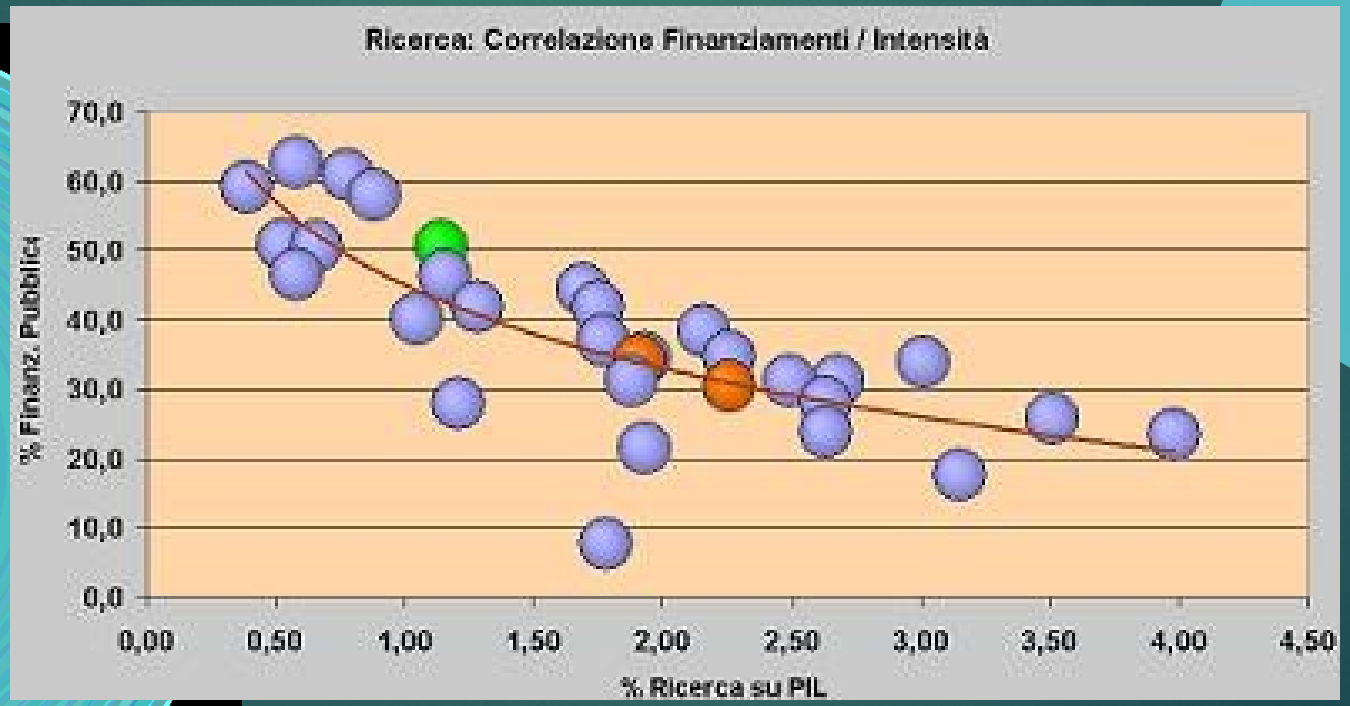
Day 4

correlation - regression

CORRELATION

IN STATISTICS, **DEPENDENCE** OR **ASSOCIATION** IS ANY STATISTICAL RELATIONSHIP, WHETHER CAUSAL OR NOT, BETWEEN TWO RANDOM VARIABLES OR BIVARIATE DATA.



Correlation coeffcient = r

$$r = \frac{\Sigma (\bar{x}-x)(\bar{y}-y)}{\sqrt{\Sigma (\bar{x}-x)^2 \Sigma (\bar{y}-y)^2}}$$

Example 1 …

weight	height
56	156
53	159
61	
67	16[illegible]
65	175
71	173
73	178
78	180
78	187
83	190
86	192

Correlation Coefficient
0,960777

Example 2 …

price	age
18.000	0,5
17.000	1
16.500	2
12.000	2,5
11.000	3
9.000	5
6.000	6
6.000	7
3.500	8
1.000	12

Correlation Coefficient

-0,95892

Example 3 …

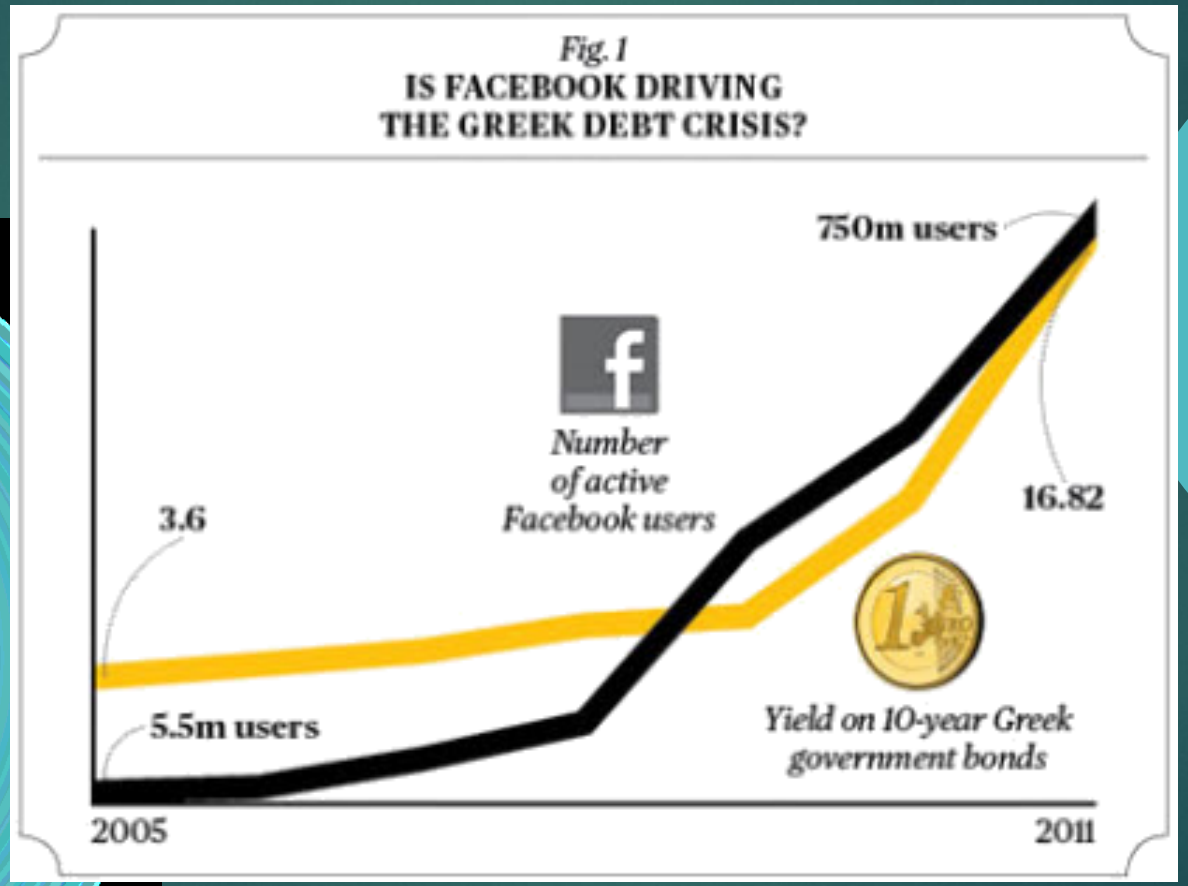
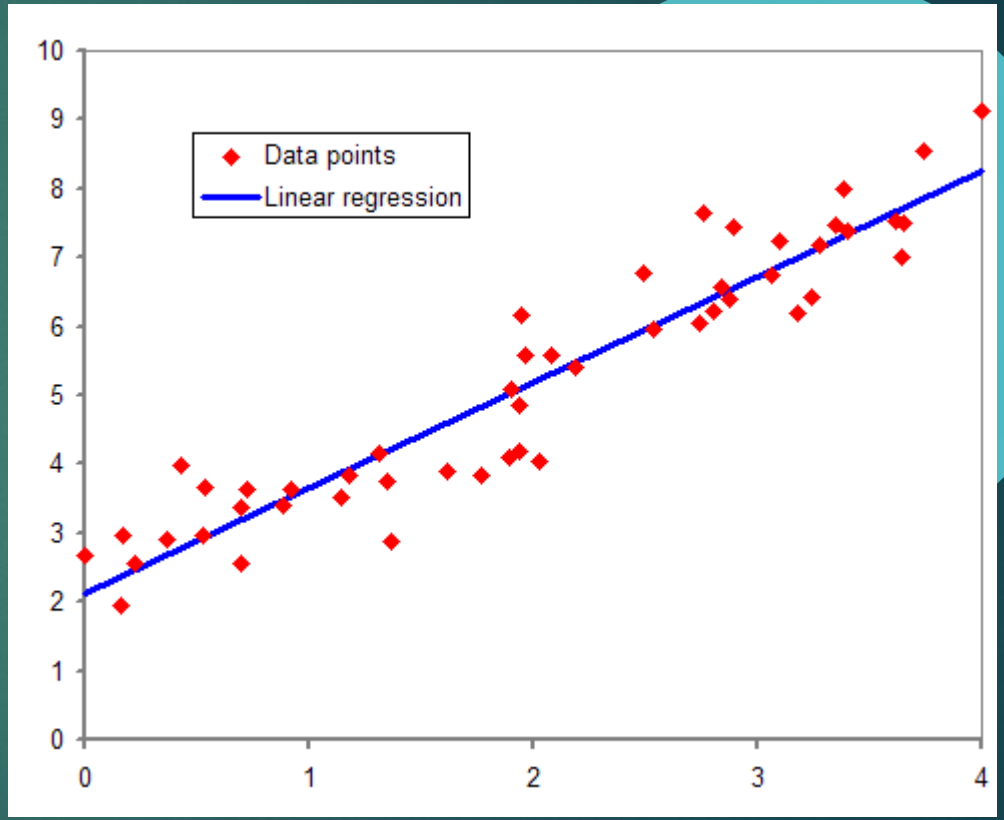


