# Aplicación del Modelo de Regresión Logística

Autor: Carlos Carreño, ccarrenovi@gmail.com , *Dedicado al Arquitecto del Universo!.*

## Introducción

El modelo de regresión logística es un modelo estadístico utilizado para calcular la probabilidad, que una variable categórica y dicotómica tome uno de los dos valores posibles. En la regresión logística la probabilidad de que la variable dependiente tome uno de los valores posibles depende de una o más variables independientes o explicativas esto se denota con *P*(Y=k/X=x).

Cuando la probabilidad P depende de una sola variable explicativa estamos en el caso de la regresión logística simple y cuando depende de más de una variable estamos en el caso de la regresión logística multivariable.

Para modelar la distribución de probabilidad se utiliza la función sigmoideo.

$$P(Y=k)=\frac{1}{1+e^{-f(x)}}\ ;k=\{1,0\}$$

$$f(x)=\beta_0+\beta_1 x$$

$$P(Y=k)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$P(Y=k)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\ \text{.....(1)}$$

Realizando operación simples se puede demostrar que:

$$\ln\left(\frac{P}{1-P}\right)=\ \beta_0+\beta_1 x\ \text{.....(2)}$$

A la ecuación (1) se le conoce como función logarítmica de pares o ODDS.

Exponenciando la ecuación (1) obtenemos.

$$\frac{P}{1-P}=e^{\beta_0}e^{\beta_1 x}\text{.....(3)}$$

Si P es la probabilidad del éxito $q=1-P$ es la probabilidad del fracaso, la razón $\frac{P}{q}$ se denomina razón de probabilidad u ODD.

La regresión logística modela la probabilidad que la variable dependiente o variable de respuesta *Y* sea un éxito $(Y=1)$ en función de los valores que tomen los coeficientes o predictores $\beta_o$ y $\beta_1$ utilizando la función $\log(ODD)$. Utilizando R podemos analizar la relación entre *P* y $\log(\frac{p}{1-p})$.

## Razón de Probabilidad

1. Construye el data frame que contenga la probabilidad p de éxito, q de fracaso, odd y log.

```
p = seq(0.001,1,0.001)
```

```
q = 1-p
datos = data.frame(p=p,q=q,odd=p/q,log=log(p/q))
View(datos)
```

| | p | q | odd | log |
|---|---|---|---|---|
| **493** | 0.493 | 0.507 | 0.9723866 | -0.028001830 |
| **494** | 0.494 | 0.506 | 0.9762846 | -0.024001152 |
| **495** | 0.495 | 0.505 | 0.9801980 | -0.020000667 |
| **496** | 0.496 | 0.504 | 0.9841270 | -0.016000341 |
| **497** | 0.497 | 0.503 | 0.9880716 | -0.012000144 |
| **498** | 0.498 | 0.502 | 0.9920319 | -0.008000043 |
| **499** | 0.499 | 0.501 | 0.9960080 | -0.004000005 |
| **500** | 0.500 | 0.500 | 1.0000000 | 0.000000000 |
| **501** | 0.501 | 0.499 | 1.0040080 | 0.004000005 |
| **502** | 0.502 | 0.498 | 1.0080321 | 0.008000043 |
| **503** | 0.503 | 0.497 | 1.0120724 | 0.012000144 |
| **504** | 0.504 | 0.496 | 1.0161290 | 0.016000341 |
| **505** | 0.505 | 0.495 | 1.0202020 | 0.020000667 |
| **506** | 0.506 | 0.494 | 1.0242915 | 0.024001152 |
| **507** | 0.507 | 0.493 | 1.0283976 | 0.028001830 |
| **508** | 0.508 | 0.492 | 1.0325203 | 0.032002731 |
| **509** | 0.509 | 0.491 | 1.0366599 | 0.036003889 |
| **510** | 0.510 | 0.490 | 1.0408163 | 0.040005335 |

