Logistic regression

STATS – Modelação Estatística

PhD Programme in Health Data Science

Cristina Costa Santos & Andreia Teixeira

Logistic Regression

Example:

A study conducted by Payne, 1987, comprised 2074 children less than 1 year-old and its goal was to relate the incidence of pulmonar infections with the type of milk being administered and the sex of the child.

	Only Formula Milk	Breast Feeding with Supplement	Only Breast Feeding
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

Logistic Regression

Example:

A study conducted by Payne, 1987, comprised 2074 children less than 1 year-old and its goal was to relate the incidence of pulmonar infections with the type of milk being administered and the sex of the child.

	Only Formula Milk	Breast Feeding with Supplement	Only Breast Feeding
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

There are **6** covariate patterns.
Binary response: each studied children either has or not a pulmonar infection
Explanatory variables: sex (2 categories) and **type of milk** (3 categories – 2 dummies)

Logistic Regression

```
pinf<-rep(c(1,0),e=6)
pinf
sex<-rep(rep(c("boy","girl"),e=3),2)
food<-rep(c("Formula","Supp","Breast"),4)
val=c(77,19,47,48,16,31,458-77,147-19,494-47,384-48,127-16,464-31)

base0<-data.frame(pinf,sex,food,val)
base0
View(base0)
base0$food <- as.factor(base0$food)
base0$food=relevel(base0$food, ref="Formula")
```

Logistic Regression

```
pinf<-rep(c(1,0),e=6)
pinf
sex<-rep(rep(c("boy","girl"),e=3),2)
food<-rep(c("Formula","Supp","Breast"),4)
val=c(77,19,47,48,16,31,458-77,147-19,494-47,384-48,127-16,464-31)

base0<-data.frame(pinf,sex,food,val)
base0
View(base0)
base0$food <- as.factor(base0$food)
base0$food=relevel(base0$food, ref="Formula")
```

```
mod3<-glm(pinf~sex+food,data=base0,family="binomial",weights = val)
summary(mod3)
```

Logistic Regression

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	$<$0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	$<$ 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Logistic Regression

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	$<$0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	$<$ 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Start by noting the reference categories:

- The reference category for **sex** is "being a boy"
- The reference category for **type of feeding** is "only formula milk"

Logistic Regression

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	<0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	< 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Start by noting the reference categories:

- The reference category for **sex** is "being a boy"
- The reference category for **type of feeding** is "only formula milk"

Interpretation:

- $\widehat{\beta_0} = -1.613$ **-> the odds** of pulmonary infection in boys receiving only adapted milk is $e^{-1.613} = 0.20$. The probability of not having an infection is 5 times greater than the probability of having an infection.

Logistic Regression

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	<0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	< 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Start by noting the reference categories:

- The reference category for **sex** is "being a boy"
- The reference category for **type of feeding** is "only formula milk"

Interpretation:

- $\widehat{\beta_0} = -1.613$ **-> the odds** of pulmonary infection in boys receiving only adapted milk is $e^{-1.613} = 0.20$. The probability of not having an infection is 5 times greater than the probability of having an infection.
- $\hat{\beta}_{sexGirl} = -0.313 \rightarrow OR(infection|Girl\ vs\ Boy) = e^{-0.313} = 0.73$

The odds of infection among girls is 0.73 times the odds of infection among boys, for the same type of feeding. Equivalently, the odds of infection among boys is 1/0.73 = 1.37 times the odds among girls (therefore 37% higher).
Being a boy is positively associated with the infection (while being a girl is negatively associated).

Logistic Regression

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	<0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	< 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Start by noting the reference categories:

- The reference category for **sex** is "being a boy"
- The reference category for **type of feeding** is "only formula milk"

Interpretation:

- $\widehat{\beta_0} = -1.613$ **-> the odds** of pulmonary infection in boys receiving only adapted milk is $e^{-1.613} = 0.20$. The probability of not having an infection is 5 times greater than the probability of having an infection.
- $\hat{\beta}_{sexGirl} = -0.313 \rightarrow OR(infection|Girl\ vs\ Boy) = e^{-0.313} = 0.73$

The odds of infection among girls is 0.73 times the odds of infection among boys, for the same type of feeding.
Equivalently, the odds of infection among boys is 1/0.73 = 1.37 times the odds among girls (therefore 37% higher).
Being a boy is positively associated with the infection (while being a girl is negatively associated).

