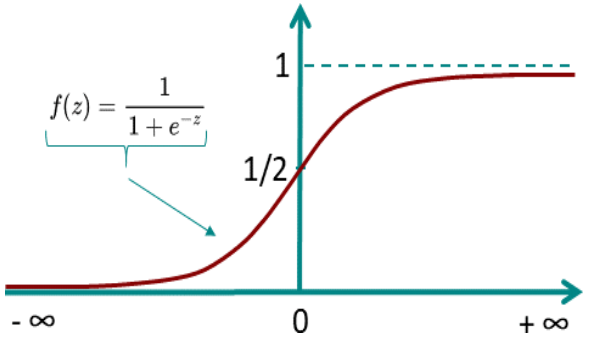
Regresión Logística con Stata

Luis Guillen Grados

Chief Data Science

lguillen@geoanalytics.pe

lguilleng@gmail.com



Fuente: https://datatab.es/tutorial/logistic-regression

medium.com/@lguilleng

linkedin.com/in/datascientistlg

Regresión Logística

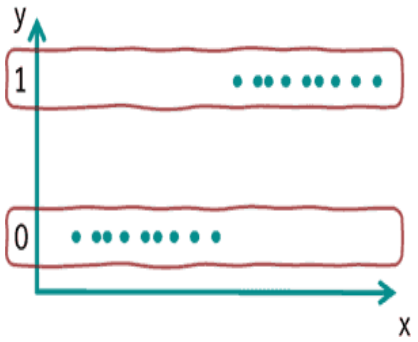
La regresión logística es un caso especial del análisis de regresión y se utiliza cuando la variable dependiente tiene una escala nominal. Este es el caso, por ejemplo, de la variable decisión de compra con los dos valores "compra un producto" y "no compra un producto".

El análisis de regresión logística es, por tanto, la contrapartida de la regresión lineal, en la que la variable dependiente del modelo de regresión debe tener al menos una escala de intervalo.

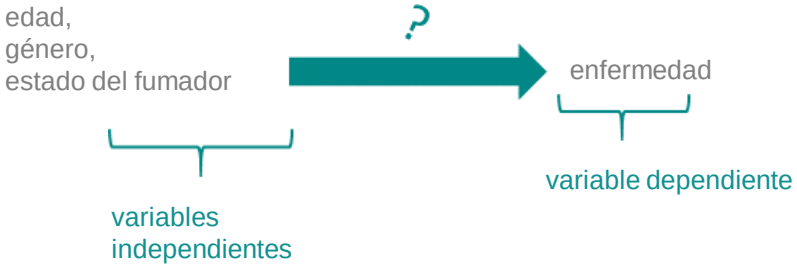
Con la regresión logística, es posible explicar la variable dependiente o estimar la probabilidad de ocurrencia de las categorías de la variable.

¿Qué es una regresión logística?

En la forma básica de la regresión logística, se pueden predecir **variables dicotómicas (0 ó 1)**. Para ello, se estima la probabilidad de que se produzca el **valor 1 (=característica presente)**.



En medicina, por ejemplo, una aplicación frecuente es averiguar qué variables influyen en una enfermedad. En este caso, 0 podría significar "no enfermo" y 1 "enfermo". Posteriormente, se podría examinar la influencia de la edad, el sexo y el hábito de fumar (fumador o no) en esta enfermedad concreta.



Ejemplo empresarial:
Un minorista en línea necesita predecir qué producto es más probable que compre un cliente determinado. Para ello, se recibe un conjunto de datos con visitantes anteriores y sus compras en la tienda online.

Ejemplo médico:
Quiere investigar si una persona es susceptible de contraer una determinada enfermedad o no. Para ello, recibe un conjunto de datos con personas enfermas y no enfermas, así como otros parámetros médicos.

Ejemplo político:
¿Votaría una persona al partido A si hubiera elecciones el próximo fin de semana?

- *Los conceptos presentados de la lámina 2 a la 6, han sido tomados de la publicación de DataTab https://datatab.net/tutorial*
- *La traducción es propia*

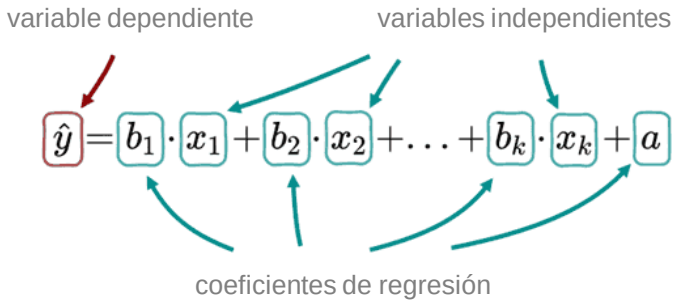
Regresión logística y probabilidades

En la regresión lineal, las variables independientes (por ejemplo, la edad y el sexo) se utilizan para estimar el valor específico de la variable dependiente (por ejemplo, el peso corporal).

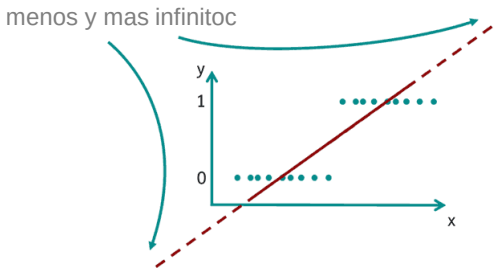
En la regresión logística, en cambio, la variable dependiente es dicotómica (0 ó 1) y se estima la probabilidad de que se dé la expresión 1. Volviendo al ejemplo anterior, esto significa: Qué probabilidad hay de que se presente la enfermedad si la persona considerada tiene una edad, un sexo y un hábito tabáquico determinados.

Calcular la regresión logística

Para construir un modelo de regresión logística, se parte de la ecuación de regresión lineal.



Sin embargo, si simplemente se calculara una regresión lineal para resolver una regresión logística, gráficamente se obtendría el siguiente resultado:

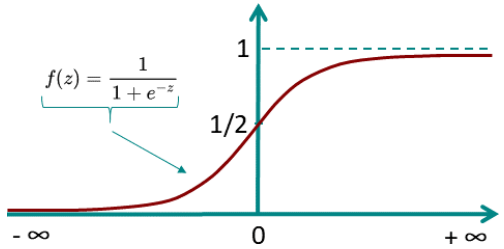


Sin embargo, como puede verse en el gráfico, ahora pueden darse **valores entre más y menos infinito**. El objetivo de la regresión logística, sin embargo, es estimar la probabilidad de ocurrencia y no el valor de la variable en sí. Por lo tanto, la ecuación debe transformarse.

Para ello, es necesario restringir el intervalo de valores para la predicción al intervalo entre 0 y 1. Para garantizar que sólo son posibles valores entre 0 y 1, se utiliza la **función** logística ***f***.

Función logística

El modelo logístico se basa en la función lógica. Lo especial de la función logística es que para valores entre menos y más infinito, siempre asume sólo valores entre 0 y 1.

