Predictive Methods
Logistic Regression

K. Gibert [1]

[1]*Department of Statistics and Operation Research*

Knowledge Engineering and Machine Learning group
Universitat Politècnica de Catalunya, Barcelona

Logistic regression

Assessing the effect of continuous variables on a dichotomous outcome

Response variable : Binary/dichotomous

(or a proportion, ordinal variable, nominal variable)

Examples:

Buy a product, Pass a course, Obtain acredit, Level the preference for a service
Having Alzheimer's disease, Responding to a chemiotherapy, Smoking in high school, Evacuate before a hurricane

Other than: tumor size, daily packs of cigarretes, Final course score

Score and Mortality in Sepsis

30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

Formalization:

Target population: septic patients

Response variable: mortality after 30 days

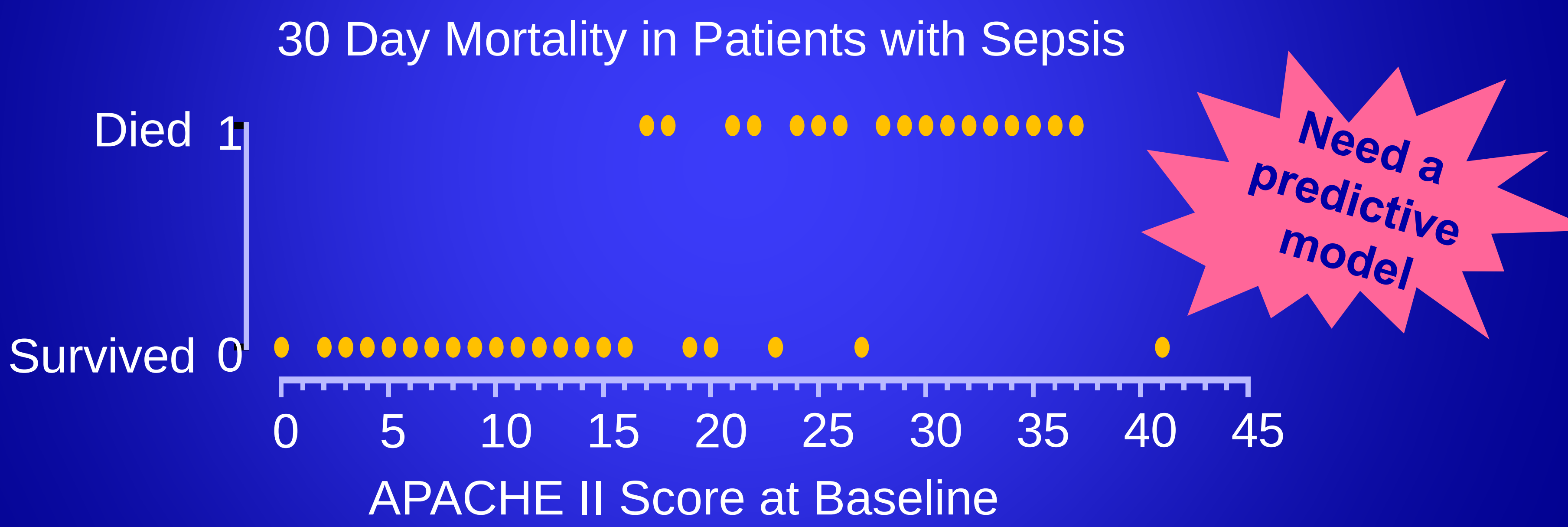
1 means patient died after 30 days

0 means patient survived after 30 days

Explanatory variable: Score obtained in APACHE II Scale at day 1

Score and Mortality in Sepsis

30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

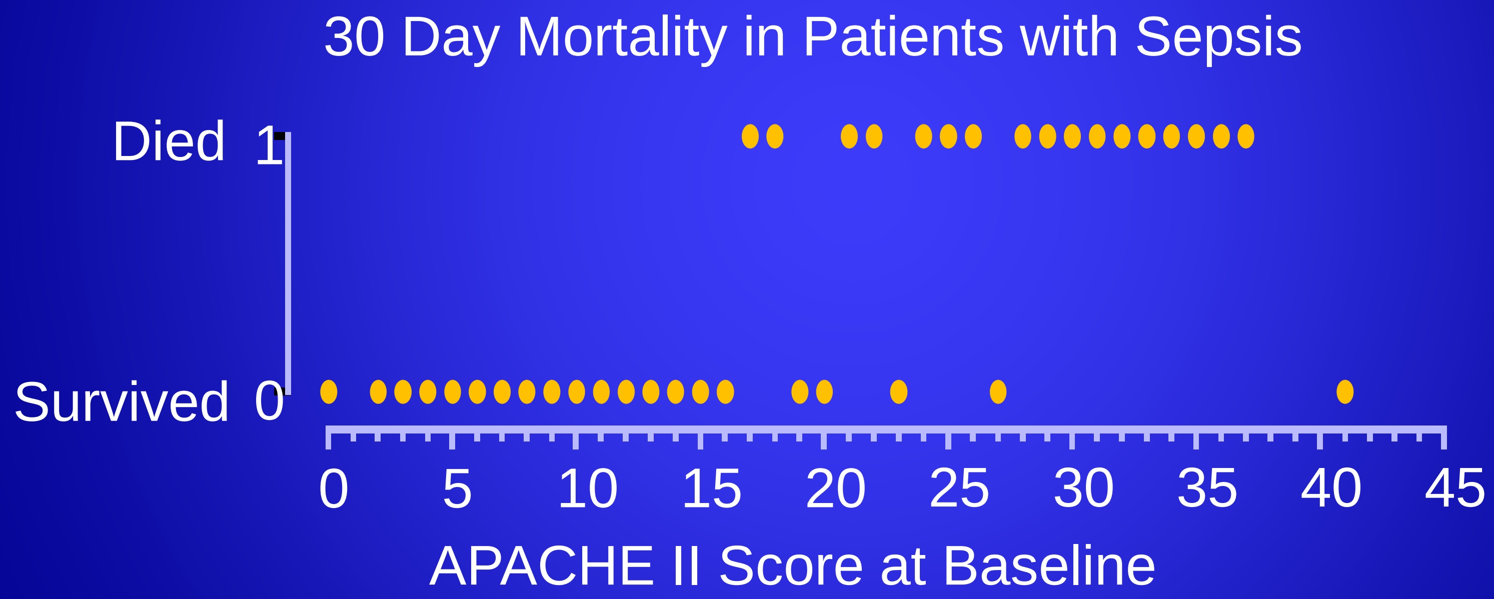


Compare mean score of dead and non-dead? HOW? T-test? ANOVA?

DO NOT ALLOW PREDICTIONS!!!!

Score and Mortality in Sepsis

30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.



Compare mean score of dead and non-dead? HOW? Linear Regression?

