Modelli lineari generalizzati

Statistica Aziendale–prof.ssa Maria Grazia Pittau
grazia.pittau@uniroma1.it

Dipartimento di Scienze Statistiche - Sapienza Università di Roma

a.a. 2024-2025

Modelli lineari generalizzati I

I **modelli lineari generalizzati** rientrano in una metodologia di analisi statistica che include, come casi speciali, i modelli di regressione lineare e logistica. La **regressione lineare** prevede direttamente valori continui y in base a un *predittore lineare* $X_i^{'}\beta = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$.
La **regressione logistica** prevede la probabilità che y sia pari a 1, $\Pr(y = 1)$, quando y può assumere solamente valore 0 o valore 1 (dati binari) in base ad un predittore lineare con trasformazione logistica-inversa (definita `invlogit` in questa sede). Un **modello lineare generalizzato** coinvolge:

1. Un vettore di dati $y = (y_1, \ldots, y_n)$
2. Una matrice di predittori X e un vettore di coefficienti β, che formano un predittore lineare $X_i^{'}\beta$
3. **Una funzione link** G, che da luogo ad un vettore di dati trasformati $G(y) = (X_i^{'}\beta)$ che vengono usati per modellare i dati
4. Una distribuzione dei dati, $p(y|\hat{y})$

Modelli lineari generalizzati II

5. Possibilmente altri parametri, come varianze, parametri di *sovradispersione* e valori soglia *cutpoints*, relativi ai predittori, alla funzione link e alla distribuzione dei dati.

Le opzioni in un modello lineare generalizzato sono la trasformazione G e la distribuzione dei dati p.

- Nella *regressione lineare*, la trasformazione è la funzione identità (ovvero, $G(y) \equiv I(y)$) e la distribuzione dei dati è una distribuzione normale, con deviazione standard σ stimata in base al vettore dei dati.
- Nella *regressione logistica*, la trasformazione è la funzione logit $G(y) = \text{logit}(y) = X_i^{'}\beta$ e la distribuzione dei dati è definita dalla probabilità tipica dei dati binari: $\Pr(y=1) = \hat{y}$.
- Nella *regressione di Poisson*, la trasformazione è la funzione logaritmica, $G(y) = \log(y) = X_i^{'}\beta$ e la distribuzione dei dati è definita dalla probabilità dei dati di conteggio

I modelli lineari generalizzati che studieremo I

In questo corso veranno studiati oltre al **modello logistico** e **modello probit**, nel seguito brevemente riassunti, anche altre classi di modelli lineari generalizzati:

1. **modello logistico** in cui la variabile risposta di tipo binaria è funzione di un predittore lineare.
2. **modello probit** che coincide con il modello di regressione logistico a meno della funzione logistica che viene sostituita dalla funzione di distribuzione cumulata normale o, equivalentemente con la distribuzione normale invece della distribuzione logistica per gli errori nella formulazione in termini di variabili latenti

I modelli lineari generalizzati che studieremo II

3. **Modelli logit e probit multinomiali** sono estensioni dei modelli di regressione logistica e probit per dati categorici con più di due categorie, per esempio risposte da indagini del tipo "Molto d'accordo", "D'accordo", "Indifferente", "Disaccordo", "Fortemente in disaccordo". Questi modelli usano la trasformazione logit o probit e la distribuzione multinomiale ma richiedono parametri addizionali per modellare le molteplici sfaccettature dei dati.
   I modelli multinomiali sono ulteriormente classificati come *ordinati* (per esempio "Molto d'accordo", . . . , "Fortemente in disaccordo") o *non ordinati* (per esempio diversi credi religiosi).

I modelli lineari generalizzati che studieremo III

4. Il **modello di Poisson** che viene usato per modellare una variabile risposta di conteggio; ovvero numero di volte in cui si verifica un certo evento:
   - numero di visite al supermercato in una settimana,
   - numero di volte in cui si fa ricorso ad un certo servizio,
   - numero di assicurazioni stipulate nell'ultimo anno,
   - numero di brevetti registrati da un'impresa in un anno,
   - numero di fasi del processo produttivo esternalizzate,
   - numero di pezzi difettosi prodotti da un certo macchinario,
   - numero di assunzioni di personale.

   Quindi la variabile risposta y_i può essere uguale a 0, 1, 2, ... La trasformazione G è la trasformazione logaritmica. In questo modo se $G(y) = \log(y) = X_i^{'}\beta$, siamo sempre sicuri che i valori stimati $\theta_i = \hat{y}_i = \exp(X_i^{'}\beta)$ siano sempre positivi. La distribuzione dei dati è una Poisson e θ_i il parametro di Poisson.

I modelli lineari generalizzati che studieremo IV

5. Il **modello logistico-binomiale** viene usato per tutti quei dati puntuali y_i che rappresentano il numero di successi in un numero n_i di prove. (Il numero n_i di prove per ciascun dato i, non è pari al numero di dati complessivo n.) In questo modello la trasformazione $G(\cdot)$ è la funzione logit e la distribuzione dei dati la binomiale. Come nella regressione di Poisson, il modello binomiale viene notevolmente migliorato se si include un parametro per la *sovradispersione*.

Stima dei modelli lineari generalizzati in R I

- A causa della varietà delle opzioni, i modelli lineari generalizzati sono in generale più complicati da stimare rispetto ai modelli di regressione lineare e logistica.
- Il punto di partenza in R è la funzione `glm()`, che rappresenta una generalizzazione della funzione `lm()` per la stima dei modelli lineari e che viene usata sia per la stima del modello di regressione logistico che per tutti gli altri modelli che studieremo im questa sede.
- Possiamo usare `glm()` direttamente per la stima del modello logistico-binomiale, probit, e per il modello di regressione di Poisson e anche per correggere, quando necessario, per la sovradispersione.
- Modelli di regressione logistici e modelli probit ordinati possono essere stimati anche usando la funzione `polr()`, modelli probit non ordinati usando il pacchetto, sempre in R, `mnp`