serie di variabili, perlopiù di carattere socio-demografico. La scelta di selezionare solamente gli over 50 si giustifica in un'ottica che li vede come i detentori di maggiori disponibilità finanziarie e di una situazione familiare più stabile.

I valori assunti dai caratteri socio-demografici sono riferiti al capofamiglia, operazione necessaria in quanto i dati relativi alle forme di risparmio sono aggregati per nucleo familiare; è ragionevole ipotizzare il capofamiglia come il maggiore percettore di reddito all'interno della famiglia stessa, anche se questa è solo una delle soluzioni percorribili. Nell'indagine i questionari somministrati al campione selezionato fanno proprio riferimento al capofamiglia, destinatario della sua compilazione.

Il questionario utilizzato nella rilevazione è predisposto seguendo una struttura modulare. Si compone di una parte di base, in cui sono rilevati i fenomeni che interessano tutte le famiglie, e di diversi allegati che riguardano soltanto alcuni sottoinsiemi di famiglie. In accordo con le teorie dell'analisi di mercato per la predisposizione di un questionario (*Bassi, 2008*), la rilevazione si è dotata di un sistema di imputazione di dati mancanti (comunque di entità modesta), che si è resa necessaria per consentire il calcolo di tutte le elaborazioni statistiche. Una ulteriore

considerazione va fatta in merito alla qualità dei dati disponibili, perché un aspetto che può influire su di essa riguarda la reticenza delle famiglie a dichiarare le proprie fonti di reddito. E' plausibile quindi che talvolta la risposta sia distorta, specialmente in riferimento ad argomenti delicati quali la ricchezza e il reddito (e in particolar modo per i lavoratori indipendenti). Per questo motivo è stato chiesto agli intervistatori un presunto giudizio sull'attendibilità delle risposte, il quale, affiancato da informazioni riguardanti le dichiarazioni al fisco, ha permesso di individuare una sistematica sottostima della ricchezza di natura finanziaria e dei redditi da interessi e dividendi (*www.bancaditalia.it*).

Le variabili relative alle attività finanziarie sono disponibili sia quantitativamente, quindi come ammontare di risorse finanziarie destinate all'investimento, sia in forma binaria, quindi sotto forma di possesso/non possesso di risorse finanziarie da investire per ciascuna tipologia di investimento. Per ragioni di praticità e semplicità di interpretazione dei risultati, queste variabili sono state accorpate in 5 principali categorie, rispettivamente: depositi bancari/postali a c/c o risparmio, titoli di Stato (BOT, CCT, BTP, CTZ e altri), fondi comuni e obbligazioni, azioni e quote di società (quotate in borsa e non, a responsabilità limitata e di persone) e gestioni patrimoniali. Le diverse tipologie si differenziano soprattutto, dal punto di vista degli investitori, in relazione al tasso di interesse che garantiscono e al loro relativo rischio (*Cerbioni, 2006*).

Scegliere la forma migliore per impiegare il proprio risparmio non è semplice; la selezione deve essere fatta tra varie possibilità, e per avvicinarsi alla scelta migliore si necessita di una disamina oggettiva delle caratteristiche di durata, di liquidità, di reddito, di rischio e fiscali di ciascuna soluzione (*www.unioneconsulenti.it*).

Le variabili ritenute idonee per spiegare le diverse categorie di investimenti sopra citati sono, come già accennato in precedenza, prevalentemente di ambito socio-demografico. Si terrà quindi conto di: sesso, età, titolo di studio, status del lavoratore, settore di attività e dell'area geografica di residenza. Inoltre si ritiene plausibile che sia il reddito disponibile netto che la propensione al rischio dell'investimento possano

essere variabili influenti sulla decisione o meno di destinare risorse finanziarie in un qualche tipo di forma.

Una gran parte di queste variabili è di carattere qualitativo e risulta dunque necessario esplicare le diverse categorie che compongono questo tipo di variabili:

- Status del lavoratore: 1.operaio 2.impiegato 3.dirigente/direttivo 4.imprenditore/libero professionista 5.altro autonomo 6.pensionato 7.non occupato.
- Settore di attività: 1.agricoltura 2.industria 3.servizi pubblici 4.altri settori 5.nessun settore.
- Titolo di studio: 0.nessuno 1.elementare 2.media 3.diploma 4.laurea
- Area geografica: 1.nord 2.centro 3.sud e isole

La propensione al rischio è suddivisa in 4 categorie, la prima identifica i capifamiglia con alta propensione al rischio e contemporanee alte prospettive di guadagno, mentre all'interno dell'ultima categoria ci saranno gli individui con una bassa propensione al rischio e contemporanne basse prospettive di guadagno. Le rimanenti variabili, età del capofamiglia e reddito disponibile netto sono quantitative. In riferimento al reddito disponibile netto (calcolato sull'anno) è opportuno ricordare che esso è definito come la somma del reddito da lavoro dipendente, pensioni e trasferimenti netti, reddito netto da lavoro autonomo e reddito da capitale, in accordo con il piano di aggregazione delle variabili che riguarda il conto del reddito (*Cerbioni, 2006*).

Si potrebbero prendere in considerazione per l'analisi anche delle ulteriori variabili, però sarebbero risultate difficilmente trattabili in riferimento ai risultati raggiunti, o avrebbero avuto poco senso in termini di definizione di possibili applicazioni del modello di sintesi della trattazione; per questi motivi sono state escluse anche se presenti nel data set originario reso disponibile dalla Banca d'Italia.

3- ANALISI DESCRITTIVE DEL FENOMENO

3.1- Distribuzione dei redditi

Innanzitutto si vuole analizzare la distribuzione dei redditi degli over 50, in particolare fornire una misura della concentrazione della ricchezza tra gli individui, tramite l'indice di concentrazione di Gini, distinguendo in prima istanza per area geografica e poi per status del lavoratore. L'obiettivo è cercare di evidenziare dei differenziali di reddito, e quindi intuitivamente di disponibilità a investire, in riferimento a questi due diversi aspetti.

L'ordinamento dei redditi ci permette il calcolo dell'indice di Gini sull'intero campione di capofamiglia ultracinquantenni, che presenta un valore di 0.346 (vedi **Figura 1**). Il risultato è molto simile al valore di 0.349 che è scaturito dall'intera indagine sulle famiglie effettuata dalla Banca d'Italia (*www.bancaditalia.it*). Non vi sono dunque differenze rilevanti in termini di concentrazione se ci condizioniamo al data set delle famiglie con capofamiglia ultracinquantenne.

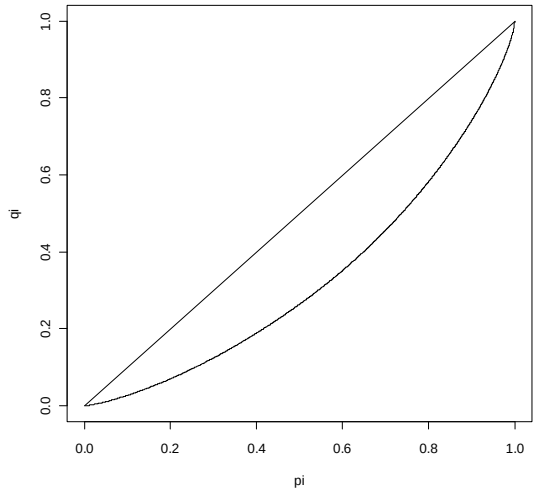

Figura 1, Curva di Lorenz riferita all'indice di concentrazione di Gini calcolato sul campione di interesse

Analizziamo l'eventuale emergere di differenze nella misura della concentrazione al variare dell'area geografica, quindi per le tre categorie nord (0.334), centro (0.333) e sud/isole (0.340). Da questi risultati si può notare come non vi siano differenze rilevanti tra le diverse aree geografiche, l'entità delle differenze è infatti piuttosto ridotta.

Analizzando invece le differenze in media rispetto alle tre categorie notiamo come il reddito medio al nord (34343.26 €) sia notevolmente superiore rispetto a quello del sud (23986.85 €); questo a testimonianza della diversa entità di entrate mensili all'interno del nucleo familiare al variare dell'area geografica di riferimento. In centro Italia il reddito medio è invece di 33490.16 €, quindi un valore molto vicino a quello del nord, e una conseguente capacità di spesa simile. Ciò è inoltre confermato dai valori di alcuni indicatori economici di povertà relativa, come per esempio la percentuale di invidui al di sotto della soglia in base al reddito equivalente e al consumo equivalente (*www.bancaditalia.it*).

Eseguendo un t test (*Azzalini, 2001*) è confermata la significativa differenza in media tra il reddito del nord e quello del sud, in ipotesi di uguaglianza di varianze delle due popolazioni sottoposte al test (t = 14.4543, df = 3928, p-value < 2.2e-16). Il p-value infatti ha un valore estremamente basso.

Si ritiene ragionevole analizzare inoltre eventuali differenze di reddito disponibile netto per le diverse occupazioni delle unità del campione, quindi rispetto alla variabile status del lavoratore. Tralasciando per il momento pensionati e disoccupati, concentriamo l'attenzione sui differenziali di reddito per tipo di occupazione. Come potevamo aspettarci, le famiglie con capofamiglia operaio hanno il reddito medio annuo più basso (30357.77 €); per i capofamiglia impiegati invece le entrate annuali corrispondono a 40527.09 €. E' invece molto interessante notare come il reddito medio sia superiore per i dirigenti (67340.92 €) rispetto ai liberi professionisti/imprenditori (60024.32 €). Si rammenti anche però che fenomeni di sottostima dei redditi sono ovviamente più frequenti nel caso si tratti di liberi professionisti; infatti accade che emergano dati discordanti tra rilevazioni sulle famiglie e rilevazioni fiscali (*www.bancaditalia.it*).

Anche in questo caso le differenze in media risultano significative e ciò è confermato dai t test eseguiti a due a due sui diversi sottocampioni. Soltanto per quanto concerne la differenza tra redditi medi dei dirigenti e dei liberi professionisti/imprenditori viene accettata l'ipotesi nulla di uguaglianza delle medie (t = 1.2953, df = 258, p-value = 0.1964), a conferma dei dubbi sorti in precedenza sulle maggiori disponibilità

finanziarie, come entrate all'anno, dei dirigenti rispetto ai liberi professionisti/imprenditori.

La diversa dinamica dei redditi per condizione professionale ha avuto anch'essa un impatto sulla povertà relativa degli individui; facendo riferimento infatti alle rilevazioni degli anni precedenti dell'indagine sui bilanci delle famiglie italiane, la quota di lavoratori dipendenti (categorie 1,2 e 3) sotto la soglia di povertà è aumentata di 0.4 punti percentuali dal 2000, attestandosi su un livello del 6.3% nel 2006 (*www.bancaditalia.it*).

In conclusione, per quanto riguarda il reddito disponibile netto (valore medio 30814.85 €) per tutto il campione di interesse abbiamo potuto constatare come esso si distribuisca non uniformemante rispetto ad alcune variabili categoriali; esso costituisce un punto cardine di questa trattazione in quanto le famiglie, per poter destinare risorse finanziarie a varie forme di investimento, devono innanzitutto possederle.

3.2- Diffusione delle attività finanziarie

Nel 2006 l'87.8% delle famiglie con capofamiglia oltre i 50 anni possedeva un deposito bancario/postale a c/c o a risparmio, il 10.8% titoli di stato, l'11.9% fondi comuni e obbligazioni, il 6.2% azioni e quote di società e solo l'1.4% gestioni patrimoniali (vedi **Figura 2**).

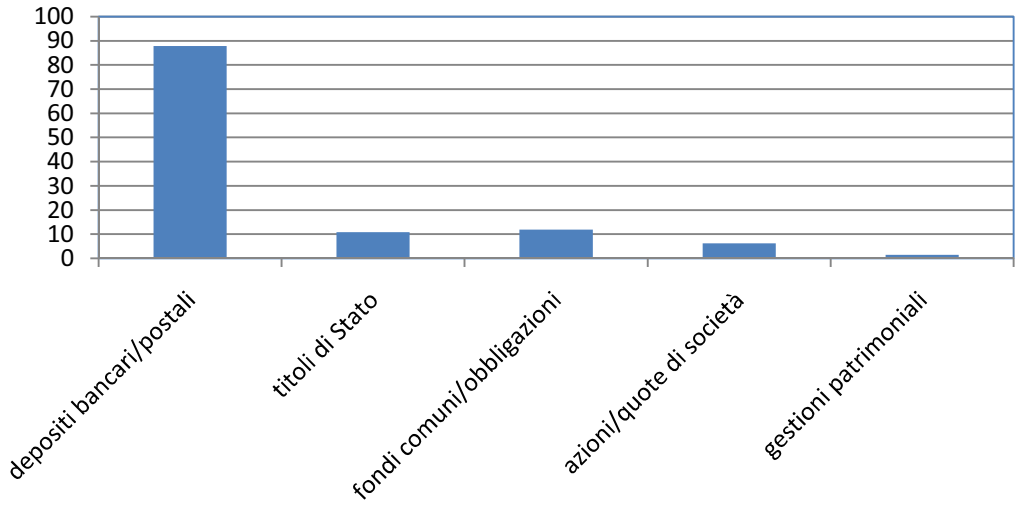

Figura 2, diffusione delle attività finanziarie

Passiamo ora ad analizzare come la diffusione delle attività finanziarie possa presentare delle connessioni con alcune delle variabili descritte precedentemente. Sarebbe infatti logico aspettarsi infatti, una relazione crescente della diffusione di queste attività al crescere per esempio del reddito e del titolo di studio conseguito.

Si possono inoltre verificare eventuali differenze in termini di utilizzo di strumenti finanziari al variare dell'area geografica e dello status del lavoratore.

Calcolando innanzitutto il reddito disponibile netto medio delle famiglie condizionato al possesso/non possesso di una determinata attività finanziaria, possiamo dare delle delucidazioni che spieghino come i capofamiglia con a disposizione un reddito più elevato, siano poi anche quelli che hanno investito più risorse attraverso i vari strumenti finanziari. Si nota appunto come nessuna forma di investimento faccia eccezione: agli investitori di ogni attività finanziaria corrispondono livelli di reddito più elevati (vedi **Tabella 1**).

Forma di investimento	**Possiede**	**Non possiede**
depositi bancari/postali	33069.19 €	14645.29 €
titoli di Stato	42261.97 €	29431.58 €
fondi comuni e obbligazioni	51735.12 €	28000.57 €
azioni/quote di società	55089.08 €	29211.01 €
gestioni patrimoniali	67509.93 €	30280.67 €

Tabella 1, reddito disponibile netto medio per possesso di attività finanziarie

La diffusione di tutte le attività finanziarie è dunque crescente al crescere del reddito disponibile netto.

Evidenziamo ora il secondo aspetto della questione, cioè si vuole verificare se sussiste una relazione di fondo tra la diffusione degli strumenti finanziari presso le famiglie e il titolo di studio posseduto dal capofamiglia. I risultati sono sintetizzati, codificando, per ogni attività finanziaria, la rispettiva percentuale di investitori per titolo di studio conseguito (vedi **Tabella 2**); si ricordi che la variabile studio è categoriale e assume 5 diverse modalità, rispettivamente 0 per chi non è in possesso

di alcun titolo di studio, 1 per la licenza elementare, 2 per la licenza media, 3 per il conseguimento del diploma e infine 4 per il conseguimento della laurea.

