Análisis de la deviance

El análisis de la deviance es una generalización del análisis de la varianza para los GLM obtenido para una secuencia de modelos anidados (cada uno incluyendo más términos que los anteriores).

Dada una secuencia de modelos anidados usamos la deviance como una medida de discrepancia y podemos formar una tabla de diferencias de deviances.

Sean $M_{p_1}, M_{p_2}, \ldots, M_{p_r}$ una sucesión de modelos anidados de dimensión $p_1 < p_2 < \ldots < p_r$ y matrices de diseño $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}, \ldots, \mathbf{X}_{p_r}$ y deviances $D_{p_1} > D_{p_2} > \ldots > D_{p_r}$.

Suponemos en todos ellos la misma distribución y la misma función link.

Las desigualdades entre las deviances no se verifican necesariamente entre los estadísticos χ^2 de Pearson.

La diferencia $D_{p_i} - D_{p_j}$, $p_j > p_i$, es interpretada como una medida de la variación de los datos explicada por los términos que están en M_{p_j} y no están en M_{p_i}, incluidos los efectos de los términos de que están en M_{p_i} e ignorando los efectos cualquier término que no está en M_{p_j}.

De esta manera, si $D_{p_i} - D_{p_j} > \chi^2_{p_j - p_i, \alpha}$ los efectos de los términos que están en M_{p_j} y no están en M_{p_i} son significativos.

Cada secuencia de modelos corresponde a una tabla de análisis de la varianza diferente. La secuencia de los modelos estará determinada por el interés del investigador.

Residuos

Para el GLM necesitamos extender la noción de residuo a todas las distribuciones que pueden reemplazar a la Normal.

Residuos de Pearson

Los residuos de Pearson se definen como

$$r_i^p = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{Var}(y_i)}},$$

siendo $\widehat{Var}(y_i) = a(\Phi)V(\widehat{\mu}_i)$.

Qué nos queda en el caso Poisson ?

Recordemos que si

$Y \sim P(\mu)$, entonces $E(Y) = \mu = Var(Y)$.

$$\frac{y - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$$

Residuos Deviance

La deviance D puede escribirse como una suma

$$D = \sum_{i=1}^{n} d_i \ ,$$

donde las d_i son los que se conocen como componentes de la *deviance*. Esta medida, tal como vimos suele usarse como una medida de discrepancia en un GLM y en ese sentido d_i es la contribución de cada dato.

Podemos definir los *residuos deviance* como

$$r_i^d = sg(y_i - \widehat{\mu}_i)\sqrt{d_i}\,.$$

Para el caso Poisson, recordemos que

$$\begin{aligned} P(Y=y) &= e^{-\mu}\frac{\mu^y}{y!} = \exp\left(y\log\mu - \mu - \log y!\right) \\ \ell(\mu, y) &= y\log\mu - \mu - \log y! \end{aligned}$$

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

Cada residuo de la deviance resulta

$$r_i^d = sg(y_i - \widehat{\mu}_i)\{2(y_i\log(y_i/\widehat{\mu}_i) - y_i + \widehat{\mu}_i)\}^{1/2}$$

Residuos de Anscombe

Una desventaja de ri^P es que en general su distribución para datos no Normales es asimétrica y por lo tanto no es de esperar que posean propiedades similares a las que poseen bajo normalidad.

Anscombe definió unos residuos baasados en una función $A(y)$ en lugar de y, de manera que la distribución de $A(Y)$ sea tan Normal como sea posible.

Wedderburn demostró que para funciones de verosimilitud en el GLM, la función $A(.)$ es

$$A(.) = \int \frac{d\mu}{V^{1/3}(\mu)}\,.$$

Por ejemplo, en el caso Poisson queda

$$A(.) = \int \frac{d\mu}{\mu^{1/3}} = \frac{3}{2}\mu^{2/3}\,.$$

entonces basaría mos los residuos en $y^{2/3} - \mu^{2/3}$.

La transformación que "normaliza"no es la misma necesariamente que la que

estabiliza varianzas. Por lo tanto, debemos dividir por el desvío de $A(Y)$. Una aproximación de primer orden a esta varianza es $A'(\mu)\sqrt{V(\mu)}$.

En el caso Poisson, resulta

$$r_i^A = \frac{\frac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/6}}\,.$$

Si bien, los residuos de Anscombe y de la deviance parecen muy diferentes, los valores que toman para y y μ dados son muy similares, tal como se muestra en la siguiente tabla.

Caso Binomial

En el caso de la distribución binomial quedaría

$$r_i^p = \frac{y_i - n_i\widehat{\Pi}_i}{\sqrt{n_i\widehat{\Pi}_i(1-\widehat{\Pi}_i)}}\,,$$

$$r_i^d = 2sg(y_i - \widehat{\Pi}_i)\left[y_i \log\left(\frac{y_i}{n_i\widehat{\Pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\widehat{\Pi}_i}\right)\right]$$

Veamos otro ejemplo:

Collett (1991) reporta los datos de un experimento sobre toxicidad en gusanos de la planta de tabaco dosis de *pyrethroid trans–cypermethrin* al que los gusanos empezaron a mostrar resistencia. Grupos de 20 gusanos de cada sexo fueron expuestos a por 3 días al *pyrethroid* y se registró el número de gusanos muertos o knockeados en cada grupo.

Los resultados se muestran en la siguiente tabla.

	dosis (μg)					
sexo	1	2	4	8	16	32
Machos	1	4	9	13	18	20
Hembras	0	2	6	10	12	16

Cuadro 1: Gusanos del tabaco

Ajustamos un modelo de regresión logística usando $\log_2(dosis)$, dado que las dosis so potencias de 2.

Para procesar con S–plus usamos los comandos

```
 options(contrasts=c("contr.treatment", "contr.poly"))
  ldose<- rep(0:5,2)
  numdead<- c(1,4,9,13,18,20,0,2,6,10,12,16)
  sex<- factor(rep(c("M","F"),c(6,6)))
  SF<- cbind(numdead,numalive=20-numdead)

  contrasts(sex)
  M
F 0
M 1
```

Comenzaremos por un gráfico

```
plot(2^ldose, probas,type="n",xlab="dosis",ylab="prob")
lines(2^ldose[sex=="M"],type="p", probas[sex=="M"],col=6)
ines(2^ldose[sex=="F"], probas[sex=="F"],type="p",col=8)
```

Queremos investigar la posibilidad de que haya diferentes pendientes para los dos sexos. Para ello plantearemos y ajustaremos el modelo

