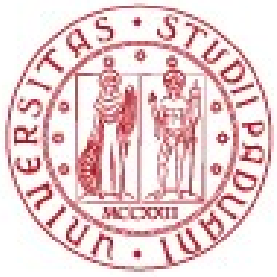
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche



Modelli statistici per la valutazione dell'efficacia e l'ottimizzazione dell'automatizzazione delle campagne di marketing tramite disegno di regressione con discontinuità

Relatore Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Correlatore Dott. Angelo Basile
Alkemy S.p.A

Laureando Gabriele Massaro
Matricola 2024562

Anno Accademico 2022/2023

Indice

Introduzione

L'efficacia di una campagna di marketing è un aspetto cruciale per il successo di un'azienda. Tuttavia, misurare l'impatto di una campagna di marketing può essere un compito arduo a causa della molteplicità di fattori che possono influenzare le scelte dei consumatori. In questo contesto, il *Regression Discontinuity Design* (RDD) è uno strumento sempre più utilizzato per analizzare gli effetti di una campagna di marketing. Il *Regression Discontinuity Design* è un metodo di valutazione che si basa su una discontinuità nei dati per identificare gli effetti causali di un trattamento o come nel caso di questa trattazione, di una campagna. In questa relazione, verrà presentato un caso di studio in cui viene applicata la metodologia per valutare l'efficacia di una campagna di marketing effettuata per uno specifico prodotto di investimento. Saranno discusse le metodologie utilizzate per l'analisi dei dati e i risultati ottenuti. L'obiettivo della relazione è mostrare come il metodo possa essere un valido strumento per valutare l'efficacia o meno di una campagna di marketing, fornendo informazioni preziose per i professionisti del settore.

Nel Capitolo 1 sarà introdotto il metodo e verrà descritta l'idea che ne sta alla base. Successivamente verranno definite le diverse tipologie del disegno ed il loro funzionamento.

Nel Capitolo 2 verranno presentate le classiche analisi grafiche di *Regression Discontinuity Design*. Vengono inoltre illustrati brevemente i metodi applicati per la loro rappresentazione.

Nel Capitolo 3 si descriverà accuratamente la procedura di stima del metodo, si darà una descrizione generale dei principali ingredienti e si forniranno le metodologie utilizzare per la scelta del valore ottimale di questi.

Nel Capitolo 4 si fornirà una descrizione minuziosa delle procedure inferenziali e di convalida del *Regression Discontinuity Design*. In particolare verranno discusse le problematiche legate all'inferenza convenzionale e verranno descritte metodologie *ad hoc* per il disegno. Vengono inoltre descritti alcuni test utili alla verifica della validità del

metodo.
Nel Capitolo 5 verranno eseguiti vari passi. In primo luogo si adatteranno dei modelli non parametrici univariati che considereranno un'unica variabile esplicativa. Successivamente verranno adattati altri modelli nei quali la nuova variabile risposta utilizzata saranno i residui di una precedente non regressione non parametrica adattata sulla vecchia risposta utilizzando due covariate differenti. Si adatteranno poi quei modelli per i residui utilizzando una nuova covariata, al fine di isolare gli effetti di quest'ultima. Verrà poi applicata una seconda metodologia concernente la metodologia in esame e in ultima istanza verranno effettuati test per la validità del disegno.
Nel Capitolo 6 si valuteranno i risultati ottenuti e si forniranno test per valutare la significatività delle stime ottenute. Verrà effettuato inoltre un confronto con un diverso valore di un elemento chiave della metodologia. Lo studio si conclude con una riflessione generale sui risultati ottenuti e con delle proposte mirate a fornire possibili miglioramenti.

Capitolo 1

Regression Discontinuity Design

1.1 Introduzione

Tipicamente, nei disegni sperimentali, ossia quegli studi nei quali si vuole valutare l'efficacia di un trattamento o di una campagna, vi è la necessità di disporre di due gruppi, uno nel quale alle unità che ne fanno parte viene assegnato il trattamento, detto gruppo trattamento ed uno nel quale questo non viene assegnato, detto gruppo di controllo (van Leeuwen et al., 2018). Generalmente i due gruppi differiscono solamente per lo status di trattamento, per il resto le unità presentano le stesse condizioni e caratteristiche. Il campo nel quale vi è una maggiore applicazione di studi sperimentali è senz'altro quello della medicina, nel quale solitamente si è interessati a valutare l'efficacia di determinati trattamenti o farmaci. Solitamente la divisione nei due gruppi viene effettuata in modo casuale, ma talvolta effettuare una *randomizzazione* non è utile per raggiungere gli obiettivi preposti. Un esempio sono le campagne di marketing. Un metodo utile in circostanze come quella menzionata precedentemente, poiché consente di identificare in modo accurato gli effetti delle campagne senza che le unità vengano assegnate in maniera casuale ai due gruppi, è il *Regression Discontinuity Design*. Il metodo, introdotto per la prima volta nel 1960 da Donald L. e Thistlethwaite, è un approccio rigoroso non sperimentale che può essere utilizzato per stimare gli effetti di una campagna o di un trattamento, tramite l'applicazione di un meccanismo di assegnazione del trattamento basato sul valore assunto da una variabile di punteggio, che ha una distribuzione continua (Thistlethwaite & Campbell, 1960). Il meccanismo di assegnazione del trattamento o della campagna consiste generalmente nella seguente procedura: si identifica una variabile quantitativa, detta variabile di punteggio o *score*, e successivamente si fissa un punto di taglio designato. Se un'unità presenta valori dello

score maggiori o uguali al *cutoff* verrà assegnata al gruppo di coloro che riceveranno la campagna o il trattamento, viceversa verrà assegnata al gruppo di controllo (Trochim, 1990).
Per comprendere meglio quanto detto facciamo un esempio molto semplice: assumiamo che il trattamento in questo caso sia un'assegnazione di premi per merito in base ad un punteggio ottenuto ad una prova scritta. Si vuole quindi studiare l'impatto dei premi per merito sui futuri risultati accademici. Il *Regression Discontinuity Design* opera in questa maniera: posto il *cutoff*, ad esempio, ad un valore pari a 6, a tutti gli studenti che otterrano una votazione maggiore o uguale a 6 verrà erogato il premio per merito, viceversa per coloro che hanno ottenuto un punteggio inferiore al valore della soglia, questo non verrà assegnato. Successivamente, si andranno a confrontare i due gruppi in termini di prestazioni raggiunte, per valutare se l'azione effettuata ha avuto o meno un effetto e nel caso in cui questo sia presente, valutarne la direzione e la magnitudine. In questo modo, quindi, si verranno a creare i due gruppi, trattamento e controllo. L'idea di Thistlethwaite & Campbell (1960) era di confrontare quelle unità che avevano ottenuto punteggi leggermente al di sotto e al di sopra della soglia, poichè ritenute unità molto simili tra loro se non per lo status di trattamento, ossia l'aver ricevuto o meno un premio.
Un ulteriore esempio utile a comprendere la logica sottostante il metodo potrebbe essere quello di assegnare un determinato farmaco a coloro che presentano un valore corporeo, come ad esempio la pressione sanguigna, oltre un certo valore. Il criterio del punto di taglio fa quindi sì che un programma o trattamento venga assegnato a coloro che più ne hanno bisogno o lo meritano maggiormente.
Le principali componenti di un disegno di *Regression Discontinuity Design*, senza le quali la metodologia non potrebbe essere applicata, sono:

- lo *score*, o variabile di punteggio;
- il *cutoff*, valore che determina la divisione in gruppo di trattamento e gruppo di controllo;
- lo *status* di trattamento, variabile che indica se il trattamento è stato assegnato o meno.

La variabile *score* può essere qualsiasi variabile continua misurata prima del trattamento. Per studiare gli effetti causali tramite metodologia RDD, non solo le tre quantità devono esistere ed essere ben definite, ma la relazione tra queste deve soddisfare delle condizioni, obiettive e verificabili.

In Figura 1.1 viene mostrato un grafico attraverso il quale si vuole evidenziare la logica ed il funzionamento del *Regression Discontinuity Design*.

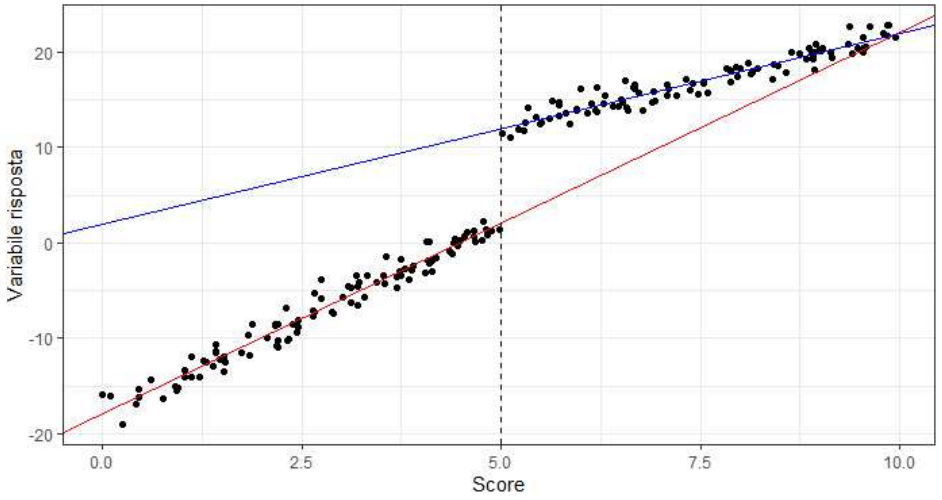


FIGURA 1.1: Logica del disegno

Si può osservare un salto verso l'alto in corrispondenza della soglia. Ciò significa che, in un intorno del *cutoff*, le unità a cui è stata assegnata la campagna o il trattamento hanno ottenuto valori della variabile risposta maggiori rispetto a coloro che invece sono stati assegnati al gruppo dei non trattati. La dimensione del salto che si verifica in corrispondenza della soglia, ossia la distanza tra le due curve al *cutoff*, è, in questo caso, una misura diretta dell'effetto causale del trattamento sulla variabile risposta (Maciejewski & Basu, 2020). Una grande differenza tra il *Regression Discontinuity Design* e gli studi sperimentali è che, in quest'ultimi, l'assegnazione al gruppo di trattamento o controllo viene fatta *ex ante* e le unità vengono semplicemente assegnate dal ricercatore, di solito in modo randomizzato, ad uno dei due gruppi. Nel disegno RDD, invece, l'assegnazione viene fatta *ex post* (Cook & Wong, 2008). Il metodo di assegnazione dei soggetti differenzia quindi il *Regression Discontinuity Design* da tutti gli altri studi randomizzati e da tutti gli altri studi quasi-sperimentali.

Le tre componenti del metodo - *score*, *cutoff* e status di trattamento - definiscono il disegno in generale e identificano la sua caratteristica più importante: nel metodo, a differenza di altri studi non sperimentali, l'assegnazione del trattamento segue una regola nota e, quindi, verificabile empiricamente. L'uso di questa metodologia è ancora basso e ciò può essere attribuito a più fattori. Il primo concerne certamente il fatto che il metodo è stato ideato per la prima volta nel 1958 senza essere mai testato fino a metà del 1970. Una seconda ragione risiede nel fatto che questo, in molti casi, non può essere utilizzato a causa dell'impossibilità di soddisfare uno o più dei suoi criteri chiave, ossia

il determinare dei punteggi-soglia tramite indicatori quantitativi. Uno dei vantaggi del disegno risiede nel fatto che questo rispetta i canoni etici, in particolar modo l'assegnazione di un trattamento o programma a chi ne ha maggiormente bisogno. Dal punto di vista metodologico, un punto a favore della metodologia RDD consiste nel fatto che l'inferenza tratta da un disegno ben implementato è confrontabile, in termini di validità interna, alle conclusioni tratte da esperimenti randomizzati. Il metodo non è, quindi, semplicemente un'ulteriore strategia per valutare l'efficacia di un trattamento, ma addirittura l'inferenza causale derivante da questa metodologia sembra essere potenzialmente più credibile di quella fornita dai classici esperimenti. Da qui scaturisce il vasto uso in tempi recenti del *Regression Discontinuity Design* (Cook & Wong, 2008). Il campo in cui si riscontra maggiore applicazione di tale metodologia è certamente la medicina, a causa dell'abbondanza di dati a disposizione e di indicatori quantitativi, quest'ultimi tali da permettere di mettere in atto il criterio di selezione tramite punteggio-soglia, ma la metodologia viene sempre più spesso applicata in altri campi di ricerca come ad esempio l'economia, la politica pubblica o la psicologia.

1.2 Tipologie di *Regression Discontinuity* e notazione adottata

La letteratura a disposizione ci fornisce due tipologie diverse di *Regression Discontinuity Design*: *Sharp RD* e *Fuzzy RD*. Il *Fuzzy RD*, a sua volta, si divide in: *Fuzzy RD no-shows* e *Fuzzy RD no-shows and crossovers*. Nel prosieguo di questa trattazione assumeremo la seguente notazione:

- Y : risultato (*outcome* ottenuto), cioè la variabile risposta;
- $Y_i(1)$: *outcome* osservato per le unità appartenenti al gruppo trattamento;
- $Y_i(0)$: *outcome* osservato per le unità appartenenti al gruppo di controllo;
- W : status di trattamento;
- X : variabile di punteggio o *score*;
- Z : status di superamento della soglia;
- c : soglia o *cutoff*.

1.2.1 Modello *Sharp RD*

Consideriamo il modello per l'i-esima osservazione:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

e definiamo

$$W_i = \begin{cases} 1 & se \quad X_i \geq c \\ 0 & se \quad X_i < c \end{cases}$$

in modo tale che la ricezione di un trattamento W_i è determinata da una certa soglia o *cutoff* c di una variabile continua X_i (Angrist & Pischke, 2009). Il termine u_i rappresenta la componente erratica del modello. Il modello precedentemente specificato è detto *Sharp RD* poichè l'assegnazione del trattamento è una funzione deterministica e discontinua al *cutoff*: tutte le osservazioni con $X_i < c$ non ricevono il trattamento, al contrario di tutte le osservazioni con $X_i \geq c$.
In Figura 1.2 si mostra la divisione in gruppo di trattamento e controllo tramite modello *Sharp RD*.

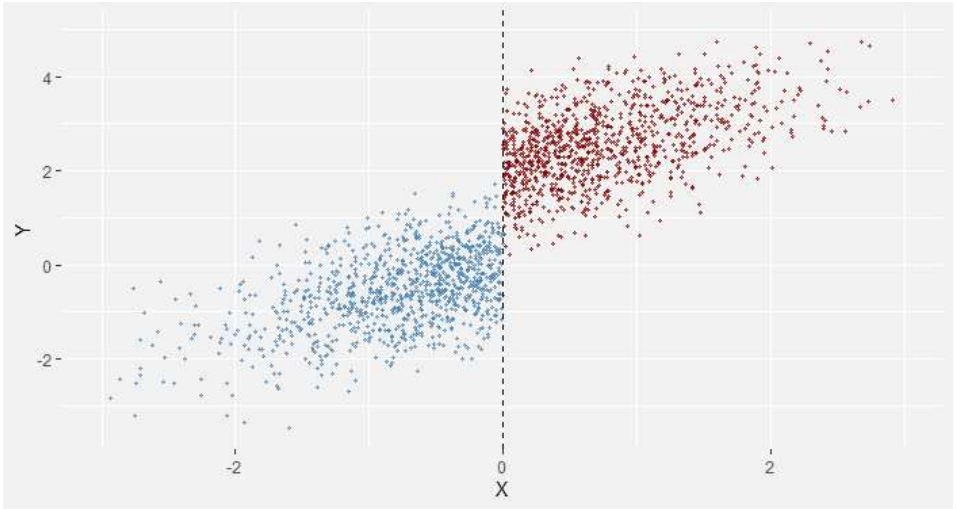


FIGURA 1.2: Divisione in trattamento-controllo: Sharp RD

Dal grafico si nota come vi sia una perfetta divisione nei due gruppi trattamento e controllo, a sinistra della soglia vi sono solo osservazioni blu (unità alle quali non viene assegnato il trattamento), mentre alla destra di questa vi sono solo osservazioni rosse (unità alle quali il trattamento è assegnato). Si ha quindi che la probabilità di ricevere il trattamento, condizionatamente al fatto che si sia superata la soglia, è pari a 1, mentre è 0 se ci si condiziona al fatto che non si sia raggiunto il *cutoff*.

In Figura 1.3 si mostra, quindi, la probabilità condizionata di ricezione del trattamento.

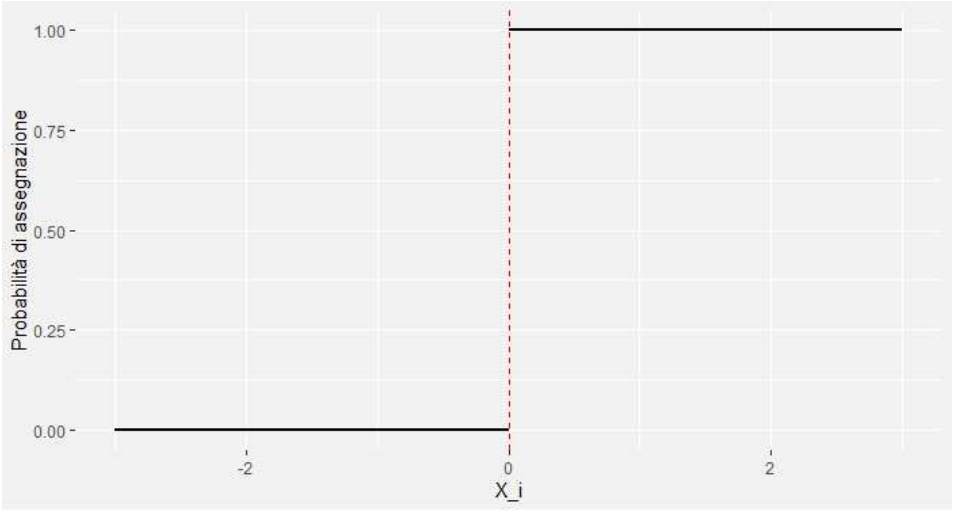


FIGURA 1.3: Sharp RD: Probabilità condizionata di ricevere il trattamento

Il grafico mostra quanto appena detto: fissata la soglia pari a 0, si ha che la probabilità di aver assegnato il trattamento è pari a 0 per tutte le unità con variabile di *score* minore di zero, mentre questa è pari a 1 per valori maggiori o uguali a quello della soglia. Nel disegno *Sharp Regression Discontinuity* ci si concentra quindi sulla discontinuità nel valore atteso condizionato per determinare un effetto causale medio del trattamento:

$$\lim_{x \downarrow c} \mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i \mid X_i = x\right]$$

che viene interpretato come l'effetto causale medio del trattamento nel punto di discontinuità:

$$\tau_{\text{SRD}} = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right].$$

Riprenderemo successivamente questa formulazione dopo aver accennato ulteriori concetti base.

Un ulteriore grafico che può essere utile per comprendere la logica sottostante il funzionamento del metodo *Sharp* viene mostrato in Figura 1.4.

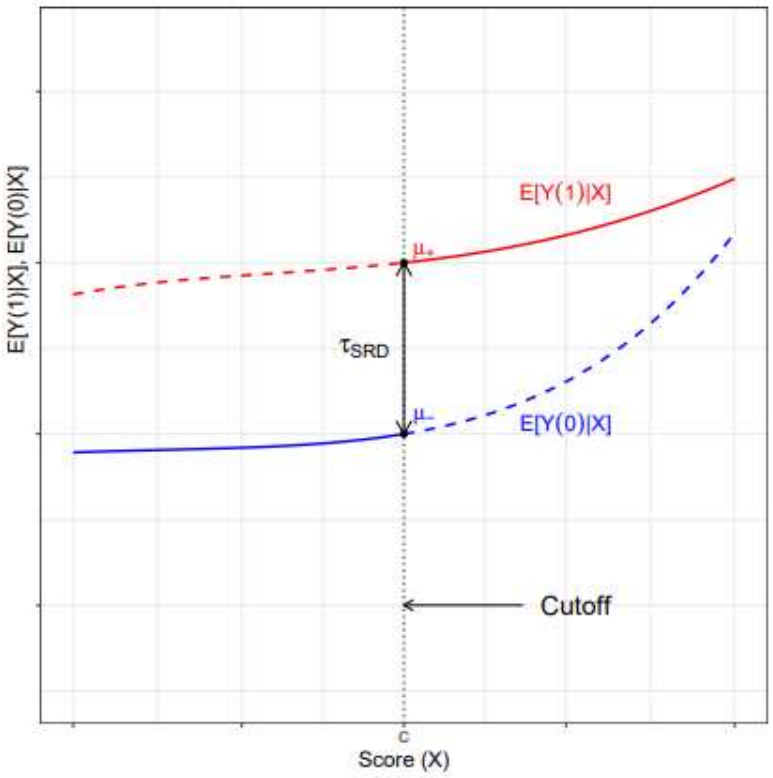


FIGURA 1.4: $E[Y(1)|X]$ vs X e $E[Y(0)|X]$ vs X

Il grafico mostra i valori attesi dei potenziali *outcome* condizionatamente al valore di X_i nell'asse delle ordinate, mentre il valore della variabile di *score* nelle ascisse.
Si può notare che:

- $\mathbb{E}\left[Y_i(1) \mid X_i = x\right]$ non viene osservato alla sinistra della soglia;
- $\mathbb{E}\left[Y_i(0) \mid X_i = x\right]$ non viene osservato per valori di X_i maggiori o uguali al *cutoff*.

Quindi, la media dell'*outcome* osservato, dato il valore X_i, è la seguente:

$$\mathbb{E}\left[Y_i \mid X_i\right] = \begin{cases} \mathbb{E}\left[Y_i(0) \mid X_i\right] & \text{se } X_i < c \\ \mathbb{E}\left[Y_i(1) \mid X_i\right] & \text{se } X_i \geq c \end{cases}$$

Abbiamo assunto precedentemente che l'*outcome* può assumere due possibili valori: $Y_i(1)$, ossia l'*outcome* che si otterrebbe nel gruppo del trattamento, e $Y_i(0)$, ossia l'*outcome* che si otterrebbe appartenendo al gruppo di controllo. Queste due variabili vengono chiamate *outcome potenziali* poiché, nonostante la variabile Y possa assumere

entrambi i valori, se ne può osservare solamente uno dei due (Imbens & Rubin, 2015). Se si osserva $Y_i(1)$ allora $Y_i(0)$ rimane latente e viceversa. Il fatto di non poter mai riuscire ad osservarle congiuntamente è il principale problema dell'inferenza causale ed è quindi sconosciuto l'effetto del trattamento a livello individuale.
In formule, l'*outcome* viene espresso come:

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{se } X_i < c \\ Y_i(1) & \text{se } X_i \geq c \end{cases}$$

Questa formulazione mostra un caso estremo di mancanza di un supporto comune: le unità nel gruppo di trattamento e nel gruppo di controllo non possono assumere lo stesso valore di X_i. Come detto precedentemente, l'effetto medio del trattamento, condizionatamente al valore della variabile di punteggio, è dato dalla distanza verticale tra le due curve di regressione ad x. Il problema principale è che questa distanza non può essere direttamente stimata a causa del fatto che le due curve non potranno mai essere osservate contemporaneamente per il medesimo valore di x. Per risolvere questo problema, immaginiamo di disporre di unità con punteggio esattamente uguale al valore di c ed unità con punteggio leggermente inferiore a questo, ossia con un punteggio pari a $c-\epsilon$ dove $\epsilon > 0$ piccolissimo. Così facendo, si assune che le unità con $X_i = c$ e $X_i = c-\epsilon$ siano praticamnte identiche, eccezion fatta che per lo status di trattamento. In questo modo si potrebbe calcolare la distanza tra le due curve di regressione in corrispondenza della soglia utilizzando l'*outcome* osservato. Nel grafico precedente, la distanza verticale in corrispondenza di $Xi = c$ è così formulata:
$\mathbb{E}\left[Y_i(1) \mid X_i = c\right] - \mathbb{E}\left[Y_i(0) \mid X_i = c\right] \equiv \mu_+ - \mu_-$,
ossia l'effetto del trattamento stimato tramite *Sharp Regression Discontinuity.*
In precedenza abbiamo definito formalmente l'effetto del trattamento come:

$$\tau_{\text{SRD}} \equiv \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right]$$

Questo misura l'effetto del trattamento per quelle unità con $X_i = c$ e risponde, quindi, alla seguente domanda: 'quale sarebbe il cambiamento dell'*outcome* medio per le unità con punteggio $X_i = c$ se cambiassimo il loro status di trattamento e si assegnasse il trattamento?'.
Questo parametro è, per costruzione, di natura locale e quindi in assenza di ulteriori assunzioni non è informativo per quanto concerne gli effetti del trattamento in corrispondenza di punti diversi dalla soglia. L'assunzione di comparabilità tra unità con punteggio molto simile ma a lati diversi dalla soglia è il concetto fondamentale su cui si basa il

Regression Discontinuity Design. L'idea fu formalizzata da Hahn et al. (2001) tramite l'utilizzo di assunzioni di continuità. Fu dimostrato che se le funzioni di regressione, viste come funzioni di x, sono continue ad $x = c$, allora si ha che:

$$\mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right] = \lim_{x \downarrow c} \mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i \mid X_i = x\right].$$

Il concetto di discontinuità nel contesto dei disegni di *Regression Discontinuity* significa che man mano che il valore di x si avvicina sempre più al valore del *cutoff* c, le funzioni delle medie dei potenziali *outcome*, $\mathbb{E}\left[Y_i(0) \mid X_i = x\right]$ e $\mathbb{E}\left[Y_i(1) \mid X_i = x\right]$, si avvicinano sempre di più al loro valore al punto di taglio, $\mathbb{E}\left[Y_i(0) \mid X_i = c\right]$ e $\mathbb{E}\left[Y_i(1) \mid X_i = c\right]$. In questo modo, la continuità giustifica la maniera di operare del disegno *Sharp RD* descritta precedentemente, ossia il concentrarsi sulle osservazioni appena superiori ed inferiori alla soglia in un piccolo intorno di quest'ultima. In generale, vi sono due assunzioni chiave per quanto riguarda i *matching-type estimators*, tra cui è compreso il *Regression Discontinuity Design*.
La prima, detta *unconfoundedness assumption*, implica che

$$Y_i(0), Y_i(1) \perp W_i \mid X_i$$

ossia che gli *outcome* potenziali siano indipendenti dalla modalità che assume lo status di trattamento, condizionatamente al valore i-esimo dello *score*. Questa viene sempre soddisfatta perchè, condizionatamente alle covariate, non vi è variazione nel trattamento.
La seconda ipotesi, detta *overlap assumption*, invece implica che

$$0 < \Pr\left(W_i = 1 \mid X_i = x\right) < 1$$

Quest'ultima viene violata perchè per ogni valore di X_i la probabilità di ricevere il trattamento non è mai compresa tra 0 e 1, poiché come si è visto in precedenza, il metodo assegna solamente probabilità pari ad 1 di ricevere il trattamento se il valore della variabile *score*ha superato la soglia, oppure probabilità pari a 0 se ciò non si è verificato. Data l'impossibilità di soddisfare le assunzioni alla base dei classici *matching-type estimators*, focalizziamoci nuovamente sull'effetto causale medio quando $X = c$:

$$\tau_{\text{SRD}} = \mathbb{E}[Y(1) - Y(0) \mid X = c] = \mathbb{E}[Y(1) \mid X = c] - \mathbb{E}[Y(0) \mid X = c]$$

Secondo il modello Sharp RD, non ci sono unità con $X = c$ per le quali osserviamo $Y_i(0)$ e sfrutteremo quindi il fatto che osserviamo unità con valori della variabile di punteggio

arbitrariamente vicini a c.