Titolo di studio	Depositi bancari/postali	Titoli di Stato	Fondi comuni e obbligazioni	Azioni/quote di società	Gestioni patrimoniali
nessuno	63.5%	2.4%	0.5%	1%	0%
elementare	81.9%	6.5%	5.6%	1.4%	0.4%
media	93.9%	11.5%	13.6%	6%	1.3%
diploma	96.7%	17.7%	19.2%	12.5%	2.8%
laurea	99.2%	19.4%	29.5%	18.5%	5.1%

Tabella 2, percentuale di investitori per titolo di studio

Si può notare, con grande evidenza, come il possesso da parte del capofamiglia di un titolo di studio sempre più elevato accresca notevolmente la percentuale di individui disposti ad investire in una qualche forma. Senza alcuna eccezione dunque, i capofamiglia che hanno conseguito un titolo di studio rilevante sono anche quelli che più sono disposti a investire le proprie risorse, forse perché possiedono le competenze e le conoscenze necessarie. Una particolare nota va sottolineata per i laureati, che in riferimento alle attività finanziarie più rischiose (fondi comuni e obbligazioni, azioni e quote di società e gestioni patrimoniali), evidenziano i "salti" maggiori: infatti il 29.5%, il 18.5% e il 5.1% dei laureati investe rispettivamente nelle tre categorie sopra citate contro delle quote assai inferiori per i possessori di un diploma (19.2%, 12.5% e 2.8%). Questo a testimonianza del fatto che per effettuare degli investimenti così detti rischiosi sono necessarie delle competenze più complesse e articolate dell'ambito economico-finanziario.

Si rammenti che fondi comuni e obbligazioni sono da ritenersi comunque meno rischiosi delle azioni e delle gestioni patrimoniali, sono però "simili" se si pensa alle competenze necessarie da possedere per poter trattare attività finanziarie di quella tipologia.

Si tratta ora di indagare l'analoga composizione degli investitori in riferimento però ora, non più alla variabile titolo di studio, ma allo status del lavoratore (vedi **Tabella 3**).

Status del lavoratore	Depositi bancari/postali	Titoli di Stato	Fondi comuni e obbligazioni	Azioni/quote di società	Gestioni patrimoniali
operaio	88.1%	3.4%	6.4%	3.7%	0%
impiegato	96.8%	13.2%	16.6%	9.7%	1.6%
dirigente	100%	17.9%	32.1%	19.3%	6.4%
imprenditore/libero prof.	99.2%	15%	29.2%	20.8%	4.2%
altro autonomo	94.8%	8%	14.4%	6.9%	2.9%
pensionato	86.7%	11.2%	10.9%	5.2%	1.2%
non occupato	78.1%	8.1%	6.8%	3.9%	1%

Tabella 3, percentuale di investitori per status del lavoratore

Le famiglie con capofamiglia dirigente o imprenditore/libero professionista sono anche quelle all'interno delle quali la diffusione delle attività finanziarie è nettamente più elevata rispetto alle altre categorie. Si noti come addirittura sia i pensionati che i non occupati investano in misura maggiore rispetto agli operai, fatta eccezione per i depositi bancari/postali. Anche a riguardo delle forme di investimento più rischiose le differenze che si presentano sono più marcate: la percentuale di dirigenti e imprenditori/liberi professionisti che investono in fondi comuni e obbligazioni è quasi il doppio di quella degli impiegati. Lo stesso dicasi per le azioni e le quote di società dove addirittura la percentuale di dirigenti e liberi professionisti/imprenditori è più del doppio di quella degli impiegati.

Riguardo al possesso di depositi bancari e postali a c/c o a risparmio si delineano differenze meno nette al variare delle diverse categorie professionali.

Passiamo ora ad analizzare, in ultima istanza la composizione in relazione all'area geografica di residenza delle famiglie (vedi **Tabella 4**).

Area geografica	Depositi bancari/postali	Titoli di Stato	Fondi comuni e obbligazioni	Azioni/quote di società	Gestioni patrimoniali
nord	95.7%	17.6%	18.7%	10%	2%
centro	90.2%	8.6%	9.7%	4.8%	1.5%
sud	74.8%	2.5%	3.5%	1.8%	0.6%

Tabella 4, percentuale di investitori per area geografica di residenza

Le famiglie residenti al nord si caratterizzano per una propensione all'investimento più marcata in relazione a tutte le diverse categorie di attività finanziarie, seguite da quelle del centro e poi da quelle del sud. Questo aspetto potrebbe essere collegato ai differenziali di reddito che in precedenza avevamo ravvisato nelle tre differenti macroaree in cui è stata suddivisa l'Italia, a supporto di ulteriori informazioni utili all'analisi.

A conclusione di queste prime analisi di tipo descrittivo si può desumere che le famiglie con capofamiglia oltre i 50 anni hanno, in generale, una scarsa propensione ad investire in attività finanziarie rischiose (per esempio azioni e quote di società e gestioni patrimoniali), mentre i depositi bancari e postali restano molto diffusi. Inoltre si sono evidenziati degli aspetti cruciali per lo svolgimento futuro della trattazione, come la relazione esistente tra reddito disponibile netto e disponibilità ad investire, oltre che differenziali di reddito per area geografica.

Infine si sono ottenute informazioni utili riguardo alla distribuzione della percentuale di risparmiatori in relazione ad alcune variabili categoriali, il che ha permesso di individuare già alcune caratteristiche dei capifamiglia connesse con la probabilità di aver investito in una qualche attività finanziaria.

4- MODELLI PER IL POSSESSO DELLE FORME DI RISPARMIO

Utilizziamo un modello di regressione multiplo nella usuale forma del tipo $Yi = \beta 1xi1 + \ldots + \beta pxip + \varepsilon i;$ dove εi è il termine d'errore e rappresenta la componente stocastica del modello. Le variabili risposte che abbiamo a disposizione sono di tipo dicotomico, cioè le variabili *Yi* assumono valori in {0;1}. Specificatamente sono variabili aleatorie di tipo Bernoulli e hanno la seguente distribuzione di probabilità: $P(Yi=y)=\pi i$ se $y=1$ e $P(Yi=y)=1-\pi i$ se $y=0$.

Si vuole modellare la media della variabile Y in funzione delle p variabili esplicative; la media di *Yi* è quindi $E(Yi)=\pi i$ e si pone $g(E(Yi))=g(\pi i)=\beta xi=\beta 1xi1+\ldots+\beta pxip$, dove g(.) è la funzione legame, nota.

Si specifica quindi una relazione lineare tra le variabili esplicative e un'opportuna trasformazione della media; le funzioni legame di uso più frequente sono la funzione logit (legame canonico), probit e cloglog. Auspicabilmente questa trasformazione garantisce che sia rispettato il campo di variazione.

Si ammette inoltre una specifica forma di eteroschedasticità, in quanto $Var(Yi)=\pi i*(1-\pi i)$ per $i=1,\ldots,n$. In sintesi, le ipotesi alla base di questi modelli sono:

(1) $Y1,\ldots,Yn$ variabili indipendenti e $Yi \sim Bernoulli(\pi i)$.

(2) $X=(x1,\ldots,xn)$ matrice non stocastica di dimensioni (nxp) e a rango pieno. Inoltre $g(\pi)=X\beta$.

In accordo con la teoria dei modelli lineari generalizzati, le equazioni di verosimiglianza per questi modelli non hanno soluzione esplicita, e andranno quindi risolte tramite metodi iterativi, in particolare l'algoritmo dei minimi quadrati pesati iterati (Newton-Raphson).

Per i risultati inferenziali ci si basa sul risultato generale di normalità asintotica dello stimatore di massima verosimiglianza (*Azzalini, 2001*).

Nei modelli logistici, l'interpretazione dei coefficienti e degli effetti marginali delle variabili sul possesso delle forme di investimento è in termini di rapporti di quote, misura che approssima il rischio relativo sotto certe condizioni: bassa quota dell'evento Y=1 e campionamento retrospettivo. Riguardo alla prima condizione, essa non è rispettata nel caso del possesso dei depositi bancari e postali (la quota

dell'evento Y=1 è dell'87.8%) mentre lo è sufficientemente per le restanti attività finanziarie esaminate (la quota più alta è quella riguardo al possesso di fondi comuni e obbligazioni ed è pari all'11.9%).

4.1- Modello per il possesso di depositi bancari e postali a conto c/c o a risparmio

I depositi bancari e postali sono caratterizzati da rendimenti molto bassi e sono ulteriormente decurtati dalla notevole incidenza della ritenuta d'imposta sugli interessi maturati. E' spesso consigliato e ritenuto opportuno depositarvi solo una parte del capitale da investire: quella necessaria per le consuete spese familiari o per la gestione dell'attività economica (*www.UnioneConsulenti.it*).

Analizziamo innanzitutto che relazione sussiste tra il possesso/non possesso di un deposito bancario/postale e alcune delle variabili esplicative presenti nel data set considerato.

Iniziamo col valutare la relazione tra la probabilità di possedere un deposito bancario/postale a c/c o a risparmio e il reddito; da una prima analisi esplorativa (boxplot) vediamo come i possessori dei depositi siano caratterizzati da un reddito più elevato (vedi **Figura 3**).

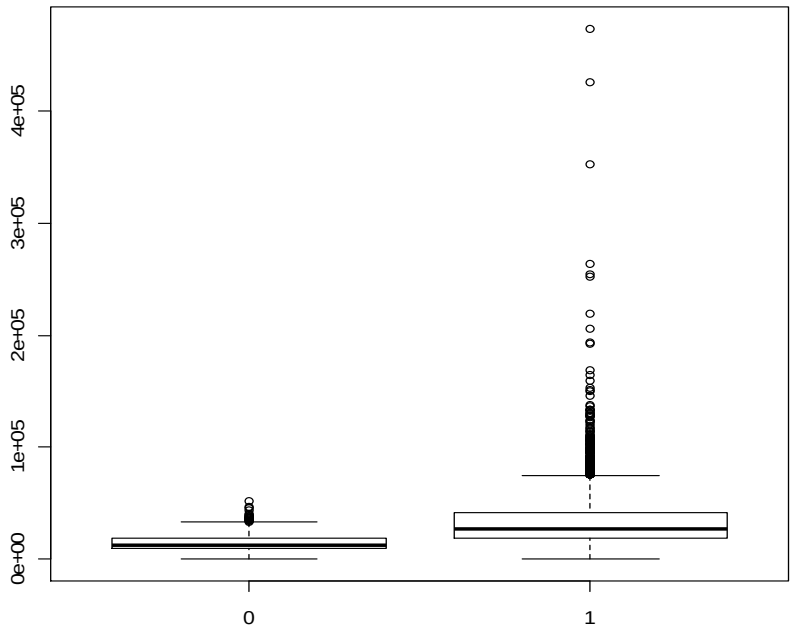

Figura 3, boxplot (possesso deposito bancario/postale, reddito)

Andiamo a stimare un primo modello con il solo reddito disponibile netto come variabile esplicativa, in modo da valutare se questa tendenza dei possessori di redditi elevati influisce realmente sulla probabilità di possedere un deposito bancario/postale. La funzione legame utilizzata è la logit (legame canonico).

mod0<-glm(formula = pdepos ~ reddito, family = binomial)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	*-8.944e-01*	*1.169e-01*	*-7.654*	*1.95e-14*	***
reddito	*1.394e-04*	*6.656e-06*	*20.944*	*< 2e-16*	***

Residual deviance: 2870.1 on 5016 degrees of freedom

Il modello stimato presenta sia l'intercetta che il parametro associato al reddito significativi ad un livello di significatività dell' 1%, il che porta quindi a rifiutare le ipotesi dei test sulla nullità dei singoli parametri.

E' da escludere una relazione quadratica tra le variabili, in quanto effettuando il test associato alla differenza tra le devianze dei due modelli annidati, il risultato ci porta ad accettare l'ipotesi nulla del modello ridotto (*Azzalini, 2001*).

Model 1: pdepos ~ reddito

Model 2: pdepos ~ poly(reddito, 2)

	Resid. Df	Resid. Dev	Df	Deviance	P(>|Chi|)
1	*5016*	*2870.1*			
2	*5015*	*2867.3*	*1*	*2.7938*	*0.09463*

E' preferibile quindi ipotizzare una relazione di tipo lineare; interpretando i parametri del modello stimato in precedenza si può concludere che il reddito ha un effetto positivo sulla probabilità di possedere un deposito bancario/postale, come ci si aspettava. Numericamente, la quota dell'evento possedere il deposito è, per esempio, 1.15 volte più alta per una famiglia che dispone di un reddito disponibile netto di 1000 € in più, a parità di altre condizioni. L'esponenziale del coefficiente, quando si trattano variabili quantitative in un modello logit, esprime la variazione percentuale della quota di Y=1 (successo in generale, possesso del deposito bancario/postale in questo caso) con un aumento unitario della variabile esplicativa (*Azzalini, 2001*).

*> exp(0.0001393989*1000)*

[1] 1.149583

Dal confronto tra il reddito e le probabilità stimate dal modello vediamo chiaramente come la probabilità cresca al crescere del reddito, avvicinandosi ad uno per redditi annui superiori a 40000 € (vedi **Figura 4**).

> predict(mod0,newdat=data.frame(reddito=40000),type="response")

0.9908207

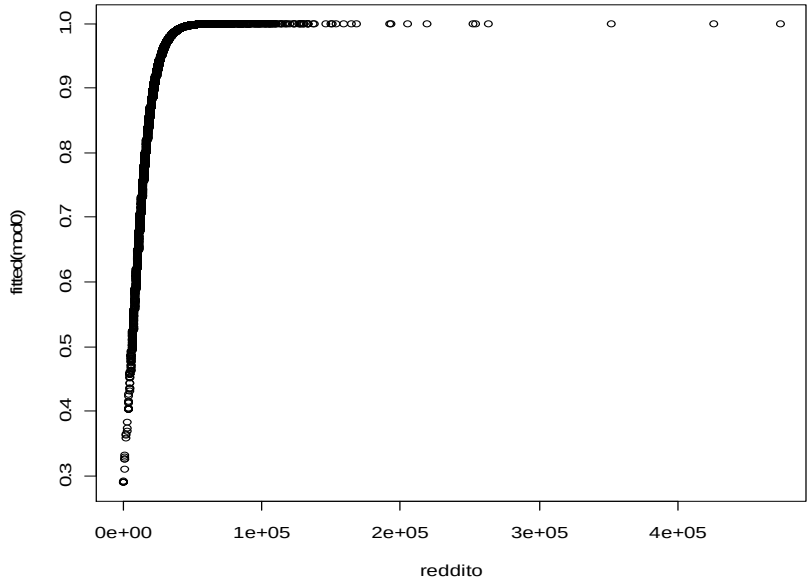

Figura 4, grafico tra reddito e valori stimati dal modello *mod0*

Effettuando il test sulla devianza residua, che ci fornisce una valutazione della bontà del modello otteniamo risultati soddisfacenti, in quanto l'ipotesi nulla di bontà del modello è ampiamente accettata (il p-value è approssimativamente 1).

Giunti a questo punto si può concludere che la probabilità di possedere un deposito bancario/postale cresce al crescere del reddito, e in particolare cresce molto velocemente ad 1 anche per redditi non eccessivamente elevati.

Né l'utilizzo di diverse funzioni legame né delle trasformazioni sulla variabile esplicativa portano ad un miglioramento del modello stesso (la devianza residua risulta più elevata spendendo gli stessi gradi di libertà).

Proviamo ora ad aggiungere ulteriori variabili esplicative al modello stimato e passare dunque ad un modello di regressione multiplo.

Introduciamo nel modello le seguenti variabili: sesso, status del lavoratore, età e area geografica di residenza. Il software R, di default, parametrizza le variabili categoriali (sesso, status del lavoratore e area geografica) rispetto al parametro d'angolo, quindi rispetto alla prima modalità assunta da ogni predittore categoriale.

