Taller 3: Modelo logístico

Andrés Felipe Palomino - David Stiven Rojas
Códigos: 1922297 - 1924615
Universidad del Valle

2 de junio de 2023

Universidad del Valle

Ejercicio 1: La base de datos AbusoDrogas.txt contiene información de 575 pacientes que fueron sometidos a un tratamiento para la drogadicción. El objetivo del estudio es determinar si la eficiencia de un programa para reducir el consumo de drogas depende de la duración del mismo (larga y corta duración), además de otras variables observadas. Luego del tratamiento, se observa si, luego de un año, el paciente sufrió alguna recaída en el abuso de drogas (variable respuesta).

Variables: **DFREE**(y_i)**, AGE**(x_1)**, BECK**(x_2)**, NDRUGTX**(x_3)**, TREAT**(x_4).

1. Ajuste un modelo logístico para la variable DFREE y las otras variables como regresoras. Exprese el modelo de forma clara e interprete los coeficientes estimados. ¿estos resultados sugieren que un tipo de tratamiento es más efectivo que el otro?

Sea el evento: A= {El paciente es libre de drogas por un año}. Con una probabilidad de éxito para A $\pi_i \in (0,1)$

$$y_i = \begin{cases} 1 \quad si \quad \text{ocurre A} \\ \\ 0 \quad si \quad \text{ocurre A'} \end{cases}$$

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9615	0.5732	-3.42	0.0006
AGE	0.0332	0.0161	2.07	0.0386
BECK	-0.0051	0.0105	-0.48	0.6292
NDRUGTX	-0.0848	0.0259	-3.28	0.0010
TREAT	0.4563	0.1959	2.33	0.0199

Tabla 1: Resumen del modelo logístico.

Adicional a el resumen presentado en la anterior tabla contamos con una devianza de 632.2 y un AIC de 642.21. En general encontramos en la Tabla 1 que la covariable BECK que hace enfásis al índice de depresión de Beck no presenta un aporte significativo dentro del modelo si ya hemos incluido el otro conjunto de variables, la interpretación de estos β_i se realizará a partir del Ods.

Covariable	β_i	Probabilidad
AGE	0.03	1.03
BECK	-0.01	0.99
NDRUGTX	-0.08	0.92
TREAT	0.46	1.58

Tabla 2: Plausibilidad generada por cada β_i

En la Tabla 2 observamos que por cada año que aumente la edad del paciente la probabilidad de que su condición de estar libre de drogas por un año aumenta en un 3 % si dejamos las demas covariables constantes, por cada valor que aumente el índice de Beck es decir el paciente tenga mayor depresión la probabilidad de que este libre de drogas por un año disminuirá en un 1 %.

Por cada número de tratamientos anteriores contra las drogas que se tengan la probabilidad de estar libre disminuirá en un 8 % y por último podemos observar que el tratamiento si presenta una mejora considerable en la probabilidad del evento A puesto que el estar en este aumenta en un 58 % la probabilidad, claramente cada una de estas interpretaciones es con las demás covariables constantes.

2. Ahora considere en el modelo las interacciones entre TREAT y las otras covariables. A partir de una prueba de hipótesis, ¿el aporte de estas interacciones es significativo? En caso de que lo sean, interprete los coeficientes asociados.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0528	0.8676	-1.21	0.2249
AGE	0.0082	0.0249	0.33	0.7403
TREAT	-1.0993	1.1371	-0.97	0.3336
BECK	-0.0029	0.0160	-0.18	0.8551
NDRUGTX	-0.1222	0.0465	-2.63	0.0086
AGE:TREAT	0.0441	0.0328	1.34	0.1789
TREAT:BECK	-0.0046	0.0213	-0.21	0.8303
TREAT:NDRUGTX	0.0559	0.0556	1.01	0.3148

Tabla 3: Resumen del modelo con interacciones

En primera instancia podemos observar que los valores p asociados a los valores z cuando se considera la interacción nos describen aporte no significativos. Al realizar la siguiente prueba de hipótesis con la prueba razón de verosimilitudes:

$$H_0 : \beta_5 = \beta_6 = \beta 7 = 0$$

$$H_1 : \beta_i \neq \beta j$$

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	570	632.21			
2	567	628.46	3	3.75	0.2901

Tabla 4: Resumen de prueba de hipotésis razón de verosimilitudes

Obtenemos un valor p asociado al estadístico chi cuadrado correspondiente a 0.2901 lo que nos indica que todos estos coeficientes son significativamente iguales a 0 por lo cual el efecto de la interacción del tratamiento con las otras covariables no presenta un aporte significativo, por lo cual no realizaremos su interpretación.
3. Considere diferentes funciones de enlace, ¿cuál proporciona mejor ajuste?

