# Tesi di laurea

Federico Tedeschi

# Indice

# Capitolo 1

# Introduzione

L'istruzione è una componente fondamentale del capitale umano (Becker, 1993), così come l'istruzione dei genitori è parte integrante del proprio *background* familiare. La letteratura in materia, infatti, dimostra come essa abbia un'incidenza diretta sui salari in tutto il mondo occidentale (vedi Angrist e Newey, 1991 e Flabbi, 1999). La mia tesi si occupa, oltre che dell'istruzione, anche del suo possibile impatto sulle decisioni di fecondità. In particolare, intendo valutare se due riforme del sistema scolastico italiano abbiano influenzato i conseguimenti scolastici degli individui che ne sono stati coinvolti.

La prima legge in questione è la n.1859 del 31/12/1962, che sancisce l'obbligatorietà del conseguimento della licenza media (con la possibilità, però, di venire prosciolti da tale obbligo a condizione di avere raggiunto 15 anni di età, con otto di frequenza scolastica). Contestualmente, viene cancellata da tale normativa la distinzione tra "scuola media" e "scuola di avviamento professionale".

L'altra legge, invece, è la n.910 dell'11/12/1969 ("Provvedimenti urgenti per l'università"): essa stabilisce la liberalizzazione degli accessi universitari, ossia la possibilità di iscrizione a qualsiasi facoltà universitaria per tutti i detentori di un diploma di maturità.

Il primo passo della mia tesi è relativo all'individuazione delle coorti coinvolte da queste riforme. Successivamente, valuto se tali classi (almeno le prime) abbiano realmente avuto un aumento dell'istruzione attribuibile alle riforme stesse (dato che entrambe, una istituendo un obbligo, l'altra rimuovendo un divieto, hanno, come effetto eventuale, l'incremento dell'istruzione degli individui che ne sono interessati), analizzando i risultati separatamente per uomini e donne.

In seguito, mi propongo di valutare, nel caso in cui l'esito dell'analisi sulle donne si riveli positivo, se tali coorti siano influenzate da queste riforme anche

in relazione alle scelte di fecondità. Infine, cerco di riscontrare se le riforme abbiano toccato, in qualche misura, anche i figli degli individui colpiti dalle stesse. In questo modo, si potrebbe valutare se scelte di fecondità o il livello d'istruzione dei figli sia direttamente influenzato da quello dei genitori.
Nel Capitolo 2 parlerò dell' "Indagine longitudinale sulle famiglie italiane", da cui provengono i dati a mia disposizione; nel Capitolo 3 delle riforme scolastiche di cui ho appena trattato, nel Capitolo 4 dei modelli utilizzati per stimare gli effetti delle riforme in questione sulle variabili descritte in precedenza. I Capitoli 5, 6 e 7 sono invece relativi alla elaborazione pratica di modelli (rispettivamente, su istruzione conseguita, scelte di fecondità e livello d'istruzione dei figli), mentre il Capitolo 8 è la Conclusione.

# Capitolo 2

# Indagine longitudinale sulle famiglie italiane

La "Indagine longitudinale sulle famiglie italiane"(vedi Pisati, 2000; Schizzerotto, 2002) è uno studio *panel* prospettico effettuato ad anni alterni. Le rilevazioni a mia disposizione sono quelle relative al 1997 e al 1999. Nel corso della prima rilevazione sono state raccolte le informazioni retrospettive relative a tutti gli eventi rilevanti accaduti ai membri del campione (gli individui adulti facenti parte delle famiglie sorteggiate) fino al momento dell'intervista, mentre nella seconda rilevazione gli individui sono stati classificati in due categorie: quelli già intervistati nel corso della *Prima rilevazione* e coloro che, invece, venivano intervistati per la prima volta in occasione della *Seconda rilevazione*. Infatti, mentre i primi devono semplicemente fornire ragguagli circa eventuali "nuovi episodi" occorsi in quel biennio, i "nuovi arrivati", evidentemente, hanno dovuto sottoporsi alla stessa serie di domande ricevute dagli intervistati della *Prima rilevazione*. Tale tipo di indagine si propone come strumento d'analisi relativamente a fenomeni di mutamento sociale. Questo fa sì che, all'interno del campione a mia disposizione, alcuni individui siano stati intervistati in entrambe le occasioni, altri solo nel '97, altri ancora solo nel '99.
L'obiettivo dell'indagine era quello di seguire la storia, a partire dal '97, di 4956 famiglie che fossero rappresentative della popolazione italiana (attraverso una procedura di campionamento "a due stadi", che selezionasse prima il comune, quindi la famiglia). Tuttavia, è stato selezionato anche un campione di riserva, nel caso che alcune famiglie risultassero irreperibili, o che rispondesse una parte insufficiente dei componenti della stessa (meno di 4 membri, nel caso di famiglie composte da 6 o più persone, meno di 3, per famiglie di 4 o 5 membri, meno di 2, per famiglie di 2 o 3 individui). Complessivamente, mi trovo ad avere a disposizione 5451 famiglie intervistate, 5397 delle quali

formate da soli cittadini italiani, 25 da cittadini stranieri, 29 composte sia da persone con cittadinanza straniera, che da italiani. Relativamente agli individui interpellati, invece, essi sono 11132, di cui 5313 maschi e 5819 femmine. Tuttavia, l'oggetto del mio interesse, come spiegato nell'Introduzione, sono principalmente le persone di nazionalità italiana nate tra il 1930 e il 1971: 7711 individui, dei quali 3735 uomini e 3976 donne. Solo alla fine, quando cercherò di valutare l'impatto eventuale dell'istruzione dei genitori su quella dei figli, mi baserò anche su informazioni provenienti da individui nati nel decennio successivo (ma avrò, complessivamente, un campione più ridotto, perchè limiterò la mia analisi a individui con almeno un genitore nato tra il 1936 e il 1956).
Ogni intervistato delle due rilevazioni è stato sottoposto a un questionario, formato da un "Foglio di famiglia" e da sette sezioni tematiche distinte. Esse sono:

- Mobilità geografica
- Istruzione e formazione professionale
- Lavoro
- Famiglia
- Risorse familiari
- Sostegni e sussidi
- Servizi sociali e assistenziali

Tali sezioni si suddividono a loro volta in schede, ognuna delle quali contiene una serie di domande tra loro strettamente collegate (nel senso che la risposta ad alcune determina quali siano i successivi interrogativi da porre all'intervistato). Ancora, alcune sezioni ("Foglio di famiglia", "Risorse familiari", "Servizi sociali e assistenziali") contengono domande che vengono poste solo al capofamiglia; le altre, invece, sono rivolte a tutti i membri adulti della convivenza. Questo fa sì che, evidentemente, non ci siano individui intervistati dopo il 1981.
Le sezioni che interessano la mia tesi sono: "Istruzione e formazione professionale" e "Famiglia", che riguardano tutti i membri maggiorenni della famiglia. In particolare, della prima sezione ho preso in esame esclusivamente le schede relative all'istruzione; della seconda, invece, la "Scheda di base" e quella "Figli e assistenza" (occupandomi esclusivamente delle domande relative ai propri genitori nella prima, ai propri figli nella seconda).

# Capitolo 3

# Riforme scolastiche negli anni '60

Come spiegato nell'Introduzione, la mia tesi sfrutta, soprattutto, l'entrata in vigore di due riforme che hanno avuto luogo negli anni '60, relative al sistema scolastico italiano. In questo capitolo, mi propongo di descriverle meglio e illustrare più diffusamente il contesto in cui esse hanno avuto luogo.
Nel secondo dopoguerra, il problema della scuola fu subito all'attenzione dei partiti politici, tanto che fu oggetto di dibattito già durante le elezioni del '46 per l'Assemblea Costituente. La stessa Costituzione repubblicana, peraltro, sancisce, nel primo comma dell'articolo 34, che l'istruzione inferiore sia obbligatoria e abbia durata di otto anni.
In realtà, già la "riforma Gentile" (promulgata nel 1923) stabiliva l'obbligo scolastico fino al quattordicesimo anno d'età. Tuttavia, le possibilità di prosecuzione degli studi dopo il conseguimento della licenza elementare erano numerose. Una di esse, addirittura, il "corso integrativo", di fatto, restava nell'ambito della scuola primaria: esso, infatti, era affidato a insegnanti elementari, e si proponeva il proseguimento delle materie "delle classi quarta e quinta, convenientemente approfondite con ampie letture" (Cives, 1990). Accanto ad esso, vi era la "scuola complementare", ricalcata sul canale della preesistente "scuola tecnica". Entrambe queste possibilità non prevedevano sbocchi ulteriori. Accanto ad esse, vi erano i corsi inferiori delle scuole di secondo grado (licei, istuti magistrali, istituti tecnici) destinati, ovviamente, a chi fosse stato intenzionato a proseguire gli studi oltre i quattordici anni. Si succedettero, poi, in epoca fascista, vari uomini a capo del "Ministero dell'Educazione Nazionale" (Casati, Fedele, Belluzzo, De Vecchi, Bottai). Di fatto, però, due cose restarono sempre immutate: la netta distinzione tra individui destinati alla prosecuzione degli studi in istituti superiori e ragazzi che si sarebbero fermati alla "licenza media" e il fatto che l'obbligo fosse sempre

rimasto sulla carta: non furono mai effettuati seri controlli, nè realizzate le strutture adeguate, affinchè tutti i bambini in uscita dalle scuole elementari proseguissero effettivamente gli studi.
Il fatto che l'istruzione obbligatoria fino alla conclusione delle scuole medie inferiori venisse formalizzata nella Costituzione, invece, costituiva un netto passo avanti. Tuttavia, prima di arrivare ad una legge in tal senso, si dovette aspettare per quasi 17 anni. Infatti, il dibattito politico fu molto acceso attorno a due questioni: l'unificazione o meno delle due scuole secondarie inferiori allora esistenti ("scuola di avviamento professionale" e "scuola media") e l'insegnamento del latino. L'importanza di questi due problemi era relativa alla necessità o meno di operare una scelta che condizionasse il proprio futuro prima dei quattordici anni, e all'insegnamento di una materia che si sarebbe rivelato utile solo a coloro i quali avessero proseguito gli studi in un liceo. La legge n.1859 del 31/12/1962 (che entrò in vigore, dunque, nel 1963), sanciva, come previsto dalla Costituzione, l'obbligo di terminare le scuole medie inferiori. Tuttavia, veniva stabilito che, pur in assenza dell'assolvimento dell'obbligo, si potesse venir prosciolti da esso, nel caso in cui si fossero comunque effettuati almeno otto anni di frequenza scolastica, e si fosse già raggiunto il quindicesimo anno di età. Peraltro, l'obbligo aveva effetto retroattivo (cioè, poteva riferirsi a individui già usciti dal sistema scolastico senza aver conseguito la licenza media) solo per coloro i quali non avessero ancora compiuto i 15 anni. Al contempo, si istituiva la "scuola media unificata", che prevedeva, nel corso del secondo anno, l'insegnamento di "elementi di latino"; lo studio della lingua antica, poi, proseguiva facoltativamente nel corso del terzo anno. Se, dunque, l'unificazione prolungò il periodo di indifferenziazione degli studi, è comunque da sottolineare che la divisione che aveva luogo durante l'ultimo anno di medie tra coloro i quali proseguivano nello studio del latino e chi invece decideva di interromperlo, anticipava di un anno (anche se non in maniera vincolante) la distinzione tra gli individui che avrebbero affrontato gli studi liceali (e poi, verosimilmente, frequentato l'università) e quelli che, al contrario, avrebbero intrapreso un altro tipo di studi (e quindi, probabilmente, non si sarebbero iscritti alla scuola terziaria), oppure si sarebbero addirittura fermati al conseguimento della licenza media. L'insegnamento del latino nelle scuole medie, comunque, venne abolito nel '77. Fino a quell'anno,tuttavia, il primo momento decisionale circa la carriera scolastica di un individuo risultò differito non di tre, ma di due anni (pur essendoci la possibilità di effettuare una scelta di tipo diverso l'anno successivo).
In seguito all'approvazione della legge di cui ho appena parlato, il parlamento si concentrò sui livelli scolastici successivi. Relativamente alla scuola superiore, non vi furono, però, novità salienti: accadde persino che non videro

la luce progetti di legge già approvati da uno dei due rami del parlamento. Le proposte principali di modifica della scuola media superiore erano due: la prima prevedeva l'innalzamento dell'obbligo di uno o due anni, mentre la seconda si riferiva alla possibilità di modificare l'indirizzo di studi scelto in corso d'opera, tramite l'istituzione di una scuola unitaria che, al suo interno, prevedesse una differenziazione anno dopo anno, cercando di garantire la maggiore flessibilità possibile. Di fatto, l'unica riforma adottata fu quella del '69 (decreto legge del 15/02, che poi verrà convertito in legge due mesi dopo, e prorogato nel '71), introdotta a titolo sperimentale, ma rimasta in vigore per ben trent'anni, relativa agli esami di maturità.
Per quanto riguarda l'istruzione terziaria, è da rilevare che l'Italia, negli anni '60, cominciava a fronteggiare un problema che, in assenza di provvedimenti, si sarebbe rivelato via via più drammatico: la necessità di personale qualificato che, evidentemente, l'università non era ancora in grado di "generare", se non in piccola parte, a causa del suo carattere poco professionalizzante. Questo, peraltro, fu l'argomento del primo capitolo del rapporto conclusivo di una Commissione di Indagine sulla scuola, che operò nei mesi a cavallo tra il '62 e il '63. Tuttavia, le proposte di tale commissione relativamente all'istruzione terziaria (istituzione di diplomi di laurea e corsi di dottorato, maggiore flessibilità dei piani di studio, più apertura tra le discipline e verso l'estero) vennero bocciate, nonostante fassero state alla base di un progetto di legge discusso per tre anni alla Camera, senza mai giungere all'approvazione. Stesso destino ebbe un progetto di riforma analogo, dibattuto in parlamento tra il '69 e il '71. Così, gli unici provvedimenti effettivamente adottati in quel periodo sono quelli relativi alla concessione dell' "assegno di studio universitario" (legge 14/12/1963, n.80), la liberalizzazione dei piani di studio e, soprattutto, quella degli accessi (queste due, incluse nei "Provvedimenti urgenti per l'università", legge 11/12/1969, n.910). Quest'ultima, in particolare, è quella che interessa la mia tesi: se, infatti, la riforma Gentile limitava l'accesso all'istruzione terziaria ai soli studenti provenienti dai licei[1], dal '69 in poi la possibilità d'iscrizione è stata estesa a tutti i diplomati di qualsiasi scuola media superiore (eventualmente, dopo la frequenza degli opportuni anni integrativi, nel caso di scuole di durata inferiore ai cinque anni). In realtà, già otto anni prima, la legge 685 del 21/7/1961 aveva previsto la possibilità d'iscrizione a numerose facoltà scientifiche per gli studenti provenienti dall'Istituto tecnico.

[1]Vigeva qualche eccezione: la possibilità d'iscrizione alla facoltà di "Agraria" per gli studenti provenienti dall' "Istituto tecnico"; a quella di "Scienze economiche" per i diplomati in tutti gli istitui tecnici; alla facoltà di "Architettura" per gli studenti del "Liceo artistico". Al contempo, le facoltà di "Giurisprudenza" e "Lettere e filosofia" erano precluse anche ai diplomati del "Liceo scientifico" e riservate, quindi, solo a quelli del "classico"

# Capitolo 4

# Metodi di stima utilizzati

## 4.1 Limiti dell'approccio ai minimi quadrati ordinari

Il metodo più immediato per valutare l'impatto di un'unità aggiuntiva di una qualsiasi variabile che si voglia considerare come esplicativa su un'altra che si ritenga la variabile risposta è senz'altro la regressione ai minimi quadrati ordinari ($MQO$). Tuttavia, si deve considerare che tale metodo condurre molto facilmente a stime distorte dei parametri che lo caratterizzano (e che sono i parametri d'interesse), ossia a stime che, mediamente, abbiano un valore più alto o più basso di quello del parametro.
Parlando del caso oggetto della mia tesi, in particolare, per valutare se un anno d'istruzione supplementare abbia o meno un impatto diretto sul numero di figli generati o adottati[1] entro i 35 anni[2], oppure sul momento in cui nasce il primogenito, servirsi di una regressione MQO per la stima dei parametro d'interesse significherebbe esporsi a rischi assai probabili di distorsione. Infatti, se madri più istruite operano scelte di fecondità diverse dalla media femminile, questo potrebbe essere dovuto a fattori non determinati dall'istruzione, e tuttavia correlati con essa. Infatti, le differenze relative ai conseguimenti di istruzione, verosimilmente, non sono casuali, ma dovute a differenti abilità e vincoli (Card, 1999 e 2001).
Chiamando "$S$" la variabile che rappresenta l'istruzione, si può utilizzare tale $S$ come esplicativa per la variabile dipendente "numero di figli", che chiamo:

[1]in seguito, quando si farà riferimento al nascita dei figli, s'intenderà il momento effettivo della loro procreazione nel caso di figli biologici, quello dell'adozione per i figli adottati

[2]anche qui, ogni volta che si parlerà della variabile "numero di figli", s'intenderà il numero di figli generati o adottati prima di compiere il trentacinquesimo anno

$"F"$ $\epsilon$ riassume l'insieme dei fattori che, a parità d'istruzione, determinano il numero dei figli di una donna. Supporre l'incorrelazione tra $S$ ed $\epsilon$ significa, di fatto, ipotizzare che l'istruzione non sia correlata (senza esserne la causa diretta) a nessun altro fattore che, a sua volta, sia in grado di determinare $F$, data l'istruzione stessa.

Tuttavia, effettuare la regressione MQO porterebbe verosimilmente a stime distorte del parametro d'interesse. Ipotizzando un legame lineare tra le variabili, la regressione è formalmente del tipo:

$$F = \alpha + \beta S + \epsilon \tag{4.1}$$

Illustro ora il significato dei parametri utilizzati.$\beta$ è il parametro d'interesse, che deve determinare, *ceteris paribus*, la variazione del numero di figli dovuta a un anno in più d'istruzione. Affinchè non ci sia distorsione di $\beta$, occorre che sia:

$$E[S \cdot \epsilon] = 0 \quad , \tag{4.2}$$

cioè che variabile esplicativa e termine d'errore siano incorrelati (dunque, che la $S$ sia esogena). La stima di $\beta$ coi MQO, infatti, è:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(S_i - \overline{S})(F_i - \overline{F})}{\sum_{i=1}^{n}(S_i - \overline{S})^2} \quad , \tag{4.3}$$

ove $\hat{\beta}$ è la stima di $\beta$, $"n"$ è la numerosità campionaria, $\overline{S}$ e $\overline{F}$ le medie campionarie, rispettivamente di $S$ ed $F$[3]. Di fatto, si tratta del rapporto tra la covarianza campionaria tra $S$ ed $F$, e la varianza campionaria di $S$. Dalla (4.1), si ricava che tale rapporto è pari a:

$$\beta + \frac{\sum_{i=1}^{n}(S_i - \overline{S})\epsilon_i}{\sum_{i=1}^{n}(S_i - \overline{S})^2} \tag{4.4}$$

Ora, è chiaro che la regressione MQO, per costruzione, rispetterà l'equazione (4.2); tuttavia, verosimilmente, in questo modo stimerà un parametro $\beta$ che non avrà lo stesso significato da noi richiesto: considererà, infatti, non solo quanto incida l'istruzione direttamente, ma anche quanto sia correlata al numero di figli tramite altri fattori. Ad esempio, è facile pensare che donne più istruite siano mediamente più ricche già alla nascita, e che questo (un cosiddetto "*background* economico" migliore) le porti ad avere redditi e patrimoni più cospicui in seguito, a prescindere dall' istruzione ricevuta, e che ciò abbia un effetto sul numero di figli. In realtà, qualsiasi variabile che

[3]da questo punto in poi, utilizzerò questi simboli ("$\hat{\beta}$", "n","$\epsilon$" e "$-$") sempre col significato assunto ora

sia correlata con l'istruzione (senza essere determinata da essa) e che, a sua volta, abbia un impatto diretto sul numero di figli, è una fonte di distorsione. Peraltro, esiste un'ulteriore possibile motivo di non correttezza della stima di $\beta$, qualora l'istruzione venga misurata in maniera imperfetta (vedi Ashenfelter e Rouse, 1998). Se si utilizzasse come variabile $S$, ad esempio, il numero di anni d'istruzione, è chiaro che si renderebbero necessarie alcune semplificazioni. Ad esempio, si può facilmente concordare che gli anni persi nel conseguire un determinato livello d'istruzione non siano da prendere in considerazione (infatti, pare paradossale attribuire, tra due diplomati, un valore di educazione scolastica superiore a chi abbia dovuto ripetere più anni). Nel contempo, però, ci sono alcune questioni che paiono essere più ostiche. Per esempio: come valutare un ultimo anno di studi di una persona che non sia poi riuscita a terminarlo con successo? Come considerare gli anni universitari di quegli individui che non abbiano poi conseguito la laurea? Peraltro, se si può supporre che chiunque sia in grado di ricordare il proprio titolo di studi, qualcuno potrebbe trovare problematico ricostruire con precisione la propria carriera scolastica. Per tali motivi, saremmo, in questo caso, in presenza di "errori di misura". Anche ammettendo che tali errori abbiano media nulla (ipotizzando, cioè, che eventuali risposte mancanti o inesatte non abbiano alcuna componente sistematica, e che esista un numero "vero" di anni d'istruzione attribuibile ad ogni persona) dovremmo comunque effettuare la regressione su una variabile esplicativa misurata, appunto, con errore. Quello che si verifica, in pratica, è che non osserviamo più $S$, ma, nella migliore delle ipotesi, $S+ \epsilon_1$, ove tale "$\epsilon_1$" è un termine d'errore incorrelato sia con $S$, sia con $F$. Ricordando che la (4.3) non è altro che un rapporto tra covarianza (tra $S$ ed $F$) e varianza (di $F$) campionarie (come spiegato in precedenza), è evidente che, mentre l'inserimento d'un errore di misura di tale tipo non influenza la covarianza campionaria tra $S$ ed $F$ (per la sua incorrelazione con $F$, per l'appunto), se non aumentando la variabilità della stima, esso agirà, invece, sul denominatore. La varianza teorica di $(S + \epsilon_1)$, infatti, per l'incorrelazione tra le due stesse variabili, sarà maggiore della varianza della sola $S$. Questo fa sì che, a livello campionario, a fronte di un numeratore che si modificherà solo per effetto del caso, ci sia un denominatore che, plausibilmente, sovrastimerà la varianza di $S$. Dunque, siamo in presenza di una forma di distorsione verso lo 0. E' chiaro che, al di là dell'incremento della variabilità determinato da questi fattori, non c'è nessun motivo per ritenere che queste distorsioni si compensino esattamente. Se, poi, la seconda è eliminabile riconducendoci a una differente definizione di $S$, più complicato appare la soluzione del problema dell'endogeneità di $S$.

Se, come variabile oggetto di studio, abbiamo invece il momento della nascita del primo figlio, siamo di fronte a un problema di "analisi di sopravvivenza"

(Cox, Oakes, 1984), ove ciò che interessa è la probabilità di caduta (in questo caso, nascita del primo figlio) al tempo $t$, dato che, fino a quel momento, non si ha ancora procreato (vedi Appendice). In ogni caso, pare più agevole, intuitivo ed oggettivo effettuare verifiche di altro tipo, fissando determinati momenti della vita della donna in cui verificare, appunto, la probabilità di essere "sopravvissute" (ovvero, non avere ancora procreato). In questo caso, ci troviamo palesemente di fronte a una variabile binomiale, che assume valore "1" nel caso in cui non si siano ancora generati figli, "0" altrimenti. Occorre, dunque, effettuare una regressione logistica (anche questo metodo sarà approfondito in Appendice). Anche qui, valgono in modo del tutto analogo al caso del numero di figli le obiezioni ai MQO derivanti da possibili fonti di distorsione: si può ipotizzare, infatti, che genitori più istruiti tendenzialmente procreino più tardi il primo figlio a causa di fattori non direttamente determinati dall'istruzione stessa (ad esempio, il *background* familiare).
In relazione, invece, all'istruzione dei figli, essa può essere misurata, evidentemente, in tanti modi quanto può esserlo quella dei genitori (definiamola $S_g$). In questo caso, dunque, $S$ non è più variabile esplicativa, ma dipendente. Anche qui, sussistono gli stessi problemi già visti, in quanto si può sostenere una correlazione tra istruzione dei genitori e quella dei figli non dovuta a un legame diretto causa-effetto, se si assume che famiglie più ricche generino figli destinati ad avere un $S$ maggiore, indipendentemente da $S_g$. Infatti, provenire da una famiglia più abbiente può determinare sia una maggiore capacità di sfruttare l'istruzione (cioè, più benefici di un anno ulteriore di scolarizzazione), sia redditi potenziali futuri più alti anche a parità di $S$ (quindi, maggiori costi dell'istruzione in termini di mancati guadagni): da un lato, quindi, uno stimolo allo studio; dall'altro, un deterrente. Un'altra possibile fonte di distorsione (ma, stavolta, a senso unico, cioè indubbiamente verso l'alto) è data dal fatto che genitori più istruiti avranno, verosimilmente, capacità maggiori, che hanno reso loro possibile (o, comunque, meno oneroso di quanto lo sarebbe stato per chi s'è fermato prima) raggiungere quel livello d'istruzione. Ora, è del tutto plausibile che tali capacità possano trasmettersi per via ereditaria. In questo caso, quindi, ci sarebbe una correlazione tra $S_g$ e l'abilità individuale dei figli (che, a sua volta, incide su $S$) come ulteriore elemento di esogeneità dell'istruzione dei genitori. Per quanto riguarda potenziali errori di misura, poi, vale la pena di notare che, in questo caso, anche la variabile dipendente potrebbe esserne affetta. Tuttavia, questo problema, oltre ad essere aggirabile tramite opportuna scelta di $S$, condurrebbe a una maggiore variabilità della stima (relativa al calcolo della covarianza campionaria tra variabile dipendente ed esplicativa), ma non ad una ulteriore distorsione.

