

Logistic Regression

Data Mining course
Master in Information Technologies
Enginyeria Informàtica

Tomàs Aluja



UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Logistic regression

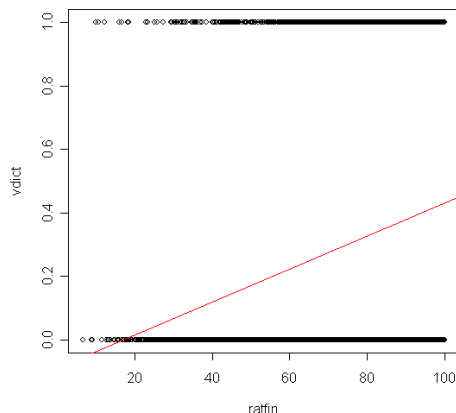
Response variable : Binary

(or a proportion, ordinal variable, nominal variable)

Examples: To buy a product, to pass a course, to obtain a credit, to level the preference for a service, ...

```
> l1 = lm(as.vector(dict) ~ ratfin)
```

Linear Model fit



$$y_i = \begin{cases} 1 & \text{if } + \text{ with } p_i \\ 0 & \text{if } - \text{ with } (1 - p_i) \end{cases}$$

$$E[y_i / x_{i1}, \dots, x_{ip}] = \hat{y} = b_0 + b_1 x_1 + \dots + b_p x_p$$

But with a linear fit we have: $-\infty < \hat{y} < \infty$

For each individual we know y_i , but we would need an estimation of p_i

Violation of linear model hypothesis

How can we model binary responses?

The response is binary 0/1

$$y_i = \begin{cases} 1 & \text{Prob}_i(1) = p_i, \\ 0 & \text{Prob}_i(0) = 1 - p_i. \end{cases}$$

$$y_i = r(\beta_0 + \beta_l x_{il} + \dots + \beta_p x_{ip}) + \varepsilon_i = E[y_i | x_{il}, \dots, x_{ip}] + \varepsilon_i$$

$$E[y_i | x_{il}, \dots, x_{ip}] = 1 \times p_i + 0 \times (1 - p_i) = p_i$$

Lets do a transformation of the linear component : $\beta'x_i$

In such a way that $r(\beta'x_i)$ mapping function in the interval 0:1 (=probability)

$$p_i = r(\beta_0 + \beta_l x_{il} + \dots + \beta_p x_{ip})$$

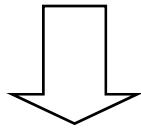
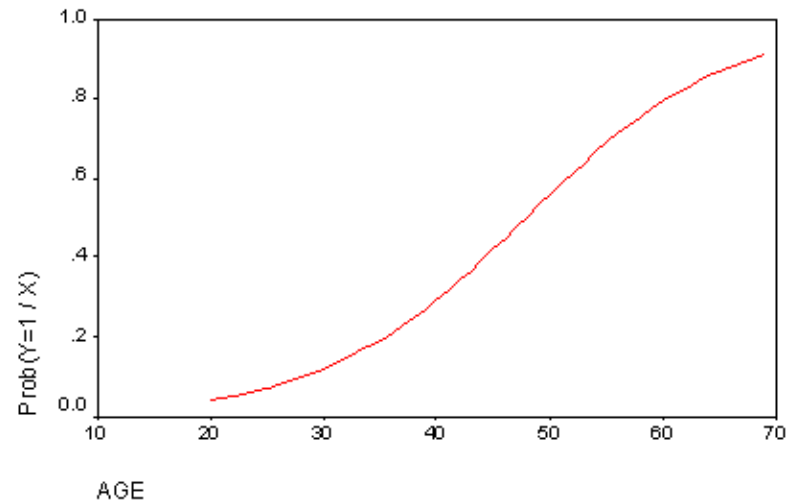
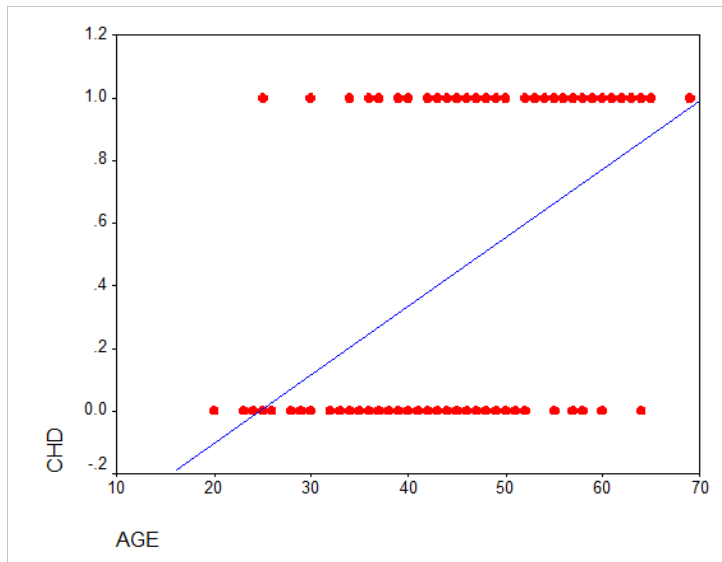
$$\varepsilon_i \text{ binomial} \quad \varepsilon \sim B(n_i, p_i)$$