Fig. 1
IS FACEBOOK DRIVING
THE GREEK DEBT CRISIS?
750m users
Number
of active
Facebook users
3.6
16.82
5.5m users
Yield on 10-year Greek
government bonds
2005
2011

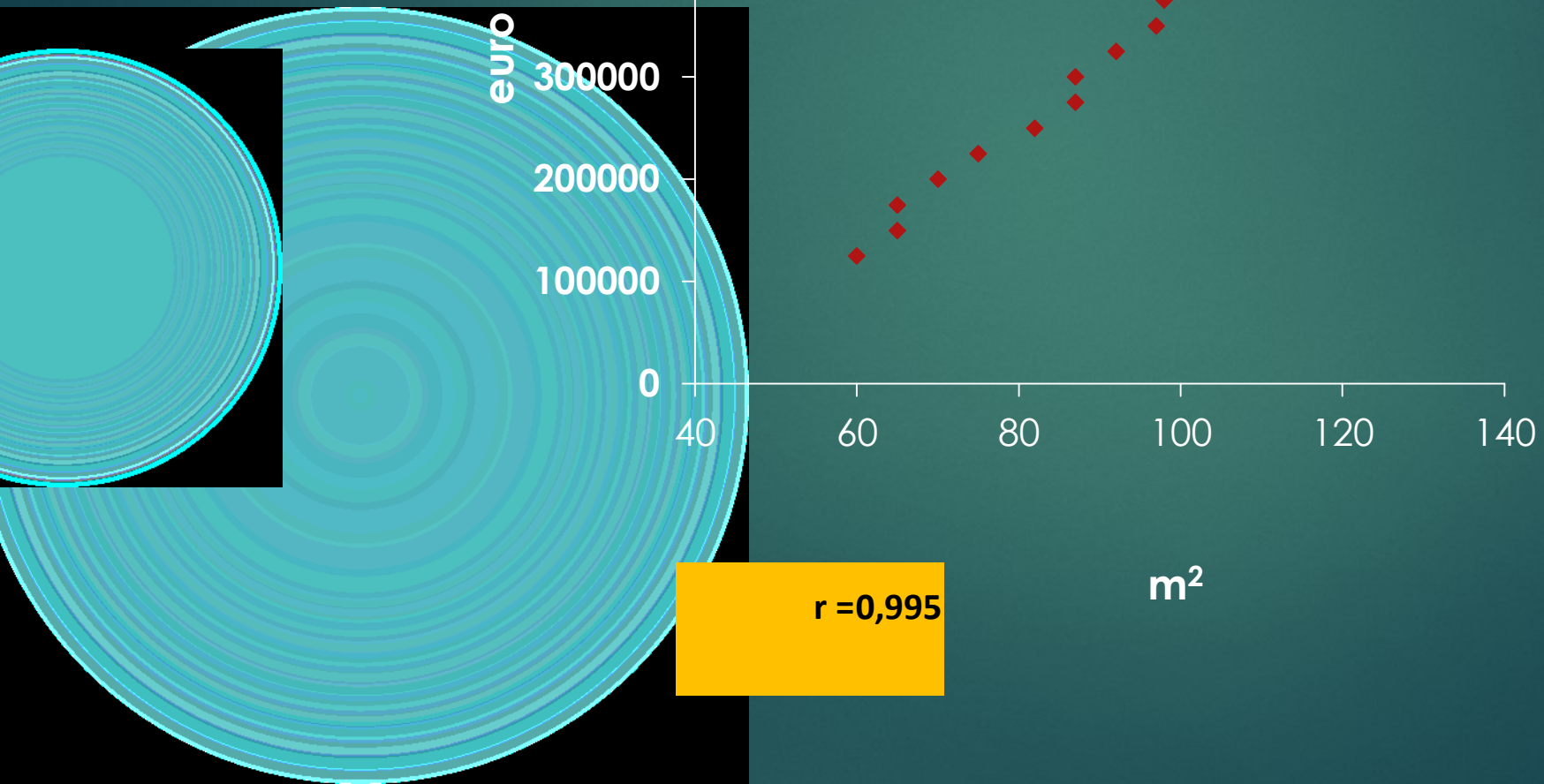
Linear regression

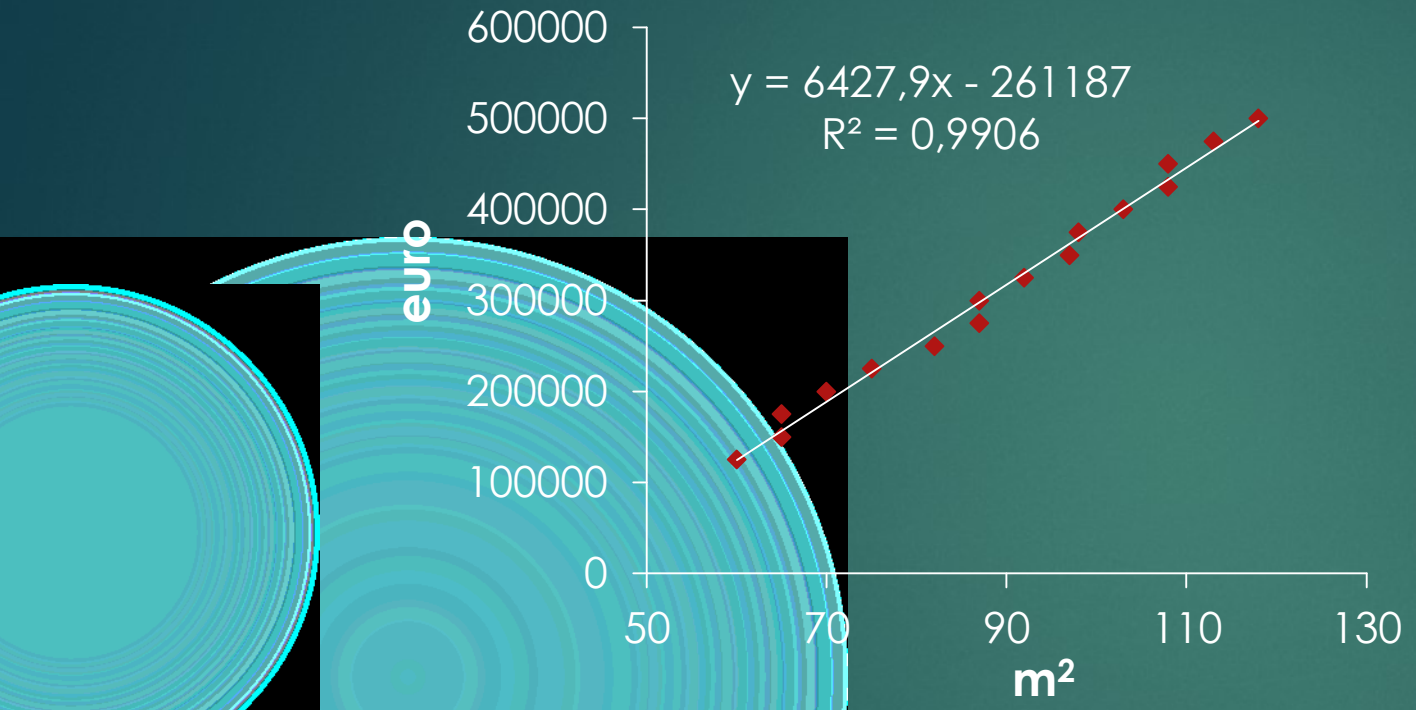
In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').



area	price
60	125000
65	150000
65	175000
70	200000
75	225000
82	250000
87	275000
	00000
92	25000
97	50000
98	75000
103	00000
108	25000
108	50000
113	475000
118	500000