- $\hat{\beta}_{foodBreast} = -0.669$

The odds of infection among children being breastfed is $e^{-0.669} = 0.51$ times the odds of infection among children being fed only with formula milk.
Equivalently, in comparison with formula milk, breast-feeding reduces the odds of infection by approximately half.

Model evaluation

Goodness of fit (sometimes referred to as calibration): Comparing the observed outcome with that predicted by the model

Discrimination: Given the covariates of the model, what is the ability to classify an individual

Goodness of fit

Hosmer-Lemeshow test:

- To compare the predicted values with observed values
- ISSUE: the observations are binary, but the model predicts probabilities!
  - We calculated for each individual the P(y=1|x) predicted by the model;
  - We consider the deciles (10th, 20th,...,90th percentiles) of the probabilities predicted by the model;
  - In each decile we can calculate the number of individuals with observed event and compare it with the expected.

Goodness of fit

Hosmer-Lemeshow test:

- To compare the predicted values with observed values
- ISSUE: the observations are binary, but the model predicts probabilities!
  - We calculated for each individual the P(y=1|x) predicted by the model;
  - We consider the deciles (10th, 20th,...,90th percentiles) of the probabilities predicted by the model;
  - In each decile we can calculate the number of individuals with observed event and compare it with the expected.

```
library (glmtoolbox)
hltest(mod3)
```

Discrimination

Sensitivity = P (subject i rated 1 / y_i^{obs}=1)
Specificity = P (subject i rated 0 / y_i^{obs} =0)

But to classify a subject based on a predicted probability (continuous variable) we need a rule (a cut off):

For example: 0 if P(y=1/x)<=0.5 and 1 if P(y=1/x)>0.5

Discrimination

But...

we can consider other cutoff points as a classification criterion, instead of 0.5.

Regardless of the criteria used, a single criterion is not enough to give a correct idea of the model's ability to discriminate..

ROC curve

ROC

Sensitivity and specificity are characteristics that are difficult to balance, that is, it is difficult to increase the sensitivity and specificity of a test at the same time.

A ROC curve (receiver operator characteristic curve) is a way of representing the normally antagonistic relationship between sensitivity and specificity of a quantitative diagnostic test along a continuum of "cutoff" values.

ROC

Sensitivity and specificity are characteristics that are difficult to balance, that is, it is difficult to increase the sensitivity and specificity of a test at the same time.

A ROC curve (receiver operator characteristic curve) is a way of representing the normally antagonistic relationship between sensitivity and specificity of a quantitative diagnostic test along a continuum of "cutoff" values.

To build a ROC curve, draw a diagram that represents the sensitivity as a function of the proportion of false positives (1- Specificity) for a set of "cutoff point" values.

ROC

Sensitivity and specificity are characteristics that are difficult to balance, that is, it is difficult to increase the sensitivity and specificity of a test at the same time.

A ROC curve (receiver operator characteristic curve) is a way of representing the normally antagonistic relationship between sensitivity and specificity of a quantitative diagnostic test along a continuum of "cutoff" values.

To build a ROC curve, draw a diagram that represents the sensitivity as a function of the proportion of false positives (1- Specificity) for a set of "cutoff point" values.

```
predicted3 <- predict(mod3, base0, type="response", se.fit=TRUE)
roc3 <- roc(base0$pinf ~ predicted3$fit, plot=TRUE, print.auc=TRUE)
auc3 <- auc(base0$pinf, predicted3$fit)
ci.auc3 <- ci.auc(roc3, conf.level=0.95)
```

