

Modelos para el aprendizaje automático



Resumen

- Introducción a factores de riesgo: RR y OR
- Regresión logística simple
- Interpretación de los coeficientes
- Estimación MV

Introducción a factores de riesgo

¿Cómo identificamos factores de riesgo?

E: el bebé presenta bajo peso

F: madre fuma

		Bajo peso		
		Si	No	
Tabaquismo madre	Si	136	225	361
	no	177	2216	2393

Una opción: **evaluar riesgos**

$$\text{PREVALENCIA de Bajo Peso} = \frac{136 + 177}{136 + 177 + 225 + 2216} = 0.1136$$

$$RR = \frac{136 / 361}{177 / 2393} = 5.09 = \text{Riesgo Relativo}$$

riesgo de BP si Fuma
riesgo de BP si NO FUMA

Si RR = 1
No hay
riesgo!

¿Cómo identificamos factores de riesgo?

		Efecto presente?	
		1 = si	0 = no
Factor presente?	1 = si	a	b
	0 = no	c	d

Si la tabla corresponde frecuencias observadas, usamos estas para estimar el riesgo de padecer la enfermedad si está presente el factor de riesgo versus si no lo está. Esto corresponde al **Riesgo Relativo**.

También puede verse la chance de enfermarse versus no enfermarse entre los que presentan el factor de riesgo versus los que no. Esto es el **Odds Ratio**

¿Cómo identificamos factores de riesgo?

E: el bebé presenta bajo peso

F: madre fuma

		Bajo peso	
		Si	No
Tabaquismo madre	Si	136	225
	no	177	2216

Otra opción: evaluar chances

$$\text{Odd}_{Fuma} = \frac{\overset{\text{\% de BP en Fuma}}{136 / 361}}{\underset{\text{\% de NOBP en Fuma}}{225 / 361}} = \frac{136}{225} = 0.60444$$

$$\text{OR} = \frac{\text{Chance de BP vs noBP si fuma}}{\text{Chance de BP vs noBP si no fuma}} = \frac{136 / 225}{177 / 2216} = 7.56 = \text{Odds Ratio}$$

Definiciones

Riesgo relativo (RR)

Compara la incidencia entre los que tienen factor de riesgo y quienes no lo tienen.

Razón de chance, oportunidad relativa,... Odds ratio (OR)

Compara chances de los que tienen factores vs los que no

En población, los def son:

$$RR = \frac{P(E / F)}{P(E / noF)}$$

$$OR = \frac{\frac{P(E / F)}{P(noE / F)}}{\frac{P(E / noF)}{P(noE / noF)}} = \frac{\frac{P(E / F)}{1 - P(E / F)}}{\frac{P(E / noF)}{1 - P(E / noF)}}$$

Handwritten notes: "Odds(F)" next to the first fraction, and "Odds(noF)" next to the second fraction.

Odds y Odds Ratios

- RR no se aplica en estudios de caso-control. OR sí.
- $OR = RR * [P(\text{noE/noF}) / P(\text{noE/F})]$
- Los Odds toman valores de 0 a +infinito
- $\log(\text{Odd})$ puede tomar cualquier valor real y puede modelarse tal como hacíamos en regresión, para una variable continua.

¿Cómo interpretarlos?

- Odds ratio: $OR = 1$ significa que el factor F no representa riesgo.
- $OR > 1$ significa que F es un 'factor dañino' (aunque depende del contexto).
- $OR < 1$ significa que F es un 'factor de protección'

Cálculos en R

Limitaciones

- Este enfoque permite sólo 2 categorías.
- No permite trabajar con varios factores (variables indep).
- No permite variables independientes continuas.

La respuesta a esto es la modelización de ODDS mediante una función de la(s) variable(s) independiente(s).

Regresión logística

Bibliografía:

- Chatterjee, S.; Hadi, A.; Price, B. “Regression Analysis by Example”. Wiley (Introductoria)
- Montgomery, Peck y Vining. “Introducción al Análisis de Regresión Lineal” (Introductoria)
- Hosmer, D.; Lemeshow, S. (2000). “Applied logistic regression” (Wiley Series in Probability)

Ejemplo (H-L): Porqué no modelar la relación para una variable respuesta dicotómica con regresión?

$$CHD = Y = \begin{cases} 1 & \text{si presento incidente cardiaco} \\ 0 & \text{si no} \end{cases}$$

