Index Class 25 - 11/02/2020

1. Classification problems
2. Types of Errors
3. Metrics of classification problems
4. Odds
5. Logistic regression
6. Significance of variables

Classification problems

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs.

The prediction outcome is normally in the form of class probability.

- 1 the event is predicted to happen
- 0 the event is not predicted to happen

Metrics - Types of errors

- True positives: predicted = 1, real = 1
- True negative: predicted = 0, real = 0
- False positive: predicted = 1, real = 0
- False negative: predicted = 0, real = 1

Metrics - Types of errors

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Metrics - Confusion Matrix

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Metrics - Types of errors

$$\text{Accuracy (ACC)} = \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$$

True positive rate (TPR), Recall, Sensitivity, probability of detection

$$= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$$

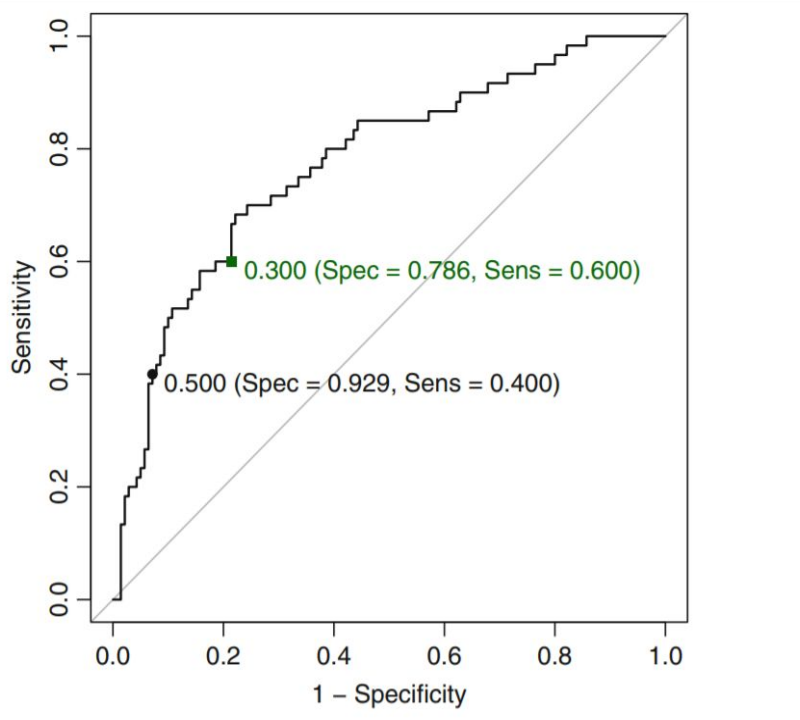
Specificity (SPC), Selectivity, True negative rate (TNR)

$$= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$$

Metrics - ROC Curve

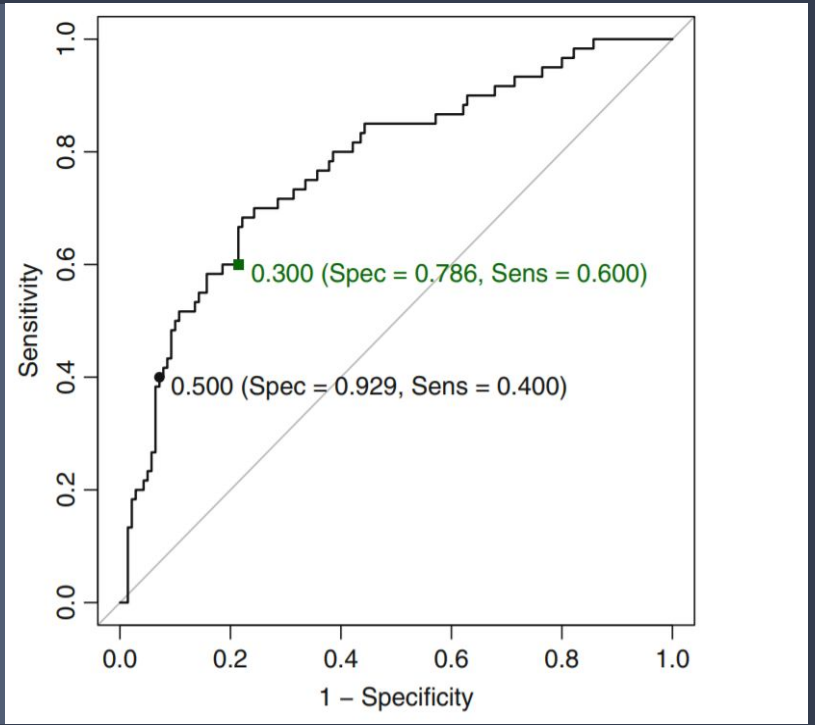
A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination **threshold** is varied.

