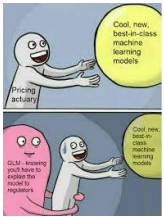
Estadística Aplicada III

Modelos lineales generalizados

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 7

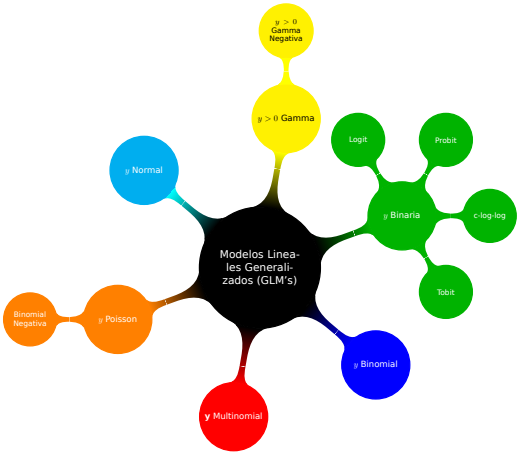


Definición y características de los Modelos Lineales Generalizados

- Estos modelos extienden a los modelos lineales para acomodar respuestas no normales y transformaciones a linealidad.
- Los modelos fueron desarrollados en 1972 por John Nelder y Robert Wedderburn, ampliando considerablemente el alcance de los modelos normales.
- También representaron un hito en el uso de los recursos computacionales utilizando un método similar al de Newton-Raphson, *mínimos cuadrados ponderados iterativos (IWLS)* mucho más eficiente que utilizar los métodos usuales de máxima verosimilitud.

Modelos Lineales Generalizados

- Estos modelos tienen por objeto modelar de manera unificada variables de respuesta categóricas, y continuas, con ciertas distribuciones específicas (aquellas que pertenecen a la familia exponencial) a través de su dependencia a combinaciones lineales de variables predictoras o de respuesta.



Características de los GLM's

1. **Supuesto distribucional**: Tenemos observaciones $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, con variables $\mathbf{x}_i \in \mathbb{R}^p$ y queremos estudiar $y_i|\mathbf{x}_i$. En general, se supone que $y|\mathbf{x}$ pertenece a una familia exponencial simple. La dependencia de y en $\mathbf{x}$ es a través de combinaciones lineales $\beta'\mathbf{x}$ como funciones de la media condicional, $\mathrm{E}(y_i|\mathbf{x}_i) = h(\beta'\mathbf{x}) = \mu_i$.
2. **Supuesto estructural**: La esperanza μ_i se relaciona al predictor lineal $\eta_i = \beta_i\mathbf{x}_i$ a través de una función h:

$$\mu_i = h(\eta_i) = h(\beta'\mathbf{x}_i) \text{ o } \eta_i = g(\mu_i)$$

donde:

- h es la *función media kernel*, una función uno a uno , suficientemente suave.
- g es la *función liga*, inversa de h

La distribución de y tiene la forma siguiente:

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

o

$$\log f(y_i|\theta_i, \phi, w_i) = \frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)$$

donde:

- θ_i es el *parámetro natural*, función de la media μ: $\theta_i = \theta(\mu_i)$ y se determina de manera única a través de la relación $\mu = b'(\theta)$.
- $\phi > 0$ es un parámetro de escala o de dispersión adicional. En algunas familias, es un parámetro fijo y en otras es un parámetro desconocido que tiene que estimarse.
- b y c son funciones, que dependen del tipo de familia exponencial.
- w_i es un peso. Si los datos no están agrupados, usualmente $w_i = 1, i = 1, \ldots, n$. Si los datos se agrupan en g grupos, usualmente $w_i = n_i, i = 1, \ldots, g$ para el promedio de observaciones, y $w_i = 1/n_i$ para la suma de observaciones. En algunos libros, a veces reemplazan $\frac{\phi}{w_i}$ por $a(\phi)$ en la ecuación dada.
- La varianza condicional de y dado **x** es de la forma $var(y|\mathbf{x}_i) = \sigma^2(\mu_i) = \frac{\phi\nu(\mu_i)}{w_i}$ donde ν se determina de manera única para la familia exponencial específica a través de la relación $\nu(\mu) = b''(\theta)$. Entonces la especificación de la estructura de la media implica cierta estructura para la varianza.

Si la función media y la función varianza se especifícan por separado, entonces aunque se siguen cumpliendo muchas relaciones y procedimientos, ya no se cumple el supuesto de familia exponencial. En su lugar se obtienen *modelos cuasi-verosímiles*, que veremos más adelante.

Ejemplo. [Normal]

Consideremos el caso normal univariado: $y|\mathbf{x} \sim \mathcal{N}\left(\mu, \sigma^2\right)$. En este caso, directamente $\mu = \beta'\mathbf{x}$.

$$\begin{aligned}
\log f(y) &= -\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \\
&= \frac{-(y^2 - 2y\mu + \mu^2)}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \\
&= \frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \\
&= \frac{y\mu - \mu^2/2}{\sigma^2} + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)
\end{aligned}$$

Entonces:

- $\theta = \mu$, $\phi = \sigma^2$, $w_i = 1$
- $b(\theta) = b(\mu) = \frac{\mu^2}{2}$
- $c(y, \phi, w_i) = -\left(\frac{y^2}{2\sigma^2} + \log(2\pi\sigma^2)\right)$

□

Ejemplos de GLM's

Ejemplo. [Poisson]

$Y|\mathbf{x} \sim \mathcal{P}(\lambda)$. Aquí $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$

$$\begin{aligned} \log f(y) &= -\lambda + y\log(\lambda) - \log(y!) \\ &= \frac{y\log(\lambda) - \lambda}{1} + (-\log(y!)) \end{aligned}$$

Entonces:

- $\theta = \log(\lambda)$, $\phi = 1$, $w_i = 1$.
- $b(\log(\lambda)) = \lambda$, por lo que $b(\theta) = e^\theta$
- $c(y, \phi, w_i) = -\log(y!)$

□

Ejemplo. [Binomial]

$Y|\mathbf{x} \sim \textbf{Bin}\,(m,p)$. Aquí $P(Y=y) = \binom{m}{y} p^y (1-p)^{m-y}$

$$\begin{aligned}
\log f(y) &= \log \binom{m}{p} + y \log(p) + (m-y)\log(1-p) \\
&= y \log\left(\frac{p}{1-p}\right) + m\log(1-p) + \log\binom{m}{p} \\
&= \frac{\left(\frac{y}{m}\log\left(\frac{p}{1-p}\right) + \log(1-p)\right) m}{1} + \log\binom{m}{p}
\end{aligned}$$

Entonces la variable que se considera es $\frac{y}{m}$ y se tienen como parámetros :

- $\theta = \log\left(\frac{p}{1-p}\right)$, lo que implica que $p = \frac{e^\theta}{1+e^\theta}$; $\phi = 1$, $w_i = m$.
- $b(\theta) = \log(1-p) = \log\left(1 - \frac{e^\theta}{1+e^\theta}\right) = -\log(1+e^\theta)$.
- $c(y, \phi, w_i) = \log\binom{m}{y}$

□

Ejemplo. [Bernoulli]

Previamente vimos el caso Bernoulli, que no se repetirá aquí. Sólo recordar:

- En el caso del logit, se definía $\log\left[\frac{p}{1-p}\right] = \beta'\mathbf{x}$ o $p = \frac{1}{1+\exp(-\beta'\mathbf{x})}$
- El modelo *probit* (o *normit*) define $p = \Phi(\beta'\mathbf{x})$ Se puede pensar en un modelo en donde hay una variable continua latente L que depende de los predictores, L que es la que define la probabilidad a través de un umbral $L_0 = \beta'\mathbf{x}$: $Y = 1$ si $L \leq L_0$ y entonces:

$$P(Y = 1|\mathbf{x}) = P(L \leq L_0) = \Phi(\beta'\mathbf{x})$$

- El modelo basado en la función liga conocida como log-log complemetaria establece que $\log(-\log(1-p)) = \beta'\mathbf{x}$.
- El modelo *tobit* desarrollado por el premio Nobel James Tobin, que introduce el concepto de muestra censurada.

