Logit Analysis

Using vttown.dta

Albert Satorra, UPF

Logit Regression

The theory of Logit regression

The logistic regression model is the one for which the probability p_i of response = 1 in case ith is equal to

$$\log (p_i/(1-p_i)) = x_i' \text{ beta}$$

where x_i is a p-vector of covariates for individual ith, and beta is a p-vector of regression coefficients (typically, the first component of x_i is the constant equal to 1). That is, x_i beta is equal to the log of the odds ratio. It is this that motivates the name "logistic".

Estimation is undertaken by the ML method. Given the data (y_i, x_i), i=1,2, ..., n, the Likelihood function is

$$L = \prod_{i=1}^{n} p^{yi} (1 - p_i)^{1-yi}, \text{ where } p_i = e^{hi}/ (1 + e^{hi}),$$

with $h_i = x_i'$ beta. The ML estimate of beta is the parameter value that maximizes L as a function of beta. In fact, the maximization is undertaken by way of minimizing -log L [1].

Interpretation of the coefficients

A linear regression model for the logit of the probability of Y = 1

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x1_i + \beta_2 x2_i$$

$$\text{Odds} = \pi_i/(1-\pi_i) = \exp(\beta_0 + \beta_1 x1_i + \beta_2 x2_i) = \exp(\beta_0)\exp(\beta_1 x1_i)\exp(\beta_2 x2_i) = \exp(\beta_0)\exp(\beta_1 x1_i)(\exp(x2_i))^{\beta_2}$$

So when x2 changes from x0 ---> x0 + 1 , the odds change from

$$A := \exp(\beta_0)\exp(\beta_1 x1_i)(\exp(x0))^{\beta_2}$$
$$\text{---}>\exp(\beta_0)\exp(\beta_1 x1_i)(\exp(x0+1))^{\beta_2}$$
$$= \exp(\beta_0)\exp(\beta_1 x1_i)(\exp(x0))^{\beta_2}(\exp(1))^{\beta_2}$$
$$= A\exp(\beta_1)$$

So the change is the multiplying factor $\exp(\beta_1)$.

The value $\exp(\beta_1)$ is call the odds-ratio, and its estimate is reported when in Stata we invoke "logistic" instead of "logit" . The appropriate s.e. for the odds are produced. The terminology "odds ratio" follow from the fact that $\exp(\beta_1)$ is the ratio of the odds at two values of X2 when X2 is 0 or 1:
Indeed

Y	0	1
X2		
0	1- p0	p0
1	1-p1	p1

The odds ratio is $p1/(1-p1) : p0/(1-p0)$

$$= [\exp(\beta_1)\, p0/(1-p0)] : p0/(1-p0) = \exp(\beta_1)$$

Note that $p1/(1-p1) : p0/(1-p0) = p1(1-p0) : p0(1-p1)$,

Thus, in % change, we have

$$100 \times (A \exp(\beta_1) - A)/A = 100 \times (\exp(\beta_1) - 1)$$

For example, when $\exp(\beta_1) = .58$ then $\exp(\beta_1) - 1 = -.42$, so we can say that Odds decline by 42 % when x2 increase one unit.
When $\exp(\beta_1) = 1.42$ then $\exp(\beta_1) - 1 = .42$ and we can say that the odds increase by 42 % when x2 increases one unit.

Odds ratio

- The most common way of interpreting a logit is to convert it to an odds ratio using the exp() function. One can convert back using the ln() function.

- An odds ratio above 1.0 refers to the odds that Y = 1 in binary logistic regression. The closer the odds ratio is to 1.0, the more the independent variable's categories (ex., male and female for gender) are independent of the dependent variable, with 1.0 representing no association.

For instance:

If in the logit regresion, $b1 = 2.303$, the corresponding odds ratio is $\exp(2.303) = 10$, then we may say that when the independent variable increases one unit, the odds that the dependent = 1 increase by a factor of 10 (i.e., an increase of 100(10-1) per cent, 900 %) when other variables are controlled. If b1 = -1.5 then the odds of Y = 1 decrease by a factor of $\exp(-1.5) = 0.22$, i.e. a decrease of 100(0.22 - 1) per cent (-88%).

In SPSS, odds ratios appear as "Exp(B)" in the "Variables in the Equation" table.

Pseudo R2 (measures of overall fit)

1) One pseudo-R2 (Aldrich and Nelson, 1984, see Hamilton, p. 233) is just

$$\text{pseudo-R2} = 2(\log L - \log L0)/ [2(\log L - \log L0) + n]$$

where n is the sample size.

2) McFadden (1974) proposes

$$\text{pseudo-R2} = 1 - (\log L0 / \log L)^{1/n}$$

3) The SAS program uses the following pseudo-R2:

$$\text{pseudo R2} := (2(\log L - \log L0) - 2k) / -2 \log L0$$

where L is the likelihood under the full model and L0 is the likelihood with the intercept only, and k is the number of regressors (not including the intercept).

Partial correlation and Wald Test of SPSS:

R statistic (bounded between -1 and 1)

R := sqrt [(Wald statistic - 2k) / -2 log L0]

where k is the number of regressors considered (usually just one variable) and L0 is the likelihood in the case of only the intercept. With the sign of the regression coefficient.

In SPSS instead of z-test we have Wald test which is Wald = $(\text{estimate}/\text{se})^2$

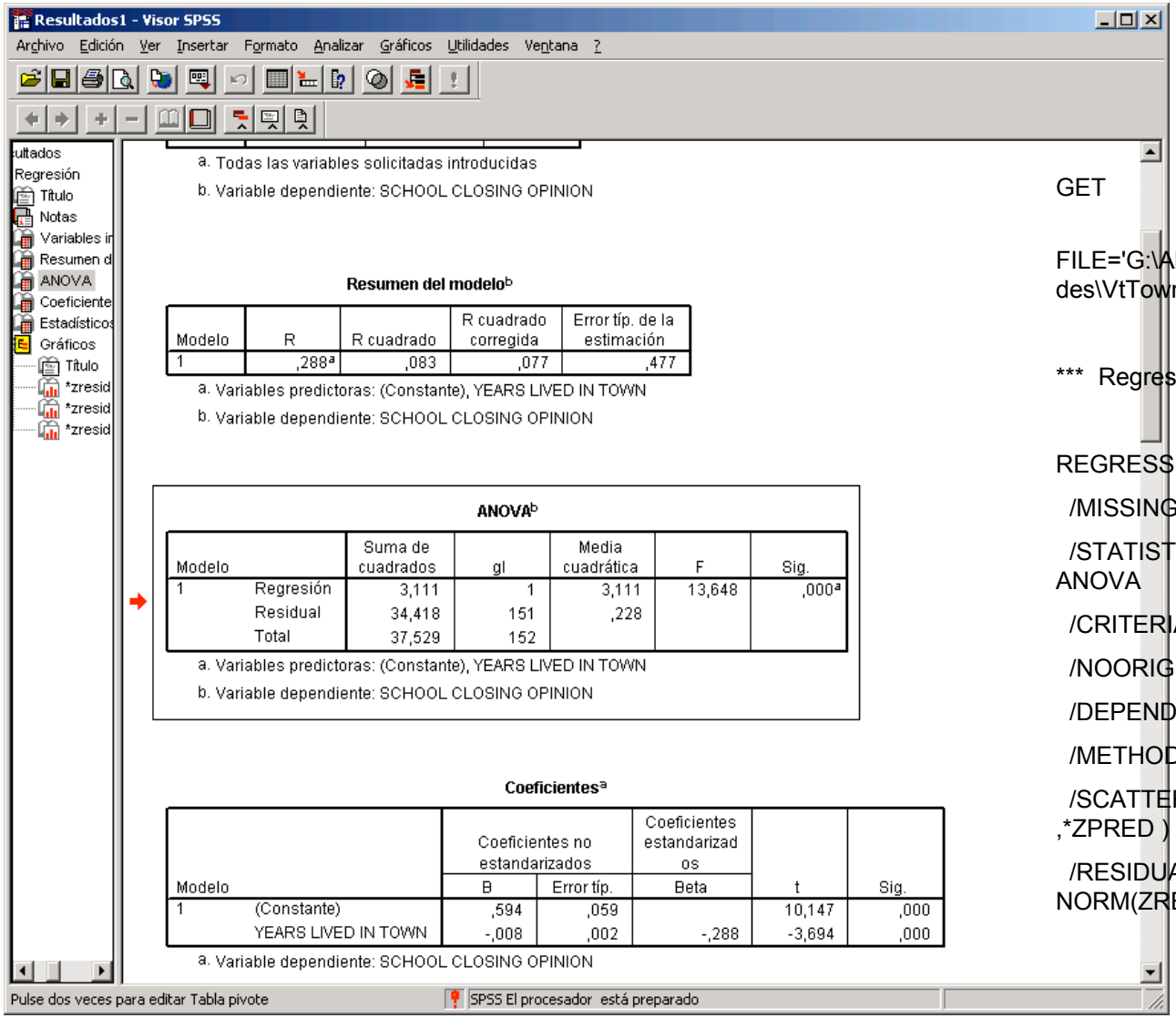
Dades Vermont town:

Survey data from a small Vrmont town where toxic wastes had contaminated the grounds of two public schools. Some people on town think that the school should be closed until proven safe. Others opposed this costly step and said that the schools should stay open.

School: 1= close , 0 = open
Lived: years respondent had lived in hte wown
Eudc: respondent's education, in years
Contam: whether respondent believed his/her own property or water had been affected by the chemical contamination
Meetings: whether respondent attended meetings of the Health and Safety Committee, a citizen's group that organized in response to the contamination crisis.

```
. use "C:\DATA\Hamilt2\SWS5\Vttown.dta", clear
(VT town survey (Hamilton 85))

. describe

Contains data from C:\DATA\Hamilt2\SWS5\Vttown.dta
  obs:           153                          VT town survey (Hamilton 85)
 vars:             7                          16 Dec 1996 09:36
 size:         1,683 (99.7% of memory free)
-------------------------------------------------------------------------------
   1. gender      byte   %8.0g       sexlbl     Respondent's gender
   2. lived       byte   %8.0g                  Years lived in town
   3. kids        byte   %8.0g       kidlbl     Have children <19 in town?
   4. educ        byte   %8.0g                  Highest year school completed
   5. meetings    byte   %8.0g       kidlbl     Attended 2 or more meetings
   6. contam      byte   %8.0g       contamlb   Believe own property/water cont
   7. school      byte   %8.0g       close      School closing opinion
-------------------------------------------------------------------------------
Sorted by:  gender
```

