Unidad 10: Regresión Logística

Regresión Logística Binaria

Supongamos que usted es asesor del equipo de campaña electoral del candidato a la presidencia, Julio Guzmán y quiere saber cuáles van a ser las posibilidades de ganar del candidato de acuerdo a algunas características que usted, después de haber realizado una revisión de la literatura sobre comportamiento electoral y elecciones, cree que influyen en la elección del votante sobre el candidato. En este primer momento, nos damos cuenta que tenemos dos posibilidades que queremos calcular, la probabilidad que alguien vote por el candidato o que no, esto es conocido como probabilidad de éxito y probabilidad de fracaso.

¿Existe alguna herramienta estadística que esté diseñada para responder a esta interrogante? Su variable dependiente cuenta con dos categorías de respuesta: es la variable dicotómica, por lo que ahora trabajaremos con una *regresión logística binaria*. Para este modelo es necesario que la variable dependiente tenga dos características: que sea e*xclusivas* (si cumplen una característica no cumplen la otra) y e*xhaustivas* (estas dos posibilidades son las únicas posibles).

A continuación, trabajaremos con una base de datos (simulada) sobre la votación de Julio Guzmán. Esta data contiene una variable dicotómica, que se llama (voto), cuyas categorías son 1 y 0, 1 es a favor del candidato y 0 en contra. En ese sentido, mediremos el efecto de tres variables: *Ingresos*, *Hijos* y *Edad*.

```
install.packages("VGAM") #instalando el paquete

library("VGAM") #llamando a la librería

data <- read.csv("datavotguzman.csv", sep=";", header=TRUE) #cargando la data

modelo <- glm (voto ~ Ingresos + Hijos + Edad , family ="binomial", data=data)  #creando el objeto modelo para nuestra variable binomial

summary(modelo)
```

Call:

glm(formula = voto ~ Ingresos + Hijos + Edad, family = "binomial", data = data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.52973	-0.19666	0.04036	0.24529	0.97255

Coefficients:

	Estimate	Std. Error	z value	**Pr(>|z|)**
(Intercept)	-10.360116	4.159874	-2.490	**0.0128** *
Ingresos	0.003488	0.001582	2.205	**0.0275** *

Hijos	-0.048122	0.428916	-0.112	**0.9107**
Edad	0.050122	0.059329	0.845	**0.3982**

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 41.054 on 29 degrees of freedom

Residual deviance: 10.783 on 26 degrees of freedom

AIC: 18.783

Number of Fisher Scoring iterations: 7

Luego de pedir el modelo, tenemos que observar la relación entre las variables independientes y la dependiente, para definir si se utilizará definitivamente o si se descartará**.** Como ya lo vimos en la regresión lineal, tenemos que ver la significancia del Pr(>|z|) de cada variable predictora, además de observar la referencia de los asteriscos (*). En ese sentido, observamos que la única variable que tiene efecto para el modelo es la Ingresos; mientras que las variables Hijos y Edad no son influyentes. Por lo que el modelo debe plantearse de la siguiente manera:

voto= -10.360116 + 0.003488 (Ingresos)

Finalmente, podemos decir que cuando la variable ingresos aumenta, también crece la probabilidad de votar por Julio Guzmán, dada la relación directa que existe. Como hemos podido observar, esta es una forma de realizar los modelos con variables categóricas.

Regresión Logística Multinomial

Si queremos modelar la ubicación ideológica de las personas, de acuerdo a algunas características que podamos recoger. Sabiendo que la ubicación ideológica, es un concepto ambiguo y difícil de medir, optaremos por categorizarla en tres opciones: *derecha*, *centro* e *izquierda*. Realizar esta medición involucra que trabajemos con una *regresión logística multinomial,* esta se caracteriza por tener en su variable independiente a las variables categóricas y se formula, a partir de observar la probabilidad de que se pueda elegir una categoría respecto a otra que es de referencia, procedimiento similar al de una regresión logística binomial. En este caso, se puede elegir la categoría de referencia izquierda, y se compara con las otras dos (derecho y centro). Entonces los resultados serán la probabilidad de encontrar en derecho respecto a encontrarse en izquierda y la probabilidad de encontrarse en centro respecto a izquierda.

```
install.packages(“VGAM”) #instalando el paquete

library(“VGAM”) #llamando a la librería

data <- read.csv("datavotguzman.csv", sep=";", header=TRUE) #cargando la data
```

```
modelo <- glm (Educa ~ Ingre+ Edad , family ="multinomial", data=data)  #creando el objeto modelo para nuestra variable multinomial

summary(modelo)
```