Figura 1.- Data frame de datos

2. Grafica *p* vs *odd*.

```
plot(datos$p~datos$odd,xlab="ODD",ylab="Probabilidad")
```

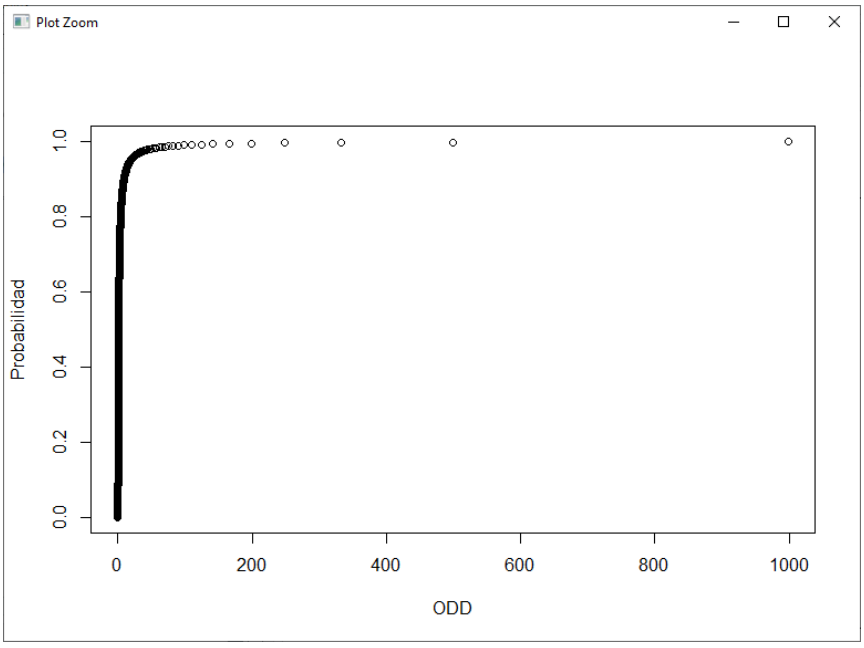



Figura 2.- Probabilidad vs ODD

3. Grafica *p* vs *log(odd)*

```
plot(datos$p~datos$log,xlab="log(ODD)",ylab="Probabilidad")
```

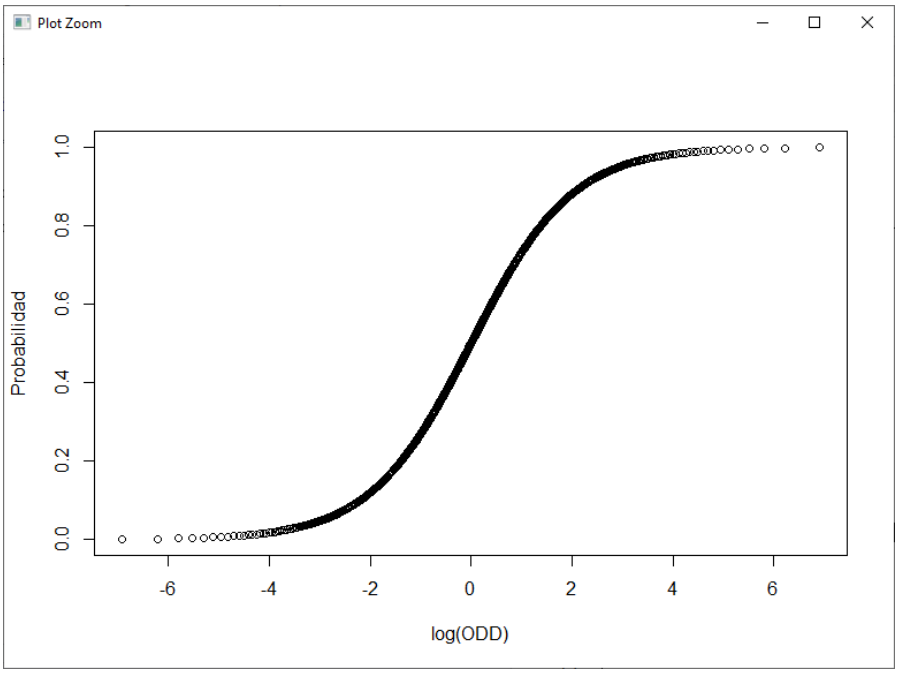



Figura 3.- Probabilidad vs Log(ODD)

Como se puede apreciar de la figura 3 y teniendo en cuenta la ecuación(2) se usa la función logarítmica de los ODD para calcular la probabilidad de la razón de éxitos entre fracasos, también se puede notar que la probabilidad está acotada por el intervalo $[0,1]$ como sabemos, $\log(ODD)$ puede tomar valores de entre $[-\infty, +\infty]$.

