Università degli Studi di Milano - Bicocca
**Dipartimento di Economia e Statistica**
**Corso di Laurea Triennale in Scienze Statistiche ed Economiche**


# Regressione beta e i suoi sviluppi

**Relatore:** Prof.ssa Sonia Migliorati

**Relazione della prova finale di:**
Giacomo Morzenti
Matricola 837395

**Anno Accademico 2020-2021**

# Indice

# 1 Introduzione

In questo elaborato è stato approfondito l'argomento d'interesse della regressione **beta**, il quale è un modello che permette di analizzare variabili risposta che si distribuiscono su un intervallo limitato. In particolare, dal momento che la variabile dipendente è una percentuale o probabilità, si prestano bene all'analisi tutti i fenomeni in cui bisogna modellizzare una proporzione come tassi, guarigioni o risultati elettorali. L'applicazione proposta in questo elaborato riguarda i risultati elettorali delle elezioni presidenziali americane che si sono svolte nel 2020, in cui i due candidati a sfidarsi erano Biden e Trump, rispettivamente per il partito democratico e quello repubblicano. La variabile risposta è stata considerata come i voti in percentuale presi da Biden per ogni contea, suddivisione di Stato. Per le covariate su cui sono stati applicati i modelli di regressione sono state scelte variabili che descrivessero caratteristiche della popolazione a livello di ricchezza, educazione ed etnia. L'obiettivo dello studio è trovare in che modo e in che quantità le variabili demografiche vanno a influire sui risultati elettorali.

L'elaborato si sviluppa nella seguente maniera: i primi tre capitoli sono dedicati alla parte teorica, in cui sono presentati in ordine il modello di regressione beta, il modello di regressione flexible beta e gli effetti casuali. Questa parte teorica è seguita da una parte descrittiva riguardante il dataset preso in analisi. Infine vi è un capitolo contenente le applicazione con i relativi risultati, seguito a sua volta dalle conclusioni. Tutti i modelli sono stati stimati attraverso l'utilizzo del programma **R**. I dati e lo script sono consultabili all'indirizzo: https://github.com/GiacomoMorzenti/Beta-regression- .

# 2 Beta Regression

## 2.1 Introduzione

I modelli di regressione sono utilizzati per analizzare e spiegare un dato di interesse in funzione di una o più variabili; uno dei più comuni è il modello di regressione lineare. Tuttavia quest'ultimo non è adatto alle situazioni in cui la variabile risposta è ristretta all'intervallo (0,1), in quanto potrebbe produrre valori al di fuori di questo range.
Una possibile soluzione è quella di trasformare la variabile dipendente così che assuma valori su tutto $\mathbb{R}$, e quindi modellare la media della variabile trasformata come predittore lineare costruito su un gruppo di variabili esogene. Questo approccio, tuttavia, presenta due problematiche:

- i parametri del modello non sono più facilmente interpretabili a causa della trasformazione applicata alla variabile risposta originale.
- tipicamente le distribuzioni di proporzioni presentano asimmetria, quindi fare inferenza basata sull'assunzione di normalità è scorretto.

Il modello di regressione beta, proposto da **Ferrari e Cribari-Neto (2004)**[1], è fatto su misura per situazioni dove la variabile dipendente (y) è misurata sull'intervallo unitario standard, ed è basato sull'assunzione che la variabile risposta sia distribuita come una variabile casuale **beta**.

## 2.2 La distribuzione beta

La distribuzione beta è molto flessibile per modellare proporzioni dal momento che la sua densità può avere forme diverse in base ai due valori dei parametri che ne caratterizzano la distribuzione. La funzione di densità della beta è data da:

$$\pi(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \qquad 0 < y < 1 \tag{2.1}$$

dove p > 0, q > 0 e $\Gamma(\cdot)$ è la funzione Gamma. La media e la varianza di Y sono rispettivamente:

$$E(Y) = \frac{p}{p+q} \qquad , \qquad var(Y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{2.2}$$

La distribuzione uniforme è un caso particolare della beta (2.1), ossia quando $p = q = 1$. Dal momento che la variabile casuale beta non è riconducibile alla famiglia **DE1** (dispersione esponenziale di ordine 1), non possono essere utilizzate le parametrizzazioni canoniche dei "Generalized linear model" (**GLM**). Per questo viene proposta una particolare parametrizzazione in cui siano presenti il parametro della media $\mu$ e il parametro di dispersione $\phi$. Ponendo $\mu = \frac{p}{p+q}$ e $\phi = p+q$ si ottengono i seguenti risultati per la media e per la varianza:

$$E(Y) = \mu, \qquad var(Y) = \frac{V(\mu)}{1+\phi} \tag{2.3}$$

Dove $V(\mu) = \mu(1-\mu)$, così che $\mu$ sia la media della variabile risposta e $\phi$ possa essere interpretato come un parametro di precisione, nel senso che, per un valore fissato di $\mu$, più grande è $\phi$, minore sarà la varianza. La funzione di densità con la nuova parametrizzazione è la seguente:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(1-\mu)\phi} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \;\; 0 < y < 1 \tag{2.4}$$

Dove $0 < \mu < 1$ e $\phi > 0$. È noto che al variare dei valori dei due parametri la funzione di densità (2.4) può avere forme diverse (1). In particolare è simmetrica quando $\mu = 0.5$. Inoltre la dispersione della distribuzione, per valori di $\mu$ fissati, diminuisce al crescere di $\phi$, il quale appunto è il parametro che rappresenta la precisione.

Il modello di regressione che viene proposto assume che la risposta sia ristretta all'intervallo $(0, 1)$, tuttavia attraverso una semplice trasformazione può essere generalizzato ai casi in cui la risposta è ristretta ad un intervallo $(a, b)$, dove $a$ e $b$ sono due scalari noti, con $a < b$.

## 2.3 Specificazione del modello

Siano $y_1, \ldots, y_n$ variabili casuali indipendenti, dove ogni $y_i$ , con $i = 1, \ldots, n$ , segue la distribuzione di densità della beta con media $\mu_i$ e precisione $\phi_i$. La media di $y_i$ è legata al predittore lineare, e quindi alle covariate, attraverso una funzione link. Quindi possiamo scrivere:

$$g_1(\mu) = \eta_i = x_i^T \beta \tag{2.5}$$

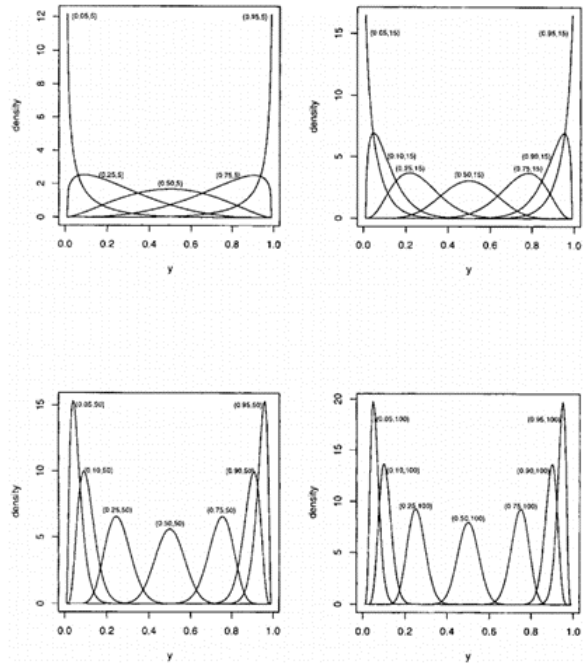



Figura 1: Le forme differenti che assume la distribuzione beta al variare del parametro della media $\mu$ e del parametro di precisione $\phi$

dove $\eta_i$ è il predittore lineare, formato rispettivamente da $x_i$ , il vettore delle covariate, e da $\beta$, il vettore dei parametri; $g_1(\cdot)$ rappresenta la funzione link. Questa funzione link è strettamente monotona e differenziabile due volte, e permette di mappare l'intervallo $(0, 1)$ di $\mu_i$ sulla linea dei valori reali. Le funzioni link più comuni sono il **logit** $g(\mu) = log(\frac{\mu}{1-\mu})$, il **probit** $g(\mu) = \Phi^{-1}(\mu)$, dove $\Phi(\cdot)$ è la funzione di ripartizione della normale standard, il **log-log link** $g(\mu) = -log(-log(\mu))$, il **clog-log link** $g(\mu) = log(-log(1-\mu))$ oppure qualunque inversa di una funzione di distribuzione cumulativa. Una funzione particolarmente utile a livello interpretativo è il logit, il quale è così espresso:

$$\mu_t = \frac{e^{x_t^T \beta}}{1 + e^{x_t^T \beta}} \tag{2.6}$$

dove $x_t^T = (x_{t1}, \ldots, x_{tk}), t = 1, \ldots, n$. Si supponga che il valore dell'i-esimo regressore aumenti di c unità e che tutte le altre variabili indipendenti rimangano immutate. Si definisca $\mu'$ la media di $y$ sotto il valore della nuova covariata, dove $\mu$ denota la media di $y$ sotto il

valore originale della covariata. Allora è possibile mostrare che:

$$e^{c\beta_i} = \frac{\mu'/(1-\mu')}{\mu/(1-\mu)} \tag{2.7}$$

$exp(c\beta_i)$ è pari all'**odds ratio**.

Il parametro di precisione $\phi_i$ può essere tenuto costante (Ferrari and Cribari-Neto, 2004)[1] oppure a sua volta essere legato ad un altro predittore lineare (Cepeda-Cuervo, 2001; Cepeda and Gamerman, 2005; Smithson and Verkuilen, 2006):

$$g_2(\phi_i) = \zeta_i = z_i^T \gamma \tag{2.8}$$

dove $z_i$ è il vettore delle covariate (non per forza uguale a $x_i$ (2.5)) e $\gamma$ sono i parametri della regressione. L'unica differenza dalla formula della media sta in $g_2(\cdot)$ la quale è una funzione link diversa da $g_1$(2.5) in quanto questa volta l'insieme di partenza da mappare sui reali è $(0, +\infty)$, cioè il supporto di $\phi$. Una possibile funzione link è il logaritmo.

## 2.4 Stima dei coefficienti

L'inferenza sui parametri di questo modello, $\beta$ per la (2.5) e $\gamma$ nel caso venga modellato anche il parametro di precisione (2.8), è svolta attraverso il metodo di **massima verosimiglianza** (ML). La funzione di verosimiglianza è così calcolata:

$$\begin{aligned} l_t(\mu_t, \phi) = log\Gamma(\phi) - log\Gamma(\mu_t\phi) - log\Gamma((1-\mu_t)\phi) \\ + (\mu_t\phi - 1)logy_t + (1-\mu_t)\phi - 1)log(1-y_t) \end{aligned} \tag{2.9}$$

dove $\mu_t$ è calcolato attraverso il predittore lineare. Siano $y_t^* = log(y_t/(1-y_t)$ e $\mu_t^* = \Psi(\mu_t\phi) - \Psi((1-\mu_t)\phi)$. La funzione score, ottenuta differenziando la log-verosimiglianza rispetto ai parametri ignoti, è data da $(U_\beta(\beta,\phi)^T, U_\phi(\beta,\phi)^T)$, dove

$$U_\beta(\beta,\phi) = \phi X^T T(y^* - \mu^*)$$

con X matrice $nxk$ con la t-esima riga pari a $x_t^T$, $y^* = (y_1^*, \ldots, y_n^*)$, $T = diag(1/g'(\mu_1), \ldots, 1/g'(\mu_n))$ e $\mu^* = (\mu_1^*, \ldots, \mu_n^*)$ e

$$U_\phi(\beta,\phi) = \sum_{t=1}^{n} (\mu_t(y_t^* - \mu_t^* + log(1-y_t) - \Psi((1-\mu_t)\phi) + \Psi(\phi))$$