$$y_i = p_i + \varepsilon_i$$

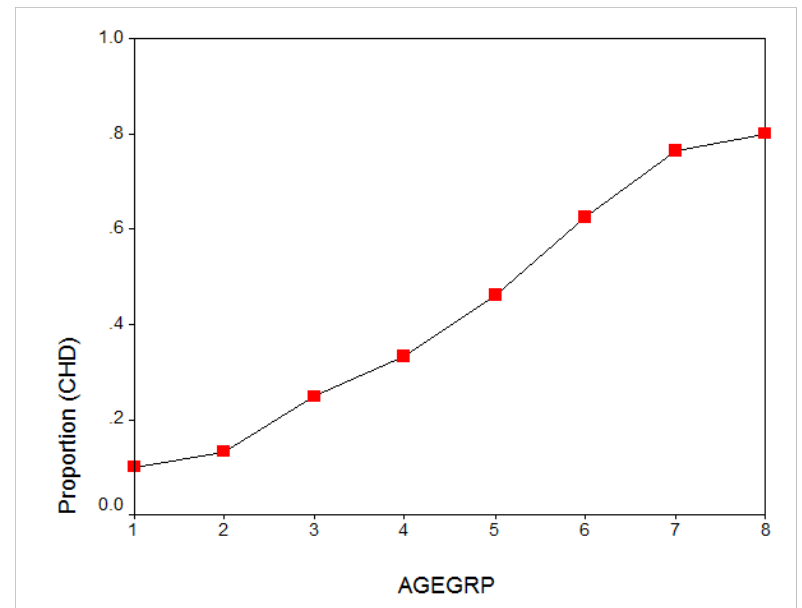
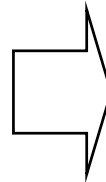
Generalized Linear Model

n_i number of observations in individual i

p_i probability of $y=1$ for individual i



Age Group	n	CHD absent	CHD present	Mean (Proportion)
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Total	100	57	43	0.43



Which r function to choose?

Logistic function

$$p_i = r(\beta' x_i) = \frac{1}{1 + \exp^{-\beta' x_i}}$$

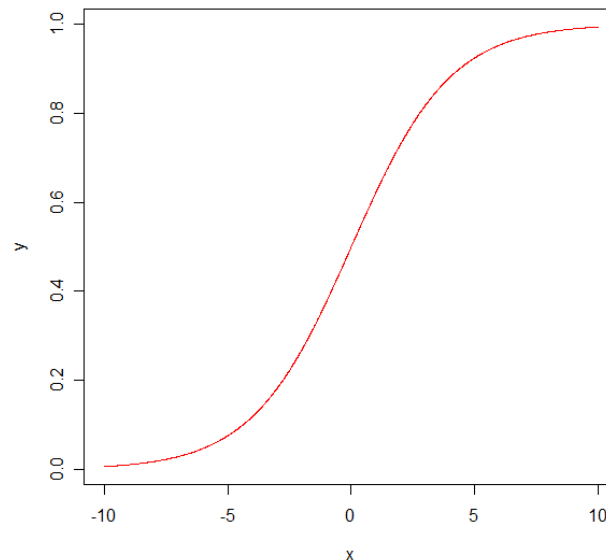
The log odds (=logit) is a linear function of the predictors

$$\ln \frac{p_i}{1 - p_i} = \beta' x_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

$$\ln \frac{P(+ / x_i)}{P(- / x_i)} = \beta' x_i$$

Logistic function is very close to the inverse of the normal distribution function (probit function)