´

□

Tabla resumen de GLM's

Nombre	Distribución	$\theta(\mu)$	$b(\theta)$	ϕ	w_i	$V(\mu)$
Normal	$\mathcal{N}(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2	1	1
Bernoulli	**Bernoulli** (p)	$\log\left(\frac{p}{1-p}\right)$	$\log(1+e^\theta)$	1	1	$n\mu(1-\mu)$
Binomial	**Bin** (m, p)	$\log\left(\frac{p}{1-p}\right)$	$\log(1+e^\theta)$	1	m	$\mu(1-\mu)$
Poisson	$\mathcal{P}(\lambda)$	$\log(\lambda)$	$\exp(\theta)$	1	1	μ
Gamma	$\mathcal{G}(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}	1	μ^2
Gamma inversa	$IG(\mu, \sigma^2)$	$1/\mu^2$	$-2(-2\theta)^{1/2}$	σ^2	1	μ^3

- Para cada familia exponencial existe una función *liga natural o canónica*: es la que relaciona al parámetro natural θ directamente al predictor lineal $\eta = \beta'\mathbf{x}$:

$$\theta = \theta(\mu) = \eta = \beta'\mathbf{x}$$

es decir, $g(\mu) \equiv \theta(\mu)$.

Algunas características de la representación exponencial I

- La ventaja de expresar diversas familias de distribuciones en una forma exponencial común es que las propiedades generales de las familias exponenciales se puede aplicar a los casos individuales. En general, se tienen estas características:
  1. La varianza condicional de Y dado **x** es una función de su media y posiblemente del parámetro de dispersión ϕ: $\text{Var}(Y|\mathbf{x}) = \phi v(\mu)$
  2. $b'(\theta) = \mu$, la función media $\mu = \text{E}(Y)$.
  3. La función varianza está dada por $V(Y) = \frac{\phi}{m} b''(\theta) = \frac{\phi}{m} v(\mu)$. Por ejemplo, para la distribución normal, se tiene
$$b'(\theta) = \theta = \mu, \quad \phi b''(\theta) = \phi = \sigma^2, \quad v(\mu) = 1$$

Estimación de GLM’s

Estimación por máxima verosimilitud de θ I

- Dado que la densidad de y es un elemento de la familia exponencial, La log-verosimilitud para y_i vimos que se representa fácilmente:

$$l(\theta, \phi|\mathbf{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i) \right)$$

- Supongamos que el GLM usa una función liga g tal que $g(\mu_i) = \eta_i = \beta' \mathbf{x}_i$. Para obtener las ecuaciones normales para estimar los parámetros del modelo, requerimos diferenciar la log-verosimilitud con respecto a cada coeficiente. Denotando con l_i el $i-$ésimo sumando de la log-verosimilitud y usando la regla de la cadena:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}$$

- Como $b'(\theta) = \mu$ y $b''(\theta) = v(\mu)$, entonces debemos tener que $\frac{d\mu}{d\theta} = b''(\theta) = v(\mu)$. Además, $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$, por lo que podemos escribir la regla de la cadena anterior como:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\phi v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} x_{ij}$$

Estimación por máxima verosimilitud de θ II

- Entonces las ecuaciones normales quedan de la forma:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi v(\mu_i)} \frac{d\mu_i}{\eta_i} x_{ij} = 0$$

Las ecuaciones tienes que resolverse numéricamente, porque son funciones no lineales de los parámetros.

Mínimos cuadrados ponderados iterativos (IWLS) I

- Si definimos $Z_i = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i} = \eta_i + (y_i - \mu_i)g'(\mu_i)$, entonces al tomar la media, tenemos que $\mathsf{E}(Z_i) = \eta_i$ y $\mathsf{Var}(Z_i) = (g'(\mu_i))^2\phi v(\mu_i)$.
- de las igualdades anteriores, podemos ver que si se pueden calcular las Z_i, podemos ajustar el modelo usando mínimos cuadrados ponderados de la regresión de Z's en las Xs. Pero no conocemos los valores de las μ_i y η_i que dependen de los coeficientes que queremos estimar, por lo que parece que no vamos a ningún lado.
- De esta idea, Nelder y Wedderburn sugirieron el método IWLS para que el procedimiento circular, se hiciera un procedimiento iterativo:
  1. Comienza con valores iniciales de las $\hat{\mu}_i$ y de $\hat{\eta}_i = g(\hat{\mu}_i)$, que denotamos como $\hat{\mu}_i{}^{(0)}$ y $\hat{\eta}_i{}^{(0)}$. Una elección simple para comenzar es $\hat{\mu}_i{}^{(0)} = Y_i$, por ejemplo.
  2. En cada iteración k, calcular la variable Z usando los valores de $\hat{\mu}$ y $\hat{\eta}$ de la iteración previa:

$$Z_i^{(k-1)} = \eta_i^{(k-1)} + \left(Y_i - \mu_i^{(k-1)}\right) g'\left(\mu_i^{(k-1)}\right)$$

junto con los pesos

$$W_i^{(k-1)} = \frac{1}{\left[g'\left(\mu_i^{(k-1)}\right)\right]^2 \psi v\left(\mu_i^{(k-1)}\right)}$$

Mínimos cuadrados ponderados iterativos (IWLS) II

3. Ajusta mínimos cuadrados ponderados de la regresión de $Z^{(k-1)}$ en las X's, usando las $W^{(k-1)}$ como pesos:

$$\mathbf{b}^{(k)} = \left(\mathbf{X}'\mathbf{W}^{k-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(k-1)}\mathbf{z}^{(k-1)}$$

4. Repetir los pasos 2 y 3 hasta que los coeficientes de regresión se estabilicen. En ese punto **b** converge a los estimadores máximo-verosímiles de las β's.

Estimación del parámetro de dispersión

- No se requiere estimar ϕ para estimar los coeficientes de regresión en un GLM. Usualmente se estima a través del método de momentos, ya que como

$$\text{Var}(Y_i) = \phi v(\mu_i)$$

se obtiene como estimador:

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

y la matriz de covarianza asintótica de los coeficientes se obtiene de la última iteración del proceso IWLS como:

$$\hat{V}(\mathbf{b}) = \hat{\phi}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

Pruebas de hipótesis en GLM's I

- Para probar la hipótesis nula $H_0 : \beta_j = b$, se calcula la estadística de Wald $W_0 = \frac{\hat{\beta}_j - b}{se(\hat{\beta}_j)}$ que bajo H_0 tiene una distribución $\mathcal{N}(0, 1)$. Pero para modelos con un parámetro de dispersión estimado, se utiliza $W_0 \sim t_{n-p-1}$.
- El equivalente a ANOVA en GLM es el *análisis de devianza* como ya se ha comentado. En este contexto, la devianza residual para un GLM es

$$D_m = 2(l_s - l_m)$$

donde l_m es la log-verosimilitud optima bajo el modelo considerado y l_s es la log-verosimilitud óptima bajo el modelo saturado, que tiene un parámetro por cada observación y consecuentemente ajustan los datos tanto como es posible.

