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

El siguiente ejemplo corresponde al TP4 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!!

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

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

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 5: 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!!!!

Algunas herramientas de diagnóstico

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:

- elección incorrecta de la función de enlace
- 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 covariable debería transformarse, por ejemplo, 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, no estaremos estimando consistentemente la matriz de covarianza de $(\widehat{\boldsymbol{\beta}})$.

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 anásis 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 necesariamente 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 $\dfrac{y_i}{n_i}$ 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 $\boldsymbol{\mu} = E(\mathbf{Y}) = (\mu_1, \ldots, \mu_n)'$ y matriz de covarianza $\Sigma_{\mathbf{Y}} = \sigma^2 \mathbf{V}(\mu)$, donde $\mathbf{V}(\mu)$ es definida positiva cuyos elementos son funciones conocidas de $\boldsymbol{\mu}$ y σ^2 es una constante de proporcionalidad. Como asumiremos que Y_i son independientes $\mathbf{V}(\mu)$ es diagonal.

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

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

Omitamos la contante de proporcionalidad σ.

Sean $\mathbf{y} = (y_1, \ldots, y_n)'$ el vector de observaciones. Para cada y_ℓ definimos

$L^*(\mu_\ell, y_\ell)$, 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 suponiendo 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, es suponer una relación media–varianza 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

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

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

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

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**

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