Sotto le condizioni di regolarità per la stima di massima verosimiglinza, quando la numerosità campionaria è sufficientemente grande, si ha

$$\begin{pmatrix}\hat{\beta}\\ \hat{\phi}\end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix}\beta\\ \phi\end{pmatrix}, K^{-1}\right)$$

dove $\hat{\beta}$ e $\hat{\phi}$ sono gli stimatori di massima verosimiglianza rispettivamente per $\beta$ e $\phi$, e $K^{-1}$ è l'inversa della matrice di informazione di Fisher.
Gli stimatori di massima verosimiglianza per $\beta$ e $\phi$ sono ottenuti dalle equazioni $U_\beta(\beta, \phi) = 0$ e $U_\phi(\beta, \phi) = 0$, le quali non hanno una forma chiusa. Quindi gli stimatori devono essere calcolati attraverso una massimizzazione numerica della funzione di log-verosimiglianza, utilizzando un algoritmo di ottimizzazione non lineare. L'algoritmo necessita la specificazione dei valori iniziali da inserire nello schema iterativo. Per il vettore dei $\beta$ possono essere usati gli stimatori dei minimi quadrati ottenuti dalla regressione sulla risposta trasformata $g(y_1), \ldots, g(y_n)$ su X, cioè $(X^TX)^{-1}X^Tz$ dove $z = (g(y_1), \ldots, g(y_n))^T$. Per $\phi$, il valore iniziale proposto è

$$\frac{1}{n}\sum_{t=1}^{n}\frac{\hat{\mu}_t(1-\hat{\mu}_t)}{\hat{\sigma}_t^2} - 1$$

dove $\hat{\mu}_t$ è ottenuto applicando $g^{-1}(\cdot)$ al t-esimo valore stimato dalla regressione lineare di $g(y_1), \ldots, g(y_n)$ su X e $\hat{\sigma}_t^2 = \hat{e}^T\hat{e}/((n-k)(g'(\hat{\mu}_t))^2)$ dove $\hat{e} = z - X(X^TX)^{-1}X^Tz$ è il vettore dei residui della regressione lineare applicata sulle risposte trasformate.

## 2.5 Diagnostica

Dopo aver stimato il modello è fondamentale svolgere analisi diagnostiche per controllare la bontà del modello e l'adattamento ai dati. Per iniziare, una misura della varianza spiegata dal modello può essere ottenuta attraverso lo pseudo $R^2(R_p^2)$, definito come il quadrato del coefficiente di correlazione tra $\hat{\eta}$ e $g(y)$. Si noti che $0 \leq R_p^2 \leq 1$, dove $R_p^2 = 1$ rappresenta perfetta corrispondenza tra $\hat{\eta}$ e $g(y)$ e quindi tra $\hat{\mu}$ e $y$. All'interno della letteratura [1][4] sono proposti diversi residui, tra cui:

- **Residui grezzi** Definiti come $y_t - \hat{\mu}_t$. Nel caso del modello beta questi residui sono inadatti in quanto non tengono conto della sua natura eteroschedastica.

- **Residui standardizzati** Definiti come:

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{var}(y_t)}}$$

dove $\hat{var}(y_t) = \hat{\mu}_t(1 - \hat{\mu}_t)/(1 + \hat{\phi})$. Qui, $\hat{\mu}_t = g^{-1}(x_t^T\hat{\beta})$, $\hat{\beta}$ e $\hat{\phi}$ sono gli stimatori di massima verosimiglianza rispettivamente di $\beta$ e $\phi$.

- **Residui di devianza** La devianza è misurata come due volte la differenza tra il massimo della verosimiglianza ottimale (cioè quella del modello saturo) e quello ottenuto dal modello in analisi. Sia definita quindi $D(y; \hat{\mu}, \hat{\phi}) = \sum_{t=1}^{n} 2(l_t(\overline{\mu}_t, \hat{\phi}) - l_t(\hat{\mu}_t, \hat{\phi})$, dove $\hat{\phi}$ è la stima ML di $\phi$, $\hat{\mu}$ è la stima ML di $\mu_t$ riferita al modello in analisi e $\overline{\mu}$ è la stima di $\mu_t$ riferita al modello saturo. Si noti che $D(y, \hat{\mu}, \phi) = \sum_{t=1}^{n}(r_t^d)^2$, dove

$$r_t^d = sign(y_t - \hat{\mu}_t)\{2[l_t(\overline{\mu}_t.\hat{\phi} - l_t(\hat{\mu}_t, \hat{\phi}]\}^{1/2}.$$

Ogni osservazione t-esima contribuisce la quantità $(r_t^d)^2$ alla devianza. $r_t^d$ è detto residuo di devianza.

- **Residui di Pearson** Definiti come

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{var}(y_t)}}$$

dove $\hat{\mu}_t = g^{-1}(x_t^T\hat{\beta})$ e $\hat{var}(y_t) = \hat{\mu}_t(1 - \hat{\mu}_t)/(1 + \hat{\phi})$.

Un grafico di questi residui rispetto all'indice delle osservazioni (t) non dovrebbe mostrare andamenti particolari, quindi una dispersione uniforme dei punti. Inoltre se ci fosse un trend sistematico nel grafico dei residui rispetto al predittore lineare $\hat{\eta}_t$, sarebbe un segnale di un'errata specificazione della funzione link. Essendo ignota la distrubuzione dei residui, è preferibile l'utilizzo della tecnica "half-past plots". appronfondito in [5]. L'identificazione dei punti influenti viene svolta attraverso i "Generalized Leverage" [6] e la misura della distanza di Cook [7].

# 3 Flexible beta

## 3.1 Introduzione

Sebbene la distribuzione beta sia dotata di un'ottima flessibilità, questa non riesce a gestire i casi di code pesanti e di bimodalità. Per risolvere questi casi inizialmente si è pensato di utilizzare misture generiche di distribuzioni beta, in quanto le misture permettono di avere una maggiore accuratezza nell'adattamento ai dati e una maggiore robustezza. Tuttavia questi benefici vengono al costo di una minor trattabilità, in quanto si va incontro al problema di non identificabilità e di funzioni di verosimiglianza illimitate.
Come soluzioni per gestire questo trade-off tra flessibilità e trattabilità è stato proposto il modello flexible beta (FB)[2][3], il quale è una mistura speciale di due distribuzioni beta con media arbitraria e varianza comune. Questo modello permette di gestire situazioni con code pesanti e di multi-modalità, senza andare a perdere l'identificabilità e garantendo funzioni di verosimiglianze limitate.

## 3.2 La distribuzione flexible beta

La distribuzione flexible beta è definita come una speciale mistura di due distribuzioni beta $Y \sim pbeta(\alpha_1 + \tau, \alpha_2) + (1-p)beta(\alpha_1, \alpha_2 + \tau)$, la quale dipende da quattro parametri $(\alpha_1, \alpha_2, \tau, p)$ dove $0 < p < 1$ $\alpha_1 > 0, \alpha_2 > 0$ e $\tau > 0$. Se $Y$ è una flexible beta allora la sua funzione di densità di probabilità per $0 < y < 1$ è definita come:

$$f_{FB}(y; \alpha_1, \alpha_2, \tau, p) =$$

$$\frac{\Gamma(\alpha_1 + \alpha_2 + \tau)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1}\left[p\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + \tau)}y^T + (1-p)\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_2 + \tau)}(1-y)^T\right]$$

I primi due momenti della FB sono pari a:

$$E(Y) = \frac{\alpha_1 + \tau p}{\phi} \tag{3.1}$$

$$Var(Y) = \frac{E(Y)(1 - E(Y)) + \tau^2 p(1-p)/\phi}{\phi + 1} \tag{3.2}$$

dove $\phi = \alpha_1 + \alpha_2 + \tau$. La distribuzione FB è una mistura di due distribuzioni beta con parametro di precisione comune $\phi = \alpha_1 + \alpha_2 + \tau$ e medie distinte $\lambda_1 > \lambda_2$:

$$f^*_{FB}(y; \lambda_1.\lambda_2, \phi, p) = p f^*_B(y; \lambda_1, \phi) + (1-p) f^*_B(y; \lambda_2, \phi) \tag{3.3}$$

dove $f^*_B$ è la beta con parametri di media e precisione, e

$$\lambda_1 = \frac{\alpha_1 + \tau}{\phi} \qquad\qquad \lambda_2 = \frac{\alpha_1}{\phi} \tag{3.4}$$

Come è possibile notare da queste formule, la struttura della mistura speciale che definisce la FB è di semplice interpretazione. Inoltre la distribuzione FB è in grado di estendere la varietà di forme della beta principalmente in termini di **bimodalità, asimmetria e comportamento delle code**. Infatti la distribuzione beta è caratterizzata da limiti in zero per entrambi gli estremi (0 e 1), tranne nel caso in cui la distribuzione beta coincide con quella uniforme. Invece questo non è valido per la FB, la quale può avere uno o entrambi i limiti agli estremi con valori strettamente positivi.

Inoltre la struttura della FB la rende, sia dal lato teorico che da quello computazionale, molto trattabile; infatti questa rispetta la proprietà di **identificabilità** in senso forte, per cui due elementi della famiglia parametrica FB sono uguali se solo se hanno gli stessi parametri. Questa proprietà risolve il problema dell'invarianza per la permutazione delle componenti e quindi il problema dell'etichettamento.

Per definire un modello di regressione con risposta FB, è conveniente introdurre una differente parametrizzazione di questa che includa esplicitamente la media e altri parametri di facile interpretazione. Una possibile soluzione è la seguente:

$$\begin{cases} \mu = E(Y) = p\lambda_1 + (1-p)\lambda_2 \\ \phi = \phi \\ \hat{w} = \lambda_1 - \lambda_2 \\ p = p \end{cases}$$

dove $\mu$ è la media, $\hat{w}$ è una misura di distanza tra le due componenti della mistura, $p$ è la proporzione della mistura e $\phi$ ricopre il ruolo del parametro di precisione (al suo crescere, Var(Y) (3.2) diminuisce).

Analizzando lo spazio parametrico, notiamo che $\phi$ è libero di muoversi nell'insieme dei reali positivi, invece $\mu, p$ e $\hat{w}$ sono legati da vincoli. Ai fini di inferenza sui parametri, si è deciso di normalizzare $\hat{w}$

per farlo muovere nell'intervallo (0,1), sostituendolo con:

$$w = \frac{\hat{w}}{min\{\frac{\mu}{p}, \frac{1-\mu}{1-p}\}}$$

in questo modo i parametri $p, \mu$ e $w$ variano nell'intervallo $(0, 1)$ e $\phi > 0$.

## 3.3 Specificazione del modello

Si consideri un vettore di variabili indipendenti $Y_T = (Y_1, \ldots, Y_i, \ldots, Y_n)$ che assumono valori nell'intervallo $(0, 1)$. Si definisce il modello di regressione FB, assumendo che ogni $Y_i$ sia indipendentemente distribuita come una flexible beta $Y_i \sim FB(\mu_i, \phi, w, p)$, come segue

$$g(\mu_i) = x_i^T \beta \qquad i = 1, \ldots, n \tag{3.5}$$

dove $\mu_i$ è la media di $Y_i$, $X_i^T$ è il vettore delle covariate, $\beta^T$ è il vettore dei parametri di regressione e $g(\cdot)$ è una funzione link, strettamente monotona e differenziabile due volte. Vengono applicati gli stessi ragionamenti in riferimento alla funzione link del capitolo precedente.
Il modello così descritto si concentra solo sul modellare il parametro della media; tuttavia, essendo la varianza in funzione della media (3.2), pure questa varierà al variare delle covariate, portando ad eteroschedasticità. Questa forma di eteroschedasticità può essere definita naturale, in quanto tiene in considerazione il fatto che il massimo della varianza di una variabile casuale con supporto in $(0, 1)$ dipenda dalla sua media $\mu$, sia pari a $\mu(1-\mu)$. Tuttavia in certi casi potrebbe essere preferibile modellare la varianza in funzione di alcune covariate. Questo è possibile nella regressione FB in quanto $\mu$ e $\phi$ non sono vincolati tra di loro. La regressione per il parametro di dispersione può essere definita come $h(\phi_i) = z_i^T \delta$, dove $h(\cdot)$ è una funzione link appropriata, $z_i^T$ è il vettore delle covariate e $\delta^T$ è il vettore dei parametri di regressione.