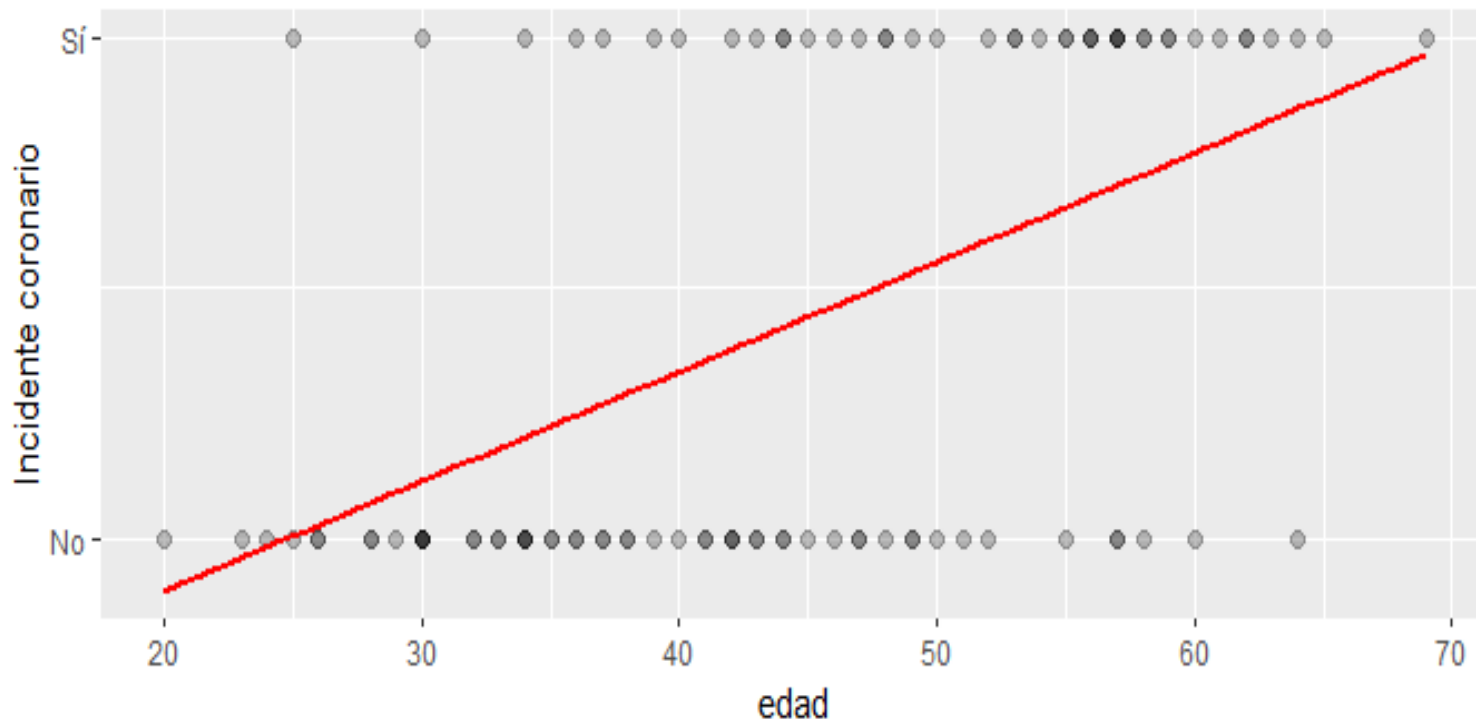
$$X = AGE$$

**Table 1.1 Age and Coronary Heart Disease (CHD)
Status of 100 Subjects**

ID	AGE	AGRP	CHD	ID	AGE	AGRP	CHD
1	20	1	0	51	44	4	1
2	23	1	0	52	44	4	1
3	24	1	0	53	45	5	0
4	25	1	0	54	45	5	1
5	25	1	1	55	46	5	0

Ejemplo (H-L, tabla1.1):

¿Porqué no modelar la relación para una variable respuesta dicotómica con regresión?



Ejemplo (H-L)

Definamos la variable Y como indicadora de presencia de Efecto (en el ejemplo sería: incidente coronario).

Esto es, $Y=1$ indica Efecto

Entonces, la probabilidad de presentarse el efecto para determinado valor de x (edad) es

$$\pi(x) = P(Y = 1|x)$$

Se propone entonces modelar esta probabilidad.

Ejemplo (H-L):

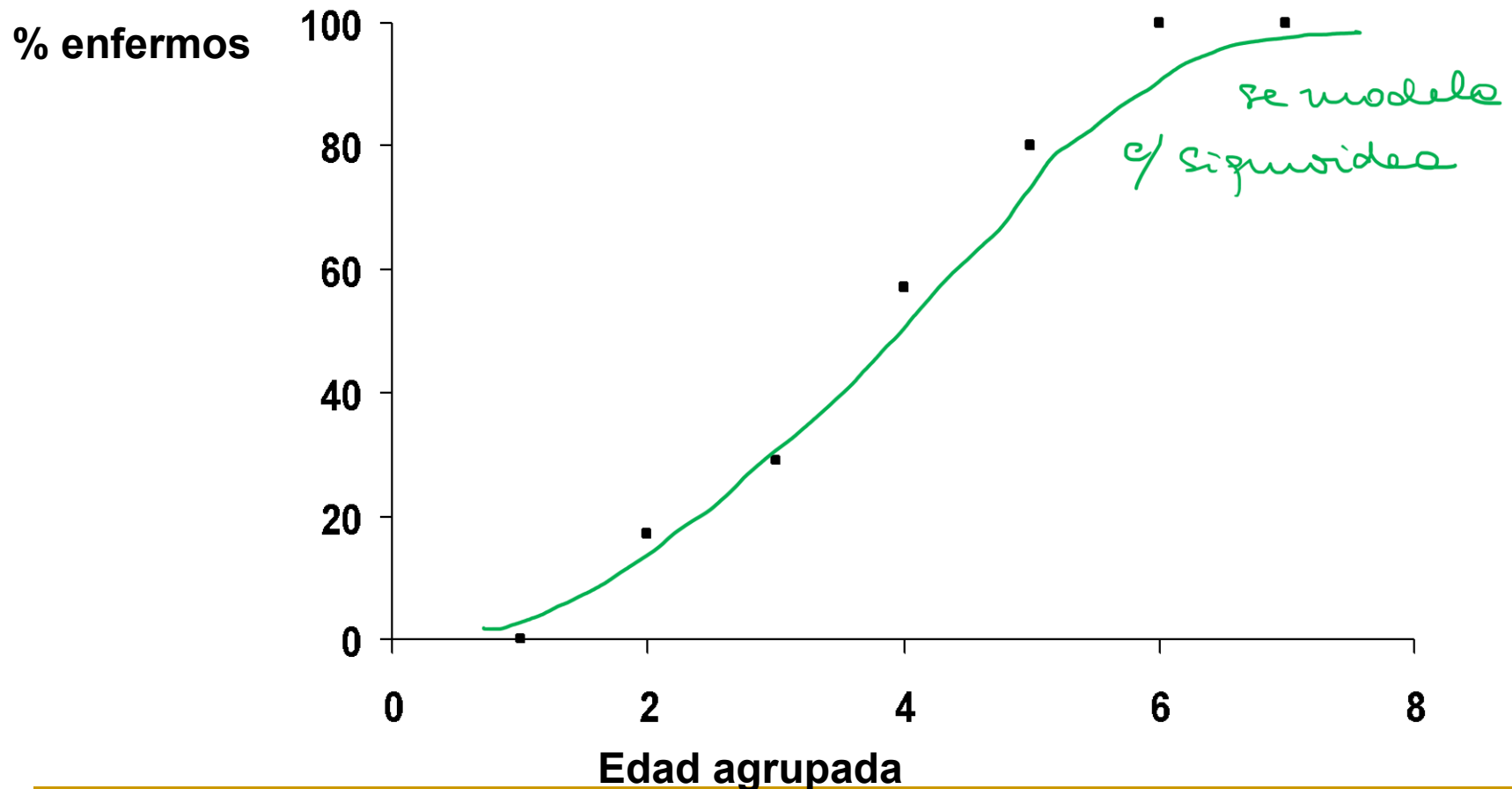
Porcentaje (%) de enfermos en cada grupo etario

		Diseased ($\gamma=i$)		% enfermos del total del nivel de edad
Age group	# in group	#	%	
20 - 29	5	0	0 %	←
30 - 39	6	1	17 %	
40 - 49	7	2	29	
50 - 59	7	4	57	← $4/7 = 0.57$
60 - 69	5	4	80	↑ los valores se dibujan
70 - 79	2	2	100	
80 - 89	1	1	100	