La stima del modello descritto produce il seguente output:

mod1<-glm(formula = pdepos ~ statuslav + reddito + eta + area + sesso, family = binomial)

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*2.293e+00*	*4.482e-01*	*5.115*	*3.13e-07 ****
statuslav2	*9.056e-01*	*3.690e-01*	*2.454*	*0.014123 **
statuslav3	*1.309e+01*	*2.494e+02*	*0.053*	*0.958124*
statuslav4	*1.754e+00*	*1.055e+00*	*1.662*	*0.096456 .*
statuslav5	*4.757e-01*	*4.241e-01*	*1.122*	*0.262042*
statuslav6	*7.824e-01*	*2.306e-01*	*3.393*	*0.000691 ****
statuslav7	*2.338e-01*	*2.514e-01*	*0.930*	*0.352357*
reddito	*1.162e-04*	*7.095e-06*	*16.379*	*< 2e-16 ****
eta	*-3.376e-02*	*6.200e-03*	*-5.445*	*5.17e-08 ****
area2	*-9.216e-01*	*1.545e-01*	*-5.964*	*2.46e-09 ****
area3	*-1.637e+00*	*1.270e-01*	*-12.884*	*< 2e-16 ****
sesso2	*-1.746e-01*	*1.067e-01*	*-1.635*	*0.101976*

Residual deviance: 2632.1 on 5006 degrees of freedom

I risultati ottenuti confermano la significatività della variabile reddito. Anche l'altra variabile quantitativa, l'età, risulta significativa ad un elevato livello di significatività. Il valore negativo del parametro associato ad essa ci informa dell'influenza negativa del crescere dell'età in relazione al possesso del deposito bancario/postale.

Precisamente risulta circa 0.97 volte meno probabile che la famiglia possieda il deposito in seguito ad un aumento di un anno dell'età del capofamiglia, ceteris paribus.

> exp(-0.0337606962)

[1] 0.9668028

La variabile categoriale area geografica è significativa in tutte le sue componenti e ci conferma (vedi analisi descrittive) come risiedere al centro e al sud in particolare, diminuisca la probabilità di possedere il deposito, rispetto al risiedere al nord.

Quando si trattano variabili categoriali, l'esponenziale di ciascun coefficiente ci fornisce una misura del rapporto di quote tra le modalità della categoria e la modalità assunta da R come parametro d'angolo: infatti in questo caso risulta che, al netto delle altre variabili esplicative, la quota di possedere il deposito al centro è 0.4 volte inferiore rispetto al possederlo al nord; analogamente per le famiglie del sud la quota è 0.19 volte inferiore, sempre rispetto al nord (parametro d'angolo). (*Azzalini, 2001*).

La variabile status del lavoratore risulta invece poco significativa in quasi tutte le sue componenti, a conferma delle differenze meno nette riscontrate nell'analisi descrittiva (le percentuali di possessori di deposito bancario/postale al variare dello status del lavoratore sono molto più simili che per le altre forme di investimento); si può dunque ritenere questa variabile poco influente sulla probabilità di detenere il deposito.

Conclusioni simili si potrebbero protrarre anche per la variabile sesso, che però si trova in una situazione limite (il p-value associato al test sulla nullità del parametro ha un valore "a cavallo" tra la regione di rifiuto e quella di accettazione del test); sembra quindi azzardato considerarla per il momento una variabile non influente sulla risposta (il sesso femminile ha la tendenza a diminuire la probabilità in esame con un coefficiente pari circa a -0.1746).

Proviamo a introdurre ora tre ulteriori variabili nello studio: il settore di attività, la propensione al rischio e il titolo di studio; le ultime due variabili citate sono di tipo categoriale ordinale e di default sono fattorizzate da R con la funzione *contr.poly* (dove i coefficienti dei contrasti lineare, cubico, ecc. sono combinazioni lineari delle

medie pesati dai polinomi ortogonali). Questa codifica ha lo scopo di mettere in evidenza trend (lineari, quadratici, cubici, ecc.) nelle variabili categoriali ordinali (*Azzalini, 2001*).

La variabile rischio non è significativa in nessuna delle sue componenti e si opta per la sua eliminazione da questo modello, anche perché i depositi bancari e postali sono l'attività finanziaria in assoluto meno rischiosa, e non avrebbe senso includerla. Inoltre, la sua inclusione, potrebbe evidenziare il suo carattere endogeno (presenza di correlazione tra parametro e termine d'errore) e quindi la sua ridondanza in termini di informazioni aggiuntive in realtà già contenute nella variabile risposta.

Il settore di attività risulta troppo correlato con altre variabili esplicative (valore del *vif* oltre la soglia di 5 considerata critica), e si preferisce non includerlo nell'analisi (*Bracalente, 2009*).

Per ciò che riguarda la variabile titolo di studio invece, essa è significativa nella sola componente lineare. Questo significa che si potrebbero escludere le componenti quadratiche, cubiche e alla quarta e che la probabilità di possedere il deposito bancario/postale aumenta linearmente con i titoli di studio più elevati.

Rimuovendo dunque le componenti superflue, e decidendo di non considerare settore e rischio perché portatrici di informazioni ridondanti, si ottiene il seguente output:

mod2<-glm(formula = pdepos ~ statuslav + reddito + eta + area + sesso + studio, family = binomial)

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*1.033e+00*	*5.164e-01*	*2.001*	*0.045379 **
statuslav2	*5.723e-01*	*3.736e-01*	*1.532*	*0.125538*
statuslav3	*1.277e+01*	*2.524e+02*	*0.051*	*0.959647*
statuslav4	*1.431e+00*	*1.051e+00*	*1.361*	*0.173476*
statuslav5	*4.345e-01*	*4.240e-01*	*1.025*	*0.305416*
statuslav6	*7.989e-01*	*2.313e-01*	*3.453*	*0.000554 ****
statuslav7	*2.185e-01*	*2.519e-01*	*0.868*	*0.385628*
reddito	*1.075e-04*	*7.269e-06*	*14.792*	*< 2e-16 ****
eta	*-2.580e-02*	*6.464e-03*	*-3.991*	*6.58e-05 ****

*area2 -8.802e-01 1.548e-01 -5.686 1.30e-08 ****

*area3 -1.539e+00 1.288e-01 -11.944 < 2e-16 ****

sesso2 -1.376e-01 1.079e-01 -1.276 0.201955

*studio 3.415e-01 6.670e-02 5.121 3.04e-07 ****

Residual deviance: 2604.4 on 5005 degrees of freedom

Lo status del lavoratore rimane dunque una variabile poco influente, e il sesso ora risulta non significativo senza più alcun dubbio.

Il test sulla devianza residua ci conduce ad accettare il modello (p-value vicino ad 1): il modello "fitta" bene i dati. Non sono inoltre presenti problemi di multicollinearità, in quanto i valori del *variance inflation factor* sono tutti inferiori a 2.

Una ulteriore valutazione della bontà di un modello (per variabili dicotomiche) può essere effettuata confrontando, in una tabella di frequenze, i valori classificati correttamente dal modello con quelli errati. Si mettono a confronto dunque le probabilità stimate dal modello con i valori osservati; per permettere questa analisi però si assume che i valori stimati superiori a 0.5 siano assimilabili a 1 e quelli inferiori a 0.5 assimilabili a 0 (altrimenti il confronto non sarebbe possibile).

> table(predict(mod4,type="response")>0.5, pdepos)

0 1

FALSE 168 107

TRUE 446 4297

Gli elementi sulla diagonale principale sono quelli classificati correttamente dal modello, e quindi una misura di affidabilità dello stesso è fornita da 0.8897967 (ottenuta dal rapporto tra la somma di 168+4297 e la somma di 107+446), che varia tra 0 ed 1. Il modello è molto buono anche sotto questo profilo perché classifica quasi l'89% dei dati in maniera esatta (*Ventura*).

Da precisare inoltre che l'utilizzo di diverse funzioni legame non comporta dei miglioramenti nel modello stimato; è da escludere poi il passaggio ad un modello di quasi verosimiglianza in quanto la stima del parametro di dispersione ψ sarebbe

1.003961; si assume dunque come valido un modello logistico che per definizione vincola il parametro di dispersione ad essere pari ad 1 (*Azzalini, 2001*).

Anche l'utilizzo di trasformate di variabili esplicative quantitative come reddito ed età non migliora l'adattamento ai dati perché produce parametri non significativi ed una peggiore classificazione delle probabilità stimate.

In conclusione, il possesso di un deposito bancario/postale a c/c o a risparmio dipende significativamente dall'area geografica, dall'età, dal reddito e dal titolo di studio conseguito, nelle seguenti forme:

Un aumento del reddito annuo di 1.000 €, ceteris paribus, aumenta la quota di possedere il deposito dell'11%; un aumento di 1 anno dell'età del capofamiglia la diminuisce del 2.6%; abitare al nord la aumenta rispetto al centro e ancor di più rispetto al sud e un titolo di studio di un livello superiore detenuto dal capofamiglia la aumenta del 40% circa.

Queste informazioni risulteranno utili in quanto permetteranno agli enti di riferimento (poste e banche) di capire quali caratteristiche della famiglia saranno compatibili con un investimento della stessa, per permettere delle azioni correttive in grado di ampliare la già larga diffusione dei depositi bancari e postali.

4.2- Modello per il possesso di titoli di Stato

I titoli di Stato sono titoli emessi dallo Stato per finanziare il debito pubblico, generalmente non garantiscono un rendimento molto elevato, ma rappresentano un investimento sicuro per chi desidera investire il proprio risparmio senza correre rischi. Il più noto è il classico BOT (Buono Ordinario del Tesoro); ha un taglio minimo di 1000 € e una scadenza a breve termine (tre, sei e dodici mesi). Per investimenti a medio e lungo periodo (tre, cinque, dieci e trenta anni) è più indicato il BTP (Buono del Tesoro Poliennale); il tasso è fisso e gli interessi sono pagati ogni sei mesi. Altri titoli di Stato emessi sono poi i CTZ e i CCT (*www.UnioneConsulenti.it*).

La stima di un primo modello per il possesso di titoli di Stato da parte della famiglia, in relazione a sesso, età, reddito e status del lavoratore fornisce già indicazioni molto rilevanti. E' preferibile però, in questo caso, utilizzare una funzione legame diversa

da quella canonica: la stima di un modello di tipo probit infatti, produce una minore devianza residua (3272.5 contro 3278.5) al pari dei gradi di libertà spesi. Di seguito è riportato l'output di questo primo modello descritto.

mod0<-glm(formula = pos.c ~ reddito + eta + sesso + statuslav, family = binomial(link = probit))

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	*-1.781e+00*	*2.263e-01*	*-7.871*	*3.52e-15*	*****
reddito	*9.702e-06*	*1.110e-06*	*8.744*	*< 2e-16*	*****
eta	*-6.738e-03*	*3.129e-03*	*-2.154*	*0.03127*	***
sesso2	*2.620e-05*	*5.469e-02*	*0.000*	*0.99962*	
statuslav2	*6.212e-01*	*1.588e-01*	*3.913*	*9.12e-05*	*****
statuslav3	*5.565e-01*	*1.912e-01*	*2.910*	*0.00361*	****
statuslav4	*5.025e-01*	*2.026e-01*	*2.480*	*0.01314*	***
statuslav5	*2.381e-01*	*2.015e-01*	*1.181*	*0.23751*	
statuslav6	*7.497e-01*	*1.457e-01*	*5.147*	*2.65e-07*	*****
statuslav7	*5.246e-01*	*1.687e-01*	*3.110*	*0.00187*	****

Residual deviance: 3272.5 on 5008 degrees of freedom

Il modello si adatta bene ai dati, in quanto il p-value riferito al test sulla devianza residua è prossimo a 1. Riguardo alle variabili possiamo già subito evidenziare l'importanza del reddito disponibile netto, variabile significativa e che produce un impatto positivo sulla probabilità di detenere titoli di Stato.

Anche l'età, significativa ad un livello del 5%, si conferma un deterrente della probabilità di possedere titoli di Stato (nel senso che assumeva un comportamento simile anche nel modello per il possesso di depositi postali/bancari), al crescere della stessa. A parità di altre condizioni, si evidenziano poi differenze significative riguardo allo status del lavoratore, in particolare si dimostra come gli impiegati, i dirigenti e i liberi professionisti hanno un'influenza positiva sulla risposta rispetto agli operai (parametro d'angolo): la variabile è significativa in quasi tutte le sue componenti. A differenza di quanto visto per i depositi bancari/postali si vede come

lo status del lavoratore sia ora una variabile influente per questa nuova forma di investimento considerata.

La variabile sesso risulta non significativa. Analizzando la relazione tra reddito e probabilità di possedere titoli di Stato, possiamo però vedere come una relazione di tipo cubico ipotizzabile tra le variabili sia preferibile, e ciò è confermato dal migliore adattamento del modello ai dati. Infatti, tramite la funzione *anova* per il confronto tra modelli generalizzati annidati si ottiene:

Model 1: pos.c ~ reddito + eta + sesso + statuslav

Model 2: pos.c ~ poly(reddito, 3) + eta + sesso + statuslav

	Resid. Df	*Resid. Dev*	*Df*	*Deviance*	*P(>|Chi|)*
1	*5008*	*3272.5*			
2	*5006*	*3159.1*	*2*	*113.39*	*< 2.2e-16 ****

Si accetta dunque il modello più esteso, con la variabile reddito presente con un polinomio di terzo grado (le componenti sono tutte significative); la devianza residua del modello si riduce.

Introduciamo ulteriori variabili nel modello, come propensione al rischio, titolo di studio, settore di attività e area geografica di residenza. Notiamo subito come la variabile propensione al rischio non aggiunga ulteriori informazioni al modello (questa variabile ha ancora poco senso analizzarla in quanto i titoli di Stato sono investimenti caratterizzati da bassissimo rischio). Riguardo al settore di attività, si opta per una sua eliminazione dal modello in quanto si diagnosticano, in seguito al suo inserimento, problemi di multicollinearità (valore del *vif* superiore a 5), e ciò potrebbe comportare un effetto distorsivo sulle stime dei parametri (difficilmente interpretabili per la varianza troppo alta dei coefficienti).

La variabile *studio* è invece significativa nella sola componente lineare, quindi la introduciamo nel modello considerandola così com'è stata definita. Si notano inoltre differenze significative per la probabilità di possesso di titoli di Stato al variare dell'area geografica: la probabilità diminuisce per le famiglie del centro e diminuisce sensibilmente per quelle del sud, sempre rispetto alle famiglie del nord che fanno da termine di confronto.

Con l'introduzione di queste nuove variabili l'età non è più significativa e si ritiene perciò non influente sulla variabile risposta.

Si assume come modello definitivo quindi un modello probit, con la variabile reddito disponibile netto presente fino alla componente cubica. L'eventuale stima di un modello quasi binomiale porta a una stima di ψ pari a 0.8982724; dunque si preferisce mantenere il modello descritto in precedenza. Di seguito è riportato l'output del modello assunto.

mod1<-glm(formula = pos.c ~ poly(reddito, 3) + eta + sesso + statuslav + studio + areageog, family = binomial(link = probit))

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-2.378642*	*0.261761*	*-9.087*	*< 2e-16 ****
poly(reddito, 3)1	*14.490335*	*2.099574*	*6.902*	*5.14e-12 ****
poly(reddito, 3)2	*-11.499103*	*2.056481*	*-5.592*	*2.25e-08 ****
poly(reddito, 3)3	*8.073954*	*1.932281*	*4.178*	*2.93e-05 ****
eta	*0.002273*	*0.003375*	*0.673*	*0.500698*
sesso2	*0.093812*	*0.059555*	*1.575*	*0.115209*
statuslav2	*0.550980*	*0.172607*	*3.192*	*0.001412 ***
statuslav3	*0.517357*	*0.200829*	*2.576*	*0.009992 ***
statuslav4	*0.400907*	*0.210920*	*1.901*	*0.057334 .*
statuslav5	*0.219821*	*0.209875*	*1.047*	*0.294919*
statuslav6	*0.711171*	*0.157377*	*4.519*	*6.22e-06 ****
statuslav7	*0.706820*	*0.184196*	*3.837*	*0.000124 ****
studio	*0.177122*	*0.028886*	*6.132*	*8.69e-10 ****
areageog2	*-0.465224*	*0.065387*	*-7.115*	*1.12e-12 ****
areageog3	*-0.937673*	*0.079621*	*-11.777*	*< 2e-16 ****

Residual deviance: 2929.8 on 5003 degrees of freedom

Il modello rimane molto buono sotto il profilo dell'analisi della devianza residua (p-value prossimo a 1); non si diagnosticano ulteriori problemi di multicollinearità tra le

variabili esplicative, infatti il *vif* più alto si riscontra per la variabile *statuslav* ma con un valore comunque contenuto e da ritenersi ammissibile (2.044318).

