Modelos Estadísticos I (Fecha: 28/05/2021)

Tarea #8

Estudiante: Roberto Vásquez Martínez *NUA:* 424662

Problema 1

Considere una muestra aleatoria con $Y_1, \ldots, Y_N$ con la distribución exponencial

$$f(y_i; \theta_i) = \theta_i \exp\{-\theta_i y_i\}.$$

Obtenga la deviance comparando el modelo maximal con diferentes valores de θ_1 para cada Y_i, y el modelo con $\theta_i = \theta$ para toda i.

(Solución)

(a) Primero supongamos el caso con diferentes valores de θ_i para cada $i = 1, 2, \ldots, N$.

Notamos que

$$\begin{aligned} f(y_i; \theta_i) &= \exp\left\{\frac{\theta_i y_i - \log\theta_i}{-1}\right\} \\ &= \exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a(\varphi)} - c(y_i, \varphi)\right\}, \end{aligned}$$

por lo que

$$\begin{aligned} b_i(\theta_i) &= \log\theta_i \\ a_i(\varphi) &= -1 \\ c_i(y_i, \varphi) &= 0. \end{aligned}$$

En este caso el parámetro de dispersión es $\varphi = 1$, como consideramos

$$a_i(\varphi) = \frac{\varphi}{w_i},$$

luego $w_i = -1$ para cada $i = 1, 2, \ldots, N$.

Sea $\tilde{\theta}_i$ es la estimación para el modelo maximal y $\hat{\theta}_i$ la estimación para el modelo seleccionado tenemos que

$$\hat{\mu}_i = \mathbb{E}[Y_i] = \frac{1}{\hat{\theta}_i} \quad \Rightarrow \quad \hat{\theta}_i = \frac{1}{\hat{\mu}_i},$$

y considerando $y_1, \ldots, y_N$ realizaciones de las variables $Y_1, \ldots, Y_N$

$$\tilde{\theta}_i = \frac{1}{y_i}.$$

Por lo tanto la deviance viene dada por

$$\begin{aligned} D =& 2 \cdot \left[\sum_{i=1}^{N} w_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right] \right] \\ =& -2 \cdot \left[\sum_{i=1}^{N} y_i \left(\frac{1}{y_i} - \frac{1}{\hat{\mu}_i} \right) - \log \frac{1}{y_i} + \log \frac{1}{\hat{\mu}_i} \right] \\ =& 2 \cdot \left[\sum_{i=1}^{N} y_i \left(\frac{1}{\hat{\mu}_i} - \frac{1}{y_i} \right) + \log \frac{1}{y_i} - \log \frac{1}{\hat{\mu}_i} \right] \\ =& 2 \cdot \left[\sum_{i=1}^{N} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) - \log \frac{y_i}{\hat{\mu}_i} \right], \end{aligned}$$

que es lo que queríamos obtener.

(b) Ahora obtendremos la deviance en el caso en el cual $\theta_i = \theta$ para toda $i = 1, 2, \ldots, N$.

En este caso se tiene que $\hat{\theta}_i = \frac{1}{\hat{\mu}}$ donde

$$\mathbb{E}[Y_i] = \mu,$$

pues $Y_1, \ldots, Y_N$ son variables aleatorias idénticamente distribuidas exponencial con parámetro θ.

Del inciso anterior tenemos que en este caso la deviance viene dada por

$$\begin{aligned} D =& 2 \cdot \left[\sum_{i=1}^{N} \left(\frac{y_i - \hat{\mu}}{\hat{\mu}} \right) - \log \frac{y_i}{\hat{\mu}} \right] \\ =& 2 \cdot \left[\sum_{i=1}^{N} \left\{ \frac{y_i}{\hat{\mu}} \right\} - N - \sum_{i=1}^{N} (\log y_i - \log \hat{\mu}) \right] \\ =& 2 \cdot \left(\sum_{i=1}^{N} \left\{ \frac{y_i}{\hat{\mu}} - \log y_i \right\} - (N + N \log \hat{\mu}) \right), \end{aligned}$$

que es lo que queríamos obtener □

Problema 2

Sea $l(\mathbf{b}_{\min})$ el valor máximo de la función de logverosimilitud para el modelo minimal con predictor lineal $x^T\beta = \beta_1$ y consideremos un modelo más general con predictor lineal $x^T\beta = \beta_1 + \beta_2 x_1 + \cdots + \beta_p x_{p-1}$.