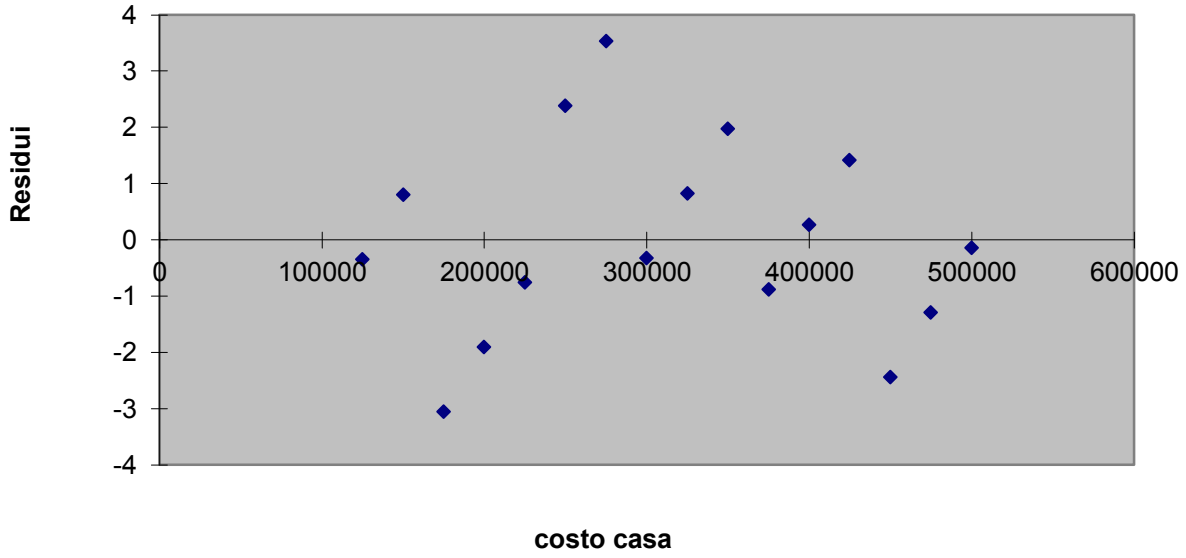
r =0,995


600000
500000
400000
300000
200000
100000
0
euro
40
60
80
100
120
140
m²
r =0,995



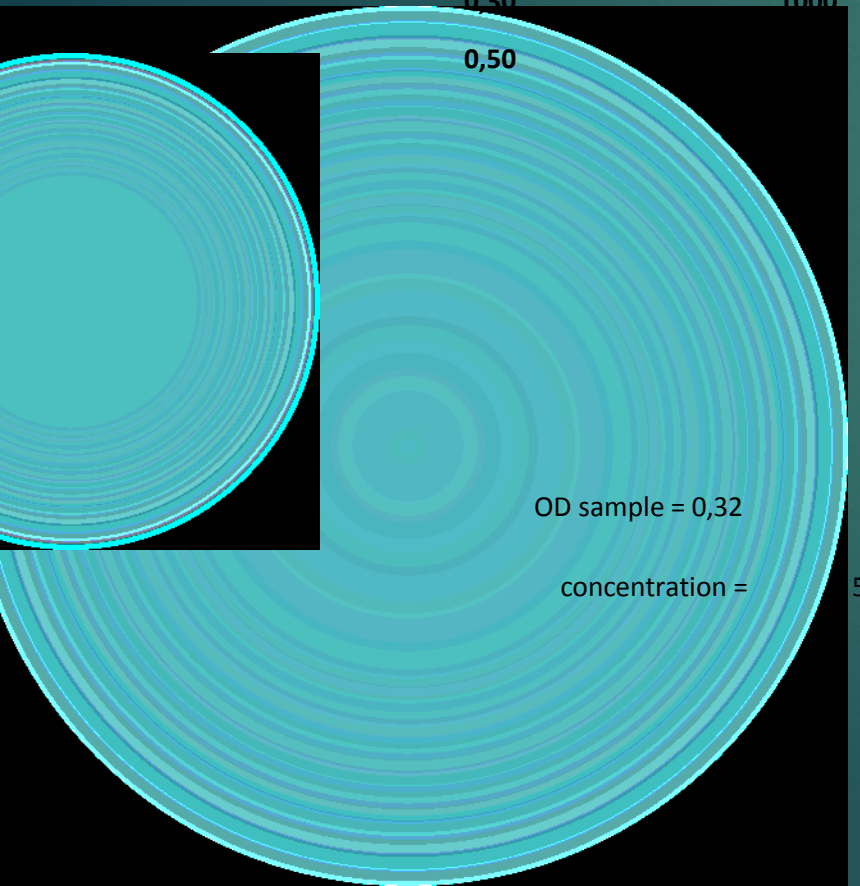
To evaluate the statistical significance of the model as a whole it comes used the test F based on the relationship between variance explained by the model and residual variance.
If the observed p-value is less than the theoretical p-value (usually 0.05) the The model used explains a significant proportion of variance in the phenomenon.