L'andamento particolare del reddito, a confronto con la probabilità di possedere titoli di Stato, ci mostra come questa cresca fino ad un certo livello (circa 100000 € annui di reddito) per poi diminuire oltre questa soglia, come si vede dalla **Figura 5.** I valori stimati si riferiscono al modello con il solo polinomio del reddito in funzione della probabilità di possedere titoli di Stato.

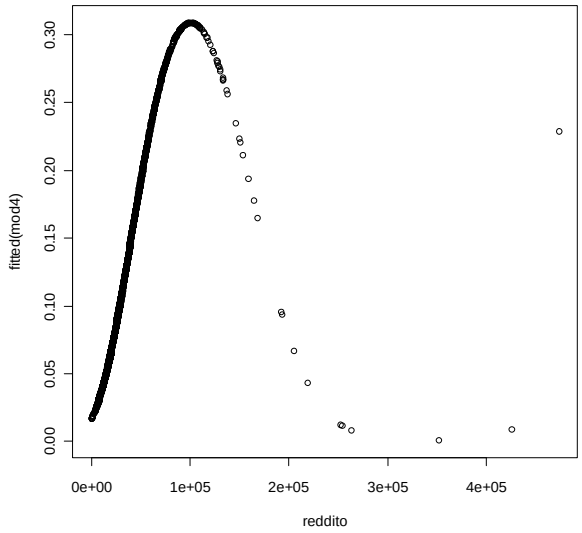

Figura 5, grafico tra reddito e valori stimati dal modello con il polinomio del reddito

Il possesso di un titolo di studio di livello superiore aumenta la probabilità di aver investito in titoli di Stato.

> exp(0.177122)

[1] 1.193777

Come già detto in precedenza, poi, è più probabile che la famiglia abbia investito se risiede al nord, rispetto alle altre due aree geografiche. L'aumento di probabilità non è però esprimibile quantitativamente in quanto si sta lavorando con un modello probit, per il quale non valgono le interpretazioni in termini di rapporto di quote o rischio relativo.

Lo stesso si può dire per la variabile status del lavoratore, in particolare gli operai vanno a diminuire la probabilità di possedere titoli di Stato, mentre la probabilità

aumenta in maniera considerevole in particolare per impiegati, dirigenti e pensionati, ceteris paribus.

Come per il modello stimato del possesso di depositi bancari e postali, si fornisce una misura di bontà del modello basata sulla corretta classificazione, mettendo a confronto valori osservati e valori stimati del modello *mod1*.

```
> table(predict(mod3,type="response")>0.5 ,pos.c)
      pos.c
             0    1
 FALSE   4476  540
 TRUE       1    1
```

Il modello classifica correttamente circa l'89.2% dei dati e risulta quindi molto affidabile nel determinare, definite certe caratteristiche del capofamiglia, se quest'ultimo può essere o meno un potenziale investitore in titoli di Stato di varia natura.

4.3- Modello per il possesso di fondi comuni e obbligazioni

Le obbligazioni sono titoli emessi dalle società per azioni o in accomandita per azioni al fine di finanziarsi. Fruttano un interesse annuo, semestrale, o trimestrale e prevede il rimborso alla scadenza. L'oscillazione del prezzo dell'obbligazione è inferiore a quello delle azioni e, quindi, il rischio connaturato al loro acquisto è minore rispetto ai titoli azionari ma è maggiore rispetto ai titoli di Stato.

I fondi comuni sono di svariate tipologie; due categorie principali consistono nei fondi comuni mobiliari e immobiliari. Per i primi i fondi si classificano in base all'incidenza di azioni, obbligazioni e titoli a reddito fisso che li compongono; quelli a maggior componente azionaria sono più rischiosi ma offrono migliori possibilità di guadagno.

Quelli immobiliari invece sono investimenti di denaro nell'acquisto di proprietà edilizie, i fabbricati che fanno parte del fondo vengono affittati per un determinato periodo e successivamente rivenduti; l'eventuale utile viene diviso tra i partecipanti al

fondo. In genere si tratta di fondi chiusi, ovvero è possibile acquistarne una quota solo nella fase di collocamento sul mercato (*www.unioneconsulenti.it*).

Si stima ora un primo modello che spiega il possesso di investimenti in fondi comuni e obbligazioni in relazione a reddito, sesso, età e status del lavoratore.

mod0<-glm(formula = pos.d ~ reddito + statuslav + eta + sesso, family = binomial)

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-2.111e+00*	*4.437e-01*	*-4.758*	*1.96e-06 ****
reddito	*3.416e-05*	*2.185e-06*	*15.633*	*< 2e-16 ****
statuslav2	*8.004e-01*	*2.719e-01*	*2.943*	*0.00325 ***
statuslav3	*7.744e-01*	*3.133e-01*	*2.472*	*0.01343 **
statuslav4	*9.374e-01*	*3.286e-01*	*2.853*	*0.00433 ***
statuslav5	*3.853e-01*	*3.349e-01*	*1.151*	*0.24987*
statuslav6	*1.167e+00*	*2.517e-01*	*4.635*	*3.58e-06 ****
statuslav7	*7.156e-01*	*3.201e-01*	*2.235*	*0.02539 **
eta	*-3.027e-02*	*6.550e-03*	*-4.621*	*3.82e-06 ****
sesso2	*-5.173e-01*	*1.187e-01*	*-4.360*	*1.30e-05 ****

Residual deviance: 3148.0 on 5008 degrees of freedom

Il modello logistico è quello preferibile rispetto all'utilizzo di diverse funzioni legame; è molto buono in termini di devianza residua e dunque si accetta l'ipotesi nulla sulla bontà dello stesso. Una nota particolare va fatta sulla variabile sesso che è significativa e mostra come il sesso femminile influisca negativamente sulla probabilità di aver investito in fondi comuni e obbligazioni, al netto delle altre esplicative.

La variabile età influisce negativamente, ed è ragionevole ipotizzare un legame lineare con la risposta. Per quanto riguarda il reddito invece, è preferibile considerare una relazione polinomiale di quarto grado, il che consente un adattamento migliore ai dati.

L'andamento dei valori stimati all'aumentare del reddito (vincolando le variabili categoriali a valori nulli e condizionandoci ad una età precisa) dimostra questa

tendenza: si presenta una curvatura verso il basso tra i 100000 € e i 200000 €, poi la probabilità di aver investito torna a salire per livelli di reddito superiori all'ultima soglia di reddito citata.

La categoria operai è quella caratterizzata dalla probabilità più bassa di avere effettuato l'investimento, probabilità che invece aumenta leggermente per gli impiegati, ancor di più per i dirigenti e in maniera sostanziale per i liberi professionisti, ceteris paribus . Lo status del lavoratore è significativo in quasi tutte le sue componenti.

Introducendo le altre variabili che si sono viste in precedenza, il modello rimane molto buono, anche se per il reddito si vede che una relazione cubica sembra sufficiente per spiegare la risposta. Il sorgere di eventuali problemi di correlazione tra variabili esplicative viene meno con l'eliminazione della variabile settore, come avvenuto per i precedenti modelli; a questo punto si può concludere che l'informazione fornita da questa variabile è contenuta in altre variabili esplicative, e dunque non ha senso di essere considerata.

La variabile *studio* è significativa nelle componenti lineare e quadratica, quindi si opta per tenere in considerazione solo i rispettivi parametri di riferimento. L'area geografica di residenza è significativa ad un elevato livello di significatività ed il suo effetto sulla risposta è molto simile a quello diagnosticato per il modello relativo al possesso dei titoli di Stato.

La propensione al rischio è da considerarsi nella sola componente lineare, e in questo caso il suo effetto è di riduzione della probabilità di aver investito in fondi comuni e obbligazioni al diminuire della propensione a rischiare; il risultato ottenuto inizia a dimostrare l'importanza di questa variabile. Va tenuto però conto del fatto che spesso fondi comuni e obbligazioni sono gestiti da professionisti, e che comunque sono investimenti meno rischiosi per esempio delle azioni e delle gestioni patrimoniali.

Output del modello *mod1*:

mod1<-glm(formula = pos.d ~ poly(reddito, 3) + statuslav + eta + sesso + rischio + areageog + poly(studio, 2), family = binomial)

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*

(Intercept) -0.75607 0.49583 -1.525 0.127300
*poly(reddito, 3)1 46.03981 3.74913 12.280 < 2e-16 ****
*poly(reddito, 3)2 -21.08020 3.05763 -6.894 5.41e-12 ****
*poly(reddito, 3)3 11.99741 3.00141 3.997 6.41e-05 ****
*statuslav2 0.62660 0.28410 2.206 0.027413 **
*statuslav3 0.73219 0.32306 2.266 0.023424 **
*statuslav4 0.67511 0.33380 2.022 0.043126 **
statuslav5 0.24831 0.33726 0.736 0.461570
*statuslav6 0.99582 0.25855 3.852 0.000117 ****
*statuslav7 0.85860 0.33330 2.576 0.009994 ***
eta -0.01201 0.00681 -1.763 0.077926 .
*sesso2 -0.35428 0.12342 -2.870 0.004099 ***
*rischio -0.35546 0.06506 -5.464 4.66e-08 ****
*areageog2 -0.76133 0.12637 -6.025 1.70e-09 ****
*areageog3 -1.49209 0.15393 -9.693 < 2e-16 ****
*poly(studio, 2)1 25.93424 5.02721 5.159 2.49e-07 ****
*poly(studio, 2)2 -11.37543 4.31621 -2.636 0.008401 ***
Residual deviance: 2856.8 on 5001 degrees of freedom

Il modello logistico rimane quello preferibile in termini di devianza residua, e non è nemmeno necessario passare a un modello di quasi verosimiglianza (ψ=0.8755954).
La probabilità di aver investito diminuisce del 30% per il sesso femminile rispetto al sesso maschile, aspetto che differisce dalle precedenti analisi dove questa variabile risultava non significativa; il dato va sempre interpretato al netto dei restanti regressori. L'età ha un debole effetto e non sarebbe azzardato considerarla ininfluente sulla risposta.
L'adattamento dei dati è molto buono e ciò è confermato dai due strumenti utilizzabili ai fini di valutarne la bontà (test sulla devianza residua e tabella di corretta/errata classificazione).
> 1-pchisq(2856.8,5001)

[1] 1

> table(predict(mod3,type="response")>0.5, pos.d)

	0	*1*
FALSE	*4350*	*506*
TRUE	*73*	*89*

Come si vede il test del log rapporto di verosimiglianza produce un p-value prossimo a 1 e l'analisi della tabella di frequenza che classifica gli elementi "corretti" ed "errati" del modello fornisce un indice di 0.8846154: il modello classifica correttamente oltre l'88% dei valori osservati.

Le società che emettono obbligazioni potranno dunque sfruttare queste informazioni rilevanti per individuare le aree, le professioni e quindi le famiglie che dispongono delle caratteristiche idonee ritenute necessarie per poter effettuare un investimento di questo tipo. Sarà utile dunque, ai fini ottenere un vantaggio competitivo, capire chi potrebbe essere un potenziale investitore per fornirgli una corretta informazione, sia intravvedere delle possibili applicazioni per fidelizzare i clienti già acquisiti.

4.4- Modello per il possesso di azioni e quote di società

Le azioni sono tra gli strumenti finanziari più rischiosi, ma possono garantire rendimenti consistenti a chi riesce a utilizzarli con abilità e a identificare il momento giusto nel quale acquistare e vendere. L'andamento di un titolo azionario dipende, soprattutto nel breve periodo, da molti fattori: andamento generale dell'economia, redditività dell'azienda, movimenti speculativi, eventi politici, sviluppo della tecnologia e altri accadimenti. Per cercare di mettersi al riparo da crolli improvvisi, è necessario seguire costantemente il mercato. L'investitore non professionale dovrebbe quindi avvicinarsi alle azioni con cautela e in un'ottica di medio lungo periodo (*www.unioneconsulenti.it*).

La maggiore evidenza a riguardo di questo tipo di investimenti, come si è visto dalle analisi descrittive, comporta una percentuale molto bassa delle famiglie disposte ad investire in azioni e quote di società. Si va ora a verificare quali siano quelle

caratteristiche di famiglie e relativo capofamiglia ultracinquantenne che influenzano il possesso o meno di questa forma di investimento.

Analogamente alla procedura adottata sinora, il seguente output si riferisce alla stima del modello con la probabilità di aver investito in azioni e quote di società come variabile risposta, in relazione a sesso, età, reddito e status del lavoratore.

mod0<-glm(formula = pos.e ~ reddito + eta + sesso + statuslav, family = binomial(link = probit))

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-1.230e+00*	*2.727e-01*	*-4.513*	*6.40e-06 ****
reddito	*1.186e-05*	*1.243e-06*	*9.539*	*< 2e-16 ****
eta	*-1.643e-02*	*4.182e-03*	*-3.928*	*8.56e-05 ****
sesso2	*-3.639e-01*	*7.643e-02*	*-4.761*	*1.93e-06 ****
statuslav2	*4.122e-01*	*1.615e-01*	*2.552*	*0.010697 **
statuslav3	*4.454e-01*	*1.907e-01*	*2.335*	*0.019542 **
statuslav4	*6.469e-01*	*1.940e-01*	*3.335*	*0.000853 ****
statuslav5	*5.771e-02*	*2.108e-01*	*0.274*	*0.784317*
statuslav6	*4.620e-01*	*1.486e-01*	*3.108*	*0.001881 ***
statuslav7	*4.094e-01*	*1.856e-01*	*2.206*	*0.027413 **

Residual deviance: 2043.5 on 5008 degrees of freedom

La funzione legame preferibile è quella probit, che produce un effetto di minore devianza residua nel modello. La relazione ottimale per le variabili quantitative età e reddito è quella di tipo lineare, in particolare per il reddito, per il quale nei modelli precedenti si era spesso notato come la relazione poteva essere di tipo polinomiale. Il grafico tra reddito e valori stimati dal modello ci può confermare che l'ipotesi di un andamento lineare tra le variabili risulta sufficiente, nonostante la presenza di una curvatura (comunque di debole entità) per valori elevati della distribuzione dei valori "fittati" dal modello (vedi **Figura 6**).

