Tarea 8: Modelos Estadísticos I.

Rojas Gutiérrez Rodolfo Emmanuel

29 de mayo de 2021

Observación 1. *Se dice que una distribución, absolutamente continua o discreta, pertenece a la familia de modelos de dispersión si su función de densidad o masa probabilidades f, admite la siguiente descomposición*

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(\theta,\phi)\right\},$$

sobre su soporte, se destaca que a θ se le conoce como el parámetro canónico de la distribución. Además, en clase se vio que suele ser el caso en diversos modelos de esta familia que $a(\phi) = \phi/\omega$, donde ω es una constante conocida. Esto último es relevante, debido a que también se vio que si tenemos una muestra aleatoria $Y_1,\ldots,Y_N$, donde para cada $i \in \{1,\ldots,N\}$ se tiene que Y_i tiene función de densidad:[1]

$$f(y_i|\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(\theta_i,\phi)\right\},$$

con $a_i(\phi) = \phi/\omega_i$ para alguna ω_i conocida. Entonces, la Deviance para algún modelo en el que se hagan ciertos supuestos sobre los parámetros $(\theta_1,\ldots,\theta_N)$, queda dada por

$$D = 2\sum_{i=1}^{N} \omega_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) + b(\hat{\theta}_i))\right],$$

donde, para $i \in \{1,\ldots,N\}$ se tiene que $\tilde{\theta}_i$ es el estimador de máxima verosimilitud[2] *de θ_i, bajo el modelo maximal que supone todos los parámetros θ distintos, y $\hat{\theta}_i$ es el estimador de máxima verosimilitud de θ_i bajo el modelo de interés.* △

Ejercicio 1:

> Considere una muestra aleatoria $Y_1,\ldots,Y_N$ con distribución exponencial
>
> $$f(y_i|\theta_i) = \theta_i \exp\{-\theta_i y_i\}.$$
>
> Obtenga la deviance comparando el modelo máximal con diferentes valores de θ_i para cada Y_i con el modelo con $\theta_i = \theta$ para cada $i \in \{1,\ldots,N\}$.

[1] O masa de probabilidades según sea el caso.

[2] El EMV del parámetro canónico de cada una de las i distribuciones.

Solución. Primeramente, observe que para $i \in \{1, \ldots, N\}$ se tiene que

$$\begin{aligned} f(y_i|\theta_i) &= \exp\{\ln(\theta_i)\} \exp\{-\theta_i y_i\} \\ &= \exp\{-\theta_i y_i + \ln(\theta_i)\}, \ y_i > 0. \end{aligned} \tag{1}$$

Ahora, defina

$$\psi_i = -\theta_i, \ b(\psi_i) = -\ln(-\psi_i) = -\ln(\theta_i), \ c(y_i, \phi) = 0 \text{ y } a_i(\phi) = \phi/\omega_i \text{ donde } \omega_i = 1 \text{ y } \phi = 1, \tag{2}$$

entonces, la densidad de la distribución exponencial con parámetro de forma θ_i, puede expresarse como:

$$f(y_i|\psi_i, \phi) = \exp\left\{\frac{y_i\psi_i - b(\psi_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

Por lo que, la distribución exponencial con parámetro de forma θ_i es un miembro de la familia exponencial de modelos de dispersión, con parámetro canónico $\psi_i = -\theta_i$. Luego, denote por $\mathbf{y} = (y_1, \ldots, y_N)$ a la muestra aleatoria observada de las Y_i, y note que de (1) se sigue que la log-verosimilitud para la misma, esta dada por

$$l(\theta_1, \ldots, \theta_N|\mathbf{y}) = \sum_{k=1}^{N} \left[y_j(-\theta_j) + \ln(\theta_j)\right], \tag{3}$$

De este modo, para el modelo máximal que supone un parámetro distinto para cada[3] Y_i, obtenemos las siguientes ecuaciones Score a partir de la log-verosimilitud en (3)

$$0 = \frac{\partial l(\theta_1, \ldots, \theta_N|\mathbf{y})}{\partial \theta_i} = -y_i + \frac{1}{\theta_i}, \ i \in \{1, \ldots, N\}.$$

Resolviendo dicho sistema de ecuaciones Score, se obtiene la siguiente estimación por máxima verosimilitud para el parámetro θ_i:

$$\widetilde{\theta}_i = \frac{1}{y_i}, \ i \in \{1, \ldots, N\}. \tag{4}$$

Y bajo la invarianza de los estimadores de máxima verosimilitud, se sigue que

$$\widetilde{\psi}_i = -\frac{1}{y_i}, \ i \in \{1, \ldots, N\}.$$

Por último, si suponemos un modelo en el que[4] $\theta_i = \theta > 0$ para cada $i \in \{1, \ldots, N\}$, entonces, la log-verosimilitud en (3) se simplifica de la siguiente forma

$$l(\theta|\mathbf{y}) = \sum_{k=1}^{N} \left[y_i(-\theta) + \ln(\theta)\right],$$

y por el curso de Inferencia Estadística I, sabemos que la log-verosimilitud anterior se maximiza en

$$\hat{\theta}_i = \hat{\theta} = 1/\overline{y} \text{ con } i \in \{1, \ldots, N\},$$

[3] Observe que esto es equivalente a suponer un parámetro canónico distinto, para cada Y_i por la igualdad $\psi_i = -\theta_i$.

[4] Observe que esto es equivalente a suponer un solo parámetro canónico, para cada Y_i, además esto también es equivalente a suponer que todas las Y_i provienen del mismo modelo.

donde $\overline{y} = (1/N)\sum_{i=1}^{N} y_i$. De lo que se sigue, nuevamente por la invarianza de los estadísticos de máxima verosimilitud, que

$$\widehat{\psi}_i = -1/\overline{y} \text{ con } i \in \{1, \ldots, N\}\,. \tag{5}$$

De este modo, se concluye de la Observación 1 y de (2), que la Deviance para el modelo de interés $\theta_i = \theta$ para cada $i \in \{1, \ldots, N\}$, queda dada por

$$D = 2\sum_{i=1}^{N} \omega_i \left[y_i(\widetilde{\psi}_i - \widehat{\psi}_i) - (b(\widetilde{\psi}_i) - b(\widehat{\psi}_i)) \right],$$

donde $\widehat{\psi}_i$ y $\widetilde{\psi}_i$ con $i \in \{1, \ldots, N\}$ como fueron dados en (4) y (5). Por último, recordando que $\omega_i = 1$ para cada $i \in \{1, \ldots, N\}$ y que $b(\psi) = -\ln(-\psi)$, se obtiene la siguiente igualdad para la deviance

$$\begin{aligned} D &= 2\sum_{i=1}^{N} \left[y_i \left(\frac{1}{\overline{y}} - \frac{1}{y_i} \right) + \ln\left(\frac{1}{y_i} \right) - \ln\left(\frac{1}{\overline{y}} \right) \right] = 2\sum_{i=1}^{N} \left[-1 - \ln\left(\frac{y_i}{\overline{y}} \right) + \frac{y_i}{\overline{y}} \right] \\ &= -2N - 2\sum_{i=1}^{N} \left[\ln\left(\frac{y_i}{\overline{y}} \right) \right] + \frac{2N\overline{y}}{\overline{y}} = -2\sum_{i=1}^{N} \left[\ln\left(\frac{y_i}{\overline{y}} \right) \right], \end{aligned}$$

es decir

$$D = -2\sum_{i=1}^{N} \left[\ln\left(\frac{y_i}{\overline{y}} \right) \right].$$

Finalmente, se destaca que bajo la hipótesis nula $\mathcal{H}_0 : \theta_i = \theta$ para cada $i \in \{1, \ldots, N\}$ se tiene que

$$D \sim \chi^2_{(N-1)}, \text{ de manera aproximada.}$$

Por lo que, si $\chi^2_{N-1,1-\alpha}$ denota al cuantil $1-\alpha$ de una distribución Ji-cuadrada con $N-1$ grados de libertad, rechazariamos dicha hipotesis bajo un nivel de significania de $100 \cdot \alpha\,\%$, en caso de que

$$D = -2\sum_{i=1}^{N} \left[\ln\left(\frac{y_i}{\overline{y}} \right) \right] \geq \chi^2_{N-1,1-\alpha}.$$

■

Ejercicio 2:

Sea $l(b_{\text{mín}})$ el valor máximo de la función de logverosimilitud para el modelo minimal con predictor lineal $x'\beta = \beta_1$ y sea lineal $x'\beta = \beta_1 + \beta_2 x_1 + \ldots \beta_p x_{p-1}$

a) Pruebe que la estadística ji-cuadrada es

$$C = 2\left[l(b) - l(b_{\text{mín}}) \right] = D_0 - D_1,$$

donde D_0 es la deviance para el modelo minimal y D_1 para el modelo más general.

b) Deducir que si $\beta_2 = \beta_3 = \ldots = \beta_p = 0$ entonces C tiene la distribución Ji-cuadrada central $p-1$ grados de libertad.

