US General Social Survey

Women Fertility

A. Cola, M. Simonetti, R. Urso

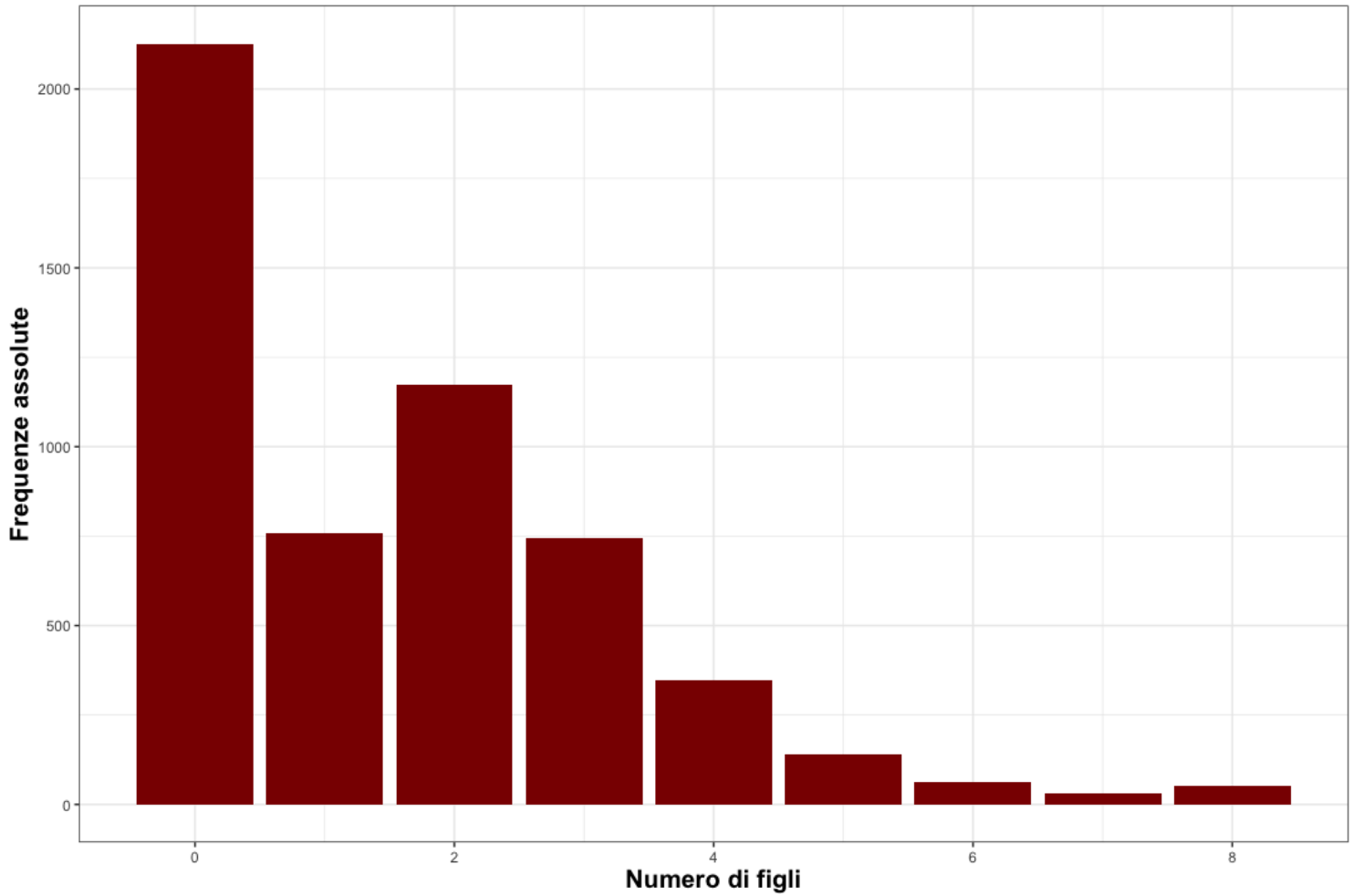
30 novembre 2022

Introduzione

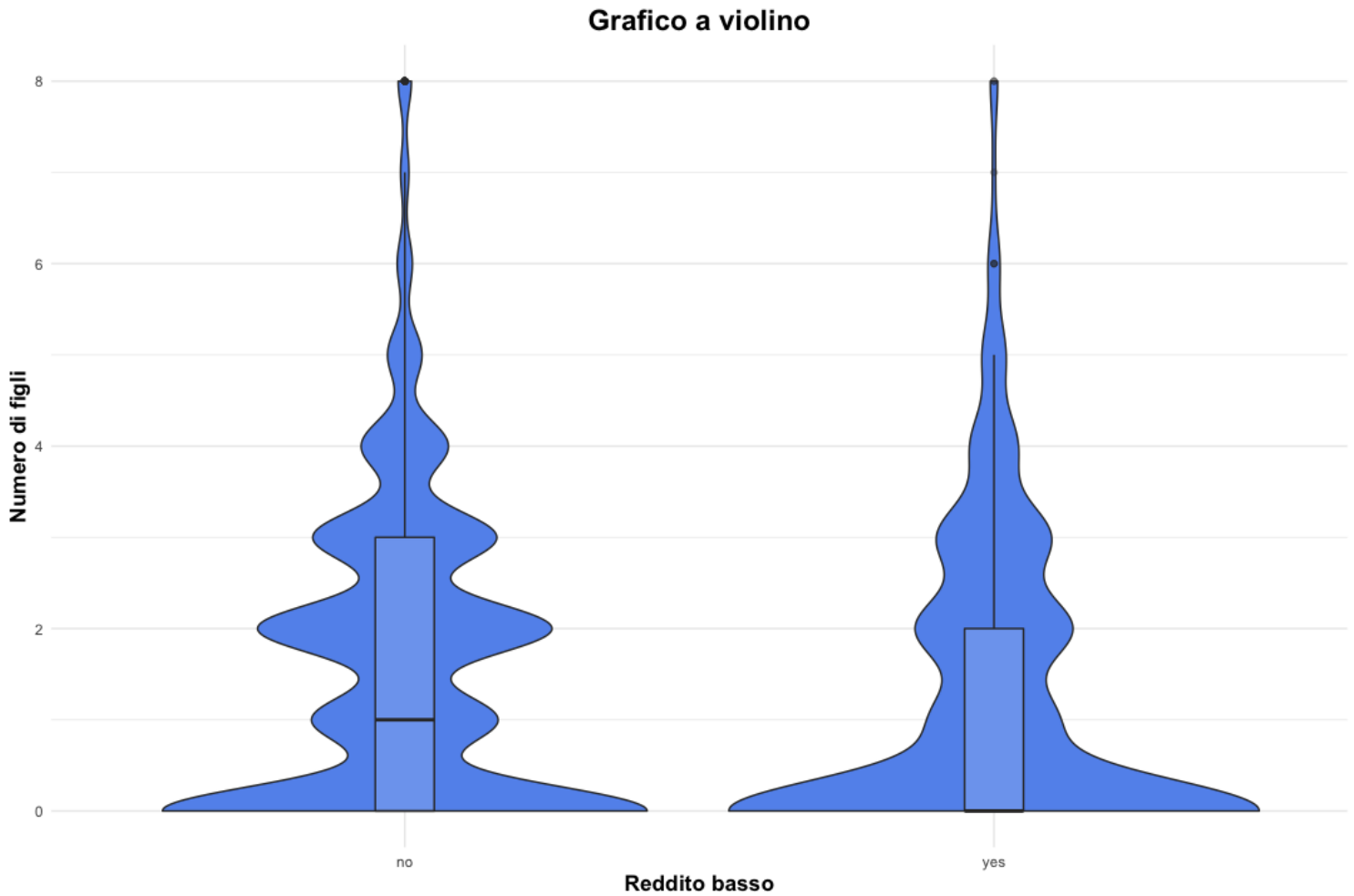
- Il dataset sul quale si intende condurre l'analisi riguarda un campione di 5439 donne rispondenti allo US General Social Survey (GSS).
- Sulle osservazioni sono state rilevate 9 variabili: numero di figli, età della rispondente, anni di educazione, numero di fratelli/sorelle, età prima nascita, etnia, residenza, il possesso di un reddito basso e l'essere o meno immigrate.
- Si vuole pervenire alla stima di un modello che meglio riesca a spiegare i fattori maggiormente influenti sul numero medio di figli.