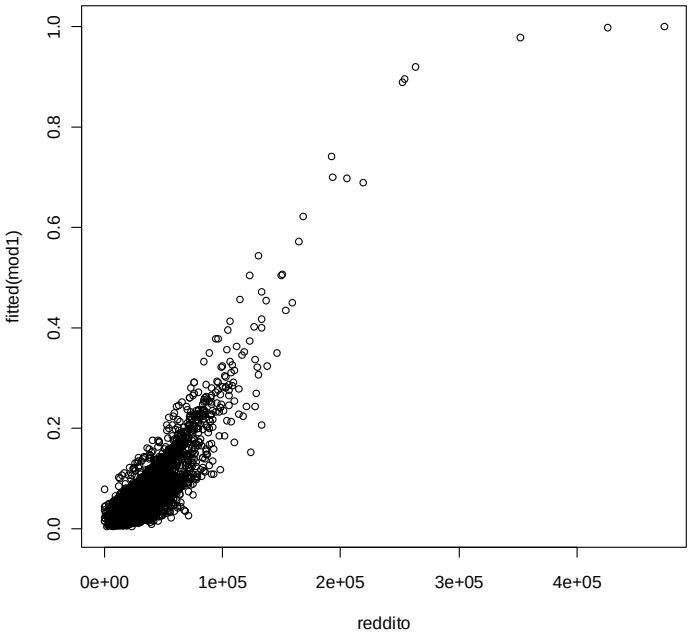

Figura 6, grafico tra reddito e valori stimati dal modello *mod0*

Le stime dei coefficienti relativi a sesso, età e reddito sono significative ad un livello di significatività dell'1%, e i valori assunti da tali coefficienti permettono di identificare una relazione crescente della probabilità di investire in azioni e quote di società in relazione al reddito e una relazione inversa in relazione all'età. Il sesso femminile va a diminuire la probabilità di possedere azioni rispetto al sesso maschile.
Lo status del lavoratore è significativo in quasi tutte le sue componenti e identifica la categoria degli imprenditori e dei liberi professionisti come i più attivi negli investimenti di questo genere. In questo primo modello "parziale", la probabilità di aver investito aumenta, per gli appartenenti alla categoria sopra citata, rispetto alla categoria operai.
L'introduzione di ulteriori variabili esplicative come propensione al rischio, area geografica e titolo di studio conferma questa tendenza che si estende anche alla categoria dei pensionati, l'altra categoria che rimane con coefficiente significativamente diverso da zero e con segno positivo.

La variabile propensione al rischio assume ora un valore decisamente elevato, come dimostrano i boxplot delle varie categorie della variabile utilizzando come valori medi i valori stimati dal modello *mod1* definito in seguito (vedi **Figura 7**).

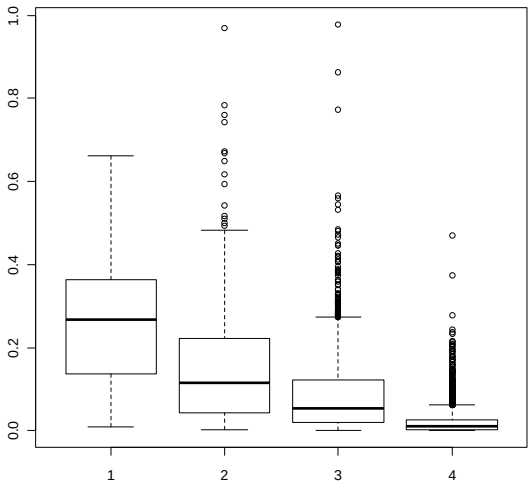

Figura 7, boxplot (propensione al rischio, possesso azioni e quote di società)

E' facile notare come gli individui con un'alta propensione al rischio siano quelli con le probabilità stimate di aver investito più elevate; l'effetto di questa variabile è coerente con la natura della variabile risposta che coinvolge il possesso di investimenti per definizione ad alto rischio (e contemporanee alte prospettive di guadagno).

Di seguito è riportata la stima del modello definitivo per il possesso di azioni e quote di società.

mod1<-glm(formula = pos.e ~ reddito + eta + sesso + statuslav + areageog + studio + rischio, family = quasibinomial(link = probit))

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*-8.982e-01*	*3.080e-01*	*-2.917*	*0.00355 **
reddito	*6.628e-06*	*1.197e-06*	*5.539*	*3.20e-08 ***
eta	*-8.333e-03*	*4.065e-03*	*-2.050*	*0.04043 *
sesso2	*-3.157e-01*	*7.546e-02*	*-4.184*	*2.91e-05 ***
statuslav2	*2.217e-01*	*1.623e-01*	*1.366*	*0.17196*
statuslav3	*1.658e-01*	*1.879e-01*	*0.882*	*0.37765*

statuslav4	*4.256e-01*	*1.881e-01*	*2.262*	*0.02372 **
statuslav5	*-4.615e-02*	*2.069e-01*	*-0.223*	*0.82353*
statuslav6	*3.430e-01*	*1.480e-01*	*2.317*	*0.02052 **
statuslav7	*4.435e-01*	*1.865e-01*	*2.378*	*0.01746 **
areageog2	*-3.362e-01*	*7.794e-02*	*-4.314*	*1.63e-05 ****
areageog3	*-7.730e-01*	*8.808e-02*	*-8.777*	*< 2e-16 ****
studio	*2.757e-01*	*3.339e-02*	*8.258*	*< 2e-16 ****
rischio	*-3.897e-01*	*3.871e-02*	*-10.066*	*< 2e-16 ****

Residual deviance: 1798.5 on 5004 degrees of freedom

(Dispersion parameter for quasibinomial family taken to be 0.8357821)

La scelta di passare ad un modello di quasi verosimiglianza è stata dettata dal fatto che il parametro di dispersione ψ risulta piuttosto basso (soprattutto in confronto ai precedenti modelli) e per la sua capacità di mantenere la variabile età significativa ad un livello di significatività del 5%. Da ricordare che le stime dei parametri non variano con il passaggio alla quasi binomiale, ma si modificano soltanto gli standard error e quindi conseguentemente i p-value dei test sui coefficienti (*Azzalini, 2001*).

Le variabili rischio e studio risultavano significative nelle sole componenti lineari e quindi si è preferito considerarle così come definite. Al diminuire della propensione al rischio si ottiene un effetto di riduzione della probabilità di aver effettuato investimenti in azioni. Parallelamente, il passaggio ad un titolo di studio più elevato aumenta la medesima probabilità. (ceteris paribus).

E' da confermare quanto detto riguardo allo status del lavoratore, anche se in realtà con il nuovo modello emerge una relazione più debole con la risposta. Lo stesso dicasi per l'età: al crescere dell'età del capofamiglia la probabilità in esame si riduce ma l'impatto di essa è debole.

Un incremento del reddito disponibile netto produce un aumento di probabilità di detenere azioni e quote di scocietà. Il modello stimato utilizza una funzione legame probit, e dunque non si possono interpretare i rapporti tra le quote dei diversi eventi e le relative approssimazioni al rischio relativo.

In conclusione, prendendo in esame i coefficienti di riferimento della variabile area geografica, si notano delle differenze sostanziali tra le diverse aree, differenze più marcate rispetto al modello precedente che analizzava il possesso di fondi comuni e obbligazioni. Al netto delle altre variabili esplicative si evidenzia come abitare al nord sia un fattore incrementale della probabilità di aver investito in azioni e quote di società. Rispettivamente risiedere al centro e al sud provoca una diminuzione della probabilità rispetto al risiedere al nord.

Per valutare l'affidabilità e la bontà del modello, che non presenta oltretutto problemi di multicollinearità, ci si riferisce alla sola analisi della corretta/errata classificazione delle probabilità. Il test sulla devianza residua non è interpretabile in quanto, essendo la devianza non normalizzata, non può essere utile ai fini di valutare l'adeguatezza del modello perché essa varia simultaneamente con il parametro di dispersione che è stato stimato (si rammenti infatti che siamo sotto ipotesi di quasi verosimiglianza).

Si riportano qui i risultati dell'analisi della tabella di frequenza per la corretta classificazione dei valori stimati a confronto con quelli osservati.

```
> table(predict(mod1,type="response")>0.5,pos.e)
       pos.e
             0    1
  FALSE   4697  297
  TRUE      10   14
> (4697+14)/5018
[1] 0.9388202
```

Il modello classifica correttamente quasi il 94% dei dati, il risultato è molto soddisfacente e si ripercuote sull'affidabilità dello stesso in termini di predizioni.

Da questo modello emerge chiaramente come la propensione al rischio dei capifamiglia svolga un ruolo importante nel discriminare tra potenziali investitori, come del resto il sesso; entrambi gli aspetti risultano specifici per le attività finanziarie analizzate in questo paragrafo. Sarà poi compito delle aziende sfruttare questi risultati al fine di coinvolgere nuovi investitori, puntando anche a tranquillizzarli riguardo al rischio connesso all'investimento, in modo da creare un

clima di maggiore fiducia intorno all'impresa e al valore delle proprie azioni/quote di capitale sociale.

4.5- Modello per il possesso di gestioni patrimoniali

Le gestioni patrimoniali, come si è visto dalle precedenti analisi descrittive, sono le forme di investimento meno diffuse tra le famiglie italiane (1.4%). In un mercato finanziario maturo come quello italiano, che offre molte possibilità d'investimento, i risparmiatori che vogliono ottimizzare i risultati dell'investimento del loro patrimonio possono rivolgersi alle banche per trovare la soluzione migliore per le loro esigenze; questo tipo di servizio prende il nome di gestione patrimoniale. Nelle gestioni patrimoniali, il gestore traccia un profilo del patrimonio del cliente, e in particolare della sua capacità e tolleranza al rischio. Poi, sulla base di tale profilo, definisce insieme al cliente gli obiettivi di investimento. Infine crea un portafoglio di attività finanziarie che corrisponda agli obiettivi prefissati. Per il servizio di gestione il cliente sostiene dei costi di gestione della gestione patrimoniale, di performance della gestione patrimoniale e di sottoscrizione di uscita dalla gestione (*www.unioneconsulenti.it*).

Si ricordi che anche le gestioni patrimoniali sono da considerarsi investimenti di un rischio finanziario rilevante e di corrispondenti opportunità di guadagno elevate.

La stima di un modello logistico per la variabile risposta possesso di gestioni patrimoniali da luogo a dei risultati molto diversi dalle precedenti forme di investimento: al variare infatti di diverse funzioni legame e alla progressiva aggiunta di nuove variabili nel modello si ottengono sempre parametri non significativi per tutte le variabili esplicative eccetto il reddito. Sembra proprio che l'unica variabile che influisce sulla decisione o meno di affidare il proprio patrimonio ad una gestione patrimoniale sia il reddito disponibile netto. Come si può facilmente notare dal seguente output del modello stimato tenendo conto di tutte le variabili considerate sinora nella trattazione.

mod0<-glm(formula = pos.f ~ reddito + rischio + studio + eta + sesso + statuslav + areageog, family = binomial)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.352e+01	9.499e+02	-0.025	0.9802	
reddito	1.273e-05	3.093e-06	4.118	3.83e-05	***
rischio.L	-5.401e-01	7.196e-01	-0.751	0.4529	
rischio.Q	-4.269e-01	5.566e-01	-0.767	0.4430	
rischio.C	-3.461e-01	3.238e-01	-1.069	0.2851	
studio.L	1.075e+01	5.365e+02	0.020	0.9840	
studio.Q	-7.934e+00	4.534e+02	-0.017	0.9860	
studio.C	4.110e+00	2.682e+02	0.015	0.9878	
studio^4	-1.502e+00	1.014e+02	-0.015	0.9882	
eta	-2.210e-03	1.685e-02	-0.131	0.8956	
sesso2	1.911e-03	3.076e-01	0.006	0.9950	
statuslav2	1.563e+01	9.346e+02	0.017	0.9867	
statuslav3	1.637e+01	9.346e+02	0.018	0.9860	
statuslav4	1.598e+01	9.346e+02	0.017	0.9864	
statuslav5	1.606e+01	9.346e+02	0.017	0.9863	
statuslav6	1.619e+01	9.346e+02	0.017	0.9862	
statuslav7	1.619e+01	9.346e+02	0.017	0.9862	
areageog2	-1.180e-01	3.074e-01	-0.384	0.7010	
areageog3	-7.065e-01	3.610e-01	-1.957	0.0503	.

Risulta dunque evidente la particolarità di questa variabile risposta: la probabilità di investire in una gestione patrimoniale è funzione del solo reddito disponibile netto, in quanto per le restanti variabili si accettano le ipotesi nulle sulla nullità dei parametri della regressione.

A questo punto, al fine di prevedere la distribuzione di probabilità di questa variabile risposta, andiamo a stimare un modello con il solo reddito a fungere da regressore.

Risulta preferibile ipotizzare una relazione quadratica tra il reddito e la probabilità di aver investito in gestioni patrimoniali, come testimonia il confronto tra i due modelli annidati che porta a rifiutare il modello più semplice (funzione *anova*).

> anova(mod2,mod3,test="Chisq")

Analysis of Deviance Table

Model 1: pos.f ~ reddito

Model 2: pos.f ~ poly(reddito, 2)

Resid. Df Resid. Dev Df Deviance P(>|Chi|)

1 5016 679.91

*2 5015 662.44 1 17.475 2.911e-05 ****

La stima del modello con la variabile reddito presente fino a un polinomio di secondo grado produce il seguente risultato (il modello probit risulta più appropriato del modello logistico per la minore devianza prodotta):

mod1<-glm(formula = pos.f ~ poly(reddito, 2), family = binomial(link = probit))

Estimate Std. Error z value Pr(>|z|)

*(Intercept) -2.34740 0.05735 -40.931 < 2e-16 ****

*poly(reddito, 2)1 22.53041 2.31053 9.751 < 2e-16 ****

*poly(reddito, 2)2 -7.23434 1.78364 -4.056 4.99e-05 ****

Residual deviance: 662.44 on 5015 degrees of freedom

Il modello produce stime significative nei parametri ad un livello di confidenza dell'1% e risulta ottimo in termini di devianza residua. I valori stimati dal modello sono crescenti fino ad un livello di reddito pari a 300000 € annui, per poi decrescere oltre tale soglia (vedi **Figura 8**).

Si noti in particolare, di come la probabilità resti molto bassa per redditi inferiori ai 100000 €, e si alza considerevolmente solo oltre quel livello.

Le altre variabili sono ininfluenti e dunque si assume che la probabilità di aver investito in gestioni patrimoniali sia funzione del solo reddito.

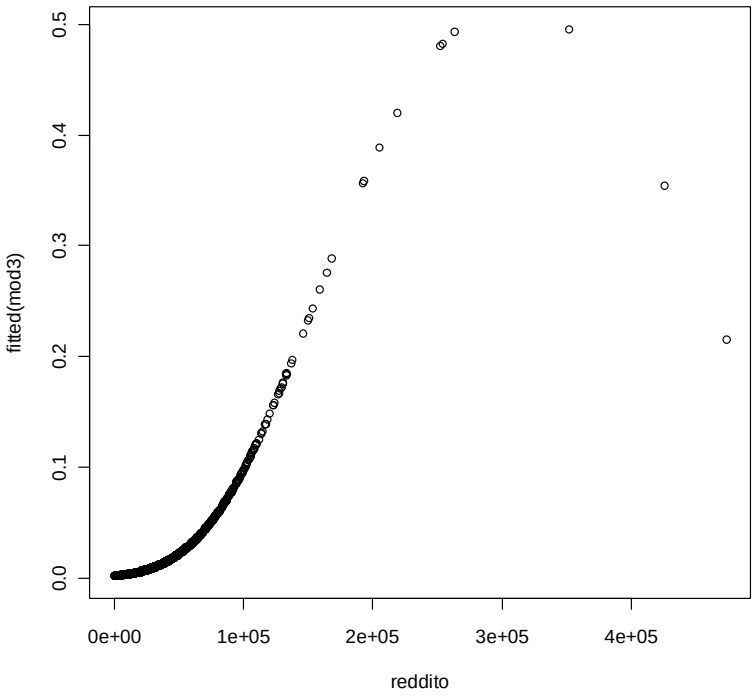

Figura 8, grafico tra reddito e valori stimati dal modello *mod1*

La peculiarità del risultato ottenuto può essere fortemente influenzata dalla scarsa numerosità dei possessori di gestioni patrimoniali, che ricordiamo costituiscono soltanto l'1.4% del campione.

