Experimental Design and Statistical Methods

Workshop

LOGISTIC REGRESSION

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Logistic regression model
  - Logit transformation
  - Odds ratio
  - Link function
  - ML estimation
- Sensitivity and specificity
- ROC curves

- Basic commands
  - glm
  - hoslem.test
  - round
  - paste
  - prop.table
- Libraries
  - epicalc
  - ResourceSelection

Data

We are interested in understanding and assessing the relationship between **arthrosis** and **age**.

Occurrence or non-occurrence are signified as **1** and 0, respectively.

In this example, the response variable is not a continuous variable, but a **binary variable** following a Bernoulli distribution:

$$P(y) = p^y(1-p)^{(1-y)}, \text{ with } y = 0,\ 1$$

As before, we postulate that the **probability of occurrence** is a function of a regressor variable x_1, or, more generally, of a set x_1, x_2, …, x_k regressor variables.

In this case, homogeneity of variances and normality of errors are not satisfied: usual *F* and *t* tests are not valid.

Age (years)	Occurrence
32	0
79	1
44	0
36	0
76	1
79	1
53	0
52	0
23	0
24	0
77	1
23	0
28	0
45	1
89	1
27	0
52	1
53	0
35	0
33	1
78	1
22	0
75	1
34	0
69	1

Logistic regression model

In the general case of k regressors we can write the **logistic model**:

$$P(y_i = 1 \mid \mathbf{x}_i) = p_i = \frac{1}{1 + e^{-\mathbf{x'}_i \boldsymbol{\beta}}} \quad \text{where} \quad \mathbf{x'}_i \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

Taking natural logarithms (base e) of p_i, we can linearize the logistic model and then obtain the **logit transformation** (or log odds):

$$\operatorname{logit}(p_i) = \ln\left[\underbrace{\frac{p_i}{1 - p_i}}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

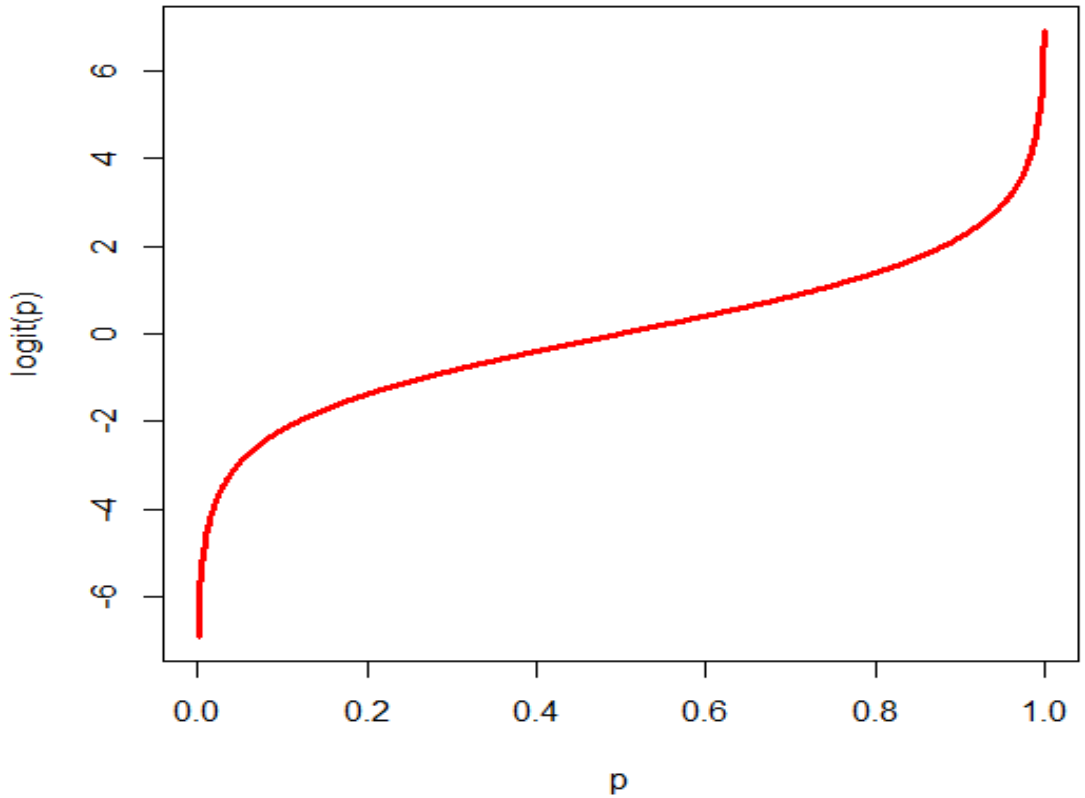
odds of an event, occurrence of arthrosis in this case