$$logit(\pi_i) = \beta_0 + \beta_1 sex + \beta_2 ldose + \beta_3\, sex{:}ldose$$

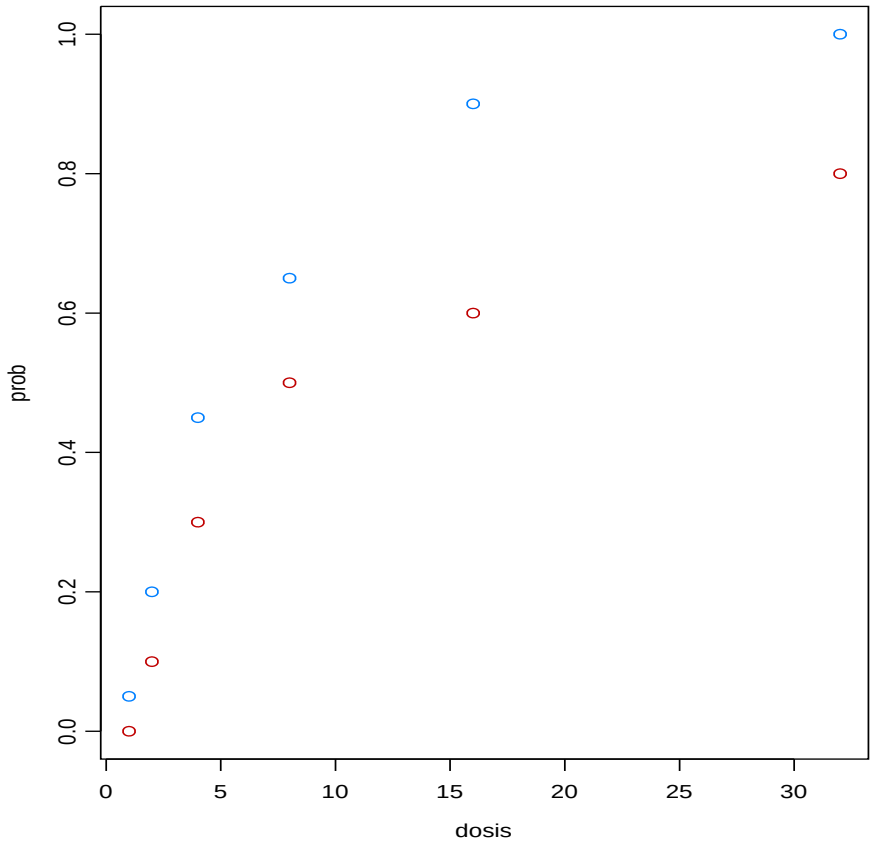


Figura 1: Gusanos del tabaco

de mannera que cuando $sex = M$, para $ldose = 3$ tendríamos

$$logit(\pi_{3,i)} = \beta_0 + \beta_1 + (\beta_2 + \beta_3)3$$

en cambio si $sex = F$, para $ldose = 3$

$$logit(\pi_{3,i}) = \beta_0 + \beta_2 3$$

Para ello hacemos

```
salida.gusanos<- glm(SF~sex*ldose, family=binomial)

summary(salida.gusanos)
```

```
Call: glm(formula = SF ~ sex * ldose, family = binomial)

Coefficients:
                Value Std. Error    t value
(Intercept) -2.9935414  0.5525295 -5.4178852
        sex  0.1749865  0.7781556  0.2248733
      ldose  0.9060363  0.1670577  5.4234939
  sex:ldose  0.3529131  0.2699444  1.3073547

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 4.993727 on 8 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

Aparentemente de la lectura de la tabla el efecto del sexo parece no significativo, sin embarg debemos ser cuidadosos al interpretar esto. Dado que estamos ajustando distintas pendientes para cada sexo, el test individual para este parámetro testea la hipótesis de que las curvas no difieren cuando la log

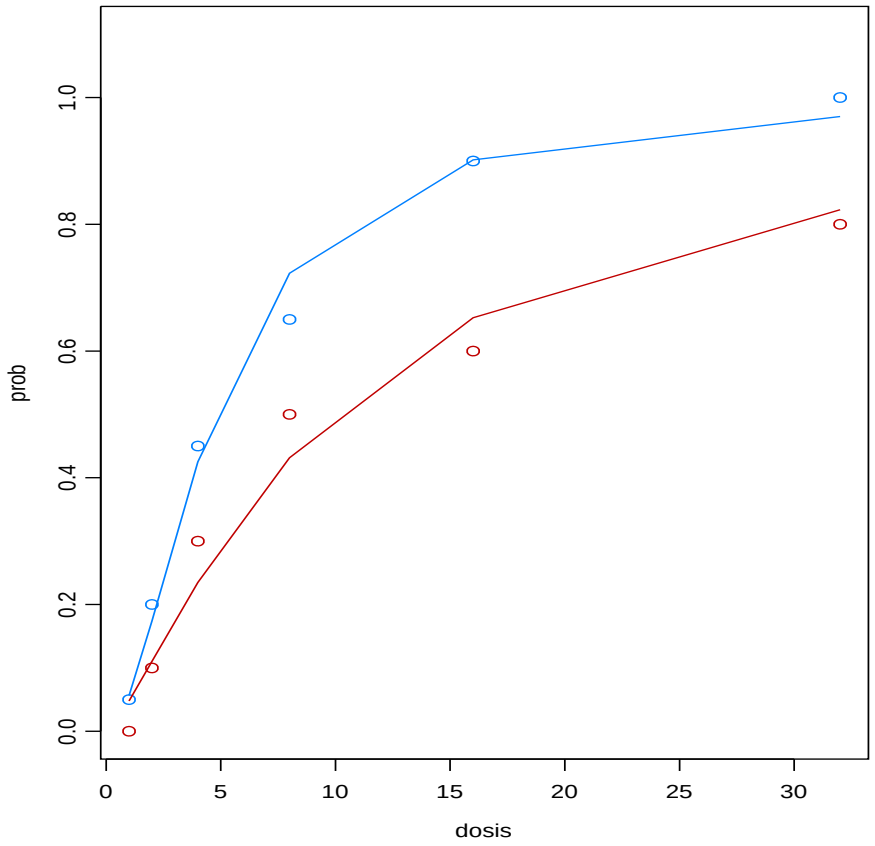


Figura 2: Gusanos del tabaco

dosis es 0. Vamos a reparametrizar de manera de incluir la intercept en una dosis central (8).

```
salida2<- glm(SF~sex*I(ldose-3), family=binomial)
summary(salida2)

Coefficients:
                     Value Std. Error   t value
     (Intercept) -0.2754324  0.2304895 -1.194989
             sex  1.2337257  0.3769412  3.272992
    I(ldose - 3)  0.9060363  0.1670577  5.423494
sex:I(ldose - 3)  0.3529131  0.2699444  1.307355

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 4.993727 on 8 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

que muestra una diferencia significativa entre los dos sexos en la dosis 8. El mod-

elo ajusta muy bien (1-pchisq(4.993727,8)= 0.7582464). Comparamos distintos modelos mediante la instrucción ANOVA

```
 anova(salida.gusanos, test="Chisq")
Analysis of Deviance Table

Binomial model

Response: SF

Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev   Pr(Chi)
     NULL                    11   124.8756
      sex  1   6.0770         10   118.7986 0.0136955
    ldose  1 112.0415          9     6.7571 0.0000000
sex:ldose  1   1.7633          8     4.9937 0.1842088
```

Ahora ajustamos una pendiene para cada sexo:

```
salida3.gusanos<- glm(SF~sex+ldose-1, family=binomial)
summary(salida3.gusanos)

Coefficients:
          Value Std. Error   t value
 sexF -3.473154  0.4682939 -7.416612
 sexM -2.372411  0.3853875 -6.155911
ldose  1.064214  0.1310130  8.122959

    Null Deviance: 126.2269 on 12 degrees of freedom
Residual Deviance: 6.757064 on 9 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

Interpretación de los coeficientes

Supongamos que tenemos una variable independiente que también es dicotómica

Nuestro modelo será

$$\text{logit}(\pi) = \beta_0 + \beta_1 x$$

donde $X = 0$ ó $X = 1$.

Los valores de nuestro modelo son

	$X = 1$	$X = 0$
$Y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$Y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$

Cuadro 2: Variables dicotómicas

El *odds ratio* es

$$\theta = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}$$

que resulta

$$\theta = e^{\beta_1}$$

por lo tanto el logaritmo del *odds ratio* es

$$\log \theta = \beta_1$$

y un intervalo de confianza para θ será

$$\exp(\widehat{\beta}_1 \pm z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\beta}_1)})$$

Consideremos el caso de una variable cualtitativa que toma varios valores, como en la siguiente situación

	blanco	negro	hispanico	otros	Total
Presente	5	20	15	10	50
Ausente	20	10	10	10	50
Total	25	30	25	20	100
θ	1	8	6	4	

Cuadro 3: Ejemplo hipótetico

```
options(contrasts=c("contr.treatment", "contr.poly"))
  yy<- c(5,20,15,10)
  nn<- c(25,30,25,20)
  color<- factor(rep(c("blanco","negro","hipanico","otros"),c(1,1,1,1)))
  SF<- cbind(yy,nyy=nn-yy)

  contrasts(color)

            Variables de Diseno
               D1     D2    D3
          hipanico negro otros
  blanco         0     0     0
hipanico         1     0     0
```

```
   negro         0     1     0
   otros         0     0     1

Call: glm(formula = SF ~ color, family = binomial)
Coefficients:
                  Value Std. Error   t value
  (Intercept) -1.386294  0.4999999 -2.772589
colorhipanico  1.791759  0.6454971  2.775782
   colornegro  2.079442  0.6324554  3.287886
   colorotros  1.386294  0.6708203  2.066566

    Null Deviance: 14.04199 on 3 degrees of freedom
Residual Deviance: 0 on 0 degrees of freedom
```

Veamos que

$$\exp(2.079442) = 8.000004$$
$$\exp(1.791759) = 5.999997$$
$$\exp(1.386294) = 3.999999$$

Observemos además que como

$$\text{logit}(\pi) = \beta_0 + \beta_{11}D_1 + \beta_{12}D_2 + \beta_{13}D_3$$

$$\begin{aligned}
\log \widehat{\theta}(negro, blanco) &= \\
&= \beta_0 + \beta_{11}(D_1 = 0) + \beta_{12}(D_2 = 1) + \beta_{13}(D_3 = 0) \\
&\quad - [\beta_0 + \beta_{11}(D_1 = 0) + \beta_{12}(D_2 = 0) + \beta_{13}(D_3 = 0)] \\
&= \beta_{12}
\end{aligned}$$

y en base a la distribución asintótica de los parámetros podemos obtener un intervalo de confianza para $\theta(negro, blanco)$.

Qué podemos hacer cuando la variable es continua o discreta con muchos valores posibles?

El siguiente ejemplo corresponde al TP3 y se ha registrado la variable edad en forma discreta. Las vaiable independiente es **Age** y la dependiente **Low**. Primero consideraremos los cuartiles de la variable.

