Tema 4: Otros Métodos de Análisis de Datos Cuantitativos y Cualitativos

Metodología de la Investigación en Fisioterapia

Miguel González Velasco
Departamento de Matemáticas. Universidad de Extremadura

Índice

Análisis de la Covarianza

- El Análisis de la Covarianza (ANCOVA) combina elementos de regresión y de análisis de la varianza.
  - Variable respuesta continua.
  - Una o más variables explicativas o regresoras cuantitativas (covariables).
  - Uno o varios factores explicativos.
- Se trata de establecer la relación lineal que une la variable respuesta con las variables explicativas para cada uno de los niveles del factor o factores explicativos.

Análisis de la Covarianza

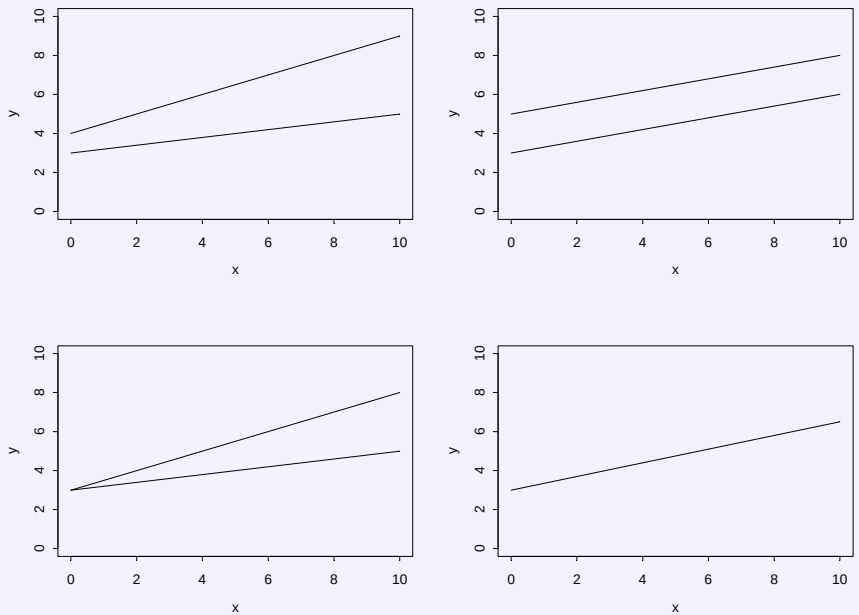
Ejemplo:

- Supongamos que queremos modelar en una determinada población el peso (variable respuesta) como función de la edad (variable explicativa cuantitativa) y el sexo (factor explicativo).
- Modelo Maximal: cuatro parámetros

$$peso_{hombre} = \alpha_{hombre} + \beta_{hombre} \times edad_{hombre} + error$$

$$peso_{mujer} = \alpha_{mujer} + \beta_{mujer} \times edad_{mujer} + error$$

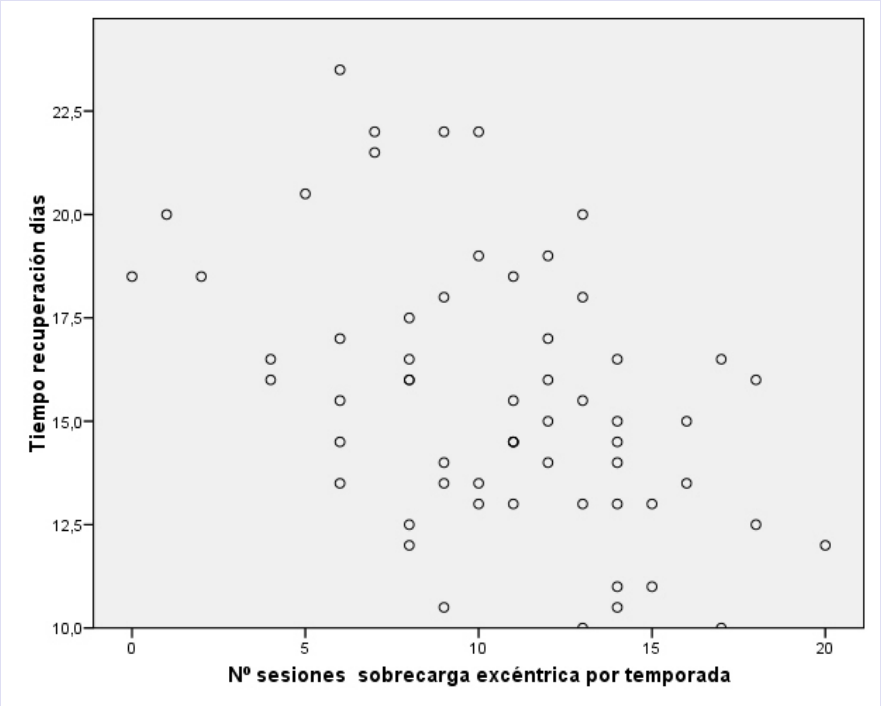
- Modelos posibles:



ANCOVA: Factor con dos categorías

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y: tiempo de recuperación tras una rotura fibrilar.
- Variables explicativas:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.



ANCOVA: Factor con dos categorías

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y: tiempo de recuperación tras una rotura fibrilar.
- Variables explicativas:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,465[a]	,216	,203	2,9006

a. Variables predictoras: (Constante), Nº sesiones sobrecarga excéntrica por temporada

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	134,752	1	134,752	16,016	,000[a]
	Residual	487,994	58	8,414		
	Total	622,746	59			

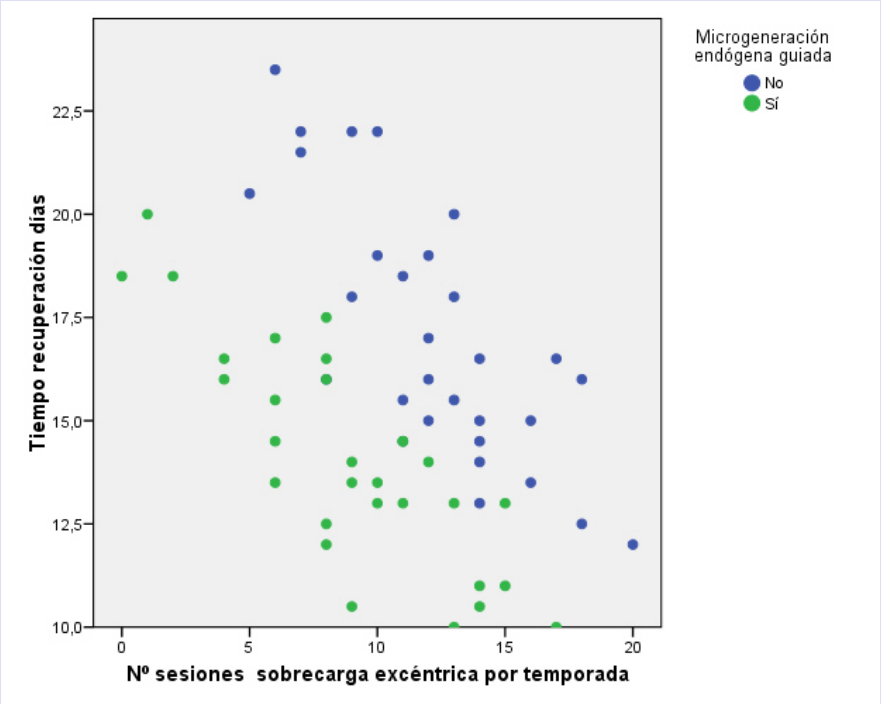
a. Variables predictoras: (Constante), Nº sesiones sobrecarga excéntrica por temporada

b. Variable dependiente: Tiempo recuperación días (escenario 1)

ANCOVA: Factor con dos categorías

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y: tiempo de recuperación tras una rotura fibrilar.
- Variables explicativas:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
  - Factor: Microgeneración endógena guiada (MEG) –Dos niveles: 0-No, 1-Sí.