	Logit				Probit				Cloglog			
β_i	Est	Std	Z	P-Val	Est	Std	Z	P-Val	Est	Std	Z	P-Val
(Intercept)	-1.96	0.57	-3.42	0.00	-1.19	0.34	-3.51	0.00	-1.97	0.49	-4.04	0.00
AGE	0.03	0.02	2.07	0.04	0.02	0.01	2.04	0.04	0.03	0.01	2.07	0.04
BECK	-0.01	0.01	-0.48	0.63	-0.00	0.01	-0.50	0.62	-0.00	0.01	-0.44	0.66
NDRUGTX	-0.08	0.03	-3.28	0.00	-0.05	0.01	-3.29	0.00	-0.07	0.02	-3.23	0.00
TREAT	0.46	0.20	2.33	0.02	0.27	0.12	2.35	0.02	0.39	0.17	2.31	0.02

Tabla 5: Resumen del ajuste del modelo con diferentes enlaces

Enlace/ Criterio	Logit	Probit	CLogLog
AIC	642.21	642.70	642.21
Devianza	632.21	632.70	632.21

Tabla 6: Criterios de información y bondad de ajuste

En la Tabla 5 y 6 observamos que los valores p y las estimaciones puntuales de cada β_i son muy similares, independiente del tipo de enlace utilizado, ideas que son corroboradas también por la Devianza y AIC.

Si buscáramos observar el poder predictivo del modelo, evidenciamos que los valores de las métricas de desempeño, Sensibilidad y Especificidad para diferentes de puntos cohorte, casi que se sobreponen en la curva de ROC, como se evidencia en la Figura 1.

En la Tabla 7 observamos diferencias casi que nulas en las métricas de desempeño junto con el punto de corte y AUC, este primero fue escogido utilizando un criterio que maximiza la probabilidad bajo curva del AUC, puesto que no teníamos se contaba con el apoyo de un experto en el tema para declinarnos por un modelo más especifico o sensible.

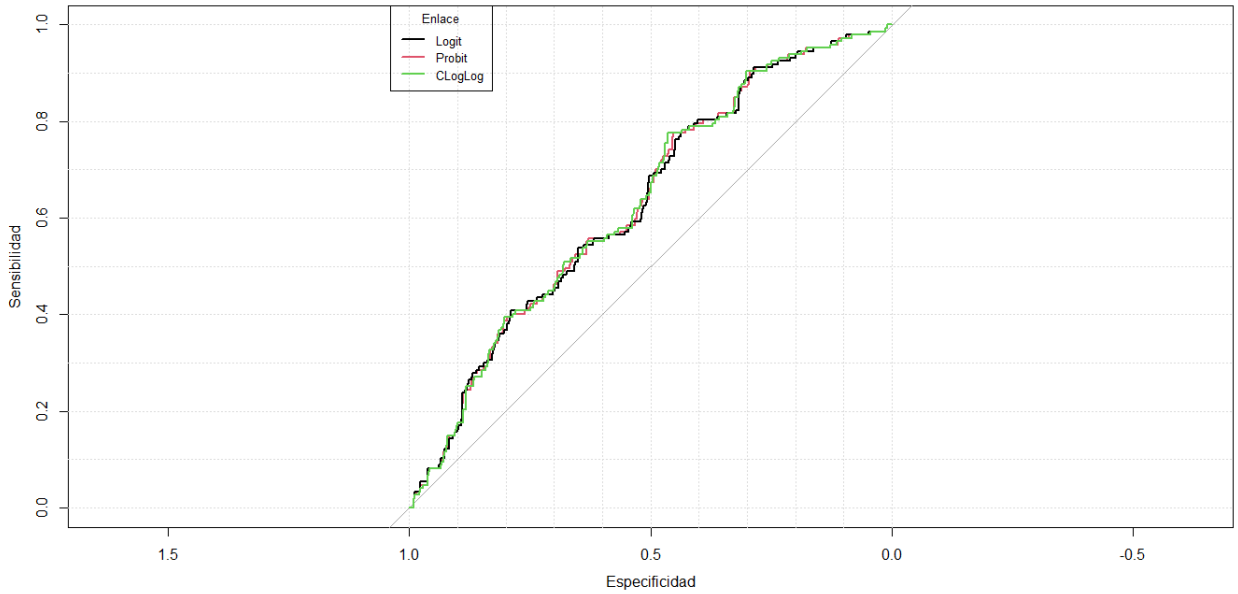

Figura 1: Curvas de ROC para los diferentes enlaces

Párametro	Logit	Probit	CLogLog
AUC	0.63	0.61	0.63
Sensibilidad	0.45	0.43	0.46
Especificidad	0.78	0.78	0.78
Punto de corte	0.23	0.23	0.23

Tabla 7: Métricas de desempeño del modelo

En general, observamos que los diferentes tipos de enlaces no presentan mejoras en ningún aspecto dentro del modelo y, por el contrario, presentamos desempeños casi que idénticos en todos los aspectos posibles, por lo cual nos declinamos por el modelo con enlace Logit, puesto que este tiene una interpretabilidad más compacta de sus β_i a través del uso Ods y en general por la parsimonia del modelo, dónde siempre buscaremos modelos menos complejos de realizar.

	ϕ
Logit	1.11
Probit	1.11
CLogLog	1.11

Tabla 7: Indicador de sobre dispersión del modelo

Por último, utilizamos los residuos de Pearson para evaluar si existe evidencia de sobre dispersión en el modelo y notamos que no es así para ningún enlace.