

donde

$$\theta = \int \frac{d\mu}{V(\mu)}$$

y b es alguna función de θ .

De manera, que el supuesto más débil que podemos hacer, al establecer una relación media–varianza, es que la distribución pertenezca a una familia exponencial.

Sea $\mathbf{Y} = (Y_1, \dots, Y_n)'$ un vector de variables aleatorias con media

$$\boldsymbol{\mu} = E(\mathbf{Y}) = (\mu_1, \dots, \mu_n)'$$

y matriz de covarianza

$$\Sigma_{\mathbf{Y}} = \sigma^2 \mathbf{V}(\boldsymbol{\mu}),$$

donde $\mathbf{V}(\boldsymbol{\mu})$ es definida positiva y además sus elementos son funciones conocidas de $\boldsymbol{\mu}$ y σ^2 es una constante de proporcionalidad.

Si las Y_i 's son independientes tendremos que

$$\mathbf{V}(\boldsymbol{\mu}) = \text{diag}(V(\mu_1), \dots, V(\mu_n)).$$

En general, tendremos que $\mu = g(\cdot)$ es una función conocida de p parámetros $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$. Como ha ocurrido hasta ahora, es usual que esta función tenga una componente lineal que involucre una matriz de diseño $\mathbf{X} \in \mathbb{R}^{n \times p}$, de manera que

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

Sean $\mathbf{y} = (y_1, \dots, y_n)'$ el vector de observaciones. Para cada y_k definimos $\ell^*(\mu_k, y_k)$, como

$$\frac{\partial \ell^*(\mu_k, y_k)}{\partial \mu_k} = \frac{y_k - \mu_k}{V(\mu_k)} \quad (7)$$

donde $\text{Var}(Y_k) = \sigma^2 V(\mu_k)$.

Omitamos la constante de proporcionalidad σ .

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

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

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

$$\ell^*(\boldsymbol{\mu}, \mathbf{y}) = \sum_{k=1}^n \ell^*(\mu_k, y_k).$$

Scores basados en ℓ^*

Se pueden definir los scores basados en ℓ^* que serán los *quasi-scores* como

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

De lo anterior obtenemos que

$$\mathbf{U}^*(\boldsymbol{\beta}) = \mathbf{D}' V^{-1}(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu}) / \sigma^2$$

donde $\mathbf{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$ es una matriz de $n \times p$ ($D_{ir} = \frac{\partial \mu_i}{\partial \beta_r}$).

Tenemos que

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

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

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

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

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

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

Aplicando el método iterativo, si $\boldsymbol{\beta}_0$ es un valor inicial el del paso siguiente lo obtenemos:

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

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

$$\widetilde{\boldsymbol{\beta}} \stackrel{(a)}{\approx} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{D}'\mathbf{V}^{-1}(\boldsymbol{\mu})\mathbf{D})^{-1})$$

y que la deviance para el modelo de quasi-verosimilitud

$$D(\mathbf{y}, \tilde{\boldsymbol{\mu}}) = 2 [\ell^*(\mathbf{y}, \mathbf{y}) - \ell^*(\tilde{\boldsymbol{\mu}}, \mathbf{y})] \stackrel{(a)}{\approx} \sigma^2 \chi_{n-p}^2$$

Cuando σ^2 no es conocido propone estimarlo como

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

donde χ^2 es el estadístico generalizado de Pearson.

Volviendo al caso Binomial

En el modelo binomial, sobredispersión significa que

$$V(Y_i) = \sigma^2 \mu_i(1 - \mu_i) / n_i,$$

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 cambiará a $\sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$. 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

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

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 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 procedimiento 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 considerablemente mayor a 1, entonces seguramente convendrá ajustar por sobredispersión, a menos que las observaciones sean no agrupadas ($n_i = 1$). El

punto de corte no es claro, Halekoh y Højsgaard (2007) sugieren preocuparse cuando el valor excede 2.

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). Ver Cuadro 4.