Per giustificare quanto detto in precedenza facciamo un'ipotesi di lisciamento, tipicamente formulata in termini di valori attesi condizionati (Lee & Lemieux, 2010):

Assunzione 2.1 (Continuità della funzione di regressione condizionata)

$$\mathbb{E}[Y(0) \mid X = x] \quad , \mathbb{E}[Y(1) \mid X = x],$$

sono continue in x.

Più generalmente, si potrebbe assumere che la funzione di distribuzione condizionata è continua. Data la funzione di distribuzione condizionata, $F_{Y(w)|X}(y \mid x) = \Pr(Y(w) \leqslant y \mid X = x)$, la versione generalizzata dell'assunzione diventa:

Assunzione 2.2 (Continuità della funzione di distribuzione condizionata)

$$F_{Y(0)|X}(y \mid x) \text{ e } F_{Y(1)|X}(y \mid x),$$

sono continue in x per tutte le y. Queste assunzioni sono più forti di quanto richiesto poiché utilizzeremo la continuità solo in $x = c$, ma è raro che sia ragionevole assumere continuità solo per uno specifico valore della covariata.
Sotto entrambe le assunzioni,

$$\mathbb{E}[Y(0) \mid X = c] = \lim_{x \uparrow c} \mathbb{E}[Y(0) \mid X = x] = \lim_{x \uparrow c} \mathbb{E}[Y(0) \mid W = 0, X = x] = \lim_{x \uparrow c} \mathbb{E}[Y \mid X = x]$$

e quindi in maniera analoga

$$\mathbb{E}[Y(1) \mid X = c] = \lim_{x \downarrow c} \mathbb{E}[Y \mid X = x].$$

Pertanto, l'effetto medio del trattamento a c, τ_{SRD}, soddisfa:

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathbb{E}[Y \mid X = x] - \lim_{x \uparrow c} \mathbb{E}[Y \mid X = x]$$

La stima è quindi la differenza di due funzioni di regressione in un punto (Imbens & Lemieux, 2008).

1.2.2 Modello *Fuzzy RD*

Prima di iniziare a descrivere le caratteristiche principali del disegno, è importante introdurre tre definizioni utili per il prosieguo della trattazione:

- Si definisce *complier* un'unità che riceve il trattamento se il suo valore associato della variabile punteggio è maggiore o uguale alla soglia. Viceversa, viene assegnato al gruppo di controllo se lo *score* ottenuto è inferiore al punto di taglio. Si ha quindi che un'unità *complier* viene influenzata dal valore della soglia fissata;

- Si definisce *nevertaker* un'unità tale che, a prescindere dal valore della soglia, fa sempre parte del gruppo di controllo, poiché il suo valore della variabile di *score* sarà sempre inferiore alla soglia e quindi il valore di quest'ultima non influenza lo status di trattamento;

- In ultima istanza, per *alwaystaker* si intende un'unità tale che, a prescindere dal valore della soglia, fa sempre parte del gruppo di trattamento, poiché il suo valore della variabile di *score* sarà sempre superiore alla soglia e quindi il valore di quest'ultima non influenza lo status di trattamento.

Finora abbiamo supposto che l'attraversamento della soglia determinasse la ricezione del trattamento in modo che il salto delle funzioni di regressione in corrispondenza del *cutoff* possa essere considerato come l'effetto causale del trattamento. Quando superare la soglia non è più l'unica causa per cui si riceve un trattamento, si ha che lo status di trattamento W non è più una funzione deterministica di X (Hahn et al., 2001). Sarebbe più utile pensare al valore c come una soglia nella quale la probabilità di ricevere un trattamento fa un salto che può essere dovuto a variabili non osservabili che hanno un impatto sulla probabilità di essere trattati. Pertanto, X sarà correlata con l'errore u e diventa più difficile stimare in modo consistente l'effetto del trattamento. In questa impostazione, usare la metodologia *Fuzzy RD* può essere una soluzione al problema: data la variabile binaria Z_i che indica se si è superata o meno la soglia:

$$Z_i = \begin{cases} 1 & se \quad X_i \geq c \\ 0 & se \quad X_i < c \end{cases}$$

si assume che Z_i sia correlata con Y_i solo tramite l'indicatore di trattamento W_i. In questo modo Z_i e u_i sono incorrelate ma Z_i influenza la ricezione del trattamento, quindi è correlata con W_i. Nel disegno *Fuzzy* è importante fare una distinzione: essere assegnati al gruppo di trattamento non equivale ad averlo ricevuto.

Questa distinzione è molto importante poiché non essere un'unità *complier* induce ulteriori complicazioni e richiede assunzioni più forti per comprendere gli effetti del trattamento che si sta valutando. Vi sono due tipologie di modello Fuzzy RD (Jacob et al., 2012): *FRD no-shows* e *FRD no-shows and cross-over*. Rispetto al disegno *Sharp RD*, quindi, il metodo *Fuzzy* prevede che vi sia un salto più piccolo in corrispondenza della soglia e che la probabilità di ricevere il trattamento cresca all'aumentare del punteggio ottenuto nella variabile punteggio:

$$\lim_{x \downarrow c} \Pr\left(W_i = 1 \mid X_i = x\right) \neq \lim_{x \uparrow c} \Pr\left(W_i = 1 \mid X_i = x\right).$$

In questa sezione si interpreta come effetto causale medio del trattamento il rapporto tra il salto nella regressione della risposta sulle covariate ed il salto nella regressione dello status di trattamento sulle covariate. In modo formale viene così espresso (Imbens & Lemieux, 2008):

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y \mid X = x] - \lim_{x \uparrow c} \mathbb{E}[Y \mid X = x]}{\lim_{x \downarrow c} \mathbb{E}[W \mid X = x] - \lim_{x \uparrow c} \mathbb{E}[W \mid X = x]}$$

che, date le definzioni precedenti di *complier*, *nevertaker* e *alwaystaker*, può essere riscritto come:

$$\tau_{\text{FRD}} = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid i \text{ è un } \textit{complier} \text{ e } X_i = c\right].$$

La stima è quindi l'effetto medio del trattamento, ma solamente nel caso in cui l'unità è un *compliers* ed $X_i = c$.
Analizziamo nel dettaglio adesso la modalità di assegnazione di un trattamento nelle due tipologie di disegno. Il primo metodo prevede che nessuna unità con $X_i < c$ riceva il trattamento. Alle unità con $Xi \geq c$, viene invece assegnato il trattamento in modo casuale. In questo modo, lo status di trattamento è determinato solo parzialmente dal valore di X_i. Le osservazioni con punteggio maggiore o uguale alla soglia ($X_i \geq c$) che non hanno ricevuto il trattamento vengono denominate *no-shows*.

Si provvede a fornire in Figura 1.5 un grafico atto a spiegare quanto appena detto.

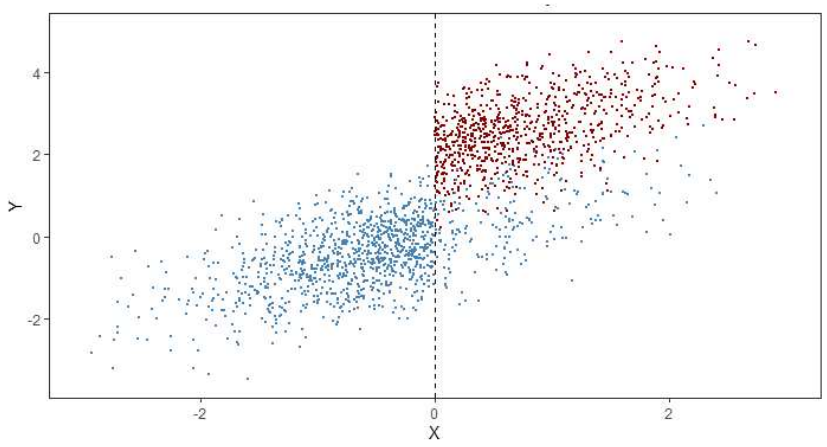


FIGURA 1.5: Divisione in trattamento-controllo: Fuzzy RD no-shows

Si può infatti osservare che alcune unità siano presenti nel gruppo di osservazioni che si trovano alla destra della soglia. Ciò accade perché il superamento della soglia non garantisce l'assegnazione del trattamento e quindi questa probabilità, che è pari a 1 nel caso di disegno *Sharp RD*, assume adesso valori minori di 1. Il secondo metodo prevede un'ulteriore caratteristica. Oltre ad assumere una probabilità di assegnazione del trattamento minore di 1 a quelle unità con $X_i \geq c$, assegna una probabilità maggiore di 0 a quelle unità con $X_i < c$. Tramite Figura 1.6 si mostra graficamente quanto descritto in precedenza.

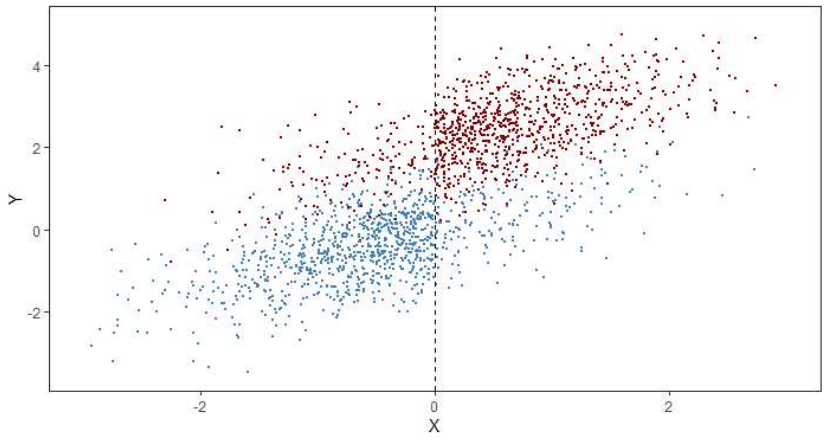


FIGURA 1.6: Divisione in trattamento-controllo: Fuzzy RD no-shows and cross-over

Si può notare, infatti, come siano presenti sia unità rosse alla sinistra della soglia, ossia quelle con $X_i \geq c$ che non hanno ricevuto trattamento, sia unità blu alla destra della soglia, cioè quelle con $X_i < c$, a cui è stato assegnato il trattamento.

Le unità con valore di $X_i < c$ a cui viene assegnato il trattamento sono dette *crossover*. Si mostra adesso in Figura 1.7 il grafico relativo alla probabilità di assegnazione del trattamento per il metodo che considera solo i *no-shows*.

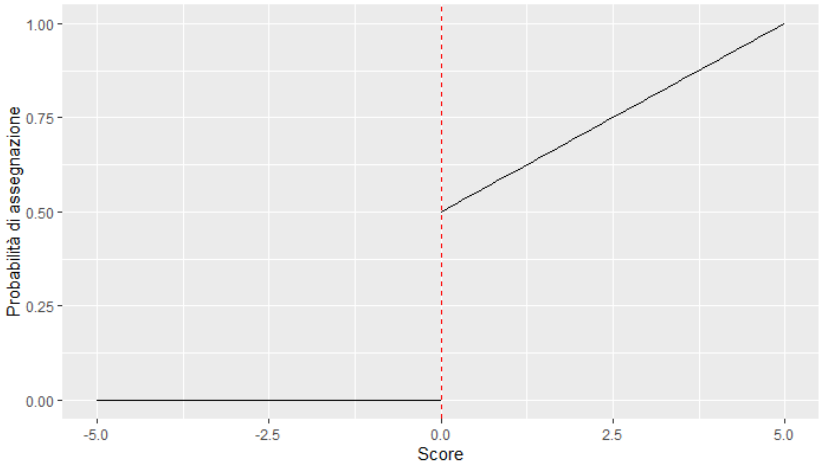


FIGURA 1.7: Divisione in trattamento-controllo: Fuzzy RD no-shows

Nel grafico in questione si nota quindi una probabilità di assegnazione pari a 0 per tutte le unità che non hanno raggiunto la soglia, viceversa una probabilità crescente per le osservazioni alla destra della soglia. In particolare, la probabilità di assegnazione del trattamento, per le unità alla destra della soglia, assume valore 0.5 in corrispondenza di questa fino ad arrivare ad 1 per punteggi uguali a 5. Infine in Figura 1.8, si mostra il grafico relativo alla probabilità di assegnazione del trattamento per il metodo che considera anche i *crossover*:

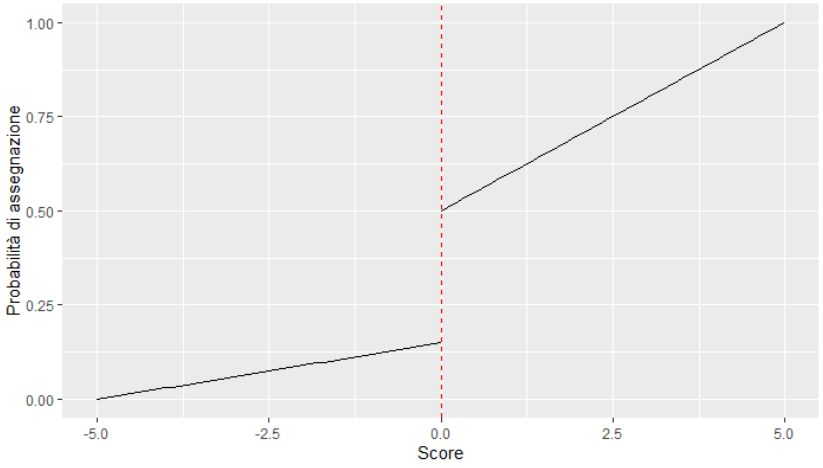


FIGURA 1.8: Divisione in trattamento-controllo: Fuzzy RD no-shows and cross-over

In questo caso, per le unità che si trovano alla sinistra della soglia, si ha una probabilità di assegnazione del trattamento che va da 0, per unità con valori dello *score* pari a 5, a 0.15 per le unità con valore dello *score* in prossimità della soglia. Per le unità che hanno invece almeno raggiunto la soglia, la probabilità di assegnazione è la stessa di quella rappresentata in Figura 1.7.

Capitolo 2

Analisi grafica

2.1 Grafici

In questo capitolo verrano mostrate le tecniche tramite le quali effettuare delle analisi grafiche nel contesto del *Regression Discontinuity Design*. Un primo grafico che permette di visualizzare la relazione tra variabile risposta e variabile di punteggio è un semplice *scatterplot*. Tramite questo grafico si possono individuare le unità che si trovano a destra ed a sinistra della soglia. Nel contsto di riferimento, uno degli svantaggi dello *scatterplot* consiste nella difficoltà di disporre di una visualizzazione diretta di salti o punti di discontinuità, che intercorrono nella relazione tra la variabile risposta e lo *score*, solo tramite i dati grezzi. Per poter osservare un salto, si potrebbero, ad esempio, adattare due rette di regressione e valutare, come detto nella sezione precedente, lo scarto che si ha tra le due intercette in corrispondenza della soglia. Mostriamo, quindi, in Figura 2.1 un esempio semplice di *scatterplot*.

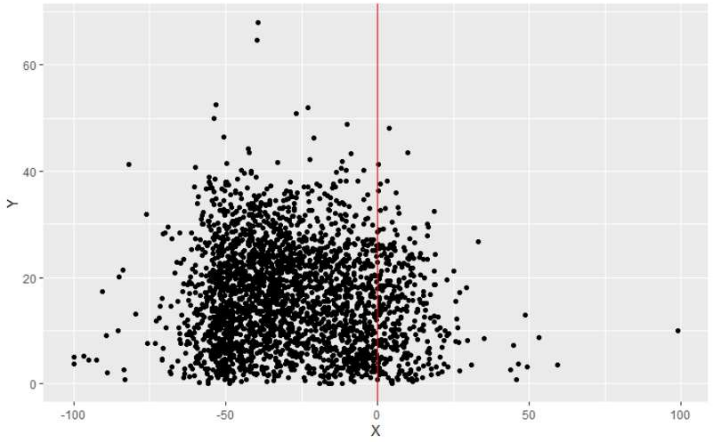


FIGURA 2.1: *Scatterplot*

Come detto in precedenza, dal grafico non risulta semplice notare punti di discontinuità in prossimità della soglia, obiettivo per il quale il metodo viene applicato. Ciò non accade nemmeno in situazioni nelle quali in realtà vi è una grande discontinuità al *cutoff*. Una mossa utile potrebbe essere quella di aggregare i dati prima di visualizzarli graficamente. Un tipico grafico, atto a mostrare il funzionamento della metodologia di *Regression Discontinuity Desing*, detto *RD Plot*, presenta due elementi chiave: il primo è l'adattamento polinomiale globale, mentre il secondo è rappresentato da medie campionarie locali, rappresentate graficamente da punti. Per adattamento polinomiale globale si intende un'approssimazione delle ignote funzioni di regressione basate su una regressione adattata tramite polinomio, separatamente per le osservazioni alla destra ed alla sinistra della soglia, tramite l'utilizzo dei semplici dati grezzi. Le medie campionarie locali sono calcolate successivamente alla divisione della variabile di punteggio in intervalli disgiunti, detti *bins*. All'interno di ogni intervallo disgiunto viene calcolata la media della variabile risposta. La combinazione di questi due elementi in un unico grafico permette di osservare la pendenza delle due funzioni di regressione, una per il gruppo di controllo ed una per il gruppo trattamento, riuscendo allo stesso tempo a visualizzare il comportamento locale dei dati per osservare l'effetto del trattamento. Una condizione necessaria che deve essere rispettata nella costruzione dei *bins* è che questi non debbano contenere unità appartenenti a gruppi diversi (Cattaneo et al., 2019). Una volta descritte e definite le componenti chiave per un grafico atto a rappresentare la metodologia del *Regression Discontinuity Design*, si fornisce un esempio nel quale la variabile punteggio viene ripartita in 40 intervalli disgiunti, 20 per ogni lato della soglia:

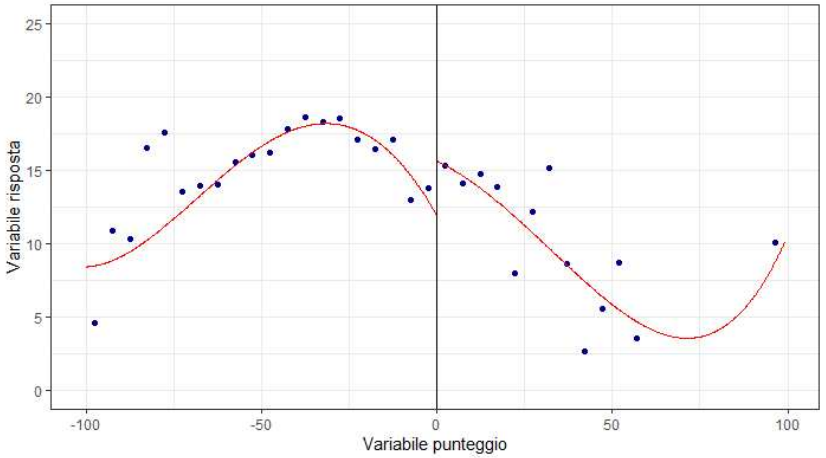


FIGURA 2.2: RD plot

In Figura 2.2 si mostra l'adattamento di due polinomi, uno per ogni gruppo di assegnazione. Osservando bene si nota un salto positivo in corrispondenza della soglia. Ciò lascia quindi presagire che vi sia un effetto positivo del trattamento. Le informazioni tratte dallo *scatterplot* e dall'*RD plot* sono, quindi, di natura differente. Utilizzare quest'ultimo, dopo aver diviso le unità in intervalli disgiunti, ci permette di valutare meglio l'andamento delle funzioni di regressione e soprattutto ci aiuta a notare eventuali punti di discontinuità in corrispondenza della soglia, obiettivo principale della metodologia. Generalmente si distinguono due diverse tipologie di intervalli disgiunti:

- *bins* che hanno stessa lunghezza, che non contengono lo stesso numero di unità;
- *bins* che contengono lo stesso numero di osservazioni, che potrebbero però, nella maggior parte dei casi, avere lunghezze diverse.

La prima tipologia viene definita *evenly-spaced bins* (Cattaneo et al., 2019) che abbrevieremo con ES, la seconda invece viene denotata con *quantile-spaced bins* (Cattaneo et al., 2019), che abbrevieremo con QS. La differenza principale tra le due tipologie di intervalli disgiunti risiede nella variabilità delle stime della media locale in ogni intervallo. Gli intervalli *ES* nonostante abbiano la stessa lunghezza, non contengono lo stesso numero di osservazioni, a meno che queste non siano uniformemente distribuite. Al contrario invece i *bins* QS contengono quasi sempre lo stesso numero di unità. Forniamo adesso, in Figura 2.3, un esempio grafico delle due tipologie di *bins*.