Analisi esplorativa

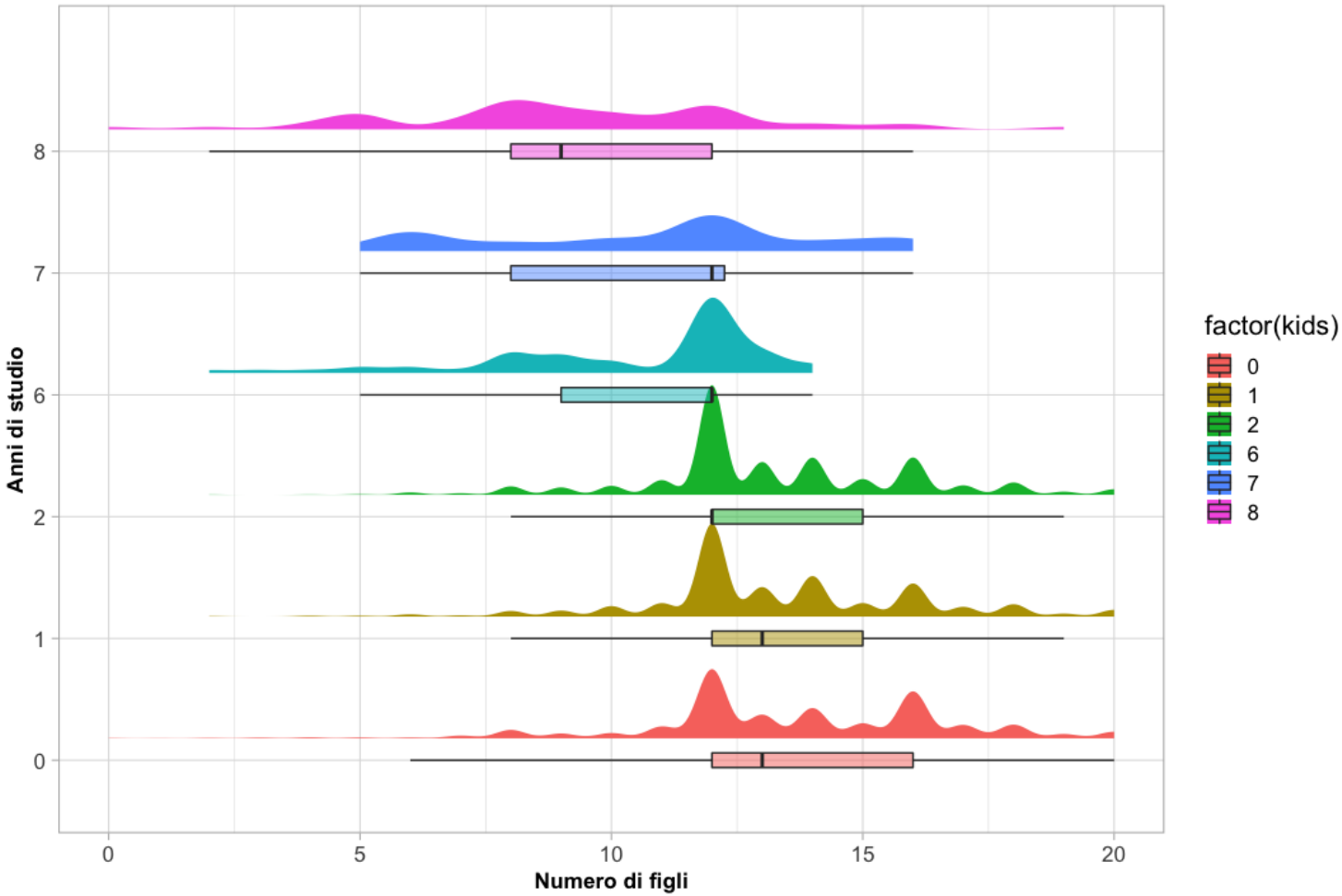
- Il primo passo consiste nella visualizzazione della distribuzione di frequenza della variabile figli.



Analisi esplorativa



Analisi esplorativa



Associazione tra variabili

- Per valutare l'associazione tra le variabili sono stati calcolati l'indice di correlazione di Pearson e il coefficiente di correlazione Punto-Biseriale.

Specificazione modello

- Il numero di figli Y_i è una variabile di conteggio la quale ha supporto $\mathbb{N}_0$. Per tale ragione si è scelto di specificare il seguente modello di Poisson:

$$log(\lambda_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, 2, \ldots, n; \tag{1}$$

dove λ_i è il valor medio di $Y_i \sim Po(\lambda_i)$, p è il numero delle variabili esplicative ed n è il numero delle osservazioni.

Stima modello

- Per procedere correttamente alla stima del modello è stato utilizzato un approccio forward, che ha segnalato la presenza di due interazioni statisticamente significative.
- Il criterio per la selezione delle interazioni è stato quello della devianza residua, in base al quale è stata inserita, ad ogni passo, l'interazione che generava la devianza più bassa.

