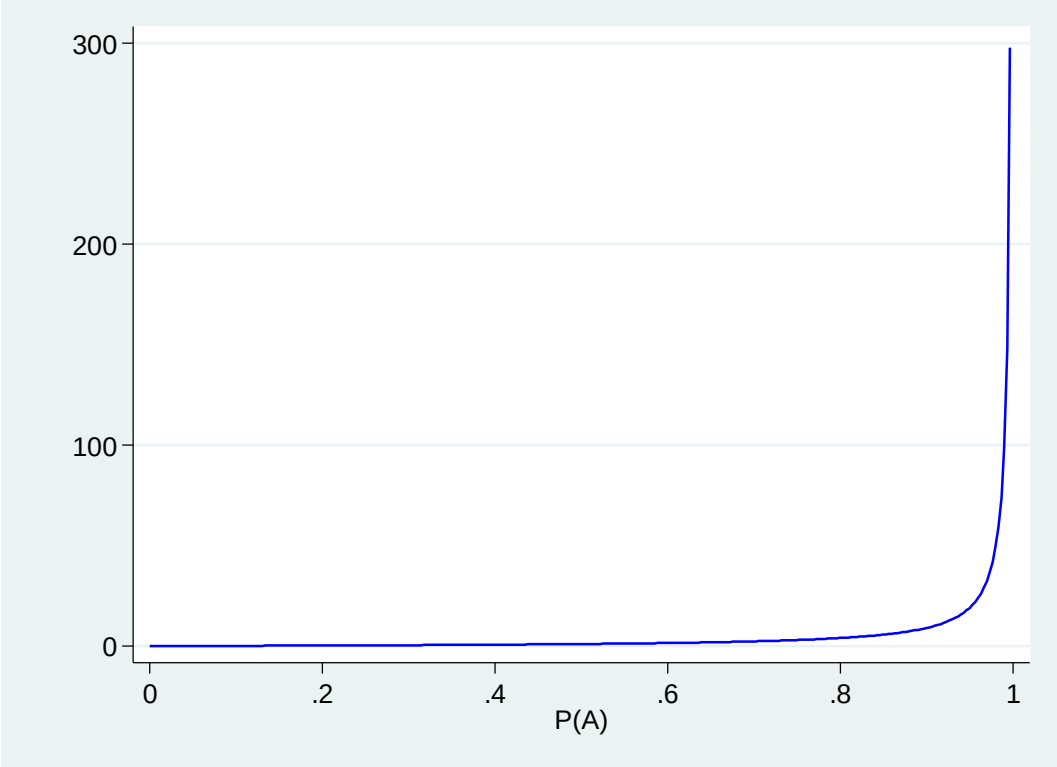
Regresión Logística

La distribución de probabilidades logística

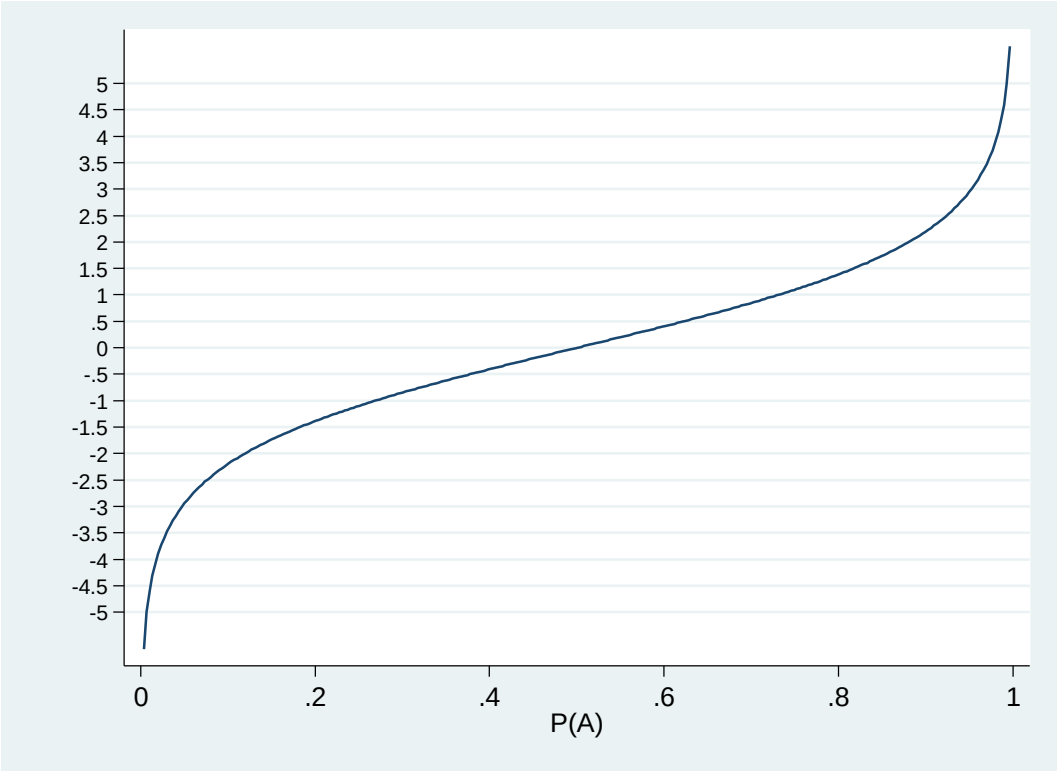
Supongamos que estamos interesados en la ocurrencia de un evento "A", cuya probabilidad de aparición es "P", es decir: $P(A) = P$ y por consiguiente la probabilidad de que "A" no ocurra es $P(A') = 1 - P$, sin embargo sabemos que la ocurrencia de A, y por ende su probabilidad, está relacionada con el valor que tome una variable aleatoria X, esto es $P(A) = P(X \leq x)$: por ejemplo si A: una persona muere y X es la edad de la persona, es razonable pensar que $P(morir) = P(Edad \leq edad)$. Notar que $P(A) = F(X)$, donde F(X) es la función de distribución de probabilidades de X. El problema fundamental es como relacionar la probabilidad de la aparición del evento "A" , con los posibles valores de la variable X.

