1. Modelos para conteos

En esta sección se modelarán conteos, que resultan de diversas situaciones, por ejemplo: el número de accidentes en una carretera, el número de caries que tiene una persona, el número de quejas por cierto servicio, etc.

Comunmente la distribución Poisson $Po(\mu)$ se utiliza para modelarlos, cuya función de densidad es:

$$f(y) = \frac{\mu^y \exp^{-\mu}}{y!}, y = 0, 1, 2, \ldots,$$

μ representa la media de las ocurrencias del evento de interés.

Esta distribución tiene como característica que $E(y) = \mu = Var(y)$, esto significa que si la media crece (decrece) la varianza crece (decrece) también.

A menudo μ se representa a través de una tasa, por ejemplo el número promedio de consumidores que compra un producto de cada 100 que entran a la tienda; μ puede definirse de varias formas, por ejemplo pensando en en el caso de accidentes en motocicleta, ésta puede definirse como el número de accidentes por cada 1000 de población, o como el número de accidentes por cada 1000 personas tienen licencia de manejo, o como el número de accidentes por cada 1000 vehículos automotores, etc. Otro ejemplo es el de conteo de ciclones, debe especificarse la temporada, es decir el tiempo de exposición. Para evitar confusiones es muy importante indicar en cada caso claramente el nivel de **exposición**.

En general se tienen dos tipos de situaciones, la primera en la que al modelar μ debe tomarse en cuenta los distintos niveles de exposición y que las variables explicativas podrán ser categóricas o continuas y entonces es que se utilizan los modelos de **Regresión Poisson**; la segunda, en la que el nivel de exposición es el mismo para todos y las variables explicativas son categóricas, de manera que la información puede resumirse en una tabla de contingencia, en este caso se usan los modelos **Loglineales**.

1.1. Regresión Poisson

En este caso las variables $Y_1, Y_2, \ldots, Y_N$ son independientes, cada Y_i representa el número de eventos ocurridos de un total de expuestos n_i con cierto patrón de covariables i.

La esperanza de Y_i puede expresarse como

$$E(Y_i) = \mu_i = n_i\theta_i$$

¿Qué representa la n_i? Por ejemplo si Y_i es el número de robos para cierto modelo y marca de coche en una compañía de seguros, es obvio pensar que el número de siniestros depende del número n_i de pólizas vendidas para ese tipo de coche, además se tienen otras variables que afectan la tasa de ocurrencia θ_i como pueden ser, la antigüedad del coche, zona dónde es usado, si se guarda en cochera, etc.

La dependencia de la tasa de ocurrencia de las variables explicativas es de la forma:

$$\theta_i = \exp^{x_i'\beta}$$

el modelo de regresión Poisson es de la forma:

$$E(Y_i) = \mu_i = n_i \exp^{x_i'\beta};$$

donde x_i es el $i-$ésimo renglón de la matriz X,

al sacar logaritmo de ambos lados se tiene:

$$\log(\mu_i) = \log(n_i) + x_i'\beta$$

A $\log(n_i)$ se le conoce como **offset**, que es una constante conocida que especifica **los diferentes niveles de exposición** y se incorpora en la estimación de los parámetros.

1.2. Interpretación de los parámetros

En el caso mś sencillo, cuando se tiene una sola variable explicativa binaria, en donde $x_j = 1$ indica presencia del atributo j, entonces la razón de tasas (rate ratio) RR para comparar la presencia del atributo versus la ausencia del mismo es:

$$RR = \frac{\mu_{j1}}{\mu_{j0}} = \frac{E(Y_i|x_j = 1)}{E(Y_i|x_j = 0)} = \exp^{\beta_j}$$

desde luego que se hizo permanecer el resto de las variables explicativas iguales para ambos niveles de x_j.

En este modelo la interpretación de los parámetros β es en función de la **razón de tasas de ocurrencia** RR por sus siglas en inglés *Rate Ratio* y NO del cociente de momios OR ni tampoco riesgo relativo.

Algo muy similar ocurre si se considera ahora una variable continua x_k, entonces al incrementar en una unidad esta variable, repercutirá en un efecto multiplicativo de tamaño $\exp^{\beta_k}$ sobre la tasa μ.

1.3. Distribución de los parámetros

Igual que para el modelo logístico ocurre que

$$\frac{\hat{\beta}_j - \beta_j}{s.e(\hat{\beta}_j)}$$

se distribuye aproximadamente como $N(0,1)$, esta estadística es la utilizada para pruebas de hipótesis y para construir intervalos de confianza.

1.4. Residuales

El valor esperado de ocurrencias del evento está dado por:

$$\hat{Y} = \hat{\mu}_i = n_i \exp(x_i'\hat{\beta}) = e_i$$

Se definen los **residuales de Pearson** como:

$$r_i = \frac{o_i - e_i}{\sqrt{(e_i)}},\ i = 1, \ldots, n$$

estos tienen una distribución asintótica normal $N(0,1)$ y por tanto

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

tiene una distribución asintótica Ji cuadrada con $n-p$ grados de libertad.

Los **residuales estandarizados**, toman en cuenta la posición del dato dentro de la nube de datos a través de la cantidad h_i:

$$r_i^* = \frac{o_i - e_i}{\sqrt{(e_i)}\sqrt{(1-h_i)}},$$

donde h_i es el i-ésimo elemento de la diagonal de $H = X'(X'X)^{-1}X$.