ANCOVA: Factor con dos categorías

Modelo:

- Variable respuesta: Y
- Variable Explicativa Cuantitativa: X_1
- Factor Explicativo con dos categorías: A y B
  - Variable "dummy"(falsa): $X_2 = 1$ si ocurre A; $= 0$ si no ocurre A

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

- $B \rightsquigarrow Y = \beta_0 + \beta_1 X_1 + \varepsilon$
- $A \rightsquigarrow Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \varepsilon$

Contraste de Hipótesis:

- $H_0 : \beta_3 = 0$ (rectas paralelas) vs $H_1 : \beta_3 \neq 0$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

- $H_0 : \beta_2 = 0$ (misma recta) vs $H_1 : \beta_2 \neq 0$

ANCOVA: Factor con dos categorías

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y: tiempo de recuperación tras una rotura fibrilar.
- Variables explicativas:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
  - Factor: Microgeneración endógena guiada (MEG) –Dos niveles: 0-No, 1-Sí.

Pruebas de los efectos inter-sujetos

Variable dependiente: Tiempo recuperación días (escenario 1)

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial
Modelo corregido	435,308[a]	3	145,103	43,352	,000	,699
Intersección	3204,675	1	3204,675	957,448	,000	,945
W1	304,546	1	304,546	90,988	,000	,619
MEG	65,951	1	65,951	19,704	,000	,260
MEG * W1	4,063	1	4,063	1,214	,275	,021
Error	187,438	56	3,347			
Total	15208,750	60				
Total corregida	622,746	59				

a. R cuadrado = ,699 (R cuadrado corregida = ,683)

Estimaciones de los parámetros

Variable dependiente: Tiempo recuperación días (escenario 1)

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%.		Eta al cuadrado parcial
					Límite inferior	Límite superior	
Intersección	18,708	,787	23,768	,000	17,131	20,285	,910
W1	-,514	,080	-6,442	,000	-,674	-,354	,426
[MEG=0]	6,267	1,412	4,439	,000	3,438	9,095	,260
[MEG=1]	0[a]	.	.	.	.	.	.
[MEG=0] * W1	-,134	,122	-1,102	,275	-,378	,110	,021
[MEG=1] * W1	0[a]	.	.	.	.	.	.

a. Al parámetro se le ha asignado el valor cero porque es redundante.

ANCOVA: Factor con dos categorías

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y: tiempo de recuperación tras una rotura fibrilar.
- Variables explicativas:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
  - Factor: Microgeneración endógena guiada (MEG) –Dos niveles: 0-No, 1-Sí.

Pruebas de los efectos inter-sujetos

Variable dependiente: Tiempo recuperación días (escenario 1)

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial
Modelo corregido	431,245[a]	2	215,622	64,180	,000	,692
Intersección	3391,868	1	3391,868	1009,585	,000	,947
W1	300,707	1	300,707	89,505	,000	,611
MEG	296,493	1	296,493	88,251	,000	,608
Error	191,501	57	3,360			
Total	15208,750	60				
Total corregida	622,746	59				

a. R cuadrado = ,692 (R cuadrado corregida = ,682)

Estimaciones de los parámetros

Variable dependiente: Tiempo recuperación días (escenario 1)

