# Análisis de datos Categóricos

## Regresión logística

Ms Carlos López de Castilla Vásquez

Universidad Nacional Agraria La Molina

2014-2

# Introducción

- Para una variable aleatoria respuesta $Y$ y una variable explicativa $X$, sea:

$$\pi(x) = \Pr(Y = 1|X = x) = 1 - \Pr(Y = 0|X = x)$$

- El modelo de regresión logística es:

$$\pi(x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$$

que es equivalente a:

$$\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x$$

# Interpretación de parámetros

- El signo del coeficiente $\beta$ determina si $\pi(x)$ aumenta o disminuye conforme $x$ aumenta.
- Si $\beta = 0$ entonces $Y$ es independiente de $X$.
- El odds se incrementa de forma proporcional a $e^{\beta}$ por cada unidad adicional en $x$.
- El parámetro $\alpha$ no suele ser de mayor interés.
- Si $\pi(x) = 1/2$ entonces $x = -\alpha/\beta$. El valor anterior es llamado LD50 y corresponde a la dosis con un 50 % de posibilidades de tener resultados letales.

# Graficando las proporciones

- Antes de estimar el modelo hay que observar la data para verificar si el modelo de regresión logístico es apropiado.
- Resulta útil graficar las proporciones muestrales $p_i = y_i/n_i$ versus $x$.
- El logit muestral $i$ es:

$$\log \frac{p_i}{1-p_i} = \log \frac{y_i}{n_i - y_i}$$

- Una alternativa es usar:

$$\log \frac{y_i + 1/2}{n_i - y_i + 1/2}$$

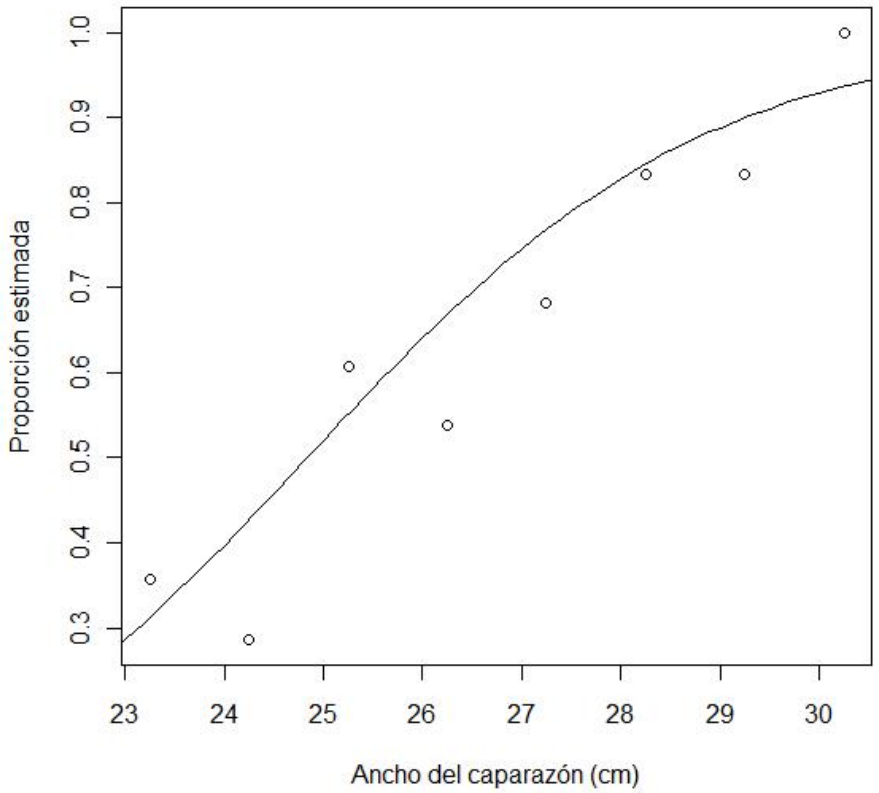
# Ejemplo: Cangrejo de herradura

- Se considera nuevamente la data correspondiente al número de satélites del cangrejo hembra de herradura.
- La variable respuesta binaria es $Y = 1$ si el cangrejo hembra tiene al menos un satélite y $Y = 0$ si no tiene satélites.
- La variable explicativa $X$ es el ancho del caparazón del cangrejo hembra.
- El modelo logístico estimado es:

$$\hat{\pi}(x) = \frac{\exp\{-12{,}351 + 0{,}497x\}}{1 + \exp\{-12{,}351 + 0{,}497x\}}$$