Logistic regression

Response variable : Binary

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

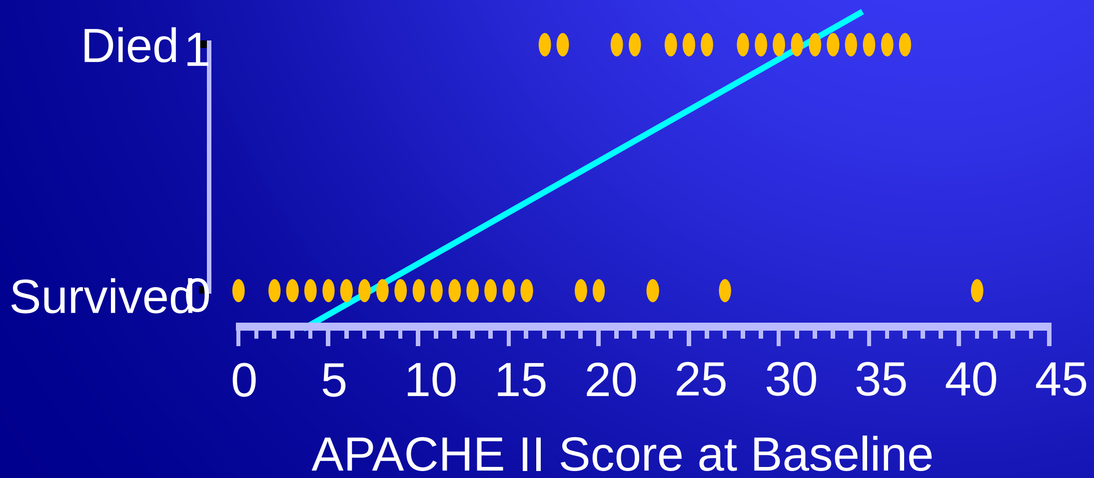
Does not work!!!

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

Linear Model fit

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

30 Day Mortality in Patients with Sepsis



$-\infty < \hat{y} < \infty$

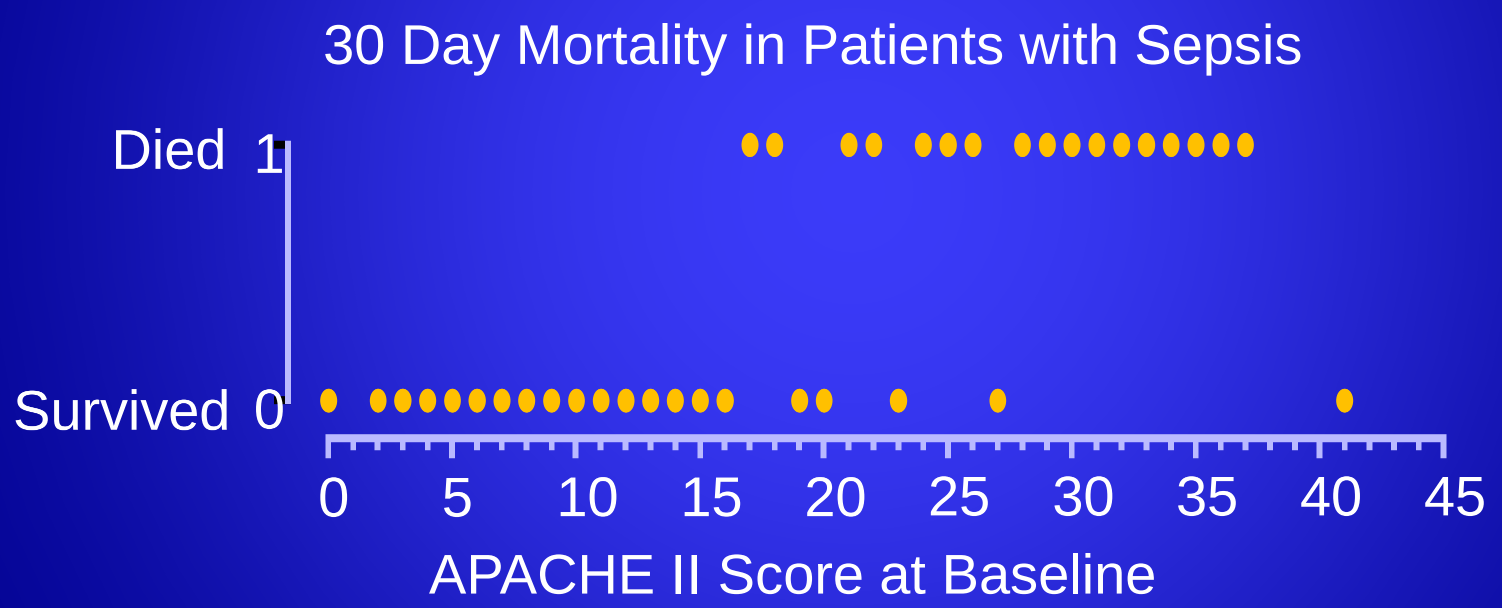
(continuous prediction $\hat{y}$ senseless
$\hat{y} \notin [0,1]$ senseless)
Error non normal (Bernoulli)