This logit transformation is a **link function** relating the probability of occurrence to several explanatory variables.

Prevents of obtaining out of range probability predictions: calculating the reverse transformation always gives a (positive) probability.

Logit function of *p*

```
> p<-seq(0,1,0.001)
> plot(p, log(p/(1-p)), type="l", col="red", lwd="3", ylab="logit(p)")
```



Logit (p) monotonically increases with p (never decreases). Also, for positive β, as x increases, p and logit(p) increase. The reverse is true for negative values of β.

Logit(p) is linear in a wide range (0.20-0.80).

A first look to the data

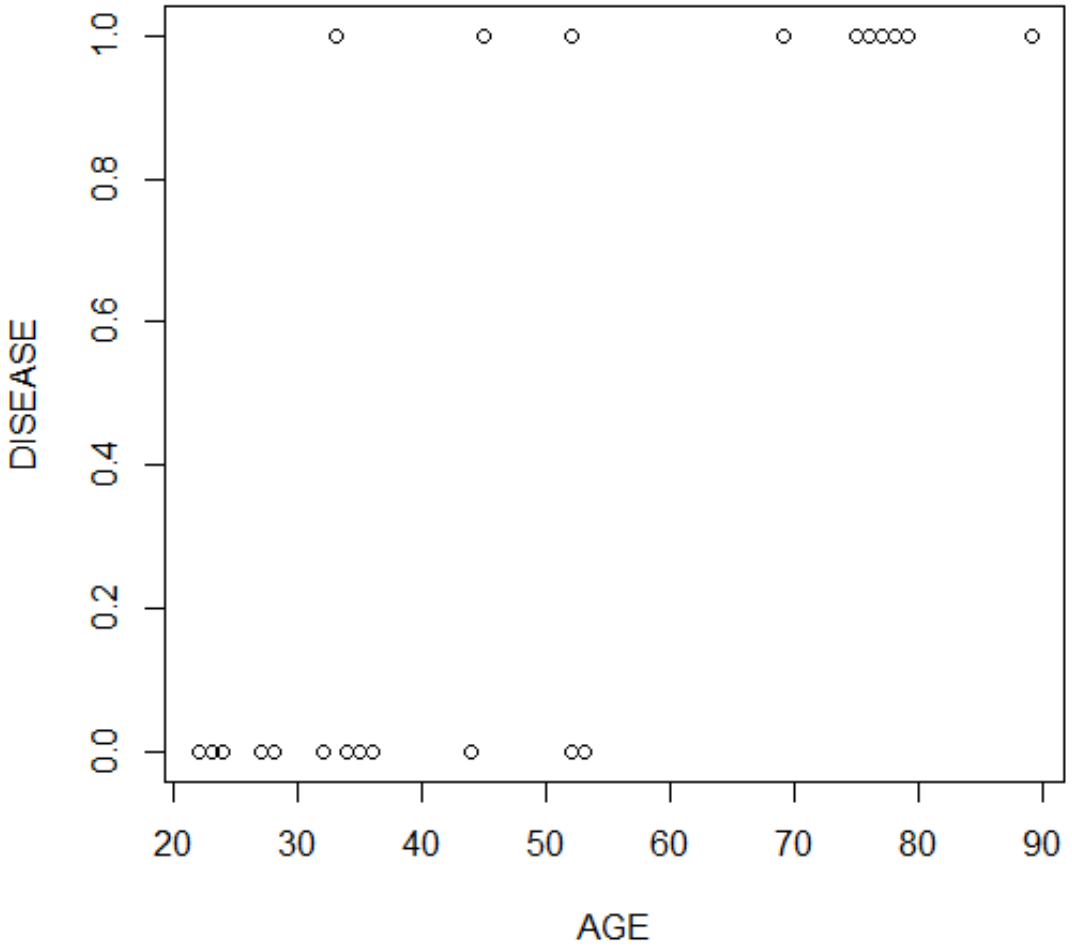
```
> DISDATA<-read.table("agedislog.txt", header=TRUE)
> attach(DISDATA)

> print(paste("Number of observations =", nrow(DISDATA)),
quote=FALSE)

[1] Number of observations = 25

> T<-table(DISDATA$DISEASE)
> PT<-prop.table(table(DISDATA$DISEASE))
> O<-round(PT/(1-PT), digits=3)
> cbind(N_obs.=T,Frequency=PT, Odds=O)

  N_obs. Frequency  Odds
0     14      0.56 1.273
1     11      0.44 0.786
```

