Regressione per dati di conteggio

(Regressione di Poisson e altro)

L. Egidi

Autunno 2021

Università di Trieste
Corso di laurea magistrale in Scienze Statistiche ed Attuariali

Modelli per dati di conteggio

Regressione di Poisson

Oltre la regressione di Poisson

Modelli per dati di conteggio

Introduzione

Un caso di interesse è quello in cui la risposta y è un conteggio.
La variabile y può assumere pertanto i valori 0, 1, 2, ...

Le variabili di conteggio sono estremamente rilevanti per le applicazioni attuariali, ad esempio sono alla base dei modelli per studiare la frequenza dei sinistri.
Esempi rilevanti di variabili dipendenti di conteggio sono:

- il numero dei sinistri
- il numero di coloro che vanno in pensione
- il numero di coloro che riscattano una polizza
- il numero di clienti che contattano il call center o accedono al portale web di un'impresa
- il numero di coloro che hanno una determinata malattia

Di solito i conteggi rilevati si riferiscono a eventi che occorrono in un determinato intervallo di tempo (o di spazio)

Regressione per dati di conteggio

- Anche in questo caso, si vuole costruire un modello statistico il cui obiettivo è prevedere (spiegare) il numero medio di eventi
- Si cercherà di valutare in che modo la distribuzione dei conteggi y_i, caratterizzata dalla sua media μ_i, dipenda dalle caratteristiche osservate della i-ma unità (come, ad esempio, il sesso, l'età, il livello di istruzione, il reddito, etc..)
- In sostanza il modello cercherà di valutare il legame tra tali variabili esplicative e il numero medio di conteggi
- Si può scegliere fra vari modelli distributivi per una variabile di conteggio
- Il più semplice, e forse il più noto, modello stocastico per la distribuzione di una variabile di conteggio è quello di Poisson

Distribuzione di Poisson

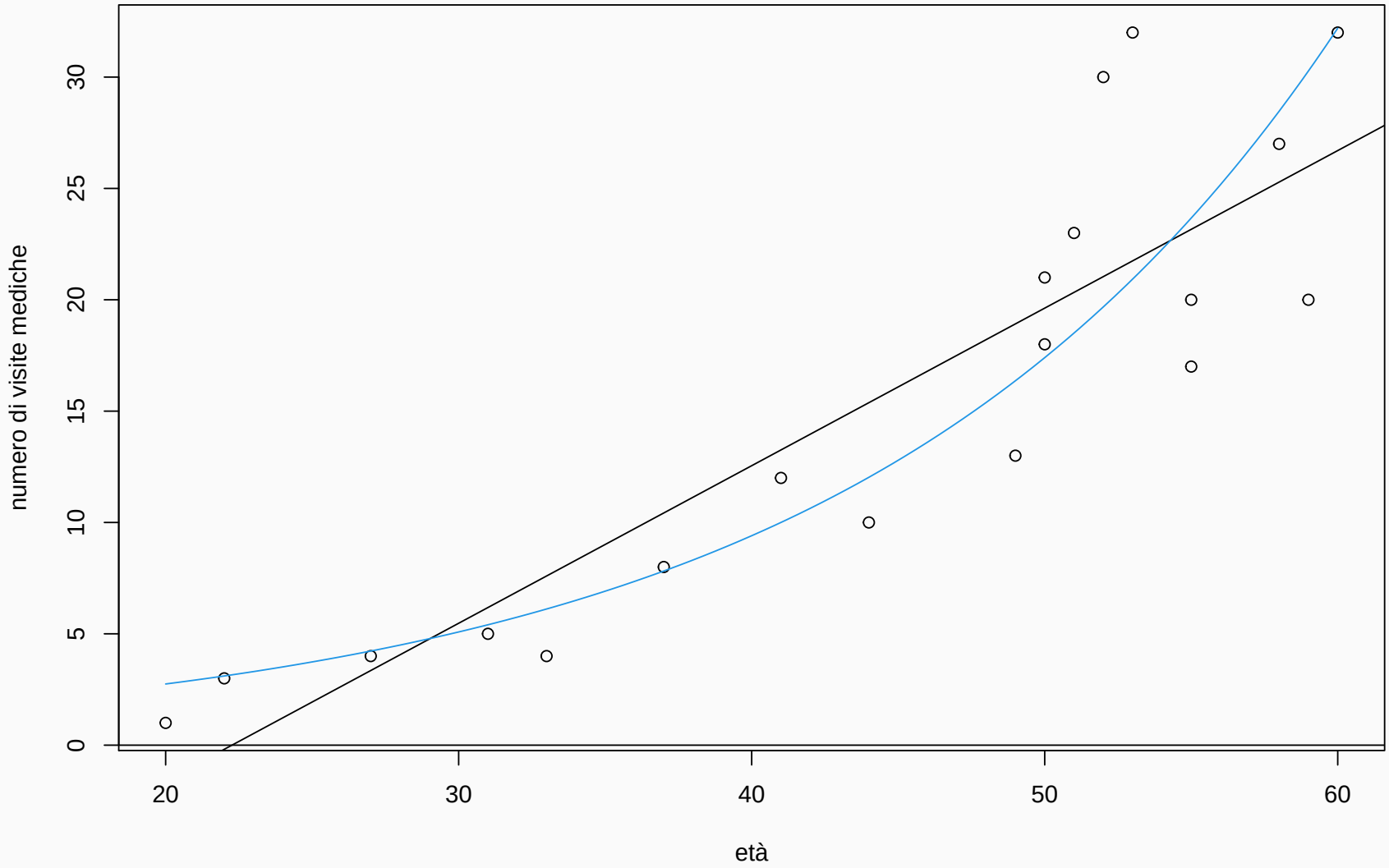
La funzione di probabilità per una variabile aleatoria Y di Poisson definita per $y = 0, 1, 2, 3, \ldots$ è

