Università degli Studi di Padova
Corso di Laurea in Statistica, Economia e Finanza

Tesi di laurea:

INDAGINE SULLA FREQUENZA DEI CONTROLLI NEGLI AUTOBUS: REALTA' E ASPETTATIVE DEI PASSEGGERI

Investigation about the frequency of checks in buses: realities and expectation of passengers

Relatore: Prof. Alessandro Bucciol
Dipartimento di Scienze Economiche

Laureanda: Camilla Lincetto

Anno Accademico 2012/2013

Al relatore Alessandro Bucciol

per la sua disponibilità e competenza.

Ai miei genitori e ai miei nonni

a cui devo tutto quel che sono,

senza di loro non sarei arrivata fin qui.

A mia sorella che sa sempre come spronarmi.

Al mio Davide che mi ha sempre sostenuta con amore e

mi ha dato la determinazione per non arrendermi mai.

A chi mi vuole bene.

INDICE

CAPITOLO 1

IL COMPORTAMENTO FRAUDOLENTO

1.1 Obiettivo dello studio

L'obiettivo del presente elaborato è quello di analizzare il comportamento delle persone nella quotidiana attività di utilizzo di un mezzo pubblico a Reggio Emilia. L'essere umano si trova costantemente di fronte ad una scelta in qualsiasi azione egli compie: essere onesto o disonesto. Per indagare da un punto di vista scientifico questo dilemma, lo studio condotto analizza un campione di viaggiatori delle linee locali degli autobus di Reggio Emilia. Si è proceduto alla raccolta di dati riguardanti coloro i quali avessero appena viaggiato su una delle linee 1-13. Le interviste sono state condotte in tre fermate: "viale Allegri" (situato in centro, in una zona a traffico limitato, e punto in cui fermano molte linee), l'"Ospedale" (luogo dove le persone hanno bisogno di andare per necessità) e la "Stazione FS" (snodo fondamentale per raggiungere tutte le zone della città con le molte linee che vi fermano).

1.2 Il comportamento fraudolento in generale

E' molto difficile definire in modo sufficientemente chiaro ed esaustivo un comportamento non etico, essendo tale argomento mutevole ed interpretabile. Possiamo definire comportamento etico un insieme di principi di comportamento o di azione condotta in accordo con la morale o l'equità, e che

rispetti i principi di giustizia e di onestà. Decisioni non etiche sono purtroppo realtà quotidiane in tutte le culture, la storia e in ogni ambito della vita, a partire dalle relazioni personali all'economia, passando per la giustizia, la sanità, eccetera.

Un aspetto molto interessante riguarda la consapevolezza che i comportamenti non etici non sono equamente distribuiti tra la popolazione ma, come è stato dimostrato, ciò non dipende dallo stato sociale. Secondo una serie di studi in diversi ambiti si è rilevata essere proprio la classe sociale più alta a presentare una maggior frequenza di comportamenti non etici.

Secondo gli autori, questo fatto può essere spiegato da fattori sia strutturali sia psicologici: l'indipendenza dagli altri comporta una minore esposizione ai rischi di comportamenti inappropriati, mentre tra i fattori psicologici assume un ruolo predominante il desiderio di affermazione di sé il quale può precludere la percezione degli effetti delle azioni condotte.

Facendo riferimento al principio di onestà o comportamenti non etici in campo economico sono ormai termini di uso comune frodi, fallimenti, paradisi fiscali e l'evasione fiscale. Questi comportamenti non etici, in particolare modo l'evasione fiscale, sono sempre stati uno dei temi rilevanti delle società moderne. Soprattutto in un periodo di crisi come quello che stiamo vivendo questi problemi diventano letteralmente un'emergenza nazionale e una priorità assoluta nell'agenda dei governi. Proprio in tal senso è interessante riconoscere con quale velocità il nuovo governo italiano sta agendo, guidato dal primo ministro Mario Monti, il quale il 16 novembre 2011, ha adottato in Italia misure di austerità finalizzate a migliorare rapidamente le condizioni economiche italiane e soprattutto a recuperare la fiducia dei mercati verso il nostro Paese.

Di tutte le manovre messe in atto, probabilmente le azioni più spettacolari e anche più redditizie sono una serie di controlli fiscali nelle più importanti città d'Italia. E' stata costituita una task force nei confronti degli evasori fiscali: cittadini, esercizi commerciali, società, etc. Da questi controlli è emerso che il 50% di negozi o attività commerciali non emette alcuna ricevuta, raggiungendo il picco del 80% in alcuni distretti. Questo aumento esponenziale dei controlli

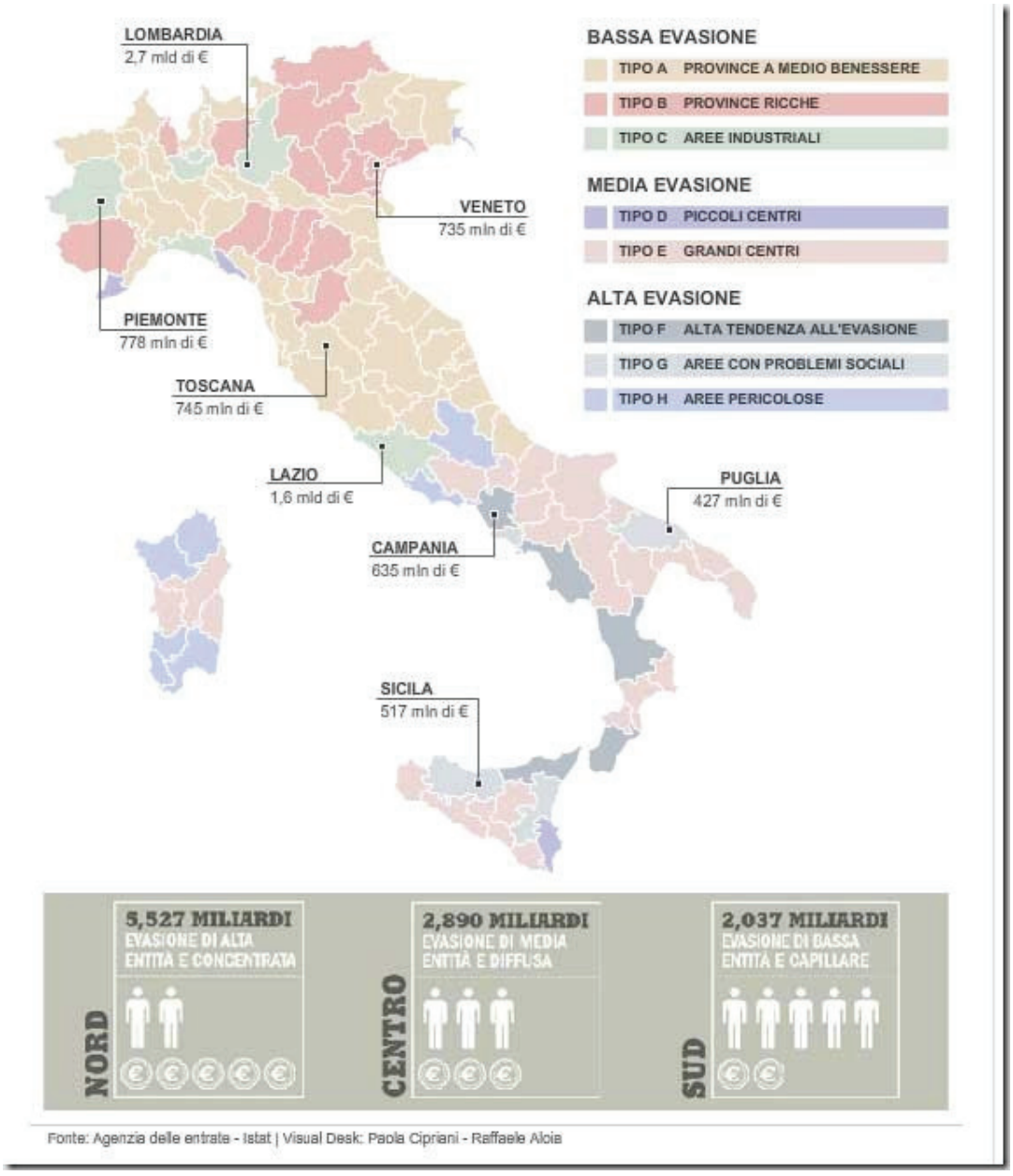
fiscali si è rivelato avere un doppio effetto: da un lato ha un effetto diretto ed immediato in quanto l'evasore può essere colto sul fatto, in secondo luogo lancia un forte messaggio intimidatorio a tutti cioè che i controlli sono efficaci ed in corso di attuazione. Quest'ultimo aspetto è fondamentale poiché la certezza del controllo fiscale può avere un effetto a lungo termine sulla situazione economica di un Paese. Se il controllo fiscale è realmente applicato e al tempo stesso la pena inflitta è abbastanza forte, allora l'evasione fiscale viene realmente scoraggiata. Una strategia adottata dal nuovo governo per implementare il controllo fiscale, è un minore impiego di denaro liquido sostituito in maniera maggiore da un più ampio utilizzo dei pagamenti elettronici; infatti, le transazioni superiori a 1000 euro si possono effettuare o tramite assegno non trasferibile, o bonifico, o con la carta di credito. Tutto ciò è dovuto alla consapevolezza che l'utilizzo di denaro contante facilita l'evasione, rendendo possibile una dichiarazione non reale del reddito. Con i suddetti pagamenti, invece, ciò non è possibile perché le transazioni sono facilmente tracciabili.

La situazione finanziaria che l'Italia si trova ad affrontare, come molti altri Paesi, è strettamente correlata con il fenomeno dell'evasione fiscale. E' convinzione comune che un evasore fiscale comporta una perdita per tutta la comunità, o che "i pagamenti fiscali sono un diritto ed un obbligo per tutti ". Dunque è necessario, soprattutto in periodi di crisi quale quello che stiamo vivendo, che ognuno dia il proprio contributo con l'obiettivo di ripristinare la situazione finanziaria del Paese in toto. Il Governo Italiano ha trasmesso attraverso i *mass media* spot pubblicitari finalizzati a dissuadere le persone dall'evasione e addirittura su Internet si possono facilmente trovare siti a cui tutti possono accedere per denunciare le frodi appena avvenute. L'uso di questi siti è estremamente semplice e, grazie alla collaborazione di Google Maps, è possibile segnalare e specificare il nome del business, nonché la data, l'ora, l'indirizzo e la quantità evasa e visualizzare il risultato su una mappa di Google. Siti quali www.evasory.info sono destinati a servire lo scopo indicato, mentre altri, come www.nonevado.it hanno lo scopo di segnalare le imprese che, al contrario,

hanno dimostrato di emettere ricevute. In entrambi i casi, l'obiettivo finale è lo stesso: distinguere i comportamenti non etici da quelli etici. Questi siti sono diventati molto popolari in un tempo molto breve. In particolare, il primo sito citato, permette di riferire l'illecito direttamente alle autorità competenti.

Questi progetti, non creati da agenzie governative, ma da cittadini comuni, sono un chiaro segnale di consapevolezza della gravità e delle implicazioni nella società del fenomeno dell'evasione fiscale.

Secondo l'agenzia delle entrate, l'evasione fiscale interessa tutte le regioni italiane con un andamento rappresentato nella seguente figura:



Da quanto emerge da una nuova indagine effettuata per conto di "Contribuenti.it Magazine" dell'Associazione Contribuenti Italiani, con un'evasione fiscale in crescita del 14,1% nei primi 6 mesi del 2012, l'Italia si conferma al primo posto in Europa con un economia sommersa del 21% del prodotto interno lordo pari a 340 miliardi di euro l'anno, essa è seguita da Grecia (20,8%) e Romania (19,1%). Nell'indagine si precisa che le imposte sottratte all'erario sono nell'ordine dei 180,9 miliardi di euro l'anno conteggiando sia quelle dirette sia indirette. L'economia sommersa dell'Italia è risultata circa il doppio di quella della Francia e della Germania.
Al fine di migliorare la situazione finanziaria, altro strumento nelle mani del governo è l'aumento della tassazione. Secondo il rapporto di Confcommercio, l'organismo di rappresentanza delle imprese impegnate nel commercio, nel turismo e nei servizi, l'Italia è diventata nel 2012 il quinto paese con il più alto tasso fiscale del mondo, raggiungendo il 45,2%.
In un'ottica di crescita del Paese è chiaro che un aumento delle tasse possa essere soltanto una soluzione temporanea, poiché comporta serie conseguenze come un globale impoverimento con il seguente calo dei consumi e la perdita di posti di lavoro. Potrebbe inoltre innescare un circolo vizioso, spingendo più persone ad evadere le tasse ritenendole eccessive.

Riguardo ai così detti "paradisi fiscali", essi sono Stati con una esigua tassazione sul reddito (con ciò si intendono anche i depositi bancari) e sui consumi. Ciò attira molto capitale proveniente da altri Paesi con imposte più elevate, fornendo in cambio una richiesta di contribuzione estremamente favorevole. Dal punto di vista del contribuente, per riportarci all'originaria definizione statunitense di paradiso fiscale, "tax haven", si tratta di un rifugio dall'alta tassazione sui redditi che spesso si accompagna a regimi in cui è fortemente tutelato e garantito l'anonimato dei proprietari dei capitali e delle società che li possiedono.
Ci sono molti paradisi fiscali generalmente accomunati da:

- Mancanza di effettivo scambio di informazioni fiscali con le autorità fiscali

estere.

- Mancanza di trasparenza nel funzionamento delle disposizioni legislative, giuridiche o amministrative.
- Nessun requisito di una presenza sostanziale locale.
- Auto-promozione come centro finanziario offshore.
- Tassazione scarsa e a volte nulla.

Il motivo per cui si potrebbe essere interessati ad aprire un conto in un paradiso fiscale non riguarda solo ragioni economiche, ma soprattutto motivi di privacy, essendo il segreto bancario applicato rigorosamente. Per un estraneo, è praticamente impossibile venire a conoscenza dei movimenti di capitale. Il governo italiano precedente ed attuale ha provato a fare alcuni accordi con i paradisi fiscali vicini come la Svizzera e San Marino, al fine di porre un maggior controllo in tale ambito.

Ora più che mai, si registra anche un significativo aumento di comportamenti fraudolenti sia a discapito delle assicurazioni, per richieste di danni fisici e/o morali inesistenti o fin troppo pretestuosi, sia a discapito di enti come l'INPS e addirittura dello Stato Italiano per quanto riguarda l'assegnazione delle pensioni di invalidità. Si stima, infatti, che in Italia il 30% del totale dei beneficiari non ne avrebbe diritto. L'INPS ne risente dal punto di vista della quantità di pensionati cui assegnare l'accompagnatoria, mentre lo Stato non percepisce le tasse in quanto tali pensioni sono esenti ed, inoltre, i finti invalidi usufruiscono di benefici di vario genere a discapito dei veri bisognosi.

Altro comportamento non etico diffuso è sicuramente l’infedeltà. Si assiste ad una crescita esponenziale di questo problema. Ciò è stato correlato all’evoluzione ed al cambiamento della società. Per esempio la diffusione di moderne tecnologie come telefoni cellulari e i social network rivestono un ruolo importante nel fenomeno del tradimento. Questi strumenti permettono di semplificare e moltiplicare la comunicazione e le possibili corrispondenze, ma

hanno anche cambiato il modo di tradire. In particolare, i social network hanno rivoluzionato la materia in quanto permettono di stabilire una discussione riservata ed eliminare qualsiasi barriera fisica.

I *social network* (legati al 66% dei divorzi) si rivelano essere sia l'evento scatenante sia causa una relazione extraconiugale sia anche la modalità con cui viene scoperto il tradimento[1].

In Italia, sette uomini su dieci hanno tradito il loro rispettivo partner di cui il 66% con un collega di lavoro[2]. La percentuale riguardante le donne è sovrapponibile a quella del genere maschile. L'Infedeltà è ormai così comune che ora esistono reti sociali apposite per esperienze extraconiugali. Detto questo, non sorprende che in Italia il 10% dei neonati non sono geneticamente correlati al padre[3].

Altro ambito in cui è sempre più frequente avere notizia di comportamenti non etici è lo sport. Dagli illeciti sportivi al crescente uso di doping si sta perdendo il suo vero significato. C'è una nozione che sembra costituire il cuore dell'etica dello sport, la nozione di fair play. E' difficile tradurre questo termine e spesso anche i regolamenti italiani delle discipline sportive preferiscono riportarlo in inglese. Non è neanche facile darne una definizione esaustiva: denota, oltre che un valore, una sorta di atteggiamento mentale fondamentale, il "giusto spirito" con cui praticare lo sport. Si evince ciò dalla Dichiarazione sul Fair Play del Consiglio internazionale dello sport e dell'educazione fisica, del 1976, fatta propria dal CIO (International Olimpic Committee), che ne ha dato la seguente caratterizzazione: a) onestà, franchezza e atteggiamento fermo e dignitoso verso chi non si comporta con fair play; b) rispetto per i compagni di squadra; c) rispetto per gli avversari, sia quando vincono, sia quando perdono, con la consapevolezza che l'avversario è un partner necessario; d) rispetto per gli arbitri, mostrato attraverso l'effettivo sforzo di collaborare con loro.