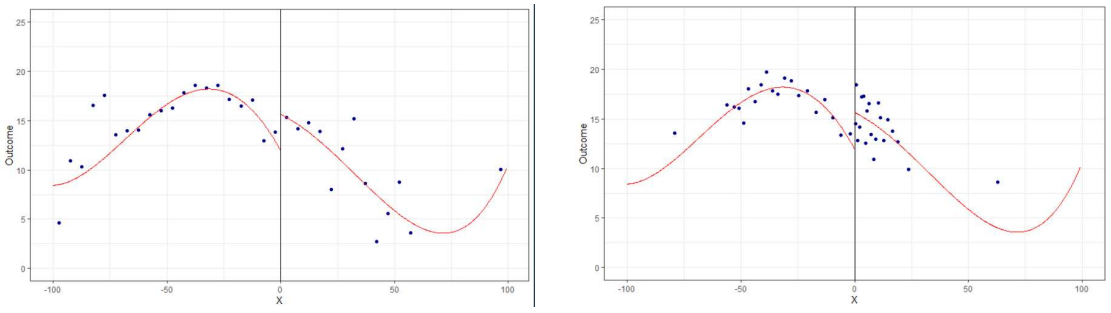


FIGURA 2.3: ES bins (sinistra) - QS bins (destra)

In questo esempio la differenza tra le due tipologie di intervalli, in termini di posizionamento dei *bins*, è evidente. Il salto che si nota in corrispondenza della soglia è invece uguale poiché cambia la collocazione geografica degli intervalli disgiunti, ma l'adattamento polinomiale locale rimane lo stesso e conseguentemente anche l'effetto del trattamento sarà lo stesso. Ad esempio è possibile notare come nell'intervallo di X che

va da -100 a -90, il numero di *bins QS* sia in numero molto inferiore rispetto alla tipologia *ES*. La differenza risiede nel fatto che in quell'intervallo considerato sono presenti pochissime osservazioni, in particolare queste ammontano a 6. Ciò fa si che, utilizzando *bins ES*, si abbiano stime delle medie locali con alta varianza. Il problema viene attenutato tramite l'utilizzo dei *bins QS*, poiché ogni intervallo contiene, per costruzione, lo stesso numero di unità. Una volta scelta la tipologia di intervalli disgiunti da utilizzare per visualizzare graficamente la relazione tra variabile risposta e *score*, rimane solamente da decidere la quantità di *bins* da osservare alla destra e alla sinistra della soglia. Per far ciò esistono in particolare due metodi: la prima denominata *Integrated Mean Squared Error Method*(*IMSE*), in italiano Metodo dell'errore quadratico medio integrato e la seconda denominata *Mimicking Variance Method*, in italiano Metodo di imitazione della varianza. Il primo metodo proposto ha come obiettivo quello di minimizzare un'approssimazione dell'errore quadratico medio integrato dello stimatore della media locale.
Distinguiamo due casistiche, una nella quale si sceglie un alto numero di *bins* ed un'altra dove invece il numero scelto risulta basso. Nel primo caso si avrà una bassa distorsione ma contemporaneamente, aumentare il numero di *bins* fa sì che vi sia un numero minore di osservazioni all'interno di ogni intervallo disgiunto e che quindi la variabilità aumenti. Viceversa, diminuendone il numero, si avrà un aumento in distorsione ed una diminuzione in variabilità. Si cerca quindi, per ogni lato del *cutoff*, il numero ottimale di intervalli disgiunti al fine di trovare un compromesso tra varianza e distorsione. Vengono mostrati adesso due grafici in cui vengono mostrate le differenze nel numero di *bins* scelti dal metodo per quanto concerne gli intervalli *evenly-spaced* e *quantile-spaced*:

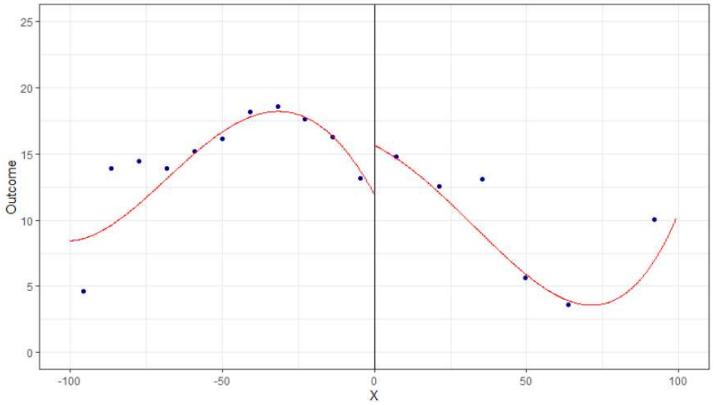


FIGURA 2.4: IMSE Method - Bins ES

In Figura 2.4 è possibile notare come il numero di intervalli disgiunti sia diverso nei due lati della soglia. Nell'esempio in questione si hanno 11 punti, ossia medie locali, a

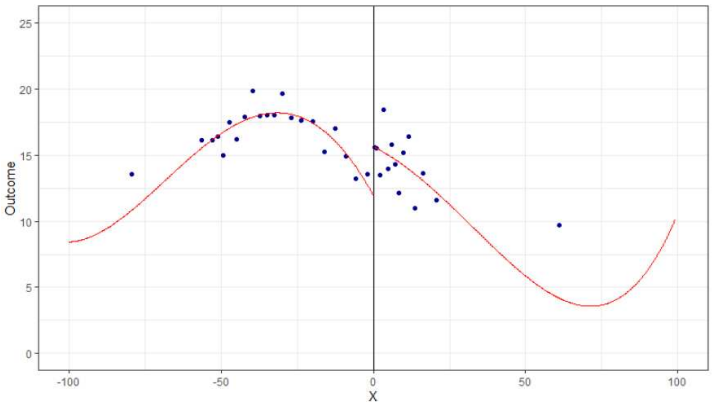


FIGURA 2.5: IMSE Method - Bins QS

sinistra del *cutoff*, mentre alla destra di queste ve ne sono 7.
Per quanto riguarda i *bins QS* invece, si ha in generale un maggior numero di medie locali stimate ed in particolare si osservano 21 *bins* alla sinistra della soglia e 14 alla destra. Ciò viene descritto in Figura 2.5.
Il secondo metodo, invece, tende a scegliere il numero di *bins* in modo tale che la variabilità delle medie locali stimate in ogni *bins* sia approssimativamente uguale alla variabilità rappresentata nello *scatterplot* iniziale. Si mostra adesso, in Figura 2.6, il grafico relativo al metodo MV per *bins* ES.

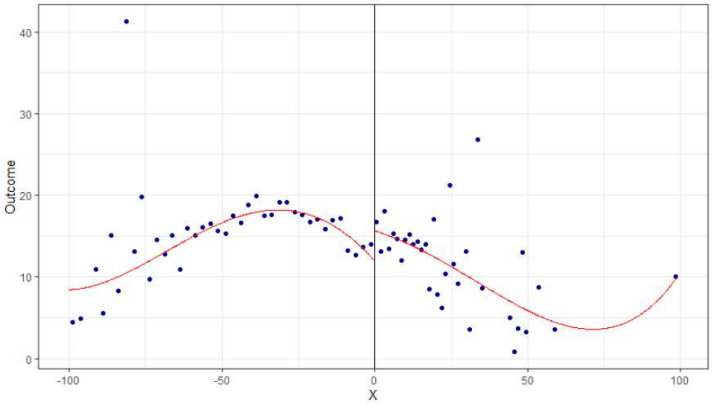


FIGURA 2.6: MV Method - Bins ES

In questo caso si ha un numero maggiore di *bins* con $J_- = 40$ e $J_+ = 75$, mentre col metodo IMSE si aveva $J_- = 11$ e $J_+ = 7$, dove J_- e J_+ indicano, rispettivamente, il numero di intervalli disgiunti a sinistra e a destra della soglia.

Infine si mostra in Figura 2.7, il grafico relativo allo stesso metodo, questa volta però per *bins* QS.

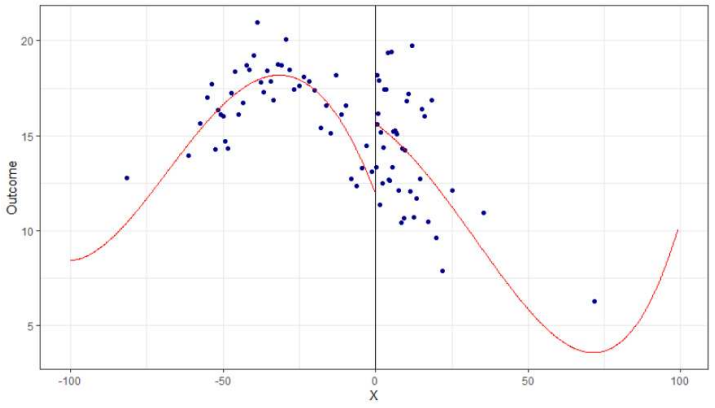


FIGURA 2.7: MV Method - Bins QS

Anche con tipologia di *bins* QS è possibile osservare un aumento di questi, ma ridotto rispetto al caso ES. In particolare, col metodo MV si ha $J_- = 44$ e $J_+ = 41$ contro $J_- = 21$ e $J_+ = 14$ del metodo IMSE.

Capitolo 3

Procedura di stima del *Regression Discontinuity Design*

3.1 Introduzione alla procedura di stima

Generalmente, nel *Regression Discontinuity Design*, la stima dell'effetto medio del trattamento, dato dalla differenza tra le due funzioni di regressione in corrispondenza della soglia, viene effettuata tramite metodi polinomiali locali al *cutoff*, utilizzati per approssimare le due funzioni di regressione, $\mathbb{E}[Y_i(1) \mid X_i = x]$ e $\mathbb{E}[Y_i(0) \mid X_i = x]$, una per ogni lato della soglia. Utilizzando un polinomio per l'approssimazione delle due curve di regressione, si va incontro a due possibili casi:

- stima globale e parametrica, nel caso in cui si utilizzassero tutte le osservazioni a disposizione nel campione;
- stima locale e non parametrica, se si utilizzasse invece solo una parte delle osservazioni.

Per ottenere una stima del parametro d'interesse $\hat{\tau}_{\text{SRD}}$ è quindi necessario approssimare $\mathbb{E}[Y_i(1) \mid X_i = x]$ e $\mathbb{E}[Y_i(0) \mid X_i = x]$. Una volta approssimate, si può poi procedere con la fase di stima e con la relativa procedura inferenziale che ne deriva. Il punto fondamentale è come approssimare le funzioni di regressione, essendo queste ignote.
L'approssimazione deve avvenire lungo i bordi, ed è risaputo che effettuare quest'operazione sulla frontiera risulti più complesso. Considere tutte le osservazioni a disposizione nel campione porterebbe ad una buona approssimazione generale, ma ciò non avviene lungo i bordi. Utilizzare solo una parte delle osservazioni a disposizione garantisce allora stime e procedure inferenziali migliori per il caso in questione (Avery, 2013).

Tramite approccio locale si tende ad utilizzare un'approssimazione tramite un polinomio di basso grado, tendenzialmente lineare o quadratico, che è essenzialmnte più robusto e meno sensibile al sovradattamento lungo i bordi. Si utilizzano quindi solo osservazioni che si trovano in un intorno della soglia. Le proprietà statistiche della stima e dell'inferenza derivante dall'approssimazione locale polinomiale dipendono dall'accuratezza dell'approssimazione nell'intorno del *cutoff*, che è controllata dalla dimensione di quest'ultimo.
Si può implementare, quindi, ad esempio, una regressione lineare utilizzando solo le osservazioni vicine al *cutoff*, separatamente per le unità appartenenti al gruppo di trattamento e al gruppo di controllo. In particolare, vengono utilizzate le osservazioni comprese tra $c - h$ e $c + h$ con $h > 0$, detta *ampiezza di banda*, che determina la dimensione dell'intorno della soglia all'interno del quale sono presenti le osservazioni che saranno poi utilizzate nel processo di stima.
Le osservazioni utilizzate in fase di stima potrebbero anche non avere tutte lo stesso peso. Ciò avviene nel caso in cui si decidesse di adottare un sistema che assegna pesi maggiori alle unità più vicine alla soglia e pesi minori a quelle più lontane. I pesi vengono determinati da una funzione nucleo, $K(\cdot)$.
Per applicare correttamente una regressione polinomiale locale è quindi necessario scegliere, tramite tecniche che vedremo successivamente, tre elementi fondamentali:

- la funzione nucleo, $K(\cdot)$;
- l'ordine del polinomio, p;
- l'ampiezza di banda, h.

Per ottenere quindi una stima dell'effetto medio del trattamento tramite approssimazione polinomiale locale bisogna eseguire 5 step:

1. Si sceglie l'ordine p del polinomio e la funzione nucleo $K(\cdot)$;
2. Si sceglie l'ampiezza di banda, h;
3. Si ottiene la stima $\hat{\mu}_+$ per le unità con $X_i \geq c$;
4. Si ottiene la stima $\hat{\mu}_-$ per le unità con $X_i < c$;
5. Si calcola la stima come $\hat{\tau}_{\text{SRD}} = \hat{\mu}_+ - \hat{\mu}_-$.

Un grafico utile a spiegare quanto detto finora viene mostrato in Figura 3.1.

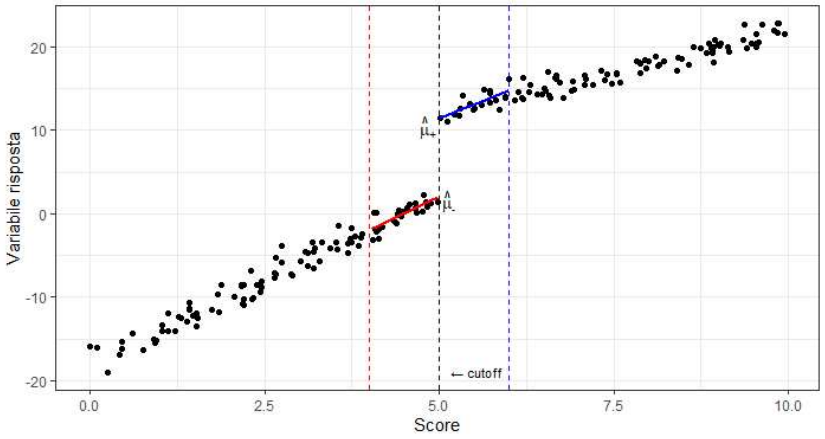


FIGURA 3.1: RDD: stima

Nell'esempio si è adattato un polinomio di grado 1, ossia una retta, all'interno dei due intorni della soglia, $c-h$ e $c+h$. Le unità con valori di X_i esterni all'intervallo $[c-h, c+h]$ non sono state utilizzate per adattare i polinomi. L'effetto medio del trattamento in questo caso è dato dalla distanza verticale tra $\hat{\mu}_+$ e $\hat{\mu}_-$.

3.1.1 Scelta della funzione nucleo e dell'ordine del polinomio

La funzione nucleo $K(\cdot)$ assegna pesi non negativi ad ogni osservazione trasformata $(X_i - c)/h$. Il sistema di assegnazione dei pesi è basato sulla distanza tra la variabile di *score* X_i ed il *cutoff* c. In generale, si tende ad assegnare via via un peso maggiore, man mano che il valore di X_i si avvicina al valore di c. Nel disegno di *Regression Discontinuity Design* la funzione che risulta avere migliori prestazioni e risultati è quella triangolare che, utilizzata congiuntamente ad un'ampiezza di banda h che ottimizza l'errore quadratico medio, porta a stime con proprietà ottimali (Cattaneo et al., 2019). La funzione nucleo triangolare assegna peso pari a 0 a osservazioni che presentano un valore di X_i esterno all'intervallo $[c-h, c+h]$, mentre assegna un peso positivo a quelle con valore della variabile di *score* compreso nell'intervallo precedentemente menzionato. Le osservazioni alle quali viene attribuito il peso maggiore sono quelle il cui valore associato è esattamente uguale a quello della soglia. Questo valore va poi diminuendo in modo simmetrico e lineare, all'aumentare della distanza tra il valore della variabile di *score* e il valore del *cutoff*. Le caratteristiche ottimali di questo tipo di funzione nucleo nel contesto del *Regression Discontinuity Design* sono molteplici:

- il *kernel* triangolare è simmetrico rispetto al punto di discontinuità, il che significa che assegna lo stesso peso alle osservazioni a sinistra e a destra del punto di soglia:

ciò garantisce che la stima del trattamento effettivo sia priva di distorsioni a causa di asimmetrie nella distribuzione delle osservazioni;

- ha un'ampiezza di banda relativamente stretta: ciò significa che le osservazioni vicine alla soglia hanno un peso maggiore rispetto ai punti più distanti. Questo è utile perché i punti vicini al punto di soglia sono quelli che influenzano maggiormente la stima del trattamento effettivo;

- ha una derivata continua al punto di soglia, quindi la stima del trattamento effettivo risulta più regolare e meno sensibile ai piccoli cambiamenti nei dati vicini al punto di soglia.

Nonostante le proprietà asintotiche ottimali della funzione nucleo triangolare, molte volte si preferisce utilizzare la funzione nucleo uniforme, la quale assegna peso pari a 0 alle osservazioni fuori da $[c-h, c+h]$ e peso positivo ed equivalente per tutte le osservazioni all'interno dell'intervallo. Utilizzare una funzione nucleo di questo tipo equivale ad adattare una semplice regressione lineare senza pesi, utilizzando solamente le osservazioni comprese nell'intervallo $[c-h, c+h]$. Questa funzione nucleo ha come vantaggio quello di minimizzare la varianza asintotica dello stimatore, sotto specifiche condizioni. Un'ulteriore funzione nucleo utilizzata è quella di *Epanechnikov*, che assegna peso nullo alle osservazioni esterne all'intervallo considerato, mentre per quelle contenute al suo interno assegna un peso che decade in maniera quadratica all'aumentare della distanza di X_i dalla soglia c. In Figura 3.2 viene spiegato graficamente il funzionamento del sistema di assegnazione dei pesi delle tre funzioni nucleo considerate.

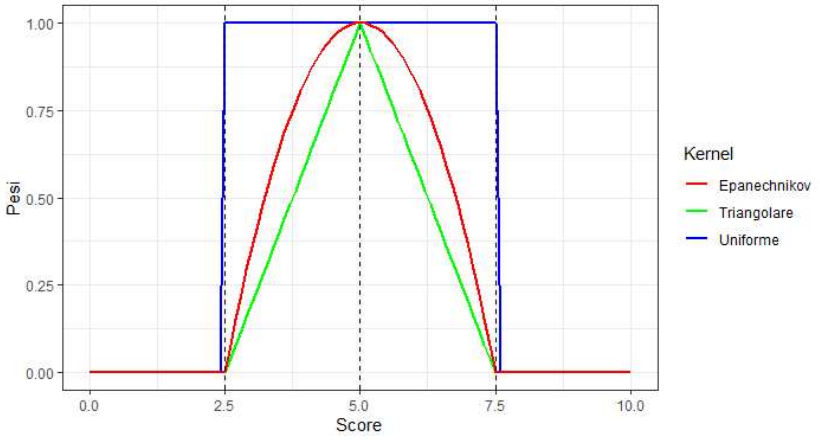


FIGURA 3.2: Sistema di assegnazione dei pesi: Kernel

Un altro tema importante da discutere è quello che concerne la scelta del grado del polinomio. Un polinomio di grado 0, ossia una costante, non ha buone proprietà lungo

i bordi, punto in cui ci interessa stimare l'effetto del trattamento, mentre un polinomio di alto grado presenta invece buone proprietà. Queste, nel contesto di adattamenti polinomiali locali, corrispondono generalmente a continuità, derivabilità, flessibilità e robustezza sulla frontiera. Potrebbe bastare questo per poter affermare che sia sufficiente scegliere un alto grado del polinomio, tuttavia si va incontro ad un altro problema. Per una data ampiezza di banda h, aumentare l'ordine del polinomio aumenta l'accuratezza della stima (diminisce quindi la distorsione dello stimatore), ma così facendo si avrebbe un aumento della varianza di questo. Si hanno quindi due casi estremi: un polinomio di basso grado comporta un'alta distorsione ed una bassa varianza, situazione che viene denominata sottoadattamento, mentre un polinomio di alto grado comporta una bassa distorsione ed un'alta varianza, situazione che viene denominata sovradattamento. Bisogna quindi trovare, anche in questo caso, un compromesso tra distorsione e varianza. In alcuni studi si è visto come un polinomio di grado 1 porti ad un buon compromesso in termini di stabilità, precisione e semplicità (Hardle et al., 1993), (Fan & Gijbels, 1995). Nonostante un polinomio lineare sia poco flessibile, una scelta appropriata dell'ampiezza di banda h porta ad una buona approssimazione delle ignote funzioni di regressione.

3.1.2 Selezione dell'ampiezza di banda

Il parametro di lisciamento h controlla l'ampiezza dell'intervallo all'interno del quale si troveranno le osservazioni utilizzate per l'adattamento del polinomio di grado p scelto precedentemente, che approssima le ignote funzioni di regressione. La scelta di h è fondamentale per l'analisi di *Regression Discontinuity* poichè ha un'influenza diretta sulle proprietà della stima e delle procedure inferenziali. In Figura 3.3 si mostra un grafico utile a comprendere cosa succede all'aumentare ed al diminuire del valore di h. Considerato $h_1 < h_2$ si hanno due intervalli, $[c - h_1; c + h_1]$ e $[c - h_2; c + h_2]$, dove nel primo intervallo è contenuto un numero minore di osservazioni rispetto al secondo. In quest'ultimo, si avrà un effetto stimato del trattamento pari a $\hat{\mu}_+ (h_2) - \hat{\mu}_- (h_2)$, che nel grafico è dato dalla distanza tra i punti c e d. L'effetto stimato sembra essere decisamente differente dal vero effetto del trattamento τ_{SRD}. Restringendo l'ampiezza dell'intervallo, notiamo come l'approssimazione migliori in modo concreto e che adesso, l'effetto stimato corrisponde a $\hat{\mu}_+ (h_1) - \hat{\mu}_- (h_1)$, che nel grafico viene visualizzato come la distanza verticale tra a e b. In questo caso la stima dell'effetto sembra avvicinarsi decisamente al vero effetto del trattamento. Restringendo quindi l'intervallo all'interno del quale adattare il polinomio di grado p scelto, si ha un miglioramento dell'accuratezza della stima e quindi dell'approssimazione. Non bisogna però pensare che basti restringere l'intervallo $[c - h, c + h]$ per ottenere la miglior stima possibile. Da un lato, ridurre il

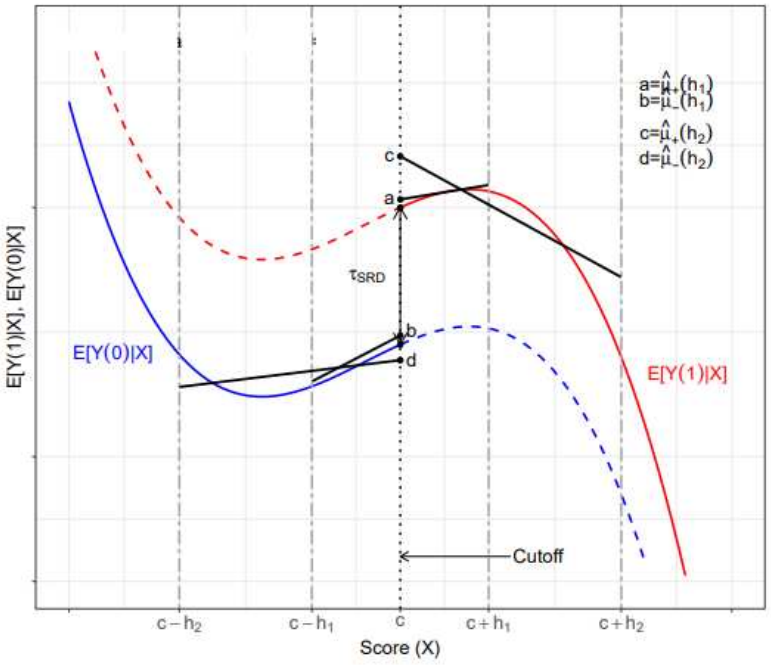


FIGURA 3.3: Distorsione e varianza per diversi valori di h

valore di h comporta un aumento dell'accuratezza e contemporaneamente un aumento della variabilità dei coefficienti stimati, dato che il numero delle osservazioni disponibili per la procedura di stima sarà molto basso. Dall'altra parte, aumentare h porta ad una diminuzione dell'accuratezza dell'approssimazione, ma allo stesso tempo si avrà una diminuzione nella varianza dei coefficienti. Risulta necessario, quindi, trovare un compromesso tra varianza e distorsione. L'approccio maggiormente utilizzato per la scelta del parametri di lisciamento h consiste nel minimizzare l'errore quadratico medio dello stimatore degli effetti del trattamento. Sappiamo che l'errore quadratico medio è dato dalla somma della distorsione al quadrato e della varianza dello stimatore, in formule:

$$\text{MSE}\left(\hat{\tau}_{\text{SRD}}\right) = \textit{Distorsione}^2\left(\hat{\tau}_{\text{SRD}}\right) + \text{Varianza}\left(\hat{\tau}_{\text{SRD}}\right) = B^2 + V,$$

dove la distorsione e la varianza approssimati sono pari a:

- $B = h^{2(p+1)}\mathcal{B}$;
- $V = \frac{1}{nh}\mathcal{V}$.