$$Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

- μ è il valore atteso cioè $E(Y) = \mu > 0$
- inoltre si ricorda che $V(Y) = E(Y) = \mu$

Regressione di Poisson

Un primo esempio: numero di visite mediche ed età



- Una regressione lineare si rivelerebbe inappropriata (anche perchè predice valori negativi)
- La curva blu fornisce una approssimazione molto migliore

Regressione di Poisson

Nel grafico sopra la curva blu, come in altri modelli di regressione, rappresenta il valore di μ_i per ogni valore della variabile esplicativa *età*. Assumiamo che la funzione $r(\cdot)$ tale che $\mu_i = r(\beta_0 + \beta_1 x_i)$ sia l'esponenziale (che assicura che μ_i possa assumere solo valori positivi):

$$\mu_i = E(y_i) = e^{\beta_0 + \beta_1 x_i}$$

Generalizzando quanto visto in questo esempio, per costruire un modello di regressione di Poisson:

1. assumiamo che che alla età x_i il numero di visite mediche abbia distribuzione di Poisson di parametro (media) μ_i. Ovvero $y_i \sim Pois(\mu_i)$
2. definiamo il **predittore lineare**
$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^T \beta$$
per un vettore $\boldsymbol{x}_i$ di variabili esplicative
3. assumiamo che
$$E(y_i) = \mu_i = r(\eta_i) = r(\boldsymbol{x}_i^T \beta)$$
con una opportuna funzione $r(\cdot)$, continua e invertibile. Pertanto è definita pure la **funzione legame** $g(\cdot) = r^{-1}(\cdot)$ e si ha equivalentemente
$$g(E(y_i)) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \beta$$

Regressione di Poisson: interpretazione dei parametri

- Se la funzione risposta è l'esponenziale questo agevola l'interpretazione dei parametri.
- Per il modello dell'esempio il parametro di interesse sarebbe β_1 associato alla covariata *età*.
- β_1 può essere interpretato come la variazione proporzionale nella media del numero di visite che corrisponde a un cambio unitario nell'*età*. Moltiplicato per 100 si può interpretare come la variazione percentuale in y per un aumento unitario di x.
- Nell'esempio si era ottenuto $\beta_0 = -0.220$ e $\beta_1 = 0.062$. Il modello prevede quindi che se si incrementa l'età di un anno il numero di visite mediche aumenta del 6.2%.

