Regresión Logística
Aplicación a Modelos Predictivos en CCR

Juan Manuel García Torrecillas
MF en Servicio de Urgencias. HU Torrecárdenas
Coordinador de Investigación. Proyecto FIS PI19/01931
Instituto de Salud Carlos IIII. CIBERESP (Madrid, España)

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD
Instituto de Salud Carlos III
UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"

Análisis Multivariante

Cuando se quiere valorar el efecto de un número extenso de factores de riesgo -predictores- sobre una variable dependiente ES NECESARIO utilizar técnicas estadísticas que puedan valorar el efecto individual de cada factor de riesgo sobre la VD, pero quitando la confusión aportada por las demás variables



Análisis Multivariantes

- **Son aproximaciones a funciones matemáticas capaces de predecir el comportamiento de la VD** a partir de otras variables (VI)

- Estos modelos se usan cada vez más frecuentemente en epidemiología

- **Cuando la VD es dicotómica (presencia/ausencia de enfermedad):**
  - Valor 0: Ausencia
  - Valor 1: Presencia
  - p= probabilidad de que aparezca la enfermedad/evento
  - 1-p=probabilidad de que no aparezca
  - Si la VD depende del tiempo→ Regresión de Cox
  - Si la VD no dependen del tiempo→ Regresión Logística

Según sea la VD…

- VD = Continua → Regresión Logística Múltiple
- VD = Dependiente del tiempo → Regresiones de Cox
  → Supervivencia
- VD = Dicotómica → Regresión Logística Binaria

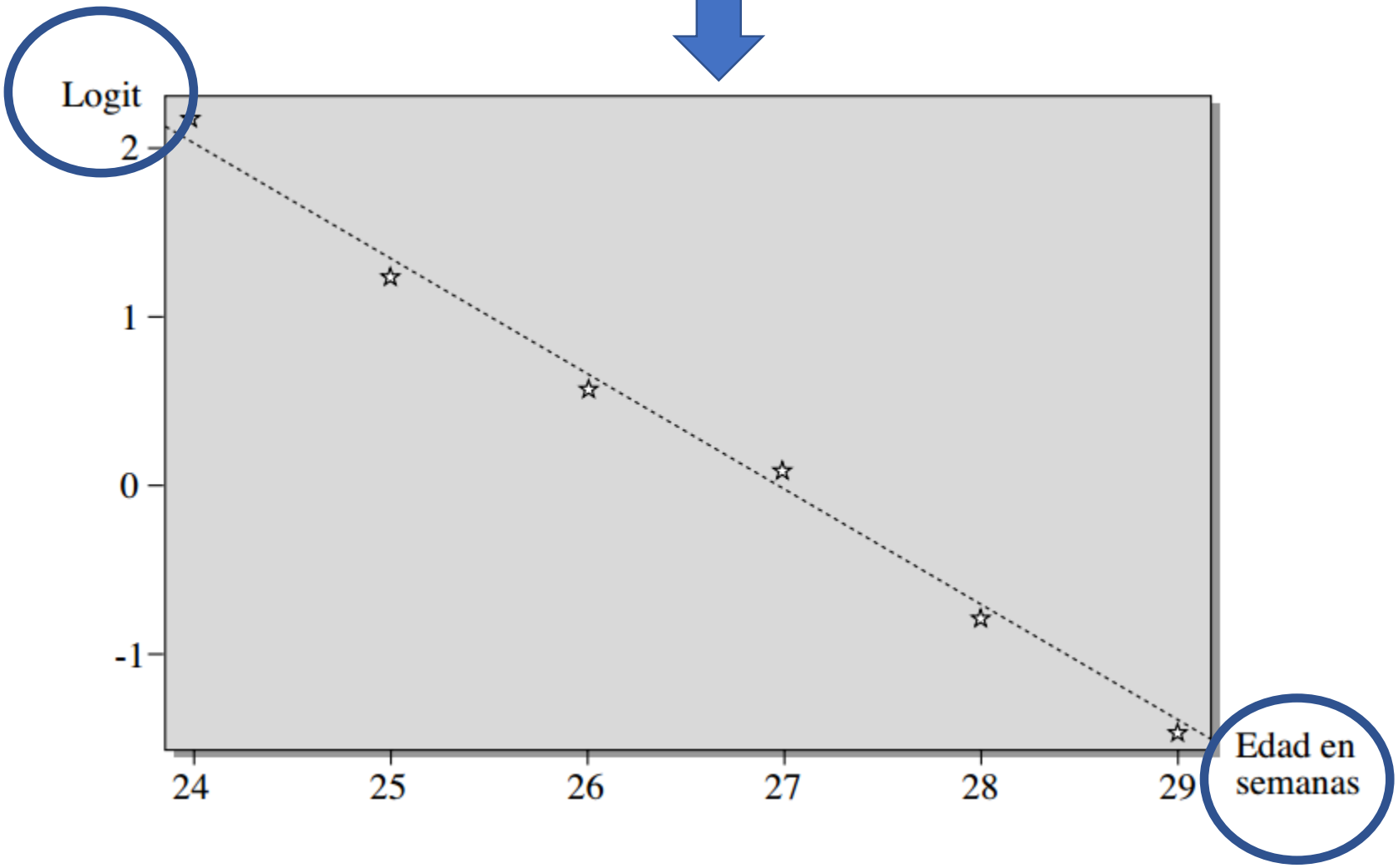
Asunciones de los modelos de Regresión

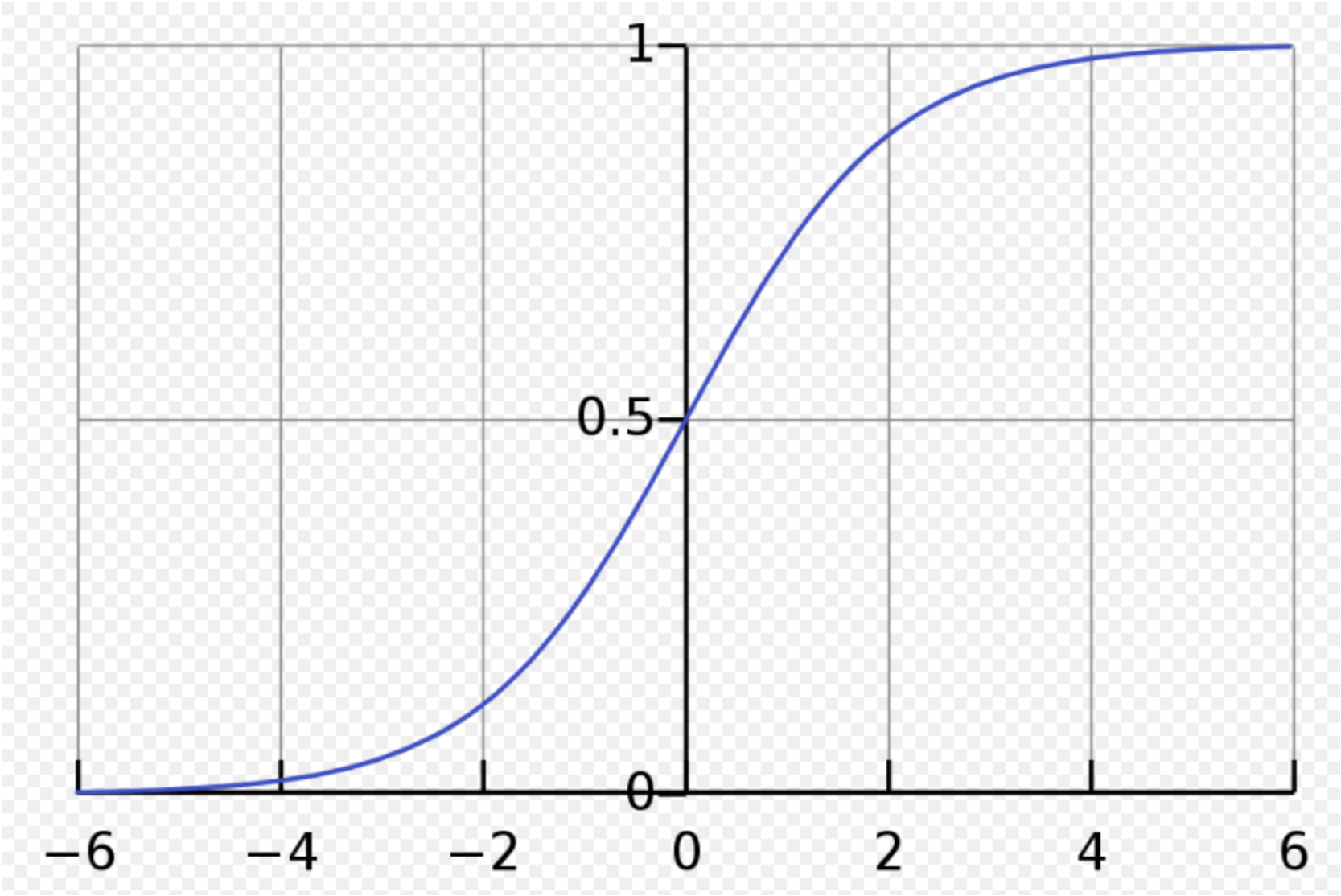
- **Asunción de Linealidad**
  - Aplica a las VI continuas del modelo
  - → **Cada cambio continuo en la VI provoca un cambio lineal en la VD**
  - En RL: Trazamos en un eje el Logit de la probabilidad obtenido por cada cambio en la VI. Debe dar una línea recta.