Ci possono essere situazioni in cui si vogliono far dipendere i parametri dei pesi p e la distanza tra i gruppi w dalle covariate, e questo è possibile nella stessa maniera proposta precedentemente.

Come conseguenza della identificabilità della distribuzione FB, il modello FBR è identificabile sotto assunzione che la matrice del disegno ha rango pieno. Inoltre, in condizioni generali, la funzione di verosimiglianza del modello FBR è superiormente limitata.

La funzione di verosimiglianza per il modello FBR calcolata su un campione di n osservazioni indipendenti $y^T = (y_1, \ldots, y_i, \ldots, y_n)$ è

uguale a:

$$L(\eta|y) = \prod_{i=1}^{n} f^*_{FB}(y_i|\mu_i, \phi, w, p) \tag{3.6}$$

dove $\eta = (\beta, \phi, w, p), \mu_i = g^{-1}(x_i^T\beta)$ e $f^*_{FB}(y|\mu, \phi, w, p)$ che è data da (3.3) con $\alpha_1$ e $\alpha_2$ presi dalla (3.4).

Dal momento che l'allocazione di ogni i-esima osservazione ad una delle due componenti della mistura è ignota, non vi è una soluzione chiusa al problema di stima. Quindi è necessaria una soluzione numerica che viene raggiunta passando a stime bayesiane, approfondite in [2].

# 4 Effetti casuali

## 4.1 Introduzione

I modelli ad **effetti casuali**, o più comunemente chiamati modelli a multilivello, sono regressioni lineari o lineari generalizzate in cui i parametri (coefficienti di regressione) sono a loro volta costituiti da un modello di probabilità. Questo secondo modello ha dei parametri propri che vengono stimati a loro volta attraverso i dati.

Le due componenti chiave del modello a multilivello sono i coefficienti variabili, e un modello per questi coefficienti. La caratteristica che contraddistingue i modelli a multilivello dalle regressioni classiche è la possibilità di modellare la varianza tra i gruppi. Infatti l'utilizzo principale dei modelli ad effetti casuali è svolto in presenza di **dati raggruppati**, come nell'applicazione proposta più avanti in cui le contee (americane) sono raggruppate per Stato.

Esistono 3 possibili modelli multilivello: intercetta variabile, coefficienti variabili o la combinazione di questi due (sia intercetta che coefficienti variabili).

## 4.2 Esempio esplicativo con modelli lineari

A fine esplicativo introduco l'esempio preso da [8] di uno studio effettuato sugli studenti di diverse scuole, in cui si cerca di prevedere il voto $y$ conseguito a una prova standard attraverso una regressione sui risultati di prove precedenti $x$ e ulteriori informazioni. In questo esempio i due livelli fanno riferimento agli studenti e alle scuole. Il modello multilivello permette di fare variare i parametri della regressione a livello degli studenti in base alla scuola di appartenenza. Assumiamo una singola covariata al livello studente $x$ (risultato ad una prova precedente) e una al livello scuola $u$ (reddito medio dei genitori). Il modello ad intercetta variabile è quello in cui le regressioni hanno i coefficienti dei parametri uguali per ognuna delle scuole, e solo l'intercetta varia. Usando la notazione i per identificare gli studenti e j[i] per la scuola j che contiene lo studente i abbiamo i due modelli rispettivamente

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \quad \text{per gli studenti} \quad i = 1, \ldots, n \tag{4.1}$$

$$\alpha_j = a + bu_j + \eta_j \quad \text{per le scuole} \quad j = 1, \ldots, J \tag{4.2}$$

dove $x_i$ e $u_j$ sono i predittori rispettivamente a livello degli studenti e delle scuole, e $\epsilon_i$ e $\eta_j$ sono gli errori indipendenti per i due livelli. Più complicato è il modello in cui sia l'intercetta che i parametri relativi

alle covariate (qui $\beta$) variano per ogni scuola, per cui si inserirebbe un ulteriore modello per la regressione di $\beta_j$. In quest'ultimo caso le covariate influirebbero in maniera differente in base al cambiamento dei parametri associati rispettivamente ad ogni scuola.

## 4.3 Effetti casuali nella beta Regression

Senza perdita di generalità [9], consideriamo il caso in cui $i = 1 \ldots, n_j$ osservazioni sono raggruppate in $j = 1 \ldots, N$ gruppi. Definiamo $b_j$ il vettore di effetti casuali per il gruppo j. Possiamo inserire gli effetti casuali al modello di regressione beta tale che

$$g(\mu_{ij}) = x_{ij}^T \beta + z_{ij}^T b_j \qquad \text{con} \quad b_j \sim N(0, G)$$

dove $z_{ij}^T$ è un vettore di covariate e G è la matrice definita positiva delle covarianze degli effetti casuali. L'assunzione di normalità degli effetti casuali è la più comune e conveniente, ma si possono assumere anche distribuzioni diverse. Per un modello con solo intercetta variabile, $b_j$ sarà uno scalare, invece per i modelli con intercetta e anche coefficienti variabili $b_j$ sarà un vettore. Nel primo caso $z_{ij}$ sarà pari a 1, nel secondo caso sarà un vettore con sempre 1 al primo posto (per l'intercetta) e poi i coefficienti che si vogliono far variare.

I parametri del modello sono stimati attraverso la massimizzazione della verosimiglianza marginale che è ottenuta integrando gli effetti inosservati bj dalla funzione di verosimiglianza.

# 5 Analisi esplorativa

## 5.1 Introduzione

Il dataset su cui verranno applicati i modelli visti nei capitoli precedenti riguarda le elezioni presidenziali americane del 2020, in cui Joe Biden rappresenta il Partito Democratico e Donald Trump quello Repubblicano. Il dataset creato ad hoc è composto dalla variabile risposta, che rappresenta la percentuale dei voti di Biden[1], e dalle covariate, che rappresentano le caratteristiche della popolazione[2]. Tutti i dati sono raccolti a livello delle contee, cioè le suddivisioni al di sotto degli Stati negli USA.

## 5.2 Descrizione e analisi delle variabili

Il dataset in analisi è composto da 3089 osservazioni (contee) e 18 variabili. Le variabili sono state raccolte da dataset diversi, per cui vi è stata la necessità di creare una variabile chiamata *chiave* per identificare univocamente le contee. Questa è composta da due parti: la prima costituita dall'abbreviazione dello Stato e la seconda dal nome della contea (es. CA.Los Angeles). Infatti è molto comune negli US utilizzare gli stessi nomi per contee all'interno di Stati diversi [3].
Su tutte le covariate numeriche si è deciso di applicare la standardizzazione

$$x_{stand} = \frac{x - \mu_x}{\sigma_x}$$

dove $\mu_x$ e $\sigma_x$ sono rispettivamente la media e la deviazione standard della variabile x. La standardizzazione è stata necessaria per permettere l'utilizzo di librerie pesanti (rSTAN) con stime bayesiane che verranno utilizzate nella parte di stima dei modelli con effetti casuali. Inoltre la standardizzazione delle variabili permette anche di confrontare meglio l'influenza delle covariate sulla variabile risposta, a parità di deviazione standard. L'interpretazione dei coefficienti verrà approfondita nella fase applicativa. Le variabili sono così descritte:

- **STATE.x** Rappresenta lo Stato di appartenza della contea; è stato necessario tenere questa variabile in quanto sarà utilizzata per il clustering. Questa è una variabile fattoriale con 49 valori,

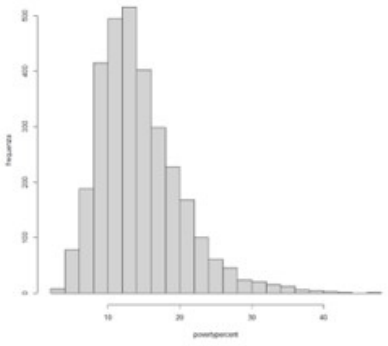
[1] dati raccolti da *https* : *//en.wikipedia.org/wiki/*2020*_United_States_presidential_election#*

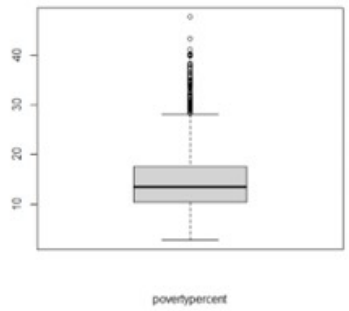
[2] dati raccolti dal sito del governo americano, https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

[3] Ad esempio esistono ben 32 contee sono chiamate "Washington"

infatti è stato scartato lo stato dell'Alaska in quanto l'unico a non avere una suddivisione in contee.

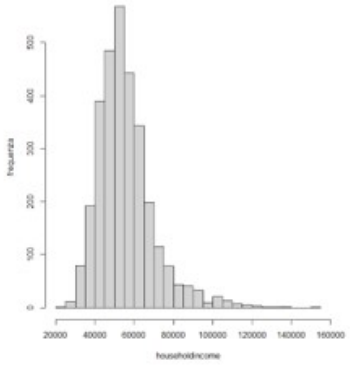
- **poverty_percent** La percentuale della popolazione che risiede nella fascia della povertà. Come possiamo notare dalla distribuzione, nonostante gli USA siano un paese generalmente ricco, vi è una media della percentuale della popolazione considerata povera intorno al 14%, dove il massimo in alcune contee raggiunge quasi il 50%.

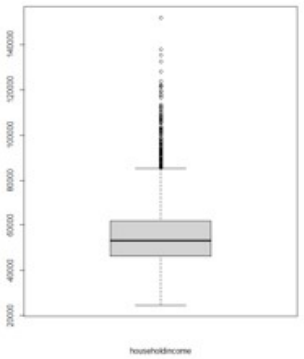



| Min | Mean | Max | S.D. |
|---|---|---|---|
| 2.70 | 14.47 | 47.70 | 5.79 |

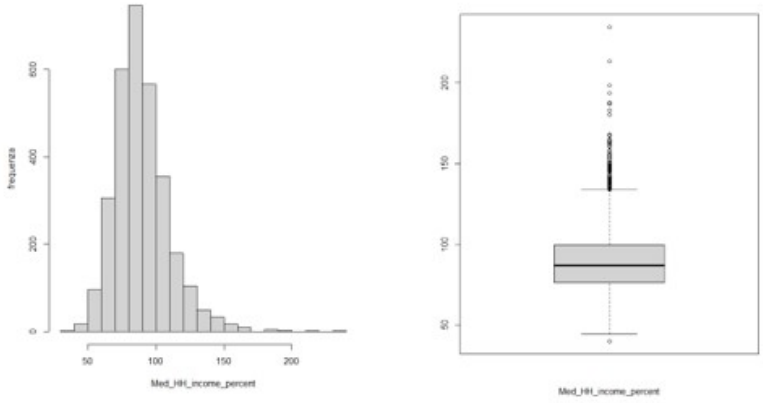
- **household_income** La mediana della distribuzione del reddito famigliare della contea in dollari. Come possiamo notare dai grafici, vi sono numerosi outlier dalla parte destra della distribuzione, che ci fanno notare la presenza di numerose contee con una mediana del reddito famigliare molto elevata.



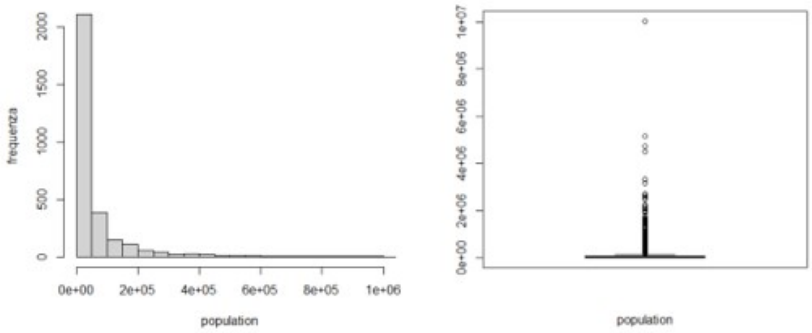
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 24 732 | 55 577 | 151 806 | 14 421 |

- **Med_HH_income_percent** La percentuale della mediana del reddito famigliare rispetto alla mediana del reddito famigliare dello Stato in cui è situata la contea. Questa variabile ci permette di capire la situazione economica della contea in relazione a quella dello Stato di appartenenza.