Regressione di Poisson: stima dei parametri

- L'ipotesi distributiva Poisson per y_i consente di usare il metodo della massima-verosimiglianza.
- La log-verosimiglianza, assumendo l'indipendenza delle osservazioni, risulta:

$$\log(L(\beta)) = \ell(\beta) = \sum_{i=1}^{n}(y_i \log(\mu_i) - \mu_i)$$

  in cui si è omessa la costante $-n\log(y_i!)$ che non dipende da β
- Utilizzando la funzione legame $\log(\mu_i) = \eta_i$ si ha

$$\ell(\beta) = \sum_{i=1}^{n} \ell_i(\beta) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \beta - \exp(\boldsymbol{x}_i^T \beta) = \sum_{i=1}^{n}(y_i \eta_i - \exp(\eta_i))$$

  Si può quindi calcolare la funzione score e porla pari a 0

$$s(\beta) = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \exp(\eta_i)) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \mu_i) \quad \text{e si ponga poi} \quad s(\beta) = 0$$

  è agevole poi ottenere l'informazione di Fisher che è pari a
  $i(\beta) = E(s(\beta)s(\beta)^T) = \sum_{1=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \mu_i$
- La soluzione delle equazioni di verosimiglianza non è diretta e richiede l'uso di algoritmi numerici (ad es. Newton-Raphson)

Inferenza nella regressione di Poisson: verifica sui singoli parametri

In virtù delle usuali assunzioni di regolarità che sono valide nel caso Poisson, si ha che asintoticamente

$$\hat{\beta} \dot{\sim} N(\beta, I(\beta)^{-1}) \qquad \text{con matrice var-covar stimata da} \qquad I(\hat{\beta})^{-1}$$

Siamo interessati al solito a:

- verificare ipotesi sui singoli β_j. Del tipo $H_0 : \beta_j = 0$ contro $H_1 : \beta_j \neq 0$
- verificare la bontà del modello complessiva
- confrontare modelli nidificati

Possiamo quindi valutare se un dato β_j associato alla covariata x_j è grande abbastanza guardando a $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$. Se il valore assoluto di tale rapporto è grande allora x_j ha un significativo impatto sulla media di y.

Sotto l'ipotesi nulla che il parametro sia pari a 0 tale rapporto è, per n sufficientmente grande, approssimabile con una gaussiana standard.

Per un giudizio veloce si guarda se il valore assoluto del rapporto è maggiore di 2. Per una valutazione più precisa si guarda al valore p associato.

Inferenza nella regressione di Poisson: performance complessiva del modello

Si può misurare la differenza fra il valore della verosimiglianza nel modello completo $L_C = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ e

- il modello nullo, che contiene solo l'intercetta $L_0 = L(\hat{\beta}_0)$,
- oppure un modello alternativo a quello stimato ottenuto ponendo a 0 alcuni parametri (modello ridotto)
  $L_R = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}, 0, \ldots, 0, 0)$.
  Nell'ultima espressione $p - k$ parametri sono fissati pari a 0
- Il confronto fra tali verosimiglianze (o log-verosimiglianze) ci può aiutare a scegliere il modello
- La differenza fra $\ell_C = \log(L_C)$ e $\ell_0 = \log(L_0)$ ci fornisce una prima indicazione: se essa è piccola allora il modello completo non è supportato dai dati