- **Asunción de Aditividad**
  - Los predictores del modelo se relacionan con la VD en modo aditivo
  - →Cada factor que incluyo induce un aumento en la variable resultado
  - **El efecto aportado por cada FR se suma al aportado por las demás**

- **Propiedad Distributiva**
  - La **VD sigue una distribución Normal** (En R. Lineal Múltiple, no en logística)
  - Tampoco es necesaria en la Regresión de Cox

Asunción de Linealidad





Fuente: https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica#/media/File:Logistic-curve.svg

Regresión logística Binaria

Situación de Logística:

- Una VD con únicamente dos opciones (dicotómica).
- Se quiere calcular la probabilidad de que ocurra una u otra a partir de una serie de informaciones (variables independientes).

Ejemplos:

- Cursa alta del hospital Vivo vs Fallecido
- A 30 días del alta reingresa: Sí vs No

Regresión logística

- Mantiene cierta analogía con el análisis de regresión múltiple excepto que la variable dependiente es dicotómica.

- Es algo así como un caso particular del análisis discriminante, en el que la variable dependiente tiene dos estados, los supuestos son menos restrictivos y se admiten variables independientes categóricas.

Ejemplo de estudio activo

- Estamos interesados en conocer qué factores o variables predictoras se asocian a la probabilidad de fallecer durante el ingreso de paciente con cáncer colorrectal que son admitidos para cirugía.

- VD = Exitus: Sí o No
- VI = Predictoras = Cuantitativas y Cualitativas

Concepto matemático (mínimo…)

- El análisis de RL se basa en intentar explicar una fx matemática sigmoidea (distribución logística), cuya VD sólo puede tomar los valores comprendidos entre 0 y 1.
- Expresión matemática:
  - $F(p) = 1 / 1 + e^{-z}$
    - F(p) = Distribución logística, proporciona valores continuos de "p" entre el 0 y el 1
    - e = número e
    - Z = Sumatorio de los productos de cada VI (Xi) por un coeficiente, según la fórmula:

$$Z = \beta_0 + \beta_{1*} X_1 + \beta_{2*} X_2 + \beta_{3*} X_3 + \ldots + \beta_{i*} X_i$$

En donde,

$\beta_0 \ldots \beta_i$ representan a los coeficientes y,

$X_0 \ldots X_i$ representan a las variables independientes o explicativas del modelo

- P = probabilidad de fallecer $P = \beta_0 + \beta_1 X$

- Donde X es la predictora y Beta = Coeficiente y/o Intercepto.

- Problemas clave con este modelo:
  - Tras el cálculo puede dar valores por debajo o por encima de 1, imposibles¡

- Modo de evitar el problema: Transformación Logit
  - Transformar la probabilidad de presentar la enfermedad, VD….
  - Esto = Transformación Logit = logaritmo de la razón de presentar la características P/(1-P)==

$$log \frac{P}{1 - P} = \beta_0 + \beta_1 X$$

- La nueva expresión no tiene el anterior problema, esta ecuación (Logit) sólo va entre [0,1] como probabilidad:

$$log \frac{P}{1 - P} = \beta_0 + \beta_1 X$$

A $\log[P/(1-P)]$ se le llama **logit(*P*)**

- ¿Cómo interpreto en RL el cambio en 1 unidad de β_1?

Si ocurre que el cambio es sólo de 1 unidad, entonces Log(OR) = β_1

O lo que es igual: $O.R. = e^{\beta_1}$

- En consecuencia, en una RL el Beta(1) es el logaritmo de la OR entre dos individuos que difieren en una unidad en términos de la variable predictora
- Criterio de estimación de los parámetros en una RL = Máxima Verosimilitud:
  - Es más complejo que los mínimos cuadrados de la Regresión Simple
  - No hay fórmula específica, es un cálculo matemático complejo
- Salida típica de regresión logística

Variable	Estimación	Error estandar
Constante	18.36	2.375
Edad gestacional	-0.6819	0.08908



$$\log \frac{P}{1-P} = 18.36 - 0.6819\ Edad$$

TABLA 1. DESCRIPTIVOS

Edad (media ± ds)	69,56±11,67	EPOC n(%)	6301 (6,8)
Estancia (media ± ds)	12,50±6,74	C. Isquem (n%)	2608 (2,8)
Estancia preoperatoria (media ± ds)	2,51±3,96	HTA	38409 (41,4)
NDA (media ± ds)	5,96±3,36	Obesidad	4157 (4,5)
NPA (media ± ds)	3,77±2,94	Ictus n(%)	1203 (1,3)
Sexo (Hombres)	55735 (60,1)	Insuf Renal	4019 (4,3)
Tipo de ingreso (programado)	71797 (77,4)	Anemia	10105(10,9)
Reingresos n (%)	15111 (16,3)	FA	7082 (7,6)
Mortalidad n (%)	3930 (4,2)	Diabetes	17210(18,6)
		Dislipemia	18826 (20,3)
			N=92768

Elección de VA: Bivariante

2A.BIVARIANTE PARA LAS CUANTITATIVAS

VARIABLES	EXITUS	VIVOS	IC Dif 95%	p
Edad $\bar{X}$(DS)	77,44(10,22)	69,22(11,61)	[7,85-8,59]	<0.0001
Estancia $\bar{X}$ (DS)	14,21(8,77)	12,43(6,22)	[1,57-2,00]	<0,0001
Estancia Preoperatoria $\bar{X}$ (DS)	4,24(6,1)	2,44(3,83)	[1,67-1,92]	<0,0001
NDA $\bar{X}$ (DS)	9,44(3,53)	5,81(3,27)	[3,53-3,74]	<0,0001
NPA $\bar{X}$ (DS)	6,41(4,41)	3,65(2,80)	[2,66-2,85]	<0,0001

2.B. BIVARIANTE PARA LAS CUALITATIVAS

VARIABLES		Exitus n(%)	OR	IC 95% OR	p
Sexo	Mujer	1428(3,9)	1		
	Hombre	2502(4,5)	1,17	1,09-1,25	<0,0001
Laparoscopia (n,%)	No	3848(4,5)	1		
	Si	82(1,1)	4,17	3,35-5,21	<0,0001
Tipo de Ingreso	IP	1628(2,3)	1		
	IU	2299(11,0)	5,32	4,99-5,68	<0,0001
Reingreso	No	3144(4,0)	1		
	Si	786(5,2)	1,3	1,20-1,41	<0,0001
Localización Recto	No	3399(4,8)	1		

	Si	910(4,2)	0,044	0,910-1,059	0,63
Localización Unión recto-sigma	No	3524(4,3)	1		
	Si	406(4,1)	0,95	0,855-1,055	0,34
Localización Cólon Descendente	No	3749(4,2)	1		
	Si	181(4,5)	1,06	0,91-1,23	0,48
Localización Cólon Transverso	No	3715(4,2)	1		
	Si	215(5,4)	1,32	1,14-1,51	<0,0001
EPOC	No	3470(4,0)	1		
	Si	455(7,2)	1,9	1,68-2,06	<0,0001
C. ISQUEMICA	No	3711(4,1)	1		
	Si	192(7,4)	1,84	1,58-2,14	<0,0001
HTA	No	2392(4,4)	1		
	Si	1534(4,0)	0,9	0,84-0,96	0,002
Valvulopatías	No	3773(4,2)	1		
	Si	138(7,6)	1,89	1,59-2,26	<0,0001
Ictus	No	3781(4,1)	1		
	Si	130(10,8)	2,81	2,33-3,38	<0,0001
Insuf Renal	No	2807(3,2)	1		
	Si	1014(25,2)	10,29	9,50-11,16	<0,0001

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetro (1)
Tipo de ingreso	Urgente	20738	0,000
	Programado	71435	1,000
ACV	No	90979	0,000
	Si	1194	1,000
Insuf Renal	No	88170	0,000
	Si	4003	1,000
EPOC	No	85923	0,000
	Si	6250	1,000