	Devianza residua
Modello Minimale Adeguato	9289.6
Modello con età:educazione	9107.9
Modello con età:fratelli	9094.6

Stima modello

Variabili	Parametri stimati	Errori standard	Test Wald (p-value)
Intercetta	1.66	0.166	$<$ 2e-16
Età	-0.008	0.003	0.002
Educazione	-0.170	0.012	$<$ 2e-16
Fratelli	0.052	0.009	3.04e-09
Etnia	-0.233	0.027	$<$ 2e-16
Età:Edu	0.003	0.0002	$<$ 2e-16
Età:Fratelli	-0.0006	0.0002	0.000232

- Per avere un'idea della performance del modello, è stato effettuato un test chi quadrato di Pearson. Il valore della statistica test è risultato pari a ***7817*** (df 5432), conducendo a rifiutare l'ipotesi nulla di adattamento del modello ai dati.
- Una delle possibili cause della mancanza di adattamento è l'*overdispersion*, la quale si verifica quando $Var(Y) > \mathbb{E}(Y)$.
- Nel caso in esame si osserva:

Media	Varianza
1.55	2.81

il che suggerisce di procedere con la specificazione di un diverso modello di regressione.

Regressione Binomiale Negativa

- Il modello di regressione con risposta binomiale negativa sarà specificato da $\mu_i = exp(\sum_{j=1}^{p} \beta_j x_{ij})$, avendo supposto $Y_i \sim NB(\mu_i, \theta)$.

Variabili	Parametri stimati	Errori standard	Test Wald (p-value)
Intercetta	-0.015	0.091	0.867
Età	0.021	0.0008	$<$ 2e-16
Educazione	-0.036	0.005	1.28e-12
Fratelli	0.021	0.0045	2.54e-06
Etnia	-0.252	0.036	2.27e-12

- Il valore stimato del parametro di dispersione θ è risultato pari a 2.33

- Anche in questo caso è stato effettuato un test sulla bontà di adattamento che, questa volta, *non* ha condotto al rifiuto dell'ipotesi nulla ($X^2 = 4764$, df $= 5434$).
- Tuttavia, dal momento che la distribuzione empirica della risposta mostra un'inflazione di zeri, risulta opportuno procedere con la stima di un modello che sia in grado di tenere conto di questo aspetto.

Regressione con inflazione di zeri

- Si supponga

$$Y_i \sim \begin{cases} 0 & \text{con probabilità } \phi_i \\ P(\lambda_i) & \text{con probabilità } 1-\phi_i. \end{cases}$$

Un modello di Poisson con inflazione di zeri (ZIP) sarà un modello mistura tra una distribuzione degenere in 0 e una distribuzione di Poisson. Quindi si avrà:

$$\Pr(Y_i = j) = \phi_i + (1-\phi_i)\frac{e^{-\lambda_i}\lambda_i^j}{j!} \quad j = 0, 1, 2, \ldots$$

- Se nel modello si inseriscono covariate è possibile esprimere i parametri in loro funzione:

$$\text{logit}(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = \boldsymbol{x_i}\beta; \quad \log(\lambda_i) = \boldsymbol{x_i}\gamma$$

Regressione Binomiale Negativa con inflazione di zeri(ZINB)

Variabili	Parametri stimati	Errori standard	Test Wald (p-value)
Count model			
Intercetta	0.748	0.084	< 2e-16
Età	0.013	0.0008	< 2e-16
Educazione	-0.041	0.004	< 2e-16
Fratelli	0.011	0.0036	0.00157
Etnia	-0.150	0.030	9.1e-07
Zero-inflation model			
Intercetta	-0.01	0.17	0.94905
Età	-0.03	0.003	< 2e-16
Residenza	0.168	0.078	0.031
RedditoBasso	0.986	0.101	< 2e-16
Etnia	0.50	0.106	2.54e-06
Fratelli	-0.043	0.014	0.00184

- Ai fini interpretativi, è conveniente notare che:

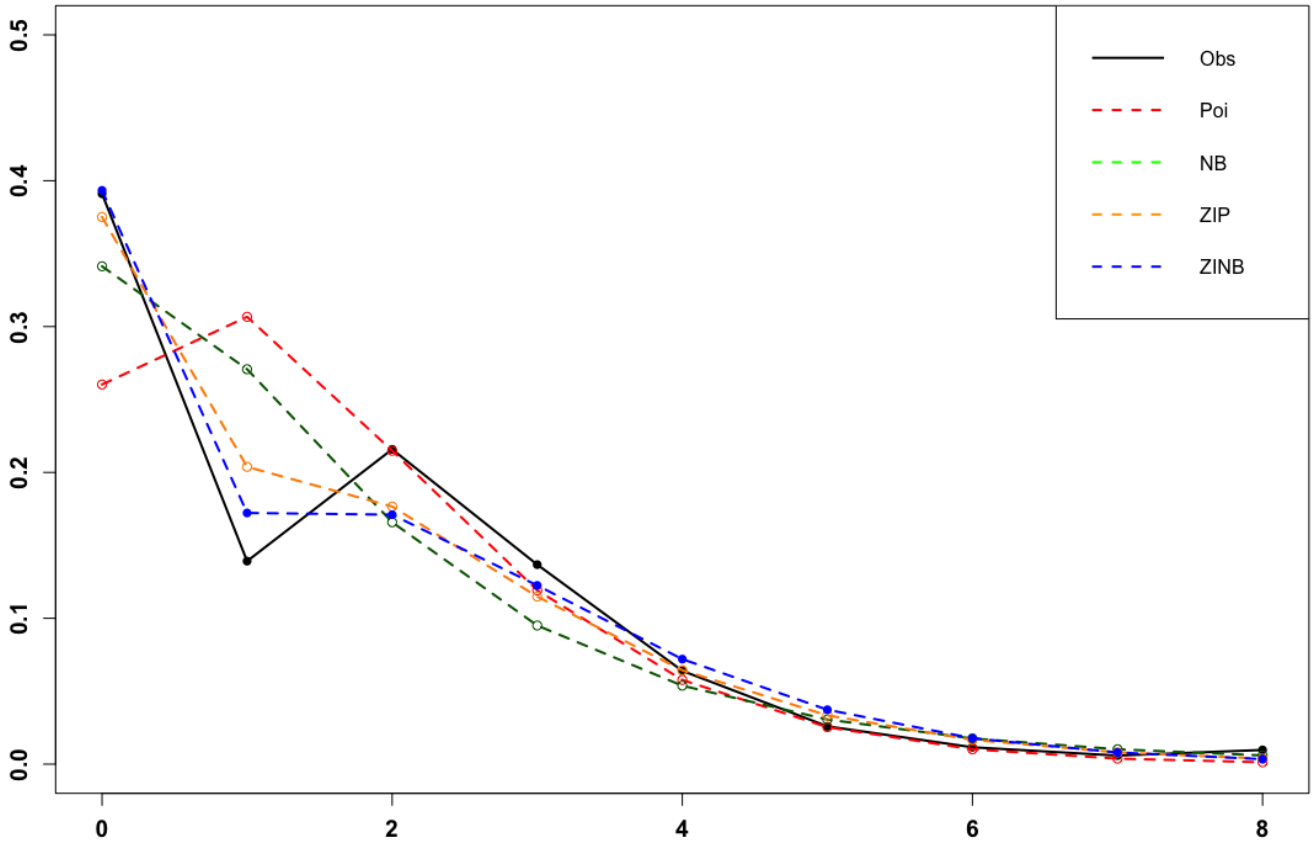
$$\lambda_i = exp(\sum_{j=1}^{p} \beta_j x_{ij}) = (e^{\beta_1})^{x_{i1}} \cdots (e^{\beta_p})^{x_{ip}}, \quad i = 1, 2, \ldots, n. \qquad (2)$$

  Ciò significa che, a parità delle altre variabili, un incremento unitario di x_{ij} determina un incremento del valore medio λ_i misurato dal fattore moltiplicativo e^{β_j}.

- Dai parametri stimati per la componente count è quindi possibile dedurre che un incremento unitario degli anni di studio determina una diminuzione del numero medio di figli pari al 4%;
- Dai parametri stimati per la componente zero-inflation si deduce che l'odds di non avere figli raddoppia per coloro che avevano un reddito basso a 16 anni.

Confronto modelli

- Un primo confronto tra modelli consiste nel comparare le frequenze relative osservate dei valori della Y_i con le frequenze teoriche stimate dai modelli.



Conclusioni

- In conclusione, è stato effettuato un ulteriore confronto tramite gli indici AIC e BIC.

	Indice AIC	Indice BIC
Modello Poisson	18296	18329
Modello Binomiale Negativa	17644	17684
Modello ZIP	17069	17129
Modello ZINB	16950	17030

Riferimenti

- Salvan, A., Sartori, N., Pace, L. (2020). Modelli lineari generalizzati. Springer, Milano.
- Agresti, A. (2018). An introduction to categorical data analysis. John Wiley Sons.