ROC curve – example 1

Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

99	+
81	-
75	+
60	+
50	-
40	+
35	-
32	-
10	-
2	-

ROC curve – example 1

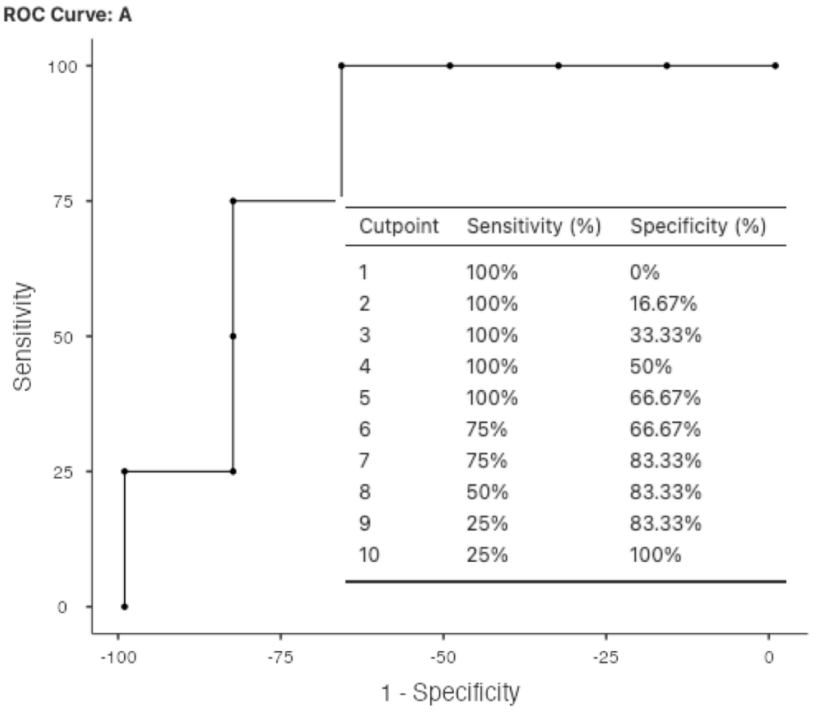
Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	+	+	+	+	+	+	+	+	+	+	+	-
81	-	+	+	+	+	+	+	+	+	+	-	-
75	+	+	+	+	+	+	+	+	+	-	-	-
60	+	+	+	+	+	+	+	+	-	-	-	-
50	-	+	+	+	+	+	+	-	-	-	-	-
40	+	+	+	+	+	+	-	-	-	-	-	-
35	-	+	+	+	+	-	-	-	-	-	-	-
32	-	+	+	+	-	-	-	-	-	-	-	-
10	-	+	+	-	-	-	-	-	-	-	-	-
2	-	+	-	-	-	-	-	-	-	-	-	-

ROC curve – example 1

Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	+	+	+	+	+	+	+	+	+	+	+	-
81	-	+	+	+	+	+	+	+	+	+	-	-
75	+	+	+	+	+	+	+	+	+	-	-	-
60	+	+	+	+	+	+	+	+	-	-	-	-
50	-	+	+	+	+	+	+	-	-	-	-	-
40	+	+	+	+	+	+	-	-	-	-	-	-
35	-	+	+	+	+	-	-	-	-	-	-	-
32	-	+	+	+	-	-	-	-	-	-	-	-
10	-	+	+	-	-	-	-	-	-	-	-	-
2	-	+	-	-	-	-	-	-	-	-	-	-



