Taller GLM - Grajales

Yosef Guevara Salamanca

1 Expresar la funcion de probabilidad Binomial de la familia exponencial indique la canonica

$$La\ forma\ de\ la\ familia\ Exponencial\ es:$$

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

$$La\ distribución\ binomial\ esta\ dada\ por:$$

$$f(x,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$log(f(y)) = log\left(\binom{n}{x} p^x (1-p)^{n-x}\right)$$

$$log(f(y)) = log\binom{n}{x} + log(p^x) + (n-x)\,log(1-p)$$

$$Se\ genera\ la\ función\ exponencial\ a\ ambos\ lados.$$

$$exp\left\{log\binom{n}{x} + xlog(p) + nlog(1-p) - xlog(1-p)\right\}$$

$$exp\left\{x\,[log(p) - log(1-p)] + nlog(1-p) + log\binom{n}{x}\right\}$$

$$exp\left\{xlog\left(\frac{p}{1-p}\right) + nlog(1-p) + log\binom{n}{x}\right\}$$

$$Finalmente\ se\ obtiene\ que:\ \theta = log\left(\frac{p}{1-p}\right)$$

$$\theta = log\left(\frac{p}{1-p}\right)$$

$$p = \frac{e^\theta}{1+e^\theta}$$

$$1-p = \frac{1+e^\theta - e^\theta}{1+e^\theta} = \frac{1}{1+e^\theta} = \left(1+e^\theta\right)^{-1}$$

$$nlog(1-p) = nlog\left[\left(1+e^\theta\right)^{-1}\right] = -nlog\left(1+e^\theta\right)$$

$$f\,(y) = \; exp\left\{x\theta \; - \; nlog\left(1+e^{\theta}\right) + log\binom{n}{x}\right\}$$

$$Donde\;\theta = \; log\left(\frac{p}{1-p}\right)\; entonces\,p \; = \; \frac{e^{\theta}}{1\;+\;e^{\theta}}$$

$$b\,(\theta) \; = \; nlog\left(1+e^{\theta}\right);\;\; a\,(\phi) = 1;\; c\,(y,\phi) = log\binom{n}{x}$$

2 Expresar la funcion de probabilidad Poisson de la familia exponencial indique la canonica

$$Distribución\;Poisson:\;\; F\,(y,\lambda) \; = \; \frac{\lambda^{\,x}e^{-\lambda}}{y!};\;\; \lambda \; > 0\,;\; y \; = \; 0,1,2....$$

$$Sabiendo\,que\,la\,familia\,exponencial\,esta\,definida\,por:$$

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

$$log\;\;(f\,(y)) = \; ylog\lambda \; - \lambda - \; log\,(y!)$$

$$log\;\;(f\,(y)) = \; exp\,\{ylog\lambda \;\; - \lambda \; - \; log\,(y!)\}$$

$$\theta \;\; = \; log\,\lambda\,;\; b\,(\theta\,) \; = \; \lambda\,;\; \phi \; = \; 1;\;\; c\,(y,\phi\,) \; = \; -log\,(y!)$$