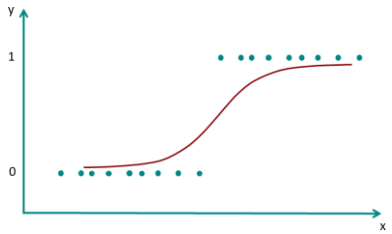


Por tanto, la función logística es perfecta para describir la **probabilidad *P(y=1)*. Si ahora se aplica la función logística a la ecuación de regresión superior, el resultado es**

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Esto asegura ahora que no importa en qué rango se encuentren los valores de x, sólo saldrán números entre 0 y 1. La nueva gráfica tiene ahora este aspecto:



La probabilidad de que para unos valores dados de la variable independiente la variable dependiente dicotómica y sea 0 ó 1 viene dada por

$$P(y=1|x_1,\ldots,x_n) = \frac{1}{1+e^{-(b_1\cdot x_1+\ldots+b_k\cdot x_k+a)}}$$
$$P(y=0|x_1,\ldots,x_n) = 1 - \frac{1}{1+e^{-(b_1\cdot x_1+\ldots+b_k\cdot x_k+a)}}$$

Para calcular la probabilidad de que una persona esté enferma o no mediante la regresión logística del ejemplo anterior, *primero hay que determinar*los parámetros del modelo *b1, b2, b3 y a. Una vez determinados, la ecuación del ejemplo anterior es*

$$P(\text{Diseased}) = \frac{1}{1+e^{-(b_1\text{Age}+b_2\text{Gender}+b_3\text{Smoking status}+a)}}$$

Método de máxima verosimilitud

Para determinar los parámetros del modelo de la **ecuación de regresión** logística, se aplica el método de máxima **verosimilitud**. El método de máxima verosimilitud es uno de los varios métodos utilizados en estadística para estimar los parámetros de un modelo matemático. Otro estimador muy conocido es el método de los mínimos cuadrados, que se utiliza en la regresión logística.

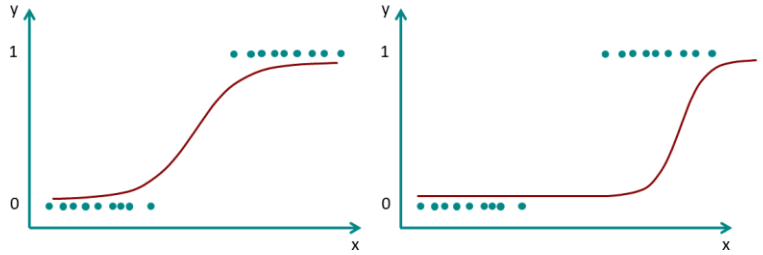
La función de verosimilitud

Para entender el método de **máxima verosimilitud**, introducimos la función de **verosimilitud L**. *L* es una función de los parámetros desconocidos del modelo, que en el caso de la regresión logística son *b1,... bn , a. Por lo tanto, también podemos escribir L(b1,... bn, a) o L(θ) si los parámetros se resumen en θ.*

L (θ) indica ahora la probabilidad de que se produzcan los datos observados. Con el cambio de θ, cambia la probabilidad de que los datos ocurran como se observan.

Con la función logística dada, la probabilidad es "alta" de que ocurran los puntos determinados → El valor de L(θ) el "alto".

Con la función logística dada, la probabilidad es "baja" de que ocurran los puntos determinados → El valor de L(θ) el "bajo".



Estimación de máxima verosimilitud

El **estimador** de máxima verosimilitud puede aplicarse a la estimación de modelos complejos tanto lineales como no lineales. En el caso de la regresión logística, el objetivo es estimar los parámetros *b1,... bn, a, que maximizan la denominada* ***función de verosimilitud logarítmicaLL****(θ). La función de verosimilitud logarítmica es simplemente el logaritmo de L(θ).*

Para esta optimización no lineal, se han establecido diferentes algoritmos a lo largo de los años, como el Stochastic Gradient Descent.

Interpretación de los resultados

La relación entre las variables dependientes e independientes en la regresión logística no es lineal. Por lo tanto, los coeficientes de regresión no pueden interpretarse del mismo modo que en la regresión lineal. Por este motivo, en la **regresión** logística se interpretan **las probabilidades**.

> Regresión lineal:
> Una variable independiente se considera buena si está fuertemente correlacionada con la variable dependiente.
>
> Regresión logística:
> Se dice que una variable independiente es buena si permite distinguir significativamente entre sí los grupos de la variable dependiente.

Las probabilidades se calculan relacionando las dos probabilidades de que y sea "1" y de que y sea "no 1".

$$odds = \frac{p}{1-p}$$

Este cociente puede tomar cualquier valor positivo. Si ahora se logaritma este valor, los valores entre menos y más son infinitamente posibles

$$z = Logit = ln(\frac{p}{1-p})$$

Estas probabilidades logarítmicas suelen denominarse "logits".

Pseudo-R al cuadrado

En una regresión lineal, el coeficiente de determinación R2 indica la proporción de la varianza explicada. En la regresión logística, la variable dependiente se escala nominal u ordinalmente y no es posible calcular una varianza, por lo que el coeficiente de determinación no puede calcularse en la regresión lógica.

Sin embargo, para hacer una afirmación sobre la calidad del **modelo** de regresión logística, se han establecido los denominados pseudocoeficientes de determinación, también llamados pseudo-R al cuadrado. **Los pseudocoeficientes** de determinación se construyen de tal forma que se sitúan entre 0 y 1 al igual que el coeficiente de determinación original. Los coeficientes de determinación más conocidos son el R-cuadrado de **Cox y Snell** y el **R-cuadrado de Nagelkerke**.

Modelo nulo

Para calcular la R-cuadrado de Cox y Snell y la R-cuadrado de Nagelkerke, se necesita la verosimilitud del llamado modelo nulo *L0* y la verosimilitud *L1* del modelo calculado. El modelo cero es un modelo en el que no se incluyen variables independientes, *L1* es la verosimilitud del modelo con las variables dependientes.

R-cuadrado de Cox y Snell

En el **R cuadrado** de Cox y Snell, se compara la relación de la función de verosimilitud del modelo cero *L0* y *L1*. Cuanto mejor sea el modelo completo en comparación con el modelo cero, menor será la relación entre *L0* y *L1. El cuadrado R de Cox y Snell se obtiene con*

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n}$$

R cuadrado de Nagelkerkes

La medida de pseudodeterminación de Cox y Snell no puede llegar a ser uno incluso con un modelo con una predicción perfecta, esto corrige el **R-cuadrado de** Nagelkerkes. El pseudocoeficiente de determinación de Nagelkerkes se convierte en uno si el modelo completo da una predicción perfecta con una probabilidad de 1.

$$R^2_N = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - L_0^{2/n}}$$

Prueba Chi2 y regresión logística

En el caso de la regresión logística, la prueba Chi-cuadrado le indica si el modelo es significativo en su conjunto o no.

Chi-Squared Test

Chi2	df	p
8.79	3	.032

Aquí se comparan dos modelos. En un modelo se utilizan todas las variables independientes y en el otro modelo no se utilizan las variables independientes.

Modelo 1
Con variables independientes

$$\frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Modelo 2
Sin variables independientes

$$\frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Ahora la prueba de Chi-cuadrado compara lo buena que es la predicción cuando se utilizan las variables dependientes y lo buena que es cuando no se utilizan las variables dependientes.

La prueba Chi2 nos dice ahora si hay una diferencia significativa entre estos dos resultados. La hipótesis nula es que ambos modelos son iguales. Si el valor p es inferior a 0,05, se rechaza esta hipótesis nula.

Estimación con STATA

Como ejemplo de **regresión logística**, examinemos el comportamiento de compra en una tienda en línea. El objetivo es determinar los factores de influencia que llevan a una persona a comprar, inmediatamente o más adelante, en la tienda en línea después de visitar el sitio web. La tienda online proporciona los datos recogidos para este fin. Por tanto, la variable dependiente tiene las tres características siguientes:

- Comprar ahora
- Comprar más tarde
- No comprar nada

Como variables independientes se dispone del sexo, la edad y el tiempo pasado en la tienda en línea.

