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).

¿Por qué podría ocurrir sobredispersión?

Supongamos que tenemos una variable dicotómica $Z \sim Bi(1, \pi)$ y que

$$\begin{aligned} Y|_{Z=0} &\sim P(\lambda_0) \\ Y|_{Z=1} &\sim P(\lambda_1) \end{aligned}$$

entonces,

$$\begin{aligned} E(Y) &= \pi\lambda_1 + (1-\pi)\lambda_0 = \mu \\ V(Y) &= E(V(Y|Z)) + V(E(Y|Z)) \\ &= \mu + (\lambda_0 - \lambda_1)^2\pi(1-\pi) \end{aligned}$$

La sobredispersión puede ser tratada de dos formas:

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

En el primer caso, por ejemplo, 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

Con frecuencia es posible caracterizar los dos primeros momentos de una variable de respuesta con distribución desconocida

$$E(Y_i) = \mu_i(\boldsymbol{\beta})$$
$$Var(Y_i) = \sigma^2 V(\mu_i)$$

donde σ es desconocida y V tiene una forma funcional conocida.

La función

$$U_i = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

tiene las siguientes propiedades:

$$E(U) = 0$$
$$Var(U) = \frac{1}{\sigma^2 V(\mu)}$$
$$-E(\frac{\partial U}{\partial \mu}) = \frac{1}{\sigma^2 V(\mu)}$$

luego la integral

$$Q(\mu, y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt \ ,$$

en caso de existir, se comporta aproximadamente como un log–likelihood y se conoce como quasi–likelihood.

Se puede demostrar (Wedderburn, 1974) que la función de log–verosimilitud coincide con la quasi-verosimilitud si y sólo si pertenece a una familia exponencial.

Teorema: Para una observación Y el likelihood ℓ tiene la propiedad

$$\frac{\partial \ell(\mu, y)}{\partial \mu} = \frac{y - \mu}{V(\mu)}$$

donde $\mu = E(Y)$ y $V(\mu) = Var(Y)$ si y solo si la densidad de Y puede ser escrita como

$$exp(y\theta - b(\theta))$$

donde

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

y b es alguna función de θ.

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

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

y matriz de covarianza

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

donde $\mathbf{V}(\mu)$ 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

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

En general, tendremos que $\mu = g(.)$ es una función conocida de p parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Como ha ocurrido hasta ahora, 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_k definimos $\ell^*(\mu_k, y_k)$, como

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

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

Omitamos la contante 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

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

De lo anterior obtenemos que

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

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

Tenemos que

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

McCullagh (1983) mostró bajo condiciones generales que **asintóticamente** $U^*(\boldsymbol{\beta})$ se comporta como una $N_p(0, \mathbf{D}'\, V^{-1}(\boldsymbol{\mu})\, \mathbf{D}/\sigma^2)$.

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

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

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:

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

Si llamamos $\widetilde{\boldsymbol{\beta}}$ al estimador resultante del proceso iterativo, McCullagh (1983) probó que asintóticamente $\widetilde{\boldsymbol{\beta}}$ puede aproximarse por una $N_p(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{D}' V^{-1}(\boldsymbol{\mu}) \mathbf{D}\right)^{-1})$ y que la deviance para el modelo de quasi–verosimilitud $D(\mathbf{y}, \widetilde{\boldsymbol{\mu}}) = 2\ [\ell^*(\mathbf{y}, \mathbf{y}) - \ell^*(\widetilde{\boldsymbol{\mu}}, \mathbf{y})] \overset{(a)}{\sim} \sigma^2 \chi^2_{n-p}$

El estimador convecional de σ^2 es tipo momentos y está dado por

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

donde χ^2 es el estadístico generalizado de Pearson.

Volviendo al caso Binomial

En el modelo binomial, sobredispersión significa que

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

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 $\sigma^2(\mathbf{X'WX})^{-1}$.

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. 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 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 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.