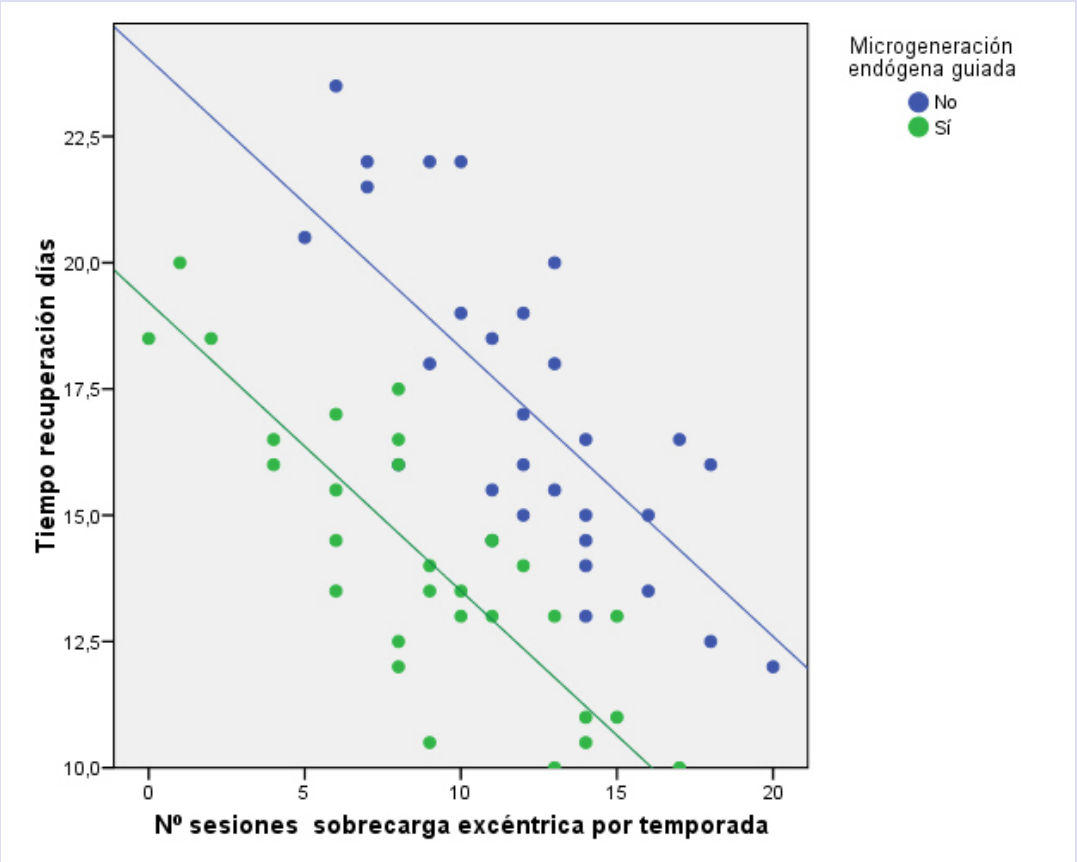
Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%.		Eta al cuadrado parcial
					Límite inferior	Límite superior	
Intersección	19,222	,635	30,272	,000	17,951	20,494	,941
W1	-,572	,060	-9,461	,000	-,692	-,451	,611
[MEG=0]	4,817	,513	9,394	,000	3,790	5,844	,608
[MEG=1]	0[a]	.	.	.	.	.	.

a. Al parámetro se le ha asignado el valor cero porque es redundante.