[1] http://www.citydiscount.it/social-network-invadono-la-coppia-il-tradimento-su-facebook/
[2] http://www.universy.it/2011/03/italia-record-di-tradimenti-i-colleghi-sono-gli-sfasciafamiglie/
[3] http://lnx.papaseparati.org/psitalia/genetica-forense/il-10-dei-bambini-italiani-non-figlio-del-pap-presunto.html

Ci troviamo al giorno d'oggi ad osservare il mondo sportivo invaso da meri interessi economici che oscurano il vero significato intrinseco dell'attività sportiva. L'aumento degli illeciti collegati alle scommesse sportive e al riciclo di capitali è esponenziale e l'immagine sana dello sport sembra un lontano ricordo. Ogni giorno si legge sui giornali di arresti per frode sportiva di atleti o manager del settore che senza scrupoli hanno aggirato leggi e regolamenti lucrando su tifosi e appassionati. Per queste persone conta solo guadagnare e vincere ricorrendo a qualsiasi mezzo come anche all'uso di sostanze dopanti. L'etimologia del termine Doping presumibilmente risale al "dop", sostanza alcolica assunta dai guerrieri zulu per eccitarsi prima della battaglia. Da ciò il termine "doping" che, almeno nell'accezione del mondo sportivo, significa "l'uso improprio di sostanze o metodi atti ad aumentare artificialmente le prestazioni fisiche mediante l'incremento delle masse muscolari o della resistenza alla fatica". Nel 2000 la legge 376 ("Disciplina della tutela sanitaria delle attività sportive e della lotta contro il doping") ha esteso tale formulazione ai " farmaci, sostanze e pratiche idonee a modificare la condizioni psicofisiche o biologiche dell'organismo al fine di alterare le prestazioni agonistiche degli atleti". (Presidenza del Consiglio dei Ministri, COMITATO NAZIONALE PER LA BIOETICA, ETICA SPORT E DOPING, 2010)

Fonte: "The Complete Book of the Olympics", Wallechinsky and Loucky; the Economist

Dunque in una società come l'attuale chi può dire di non aver mai assistito ad un comportamento fraudolento di qualsiasi genere?
Si pensi a quante volte, ad esempio, si è assistito ad un controllo dei biglietti su un mezzo di trasporto pubblico; a chi viaggia abitualmente su treni o autobus sarà sicuramente capitato che i controllori abbiano colto un viaggiatore senza titolo di viaggio convalidato e che, di conseguenza, lo abbiano multato. Chi ha avuto la possibilità di recarsi all'estero, si sarà reso conto che questo comportamento fraudolento, perlomeno nei Paesi Occidentali, è raro perché vi è una sensibilizzazione maggiore alla civiltà e all'educazione. All'estero maggior attenzione è dedicata ai controlli e, di fatto, chi sale a bordo sui mezzi pubblici senza titolo di viaggio viene immediatamente identificato o dal conducente o da un suo collega; c'è, inoltre, maggior deterrenza in quanto si oblitera il biglietto subito all'ingresso e, nel caso non si possedesse, si ha la possibilità acquistandolo dal conducente.
In Italia spesso e volentieri non sono venduti i biglietti a bordo; inoltre il conducente non è affiancato quasi mai da un collega che controlli le obliterazioni. Sarà argomento di tale tesi percepire le aspettative dei viaggiatori sulla frequenza dei controlli sui trasporti pubblici di Reggio Emilia in funzione del possesso o meno del titolo di viaggio.

1.3 Cosa può condizionare le aspettative?

I motivi per cui un viaggiatore non dovrebbe possedere il biglietto sono di varia natura, la prima principalmente economica. Infatti, il costo del biglietto o dell'abbonamento, se si tratta di viaggiatori abituali, potrebbe essere elevato; a causa della crisi che ha colpito l'Italia verso la fine del 2009, le persone si sono impoverite e cercano di risparmiare quanto più possibile.

Dal punto di vista sociale, una persona non convalida il biglietto perché in presenza di amici o conoscenti vuole affermare se stessa dimostrandosi superiore al rischio; oppure per provare l'ebbrezza del rischio; o, addirittura, si può anche ipotizzare che un viaggiatore abituale conosca la scarsa frequenza dei controlli nel suo percorso e che perciò decida di sfidare la sorte.

Ci si potrebbe porre la questione se ci sono dei fattori esogeni, quindi degli elementi esterni cui non si può dare una spiegazione, che influenzano tali attese. Da un punto di vista logico ci si immagina che un viaggiatore senza biglietto convalidato abbia maggior timore di subire un controllo e perciò di essere multato; psicologicamente, egli percepisce di viaggiare in modo illecito e, di conseguenza, sente gravare su di sé il rischio di subire un controllo.

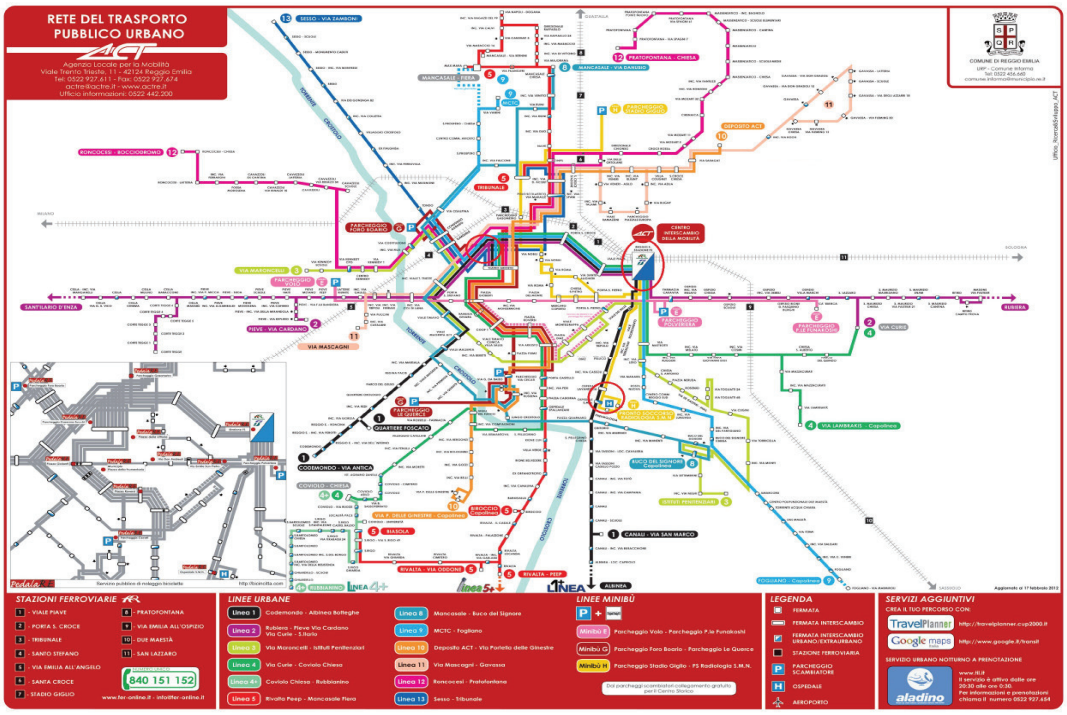
Indagheremo attraverso diverse variabili, demografiche e personali del viaggiatore oggetto delle rilevazioni, come può cambiare l'aspettativa sulla frequenza dei controlli da parte degli intervistati.

CAPITOLO 2

INDAGINE SUI PASSEGGERI DEGLI AUTOBUS DI REGGIO EMILIA

2.1 Come sono stati raccolti i dati

Ai fini dell'indagine di cui la tesi tratterà, sono stati campionati i viaggiatori delle linee degli autobus che fanno servizio a Reggio Emilia; in maniera arbitraria, si è scelto di raccogliere i dati delle persone che avessero appena viaggiato su una delle linee 1-13. Sono state anche definite tre fermate su cui appostarsi per fare le interviste ai passeggeri disposti a rispondere alle domande; ovvero "viale Allegri" (situato in centro, in una zona a traffico limitato, e punto in cui fermano molte linee), l'"Ospedale" (luogo dove per necessità le persone hanno bisogno di andare) e la "Stazione FS" (snodo fondamentale per raggiungere tutte le zone della città con le molte linee che vi fermano).



L'intervista è uno degli strumenti più importanti del metodo qualitativo, con tale termine si intende un insieme di tecniche utilizzate in ambito disciplinare; in questo caso si è utilizzato un tipo di intervista "strutturato" che consiste in una traccia di domande predefinita dove il margine dell'intervistato è nullo, perciò non era possibile tener conto dell'emotività e del linguaggio non verbale della persona intervistata.

Per evitare una distorsione nelle risposte, gli intervistatori dichiaravano fin da subito di non essere controllori e di voler sottoporre un questionario al solo fine di un'indagine sui viaggiatori.

Le persone disposte a rispondere alle domande entravano a far parte del campione, mentre chi non accettava era scartato; era richiesta l'esibizione del biglietto convalidato. Se l'intervistato lo mostrava gliene veniva procurato uno nuovo; se, invece, non risultava averlo, era catalogato come "non avente il biglietto". Viene proposta, ora, una tabella riassuntiva in cui le domande dei questionari avranno a fianco i nomi delle variabili cui saranno associate e il tipo di parametro utilizzato per esse.

DOMANDA	**VARIABILE CORRISPONDENTE**	**TIPO DI VARIABILE UTILIZZATA**
Giorno dell'intervista	day	Fattoriale con 7 livelli
Ora di arrivo	time	Numerica
Clima	weather	Fattoriale con 3 livelli
Se soleggiato, quanto caldo	temp	Fattoriale con 5 livelli
Linea su cui si è viaggiato	bus_line	Numerica
Fermata dell'autobus	stop	Fattoriale con 3 livelli
L'autobus era in orario?	in_time	Dummy (1=si)
Biglietto convalidato?	ticket	Dummy (1=si)
Sesso dell'intervistato	male	Dummy (1=maschio)
Età dell'intervistato	age	Fattoriale con 5 livelli
Nazionalità dell'intervistato	nationality	Dummy (1=italiana)
Etnia dell'intervistato	ethnicity	Fattoriale con 6 livelli
L'intervistato viaggia da solo?	alone	Dummy (1=si)

Viaggia con effetti personali?	belonging	Dummy (1=si)
Come è vestito l'intervistato?	dress_code	Fattoriale con 3 livelli
Tipo di biglietto esibito	ticket_type	Fattoriale con 3 livelli
Zone coperte dal biglietto	zone	Fattoriale con 4 livelli

Tabella 1: variabili del primo questionario.

Nel caso in cui l'intervistato rispondeva che l'autobus era in ritardo, veniva domandato a quanti minuti ammontava il ritardo, tale osservazione era collocata nella variabile numerica *delay*.

Inoltre, se la persona dichiarava di viaggiare insieme con qualcuno, veniva chiesto con chi e quanti erano, le variabili numeriche generate da queste risposte sono state: *friends* (amici), *relatives* (parenti), *colleagues* (colleghi), *other* (altro) e in quest'ultimo caso veniva chiesto di specificare chi.

Inoltre, per favorire un'indagine più approfondita, a prescindere se avessero il biglietto oppure no, se gli intervistati accettavano di rilasciare un ulteriore intervista, veniva consegnato loro un altro biglietto. La seconda intervista

riguardava aspetti più specifici; per le variabili generate da questo secondo questionario si rimanda alla tabella qui di seguito.

DOMANDA	**VARIABILE CORRISPONDENTE**	**TIPO DI VARIABILE UTILIZZATA**
Viaggia per lavoro o piacere	travel	Dummy (1=lavoro)
Quanti giorni prendi l'autobus in una settimana tipo?	n_day	Numerica
Minuti trascorsi sull'autobus in una giornata tipo	minutes	Numerica
Sei soddisfatto del servizio?	satisfy	Dummy (1=si)
Hai mezzi di trasporto alternativi?	alternative	Dummy (1=si)
Hai mai ricevuto una multa?	fine	Dummy (1=si)
Conosci qualcuno che ha ricevuto una multa?	fine_oth	Dummy (1=si)
Stima il massimo ammontare della multa per chi non ha il biglietto	fine_cost	Numerica (in euro)
Stima la frequenza dei controlli sugli autobus	control	Numerica (in percentuale)
Stima la percentuale di persone senza biglietto	evasion	Numerica (in percentuale)
Titolo di studio	education	Fattoriale con 5 livelli
Occupazione	occupation	Fattoriale con 9 livelli
Come descriveresti la tua attitudine al rischio?	risk	Fattoriale con 5 livelli
Reddito mensile percepito	income	Fattoriale con 4 livelli

Tabella 2: variabili del secondo questionario.

Se l'intervistato dichiarava di non essere soddisfatto del servizio effettuato dal mezzo di trasporto, veniva chiesto il motivo e dalle risposte sono state create le variabili: *late* (ritardo), *dirty* (sporcizia), *expensive* (servizio caro), *crowded* (autobus affollati) e *reason_oth* (altro).

Lo stesso si è fatto per le risposte sui mezzi di trasporto alternativi, dalle risposte fornite sono state generate le variabili: *car* (automobile), *scooter* (motorino), *bike* (moto), *bicycle* (bicicletta) e *altern_other* (altro).

Sul totale di 548 osservazioni derivanti dalla prima intervista, approssimativamente 160 persone hanno risposto al secondo questionario, dei quali 103 individui hanno esibito il titolo di viaggio obliterato. Per un'analisi più accurata sui dati si rimanda al paragrafo successivo.

2.2 La città di Reggio Emilia

Reggio Emilia è una delle nove province della regione italiana Emilia Romagna. Si è sentito molto parlare di questi territori verso la fine del maggio del 2012 perché la popolazione è stata colpita da due forti eventi sismici che hanno messo a dura prova tutta la regione; d'altro canto, nonostante la situazione fosse grave, tutta la popolazione ha dimostrato una forza e una coalizione ammirevole e, seppur messa in ginocchio da ingenti danni, l'intera regione non ha esitato di rimettersi subito al lavoro e portare aiuti e beni di prima necessità a chi una casa non l'aveva più.
A parte l'evento eccezionale che ha colpito quasi tutta la regione, non è un mistero quanto siano operosi gli emiliani. Si pensi a quante industrie risiedono nel territorio; la maggior parte appartengono al settore agro-alimentare (lattiero-casearia, enologica, degli insaccati, degli zuccherifici, etc.), ma non meno importante è il settore tessile e, sopratutto, meccanico. Quest'ultimo si è guadagnato un prestigio a livello mondiale grazie alla casa automobilistica "Ferrari" che ha sede a Maranello (in provincia di Modena).
Inoltre, è particolarmente sviluppato il settore terziario dell'Emilia Romagna, a

cominciare dagli scambi commerciali. Infatti, si trova in una posizione dell'Italia altamente strategica perché ha diversi porti marittimi, una rete ferroviaria importante e uno scalo aereo internazionale a Bologna.

La regione ha, in assoluto, la più omogenea distribuzione della popolazione d'Italia. Non solo non esiste una vera metropoli che sia più abitata delle altre città, ma al contrario tutti i capoluoghi di provincia superano la soglia di 100.000 abitanti senza arrivare ai 200.000, ovviamente fatta eccezione per Bologna che conta più di 380.000 abitanti. Pur essendosi verificata, qui come in tutta Italia, una forte emigrazione dalle campagne alle città, lo sviluppo urbano è stato tenuto sotto controllo dall'amministrazione regionale. Il ricco passato storico di quasi tutte le città emiliane e il forte senso dell'orgoglio locale hanno evitato l'eccessiva preponderanza di un centro rispetto agli altri.

A supporto di quanto detto fin ora, il quotidiano "Il Sole 24 Ore" ha pubblicato un'indagine condotta dal Centro Studi Sintesi nella quale si mette a confronto il reddito disponibile con il tenore di vita delle famiglie. Laddove le due variabili sono estreme il cosiddetto "rischio-evasione"aumenta in quanto un reddito minimo non riesce a far fronte ad un tenore di vita elevato. Ponendo la media nazionale pari a 100, a livello locale se il punteggio ottenuto è più alto significa che i consumi sono "giustificati" dai redditi, mentre se il punteggio è inferiore alla soglia indica che, mediamente, si spende di più di quanto si dichiara al fisco. E' emerso che l'Emilia-Romagna si è classificata al primo posto d'Italia con 147 punti, un risultato che indica che c'è minor rischio di "nero" potenziale, in particolare, a Reggio Emilia sono stati assegnati 116 punti, 15 punti in più rispetto a sei anni fa.

Entrando nello specifico di Reggio Emilia, si contano circa 172.000 abitanti, che ne fanno una città di media grandezza; essa è in vetta alle classifiche delle città più prospere e vivibili dell'Italia e conosciuta per l'operosità e l'ingegno dei suoi abitanti.

Un'inchiesta, pubblicata sul sito Internet dall'ufficio stampa del comune della città in data 2 gennaio 2012 e realizzata da "Italia Oggi" (un quotidiano economico, giuridico e politico) in collaborazione con l'Università "La Sapienza",

colloca la città di Reggio Emilia al quinto posto, dopo Trento e Bolzano (entrambe autonome), Pordenone e Mantova, tra le città italiane per la qualità della vita. Tra le voci che concorrono a segnare gli ottimi livelli di qualità della vita, nonostante sia stata anch'essa colpita dalla crisi economica della Storia recente, vi sono il verde e l'attenzione all'ambiente, l'ecosostenibilità, i servizi, la vivacità economica (calcolata sulla base del quadro attuale del sistema imprenditoriale delle province italiane, ad esempio la densità imprenditoriale, il tasso di crescita delle imprese, il grado di innovazione, il mercato del lavoro, il tasso di imprenditoria femminile, etc, e sulle previsioni del PIL di medio periodo). A riprova dei buoni standard di vita l'inchiesta ha evidenziato che il tenore di vita dei reggiani, calcolato in conformità a: redditi, capacità di spesa, depositi bancari e costo degli alloggi, è quinto su base nazionale.

2.3 Statistiche descrittive

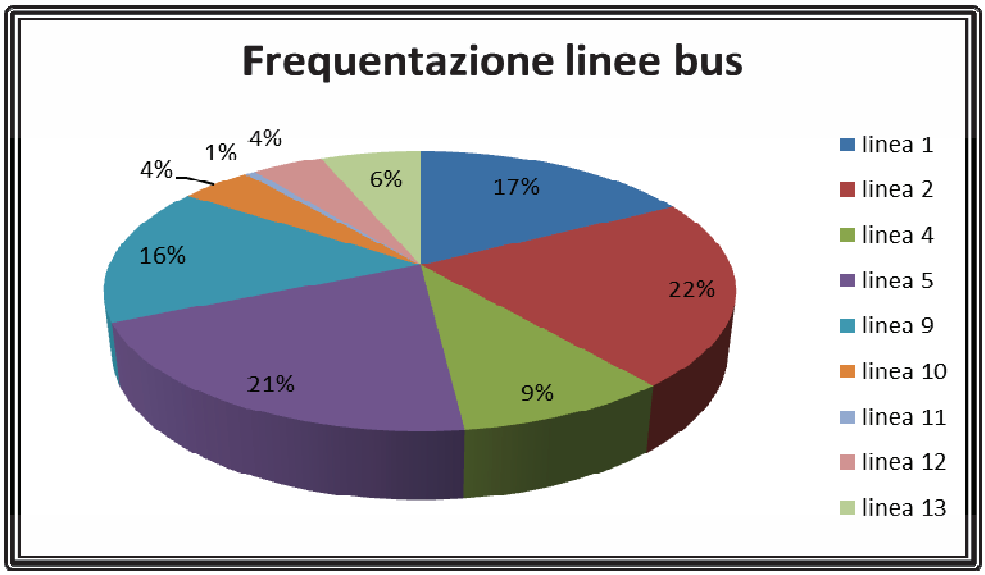
Dato che lo scopo dell'analisi qui proposta è studiare la relazione che intercorre tra la percezione della frequenza dei controlli sugli autobus e le risposte degli intervistati ai questionari si sono elaborati i dati delle persone che hanno risposto ad entrambe le interviste. A tal proposito è stato utilizzato il software statistico Stata, un moderno e ricco programma per le analisi statistiche, la creazione di grafici e la manipolazione di dati.
La numerosità campionaria risulta essere di circa 160 osservazioni.