$$p_i = \frac{1}{1 + \exp^{-\beta' x_i}} \cong \Phi^{-1}(\beta' x_i)$$

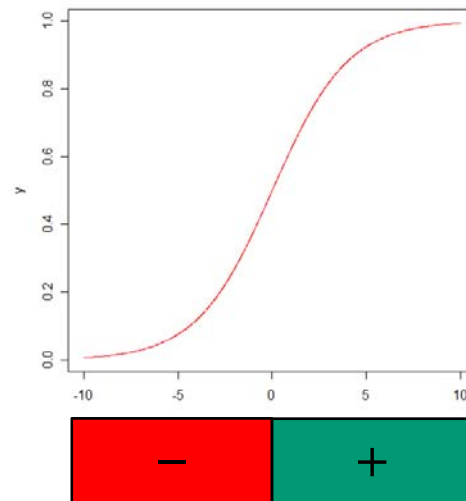


Interpretation of the logistic function

The linear predictor can be interpreted as the propensity to choose the + event

$$\eta_i = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

If η_i is greater than a threshold (i.e. 0) then the individual choose +, otherwise -



In economy the propensity can be formulated as the difference of two *utilities*

$$y_i / x_1, \dots, x_p = 1 \quad \text{if } u(1) > u(0)$$

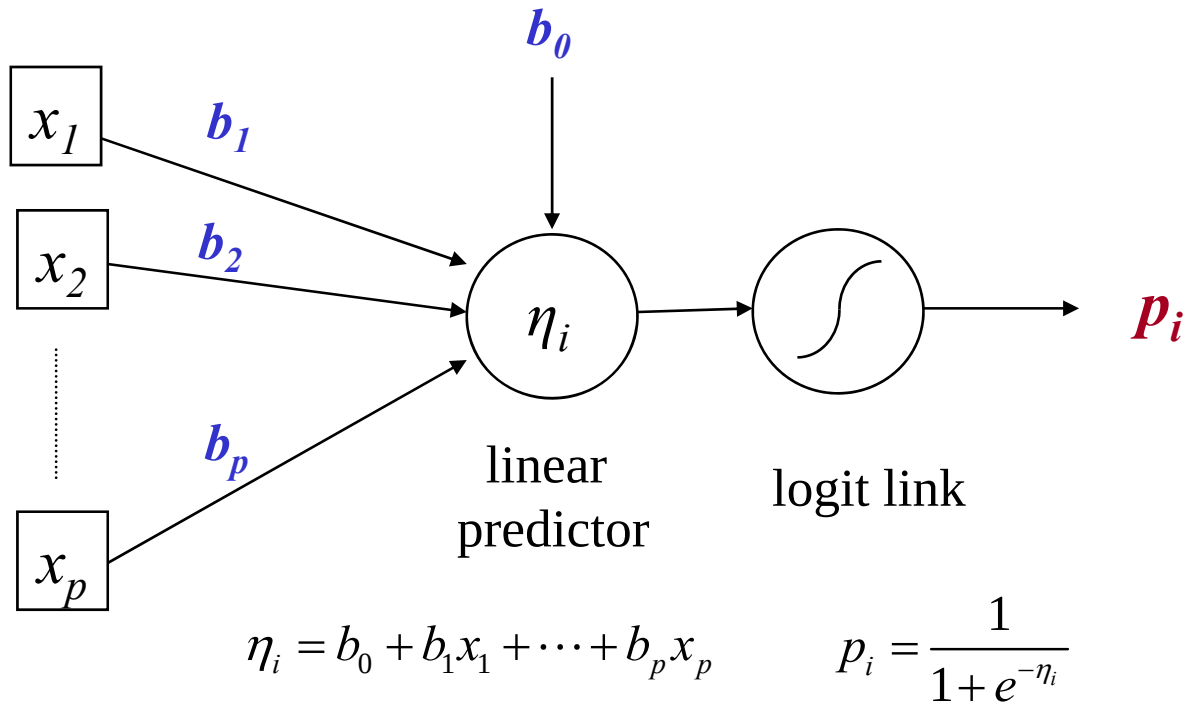
$$y_i / x_1, \dots, x_p = 0 \quad \text{if } u(0) > u(1)$$

Hence, supposing a linear model for the propensity

$$u(1) - u(0) = \beta' x_i + \varepsilon_i$$

$$P(y_i = 1 / x_{i1}, \dots, x_{ip}) = P(u(1) - u(0) > 0) = P(\beta' x_i > -\varepsilon_i)$$

