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^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.

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.

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

$$n'_k = \text{número total de sujetos en el grupo } k$$

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

$$\overline{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \, \widehat{\pi}_j}{n'_k}$$

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} especialmente si hay muchas probabilidades estimadas pequeñas (inferiores a 0.20).

Ejemplo

Consideremos ejemplo de *Bajo Peso en Recién Nacidos* del TP4 dados en el archivo birthwt. Los datos corresponden a un estudio de factores de riesgo de bajo peso en recién nacidos. Los datos fueron recogidos en el Baystate Medical Center, Springfield, Massachusetts, en 1986. El siguiente cuadro describe las

variables consideradas.

Variable	**Nombre**
Código de identificación	ID
Bajo peso al nacer (0= peso al nacer $\geq$ 2500gr) (1= peso al nacer $<$ 2500gr)	LOW
Edad de la madre en años	AGE
Peso en libras en la última menstruación	LWT
Raza(1=Blanca, 2=Negra, 3=Otros)	RACE
Fuma durante el embarazo (1= Si, 0=No)	SMOKE
Historia de trabajo prematuro (premature labor) (0=ninguna, 1=uno, 2=dos, etc.)	PTL
Historia de hipertensión (1=si, 0=no)	HT
Presencia de irritación uterina (1=si, 0=no)	UI
Número de visitas al médico durante el primer trimestre (0=ninguna, 1=uno, 2=dos, etc.)	FTV
Peso al nacer en gramos	BWT

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 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

		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

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("C:\\Users\\Ana\\GLM\\web_2010\\birthwt.txt",header=T)
attach(birth)
names(birth)
 [1] "id"    "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"    "ui"    "ftv"   "bwt"
frace<-factor(race)
lwd<- 1*(lwt<110)
lwdf<- factor(lwd)
ptd<-1*(ptl==0)
ptd=1-ptd
ptdf<-factor(ptd)

sal.int<- glm(low~age+frace+smoke+ht+ui+lwdf+ptd+age*lwdf+smoke*lwdf, family=binomial)
summary(sal.int)

Call:
glm(formula = low ~ age + frace + smoke + ht + ui + lwdf + ptd +
    age * lwdf + smoke * lwdf, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2222  -0.8072  -0.4431   0.8972   2.2821
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.51175    1.08754  -0.471  0.63795
age         -0.08398    0.04557  -1.843  0.06533 .
frace2       1.08310    0.51892   2.087  0.03687 *
frace3       0.75968    0.46403   1.637  0.10161
smoke        1.15313    0.45844   2.515  0.01189 *
ht           1.35922    0.66147   2.055  0.03989 *
ui           0.72817    0.47948   1.519  0.12885
lwdf1       -1.72995    1.86831  -0.926  0.35447
ptd          1.23158    0.47139   2.613  0.00898 **
age:lwdf1    0.14741    0.08286   1.779  0.07523 .
smoke:lwdf1 -1.40738    0.81868  -1.719  0.08560 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 192.01  on 178  degrees of freedom

Number of Fisher Scoring iterations: 5
```

```
1-pchisq(192.01 ,178)
[1] 0.2239225

library(ResourceSelection)
hoslem.test(sal.int$y,fitted(sal.int),g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  sal.int$y, fitted(sal.int)
X-squared = 4.7843, df = 8, p-value = 0.7804
```

```
hl<- hoslem.test(sal.int$y,fitted(sal.int),g=10)
cbind(hl$observed,hl$expected)
```

```
                y0 y1     yhat0     yhat1
[0.0135,0.0736] 19  0 18.074237  0.925763
(0.0736,0.0962] 18  1 17.430816  1.569184
(0.0962,0.156]  15  4 16.592368  2.407632
(0.156,0.209]   17  2 15.510367  3.489633
(0.209,0.278]   14  6 14.982922  5.017078
(0.278,0.331]   12  6 12.433818  5.566182
(0.331,0.425]   12  6 11.183784  6.816216
(0.425,0.491]    9 10 10.430386  8.569614
(0.491,0.614]   10  9  8.482910 10.517090
(0.614,0.937]    4 15  4.878392 14.121608
```

Evaluemos qué pasa si cambiamos el número de grupos:

```
 for(i in 10:15){
   print(hoslem.test(sal.int$y,fitted(sal.int),g=i)$p.value)
   }
[1] 0.7803584
[1] 0.9810246
[1] 0.9074916
[1] 0.8973367
[1] 0.9875517
[1] 0.7612329
```

Simulemos un ejemplo

```
pvaluesB=NULL
pvaluesM=NULL

for (i in 1:1000) {
n <- 100
x <- rnorm(n)
xb <- x^2
pr <- exp(xb)/(1+exp(xb))
y <- 1*(runif(n) < pr)

modB <- glm(y~xb, family=binomial)
pvaluesB[i] <- hoslem.test(modB$y, fitted(modB), g=10)$p.value

modM <- glm(y~x, family=binomial)
pvaluesM[i] <- hoslem.test(modM$y, fitted(modM), g=10)$p.value
}

par(mfrow=c(1,2))
hist(pvaluesB)
hist(pvaluesM)

c(mean(pvaluesB>0.20),mean(pvaluesM<0.20))
```