Regresión Logística

Adriano Jonathan[1], Jiménez Kevin[2]
, and Vega Anthony[3]

Universidad Politécnica Salesiana,Departamento de Ingenieria en Ciencias de la Computación, Quito, Ecuador
jadriano@est.ups.edu.ec,kjimenezp1@est.ups.edu.ec
avegac@est.ups.edu.ec

Abstract
In this document, a logistic regression model will be developed, which is a statistical instrument for multivariate analysis, for both explanatory and predictive use. Its use is useful when there is a dichotomous dependent variable (an attribute whose absence or presence we have scored with the values zero and one, respectively) and a set of predictive or independent variables, which can be quantitative (which are called covariates or covariates). or categorical. In the latter case, it is required that they be transformed into "dummy" variables, that is, simulated variables1. The purpose of the analysis is to: predict the probability that a certain "event" will happen to someone: for example, being unemployed = 1 or not being unemployed = 0, being poor = 1 or not poor = 0, receiving a sociologist = 1 or not received = 0). Determine which variables weigh more to increase or decrease the probability that the event in question will happen to someone This assignment of the probability of occurrence of the event to a certain subject, as well as the determination of the weight that each of the dependent variables in this probability , are based on the characteristics of the subjects to whom, indeed, these events occur or not. This logistic regression model will be used to determine which variables are capable of evaluating people's survival after the sinking of the Titanic, thus identifying the reasons that explain why some people have survived and others have not. That is, what variables increased the probability of surviving an accident.

Keywords: Limpieza de datos,regresión,significancia, Comprobación, modelo,outliers, valores influyentes,PassengerId, Survived, Pclass, Name, Sex, Age, Sibsp, Parch

1. Introducción

El modelos de regresión se incluyen un conjunto de técnicas estadísticas que tratan de explicar cómo se modifica la variable dependiente o resultado, cuando cambian otra u otras variables, denominadas independientes o predictores, este modelo de regresión logística se desarrollara en el lenguaje de programación R. Lo que caracteriza en principio a las distintas clases de modelos de regresión es la naturaleza de la variable dependiente; así, con variables continuas la clase de modelos de regresión lineal es la más utilizada; con variables dicotómicas lo es el modelo de regresión logística. La regresión

logística es uno de los instrumentos estadísticos más expresivos y versátiles de que se dispone para el análisis de datos.

2. Metodología

Modelo de Regresión Logística (clasificación), Dataset del Titanic,

2.1. Variabres

•Variables a utilizar
• PassengerId: identificación del pasajero (del 1 al 891)
• Survived: Es la variable dependiente, codificada como 0 si el individuo no sobrevivió y como 1 si el individuo sí lo hizo. Las variables independientes que se utilizarán tendrán que ver con las condiciones de los pasajeros a bordo. Éstas son:
• Pclass: Clase del pasaje (primera clase =1, segunda clase=2, tercera clase = 3)
• Name: Nombre del Pasajero.
• Sex: género del pasajero (male=masculino, female=femenino), para usar esta información se podría codificar como 0 para mujeres y 1 para hombres.
• Age: Edad del pasajero
• Sibsp: Número de hermanos/cónyuges a bordo (0, 1, 2, 3 o más)
• Parch: Número de padres/hijos que acompañaban al individuo (0, 1, 2, 3 o más)

3. Experimentación

En esta sección vamos a analizar el rendimiento de datos reales que contiene nuestro archivo csv, con el fin de saber qué variables incrementaban la probabilidad de sobrevivir a un accidente. en este caso el titanic

Ingreso de Dataset del Titanic

```
#CARGAR Y MOSTRAR BDD
train_titanic <- read_csv("D:/as/period58UPS/ESTADISTICA/segundo_parcial/taller/train_titanic.csv")
View(train_titanic)
```

Selección de variables a utilizar

```
y = train_titanic$Survived
x1= train_titanic$Pclass
x2= train_titanic$Sex
x3= train_titanic$Age
x4= train_titanic$SibSp
x5 = train_titanic$Parch
x6 = train_titanic$Fare
```

0 para mujeres y 1 para hombres.

```
x2 = (x2=='male')*1
```

Limpieza de datos

```
train_titanic2 =  train_titanic[! is.na(train_titanic$Age),]
y2 = train_titanic2$Survived
x12= train_titanic2$Pclass
x22= train_titanic2$Sex
x32= train_titanic2$Age
x42= train_titanic2$SibSp
x52 = train_titanic2$Parch
x62 = train_titanic2$Fare
```