$$Link\;\;Canónico\;\;Poisson:$$

$$g\,(\mu) \; = \; \theta;\; \lambda = e^{\theta}$$

$$log\,(\lambda) = \theta,\; log\,es\,canonico$$

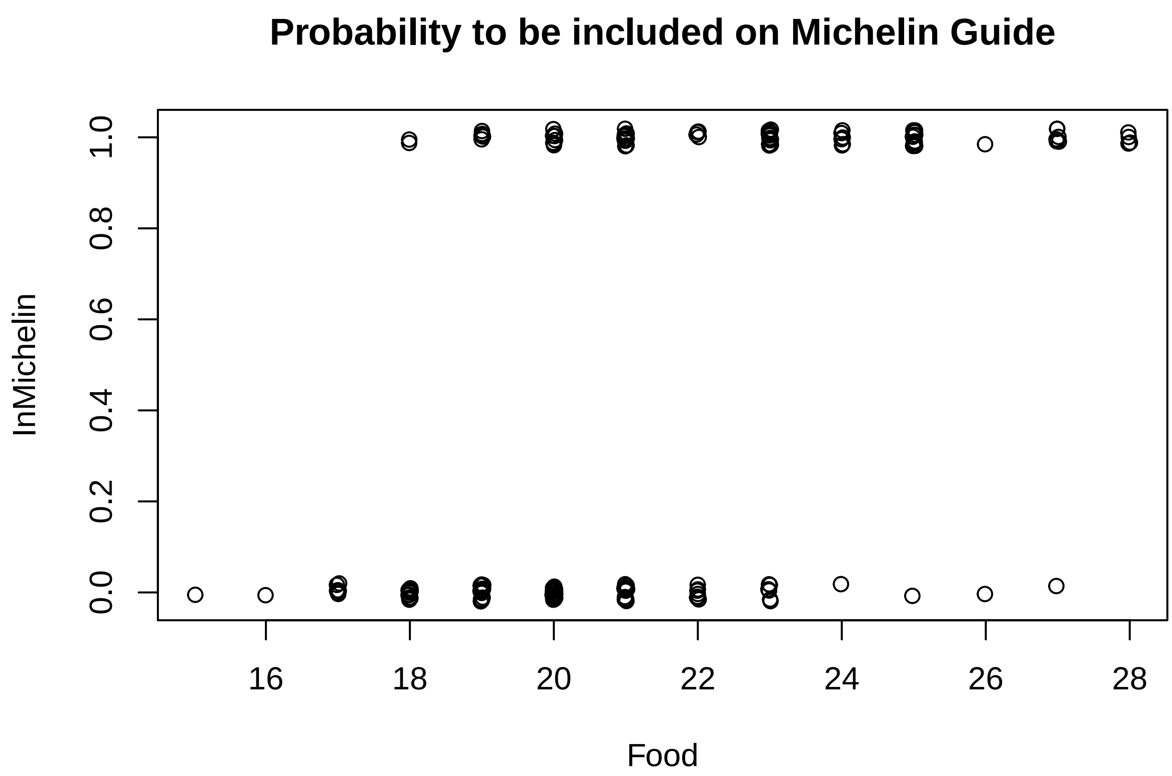
```r
michelin <- read.csv("MichelinNY.csv", header=T,sep=";")
head(michelin)
```

3 Respuesta binomial mismo grafico para comparar funciones Logit, probit, cuchy y clog

```
##   InMichelin Restaurant.Name Food Decor Service Price
## 1          0  14 Wall Street   19    20      19    50
## 2          0             212   17    17      16    43
## 3          0        26 Seats   23    17      21    35
## 4          1              44   19    23      16    52
## 5          0               A   23    12      19    24
## 6          0          A.O.C.   18    17      17    36
```