Non linearity

Violation of linear model hypothesis

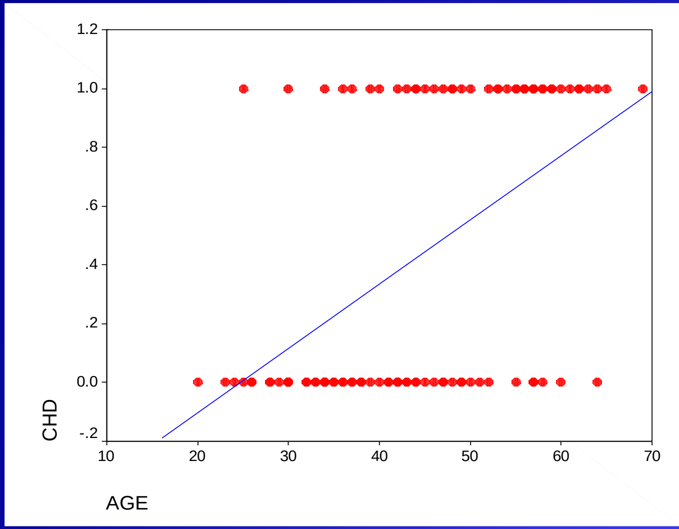
Score and Mortality in Sepsis

30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

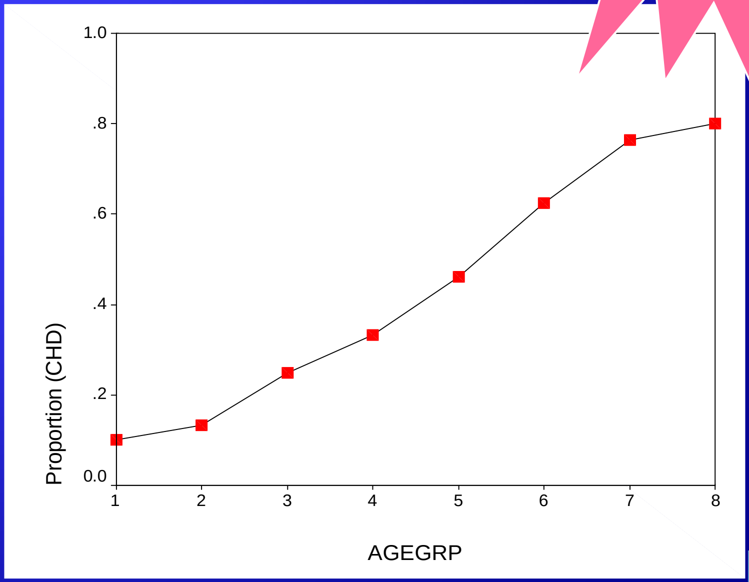


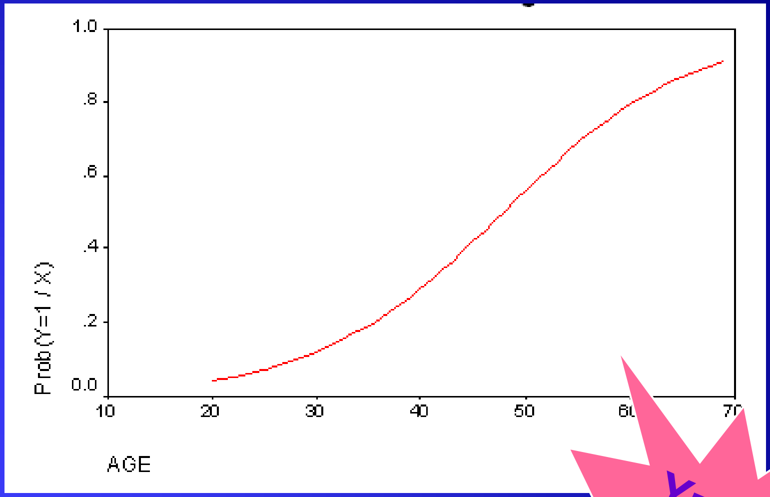
Compare mean score of dead and non-dead? HOW? Linear Regression?

Reformulate



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

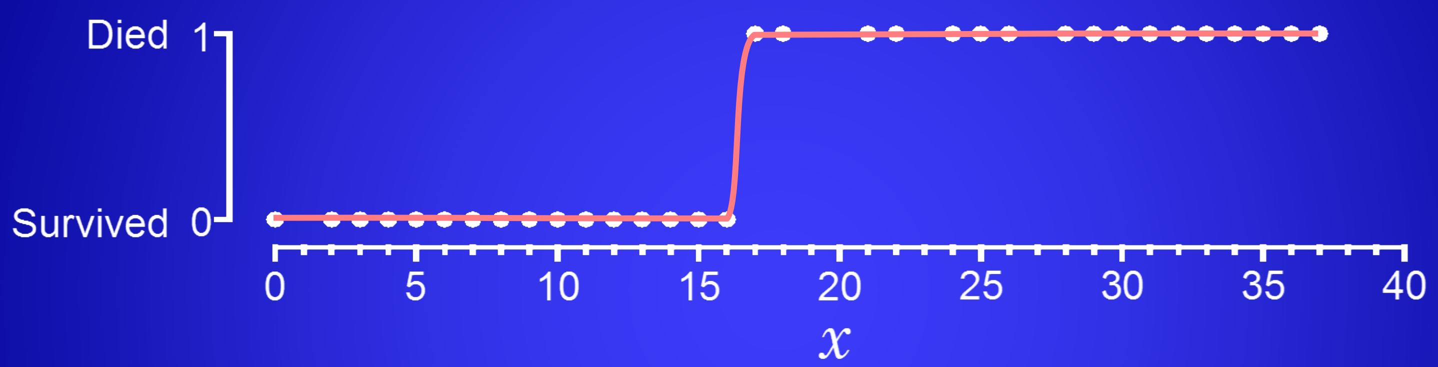




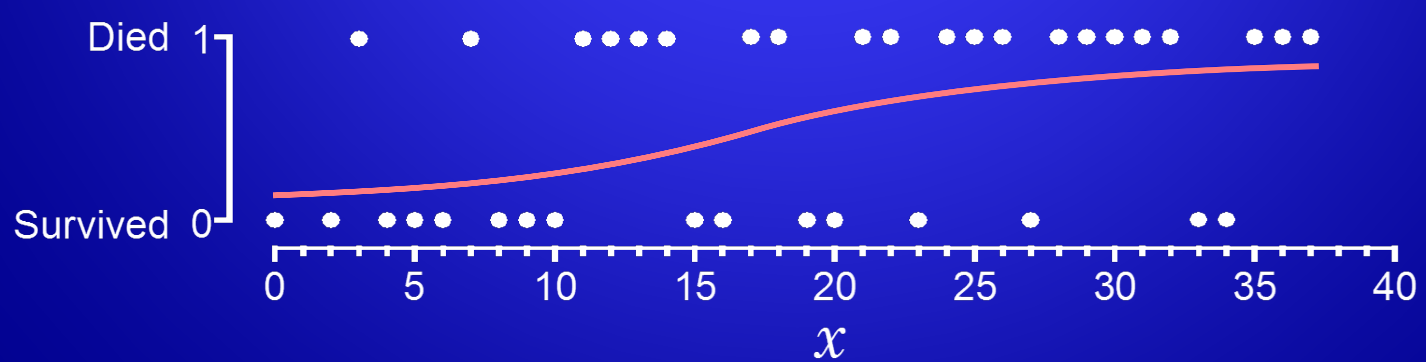
Y= freq of Dead

Find best curve to fit the data.

Sharp cut off point between live or die



Lengthy transition from survival to death



How can we model binary responses?