Niveles →
del factor edad

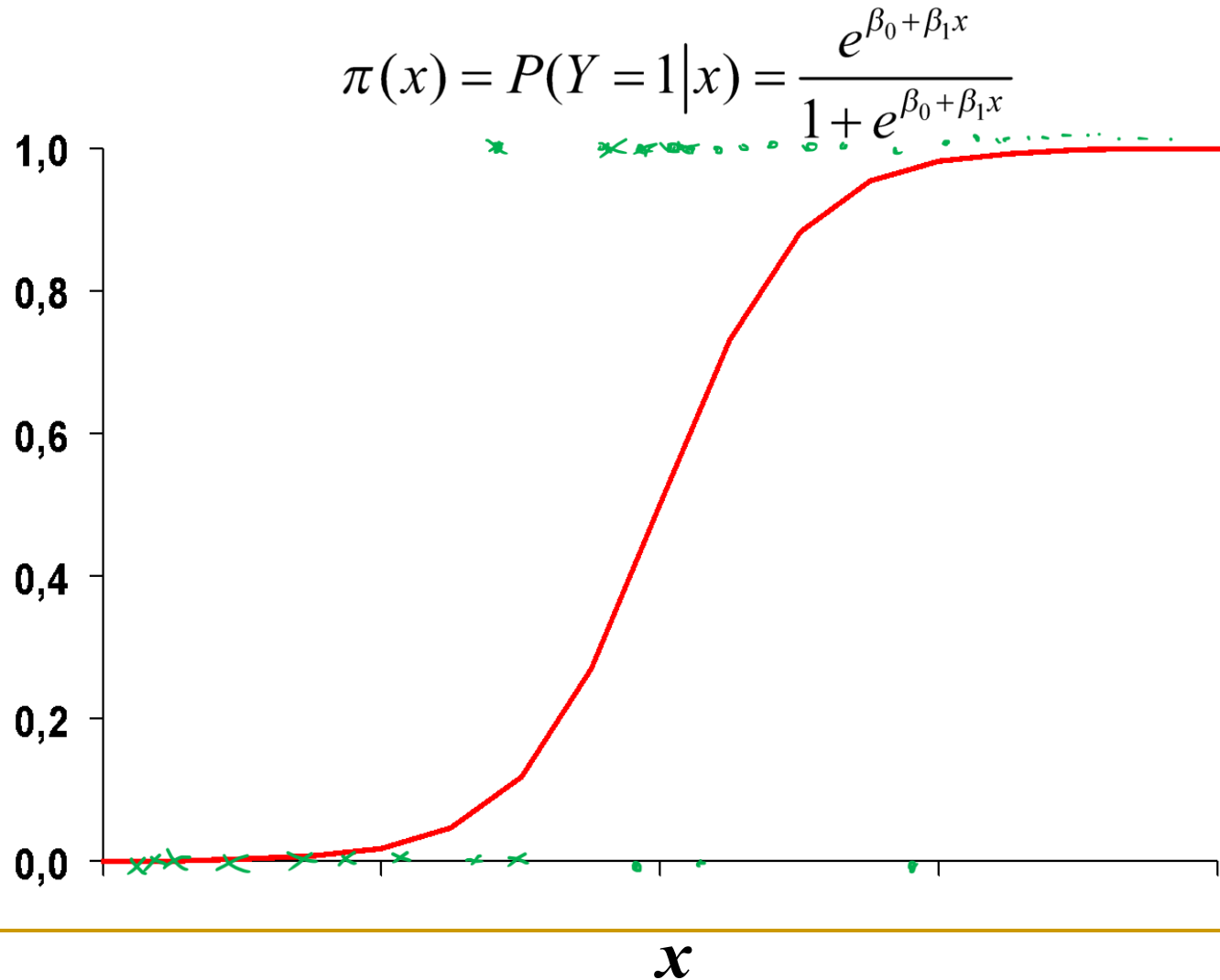
Ejemplo (H-L):

El gráfico sugiere que estas proporciones pueden modelarse con una función particular.



Un modelo posible: Función Logística

Probabilidad
de $Y=1$



Características de la función logística

- Es monótona (creciente ó decreciente), dependiendo del signo de β_1
- Es *casi* lineal en el rango donde $\pi(x)$ está entre 0.2 y 0.8 y gradualmente se acerca a 0 ó 1 en los límites del rango de x

Esto es: $\pi'(x) \approx (1-\pi(x)) \pi(x) \beta_1$

- $\pi(x) = 0.5$ cuando $x = -\beta_0 / \beta_1$ (para 1 regresor)
- El “odd” en el nivel x es función de $\pi(x)$

Modelo de Regresión Logística simple

Se propone modelar la probabilidades condicionales mediante una función logística.

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

No modelo “Y” sino **la probabilidad de Y=1 según sea el valor de “x”**

Transformación logit

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \Rightarrow g(\mathbf{x}) = \ln \left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right] = \beta_0 + \beta_1 x$$

Handwritten notes:
- A green arrow points from the handwritten \ln to the fraction inside the log, with the text $\equiv \log$.
- The entire log expression is circled in red.
- Below the right-hand side, it says "función lineal de x !!!" in green.

Logit = log(odd(x))

Esto es

$$\pi(\mathbf{x}) = P(Y = 1 / x)$$

$$e^{\beta_0 + \beta_1 x} = e^{g(\mathbf{x})} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = ODD(x) = \frac{P(Y = 1 / x)}{P(Y = 0 / x)}$$

Lo que corresponde a probabilidad de respuesta positiva / prob de respuesta negativa en un cierto nivel x del factor.

Observaciones

- El ajuste se hace modelando los “logits” = $g(x)$
- Y/x es una variable aleatoria Bernoulli con $p = \pi(x)$

- En Regresión lineal : $Y = \underbrace{E(Y/x)}_{\beta_0 + \beta_1 x_1 + \dots} + \varepsilon$, con $\varepsilon \sim N(0; \sigma^2)$. ¿Supuestos?

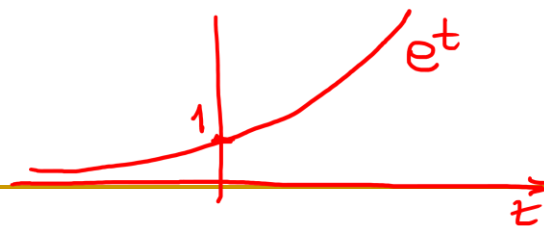
- En Regresión Logística: los errores son dicotómicos con probabilidades $\pi(x)$ y $(1-\pi(x))$. ¿Supuestos?
 - Y sea dicotómica
 - indep de las observaciones
 - (multicolinealidad)

Interpretación de los Coeficientes

$$e^{\beta_0 + \beta_1(x+1)} = e^{\beta_0 + \beta_1 x} \cdot e^{\beta_1} = e^{\beta_1} \cdot \left(odds(x) = \frac{\pi(x)}{1 - \pi(x)} \right) e^{\beta_0 + \beta_1 x}$$

$$\frac{odds(x+1)}{odds(x)} = e^{\beta_1} = OR(x)$$

- e^{β} representa el cambio en los odds al incrementar x en 1 unidad.
- Si $\beta_1 = 0$, x no tiene efecto sobre Y ($e^{\beta}=1$), o sea la chance es la misma en cualquier nivel de x .
- If $\beta_1 > 0$, la chance de ocurrir $Y=1$ crece con x ($e^{\beta} > 1$) *como* E_f de H-L
- If $\beta_1 < 0$, la chance de ocurrir $Y=1$ decrece con x ($e^{\beta} < 1$) *como el ej de occid de tránsito*



Ejemplo 1 (H-L):

Los datos dan la edad en años (AGE) y presencia/ausencia de evidencia de enfermedad coronaria (CHD) para 100 personas seleccionadas para participar en el estudio.