Call:

vglm(formula = Educa ~ Ingre + Edad, family = multinomial, data = data)

Pearson residuals:

	Min	1Q	Median	3Q	Max
log(mu[,1]/mu[,4])	-2.084	-0.5821	-0.3061	0.8753	1.921
log(mu[,2]/mu[,4])	-1.613	-0.4999	-0.1927	0.3938	3.319
log(mu[,3]/mu[,4])	-1.406	-0.5243	-0.2180	-0.1459	2.275

Coefficients:

	Estimate	Std. Error	z value	**Pr(>|z|)**	
(Intercept):1	4.942e+00	2.790e+00	**1.771**	**0.0765**	.
(Intercept):2	7.441e+00	2.993e+00	**2.486**	**0.0129**	*
(Intercept):3	3.099e+00	2.879e+00	**1.076**	**0.2818**	
Ingre:1	-8.967e-04	6.646e-04	**-1.349**	**0.1772**	
Ingre:2	-7.351e-04	6.858e-04	**-1.072**	**0.2837**	
Ingre:3	4.092e-05	6.666e-04	**0.061**	**0.9511**	
Edad:1	-4.018e-02	4.773e-02	**-0.842**	**0.3999**	
Edad:2	-1.153e-01	5.378e-02	**-2.143**	**0.0321**	*
Edad:3	-5.985e-02	4.635e-02	**-1.291**	**0.1966**	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors: 3

Names of linear predictors: log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]), log(mu[,3]/mu[,4])

Dispersion Parameter for multinomial family: 1

Residual deviance: 68.9402 on 81 degrees of freedom

Log-likelihood: -34.4701 on 81 degrees of freedom

Number of iterations: 5

De los resultados que hemos obtenido centraremos la atención en las columnas *z value* y *p value*. Así verificamos que de las tres variables sólo tiene efecto la variable *Edad*, y en una sola categoría. Así, *Edad2* favorece a la categoría dos que es centro, eso quiero decir que es directa en relación de comparación o inversa en relación a la modalidad de referencia. De esta manera funcional el Modelamiento Logístico Multinomial.

Formas de Selección de variables

FORWARD (HACIA ADELANTE)

Se empieza solo con el intercepto y se van agregando una a una las ”principales”variables predictoras (de ser estas significativas) hasta alcanzar un conjuntó optimo .

BACKWARD (HACIA ATRÁS)

Se empieza con todas las potenciales variables predictoras y se van eliminando una a una las ”menos importantes”(de ser estas no significativas) hasta quedarse con un conjunto óptimo.

Ejercicios

En la base de datos “datavotPPK.csv”, encontrarás una data simulada de 173 casos que serán evaluadas en 5 variables, una primera variable dicotómica que observa la preferencia o no preferencia por PPK; la segunda variable observa los ingresos de las personas; una tercera variable categórica que mide el nivel de estudios entre 1 y 5; la cuarta variable refiere a la zona en la que vive y, finalmente, una quinta variable que nos hace referencia a la Edad de las personas.

1. A continuación, tendrás que realizar una regresión binomial que tendrá como variable dependiente el voto por PPK. Para ello utilizaras, en principio, como variables independientes el nivel de ingresos, nivel de estudios, la zona de residencia y la edad. Después, podrás proceder con el procedimiento de FORWARD para como todos los modelos y las variables explicativas.

2. Como segundo ejercicio, tendrás que elaborar una regresión multinomial con la variable Zona de residencia como dependiente; además de realizar procedimiento de BACKWARD con todas las variables para comparar el AIC de cada regresión.