	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 4: Proporción de zanahorias dañadas

Graficamos los logit empíricos para ver si hay alguna tendencia. El gráfico parace

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)
```

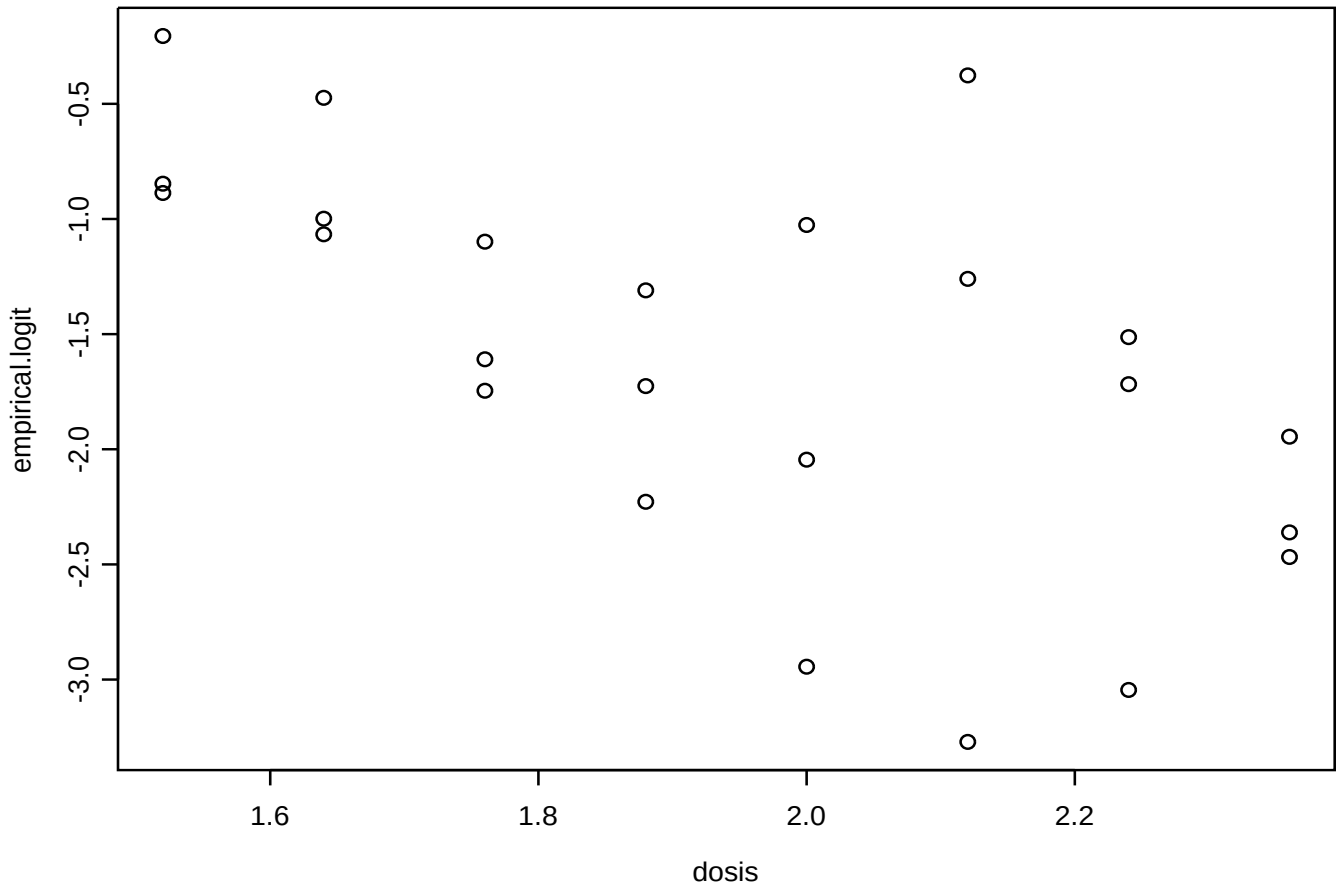

Figura 4: logits empíricos vs. dosis

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
```

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

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
```

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

De todos modos hay residuos grandes, intentemos un ajuste con un modelo maximal: incluimos un intercept, dos variables dummies para distinguir bloque y 7 variables para distinguir los niveles de log(dosis). Este es un modelo sin interacciones en el que no se asume una forma funcional entre dosis (lineal o la que fuera) y respuesta.

```
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
```

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

```
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
```

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