(a) Pruebe que la estadística chi-cuadrada es

$$C = 2[l(\mathbf{b}) - l(\mathbf{b}_{\text{mín}})] = D_0 - D_1,$$

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

(Solución)
Sabemos que según la prueba de razón de verosimilitud la estadística C a la que se refiere el problema es en efecto

$$C = 2[l(\mathbf{b}) - l(\mathbf{b}_{\text{mín}})],$$

en donde se compara el modelo más general con el modelo minimal.

Si $l(\mathbf{b}_{\max})$ es la logverosimilitud del modelo maximal entonces la deviance para el modelo minimal es

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

mientras que la deviance para el modelo más general es

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

Por lo tanto

$$\begin{aligned} D_0 - D_1 =& 2\left[l(\mathbf{b}_{\text{máx}}) - l(\mathbf{b}_{\text{mín}})\right] - 2\left[l(\mathbf{b}_{\text{máx}} - l(\mathbf{b})\right] \\ =& 2\left[l(\mathbf{b}) - l(\mathbf{b}_{\text{mín}})\right] \\ =& C, \end{aligned}$$

que es lo que queríamos ver.

(b) Deducir que si $\beta_2 = \beta_3 = \cdots = \beta_p = 0$, entonces C tiene la distribución chi-cuadrada central con $(p-1)$ grados de libertad.

(Solución)
Notamos que si β es el verdadero valor de los parámetros en el modelo con p parámetros y $\mathbf{b}$ es la estimación de máxima verosimilitud entonces

$$Z_1 = 2\left[l(\mathbf{b}) - l(\beta)\right] \sim \chi^2_p \text{ asintóticamente,} \tag{1.1}$$

de manera similar si $\beta_{\text{mín}}$ es el verdadero valor del los parámetros en el modelo minimal entonces

$$Z_2 = 2\left[l(\mathbf{b}_{\text{mín}}) - l(\beta_{\text{mín}})\right] \sim \chi^2_1 \text{ asintóticamente.} \tag{1.2}$$

Como $\beta_2 = \beta_3 = \cdots = \beta_p = 0$, entonces el modelo con p parámetros en realidad tiene sólo un parámetro, por lo que $l(\beta) = l(\beta_{\text{mín}}) = 0$, luego

$$\begin{aligned} Z_1 - Z_2 =& 2\left[l(\mathbf{b}) - l(\beta)\right] - 2\left[l(\mathbf{b}_{\text{mín}}) - l(\beta_{\text{mín}})\right] \\ =& 2\left[l(\mathbf{b}) - l(\mathbf{b}_{\text{mín}})\right] + 2\left[l(\beta_{\text{mín}}) - l(\beta)\right] \\ =& 2\left[l(\mathbf{b}) - l(\mathbf{b}_{\text{mín}})\right] \\ =& C. \end{aligned}$$

De lo anterior, (1.1) y (1.1) concluimos que

$$C \sim \chi^2_{p-1} \text{ aproximadamente,}$$

que es lo que queríamos probar. □

Problema 3

(Solución)
El modelo que proponemos para describir la relación entre la radiación y la tasa de mortalidad es el GLM Binomial. Tenemos 6 grupos, sea Y_i el número de casos de leucemia en cada grupo y n_i el total de observaciones por grupo con $i = 1, 2, \ldots, 6$. Si $P_i = Y_i/n_i$ y $Y_i \sim \text{Bin}(n_i, \pi_i)$ dado la variable explicativa X_i, entonces nuestro modelo tiene el siguiente predictor lineal

$$g(\pi_i) = x_i^T \beta,$$

en donde vamos a considerar x_i un vector de variables dummy. Si X tiene como renglones a los vectores x_i, codificaremos la pertenencia a cada uno de los grupos con la siguiente matriz.