ROC curve – example 2

Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

99	+
81	+
75	+
60	+
50	+
40	-
35	-
32	-
10	-
2	-

ROC curve – example 2

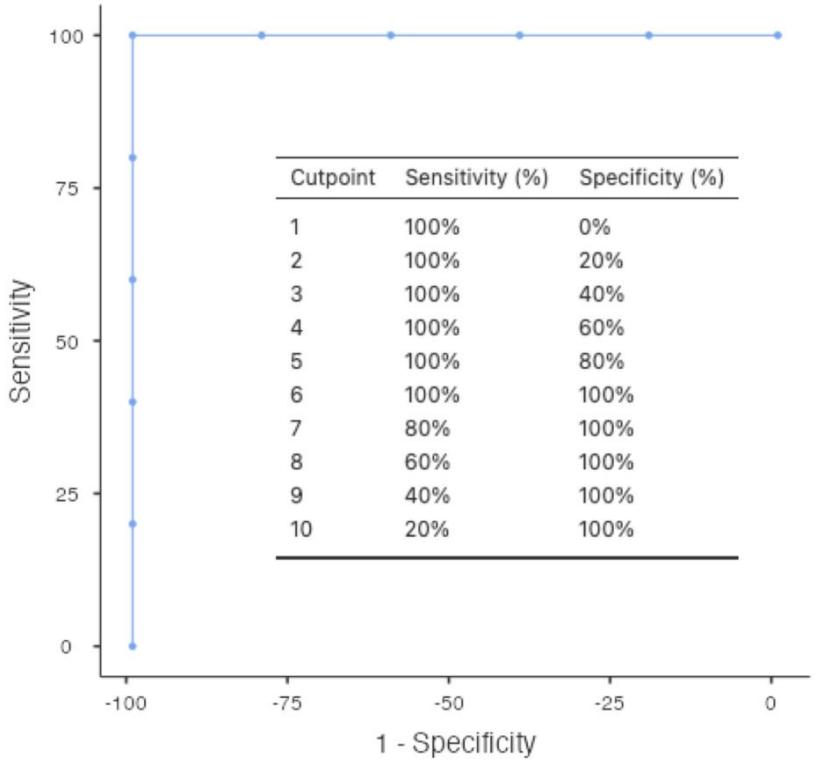
Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	+	+	+	+	+	+	+	+	+	+	+	-
81	+	+	+	+	+	+	+	+	+	+	-	-
75	+	+	+	+	+	+	+	+	+	-	-	-
60	+	+	+	+	+	+	+	-	-	-	-	-
50	+	+	+	+	+	+	-	-	-	-	-	-
40	-	+	+	+	+	+	-	-	-	-	-	-
35	-	+	+	+	+	-	-	-	-	-	-	-
32	-	+	+	+	-	-	-	-	-	-	-	-
10	-	+	+	-	-	-	-	-	-	-	-	-
2	-	+	-	-	-	-	-	-	-	-	-	-

ROC curve – example 2

Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	+	+	+	+	+	+	+	+	+	+	+	-
81	+	+	+	+	+	+	+	+	+	+	-	-
75	+	+	+	+	+	+	+	+	+	-	-	-
60	+	+	+	+	+	+	+	-	-	-	-	-
50	+	+	+	+	+	+	-	-	-	-	-	-
40	-	+	+	+	+	+	-	-	-	-	-	-
35	-	+	+	+	+	-	-	-	-	-	-	-
32	-	+	+	+	-	-	-	-	-	-	-	-
10	-	+	+	-	-	-	-	-	-	-	-	-
2	-	+	-	-	-	-	-	-	-	-	-	-



Cutpoint	Sensitivity (%)	Specificity (%)
1	100%	0%
2	100%	20%
3	100%	40%
4	100%	60%
5	100%	80%
6	100%	100%
7	80%	100%
8	60%	100%
9	40%	100%
10	20%	100%

ROC curve – example 3

Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

99	-
81	+
75	-
60	+
50	-
40	+
35	-
32	+
10	-
2	+

ROC curve – example 3

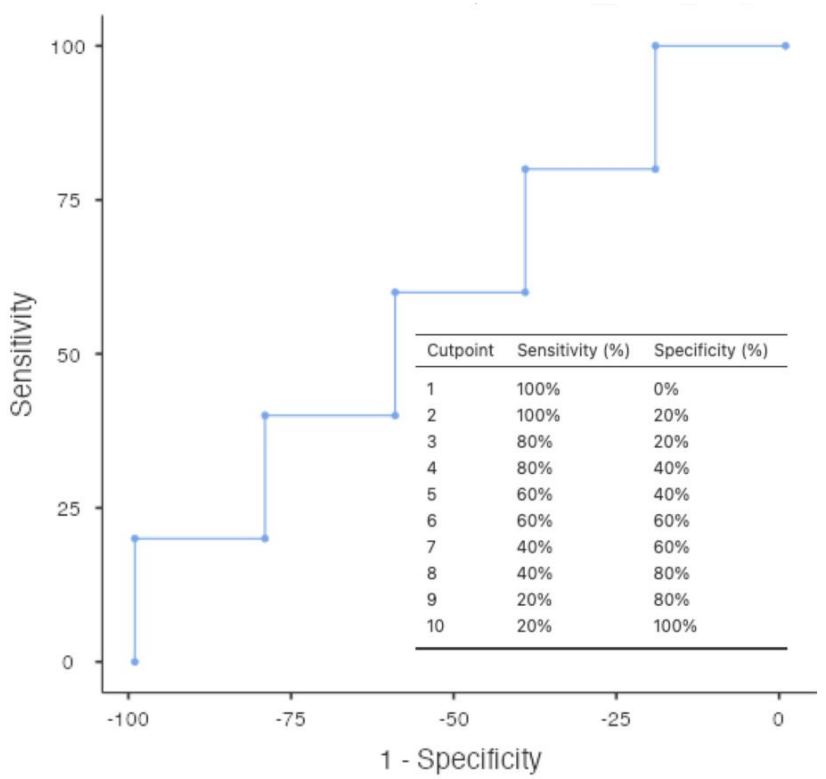
Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	-	+	+	+	+	+	+	+	+	+	+	-
81	+	+	+	+	+	+	+	+	+	+	-	-
75	-	+	+	+	+	+	+	+	+	-	-	-
60	+	+	+	+	+	+	+	+	-	-	-	-
50	-	+	+	+	+	+	+	-	-	-	-	-
40	+	+	+	+	+	+	-	-	-	-	-	-
35	-	+	+	+	+	-	-	-	-	-	-	-
32	+	+	+	+	-	-	-	-	-	-	-	-
10	-	+	+	-	-	-	-	-	-	-	-	-
2	+	+	-	-	-	-	-	-	-	-	-	-