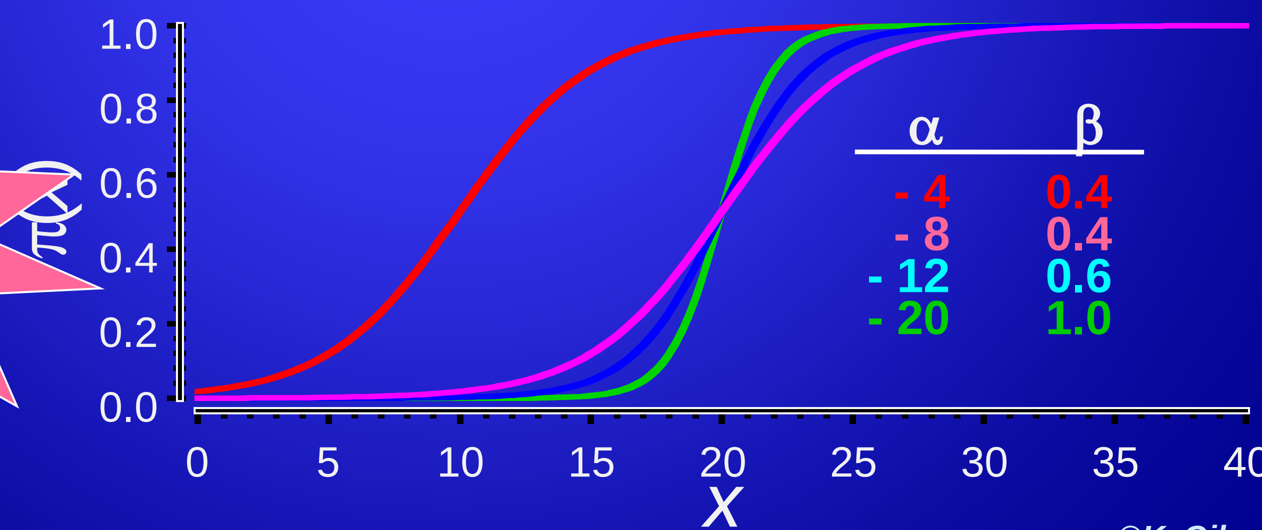
Response is binary 0/1

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

Modelling: Family of sigmoidal curves

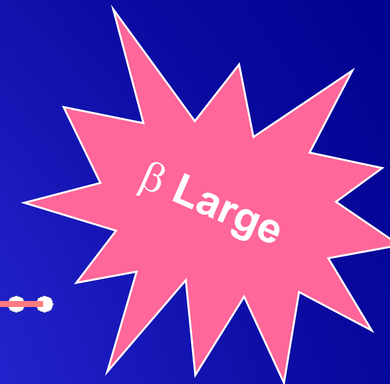
$$\pi(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

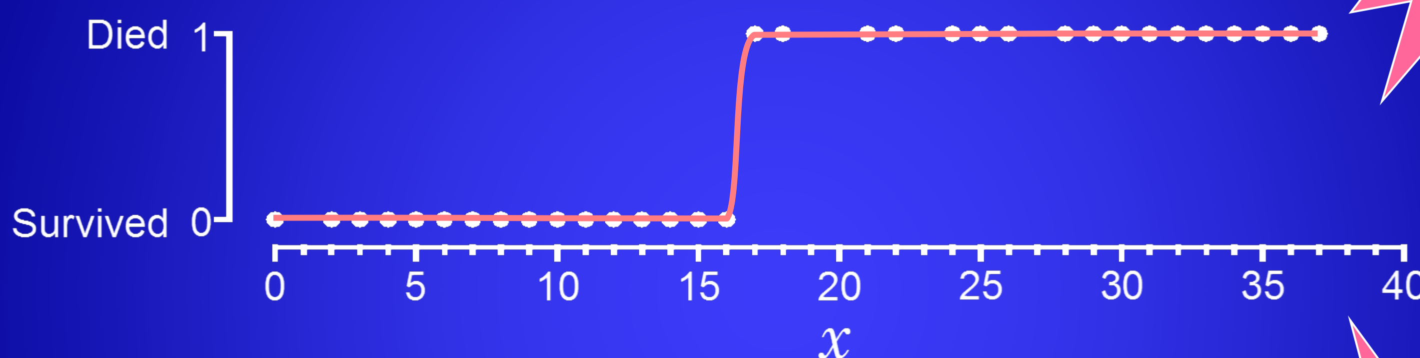
β controls how fast $\pi(x)$ rises from 0 to 1.



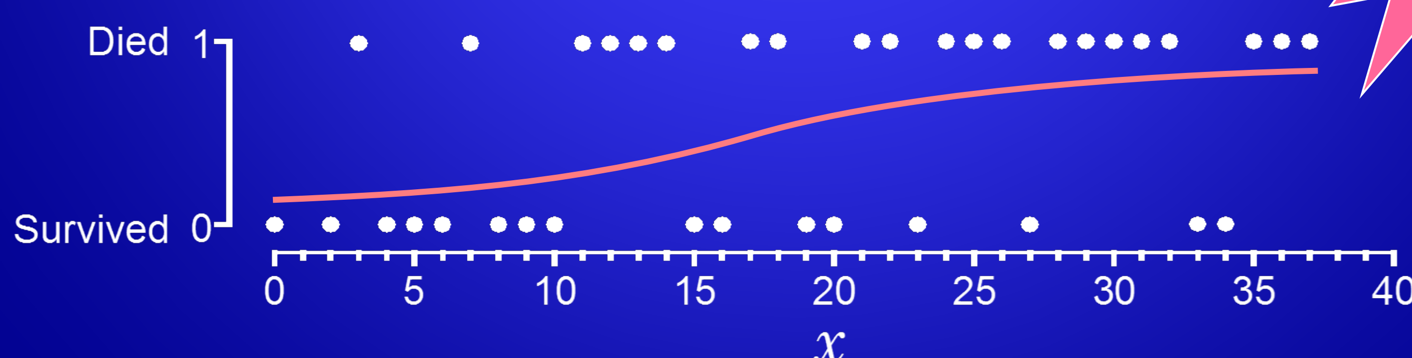
Find best curve to fit the data.

Sharp cut off point between live or die



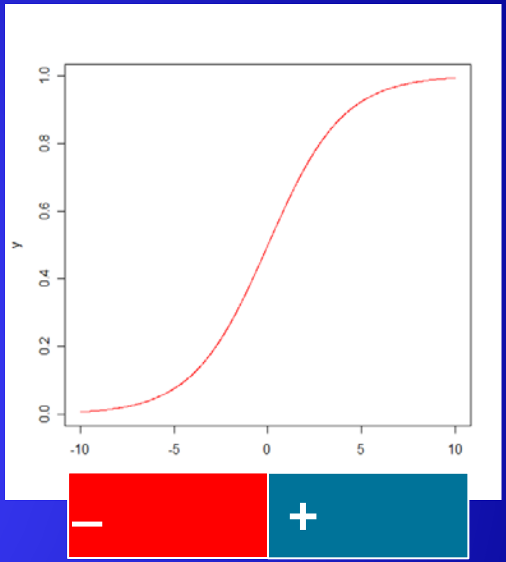


Lengthy transition from survival to death



Interpretation of the logistic function

Propensity of the + event



Decision rule:

Determine a threshold ℓ (i.e 0.5)

If η_i > ℓ then consider i propense to +,
otherwise assign –

Threshold low: conservative model

In economy the propensity to buy/invest is associated to a user choice
In health propensity is associated to desease
In survival analysis is associated to survival

Transformation

Probability of dead (Y=1) given x

$$\pi(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Probability of non dead (Y=0) given x

