Regresión de Poisson

Sirve para modelar datos de tipo de conteo que no están dados en forma de proporciones.

Casos típicos de datos de Poisson o que provienen de un proceso tipo Poisson en los que el límite superior de ocurrencias es infinito se encuentran en la práctica, por ej.:.

1. el número de partículas radioactivas emitidas en un intervalo de tiempo
2. en estudios de comportamiento el número de incidentes en intervalos de longitud especificada

El modelo Poisson asume que

$$E(Y_i) = Var(Y_i) = \mu_i$$

y como ya hemos mencionado es un supuesto que puede ser restrictivo, pues con frecuencia los datos reales exhiben una variación mayor que la que permite este modelo.

Regresión de Poisson

$$Y_i \sim P(\mu_i), \quad i = 1, \ldots, n$$

Queremos relacionar las medias μ_i con covariables $\mathbf{x}_i$.

Recordemos que si $Y \sim P(\mu)$

$$P(Y = y) = e^{-\mu} \frac{\mu^y}{y!} = \exp\left(y \log \mu - \mu - \log y!\right)$$

por lo tanto

$$\begin{aligned} \theta &= \log \mu \quad b(\theta) = e^{\theta} \\ \phi &= 1 \qquad a(\phi) = 1 \\ c(y, \phi) &= -\log y! \end{aligned}$$

Link natural es $\eta = \log \mu$, que asegura que el valor predicho de μ será no negativo.

Ajuste del modelo Cuando se usa el link log tenemos, para cada observacón se postula

$$\log(\mu_i) = \mathbf{x}_i^t \boldsymbol{\beta}$$

En este caso, Newton–Raphson y Fisher–scoring coinciden. Mediante el algoritmo iterativo calculamos:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$$

donde

$$\mathbf{W} \quad = \quad diag(\mu_i)$$

y la variable de trabajo

$$\mathbf{z} = \boldsymbol{\eta} + \left(\frac{\partial \eta}{\partial \mu}\right)(\mathbf{y} - \boldsymbol{\mu})$$

y $\dfrac{\partial \eta}{\partial \mu} = \dfrac{1}{\mu}$

Después de la estimación

El logaritmo de la función de verosimilitud es, salvo constantes,

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

Si usamos el link log, entonces $\log \mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ y la **deviance** queda

$$D = 2\sum_{i=1}^{n}\left(y_i \log \frac{y_i}{\mu_i} - (y_i - \mu_i)\right)$$

Después de la estimación

Cuando el modelo tiene intercept, resulta

$$\log \mu_i = \beta_0 + \sum_{i=1}^{p} x_{ij}\beta_j, i = 1, \ldots, n$$

$$\frac{\partial D}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \mu_i) \, .$$

Si consideramos los valores predichos con el estimador de máxima verosimilitud, $\widehat{\mu}_i$

$$\sum_{i=1}^{n} \widehat{\mu}_i = \sum_{i=1}^{n} y_i$$

y por lo tanto la deviance se simplifica a :

$$D = 2 \sum_{i=1}^{n} y_i \, \log \frac{y_i}{\mu_i} \, .$$

Residuos

En general se definen los siguientes residuos:

- Residuos de Respuesta: $Y_i - \widehat{\mu}_i$
- Residuos de Pearson: $\frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}$

  Tienen media 0 y varianza constante si la función de varianza es la correcta.
- Residuos de Trabajo: $(Y_i - \widehat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}$
- Residuos de Deviance: $sgn(Y_i - \widehat{\mu}_i)\sqrt{d_i}$ siendo d_i componente deviance

  Generalmente, más normales que los de Pearson, es decir menos asimétricos. Utiles para visualizar outliers.

Residuos Poisson

Residuos deviance:

$$r_i^d = sg(y_i - \widehat{\mu}_i)\left\{2(y_i \log(y_i/\widehat{\mu}_i) - y_i + \widehat{\mu}_i)\right\}^{1/2}$$

Residuos de Pearson:

$$r_i^p = \frac{y - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$$

Offset

En el caso de la regresión Poisson es frecuente que aparezca una covariable en el predictor lineal cuyo coeficiente no es estimado pues se asume como 1: esta variable es conocida como **offset**.