1-Specificity = False Positive Rate
Sensitivity = True Positive Rate



Metrics - AUC

The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.



Ratio or Odds

Tabla 1. Participación en una huelga por sexo. Valores absolutos

	Hombre	Mujer	Total
No han participado	960	1057	2017
Han participado	261	206	468
Total	1221	1263	2484

468/2484=0.1884->18.84%
468/2017=0.2320 (Ratio)
Ratio : How many times A is greater than B

Odds

This is called odds:

$$Odd = \frac{p}{q} = \frac{p}{(1-p)}$$

	Hombre	Mujer
$Odd_{Participa/No\ participa}$	0,272	0,195

Odds

	Hombre	Mujer
$Odd_{Participa/No\ participa}$	0,272	0,195

$$Mujeres/Hombres = 0{,}195/0{,}272 = 0{,}717$$

$$Odd\ Ratio = OR = \frac{Odd_A}{Odd_B} = \frac{\frac{p_A}{(1-p_A)}}{\frac{p_B}{(1-p_B)}}$$

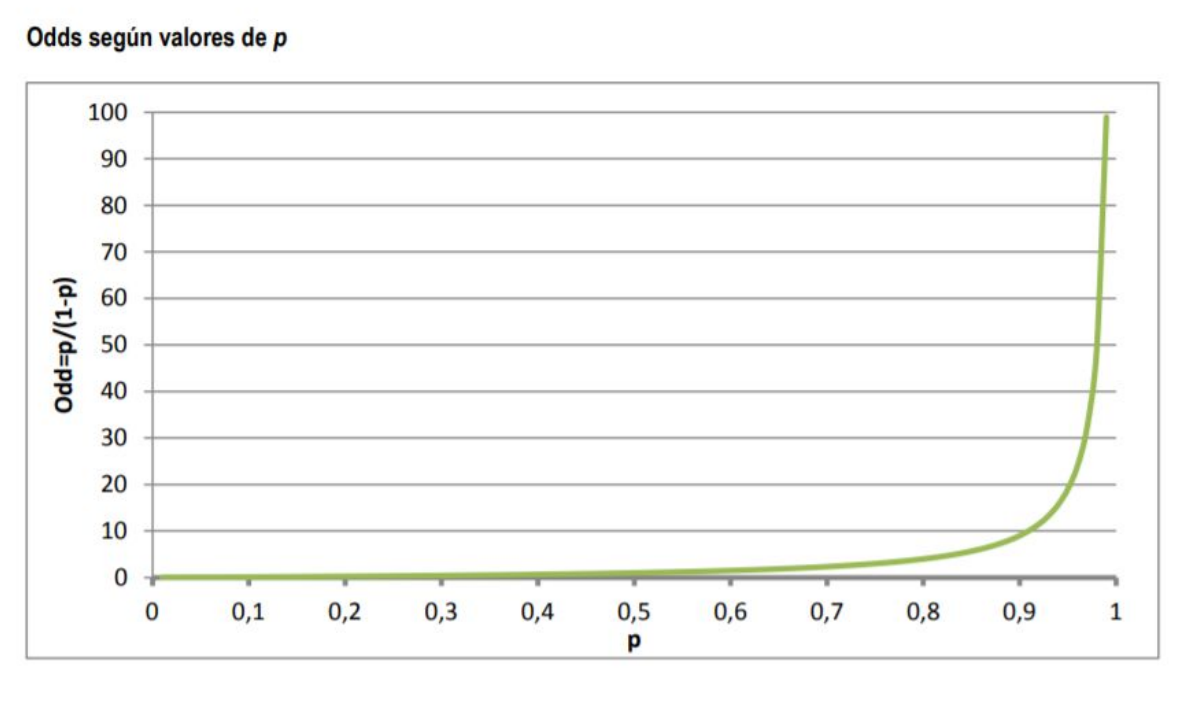
Odds/Proportion

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

$$p = \frac{Odd}{(1+Odd)}$$

$$\text{Logit}=\text{Ln}(\text{Odd})=\text{Ln}\left(\frac{\text{p}}{1-\text{p}}\right)$$