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

Odds of dead: prob of dead vs non dead

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

Linear transformation logit ($\pi(y|x)$)= *ln(odds)*

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \alpha+\beta x.$$

Reduction to Multiple Linear Regression

Multiple logistic regression

Several independent variables

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K$$

✓β_0 = log odds ratio for X=0 *(baseline odds ratio, moves curve left/right)*

✓β_κ = log odds ratio associated with X_k (Steepness of curve)

increase of log-odds when X_k increases one unit and

$X \neq X_k$ keep constant

(marginal unitary effect of Xk on log odds)

✓e^{β_κ} = unitary marginal odds ratio

Regressors numerical or dummy

Interpreting the coefficients of a logistic regression

Lets take one predictor *x=0,1*

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

Likewise, …

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

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 being in class + against -, when passing from *x=0* to *x=1*

Interpreting the coefficients of a logistic regression

Lets take one predictor *x=0,1* (i.e. *0*=non-married, *1*=married)

CREDSCO application:

Response variable: Dictamen
Regressor: Civil Status (dummies)

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

The odds for a married person, express how more likely a married person is to have a positive dictamen rather than negative

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

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*

Multiple logistic regression

Several independent variables

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K$$

$$\left[\frac{P(y|x)}{1-P(y|x)}\right] = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots}$$

Change in probability no constant with cnt changes in X

✓Assumptions:

Non assumed normality, linearity, homokedasticity

Discriminant analysis more powerful when assumptions ...ld

Sensitive to outliers

✓Good practice guidelines:

10 cases minimum per regressor (Hosmer and Lemeshow)

50 cases minimum per regressor for stepwise

Group avoiding multicolinearity is better (separability)

Multiple logistic regression

The logit function: $\pi \in [0,1]$ $\text{logit}(\pi) = \log(\pi / (1-\pi))$

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

$$E(y_i) = p_i = \pi(x_i)$$

x_i the APACHE II score of the i^{th} patient
π, probability of dying with a certain APACHE II score

Logistic regression equation can be rewritten as

$$logit(E(y_i)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K$$

Link function

Fitting the model

Estimate the coefficients of the linear equation by ordinary methods:

Maximum likelihood estimation

- Model selection:
  – Complete model (no-viable with big K)
  – Hierarchical method

    (enter control variables before predictors affected by them)
  – Stepwise method

    (enter first more signifficant variables)

  – Contribution: chi2

Maximum Likelihood Estimation (MLE) remainder

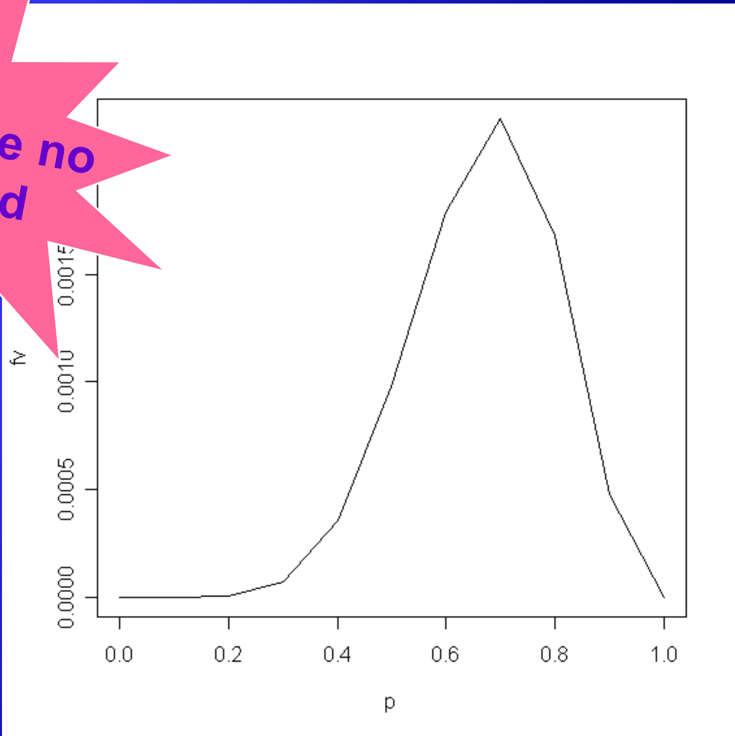
Choose as estimates of the parameters those who maximize the probability of the observed data

$$Max\ L(\theta) = \Pr(x_1,\ldots x_n/\theta) = \Pr(x_1/\theta)\times\cdots\times\Pr(x_n/\theta)$$

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

Convergence no guaranteed

```
> 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

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

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

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

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

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_i^n (y_i \mathbf{x}_i - \frac{e^{\beta' x_i}}{1+e^{\beta' x_i}} \mathbf{x}_i)$$

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

$$\frac{\partial^2 l(\beta)}{\partial \beta \, \partial \beta'} = \sum_i^n - \frac{e^{\beta' x_i}}{(1+e^{\beta' x_i})^2} \mathbf{x}_i \mathbf{x}_i'$$

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

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

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

Iterate till convergence

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

Model inference

- The Wald statistic for the β_k coefficient is:

$$\left(\frac{\hat{\beta}_k}{S_{\hat{\beta}_k}}\right)^2 \sim \chi^2_1$$

if pvalue<0.05 keep the term

- The "Partial R" is

$$R = \{[(Wald-2)/(-2LL(\alpha)]\}^{1/2}$$

Warning: Multicolinearity

- Determine the significant regressors

Model assessment/validation

R² non reliable

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, is similar to the sum of residual squares in line

Smaller dev
Better model

$$D = -2\log\frac{L(\beta_{cur})}{L(\beta_{sat})} = -2\sum_{i=1}^{n}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) \sim \chi^2_{\nu=n-p-1}$$

Null deviance: Deviance of the null model (just with constant term)
Residual deviance: Deviance of the proposed model
AIC: Deviance with complexity penalization (2p)

Standard errors >2
Points numerical problems

Confusion matrix also an alternative

25% improvement over Accuracy of random assignment

By chance accuracy

Pseudo-R^2

- One psuedo-R^2 statistic is the McFadden's-R^2 statistic:

$$\text{McFadden's-}R^2 = 1 - [LL(\alpha,\beta)/LL(\alpha)]$$

$$\{= 1 - [-2LL(\alpha, \beta)/-2LL(\alpha)] \text{ (from } \textit{SPSS} \text{ printout)}\}$$

- where the R^2 is a scalar measure which varies between 0 and (somewhat close to) 1 much like the R^2 in a LP model.

Structural Breaks

- You may have structural breaks in your data. Pooling the data imposes the restriction that an independent variable has the same effect on the dependent variable for different groups of data when the opposite may be true.
- You can conduct a likelihood ratio test:

  LR[i+1] = -2LL(pooled model)

  [-2LL(sample 1) + -2LL(sample 2)]

  where samples 1 and 2 are pooled, and i is the number of independent variables.