ROC curve – example 3

Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	-	+	+	+	+	+	+	+	+	+	+	-
81	+	+	+	+	+	+	+	+	+	+	-	-
75	-	+	+	+	+	+	+	+	+	-	-	-
60	+	+	+	+	+	+	+	+	-	-	-	-
50	-	+	+	+	+	+	+	-	-	-	-	-
40	+	+	+	+	+	+	-	-	-	-	-	-
35	-	+	+	+	+	-	-	-	-	-	-	-
32	+	+	+	+	-	-	-	-	-	-	-	-
10	-	+	+	-	-	-	-	-	-	-	-	-
2	+	+	-	-	-	-	-	-	-	-	-	-



Cutpoint	Sensitivity (%)	Specificity (%)
1	100%	0%
2	100%	20%
3	80%	20%
4	80%	40%
5	60%	40%
6	60%	60%
7	40%	60%
8	40%	80%
9	20%	80%
10	20%	100%

ROC curve – example 4

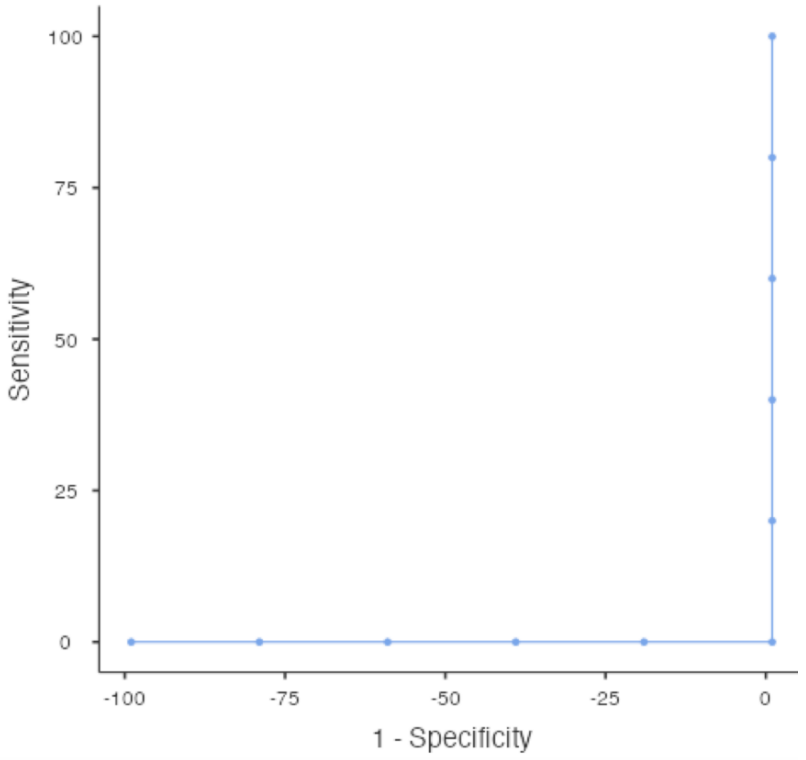
Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