r = 0,995313
p<0.0001


Residui
4
3
2
1
0
-1
-2
-3
-4
0
100000
200000
300000
400000
500000
600000
costo casa

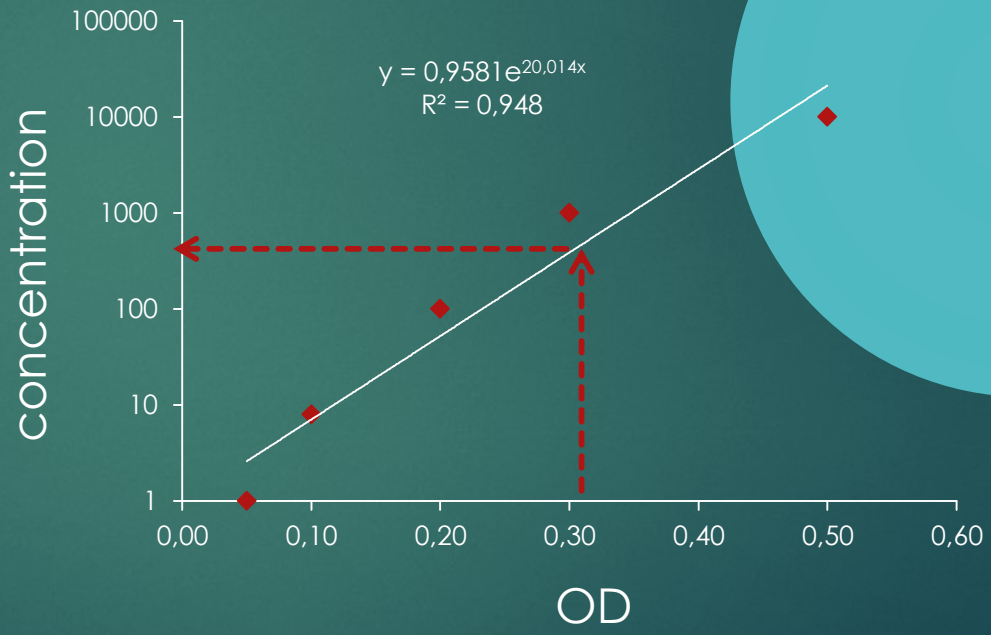
Instrument calibration

OD	concentration
0,05	1
0,10	8
0,20	100
0,30	1000
0,50	



OD sample = 0,32

concentration = 579,2416



Multivariate linear regession

price	nearby pharmacies	Distance to the City centre	floor	year	Distance to the highway	are
125000	6	250	6	1954	200	60
[illegible]	3	300	4	1936	250	60
[illegible]	4	250	4	1965	180	65
200[illegible]	3	220	3	1968	150	70
225000	1	180	1	1973	190	55
250000	4	169	4	2004	80	82
275000	3	300	5	1998	110	87
300000	1	120	3	2000	120	78
325000	5	89	2	1994	90	83
350000	5	24	1	1993	90	67
375000	3	80	1	2001	100	98
400000	2	100	5	2005	200	90
425000	1	60	3	1958	78	95
450000	5	120	2	1935	80	108
475000	6	25	1	2000	34	89
500000	3	10	1	2010	5	94
	[illegible],004187254	**-0,855097664**	**-0,57074**	**0,39940657**	**-0,775[illegible]07084**	**0,8[illegible]7505543**
	NS	**p<0.0001**	**0,0209**	**NS**	**0,0[illegible]04**	**p<0.0001**

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0,961953367
R al quadrato	0,925354281
R al quadrato corretto	0,875590468
Errore standard	41981,76232
Osservazioni	16

ANALISI
NZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
ssione	6	1,96638E+11	32772964116	18,5949233	0,000133688
o	9	15862215305	1762468367		
	15	2,125E+11			

	Coefficienti
Intercetta	285306,71
n° farmacie vicine	-2867,821061
distanza centro	-632,1291229
piano	-5082,126583
anno costruzione	-84,46606416
distanza viabilità	-44,39216187
superficie	3936,412268

Example…

age	cholesterol	Glycemia	Pressure (min)	Card freq	Life style (1 - 5)
56	321	188	188	186	1
67	300	200	200	169	2
68	342	156	190	184	1
71	287	189	156	160	1
73	220	177	180	120	2
		145	138	134	3
77		166	100	80	4
81		123	77	110	5
84		110	80	90	2
87	17	86	92	123	5
96	160	80	60	53	5
	-0,8878	-0,8862	-0,8845	-0,8647	0,7872

r vs. anni di vita

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	**0,95892**
R al quadrato	**0,91952**

	Coefficienti
Intercetta	111,6407
colesterolo	-0,01621
glicemia	-0,14676
pressione min	0,027452
freq card	-0,11672
stile di vita (1 - 5)	0,378917

age	Calculated age
56	
67	
68	6,
71	65,
73	73,7
75	75,5
77	78,8
81	81,8
84	85,0
87	86,2
96	94,6