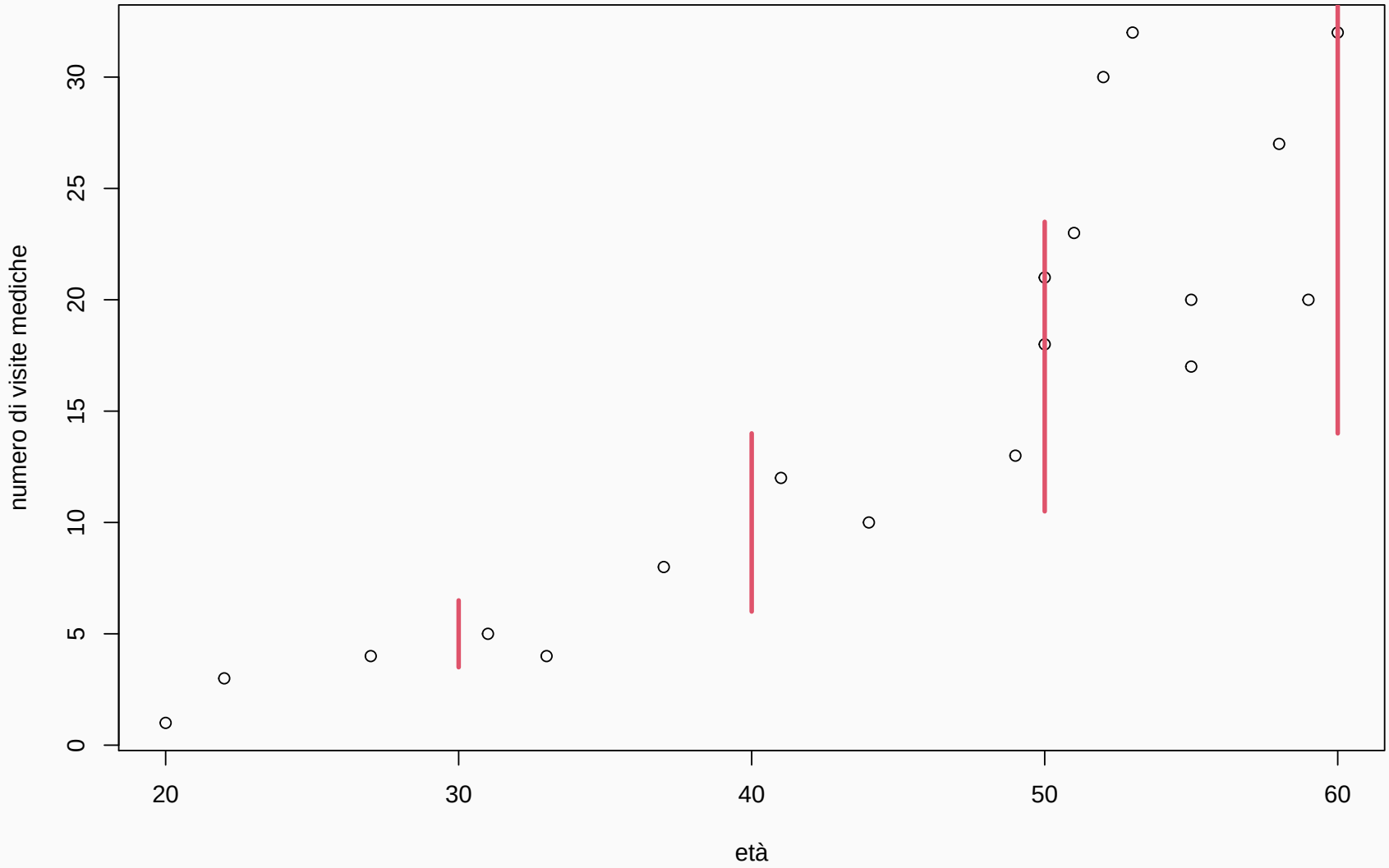
Inferenza nella regressione di Poisson: performance complessiva del modello

- Lo stesso vale per il confronto fra $\ell_C = \log(L_C)$ e $\ell_R = \log(L_R)$, se la differenza è piccola allora si preferisce il modello più parsimonioso (con meno parametri)
- Per decidere se la differenza che otteniamo è piccola si può considerare la statistica $W = 2(\ell_C - \ell_R)$ che deriva da un test del rapporto di verosimiglianza e, sotto l'ipotesi di nullità per i parametri aggiuntivi nel modello completo, si distribuisce asintoticamente come un χ^2 con $p - k$ g.d.l.
- Possiamo quindi calcolare il p-value $p = Pr(W \geq W_{oss})$ dove $W \dot{\sim} \chi^2$ con $p - k$ g.d.l.. Se questo valore è molto piccolo i dati indicano che il modello con la restrizione di nullità è poco plausibile
- Si può in alternativa usare il crierio di Akaike

$$AIC = -2\ell_c + 2p$$

per confrontare anche modelli non nidificati. Si sceglie il modello per cui AIC è più piccolo

Ancora sull'esempio delle visite mediche



Le linee rosse verticali illustrano il fatto che il numero di visite mediche si presenta più disperso al crescere della media. Il modello tiene conto di questo?