p = 11/25 = 0.44

Odds of arthrosis
= 0.44/(1-0.44) = 0.786

The data in a graphic

```
> Plot(AGE, DISEASE)
```



Arthrosis is more common in aged people

Likelihood function (Searle et al., 1992)

Suppose a vector of random variables, $\mathbf{x}$, has density function $f(\mathbf{x})$. Let be $\boldsymbol{\theta}$ the vector of parameters involved in $f(\mathbf{x})$. Then $f(\mathbf{x})$ is a function both of $\mathbf{x}$ and $\boldsymbol{\theta}$. As a result, it can be thought of in at least two different contexts:

1. The first is as a **density function**, in which case, $\boldsymbol{\theta}$ is assumed to be known. We use the symbol $f(\mathbf{x}|\boldsymbol{\theta})$ in place of $f(\mathbf{x})$ to emphasize that $\boldsymbol{\theta}$ is being taken as known.
2. A second context is where $\mathbf{x}$ represents a known vector of data and where $\boldsymbol{\theta}$ is unknown. Then $f(\mathbf{x})$ will be a function of just $\boldsymbol{\theta}$. It is called the **likelihood function** for the data $\mathbf{x}$, and because in this context $\boldsymbol{\theta}$ is unknown and $\mathbf{x}$ is known, we use the symbol $L(\boldsymbol{\theta}\,|\mathbf{x})$. Nowadays, mathematically both functions represent the same:

$$f(\mathbf{x}|\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\mathbf{x})$$

Maximun Likelihood estimation

The likelihood function is the foundation of the widely used method of estimation, known as **maximum likelihood (ML) estimation**. It yields estimators that have many good properties

The essence of the ML method is to view $L(\boldsymbol{\theta}|\mathbf{x})$ as a function of the mathematical variable $\boldsymbol{\theta}$ and to derive $\hat{\boldsymbol{\theta}}$ as that value of $\boldsymbol{\theta}$ which maximizes $L(\boldsymbol{\theta}|\mathbf{x})$.

The only proviso is that this maximization must be carried out within the range of permissible values for $\boldsymbol{\theta}$. For example, the estimate of a variance component must be only positive.

We proceed differentiating L with respect to $\boldsymbol{\theta}$ and equating the derivative to 0. But maximizing L is equivalent to maximizing the natural logarithm of L, which we denote by l, and it is often easier to use l than L.

Note that the ML method of estimation requires a known distribution relating the parameters to the data.

Likelihood ratio test and Deviance

A **likelihood-ratio test** can be used under full ML. The use of such a test is a quite general principle for statistical testing. In hierarchical linear models, the **deviance test** is mostly used for multiparameter tests and for tests about the random part of the model. This approach is based on estimating two models, M_0 and M_1. It is assumed that M_0 excludes the effects hypothesized to be null, while these effects are included in M_1.

For each model, a **deviance statistic**, equal to **-2 ln *L*** for that model, is computed. The deviance can be regarded as a measure of lack of fit between model and data. In general, the larger the deviance, the poorer the fit to the data.

The **difference between the deviances** has a large-sample chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated. Large values of the chi-square statistic are taken as evidence that the null hypothesis is implausible.

An approximate rule for evaluating the chi-square is that the difference in deviance between models should be at least twice as large as the difference in the number of parameters estimated.