ANCOVA: Factor con dos categorías

Ejemplo: `Rotura fibrilar`

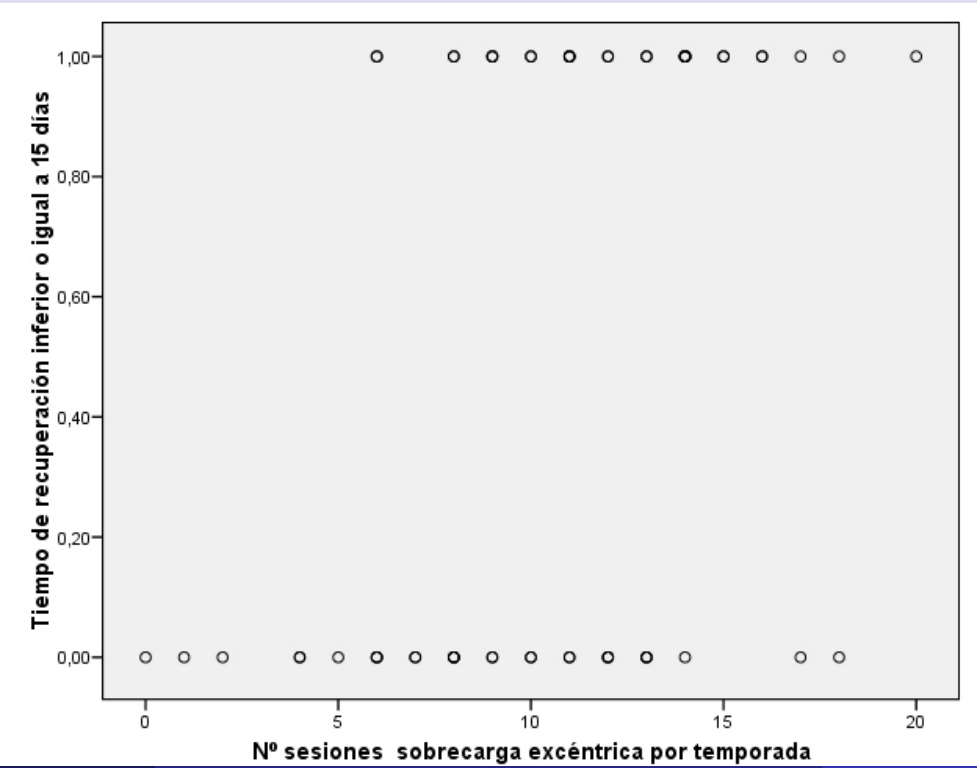
- Variable respuesta Y: tiempo de recuperación tras una rotura fibrilar.
- Variables explicativas:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
  - Factor: Microgeneración endógena guiada (MEG) –Dos niveles: 0-No, 1-Sí.



Modelo de Regresión Logística Simple

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y:
  - 1 si el tiempo de recuperación tras una rotura fibrilar es menor o igual a 15 días.
  - 0 si el tiempo de recuperación tras una rotura fibrilar es mayor de 15 días.
- Variable explicativa:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.



Modelo de Regresión Logística Simple

Definición

Variable Respuesta Dicotómica: $Y = 1$ si ocurre un determinado suceso; $Y = 0$ si no ocurre dicho suceso.

- Y sigue una distribución de Bernoulli.
- Suceso = muerte, curación, enfermedad, recuperación,...

Variable Regresora: X.

$$p_x = E[Y|X = x] \quad \text{prob. de que ocurra el suceso cuando } X = x$$

$$p_x = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))} \Longleftrightarrow \log\left(\frac{p_x}{1 - p_x}\right) = \beta_0 + \beta_1 x$$

$$\frac{p_x}{1 - p_x} = \exp(\beta_0 + \beta_1 x)$$

Modelo de Regresión Logística Simple

Muestra

Muestra aleatoria de la población bidimensional (X, Y)

Y	y_1	y_2	y_3	$\ldots$	y_n
X	x_1	x_2	x_3	$\ldots$	x_n

Inferencia

- Estimación: parámetros β_0, β_1
- Test de Hipótesis:
  - Adecuación del Modelo: Test de Hosmer-Lemeshow
  - Parámetros del modelo: Para cada $i = 0,\ 1, \qquad H_0 : \beta_i = 0$
- Ajuste del Modelo: Coeficiente R^2 de Nagelkerke.

Modelo de Regresión Logística Simple

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y:
  - 1 si el tiempo de recuperación tras una rotura fibrilar es menor o igual a 15 días.
  - 0 si el tiempo de recuperación tras una rotura fibrilar es mayor de 15 días.
- Variable explicativa:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
- Test de Hosmer-Lemeshow: H_0: Modelo adecuado vs H_1: Modelo inadecuado

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	5,166	8	,740

- Coeficiente R^2 de Nagelkerke: Se encuentra entre 0 y 1 y se interpreta como el coeficiente R^2 en regresión lineal.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	71,796[a]	,173	,230