2.3.1 Medie e frequenze del campione

Dalle interviste è emerso che a proposito dei dati sulle caratteristiche del giorno

in cui le persone hanno viaggiato, l' 82% risulta che viaggiasse tra lunedì e venerdì, mentre il restante 18% nel week-end; a tal fine si è utilizzata una variabile dummy "weekend", la quale vale 1 se il giorno dichiarato è sabato o domenica e 0 altrimenti. Le condizioni atmosferiche, catalogate grazie ad una variabile fattoriale "weather" che valesse 1 nel caso in cui ci fosse il sole, 2 se nuvoloso e 3 se piovesse, sembrano influenzare l'affluenza negli autobus, infatti, dal primo questionario emerge che il 37% dei viaggiatori ha dichiarato che ci fosse il sole, il 45% che fosse nuvoloso e il 18% che piovesse, perciò più della metà degli intervistati ha viaggiato con un tempo atmosferico incerto se non addirittura piovoso.

Analizzando i dati concernenti le linee dell'autobus, si è registrata una frequenza maggiore, pari a 21.6%, sulla linea 2 che percorre il tragitto Rubiera-S. Ilario, in pratica attraversa la città sull'asse est-ovest; segue la linea 5 con una frequenza pari a 20.4%, essa attraversa la città sull'asse nord-sud. Le linee che non sono state percorse da alcun intervistato sono la 3 (collega la parte ovest della città con gli Istituti Penitenziari restando fuori dal centro città) e la 8 (i cui capolinea sono Mancasale-via Danubio, a nord, e Buco del Signore, a sud, passando esternamente al centro); la linea che ha assunto la frequenza minima diversa da zero è la 11 con 0.6%, essa collega la zona ovest di Reggio Emilia con la località di Gavassa, a nord-est.

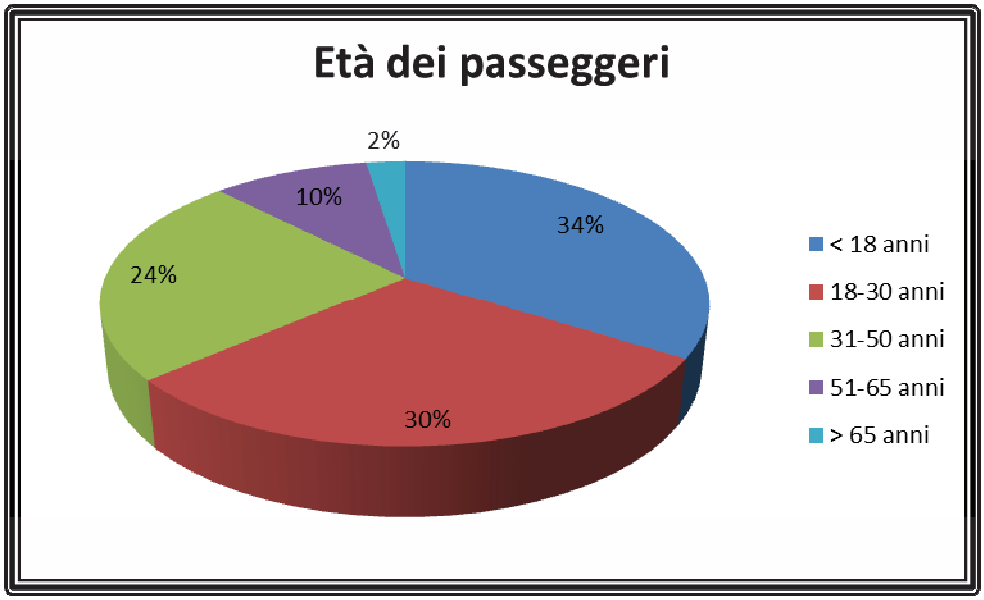


A conferma di quanto è stato appena illustrato, i dati evidenziano che l' 85.8% dei viaggiatori si trovava nella fermata strategica di viale Allegri, quindi zona centro di Reggio Emilia; la stazione ferroviaria FS segue con il 13.6% e l'ospedale ha registrato lo 0.6% di frequenza.

Inoltre, si è evidenziato che il 57% degli autobus risultava in orario. Per gli autobus che erano in ritardo l'intervistato ha dichiarato che in media l'orario di arrivo era stato ritardato di quasi due minuti; il valore massimo attribuito al ritardo di un autobus è di 15 minuti ed è un dato fornito solo da una persona, mentre il valore minimo è di 0 minuti, ovvero l'autobus è arrivato in orario, ed è stato notificato da 92 persone che equivalgono al 57.14% degli intervistati.

Il 65% dei viaggiatori intervistati ha esibito il biglietto convalidato, quindi si evince che il 35% viaggia in maniera illecita sui mezzi di trasporto pubblici di Reggio Emilia ed è un risultato significativo; inoltre, sul totale degli intervistati il 44% è di sesso femminile, si evince che il campione è abbastanza omogeneo al suo interno dal punto di vista del sesso.

Il 34% del campione totale risulta avere meno di 18 anni; si può supporre che molto probabilmente essi utilizzino l'autobus come mezzo di trasporto da e per la scuola. Un'altra percentuale cospicua, pari a quasi il 30%, risulta avere tra i 18 e i 30 anni, il 24% dichiara di avere tra i 31 e i 50 anni, il 10% tra i 51 e i 65 anni e il 2% risulta in età pensionabile, quindi oltre i 65 anni. Per poter individuare queste 5 fasce d'età si è costruita una variabile fattoriale a 5 livelli ognuno dei quali indicava un determinato range degli anni. Di seguito è possibile vedere un diagramma circolare, o più comunemente un grafico a torta, utilizzato in statistica descrittiva per rappresentare variabili misurate su classi di categorie affinché non sia stabilito, anche involontariamente, un ordine inesistente nelle categorie.

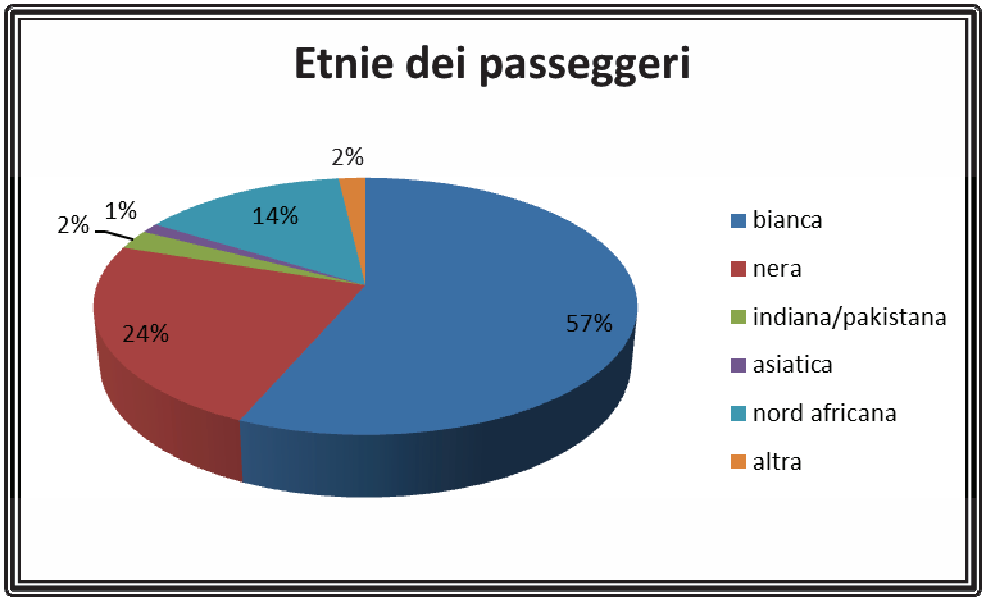


E' emerso che il 47.5% è di nazionalità italiana. Confrontando gli individui del campione a seconda della nazionalità e se avessero o no un titolo di viaggio regolare, come si può notare anche dalla tabella sottostante, è emerso che gli italiani che non hanno il biglietto sono l'11.18% e quelli che lo hanno sono il 36.65%, mentre gli stranieri che non hanno il biglietto sono pari al 23.6% e quelli che lo posseggono sono il 28.57%. Perciò, coloro che viaggiano irregolarmente sono in particolar modo gli stranieri e in maniera regolare sono di più gli italiani.

	nationality		
ticket	0	1	Total
0	38	18	56
	23.60	11.18	34.78
1	46	59	105
	28.57	36.65	65.22
Total	84	77	161
	52.17	47.83	100.00

Output 1: tabella di frequenza per variabili *ticket* (si possiede o no il biglietto) e *nationality* (nazionalità italiana o straniera).

L'intervistatore era incaricato, poi, di riportare l'etnia cui apparteneva l'intervistato. Analizzando la variabile "*ethnicity*", anch'essa fattoriale, che classifica in 6 livelli diverse etnie, si evince che la maggior parte (56.5%) degli intervistati è di etnia bianca, il 23.6% era nera, il 14.3% nord-africana e a seguire c'è l'etnia indiana/pakistana (2.5%), asiatica (1.2%) e non specificata (1.9%).



Il 54% degli intervistati ha dichiarato di viaggiare da solo; qualora l'intervistato rispondesse di viaggiare insieme con qualcuno, si ricorda che nel questionario veniva richiesto di specificare con chi. Di seguito vengono riportate, quindi, le risposte del campione. Tra coloro che viaggiavano insieme con qualcuno, il 35% ha risposto di viaggiare con amici, di cui il 21.4% ne dichiarava solo uno e il numero massimo di amici dichiarati è stato 4 (3.7%). Il 9.8% viaggiava con i propri parenti, nella maggior parte dei casi i viaggiatori sono accompagnati da un solo parente, emerge che al massimo i viaggiatori fossero in compagnia di tre parenti. Sul totale del campione non ci sono state persone che abbiano risposto di essere accompagnati da un collega o qualcun altro. Il più delle volte coloro che hanno risposto di essere in compagnia di qualcuno si trovavano nella

fermata di viale Allegri.

Veniva domandato anche se i viaggiatori avessero effetti personali con sé, in quanto nel caso fossero studenti, essi avrebbero dovuto avere uno zaino o qualcosa di simile, oppure, qualora fossero dei turisti, una valigia, nel caso di lavoratori una borsa ventiquattrore, etc. E' emerso che solo il 7% portasse con sé effetti personali, dei quali il 6.3% si trovava in viale Allegri, questo è un dato pressoché superficiale dato che la maggioranza delle osservazioni era in quella fermata.

Successivamente si è analizzato come gli individui del campione definissero il loro abbigliamento. Si è riscontrato che l' 86% si definisce casual, il 9.5% elegante e adatto al lavoro, il 3.2% ritiene che il suo look sia povero e l'1.3% etnico. E' importante considerare che poco più della metà di chi si ritiene casual non è di nazionalità italiana; all'interno dei gruppi "elegante" e "povero" vi è una discreta omogeneità a livello di nazionalità e tutti coloro che hanno un abbigliamento etnico non sono italiani.

Sul totale di chi ha risposto, il 56.3% aveva l'abbonamento; i rimanenti avevano un biglietto di singola corsa (24.3%) o multipla (19.4%).

Più interessanti sono le risposte al secondo questionario, in quanto la persona intervistata esprime alcune sue caratteristiche e attitudini, come ad esempio se è un viaggiatore abituale, se ha mai ricevuto una multa, che grado di istruzione ha conseguito, che lavoro fa, etc. Vediamo nello specifico come ha risposto il campione.

Il 66.5% degli intervistati sposta per lavoro; se si confronta l'età con il parametro "*travel*", la quale è una variabile dummy che identifica se l'intervistato viaggia per lavoro o per piacere, si nota che il 24% di coloro che

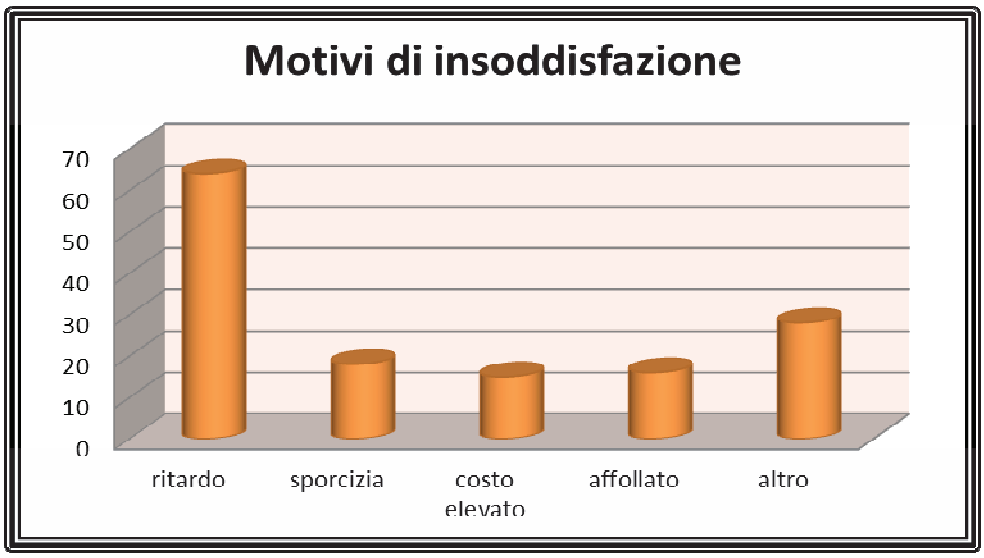
viaggiano per lavoro hanno meno di 18 anni, il 18.5% ha tra i 18 e i 30 anni, il 17.9% tra i 31 e i 50 anni, il 6.1% tra i 51 e i 65 anni, come naturale che sia nessuno con età maggiore di 65 anni viaggia per lavoro. Rilevante è che il 10% degli intervistati risulti essere minorenne e viaggiare per piacere.

In una settimana tipo è emerso che i viaggiatori utilizzano l'autobus per il proprio spostamento in media quasi 6 giorni su sette, la percentuale che dichiara di viaggiare 6 giorni su 7 è del 32.3% sul totale del campione e quasi il 35% afferma di viaggiare 7 giorni su 7 in una settimana tipo. Confrontando l'età con la quantità di giorni in cui in media si usufruisce dell'autobus si evince che il 58.2% di coloro che rientrano nella fascia d'età minorenne viaggia 6 giorni su 7, mentre chi ha meno di 18 anni e viaggia 7 giorni su 7 è pari al 30.9%.

Si è poi pensato di definire una nuova variabile che interpreti come viaggiatore occasionale una persona che in media viaggia massimo 4 giorni alla settimana. I dati mostrano che solo il 18.3% del campione è un viaggiatore occasionale.
In una giornata tipo, gli intervistati trascorrono sull'autobus in media quasi 39 minuti. Osservando le frequenze relative più alte registrate, si è notato che 16 individui passano 10 minuti, 19 dicono 20 minuti, 35 restano mezz'ora sull'autobus, 22 trascorrono 40 minuti, 26 un'ora e, infine, 7 individui dichiarano di rimanere due ore.

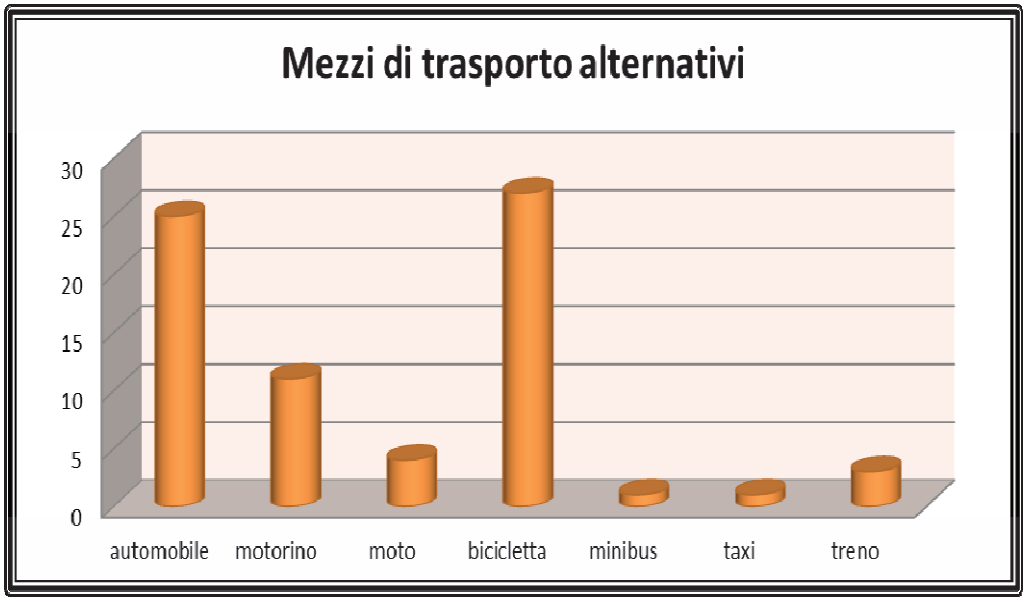
Poco più della metà (54.2%) non è soddisfatto del servizio; a questo proposito nel questionario si chiedeva di specificare le cause che determinavano l'insoddisfazione del viaggiatore, si poteva fornire più di una risposta. I motivi che sono stati menzionati di più sono il ritardo (al primo posto con 64 risposte su 164 totali), la sporcizia, l'affollamento dei mezzi di trasporto, e il costo eccessivo del servizio. Vi era la possibilità di esprimere anche altre cause, sono stati allora menzionate le corse cancellate, l'inattendibilità e, di conseguenza, i cambi da una linea all'altra sbagliati, tragitti troppo lunghi, scarso numero di corse, autisti maleducati, poco comfort dei mezzi, ambienti troppo caldi, tabelle

orarie sbagliate e da modificare, infine un individuo desidererebbe anche che gli autobus mantenessero il servizio feriale anche la domenica.



A questo punto è stato domandato se il campione avesse delle possibilità diverse rispetto il mezzo di trasporto pubblico: solo il 42% ha risposto affermativamente. Ciò può dipendere dal fatto che per raggiungere il luogo di lavoro bisogna passare per la zona a traffico limitato e non si possiede il permesso, oppure si tratta di ragazzini che raggiungono la scuola in autobus e i genitori non comprano loro il motorino o non li lasciano andare in bicicletta perché le strade sono pericolose.

Nel caso la risposta fosse stata affermativa, si è indagato su quale mezzo di trasporto potessero far affidamento oltre all'autobus, anche in questo caso potevano dare più di una risposta. L'alternativa maggiormente presa in considerazione è stata la bicicletta, poi la macchina, lo scooter, la moto e c'è anche chi ha menzionato il treno o il taxi oppure ancora il minibus.



In seguito, il questionario ha approfondito l'argomento delle sanzioni: il 46% ha dichiarato di aver ricevuto almeno una multa e, in generale, il 74% conosce qualcuno che è stato sanzionato. In totale il 38% ha sia ricevuto personalmente la sanzione sia ha conoscenti che sono stati multati. Confrontando la variabile legata al sesso con il fatto se si ha mai ricevuto una multa, si nota che il 27.44% di chi è stato sanzionato appartiene alla categoria maschile, mentre la quota sanzionata della categoria femminile è il 18.9%, non vi è, invece, molta disparità tra i due sessi nel caso non siano stati multati.

	fine		
male	0	1	Total
0	41	31	72
	25.00	18.90	43.90
1	47	45	92
	28.66	27.44	56.10
Total	88	76	164
	53.66	46.34	100.00

Output 2: tabella di frequenza tra le variabili sesso e *fine* (si è ricevuta almeno una multa).