Akaike information criterion

The **Akaike information criterion** (**AIC**) is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model. AIC does not provide a test of a model in the sense of testing a null hypothesis; i.e. AIC can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any warning of that.

In the general case, $AIC = 2k - 2\ln L = 2k + Deviance$

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model.

Given a set of candidate models for the data, ***the preferred model is the one with the minimum AIC value***. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting.

(From Wikipedia)

Fitting the logistic regression model – 1 -

```
> DISDATA.GLM <- glm(DISEASE~ AGE, family=binomial())
> anova(DISDATA.GLM, test="Chisq")

Analysis of Deviance Table

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    24     34.296
AGE   1   18.514         23     15.783 1.687e-05 ***
```

The Analysis of Deviance Table is similar to an ANOVA table.
Model deviance in GLM setting is an analogous concept to the residual sum of squares in the regular linear regression model.
We contrast the residual deviances of two models, one with the intercept fitted (NULL) and the second with the full model fitted (AGE). The p-value of the Chi-square statistic is interpreted as in ANOVA.

Fitting the logistic regression model – 2 -

```
> summary(DISDATA.GLM)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.4194     2.3266  -2.759  0.00579 **
AGE           0.1239     0.0459   2.699  0.00696 **

    Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 15.783  on 23  degrees of freedom
AIC: 19.783

Number of Fisher Scoring iterations: 5
```

$$\left(\frac{\hat{\beta}}{s.e.(\hat{\beta})}\right)$$

```
> 2*(1-pnorm(2.699))
[1] 0.006954818
```

0.1239 is the **increment in the logit of *p*** for 1 year increase in age. It is significant. The probability of having arthrosis augments with age (beta for AGE > 0).

AIC = 15.783 + 2*2 (Residual deviance + 2k, k being Intercept and AGE)

The parameters of the logistic model are estimated by **Maximum Likelihood** using an iterative procedure with different optimization techniques, Fisher scoring in this case.

Fitting the logistic regression model – 3 -

```
> exp(coef(DISDATA.GLM))  # exponentiated coefficients

(Intercept)         AGE
0.001629669 1.131874986

> library(epicalc)
> logistic.display(DISDATA.GLM)

Logistic regression predicting DISEASE

                    OR(95%CI)          P(Wald's test) P(LR-test)
AGE (cont. var.) 1.13 (1.03,1.24)  0.007              < 0.001

Log-likelihood = -7.8915
AIC value = 19.7829
```

1.132 is the **increment in the odds of *p*** for 1 year increase in age

Note that $e^{0.1239} = 1.132$ and $ln(1.132) = 0.1239$

Fitting the logistic regression model – 4 -

```
> library(ResourceSelection)
> hoslem.test(DISDATA.GLM$y, fitted(DISDATA.GLM))

Hosmer and Lemeshow goodness of fit (GOF) test

X-squared = 6.3015, df = 8, p-value = 0.6135
```

The Hosmer-Lemeshow test of goodness-of-fit divides subjects into deciles based on predicted probabilities, then computes a chi-square from observed and expected frequencies. It tests the null hypothesis that there is no difference between the observed and predicted values of the response variable. Therefore, when the test is not significant, as in this example, we can not reject the null hypothesis and then we say that the model fits the data well.

Logistic regression diagnostics

There are three basic building blocks for logistic regression diagnostic:

1. **Pearson residuals** are defined as the standardized difference between the observed frequency and the predicted frequency. They measure the relative deviations between the observed and fitted values.
2. **Deviance residuals** are the signed square roots of the contributions to the **Deviance** statistic. They are parallel to the raw residual in OLS regression, where the goal is to minimize the sum of squared residuals. The Deviance statistic measures the difference between the log likelihood under the model and the maximum likelihood that is achievable.
3. **Hat diagonal** technically it is the diagonal of the hat matrix, measures the leverage of an observation. It is also sometimes called the Pregibon leverage.

As in lm, R also gives Dfbetas, Dffits, Cov. ratio and Cook distance.