Le quantità $\mathcal{B}$ e $\mathcal{V}$ sono definite rispettivamente come $\mathcal{B} = \mathcal{B}_+ - \mathcal{B}_-$ e $\mathcal{V} = \mathcal{V}_+ - \mathcal{V}_-$, dove:

- $\mathcal{B}_- \approx \mu_-^{(p+1)} K_-, \quad \mathcal{B}_+ \approx \mu_+^{(p+1)} K_+,$

- $\mathcal{V}_- \approx \frac{\sigma_-^2}{f} Q_-, \quad \mathcal{V}_+ \approx \frac{\sigma_+^2}{f} Q_+$

In particolare :

- le derivate $\mu_+^{(p+1)} = \lim_{x \downarrow c} \frac{d^{p+1}\mathbb{E}[Y_i(1)|X=x]}{dx^{p+1}}$ e $\mu_-^{(p+1)} = \lim_{x \uparrow c} \frac{d^{p+1}\mathbb{E}[Y_i(0)|X=x]}{dx^{p+1}}$, dipendono dalla curvatura delle ignote funzioni di regressione ;
- K_+ e K_- sono costanti;
- f rappresenta la densità dello *score* al *cutoff*;
- $\sigma_+^2 = \lim_{x \downarrow c} \mathbb{V}\left[Y_i(1) \mid X_i = x\right]$ e $\sigma_-^2 = \lim_{x \uparrow c} \mathbb{V}\left[Y_i(0) \mid X_i = x\right]$;
- Q_+ e Q_- sono costanti.

Si nota come distorsione e varianza dello stimatore degli effetti del trattamento dipendano, nella loro formulazione, dal valore dell'ampiezza di banda. In particolare, la distorsione dipende dal valore di h al numeratore, viceversa la varianza al denominatore, e quindi in maniera inversa. Si ha quindi che:

- se $h \to 0$ allora la distorsione tenderà a zero, mentre la varianza tenderà a infinito;
- se $h \to \infty$ allora la distorsione tenderà ad infinito e contemporaneamente la varianza tenderà a zero.

È quindi necessario trovare un valore di h tale da giungere ad un compromesso tra varianza e distorsione, al fine di minimizzare l'errore quadratico medio. Si avrà quindi il seguente problema di minimizzazione:

$$\min_{h>0} \left(h^{2(p+1)} \mathcal{B}^2 + \frac{1}{nh} \mathcal{V} \right)$$

la cui soluzione è l'ampiezza di banda h ottimale (Cattaneo et al., 2019), che si dimostra essere data dalla seguente formulazione:

$$h^{\text{opt}} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2} \right)^{1/(2p+3)} n^{-1/(2p+3)}.$$

Questo valore è proporzionale a $n^{-1/(2p+3)}$ ed il suo valore aumenta con $\mathcal{V}$ e diminuisce con $\mathcal{B}$. Un incoveniente che si riscontra in alcune applicazioni si ha quando le stime delle distorsioni ai due lati della soglia sono prossime a 0, che causano scarsi risultati per i due valori di h scelti. Per ovviare al problema si include un termine di regolarizzazione,

$\mathcal{R}$, al fine di evitare che il denominatore sia troppo piccolo in campioni con dimensione campionaria bassa. Nel caso di un'ampiezza di banda comune il valore di h^{opt} diventa:

$$h_*^{\text{opt}} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2 + \mathcal{R}} \right)^{1/(2p+3)} n^{-1/(2p+3)},$$

dove $\mathcal{R}$, essendo a denominatore, comporta un valore ottimo di h minore (Cattaneo et al., 2019).
Si fornisce adesso un'ulteriore motivazione teorica del perchè sia necessario arrivare a trovare un compromesso tra varianza e distorsione.
Scegliendo $h > h^{\text{opt}}$, si ha che al diminuire del valore di h:

1. la distorsione diminuisce;
2. la varianza aumenta.

Al diminuire del valore dell'ampiezza di banda si avrà una riduzione dell'errore quadratico medio, dovuta alla diminuzione della distorsione, che sarà maggiore in modulo dell'aumento che si ha a causa dell'incremento della variabilità. Nel caso in cui $h > h^{\text{opt}}$ è quindi possibile ridurre la distorsione senza che l'errore quadratico medio aumenti.
Al contrario, quando si ha $h = h^{\text{opt}}$, aumentare o diminuire il valore di h porta necessariamente ad un incremento dell'errore quadratico medio. In molte applicazioni è molto utile scegliere due diverse ampiezze di banda, h_- e h_+, una per la parte sinistra ed una per la parte destra, rispetto al *cutoff*. Così facendo si avranno due approssimazioni dell'EQM, una per ogni lato rispetto alla soglia, separatamente per ogni stima. Vengono quindi scelti due diversi valori di h per $\hat{\mu}_+$ e $\hat{\mu}_-$, che verranno utilizzati per ottenere lo stimatore degli effetti del trattamento. In pratica, questo equivale a scegliere un intorno asimmetrico del *cutoff* c, $[c - h_-; c + h_+]$, dove:

- h_- è l'ampiezza di banda del gruppo controllo;
- h_+ è l'ampiezza di banda del gruppo trattamento.

In questo caso la soluzione al problema di minimizzazione dell'errore quadratico medio è data da:

- $h_{\text{opt}-} = \left(\frac{\mathcal{V}_-}{2(p+1)\mathcal{B}_-^2} \right)^{1/(2p+3)} n_-^{-1/(2p+3)}$
- $h_{\text{opt}+} = \left(\frac{\mathcal{V}_+}{2(p+1)\mathcal{B}_+^2} \right)^{1/(2p+3)} n_+^{-1/(2p+3)}$

rispettivamente per il gruppo controllo ed il gruppo trattamento. La scelta di utilizzare due ampiezze di banda differenti è molto rilevante nel momento in cui la distorsione e/o la varianza dei due gruppi differiscono sostanzialmente, ad esempio nel caso di:

- differenti curvature delle funzioni di regressione;
- differenti valori attesi condizionati;
- differenti varianza condizionate.

Capitolo 4

Inferenza e validità del disegno

4.1 Problematiche

La prima parte di questa sezione concerne la definizione e l'implementazione di test di ipotesi e la costruzione di relativi intervalli di confidenza. Alcune problematiche relative alle procedure inferenziali nascono a causa dell'utilizzo del valore ottimo dell'ampiezza di banda. Ciò accade perché, essenzialmente, gli obiettivi e gli scopi delle procedure di stima ed inferenza sono differenti. L'inferenza convenzionale, ad esempio, assume che il modello adattato sia correttamente specificato e che, quindi, la distorsione sia trascurabile. Nel Capitolo precedente si è visto come in realtà la distorsione sia non nulla ed, in particolare, il valore di h_{opt} è stato ottenuto tramite il compromesso tra varianza e distorsione. Ciò significa che è necessario valutare diversi approcci inferenziali per quanto concerne l'utilizzo del valore ottimo dell'ampiezza di banda. Per ovviare a questo problema si suggeriscono due approcci:

- usare h_{opt} sia per la fase di stima che per la fase di inferenza, modificando le classiche statistiche test;
- usare h_{opt} solamente in fase di stima, per poi scegliere un altro valore dell'ampiezza di banda per la procedura inferenziale.

4.2 Utilizzo di h_{opt} per l'inferenza

In questa prima parte ci si focalizza su procedure inferenziali che utilizzano il valore del parametro di regolazione scelto nella precedente fase di stima. Lo stimatore $\hat{\tau}_{\text{SRD}}$ ha

distribuzione approssimata pari a:

$$\frac{\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} - B}{\sqrt{V}} \overset{a}{\sim} \mathcal{N}(0,1)$$

dove B e V sono rispettivamente distorsione e varianza dello stimatore polinomiale locale di ordine p. Rispetto alla regressione standard, il termine di distorsione figura in maniera esplicita. Un intervallo di confidenza asintotico al 95% è dato, in modo approssimato, da:

$$\text{CI} = \left[(\hat{\tau}_{\text{SRD}} - B) \pm 1.96 \cdot \sqrt{V}\right].$$

A meno che B non sia trascurabile, qualsiasi procedura inferenziale che ignori il termine di distorsione porterà ad un'inferenza errata, dato che l'intervallo di confidenza dipende dal suo valore. Descriviamo tre strategie diverse per fare inferenza sul parametro d'interesse, basate sulla distribuzione approssimata di $\hat{\tau}_{\text{SRD}}$.

4.2.1 Inferenza convenzionale

La prima strategia proposta assume che la distorsione sia trascurabile, ossia $B = 0$. Ciò appare insensato dal punto di vista metodologico poiché nel caso in cui la distorsione sia nulla, il valore dell'ampiezza di banda ottimale, h_{opt}, non potrebbe essere selezionato tramite il compromesso tra varianza e distorsione. Questa strategia valuta l'approccio polinomiale locale come se fosse un metodo parametrico, assumendo quindi che il modello sia correttamente specificato. Nel caso in cui la distorsione sia nulla, questa strategia potrebbe essere sensata da attuare. La distribuzione approssimata degli effetti del trattamento sarebbe (Cattaneo et al., 2019):

$$\frac{\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}}{\sqrt{V}} \overset{a}{\sim} \mathcal{N}(0,1)$$

e di conseguenza l'intervallo di confidenza associato:

$$\text{CI}_{\text{us}} = \left[\hat{\tau}_{\text{SRD}} \pm 1.96 \cdot \sqrt{V}\right].$$

L'intervallo di confidenza risultante dal considerare la distorsione trascurabile è equivalente a quello che si otterrebbe con la procedura ai minimi quadrati e per tale ragione lo denotiamo con intervallo di confidenza *convenzionale* (Cattaneo et al., 2019). Utilizzando questo tipo di intervallo si assume implicitamente che il polinomio scelto fornisca un'approssimazione esatta delle due funzioni di regressione, $\mathbb{E}\left[Y_i(0) \mid X_i\right]$ e $\mathbb{E}\left[Y_i(1) \mid X_i\right]$.

Ciò non è credibile in quanto sappiamo che le due funzioni sono ignote e quindi l'assunzione fatta in precedenza non è verificabile. Utilizzare questo intervallo in situazioni nelle quali la distorsione non è trascurabile porterebbe a sovra-rigettare l'ipotesi nulla di effetto del trattamento nullo. Un'alternativa consiste nell'utilizzare sempre l'intervallo convenzionale, utilizzando questa volta un'ampiezza di banda minore di quella ottima scelta precedentemente tramite il compromesso tra varianza e distorsione. In pratica si selezione l'ampiezza di banda ottima h_{opt}, successivamente si seleziona un valore minore di quello scelto al punto precedente ed infine si costruisce un intervallo di confidenza convenzionale con questo nuovo valore scelto. La giustificazione teorica è che per valori di $h < h_{opt}$, il termine di distorsione diventa trascurabile nella distribuzione approssimata dello stimatore risultante. Lo svantaggio di questa procedura è che non vi sono criteri chiari su come comprimere l'ampiezza di banda. A prescindere da come viene ridotto il valore di h, in generale, la procedura porta ad una perdita di potenza statistica poiché un h più piccolo comporta un numero di osservazioni minore, all'interno dell'intervallo considerato, utilizzato per le procedure di stima e inferenza.

4.2.2 *Standard Bias Correction*

Questa procedura, a differenza di quella basata sull'inferenza convenzionale, fonda la procedura inferenziale sulla stessa distribuzione approssimata dello stimatore degli effetti del trattamento, utilizzando però una stima del termine di distorsione. L'approccio viene denominato con *correzione della distorsione standard*, dall'inglese *standard bias correction* (Cattaneo et al., 2019). Si ottiene la stima di B, stimando precedentemente $\mu_+^{(p+1)}$ e $\mu_-^{(p+1)}$ tramite un polinomio locale di ordine $q \geq p+1$. Fatto ciò si può costruire l'intervallo come segue:

$$\mathrm{CI}_{\mathrm{bc}} = \left[\left(\hat{\tau}_{\mathrm{SRD}} - \hat{B}\right) \pm 1.96 \cdot \sqrt{V}\right]$$

Il termine di distorsione, dipende dalle curvature delle ignote funzioni di regressione, $\mathbb{E}\left[Y_i(0) \mid X_i\right]$ e $\mathbb{E}\left[Y_i(1) \mid X_i\right]$, ottenute tramite le loro derivate di ordine $p+1$ in corrispondenza della soglia. Per stimare le derivate si utilizzano polinomi di grado p+1 o superiore. Per far ciò è necessario disporre di una nuova ampiezza di banda, che nella trattazione denomineremo con b. Si ha quindi che:

- la stima dell'effetto del trattamento necessita dell'ampiezza di banda h;
- la stima del termine di distorsione necessita dell'ampiezza di banda b.

Il rapporto tra queste due ampiezze di banda, che denoteremo con $\rho = h/b$, misura la variabilità della stima, relativamente allo stimatore degli effetti del trattamento. In generale, l'approccio richiede che il rapporto tra le due ampiezze di bande sia molto piccolo, ossia che $\rho = h/b \to 0$. Viene escluso il caso in cui queste siano uguali, ovvero il caso in cui $\rho = 1$. Il metodo permette quindi di utilizzare ampiezze di banda più larghe, che portano a risultati inferenziali validi. Nonostante ciò, non si hanno buone prestazioni nelle applicazioni poiché stimando il termine di distorsione viene introdotta ulteriore variabilità che non viene inglobata nel termine di varianza V utilizzato per la costruzione dell'intervallo.

4.2.3 *Robust Bias Correction*

Il problema appena discusso viene risolto col metodo che viene denominato *robust bias correction*, in italiano *correzione robusta della distorsione* (Cattaneo et al., 2019). Questo approccio porta a procedure inferenziali valide, anche nel caso in cui venga utilizzato h_{opt}. L'approccio rimane inoltre valido anche nel caso in cui $h = b$, ed implica quindi che vengano usati gli stessi dati sia per la procedura di stima che per quella inferenziale. Gli intervalli che ne derivano si basano sulla stessa quantità di quello precedente, ossia $\hat{\tau}_{\text{SRD}} - \hat{B}$. La differenza risiede nel fatto che in questa procedura viene utilizzata una nuova varianza asintotica, V_{rbc}, che diversamente da quella utilizzata nei precedenti intervalli descritti, incorpora la variabilità addizionale introdotta nella fase di stima del termine di distorsione. In questo modo, questa nuova varianza risulta essere maggiore di quella ottenuta con metodi convenzionali usando la stessa ampiezza di banda.
L'approccio porta al seguente intervallo di confidenza:

$$\text{CI}_{\text{rbc}} = \left[\left(\hat{\tau}_{\text{SRD}} - \hat{B}\right) \pm 1.96 \cdot \sqrt{V_{\text{rbc}}}\right].$$

Come CI_{bc}, anche CI_{rbc} è costruito attorno alla stima corretta, $\hat{\tau}_{\text{SRD}} - \hat{B}$ e non attorno alla sola stima $\hat{\tau}_{\text{SRD}}$. Per riepilogare quanto detto finora mostriamo i tre intervalli di confidenza descritti in Tabella 4.1, nella quale vengono riportati nell'ordine: le tipologie di intervalli descritti, le quantità su cui questi sono centrati ed i relativi errori standard.

Convenzionale: CI_{us}	$\hat{\tau}_{\text{SRD}}$	$\sqrt{\hat{V}}$
Bias-Corrected: CI_{bc}	$\hat{\tau}_{\text{SRD}} - \hat{B}$	$\sqrt{\hat{V}_{\text{bc}}}$
Robust bias-corrected: CI_{rbc}	$\hat{\tau}_{\text{SRD}} - \hat{B}$	$\sqrt{\hat{V}_{\text{rbc}}}$

Tabella 4.1: Tipologie di intervalli

Riepilogando quanto detto quindi:

- L'intervallo di confidenza convenzionale ignora il termine di distorsione ed è centrato sullo stimatore ottenuto tramite approssimazione polinomiale locale ed usa l'errore standard convenzionale $\sqrt{\hat{V}}$;
- L'intervallo con correzione della distorsione, $\mathrm{CI}_{\mathrm{bc}}$, sottrae la stima del termine di distorsione alla stima dell'effetto del trattamento e per l'errore standard utilizza una varianza che ignora la variabilità introdotta dalla stima del termine di distorsione;
- L'intervallo $\mathrm{CI}_{\mathrm{rbc}}$ è centrato sulla stessa quantità di $\mathrm{CI}_{\mathrm{bc}}$ ma utilizza un termine di varianza che incorpora il *bias* introdotto dalla stima del termine di distorsione.

4.2.4 Utilizzare diverse ampiezze di banda per stima ed inferenza

Abbiamo descritto precedentemente i problemi in cui si incorre nel momento in cui si utilizza h_{opt} per la costruzione di intervalli di confidenza convenzionali. Si descrive adesso un'alternativa che consiste nell'uso di due differenti valori dell'ampiezza di banda, uno relativo alla procedura di stima ed un secondo per quella inferenziale. Un criterio ottimale associato a proprietà di robustezza degli intervalli di confidenza è la minimizzazione dell'errore di copertura, ossia la differenza tra la copertura empirica dell'intervallo di confidenza e il suo livello nominale. Ad esempio, se un intervallo di confidenza al 95% contiene il 90% dei veri parametri, allora l'errore di copertura sarà pari al 5%. Minimizzare questa componente erratica durante la procedura inferenziale, è analogo al minimizzare l'errore quadratico medio in fase di stima. Un approccio alternativo consiste allora nell'utilizzare due ampiezze di banda, dove ogni parametro ha uno scopo differente ed in particolare uno concerne la procedura di stima, l'altro quella inferenziale. Si procede nella seguente maniera:

1. si stima l'effetto del trattamento utilizzando h_{opt};
2. si costruiscono intervalli di confidenza con un'altra ampiezza di banda, h_{cer}, scelta in modo tale da minimizzare il tasso di errata copertura dell'intervallo ottenuto col metodo di correzione robusta della distorsione.

Si avrà quindi che h_{opt} minimizza l'errore quadratico medio, mentre h_{cer} minimizza il tasso di errata copertura dell'intervallo $\mathrm{CI}_{\mathrm{rbc}}$. Si può dimostrare che h_{cer} ha un veloce

tasso di decadimento rispetto ad h_{opt}, e ciò implica che, per campioni sufficientemente grandi, $h_{cer} < h_{opt}$.

4.3 Convalida e falsificazione del disegno RD

Una delle caratteristiche principali del *Regression Discontinuity Design*, nonché uno dei principali vantaggi, è che il meccanismo attraverso il quale una campagna o un trattamento viene assegnato è noto e basato su caratteristiche osservabili. Il meccanismo di assegnazione del trattamento basato sullo *score* e sul *cutoff* non è tuttavia autosufficiente per garantire che le assunzioni richieste vengano soddisfatte. Poniamo il caso in cui le unità oggetto di studio abbiano la possibilità di conoscere il valore della soglia che determina lo status di trattamento. Queste potrebbero manipolare o cambiare i loro punteggi, soprattutto se questi si trovano appena al di sotto della soglia, in modo tale da poter usufruire del trattamento. Per valutare la validità di un disegno di *Regression Discontinuity* vi sono diversi metodi empirici, che sotto determinate assunzioni, possono fornire utile evidenza riguardo la plausibilità delle assunzioni del disegno. In particolare, tratteremo tre tipologie di test che verificano la validità del disegno, che denomineremo con:

1. densità della variabile punteggio;
2. soglie artificiali; ;
3. sensibilità alle osservazioni vicine alla soglia;

4.3.1 Densità della variabile di punteggio

Questo test valuta se sono presenti differenze significative riguardo il numero di unità che si trovano in corrispondenza della soglia. L'assunzione sottostante è che, se le unità non hanno la possibilità di manipolare in modo preciso il valore del loro punteggio ottenuto, allora il numero di unità assegnate al gruppo di trattamento, appena sopra la soglia, dovrebbe essere simile al numero di osservazioni assegnate al gruppo di controllo, che si trovano appena sotto la soglia. Applicazioni nelle quali si nota invece un salto brusco del numero di osservazioni all'oltrepassare della soglia non sono credibili (McCrary, 2008). In Figura 4.1 si mostrano due casi, uno in cui si ha un numero simile di osservazioni, ossia un disegno valido, ed un altro in cui vi è una netta differenza tra osservazioni appartenenti ai due gruppi appena in corrispondenza del *cutoff*.

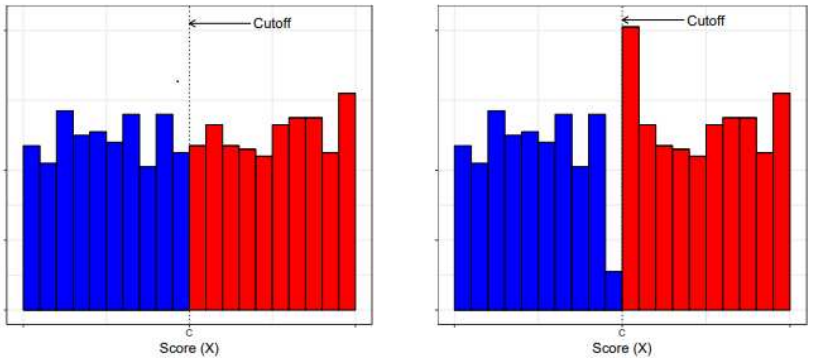


FIGURA 4.1: Test densità variabile di punteggio

Nel grafico di destra si nota un grande salto per quanto concerne il numero di osservazioni dei due gruppi, ciò rende quindi il disegno poco credibile. Oltre ad un mero test grafico, per essere più formali, si può procedere all'esecuzione di un test, detto di densità. Una possibile strategia è scegliere un piccolo intorno della soglia e applicare un test di Bernoulli all'interno di questo intervallo, con probabilità di successo pari a 0.5. Questa strategia verifica se il numero di osservazioni facenti parte del gruppo di trattamento nell'intorno scelto è compatibile con ciò che sarebbe stato osservato se le unità fossero state assegnate al gruppo trattamento con una probabilità del 50%. Non vi sono metodologie specifiche per quanto concerne la scelta dell'intorno della soglia. Una maniera di procedere consiste nel condurre ripetutamente il test per diversi intorni innestati attorno al *cutoff*. Un approccio complementare consiste nel testare l'ipotesi nulla che la densità della variabile di *score* è continua in corrispondenza della soglia. L'implementazione di questo test richiede la stima della densità delle osservazioni vicine al *cutoff*, separatamente per le osservazioni alla destra ed alla sinista della soglia. Si avrà dunque il seguente sistema di ipotesi:

$$\begin{cases} H_0: & \text{non vi è manipolazione di } X_i \\ H_1: & \text{vi è manipolazione di } X_i \end{cases}$$

Rigettare H_0 significa affermare che sia presente alterazione nei valori di X_i e che quindi, il disegno non è valido.

4.3.2 Soglie artificiali

In questo test, i ricercatori modificano artificialmente il *cutoff*, ad esempio spostandolo leggermente a destra o a sinistra del suo vero valore, e verificano se i risultati dell'analisi cambiano o meno in modo significativo (Valentim et al., 2021). Nel test con soglie artificiali quindi, la validità del disegno può essere valutata verificando se i risultati ottenuti, utilizzando la soglia scelta o quella artificiale, sono simili o meno, poiché scegliendo valori vicini a quelli del valore originale, ci si aspetta che l'effetto del trattamento sia simile per le due analisi. Se invece i risultati dell'analisi cambiassero in modo significativo a seguito di variazioni nel valore della soglia, allora ciò potrebbe indicare che il disegno non è valido o che ci sono problemi di robustezza. La scelta del valore della soglia artificiale dipende dall'obiettivo dell'analisi e dalle caratteristiche dei dati. Un approccio comune consiste nell'utilizzare un *cutoff* con valore molto vicino a quello originale. In questo modo, è possibile verificare se i risultati cambiano significativamente quando si utilizza una soglia leggermente diversa. In generale, la scelta del valore della soglia artificiale dovrebbe essere effettuata in modo da garantire che l'analisi sia robusta e che i risultati non siano influenzati da scelte arbitrarie della soglia.

4.3.3 Sensibilità alle osservazioni vicine alla soglia

Questo approccio di falsificazione cerca di indagare quanto i risultati siano sensibili alle unità che si trovano molto vicine al *cutoff* (Van der Klaauw, 2008). L'idea alla base è quindi quella di escludere tali unità e quindi ripetere le procedure di stima e di inferenza utilizzando il campione rimanente. Questa idea viene indicata con l'espressione *donut hole*. Anche nel caso in cui non vi sia il sospetto di un'effettiva manipolazione della variabile X_i, questo approccio risulta essere molto utile, poiché permette di valutare la sensibilità dei risultati ottenuti dopo aver tolto quelle unità che nella stima polinomiale locale sono le più influenti, ossia quelle più vicine alla soglia. Bisognerà utilizzare una nuova ampiezza di banda ed inoltre è importante sottolineare come non sia scontato che, eliminando le osservazioni in prossimità della soglia, le unità incluse nei processi di stima ed inferenza siano in numero inferiore a quelle utilizzate con tutto il campione a disposizione, perchè avendo scelto un nuovo h, potrebbero essere incluse più unità rispetto alla prima fase. Pertanto, se la rimozione delle osservazioni vicino alla soglia non influisce significativamente sui risultati, allora il disegno è robusto e i risultati ottenuti sono affidabili. In alternativa, se questa rimozione porta a risultati significativamente diversi, ciò potrebbe indicare che il disegno non è valido o che ci sono problemi di robustezza.