$$X = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix}$$

En clase vimos, que las funciones liga para este tipo de estudios clínicos usuales podrían ser logit, probit y clog-log probaremos con cada una de estas ligas, cabe aclarar que el predictor lineal que estamos considerando tiene intercepto, es decir a la matriz de diseño le agregamos una columna de 1's.

El resultado para la liga logit fue

Dep. Variable:	['y1', 'y2']	**No. Observations:**	6
Model:	GLM	**Df Residuals:**	2
Model Family:	Binomial	**Df Model:**	3
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-11.654
Date:	Fri, 28 May 2021	**Deviance:**	1.6421
Time:	22:24:27	**Pearson chi2:**	1.58
No. Iterations:	6		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-3.5717	0.247	-14.442	0.000	-4.056	-3.087
x1	-0.1172	0.516	-0.227	0.820	-1.128	0.894
x2	0.5444	0.230	2.368	0.018	0.094	0.995
x3	0.8903	0.212	4.201	0.000	0.475	1.306

El resultado para la liga probit fue

Dep. Variable:	['y1', 'y2']	**No. Observations:**	6
Model:	GLM	**Df Residuals:**	2
Model Family:	Binomial	**Df Model:**	3
Link Function:	probit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-12.064
Date:	Fri, 28 May 2021	**Deviance:**	2.4625
Time:	22:27:59	**Pearson chi2:**	2.36
No. Iterations:	7		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.9340	0.109	-17.665	0.000	-2.149	-1.719
x1	-0.0366	0.218	-0.168	0.867	-0.463	0.390
x2	0.2788	0.119	2.341	0.019	0.045	0.512
x3	0.4525	0.105	4.307	0.000	0.247	0.658

Finalmente el resultado para la liga clog-log fue

Dep. Variable:	['y1', 'y2']	**No. Observations:**	6
Model:	GLM	**Df Residuals:**	2
Model Family:	Binomial	**Df Model:**	3
Link Function:	cloglog	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-11.529
Date:	Fri, 28 May 2021	**Deviance:**	1.3923
Time:	22:29:25	**Pearson chi2:**	1.34
No. Iterations:	7		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-3.5712	0.241	-14.822	0.000	-4.043	-3.099
x1	-0.1301	0.508	-0.256	0.798	-1.126	0.866
x2	0.4989	0.212	2.351	0.019	0.083	0.915
x3	0.8403	0.202	4.158	0.000	0.444	1.236

Nos quedamos con aquella que tiene el mejor comportamiento respecto a la deviance que es con la liga clog-log. A continuación graficamos la media estimada para le proporción de casos de leucemia para cada uno de los grupos, así como las proporciones observadas.

