Università degli studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Tesi di Laurea

Effetto delle differenze di reddito sulle elezioni presidenziali statunitensi

Relatore: Dott. Pietro Belloni
Dipartimento di Scienze Statistiche

Laureando: Enrico Fraulini
Matricola 2053208

Anno Accademico 2023/2024

A Silvia, Sveva, Andrea e Alberto,
per le chiacchiere, gli spritz e
l'amicizia costruita in questi
(faticosi) anni.

Indice

Abstract

Valutare quali sono i fattori che influenzano le intenzioni di voto dei cittadini è complicato a causa delle tante variabili da tenere in considerazione. Partendo dai dati per contea dell'elezione per il presidente degli Stati Uniti del 2020, l'obiettivo di questa analisi è stato studiare la relazione tra la situazione economica e il voto, correggendo per alcuni confondenti e includendo la variabilità spaziale. È stata analizzata l'associazione tra due possibili trattamenti, reddito mediano o indice di Gini, con la percentuale di voti ottenuta dal Partito Democratico. A questo scopo è stato implementato un algoritmo di matching basato sull'estensione del propensity score per trattamenti continui, stimato partendo dalla stima del modello di Besag-York-Mollié. Si mostra che, una volta avvenuta la correzione per i confondenti, le associazioni dei trattamenti con la risposta cambiano. Se in un primo momento entrambi mostrano un'associazione positiva e significativa con il voto, dopo la correzione il reddito mediano mostra un'associazione negativa e significativa mentre l'indice di Gini non presenta alcuna associazione significativa con la risposta.

Introduzione

Il fattore economico ricopre un aspetto importante della vita di ogni individuo poichè fortemente connesso alla qualità della vita, sia in termini di salute che di soddisfazione personale. Per questa ragione ogni persona è incentivata a tutelare i propri interessi all'interno della società e questo si riflette nella scelta dei rappresentanti politici. Questa scelta risulta influenzata anche da tanti altri fattori, come l'etnia, il titolo di studio e in generale dal proprio vissuto e dal contesto socio-culturale in cui si è immersi, compreso il luogo in cui si vive. Questo comporta grandi difficoltà nel determinare e quantificare quale impatto ha il fattore economico sull'intenzione di voto dei cittadini. Utilizzando metodi sviluppati originariamente in ambito epidemiologico è però possibile arginare il problema valutando se e come il livello di disuguaglianza economica influenza il voto delle elezioni. Sebbene siano valutazioni generiche, esse assumono un particolare significato nel caso statunitense, nel quale gli Stati sono caratterizzati da ampie divisioni socio-culturali: per questa ragione è stato deciso di considerare le elezioni statunitensi del 2020. Le domande a cui si vuole rispondere nel corso dello studio sono focalizzate sull'aspetto economico: il reddito mediano di una contea presenta un'associazione con il voto espresso alle elezioni? Il livello di disuguaglianza economica, descritta dall'indice di Gini, ha un effetto sul voto delle elezioni? Oppure sono associazioni spurie, determinate in realtà da relazioni con altre variabili? Oltre ad essere un tema di grande interesse, la metodologia sviluppata per trovare alcune di queste risposte può essere utilizzata e riproposta anche per altre ricerche, sia legate alle scienze politiche, sia in ambito economico o medico.

Nel Capitolo 1 viene presentato il problema alla luce della letteratura prodotta

fino ad ora, e si descrive l'insieme di dati e variabili utilizzate. Nel Capitolo 2 è stata riportata la metodologia, dando particolare rilevanza agli elementi di novità presenti nello studio, come la stima del propensity score per trattamenti continui utilizzando il modello di Besag-York-Mollié per dati spaziali. Nel Capitolo 3 si descrivono i risultati ottenuti, sia per quanto riguarda l'associazione tra il reddito mediano e il voto che per la relazione tra l'indice di Gini e il voto. Nel Capitolo 4 si conclude fornendo un'interpretazione dei risultati, cercando di dare una risposta alle domande formulate in precedenza.

Capitolo 1

Dati e presentazione dello studio

I fattori che influenzano il voto di ogni cittadino sono molteplici e spesso difficili da quantificare. Sebbene tale considerazione possa essere valutata e riproposta per ogni paese al mondo - ognuno di essi definito da un insieme di molti individui con caratteristiche e vite diverse - assume una particolare rilevanza nel caso statunitense, dove le differenze socio-culturali tra uno stato e l'altro sono nette e, alle volte, conflittuali. L'analisi proposta si pone l'obiettivo di valutare l'effetto delle disuguaglianze economiche sul voto nelle elezioni statunitensi del 2020 (Biden contro Trump), utilizzando metodi sviluppati originariamente in ambito epidemiologico per considerare l'effetto di altre variabili che possono fungere da confondenti nella relazione tra il trattamento (reddito mediano e indice di Gini) e la risposta (voto nelle elezioni statunitensi del 2020).

Prima di descrivere l'insieme dei dati utilizzati per lo studio, si ritiene utile fornire una panoramica del funzionamento del sistema elettorale statunitense e della relazione che potrebbe sussistere tra il reddito ed il voto alle elezioni. Si include inoltre una sezione (Paragrafo 1.3) nella quale si sottolinea l'importanza degli effetti spaziali per l'analisi di dati elettorali.

1.1 Il sistema elettorale statunitense, in breve

Le elezioni statunitensi si tengono ogni quattro anni il martedì successivo al primo lunedì di novembre. La campagna elettorale comincia ben prima: per legge ogni partito è obbligato ad organizzare delle primarie, ovvero una consultazione elettorale interno al partito che decreterà il candidato del partito alla presidenza. Le primarie dei rispettivi partiti iniziano nel febbraio dell'anno delle elezioni.

Il giorno delle elezioni (o nei giorni precedenti, se i cittadini richiedono di votare per posta) ogni persona esprime la preferenza per un presidente ed un vicepresidente. Questi 'voti popolari' determinano l'elezione di un gruppo di persone, chiamati 'grandi elettori': essi formano il collegio elettorale e tramite il proprio voto decretano l'elezione effettiva del presidente e del vicepresidente. Ad ogni stato è assegnato un certo numero di grandi elettori, in base alla propria rappresentazione al Congresso (organo legislativo statunitense, che comprende il Senato e la Camera dei Rappresentanti). Il candidato presidente che ottiene la maggioranza dei voti nello stato otterrà tutti i voti dei grandi elettori di quello stato (solo Maine e Nebraska seguono regole diverse[1]). Ci sono 538 grandi elettori totali: chi supera la maggioranza di essi (270) vince le elezioni e diventa presidente.

Data la natura maggioritaria del sistema elettorale ogni elezione viene contesa soprattutto in alcuni Stati, chiamati *flipped states*, che sono in bilico e determinanti per la vittoria di uno dei due candidati. Gli Stati decisivi per le elezioni del 2020 furono Michigan, Wisconsin, Arizona, Pennsylvania e Georgia in cui il Partito Democratico ottenne la maggioranza dei voti (Stati repubblicani al momento dell'elezione, figura 1.1).

Gli altri Stati rappresentano invece delle roccaforti per i due partiti, e in essi l'esito delle elezioni si ritiene abbastanza scontato. La California ad esempio è sal-

[1] Lo stato del Nebraska è diviso in tre distretti congressuali, e gli sono assegnati cinque grandi elettori totali. Due di questi sono assegnati in blocco al vincitore del voto popolare nello stato, mentre ognuno dei restanti tre è assegnato al candidato che ottiene la maggioranza dei voti nei tre distretti congressuali. Analogamente il Maine è suddiviso in due distretti congressuali e gli sono assegnati quattro grandi elettori: due vanno a chi vince il voto popolare, mentre ognuno dei restanti due va al partito che ha ottenuto la maggioranza in ciascuno dei due distretti congressuali.

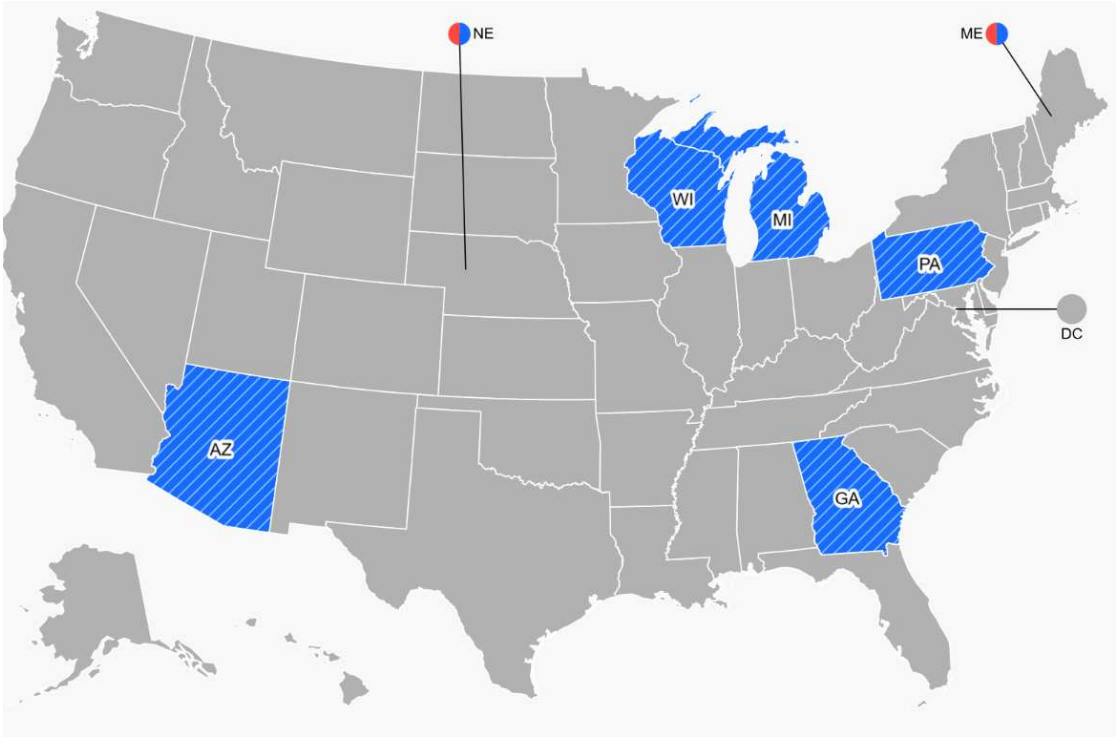


Figura 1.1: Cartina degli Stati Uniti: in blu sono rappresentati i *flipped states* per le elezioni del 2020 (CNN 2020).

damente in mano al Partito Democratico, mentre in Texas è solida la maggioranza repubblicana. Ovviamente con il tempo anche la rilevanza dei partiti negli Stati può cambiare, e California e Texas in questo caso sono un ottimo esempio: per diverse cause (sovraffollamento, altissimo costo della vita, possibilità di lavorare da remoto, solo per citarne alcune) molte persone stanno emigrando dallo stato californiano verso il Texas, modificando la struttura socio-demografica e portando diversi cambiamenti all'interno dello stato (F. Noel Perry, Colleen Kredell, Marcia E. Perry, Stephanie Leonard 2019). Non è escluso quindi che nei prossimi anni si possano osservare variazioni anche in questa direzione e che le maggioranze partitiche nei diversi Stati si modifichino, portando cambiamenti sostanziali all'interno nella struttura sociale attuale degli Stati Uniti.

1.2 L'importanza del reddito sul voto

La politica e le elezioni sono un momento di coesione sociale, che contribuiscono alla condivisione di un sentimento democratico comune. Il modo in cui vengono vissute da ogni individuo è però diverso, e dipende anche dal proprio status all'interno della società. Erikson (2015) ha mostrato ad esempio che le persone con redditi più bassi sono meno incentivate a votare o a possedere le competenze e le informazioni utili ad entrare in politica. Alcune ricerche hanno mostrato inoltre che a causa di questa sproporzione in termini di competenza e del minor tasso di affluenza, gli individui con redditi più bassi hanno un minor peso sugli esiti elettorali (Winters e Page 2009, Hacker e Pierson 2010, Gilens e Page 2014). Valutare come il reddito influenza il voto di ogni cittadino è una questione ancora più complicata e alle volte controintuitiva: si osserva ad esempio che le classi dirigenti tendono a favorire le classe più povere, agevolando una maggior liberalizzazione e politiche redistributive. D'altra parte, gli elettori con redditi alti sono anche quelli con le capacità di rallentare tali cambiamenti. Hersh e Nall (2016) hanno anche riportato come sia molto complicato distinguere il contributo dell'etnia da quello economico:

> "*è stato riscontrato che la relazione tra reddito e appartenenza politica tende a essere più forte del previsto in luoghi con alti livelli di diversità razziale e un passato di sfruttamento razziale locale, mentre risulta più debole altrove. Si conclude che il legame tra reddito e orientamento politico è strettamente connesso alla razza.*"

Nel corso degli anni molta letteratura si è concentrata inoltre sull'aumento delle diseguaglianze, in termini di reddito e ricchezza, e tutto ciò che esse comportano, come una discrepanza sempre più marcata nella qualità della vita (Deininger e Squire 1996, Gelman et al. 2005, Galbraith e Hale 2008, Galbraith e Choi 2020). Da un lato ci si aspetterebbe che una grande disuguaglianza all'interno di uno stato porti una maggiore affluenza alle urne: gli individui più poveri potrebbero avere l'obiettivo di fare valere la propria voce dando fiducia al partito politico che propone politiche maggiormente redistributive, mentre la classe più benestante

avrebbe l'interesse di votare il candidato che propone tagli delle tasse e misure regressive. L'altra ipotesi è invece che un livello di disuguaglianza sempre più ampio deprima l'affluenza alle urne: la disgregazione sociale porta alla sfiducia nella politica e gli individui appartenenti ad ogni classe sociale perdono la speranza che essa possa risolvere i propri problemi. Galbraith e Hale (2008) con un'analisi per le elezioni del 1992 e 2004 hanno mostrato che un grande livello di disuguaglianza è associata negativamente all'affluenza e invece positivamente al voto per il Partito Democratico. Nelle aree più omogenee, suburbane e rurali, a prevalenza bianca ottiene la maggioranza dei voti il Partito Repubblicano, mentre sembra esserci una concordanza tra la classe ricca urbana, le minoranze etniche, e le persone con redditi più bassi a favore del Partito Democratico.

Valutare l'effetto del reddito sul voto dei cittadini è complicato anche per gli effetti contraddittori che sembrano emergere dalle analisi. Gelman et al. (2005) notarono ad esempio come ci fosse una forte tendenza delle persone con un reddito alto a votare repubblicano, mentre i luoghi con redditi mediani alti (la California, per fare un esempio) tendessero ad avere una prevalenza democratica. La risoluzione del paradosso (Galbraith e Hale 2008) mostrò che la pendenza della retta voto/reddito per ogni stato era fortemente associata alla separazione fisica di ricchi e poveri all'interno dello stesso stato. La retta era più piatta (con effetto del reddito sempre più trascurabile) nei luoghi in cui ricchi e poveri tendevano a vivere in contee diverse, mentre aveva un'inclinazione maggiore nelle aree in cui persone con redditi molto diversi vivevano nelle stesse contee. I primi sono caratterizzati da una chiara prevalenza democratica (luoghi con forte disuguaglianza, ma con un reddito mediano alto) mentre i secondi hanno una base solidamente repubblicana (aree con minore disuguaglianza, ma reddito mediano più basso).

Negli ultimi anni il problema di valutare l'effetto dei singoli fattori (e quindi del reddito e della disuguaglianza economica) sul voto è stato anche accentuato da una schizofrenica campagna elettorale. Sia nel 2016 che nel 2020 la campagna di Trump proponeva esplicitamente una politica economica indirizzata alla base elettorale maggiormente benestante, proponendo tagli delle tasse per le persone più ricche. D'altra parte, la retorica utilizzata puntava invece all'odio e alla frustra-

zione della classe sociale industriale, proponendo una resurrezione - immaginaria, nei fatti - dell'industria manifatturiera. Il Partito Democratico invece, nonostante proponga politiche maggiormente redistributive a favore delle classi meno abbienti, viene percepito dall'opinione pubblica maggiormente vicino alle élite e alle classi dirigenti. Soprattutto nelle elezioni del 2016, probabilmente anche a causa dei frequenti interventi per banche d'investimento e a al suo passato da senatrice dello stato di New York, Hillary Clinton venne accostata maggiormente a Wall Street, rispetto al proprio sfidante (Galbraith e Choi 2020).

Alla luce dei tanti aspetti che coinvolgono e si intrecciano con la situazione socio-economica degli individui, è interessante valutare se ci sia un'effettiva relazione tra il reddito e gli esiti elettorali, o se essa risulti emergere solo a causa di altri fattori, come l'etnia o il livello di istruzione.

1.3 Effetto della spazialità

Diverse ricerche hanno mostrato nel corso degli anni il ruolo fondamentale della geografia nel contesto elettorale. In ambito europeo, ad esempio, O'Loughlin, Flint e Anselin (1994) descrissero come il voto nel 1930 della Germania nazista fu fortemente influenzato dalle differenze regionali: a livello locale ogni regione ebbe esperienza della politica nazista in modo diverso. Altre ricerche si sono concentrate invece sulle elezioni britanniche (Cutts et al. 2014), sull'aumento del voto populista in Austria (Essletzbichler et al. 2021) o sulle elezioni in Repubblica Ceca tra il 2006 e il 2017 (Lysek, Pánek e Lebeda 2021).

Anche nel caso statunitense si è mostrato che i pattern spaziali sono fondamentali per la comprensione degli esiti elettorali e che stanno acquisendo anche più importanza nel tempo. Kim, Elliott e Wang (2003), soffermandosi sulle elezioni tra il 1988 ed il 2000, osservarono che globalmente le contee tendevano sempre di più a raggrupparsi secondo il supporto allo stesso partito. Scrivono infatti che

> "*la formazione di cluster geografici caratterizzati dal supporto politico e il conseguente rafforzamento delle rispettive basi elettorali suggeriscono*

> *che essi diventeranno sempre più polarizzati in termini di partito ed ideologia.*"

Questa tendenza è supportata anche da ricerche successive, dove si conferma che sia le elezioni presidenziali che quelle per il Senato sono diventate maggiormente polarizzate, se confrontate a quelle tenute nel secondo dopoguerra. Si riporta inoltre quanto sia determinante utilizzare dati per contea nella valutazione dei cambiamenti elettorali, rispetto ad un'analisi condotta solo sugli Stati (Amlani e Algara 2021). Siccome alcuni dei fattori che influenzano il voto dei cittadini agiscono globalmente, mentre altri solo localmente, è necessario tenere conto nelle analisi della struttura spaziale. Stewart Fotheringham, Li e Wolf (2021) sottolineano ad esempio che

> "*sebbene la geografia a livello statale sia utile per comprendere gli esiti elettorali negli Stati Uniti (cioè, chi vince alla fine), essa oscura nei fatti la diversità dei fattori socioeconomici e contestuali che guidano l'esito (perché vincono). La scala e la struttura geografica della variazione possono essere differenti a seconda del fattore socioeconomico in analisi. Inoltre, gli effetti contestuali variano significativamente da uno stato all'altro. Essi sono alla base di molti Stati contesi, avendo spesso un impatto positivo in una parte dello stato ma un impatto negativo altrove*".