- En los GLMs que tienen $\phi = 1$, la prueba LRT es simplemente la diferencia en las devianzas residuales para modelos anidados: si tenemos un modelo 0 con $p_0 + 1$ coeficientes para H_0, que está anidado en un modelo 1 para H_1 con $p_1 + 1$ coeficientes, con $p_0 < p_1$ entonces la prueba es

$$G_0^2 = D_0 - D_1 \sim \chi^2_{(p_1 - p_0)}$$

- Cuando se tiene que estimar ϕ, se usa una prueba F

$$F_0 = \frac{\frac{D_0 - D_1}{p_1 - p_0}}{\hat{\phi}} \sim F_{p_1 - p_0, n-p-1}$$

  tomando $\hat{\phi}$ es el parámetro estimado en el modelo más grande ajustado a los datos (que no necesariamente es el modelo 1).
- La devianza residual dividida por la dispersión estimada es lo que se reporta como la *devianza escalada*.
- El número $R^2 = 1 - \frac{D_1}{D_0}$ representa la proporción de la devianza nula que se explica por el modelo.

Ejemplo: vacas yacentes I

- Justo antes o después de parir algunas vacas son incapaces de soportar su propio peso, y se vuelven yacentes. Algunas vacas con esta condición se recuperan pero otras no y es de interés comprender la probabilidad de sobrevivencia, que varia con las características de las vacas. El conjunto de datos 'downer' contiene datos de un estudio de 435 vacas yacentes que se realizó en Nueva Zelanda entre 1983 y 1984. Las variables son:
  - `ast`: suero asparate amino-transferencia IU/l a 30 grados
  - `calving`: 0 si la condición ocurre antes de parir o 1 si es post-parto
  - `ck`: Suero creatina fosfokinasa
  - `daysrec`: días yacente cuando las medidas fueron tomadas, redondeado hacia abajo al día más cercano
  - `inflamat`: Inflamación, 1 si está presente y 0 ausente
  - `myopathy`: Desorden muscular, 1 si está presente
  - `outcome`: 1 si sobrevive 0 si muere o se sacrifica
  - `pcv`: Volumen de hematrocitos en
  - `urea`: nivel de urea

Ejemplo: vacas yacentes II

```
library(alr4)
library(dplyr)
data("Downer")
head(Downer)

  calving daysrec    ck  ast urea pcv inflamat myopathy  outcome
1   after       1  3000  590 10.9  41       no     <NA> survived
2   after       4 12100 1240 23.8  NA     <NA>  present survived
3   after       0   600  207  8.1  31      yes     <NA>     died
4   after       1   590  113 14.6  43     <NA>     <NA>     died
5   after       2  1800  243 12.3  48     <NA>   absent     died
6  before       6   380  150   NA  28      yes   absent     died
```

- Primero que nada tratemos de entender los datos

Ejemplo: vacas yacentes III

```
summary(Downer)
```

```
   calving       daysrec             ck              ast              urea             pcv       
 before:107   Min.   : 0.000   Min.   :   13   Min.   :  33.0   Min.   : 1.000   Min.   :13.00  
 after :324   1st Qu.: 0.000   1st Qu.:  560   1st Qu.: 123.0   1st Qu.: 5.625   1st Qu.:32.00  
 NA's  :  4   Median : 1.000   Median : 1760   Median : 240.0   Median : 7.600   Median :35.00  
              Mean   : 1.947   Mean   : 5352   Mean   : 398.4   Mean   : 9.803   Mean   :35.56  
              3rd Qu.: 3.000   3rd Qu.: 5467   3rd Qu.: 492.0   3rd Qu.:10.975   3rd Qu.:40.00  
              Max.   :20.000   Max.   :71000   Max.   :2533.0   Max.   :50.000   Max.   :61.00  
              NA's   :3        NA's   :22      NA's   :6        NA's   :169      NA's   :260    
 inflamat      myopathy       outcome    
 no  : 38   absent :127   died    :269  
 yes : 98   present: 95   survived:166  
 NA's:299   NA's   :213                 
```

- Hay muchos datos faltantes de algunas variables. La variable inflamación sólo se midió durante el segundo año del estudio, por eso sólo hay 136 de las 435 observaciones.
- También notamos que `ck` tiene un mínimo de 13 y un máximo de 71,000. El cociente es mayor que 1000, por lo que, si recuerdan se recomienda hacer una transformación a logaritmos. Para efectos de interpretación, en este caso conviene usar el logaritmo de 2. Lo mismo con `ast`.

Ejemplo: vacas yacentes IV

- ¿Qué fracción de las vacas sobrevivieron?

```
table(Downer$outcome)/length(Downer$outcome)
```

```
     died  survived
0.6183908 0.3816092
```

- Algunas preguntas, por ejemplo, las vacas con desorden muscular sobreviven igual? Estamos preguntando por la distribución condicional de `outcome` | `myopathy`. Vemos que muy pocas vacas con miopatía sobreviven.

```
with(Downer,prop.table(table(outcome,myopathy),margin=2)) #usa las columnas como los totales.
```

```
          myopathy
outcome        absent    present
  died     0.61417323 0.93684211
  survived 0.38582677 0.06315789
```

- Con las consideraciones hechas, ajustemos un primer modelo considerando por ejemplo `log2ck`, una sola variable para interpretar los coeficientes:

Ejemplo: vacas yacentes V

```
downer <- Downer %>%
          mutate(outcome = ifelse(outcome=="died",0,1),
                 myopathy = factor(myopathy),
                 inflamat = factor(inflamat),
                 log2ck = log2(ck),
                 log2ast = log2(ast))

m0 <- glm(outcome ~ log2ck, data = downer, family = "binomial")
summary(m0)


Call:
glm(formula = outcome ~ log2ck, family = "binomial", data = downer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1337  -0.8811  -0.5608   1.0588   1.9935

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.00065    0.58089   6.887 5.69e-12 ***
log2ck      -0.42402    0.05497  -7.714 1.22e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 550.49  on 412  degrees of freedom
Residual deviance: 475.18  on 411  degrees of freedom
  (22 observations deleted due to missingness)
AIC: 479.18

Number of Fisher Scoring iterations: 3
```

Ejemplo: vacas yacentes VI

- Del ajuste, vemos que $\eta = 4 - 0.42\texttt{log2ck}$. Entonces si se incrementa una unidad `log2ck` (o en este caso, si `ck` se duplica) entonces el logaritmo natural de los momios decrece por 0.42 o bien, los momios se multiplican por $\exp(-0.42) = 0.65$. Esto quiere decir que una vaca con $CK = 1000$ tiene momios de sobrevivencia que son 0.65 veces los momios de una vaca con $\texttt{ck} = 500$.
- Considerando un modelo con más predictores:

Ejemplo: vacas yacentes VII

```
m1 <- glm(outcome ~ calving + myopathy + daysrec + log2ast + log2ck, data = downer, family = "binomial")
summary(m1)


Call:
glm(formula = outcome ~ calving + myopathy + daysrec + log2ast +
    log2ck, family = "binomial", data = downer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3023  -0.9208  -0.3829   1.1482   2.3898

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.23651    1.61117   0.767  0.44281
calvingafter     -0.30333    0.40767  -0.744  0.45684
myopathypresent -1.84066    0.61180  -3.009  0.00262 **
daysrec          -0.05460    0.11234  -0.486  0.62696
log2ast           0.03883    0.26384   0.147  0.88300
log2ck           -0.15259    0.14448  -1.056  0.29091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 245.10  on 215  degrees of freedom
Residual deviance: 211.27  on 210  degrees of freedom
  (219 observations deleted due to missingness)
AIC: 223.27

Number of Fisher Scoring iterations: 5
```