Graficamos los logit empíricos para ver si hay alguna tendencia. El gráfico parece mostrar que hay una asociación negativa, posiblemente lineal. Este gráfico no tiene en cuenta efectos de bloque.

```
empirical.logit_log((y+.5)/(n-y+.5))  
plot(dosis, empirical.logit)
```

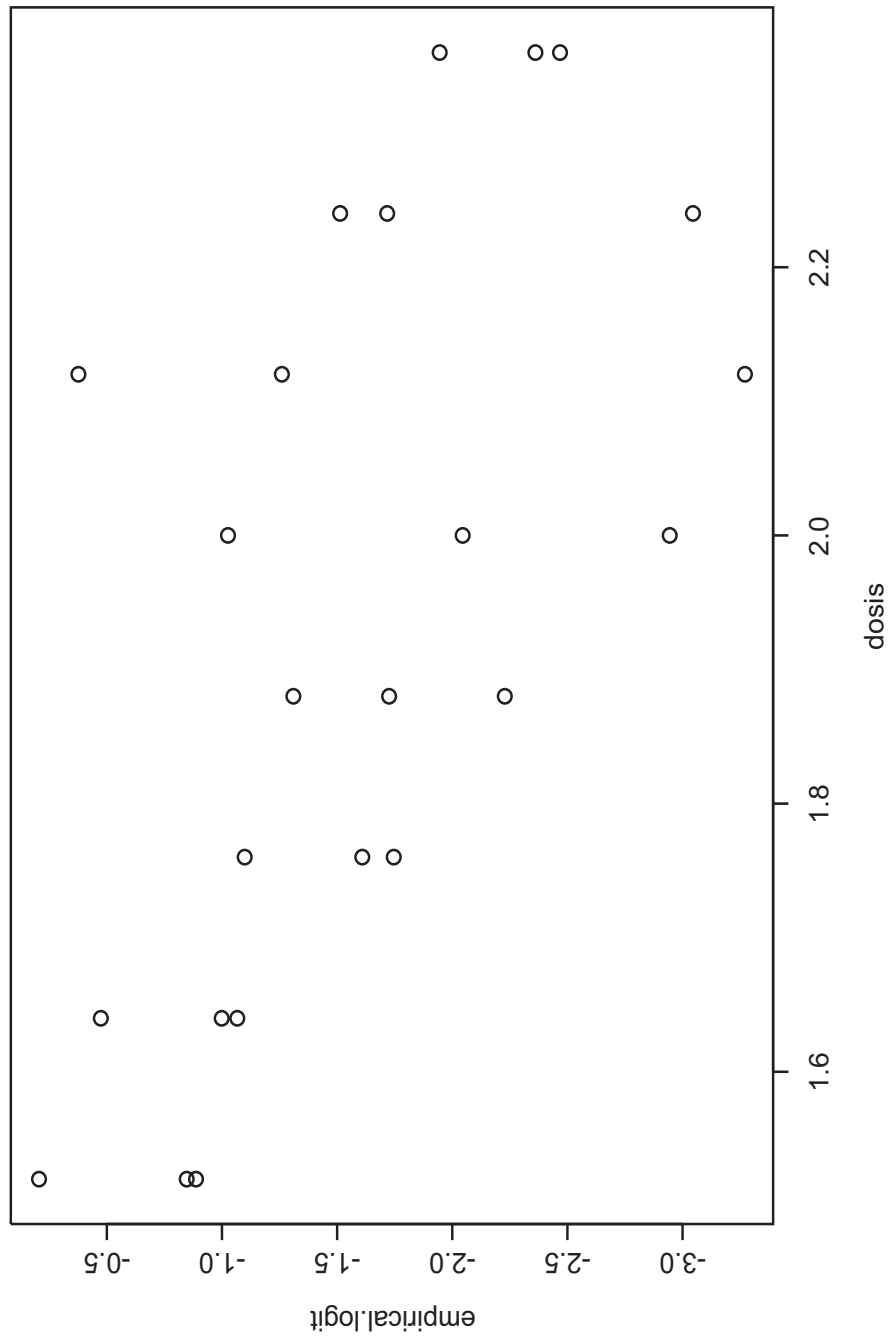


Figura 4: logits empíricos vs. dosis

log(dosis)	Bloque		
	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 4: Proporción de zanahorias dañadas

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

```
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
```

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

```
summary(sal.ini)
```

```
Coefficients:
```

	Value	Std. Error	t value
(Intercept)	1.4859774	0.6549929	2.268693
C(fbloque, mat1)1	0.5341296	0.2315660	2.306598
C(fbloque, 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
```

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

La falta de ajuste podría deberse a distintos factores como que el efectos de la dosis no es lineal, a que hay interacción entre bloque y dosis o a sobredispersión.

Si ajustamos con quasi-verosimilitud el modelo bloque + log(dosis) queda:

```
sal.quasi<-glm(formula = sf~ C(fbloque,mat1)+dosis,family = quasi(link = logit,
variance = "mu(1-mu)"), data = carrots, na.action = na.exclude,
control = list(epsilon = 0.0001, maxit = 50, trace = F))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.919979374228144	-1.021535944114662	-0.3239372066298342	1.060182840478863	3.432377384210175

Coefficients:

	Value	Std. Error	t value
(Intercept)	1.4802190917903660	0.9390873185627804	1.576231584146783
C(fbloque, mat1)1	0.5423815792609435	0.3316511148307153	1.635398028249630
C(fbloque, mat1)2	0.8432621874959400	0.3234003284647346	2.607487108931289
dosis	-1.8173779876929670	0.4919685839291054	-3.694093580485330

(Dispersion Parameter for Quasi-likelihood family taken to be 2.052915304094748)

Null Deviance: 83.34425516168993 on 23 degrees of freedom

Residual Deviance: 39.97574975150014 on 20 degrees of freedom

```
resid.pearson<-residuals.glm(sal.quasi,type="pearson")
```

```
> sum(resid.pearson*resid.pearson)/20
```

```
[1] 2.052915304094748
```

$P(X_{20}^2 > 39.97574975150014 / 2.052915304094748) = 0.491319827$

De todos modos hay residuos grandes, intentemos un ajuste con un modelo maximal: incluimos un intercept, dos variables dummies para bloque y 7 variables para distinguir los niveles de log(dosis)

```
mat2<- rep(rep(0,7),8)
dim(mat2)<- c(8,7)
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(fbloque,mat1)+C(fdosis,mat2),family=binomial,x=T)
summary.glm(sal)
```

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

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

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

Lo ajustamos con quasi-verosimilitud:

```
Call: glm(formula = sf ~ C(fbloque, 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

Ahora ajustamos el modelo inicial usando esta estimación de σ^2 :

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

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

	Value	Std. Error	t value
(Intercept)	1.4859774	0.8853604	1.678387
C(fbloque, mat1)1	0.5341296	0.3130101	1.706430
C(fbloque, 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
```

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

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 $\sum h_{ii} = p$, 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

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_{m-p}^2 , 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 $n_j \hat{\pi}_j \geq 5$ y $n_j(1 - \hat{\pi}_j) \geq 5$ para la mayoría de las celdas. Por ejemplo, podríamos tener hasta un 20 % de estos valores menores a 5, pero ninguno menor que 1.

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: $\frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}$.

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.

$$\widehat{C} = \sum_{k=1}^g \frac{(o_k - n'_k \pi_k)^2}{n'_k \pi_k (1 - \pi_k)}$$

n'_k = número total de sujetos en el grupo k

$$o_k = \sum_{j=1}^{c_k} y_j$$

$$\pi_k = \sum_{j=1}^{c_k} \frac{m_j \pi_j}{n'_k}$$

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_{g-2}^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_{g-2}^2 especialmente si hay muchas probabilidades estimadas pequeñas (inferiores a 0.20).

Ejemplo

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

En el último ajuste que haremos siguiendo la práctica obtendremos la tabla que sigue:

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

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

Aplicando el método propuesto basado en los percentiles de las probabilidades estimadas Hosmer y Lemeshow (1989) reportan un valor de $\widehat{C} = 5.23$ que al ser comparado con una χ^2_8 tiene un percentil 0.73, lo que indica que el modelo

		Decil de Riesgo									
Peso	1	2	3	4	5	6	7	8	9	10	Total
Bajo	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	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

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.

```

options(contrasts=c("contr.treatment", "contr.poly"))

birth<- read.table("birthwt.txt",header=T)
attach(birth)

sal.int<- glm(log~age+frace+smoke+ht+ui+lwdf+ptd+age*lwdf+smoke*lwdf, family=binomial)
summary(sal.int)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.22223 -0.8071988 -0.4431412  0.8972252  2.282125

Coefficients:
            Value Std. Error    t value
(Intercept) -0.5117833  1.08640950  -0.4710777
      age    -0.0839755  0.04547579  -1.8465980
     frace2  1.0830854  0.51849984   2.0888827
     frace3  0.7596625  0.46342813   1.6392239
      smoke  1.1531109  0.45781495   2.5187270
         ht   1.3592050  0.66117785   2.0557328
         ui   0.7281654  0.47920772   1.5195193
      lwdf  -1.7298950  1.86705436  -0.9265370
         ptd  1.2315675  0.47106004   2.6144598
    age:lwdf  0.1474082  0.08278654   1.7805811
    smoke:lwdf -1.4073631  0.81830710  -1.7198471