Logistic regression diagnostics – 1 -

```
> RD<-resid(DISDATA.GLM, type="deviance") # deviance residuals
> RP<-resid(DISDATA.GLM, type="pearson") # Pearson residuals
> cbind(AGE,DISEASE,Deviance_res=RD,Pearson_res=RP)
```

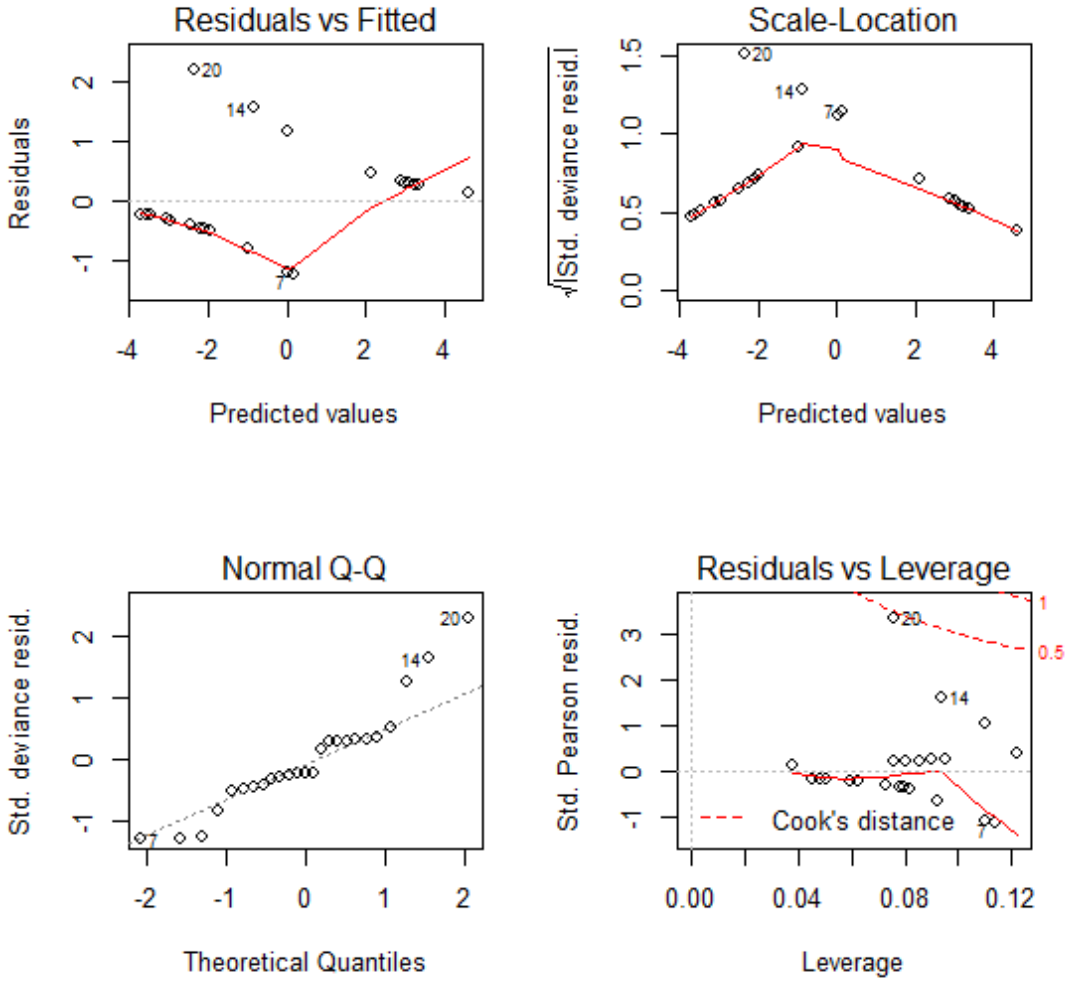
	AGE	DISEASE	Deviance_res	Pearson_res
1	32	0	-0.4058243	-0.29297130
2	79	1	0.2604552	0.18574240
3	44	0	-0.8021642	-0.61604984
4	36	0	-0.5134167	-0.37533754
5	76	1	0.3124654	0.22367052
6	79	1	0.2604552	0.18574240
7	53	0	-1.2400183	-1.07574399
8	52	0	-1.1868304	-1.01113629
9	23	0	-0.2356275	-0.16777683
10	24	0	-0.2504563	-0.17849713
11	77	1	0.2941113	0.21023718
12	23	0	-0.2356275	-0.16777683
13	28	0	-0.3192900	-0.22867998
14	45	1	1.5507097	1.52575541
15	89	1	0.1410434	0.09998129
16	27	0	-0.3005534	-0.21494578
17	52	1	1.1680186	0.98898636
18	53	0	-1.2400183	-1.07574399
19	35	0	-0.4843696	-0.35279528
20	33	1	2.2019095	3.20830533

Note that only the first 20 values are displayed.