Ejemplo: vacas yacentes VIII

- La regresión anterior se calcula con $435 - 219 = 216$ casos. El efecto de observar miopatía ($\texttt{myopathy} = 1$), manteniendo los otros predictores fijos, es multiplicar los momios de sobrevivencia por $\exp(-1.84) = 0.16$ y el efecto de doblar `ck` (agregar una unidad a `log2ck`), es multiplicar los momios de sobrevivencia por $\exp(-0.15) = 0.86$. Parece que el único coeficiente significativo del modelo es la miopatía.
- Para probar la hipótesis $H_0 : \beta_0 + \beta_1 \texttt{log2ck}$ vs. $H_1 : \beta_0 + \beta_1 \texttt{log2ck} + \beta_2 \texttt{calving} + \beta_3 \texttt{myopathy} + \beta_4 \texttt{log2ast}$, se calcula $G_0^2 = 475.18 - 211 \sim \chi^2_{411-210}$. El p-value es

```
pchisq(475.18-211,411-210,lower.tail=F)
```

```
[1] 0.001846457
```

Por lo tanto, se tiene evidencia estadística para considerar significativo el modelo más grande, lo que quiere decir que los predictores son significativos. Ahora consideremos el modelo sólo con miopatía

```
m0 <- glm(outcome ~ myopathy, data = downer, family = "binomial")
pchisq(m0$deviance - m1$deviance, m0$df.residual - m1$df.residual, lower.tail = F)
```

```
[1] 0.9843474
```

Confirmamos la significancia de miopatía como la variable más significativa para explicar la sobrevivencia.

Modelos loglineales y regresión Poisson I

- Los modelos Poisson surgen en dos contextos diferentes:
  1. Cuando se supone que la distribución condicional de la variable de respuesta dados los predictores es Poisson.
  2. Cuando se analizan asociaciones en tablas de contingencia. En las tablas de contingencia, los conteos son multinomiales, no Poisson, condicionales, pero con una interpretación adecuada de los parámetros (condicionando los totales como ya vimos), los estimadores multinomiales se pueden obtener *como si* fueran los conteos Poisson.
- Entonces se puede usar el mismo enfoque de modelos GLM se puede usar para regresión Poisson y para modelos loglineales de tablas de contingencia.
- El uso más común de la regresión Poisson, en donde la variable de respuesta es un conteo, es en el análisis de los modelos log-lineales para tablas de contingencia.
- Los modelos log-lineales tienen mucho en común con los modelos de regresión, pero hay diferencias importantes en lenguaje, notación, formas de modelos, casos especiales, y sumarizaciones.

- El modelo de regresión Poisson se puede aplicar a las tablas de contingencia de conteos, tratando cada una de las variables que categorizan los datos como un factor y los conteos de las celdas de la tabla como la variable dependiente. Por ejemplo, para una tabla de contingencia de dos variables, el modelo se puede representar como:

$$\log(\mu_{jk}) = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

donde los parámetros cumplen las restricciones: $\sum_j \alpha_j = \sum_k \beta_k = \sum_{jk} \gamma_{jk} = 0$ y $\log(\mu_{jk})$ representa el logaritmo esperado de los conteos.

Planes de muestreo para modelos loglineales I

- En general las tablas de contingencia se pueden construir de diferentes modos, pero en todos los casos el modelo Poisson puede ser usado:
  1. **Muestreo Poisson**: Tamaño de muestra total aleatorio n.
  2. **Muestreo Multinomial**: Se fija el tamaño de muestra total de antemano n y se muestrea en las diferentes celdas hasta alcanzar el tamaño de muestra. En este caso los conteos no son independientes por la restricción de que suman n. En este caso la media total queda determinada por el plan de muestreo (la ordenada al origen).
  3. **Muestreo producto-multinomial**: se muestrea tomando igual número de observaciones en alguna de las dimensiones. En este caso, la ordenada al origen y los efectos principales están determinados por el muestreo.
  4. **Muestreo fijando dos niveles**: Se muestrea un número fijo en cada combinación de dos dimensiones, para obtener esquemas de muestreo multinomiales en esas dos dimensiones. En este caso, todos los modelos que se ajusten deben contener los términos `1 + A + B + A:B = A*B`, porque están fijos por el diseño muestral.
  5. **Muestreo retrospectivo (o de control de casos)**: Supongamos que una de las dimensiones C del problema tiene categorías que ocurren de manera rara. Podemos decidir muestrar $n/2$ eventos raros de C y el resto de los no raros. Entonces C funge como una variable de respuesta. Estos modelos deben incluir `1 + C`. Los otros términos del modelo en donde aparece C nos dicen si la respuesta está relacionada con los predictores, y de qué manera.

Ejemplo: tablas de 2×2

- Estos modelos fueron propuestos por Birch en 1963 (Maximum Likelihood in three-way contingency tables, JRSS,B, 25, 220-233).
- Consideremos primero una tabla de contingencia de 2×2, de todos los doctorados otorgados en ciencias matemáticas en los EUA en 2011:

```
data("AMSsurvey")
head(AMSsurvey)
```

```
   type    sex citizen count count11
1 I(Pu)   Male      US   132     148
2 I(Pu) Female      US    35      40
3 I(Pr)   Male      US    87      63
4 I(Pr) Female      US    20      22
5    II   Male      US    96     161
6    II Female      US    47      53
```

Colapsando a una tabla donde sólo se tome `citizen` y `sex`:

```
tabla <- xtabs(count11 ~ sex + citizen, data=AMSsurvey)
tabla
```

```
        citizen
sex      Non-US  US
  Female    276 219
  Male      575 574
```

El análisis típico en una tabla de dos dimensiones es probar independencia de renglones y columnas, usando una prueba de bondad de ajuste $\chi^2_{(r-1)(c-1)}$

```
chisq.test(tabla, correct = F)
```

```
Pearson's Chi-squared test
```

Ejemplo: tablas de 3×3 I

- Incorporando el tipo de institución (`type` se refiere a grupos I para universidades públicas y privadas respectivamente, II y III para grupos II y III, IV para estadística y bioestadística y Va para matemáticas aplicadas), hay muchos más modelos que pueden ser considerados: con factores simples, interacciones de dos y tres variables.

```
tabla <- ftable(xtabs(count11 ~ type + sex + citizen, data=AMSsurvey))
tabla

             citizen Non-US  US
type  sex
I(Pr) Female             26  22
      Male               82  63
I(Pu) Female             32  40
      Male              136 148
II    Female             56  53
      Male              116 161
III   Female             30  28
      Male               61  71
IV    Female            115  55
      Male              153  89
Va    Female             17  21
      Male               27  42
```

Ejemplo: tablas de 3×3 II

Para plantear modelos razonables, se debe cumplir el *Principio de marginalidad*: Un modelo que incluye un término de orden alto (como interacciones) también debe incluir los predictores relativos de menor orden de ese término: los efectos principales que componen la interacción.
El modelo más grande es el modelo saturado, y a partir de este se pueden considerar modelos menores eliminando términos, pero cumpliendo el principio de marginalidad. Por ejemplo, podemos ejecutar los modelos en el orden siguiente:

```
mod.saturado <- glm(count ~ type*sex*citizen, family=poisson, data=AMSsurvey)
```

A partir del modelo saturado, podemos usar `Anova` para calcular todas las pruebas de modelos conformando el principio de marginalidad (usualmente se llaman pruebas tipo II)

Ejemplo: tablas de 3×3 III

```
Anova(mod.saturado)

Analysis of Deviance Table (Type II tests)

Response: count
                LR Chisq Df Pr(>Chisq)
type             233.336  5  < 2.2e-16 ***
sex              182.983  1  < 2.2e-16 ***
citizen            5.923  1  0.0149447 *
type:sex          69.135  5  1.551e-13 ***
type:citizen      24.041  5  0.0002132 ***
sex:citizen        0.538  1  0.4634559
type:sex:citizen   1.419  5  0.9221984
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- La tabla se tiene que leer de abajo hacia arriba. La triple interacción y la interacción de `sex:citizen` no son significativas, las otras son diferentes de 0.
- La prueba de los efectos principales usualmente son irrelevantes, pues corresponden a las marginales de las variables.

Ejemplo: tablas de 3×3 IV

- Si actualizamos el modelo quitando los dos últimos términos que no fueron significativos:

```
mod.1 <- update(mod.saturado, .~ . -sex:citizen - type:sex:citizen)
mod.1


Call:  glm(formula = count ~ type + sex + citizen + type:sex + type:citizen,
    family = poisson, data = AMSsurvey)

Coefficients:
        (Intercept)            typeI(Pu)               typeII              typeIII
          3.099e+00            3.417e-01            7.681e-01            5.436e-01
             typeIV               typeVa              sexMale            citizenUS
          1.531e+00           -6.296e-01            1.305e+00            2.844e-02
  typeI(Pu):sexMale       typeII:sexMale      typeIII:sexMale       typeIV:sexMale
          1.041e-01           -6.597e-01           -9.628e-01           -1.112e+00
     typeVa:sexMale  typeI(Pu):citizenUS     typeII:citizenUS    typeIII:citizenUS
         -4.363e-01            2.065e-02           -6.724e-05           -1.808e-01
   typeIV:citizenUS     typeVa:citizenUS
         -6.251e-01            1.539e-01

Degrees of Freedom: 23 Total (i.e. Null);  6 Residual
Null Deviance:     521.4
Residual Deviance: 1.957  AIC: 175.3
```

- La prueba para la devianza residual es

```
pchisq(1.9568,df=6,lower.tail = F)
```

```
[1] 0.9236285
```

- Esto quiere decir que no hay diferencia entre el modelo saturado y el ajustado, y por lo tanto el modelo ajusta muy bien los datos.

Sobredispersión en modelos Poisson I

- Un GLM en donde el parámetro de dispersión ϕ es constante (`binomial` o `poisson`), tiene devianza residual:

$$D_m = 2(l_s - l_m) \dot{\sim} \chi^2_{(n-p)}$$

donde l_s es la log-verosimilitud del modelo saturado, y l_m la del modelo ajustado (que tiene p coeficientes).

- Para una distribución $\chi^2_{(p)}$, sabemos que $\mathrm{E}(\chi^2_{(p)}) = p$. Así que si un modelo ajusta bien los datos, se esperaría que $D_m \approx gl_m$.
- En el caso de que $Dev_m > gl_m$ esto se puede deber a dos principales causas:
  1. Un modelo mal ajustado:
     - El modelo no está bien especificado.
     - Hay algunos factores explicativos que no están incorporados en los predictores.
     - posiblemente outliers.
  2. Por **sobredispersión**: la variación observada de los datos es mayor que la que estima el modelo. Esto significa que posiblemente:
     - La variabilidad de los casos a nivel individual es importante.
     - Hay correlación entre las observaciones de la variable de respuesta.
     - El diseño muestral considera conglomerados.
     - Se omiten algunas variables no observadas.
     - Exceso de ceros.

Consecuencias de la sobredispersión I

- Usualmente se obtienen estimadores consistentes de β, pero
  - Los errores estándar no son correctos
  - Se obtienen intervalos de confianza demasiado optimistas.
  - Se pueden seleccionar modelos demasiado complejos.
- Las consecuencias pueden ser potencialmente severas.

Ejemplo: Modelo de vínculos (Ornstein, 1976) I

- Los siguientes datos (Ornstein, 1976) corresponden a datos de 248 empresas canadienses, en donde cada empresa tiene un cierto número de *vínculos* (interlocks): estos son el número de funcionarios de alto nivel que comparten las empresas en el conjunto de datos.

```
library(car)
data("Ornstein") # en el paquete car
str(Ornstein)
```

```
'data.frame': 248 obs. of  4 variables:
 $ assets    : int  147670 133000 113230 85418 75477 40742 40140 26866 24500 23700 ...
 $ sector    : Factor w/ 10 levels "AGR","BNK","CON",..: 2 2 2 2 2 4 9 2 9 8 ...
 $ nation    : Factor w/ 4 levels "CAN","OTH","UK",..: 1 1 1 1 1 1 1 1 1 4 ...
 $ interlocks: int  87 107 94 48 66 69 46 16 77 6 ...
```

```
levels(Ornstein$sector)
```

```
 [1] "AGR" "BNK" "CON" "FIN" "HLD" "MAN" "MER" "MIN" "TRN" "WOD"
```

```
levels(Ornstein$nation)
```

```
[1] "CAN" "OTH" "UK"  "US"
```

- Podemos ver la gráfica de la respuesta no condicional, hay 28 empresas sin vínculos, 19 con 1, 14 con 2, y así sucesivamente:

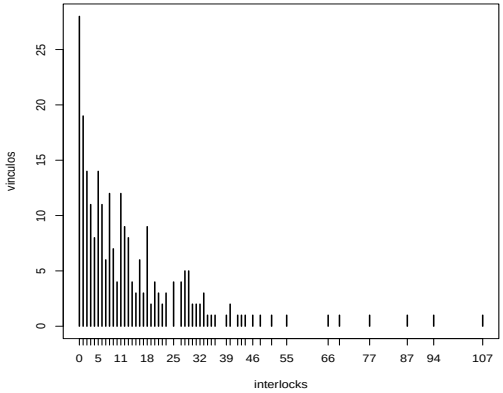
Ejemplo: Modelo de vínculos (Ornstein, 1976) II

```
vinculos <- xtabs(~ interlocks,data=Ornstein)
vinculos
```

```
interlocks
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  25
 28  19  14  11   8  14  11   6  12   7   4  12   9   8   4   3   6   3   9   2   4   3   2   3   4
 27  28  29  30  31  32  33  34  35  36  39  40  42  43  44  46  48  51  55  66  69  77  87  94 107
  4   5   5   2   2   2   3   1   1   1   1   2   1   1   1   1   1   1   1   1   1   1   1   1   1
```

```
plot(vinculos,type="h")
```

Ejemplo: Modelo de vínculos (Ornstein, 1976) III



Ejemplo: Modelo de vínculos (Ornstein, 1976) IV

- Podemos ajustar un modelo Poisson para estos datos:

```
mod1 <- glm(interlocks ~ ., family = poisson, data = Ornstein)
mod1


Call:  glm(formula = interlocks ~ ., family = poisson, data = Ornstein)

Coefficients:
(Intercept)       assets    sectorBNK    sectorCON    sectorFIN    sectorHLD    sectorMAN
  2.325e+00    2.085e-05   -4.092e-01   -6.196e-01    6.770e-01    2.085e-01    5.260e-02
  sectorMER    sectorMIN    sectorTRN    sectorWOD    nationOTH     nationUK     nationUS
  1.777e-01    6.211e-01    6.778e-01    7.116e-01   -1.632e-01   -5.771e-01   -8.259e-01