Gli autori suggeriscono quindi quanto sia rilevante tenere conto sia della geografia, sia della struttura sociale che caratterizza i diversi luoghi. Nel corso dell'analisi, tramite la costruzione della matrice dei primi vicini e l'introduzione delle altre covariate, si terrà conto di entrambi gli aspetti.

1.4 Origine dei dati e scelta dei confondenti

Il database ideale per l'analisi che si vuole condurre conterrebbe i dati individuali dei singoli cittadini, con le caratteristiche personali ed il voto espresso. Data l'evidente impossibilità di reperire tali informazioni, si è deciso di lavorare

sui dati aggregati per contea: in particolare sono state selezionate tutte le contee di tutti gli Stati americani, ad eccezione di Alaska e Hawaii, che rappresentano casi geograficamente isolati e con caratteristiche sociali peculiari.

In primo luogo, si considerano gli esiti delle elezioni statunitensi del 2020, prendendo come variabile risposta la percentuale di voti ottenuta dal Partito Democratico nelle diverse contee (MIT 2020). Per tutte le altre variabili i dati sono stati reperiti dal Census Bureau Data statunitense (Census Bureau Data 2020). Come possibili misure di disuguaglianza economica sono stati considerati:

- reddito mediano, in particolare quello che ricade sotto la voce di *household income*, definito come '*il reddito lordo di tutte le persone che occupano la stessa casa e che hanno più di 15 anni*' (Census Bureau Data 2020);

- indice di Gini (Gini 1921), definito come '*una misura riassuntiva della disuguaglianza di reddito e che rappresenta la dispersione del reddito lungo la sua intera distribuzione. Il coefficiente di Gini varia da 0, che indica perfetta uguaglianza (dove tutti ricevono una quota uguale), a 1, che rappresenta una perfetta disuguaglianza (dove solo un beneficiario o gruppo di beneficiari riceve tutto il reddito). Il coefficiente di Gini si basa sulla differenza tra la curva di Lorenz (la distribuzione cumulativa osservata del reddito) e il concetto di una distribuzione del reddito perfettamente equa'* (Census Bureau Data 2020).

Il logaritmo del reddito mediano ed il logaritmo dell'indice di Gini rappresentano i due trattamenti continui dei quali si è interessati a vedere l'effetto sul voto. Alla luce delle considerazioni precedenti (Sezione 1.2 e 1.3) e dell'obiettivo dello studio, si è deciso di includere nell'analisi tutte le variabili reperibili che potessero rappresentare dei confondenti per la relazione tra trattamento (reddito o indice di Gini) e voto. Di seguito si introducono brevemente le variabili utilizzate.

Percentuale di persone bianche

Negli Stati Uniti vi è un generale consenso sul ruolo centrale che ricopre l'etnia, sia del cittadino che del candidato, nella politica e nei risultati alle elezioni

(Leighley e Vedlitz 1999, Chong e Rogers 2005, Fraga 2018) Inoltre, come già accennato in precedenza (Sezione 1.2), l'etnia presenta una forte associazione anche con il reddito, o più in generale con il livello di disuguaglianza economica. Studi hanno mostrato ad esempio che le persone bianche o asiatiche sono predominanti nelle fasce di retto più alte; inoltre il reddito della maggior parte degli altri gruppi etnici varia tra il 50 e l'80% del corrispondente livello di reddito dei bianchi, in vari percentili della distribuzione generale del reddito stesso (Akee, Jones e Porter 2019, Chetty et al. 2020, Vo, Schleifer e Hekmatpour 2023). Nella valutazione dell'effetto causale dei due trattamenti sul voto, è quindi opportuno porre a confronto contee che presentano composizioni etniche simili. Possedendo dati aggregati e volendo includere una variabile che tenesse in considerazione la composizione etnica della popolazione, si prende la percentuale di persone bianche (*white alone*, dalla notazione del Census Bureau Data).

Percentuale di persone con assicurazione sanitaria

Il rapporto degli statunitensi con il sistema sanitario è diverso da quello europeo e la sanità è uno tra i temi più ricorrenti e discussi nelle campagne elettorali. Negli Stati Uniti non esiste un vero e proprio sistema sanitario nazionale, ma le prestazioni vengono pagate privatamente dai cittadini nel momento della necessità. Per questa ragione, nella gran parte dei casi, le persone stipulano assicurazioni sanitarie: alcune, come il Medicaid e Medicare, sono assicurazioni pubbliche, non a scopo di lucro e con prezzi molto agevolati; altre vengono invece incluse all'interno del contratto di lavoro, se la persona lavora come dipendente. Rappresentando un tema di divisione (Lake e Nie 2023) e potenzialmente associato al voto e al reddito, si decide di includere nell'analisi la percentuale di persone che, per ogni contea, possiedono un'assicurazione sanitaria (senza specificare di quale polizza assicurativa si tratti).

Percentuale di persone laureate (o con un titolo di studio superiore)

Diverse ricerche (Mayer 2010, Jerrim e Macmillan 2015, Blanden, Doepke e Stuhler 2023) nel corso degli anni hanno mostrato che c'è una forte associazione

tra il titolo di studio e il reddito; si osserva inoltre che uno dei predittori migliori per il reddito individuale è proprio il livello di istruzione (ciò implica che una riduzione della disuguaglianza economica potrebbe portare anche ad una riduzione disuguaglianza scolastica). Il livello di istruzione ha un forte impatto anche sulla partecipazione dei cittadini alla vita politica e alle elezioni (Nie, Junn e Stehlik-Barry 1996). Vi sono ricerche che mostrano come ci sia una relazione causale tra il titolo di studio e l'affluenza alle elezioni. Sondheimer e Green (2010), ad esempio, concludono affermando:

> "*prima di condurre questo studio, gli autori hanno espresso scetticismo sull'ipotesi che un aumento dell'educazione potesse essere connesso all'affluenza. [...] Tuttavia, sebbene ognuno di questi studi sia basato sua una bassa dimensione campionaria, l'evidenza sperimentale indica che il titolo di studio influenza fortemente l'affluenza alle urne.*"

Per queste ragioni si ritiene necessario includere una variabile che controlli anche l'istruzione: a questo scopo si considera, per ogni contea, la percentuale di persone che hanno una laurea o un titolo di studio superiore.

Percentuale di persone disoccupate

Tra i diversi fattori che possono influenzare il voto c'è la disoccupazione, come elemento di disagio e preoccupazione nella popolazione. Si è soliti pensare che il partito in carica sia avvantaggiato, per le elezioni successive, quando la disoccupazione nel paese è bassa, e che sia svantaggiato invece quando è alta. Wright (2012), lavorando su dati per contea, mostra che una maggior disoccupazione favorisce il candidato democratico: l'effetto risulta maggiore quando al governo vi è il Partito Repubblicano, ma vi è un'associazione positiva, anche se inferiore, quando il presidente in carica è democratico. La ricerca suggerisce quindi che la disoccupazione è vista dall'opinione pubblica statunitense come un problema maggiormente risolvibile dal Partito Democratico, indipendentemente che quest'ultimo sia in carica oppure no. Altre ricerche, sempre utilizzando dati per contea, mostrano e confermano che un aumento della disoccupazione sfavorisce il Partito Repubblicano

e stimola una maggiore affluenza alle urne (Burden e Wichowsky 2014, Helgason e Mérola 2017). Dati questi presupposti e alla luce della letteratura presente fino ad ora, si decide di includere nell'analisi la percentuale di persone disoccupate per contea, per controllare anche questo aspetto nella relazione tra i livelli di disuguaglianza economica e il voto.

Densità abitativa

Gli Stati Uniti sono un territorio molto vasto ed eterogeneo sotto l'aspetto della densità abitativa. Vi sono alcune contee in cui non si raggiunge un abitante per chilometro quadrato, mentre altre in cui si superano i settantamila abitanti. Siccome l'obiettivo dello studio è confrontare contee che presentano strutture simili, si ritiene utile includere all'interno dell'analisi una variabile che rappresenti e descriva questa eterogeneità: si prende perciò il numero di abitanti per chilometro quadrato di ogni contea.

Database finale

In conclusione, il database finale contiene quindi 3108 osservazioni (una per ogni contea) e 6 variabili: la percentuale di persone bianche, la percentuale di persone con assicurazione sanitaria, la percentuale di persone laureate o con un titolo di studio superiore, la percentuale di persone disoccupate, la variabile categoriale associata allo Stato e la densità abitativa (numero di abitanti per chilometro quadrato). Si includono inoltre tutte le coordinate geografiche necessarie per la costruzione della matrice dei primi vicini.

Capitolo 2

Propensity score generalizzato con lisciamento spaziale

Il seguente capitolo è stato suddiviso in quattro parti: nella prima sezione si introduce e discute per quali ragioni sia necessario il propensity score generalizzato e quali siano gli svantaggi per condurre lo studio desiderato. Nella sezione successiva si definisce il modello di Besag-York-Mollié, tramite il quale si vuole tenere conto degli effetti spaziali, e che sta alla base dell'algoritmo di matching presentato invece nella parte successiva (Sezione 2.3). Si conclude poi con un riepilogo di tutto il procedimento utilizzato, dalla stima del *GPS* alla costruzione della pseudo-popolazione finale, e con una breve descrizione di quali strade sia possibile intraprendere per la stima dell'exposure response function.

2.1 Propensity score generalizzato

Nella maggior parte degli studi osservazionali può essere complicato analizzare gli effetti causali per la presenza di confondenti e di modificatori d'effetto. Gli usuali metodi di regressione non seguono una procedura caratterizzata da una chiara distinzione tra la fase di progettazione e quella di analisi (Wu et al. 2022). Solo

in quest'ultima fase è consentito introdurre la variabile risposta, mentre la fase di progettazione dovrebbe includere solo il trattamento e le covariate a disposizione.

Per questa ragione i risultati ottenuti tramite i metodi di regressione non possono essere interpretati per valutare nessi causali: si sono sviluppate quindi procedure alternative che permettono di valutare l'effetto causale di un trattamento, controllando rispetto a tutte le variabili che è possibile misurare ed includere all'interno del modello. Uno strumento che ha riscosso molto successo negli anni a questo scopo è il *propensity score* (Rosenbaum e Rubin 1983). Siccome nella maggioranza dei casi è utilizzato per valutare trattamenti dicotomici, prima di introdurre l'estensione a trattamenti continui, si propone una revisione del propensity score usuale.

2.1.1 Propensity score per trattamenti dicotomici

Per valutare l'effetto di un trattamento il problema principale risiede nell'impossibilità di osservare l'evento controfattuale, ovvero di valutare cosa accadrebbe se si decidesse di non assegnare il trattamento considerato. Data n la dimensione del campione, per ogni osservazione $i \in \{1, ..., n\}$ sia Y_i la variabile risposta e T_i la covariata associata al trattamento, l'outcome osservato

$$Y_i^{OBS} = Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$$

dipende dal trattamento assegnato, mentre l'outcome potenziale (Y_{0i} o Y_{1i}) no. Si potrebbe dire che entrambi esistono, ma in due mondi diversi: quello reale e quello controfattuale. Se si decidesse ad esempio di somministrare un medicinale a un paziente, per verificare l'effetto di quest'ultimo sarebbe necessario avere un altro paziente, identico al primo, a cui non somministrare il farmaco. Nell'impossibilità di osservare il controfattuale (di non avere un altro paziente identico al primo), ci si pone l'obiettivo di cercare il sostituto migliore possibile, che condivida il maggior numero di aspetti con l'individuo di interesse. Ciò consente di arginare, per quanto possibile, la *self-selection* che produce distorsioni nei risultati (Rosenbaum e Rubin 1983, Austin 2011).

Nella maggior parte dei casi si è interessati a valutare l'effetto medio del trattamento sull'intera popolazione (ATE - *average treatment effect*)

$$ATE = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

L'effetto del trattamento sulla popolazione dei trattati (ATT - *average treatment effect on the sub-population of treated*) si può invece scrivere come

$$ATT = E(Y_1 - Y_0|T = 1) = E(Y_1|T = 1) - E(Y_0|T = 1).$$

Mentre il primo termine può essere stimato usando i risultati derivanti dal sottogruppo dei trattati, $E(Y_1|T = 1) = E(Y^{OBS}|T = 1)$, il secondo non è direttamente calcolabile perché l'outcome Y_0 non si osserva mai per la sotto-popolazione dei trattati. Da qui deriva anche la fonte di distorsione, infatti

$$E(Y_0|T = 1) \neq E(Y^{OBS}|T = 0) = E(Y_0|T = 0)$$

$$\longrightarrow \text{ BIAS} = E(Y_0|T = 1) - E(Y^{OBS}|T = 0).$$

Si osserva allora che la distorsione è dovuta alla differenza tra il gruppo dei trattati e il gruppo dei controlli, prima che venga somministrato il trattamento. In particolare, si possono individuare tre fonti di distorsione:

- B1, dovuta alla mancata sovrapposizione dei supporti nel gruppo dei trattati e dei controlli (vi sono unità nel gruppo dei trattati che non trovano un'adeguata unità di confronto nel gruppo dei controlli);
- B2, dovuta a distribuzioni differenti dei confondenti *osservati* nel gruppo dei trattati e dei controlli;
- B3, dovuta a distribuzioni differenti dei confondenti *non osservati* nel gruppo dei trattati e dei controlli.

Per ridurre il più possibile le prime due fonti di distorsione è possibile utilizzare il propensity score, definito come la probabilità che un soggetto venga trattato, condizionatamente alle covariate a disposizione

$$e(x) = \mathbb{P}(T = 1|\mathbf{x}).$$

La terza fonte di distorsione (B3) non è eliminabile in nessun caso, se non attraverso un esperimento randomizzato controllato, gold standard in tutti gli studi epidemiologici, ma anche raramente realizzabile.

2.1.2 Propensity score per trattamenti continui

Per esaminare il presente argomento, non è possibile adottare l'approccio precedente a causa della natura continua dei trattamenti in esame (reddito ed indice di Gini). Pertanto, è indispensabile considerare un'espansione del concetto di propensity score, denominata *propensity score generalizzato (GPS)*. Si definisca n la dimensione del campione e per ogni unità $j \in \{1, ..., n\}$ siano $\mathbf{x}_j = (x_{1j}, ..., x_{pj})$ l'insieme delle covariate. Sia $T_j^{oss} \in \mathbb{T} = [t^0, t^1]$ il trattamento per l'osservazione j-esima. Si introducano inoltre le seguenti definizioni (Hirano e Imbens 2004, Kennedy et al. 2017, Wu et al. 2022):

Definizione 1: gli *outcome potenziali* sono un set di variabili casuali, $Y(t)\ \forall t \in [t^0, t^1]$, in cui $Y = Y(T^{oss})\ \forall T^{oss} \in [t^0, t^1]$.

Definizione 2: il *propensity score generalizzato* è la densità di probabilità condizionata del trattamento, date le altre covariate:

$$\mathbf{e}(\mathbf{x}) = \{f_{T_j^{oss}|\mathbf{x_j}}(t|\mathbf{x}), \forall t \in [t^0, t^1]\}.$$

I singoli valori $e(t, \mathbf{x}) = f_{T_j^{oss}|\mathbf{x}_j}(t|\mathbf{x})$ sono realizzazioni di $\mathbf{e}(\mathbf{x})$ al livello del trattamento $T_j^{oss} = t$.

Definizione 3: la media causale dell'*exposure response function* (ERF) è data da

$$\mu(t) = E[Y(t)] \quad \forall t \in [t^0, t^1].$$

Si osserva che mentre nel propenisty score per trattamenti dicotomici T^{oss} assume solo due possibili valori ($T^{oss} \in \{t^0, t^1\}$), nel caso di trattamenti continui esso

può assumere tutti i valori (reali) compresi tra t^0 e t^1. Il propensity score generalizzato mantiene però la proprietà fondamentale del propensity score usuale: condizionatamente al *GPS*, la probabilità di ricevere un certo livello di trattamento è indipendente dal set di covariate $\mathbf{x}$ (Hirano e Imbens 2004, Khoshnevis, Wu e Braun 2023).

Nel contesto degli outcome potenziali (Rubin 1974) - adattati nel caso di trattamento continuo - è necessario imporre alcune assunzioni per garantire l'identificabilità (Wu et al. 2022):

Assunzione 1 (Consistenza): per ogni unità j, $T_j^{oss} = t$ implica che $Y_j^{oss} = Y_j(t)$.

Assunzione 2 (Sovrapposizione): per ogni valore di $\mathbf{x}$, la densità di probabilità condizionata di ricevere ogni possibile trattamento $t \in \mathbb{T}$ è positiva: $f_{T_j^{oss}|\mathbf{X_j}}(t|\mathbf{x}) \geq p$ per ogni t, $\mathbf{x}$, e per qualche costante $p > 0$.

Assunzione 3 (Identificabilità debole): per ogni possibile livello di trattamento t, in cui t è continuo nell'intervallo $[t^0, t^1]$, si ha $T^{oss} \perp Y(t)|X$.

L'assunzione 2 evita che il *GPS* non sia mai uguale a zero, ovvero che per ogni possibile combinazione delle covariate $\mathbf{x_j} = \mathbf{x}$ si possa stimare $\mu(t)$ per ogni livello di trattamento t senza ricorrere all'estrapolazione.

L'assunzione numero 3 può essere sostituita con un'altra assunzione più debole (Wu et al. 2022). Si definisce in primo luogo un raggio Δ_n che determina un intervallo nel dominio del trattamento ($[t - \Delta_n, t + \Delta_n]$) e che soddisfi la condizione per cui quando $n \to \infty$, allora $\Delta_n \to 0$. In tal caso:

Assunzione 3bis (Identificabilità debole locale): il meccanismo di assegnazione è localmente e debolmente identificabile se per ogni unità j e per tutti i valori $t \in \mathbb{T}$, t è distribuita in modo continuo rispetto alla misura di Lebesgue in

$\mathbb{T}$, ovvero se

$$f(Y_j(t) \mid \mathbf{x}_j, T_j^{oss} = \tilde{t}) = f(Y_j(t) \mid \mathbf{x}_j) \quad \text{per} \quad \tilde{t} \in [t - \Delta_n, t + \Delta_n]$$

dove f è una generica densità di probabilità.