en la que se observa un buen ajuste. □

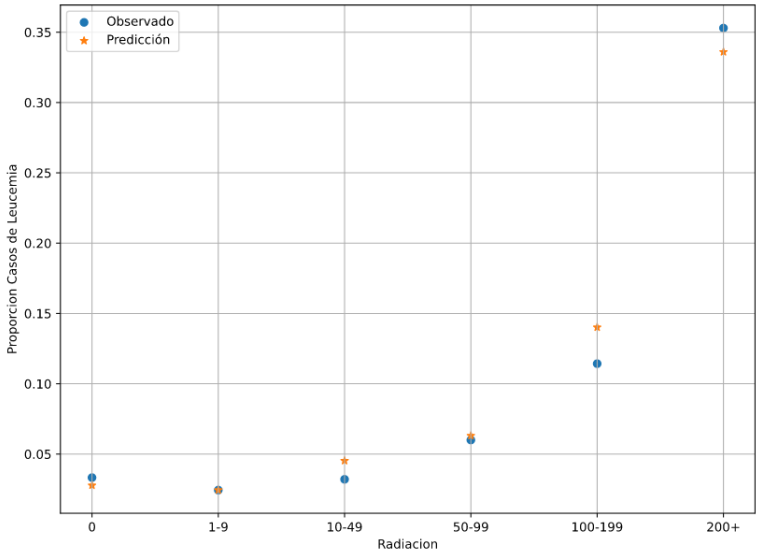

Figura 1: Proporción de casos de leucemia sobrevivientes bomba de Hiroshima

Problema 4

(Solución)
El propósito es determinar las expresiones que permiten aplicar el el método de scoring de Fisher, equivalente a IRLS. Sea f la función de probabilidad Poisson con media μ, entonces

$$f(y;\mu) = \exp\{y\log\mu - \mu - \log y!\} \text{ para } y = 0, 1, \ldots,$$

Sea $\theta = \log\mu$, en este caso consideramos el parámetro de dispersión como $\varphi = 1$. Si $b(\theta) = \exp\theta = \mu$, $a(\varphi) = 1$ y $c(y,\varphi) = -\log y!$, entonces

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

y ya que tenemos esta expresión tenemos las formas vistas en clase para la media y varianza, luego

$$V(Y) = b''(\theta) = \exp\{\theta\} = \mu.$$

Usando la liga canónica tenemos el predictor lineal siguiente

$$g(\mu) = \eta = \theta = \log\mu,$$

entonces $\frac{d\eta}{d\mu} = \frac{1}{\mu}$.

Sabemos que el proceso iterativo del scoring de Fisher se reduce a

$$\beta^{(k+1)} = (X^T DWDX)^{-1} X^T DWD(\eta^{(k)} + D^{-1}(y - \mu^{(k)})),$$

donde $D = \text{diag}(1/g'(\mu_i))$, $W = \text{diag}(1/V(Y_i))$, X es la matriz de diseño y $\eta^{(k)}, \mu^{(k)}$ son los vectores que encapsulan al predictor lineal y la media de cada variable respuesta.

En el caso de la distribución Poisson y con los cálculos ya hechos tenemos que $D = \text{diag}(\mu_i)$, $W = \text{diag}(1/\mu_i)$, por lo que si $z^{(k)} = (\eta^{(k)} + D^{-1}(y - \mu^{(k)}))$, el método de scoring de Fisher se reduce a

$$\beta^{(k+1)} = (X^T D X)^{-1} X^T D z^{(k)},$$

que es el problema de IRLS.

El ejemplo planteado lo vimos en la ayudantía, la distribución propuesta es precisamente la Poisson, y como vimos el predictor lineal adecuado es de la forma

$$\eta = g(\mu) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Calculamos intervalos de confianza y de predicción a través de la estimación asíntotica de la varianza para los estimadores y usando cuantiles Poisson con un nivel de significancia de $\alpha = 0.05$. Para poder graficar en la escala de los casos de SIDA en Bélgica usamos el Método Delta. Obtenemos el siguiente ajuste