```
Analisis de cuartiles para Age:
> summary(age)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
   14      19     23 23.24      26   45

edad<- 1*(age<19)+2*(age>= 19 & age<23) +3*(age>= 23 & age<26)+ 4*(age>=26)
table(edad)
  1  2  3  4
 35 59 41 54
```

```
table(edad,low)
   0  1
1 23 12
2 41 18
3 25 16
4 41 13

> (23*18)/(41*12)
[1] 0.8414634
> (23*16)/(25*12)
[1] 1.226667
> (23*13)/(41*12)
[1] 0.6077236

> contrasts(edad)<- contr.treatment(4)
> contrasts(edad)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

```
Edad  y    n-y
1 23.00 12.00
2 41.00 18.00
3 25.00 16.00
4 41.00 13.00

summary(glm(sf~edad,family=binomial))

Call: glm(formula = sfchd ~ raza, family = binomial)
Coefficients:
                Value Std. Error    t value
(Intercept)  0.6505876  0.3561062  1.8269484
      edad2  0.1726127  0.4547058  0.3796141
      edad3 -0.2043005  0.4788649 -0.4266349
      edad4  0.4980351  0.4776242  1.0427342

exp(-0.1726127)= 0.8414635
exp(0.2043005)= 1.226667
exp(-0.4980351)= 0.6077236
##############################################################
```

```
Intervalos de Confianza

cbind(exp(-0.1726127-1.96* 0.4547058),exp(-0.1726127+1.96* 0.4547058))

(0.3451293, 2.051581)

cbind(exp(0.2043005-1.96* 0.4788649),exp(0.2043005+1.96* 0.4788649))

(0.4798534, 3.135773)

cbind(exp(-0.4980351-1.96* 0.4776242),exp(-0.4980351+1.96* 0.4776242))

(0.238311, 1.549773)
```

Observemos que el 1 pertenece a todos los intervalos de confianza!!

Otro ejemplo

cuartil	20-34	35-44	45-54	55-64	total
Si	3	8	11	21	43
no	22	19	10	6	57
total	25	27	21	27	100
θ	1	3.1	8.1	25.7	
$\log\theta$	0.0	1.1	2.1	3.2	

Cuadro 4: Ejemplo hipótetico

```
attach(chd)
edadf<- factor(edad)
contrasts(edadf)<- contr.treatment(4)
contrasts(edadf)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

```
sf<- cbind(y,ny)
summary(glm(sf~edadf,family=binomial))

Coefficients:
                Value Std. Error   t value
(Intercept) -1.992430  0.6154535 -3.237337
     edadf2  1.127433  0.7459320  1.511442
     edadf3  2.087740  0.7547795  2.766027
     edadf4  3.245193  0.7701095  4.213937
```

Como los puntos medios de los intervalos son casi equidistantes podemos usar polinomios ortogonales.

```
contrasts(edadf)<- contr.poly(4)
contrasts(edadf)

           D1   D2         D3
          .L   .Q         .C
1 -0.6708204  0.5 -0.2236068
2 -0.2236068 -0.5  0.6708204
3  0.2236068 -0.5 -0.6708204
4  0.6708204  0.5  0.2236068

Call: glm(formula = sf ~ edadf, family = binomial)