Data: **data_logistica.xlsx**

comportamiento	sexo	edad	tiempo
Comprar ahora	Mujer	22	40
Comprar ahora	Mujer	25	23
Comprar ahora	Hombre	18	12
Comprar ahora	Hombre	45	28
Comprar ahora	Mujer	12	43
Comprar ahora	Hombre	43	23
Comprar ahora	Hombre	23	55
Comprar ahora	Hombre	33	34
Comprar después	Mujer	27	28
Comprar después	Mujer	27	15
Comprar después	Hombre	48	110
Comprar después	Hombre	34	28
Comprar después	Hombre	32	11
Comprar después	Hombre	66	32
Comprar después	Hombre	24	23
Comprar después	Mujer	34	44
No comprar	Mujer	27	55
No comprar	Mujer	34	65
No comprar	Hombre	55	56
No comprar	Hombre	65	80
No comprar	Hombre	44	66
No comprar	Mujer	88	44
No comprar	Mujer	33	65
No comprar	Mujer	43	34

```
. import excel "data_logistica", firstrow

. encode sexo, gen(csexo)
. encode comportamiento, gen(ccomportamiento)

. recode ccomportamiento (1=1) (2=1) (3=0)
. label drop ccomportamiento
. label define c 1 "Compra" 0 "No compra"
. label value ccomportamiento c

. recode csexo (1=1) (2=0)
. label drop csexo
. label define cs 1 "Hombre" 0 "Mujer"
. label value csexo cs

. logit ccomportamiento edad tiempo csexo

Iteration 0:   log likelihood =  -15.27634
Iteration 1:   log likelihood = -8.8670788
…(omitido).
Iteration 5:   log likelihood = -8.0587521

Logistic regression                                     Number of obs =     24
                                                        LR chi2(3)    =  14.44
                                                        Prob > chi2   = 0.0024
Log likelihood = -8.0587521                             Pseudo R2     = 0.4725

------------------------------------------------------------------------------
ccomportamiento | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
----------------+-------------------------------------------------------------
           edad |  -.1331409    .075555    -1.76   0.078     -.281226    .0149442
         tiempo |  -.0526772   .0311538    -1.69   0.091    -.1137375    .0083831
          csexo |   3.987094   2.283466     1.75   0.081    -.4884179    8.462606
          _cons |   6.559921   2.919022     2.25   0.025     .8387427     12.2811
------------------------------------------------------------------------------
```

El proceso de estimación culmina en la iteración 5

Significancia global, por lo menos uno es significativo

No presenta significancia estadística < 0,05

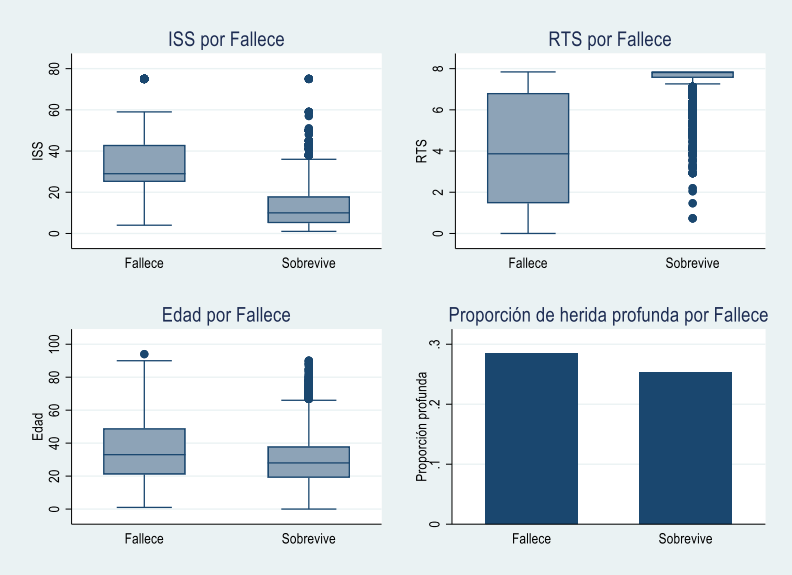
Ejemplo: Los datos de traumatología de la UNM

Los datos que se van a analizar se recogieron de 3 132 pacientes ingresados en el Centro de Traumatología de la Universidad de Nuevo México entre los años 1991 y 1994. Para cada paciente, el médico que lo atendió registró su edad, su puntuación de trauma revisada (RTS), su puntuación de gravedad de la lesión (ISS), si sus lesiones eran contundentes (es decir, el resultado de un accidente de coche: BP=0) o profundas (es decir, heridas de bala: BP=1), y si finalmente sobrevivieron a sus lesiones (MUERTE = 1 si murieron, MUERTE = 0 si sobrevivieron). Aproximadamente el 9% de los pacientes ingresados en el Centro de Traumatología de la UNM mueren finalmente a causa de sus lesiones.

El ISS es un índice global de las lesiones de un paciente, basado en aproximadamente 1 300 lesiones catalogadas en la Escala Abreviada de Lesiones. El ISS puede tomar valores desde 0 para un paciente sin lesiones hasta 75 para un paciente con 3 o más lesiones potencialmente mortales. El ISS es el índice de lesiones estándar utilizado por los centros de traumatología de todo EE.UU. El RTS es un índice de lesiones fisiológicas y se construye como una media ponderada de la presión arterial sistólica, la frecuencia respiratoria y la escala de coma de Glasgow de un paciente. El RTS puede tomar valores desde 0 para un paciente sin signos vitales hasta 7,84 para un paciente con signos vitales normales.

Exploramos cada variable independiente (ISS, EDAD, BP y RTS) vs dependiente (fallece)

```
graph box iss, over(fallece, relabel(1 "Sobrevive" 2 "Fallece" ) descending)
ytitle(ISS) title(ISS por Fallece) name(iss)

graph box rts, over(fallece, relabel(1 "Sobrevive" 2 "Fallece" ) descending)
ytitle(RTS) title(RTS por Fallece) name(rts)

graph box edad, over(fallece, relabel(1 "Sobrevive" 2 "Fallece" ) descending)
ytitle(Edad) title(Edad por Fallece) name(edad)

graph bar bp,over(fallece,relabel(1 "Sobrevive" 2 "Fallece") descending)
ytitle("Proporción profunda") title("Proporción de herida profunda por Fallece")
name(bp)

graph combine iss rts edad bp
```



Significancia estadística para cada coeficiente estimado

Por lo menos uno de los factores es importante par explicar la probabilidad de fallecer

```
. logistic fallece iss bp rts edad, coef

Logistic regression                               Number of obs =   3,132
                                                  LR chi2(4)    = 933.34
                                                  Prob > chi2   = 0.0000
Log likelihood = -446.01414                       Pseudo R2     = 0.5113

------------------------------------------------------------------------------
     fallece | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         iss |   .0651794   .0071603     9.10   0.000     .0511454    .0792135
          bp |   1.001637   .2275466     4.40   0.000     .5556543    1.447621
         rts |  -.8126968   .0537067   -15.13   0.000      -.91796   -.7074335
       edad |    .048616   .0052318     9.29   0.000     .0383619    .0588701
       _cons |  -.5956074   .4344009    -1.37   0.170    -1.447017    .2558028
------------------------------------------------------------------------------
```

$$\text{Log}\left(\frac{\hat{p}}{1-\hat{p}}\right) = -.596 + .065iss + 1.002bp - .813rts + .049edad$$

donde $\hat{p}$ es la probabilidad estimada de fallecer

OR significativos (diferentes de 1)

```
. logistic fallece iss

Logistic regression                               Number of obs =   3,132
                                                  LR chi2(4)    = 933.34
                                                  Prob > chi2   = 0.0000
Log likelihood = -446.01414                       Pseudo R2     = 0.5113

------------------------------------------------------------------------------
     fallece | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         iss |   1.067351   .0076426     9.10   0.000     1.052476    1.082435
          bp |   2.722737   .6195495     4.40   0.000     1.743081    4.252983
         rts |     .44366   .0238275   -15.13   0.000     .3993328    .4929076
       edad |   1.049817   .0054924     9.29   0.000     1.039107    1.060637
       _cons |   .5512277   .2394538    -1.37   0.170     .2352709    1.291498
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

La probabilidad de morir por una lesión profunda (herida de bala) es 2.72 veces mayor que la probabilidad de morir por una lesión contundente

