# Predictive Methods
# Logistic Regression

***K. Gibert***

*Department of Statistics and Operation Research*

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

*karina.gibert@upc.edu*
*www.eio.upc.edu/homepages/karina*

# 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

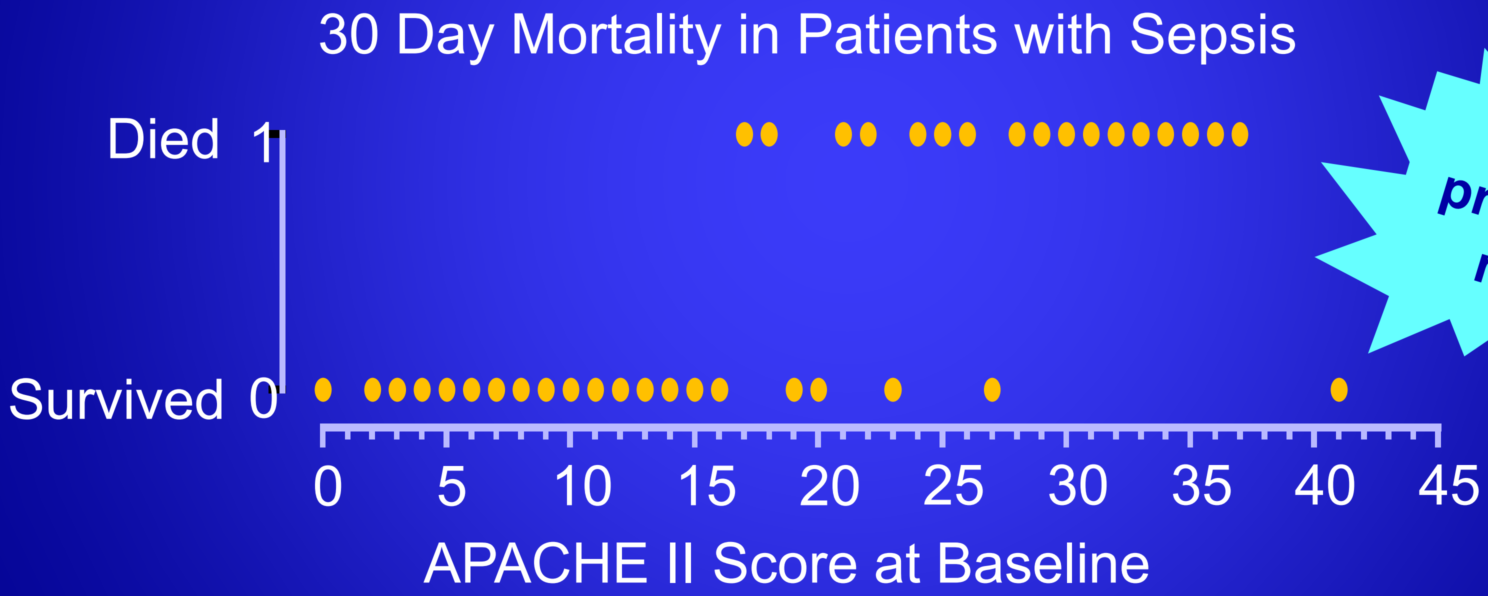
1 means patient died after 30 days

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.



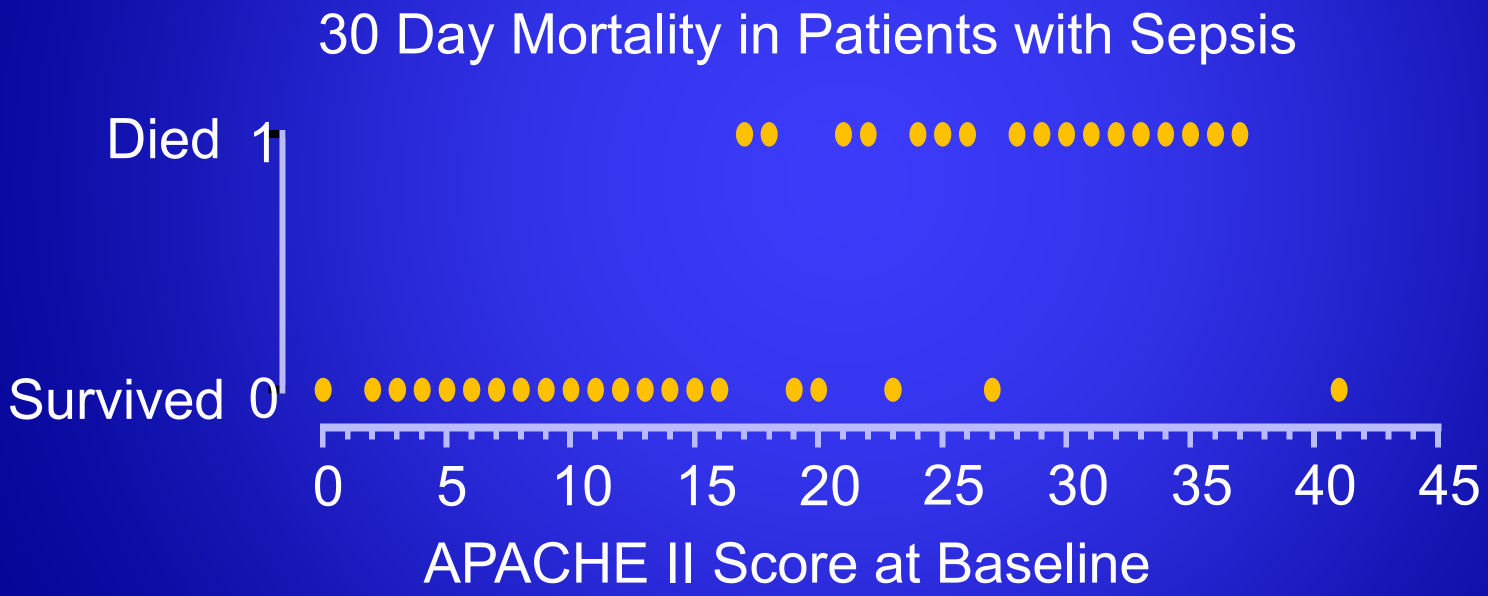
Need a predictive model

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**

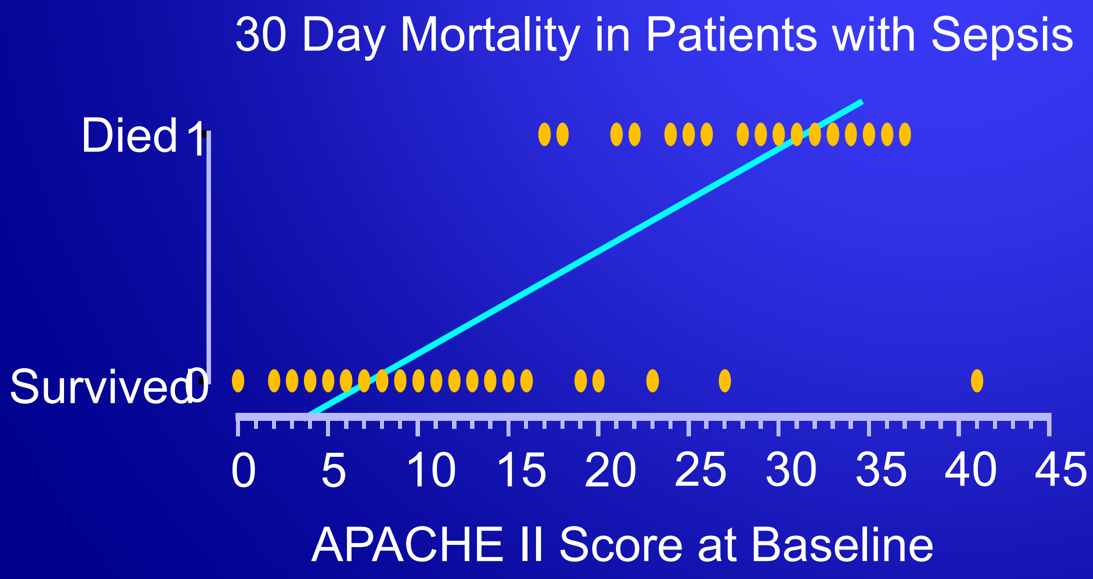
$$y_i = \begin{cases} 1 \; if \; + & \text{with } p_i \\ 0 \; if \; - & \text{with } (1-p_i) \end{cases}$$