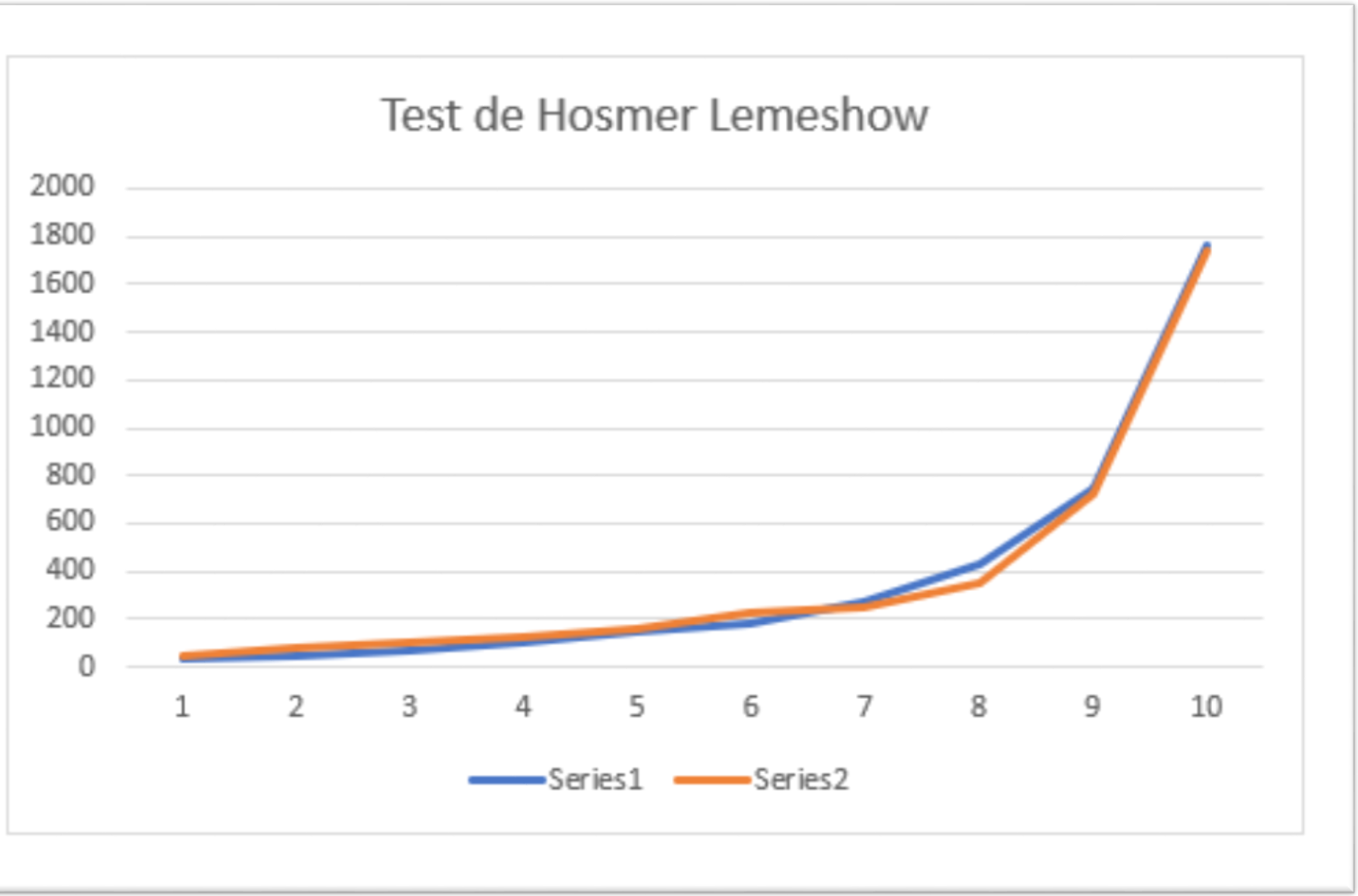
Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Edad	0,057	0,002	930,312	1	0,000	1,059	1,055	1,063
	EPOC(1)	0,360	0,056	40,650	1	0,000	1,434	1,283	1,602
	ACV(1)	0,628	0,103	36,906	1	0,000	1,874	1,530	2,294
	Insuf Renal(1)	1,983	0,045	1977,838	1	0,000	7,261	6,654	7,924
	Tipo de ingreso(1)	-1,543	0,036	1873,932	1	0,000	0,214	0,199	0,229
	Constante	-6,699	0,147	2087,579	1	0,000	0,001		

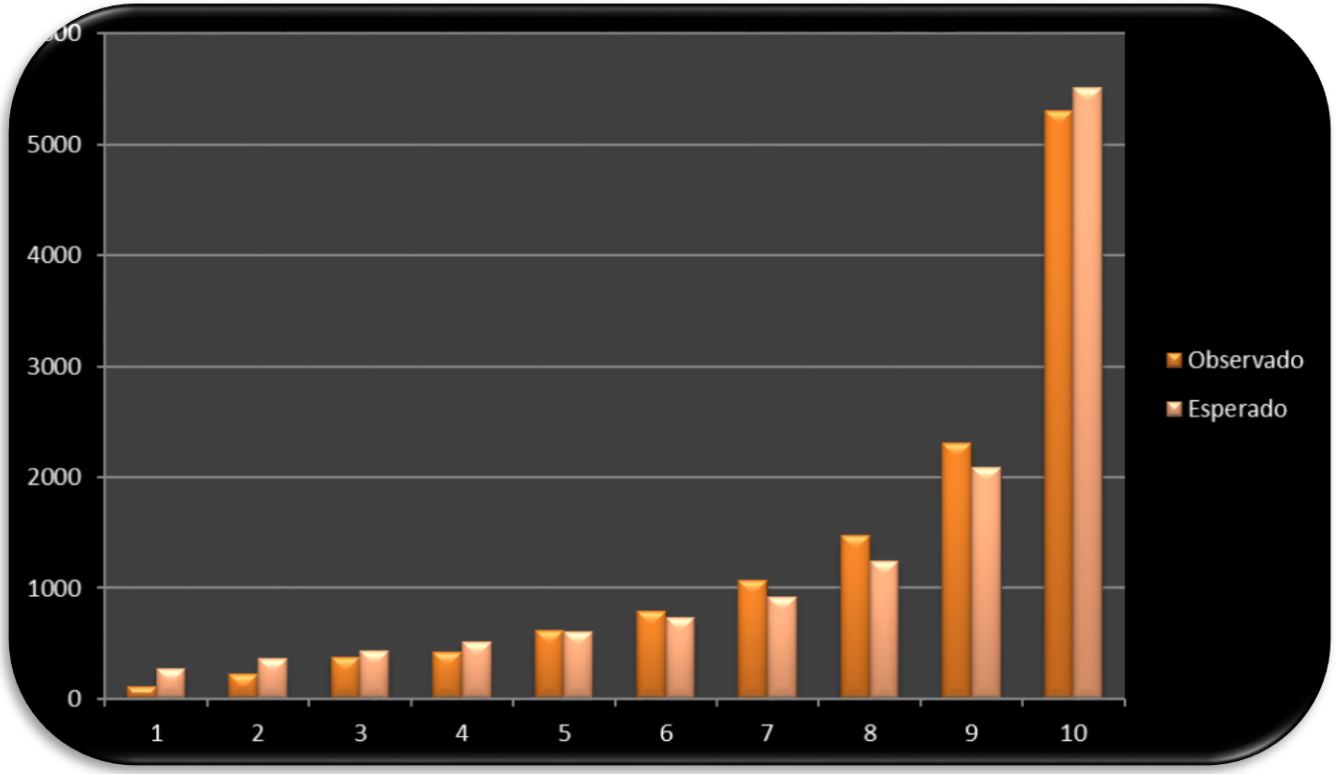
a. Variables especificadas en el paso 1: Edad, EPOC, ACV, Insuf Renal, Tipo de ingreso.