Capitolo 5

Applicazione: analisi dell'efficacia di una campagna di marketing

5.1 Introduzione, contesto e obiettivi dell'analisi

L'analisi effettuata concerne l'utilizzo della metodologia *Regression Discontinuity Design*, allo scopo di valutare l'efficacia di una campagna di marketing. In particolare si andrà a valutare ed analizzare la probabilità di acquisto di un particolare prodotto. Tramite le prime analisi esplorative si osserverà infatti come le due modalità della variabile risposta dicotomica *successo della campagna* siano decisamente sbilanciate. Questo si rifletterà anche sulla probabilità di acquisto del prodotto, che assumerà valori bassissimi, tendenti quasi a zero. Ciò significa che le differenze riscontrabili tra i due gruppi saranno molto basse, poiché come detto, a prescindere dal fatto che un'unità si trovi alla destra o alla sinistra della soglia, la probabilità di acquisto rimarrà molto bassa. La variabile di punteggio, in questo caso specifico, rappresenta una valutazione del cliente da parte di uno specifico modello di *propensity*, modello utilizzato dall'azienda per l'individuazione dei migliori clienti, ed, in particolare, in questo studio la soglia è stata fissata ad un valore pari ad 1, poiché l'azienda ha ritenuto questo essere il giusto valore al fine di discriminare i migliori clienti dal resto delle unità presenti. Il primo obiettivo dell'analisi è, quindi, quello di verificare se il *Regression Discontinuity Design* risulta essere un metodo adatto al caso in questione, avendo a disposizione come variabile risposta una variabile dicotomica e non una variabile quantitativa, come di solito accade nei contesti in cui si usa il metodo e come è stato mostrato nelle precedenti sezioni teoriche. In secondo luogo si vuole verificare se la campagna di marketing effettuata presenta un effetto, ciò significa valutare se aver assegnato o meno la campagna comporta un

cambiamento nelle probabilità di acquisto del prodotto da parte dei clienti. Si andrà in terzo luogo a valutare se esistono differenze significative in termini di probabilità di acquisto, in corrispondenza della soglia. Tutte le analisi effettuate sono state svolte col *software* `R`.

5.2 I dati

L'insieme di dati sul quale sono state effettuate le analisi concerne in particolar modo una campagna di marketing effettuata da una banca al fine di promuovere uno specifico prodotto di investimento. Inizialmente si avevano a disposizione due *dataset*: il primo contenente alcune variabili esplicative, il secondo contentente la variabile risposta *successo della campagna* e la corrispondente data di apertura del contratto relativo alla prestazione acquisita. Il *dataset* iniziale conteneva 109.412 osservazioni relative a 14 variabili. Di seguito si fornisce una breve descrizione delle variabili contenute al suo interno:

1. ***id***: identificativo del cliente;
2. ***prob***: probabilità di acquisto del prodotto da modello di *propensity* della banca;
3. ***percentile***: percentile della variabile *prob*;
4. ***score***: punteggio ottenuto dal cliente;
5. ***dtms esito***: data del contatto commerciale;
6. ***cod trt***: identificativo della campagna commerciale dove solo il codice 1693 è quello associato al modello di *propensity*;
7. ***trt modello***: 1 se *codtrt* è quello del modello, 0 altrimenti;
8. ***cod canale delivery***: canale di contatto utilizzato per la campagna;
9. ***cod cell package***: campo tecnico di campagna;
10. ***contatto maggio***: 1 se il cliente è stato contattato a maggio sul prodotto specifico da qualsiasi campagna, 0 altrimenti;
11. ***contatto mod maggio***: 1 se il cliente è stato contattato a maggio sul prodotto specifico, dalla campagna derivante dal modello di *propensity* del mese precedente, 0 altrimenti;

12. ***successo***: 1 successo della campagna, 0 insuccesso della campagna;
13. ***data apertura contratto***: data apertura del contratto;
14. ***des cluster***: non si ha una descrizione della variabile.

La prima azione effettuata sull'insieme di dati è stata quella di eliminare le variabili *des cluster* e *cod cell package*, poichè non ritenute utili ai fini dell'analisi, sotto suggerimento dell'azienda. Successivamente è stata creata la variabile dicotomica W, ossia lo status di trattamento, dove la modalità 1 si riferisce al fatto che l'unità in questione abbia almeno raggiunto la soglia, viceversa la modalità 0 indica il fatto questa non sia stata raggiunta. La tipologia di disegno applicata in questa analisi è il *Fuzzy RD*, si assegna quindi la campagna solo ad una percentuale di unità che hanno raggiunto la soglia, in maniera casuale. Sono stati osservati numerosi dati mancanti che in questo specifico contesto corrispondevano ad una mancata azione di marketing, motivo per il quale sono state eliminate le relative osservazioni. Alla fine di queste semplici operazioni preliminari sui dati, il *dataset* finale su cui verranno svolte tutte le successive analisi risulta composto da: 61.547 osservazioni relative a 13 variabili. Molte di queste variabili però non potranno essere utilizzate per le analisi in questione. In particolare alcune di queste sono state utilizzate per costruire la variabile *score*, come *percentile* e *prob*. Altre variabili, quali *cod cell package* e *des cluster* sono state ritenute dall'azienda variabili da non utilizzare. La variabile relativa alla data di apertura del contratto non è stata invece ritenuta utile ai fini dell'analisi. Nello specifico le uniche covariate utilizzabili, oltre alla variabile risposta, sono quelle relative al canale di contatto del cliente e al precedente contatto nel mese di maggio.

5.3 Analisi esplorative

5.3.1 Variabile risposta

Il grafico riportato in Figura 5.1 mostra la distribuzione di frequenza della variabile risposta:

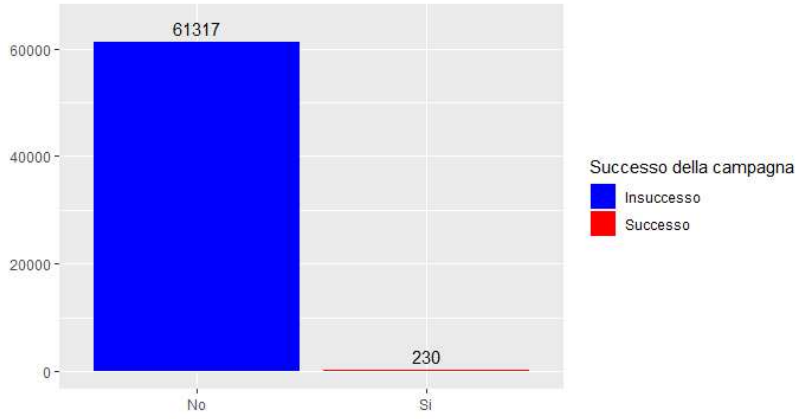


FIGURA 5.1: Istogramma variabile risposta

Dall'istogramma è possibile notare che il successo della campagna si manifesta, nel campione a disposizione, solamente per 231 volte rispetto ad un totale di 61.547 unità. Questo grafico riesce, quindi, in maniera molto semplice ad evidenziare la presenza di una classe rara che ha generato non poche difficoltà nella gestione dei dati. Per rendere ancor di più l'idea, nella Tabella 5.1 vengono mostrate le frequenze relative per la variabile risposta.

	Percentuale
Campagne non di successo	99.63%
Campagne di successo	0.37%

TABELLA 5.1: Frequenze relative variabile risposta

Da essa si può osservare come solamente nello 0.37% dei casi si è registrato l'acquisto del prodotto da parte del cliente. Avendo osservato un tale disequilibrio tra le due modalità, ci si aspetta che le probabilità di successo stimate dai modelli saranno molto basse, coerentemente con quanto detto precedentemente.

5.3.2 Raggiungimento della soglia

Come detto in precedenza, l'azienda ha deciso di fissare il valore del *cutoff* pari ad 1 e ciò permette di discriminare, all'interno dell'insieme di dati a disposizione, i

clienti migliori dal resto delle unità incluse. Denominiamo con W la variabile *status di trattamento*, dove $W = 1$ se il valore associato della variabile punteggio è almeno pari ad 1, viceversa si avrà $W = 0$. In Figura 5.2 viene fornita la frequenza assoluta dei due gruppi formatisi tramite il meccanismo di assegnazione della campagna:

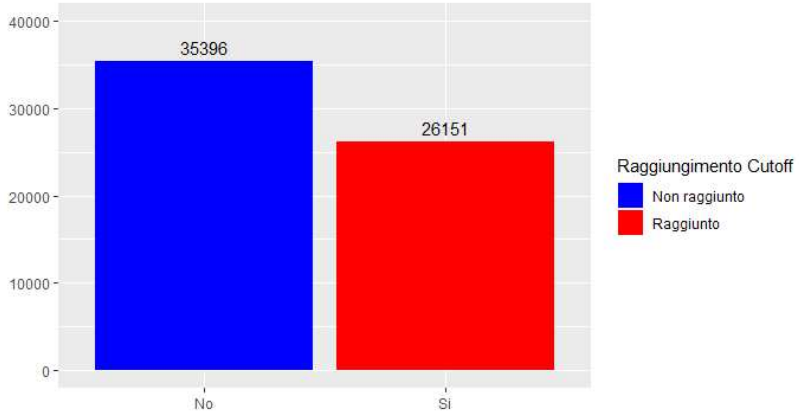


FIGURA 5.2: Raggiungimento *cutoff*

Da essa si nota che sono 35.396 le unità del nostro insieme di dati che possiedono un valore della variabile *score* inferiore ad 1, ossia quella che non hanno ricevuto la campagna di marketing, non avendo raggiunto la soglia prefissata. Sono invece 26.151 quelle che hanno almeno raggiunto la soglia. La tipologia di disegno applicato in questo caso è il *Fuzzy RD*, ragion per cui solo alcune di queste 26.151 unità riceveranno la campagna di marketing. In particolare di queste, si è verificato come solo 3.298 unità abbiano ricevuto la campagna. Potrebbe risultare di più semplice lettura osservare le percentuali di unità appartenenti ad ognuno dei due gruppi, fornite dalla Tabella 5.2. Essa ci mostra come le numerosità dei due gruppi siano abbastanza simili, anche se,

	Percentuali
Unità che non hanno raggiunto la soglia	57.52%
Unità che hanno almeno raggiunto la soglia	42.48%

TABELLA 5.2: percentuali di raggiungimento della soglia

visto che l'obiettivo sarebbe quello di discriminare i migliori clienti dal resto, ci si poteva aspettare una percentuale minore di unità appartenenti al gruppo di coloro che hanno ricevuto la campagna. Un'altra informazione utile potrebbe essere quella di quantificare la distribuzione di frequenza dei successi nei due gruppi, ottenuti grazie alla divisione in controllo e trattamento per mezzo della variabile *status di trattamento*. Le numerosità dei due gruppi sono quelle indicate nel grafico a barre precedente: 35.396 unità nel

gruppo controllo e 26.151 unità nel gruppo trattamento. La Tabella 5.3 mostra le frequenze di campagne di successo ed insuccesso nei due gruppi, e in termini percentuali

	Campagne non di successo	Campagne di successo
Gruppo controllo	35295	101
Gruppo trattamento	26022	129

Tabella 5.3: Frequenza di successi nei gruppi trattamento e controllo

nella Tabella 5.4.

	Campagne non di successo	Campagne di successo	
Gruppo controllo	99.72%	0.28%	100%
Gruppo trattamento	99.52%	0.48%	100%

Tabella 5.4: Percentuali di successi nei gruppi trattamento e controllo

Nonostante il numero di campagne di successo sia molto basso, in entrambi i gruppi si possono scovare delle piccole differenze tra questi, anche solo grazie a questa semplice tabella. Notiamo, infatti, come la frequenza di successi nel gruppo di unità che hanno almeno raggiunto la soglia sia maggiore rispetto a quella relativa alle unità che non hanno raggiunto il *cutoff*.

5.3.3 Modalità di contatto

Successivamente si è deciso di valutare la distribuzione di frequenza dei canali di comunicazione all'interno del sottoinsieme di dati contenente solamente le informazioni relative ai clienti per i quali si è registrato un successo della campagna e ne viene fornita una rappresentazione grafica in Figura 5.3.

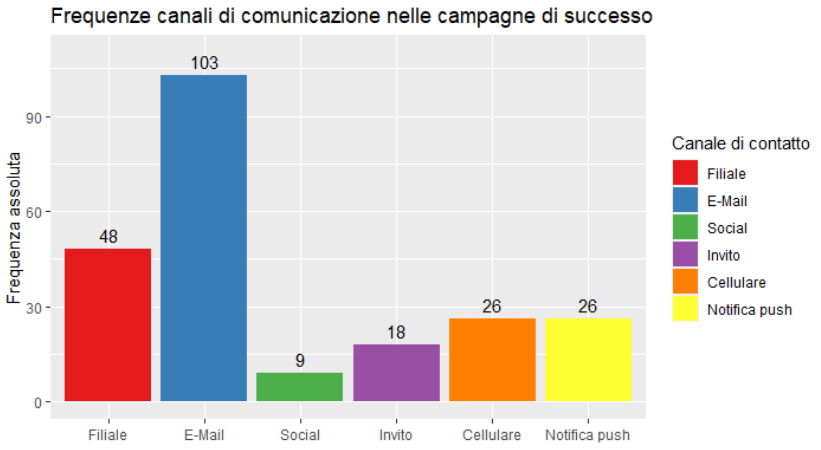


FIGURA 5.3: Canali di comunicazione

Si osserva come nel *dataset* relativo alle sole campagne di successo, l'*e-mail* sia la modalità di contatto più frequente, seguita dal contatto tramite filiale. Successivamente si è deciso di valutare la stessa distribuzione di frequenza ma dividendo ulteriormente l'insieme di dati precedente in due *dataset*, a seconda che la campagna sia stata assegnata o meno. In questo modo si vuole provare a scovare la presenza di eventuali canali che potrebbero propendere maggiormente all'interno di uno dei due gruppi. I risultati vengono mostrati in Figura 5.4:

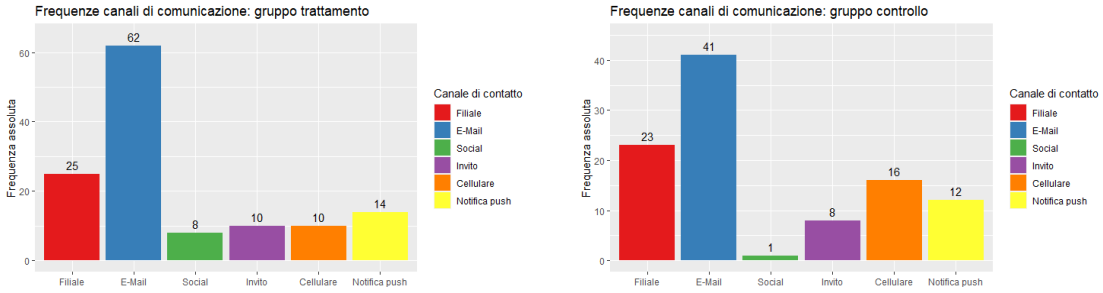


FIGURA 5.4: Canali di comunicazione nei gruppi trattamento e controllo

Osservando la distribuzione di frequenza dei canali di comunicazione nei gruppi trattamento e controllo non si notano particolari differenze tra i due gruppi. L'unica nota

interessante potrebbe essere quella che la modalità di comunicazione *social* risulta presente solo una volta, nel gruppo di coloro che non hanno raggiunto la soglia, rispetto alle 8 volte in cui risulta nel gruppo trattamento. Un'ulteriore differenza riscontrabile risiede nel fatto che la notifica *push*, ad esempio, passa da essere la terza modalità più frequente nel gruppo controllo, alla quarta nel gruppo a cui è stata assegnata la campagna.

5.3.4 Precedente contatto a maggio da modello di *propensity* o da altre campagne

Un'altra analisi utile per capire al meglio le caratteristiche dei nostri dati è quella sulle distribuzioni di frequenza, nell'insieme di dati che include solamente le informazioni relative alle unità per le quali si è giunti ad un successo per la campagna, relative:

- all'aver contattato il cliente a maggio tramite qualsiasi modello;
- all'aver contattato il cliente a maggio tramite il modello di *propensity*, ossia il modello tramite il quale l'azienda ha deciso di fissare il *cutoff* ad 1.

Il grafico a barre riportato in Figura 5.5 mostra le distribuzioni di frequenza relativa alla prima delle due variabili precedentemente menzionate.

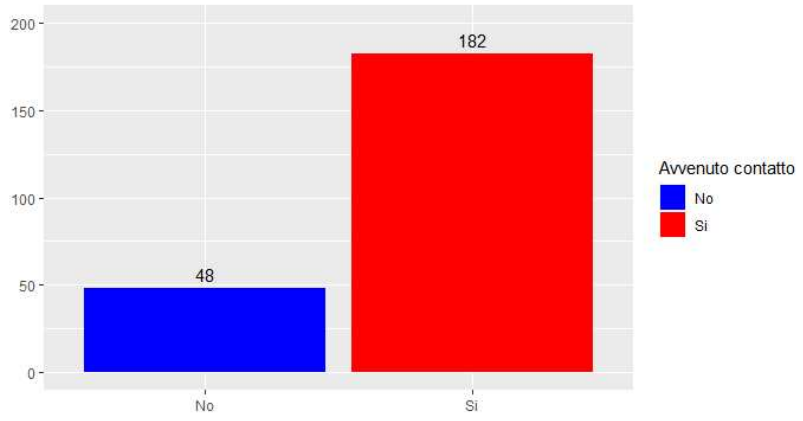


FIGURA 5.5: Contatto a maggio nell'insieme delle campagne di successo

Si nota come delle 231 campagne di successo presenti all'interno del nostro *dataset*, ben 182 siano derivanti da un precedente contattato avvenuto nel mese di maggio. Si può quindi pensare che 48 sia il numero di clienti che ha effettuato un acquisto che può essere definito spontaneo.

Il contrario accade per quanto riguarda sempre un contatto a maggio, ma secondo quanto indicato dal modello di *propensity*.

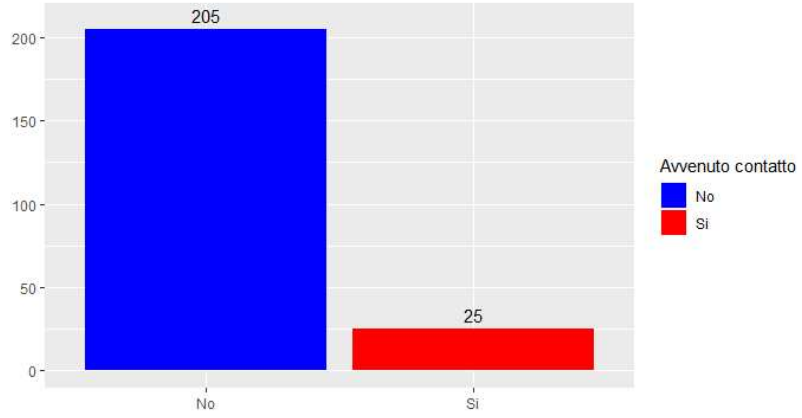


FIGURA 5.6: Contatto a maggio da modello nell'insieme delle campagne di successo

In Figura 5.6 si mostra infatti come delle 230 campagne di successo, solo 25 derivano da un precedente contatto a maggio dal modello di *propensity*.

5.3.5 Variabile punteggio

In ultima istanza si analizza la distribuzione di frequenza della variabile punteggio. In Figura 5.7 viene riportato l'istogramma relativo alla distribuzione di frequenza della variabile punteggio.

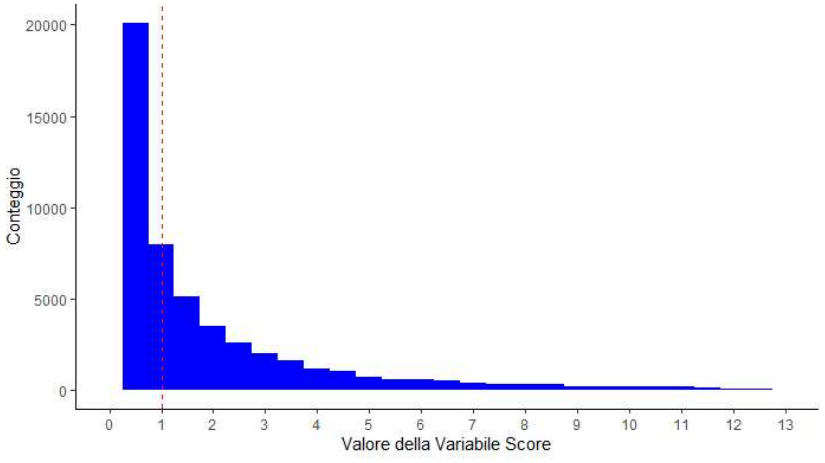


FIGURA 5.7: Istogramma variabile punteggio

Da essa si nota come la maggior parte delle unità abbiano ottenuto un punteggio compreso tra 0 e 3. In particolare sono pochissime le unità che hanno ottenuto punteggi

maggiori di 5. Successivamente si valuta se la numerosità dei due gruppi creatisi sia simile in un intorno della soglia, in modo tale da verificare se il disegno è valido o se vi è stata manipolazione della variabile di punteggio, come mostrato nella corrispondente sezione teorica. Il seguente grafico ci permette di valutare quanto detto.

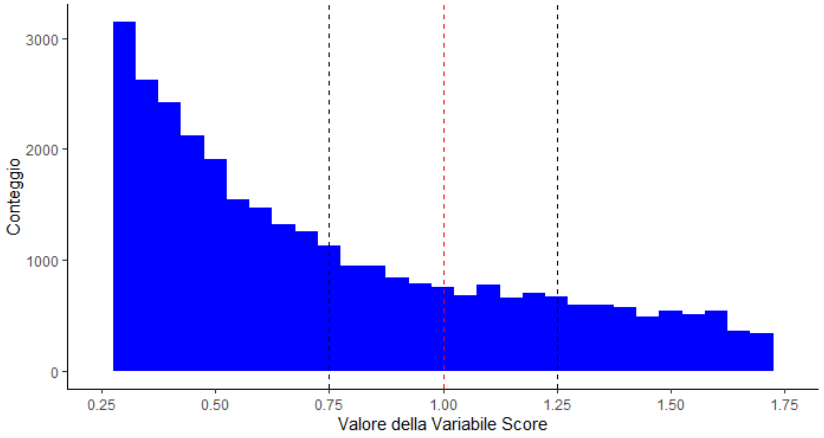


FIGURA 5.8: Verifica validità del disegno: primo intorno della soglia

Dalla Figura 5.8 si può vedere come le unità, in un primo intorno della soglia ($[0.75; 1.25]$) delimitato dalle linee nere tratteggiate, appartenti al gruppo di controllo $[0.75; 1)$ e al gruppo trattamento $[1; 1.25]$, abbiano numerosità simili, ragion per cui non sembra esserci stata manipolazione della variabile di punteggio. Successivamente, si è provato ad allargare l'intervallo attorno alla soglia, per confermare quanto appena detto. Si mostra quindi la Figura 5.9.

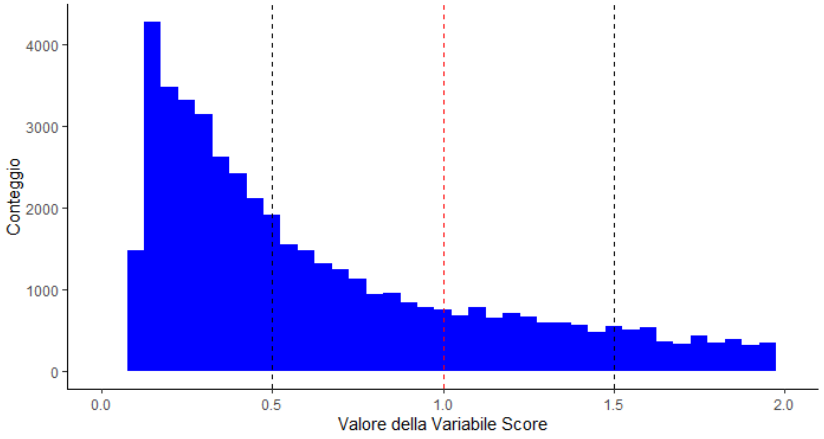


FIGURA 5.9: Verifica validità del disegno: secondo intorno della soglia

Si nota come le numerosità siano diverse ed in particolare sembrano esserci più unita a sinistra della soglia. Ciò non rende invalido il disegno perchè, nonostante le differenze

non siano minimali, questi risultati appaiono coerenti con la distribuzione della variabile punteggio. Inoltre, la non validità avrebbe portato ad avere più osservazioni nel gruppo di trattamento e non di controllo, al contrario di quanto accade nel caso specifico che stiamo considerando.