## 4.2 Utilizzo di "variabili strumentali"

Per aggirare i limiti evidenziati dai MQO, servirebbe una "variabile strumentale" ($IV$, vedi Bowden e Turkington, 1984). Chiamando "$X$" la variabile esplicativa e "$Y$" la dipendente, infatti, abbiamo visto come un'espressione del tipo della (4.3), naturalmente sostituendo $X$ a $S$ e $Y$ a $F$, condurrebbe alla distorsione illustrata in (4.4), qualora ci fosse correlazione tra $X$ ed $\epsilon$. Al contrario, inserendo una nuova variabile, che chiamo $Z$, che abbia determinate caratteristiche, è possibile ottenere stime corrette (ossia, non distorte). $Z$, per essere una IV, deve avere due proprietà: essere correlata con $X$ e, al contempo, incorrelata con $\epsilon$. Lo stimatore IV (c ioè quello ottenibile tramite $Z$) è il seguente:

$$\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum_{i=1}^{n}(X_i - \overline{X})(Z_i - \overline{Z})} \tag{4.5}$$

Stavolta, quindi, siamo di fronte a un rapporto tra due covarianze campionarie: quella tra $Z$ ed $Y$ e quella tra $X$ e $Y$. Ricordando la (4.1), ed applicando di nuovo le sostituzioni tra $X$ ed $S$, $Y$ ed $F$, si ottiene che:

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})\epsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})(Z_i - \overline{Z})} \tag{4.6}$$

Risultano evidenti, a questo punto, le richieste di correlazione tra $X$ e $Z$, e, al contempo, d'incorrelazione tra $Z$ ed $\epsilon$. La prima condizione si rende necessaria perchè, in caso contrario, il denominatore dell'equazione (4.5) convergerebbe a 0 (e il rapporto, quindi, non potrebbe convergere). A livello teorico, $\beta$ è, analogamente a quello che è lo stimatore rispetto ai momenti campionari, pari al rapporto tra le covarianze, ed una frazione ha senso solo se il denominatore è non nullo. Anche a livello intuitivo, non si capisce a cosa possa servire una variabile incorrelata con tutte le altre: l'incorrelazione con la variabile esplicativa e il termine d'errore, infatti, implica anche quella con la variabile risposta, come si evince dalla (4.1). Quello che accade con le IV, d'altronde, è che si vuole misurare l'effetto della variabile esplicativa sulla risposta tramite la variabile strumentale (come dice la parola stessa): siccome $Z$ influenza $Y$ solo tramite $X$ (essendo incorrelata con $\epsilon$), si può ricavare l'impatto di un'unità aggiuntiva di $X$ su $Y$, di fatto, come il rapporto tra gli impatti di $Z$, rispettivamente, su $Y$ e $X$. Ovviamente, in caso d'incorrelazione tra $X$ e $Z$, questa operazione perde di significato. La condizione di incorrelazione della variabile strumentale col termine d'errore, invece, si rende necessaria per garantire la correttezza dello stimatore, come si evince dall'equazione (4.6). E' evidente come non sia semplice reperire variabili con tali caratteristiche. Il *background* familiare, infatti (ammesso che possa essere ben approssimato) incide, verosimilmente, sulle variabili risposta indipendentemente dall'istruzione, nella misura in cui influenza reddito e patrimonio dell'individuo. Un test come quello del "Q.I." (quoziente intellettivo), invece, è considerato poco attendibile (vedi Willis e Rosen, 1979) Peraltro, almeno relativamente all'istruzione dei figli, data che probabilmente l'intelligenza ha una componente ereditaria, avrebbe un impatto non dovuto all'istruzione dei genitori, se pensiamo che l'intelligenza d'un individuo possa influenzare il suo livello d'istruzione. Per ottenere una variabile strumentale, servirebbe dunque un esperimento casuale, in grado di incidere in modo diretto unicamente sul livello d'istruzione degli individui. Pur non essendosi mai verificato, ovviamente, un sorteggio di tale tipo, è però possibile reperire variabili analoghe. Si potrebbe, per esempio, ricorrere alla zona geografica di nascita o di residenza all'età di 14 anni, sapendo che, nel nostro paese, a differenti zone corrispondono differenti livello di reddito e di istruzione. Tuttavia, in questo modo, non si terrebbe conto del fatto che la variabile "zona geografica" sarebbe, in realtà, correlata con altri fattori, come, ad esempio, le differenze di reddito, potenzialmente in grado di influenzare le scelte relative ai figli e alla loro istruzione. Un altro elemento simile a un sorteggio (un cosiddetto "pseudoesperimento naturale") s'è verificato, in Italia, per via legislativa. Infatti, in Italia, negli anni '60, sono state introdotte alcune riforme, che hanno, verosimilmente, influenzato le scelte degli individui. La prima, infatti, nel '62, ha unificato la "scuola media inferiore" e la "scuola d'avviamento professionale", abolendo di fatto quest'ultima, e rendendo obbligatorio il conseguimento della "licenza media" o, in alternativa, la permanenza nel sistema scolastico fino ai 15 anni. Questa riforma non avrebbe dovuto riguardare le coorti antecedenti al '49, avendo esse già raggiunto i 15 anni (o, tutt'al più, facendolo proprio nel '63, anno effettivo d'entrata in vigore della riforma, per quanto riguarda i nati nel '48). I nati nel '49, invece, subiscono la riforma solo nella misura in cui non abbiano ancora ottenuto la licenza media a 14 anni (per lo più, quindi, sono persone che sono già uscite dal sistema scolastico, o hanno già conseguito la licenza stessa). Così, questa riforma diventa via via più vincolante coorte dopo coorte, fino alla classe '52: per gli individui di questa classe, infatti, una volta ottenuta la licenza elementare, c'è l'obbligo di proseguire gli studi, mentre per quelli delle classi precedenti, per lo più, si trattava di trovarsi costretti a terminare la scuola media già iniziata, o a iscriversi alla stessa dopo aver interrotto gli studi (e quanto tale obbligo fosse poi effettivo è tutto da verificare). Dal punto di vista della coercizione allo studio, comunque, l'effetto sulla coorte '52 è il medesimo delle classi successive. Tuttavia, dato che è plausibile supporre che l'istruzione sarebbe comunque aumentata nel corso del tempo (seguendo, diciamo così, una sua evoluzione "naturale") anche in assenza della riforma,

probabilmente l'effetto più forte della riforma stessa si può rilevare proprio per la classe '52, mentre va verosimilmente scemando per le coorti successive. L'altra riforma, invece, che entra in vigore dal '69, liberalizza l'accesso all'università (è utilizzata come strumento da Brunello e Miniaci, 1999, e Brunello, Comi e Lucifora, 2001, per stimare l'impatto dell'istruzione sul reddito). Infatti, mentre prima era consentita l'iscrizione alla stessa solo per gli individui provenienti dai "licei" (con alcune eccezioni che vedremo in seguito), questa riforma non poneva alcun limite di accesso, al di là del possesso di un diploma di scuola media superiore quinquennale. Quest'ultima riforma, quindi, coinvolge approssimativamente le stesse classi della precedente (in maniera piena, comunque, i nati dal '51 in poi, i quali, se hanno conseguito un diploma di scuola superiore, hanno tutti avuto poi la possibilità di iscriversi immediatamente all'università), anche se non si tratta più di un obbligo universale, ma di una possibilità offerta a uno specifico gruppo di persone, ovvero chiunque scelga di frequentare una scuola superiore diversa da un liceo. Tale riforma, peraltro, pur coinvolgendo la popolazione in misura presumibilmente minore, proprio in quanto riguarda solo una frazione di essa, protrae il suo effetto nel tempo. Infatti, mentre l'estensione dell'obbligo non fà altro che anticipare, in pratica, qualcosa che verosimilmente sarebbe avvenuto comunque nel tempo, la seconda continua tuttora a garantire la possibilità di proseguire gli studi in qualsiasi facoltà unuiversitaria a tutte le persone in possesso di un diploma di maturità.
Qualora introducessimo la variabile "$I$" (quella binaria indicante l'esposizione o meno alle riforme) non potremmo considerarla uno strumento. Infatti, il valore di $I$ è legato al fattore temporale: in pratica, il valore di $I$ è fortemente condizionato dalla coorte di appartenenza. Quindi, qualsiasi $\epsilon$ rispetto a cui tale variabile dovrebbe essere esogena, in realtà è correlata con essa, se lo è con "$T$" (variabile che misura la coorte di appartenenza). Dunque, nel nostro caso specifico, è facile supporre che la richiesta esogeneità non ci sia, perchè ci si può aspettare che persone nate dopo, a prescindere dall'istruzione ricevuta, subiscano l'influenza dei mutamenti socio-culturali relativamente alle decisioni riguardo alla prole. In pratica, essendo

$$E[T \cdot \epsilon] \neq 0 \tag{4.7}$$

si ha:

$$E[I \cdot \epsilon] \neq 0 \tag{4.8}$$

Nel paragrafo successivo, illustrerò come sia possibile, tramite il *"regression discontinuity design"* (vedi Hahn, Todd e Van der Klaw, 2001), servirsi ugualmente delle riforme in maniera strumentale, dopo averne valutata l'eventuale efficacia sul livello d'istruzione delle classi interessate.

## 4.3 Valutazione d'impatto

La mia tesi è fondamentalmente costituita dalla valutazione d'impatto di una *politica*. Con questo termine, mi riferisco ad una o più azioni intraprese da un'autorità pubblica, con lo scopo di ottenere certi risultati. Ora, in questo caso, è evidente come l'approvazione delle riforme del sistema scolastico sia una politica, il cui scopo è, palesemente, quello di aumentare l'istruzione degli individui. La valutazione d'impatto, appunto, si propone di stabilire se la politica abbia avuto un effetto sulle variabili osservate (ossia, se si siano prodotti cambiamenti che non si sarebbero osservati altrimenti).
Ciò che si vuole scoprire è la differenza tra un evento fattuale (osservato) e uno controfattuale (ipotetico). Per effettuare tale analisi, occorre che:

- la politica sia ben definita;
- la popolazione su cui viene realizzato l'intervento sia ugualmente ben definita
- l'intervento possa modificare almeno una caratteristica della popolazione, misurabile sia in presenza che in assenza della politica

Ora, la valutazione d'impatto viene effettuata, normalmente, quando sussistono dubbi sull'efficacia di un intervento, onde verificare se valga la pena di insistere o se non sia il caso di modificare (o ritirare) la legge relativa, oppure qualora si voglia effettuare o modificare un intervento, si possono preventivamente verificare i risultati ottenuti in situazioni analoghe. Nel mio caso, invece, mi servo della valutazione d'impatto per un'analisi che non ha a che fare con l'eventuale ritiro di una legge, e la utilizzo poi strumentalmente, per capire se ci sia modo di verificare un impatto di tipo diverso, ovvero quello dell'istruzione sulle scelte relative ai figli.
Vediamo ora come esprimere una valutazione d'impatto da un punto di vista formale. Chiamo:

- $Y$ la variabile obiettivo, cioè la caratteristica delle unità della popolazione che potrebbero essere modificate dall'intervento.
- $Y_i^T$ il risultato relativo a tale caratteristica ottenuto dall'i-esima unità nel caso in cui sia sottoposta all'intervento
- $Y_i^{NT}$ il risultato che, invece, l'i-esima unità sperimenterebbe in caso di mancata esposizione

(4.9)

L'impatto della politica sull'i-esima unità è dunque pari a: $\alpha_i = Y_i^T - Y_i^{NT}$ Una politica è detta *universale* qualora coinvolga tutti i membri della popolazione, *non universale* altrimenti. Nel nostro caso, la politica è senz'altro non universale: infatti, la sua retroattività, ammesso che ci sia, è limitata nel tempo, quindi non tutti gli individui vengono coinvolti (per motivi angrafici).

### 4.3.1 Il problema fondamentale

Chiamando $I$ la variabile binaria che assegna all'individuo valore 1 qualora sia esposto all'intervento, 0 altrimenti, $I_i$ informa se l'i-esima unità è o meno tra gli esposti, dunque qual è l'evento osservabile in relazione a quell' unità. Tale evento è detto anche *fattuale*, mentre quello non verificatosi è chiamato *controfattuale*. La (4.9) esprime sempre una differenza tra due variabili, di cui solo una osservabile, dunque essa, a sua volta, non è mai osservabile. Questo è stato definito da Holland (1986) come *"The fundamental problem of causal inference"*.
E' necessario, quindi, spostare l'attenzione dai singoli individui a gruppi (solitamente, la popolazione degli esposti, o, qualora non sia possibile, una frazione di essi), per valutare l'impatto medio della politica su di essi. Si rinuncia, quindi, a scoprire i singoli $\alpha$, e si calcola: $E\left[\alpha|I=1\right]$, o, comunque, un impatto medio su un gruppo. Tipicamente, può venire in aiuto il confronto tra esposti e non esposti (o, nel caso di politica universale, tra prima e dopo l'intervento).
Ancora una volta, però, c'è una differenza tra ciò che si osserva e ciò che si cerca. Infatti, per ogni individuo, si ha a disposizione: $Y^T|I=1$ e $Y^{NT}|I=0$ (a seconda che sia esposto o meno al trattamento), ovvero il solo risultato fattuale. Tuttavia, è:

$$E\left[\alpha|I=1\right] = E\left[Y^T|I=1\right] - E\left[Y^{NT}|I=1\right] \tag{4.10}$$

da cui:

$$E\left[\alpha|I=1\right] = E\left[Y^T|I=1\right] - E\left[Y^{NT}|I=0\right] + E\left[Y^{NT}|I=0\right] - E\left[Y^{NT}|I=1\right]. \tag{4.11}$$

Con opportune trasformazioni, la (4.11) diventa:

$$E\left[Y^T|I=1\right] - E\left[Y^{NT}|I=0\right] = E\left[\alpha|I=1\right] + E\left[Y^{NT}|I=1\right] - E\left[Y^{NT}|I=0\right] \tag{4.12}$$

La (4.12) illustra come la differenza che osserviamo tra esposti e non esposti è dovuta non solo all'impatto della politica, ma anche alle eventuali differenze tra i due gruppi che osserveremmo anche in assenza della stessa. Calcolando la semplice differenza tra i due gruppi, quindi, incapperemmo, qualora essi siano diversi in relazione a caratteristiche che influenzino la variabile d'interesse, in un problema di *selection bias*: avremmo, cioè, una distorsione dovuta a un problema di selezione, considerando come impatto della politica qualcosa che in realtà si sarebbe verificato ugualmente. Nel nostro caso, tale problema di selezione si presenta, in quanto esiste differenza tra esposti e non esposti: la coorte di nascita, che fa sì che gli esposti, essendo nati dopo, avrebbero comunque, verosimilmente, ottenuto risultati diversi dai non esposti.

### 4.3.2 Il matching

Le differenze non determinate dal trattamento tra individui esposti e non esposti sono dovute (a parte l'effetto della variabilità campionaria) al fatto che persone con determinate caratteristiche abbiano una probabilità più alta di essere esposte alla politica rispetto ad altre. Tali caratteristiche sono rilevanti solo qualora, oltre ad incidere sulla probabilità di esposizione, influenzino anche i risultati potenziali. Le variabili che, invece, incidono nell'esposizione o meno al trattamento essendo tuttavia incorrelate con $Y^T$ ed $Y^{NT}$ non sono da tenere in considerazione.

Qualora queste caratteristiche siano note ed osservabili, è possibile operare il *matching* (vedi Heckmann, Hichimura e Todd, 1998), ossia confrontare individui identici (o quasi) rispetto a tali caratteristiche. In particolare, tale operazione consiste nell'associare ad ogni esposto un non esposto che sia il più possibile simile ad esso. Per far questo, occorrono queste due condizioni:

- Il gruppo dei non esposti deve essere nettamente superiore, per numerosità, a quello degli esposti, per garantire una buona scelta di accoppiamento anche per quegli individui con caratteristiche rare nei non esposti e frequenti negli esposti
- Chiamando $x$ una caratteristica di interesse, deve essere:

$$0 < Pr(I = 1|x) < 1 \quad \forall x \in \mathbf{X}$$

La seconda condizione deve valere per tutte le caratteristiche che, come illustrato in precedenza, siano correlate con la probabilità di esposizione al trattamento e, contemporaneamente, influenzino gli esiti potenziali, ovvero per quel vettore $\overline{x}$ tale che:

$$(Y^T, Y^{NT}) \perp I | x$$

Se ciò si verifica, vale la *strong ignorability* (Rosenbaum e Rubin, 1983), ovvero: le caratteristiche d'interesse assumono valori tali da essere riscontrati

in entrambi i gruppi. In realtà, se sono presenti solo nei non esposti, il problema è relativo (a meno che non si voglia stimare l'impatto che avrei avuto per i non esposti), perchè è sufficiente escludere gli individui con tali valori dal gruppo di controllo. Qualora, invece, ci siano caratteristiche che sono presenti solo nel gruppo degli esposti, per quegli individui non è stimabile l'impatto, mancando un gruppo di controllo. In realtà, qualora le possibili fonti di selezione siano più di una, sarebbe sufficiente confrontare individui con caratteristiche tali da avere la stessa probabilità di esposizione, chiamata"*propensity score*"(introdotto da Rosenbaum e Rubin, 1983).
Nel mio caso, gli individui differiscono per la coorte. Questo determina differenze in una serie di variabili, a loro volta relative alle scelte che condizionano l'istruzione. Quello che potrei fare, sarebbe un confronto al margine, ovvero tra le prime persone colpite dalla riforma e le ultime non esposte (o che hanno una probabilità molto bassa di essere condizionate dalle riforme). Quello che accade, però, è che la numerosità campionaria per genere e coorte è estremamente ridotta (spesso inferiore alle 100 unità), quindi un confronto tra donne o uomini di due anni consecutivi difficilmente si rivelerà significativo. Il *matching*, dunque, pare una strada poco praticabile.

### 4.3.3 Il regression discontinuity design

Come già illustrato in precedenza, quello che servirebbe sarebbe un esperimento casuale. In quel caso, infatti, sarebbe:

$$(Y^T, Y^{NT}) \perp I \tag{4.13}$$

Questo significa che il fatto di essere esposti o meno non dà nessuna informazioni sui risultati potenziali, dunque che i due gruppi sono statisticamente uguali. Questo ci permette di identificare la distribuzione dell'evento controfattuale degli esposti mediante quello fattuale dei non esposti. In pratica, il risultato dei non esposti può essere considerato quello che avrebbero ottenuto gli individui trattati, qualora non fossero stati esposti alla politica. Formalmente:

$$E\left[Y^{NT}|I=1\right] = E\left[Y^{NT}|I=0\right] \tag{4.14}$$

Da (4.12) e (4.14), si ricava:

$$E\left[\alpha|I=1\right] = E\left[Y^T|I=1\right] - E\left[Y^{NT}|I=0\right],$$

a causa dell'annullamento del *selection bias*.
Il *regression discontinuity design* (*rdd*) si applica quando la partecipazione al programma dipende da una variabile continua e osservabile $S$. Se, in un

punto del supporto di $S$, si verifica una discontinuità, lì c'è la possibilità di applicare il *rdd*. Formalmente, chiamando $\overline{s}$ tale punto, è:

$$Pr(I(\overline{s}^+) = 1) \neq Pr(I(\overline{s}^-) = 1) \tag{4.15}$$

, intendendo con $\overline{s}^+$ e $\overline{s}^-$, rispettivamente, intorno destro e sinistro di $\overline{s}$ . Trochim(1984) distingue il *rdd* in due tipi: *sharp* e *fuzzy*. Nel primo caso, al passaggio della soglia, la probabilità di esposizione al trattamento passa da 0 a 1, dunque $I$ è una funzione deterministica di $S$.

$$I = \begin{cases} 1 & \text{se } S > \overline{s} \\ 0 & \text{altrimenti} \end{cases}$$

Quando invece, la discontinuità nel valore di soglia è di dimensione inferiore di 1, e dunque l'esposizione al trattamento non è più una funzione deterministica di $S$, si parla di *fuzzy rrd*.
Considero $Y^T$ ed $Y^{NT}$ come $Y^T(S)$ ed $Y^{NT}(S)$, ossia considero i due esiti possibili di ogni individuo come funzione della variabile $S$, mantenendo costanti le altre caratteristiche. Affinchè sia applicabile il *sharp rrd*, occorre, per stimare l'impatto sugli esposti marginali, che $Y^{NT}$ sia una funzione continua in un intorno di $\overline{s}$. Ciò equivale a dire che:

$$E\left[Y^{NT}(\overline{s}^+)\right] - E\left[Y^{NT}(\overline{s}^-)\right] = 0$$

e quindi che:

$$E\left[Y(\overline{s}^+)\right] - E\left[Y(\overline{s}^-)\right] = E\left[\alpha(\overline{s})\right]$$

In pratica, la differenza che si riscontra tra esposti e non esposti marginali, se la funzione che descrive l'esito potenziale da non esposti al variare di $s$ è continua nel valore di soglia, non può essere data dalla variazione di $S$ (che è infinitesimale). D'altronde, essendo $S$ l'unico motivo di *selection bias*, è evidente che la discontinuità eventuale non può essere altro che l'impatto della riforma sugli esposti marginali. Con un ragionamento del tutto analogo, ponendo la richiesta di contuità di $Y^T$ nel punto di soglia, si può stimare l'impatto sui non esposti marginali (misurando, quindi, quanto essi avrebbero guadagnato, se fossero stati esposti alla politica). Risulta evidente che, qualora entrambe le condizioni siano rispettate, il salto osservato può essere interpretato in entrambi i sensi: impatto effettivamente "subito" dagli esposti marginali da un lato; mancato impatto, invece, per quanto riguarda i non esposti marginali.
Quello che queste condizioni comportano è che, in un intorno di $\overline{s}$, si possa ritenere, in pratica, che l'esposizione o meno al trattamento sia casuale (vedi Thistlewaite e Campbell,1960), ovvero che:

$$(Y^T, Y^{NT}) \perp \quad I | S = \overline{s}$$

Si parla, in questo caso, di prova "quasi sperimentale"(Cook e Campbell, 1979; Rubin, 1977).
Nel caso di *fuzzy rdd*, queste due condizioni continuano a valere. Tuttavia, è possibile stimare l'impatto solo per quegli individui che, al superare la soglia, modificano il loro comportamento. A quel punto, l'impatto medio per gli individui in un intorno della soglia sarebbe pari a:

$$E\left[\alpha(\overline{s})|I(\overline{\{s}^{+})\neq I(\overline{s}^{-})\right]Pr(I(\overline{s}^{+})\neq I(\overline{s}^{-}))$$

In pratica, tale discontinuità è data dal prodotto dell'impatto sui *complier* (coloro, cioè, che effettivamente modificano il loro comportamento al passare da una parte all'altra della soglia) per la frazione di *complier* all'interno degli individui che si trovano nell'intorno della soglia (o, se si preferisce, la probabilità di essere tale, per $S = \overline{s}$). Per il *fuzzy rdd*, comunque, vale una condizione ulteriore; ovvero il fatto che l'esposizione al trattamento($I(s)$) sia monotona in un intorno di $\overline{s}$ (considerando, quindi, l'esposizione come una funzione di $S$, in mancanza di informazioni sulle altre variabili che determinano se l'individuo sarà esposto o meno).
Nel mio caso, comunque, nonostante sia difficile dividere nettamente gli individui tra "esposti" e "non esposti", e nonostante l'esposizione possa avvenire secondo diversi gradi, è possibile valutare se esistano delle discontinuità in determinati punti, e ritenere che la situazione sia simile a quella di un *fuzzy rdd*.

### I modelli utilizzati

Nell' ambito del *rdd*, utilizzerò i M.Q.O., oppure un modello di regressione che tenga conto della natura discreta della variabile risposta (di tipo *binomiale*, *poisson* o *quasipoisson* ). Illustrerò tali modelli in Appendice.

## 4.4 Metodi scelti

Avendo in precedenza descritto l'influenza che mi aspetto le riforme abbiano sulle differenti coorti, si può ritenere che non solo la classe '52 sia la più colpita dalle stesse, ma che il passaggio tra la classe '51 e la '52 sia l'unico in cui sia osservabile una differenza notevole nel livello d'istruzione. Infatti, pur essendo possibile che anche le classi precedenti abbiano risentito delle riforme, è anche vero che, esprimendo l'istruzione come funzione dell'anno di nascita, si può descriverla tramite modelli parametrici che prevedano un unico punto di discontinuità, a cavallo tra il '51 e il '52. In questo modo, infatti, mi propongo di stimare l'impatto delle riforme sulle classi '51 e '52, ovvero il

guadagno della seconda coorte sulla precedente (in termini d'istruzione) per effetto del fatto che questa subisce in pieno l'effetto delle riforme stesse.
Per ognuna delle variabili risposta con cui ho misurato l'istruzione degli individui e le loro scelte relative ai figli, ho utilizzato, come variabili esplicative, oltre all'anno di nascita, anche le potenze successive di questa variabile (la seconda, terza e quarta), stimando dapprima separatamente i modelli relativamente al periodo prima e dopo le riforme (comportandomi come se le riforme riguardassero esclusivamente le coorti dal '52 in poi) e utilizzando solo gli individui nati tra il 1930 e il 1971 (separando, peraltro, nell'analisi, uomini e donne, in linea con la letteratura precedente in materia). In particolare, la scelta del 1930 come coorte iniziale è dovuta a vari fattori. Prima di tutto, per le coorti precedenti, si evidenzia una graduale diminuzione della numerosità campionaria, che renderebbe i risultati relativi a tali classi via via più incerti. Poi, dal momento che ricerco un modello che risulti relativamente stabile nei suoi due periodi (appunto, prima e dopo le riforme), se uno di questi risulta troppo lungo, la stima richiederebbe probabilmente un numero elevato di parametri (e perderebbe di senso, non essendo ipotizzabile la minima stabilità). La sceltà del '71 come ultima coorte presa in esame, invece, è vincolata dalla necessità di "consentire" agli individui di portare a termine la propria istruzione: tenendo conto che le prime interviste hanno luogo nel '97, ho ritenuto opportuno considerare solo le classi di individui intervistate nel corso di un anno in cui ne avrebbero compiuti almeno 26, pur comprendendo che ciò potrebbe aver portato a una piccola sottostima per le ultime classi, in particolare la '71, relativamente al possesso o meno del titolo universitario. Alla fine, utilizzo i due "submodelli" ottenuti come punto di partenza nella stima del modello completo.
La potenza "i-esima" della coorte di nascita è calcolata come $(coorte\text{-}1951.5)^i$. In questo modo, valuto tali variabili separatamente per i due submodelli, a partire da valori prossimi allo 0 e potendo così osservare la significatività dei parametri risentendo in misura minore dell'arbitrarietà dei valori iniziali (1930 e 1952 rispettivamente). Valuto i modelli migliori sulla base del *fit* del modello pienamente parametrizzato (quello, in pratica, a cui si fa corrispondere ad ogni anno il valore medio campionario della coorte relativamente alla variabile risposta) e della significatività dei singoli parametri. Nel far questo, non seguo alcun presupposto teorico aprioristico, se non l'idea che difficilmente una coorte possa avere un livello d'istruzione complessivo minore della precedente. Quindi, mi baso sui due sottomodelli stimati come punto di partenza per la stima del modello complessivo. Relativamente ad esso, ho introdotto la variabile *riforma* che assume valore 1 per i nati dal '52 in poi, 0 altrimenti. Anche le interazioni tra le potenze della coorte e la riforma, peraltro, sono incluse come variabili esplicative, per permettere

la massima flessibilità al modello. D'altronde, escluderli non avrebbe senso. Presupporre che il modello sia stabile tra prima e dopo la riforma, infatti, significherebbe, relativamente alla stima dell'impatto della stessa, che si ritiene esso resti costante per tutte le coorti coinvolte (cosa che ho spiegato essere poco plausibile). Di fatto, applico un *regression discontinuity design*.
Questo modo di procedere presenta tre limiti. Il primo è che, per ottenere modelli con un buon *fit*, talvolta essi sono eccessivamente flessibili. Questo potrebbe condurre, in alcuni casi, a modelli implausibili. Qualora questo risulti eclatante, è possibile utilizzare modelli alternativi. Ad esempio, i risultati relativi alle donne che raggiungono la licenza media o il diploma superiore nel '51 sono estremamente bassi. Un modello che preveda l'eventualità di una discontinuità proprio tra qualla coorte e la successiva rischia di presentare salti evidentemente dovuti a variabilità campionaria (le donne italiane nate nel '51 nel campione sono solo 78, quindi è del tutto accettabile che la stima campionaria del valore della percentuale di queste in possesso di un dato titolo possa differire sensibilmente dal valore analogo della popolazione). Per questo, in tali casi, ho scelto modelli che, pur avendo un *fit* complessivo peggiore, avessero comunque un adattamento accettabile, e non risentissero in maniera eccessiva di tali valori anomali. Il secondo problema è dato dal fatto che, in tali modelli, non si prende in considerazione l'eventualità che le riforme possano colpire anche le coorti immediatamente precedenti alla '52 (in particolare, quelle del '49, del '50 e del '51): anche se questo, verosimilmente, accade in misura poco sensibile, è altrettanto vero che tale elemento può portare a una distorsione verso il basso della stima dell'impatto sulla coorte '52 (non tenendo conto che parte di tale impatto potrebbe già essere incorporato nelle classi precedenti). Infine, il terzo problema è relativo al fatto che, ammettendo la possibilità di includere nelle esplicative una potenza della coorte senza utilizzare tutte quelle di ordine inferiore, il modello dipende anche dal valore di intercetta scelto (di fatto, ciò equivale a fissare vincoli arbitrari).
Sulla base di queste considerazoni, effetto due controlli sui risultati ottenuti. In entrambi, impongo che i momenti dell'anno di nascita impiegati come variabili esplicative siano solo i primi due, ed escludo le interazioni di questi con *riforma* dalle variabili presenti nel modello. In particolare, calcolo il momento secondo come $(coorte\text{-}1929)^2$ (non essendoci problemi di valutazione della significatività dei parametri ed utilizzando appunto solo le prime due potenze della coorte e *riforma* come esplicative, la scelte dell'intercetta, in questo caso, non pone vincoli sul modello). Ora, diventa chiaro come ciò che ipotizzo è l'esistenza di una stabilità dell'andamento dell'istruzione in assenza di un impatto delle riforme (stabilità che, eventualmente, verrebbe alterata, qualora *riforma* si rivelasse significativo). Tale assunzione, in realtà,

è anch'essa arbitraria; tuttavia, un modello di tipo quadratico pare sensato, dal momento che un andamento lineare non può essere seguito per periodi troppo lunghi (infatti, per quanto riguarda ad esempio variabili relative all'educazione scolastica, ci si aspetta che a periodi di forte crescita iniziale segua poi un rallentamento e un assestamento, quando l'istruzione diventa di massa).
Il primo controllo che effettuo è questo: stimo il modello considerando tre possibili variabili esplicative: le prime due potenze della coorte e *riforma*, inserendo quest'ultima, però, solo qualora risulti statisticamente significativa. In questo modo, si vuole verificare, appunto, se la riforma ha avuto un impatto supponendo che, in assenza di essa, l'andamento sarebbe rimasto quadratico e stabile. Il secondo controllo, invece, tiene in considerazione anche il fatto che la riforma potrebbe aver avuto effetto anche sulle coorti '49, '50 e '51. Infatti, viene stimato un modello quadratico sulle sole osservazioni fino alla coorte '48, per poi effettuare una previsione sulle coorti successive, con l'utilizzo di bande di confidenza. Quello che mi aspetto è che, se la riforma non ha effetto, i valori relativi alle coorti successive alla '48 siano, in gran parte, interni alle bande di confidenza (interpretando, quindi, un risultato contrario come effetto della riforma). Devo ricordare, peraltro, che l'utilizzo di un *trend* quadratico ha lo scopo di smorzare un andamento crescente, non di invertirlo: se, quindi, il modello scelto dovesse portare a valori previsti decrescenti prima della coorte '71, tale previsione perderebbe di significato e, a quel punto, sarebbe preferibile utilizzare un modello lineare. Sia in questo caso, sia in quello in cui il momento secondo abbia un effetto assai ridotto sul *trend* rispetto alla componente lineare, si darà vita, probabilmente, a previsioni troppo ottimistiche ("confidando" sul fatto che il trend continui ad essere lineare, quando, di fatto, per quanto detto prima, è facile pensare che, almeno relativamente all'istruzione, essa, dal '49 in poi, in assenza della riforma, pur continuando a crescere, l'avrebbe fatto a ritmi più lenti dei precedenti).
Di fatto, le stime iniziali si basano sull'idea di costruire il modello migliore dal punto di vista dell'adattamento ai dati, seguendo unicamente il principio di presentare un andamento monotono non decrescente, ed esaltano la flessibilità senza prevedere alcuna stabilità del modello stimato sul periodo pre-riforma, nè alcun presupposto teorico a priori; i controlli, invece, si basano sull'ipotesi di un andamento quadratico stabile (poi corretto nel caso che conduca a valori decrescenti prima della coorte '71), e non pretendono di creare modelli con un buon adattamento ai dati; peraltro, anche la condizione di significatività dei parametri associati diventa meno importante. Inoltre, i controlli, prevedendo l'uso delle prime due potenze della classe di appartenenza(o della sola coorte di nascita, talvolta, nel secondo caso), non

risentono del problema dell'arbitrarietà della scelta del valore d'intercetta (se non per quanto riguarda la significatività del parametro associato alla classe di nascita, che, però, non è l'oggetto d'interesse). Di fatto, verifico i risultati ottenuti con modelli estremamente flessibili e che si adattano bene ai dati con controlli che, invece, presuppongono grande stabilità e meno arbitrarietà.
In definitiva, comunque, l'idea è quella di ottenere, tramite rapporto, l'impatto di una unità d'istruzione aggiuntiva sulle variabili dipendenti, analogamente a quanto esposto in (4.5), considerando, tuttavia, che non può essere effettuato un semplice rapporto tra le covarianze utilizzando l'intera popolazione campionaria, non essendo *riforma*, come spiegato in precedenza, una IV, se non al margine. La tesi, dunque, si può dividere in due fasi, entrambe caratterizzate da una "verifica d'impatto". Il primo *step* è relativo all'impatto delle riforme scolastiche sull'istruzione dei genitori, il secondo passo, invece, acquista significato solo qualora il primo impatto risulti significativo su almeno una variabile relativa all'istruzione, e vuole stimare l'eventuale effetto dell'istruzione sulle scelte relative ai figli e alla loro eventuale istruzione. D'altro canto, effettuando un confronto tra le sole coorti '51 e '52 (tramite *matching*), test su due campioni o modelli che prevedano l'uso della coorte di nascita come variabile esplicativa, ridotti ovviamente a queste due sole classi) per minimizzare questo problema (oltre alla potenziale distorsione verso il basso dovuta alla non considerazione dell'impatto delle riforme sulle coorti immediatamente precedenti alla '52, in particolare la '51), data la ridotta numerosità campionaria a mia disposizione, difficilmente si potrebbe apprezzare un effetto delle riforme, anche qualora esso fosse presente.