```
. estat gof

Goodness-of-fit test after logistic model
Variable: fallece

      Number of observations =    3,132
Number of covariate patterns =    2,096
          Pearson chi2(2091) = 2039.73
                Prob > chi2 =    0.7849
```

Ho: El modelo se encuentra bien especificado

Ejemplo: Estudio de la eficacia de un programa de aprendizaje

Spector y Mazzeo (1980) utilizaron datos de la tabla siguiente, para analizar si una nueva metodología didáctica resultaba eficaz en la enseñanza de la economía. En su estudio, la variable dependiente es MEJORA, variable que indica si mejoró o no la nota del alumno tras un período de aprendizaje. El resto de variables son CM, media de las calificaciones pasadas del alumno, NP, nota del alumno en un examen previo al período de aprendizaje; y PSI, variable binaria que indica si en el período de aprendizaje el alumno estudió con el nuevo método didáctico o no.

archivo:
eficacia_programa.xlsx

Greene, W. H. (1998). Análisis Econométrico, Tercera Edición Prentice Hall. *Version en Castellano,(1999), Madrid*.

obs	cm	np	psi	mejora
1	2.66	20	0	0
2	2.89	22	0	0
3	3.28	24	0	0
4	2.92	12	0	0
5	4.00	21	0	1
6	2.86	17	0	0
7	2.76	17	0	0
8	2.87	21	0	0
9	3.03	25	0	0
10	3.92	29	0	1
11	2.63	20	0	0
12	3.32	23	0	0
13	3.57	23	0	0
14	3.26	25	0	1
15	3.53	26	0	0
16	2.74	19	0	0
17	2.75	25	0	0
18	2.83	19	0	0
19	3.12	23	1	0
20	3.16	25	1	1
21	2.06	22	1	0
22	3.62	28	1	1
23	2.89	14	1	0
24	3.51	26	1	0
25	3.54	24	1	1
26	2.83	27	1	1
27	3.39	17	1	1
28	2.67	24	1	0
29	3.65	21	1	1
30	4.00	23	1	1
31	3.10	21	1	0
32	2.39	19	1	1

```
. import excel "eficacia_programa", firstrow

. label var cm "media de las calificaciones pasadas del alumno"

. label var np "nota del alumno en un examen previo al periodo de aprendizaje"

. label var psi "1 = estudió con el nuevo método didáctico, 0 = no estudió con
el nuevo método didáctico"

. label var mejora "1 = mejoró la nota del alumno, 0 = no mejoró la nota del
alumno"

. logit mejora cm np psi

Iteration 0:   log likelihood =  -20.59173
Iteration 1:   log likelihood = -13.259769
Iteration 2:   log likelihood = -12.894607
Iteration 3:   log likelihood = -12.889639
Iteration 4:   log likelihood = -12.889634
Iteration 5:   log likelihood = -12.889634

Logistic regression                                     Number of obs =     32
                                                        LR chi2(3)    =  15.40
                                                        Prob > chi2   = 0.0015
Log likelihood = -12.889634                             Pseudo R2     = 0.3740

------------------------------------------------------------------------------
      mejora | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          cm |   2.826113   1.262941     2.24   0.025     .3507936    5.301432
          np |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
         psi |   2.378688   1.064564     2.23   0.025     .2921801    4.465195
       _cons |  -13.02135   4.931324    -2.64   0.008    -22.68656   -3.356129
------------------------------------------------------------------------------

. sum cm np psi mejora

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
          cm |         32    3.117188    .4667128       2.06          4
          np |         32     21.9375    3.901509         12         29
         psi |         32       .4375    .5040161          0          1
      mejora |         32      .34375    .4825587          0          1
```

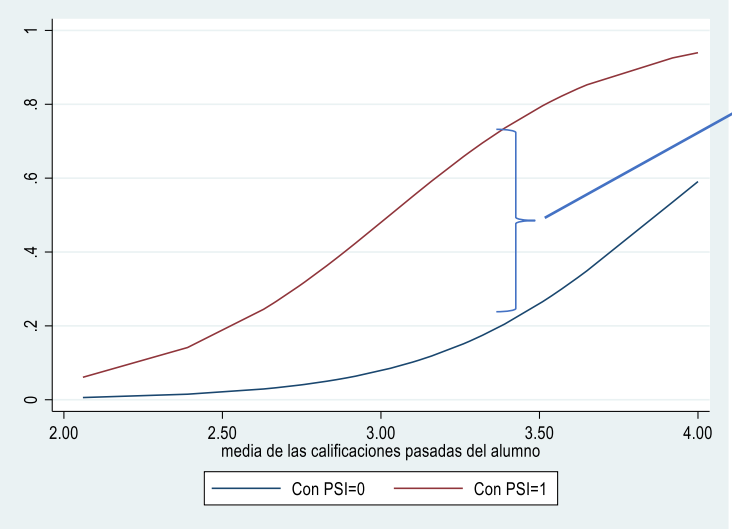
¿Cuál es el efecto marginal del nuevo método didáctico?

Para calcular el efecto marginal, utilizamos los coeficientes del modelo logit estimado. Con estos coeficientes estimamos las probabilidades en función de la variable CM y fijamos el valor de la variable NP igual a su media muestral.

$$PSI = 0; Prob(mejora = 1) = f[-13.021 + 2.826CM + 0.095(21.938)]$$

$$PSI = 1; Prob(mejora = 1) = f[-13.021 + 2.826CM + 0.095(21.938) + 2.379]$$

```
. g pmejora_psi0 = 1/(1+ 1/exp(-13.021 + 2.826*cm + 0.095*21.938))
. g pmejora_psi1 = 1/(1+1/exp(-13.021 + 2.826*cm + 0.095*21.938 +2.379))
. label var pmejora_psi0 "Con PSI=0"
. label var pmejora_psi1 "Con PSI=1"

. twoway (scatter pmejora_psi0 cm, connect(l) msymbol(i)) (scatter
pmejora_psi1 cm, connect(l) msymbol(i))
```



La probabilidad de que la nota de un estudiante mejore tras seguir un curso con la nueva metodología es mucho mayor en alumnos con alta calificación media que en alumnos con baja calificación media

¿Qué tan bueno es el ajuste del modelo?

```
. estat gof

Goodness-of-fit test after logistic model
Variable: mejora

      Number of observations =     32
Number of covariate patterns =     32
            Pearson chi2(28) =  27.26
                 Prob > chi2 = 0.5043
```

Cuando el número de covariables se acerca al número de observaciones la prueba de Pearson no es válida. En su lugar, podemos usar el Hosmer-Lemeshow

El número de grupos debe ser mayor al número de predictores