Model Lineal de Probabilitat




```
GET

FILE='G:\Albert\Web\Metodes2005\Dades\VtTown.sav'.

***  Regressió lineal

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT school
  /METHOD=ENTER lived
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

Residus ?

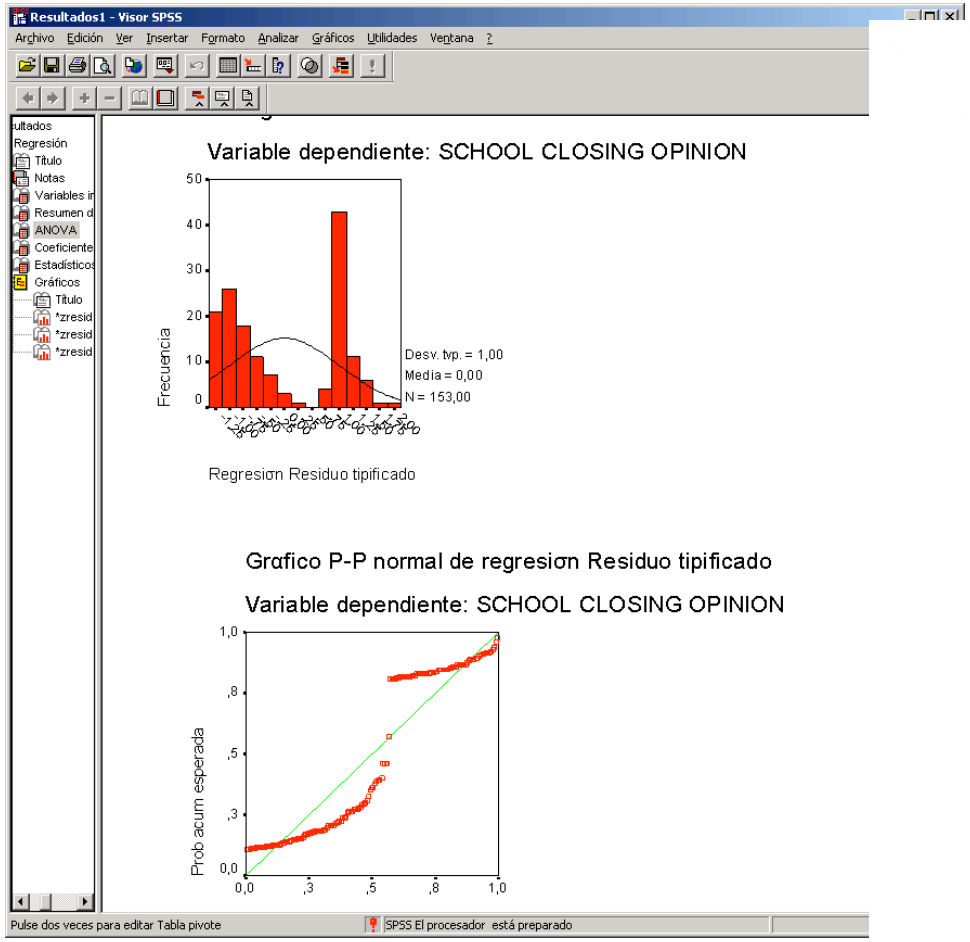
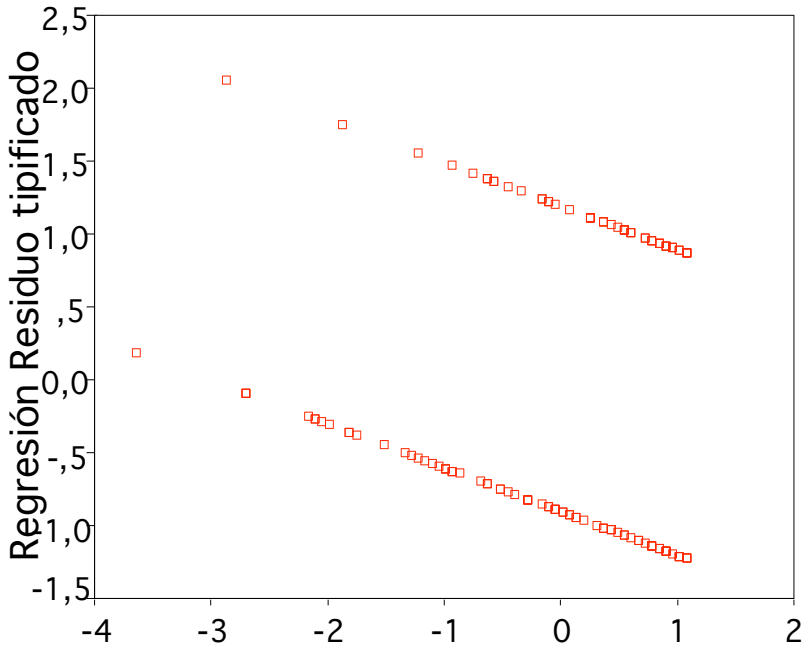
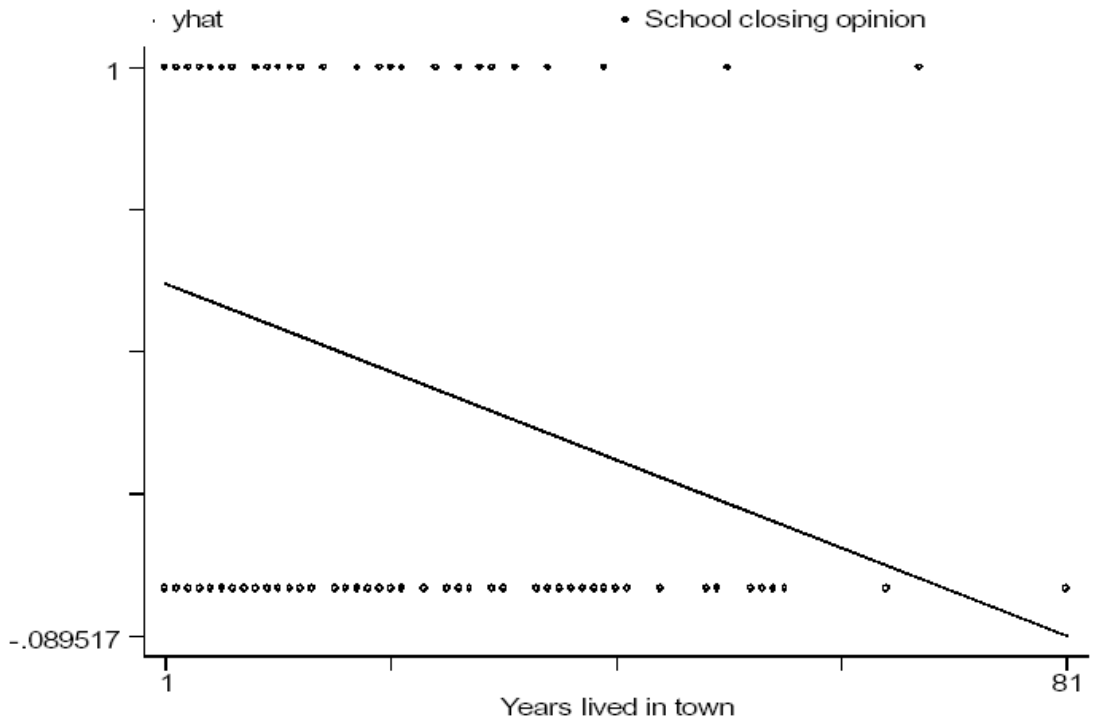


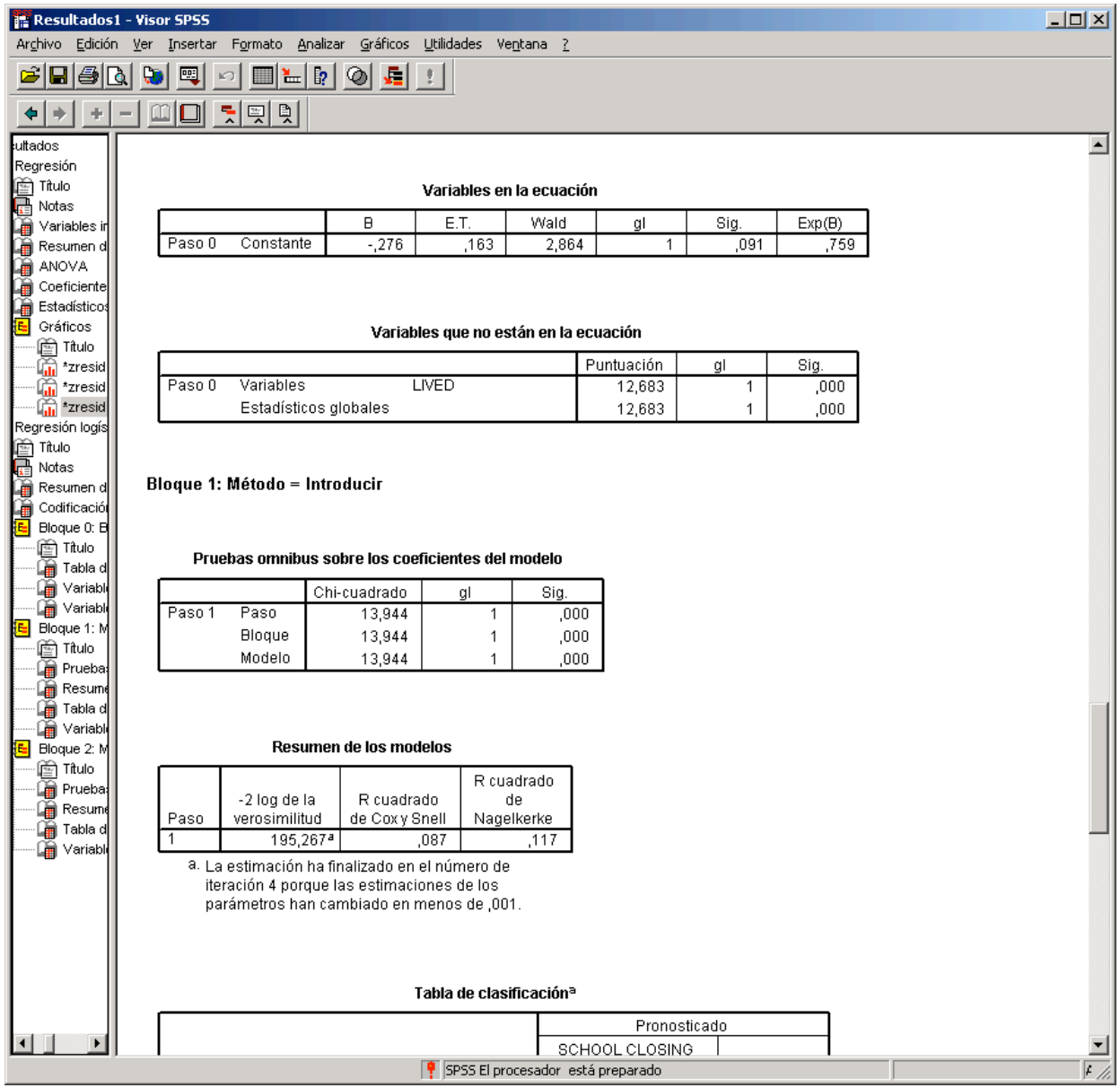
Gráfico de dispersión

Variable dependiente: SCHOOL