# Capitolo 5

# Impatto delle riforme sul livello d'istruzione

In questo capitolo, l'interesse sarà rivolto alla ricerca dell'eventuale impatto delle riforme scolastiche su variabili relative al livello d'istruzione degli individui. Essendo le variabili risposta essenzialmente binarie, utilizzerò, per i modelli iniziali, le stime dei coefficienti di regressione ai minimi quadrati generalizzati[1] ($GLM$) di un modello binomiale.

## 5.1 Scuola media inferiore

In questo paragrafo, analizzerò l'eventuale impatto delle riforme sul conseguimento della licenza media. Ovvio aspettarsi un'influenza della sola prima riforma, dato che risulta difficile pensare come la liberalizzazione del sistema universitario possa influenzare il comportamento degli individui in relazione all'istruzione media inferiore, tanto più considerando che la riforma universitaria entra in vigore nel '69, mentre le coorti '51 e '52 (quelle tra cui verifico l'eventuale discontinuità) si trovano ad iscriversi alle scuole medie (in assenza di incidenti di percorso), rispettivamente, nel '62 e nel '63, e in gran parte, nel '69, avranno quindi già conseguito il titolo di media inferiore (oppure saranno usciti dal sistema scolastico). A maggior ragione, le coorti dal '49 al '51, che potrebbero essere colpite dall' estensione dell'obbligo, difficilmente saranno condizionate in questo senso dalla riforma universitaria.

[1] vedi Appendice

### 5.1.1 Modello stimato

Appare più semplice la stima del modello relativo ai maschi, dal momento che, come annunciato nell'Introduzione, le femmine sono caratterizzate da un forte valore anomalo relativo al '51 (51%, contro il 63 % dell'anno precedente), cosa che creerà qualche problema nella scelta del modello. D'ora in avanti, comunque, chiamerò in questo modo le variabili che compariranno nei modelli: $x$ l'anno di nascita, $x^2$, $x^3$ e $x^4$, rispettivamente, le potenze, dalla seconda alla quarta, di questa variabile, *riforma*, ovviamente, la variabile binaria relativa alla riforme scolastiche (assumente valore 1 per i nati dopo il 1951, 0 altrimenti). Per illustrare i risultati ottenuti, riporterò l'*output* fornito da "R"[2] (Iacus, Masarotto, 2003). I primi modelli sono di tipo binomiale. I parametri riportati dall'*output*, quindi, non stimano l'impatto delle variabili sulla probabilità, ma vanno interpretati nell'ottica di tale distribuzione. Nell' ultima colonna è riportata la significatività dei parametri osservati (vedi Appendice). Indicativamente, includerò parametri significativi al 5%. Non escluderò mai, invece, il parametro di intercetta dalla stima del modello.
Relativamente ai maschi, l'unica scelta praticabile pare quella di un modello che preveda solamente due variabili esplicative: $x$ e *riforma*. Infatti:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.225355   0.090270  13.574  < 2e-16 ***
x           0.082409   0.007165  11.501  < 2e-16 ***
riforma     0.436850   0.164810   2.651  0.00803 **
```

Come si vede, il coefficiente *riforma* risulta altamente significativo. Verifico il *fit* complessivo del modello, contrapponendolo a quello pienamente parametrizzato. Anche qui, il valore d'interesse è quello dell'ultima colonna, che valuta la significatività del modello che fa corrispondere ad ogni coorte il valore campionario rispetto al modello stimato.

```
Analysis of Deviance Table

Model 1: y ~ x + riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3732     3367.4
2      3693     3322.0   39     45.5       0.2
```

[2] vedi Appendice

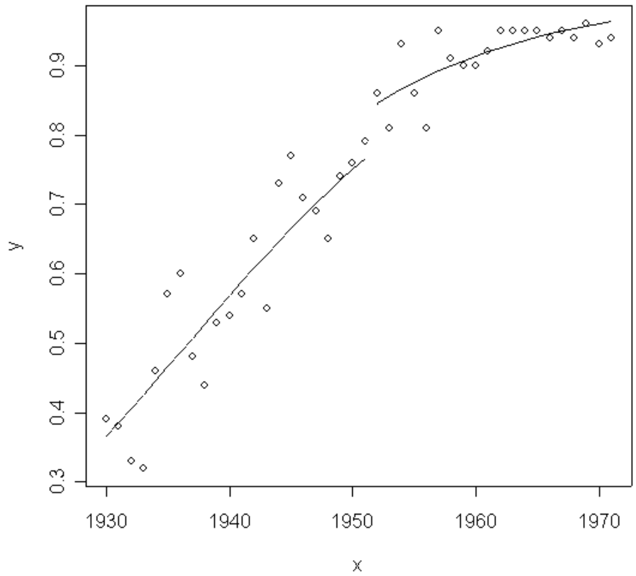


Figura 5.1: Conseguimento licenza media maschi

Il modello pare, dunque, avere un buon andamento generale. Per verificare l'entità della discontinuità (interpretandola come effetto delle riforme sulla coorte '52, o della mancata esposizione alle stesse della classe '51), calcolo i valori che riscontro in corrispondenza dell'intorno destro e di quello sinistro del valore di $x$ "1951.5". Sostituisco, in sostanza, alla formula della binomiale i parametri stimati, ottenendo: Ottengo:

```
0.8405338-0.7730046=0.0675292,
```

stimando così come impatto della riforma un aumento immediato di circa il 6.75% di maschi in possesso di licenza media (vedi grafico 5.1).
La complessità della stima relativa all'effetto della riforma sul conseguimento del diploma di licenza media per le femmine è sintetizzata dal fatto che il modello con l'adattamento migliore, come vedremo, prevede un netto calo nell'istruzione tra il '50 e il '51, cosa del tutto inverosimile. Il modello migliore che si possa ottenere (dal punto di vista del *fit* complessivo) prevede l'utilizzo di ben cinque variabili esplicative, come si può vedere.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  4.320e-01  7.485e-02   5.771 7.87e-09 ***
x^3            9.339e-04  1.368e-04   6.829 8.55e-12 ***
x^4          3.583e-05  6.713e-06   5.338 9.41e-08 ***
riforma      6.486e-01  1.398e-01   4.638 3.52e-06 ***
riforma:x^2    2.541e-02  4.045e-03   6.282 3.35e-10 ***
x^3:riforma   -2.659e-03  3.994e-04  -6.656 2.81e-11 ***
```

Di seguito, il confronto col modello pienamente parametrizzato:

```
Analysis of Deviance Table

Model 1: y ~ x^3 + x^4 + riforma + x^2:riforma + x^3:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3970     3809.0
2      3934     3770.6   36     38.3       0.4
```

Esiste, però, anche un modello accettabile che non prende in considerazione l'impatto delle riforme.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.977e-01  8.659e-02   8.057  7.8e-16 ***
x            1.225e-01  9.705e-03  12.618  < 2e-16 ***
x^2            1.941e-02  5.964e-03   3.254  0.00114 **
x^3            2.153e-03  6.938e-04   3.103  0.00192 **
x^4            6.178e-05  2.048e-05   3.017  0.00255 **
x^3:riforma   -4.324e-03  1.376e-03  -3.143  0.00167 **
```

Come si vede, il *fit* complessivo, pur accettabile, è inferiore a quello del modello precedente, e il criterio di Akaike[3] è superiore.

```
Analysis of Deviance Table

Model 1: y ~ x + x^2+ x^3 + x^4 + x^3:riforma
Model 2: y ~ factor(x)
```

[3] verrà illustrato meglio in Appendice, comunque tiene in considerazione quanto i dati siano plausibili se quello fosse il modello vero, calcolando la log-verosimiglianza, e penalizza, al contempo, la sovraparametrizzazione

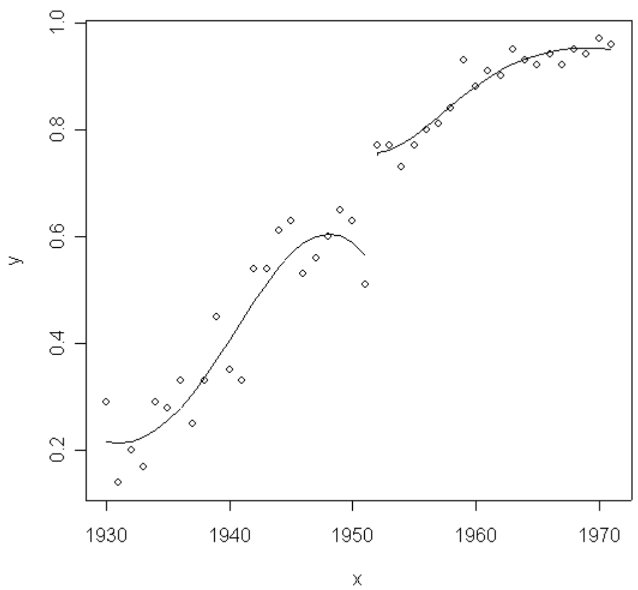


Figura 5.2: Conseguimento licenza media femmine 1

```
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3970     3815.7
2      3934     3770.6   36     45.1        0.1
```

Il primo modello prevede una discontinuità pari a:

`0.7415423- 0.5897665=0.1517758.`

Tuttavia, come si evince dal grafico (5.2), tale salto è decisamente condizionato dall'evidente calo proprio in prossimità della coorte '51. Peraltro, ristimando i dati relativi alle coorti pre-riforma senza prendere in considerazione l'ultima classe, si può osservare come un modello lineare non sia del tutto fuori luogo.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.860383   0.101456    8.48   <2e-16 ***
x           0.108951   0.008762   12.44   <2e-16 ***
Analysis of Deviance Table
```

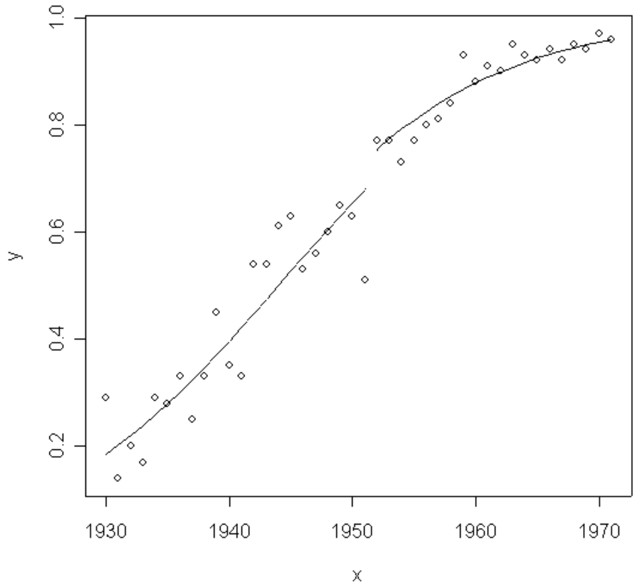


Figura 5.3: Conseguimento licenza media femmine 2

```
Model 1: y ~ x
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      1790    2280.74
2      1771    2251.48   19    29.26      0.06
```

Dato che, per il periodo successivo alla riforma, l'ipotesi di andamento lineare appare ancora più plausibile, stimo il modello complessivo senza utilizzare le potenze di $x$, arrivando ad un nuovo modello che, pur prevedendo la variabile *riforma* al limite della significatività, ha il pregio di essere estremamente parsimonioso (due soli variabili esplicative), di non considerare completamente "fisiologico" il passaggio dal 51% di donne in possesso della licenza media nate nella coorte '51 al 77 % riscontrato nelle due coorti successive e, al contempo, di non essere eccessivamente condizionato dal valore anomalo del '51 (come si evince dall'andamento del grafico 5.3).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.797402   0.082992   9.608   <2e-16 ***
```

```
x            0.105894   0.006943  15.252   <2e-16 ***
riforma      0.278630   0.147160   1.893   0.0583
AIC: 3830
Analysis of Deviance Table

Model 1: y ~ x + riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3973     3824.0
2      3934     3770.6   39     53.3       0.1
```

Tale modello, peraltro, non solo presenta un adattamento discreto ai dati, ma ha un *fit* decisamente migliore di quello che otterrei escludendo la variabile *riforma* (pur non significativa al 5 %), nonchè un criterio di Akaike inferiore.

```
AIC: 3831.6
Analysis of Deviance Table

Model 1: y ~ x
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3974     3827.6
2      3934     3770.6   40     56.9   0.04015
```

La stima dell'impatto data dal modello che, in ultima analisi, pare il più sensato, è dunque:

```
0.7457423-0.6894185=0.0563238,
```

quindi del tutto comparabile con quella relativa ai maschi.

### 5.1.2 Stima modello quadratico

Effettuo, poi, le nuove stime, basate su modelli più semplici, in quanto, come spiegato in precedenza, mi limito a $x$, $x^2$ e *riforma* come potenziali variabili esplicative. Anche qui, l'analisi relativa ai maschi si rivelerà di più facile interpretazione. Infatti, essa prevede comunque la presenza di discontinuità.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.863e+02  2.922e+01  -6.377 1.81e-10 ***
x            9.623e-02  1.509e-02   6.378 1.80e-10 ***
x^2          -4.191e-04  4.002e-04  -1.047  0.29498
```

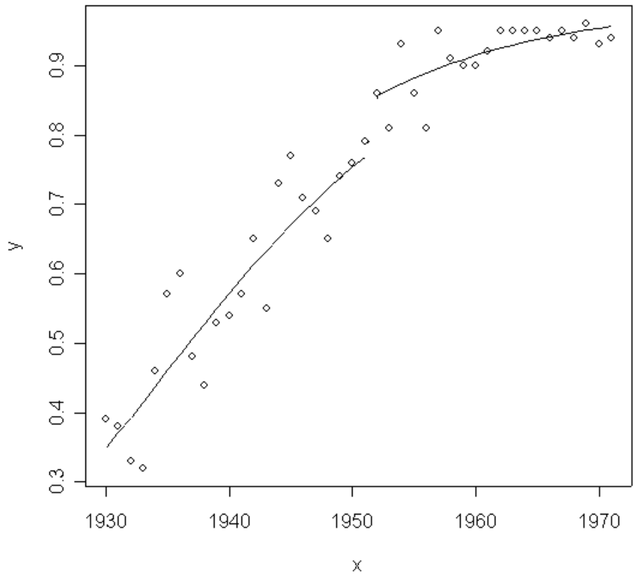


Figura 5.4: Conseguimento licenza media maschi (modello quadratico)

```
riforma      5.113e-01  1.814e-01   2.819  0.00482 **
```

Eliminando la $z$ (in quanto non significativa) ci ricondurremmo allo stesso modello scelto in precedenza (che, infatti, prevedeva solo $x$ e *riforma* come variabili esplicative), confermando, così, la presenza di discontinuità. Il modello completo, invece, prevederebbe un salto pari a:

```
0.8514167- 0.7745888=0.0768279,
```

di circa un punto percentuale superiore, dunque, a quello ottenuto tramite il modello che esclude la $x^2$ (vedi grafico 5.4). Il modello completo relativo alle donne, invece, non contempla una discontinuità significativa:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.876e+02  2.883e+01  -6.507 7.68e-11 ***
x            9.645e-02  1.488e-02   6.482 9.05e-11 ***
x^2          2.675e-04  3.751e-04   0.713    0.476
riforma      2.450e-01  1.540e-01   1.591    0.112
```

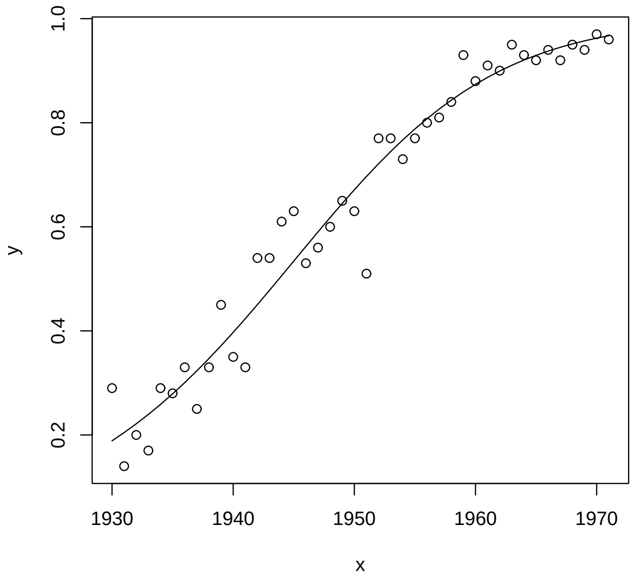


Figura 5.5: Conseguimento licenza media femmine (modello quadratico)

E' di conforto, però, osservare che la non significatività di $x^2$ permarrebbe anche in assenza della riforma (vedi grafico 5.5).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.919e+02  2.902e+01  -6.610 3.84e-11 ***
x            9.865e-02  1.498e-02   6.585 4.56e-11 ***
x^2            4.516e-04  3.628e-04   1.245    0.213
```

Anche qui, si osservi quindi come, eliminando la $x^2$, ci si ricondurrebbe al modello scelto in precedenza.

### 5.1.3 Controllo basato sulla previsione

Effettuiamo, poi, il secondo controllo, quello che, stimando un modello sulle coorti fino al '48 (quelle, cioè, che si presume non siano toccate dalla riforma) servendosi delle esplicative $x$ e $x^2$, determina, come spiegato in precedenza, delle bande di confidenza. Anche qui, si può osservare (grafici 5.6 e 5.7 come valori al di fuori di tali bande siano riscontrabili quasi solo relativamente ai maschi) Il fatto può essere spiegato dalla variabilità campionaria relativa alle donne, che fa sì che, nel modello stimato, prevalga la componente lineare (mentre, negli uomini, quella quadratica determina previsioni meno ottimistiche). Le indicazioni ottenute, complessivamente, fanno propendere per un effetto delle riforme sul conseguimento della licenza media sia per gli uomini che per le donne (che stimo intorno ai 7 punti per i maschi, intorno ai 6 per le femmine, almeno relativamente alla classe '52) anche se l'andamento campionario più incoerente relativo a queste ultime determina maggiori difficoltà interpretative. Questo può essere spiegato dal fatto che le donne nate prima degli anni '60 (le cui coorti, sia maschili che femminili, si stabilizzano su valori costantemente superiori al 90%) hanno tassi di conseguimento della licenza media costantemente più bassi di quelli degli uomini, cosa che rende più vicini tali valori al 50% (almeno a partire dagli anni '40), aumentando, di fatto, la variabilità campionaria (che, in una *binomiale*, è massima per il valore di probabilità "0.5"). Negli anni '30, invece, ho costantemente a disposizione meno di 100 osservazioni sia per gli uomini che per le donne, quindi non c'è nessuna sorpresa nel riscontrare talvolta forti differenze campionarie tra anni consecutivi.

## 5.2 Scuola media superiore

Anche qui, mi aspetto che la liberalizzazione degli accessi universitari possa incidere poco sul raggiungimento di un titolo post-media [4]. Verosimilmente, infatti, saranno in pochi a essere spronati al raggiungimento del diploma quinquennale (nel caso di istituti tecnici o professionali, o di corsi di breve durata) semplicemente dalla consapevolezza di potere poi eventualmente proseguire gli studi all'università. Analogamente a quanto detto per le scuole medie, poi, l'iscrizione a un istituto superiore, per le coorti '51 e '52, in assenza di ritardi acquisiti durante la scuola dell'obbligo, avviene, rispettivamente, nel '65 e nel '66, quindi prima della liberalizzazione degli accessi all'università.
Relativamente all'estensione dell'obbligo, d'altra parte, m'aspetto allo stesso modo che sia difficile che individui che raggiungono la licenza media perchè costretti a farlo dalla legge siano poi decisi a proseguire gli studi fino al conseguimento di un titolo di scuola superiore (anche se la cosa non è da escludere, soprattutto in relazione ai diplomi e agli attestati di qualifica).
Spiegazioni più plausibili di un eventuale impatto delle riforme su un aumento dei diplomati appaiono quindi altre. Il fatto che un titolo di scuola media

[4] sia qui che in seguito, parlerò di "scuola superiore" con questa accezione, includendo anche diplomi e corsi di qualifica professionale

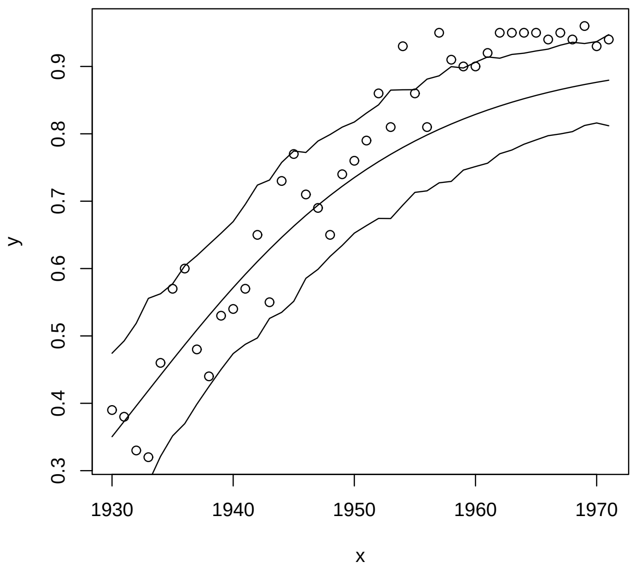


Figura 5.6: Conseguimento licenza media maschi (modello previsivo)

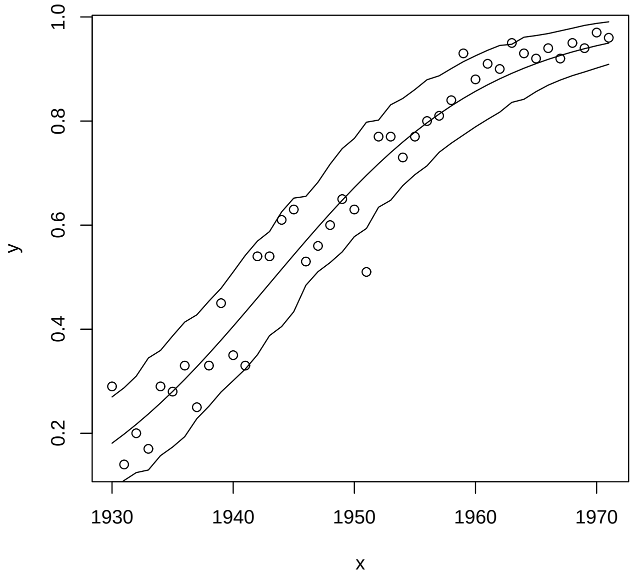


Figura 5.7: Conseguimento licenza media femmine (modello previsivo)

inferiore cominci a perdere di valore come "segnale" nei confronti del mercato del lavoro potrebbe essere una di queste: in sostanza, diventa necessario, dopo la riforma, almeno conseguire un titolo di studi di scuola superiore, data la saturazione dei livelli scolastici inferiori, qualora si voglia acquisire un vantaggio competitivo iniziale (in termini di probabilità di assunzione, o di ottenimento di incarichi più remunerativi) sugli individui meno istruiti. Un altro elemento da prendere in considerazione è l'istituzione della "scuola media unificata": un gruppo di individui che avrebbero scelto la scuola di avviamento professionale, e quindi non avrebbero avuto possibilità di proseguire gli studi conseguito il titolo, hanno, invece, l'opportunità di scegliere, dopo la scuola media (e senza avere alcun "handicap" relativamente all'istruzione nei confronti di nessuno) se iscriversi o meno ad una scuola superiore.

### 5.2.1 Modello stimato

La variabile che prendo in esame, dunque, è ancora di tipo binomiale; scelgo quella relativa al conseguimento di un qualsiasi tipo di diploma di scuola superiore, indipendentemente dal fatto che si tratti di un diploma di maturità (corsi quinquennali), di un diploma quadriennale, oppure di un diploma o attestato di qualifica (corsi di durata biennale o triennale). Tutto quanto detto in precedenza a proposito di variabili esplicative e modelli, dunque, continua a valere: l'unica differenza è nella variabile risposta. Anche in questo caso, la coorte del '51 presenta un valore anomalo nelle donne (0.33, contro lo 0.44 dell'anno precedente), cosa che, nuovamente, rende più problematica l'analisi del risultato relativo alle femmine.
Per quanto riguarda gli uomini, invece, risulta arduo stimare un modello che consideri la riforma come significativa. Il modello scelto presenta 5 variabili esplicative e ha un ottimo adattamento ai dati, tuttavia risente eccessivamente del valore anomalo del '71, che è estremamente elevato (vedi grafico 5.8).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.425e-01  6.110e-02  -2.332 0.019681 *
x^2          -1.948e-02  4.073e-03  -4.783 1.73e-06 ***
x^3           -1.750e-03  5.088e-04  -3.439 0.000583 ***
x^4        -4.834e-05  1.607e-05  -3.009 0.002620 **
x^2:riforma    3.461e-02  7.717e-03   4.484 7.31e-06 ***
x^4:riforma    1.040e-04  3.280e-05   3.172 0.001516 **

Analysis of Deviance Table
```

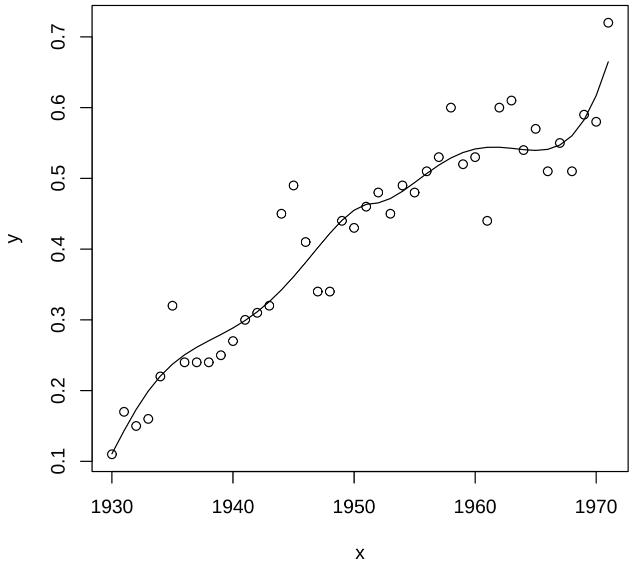


Figura 5.8: Conseguimento diploma superiore maschi

```
Model 1: y ~ x^2 + x^3 + x^4 + x^2:riforma + x^4:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3729     4828.9
2      3693     4797.5   36     31.5       0.7
```

Relativamente alle femmine, l'unico modello con un *fit* discreto prevede la presenza di 5 variabili esplicative, tra cui la riforma.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.916e-01  7.882e-02  -7.507 6.07e-14 ***
x^3            8.340e-04  1.662e-04   5.019 5.19e-07 ***
x^4           3.054e-05  8.333e-06   3.665 0.000247 ***
riforma      3.375e-01  1.276e-01   2.644 0.008181 **
riforma:x^2    1.116e-02  3.103e-03   3.598 0.000321 ***
x^3:riforma   -1.859e-03  4.471e-04  -4.158 3.21e-05 ***
AIC: 4792.1
Analysis of Deviance Table
```

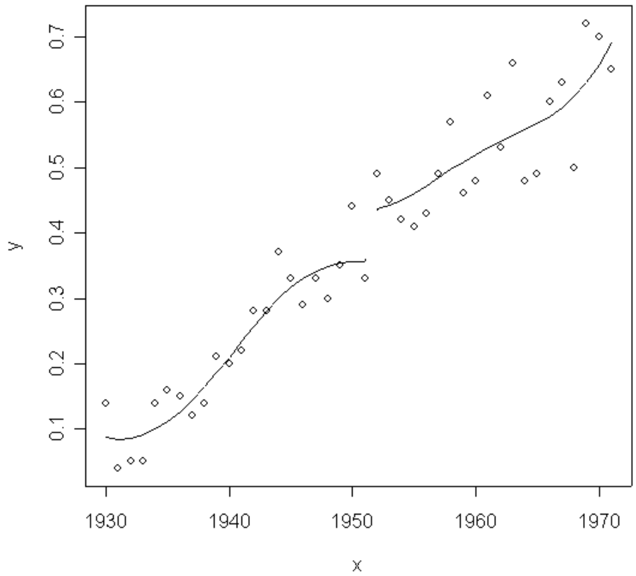


Figura 5.9: Conseguimento diploma superiore femmine 1

```
Model 1: y ~ x^3 + x^4 + riforma + x^2:riforma + x^3:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3970     4780.1
2      3934     4729.9   36     50.3       0.1
```

Tuttavia (vedi grafico 5.9) un modello di tale tipo prevederebbe una discontinuità di:

0.4368146-0.3562678=0.0805468.