5- MODELLI PER L'AMMONTARE DELLE FORME DI RISPARMIO

Si consideri un modello di regressione lineare multipla nella usuale forma del tipo $Yi = \beta 1xi1 + \ldots + \beta pxip + \varepsilon i;$ dove εi è il termine d'errore e rappresenta la componente stocastica del modello. La variabile risposta *Yi* è quantitativa, e misura l'ammontare delle risorse finanziarie destinate alle varie tipologie di forme di risparmio; per ogni *Yi* si assume $Yi \sim N(\mu i, \sigma^2)$, cioè ogni *Yi* si distribuisce come una variabile casuale normale di media μi e varianza σ^2.

Si vuole modellare la media della variabile *Y* in funzione delle p variabili esplicative, la media della variabile *Yi* risulta quindi $E(Yi) = \mu i = \beta 1xi1 + \ldots + \beta pxip$, quindi espressa come combinazione lineare delle variabili esplicative.

In sintesi, le ipotesi alla base di questi modelli sono:

(1) $Yi = \beta 1xi1 + \ldots + \beta pxip + \varepsilon i = \mu i + \varepsilon i$ dove $\varepsilon i \sim N(0, \sigma^2)$; quindi gli errori si assumono normali e indipendentemente distribuiti, con media 0 e varianza costante.

(2) X=(x1,…,xn) matrice non stocastica di dimensioni (nxp) e a rango pieno.

(3) Ogni parametro βj associato alle variabili esplicative deve entrare in maniera lineare nel modello.

Riguardo all'assunzione di normalità, si dovranno effettuare dei test sui residui del modello, per verificare la validità di questa ipotesi; inoltre risulteranno necessari dei test che verifichino l'omoschedasticità del modello, perché in assenza di questa ipotesi possono sorgere problemi di distorsione di stima della varianza e di conseguenza degli errori standard, e ciò può invalidare i test di significatività dei coefficienti. Le stime sono ottenute massimizzando la funzione di verosimiglianza; in generale questo metodo risulta robusto, ossia piccole variazioni delle ipotesi del modello non invalidano l'inferenza o le conclusioni a cui esso conduce (*Azzalini, 2001*).

5.1- Modello per l'ammontare dei depositi bancari e postali a c/c o a risparmio

Si vuole a questo punto valutare, a confronto con i rispettivi modelli per il possesso delle varie forme di investimento, se le stesse variabili che discriminano tra potenziali

investitori e non, delle varie attività finanziarie, sono le stesse che descrivono l'ammontare di risorse impiegate nell'investimento. Per tutte le analisi che seguono, i data set utilizzati si condizionano al possesso della rispettiva forma di investimento; si tiene conto cioè, per esempio in questo primo caso, solo delle famiglie che possiedono un deposito bancario o postale a c/c o a risparmio.

Si va dunque a stimare un modello di regressione lineare multipla con l'ammontare dei depositi bancari e postali come variabile risposta, in funzione delle solite variabili esplicative di cui si è ampiamente discusso. Si decide di lavorare con una trasformata della variabile risposta, quindi si prende la radice quadrata dell'ammontare dei depositi per procedere su una scala più ridotta e perché questa trasformazione è spesso utile in caso di non normalità degli errori (l'effetto dei regressori sarà dunque da misurare sulla radice quadrata dell'ammontare di risorse destinate all'investimento). Altre trasformazioni solitamente di uso comune, come la trasformata logaritmica, sembrano meno adatte della radice quadrata in quanto, in fase di diagnostica dei residui producono risultati meno confortanti (in riferimento alle ipotesi sottostanti il modello).

Dopo una prima analisi esplorativa si nota come le singole relazioni tra la variabile risposta e alcune variabili esplicative non evidenziano grosse differenze in media tra le variabili categoriali (vedi **Figura 10, 11 e 12**). Anche la relazione tra reddito e ammontare investito nei depositi è graficamente di difficile interpretazione; non sembra esserci una relazione ben definita (vedi **Figura 9**).

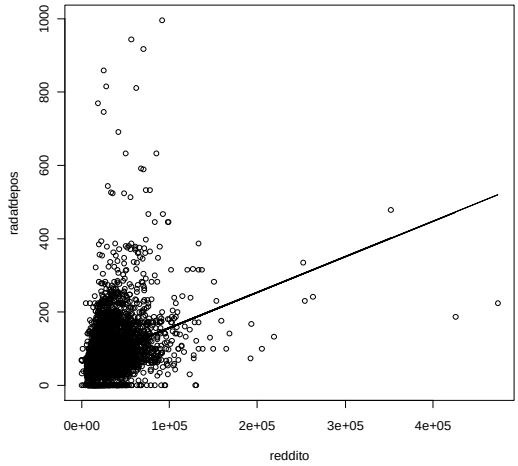

Figura 9, relazione tra reddito e ammontare investito in depositi bancari/postali (radice quadrata)

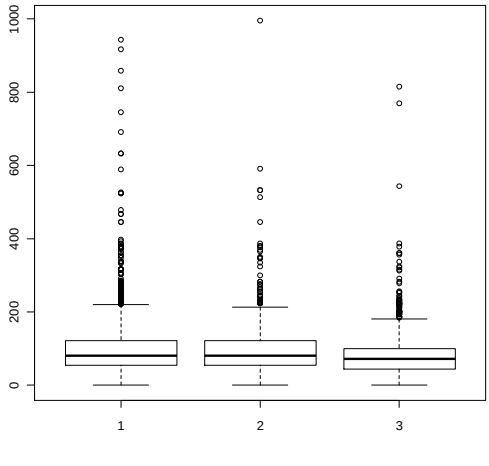

Figura 10, relazione tra area geografica e ammontare investito in depositi bancari/postali (radice quadrata)

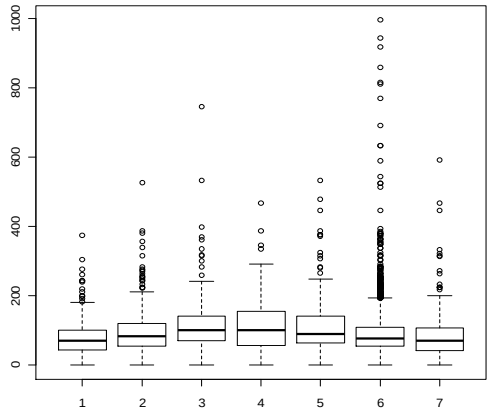

Figura 11, relazione tra status del lavoratore e ammontare investito in depositi bancari/postali (radice quadrata)

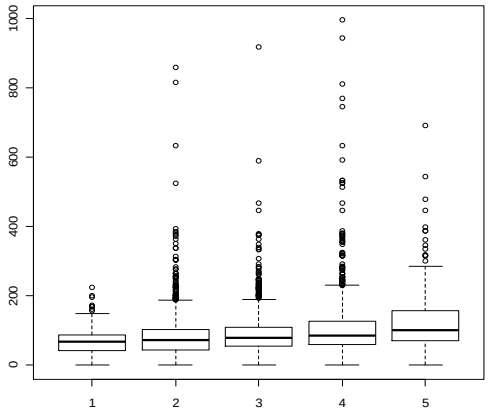

Figura 12, relazione tra titolo di studio e ammontare investito in depositi bancari/postali (radice quadrata)

Già prima dell'analisi preliminare è stata esclusa la variabile rischio, perché si stanno trattando risorse finanziarie investite in una forma di risparmio a bassissimo rischio.
Dai grafici sopra riportati si evince che non sussistono grandi differenze in media al variare dell'area geografica, dello status del lavoratore e del titolo di studio; anche se per esempio per la categoria di imprenditori e quella dei laureati i valori medi sono più elevati rispetto alla media generale (rispettivamente 20463.72 € e 22600.38 € con una media generale di 14131.84 €). Queste differenze, che come si è detto in precedenza non sono di ampio respiro, emergono comunque se si analizzano dei modelli di analisi della varianza rispetto a ciascuna variabile categoriale. Per esempio per la variabile *statuslav* si ottengono i seguenti risultati:

lm(formula = radafdepos ~ statuslav)

(Intercept)	*77.444*	*4.363*	*17.748*	*< 2e-16 ****
statuslav2	*17.157*	*5.830*	*2.943*	*0.00327 ***
statuslav3	*42.927*	*7.638*	*5.620*	*2.03e-08 ****
statuslav4	*37.334*	*8.079*	*4.621*	*3.93e-06 ****
statuslav5	*38.082*	*7.238*	*5.261*	*1.50e-07 ****
statuslav6	*13.753*	*4.567*	*3.011*	*0.00262 ***
statuslav7	*5.744*	*6.114*	*0.940*	*0.34751*

Residual standard error: 74.18 on 4397 degrees of freedom
Multiple R-squared: 0.0144, Adjusted R-squared: 0.01306
F-statistic: 10.71 on 6 and 4397 DF, p-value: 7.501e-12

Il problema però è che queste differenze, pur essendo significative, non hanno utilità nel determinare l'ammontare investito, infatti il coefficiente di determinazione lineare è prossimo allo zero. Dunque il modello non si adatta per niente ai dati. Questa tendenza si estende anche alle altre variabili studio e area geografica (analizzando sempre i modelli con una sola variabile esplicativa e svolgendo dunque delle analisi della varianza ad un fattore).

Per quanto riguarda il reddito si può subito verificare come una relazione lineare spieghi in maniera grossolana il fenomeno (come si vede dalla **Figura 9** al quale è stata aggiunta la retta stimata). Una relazione quadratica sembra essere preferibile, soprattutto perché tramite la funzione *anova* si accetta il modello con più parametri. Persiste tuttavia questa scarsa capacità del modello di spiegare i dati (nel modello con il solo reddito inserito come polinomio di secondo grado il coefficiente di determinazione lineare è pari a 0.1077 e varia tra 0 ed 1).

Vista la già scarsa capacità de modello di adattarsi bene ai dati, si opta per togliere dall'analisi le variabili sesso ed età del capofamiglia, che nel passaggio alla stima di un modello di regressione lineare multipla risultano non significative; la scelta è supportata inoltre dall'utilizzo della funzione di R *stepAIC*, che sceglie automaticamente le variabili da includere nel modello. Il criterio si basa sulla minimizzazione della quantità AIC=n*log(RSS)+2p dove RSS è la somma quadratica dei residui e p è il numero di regressori (*Salvan, 2001*).

Il modello preferibile risulta dunque dato dal seguente output:

mod0<-lm(formula = radafdepos ~ poly(reddito, 2) + statuslav + studio + areageog + studio)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*70.434*	*5.535*	*12.726*	*< 2e-16 ****
poly(reddito, 2)1	*1480.843*	*82.756*	*17.894*	*< 2e-16 ****

*poly(reddito, 2)2 -519.524 72.775 -7.139 1.1e-12 ****

statuslav2 3.521 5.646 0.624 0.53296

statuslav3 1.30 7.523 0.173 0.86270

statuslav4 5.762 7.798 0.739 0.45997

*statuslav5 22.023 6.897 3.193 0.00142 ***

*statuslav6 17.777 4.335 4.101 4.2e-05 ****

*statuslav7 12.527 5.808 2.157 0.03108 **

*studio 3.600 1.186 3.036 0.00241 ***

areageog2 -4.045 2.698 -1.499 0.13394

*areageog3 -7.084 2.584 -2.741 0.00615 ***

Residual standard error: 70.24 on 4392 degrees of freedom

Multiple R-squared: 0.1174, Adjusted R-squared: 0.1152

F-statistic: 53.1 on 11 and 4392 DF, p-value: < 2.2e-16

L'analisi congiunta delle variabili evidenza le reali influenze delle variabili sull'ammontare dei depositi: in particolare per lo status del lavoratore la categoria che presenta una significativa differenza dalle altre è quella dei pensionati. Per loro infatti, si ha un aumento medio di ammontare di risorse pari a circa 316 € (il parametro di riferimento per i pensionati vale 17.777 e andrà elevato al quadrato per la trasformazione applicata sulla variabile risposta, si ricordi che ciò vale per tutti i parametri del modello) rispetto agli operai. I residenti al sud, altra variabile significativa, possiedono mediamente risparmi inferiori di 50 € circa rispetto al nord; per il centro invece non ci sono differenze significative con il nord. L'ammontare cresce anche in seguito ad un incremento del titolo di studio conseguito, a parità di altre condizioni.

I risultati divergono per alcuni aspetti dal modello stimato per il possesso di depositi bancari e postali, in particolare per ciò che concerne l'età, qui non più influente, e i residenti del centro, che qui non presentano significative differenze dal nord.

Il modello comunque, come si è già detto, descrive poco della variabile in oggetto, ed ha dunque una scarsa capacità previsiva (il che poteva essere intuito già dal calcolo di

alcune correlazioni piuttosto basse). L'ipotesi di omoschedasticità non sembra una forzatura mentre la debolezza del modello si riscontra anche nell'ipotesi di normalità dei residui, scarsamente rispettata (qq plot). Non si riscontrano problemi di multicollinerità tra regressori una volta eliminata la variabile *settore*.

5.2- Modello per l'ammontare dei titoli di Stato

Gli ammontari investiti in titoli di Stato non presentano scostamenti significativi dalla media generale (34690.91 €) al variare dello status del lavoratore e del sesso del capofamiglia, ciò è confermato anche dalle analisi della varianza ad un fattore eseguite per entrambe le variabili. Esse rimangono poi totalmente ininfluenti anche nel passaggio ad una analisi congiunta di tutte le variabili. La propensione al rischio è ancora da ritenersi esclusa a priori per la sua inadeguatezza in riferimento ai titoli di Stato, investimenti a rischio esiguo.

Per ragioni di praticità, si lavora ancora con la radice quadrata della variabile risposta. In questo caso la relazione con il reddito risulta più "linearizzabile", e non sembra azzardato rinunciare all'introduzione di un polinomio di grado più elevato. E' utile specificare inoltre che una trasformazione logaritmica risulta essere meno adatta di quella utilizzata.

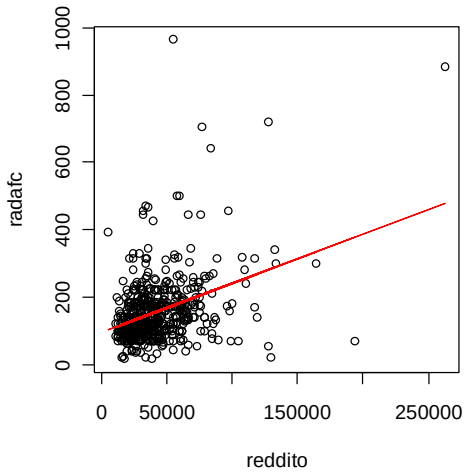

Figura 13, relazione tra reddito e ammontare investito in titoli di Stato (radice quadrata)

La **Figura 13** mostra la correlazione tra il reddito e l'ammontare (radice quadrata), che risulta comunque piuttosto debole e di diretta proporzionalità:

> cor(radafc,reddito)

[1] 0.3705787

Le altre variabili che si è scelto di includere nel modello sono l'età, l'area geografica di residenza e il titolo di studio (nella sola componente lineare). Di seguito è riportato l'output della regressione.

mod0<-lm(formula = radafc ~ reddito + eta + areageog + studio)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*-1.648e+01*	*3.500e+01*	*-0.471*	*0.63806*
reddito	*1.404e-03*	*1.708e-04*	*8.224*	*1.5e-15 ****
eta	*1.309e+00*	*4.287e-01*	*3.054*	*0.00237 ***
areageog2	*2.870e+01*	*1.033e+01*	*2.778*	*0.00567 ***
areageog3	*-1.802e+01*	*1.522e+01*	*-1.184*	*0.23697*
studio	*7.398e+00*	*4.244e+00*	*1.743*	*0.08186 .*

Residual standard error: 90.08 on 535 degrees of freedom

Multiple R-squared: 0.1719, Adjusted R-squared: 0.1642

F-statistic: 22.22 on 5 and 535 DF, p-value: < 2.2e-16

Il reddito influisce significativamente sulla risposta, anche se debolmente: un aumento del reddito annuo di 50.000 € produce infatti un incremento medio degli investimenti in titoli di Stato per un ammontare di soli 4931.26 €. Solo variazioni consistenti nei redditi inducono quindi ad optare per una destinazione maggiore di risorse in titoli di Stato.