Ajuste del modelo logístico:

```
modelo1 <- glm(CHD ~ AGE, data = dataHL, family = binomial())
```

```
summary(modelo1)
```

Coefficients:

	<i>Estimate</i>	<i>Std.Error</i>	<i>z value</i>	<i>Pr(> z)</i>
(Intercept)	-5.30945	1.13365	-4.683	2.82e-06 ***
AGE	0.11092	0.02406	4.610	4.02e-06 ***

$\hat{\beta}_0$
 $\hat{\beta}_1$

¿Cómo interpretamos el odd de AGE?

$$\hat{\beta}_{AGE} = 0,11$$

$$OR(x) = \frac{odds(x+1)}{odds(x)} = e^{0,11} = 1,12$$

Por ejemplo, si $x = AGE = 50 \Rightarrow$ *40 años... 62 años... cualquier k!!*

$$OR(50) = \frac{ODD(\overbrace{51}^{x+1})}{ODD(\underbrace{50}_x)} = 1,12$$

$$\underbrace{ODD(\overbrace{51}^{x+1})}_{x+1} = 1,12 * \underbrace{ODD(\underbrace{50}_x)}_x$$

chance de
incidente coronario = $1,12 * \text{chance de}$
(vs no incid) IC (vs...)

Si tiene 51 años

Si tiene 50

O sea, con 1 año adicional *(a partir de cualquier x)* aumenta un 12% la

Chance de incidente coronario.

Ejemplo 2

X: # horas de asistencia a programas de educación vial

Y: es responsable de accidente de tránsito? $Y=1$ corresponde a SI

		coef estimados ↓ B	E.T.	Wald	gl	Sig.	e^B Exp(B)
X	$\hat{\beta}_1 = -0,363$	-,363	,051	51,071	1	,000	,695
Constante	$\hat{\beta}_0 = 7,925$	7,925	1,244	40,611	1	,000	2764,595

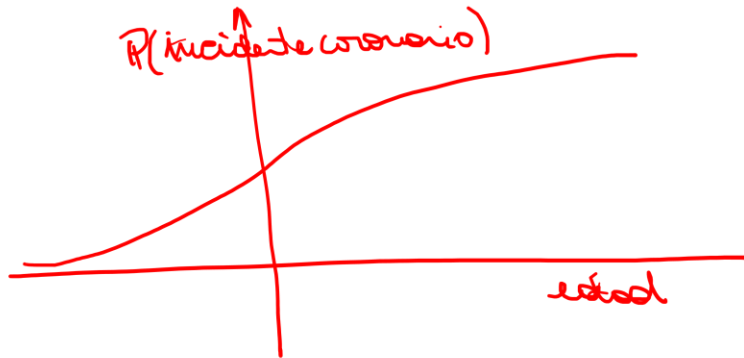
$$\hat{\beta}_1 = -0,363 \Rightarrow e^{\hat{\beta}_1} = 0,695 = OR$$

$$ODD(X+1) = 0,695 * ODD(X)$$

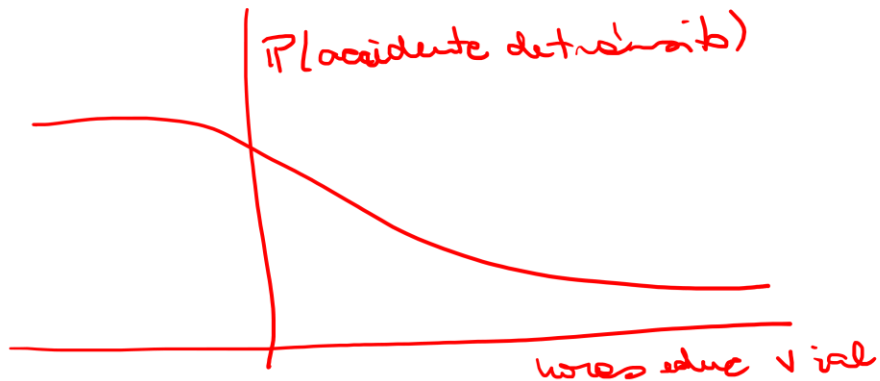
Por cada hora adicional de asistencia vial, la chance de accidente de tránsito disminuye aprox un 30%.

Oss : logistica p/1 variable

Ej H-L:

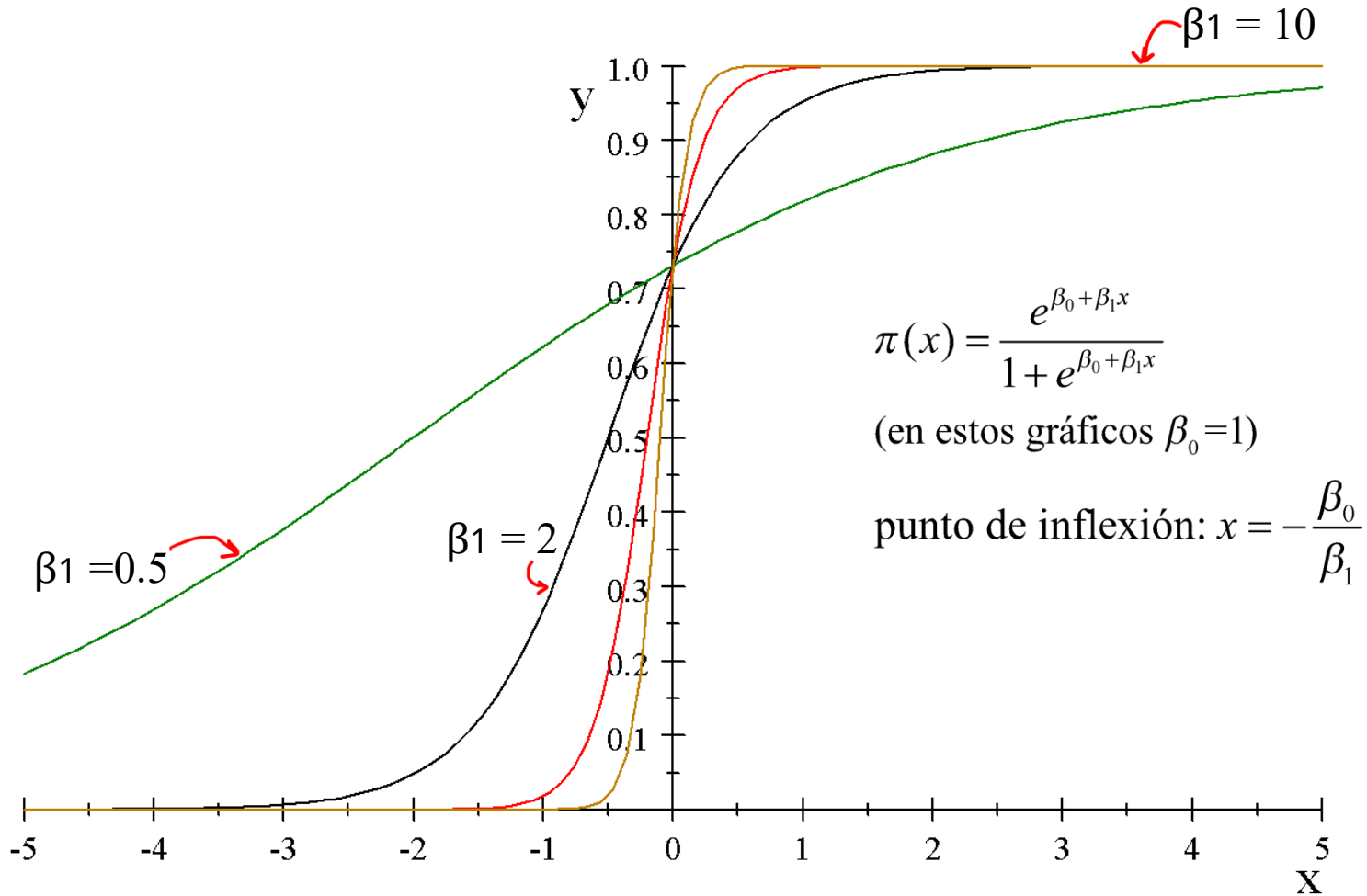


$\beta_1 > 0 \rightarrow$ creciente



$\beta_1 < 0 \rightarrow$ decreciente

Cómo cambia la curva logística cuando crece β ?



Regresión logística múltiple

Bibliografía:


- Chatterjee, S.; Hadi, A.; Price, B. “Regression Analysis by Example”. Wiley (Introductoria)
- Montgomery, Peck y Vining. “Introducción al Análisis de Regresión Lineal” (Introductoria)
- Hosmer, D.; Lemeshow, S.(2000). “Applied logistic regression” (Wiley Series in Probability)

Resumen

- Regresión logística múltiple
- Interpretación de los coeficientes
- Estimación máximo verosímil
- Inferencia en el modelo RL: significación de variables, IC, comparación de modelos.
- Medidas de ajuste: pseudoR^2 , test de Hosmer-Lemeshow
- Métodos de selección de variables
- Medidas Diagnósticas en RL
- RL para clasificación

Modelo de Regresión Logística múltiple

Se propone modelar la probabilidades condicionales mediante una función logística. Para el caso de p predictores sería:

$$\pi(\underline{x}) = P(Y = 1 / \underbrace{x_1, \dots, x_p}_{\text{predictors}}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$


Transformación logit

$$g(\mathbf{x}) = \ln \left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

Handwritten annotations in red:

- An arrow points from $\pi(\mathbf{x})$ to $P(Y=1/x)$.
- An arrow points from $1 - \pi(\mathbf{x})$ to $P(Y=0/x)$.
- A bracket under $1 - \pi(\mathbf{x})$ is labeled $ODD(x)$.

Esto es

$$\pi(\mathbf{x}) = P(Y = 1 / x)$$

$$e^{g(\mathbf{x})} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = ODD(x) = \frac{P(Y = 1 / x)}{P(Y = 0 / x)}$$

$ODD(x)$ = probabilidad de respuesta positiva / prob de respuesta negativa en un cierto nivel \mathbf{x} de factores.

Observaciones

- Las variables predictoras pueden ser continuas o categóricas.
- Los coeficientes β_i son interpretables.
- Permiten ver cómo se incrementan los odds ratios.

Interpretación de los Coeficientes

- En RLM, los coeficientes muestran el cambio en la respuesta media por cada aumento de una unidad en la variable predictora (manteniendo las demás constantes)
- En R Logística, la relación es

$$\frac{odds(x+1)}{odds(x)} = e^{\beta} = OR(x) \quad \left(odds(x) = \frac{\pi(x)}{1 - \pi(x)} \right)$$

Interpretación de los Coeficientes

$$\begin{aligned}
 & \frac{e^{\beta_0 + \beta_1 x_i + \dots + \beta_i (x_i + 1) + \dots}}{e^{\beta_0 + \beta_1 x_i + \dots + \beta_i x_i + \dots}} = e^{\beta_i} \\
 & \Rightarrow \frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)} = e^{\beta_i} = OR(x) \quad \left(\text{odds}(x_i) = \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = e^{\beta_0 + \beta_1 x_i + \beta_2 x_2 + \dots}
 \end{aligned}$$

\downarrow
 $\text{odds}(x_i + 1)$ significa aumentar sólo la componente x_i en 1

- e^{β} representa el cambio en los odds al incrementar x en 1 unidad, dejando las demás variables predictoras fijas.
- Si $\beta_i = 0$, x_i no tiene efecto sobre Y ($e^{\beta}=1$), o sea la chance es la misma en ambos niveles de Y .
- If $\beta_i > 0$, la chance de ocurrir $Y=1$ crece con x_i ($e^{\beta} > 1$)
- If $\beta_i < 0$, la chance de ocurrir $Y=1$ decrece con x_i ($e^{\beta} < 1$)

Interpretación de los coeficientes

- Si $\beta_i = 0$, x_i no tiene efecto sobre el OR (no es factor de riesgo)
- Si $\beta_i \neq 0$, $e^{\beta_i(x_i+1)} = e^{\beta_i x_i} e^{\beta_i}$

$$Ej : \beta_i = 0.5 \Rightarrow e^{\beta_i} = 1.65$$

Entonces un cambio de 1 unidad en x_i aumenta un 65% el Odd (aumenta la chance de tener enfermedad respecto a no tenerla)

Dependiendo de las otras regresores "x" fijos

Recordar que
$$e^{\beta_i} = \frac{e^{\beta_i(x_i+1)}}{e^{\beta_i x_i}}$$

Ejemplo (UCI <https://archive.ics.uci.edu/>)

En el estudio de factores de riesgo asociados con Alzheimer se quiere determinar si los incidentes de demencia pueden relacionarse con el consumo de vino y otras variables: AGE (edad en años), WINE (consumo de vino: 0= no consume; 1= si), MMSE (mini-mental, examen de estado mental, con puntaje de 0 a 30) y T3DEMEN (incidentes de demencia: 1=sí; 0=no).