Tabla de contingencia para la prueba de Hosmer y Lemeshow

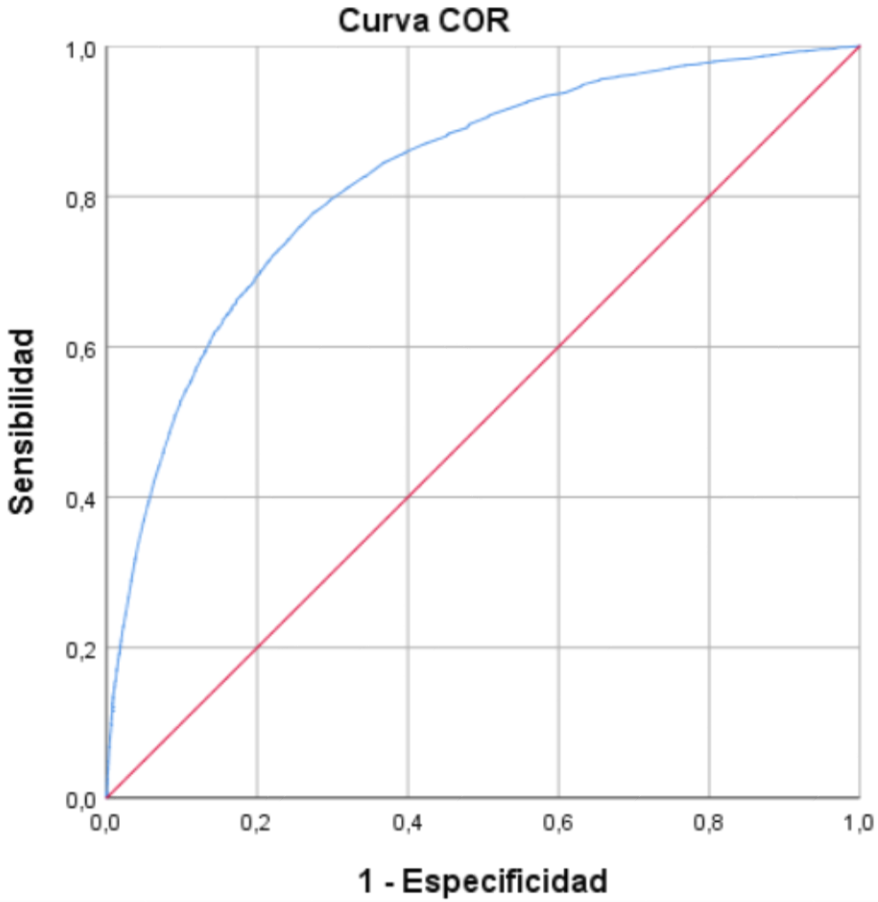
		Exitus = No		Exitus = Si		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	9147	9143,381	38	41,619	9185
	2	9543	9517,958	51	76,042	9594
	3	9660	9617,192	64	106,808	9724
	4	8555	8533,590	101	122,410	8656
	5	8997	8981,431	147	162,569	9144
	6	9842	9805,995	186	222,005	10028
	7	8626	8646,491	267	246,509	8893
	8	8761	8839,697	431	352,303	9192
	9	8490	8508,079	747	728,921	9237
	10	6758	6785,186	1762	1734,814	8520


Test de Hosmer Lemeshow
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
Series1
Series2

Test de Hosmer Lemeshow



p<0,005



Área bajo la curva

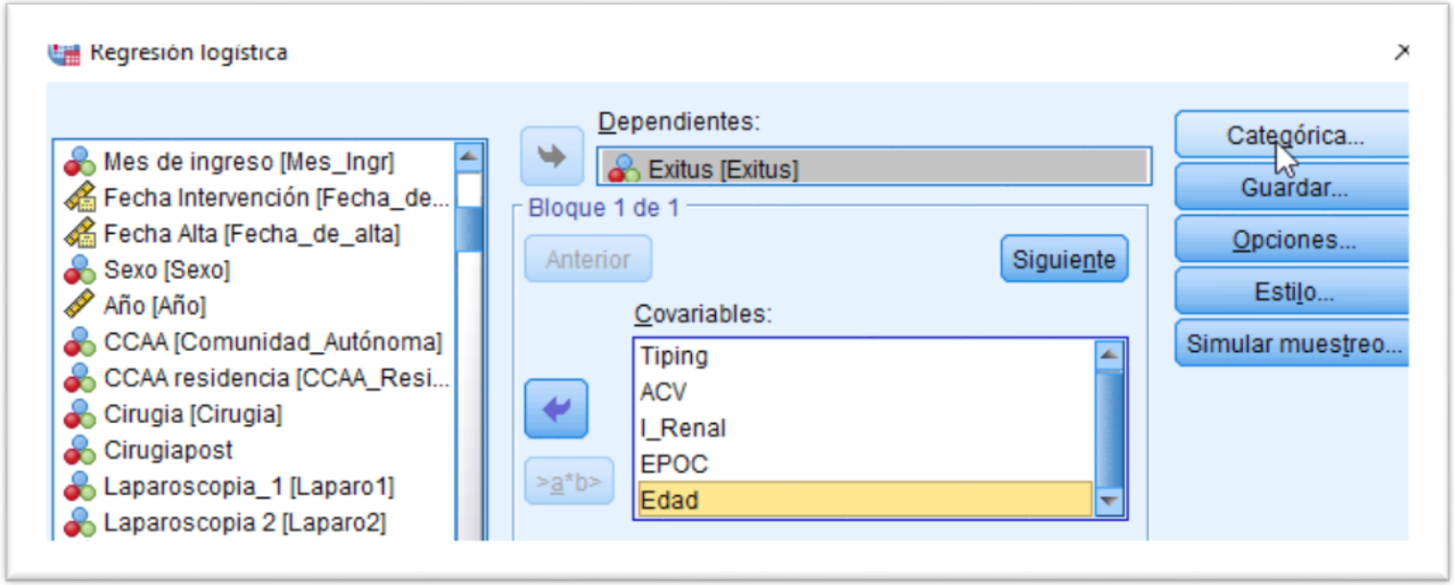
Variables de resultado de prueba:

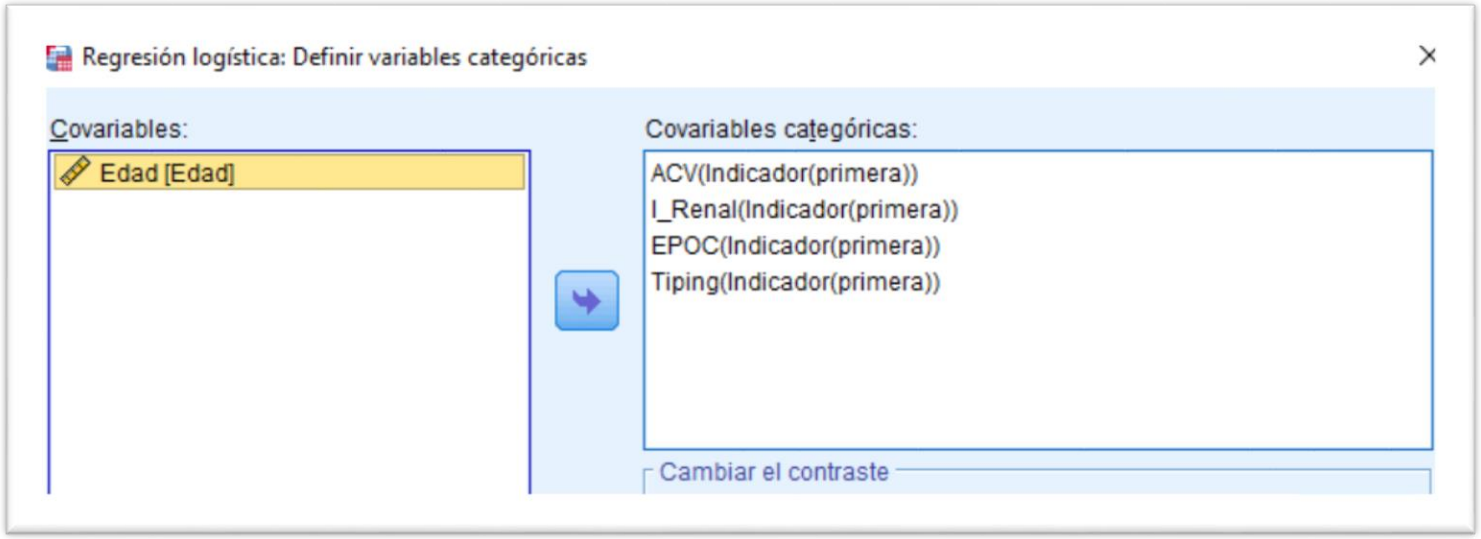
Área	Desv. Error[a]	Significación asintótica[b]	confianza asintótico Límite inferior	Límite superior
0,825	0,003	0,000	0,818	0,832

Las variables de resultado de prueba: Probabilidad pronosticada

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5





Guardar probabilidad. Ejecutar Tb Hosmer-Lemeshow

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Tipo de Ingreso(1)	-1,497	,015	9457,665	1	,000	,224
	ACV(1)	,320	,043	55,796	1	,000	1,377
	Insuf Renal(1)	1,200	,019	3857,824	1	,000	3,321
	EPOC(1)	,052	,024	4,526	1	,033	1,053
	Edad	,020	,001	1104,397	1	,000	1,020
	Constante	-3,116	,044	4944,599	1	,000	,044

a. Variables especificadas en el paso 1: Tipo de Ingreso, ACV, Insuf Renal, EPOC, Edad.