Pensare che, nel caso delle femmine, la riforma della scuola media abbia un effetto superiore sul livello scolastico successivo che sul conseguimento della stessa licenza media, è quantomeno improbabile. Questo modello, di fatto, appare molto simile a quello che prevedeva una discontinuità di 15 punti percentuali relativamente al raggiungimento, sempre da parte delle donne, del titolo di scuola media inferiore.
Il modello migliore ottenibile senza considerare l'impatto della riforma ha un adattamento non straordinario, presenta ovviamente un criterio di Akaike superiore al modello appena esposto, e prevede, per giunta, la presenza di

un parametro significativo solo al 7.6%. Tuttavia, presenta dei pregi: è relativamente parsimonioso (presentando 3 variabili esplicative), nonchè stabile e plausibile, dato che prevede, come variabili esplicative, solo la coorte di nascita e le sue potenze seconda e terza.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.818e-01  5.014e-02  -7.614 2.65e-14 ***
x            5.939e-02  7.406e-03   8.019 1.06e-15 ***
x^2         -1.233e-03  3.111e-04  -3.963 7.40e-05 ***
x^3          5.059e-05  2.851e-05   1.775    0.076 .
 AIC: 4793.1
Analysis of Deviance Table

Model 1: y ~ x + x^2 + x^3
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3972     4785.1
2      3934     4729.9   38     55.2   0.03531
```

Volendo escludere dal modello $x^3$, pur ottenendo un modello ancora più semplice e parsimonioso , peggiorerei l'adattamento complessivo del modello (e il modello pienamente parametrizzato passerebbe da una significatività del 3.5%, già piuttosto alta, ad una inferiore al 2.4%) e otterrei un criterio di Akaike superiore.

```
AIC: 4794.3
 Analysis of Deviance Table

Model 1: y ~ x + x^2
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3973     4788.3
2      3934     4729.9   39     58.4   0.02368
```

Il modello preferibile, pare, quindi, quello che prevede, come regressori, $x$,$x^3$ e $x^2$: analogamente a quanto detto per il modello scelto a proposito della scuola media, esso non risente in maniera eccessiva del dato anomalo del '51 (vedi grafico 5.10).

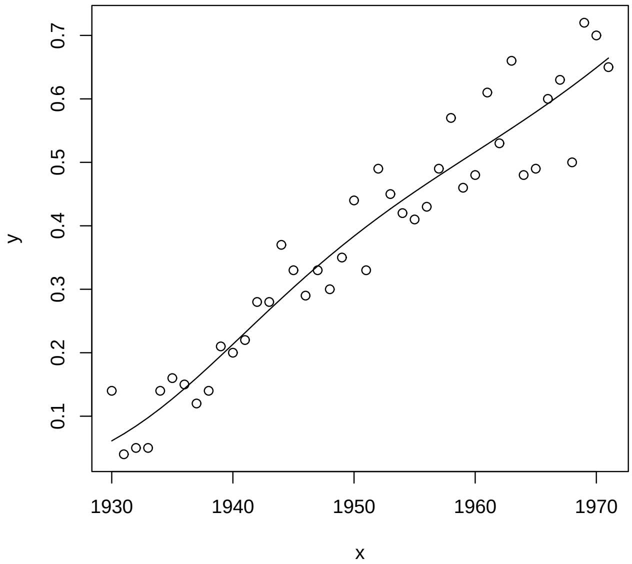


Figura 5.10: Conseguimento diploma superiore femmine 2

### 5.2.2 Stima modello quadratico

Passando alla stima del modello quadratico, stavolta, non ci sono ambiguità. Anche per le donne, infatti, il risultato conferma immediatamente la scelta di un modello che non preveda l'impatto delle riforme: per entrambi i sessi, infatti, la variabile *riforma* si rivela non significativa, se inclusa con $x$ e $x^2$ tra i regressori. Per le donne, infatti, il modello si presenta così:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.471e+02  3.180e+01  -7.769 7.91e-15 ***
x            1.267e-01  1.640e-02   7.727 1.10e-14 ***
x^2           -1.182e-03  3.040e-04  -3.887 0.000101 ***
riforma     -5.948e-02  1.344e-01  -0.443 0.658072
```

Non ci sono dubbi: $x$ e $x^2$ sono estremamenti significativi, mentre *riforma* non lo è per nulla (il coefficiente stimato relativo, peraltro, è addirittura negativo, quindi opto per la sua esclusione). Il grafico relativo, il (5.11), è espressione di un modello visto prima, quello che prevedeva, come regressori, solo $x$ e $x^2$, per l'appunto. Esso, peraltro, è simile al modello scelto in precedenza

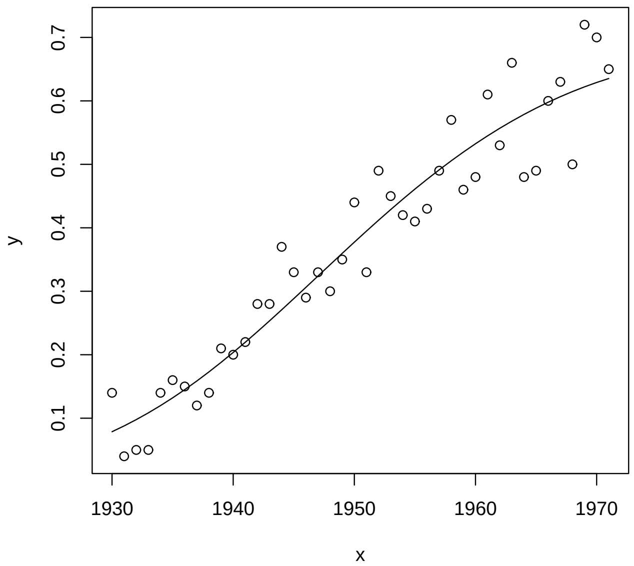


Figura 5.11: Conseguimento diploma superiore femmine (modello quadratico)

(differendo solo per l'esclusione di $x^3$, variabile peraltro non significativa al 5%, tra i regressori) e, dal confronto col modello pienamente parametrizzato (significativo al 2.37%) si rivela essere non lontanissimo dai dati.
Per gli uomini, le considerazioni da fare sono praticamente identiche.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.898e+02  2.833e+01  -6.698 2.11e-11 ***
x            9.744e-02  1.462e-02   6.667 2.61e-11 ***
x^2          -1.080e-03  2.846e-04  -3.794 0.000148 ***
riforma     -1.164e-02  1.321e-01  -0.088 0.929811
```

Il modello cui si riferisce il grafico (5.12) è quindi questo:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.889e+02  2.628e+01  -7.187 6.65e-13 ***
x            9.696e-02  1.356e-02   7.153 8.48e-13 ***
x^2          -1.079e-03  2.844e-04  -3.794 0.000148 ***
```

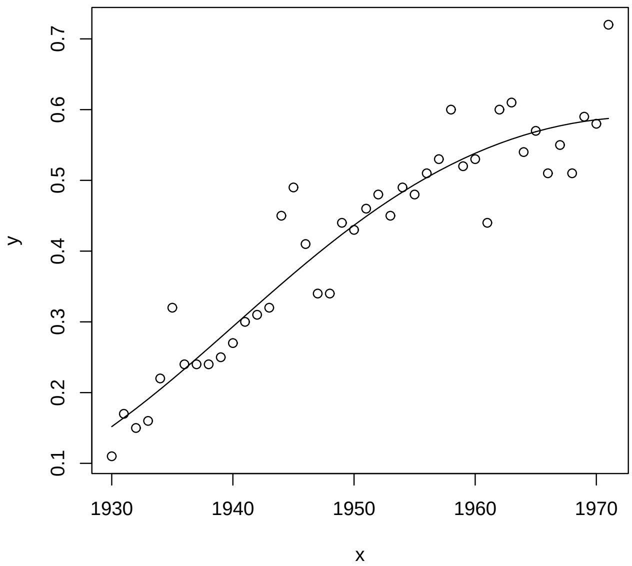


Figura 5.12: Conseguimento diploma superiore maschi (modello quadratico)

### 5.2.3 Controllo basato sulla previsione

Si può osservare subito come la stima di un modello quadratico relativamente ai dati fino al '48 porti ad andamenti decrescenti già a partire dagli anni '50 per entrambi i sessi (vedi grafici 5.14 e 5.13). Per quanto riguarda la stima di un modello lineare, si osservi come essa porti a valori addirittura inferiori a quelli previsti quasi immediatamente per le femmine (vedi 5.15), a partire dagli anni '60 per i maschi (vedi 5.15). Come spiegato nell'Introduzione, questo non è per nulla sorprendente: utilizzare un modello lineare stimato sulle coorti pre-riforma significa supporre che l'andamento crescente si mantenga costante, cosa assolutamente controintuitiva (ancora di più per le donne, che hanno avuto tassi d'incremento più alti per quelle classi). In definitiva, non c'è alcuna evidenza che la riforma abbia un effetto sul conseguimento del titolo di scuola superiore, e l'unico elemento che crea qualche difficoltà nella modellazione pare essere il forte valore anomalo riscontrato per la coorte femminile del '51.

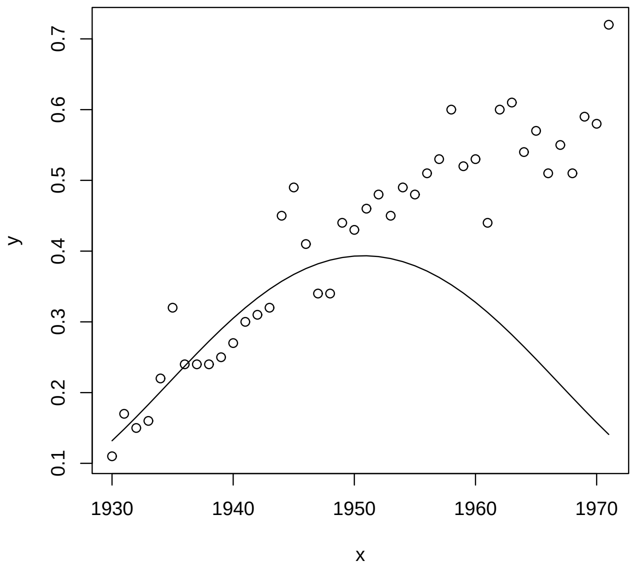


Figura 5.13: Conseguimento diploma superiore maschi (modello previsivo 1)

## 5.3 Istruzione terziaria

Valuto ora l'eventuale impatto delle riforme scolastiche sul conseguimento di titoli post-secondari (lauree, diplomi universitari o di Accademia, "Isef", "Scuola per assistenti sociali")[5], che considero nel loro complesso (in più dell'80% dei casi, comunque, si tratta di lauree). Relativamente alla prima riforma, se era da ritenersi improbabile che uno studente che raggiunge la licenza media perchè costretto a farlo dalla legge prosegua poi gli studi fino al diploma, a maggior ragione lo sarà in relazione alla laurea. Per quanto riguarda l'eventuale spostamento in avanti del titolo necessario per l'autosegnalazione al mercato del lavoro, è palese che m'aspetterei un impatto prima di tutto sui diplomi, e di conseguenza (quindi, indirettamente) sulle lauree. Circa la possibilità che uno studente che, in assenza della riforma, si sarebbe iscritto alla scuola d'avviamento professionale, prosegua poi gli studi fino alla laurea, anch'essa è evidentemente condizionata al fatto che questi, prima, raggiunga il diploma (si tratta, di nuovo, di un'eventualità più remota). Se a ciò, poi, aggiungiamo che non è stato riscontrata un'influenza significativa

[5] d'ora in poi, quando utilizzerò il termine "laurea", mi riferirò indistintamente a questi titoli

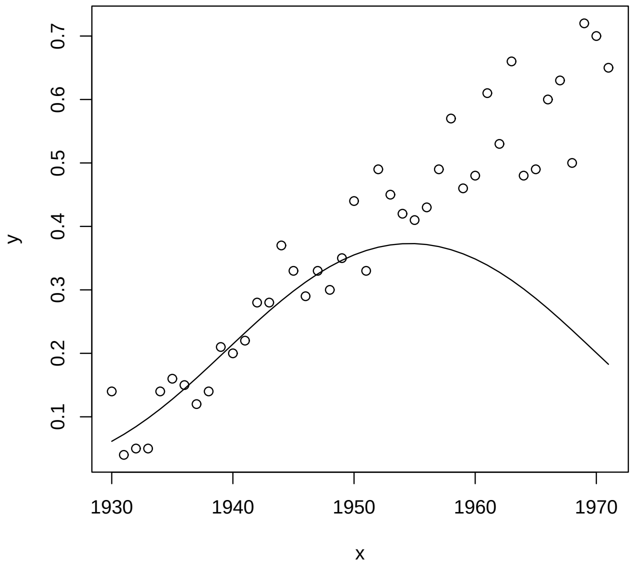


Figura 5.14: Conseguimento diploma superiore femmine (modello previsivo 1)

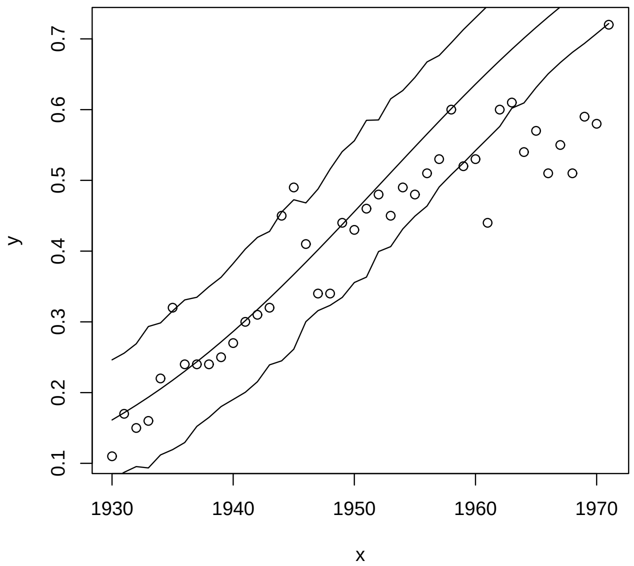


Figura 5.15: Conseguimento diploma superiore maschi (modello previsivo 2)

delle riforme sulla percentuale dei diplomati, risulta palese come la riforma delle scuole medie abbia difficilmente condizionato la probabilità di laurearsi degli individui coinvolti (come peraltro appare intuitivo).
E' di interesse, evidentemente, valutare invece se possa esserci un impatto della riforma universitaria sulla percentuale di laureati complessiva (come parrebbe auspicabile). Effettuerò le stesse operazioni svolte per i livelli scolastici inferiori, per poi commentare i dati anche alla base di queste previsioni.

### 5.3.1 Modello stimato

Anche per quanto riguarda l'istruzione terziaria, i modelli scelti non prevedono una discontinuità tra il '51 e il '52. Stavolta, tuttavia, l'analisi risulta più agevole per le donne. L'elemento che colpisce nei risultati femminili, stavolta, non è più il risultato anomalo del '51, bensì i valori elevati riscontrati nelle classi tra il '49 e il '54 (ne parlerò più diffusamente in seguito). Il modello che prevederebbe un impatto della riforma, comunque, è il seguente.

```
 Call:
glm(formula = y ~ x + riforma + x:riforma, family = binomial)
```

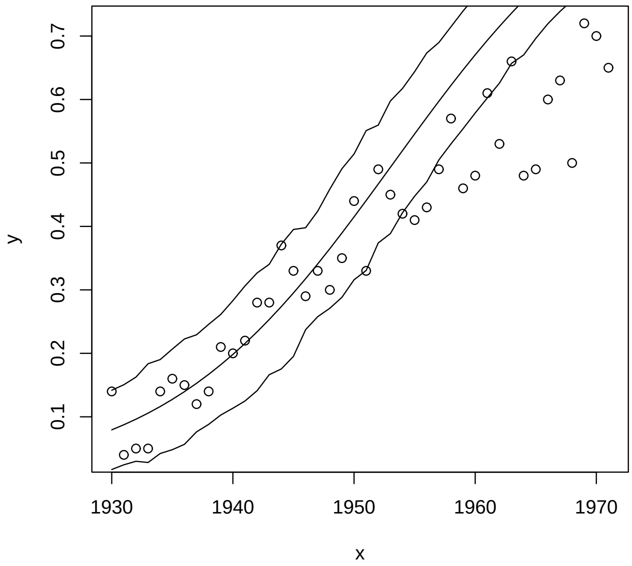


Figura 5.16: Conseguimento diploma superiore femmine (modello previsivo 2)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6393  -0.5378  -0.4669  -0.2850   2.8379

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.79038    0.15811 -11.323  < 2e-16 ***
x            0.10318    0.01769   5.832 5.49e-09 ***
riforma     -0.42748    0.21032  -2.032  0.04210 *
x:riforma   -0.06556    0.02087  -3.141  0.00169 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
AIC: 2554.7
```

Esso, peraltro, presenta un ottimo adattamento, e non fa uso delle potenze di $x$: ci sono solo tre variabili esplicative, e, sia prima che dopo la riforma, è previsto un andamento lineare.

```
Analysis of Deviance Table

Model 1: y ~ x + riforma + x:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3972    2546.68
2      3934    2519.26   38    27.42      0.90
```

Tuttavia, la variabile *riforma* è significativa solo al 4.21% e, soprattutto, il coefficiente associato al parametro è negativo. La cosa, da un punto di vista teorico, è estremamente implausibile, in quanto indicherebbe un effetto negativo delle riforme sulla possibilità di pervenire alla laurea. Peraltro, è possibile, sostituendo $x^3$ a *riforma*, ottenere un modello che ponga meno dubbi sulla significatività delle variabili ($x^3$ è significativo all'1.5%) e che abbia un valore AIC più basso a parità di parsimonia (utilizzo sempre 3 parametri), quindi un migliore adattamento ai dati.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.9802079  0.1070918 -18.491  < 2e-16 ***
x            0.0630354  0.0158803   3.969  7.2e-05 ***
x^3            0.0001083  0.0000445   2.433  0.01496 *
x:riforma   -0.0697849  0.0216279  -3.227  0.00125 **
```

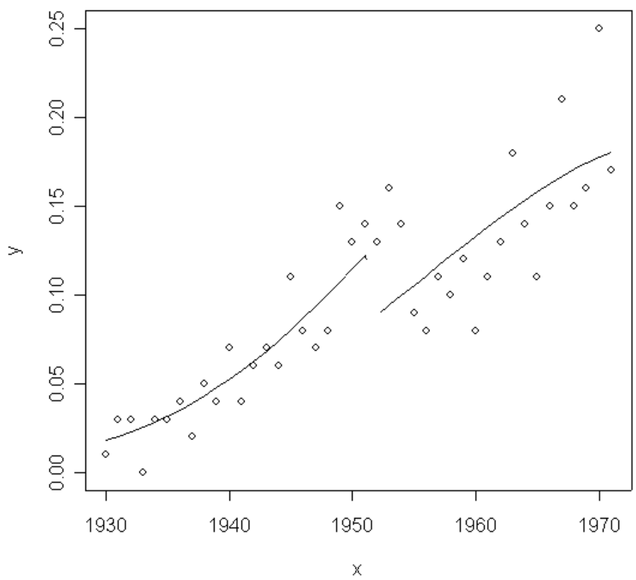


Figura 5.17: Conseguimento laurea femmine

```
AIC: 2552.7
Analysis of Deviance Table

Model 1: y ~ x + x^3 + x:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3972    2544.71
2      3934    2519.26   38    25.45      0.94
```

Ora, se il modello che prevede un salto negativo può essere immediatamente scartato, il grafico relativo all'ultimo modello (il 5.17) pone in evidenza, comunque, come i valori campionari dal'49 al '54 siano costantemente superiori, e quelli del quinquennio successivo invece al di sotto, rispetto alla stima dei valori della popolazione. Questo è evidentemente dovuto, come preannunciato, al fatto che, dal '49 al '54, si siano riscontrati tassi di conseguimento della laurea nettamente più alti che negli anni immediatamente precedenti e successivi. Cercherò di analizzare meglio la questione successivamente.
Relativamente agli uomini, invece, noto come il periodo successivo alla riforma sia caratterizzato da una sostanziale mancanza di trend.

```
anova(mod.A,mod.B,test="Chisq")
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      1958    1457.10
2      1939    1443.21   19    13.89      0.79
```

Valuto, quindi, la possibilità di costruire un modello complessivo che tenga conto di questo fatto, utilizzando la variabile: *prima* al posto di *riforma*, ossia la sua variabile complementare (vale 1 quando *riforma* vale 0, e viceversa). Questo, infatti, consente di imporre il vincolo di andamento costante dal '52 in poi, senza imporre alcuna restrizione sul periodo successivo. E' possibile ottenere un modello di questo tipo che preveda una significatività di *prima* al 6.9%, con un buon adattamento ai dati.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.969e+00  6.891e-02 -28.572  < 2e-16 ***
prima       -2.291e-01  1.260e-01  -1.817  0.06915 .
prima:x^3      1.104e-04  3.807e-05   2.900  0.00373 **
AIC: 2467.4
Analysis of Deviance Table

Model 1: y ~ prima + x^3:prima
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3732    2461.43
2      3693    2418.71   39    42.72      0.31
```

Escludendo da tale modello l'impatto della riforma, ottengo invece:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.041e+00  5.787e-02 -35.270  < 2e-16 ***
x^3:prima      1.438e-04  3.443e-05   4.176 2.97e-05 ***
AIC: 2468.8
```

Questo modello presenta un AIC più basso, e, al tempo stesso, un peggiore adattamento ai dati.

```
 Analysis of Deviance Table
```

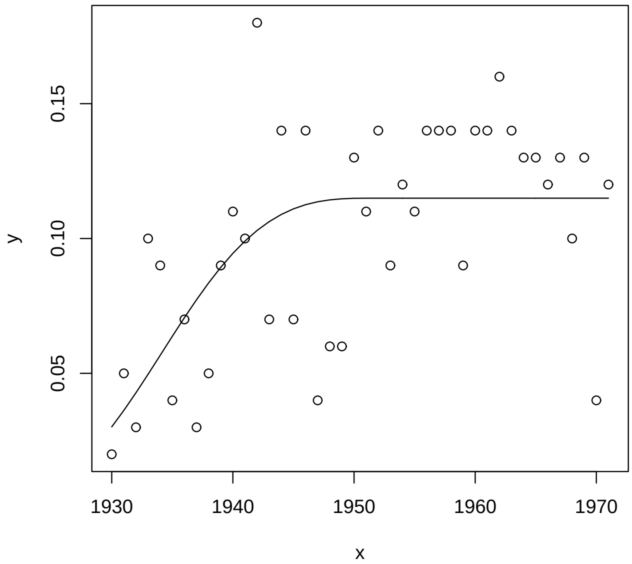


Figura 5.18: Conseguimento laurea maschi

```
Model 1: y ~ x^3:prima
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3733    2464.80
2      3693    2418.71   40    46.09      0.23
```

Tuttavia, sarebbe possibile utilizzare un modello meno parsimonioso, utilizzando anche $x$ e $x^3$. Così facendo, però, verrebbe meno l'andamento costante del periodo post-riforme (vedi grafico 5.18), perdendo quindi di senso l'utilizzo della variabile *prima*; peraltro, pur migliorando il valore dell' AIC ed essendo l'adattamento complessivo ai dati il migliore in assoluto, i due parametri aggiuntivi risultano significativi, considerati insieme, solo al 9%.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.037e+00  6.995e-02 -29.117  < 2e-16 ***
x            2.463e-02  1.153e-02   2.137  0.03261 *
x^3          -9.836e-05  4.905e-05  -2.005  0.04494 *
x^3:prima     1.607e-04  4.916e-05   3.268  0.00108 **
AIC: 2468
```

```
Analysis of Deviance Table

Model 1: y ~ x + x^3 + x^3:prima
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3731    2459.99
2      3693    2418.71   38    41.28      0.33

Analysis of Deviance Table

Model 1: y ~ x^3:prima
Model 2: y ~ x + x^3 + x^3:prima
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3733    2464.80
2      3731    2459.99    2     4.81      0.09
```

Tramite i controlli, cercherò di chiarire meglio questi punti.

### 5.3.2 Stima modello quadratico

Limitandomi all'uso di $x$, $x^2$ e *riforma* come variabili esplicative, ottengo un'indicazione di non significatività della riforma relativamente agli uomini, mentre resta aperta la questione dell'andamento delle donne. Infatti, il modello per gli uomini è del tipo:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.471e+02  4.709e+01  -3.125  0.00178 **
x            7.467e-02  2.429e-02   3.075  0.00211 **
x^2          -1.379e-03  4.617e-04  -2.986  0.00283 **
riforma      2.062e-01  2.069e-01   0.997  0.31890
```

Essendo la variabile *riforma* decisamente non significativa, ristimo il modello servendomi di due sole variabili esplicative:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.663e+02  4.333e+01  -3.839 0.000124 ***
x            8.457e-02  2.235e-02   3.785 0.000154 ***
x^2         -1.421e-03  4.622e-04  -3.075 0.002104 **
```

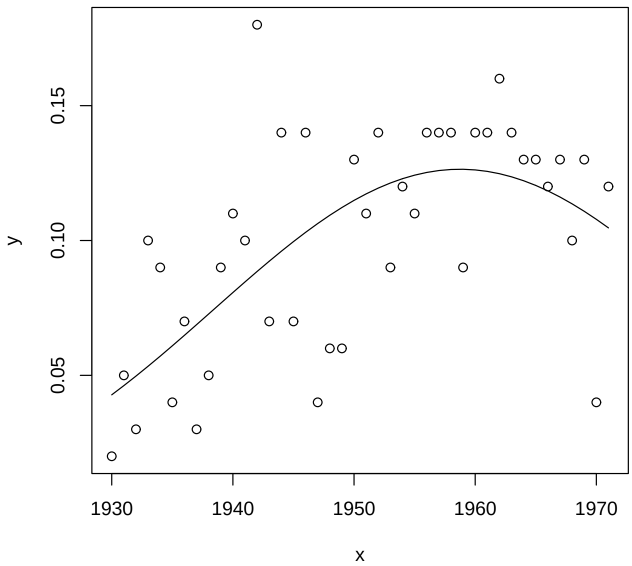


Figura 5.19: Conseguimento laurea maschi (modello quadratico)

Il fatto che il modello assuma un andamento decrescente negli ultimi anni (vedi grafico 5.19), testimonia come non ci sia, di fatto, un trend crescente negli anni successivi alla riforma, dunque va a sostegno della stima di un andamento costante.
Relativamente alle donne, invece, si conferma la significatività del parametro relativo a *riforma*, così come la sua negatività.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.484e+02  5.620e+01  -4.420 9.87e-06 ***
x            1.266e-01  2.897e-02   4.371 1.24e-05 ***
x^2           -1.289e-03  5.031e-04  -2.562   0.0104 *
riforma     -4.270e-01  2.165e-01  -1.973   0.0485 *
AIC: 2558
```

Tuttavia, tale significatività, ancora una volta, è tutt'altro che impressionante (4.85%), pur essendo l'AIC più basso del modello che esclude *riforma* dalle esplicative.

```
Coefficients:
```

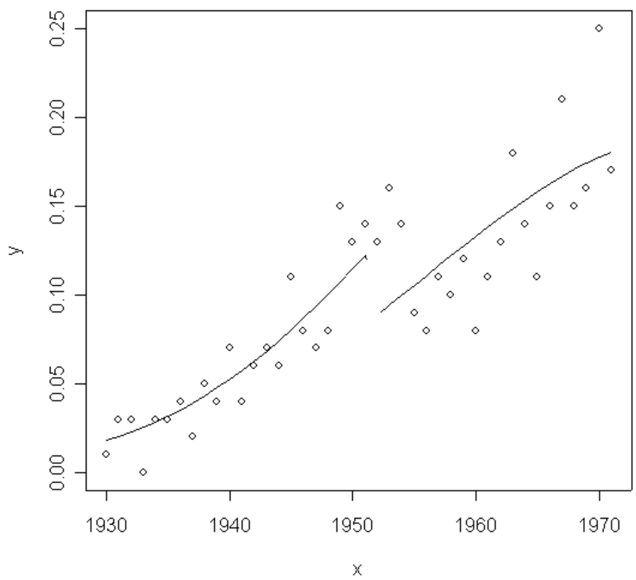


Figura 5.20: Conseguimento laurea femmine (modello quadratico 1)