**Does not work!!!**

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

**Linear Model fit**

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



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

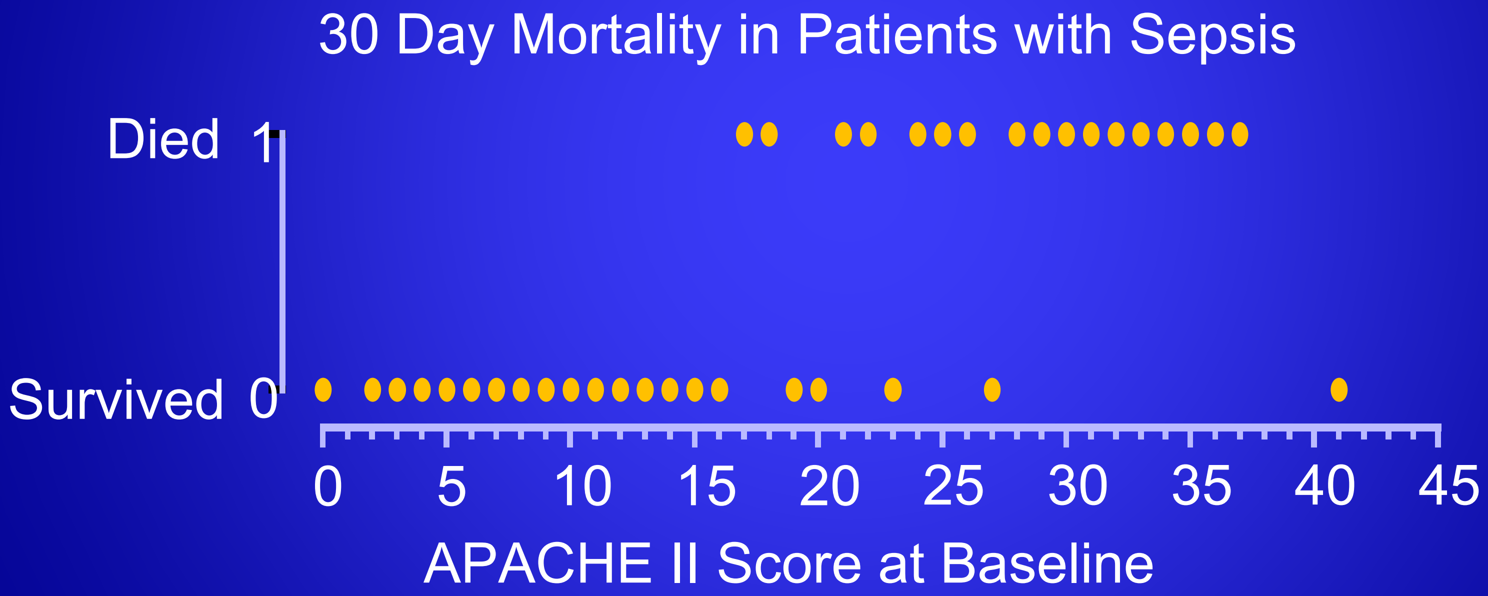
*(continuous prediction ŷ senseless*
*ŷ ∉ [0,1] senseless )*
*Error non normal (Bernoulli)*

*Non linearity*

Violation of linear model hypothesis

# Score and Mortality in Sepsis

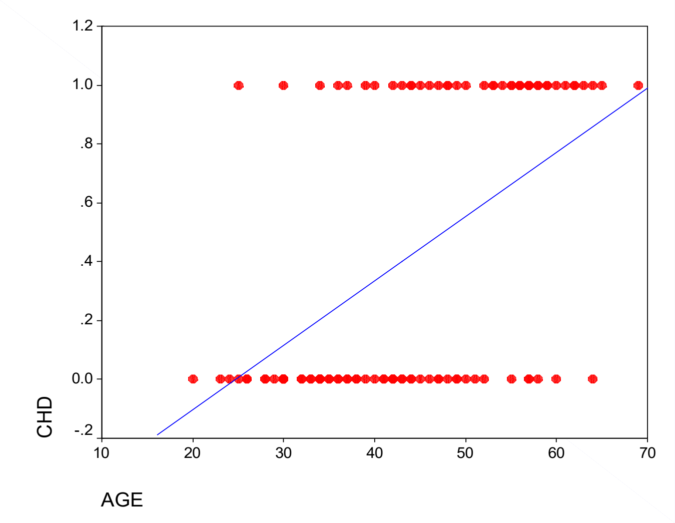
30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.



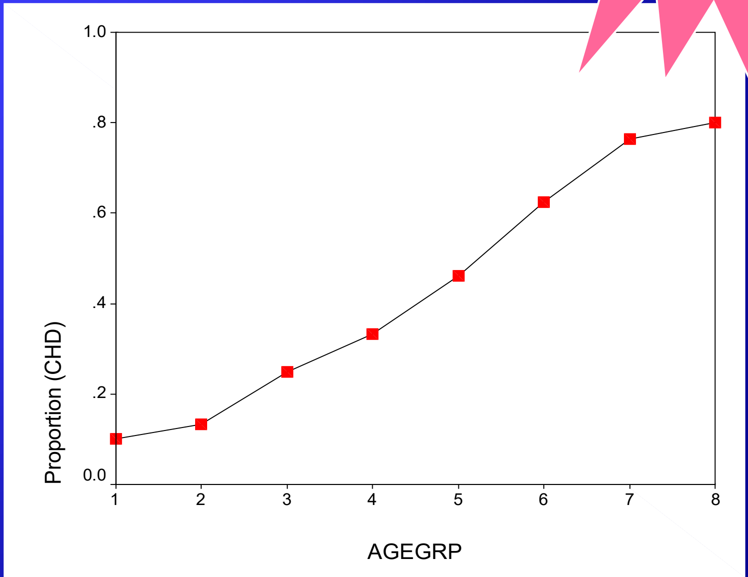
Compare mean score of dead and non-dead? HOW? Linear Regression?