Regressione di Poisson: la sovradispersione

- Una caratteristica saliente della distribuzione di Poisson è che la sua media è pari alla sua varianza
- Il modello di regressione di Poisson riguarda la media ma impone anche che la varianza di y_i vari di conseguenza: implicitamente si introduce una forma di eteroschedasticità
- In molti casi questo riduce la flessibilità del modello e la sua capacità di descrivere i dati osservati
- Un semplice modo per controllare l'appropriatezza del modello è quindi quello di verificare se i dati riflettono la richiesta

$$\text{media}(Y_i) = \text{varianza}(Y_i)$$

Regression di Poisson: controllo dei residui

- Possiamo calcolare i residui di una regressione di Poisson analogamente a quanto si fa per una regressione normale
- Possiamo dapprima ottenere la previsioni $\hat{\mu}_i$ attraverso il modello stimato. Più precisamente:

$$\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{p-1} x_{ip-1}}$$

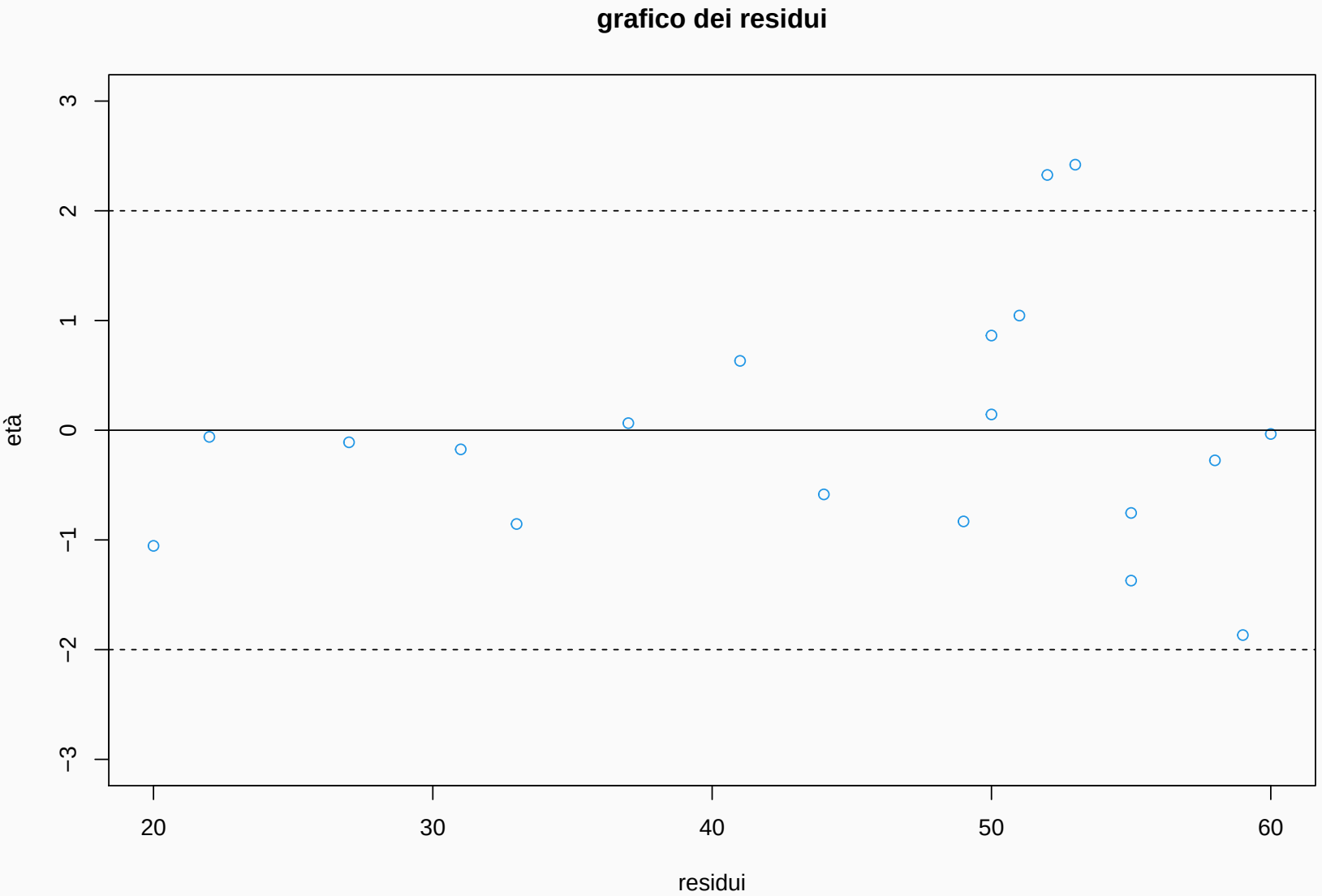
Si possono quindi ottenere i residui misurando le differenze fra i valori medi previsti e i valori osservati e dividere questi per la deviazione standard stimata (così da renderli comparabili e standardizzati: il modello è eteroschedastico)