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.938e+02  4.696e+01  -4.128 3.66e-05 ***
x            9.848e-02  2.420e-02   4.069 4.73e-05 ***
x^2          -1.046e-03  4.737e-04  -2.207   0.0273 *
AIC: 2559.9
```

Entrambi i grafici, comunque (5.20 e 5.21), mettono di nuovo in risalto l'anomalia dei valori tra il '49 e il '54. Effettuo ora l'ultimo controllo, prima di trarre le conclusioni.

### 5.3.3 Controllo basato sulla previsione

Un modello basato sulla previsione, dato l'andamento costante del periodo post- riforma, non ha molto senso, se effettuato sugli uomini. Ancora una volta, la previsione basata su un *trend* quadratico conduce immediatamente a valori decrescenti (vedi grafico 5.22). Usando il modello lineare, tuttavia, il fatto che i valori comincino a rivelarsi inferiori al limite della significatività, rispetto a quelli previsti, solo per le ultime coorti esaminate (vedi grafico 5.23) dimostra come il trend di crescita si fosse interrotto, di fatto, già prima delle coorti post-riforma (si può osservare, infatti, come, dopo la crescita

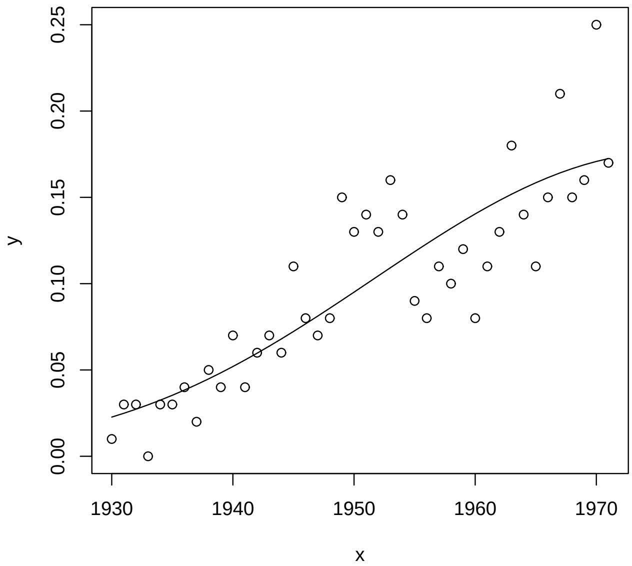


Figura 5.21: Conseguimento laurea femmine (modello quadratico 2)

degli anni '30, già le coorti del decennio successivo comincino a stabilizzarsi, fatta salva una forte variabilità campionaria).
Per quanto riguarda le donne, invece, la $x^2$ si verifica così poco significativa da non incidere sul trend crescente (vedi grafico 5.24). Ancora una volta, spiccano i valori tra il '49 e il '54, gli unici in linea con le previsioni stesse (mentre gli altri sono al di sotto delle bande di confidenza, a testimonianza di quanto la crescita del periodo pre-riforma sia stata molto superiore a quella successiva). Questo, evidentemente, dimostra, una volta di più, che è necessario ricercare una spiegazione agli elevati risultati di queste coorti.

### 5.3.4 Analisi risultati femminili '49/'54

La prima possibile motivazione dei risultati riscontrati nelle coorti femminili è che i valori più alti di raggiungimento della laurea tra le classi '49/'54 siano dovuti al caso. Per verificare questa eventualità, restringo il campione a queste classi e a quelle immediatamenente precedenti e successive (quindi, dal '46 al '57), inserendo come unica variabile esplicativa una indicatrice, che assume valore 1 per le classi '49/'54, 0 per le altre. Chiamando $x$ tale variabile, essa risulta altamente significativa.

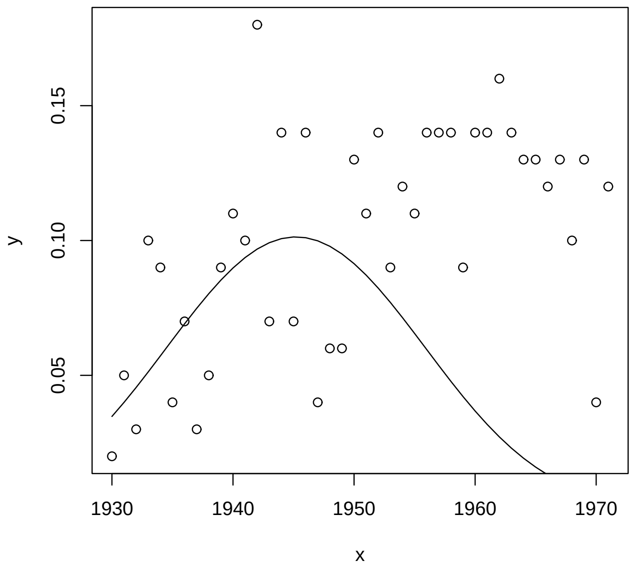


Figura 5.22: Conseguimento laurea maschi (modello previsivo 1)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3863     0.1409  -16.93  < 2e-16 ***
x             0.5629     0.1864    3.02  0.00253 **
```

La differenza tra le medie dei due gruppi considerati, infatti, è di

```
0.1390264 - 0.08422337 = 0.05480303,
```

circa il 5.5%, veramente un'enormità, dato che nelle coorti vicine il valore è di circa 8.4%.
Dovendo ritenere l'incremento delle lauree conseguite in quel periodo non come effetto del caso, occorre trovare una spiegazione alternativa del fenomeno. Prima di farlo, però, pare opportuno ricordare che, tra il possesso di un qualsiasi titolo di scuola superiore e quella di uno di scuola terziaria, esistono due passaggi intermedi, ovvero: il possesso di un diploma di maturità e l'iscrizione all'università, subordinata, appunto, al conseguimento di tale diploma.

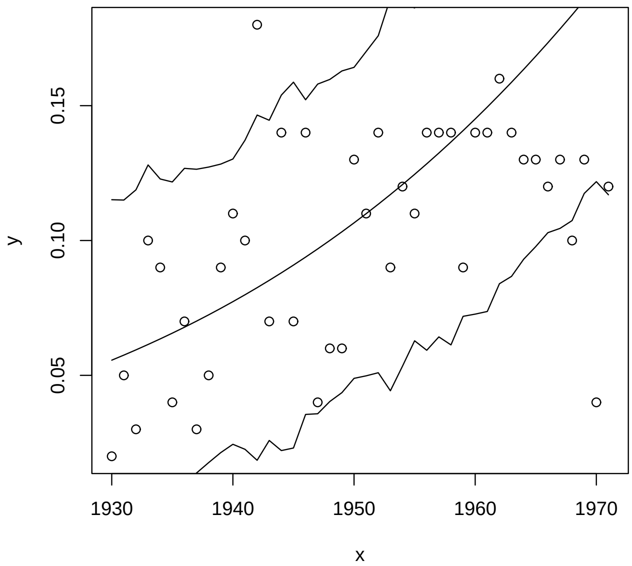


Figura 5.23: Conseguimento laurea maschi (modello previsivo 2)

### Il diploma di maturità

Il grafico riguardante il conseguimento del titolo di maturità pare evidenziare una discontinuità proprio a partire dal '49. Stimo il modello relativo, con le modalità precedenti, prevedendo l'eventuale salto in quel punto. E' da notare che, se per i livelli d'istruzione più bassi il problema riguardo ai risultati femminili era relativo al valore anomalo del '51, stavolta i valori estremamente bassi vicino alla soglia sono dati dagli scarsi conseguimenti di due coorti, la '47 e la '48. Ricercando un modello che, al contempo, preveda la discontinuità, non risenta eccessivamente di questi due valori, abbia un *fit* discreto e sia relativamente parsimonioso, ottengo:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.337e+00  9.367e-02 -14.276  < 2e-16 ***
x^2          -5.119e-03  8.462e-04  -6.050 1.45e-09 ***
riforma      5.096e-01  1.615e-01   3.155 0.001605 **
riforma:x    6.473e-02  1.854e-02   3.491 0.000482 ***
riforma:x^3    2.201e-04  3.898e-05   5.646 1.64e-08 ***
AIC: 4449.4
```

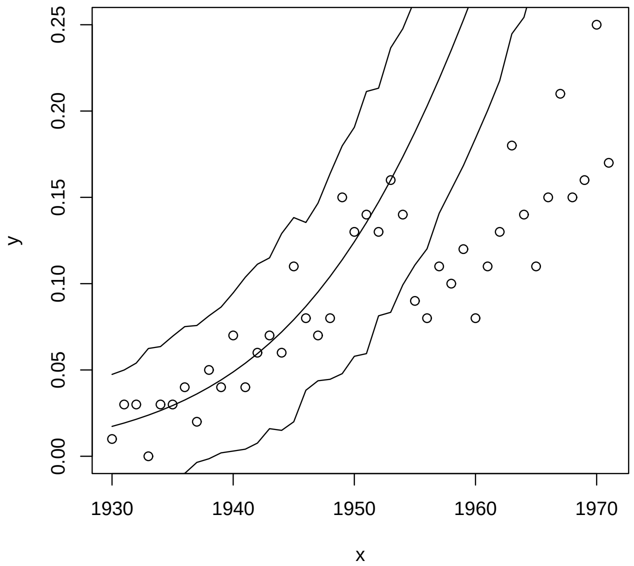


Figura 5.24: Conseguimento laurea femmine (modello previsivo )

```
Analysis of Deviance Table

Model 1: y ~ x^2 + riforma + x:riforma + x^3:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3971     4439.4
2      3934     4392.8   37     46.6       0.1
```

Tale modello prevederebbe un salto comunque notevole in coincidenza della discontinuità, come evidenziato dal grafico (5.25).

```
 0.3041951- 0.2080038= 0.0961913
```

L'unico modello accettabile che, invece, non preveda la discontinuità, è:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.1318118  0.0708038 -15.985  < 2e-16 ***
x^2          -0.0145693  0.0025890  -5.627 1.83e-08 ***
x^3         -0.0005551  0.0001646  -3.372 0.000746 ***
x:riforma    0.1698322  0.0205242   8.275  < 2e-16 ***
```

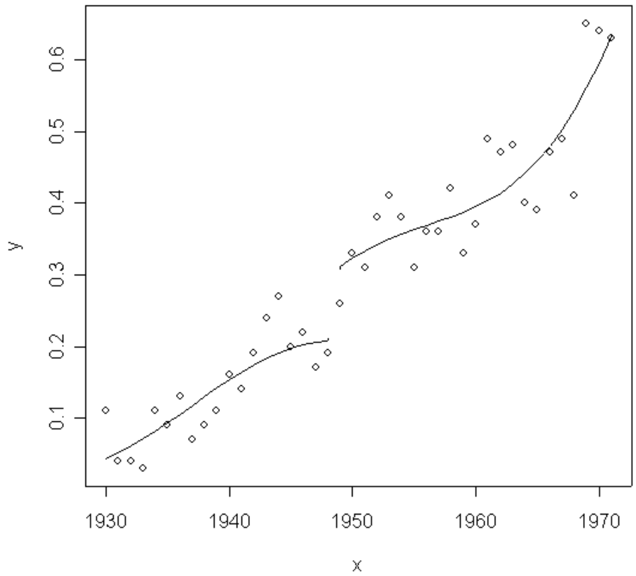


Figura 5.25: Conseguimento maturità femmine 1

```
x^3:riforma     0.0010205  0.0002472   4.129 3.64e-05 ***
AIC: 4448.3
Analysis of Deviance Table

Model 1: y ~ x^2 + x^3 + x:riforma + x^3:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3971     4438.3
2      3934     4392.8   37     45.5       0.2
```

Il secondo modello (vedi grafico 5.26) presenta un andamento leggermente migliore del primo; la differenza fra i due è che uno prevede la riforma, l'altro la potenza terza della coorte, come variabile esplicativa.
Utilizzando il modello predefinito, noto come nè il momento secondo, nè la riforma risultino significativi.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.693e+02  3.947e+01  -4.289 1.79e-05 ***
x            8.635e-02  2.036e-02   4.241 2.22e-05 ***
```

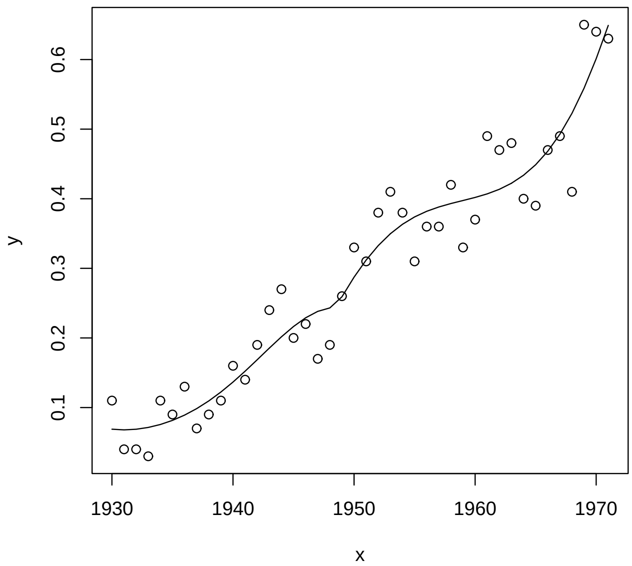


Figura 5.26: Conseguimento maturità femmine 2

```
x^2           -5.033e-04  3.525e-04  -1.428    0.153
riforma      2.122e-01  1.581e-01   1.342    0.180
```

E' possibile ristimare il modello senza riforma.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.031e+02  3.101e+01  -6.549 5.79e-11 ***
x            1.038e-01  1.599e-02   6.493 8.41e-11 ***
x^2         -7.143e-04  3.199e-04  -2.233   0.0255 *
AIC: 4460.3
```

Tuttavia, si può prevedere un modello che abbia, come esplicative, $x$ e *riforma*.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.156e+02  1.140e+01 -10.136   <2e-16 ***
x            5.863e-02  5.873e-03   9.983   <2e-16 ***
riforma      3.134e-01  1.428e-01   2.195   0.0282 *
```

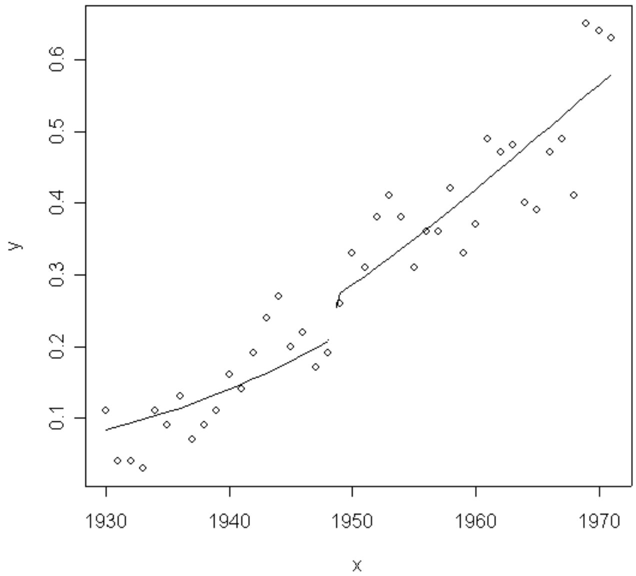


Figura 5.27: Conseguimento maturità femmine (modello quadratico 1)

```
AIC: 4460.5
```

Ancora una volta, ci troviamo di fronte a due modelli con adattamento ai dati pressochè identico, di cui uno prevede, tra le esplicative, la variabile *riforma*, l'altro una variabile altrettanto significativa (in questo caso, non più la potenza terza, ma la seconda). I grafici relativi sono il (5.27) e il 5.28).
Stimando una previsione sui valori successivi al '48, vediamo come ancora una volta il modello quadratico conduca a valori decrescenti già a partire dagli anni '50 (vedi grafico 5.29). Utilizzando il modello lineare, invece, possiamo notare come solo nella seconda metà degli anni '60 comincino a esserci valori al di sotto delle bande di confidenza (vedi grafico 5.30).
In conclusione, se la scelta del modello metteva il conseguimento di un diploma qualsiasi e il raggiungimento del titolo di maturità in posizioni simili (due modelli accettabili, di cui uno prevedeva un salto notevole, l'altro nessuna discontinuità), i controlli fanno propendere decisamente per l'assenza di impatto nella primo situazione, mentre pare plausibile che ci sia una discontinuità tra la coorte '48 e la '49 nel caso del titolo quinquennale.

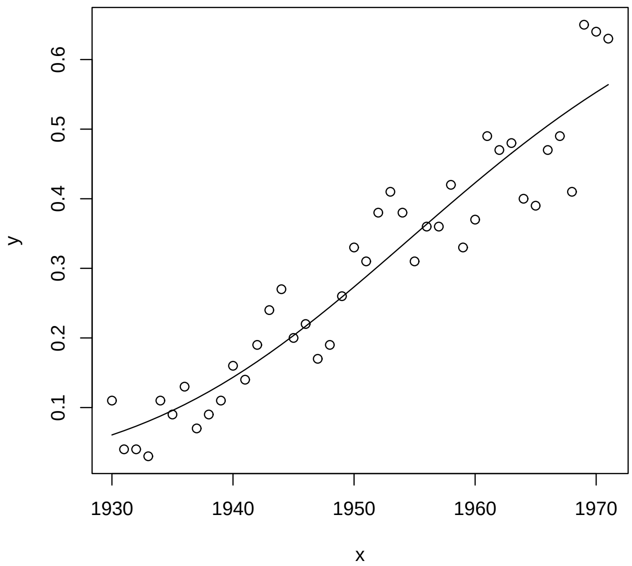


Figura 5.28: Conseguimento maturità femmine (modello quadratico 2)

### Le iscrizioni all'università

Passo, ora, a valutare la percentuale di iscrizioni femminili all'università per anno di nascita, fissando ancora l'eventuale discontinuità tra le classi '48 e '49. Il modello che prevede l'impatto della riforma si rivela decisamente stabile e parsimonioso, prevedendo solo tre variabili esplicative: oltre a *riforma*, le potenze seconda e terza della coorte di nascita.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0534957  0.1134103 -18.107  < 2e-16 ***
x^2          -0.0025746  0.0006477  -3.975 7.03e-05 ***
x^3           0.0001803  0.0000295   6.113 9.80e-10 ***
riforma      0.9324189  0.1213751   7.682 1.56e-14 ***
AIC: 3589.6
Residual deviance: 3581.6  on 3972  degrees of freedom
```

Tale modello ha anche un buon adattamento ai dati, tuttavia sovrastima verosimilmente l'impatto delle riforme, come si evince dal grafico (5.31) , che assume un andamento calante a partire proprio dal '49, per qualche anno.

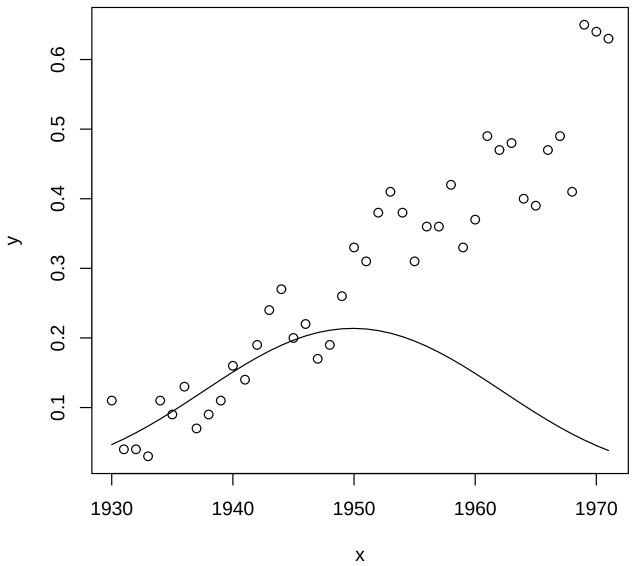


Figura 5.29: Conseguimento maturità femmine (modello previsivo 1)

```
 Analysis of Deviance Table

Model 1: y ~ x^2 + x^3 + riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3972     3581.6
2      3934     3537.0   38     44.6       0.2
```

Il salto complessivo, infatti, è pari a:

```
0.2458116- 0.1136996= 0.132112
```

.Si prevederebbe, in pratica, più di un raddoppio delle iscrizioni femminili all'università per coorte a causa della sola riforma universitaria. E' evidente come tale grafico risenta dei valori anomali (troppo bassi) in un intorno della soglia. Tuttavia, qualsiasi modello che abbia un adattamento discreto ai dati presenta questo problema. Il modello che non preveda la significatività della riforma, invece, è questo:

```
Call:
glm(formula = y ~ x + x^2 + x^3 + x^2:riforma, family = binomial)
```

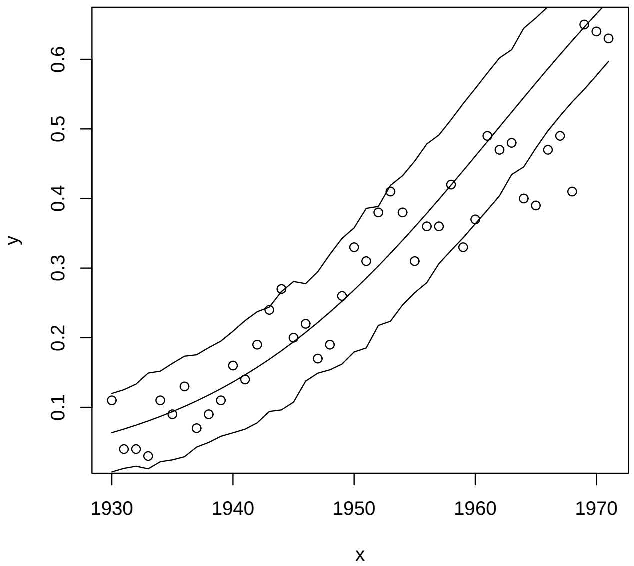


Figura 5.30: Conseguimento maturità femmine (modello previsivo 2)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0500  -0.7533  -0.5200  -0.2686   2.8320

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5985707  0.0740441 -21.589  < 2e-16 ***
x            0.1581707  0.0317774   4.977 6.44e-07 ***
x^2           0.0111061  0.0043474   2.555 0.010630 *
x^3           0.0005162  0.0001373   3.759 0.000170 ***
x^2:riforma  -0.0271989  0.0084854  -3.205 0.001349 **
Residual deviance: 3585.2
AIC: 3595.2
Analysis of Deviance Table

Model 1: y ~ x + x^2 + x^3 + x^2:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
```

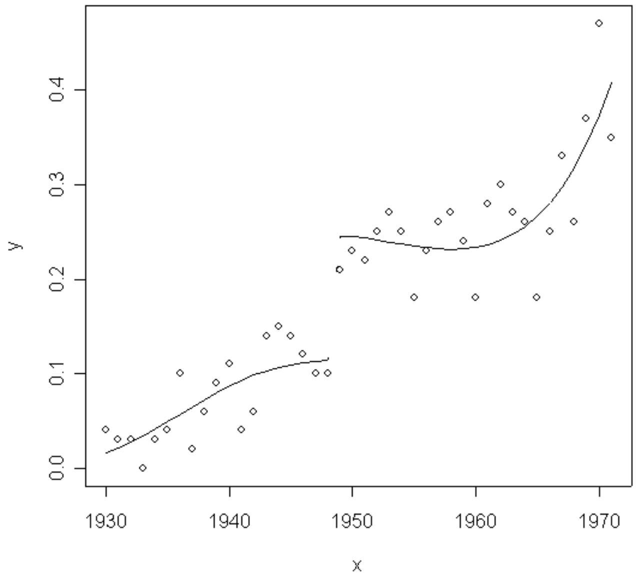


Figura 5.31: Iscrizioni università femmine

```
1       3971      3585.2
2       3934      3537.0   37      48.2        0.1
```

Esso, in pratica, presenta $x$ e $x^2$:*riforma* come variabili esplicative al posto di *riforma*. Tuttavia, non solo ha un adattamento ai dati peggiore e un coefficiente A.I.C. più alto, ma addirittura, pur avendo un parametro in più, è caratterizzato da una varianza dei residui maggiore rispetto al modello precedentemente adottato. Circa il primo controllo, si noti come esso preveda ancora la significatività della riforma.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.3905575  0.2675293 -12.674  < 2e-16 ***
x            0.0920818  0.0261552   3.521 0.000431 ***
x^2           -0.0009241  0.0004408  -2.096 0.036073 *
riforma      0.4320028  0.1963683   2.200 0.027810 *
```

Esso, peraltro, pare presentare una discontinuità più verosimile:

```
0.1802952-0.1249517= 0.0553435.
```

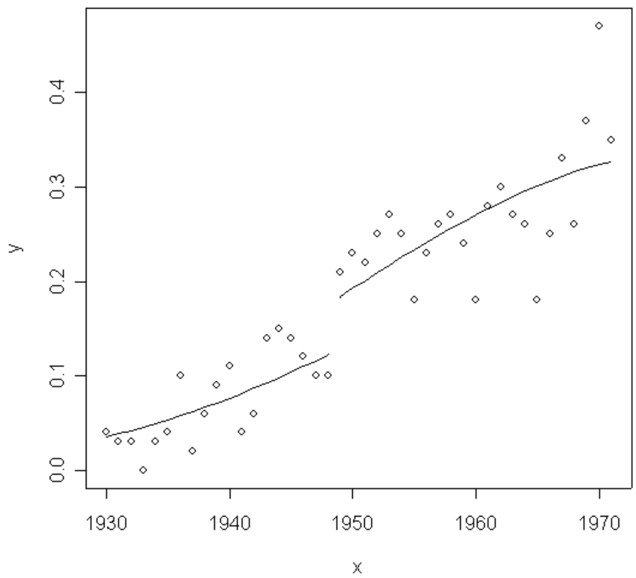


Figura 5.32: Iscrizioni università femmine (modello quadratico)

Come si vede dal grafico (5.32), ciò è dato dal fatto che l'andamento risulta costantemente crescente. Peraltro,è plausibile pensare che il modello sottostimi sia i valori pre-riforma (risentendo anch'esso del calo riscontrato tra il '48 e il '49), sia quelli post (dato che i valori campionari per le coorti immediatamente successive sono costantemente superiori a quelli teorici), e che l'entità della discontinuità misurata, quindi, non si discosti tanto dalla realtà. Sostituendo a $x^2$ la variabile *x:riforma*, peraltro, è possibile stimare un modello simile che, in più, non utilizza le potenze della coorte come esplicative, ed ha un adattamento accettabile, prevedendo un salto analogo

```
(0.2006495-0.1454242) =0.0552253

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.77095    0.15251 -11.612  < 2e-16 ***
x            0.09202    0.01863   4.941 7.78e-07 ***
riforma      0.38871    0.18300   2.124  0.03367 *
x:riforma   -0.06120    0.01995  -3.068  0.00215 **
AIC: 3599.4
Analysis of Deviance Table

Model 1: y ~ x + x:riforma + riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3972     3591.4
2      3934     3536.8   38     54.6   0.03948
```