Degrees of Freedom: 247 Total (i.e. Null);  234 Residual
Null Deviance:     3737
Residual Deviance: 1887  AIC: 2813
```

- Interesa analizar la relación del número de vínculos con respecto a otras características de las empresas: sus activos, la nación controladora (4 naciones), y el sector de operación de la empresa (hay 10 sectores). La nación base es Canadá y el sector base es la agricultura (son los valores de las variables categóricas que no están incluídas en los datos).
- En este ejemplo se tiene evidencia de sobredispersión, ya que la devianza residual es 1887 con 234 grados de libertad.

Ejemplo: Modelo de vínculos (Ornstein, 1976) V

- Los coeficientes en este modelo se interpretan como efectos en la escala log del conteo, así que hay que exponenciar los coeficientes para producir los efectos multiplicativos en la escala de conteo:

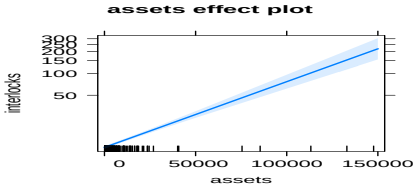
```
exp(coef(mod1))
```

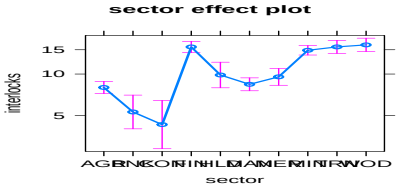
```
(Intercept)      assets   sectorBNK   sectorCON   sectorFIN   sectorHLD   sectorMAN   sectorMER
 10.2223811   1.0000209   0.6641933   0.5381753   1.9679164   1.2317801   1.0540063   1.1944549
  sectorMIN   sectorTRN   sectorWOD   nationOTH    nationUK    nationUS
  1.8609104   1.9695952   2.0371514   0.8494158   0.5615318   0.4378261
```

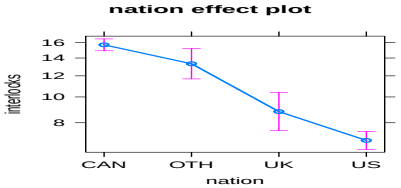
- Podemos interpretar de la siguiente manera: Para una empresa gringa, mantiene en promedio 43.78 % de vinculos menos que las canadienses (que es la categoría base).
- Para ver los efectos, podemos usar el paquete `effects`. El eje vertical está en la escala del predictor lineal (logaritmo), pero las marcas son las etiquetas de la respuesta.

```
library(effects)
plot(allEffects(mod1),cex=0.5)
```

Ejemplo: Modelo de vínculos (Ornstein, 1976) VI







Modelos de quasi-verosimilitud: modelo quasiPoisson

- Para resolver el problema de sobredispersión en un modelo Poisson hay dos opciones:
  1. Ajustar un modelo quasi-Possion, que introduce un parámetro de dispersión en la función varianza: $Var(Y_i|\eta_i) = \phi\mu_i$. De esta manera, si $\phi > 1$, la varianza condicional de Y se incrementará más rápido que su media. Como no hay una familia exponencial para este caso, y el GLM no implica una distribución particular para la variable de respuesta, entonces no se puede usar máxima verosimilitud, pero si se aplican los métodos usuales, se le llama *quasi* máxima verosimilitud.
  2. Usar un modelo con respuesta binomial negativa: se supone un modelo Poisson para $Y|\mu^*$ y μ^* es una variable aleatoria que tiene distribución gamma con media μ y parámetro de escala ω. Entonces los conteos observados siguen una distribución binomial negativa, con media $\mathrm{E}(Y) = \mu$ y varianza $\mathrm{Var}(Y) = \mu + \mu^2/\omega$.
- Consideraremos en lo que sigue el modelo de sobredispersión. Los estimados de quasi-verosimilitud son idénticos a los de máxima verosimilitud, pero los errores estándar de los coeficientes cambian: si $\tilde{\phi}$ es el parámetro estimado para el modelo, entonces:

$$se(\hat{\beta}_i^{qP}) = \tilde{\phi}^{1/2} se(\hat{\beta}_i^{P})$$

- El estimador que se usa para el parámetro de dispersión es el de momentos que ya se mencionó antes:

$$\tilde{\phi} = \frac{1}{n-k} \sum \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Ejemplo: Modelo quasiPoisson para los datos de vínculos I

- Repetimos el ejercicio cambiando la familia de `poisson` a `quasipoisson`.
- Esto no cambia los valores del ajuste, pero notemos que los errores ahora ya no son z sino t, lo que cambia la significancia de los coeficientes.
- También estima el parámetro de dispersión que ya no es 1, sino $\tilde{\phi} = 7.9439$ y entonces cada error estándar se multiplica por $\sqrt{\tilde{\phi}} = \sqrt{7.9439} = 2.8184925$.

```
modqP <- glm(interlocks ~ ., family = quasipoisson, data = Ornstein)
modqP


Call:  glm(formula = interlocks ~ ., family = quasipoisson, data = Ornstein)

Coefficients:
(Intercept)       assets    sectorBNK    sectorCON    sectorFIN    sectorHLD    sectorMAN
  2.325e+00    2.085e-05   -4.092e-01   -6.196e-01    6.770e-01    2.085e-01    5.260e-02
  sectorMER    sectorMIN    sectorTRN    sectorWOD    nationOTH     nationUK     nationUS
  1.777e-01    6.211e-01    6.778e-01    7.116e-01   -1.632e-01   -5.771e-01   -8.259e-01

Degrees of Freedom: 247 Total (i.e. Null);  234 Residual
Null Deviance:      3737
Residual Deviance: 1887  AIC: NA
```

Otro enfoque para sobredispersión: regresión binomial negativa I

- Recordemos lo siguiente: $Z \sim \mathcal{G}(\alpha, \beta)$ si tiene como densidad:

$$f(z) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} z^{\alpha-1} e^{-z/\beta}$$

  donde α es el parámetro de escala y β el de forma. Con esta parametrización, $\mathsf{E}(Z) = \alpha\beta$ y $\mathsf{Var}(Z) = \alpha\beta^2$.

- Otra manera de pensar el problema de sobredispersión es permitir que el parámetro de la distribución Poisson tenga su propio comportamiento. Usualmente se considera el siguiente modelo jerárquico:

$$\begin{aligned} Y|\lambda &\sim \mathcal{P}(\lambda) \\ \lambda &\sim \mathcal{G}(\omega, \mu_i/\omega) \end{aligned}$$

  de tal forma que $\mathsf{E}(\lambda) = \mu_i = \omega\mu_i/\omega$.

- A partir de este modelo jerárquico, ¿cuál es la distribución de Y_i?

$$P(Y = y) = \int_0^{\infty} f_Y(y|\lambda) f_\lambda(\lambda)\, d\lambda \quad = \quad \int_0^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} \frac{\omega^{\omega}\lambda^{\omega-1}e^{-\lambda\omega/\mu_i}}{\mu_i^{\omega}\Gamma(\omega)}\, d\lambda$$

$$\ldots$$

Otro enfoque para sobredispersión: regresión binomial negativa II

- Entonces Y sigue la generalización de una distribución binomial negativa, que es la distribución de Polya:

$$P(Y = y_i) = \frac{\Gamma(y_i + \omega)}{y_i! \Gamma(\omega)} \left(\frac{\omega}{\omega + \mu_i} \right)^{\omega} \left(\frac{\mu_i}{\omega + \mu_i} \right)^{y_i}$$

  que tiene media $\mathrm{E}(Y_i) = \mu_i$ y varianza $\mathrm{Var}(Y_i) = \mu_i + \mu_i^2/\omega$.

- En el contexto de los modelos GLM's, el parámetro ω de la binomial negativa se supone conocido. Se puede construir un grid de valores para estimar aquel valor de ω que minimice el AIC.

Ejemplo: datos de vínculos con Binomial Negativa I

```
library(MASS)
modbn <- glm(interlocks ~ ., family = negative.binomial(1), data = Ornstein)
theta <- seq(0.5, 2.5, by=0.5) # grid
aics <- rep(0,5)
for (i in seq(along=theta)) aics[i] <- AIC(update(modbn, family=negative.binomial(theta[i])))
rbind(theta,aics)

          [,1]     [,2]     [,3]     [,4]     [,5]
theta    0.500    1.000    1.500    2.000    2.500
aics  1789.075 1721.215 1716.612 1730.192 1750.512
```

Entonces el mínimo AIC se tiene alrededor de $\omega = 1.5$, que nos da la estimación siguiente:

Ejemplo: datos de vínculos con Binomial Negativa II

```
modbnopt <- update(modbn,family=negative.binomial(1.5))
summary(modbnopt)


