Matemática: Una publicación del ICM – ESPOL
2017, Vol.15, No. 1 **Publicado: 2017-04-03**

MODELO DE REGRESIÓN LOGÍSTICA APLICADO A LA PROBABILIDAD DE FALLECIMIENTO DE LOS HOMBRES DE LA COMUNIDAD VALENCIANA

LOGISTIC REGRESSION MODEL APPLIED TO THE PROBABILITY OF DEATH OF THE MEN OF THE VALENCIAN COMMUNITY

Plata Wendy[1], Yambay Karen[2]

Resumen: *En este artículo se presenta una aplicación de los modelos de regresión logística a fin de explicar la probabilidad de fallecimiento en función de la edad de los hombres de la Comunidad Valenciana en el año 2000. Para el efecto, en primer lugar se desarrollaron los principios teóricos que conforman la base de este estudio, para luego con el soporte del software estadístico R, efectuar el ajuste de los datos a una curva logística, para en base a evidencia estadística determinar la significancia de los coeficientes del modelo y calcular la razón de posibilidades o chances. Los principales resultados obtenidos han sido que un Modelo de Regresión Logística ajustó la probabilidad de fallecimiento de los Hombres de la Comunidad Valenciana para el año 2000, siendo la edad un factor influyente al momento de calcular la probabilidad de fallecer.*
Palabras claves: *R, regresión logística, modelos de regresión, modelos no lineales, residuos, leverages, binomial, razón de chances*

Abstract: This *paper presents an illustration of logistic regression models in order to explain the death probability versus men's age in Valencian Community in 2000. First, theoretical principles are developed. Then using R software, a logistic adjustment has been made with the purpose of identifying significance of coefficients based on statistical evidence, and finally calculating odds ratio. The main results have been that a Logistic Regression Model fixed the probability of dying in men´s Valencian Comunity in 2000 and the men´s age is an important factor when probability of dying is computed.*
Keywords: *R, logistic regression, regression models, nonlinear models, residuals, leverages, binomial, odds ratio.*

Recibido: Noviembre 2016
Aceptado: Marzo 2017

1. INTRODUCCIÓN

El presente trabajo está orientado a presentar una ilustración del modelamiento estadístico de datos usando la Regresión Logística [1]. Para el efecto, se consideraron los datos de población y defunciones de los Hombres de la Comunidad Valenciana registrados en el año 2000, para edades comprendidas entre 0 y 100 años con la finalidad de identificar los cambios que se pudieran observar en la probabilidad de fallecimiento a medida que la edad de los hombres aumenta; así como también, ajustar un modelo donde se logre explicar la probabilidad de fallecimiento en función de la edad, para posteriormente determinar el chance o posibilidad de fallecer y la razón de chances, es decir, "odds" y "odds ratio", respectivamente [2].

2. METODOLOGÍA

En este trabajo se inició con la revisión de los fundamentos teóricos de la Regresión Logística como parte de los Modelos Lineales Generalizados.

Posteriormente, se generó un gráfico *"Probabilidad de fallecimiento de Hombres vs Edad"* con el objetivo de identificar el cambio de la probabilidad a medida que aumentaba la edad de los hombres. Se observó la curva obtenida en el gráfico anterior; y, con esta información se procedió a realizar un ajuste de la curva, a través de modelo de regresión logística de la familia Binomial [2]. Con los resultados de las estimaciones del modelo se determinó la significancia de los parámetros; y, dado que la regresión logística involucra una probabilidad de éxito, se calculó la razón de chances o posibilidades de fallecimiento en los hombres de la Comunidad Valenciana; para finalizar, se efectuó el análisis de los gráficos de diagnóstico del modelo ajustado [3].