A graphical representation of the logistic regression



But, how to estimate the $b_0, b_1, \dots, b_p$

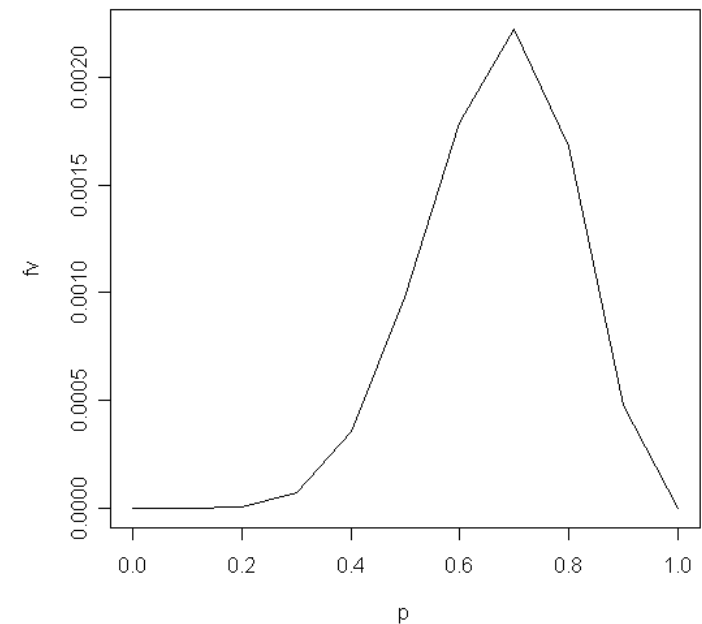
Maximum Likelihood Estimation (MLE) remainder

Choose as estimates of the parameters those who miximizes the probability of the oserved data

$$\text{Max } L(\theta) = \Pr(x_1, \dots, x_n / \theta) = \Pr(x_1 / \theta) \times \dots \times \Pr(x_n / \theta)$$

A silly exemple, estimate the probability of heads in 10 coin tosses if we get 7 heads

```
> n = 10
> n1 = 7
> n0 = n - n1
> p = seq(from=0, to=1, by=0.1)
> p
 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> fv = p^n1*(1-p)^n0
> fv
0.0000000000 0.0000000729 0.0000065536
0.0000750141 0.0003538944 0.0009765625
0.0017915904 0.0022235661 0.0016777216
0.0004782969 0.0000000000
> plot(p, fv, type="l")
```



MLE of the Logistic Regression

$$L(\beta) = \Pr((y_1, x_1), \dots, (y_n, x_n)) = \prod_{i=1}^n \Pr(y_i / x_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

$$\log L(\beta) = l(\beta) = \sum_i^n \log p_i = \sum_i^n (y_i \log p_i + (1 - y_i) \log(1 - p_i))$$

$$p_i^{y_i} (1 - p_i)^{1-y_i} = \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) = \left(e^{-\beta' x_i} \right)^{y_i} \left(1 + e^{\beta' x_i} \right)$$

$$l(\beta) = \sum_i^n (y_i \beta' x_i + \log(1 + e^{\beta' x_i}))$$

$$\frac{\partial l(\beta)}{\partial \beta} = X'(y - p)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} = -X' W X$$

$$W = \begin{bmatrix} \ddots & & \\ & p_i(1 - p_i) & \\ & & \ddots \end{bmatrix}$$

MLE of the Logistic Regression

Newton-Raphson

$$\beta^{t+1} = \beta^t - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \left(\frac{\partial l(\beta)}{\partial \beta} \right)$$

$$\beta^{t+1} = \beta^t + (X'WX)^{-1} X'(y - p) = (X'WX)^{-1} X'Wz$$

$$z = X\beta^t + W^{-1}(y - p)$$

Iterated Reweighted Least Squares (IRLS algorithm)

Initialize $\beta_0 = \log(n_+/n_-)$ $\beta_j = 0, j=1, \dots, p$ (null model)

Iterate till convergence

- Estimate p and W
- Calculate z
- Update β by weighted regression

How can we assess the quality of the logistic regression?