a) y **b)**

Solución. Primeramente, suponga que se cuenta con N observaciones, que $1 < p < N$ y denote por $l(b)$ al valor máximo de la log-verosimilitud para el modelo con predictor lineal $x'\beta = \beta_1 + \beta_2 x_1 + \ldots + \beta_p x_{p-1}$. La deviance para el modelo minimal, se definió como la deviance nula y se sabe que está dada por:

$$D_0 = 2\left[l(b_{\text{máx}}) - l(b_{\text{mín}})\right],$$

donde, $l(b_{\text{máx}})$ denota al valor máximo de la verosimilitud para el modelo saturado, y $l(b_{\text{mín}})$ como en el enunciado. Por otro lado, la deviance para el modelo más general con predictor lineal $x'\beta = \beta_1 + \beta_2 x_1 + \ldots \beta_p x_{p-1}$, está dada por:

$$D_1 = 2\left[l(b_{\text{máx}}) - l(b)\right].$$

Así, por la teoría distribucional de razón de verosimilitudes, se sabe que bajo la hipótesis nula $\mathcal{H}_0 : \beta_2 = \ldots = \beta_p = 0$, $D_0 \sim \chi^2_{N-1}$ y que $D_1 \sim \chi^2_{N-p}$, y[5] por ende, se sigue bajo dicha hipótesis nula que

$$D_0 - D_1 \sim \chi^2_{(N-1)-(N-p)} = \chi^2_{p-1}.$$

El resultado es ahora claro al notar que

$$D_0 - D_1 = 2\left[l(b_{\text{máx}}) - l(b_{\text{mín}})\right] - 2\left[l(b_{\text{máx}}) - l(b)\right] = 2\left[l(b) - l(b_{\text{mín}})\right] = C.$$

Donde la anterior es la estadística Ji cuadrada solicitada.

■

Ejercicio 3:

> El número de muertes por leucemia y otros tipos de cáncer entre supervivientes de la bomba atómica de Hiroshima se muestran en la Tabla en la figura 1, clasificados por la dosis de radiación recibida. Los datos se refieren a muertes durante el periodo 1950-1959, entre supervivientes cuyas edades estaban entre 25 y 64 años en 1950. Obtener un modelo adecuado para describir la relación dosis –respuesta entre la radiación y las tasas de mortalidad proporcionales para leucemia.

Tabla 1. Muertes por leucemia y otros tipos de cánceres, clasificadas por las dosis de radiación recibida de la bomba atómica de Hiroshima.

Muertes	Dosis de radiación (rads)					
	0	1-9	10-49	50-99	100-199	200+
Leucemia	13	5	5	3	4	18
Otros cánceres	378	200	151	47	31	33
Total	391	205	156	50	35	51

Figura 1: Datos: Ejercicio 3

[5] Ambas de manera aproximada, provistos de que N se suficientemente grande. Y el modelo sea razonable.

Solución. Primeramente, denote por $y = \begin{pmatrix} 13 & 5 & 5 & 3 & 4 & 18 \end{pmatrix}'$ al vector columna de observaciones de casos de Leucemia por grupo de exposición y por $T = \begin{pmatrix} 391 & 205 & 156 & 50 & 35 & 51 \end{pmatrix}'$ al vector columna con el total de individuos por grupo de exposición, entonces, el vector columna π con entradas:

$$\pi_i = y_i/T_i,\ i \in \{1, \ldots, 6\},$$

nos da la proporción de individuos que murieron en cada grupo a causa de la leucemia, dado que fallecieron por algún tipo de cáncer. Utilizando dicho vector, se construyo la gráfica en la figura 2. En dicha gráfica, podemos notar que existe una relación casi creciente entre la proporción de

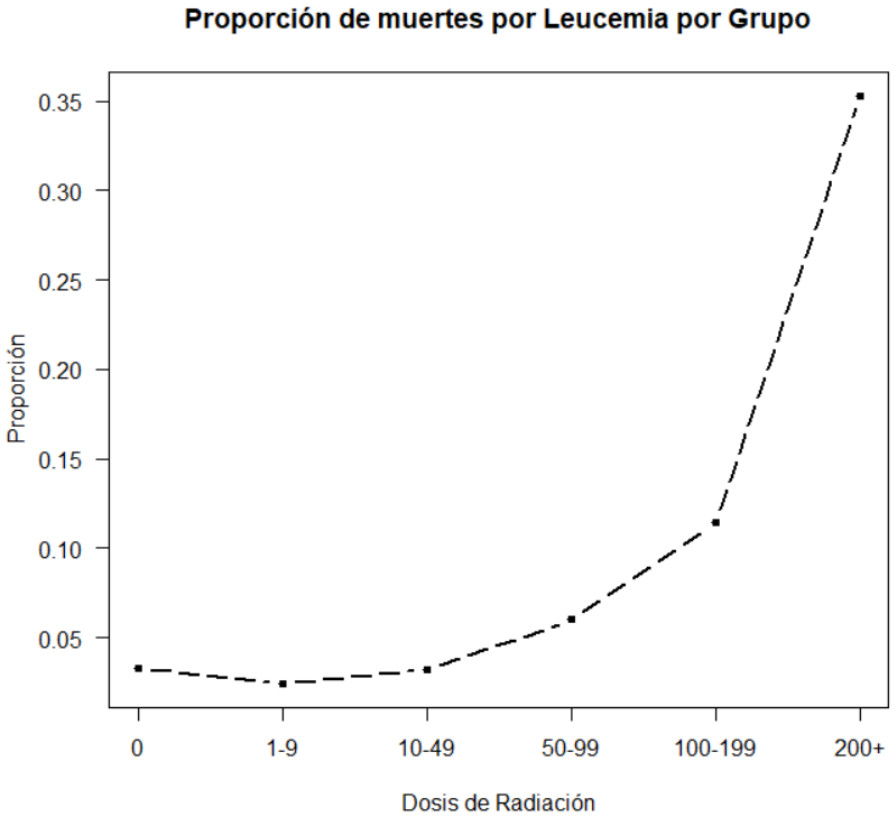

Figura 2: Proporción del número de muertes por leucemia, por grado de exposición a la radiación al que estuvieron expuestos los individuos.

muertes por leucemia, respecto a muertes por otros tipos de cáncer, y la dosis de radiación a la que estuvieron expuestos los distintos individuos en cada grupo. Ahora, piense a la proporción de muertes por leucemia y no por otro tipo de cáncer por cada nivel de radiación, como una estimación de la probabilidad de morir por Leucemía y no por otro cáncer, dado que la causa de muerte fue cáncer y que se estuvo expuesto a un cierto nivel de radiación, bajo este contexto, se propone modelar la relación dosis-respuesta entre la dosis de radiación y las tasas de mortalidad para leucemia, de

la siguiente manera. Sea $x = (0, 1, 10, 50, 100, 200)'$ el vector columna con los extremos izquierdos de cada uno de los intervalos de dosis de radiación, a las que pertenecen las distintas proporciones de muertes por leucemia, se propone un modelo lineal generalizado de la familia binomial, para modelar la probabilidad de morir por leucemia dado que la causa de muerte fue cáncer y dado que se estuvo expuesto a cierto nivel de radiación, es decir, se ajustará el siguiente GLM binomial:

$$g(\pi_i) = \beta_0 + \beta_1 x_i, \ i \in \{1, \ldots, 6\}, \tag{6}$$

donde g es una función liga, la cual se elegirá entre las tres más utilizadas para este tipo de modelos: la logit, la probit o la logaritmo complementaria. Ahora, hay dos razones para la elección de este modelo:

1. Se intento ajustar un modelo con variables Dummy, una por cada grupo, pero al final el modelo terminaba siendo el modelo saturado[6], y a pasar de ello, no obtenía un mejor AIC que ninguno de los modelos que se presentarán[7] a continuación.

2. De acuerdo con Dunn y Smyth (2018), cuando las categorías que separan a los datos son intervalos numéricos de longitud distinta, conviene tomar un valor representativo de cada intervalo y cambiar la variable categórica por una cuantitiva, que es lo que se esta haciendo. Además, se menciona que en casos como el nuestro en el que el último intervalo solo posee un extremo izquierdo[8], es preferible tomar dicho valor como representativo para cada intervalo, lo que es contra-intuitivo dado que la elección mas natural sería el promedio, sin embargo, la última categoría no posee un promedio bien definido.

Hechas las observaciones anteriores, se procedió a realizar el ajuste del modelo en (6) con la función *glm* de *R*, para cada una de las funciones liga mencionadas, los resultados obtenidos se presentan en la Tabla 1.

Liga	$\hat{\beta}_0$	$\hat{\beta}_1$	Deviance Residual	DF	AIC
Logit	-3.4890	0.0144	0.4321	4	26.0971
Probit	-1.8959	0.0075	0.5324	4	26.1974
Comp log-log	-3.4949	0.0133	0.4434	4	26.1084

Cuadro 1: Resultados de regresión para el modelo en (6).