Da notare che l'età, non influente nel determinare il possesso dei titoli, ora è significativa, e un aumento della stessa di una decina d'anni porta ad un incremento medio dell'investimento di 171.39 €. Quantitativamente, anche questo effetto è da considerarsi molto debole. L'inserimento dell'area geografica nel modello sembra un po' una forzatura; si può dire però a riguardo che i residenti del centro sono più

disponibili di quelli del nord ad investire in titoli (per la prima volta), mentre il sud, anche se il parametro associato non è significativo, presenta il solito segno negativo.
Il titolo di studio, che si trova in una situazione limite in termini di significatività, denota un effetto più marcato sulla risposta: per la categoria laureati infatti l'ammontare di investimenti aumenta mediamente di 1368.34 € rispetto a chi non posside alcun titolo di studio. Il risultato però, è da prendere con le pinze, proprio per il dubbioso rifiuto del test sulla nullità del parametro associato a *studio*.
Gli effetti perlopiù deboli e poco significativi si ripercuotono sul valore molto basso del coefficiente di determinazione lineare (0.1719); il modello ha una scarsa capacità di adattarsi ai dati.
Si riscontrano comunque delle differenze tra l'influenza delle variabili per modellare l'ammontare e l'influenza delle stesse per determinarne il possesso.
Anche le analisi grafiche di diagnostica del modello ne testimoniano la debolezza, come si può vedere sia dal plot tra residui e valori stimati sia dal qqplot (vedi **Figura 14 e 15**).

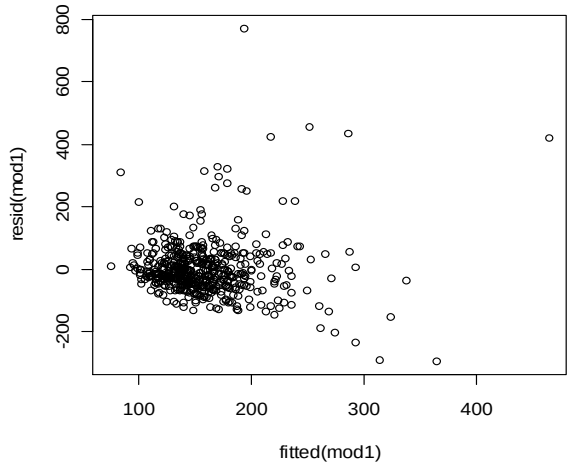

Figura 14, grafico tra valori stimati e residui del modello *mod0*

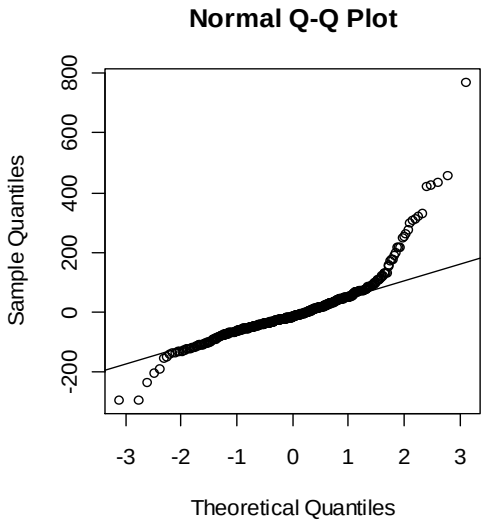

Figura 15, normal qq plot ddel modello *mod0*

Nemmeno delle trasformazioni sui predittori o altre trasformate della variabile risposta (logaritmo) del modello riescono a migliorare la solidità dell ipotesi di omoschedasticità e di normalità dei residui e la bontà di adattamento ai dati.
L'ammontare di risorse investite in tioli di Stato dipende in maniera debole da un numero ristretto delle variabili considerate nella trattazione.

5.3- Modello per l'ammontare di fondi comuni e obbligazioni

Delle dinamiche simili al modello appena analizzato persistono anche nella distribuzione degli investimenti effettuati in fondi comuni e obbligazioni. Si vedano per esempio le correlazioni tra la radice quadrata dell'ammontare investito e le due variabili esplicative quantitative età e reddito.

> cor(reddito,radafd)

[1] 0.3099644

> cor(eta,radafd)

[1] 0.08675601

Le correlazioni sono positive e deboli (la radice quadrata si conferma però la trasformazione migliore), in particolare per quel che riguarda l'età. Applichiamo una regressione lineare multipla con tutte le esplicative all'ammontare investito in fondi comuni e obbligazioni. Da alcune analisi della varianza e tramite la funzione *stepAIC* si può notare come le variabili candidate ad essere escluse dal modello sono sesso e status del lavoratore.
Non si evidenziano infatti differenze in media significative al variare dell'appartenenza del capofamiglia a una diversa categoria di *sesso* e *statuslav*. Eliminando queste due variabili e andando a stimare il modello denominato *mod0* si ottiene il seguente output (la variabile studio si considera numerica perché la sua significatività si ferma alla componente lineare).

lm(formula = radafd ~ eta + reddito + rischio + areageog + studio)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	*-4.401e+01*	*3.569e+01*	*-1.233*	*0.2180*	
eta	*2.002e+00*	*4.794e-01*	*4.177*	*3.40e-05*	***

*reddito 6.861e-04 1.205e-04 5.695 1.96e-08 ****

rischio.L 1.155e+01 2.297e+01 0.503 0.6154

*rischio.Q -4.578e+01 1.772e+01 -2.584 0.0100 **

rischio.C 1.126e+01 1.053e+01 1.069 0.2853

areageog2 1.677e+01 1.072e+01 1.565 0.1182

*areageog3 -2.941e+01 1.410e+01 -2.086 0.0374 **

*studio 1.842e+01 4.491e+00 4.102 4.68e-05 ****

Residual standard error: 98.18 on 586 degrees of freedom

Multiple R-squared: 0.1697, Adjusted R-squared: 0.1584

F-statistic: 14.97 on 8 and 586 DF, p-value: < 2.2e-16

L'eliminazione di *sesso* e *statuslav* permette di ottenere parametri significativi riguardo alle restanti variabili e l'adattamento del modello ai dati, pur restando abbastanza scarso, è migliore che nei precedenti modelli per l'ammontare delle risorse investite. Il coefficiente di determinazione lineare vale 0.1697 e si noti come lo stesso indice "corretto" sia migliore rispetto al modello considerato con tutte le variabili (0.1584 contro 0.1538). Il coefficiente di determinazione lineare corretto si impiega per confrontare la bontà di modelli differenti, anche applicati alle stesse osservazioni, ma che coinvolgono un differente numero di variabili; il suo utilizzo si giustifica in quanto l'indice di determinazione lineare non può mai diminuire, anche se viene introdotta una variabile esplicativa totalmente non significativa (*Bracalente, 2009*).

Ora analizziamo gli effetti dei regressori sull'ammontare di risorse che le famiglie destinano a fondi comuni e obbligazioni. Un aumento del reddito di 50.000 € provoca un incremento medio degli investimenti di 1176.83 €; i residenti del sud investono risorse finanziarie mediamente inferiori di 8435.93 € mentre per il nord e il centro non si rilevano particolari effetti (entrambi i coefficienti non significativi); un invecchiamento di dieci anni dell'età del capofamiglia produce un aumento dell'investimento di 400.96 € (effetto debole come testimonia la bassa correlazione tra l'età e la variabile dipendente) e infine riguardo allo studio si nota come i laureati

investano mediamente 8482.66 € in più delle altre categorie. Il rischio è significativo nella sola componente quadratica e la sua analisi non è di facile interpretazione. Tutte le conclusioni riguardo agli effetti dei regressori sono effettuate al netto delle altre variabili esplicative presenti nel modello.

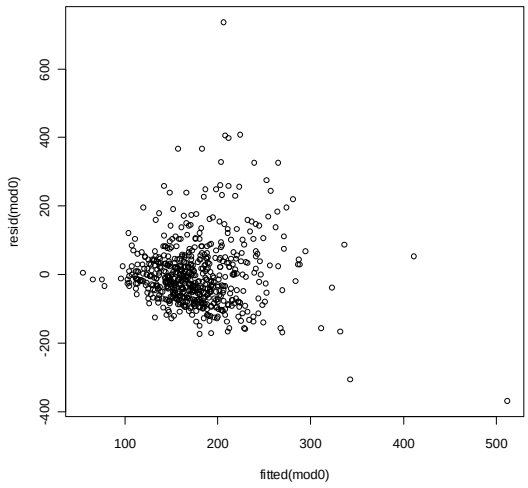

Figura 16, grafico tra valori stimati e residui del modello

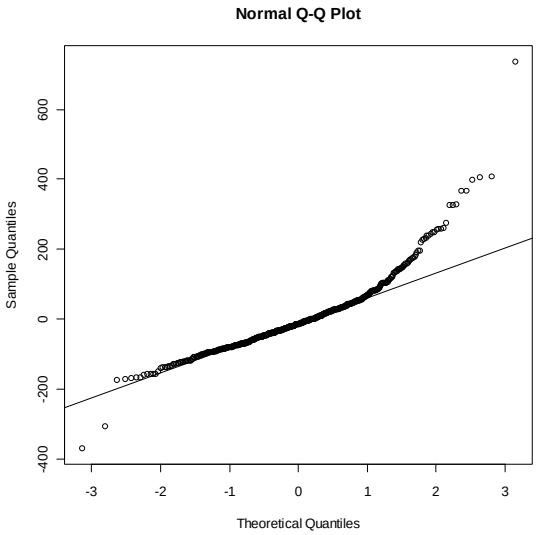

Figura 17, normal qqplot del modello

Le **Figure 16 e 17**, utili in fase di diagnostica del modello e quindi di verifica delle ipotesi su cui è basato, mostrano come l'omoschedasticità e la linearità non siano pienamente rispettate ma ci si può accontentare (preoccupa la coda destra del qqplot che mostra un andamento sistematico di allontanamento dalla normalità). Anche riguardo a questo aspetto la trasformata logaritmica non risolve i problemi diagnosticati.

5.4- Modello per l'ammontare di azioni e quote di società

L'investimento medio in azioni e quote di società è di 48670.71 €, quindi si tratta di investimenti di un certo peso. Si tratta infatti di attività finanzarie volte alla ricerca di guadagni sperati piuttosto elevati. Dopo una prima analisi esplorativa che conferma la usuale relazione crescente tra reddito e ammontare degli investimenti si va a calcolare la correlazione che sussiste tra queste due quantità (l'ammontare è sempre espresso sotto radice quadrata).

> cor(reddito,radafe)

[1] 0.4173831

Le variabili sono correlate tra loro e dunque sarà logico aspettarsi che le famiglie con redditi più elevati siano quelle che investono un maggior numero di risorse. L'età invece, che nei precedenti modelli aveva una correlazione molto debole con le variabili risposte, in questo caso quasi si annulla: probabilmente non influirà sulla variabile dipendente.

> cor(eta,radafe)

[1] 0.03851096

La stima del modello che considera tutte le variabili produce i risultati che seguono; si tenga sempre conto del fatto che la variabile risposta è una trasformata della variabile originaria (radice quadrata) e che altre trasformazioni solitamente utili non forniscono risultati migliori.

mod0<-lm(formula = radafe ~ sesso + studio + statuslav + rischio + eta + reddito + areageog)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*1.696e+02*	*7.459e+01*	*2.273*	*0.023728 **
sesso2	*-1.107e+01*	*2.055e+01*	*-0.539*	*0.590434*
studio.L	*-1.014e+02*	*4.200e+01*	*-2.413*	*0.016432 **
studio.Q	*8.509e+01*	*3.645e+01*	*2.334*	*0.020246 **
studio.C	*-6.724e+01*	*2.580e+01*	*-2.606*	*0.009616 ***
studio^4	*3.780e+01*	*1.795e+01*	*2.106*	*0.036071 **
statuslav2	*1.542e+01*	*4.148e+01*	*0.372*	*0.710276*
statuslav3	*6.488e+00*	*4.474e+01*	*0.145*	*0.884814*
statuslav4	*9.994e+00*	*4.385e+01*	*0.228*	*0.819882*
statuslav5	*-2.949e+01*	*5.095e+01*	*-0.579*	*0.563175*
statuslav6	*3.897e+01*	*3.925e+01*	*0.993*	*0.321589*
statuslav7	*5.105e+01*	*4.872e+01*	*1.048*	*0.295634*
rischio	*-3.458e+01*	*1.001e+01*	*-3.455*	*0.000632 ****
eta	*3.106e-01*	*1.028e+00*	*0.302*	*0.762672*

*reddito 1.556e-03 2.146e-04 7.250 3.7e-12 ****

areageog2 1.315e+01 1.890e+01 0.696 0.487241

areageog3 -2.703e+01 2.450e+01 -1.103 0.270862

Residual standard error: 120.8 on 294 degrees of freedom

Multiple R-squared: 0.2481, Adjusted R-squared: 0.2071

F-statistic: 6.062 on 16 and 294 DF, p-value: 1.5e-11

Si è deciso di valutare l'effetto di tutte le variabili proprio per testimoniare la particolarità di questo modello in cui si riscontra la scarsa influenza di età, area geografica, sesso e status del lavoratore. Assumono invece fondamentale importanza la propensione al rischio (nella sola componente lineare) e il titolo di studio (significativo in tutte le sue componenti).

L'eliminazione delle variabili sopra definite come non in grado di aggiungere informazioni sull'ammontare degli investimenti in azioni produrrebbe un miglioramento del modello nel senso che si otterrebbero tutti i parametri significativi ed un aumento del coefficiente di determinazione lineare corretto (che permette il confronto tra bontà di modelli con numero diverso di variabili).

Il reddito svolge, come ci si aspettava, un ruolo decisamente di prim'ordine in questo caso: un aumento del reddito di 50.000 € fa investire mediamente 6053.24 € in più a parità di altre condizioni.

*>(coef(mod0)[15]*50000)^2*

[1] 6053.237

Una nota particolare va spesa riguardo alla variabile categoriale ordinale titolo di studio, per la quale si mette in luce un trend stimato da un polinomio di quarto grado; questo perché si nota che gli investitori senza nessun titolo di studio sono solamente quattro e hanno investito mediamente 187160.8 €. La "povertà" numerica di questa categoria porta dunque a risultati che potrebbero essere fuorvianti, nel senso che queste quattro famiglie potrebbero rappresentare dei casi a parte, degli outliers. Se si analizzano le altre categorie infatti, si può notare come gli investimenti medi crescano

al crescere del titolo di studio, in maniera non lineare (sembra sufficiente un trend cubico per valutarne la relazione).

Tramite le analisi grafiche e il test specifico di Bonferroni che verifica la presenza di outliers abbiamo la conferma della stranezza dei valori osservati prima descritti.

> outlierTest(mod0)

	rstudent	*unadjusted p-value*	*Bonferonni p*
100	*6.735435*	*8.6369e-11*	*2.6861e-08*
232	*4.854782*	*1.9634e-06*	*6.1062e-04*
183	*3.855698*	*1.4189e-04*	*4.4127e-02*
297	*3.841457*	*1.4996e-04*	*4.6636e-02*

Tre di questi quattro valori identificati si riferiscono a famiglie con capofamiglia senza alcun titolo di studio; si possono notare bene anche dalla **Figura 18** (nonostante non escano dalle bande della distanza di Cook e quindi non influiscano pesantemente sull'accuratezza del modello). Una delle soluzioni possibili in questi casi potrebbe essere la rimozione delle osservazioni ritenute anomale.