Coefficients:
                  Value Std. Error     t value
(Intercept) -0.37733861  0.2451542 -1.53918882
    edadf.L  2.39167304  0.5341423  4.47759570
    edadf.Q  0.01501003  0.4903084  0.03061345
    edadf.C  0.08145331  0.4421501  0.18422094
```

En este caso sólo el coeficiente que corresponde al término lineal es significativo!!!!

Como en regresión lineal al graficar los residuos vs. el predictor lineal $\widehat{\eta}$ esperamos encontrar una banda horizontal, más o menos paralela al eje de las abscisas alrededor del 0.

Podríamos encontrar una curvatura o un ancho de la banda variable.

Una curvatura podría sugerir:

1. elección incorrecta de la función de enlace
2. omisión de algún término no lineal de una covariable

El ancho de banda variable puede sugerir que la función de varianza es incorrecta.

También estos gráficos pueden ayudar a detectar residuos muy grandes, es decir mayores que 2 ó 3.

Otra posibilidad es graficar los residuos vs. cada covariable por separado, tal como lo hacíamos en Modelo Lineal.

Una curvatura en este gráfico nuevamente puede sugerir que la variable en estudio puede entrar en el modelo como x^2, o $\sqrt{x}$ o $\log x$.

Problemas con la función de varianza

Como en el modelo lineal el gráfico del valor absoluto de los residuos vs. $\widehat{\mu}$ puede ser útil para detectar problemas en la función de varianza.

Un gráfico sin ninguna tendencia indicaría una función de varianza correcta. En cambio, por ejemplo, una tendencia positiva sugeriría utilizar una función de varianza que aumente más rapidamente. Debemos tener en cuenta que dentro de una familia particular de distribuciones no es posible cambiar la función de varianza, sino que ésta está fijada por el modelo.

En el GLM la situación es muy parecida a la del Modelo Lineal: si la función de varianza no es la correcta el estimador de $\boldsymbol{\beta}$ será asintóticamente insesgado y normal, pero no eficiente. Así mismo, tendremos problemas con $Var(\widehat{\boldsymbol{\beta}})$.

Estimación e interpretación de los coeficientes en presencia de interacción

Como ya hemos visto en el ejemplo de toxicidad es posible que haya interacción entre dos variables independientes.

En este caso , cómo se estiman los odds ratios y se calculan sus intervalos de confianza? Por simplicidad supondremos que tenemos sólo dos variables.

Consideremos el caso en que tenemos un factor de riesgo F, una covariable X y su interacción $F \times X$. El logit para el caso en que $F = f$ y $X = x$ será

$$\text{logit}(f, x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f x$$

Si fijamos $X = x$ los log odds de $F = f_1$ versus $F = f_0$ será

$$\begin{aligned} \log \theta(F = f_1, F = f_0, X = x) &= \text{logit}(f_1, x) - \text{logit}(f_0, x) \\ &= \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0) \end{aligned}$$

por lo tanto

$$\theta(F = f_1, F = f_0, X = x) = e^{\beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0)}\,.$$

Para calcular un intervalo de confianza necesitamos estimar la varianza de estimador:

$$\begin{aligned} &\widehat{Var}(\log\widehat{\theta}(F = f_1, F = f_0, X = x)) = \\ = \; &[f_1 - f_0]^2 \widehat{Var}(\widehat{\beta}_1) + [x(f_1 - f_0)]^2 \widehat{Var}(\widehat{\beta}_3) + 2x(f_1 - f_0)^2 \, \widehat{Cov}(\widehat{\beta}_1, \widehat{\beta}_3). \end{aligned}$$

Un intervalo de de confianza de nivel aproximado para θ puede ser calculado como

$$\exp[\widehat{\beta}_1(f_1 - f_0) + \widehat{\beta}_3 x(f_1 - f_0) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\log\widehat{\theta}(F = f_1, F = f_0, X = x))}]$$

Si F es un factor dicotómico y $f_1 = 1$ y $f_2 = 0$, entonces estas expresiones se simplifican a

$$\log\theta(F = 1, F = 0, X = x) \; = \; \beta_1 + \beta_3 x$$

por lo tanto

$$\theta(F=1,F=0,X=x)=e^{\beta_1+\beta_3\,x}$$

la varianza de estimador

$$\widehat{Var}(\widehat{\beta}_1)+x^2\widehat{Var}(\widehat{\beta}_3)+2x\;\widehat{Cov}(\widehat{\beta}_1,\widehat{\beta}_3).$$

y el intervalo de de confianza de nivel aproximado

$$\exp\left[\widehat{\beta}_1+\widehat{\beta}_3 x\pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}(\log\widehat{\theta}(1,0,X=x))}\right]$$

Algunas estrategias para construir un modelo en regresión logística

Hosmer y Lemeshow (1989) sugieren algunas estrategias a la hora de ajustar un modelo de regresión logística. Enumeramos algunas de ellas:

- Recomiendan comenzar por un análisis cuidadoso de cada variable a través de un ajuste univariado. Para variables nominales, ordinales y continuas con muy pocos valores sugieren hacerlo a través de una tabla de contingencia

para la respuesta ($y = 0, 1$) y los k valores de la variable independiente. Además de realizar un test de ajuste global (cociente de verosimilitud), para aquellas variables que exhiben un moderado nivel de asociación, proponen estimar los odds ratios usando uno de los niveles como referencia.

- En este punto sugieren tener mucho cuidado con aquellas tablas de contingencia que tienen alguna casilla con 0. Una estrategia para evitar esto puede ser colapsar algunas categorías de la variable independiente de alguna manera razonable o eliminar la categoría completamente.
- Cuando la variable es continua puede hacerse un gráfico suavizado, dividiendo a la variable independiente en clases o intervalos. Hemos visto las versiones más sencillas de estos plots. Si la escala es logit servirá para evaluar graficamente la importancia de la variable y si la escala es apropiada.
- Una vez realizado el análisis univariado seleccionan las variables para un análsis multivariado. Recomiendan como candidato para la regresión multivariada a toda variable que en el test univariado tenga un p–valor < 0.25, así como a toda variable que se sepa es importante desde el punto de vista biológico (o del problema).

  Una vez que todas estan variables han sido identificadas, comienzan con un

modelo multivariado que las contiene a todas.

Este punto de corte 0.25 fue sugerido por Mickey and Greenland (1989). El uso de un punto tan grande (el usual es 0.05)tiene la desventaja de que pueden introducirse variables de dudosa importancia.

Un problema de la aproximación por los modelos univariados es que variables que están en forma individual débilmente asociadas con la respuesta pueden ser predictores importantes cuando se consideran en forma conjunta.

Por este motivo, debe revisarse la incorporación de todas las variables antes de arribar a un modelo final.

- La importancia de cada variable en el modelo multivariado puede ser evaluada a través del estadístico de Wald de cada una y una comparación del coeficiente estimado del modelo multivariado con el coeficiente estimado en el modelo univariado que sólo contiene esa variable.

  Hosmer y Lemeshow sugieren eliminar las variables que no contribuyen al modelo cuando nos basamos en estos criterios y ajustar un nuevo modelo. Proponen comparar los coeficientes estimados de las variables que quedan

en el nuevo modelo con los estimados en el viejo modelo. En particular, deberíamos preocuparnos por aquellas variables que cambian mucho en magnitud. Esto podría indicar que algunas de las variables eliminadas son importantes en el efecto de las variables restantes en el ajuste.

Este procedimiento de eliminación, reajuste y verificación continúa hasta que parezca que las variables importantes han sido incluidas y las excluidas son las biológica o estadísticamente sin importancia.

- En general, la decisión de comenzar con todas las variables posibles depende de la cantidad de observaciones. Cuando los datos no son adecuados para soportar este análisis, podría llegarse a resultados inestables: los estadísticos de Wald no serían adecuados para la selección de las variables. En este caso habría que refinar los resultados del análsis univariado y ver que es lo relevante desde el punto de vista científico.
- Un análisis alternativo puede ser utilizar un *método stepwise* en el que las variables son incluidas o excluidas secuencialmente de manera de poder identificar un modelo *full* y luego proceder como hemos descripto.
- Para las variable continuas deberemos chequear el supuesto de linealidad.

Box–Tidwell (1962) sugieren incorporar un término de la forma $xln(x)$ y ver si su coeficiente es significativo o no. Un coeficiente significativo daría evidencias de no linealidad. Sin embargo, advierten sobre la falta de potencia del método para detectar pequeños apartamientos de la linealidad.

- Una vez que obtenemos un modelo que creemos que contiene las variables esenciales deberemos considerarar la necesidad de incorporar interacciones entre ellas. Sugieren incorporar la interacción y evaluar su significación en términos del cociente de verosimilitud. Ellos recomiendan no incorporar interacciones cuyo único efecto es agrandar los errores standard sin cambiar el valor estimado. En su experiencia para que una interacción cambie el valor estimado y los estimadores por intervalo el coeficiente estimado de la interacción debe ser al menos moderadamente significativo.

Observaciones Agrupadas en el caso Binomial

Como hemos visto cuando las variables son discretas puede haber replicaciones. Podemos encontrar que algunas de nuestras n observaciones toman el mismo valor en x_i. Si llamamos $x_1^*, \ldots, x_m^*$ a los valores distintos de las covariables (sin tener en cuenta las repeticiones), $m \leq n$, podemos comprimir los valores de las respuesta en

$$y_i^* = \sum_{j:x_j=x_i^*} y_j \quad n_i^* = \sum_{j:x_j=x_i^*} n_j \, .$$

Si los n_i^* son grandes podremos tener estadísticos de bondad de ajuste X^2 o G^2 bien aproximados. Como ya observamos, estos estadísticos tendrán $m - p$ grados de libertad en lugar de $n - p$.

Si el modelo es cierto, al colapsar los valores con igual x_i no hay pérdida de información al sumar las Y_i's correspondiente. Sin embargo, si el modelo no es cierto, las Π_i's de observaciones con igual x_i's no serán necesariamente idénticas y en ese caso no será fácil detectar apartamientos al modelo.

El hecho de agrupar observaciones también puede limitar la posibilidad de detectar sobredispersión, que ocurre cuando las variables Y_i's tienen varianza

mayor que $n_i\Pi_i(1-\Pi_i)$.

Una posibilidad para detectar sobredispersión es examinar $\frac{y_i}{n_i}$ en observaciones con igual x_i, lo que no se puede hacer si se agrupa.

La falta de ajuste del modelo se pude deber a:

- covariables omitidas
- función link incorrecta
- presencia de outliers
- sobredispersión

Sobredispersión

Algunas veces la falta de ajuste se debe a sobredispersión, que es un fenómeno que no conocíamos en el contexto del modelo lineal clásico, pues σ no está sujeta a una relación con los β's.

Cuando tenemos respuestas dicretas, como la Binomial o la Poisson la media y la varianza están fuertemente ligadas y puede ocurrir sobredispersión (o eventualmente subdispersión, pero este fenómeno es menos frecuente).

La sobredispersión puede ser tratada de dos formas:

- sumergir a la variable de respuesta en un modelo que contemple una distribución más rica y que contemple una dispersión mayor
- usar la teoría de quasi–verosimilitud.

En el primer caso, por ejemplo, si tenemos un modelo Binomial podría mos ampliarlo a un Beta–Binomial y si tenemos un Poisson podríamos considerar un modelo Binomial Negativo.

En el segundo caso, la quasi–verosimilitud permite establecer una relación media–varianza sin suponer una distribución determinada para las respuestas.

Quasi–verosimilitud

Sea $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ un vector de variables aleatorias con media $\mu = E(\mathbf{Y}) = (\mu_1, \ldots, \mu_n)'$ y matriz de covarianza $\Sigma_{\mathbf{Y}} = \sigma^2 V(\mu)$, donde $V(\mu)$ es definida positiva cuyos elementos son funciones conocidas de μ y σ^2 es una constante de proporcionalidad. $V(\mu)$ recibe el nombre de **función de covarianza**.

Si las Y_i's son independientes tendremos que

$$V(\mu) = diag(V(\mu_1), \ldots, V(\mu_n)) \,.$$

En general tendremos que $\mu = g(.)$ es una función conocida de p parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Es usual que esta función tenga una componente lineal que involucre una matriz de diseño $\mathbf{X} \in I\!R^{n \times p}$, de manera que

$$\mu = g(X\boldsymbol{\beta}) \,.$$

Sean $\mathbf{y} = (y_1, \ldots, y_n)'$ el vector de observaciones. Para cada y_ℓ definimos la función de quasi–verosimilitud, $L^*(\mu_\ell, y_\ell)$, como

$$\frac{\partial L^*(\mu_\ell, y_\ell)}{\partial \mu_\ell} \;=\; \frac{y_\ell - \mu_\ell}{V(\mu_\ell)} \tag{8}$$

donde $Var(Y_\ell) = \sigma^2 V(\mu_\ell)$

El logaritmo de la función de quasi–verosimilitud para las n observaciones se define a través del sistema de ecuaciones diferenciales:

$$\frac{\partial L^*(\boldsymbol{\mu}, \mathbf{y})}{\partial \boldsymbol{\mu}} \;=\; V^{-1}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})$$

Como en este caso estamos suponinedo que las observaciones son independientes obtendremos que

$$L^*(\boldsymbol{\mu}, \mathbf{y}) \;=\; \sum_{\ell=1}^{n} L^*(\mu_\ell, y_\ell)\,.$$

Integrando $\dfrac{y_\ell - \mu_\ell}{V(\mu_\ell)}$ respecto de μ_ℓ nos queda

$$L^*(\mu_\ell, y_\ell) \;=\; y_\ell\, \theta_\ell - b(\theta_\ell) + c(y_\ell, \phi)$$

donde

$$\begin{aligned} \theta_\ell &= \int V^{-1}(\mu_\ell) d\mu_\ell \\ b'(\theta_\ell) &= \mu_\ell \\ b''(\theta_\ell) &= \frac{\partial \mu_\ell}{\partial \theta_\ell} = V(\mu_\ell) \end{aligned}$$

Por lo tanto, la densidad de Y_ℓ puede escribirse como una familia exponencial a un parámetro. La recíproca también es cierta. Luego, suponer que las observaciones tienen una distribución en una familia exponencial exponencial, simplemente es suponer una relación varianza–media en los datos.

Suponer una relación en los datos puede ser difícil, sin embargo una relación media–varianza puede ser más fácilmente postulada.

En la siguiente tabla vemos algunos ejemplos:

Propiedades

Sea $L^*_\ell = L^*(\mu_\ell, y_\ell)$ la log–quasi–verosimilitud de una única observación . Entonces

1. $E(\dfrac{\partial L^*_\ell}{\partial \beta_j}) = 0$

2. $E(\dfrac{\partial L^*_\ell}{\partial \beta_j}\dfrac{\partial L^*_\ell}{\partial \beta_k}) = -\sigma^2 E(\dfrac{\partial^2 L^*_\ell}{\partial \beta_j \partial \beta_k}) = \sigma^2 V^{-1}(\mu_\ell)\dfrac{\partial \mu_\ell}{\partial \beta_j}\dfrac{\partial \mu_\ell}{\partial \beta_k}$

La cantidad de **2.** es una medida de la información cuando sólo se conoce la relación media–varianza.

Scores basados en L^*

Se pueden definir los scores basados en L^* que serán los *quasi–scores* como

$$U^*(\boldsymbol{\beta}) = \frac{\partial L^*(\boldsymbol{\mu}, \mathbf{y})}{\partial \boldsymbol{\beta}}\,.$$

De lo anterior obtenemos que

$$U^*(\boldsymbol{\beta}) = \mathbf{D}'\, V^{-1}(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})$$

donde $\mathbf{D} = \dfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$ es una matriz de $n \times p$.

Tenemos que

$$\begin{aligned} E[U^*(\boldsymbol{\beta})] &= 0 \\ \Sigma_{U^*(\boldsymbol{\beta})} &= \sigma^2 \mathbf{D}'\, V^{-1}(\boldsymbol{\mu})\, \mathbf{D} \end{aligned}$$

Observemos que $U^*(\boldsymbol{\beta})]$ es una suma de v.a. con media 0 y varianza finita. McCullagh (1983) mostró bajo condiciones más generales que **asintóticamente**

$$U^*(\boldsymbol{\beta}) \overset{(a)}{\sim} N_p(0, \sigma^2 \mathbf{D}'\, V^{-1}(\boldsymbol{\mu})\, \mathbf{D})\,.$$

Estimación e Inferencia por MQV

La log–quasi–verosimilitud puede ser utilizada de la misma forma que la log–verosimilitud.

La estimación por MQV consiste en resolver el sistema

$$\frac{\partial L^*(\boldsymbol{\mu}, \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{D}' V^{-1}(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu}) = 0$$

Notemos que en esta instancia no es necesario conocer ni $L^*(\boldsymbol{\mu}, \mathbf{y})$ ni σ^2.

Si aplicamos Fisher–scoring, si $\boldsymbol{\beta}_0$ es un valor inicial el del paso siguiente lo obtenemos:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \left[\mathbf{D}_0' V^{-1}(\boldsymbol{\mu}_0)\mathbf{D}_0\right]^{-1} \mathbf{D}_0' V^{-1}(\boldsymbol{\mu}_0)(\mathbf{Y} - \boldsymbol{\mu}_0)$$

Si llamamos $\widetilde{\boldsymbol{\beta}}$ al estimador resultante, McCullagh (1983) probó que asintóticamente

$$\widetilde{\boldsymbol{\beta}} \overset{(a)}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{D}' V^{-1}(\boldsymbol{\mu})\mathbf{D}\right)^{-1})$$

y que la deviance para el modelo de quasi–verosimilitud

$$D(\mathbf{y}, \widetilde{\boldsymbol{\mu}}) = 2\ [L^*(\mathbf{y}, \mathbf{y}) - L^*(\widetilde{\boldsymbol{\mu}}, \mathbf{y})] \overset{(a)}{\sim} \sigma^2 \chi^2_{n-p}$$

Cuando σ^2 no es conocido propone estimarlo como

$$\widetilde{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \widetilde{\mu_i})^2 / V_i(\widetilde{\mu_i}) = \chi^2 / n - p$$

donde χ^2 es el estadístico generalizado de Pearson.

Volviendo al caso Binomial

En el modelo binomial, sobredispersión significa que

$$V(Y_i) = \sigma^2\ \mu_i(n_i - \mu_i)/n_i\,,$$

con $\sigma^2 > 1$.

Si especificamos esta función de varianza, el método de quasi–likelihood da lugar al mismo estimador que máxima verosimilitud usando el algoritmo de Fisher–scoring, sin embargo la matriz de covarianza si cambiará a $\sigma^2(\mathbf{X'WX})^{-1}$.

Los tests para modelos anidados pueden basarse en G^2/σ^2 comparando con una distribución χ^2 con tantos grados de libertad como la diferencia entre la cantidad de parámetros de ambos modelos.

Estimación de σ^2

Como vimos

$$\widetilde{\sigma}^2 = \chi^2/n - p$$

que es el estadístico de Pearson común que usamos para evaluar la bondad del ajuste.

Si el modelo es válido, éste es un estimador consistente de σ^2, mientras que el equivalente basado en $G^2/n - p$ no lo es. Cuando hay importantes covariables omitidas, χ^2 puede crecer mucho y por lo tanto, σ^2 podría ser sobreestimado. Por ello, algunos autores recomiendan estimar a σ^2 bajo un **modelo maximal** que incluya todas las covariables que nos interesan, pero que que no sea el saturado.

Qué pasa si los datos son no agrupados ($n_i = 1$)?

Mccullagh y Nelder (1989) dicen que en este caso no es posible la sobredispersión, en tanto el único modelo que sostiene como valores posibles 0 o 1 es el Bernoulli.

Por lo tanto, cuando las observaciones no están agrupadas asumimos que $\sigma^2 = 1$.

Schafer (2000) recomienda que antes de hacer el procedimento de selección de variables, se ajuste un modelo maximal y se calcule $X^2/n - p$. Si este valor es cercano a 1 (1.05, 1.10), entonces ajustar por sobredispersión no tendrá demasiado impacto en los tests y podemos tomar $\sigma^2 = 1$. En cambio, si $X^2/n - p$ es considerablementemayor a 1, entonces seguramente convendrá ajustar por sobredispersión, a menos que las observaciones sean no agrupadas ($n_i = 1$).

Ejemplo

McCullagh y Nelder (1989) presentan los resultados de un experimento con tres bloques en que interesa relacionar la proporción de zanahorias dañadas por un insecticida y el logaritmo de la dosis recibida (8 dosis distintas).

	Bloque		
log(dosis)	1	2	3
1.52	10/35	17/38	10/34
1.64	16/42	10/40	10/38
1.76	8/50	8/33	5/36
1.88	6/42	8/39	3/35
2.00	9/35	5/47	2/49
2.12	9/42	17/42	1/40
2.24	1/32	6/35	3/22
2.36	2/28	4/35	2/31

Cuadro 5: Proporción de zanahorias dañadas

Si proponemos un modelo aditivo sencillo de bloque + log(dosis) nos queda:

```
sal.ini<-glm(sf~C(bloque,mat1)+dosis,family=binomial,x=T)
summary(sal.ini)

Call: glm(formula = sf ~ C(bloque, mat1) + dosis, family = binomial, x = T)

Coefficients:
                     Value Std. Error   t value
     (Intercept)  1.4859774  0.6549929  2.268693
C(bloque, mat1)1  0.5341296  0.2315660  2.306598
C(bloque, mat1)2  0.8349701  0.2258107  3.697655
           dosis -1.8160247  0.3431103 -5.292831

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 82.86444 on 23 degrees of freedom

Residual Deviance: 39.8044 on 20 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

$P(X^2_{20} > 39.8044) = 0.005287607$

```
attach(carrot)
sf<- cbind(y,ny)
mat1<- c(1,0,0,0,1,0)
dim(mat1)<- c(3,2)
mat1
     [,1] [,2]
[1,]    1    0
[2,]    0    1
[3,]    0    0
> mat2<- rep(rep(0,7),8)
> dim(mat2)<- c(8,7)
> mat[1,1]<- 1
> mat[2,2]<- 1
> mat[3,3]<- 1
> mat[4,4]<- 1
> mat[5,5]<- 1
> mat[6,6]<- 1
> mat[7,7]<- 1
> mat2[1,1]<- 1
> mat2[2,2]<- 1
> mat2[3,3]<- 1
```

```
> mat2[4,4]<- 1
> mat2[5,5]<- 1
> mat2[6,6]<- 1
> mat2[7,7]<- 1
> mat2
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    0    0    0    0    0    0
[2,]    0    1    0    0    0    0    0
[3,]    0    0    1    0    0    0    0
[4,]    0    0    0    1    0    0    0
[5,]    0    0    0    0    1    0    0
[6,]    0    0    0    0    0    1    0
[7,]    0    0    0    0    0    0    1
[8,]    0    0    0    0    0    0    0
```

```
sal<-glm(sf~ C(bloque,mat1)+C(fdosis,mat2),family=binomial,x=T)
```

```
> summary.glm(sal)

Call: glm(formula = sf ~ C(bloque, mat1) + C(fdosis, mat2),
```

```
family = binomial, x = T)

Coefficients:
                      Value Std. Error    t value
     (Intercept) -2.9028802  0.4060609 -7.1488796
C(bloque, mat1)1  0.5487605  0.2341507  2.3436216
C(bloque, mat1)2  0.8435511  0.2281013  3.6981423
C(fdosis, mat2)1  1.7664710  0.4247983  4.1583756
C(fdosis, mat2)2  1.5579991  0.4227610  3.6852955
C(fdosis, mat2)3  0.8635407  0.4440486  1.9446985
C(fdosis, mat2)4  0.6318727  0.4560345  1.3855810
C(fdosis, mat2)5  0.4318233  0.4582406  0.9423506
C(fdosis, mat2)6  1.1185155  0.4315037  2.5921341
C(fdosis, mat2)7  0.2670066  0.5015928  0.5323173

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 82.86444 on 23 degrees of freedom

Residual Deviance: 27.13288 on 14 degrees of freedom
```

```
Number of Fisher Scoring Iterations: 4
```