3. FUNDAMENTOS TEÓRICOS

Sea Y una variable binaria que toma valores 1 o 0, correspondientes a una probabilidad asociada con éxito o fracaso; tal que:

$$Y \sim Binomial(n, p)$$

Bajo estas condiciones, la probabilidad p es una valor entre cero y uno; y, por lo tanto, el logaritmo natural de $p/1-p$ es un valor entre cero e infinito.

$$0 < p < 1$$

$$-\infty < \log(\theta) = \log\left(\frac{p}{1-p}\right) < \infty$$

[1] *Plata Wendy, Profesora del Departamento de Matemáticas, Facultad de Ciencias Naturales y Matemáticas, ESPOL. (e-mail: wplata@espol.edu.ec).*

[2] *Yambay Karen, Profesora del Centro de Lenguas Extranjeras, Facultad de Ciencias Sociales y Humanísticas, ESPOL. (e-mail: kayambay@espol.edu.ec)*

Donde $\theta = \frac{p}{1-p}$ se denomina *chance* o *posibilidad*, siendo así que $\eta = \log(\theta)$, entonces:

$$\eta = \log\left(\frac{p}{1-p}\right)$$

Si $\eta = \beta_0 + \beta_1 X = \log(\theta)$, entonces

$$\theta = e^{\beta_0+\beta_1 X}$$

Por lo tanto, cuando X aumenta en una unidad, la razón de chances o razón de posibilidades Ω, queda en función de β_1, como se observa a continuación:

$$x+1 \quad \rightarrow \quad \theta = e^{\beta_0+\beta_1 X+1}$$

$$x \quad \rightarrow \quad \theta = e^{\beta_0+\beta_1 X}$$

$$\Omega = \frac{\theta_{x+1}}{\theta_x} = e^{\beta_1}$$

La razón de chance $\Omega = e^{\beta_1}$ es el resultado de incrementar X en una unidad. Por consiguiente, en el Modelo de Regresión Logística en el que interviene una sola variable de explicación, la prueba de hipótesis consiste únicamente en verificar la significancia de β_1, lo cual es equivalente a verificar si la razón de chances Ω es igual a uno.

$$H_0: \beta_1 = 0 \quad equivale\ a \quad H_0: \Omega = 1$$

Si no se rechaza H_0, esto quiere decir que, no existe efecto alguno en la variable de respuesta al incrementar la variable de explicación en una unidad [1].

1. AJUSTE DEL MODELO DE REGRESIÓN LOGÍSTICA

Con el propósito de ilustrar un ejemplo práctico de un modelo de regresión logística, se utilizaron los datos de población y defunciones de los Hombres de la Comunidad Valenciana registrados en el año 2000, para edades comprendidas entre 0 y 100 años, definiendo una variable de respuesta Binomial [6], tal que:

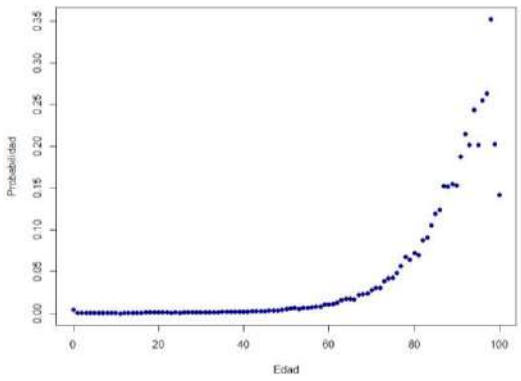
$$Y \sim Binomial(n, p)$$

Donde p corresponde a la *Probabilidad de fallecimiento de Hombres en la Comunidad Valenciana para el año 2000.*

La estructura de la base de datos contaba con las variables *edad, año, cantidad de defunciones* y *cantidad de hombres vivos*. Dado que, la probabilidad de interés es la de fallecimiento, se definió como ***Éxito:*** Defunciones y ***Fracaso:*** Hombres vivos, siendo $\boldsymbol{n_i}$ la suma de la *cantidad de defunciones* y la *cantidad de hombres vivos* para cada *edad* $i = 0,1,2,\ldots 100$; con estos datos, se calculó la $\boldsymbol{p}$ ***= Probabilidad de fallecimiento de los Hombres en la Comunidad Valenciana.***

Con el soporte del software estadístico R [4], se procedió a graficar p en función de la variable *edad*. Véase Gráfico 1.