Constante (Beta 0)

Beta1, Beta2, Beta n…

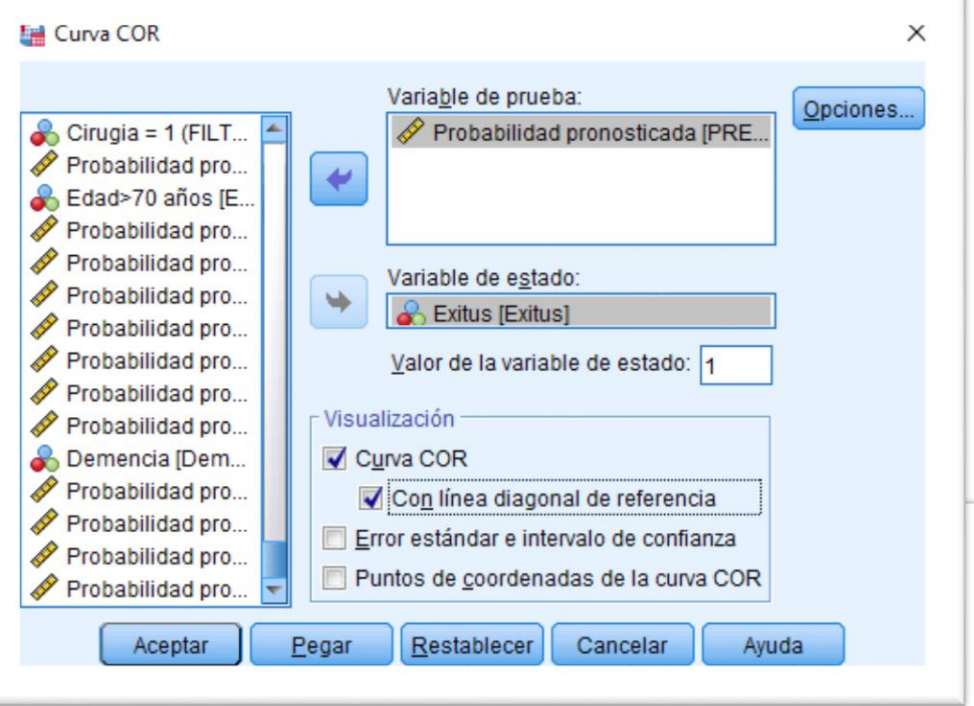
OR

Con los Coeficientes

Y* = -3,116 + (-1,497* Tiping) + (0,320*ACV) + (1,20*IR) + (0,52*EPOC) + (0,020*Edad)

Con las OR

Y* = -3,116 + 0,224* Tiping) + (1,377*ACV) + (3,321*IR) + (1,053*EPOC) + (1,020*Edad)

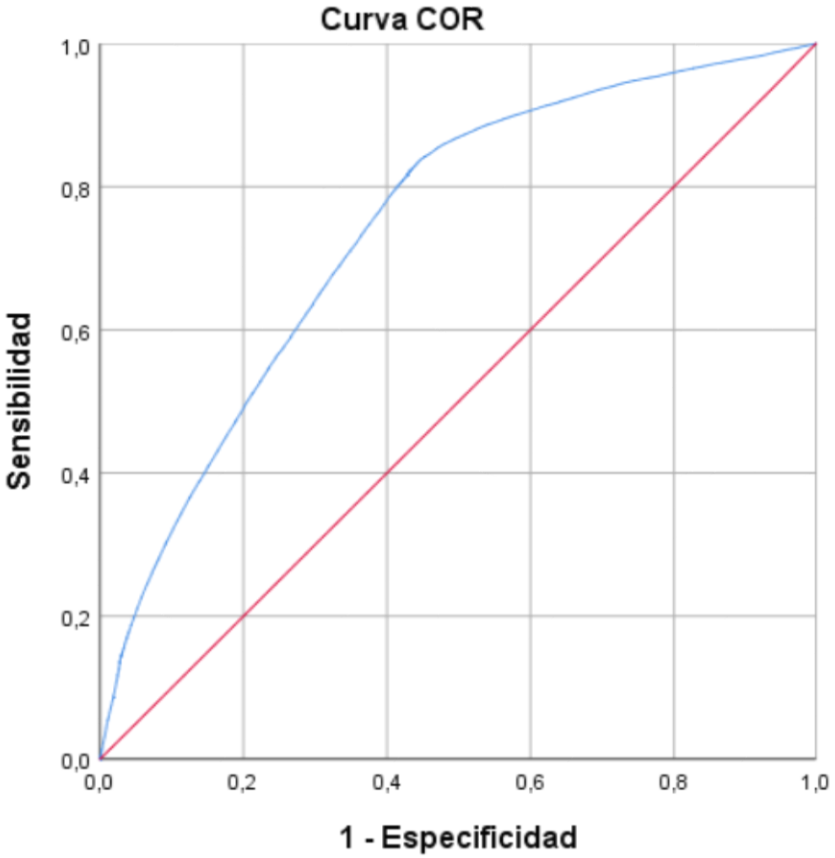


Área bajo la curva

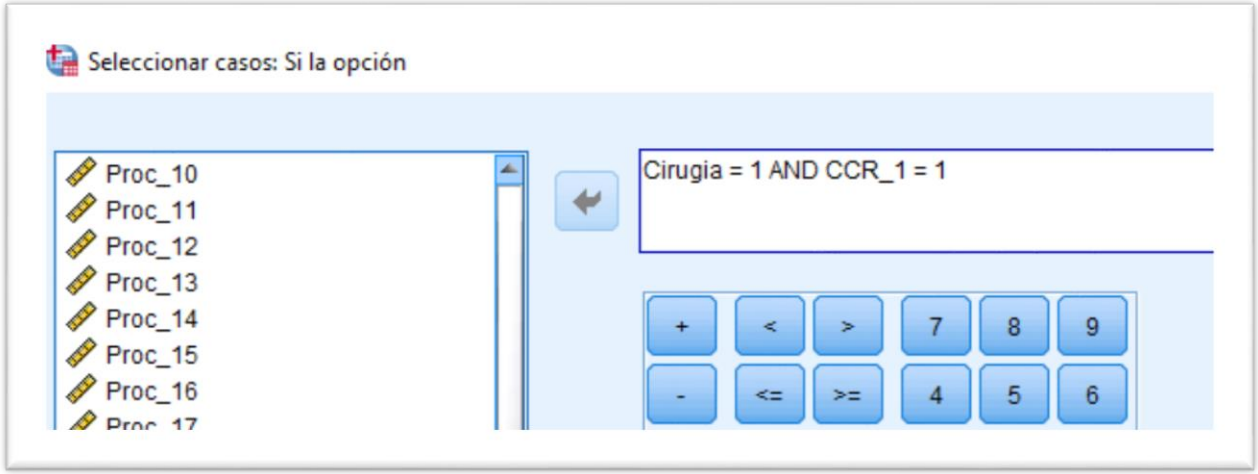
Variables de resultado de prueba

Área
,743

Las variables de



Seleccionamos: Sólo operados y Sólo CCR
en primera posición diagnóstica



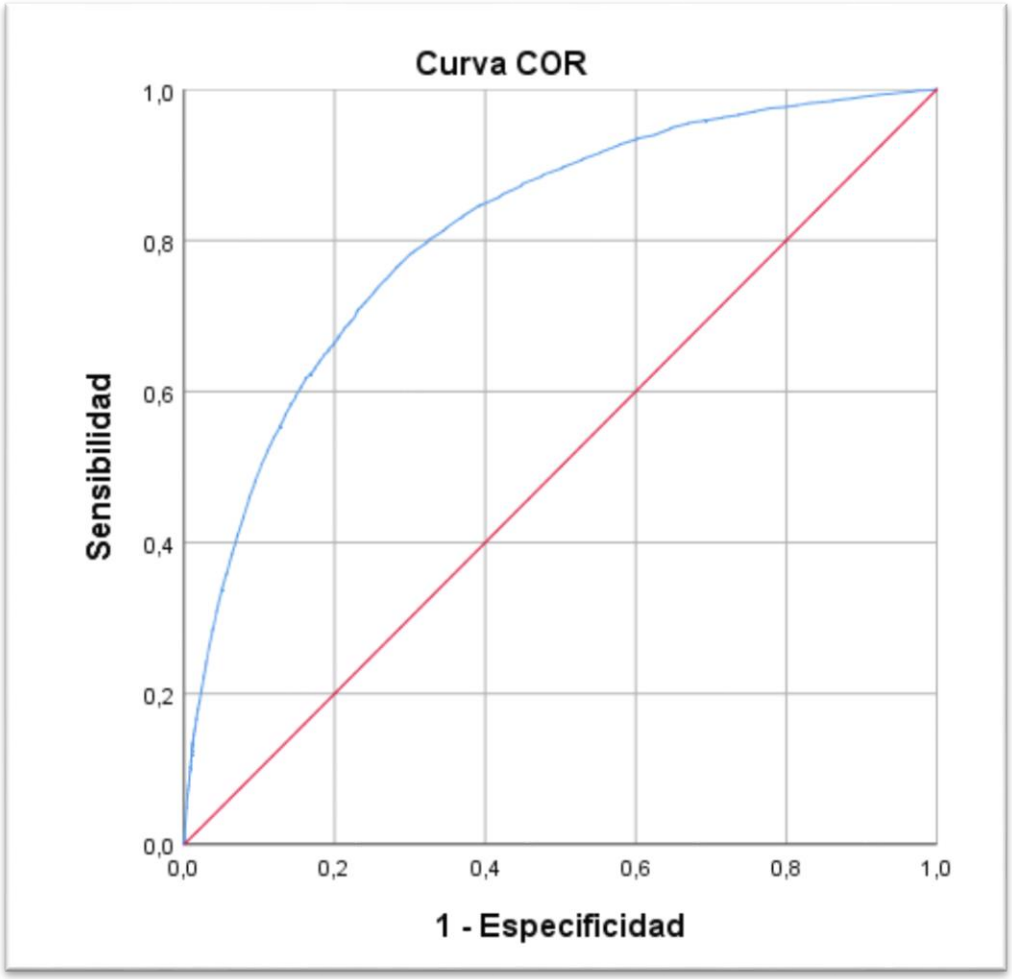
Nuevos valores de la OR en la RL

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Tipo de Ingreso(1)	-1,457	,029	2527,320	1	,000	,233
	ACV(1)	,508	,085	35,556	1	,000	1,662
	Insuf Renal(1)	1,870	,036	2704,548	1	,000	6,490
	EPOC(1)	,338	,046	53,774	1	,000	1,402
	Edad	,052	,002	1186,243	1	,000	1,053
	Constante	-6,229	,116	2867,837	1	,000	,002

a. Variables especificadas en el paso 1: Tipo de Ingreso, ACV, Insuf Renal, EPOC, Edad.