Deviance

Likelihood ratio test of the proposed model respect to the saturated model (the one providing a perfect fit $p_i = y_i$). It can be interpreted as a proximity measure of the model fit respect to the data, in a similar fashion as the sum of residual squares in linear models.

$$D = -2 \sum_{i=1}^n (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \quad D = -2 \log L(\beta) \sim \chi^2_{v=n-p-1}$$

Null deviance: Deviance of the null model (just with constant term)

Residual deviance: Deviance of the proposed model

AIC: Deviance with complexity penalization (+2p)

```

> learn <- sample(1:n, round(0.67*n))
> l3 = glm(dict ~ edat+ratfin+tiptreb, family = binomial, data = dd[learn,])
> summary(l3)
      glm(formula = dict ~ edat + ratfin + tiptreb, family = binomial(link = logit),
        data = dd[learn, ])

Deviance Residuals:
      Min       1Q   Median       3Q      Max
-2.1157  -1.0444   0.4602   1.0010   1.9476

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.515779   0.875162  -0.589 0.555625
edat         0.033935   0.010838   3.131 0.001742 **
ratfin      -0.033892   0.006085  -5.569 2.56e-08 ***
tiptrebauton 1.619291   0.662626   2.444 0.014536 *
tiptrebfixe  2.231853   0.657498   3.394 0.000688 ***
tiptrebtemp  0.562770   0.766715   0.734 0.462948
---

Null deviance: 563.92  on 406  degrees of freedom
Residual deviance: 489.35  on 401  degrees of freedom
AIC: 501.35

Number of Fisher Scoring iterations: 4

```

Could we simplify the model?

```
> step(l3)
Start:  AIC= 501.35
dict ~ edat + ratfin + tiptreb
      Df Deviance   AIC
<none>    489.35 501.35
- edat    1   499.60 509.60
- tiptreb 3   520.95 526.95
- ratfin  1   525.10 535.10
```

```
Call:  glm(formula = dict ~ edat + ratfin + tiptreb, family = binomial(link = logit),
data = dd[learn, ])
```

Coefficients:

(Intercept)	edat	ratfin	tiptrebauton	tiptrebfixe	tiptrebtemp
-0.51578	0.03394	-0.03389	1.61929	2.23185	0.56277

```
Degrees of Freedom: 406 Total (i.e. Null); 401 Residual
```

```
Null Deviance:      563.9 Residual Deviance: 489.3      AIC: 501.3
```

Interpreting the coefficients of a logistic regression

Lets take one predictor $x=0,1$ (i.e. 0 =single, 1 =married)

The odds for a married person, express how much likely is having a positive dictamen rather than a negative dictament being married

Likewise, ...

$$\frac{\Pr(+ / x = 1)}{\Pr(- / x = 1)} = e^{\beta_0 + \beta_1}$$

$$\frac{\Pr(+ / x = 0)}{\Pr(- / x = 0)} = e^{\beta_0}$$

The ODDS RATIO

$$\frac{\Pr(+ / 1) / \Pr(- / 1)}{\Pr(+ / 0) / \Pr(- / 0)} = e^{\beta_1}$$

The exponential of the β_1 coefficient measures the change in the odds of + against -, when passing from $x=0$ to $x=1$

The obtained model:

$$\log \frac{p_i}{1-p_i} = -0.51578 + 0.03394 \text{edat} - 0.03389 \text{ratfin} + 1.61929 \text{auton} + 2.23185 \text{fixe} + 0.56277 \text{temp}$$

i: edat=25, ratfin=40, temp=1

$$\log \frac{p_i}{1-p_i} = -0.51578 + 0.03394 \times 25 - 0.03389 \times 40 + 0.56277 = -0.46011 \quad p_i = 0.387$$

i': edat=26, ratfin=40, temp=1

$$\log \frac{p_{i'}}{1-p_{i'}} = -0.51578 + 0.03394 \times 26 - 0.03389 \times 40 + 0.56277 = -0.42617 \quad p_{i'} = 0.395$$