5.4 Istogrammi su intorni del *cutoff*

In questo paragrafo ci si concentra sull'analisi della distribuzione di frequenza delle campagne di successo in base ai valori del *cutoff*, per poi andare nello specifico a valutare tale distribuzione in un intorno della soglia. Si sottoseleziona quindi un *dataset*, in cui sono contenute solo le unità corrispondenti a campagne di successo. Gli intervalli in cui è stata ripartita la variabile continua *score* sono di lunghezza diversa perché utilizzando un'unica misura si sarebbero avuti intervalli con numerosità troppo diverse. In prima istanza viene mostrato in Figura 5.10 un grafico a barre che mostra la distribuzione delle campagne di successo in base ai valori assunti dalla variabile punteggio.

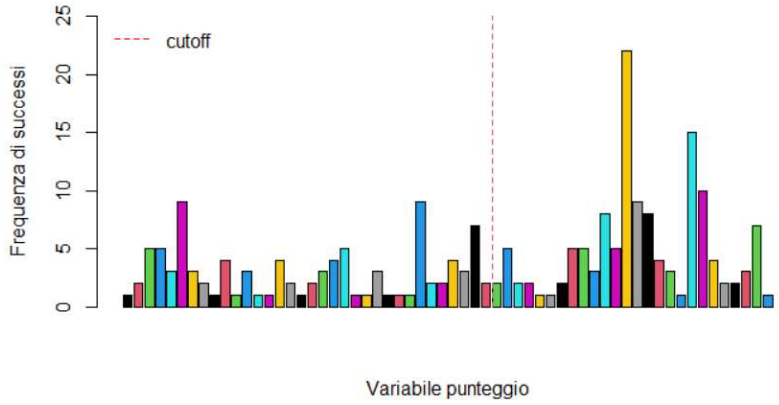


FIGURA 5.10: Frequenze dei successi per valori della variabile punteggio

Si può osservare come la frequenza delle campagne di successo sia abbastanza simile per valori della variabile punteggio, per poi avere un picco in corrispondenza di valori dello *score* pari a 3. La linea verticale rossa tratteggiata indica la soglia che, come detto precedentemente, è stata fissata pari ad 1. Successivamente, in Figura 5.11, si valuta la stessa distribuzione ma in un intorno $I_1 = [0.5; 4]$ del *cutoff*, quindi per un sotto-intervallo di valori relativi allo *score*.
Il grafico in questione non aggiunge molte informazioni rispetto a quanto detto già prima. Anche in questa circostanza si nota inizialmente una distribuzione simile per poi

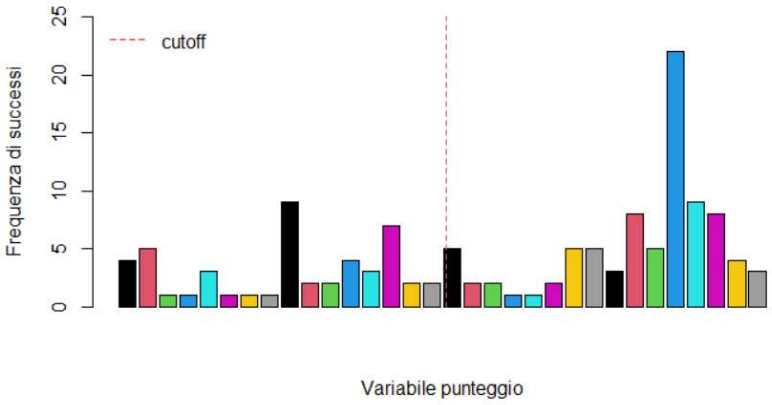


FIGURA 5.11: Frequenze dei successi per valori della variabile punteggio: primo intorno

osservare un picco (come nel caso precedente), sempre in corrispondenza di valori della variabile punteggio in corrispondenza di 3. In ultima istanza si restringe ulteriormente l'intorno, come in Figura 5.12 in cui viene considerato l'intorno $I_2 = [0.75; 1.5]$, per valutare se sia possibile ottenere informazioni più rilevanti:

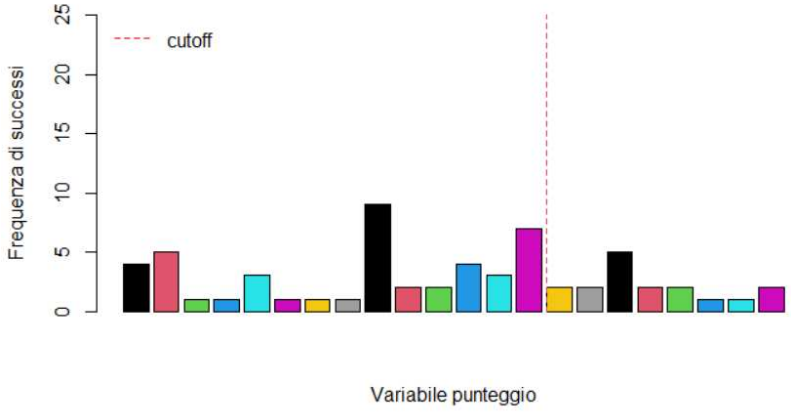


FIGURA 5.12: Frequenze dei successi per valori della variabile punteggio: secondo intorno

Qui è possibile notare come la distribuzione nel nuovo intorno considerato sia molto simile, alla destra e alla sinistra della soglia. Si può allora ipotizzare che la scelta di fissare la soglia ad 1 possa non comporti un effetto significativo nelle probabilità di successo della campagna. Ciò andrà chiaramente valutato nella successiva fase di analisi.

5.5 Modelli univariati

Per valutare l'effetto della campagna di marketing, in questo paragrafo, verranno adattati vari modelli, contenenti ognuno un'unica variabile esplicativa, lo *score*. In particolare verranno adattati, ad eccezione del primo, tutti modelli non parametrici e locali, allo scopo di modellare in modo accurato la relazione tra la risposta e la variabile esplicativa, considerando un intorno locale dei punti. Nell'ordine si mostreranno i risultati ottenuti tramite i seguenti modelli:

1. Modello di regressione logistica;
2. Modello di regressione logistica locale;
3. Modello di regressione *loess* logistico;
4. *Splines* di regressione logistiche;
5. *Splines* di lisciamento logistiche;
6. Albero di regressione.

Per ogni modello adattato verrà inoltre fornito il grafico tramite il quale si valuterà l'effetto della campagna di marketing, rappresentante le probabilità di acquisto relativamente ai valori della variabile punteggio, per ogni lato della soglia. Le stime dell'effetto della campagne ed i relativi intervalli di confidenza ottenuti verranno mostrati nel seguente Capitolo, in modo tale da fornire conclusioni legate ai risultati delle analisi effettuate. Si ricorda inoltre che la tipologia di disegno adottata dall'azienda è il *Fuzzy Regression Discontinuity*. Come visto nella sezione teorica concernente le varie tipologie di disegno di *Regression Discontinuity*, questa tipologia non assegna il trattamento a nessuna unità che non abbia almeno raggiunto la soglia. Tra le unità che invece presentano un punteggio almeno pari al *cutoff*, il disegno ne seleziona causalmente una porzione alla quale assegnare la campagna. Si ha dunque che solamente ad una parte delle unità che hanno raggiunto la soglia viene assegnata la campagna. In particolare si è visto come queste siano in numero pari a 3.298 (12.61%) su un totale di 26.151 unità con punteggio almeno pari alla soglia.

5.5.1 Regressione logistica

Il primo modello adattato è, come detto precedentemente, il modello di regressione logistica, ovvero il più semplice dei modelli di regressione per quanto concerne una variabile risposta dicotomica. L'approccio del modello è di tipo parametrico e globale. Il modello prevede come variabile risposta il successo della campagna e come variabile esplicativa lo *score*. Tramite questo modello, si giunge al grafico riportato in Figura 5.13, attraverso il quale si può dare una valutazione dell'effetto della campagna di marketing.

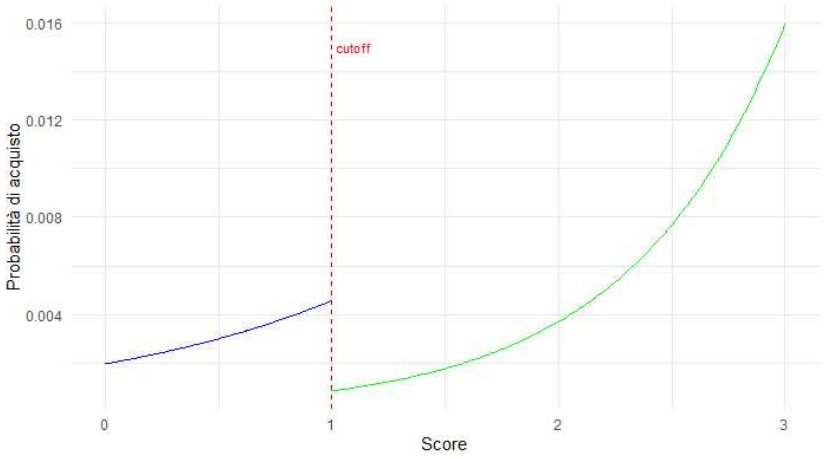


FIGURA 5.13: Valutazione efficacia: regressione logistica

Come possibile osservare dal grafico, la probabilità di acquisto al *cutoff* sembra essere maggiore per quelle unità che non hanno ricevuto la campagna. In generale sembra esserci un piccolo effetto della campagna ed in particolare questo sembra essere negativo. Ciò si può evincere dal fatto che il salto che si nota in corrispondenza del *cutoff* è direzionato verso il basso.

5.5.2 Regressione logistica locale

Il secondo modello adattato è la regressione logistica locale. Anche questa modellazione segue un approccio non parametrico e locale. Il parametro di regolazione rappresentato dall'ampiezza di banda è stato selezionato tramite convalida-incrociata ed il suo valore è pari a 0.93. È stato ritenuto opportuno utilizzare lo stesso valore di h in entrambi gli intorni della soglia. Per adattare il modello è stata utilizzata la funzione *sm.binomial* della libreria *sm* (Bowman & Azzalini, 1997). Tramite regressione logistica locale si è giunti al risultato mostrato in Figura 5.14.

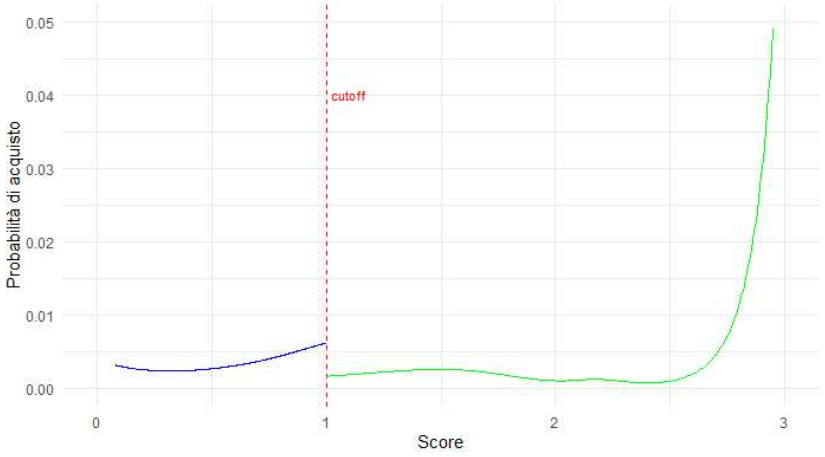


FIGURA 5.14: Valutazione efficacia: regressione logistica locale

Il modello di regressione logistica locale, come il precedente modello, ci porta ad affermare che sembra essere presente un piccolo effetto negativo della campagna. Tutto ciò relativamente alle osservazioni in corrispondenza della soglia. Guardando solamente i grafici questo sembra essere più piccolo di quello previsto dal modello parametrico, ma bisogna attenzionare il fatto che i due grafici abbiano diversi valori degli assi e che quindi in realtà il salto ottenuto tramite regressione logistica locale sembra essere più pronunciato.

5.5.3 *Loess* logistico

Successivamente, si decide di adattare un modello di regressione *loess* logistico. Nei metodi non parametrici e locali adattati finora si è imposta un'ampiezza di banda che, dopo essere stata selezionata tramite convalida incrociata, è stata mantenuta fissa. Dal momento che la variabilità della stima dipende inversamente dalla densità della variabile punteggio, in molti casi è vantaggioso permettere che h sia variabile. La regressione *loess* viene effettuata prefissando una percentuale costante di punti rilevanti da includere (Azzalini & Scarpa, 2009): questo significa che l'ampiezza di banda viene ampliata o contratta, in base alla concentrazione locale dei valori dello *score*. In generale, i *kernel* con supporto limitato sono più adatti alla regressione *loess*, visto che distinguono chiaramente tra punti utilizzati e non utilizzati. Come *default*, `R` usa il kernel biquadratico per la stima. Dal momento che si rischia di ottenere una stima poco robusta rispetto agli outlier e influenzata da osservazioni lontane, la procedura di stima non si basa sui minimi quadrati, bensì su un procedimento di stima robusta. Il parametro di regolazione in questo caso è lo *span*, ossia la proporzione dei punti dati utilizzati per stimare la regressione locale in ogni punto. In altre parole, lo *span* controlla la dimensione dell'ampiezza di banda utilizzata per calcolare la stima locale. In generale, valori di *span* più elevati portano a una stima più liscia, ma possono mascherare le variazioni locali dei dati, mentre valori di *span* più bassi portano a una stima più dettagliata ma più "rumorosa". Una differenza tra regressione logistica locale e *loess* logistico è che la prima si concentra sulla stima della probabilità di appartenenza alla classe dicotomica (in questo caso successo o insuccesso della campagna) in punti specifici, mentre il *loess* logistico fornisce una stima continua della probabilità lungo l'intera gamma di valori della variabile esplicativa. Per adattare il modello si utilizza la funzione *gam* (Hastie, 2011) dell'omonimo pacchetto di `R`, specificando l'opzione *lo* relativamente alla variabile punteggio, in modo tale da adattare il *loess*. Il parametro di regolazione *span* è stato ottenuto tramite convalida-incrociata.

Il risultato ottenuto è osservabile in Figura 5.15:

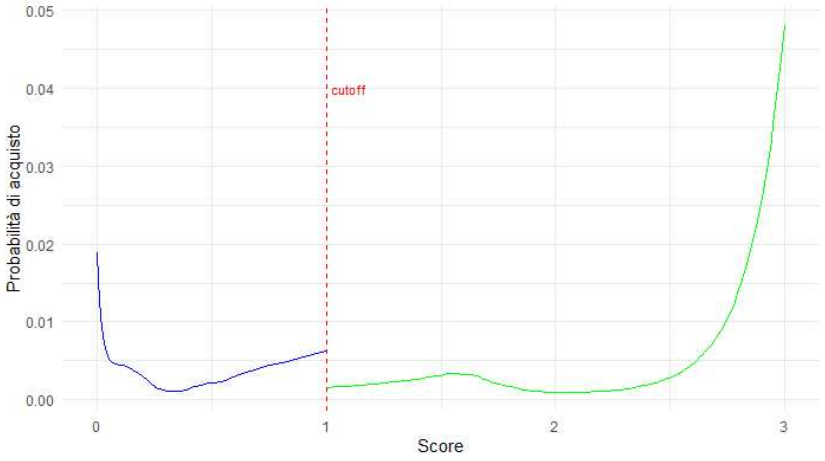


FIGURA 5.15: Valutazione efficacia: *loess* logistico

Tramite questo modello di regressione si evidenzia ancora una volta il salto che avviene in corrispondenza del *cutoff*. L'*output* ottenuto è coerente con quelli dei precedenti modelli adattati. Rispetto al modello di regressione logistica locale, le curve di regressione stimate sembrano essere meno lisce, ciò è dovuto al valore dello *span* selezionato.

5.5.4 *Splines* di regressione

Si passa adesso all'adattamento di *splines* di regressione. La stima della probabilità di acquisto è stata ottenuta utilizzando una funzione logistica, che si basa sulla relazione stimata dalla *spline di regressione*. La funzione logistica prende in *input* il valore stimato dalla *spline* di regressione e restituisce una stima della probabilità di acquisto. Per adattare la *splines* di regressione si è utilizzata la funzione `glm` del pacchetto `stats`, specificando una *spline* cubica per la variabile punteggio. Per quanto concerne questo metodo si ottengono le probabilità stimate riportate in Figura 5.16.

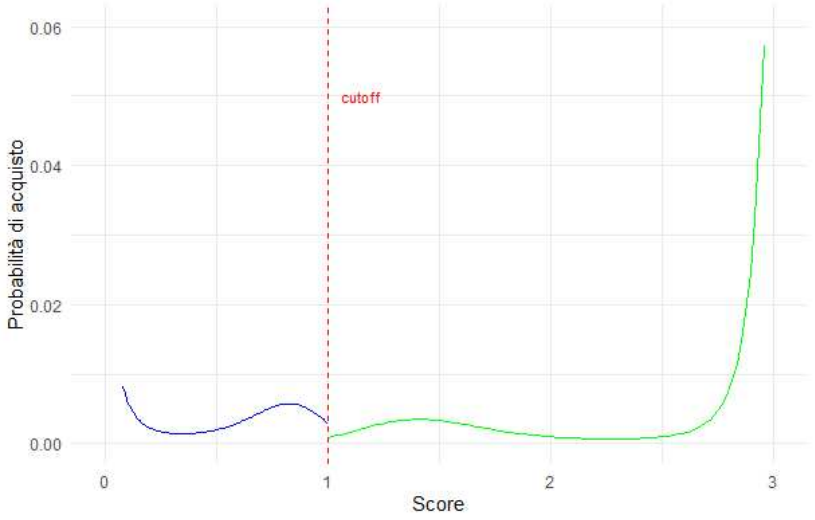


FIGURA 5.16: Valutazione efficacia: splines di regressione

Dal grafico si nota sempre il solito effetto negativo della campagna, ma ulteriormente ridotto rispetto a quanto stimato precedentemente. Un'ulteriore osservazione riguarda la flessione della curva di regressione relativa alla parte destra della soglia in corrispondenza di valori della variabile *score* pari a 2.5. Subito dopo la flessione, avvicinandosi a valori dello *score* pari a 3, si ha un deciso aumento della probabilità di acquisto. Ciò sembra essere coerente per tutti i modelli adattati finora e con quanto detto nelle precedenti fasi di analisi. Per valutare se il valore associato dello *score* pari a 3 sia effettivamente un picco, si dovrebbe valutare l'andamento della seconda curva di regressione per valori dello *score* maggiori di 3 e valutare se questa curva continua a crescere o meno.

5.5.5 *Splines* di lisciamento logistiche

Successivamente si adattano delle *splines* di lisciamento logistiche. Le *splines* di lisciamento consentono di rappresentare la relazione tra le variabili in modo flessibile, ma regolare, senza dover fare assunzioni sul tipo di relazione funzionale tra le variabili. Queste hanno la proprietà di ridurre l'effetto dei valori estremi, ovvero di attenuare l'effetto delle osservazioni che si sulla frontiera. Questa proprietà è particolarmente utile quando si utilizza il *Regression Discontinuity Design*, in cui si cerca di stimare l'effetto di una campagna in corrispondenza della soglia. Per adattare il modello si utilizza la funzione `gam` dell'omonima libreria di `R`, ottenendo il grafico in Figura 5.17.

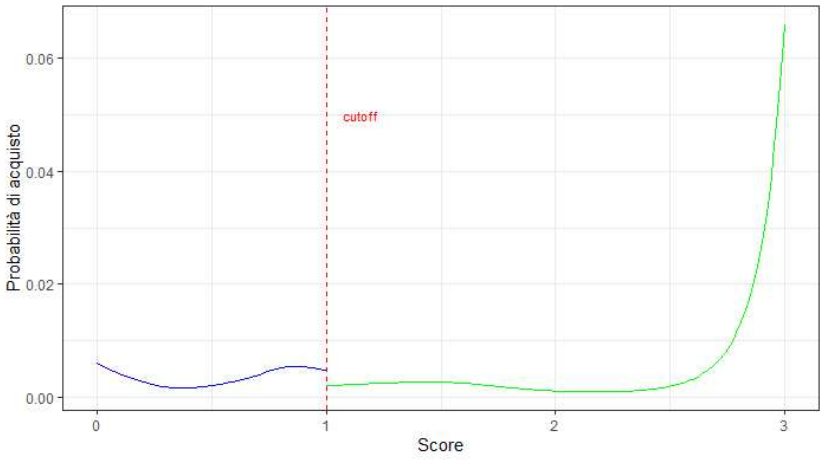


FIGURA 5.17: Valutazione efficacia: *Splines* di lisciamento logistiche

Le conclusioni che si possono trarre osservando il grafico sono le stesse evidenziate dai precedenti modelli adattati. Tramite *splines* di lisciamento logistiche si può verificare come l'effetto stimato della campagna sia minore rispetto ad altri modelli. In particolare, la probabilità di acquisto sembra essere quasi costante per valori della variabile punteggio compresi tra 1 e 2.5.

5.5.6 Albero di regressione

In ultima istanza si decide di adattare un albero di regressione, soprattutto per la facilità di interpretazione che questo fornisce. Si utilizza sempre come unica covariata lo *score*. È necessario scegliere il numero ottimale di foglie per ogni albero adattato, in modo tale da minimizzare la devianza del modello. In questo caso è necessario adattare un albero per ogni lato della soglia. Dopo aver scelto il numero ottimale di foglie per ogni albero, si effettuano le previsioni per ogni lato della soglia. I grafici in Figura 5.18 e Figura 5.19 mostrano la scelta del numero ottimo di foglie, J_1 e J_2, rispettivamente per le unità a sinistra e a destra della soglia.

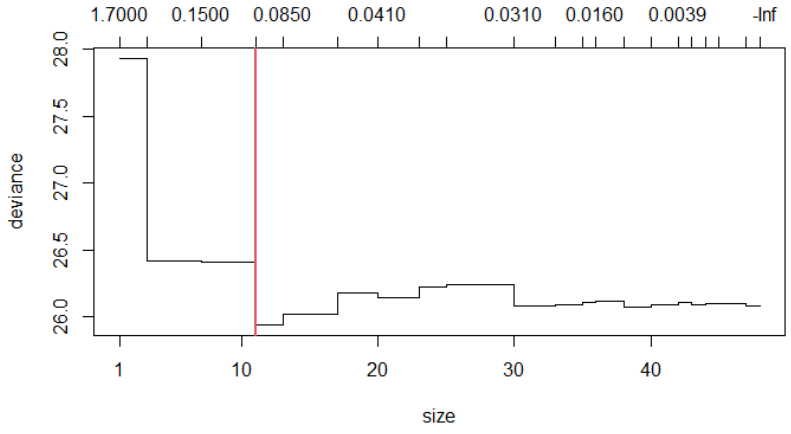


FIGURA 5.18: Numero ottimo di foglie per il primo intorno

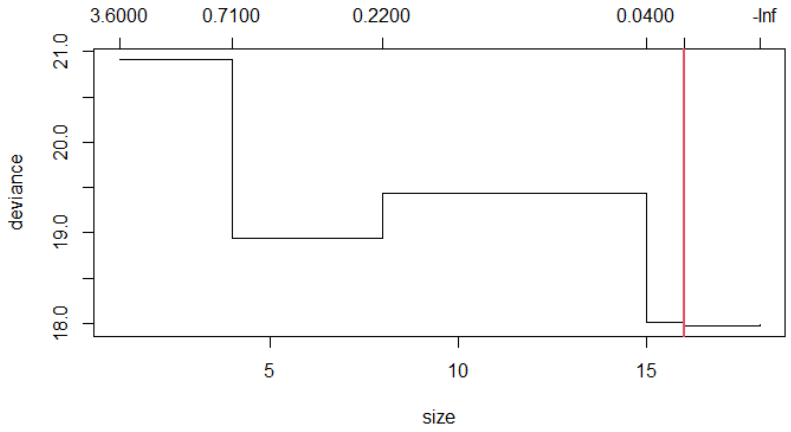


FIGURA 5.19: Numero ottimo di foglie per il secondo intorno

Da questi si può notare come il numero ottimo di foglie per i due alberi di regressione sia rispettivamente 11 e 16.