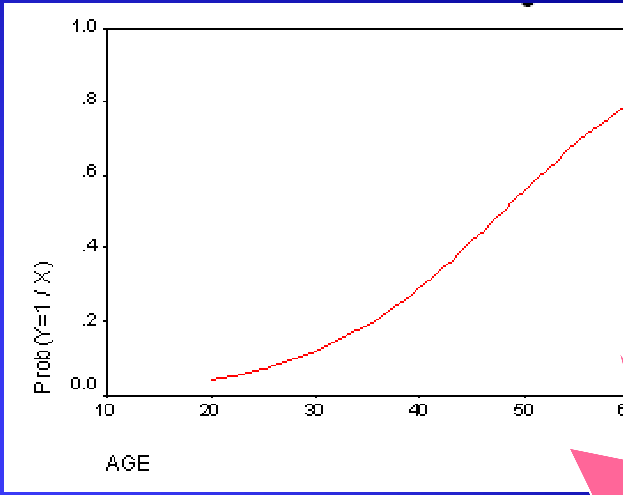
©K. Gibert

# 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 |

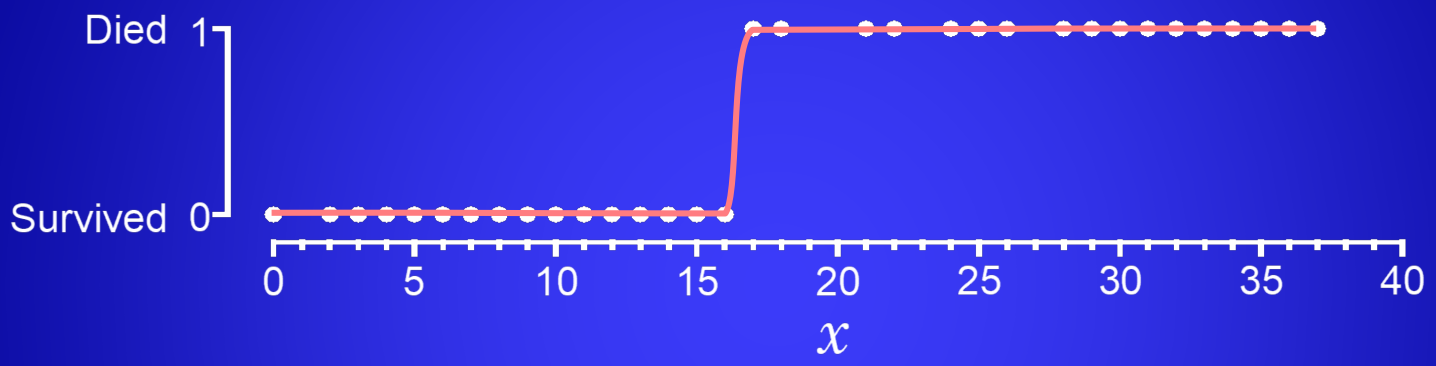




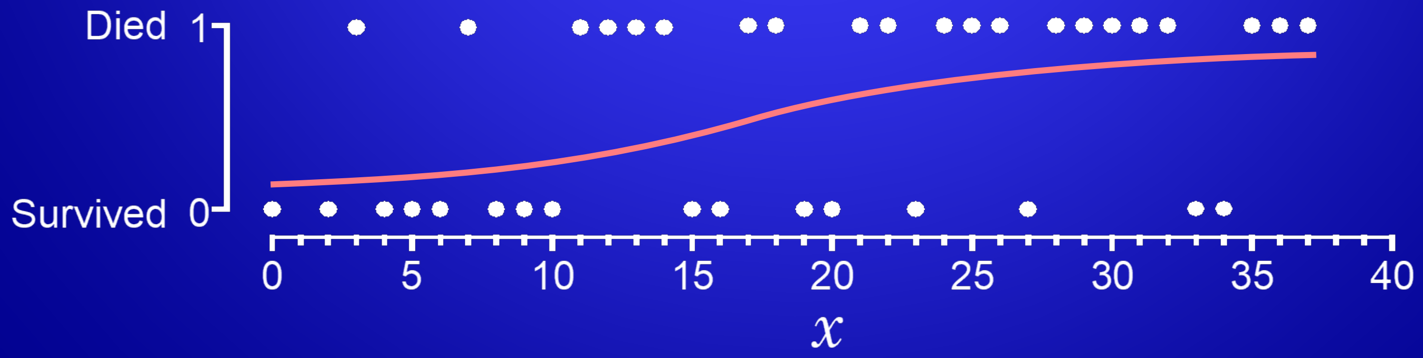
Y= freq of Dead

# Find best curve to fit the data.

Sharp cut off point between live or die



Lengthy transition from survival to death



©K. Gibert

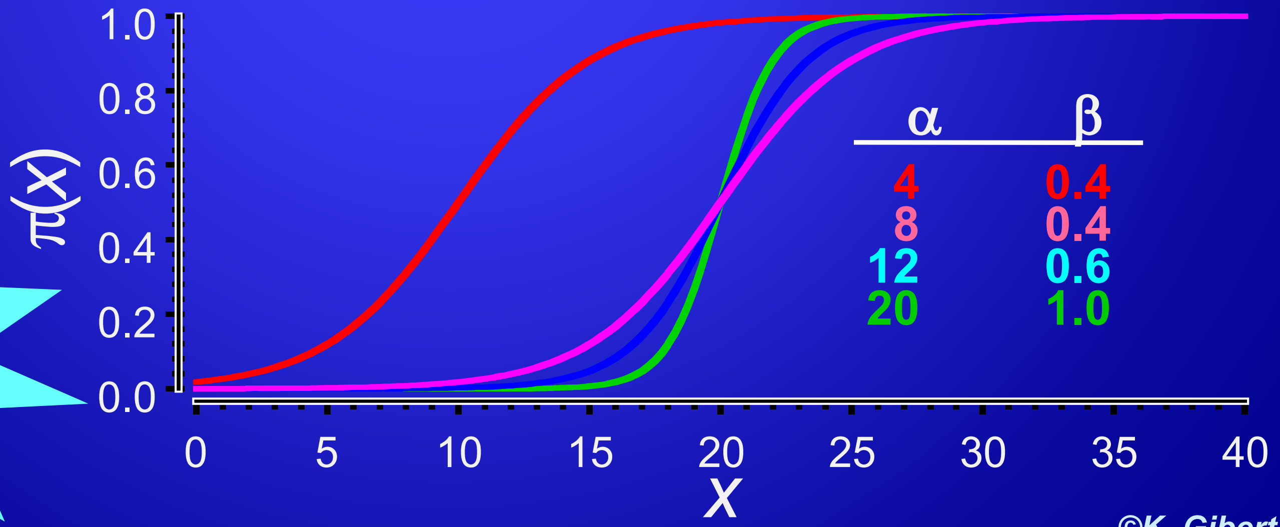
# How can we model binary responses?

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

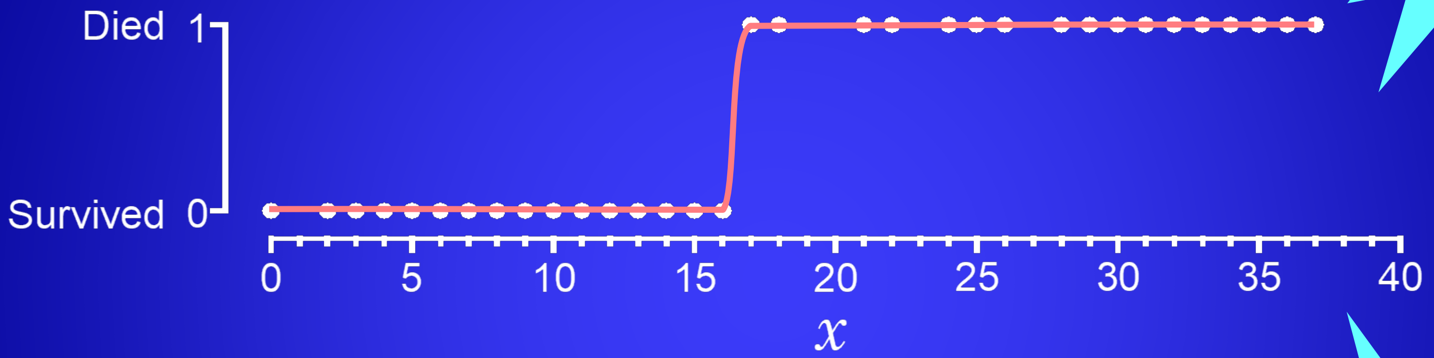
$$\pi(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$



β controls how fast π(x) rises from 0 to 1.

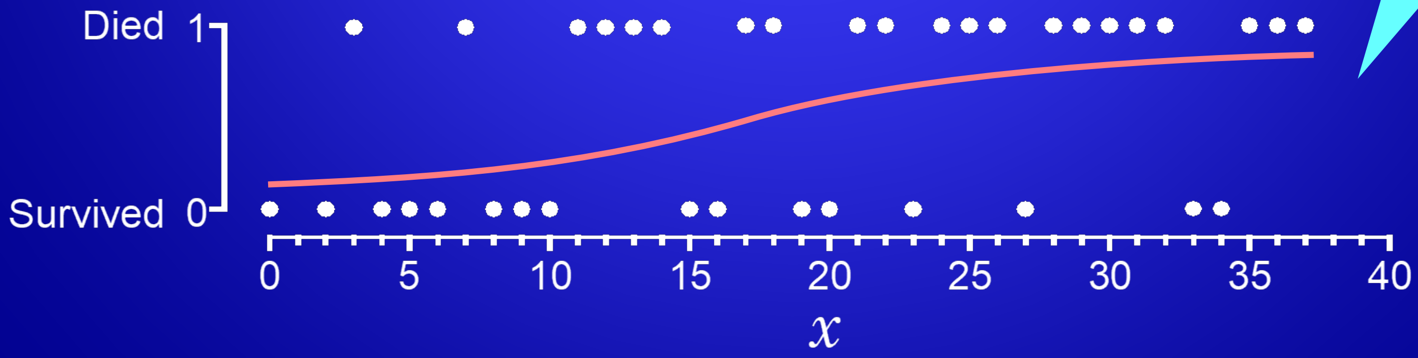
# Find best curve to fit the data.

Sharp cut off point between live or die



$\beta$ Large

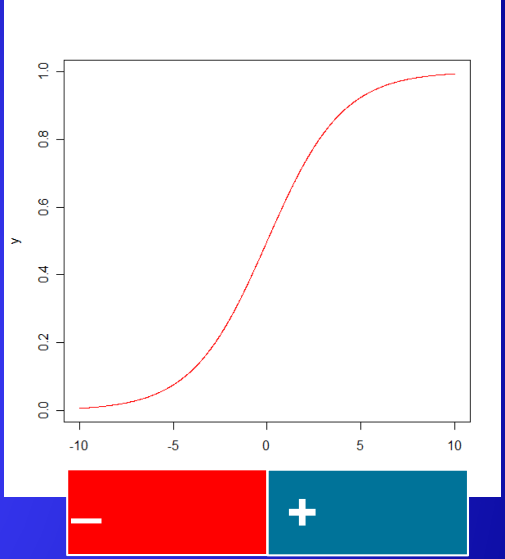
Lengthy transition from survival to death



$\beta$ Small

# Interpretation of the logistic function

Propensity of the + event



***Decision rule:***

***Determine a threshold ℓ (i.e 0.5)***

***If $\eta_i$ > 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

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

✓$\beta_0$ = log odds ratio for X=0 *(baseline odds ratio, moves curve left/right)*

✓$\beta_\kappa$ = 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)*

✓$e^{\beta_\kappa}$ = 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 *$\beta_1$* 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)

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

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

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

The ODDS RATIO

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

**The exponential of the *β₁* coefficient measures the change in the odds of + against -, when passing from *x=0* to *x=1***

# Multiple logistic regression

Several independent variables

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

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

Change in probability no constant with constant changes in X

✓Assumptions:

Non assumed normality, linearity, homokedasticity, independency

Discriminant analysis more powerful when assumptions hold

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)

©K. Gibert

# Multiple logistic regression

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

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

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

$x_i$ the APACHE II score of the $i^{\text{th}}$ patient
$\pi$, 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