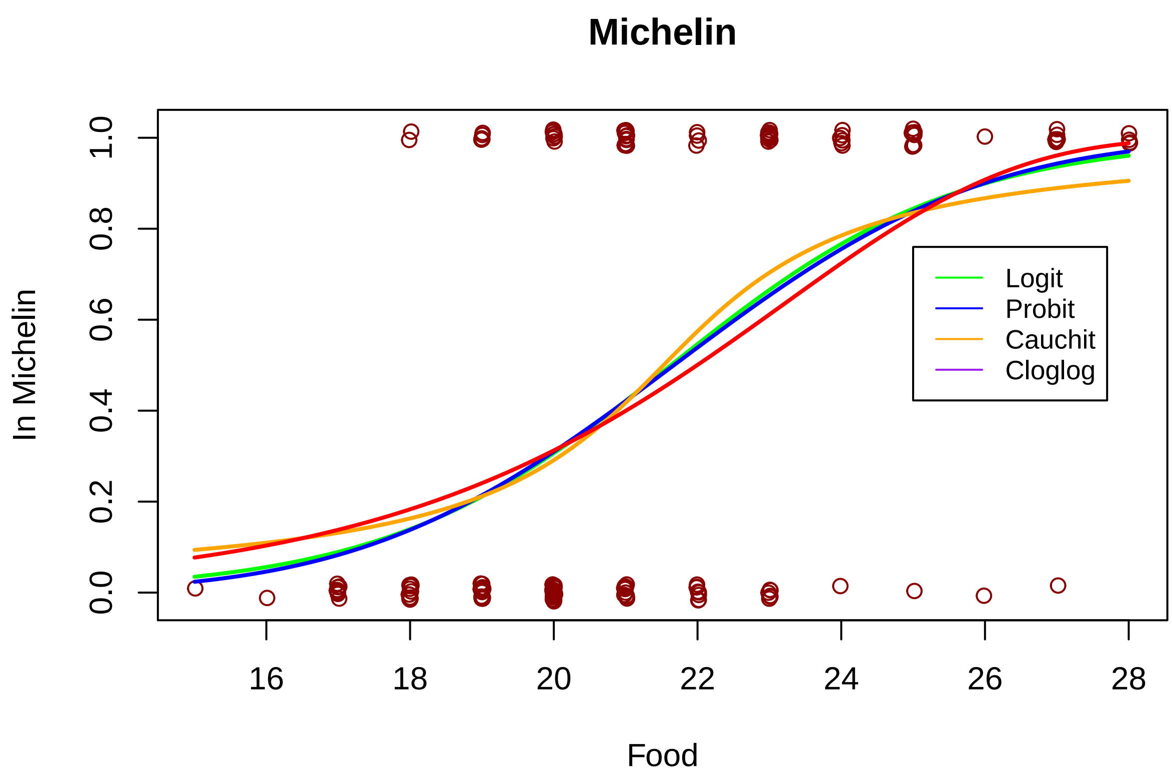
```r
attach(michelin)
```

```r
plot(jitter(InMichelin, 0.1) ~ jitter(Food, 0.1),
  main = "Probability to be included on Michelin Guide",
  xlab = "Food", ylab = "InMichelin")
```




```r
# logit
logit <- glm(InMichelin ~ Food, data = michelin, family = binomial("logit"))

# Probit
probit <-glm(InMichelin ~ Food, family = binomial("probit"))

# Cacchy
cauchit <- glm(InMichelin ~ Food, family = binomial("cauchit"))

# Cloglog
cloglog <-glm(InMichelin ~ Food, family = binomial("cloglog"))
```

```r
plot(jitter(InMichelin,0.1) ~ jitter(Food,0.1), col="red4",
     main="Michelin", ylab="In Michelin", xlab="Food")
newdat  <- data.frame(Food=seq(min(Food), max(Food),len =164))
newdat$logit = predict(logit , newdata=newdat , type="response")
newdat$probit = predict(probit , newdata=newdat , type="response")
newdat$cauchit = predict(cauchit , newdata=newdat , type="response")
newdat$cloglog = predict(cloglog , newdata=newdat , type="response")
lines(logit ~ Food , newdat , col="green", lwd =2)
lines(probit ~ Food, newdat , col="blue", lwd =2)
lines(cauchit ~ Food, newdat , col="orange", lwd=2)
lines(cloglog ~ Food, newdat , col="red", lwd=2)
legend (25, 0.76,  legend=c("Logit", "Probit", "Cauchit","Cloglog"),
        col=c("green", "blue", "orange", "purple"), lty=1, cex =0.8)
```



Semejanza entre **logit** y **probit** ambas graficas se parecen mucho pues su objetivo general es re escalar cualquier numero de tal manera que se genere un intervalo de predicción que caiga entre 0 y 1 basado en una distribución de probabilidad.