Modelo de Regresión Logística Simple

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y:
  - 1 si el tiempo de recuperación tras una rotura fibrilar es menor o igual a 15 días.
  - 0 si el tiempo de recuperación tras una rotura fibrilar es mayor de 15 días.
- Variable explicativa:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
- Estimación de los parámetros (β_0, β_1) y test de hipótesis (para cada $i = 0,\ 1$, $H_0 : \beta_i = 0$)

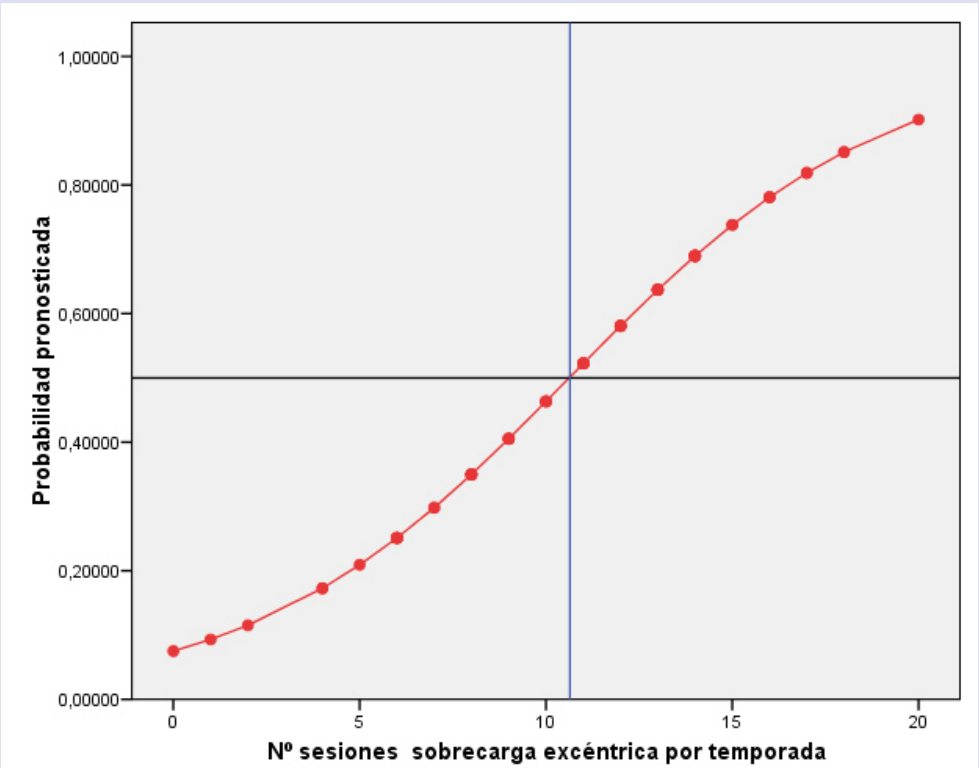
		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	W1	,236	,080	8,754	1	,003	1,267	1,083	1,482
	Constante	-2,512	,901	7,773	1	,005	,081		

a. Variable(s) introducida(s) en el paso 1: W1.

Modelo de Regresión Logística Simple

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y:
  - 1 si el tiempo de recuperación tras una rotura fibrilar es menor o igual a 15 días.
  - 0 si el tiempo de recuperación tras una rotura fibrilar es mayor de 15 días.
- Variable explicativa:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
- Representación modelo ajustado:



Modelo de Regresión Logística Simple

Ejemplo: `Rotura fibrilar`

- Variable respuesta Y:
  - 1 si el tiempo de recuperación tras una rotura fibrilar es menor o igual a 15 días.
  - 0 si el tiempo de recuperación tras una rotura fibrilar es mayor de 15 días.
- Variable explicativa:
  - Variable cuantitativa X: número de sesiones de sobrecarga excéntrica por temporada.
- Clasificación:

Tabla de clasificación [a]

Observado			Pronosticado		
			Y2		Porcentaje correcto
			,00	1,00	
Paso 1	Y2	,00	19	11	63,3
		1,00	9	21	70,0
	Porcentaje global				66,7

a. El valor de corte es ,500

Modelo de Regresión Logística Múltiple

Definición

Variable Respuesta Dicotómica: $Y = 1$ si ocurre un determinado suceso; $Y = 0$ si no ocurre dicho suceso.

