UNIVERSIDAD DE PUERTO RICO
RECINTO DE RIO PIEDRAS
FACULTAD DE ADMINISTRACION DE EMPRESAS
Instituto de Estadística y Sistemas Computadorizados de Información

REGRESIÓN LOGÍSTICA

USANDO R

Preparado por:
José Carlos Vega Vilca, Ph.D.
Jose.vega23@upr.edu

REGRESIÓN LOGÍSTICA

Es la relación funcional entre una variable dependiente cualitativa (dos o más categorías) y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial un modelo de tipo exponencial. Una transformación logarítmica (*logit*) permite su uso como una función lineal.

Ejemplo 1.- Se llevó a cabo un estudio con 650 clientes de una Institución Bancaria para establecer la relación entre los que poseen "casa propia" y el "Nivel de Ingresos". Los resultados fueron

		Nivel de Ingresos		
		Alto	Bajo	
Casa Propia	SI	200	50	250
	NO	100	300	400
		300	350	**650**

Definición de Odds y Odds Ratio

Entre los que tienen **Ingresos Altos**. ¿Cuál es la probabilidad de tener casa propia	
Entre los que tienen **ingresos Altos**. ¿Cuál es la probabilidad de NO tener casa propia	
EL odds de tener casa propia, con **Ingresos Altos**	
Entre los que tienen **Ingresos Bajos**. ¿Cuál es la probabilidad de tener casa propia	
Entre los que tienen **ingresos Bajos**. ¿Cuál es la probabilidad de NO tener casa propia	
El odds de tener casa propia, con **Ingresos Bajos**	
El Odds Ratio (OR) de tener casa propia	

MODELO DE REGRESION LOGISTICA SIMPLE

Sea *Y* una variable binomial puntual, donde se cumple:

$P(Y = éxito) = p$ y $P(Y = fracaso) = 1 - p = q$.

Sea *X* una variable aleatoria o no, continua o discreta. El modelo de regresión expresa el logaritmo del *odds* para un valor de $X = x$.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \beta_0 + \beta_1 X$$

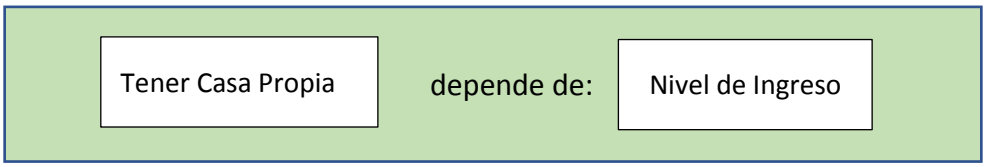
También:

$$log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X$$

Relación de las variables en estudio

La variable dependiente: Y: Casa Propia ➔ SI (1), NO (0)

La variable independiente: X: Nivel de Ingresos ➔Alto (1), Bajo (0)



Datos:

		Nivel de Ingresos		
		Alto	Bajo	
Casa Propia	SI	200	50	250
	NO	100	300	400
		300	350	**650**

```
  casa ingreso  frec
   Y=1     X=1   200
   Y=1     X=0    50
   Y=0     X=1   100
   Y=0     X=0   300
```

```
casa=c(rep(1,250),rep(0,400))
ingre=c(rep(1,200),rep(0,50),rep(1,100),rep(0,300))
```

```
mode=glm(casa~ingre,family=binomial)
summary(mode)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7918     0.1528  -11.73   <2e-16 ***
ingre         2.4849     0.1958   12.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Presentación del Modelo

$$log\left[\frac{P(Y=1)}{P(Y=0)}\right] = -1.7918 + 2.4849\ Ingresos$$

Odds Ratio: OR = $\exp(2.4849) = 12.00$

Cuando se comparan a los clientes de Ingresos Altos y Bajos. Si entre los clientes de Ingresos Bajos se puede encontrar UNO con casa propia, entonces entre los clientes de Ingresos Altos se podrá encontrar DOCE con casa propia, en promedio.

Usando el modelo, calcular:

El odds de tener casa propia, con Ingresos Altos	
El odds de tener casa propia, con Ingresos Bajos	

Cálculo del OR