Supongamos que tenemos $y_1, y_2, \ldots, y_n$ variables independientes que corresponden al número de eventos observados entre n_i expuestos (*exposure*)para la el $i-$ésimo valor de la covariable. Por ejemplo, y_i es el número de reclamos de seguro de autos de una determinada marca y año. El valor esperado de y_i puede escribirse como

$$\mu_i = E(y_i) = n_i \lambda_i \, ,$$

es decir que depende del número de autos asegurados y la tasa media de reclamos.

Offset

Podríamos creer que es λ_i, y no μ_i, quien depende de variables tales como años del auto y lugar donde se usa. Bajo un modelo con link log tenemos que

$$\log \mu_i = \log\, n_i + \mathbf{x}_i'\boldsymbol{\beta} = o_i + \mathbf{x}_i'\boldsymbol{\beta}\,,$$

donde o_i recibe el nombre de offset.

En `R` lo especificaríamos así

```
glm(y modelo+lugar, offset = cantautos, family = poisson)
```

Función de Varianza

Este modelo asume que

$$E(Y_i) = Var(Y_i) = \mu_i$$

sin embargo es posible que un conjunto de datos tengan una dispersión mayor.

Cuando los datos exhiben sobredispersión, se puede tomar uno de los siguientes caminos:

1. Suponer que $Var(Y_i) = \sigma^2 \mu_i$ y estimar σ^2 usando un modelo de quasi–verosimilitud.
2. Sumergir a la variable de respuesta en una familia de distribuciones que contemple una dispersión mayor: *Binomial Negativa*

Binomial Negativa

Si

$$\begin{aligned} Y|\lambda &\sim P(\lambda) \\ \lambda &\sim \Gamma(\alpha, \beta) \end{aligned}$$

donde

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\,\lambda^{\alpha-1}e^{-\lambda/\beta}I_{[0,\infty)}(\lambda)\,,$$

entonces

$$Y \quad : \quad P(Y=y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\,y!}\left(\frac{\beta}{1+\beta}\right)^{y}\left(\frac{1}{1+\beta}\right)^{\alpha}$$

Binomial Negativa

La media y la varianza de Y son:

$$E(Y) = E(E(Y|\lambda)) = E(\lambda) = \alpha\,\beta$$

$$\begin{aligned} Var(Y) &= E(Var(Y|\lambda)) + Var(E(Y|\lambda)) \\ &= Var(\lambda) + E(\lambda) = \alpha\,\beta + \alpha\,\beta^2 \end{aligned}$$

La distribución **BN** suele parametrizarse en términos de $\mu = \alpha\beta$ y $\kappa = 1/\alpha$ como

$$P(Y = y) = \frac{\Gamma(\kappa^{-1} + y)}{\Gamma(\kappa^{-1})\, y!} \left(\frac{\kappa\mu}{1 + \kappa\mu}\right)^{y} \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa}.$$

Binomial Negativa

En este caso, diremos que $Y \sim BN(\mu, \kappa)$. Con esta parametrización resulta

$$\begin{aligned} E(Y) &= \mu \\ Var(Y) &= \mu + \kappa\mu^2 \,, \end{aligned}$$

por lo tanto, en una BN la varianza es mayor que la media.

Esto nos sugiere que si sospechamos que hay subdispersión deberíamos elegir el camino de quasi–verosimilitud, pues la BN no puede tratar este problema.

¿Cómo ajustamos una distribución BN?

Salvo constantes el log–likelihood resulta

$$\ell = \log \Gamma(\kappa^{-1} + y) - \log y! + y \log \left(\frac{\kappa\mu}{1+\kappa\mu} \right) + \kappa^{-1} \log \left(\frac{1}{1+\kappa\mu} \right)$$

Para κ fijo, esta distribución pertenece a una familia exponencial a un parámetro con

$$\theta = \log \left(\frac{\kappa\mu}{1+\kappa\mu} \right) .$$

Si κ es conocido, se puede computar el estimador de $\boldsymbol{\beta}$ mediante un procedimiento iterativo. El problema es que en general κ es desconocido y por lo tanto, se debe estimar en forma simultánea ambos parámetros.

En `R`

```
library(MASS)
glm.nb(y~ x1+x2,data=misdatos)
```