Nueva Curva ROC



Mejoría destacable en la capacidad discriminativa del modelo

Área bajo la curva

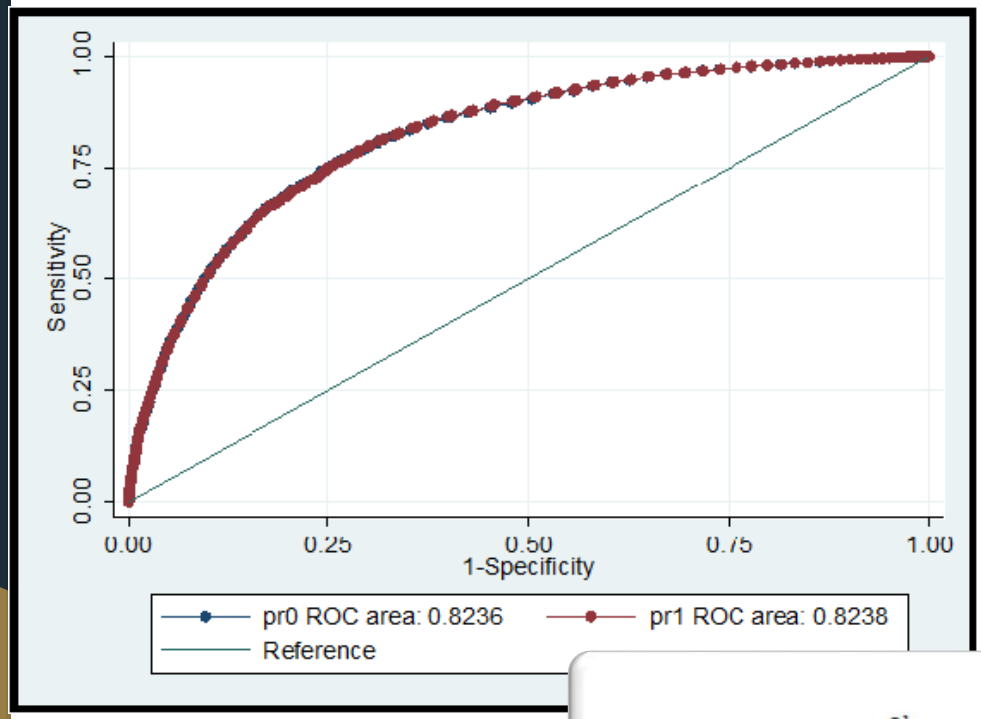
Variables de resultado de prue
Área
,813

Capacidad Discriminativa

- Representa todos los pares posibles de individuos formados por un individuo en el que ocurrió el evento y otro en el que no ocurrió. Indica la proporción en los que el modelo es capaz de señalar al individuo en el que ocurrió el evento con mayor probabilidad.

- **Capacidad de diferenciar a los individuos que padecerán un evento de los que no**

- AUC ó Estadístico C de Harrel: Se considera aceptable a partir de 0,70



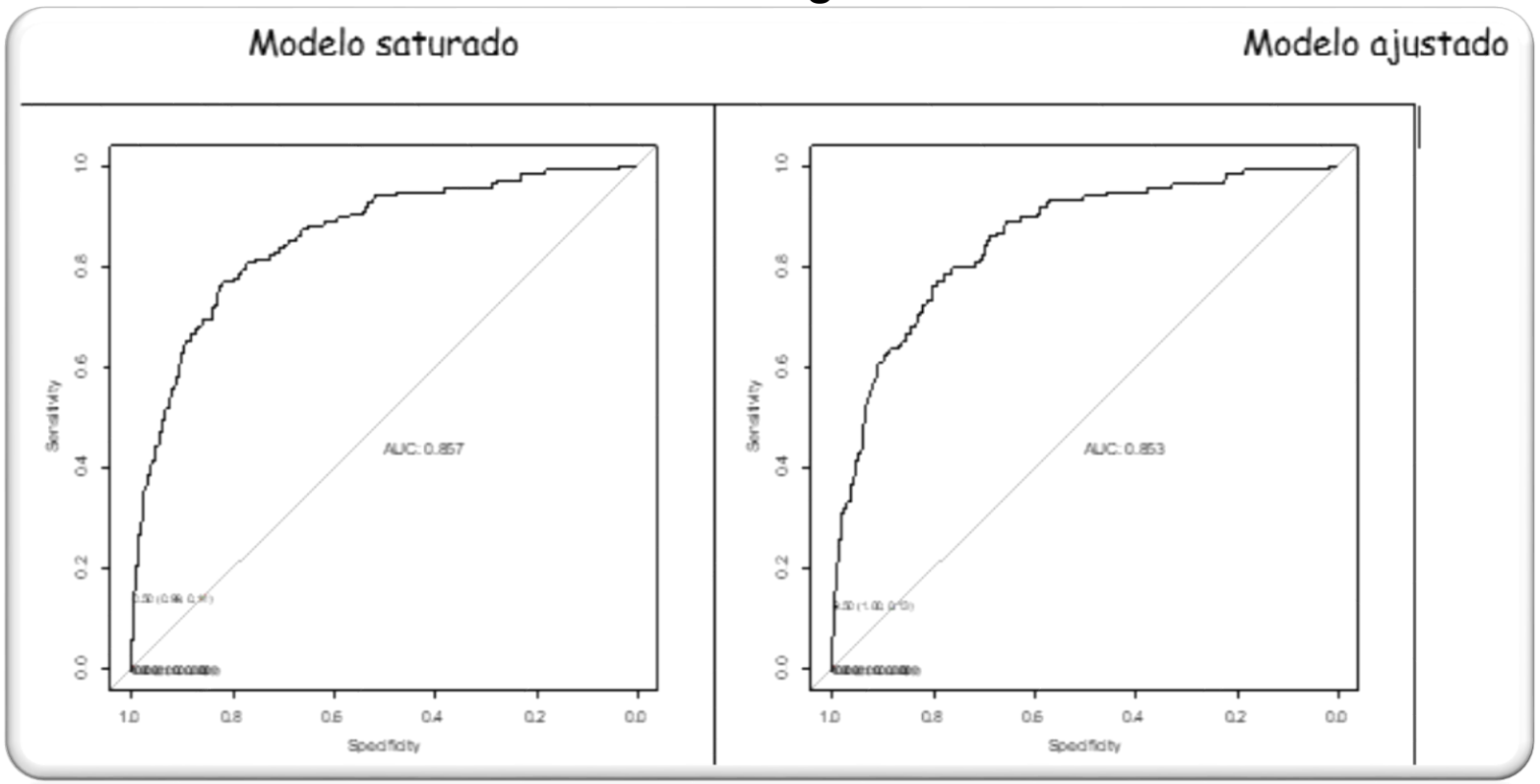
Discriminación del Modelo

- Curvas ROC
- Comparación entre Cohortes

	Obs	ROC Area	Std. Err.	-Asymptotic Normal- [95% Conf.	Interval]
pr0	102,685	0.8236	0.0030	0.81777	0.82947
pr1	102,685	0.8238	0.0030	0.81800	0.82965