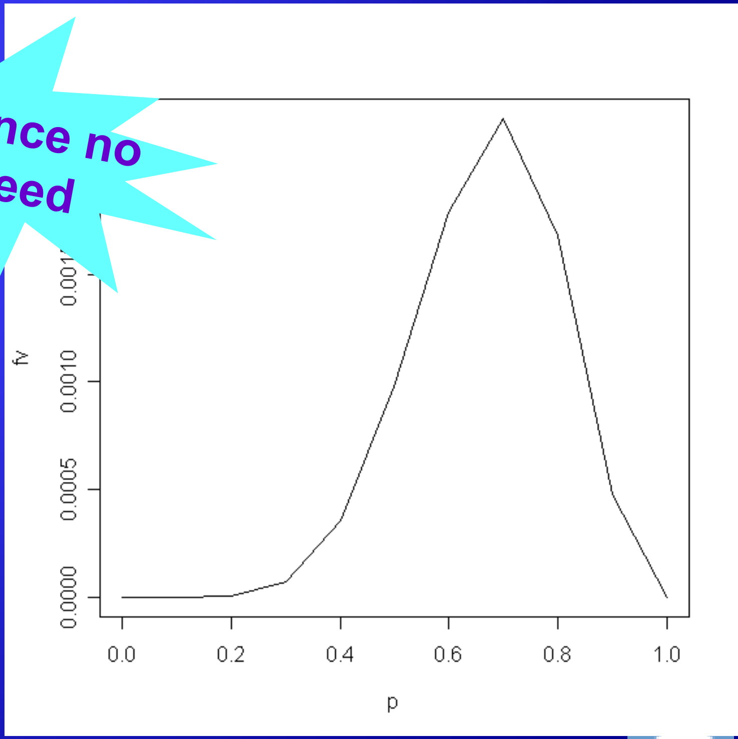
Choose as estimates of the parameters those who maximize the probability of the observed data

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

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

```
> 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")
```

Convergence no guaranteed



# MLE of the Logistic Regression

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

$$\log L(\beta) = l(\beta) = \sum_{i}^{n} \log p_i = \sum_{i}^{n} (y_i \log p_i + (1-y_i)\log(1-p_i))$$

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

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

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

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

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

$$\downarrow$$

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

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

# MLE of the Logistic Regression

*Newton-Raphson*

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

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

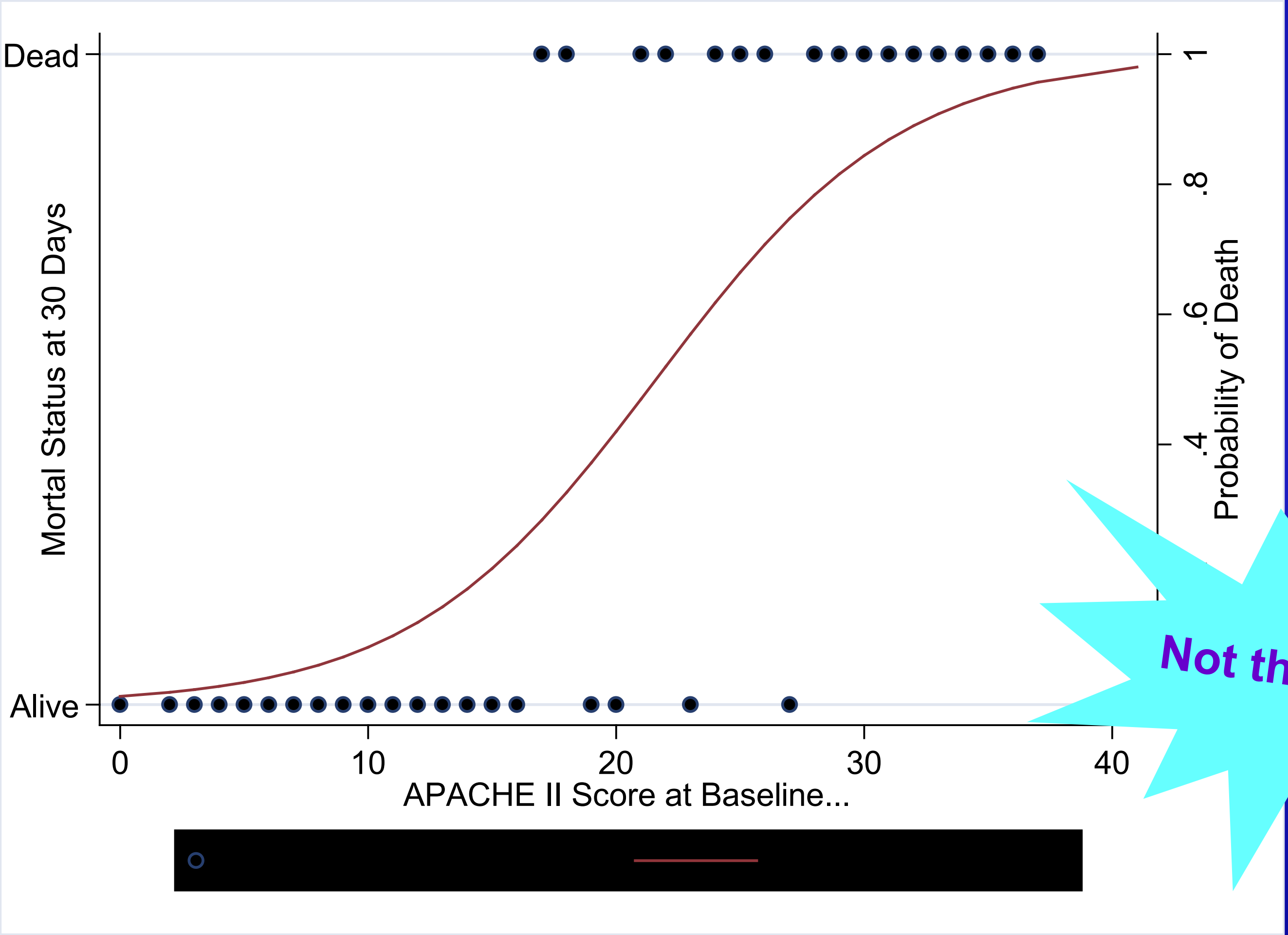
**Iterated Reweighted Least Squares (IRLS algorithm)**

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

# Joint representation of observed data and fits



Not that useful

# Model inference

- The Wald statistic for the $\beta_k$ 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 = $\sqrt{\frac{Wald-2}{2LL(\alpha)}}$

LL(α): log likelihood of the null model (only constant term)

Warning: Multicolinearity

- Determine the significant regressors

95%-IC for coeficients: $[e^{-\hat{\beta}_k} e^{\pm \chi^2{}_{1,0.05} S_{\hat{\beta}_k}}]$

For very large n, normal approach works $[e^{-\hat{\beta}_k} e^{\pm 1.96 S_{\hat{\beta}_k}}]$

# Model inference

Wald Confidence Interval for $\pi$ $(\pi(x_i))$::

$$IC(\pi, 1-\alpha) = [\hat{\pi} \pm z_{1-\alpha}\, S_{\hat{\pi}}]$$

**n large**

$n\pi > 5$,

$n\,(1-\pi) > 5$

Wilson Confidence Interval

$$IC(\pi, 1-\alpha) = [\hat{\pi} + \frac{1}{2n} {z_{1-\alpha}}^2 \pm z_{1-\alpha} \sqrt{\frac{1}{n}\hat{\pi}(1-\hat{\pi}) + \frac{1}{4n^2} {z_{1-\alpha}}^2}]$$

**Larger intervals**

$$1-\alpha = 0.95 \rightarrow z_{1-\alpha} = 1{,}96$$

©K. Gibert

# Model assessment/validation

Numerical (eval training/test):

- *Confusion matrix*
- *Goodness of fit indicators*
  - *Deviance*
  - *Pseudo-$R^2$*
  - *AIC (Akaike information criterion)*

**$R^2$ non reliable**

Graphical:

- *Residuals plots*
- *ROC curve*

*Simple/cross validation (generalization error)*

# Model assessment/validation

## Deviance

Ratio between Likelihood of the proposed model and the perfect one $p_i$ = $y_i$ (saturated).

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

Measures proximity between model fit and data,

*Same role as sum of residual squares in linear models.*

Perfect model: Deviance=0 ($H_0$)

Signifficance: Deviance too big : Model invalid

# Model assessment/validation

## Deviance

Ratio between Likelihood of the proposed model and the perfect one $p_i = y_i$ (saturated).