Assunzione 4 (Lisciamento): per ogni unità j e per ogni $t \in \mathbb{T}$,

$$\mu_{GPS}(t, e) \equiv E[Y_j(t) \mid e(T_j^{oss}, \mathbf{X_j}) = e, T_j^{oss} = t].$$

è continua, secondo la definizione lipschitziana, rispetto a t per ogni valore di e. Ciò implica che

$$|\mu_{GPS}(t, e) - \mu_{GPS}(t', e)| \leq B|t - t'| \quad \forall t, t' \in \mathbb{T}, \forall e, \text{e per qualche costante } B.$$

Sotto le assunzioni *1-4* valgono poi due lemmi (Wu et al. 2022, Kim et al. 2018):

Lemma 1 (Identificabilità debole locale dato il GPS): si supponga che il meccanismo di assegnazione soddisfi la condizione di identificabilità debole locale. Allora per ogni unità j e per ogni $t \in \mathbb{T}$

$$f(Y_j(t) \mid e(\tilde{t}, \mathbf{x}_j), T_j^{oss} = \tilde{t}) = f(Y_j(t) \mid e(\tilde{t}, \mathbf{x}_j)) \quad \text{con} \quad \tilde{t} \in [t - \Delta_n, t + \Delta_n]$$

dove f è una generica densità di probabilità.

Lemma 2 (media causale ERF) : per ogni $t \in \mathbb{T}$

$$\mu(t) = E[Y_j(t)] = \lim_{\Delta_n \to 0} E[E\{Y_j^{obs} \mid e(T_j^{oss}, \mathbf{x}_j), T_j^{oss} \in [t - \Delta_n, t + \Delta_n]\}]$$

Sotto le assunzioni *1-4*, i lemmi *1-2* garantiscono che, per ogni t, sia possibile calcolare il valore atteso di $Y_j(t)$.

Data la potenzialità nell'aggiustare la distorsione prodotta dalla *self-selction*, da quando è stato introdotto, il propensity score generalizzato è stato già utilizzato in molti ambiti, come quello medico (Cecchini e Smith 2018, Lei et al. 2021) ed economico (Doyle 2011, Li et al. 2019).

2.2 Modello di Besag-York-Mollié

Data l'importanza ricoperta dalla spazialità ed in generale dalle aree vicine sul voto, si è deciso di utilizzare il modello di Besag-York-Mollié (Besag, York e Mollié 1991) per tenere in considerazione anche questo aspetto nel corso dell'analisi. Anche se inizialmente pensato e costruito per dati dicotomici o discreti, esso è stato ampiamente utilizzato nel corso degli anni e generalizzato anche per risposte continue (Morales-Otero e Núñez-Antón 2021, Schündeln et al. 2021, Bakka et al. 2018). Data Y la variabile risposta, n la dimensione del campione e $\mathbf{x_j} = (x_{j1}, x_{j2}, ..., x_{jp})^T$ il vettore p-dimensionale che include le covariate, allora si definisce

$$Y_j \sim N(\mu_j, \sigma^2), \quad j = 1, 2, ..., n \tag{2.1}$$

dove

$$\mu_j = \mathbf{x}_j^T \boldsymbol{\beta} + u_j + v_j,$$

in cui $\mathbf{x}_j^T \boldsymbol{\beta}$ è il termine che rappresenta gli effetti fissi, u_i è l'effetto casuale della struttura spaziale e v_j è l'effetto casuale non strutturato. Per definire gli effetti spaziali u_j si determina innanzitutto la matrice di adiacenza Q. Essa è definita come una matrice quadrata binaria in cui colonne e righe sono date dalle aree geografiche. Ogni elemento $Q_{i,j}$ assume il valore 1 quando vi sono uno o più punti confinanti tra l'area i e l'area j, mentre in ogni altro caso $Q_{i,j} = 0$. Ovvero

$$Q_{i,j} = \begin{cases} 1 & \text{se l'area i confina con l'area j} \\ 0 & \text{altrimenti.} \end{cases}$$

In particolare, gli effetti spaziali vengono modellati tramite un CAR (*conditional* auto regressive), che restituisce i dati lisciati in accordo con la struttura di adiacenza data dalla matrice dei vicini Q (Besag 1974):

$$u_i | u_{-i} \sim N(\bar{u}_{\delta_i}, \frac{\sigma_u^2}{n_{\delta_i}})$$

dove u_{-i} è u senza la i-esima aerea, $\bar{u}_{\delta_i} = n_{\delta_i}^{-1} \sum_{j \in \delta_i} u_j$, δ_i è l'insieme dei vicini e n_{δ_i} è il numero di vicini dell'area i-esima. Inoltre si modella l'effetto casuale non strutturato come

$$v_i \sim N(0, \sigma_v^2).$$

2.3 Matching e pseudo-popolazione

Come descritto precedentemente (Sezione 2.1.1) una delle principali fonti di distorsione negli studi osservazionali è lo sbilanciamento delle covariate *prima* del trattamento, ovvero la differente distribuzione delle stesse covariate per diversi valori di T. Negli studi randomizzati controllati il problema è risolto dal fatto che le unità ricevono diversi valori del trattamento in modo casuale: ciò comporta per ogni $T \in [t^0, t^1]$ le covariate avranno distribuzioni simili (ovvero bilanciate). Nel caso degli studi osservazionali è possibile invece arginare il problema tramite la procedura di matching, che ha l'obiettivo di creare un nuovo database in cui la distribuzione delle covariate per ogni livello di trattamento è il più bilanciata possibile (Wu et al. 2022, Khoshnevis, Wu e Braun 2023). Nel caso di un trattamento dicotomico è abbastanza immediato pensare di confrontare due soggetti, uno trattato ($T = 1$) ed uno non trattato ($T = 0$), con simili livelli di propensity score (ovvero con distribuzione delle covariate simili). Invece, nel caso in cui T sia una variabile continua, non è possibile definire unità trattate e non trattate: si rende quindi necessario un nuovo approccio.

2.3.1 Controllo del bilanciamento delle covariate

Dato il database iniziale, l'obiettivo dell'algoritmo di matching è costruire un nuovo database, analogo a quello di partenza, ma con l'aggiunta di una nuova colonna in cui si definiscono i pesi da attribuire ad ogni osservazione (Khoshnevis, Wu e Braun 2023). Questo nuovo database sarà chiamato *pseudo-popolazione*.

Prima di procedere con la descrizione dell'algoritmo, è opportuno introdurre una misura di 'qualità' per la pseudo-popolazione che si otterrà alla fine. Ricordando che l'obiettivo è ottenere un database in cui la distribuzione delle covariate sia il più simile possibile per ogni valore di $T \in [t^0, t^1]$, una possibile misura avviene tramite il calcolo della correlazione assoluta tra le singole covariate e il trattamento. A questo scopo si definisce l'intervallo

$$\{t^{(1)} = \min(t), t^{(2)} = \min(t) + \frac{\max(t) - \min(t)}{K}, ..., t^{(K+1)} = \max(t)\} \in [\min(t), \max(t)]$$

dove K è il numero totale di blocchi e $\mathbb{T}_k = [t_{(k)}, t_{(k+1)})$. Sia con r_k il numero totale di unità all'interno del blocco $\mathbb{T}_k$. Si supponga inoltre che l'unità i-esima nel k-esimo blocco $\mathbb{T}_k$ abbia trattamento t_{ik} e covariate $\mathbf{x}_{ik}$, e che compaia n_{ik} volte nella pseudo-popolazione (n_{ik} rappresenta quindi la frequenza finale, ovvero il numero di volte in cui è avvenuto un matching tra l'osservazione i-esima e quella fittizia). Prima di definire la correlazione si ortogonalizzano e centrano sullo zero le covariate e il trattamento:

$$\mathbf{x}_{ik}^{\dagger} = \mathbf{S_x}^{-1/2}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_{ik}), \quad t_{ik}^{\dagger} = S_t^{-1/2}(t_{ik} - \bar{t}_{ik}),$$

dove

$$\bar{\mathbf{x}}_{ik} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}\mathbf{x}_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}}, \quad \mathbf{S}_x = \frac{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}(\mathbf{x}_{ik} - \bar{\mathbf{x}_{ik}})(\mathbf{x}_{ik} - \bar{\mathbf{x}_{ik}})^T}{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}},$$
$$\bar{t}_{ik} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}t_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}}, \quad S_t = \frac{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}(t_{ik} - \bar{t}_{ik})^2}{\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}}.$$

A tal punto è possibile definire una misura globale ed una locale per la correlazione assoluta a blocchi.

Misura globale: per una qualità ottimale della pseudo-popolazione sarebbe necessario che la correlazione tra le covariate e il trattamento fosse, in media, uguale a zero. Nella pratica si assume che la condizione di bilanciamento delle covariate sia soddisfatta quando

$$\left|\sum_{k=1}^{K}\sum_{i=1}^{r_k} n_{ik}\mathbf{x}_{ik}^{\dagger}t_{ik}^{\dagger}\right| < \boldsymbol{\epsilon}_1 \tag{2.2}$$

dove ogni elemento del vettore p-dimensionale $\boldsymbol{\epsilon}_1$ è uguale ad una soglia prefissata, ad esempio 0.1 (Zhu, Coffman e Ghosh 2015, Wu et al. 2022).

Misura locale: per una qualità 'ottima' della pseudo-popolazione, in ogni blocco, la correlazione tra le covariate e il trattamento dovrebbe essere, in media, uguale a zero. Nella pratica si assume che la condizione di bilanciamento delle covariate

sia soddisfatta quando

$$\left| \frac{\sum_{i=1}^{r_k} n_{ik} \mathbf{x}_{ik}^{\dagger}}{\sum_{i=1}^{r_k} n_{ik}} - \frac{\sum_{k' \neq k} \sum_{i=1}^{r_{k'}} n_{ik'} \mathbf{x}_{ik'}^{\dagger}}{\sum_{k' \neq k} \sum_{i=1}^{r_{k'}} n_{ik'}} \right| < \boldsymbol{\epsilon}_2 \tag{2.3}$$

dove ogni elemento del vettore p-dimensionale $\boldsymbol{\epsilon}_2$ è uguale ad una soglia prefissata, ad esempio 0.2 (Harder, Stuart e Anthony 2010, Wu et al. 2022).

2.3.2 Algoritmo di matching

L'algoritmo è composto dai seguenti passaggi:

1. si stima il modello di Besag-York-Mollié e se ne ottengono i valori interpolati;
2. si calcola il *GPS* tramite la densità di una distribuzione normale

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\mathbf{t} - \boldsymbol{\mu}}{\sigma} \right)^2 \right\}$$

   dove $\mathbf{t}$ è il vettore relativo al trattamento e $\boldsymbol{\mu}$ è il vettore dei valori predetti dal *BYM*. Si indicherà con $\hat{e}(t_j, \mathbf{x}_j)$ la stima di *GPS* ottenuta per un'arbitraria unità j, con livello di trattamento t_j e covariate $\mathbf{x}_j$;
3. si definisce una griglia di valori che suddivide il dominio del trattamento:

$$[t^{(m)} - \Delta_n, t^{(m)} + \Delta_n] \quad \text{m} = 1,2,...,\text{M}.$$

   Per la scelta di Δ_n si veda la Sezione 2.3.3.
4. a tal punto, per ogni valore di m, si esegue il seguente ciclo:
   4.1. dato m si crea un nuovo insieme ipotetico di osservazioni (j') con identiche covariate, ma aventi tutte lo stesso valore di trattamento $t^{(m)}$;
   4.2. per ogni osservazione j', si cerca un'osservazione appartenente al dataset originale j tale che

- j abbia valore di trattamento t_j tale che

$$t_j \in [t^{(m)} - \Delta_n, t^{(m)} + \Delta_n];$$

- j sia il più vicino a j' rispetto ad una metrica bidimensionale che include il trattamento e il *GPS* stimato.

4.3. si crea un sottoinsieme con tutte le osservazioni j che soddisfano la condizione

$$|t_j - t^{(m)}| \leq \Delta_n;$$

4.4. si calcola il *GPS* per ogni osservazione j' e si sceglie poi il valore di j che minimizza la distanza L_1 tra j e j', dove lo spazio bidimensionale è determinato dal valore di trattamento e *GPS*. In particolare

$$j_{GPS}(e^{(m)}_{j'}, t^{(m)}) = \underset{j:t_j \in [t^{(m)}-\Delta_n, t^{(m)}+\Delta_n]}{\operatorname{argmin}} ||(\lambda \hat{e^*}(t_j, \mathbf{x}_j), (1-\lambda)t^*_j) - (\lambda \hat{e^{(l)*}_{j'}}, (1-\lambda)t^*_j)|| \tag{2.4}$$

dove $||\cdot||$ è una metrica bidimensionale, $\lambda \in [0, 1]$ un parametro di scala che assegna i pesi alle due dimensioni e Δ_n è il raggio che definisce il dominio del trattamento, definito in precedenza. Per la scelta del parametro di scala λ si veda la Sezione 2.3.3.

4.5. trovato il valore j che minimizza questa distanza, si associa all'osservazione j' l'indice di riga di j. In altri termini si impone

$$\hat{Y}_{j'}(w^{(m)}) = Y^{obs}_{j_{GPS}};$$

Semplificando, se si ottengono ad esempio:

$$a = \begin{bmatrix} 0.5 \\ 3 \\ 10 \\ 0.7 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 2 \\ 16 \\ 8 \\ 0.5 \\ 3 \end{bmatrix}$$

dove a è il *GPS* normalizzato per il sottoinsieme e b è il *GPS* normalizzato per l'insieme completo, allora il vettore di associazione sarà $b_{index} = (4, 2, 3, 3, 1, 2)$;

4.6. si costruisce una tabella di frequenza per gli indici di riga ottenuti. Facendo riferimento all'esempio precedente si ottiene la Tabella 2.1.

	id	N
1	1	1
2	2	2
3	3	2
4	4	1

Tabella 2.1: Tabella di frequenza per gli indici di riga.

5. si sommano le m tabelle di frequenza ottenute dal ciclo del punto *4*. Il passaggio è particolarmente rilevante perché i risultati di questa somma rappresentano i pesi della pseudo-popolazione finale;

6. data la pseudo-popolazione si effettua un test di bilanciamento delle covariate (Sezione 2.3.1) e si possono verificare due situazioni:

   - la correlazione soddisfa le condizioni date in input: l'algoritmo si ferma e restituisce la pseudo-popolazione finale;
   - la correlazione non soddisfa le condizioni date in input: l'algoritmo applica una trasformazione sulla covariata x_j che presenta la correlazione massima con il trattamento (di default x_j^2 o x_j^3) e riparte dalla stima del propensity score generalizzato.

Nella descrizione dell'algoritmo si sono introdotti due parametri, Δ_n e λ: nella sezione successiva si delineano alcune linee guida per la determinazione del loro valore ottimo.

2.3.3 Selezione degli iperparametri Δ_n e λ

Nell'algoritmo in Sezione 2.3.2 è stato definito un valore di Δ_n tramite cui è prodotta la suddivisione del dominio del trattamento, per poi proseguire con la procedura di matching e la creazione della pseudo-popolazione. Non esistono

metodi con una base teorica solida per la scelta di questo parametro, ma è possibile procedere in modo analogo a quando è necessario selezionare una larghezza di banda nei metodi non parametrici di lisciamento: si considera una griglia di valori e si sceglie il termine che fornisce i risultati i migliori (Khoshnevis, Wu e Braun 2023). In questo caso il procedimento è analogo, e la qualità dei risultati è determinata dalla correlazione massima che si osserva tra trattamento e covariate nella pseudo-popolazione. Si procede come di seguito:

1. si determina una griglia di valori candidati per Δ_n relativamente piccoli;
2. per ogni valore della griglia si eseguono i successivi due passi:
   2.1. si costruisce la pseudo-popolazione per ogni particolare valore di Δ_n;
   2.2. si calcola la massima correlazione assoluta (tra trattamento e covariate) per la pseudo-popolazione prodotta;
3. si prende il valore di Δ_n che minimizza la correlazione assoluta;
4. passo facoltativo: se è stata scelta una griglia iniziale di valori non molto fitta, può essere opportuno effettuare uno zoom nell'intorno del Δ_n considerato - definendo quindi una nuova griglia - e poi ripetere il punto 2.

Il parametro λ (parametro di scala) regola l'importanza da assegnare al valore di *GPS* e di trattamento per calcolare la distanza che separa due osservazioni. Esso dovrebbe assumere quindi un valore vicino ad 1, affinché si dia particolare rilevanza al *GPS* nella procedura di matching (Equazione 2.4). Analogamente a come fatto per Δ_n, si costruisce una possibile griglia di valori per λ e poi si procede con lo stesso algoritmo.

2.3.4 Possibili miglioramenti

Nel caso in cui le correlazioni assolute tra il trattamento e le covariate siano ancora troppo grandi si tenta di arginare il problema in diversi modi (Khoshnevis, Wu e Braun 2023). È possibile:

1. ridurre l'analisi ad un supporto limitato del trattamento;
2. ridurre l'analisi ad un supporto limitato di *GPS*;
3. introdurre ulteriori possibili trasformazioni delle covariate (si ricorda - Sezione 2.3.2 - che per default si propongono come trasformazioni solo il quadrato (x^2) ed il cubo (x^3) della variabile).

Sebbene possano portare effettivi miglioramenti nella pseudo-popolazione, i punti *1-2* impongono anche inevitabili compromessi nello studio e nell'interpretazione dei risultati. La riduzione del supporto del trattamento, ad esempio, costringe ad eliminare parte delle osservazioni disponibili e produce delle variazioni nella pseudo-popolazione rispetto al database originale.

2.4 Riepilogo del procedimento e stima dell'exposure response curve

2.4.1 L'algoritmo, in breve

Dati i numerosi passi per raggiungere il risultato finale, si ritiene utile riassumere tutto il procedimento (rappresentato in modo dettagliato in Figura 2.1) in pochi punti che possano aiutare a fare chiarezza:

1. valutazione del bilanciamento delle covariate sul database di partenza;
2. stima del modello di Besag-York-Mollié e del *GPS*;
3. costruzione della pseudo-popolazione tramite l'algoritmo di matching;
4. test di valutazione del bilanciamento delle covariate per la pseudo-popolazione:
   - test passato $\longrightarrow$ viene restituita la pseudo-popolazione prodotta;
   - test non passato $\longrightarrow$ si applica una trasformazione alla covariata con maggiore correlazione con il trattamento e si riparte dal punto 2.