Los **residuales deviance**:

$$d_i = \text{signo}(o_i - e_i)\sqrt{2\left[o_i \log(\frac{o_i}{e_i}) - (o_i - e_i)\right]},$$

llamados así por su contribución a la Deviance:

$$D = 2\sum \left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i)\right] = \sum d_i^2,$$

cuya distribución es asintóticamente Ji cuadrada con $n-p$ grados de libertad.

En estos modelos la X^2 y D se usan directamente como medidas de carencia de ajuste del modelo.

1.5. Sobredispersión

Hay ocasiones en que no se logra que el modelo ajuste bien a los datos, esto ocurre generalmente por dos motivos

- Cuando el modelo no está bien especificado, es decir hay variables predictoras que no se han incluído en el modelo, o que hay relaciones no lineales.
- Cuando existe sobredispersión, $E(y) < Var(y)$

1.5.1. Modelo Quasipoisson

Suponiendo que se modelo bien la parte sistemática, se puede relajar la condición $E(y) = Var(y)$ y hacer que $Var(y_i|\eta_i) = \phi\mu_i$, cuando $\phi < 1$ se tiene baja-dispersión y si $\phi > 1$ se tiene sobredispersión, hacer esto lleva al modelo conocido como **Quasipoisson**

Para usar este modelo se requiere una estimación previa de ϕ, un posible estimador es:

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

El modelo entonces tiene los mismos coeficientes que el modelo Poisson, peron la inferencia se ajusta por la sobredispersión. Los errores estandar de los coeficientes se multiplican por $\phi^{1/2}$. En R esto se hace usando `family=quasipoisson`

1.5.2. Modelo con distribución Binomial Negativa

La distribución Binomial Negativa es una generalización de la Poisson que permite la sobredispersión.

En esta distribución $E(y_i) = \mu_i$ y $Var(y_i) = \mu_i + \mu_i^2/\theta = \mu_i + \alpha\mu_i^2$

y cuando $\alpha \to 0$, $Var(y_i) \to \mu_i$, esto es la Binomial Negativa tiende a la Poisson.

Si la θ se conoce, en R se puede usar `glm` especificando por ejemplo `family=negative.binomial(theta=1)`. Pero la mayoría de las veces no se conoce y entonces no es un caso de glm, pero se puede estimar por máxima verosimilitud tanto a las $\beta's$ y como a θ con `glm.nb()` del paquete MASS.

1.5.3. Exceso de ceros

Hay ocasiones en que hay demasiados ceros, que propician un mal ajuste, por ejemplo en estudios de servicios de salud encuentran que mucha gente NUNCA acude al hospital en un periodo dado, en el caso de seguros, hay muchos asegurados que no presentan reclamaciones en el año, esto puede explicarse por el pago de deducible.

Dos son los modelos que veremos que lidian con el exceso de ceros.

Modelo ZIP por sus siglas en inglés *Zero Inflated Poisson*. Este modelo supone que los conteos surgen de una mezcla de dos clases de observaciones: los ceros estructurales que siempre $y_i = 0$ y el resto de conteos y_i que ocasionalmete tomara el valor de cero. Este modelo tiene dos componentes

1. Un modelo para el evento binario que define si pertenece a los ceros o no. Es un modelo logístico para la probabilidad π_i de pertenecer a los ceros, con variables predictoras $z_1, z_2, \ldots z_q$

$$logit(\pi_i) = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \ldots + \gamma_q z_{iq}$$

2. Un modelo Poisson para la otra clase, que pueden ser conteos ceros o positivos, con variables predictoras $x_1, x_2, \ldots x_p$

$$log(\mu(x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

Se puede mostrar que

$$P(y_i = 0|x, z) = \pi_i + (1 - \pi_i)e^{-\mu_i}$$

$$P(y_i|x, z) = (1 - \pi_i) \times \left[\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}\right],\ y_i \geq 0$$

Modelo Hurdle tiene un proceso que genera $y_i = 0$ y otro que genera $y_i \geq 1$.

- $$P(y_i = 0|x, z) = \pi_i,\ y_i = 0$$

- $$P(y_i|x,z) = \psi\left[f_2(y_i)\right] = \frac{1-f_1(0)}{1-f_2(0)}\left[\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}\right] = \frac{(1-\pi_i)}{1-e^{-\mu_i}}\times\left[\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}\right],\ y_i \geq 0$$

  En este caso la valla es uno, primero se presenta la probabilidad de no pasar la valla y después la probabilidad de que el conteo sea mayor o igual a uno, ψ es la probabilidad de pasar la valla.

1.6. Modelos Loglineales

Los modelos loglineales para tablas de $2\times$ 2 describen las asociaciones entre dos variables **discretas** digamos X y Y. El modelo loglineal nos dice cuan grande es el conteo de la celda dependiendo de los niveles de las dos variables categóricas.

Las frecuencias de tabla pueden calcularse como $\mu_{ij} = n\pi_{ij}$ y si consideramos que las variables X y Y son independientes se tiene que $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ entonces

$$\mu_{ij} = n\,\pi_{i\bullet}\pi_{\bullet j}$$

.

Si sacamos logaritmo de ambos lados de la igualdad tendremos

$$\log(\mu_{ij}) = \log(n) + \log(\pi_{i\bullet}) + \log(\pi_{\bullet j})$$

el modelo podemos verlo como:

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Es importante hacer notar que en el conjunto de los términos $\{\lambda_i^X\}$hay un término redundante y sólo $I-1$ de ellos son desconocidos. Análogamente para el conjunto $\{\lambda_j^Y\}$ hay un término redundante y solo hay $J-1$ parámetros desconocidos. Puede haber distintas parametrizaciones y no hay un único conjunto de parámetros. Deben imponerse diferentes condiciones para obtener los parámetros.