Valutando, infine, le previsioni basate sui dati delle coorti fino al '48, ancora una volta si deve ricorrere al modello lineare, essendo quello quadratico decrescente già dalle classi immediatamente successive al '48 (vedi grafici 5.33 e 5.34). Si vede come i valori siano al di sotto delle bande di confidenza solo dagli anni '60; così come quelli dal '49 al '55 si rilevino vicini alla banda superiore. Si può ritenere, in conclusione, che la discontinuità tra il '48 e il '49 sia più evidente rispetto a quanto visto in occasione del conseguimento del diploma di maturità. Un ulteriore conferma di questo fatto è che il modello con tutte e 9 le variabili esplicative preveda proprio il parametro associato alla riforma universitaria come l'unico significativo.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.336e+00  4.110e-01  -5.684 1.32e-08 ***
```

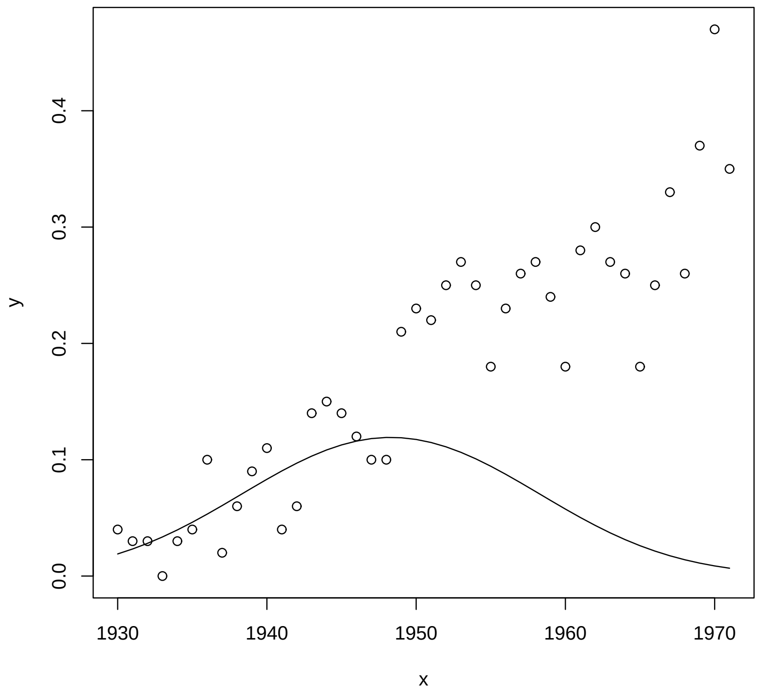


Figura 5.33: Iscrizioni università femmine previsto 1

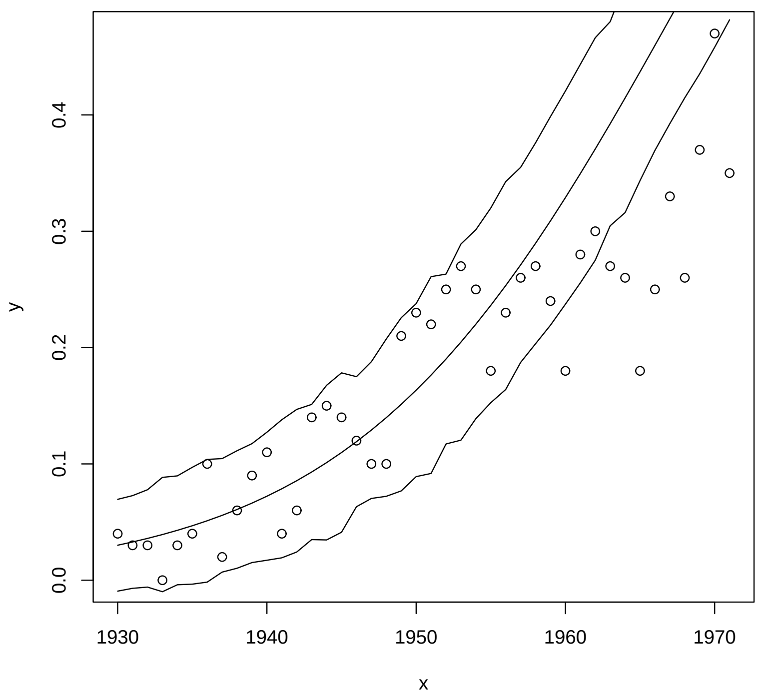


Figura 5.34: Iscrizioni università femmine previste 2

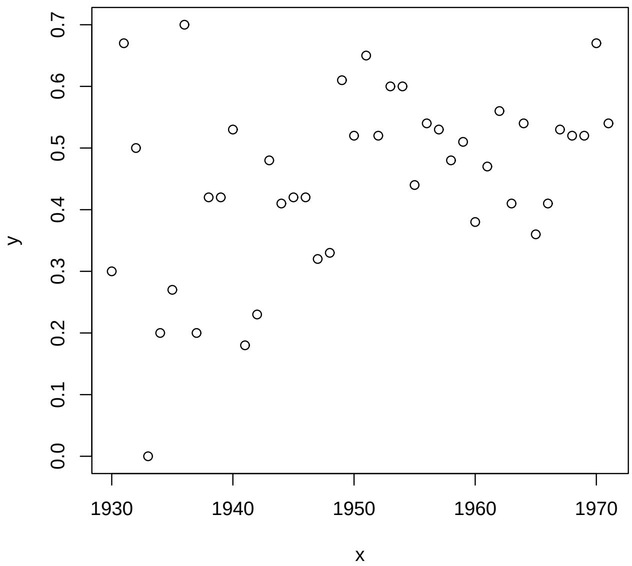


Figura 5.35: Iscrizioni università femmine diplomate

| | | | | | |
|---|---|---|---|---|---|
| x | -2.474e-01 | 3.262e-01 | -0.758 | 0.448 | |
| x^2 | -4.306e-02 | 7.511e-02 | -0.573 | 0.566 | |
| x^3 | -1.557e-03 | 6.352e-03 | -0.245 | 0.806 | |
| x^4 | -6.767e-06 | 1.753e-04 | -0.039 | 0.969 | |
| riforma | 1.012e+00 | 4.903e-01 | 2.064 | 0.039 | * |
| x:riforma | 3.262e-01 | 3.610e-01 | 0.904 | 0.366 | |
| x^2:riforma | 3.309e-02 | 7.968e-02 | 0.415 | 0.678 | |
| x^3:riforma | 1.948e-03 | 6.575e-03 | 0.296 | 0.767 | |
| x^4:riforma | 6.018e-06 | 1.790e-04 | 0.034 | 0.973 | |

Peraltro, limitando l'analisi al comportamento delle donne in possesso di un diploma di maturità, si può osservare dal grafico (5.35) come non ci sia un andamento monotono nè prima nè dopo la soglia. Tuttavia, limitando l'analisi alle nate tra il '46 e il '57, si può vedere che la percentuale di donne iscritte a un corso universitario sia significativamente superiore tra le coorti '49/'54 che in quelle immediatamente precedenti e successive. Inserendo, infatti, nella regressione binomiale la variabile indicatrice *gruppo* (che assume valore 1 per le coorti centrali, 0 altrimenti), ottengo una evidente significatività di tale parametro.

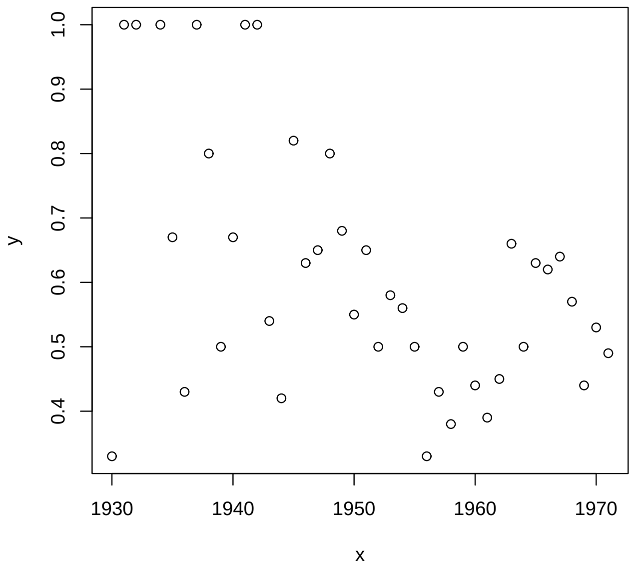


Figura 5.36: Iscrizioni università femmine diplomate

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2418     0.1298  -1.864  0.06239 .
x             0.5320     0.1848   2.879  0.00399 **
```

La differenza nelle medie tra i due gruppi, d'altronde, è netta:

`0.5720451-0.4398428= 0.1322023.`

Valutando, invece, quante, tra le iscritte all'università, raggiungano poi un titolo di istruzione terziaria, si può osservare, pur nell'ambito, ancora, dell'assenza di monotonia, il calo di conseguimento del titolo a partire dal '55 (vedi grafico 5.36). Confrontando, stavolta, le coorti '49/'54 con quelle '55/'60, otteniamo:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4551     0.1664   2.734  0.00625 **
x            -0.5064     0.2310  -2.192  0.02838 *
```

La significatività, pur essendo meno schiacciante di prima, è del livello del 2.84%. La differenza, comunque, è di circa il 12.5%:

```
0.6118511-0.4871778= 0.1246733.
```

Quello che si può concludere, dunque, è che i risultati di conseguimento del titolo di laurea notevolmente più alti a quelli precedenti e successivi tra le donne nate nel periodo '49/'55 sono dovuti a tre fattori (anche se, circa il primo di essi, esiste qualche dubbio).

- Vi è probabilmente una discontinuità tra le coorti '48 e '49, per quanto riguarda la probabilità di conseguimento di un diploma di maturità
- Tra tali diplomate, vi è un tasso più alto di iscrizione all'università delle coorti in questione rispetto a quelle '46/'48 e '55/'57. Il risultato è che la discontinuità nel tasso d'iscrizione all'università nella popolazione è nettamente più significativa di quanto lo fosse nel caso del titolo superiore quinquennale.
- All'interno delle iscritte all'università, le classi che stiamo esaminando hanno conseguimenti che, pur in linea con quelli dei periodi precedenti, sono decisamente superiori a quelli ottenuti dalle iscritte nate nel periodo successivo ('55/'60).

## 5.4 Anni d'istruzione complessiva

### 5.4.1 Stima numero anni di corso

Per quanto riguarda la stima degli anni medi di istruzione per coorte, è stato ovviamente necessario ricorrere a qualche approssimazione (come anticipato nell'Introduzione). Ho attribuito 5 anni per il possesso del titolo di scuola elementare e 3 per quello di scuola media. Riguardo alla scuola superiore, ho assegnato il numero di anni di corso relativi qualora esso fosse desumibile dalle risposte dell'individuo; 4.5 anni nel caso in cui non fosse chiaro se avesse conseguito un titolo quadriennale o quinquennale, 2.5 qualora il dubbio fosse tra titolo biennale o triennale, 4 anni nel caso in cui non si potesse risalire alla durata del corso. Tuttavia, queste situazioni sono relative solo a una manciata di individui, dunque tali approssimazioni non incidono più di tanto sul modello complessivo. Relativamente all'università, ho proceduto attribuendo il numero di anni di corso relativi (assegnando valore 2.5 per quei diplomi la cui durata non fosse chiara, e 5 anni per i corsi di laurea). Anche nei casi di corsi post-laurea, a volte, si è reso necessario attribuire un valore intermedio, qualora non si riuscisse a stabilire la natura biennale o triennale dei corsi. Circa l'iscrizione a istituti senza il raggiungimento del titolo, ad eccezione dell'università, mi sono basato sull'ultima classe frequentata, attribuendo un numero d'anni d'istruzione pari al numero di tale classe, sottrendo poi 0.5. Per esempio, chi avesse frequentato fino alla terza classe delle superiori, otterrebbe un punteggio di 2.5 (che poi darebbe un totale di 10.5 anni d'istruzione). Nel caso tale classe, però, fosse l'ultima del corso, non ho conteggiato il mezzo punto, essendo chiaro che essa non è stata portata a termine con successo (ad esempio, assegno 4 anni a chi ha frequentato l'ultima classe della scuole superiori senza diplomarsi). Relativamente agli iscritti all'università che non han conseguito il titolo, invece, ho assegnato mezzo anno a chi non ha sostenuto esami; un anno a chi invece ha superato almeno un esame.

### 5.4.2 Modello stimato

La stima dei modelli risulta più semplice che nei casi precedenti. Qui, infatti, siamo nel caso del modello lineare (la minimizzazione, però, avviene non in relazione ai MQO, ma con pesi pari alle frequenze osservate per anno). La regressione del numero medio di anni d'istruzione per coorte sulla classe stessa pare seguire un *trend* lineare, per entrambi i sessi e sia prima che dopo la soglia (che fisso nuovamente tra il '51 e il '52). Testando, infatti, il modello lineare contro un modello non parametrico dipendente da un parametro $h$[6](vedi Azzalini e Bowman, 1997), si vede dai quattro grafici relativi (5.37, 5.38, 5.39, 5.39) come la curva non scenda mai sotto la soglia dello 0.05% (quindi, come l'ipotesi di linearità non verrebbe mai rifiutata). Anche se tali test tengono in considerazione solo i valori medi per anno, non considerando le frequenze, tali risultati sono comunque indicativi. Osservando, invece, i due grafici che si riferiscono al periodo completo (5.41 e 5.42), si nota come, per valori di $h$ oltre una certo livello, tale ipotesi verrebbe rifiutata. Stimiamo i due modelli utilizzando dunque tre parametri:$x$, *riforma* e *x:riforma*. Questa volta, utilizzo un modello lineare, quindi l'*output* fornirà una stima dell'effettiva incidenza delle variabili esplicative sul numero di anni di istruzione. Il parametro: *riforma* risulta comunque non significativo. Per le donne, infatti, è:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.66044    0.21447  45.043  < 2e-16 ***
x            0.21313    0.01785  11.938 1.99e-14 ***
riforma      0.27122    0.30313   0.895 0.376563
```

[6]i modelli non parametrici, e questo in particolare, verranno meglio approfonditi in Appendice

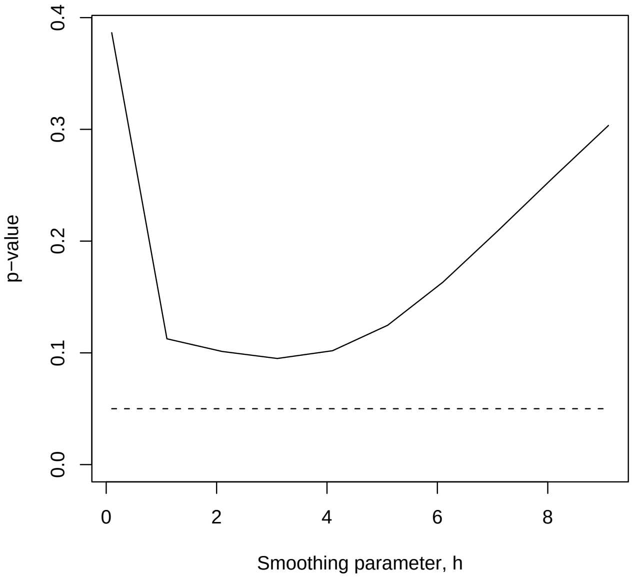


Figura 5.37: Test linearità femmine pre-riforma

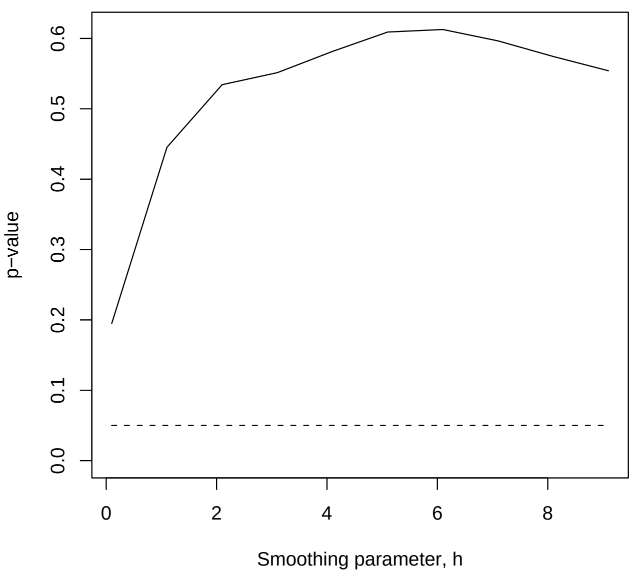


Figura 5.38: Test linearità femmine post-riforma

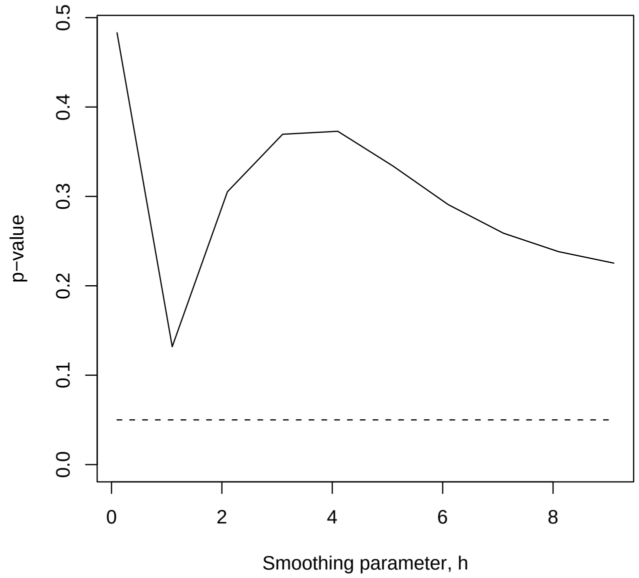


Figura 5.39: Test linearità maschi pre-riforma

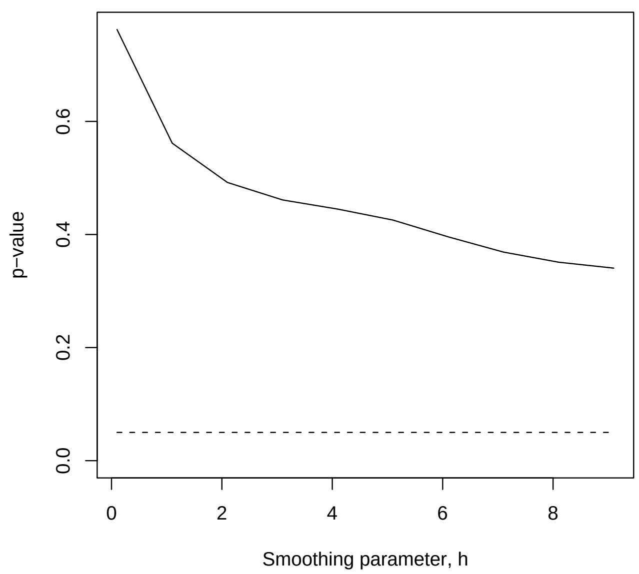


Figura 5.40: Test linearità maschi post-riforma

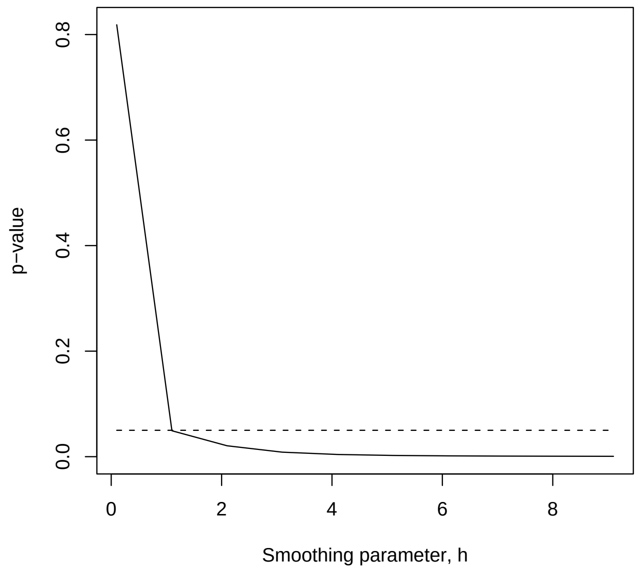


Figura 5.41: Test linearità femmine totale

```
x:riforma   -0.09518    0.02548  -3.735 0.000615 ***
```

mentre per gli uomini è:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.38582    0.18355  56.582  < 2e-16 ***
x            0.15762    0.01560  10.102 2.57e-12 ***
riforma      0.45625    0.26458   1.724   0.0928 .
x:riforma   -0.11971    0.02244  -5.335 4.64e-06 ***
```

Ristimo, quindi, i due modelli in assenza del parametro *riforma* (vedi grafici 5.44 e 5.43). Per le donne, è:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.79621    0.15118   64.80  < 2e-16 ***
x            0.22278    0.01418   15.71  < 2e-16 ***
x:riforma   -0.09481    0.02542   -3.73 0.000608 ***
Multiple R-Squared: 0.9513,     Adjusted R-squared: 0.9488
```

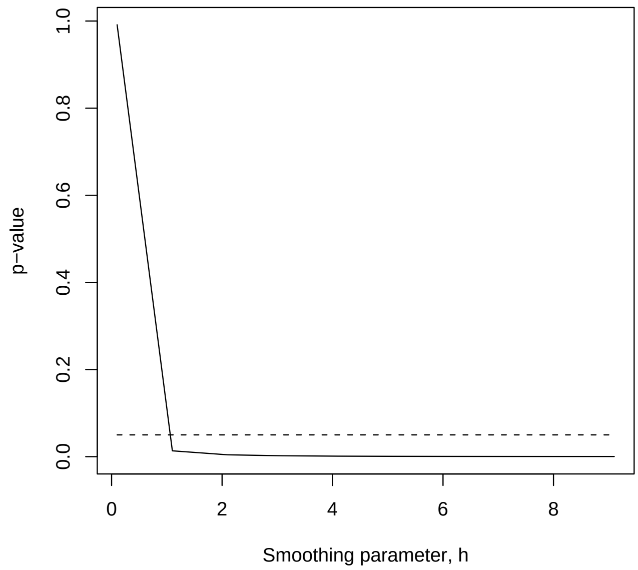


Figura 5.42: Test linearità maschi totale

mentre per gli uomini:

```
 Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.60541    0.13550  78.268  < 2e-16 ***
x            0.17344    0.01294  13.408 3.47e-16 ***
x:riforma   -0.11803    0.02298  -5.137 8.11e-06 ***
Multiple R-Squared: 0.9126,     Adjusted R-squared: 0.9082
```

Vediamo come i valori di $R^2$ ed $R^2$ *aggiustato* (misure della capacità esplicativa del modello) risultino intorno a 0.9 per gli uomini, 0.95 per le donne. Negli uomini, tra l'altro, il parametro *riforma* risulterebbe significativo, se si usasse come soglia il 10%. I risultati, in realtà, non sono sorprendenti: abbiamo visto come, per le donne, le riforme comincino a produrre risultati già a partire dalla coorte 1949. Peraltro, riferendoci agli uomini, un aumento di circa il 6-7% dei conseguimenti della licenza media, qualora assumiamo che questo sia l'unico impatto della riforme scolastiche e che questo gruppo di persone, in assenza delle riforme stesse, non si sarebbe nemmeno iscritto alla scuola media, produrebbe una aumento di circa 0.2 anni d'istruzione, difficilmente apprezzabile in un campione non numerosissimo.

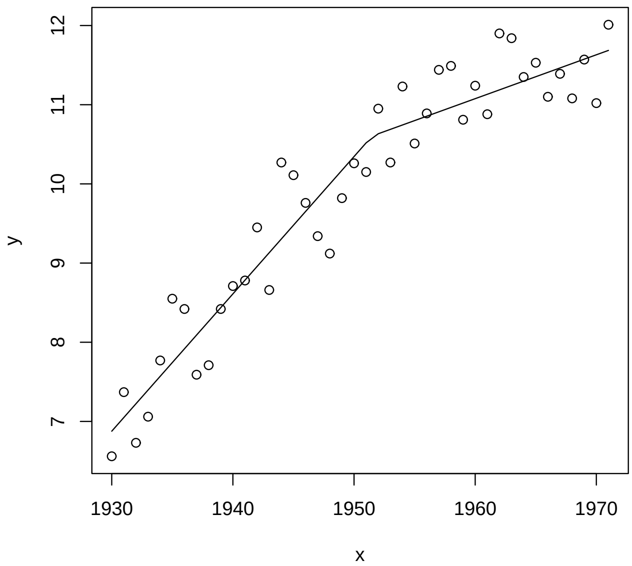


Figura 5.43: Anni istruzione maschi

### 5.4.3 Modello quadratico

Anche utilizzando il modello quadratico, il parametro *riforma* risulta assolutamente non significativo per le donne.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.943e+02  5.667e+01  -8.722 1.32e-10 ***
x            2.587e-01  2.925e-02   8.845 9.20e-11 ***
x^2          -2.107e-03  5.989e-04  -3.518  0.00114 **
riforma      2.933e-01  3.080e-01   0.952  0.34695
```

Così, il grafico (5.45) è espressione del modello :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.131e+02  5.307e+01  -9.669 6.57e-12 ***
x            2.684e-01  2.739e-02   9.801 4.51e-12 ***
x^2          -2.086e-03  5.978e-04  -3.489  0.00122 **
```

Ancora una volta, peraltro, il parametro *riforma* sarebbe significativo per gli uomini alla soglia del 10%, non a quella del 5%.

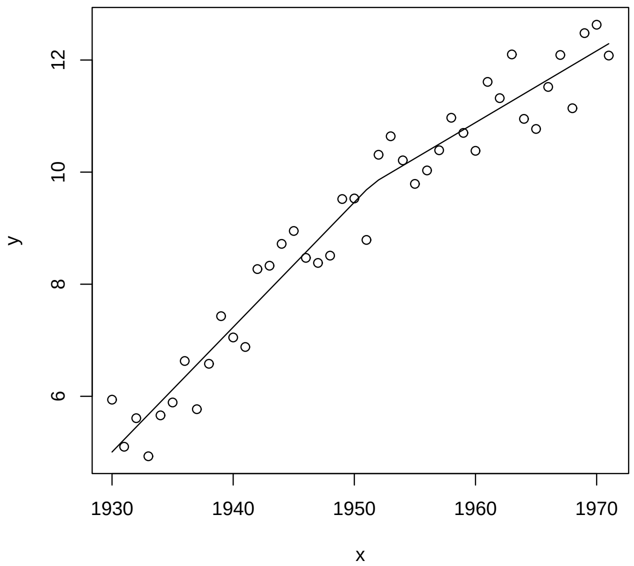


Figura 5.44: Anni istruzione femmine

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.351e+02  4.832e+01  -9.004 5.80e-11 ***
x           2.289e-01  2.493e-02   9.181 3.47e-11 ***
x^2          -2.942e-03  5.112e-04  -5.754 1.24e-06 ***
riforma     4.634e-01  2.558e-01   1.811    0.078 .
```

Così, anche per i maschi si utilizza il modello a due variabili esplicative (vedi grafico 5.46).

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.642e+02  4.689e+01  -9.899 3.41e-12 ***
x           2.439e-01  2.420e-02  10.081 2.03e-12 ***
x^2          -2.899e-03  5.254e-04  -5.517 2.43e-06 ***
```

### 5.4.4 Previsione

Nuovamente, vediamo come il modello previsivo non sia buono basandosi su di un modello quadratico (vedi grafici 5.47 e 5.48): per quanto riguarda le

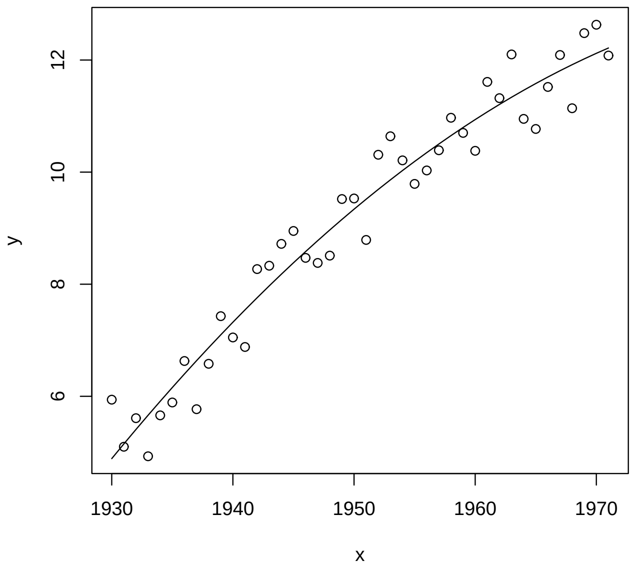


Figura 5.45: Anni istruzione femmine (modello quadratico)

donne, infatti, le bande di confidenza diventerebbero eccessivamente larghe, mentre per gli uomini si presenterebbe l'usuale problema di un andamento calante già a partire dalle coorti degli anni '50. Si noti, comunque, come i valori nel periodo post-riforma si trovino al di sotto delle bande di confidenza solo a partire dagli anni '60, elemento di conferma che in realtà un qualche effetto delle riforme ci sia sostanzialmente stato (vedi grafici 5.49 e 5.50).