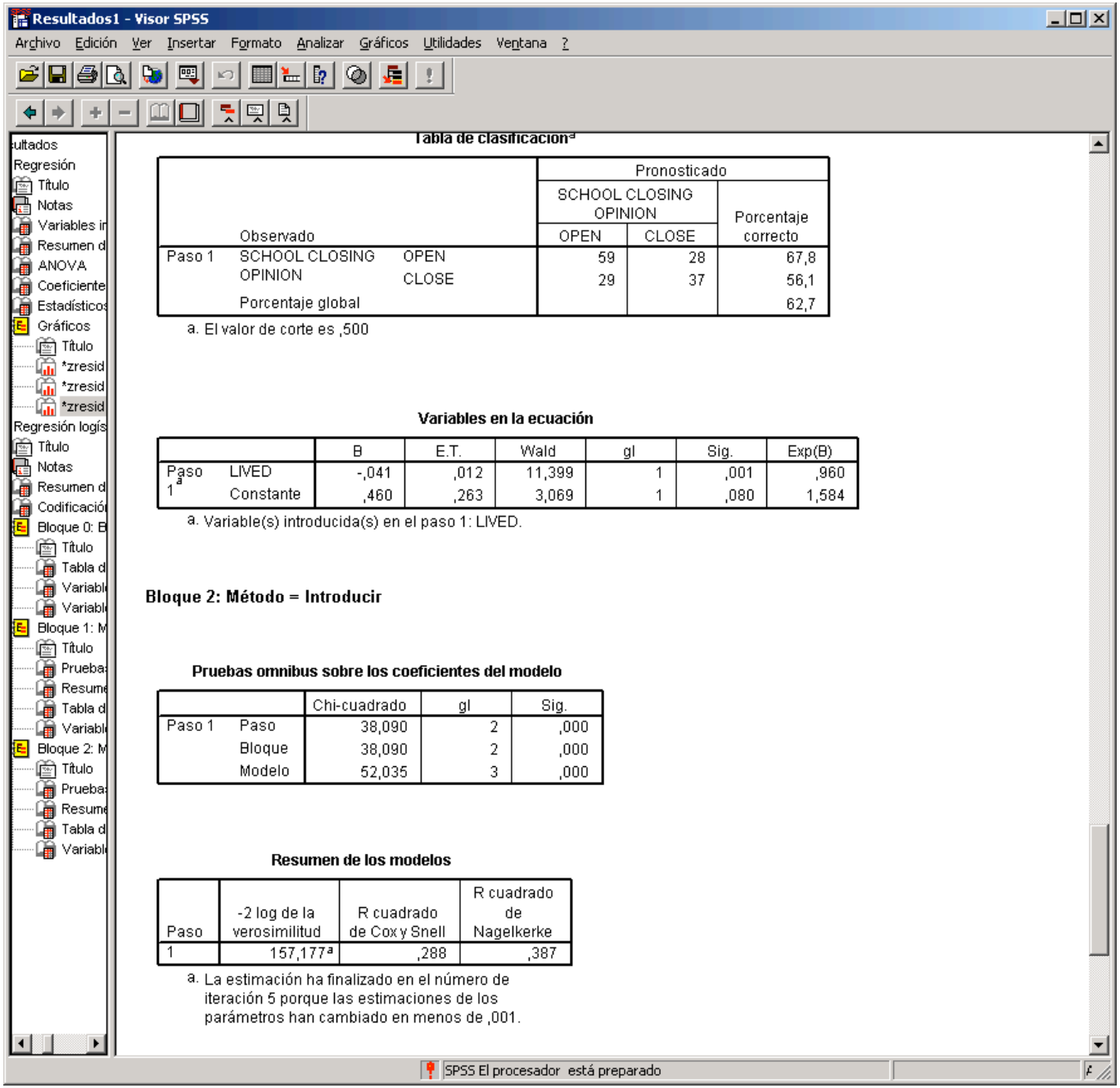


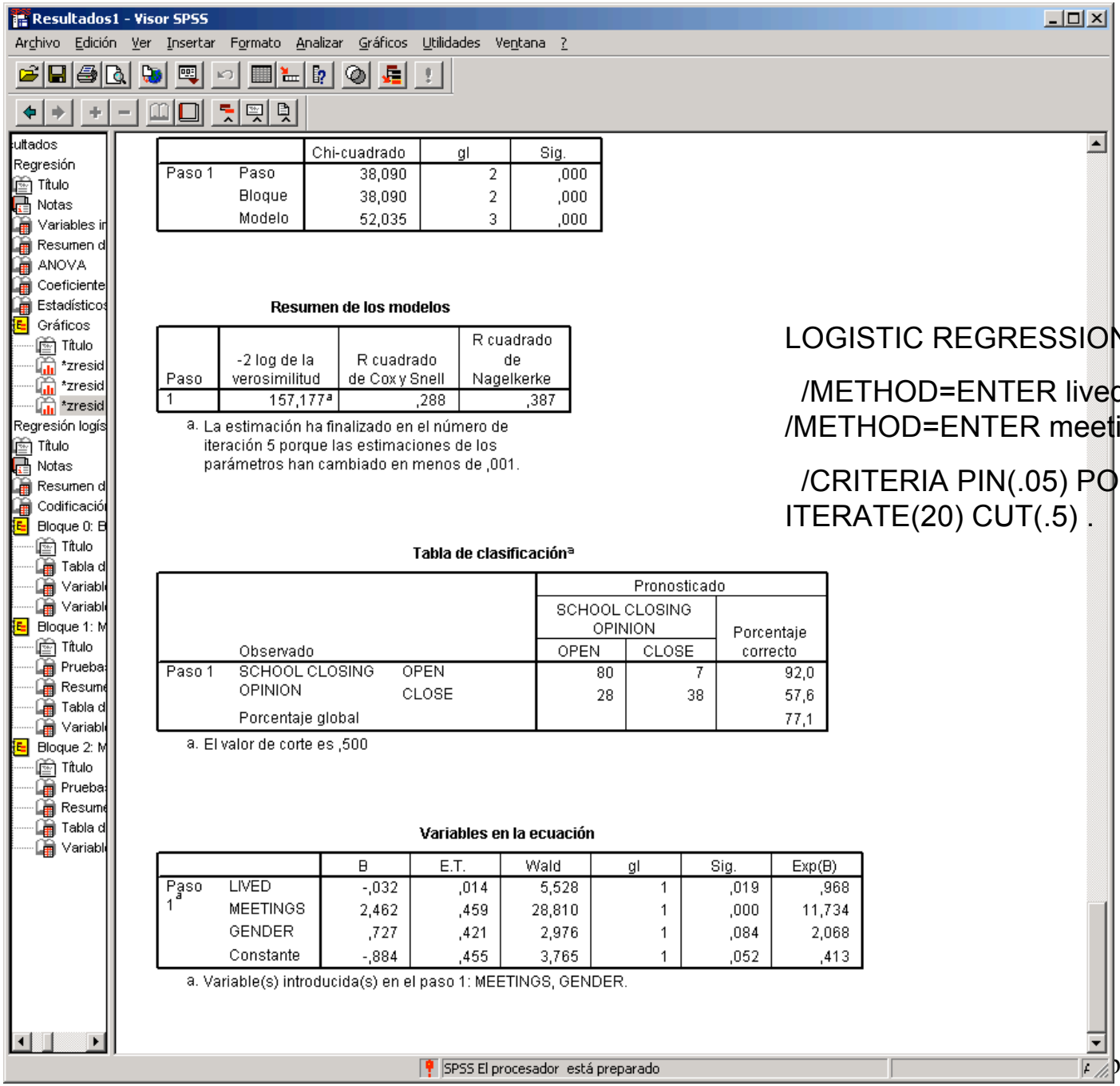
```
  school |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+------------------------------------------------------------------
   lived |  -.0084379    .002284     -3.694   0.000      -.0129506   -.0039252
   _cons |    .593954   .0585363     10.147   0.000       .4782981    .7096099
------------------------------------------------------------------------------