Inoltre, indagando ancora con tali variabili e confrontandole con i casi "l'intervistato ha mostrato il biglietto convalidato" / "l'intervistato non ha mostrato il biglietto convalidato", si evince che il sesso femminile è più onesto, in quanto le donne intervistate nella maggior parte dei casi hanno esibito il regolare titolo di viaggio e, rispetto agli uomini, erano numericamente inferiori coloro che hanno ricevuto almeno una volta la multa. Vediamo di seguito i dettagli delle due analisi:

Biglietto non esibito

	fine	
male	0	1	Total
0	3	11	14
	5.36	19.64	25.00
1	14	28	42
	25.00	50.00	75.00
Total	17	39	56
	30.36	69.64	100.00

Output 3: tabella di frequenza tra le variabili sesso e *fine* (si è ricevuta almeno una multa) nel caso in cui non si possedeva il biglietto.

Biglietto esibito

	fine	
male	0	1	Total
0	36	19	55
	34.29	18.10	52.38
1	33	17	50
	31.43	16.19	47.62
Total	69	36	105
	65.71	34.29	100.00

Output 4: tabella di frequenza tra le variabili sesso e *fine* (si è ricevuta almeno una multa) nel caso in cui si possedeva il biglietto.

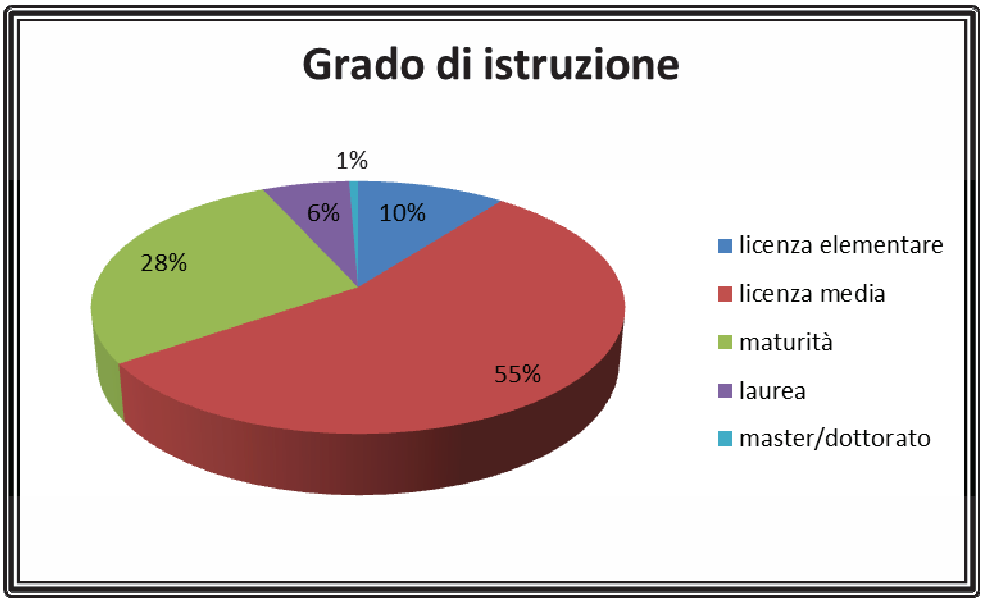
Si è, quindi, richiesto di stimare l'ammontare in euro della multa per mancata

esibizione del titolo di viaggio convalidato. Ne è risultato che in media si reputa che la multa sia pari a € 44.60. Curiosa è la disparità tra le risposte degli intervistati, tra i quali c'è chi ritiene che la sanzione possa valere € 5.00 (valore minimo assunto da tale variabile) e chi € 120.00 (valore massimo). Ovviamente è importante andare a vedere i valori dichiarati da chi ha effettivamente ricevuto la multa e chi no. Ne emerge una grande variabilità tra le risposte. Infatti, il *range* delle risposte va da un minimo di € 10 euro a un massimo di € 100; in media risulta che chi ha già ricevuto la multa stima che essa sia pari a € 42.54. Sul sito dei trasporti di Reggio Emilia (www.actre.it) sono disponibili gli importi delle sanzioni amministrative, attualmente la multa base è pari a € 40 più il costo del biglietto, ma in caso di ritardo nel pagamento di sei giorni la sanzione aumenta a € 50 più il costo del biglietto oppure oltre due mesi dalla notifica si arriva a € 150 più le spese postali e il titolo di viaggio.

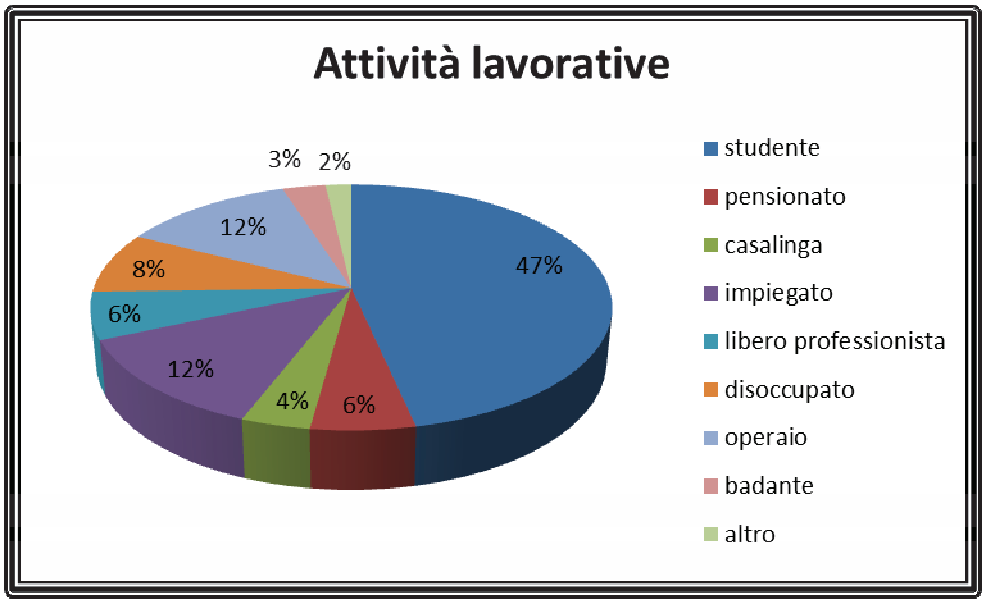
Alla domanda di stimare la frequenza dei controlli sui biglietti dalle persone preposte del servizio pubblico, è emerso che in media la frequenza è del 30.13%; distinguendo tra chi ha il biglietto convalidato e chi no, è emerso che un maggior numero di persone che viaggiano irregolarmente sui mezzi di trasporto prevedono che vi sia il rischio di controlli sui biglietti di oltre il 50%, più omogenea è la distribuzione della percezione dei controlli da parte di chi viaggia regolarmente sull'autobus.

Un risultato altrettanto indicativo è ciò che si riscontra quando viene richiesto agli intervistati di stimare la percentuale di persone senza biglietto sugli autobus, complessivamente si reputa che l'evasione sia pari al 59.24% dei viaggiatori. Andando a vedere le stime nel dettaglio tra chi viaggia con un titolo di viaggio valido e chi invece no, si viene a scoprire che, se gli intervistati hanno esibito un biglietto regolare, essi in media ritengono che il 56.35% dei viaggiatori sia irregolare; mentre aumenta la percezione di chi non esibisce il titolo di viaggio, infatti, in media si stima che l'evasione sia pari al 63.31%.

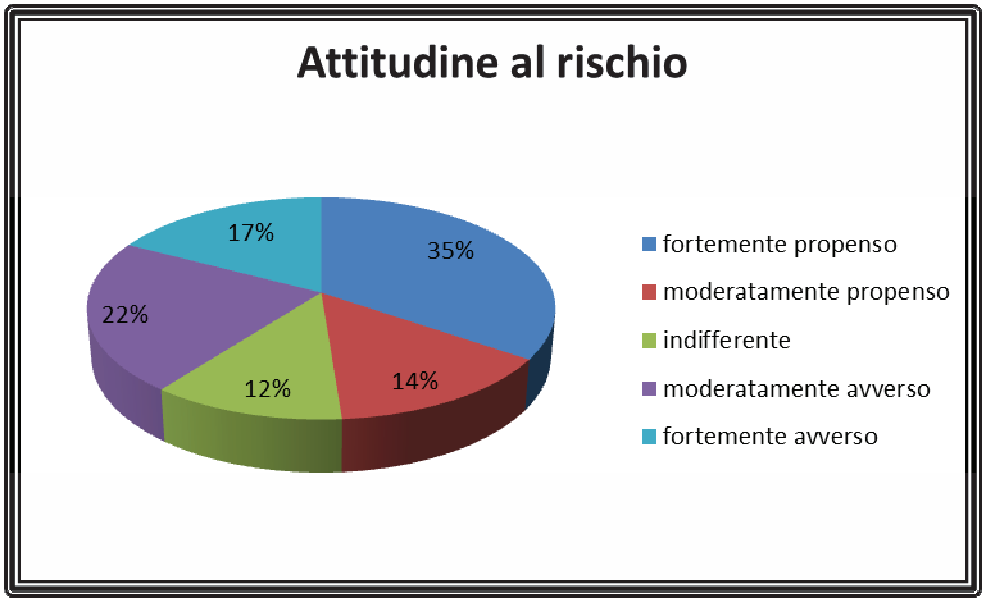
In ultimo, ma non meno importante ai fini dell'indagine, sono state richieste agli intervistati alcune caratteristiche personali, tra le quali il grado di istruzione: il 10.49% degli individui campionati ha la licenza di scuola elementare, il 54.94% la licenza di scuola media, il 27.78% possiede il diploma di maturità, solo il 6.17% si è laureato e, infine, solo una dichiara di aver come titolo di studio il master/dottorato.



E' stato domandato loro anche l'occupazione svolta al momento. Quasi la metà degli individui (46.58%) ha dichiarato di essere studente; le categorie degli operai e degli impiegati sono rappresentate entrambe dal 12.42% del campione, l' 8.07% è disoccupato, il 6.21% è libero professionista, il 5.59% è pensionato, il 3.73% è una casalinga (tutti gli intervistati di questa categoria sono di sesso femminile), il 3.11% è un/una badante, infine l'1.86% non ha specificato la propria professione.



Si chiedeva in che modo l'intervistato avrebbe descritto la propria attitudine verso il rischio, venivano proposte cinque possibili risposte (da "fortemente avverso" a "fortemente propenso") e la persona doveva scegliere quella che riteneva più adeguata alla sua abitudine. Il 17.28% del campione si è dichiarato fortemente avverso al rischio, il 22.22% ha poca attitudine, ma l'attenzione è richiamata dalle significative percentuali a favore del rischio, infatti, il 14.2% si definisce moderatamente propenso e addirittura il 34.57% è fortemente propenso; mentre si ritiene indifferente al rischio l' 11.73% del campione totale.

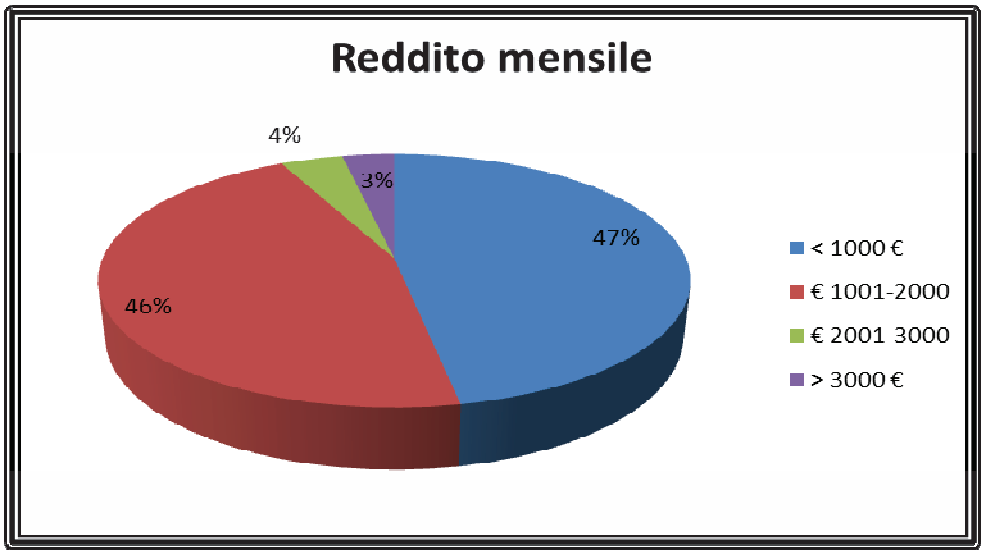


Infine, il questionario richiedeva di dire a quale fascia di reddito netto mensile percepito si apparteneva: ne è risultato che il 46.98% del campione intervistato dichiara di percepire un reddito mensile pari a meno di € 1000, il 45.64% tra € 1001 e € 2000, il 4.03% tra € 2001 e € 3000 e, infine, il 3.36% oltre € 3000. A prima vista si potrebbe pensare che chi percepisce un reddito inferiore a € 1000 sia uno studente. Invece si riscontra che solo l'11.18% è uno studente con tale reddito e ben il 34.46% non è uno studente ma percepisce € 1000. Per una maggior chiarezza, viene proposta la tabella riassuntiva, dove il reddito, *income*, è una variabile fattoriale a quattro livelli dove 1 indica la fascia meno di € 1000 e così via, mentre *student* è una dummy che indica con 1 se l'intervistato è uno studente:

income	student 0	1	Total
1	51 34.46	18 12.16	69 46.62
2	28 18.92	40 27.03	68 45.95
3	1 0.68	5 3.38	6 4.05
4	1 0.68	4 2.70	5 3.38
Total	81 54.73	67 45.27	148 100.00

Output 5: tabella di frequenza tra le variabili reddito e studente.

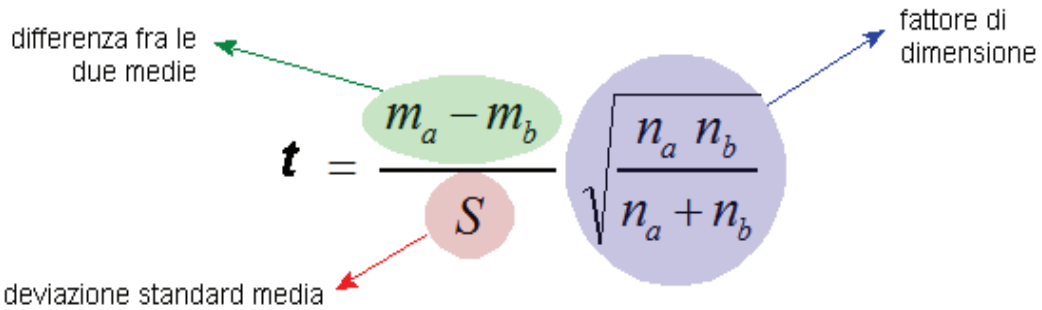
Si evidenzia quindi che la maggior parte della popolazione intervistata percepisce mensilmente meno di € 2000, tale dato può essere interessante in quanto mettendolo in relazione alle motivazioni date per aver definito il servizio insoddisfacente, emerge che coloro che hanno definito i mezzi pubblici costosi rientrano tutti nelle prime due fasce di reddito.



2.3.2 Test T di Student

A completamento delle statistiche descrittive appena proposte, è opportuno proporre dei test statistici volti a verificare l'ipotesi nulla secondo cui due campioni hanno le rispettive medie uguali rispetto ad una variabile contro l'ipotesi alternativa, secondo cui le medie dei due campioni sono significativamente diverse. Per questa analisi il test statistico idoneo è il test "T di Student", il quale in generale può essere di due tipologie: ad un campione o a due campioni.

Ai fini dell'indagine si considera il test T a due campioni, il quale prevede il calcolo della statistica test t come segue:



dove la deviazione standard, *S*, è la radice quadrata della varianza che si ottiene sommando le devianze dei due campioni e dividendo per la somma dei gradi di libertà, i quali si calcolano sottraendo 2 alla somma delle due numerosità campionarie definite con *n* (con i termini *a* e *b* si indicano i due campioni).
Una volta calcolato il valore t, esso va confrontato con i valori tabulati della distribuzione T di Student; da tale confronto si potrà stabilire se la differenza tra le due medie è nulla e dovuta al caso oppure se è statisticamente significativa. La verifica d'ipotesi si basa sulla determinazione del valore p che misura la probabilità di estrarre campioni caratterizzati da un valore della statistica t più elevati di quello osservato per i campioni in esame. Valori molto bassi del p-value e, quindi, inferiori al livello di significatività arbitrariamente prescelto indicano che sotto l'ipotesi nulla il risultato campionario osservato è molto anomalo e fanno propendere per la decisione di rifiutare l'ipotesi stessa.
Solitamente si utilizza un livello di significatività pari al 5%; tale termine si definisce quel valore soglia tale per cui l'ipotesi nulla viene rifiutata quando questa è vera con quel livello di probabilità.

Si considerino ora i dati ricavati dalle interviste eseguite a Reggio Emilia; si ricordi, inoltre, che lo scopo di questa indagine è valutare la percezione della frequenza dei controlli in relazione alle altre variabili ricavate dalle risposte fornite dagli intervistati. Dalle analisi precedenti si è riscontrato un comportamento maggiormente etico da parte delle donne, le quali, a differenza degli uomini, erano in possesso di biglietto regolare rispetto a quelle senza. Potrebbe essere interessante, quindi, verificare se c'è una differenza delle percezioni tra il genere femminile e il genere maschile in merito alla frequenza dei controlli. L'output fornito dal software Stata è il seguente:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      69    29.82609    3.051976    25.35162    23.73596    35.91621
       1 |      92     30.3587    2.675988    25.66718    25.04317    35.67422
---------+--------------------------------------------------------------------
combined |     161    30.13043    2.006069    25.45416    26.16865    34.09222
---------+--------------------------------------------------------------------
    diff |           -.5326087    4.066223               -8.563384    7.498167
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.1310
Ho: diff = 0                                     degrees of freedom =      159

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.4480         Pr(|T| > |t|) = 0.8960          Pr(T > t) = 0.5520
```

Output 6: confronto delle medie dei due sessi rispetto la percezione dei controlli.

Il test T di Student, con un p-value pari a 0.896 (>0.05), non rifiuta l'ipotesi nulla di uguaglianza tra le medie delle percezioni maschili e femminili; l'assenza di disparità delle percezioni appare evidente anche osservando i valori calcolati delle medie: in media la frequenza dei controlli secondo le donne è pari al 29.83%, mentre secondo gli uomini è del 30.36%. Nonostante le donne siano più oneste, non cambia la percezione all'interno dei due sessi.

Si valuta ora la percezione della frequenza dei controlli dal punto di vista di chi possiede e chi no il titolo di viaggio convalidato. Da un punto di vista logico e morale, si presume che chi viaggia in maniera illecita tema di subire un controllo e di ricevere una sanzione, quindi dovrebbe dichiarare una percentuale maggiore. D'altro canto, ci si aspetta che gli intervistati onesti, viaggiando regolarmente, siano meno influenzati dalla preoccupazione dei controlli:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      55        36.4    3.063921    22.72264    30.25721    42.54279
       1 |     103    27.41748    2.607357    26.46178     22.2458    32.58916
---------+--------------------------------------------------------------------
combined |     158     30.5443    2.029743    25.51348    26.53518    34.55343
---------+--------------------------------------------------------------------
    diff |            8.982524     4.21357                .6595123    17.30554
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.1318
Ho: diff = 0                                     degrees of freedom =      156

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9827         Pr(|T| > |t|) = 0.0346          Pr(T > t) = 0.0173
```

Output 7: confronto delle medie di chi ha il biglietto e chi no rispetto la percezione dei controlli.