$P(X^2_{14} > 27.13288) = 0.0185$

```
resid.pearson<-residuals.glm(sal,type="pearson")
> sum(resid.pearson*resid.pearson)/14
[1] 1.82712

  *** Generalized Linear Model ***

Call: glm(formula = sf ~ C(bloque, mat1) + C(fdosis, mat2),
family = quasi(link = logit, variance = "mu(1-mu)"), data =
carrot, na.action = na.exclude, control = list(
epsilon = 0.0001, maxit = 50, trace = F))



Coefficients:
                      Value Std. Error    t value
     (Intercept) -2.9028802  0.5488766 -5.2887665
C(bloque, mat1)1  0.5487605  0.3165038  1.7338196
C(bloque, mat1)2  0.8435511  0.3083269  2.7358988
```

```
C(fdosis, mat2)1  1.7664710  0.5742042  3.0763810
C(fdosis, mat2)2  1.5579991  0.5714503  2.7263947
C(fdosis, mat2)3  0.8635407  0.6002250  1.4386948
C(fdosis, mat2)4  0.6318727  0.6164264  1.0250578
C(fdosis, mat2)5  0.4318233  0.6194084  0.6971543
C(fdosis, mat2)6  1.1185155  0.5832679  1.9176700
C(fdosis, mat2)7  0.2670066  0.6780082  0.3938103
```