Se puede apreciar las siguientes reglas en la regresión logística:

- *P(éxito) > P(fracaso)* si *log(ODD) > 0*
- *P(éxito) < P(fracaso)* si *log(ODD) < 0*
- *P(éxito) = P(fracaso)* si *log(ODD) = 0*

## Estimación de Parámetros

La estimación de los parámetros $\beta_0$ y $\beta_1$ se realiza utilizando el método de máxima verosimilitud, en este método los valores de los parámetros son aquellos que maximizan la probabilidad de obtener los datos observados.

## Creación del Modelo de Regresión Logística Simple

Analizaremos la relación de la probabilidad de tener un atraso en el pago de la cuota de la tarjeta de crédito “default” respecto del promedio de saldo mensual de la tarjeta al cual llamaremos “balance”.

1. Construye el data frame y prepara los datos

```
library(tidyverse)
library(ISLR)

datos = select(Default, default, balance)

# Recodificando NO=0 y Si=1
datos = mutate(datos, default = recode(default, "No"=0,"Yes"=1))
```

2. Crea el modelo

```
modelo = glm(default~balance, data=datos, family = "binomial")
```

3. Muestras el valor de los parámetros del modelo

```
summary(modelo)
```

```
Call:
glm(formula = default ~ balance, family = "binomial", data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <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: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

## Validación del Modelo

En regresión logística se evalúa la significancia del modelo en su conjunto mediante el estadístico ***p-value***. Se puede usar uno de los siguientes métodos:

- ***Wald chi-square***: está muy expandido pero pierde precisión con tamaños muéstrales pequeños.
- ***Likelihood ratio***: usa la diferencia entre la probabilidad de obtener los valores observados con el modelo logístico creado y las probabilidades de hacerlo con un modelo sin relación entre las variables.

1. Calcula el estadístico del error

```
error = with(modelo,null.deviance - deviance )
print(error)
[1] 1324.198
```

2. Calcula el p-value y contrástalo contra $\propto = 0.05$

```
# número de variables
k=1
p_value = with(modelo, pchisq(error,k,lower.tail = FALSE))
print(p_value)
[1] 6.232869e-290
```

Como puedes apreciar en valor de p-value=6.23e-290 es mucho menor que $\propto = 0.05$, por lo tanto rechazamos la hipótesis nula $H_o$ y aceptamos la hipótesis $H_1$que indica que la variable balance si es significativa para el modelo.

## Las Predicciones

Una vez estimados los coeficientes $\beta_0$ y $\beta_1$ podemos usar la ecuación (1)

$$P(Y = k) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

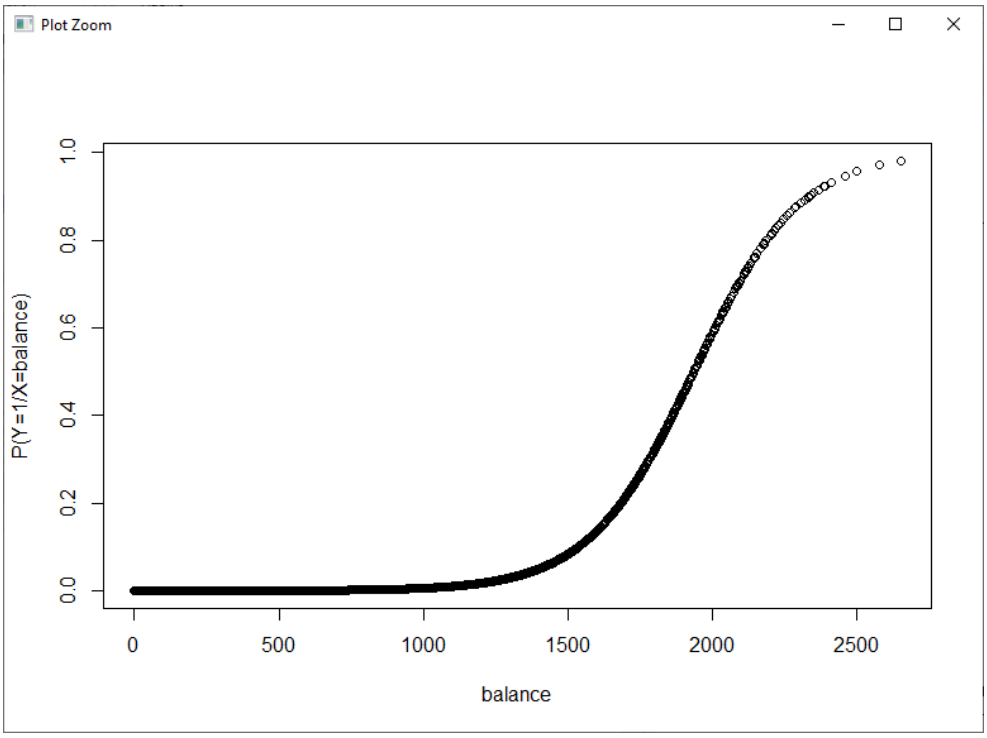
para calcular la probabilidad de que la variable de respuesta Y pertenezca al nivel de referencia (éxito), para un valor determinado de la variable predictora o explicativa que puede ser continua en otras palabras realizar las predicciones.

1. Calcula la probabilidad de la Variable Respuesta