Observation 20 (a 33 year old person with arthrosis) can be an outlier.

Logistic regression diagnostics – 1b -

```
> layout(matrix(c(1,2,3,4),2,2))
> plot(DISDATA.GLM)
```



This confirms graphically what was suggested in the previous slide.

Logistic regression diagnostics – 2 -

```
> influence.measures(DISDATA.GLM)
```

	dfb.1_	dfb.AGE	dffit	cov.r	cook.d	hat	inf
1	-0.12843	0.1075	-0.1406	1.153	0.003654	0.0731	
2	-0.06930	0.0833	0.0919	1.172	0.001533	0.0759	
3	-0.18425	0.1027	-0.3248	1.099	0.021318	0.0925	
4	-0.16204	0.1277	-0.1907	1.148	0.006848	0.0819	
5	-0.08883	0.1087	0.1223	1.185	0.002729	0.0903	
6	-0.06930	0.0833	0.0919	1.172	0.001533	0.0759	
7	0.00916	-0.1726	-0.5899	0.977	0.083612	0.1136	
8	-0.02789	-0.1260	-0.5509	0.994	0.071096	0.1101	
9	-0.06272	0.0564	-0.0643	1.140	0.000748	0.0482	
10	-0.06851	0.0612	-0.0705	1.142	0.000901	0.0510	
11	-0.08196	0.0996	0.1113	1.181	0.002256	0.0854	
12	-0.06272	0.0564	-0.0643	1.140	0.000748	0.0482	
13	-0.09574	0.0833	-0.1007	1.150	0.001854	0.0623	
14	0.34760	-0.1721	0.6784	0.835	0.132951	0.0938	
15	-0.02740	0.0317	0.0336	1.133	0.000203	0.0377	
16	-0.08834	0.0774	-0.0923	1.148	0.001554	0.0595	
17	0.02740	0.1238	0.5412	1.001	0.068016	0.1101	
18	0.00916	-0.1726	-0.5899	0.977	0.083612	0.1136	
19	-0.15400	0.1236	-0.1772	1.150	0.005881	0.0800	
20	0.85208	-0.7043	0.9462	0.527	0.455211	0.0756	*

Note that only the first 20 values are displayed.

Observation 20 is influential (*), as the estimates of the different regression diagnostics are higher than the critic values.

Anyway, we will kept it in subsequent calculus.

Prediction – 1 -

```
> PRED.PROB<-round(predict(DISDATA.GLM, type="response"), digits=3)
> PRED.DIS<-cut(PRED.PROB, breaks=c(0,0.5,1), labels=c("0","1"))
> data.frame(AGE, DISEASE, PRED.PROB, PRED.DIS)
```

	AGE	DISEASE	PRED.PROB	PRED.DIS
1	32	0	0.079	0
2	79	1	0.967	1
3	44	0	0.275	0
4	36	0	0.123	0
5	76	1	0.952	1
6	79	1	0.967	1
7	53	0	0.536	1
8	52	0	0.506	1
9	23	0	0.027	0
10	24	0	0.031	0
11	77	1	0.958	1
12	23	0	0.027	0
13	28	0	0.050	0
14	45	1	0.300	0
15	89	1	0.990	1
16	27	0	0.044	0
17	52	1	0.506	1
18	53	0	0.536	1
19	35	0	0.111	0
20	33	1	0.089	0
21	78	1	0.962	1
22	22	0	0.024	0
23	75	1	0.946	1
24	34	0	0.099	0
25	69	1	0.894	1

$$\hat{p}_{(y|x=32)} = \frac{1}{1+e^{-(-6.4194+0.1239\times 32)}} = 0.079$$

In the second line of the program we state that when the predicted probability is ≥ 0.5, the individual is expected to present arthrosis.