Scelto il numero ottimale di foglie, si mostrano in Figura 5.20 e Figura 5.21 i grafici relativi agli alberi ottimi:

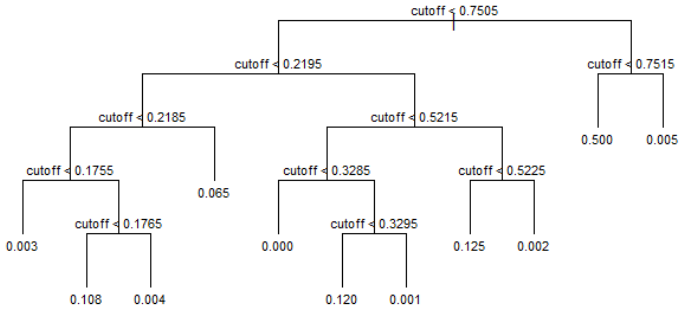


FIGURA 5.20: Albero ottimo per il primo intorno

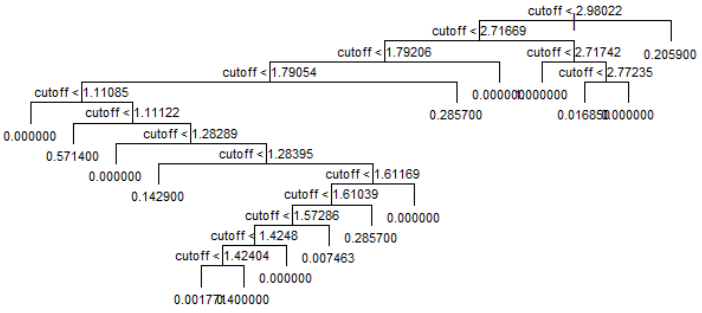


FIGURA 5.21: Albero ottimo per il secondo intorno

Come è possibile notare dai precedenti grafici, l'interpretazione di un modello ad albero è molto semplice. Il funzionamento è il seguente: se entra una nuova osservazione, si valuta il suo valore relativo alla variabile *score* e si scende nella direzione fornita dall'albero in base al valore che si osserva. Una volta fatti degli *split* si giunge alla probabilità che ha quella unità di acquistare il prodotto, ossia qual è la probabilità che la campagna abbia successo.

Infine, si mostra in Figura 5.22 il grafico atto a valutare l'efficacia della campagna.

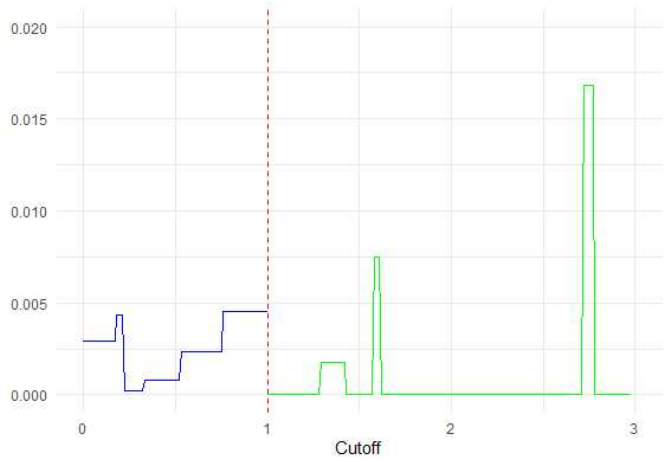


FIGURA 5.22: Valutazione dell'efficacia: albero di regressione

Anche nel caso degli alberi di regressione si nota un salto in corrispondenza della soglia, che corrisponde ad un effetto negativo della campagna. Una possibile ipotesi già fornita precedentemente è che i veri clienti ottimali, identificati in questo caso come le unità aventi un valore dello *score* pari ad almeno 1, siano in realtà coloro che abbiano un valore relativo alla variabile punteggio pari ad almeno 3.

5.6 Modelli non parametrici con residui

L'ultima fase di analisi concerne una fase di modellazione a due stadi. Nella prima fase si adatta modello non parametrico in cui viene utilizzata la variabile successo della campagna come risposta, mentre come covariate si utilizzano *cod canale delivery* e *contatto mod maggio*. Si adatta il modello e si estraggono i residui. Nella seconda fase si prendono i residui ottenuti nella precedente fase e si utilizzano come nuova variabile risposta. Vengono quindi adattati modelli non parametrici quali *loess* e *splines* di lisciamento, utilizzando come variabile esplicativa lo *score*. La scelta è ricaduta su questi modelli perchè, nella classe dei modelli non parametrici relativi a contesti di regressione con l'utilizzo di un'unica variabile esplicativa, forniscono le migliori prestazioni sulla frontiera, relativamente al compromesso varianza-distorsione (Hastie & Tibshirani, 1987). In questa sezione, a differenza di quanto fatto in precedenza, si modella una variabile risposta quantitativa e non più dicotomica. L'utilizzo dei residui come variabile risposta nel nuovo modello di regressione comporta non pochi vantaggi. In primo luogo, utilizzare questi come variabile dipendente nella seconda regressione può aiutare a controllare per effetti di confondimento. Ciò significa che è possibile isolare l'effetto di una variabile indipendente specifica sulla variabile dipendente, controllando per altri fattori che potrebbero influenzare la relazione tra le due variabili. Ciò può aiutare a identificare relazioni non lineari tra le variabili, poiché tali relazioni potrebbero non essere state catturate dalla prima regressione. Un'ulteriore motivazione che giustifica la strada percorsa è che l'utilizzo dei residui come variabile risposta nel nuovo modello può migliorare la precisione del modello predittivo. I residui contengono informazioni sulle fluttuazioni casuali della risposta che non sono spiegate dalle variabili esplicative precedenti e l'utilizzo di questi dati può fornire un'indicazione più accurata della variazione di Y dovuta alla nuova variabile esplicativa, che in questo caso è lo *score*. L'obiettivo è quindi quello di depurare gli effetti delle due covariate categoriche da quello della variabile di punteggio e andare a valutare se l'effetto della campagna sia presente o meno. Per mera formalità si forniscono i passi di modellazione:

- Passo 1: Si modello di regressione che prevede y (successo della campagna) come variabile risposta e $x1$ (*cod canale delivery*) e $x2$ (*contatto mod maggio*) come esplicative: $\hat{y} = f(x_1, x_2)$;
- Passo 2: Si estraggono i residui $z = y - \hat{y}$;
- Passo 3: Si utilizzano i residui z come nuova variabile risposta per l'adattamento di modelli non parametrici: $\hat{z} = g(score)$.

5.6.1 *Loess* su residui di albero di regressione

In prima istanza viene adattato un albero di regressione che prevede come variabile risposta il successo della campagna. Una volta adattato, vengono estratti i residui come differenza tra valore osservato e valore previsto dall'albero di regressione. Il primo modello non parametrico adattato che prevede i residui come variabile esplicativa è il *loess*. Il metodo è una variante della regressione locale, che esprime il parametro di lisciamento attraverso la frazione di osservazioni rilevanti per la stima ad una determinata ascissa che viene tenuta costante. Nel *loess*, il parametro di lisciamento è quindi regolato dalla frazione di punti utilizzati, essendo l'ampiezza di banda determinata da tale frazione (Azzalini & Scarpa, 2009). In questo modo si vuole valutare se, isolando l'effetto della campagna dal resto delle variabili esplicative, i risultati cambiano. L'obiettivo è quindi quello di verificare se, al netto delle altre covariate, la campagna abbia un effetto significativo o meno. Si mostrano in Figura 5.23 i risultati ottenuti tramite l'adattamento del modello precedentemente menzionato.

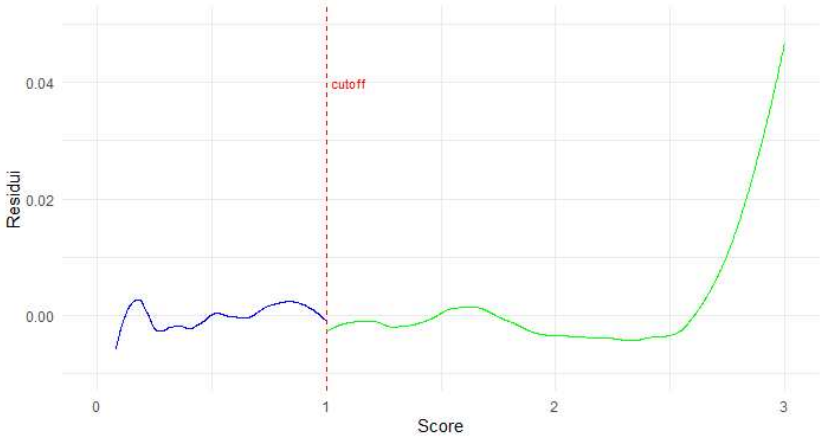


FIGURA 5.23: Valutazione dell'efficacia: *loess* su residui albero

Dal grafico si nota che il salto in corrispondenza della soglia sia presente, ma più piccolo rispetto a quanto ottenuto con i modelli utilizzati nei precedenti paragrafi. È bene osservare che il grafico si riferisce all'effetto parziale della covariata *score*. Il parametro di regolazione (*span*), ottenuto tramite convalida incrociata, ha un valore pari a 0.42. Si può infatti notare la differenza tra questo grafico e quello riportato in Figura 5.15. Le curve di regressione stimate utilizzando *loess* logistico, che prevedevano un valore maggiore di *span* (0.56) rispetto al modello *loess* sui residui, sono infatti più lisce. Stimare quidi un nuovo modello di regressione *loess* su una risposta quantitativa con variabile esplicativa lo *score*, porta a stime più piccole dell'effetto della campagna.

5.6.2 *Splines* di lisciamento su residui di alberi di regressione

Dopo aver adattato un lisciatore *loess*, si passa all'adattamento di *splines* di lisciamento. Anche in questo caso la variabile risposta sono i residui ottenuti tramite albero di regressione. Si sta quindi ancora una volta valutando se vi è un effetto parziale della campagna, al netto di *cod canale delivery* e *contatto mod maggio*, le due variabili esplicative.

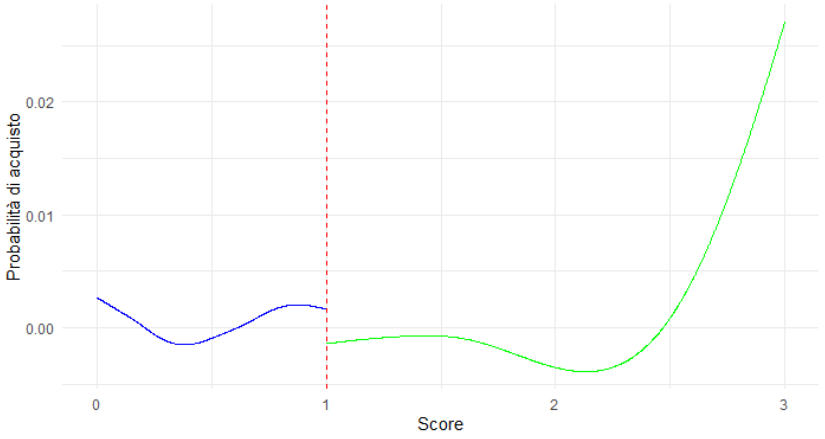


FIGURA 5.24: Valutazione dell'efficacia: *splines* di lisciamento su residui albero

Dalla Figura 5.24 è possibile notare come anche in questa circostanza sia presente un salto verso il basso. Possiamo concludere allora che, nonostante si sia isolato l'effetto delle due covariate categoriali, la campagna sembra non avere effetto. In particolare questo sembra essere minore di quello previsto dal *loess*.

5.6.3 *Loess* su residui di *splines* di lisciamento

Per un confronto con i modelli *loess* e *splines* di lisciamento adattati sui residui dell'albero di regressione, si adattano gli stessi modelli su altri residui, questa volta ottenuti successivamente all'aver adattato *splines* di lisciamento che considerano il successo della campagna come variabile risposta e *cod canale delivery* e *contatto mod maggio* come variabili esplicative. Si vuole ancora una volta isolare l'effetto della campagna dalle precedenti variabili menzionate e valutare se questa ha un effetto. Per fare ciò si adatta in primo luogo un modello *loess*, che come detto precedentemente allarga o contrae l'ampiezza di banda in base alla concentrazione locale dei punti.

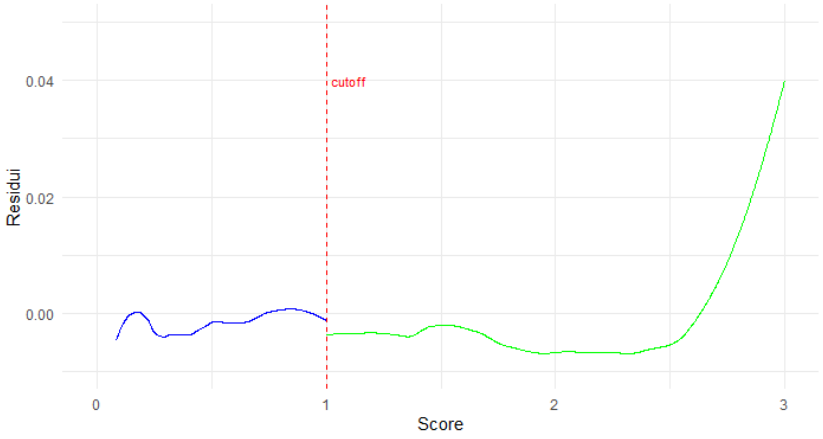


FIGURA 5.25: Valutazione dell'efficacia: *loess* su residui *splines*

In Figura 5.25 viene mostrato quanto ottenuto tramite *loess*. I risultati sembrano essere molto simili a quelli ottenuti in Figura 5.23, in cui si nota un piccolissimo salto direzionato verso il basso. Anche in questa situazione, isolare l'effetto della campagna non ha portato a conclusioni differenti rispetto a quelle ottenute in precedenza. Sembra dunque che la campagna non abbia avuto nessun effetto. Si ricorda ancora una volta che in Figura 5.25, come nei precedenti due grafici, viene mostrato l'effetto parziale della variabile punteggio.

5.6.4 *Splines* di lisciamento su residui dello stesso modello

Infine si adattano *splines* di lisciamento che prevedono come risposta i residui ottenuti dal medesimo lisciatore. Vengono spiegati brevemente i passi eseguiti. Si adattano *splines* di lisciamento con risposta il successo della campagna e come esplicative *cod canale delivery* e *contatto mod maggio*. Si estraggono i residui da questo e si riadattano nuovamente delle *splines* di lisciamento, questa volta però con residui come variabile risposta e *score* come covariata. Si giunge ai seguenti risultati:

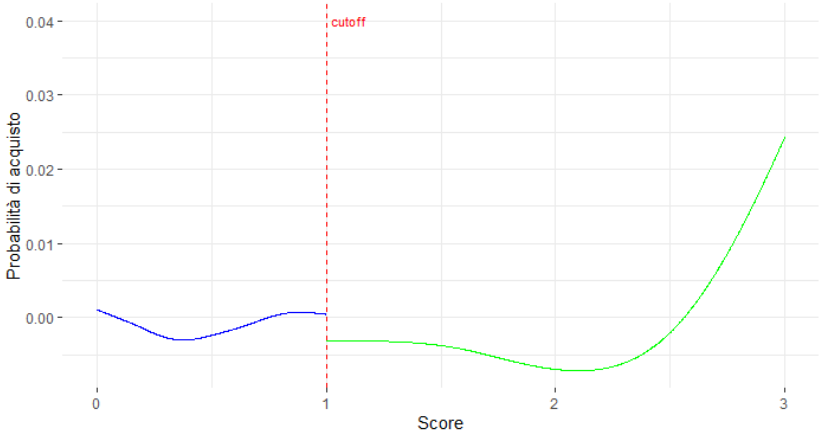


FIGURA 5.26: Valutazione dell'efficacia: *splines* di lisciamento su residui *splines*

In Figura 5.26 il salto che si osserva in corrispondenza della soglia sembra essere più piccolo di quello che si osserva nel grafico corrispondente alle *splines* di lisciamento che utilizzano i residui dell'albero di regressiome come variabile dipendente. In generale, non sembra essere presente effetto parziale. Il fatto che non si verifichi mai un effetto della campagna potrebbe essere dovuto al fatto che si sia sbagliata la scelta della tipologia di disegno da adottare. In alternativa si possono sottoselezionare solo le unità alla destra della soglia che hanno ricevuto la campagna (seguire quindi una tipologia di disegno *Sharp*), e valutare se i risultati ottenuti cambiano o se bisogna rivedere proprio la struttura della campagna.

5.7 Valutazione effetto della campagna tramite disegno *Sharp*

Dopo aver osservato che, tramite disegno *Fuzzy RD* non si è raggiunto un effetto significativo della campagna, si tenta di valutare l'effetto della campagna tramite la seconda tipologia di *Regression Discontinuity*, il disegno *Sharp*. Si ricorda che questa tipologia di disegno, a differenza del modello *Fuzzy*, assegna la campagna di marketing a tutte le unità che hanno almeno raggiunto la soglia. Si ha quindi che l'assegnazione del trattamento è una funzione deterministica e discontinua al *cutoff*. Ciò viene spiegato in modo conciso in Figura 1.3. Per applicare la metodologia in questione è necessario rimuovere precedente dal *dataset* tutte le osservazioni con punteggi maggiori o uguali alla soglia alle quali non era stata assegnata la campagna, ossia bisogna rimuovere le unità *no-shows*. Eliminate queste unità si ripetono le analisi effettate in precedenza, fissando sempre la soglia pari ad 1. Così facendo si avrà che la numerosità del gruppo di controllo sarà molto minore rispetto a quella prevista dal disegno *Fuzzy*. Si passa infatti dal disporre di 26.151 unità all'averne solamente 3.298. In particolare le analisi concernono i seguenti modelli: regressione logistica locale e *loess* logistico.

Regressione logistica locale

Si adatta in prima istanza il modello di regressione locale logistica, il quale fornisce i seguenti risultati mostrati in Figura 5.27.

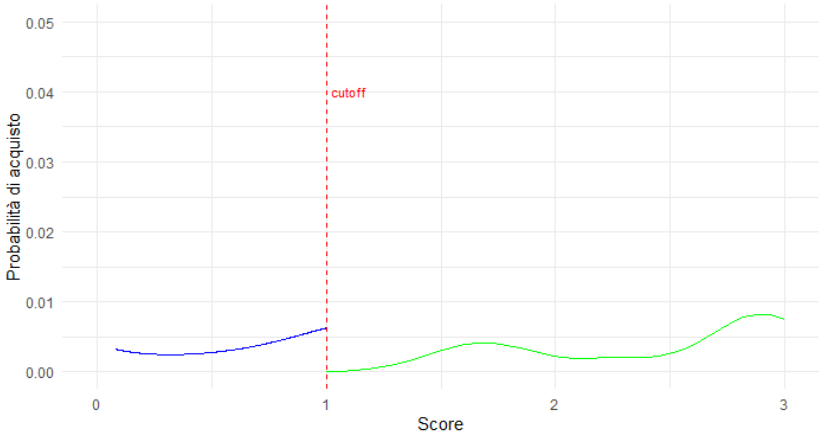


FIGURA 5.27: Valutazione dell'efficacia: Regressione locale logistica (*Sharp RD*)

Nonostante si sia utilizzata una differente tipologia di disegno persiste il salto verso il basso che si era osservato nei precedenti modelli adattati seguendo la metodologia *Fuzzy*.

Loess logistico

Si adatta adesso un *loess* logistico allo scopo di valutare se le conclusioni concernenti la valutazione dell'efficacia della campagna tramite *Sharp RD* sono le stesse ottenute col modello di regressione logistica locale. In Figura 5.28 vengono mostrati i risultati ottenuti.

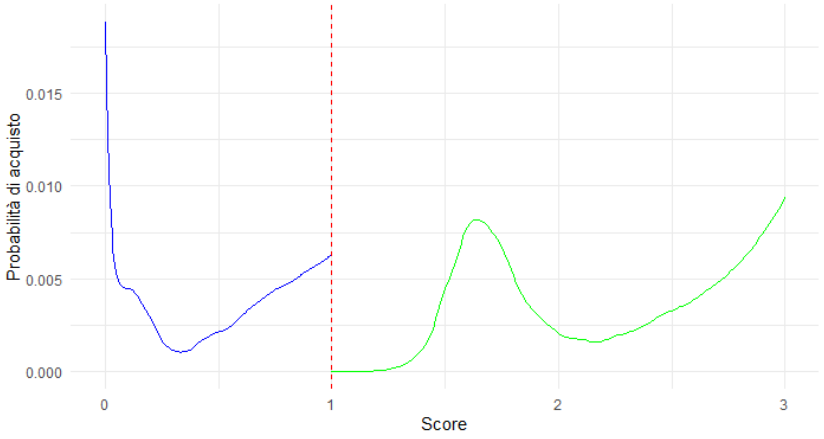


FIGURA 5.28: Valutazione dell'efficacia: *Loess* logistico (*Sharp RD*)

Anche tramite modello *loess* logistico si giunge alle stesse conclusioni tratte durante tutte le analisi. Sembra esserci un salto in corrispondenza del *cutoff* verso il basso. Ciò lascia presagire che, anche se si fosse adotatta una tipologia di disegno diversa dal *Fuzzy*, non si sarebbe arrivati al risultato sperato.

5.8 Test di validità

Come visto nel Capitolo 4, si applicano i test relativi a soglie artificiali e sensibilità alle osservazioni vicine alla soglia, al fine di verificare se il disegno *Fuzzy* sia valido o meno. Vengono forniti solamente grafici, relativi alla modellazione per un'unica covariata. In particolare il test viene effettuato solo per due tipologie di modelli: il modello di regressione logistica locale ed il *loess* logistico, avendo constatato nella parte di modellazione della variabile risposta successo della campagna che i risultati a cui si giunge sono molto simili.

Soglie artificiali

In prima istanza si applica il test che concerne le soglie artificiali. Si decide di utilizzare due differenti soglia, $c = 0.8$ e $c = 1.2$, scelte in modo arbitrario. I risultati ottenuti vengono mostrati successivamente in Figura 5.29 e Figura 5.30.

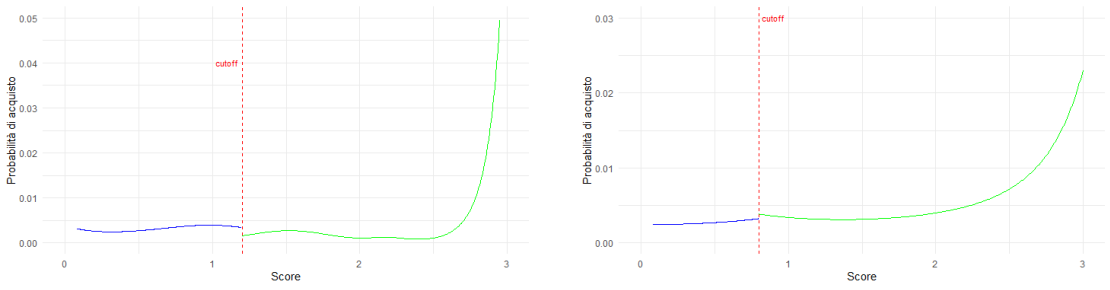


FIGURA 5.29: Logistica locale per soglia = 1.2 (sinistra) e soglia = 0.8 (destra)

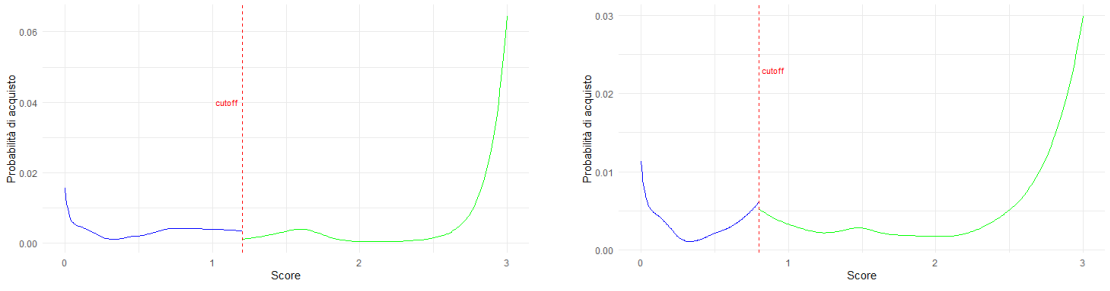


FIGURA 5.30: *Loess* logistico per soglia = 1.2 (sinistra) e soglia = 0.8 (destra)

In tutti i grafici è possibile osservare un piccolo salto verso il basso, eccezion fatta per il modello di regressione logistica locale con soglia fissata a 0.8, nel quale si nota un piccolissimo salto verso l'alto. Non essendoci differenze nette rispetto a quanto visto nelle analisi con soglia pari a 1, tramite test con soglie artificiali, si può supporre il disegno sia valido.

Sensibilità alle osservazioni vicine alla soglia

In secondo luogo si utilizza il test che valuta la sensibilità alle osservazioni vicine alla soglia. Vengono eliminate, rispettivamente a sinistra e destra della soglia, le osservazioni con $0.95 < score < 1$ e $\leq score < 1.05$. Si riadattano quindi i modelli e si verifica la robustezza dei risultati ottenuti inizialmente. I risultati ottenuti vengono mostrati in Figura 5.31 e in Figura 5.32.