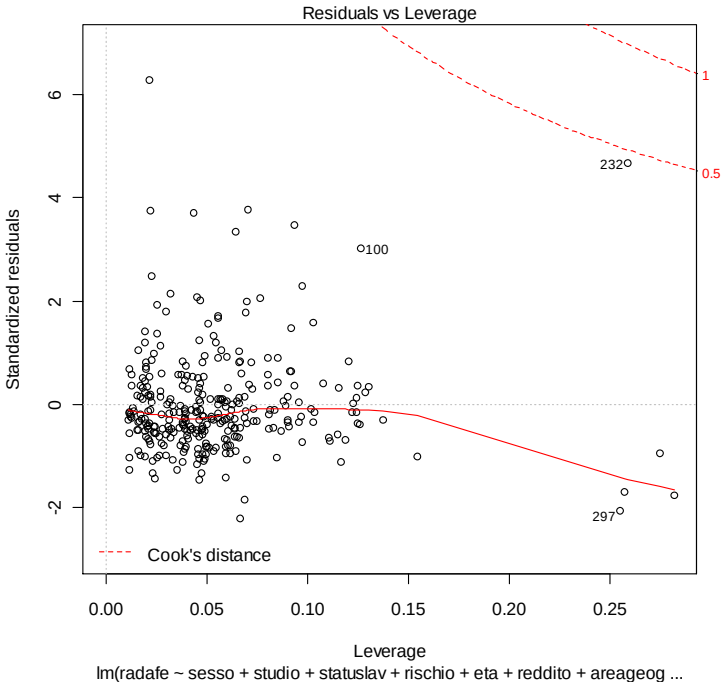

Figura 18, leverage versus standardized residuals

Il modello si adatta discretamente ai dati e “spiega” il 24.81% della variabilità; le analisi grafiche dei residui forniscono risultati non ottimali ma da considerarsi sufficienti ai fini dell’attendibilità delle conclusioni tratte (la variabiltà dei residui presenta un lieve trend e la coda destra del qqplot devia sistematicamente dalla distribuzione normale).

5.5- Modello per l’ammontare delle gestioni patrimoniali

L’investimento medio in gestioni patrimoniali corrisponde ad un ammontare di risorse pari a 71805.72 €. Sono perciò investimenti rilevanti e come si è già detto, caratterizzati da un elevato rischio ma anche da alte prospettive di guadagno.
La particolarità delle gestioni patrimoniali che si era evidenziata nell’analizzare il possesso o meno dell’investimento, emerge anche nella stima del modello che descrive il rispettivo ammontare di risorse impiegato. Notiamo infatti come il reddito sia molto correlato con la variabile risposta.

> cor(reddito,af.f)

[1] 0.7121441

Le restanti variabili invece, sia che si considerino dei modelli “intermedi”, sia che consideri il modello stimato con tutte le variabili esplicative, rimangono non significative e non aggiungono ulteriore informazione, come era accaduto anche per il possesso/non possesso delle gestioni patrimoniali. Anche l’uso della funzione *stepAIC* ci porta alle stesse conclusioni: l’unica variabile di cui ha senso tenere conto in questo caso è il reddito disponibile netto. Di seguito è riportato l’output delle iterazioni dell’algoritmo utilizzato dalla funzione stepAIC, che all’ultimo passo stima il modello con il solo reddito funzione dell’ammontare investito in gestioni patrimoniali.

> stepAIC(mod0)
Start: AIC=1672.14
af.f ~ reddito + eta + sesso + statuslav + areageog + studio + rischio

```
           Df  Sum of Sq        RSS    AIC
- statuslav  5 2.3503e+10 6.3535e+11 1664.9
- areageog   2 8.4905e+09 6.2034e+11 1669.1
- sesso      1 1.3804e+08 6.1199e+11 1670.2
- rischio    1 9.1236e+08 6.1276e+11 1670.2
- studio     1 2.8109e+09 6.1466e+11 1670.5
- eta        1 1.3154e+10 6.2500e+11 1671.7
<none>                   6.1185e+11 1672.1
- reddito    1 2.5234e+11 8.6419e+11 1695.0
Step:  AIC=1664.86
af.f ~ reddito + eta + sesso + areageog + studio + rischio
           Df  Sum of Sq        RSS    AIC
- areageog  2 5.7957e+09 6.4115e+11 1661.5
- sesso     1 6.1872e+08 6.3597e+11 1662.9
- studio    1 5.0614e+09 6.4041e+11 1663.4
- rischio   1 5.9993e+09 6.4135e+11 1663.5
- eta       1 8.0256e+09 6.4338e+11 1663.8
<none>                   6.3535e+11 1664.9
- reddito   1 5.6753e+11 1.2029e+12 1708.8
Step:  AIC=1661.51
af.f ~ reddito + eta + sesso + studio + rischio
          Df  Sum of Sq        RSS    AIC
- sesso    1 3.7659e+08 6.4153e+11 1659.5
- studio   1 4.2887e+09 6.4544e+11 1660.0
- rischio  1 4.4311e+09 6.4558e+11 1660.0
- eta      1 1.0771e+10 6.5192e+11 1660.7
<none>                   6.4115e+11 1661.5
- reddito  1 5.7636e+11 1.2175e+12 1705.7

Step:  AIC=1659.55
```

af.f ~ reddito + eta + studio + rischio

Df Sum of Sq RSS AIC

- studio 1 4.6584e+09 6.4618e+11 1658.1

- rischio 1 5.0984e+09 6.4662e+11 1658.1

- eta 1 1.1079e+10 6.5260e+11 1658.8

<none> 6.4153e+11 1659.5

- reddito 1 5.7926e+11 1.2208e+12 1703.9

Step: AIC=1658.07

af.f ~ reddito + eta + rischio

Df Sum of Sq RSS AIC

- rischio 1 6.6457e+09 6.5283e+11 1656.8

- eta 1 1.3112e+10 6.5930e+11 1657.5

<none> 6.4618e+11 1658.1

- reddito 1 6.4759e+11 1.2938e+12 1706.1

Step: AIC=1656.81

af.f ~ reddito + eta

Df Sum of Sq RSS AIC

- eta 1 1.2396e+10 6.6522e+11 1656.2

<none> 6.5283e+11 1656.8

- reddito 1 6.7186e+11 1.3247e+12 1705.8

Step: AIC=1656.16

af.f ~ reddito

Df Sum of Sq RSS AIC

<none> 6.6522e+11 1656.2

- reddito 1 6.8452e+11 1.3497e+12 1705.1

Call:

lm(formula = af.f ~ reddito)

Coefficients:

(Intercept) reddito

-30202.826 1.511

Il modello che si aasume per l'ammontare impiegato nelle gestioni patrimoniali fornisce quindi il seguente output di regressione, dove però si è optato per inserire il reddito come polinomio di quarto grado. Inoltre la variabile risposta non ha subito alcuna trasformazione ed esprime direttamente l'effettivo ammontare investito (eventuali trasformate non portavano a significativi miglioramenti del modello stimato).

mod2<-lm(formula = af.f ~ poly(reddito, 4))

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*71806*	*6471*	*11.097*	*< 2e-16 ****
poly(reddito, 4)1	*827360*	*54905*	*15.069*	*< 2e-16 ****
poly(reddito, 4)2	*333963*	*54905*	*6.083*	*6.37e-08 ****
poly(reddito, 4)3	*467723*	*54905*	*8.519*	*2.80e-12 ****
poly(reddito, 4)4	*364628*	*54905*	*6.641*	*6.61e-09 ****

Residual standard error: 54900 on 67 degrees of freedom

Multiple R-squared: 0.8504, Adjusted R-squared: 0.8414

F-statistic: 95.19 on 4 and 67 DF, p-value: < 2.2e-16

La stima di un trend polinomiale fino al quarto grado è supportata dal risultato della funzione *anova,* che mettendo a confronto il modello con la relazione lineare e quello con il polinomio, porta a rifiutare il modello con meno parametri.

> anova(mod1,mod2)

Analysis of Variance Table

Model 1: af.f ~ reddito

Model 2: af.f ~ poly(reddito, 4)

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	*Pr(>F)*
1	*70*	*6.6522e+11*				
2	*67*	*2.0198e+11*	*3*	*4.6325e+11*	*51.224*	*< 2.2e-16 ****

Nonostante sia una sola variabile che descrive il comportamento delle risorse investite nell'attività finanziaria considerata in questo paragrafo, l'adattamento del modello ai dati è sorprendente.

Il coefficiente di determinazione lineare vale 0.8504, la bontà del modello è quasi ottimale dal punto di vista della varianza spiegata. In riferimento alle ipotesi sottostanti al modello invece, le analisi grafiche sono molto meno confortanti di quanto si è visto sinora (vedi **Figura 19 e 20**).

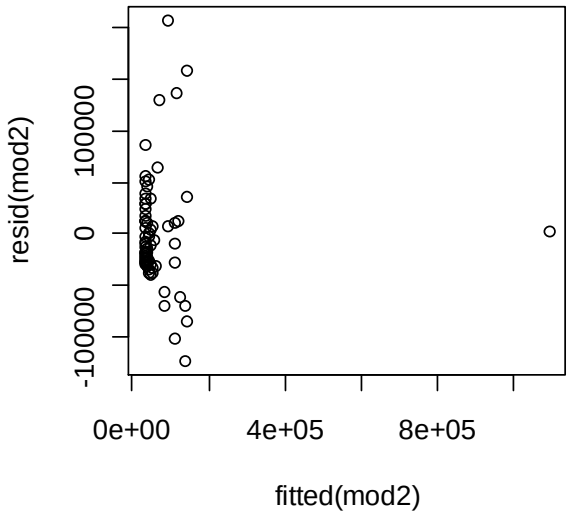

Figura 19, grafico tra valori stimati e residui del modello *mod2*

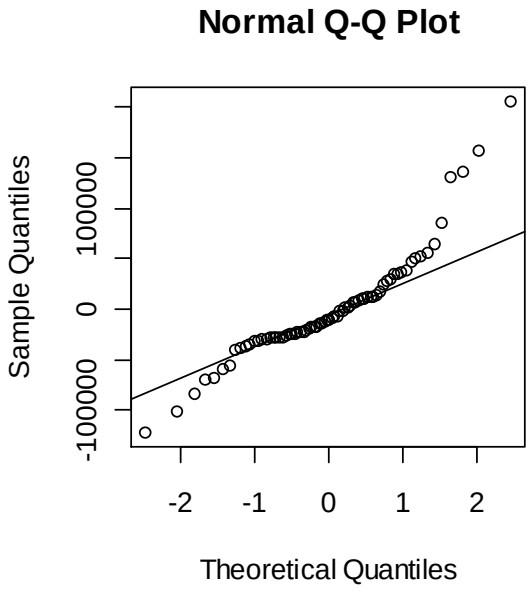

Figura 20, normal qqplot del modello *mod2*

Sono poco rispettate sia l'ipotesi di omoschedasticità sia l'ipotesi di normalità dei residui, ed eventuali trasformate non migliorano entrambi gli aspetti.

Si rammenti che la bontà di adattamento di questo modello, e anche le analisi grafiche, soffrono della bassa numerosità campionaria (il numero di investitori in gestioni patrimoniali è di sole 72 unità).

6- CONCLUSIONI

Le analisi svolte in questa trattazione dimostrano innanzitutto la tendenza degli Italiani ultracinquantenni ad investire poco in attività finanziarie rischiose, o comunque alternative ai depositi bancari e postali a c/c o a risparmio. Poco in riferimento alla percentuale di investitori sul totale del campione; quantitativamente invece gli investimenti effettuati, soprattutto per le azioni e le getioni patrimoniali hanno valori medi elevati e quindi un certo peso.

Le stime dei modelli a risposta binaria hanno portato a conclusioni molto interessanti dal punto di vista della corretta classificazione dei dati; si è riusciti dunque, attraverso le variabili utilizzate a discriminare tra potenziali investitori e non. Diversamente, i modelli finalizzati a descrivere le relazioni tra le variabili e il rispettivo ammontare investito nelle varie attività finanziarie, descrivono poco del fenomeno, probabilmente troppo complesso e che necessiterebbe l'inclusione di ulteriori fattori.

Emerge chiaramente un filo conduttore in questo lavoro, e cioè l'importanza del reddito: questa variabile risulta essere la più determinante in riferimento a tutti gli aspetti analizzati e ci mostra come una forte disponibilità di risorse permette di investire e con somme di una certa entità. In particolare è la sola caratteristica che influisce sul possesso e sulle quantità investite nelle gestioni patrimoniali.

L'area geografica di residenza ha una forte capacità discriminante per ciò che concerne i depositi bancari e postali: al nord i risparmi sono considerevoli rispetto al sud ma molto simili al centro Italia. Lo stesso tipo di effetto è ravvisabile negli investimenti in titoli di Stato. I potenziali investitori in fondi comuni, obbligazioni e azioni sono fortemente condizionati al possesso di un titolo di studio elevato (diploma e laurea); dal crescere dell'età del capofamiglia (negativamente), e dal sesso (le donne sono meno disposte ad investire). Gli effetti non sono però più ravvisabili se si fa riferimento all'ammontare di risorse investito, ad eccezione del titolo di studio, che mantiene la sua significatività.

Le categorie di dirigenti, imprenditori e pensionati generalmente tendono a possedere attività finanziarie quali i titoli di Stato e le obbligazioni più degli altri gruppi di

lavoratori e dei disoccupati. Anche questo effetto viene a mancare se si passa ai modelli per l'ammontare investito.

Le azioni dipendono in maniera sostanziale anche dalla propensione al rischio: tendono ad investire infatti solamente gli individui con alta propensione al rischio e contemporanee alte prospettive di guadagno.

Gli enti, le istituzioni e le aziende che trattano le attività finanziarie degli Italiani potranno disporre di questi strumenti in modo da individuare i potenziali investitori e il loro peso come clienti.

Definire un potenziale investitore in relazione a delle sue caratteristiche specifiche e influenti può essere utile al fine di creare delle strategie di marketing volte a proporre politiche di differenziazione di offerta (per esempio riguardo al tasso di interesse).

Questo approccio si giustifica in un'ottica che vede il cliente, nel nostro caso l'investitore, al centro del circuito del *relationship marketing*; il che comporta il passaggio a relazioni di tipo *one to one* con il cliente. Con ogni investitore si instaura un rapporto specifico e differenziato in relazione alle sue caratteristiche, con l'obiettivo di puntare alla sua fidelizzazione se già acquisito. Inoltre, sempre secondo questo approccio, la storia del cliente e il suo rapporto con l'impresa vengono memorizzati nel *customer data base*, allo scopo di monitorare eventuali cambiamenti nel tempo (*Grandinetti, 2008*).

BIBLIOGRAFIA

- A. Azzalini. *"Inferenza statistica: una presentazione basata sul concetto di verosimiglianza"* (2001, Springer Italia, Milano).
- F. Bassi. *"Analisi di mercato. Strumenti e statistiche per le decisioni di marketing"* (2008, Carocci, Roma).
- B. Bracalente, M. Cossignani, A. Mulas. *"Statistica aziendale"* (2009, Mc Graw-Hill, Milano).
- F. Cerbioni, L.Cinquini, U. Sostero. *"Contabilità e bilancio"* (2006, Mc Graw-Hill, Milano).
- R. Grandinetti. *"Marketing. Mercati, prodotti e relazioni"* (2008, Carocci, Roma).
- L. Pace, A. Salvan. *"Introduzione alla Statistica"* (2001, Cedam, Padova)

SITOGRAFIA

- www.bancaditalia.it
- www.unioneconsulenti.it