Odds/Proportion

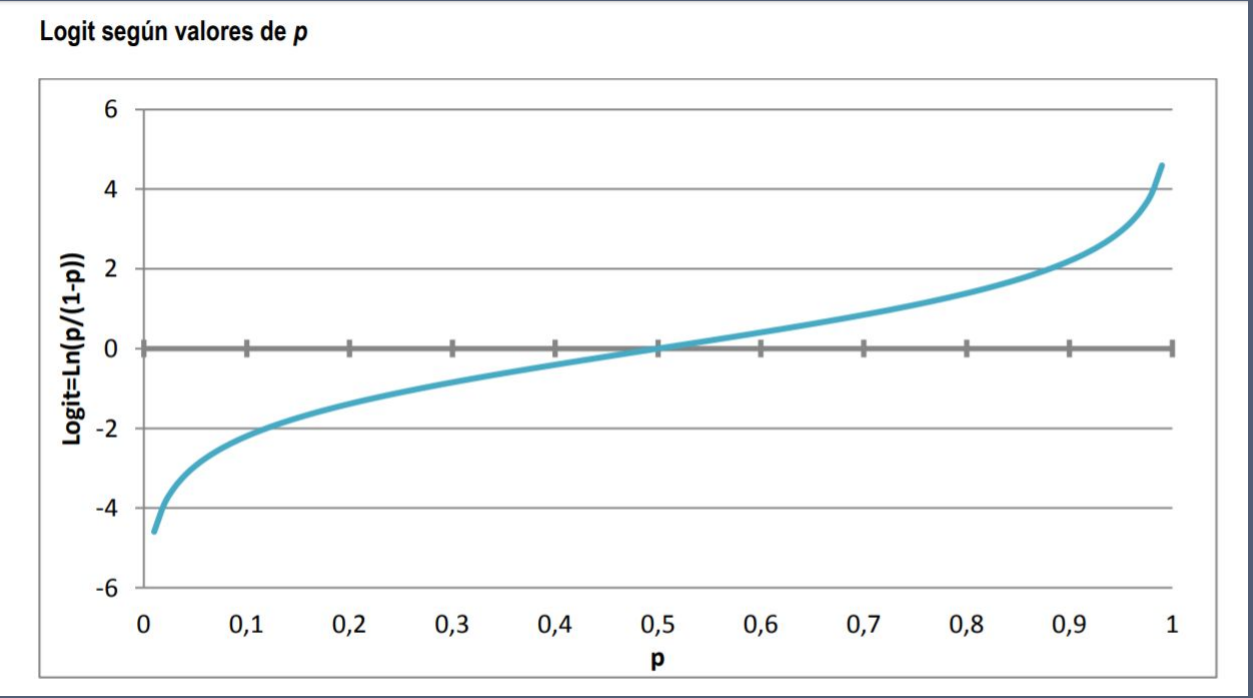


Odds/Proportion

-Logit is symmetric

Odd=0,3/0,7=0,429	Logit=Ln(0,3/0,7)=-0,847
Odd=0,7/0,3=2,333	Logit=Ln(0,7/0,3)=+0,847

Odds/Proportion



Logistic Regression

$$z=\alpha+\beta x$$

$$\mathrm{Ln}\left(\frac{p}{1-p}\right)=\alpha+\beta x$$

$$\left(\frac{p}{1-p}\right)=e^{(\alpha+\beta x)}$$

$$p=\frac{1}{1+e^{-(\alpha+\beta x)}}$$

Logistic Regression

Tabla 3. Participación en una huelga por sexo. Datos expresados en proporciones

	Hombre	Mujer	Total
No Huelguista	0,78624079	0,83689628	0,81199678
Huelguista	0,21375921	0,16310372	0,1884058
	1	1	1

$$\text{Logit}=\text{Ln}\left(\frac{0{,}21375921}{0{,}78624079}\right)=\text{Ln}\,(0{,}271875)=-1{,}30241288$$

$$\text{Logit}=\text{Ln}\left(\frac{0{,}16310372}{0{,}83689628}\right)=\text{Ln}\,(0{,}1948912)=-1{,}63531382$$

$$z=\alpha+\beta x$$

$$\text{Ln}\left(\frac{p}{1-p}\right)=\alpha+\beta x \qquad p=\frac{1}{1+e^{-(\alpha+\beta x)}}$$

X=0 -> Logit = alpha = -1.30
X=1 -> Logit = alpha+beta= -1.63
=> beta = -0.33

$$\alpha+\beta x=-1{,}302-0{,}333x$$

Logistic Regression

$$p = \frac{1}{1 + e^{-(\alpha+\beta x)}} = \frac{1}{1 + e^{-(-1,302)}} = \frac{1}{1 + 3,678} = 0,214$$

$$p = \frac{1}{1 + e^{-(\alpha+\beta x)}} = \frac{1}{1 + e^{-(-1,302-0,333)}} = \frac{1}{1 + 5,129} = 0,163$$

La constante α=-1,302 $\quad e^{\alpha} = e^{-1,302} = 0,272 = \text{Odd}_{\text{hombres}}$

La pendiente β=0,642 $\quad e^{\beta} = e^{-0,333} = 0,717 = \text{OR}$

Another Example

Tabla 4. Participación en una huelga por nivel de estudios. Datos absolutos

	Nivel de Estudios		
	Hasta Básicos	Medios	Universitarios
No hacen huelga	1116	554	344
Sí hacen huelga	138	182	145
Total	1254	736	489