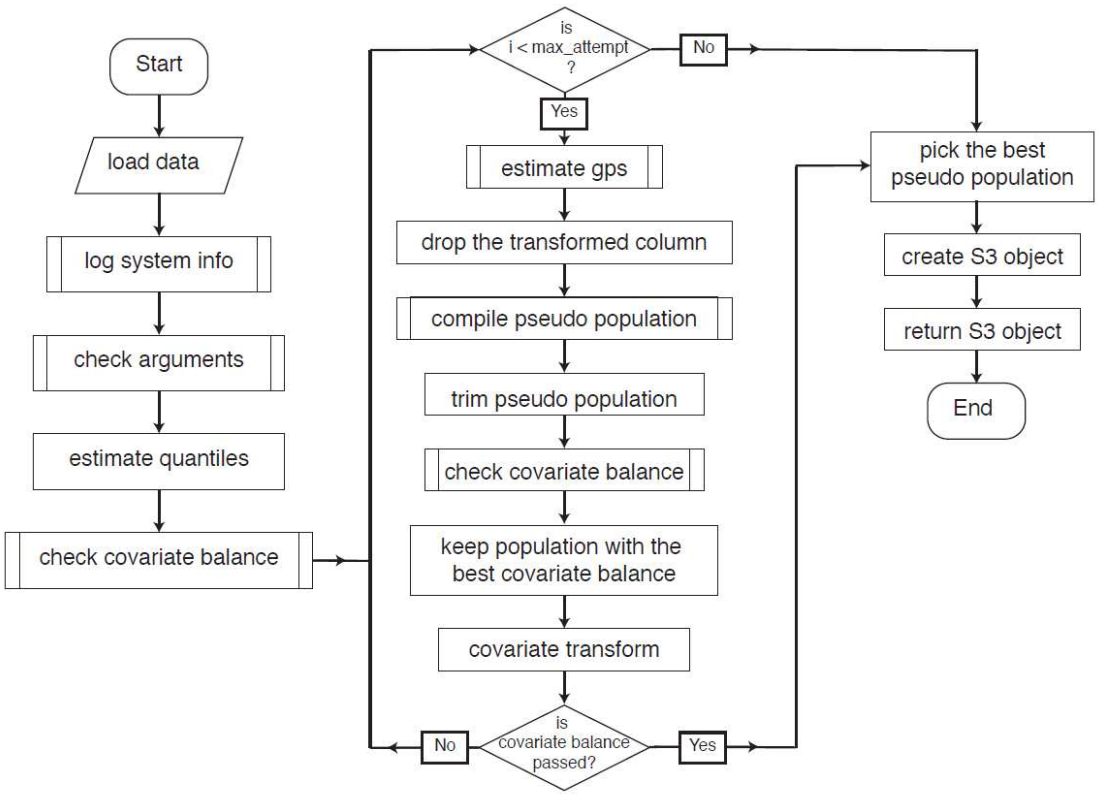


Figura 2.1: Algoritmo completo per l'implementazione della procedura di matching e per la generazione della pseudo-popolazione (da Khoshnevis, Wu e Braun 2023).

2.4.2 Exposure response curve

Se la pseudo-popolazione ottenuta soddisfa le condizioni di bilanciamento delle covariate, è possibile condurre un'analisi sul nuovo database e ricavare dei risultati che possono essere anche interpretati nel contesto di nessi di causalità. L'andamento della risposta in funzione del trattamento può essere ottenuto tramite metodi parametrici, semi-parametrici o non parametrici. A tal punto la scelta è completamente a carico dell'analista e dipenderà dai dati a disposizione.

Capitolo 3

Analisi sul voto

Attraverso la metodologia descritta nel Capitolo 2 si intende condurre un'analisi per valutare l'effetto del trattamento (reddito mediano o indice di Gini) sul voto dei cittadini nelle elezioni statunitensi nel 2020. Dopo una prima analisi esplorativa (Paragrafo 3.1), segue una presentazione dei risultati ottenuti, che si articola in due parti, definite dall'utilizzo dei due diversi trattamenti. Nella prima il trattamento è il logaritmo del reddito mediano, mentre nella seconda è definito dal logaritmo dell'indice di Gini (la risposta e le covariate rimangono invece le medesime in entrambi i casi). Come visto nel Paragrafo 2.3.4 quando la pseudo-popolazione non soddisfa i criteri richiesti per il bilanciamento delle covariate è possibile produrre qualche miglioramento attraverso la riduzione del dominio del trattamento. Per questa ragione si sono presentati tre diversi scenari: nel primo il trattamento è delimitato tra il 1° ed il 99° percentile, nel secondo tra il 5° ed il 95° percentile e nell'ultimo tra il 10° e 90° percentile. Per la discussione e l'interpretazione dei risultati si rimanda invece al Capitolo 4.

3.1 Analisi esplorativa

Prima di procedere con la presentazione dei risultati è di interesse valutare la distribuzione della risposta, dei trattamenti e delle covariate a disposizione.

3.1.1 Risposta: percentuale di voti al Partito Democratico

In primo luogo è di interesse valutare come si distribuiscono i voti del Partito Democratico in tutte le contee degli Stati Uniti. In Figura 3.1a si osserva che le regioni a maggioranza democratica sono soprattutto a nord-est e sud-ovest. Inoltre in alcuni stati, come California, New York e Virginia, la prevalenza del Partito Democratico è molto forte: in alcune contee il partito raggiunge più dell'80% dei voti.

Al contrario, nella zona centrale le percentuale dei voti diminuiscono. Il Texas ad esempio (che include 254 contee) è uno stato a forte maggioranza repubblicana e in alcune contee il Partito Democratico, consapevole di raccogliere pochissimi consensi, non si presenta nemmeno (contee completamente bianche in Figura 3.1a). Situazioni analoghe si ripresentano anche in Utah, Kansas ed Oklahoma.

3.1.2 Trattamenti: log-reddito mediano e log-indice di Gini

Le distribuzioni dei due trattamenti sono riportate in Figura 3.1b e Figura 3.2a; in entrambi i casi si è preferito mostrare, per chiarezza espositiva, il logaritmo delle due quantità. La contea di Whitley County, in Kentucky, ha il reddito mediano più basso osservato, pari a 22.292 dollari l'anno, mentre la contea con il reddito mediano più alto, di 147.111 dollari annui, è Loudoun County in Virginia. Le contee con i redditi mediani più alti si trovano a nord est e sud ovest degli USA: in particolare in California, Pennsylvania e New Jersey. Inoltre anche certe aree del Texas e della Florida sono molto ricche, insieme a zone del South Dakota e dell'Illinois.

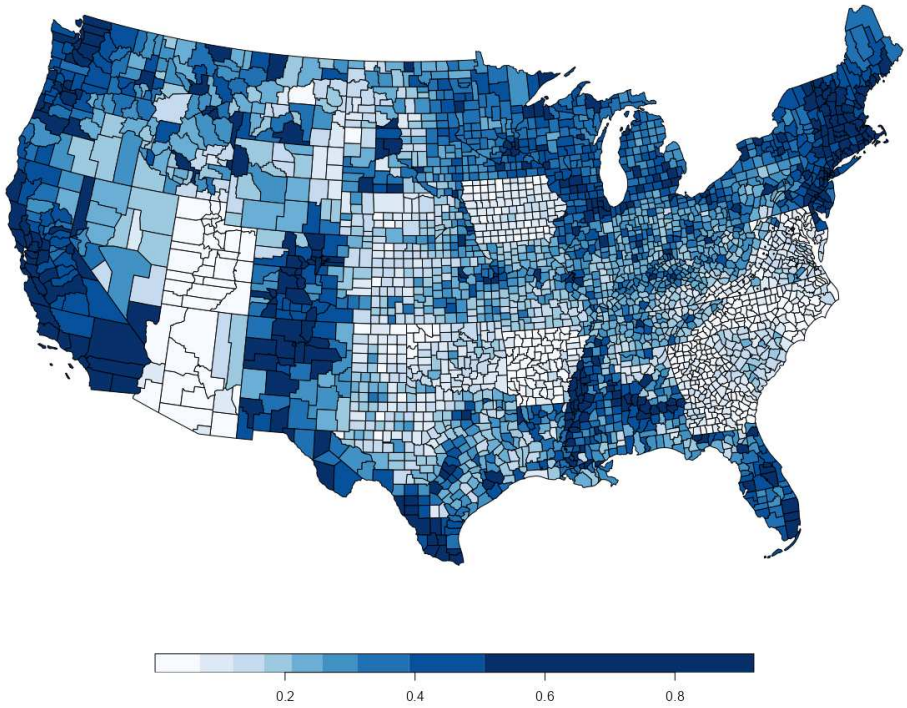
L'indice di Gini mostra invece una distribuzione differente: i valori più alti si osservano prevalentemente a sud-est e in alcune aree a sud-ovest degli USA. La contea con il livello massimo di disuguaglianza, Harding County, si trova in New Mexico e ha un indice di Gini pari a 0.696; quella invece maggiormente uniforme è Kenedy County in Texas, con indice di Gini pari a 0.291.

3.1.3 Variabili utilizzate all'interno della procedura di matching

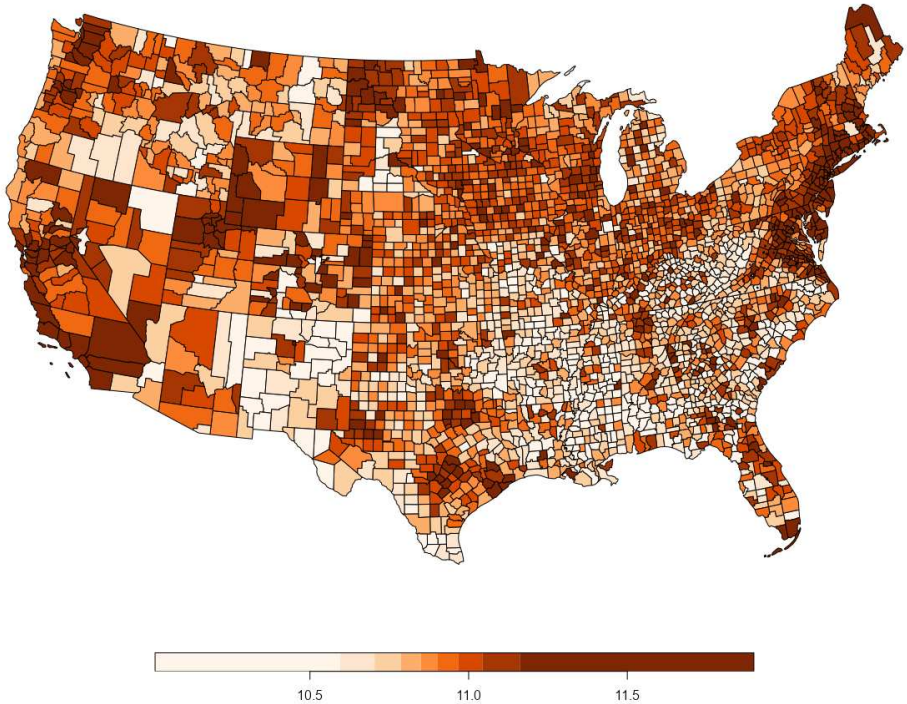
Le variabili utilizzate all'interno della procedura di matching sono lo stato a cui appartiene la contea, la densità abitativa, la percentuale di persone laureate, la percentuale di persone con un'assicurazione sanitaria, la percentuale di disoccupati e la percentuale di persone bianche.

In Figura 3.2b si riporta la distribuzione del logaritmo della densità abitativa per le diverse contee. Si osserva innanzitutto che c'è un gradiente molto ampio tra le zone degli Stati Uniti: essendo un territorio vastissimo vi sono alcune contee in cui la densità abitativa non raggiunge un abitante per chilometro quadrato, mentre altre nelle quali si superano i settantamila abitanti. Oregon, Colorado, Arizona, Minnesota e Oklahoma sono alcuni degli stati meno densamente popolati, mentre Connecticut, New Jersey e New York alcuni di quelli più popolosi (tutti a nord est degli Stati Uniti).

Per quanto riguarda le altre variabili (Figura 3.3) si osserva che la percentuale di laureati e di persone bianche sono le due covariate che variano di più tra le contee: la prima si muove nell'intervallo [0, 79.3], per la seconda il dominio è definito da [0.05, 1]. Per la percentuale di persone con assicurazione sanitaria (supporto dato da [57.4, 99.5]) vi sono alcune contee nelle quali quasi tutte le persone possiedono un'assicurazione sanitaria, mentre altre in cui tale percentuale si ferma poco prima del 60%. La percentuale di persone disoccupate è la covariata con minor variabilità, i cui valori sono maggiormente centrati intorno alla media (supporto: [0, 32.2]).

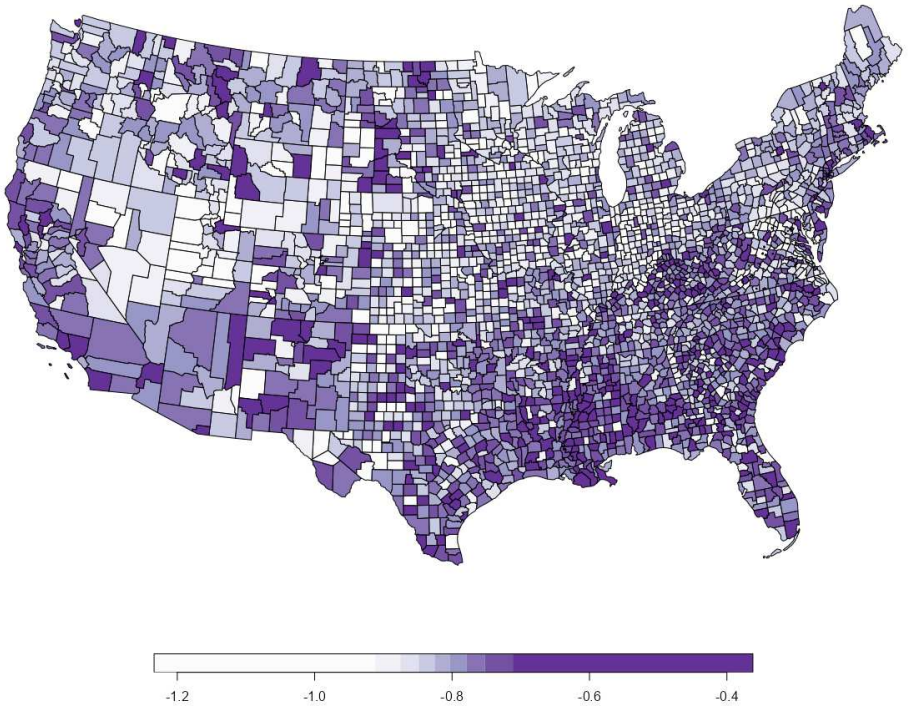


(a) Distribuzione spaziale della percentuale di voti ottenuta dal Partito Democratico alle elezioni statunitensi del 2020.

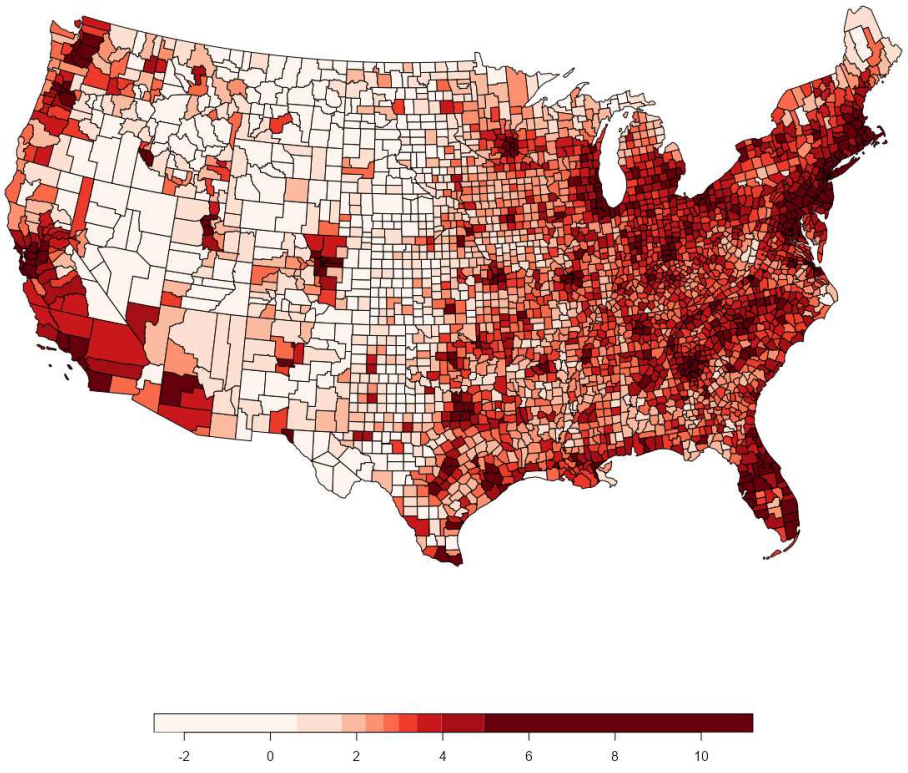


(b) Distribuzione spaziale del logaritmo del reddito mediano.

Figura 3.1: Distribuzione spaziale della percentuale di voti ottenuta dal Partito Democratico (a) e del logaritmo del reddito mediano (b).



(a) Distribuzione spaziale del logaritmo dell'indice di Gini.



(b) Distribuzione spaziale del logaritmo della densità abitativa.

Figura 3.2: Distribuzione spaziale del logaritmo dell'indice di Gini (a) e del logaritmo della densità abitativa (b).

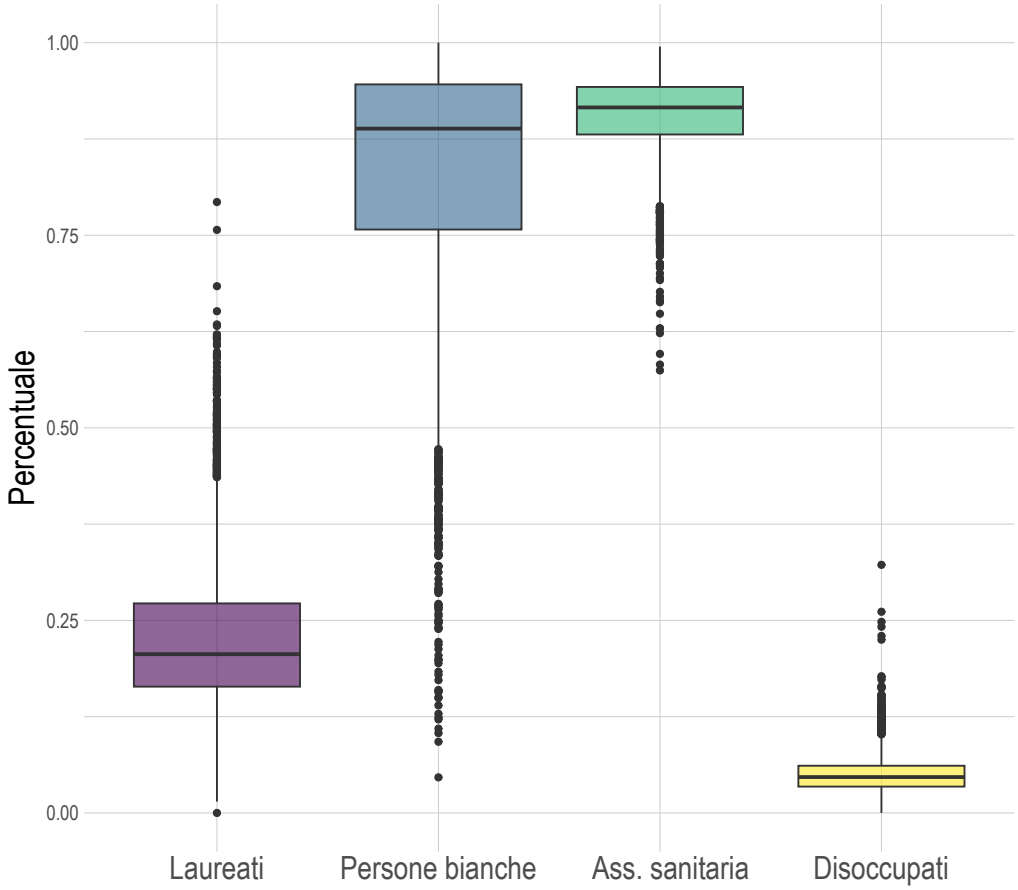


Figura 3.3: Distribuzione delle variabili associate alla percentuale di persone laureate, alla percentuale di persone bianche, alla percentuale di persone con assicurazione sanitaria e alla percentuale di persone disoccupate.