```
> 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
```

The obtained model:

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

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

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

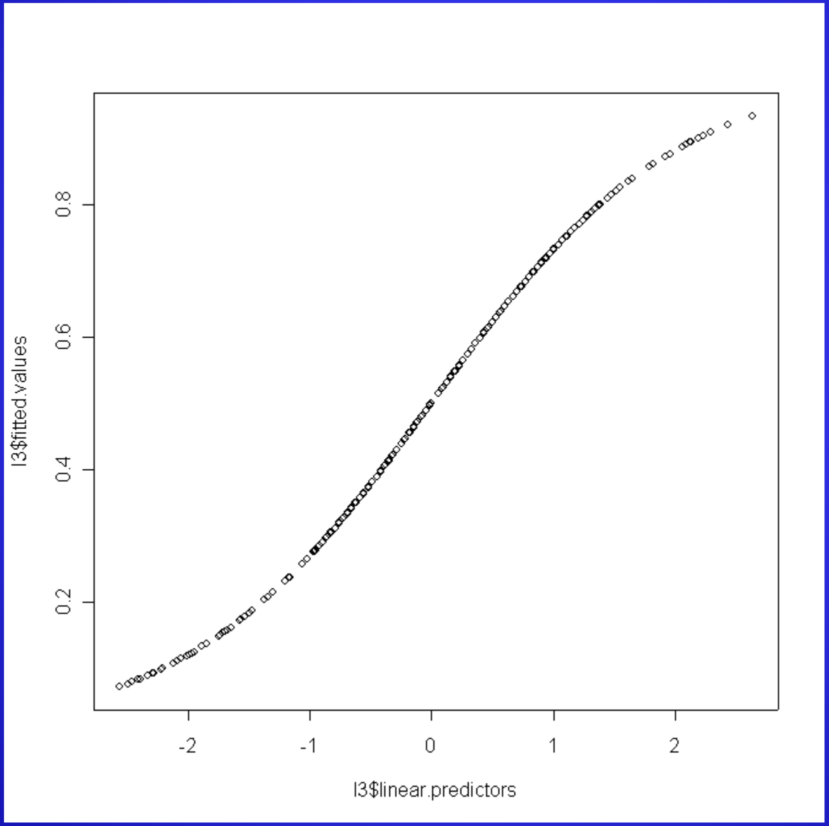
efecto de la edat: $\log \frac{p_{i'}}{1-p_{i'}} - \log \frac{p_i}{1-p_i} = 0.03394 \qquad \frac{p_{i'} / 1-p_{i'}}{p_i / 1-p_i} = e^{0.03394} = 1.0345$

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
```

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

$$E\left[\chi^2_\nu\right] = \nu$$

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%

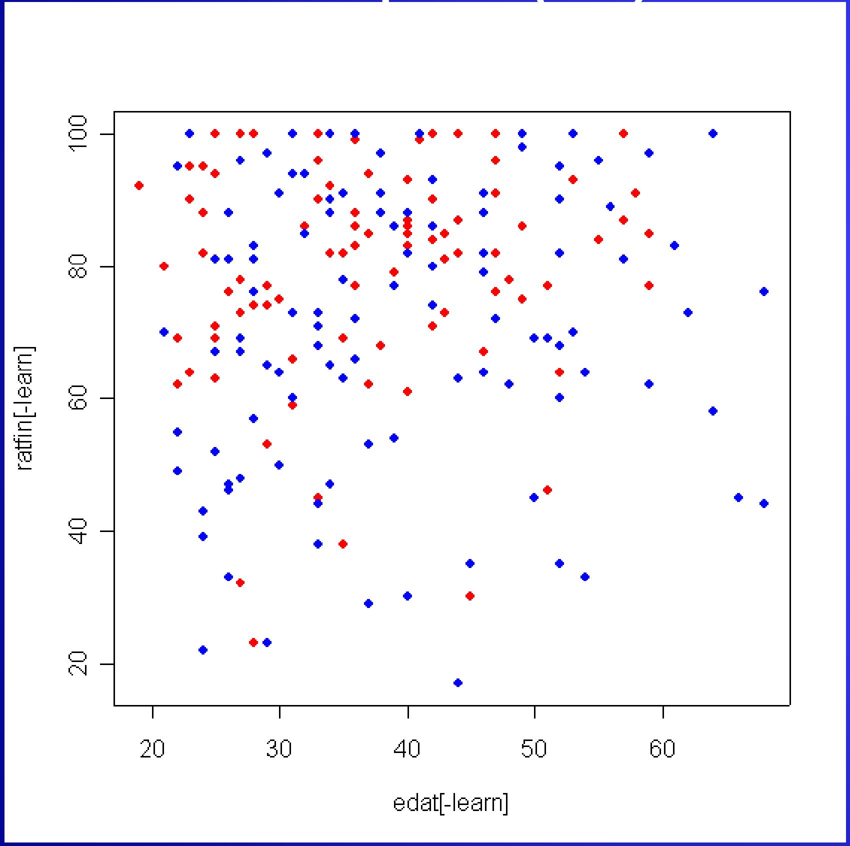
GE in *test*

```
> 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
```