Creamos el vector eliminando los NA que aparecen en la variable fecha, dejando 714 observaciones.
0 para mujeres, 1 para hombres.

```
x22 = (x22=='male')*1

mm = cbind(y2,x12,x22,x32,x42,x52,x62)
```

Vemos si hay una correlación alta entre x32 (edad) con otra variable

```
            y2         x12         x22         x32         x42         x52         x62
y2   1.00000000 -0.35965268 -0.53882559 -0.07722109 -0.01735836  0.09331701  0.26818862
x12 -0.35965268  1.00000000  0.15546030 -0.36922602  0.06724737  0.02568307 -0.55418247
x22 -0.53882559  0.15546030  1.00000000  0.09325358 -0.10394968 -0.24697204 -0.18499425
x32 -0.07722109 -0.36922602  0.09325358  1.00000000 -0.30824676 -0.18911926  0.09606669
x42 -0.01735836  0.06724737 -0.10394968 -0.30824676  1.00000000  0.38381986  0.13832879
x52  0.09331701  0.02568307 -0.24697204 -0.18911926  0.38381986  1.00000000  0.20511888
x62  0.26818862 -0.55418247 -0.18499425  0.09606669  0.13832879  0.20511888  1.00000000
```

No se encuentra alta correlación con ninguna otra variable. Se toma la correlación más alta, en este caso con la x12 con una correlación de -0.3692

Hacemos la ecuación de regresión con las variables de x32 en base a x12.

```
> reg

Call:
lm(formula = x32 ~ x12)

Coefficients:
(Intercept)          x12
     44.011       -6.399
```

Desarrollo de bucle que buscara y reemplaza los valores de la x3 en base a la regresión obtenida **Obtenemos la matriz aleatoria de los datos.**

```
# A tibble: 891 x 12
   PassengerId Survived Pclass Name                    Sex      Age SibSp Parch Ticket   Fare Cabin Embarked
         <dbl>    <dbl>  <dbl> <chr>                   <chr>  <dbl> <dbl> <dbl> <chr>   <dbl> <chr> <chr>
 1         836        1      1 "Compton, Miss. Sara R~ female    39     1     1 PC 17~  83.2 E49   C
 2         679        0      3 "Goodwin, Mrs. Frederi~ female    43     1     6 CA 21~  46.9 NA    S
 3         129        1      3 "Peter, Miss. Anna"     female    25     1     1 2668    22.4 F E69 C
 4         509        0      3 "Olsen, Mr. Henry Marg~ male      28     0     0 C 4001  22.5 NA    S
 5         471        0      3 "Keefe, Mr. Arthur"     male      25     0     0 323592   7.25 NA    S
 6         299        1      1 "Saalfeld, Mr. Adolphe" male      38     0     0 19988   30.5 C106  S
 7         270        1      1 "Bissette, Miss. Ameli~ female    35     0     0 PC 17~ 136.   C99   S
 8         187        1      3 "O'Brien, Mrs. Thomas ~ female    25     1     0 370365  15.5 NA    Q
 9         307        1      1 "Fleming, Miss. Margar~ female    38     0     0 17421  111.   NA    C
10         597        1      2 "Leitch, Miss. Jessie ~ female    31     0     0 248727  33    NA    S
# ... with 881 more rows
```

Matriz con números de filas aleatorizadas. **Creamos las variables de entrenamiento para el modelo de regresión logística**

```
yE = training$Survived
x1E = training$Pclass
x2E = training$Sex
x3E = training$Age
x4E = training$SibSp
x5E = training$Parch
x6E = training$Fare
x2E = (x2E=='male')*1
```

Modelo de regresión logística con las variables de entrenamiento

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.060479   0.668082   7.575 3.60e-14 ***
x1E         -1.089893   0.174935  -6.230 4.66e-10 ***
x2E         -2.709938   0.235541 -11.505  < 2e-16 ***
x3E         -0.045170   0.009686  -4.663 3.11e-06 ***
x4E         -0.307330   0.122754  -2.504   0.0123 *
x5E         -0.241950   0.141691  -1.708   0.0877 .
x6E          0.003776   0.002774   1.361   0.1735
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 834.27  on 623  degrees of freedom
Residual deviance: 559.17  on 617  degrees of freedom
AIC: 573.17