Note that in some cases there is a discrepancy between the observed value and the prediction of disease.

We can try to define another cutpoint do decide whether the individual will present or not the disease, as shown in the next slide.

Prediction – 2 -

```
> T<-table(DISDATA.GLM$y, fitted(DISDATA.GLM)>.5)
> TSN<-setNames(T, rep(" ",length(T)))
> colnames(TSN)<-c("Healthy", "Arthrosis"); TSN
```

```
  Healthy Arthrosis
0      11         3
1       2         9
```

False positives

False negatives

80% of the individuals (20/25) were well classified

```
> T<-table(DISDATA.GLM$y, fitted(DISDATA.GLM)>.3)
```

```
  Healthy Arthrosis
0      11         3
1       1        10
```

84% of the individuals (21/25) were well classified

```
> T<-table(DISDATA.GLM$y, fitted(DISDATA.GLM)>.7)
```

```
  Healthy Arthrosis
0      14         0
1       3         8
```

88% of the individuals (22/25) were well classified

Which cutpoint would fit better the predicted results to the observed ones?

Prediction – 3 -

Be careful:

When a prediction model is developed and it is used on the same data set to predict the accuracy of the model, we get biased results.

It is recommended to use "**jackknife**" methods: fit the model removing a particular observation and then use the fitted model to predict this lacking observation.

This is repeated for each of the observations.

Sensitivity and specificity

The accuracy of the classification is measured by its **sensitivity** (the ability to predict an **event** correctly) and **specificity** (the ability to predict a **nonevent** correctly).

Sensitivity is the proportion of **event** responses that were predicted to be **events**. For a probability of 0.5, 8 of 11 observations were classified correctly, i.e. 72.7%.

Specificity is the proportion of **nonevent** responses that were predicted to be **nonevents**. For a probability of 0.5, 11 of 14 observations were classified correctly, i.e. 78.6%.

Ideally we are interested in high values of both sensitivity and specificity

Prediction – 3 -

```
> PSSN<-setNames(PSS, rep(" ",length(PSS)))
> colnames(PSSN)<-c("Prob.cutpoint", "1-Specificity", "Sensitivity"); PSSN
```

	Prob.cutpoint	1-Specificity	Sensitivity
1	0.0242	1.0000000	1.00000000
2	0.0273	0.9285714	1.00000000
3	0.0308	0.7857143	1.00000000
4	0.0441	0.7142857	1.00000000
5	0.0496	0.6428571	1.00000000
6	0.0790	0.5714286	1.00000000
7	0.0885	0.5000000	1.00000000
8	0.0990	0.5000000	0.90909091
9	0.1106	0.4285714	0.90909091
10	0.1234	0.3571429	0.90909091
11	0.2751	0.2857143	0.90909091
12	0.3004	0.2142857	0.90909091
13	0.5055	0.2142857	0.81818182
14	0.5364	0.1428571	0.72727273
15	0.8935	0.0000000	0.72727273
16	0.9464	0.0000000	0.63636364
17	0.9523	0.0000000	0.54545455
18	0.9576	0.0000000	0.45454545
19	0.9624	0.0000000	0.36363636
20	0.9666	0.0000000	0.27272727
21	0.9901	0.0000000	0.09090909

The probability cutpoints are those of the predicted probabilities, ties not included (ages 53 and 79 were repeated).

The values change according the specific cutpoints assumed. As you increases the probability cutpoint from 0 to 1, sensitivity decreases and specificity increases.

In our case, the better cutpoint correspond to a probability of 0.3004, with the highest combination of sensitivity and specificity values.

Receiver Operating Characteristic (ROC) curves

ROC curves are used to evaluate and compare the performance of diagnostic tests and to evaluate model fit.

An ROC curve is just a plot of the proportion of true positives – sensitivity- (events predicted to be events) versus the proportion of false positives -1-specificity- (nonevents predicted to be events).