r =0,95892

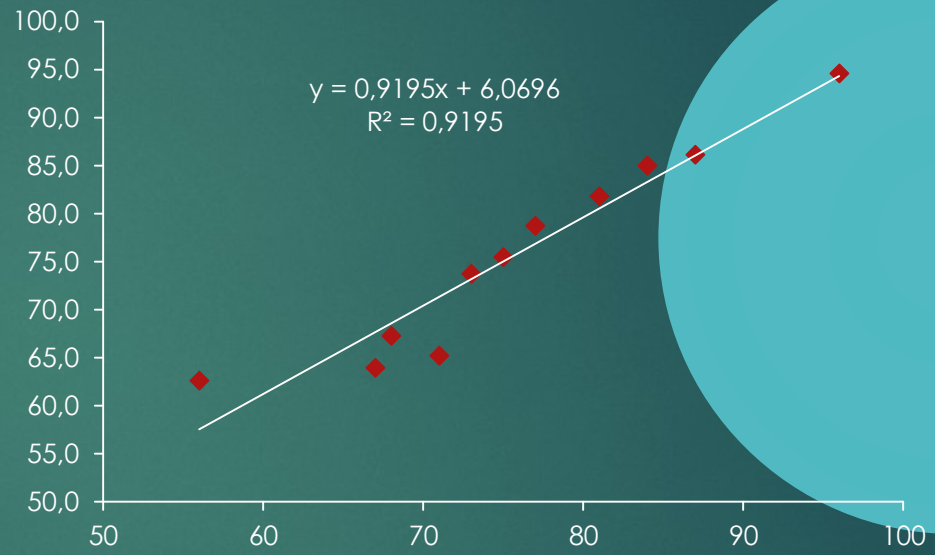


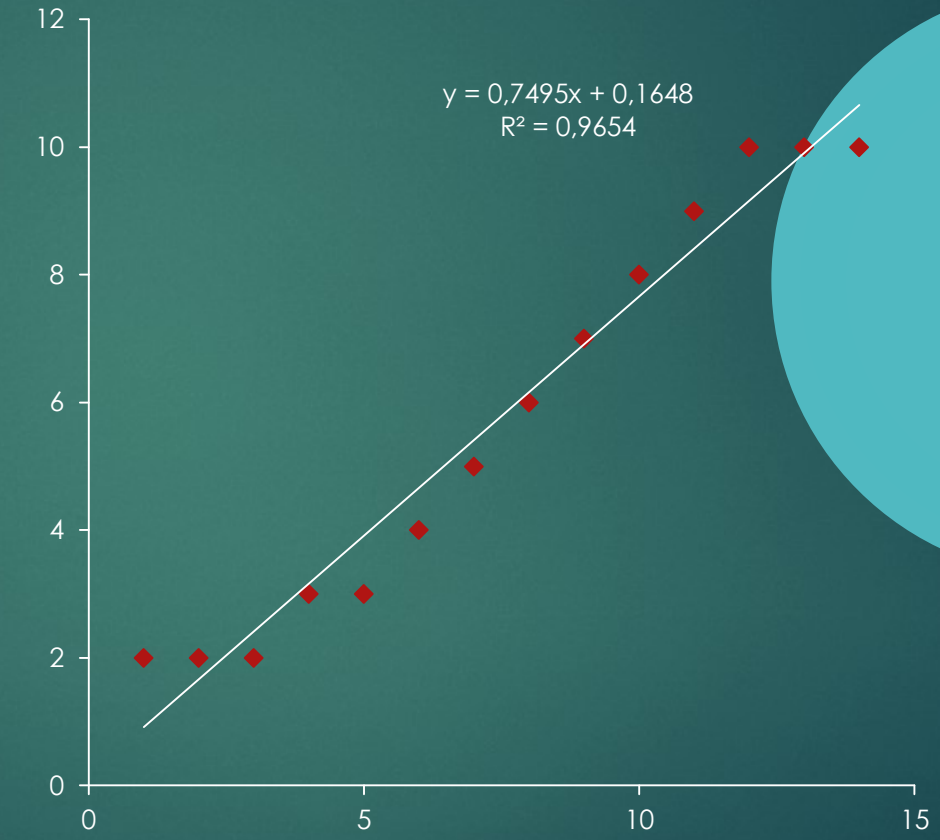
Table 2 *p* value of correlation of MCV with other hematological parameters in CTR and MICRO group of dogs assessed by a linear multivariate regression model

	CTR	MICRO
WBC	NS	$p<0.05$
RBC	$p<0.001$	$p<0.001$
Hgb	$p<0.05$	NS
Hct	NS	$p<0.001$
MCH	$p<0.001$	NS
MCHC	$p<0.001$	NS
RDW	NS	$p<0.001$
PLT	NS	NS
MPV	NS	$p<0.05$

NS, non significant=$p>0.05$

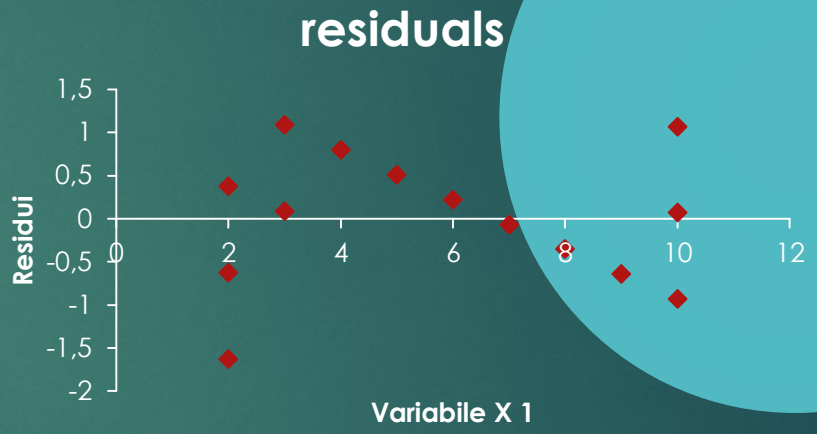
ANALYTICAL METHODS

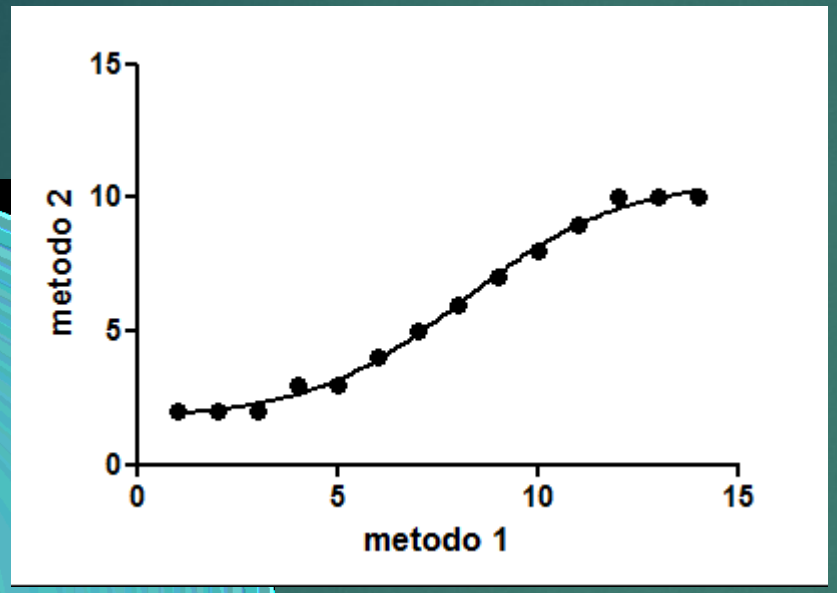
method 1	method 2
1	2
2	2
3	2
6	
7	
8	
9	7
10	8
11	9
12	10
13	10
14	10



OUTPUT RIEPILOGO

	egressione
R multip	0,982562
R al quadrat	0,965428
R al quadrato c	0,962547
Errore standard	0,809582
Osservazioni	14