```
. estat gof, group(10)

note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: mejora

Number of observations =     32
Number of groups =     10
Hosmer-Lemeshow chi2(8) =   7.45
Prob > chi2 = 0.4887
```

Capacidad Predictiva

Se resume la capacidad predictiva de un modelo de regresión logística mediante el concepto de sensibilidad:

$$P(\hat{y} = 1/y = 1)$$

y mediante el concepto de especificidad:

$$P(\hat{y} = 0/y = 0)$$

Es decir, la predicción de éxito cuando es cierto se denomina sensibilidad y la predicción de un fracaso cuando es, a su vez cierto, se denomina especificidad.

Matriz de Confusión

Condición verdadera	Condición estimada: 1	Condición estimada: 0
1	a	b
0	c	d

Sensibilidad a / (a + b)

Especificidad d / (c + d)

PPV a / (a + c)

NPV d / (b + d)

Tasa de VERDADEROS POSITIVOS (TP)

Si lo que nos interesa es evitar falsos negativos, debemos elegir SENSIBILIDAD ALTA.

Tasa de VERDADEROS NEGATIVOS (TN)

Si lo que nos interesa es identificar los verdaderos negativos**, (evitar falsos positivos) debemos elegir ESPECIFIDAD ALTA**.

Sensibilidad: proporción de verdaderos positivos que el modelo o prueba identifica correctamente

Especificidad: proporción de verdaderos negativos correctamente identificados.

PPV: Valor predictivo positivo, es la proporción de resultados positivos con el modelo que están correctamente estimados.

VPN: Valor predictivo negativo, es la proporción de resultados negativos con el modelo que están correctamente estimados.

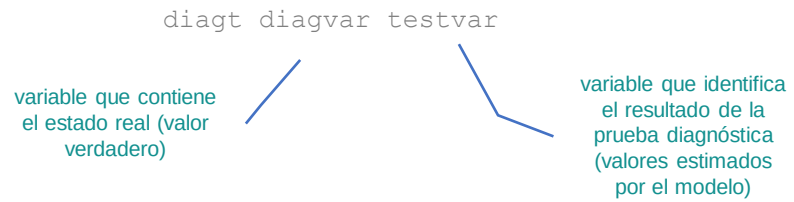
EJEMPLOS

Prueba de dopping en un deportista. Si un falso positivo significa expulsarlo de la competición injustamente, debemos evitar esta situación. → Evitar falsos positivos → ESPECIFIDAD ALTA

Prueba de Covid-19. Si un falso negativo pone en riesgo la vida de la persona, pues no recibe el tratamiento a tiempo, debemos evitar esta situación. → Evitar falsos negativo → SENSIBILIDAD ALTA

Cálculo de la Especificidad y Sensibilidad con Stata

El comando *diagt* muestra varias estadísticas de resumen para las pruebas de diagnóstico: sensibilidad, especificidad y valores predictivos, a partir de una tabla de 2x2. (No es un comando base de Stata. Requiere instalación)




```
. quietly logit mejora cm np psi

. predict prob_mejora
(option pr assumed; Pr(mejora))

. g mejora_fit=prob_mejora>=0.5

. label var mejora_fit "Condición estimada"

. ssc install diagtest

. diagt mejora mejora_fit
```

1 = mejoró la nota del alumno, 0 = no mejoró la nota del alumno	Condición estimada Pos.	Neg.	Total
Abnormal	8	3	11
Normal	3	18	21
Total	11	21	32

True abnormal diagnosis defined as mejora = 1

			[95% Confidence Interval]	
Prevalence	Pr(A)	34.4%	18.6%	53.2%
Sensitivity	Pr(+|A)	72.7%	39.0%	94.0%
Specificity	Pr(-|N)	85.7%	63.7%	97.0%
ROC area	(Sens. + Spec.)/2	0.79	0.63	0.95
Likelihood ratio (+)	Pr(+|A)/Pr(+|N)	5.09	1.68	15.42
Likelihood ratio (-)	Pr(-|A)/Pr(-|N)	0.32	0.12	0.85
Odds ratio	LR(+)/LR(-)	16.00	2.80	91.76
Positive predictive value	Pr(A|+)	72.7%	39.0%	94.0%
Negative predictive value	Pr(N|-)	85.7%	63.7%	97.0%

La sensibilidad y especificidad varían de acuerdo al valor de corte para la clasificación en positivos y negativos (con la condición de interés / sin la condición de interés). Para el cálculo de la sensibilidad y especificidad del cuadro anterior, se utilizó como corte de clasificación el valor 0,5. Si tomamos como punto de corte el valor de 0,3, se obtendría los siguientes resultados:

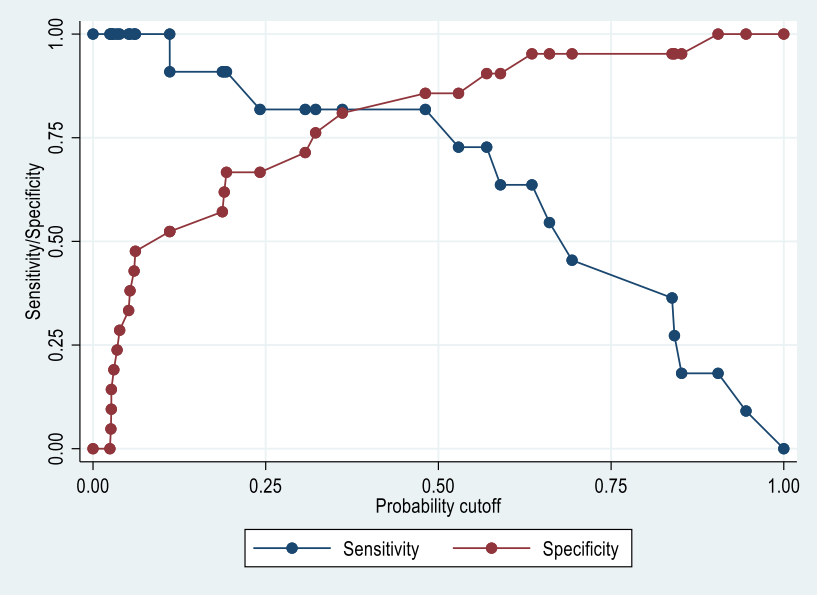
```
. g mejora_fit2=prob_mejora>=0.3

. diagt mejora mejora_fit2, notable
True abnormal diagnosis defined as mejora = 1
```

			[95% Confidence Interval]	
Prevalence	Pr(A)	34.4%	18.6%	53.2%
Sensitivity	Pr(+|A)	81.8%	48.2%	97.7%
Specificity	Pr(-|N)	71.4%	47.8%	88.7%
ROC area	(Sens. + Spec.)/2	0.77	0.61	0.92
Likelihood ratio (+)	Pr(+|A)/Pr(+|N)	2.86	1.38	5.95
Likelihood ratio (-)	Pr(-|A)/Pr(-|N)	0.25	0.07	0.92
Odds ratio	LR(+)/LR(-)	11.25	2.01	60.06
Positive predictive value	Pr(A|+)	60.0%	32.3%	83.7%
Negative predictive value	Pr(N|-)	88.2%	63.6%	98.5%

Evaluar la sensibilidad y especificidad en umbrales (valor de corte) entre 0 y 1

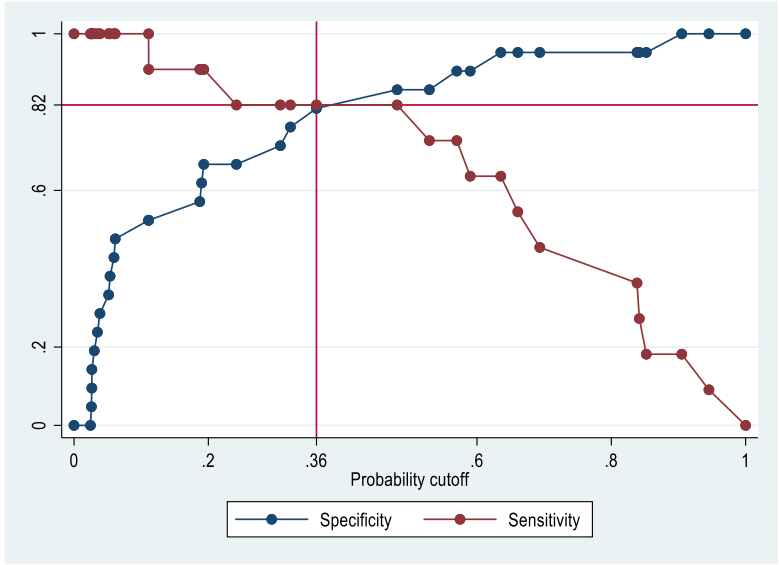
El comando ***lsens*** traza la sensibilidad y la especificidad; traza tanto la sensibilidad como la especificidad frente al límite de probabilidad c.