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

Measures proximity between model fit and data,

*Same role as sum of residual squares in linear models.*

Perfect model: Deviance=0 ($H_0$)

Signifficance: Deviance too big : Model invalid

# Model assessment/validation

By chance accuracy

## Null Deviance ($D_0$)

Deviance of the null model (just with constant term)

improvement over Accuracy of random assignment

## Residual deviance ($D_e$) :

Deviance of the proposed model

***$D_0 - D_e \sim \chi^2_{\nu 0 \text{ - } \nu e}$ added explanatory capacity of variables in model***

## AIC:

Deviance with complexity penalization *(+2p)*

Standard errors >2

Point numerical problems

# Pseudo-$R^2$

- McFadden's-$R^2$ statistic (a pseudo-$R^2$) :

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

$R^2$ in [0,1], 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: $\log \frac{p_{i'}}{1-p_{i'}} - \log \frac{p_i}{1-p_i} = 0.03394 \qquad \frac{p_{i'}/1-p_{i'}}{p_i/1-p_i} = e^{0.03394} = 1.0345$

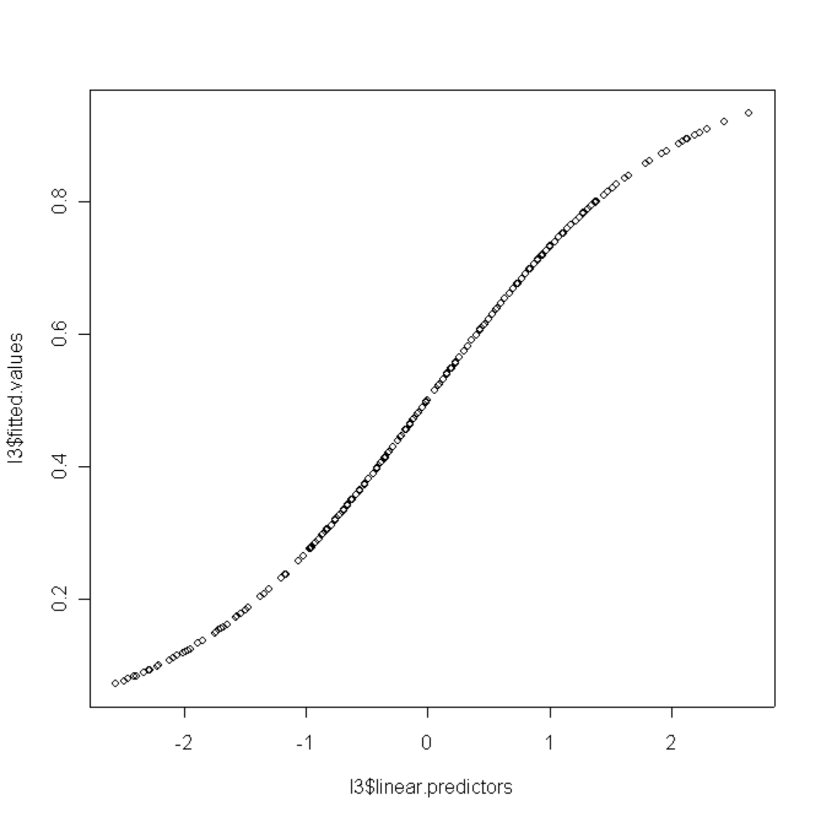
**Interpret the coefficients**

```
> exp(l3$coefficients)
 (Intercept)          edat          ratfin  tiptrebauton    tiptrebfixe    tiptrebtemp

   0.5970355     1.0345176     0.9666757      5.0495067      9.3171141      1.7555279
```

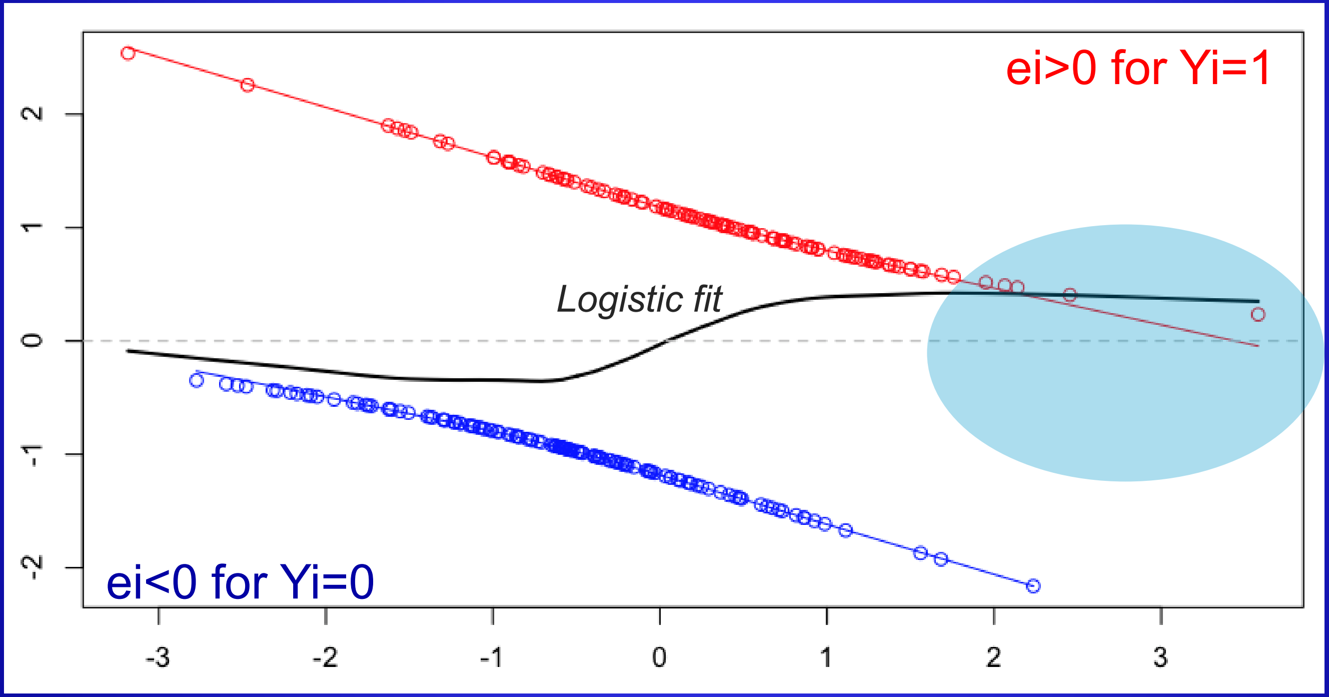
# Plot of the linear predictor and the estimated probabilities

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



# *Residuals analysis*

```
plot(predict(reg),residuals(reg))
abline(h=0,lty=2,col="grey")
```



©K. Gibert

# 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
```

$$Deviance_1 - Deviance_2 \sim \chi_{\nu_1-\nu_2} \qquad E\left[\chi_\nu^2\right]=\nu$$

# 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%***

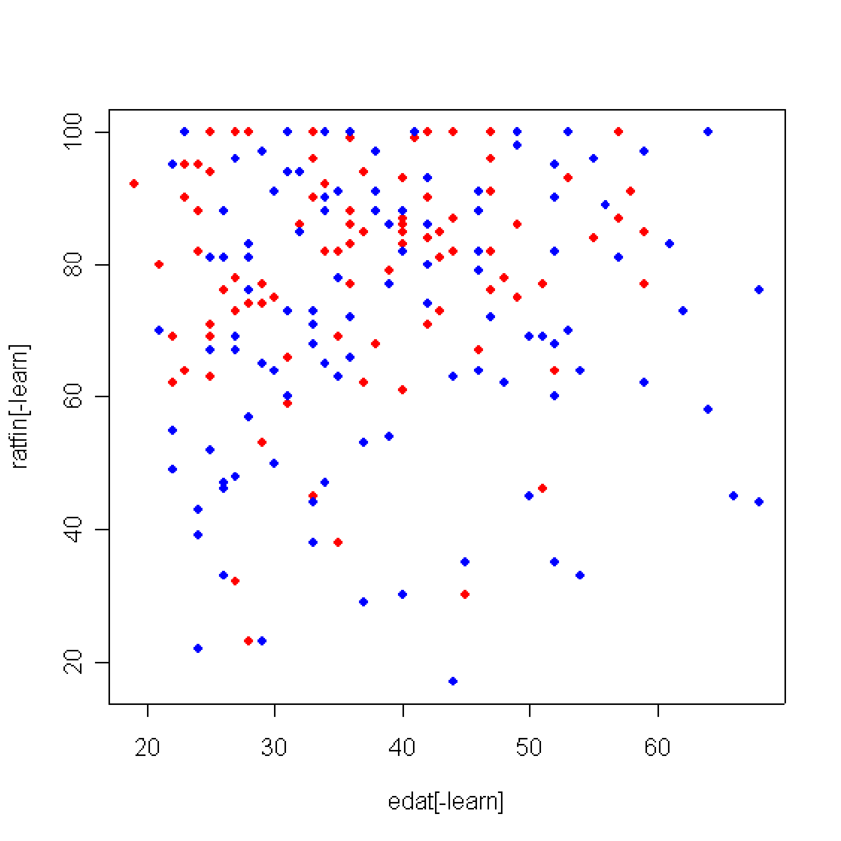
## GE in *test*