Pese a que los AIC obtenidos son muy similares para cada una de las funciones liga propuestas, el AIC del modelo con la liga logit es por poco el más pequeño de todos,[9] por lo que se propone este modelo para modelar la relación dosis respuesta solicitada. Un resumen completo sobre las estimaciones de los coeficientes de este modelo se presenta en la Tabla 2:

[6] Ya que incluía 6 parámetros, el mismo número que el número de observaciones.
[7] Sin importar la elección de la función liga.
[8] Entiéndase finito.
[9] Además, a pesar de que aquí no se usará, se sabe que el modelo binomial con la liga logit, tiene una ventaja interpretativa, ya que surge el concepto de momios.

Coeficiente	Estimación	z-valor	p-valor
β_0	-3.4890	-17.098	$< 2 \cdot 10^{-16}$
β_1	0.0144	7.932	$2.15 \cdot 10^{-15}$

Cuadro 2: Análisis de regresión para el modelo en (6) con liga logit.

En la tabla 2 es posible observar que ambos coeficientes resultan significativos, bajo un nivel de significancia del 5 %, más aún, en la Tabla 1 se obtuvó que este modelo cuenta con una Deviance residual 0.4321 la cual tiene 4 grados de libertad, dado que la Deviance residual no es mayor que sus grados de libertad, se descarta el tener que ajustar un modelo que considere sobre dispersión. Finalmente, se deja en la figura 3, una gráfica de este modelo con los valores ajustados para la probabilidad π dados diversos valores de niveles de exposición, con intervalos de confianza en lineas azules al 95 % y las proporciones π_i en puntos rojos. Es importante destacar, que en dicha gráfica se preserva el patrón casi creciente observado en la gráfica en la figura 2, y que el ajuste del modelo al menos visualmente parece razonable.

■

Ejercicio 4:

En la clase vimos el procedimiento iterativo IRLS de estimación de los parámetros de los modelos lineales generalizados, para el modelo de regresión logística. En este caso, para obtener el algoritmo de estimación, se consideraban las siguientes relaciones:

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right),\ \frac{\partial \eta}{\partial \mu} = g'(\mu) = \frac{1}{\mu(1-\mu)},\ V(\mu) = \frac{\mu(1-\mu)}{n} \text{ y } w = n\mu(1-\mu).$$

Haga lo siguiente:

a) Haga el desarrollo de las relaciones anteriores para el modelo de regresión Poisson.

b) Obtenga el algoritmo correspondiente a la regresión Poison.

c) En la tabla en la figura 4 se muestran los números de casos de AIDS que se presentaron en los años señalados, en Bélgica a partir de 1981. Se asume que estos datos siguen un modelo de regresión de Poisson, donde la variable explicativa son los años y la respuestas los números de casos por año.

Observación 2. *Para este ejercicio, se considerará una liga Canónica puesto que ninguna función liga ha sido especificada. Además, para este ejercicio tome en cuenta que la notación* $\mathbb{N}_0$ *será utilizada para hacer referencia al conjunto* $\mathbb{N} \cup \{0\}$. △

Solución. **a)** Inicialmente, considere la función masa de probabilidades de una variable aleatoria Poisson de media $\mu > 0$:

$$p(y|\mu) = \frac{\exp\{-\mu\}\,\mu^y}{y!},\ y \in \mathbb{N}_0,$$

y, note que

$$\frac{\mu^y}{y!} = \exp\{y\log(\mu) - \log(y!)\},\ y \in \mathbb{N}_0.$$

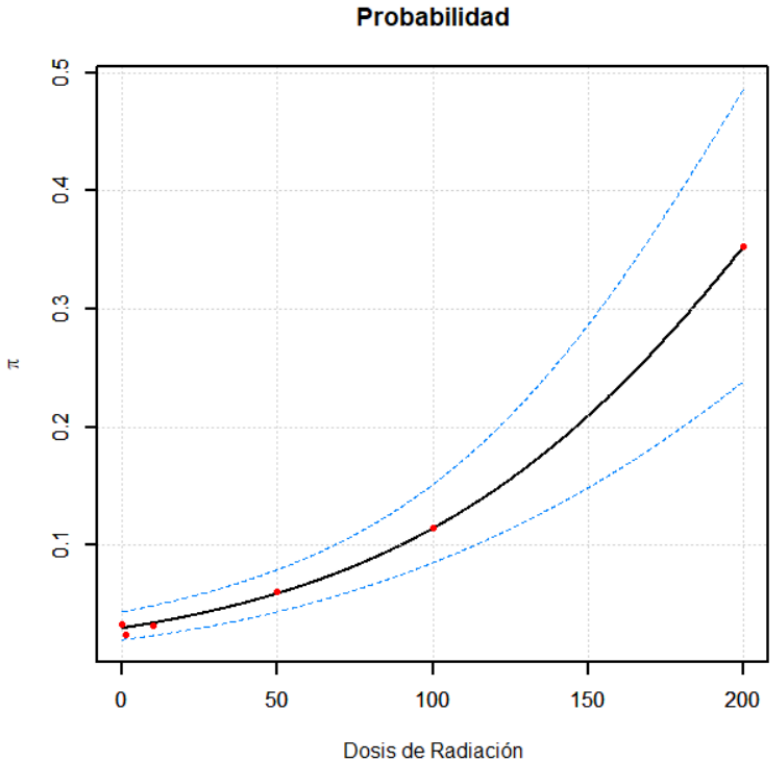

Figura 3: Probabilidad de morir por leucemia dado que la causa de muerte fue algún tipo de cáncer, bajo diferentes niveles de exposición a la radiación. Proporciones observadas en puntos rojos, e intervalos de confianza al 95 % en lineas azules.

Tabla.-Números de casos de AIDS que ocurrieron anualmente de 1981 a 1993 en Bélgica.

Año	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
Casos	12	14	33	50	67	74	123	141	165	204	253	246	240

Figura 4: Datos: Ejercicio 4

De las igualdades anteriores se concluye que

$$p(y|\mu) = \exp\left\{y\log(\mu) - \log(y!) - \mu\right\},\ y \in \mathbb{N}_0.$$

Así, tome $a(\phi) = \phi = 1,\ \theta = \log(\mu),\ \mu = b(\theta) = e^{\theta}$ y $c(y,\phi) = -\log(y!)$ entonces

$$p(y|\mu) = p(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},\ y \in \mathbb{N}_0.$$

Por lo que, se concluye que la distribución Poisson es un miembro de la Familia Exponencial de Modelos de Dispersión, con:

$$\begin{aligned} &\text{Parámetro canónico } \theta = \log(\mu). \\ &\text{Parámetro de dispersión } \phi = 1. \\ &\text{Función de cumulantes } \mu = b(\theta) = e^{\theta}. \end{aligned} \tag{7}$$

Además, se tiene que

$$V(\mu) = \frac{\partial\mu}{\partial\theta} = \frac{\partial}{\partial\theta}e^{\theta} = e^{\theta} = e^{\ln(\mu)} = \mu. \tag{8}$$

Ahora, denote por $g : (0,\infty) \to \mathbb{R}_+$ a la liga canónica, la cual por (7) esta dada por

$$g(\mu) = \theta = \ln(\mu). \tag{9}$$

Y sea η un predictor lineal tal que

$$g(\mu) = \eta.$$

Entonces

$$\frac{\partial\eta}{\partial\mu} = g'(\mu) = \frac{1}{\mu}. \tag{10}$$

Finalmente, los pesos de regresión quedan dados por

$$w = \frac{1}{V(\mu)(\partial\eta/\partial\mu)} = \frac{1}{[g'(\mu)]^2 V(\mu)} = \frac{\mu^2}{\mu} = \mu.$$

b) Suponga que tiene $y_1,\ldots,y_N$ observaciones de conteos, donde y_i proviene de una distribución Poisson de media μ_i, $i \in \{1,\ldots,N\}$, y suponga que se quiere ajustar un GLM Poisson con liga canónica, utilizando una matriz X de covariables $n \times p$ con $p < n$. Se utilizarán las relaciones obtenidas en el inciso **a)**, así como el algoritmo IRLS proporcionado por el Doctor Villa en sus notas de clase, para particularizar dicho algoritmo al modelo de regresión Poisson. Para ello, denote por $\mathbf{x}_i'$ a la i-ésima fila de X y por $\beta = (\beta_0,\ldots,\beta_{p-1})$ al vector de coeficientes del modelo, entonces, el algoritmo IRLS para la estimación del vector de coeficientes β del modelo de regresión Poisson, queda dado como:

1. Sea $k = 0$. Se toman como semilla inicial del algoritmo los siguientes valores

$$\hat{\mu}_i^k = y_i \text{ y } \eta_i^k = g(\hat{\mu}_i^k) = \ln(\hat{\mu}_i^k),\ i \in \{1, \ldots, N\}$$

En caso, de que la elección realizada cause problemas con la definición de la función liga, se puede añadir una pequeña perturbación a los datos. Por ejemplo, en el caso de conteos Poisson es possible que alguna $y_i = 0$, lo que implicaría al llevar a cabo el procedimiento anterior, que se este tomando el logaritmo natural de cero o que se este dividiendo entre cero, por ello, una mejor elección será

$$\hat{\mu}_i^k = \begin{cases} y_i + 0.1 & \text{si } y_i = 0, \\ y_i & \text{si } y_i > 0. \end{cases} \text{ y } \hat{\eta}_i^k = g(\hat{\mu}_i^k) = \ln(\hat{\mu}_i^k),\ i \in \{1, \ldots, N\}.$$

donde, en la última igualdad se ha utilizado (9).