. predict yhat

. graph yhat school lived, c(l.) s(.o)

.
```





Albert Satorra, UPF


Resultados1 - Visor SPSS
Archivo Edición Ver Insertar Formato Analizar Gráficos Utilidades Ventana ?
Tabla de clasificación[a]
Pronosticado
SCHOOL CLOSING OPINION
OPEN
CLOSE
Porcentaje correcto
Observado
Paso 1 SCHOOL CLOSING OPINION OPEN 59 28 67,8
CLOSE 29 37 56,1
Porcentaje global 62,7
a. El valor de corte es ,500
Variables en la ecuación
B E.T. Wald gl Sig. Exp(B)
Paso 1[a] LIVED -,041 ,012 11,399 1 ,001 ,960
Constante ,460 ,263 3,069 1 ,080 1,584
a. Variable(s) introducida(s) en el paso 1: LIVED.
Bloque 2: Método = Introducir
Pruebas omnibus sobre los coeficientes del modelo
Chi-cuadrado gl Sig.
Paso 1 Paso 38,090 2 ,000
Bloque 38,090 2 ,000
Modelo 52,035 3 ,000
Resumen de los modelos
Paso
-2 log de la verosimilitud
R cuadrado de Cox y Snell
R cuadrado de Nagelkerke
1 157,177[a] ,288 ,387
a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.
SPSS El procesador está preparado



		Chi-cuadrado	gl	Sig.
Paso 1	Paso	38,090	2	,000
	Bloque	38,090	2	,000
	Modelo	52,035	3	,000

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	157,177[a]	,288	,387

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación[a]

			Pronosticado		
			SCHOOL CLOSING OPINION		Porcentaje correcto
	Observado		OPEN	CLOSE	
Paso 1	SCHOOL CLOSING OPINION	OPEN	80	7	92,0
		CLOSE	28	38	57,6
	Porcentaje global				77,1

a. El valor de corte es ,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	LIVED	-,032	,014	5,528	1	,019	,968
	MEETINGS	2,462	,459	28,810	1	,000	11,734
	GENDER	,727	,421	2,976	1	,084	2,068
	Constante	-,884	,455	3,765	1	,052	,413

a. Variable(s) introducida(s) en el paso 1: MEETINGS, GENDER.