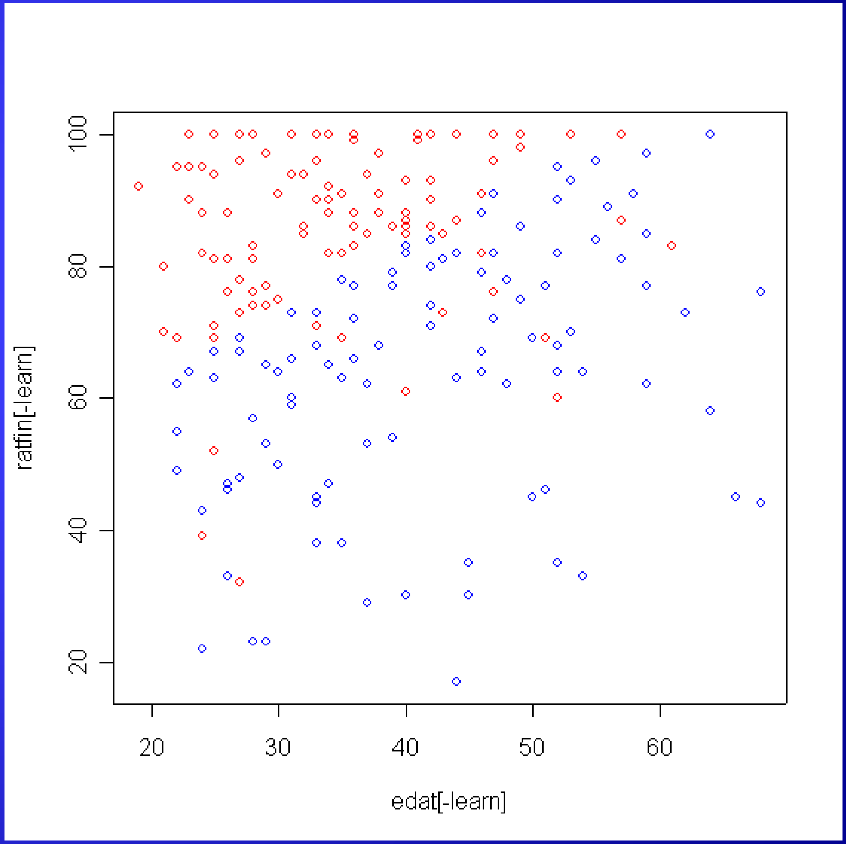
$P_{acierto}$=62.0%

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

Actual response *(test)*



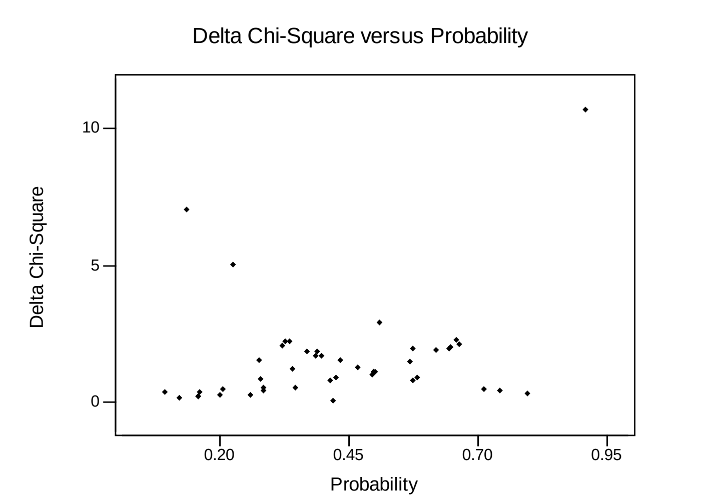
Prediction *(test)*

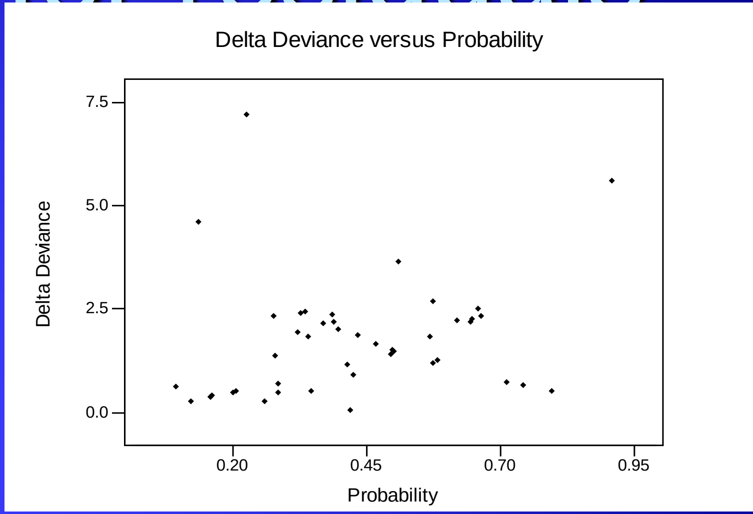


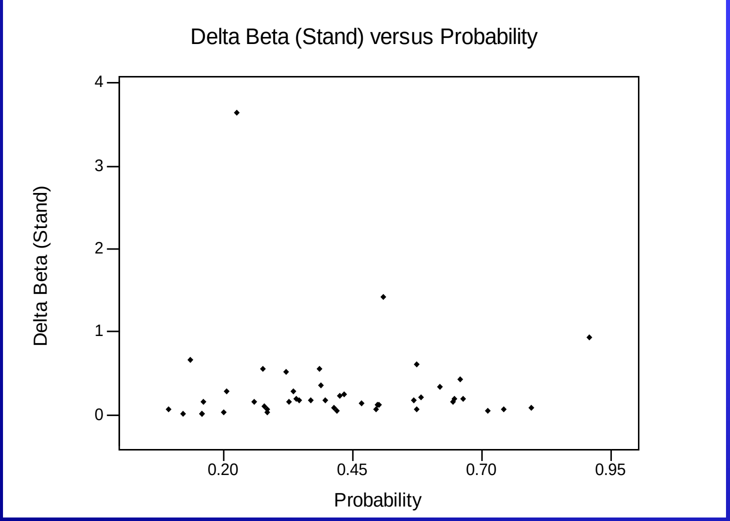
Regression Diagnostics

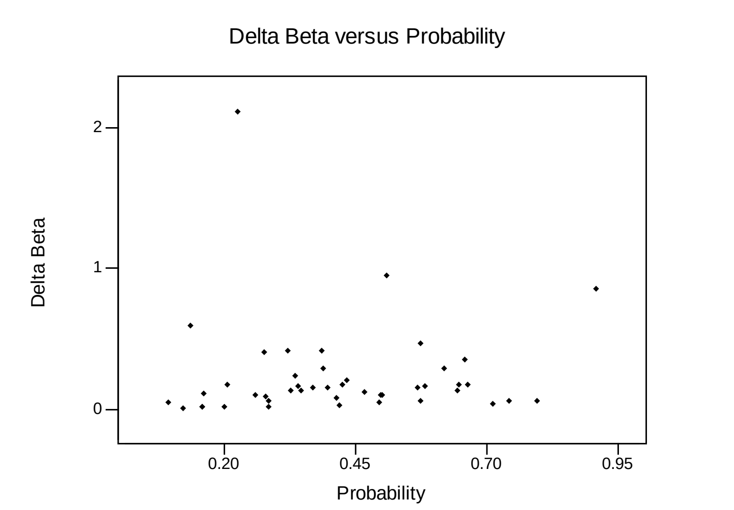
- In logistic regression Residual = 1− Estimated probability. Residuals for each subject are calculated standardised and plotted against probability. Eight diagnostic plots are available, four dealing with residuals and four with leverage.
- These plots are demonstrated in the slides that follow.
- ROC and concentration curves