2. Se calculan los pseudo-datos

$$\begin{aligned} \zeta_i^{k+1} &= \hat{\eta}_i^k + \frac{g'(\hat{\mu}_i^k)(y_i - \hat{\mu}_i^k)}{\alpha(\hat{\mu}_i^k)} \\ &= \ln(\hat{\mu}_i^k) + \frac{(y_i - \hat{\mu}_i^k)}{\hat{\mu}_i^k \alpha(\hat{\mu}_i^k)},\ i \in \{1, \ldots, N\}, \end{aligned}$$

donde, en la última igualdad se ha utilizado (10). Y los pesos iterativos

$$\begin{aligned} w_i^{k+1} &= \frac{\alpha(\hat{\mu}_i^k)}{(g'(\hat{\mu}_i^k))^2 V(\hat{\mu}_i^k)} = \frac{\alpha(\hat{\mu}_i^k)}{(g'(\hat{\mu}_i^k))^2 V(\hat{\mu}_i^k)} \\ &= \alpha(\hat{\mu}_i^k)\mu_i^k,\ i \in \{1, \ldots, N\}. \end{aligned}$$

donde, en la última igualdad se ha utilizado (10) y (8). Cuando, $\alpha(\mu) = 1$ se le denomina peso de Fischer, y el método anterior coincide con el Scoring de Fischer.[10]

3. Encontrar $\hat{\beta}^{k+1}$; que es la solución al problema de mínimos cuadrados ponderados siguiente:

$$\hat{\beta}^{k+1} = argmin_\beta \left[\sum_{i=1}^{N} w_i^{k+1} (\zeta_i^{k+1} - \mathbf{x}_i' \beta)^2 \right].$$

Luego, se actualizan[11]

$$\hat{\eta}^{k+1} = X\hat{\beta}^{k+1},\ \hat{\mu}^{k+1} = g^{-1}\left(\hat{\eta}^{k+1}\right) = \exp\left\{\hat{\eta}^{k+1}\right\},$$

donde, en la última igualdad se ha utilizado (9).

4. Se calcula la Deviance total con los valores ajustados $\hat{\mu}^{k+1} = (\hat{\mu}_1^{k+1}, \ldots, \hat{\mu}_N^{k+1})$:

$$D(y, \hat{\mu}^{k+1}) = 2 \sum_{j=1}^{n} \left\{ y_j \ln(y_j / \hat{\mu}_j^{k+1}) - (y_j - \hat{\mu}_j^{k+1}) \right\}.$$

[10] Qué es el método que llevaremos finalmente a cabo.

[11] Donde, $\exp\left\{\hat{\eta}^{k+1}\right\}$ debe interpretarse como aplicar la función exponencial a cada entrada del vector $\hat{\eta}^{k+1}$.

Si:[12]

$$\frac{\left|D(y,\hat{\mu}^{k+1}) - D(y,\hat{\mu}^{k})\right|}{|D(y,\hat{\mu}^{k+1})| + 0.1} < \varepsilon,$$

donde ε suele ser una cantidad pequeña.[13] El algoritmo acaba después de $k+1$ iteraciones, y la estimación de máxima verosimilitud del vector β es $\hat{\beta}^{k+1}$. En caso contrario, regrese al paso dos y actualice cada k por $k+1$, en otras palabras, se repite con $\hat{\eta}^{k+1} = (\hat{\eta}_1^{k+1}, \ldots, \hat{\eta}_N^{k+1})$ y $\hat{\mu}^{k+1} = (\hat{\mu}_1^{k+1}, \ldots, \hat{\mu}_N^{k+1})$ todos los pasos realizados desde el paso dos con $\hat{\eta}^{k} = (\hat{\eta}_1^{k}, \ldots, \hat{\eta}_N^{k})$ y $\hat{\mu}^{k} = (\hat{\mu}_1^{k}, \ldots, \hat{\mu}_N^{k})$. Y así de manera iterativa hasta que la condición de las deviances se cumpla.

c) Para este inciso se programó en R el algoritmo anterior, el código puede ser visto en el Script Ejercicio4.R adjunto a este trabajo. De este modo, denote por y al vector de observaciones de casos de AIDS por año, y por x al vector de datos con los años en que ocurrieron dichos casos codificados de la siguiente manera: $1981 = 0,\ 1982 = 1, \ldots,\ 1992 = 11$ y $1993 = 12$.[14] El objetivo de este ejercicio, es plantear un modelo de regresión Poisson, donde la respuesta es y y x es la variable explicativa. Con ello en mente, se deja en la figura 5 una gráfica del logaritmo natural del número de casos de $AIDS$, contra el año codificado en el que ocurrieron. La idea de está gráfica, es darnos una idea de como poder establecer una relación para el GLM que se ajustará. Primeramente, se observa una relación creciente la mayor parte del tiempo, y existe una dispersión baja en las observaciones, por lo que, parece razonable el supuesto de que las observaciones siguen un modelo de regresión Poisson, donde la relación entre la liga canónica y el predictor lineal, sea también lineal en x. Con esto en mente, se utilizó el procedimiento IRLS programado para estimar los coeficientes del modelo

$$\ln(\mu_i) = \beta_0 + \beta_1 x_i,\ i \in \{1, \ldots, 13\}, \tag{11}$$

donde $\mu_i = \mathbb{E}[y_i|x_i]$. Lo que arrojó los resultados presentados en el Cuadro 3:

Iter $= k+1$	$\hat{\beta}_0^{k+1}$	$\hat{\beta}_1^{k+1}$	Deviance Total
1	3.426	0.195	82.491
2	3.345	0.202	80.688
3	3.343	0.202	80.686
4	3.343	0.202	80.686

Cuadro 3: Resultados del Procedimiento IRLS, para la estimación de los coeficientes del modelo en (11).

En el Cuadro 3, puede observarse que el método convergió después de 4 iteraciones, que las estimaciones de los coeficientes están dadas por

$$\hat{\beta}_0 = \hat{\beta}_0^4 = 3.343,\ \hat{\beta}_1^4 = \hat{\beta}_0 = 0.202,$$

y que la Deviance Residual de este modelo tiene un valor de

$$D = 80.686.$$

[12] El siguiente criterio es usado por el software R para declarar convergencia del Método. Donde, $D(y,\hat{\mu}^k)$ es la Deviance total calculada con los valores ajustados $\hat{\mu}^k = (\hat{\mu}_1^k, \ldots, \hat{\mu}_N^k)$.

[13] Por default R utiliza $\varepsilon = 10^{-8}$.

[14] Esto último por razones computacionales.

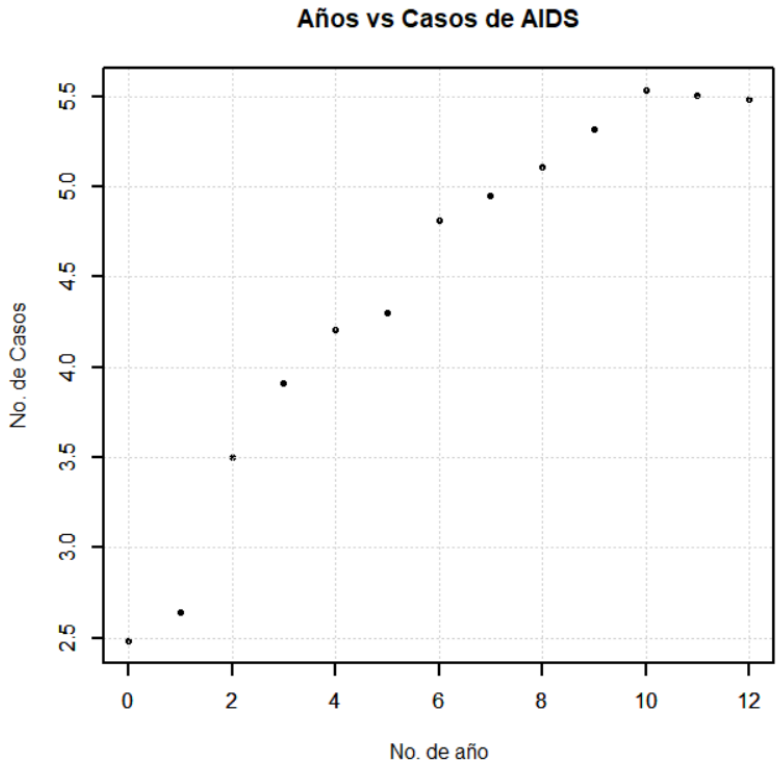

Figura 5: Gráfica de dispersión, casos de AIDS vs Año codificado de ocurrencia.