3.2 Risultati

Per la generazione della pseudo-popolazione e l'analisi dei risultati si segue un procedimento comune per entrambi i trattamenti. In primo luogo si imposta il test di bilanciamento: esso risulta soddisfatto quando la correlazione assoluta massima tra le covariate e il trattamento è inferiore a 0.1, ovvero quando tutte le covariate hanno una correlazione assoluta con il trattamento minore di 0.1. L'algoritmo si arresta quando tale test è soddisfatto, oppure quando si raggiunge il numero massimo di iterazioni possibili (impostato uguale a 10). Successivamente si valutano diversi valori per gli iperparametri Δ_n e λ e si selezionano quelli associati alla pseudo-popolazione che presenta la minor correlazione assoluta massima tra le covariate e il trattamento. In questa fase si è osservato che λ era poco rilevante e che si ottenevano i risultati migliori (e con costi computazionali inferiori) fissando $\lambda = 1$, ovvero dando maggiore peso possibile al *GPS* nella procedura di matching: per questa ragione si è scelto questo valore per tutti i risultati successivi.

Ottenuta la pseudo-popolazione migliore si determinano le correlazioni aggiustate tra le covariate e il trattamento e si valuta se il test di bilanciamento risulta soddisfatto: in caso affermativo si può procedere con l'analisi, altrimenti è necessario effettuare alcune modifiche, come la riduzione del dominio del trattamento, e ricalcolare la pseudo-popolazione. Nel corso di questo studio si è deciso comunque, per completezza e in termini di confronto, di proseguire l'analisi anche nel caso in cui il test non risultasse soddisfatto.

Per valutare la relazione tra il voto al Partito Democratico e il trattamento si è in primo luogo stimato un modello lineare, senza covariate, sul database iniziale:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2) \tag{3.1}$$

dove y_i rappresenta la percentuale di voti al candidato democratico nella contea i-esima, β_0 l'intercetta, x_i il trattamento, β_1 il coefficiente associato a quest'ultimo e ϵ_i l'errore statistico.

Successivamente, si stima un modello analogo, ma utilizzando la pseudo-popolazione e i pesi calcolati tramite la procedura di matching. Si confronteranno quindi, di

volta in volta, due modelli: il modello lineare, senza covariate e senza pesi, calcolato utilizzando tutti i dati a disposizione, e il modello lineare pesato che tiene conto delle covariate nella costruzione della pseudo-popolazione.

3.2.1 Risultati: relazione tra reddito mediano e voto

Il primo trattamento preso in considerazione è il logaritmo del reddito mediano. Come visto in precedenza (Sezione 2.3.4), quando la pseudo-popolazione non soddisfa il test di bilanciamento delle covariate è possibile prendere alcuni accorgimenti per migliorarla in termini di correlazione assoluta tra trattamento e covariate. In particolare, seppur con delle controindicazioni, il modo più efficace è limitare il dominio del trattamento: per questa ragione vengono presentati tre diversi risultati, con restrizioni sul logaritmo del reddito mediano differenti.

Prima di procedere, si stima il modello lineare in Equazione 3.1 assumendo come trattamento (x_i, nel modello) il logaritmo del reddito mediano: esso fungerà da termine di confronto per i risultati successivi (Tabella 3.1). Si ottiene $\beta_1 = 0.090$ con p-value associato minore di 0.001, che suggerisce un aumento della percentuale di voto quando aumenta il reddito mediano nella contea.

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001

Tabella 3.1: Stima del modello lineare senza covariate sul database inziale.

Dominio del trattamento tra il 1° e 99° percentile

In primo luogo si prende un intervallo tra il 1° ed il 99° percentile, tale che il supporto del trattamento passi da $[10.01, 11.90]$ totale a $[10.29, 11.56]$ ridotto. La miglior pseudo-popolazione ottenuta in tal caso è quella corrispondente al valore di $\Delta_n = 0.04$ (si ricorda che è sempre stato fissato $\lambda = 1$). Dopo un massimo di 10 iterazioni il test di bilanciamento delle covariate non risulta soddisfatto, come mostrato in Figura 3.4. In particolare la correlazione massima aggiustata è 0.143

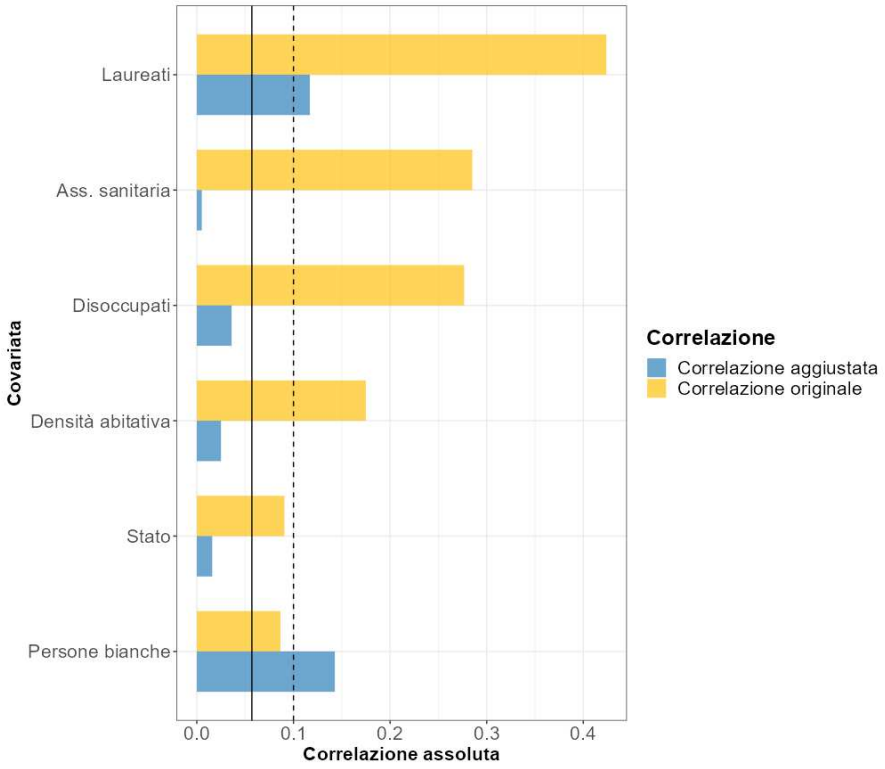


Figura 3.4: Correlazione assoluta tra il log-reddito mediano e le covariate quando il dominio del trattamento è tra il 1° e il 99° percentile. In giallo è riportata la correlazione originale (calcolata sul database iniziale), in blu la correlazione aggiustata (calcolata sulla pseudo-popolazione). La linea nera continua è posta in corrispondenza della media delle correlazioni assolute, quella tratteggiata in corrispondenza della correlazione massima data in input per il test di bilanciamento.

(partendo da quella originale uguale a 0.424) e quella media 0.069 (partendo da quella originale uguale a 0.223).

Sebbene il test di bilanciamento non risulti soddisfatto, si stima il modello lineare pesato per la pseudo-popolazione (Figura 3.5, Tabella 3.2). I coefficienti dei modelli sono entrambi positivi e significativi, anche se la pendenza della retta è inferiore per il primo modello, mostrando una minor associazione tra il voto e il reddito mediano.

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001
Modello lineare pesato	0.035	< 0.001

Tabella 3.2: Confronto tra il modello lineare senza covariate (stimato sul database inziale), e il modello lineare pesato (stimato sulla pseudo-popolazione).

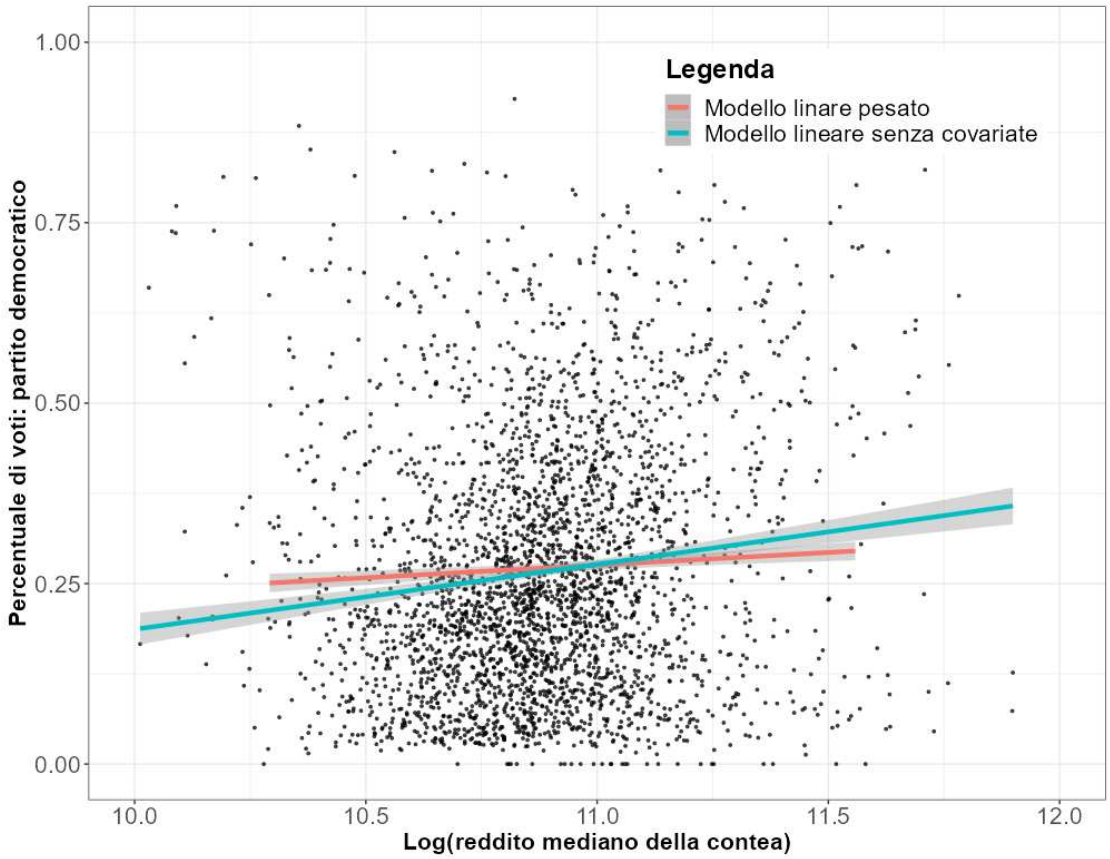


Figura 3.5: Modelli lineari per la relazione tra il log-reddito mediano e il voto al Partito Democratico. In rosso si riporta il modello lineare pesato (con i pesi ottenuti tramite la costruzione della pseudo-popolazione), mentre in azzurro il modello lineare classico.

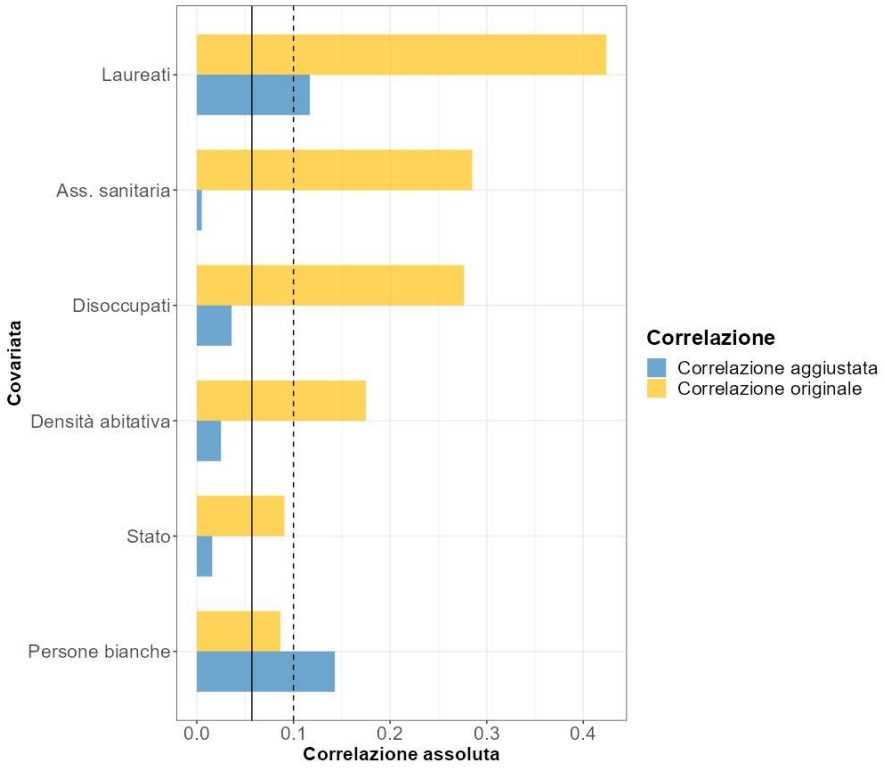


Figura 3.6: Correlazione assoluta tra il log-reddito mediano e le covariate quando il dominio del trattamento è tra il 5° e il 95° percentile. In giallo è riportata la correlazione originale (calcolata sul database iniziale), in blu la correlazione aggiustata (calcolata sulla pseudo-popolazione). La linea nera continua è posta in corrispondenza della media delle correlazioni assolute, quella tratteggiata in corrispondenza della correlazione massima data in input per il test di bilanciamento.

Dominio del trattamento tra il 5° e 95° percentile

Riducendo il dominio del trattamento tra il 5° e il 95° percentile, si passa da un supporto totale di [10.01, 11.90] a quello limitato tra [10.47, 11.32]. In modo analogo a quanto fatto in precedenza, dopo aver calcolato la pseudo-popolazione (con $\Delta_n = 0.01$) si è determinata la correlazione assoluta tra covariate e trattamento (Figura 3.6).

In questo frangente la correlazione massima aggiustata è 0.138 (correlazione originale uguale a 0.378), mentre quella media 0.069 (correlazione originale uguale a 0.198). Si nota quindi che dopo 10 iterazioni non risulta ancora soddisfatta

la condizione relativa al bilanciamento delle covariate; inoltre il miglioramento rispetto al caso precedente è minimo, in quanto la correlazione assoluta massima si riduce solo di 0.005.

Il confronto tra i due modelli produce il risultato riportato in Tabella 3.3 e in Figura 3.7. In questo caso il coefficiente del modello lineare pesato, oltre a ridursi in valore assoluto, non risulta significativo.

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001
Modello lineare pesato	0.026	0.058

Tabella 3.3: Confronto tra il modello lineare senza covariate (stimato sul database iniziale), e il modello lineare pesato (stimato sulla pseudo-popolazione).

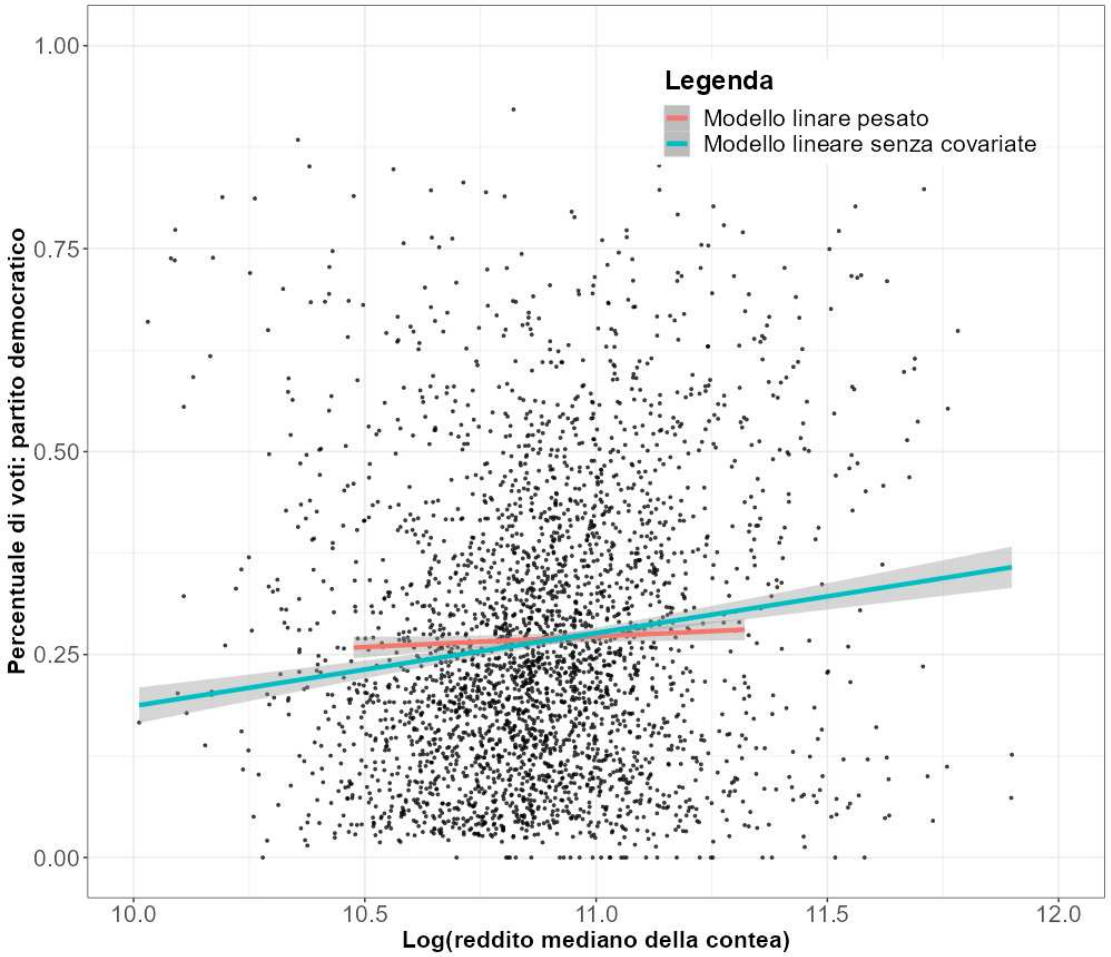


Figura 3.7: Modelli lineari per la relazione tra il log-reddito mediano e il voto al Partito Democratico. In rosso si riporta il modello lineare pesato (con i pesi ottenuti tramite la costruzione della pseudo-popolazione), mentre in azzurro il modello lineare classico.