Il risultato del test in questo caso ha riscontro positivo rispetto le aspettative, infatti, dato un livello di significatività pari al 5%, il p-value calcolato (0.0346) è inferiore, ne consegue che l'ipotesi nulla di uguaglianza delle medie è rifiutata. Gli intervistati che non possiedono il biglietto percepiscono una frequenza dei controlli pari al 36.4%, mentre chi viaggia regolarmente pensa che i controlli siano pari al 27.42%, questi dati sono in linea con quanto detto precedentemente, anche se si può affermare che la percentuale fornita dai viaggiatori disonesti è relativamente bassa. Un range percentuale che possa essere definito ragionevole per quanto concerne i controlli potrebbe essere tra il 20 e il 40%, si provvederà a costruire un modello di regressione Probit con variabile dipendente una dummy che indichi con 1 la percezione della frequenza dei controlli tra il 20 e il 40% e 0 altrimenti.

Dato che nel campione analizzato non c'è una numerosità prevalente a livello di nazionalità - infatti, solo il 47.5% degli individui è italiano - si è ritenuto opportuno confrontare anche la percezione della frequenza dei controlli tra i sottocampioni italiano e straniero. Si è ottenuto:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      82    37.80488    2.783324    25.20408    32.26694    43.34282
       1 |      77    22.48052    2.681078    23.52636    17.14069    27.82035
---------+--------------------------------------------------------------------
combined |     159    30.38365    2.023325    25.51316    26.38739     34.3799
---------+--------------------------------------------------------------------
    diff |            15.32436    3.873013                7.674426    22.97429
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   3.9567
Ho: diff = 0                                     degrees of freedom =      157

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0001          Pr(T > t) = 0.0001
```

Output 8: confronto delle medie delle popolazioni italiana e straniera rispetto la percezione dei controlli.

Il test rifiuta l'ipotesi nulla con un p-value pari a 0.0001, il che comporta una significativa differenza tra la percezione degli italiani e degli stranieri; a riprova di ciò si possono notare cosa dichiarano in media i due sottocampioni. Gli intervistati che non sono di nazionalità italiana ritengono che in media la frequenza dei controlli sia del 37.8%, un risultato nettamente superiore rispetto a quanto dichiarato dagli italiani, secondo i quali in media la frequenza è solo del 22.48%.

Ci si è poi domandato se l'aver ricevuto almeno una volta la multa possa influire sulla percezione dei controlli. In realtà emerge che non vi è differenza: chi ha ricevuto una sanzione dichiara in media una percezione del 31.5%, mentre chi non è mai stato multato quasi del 29%.

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      86    28.95349    2.793088    25.90204    23.40008     34.5069
       1 |      75       31.48     2.89089    25.03584    25.71978    37.24022
---------+--------------------------------------------------------------------
combined |     161    30.13043    2.006069    25.45416    26.16865    34.09222
---------+--------------------------------------------------------------------
    diff |           -2.526512    4.029182               -10.48413    5.431108
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.6271
Ho: diff = 0                                     degrees of freedom =      159

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.2658         Pr(|T| > |t|) = 0.5315          Pr(T > t) = 0.7342
```

Output 9: confronto delle medie di chi è stato multato e chi no rispetto la percezione dei controlli.

Considerando la stima della percentuale di persone senza biglietto, si propone ora il test T di Student per il confronto tra le medie stimate dai due sessi.

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      67    56.67164    3.219006     26.3487    50.24469     63.0986
       1 |      91    61.13187    2.505048    23.89663    56.15515    66.10858
---------+--------------------------------------------------------------------
combined |     158    59.24051    1.987452    24.98189    55.31491     63.1661
---------+--------------------------------------------------------------------
    diff |           -4.460226    4.018604               -12.39812    3.477672
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.1099
Ho: diff = 0                                     degrees of freedom =      156

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1344         Pr(|T| > |t|) = 0.2688          Pr(T > t) = 0.8656
```

Output 10: confronto tra sesso femminile e maschile per la stima delle persone che viaggiano senza biglietto.

Con un p-value nettamente superiore al livello di significatività prescelto (0.05), si accetta l'ipotesi nulla di uguaglianza tra i due generi, infatti, se le donne dichiarano che in media le persone senza biglietto sono una quota pari al 56.67%, gli uomini aumentano la percezione al 61.13%. Perciò non c'è una differenza significativa, ma non è irrilevante il fatto che entrambe le stime superino il 55%, ciò significa che l'aspettativa di viaggiatori illeciti è elevata.

Mantenendo di riferimento la variabile che indica la stima percentuale di viaggiatori irregolari e mettendola in relazione al fatto se si possiede o meno il titolo di viaggio convalidato, il software Stata propone tale risultato:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      54    63.31481    3.159937     23.2207    56.97678    69.65285
       1 |     101    56.34653    2.553485     25.6622    51.28049    61.41258
---------+--------------------------------------------------------------------
combined |     155    58.77419    2.006912    24.98585    54.80956    62.73882
---------+--------------------------------------------------------------------
    diff |             6.96828    4.188161               -1.30581    15.24237
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.6638
Ho: diff = 0                                     degrees of freedom =      153

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9509         Pr(|T| > |t|) = 0.0982          Pr(T > t) = 0.0491
```

Output 11: confronto delle medie di chi ha il biglietto e chi no rispetto la stima percentuale dei viaggiatori irregolari.

Nonostante a un livello di significatività del 5% si accetti l'ipotesi nulla di uguaglianza tra le medie, se si scegliesse un livello pari al 10% si avrebbe un riscontro della differenza tra chi ha esibito il titolo di viaggio e chi no. Si noti, infatti, che la percezione di chi viaggia illegalmente è del 63.3%, tale dato è superiore di qualche punto percentuale rispetto a quanto valuta in media una persona che viaggia regolarmente (56.35%).

Si è poi riscontrata l'assenza di significativa differenza tra italiani e stranieri nel determinare la percentuale di persone che viaggiano senza titolo di viaggio; infatti, se gli italiani in media stimano che siano il 58.78% dei viaggiatori, gli stranieri dicono il 59%.

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      80     59.0375    2.848853    25.48092      53.367      64.708
       1 |      76    58.77632    2.818896    24.57457    53.16079    64.39184
---------+--------------------------------------------------------------------
combined |     156    58.91026    1.998642    24.96304    54.96216    62.85835
---------+--------------------------------------------------------------------
    diff |            .2611842    4.011506               -7.663498    8.185866
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.0651
Ho: diff = 0                                     degrees of freedom =      154

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5259         Pr(|T| > |t|) = 0.9482          Pr(T > t) = 0.4741
```

Output 12: confronto delle medie degli italiani e degli stranieri rispetto la stima percentuale dei viaggiatori irregolari.

Inaspettatamente si è evidenziata una differenza significativa tra le medie delle stime di viaggiatori senza biglietto fornite dagli intervistati soli e da quelli che erano in compagnia di qualcuno. Di seguito è riportato il calcolo del test:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      70    65.17143     2.88194    24.11204    59.42212    70.92074
       1 |      84    53.08333    2.655017    24.33363    47.80261    58.36405
---------+--------------------------------------------------------------------
combined |     154    58.57792    2.006288    24.89738    54.61432    62.54153
---------+--------------------------------------------------------------------
    diff |             12.0881    3.921788                4.339843    19.83635
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   3.0823
Ho: diff = 0                                     degrees of freedom =      152

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9988         Pr(|T| > |t|) = 0.0024          Pr(T > t) = 0.0012
```

Output 13: confronto delle medie di chi viaggiava solo e chi in compagnia rispetto la stima percentuale dei viaggiatori irregolari.

Il test T di Student riporta un p-value pari a 0.0024 che porta a rifiutare l'ipotesi nulla di uguaglianza delle medie con un livello di significatività del 5%, si potrebbe rifiutare anche a un livello soglia dell' 1%. Questo risultato indica una forte disparità tra le stime, infatti, chi ha risposto di aver viaggiato da solo propone una stima del 53.1%, invece chi era in compagnia del 65.2%.

Nel prossimo capitolo si approfondirà l'analisi ai dati tramite modelli di regressione lineare multivariati e poi si proverà con i modelli Probit per stimare prima cosa porta a una stima ragionevole della frequenza dei controlli e dopo a una stima elevava della stessa.

CAPITOLO 3

ANALISI STATISTICA DEI DATI

3.1 Modello di regressione lineare (OLS)

In molte applicazioni si raccolgono dati relativi a una variabile d'interesse quantitativa, detta *Y*, e inoltre, sulla *i*-esima unità statistica, *i=1,…,n*, sono rilevati anche i valori di *k* variabili concomitanti, esse possono essere quantitative o esprimere in forma quantitativa livelli di un fattore (variabili fattoriali possono essere il sesso o il mezzo di trasporto). Spesso, poi, si vuole studiare la relazione intercorrente tra le variabili che caratterizzano il fenomeno osservato. Nello studio empirico delle relazioni tra variabili l'uso di metodi statistici gioca un ruolo centrale, in quanto per studiare un fenomeno o un evento si ricorre prevalentemente al modello di regressione. Esso consiste nell'uguagliare una variabile risposta (Y) alla somma di due termini, ovvero la componente sistematica che esprime la relazione con le variabili esplicative (*k* variabili concomitanti) e la componente accidentale, o più comunemente d'errore, la quale rappresenta gli scostamenti di natura casuale tra Y e la componente sistematica.
L'analisi che verrà proposta in seguito farà riferimento a un modello di regressione lineare, il quale prevede che nella componente sistematica vi sia una relazione lineare tra le variabili esplicative e la variabile risposta.

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon \qquad i=1,\ldots,n$$

Come più volte ripetuto, lo scopo di questa indagine è capire in che modo la

frequenza dei controlli sugli autobus possa essere influenzata dalle variabili ricavate grazie alle due interviste sottoposte al campione. Cercando di calcolare un modello appropriato, si devono selezionare dei parametri che possano essere adeguati al parametro di riferimento. Tra questi si ritiene logico inserire la variabile che indica se l'intervistato ha esibito o meno il biglietto convalidato, *ticket*, e la dummy che indica con 1 se la persona era italiana oppure 0 se era straniera. Potrebbe essere interessante verificare se la variabile numerica associata alla risposta "quanti giorni in una settimana tipo prendi l'autobus" influisce sulla stima della frequenza dei controlli, poiché se una persona viaggia spesso potrebbe conoscere meglio l'assiduità con cui i controllori svolgono la loro mansione. Ipoteticamente anche le condizioni meteorologiche e il clima possono influenzare la percezione dei viaggiatori, infatti il bel tempo può favorire lo spostamento dei controllori da un autobus all'altro. Nonostante il test T di Student porti a escludere che il sesso femminile e quello maschile abbiano percezioni diverse sui controlli, la variabile sesso verrà inserita lo stesso a scopo cautelativo. Poiché la specificazione conterrebbe tanti parametri simili tra loro da stimare sulla base di poche osservazioni, per trovare un effetto delle variabili principali sulla stima della frequenza dei controlli, vengono generate ulteriori variabili che rendano l'interpretazione più semplice e immediata. Nel dettaglio vengono generate:

- *young*: età minore di 30 anni;
- *middleage*: età tra i 30 e i 50 anni;
- *old*: età maggiore di 50 anni;
- *risk_concerned*: se l'individuo è propenso o fortemente propenso al rischio;
- *risk_unconcerned*: se l'individuo è avverso o fortemente avverso al rischio;
- *loweduc*: se il grado di istruzione è licenza elementare o media;
- *higheduc*: se il grado di istruzione è diploma di maturità, laurea o master/dottorato;
- *warmday*: temp inferiore al livello 3;

- *warm*: temp uguale al livello 3;
- *cold*: temp superiore al livello 3;
- *cloudy*: giorno nuvoloso;
- *sunny*: giorno soleggiato;
- *rainy*: giorno piovoso;
- *dress_elegant*: individuo vestito elegante;
- *dress_poor*: individuo vestito povero;
- *evasionhigh*: percentuale stimata dei viaggiatori senza biglietto superiore al 75%;
- *finehigh*: costo stimato della multa superiore a 50 euro;
- *lowincome*: reddito mensile percepito minore o uguale a 2000 euro;
- *highincome*: reddito mensile percepito superiore a 2000 euro.

Si esclude, invece, che l'occupazione e l'etnia possano alterare la frequenza attesa dei controlli.

Per trovare la miglior specificazione, viene utilizzato il metodo di selezione *stepwise*, utilizzato soprattutto in presenza di molteplici variabili per la stima di modelli multivariati e/o generalizzati (*glm*). Esso nasce dalla necessità di selezionare un sottoinsieme "ottimo" tra un gran numero di variabili esplicative per la costruzione di un modello efficiente.
In merito alle scelte del criterio di analisi *stepwise*, si hanno tre possibilità:

- la "*forward selection*", con cui si costruisce il modello ottimo aggiungendo una a una le variabili partendo da zero;
- la "*backward elimination*", dove partendo da un modello "completo" si eliminano via via le variabili che risultano non significative e apportano uno scarso contributo predittivo;
- la "*stepwise regression analysis*", che combina i due procedimenti sopra descritti.

In questa analisi si procederà con la *backward selection*.

A tal fine è necessario prevedere una penalizzazione crescente al decrescere del numero di parametri, in quanto vi è un *trade-off* tra la complessità di un modello stimato e l'adattamento del modello ai dati. Esistono due criteri adatti a quantificare la bontà di adattamento del modello: il criterio di Akaike (AIC) e il criterio di Schwartz (BIC). Essi sono calcolati in questo modo:

$$AIC = -2log\, L_k + 2k \qquad BIC = -2log\, L_k + k\, log(n)$$

Dove L_k è il valore della funzione di verosimiglianza calcolata in corrispondenza delle stime di massima verosimiglianza dei parametri del modello a k parametri e k è il numero di parametri inseriti nel modello, inoltre la quantità 2k agisce da fattore di penalizzazione in termini di numero delle esplicative.
La regola è quella di preferire i modelli con l'AIC e/o il BIC più basso. Quando essi aumentano, la bontà di adattamento del modello viene meno. Tra i due, il BIC penalizza maggiormente i parametri aggiuntivi.

Una regola importante per poter stimare un modello di regressione lineare è che in presenza di variabili fattoriali bisogna evitare la collinearità tra i parametri; altrimenti avviene la così detta "trappola delle dummy". Le possibili scelte per raggirare il problema in un modello lineare semplice sono o eliminare la costante dalla stima oppure eliminare un livello delle variabili fattoriali.
In questo caso si è preferito mantenere la costante e inserire nella specificazione del modello j-1 livelli per ciascun parametro.

La prima specificazione, come anticipato, prevede l'inserimento di tutte le variabili che possono avere effetto sulla stima percentuale dei controlli sugli autobus:

```
control ~ young + old + alone + alternative + risk_unconcerned + rainy +
cloudy + cold + higheduc + dress_elegant + dress_poor + evasionhigh + finehigh
+ fine_cost + in_time + highincome + male + delay + n_day + nationality +
occasional_trav + satisfy + ticket + minutes
```

```
      Source |       SS       df       MS              Number of obs =     125
-------------+------------------------------           F( 24,   100) =    2.11
       Model |   25017.047    24  1042.37696           Prob > F      =  0.0055
    Residual |   49393.753   100   493.93753           R-squared     =  0.3362
-------------+------------------------------           Adj R-squared =  0.1769
       Total |     74410.8   124  600.087097           Root MSE      =  22.225
------------------------------------------------------------------------------
         control |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
           young |   .9865078   6.423157     0.15   0.878    -11.75685    13.72987
             old |   12.83065   8.471411     1.51   0.133     -3.97639    29.63768
           alone |   3.150497   5.609723     0.56   0.576    -7.979034    14.28003
     alternative |   .4372993   4.668029     0.09   0.926    -8.823937    9.698535
risk_unconcerned |    1.32761   4.591771     0.29   0.773    -7.782333    10.43755
           rainy |  -8.020779   7.045709    -1.14   0.258    -21.99926    5.957707
          cloudy |  -5.016653   7.001534    -0.72   0.475     -18.9075    8.874192
            cold |   7.077587    7.59733     0.93   0.354    -7.995299    22.15047
        higheduc |   1.332268   5.298402     0.25   0.802     -9.17961    11.84415
   dress_elegant |  -10.17392   8.172695    -1.24   0.216    -26.38832     6.04047
      dress_poor |  -11.53237   12.62688    -0.91   0.363    -36.58373    13.51899
     evasionhigh |   7.445922   4.657769     1.60   0.113     -1.79496     16.6868
        finehigh |  -3.213413   6.143902    -0.52   0.602    -15.40274    8.975913
       fine_cost |   .2550048   .1810809     1.41   0.162    -.1042545     .614264
         in_time |  -12.71324   7.020828    -1.81   0.073    -26.64236    1.215883
      highincome |  -1.702714   7.902558    -0.22   0.830    -17.38116    13.97574
            male |  -9.517749   4.634079    -2.05   0.043    -18.71163   -.3238688
           delay |  -1.441649   1.175487    -1.23   0.223    -3.773783    .8904843
           n_day |   1.971795   3.110397     0.63   0.528    -4.199144    8.142734
     nationality |  -20.34678   5.166386    -3.94   0.000    -30.59674   -10.09682
 occasional_trav |  -5.375324   13.97047    -0.38   0.701    -33.09235     22.3417
         satisfy |   .2377369   4.592465     0.05   0.959    -8.873583    9.349057
          ticket |  -13.47831   4.910251    -2.74   0.007     -23.2201   -3.736509
         minutes |   .1954049   .8011916     0.24   0.808    -1.394136    1.784946
           _cons |   35.96264   24.05881     1.49   0.138    -11.76935    83.69464
------------------------------------------------------------------------------
```

Output 14: modello OLS "completo".