Luego ¿Cómo hacer para que la P(A) dependa linealmente de X?, la respuesta directa a este problema sería proponer: $P(A) = \alpha + \beta \cdot X$, sin embargo esta propuesta no es satisfactoria ya que $P(A) \in [0,1]$ y la función lineal puede tomar cualquier valor real. Si deseamos perseverar en la asociación lineal de la P(A) con X, debemos pensar en una transformación de P(A) que garantice que tome valores en todos los reales. Las propuestas que resuelven el problema son muchas, sin embargo la más útil es la siguiente:

- Si consideramos el Odds del evento A, es decir $Odds(A) = \frac{P(A)}{1-P(A)}$ y lo evaluamos para todos los posibles valores de P(A), obtenemos la siguiente función:



Observamos, como es sabido que el Odds, puede tomar cualquier valor real positivo, esto nos ilumina a considerar el logaritmo del Odds, ya que la función logaritmo tiene dominio en los reales positivos pero su recorrido son todos los reales, como se observa en el siguiente gráfico:



- Así entonces proponemos la relación:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta \cdot X$$

Que nos lleva a :

$$P(A) = F(X) = \frac{e^{\alpha+\beta\cdot X}}{1+e^{\alpha+\beta\cdot X}}$$

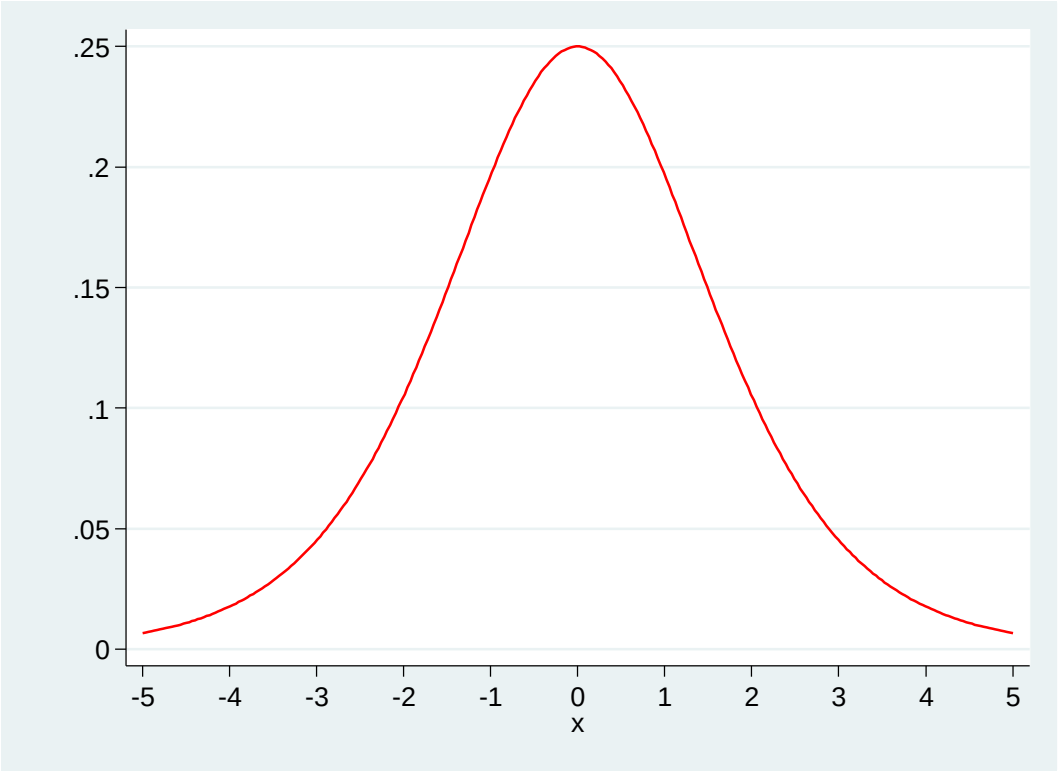
De donde deducimos que la función densidad de probabilidades es:

$$f(X) = \frac{\beta e^{\alpha+\beta\cdot X}}{(1+e^{\alpha+\beta\cdot X})^2}$$

Particularmente si consideramos α=0 y β=1, la función densidad de probabilidades es:

$$f(X) = \frac{e^{X}}{(1+e^{X})^2}$$

Cuyo gráfico es el siguiente:



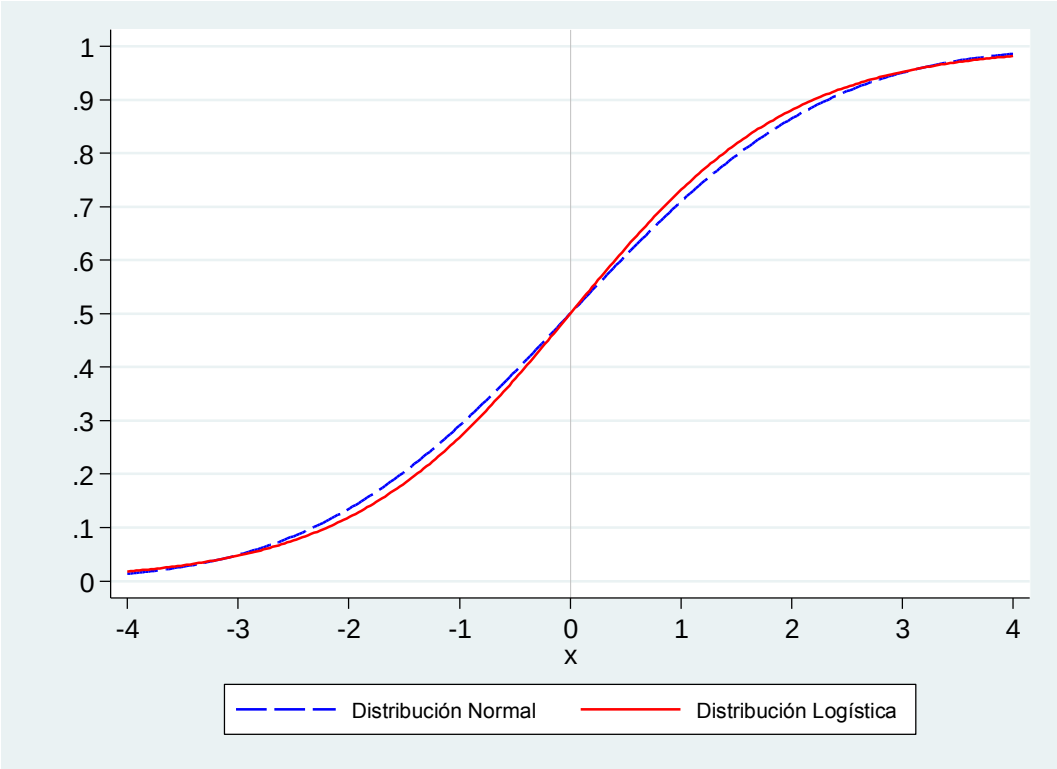
La esperanza y la varianza de la distribución logística estándar son respectivamente:

$$E[X] = 0$$
$$Var[X] = \frac{\pi^2}{3}$$

En consecuencia para la distribución logística de parámetros α y β se tiene:

$$E[X] = \alpha$$
$$Var[X] = \frac{(\beta\pi)^2}{3}$$

Usando estos resultados se encuentra un hecho sorprendente: la función de distribución de la logística estándar, difiere muy poco con la función de distribución de la $N(0, \frac{\pi^2}{3})$, como lo muestra el siguiente gráfico:



Para la distribución logística estándar se verifica:

- $1 - F(X) = \frac{1}{1+e^X}$
- $f(X) = F(X)[1 - F(X)]$

La regresión logística

Nos interesa modelar la aparición de un evento, A, explicándolo por un perfil definido como una combinación lineal de variables:

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p$$

La respuesta la codificamos de la siguiente forma:

$$Y = \begin{cases} 1, si\ el\ evento\ A\ aparece \\ 0, si\ el\ evento\ A\ no\ aparce \end{cases}$$

Definiendo $P(Y = 1|X) = P(A) = \pi(X)$, es claro que la distribución de probabilidades de Y es Bernoullí con probabilidad de éxito $\pi(X)$, es decir la función de cuantía de probabilidades es:

$$P(Y = y) = (1 - \pi(X))^{1-y}\pi(X)^y, con\ y = 0{,}1$$

Al asumir que $\pi(X) = F(X)$ donde F(X) es la función de distribución logística evaluada en el perfil Xβ, la cuantía de probabilidades de Bernoullí se puede escribir como:

$$P(Y = y \mid X) = \left(\frac{1}{1 + e^{X\beta}}\right)^{1-y} \left(\frac{e^{X\beta}}{1 + e^{X\beta}}\right)^{y}, con\ y = 0{,}1$$

Por lo tanto si se tiene una muestra aleatoria de "n" perfiles asociados a sus respectivas respuestas "y", la función de verosimilitud que estima los parámetros **β** del modelo es:

$$L = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{X_i\beta}}\right)^{1-y_i} \left(\frac{e^{X_i\beta}}{1 + e^{X_i\beta}}\right)^{y_i}, con\ y_i = 0{,}1$$

Esta función de verosimilitud corresponde al modelo logístico de respuesta binaria. Los parámetros hay que estimarlos mediante el método iterativo de Newton-Raphson, como se revisó en el capítulo I.

Como se estableció anteriormente:

$$ln\left(\frac{P(Y = 1\ |X)}{1 - P(Y = 1\ |X)}\right) = ln(Odds(Y = 1|X\beta)) = X\beta$$

Esta relación permite comparar dos perfiles: **X** y **X'** pues al evaluar la expresión anterior en cada uno de estos perfiles y luego restar estas ecuaciones se obtiene:

$$ln(Odds(Y = 1|X)) = X\beta$$
$$ln\big(Odds(Y = 1|X')\big) = X'\beta$$

$$\Downarrow$$

$$ln(Odds(Y = 1|X)) - ln(Odds(Y = 1|X')) = X\beta - X'\beta = (X - X')\beta$$

O equivalentemente:

$$ln\left(\frac{Odds(Y = 1|X)}{Odds(Y = 1|X')}\right) = \ln\,(OR) = X\beta - X'\beta = (X - X')\beta$$

Por lo tanto β, es el cambio del *ln(OR)* por cambio de perfil, de donde se deduce que:

$$OR = e^{(X-X')\beta}$$

Si X es una variable dicotómica, por ejemplo X=1 y X=0 denoten exposición y no exposición respectivamente, la expresión del OR es:

$$OR = e^{(X-X')\beta} = e^{(1-0)\beta} = e^{\beta}$$

Cuya interpretación ya es conocida.