```
LOGISTIC REGRESSION VAR=school
  /METHOD=ENTER lived
/METHOD=ENTER meetings gender
  /CRITERIA PIN(.05) POUT(.10)
ITERATE(20) CUT(.5) .
```

Logit analysis

```
library(foreign)
data=read.dta("E:/Albert/COURSES/cursDAS/AS2003/data/vttown.dta")
help(glm)
help(family)
attach(data)
names(data)
[1] "gender"   "lived"    "kids"     "educ"     "meetings" "contam"
"school"
results = glm(school ~lived + meetings, family=binomial)
results
Call:  glm(formula = school ~ lived + meetings, family = binomial)
Coefficients:
(Intercept)        lived  meetingsyes
   -0.34850     -0.03575      2.36881

Degrees of Freedom: 152 Total (i.e. Null);  150 Residual
Null Deviance:      209.2
Residual Deviance: 160.3        AIC: 166.3
fv=results$fitted.values
re=results$residuals
plot(fv, re)
logit = results$linear.predictor
```

Logit analysis

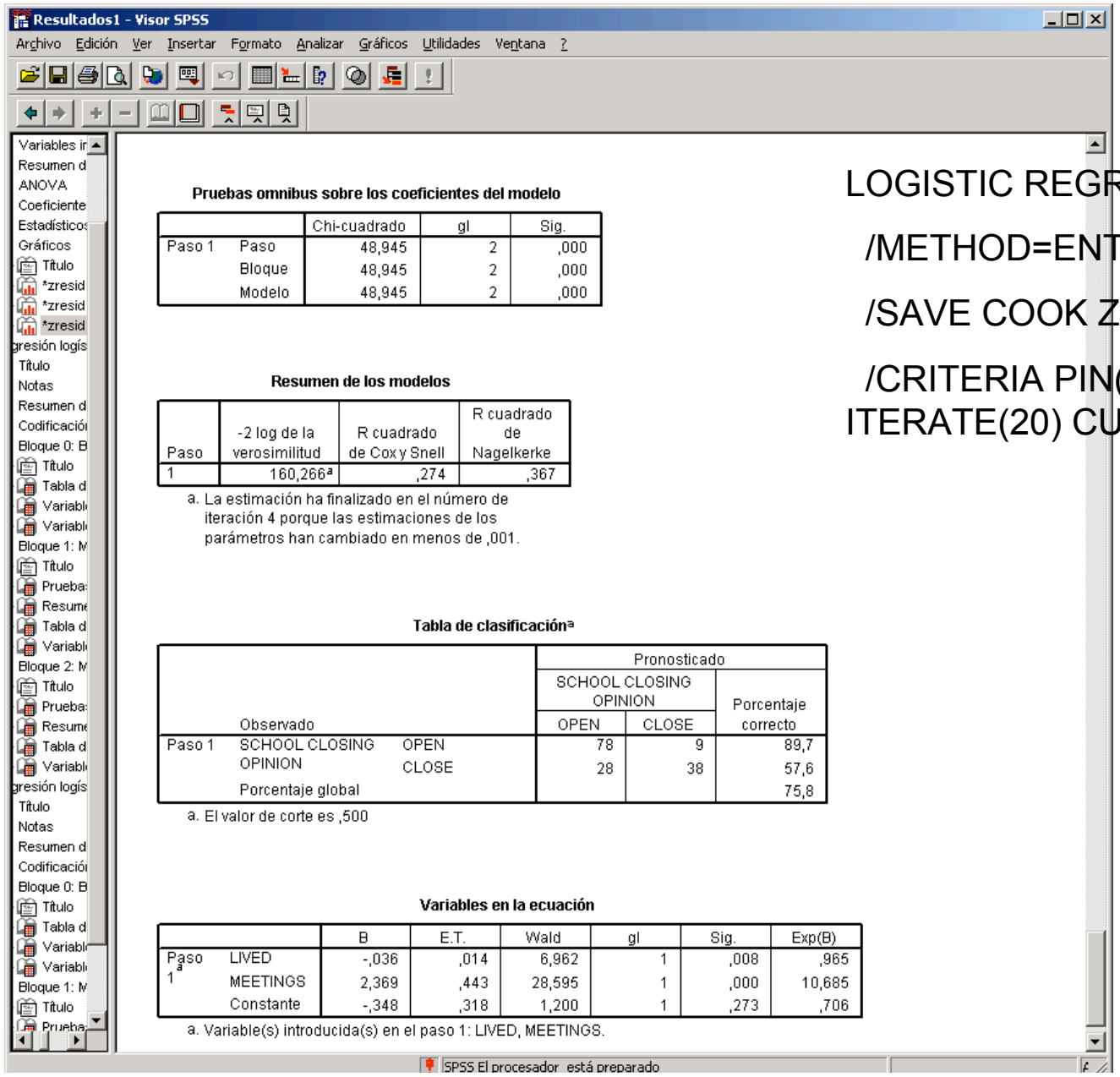
```
summary(results)
Call:
glm(formula = school ~ lived + meetings, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0559  -0.8567  -0.5140   0.6189   2.3832
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34850    0.31796  -1.096   0.2731
lived       -0.03575    0.01352  -2.644   0.0082 **
meetingsyes  2.36881    0.44251   5.353 8.65e-08 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 209.21  on 152  degrees of freedom
Residual deviance: 160.27  on 150  degrees of freedom
AIC: 166.27

Number of Fisher Scoring iterations: 3
```

Residual deviance = -2 log L

The test of the significance of the model is