Los cálculos anteriores se constataron con la librería *glm* de *R* arrojando resultados idénticos, además, haciendo uso de dicha librería se obtuvo un AIC para este modelo de:

$$AIC = 166.4. \tag{12}$$

Por otro lado, observando la gráfica en la figura 5, se notó que parece existir cierta concavidad en la misma, por lo que se decidió agregar un término $\sqrt{x}$ al modelo, con el objetivo de ver si mejoraba en algo el ajuste del mismo. Por lo que, nuevamente se utilizó el algoritmo programado para estimar los coeficientes del modelo:

$$\ln(\mu_i) = \delta_0 + \delta_1 x_i + \delta_2 \sqrt{x_i},\ i \in \{1, \ldots, 13\}, \tag{13}$$

donde $\mu_i = \mathbb{E}[y_i|x_i]$. Obteniendo los resultados presentados en el Cuadro 4.

Iter $= k+1$	$\hat{\delta}_0^{k+1}$	$\hat{\delta}_1^{k+1}$	$\hat{\delta}_2^{k+1}$	Deviance Total
1	1.894	-0.072	1.333	26.957
2	1.692	-0.1005	1.4851	26.324
3	1.678	-0.103	1.496	26.322
4	1.678	-0.103	1.496	26.322

Cuadro 4: Resultados del Procedimiento IRLS, para la estimación de los coeficientes del modelo con el termino adicional $\sqrt{x}$.

En el Cuadro 4, puede notar que el método convergió después de 4 iteraciones, que las estimaciones de los coeficientes de este modelo están dadas por:

$$\hat{\delta}_0 = \hat{\delta}_0^4 = 1.678,\ \hat{\delta}_1 = \hat{\delta}_1^4 = -0.103,$$
$$\hat{\delta}_2 = \hat{\delta}_2^4 = 1.496,$$

y que la Deviance Residual del mismo tiene un valor de

$$D = 26.322. \tag{14}$$

Nuevamente, los cálculos anteriores se compararon con los obtenidos haciendo uso de la librería *glm* de *R*, lo que corroboró los cálculos obtenidos, además, ocupando dicha librería se obtuvo un AIC para este modelo de:

$$AIC = 114.01.$$

Los AIC se calcularon para comparar los modelos expuestos. En este caso, resulta evidente que el modelo en (13) posee un mejor ajuste, al ver que el AIC del mismo es de 114.01, lo cual es bastante menor que el AIC del modelo (11) presentado en (12). Adicionalmente, se agrega el comentario de que haciendo uso de *R* todos los coeficientes del modelo (13), resultaron significativos bajo un nivel de significancia del 5 %.[15] Finalmente, se deja en la figura 6 una gráfica de los valores ajustados por el modelo (13), con sus respectivos intervalos de confianza al 95 % en lineas azules, y las observaciones en puntos negros. En dicha gráfica se puede ver un ajuste más que razonable del modelo ajustado.

■

Ejercicio 5:

En un estudio médico se cuentan los números de pólipos en pacientes que padecen poliposis adenomatosa familiar, después de recibir un tratamiento con un nuevo medicamento, o de un placebo. En la Tabla en la figura 7 se presentan los datos asociados al estudio. Haga lo siguiente

a) Grafique los datos y coméntelos.

b) Encuentre un modelo MLG Poisson adecuado para modelar los datos y justifique que hay sobredispersion.

[15] Además, los grados de libertad de la deviance residual en (14), se obtuvieron con *R* arrojando un resultado de 10. Y dado que, la deviance residual no resulta mucho mayor a sus grados de libertad no existe indicio de sobre-dispersión, y los *p*-valores resultantes se pueden tomar en cuenta sin problema alguno.

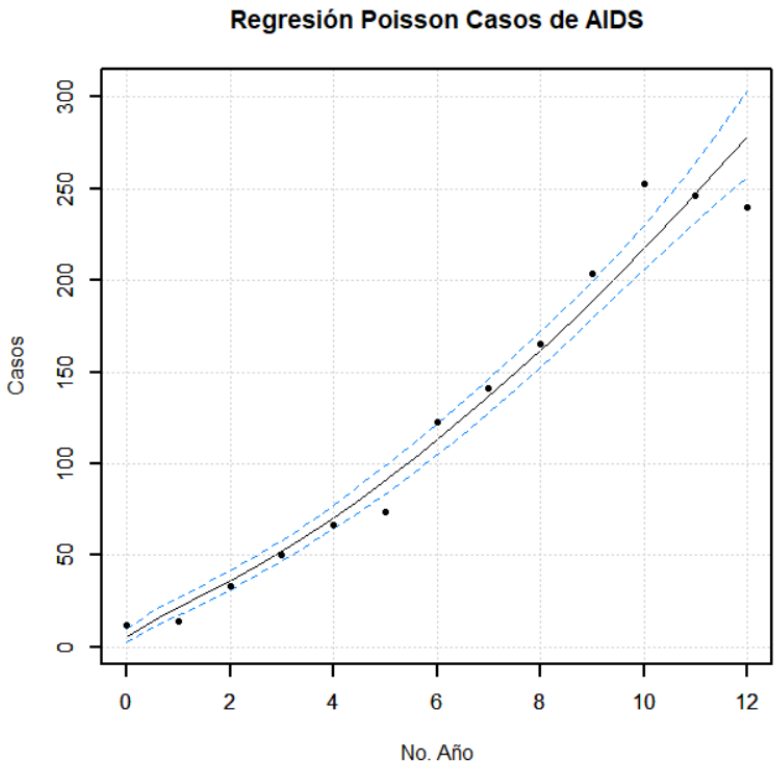

Figura 6: Valores ajustados por el modelo (13), mas intervalos de confianza al 95 % en lineas azules, y observaciones en puntos negros.

c) Ajuste un modelo cuasi-Poisson a los datos.

d) Ajuste un MLG binomial negativo a los datos.

e) Elija un modelo final y comente su elección.

Solución. **a)** Primeramente, es natural suponer que la distribución del número de pólipos en cada paciente, debería ser distinta de acuerdo a la edad del paciente, ya que la poliposis podría ser mas agresiva o menos agresiva de acuerdo a la edad, y al grupo en el que este se encuentre, esto último en caso de que el nuevo medicamento sea efectivo.[16] Por ende, se etiquetó al grupo que recibe tratamiento como el grupo 0 y al grupo que recibe el placebo como el grupo 1, y con ello se realizó la Boxplot presentada en la figura 8. En este Boxplot, pareciera existir cierta evidencia de que nuestra hipótesis sobre que el grupo al que pertenece el paciente influye en la distribución del número de pólipos que posee, cuando menos debería ser tomada en cuenta, ya que la caja para el

[16] Ya que debería haber un menor número de Pólipos en el grupo que recibe el tratamiento.

Tabla.-Número de pólipos en los grupos tratamiento y placebo del estudio del Problema 5.

Grupo Tratamiento				Grupo Placebo			
Número	Edad	Número	Edad	Número	Edad	Número	Edad
1	22	17	22	7	34	44	19
1	23	25	17	10	30	46	22
2	16	33	23	15	50	50	34
3	23			28	18	61	13
3	23			28	22	63	20
4	42			40	27		

Figura 7: Datos ejercicio 5.

grupo 0 se encuentra ubicada algo por debajo de la caja para el grupo 1. La posibilidad de que la distribución del número de pólipos también dependa de la edad, se puede explorar observando el gráfico presentado en la figura 9. En está figura, se puede apreciar un gráfico de dispersión, con el promedio del numero de pólipos por edad y por grupo. En dicho gráfico, es posible observar cierta tendencia decreciente del numero promedio de pólipos en ambos grupos conforme la edad avanza. Por lo que, en el siguiente inciso se considerarán ambas covariables para construir un modelo de regresión Poisson.

b) Se ajustará un modelo de regresión Poisson para los conteos de Polipos, y se consideraran como covariables a la Edad de los pacientes y el grupo codificado al pertenecen estos, además, se utilizará por función liga a la liga logaritmica, la cual corresponde a la liga canónica de este tipo de modelos, en otras palabras, se ajustará el modelo:

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},\ i \in \{1, \ldots, 20\}\,, \tag{15}$$

donde, para $i \in \{1, \ldots, 20\}$ se tiene que $\mu_i = \mathbb{E}[y_i | x_{1i}, x_{2i}]$, y_i es el número de pólipos en el paciente i, x_{1i} es la edad de dicho paciente y x_{2i} el grupo al que pertenece. El ajuste del modelo anterior, se realizó haciendo usó de la función glm de R lo que arrojó los resultados presentados en la tabla 5.

Coeficiente	Estimación	z-valor	p-valor
β_0	3.1699	18.84	$< 2 \cdot 10^{-16}$
β_1	-0.0388	-6.52	$7.02 \cdot 10^{-11}$
β_2	1.3591	11.55	$< 2 \cdot 10^{-16}$

Cuadro 5: Análisis de regresión para el modelo en (15) con liga logarítmica.

En principio, pareciera que el modelo anterior es bastante bueno ya que todos los coeficientes resultan significativos, bajo un nivel de significancia del 5 %. Sin embargo, los problemas empiezan al ver que el AIC arrojado por R para este modelo es de 273.88, un valor que a priori pareciera ser grande. Igualmente, se calculó con R la deviance residual para este modelo junto con sus grados de libertad, arrojando un resultado de 179.54 y 17 respectivamente, de este modo, dado que la deviance residual es bastante mayor a sus grados de libertad, se cuenta con un primer indicio para sospechar sobre la existencia de sobre-dispersión[17] en los datos modelados, por lo que, el modelo

[17]Lo que podría ser una explicación del mal ajuste del modelo reflejado por el AIC.