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

Regression di Poisson: controllo dei residui

- il controllo dei residui può essere utile a verificare se le assunzioni fatte sono ragionevoli
- i residui dovrebbero avere valori che si muovono attorno allo 0, dovrebbero essere non troppo elevati (li abbiamo standardizzati e la loro varianza è unitaria) e non mostrare nessun pattern se guardati in relazione ai valori previsti di y_i
- se il campione è sufficientemente elevato (anche n=50 può bastare) essi dovrebbero comportarsi come valori estratti dalla gaussiana standard. La gran parte dei residui dovrebbe quindi stare fra -2 e 2 (circa il 95%).
- se invece osserviamo un gran numero di residui che hanno valore assoluto anche di molto superiore a 2 questo è un sintomo che in realtà per i dati media$(Y_i) <$ varianza(Y_i)
- questa situazione è detta di sovradispersione e indica che il modello di Poisson potrebbe non essere appropriato
- si noti che lo stesso controllo potrebbe portarci a diagnosticare sottodispersione (meno frequente nella pratica)

Media e varianza: dati sulle visite mediche



Si notino i due residui fuori della bande $[-2, +2]$

Regressione di Poisson: tenere conto dell'esposizione (un modello per i tassi)

- Il modello di base per la regressione di Poisson può esser esteso per tenere conto del fatto che i conteggi sono ottenuti in condizioni differenti per le diverse unità
- se volessimo costruire un modello per il numero y_i di incidenti avvenuti nella stessa strada dovremmo tenere conto del fatto che a parità di altre condizioni potrei variare il periodo di osservazione per gli incidenti.
- Il tempo e_i entro il quale si osserverà il numero di incidenti costituisce una variabile di esposizione ed è necessario tenerne conto nel modello.
- Potrebbe ad esempio esser sensato fare un modello per i tassi invece che per i conteggi, cioè potremmo scrivere che, definito $\mu_i = E(y_i)$,

$$\log\left(\frac{\mu_i}{e_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{ip-1} x_{ip-1}$$

- ma ciò equivale a porre

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{ip-1} x_{ip-1} + \log(e_i)$$

- quindi questo equivale a inserire la variabile aggiuntiva $\log(e_i)$ nel modello di Poisson fra le covariate ma imponendo che il suo coefficiente sia pari a 1
- La covariata speciale $\log(e_i)$ è detta **offset**

Oltre la regressione di Poisson

Oltre la regressione di Poisson

- Il modello di Poisson è senza dubbio il più semplice per modellare conteggi
- Tuttavia le variabili aleatorie di Poisson non sono le uniche che possono essere usate per descrivere una variabile di conteggio
- Modelli leggermenti più complessi, per esempio il modello Binomiale Negativo può rivelarsi più flessibile per trattare situazioni (ad esempio la sovradispersione) che si incontrano di frequente nella pratica
- I modelli di Poisson possono esser poi semplificati per fronteggiare situazioni non standard e dati con parttern più complessi. Ad esempio:
  - il caso di conteggi in cui non ci sarà mai la possibilità di osservare 0 (Poisson troncata)
  - il caso in cui il conteggio si ha solo per una porzione del campione perchè per i restanti si sa che solo il valore 0 è possibile (modelli con inflazione di zeri)
- Modelli più complessi sono inoltre necessari per tenere conto di altre situazioni in cui le condizioni teoriche che danno luogo a conteggi compatibili con la Poisson non sono presenti

Modelli per dati sovradispersi

- Il modello di Poisson può rivelarsi troppo restrittivo (in particolare a causa della relazione $E(Y_i) = Var(Y_i)$
- Spesso i dati riveleranno che la varianza osservata è superiore a quella implicata dalla Poisson (*sovradispersione*) oppure, meno frequentemente, essa è inferiore alla media (*sottodispersione*)

