Regresión de Poisson

La regresión de Poisson es una de las aplicaciones más importantes de GLM.

En este caso estamos interesados en datos de tipo de conteo que no están dados en forma de proporciones. Ejemplos 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 ejemplo, el número de partículas radioactivas emitidas en un intervalo de tiempo o en estudios de comportamiento el número de incidentes en intervalos de longitud especificada.

Aún en los estudios más cuidados puede haber apartamientos al modelo Poisson. Por ejemplo, un contador Geiger tiene un *dead–time* después de la llegada de una partícula, éste es un lapso durante el cual no puede detectar más partículas. Luego cuando la tasa de emisión de partículas es alta, el efecto de *dead–time* lleva a apartamientos notables del modelo de Poisson para el número de ocurrencias registradas. Así también, por ejemplo, si estamos realizando un estudio de la conducta de un chimpancé y contamos el número de ocurrencias

de cierto evento es factible que estas se registren en grupos.

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.

Asumiremos que

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

y como siempre deseamos relacionar las medias μ_i con covariables $\mathbf{x}_i$.

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

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

por lo tanto

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

Luego, el link natural es $\eta = \log \mu$, que asegura que el valor predicho de μ_i será no negativo. Cuando se utiliza en el modelo Poisson este link suele llamárselo *modelo loglineal*, sin embargo esta denominación, como veremos, se utiliza en el contexto de tablas de contingencia.

Ajuste del modelo

Cuando se usa el link log 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

$$\mathbf{W} = diag\left(V_i^{-1}\ \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$$

y la variable de trabajo

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

¿Qué resulta en el caso en que $\eta = \log \mu$?

Como $\dfrac{\partial \eta}{\partial \mu} = \dfrac{1}{\mu}$, resulta

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

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}\,.$$

Después de la estimación

Salvo constantes, tenemos que el logaritmo de la función de verosimilitud es

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

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

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

Notemos que si el modelo tiene intercept,

$$\log \mu_i = \beta_1 + \sum_{i=2}^{p} x_{ij}\beta_j,\, i = 1, \ldots, n$$

$$\frac{\partial D}{\partial \beta_1} = \sum_{i=1}^{n} (Y_i - \mu_i) \, .$$

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 :

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

Podemos definir los residuos deviance como:

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

y los residuos de Pearson como:

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

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 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. 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.

Por ejemplo, si Y_i es el número de muertes por cancer en el año 2001 en una determinada población, parece razonable ajustar por el tamaño de la población.

Ejemplo: Médicos Ingleses: fumadores y muerte coronaria (Annette Dobson (1990))

edad	smoke	y	pop
1	1	32	52407
2	1	104	43248
3	1	206	28612
4	1	186	12663
5	1	102	5317
1	0	2	18790
2	0	12	10673
3	0	28	5710
4	0	28	2585
5	0	31	1462

```
plot(edad,(y/pop)*100000,type="n") text(edad,(y/pop)*100000,
c("*","o")[factor(smoke)]) logpop<-log(pop) edad2<- edad*edad
smkage<- edad*smoke dmat<- cbind(smoke,edad,edad2,smkage,rep(1,10))
```