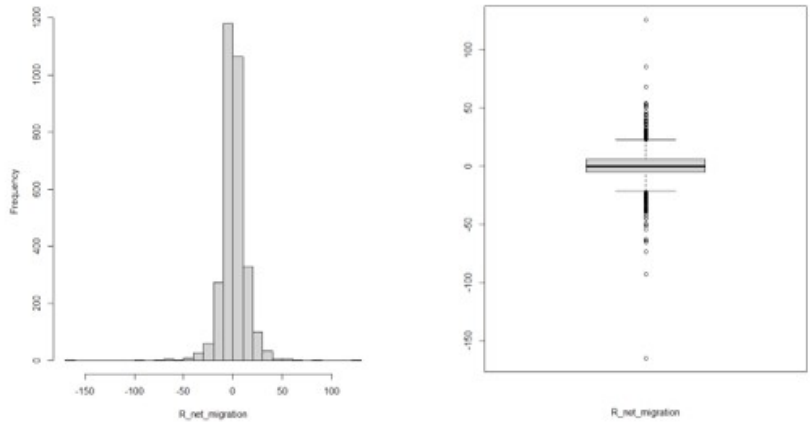

| Min | Mean | Max | S.D. |
|---|---|---|---|
| 39.92 | 89.49 | 234.52 | 19.86 |

- **population** La popolazione residente di ogni contea. Nella maggior parte delle contee la popolazione non supera i diecimila abitanti, tuttavia vi sono diversi outlier di contee con popolazioni sopra il milione, tra cui Los Angeles in California che supera addirittura i dieci milioni di abitanti.



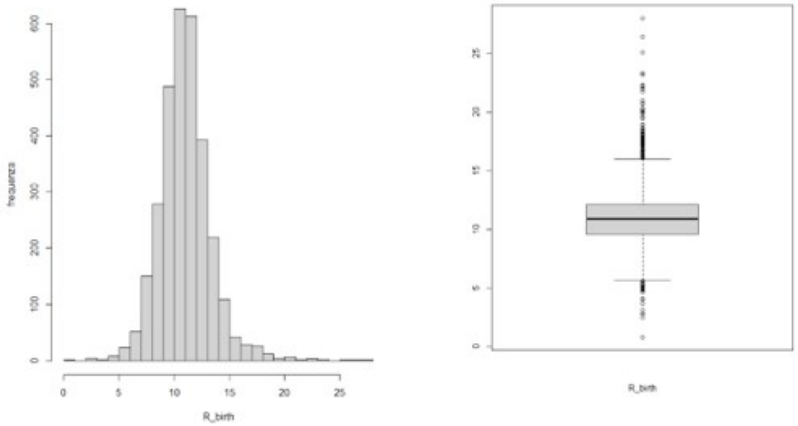
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 169 | 105 017 | 10 039 107 | 335 785 |

- **R_net_migration** Il tasso di immigrati (positivo) o emigranti (negativo) rispetto alla popolazione della contea.



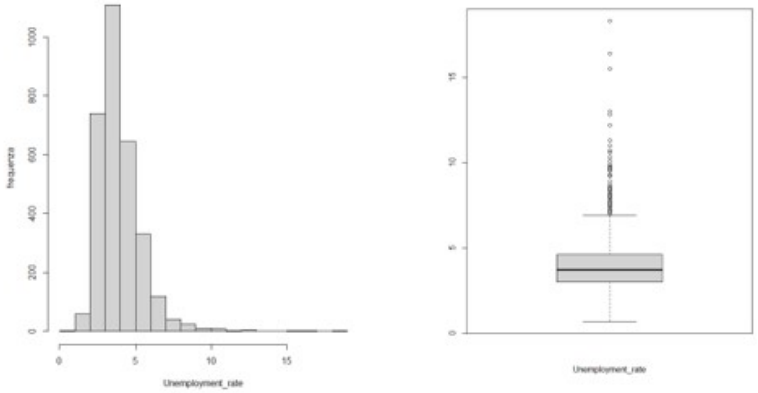

| Min | Mean | Max | S.D. |
|---|---|---|---|
| -165.38 | 0.39 | 126.18 | 12.33 |

- **R_birth** Il tasso di natalità.



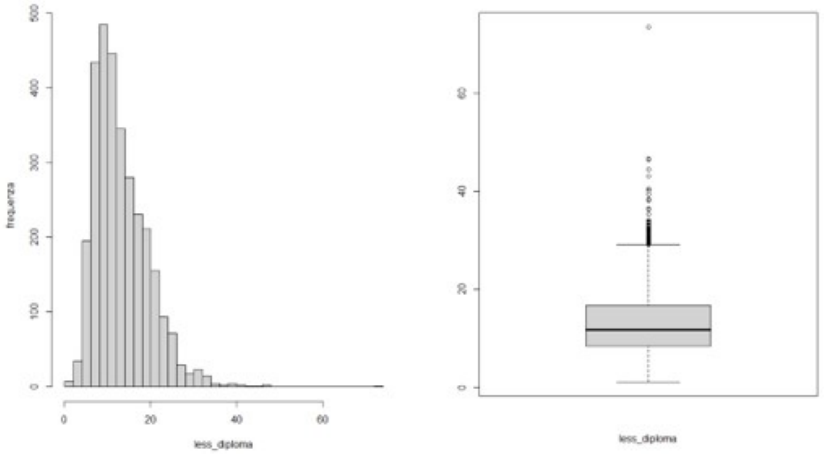

| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.79 | 10.97 | 27.99 | 2.33 |

- **unemployment_rate** Il tasso di disoccupazione. Si noti che il tasso di disoccupazione medio per contea negli USA è minore della metà di quello italiano.




| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.70 | 3.96 | 18.3 | 1.39 |

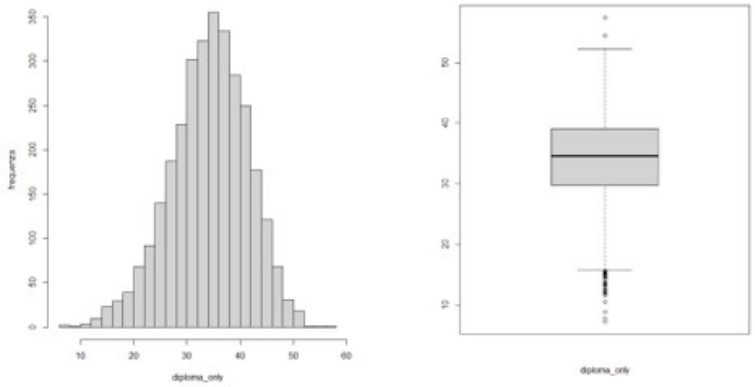
- **less_diploma** La percentuale di popolazione che ha un livello di educazione inferiore al diploma superiore. Il livello massimo di questa variabile è riscontrato nella contea Kenedy in Texas, la quale ha quasi tre quarti della popolazione senza nemmeno un diploma.[4]




[4]Tuttavia bisogna considerare anche che questa contea è la terza dal basso per popolazione, infatti conta solo poco più di quattrocento abitanti

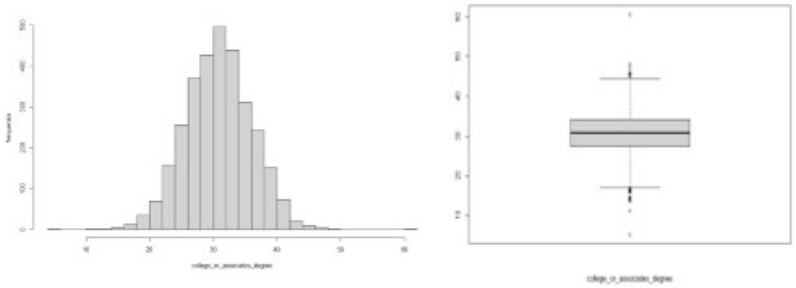
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 1.12 | 13.08 | 73.56 | 6.26 |

- **diploma_only** La percentuale di popolazione che risulta avere un livello di istruzione pari al diploma superiore e non oltre.



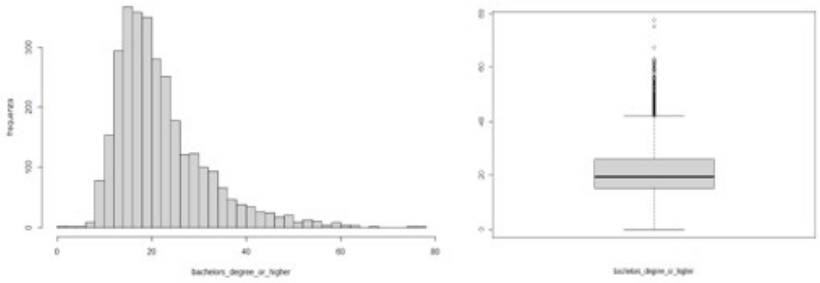
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 7.26 | 34.17 | 57.43 | 7.20 |

- **college_or_associates_degree** La percentuale di popolazione che ha conseguito una laurea che richiede il raggiungimento di sessanta crediti. La durata media di questo percorso è di circa due anni per uno studente full-time.



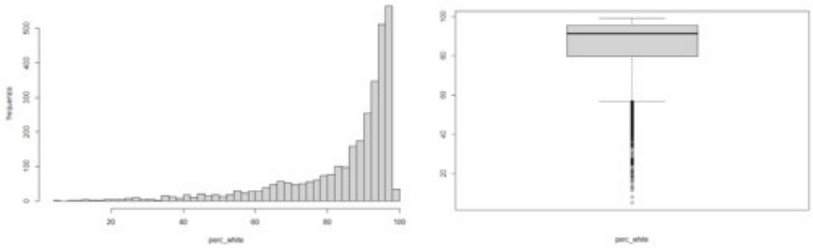
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 5.24 | 30.83 | 60.56 | 5.20 |

- **bachelor_degree_or_higher** La percentuale di popolazione che ha conseguito una laurea che richiede il raggiungimento di centoventi crediti, quindi con durata media di quattro anni, oppure un titolo superiore a questo (dottorato, master etc.).



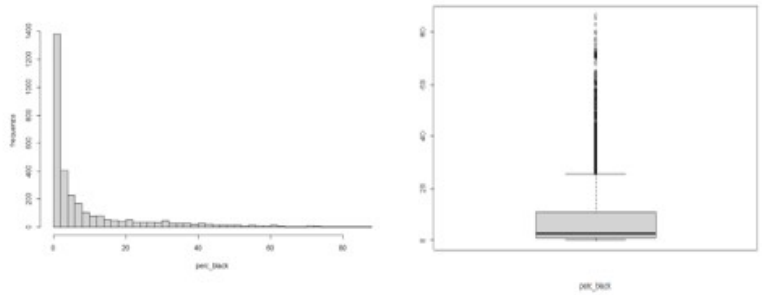
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0 | 21.92 | 77.56 | 9.52 |

- **perc_white** La percentuale di popolazione con la pelle bianca.



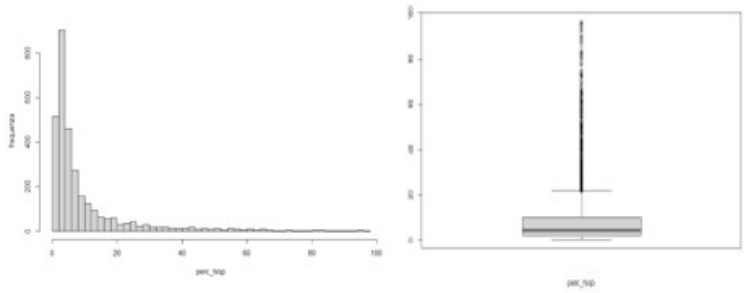
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 5.35 | 84.74 | 99.04 | 15.85 |

- **perc_black** La percentuale di popolazione con la pelle nera.