Number of Fisher Scoring iterations: 5
```

TEST (30 % de los datos) -¿267 datos

```
# A tibble: 267 x 12
   PassengerId Survived Pclass Name                Sex      Age SibSp Parch Ticket   Fare Cabin Embarked
         <dbl>    <dbl>  <dbl> <chr>               <chr>  <dbl> <dbl> <dbl> <chr>   <dbl> <chr> <chr>
 1         558        0      1 Robbins, Mr. Victor male    38       0     0 PC 17~ 228.   NA    C
 2         583        0      2 Downton, Mr. William J~ male    54       0     0 28403   26    NA    S
 3         657        0      3 Radeff, Mr. Alexander   male    25       0     0 349223   7.90 NA    S
 4         496        0      3 Yousseff, Mr. Gerious   male    25       0     0 2627    14.5  NA    C
 5         207        0      3 Backstrom, Mr. Karl Al~ male    32       1     0 31012~  15.8  NA    S
 6         592        1      1 Stephenson, Mrs. Walte~ female  52       1     0 36947   78.3  D20   C
 7         394        1      1 Newell, Miss. Marjorie  female  23       1     0 35273  113.   D36   C
 8         631        1      1 Barkworth, Mr. Algerno~ male    80       0     0 27042   30    A23   S
 9         768        0      3 Mangan, Miss. Mary      female  30.5     0     0 364850   7.75 NA    Q
10         827        0      3 Lam, Mr. Len            male    25       0     0 1601    56.5  NA    S
# ... with 257 more rows
```

Creamos las variables de prueba

```
yP = test$Survived
x1P = test$Pclass
x2P = test$Sex
x3P = test$Age
x4P = test$SibSp
x5P = test$Parch
x6P = test$Fare

x2P = (x2P=='male')*1
```

Ecuación de regresión y probamos con las variables de test

```
> yest1
  [1] 1 0 0 0 0 1 1 0 1 0 0 0 0 1 0 1 0 0 1 1 0 1 1 1 0 0 1 0 0 0 0 1 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0
 [52] 0 1 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0
[103] 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 0 1 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 1 1
[154] 1 1 0 0 1 1 1 0 0 0 0 0 1 1 0 0 1 1 0 0 0 1 0 1 1 0 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1
[205] 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0
[256] 1 1 1 0 1 0 0 0 0 0 0 0
> error
  [1] 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0
 [52] 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1
[103] 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
[154] 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 0
[205] 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1
[256] 1 1 0 1 1 1 0 1 1 1 1 1
>
```

Modelo de regresión

```
> accuracy = sum(error)/length(y2)
> accuracy
[1] 0.3067227
>
```

Hombre		La probabilidad estimada de que el pasajero hombre haya sobrevivido es del 25.5 %
Mujer		La probabilidad estimada de que la pasajera mujer haya sobrevivido es del 83.7 %
Pasajero		La probabilidad estimada de que el pasajero sobreviva siendo hombre es de 0.067 veces sobre la posibilidad de que sobreviva siendo mujer. %
Jack		La probabilidad de que Jack haya sobrevivido es de 0.30 %
Jack y Rose		La probabilidad de que Jack haya sobrevivido con Rose 24 %

Cuadro 1: PROBABILIDADES

El modelo es bastante malo, ya que solo tiene una precisión del 30.67 %

Prueba segun ciertos parámetros cambiando los valores de la variable con mayor correlación con Y, para comprobar el odds.

4. Conclusiones

Según nuestro modelo y las pruebas realizadas podemos concluir que la probabilidad de que una mujer haya sobrevivido es superior a la probabilidad de que un hombre haya sobrevivido.
Segun el modelo de regresión del presente deber podemos observar que con estos datos y la herramienta se puede realizar una infinidad de predicciones y exactas.
Segun el modelo de regresión que hace referencia al Titanic podemos observar que con estos datos y la herramienta R se puede realizar predicciones y ver las probabilidades de ciertos eventos que sucedieron en el mismo

Referencias

[1] Métodos de Regresión y Clasificación Lineales Alvaro J. Riascos Villegas Junio de 2019
Clasificación y regresión logística José R. Berrendero Universidad Autónoma de Madrid

[2] Ramsey, J. B. 1969. Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society, Series

[3] Points of Significance: Association, correlation and causation. Naomi Altman & Martin Krzywinski Nature Methods

[4] Barahona, p. (2012). Factores determinantes del rendimiento académico de los estudiantes de la Universidad de Atacama. (Tesis de Maestría). Universidad de Atacama. Chile.

[5] Hernández, D. (2011). Impacto de la didáctica en el rendimiento académico universitario. (Tesis de Licenciatura). Universidad Nacional Federico Villarreal. Perú.

[6] RIAL, A.; VARELA, J. y ROJAS, A. (2001). Depuración y Análisis Preliminares de Datos en SPSS. Sistemas Informatizados para la Investigaci´on del Comportamiento. RA-MA.

https://agenciab12.com/noticia/que-son-regresion-clasificacion-machine-learning