La gran diferencia entre **logit** y **probit** es que en el modelo **logit** los errores siguen una distribución logistica acumulativa, mientras que el **probit** se asume que los errores siguen una distribución normal acumulativa

4 Dar un ejemplo aplicado de regresión logística simple. Interpretar el odds ratio.

Para este ejercicio vamos a hacer uso de la base de datos MichelinNY,la cual contiene una lista de 164 restaurante en Nueva York que fueron o no añadidos a la lista de restaurantes con estrellas michelin, la variable respuesta del DB es en InMichelin, donde 0 significa que el restaurante no se encuentra en la gula y 1 que si se encuentra en la guia. La covariable a utilizar es Price (Precio del restaurante)

```
michelin.price <- glm(InMichelin ~ Price, data = michelin, family = binomial("logit"))
summary(michelin.price)
```

```
##
## Call:
## glm(formula = InMichelin ~ Price, family = binomial("logit"),
##     data = michelin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3373  -0.7884  -0.3800   0.8998   2.0553
##
```

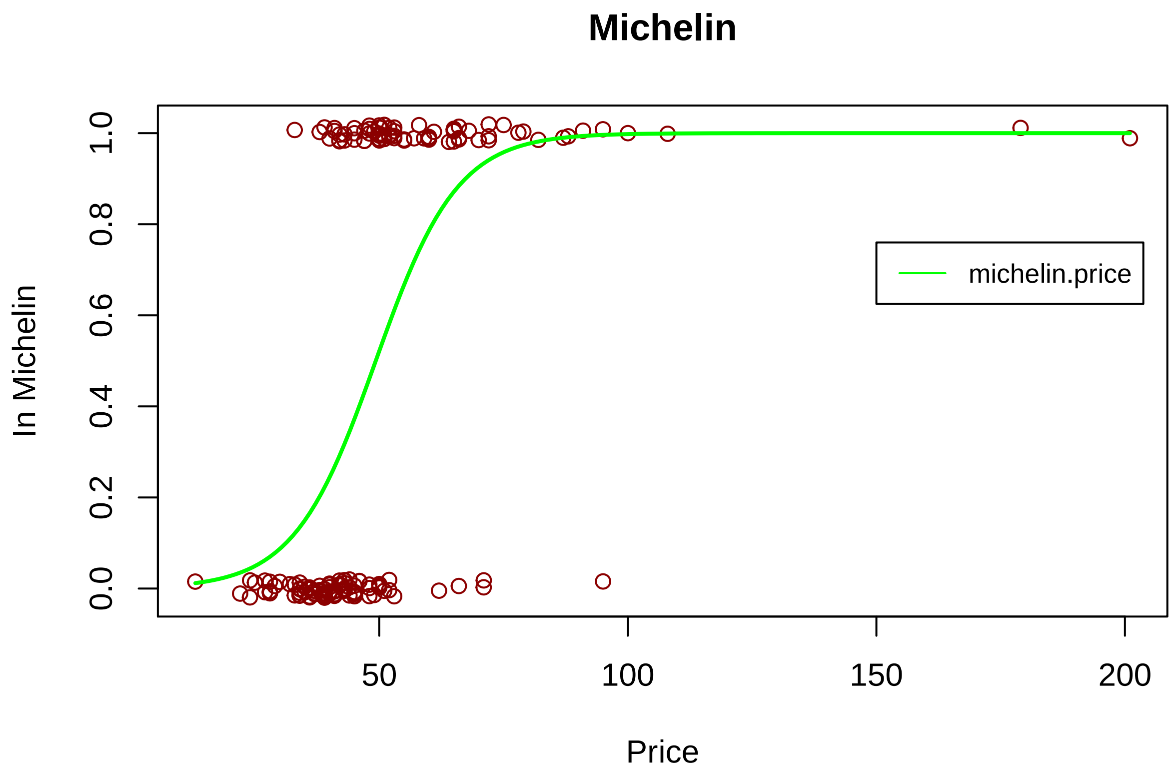
```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.00082    1.03153  -5.817 5.98e-09 ***
## Price        0.12174    0.02184   5.576 2.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.79  on 163  degrees of freedom
## Residual deviance: 161.13  on 162  degrees of freedom
## AIC: 165.13
##
## Number of Fisher Scoring iterations: 6
```