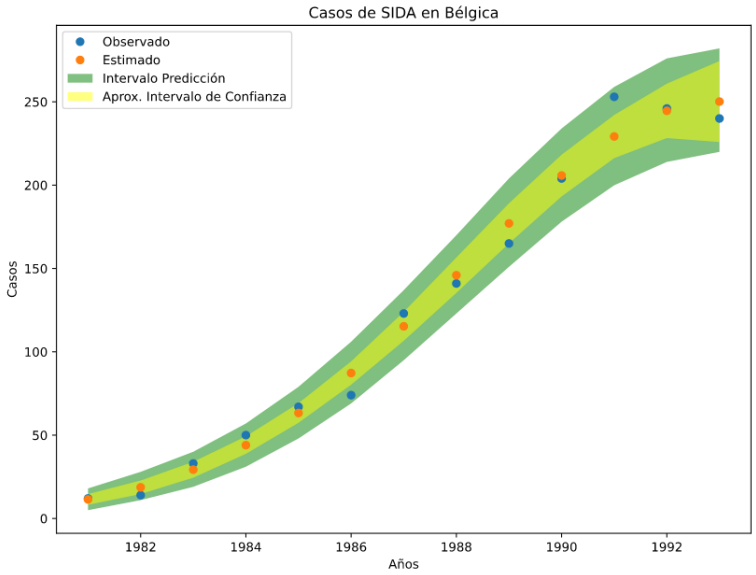

Figura 2: Casos de SIDA en Bélgica

Problema 5

(a)

(Solución)

Hacemos las gráficas de los datos para cada grupo

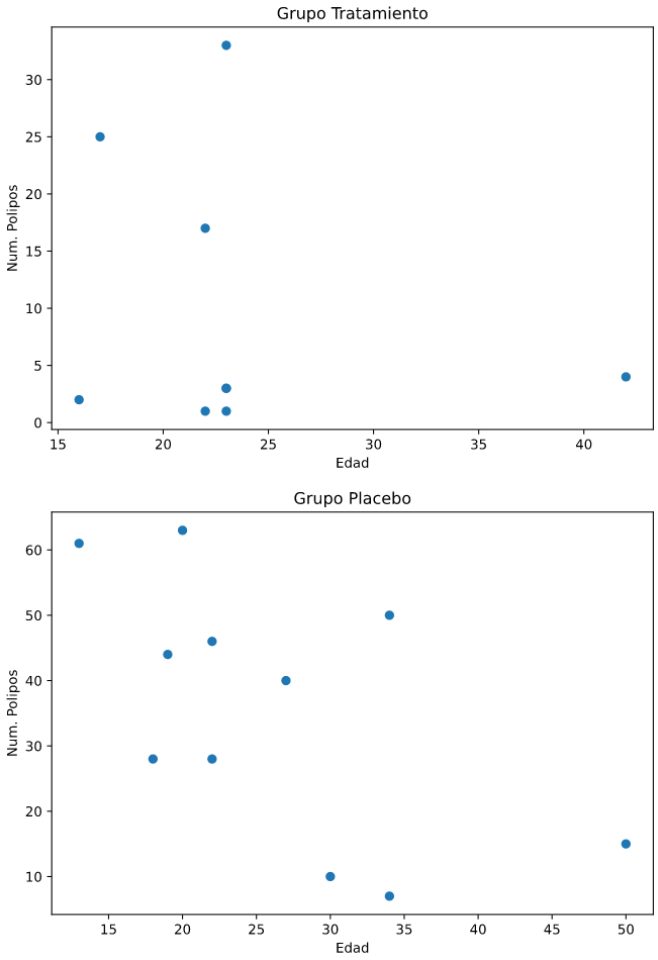

Figura 3: Número de Polipos respecto a la edad

En esta gráfica observamos gran disparidad entre el número de polipos en pacientes de edades similares lo que podría indicar sobredispersión.

(b)

Ajustaremos a cada grupo el modelo de dispersión Poisson.

Para el grupo en tratamiento tenemos lo siguiente

Dep. Variable:	y	**No. Observations:**	9
Model:	GLM	**Df Residuals:**	7
Model Family:	Poisson	**Df Model:**	1
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-66.240
Date:	Fri, 28 May 2021	**Deviance:**	101.53
Time:	23:38:37	**Pearson chi2:**	108.
No. Iterations:	5		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	3.2619	0.446	7.314	0.000	2.388	4.136
x1	-0.0430	0.020	-2.166	0.030	-0.082	-0.004

Y para el grupo con un placebo es

Dep. Variable:	y	**No. Observations:**	11
Model:	GLM	**Df Residuals:**	9
Model Family:	Poisson	**Df Model:**	1
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-67.673
Date:	Fri, 28 May 2021	**Deviance:**	77.958
Time:	23:39:15	**Pearson chi2:**	74.3
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	4.5191	0.153	29.467	0.000	4.219	4.820
x1	-0.0384	0.006	-6.149	0.000	-0.051	-0.026

Podemos ver que en ambos modelos la deviance resulta mucho mayor a los grados de libertad residuales esto indica o que el modelo no es el adecuado o un problema de sobredispersión.