```
exp(cbind(OR=coef(mode),confint(mode)))
                    OR     2.5 %     97.5 %
(Intercept)  0.1666667 0.1221501  0.2225914
ingre       12.0000000 8.2357850 17.7587357
```

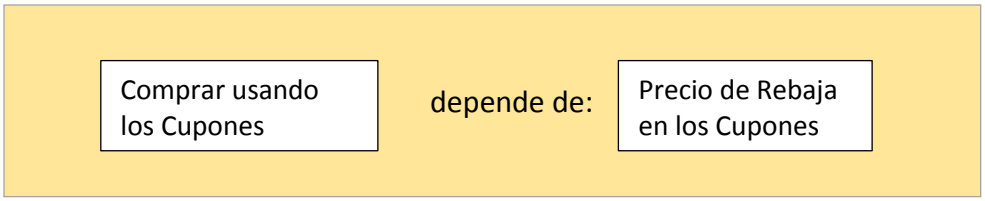
Ejemplo 2.- En una promoción de ventas: una tienda ha entregado 1000 cupones en hogares de nuevos clientes, para rebajar el precio de la compra en una próxima oportunidad. Los resultados se muestran en la siguiente tabla

Cupón	Número de Hogares	Número de clientes que compraron	No compraron	Proporción de Cupones Usados
5	200	30	170	0.150
10	200	55	145	0.275
15	200	70	130	0.350
20	200	100	100	0.500
30	200	137	63	0.685
	1000	**392**	**608**	

Relación entre variables

La variable dependiente: Comprar usando los cupones

La variable independiente: Precio de Rebaja en los Cupones




```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.044348   0.160977  -12.70   <2e-16 ***
Cupon        0.096834   0.008549   11.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Presentación del Modelo

$$log\left(\frac{P(éxito)}{P(fracaso)}\right) = -2.0443 + 0.0968\ PrecioCupón$$

Odds Ratio: OR = $\exp(0.096834) = 1.102$ Entre las personas que compraron con el cupón, por cada dólar adicional en dicho cupón, las persona que compran se incrementan en promedio en 10.2%

También: Si comparamos a las persona que compraron con cupones de \$5 y personas que compraron con cupones de \$20. En este último grupo la posibilidad de compra es $1.102^{15} = 4.92$ veces, respecto al primer grupo.

MODELO DE REGRESION LOGISTICA MULTIPLE

Sea *Y* una variable binomial puntual, donde se cumple:
$P(Y = éxito) = p$ y $P(Y = fracaso) = 1 - p = q$.

Sea $X_1, X_2, \cdots, X_k$, variables aleatorias o no, continuas o discretas. El modelo de regresión expresa el logaritmo del *odds* para $X_1 = x_1, \cdots, X_k = x_k$.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

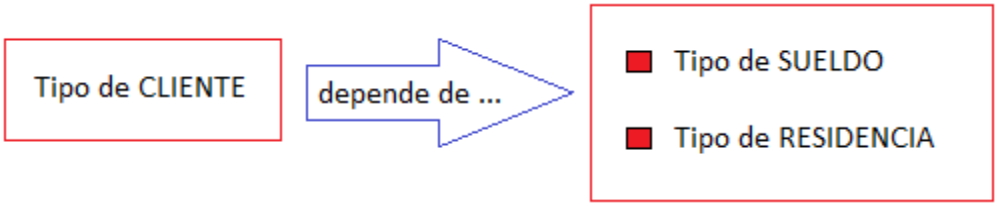
Ejemplo:
Estudiar, mediante un modelo de regresión logística, la posible asociación entre el tipo de cliente de una institución bancaria con las variables: tipo de sueldo y ambiente de residencia. Las variables en estudio son:

- Tipo de cliente: BUENO (1) y MALO (0)
- Tipo de sueldo: ALTO (1) y bajo (0)
- Tipo de residencia: URBANO (1) y RURAL (0)

Los resultados se resumen en la tabla siguiente:

	SUELDO.ALTO(1)		SUELDO.BAJO(0)	
Medio →	Urbano(1)	Rural(0)	Urbano(1)	Rural(0)
BUENO(1)	32	1	15	2
MALO(0)	15	10	15	10

Relación de variables



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2012     0.6744  -3.264 0.001099 **
sueldo        0.5727     0.4475   1.280 0.200620
residencia    2.3086     0.6651   3.471 0.000518 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Presentación del Modelo

$$log\left(\frac{p}{1-p}\right) = -2.2012 + 0.5727\ sueldo + 2.3086\ residencia$$

```
                   OR      2.5 %    97.5 %
(Intercept)  0.1106686 0.02384566  0.361456
sueldo       1.7729671 0.73905725  4.311259
residencia  10.0599158 3.09041497 45.586893
```

MODELO DE REGRESIÓN LOGISTICA NOMINAL

Este modelo es usado cuando no hay un orden natural en las categorías de la variable respuesta. Aquí una categoría es elegida arbitrariamente como la categoría de referencia. Supongamos que ésta es la primera categoría, entonces la probabilidad de clasificar una observación en una de las *G* clases es obtenida del modelo:

$$log\left[\frac{P(Y=g)}{P(Y=1)}\right] = c_g + \beta_{1g}X_1 + \beta_{2g}X_2 + \cdots + \beta_{pg}X_p$$

$$g = 2, 3, \cdots, G$$

Ejemplo:

Un estudio sobre un conjunto de datos de *iris*, las mediciones en centímetros de las variables longitud y ancho de sépalo, y longitud y ancho de pétalo, respectivamente, para 50 flores desde cada una de tres especies de iris. Las especies son: “setosa”, “versicolor”, “virginica”

- X1: longitud de sépalo
- X2: ancho de sépalo
- X3: longitud de pétalo
- X4: ancho de pétalo
- Y: especie

```
Coefficients:
  (Intercept)         X1          X2        X3         X4
2    18.69037 -5.458424  -8.707401 14.24477 -3.097684
3   -23.83628 -7.923634 -15.370769 23.65978 15.135301
```

Presentación del modelo

El modelo de regresión logística para los datos analizados está formado por el siguiente sistema de dos ecuaciones:

$$log\left[\frac{P(Y=2)}{P(Y=1)}\right] = \quad 18.69 - 5.46\,X1 - 8.71\,X2 + 14.24\,X3 - 3.10\,X4$$

$$log\left[\frac{P(Y=3)}{P(Y=1)}\right] = -23.84 - 7.92\,X1 - 15.37\,X2 + 23.66\,X3 + 15.14\,X4$$

Cálculo de la Tasa de Error

1) Tasa de Error Aparente (TEA)

$$TEA = \frac{2}{150} \times 100 = 1.33\%$$

	Grupo verdadero		
pre	**1**	**2**	**3**
1	50	0	0
2	0	49	1
3	0	1	49

2) Tasa de error por validación cruzada, dejando el 30% fuera

```
$N.errores
[1] 3

$N.casos
[1] 45

$TEVC
[1] 0.06666667

$predicion
 [1] 2 2 2 2 1 1 1 1 1 2 1 1 2 3 3 1 2 3 3 3 3 1 1 1 1 3 2 2 1 1 2 2 1 2 2 2 3 2
[39] 3 2 2 2 1 3 2
Levels: 1 2 3
```

Clasificación de DOS Nuevos Sujetos

El modelo de regresión logística estimado podrá ubicar nuevas flores, en base a sus características. Supongamos dos nuevas flores con las siguientes características de longitud y ancho de sépalo y pétalo:

Flor	X1	X2	X3	X4
1	7.0	3.2	4.7	1.4
2	7.2	3.0	5.8	1.6

```
                1           2            3
Flor1  2.427101e-07 0.99998774 1.201699e-05
Flor2  1.067125e-15 0.02892881 9.710712e-01
```

- La flor 1, tiene la probabilidad más alta de ser clasificado en el grupo 2, es decir en el grupo “versicolor”.
- La flor 2, tiene la probabilidad más alta de ser clasificado en el grupo 3, es decir en el grupo “virginica”.

MODELO DE REGRESIÓN LOGISTICA ORDINAL

Este modelo es usado cuando hay un obvio orden natural en las categorías de la variable respuesta. Hay varios modelos diferentes en regresión logística ordinal; aquí será usado el llamado modelo de chances proporcionales. La probabilidad de clasificar una observación en una de las *G* clases, según este modelo, es obtenido de:

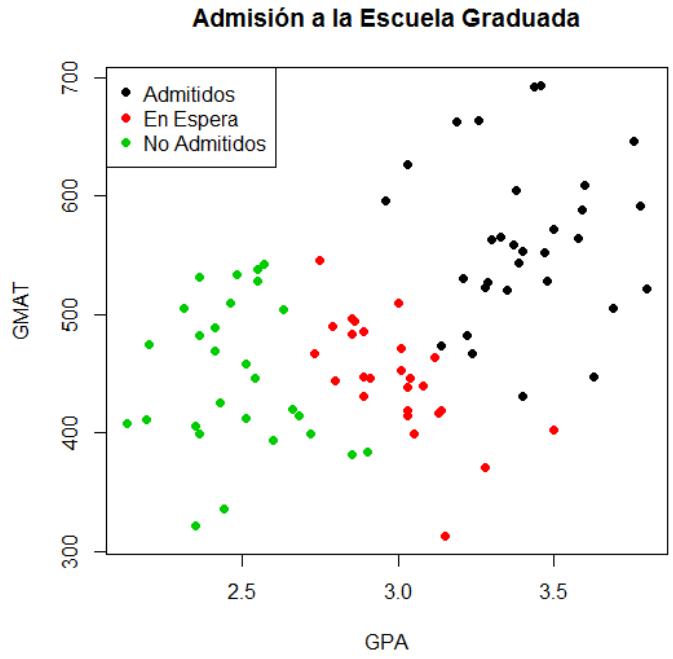
$$log\left[\frac{P(Y \geq g)}{P(Y < g)}\right] = c_g + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$g = 2, 3, \cdots G$$