La novedad es que si X es una variable contínua y comparamos el perfil X con el perfil X+1, la expresión que define el OR entre perfiles es:

$$OR = e^{(X-X')\beta} = e^{(X+1-X)\beta} = e^{\beta}$$

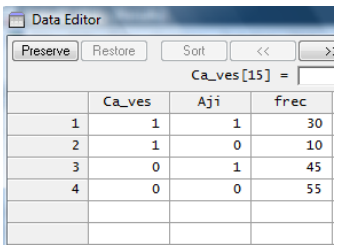
Que representa el cambio de riesgo cuando la variable X se incrementa en "una unidad".

STATA provee dos comandos equivalentes para estimar un modelo de regresión logística binaria:

1) logit respuesta variables, que reporta los β estimados
2) logistic respuesta variables, que reporta los respectivos OR

Ejemplo 1: Estimar la fuerza de la asociación en la siguiente tabla:

	Cáncer de vesícula	Control	
Consumo ají rojo	30	45	75
No consumo de ají rojo	10	55	65
	40	100	140



	Ca_ves	Aji	frec
1	1	1	30
2	1	0	10
3	0	1	45
4	0	0	55

```
. cc Ca_ves Aji [freq= frec]
                                                              Proportion
                 |   Exposed   Unexposed  |      Total      Exposed
-----------------+------------------------+------------------------
           Cases |        30          10  |         40       0.7500
        Controls |        45          55  |        100       0.4500
-----------------+------------------------+------------------------
           Total |        75          65  |        140       0.5357
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |          3.666667      |    1.528998    9.270422 (exact)
 Attr. frac. ex. |          .7272727      |    .3459769    .8921301 (exact)
 Attr. frac. pop |          .5454545      |
                 +-------------------------------------------------
                               chi2(1) =    10.34  Pr>chi2 = 0.0013


. logistic Ca_ves Aji [freq= frec]

Logistic regression                               Number of obs   =        140
                                                  LR chi2(1)      =      10.75
                                                  Prob > chi2     =     0.0010
Log likelihood = -78.381871                       Pseudo R2       =     0.0642

------------------------------------------------------------------------------
      Ca_ves | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Aji |   3.666667   1.528328     3.12   0.002     1.619856    8.299778
------------------------------------------------------------------------------
. logit Ca_ves Aji [freq= frec]

Iteration 0:   log likelihood = -83.757742
Iteration 1:   log likelihood = -78.492222
Iteration 2:   log likelihood = -78.382053
Iteration 3:   log likelihood = -78.381871

Logistic regression                               Number of obs   =        140
                                                  LR chi2(1)      =      10.75
                                                  Prob > chi2     =     0.0010
Log likelihood = -78.381871                       Pseudo R2       =     0.0642

------------------------------------------------------------------------------
      Ca_ves |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Aji |   1.299283   .4168167     3.12   0.002     .4823373    2.116229
       _cons |  -1.704748   .3437741    -4.96   0.000    -2.378533   -1.030963
------------------------------------------------------------------------------

. dis exp( 1.299283)
3.6666667
```

Ejemplo 2: Estimar la fuerza de la asociación de la glicemia con la mortalidad intrahospitalaria por IAM (en la base de datos glicemiamorintra.dta)

```
. logistic  morintra glicemia

Logistic regression                               Number of obs   =        350
                                                  LR chi2(1)      =      14.68
                                                  Prob > chi2     =     0.0001
Log likelihood = -112.89033                       Pseudo R2       =     0.0611

------------------------------------------------------------------------------
    morintra | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    glicemia |   1.005852   .0015213     3.86   0.000     1.002875    1.008838
------------------------------------------------------------------------------
```

El efecto sexo:

```
. logistic  morintra sexo

Logistic regression                               Number of obs   =        350
                                                  LR chi2(1)      =       7.76
                                                  Prob > chi2     =     0.0053
Log likelihood = -116.35056                       Pseudo R2       =     0.0323

------------------------------------------------------------------------------
    morintra | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |        2.7   .9455727     2.84   0.005     1.359135    5.363705
------------------------------------------------------------------------------
```

El efecto glicemia ajustando por sexo:

```
. logistic  morintra  glicemia sexo

Logistic regression                               Number of obs   =        350
                                                  LR chi2(2)      =      21.44
                                                  Prob > chi2     =     0.0000
Log likelihood = -109.51338                       Pseudo R2       =     0.0891

------------------------------------------------------------------------------
    morintra | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    glicemia |   1.005747   .0015373     3.75   0.000     1.002738    1.008764
        sexo |   2.593596   .9369891     2.64   0.008     1.277581    5.265215
------------------------------------------------------------------------------
```

Evaluación de la posible interacción de la glicemia con sexo:

```
. logistic  morintra  glicemia sexo  glic_sexo

Logistic regression                               Number of obs   =        350
                                                  LR chi2(3)      =      21.73
                                                  Prob > chi2     =     0.0001
Log likelihood = -109.36633                       Pseudo R2       =     0.0904

------------------------------------------------------------------------------
    morintra | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    glicemia |   1.006353   .0019239     3.31   0.001     1.002589     1.01013
        sexo |   3.702559   2.767778     1.75   0.080     .8554656    16.02513
   glic_sexo |   .9982576   .0032058    -0.54   0.587     .9919941    1.004561
------------------------------------------------------------------------------
```