Además graficando la varianza de cada observación respecto a la media estimada se tiene lo siguiente

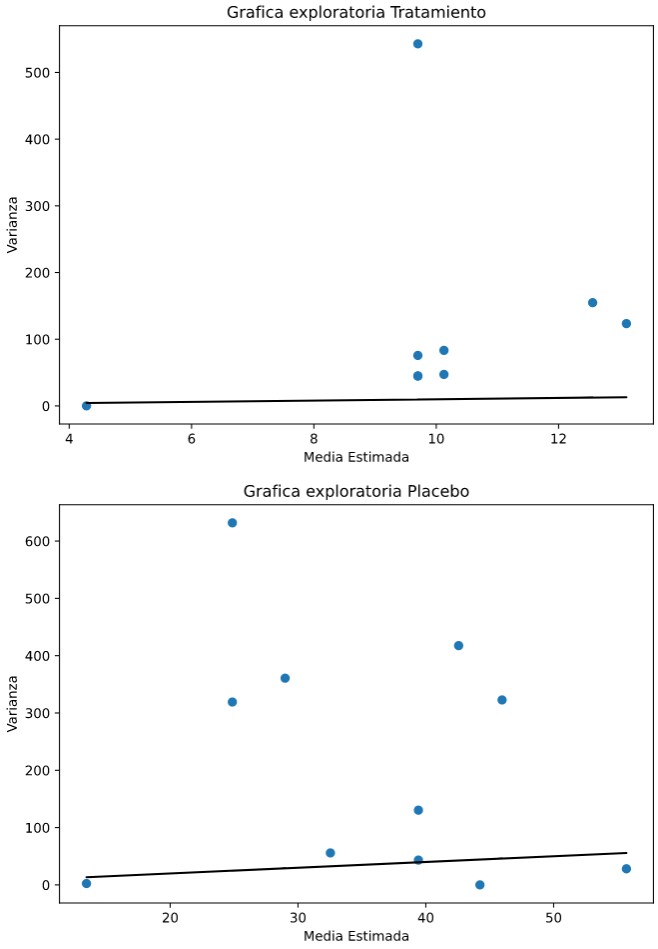

Figura 4: Varianza respecto a la media

Aquí observamos que la varianza es mayor que la media, lo que da evidencia de un problema de sobredispersión en el modelo Poisson GLM.

(c)

(Solución)
A continuación ajustaremos un modelo quasi-Poisson, las estimaciones serán iguales que en el Poisson GLM pero los errores estándar de los parámetros serán diferentes.

Para el grupo en tratamiento obtenemos

```
    Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.834  -3.585  -2.522   1.967   5.849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.26186    1.74929   1.865    0.104
Edad        -0.04303    0.07791  -0.552    0.598

(Dispersion parameter for quasipoisson family taken to be 15.38385)

    Null deviance: 107.22  on 8  degrees of freedom
Residual deviance: 101.53  on 7  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Y para el grupo placebo

```
    Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2406  -2.3890   0.4149   1.1421   4.4258

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.51912    0.44072   10.25  2.9e-06 ***
Edad        -0.03840    0.01795   -2.14    0.061 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 8.258141)

    Null deviance: 121.341  on 10  degrees of freedom
Residual deviance:  77.958  on  9  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

En estos ajuste la deviance parece no ajustar a la χ^2 correspondiente.

(d)

A continuación ajustamos el modelo binomial negativo.