El modelo ajustado esta dado por:

$$\hat{\theta}\left(x\right)=\frac{1}{1+exp\left(-\left(-6.00082+0.12174\cdot\quad Price\right)\right)}$$

Al graficarlo tenemos que:

```r
plot(jitter(InMichelin,0.1) ~ jitter(Price,0.1), col="red4",
     main="Michelin", ylab="In Michelin", xlab="Price")
newdat  <- data.frame(Price=seq(min(Price), max(Price),len =164))
newdat$michelin.price = predict(michelin.price , newdata=newdat , type="response")
lines(michelin.price ~ Price , newdat , col="green", lwd =2)
legend (150, 0.76,  legend=c("michelin.price"),col=c("green"), lty=1, cex =0.8)
```



La funcion linear de x **logit** de los odds estimados esta dada por:

$$\left(\frac{\theta\left(x\right)}{1-\theta\left(x\right)}\right) = exp\left(B_0 + B_1 x\right) \; = \; exp\left(-6.00082 \; + 0.12174 \; \cdot Price\right)$$

Como el predictor lineal solo tiene una variable regresora B1 = 0.12174, se puede calcular el odds ratio de tal manera que por cada unidad de precio adicional la posibilidad de ser incluido en la lista michelin se incrementa en exp(0.12174) = 1.13, es decir un 13%

A su vez si analizamos 2 valores cualquiera dentro del rango de precio de los restaurante tenemos que:

```
b.0 <- -6.00082
b.1 <- 0.12174

odds.ratio.1 <- exp(b.0 + b.1)
odds.ratio.5 <- exp(b.0 + b.1 * 5)

odds.ratio.1/odds.ratio.5
```

```
## [1] 0.6144916
```

El cociente de probabilidad nos dice que es 0.64 veces menos probable ingresar en la guia michelin si el Precio es 1 en lugar de 5

5 Dar un buen ejemplo de regresión logística multiple para razón de verosimilitud. Comparar modelos anidados.

```r
modelo.full <- glm(InMichelin ~ Food + Decor + Service + Price, family = binomial)
summary(modelo.full)
```

```
##
## Call:
## glm(formula = InMichelin ~ Food + Decor + Service + Price, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3923  -0.6723  -0.3810   0.7169   1.9694
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.19745    2.30896  -4.850 1.24e-06 ***
## Food          0.40485    0.13146   3.080  0.00207 **
## Decor         0.09997    0.08919   1.121  0.26235
## Service      -0.19242    0.12357  -1.557  0.11942
## Price         0.09172    0.03175   2.889  0.00387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.79  on 163  degrees of freedom
## Residual deviance: 148.40  on 159  degrees of freedom
## AIC: 158.4
##
## Number of Fisher Scoring iterations: 6
```

Vemos que la covariable Decor y Service no son significativa para el modelo por lo que son descartadas para el modelo reducido

```r
modelo.food.price <- glm(InMichelin ~ Food + Price , family = binomial)
modelo.food <- glm(InMichelin ~ Food, family = binomial)
```

Podemos usar la desviacion del modelo para probar hipótesis sobre subconjuntos de los paramteros del modelo mediante el cociente de verosimilitud dado por:

Donde Ho: El modelo reducido no es mejor pues no aporta más información

$$G = \ -2ln\left(L\left(\frac{Modelo\,reducido}{Modelo\,saturado}\right)\right) \ = \ \chi^2$$

con alpha = 0.05

```r
G <- -2*(logLik(modelo.food)-logLik(modelo.food.price))
chisq.critico <-qchisq (0.95 ,1)
chisq.critico
```

```
## [1] 3.841459
```

```r
G
```

```
## 'log Lik.' 23.18121 (df=2)
```

```r
pchisq(G,1, lower.tail = FALSE)
```

```
## 'log Lik.' 1.474305e-06 (df=2)
```

como G > chisq.critico, se rechaza la hipotesis nula es decir el modelo que solo contiene la covariable Food no es mejor que el que contiene a las covariable Food y Price