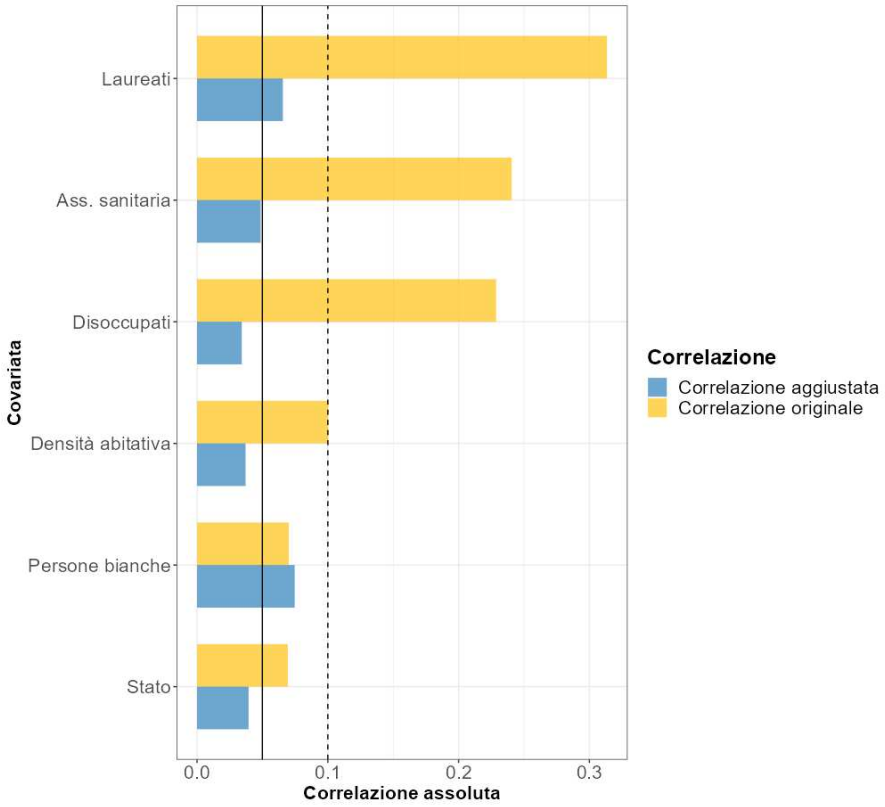


Figura 3.8: Correlazione assoluta tra il log-reddito mediano e le covariate quando il dominio del trattamento è tra il 10° e il 90° percentile. In giallo è riportata la correlazione originale (calcolata sul database iniziale), in blu la correlazione aggiustata (calcolata sulla pseudo-popolazione). La linea nera continua è posta in corrispondenza della media delle correlazioni assolute, quella tratteggiata in corrispondenza della correlazione massima data in input per il test di bilanciamento.

Dominio del trattamento tra il 10° e 90° percentile

Riducendo il dominio del trattamento tra il 10° e il 90° percentile, si passa da un supporto totale di [10.01, 11.90] a quello limitato tra [10.58, 11.18]. In modo analogo a quanto fatto in precedenza, dopo aver calcolato la pseudo-popolazione (con $\Delta_n = 0.03$) si è determinata la correlazione assoluta tra covariate e trattamento (Figura 3.8).

La correlazione massima aggiustata risulta 0.075 (originale uguale a 0.313), mentre quella media 0.050 (originale uguale a 0.170). Dopo un'unica iterazione la

condizione di bilanciamento è soddisfatta e l'algoritmo si arresta.

Anche in questo caso si riporta la stima dei due modelli (Tabella 3.4 e Figura 3.9). Si osserva che i coefficienti, entrambi significativi, mostrano un andamento contrario: nel modello lineare senza covariate all'aumentare del reddito mediano aumenta anche la percentuale di voti al partito (pendenza della retta positiva), mentre nel modello lineare pesato si riscontra l'andamento opposto (pendenza della retta negativa).

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001
Modello lineare pesato	-0.078	< 0.001

Tabella 3.4: Confronto tra il modello lineare senza covariate (stimato sul database iniziale) e il modello lineare pesato (stimato sulla pseudo-popolazione) quando il dominio del trattamento è tra il 10° e il 90° percentile.

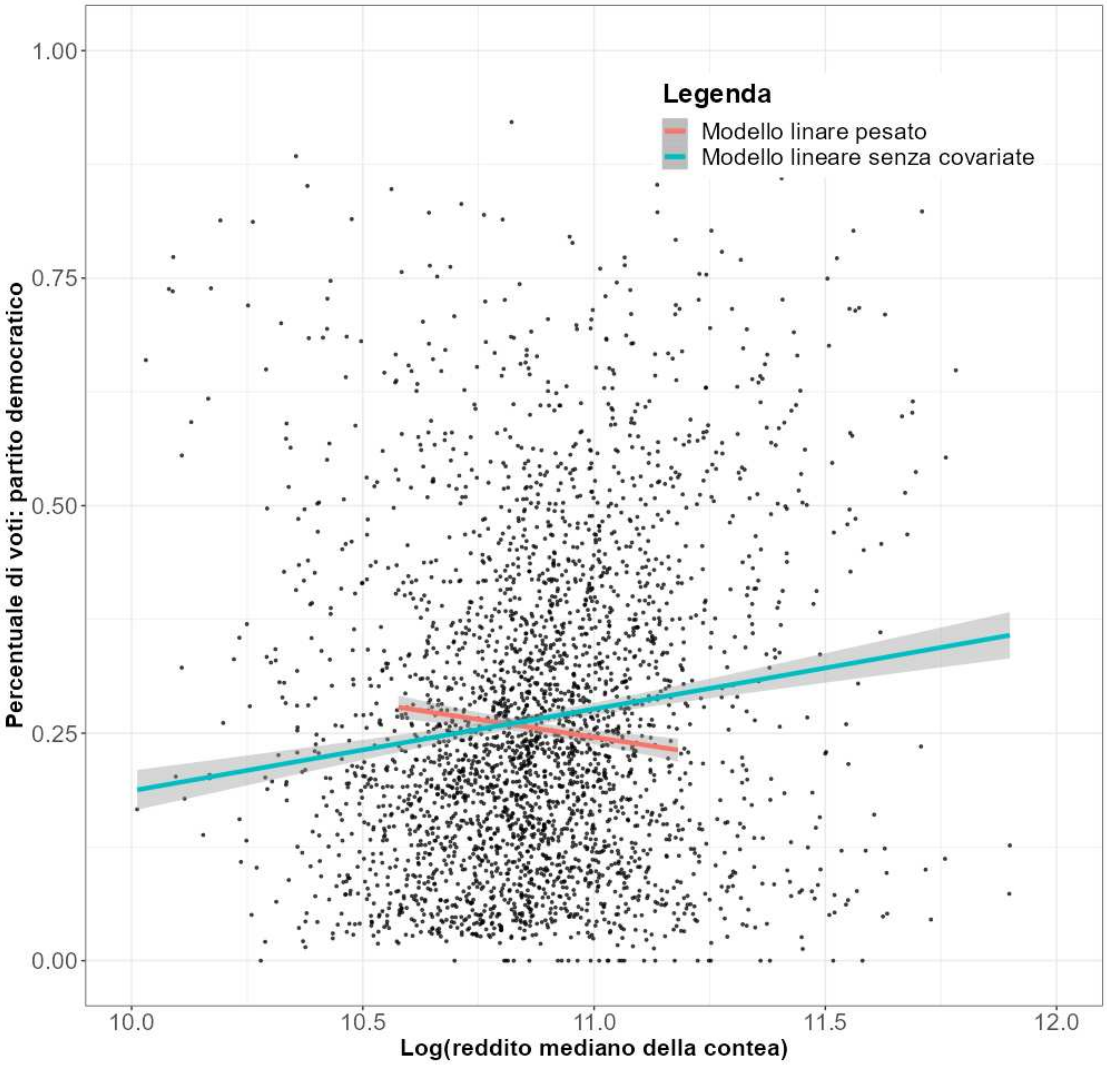


Figura 3.9: Modelli lineari per la relazione tra il log-reddito mediano e il voto al Partito Democratico quando il dominio del trattamento è tra il 10° e il 90° percentile. In rosso si riporta il modello lineare pesato (con i pesi ottenuti tramite la costruzione della pseudo-popolazione), mentre in azzurro il modello lineare classico.

3.2.2 Risultati: relazione tra indice di Gini e voto

Per il secondo trattamento, il logaritmo dell'indice di Gini, si segue una struttura analoga alla precedente.

La stima del modello lineare in Equazione 3.1 assumendo come trattamento il log-indice di Gini fornisce i risultati in Tabella 3.5. Si ottiene $\beta_1 = 0.226$ con p-value associato minore di 0.001, che suggerisce un aumento della percentuale di voto quando aumenta l'indice di Gini nella contea.

Modello	β_1	p-value
Modello lineare senza covariate	0.226	< 0.001

Tabella 3.5: Stima del modello lineare senza covariate sul database inziale.

Dominio del trattamento tra il 1° e 99° percentile

Preso l'intervallo tra il 1° ed il 99° percentile, il supporto del trattamento passa da $[-1.234, -0.362]$ totale a $[-0.998, -0.602]$ ridotto. La miglior pseudo-popolazione ottenuta in tal caso è quella corrispondente al valore di $\Delta_n = 0.01$ (si ricorda che è sempre stato fissato $\lambda = 1$). In tal caso, come mostrato in Figura 3.10, il test di bilanciamento risulta soddisfatto dopo un'unica iterazione. In particolare la correlazione massima aggiustata è 0.085 (partendo da quella originale uguale a 0.281) e quella media 0.036 (partendo da quella originale uguale a 0.152).

I risultati della stima del modello lineare pesato per la pseudo-popolazione sono riportati in Figura 3.5 e Tabella 3.6. A differenza del coefficiente per il modello lineare senza covariate (significativo e positivo), il coefficiente del modello lineare pesato risulta negativo e non significativo: ciò suggerisce che quando si corregge per i possibili confondenti, non vi è più associazione tra il livello di disuguaglianza economica e il voto.

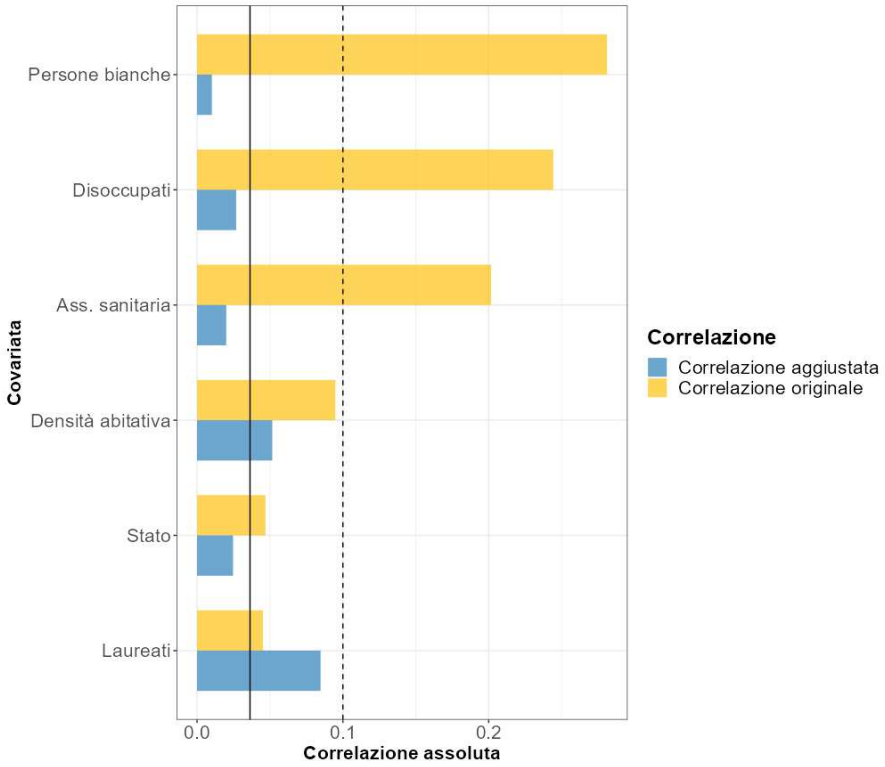


Figura 3.10: Correlazione assoluta tra il log-indice di Gini e le covariate quando il dominio del trattamento è tra il 1° e il 99° percentile. In giallo è riportata la correlazione originale (calcolata sul database iniziale), in blu la correlazione aggiustata (calcolata sulla pseudo-popolazione). La linea nera continua è posta in corrispondenza della media delle correlazioni assolute, quella tratteggiata in corrispondenza della correlazione massima data in input per il test di bilanciamento.

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001
Modello lineare pesato	-0.071	0.006

Tabella 3.6: Confronto tra il modello lineare senza covariate (stimato sul database inziale), e il modello lineare pesato (stimato sulla pseudo-popolazione).

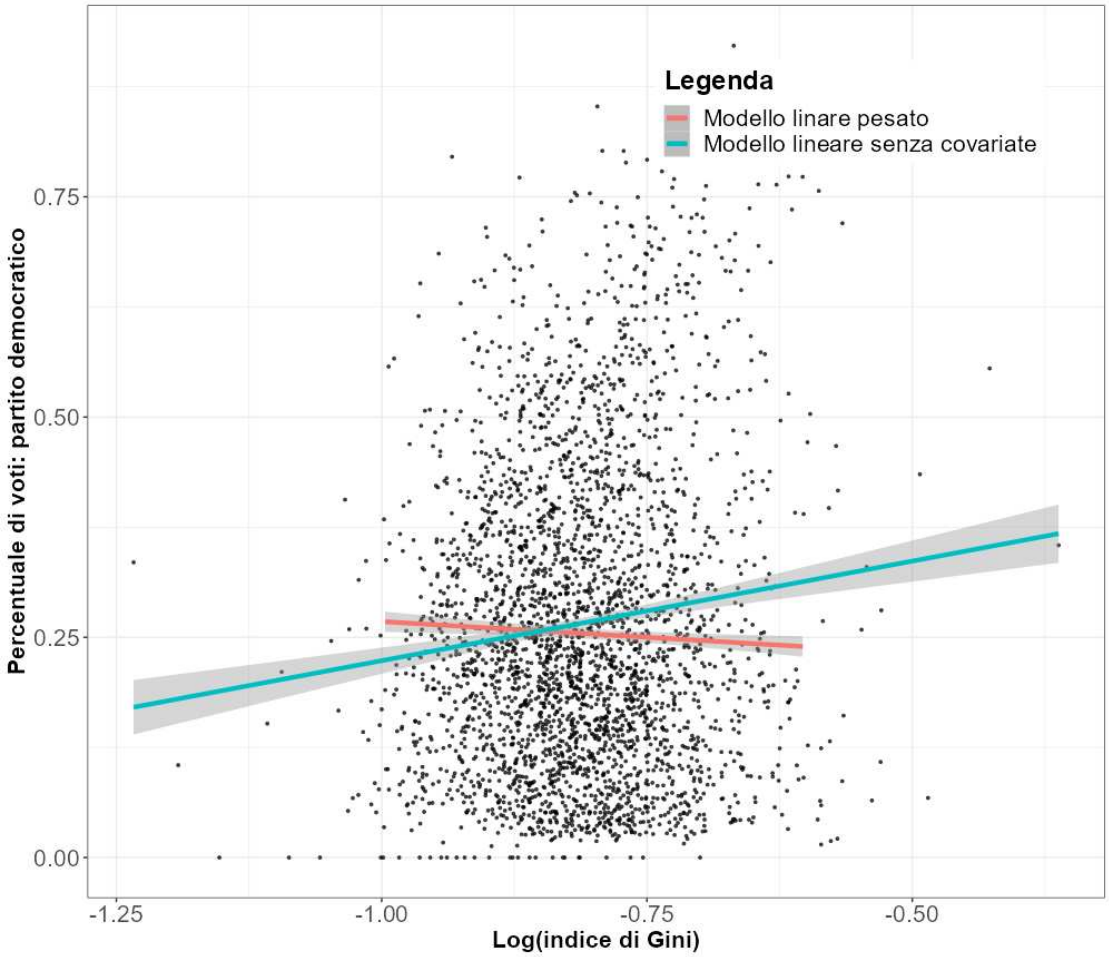


Figura 3.11: Modelli lineari per la relazione tra il log-indice di Gini e il voto al Partito Democratico. In rosso si riporta il modello lineare pesato (con i pesi ottenuti tramite la costruzione della pseudo-popolazione), mentre in azzurro il modello lineare classico.

Dominio del trattamento tra il 5° e 95° percentile

Sebbene il test di bilanciamento fosse già soddisfatto prendendo l'intervallo del trattamento tra il 1° e il 99° percentile, a scopo esplicativo e di confronto con il caso precedente, si decide di valutare anche le riduzioni del dominio del trattamento. Preso l'intervallo tra il 5° e il 95° percentile, si passa da un supporto totale di $[-1.234, -0.362]$ a quello limitato tra $[-0.945, -0.668]$. La pseudo-popolazione è stata calcolata con $\Delta_n = 0.03$; la correlazione tra le covariate e il trattamento è riportata in Figura 3.12.

La correlazione massima aggiustata risulta 0.068 (correlazione originale uguale a 0.265), mentre quella media 0.034 (correlazione originale uguale a 0.133). Anche in questo caso il test di bilanciamento risulta soddisfatto dopo una sola iterazione. Rispetto al caso precedente la correlazione massima si riduce di 0.017.

Il confronto tra i due modelli produce il risultato riportato in Tabella 3.7 e in Figura 3.13. I risultati sono analoghi ai precedenti: il coefficiente del modello lineare pesato risulta negativo e non significativo.

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001
Modello lineare pesato	-0.053	0.187

Tabella 3.7: Confronto tra il modello lineare senza covariate (stimato sul database iniziale), e il modello lineare pesato (stimato sulla pseudo-popolazione).