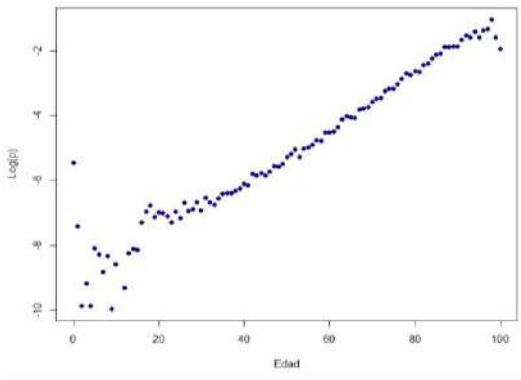
Gráfico 1: Probabilidad de Fallecimiento Hombres en la Comunidad Valenciana – Año 2000



En el Gráfico 1 se puede apreciar que a edades menores que 50 años la probabilidad de fallecer es muy baja mientras que partir de 50 años la probabilidad de fallecer aumenta; y, cuando un hombre cumple 98 años tiene la más alta probabilidad de morir, mientras que la probabilidad de morir a los 99 es menor que la de morir a los 98.

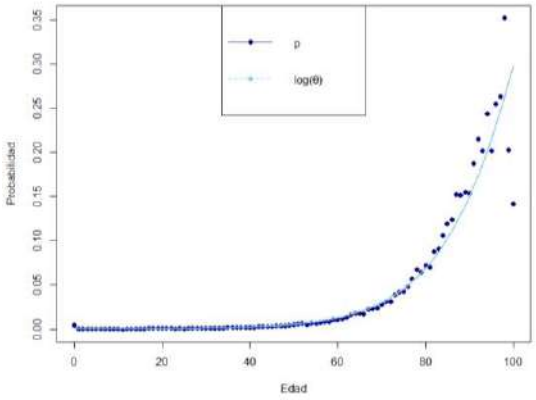
Una forma de visualización alternativa de la probabilidad de fallecimiento es a través del cálculo de su logaritmo natural, donde la protuberancia de la curva entre las edades 18 y 20 podría atribuirse a fallecimientos ocurridos a temprana edad por accidentes de tránsito [5]. Véase Gráfico 2.

Gráfico 2: Logaritmo de la Probabilidad de Fallecimiento de Hombres



Una vez que fueron analizados los datos sin ajustar, se procedió a elaborar un modelo de regresión logística, como se observa en el Gráfico 3.

Gráfico 3: Ajuste del Modelo de Regresión Logística Probabilidad de Fallecimiento de Hombres



Con la curva de regresión logística $log(\theta)$ [7] se trató de explicar la *Probabilidad de fallecimiento en función de la edad de los Hombres*, este ajuste se lo realizó con el uso del comando:

$$\text{“}cbind(y, n-y)\text{”}$$

Donde $\boldsymbol{y}$ corresponde al número de éxitos y $(\boldsymbol{n-y})$ es la cantidad de fracasos; además, para el ajuste del modelo se asoció a la variable de respuesta Y con la familia Binomial. Los resultados del ajuste se muestran en el Cuadro 1.

Cuadro 1: Resultados de Modelo Ajustado con Regresión Logística

```
glm(formula = cbind(def, pob) ~ edad, family = binomial, data =
Tabla)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.6474  -1.4239   0.2611   1.8554  23.7849
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.583809   0.036406  -263.2   <2e-16 ***
edad         0.087252   0.000506   172.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48691.7  on 100  degrees of freedom

Residual deviance:  1084.5  on  99  degrees of freedom

AIC: 1703.6      Number of Fisher Scoring iterations: 5
```

De los resultados del Cuadro 1 se pudo referir que el coeficiente de la variable *edad* es altamente significativo, dado que existe evidencia abrumadora para no rechazar la hipótesis alterna que postula $H_1: \beta_1 \neq 0$, ya que el valor p de la prueba es muy pequeño; por lo tanto, existe evidencia estadística para inferir que la variable edad explica la *Probabilidad de fallecimiento de los Hombres en la Comunidad Valenciana*.

El modelo ajustado resultante de la Regresión Logística es:

$$\eta = -9.583809 + 0.087252 \text{ Edad}$$

Al calcular la razón de chances o razón de posibilidades, se tiene:

$$\Omega = exp(0.087252) = 1.091172$$

Lo cual significa que la posibilidad de que un hombre fallezca aumenta en 9,11% cada vez que su edad aumenta en un año.

2. INTERPRETACIÓN DE GRÁFICOS DE DIAGNÓSTICO

Al construir modelos de regresión se generan los denominados *gráficos de diagnóstico*, además de los resultados expuestos en la anterior sección.