```

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 234.672 on 188 degrees of freedom

Residual Deviance: 192.0123 on 178 degrees of freedom

```
hosmerlem <- function (y, yhat, g = 10)
{
  cutyhat <- cut(yhat, breaks = quantile(yhat, probs = seq(0,1, 1/g)), include.lowest = T)
  obs <- xtabs(cbind(1 - y, y) ~ cutyhat)
  expect <- xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
  chisq <- sum((obs - expect)^2/expect)
  P <- 1 - pchisq(chisq, g - 2)
  c("X^2" = chisq, Df = g - 2, "P(>Chi)" = P)
}

hosmerlem(sal.int$y,sal.int$fit, g = 10)
      X^2      Df  P(>Chi)
4.7843350 8.0000000 0.7803584
```

```

> obs
cutyhat      V1 V2
[0.0135,0.0736] 19 0
(0.0736,0.0962] 18 1
(0.0962,0.156]  15 4
(0.156,0.209]  17 2
(0.209,0.278]  14 6
(0.278,0.331]  12 6
(0.331,0.425]  12 6
(0.425,0.491]   9 10
(0.491,0.614]  10 9
(0.614,0.937]   4 15

> expect
[0.0135,0.0736] 18.074237 0.925763
(0.0736,0.0962] 17.430816 1.569184
(0.0962,0.156]  16.592368 2.407632
(0.156,0.209]  15.510367 3.489633
(0.209,0.278]  14.982922 5.017078
(0.278,0.331]  12.433818 5.566182
(0.331,0.425]  11.183784 6.816216
(0.425,0.491]  10.430386 8.569614
(0.491,0.614]   8.482910 10.517090
(0.614,0.937]   4.878392 14.121608

```

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 se encuentran en la práctica. Por ejemplo, el número de partículas radioactivas emitidas en un intervalo de tiempo o en estudios de comportamiento el 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 apartamientos notables del modelo de Poisson para el número de ocurrencias registradas. Así también, 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

$$E(Y_i) = \text{Var}(Y_i) = \mu_i$$

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

Asumiremos que

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

y como siempre deseamos relacionar las medias μ_i con covariables \mathbf{x}_i .

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

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

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. Mediate el algoritmo iterativo calculamos:

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

donde

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

y la variable de trabajo

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

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

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

$$\mathbf{W} = \text{diag}(\mu_i)$$

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}.$$

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,

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

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

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

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

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:

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

y los residuos de Pearson como:

$$r_i^p = \frac{Y - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

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, \dots, Y_n variables independientes que corresponden al número de eventos observados entre n_i expuestos (*exposure*) para la 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 asegurados 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.