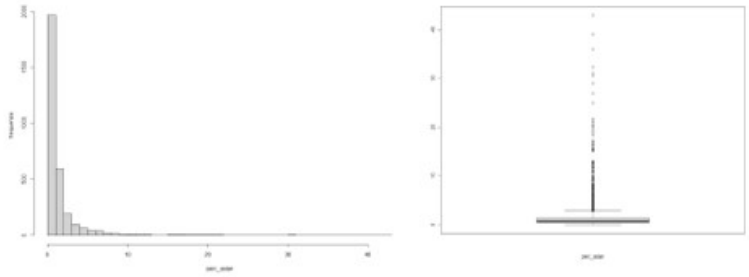
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.00 | 9.38 | 86.59 | 14.45 |

- **perc_hisp** La percentuale di popolazione con origini ispaniche.



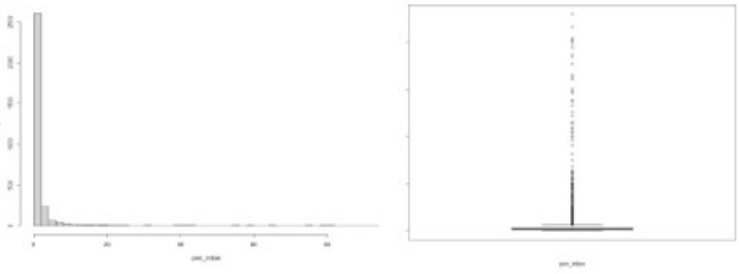
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.65 | 9.79 | 96.35 | 13.90 |

- **perc_asian** La percentuale di popolazione asiatica.



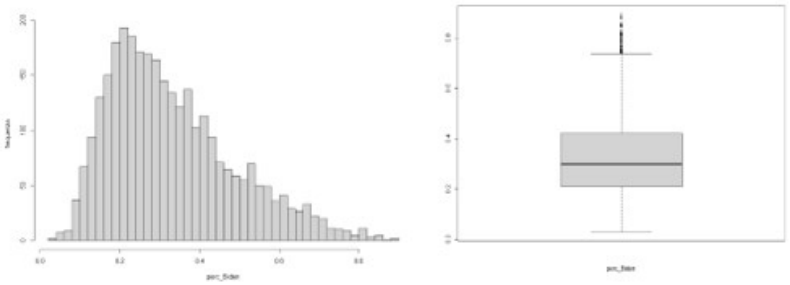
| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.00 | 1.52 | 42.94 | 2.74 |

- **perc_indian** La percentuale di popolazione indiana.



| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.00 | 2.12 | 92.38 | 6.74 |

- **perc_Biden** La variabile risposta che rappresenta la percentuale dei voti di Biden ottenuti all'elezioni presidenziali del 2020. La distribuzione mostra asimmetria sul lato destro, per cui possibilimente la FB potrebbe gestire meglio questa sua particolarità. Inoltre non vi sono contee con valori sugli estremi della dsitribuzione (0 e 1), per cui non sarà l'implementazione di un metodo ad hoc per gestire tali valori estremi.



| Min | Mean | Max | S.D. |
|---|---|---|---|
| 0.03 | 0.33 | 0.89 | 0.16 |

Per eliminare il problema di multicollinearità si è deciso di eliminare la variabile *perc_white* in quanto somma ad uno assieme alle altre variabili relative alla popolazione, e la variabile *bacherlor_degree_or_higher* che somma ad uno assieme alle altre variabili relative all'educazione.

## 5.3 Analisi delle correlazioni

Infine, attraverso la matrice di correlazione (Figura 2), si è trovata una correlazione abbastanza alta tra le variabili economiche (*poverty_percent, household_income* e *Med_HH_income_percent*), tuttavia si è preferito mantenere le variabili in quanto la correlazione non era eccessivamente alta (non supera lo 0.8).

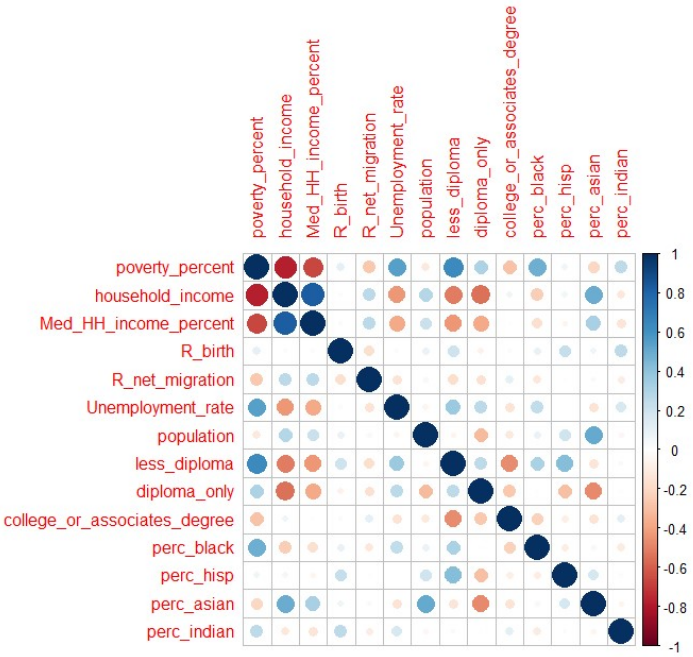



Figura 2: La matrice di correlazione delle variabili

# 6 Applicazioni e risultati

## 6.1 I limiti del modello lineare

Come già detto precedentemente, il modello lineare in presenza di una variabile risposta con valori limitati nell'intervallo $(0, 1)$ presenta diverse problematiche; proprio a causa di queste problematiche è stato creato il modello beta regression e le sue relative estensioni.

In questa sezione andremo a mostrare i limiti del modello lineare applicato alla variabile trasformata attraverso la funzione **logit** $g(y) = log(\frac{y}{1-y})$ che permette di mappare i valori della variabili risposta dall'intervallo unitario alla scala dei reali.
Come metodo di selezione delle variabili si è deciso di utilizzare l'AIC[5] in entrambe le direzioni, partendo dal modello saturo con tutte le variabili selezionate nel capitolo precedente. La selezione delle variabili così descritta porta a mantenere tutte le variabili tranne *R_net_migration*.
I risultati del modello lineare sono i seguenti

| Variabili | Stima | Standard error | p-value |
|---|---|---|---|
| Intercetta | -0.783 | 0.008 | 0.000 |
| poverty_percent | -0.096 | 0.019 | 0.000 |
| household_income | 0.095 | 0.022 | 0.000 |
| Med_HH_income_percent | -0.195 | 0.015 | 0.000 |
| R_birth | -0.025 | 0.009 | 0.006 |
| Unemployment_rate | 0.183 | 0.009 | 0.000 |
| population | 0.034 | 0.009 | 0.000 |
| less_diploma | -0.377 | 0.014 | 0.000 |
| diploma_only | -0.259 | 0.012 | 0.000 |
| college_or_associates_degree | -0.265 | 0.011 | 0.000 |
| perc_black | 0.424 | 0.011 | 0.000 |
| perc_hisp | 0.166 | 0.011 | 0.000 |
| perc_asian | 0.089 | 0.011 | 0.000 |
| perc_indian | 0.149 | 0.009 | 0.000 |

Da questi tabella si può dedurre che la variabile più influente, a parità di deviazione standard, è *perc_black*, la quale rappresenta la suddivisione etnica della popolazione, e a seguire *less_diploma, diploma_only* e *college_or_associates_degree* che rappresentano le tre variabili relative all'educazione della popolazione. Si noti inoltre che tutte le variabili risultano significative per alpha minore di 0.001 tranne *R_birth*, la quale risulta significativa per alpha minore di 0.006.

[5] Akaike information criterion

Tuttavia vi sono diverse problematiche nella stima di questo modello. Innanzitutto, i valori fittati dal modello non rispettano il range della variabile risposta originaria:

| Min | Mean | Max . |
|---|---|---|
| -2.877 | -0.784 | 2.23 |

Inoltre non sono rispettate diverse assunzioni alla base del modello lineare, tra cui:

- autocorrelazione fra i residui. Infatti il Durbin-Watson test [10] riporta un valore di DW=1.305, con p-value $<$ 2.2e-16.
- normalità dei residui. In questo caso è stato applicato lo Shapiro-Wilk Test [11] che ha riportato un valore della statistica SW=0.98 con un p-value $<$ 2.2e-16.
- omoschedasticità dei residui. Qui è stato utilizzato il Breusch and Pagan test [12], il quale ha riportato una statistica test pari a BP=200.19, con il relativo p-value$<$ 2.2e-16.

## 6.2 Beta regression model

Come si è potuto notare dai risultati ottenuti, il modello lineare non è adatto alle situazioni in cui la variabile risposta è limitata in un intervallo (unitario o non); proprio per questi casi è stato creato il modello di regressione beta. Vengono quindi ora riportati i risultati ottenuti con l'applicazione della libreria **betareg** di R, che ha permesso la stima del modello di regressione beta, che utilizza la statistica inferenziale spiegata nella parte teorica precendete. La selezione delle variabili è stata svolta ancora secondo AIC, la quale ha portato alla rimozione della variabile *R_net_migration*, mantenendo tutte le altre variabili.

Come possiamo notare dalla tabella dei coefficienti, vi è una correlazione positiva fra la variabile risposta (percentuale dei voti di Biden) e le variabili *household_income, Unemployment_rate, population* e tutte le variabili riguardanti la percentuale di popolazione[6]. Invece le variabili che sono correlate negativamente con la variabile risposta

[6]Queste variabili riguardano principalmente le minoranze della popolazione, come gli indiani o gli asiatici. Infatti ci aspettiamo che la maggioranza della popolazione, la popolazione bianca, avesse favoritismi nei confronti dell'altro candidato, Trump. Questo in quanto la popolazione bianca era altamente correlata negativamente con quella nera.

sono *poverty_percent, Med_HH_income_percent, R_birth* e tutte e tre quelle relative all'educazione. Il coefficiente pseudo $R^2$ risulta essere pari a 0.653.

| Variabili | Stima | Standard error | p-value |
|---|---|---|---|
| Intercetta | -0.746 | 0.008 | 0.000 |
| poverty_percent | -0.084 | 0.018 | 0.000 |
| household_income | 0.092 | 0.014 | 0.000 |
| Med_HH_income_percent | -0.194 | 0.014 | 0.000 |
| R_birth | -0.026 | 0.008 | 0.002 |
| Unemployment_rate | 0.168 | 0.009 | 0.000 |
| population | 0.033 | 0.009 | 0.000 |
| less_diploma | -0.363 | 0.014 | 0.000 |
| diploma_only | -0.256 | 0.011 | 0.000 |
| college_or_associates_degree | -0.254 | 0.011 | 0.000 |
| perc_black | 0.403 | 0.009 | 0.000 |
| perc_hisp | 0.165 | 0.010 | 0.000 |
| perc_asian | 0.089 | 0.011 | 0.000 |
| perc_indian | 0.144 | 0.009 | 0.000 |
| phi | 26.156 | 0.656 | 0.000 |

Il modello di regressione beta è stato stimato attraverso l'utilizzo della funzione link **logit**, la quale permette di fare un'interpretazione dei coefficienti in termini di odds ratio. Infatti $exp(\beta)$ risulta essere pari all'odds ratio[7]. L'interpretazione dell'odds ratio è tale che: considerando una variabile $x_1$, il cui coefficiente di regressione è pari a $\beta_1$, l'odds ratio (OR) è pari a $\exp(\beta_1)$ e OR-1 è pari alla variazione percentuale della variabile risposta data dalla variazione unitaria della variabile indipendente a parità delle rimanenti covariate. Per esempio, se $\beta_1 = 0.1$, allora $OR = exp(0.1) = 1.10$, per cui al variare di un'unità di $x_1$ si otterebbe una variazione della variabile risposta pari a $OR - 1 = 1.10 - 1 = 0.1 = 10\%$.