# Ejemplo: Cangrejo de herradura

Figura 1: Proporciones estimadas y observadas



# Inferencia

- Para el modelo con un solo predictor:

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x$$

las pruebas de significancia se enfocan en $H_0 : \beta = 0$, la hipótesis de independencia.

- Se pueden utilizar la prueba de Wald, scores y razón de verosimilitud.
- Para muestras grandes las tres pruebas anteriores dan resultados similares.

# Inferencia

- Los intervalos de confianza suelen ser más eficientes. El intervalo de Wald es:

$$\hat{\beta} \pm z_{1-\alpha/2}\text{EE}(\hat{\beta})$$

- Un intervalo de confianza para $\text{logit}\pi(x_0)$ es:

$$\hat{\alpha} + \hat{\beta}x_0 \pm z_{1-\alpha/2}\text{EE}(\hat{\alpha} + \hat{\beta}x_0)$$

donde EE es la raíz cuadrada de:

$$\text{Var}(\hat{\alpha} + \hat{\beta}x_0) = \text{Var}(\hat{\alpha}) + \text{Var}(\hat{\beta}x_0) + 2x_0\text{Cov}(\hat{\alpha}, \hat{\beta})$$

# Ejemplo: Cangrejo de herradura

## Modelo 1

```
> Cangrejo <- read.table(file="G://Cangrejo.txt", header=T)
> attach(Cangrejo)
> Sat <- ifelse(Sat>0, 1, 0)
> modelo1 <- glm(Sat ~ Ancho, family=binomial(link=logit))
```

## Matriz de varianza-covarianza

```
> vcov(modelo1)
```

# Regresión logística múltiple

- El modelo general de regresión logística es:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}$$

- La devianza es:

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]$$

y tiene la forma:

$$D = 2\sum_i o_i \log\frac{o_i}{e_i}$$

# Regresión logística múltiple

- Las frecuencias observadas $o_i$ son $y_i$ y $n_i - y_i$, mientras que las frecuencias esperadas $e_i$ son:

$$\hat{y}_i = n\hat{\pi}_i \quad \text{y} \quad n_i - \hat{y}_i$$

obtenidas usando el modelo estimado.

- La prueba de bondad de ajuste del modelo puede ser llevada a cabo usando la aproximación:

$$D \sim \chi^2_{N-p}$$

donde $N$ es el número de datos y $p$ el número de parámetros a estimar en el modelo.

# Ejemplo: Diabetes

- Se tiene información proveniente de un estudio con 768 pacientes mujeres del Instituto Nacional de enfermedades Digestivas, Diabetes y de Riñón.
- Las variables independientes involucradas son: número de embarazos, concentración de glucosa en plasma en una prueba de tolerancia oral (mmol/L), presión arterial diastólica (mmHg), grosor del pliegue del tríceps (mm), suero de insulina en dos horas (muU/ml), índice de masa corporal, función pedigrí de diabetes, edad (años).
- La variable respuesta diabetes cuyo valor 1 es interpretado como *prueba de diabetes positiva*.