Come si può vedere dall'output 14, la nazionalità, la variabile che identifica se l'intervistato avesse il titolo di viaggio convalidato e il sesso sono parametri significativamente diversi da 0 e per questo motivo influenzano la stima dei controlli sugli autobus.

Dal modello si evince che se l'intervistato è di nazionalità italiana, la stima che fornirà agli intervistatori sarà in media, a parità di altre variabili, minore di 20.35 punti percentuali rispetto a un altro intervistato di nazionalità straniera.

Era da immaginare anche il risultato che proviene dalla variabile *ticket*, chi ha esibito un titolo di viaggio regolare, sottostima in media di quasi il 13.5% la frequenza dei controlli, rispetto chi invece viaggia illegalmente.
Il parametro associato al sesso indica, invece, che gli intervistati uomini sottostimano di quasi il 9.52% la variabile risposta rispetto le donne.
Il valore denominato *R-squared* (coefficiente di determinazione) misura la frazione della varianza della risposta spiegata dal modello, esso varia tra 0, quando il modello non spiega la variabile risposta, e 1, quando il modello spiega perfettamente la variabile risposta. In questo caso indica che circa il 34% della variabilità della stima della frequenza dei controlli è spiegato dalle variabili della specificazione.

Si calcolano i criteri di Akaike e di Schwartz per questo modello:

```
Measures of Fit for regress of control

Log-Lik Intercept Only:     -576.682     Log-Lik Full Model:         -551.071
D(100):                     1102.143     LR(24):                       51.222
                                         Prob > LR:                     0.001
R2:                            0.336     Adjusted R2:                   0.177
AIC:                           9.217     AIC*n:                      1152.143
BIC:                         619.311     BIC':                         64.657
```

Output 15: calcolo dei criteri di bontà di adattamento ai dati del modello.

Si riscontra un elevato valore di AIC e di BIC, quindi si ravvede la necessità di cercare una specificazione più parsimoniosa e migliore.

In maniera iterativa sono state eliminate in ordine le variabili: *satisfy, alternative* e *young*. Alla fine è stato, quindi, ottenuto il modello:

```
control ~ old + alone + risk_unconcerned + rainy + cloudy + cold + higheduc +
dress_elegant + dress_poor + evasionhigh + finehigh + fine_cost + in_time +
highincome + male + delay + n_day + nationality + occasional_trav + ticket +
minutes
```

```
      Source |       SS       df       MS              Number of obs =     125
-------------+------------------------------           F( 21,   103) =    2.48
       Model |  24999.0059    21  1190.42885           Prob > F      =  0.0013
    Residual |  49411.7941   103  479.726157           R-squared     =  0.3360
-------------+------------------------------           Adj R-squared =  0.2006
       Total |     74410.8   124  600.087097           Root MSE      =  21.903
------------------------------------------------------------------------------
         control |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
             old |   12.12171   7.393503     1.64   0.104    -2.541561    26.78498
           alone |   2.973439   5.359454     0.55   0.580    -7.655775    13.60265
risk_unconcerned |   1.395543   4.472793     0.31   0.756    -7.475187    10.26627
           rainy |  -8.055079   6.665388    -1.21   0.230    -21.27431    5.164147
          cloudy |  -5.025424   6.818752    -0.74   0.463    -18.54881    8.497962
            cold |   7.048607     7.4434     0.95   0.346    -7.713621    21.81084
        higheduc |   1.270055   5.164579     0.25   0.806     -8.97267    11.51278
   dress_elegant |  -10.29513   7.491093    -1.37   0.172    -25.15195    4.561683
      dress_poor |  -11.46898   12.40933    -0.92   0.358    -36.07995      13.142
     evasionhigh |   7.520729   4.527037     1.66   0.100    -1.457581    16.49904
        finehigh |  -3.422919   5.850038    -0.59   0.560    -15.02509    8.179251
       fine_cost |   .2557054   .1746185     1.46   0.146    -.0906091      .60202
         in_time |  -12.57631   6.842991    -1.84   0.069    -26.14777    .9951441
      highincome |  -1.715386   7.750432    -0.22   0.825    -17.08654    13.65577
            male |  -9.372926   4.485526    -2.09   0.039    -18.26891   -.4769429
           delay |  -1.430581   1.156661    -1.24   0.219    -3.724545    .8633825
           n_day |   1.936865   3.003019     0.64   0.520    -4.018915    7.892645
     nationality |  -20.28621   5.015029    -4.05   0.000    -30.23233   -10.34008
 occasional_trav |  -5.439947   13.70274    -0.40   0.692    -32.61609     21.7362
          ticket |  -13.52016     4.8288    -2.80   0.006    -23.09695   -3.943377
         minutes |   .1894637   .7820408     0.24   0.809     -1.36153    1.740457
           _cons |   37.17916   22.65572     1.64   0.104     -7.75312    82.11144
------------------------------------------------------------------------------
```

Output 16: modello ridotto.

Non è stata raggiunta la significatività da altre variabili oltre quelle già segnalate precedentemente perciò l'effetto che provocano sulla variabile risposta è pressoché invariato.

Anche il coefficiente di determinazione, *R-squared*, non cambia rispetto il modello completo (output 14), ma facendo riferimento a *R-squared adj* (coefficiente di determinazione corretto), che calcola la bontà del modello a livello di parsimonia dei parametri, esso ora è pari a 0.2006, mentre prima era 0.1769: è migliorato.

La bontà di adattamento ai dati è migliorata anche se non di molto, infatti si

osservi l'output:

```
Measures of Fit for regress of control

Log-Lik Intercept Only:     -576.682     Log-Lik Full Model:         -551.094
D(103):                     1102.188     LR(21):                       51.176
                                         Prob > LR:                     0.000
R2:                            0.336     Adjusted R2:                   0.201
AIC:                           9.170     AIC*n:                      1146.188
BIC:                         604.872     BIC':                         50.218
```

Output 17: : calcolo dei criteri di bontà di adattamento ai dati del modello ridotto (output 16).

Si calcola il test di White per verificare l'omoschedasticità:

```
White's general test statistic :       125  Chi-sq(124)  P-value =  .4579
```

Output 18: test di White per il modello ridotto (output 16)

L'ipotesi nulla di omoschedasticità viene accettata, la componente erratica ha varianza uguale.

Il modello ridotto appena proposto migliora leggermente l'adattamento ai dati, d'altro canto, riducendo il numero di variabili nella specificazione, l'AIC tornerebbe ad assumere un valore alto e anche il coefficiente di determinazione si ridurrebbe a 0.1361. Infatti, procedendo con il metodo *stepwise*, i criteri peggiorano nonostante si raggiunga la significatività di tutti i parametri. Il modello che si raggiunge è:

```
control ~ n_day + nationality

      Source |       SS       df       MS              Number of obs =     159
-------------+------------------------------           F(  2,   156) =   12.29
       Model |  13996.6097     2  6998.30486           Prob > F      =  0.0000
    Residual |  88848.9878   156  569.544793           R-squared     =  0.1361
-------------+------------------------------           Adj R-squared =  0.1250
       Total |  102845.597   158  650.921503           Root MSE      =  23.865
------------------------------------------------------------------------------
     control |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       n_day |   3.215925   1.122948     2.86   0.005     .9977792    5.434071
 nationality |  -15.38955   3.787198    -4.06   0.000    -22.87036   -7.908747
       _cons |   20.07807   6.727604     2.98   0.003     6.789121    33.36702
------------------------------------------------------------------------------
```

Output 19: modello con tutte le variabili significative.

```
Measures of Fit for regress of control

Log-Lik Intercept Only:     -740.142     Log-Lik Full Model:         -728.511
D(156):                     1457.023     LR(2):                       23.260
                                         Prob > LR:                    0.000
R2:                            0.136     Adjusted R2:                  0.125
AIC:                           9.201     AIC*n:                     1463.023
BIC:                         666.274     BIC':                       -13.122
```

Output 20: criteri AIC e BIC per il modello dell'output 19.

Come anticipato, si perde molto in termini di bontà di adattamento ai dati, in quanto con il modello ridotto riportato all'output 16 si era raggiunto un AIC pari a 1146.188, e non ne vale la pena dato il significativo aumento dei due indici di riferimento.

3.2 Modello OLS con trasformazione logaritmica

Un modello di regressione lineare senza trasformazioni non è sempre corretto; infatti in presenza di parametri che possono assumere valori di grandezza differenti andrebbe verificato se il modello può essere migliorato grazie ad alcune trasformazioni.

Vediamo che la variabile dipendente e la variabile *evasion*, poiché sono stime percentuali, assumono un *range* che va da 0 a 100, il quale, quindi, passa da unità a centinaia. Si prova a trasformare la variabile control con il log(1+control), denominata lcontrol; si è reso necessario aggiungere una unità alla trasformazione in quanto se le osservazioni avessero assunto il valore 0 ci sarebbero stati problemi con il dominio della funzione logaritmo. Si stima un nuovo modello con la stessa specificazione iniziale di prima. Ne risulta:

```
lcontrol ~ young + old + alone + alternative + risk_unconcerned + rainy +
cloudy + cold + higheduc + dress_elegant + dress_poor + evasionhigh + finehigh
+ fine_cost + in_time + highincome + male + delay + n_day + nationality +
occasional_trav + satisfy + ticket + minutes
```

```
      Source |       SS       df       MS              Number of obs =     125
-------------+------------------------------           F( 24,   100) =    2.95
       Model |  63.5267914    24  2.64694964           Prob > F      =  0.0001
    Residual |  89.7035972   100  .897035972           R-squared     =  0.4146
-------------+------------------------------           Adj R-squared =  0.2741
       Total |  153.230389   124  1.23572894           Root MSE      =  .94712
------------------------------------------------------------------------------
        lcontrol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
           young |   .0238195   .2737269     0.09   0.931    -.5192468    .5668858
             old |   .4023628   .3610145     1.11   0.268    -.3138797    1.118605
           alone |   -.126309   .2390619    -0.53   0.598     -.600601    .3479829
     alternative |  -.0866569    .198931    -0.44   0.664    -.4813303    .3080165
risk_unconcerned |   .0135693   .1956812     0.07   0.945    -.3746567    .4017953
           rainy |   .0234974   .3002573     0.08   0.938    -.5722046    .6191993
          cloudy |   .2263528   .2983748     0.76   0.450    -.3656142    .8183199
            cold |   .0648782    .323765     0.20   0.842    -.5774624    .7072187
        higheduc |  -.0768383   .2257947    -0.34   0.734    -.5248087     .371132
   dress_elegant |  -.3812026   .3482845    -1.09   0.276    -1.072189     .309784
      dress_poor |  -.0761189   .5381023    -0.14   0.888    -1.143698    .9914606
     evasionhigh |   .3624582   .1984938     1.83   0.071    -.0313478    .7562641
        finehigh |  -.4145028   .2618262    -1.58   0.117    -.9339586     .104953
       fine_cost |   .0166055   .0077169     2.15   0.034     .0012954    .0319155
         in_time |  -.5701074    .299197    -1.91   0.060    -1.163706    .0234909
      highincome |  -.1753943   .3367725    -0.52   0.604    -.8435413    .4927528
            male |  -.5566258   .1974842    -2.82   0.006    -.9484288   -.1648228
           delay |  -.0604377   .0500941    -1.21   0.230     -.159823    .0389477
           n_day |    .075242   .1325515     0.57   0.572    -.1877365    .3382205
     nationality |  -.9798359   .2201688    -4.45   0.000    -1.416644   -.5430273
 occasional_trav |   -.241176   .5953605    -0.41   0.686    -1.422354    .9400023
         satisfy |   .0528262   .1957108     0.27   0.788    -.3354584    .4411109
          ticket |  -.6507927   .2092534    -3.11   0.002    -1.065946   -.2356398
         minutes |  -.0026286   .0341433    -0.08   0.939    -.0703679    .0651107
           _cons |    3.42409   1.025281     3.34   0.001     1.389961    5.458219
------------------------------------------------------------------------------
```

Output 21: modello OLS con variabile dipendente il logaritmo di 1 più la frequenza dei controlli.

Oltre a *ticket*, la nazionalità e il sesso, dopo la trasformazione della variabile dipendente, anche la variabile *fine_cost*, che è una stima della multa, assume significatività al 5%.

Gli uomini dichiarano una stima della frequenza dei controlli inferiore di 42.7 punti percentuali rispetto alle donne.

La nazionalità anche ora ha un ruolo rilevante: gli intervistati italiani sottostimano i controlli del 62.5% rispetto agli stranieri.

Inoltre, la parte del campione che ha esibito un titolo di viaggio regolare ritiene che i controllori eseguano verifiche sui biglietti pari al 47.8% in meno di quelli che viaggiano irregolarmente.
Infine, l'aumento di un euro nella stima della multa porta a ritenere che i controlli aumentino dell' 1.7%.
Ora il coefficiente di determinazione è pari a 0.4146, ovvero circa il 41.5% della variabilità della variabile dipendente è spiegata da questo modello.

Per questo modello l'indice di Akaike e di Schwartz risultano:

```
Measures of Fit for regress of lcontrol

Log-Lik Intercept Only:     -190.094     Log-Lik Full Model:         -156.630
D(100):                      313.259     LR(24):                       66.929
                                         Prob > LR:                     0.000
R2:                            0.415     Adjusted R2:                   0.274
AIC:                           2.906     AIC*n:                       363.259
BIC:                        -169.572     BIC':                         48.951
```

Output 22: AIC e BIC per il modello "completo" con variabile dipendente log(1+control).

Come fatto precedentemente, si procede con il metodo di selezione *stepwise* e, dopo aver tolto dalla specificazione in ordine le variabili *risk_unconcerned, rainy, minutes, dress_poor, cold, young, occasional_trav, satisfy, higheduc* e *alternative*. Pertanto si ottiene il modello:

```
lcontrol ~ old + alone + cloudy + dress_elegant + evasionhigh + finehigh +
fine_cost + in_time + highincome + male + delay + n_day + nationality + ticket
```

```
      Source |       SS       df       MS              Number of obs =     127
-------------+------------------------------           F( 14,   112) =    5.60
       Model |  63.6129778    14  4.54378413           Prob > F      =  0.0000
    Residual |  90.9226462   112  .811809341           R-squared     =  0.4116
-------------+------------------------------           Adj R-squared =  0.3381
       Total |  154.535624   126  1.22647321           Root MSE      =    .901
------------------------------------------------------------------------------
     lcontrol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
          old |   .3820446   .2597928     1.47   0.144    -.1327015    .8967907
        alone |  -.1629301    .191367    -0.85   0.396    -.5420993    .2162392
       cloudy |   .2506057   .1854098     1.35   0.179    -.1167601    .6179715
dress_elegant |  -.3771943   .2952978    -1.28   0.204    -.9622891    .2079005
  evasionhigh |   .3651905   .1822504     2.00   0.048     .0040848    .7262962
     finehigh |  -.4612107   .2343261    -1.97   0.052    -.9254978    .0030764
    fine_cost |   .0173691   .0071341     2.43   0.016     .0032339    .0315043
      in_time |  -.5610317   .2601401    -2.16   0.033    -1.076466   -.0455974
   highincome |  -.1656724   .3103516    -0.53   0.595    -.7805943    .4492496
         male |  -.5573138   .1782921    -3.13   0.002    -.9105767   -.2040509
        delay |  -.0572857    .046179    -1.24   0.217    -.1487834     .034212
        n_day |    .132167   .0506168     2.61   0.010     .0318762    .2324577
  nationality |  -.9660198   .1864523    -5.18   0.000    -1.335451   -.5965884
       ticket |  -.6610907    .184549    -3.58   0.001    -1.026751   -.2954306
        _cons |   3.028494   .4489116     6.75   0.000     2.139033    3.917955
------------------------------------------------------------------------------
```

Output 23: modello OLS "ridotto" con variabile dipendente il logaritmo di 1 più la frequenza dei controlli.

Si riscontra un significativo miglioramento delle stime eliminando le variabili che portano uno scarso contributo predittivo.

Dalle stime si evince chi ritiene che i viaggiatori senza biglietto superino il 75% sul totale viaggiatori sovrastima una frequenza dei controlli pari al 44.1%.

Un euro in più nella stima della multa porta ad aumentare la percezione delle verifiche dei titoli di viaggio dell' 1.7%.

Se l'autobus risulta in orario, si presume che i controlli siano meno frequenti di del 43%.

Come nel modello precedente, la variabile sesso rimane significativa, si osserva che gli uomini sottostimino la frequenza del 42.7% rispetto le donne.

Inoltre, un giorno in più in cui si utilizza l'autobus per spostarsi induce gli intervistati a ritenere che i controlli siano più frequenti del 14.1%.

Mantengono, poi, lo stesso effetto negativo sia la variabile che si riferisce alla

nazionalità sia la dummy che indica con 1 gli intervistati che hanno esibito un titolo di viaggio regolare. Gli italiani dichiarano una stima inferiore del 61.9% rispetto agli stranieri e chi ha un biglietto obliterato sottostima la variabile dipendente del 48%.
Nonostante il coefficiente di determinazione sia peggiorato, si può notare, invece, che *R-squared* corretto è migliorato: ora è pari a 0.3381 e prima era 0.2741.

Vediamo ora la bontà di adattamento ai dati:

```
Measures of Fit for regress of lcontrol

Log-Lik Intercept Only:     -192.666     Log-Lik Full Model:         -158.985
D(112):                      317.970     LR(14):                       67.363
                                         Prob > LR:                     0.000
R2:                            0.412     Adjusted R2:                   0.338
AIC:                           2.740     AIC*n:                       347.970
BIC:                        -224.579     BIC':                          0.456
```

Output 24: AIC e BIC per il modello dell'output 23.

Si può notare un miglioramento di entrambi gli indici che prima erano pari a AIC= 363.259 e BIC= -169.572; inoltre, dal test per la verifica di omoschedasticità si conviene che si accetta l'ipotesi nulla e viene pertanto esclusa l'eteroschedasticità.

```
White's general test statistic :  108.8264  Chi-sq(101)  P-value =  .2797
```

Output 25: test di White applicato al modello dell'output 23.

Se si volesse cercare un modello con tutti i predittori significativi, si otterrebbe un indice di AIC molto elevato.
Quest'ultimo modello ha la specificazione:

```
lcontrol ~ male + n_day + nationality + ticket
```

```
      Source |       SS       df       MS              Number of obs =     158
-------------+------------------------------           F(  4,   153) =   10.54
       Model |    40.65059     4  10.1626475           Prob > F      =  0.0000
    Residual |  147.574371   153  .964538369           R-squared     =  0.2160
-------------+------------------------------           Adj R-squared =  0.1955
       Total |  188.224961   157   1.1988851           Root MSE      =  .98211
------------------------------------------------------------------------------
    lcontrol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -.370816   .1681235    -2.21   0.029    -.7029591    -.038673
       n_day |    .120926   .0462713     2.61   0.010     .0295128    .2123392
 nationality |  -.7747979   .1642016    -4.72   0.000    -1.099193   -.4504029
      ticket |    -.48466   .1730655    -2.80   0.006    -.8265666   -.1427535
       _cons |    3.24376   .3182936    10.19   0.000     2.614942    3.872578
------------------------------------------------------------------------------
```

Output 26: modello ridotto con tutte le variabili significative ma indici di AIC e BIC elevati.