## 5.5 Altri possibili effetti

Valuto, per verificare quanto sia stata rispettata la normativa del '62 sin da subito, se ci sia stato un effetto anche sulle iscrizioni alle scuole medie. Prima di tutto, verifico la percentuale immediata di iscritti alle scuole medie per anno di conseguimento della licenza elementare, poi la percentuale di persone che, non essendosi iscritte immediatamente, lo facciano l'anno successivo. Infine, verifico la percentuale di iscritti alla scuola media che abbiano poi conseguito la licenza relativa. Continuo, quindi, ad operare con variabili binarie. Tuttavia, mi concentro non più sull'anno di nascita, ma sull'anno di conseguimento della licenza elementare o di iscrizione alla scuola media,

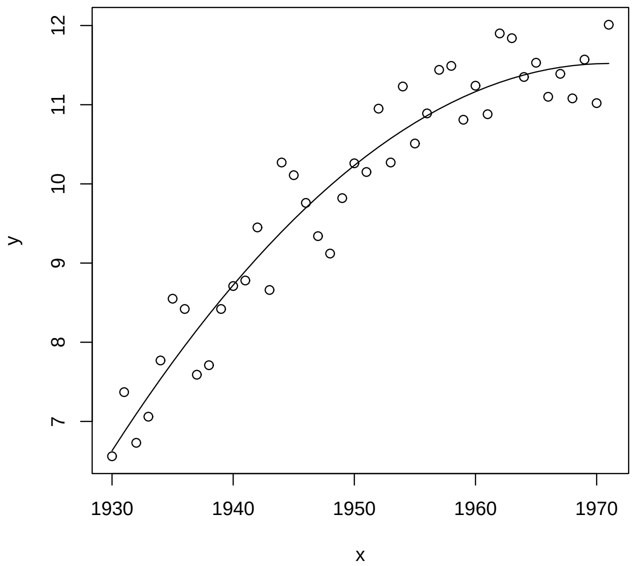


Figura 5.46: Anni istruzione maschi (modello quadratico)

utilizzando come popolazione gli anni compresi tra il '41 e l'82, e l'eventuale discontinuità tra il '72 e il '73. L'unico effetto significativo, seppure *border line*, sembra quello sulle iscrizioni alla scuola media per anno di iscrizione, relativamente alle donne.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.629448   0.094873   6.635 3.25e-11 ***
x           0.069571   0.008577   8.112 4.99e-16 ***
riforma     0.296830   0.150793   1.968  0.04902 *
x:riforma   0.047181   0.015231   3.098  0.00195 **
```

Il salto risulta pari a

```
0.7163196- 0.6523643 = 0.0639553
```

(vedi grafico 5.51): il fatto che più di 6 punti percentuali di salto risultino significativi solo al limite è evidentemente dovuto alla ridotta numerosità campionaria (d'altra parte, la popolazione qui è ridotta alle iscritte alle scuole medie).

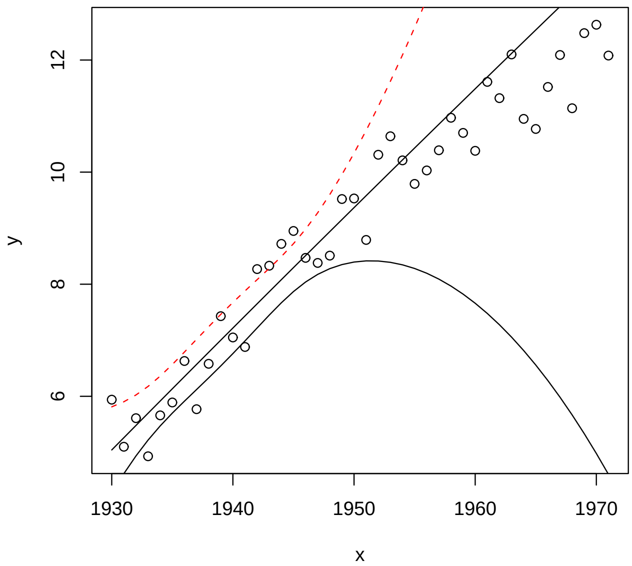


Figura 5.47: Anni istruzione femmine (modello previsivo 1)

## 5.6 Conclusioni

Relativamente all'effetto delle riforme scolastiche, si può affermare quanto segue. La prima, pur non essendo stata fatta rispettare in modo pieno sin da subito[7] (come dimostra il fatto che le iscrizioni alla scuola media per che ha appena conseguito la licenza elementare mantengono un tasso inferiore al 90% fino alla fine degli anni '60, e che chi non effettua l'iscrizione immediata difficilmente viene poi costretto a farlo l'anno successivo), produce comunque qualche effetto. Quelli evidenti sono dati da un incremento delle persone che conseguono la licenza media per coorte di nascita (sia per gli uomini che per le donne). Probabilmente, poi, per le donne, c'è anche un incremento delle iscrizioni immediate alla scuola media dato il conseguimento della licenza elementare, nonchè un aumento di ottenimenti del titolo di maturità. Quest'ultimo, oltre a non essere schiacciantemente significativo, potrebbe essere dovuto anche a un effetto della liberalizzazione degli accessi universitari, considerando anche che non vi è un analogo aumento dei conseguimenti di diplomi di scuola superiore (intendendo anche quelli di durata inferiore al

[7] come riscontrato anche da Checchi, 1997

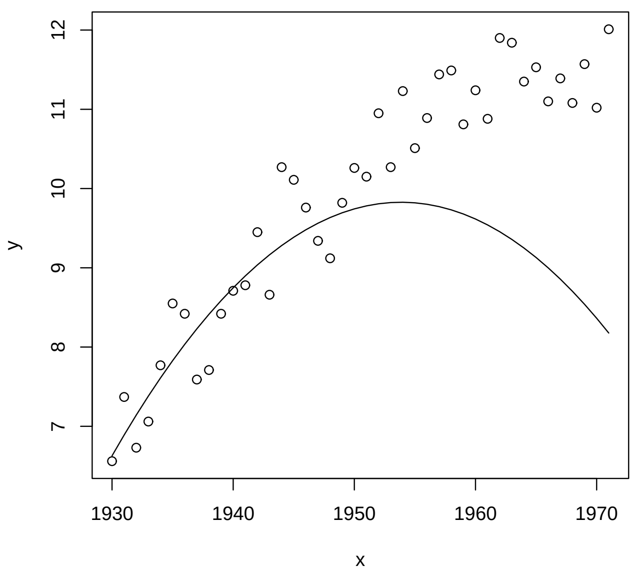


Figura 5.48: Anni istruzione maschi (modello previsivo 1

quinquennio). Si potrebbe ritenere, però, che ciò sia dovuto anche a una saturazione dei livelli d'istruzione inferiore, che abbia così costretto le donne a conseguire un diploma di maturità per "segnalarsi" al mercato del lavoro.
Riguardo alla riforma universitaria, invece, questa pare aver prodotto effetti solo sulle donne. Al di là, infatti, dell'eventuale maggior probabilità di conseguimento del titolo di maturità, di cui ho appena trattato, si può osservare come le coorti tra il '49 e il '54 abbiano un tasso d'iscrizione all'università particolarmente elevato. Il fatto, poi, che queste classi siano poi seguite da altre con tassi di "mortalità universitaria" piuttosto elevati, rispetto agli anni precedenti, fa sì che il risultato di tali coorti (riguardo all'ottenimento di un titolo universitario) risulti particolermente superiore a quello delle classi immediatamente precedenti e successive.
Si può ritenere che il fatto che le donne abbiano verosimilmente risentito di più delle riforme universitarie, complessivamente, sia dovuto alla partenza da una situazione di minore istruzione (riferendomi alle prime classi toccate dalla riforma). Inoltre, si può pensare che le coorti tra il '49 e il '56, avvenendo la liberalizzazione degli studi quando esse hanno, per lo più, già deciso quale scuola superiore intraprendere, risentano della riforma in quanto essa può modificare i loro progetti in relazione all'istruzione, mentre la stessa cosa

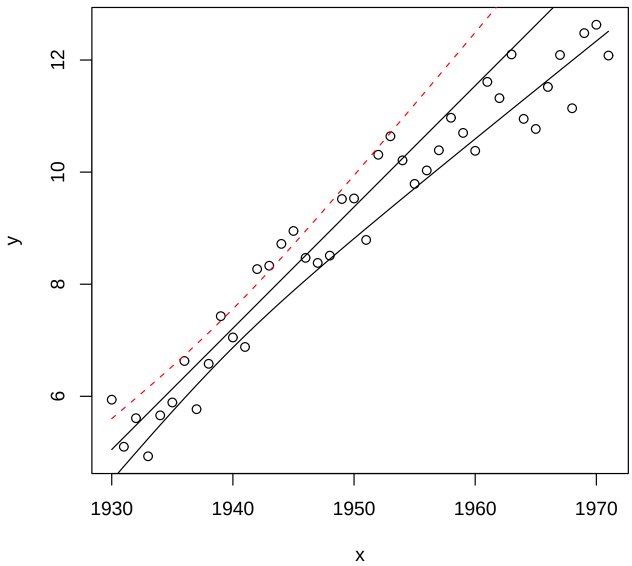


Figura 5.49: Anni istruzione femmine (modello previsivo 2)

non succede, invece, per chi sa già a priori che la scelta della scuola superiore non sarà vincolante per un'eventuale prosecuzione degli studi (Brandolini e Cipollone, 2002).

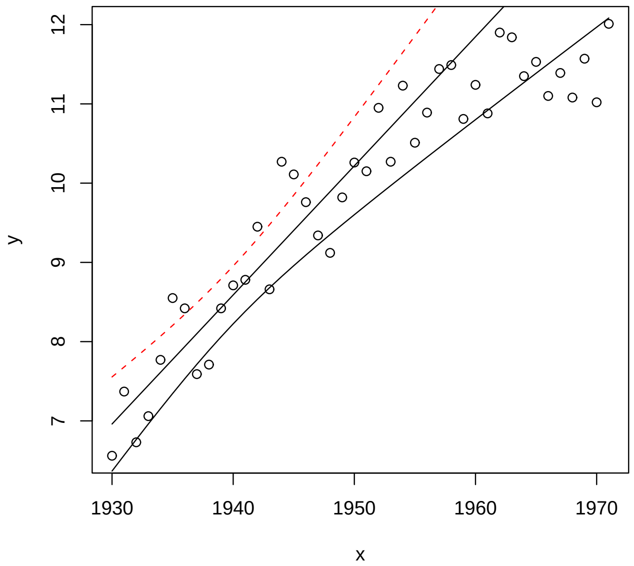


Figura 5.50: Anni istruzione maschi (modello previsivo 2)

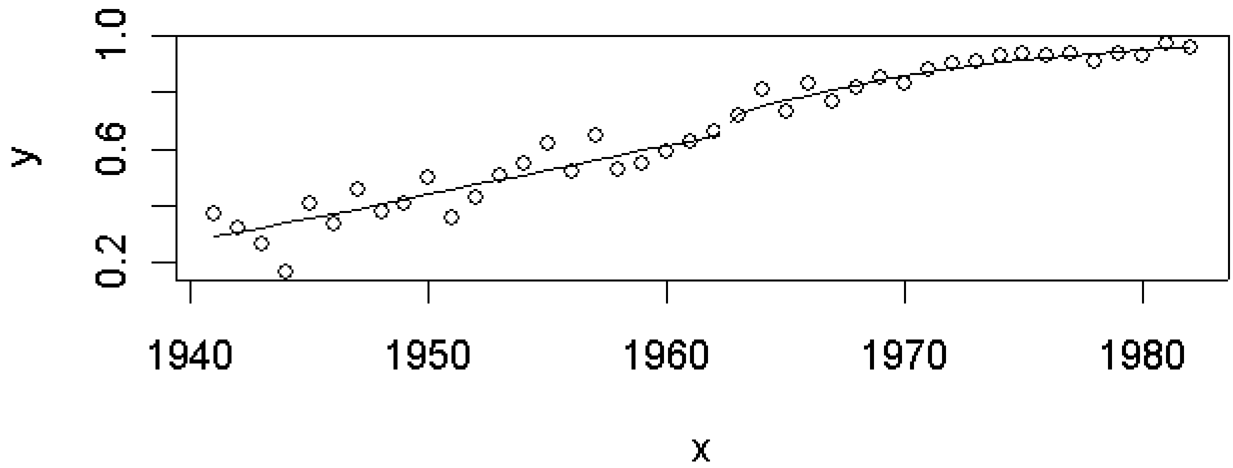


Figura 5.51: Conseguimento licenza medie femmine

# Capitolo 6

# Verifica dell' impatto dell'istruzione sulle scelte di fecondità

## 6.1 Introduzione

Passo, ora, a valutare se le prime donne ad aver risentito dell'effetto delle riforme siano state condizionate dalle stesse anche in relazione alle scelte relative ai figli. Mi concentro, dapprima, sul momento della nascita (o adozione) del primogenito. Innanzitutto, valuto un confronto tra le nate nel '51 e quelle del '52; quindi, rivolgo l'attenzione alla variabile binomiale che esprime se le donne abbiano già un figlio o no, rispettivamente, al compimento dei 20, 25, 30 e 35 anni. Il tempo che passa prima che accada un evento è, di fatto, un dato "di sopravvivenza" (vedi Appendice). R consente di effettuare confronti tra due gruppi per vedere se è plausibile affermare che la distribuzione dei tempi di sopravvivenza sia la stessa (vedi Klein, Moeschberger, 1997). Utilizzando come gruppi, come preannunciato, le nate nel '51 e nel '52, non risulta esserci alcuna differenza significativa.

```
Call:
survdiff(formula = Surv(age, figlio) ~ nata)

            N Observed Expected (O-E)^2/E (O-E)^2/V
nata=1951  78       66     67.9    0.0549    0.0946
nata=1952 111       99     97.1    0.0384    0.0946

 Chisq= 0.1  on 1 degrees of freedom, p= 0.758
```

Il risultato, dunque, è assolutamente non significativo (il valore che interessa è il *p-value* nell'ultima riga). Nei paragrafi successivi, amplio l'analisi a tutte le coorti tra il '30 e il '71, utilizzando variabili binomiali, per verificare l'eventuale presenza di discontinuità tra le due classi consecutive.

## 6.2 Nascita del primo figlio

In questo capitolo, valuto se una donna ha o meno un figlio al momento del compimento di un dato anno d'età. Ho a disposizione anno e mese sia della nascita della madre, che di quella del primo figlio. Considero una madre che faccia un figlio nello stesso mese in cui compie $x$ anni come se, al suo $x$-esimo compleanno, essa avesse già avuto il figlio. Ovviamente, questi casi sono ridotti, quindi i risultati sarebbero stati pressochè identici qualora avessi operato la scelta opposta.

### 6.2.1 Ventenni

Stimo, con il metodo consueto, un modello per la variabile binomiale assumente valore 1 nel caso in cui la madre, al ventesimo compleanno, non abbia avuto figli, 0 altrimenti. Il modello che prevede *riforma* come parametro significativo contempla, nello stesso momento, un andamento costante per il periodo pre-riforma, ha solo due esplicative ed ha un ottimo adattamento ai dati.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.6955429  0.0951134  28.340  < 2e-16 ***
riforma     -0.6910563  0.1475006  -4.685 2.80e-06 ***
riforma:x^2    0.0040843  0.0008117   5.032 4.86e-07 ***
AIC: 2003.9
Analysis of Deviance Table

Model 1: y ~ riforma + x^2:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3968    1997.87
2      3929    1963.47   39    34.40      0.68
```

Il modello che, invece, non prevede discontinuità ha una variabile esplicativa in più, prevedendo la coorte e la potenza seconda al posto di *riforma*.

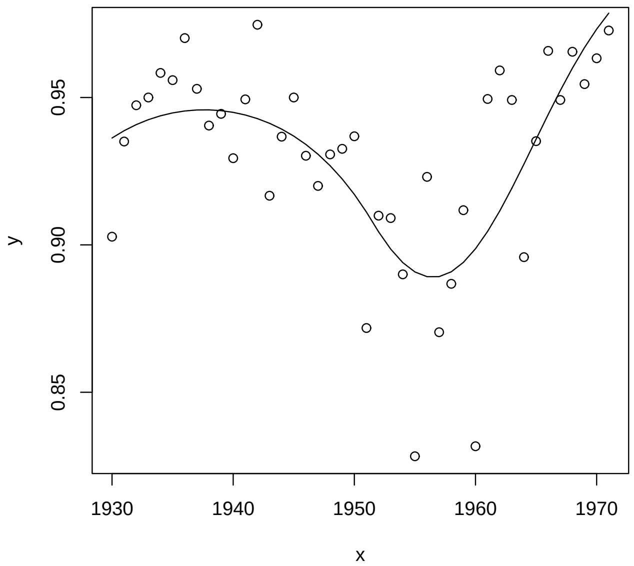


Figura 6.1: Figli 20 anni

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.286787   0.086699  26.376  < 2e-16 ***
x           -0.082587   0.020885  -3.954 7.68e-05 ***
x^2           -0.002976   0.001345  -2.212    0.027 *
x^2:riforma    0.011251   0.002738   4.110 3.96e-05 ***
AIC: 2001.2
```

Il modello pare addirittura preferibile al precedente, quanto ad AIC e adattamento globale ai dati (vedi grafico 6.1).
Il modello quadratico prevede la significatività del parametro *riforma* (vedi grafico 6.2).

```
 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.419e+02  5.096e+01   2.785  0.00536 **
x           -7.192e-02  2.629e-02  -2.735  0.00623 **
x^2            2.227e-03  5.326e-04   4.180 2.91e-05 ***
riforma     -7.214e-01  2.330e-01  -3.097  0.00196 **
```

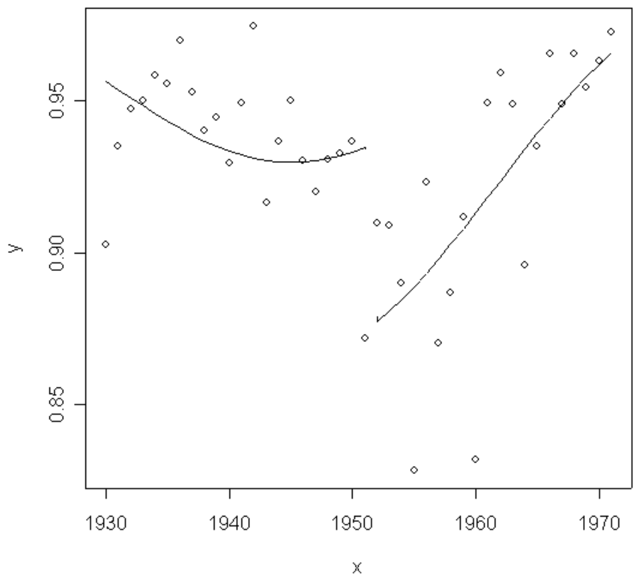


Figura 6.2: Figli 20 anni (modello quadratico)

Procedere a previsioni avrebbe poco senso, dato che è chiara l'inversione di tendenza che avviene a partire dalla metà degli anni '50 (da queste coorti in avanti, si comincia a posticipare, non più ad anticipare, la nascita del primo figlio).
Ciò che stupisce è che il parametro associato alla riforma abbia, sia nel modello che prevede il salto, sia in quello di controllo, valore negativo. L'idea sarebbe che una riforma che aumenta l'istruzione produca, a causa di ciò, un'anticipazione della nascita del primo figlio. Ciò appare piuttosto controintuitivo. In realtà, si può notare come il modello sia fortemente influenzato dal fatto che i valori negli anni '50 (quindi, l'ultimo pre-riforma e i primi al di là della soglia) risultino quasi costantemente più bassi di quelli delle coorti di fine anni '40 e dei primi anni '60. Limitandoci a queste classi (dalla '46 alla '65), tale differenza risulta significativa al 6.64%.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4435     0.1669  14.637   <2e-16 ***
x            -0.3768     0.2053  -1.836   0.0664 .
```

Il fatto che tali valori siano più bassi di quelli precedenti e successivi, comunque, può essere tranquillamente spiegato da un modello continuo, prima calante poi crescente, come quello adottato.

### 6.2.2 Venticinquenni

Spostando a 25 anni la soglia in cui valutare se si ha già procreato o meno, vediamo come il modello che prevede la presenza della variabile *riforma* tra le esplicative sia piuttosto semplice.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.082116   0.089688   0.916  0.35989
x           -0.022258   0.007549  -2.949  0.00319 **
riforma     -0.394547   0.128548  -3.069  0.00215 **
x:riforma    0.120556   0.011357  10.615  < 2e-16 ***
AIC: 5106.8
```

Esso, infatti, non utilizza le potenze della classe di appartenenza, e ha, complessivamente, tre variabili esplicative. Anche l'adattamento generale ai dati è discreto.

```
Analysis of Deviance Table
```

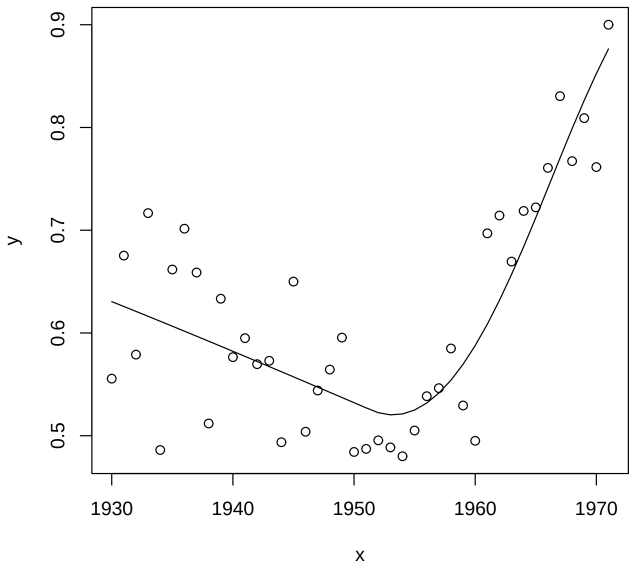


Figura 6.3: Figli 25 anni

```
Model 1: y ~ x + riforma + x:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3967     5098.8
2      3929     5048.7   38     50.1       0.1
```

Il modello che, invece, non prevede l'impatto delle riforme, ha due soli parametri.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0984322  0.0473805   2.077   0.0378 *
x           -0.0202801  0.0046874  -4.327 1.51e-05 ***
x^2:riforma    0.0059331  0.0005592  10.610  < 2e-16 ***
AIC: 5105.5
```

In pratica, si sostituisce $x^2$*:riforma* a *x:riforma* e a *riforma*. Il modello risulta più parsimonioso, con un AIC inferiore e un *fit* analogo, dunque preferibile (vedi grafico 6.3).

```
Analysis of Deviance Table
```

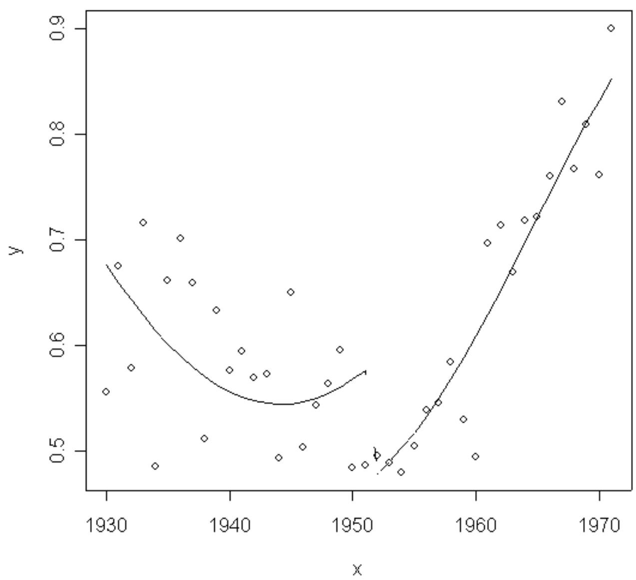


Figura 6.4: Figli 25 anni (modello quadratico)

```
Model 1: y ~ x + x^2:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3968     5099.5
2      3929     5048.7   39     50.8       0.1
```

Col modello quadratico, otteniamo nuovamente la significatività di *riforma*, cui è associato un parametro negativo (vedi grafico 6.4).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.635e+02  2.441e+01   6.698 2.11e-11 ***
x           -8.433e-02  1.260e-02  -6.695 2.16e-11 ***
x^2          2.783e-03  2.712e-04  10.260  < 2e-16 ***
riforma     -4.352e-01  1.292e-01  -3.369 0.000755 ***
```

Di nuovo, non avrebbe senso procedere a previsioni basate sull'andamento nei primi anni, dal momento che il trend, a partire almeno dalla metà degli anni '50, presenta un'inversione di tendenza.

La situazione è del tutto analoga a quella vista usando come soglia i 20 anni. Stavolta, i valori più bassi dei circostanti sono quelli relativi alle coorti tra il '50 e il '55. Confrontandole con le classi '47/'49 e '56/'58, come fatto in precedenza, si ottiene:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.24640    0.08093   3.045  0.00233 **
x           -0.28493    0.11644  -2.447  0.01440 *
```

Stavolta, il fatto che tali classi abbiano valori inferiori risulta significativo addirittura all'1.44%. Di nuovo, lo stesso modello stimato pone queste coorti nel punto di svolta (dunque, come quelle con la percentuale di "sopravvivenza" più bassa), quindi la discontinutà riscontrata da alcuni modelli pare dovuta, sostanzialmente, al fatto che proprio le coorti più vicine all'eventuale salto siano quelle con minore probabilità di non avere procreato al compimento dei 25 anni.

### 6.2.3 Trentenni

Relativamente alla verifica ai 30 anni, invece, occorre premettere che posso considerare solo le donne nate fino al '66, dal momento che le prime interviste delle unità campionarie risalgono al gennaio '97. Sia nel periodo pre-riforma che in quello successivo, comunque, pare che un modello lineare sia adeguato. Stimo, quindi, un modello che escluda le potenze della coorte di nascita dalle esplicative.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.304019   0.104400 -12.491  < 2e-16 ***
x           -0.027150   0.008387  -3.237  0.00121 **
riforma     -0.047733   0.157042  -0.304  0.76116
x:riforma    0.120324   0.015249   7.891 3.00e-15 ***
```

Il parametro associato *riforma* risulta evidentemente non significativo. Di conseguenza, ristimo il modello senza tale variabile (vedi grafico 6.5).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.32521    0.07799 -16.992  < 2e-16 ***
x           -0.02862    0.00686  -4.172 3.02e-05 ***
x:riforma    0.11924    0.01481   8.050 8.29e-16 ***
```

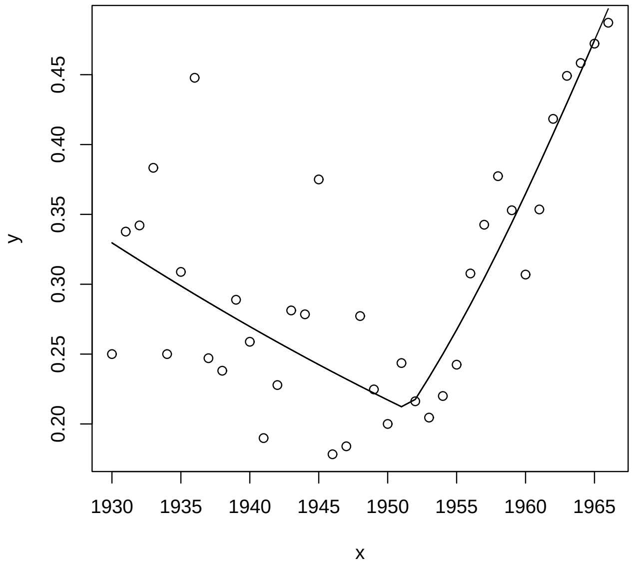


Figura 6.5: Figli 30 anni

L'adattamento ai dati, peraltro, risulta ottimo.

```
Analysis of Deviance Table

Model 1: y ~ x + x:riforma
Model 2: y ~ factor(x) + factor(x):riforma
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3405     4091.2
2      3371     4054.8   34     36.4       0.4
```

Anche il modello quadratico, peraltro, non presenta problemi interpretativi.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.743e+02  2.871e+01   6.069 1.28e-09 ***
x           -9.058e-02  1.483e-02  -6.110 9.98e-10 ***
x^2           2.912e-03  3.841e-04   7.581 3.43e-14 ***
riforma     -1.383e-01  1.579e-01  -0.875    0.381
```

Risultando non significativo l'effetto delle riforme, ristimo usando solo $x$ e $x^2$ (vedi grafico 6.6) .

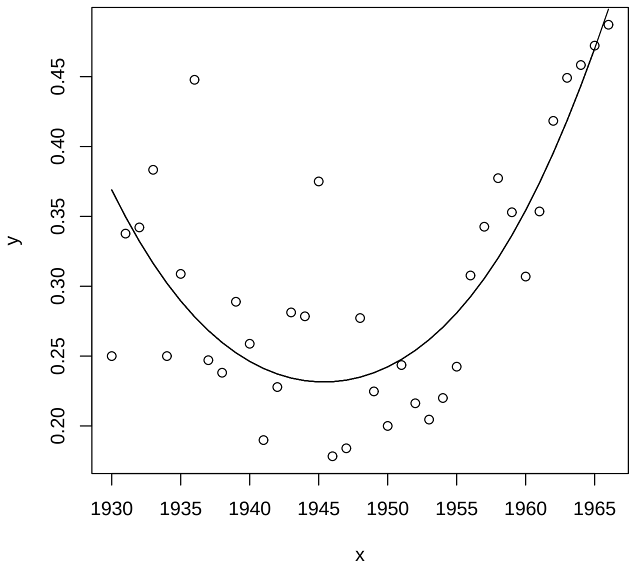


Figura 6.6: Figli 30 anni (modello quadratico

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.768e+02  2.857e+01   6.189 6.06e-10 ***
x           -9.189e-02  1.475e-02  -6.229 4.70e-10 ***
x^2           2.806e-03  3.639e-04   7.710 1.25e-14 ***
```

Ancora una volta, si verifica un'inversione del trend nelle coorti centrali, che rende insensato un modello basato sulla previsione.

### 6.2.4 Trentacinquenni

Stimo, infine, un modello del tipo di quelli visti finora, valutando, però, l'eventuale nascita di un figlio ai 35 anni d'età. Per far questo, devo ulteriormente ridurre le coorti d'interesse, non andando oltre a quella del '61. Non c'è modo di trovare modelli accettabili che prevedano la significatività di *riforma*. Il modello migliore, comunque, è il seguente.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.748e+00  6.955e-02 -25.132  < 2e-16 ***
```

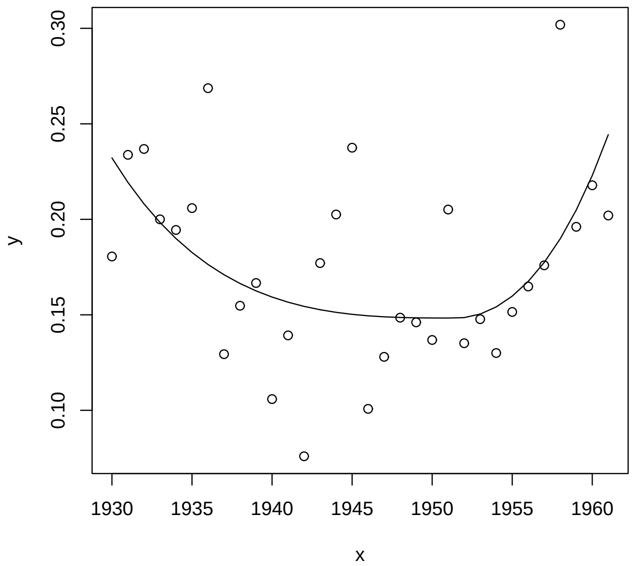


Figura 6.7: Figli 35 anni

```
x^3          -5.550e-05  1.898e-05  -2.923 0.003463 **
x^2:riforma    7.382e-03  2.097e-03   3.520 0.000431 ***
```

Esso prevede due soli parametri, ed ha un buon adattamento ai dati (vedi grafico 6.7).