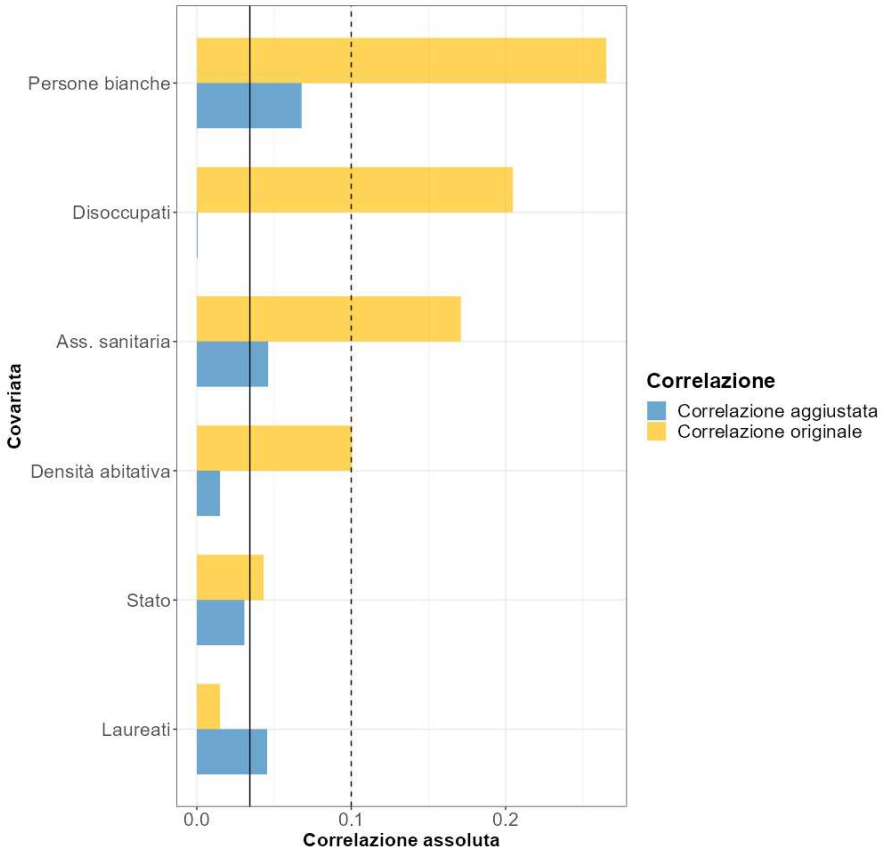


Figura 3.12: Correlazione assoluta tra il log-indice di Gini e le covariate quando il dominio del trattamento è tra il 5° e il 95° percentile. In giallo è riportata la correlazione originale (calcolata sul database iniziale), in blu la correlazione aggiustata (calcolata sulla pseudo-popolazione). La linea nera continua è posta in corrispondenza della media delle correlazioni assolute, quella tratteggiata in corrispondenza della correlazione massima data in input per il test di bilanciamento.

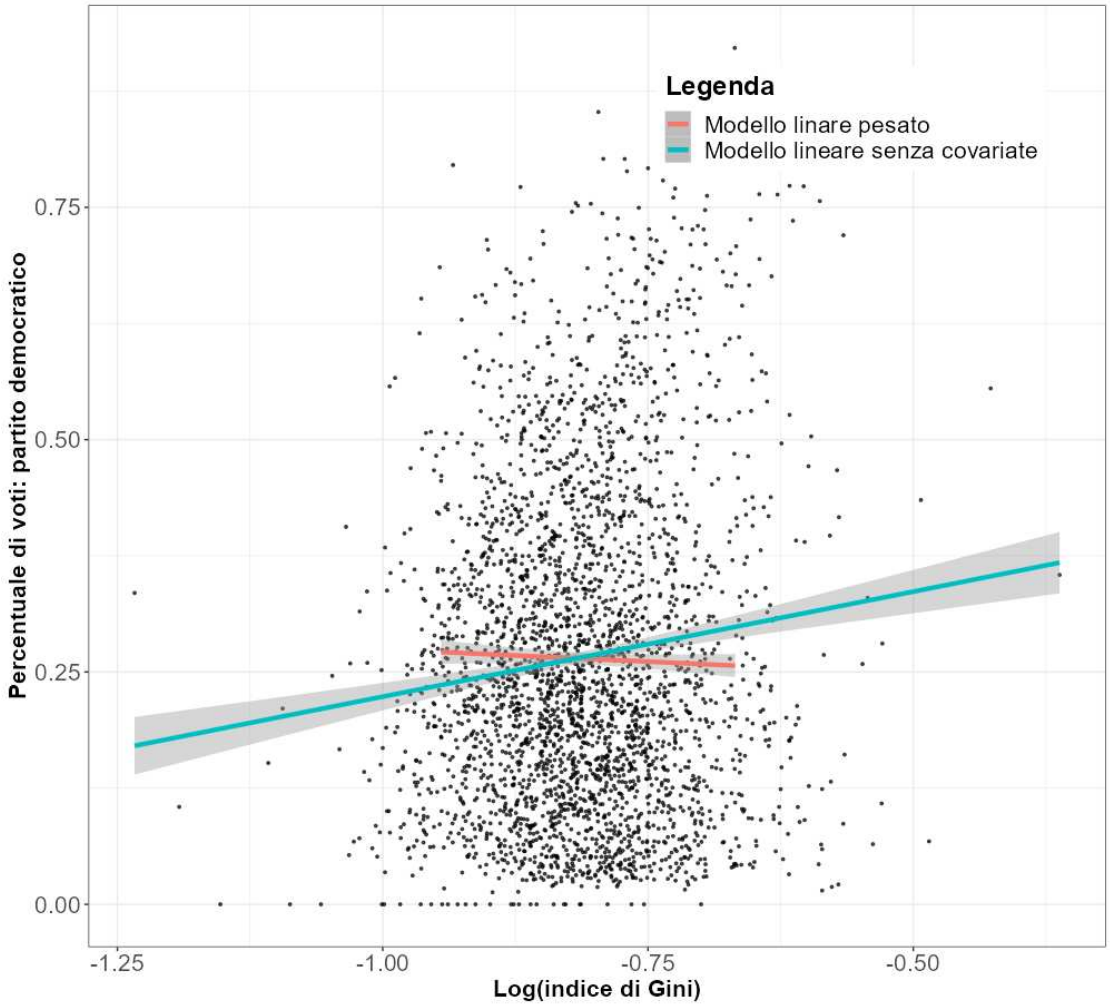


Figura 3.13: Modelli lineari per la relazione tra il log-indice di Gini e il voto al Partito Democratico. In rosso si riporta il modello lineare pesato (con i pesi ottenuti tramite la costruzione della pseudo-popolazione), mentre in azzurro il modello lineare classico.

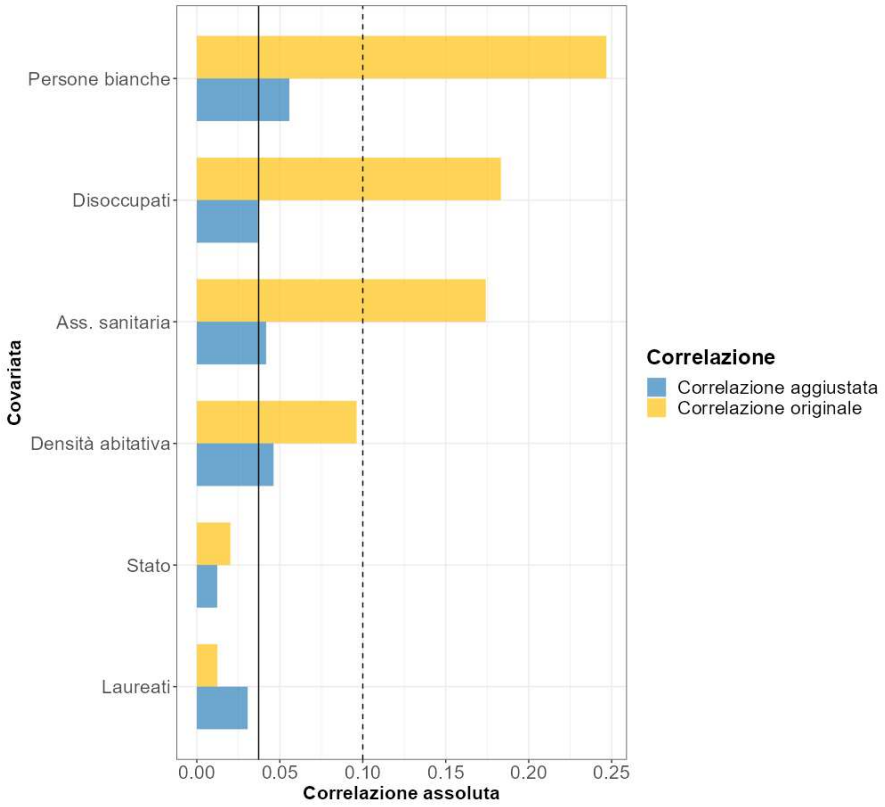


Figura 3.14: Correlazione assoluta tra il log-indice di Gini e le covariate quando il dominio del trattamento è tra il 10° e il 90° percentile. In giallo è riportata la correlazione originale (calcolata sul database iniziale), in blu la correlazione aggiustata (calcolata sulla pseudo-popolazione). La linea nera continua è posta in corrispondenza della media delle correlazioni assolute, quella tratteggiata in corrispondenza della correlazione massima data in input per il test di bilanciamento.

Dominio del trattamento tra il 10° e 90° percentile

Riducendo il dominio del trattamento tra il 10° e il 90° percentile, si passa da un supporto totale di $[-1.234, -0.362]$ a quello limitato $[-0.916, -0.706]$. La pseudo-popolazione è stata calcolata con $\Delta_n = 0.03$; la correlazione tra le covariate e il trattamento è riportata in Figura 3.14.

La correlazione massima aggiustata è 0.056 (originale uguale a 0.247), mentre quella media 0.037 (originale uguale a 0.122). Dopo un'unica iterazione la condizione di bilanciamento è soddisfatta e l'algoritmo si arresta.

Anche in questo caso si riporta la stima dei due modelli (Tabella 3.8 e Figura 3.15). Anche in questo caso il coefficiente β_1 è negativo e non significativo, mostrando come non vi sia associazione tra il livello di disuguaglianza economica e il voto al Partito Democratico.

Modello	β_1	p-value
Modello lineare senza covariate	0.090	< 0.001
Modello lineare pesato	-0.006	0.923

Tabella 3.8: Confronto tra il modello lineare senza covariate (stimato sul database inziale) e il modello lineare pesato (stimato sulla pseudo-popolazione) quando il dominio del trattamento è tra il 10° e il 90° percentile.

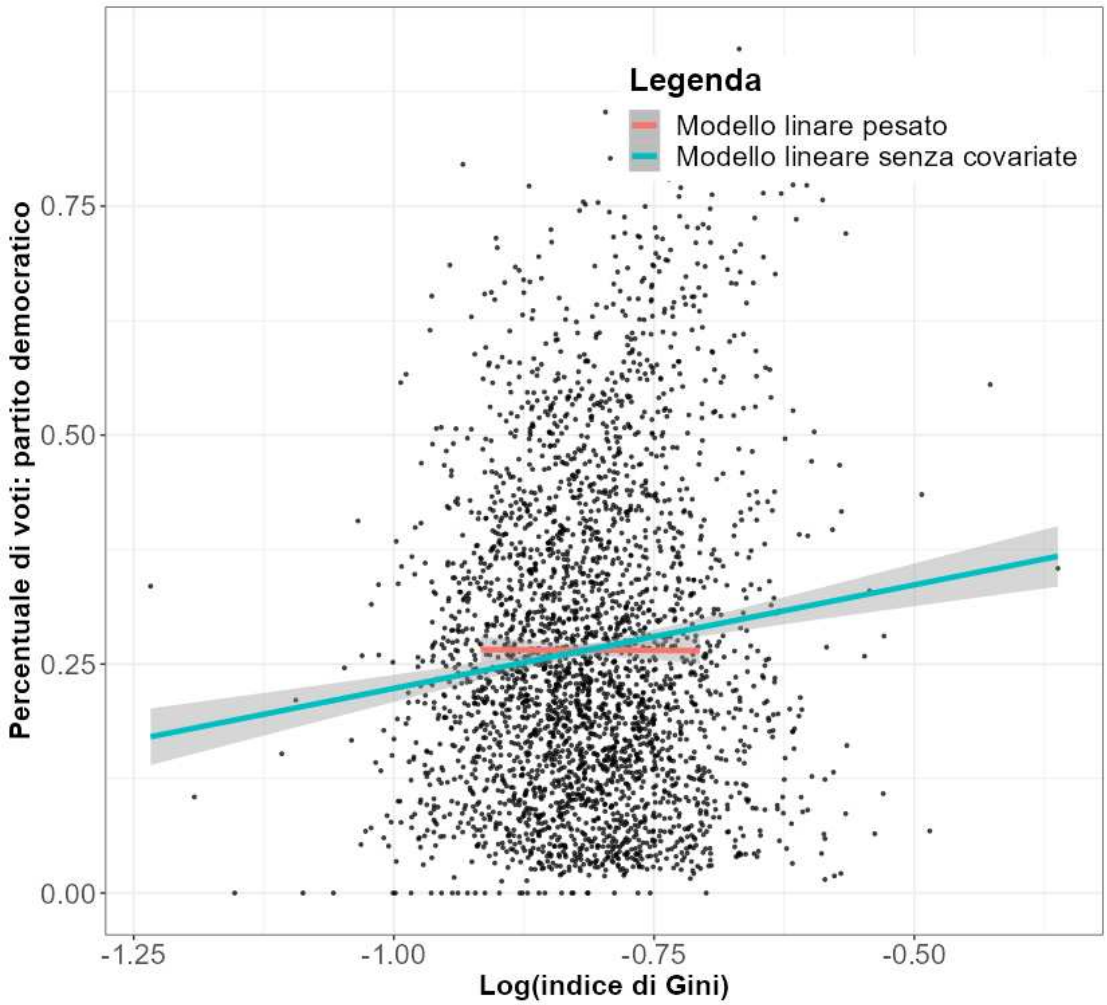


Figura 3.15: Modelli lineari per la relazione tra il log-indice di Gini e il voto al Partito Democratico quando il dominio del trattamento è tra il 10° e il 90° percentile. In rosso si riporta il modello lineare pesato (con i pesi ottenuti tramite la costruzione della pseudo-popolazione), mentre in azzurro il modello lineare classico.

Capitolo 4

Discussione dei risultati e limiti dello studio

Successivamente all'analisi discussa nel Capitolo 3, si propone una valutazione ed un'interpretazione dei risultati ottenuti per la relazione tra risposta e i due diversi trattamenti. Infine, si fornisce una descrizione dei limiti dello studio condotto tramite l'utilizzo del propensity score generalizzato e della procedura di matching.

4.1 Discussione dei risultati

L'analisi condotta ha previsto l'utilizzo di due trattamenti, il log-reddito mediano e log-indice di Gini. Oltre a presentare i risultati separatamente, si propone un confronto tra le due misure di disuguaglianza economica e la loro relazione con la percentuali di voti ottenuta dal Partito Democratico.

4.1.1 Relazione tra reddito mediano e voto

Il modello lineare presentato in Tabella 3.1, mostra un'associazione positiva significativa tra il reddito mediano e il voto al Partito Democratico. I tre risultati presentati, a causa di una limitazione differente del supporto, suggeriscono anche valutazioni differenti. Nel primo caso, per la limitazione del supporto tra il 1° e 99°

percentile, l'associazione tra trattamento e voto rimane positiva, ma è importante sottolineare che il modello è stato stimato nonostante la condizione per il bilanciamento delle covariate non fosse soddisfatta (Figura 3.4). Ciò implica che anche all'interno della pseudo-popolazione la distribuzione delle covariate non è la stessa per ogni livello del trattamento: l'effetto che le covariate potrebbero ricoprire agendo da confondenti è solo parzialmente corretto e potrebbe portare ancora ad una distorsione nei risultati. Quando il dominio del trattamento risulta compreso tra il 5° e il 95° percentile è possibile fare considerazioni analoghe, anche se in questo caso il coefficiente del modello lineare pesato risulta non significativo, suggerendo una mancata associazione tra il reddito mediano e il voto.

Affinché il test di bilanciamento risulti soddisfatto è necessario limitare il supporto del trattamento tra il 10° e 90° percentile. In questo caso la correlazione massima assoluta riscontrata nella popolazione è pari a 0.075 (correlazione tra il reddito mediano e la percentuale di persone bianche). In tal caso, è possibile assumere che gli effetti dei confondenti siano corretti, e dare un'interpretazione dei risultati anche nel contesto di nesso di causalità. Il confronto tra il modello lineare senza covariate stimato su tutto l'insieme di dati e il modello lineare pesato stimato sulla pseudo-popolazione porta a osservare due risultati opposti: mentre il coefficiente della prima retta è positivo e significativo, mostrando un aumento della percentuale dei voti alla crescita del reddito mediano, il coefficiente associato alla seconda è negativo e significativo, mostrando una diminuzione della percentuale dei voti all'aumentare del reddito mediano. Questi andamenti suggeriscono che se non si considerano i possibili confondenti nella relazione tra il reddito mediano e il voto, si corre il rischio di trarre conclusioni errate, e di osservare associazioni contrarie rispetto a quelle reali.

4.1.2 Relazione tra indice di Gini e voto

L'associazione tra l'indice di Gini e il voto al Partito Democratico, stimata tramite il modello lineare riportato in Tabella 3.5, risulta significativa e positiva. In questo caso, per ottenere una pseudo-popolazione che soddisfacesse il bilanciamento delle covariate, è stato sufficiente limitare il dominio del trattamento tra il

1° e il 99° percentile. La correlazione massima assoluta è infatti pari a 0.085. Anche utilizzando supporti del trattamento più stretti, i risultati ottenuti vanno tutti nella stessa direzione, mostrando come non vi sia alcuna associazione tra l'indice di Gini e il voto. I coefficienti ricavati risultano tutti non significativamente diversi da zero, anche se si osserva un aumento progressivo del p-value con la diminuzione del dominio del trattamento.

Nella valutazione della relazione tra il livello di disuguaglianza economica e il voto, la mancata correzione per i confondenti porta a trarre conclusioni errate, mostrando un'associazione che invece non risulta significativa.

4.1.3 Confronto tra i trattamenti

I risultati per entrambi trattamenti utilizzati hanno mostrato che è di fondamentale importanza la considerazione dei confondenti per valutare la relazione con la risposta. Nel caso del reddito mediano l'associazione viene ribaltata, mentre per l'indice di Gini non risulta più significativa. L'approccio utilizzato, tramite il propensity score generalizzato e la procedura di matching, si è dimostrato molto utile ed efficace, seppure nei suoi limiti (sezione 4.2). La generazione della pseudo-popolazione consente di limitare il problema della differente distribuzione delle covariate per i diversi livelli di trattamento, arginando quindi la distorsione nei risultati.

4.2 Limiti dello studio

Nel corso dell'analisi sono state riscontrate difficoltà e criticità, alcune delle quali imposte dalla reperibilità dei dati e altre dalla procedura di matching. Come accennato nel Paragrafo 1.4, l'utilizzo di dati aggregati per contea non è l'ideale nella valutazione dell'effetto dei trattamenti sul voto: sarebbe necessario un database composto da i dati personali dei cittadini, con le caratteristiche individuali ed il voto espresso. Data l'irreperibilità di quest'ultimi, si sono utilizzati i dati in forma aggregata. Inoltre, una futura analisi potrebbe comprendere anche variabili diverse da quelle qui utilizzate. Ad esempio, per quanto riguarda l'etnia, potrebbe

essere opportuno utilizzare il *diversity index*[1], invece che utilizzare la percentuale di persone bianche.