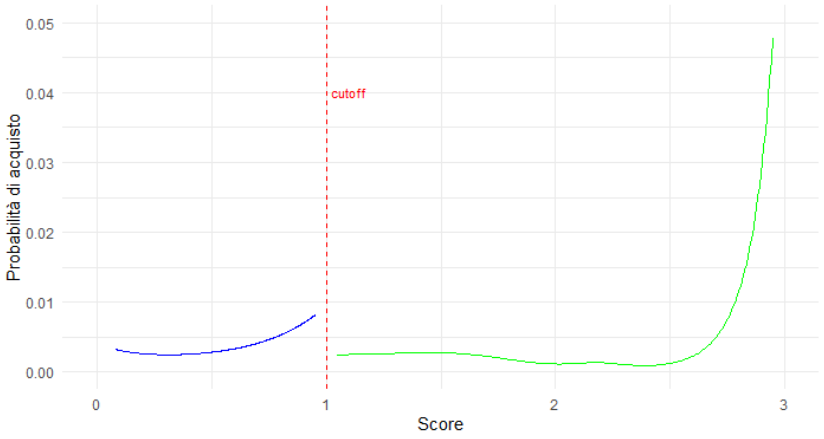


FIGURA 5.31: Regressione logistica locale

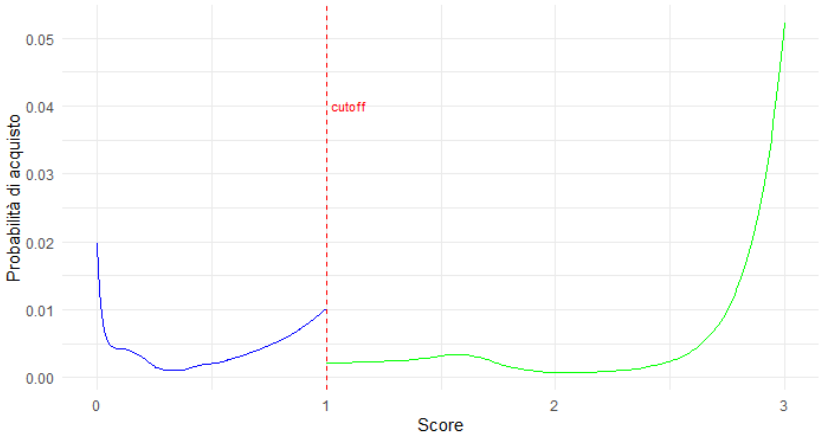


FIGURA 5.32: *Loess* logistico

In tutti i grafici soprastanti si osserva un salto direzionato verso il basso, coerentemente con quanto osservato nell'analisi principale. Anche in questo caso ci si aspetta dunque che il test suggerisca che il disegno è valido.

Capitolo 6

Conclusioni

6.1 Stime dell'effetto della campagna e relativi intervalli di confidenza

Modelli con risposta successo della campagna ed esplicativa lo *score*

In questa sezione si provvede a mostrare le tabelle relative ai valori stimati dell'effetto della campagna per i vari modelli adattati che utilizzano come variabile risposta il successo della campagna e come unica variabile esplicativa lo *score*. In particolare si mostrano stime e relativi intervalli di confidenza ottenuti con modello di regressione logistico locale e modello *loess* logistico. Non risulta utile presentare le stime e gli intervcalli di confidenza per ogni modello adattato poiché i risultati ottenuti, come visto nel Capitolo 5, sono molto simili tra loro. Come visto nelle precedenti sezioni teoriche, le stime dell'effetto della campagna vengono ottenute stimando la funzione di regressione per entrambi i gruppi (trattamento e controllo), e calcolando la differenza tra le stime al punto di discontinuità, ossia in corrispondenza della soglia. Come illustrato in Figura 3.1, la stima dell'effetto, in questo caso della campagna di marketing, corrisponde alla differenza dei valori delle funzioni di regressione nel punto di soglia. In primo luogo si forniscono in Tabella 6.1 le stime dell'effetto della campagna per quanto concerne i modelli che prevedono come unica variabile esplicativa lo *score*.
Tra i due modelli, il modello di regressione *loess* logistico è quello che stima il maggiore effetto negativo del trattamento. Ciò significa che, tramite questo modello, assegnare la campagna comporta una diminuzione della probabilità di acquisto, ossia una diminuzione della probabilità di successo al *cutoff*, pari a 0.47%. Dopo aver osservato i

Modello adattato	Stima effetto	Limite inferiore	Limite superiore
Logistico locale	-0.45%	-1.19%	+0.28%
Loess logistico	-0.47%	-1.69%	+0.75%

TABELLA 6.1: Stime e IC modelli per successo della campagna con *score*

risultati descritti in Tabella 6.1, si può affermare che tutti i modelli stimano un piccolo effetto negativo della campagna. Ciò lascia presagire che questa debba essere in qualche modo rivista ed in particolare sarebbe necessario cambiare il valore della soglia, è dire, assegnare la campagna solo ad unità con valore dello *score* pari ad un altro valore, presumibilmente vicino a 3. Inoltre, in tutti gli intervalli di confidenza considerati è compreso lo 0, quindi l'effetto della campagna è da considerarsi nullo. Si può quindi concludere affermando che la campagna non ha comportato un effetto, negativo o positivo che sia, nella probabilità di acquisto da parte del cliente. È quindi necessario rivedere il meccanismo di assegnazione della campagna o essa stessa, magari attuando una strategia differente.

Modelli con risposta i residui dell'albero ed esplicativa lo *score*

La Tabella 6.2 mostra le stime dell'effetto della campagna ottenute tramite i modelli che utilizzano i residui dell'albero di regressione come esplicativa e lo *score* come unica covariata. Si ricorda che, dopo aver adattato un albero di regressione con successo della campagna come variabile risposta e *cod canale delivery* e *contatto mod maggio* come esplicative, si sono estratti i residui di questa regressione e si sono utilizzati come nuova variabile risposta, nei modelli *loess* e *splines* di lisciamento, usando come unica covariata lo *score*. Ciò è stato fatto al fine di isolare l'effetto dello *score* dalle variabili concomitanti (*cod canale delivery e contatto mod maggio*).

Modello adattato	Stima effetto	Limite inferiore	Limite superiore
Loess	-0.17%	-0.54%	+0.20%
Splines di lisciamento	-0.31%	-0.94%	+0.19%

TABELLA 6.2: Stime effetti della campagna e IC per modelli su residui dell'albero

In questo caso, l'effetto parziale negativo maggiore viene stimato dal modello che utilizza *splines* di lisciamento. Anche in questo caso gli intervalli di confidenza includono lo zero e si può quindi concludere che anche per quanto concerne la modellazione dei residui di una precedente regressione, non si nota, in questo caso, un effetto parziale della campagna.

Modelli con risposta i residui delle *splines* ed esplicativa lo *score*

Successivamente si forniscono in Tabella 6.3 le stime ed i relativi intervalli di confidenza ottenuti per i modelli che prevedono come variabile risposta i residui ottenuti successivamente all'aver adattato *splines* di lisciamento e come variabile risposta lo *score*. I modelli adattati sono ancora una volta il *loess* e le *splines* di lisciamento.

Modello adattato	Stima effetto	Limite inferiore	Limite superiore
Loess	-0.24%	-0.66%	+0.19%
Splines di lisciamento	-0.35%	-0.82%	+0.21%

TABELLA 6.3: Stime effetti della campagna e IC per modelli su residui delle *splines*

Come possibile osservare, le *splines* di lisciamento stimano un effetto parziale negativo maggiore rispetto a quanto ottenuto con i *loess*. Nonostante ciò, in entrambi gli intervalli di confidenza è contenuto lo zero, per cui, ancora una volta, l'effetto della campagna è da considerarsi nullo. Si è visto quindi che adattando:

- modelli univariati che considerano il successo della campagna come risposta e lo *score* come esplicativa;
- modelli univariati che considerano come esplicativa i residui ottenuti da un albero di regressione che prevedeva due differenti covariate categoriali e come esplicativa lo *score* (isolando dunque l'effetto di quest'ultimo da quello delle due precedenti covariate);
- modelli univariati che considerano come esplicativa i residui ottenuti da *splines* di lisciamento che prevedeva due differenti covariate categoriali e come esplicativa lo *score*

in nessun caso sia presente un effetto della campagna. Si può quindi affermare che, probabilmente, la campagna andrebbe rivista e magari andrebbe cambiata l'azione di marketing effettuata.

6.2 Stime e intervalli di confidenza per disegno *Sharp RD*

Come analizzato nel Capitolo 5, si è anche provato a valutare se, assegnando la campagna a tutti i clienti che avevano almeno raggiunto la soglia, applicando quindi la metodologia *Sharp RD*, i risultati cambiassero e se fosse quindi presente un effetto. Dai grafici in Figura 5.27 e Figura 5.28 si era visto come in realtà non sembrerebbero esserci variazioni sostanziali dei risultati. Questi ultimi vengono forniti in Tabella 6.4.

Modello adattato	Stima effetto	Limite inferiore	Limite superiore
Loess	-0.68%	-1.24%	+0.29%
Splines di lisciamento	-0.63%	-1.72%	+0.46%

Tabella 6.4: Stime effetti della campagna e IC *Sharp RD*

Le stime ottenute sono piccole ed inoltre nessun intervallo di confidenza contiene lo zero. Nonostante quindi si sia applicata una metodologia diversa da quella precedente, non si notano effetti significativi della campagna.

6.3 Stime e intervalli di confidenza per test di validità del disegno

In questa sezione vengono fornite le stime ed i relativi intervalli di confidenza ottenuti per i due test di validità del disegno.

Soglie artificiali

Le stime dell'effetto della campagna ottenute sono quelle riportate in Tabella 6.5.

Modello adattato	Stima effetto ($c = 1.2$)	Stima effetto ($c = 0.8$)
Logistico locale	-0.22%	+0.06%
Loess logistico	-0.24%	-0.09%

Tabella 6.5: Stime effetto della campagna con soglie artificiali

Da essa si nota come le stime ottenute siano molto basse. Si osserva, in particolare, come la stima ottenute da logistico locale sia positiva, come valutato precedentemente dai grafici.

Infine si mostrano in Tabella 6.6 i relativi intervalli di confidenza.

Modello adattato	IC ($c = 1.2$)	IC ($c = 0.8$)
Logistico locale	[-0.81%, 0.44%]	[-0.09%, 0.36%]
Loess logistico	[-1.19%, 0.71%]	[-1.15%, 0.45%]

Tabella 6.6: Intervalli di confidenza con soglie artificiali

È possibile osservare come lo zero sia compreso all'interno di ogni intervallo, ragion per cui si considera nullo l'effetto del trattamento. I risultati ottenuti non si discostano in modo evidente da quelle ottenute con la soglia originale. Si può quindi concludere che il test relativo alle soglie artificiali conferma la validità del disegno.

Sensibilità alle osservazioni vicine alla soglia

Si forniscono adesso in Tabella 6.7 le stime dell'effetto della campagna ottenute.

Modello adattato	Stima effetto
Logistico locale	-0.56%
Loess logistico	-0.81%

Tabella 6.7: Stime effetto della campagna per secondo test di validità

Anche in questo caso, le stime ottenute sono tutte negative, è dire, l'effetto della campagna, se significativamente diverso da 0, sembra essere negativo. Infine si mostrano in Tabella 6.8 i relativi intervalli di confidenza.

Modello adattato	Limite inferiore	Limite superiore
Logistico locale	-1.43%	0.31%
Loess logistico	-2.34%	0.72%

Tabella 6.8: Intervalli di confidenza per secondo test di validità

Da essa si osserva come lo zero sia contenuto all'interno di ogni intervallo, risultato coerente con tutte le analisi precedenti. Anche il test relativo sulla sensibiilità alle osservazioni vicine alla soglia conferma la validità del disegno.

6.4 Problemi e possibili miglioramenti

In questa trattazione l'analisi di *Regression Discontinuity Design* non è stata applicata tramite i pacchetti e le funzioni già implementate del *software* `R`. Nella maggior parte delle sue applicazioni, la metodologia viene attuata avendo a disposizione una

variabile risposta quantitativa continua. Le maggiori difficoltà riscontrate dunque in quest'analisi sono dovute proprio alla variabile risposta che, nel caso in questione era dicotomica, con una modalità nettamente prevalente, in termini di frequenze assolute, sull'altra. Per valutare l'efficacia della campagna tramite *Regression Disconitnuity Design* è stato necessario quindi utilizzare differenti metodologie di stima di *data mining*. I risultati a cui si è giunti ci portano ad affermare che la campagna effettuata non ha avuto nessun effetto per quanto concerne la probabilità di acquisto del prodotto da parte dei clienti. Un altro problema riscontrato durante le analisi è stato dovuto alla carenza di variabili esplicative nel modello, ragion per la quale non è stato possibile adattare ulteriori modelli al fine di prevedere le probabilità di acquisto del prodotto da parte del cliente e di conseguenza di valutare, tramite altre tecniche, l'effetto della campagna. Le analisi potrebbero essere migliorate dunque ad esempio aggiungendo ulteriori covariate per migliorare la previsione e quindi la valutazione dell'effetto che la campagna ha avuto. Un'ipotesi sollevata in fase di analisi per quanto concerne un possibile miglioramento della campagna si riferisce ad un ipotetico cambiamento della soglia utilizzata per discriminare i migliori clienti dal restante gruppo. Vengono dunque ripetute nel paragrafo successuvo le stesse analisi precedenti, utilizzando come soglia un valore della variabile punteggio pari a 2.5. Per fare ciò è necessario prima eliminare tutte le osservazioni, che avevano ottenuto punteggi della variabile score appartenenti all'intervallo $[1; 2.5)$ a cui era stata assegnata la campagna. In questo modo la campagna di marketing non verrà assegnata alle unità che non hanno avuto un punteggio almeno pari a 2.5. Verrà assegnata invece in modo casuale a quelle unità che hanno almeno raggiunto la soglia. Importante rimembrare che ciò viene effettuato in questa maniera poiché la tipologia di disegno utilizzata è il *Fuzzy RD*. Cambiare la soglia in un disegno di *Regression Discontinuity* è un'azione forte. Ci sono alcune considerazioni da tenere a mente:

- la scelta della soglia dovrebbe essere guidata dalla teoria e dalla ricerca precedente. per essere sicuri della validità della scelta bisognerebbe disporre di molte più informazioni riguardo la campagna, i clienti, il prodotto e la modalità con la quale questa è stata fissata a 1. Non dovrebbe quindi essere scelta arbitrariamente o solo perché i risultati non sono significativi;

- cambiare la soglia potrebbe comportare la perdita di potenza, poiché si stanno escludendo alcune osservazioni che potrebbero essere utili per l'analisi;

- la scelta della soglia può influire sulla validità delle conclusioni. Se la soglia viene cambiata troppo spesso, i risultati potrebbero diventare poco affidabili.

In questo caso quindi si tenta di valutare semplicemente se cambiare la soglia comporta variazioni sostanziali nei risultati ottenuti. Nel caso in cui si verificasse che, spostando il *cutoff*, l'effetto fosse significativo, andrebbe indagato in modo minuzioso e dettagliato il *modus operandi* con la quale è stata effettuata la scelta della soglia con la quale sono state effettuate le analisi.

6.4.1 Modelli univariati con una nuova soglia

In questa sezione si provvede a fornire gli *output* dei modelli di regressione logistica locale e *loess* logistico che vengono presentati successivamente, al fine di valutare se, cambiando il valore della soglia ed in particolando ponendola pari a 2.5, si possa verificare un effetto significativo della campagna. Come detto nel paragrafo precedente, prima di adattare i vari modelli sono state eliminate le osservazioni alle quali era stata assegnata la campagna che avevano ottenuto punteggio compreso tra 1 e 2.5 (escluso). Vengono adattati solamente due modelli. Adattare più modelli risulterebbe solo oneroso dal punto vista computazionale poiché si è visto che comunque, in generale, i risultati a cui si arriva sono più o meno similari per tutti i modelli. In particolare si decide di adattare il modello di regressione logistica locale e il modello di regressione *loess* logistico. Le fondamenta teoriche sono le stesse discusse precedentemente, ragion per cui verranno mostrati nelle Figure 6.1 e 6.2 solamente i grafici prodotti e se ne darà un commento generale.

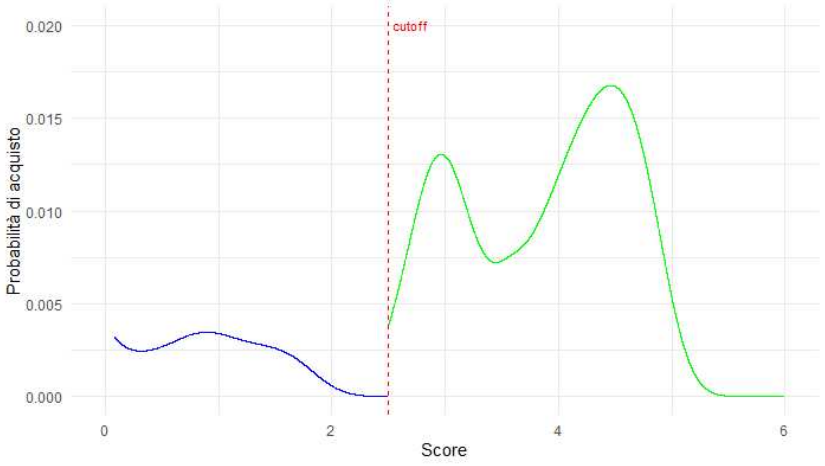


FIGURA 6.1: Valutazione efficacia: regressione logistica locale

Cambiando la soglia da 1 a 2.5 possiamo osservare, al contrario di come si era constatato precedentemente, un deciso salto verso l'alto passando dal lato sinistro al lato destro della soglia. Ciò lascia presagire che se i clienti migliori fossero stati identificati come quelli aventi punteggio pari ad almeno 2.5, la campagna avrebbe potuto avere un effetto,

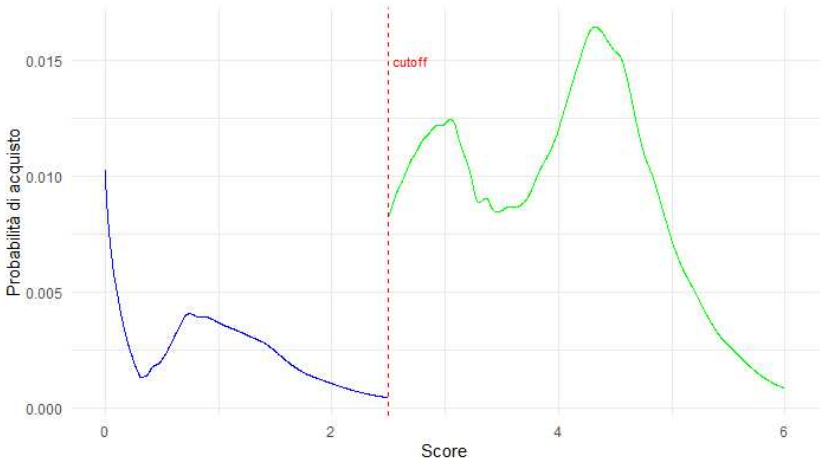


FIGURA 6.2: Valutazione efficacia: regressione *loess* logistico

per di più positivo. Si provvede adesso a fornire la Tabella 6.9 che mette a confronto le stime dell'effetto della campagna ottenute con le due diverse soglie.

Modello adattato	Stima effetto ($c = 1$)	Stima effetto ($c = 2.5$)
Logistico locale	-0.45%	+0.70%
Loess logistico	-0.47%	+0.78%

TABELLA 6.9: Differenza stime effetto della campagna con nuova soglia (unica covariata)

Dalla Tabella 6.9 è possibile notare come i risultati ottenuti con le due soglie siano decisamente differenti. Tutti i modelli adattati con la nuova soglia stimano un effetto positivo della campagna. In secondo luogo si fornisce la Tabella 6.10 relativa agli intervalli di confidenza ottenuti con i nuovi modelli, al fine di valutare se l'effetto della campagna risulta in questo caso significativo.

Modello adattato	Limite inferiore	Limite superiore
Logistico locale	0.52%	0.87%
Loess logistico	0.21%	1.36%

TABELLA 6.10: Intervalli di confidenza con nuova soglia (unica covariata)

Come è possibile notare, in tutti gli intervalli di confidenza ottenuti, non è presente in nessun caso lo zero, motivo per il quale si può affermare che in questo caso l'effetto della campagna è significativo e che quindi, assegnare una campagna a quei clienti che hanno ottenuto un punteggio minimo di 2.5, porta un effettivo aumento della probabilità di acquisto da parte del cliente stesso e conseguentemente una maggiore probabilità di successo per la campagna. Andrebbero dunque riviste molte caratteristiche dell'azione

effettuata allo scopo di valutare se l'analisi svolta possa essere un'indicazione utile per l'azienda o una mera prova analitica.

6.5 Conclusioni

Con questa trattazione si è voluto fornire una panoramica generale sul *Regression Discontinuity Design*, in particolare sulla sua procedura di stima, inferenza e validità. Da un punto di vista statistico le difficoltà principali emersi dall'analisi dei dati effettuati derivano in primo luogo dallo squilibrio delle modalità osservate nella variabile risposta. Prevedere quindi le probabilità di successo non è stato dunque semplice, a maggior ragione non disponendo di un numero alto di variabili esplicative. L'obiettivo dell'analisi, ossia la valutazione dell'efficacia della campagna di marketing, è stato perseguito mediante tecniche di *data mining*, in particolare tramite l'uso di modelli di regressione non parametrica. I parametri di regolazione dei vari modelli sono stati selezionati con procedure che seguono l'ottica del compromesso tra varianza e distorsione, come la convalida-incrociata. Si sono adattati nell'ordine:

1. modelli univariati non parametrici che prevedevano il successo della campagna come variabile risposta e lo *score* come variabile esplicativa;

2. modelli univariati non parametrici che prevedevano come variabile risposta i residui di un precedente modello non parametrico (albero di regressione e *splines* di lisciamento, adattati utilizzando il successo della campagna come risposta e le variabili *cod canale delivery* e *contatto mod maggio* come esplicative) e lo *score* come esplicativa, al fine di isolare gli effetti di quest'ultima variabile da quelli delle altre due;

3. modelli univariati applicando la metodologia *Sharp RD*, al fine di valutare se, usando una differente tipologia di *Regression Discontinuity Design*, i risultati concernenti gli effetti della campagna variassero.

I risultati ottenuti ci portano ad affermare che la campagna di marketing condotta dall'azienda non ha avuto nessun effetto. Ciò porta ad affermare che il criterio con cui è stata scelta la soglia andrebbe rivisto, allo scopo di identificare in modo più opportuno i clienti migliori, in modo tale da assegnare la campagna solo a coloro che si pensa che possano effettivamente acquistare il prodotto. Ciò garantirebbe l'efficacia della campagna e un risparmio notevole in termini di spese effettuate. Un'altra strada percorribile dall'azienda sarebbe quella di effettuare un'azione diversa da quella fatta in precedenza. Ciò potrebbe far sì che la campagna abbia un effetto significativo sulla vendita del

prodotto. Nel nostro caso si avevano a disposizioni poche variabili esplicative, ed in particolar modo nessuna di queste era relativa a caratteristiche del cliente. In conclusione si può quindi affermare che in questo caso, tramite *Regression Discontinuity Design* si è verificato come la campagna in questione non risultasse efficace ma che, con opportuni miglioramenti relativi alla fase preparatoria dell'azione da effettuare, si possano ottenere i risultati sperati.

Bibliografia

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Avery, M. (2013). Literature review for local polynomial regression. *Unpublished manuscript* .

Azzalini, A. & Scarpa, B. (2009). *Analisi dei dati e data mining*. Springer Science & Business Media.

Bowman, A. W. & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, vol. 18. OUP Oxford.

Cattaneo, M. D., Idrobo, N. & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cook, T. D. & Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique* , 127–150.

Fan, J. & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 371–394.

Hahn, J., Todd, P. & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69**, 201–209.

Hardle, W., Hall, P. & Ichimura, H. (1993). Optimal smoothing in single-index models. *The annals of Statistics* **21**, 157–178.

Hastie, T. (2011). *Gam: Generalized Additive Models*. R package version 1.06.2.

Hastie, T. & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association* **82**, 371–386.

Imbens, G. W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics* **142**, 615–635.

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jacob, R., Zhu, P., Somers, M.-A. & Bloom, H. (2012). A practical guide to regression discontinuity. *MDRC* .

Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature* **48**, 281–355.

Maciejewski, M. L. & Basu, A. (2020). Regression discontinuity design. *JAMA* **324**, 381–382.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics* **142**, 698–714.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology* **51**, 309.

Trochim, W. M. (1990). The regression-discontinuity design. *Research methodology: Strengthening causal interpretations of nonexperimental data* , 119–139.

Valentim, V., Núñez, A. R. & Dinas, E. (2021). Regression discontinuity designs: a hands-on guide for practice. *Italian Political Science Review/Rivista Italiana di Scienza Politica* **51**, 250–268.

Van der Klaauw, W. (2008). Regression–discontinuity analysis: a survey of recent developments in economics. *Labour* **22**, 219–245.

van Leeuwen, N., Lingsma, H. F., Mooijaart, S. P., Nieboer, D., Trompet, S. & Steyerberg, E. W. (2018). Regression discontinuity was a valid design for dichotomous outcomes in three randomized trials. *Journal of clinical epidemiology* **98**, 70–79.