Call:
glm(formula = interlocks ~ ., family = negative.binomial(1.5),
    data = Ornstein)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7990  -1.1440  -0.2933   0.4637   2.1526

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.256e+00  1.425e-01  15.828  < 2e-16 ***
assets       3.252e-05  5.464e-06   5.950 9.69e-09 ***
sectorBNK   -1.057e+00  5.467e-01  -1.934  0.05437 .
sectorCON   -7.297e-01  4.580e-01  -1.593  0.11248
sectorFIN    6.094e-01  2.350e-01   2.593  0.01011 *
sectorHLD    1.397e-01  3.613e-01   0.387  0.69928
sectorMAN    7.790e-02  1.919e-01   0.406  0.68515
sectorMER    2.051e-01  2.407e-01   0.852  0.39498
sectorMIN    5.217e-01  1.891e-01   2.758  0.00627 **
sectorTRN    5.957e-01  2.470e-01   2.412  0.01666 *
sectorWOD    6.537e-01  2.401e-01   2.722  0.00697 **
nationOTH    8.405e-03  2.401e-01   0.035  0.97211
nationUK    -4.791e-01  2.447e-01  -1.958  0.05140 .
nationUS    -7.862e-01  1.367e-01  -5.751 2.77e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.5) family taken to be 1.02452)

    Null deviance: 487.78  on 247  degrees of freedom
Residual deviance: 322.28  on 234  degrees of freedom
AIC: 1716.6
```

Residuales y gráficas de residuales I

- La mayoría de los diagnósticos que se calculan para los modelos lineales pueden extenderse a los GLM's. Diagnósticos aproximados se basan en mínimos cuadrados ponderados o se derivan de estadísticas que se pueden calcular fácilmente de esta solución.
- Por ejemplo, los apalancamientos (leverages) h_i se toman de la diagonal de la matriz $\mathbf{H} = (\mathbf{X}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}$ y tienen la interpretación usual.
- En los GLMs no hay residuales como en el sentido típico de los modelos lineales. En los modelos lineales, los residuales son: $\hat{y} - y$ para representar el error estadístico $\epsilon = \mathsf{E}(y|\eta) - y$. Pero en los GLM's no hay componente aditivo.
- Hay varios tipos de residuales disponibles para GLMs:
  1. **Residuales de trabajo**: son los que se obtienen del ajuste final de mínimos cuadrados ponderados.
  2. **Residuales respuesta**: $y_i - \hat{\mu}_i$. Este tipo de residuales son los usuales en el modelo gaussiano, pero no se pueden usar para diagnósticos, ya que ignoran que la varianza no es constante.
  3. **Residuales Pearson**: estos se basan en la prueba de bondad de ajuste del modelo. Son los que usualmente se usan con un GLM por su analogía directa con los modelos lineales:

$$e_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var(y_i|\mathbf{x})}/\hat{\phi}}}$$

Se obtienen en `R` con `residuals(modelo,type="pearson")`.

4. **Residuales Pearson estandarizados**: estos corrigen a los anteriores por la varianza condicional y por el leverage de las observaciones:

$$e_{PS,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(y_i|\hat{\mathbf{x}})(1-h_i)}}$$

   Los valores de h_i se toman de la última iteración del método IWLS. A diferencia de los modelos lineales, estos valores dependenden de y y de la configuración de los predictores.
5. **Residuales de la devianza**, $e_{D,i}$: son las raíces cuadradas de los componentes caso por caso de la devianza residual, poniéndoles el signo de $y_i - \hat{\mu}_i$. Se obtienen de `R` con `residuals(modelo, type="deviance")`.
6. **Residuales de la devianza estandarizados**:

$$e_{DS,i} = \frac{e_{D,i}}{sqrt\hat{\phi}(1-h_i)}$$

7. **Residuales estudentizados (aproximados)**:

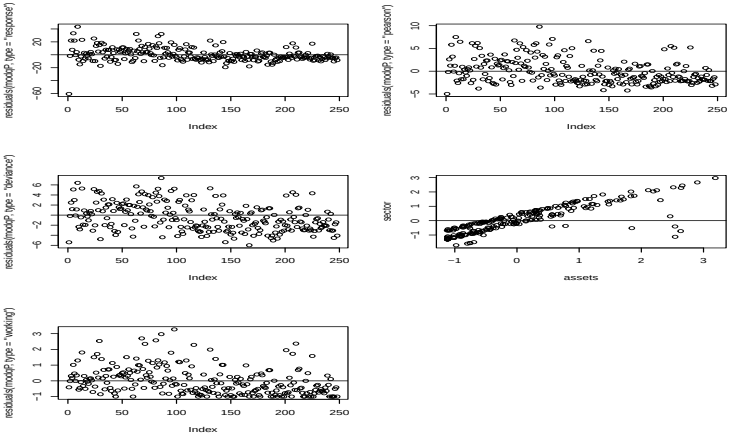
$$e_{T,i} = signo(y_i - \hat{\mu}_i)\sqrt{(1-h_i)e^2_{DS,i} + h_i e^2_{PS,i}}$$

Residuales y gráficas de residuales III

- Por ejemplo, con los datos de Ornstein, algunos de los mencionados:

```
par(mfrow=c(3,2))
plot(residuals(modqP,type = "response"));abline(h=0)
plot(residuals(modqP,type = "pearson"));abline(h=0)
plot(residuals(modqP,type = "deviance"));abline(h=0)
plot(residuals(modqP,type = "partial"));abline(h=0) # con respecto a alguno de los predictores
plot(residuals(modqP,type = "working"));abline(h=0)
```

Residuales y gráficas de residuales IV



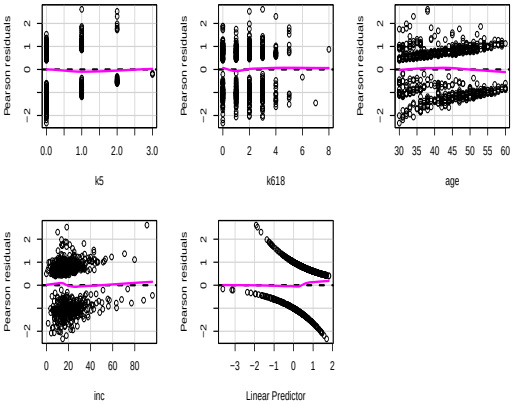
- Una aproximación a las distancias de Cook, para determinar posibles puntos de influencia, es $D_i = \frac{e^2_{PS,i}}{p+1} \times \frac{h_i}{1-h_i}$.

Ejemplo: Fuerza laboral de las mujeres I

- Los siguientes datos son una muestra de un estudio de panel de la Dinámica de Ingreso (Mroz, T. (1987): 'The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions" *Econometrica*, Vol.55, 765-799.). La respuesta es la participación de las mujeres casadas en la fuerza laboral.
  Las variables son:
  - `lfp`: participación de la esposa en la fuerza laboral
  - `k5`: número de niños menores a 5 años
  - `k618`: número de niños de 6 a 18 años
  - `age`: edad de la esposa
  - `wc`: si la esposa asistió al colegio
  - `hc`: si el esposo asistió al colegio
  - `lwg`: logaritmo del salario estimado de la esposa. El logaritmo del salario estimado de la esposa se calcula como el logaritmo de su sueldo real si trabaja, y si no trabaja, entonces se imputa como el valor de predicción de una regresión de los logs de los salarios sobre los otros predictores para mujeres en la fuerza laboral.
  - `inc`: ingreso de la familia excluyendo el ingreso de la esposa

Ejemplo: Fuerza laboral de las mujeres II

```
library(car)
data(Mroz)
mod1 <- glm(lfp ~ k5 + k618 + age + inc, family = binomial(link=logit), data = Mroz)
residualPlots(mod1,layout=c(2,3))
```



Ejemplo: Fuerza laboral de las mujeres III