Uso della quasi verosimiglianza

- Si assuma che come per un modello di Poisson sia

$$E(y_i) = \mu_i = \exp(\boldsymbol{x}_i^T \beta)$$

- ma che $Var(y_i) = \phi\mu_i$ ove $\phi > 0$ è un parametro che consente di considerare il caso di sovradispersione ($\phi > 1$) o sottodispersione ($\phi < 1$)
- Se $\phi \neq 1$ allora il modello stocastico per y_i non è Poisson
- E' ancora possibile usare la stessa equazione di stima ottenuta attraverso il metodo della massima verosimiglianza

Metodo della quasi verosimiglianza

Le soluzioni delle equazioni

$$\sum_{i=1}^{n}(y_i - \exp(\boldsymbol{x}_i^T \beta))x_{ij} = 0 \qquad \forall j$$

forniscono ancora stimatori consistenti di β se la media è correttamente specificata e se i dati sono supposti incorrelati.

- In questo caso però si parlerà di stimatori di quasi-verosimiglianza, in quanto l'equazione non è ricavata facendo riferimento a un modello distributivo.
- Tuttavia è necessario correggere gli errori standard delle stime per tenere conto della varianza specificata come $Var(y_i) = \phi\mu_i$
- Il parametro ϕ può essere stimato per esempio da

$$\hat{\phi} = \frac{1}{n-p}\sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Metodo della quasi verosimiglianza

- quindi le varianze sono stimate da $\widehat{Var(\beta_j)} = \hat{\phi} I(\hat{\boldsymbol{\beta}})^{-1}_{jj}$ ove $I(\hat{\boldsymbol{\beta}})$ è la matrice di informazione osservata
- in `R` è possibile stimare tale modello attraverso la quasi verosimiglianza specificando la famiglia `quasipoisson`
- occorre porre attenzione anche nell'uso della devianza che, come vedremo, è anch'essa inflazionata se non teniamo conto di ϕ

Regressione Binomiale negativa

- Si tratta di un modello ampiamente utilizzato per una variabile dipendente che è un conteggio

$$Pr(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k} \qquad z = k, k+1, \ldots$$

con $E(Z) = k(1-p)/p$ e $Var(Z) = k(1-p)/p^2$

- Tale distribuzione ha alcuni vantaggi se comparata con la Poisson:
  - ha due parametri e quindi maggiore flessibilità nella modellizzazione dei dati
  - la varianza è superiore rispetto alla media e quindi è adeguata nel caso di sovradispersione
  - Si può verificare che la Poisson è un caso limite della binomiale negativa (se $p \to 1$ e $k \to 0$ così che $kp \to \lambda$)
- Si può dimostrare che la binomiale negativa emerge come distribuzione di miscugli di Poisson. Questo nel caso che per ogni unità y_i sia una Poisson di media μ_i e si assume che il parametro μ_i sia determinazione di una distribuzione *Gamma*.

Regressione Binomiale negativa

- Se vogliamo costruire un modello per la Binomiale negativa conviene adottare una parametrizzazione diversa ponendo $Y = Z - k$ e $p = \frac{1}{1+\alpha}$
- ottenendo

$$Pr(Y = y) = \binom{y + k - 1}{k - 1} \frac{\alpha^y}{(1 + \alpha)^{y+k}} \qquad y = 0, 1, \ldots$$

- Risulta che
  - $E(Y) = \mu = k\alpha$
  - e che $Var(Y) = k\alpha + k\alpha^2 = \mu + \mu^2/k$
- Possiamo utilizzare la seguente funzione legame $\log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{k+\mu}$

Modelli per Poisson troncate

- Immaginiamo che la variabile di interesse sia il numero di polizze possedute dai clienti di una compagnia assicurativa
- In questo caso si tratta pur sempre di un conteggio ma la distribuzione di Poisson sarebbe inappropriata perché i clienti avranno un numero di polizze che non è mai pari a 0
- In questo caso si potrebbe immaginare che i conteggi y_i sono ancora Poisson ma condizionatamente al fatto che essi siano positivi

$$p(y_i|y_i > 0) = \frac{e^{-\mu_i}\mu_i^{y_i}}{(1 - e^{-\mu_i})y_i!} \qquad y_i = 1, 2, \ldots$$