- Y sigue una distribución de Bernoulli.

Variables Regresoras: $X_1, \ldots, X_k$.

$$X = (X_1, \ldots, X_k) \qquad x = (x_1, \ldots, x_k)$$

$$E[Y|X = x] = p_x \quad \text{prob. de que ocurra el suceso cuando } X = x$$

$$p_x = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k))} \Longleftrightarrow \log\left(\frac{p_x}{1 - p_x}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$\frac{p_x}{1 - p_x} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$$

Modelo de Regresión Logística Múltiple

Muestra

Variable Respuesta: Y

Variables Regresoras: $X_1, X_2, \ldots, X_k$

Datos:

Y	X_1	X_2	$\ldots$	X_k
Y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
Y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
Y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

Inferencia

- Estimación: parámetros $\beta_0 \ldots \beta_k$
- Test de Hipótesis:
  - Adecuación del Modelo: Test de Hosmer-Lemeshow
  - Parámetros del modelo: Para cada $i = 1, \ldots, k$, $\quad H_0 : \beta_i = 0$
- Ajuste del Modelo: Coeficiente R^2 de Nagelkerke.
- Selección de Variables: Métodos Forward y Backward.

Homogeneidad

Sea X una variable cualitativa con categorías $A_1, \ldots, A_r$, definida en s poblaciones $B_1, \ldots, B_s$. Estamos interesados en saber si la variable X se comporta igual en las s poblaciones.

Los parámetros de interés serán las probabilidades condicionadas $P(A_i|B_j)$ que representan la probabilidad de que un individuo de la población B_j presente la modalidad A_i de la variable X.

Las poblaciones serán homogéneas en lo que respecta a la variable X si

$$P(A_i|B_1) = \cdots = P(A_i|B_s) \qquad \text{para} \quad i = 1, \ldots, r$$

es decir, si las probabilidades de cada modalidad de X son iguales en todas las poblaciones.

Tablas de contingencia: Introducción

Independencia

Sea X una variable cualitativa con modalidades $A_1, \ldots, A_r$, e Y otra variable cualitativa con modalidades $B_1, \ldots, B_s$, ambas definidas en la misma población.
Estamos interesados en saber si las variables X e Y presentan relación.
Los parámetros de interés serán las probabilidades condicionadas $P(A_i|B_j)$ que representan la probabilidad de que un individuo que presenta la modalidad B_j del carácter Y presente la modalidad A_i de la variable X. También en este caso son de interés las probabilidades condicionadas $P(B_j|A_i)$.
Las variables X e Y son independientes si

$$P(A_i|B_1) = \cdots = P(A_i|B_s) \qquad \text{para} \quad i = 1, \ldots, r$$

es decir, si la probabilidad de que un individuo presente una determinada modalidad de X no depende de qué modalidad de Y presente dicho individuo.

Tablas de contingencia: Introducción

Toma de datos

En cualquiera de las situaciones anteriores resolveremos el problema en base a una muestra de T individuos. Los datos se representan en una tabla de contingencia:

	B_1	$\dots$	B_j	$\dots$	B_s	Total
A_1	O_{11}	$\dots$	O_{1j}	$\dots$	O_{1s}	F_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	O_{i1}	$\dots$	O_{ij}	$\dots$	O_{is}	F_i
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	O_{r1}	$\dots$	O_{rj}	$\dots$	O_{rs}	F_r
Total	C_1	$\dots$	C_j	$\dots$	C_s	T

Frecuencias absolutas de la muestra (valores observados)

$O_{ij} \equiv$ número de individuos que presentan simultáneamente la modalidad A_i de la variable X y la modalidad B_j de la variable Y.