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)])
```

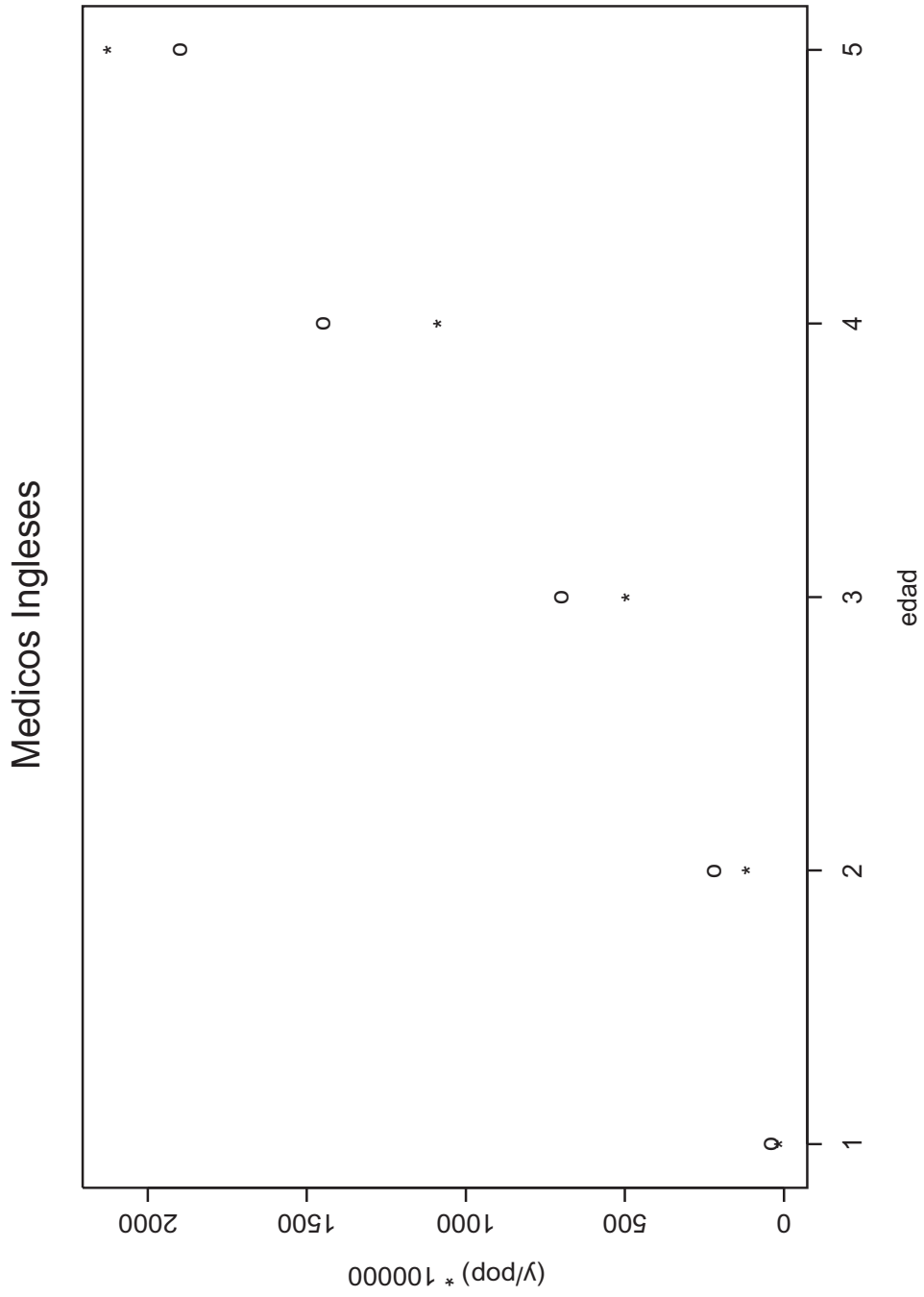


Figura 5: Médicos Ingleses

```
logpop<-log(pop)
edad2<- edad*edad
smkage<- edad*smoke
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
	-10.7917624	0.45003224	-23.979976

```
(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
```


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)

Si x_j es una variable explicativa dicotómica, tal que $x_j = 0$ si el factor está ausente y $x_j = 1$ si el factor está presente, de manera que el **rate ratio** para presencia vs. ausencia es:

$$\text{Rate Ratio} = \frac{E(Y_i|\text{presente})}{E(Y_i|\text{ausente})} = e^{\beta_j}$$

si las otras variables se mantiene fijas.

De la misma forma, si x_j es continua y se incrementa en una unidad manteniendo las otras variables fijas, el Rate Ratio sería e^{β_j} .

Función de Varianza

Este modelo asume que

$$E(Y_i) = \text{Var}(Y_i) = \mu_i$$

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

Binomial Negativa

Si

$$Y|\lambda \sim P(\lambda)$$
$$\lambda \sim \Gamma(\alpha, \beta)$$

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:

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

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

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 $Y \sim \text{BN}(\mu, \kappa)$. Con esta parametrización resulta

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

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 β 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 en 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 β .

Otra posibilidad es definir una grilla de valores para κ y maximizar el likelihood respecto de β . 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 usáramos el método de Fisher–scoring como hasta ahora:

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

donde

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

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

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}$$

$$w_i = \frac{\mu_i^2}{\mu_i + \kappa \mu_i^2}$$

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 β se estimaría mediante la fórmula habitual $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ usando $\hat{\kappa}$ en lugar de κ . Vale la pena notar que en este caso no estamos considerando la variabilidad de la estimación de κ