Si può notare che gli uomini stimano che la frequenza dei controlli sia inferiore del 31% in meno rispetto le donne.
Nel caso si viaggi un giorno in più, si è indotti a sovrastimare i controlli del 12.9%.
Come nei casi precedenti, la nazionalità provoca un effetto negativo sulla stima: gli intervistati italiani forniscono un valore inferiore del 53.9% rispetto gli stranieri.
Infine, chi possiede il biglietto valido sottostima le verifiche del 38.4%.

Il coefficiente di determinazione indica che il modello spiega male la variabile risposta rispetto al modello ridotto precedente.
Viene mostrata ora la bontà di adattamento ai dati del modello.

```
Measures of Fit for regress of lcontrol

Log-Lik Intercept Only:     -238.021     Log-Lik Full Model:         -218.800
D(153):                      437.599     LR(4):                        38.442
                                         Prob > LR:                     0.000
R2:                            0.216     Adjusted R2:                   0.195
AIC:                           2.833     AIC*n:                       447.599
BIC:                        -336.978     BIC':                        -18.192
```

Output 27: AIC e BIC per il modello dell'output 26.

Si registra un netto peggioramento degli indici rispetto al modello ridotto dell'output 23 il cui AIC valeva 347.970, inoltre, quest'ultimo ha un BIC più grande di 100 (-224.579).

3.3 Modello Probit sulla frequenza ragionevole dei controlli

Come preannunciato alla fine del secondo capitolo, verrà proposta ora la stima di un modello Probit che avrà come variabile dipendente una dummy denominata *controlrag* che indicherà con 1 una ragionevole stima della frequenza dei controlli, ovvero una percentuale tra il 20% e il 40%, 0 altrimenti. Si mantiene come metodo per selezionare la specificazione migliore la *backward selection*, il modello completo equivale a:

```
probit ( controlrag ~ young + old + alone + alternative + risk_unconcerned +
rainy + cloudy + cold + higheduc + dress_elegant + dress_poor + evasionhigh +
finehigh + fine_cost + in_time + male + delay + n_day + nationality +
occasional_trav + satisfy + ticket + minutes )
```

```
Probit regression                                 Number of obs   =        136
                                                  LR chi2(23)     =      38.32
                                                  Prob > chi2     =     0.0236
Log likelihood = -64.911671                       Pseudo R2       =     0.2279
------------------------------------------------------------------------------
      controlrag |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
           young |  -.9043989   .4088957    -2.21   0.027     -1.70582   -.1029781
             old |  -.1187839   .5524445    -0.22   0.830    -1.201555    .9639874
           alone |  -1.489068   .4267099    -3.49   0.000    -2.325404   -.6527322
     alternative |    .124987   .2808598     0.45   0.656    -.4254882    .6754621
risk_unconcerned |   .1328041   .2913621     0.46   0.649     -.438255    .7038632
           rainy |   .2783042   .4713352     0.59   0.555    -.6454957    1.202104
          cloudy |   .7970202   .4686624     1.70   0.089    -.1215413    1.715582
            cold |  -.5703223    .459746    -1.24   0.215    -1.471408    .3307633
        higheduc |   .6364553    .359295     1.77   0.076      -.06775    1.340661
   dress_elegant |  -.1213703   .4643661    -0.26   0.794    -1.031511    .7887705
      dress_poor |   .7032596   .7935822     0.89   0.376     -.852133    2.258652
     evasionhigh |  -.0493642   .2944823    -0.17   0.867    -.6265389    .5278104
        finehigh |  -.3805735   .3990683    -0.95   0.340    -1.162733    .4015859
       fine_cost |   .0236501   .0122143     1.94   0.053    -.0002895    .0475898
         in_time |   .3145929    .433204     0.73   0.468    -.5344713    1.163657
            male |  -.4932721   .3164656    -1.56   0.119    -1.113533    .1269891
           delay |    .079867   .0720868     1.11   0.268    -.0614206    .2211546
           n_day |   .0834565   .1922934     0.43   0.664    -.2934317    .4603448
     nationality |  -.5913158   .3285577    -1.80   0.072    -1.235277    .0526455
 occasional_trav |   1.006235   .8388551     1.20   0.230    -.6378904    2.650361
         satisfy |   .1118203   .2906246     0.38   0.700    -.4577935     .681434
          ticket |  -.1716752   .3281024    -0.52   0.601    -.8147441    .4713937
         minutes |   .0531308   .0565899     0.94   0.348    -.0577834     .164045
           _cons |  -1.227789   1.492944    -0.82   0.411    -4.153906    1.698329
------------------------------------------------------------------------------
```

Output 28: modello probit completo con variabile dipendente la probabilità della stima ragionevole della frequenza dei controlli.

Le uniche variabili significative al 5% sono *alone* e *young*.

Si calcolano, quindi, gli indici di bontà di adattamento ai dati per il modello completo.

```
Measures of Fit for probit of controlrag

Log-Lik Intercept Only:      -84.069     Log-Lik Full Model:          -64.912
D(112):                      129.823     LR(23):                       38.315
                                         Prob > LR:                     0.024
McFadden's R2:                 0.228     McFadden's Adj R2:            -0.058
Maximum Likelihood R2:         0.246     Cragg & Uhler's R2:            0.346
McKelvey and Zavoina's R2:     0.439     Efron's R2:                    0.263
Variance of y*:                1.781     Variance of error:             1.000
Count R2:                      0.779     Adj Count R2:                  0.286
AIC:                           1.308     AIC*n:                       177.823
BIC:                        -420.394     BIC':                         74.676
```

Output 29: AIC e BIC per il modello completo dell'output 28.

Si ricordi che un modello Probit non è interpretabile come un semplice modello di regressione lineare, infatti i coefficienti stimati non corrispondono alla vera influenza che esercita il regressore sulla variabile risposta, ma indicano semplicemente il segno della probabilità. Mentre nel modello OLS gli effetti marginali coincidono con i coefficienti, nel modello Probit essi corrispondono ai coefficienti moltiplicati per la densità della distribuzione.

Per poter interpretare i valori stimati bisogna ricorrere all'output seguente.

```
Probit regression, reporting marginal effects           Number of obs =    136
                                                        LR chi2(23)   =  38.32
                                                        Prob > chi2   = 0.0236
Log likelihood = -64.911671                             Pseudo R2     = 0.2279
------------------------------------------------------------------------------
contro~g |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
   young*|  -.3109226   .1417689    -2.21   0.027   .676471  -.588785 -.033061
     old*|  -.0370588   .1668567    -0.22   0.830      .125  -.364092  .289974
   alone*|  -.4702609   .1196681    -3.49   0.000   .544118  -.704806 -.235716
altern~e*|   .0403808   .0909737     0.45   0.656   .441176  -.137924  .218686
risk_u~d*|   .0429441   .0944262     0.46   0.649   .433824  -.142128  .228016
   rainy*|   .0943756   .1672304     0.59   0.555   .176471   -.23339  .422141
  cloudy*|   .2606183   .1528364     1.70   0.089   .433824  -.038936  .560172
    cold*|  -.1688123   .1241771    -1.24   0.215   .308824  -.412195   .07457
higheduc*|   .2153428   .1246014     1.77   0.076   .338235  -.028872  .459557
dress_~t*|  -.0377447   .1395426    -0.26   0.794   .095588  -.311243  .235754
dress_~r*|   .2612765   .3139543     0.89   0.376   .029412  -.354063  .876616
evasio~h*|    -.01579   .0937289    -0.17   0.867   .330882  -.199495  .167915
finehigh*|  -.1153645   .1123506    -0.95   0.340   .286765  -.335568  .104839
fine_c~t |   .0076072   .0038408     1.94   0.053   44.1261   .000079  .015135
 in_time*|   .0997274   .1352273     0.73   0.468   .558824  -.165313  .364768
    male*|   -.161724   .1040186    -1.56   0.119   .580882  -.365597  .042149
   delay |   .0256896   .0231854     1.11   0.268   1.83088  -.019753  .071132
   n_day |   .0268442   .0619482     0.43   0.664   5.55882  -.094572   .14826
nation~y*|  -.1875796   .1013113    -1.80   0.072   .485294  -.386146  .010987
occasi~v*|   .3662521   .3133097     1.20   0.230   .176471  -.247824  .980328
 satisfy*|   .0361114   .0940972     0.38   0.700   .441176  -.148316  .220539
  ticket*|  -.0559494   .1080538    -0.52   0.601      .625  -.267731  .155832
 minutes |   .0170898   .0182132     0.94   0.348   3.80882  -.018607  .052787
---------+--------------------------------------------------------------------
  obs. P |   .3088235
 pred. P |   .2558285  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Output 30: calcolo degli effetti marginali del modello probit completo.

Chi viaggia da solo ha una probabilità del 47% inferiore a chi viaggia in compagnia di qualcuno di dichiarare una frequenza ragionevole.
Inoltre, chi ha meno di 30 anni ha tale probabilità del 31.1% inferiore rispetto a qualcuno più anziano.

Procedendo con il metodo di selezione delle variabili prescelto, si eliminano dalla

specificazione in ordine le variabili *evasionhigh, old, dress_elegant, satisfy, alternative, ticket, risk_unconcerned, n_day, minutes, in_time, delay, dress_poor, finehigh, , rainy* e *male*, quindi si giunge agli indici di Akaike e di Schwartz più bassi per il modello:

```
probit ( controlrag ~ young + alone + cloudy + cold + higheduc + fine_cost +
nationality + occasional_trav )

Probit regression                                 Number of obs   =        143
                                                  LR chi2(8)      =      32.68
                                                  Prob > chi2     =     0.0001
Log likelihood = -71.926063                       Pseudo R2       =     0.1851
------------------------------------------------------------------------------
     controlrag |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          young |  -.7581528   .3116804    -2.43   0.015    -1.369035   -.1472705
          alone |  -1.312584   .3549783    -3.70   0.000    -2.008329   -.6168392
         cloudy |   .7755242   .3777034     2.05   0.040     .0352392    1.515809
           cold |  -.6484573   .4017797    -1.61   0.107    -1.435931    .1390164
       higheduc |   .5113443   .3058404     1.67   0.095    -.0880919    1.110781
      fine_cost |   .0133084   .0085438     1.56   0.119    -.0034372     .030054
    nationality |  -.4128086   .2559495    -1.61   0.107    -.9144605    .0888433
occasional_trav |    .747045   .3098492     2.41   0.016     .1397518    1.354338
          _cons |  -.2226978   .5121126    -0.43   0.664     -1.22642    .7810243
------------------------------------------------------------------------------
```

Ouput 31: modello Probit ridotto.

Ora si può notare che, oltre *alone* e *young*, diventano rilevanti al 5% le variabili *cloudy* e *occasional_trav*.

Si osservino gli indici di Akaike e Schwartz che sono inferiori rispetto al modello completo, dove erano rispettivamente 177.823 e -420.394:

```
Measures of Fit for probit of controlrag

Log-Lik Intercept Only:      -88.266     Log-Lik Full Model:          -71.926
D(134):                      143.852     LR(8):                        32.679
                                         Prob > LR:                     0.000
McFadden's R2:                 0.185     McFadden's Adj R2:             0.083
Maximum Likelihood R2:         0.204     Cragg & Uhler's R2:            0.288
McKelvey and Zavoina's R2:     0.354     Efron's R2:                    0.212
Variance of y*:                1.548     Variance of error:             1.000
Count R2:                      0.734     Adj Count R2:                  0.136
AIC:                           1.132     AIC*n:                       161.852
BIC:                        -521.169     BIC':                          7.024
```

Output 32: AIC e BIC calcolati sul modello ridotto dell'output 31.

Vediamo gli effetti marginali che provocano sulla variabile dipendente.

```
Probit regression, reporting marginal effects           Number of obs =    143
                                                        LR chi2(8)    =  32.68
                                                        Prob > chi2   = 0.0001
Log likelihood = -71.926063                             Pseudo R2     = 0.1851
------------------------------------------------------------------------------
contro~g |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
   young*|  -.2630578   .1089086    -2.43   0.015   .664336  -.476515 -.049601
   alone*|  -.4305714   .1063062    -3.70   0.000   .559441  -.638928 -.222215
  cloudy*|   .2599013   .1259084     2.05   0.040   .426573   .013125  .506677
    cold*|  -.1948468   .1086714    -1.61   0.107   .307692  -.407839  .018145
higheduc*|   .1748965   .1065814     1.67   0.095    .34965  -.033999  .383792
fine_c~t |   .0043873   .0027928     1.56   0.119   44.1969  -.001087  .009861
nation~y*|  -.1346535   .0821736    -1.61   0.107   .475524  -.295711  .026404
occasi~v*|   .2713338   .1173205     2.41   0.016   .188811    .04139  .501278
---------+--------------------------------------------------------------------
  obs. P |   .3076923
 pred. P |    .268401  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Output 33: effetti marginali del modello ridotto dell'output 31.

Si può notare che gli intervistati con meno di 30 anni hanno una probabilità di fornire una frequenza ragionevole dei controlli del 26.3% in meno rispetto a gli altri intervistati.

Chi ha risposto di essere solo ha la probabilità del 43% inferiore in confronto a chi era in compagnia di qualcuno.

Se il tempo della giornata in cui è stata fatta l'intervista era nuvoloso, l'intervistato aumentava la probabilità di percezione di controlli tra il 20 e il 40% quasi del 26%.

Infine, gli intervistati che risultano viaggiatori occasionali, in quanto in una settimana tipo viaggiano meno di 4 giorni, hanno una probabilità maggiore del 27.1%.

Come prima, viene proposta la stima del modello con tutti i predittori significativi, ma la bontà del modello risulta pessima in confronto al modello ridotto mostrato prima:

```
probit ( controlrag ~ alone + cloudy + cold + higheduc + occasional_trav )

Probit regression                                 Number of obs   =        158
                                                  LR chi2(5)      =      19.77
                                                  Prob > chi2     =     0.0014
Log likelihood = -86.288904                       Pseudo R2       =     0.1028
------------------------------------------------------------------------------
     controlrag |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          alone |  -.8084281   .2693476    -3.00   0.003     -1.33634   -.2805165
         cloudy |   .8443366   .3406765     2.48   0.013      .176623    1.51205
           cold |  -.7874544   .3587809    -2.19   0.028    -1.490652   -.0842567
       higheduc |   .5548164   .2724469     2.04   0.042     .0208303    1.088803
occasional_trav |   .6889534   .2767912     2.49   0.013     .1464526    1.231454
          _cons |  -.5897726   .1787731    -3.30   0.001    -.9401613   -.2393838
------------------------------------------------------------------------------
```

Output 34: modello probit con tutti I parametric significativi.

Ma, come anticipato, rispetto il modello precedente dove l'indice di Akaike e di Schwartz valevano rispettivamente 161.852 e -521.169, questo modello si adatta in maniera peggiore ai dati:

```
Measures of Fit for probit of controlrag

Log-Lik Intercept Only:      -96.175     Log-Lik Full Model:          -86.289
D(152):                      172.578     LR(5):                        19.773
                                         Prob > LR:                     0.001
McFadden's R2:                 0.103     McFadden's Adj R2:             0.040
Maximum Likelihood R2:         0.118     Cragg & Uhler's R2:            0.167
McKelvey and Zavoina's R2:     0.206     Efron's R2:                    0.117
Variance of y*:                1.259     Variance of error:             1.000
Count R2:                      0.728     Adj Count R2:                  0.085
AIC:                           1.168     AIC*n:                       184.578
BIC:                        -596.937     BIC':                          5.540
```

Output 35: AIC e BIC del modello all'output 34.

3.4 Modello Probit sulla frequenza elevata dei controlli

Inoltre, è stato ritenuto interessante osservare cosa porta a ritenere elevata la frequenza dei controlli sugli autobus. A tal scopo è stata calcolata la mediana dei valori assunti dalla variabile *control*, che risulta pari a 20%. Successivamente, è stata generata una variabile dummy denominata come *highcontrol* con valore pari a 1 se la probabilità di stimare una frequenza delle verifiche sui titoli di viaggio era superiore a questa percentuale.
Si procede con una regressione Probit e si calcola il modello completo riportato di seguito:

```
probit ( highcontrol ~ young + old + alone + alternative + risk_unconcerned +
rainy + cloudy + cold + higheduc + dress_elegant + dress_poor + evasionhigh +
finehigh + fine_cost + in_time + highincome + male + delay + n_day +
nationality + occasional_trav + satisfy + ticket + minutes )
```

Per la stima del modello si rimanda il lettore alla pagina successiva.

```
Probit regression                                 Number of obs   =        125
                                                  LR chi2(24)     =      62.13
                                                  Prob > chi2     =     0.0000
Log likelihood = -55.479346                       Pseudo R2       =     0.3589
------------------------------------------------------------------------------
     highcontrol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
           young |  -.2889905   .4604754    -0.63   0.530    -1.191506    .6135247
             old |   .4891458   .6149655     0.80   0.426    -.7161644    1.694456
           alone |  -.2218486   .3832877    -0.58   0.563    -.9730786    .5293814
     alternative |   .1544403   .3312596     0.47   0.641    -.4948166    .8036972
risk_unconcerned |  -.3245177   .3289032    -0.99   0.324    -.9691561    .3201207
           rainy |  -.8051759   .4678803    -1.72   0.085    -1.722204    .1118527
          cloudy |  -1.146693   .5049822    -2.27   0.023     -2.13644   -.1569459
            cold |   .8927732   .5439938     1.64   0.101    -.1734351    1.958982
        higheduc |  -.3050109   .3782051    -0.81   0.420    -1.046279    .4362574
   dress_elegant |  -.6400606   .6581675    -0.97   0.331    -1.930045    .6499239
      dress_poor |   .0161467   .9106605     0.02   0.986    -1.768715    1.801008
     evasionhigh |   .6440808   .3413963     1.89   0.059    -.0250438    1.313205
         finehigh |  -.6180385   .4714774    -1.31   0.190    -1.542117    .3060403
        fine_cost |   .0321685   .0132541     2.43   0.015      .006191    .0581461
          in_time |  -1.217643   .5163388    -2.36   0.018    -2.229648   -.2056371
      highincome |  -.7123787   .5670929    -1.26   0.209     -1.82386     .399103
            male |  -1.043165   .3685623    -2.83   0.005    -1.765534   -.3207961
           delay |  -.1019254   .0851855    -1.20   0.231    -.2688858    .0650351
           n_day |   .3135009   .2373807     1.32   0.187    -.1517567    .7787586
     nationality |  -1.868808   .4306074    -4.34   0.000    -2.712783   -1.024833
 occasional_trav |   1.271474   1.055392     1.20   0.228     -.797057    3.340004
         satisfy |  -.2397419   .3243982    -0.74   0.460    -.8755507    .3960668
          ticket |  -.7959719   .3476418    -2.29   0.022    -1.477337   -.1146065
         minutes |  -.0048619   .0575215    -0.08   0.933     -.117602    .1078783
           _cons |   .3926653   1.722282     0.23   0.820    -2.982945    3.768275
------------------------------------------------------------------------------
```

Output 36: modello probit completo con variabile dipendente la probabilità di una stima elevata della frequenza dei controlli sugli autobus.