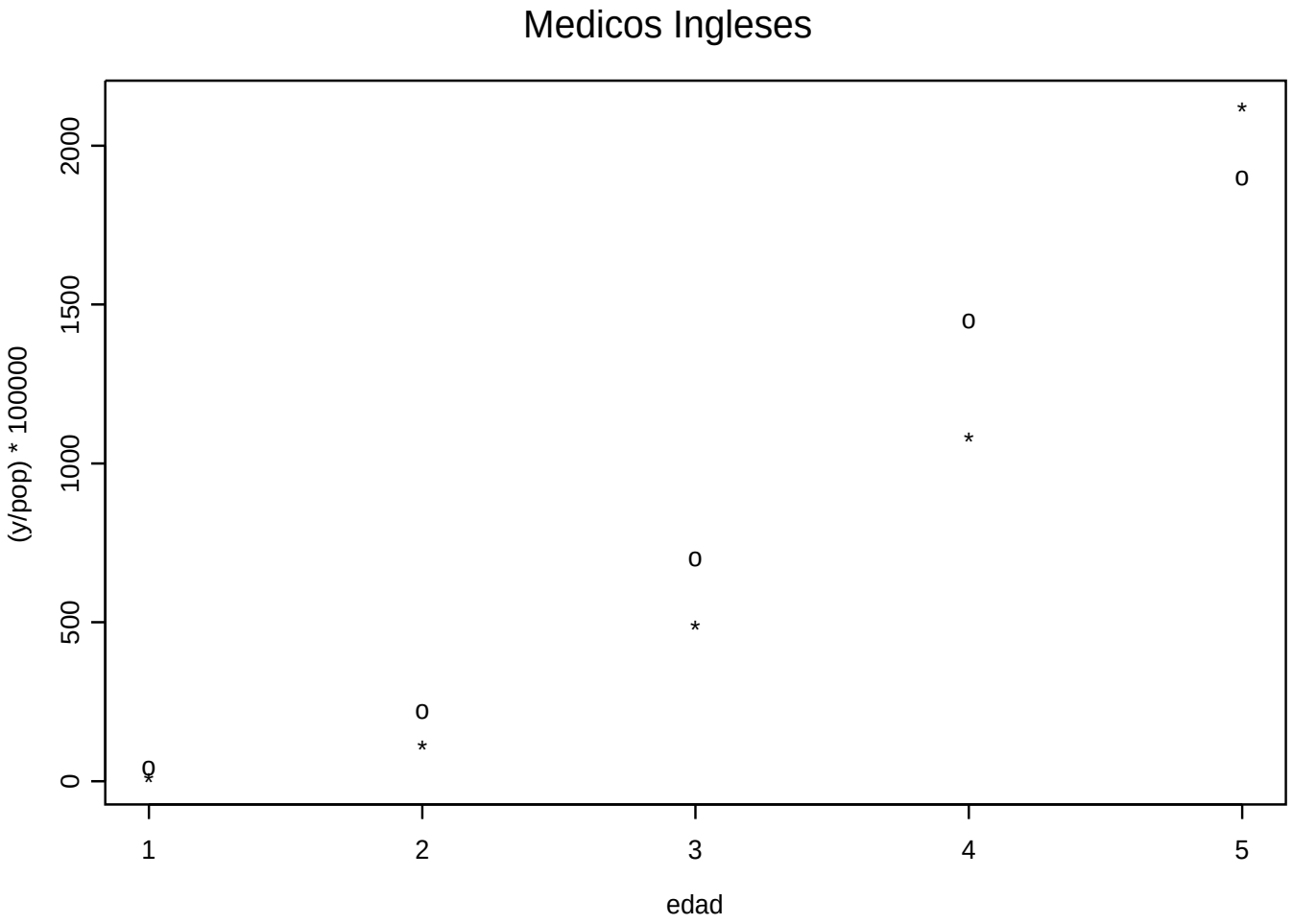

Figura 1: Medicos Ingleses

```
summary.glm(sal)

Call: glm.fit(x = dmat, y = y, offset = logpop, family = poisson)

Coefficients:
             Value Std. Error    t value
 smoke   1.4409718 0.37216161   3.871898
  edad   2.3764783 0.20793739  11.428817
 edad2  -0.1976765 0.02736679  -7.223228
smkage  -0.3075481 0.09703401  -3.169487
       -10.7917624 0.45003224 -23.979976

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: on 9 degrees of freedom

Residual Deviance: 1.63537 on 5 degrees of freedom
```

Término	Edad	$Edad^2$	Smoke	Smkage
Coef.	2.376	-0.198	1.441	-0.308
S.E.	0.208	0.027	0.372	0.097
Rate ratio	10.762	0.820	4.225	0.735
IC 95 %	(7.2,16.2)	(0.78,0.87)	(2.04,8.76)	(0.61,0.89)

Si x_j es una variable explicativa dicotómica, tal que $x_j = 0$ si el factor está ausente y $x_j = 1$ si el factor está presente, de manera que el **rate ratio** para presencia vs. ausencia es:

$$\text{Rate Ratio} = \frac{E(Y_i|presente)}{E(Y_i|ausente)} = e^{\beta_j}$$

si las otras variables se mantiene fijas.

De la misma forma, si x_j es continua y se incrementa en una unidad manteniendo las otras variables fijas, el Rate Ratio sería e^{β_j}.

Función de Varianza

Este modelo asume que

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

sin embargo es posible que un conjunto de datos tenga 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, como en el caso binonial.
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

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

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

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

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

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

Como ya vimos 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 el procedimiento iterativo que hemos visto. Sin embargo, el problema es que en general

κ es desconocido y por lo tanto se debe estimar en forma simultánea ambos parámetros.

R no considera la familia BN entre las alternativas de su procedimiento **glm**.

Una posibilidad es maximizar el likelihood aplicando el método de Newton–Raphson en forma conjunta para κ y $\boldsymbol{\beta}$.

Otra posibilidad es definir una grilla de valores para κ y maximizar el likelihood respecto de $\boldsymbol{\beta}$. Se puede graficar el máximo de la función de verosimilitud para identificar donde se alcanza el estimador de máxima verosimilitud de κ. Se podría comenzar con una grilla más o menos gruesa y luego refinarla en la zona más adecuada.

En el segundo método, para cada κ usaríamos el método de Fisher–scoring como hasta ahora:

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

donde

$$\mathbf{W} = diag\left(V_i^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right) = diag(w_1, \ldots, w_n)$$

y la variable de trabajo

$$z_i = \eta_i + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i)$$

Eventualmente si tuvieramos un offset quedaría:

$$z_i = \eta_i - o_i + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i)$$

Por la expresión de la varianza que obtuvimos resulta $V_i = \mu_i + \kappa\mu_i^2$ y si usamos el link log, como en el caso Poisson, resulta

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}$$

$$w_i = \frac{\mu_i^2}{\mu_i + \kappa\mu_i^2}$$

Observemos que la diferencia con la regresión Poisson está en los pesos w_i y no en la variable de trabajo. En este método de la grilla, la matriz de covarianza de $\boldsymbol{\beta}$ se estimaría mediante la fórmula habitual $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ usando $\widehat{\kappa}$ en lugar de κ. Vale la pena notar que en este caso no estamos considerando la variabilidad de la estimación de κ