```
> 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
```

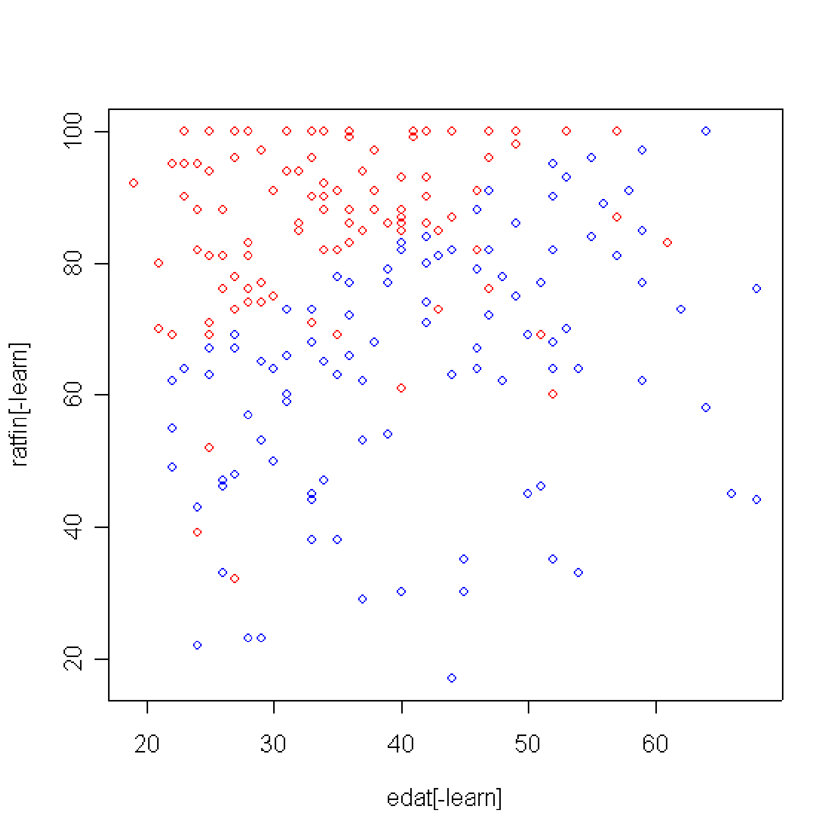
***$P_{acierto}$=62.0%***

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

**Actual response *(test)***



**Prediction (*test*)**



# Regression Diagnostics

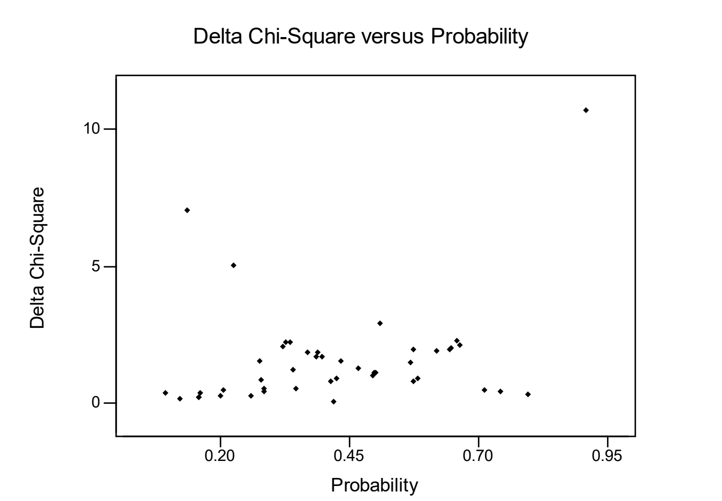
- In logistic regression Residual = 1− Estimated probability.

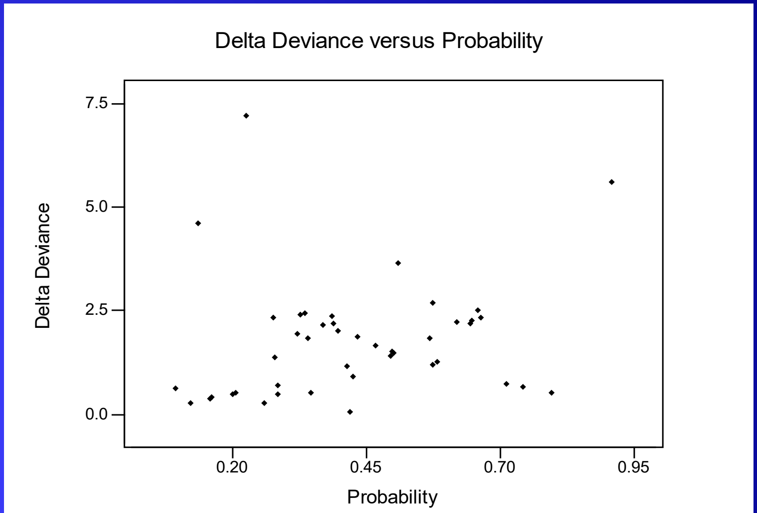
Residuals for each subject are calculated standardised and plotted against probability. Eight diagnostic plots are available, four dealing with residuals and four with leverage.

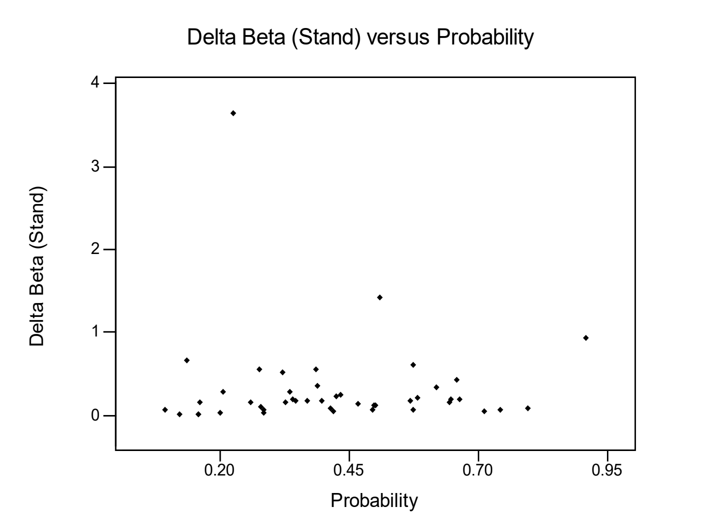
- These plots are demonstrated in the slides that follow.

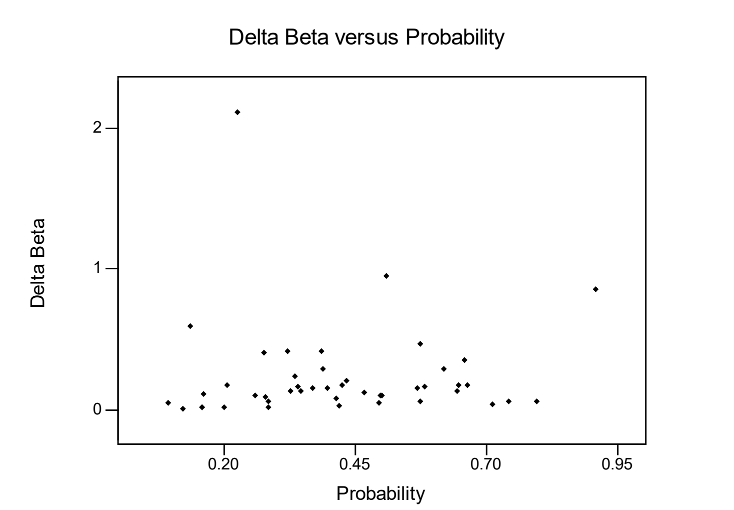
- ROC and concentration curves

# Diagnostic plots for residuals

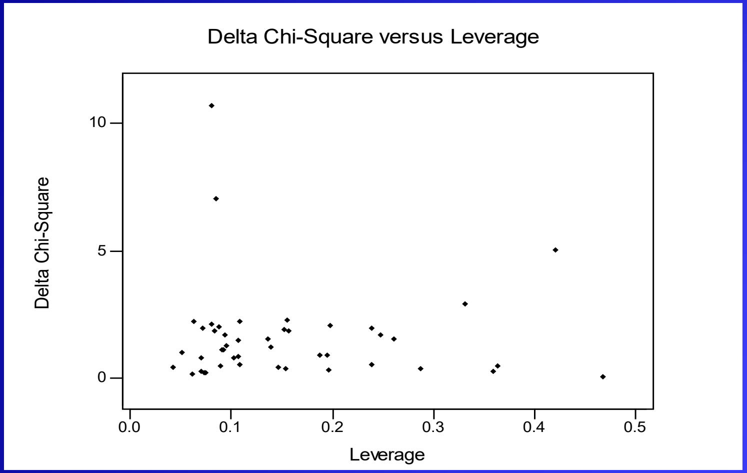


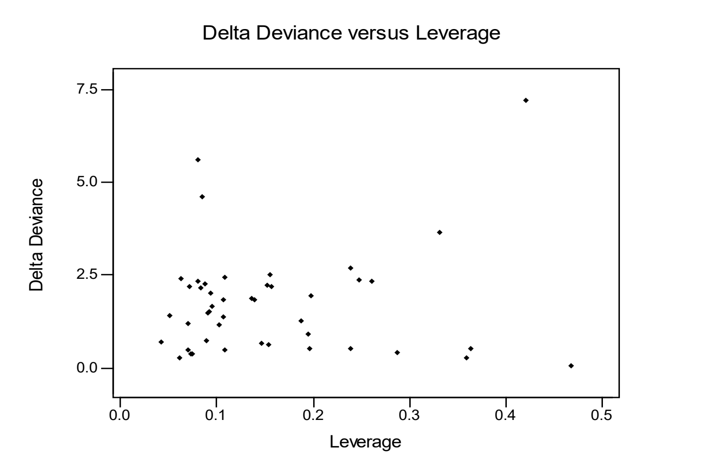


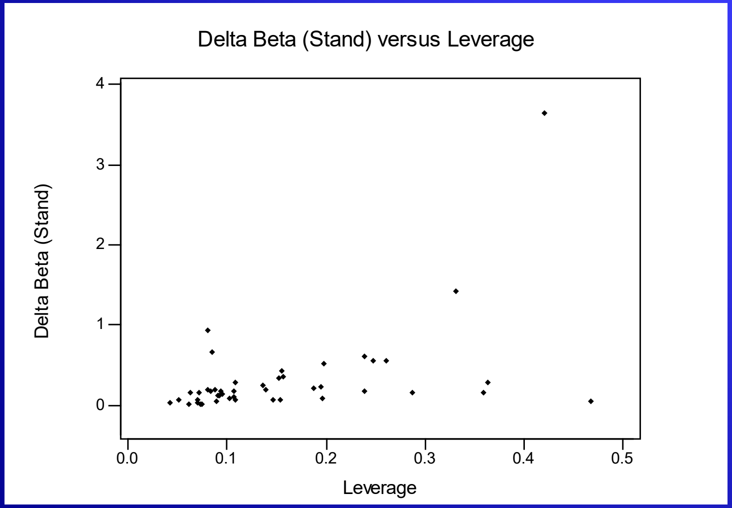


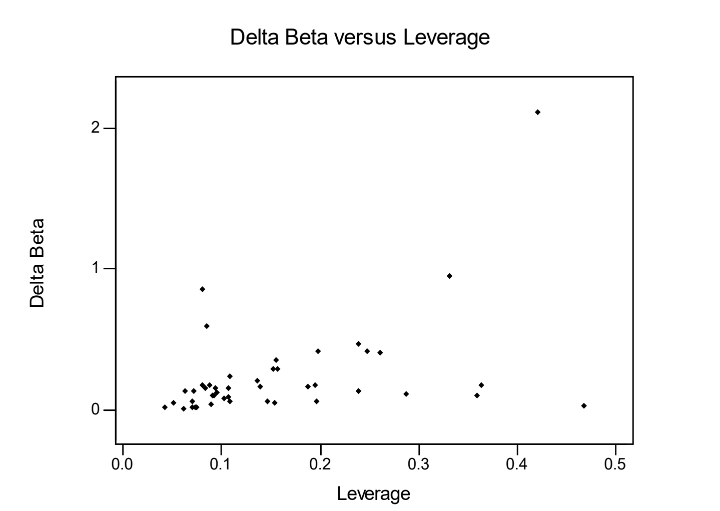


# Diagnostic plots for leverage


Delta Chi-Square versus Leverage


Delta Deviance versus Leverage


Delta Beta (Stand) versus Leverage


Delta Beta versus 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 |

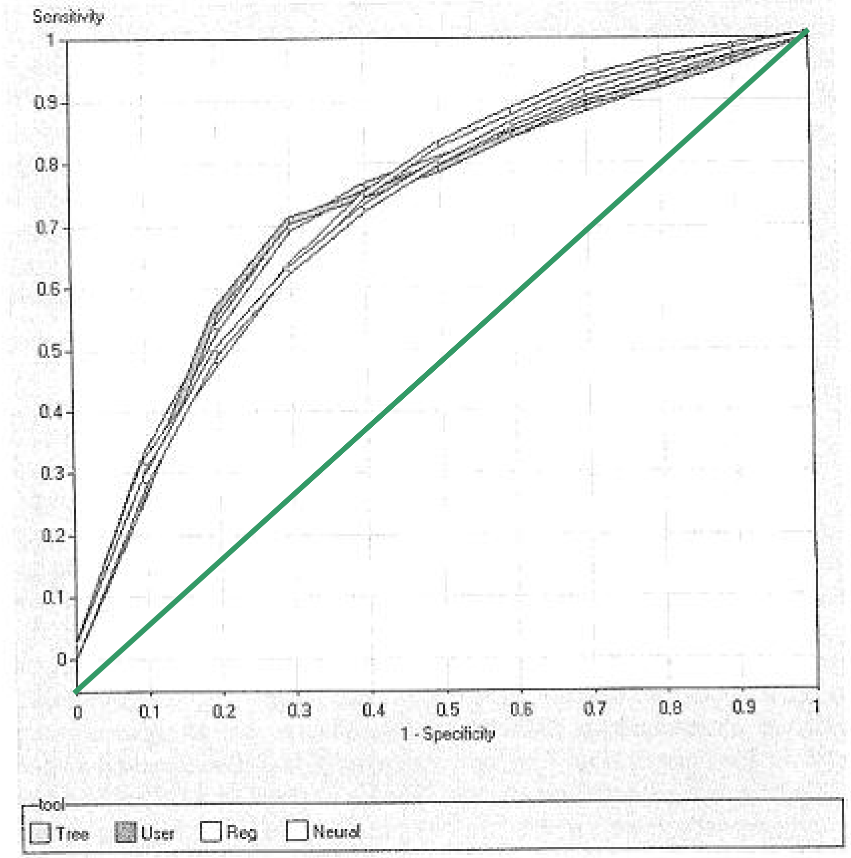


**Figure 10.5** ROC curves for the considered models. The curve called user is the MBR model.

## 2. The Binomial Distribution

Let

*m* be the number of people at risk of death

*d* be the number of deaths

$\pi$ be the probability that any patient dies.

The death of one patient has no effect on any other.

Then *d* has a **binomial distribution** with

parameters *m* and $\pi$,
mean $m\pi$, and
variance $m\pi(1-\pi)$.

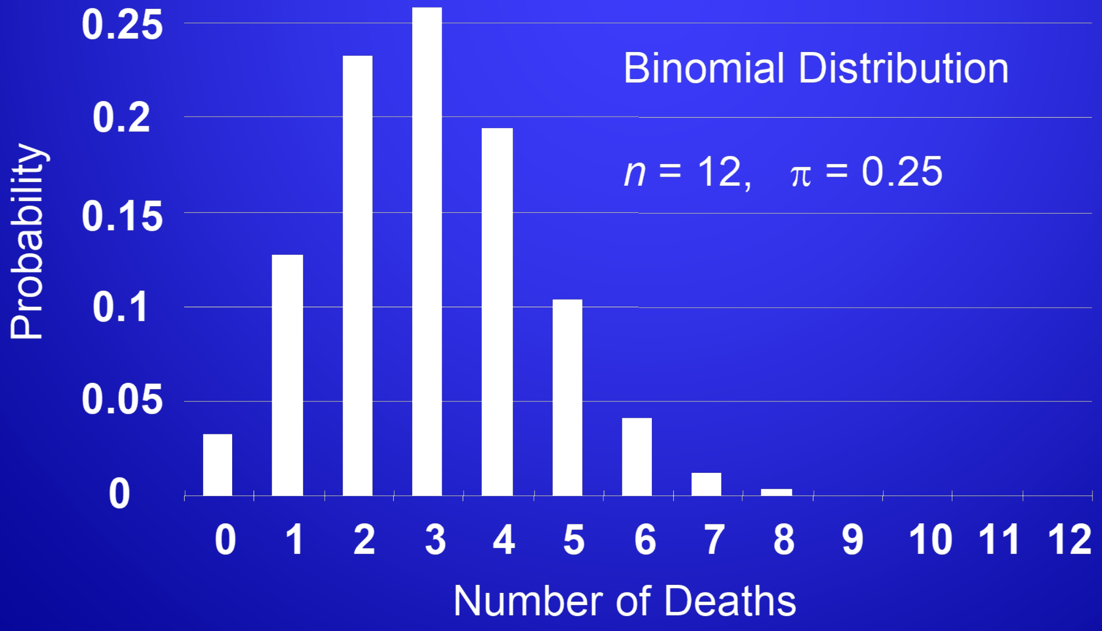
Pr[*d* deaths]

$$= \frac{m!}{(m-d)!d!} \pi^d (1-\pi)^{(m-d)} \quad : d = 0, 1, \cdots, m \qquad \{3.4\}$$

The population mean of any random variable *x* is also equal to its expected value and is written *E*(*x*). Hence

$E(d) = \pi m$ and $E(d/m) = \pi$

For *m* = 12 and π = 0.25 this distribution is as follows.