Per quanto riguarda la metodologia, uno dei limiti maggiori è determinato proprio dalla generazione della pseudo-popolazione: sebbene la procedura di matching sia ottimizzata per minimizzare la correlazione assoluta tra le covariate e il trattamento, non è detto che si riesca a trovare una pseudo-popolazione che soddisfi il bilanciamento delle covariate. In tal caso, non è garantito che la distribuzione delle covariate sia la medesima per ogni valore del trattamento. Si è visto che uno dei metodi più efficaci per arginare è la limitazione del dominio del trattamento (almeno nel caso di questa analisi, le altre possibili soluzioni presentate in 2.3.4 non hanno prodotto miglioramenti). Anche questa soluzione porta però a notevoli compromessi: nel caso del log-reddito mediano si è dovuto ridurre il dominio del trattamento tra il 10° e il 90° percentile, perdendo molte osservazioni dell'insieme di dati originale.

[1]Misura che quantifica la probabilità che due persone, scelte casualmente, appartengano a due gruppi etnici diversi. Un valore pari a zero indica che tutti nella popolazione appartengono allo stesso gruppo etnico, mentre un valore vicino a 100 indica che quasi tutti appartengono ad un gruppo etnico diverso (Census Bureau Data 2020).

Conclusioni

Lo studio condotto sulle elezioni statunitensi del 2020 ha permesso di valutare se e come il fattore economico ha influenzato il voto dei cittadini. Da una prima analisi l'associazione tra i trattamenti e il voto risultavano significative e positive, suggerendo un aumento della percentuale di voti ottenuta dal Partito Democratico alla crescita del reddito mediano o della disuguaglianza economica. La correzione dei confondenti tramite il propensity score generalizzato e la procedura di matching ha mostrato associazioni diverse: all'aumentare del reddito mediano la percentuale di voti diminuisce, mentre non vi è alcuna associazione significativa tra l'incide di Gini e i voti ottenuti dal Partito Democratico. I risultati mostrano l'efficacia della metodologia utilizzata: la generazione di una popolazione fittizia, nella quale le covariate hanno distribuzioni simili per ogni livello del trattamento, consente una valutazione del nesso causale tra le variabili.

Bibliografia

Akee, Randall, Maggie R Jones e Sonya R Porter (2019). «Race matters: Income shares, income inequality, and income mobility for all US races». In: *Demography* 56.3, pp. 999–1021.

Amlani, Sharif e Carlos Algara (2021). «Partisanship & nationalization in American elections: Evidence from presidential, senatorial, & gubernatorial elections in the US counties, 1872–2020». In: *Electoral Studies* 73, p. 102387.

Austin, Peter C (2011). «An introduction to propensity score methods for reducing the effects of confounding in observational studies». In: *Multivariate behavioral research* 46.3, pp. 399–424.

Bakka, Haakon et al. (2018). «Spatial modeling with R-INLA: A review». In: *Wiley Interdisciplinary Reviews: Computational Statistics* 10.6, e1443.

Besag, Julian (1974). «Spatial interaction and the statistical analysis of lattice systems». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225.

Besag, Julian, Jeremy York e Annie Mollié (1991). «Bayesian image restoration, with two applications in spatial statistics». In: *Annals of the institute of statistical mathematics* 43, pp. 1–20.

Blanden, Jo, Matthias Doepke e Jan Stuhler (2023). «Educational inequality». In: *Handbook of the Economics of Education*. Vol. 6. Elsevier, pp. 405–497.

Burden, Barry C e Amber Wichowsky (2014). «Economic discontent as a mobilizer: unemployment and voter turnout». In: *The Journal of Politics* 76.4, pp. 887–898.

Cecchini, Michele e Peter Smith (2018). «Assessing the dose-response relationship between number of office-based visits and hospitalizations for patients with type II diabetes using generalized propensity score matching». In: *Plos one* 13.12, e0209197.

Census Bureau Data (2020). *Census Bureau Data*. `https://data.census.gov/`. Accessed: 10-01-2024.

Chetty, Raj et al. (2020). «Race and economic opportunity in the United States: An intergenerational perspective». In: *The Quarterly Journal of Economics* 135.2, pp. 711–783.

Chong, Dennis e Reuel Rogers (2005). «Racial solidarity and political participation». In: *Political Behavior* 27, pp. 347–374.

CNN (2020). *Presidential Results*. `https://edition.cnn.com/election/2020/results/president`. Accessed: 07-03-2024.

Cutts, David et al. (2014). «With a little help from my neighbours: A spatial analysis of the impact of local campaigns at the 2010 British general election». In: *Electoral studies* 34, pp. 216–231.

Deininger, Klaus e Lyn Squire (1996). «A new data set measuring income inequality». In: *The World Bank Economic Review* 10.3, pp. 565–591.

Doyle, William R (2011). «Effect of increased academic momentum on transfer rates: An application of the generalized propensity score». In: *Economics of Education Review* 30.1, pp. 191–200.

Erikson, Robert S (2015). «Income inequality and policy responsiveness». In: *Annual Review of Political Science* 18, pp. 11–29.

Essletzbichler, Jürgen et al. (2021). «Spatial variation in populist right voting in Austria, 2013–2017». In: *Political Geography* 90, p. 102461.

F. Noel Perry, Colleen Kredell, Marcia E. Perry, Stephanie Leonard (2019). *California Migration*. `https://www.next10.org/sites/default/files/2019-06/California-Migration-Final2.pdf`. Accessed: 10-03-2024.

Fraga, Bernard L (2018). *The turnout gap: Race, ethnicity, and political inequality in a diversifying America*. Cambridge University Press.

Galbraith, James e Jaehee Choi (2020). «The consequences of economic inequality for presidential elections in the United States». In: *Structural Change and Economic Dynamics* 53, pp. 86–98.

Galbraith, James K e J Travis Hale (2008). «State income inequality and presidential election turnout and outcomes». In: *Social Science Quarterly* 89.4, pp. 887–901.

Gelman, Andrew et al. (2005). «Rich state, poor state, red state, blue state: What's the matter with Connecticut?» In: *Poor State, Red State, Blue State: What's the Matter with Connecticut*.

Gilens, Martin e Benjamin I Page (2014). «Testing theories of American politics: Elites, interest groups, and average citizens». In: *Perspectives on politics* 12.3, pp. 564–581.

Gini, Corrado (1921). «Measurement of inequality of incomes». In: *The economic journal* 31.121, pp. 124–125.

Hacker, Jacob S e Paul Pierson (2010). «Winner-take-all politics: Public policy, political organization, and the precipitous rise of top incomes in the United States». In: *Politics & Society* 38.2, pp. 152–204.

Harder, Valerie S, Elizabeth A Stuart e James C Anthony (2010). «Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research.» In: *Psychological methods* 15.3, p. 234.

Helgason, Agnar Freyr e Vittorio Mérola (2017). «Employment insecurity, incumbent partisanship, and voting behavior in comparative perspective». In: *Comparative Political Studies* 50.11, pp. 1489–1523.

Hersh, Eitan D e Clayton Nall (2016). «The primacy of race in the geography of income-based voting: New evidence from public voting records». In: *American Journal of Political Science* 60.2, pp. 289–303.

Hirano, Keisuke e Guido W Imbens (2004). «The propensity score with continuous treatments». In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164, pp. 73–84.

Jerrim, John e Lindsey Macmillan (2015). «Income inequality, intergenerational mobility, and the Great Gatsby Curve: Is education the key?» In: *Social Forces* 94.2, pp. 505–533.

Kennedy, Edward H et al. (2017). «Non-parametric methods for doubly robust estimation of continuous treatment effects». In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.4, pp. 1229–1245.

Khoshnevis, Naeem, Xiao Wu e Danielle Braun (2023). «CausalGPS: An R Package for Causal Inference With Continuous Exposures». In: *arXiv preprint arXiv:2310.00561*.

Kim, Jeongdai, Euel Elliott e Ding-Ming Wang (2003). «A spatial analysis of county-level outcomes in US Presidential elections: 1988–2000». In: *Electoral studies* 22.4, pp. 741–761.

Kim, Wooyoung et al. (2018). «The identification power of smoothness assumptions in models with counterfactual outcomes». In: *Quantitative Economics* 9.2, pp. 617–642.

Lake, James e Jun Nie (2023). «The 2020 US Presidential election and Trump's wars on trade and health insurance». In: *European Journal of Political Economy* 78, p. 102338.

Lei, Yalin et al. (2021). «Sedentary behavior is associated with chronic obstructive pulmonary disease: A generalized propensity score-weighted analysis». In: *Medicine* 100.18.

Leighley, Jan E e Arnold Vedlitz (1999). «Race, ethnicity, and political participation: Competing models and contrasting explanations». In: *The Journal of Politics* 61.4, pp. 1092–1114.

Li, Yajuan et al. (2019). «Measuring the effects of advertising on green industry sales: A generalized propensity score approach». In: *Applied Economics* 51.12, pp. 1303–1318.

Lysek, Jakub, Jiří Pánek e Tomáš Lebeda (2021). «Who are the voters and where are they? Using spatial statistics to analyse voting patterns in the parliamentary elections of the Czech Republic». In: *Journal of Maps* 17.1, pp. 33–38.

Mayer, Susan E (2010). «The relationship between income inequality and inequality in schooling». In: *Theory and research in Education* 8.1, pp. 5–20.

MIT (2020). *Election Data Lab*. `https://electionlab.mit.edu/data`. Accessed: 15-11-2023.

Morales-Otero, Mabel e Vicente Núñez-Antón (2021). «Comparing Bayesian spatial conditional overdispersion and the Besag–York–Mollié models: application to infant mortality rates». In: *Mathematics* 9.3, p. 282.

Nie, Norman H, Jane Junn e Kenneth Stehlik-Barry (1996). *Education and democratic citizenship in America*. University of Chicago Press.

O'Loughlin, John, Colin Flint e Luc Anselin (1994). «The geography of the Nazi vote: Context, confession, and class in the Reichstag election of 1930». In: *Annals of the association of American geographers* 84.3, pp. 351–380.

Rosenbaum, Paul R e Donald B Rubin (1983). «The central role of the propensity score in observational studies for causal effects». In: *Biometrika* 70.1, pp. 41–55.

Rubin, Donald B (1974). «Estimating causal effects of treatments in randomized and nonrandomized studies.» In: *Journal of educational Psychology* 66.5, p. 688.

Schündeln, Michael M et al. (2021). «Statistical methods for spatial cluster detection in childhood cancer incidence: A simulation study». In: *Cancer epidemiology* 70, p. 101873.

Sondheimer, Rachel Milstein e Donald P Green (2010). «Using experiments to estimate the effects of education on voter turnout». In: *American Journal of Political Science* 54.1, pp. 174–189.

Stewart Fotheringham, A, Ziqi Li e Levi John Wolf (2021). «Scale, context, and heterogeneity: A spatial analytical perspective on the 2016 US presidential election». In: *Annals of the American Association of Geographers* 111.6, pp. 1602–1621.

Vo, Tiffanie, Cyrus Schleifer e Peyman Hekmatpour (2023). «Asian Americans and Income Inequality: Disparities Between and Within Racial, Ethnic, and Gender Groups». In: *Sociological Perspectives* 66.6, pp. 1103–1124.

Winters, Jeffrey A e Benjamin I Page (2009). «Oligarchy in the United States?» In: *Perspectives on politics* 7.4, pp. 731–751.

Wright, John R (2012). «Unemployment and the democratic electoral advantage». In: *American Political Science Review* 106.4, pp. 685–702.

Wu, Xiao et al. (2022). «Matching on generalized propensity scores with continuous exposures». In: *Journal of the American Statistical Association*, pp. 1–29.

Zhu, Yeying, Donna L Coffman e Debashis Ghosh (2015). «A boosting algorithm for estimating generalized propensity scores with continuous treatments». In: *Journal of causal inference* 3.1, pp. 25–40.

Appendice A

Distribuzione spaziale dei confondenti

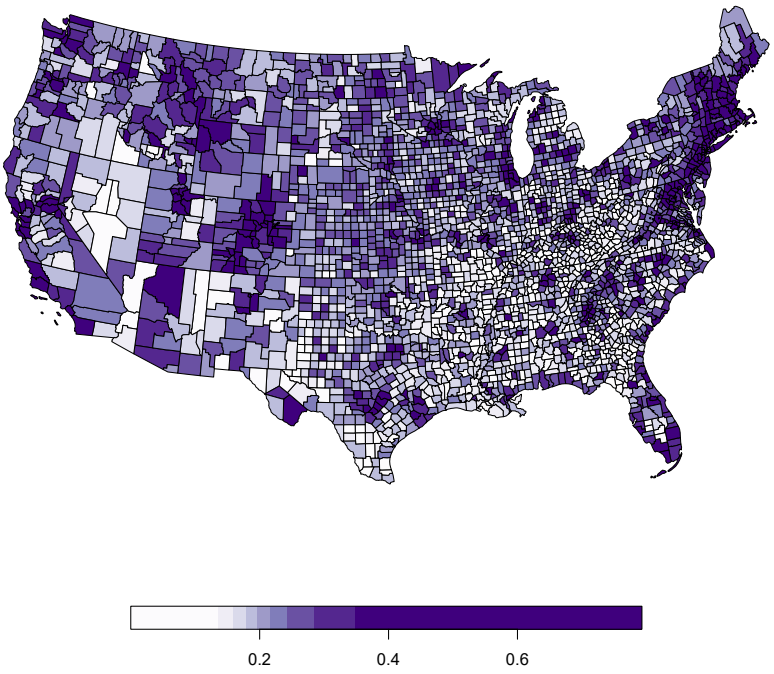
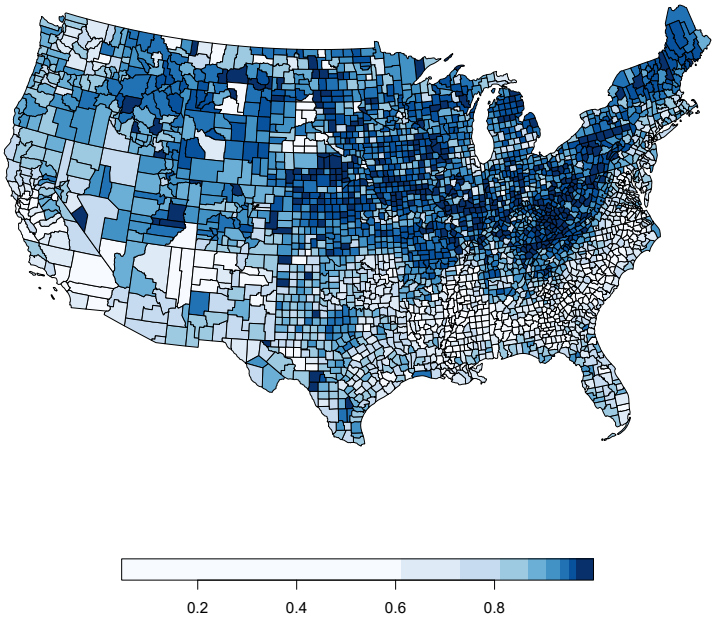
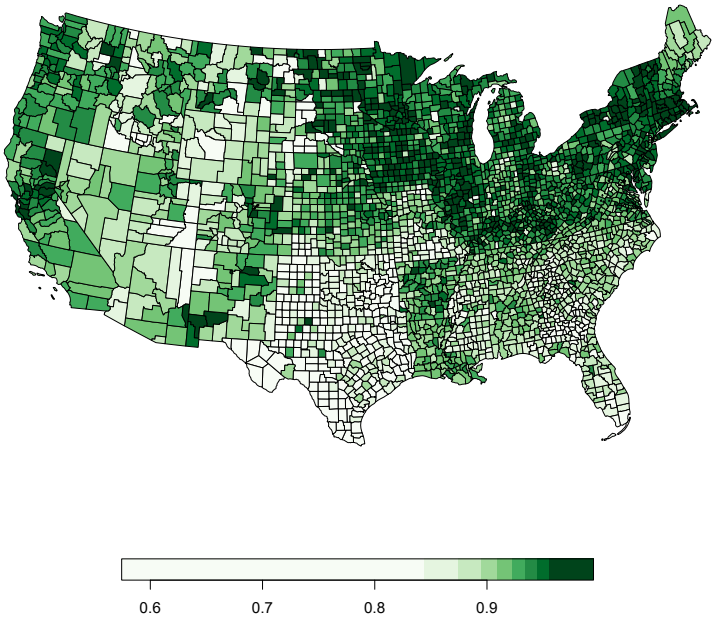


Figura A.1: Distribuzione spaziale della percentuale di persone laureate per contea.



(a) Distribuzione spaziale della percentuale di persone bianche per contea.



(b) Distribuzione spaziale della percentuale di persone che possiede un'assicurazione sanitaria.

Figura A.2: Distribuzione spaziale della percentuale di persone bianche (a) e della percentuale di persone con un'assicurazione sanitaria (b).

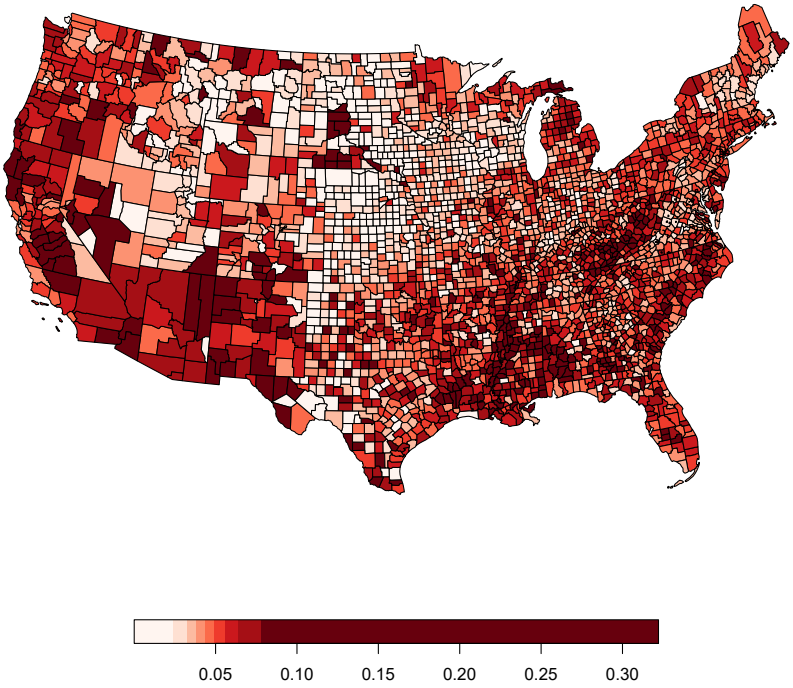


Figura A.3: Distribuzione spaziale della percentuale di persone disoccupate per contea.