Ho: area(pr0) = area(pr1)
chi2(1) = 1.60 Prob>chi2 = 0.2052

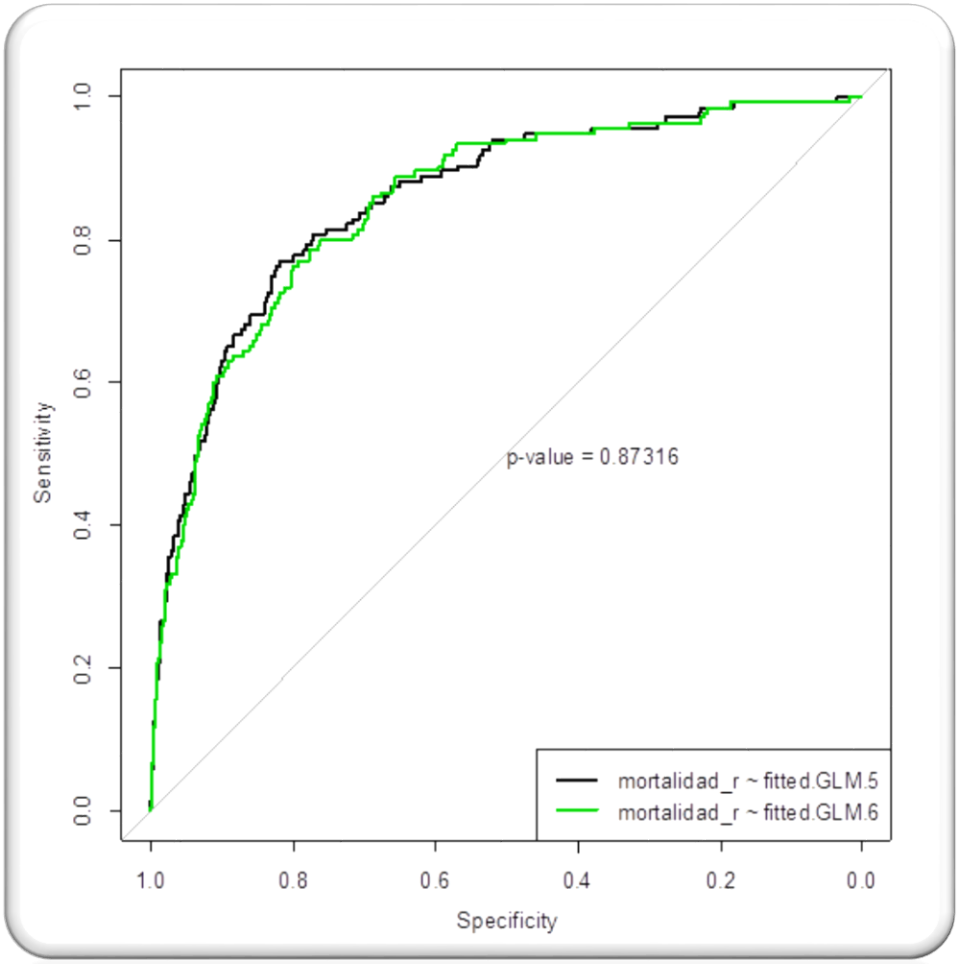
Modelos Saturados vs Ajustados
Modelo Training vs Test



Comparaciones entre Modelos

-Test de Hipótesis

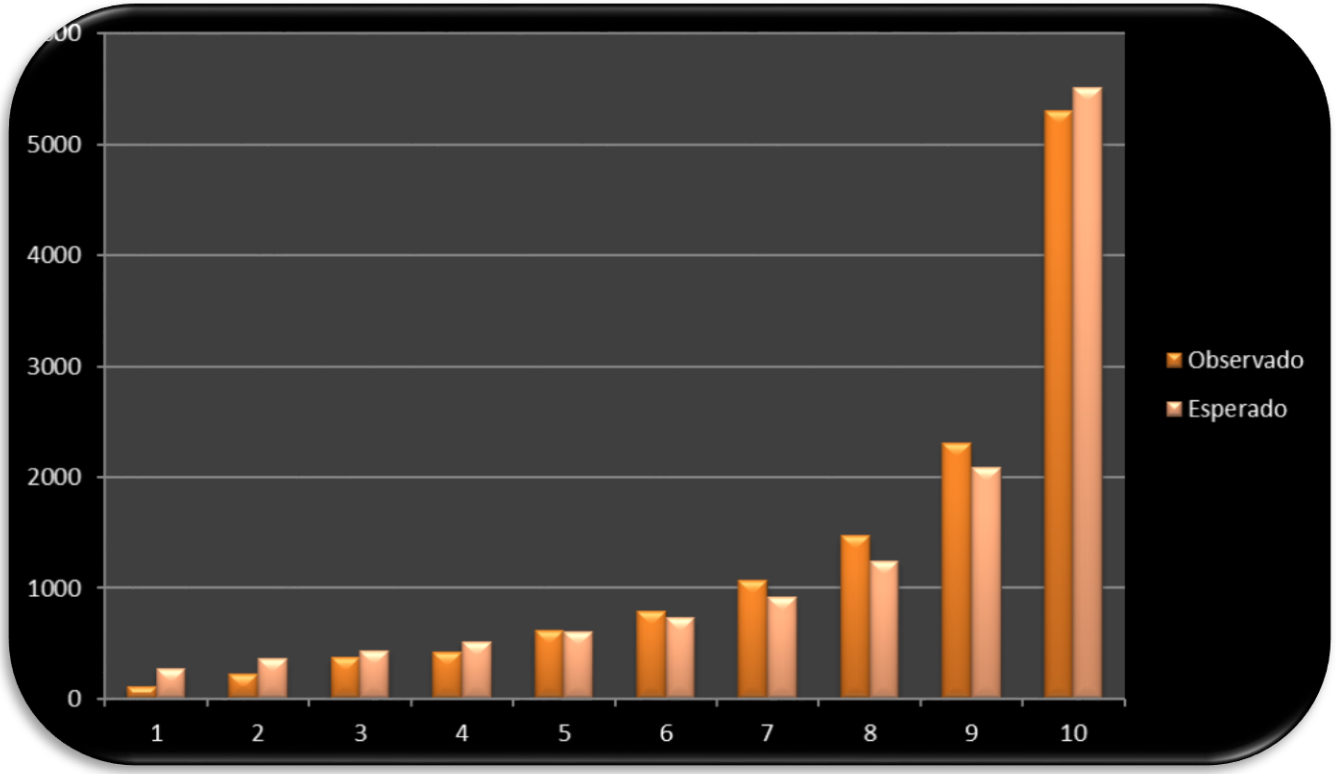
-Comparaciones ROC



Calibración del Modelo

- La prueba estadística es el test de Hosmer-Lemeshow
- Es la comparación entre las frecuencias observadas y las predichas por el modelo
- Se estructura en 10 deciles de riesgo
- Test NO significativo→ No hay diferencias observado/predicho

Test de Hosmer Lemeshow



p<0,005

MODELOS PREDICTIVOS EN CCR

¿Cuándo un Modelo Predictivo tiene “éxito”?

- Responde a un **escenario de incertidumbre** clínica
- Metodología de **validación** adecuada
- Los líderes clínicos deben **aceptarlos y aplicarlos**

¿Cuáles son las utilidades generales de los Modelos Predictivos?

- Diseño y análisis de ensayos clínicos y terapéuticos
- Investigar la variabilidad de los servicios de salud (Jarman et al)
- Ajuste por Case-Mix
- Ayuda a clínicos y pacientes en la toma de decisiones

Pronóstico y pasos en su investigación

P= estimación de la (P) ó Riesgo de ocurrencia de un determinado evento

P(Médico)= Riesgo de presentar un determinado estado de salud o de enfermar (+ frec)

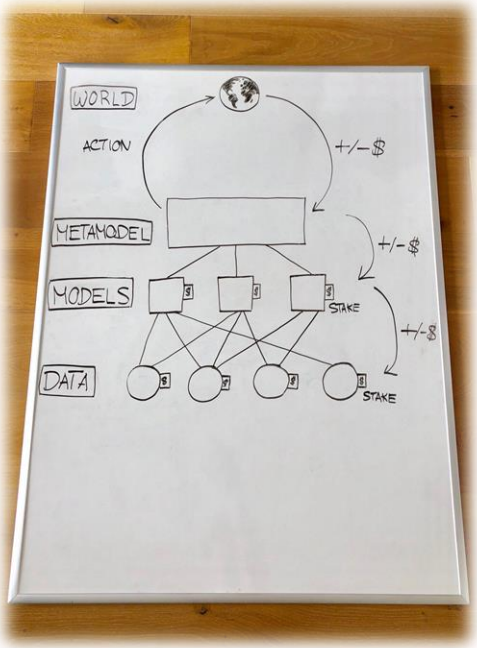
Los **Scores Pronósticos** NO SIEMPRE están ligados al riesgo de enfermar (Apgar, Fine, etc…)

Esquema de Moons: Los Modelos Pronósticos han de pasar por 3 fases

A. Estudios de Desarrollo
B. Estudios de Validación
C. Estudios de Impacto

Estudios de Desarrollo

- Estudios de los que se derivan los modelos multivariantes para la construcción del modelo
- Conformación: 1 VD y n V. Predictoras/Independientes. Factores-Pesos (Coeficientes)
- Se evalúa la capacidad discriminativa (C Statistic, AUC) y Calibración (Hosmer-Lemeshow)
- Contiene la validación interna (muestra training vs muestra test)
- Riesgo de esta etapa: Sobreajuste



Estudios de Validación

Validación aparente

Validación interna (estudio de reproductibilidad)

- Split Sampling o partición de la muestra
- Cross Validation
- Bootstraping (remuestreo)

Validación Externa (Cantalejo,2005)

- =Generabilidad
- =Comportamiento en una población “parecida” a la de la que proviene la muestra
- Temporal: Se prueba el modelo en enfermos más recientes
- Geográfica: Se prueba en otro hospital, área, etc..
- Espectro: Con pacientes más graves