Tabla 5. Participación en una huelga por nivel de estudios. Porcentajes verticales

	Nivel de Estudios		
	Hasta Básicos	Medios	Universitarios
No hacen huelga	89,0%	75,3%	70,3%
Sí hacen huelga	11,0%	24,7%	29,7%
Total	100%	100%	100%

$$OR_{Medios/Básicos} = \frac{\frac{182}{554}}{\frac{138}{1116}} = 2{,}657$$

$$OR_{Universitarios/Básicos} = \frac{\frac{145}{344}}{\frac{138}{1116}} = 3{,}409$$

	X_1	X_2
Hasta Básicos	0	0
Medios	1	0
Universitarios	0	1

$a=\text{Ln}(\text{Odd}_{básicos})=\text{Ln}(0{,}11/0{,}89)=-2{,}09$

b_1 será el $\text{Ln}(OR_{medios/básicos})=\text{Ln}(2{,}657)=0{,}977$

b_2 será el $\text{Ln}(OR_{universitarios/básicos})=\text{Ln}(3{,}409)=1{,}226$

Logistic Regression

	$Z=a+b_1x_1+b_2x_2$	$p=\frac{1}{1+e^{-z}}$
Hasta Básicos	z=-2,09+0,977(0)+1,226(0)=-2,09	$p=\frac{1}{1+e^{2,09}}=0,11$
Medios	z=-2,09+0,977(1)+1,226(0)=-1,113	$p=\frac{1}{1+e^{1,113}}=0,2473$
Universitarios	z=-2,09+0,977(0)+1,226(1)=-0,864	$p=\frac{1}{1+e^{0,864}}=0,2965$

$$p=\frac{1}{1+e^{-(a+b_1x_1+b_2x_2)}}$$

One more example

Tabla 7. Participación en huelgas según sexo y nivel de estudios. Absolutos y proporción.

Estudios	Sexo	No han realizado huelga	Sí han realizado huelga	Total	P (Han realizado huelga)
Hasta Básicos	Hombre	519	79	598	0,13210702
	Mujer	597	59	656	0,08993902
Medios	Hombre	282	103	385	0,26753247
	Mujer	272	79	351	0,22507123
Universitarios	Hombre	158	78	236	0,33050847
	Mujer	186	67	253	0,26482213

One more example

	B	Exp(B)
Estudios (1)	0,966	2,627
Estudios (2)	1,229	3,419
Sexo (1)	-0,322	0,724
Constante	-1,932	0,145

$$Y=a_o+b_1(Sexo1)+b_2(Estudios1)+b_3(Estudios2)$$

Estudios	Sexo		$p = \frac{1}{1+e^{-z}}$	P (Han realizado huelga)
Hasta Básicos	Hombre	Z=-1,932	0,12652938	0,13210702
	Mujer	Z=-1,932-0,322	0,09500499	0,08993902
Medios	Hombre	Z=-1,932+0,966	0,2756785	0,26753247
	Mujer	Z=-1,932-0,322+0,966	0,21619152	0,22507123
Universitarios	Hombre	Z=-1,932+1,229	0,33114743	0,33050847
	Mujer	Z=-1,932-0,322+1,229	0,26405461	0,26482213

Logistic Regression- Cost function

$$h(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$h(x) = \begin{cases} > 0.5, & \text{if } \theta^T x > 0 \\ < 0.5, & \text{if } \theta^T x < 0 \end{cases}$$

$$cost = \begin{cases} -log(h(x), & \text{if } y = 1 \\ -log(1 - h(x)), & \text{if } y = 0 \end{cases}$$

$$cost(h(x), y) = -ylog(h(x)) - (1 - y)log(1 - h(x))$$

Cost Function

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^i log(h(x^i)) + (1 - y^i) log(1 - h(x^i))]$$

Gradient

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h(x^i) - y^i)\, x_j^i$$

Multi classification problem

- First each class is separated in columns as a binary columns
- 1 if the row belongs to this class
- 0 otherwise

$$\hat{p}_\ell^* = \frac{e^{\hat{y}_\ell}}{\sum_{l=1}^{C} e^{\hat{y}_l}}$$

Statistical hypothesis testing

A statistical hypothesis is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables.

A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets.

P-value

The p-value or probability value or significance is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary would be greater than or equal to the actual observed results.

Normally we use a p-value of 0.05

P-value

	B	E.T.	Wald	gl	Sig.	Exp(B)
SEXO(1)	-,322	,106	9,279	1	,002	,724
ESTUDIOS			98,031	2	,000	
ESTUDIOS(1)	,966	,125	60,156	1	,000	2,627
ESTUDIOS(2)	1,229	,134	83,836	1	,000	3,419
Constante	-1,932	,103	354,159	1	,000	,145