©K. Gibert

## 3. Generalized Linear Models

Logistic regression is an example of a generalized linear model. These models are defined by three attributes: The distribution of the model's random component, its linear predictor, and its link function. For logistic regression these are defined as follows.

## a) The random component

$d_i$ *is* the **random component** of the model. In logistic regression, $d_i$ has a binomial distribution obtained from $m_i$ trials with mean $E(d_i)$. (In the sepsis example, $m_i = 1$ for all $i$.)

Stata refers to the distribution of the random component as the distributional family.

## b) The linear predictor

$\alpha + x_i \beta$ is called the **linear predictor**

## c) The link function

$E(d_i)$ is related to the linear predictor through a **link function**. Logistic regression uses a logit link function

$$\text{logit}(E(d_i)) = \alpha + x_i \beta$$

# How can we model binary responses?

The response is binary 0/1

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

**$\varepsilon_i$ bernoulli, $\hat{y}$ continuous**
Violation of linear model hypothesis

$$y_i = r(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) + \varepsilon_i = E\left[ y_i \middle| x_{i1}, \ldots, x_{ip} \right] + \varepsilon_i$$

For each individual we know *y<sub></sub>*$y_i$ , but we would need an estimation of $p_i$

**$\varepsilon_i$ binomial** $\varepsilon \sim B(n_i, p_i)$

$$E\left[ y_i \middle| x_{i1}, \ldots, x_{ip} \right] = 1 \times p_i + 0 \times (1 - p_i) = p_i$$

$$y_i = p_i + \varepsilon_i$$

$$p_i = r(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

©K. Gibert

# How can we model binary responses?

The response is binary 0/1

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

$$y_i = r(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) + \varepsilon_i = E\left[ y_i \middle| x_{i1}, \ldots, x_{ip} \right] + \varepsilon_i$$

$$E\left[ y_i \middle| x_{i1}, \ldots, x_{ip} \right] = 1 \times p_i + 0 \times (1 - p_i) = p_i$$