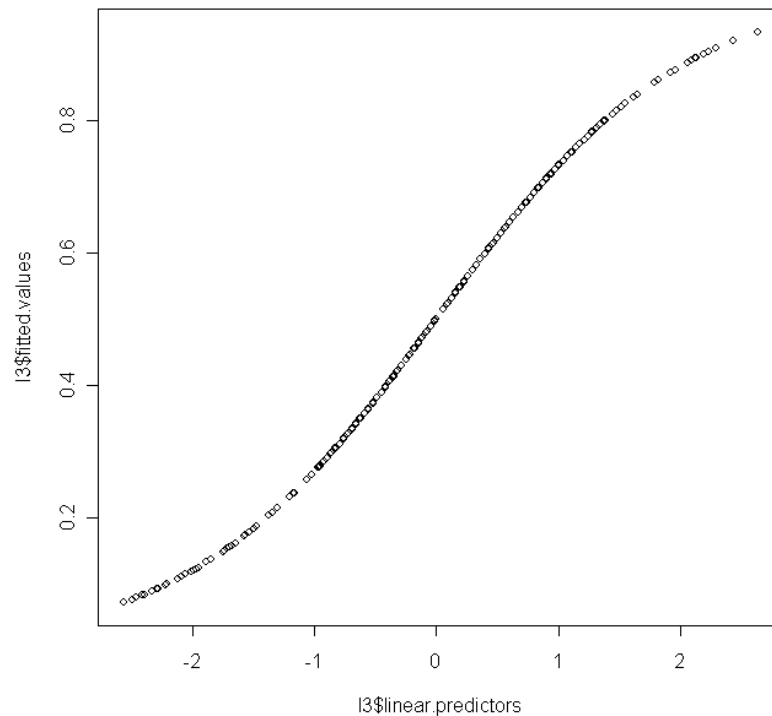
efecto de la edat: $\log \frac{p_{i'}}{1-p_{i'}} - \log \frac{p_i}{1-p_i} = 0.03394 \quad \frac{\frac{p_{i'}}{1-p_{i'}}}{\frac{p_i}{1-p_i}} = e^{0.03394} = 1.0345$

Interpret the coefficients

```
> exp(l3$coefficients)
(Intercept)      edat      ratfin  tiptrebauton  tiptrebfixe  tiptrebtemp
  0.5970355    1.0345176    0.9666757    5.0495067    9.3171141    1.7555279
```

Plot of the linear predictor and the estimated probabilities

```
> plot(l3$linear.predictors, l3$fitted.values)
```



Importance of the variables

Descomposition of the Deviance

```
> anova(l3)
```

```
Analysis of Deviance Table
```

```
Model: binomial, link: logit
```

```
Response: dict
```

```
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			406	563.92
edat	1	9.38	405	554.54
ratfin	1	33.59	404	520.95
tiptreb	3	31.60	401	489.35

$$Deviance_1 - Deviance_2 \sim \chi_{\nu_1 - \nu_2}$$

$$E[\chi_\nu^2] = \nu$$

Selecting the model

Estimate of the Generalization Error in a test sample:

Error rate in *learn*

```
> l3pred=NULL
> l3pred[l3$fitted.values<0.5]=0
> l3pred[l3$fitted.values>=0.5]=1
> table(dict[learn], l3pred)
  l3pred
    0    1
0 118  80
1  67 142
```

$P_{acierto}=63.9\%$

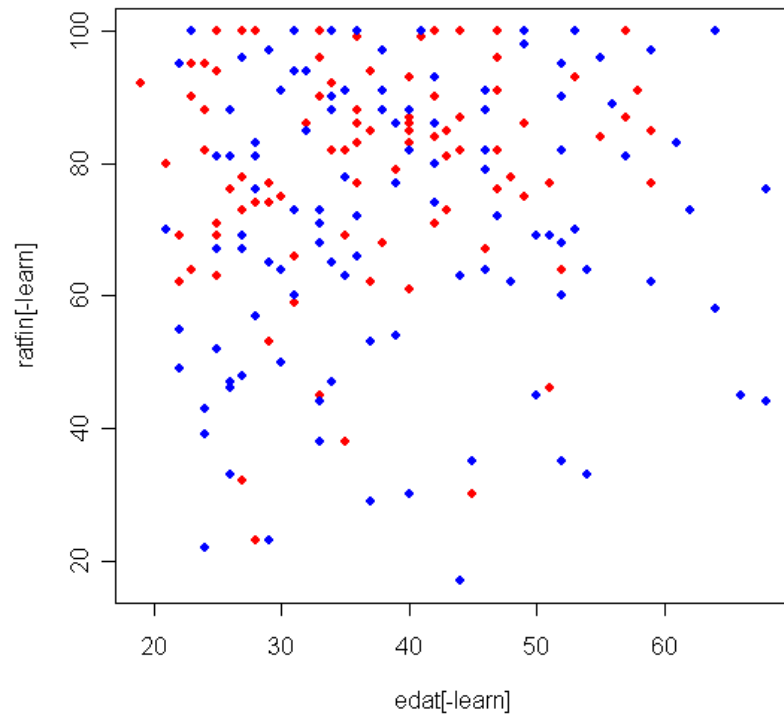
GE in *test*

