Regresión logística: interpretación constante

Luis Maldonado

Noviembre, 2018

1. La presente guía explica la constante en modelos de regresión logística. Para ello, usamos los datos del titanic. Estimamos un modelo de regresión con *survived* como variable dependiente y *sex* como variable independiente. Creamos dos versiones de la variable sexo:

```
use titanic3.dta

gen sex2=1 if sex==2
replace sex2=0 if sex==1

gen sex3=4 if sex==2
replace sex3=3 if sex==1
```

   Las etiquetas de *sex* son 2 para hombre y 1 para mujer. En el caso de *sex2*, 1=hombre y 0=mujer. Para la variable *sex3*, 4=hombre y 3=mujer.

2. En primer lugar, vamos a calcular los odds o chances para las mujeres. Para ello, hacemos lo siguiente

```
tab survived if sex==1
```

```
   Survived |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        127       27.25       27.25
          1 |        339       72.75      100.00
------------+-----------------------------------
      Total |        466      100.00
```

```
* Probabilidad
dis 339/466
.72746781

* Odds
dis (.72746781/(1-.72746781))
2.6692913

* logit
dis log(2.6692913)
.98181301
```

3. Pues bien sabemos que los odds de sobrevivir para las mujeres son 2,669. **En Stata, podemos obtener dichos odds del siguiente modo**:

```
logit survived sex2, or

Logistic regression                             Number of obs     =      1,309
                                                LR chi2(1)        =     372.92
                                                Prob > chi2       =     0.0000
Log likelihood = -684.05153                     Pseudo R2         =     0.2142

------------------------------------------------------------------------------
    survived | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sex2 |   .0884393   .0120295   -17.83   0.000     .0677432    .1154584
       _cons |   2.669291   .2777073     9.44   0.000     2.176902    3.273054
------------------------------------------------------------------------------
```

Tal como en una regresión OLS, la constante en una regresión logística corresponde al valor de la variable dependiente cuando la variable independiente es cero. En este caso, **la constante refiere a los odds para las mujeres, pues mujer=0**. Además, el coeficiente de *sex2* no denota los odds para los hombres, sino refiere a los odds ratio o *cómo cambian los odds de sobrevivir cuando comparamos hombre con mujeres.*

4. Interpretación:

   - Coeficiente de *sex2*: ser hombre disminuye los odds de sobrevivir por un factor de 0.09. Dicho de otro modo, en comparación con las mujeres, los odds de los hombres de sobrevivir son 91% más bajos.
   - Constante: los odds de sobrevivir para las mujeres son 2,67. También se puede decir del siguiente modo: esparamos encontrar 2,67 sobrevivientes por cada no sobreviviente en el grupo de las mujeres.

5. Sin embargo, el valor de la constante en el caso de una regresión logística depende del modo cómo la variable de sexo esté etiquetada. veamos los siguientes dos casos:

```
* Variable sexo original, con 2=hombre y 1=mujer
logit survived sex, or

Logistic regression                             Number of obs     =      1,309
                                                LR chi2(1)        =     372.92
                                                Prob > chi2       =     0.0000
Log likelihood = -684.05153                     Pseudo R2         =     0.2142

------------------------------------------------------------------------------
    survived | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .0884393   .0120295   -17.83   0.000     .0677432    .1154584
       _cons |   30.18217   6.814284    15.09   0.000     19.38979     46.9816
------------------------------------------------------------------------------

* Version de sexo, con 4=hombre y 3=mujer
logit survived sex3, or

Logistic regression                             Number of obs     =      1,309
                                                LR chi2(1)        =     372.92
                                                Prob > chi2       =     0.0000
Log likelihood = -684.05153                     Pseudo R2         =     0.2142

------------------------------------------------------------------------------
```

```
    survived | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sex3 |   .0884393   .0120295   -17.83   0.000     .0677432    .1154584
       _cons |   3858.863   1899.404    16.78   0.000     1470.559    10125.97
------------------------------------------------------------------------------
```

Note que el odds ratio es el mismo en los dos modelos (0.088), pero la constante cambia. Originalmente, tenemos 2,669 y ahora tenemos 30,182 y 3858,86.

6. ¿Por qué la constante no es la misma? La razón es que el modelo de regresión logística no es lineal. Este es un modelo exponencial (recuerde la formulación matemática de la regresión logística), lo que implica que los valores crecen exponencialmente dependiendo de los valores que tengan los hombre y las mujeres. Para ilustrar:

```
dis exp(0)
1

dis exp(1)
2.7182818

dis exp(2)
7.3890561
```

Este código estima valores exponenciales (anti-logaritmo) para valores de 0, 1 y 2. Tal como puede ver, los valores no son los mismos. Extrapolando este patrón al caso de una regresión logística, no es lo mismo que mujer (categoría de referencia) tenga valores de 0, 1 o 3 para una regresión logística, pues la exponenciación arroja valores distintos.

7. Entonces, **¿cuál es el modelo correcto?** Todos los modelos arrojan el beta correcto, lo que cambia es la constante. El modelo con la constante correcta es aquel que usa sex2 (hombre=1 y mujer=0). También puede obtener las estimaciones correctas con el siguiente modelo

```
logit survived i.sex, or

Logistic regression                             Number of obs     =      1,309
                                                LR chi2(1)        =     372.92
                                                Prob > chi2       =     0.0000
Log likelihood = -684.05153                     Pseudo R2         =     0.2142

------------------------------------------------------------------------------
    survived | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        male |   .0884393   .0120295   -17.83   0.000     .0677432    .1154584
       _cons |   2.669291   .2777073     9.44   0.000     2.176902    3.273054
------------------------------------------------------------------------------
```

8. Para detalles sobre odds y odds ratio ver:

https://www.stata.com/support/faqs/statistics/odds-ratio-versus-odds/