VAR	DESCRIPCIÓN	TIPO
AGE	Edad en años	Escalar
WINE	Consumo de Vino	Categórica
MMSE	Examen de Estado Mental	Escalar
HIGHBP	Presión Diastólica	Categórica
T3DEMEN	Incidentes de demencia	Categórica

Vamos a R

Ejemplo

Ajustando el modelo con algunas variables:

```
modeloD <- glm(T3DEMEN ~ AGE+WINE+MMSE, data = dataD, family = binomial())  
summary(modelo2)
```

Coefficients:

	<i>Estimate</i>	<i>Std. Error</i>	<i>z value</i>	<i>Pr(> z)</i>
<i>(Intercept)</i>	<i>-1.85787</i>	<i>2.69333</i>	<i>-0.690</i>	<i>0.490</i>
<i>AGE</i>	<i>0.11571</i>	<i>0.02890</i>	<i>4.004</i>	<i>6.23e-05 ***</i>
<i>WINE1</i>	<i>-0.46202</i>	<i>0.34957</i>	<i>-1.322</i>	<i>0.186</i>
<i>MMSE</i>	<i>-0.31753</i>	<i>0.05232</i>	<i>-6.069</i>	<i>1.29e-09 ***</i>

Ejemplo

$$\underline{x} = (\text{AGE}, \text{MMSE}, \text{WINE})$$

$$\text{WINE} = \begin{cases} 0 & \text{no consume} \\ 1 & \text{si si} \end{cases}$$

El modelo ajustado es

$$\log \left(\frac{\pi(\underline{x})}{1 - \pi(\underline{x})} \right) = -1.85 + 0.11 \text{AGE} - 0.31 \text{MMSE} - 0.46 \text{WINE}$$

$$\hat{\pi}(\underline{x}) = P(\widehat{Y=1}/\underline{x}) = \frac{e^{-1.85+0.11\text{AGE}-0.31\text{MMSE}-0.46\text{WINE}}}{1 + e^{-1.85+0.11\text{AGE}-0.31\text{MMSE}-0.46\text{WINE}}}$$

$$\frac{\hat{\pi}(\underline{x})}{1 - \hat{\pi}(\underline{x})} = e^{-1.85+0.11\text{AGE} - \dots - \dots}$$

a) ¿Cuál es el modelo ajustado para el grupo que no consume vino?

b) ¿Qué significa el coeficiente de AGE en cada grupo? Es el mismo?

¿Qué cambia en cada uno?

Rtas:

Para Wine = 0 (no consume vino)

$$\hat{\pi}(x) = \frac{e^{-1.85+0.11AGE-0.31MMSE}}{1 + e^{-1.85+0.11AGE-0.31MMSE}}$$

Para Wine = 1 (consume vino)

$$\hat{\pi}(x) = \frac{e^{-2.31+0.11AGE-0.31MMSE}}{1 + e^{-2.31+0.11AGE-0.31MMSE}}$$

El coeficiente de AGE es el mismo en ambos.
Lo que cambia es la constante.

Interpretación de los coeficientes

$$\beta_{AGE} = 0.11 \Rightarrow e^{\beta_i} = 1.12 = \text{OR}$$

Esto dice que por cada año ^{adicional} aumenta un 12% el Odd (aumenta la chance de tener incidentes de demencia respecto a no tenerlos) **manteniendo las otras variables constantes en el modelo.**

$$\beta_{MMSE} = -0.31 \Rightarrow e^{\beta_i} = 0.73$$

Ej: ^{chance de demencia} $\left\{ \begin{array}{l} AGE = 50 \\ W = 0 \\ MMSE = 23 \end{array} \right\} = 0,73 * \text{chance de demencia en } \left\{ \begin{array}{l} AGE = 50 \\ W = 0 \\ MMSE = 22 \end{array} \right\}$

Esto dice que por cada punto extra en el examen MMSE disminuye un 27% el Odd (disminuye la chance de tener incidentes de demencia respecto a no tenerlos) **manteniendo las otras variables constantes en el modelo.**

Estimación de parámetros en Regresión Logística: Máxima verosimilitud

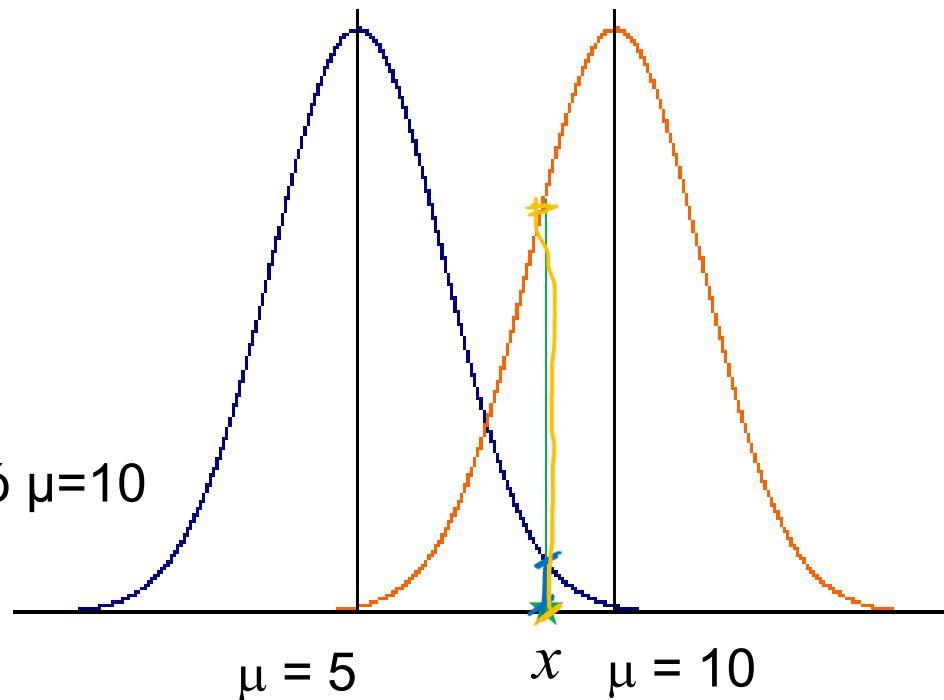
El método de mínimos cuadrados no aplica a este caso.. Pero hay un método más general: **estimación por máxima verosimilitud** (R. A. Fisher, 1890~1962).