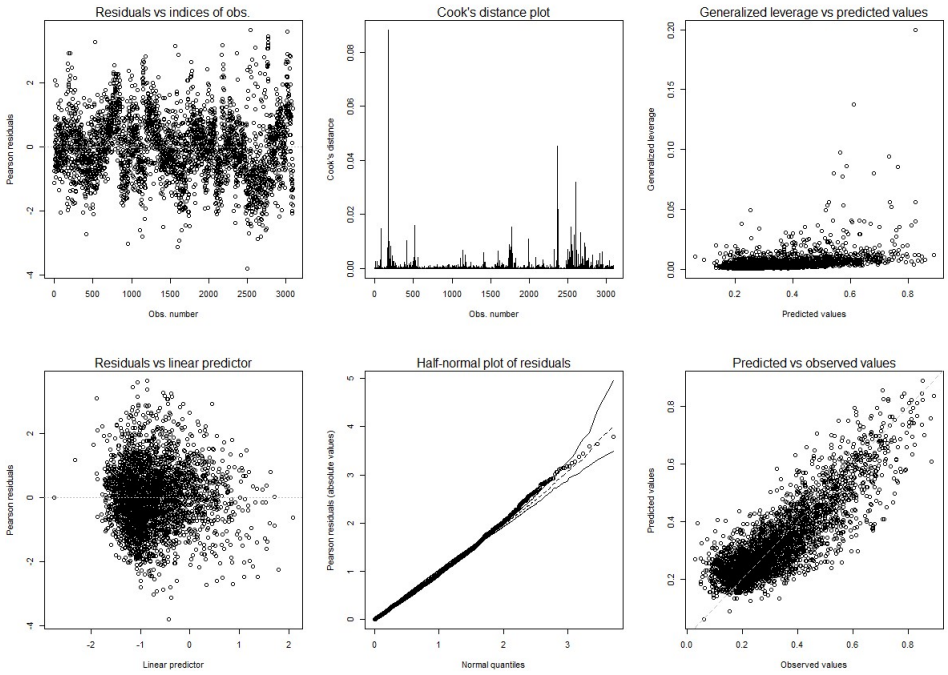
A causa della standardizzazione delle variabili, per l'interpretazione dell'OR, la variazione unitaria della variabile indipendente non sarà più unitaria, ma sarà una variazione pari alla deviazione standard della variabile originale. Per semplificare le interpretazioni riporto quindi sia la deviazione standard delle variabili, sia il loro OR.

[7]dove $\beta$ è un qualunque coefficiente relativo ad una variaibile.

| Variabili | Odds ratio | Standard Deviation |
|---|---|---|
| poverty_percent | 0.919 | 5.79 |
| household_income | 1.096 | 14 421 |
| Med_HH_income_percent | 0.824 | 19.86 |
| R_birth | 0.974 | 2.33 |
| Unemployment_rate | 1.183 | 1.39 |
| population | 1.034 | 335 785 |
| less_diploma | 0.696 | 6.26 |
| diploma_only | 0.774 | 7.20 |
| college_or_associates_degree | 0.776 | 5.20 |
| perc_black | 1.496 | 14.45 |
| perc_hisp | 1.179 | 13.90 |
| perc_asian | 1.093 | 2.74 |
| perc_indian | 1.155 | 6.74 |

A parità di Standard Deviation, la variabile indipendente che risulta più influente in termini di odds ratio per la variabile risposta è *perc_black*. L'interpretazione di quest'ultima è quindi la seguente: a fronte di un aumento di 14.45 punti percentuali della popolazione nera di una contea, si ha ragione di scommettere su un aumento dei voti di Biden del 49.6%. RIguardo alle variabili economiche, è possibile notare una correlazione positiva tra la percentuale dei voti di Biden e il reddito medio famigliare, tuttavia, grazie alla variabile *Med_HH_income_percent*, notiamo che se il reddito della famiglia è sproporzionato alla media del proprio Stato, questo andrà ad impattare negativamente sulla variabile risposta.

Riguardo alle covariate riferite all'educazione, i loro coefficienti variano in base alla variabile esplicativa che si decide di escludere, il che rende l'interpretazione di queste variabili complessa. Inoltre, tale interpretazione è difficoltosa in quanto, quando varia una variabile, di conseguenza variano pure le altre, per cui l'interpretazione con l'odds ratio risulta inesatta. Al fine dell'interpretazione di tali variabili, ho stimato a parte un modello in cui al loro posto vi è una variabile indice che riassume il contributo di tutte e quattro le variabili, ed è così costituita: ind_educ= (-3) * less_diploma + (-1)*diploma_only + college_or_associates_degree + 3 * bachelors_degree_or_higher. Tale variabile è stata standardizzata e poi inserita nel modello al posto delle variabili educative, risultando altamente significativa e altamente correlata con la variabile risposta (coefficiente pari a 0.378). Questa variabile ci permette di affermare senza nessun dubbio che all'aumentare del livello di educazione della popolazione, la variabile risposta aumenterà a sua volta.




Per la diagnostica del modello di regressione beta verranno utilizzati i grafici standard proposti dalla libreria **betareg**. La libreria propone diversi tipi di residui, in questa analisi è stato deciso di utilizzare i residui di Pearson. Il primo grafico, relativo ai residui in funzione delle osservazioni, non riporta trend sistematici, per cui non presenta problemi. Il secondo grafico presenta le distanze di Cook delle osservazioni. In riferimento all'analisi delle distanze di Cook ci sono due scuole di pensiero: la prima prende come soglia di cut off il valore 1, la seconda il valore 4/n. Nella prima nessun punto risulterebbe superiore alla soglia, nella seconda invece ci sarebbero circa 200 valori che superano la soglia. In particolare, come si denota dal grafico, vi sono 3 contee che spiccano subito alla vista per un'elevata distanza di Cook, queste sono: Los Angeles, della California, che presenta più di dieci milioni di abitanti e la sua variabile risposta ha un valore pari a 0.71 che è quasi estremo considerando la sua distribuzione, Oglala, del Sud Dakota, che è composta da una tribù di nativi americani, e che riposta una valore della variabile riposta pari a 0.88, e infine Kenedy, del Texas, la quale è la contee con il livello più alto di percentuale di popolazione con un livello di educazione inferiore al diploma (73.56). Considerando la seconda soglia di Cook (4/n=0.001), e togliendo le osservazioni che superano tale soglia, si ottiene un aumento del coefficiente pseudo $R^2$ di 0.05, arrivando a 0.70.

Il quarto grafico mostra i residui in dipendenza dal predittore lineare; essendo i punti uniformemente dispersi, vuol dire che la specificazione della funzione link (in questo caso **logit**) è da considerarsi corretta. Per completezza si è calcolato l'AIC relativo alle diverse link function:

| | logit | probit | cloglog | loglog |
|---|---|---|---|---|
| AIC | -6434.324 | -6422.616 | -6371.868 | -6407.669 |

La migliore funzione link risulta essere la funzione **logit**.

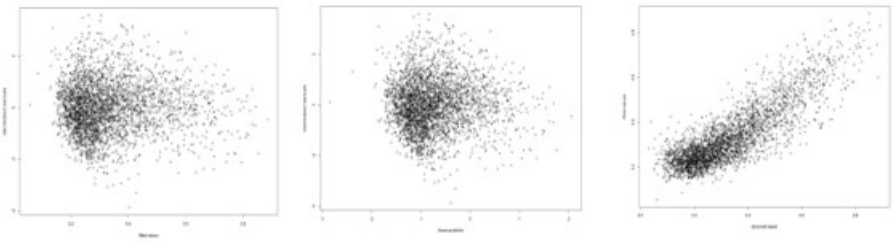
Infine gli ultimi due grafici mostrano un buon adattamento del modello ai dati, con una maggiore dispersione in prossimità dei valori più alti della variabile risposta (coda destra della distribuzione).

## 6.3 Flexible beta

Come modello alternativo a quello di regressione beta semplice è stato proposto il modello flexible beta, il quale è particolarmente adatto a gestire situazioni di bimodalità, asimmetria e comportamenti delle code anomali. Per stimare questo modello è stato utilizzato il pacchetto di R **FlexReg**, il quale utilizza metodi di stima basati sull'approccio Bayesiano attraverso l'algoritmo Hamiltonian Monte Carlo [13]. Vengono quindi riportati i risultati della stima del modello di regressione FB utilizzando la funzione link **logit**.

| Variabili | Post. Mean | 2.5% | 97.5% |
|---|---|---|---|
| Intercetta | 0.111 | -0.014 | 0.229 |
| poverty_percent | -0.083 | -0.118 | -0.048 |
| household_income | 0.086 | 0.047 | 0.124 |
| Med_HH_income_perc | -0.009 | -0.011 | -0.008 |
| R_net_migration | -0.011 | -0.028 | 0.006 |
| R_birth | -0.034 | -0.049 | -0.017 |
| Unemployment_rate | 0.166 | 0.148 | 0.184 |
| population | 0.031 | 0.013 | 0.049 |
| less_diploma | -0.373 | -0.402 | -0.346 |
| diploma_only | -0.261 | -0.283 | -0.240 |
| college_or_associates_degree | -0.254 | -0.274 | -0.234 |
| perc_black | 0.400 | 0.381 | 0.419 |
| perc_hisp | 0.173 | 0.153 | 0.193 |
| perc_asian | 0.083 | 0.062 | 0.104 |
| perc_indian | 0.143 | 0.127 | 0.159 |
| phi | 27.873 | 26.370 | 29.432 |
| p | 0.990 | 0.979 | 0.997 |
| w | 0.616 | 0.480 | 0.744 |

Per valutare la significatività delle stime ottenute dal pacchetto **FlexReg**, gli intervalli di confidenza tra il 2.5% e il 97.5% dei coefficienti non devono contenere lo zero, affinchè gli stessi coefficienti siano significativi. In particolare non risultano significativi l'intercetta e *R_net_migration*, variabile che pure nel modello di regressione beta era stata scartata. Inoltre il parametro p relativo alla mistura, mostra la prevalenza quasi totalitaria di una componente della mistura rispetto all'altra.



La diagnostica è stata svolta utilizzando i residui standardizzati proposti dalla libreria **FlexReg**. Il primo grafico rappresenta i residui standardizzati rispetto ai valori stimati; non vi sono particolari trend, per cui non mostra problemi riguardo alla regressione. Lo stesso ragionamento si applica al secondo grafico, che invece mostra i residui rispetto al predittore lineare, per cui conferma il **logit** come una buona scelta per la funzione link. Infinte il terzo grafico, il quale riporta i valori stimati rispetto ai valori osservati, ci mostra che vi è ancora una maggior dispersione dei valori nella coda destra della distribuzione.

## 6.4 Modelli con effetti casuali

L'implementazione degli effetti casuali è avvenuta attraverso il pacchetto **rstan**, il quale fa uso dell'inferenza Bayesiana attraverso i metodi Monte Carlo basati su catene di Markov (MCMC). In questo elaborato non si entrerà nello specifico degli aspetti tecnici della stima, bensì l'interesse sarà circoscritto agli aspetti interpretativi e di confronto con gli altri modelli. In particolare si è deciso di limitare gli effetti casuali alla componente dell'intercetta sia per il modello di

Tabella 1: Stime modello di regressione beta

| Variabili | Mean | 2.5% | 97.5% |
|---|---|---|---|
| Intercetta | -0.66 | -0.77 | -0.55 |
| poverty_percent | -0.04 | -0.07 | -0.01 |
| household_income | 0.10 | 0.02 | 0.18 |
| Med_HH_income_perc | -0.19 | -0.26 | -0.12 |
| R_net_migration | 0.02 | 0.01 | 0.04 |
| R_birth | -0.03 | -0.05 | -0.02 |
| Unemployment_rate | 0.06 | 0.05 | 0.08 |
| population | 0.01 | -0.01 | 0.02 |
| less_diploma | -0.31 | -0.33 | -0.29 |
| diploma_only | -0.28 | -0.30 | -0.27 |
| college_or_associates_degree | -0.22 | -0.24 | -0.20 |
| perc_black | 0.51 | 0.49 | 0.53 |
| perc_hisp | 0.31 | 0.29 | 0.34 |
| perc_asian | 0.04 | 0.02 | 0.06 |
| perc_indian | 0.18 | 0.16 | 0.19 |
| phi | 50.96 | 48.39 | 0.19 |
| sigma_u | 0.39 | 0.32 | 0.49 |

regressione beta sia per quello flexible beta. Vengono riportate quindi prima le stime della regressione beta con effetti casuali sull'intercetta, e in seguito quella della flexible beta. Sia nella regressione beta che in quella FB non risulta significativo solo il coefficiente relativo alla variabile *population*, in quanto contiene lo zero nell'intervallo di confidenza. Inoltre anche i coefficienti relativi a *R_birth* e *R_net_migration* sono al limite per essere considerati significativi. Anche in questo caso il parametro che risulta essere più influente a parità di deviazione standard è ancora *perc_black*, il quale addirittura raggiunge il valore di 0.51 in entrambi i modelli; a seguire poi vi sono *perc_hisp* e *less_diploma*.