(Dispersion Parameter for Quasi-likelihood family taken to be 1.82712)

```
    Null Deviance: 82.86444 on 23 degrees of freedom

Residual Deviance: 27.13288 on 14 degrees of freedom

Number of Fisher Scoring Iterations: 4
```

```
sal.fin<-glm(sf~C(bloque,mat1)+dosis,family=binomial,x=T)
> summary(sal.fin,dispersion=1.82712)

Call: glm(formula = sf ~ C(bloque, mat1) + dosis, family = binomial, x = T)
Coefficients:
```

```
                     Value Std. Error   t value
     (Intercept)  1.4859774  0.8853604  1.678387
C(bloque, mat1)1  0.5341296  0.3130101  1.706430
C(bloque, mat1)2  0.8349701  0.3052306  2.735538
           dosis -1.8160247  0.4637856 -3.915656

(Dispersion Parameter for Binomial family taken to be 1.82712 )

    Null Deviance: 82.86444 on 23 degrees of freedom

Residual Deviance: 39.8044 on 20 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

$$\frac{39.8044}{1.82712} \quad \rightarrow P(X^2_{20} > 21.7853) = 0.3522$$

Algunas observaciones sobre diagnóstico

Chequeo de la función link

Una manera sencilla de chequear si la función link es adecuada es graficando la *variable de trabajo* z contra el predictor lineal η. Recordemos que

$$\mathbf{z} = \eta + \frac{\partial \eta}{\partial \mu}(Y - \mu)\,.$$

El gráfico debería parecerse a una recta y una curvatura sugeriría que la función link no es la adecuada. Sin embargo, en el caso de datos binarios este plot no es adecuado.

Leverage

En regresión ordinaria, los elementos diagonales h_{ii} de la matriz de proyección

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

son llamados *leverage*. Los puntos con alto leverage son considerados como potencialmente influyentes y dado que $\Sigma\, h_{ii} = p$, se suele considerar como puntos de corte $2p/N$ o $3p/N$.

En GLM, cuando calculamos el estimador de máxima verosimilitud el rol de $\mathbf{X}$ lo cumple $\mathbf{W}^{1/2}\mathbf{X}$, como el estimador de mínimos cuadrados ponderados en un modelo lineal y por lo tanto, obtendremos los leverage a partir de la matriz

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

Observemos que una observación con un $\mathbf{x}$ lejano del centroide de puede no tener alta influencia si su peso es pequeño. El gráfico de residuos versus leverage podría ayudar a detectar algunos datos atípicos en algunas situaciones.

Más sobre Bondad de ajuste

Hemos visto que la distribución de los estadísticos de Pearson X^2 y de la deviance D bajo el supuesto de que el modelo es cierto se aproxima por una distribución χ^2_{m-p}, donde m es la mayor cantidad de parámetros que pueden ser especificados bajo el modelo saturado. El problema es que si $m \approx n$, como la distribución es obtenida cuando n tiende a ∞, tenemos que el número de parámetros crece a la misma velocidad que el número de observaciones y la aproximación en este caso no es buena.

Algunos autores sugieren utilizar la aproximación cuando n_j son suficientemente grandes como para que $n_j\widehat{\Pi}_j \geq 5$ y $n_j(1-\widehat{\Pi}_j) \geq 5$ para la mayoría de las celdas. Por ejemplo, podríamos tener hasta un 20 % de estos valores menores a 5, pero ninguno menor que 1.

McCullagh y Nelder (1989) examinan el valor esperado de la distribución de ambos estadísticos y muestran que la esperanza es menor que $m-p$, tal como debería ser si la distribución fuera χ^2_{m-p}. Estos autores dan un factor de corrección cuando $n_j\widehat{\Pi}_j$ y $n_j(1-\widehat{\Pi}_j)$ exceden 1 para cada j. Sin embargo, hay cierta controversia ya que Hosmer y Lemeshow (1989) dicen que en su experiencia, aunque limitada, este factor de corrección achica demasiado el valor esperado cuando $m \approx n$ y por lo tanto, interpretan que si $m \approx n$ el uso de $m-p$ da un estimador razonable del valor esperado de X^2 y de D, cuando el modelo es correcto.

Test de Hosmer y Lemeshow

Una forma de evitar estas dificultades con la distribución de X^2 y D cuando $m \approx n$, es agrupando los datos de alguna forma. La estrategia que proponen

Hosmer y Lemeshow (1980) y (1982) es agrupar basándose en las probabilidades estimadas.

Supongamos, por simplicidad, que $m = n$. En este caso podemos pensar en que tenemos un vector de n probabilidades estimadas, ordenadas de menor a mayor. Ellos proponen dos estrategias:

- colapsar la tabla basándose en los percentiles de las probabilidades estimadas
- colapsar la tabla basándose en valores fijos de las probabilidades estimadas.

Con el primer método, si, por ejemplo, usamos $g = 10$ grupos, en el primer grupo tendríamos los individuos con las $\frac{n}{10}$ probabilidades estimadas más pequeñas.

Con el segundo método, si $g = 10$, los grupos resultarían de usar como puntos de corte: $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}$.

El test resultante se basará en un estadístico de Pearson aplicado en cada grupo, donde la probabilidad estimada en cada grupo se computa como el promedio de las probabilidades estimadas y el número de datos observados en cada grupo es la suma de los y's correspondientes.

$$
\begin{aligned}
\widehat{C} &= \sum_{k=1}^{g} \frac{(o_k - n'_k \overline{\Pi}_k)^2}{n'_k \overline{\Pi}_k (1 - \overline{\Pi}_k)} \\
n'_k &= \text{número total de sujetos en el grupo } k \\
o_k &= \sum_{j=1}^{c_k} y_j \\
\overline{\Pi}_k &= \sum_{j=1}^{c_k} \frac{m_j \, \widehat{\Pi}_j}{n'_k}
\end{aligned}
$$

donde c_k es el número de puntos de diseño distintos en el k–ésimo grupo y m_j es el número de observaciones con dicho diseño.

Hosmer y Lemeshow (1980) muestran, mediante un estudio de simulación, que si $m = n$ y el modelo logístico estimado es el modelo correcto, $\widehat{C}$ es bien aproximado por una distribución χ^2_{g-2}. También sugieren que la aproximación es válida cuando $m \approx n$.

En un trabajo posterior Hosmer, Lemeshow y Klar (1988) muestran que el método basado en los percentiles de las probabilidades estimadas se ajusta mejor a una χ^2_{g-2}.

Ejemplo

Volvamos al ejemplo de *Bajo Peso en Recién Nacidos*.

En el último ajuste que hicimos obtuvimos la tabla que sigue.

A partir de estas estimaciones se pueden calcular las probabilidades estimadas y los correspondientes percentiles

Si aplicamos el método propuesto usando los percentiles de las probabilidades estimadas obtenemos $\widehat{C}$ =5.23 que al ser comparada con una χ^2_8 tiene un percentil 0.73, lo que indica que el modelo ajusta bien. Si inspeccionamos la tabla comprobamos que hay un solo valor esperado menor a 1 y cinco toman valores inferiores a 5. Si nos preocuparan estos valores se podrían combinar columnas adyacentes para incrementar los valores esperados en las casillas y de esta forma estar más tranquilos con respecto a la aproximación.

Variable	Coef. Estimado	SE	Coef/SE
AGE	-0.084	0.046	-1.84
RACE(1)	1.086	0.519	2.09
RACE(2)	0.760	0.460	1.63
SMOKE	1.153	0.458	2.52
HT	1.359	0.662	2.05
UI	0.728	0.480	1.52
LWD	-1.730	1.868	-0.93
PTD	1.232	0.471	2.61
AGE x LWD	0.147	0.083	1.78
SMOKE x LWD	-1.407	0.819	-1.72
Intercept	-0.512	1.088	-0.47

Regresión de Poisson

La regresión de Poisson es una de las aplicaciones más importantes de GLM.

En este caso estamos interesados en datos de tipo de conteo que no están dados en forma de proporciones. Ejemplos típicos de datos de Poisson o que provienen de un proceso tipo Poisson en los que el límite superior de ocurrencias es infinito ocurren en la práctica. Por ejemplo, número de partículas radioactivas

		Decil de Riesgo										
Peso		1	2	3	4	5	6	7	8	9	10	Total
Bajo	Obs.	0	1	4	2	6	6	6	10	9	15	59
	Esp.	0.9	1.6	2.3	3.7	5.0	5.6	6.8	8.6	10.5	14.1	59
Normal	Obs.	18	19	14	18	14	12	12	9	10	4	130
	Esp.	17.2	18.4	15.8	16.4	15.0	12.4	11.2	10.4	8.5	4.9	130
	Total	18	20	18	20	20	18	18	19	19	19	189

emitidas en un intervalo de tiempo o en estudios de comportamiento número de incidentes en intervalos de longitud especificada.

Aún en los estudios más cuidados puede haber apartamientos al modelo Poisson. Por ejemplo, un contador Geiger tiene un *dead–time* después de la llegada de una partícula, éste es un lapso durante el cual no puede detectar más partículas. Luego cuando la tasa de emisión de partículas es alta, el efecto de *dead–time* lleva a apartamiento notables del modelo de Poisson para el número de ocurrencias registradas. Por ejemplo, si estamos realizando un estudio de la conducta de un chimpancé y contamos el número de ocurrencias de cierto evento es factible que éstas se registren en grupos.

El modelo Poisson asume que

$$E(Y_i) = Var(Y_i) = \mu_i$$

y como ya hemos mencionado es un supuesto que puede ser restrictivo, ya que con frecuencia los datos reales exhiben una variación mayor que la que permite este modelo.

Asumiremos que

$$Y_i \sim P(\mu_i), \quad i = 1, \ldots, n$$

y como siempre queremos relacionar las medias μ_i con un vector de covariables $\mathbf{x}_i$.

Recordemos que si $Y \sim P(\mu)$

$$\begin{aligned} P(Y = y) &= e^{-\mu}\frac{\mu^y}{y!} \\ &= \exp\left(y \log \mu - \mu - \log y!\right) \end{aligned}$$

por lo tanto

$$\begin{aligned} \theta &= \log \mu \\ b(\theta) &= e^{\theta} \\ \phi &= 1 \\ a(\phi) &= 1 \\ c(y, \phi) &= -\log y! \end{aligned}$$

Luego, el link natural es $\eta = \log \mu$, que asegura que el valor predicho de μ_i será no negativo. Cuando se utiliza en el modelo Poisson este link suele llamárselo *modelo loglineal*, sin embargo esta denominación, como veremos, se utiliza en el contexto de tablas de contingencia.

Ajuste del modelo

Cuando se usa el link log Newton–Raphson y Fisher–scoring coinciden. Mediante el algoritmo iterativo calculamos:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$$

donde

$$\mathbf{W} = diag\left(V_i^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$$

y la variable de trabajo

$$\mathbf{z} = \boldsymbol{\eta} + \left(\frac{\partial \eta}{\partial \mu}\right)(\mathbf{y} - \boldsymbol{\mu})$$

Qué queda en el caso en que $\eta = \log \mu$?

Como $\dfrac{\partial \eta}{\partial \mu} = \dfrac{1}{\mu}$, resulta

$$\begin{aligned} \mathbf{W} &= diag(\mu_i) \\ z_i &= \eta_i + \frac{y_i - \mu_i}{\mu_i}\,. \end{aligned}$$

Después de la estimación

Salvo constantes, tenemos que el logaritmo de la función de verosimilitud es

$$\sum_{i=1}^{n} (y_i \log \mu_i - \mu_i)$$

Si usamos el link log, entonces $\log \mu_i = \mathbf{x}_i' \boldsymbol{\beta}$ y la **deviance** queda

$$D = 2 \sum_{i=1}^{n} \left(y_i \log \frac{y_i}{\mu_i} - (y_i - \mu_i) \right)$$

Notemos que si el modelo tiene intercept,

$$\log \mu_i = \beta_1 + \sum_{i=2}^{p} x_{ij} \beta_j, i = 1, \ldots, n$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} (Y_i - \mu_i) \, .$$

Si consideramos los valores predichos con el estimador de máxima verosimilitud, $\widehat{\mu}_i$

$$\sum_{i=1}^{n} \widehat{\mu}_i = \sum_{i=1}^{n} Y_i$$

y por lo tanto la deviance se simplifica a :

$$D = 2 \sum_{i=1}^{n} y_i \log \frac{y_i}{\mu_i} .$$

Podemos definir los residuos deviance como:

$$r_i^d = sg(y_i - \widehat{\mu}_i)\{2(y_i \log(y_i/\widehat{\mu}_i) - y_i + \widehat{\mu}_i)\}^{1/2}$$

y los residuos de Pearson como:

$$r_i^p = \frac{y - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$$

Offset

En el caso de la regresión Poisson es frecuente que aparezca una covariable en el predictor lineal cuyo coeficiente no es estimado pues se asume como 1: esta variable es conocida como **offset**.

Supongamos que tenemos $Y_1, Y_2, \ldots, Y_n$ variables independientes que corresponden al número de eventos observados entre n_i expuestos (*exposure*)para la el $i-$ésimo valor de la covariable. Por ejemplo, Y_i es el número de reclamos de seguro de autos de una determinada marca y año. El valor esperado de Y_i puede escribirse como

$$\mu_i = E(Y_i) = n_i \lambda_i \,,$$

es decir que depende del número de autos segurados y la tasa media de reclamos. Podríamos creer que es λ_i, y no μ_i, quien depende de variables tales como años del auto y lugar donde se usa. Bajo un modelo con link log tenemos que

$$\log \mu_i = \log\, n_i + \mathbf{x}_i' \boldsymbol{\beta} = o_i + \mathbf{x}_i' \boldsymbol{\beta} \,,$$

donde o_i recibe el nombre de offset.

Por ejemplo, si Y_i es el número de muertes por cancer en el año 2001 en una determinada población, parece razonable ajustar por el tamaño de la población.

Función de Varianza

Este modelo asume que

$$E(Y_i) = Var(Y_i) = \mu_i$$

sin embargo es posible que un conjunto de datos tengan una dispersión mayor.

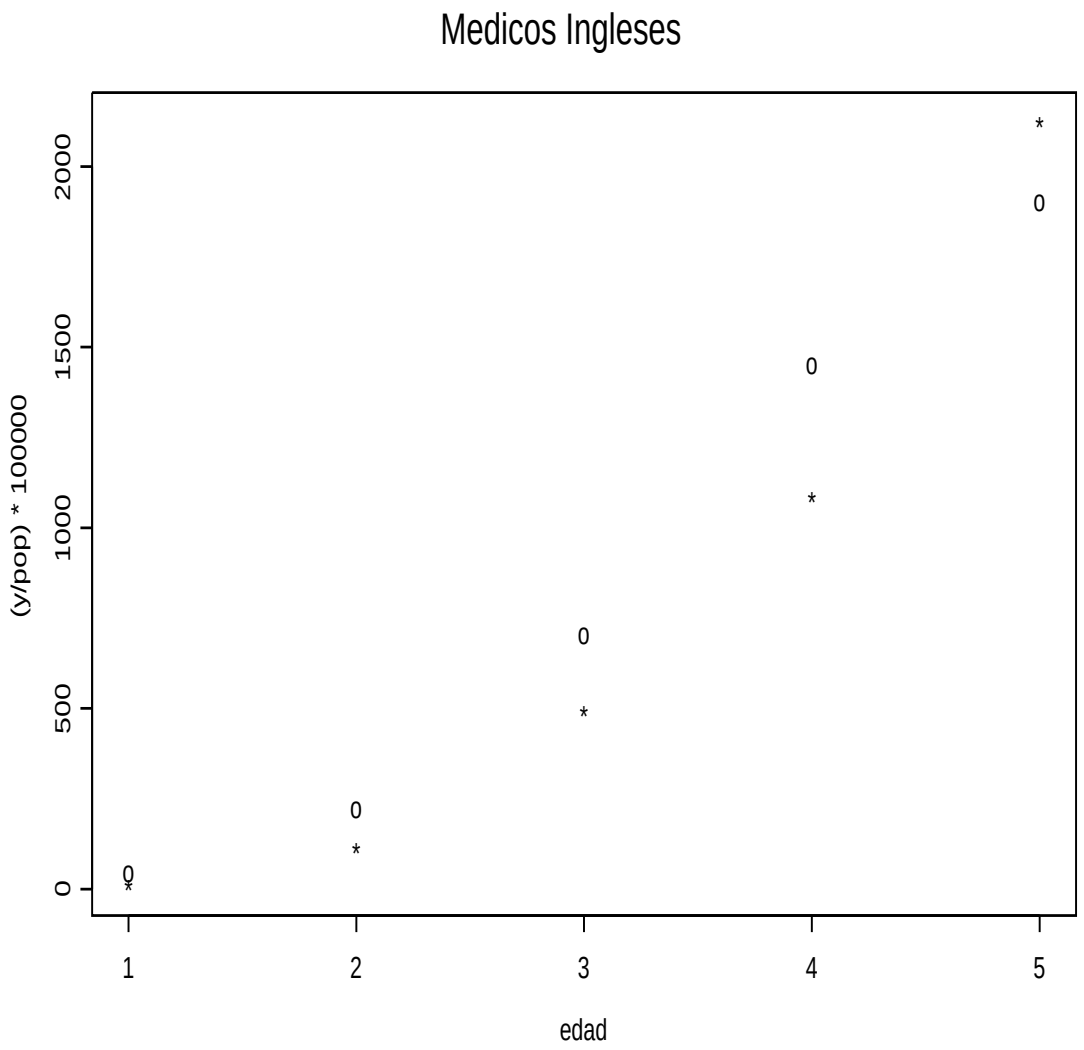
Cuando los datos exhiben sobredispersión, se puede tomar uno de los siguientes caminos:

1. Suponer que $Var(Y_i) = \sigma^2 \mu_i$ y estimar σ^2 usando un modelo de quasi–verosimilitud, como en el caso binonial.
2. Sumergir a la variable de respuesta en una familia de distribuciones que contemple una una dispersión mayor: *Binomial Negativa*

Ejemplo: Médicos Ingleses: fumadores y muerte coronaria (Annette Dobson (1990))

edad	smoke	y	pop
1	1	32	52407
2	1	104	43248
3	1	206	28612
4	1	186	12663
5	1	102	5317
1	0	2	18790
2	0	12	10673
3	0	28	5710
4	0	28	2585
5	0	31	1462