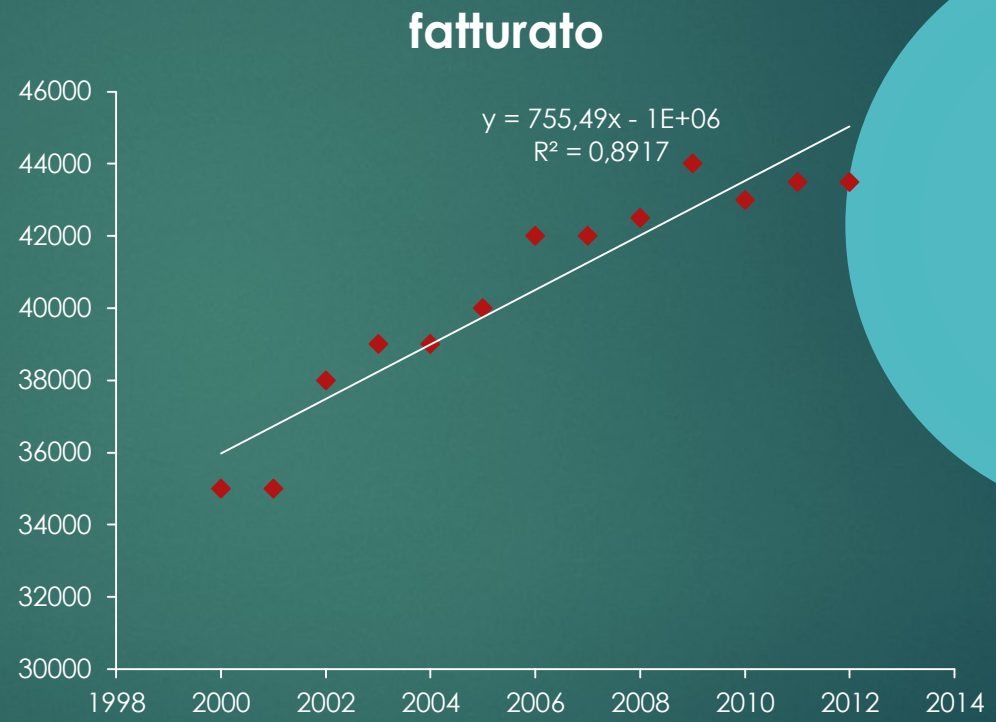




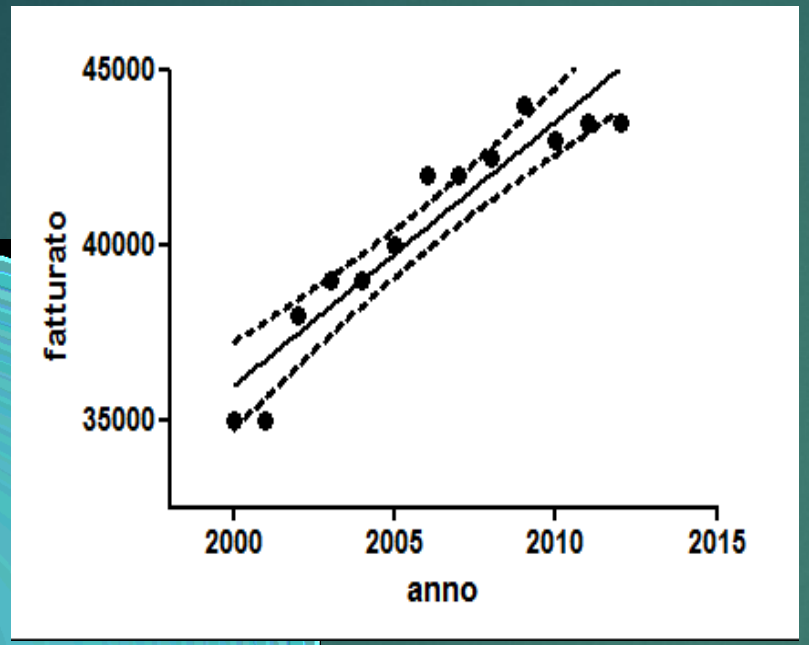
Boltzmann sigmoidal	
Best-fit values	
BOTTOM	1,688
TOP	10,68
V50	8,155
SLOPE	1,952
Std. Error	
BOTTOM	0,2276
TOP	0,3054
V50	0,1873
SLOPE	0,2040
95% Confidence Intervals	
BOTTOM	1.181 to 2.195
TOP	9.999 to 11.36
V50	7.738 to 8.572
SLOPE	1.497 to 2.406
Goodness of Fit	
Degrees of Freedom	10
R^2	0,9958
Absolute Sum of Squares	0,5560
Sy.x	0,2358
Number of points Analyzed	14

REGRESSION AND PREDICTION

anno	fatturato
2000	35000
2001	
2002	
2003	
2004	390
2005	40000
2006	42000
2007	42000
2008	42500
2009	44000
2010	43000
2011	43500
2012	43500
2012	?
2013	?



REGRESSION AND PREDICTION



Goodness of Fit	
r^2	0,89
Sy.x	1100
Is slope significantly non-zero?	
F	91
DFn, DFd	1.0, 11
P value	< 0.0001
Deviation from zero?	Significant
Data	
Number of X values	13
Maximum number of Y replicates	1
Total number of values	13
Number of missing values	0

2010,08	43582,110	967,3312	967,3312
2010,20	43672,760	982,8868	982,8868
2010,32	43763,420	998,6403	998,6403
2010,44	43854,070	1014,583	1014,583
2010,56	43944,730	1030,705	1030,705
2010,68	44035,380	1046,999	1046,999
2010,80	44126,040	1063,457	1063,457
2010,92	44216,700	1080,071	1080,071
2011,04	44307,350	1096,835	1096,835
2011,16	44398,010	1113,740	1113,740
2011,28	44488,660	1130,782	1130,782
2011,40	44579,320	1147,955	1147,955
2011,52	44669,970	1165,251	1165,251
2011,64	44760,630	1182,666	1182,666
2011,76	44851,290	1200,195	1200,195
2011,88	44941,940	1217,832	1217,832
2012,00	45032,600	1235,574	1235,574