```
datos$Predict = modelo$fitted.values
head(datos)
default   balance      Predict
1       0  729.5265 0.0013056797
2       0  817.1804 0.0021125949
3       0 1073.5492 0.0085947405
4       0  529.2506 0.0004344368
5       0  785.6559 0.0017769574
6       0  919.5885 0.0037041528
```

2. Grafica Predict vs balance

```
plot(datos$Predict~datos$balance,xlab="balance",ylab = "P(Y=1/X=balance)")
```




3. Crea nuevos datos balance y utiliza el modelo para predecir la variable de respuesta.

```
nuevos_balance = sample(seq(500,3000,15),size = 100,replace = FALSE,prob = NULL)
head(nuevos_balance)
[1] 1025  875 1400  995 2930 2825
```

```
nuevos_datos = data.frame(balance = nuevos_balance)
nuevos_datos$Predict=predict(modelo,  newdata = nuevos_datos, type = "response")
```

```
library(InformationValue)
corte = optimalCutoff(datos$default,datos$Predict,optimiseFor="Both")
print(corte)
[1] 0.03100811
```

```
nuevos_datos=mutate(nuevos_datos,
      default= ifelse(Predict > corte, 1, 0)
)

View(nuevos_datos)
```

| | balance | Predict | default |
|---|---|---|---|
| 1 | 1025 | 0.0065942524 | 0 |
| 2 | 875 | 0.0029010470 | 0 |
| 3 | 1400 | 0.0496021310 | 1 |
| 4 | 995 | 0.0055970198 | 0 |
| 5 | 2930 | 0.9957665517 | 1 |
| 6 | 2825 | 0.9924834995 | 1 |
| 7 | 2810 | 0.9918424950 | 1 |
| 8 | 2720 | 0.9866884655 | 1 |
| 9 | 1535 | 0.0988121525 | 1 |
| 10 | 2330 | 0.8967057293 | 1 |
| 11 | 1955 | 0.5247427699 | 1 |
| 12 | 590 | 0.0006066391 | 0 |

## Modelo de Regresión Logística Multivariable

El modelo de regresión logística multivariable se basa en los mismos principios que el modelo de regresión logística simple solo se amplía el número de variables explicativas o predictores, estas pueden ser categóricas o continuas. La fórmula para el cálculo de $\log(ODD)$ será:

$$logit(Y) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

La fórmula para el cálculo de la probabilidad de que la variable de respuesta tome el valor del nivel de éxito será:

$$P(Y = k) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i}}$$

## Ejercicio: Aplicación del Modelo de Regresión Logística Multivariable

Se utilizara el conjunto de datos de Framingham, el cual contiene más de 4900 observaciones con distintas variables de pacientes utilizadas para estudiar el impacto de estas variables en problemas cardiovasculares.

Realiza las siguientes actividades:

a) Crea un modelo de regresión logística a partir del conjunto de datos
b) Calcula los parámetros del modelo
c) Realiza la validación del Modelo
d) Calcula el nivel de acierto del modelo
e) Realiza predicciones sobre nuevos datos