```
plot(edad,(y/pop)*100000,type="n")
text(edad,(y/pop)*100000, c("*","o")[factor(smoke)])
logpop<-log(pop)
edad2<- edad*edad
smkage<- edad*smoke
```


Medicos Ingleses
(y/pop) * 100000
0
500
1000
1500
2000
1
2
3
4
5
edad

```
dmat<- cbind(smoke,edad,edad2,smkage,rep(1,10))

summary.glm(sal)

Call: glm.fit(x = dmat, y = y, offset = logpop, family = poisson)

Coefficients:
             Value Std. Error    t value
 smoke   1.4409718 0.37216161   3.871898
  edad   2.3764783 0.20793739  11.428817
 edad2  -0.1976765 0.02736679  -7.223228
smkage  -0.3075481 0.09703401  -3.169487

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: on 9 degrees of freedom

Residual Deviance: 1.63537 on 5 degrees of freedom
```

Binomial Negativa Recordemos que tal como probamos si

$$\begin{aligned} Y|\lambda &\sim P(\lambda) \\ \lambda &\sim \Gamma(\alpha, \beta) \end{aligned}$$

Término	Edad	$Edad^2$	Smoke	Smkage
Coef.	2.376	-0.198	1.441	-0.308
S.E.	0.208	0.027	0.372	0.097
Rate ratio	10.762	0.820	4.225	0.735
IC 95 %	(7.2 ,16.2)	(0.78,0.87)	(2.04,8.76)	(0.61,0.89)

donde

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\ \beta^{\alpha}}\, \lambda^{\alpha-1} e^{-\lambda/\beta} I_{[0,\infty)}(\lambda)\,,$$

entonces

$$Y\ :\ P(Y=y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\ y!} \left(\frac{\beta}{1+\beta}\right)^{y} \left(\frac{1}{1+\beta}\right)^{\alpha}$$

La media y la varianza de Y son:

$$E(Y) = E(E(Y|\lambda)) = E(\lambda) = \alpha\,\beta$$

$$\begin{aligned} Var(Y) &= E(Var(Y|\lambda)) + Var(E(Y|\lambda)) \\ &= Var(\lambda) + E(\lambda) = \alpha\,\beta + \alpha\,\beta^2 \end{aligned}$$

La distribución **BN** suele parametrizarse en términos de $\mu = \alpha\beta$ y $\kappa = 1/\alpha$ como

$$P(Y = y) = \frac{\Gamma(\kappa^{-1} + y)}{\Gamma(\kappa^{-1})\; y!} \left(\frac{\kappa\mu}{1 + \kappa\mu}\right)^y \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa} .$$

En este caso, diremos que $Y \sim BN(\mu, \kappa)$. Notemos que con esta parametrización

$$\begin{aligned} E(Y) &= \mu \\ Var(Y) &= \mu + \kappa\mu^2 , \end{aligned}$$

por lo tanto, en una BN la varianza es mayor que la media. Esto nos sugiere que si sospechamos que hay subdispersión deberíamos elegir el camino de quasi–verosimilitud, pues la BN no puede tratar este problema.

Cómo ajustamos una distribución BN?

Salvo constantes el log–likelihood resulta

$$\ell = \log \Gamma(\kappa^{-1} + y) - \log y! + y \log \left(\frac{\kappa\mu}{1 + \kappa\mu} \right) + \kappa^{-1} \log \left(\frac{1}{1 + \kappa\mu} \right)$$

Como ya vimos para κ fijo, esta distribución pertenece a una familia exponencial a un parámetro con

$$\theta = \log \left(\frac{\kappa\mu}{1 + \kappa\mu} \right) .$$

Si κ es conocido, se puede computar el estimador de $\boldsymbol{\beta}$ mediante el procedimiento iterativo que hemos visto. Sin embargo, el problema es que en general κ es desconocido y por lo tanto se debe estimar es forma simultánea ambos parámetros.

S–plus no considera la familia BN entre las alternativas de su procedimiento **glm**.

Una posibilidad es maximizar el likelihood aplicando el método de Newton–

Raphson en forma conjunta para κ y $\boldsymbol{\beta}$.

Otra posibilidad es definir una grilla de valores para κ y maximizar el likelihood respecto de $\boldsymbol{\beta}$. Se puede graficar el máximo de la función de verosimilitud para identificar donde se alcanza el estimador de máxima verosimilitud de κ. Se podría comenzar con una grilla más o menos gruesa y luego refinarla en la zona más adecuada.

En el segundo método, para cada κ usaríamos el método de Fisher–scoring como hasta ahora:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$$

donde

$$\mathbf{W} = diag\left(V_i^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right) = diag(w_1, \ldots, w_n)$$

y la variable de trabajo

$$z_i = \eta_i + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i)$$

Eventualmente si tuvieramos un offset quedaría:

$$z_i = \eta_i - o_i + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i)$$

Por la expresión de la varianza que obtuvimos resulta $V_i = \mu_i + \kappa\mu_i^2$ y si usamos el link log, como en el caso Poisson, resulta

$$\begin{aligned} z_i &= \eta_i + \frac{y_i - \mu_i}{\mu_i} \\ w_i &= \frac{\mu_i^2}{\mu_i + \kappa\mu_i^2} \end{aligned}$$

Observemos que la diferencia con la regresión Poisson está en los pesos w_i y no en la variable de trabajo. En este método de la grilla, la matriz de covarianza de $\boldsymbol{\beta}$ se estimaría mediante la fórmula habitual $(\mathbf{X'WX})^{-1}$ usando $\widehat{\kappa}$ en lugar de κ. Vale la pena notar que en este caso no estamos considerando la variabilidad de la estimación de κ