Non linear models: sigmoid

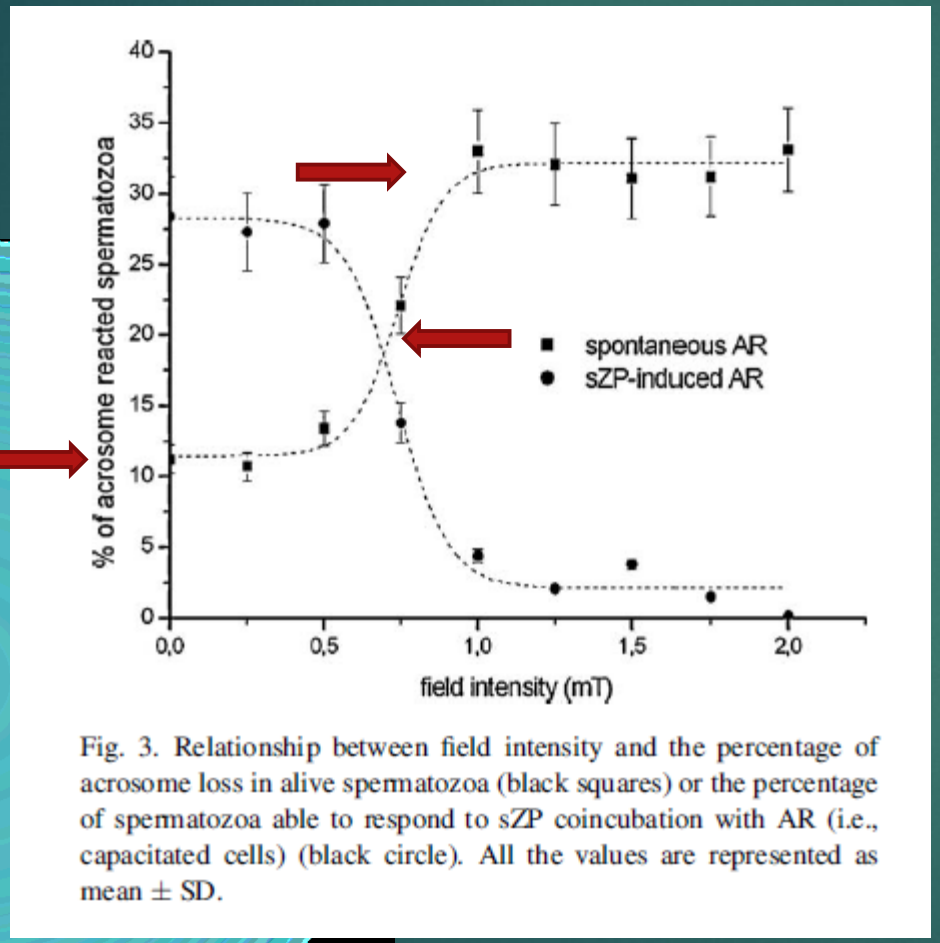
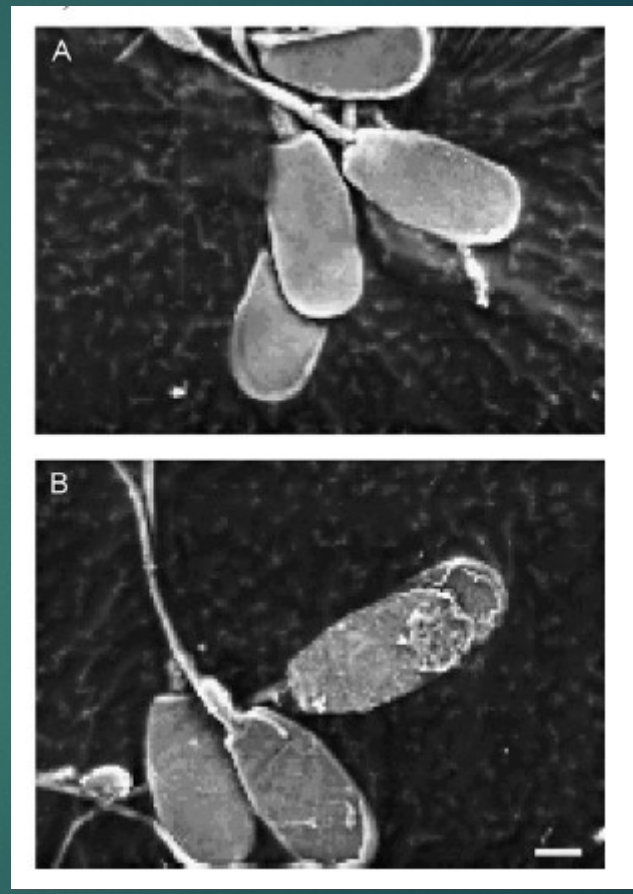
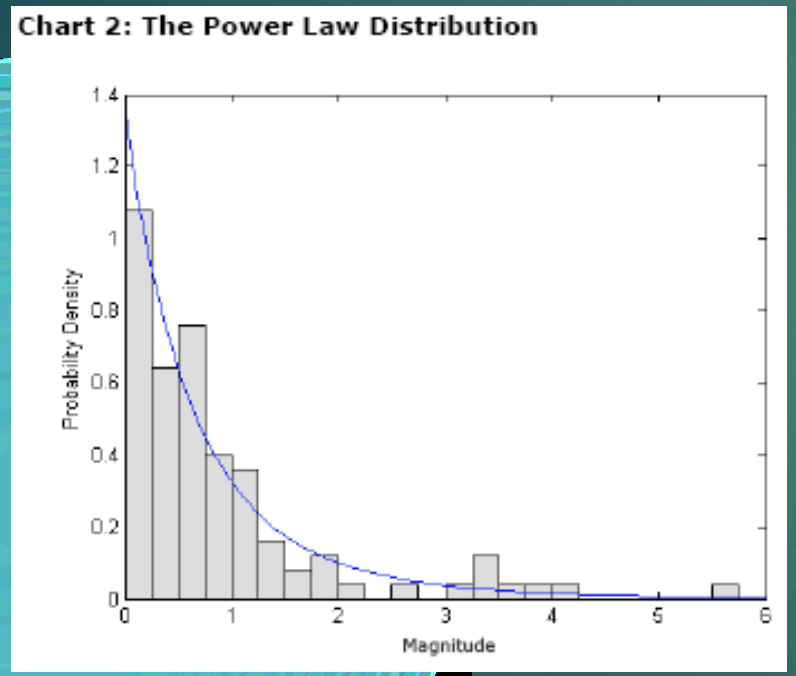


Fig. 3. Relationship between field intensity and the percentage of acrosome loss in alive spermatozoa (black squares) or the percentage of spermatozoa able to respond to sZP coincubation with AR (i.e., capacitated cells) (black circle). All the values are represented as mean ± SD.



Non linear model: power law

$$y = ax^{-b}$$



Logistic model

- La **regressione logistica** è un caso particolare di modello lineare generalizzato avente come funzione link la funzione logit. Si tratta di un modello di regressione applicato nei casi in cui la variabile dipendente y sia di tipo dicotomico riconducibile ai valori 0 e 1, come **lo sono tutte le variabili che possono assumere esclusivamente due valori: *vero o falso*,*maschio o femmina*, *vince o perde*, *sano o ammalato*, ecc**.

Logit

Da Wikipedia, l'enciclopedia libera.

Il **logit** è una funzione, che si applica a valori compresi nell'intervallo (0,1), tipicamente valori rappresentanti probabilità. Viene definito come

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)$$

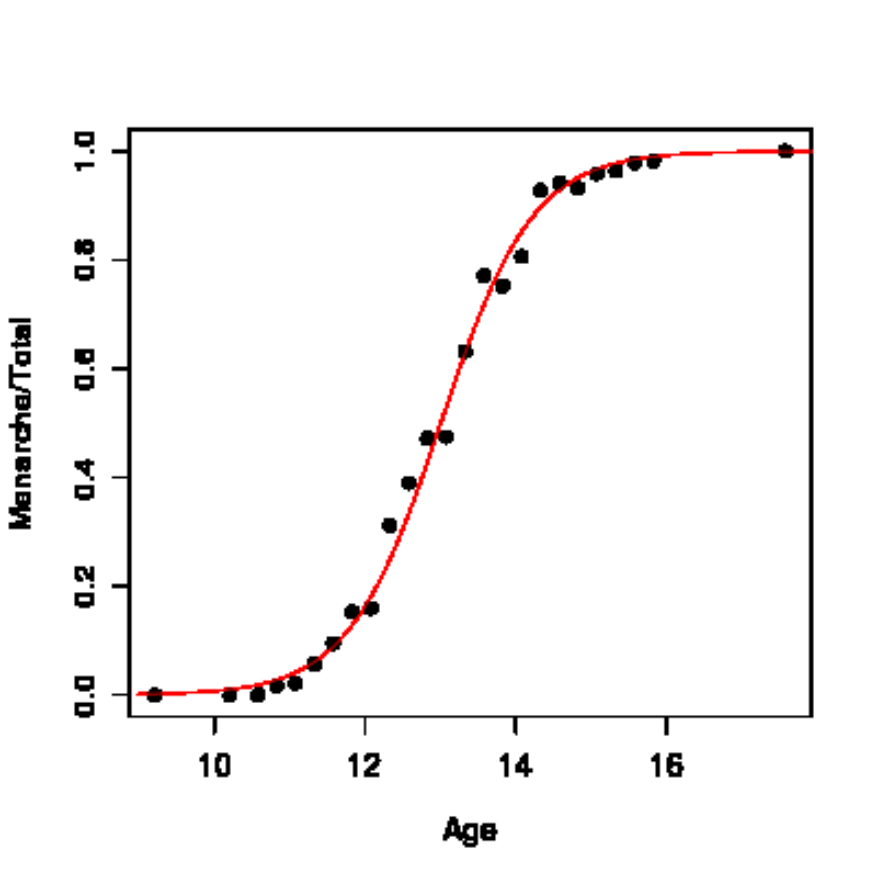
dove *ln* è il logaritmo naturale e $\frac{p}{1-p}$ è detto odds.