- E possibile seguire le stesse idee utilizzate per la regressione di Poisson
  - definiamo il predittore lineare $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$
  - usiamo una funzione legame opportuna $\log(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$
- È possibile scrivere le equazioni di verosimiglianza e reperire le stime di massima verosimiglianza (mediante opportuni algoritmi numerici)
- Concettualmente un analogo discorso può esser fatto per la Binomiale Negativa troncata in 0

Modelli di Poisson con inflazione di zeri

- In alcune situazioni applicative emerge chiaramente che il motivo dell'inadeguatezza di un modello è la eccessiva presenza di valori pari a 0 nella variabile y_i
- Se ad esempio volessimo costruire un modello per predire il numero di articoli acquistati su Amazon da coloro che entrano nel sito troverei che nella stragrande maggioranza dei casi il valore è 0. E il numero di zeri potrebbe essere eccessivamente alto per giustificare un modello di Poisson (o Binomiale negativo).
- Se ignoriamo l'inflazione di zeri ci sono conseguenze:
  - potrebbero esserci distorsioni nelle stime dei parametri e degli errori standard (in fondo usiamo un modello inadeguto)
  - l'eccessivo numero di zeri dà luogo a sovradispersione
- Tuttavia il motivo per l'eccesso di zeri potrebbe esser legato al fatto che in realtà noi osserviamo il miscuglio di due popolazioni:
  - una popolazione per la quale y_i è una variabile degenere pari a 0
  - una popolazione per la quale invece y_i segue una legge di Poisson (o anche binomiale negativa)

Verso un modello mistura per l'inflazione di zeri

- Nell'esempio richiamato sopra è come se si ipotizzasse che vi siano 2 tipi di utenti di Amazon:
  1. quelli che vanno sul sito, guardano gli articoli ma che non comprano mai su Amazon (per essi y_i è pari a 0)
  2. quelli che comprano un numero di articoli y_i che si distribuisce, ad esempio, secondo una legge di Poisson
- Assumiamo che sia π_i la probabilità di appartenere alla prima popolazione per l'individuo i-esimo.
- Nei nostri dati osserveremmo quindi più valori di y_i pari a 0 in quanto agli zeri generati dalla Poisson si aggiungono quelli per coloro che non comprano mai, pertanto avrò la seguente *distribuzione mistura*:

$$p(y_i) = \pi_i \mathrm{I}_{y_i=0} + (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},$$

che possiamo anche riscrivere come:

- $p(y_i = 0) = \pi_i + (1-\pi_i)e^{-\mu_i}$
- $p(y_i = k) = (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},\ k = 1, 2, \ldots$

Modelli con inflazione di zeri

Il modello riguarda la variabile risposta y_i la cui distribuzione è

$$p(y_i = 0) = \pi_i + (1 - \pi_i)e^{-\mu_i}$$

$$p(y_i = k) = (1 - \pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \ k = 1, 2, \ldots,$$

e che ha $E(y_i) = \mu_i(1 - \pi_i)$ $\quad$ $Var(y_i) = (1 - \pi_i)(\mu_i + \pi_i\mu_i^2)$

- possiamo poi considerare delle variabili esplicative $\boldsymbol{x}_i$ e considerare
  - $\eta_i = \boldsymbol{x}_i^T\beta$ e la funzione legame usuale $\log(\mu_i) = \eta_i$
- Inoltre se si ipotizza che la probabilità che una unità appartenga o meno a una delle due sottopopolazioni dipende da un vettore di esplicative $\boldsymbol{z}_i$ (che potrebbe anche coincidere con $\boldsymbol{x}_i$), allora si può completare il modello con:
  - $\tau_i = \boldsymbol{z}_i^T\boldsymbol{\delta}$ e
  - $\log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{z}_i^T\boldsymbol{\delta}$
- E' possibile ricavare la verosimiglianza del modello e quindi ottenere le stime dei parametri $(\boldsymbol{\delta}, \beta)$
- Un modello analogo si può assumere utilizzando la Binomiale Negativa per Y.
- E' importante notare che i modelli Binomiale negativo, quelli con troncamento in zero e quelli con inflazione di zeri non sono modelli della classe dei GLM per cui occorrerà per essi usare specifici packages e funzioni.