```
. lsens
```



Algunos investigadores, en lugar de utilizar como punto de corte 0,5, utilizan el valor de corte que iguala a la sensibilidad y la especificidad, es decir, el valor que hace que las líneas del gráfico anterior se intersecten. Para encontrar este umbral (corte) realizamos lo siguiente:

```
# Paso 1: Guardar en variables, los valores de corte, sensibilidad y
especificidad que generan la gráfica anterior

lsens, genprob(corte) gensens(sensibilidad) genspec(especificidad) nograph

# Paso 2: Crear una variable que contenga el valor absoluto de la diferencia
entre sensibilidad y especificidad

gen diferencia = abs(sensibilidad-especificidad)

# Paso 3: Ordenar ascendentemente en base a la variable DIFERENCIA

sort diferencia

# Paso 4: Capturar los valores de las variables CORTE y SENSIBILIDAD que se
encuentran en el primer registro

local x=corte[1]
local y=sensibilidad[1]

# Paso 5: Graficar la SENSIBILIDAD, ESPECIFICIDAD en función a CORTE, añadiendo
las línea vertical "x" y la línea horizontal "y", que corresponden a la
intersección de SENSIBILIDAD y ESPECIFICIDAD.

format corte %11.2gc
format sensibilidad %11.2gc
format especificidad %11.2gc
twoway (scatter especificidad corte, c(l) sort( especificidad corte)) (scatter
sensibilidad corte, c(l) sort(corte sensibilidad ) xline(`x') yline(`y') xlab(0
0.2 `x' 0.6 0.8 1) ylab(0 0.2 `y' 0.6 1))
```



El umbral (corte) donde se intersectan la sensibilidad y la especificidad es 0.36. Ahora utilizamos este valor para la clasificación y calculamos nuevamente la sensibilidad y especificidad.

```
. g mejora_fit3=prob_mejora>=0.36

. diagt mejora mejora_fit3
```

1 = mejoró la nota del alumno, 0 = no mejoró la nota del alumno	mejora_fit3 Pos.	Neg.	Total
Abnormal	9	2	11
Normal	4	17	21
Total	13	19	32

True abnormal diagnosis defined as mejora = 1

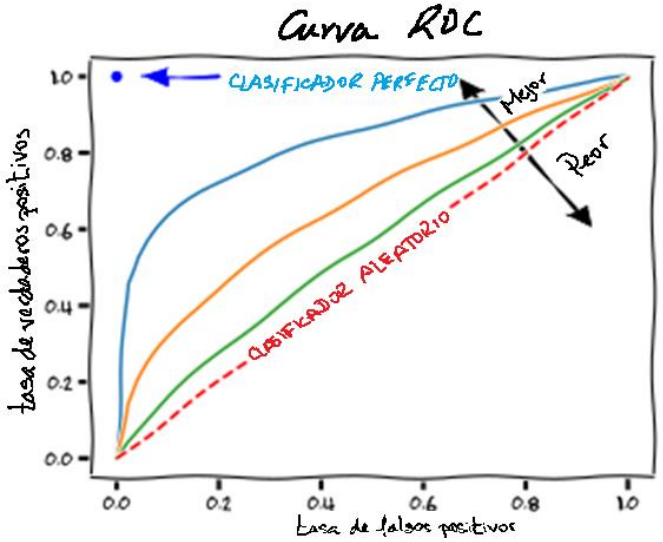
			[95% Confidence Interval]	
Prevalence	Pr(A)	34.4%	18.6%	53.2%
Sensitivity	Pr(+|A)	81.8%	48.2%	97.7%
Specificity	Pr(-|N)	81.0%	58.1%	94.6%
ROC area	(Sens. + Spec.)/2	0.81	0.67	0.96
Likelihood ratio (+)	Pr(+|A)/Pr(+|N)	4.30	1.70	10.83
Likelihood ratio (-)	Pr(-|A)/Pr(-|N)	0.22	0.06	0.80
Odds ratio	LR(+)/LR(-)	19.13	3.15	111.84
Positive predictive value	Pr(A|+)	69.2%	38.6%	90.9%
Negative predictive value	Pr(N|-)	89.5%	66.9%	98.7%

Comparar estos valores con los obtenidos anteriormente

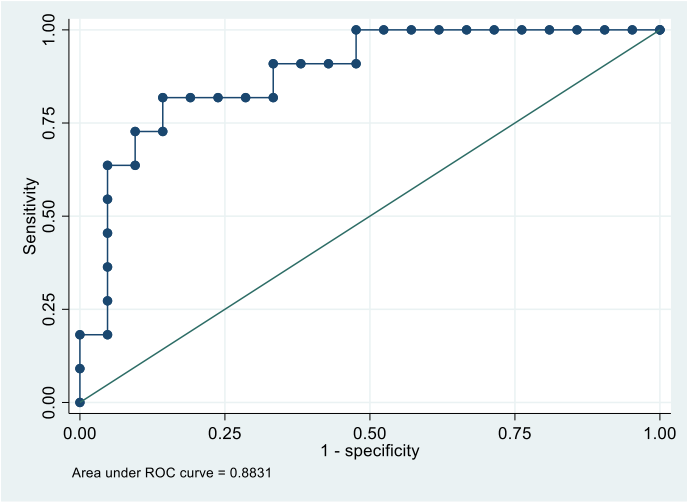
Curva ROC

Una curva de tipo receiver operating characteristic (ROC) es un gráfico en el que se representa la sensibilidad en función de (1– especificidad). Si vamos modificando los valores del valor de corte π0 y representamos la sensibilidad (en ordenadas) frente a (1– especificidad) (en abscisas) tenemos la curva ROC. Es una curva cóncava que conecta los puntos (0,0) y (1,1). Cuanto mayor sea el área bajo la curva mejores serán las predicciones.

La curva ROC ofrece un mejor resumen de la capacidad predictiva que una tabla de clasificación, porque presenta la potencia predictiva para todos los posibles valores de referencia π0. Cuando π0 está cerca de 0 casi todas las predicciones serán igual a 1, con lo cual la sensibilidad estará próxima a 1 y la especificidad estará cerca de 0. Así, el punto (1– especificidad, sensibilidad) tendrá coordenadas (1,1). Cuando π0 está cerca de 1 casi todas las predicciones serán igual a 0, con lo cual la sensibilidad estará próxima a 0 y la especificidad estará cerca de 1. Así, el punto (1–especificidad, sensibilidad) tendrá coordenadas (0,0). Para una especificidad dada (fijando un valor en el eje de abscisas), la mayor potencia predictiva corresponde a la sensibilidad más alta (mayor valor en el eje de ordenadas), de modo que cuanto mayor sea el área bajo la curva ROC mayor será la potencia de predicción.



```
. lroc

Logistic model for mejora

Number of observations =        32
Area under ROC curve   =    0.8831
```



La esquina inferior izquierda de la curva ROC es un umbral de 0 y la esquina superior derecha de 1, los cuales clasificarían todas las observaciones de la misma manera, por lo que no son buenas. Queremos que la curva ROC esté lo más cerca posible de "llenar" el área superior. El Área bajo la curva ROC (llamada AUC) es una medida de ajuste; el ajuste aquí es .8831, que es muy bueno. (El AUC varía de 0,5 a 1. 0,5 indica probabilidad, y 1 indica un ajuste perfecto. Un AUC de 1 es realmente problemático, ya que podríamos estar sobre ajustados) .

El comando estat classification

El comando 'estat classification' muestra varios estadísticos de resumen, incluyendo la tabla de clasificación. ‘estat classification' requiere que los resultados actuales de la estimación provengan de los comandos 'logistic', 'logit', 'probit' o 'ivprobit'."

Este comando permite obtener una tabla que muestra las predicciones del modelo y la precisión de las mismas en la clasificación de las observaciones. Para utilizar este comando, es necesario haber estimado previamente un modelo de regresión logística, logit, probit o ivprobit en Stata.