```
1-pchisq(209.21  -160.27 , 2)
[1] 2.359468e-11
(exp(-0.03575)-1)*100
[1] -3.511852, i.e. 3.5% decrease on the odds when lived → lived +1
```

Logit analysis



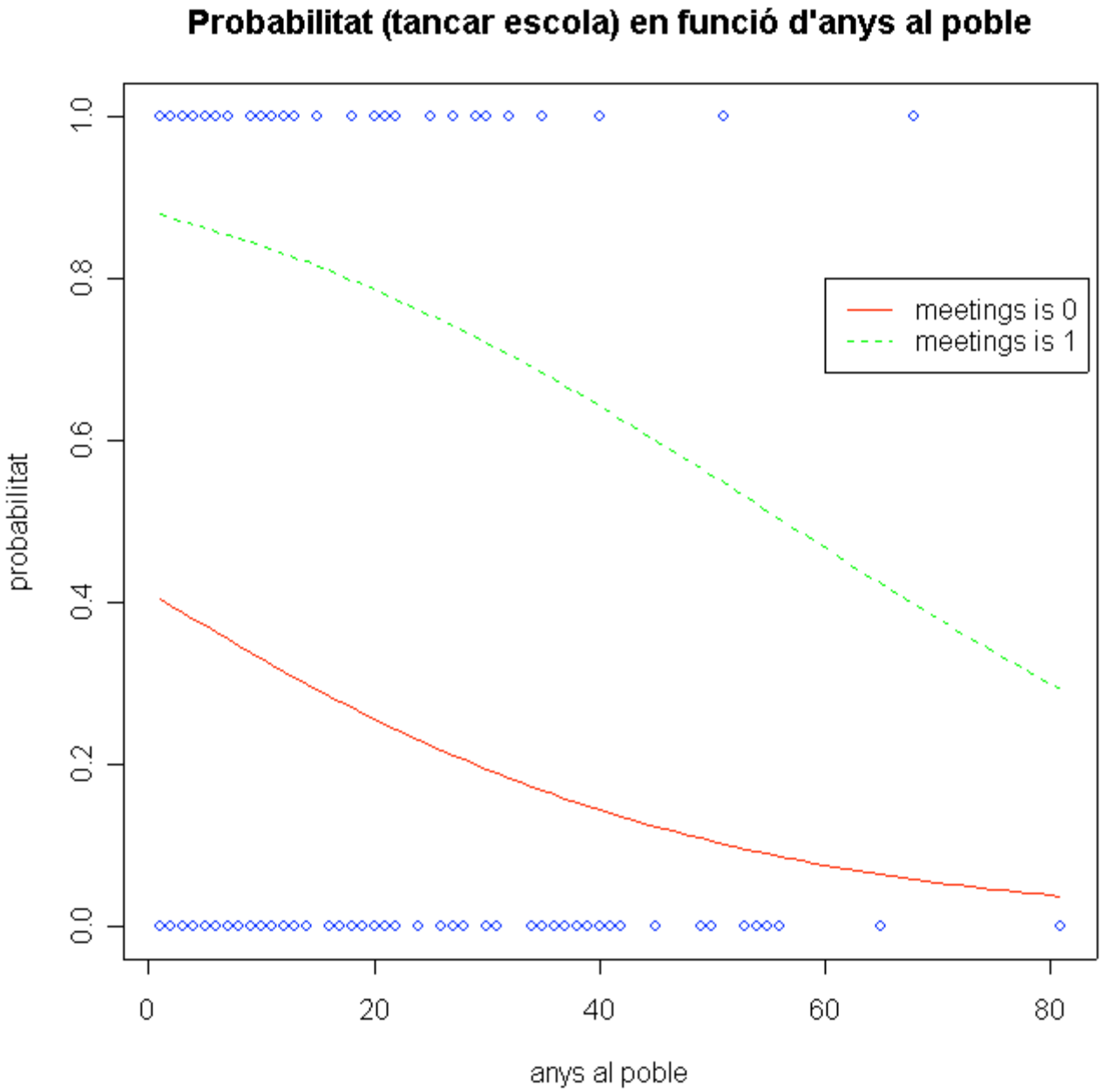

```
LOGISTIC REGRESSION VAR=school
  /METHOD=ENTER lived meetings
  /SAVE COOK ZRESID
  /CRITERIA PIN(.05) POUT(.10)
ITERATE(20) CUT(.5) .
```

Logit regression using R

```
fitlogit =glm(school     ~  lived , binomial)
summary(fitlogit)
ind=sort(lived, index.return=T)$ix
plot(lived[ind],fp[ind], type ="l", col="red", xlab="years living

hatvalues(fitlogit)
dfbetas(fitlogit)
rstudent(fitlogit)

plot(hatvalues(fitlogit),rstudent(fitlogit),type="n")
dfb=dfbetas(fitlogit)[2]
points(hatvalues(fitlogit),rstudent(fitlogit), cex = 10*dfb/max(df
 abline(h =c(-1,0,1), lty=2)
 abline(v= =c(.015,.030), lty=2)
abline(v= c(.015,.030), lty=2)
identify(hatvalues(fitlogit),rstudent(fitlogit), 1:length(rstudent
```

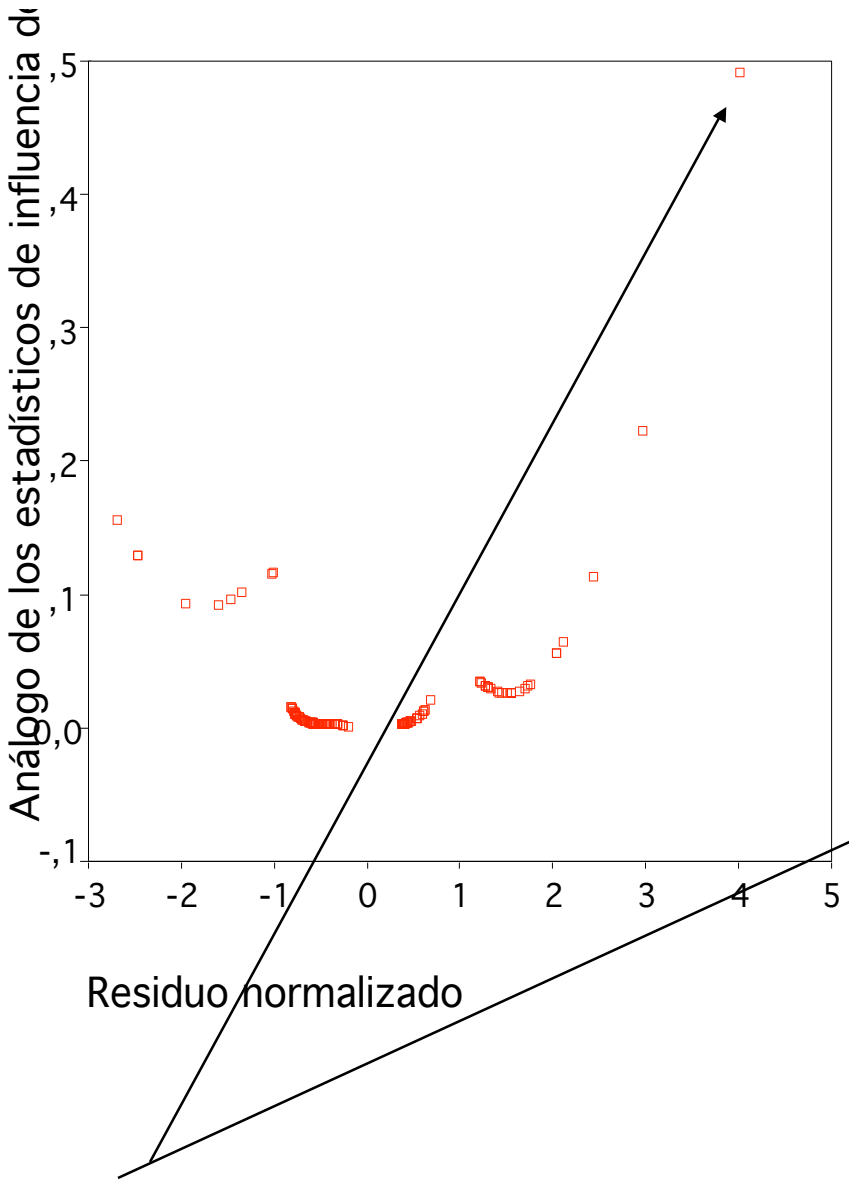
Making a conditional effect plot

```
### making a plot
x=seq(1,81,1)
logit0 =-.3485 -.0358*x +2.3688*0
logit1 =-.3485 -.0358*x +2.3688*1
p0=1/(1+exp(-logit0))
p1=1/(1+exp(-logit1))
library(foreign)
data=read.spss("I:/pol/Metodes/DADES/VtTown.sav")
names(data)
attach(data)
DS=rep(0,length(SCHOOL))
DS[SCHOOL=="CLOSE"] =1
plot(LIVED,DS, col="blue", main="Prob. en funció d'anys al poble",
cex=.8,xlab="anys al poble", ylab="probabilitat")
lines(x,p0,col="red", lty=1 )
lines(x,p1,col="green", lty=2  )
 legend(60,.8,c("meetings is 0", "meetings is 1"), lty=c(1,2), col=c("red",
"green"), cex=.8)
#### abline(lm(DS ~LIVED), col="orange")
```


Probabilitat (tancar escola) en funció d'anys al poble
probabilitat
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
anys al poble
meetings is 0
meetings is 1

orra, UPF

Cook vs residuo normalizado



VtTown.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades
Ventana ?

1 : id 1

	school	coo_1	zre_1	var
134	0	,15622	-2,69739	
135	0	,00346	-,57715	
136	1	,02981	1,70195	
137	0	,00324	-,50927	
138	1	,02636	1,44902	
139	0	,00381	-,60895	
140	0	,00399	-,61993	
141	0	,00324	-,50927	
142	0	,00326	-,42590	
143	0	,01443	-,81058	
144	0	,00957	-,75465	
145	0	,00510	-,66588	
146	1	,49143	4,01416	
147	1	,00336	,39821	
148	1	,03454	1,21182	
149	0	,00367	-,59816	
150	1	,00770	,53962	
151	1	,03343	1,23368	
152	1	,00722	,53006	
153	1	,00348	,40539	
154				

Vista de datos Vista de varia

SPSS El prc

Multinomial Logit Regression