El Gráfico 4 muestra los Residuos vs los Valores ajustados, donde se puede observar que los Errores son aleatorios y cercanos a cero y también se observa que las observaciones correspondientes a la edad 1 y 88 no se ajustan bien al modelo, pues tienen valores muy por encima de la recta, estas observaciones pudieran estar afectando la estimación del modelo.

Gráfico 4: Residuals vs Fitted

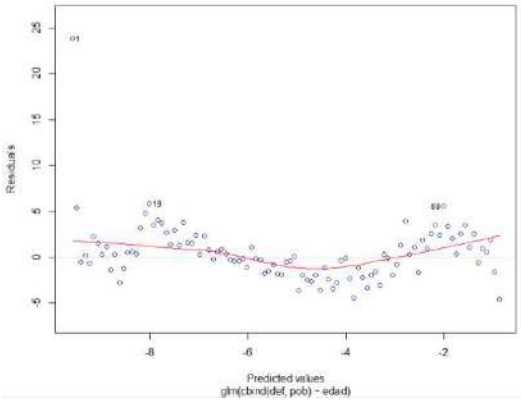
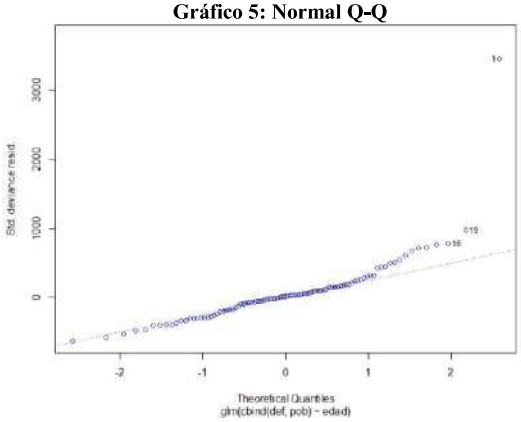


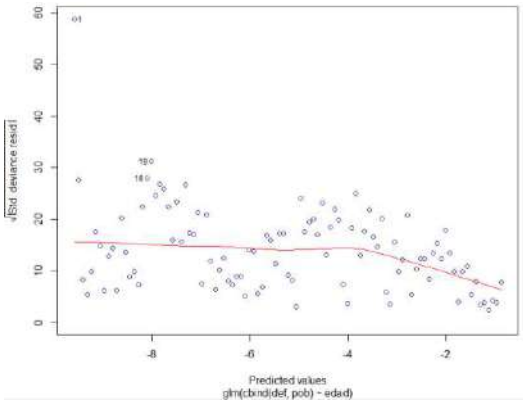
Gráfico 5: Normal Q-Q



En el Gráfico 5, que corresponde a Normal Q-Q [8] se pudo evidenciar que los Errores no se ajustan del todo a una línea recta, siendo los Errores de las edades, 1, 18 y 19 muy extremos.

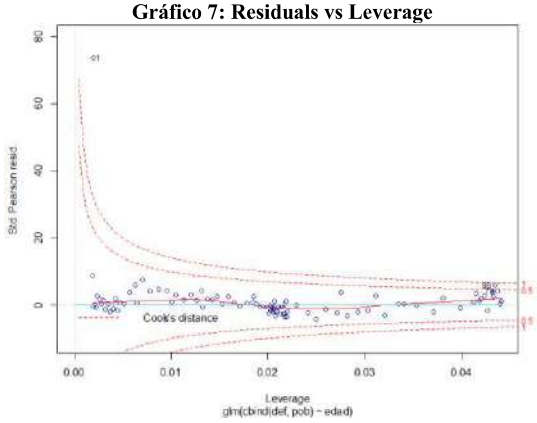
En el Gráfico 6, Scale-Location, se observa que la mayoría de Errores son apropiados, mientras que, los Errores de edades 1, 18 y 19 son grandes.

Gráfico 6: Scale-Location



Con respecto a Residuals vs Leverage del Gráfico 7, se pudo evidenciar que los Errores correspondientes a la edad 1 son valores muy por encima de las demás observaciones; y, la observación de la edad 88 puede afectar a ambos, es decir, a los Residuos y a Leverage [9], dado que es aberrante.

Gráfico 7: Residuals vs Leverage



3. CÓDIGO R

Se generó código en el lenguaje de programación R [10], a través de la interfaz de usuario R-Studio, con la finalidad de seguir los pasos propuestos en la metodología del presente estudio, es decir, cargar la base de datos a R, graficar la Probabilidad de Fallecimiento vs Edad, ajustar el modelo y verificar supuestos.