Para el grupo en tratamiento

```
    Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4174  -1.2640  -0.8597   0.5111   1.3942

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.31314    1.32718   2.496   0.0125 *
Edad        -0.04532    0.05491  -0.825   0.4091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.8763) family taken to be 1)

    Null deviance: 10.2918  on 8  degrees of freedom
Residual deviance:  9.7148  on 7  degrees of freedom
AIC: 65.398
```

y en el caso del placebo

```
    Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9804  -0.7710   0.1237   0.3745   1.5305

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.47737    0.43592  10.271   <2e-16 ***
Edad        -0.03671    0.01586  -2.315   0.0206 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(4.5273) family taken to be 1)

    Null deviance: 16.862  on 10  degrees of freedom
Residual deviance: 11.651  on  9  degrees of freedom
AIC: 97.82

Number of Fisher Scoring iterations: 1
```

A diferencia de los ajustes con el modelo quasi-Poisson aquí si parece que la deviance residual tiene el comportamiento asíntotico χ^2 deseado.

(e)

A continuación mostramos ambos ajustes gráficamente. La linea sólida es el modelo quasi-Poisson mientras que la línea segmentada es el modelo binomial negativo.

Para el grupo en tratamiento

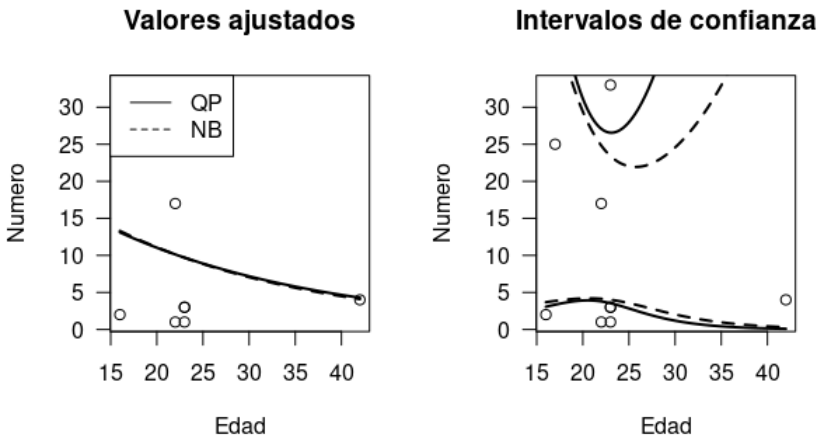

Figura 5: Ajuste grupo en tratamiento

Y para el grupo placebo tenemos lo siguiente

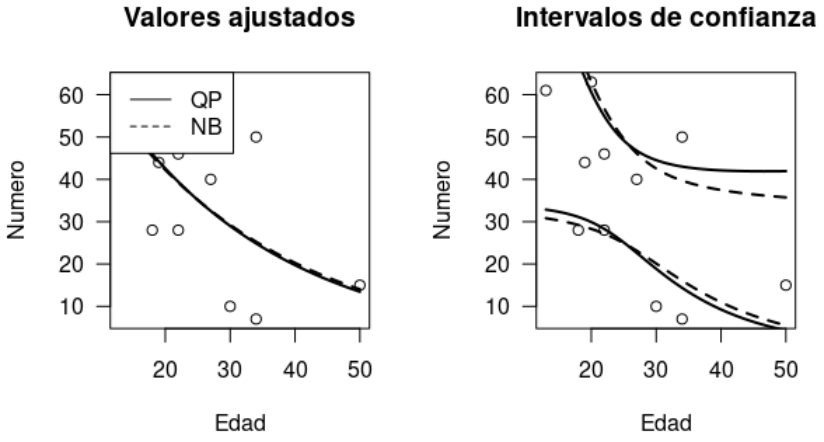

Figura 6: Ajuste grupo placebo

Observamos que el ajuste es similar en ambos casos, en el grupo placebo hay un mejor comportamiento de los intervalos de confianza en el modelo binomial negativo. Por otro lado, ya hemos mencionado que la deviance residual se comporta mejor en el caso binomial negativo y adicionalmente la edad es significativa para el grupo placebo del modelo binomial negativo. En base a estas observaciones elegimos como mejor opción al modelo binomial negativo. □