# Regresión logística

Parte 2

# Interpretación del modelo con más de un predictor

- Cuando se tiene más de un predictor, los odds ratio que se obtienen son indicadores relativos
- Permiten saber el efecto de un factor controlando el efecto de los restantes factores incluidos como predictores

# Ejemplo: Regresión logística con SMOKE y AGE

```
Call:
glm(formula = LOW ~ SMOKE + AGE, family = "binomial", data = LOWBWT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1589  -0.8668  -0.7470   1.2821   1.7925

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06091    0.75732   0.080   0.9359
SMOKE        0.69185    0.32181   2.150   0.0316 *
AGE         -0.04978    0.03197  -1.557   0.1195
---
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: 227.28  on 186  degrees of freedom
AIC: 233.28

Number of Fisher Scoring iterations: 4
```

```
> or_glm(data = LOWBWT, model = fumar.edad, incr =
list(SMOKE=1,AGE=10))
  predictor oddsratio ci_low (2.5) ci_high (97.5) increment
1     SMOKE     1.997        1.064           3.77         1
2       AGE     0.608        0.318           1.12        10
```

# Modelo de regresión logística

- Dada una variable dicotómica Y, y un predictor X, llamando

$$\pi(x) = P(Y = 1\ dado\ el\ valor\ de\ x)$$

- Se propone

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

O lo que es equivalente

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

# Estimación de parámetros

Método de máxima verosimilitud

- Si se propone

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

- Función de verosimilitud

$$\ell(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}$$

- Se maximiza el logaritmo de la función de verosimilitud

$$LL(\beta) = ln\left(\ell(\beta)\right) = \sum \{y_i ln[\pi(x_i)] + (1-y_i) ln[1-\pi(x_i)]\}$$

- Deviance. Se denomina al logaritmo de la función de verosimilitud multiplicada por menos 2

$$D = -2\mathrm{LL}$$

- La deviance tiene importantes propiedades probabilísticas y permite definir test de bondad de ajuste para el modelo

- Test de cociente de máxima verosimilitud

# Test de cociente de máxima verosimilitud

Permite comparar modelos anidados

Sea $D_1 = -2LL(modelo1)$ la deviance del modelo 1
Sea $D_2 = -2LL(modelo2)$ la deviance del modelo 2, que tiene los mismos predictores que el modelo 1 más una agregado

- Entonces

$$G = D_1 - D_2 = -2\ln\left(\frac{función\ de\ verosimilitud\ del\ modelo\ 1}{función\ de\ verosimilitud\ del\ modelo\ 2}\right)$$

Tiene distribución aproximada chi cuadrado con 1 grado de libertad

# Regresión logística con SMOKE

```
## Call:
## glm(formula = LOW ~ SMOKE, family = "binomial", data = LOWBWT)
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0197  -0.7623  -0.7623   1.3438   1.6599
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0871     0.2147  -5.062 4.14e-07 ***
## SMOKE         0.7041     0.3196   2.203   0.0276 *
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 229.80  on 187  degrees of freedom
## AIC: 233.8
##
##   predictor oddsratio ci_low (2.5) ci_high (97.5) increment
## 1     SMOKE     2.022        1.082          3.801         1
```

# Regresión logística con SMOKE y AGE

```
Call:
glm(formula = LOW ~ SMOKE + AGE, family = "binomial", data = LOWBWT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1589  -0.8668  -0.7470   1.2821   1.7925

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06091    0.75732   0.080   0.9359
SMOKE        0.69185    0.32181   2.150   0.0316 *
AGE         -0.04978    0.03197  -1.557   0.1195
---
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: 227.28  on 186  degrees of freedom
AIC: 233.28

Number of Fisher Scoring iterations: 4
```

# Test de máxima verosimilitud

```
> anova(fumar,fumar.edad)
Analysis of Deviance Table

Model 1: LOW ~ SMOKE
Model 2: LOW ~ SMOKE + AGE
  Resid. Df Resid. Dev Df Deviance
1       187     229.81
2       186     227.28  1   2.5283
```

# Selección automática de predictores

- Como en los modelos de regresión lineal, es posible realizar una selección automática de predictores
- En este caso, los criterios estarán basados en la deviance y en el AIC y sus variantes.

```
birthwt.glm <- glm(LOW ~ 1, family = binomial, data = LOWBWT)
birthwt.step <- stepAIC(birthwt.glm,
scope = list(upper = ~AGE+LWT+factor(RACE)+SMOKE+PTL+HT+UI+FTV, lower = ~ 1),
                    direction = c("both"),trace = T)
```

# Clasificación con regresión logística

# Regresión logística

- Modelo propuesto

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- Equivalente a

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

Si tengo estimadores de los parametros

- Puedo calcular una probabilidad de éxito a partir del modelo para cada individuo
- Puedo predecir el valor de *y* a partir de una probabilidad de corte
- Puedo comparar con la dependiente original