La idea es simple:

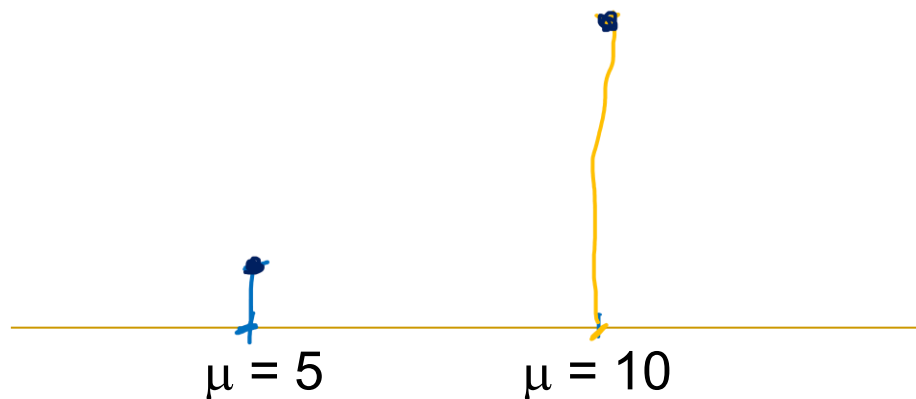
Los parámetros desconocidos de modelo se eligen de modo que maximicen la probabilidad de ocurrencia de los datos observados en la muestra.

Ejemplo bobo: Se observó el valor x .
Cuál μ es más verosímil?

$X \sim N(\mu; \sigma^2)$ con $\mu=5$ ó $\mu=10$



$L(\mu)$:
función de
verosimilitud



Función de verosimilitud (likelihood function)

Supongamos observaciones de las v.a. X_1, \dots, X_n , iid con $f(x_i|\theta)$

La función de verosimilitud de la muestra es

$$L(\theta|x_1, \dots, x_n) = f(x_1|\theta) \dots f(x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

donde θ es el vector de parámetros.

El estimador M.V. es el θ donde se da el máximo de esta función.

En regresión logística los estimadores MV no tienen forma explícita (en regresión múltiple sí!). Aquí salen por métodos numéricos.

Test de cociente de (max) verosimilitudes

Se tiene una distribución asintótica para el estadístico (Wilks, 1935) que prueba las hipótesis:

$$H_0: \theta \in \Omega_0 \text{ versus } H_1: \theta \notin \Omega_0$$

El estadístico del test sale de comparar verosimilitudes:

$$2 \ln \left[\frac{L(\hat{\theta}_{\Omega})}{L(\hat{\theta}_{\Omega_0})} \right] = 2 \ln(L(\hat{\theta}_{\Omega})) - 2 \ln(L(\hat{\theta}_{\Omega_0}))$$

Este test permite ensayar hipótesis basados en los estimadores MV para comparar modelos “anidados”, esto es: modelo completo versus modelo reducido.

Propiedades de los estimadores MV

- Los estimadores MV son asintóticamente insesgados y consistentes.
- Permiten estimar una función del parámetro usando el EMV de este (propiedad de invariancia).
- Para muestras grandes puede asumirse que la distribución de cualquier estimador MV es Normal

$$\hat{\theta}_{MV} \xrightarrow{\approx} N(\theta, \mathbf{I}_{\theta}^{-1})$$

Usando este resultado se tienen los test de Wald (con aproximación normal) que aparecen por default en la salida de R.

Análisis de Devianza

A partir de lo anterior, se define la deviancia (devianza) del modelo con parámetros θ a

$$Deviance(Mod) = -2 \log L_{\text{mod}}$$

La verosimilitud está evaluada en el máximo para el conjunto de parámetros.

El test de cociente de verosimilitudes puede verse como una diferencia de devianzas entre modelos:

$$Deviance(Mod_{Chico}) - Deviance(Mod_{Comp}) \approx \chi^2_{k-q}$$

Observaciones

- En el modelo de regresión esto equivale a los test F pues en este caso $\text{Deviance} = \text{SSE}$.
- R^2 y R^2 ajustado admiten expresiones en términos de devianzas.
- La devianza está definida para cualquier modelo paramétrico (sólo requiere tener una función de verosimilitud).
- La devianza permite definir indicadores para comparación de modelos (AIC, BIC, etc.)

Análisis de Deviance

Se utiliza para:

1. Testear la significancia del modelo *¿vale la pena modelar Y a partir de X_1, \dots, X_p ?*
2. Testear la significatividad de una variable *¿vale la pena agregar esa variable a los q' ya puse?*
3. Comparar modelos anidados en general

Comparación con deviancias

1. Para evaluar la significatividad del modelo se compara $\text{Dev}(\text{Modelo})$ con la deviancia del modelo “nulo”, esto es, con solo el término constante.
2. Para evaluar la significatividad de una variable se compara el modelo con y sin esta, utilizando las deviancias correspondientes.
3. Para comparar otros modelos, si son anidados, se usa la misma estrategia.

1- Test de significatividad del modelo

Equivalente al test F en regresión, podemos hacer test de cociente de verosimilitudes ó test de deviancias para decidir si:

$\gamma \sim 1$ (modelo = modelo nullo)

$H_0 : \beta_1 = \beta_2 = \dots = 0$ (no es signif el mod ~~o sea~~ no vale para modelar γ con X_1, \dots, X_p)

$H_1 : \text{alguno distinto de } 0$

$\gamma \sim X_1, \dots, X_p$

El estadístico de comparación de deviancias se utiliza para testear la significatividad del modelo de p-regresoras, teniendo una distribución χ^2 con p g.l.

Vamos a R

En el ejemplo de demencia: significatividad del modelo

```
modeloD <- glm(T3DEMEN ~ AGE+WINE+MMSE, data = dataD, family =binomial())  
summary(modelo2)
```

Coefficients:

	<i>Estimate</i>	<i>Std. Error</i>	<i>z value</i>	<i>Pr(> z)</i>
(Intercept)	-1.85787	2.69333	-0.690	0.490
AGE	0.11571	0.02890	4.004	6.23e-05 ***
WINE1	-0.46202	0.34957	-1.322	0.186
MMSE	-0.31753	0.05232	-6.069	1.29e-09 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 314.39 on 271 degrees of freedom

Residual deviance: 219.69 on 268 degrees of freedom

AIC: 227.69

Number of Fisher Scoring iterations: 5

En el ejemplo de demencia: significatividad del modelo

H0: modelo sólo con β_0 (modelo null)

H1: modelo completo

Se tiene:

$$\begin{aligned} \text{Dev(mod chico)} - \text{Dev(mod completo)} &= \\ = 314.39 - 219.69 &= 94.69 \end{aligned}$$



Como la distribución de referencia es χ^2 con g.l. = 3, se tiene un p-valor = 0.0000000... lo que indica rechazar la hipótesis nula (el modelo es significativo).

2-Pruebas de significatividad de los coeficientes

¿El modelo que incluye a la variable en cuestión nos dice más acerca de la variable respuesta que el modelo que no incluye esa variable?

Esto puede responderse con prueba de devianzas ó con la prueba de Wald, que da por default R.