Lets do a transformation of the linear component : *β’x<sub></sub>*$_i$

In such a way that *r(β’x$_i$)* mapping function in the interval *0:1* (=probability)

$$p_i = r(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

**$\varepsilon_\iota$ binomial** $\varepsilon$ □ $B(n_i, p_i)$

$$\boxed{y_i = p_i + \varepsilon_i}$$

**Generalized Linear Model**

*n$_i$* number of observations in individual *i*

*p$_i$* probability of *y=1* for individual *i*

# Which $r$ function to choose?

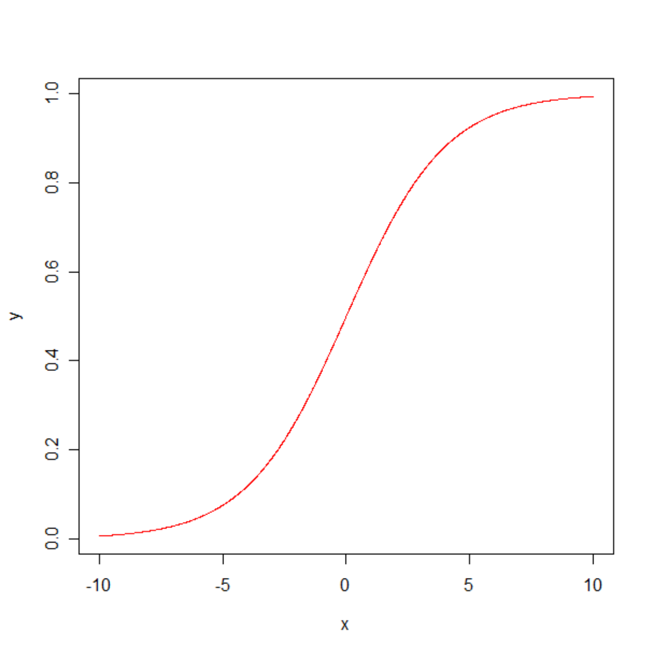
## Logistic function

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

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

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

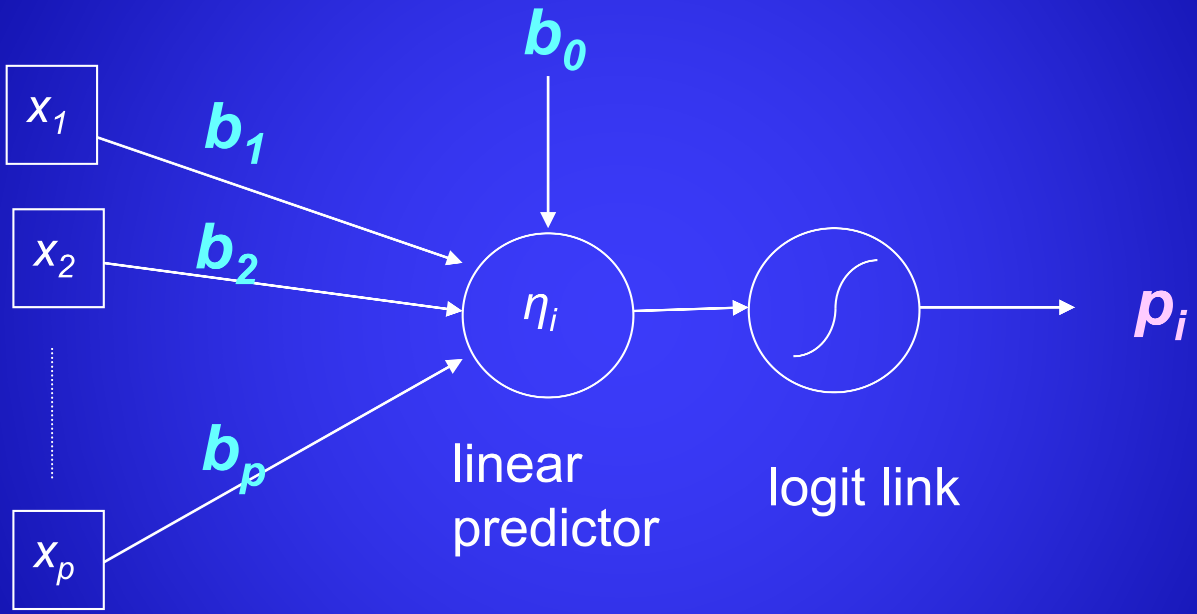
$$\ln \frac{P(+/x_i)}{P(-/x_i)} = \beta' x_i$$



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

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

# A graphical representation of the logistic regression



$$\eta_i = b_0 + b_1 x_1 + \cdots + b_p x_p \qquad p_i = \frac{1}{1+e^{-\eta_i}}$$

But, how to estimate the $b_0$, $b_1$, …, $b_p$

# Multiple logistic regression

Several independent variables

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

✓$\beta_0$ = log odds ratio for X=0 *(baseline odds ratio, moves curve left/right)*

✓$\beta_\kappa$ = 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)*

✓$e^{\beta_\kappa}$ = unitary marginal odds ratio

**Regressors numerical or dummy**

# Logistic regression

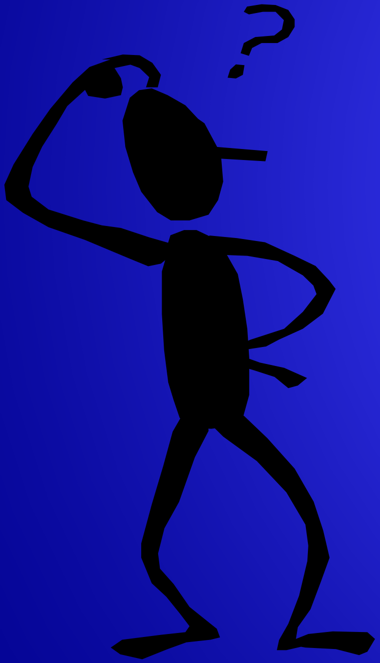
***Karina Gibert***

*Dpt. Statistics and Operation Research*

*Knowledge Engineering and Machine Learning Research group at*
*Intelligent Data Science and Artificial Intelligence Specific Research Center*

*Institut Universitari de Recerca en Ciència y Tecnologia de la Sostenibilitat*
*Universitat Politècnica de Catalunya-BarcelonaTech (Spain)*

*karina.gibert@upc.edu*
*www.eio.upc.edu/homepages/karina*



*Are there any questions?...*