In entrambi i modelli con effetti casuali possiamo notare che il parametro di dispersione *phi* è stato stimato a circa il doppio rispetto alla stima dei modelli senza gli effetti casuali. In entrambi i modelli il parametro relativo alla varianza degli errori casuali *sigma_u* è stimato pari a 0.39, il quale rappresenta la varianza dell'errore casuale che va ad influire sull'intercetta del modello.

Per confrontare i modelli stimati finora si è deciso di utilizzare

Tabella 2: Stime modello di regressione flexible beta

| Variabili | Mean | 2.5% | 97.5% |
|---|---|---|---|
| Intercetta | -0.67 | -0.78 | -0.55 |
| poverty_percent | -0.04 | -0.07 | -0.01 |
| household_income | 0.10 | 0.03 | 0.19 |
| Med_HH_income_perc | -0.20 | -0.28 | -0.12 |
| R_net_migration | 0.02 | 0.01 | 0.04 |
| R_birth | -0.04 | -0.05 | -0.02 |
| Unemployment_rate | 0.06 | 0.04 | 0.08 |
| population | 0.01 | -0.01 | 0.02 |
| less_diploma | -0.31 | -0.34 | -0.29 |
| diploma_only | -0.29 | -0.31 | -0.27 |
| college_or_associates_degree | -0.22 | -0.24 | -0.21 |
| perc_black | 0.51 | 0.50 | 0.53 |
| perc_hisp | 0.32 | 0.30 | 0.34 |
| perc_asian | 0.03 | 0.01 | 0.05 |
| perc_indian | 0.18 | 0.16 | 0.19 |
| phi | 53.88 | 51.14 | 56.65 |
| p | 1.00 | 0.99 | 1.00 |
| w | 0.62 | 0.48 | 0.75 |
| sigma_u | 0.39 | 0.32 | 0.48 |

il **WAIC** (Watanabe–Akaike information criterion)[14], il quale è la versione generalizzata dell'**AIC** (Akaike information criterion) per i modelli statistici singolari[15]. Per poter confrontare tutti e 4 i modelli (regressione beta e FB, con e senza effetti casuali), è stato stimato di nuovo il modello di regressione beta utilizzando il pacchetto **FlexReg**, in modo da poterne ricavare il WAIC. Inoltre le librarie permettono di calcolare il **LOO** (leave-one-out cross-validation) come ulteriore metodo di comparazione dei modelli.

| | WAIC | LOO |
|---|---|---|
| Beta regression | -6427.0 | -6426.9 |
| Flexible beta | -6463.5 | -6463.3 |
| Beta regression with random effects | -8444.3 | -8443.8 |
| Flexible beta with random effects | -8488.6 | -8488.0 |

Secondo entrambi i criteri di selezione, il migliore modello tra quelli proposti risulta essere il modello di regressione FB con effetti casuali.

Inoltre dalla tabella si può notare che, con o senza effetti casuali, il modello FB ottenga risultati migliori della semplice regressione beta.

# 7 Conclusioni

In questo elaborato sono stati trattati i modelli di regressione beta e flexible beta, con e senza l'aggiunta degli effetti casuali, attraverso l'uso del programma **R**. In primis è stata proposta una sintesi della teoria relativa alle distribuzioni seguita dalla specificazione dei modelli di regressione. Questi modelli poi sono stati applicati al caso d'interesse, ovvero al dataset riguardante le elezioni presidenziali americane del 2020. I dati analizzati sono stati raccolti per contea e la variabile risposta è stata considerata come la percentuale dei voti di Biden, la quale è stata regredita rispetto a variabili demografiche, suddivise principalmente in tre tipologie: educazione, economia ed etnia. In particolare le variabili esplicative in questione sono: poverty_percent, household_income, Med_HH_income_percent, R_net_migration, R_birth, Unemployment_rate, population,
less_diploma, diploma_only, colleger_or_associates_degree, perc_black, perc_hisp, perc_asian e perc_indian.

Nella parte applicativa inizialmente si è mostrato come il modello lineare, in presenza di variabile risposta limitata, presenti numerevoli problematiche e di conseguenza non sia applicabile. Quindi è stato stimato il modello di regressione beta attraverso l'utilizzo della libreria **betareg**. La stima del modello ha portato ad escludere la variabile *net_migration*, e ad identificare la variabile *perc_black* come quella maggiormente correlata con la risposta. Per l'interpretazione delle variabili relative all'educazione si è deciso di stimare un modello di regressione beta inserendo un indice che riassumesse il significato di queste, il quale ha portato ad un risultato chiaro: all'aumentare del grado di educazione della popolazione, aumenta a sua volta anche la variabile risposta. Anche il modello flexible beta porta ad escludere la variabile relativa all'immigrazione e ottiene stime simili a quelle della regressione beta.

Infine attraverso l'inserimento degli effetti casuali, sono stati stimati i modelli precedenti con l'intercetta variabile al variare dello Stato di appartenza della contea analizzata. Come è possibile notare dalle ultime tabelle sul confronto dei modelli, l'aggiunta degli effetti casuali porta ad un netto miglioramento; inoltre generalmente la regressione flexible beta porta risultati migliori della regressione beta base.

# Riferimenti bibliografici


[1] Silvia Ferrari & Francisco Cribari-Neto *beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, 31:7, 799-815, DOI: 10.1080/0266476042000214501

[2] Sonia Migliorati. Agnese Maria Di Brisco. Andrea Ongaro. *A New Regression Model for Bounded Responses* Bayesian Anal. 13 (3) 845 - 872, September 2018.https://doi.org/10.1214/17-BA1079

[3] Di Brisco AM, Migliorati S. *A new mixed-effects mixture model for constrained longitudinal data*. Statistics in Medicine. 2019;1–17. https://doi.org/10.1002/sim.8406

[4] Patrícia L. Espinheira , Silvia L.P. Ferrari & Francisco Cribari-Neto *On beta regression residuals*, Journal of Applied Statistics,(2008), 35:4, 407-419, DOI:10.1080/02664760701834931

[5] Atkinson, A. C. (1985) *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis* (New York: Oxford University Press)

[6] Wei, B.-C., Hu, Y.-Q. Fung, W.-K. (1998) *Generalized leverage and its applications*, Scandinavian Journal of Statistics, 25, pp. 25-37.

[7] Cook, R. D. (1977) *Detection of influential observations in linear regression*, Technometrics, 19, pp. 15-18

[8] Gelman, Hill *Data Analysis Using Regression* (2007)

[9] Hunger et al *Longitudinal beta regression models for analyzing health-related quality of life scores over time.* BMC Medical Research Methodology 2012 12:144.

[10] Durbin, J. e Watson, *Testing for serial correlation in least squares regression* G.S (1951) , II, Biometrica, 38, 159-179.

[11] Shapiro, S. S.; Wilk, M. B. (1965). *An analysis of variance test for normality (complete samples).* Biometrika. 52 (3–4): 591–611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384. p. 593

[12] Breusch, T. S.; Pagan, A. R. (1979). *A Simple Test for Heteroskedasticity and Random Coefficient Variation.* Econometrica. 47 (5): 1287–1294. doi:10.2307/1911963. JSTOR 1911963. MR 0545960.

[13] Gelman, A.; Carlin, J. B.; Stern, H. S. and Rubin, D. B. (2014) *Inference is dealt with a Bayesian approach based on the Hamiltonian Monte Carlo (HMC) algo-rithm* doi:10.1201/b16018

[14] Watanabe, Sumio (2010). *Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory.* Journal of Machine Learning Research. 11: 3571–3594.

[15] Watanabe, S. (2008), Accardi, L.; Freudenberg, W.; Ohya, M. (eds.), *Algebraic geometrical method in singular statistical estimation*, Quantum Bio-Informatics, World Scientific: 325–336, Bibcode:2008qbi..conf..325W, doi:10.1142/9789812793171_0024, ISBN 978-981-279-316-4.

# A Appendice

## A.1 Script di R

```
library(tidyverse)
library(dplyr)
library(corrplot)
library(betareg)
library(FlexReg)
library(rstan)
library(loo)
library(MCMCpack)
library(lmtest)
library(olsrr)


ed=dataset_education
dim(ed)
anyNA(ed)
sum(complete.cases(ed))
educ=ed[complete.cases(ed),]
anyNA(educ)
```

eliminati gli NA tolgo le osservazioni duplicate (createsi a causa di city-county)

```
anyDuplicated(educ$chiave)
anyDuplicated(ed$chiave)
sum(duplicated(educ$chiave))
doppionied=which(duplicated(educ$chiave))
doppionied
education=educ[-doppionied,]
dim(education)
anyDuplicated(education$chiave)
```

eliminati i doppioni è pronto ad essere unito agli altri

```
pop=dataset_population
dim(pop)
anyNA(pop)
popul=pop[complete.cases(pop),]
anyNA(popul)
anyDuplicated(popul$chiave)
```

```
dop.popul=which(duplicated(popul$chiave))
dop.popul
population=popul[-dop.popul,]
anyDuplicated((population$chiave))
dim(population)

pov=dataset_poverty
dim(pov)
anyNA(pov)
anyDuplicated(pov$chiave)
dop.pov=which(duplicated(pop$chiave))
dop.pov
poverty=pov[-dop.pov,]
dim(poverty)

unem=dataset_unemployment
dim(unem)
anyNA(unem)
anyDuplicated(unem$chiave)
dop.unem=which(duplicated(unem$chiave))
unemployment=unem[-dop.unem,]
anyDuplicated(unemployment$chiave)
dim(unemployment)
```