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
lrtest(modelo.food , modelo.food.price)
```

```
## Likelihood ratio test
##
## Model 1: InMichelin ~ Food
## Model 2: InMichelin ~ Food + Price
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -87.865
## 2   3 -76.274  1 23.181  1.474e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se verifica esto una ves más con el comando Lrtest, como Pr(>Chisq) = 1.474e-06< 0.05 se rechaza la hipotesis nula

6 Dar un buen ejemplo de regresión probit simple y calcular el ED50 e interpretar

```r
library(GLMsData)
data("turbines")
attach(turbines)

# Ejemplo  del  libro  Generalized  Linear  Models  with  examples  in R

mod.hours  <- glm(Fissures/Turbines ~   Hours , data=turbines , family = binomial(link = "probit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
summary(mod.hours)
```

```
##
## Call:
## glm(formula = Fissures/Turbines ~ Hours, family = binomial(link = "probit"),
##     data = turbines)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.20203  -0.13676  -0.04585   0.07425   0.33158
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.2854920  1.3088282  -1.746   0.0808 .
## Hours        0.0005897  0.0004023   1.466   0.1427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.94214  on 10  degrees of freedom
## Residual deviance: 0.27436  on  9  degrees of freedom
## AIC: 10.467
##
## Number of Fisher Scoring iterations: 5
```

De acuerdo a la ecuacion de la seccion 9.6 Median Effective Dose , ED50 el libro mencionado , pagina 344

```r
b0 =  -2.285492034
b1 = 0.0005897
```

```r
p = 0.5 # para  buscar  el x que  haga  que la  proporcion  sea  50%
g.p = qnorm(p) # esta es la  funcion  enlace  para  probit

ed50 =   (g.p-b0)/b1 # 3875.686
```

Tambien se puede hacer con una ecuacion de la base

```r
library(MASS)
dose.p(mod.hours)
```

```
##              Dose       SE
## p = 0.5: 3875.45 942.2841
```

El ED50 nos informa el tiempo de funcionamiento en el cual se esperara que el 50% de las turbinas comenzara a fallar, en este caso cuando se cumplan 3875 horas el 50% de las turbinas presentaran fisuras