```
plogit <- function(x) 1/(1+exp(-x))
eta <- seq(-10, 10, len=100)
p1 <- plogit(eta-1)
p2 <- plogit(eta+1)
p3 <- plogit(eta+4.5)
plot(c(-10,10), range(p1,p2,p3), type="n", axes=FALSE, xlab="x", ylab="Pr(y > j)")
axis(2)
box()
abline(h=c(0,1), col="gray")
lines(eta, p1, lwd=2)
lines(eta, p2, lwd=2)
lines(eta, p3, lwd=2)
coords <- locator(2)
arrows(coords$x[1], coords$y[1], coords$x[2], coords$y[2], code=1, length=0.125)
text(coords$x[2], coords$y[2], pos=3, "Pr(y > 1)")
coords <- locator(2)
arrows(coords$x[1], coords$y[1], coords$x[2], coords$y[2], code=1, length=0.125)
text(coords$x[2], coords$y[2], pos=3, "Pr(y > 2)")
coords <- locator(2)
arrows(coords$x[1], coords$y[1], coords$x[2], coords$y[2], code=1, length=0.125)
text(coords$x[2], coords$y[2], pos=3, "Pr(y > 3)")
```

1	149,382(a)	,324	,434

a La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación(a)

	Observado		Pronosticado		
			SCHOOL CLOSING OPINION		
			OPEN	CLOSE	Porcentaje correcto
Paso 1	SCHOOL CLOSING OPINION	OPEN	75	12	86,2
		CLOSE	24	42	63,6
	Porcentaje global				76,5

a El valor de corte es ,500

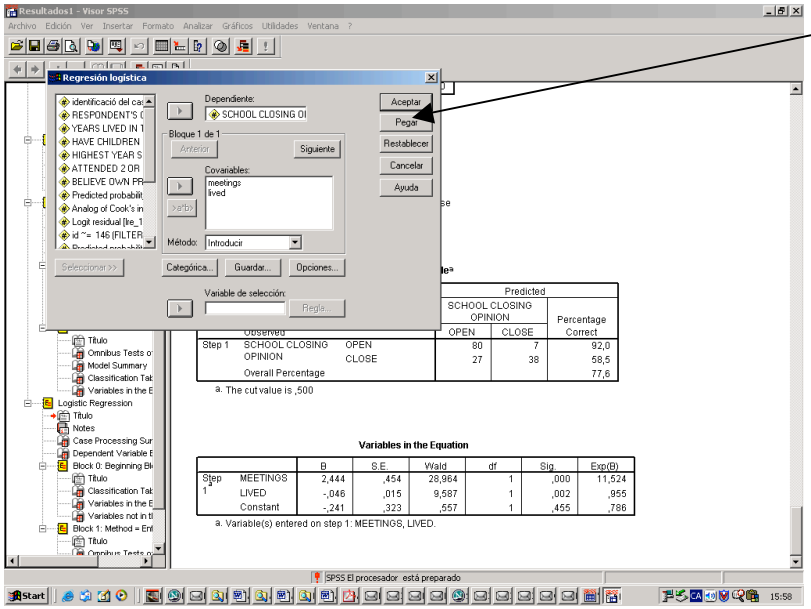
Variables en la ecuación

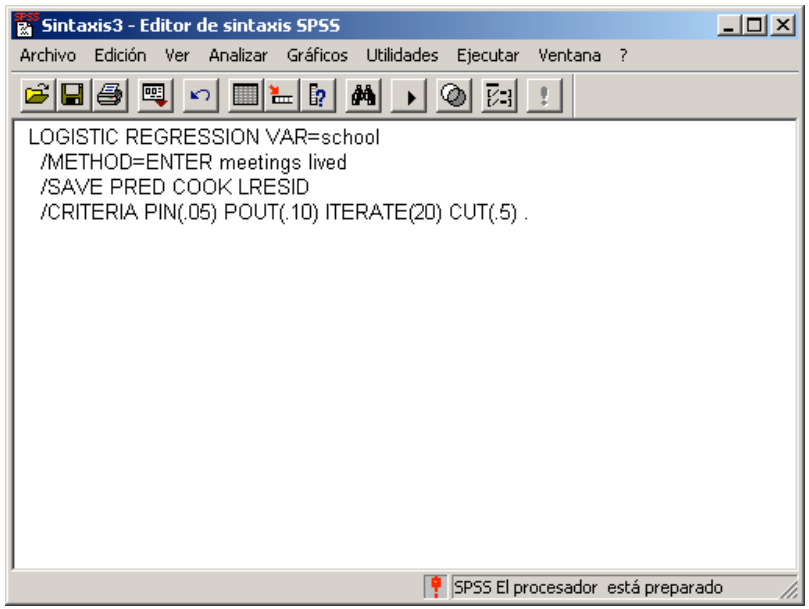
		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1(a)	LIVED	-,046	,015	9,698	1	,002	,955
	EDUC	-,166	,090	3,404	1	,065	,847
	CONTAM	1,208	,465	6,739	1	,009	3,347
	MEETINGS	2,173	,464	21,919	1	,000	8,784
	Constante	1,731	1,302	1,768	1	,184	5,649

a Variable(s) introducida(s) en el paso 1: LIVED, EDUC, CONTAM, MEETINGS.

Sintaxis de SPSS

Activa un fitxer de sintaxis,
que es pot executar parcialment

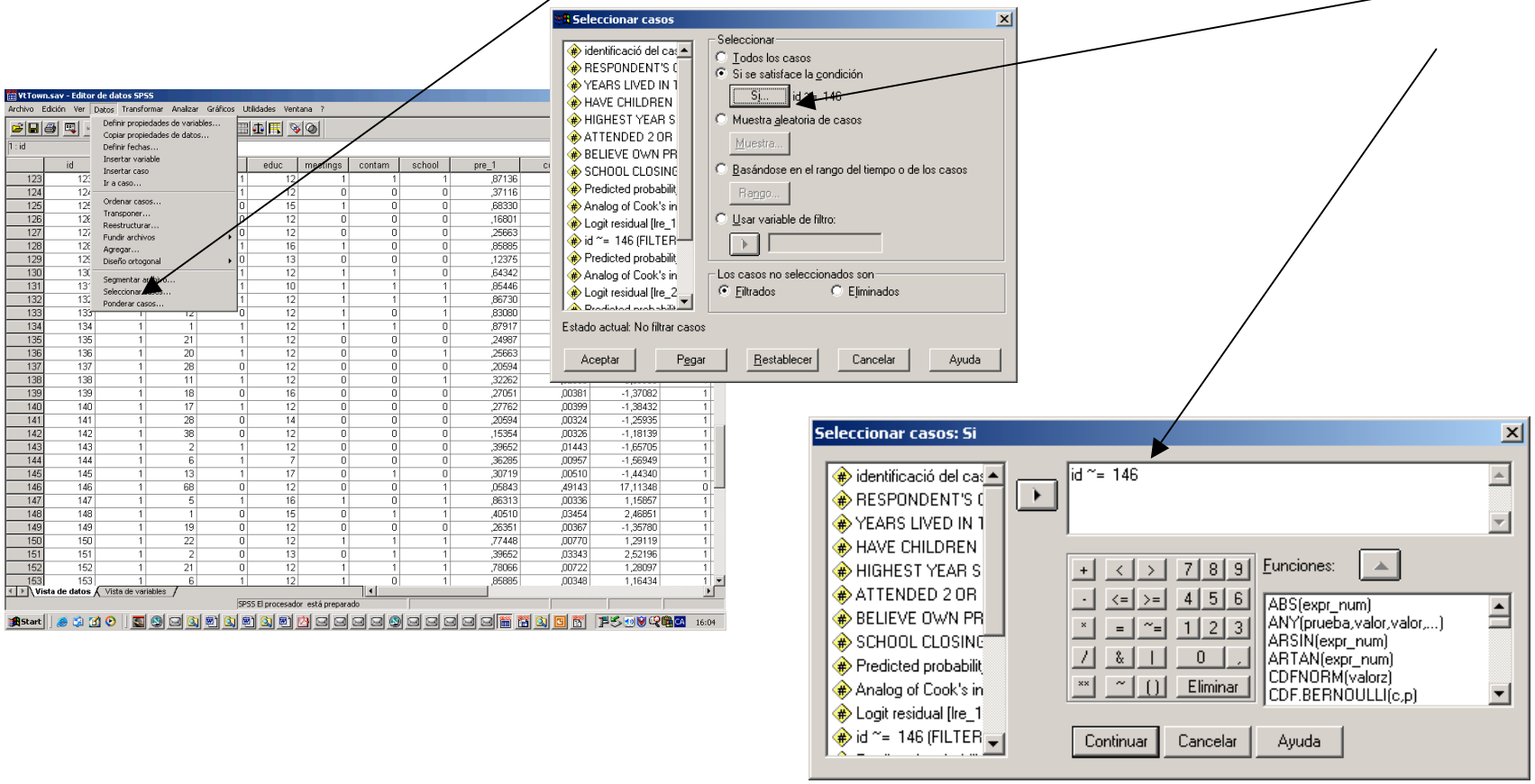




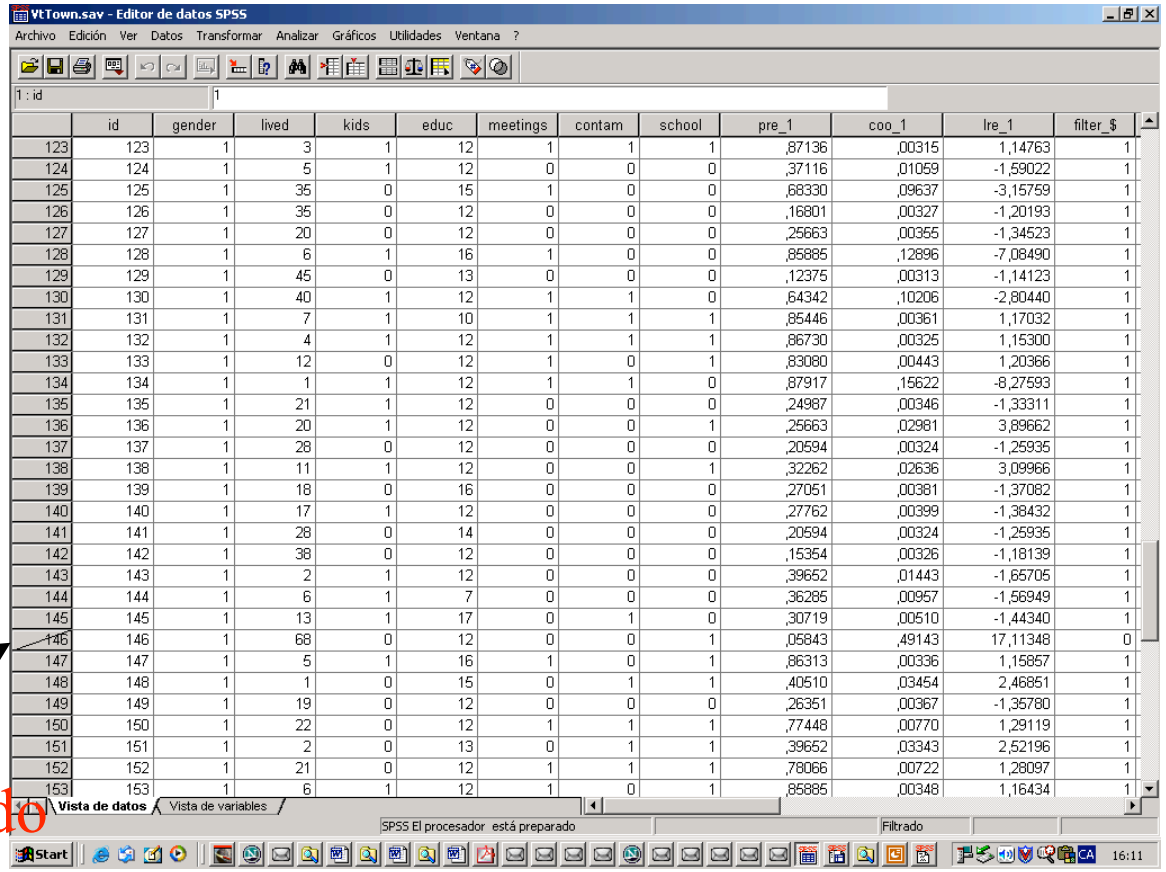
Suprimir casos en el análisis

Selecciona casos

condición



... filtrat de casos



	id	gender	lived	kids	educ	meetings	contam	school	pre_1	coo_1	lre_1	filter_$
123	123	1	3	1	12	1	1	1	,87136	,00315	1,14763	1
124	124	1	5	1	12	0	0	0	,37116	,01059	-1,59022	1
125	125	1	35	0	15	1	0	0	,68330	,09637	-3,15759	1
126	126	1	35	0	12	0	0	0	,16801	,00327	-1,20193	1
127	127	1	20	0	12	0	0	0	,25663	,00355	-1,34523	1
128	128	1	6	1	16	1	0	0	,85885	,12896	-7,08490	1
129	129	1	45	0	13	0	0	0	,12375	,00313	-1,14123	1
130	130	1	40	1	12	1	1	0	,64342	,10206	-2,80440	1
131	131	1	7	1	10	1	1	1	,85446	,00361	1,17032	1
132	132	1	4	1	12	1	1	1	,86730	,00325	1,15300	1
133	133	1	12	0	12	1	0	1	,83080	,00443	1,20366	1
134	134	1	1	1	12	1	1	0	,87917	,15622	-8,27593	1
135	135	1	21	1	12	0	0	0	,24987	,00346	-1,33311	1
136	136	1	20	1	12	0	0	1	,25663	,02981	3,89662	1
137	137	1	28	0	12	0	0	0	,20594	,00324	-1,25935	1
138	138	1	11	1	12	0	0	1	,32262	,02636	3,09966	1
139	139	1	18	0	16	0	0	0	,27051	,00381	-1,37082	1
140	140	1	17	1	12	0	0	0	,27762	,00399	-1,38432	1
141	141	1	28	0	14	0	0	0	,20594	,00324	-1,25935	1
142	142	1	38	0	12	0	0	0	,15354	,00326	-1,18139	1
143	143	1	2	1	12	0	0	0	,39652	,01443	-1,65705	1
144	144	1	6	1	7	0	0	0	,36285	,00957	-1,56949	1
145	145	1	13	1	17	0	1	0	,30719	,00510	-1,44340	1
146	146	1	68	0	12	0	0	1	,05843	,49143	17,11348	0
147	147	1	5	1	16	1	0	1	,86313	,00336	1,15857	1
148	148	1	1	0	15	0	1	1	,40510	,03454	2,46851	1
149	149	1	19	0	12	0	0	0	,26351	,00367	-1,35780	1
150	150	1	22	0	12	1	1	1	,77448	,00770	1,29119	1
151	151	1	2	0	13	0	1	1	,39652	,03343	2,52196	1
152	152	1	21	0	12	1	1	1	,78066	,00722	1,28097	1
153	153	1	6	1	12	1	0	1	,85885	,00348	1,16434	1

Caso suprimido