IMPORT BIDEN METTENDOLA COME NUMERIC

```
perc=percentuali_biden
dim(perc)
anyNA(perc)
anyDuplicated(perc$chiave)
dop.perc=which(duplicated(perc$chiave))
percentuali=perc[-dop.perc,]
anyDuplicated(percentuali$chiave)
dim(perc)

race=dataset_race
etnia=filter(race, year==2019)
dim(etnia)
anyDuplicated((etnia$chiave))
dop.etnia=which(duplicated(etnia$chiave))
dop.etnia
```

```
origin=etnia[-dop.etnia,]
dim(origin)
anyDuplicated(origin$chiave)
originni=mutate(origin, perc_white=white_pop/pop, perc_black=black_pop/pop,
perc_hisp=hisp_pop/pop, perc_asian=asian_pop/pop, perc_indian=indian_pop/pop)
summary(originni)
origini=dplyr::select(originni, chiave, perc_white, perc_black, perc_hisp,
perc_asian, perc_indian)
```

puliti tutti i dataset procedo alla unione di questi

```
dim(poverty)
dim(population)
dim(unemployment)
dim(education)
dim(perc)
dim(origini)

unione=inner_join(poverty, population, by = c("chiave" = "chiave"))
dim(unione)
unione2=inner_join(unione, unemployment, by = c("chiave" = "chia-
ve"))
dim(unione2)
unione3=inner_join(unione2, education, by = c("chiave" = "chiave"))
dim(unione3)
unione5=inner_join(unione3, origini, by= c("chiave" = "chiave"))
dim(unione5)
unione4=inner_join(unione5, percentuali, by = c("chiave" = "chia-
ve"))
dim(unione4)

anyNA(unione4)
anyDuplicated(unione4$chiave)
dop=which(duplicated(unione4$chiave))
prova=unione4[-dop,]
anyDuplicated(prova)
dim(prova)
summary(prova)
data=prova[,c(1,5,3,4,8,9,10,11,13,14,17,18,19,20,21,22,23,24,25, 27)]
summary(data)
dim(data)
```

```
dati=mutate(data,percBiden = BIDEN*0.01)

hist(dati$percBiden,breaks=100)
summary(dati$percBiden)

dati2=mutate(dati, trasfBiden= log((percBiden)/(1-percBiden)))
summary(dati2)
dim(dati2)
summary(dati2)
str(dati2)
M=cor(dati2[,3:19])
M
corrplot(M)
hist(dati2$percBiden, breaks=50,xlab= "Percentuali Biden" )

datifixed=mutate(dati2, income = householdincome / 1000 , net_migration
= NET_MIG / 1000, population = POP_ESTIMATE / 100000)
summary(datifixed)
datifixed=datifixed[,-c(2,4,5,6,18)]
datifixed=mutate(datifixed, ind_educ=(-3)*less_diploma + (-1)*diploma_only
+ college_or_associates_degree + 3*bachelors_degree_or_higher)
datiscaled=mutate(datifixed, household_income=scale(datifixed$income),
population = scale(datifixed$population),
poverty_percent=scale(datifixed$povertypercent),
R_birth = scale(datifixed$R_birth),
R_net_migration = scale(datifixed$R_NET_MIG),
Med_HH_income_percent_ =scale(datifixed$Med_HH_Income_Percent_of_State_Total),
unemployment_rate = scale(datifixed$Unemployment_rate),
less_diploma = scale(datifixed$less_diploma),
diploma_only = scale(datifixed$diploma_only),
college_or_associates_degree =
scale(datifixed$college_or_associates_degree),
perc_black = scale(datifixed$perc_black),
perc_hisp = scale(datifixed$perc_hisp),
perc_asian = scale(dati2$perc_asian),
perc_indian = scale(dati2$perc_indian),
ind_educ= scale(datifixed$ind_educ))
summary(datiscaled)
datiscaled=datiscaled[,-c(2,4,5,6,10,11,15,18,19)]
```

ANALISI ESPLORATIVA DELLE VARIABILI

```
hist(scale(dati2$povertypercent), breaks=20, xlab = "poverty_percent",
ylab="frequenza")
summary(dati2$povertypercent)
boxplot(dati2$povertypercent, xlab="povertypercent")
sd(dati2$povertypercent)

hist(dati2$householdincome, breaks=20, xlab = "householdincome",
ylab="frequenza")
boxplot(dati2$householdincome,xlab="householdincome")
summary(dati2$householdincome)
sd(dati2$householdincome)

hist(dati2$Med_HH_Income_Percent_of_State_Total, breaks=20, xlab
= "Med_HH_income_percent", ylab="frequenza")
boxplot(dati2$Med_HH_Income_Percent_of_State_Total,xlab="Med_HH_income_percent")
summary(dati2$Med_HH_Income_Percent_of_State_Total)
sd(dati2$Med_HH_Income_Percent_of_State_Total)

hist(dati2$POP_ESTIMATE, breaks=200, xlab = "population", ylab="frequenza",
xlim=c(0,1000000))
boxplot(dati2$POP_ESTIMATE,xlab="population")
summary(dati2$POP_ESTIMATE)
sd(dati2$POP_ESTIMATE)

hist(dati2$R_NET_MIG, xlab="R_net_migration", breaks=30)
boxplot(dati2$R_NET_MIG, xlab="R_net_migration")

summary(dati2$R_NET_MIG)
sd(dati2$R_NET_MIG)

hist(dati2$R_birth, breaks=20, xlab = "R_birth", ylab="frequenza")
boxplot(dati2$R_birth,xlab="R_birth")
summary(dati2$R_birth)
sd(dati2$R_birth)

hist(dati2$Unemployment_rate, breaks=20, xlab = "Unemployment_rate",
ylab="frequenza")
boxplot(dati2$Unemployment_rate,xlab="Unemployment_rate")
summary(dati2$Unemployment_rate)
sd(dati2$Unemployment_rate)
```

```
hist(dati2$less_diploma, breaks=30, xlab = "less_diploma", ylab="frequenza")
boxplot(dati2$less_diploma,xlab="less_diploma")
summary(dati2$less_diploma)
sd(dati2$less_diploma)

hist(dati2$diploma_only, breaks=30, xlab = "diploma_only", ylab="frequenza")
boxplot(dati2$diploma_only,xlab="diploma_only")
summary(dati2$diploma_only)
sd(dati2$diploma_only)

hist(dati2$perc_asian, breaks=40, xlab = "perc_asian", ylab="frequenza")
boxplot(dati2$perc_asian,xlab="perc_asian")
summary(dati2$perc_asian)
sd(dati2$perc_asian)

hist(dati2$perc_indian, breaks=40, xlab = "perc_indian", ylab="frequenza")
boxplot(dati2$perc_indian,xlab="perc_indian")
summary(dati2$perc_indian)
sd(dati2$perc_indian)
str(dati2)

dati2=mutate(dati2, perc_asian= perc_asian*100, perc_indian=perc_indian*100)

hist(dati2$percBiden, breaks=40, xlab = "perc_Biden", ylab="frequenza")
boxplot(dati2$percBiden,xlab="perc_Biden")
summary(dati2$percBiden)
sd(dati2$percBiden)
str(dati2)
```

LINEAR MODEL

```
linearmodel1=lm(trasfBiden    . -STATE.x - percBiden - less_diploma
- diploma_only -
college_or_associates_degree, datiscaled )
summary(linearmodel1)
stepmodel1=stepAIC(linearmodel1, direction="both")
summary(stepmodel1)
```

elimina R_net_migration per la selezione del modello migliore secondo

AIC

bestlmAIC1=lm(trasfBiden . -STATE.x - percBiden - R_net_migration
, datiscaled)
summary(bestlmAIC1)
summary(fitted(bestlmAIC1))
plot(fitted.values(bestlmAIC1))

diagnostica lm

bptest(bestlmAIC1)
dwtest(bestlmAIC1)
par(mfrow=c(2,2))
plot(bestlmAIC1)
boxplot(residuals(bestlmAIC1))

ols_test_normality(bestlmAIC1)

BETA REGRESSION

step(betamod, direction="both")
betamod=betareg(percBiden . -STATE.x - trasfBiden - R_net_migration
-ind_educ, datiscaled,
link = "logit")
summary(betamod)

AIC(betamod)

par(mfrow=c(2,3))
plot(betamod, which=1:6, type="pearson")

le 3 osservazioni con cook distance sopra la soglia sono Los Angeles della California, Oglala
Lakota tribu di nativi americani e
TX.Kenedy che è lo stato con il più alto tasso di persone senza nemmeno il diploma (73%)
provo a togliere i 3 stati con una cook distance elevata, ma ottengo un miglioramento trascurabile del R2 (0.02)

b=which(cooks.distance(betamod)¿0.03)

datifixed3=datiscaled[-b,]
betamod3=betareg(percBiden . -STATE.x - trasfBiden - R_net_migration, datifixed3, link = "logit")
summary(betamod3)

provo a togliere tutte quelle sopra la soglia di cook (4/n), che risultano essere circa 200, e ottengo un
miglioramento di 0.07 nel R quadro

a=which(cooks.distance(betamod)¿0.001)
datifixed2=datiscaled[-a,]
betamod2=betareg(percBiden . -STATE.x - trasfBiden - R_net_migration - ind_educ, datifixed2, link = "logit")
summary(betamod2)

plot(betamod2,which=1:6)

residualsP=residuals(betamod,type="pearson")

predlin=predict(betamod, type="link")
fitmean=predict(betamod, type="response")
plot(predlin, residualsP)

possibili link: "logit", "probit", "cloglog" , "loglog"
migliore link a livello di grafico è il "logit"

diagnostica

fitted=predict(modbeta, type="response")
residui=residuals(modbeta,type="standardized")
predittlin=predict(modbeta, type="link")

plot(fitted,residui, xlab="fitted values", ylab="standardized residuals")
plot(predittlin,residui, xlab="linear predictor", ylab="standardized residuals")
plot(datiscaled$percBiden,fitted,xlab ="observed values", ylab="fitted values")

summary(residuals(modbeta, type="standardized"))

FLEXREG

```
modbeta=flexreg(percBiden  poverty_percent + R_birth + R_net_migration
+ Med_HH_income_percent +
less_diploma + diploma_only + college_or_associates_degree + perc_black
+ perc_hisp + household_income +
population + unemployment_rate + perc_asian + perc_indian , da-
ta=datiscaled )
modbeta1=flexreg(percBiden  poverty_percent + R_birth + R_net_migration
+ Med_HH_income_percent +
less_diploma + diploma_only + college_or_associates_degree + perc_black
+ perc_hisp + household_income +
population + unemployment_rate + perc_asian + perc_indian , da-
ta=datiscaled, type="Beta")

summary(modbeta)
summary(modbeta1)

WAIC(modbeta)
WAIC(modbeta1)

STAN

n=length(datifixed$povertypercent)
y=datifixed$percBiden

X=cbind(rep(1,n),
income=scale(datifixed$income),
population = scale(datifixed$population),
povertypercent=scale(datifixed$povertypercent),
R_birth = scale(datifixed$R_birth),
R_NET_MIG = scale(datifixed$R_NET_MIG),
Med_HH_Income_Percent_of_State_Total =
scale(datifixed$Med_HH_Income_Percent_of_State_Total),
unemployment_rate = scale(datifixed$Unemployment_rate),
less_diploma = scale(datifixed$less_diploma),
diploma_only = scale(datifixed$diploma_only),
college_or_associates_degree = scale(datifixed$college_or_associates_degree),
perc_black = scale(datifixed$perc_black),
perc_hisp = scale(datifixed$perc_hisp),
perc_asian = scale(dati2$perc_asian),
perc_indian = scale(dati2$perc_indian))

livelli=dati$STATE.x
```

```
livelli=as.factor(livelli)
levels(livelli)=seq(from=1, to=49, by=1 )
str(livelli)
livelli=as.numeric(livelli)

FB_Mixed = rstan::stan_model("FB_mixed.stan")
Beta_Mixed = rstan::stan_model("Beta_Mixed.stan")

creo una lista con i dati necessari per la stima

data.stan = list(
N = n, y = y,
K = ncol(X), X = X,
sd_prior = 10, g=0.001,
J = 49, subject = livelli
)

n.iter = 5000  lunghezza della catena

fit.FB = rstan::sampling(
object = FB_Mixed,  Stan program
data = data.stan,  named list of data
chains = 1,  number of Markov chains
warmup = 0.5*n.iter,  number of warmup iterations per chain
iter = n.iter,  total number of iterations per chain
cores = 1,  number of cores (using 2 just for the vignette)
thin=1,
control = list(adapt_delta = .95),
refresh = n.iter/100  show progress every 'refresh' iterations
)

print(fit.FB, pars=c("beta", "phi", "p", "w", "sigma_u", "U"))

fit_Beta = rstan::sampling(
object = Beta_Mixed,  Stan program
data = data.stan,  named list of data
chains = 1,  number of Markov chains
warmup = 0.5*n.iter,  number of warmup iterations per chain
iter = n.iter,  total number of iterations per chain
cores = 1,  number of cores (using 2 just for the vignette)
thin=1,
control = list(adapt_delta = .8),
```

```
refresh = n.iter/100  show progress every 'refresh' iterations
)
print(fit_Beta, pars=c("beta", "phi", "sigma_u", "U"))

WAIC: Misura simile all'AIC per confrontare modelli. Piu' basso e',
meglio e'
waic(extract_log_lik(fit_Beta))
waic(extract_log_lik(fit.FB))
loo(fit_Beta)
loo(fit.FB)
```