```
. logit mejora cm np psi

<omitido>
Iteration 5:   log likelihood = -12.889634

Logistic regression                                     Number of obs =     32
                                                        LR chi2(3)    =  15.40
                                                        Prob > chi2   = 0.0015
Log likelihood = -12.889634                             Pseudo R2     = 0.3740

------------------------------------------------------------------------------
      mejora | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          cm |   2.826113   1.262941     2.24   0.025     .3507936    5.301432
          np |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
         psi |   2.378688   1.064564     2.23   0.025     .2921801    4.465195
       _cons |  -13.02135   4.931324    -2.64   0.008    -22.68656   -3.356129
------------------------------------------------------------------------------

. estat classification

Logistic model for mejora

             -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         8             3  |         11
     -     |         3            18  |         21
-----------+--------------------------+-----------
   Total   |        11            21  |         32

Classified + if predicted Pr(D) >= .5
True D defined as mejora != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   72.73%
Specificity                     Pr( -|~D)   85.71%
Positive predictive value       Pr( D| +)   72.73%
Negative predictive value       Pr(~D| -)   85.71%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   14.29%
False - rate for true D         Pr( -| D)   27.27%
False + rate for classified +   Pr(~D| +)   27.27%
False - rate for classified -   Pr( D| -)   14.29%
--------------------------------------------------
Correctly classified                        81.25%
--------------------------------------------------
```

Valor por defecto del punto de corte

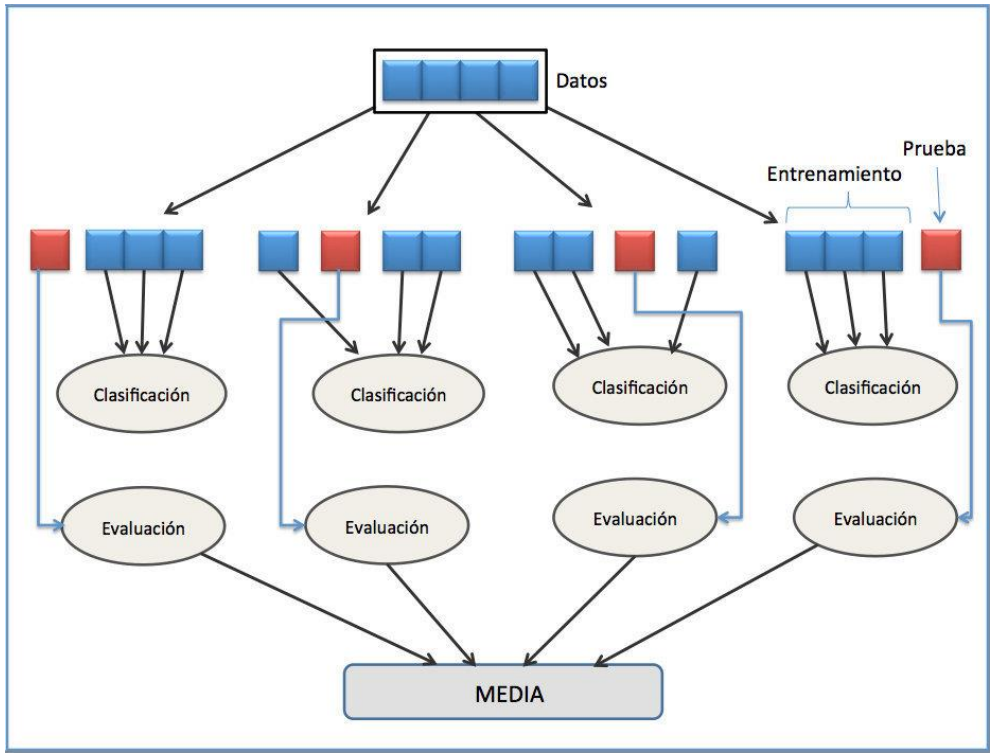
Podemos especificar el punto de corte añadiendo la opción cut(#)

```
. estat classification, cut(0.36)

Logistic model for mejora

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         9             4  |         13
     -     |         2            17  |         19
-----------+--------------------------+-----------
   Total   |        11            21  |         32

Classified + if predicted Pr(D) >= .36
True D defined as mejora != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   81.82%
Specificity                     Pr( -|~D)   80.95%
Positive predictive value       Pr( D| +)   69.23%
Negative predictive value       Pr(~D| -)   89.47%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   19.05%
False - rate for true D         Pr( -| D)   18.18%
False + rate for classified +   Pr(~D| +)   30.77%
False - rate for classified -   Pr( D| -)   10.53%
-------------------------------------------------
Correctly classified                        81.25%
-------------------------------------------------
```

¿Cómo podemos evaluar la capacidad predictiva de un modelo y su capacidad para generalizar a datos nuevos e independientes

En el caso de un modelo logit en STATA, se puede utilizar la validación cruzada para evaluar la calidad de ajuste del modelo y su capacidad para predecir los valores de una variable de respuesta binaria.

La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico cuando el conjunto de datos se ha segmentado en una muestra de entrenamiento y otra de prueba, la validación cruzada comprueba si los resultados del análisis son independientes de la partición.



https://commons.wikimedia.org/w/index.php?curid=17617674)

El comando cvauroc

Un aspecto importante del modelado predictivo (independientemente del tipo de modelo) es la capacidad de un modelo para generalizar a nuevos casos. La evaluación del rendimiento predictivo (AUC = área bajo la curva ROC) de un conjunto de variables independientes utilizando todos los casos de la muestra de análisis original, a menudo da como resultado una estimación excesivamente optimista del rendimiento predictivo. Se puede utilizar la validación cruzada K-fold para generar una estimación más realista del rendimiento predictivo.

cvauroc implementa la validación cruzada de k veces para el AUC para un resultado binario después de ajustar un modelo de regresión logístico o probit promediando las AUC correspondientes a cada pliegue y obteniendo el AUC validado de forma cruzada para obtener inferencia estadística e intervalos de confianza corregidos con sesgo de arranque del 95 % (IC). Además, cvauroc proporciona opcionalmente las probabilidades ajustadas con validación cruzada para la variable dependiente o el resultado contenido en una nueva variable llamada _fit , la sensibilidad y especificidad para cada nivel de la variable dependiente, contenida en dos nuevas variables llamadas _sen y _spe y la gráfica para las curvas medias cvAUC y k-fold ROC.

```
/* Instalación del comando */
ssc install cvauroc
```

Un ejemplo de aplicación

Usaremos un extracto de Cattaneo (2010) que examina a 4,642 bebés nacidos de forma individual en Pennsylvania entre 1989 y 1991. Nuestro objetivo es estimar la probabilidad de tener un bebé con bajo peso al nacer (lbw). Ajustaremos un modelo de regresión logística utilizando el estado civil de la madre (mmarried), la edad de la madre (mage), la educación de la madre (medu), la raza de la madre (mrace), la educación del padre (fedu), el comportamiento de fumar de la madre (mbsmoke), si la madre tuvo una visita prenatal en el primer trimestre del bebé (prenatal1) y si el bebé es el primer hijo de la madre (fbaby) como predictores independientes para lbw. Luego, para entender la capacidad predictiva de nuestro modelo elegido, calcularemos el AUC utilizando el enfoque naive clásico, basado en las probabilidades ajustadas del modelo, y lo compararemos con el AUC de la estrategia de validación interna implementada con el paquete estadístico cvauroc.