Diagnostic plots for residuals



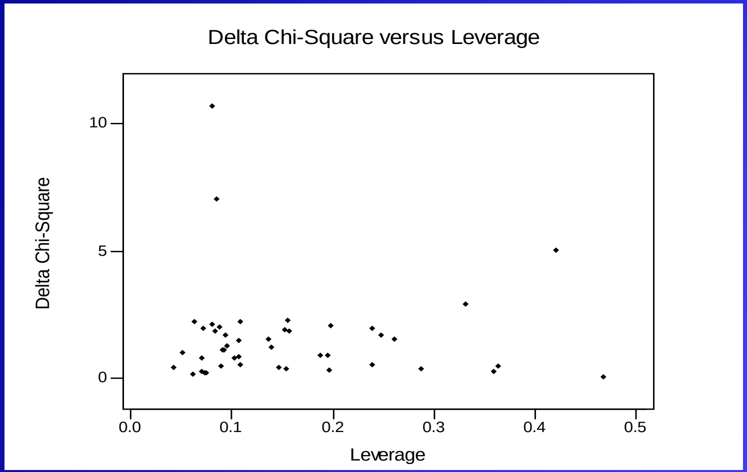


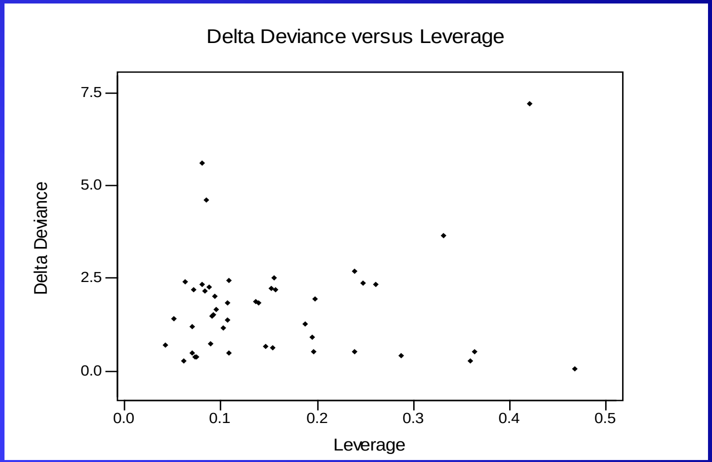


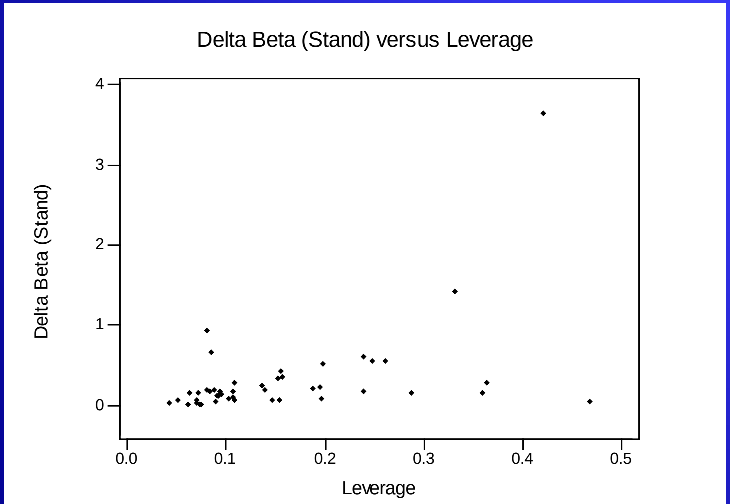


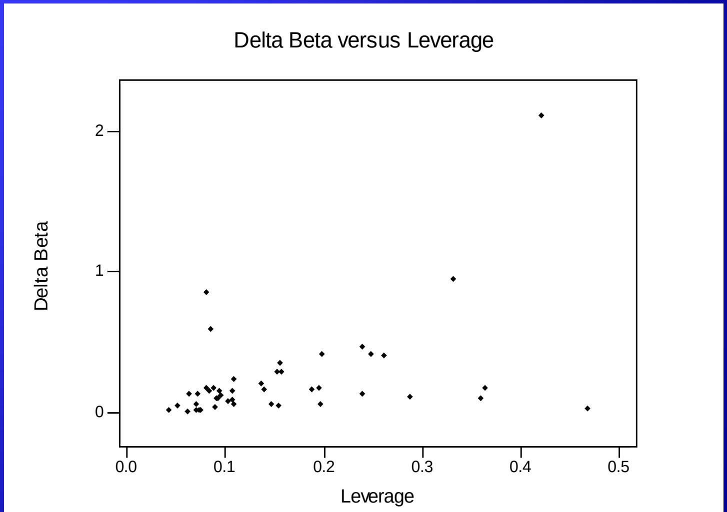
©K. Gibert

Diagnostic plots for leverage









Índex Gini de rendiment

- Àrea entre la curva ROC i la bisectriu de 45º

	Logistic Regression	RBF	CART Tree	K-NN MBR
Gini index	0,4375	0,4230	0,4445	0,5673

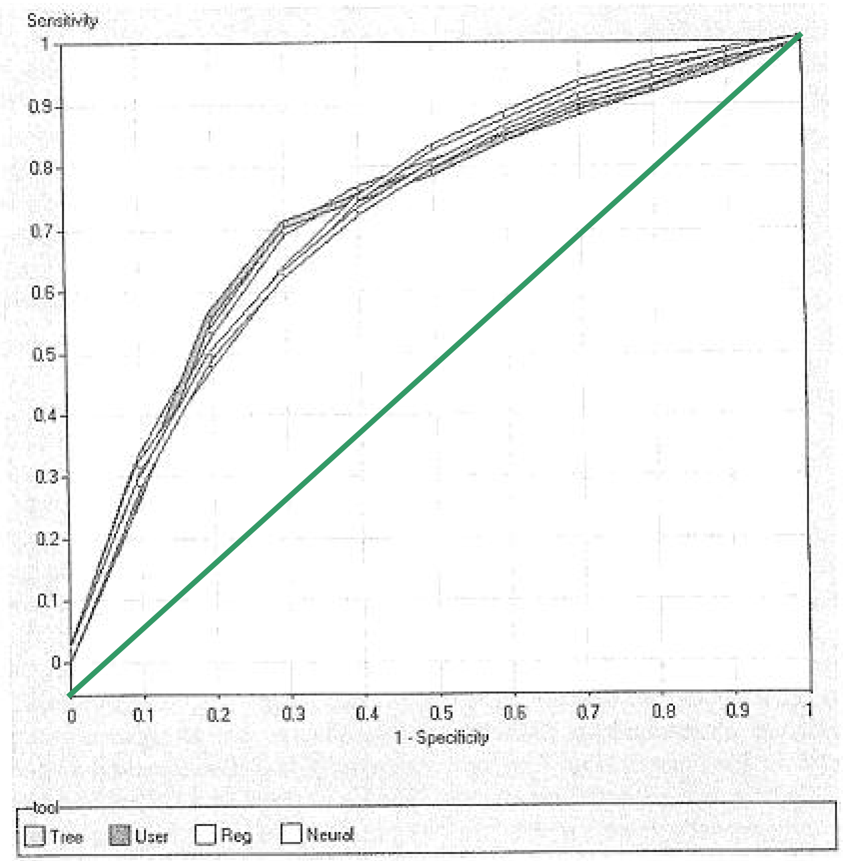


Figure 10.5 ROC curves for the considered models. The curve called user is the MBR model.