# Ejemplo: Diabetes

## Modelo 1

*> modelo1 <- glm(Diabetes ~ ., family=binomial(link=logit), data=Diabetes)*

## Modelo 2

*> modelo2 <- glm(Diabetes ~ Embarazos + Plasma + Presion + Indice + Pedigri, family=binomial(link=logit), data=Diabetes)*

## Modelo 1 versus Modelo 2

*> anova(modelo2, modelo1, test="Chisq")*

# Estadísticas de bondad de ajuste

- El estadístico chi-cuadrado de Pearson es:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \sim \chi^2_{N-p}$$

- Si $y$ es 0 o 1 entonces $X^2$ y $D$ no proporcionan una medida de bondad de ajuste que sea apropiada.
- Lo anterior también podria ocurrir si las variables explicativas son continuas.
- En cualquiera de estas situaciones es preferible usar el estadístico de *Hosmer y Lemeshow* (1980).

# Estadísticas de bondad de ajuste

- La idea es agrupar las observaciones en categorías de acuerdo a las probabilidades estimadas usando $g$ grupos cada uno con aproximadamente la misma cantidad de observaciones.
- Con 10 grupos, el primer grupo de conteos observados y sus correspondientes conteos estimados esta formado con las $n/10$ observaciones con las probabilidades más altas y así sucesivamente.
- El valor estimado es la suma de las probabilidades estimadas en cada grupo.
- Sea $y_{ij}$ la observación $j$ en el grupo definido por la partición $i$, $i = 1, 2, \cdots, g$ y $j = 1, 2, \cdots, n_i$.

# Estadísticas de bondad de ajuste

- Sea $\hat{\pi}_{ij}$ las probabilidades estimadas con la data no agrupada.
- El estadístico de *Hosmer y Lemeshow* es:

$$X^2_{HL} = \sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}\right)^2}{\left(\sum_j \hat{\pi}_{ij}\right)\left(1 - \left(\sum_j \hat{\pi}_{ij}\right)/n_i\right)}$$

  cuya distribución es aproximadamente chi-cuadrado con $g - 2$ grados de libertad.
- Si el valor es grande puede ser evidencia de una falta de ajuste en el modelo.

# Estadísticas de bondad de ajuste

- Es posible comparar el logaritmo de la verosimilitud del modelo estimado y el *modelo minimal* que es aquel donde todas las probabilidades son iguales.
- Sea $\hat{\pi}_i$ las probabilidades estimadas para $y_i$ bajo el modelo de interés.
- La estadística chi-cuadrado de razón de verosimilitud es:

$$C = 2\left(l(\hat{\pi}, \mathbf{y}) - l(\tilde{\pi}, \mathbf{y})\right) \sim \chi^2_p$$

- Otra estadística usada es:

$$\text{pseudo}\, R^2 = \frac{l(\tilde{\pi}, \mathbf{y}) - l(\hat{\pi}, \mathbf{y})}{l(\tilde{\pi}, \mathbf{y})}$$

# Residuales de Pearson

- El *residual de Pearson* es:

$$p_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}} \qquad k = 1, 2, \cdots, m$$

  tal que $X^2 = \sum p_k^2$.

- El *residual estandarizado de Pearson* es:

$$r_{p_k} = \frac{p_k}{\sqrt{(1-h_k)}}$$

  donde $h_k$ es el *leverage* obtenido de la matriz *hat*.

# Residuales de Devianza

- Los *residuales de Devianza* son:

$$d_k = \text{signo}(y_k - n_k\hat{\pi}_k)\left\{2\left[y_k \log\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k}\right)\right]\right\}$$

  donde $D = \sum d_k^2$.

- Los *residuales estandarizados de Devianza* son:

$$r_{d_k} = \frac{d_k}{\sqrt{(1 - h_k)}}$$

# Residuales

## Residual de Pearson

```
> res.pearson <- resid(modelo2, type="pearson")
```

## Residual estandarizado de Pearson

```
> res.est.p <- res.pearson/sqrt(1-lm.influence(modelo2)$hat)
```

## Residual de Devianza

```
> res.devianza <- resid(modelo2, type="deviance")
```

## Residual estandarizado de Devianza

```
> res.est.d <- res.devianza/sqrt(1-lm.influence(modelo2)$hat)
```