```
Analysis of Deviance Table

Model 1: y ~ x^3 + x^2:riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      2868    2621.64
2      2839    2587.88   29    33.75      0.25
```

Nuovamente, il modello predefinito associa a *riforma* un parametro non significativo.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.739e+02  4.492e+01   3.871 0.000108 ***
x           -9.071e-02  2.322e-02  -3.907 9.34e-05 ***
```

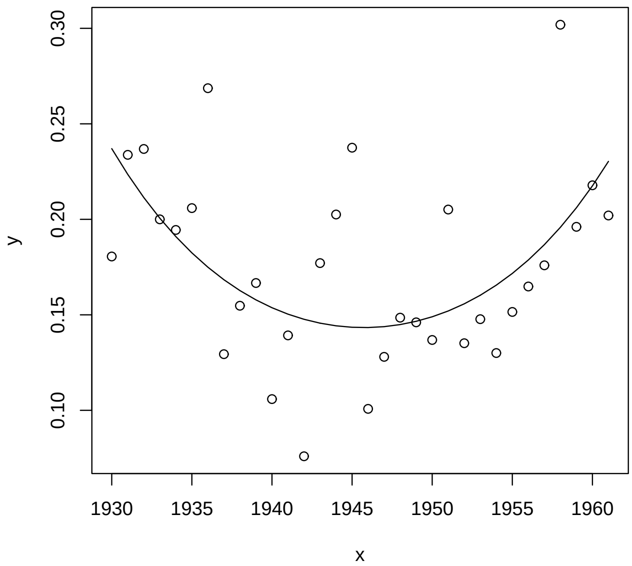


Figura 6.8: Figli 35 anni (modello quadratico)

```
x^2            2.973e-03  8.171e-04   3.638 0.000275 ***
riforma     -2.148e-01  2.350e-01  -0.914 0.360705
```

Ristimo, dunque, senza tale variabile (vedi grafico 6.8).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 160.231655  42.420072   3.777 0.000159 ***
x            -0.083629   0.021918  -3.816 0.000136 ***
x^2           0.002498   0.000631   3.958 7.54e-05 ***
```

Ancora una volta, pur non essendoci la netta inversione del *trend* riscontrata in precenza, l'andamento non pare assolutamente monotono, dunque non avrebbe senso procedere alla previsione basata sui valori fino alla coorte '48.

## 6.3 Numero di figli

Stimo un modello, ora, sul numero di figli per donna ai 35 anni. Si potrebbe trattare di una variabile di Poisson (vedi Appendice). Tuttavia, l'ipotesi di uguaglianza tra media e varianza (tipica di questa variabile) potrebbe essere


Figura 6.9: Figli per donna

troppo restrittiva. Infatti, spesso popolazioni eterogenee conducono a una sovradispersione (vedi McCullagh, Nelder, 1983). Per questo motivo, ristimo i dati usando la "quasipoisson": si lasciano, cioè, inalterate le stime, ma varia la significatività dei parametri, venendo meno il vincolo di uguaglianza tra media e varianza. I parametri stimati rappresenteranno, appunto, l'impatto delle variabili su $\lambda$, media, appunto, della Poisson. L'unico modello accettabile (che ha, tra l'altro, un *fit* elevatissimo) pare quello lineare (vedi grafico 6.9).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.630651   0.016896  37.326  < 2e-16 ***
x           -0.010225   0.001589  -6.435 1.24e-10 ***
Analysis of Deviance Table

Model 1: y ~ x
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      2494    1498.74
```

```
2      2464    1480.18   30    18.56      0.95
```

Valuto cosa accade se presento a R la variabile come appartenente alla famiglia *quasipoisson.*

```
Call:
glm(formula = y ~ x, family = quasipoisson)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.163696  -0.660181  -0.005613   0.206164   4.056235

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.63065    0.01276  49.445   <2e-16 ***
x           -0.01023    0.00120  -8.524   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.5698749)
```

Il parametro di dispersione adottato è invece bassissimo ( 0.5698749, indicato nell'ultima riga; dunque la varianza parrebbe addirittura minore della media). Vediamo, comunque, come non vari l'alta significatività di $x$. Nel modello predefinito, *riforma* risulta evidentemente non significativa.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.1365788 13.4958584  -0.455   0.6493
x            0.0035164  0.0069749   0.504   0.6142
x^2           -0.0004869  0.0002499  -1.948   0.0514 .
riforma      0.0597507  0.0671895   0.889   0.3738
```

Il fatto che, escludendo tale variabile, si ottengano comunque parametri non significativi, è dovuto alla forte vicinanza all'andamento lineare evidenziata prima (vedi grafico 6.10).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5669577 12.4620326  -0.126   0.8999
x            0.0011504  0.0064386   0.179   0.8582
x^2           -0.0003411  0.0001883  -1.811   0.0701 .
```

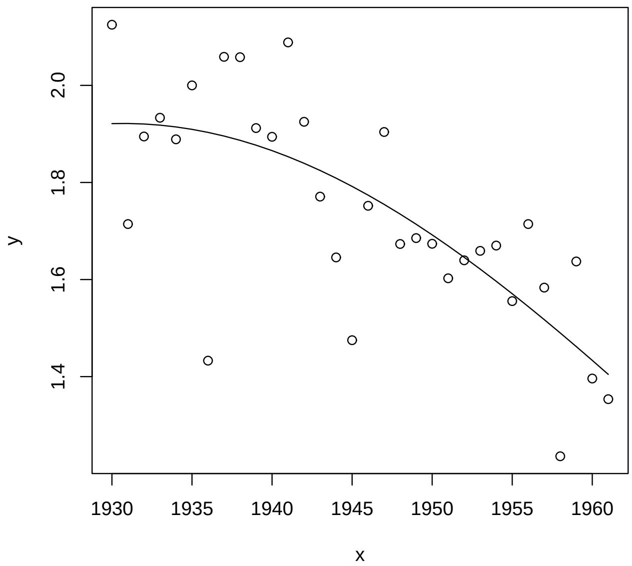


Figura 6.10: Figli per donna (modello quadratico)

La previsione, stavolta, appare invece sensata, dal momento che l'andamento decrescente appare conservarsi per tutte le coorti. Tuttavia, un paio di valori su dieci si trovano al di sotto delle bande di confidenza. Questo accade perchè il modello quadratico, in questo caso, accentua, anzichè attenuare, la tendenza dell'andamento lineare. Il grafico ( 6.11) si può quindi considerare questa un'ulteriore conferma dell'andamento lineare del numero medio di figli per donna entro i 35 anni.

## 6.4 Conclusioni

Abbiamo visto come, in relazione a nascita ed adozione di figli da parte delle donne, non ci sia alcuna evidenza di un impatto dell'istruzione sulle scelte relative a questa materia (in particolare, ho concentrato l'attenzione sulla nascita del primo figlio e il numero di figli al compimento del trentecinquesimo anno). La correlazione che normalmente si riscontra in questo campo, dunque, è verosimilmente attribuibile ad altri fattori (ad esempio, il *background* economico).

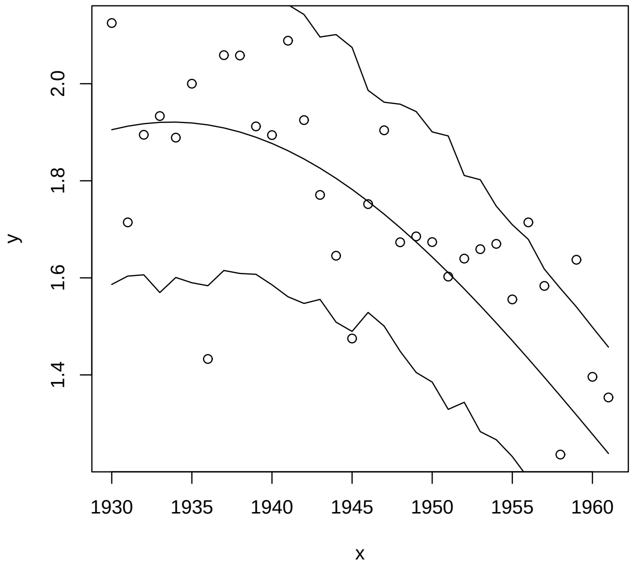


Figura 6.11: Figli per donna (modello previsivo)

# Capitolo 7

# Impatto dell'istruzione dei genitori su quella dei figli

L'ultima verifica della mia tesi voleva essere relativa all'eventuale impatto dell'istruzione dei genitori su quella dei figli. Agli intervistati viene domandato, infatti, anche l'anno di nascita dei genitori. E' dunque possibile effettuare confronti tra l'istruzione dei figli e la coorte di nascita dei genitori. Ovviamente, si possono verificare quattro tipo di impatto: i padri sui figli, le madri sui figli, i padri sulle figlie, le madri sulle figlie. La mia intenzione era verificare la potenziale discontinuità, come sempre, tra il '51 e il '52, ipotizzando che il salto riscontrato nell'intera coorte sia presente anche nel sotto gruppo dei genitori. Tuttavia, essendo gli intervistati più giovani nati nell' '81, non è possibile procedere troppo oltre con le date di nascita dei genitori. Considerando le coorti '36/'56 (in modo da avere solo figli che siano stati esposti alle riforme), si ha, comunque, non solo una ridotta numerosità campionaria (dunque una grande variabilità), ma anche una possibile distorsione. Infatti, i genitori nati tra il '52 e il '56 (nel periodo post-riforma, quindi), saranno pochi: avranno figli solo nelle ultime coorti intervistate, a patto che abbiano fatto figli in età via via più giovane, avvicinandosi al '56. In particolare, gli individui di quest'ultima classe saranno rappresentati solo a condizione di aver generato il primo figlio entro l'anno in cui essi ne compiono 25. E' plausibile che questo porti ad una distorsione sul loro livello d'istruzione rispetto all'insieme dei genitori di quella coorte (per esempio, difficilmente tali individui saranno laureati). Non ponendo limiti superiori all'anno di nascita degli intervistati, poi, sono presenti nel campione anche individui che hanno appena raggiunto i 18 anni. Valutando, come indicatore dell'istruzione dei figli, il conseguimento di un diploma di scuola superiore, avrei una distorsione verso il basso per le ultime coorti. Peraltro, servirsi del conseguimento del titolo di licenza media avrebbe poco senso, dal momento che la percentuale di individui che lo raggiungono, per i figli di genitori nati tra la fine degli anni '40 e l'inizio degli anni '50, è verosimilmente vicinissima al 100%. Pensando che, in realtà, anche valutando l'iscrizione ad una qualsiasi scuola superiore avrei ottenuto risultati estremamente elevati, ho concentrato l'attenzione sui soli iscritti a scuole di durata almeno quadriennale. Per circa il 5% degli individui, tuttavia, dalle risposte date non è stato possibile risalire al tipo di scuola superiore a cui essi si sono iscritti. Considerando, quindi, le iscrizioni sia per difetto che per eccesso, ho ottenuto otto coppie di valori (anno di nascita del genitore, iscrizione figlio). Per i problemi illustrati in precedenza, i modelli stimati appaiono del tutto non plausibili.

# Capitolo 8

# Conclusione

La mia tesi si propone di verificare il possibile impatto delle riforme del sistema scolastico sul conseguimento di istruzione, e di quest'ultimo sulle scelte relative ai figli. Le due riforme del sistema scolastico italiano in questione relative alla scuola media e all'università.
La prima di queste è la n.1859 del 31/12/1962, che stabilisce l'innalzamento dell'obbligo a otto anni, ossia fino al conseguimento della licenza media, con possibilità di proscioglimento dall'obbligo stesso a condizione di avere frequentato la scuola per otto anni, avendone compiuti 15. La stessa legge, peraltro, stabilì l'istituzione della "scuola media unificata", ponendo fine alla distinzione precedente tra "scuola media inferiore" e "scuola d'avviamento professionale".
La seconda riforma, invece, è quella della liberalizzazione degli accessi universitari (inclusa nei "Provvedimenti urgenti per l'Università", ossia nella legge n.910, dell' 11/12/1969). Questa legge dà la possibilità a tutti gli individui in possesso di un diploma di maturità di accedere a qualsiasi facoltà universitaria (in precedenza, questa prerogativa era solo dei diplomati del "Liceo Classico").
Una difficoltà relativa a queste riforme era quella di stabilire esattamente quali coorti ne venissero coinvolte. Si può ritenere, in sintesi, che le classi antecedenti al '49 non ne risentono in alcun modo, mentre quelle successive al '51 ne sono pienamente colpite. Valutando se ci siano stati impatti della riforma almeno sulle prime classi coinvolte (per far questo, mi servo del *"Regression discontinuity design"*, si può ritenere che, pur non essendo la legge del 1962 stata rispettata in maniera assoluta, essa pare comunque aver avuto un effetto sul conseguimento della licenza media per la coorte '52 (effetto che, verosimilmente, è poi scemato gradualmente) sia per gli uomini che per le donne. Per quest'ultime, probabilmente, è riscontrabile pure un effetto delle riforme sull'ottenimento del diploma di maturità. Relativamente alle donne,

comunque, si riscontrano anche effetti che, evidentemente, sono determinati unicamente dalla riforma universitaria. E' osservabile, infatti, un impatto sulle iscrizioni all'università (dovuto anche ad un incremento delle iscrizioni tra le diplomate per le coorti '49/'54). Queste classi, paiono, peraltro, risentire della riforma anche in termini di probabilità complessiva di conseguire la laurea, non avendo una *performance* peggiore delle diplomate più anziane (a differenza di quanto avverrà per le coorti successive). Per quanto riguarda il conseguimento di un titolo universitario da parte delle donne, dunque, si può pensare che l'impatto, anzichè calare gradualmente, sia forte (si passa dall'8% al 13%!) per le prime classi coinvolte, e si vanifichi a partire dalla coorte '55.

Volendo, poi, valutare se le donne risentano di questo incremento di istruzione dovuto alle riforme relativamente alle scelte relative alla procreazione ed adozione di figli, i dati non mostrano alcuna evidenza in questo senso. Per verificare, invece, se la nota correlazione, in Italia, tra istruzione dei genitori e quella dei figli sia dovuta a un impatto diretto della prima sulla seconda o sia da attribuire a fattori correlati ad entrambi, il fatto che il campione sia limitato ad individui nati non oltre il 1981 non consente di fornire dati a sufficienza per dare risposte statisticamente accettabili a questo quesito.

# Appendice

## Modelli statistici

Partiamo dal presupposto di avere a disposizione un insieme di variabili che chiameremo *esplicative* ($X$) e un gruppo di altre variabili che chiameremo *dipendenti* ($y$). L'assunto fondamentale è quello di considerare che vi sia una relazione che lega $X$ a $y$ e che dunque la variabile dipendente sia stata in qualche modo determinata da $X$. È tuttavia noto che le variabili dipendenti si distribuiscono secondo una determinata legge probabilistica: $y\ F_0$.

Tipicamente vi sono più modelli probabilistici ($F_0$) che si ritengono compatibili con $y$ e l'insieme di tali modelli è detto Modello Statistico e viene indicato con la lettera $\mathcal{F}$:

$$\mathcal{F} = F_0(\theta, \theta \in \Theta)$$

ove $\Theta$ è lo spazio parametrico.

Esistono due tipi di modelli statistici: i modelli parametrici e i modelli non parametrici. Un modello statistico si dice *parametrico* se nella precedente relazione vi è una corrispondenza biunivoca tra $\Theta \subseteq \mathbb{R}^n$ (con $n \in \mathbb{R}^n$) e $\mathcal{F}$; mentre se $\Theta$ non può essere posto in corrispondenza biunivoca il modello è un modello *non parametrico.*

Tipicamente, i modelli non parametrici si servono di stime locali, ed entrano in gioco quando i dati non paiono seguire un modello stabile. Essi possono anche essere utilizzati (come nella mia tesi) per sottoporre a *test* modelli parametrici, o come indirizzo per la scelta degli stessi. Nel caso degli anni d'istruzione, mi sono servito di un *test* sui residui di un modello lineare, per valutare l'ipotesi che essi abbiano media nulla contro quella che essi siano una funzione della variabile esplicativa (e di un parametro "di lisciamento","h", che faccio variare tra 0.1 e 10).

## Modello Lineare Generalizzato (GLM)

Un modello d'analisi è dunque un costrutto teorico scelto per rappresentare le proprietà e le relazioni assunte tra gli elementi che costituiscono i dati in esame.
Se queste relazioni sono di tipo lineare, il modello che si sceglie è un Modello Lineare del tipo:

$$y = \sum_j \beta_j x_j + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$$

ove $\varepsilon$ è il termine di errore indipendente con media 0 e varianza pari a $\sigma^2$. Dunque, è:

$$\mu - i = \sum_j \beta_j x_j \tag{8.1}$$

intendendo per $\mu - i$ il valore atteso di $y_i$ condizionatamente alle variabili esplicative. Nel caso dei GLM, invece, è:

$$\mu - i = g(\sum_j \beta_j x_j) \tag{8.2}$$

La differenza con il modello lineare classico, dunque, è data dal fatto che il modello abituale non consente di trovare il valore atteso condizionato della variabile dipendente, se non in maniera indiretta (vedi McCullagh e Nelder, 1990). La Poisson misura il numero di volte in cui si verifica un determinato evento in un dato lasso di tempo. Una *poisson* è così definita:

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{8.3}$$

, ove $x$ è l'insieme dei valori che può assumere la variabile (dunque, l'insieme dei numeri naturali, incluso lo 0). E' possibile, dunque, osservare come tale distribuzione sia interamente definita dal parametro $\lambda$. Quest'ultimo parametro, peraltro, coincide con la media e la varianza della distribuzione stessa. La "quasi poisson" invece, differisce dalla Poisson in quanto viene meno il vincolo di uguaglianza tra varianza e media Per quanto riguarda, invece, il modello binomiale, viene stimato il cosiddetto *logit*, o "regressione logistica". In pratica, chiamando $\eta$ il parametro stimato, è:

$$\mu_i = (e^{\eta_i})/(1 + e^{\eta_i}) \tag{8.4}$$

# Analisi dei dati di sopravvivenza

Come accennato nell'Introduzione, l'analisi di sopravvivenza dei dati ha come oggetto di interesse la probabilità di caduta (in genere, che si verifichi un dato evento puntuale) dopo una lunghezza temporale chiamata, appunto, "tempo di caduta", relativa a un individuo o a un gruppo di individui. I modelli relativi, in sostanza, stimano la probabilità che un dato evento si verifichi nel tempo $t$ dato che non è avvenuto in precedenza. In realtà, oltre a tale probabilità (condizionale), può essere calcolata quella marginale (ovvero la stessa probabilità, ma calcolata al tempo 0). Le due probabilità sono, in realtà, in corrispondenza biunivoca (la probabilità condizionale, infatti, non è altro che il rapporto tra la quella marginale e il complementare della funzione di ripartizione, a sua volta in relazione biunivoca con la probabilità stessa, come in tutte le variabili). Essendo la natura del tempo continua, tuttavia, usualmente si ragiona in termini di "densità". In particolare, i modelli di analisi di sopravvivenza dei dati derivano dalla distribuzione esponenziale, la cui densità si calcola come: $\rho e^{-\rho t}$, ove $\rho$ è la densità condizionale (costante). La distribuzione esponenziale è, dunque, un caso particolare tra i modelli di analisi di sopravvivenza, in cui la probabilità di caduta in un dato intervallo di tempo di durata $z$ (dato che la caduta non è avvenuta in precedenza) è costante (si parla di "assenza di memoria" del processo).

# Il criterio AIC

Il criterio di Akaike o criterio AIC, acronimo di *Asymptotic Information Criterion* è un criterio di applicabilità generale utilizzato sia nello studio delle serie storiche sia nell'analisi di regressione per evitare il rischio di sovraparametrizzazione (*overfitting*) di un modello. Esso è stato introdotto, appunto, da Akaike(1973). Un buon adattamento ai dati disponibili da parte di un modello poco parsimonioso, infatti, non corrisponde necessariamente ad una buona capacità previsiva del modello stesso.

Per ovviare a questo inconveniente il criterio AIC, che rappresenta una stima dell'indice di distanza tra modelli di Kullback-Leibler, come molti altri criteri basati sullo stesso principio assegna un 'costo' all'introduzione di ogni parametro addizionale. Questa la formulazione (Sakamoto, Ishiguro e Kitagawa, 1986) della quantità di informazione del criterio AIC, che dovrà essere minimizzata utilizzando nelle stime un appropriato numero di parametri:

$$AIC(k) = -2\left(\log L(\hat{\delta}) - k\right)$$

$k$ rappresenta il numero di parametri del modello , $\hat{\delta}$ è il vettore ($k \times 1$) contenente i parametri stimati e $L(\hat{\delta})$ è la funzione di verosimiglianza calcolata in $\hat{\delta}$ ed in ipotesi di gaussianità del *white noise*.

La quantità $k$ rappresenta una sorta di 'penalità', legata al numero di parametri del modello. Essa non varia al crescere di $n$, così, per una numerosità campionaria elevata, viene attribuita molta più importanza alla verosimiglianza che al numero di parametri utilizzati; è dunque riscontrabile una certa tendenza alla sovraparametrizzazione da parte di questo criterio.

# L'output di R

Nel corso di questo studio sono stati riportati i dati tratti dalle elaborazioni statistiche fornite dal programma R. Le prossime sezioni spiegheranno nel dettaglio il significato di ciascuna voce mostrata in *output*.

## Il modello lineare generalizzato

Per l'analisi dei risultati per il modello lineare generalizzato R dispone di due comandi: summary(x) e anova(x,y,test="Chisq"), che si riferiscono al modello x e che sono utilizzati rispettivamente per il riepilogo dei parametri stimati e per l'analisi dei residui, operando un confronto tra il modello $x$ il modello $y$, che ha senso quando il secondo modello contiene gli stessi parametri liberi del primo, più altri di cui valuto la significatività. Vediamo un esempio del primo comando:

```
Call:
glm(formula = y ~ x + z, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7222  -1.1835   0.7745   0.9959   1.4314

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.269e+00  1.475e-01   8.605  < 2e-16 ***
x           8.771e-02  3.136e-02   2.797  0.00516 **
z           7.921e-05  1.419e-03   0.056  0.95550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

```
    Null deviance: 2398.9  on 1775  degrees of freedom
Residual deviance: 2282.0  on 1773  degrees of freedom
AIC: 2288.0

Number of Fisher Scoring iterations: 4
```

Il primo blocco richiama la famiglia distributiva della variabile risposta e le variabili esplicative del modello, mentre il secondo illustra alcuni indici di posizione dei residui stimati (minimo, massimo, mediana, primo e terzo quartile). Il terzo blocco, invece, oltre a fornire una stima per i parametri che compongono il modello (intercetta e coefficienti delle variabili esplicative) aggiunge il valore dell'errore standard (seconda colonna), della statistica t di Student (terza colonna) e dell'$\alpha$ osservato (o p-value) nel test t. Quest'ultima colonna è importante per stabilire la significatività di ogni singola variabile nella spiegazione del modello: nel caso in cui tale livello fosse superiore a quello fissato in precedenza e ritenuto "soglia di sicurezza" per la bontà della nostra affermazione (ad esempio lo 0.05) si potrà ritenere trascurabile l'apporto della variabile esplicativa nella determinazione della dipendente.

Il quarto blocco riguarda la bontà del modello nel suo complesso: la prima riga si riferisce alla variabilità complessiva dei dati, la seconda a quella dei residui del modello, la terza ne indica il criterio di Akaike. In quest'ultimo blocco, l'*output* di R differisce rispetto al caso del modello lineare classico: eccone un esempio.

```
Call:
lm(formula = y ~ x, weights = freq./1870)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17059 -0.07560 -0.01940  0.09314  0.16910

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.66044    0.20557   46.99  < 2e-16 ***
x            0.21313    0.01711   12.46 7.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1068 on 20 degrees of freedom
Multiple R-Squared: 0.8858,     Adjusted R-squared: 0.8801
```

```
F-statistic: 155.1 on 1 and 20 DF,  p-value: 7.022e-11
```

In questo caso, infatti, si illustra, nella prima riga, l'errore standard del modello; nella seconda il coefficiente di determinazione multipla normale e corretto (ossia, $R^2$ e $R^2$ aggiustato), la terza il risultato del test F di Snedecor (con indicati il numero di parametri stimati e i gradi di libertà), utilizzato per stabilire la significatività dell'incremento di varianza spiegata dal modello prescelto: ancora, si tratta di valutare se tale valore è inferiore al livello di significatività prescelto (in caso contrario, il modello è assolutamente non esplicativo).

È da precisare che la potenza di calcolo di R non è adatta a calcolare cifre inferiori a $2 \cdot 10^{-16}$. Un $\alpha$ osservato inferiore sarà perciò riportato come `<2e-16`, valore comunque adeguato ai fini del nostro calcolo e delle elaborazioni statistiche in genere.

Passiamo ora ad osservare, tornando al modello lineare generalizzato, un esempio di tabella di analisi della varianza (ANOVA), utile a fornire informazioni sulla significatività complessiva di un gruppo di parametri aggiuntivi:

```
 anova(mod.2,mod.A,test="Chisq")
Analysis of Deviance Table

Model 1: y ~ x + riforma
Model 2: y ~ factor(x)
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      3732     3367.4
2      3693     3322.0   39     45.5       0.2
```

Stavolta, R, dove aver richiamato l'*input* e i modelli da confrontare (prime quattro righe), ne illustra i residui e il numero di gradi di libertà complessivi (prime due colonne dell'ultimo blocco). L'ultima riga, poi, presenta tre ulteriori colonne, indicanti il numero di vincoli in più del modello più parsimonioso e la differenza nella variabilità non spiegata. L'ultima colonna, infine, mostra il p-value del test di significatività dei parametri aggiuntivi del modello con più parametri liberi.

# Bibliografia


[1] G. Becker (1993), *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, Chicago, University of Chicago Press (prima edizione: 1975).

[2] J. D. Angrist e W. Newey (1991), "Over-identification Test in Earnings Function with Fixed Effects", *Journal of Business and Economic Statistics, Luglio, pagine 317-323.*

[3] L. Flabbi (1999), *Return of schooling in Italy: OLS, IV and Gender Differences*, Working paper "Università Bocconi".

[4] M. Pisati (2000), *La mobilità sociale*, Bologna, Il Mulino.

[5] A. Schizzerotto (2002), *Vite ineguali*, Bologna, Il Mulino.

[6] G.Cives (1990), *La scuola italiana dall'Unità ai nostri giorni*, Firenze, La Nuova Italia.

[7] D. Card (1999), "The Causal Effect of Education on Earnings", *Handbook of Labor Economics*, Amsterdam e New York, Ashenfelter e Card.

[8] D. Card (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems", *Econometrica*, vol.69, numero 5.

[9] D.R. Cox, D. Oakes (1984) *Analysis of Survival data*, Londra e New York, Chapman e Hall.

[10] R.J. Bowden, D.A. Turkington (1984), *Instrumental variables*, Cambridge, Cambridge University Press.

[11] R.J. Willis, S.Rosen (1979), "Education and Self-selection",*Journal of Political Economy"*, Vol.87, numero 5.

[12] G.Brunello, R. Miniaci (1999), "The economic Returns to Schooling for Italian men. An evaluation Based on Instrumantal Variables", *Labour Economics* 6.

[13] G. Brunello, S. Comi, C. Lucifora, *The Returns to Education in Italy: a New Look at the Evidence*, Fondazione Enrico Mattei, Working Paper 99-101.

[14] J. Hahn, P. Todd, W. Van der Klaw (2001), "Identification and Estimation of tratment effects with a regression discontinuty design", *Econometrica*, 69:201-209.

[15] P. Holland (1986), "Statistic and causal inference", *Journal of the American Statistical Association*

[16] J.J. Heckmann, H. Ichimura, P. Todd (1998), "Matching as an econometric evaluator estimator" (1998), *Review of Economic Studies*, 65: 261-294.

[17] P.R. Rosenbaum, D.B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects" *Biometrica*, 70(1): 41-55.

[18] D.L. Thistlewaite, D.B. Campbell (1960): "Regression discontinuity analysis: an alternative to the ex post facto experiment", *Journal of Educational Psichology*, Volume 51, numero 6.

[19] T.D. Cook, D.T. Campbell (1979), *Quasi experimentation. Design and analysis issues for fields settings*, Boston, Hogton Mifflin Company.

[20] D.B. Rubin (1977), "Assignment to treatment group on the basis of a covariate", *Journal of Educational Statistics*, 2: 4-58

[21] S.M. Iacus, G.Masarotto (2003), *Laboratorio di statistica con R*, Milano, McGraw-Hill.

[22] A. Checchi (1997), *Il sistema formativo in Italia*, Fondazione Di Vittorio.

[23] A. W. Bowman, A. Azzalini (1997), *Applied smoothing techniques for Data Analysis* Oxford, Oxford University Press

[24] A.Brandolini, P.Cipollone (2002), *Return to education in Italy: 1992-1997*, manoscritto.

[25] J.P. Klein, M.L. Moeschberger (1997) , *Survival analysis: techniques for censored and truncated data*, New York, Springer.

[26] P. McCullagh, J. A. Nelder (1983), *Generalized linear models*, Londra e New York, Chapman e Hall.

[27] Y. Sakamoto, M.Ishiguro, G.Kitagawa (1984), *Akaike Information Criterion Statistics*, Tokyo, D.Reidel Publishing Company.