$F_i \equiv$ número de individuos que presentan la modalidad A_i de la variable X.

$C_j \equiv$ número individuos que presentan la modalidad B_j de la variable Y.

Valores esperados

Se calculan mediante la expresión

$$E_{ij} = \frac{F_i C_j}{T}, \quad i = 1, \ldots, r;\ j = 1, \ldots, s.$$

Son los valores que cabría esperar para O_{ij} si hubiese homogeneidad de las poblaciones o independencia de las variables. Por lo tanto, bajo la hipótesis de homogeneidad o la de independencia $O_{ij} - E_{ij}$ deberían ser próximos a 0.

Si hemos comprobado que no hay independencia, para saber a cual de las categorías de X o de Y hemos de atribuir la relación entre ambas variables, hemos de buscar en la tabla los valores de $|O_{ij} - E_{ij}|$ más altos.

Tablas de contingencia: Contraste de homogeneidad

Contraste de hipótesis

$$H_0 : P(A_i|B_1) = \cdots = P(A_i|B_s) \qquad \text{para} \quad i = 1, \ldots, r$$
$$H_1 : \text{ estas probabilidades no son iguales para algún } i$$

dicho de forma intuitiva, contrastamos

H_0 : Las poblaciones son homogéneas con respecto a X

H_1 : Las poblaciones no son homogéneas con respecto a X

El valor experimental se calcula mediante la fórmula:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Rechazamos H_0 al nivel de significación α si $\chi^2 > \chi^2_\alpha((r-1)(s-1))$, siendo $\chi^2_\alpha((r-1)(s-1))$ el cuantil de orden $1-\alpha$ de la distribución chi-cuadrado con $(r-1)(s-1)$ grados de libertad.

Este test es válido si:

- Ningún E_{ij} es < 1.
- A lo sumo un 20% de los E_{ij} son < 5.

Cuando no se verifican las condiciones de validez se puede utilizar el Test Exacto de Fisher.

Tablas de contingencia: Contraste de independencia

Contraste de hipótesis

$$H_0 : P(A_i|B_1) = \cdots = P(A_i|B_s) \qquad \text{para} \quad i = 1, \ldots, r$$
$$H_1 : \text{ estas probabilidades no son iguales para algún } i$$

dicho de forma intuitiva, contrastamos

$$H_0 : \text{ Las variables } X \text{ e } Y \text{ son independientes}$$
$$H_1 : \text{ hay relación entre las variables}$$

El valor experimental se calcula mediante la fórmula:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Rechazamos H_0 al nivel de significación α si $\chi^2 > \chi^2_\alpha((r-1)(s-1))$, siendo $\chi^2_\alpha((r-1)(s-1))$ el cuantil de orden $1-\alpha$ de la distribución chi-cuadrado con $(r-1)(s-1)$ grados de libertad.

Este test es válido si:

- Ningún E_{ij} es < 1.
- A lo sumo un 20% de los E_{ij} son < 5.

Cuando no se verifican las condiciones de validez se puede utilizar el Test Exacto de Fisher.

Tablas de contingencia: Contraste de independencia

Medidas de asociación

En caso de ser significativo el contraste anterior, calculamos grado de relación con el coeficiente de contingencia de Pearson:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + T}} = \sqrt{\frac{\sum_{i,j} \frac{O_{ij}^2}{E_{ij}} - T}{\sum_{i,j} \frac{O_{ij}^2}{E_{ij}}}}$$

C toma valores entre 0 (asociación nula o independencia) y $\sqrt{(q-1)/q}$ (asociación máxima), siendo $q = \min\{r, s\}$.

Para tablas 2×2 tenemos el coeficiente ϕ definido

$$\phi = \sqrt{\frac{\chi^2}{T}} = \sqrt{\frac{(O_{11}O_{22} - O_{12}O_{21})^2}{F_1 F_2 C_1 C_2}}$$

ϕ toma valores entre 0 (asociación nula o independencia) y 1 (asociación máxima).