```
     Test stat Pr(>|Test stat|)
k5      0.6868           0.4073
k618    0.2604           0.6098
age     1.2440           0.2647
inc     1.0664           0.3018
```

- Los resultados se reportan de manera similar a los modelos de regresión lineal. En la salida se menciona que el parámetro de dispersión $\phi = 1$.
- Lo que usualmente se espera ver en estas gráficas para identificar un modelo correcto es una función media condicional constante. En este caso los residuales se ven razonables.
- La función `predict` da predicciones para GLM's. Los valores que devuelve por default son los valores ajustados:

```
# Devuelve los valores ajustados, del predictor lineal:
head(predict(mod1))
```

```
          1           2           3           4           5           6
 0.27112196  1.37323387 -0.29579209  1.22259402 -0.06131959  0.19636291
```

```
# Devuelve las probabilidades ajustadas:
head(predict(mod1,type = "response"))
```

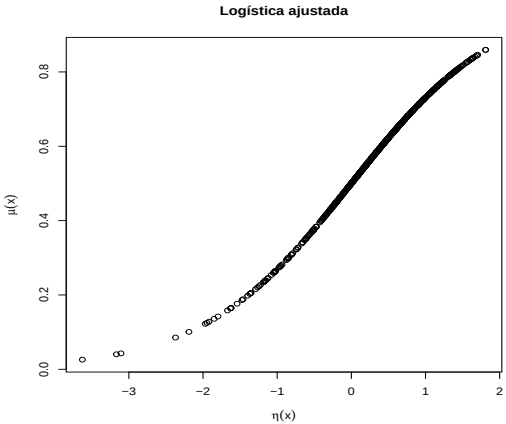
```
        1         2         3         4         5         6
0.5673683 0.7979021 0.4265865 0.7725197 0.4846749 0.5489336
```

Ejemplo: Fuerza laboral de las mujeres IV

- Podemos obtener la curva logística ajustada:

```
plot(predict(mod1), predict(mod1,type = "response"),
     main= "Logistica ajustada",
     xlab = expression(eta(x)),
     ylab = expression(mu(x)))
```

Ejemplo: Fuerza laboral de las mujeres V



Ejemplo: datos binomiales I

- Para datos con respuesta binomial, se cuentan los éxitos en n ensayos. Para especificar un modelo binomial en R se puede hacer de varias formas, como ya comentamos antes:
  1. Una matriz con dos columnas con número de éxitos Y y el número de fracasos $n - Y$.
  2. La respuesta puede ser la proporción Y/n especificando el valor de n en `weights`.
- Sin importar cómo se especifique, la función `glm` considera como respuesta a la proporción de éxitos Y/n y la media de la respuesta $\mu(\mathbf{x})$ se interpreta igual que en el modelo binario. Los siguientes datos corresponden a las votaciones de consultas populares de un nefasto presidente electo (ficticias):

```
votaciones <- data.frame(
      colonia = factor(rep(c("BJ","IZ"),c(3,3))),
      preferencia = factor(rep(c("B","M","A"),2)),
      votaron = c(91,121,64,214,284,201),
      no.votaron = c(39,49,24,87,76,25),
      logit.votacion = log(c(91,121,64,214,284,201)/c(39,49,24,87,76,25)) )
votaciones
```

```
  colonia preferencia votaron no.votaron logit.votacion
1      BJ           B      91         39      0.8472979
2      BJ           M     121         49      0.9039702
3      BJ           A      64         24      0.9808293
4      IZ           B     214         87      0.9000679
5      IZ           M     284         76      1.3182409
6      IZ           A     201         25      2.0844291
```

Ejemplo: datos binomiales II

- También podemos ver los datos como una tabla de contingencia:

```
ftable(xtabs(cbind(votaron,no.votaron) ~ colonia + preferencia ,data=votaciones))
```

```
                    votaron no.votaron
colonia preferencia
BJ      A                64         24
        B                91         39
        M               121         49
IZ      A               201         25
        B               214         87
        M               284         76
```

- Se puede ajustar este modelo agrupado, o se pueden descomponer todos los conteos en el número total de casos para ajustar un modelo logístico.

Ajustamos un modelo binomial con estas variables, considerando la interacción de la colonia y la preferencia. Noten que el modelo sólo tiene 6 datos y hay 6 parámetros a estimar, este es un ejemplo de un modelo saturado, por eso los residuales son 0:

Ejemplo: datos binomiales III

```
mod1 <- glm(cbind(votaron,no.votaron) ~ colonia*preferencia, family = binomial, data = votaciones)
summary(mod1)


Call:
glm(formula = cbind(votaron, no.votaron) ~ colonia * preferencia,
    family = binomial, data = votaciones)

Deviance Residuals:
[1]  0  0  0  0  0  0

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.98083    0.23936   4.098 4.17e-05 ***
coloniaIZ               1.10360    0.31979   3.451 0.000559 ***
preferenciaB           -0.13353    0.30647  -0.436 0.663045
preferenciaM           -0.07686    0.29320  -0.262 0.793212
coloniaIZ:preferenciaB -1.05083    0.39378  -2.669 0.007618 **
coloniaIZ:preferenciaM -0.68933    0.38421  -1.794 0.072791 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.4832e+01  on 5  degrees of freedom
Residual deviance: 4.8850e-14  on 0  degrees of freedom
AIC: 44.093

Number of Fisher Scoring iterations: 3
```

Las predicciones se obtienen de la siguiente manera:

Ejemplo: datos binomiales IV

```
# Todos los valores son exactamente los observados:
predict(mod1, type = "link")

        1         2         3         4         5         6
0.8472979 0.9039702 0.9808293 0.9000679 1.3182409 2.0844291
```

Podemos quitar las interacciones:

Ejemplo: datos binomiales V

```
mod2 <- update(mod1,.~. - colonia:preferencia)
summary(mod2)


Call:
glm(formula = cbind(votaron, no.votaron) ~ colonia + preferencia,
    family = binomial, data = votaciones)

Deviance Residuals:
       1         2         3         4         5         6
 1.27448  -0.02845  -1.71962  -0.88628   0.02167   1.32305

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.4069     0.1814   7.756 8.75e-15 ***
coloniaIZ      0.4067     0.1401   2.903  0.00369 **
preferenciaB  -0.7998     0.1887  -4.239 2.24e-05 ***
preferenciaM  -0.4981     0.1867  -2.668  0.00764 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.8319  on 5  degrees of freedom
Residual deviance:  7.1186  on 2  degrees of freedom
AIC: 47.212

Number of Fisher Scoring iterations: 4
```

Ajustando con pesos:

Ejemplo: datos binomiales VI

```
n <- with(votaciones, votaron + no.votaron)
mod3 <- glm(votaron/n ~ colonia*preferencia, family = binomial, weights = n, data = votaciones)
summary(mod3)


Call:
glm(formula = votaron/n ~ colonia * preferencia, family = binomial,
    data = votaciones, weights = n)

Deviance Residuals:
[1]  0  0  0  0  0  0

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.98083    0.23936   4.098 4.17e-05 ***
coloniaIZ               1.10360    0.31979   3.451 0.000559 ***
preferenciaB           -0.13353    0.30647  -0.436 0.663045
preferenciaM           -0.07686    0.29320  -0.262 0.793212
coloniaIZ:preferenciaB -1.05083    0.39378  -2.669 0.007618 **
coloniaIZ:preferenciaM -0.68933    0.38421  -1.794 0.072791 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.4832e+01  on 5  degrees of freedom
Residual deviance: 4.8850e-14  on 0  degrees of freedom
AIC: 44.093

Number of Fisher Scoring iterations: 3
```