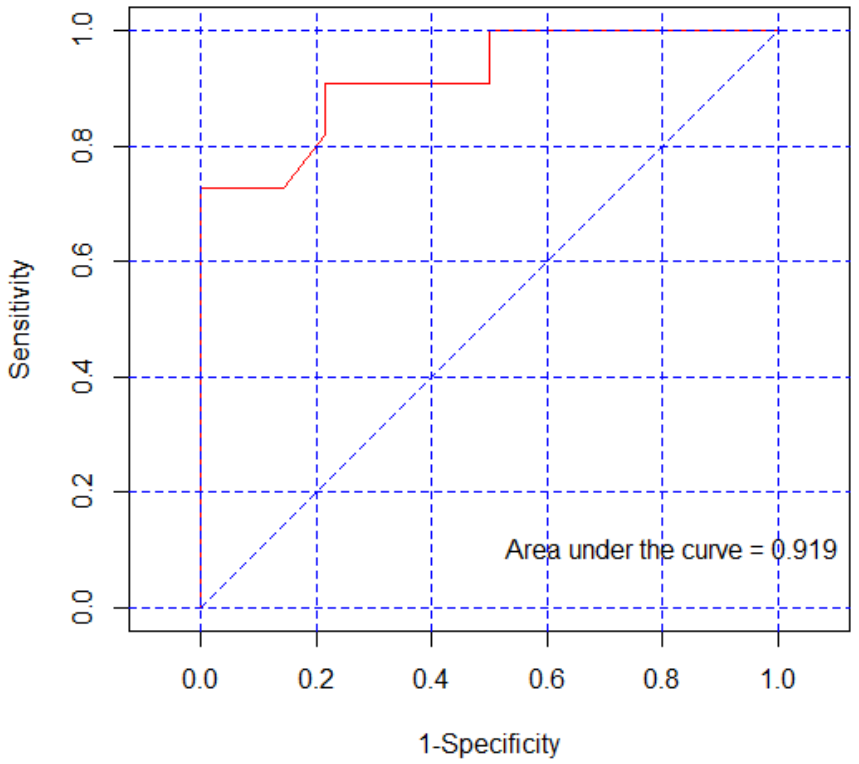
Suppose the individuals undergo a test for predicting the event and the test is based on the estimated probability of the event. Higher values of this estimated probability are assumed to be associated with the event.

A receiver operating characteristic (ROC) curve can be constructed by varying the cutpoint that determines which estimated event probabilities are considered to predict the event. For each cutpoint, two measures are calculated (see next slide).

The ROC curve is a plot of sensitivity against 1–specificity

Prediction – 4 -

```
> library epicalc
> lroc(DISDATA.GLM)
```



The ROC curve (in red) rises quickly which gives an area under the curve of 0.919, close to the maximum value (1).

Very good predictive power of our logistic regression model

The closest point to the maximum (1 for sensitivity, 1 for specificity) indicates the optimal cutpoint. In this case, 0.91 for sensitivity and 0.79 for specificity, corresponding to a **`Prob.cutpoint`** of 0.30 (see previous slide).

Final considerations

This model can accommodate **ordinal** and **nominal** responses.

The regression logistic model can include **several explanatory variables**, some of them **binary**. The table on the right presents an example with two explanatory variables and the student can work the appropriate model.

With several explanatory variables, different procedures of **variable selection** can be applied.

Other link functions can be assumed like in the Probit model, where the probability of 0 and 1 values are determined from the area under the normal distribution.

Vol	Rate	Resp
3.70	0.825	1
3.50	1.090	1
1.25	2.500	1
0.75	1.500	1
0.80	3.200	1
0.70	3.500	1
0.60	0.750	0
1.10	1.700	0
0.90	0.750	0
0.90	0.450	0
0.80	0.570	0
0.55	2.750	0
0.60	3.000	0
1.40	2.330	1
0.75	3.750	1
2.30	1.640	1
3.20	1.600	1
0.85	1.415	1
1.70	1.060	0
1.80	1.800	1
0.40	2.000	0
0.95	1.360	0
1.35	1.350	0
1.50	1.360	0
1.60	1.780	1
0.60	1.500	0
1.80	1.500	1
0.95	1.900	0
1.90	0.950	1
1.60	0.400	0
2.70	0.750	1
2.35	0.030	0
1.10	1.830	0
1.10	2.200	1
1.20	2.000	1
0.80	3.330	1
0.95	1.900	0
0.75	1.900	0
1.30	1.625	1