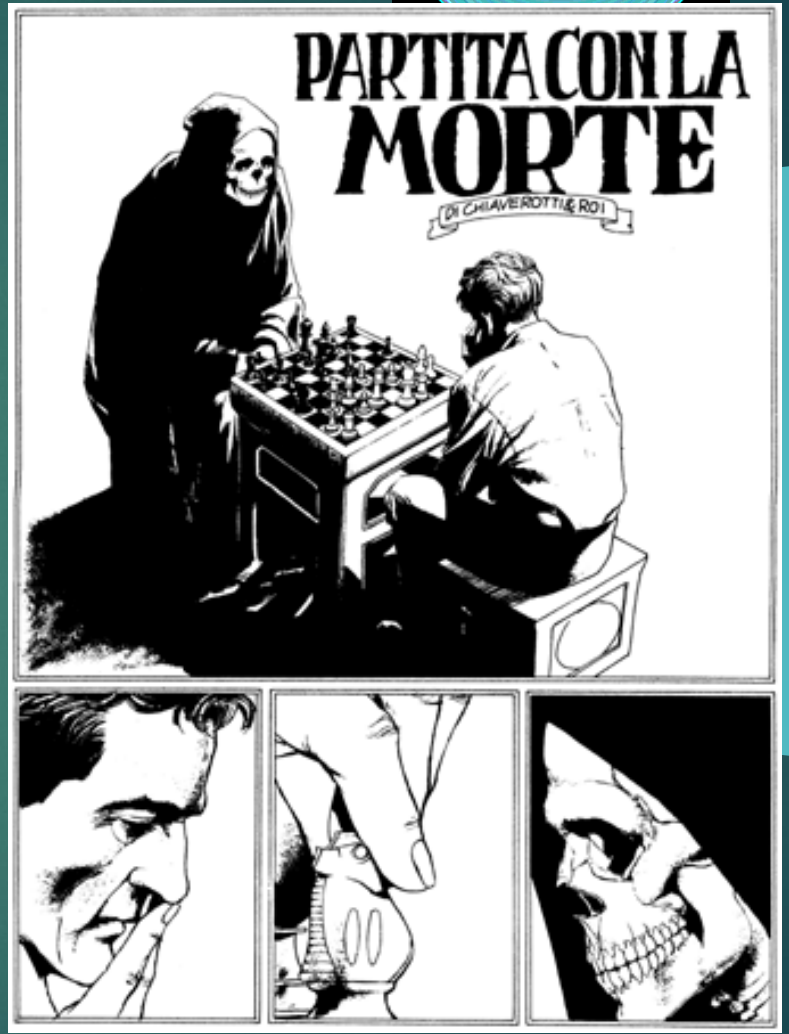
Ha come funzione inversa

$$p = \frac{e^{\operatorname{logit}}}{1 + e^{\operatorname{logit}}}$$

La funzione logit si applica ad esempio nella regressione logistica e nella variabile casuale logistica.

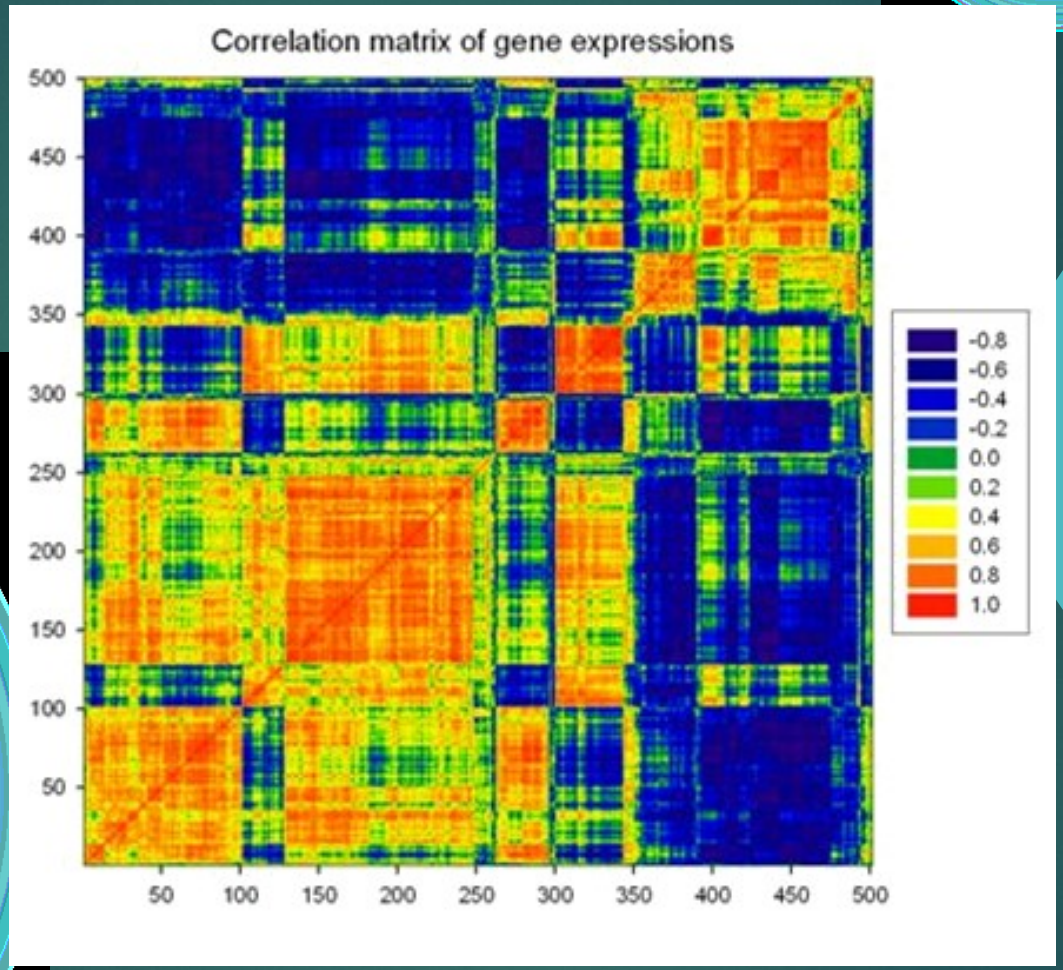
Example…


Menarche/Total
0.0
0.2
0.4
0.6
0.8
1.0
10
12
14
16
Age


PARTITA CON LA MORTE
DI CHIAVE ROTTI & ROI

important

- **Correlation by itself does not imply a cause and effect relationship!!!!!!!**


Correlation matrix of gene expressions
500
450
400
350
300
250
200
150
100
50
50
100
150
200
250
300
350
400
450
500
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0