7 Para una respuesta Binomial podemos chequear los datos usando la siguientes igualdad:

$$\sum_{i=1}^{n} = y_i = \sum {}_{i=1}^{n} = \theta_i$$

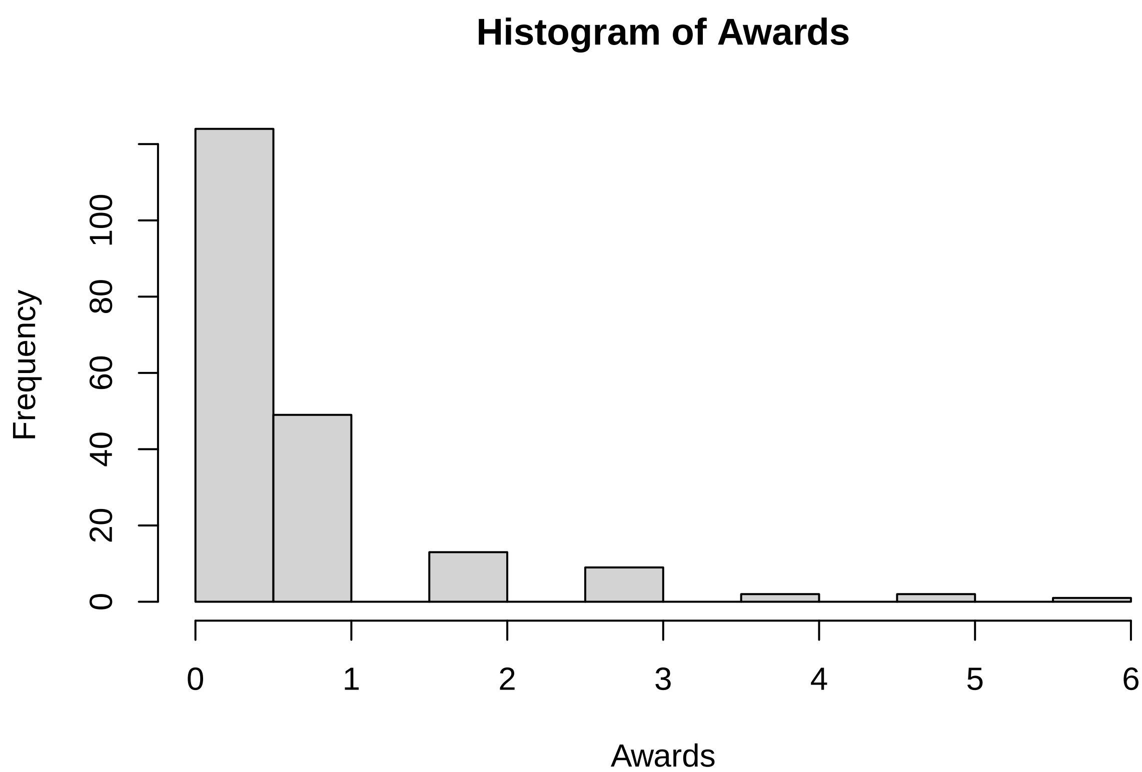
```
m.logit <- glm(InMichelin ~ Food, family = binomial)
m.probit <- glm(InMichelin ~ Food, family = binomial(link = "probit"))
m.cauchit <- glm(InMichelin ~ Food, family = binomial(link = "cauchit"))
m.cloglog <- glm(InMichelin ~ Food, family = binomial(link = "cloglog"))

sum.yi <- sum(InMichelin);
sum.pi.logit <- sum(predict(m.logit, type = "response"))
sum.pi.probit <- sum(predict(m.probit, type = "response"))
sum.pi.cauchit <- sum(predict(m.cauchit, type = "response"))
sum.pi.cloglog <- sum(predict(m.cloglog, type = "response"))

tabla <- rbind(sum.yi, sum.pi.logit, sum.pi.probit, sum.pi.cauchit, sum.pi.cloglog)
diferencia <- c(0,abs(sum.pi.logit -sum.yi), abs(sum.pi.probit -sum.yi), abs(sum.pi.cauchit -sum.yi), a
tabla <- round(cbind(tabla, diferencia),3)
colnames(tabla) <- c("Sumatoria", "Diferencia")
tabla
```

```
##                Sumatoria Diferencia
## sum.yi            74.000      0.000
## sum.pi.logit      74.000      0.000
## sum.pi.probit     73.682      0.318
## sum.pi.cauchit    74.625      0.625
## sum.pi.cloglog    73.556      0.444
```

Podemos ver que para este caso el mejor modelo a escoger es generado por el link **Probit**, pues es el más similiar a la sumatoria de los yi.

8 Dar un buen ejemplo de regresión poisson simple e interpretar el parametro de interes

```
awards <- read.csv("Awards.csv", header = T, sep=";")
attach(awards)
hist(Awards)
```



Gracias al histograma podemos ver que Awards sigue una distribución **Poisson**, es decir que esta positivamente sesgada.

```
m.awards.score <- glm(Awards ~ Score.Math, family = "poisson")
summary(m.awards.score)
```

```
##
## Call:
## glm(formula = Awards ~ Score.Math, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1853  -0.9070  -0.6001   0.3246   2.9529
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.333532   0.591261  -9.021   <2e-16 ***
## Score.Math   0.086166   0.009679   8.902   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 204.02  on 198  degrees of freedom
## AIC: 384.08
```

```
##
## Number of Fisher Scoring iterations: 6
```

El modelo de probabilidad esta dado por:

$$log\,(\theta) = B_0 + B_1 \cdot Score.Math$$
$$log\,(\theta) = -5.3338 + 0.0862 \cdot Score.Math$$

Entonces es posible de decir que por cada unidad que incremente el Score.Math los Awards aunmentaran en exp(0.0862)=1.0899, un 8%