```
/* Consideramos que un bajo peso al nacer es menor a 2500 */
use peso_nacer, clear
generate lbw = cond(bweight<2500,1,0)

/* Estiamos la probabilidad de tener un bebé con peso bajo al nacer */
quietly logistic lbw mage medu mmarried prenatal1 fedu mbsmoke mrace fbaby

/* Estimamos las probabilidades para obtener el valor de AUC del modelo */
predict fitted, pr

/* Calculamos el AUC */
roctab lbw fitted
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
4,642	0.6847	0.0172	0.65095	0.71848

```
/* Realizamos una validación cruzada */
. cvauroc lbw mage medu mmarried prenatal1 fedu mbsmoke mrace fbaby,seed(3489)
kfold(10)

1-fold (N=465).........AUC =  0.607
2-fold (N=464).........AUC =  0.700
3-fold (N=464).........AUC =  0.686
4-fold (N=464).........AUC =  0.724
5-fold (N=464).........AUC =  0.669
6-fold (N=465).........AUC =  0.689
7-fold (N=464).........AUC =  0.759
8-fold (N=464).........AUC =  0.653
9-fold (N=464).........AUC =  0.659
10-fold (N=464).........AUC =  0.616

Model:logistic

Seed:3489

------------------------------------------------------------------------
Cross-validated (cv) mean AUC, SD and Bootstrap Bias Corrected 95%CI
------------------------------------------------------------------------
cvMean AUC:                        | 0.6763
Bootstrap bias corrected 95%CI:    | 0.6431, 0.7172
cvSD AUC:                          | 0.0461
------------------------------------------------------------------------
```

La validación cruzada nos da un valor promedio para AUC de 0.6763, que es menor 0.6847 obtenido con el modelo total.

Siendo una diferencia pequeña, nos estaría indicando que, que la capacidad del modelo para generalizar a nuevos casos es prácticamente la misma.

Aplicación a la ENAHO

Supongamos que necesitamos analizar cuanto influye el gasto per cápita mensual del hogar, el ingreso per cápita mensual del hogar, la cantidad de miembros por hogar, la cantidad de perceptores de ingreso y, la pertenencia al área urbana o rural, en la condición del hogar de ser pobre.

Para analizar lo solicitado, haremos uso del archivo sumaria-2021.dta, el cual contiene variables calculadas de gasto e ingreso de los hogares, la cantidad de miembros y perceptores de ingresos del hogar y, la condición de pobreza del hogar.

```
. sum mieperho percepho gashog2d inghog2d

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    mieperho |     34,245    3.309271    1.760757          1         15
    percepho |     34,245    2.139057     1.03659          0         10
    gashog2d |     34,245    25601.36    20001.26   540.3101   535357.4
    inghog2d |     34,245    32563.53    31418.15          0   698107.7

. tab pobreza

         pobreza |      Freq.     Percent        Cum.
-----------------+-----------------------------------
   pobre extremo |      1,130        3.30        3.30
pobre no extremo |      5,210       15.21       18.51
        no pobre |     27,905       81.49      100.00
-----------------+-----------------------------------
           Total |     34,245      100.00
```

Las variables gashog2d e inghog2d son gastos e ingresos totales anuales del hogar. Para obtener un aproximado del gasto e ingreso mensual, estas variables deben ser divididas entre 12 y, para obtener sus valor per cápita, se debe dividir entre la cantidad de miembros del hogar.

```
. g gasmenp=gashog2d/12/mieperho
. g ingmenp=inghog2d/12/mieperho
. sum gasmenp ingmenp

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     gasmenp |     34,245     746.436    641.1798   45.02584   17060.33
     ingmenp |     34,245    949.6719    1027.954          0   28911.83
```

La condición de hogar pobre se obtiene en base a la variable pobreza.

```
. g pobre=pobreza<=2
. label define pobre 1 "Pobre" 0 "No pobre"
. label value pobre pobre
. tab pobre

      pobre |      Freq.     Percent        Cum.
------------+-----------------------------------
   No pobre |     27,905       81.49       81.49
      Pobre |      6,340       18.51      100.00
------------+-----------------------------------
      Total |     34,245      100.00
```

La condición del hogar de pertenecer a un área urbana o rural se obtiene en base a la variable estrato.

```
. g area=estrato<=5
. label define area 1 "urbano" 0 "rural"
. label value area area
. tab area

       area |      Freq.     Percent        Cum.
------------+-----------------------------------
      rural |     12,170       35.54       35.54
     urbano |     22,075       64.46      100.00
------------+-----------------------------------
      Total |     34,245      100.00
```

Estimando el modelo logit

```
. svyset conglome [pw= factor07 ], strata(estrato)

Sampling weights: factor07
             VCE: linearized
     Single unit: missing
        Strata 1: estrato
 Sampling unit 1: conglome
           FPC 1: <zero>
```

```
. svy: logit pobre ingmenp gasmenp mieperho percepho area

Survey: Logistic regression

Number of strata =     8                        Number of obs    =     34,245
Number of PSUs   = 5,359                        Population size = 9,903,824
                                                 Design df        =     5,351
                                                 F(5, 5347)       =    298.14
                                                 Prob > F         =    0.0000
------------------------------------------------------------------------------
             |             Linearized
       pobre | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ingmenp |   .0001054    .000122     0.86   0.388    -.0001338    .0003446
     gasmenp |  -.0371844   .0011069   -33.59   0.000    -.0393543   -.0350144
    mieperho |   .2184443    .033986     6.43   0.000     .151818    .2850707
    percepho |  -.0021428   .0603408    -0.04   0.972    -.1204354    .1161498
        area |    4.00611   .1156044    34.65   0.000     3.779478    4.232742
       _cons |   9.567089   .3058645    31.28   0.000      8.96747    10.16671
------------------------------------------------------------------------------
```

Cambios marginales para variables discretas (DERIVADA)

```
. margins, dydx(*)

Average marginal effects

Number of strata =     8                        Number of obs    =     34,245
Number of PSUs   = 5,359                        Population size = 9,903,824
Model VCE: Linearized                            Design df        =     5,351

Expression: Pr(pobre), predict()
dy/dx wrt:  ingmenp gasmenp mieperho percepho area

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ingmenp |   3.73e-06   4.32e-06     0.87   0.387    -4.72e-06    .0000122
     gasmenp |  -.0013177   7.16e-06  -184.08   0.000    -.0013317   -.0013036
    mieperho |   .0077407   .0011834     6.54   0.000     .0054207    .0100607
    percepho |  -.0000759   .0021382    -0.04   0.972    -.0042677    .0041159
        area |   .1419596   .0036628    38.76   0.000     .1347791    .1491401
------------------------------------------------------------------------------
```

Si el hogar es urbano, tiene un 14% mas de probabilidad de ser pobre

Cambios marginales para variables continuas (ELASTICIDAD)

```
. margins, eyex(*)

Average marginal effects

Number of strata =     8                        Number of obs    =     34,245
Number of PSUs   = 5,359                        Population size = 9,903,824
Model VCE: Linearized                            Design df        =     5,351

Expression: Pr(pobre), predict()
ey/ex wrt:  ingmenp gasmenp mieperho percepho area

------------------------------------------------------------------------------
             |            Delta-method
             |      ey/ex   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ingmenp |   .0933414   .1080304     0.86   0.388    -.1184422     .305125
     gasmenp |  -26.47884   .8059897   -32.85   0.000     -28.0589   -24.89877
    mieperho |   .5439836   .0845053     6.44   0.000     .3783189    .7096484
    percepho |  -.0036471   .1027078    -0.04   0.972    -.2049962     .197702
        area |   2.614901   .0740299    35.32   0.000     2.469772     2.76003
------------------------------------------------------------------------------
```

Si la cantidad de miembros por hogar aumenta en 1% la probabilidad de ser un hogar pobre aumenta en 54%