Ejemplo 3: Se aleatorizan 150 pacientes a tres tratamientos, cuya respuesta es dicotómica: mejora o no mejora, si el tratamiento 1 es el de referencia, comparar los tratamientos 2 y 3. (trestrat.dta)

```
. tab  trat mejora, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |    0: no mejora 1:
           |        mejora
      trat |         0          1 |     Total
-----------+----------------------+----------
         1 |        19         31 |        50
           |     38.00      62.00 |    100.00
-----------+----------------------+----------
         2 |        13         37 |        50
           |     26.00      74.00 |    100.00
-----------+----------------------+----------
         3 |        10         40 |        50
           |     20.00      80.00 |    100.00
-----------+----------------------+----------
     Total |        42        108 |       150
           |     28.00      72.00 |    100.00

. tab trat, gen(Trat)

       trat |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         50       33.33       33.33
          2 |         50       33.33       66.67
          3 |         50       33.33      100.00
------------+-----------------------------------
      Total |        150      100.00

. logistic   mejora Trat2 Trat3, nolog

Logistic regression                               Number of obs   =        150
                                                  LR chi2(2)      =       4.13
                                                  Prob > chi2     =     0.1266
Log likelihood = -86.876173                       Pseudo R2       =     0.0232

------------------------------------------------------------------------------
      mejora | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Trat2 |   1.744417   .7580485     1.28   0.200     .7443115    4.088329
       Trat3 |   2.451613   1.123175     1.96   0.050     .9988146    6.017539
------------------------------------------------------------------------------
```

Regresión logística en diagnóstico y pronóstico

Desde el punto de vista estadístico, tanto el análisis diagnóstico como el pronóstico, se inscriben en el llamado análisis discriminante, esto es dadas dos muestras pertenecientes a poblaciones distintas y conocida esta pertenencia, determinar el conjunto de variables "descriptoras" (perfil del sujeto) que tiene capacidad de identificar cada una de las poblaciones a las que se hace referencia. Si la pertenencia a cada una de las poblaciones en cuestión la denotamos por los códigos "0 y 1" contenidos en la variable "Y", para el perfil Xβ, podemos escribir:

$$P(Y = 1|X) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

Así, dado un perfil X, la idea es determinar si la estimación de probabilidad hecha por la distribución logística está cerca del 0 o cerca del 1. Esta decisión es posible tomarla si se escoge un punto de corte, "p" para la probabilidad estimada, de modo que si $P(Y = 1|X\beta) > p$, el sujeto será clasificado en la población "1", de lo contrario el sujeto será clasificado en la población "0". Con esta conceptualización, si la población de interés (enfermos, muertos, mejorías…) la llamamos "A" y la codificamos con "1", definimos:

S=P(Y=1 | X ϵ A) : Sensibilidad de la discriminación .
E=P(Y=0 | X ϵ A'): Especificidad de la discriminación.

Obviamente que las probabilidades complementarias definen los sucesos "Falso Negativo" y "Falso Positivo", respectivamente, es decir:

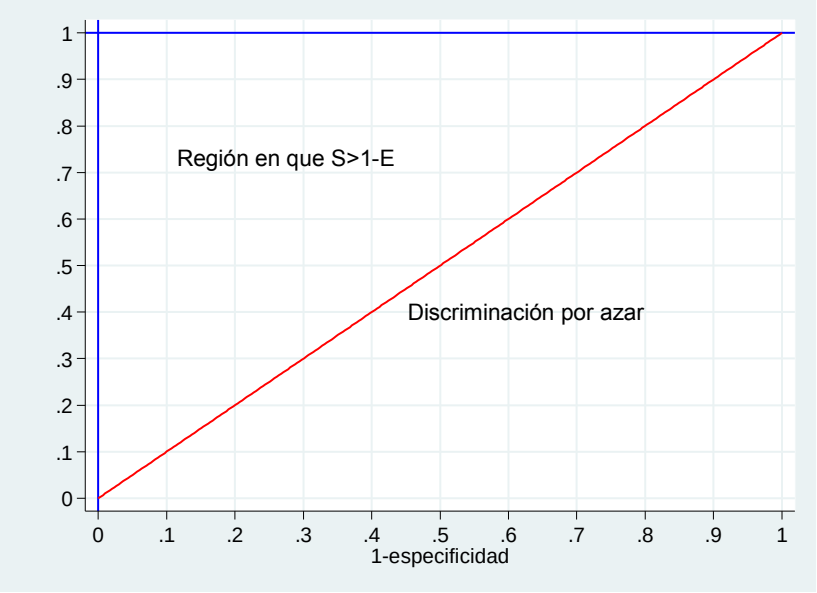
P(Y=0 | X ϵ A) = 1-S: Falso Negativo. (1-S)
P(Y=1 | X ϵ A') =1-E: Falso Positivo. (1-E)

Como nuestro interés es tener buena capacidad de clasificación, es decir alta sensibilidad y alta especificidad. Lo que se traduce en:

P(Y=1 | X ϵ A)> P(Y=1 | X ϵ A')

S>1-E

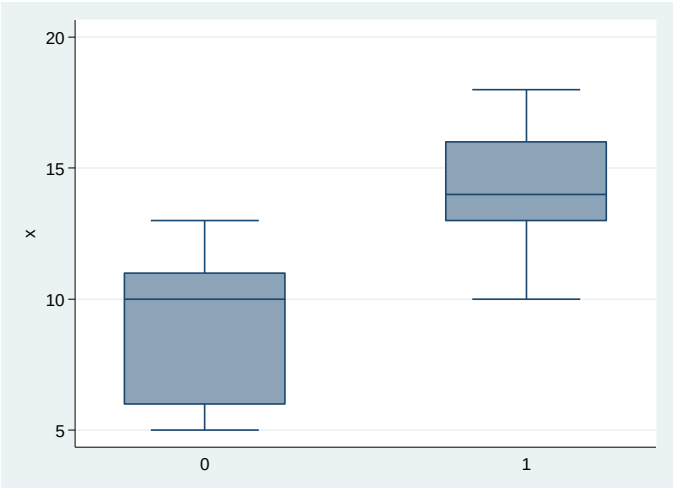
Si la discriminación fuera debida al azar se tendría S=1-E. Representando gráficamente estas relaciones, donde el eje de las abscisas es 1-E y el de las ordenadas S, se tiene:



Notar que si la discriminación fuese perfecta, es decir 100% de sensibilidad y 100% de especificidad, el punto de corte para la discriminación estaría en la intersección de las líneas azules y el área bajo la curva azul sería 1. Obviamente en cualquier aplicación real en que haya buena discriminación, esta área sería menor que 1 pero mayor a 0.5.