Sono significative le variabili *cloudy, fine_cost, in_time, male, nationality* e *ticket*.

Vediamo i valori dei criteri di Akaike e di Schwartz per questo modello.

```
Measures of Fit for probit of highcontrol

Log-Lik Intercept Only:      -86.543     Log-Lik Full Model:          -55.479
D(100):                      110.959     LR(24):                       62.128
                                         Prob > LR:                     0.000
McFadden's R2:                 0.359     McFadden's Adj R2:             0.070
Maximum Likelihood R2:         0.392     Cragg & Uhler's R2:            0.522
McKelvey and Zavoina's R2:     0.631     Efron's R2:                    0.401
Variance of y*:                2.711     Variance of error:             1.000
Count R2:                      0.776     Adj Count R2:                  0.533
AIC:                           1.288     AIC*n:                       160.959
BIC:                        -371.873     BIC':                         53.751
```

Output 37: AIC e BIC per il modello completo dell'output 36.

Vengono calcolati, quindi, gli effetti marginali:

```
Probit regression, reporting marginal effects           Number of obs =    125
                                                        LR chi2(24)   =  62.13
                                                        Prob > chi2   = 0.0000
Log likelihood = -55.479346                             Pseudo R2     = 0.3589
------------------------------------------------------------------------------
highco~l |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
   young*|  -.1148761   .1818981    -0.63   0.530      .672   -.47139  .241638
     old*|    .192021   .2317349     0.80   0.426      .136  -.262171  .646213
   alone*|  -.0882032   .1518695    -0.58   0.563       .56  -.385862  .209456
altern~e*|   .0614496    .131589     0.47   0.641      .448   -.19646  .319359
risk_u~d*|  -.1284356   .1289381    -0.99   0.324       .44   -.38115  .124278
   rainy*|  -.2973462   .1513883    -1.72   0.085      .176  -.594062  -.00063
  cloudy*|  -.4273232   .1659574    -2.27   0.023        .4  -.752594 -.102053
    cold*|    .341309   .1883518     1.64   0.101      .264  -.027854  .710472
higheduc*|  -.1204094   .1473614    -0.81   0.420      .336  -.409233  .168414
dress_~t*|  -.2386185    .217324    -0.97   0.331      .096  -.664566  .187329
dress_~r*|   .0064308   .3628345     0.02   0.986      .032  -.704712  .717573
evasio~h*|   .2524304   .1288711     1.89   0.059      .344  -.000152  .505013
finehigh*|  -.2377283   .1705754    -1.31   0.190      .272   -.57205  .096593
fine_c~t |   .0128057   .0052738     2.43   0.015   44.0652   .002469  .023142
 in_time*|  -.4573277   .1712263    -2.36   0.018      .544  -.792925  -.12173
highin~e*|  -.2616164   .1796161    -1.26   0.209      .088  -.613657  .090425
    male*|  -.3980373   .1283639    -2.83   0.005       .56  -.649626 -.146449
   delay |  -.0405747    .033912    -1.20   0.231     1.888  -.107041  .025892
   n_day |   .1247993   .0944584     1.32   0.187     5.544  -.060336  .309934
nation~y*|  -.6490889   .1111902    -4.34   0.000      .504  -.867018  -.43116
occasi~v*|   .4508391    .286533     1.20   0.228      .176  -.110755  1.01243
 satisfy*|  -.0950499   .1277715    -0.74   0.460      .408  -.345477  .155378
  ticket*|  -.3093311   .1279321    -2.29   0.022        .6  -.560073 -.058589
 minutes |  -.0019354   .0228971    -0.08   0.933     3.816  -.046813  .042942
---------+--------------------------------------------------------------------
  obs. P |        .48
 pred. P |   .4738156  (at x-bar)
```

Output 38: effetti marginali del modello completo.

Dall'output sopra riportato si evince che una giornata nuvolosa comporta una diminuzione della probabilità di stimare un'elevata frequenza dei controlli del 42.7%.
L'aumento di un euro nella stima della multa produce l'aumento della probabilità di una percezione elevata delle verifiche dell' 1.3%.
Se l'autobus risulta in orario, la probabilità diminuisce del 45.7%.
Lo stesso effetto negativo che comporta l'orario è dato anche dal fatto se l'intervistato è di sesso maschile, gli uomini hanno una probabilità del 39.8% in meno delle donne.
La nazionalità straniera, invece, apporta un effetto positivo sulla probabilità di una stima elevata della frequenza dei controlli; uno straniero ritiene che sia maggiore del 64.9% rispetto agli italiani.
Inoltre, chi ha esibito un titolo di viaggio regolare ha una probabilità inferiore del 30.9% di stimare un'elevata frequenza in confronto a chi viaggia irregolarmente.

Dopo aver eliminato ordinatamente dalla specificazione le variabili che meno contribuiscono a spiegare la variabile dipendente, cioè *dress_poor, minutes, alternative, young, alone, dress_elegant, delay, satisfy, n_day, occasional_trav* e *finehigh*, si giunge al modello che secondo la *stepwise selection* è il migliore:

```
probit ( highcontrol ~ old + risk_unconcerned + rainy + cloudy + cold +
higheduc + evasionhigh + fine_cost + in_time + highincome + male + nationality
+ ticket )
```

```
Probit regression                                 Number of obs   =        128
                                                  LR chi2(13)     =      53.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -61.801225                       Pseudo R2       =     0.3023
------------------------------------------------------------------------------
     highcontrol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
             old |   .6959977   .4488418     1.55   0.121    -.1837161    1.575712
risk_unconcerned |   -.324458    .292029    -1.11   0.267    -.8968243    .2479082
           rainy |  -.7584581   .4161487    -1.82   0.068    -1.574095    .0571783
          cloudy |  -1.075464    .414935    -2.59   0.010    -1.888722   -.2622068
            cold |   .8057247   .4502931     1.79   0.074    -.0768336    1.688283
        higheduc |  -.3553997   .3126363    -1.14   0.256    -.9681556    .2573562
     evasionhigh |   .6126505    .305495     2.01   0.045     .0138913     1.21141
       fine_cost |    .016635   .0086704     1.92   0.055    -.0003586    .0336286
         in_time |  -.8784827   .3065072    -2.87   0.004    -1.479226   -.2777397
      highincome |  -.4896826   .5288819    -0.93   0.355    -1.526272    .5469068
            male |  -.9842073   .3212929    -3.06   0.002     -1.61393   -.3544847
     nationality |  -1.650203   .3802429    -4.34   0.000    -2.395465   -.9049405
          ticket |  -.8758414    .316652    -2.77   0.006    -1.496468   -.2552149
           _cons |   1.963437   .6187965     3.17   0.002     .7506182    3.176256
------------------------------------------------------------------------------
```

Output 39: modello probit ridotto.

Ora sono sei le variabili significative: *cloudy, evasiohigh, in_time, male, nationality* e *ticket*.

Mentre i criteri di bontà di adattamento ai dati nel modello completo erano pari a AIC= 160.959 e BIC= -371.873, questo modello risulta migliore, infatti questi valori sono diminuiti:

```
Measures of Fit for probit of highcontrol

Log-Lik Intercept Only:      -88.582     Log-Lik Full Model:          -61.801
D(114):                      123.602     LR(13):                       53.562
                                         Prob > LR:                     0.000
McFadden's R2:                 0.302     McFadden's Adj R2:             0.144
Maximum Likelihood R2:         0.342     Cragg & Uhler's R2:            0.456
McKelvey and Zavoina's R2:     0.550     Efron's R2:                    0.339
Variance of y*:                2.221     Variance of error:             1.000
Count R2:                      0.734     Adj Count R2:                  0.443
AIC:                           1.184     AIC*n:                       151.602
BIC:                        -429.529     BIC':                          9.515
```

Output 40: AIC e BIC del modello ridotto.

Per capire in che modo le variabili significative condizionano la probabilità di una stima elevata della frequenza dei controlli vengono calcolati gli effetti marginali:

```
Probit regression, reporting marginal effects           Number of obs =    128
                                                        LR chi2(13)   =  53.56
                                                        Prob > chi2   = 0.0000
Log likelihood = -61.801225                             Pseudo R2     = 0.3023
------------------------------------------------------------------------------
highco~l |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
     old*|   .2667962   .1568162     1.55   0.121   .132813  -.040558   .57415
risk_u~d*|  -.1285415   .1146108    -1.11   0.267     .4375  -.353174  .096092
   rainy*|  -.2830102    .138356    -1.82   0.068   .171875  -.554183 -.011837
  cloudy*|  -.4049265   .1402431    -2.59   0.010    .40625  -.679798 -.130055
    cold*|   .3101867   .1597883     1.79   0.074   .265625  -.002993  .623366
higheduc*|   -.140278   .1214794    -1.14   0.256   .351563  -.378373  .097817
evasio~h*|   .2404198   .1157467     2.01   0.045    .34375    .01356  .467279
fine_c~t |   .0066274   .0034554     1.92   0.055   44.5949  -.000145    .0134
 in_time*|  -.3394948   .1110804    -2.87   0.004   .546875  -.557208 -.121781
highin~e*|  -.1873665   .1882339    -0.93   0.355   .085938  -.556298  .181565
    male*|  -.3773533   .1135476    -3.06   0.002   .554688  -.599902 -.154804
nation~y*|  -.5903284   .1080842    -4.34   0.000   .507813   -.80217 -.378487
  ticket*|  -.3382515   .1141172    -2.77   0.006   .609375  -.561917 -.114586
---------+--------------------------------------------------------------------
  obs. P |   .4765625
 pred. P |   .4792751  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Output 41: effetti marginali del modello ridotto.

Si può notare che se il giorno dell'intervista era nuvoloso, la probabilità di fornire una stima dei controlli di oltre il 20% diminuisce di quasi il 40.5%.
Se l'intervistato riteneva che i viaggiatori senza biglietto superassero la soglia del 75%, egli aveva una probabilità di stimare la variabile risposta del 24% in più.
Se l'autobus veniva dichiarato in orario, gli intervistati avevano una probabilità negativa del 33.9% e di poco superiore (-37.7%) era tale percentuale nel caso in cui essi fossero di genere maschile.

Gli italiani dichiaravano una stima dei controlli maggiore del 20% con una probabilità del 59% in meno.
Infine, se veniva esibito un regolare titolo di viaggio, la probabilità di una stima elevata era di meno il 33.8%.

Per correttezza si propone ora il modello cui si giungeva dopo aver eliminato tutte le variabili non significative, ma, come nei casi precedenti, si fa presente che gli indici di Akaike e di Schwartz peggiorano rispetto il modello ridotto appena trovato.

```
probit ( highcontrol ~ fine_cost + in_time + male + nationality + ticket )

Probit regression                                 Number of obs   =        143
                                                  LR chi2(5)      =      36.00
                                                  Prob > chi2     =     0.0000
Log likelihood = -81.034278                       Pseudo R2       =     0.1817
------------------------------------------------------------------------------
 highcontrol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   fine_cost |   .0164408   .0076056     2.16   0.031      .001534    .0313476
     in_time |   -.513895   .2364503    -2.17   0.030     -.977329   -.0504609
        male |  -.7474557   .2597411    -2.88   0.004    -1.256539   -.2383725
 nationality |   -1.11344   .2547945    -4.37   0.000    -1.612828   -.6140517
      ticket |  -.6806507   .2564196    -2.65   0.008    -1.183224   -.1780774
       _cons |   .9118014   .4464903     2.04   0.041     .0366965    1.786906
------------------------------------------------------------------------------
```

Output 42: modello probit con tutte le variabili esplicative significative.

```
Measures of Fit for probit of highcontrol

Log-Lik Intercept Only:      -99.033     Log-Lik Full Model:          -81.034
D(137):                      162.069     LR(5):                        35.997
                                         Prob > LR:                     0.000
McFadden's R2:                 0.182     McFadden's Adj R2:             0.121
Maximum Likelihood R2:         0.223     Cragg & Uhler's R2:            0.297
McKelvey and Zavoina's R2:     0.342     Efron's R2:                    0.216
Variance of y*:                1.519     Variance of error:             1.000
Count R2:                      0.692     Adj Count R2:                  0.362
AIC:                           1.217     AIC*n:                       174.069
BIC:                        -517.841     BIC':                        -11.182
```

Output 43: AIC e BIC calcolati sul modello dell'output 42.

Si tenga presente che il modello ridotto migliore trovato con il metodo *stepwise* risultava avere l'indice di Akaike pari a 151.602 e di Schwartz pari a -429.529.

CONCLUSIONI

L'indagine proposta è il frutto dell'elaborazione dei dati raccolti attraverso due interviste sottoposte ai viaggiatori degli autobus della città di Reggio Emilia.
Lo scopo era quello di capire cosa portava il campione ad avere delle aspettative più o meno alte sui controlli dei titoli di viaggio anche in base alle proprie caratteristiche personali.
Dopo aver illustrato le statistiche di base, ovvero le medie, alcuni test per il confronto delle stesse e le frequenze osservate, sono stati costruiti diversi modelli statistici al fine di spiegare nella maniera migliore la percezione della frequenza dei controlli in base alle informazioni disponibili.
Al fine di dichiarare quale combinazione di variabili potesse risultare efficiente nel spiegare la variabile risposta, è stato molto utile il metodo di selezione *stepwise*. Si partiva da un modello "completo", in cui si introducevano tra i predittori tutte le variabili ritenute responsabili di ciò che risultava dalla stima della frequenza dei controlli, e si eliminavano ordinatamente le variabili esplicative non significative. Si giungeva, quindi, al miglior modello quando i criteri di bontà di adattamento ai dati (AIC e BIC), calcolati per tutte le specificazioni, risultavano i più bassi.
Per primo è stato cercato il miglior modello di regressione lineare che spiegasse quali effetti avevano le variabili ritenute opportune sulla stima della frequenza dei controlli sui biglietti dell'autobus. Si è ottenuto un modello con tre variabili significative, quali la nazionalità italiana, il sesso maschile ed il fatto se l'intervistato avesse esibito un titolo di viaggio regolare. Tutti i parametri risultavano avere un effetto negativo sulla variabile risposta; infatti, si è riscontrato che gli intervistati con queste caratteristiche sottostimavano la frequenza delle verifiche sui biglietti.
Dato che la variabile dipendente, *control*, assumeva valori di grandezze diverse, si è ritenuto opportuno riprendere la stessa specificazione iniziale del primo modello stimato e porre come variabile risposta il logaritmo di *control* precedentemente sommato a uno. Utilizzando sempre lo stesso metodo di

selezione delle variabili da inserire nelle specificazioni, è stato ottenuto un modello più parsimonioso di quello "ridotto" senza trasformazioni. Inoltre, sono aumentate le variabili significative all'interno della stima. Oltre alle variabili nominate prima, le quali hanno mantenuto l'effetto negativo, sono risultate significative anche l'elevata percezione dei viaggiatori senza biglietto, la stima della multa, il fatto se l'autobus dell'intervistato era in orario e il numero medio di giorni alla settimana in cui si utilizzava il mezzo di trasporto.

La trasformazione della variabile dipendente ha dato esito positivo: sono migliorate le stime, si è raggiunto un modello parsimonioso e sono state trovate altre variabili che influiscono sulla stima della frequenza dei controlli.

Successivamente, è stata eseguita una stima di un modello Probit che calcolasse la probabilità che gli intervistati fornissero una frequenza ragionevole, quindi controlli tra il 20 e il 40%. Le variabili che influenzano tale probabilità sono risultate essere l'età inferiore ai 30 anni, se si viaggiava da soli, il tempo nuvoloso e se l'intervistato era un viaggiatore occasionale.

In ultimo, sempre attraverso un modello Probit, sono state cercate le variabili che potessero influire sulla probabilità di dichiarare una frequenza elevata dei controlli (maggiore al 20%). Ancora una volta il tempo nuvoloso, il sesso, la nazionalità italiana e se l'intervistato ha esibito il titolo di viaggio hanno un effetto sulla variabile risposta, inoltre, sono risultate significative anche l'elevata stima dei viaggiatori senza biglietto e se l'autobus dell'intervistato era in orario.

Quello che si può concludere è che il fattore sesso, se si ha esibito un biglietto regolare e la nazionalità influiscono sia sulla stima della frequenza dei controlli, sia sulla probabilità che essa sia ragionevole o elevata. Gli italiani sono risultati sottostimare sempre la frequenza in confronto agli stranieri, i quali come già detto forse sono maggiormente abituati ai controlli.

BIBLIOGRAFIA

Atienza M., Ruiz Manero J. (2004), "*Illeciti atipici: l'abuso del diritto, la frode alla legge, lo sviamento di potere"*, il Mulino, Bologna

Azzalini A. (2001), "*Inferenza Statistica. Una presentazione basata sul concetto di verosimiglianza"*, 2° edizione, Springer-Verlag Italia, Milano

Bucciol A., Landini F., Piovesan M. (2012), "*Unethical Minds: Individual Characteristics that Predict Unethical Behavior*", Department of Economics, University of Verona

Del Monte A., Papagni E. (2007), "*The determinants of corruption in Italy: regional panel data analysis*", European Journal of Political Economy, 23, pp. 379-396

Del Monte A., Papagni E. (2001), "*Public Expenditure, Corruption and Economic Growth. The Case of Italy*", European Journal of Political Economy, Vol. 17, 1-16

ISTAT (2002), "*Rapporto statistico sulla Regione Emilia-Romagna*", Monografie regionali

Pace L., Salvan A. (2001), "*Introduzione alla Statistica. Inferenza, verosimiglianza, modelli"*, Cedam, Milano

Piccolo D. (2004), "*Statistica per le decisioni"*, il Mulino, Bologna

SITOGRAFIA

Ministero delle Infrastrutture e dei Trasporti:
http://www.mit.gov.it/mit/site.php

Azienda Trasporti Pubblici Reggio Emilia:
www.actre.it

Indagine della Banca d'Italia sull'economia dell'Emilia-Romagna:
http://first.aster.it/pubblicazioni/Bancaditalia_economieregionaliEmilia-Romagna_2010.pdf

www.ilsole24ore.com

www.wallstreetitalia.com

www.reggionline.com/it