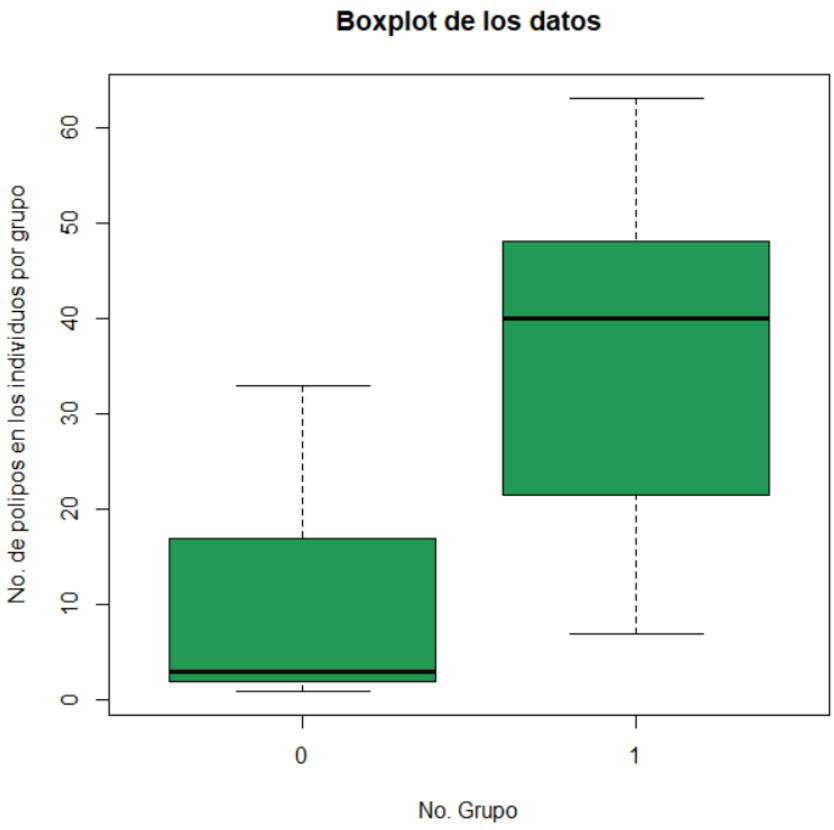

Figura 8: Boxplot conteos de pólipos en el grupo de tratamiento, codificado con 0 y en el grupo de placebo codificado con 1.

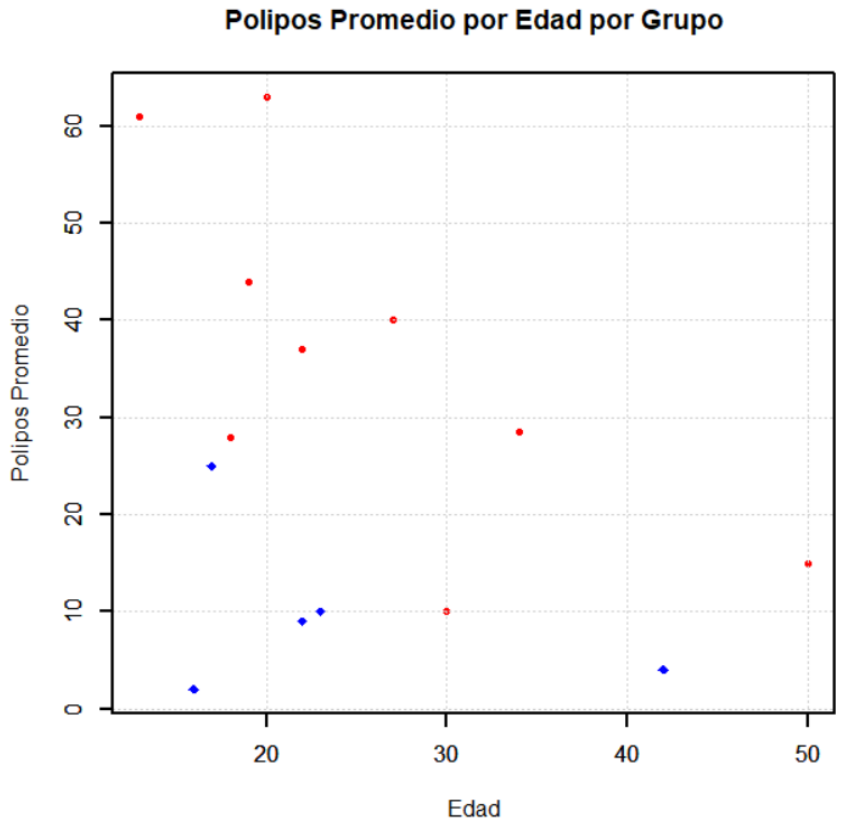

Figura 9: Gráfico de dispersión, promedio del numero de polipos por edad y por grupo. En rojo se presenta al grupo 1 y en azul al grupo 0.

de regresión Poisson podría no ser adecuado para modelar los mismos, puesto que podría estar subestimando los errores estándar en las estimaciones de los parámetros. Para confirmar la existencia de sobre-dispersión, se sugiere en Dunn y Smyth (2018) ver si el efecto de las deviances se sostiene en el modelo de regresión Poisson que considera a las covariables utilizadas en el modelo (15), más su interacción, [18] y que se descarte la existencia de ouliers que pudiesen estar inflando la varianza de los datos en algún sentido. Lo primero, se realizó nuevamente en R lo que arrojó una Deviance residual de 179.491 que cuenta con 16 grados de libertad, con lo que se concluyo que el efecto de las Deviances solamente fue acentuado. Por otra parte, para detectar posibles outliers, se hizo uso de nuestro conocimiento sobre medidas de influencia y de la función en R de nombre $influence.measures$, con la que se obtuvo el resumen presentado en la tabla 6. Los códigos en el

dfb.0	dfb.1	dfb.2	dffit	cov.r	cook.d	hat
0	0	0	0	2	4	0

Cuadro 6: Estadísticos de influencia para el modelo en (15) con liga logarítmica.

encabezado de la tabla 6, indican en este orden lo siguiente, $DFBETA_0$, $DFBETA_1$, $DFBETA_2$, $DFFIT$, $COVRATIO$, $Cook's\ D$, y valores palanca. Mientras que, los conteos que aparecen abajo son el número de observaciones que resultan influyentes de acuerdo a cada una de estas medidas. Note que, de acuerdo a los valores palanca no existe ninguna observación que se pueda considerar potencialmente influyente, sin embargo, hay cuatro valores que resultan influyentes al ser removidos bajo el criterio de la D de Cook y dos resultaron influyentes bajo la medida $COVRATIO$. Indagando un poco, se notó que para la D de Cook se trataba de las observaciones $8, 9, 10, 18$, mientras que para la $COVRATIO$ eran las observaciones $12, 19$. Teniendo esta información, se construyo la gráfica presentada en la figura 10, la cual es un gráfica de dispersión de los residuales de la Deviance estandarizados[19] contra los valores ajustados por el modelo, en la cual se resalta en color rojo los residuales de estas observaciones. Como puede verse, la mayoría de dichos residuales no son exageradamente grandes en comparación con los demás, salvo quizás el que se encuentra cercano a un valor de 6, esta observación podría ser considerada como un outlier, sin embargo, teniendo en cuenta que de manera global los residuales no son pequeños, y que las observaciones señaladas no influyen en más de un aspecto de la regresión, también podría ser una falta de ajuste del modelo Poisson, por lo que, cabe la posibilidad de que la observación con el residual mas grande sea de hecho un valor valido para los datos del fenómeno modelado. Finalmente, debido a que los demás indicadores realizados, señalan la existencia de sobre-dispersión, se cree que se cuenta con información suficiente para decretar que el modelo de regresión Poisson no es el adecuado para los datos. En los siguientes incisos, se buscará atacar el problema de sobre-dispersión en los datos.

c) Haciendo uso del comando glm de R, se obtuvo el resumen presentado en la tabla 7 del modelo en (15), utilizando esta vez una familia quasi-poisson:

[18] Que define simplemente como el producto de las covaraibles. Se aprovecha este apartado para destacar que cuando se ajusto este modelo, el coeficiente para dicho término de interacción obtuvo un p-valor de 0.82402, tan alto que se omitió de este y los siguientes modelos. Ya que os mismos solo tenderán a inflar dicho p-valor.

[19] Ver Dunn y Smyth (2018) sección 8.8.3.