```
> l3t = predict(l3, dd[-learn,])
> pt = 1/(1+exp(-l3t))
> l3predt = NULL
> l3predt[pt<0.5]=0
> l3predt[pt>=0.5]=1
> table(dict[-learn], l3predt)
  l3predt
    0    1
0  65  40
1  36  59
```

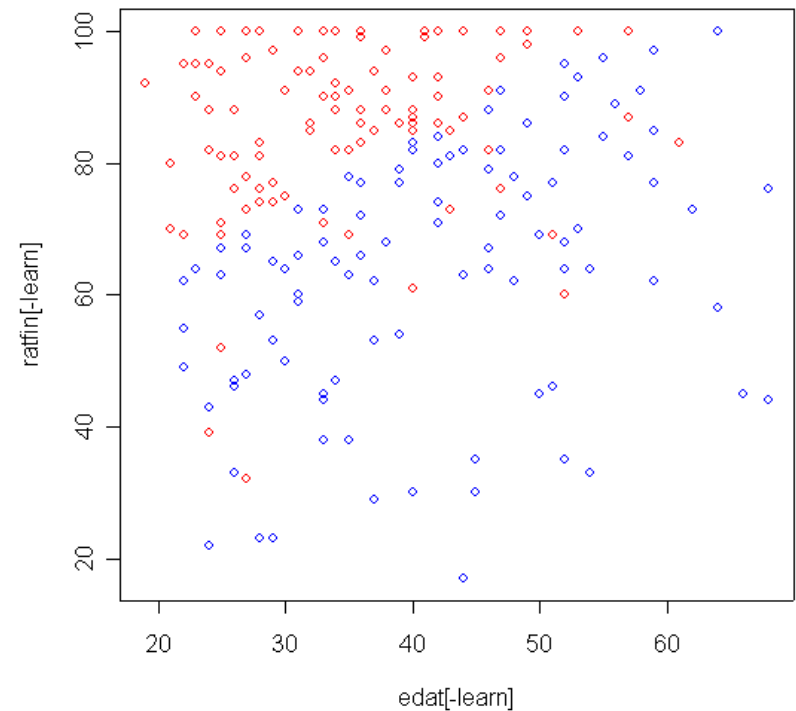
$P_{acierto}=62.0\%$

Graphical comparison of the real response respect to the predicted in the test sample

Actual response (*test*)

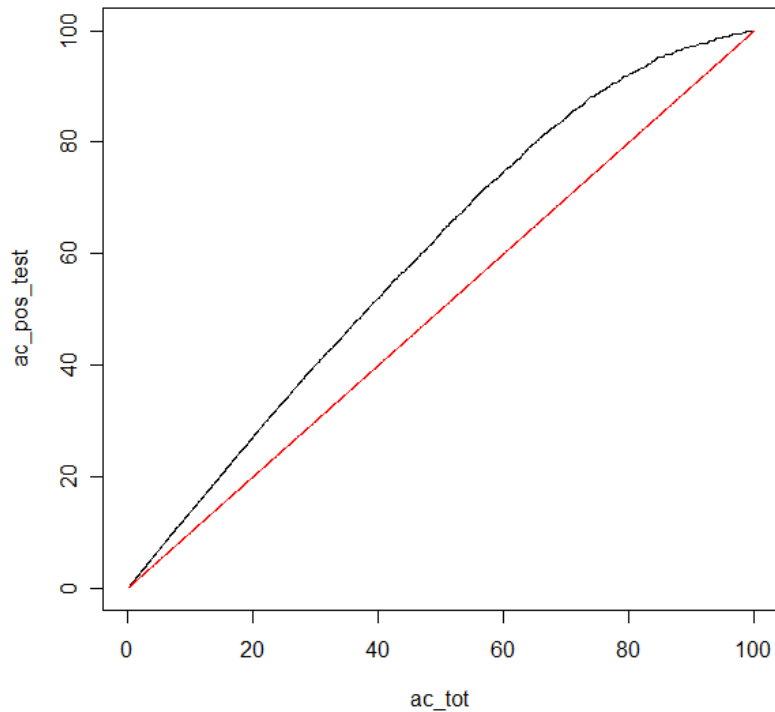


Prediction (*test*)



ROC and Concentration curves

Concentration Curve



ROC Curve