Observemos el siguiente ejemplo: Supongamos que la va X discrimina dos poblaciones, disponemos de 10 valores para la población “1” y 10 valores para la población “0”, los datos son los siguientes:

Id	y	x
1	0	6
2	0	11
3	0	10
4	0	10
5	0	11
6	0	6
7	0	10
8	0	5
9	0	10
10	0	13
11	1	13
12	1	11
13	1	14
14	1	16
15	1	18
16	1	14
17	1	14
18	1	18
19	1	10
20	1	16



```
. logit  y x, nolog

Logistic regression                               Number of obs   =         20
                                                  LR chi2(1)      =      14.64
                                                  Prob > chi2     =     0.0001
Log likelihood = -6.5445675                       Pseudo R2       =     0.5279

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .9371287   .4217506     2.22   0.026     .1105127    1.763745
       _cons |  -10.99978   4.905968    -2.24   0.025     -20.6153   -1.384262
------------------------------------------------------------------------------
```

$$P(y=1|X)=\frac{e^{-10.99978+0.9371287\cdot X}}{1+e^{-10.99978+0.9371287\cdot X}}$$

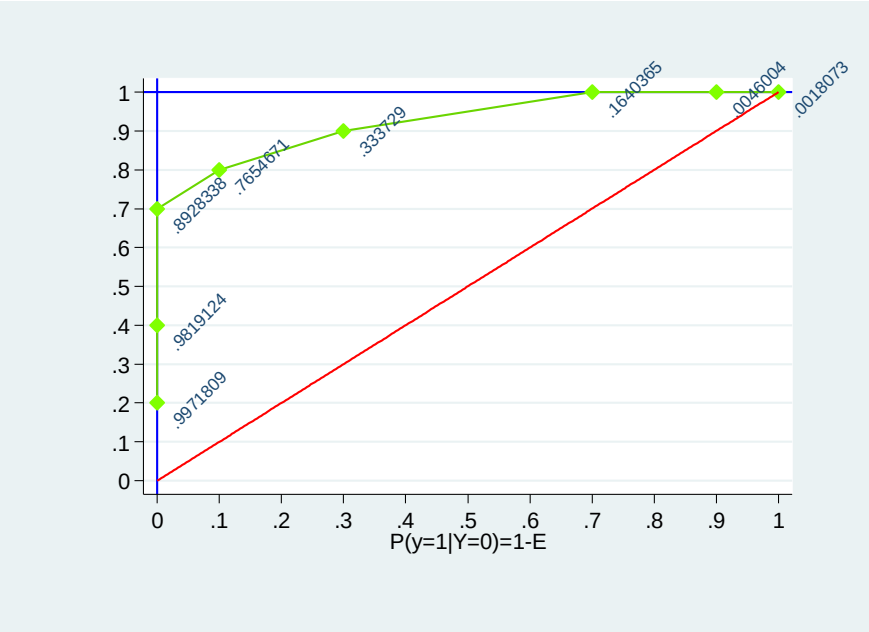
```
. predict p
(option pr assumed; Pr(y))
```

id	y	x	P(y=1|x)
1	0	6	0.004600
2	0	11	0.333729
3	0	10	0.164037
4	0	10	0.164037
5	0	11	0.333729
6	0	6	0.004600
7	0	10	0.164037
8	0	5	0.001807
9	0	10	0.164037
10	0	13	0.765467
11	1	13	0.765467
12	1	11	0.333729
13	1	14	0.892834
14	1	16	0.981912
15	1	18	0.997181
16	1	14	0.892834
17	1	14	0.892834
18	1	18	0.997181
19	1	10	0.164037
20	1	16	0.981912

La información anterior la podemos resumir en la siguiente tabla:

Pto. Corte	Número de sujetos (y=1|Y=0)	Número de sujetos(y=1|Y=1)	P(y=1|Y=0)=1-E	P(y=1|Y=1)=S
0.0018073	10	10	1.000	1.000
0.0046004	9	10	0.900	1.000
0.1640365	7	10	0.700	1.000
0.333729	3	9	0.300	0.900
0.7654671	1	8	0.100	0.800
0.8928338	0	7	0.000	0.700
0.9819124	0	4	0.000	0.400
0.9971809	0	2	0.000	0.200

Al graficar la sensibilidad versus 1-especificidad para los distintos puntos de corte se obtiene la curva ROC (*Receiver Operating Characteristic)*.