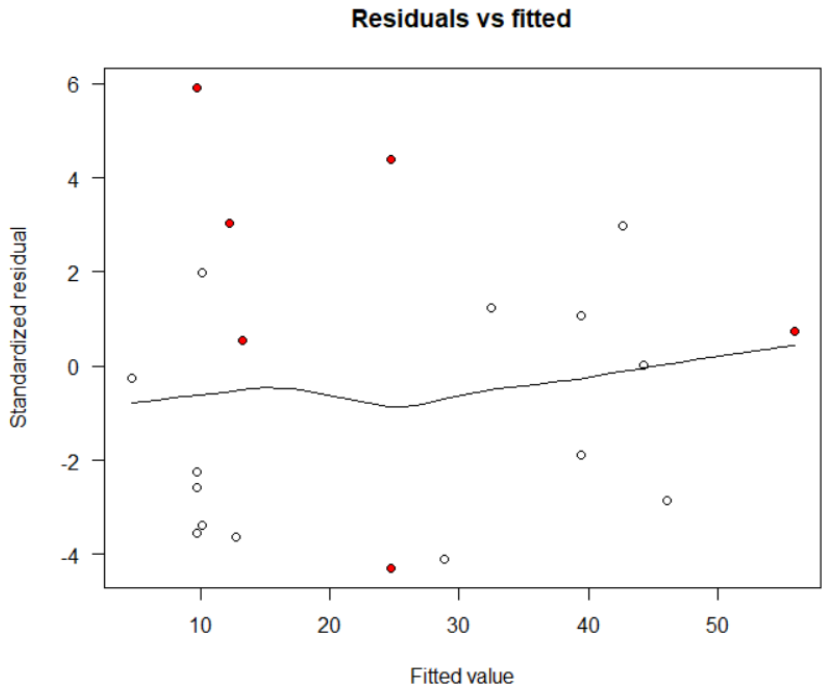

Figura 10: Gráfica de dispersión de los residuales de la Deviance estándarizados contra los valores ajustados por el modelo. Se destaca en color rojo los residuales de las observaciones bajo sospecha.

Coeficiente	Estimación	t-valor	p-valor
β_0	3.1699	5.754	$2.34 \cdot 10^{-5}$
β_1	-0.0388	-1.991	0.0628
β_2	1.3591	3.527	0.0030

Cuadro 7: Análisis de regresión para el modelo en (15) suponiendo familia cuasi-poisson y liga logarítmica.

Y una estimación del parámetro de dispersión de:[20]

$$\hat{\phi} = 10.728, \tag{16}$$

por lo que, los errores estándar se inflan por un factor de $\sqrt{\hat{\phi}} = 3.275$. Como puede notar, dichos errores estándar para las estimaciones de los coeficientes se omiten, debido a que una vez teniendo los tres modelos solicitados en este ejercicio se hará un comparativo de los mismos. Sin embargo, es importante destacar que el coeficiente β_1, aquel que pre-multiplica a la covariable edad del paciente, deja de ser significativamente distinto de cero bajo un nivel de significancia del 5 %, algo que reafirma la existencia de sobre-dispersión en los datos, ya que en el modelo de regresión Poisson este coeficiente si era significativo. A pesar de ello, se resalta que dicho coeficiente sigue manteniendo un p-valor relativamente cercano a 0.05, por lo que, a priori tampoco existen motivos para quitar la edad como covariable del modelo. Por último, dado que se intentará comparar este modelo con el que será calculado en el inciso c), y a sabiendas de que este modelo no posee un AIC por no estar basado en ningún modelo de probabilidad, se seguirá lo dicho en Dunn y Smyth (2018) subsección 10.5.3, en donde se menciona que si el modelo quasi-poisson es correcto, entonces los residuales de la deviance estandarizados deberían de tener aproximadamente una distribución normal de media cero y varianza constante uno, con ello en mente, se presenta en la fila superior de la figura 11, las siguientes gráficas: Una gráfica QQ-normal para dichos residuales y la gráfica de los valores ajustados por el modelo quasi-poisson contra dichos residuales. Ahora, observe que en la gráfica QQ se puede apreciar un ajuste razonable para hacia la cola derecha de la distribución normal estándar, sin embargo, el ajuste hacia la cola izquierda no parece ser óptimo. Sin embargo, en la gráfica de residuales contra valores ajustados si parece existir un patrón de ruido blanco, además en esta gráfica se nota de manera global, que los residuales son más pequeños que los del modelo Poissón, por lo que parece existir un mejor ajuste. Finalmente, la tercer gráfica que completa la primer fila de gráficos en la figura 11 es una gráfica de las distancias de Cook, recuerde que una observación se considera influyente bajo esta métrica cuando el valor de su distancia, es mayor que la mediana de una ditribución F con p y $n-p$ grados de libertad,[21] en es este caso dicha mediana tiene un valor de 0.8212, por lo que ninguna observación se considera influyente bajo esta métrica, con lo que se ha resuelto el problema sobre las cuatro observaciones que posiblemente eran influyentes bajo este criterio, en el modelo de regresión Poisson.[22] Como comentario adicional, este conjunto de tres gráficas se denomina gráficas de diagnóstico de acuerdo a Dunn y Smyth (2018), dado este comentario, se concluye que el modelo quasi-Poisson presenta un ajuste razonable bajo las gráficas

[20] Note que esto es otra señal de sobre-disperción, ya que la estimación del parámetro de dispersión ϕ es mayor 1.

[21] Esto se sostiene para GLM's ver capítulo 8 Dunn y Smyth (2018). En nuestro caso $p = 3$ número de parámetros en el modelo y $n = 20$ el número de observaciones.

[22] Se hace el comentario adicional, de que siguen existiendo dos observaciones influyentes bajo la $COVRATIO$, sin embargo a esta medida no le prestan demasiada atención los autores citados.

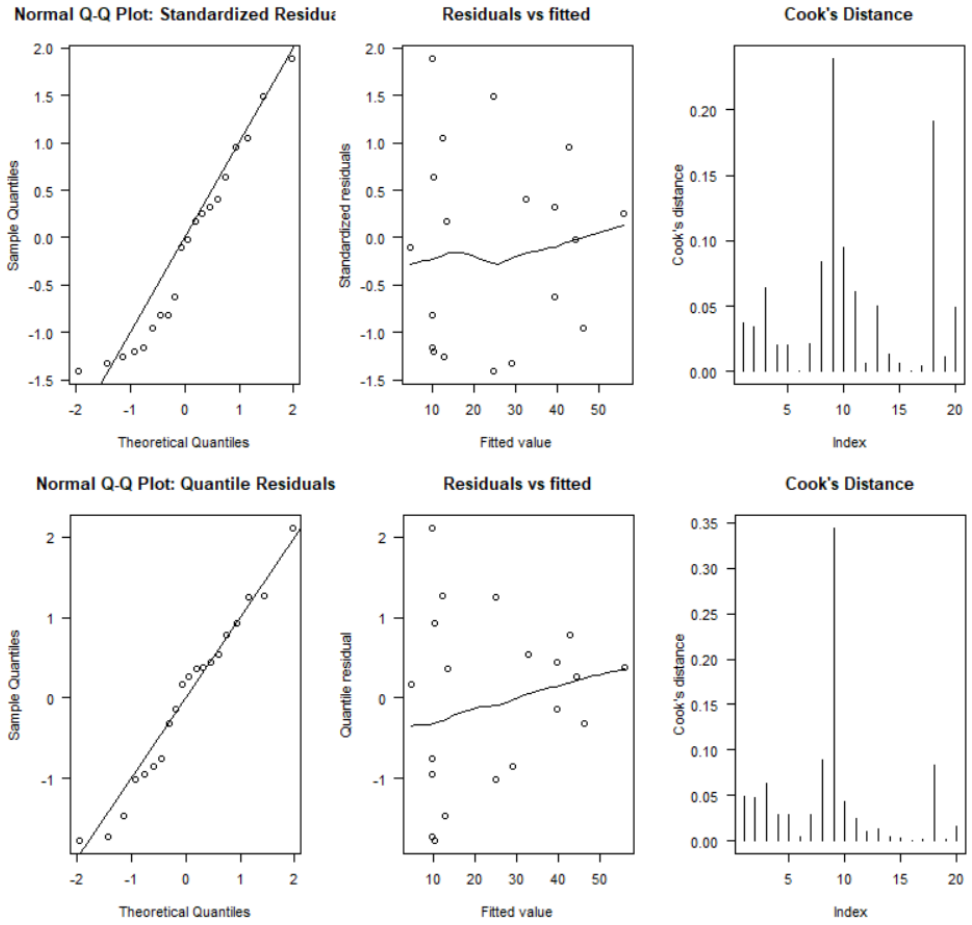

Figura 11: Gráficas de diagnósticos para los Modelos Cuasi-Poisson y Binomial Negativo.

de diagnóstico.

d) Por último, se ajustó el modelo en (15) haciendo uso de la familia binomial negativa, por medio del comando *glm.nb* de *R* los resultados se presentan en la tabla 8:

Coeficiente	Estimación	z-valor	p-valor
β_0	3.1579	5.6641	$1.478 \cdot 10^{-8}$
β_1	-0.0386	-1.8401	0.0658
β_2	1.3681	3.7073	$2.094 \cdot 10^{-4}$

Cuadro 8: Análisis de regresión para el modelo en (15) suponiendo familia binomial negativa y liga logarítmica.