99	-
81	-
75	-
60	-
50	-
40	+
35	+
32	+
10	+
2	+

ROC curve – example 4

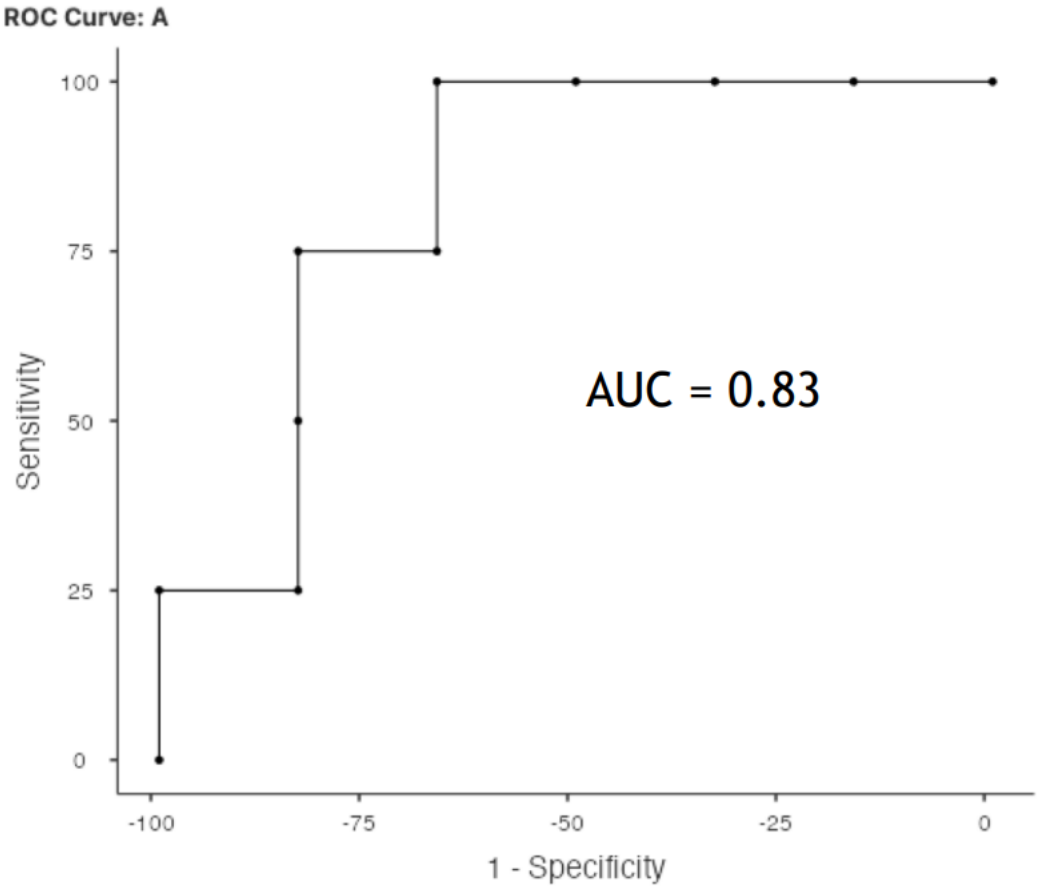
Let's suppose that we obtain with a model a probability of having a disease (left column) in these 10 individuals but you know who really has the disease (+) and who doesn't (-):

		1	2	3	4	5	6	7	8	9	10	11
99	-	+	+	+	+	+	+	+	+	+	+	-
81	-	+	+	+	+	+	+	+	+	+	-	-
75	-	+	+	+	+	+	+	+	+	-	-	-
60	-	+	+	+	+	+	+	+	-	-	-	-
50	-	+	+	+	+	+	+	-	-	-	-	-
40	+	+	+	+	+	+	-	-	-	-	-	-
35	+	+	+	+	+	-	-	-	-	-	-	-
32	+	+	+	+	-	-	-	-	-	-	-	-
10	+	+	+	-	-	-	-	-	-	-	-	-
2	+	+	-	-	-	-	-	-	-	-	-	-



Area under the curve

99	+
81	-
75	+
60	+
50	-
40	+
35	-
32	-
10	-
2	-



Area under the curve

99	+
81	+
75	+
60	+
50	+
40	-
35	-
32	-
10	-
2	-