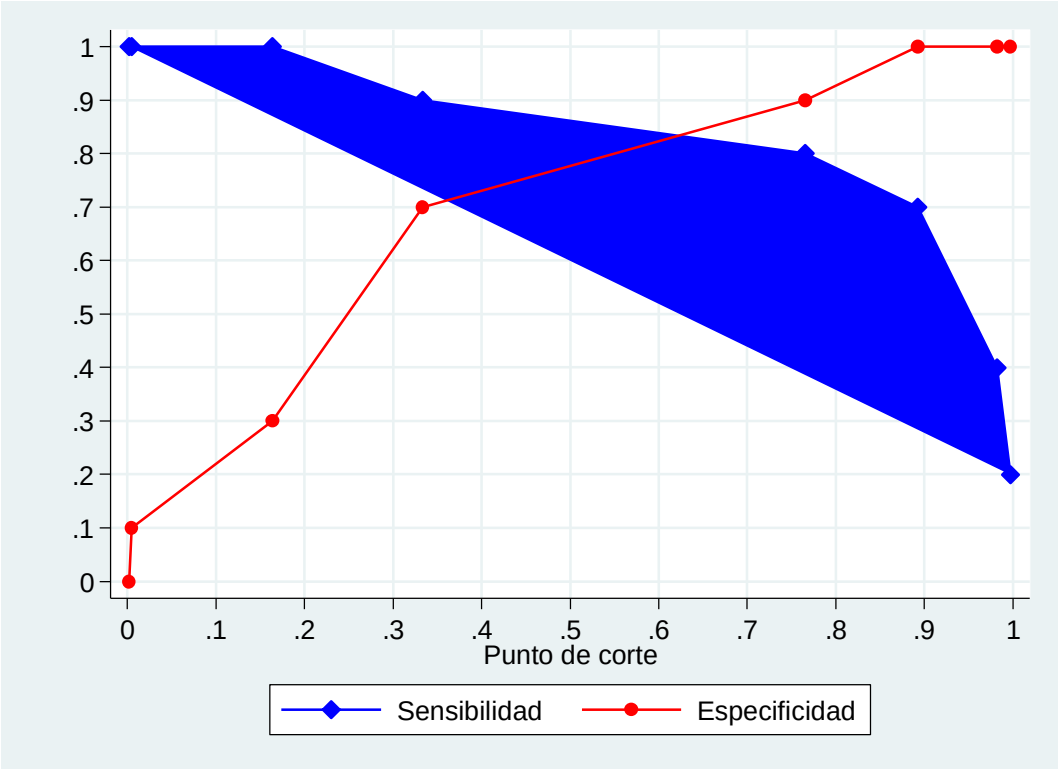


Como se sugirió, la capacidad de discriminación está dada por el área bajo la curva ROC, que en nuestro caso es 0.9250. El mejor punto de corte es aquel que está más cerca de la discriminación perfecta.

Según Hosmer y Lemeshow ("Applied Logistic Regression" Second Edition, p. 162), la capacidad de discriminación puede clasificarse según:

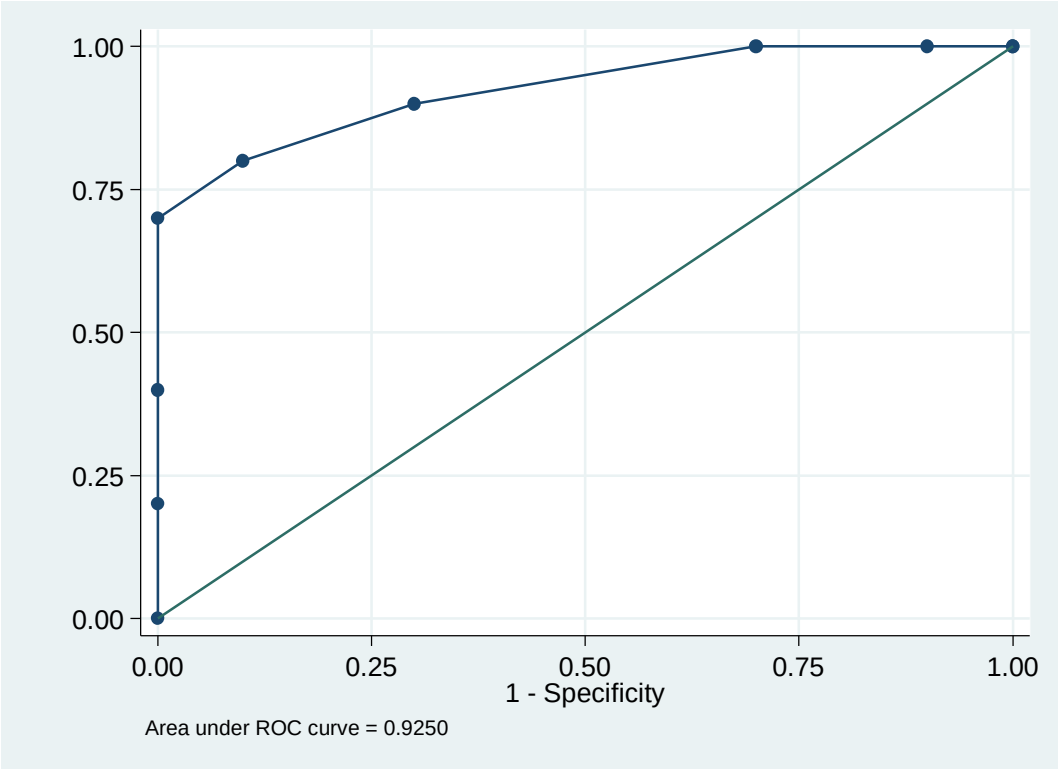
Area bajo la curva ROC	Discriminación
0.5	por azar
0.7 a 0.8	aceptable
0.8 a 0.9	muy buena
0.9 a 1	excelente

Sin embargo, para encontrar el mejor punto de corte, es mejor usar el siguiente gráfico:

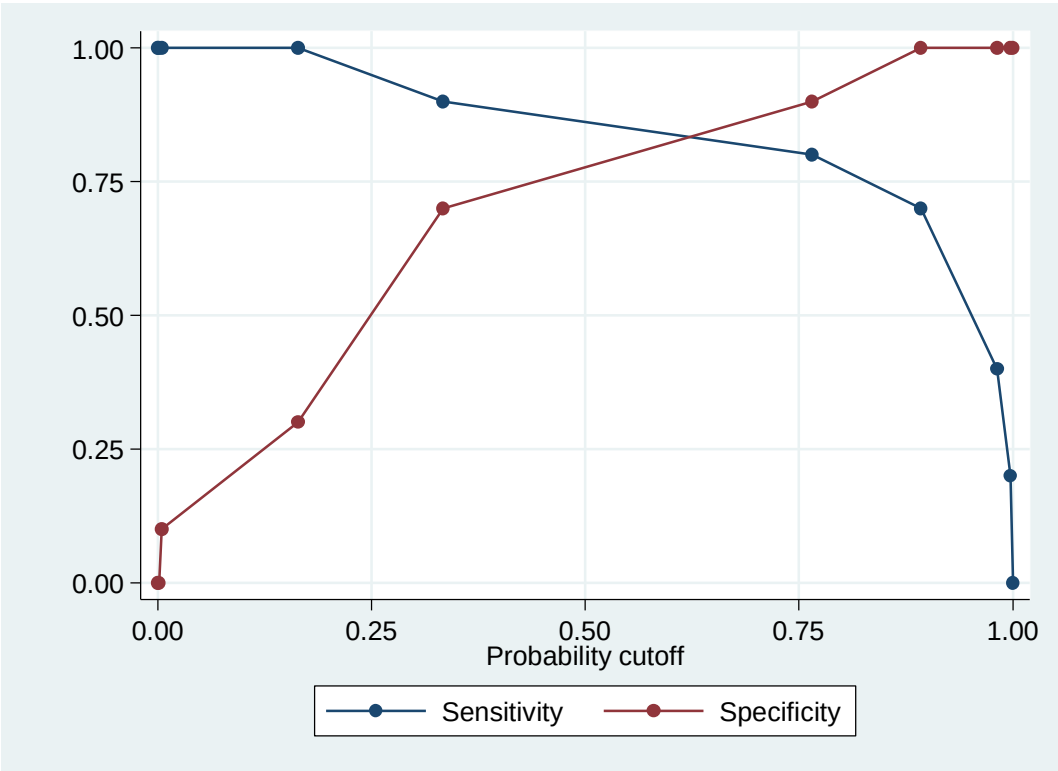


En STATA:

```
. logit  y x, nolog

Logistic regression                             Number of obs   =         20
                                                LR chi2(1)      =      14.64
                                                Prob > chi2     =     0.0001
Log likelihood = -6.5445675                     Pseudo R2       =     0.5279

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .9371287   .4217506     2.22   0.026     .1105127    1.763745
       _cons |  -10.99978   4.905968    -2.24   0.025     -20.6153   -1.384262
------------------------------------------------------------------------------

. predict p
(option pr assumed; Pr(y))

. lroc

Logistic model for y

number of observations =       20
area under ROC curve   =   0.9250
```


1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity
Area under ROC curve = 0.9250

```
. lsens
```


1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity

Recordando la definición de los Likelihood Ratios:

$$LR+= \frac{P(y=1|Y=1)}{P(y=1|Y=0)} = \frac{sensibilidad}{1-especificidad}$$

$$LR-= \frac{P(y=0|Y=1)}{P(y=0|Y=0)} = \frac{1-sensibilidad}{especificidad}$$

Podemos leer a cabalidad la siguiente salida de STATA:

```
. roctab  y p,d

Detailed report of Sensitivity and Specificity
------------------------------------------------------------------------------
                                      Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= .0018073 )   100.00%         0.00%       50.00%       1.0000
( >= .0046004 )   100.00%        10.00%       55.00%       1.1111       0.0000
( >= .1640365 )   100.00%        30.00%       65.00%       1.4286       0.0000
( >= .333729 )     90.00%        70.00%       80.00%       3.0000       0.1429
( >= .7654671 )    80.00%        90.00%       85.00%       8.0000       0.2222
( >= .8928338 )    70.00%       100.00%       85.00%                    0.3000
( >= .9819124 )    40.00%       100.00%       70.00%                    0.6000
( >= .9971809 )    20.00%       100.00%       60.00%                    0.8000
( >  .9971809 )     0.00%       100.00%       50.00%                    1.0000
------------------------------------------------------------------------------


                    ROC                    -Asymptotic Normal--
          Obs       Area     Std. Err.      [95% Conf. Interval]
        -------------------------------------------------------
          20     0.9250       0.0575        0.81231     1.00000
```

Si se ha escogido como punto de corte 0.7654671:

```
. lstat, cutoff(.7654671)

Logistic model for y

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         8             1  |          9
     -     |         2             9  |         11
-----------+--------------------------+-----------
   Total   |        10            10  |         20

Classified + if predicted Pr(D) >= .7654671
True D defined as y != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   80.00%
Specificity                     Pr( -|~D)   90.00%
Positive predictive value       Pr( D| +)   88.89%
Negative predictive value       Pr(~D| -)   81.82%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   10.00%
False - rate for true D         Pr( -| D)   20.00%
False + rate for classified +   Pr(~D| +)   11.11%
False - rate for classified -   Pr( D| -)   18.18%
--------------------------------------------------
Correctly classified                        85.00%
--------------------------------------------------
```

Para encontrar el punto de corte de la variable original X, usamos:

$$\ln\left(\frac{p}{1-p}\right) = a + b \cdot X$$

De donde:

$$X = \frac{\ln\left(\frac{p}{1-p}\right) - a}{b}$$

En nuestro caso:

$$X = \frac{\ln\left(\frac{0.7654671}{1-0.7654671}\right) - (-10.99978)}{0.9371287} = 12.999997 = 13$$

```
. dis (ln(.7654671/(1-.7654671))-( -10.99978))/ .9371287
12.999997
```