Ejemplo

Un estudio para analizar la admisión de estudiantes a una Escuela Graduada de Negocios está basado en el análisis de dos variables

X1: GPA
X2: GMAT
Y : grupo Y = 1: Admitidos, Y = 2: En Espera, Y = 3: No Admitidos



ANALISIS

```
                       Model Likelihood     Discrimination     Rank Discrim.
                          Ratio Test           Indexes           Indexes
Obs             85    LR chi2     164.67    R2        0.964    C       0.910
 1              31    d.f.             2    g        16.277    Dxy     0.821
 2              26    Pr(> chi2) <0.0001    gr 11718763.892    gamma   0.995
 3              28                          gp        0.468    tau-a   0.552
max |deriv| 5e-07                          Brier     0.019

     Coef     S.E.    Wald Z Pr(>|Z|)
y>=2 111.2380 32.0996  3.47  0.0005
y>=3  99.6138 28.8052  3.46  0.0005
X1   -26.2049  7.4663 -3.51  0.0004
X2    -0.0593  0.0193 -3.07  0.0022
```

<u>Presentación del modelo</u>

El modelo de regresión logística para los datos analizados está formado por el siguiente sistema de dos ecuaciones:

$$log\left[\frac{P(Y \geq 2)}{P(Y < 2)}\right] = 111.238 - 26.2049\, X1 - 0.0593\, X2$$

$$log\left[\frac{P(Y \geq 3)}{P(Y < 3)}\right] = 99.6138 - 26.2049\, X1 - 0.0593\, X2$$

Cálculo de la Tasa de Error

1) Tasa de Error Aparente (TEA)

$$TEA = \frac{3}{85} \times 100 = 3.53\%$$

Grupo verdadero

pre	**1**	**2**	**3**
1	30	1	0
2	1	24	0
3	0	1	28

2) Tasa de Error por Validación Cruzada, dejando el 30% fuera.

```
$N.errores
[1] 3

$N.casos
[1] 26

$TEVC
[1] 0.1153846

$predicion
 [1] 2 1 1 2 2 2 2 2 3 3 3 2 1 1 2 3 1 3 3 3 3 1 2 3 3 2
```

Clasificación de DOS Nuevos Sujetos

El modelo de regresión logística estimado podrá ubicar nuevos estudiantes en base a sus características. Supongamos dos nuevos estudiantes con las siguientes calificaciones de GPA y GMAT:

Estudiante	X1	X2
1	2.81	500
2	3.24	481

```
              Y=1         Y=2           Y=3
est1 0.0003489319 0.9746580 2.499312e-02
est2 0.8985678892 0.1014311 1.010018e-06
```

- El estudiante 1, tiene la probabilidad más alta de ser clasificado en el grupo 2, es decir en el grupo “En Espera”.
- El estudiante 2, tiene la probabilidad más alta de ser clasificado en el grupo 1, es decir en el grupo “Admitido”.

Caso de estudio 1:
Analizar los datos de una agencia bancaria mediante regresión logística, donde la variable dependiente es la valoración de crédito y las predictoras son: edad, nivel de ingreso, número de tarjetas de crédito, nivel de educación, número de préstamos de carros.

	Variables	tipo
X1	Edad	continuo
X2	Nivel de Ingreso	bajo=1, medio=2, alto=3
X3	Número de tarjetas de crédito	<5 = 1, >=5 = 2
X4	Nivel de educación	Esuperiror = 1, Univer = 2
X5	Número de préstamo de carros	<2 = 1, >=2 =2
Resp	valoración de crédito	1: Bueno, 0: Malo

Caso de estudio 2:
Analizar mediante Regresión Logística el siguiente caso
Variable dependiente: "Response"
Variables independientes: "sex", "Age"

Importance of air conditioning and power steering in cars (row percentages in brackets)*

		Response			
Sex	Age	No or little importance	Important	Very important	Total
Women	18-23	26 (58%)	12 (27%)	7 (16%)	45
	24-40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	18-30	40 (62%)	17 (26%)	8 (12%)	65
	24-40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	15 (37%)	18 (44%)	41
Total		105	94	101	300

* row percentages may not add to 100 due to rounding.

REFERENCIAS
Dobson, A. J. (2002). An Introduction to Generalized Linear Models, Second Edition. Chapman & Hall/CRC

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.