Cuadro 2: Código R usado en la metodología

```
library(lattice, pos=4)
#CARGAR DATOS A R
Tabla <-
read.csv("~/ESPOL/Curso_Estadistica_Investigadores/Proyecto_Final_R/Tabla_Mortalidad_Hombres.csv", sep=";")
View(Tabla)

#CALCULAR PROBABILIDAD DE FALLECIMIENTO
n=with(Tabla,def+pob)
p=Tabla$def/n

#GRAFICAR PROBABILIDAD DE FALLECIMIENTO VS EDAD
plot(p~edad,xlab="Edad",ylab="Probabilidad",main="Probabilidad de
Fallecimiento en Hombres",
   col="navyblue", pch=16,auto.key=list(border=TRUE),
   par.settings =
simpleTheme(pch=16),scales=list(x=list(relation='same'),y=list(relation='s
ame')),
   Tabla)
legend("top",legend=c("p","log(θ)"),col=c("navyblue","skyblue"),pch=16,l
ty=1:2)
#plot(Tabla)

#AJUSTAR MODELO DE REGRESIÓN LOGÍSTICA
modelo1=glm(cbind(def,pob)~edad,binomial,Tabla)
summary(modelo1)

# GENERAR GRÁFICOS DE DIAGNÓSTICO
plot(modelo1,col="navyblue")

# GRAFICAR CURVA DE AJUSTE DEL MODELO
curve(predict(modelo1,data.frame(edad=x),type="response"),col="skyblu
e",lwd=2,add=TRUE)

#GRAFICAR EL LOGARITMO NATURAL DE LA PROBABILIDAD DE FALLECER
plot(log(p)~edad,
   xlab="Edad",ylab="Log(p)",main="Logaritmo de la Probabilidad de
Fallecimiento en Hombres",
   col="navyblue", pch=16,auto.key=list(border=TRUE),
   par.settings =
simpleTheme(pch=16),scales=list(x=list(relation='same'),y=list(relation='s
ame')),
   Tabla)
```

4. CONCLUSIONES

Mediante el desarrollo de este trabajo se pudo evidenciar que el Modelo de Regresión Logística se ajustó perfectamente a la probabilidad de fallecimiento de los Hombres de la Comunidad Valenciana para el año 2000, dado que hubo evidencia abrumadora para no rechazar la hipótesis alterna, el valor p de la prueba resulto menor que *2e-16*. Además, se pudo comprobar que la *edad* de los hombres de la Comunidad Valenciana influye en la probabilidad de fallecer, pues se observó que la razón de chances o posibilidades resultó mayor que 1. Finalmente, se determinó que a pesar del buen ajuste obtenido con el modelo, existen observaciones que pudiesen estar afectando el ajuste del mismo.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] Agresti, A. (2002). "Categorical Data Analysis", Second Edition, Wiley & Sons Inc., University of Florida.

[2] Demidenko, E. (2013). "Mixed Models, Theory and Applications with R", Second Edition, Wiley & Sons Inc., USA.

[3] Pinheiro, J. C. & Bates, D. M. (2000). "Mixed-Effects Models in s and S-PLUS", Springer-Verlag New York, Inc., USA.

[4] Venables, W. N. & Ripley, B. D. (2002). "Modern Applied Statistics with S", Fourth Edition, Springer-Verlag New York, Inc., USA.

[5] Fernández, R. (2014). "Mortalidad por causas externas en España", a. Instituto de Salud Carlos III. Boletín Epidemiológico Semanal, Vol. 22, No. 6, 2014, Centro Nacional de Epidemiología (CNE), España.

[6] Biagini, F. & Campanino, M. (2016). "Elements of Probability and Statistics", Springer International Publishing Switzerland.

[7] Hormer, D. & Lemeshow, S. (2000). "Applied Logistic Regression", Second Edition, Wiley, Intersciencie Publication, Canada.

[8] Henrr, C. (2002). "Testing for Normality", Marcel, Dekker INC., New York.Basel

[9] Thorn, C. (2016). "The R Primer", Second Edition, CRC Press, University of Copenhagen, Denmark.

[10] Christian, P. (2010). "Introducing Method Monte Carlo with R", Springer Science+Business Media, London.