Estudios de Impacto: Cuantifican si el uso del Modelo Pronóstico obtenido realmente consigue mejorar el output buscado

INNOVATION IN PREDICTIVE MODELS: SMR METHODOLOGY

BMC Health Services Research

BioMed Central

Research article

Open Access

Standardisation of rates using logistic regression: a comparison with the direct method

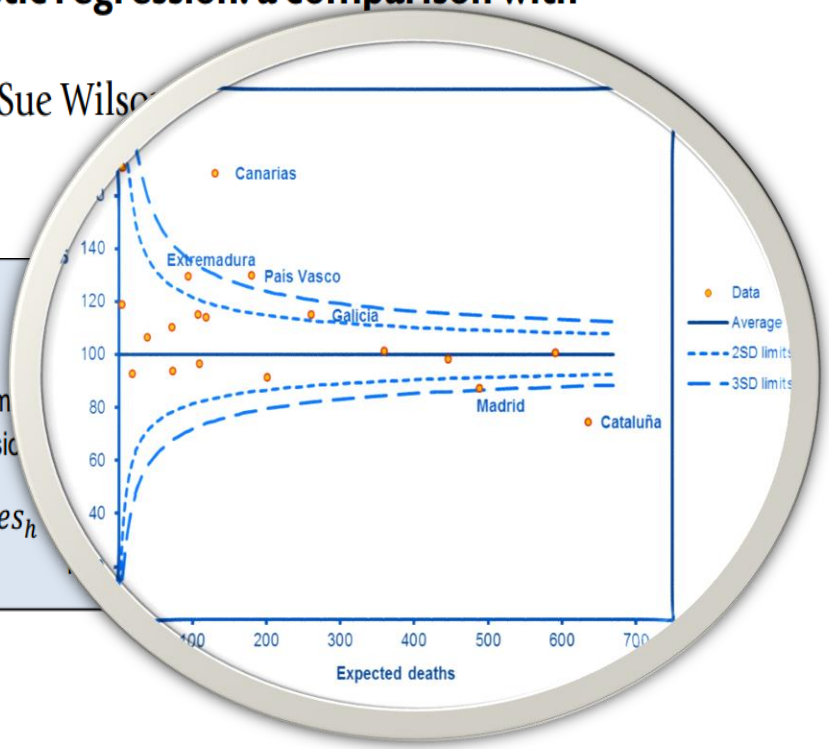
Andrea K Roalfe*, Roger L Holder and Sue Wilso

$$HSMR_h = \frac{Observed\ Deaths_h}{Predicted\ Deaths_h}$$ where,

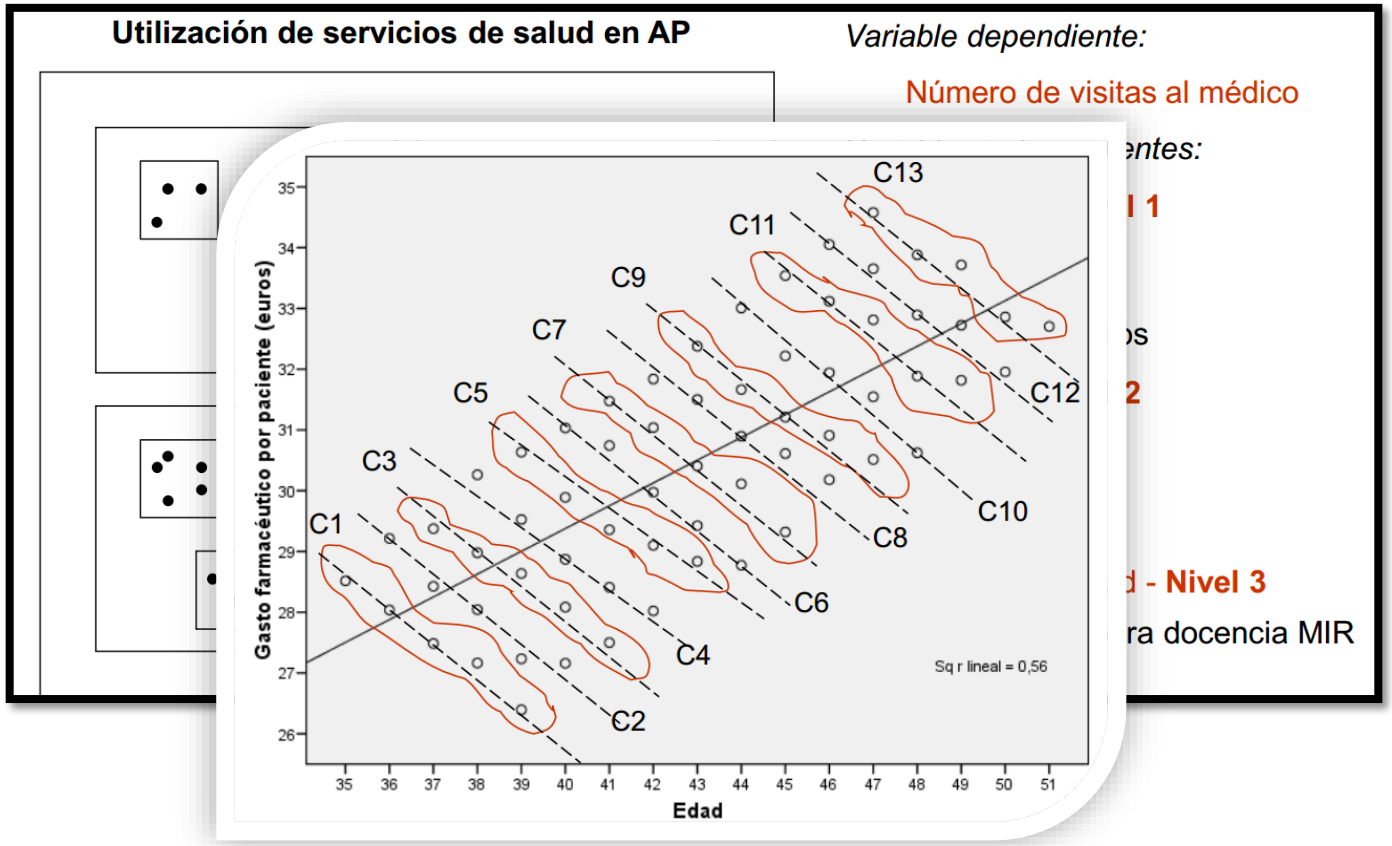
$Observed\ Deaths_h = \sum_j Numerator_h$ the sum

admissio

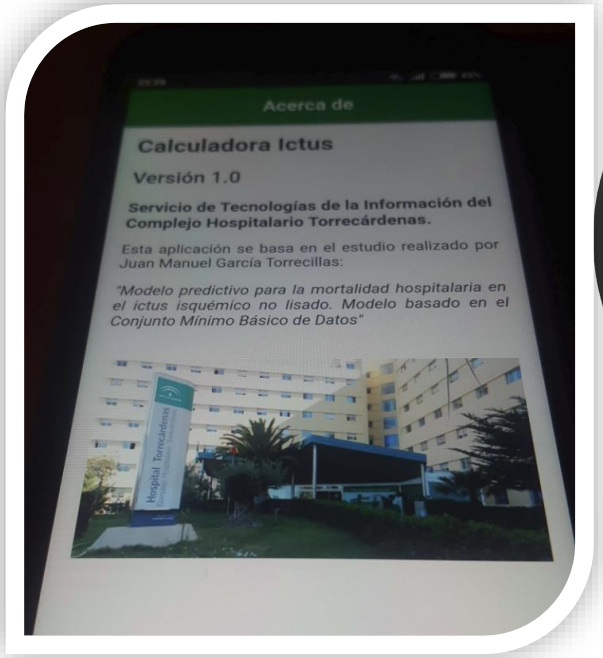
$Predicted\ Deaths_h = \sum_j Predicted\ Probabilities_h$

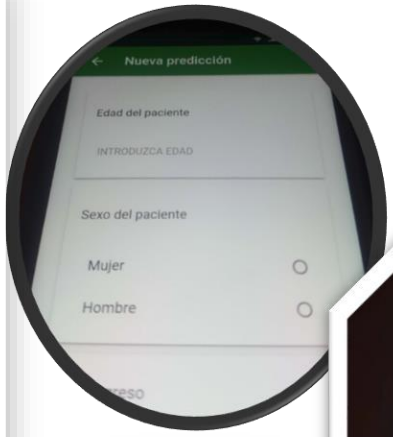


MORE INNOVATION METHODOLOGY. MULTILEVEL ANALISYS



Fuente: Ocaña Arriola, 2016





APPs for Ictus
- Quick stimation of risk
- Agile
- Easy



Regresión Logística
Aplicación a Modelos Predictivos en CCR

MUCHAS GRACIAS

Grupo de Investigación PI16/01931
Financiado Fondos FEDER/ISCIII
Proyecto Periodo 2017-2019



Muchas Gracias
por vuestra atención