Nuevamente, note que los errores estándar de este modelo han sido omitidos, sin embargo, observe que haciendo uso de esta familia también se tiene que el parámetro β_1 resulta no ser significativo, bajo un nivel de significancia del 5 %, hecho que resalta la sobre-dispersión existente en los datos, pues como ya se comento con anterioridad, este coeficiente resultaba muy significativo en el modelo de regresión Poisson originalmente planteado. Por otro lado, en la segunda fila de gráficas de la figura 11 podemos observar las gráficas de diagnóstico para este modelo, se destaca que en este caso se sugiere en Dunn y Smyth (2018) el hacer uso de los residuales de cuantiles,[23] para construir las mismas, ya que estos se distribuyen aproximadamente como una normal de media cero y varianza constante uno, en caso de que el modelo Binomial Negativo propuesto[24] sea un modelo razonable. De este modo, comencemos con el análisis de dichas gráficas. Primeramente, observe que la gráfica QQ parece mostrar un ajuste razonable de los residuales de cuantiles a la distribución normal estándar, incluso el mal ajuste hacia la cola izquierda de la distribución, que se notó en la gráfica de los residuales de la deviance estandarizados para el modelo quasi-poisson, parece verse disminuido, sin embargo, en la gráfica de residuales vs valores ajustados parece prevalecer cierta tendencia, algo que no coincide con la dispersión esperada en esta gráfica de tratarse de algo que es aproximadamente ruido blanco, sin embargo, en esta gráfica también se nota que de manera general, los residuales resultan más pequeños que los del modelo Poisson, por lo que parece existir un mejor ajuste. Finalmente, para analizar la gráfica de la D de Cook, recuerde que nuestro valor de corte es la mediana de una distribución F con 3 y 17 grados de libertad, y que ya se había comentado que dicha mediana tenía un valor de 0.8212, por lo que, al igual que en el modelo quasi-poisson, se nota que se ha eliminado el problema de la influencia que tenían ciertas observaciones en el modelo de regresión Poisson. De este modo, se concluye que el modelo binomial negativo también presenta un ajuste razonable bajo el criterio de las gráficas de diagnóstico. En el último inciso, se presentará cual de estos dos modelos se piensa podría ser mejor para modelar los datos.

e) Antes de comenzar con la selección del modelo, se deja en las tablas 9 y 10 las estimaciones de los coeficientes obtenidas por los distintos modelos probados en este ejercicio, así como sus errores estándar:

[23] Ver sección 8.3.4 de Dunn y Smyth (2018) para una definición de estos residuales.

[24] Los residuales de la deviance estandarizados también tienden a hacerlo, pero en la referencia mencionada se dice que en el caso de modelos discretos es preferible usar estos residuales. Quizá, aquí se pregunte el porque no se usaron estos residuales con el modelo quasi-poisson, y la respuesta es sencilla, no están definidos debido a que dicho modelo no se basa en ningún modelo de probabilidad.

GLM	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Poisson glm	3.170	-0.0388	1.359
Quasi-Poisson	3.170	-0.0388	1.359
Neg bin glm	3.158	-0.0386	1.368

Cuadro 9: Estimaciones de los coeficientes obtenidas con los diversos modelos ajustados.

Coeficiente	se.Poisson	se.Quasi-poisson	se.Quasi-poisson/se.Poisson	se.NB
β_0	0.168	0.551	3.275	0.558
β_1	0.006	0.020	3.275	0.021
β_2	0.118	0.385	3.275	0.369

Cuadro 10: Errores estándar de los coeficientes obtenidas con los diversos modelos ajustados.

Observe que las estimaciones de los coeficientes para los tres modelos son muy similares,[25] sin embargo, los errores estándar solo resultan parecidos en los modelos quasi-poisson y binomial negativo, ya que como era de esperarse estos errores estándar resultan mayores a aquellos en el modelo Poisson, debido a que este último no considera la sobre-dispersión en los datos. Ahora, dado que los errores estándar y las estimaciones de los coeficientes del modelo son muy similares, bajo los modelos binomial negativo y quasi-poisson, se tendería a pensar que dichos modelos son prácticamente equivalentes para modelar los datos que nos fueron dados, sin embargo, para descartar este pensamiento observe la gráfica en la figura 12. En dicha figura, se aprecia en el lado izquierdo, una gráfica de valores ajustados e intervalos de confianza al 95 % generados con el modelo quasi-Poisson, en color rojo se muestran los valores ajustados y los intervalos mencionados cuando la covariable grupo es igual a 1, en azul lo análogo pero con la covariable grupo igual a 0, también se agregan las observaciones del número de polipos siguiendo la misma lógica en los colores. Mientras que, en el lado derecho se presenta la gráfica análoga para el modelo binomial negativo. Ahora, observe que pese a que los valores ajustados por ambos modelos son prácticamente iguales, hay una diferencia en los intervalos de confianza significativa, pues el modelo binomial negativo tiende a generar intervalos de confianza mas anchos para el grupo 1, mientras que, pese a que esto es un poco más difícil de notar, el modelo quasi-poisson tiende a generar intervalos de confianza mas anchos para el grupo 0. Se resalta además, que en ambos casos el ajuste parece razonable, cosa que ya se había constatado en las gráficas de diagnostico de ambos modelos. Ahora, con todos estos elementos y dado que se tiene que elegir uno de estos modelos, me decantaría por el modelo binomial negativo, ya que este es un modelo basado en un modelo de probabilidad[26], que por lo tanto tendrá un AIC bien definido y podría ser comparado de manera más sencilla con otros modelos que pudieran proponerse, además, de que en la gráfica QQ de sus residuales se vio un mejor ajuste de los mismos a la distribución normal estándar y que viéndonos conservadores, genera los intervalos de confianza más amplios para el grupo que parece tener mayor variabilidad, lo que podría ser una ventaja considerando la sobre-dispersión en los datos.

Observación 3. *Finalmente se hace la siguiente observación, bajo los modelos cuasi-poisson y*

[25] Inclusive resultan iguales en para los modelos de regresión Poisson y Quasi-poisson, como era de esperarse.
[26] Con el que además se puede modelar la asimetría de las colas.

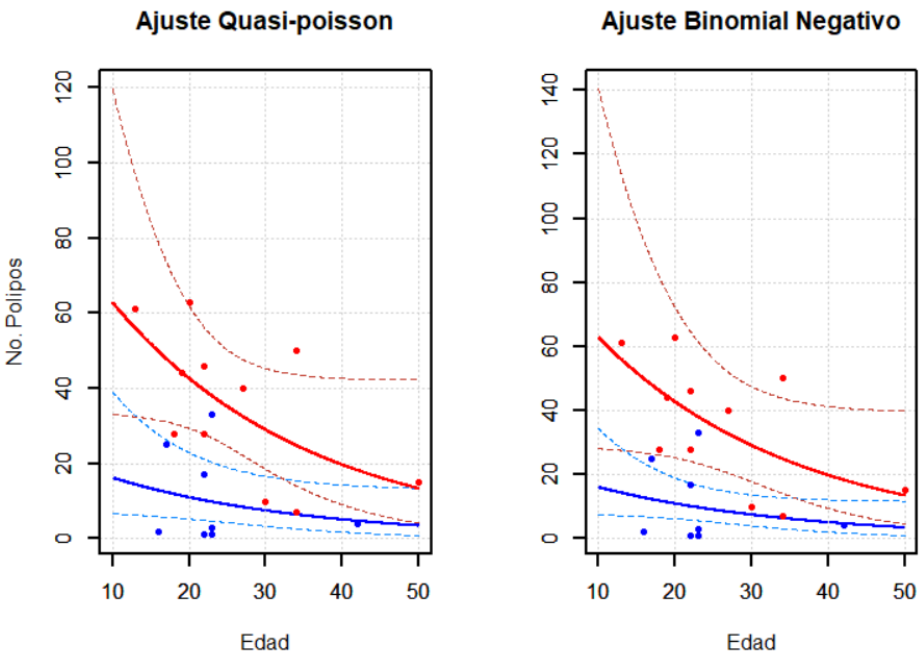

Figura 12: Gráfica en el lado izquierdo, valores ajustados e intervalos de confianza al 95 % confianza del modelo quasi-Poisson, en color rojo se muestran los valores ajustados y los intervalos mencionados cuando la covariable grupo es igual a 1, en azul lo análogo pero con la covariable grupo igual a 0. En el lado derecho, se encuentra la gráfica análoga para el modelo binimial negativo.

binomial negativo se observó que el coeficiente que pre-multiplica a la covariable edad, no resulta ser significativamente distinto de cero, bajo un nivel de significancia del 5 %*, sin embargo, el p-valor bajo mi criterio no resultaba suficientemente elevado como para aceptar la hipótesis de que este valor fuese igual a cero, y por ende remover el coeficiente del modelo. Sin embargo, en está última observación se añade que el modelo binomial negativo, que no considera dicho coeficiente consigue un AIC de* 165.8*, mientras que el modelo elegido en el inciso anterior posee un AIC de* 164.88 *y pese a que la diferencía en dichos criterios es pequeña, resulta favorable al modelo seleccionado.* △

■

1. Gracias Irving por aguantar mis dudas.

Referencias

Dunn, P. K. & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R* (1st ed. 2018). Springer New York : Imprint: Springer. https://doi.org/10.1007/978-1-4419-0118-7