9 Dar un buen ejemplo de regresión poisson multiple e interpretar al menos dos parametros de interés.

```r
m.awards.score.program <- glm(Awards ~ Score.Math + Program, family="poisson")

summary(m.awards.score.program)
```

```
##
## Call:
## glm(formula = Awards ~ Score.Math + Program, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2043  -0.8436  -0.5106   0.2558   2.6796
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.16327    0.66288  -6.281 3.37e-10 ***
## Score.Math         0.07015    0.01060   6.619 3.63e-11 ***
## ProgramGeneral    -1.08386    0.35825  -3.025  0.00248 **
## ProgramVocational -0.71405    0.32001  -2.231  0.02566 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.67  on 199  degrees of freedom
## Residual deviance: 189.45  on 196  degrees of freedom
## AIC: 373.5
##
## Number of Fisher Scoring iterations: 6
```

El modelo lineal esta dado por:

$$log\,(\theta) \;=\; B_0 \;\; + B_1 Score.Math \;+ B_2\;\; Program\,(General) + B_3\,Program\,(Vocational)$$

$$log\,(\theta) \;=\; -4.163 \;+ 0.070 Score.Math \;+ -1.083\;\; Program\,(General) + -0.714\,Program\,(Vocational)$$

Para este modelo cada vez que se incremente en una unidad **Score.Math** el logaritmo de premios incrementa levente un (0.07), cuando el programa es **General** el logaritmo de premios disminuye en un (-1.083) y cada vez que el programa es **Vocacional** el logaritmo disminuye (-0.714)

10 Dar un buen ejemplo de regresión Poisson simple termino **Offset** e interpretar el parametro de interes.

Se incluye un ejemplo del ajuste para modelos Poisson haciendo uso del **Offset**, es decir que este modelo esta escrito en el predictor lineal tal que:

$$log\,\lambda \;=\; B_0 \;+ B_1x_1 + B_2x_2 \;+\; .... \;+\; B_kx_k + \;log\,T$$

Siendo en este caso el termino offset **log T**, se usara el ejemplo del libro de dalgaad como referencia

Descripción del DataSet

This data set contains counts of incident lung cancer cases and population size in four neighbouring Danish cities by age group.

```r
library(ISwR)
data(eba1977)
attach(eba1977)
```

Para ajustar el modelo necesitamos incorporara un **offset** para que se contabilice tanto para los **age**estructuras

```r
m.city.age<-glm(cases~ age , offset = log(pop),family=poisson)
summary(m.city.age)
```

```
##
## Call:
## glm(formula = cases ~ age, family = poisson, offset = log(pop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8520  -0.6424  -0.1067   0.7853   1.5468
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.8623     0.1741 -33.676  < 2e-16 ***
## age55-59      1.0823     0.2481   4.363 1.29e-05 ***
## age60-64      1.5017     0.2314   6.489 8.66e-11 ***
## age65-69      1.7503     0.2292   7.637 2.22e-14 ***
## age70-74      1.8472     0.2352   7.855 4.00e-15 ***
## age75+        1.4083     0.2501   5.630 1.80e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 129.908  on 23  degrees of freedom
## Residual deviance:  28.307  on 18  degrees of freedom
## AIC: 136.69
##
## Number of Fisher Scoring iterations: 5
```

tenemos entonces que el modelo esta dado por, donde se contrasta con age40:54

$$log\,\lambda = -5.86 + (1.08)\,age55:59 + (1.5\;)\,age60:64 + (1.75)\,age65:69 + (1.8472\;\;)\,age70:74 + (1.4083)\,age75$$

Si realizamos la comparacion con el rango age60:64 (B2), lo cual nos quiere decir que el log-rate de indicentes de cancer es 1.5 veces más alto para este grupo

```
# Para B2 en RR esta daddo por
exp(1.5)
```

```
## [1] 4.481689
```

Podemos decir que el grupo entre age60:64 es 4.5 veces mas propenso a incidentes de cancer de pulmón