Test de Wald para significación de los coeficientes

El test tiene hipótesis

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Basado en la normalidad asintótica de los estimadores MV se tiene que:

$$W = \hat{\beta}'(X'VX)\hat{\beta} \sim \chi^2_{p+1}$$

Particularmente, se puede testear sobre un coeficiente a partir de $\left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right) \approx N(0,1)$

Son los que aparecen en la salida del summary del modelo!

Ojo! Pueden presentarse malas estimaciones por multicolinealidad ó separación.

Vamos a R

Analizando significatividad de variables con Wald

γ
`modelo3 <- glm(T3DEMEN ~ AGE+MMSE, data = dataD, family = binomial())`
`summary(modelo3)`

Coefficients:

	Estimate	Std. Error	z	value Pr(> z)
(Intercept)	-2.09432	2.70623	-0.774	0.439
<u>AGE</u>	0.11235	0.02873	3.910	9.23e-05 ***
MMSE	-0.30671	0.05136	-5.972	2.34e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 314.39 on 271 degrees of freedom

Residual deviance: **221.47** on **269** degrees of freedom

$H_0: \gamma \sim \text{MMSE}$

$H_1: \gamma \sim \text{MMSE} + \text{AGE}$

Intervalos de Confianza para los coeficientes y los odds

Basandonos en el estadístico de Wald, un IC para β_j de nivel 95% es:

$$\hat{\beta}_j \pm 1.96se(\hat{\beta}_j) \equiv \left(\hat{\beta}_j - 1.96se(\hat{\beta}_j) , \hat{\beta}_j + 1.96se(\hat{\beta}_j) \right)$$

Y el correspondiente IC para el odd ratio (¡no es simétrico!) resulta ser:

$$\left(e^{\hat{\beta}_j - 1.96se(\hat{\beta}_j)} , e^{\hat{\beta}_j + 1.96se(\hat{\beta}_j)} \right)$$

2. Significación de una variable con devianzas

Se comparan 2 modelos

$$\text{Log(odds)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (\text{modelo Completo})$$

$$\text{Log(odds)} = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad (\text{modelo Chico})$$

El estadístico basado en devianza se distribuye aprox χ^2 con $df = n^\circ$ parámetros extra en el modelo Completo.

2- En el ejemplo demencia2:

Significatividad de una variable con devianzas

$$\text{Log(odds)} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{MMSE} + \beta_3 \text{WINE}$$

(modelo Completo)

$$\text{Log(odds)} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{MMSE}$$

(modelo Chico)

H0: modelo chico

H1: modelo completo

Vamos a R

2- En el ejemplo de demencia: ajuste modelo sin Wine

```
modelo3<- glm(T3DEMEN ~ AGE+MMSE,data = dataD, family = binomial())  
summary(modelo3)
```

Coefficients:

	<i>Estimate</i>	<i>Std. Error</i>	<i>z</i>	<i>value</i>	<i>Pr(> z)</i>
<i>(Intercept)</i>	-2.09432	2.70623	-0.774	0.439	
<i>AGE</i>	0.11235	0.02873	3.910	9.23e-05	***
<i>MMSE</i>	-0.30671	0.05136	-5.972	2.34e-09	***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 314.39 on 271 degrees of freedom
*Residual deviance: **221.47** on **269** degrees of freedom*

2- Significatividad de una variable con devianzas

$$\text{Log(odds)} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{MMSE} + \beta_3 \text{WINE}$$

(modelo Completo)

$$\text{Log(odds)} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{MMSE}$$

(modelo Chico)

H0: modelo chico

H1: modelo completo

Se tiene:

$$\begin{aligned} \text{dev(mod chico)} - \text{dev(mod completo)} &= \\ &= 221.47 - 219.69 = 1.78 \end{aligned}$$

Como la distribución de referencia es χ^2 con g.l. = 1, se tiene un p-valor = 0.1821 lo que indica no rechazar la hipótesis nula (elegimos el modelo chico) o sea « no es significativo agregar wine al modelo »

3- Comparación de modelos anidados

$$\text{Log(odds)} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{MMSE} + \beta_3 \text{WINE}$$

(modelo grande)

$$\text{Log(odds)} = \beta_0 + \beta_1 \text{AGE}$$

(modelo Chico)

H0: modelo chico

H1: modelo grande

```
> anova(modeloD3,modeloD,test = "Chisq")
```

Analysis of Deviance Table

Model 1: T3DEMEN ~ AGE

Model 2: T3DEMEN ~ AGE + WINE + MMSE

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	270	269.24			
2	268	219.69	2	49.549	1.74e-11 ***

Pruebas para evaluación del modelo

Test de Hosmer- Lemeshow

H0: el modelo ajusta a los datos H1: no ajusta

La prueba agrupa los datos en 10 grupos de igual tamaño, a partir de sus probabilidades estimadas. Luego compara las frecuencias observadas con las esperadas en cada partición (similar al estadístico de Pearson). Puede aplicarse con predictores continuos.

El estadístico se distribuye también χ^2 con

g.l.= cantidad de grupos $(10) - 2 = 8$

OJO!! Se busca tener p-valor alto

Estadístico de H-L

$$\hat{C} = \sum_{j=1}^{10} \frac{(O_j - n_j \bar{\pi}_j)^2}{n_j \bar{\pi}_j (1 - \bar{\pi}_j)} \sim \chi_8^2$$

n_j = # sujetos en el grupo j

O_j = # sujetos en el decil j

$\bar{\pi}_j$ = promedio ponderado de las $\hat{\pi}_i$ estimadas en todos los que corresponden al grupo j

OBS: La prueba HL con diez intervalos puede fallar (lo que resulta en un valor p bajo..) en un modelo con pocos predictores. Se puede probar reducir el número de intervalos con la opción `g` en R.

Pseudo R cuadrados

Tratan de copiar la interpretación del coeficiente de determinación en RLM, pero los valores en RL son usualmente bajos, aún ante un buen modelo (H-L. pag 167).

➤ Nagelkerke (toma valores entre 0 y 1)

$$R_{Nag}^2 = \frac{1 - [L(L(\text{mod } \beta_0)/L(\text{mod } C \text{ omp}))]^{2/n}}{1 - [LL(\text{mod } \beta_0)]^{2/n}}$$

➤ Cox-Snell (no llega a 1)

$$R_{Cox-Snell}^2 = 1 - \left[\frac{L(\text{mod } \beta_0)}{L(\text{mod } C \text{ omp})} \right]^{2/n}$$

Diagnósticos: evaluando puntos influyentes y outliers

- Residuos de Pearson y de deviancias: para ver puntos de mal ajuste.
- Leverage: Para detectar observaciones outliers.
- Distancia de Cook para detectar puntos influyentes.

Vamos a R