- Se puede definir como “1” a aquellos individuos con probabilidad de éxito mayor que 0.5
- O bien, se puede calcular la utilizar probabilidad en termino de los datos del problema

- Ambas opciones pueden utilizarse en R

# Matriz de confusión

**Predicción**

| **Variable original** | **Éxito** | **Fracaso** | |
|---|---|---|---|
| **Éxito** | A | B | A+B |
| **Fracaso** | C | D | C+D |
| | A+C | B+D | TOTAL |

$$Precisión = \frac{\#\ de\ individuos\ bien\ clasificados}{\#\ total\ de\ individuos} = \frac{A+D}{TOTAL}$$

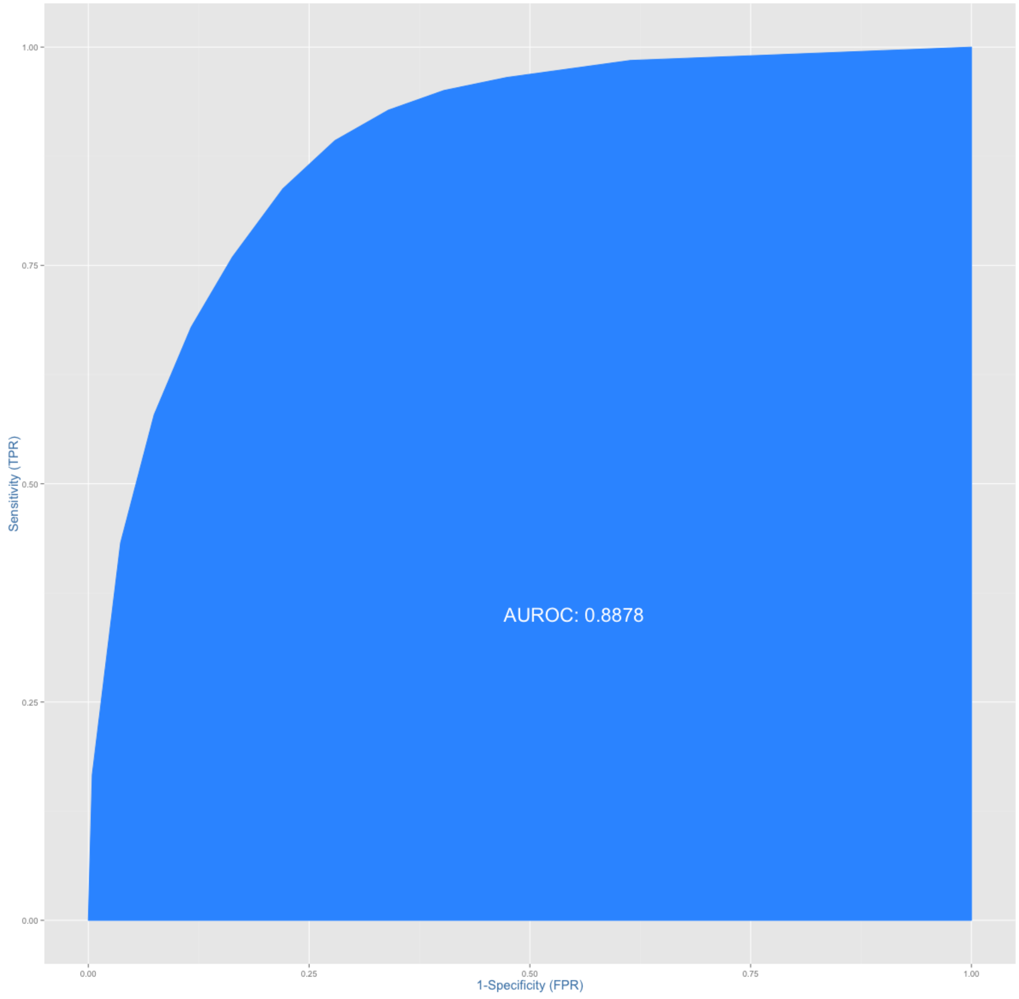
Error de clasificación: $\frac{C+B}{TOTAL}$

$$Sensitividad = \frac{\#\ de\ individuos\ con\ Y=1\ bien\ clasificados}{\#\ de\ individuos\ con\ Y=1} = \frac{A}{A+B}$$

$$Especificidad = \frac{\#\ de\ individuos\ con\ Y=0\ bien\ clasificados}{\#\ de\ individuos\ con\ Y=0} = \frac{D}{C+D}$$

# Clasificación

- Curvas ROC
- Mientras mayor es el área, mejor es el caracter predictor del modelo



# Pseudo r cuadrados

- *McFadden*

$$R^2_{Mcf} = 1 - \frac{ln(Lm)}{ln(L0)}$$

- Cox Snell

$$R^2_{CS} = 1 - \left(\frac{L0}{Lm}\right)^{2/n}$$

- Nagelkerke

$$R^2_N = \frac{R^2_{CS}}{1 - L0^{2/n}}$$