ROC analysis, a method to assess binary decision rules

David Makowski
INRA
Outline

1. What is a binary decision rule?

2. ROC analysis, a method to assess the accuracy of binary decision rules

3. An example: assessment of decision rules for the control of sclerotinia

4. Exercise with R:
   - Assessment of models for categorizing soft wheat fields according to their grain protein content
1. What is a binary decision rule?
1. What is a binary decision rule?

Decision rule

« A rule for **taking decisions** in function of some **variables** ». 
1. What is a binary decision rule?

**Binary decision rule**

« A rule for choosing among two decisions ». 

Variables → Decision rule → Decision 1 or Decision 2
1. What is a binary decision rule?

Examples of binary decision rules

- Apply a chemical treatment / No chemical treatment
- Sow cultivar 1 / Sow cultivar 2
- Apply fertilizer / No application

...
1. What is a binary decision rule?

Binary decision rule based on an indicator and a decision threshold

« I apply fungicide if the indicator is higher than a decision threshold »

« I apply fungicide if \( I \geq S \). No application otherwise »

• Indicator \( I = \) measure or model prediction (ex: % diseased organs).

• Threshold \( S = \) Numerical value (ex: 20%).
1. What is a binary decision rule?

Optimization of

« I apply fungicide if \( I \geq S \). No application otherwise »

Two practical problems:

• Choose the best threshold \( S \) for a given indicator \( I \).

• Choose the best indicator among several candidates.
A framework for assessing binary decision rules

1. Define a series of indicators (measured variables and/or models).

2. Define the range of variation for the threshold $S$ associated to each indicator (e.g. 0-100 % of diseased flowers).

3. Define one or several criteria for assessing the decision rules (i.e. the combinations of all possible $I$ and $S$).

4. Estimate the values of the criteria for each rule.

5. Choose the « best » rule.
2. ROC analysis

ROC = Receiver Operating Characteristic
2. ROC analysis

**ROC analysis**

**Notations**

$Y$: a random variable taking the value 0 or 1 for a negative and positive response respectively.

$I$: a variable corresponding to the output of a given indicator.

$S$: a decision threshold.

**Examples for $Y$**

$Y = 0$ if the yield loss due to the disease is small, $Y=1$ otherwise.

$Y = 0$ if the percentage of diseased plants at harvest $< 10\%$, $Y=1$ otherwise.

$Y = 0$ if weed biomass $< 0.15$ t/ha, $Y=1$ otherwise.
2. ROC analysis

**ROC analysis**

*n* plots with \( Y=0 \) (e.g. % diseased plants at harvest < 10%).

\( m \) plots with \( Y=1 \) (e.g. % diseased plants at harvest \( \geq \) 10%).

(i). Determine the value of the indicator \( I \) for each plot.

(ii). Define a decision threshold \( S \).

(iii). **Sensitivity** = \( \text{Prob}(I \geq S \mid Y=1) = 1 – \text{False negative rate} \)

(iv). **Specificity** = \( \text{Prob}(I < S \mid Y=0) = 1 – \text{False positive rate} \)

(v). **ROC curve**: Sensitivity (S) versus 1 − Specificity (S)

(vi). Estimate the area under the ROC curve (**AUC**) for each indicator \( I \).

If AUC~0.5, the indicator is not useful (not better than random decisions).
2. ROC analysis

ROC analysis

Number of plots

Value of $I$

Threshold $S$

Plots with high incidence at harvest

Plots with low incidence at harvest

Specificity

Sensitivity
2. ROC analysis

![ROC Curve Diagram]

- **Sensitivity**
- **Specificity**
- **Optimal values**
- **AREA**
2. ROC analysis

Area Under the Curve = Probability of correctly ranking two plots

Optimal values

1 - Specificity

Sensitivity
3. An example: assessment of decision rules for the control of sclerotinia
3. An example

*Sclerotinia sclerotiorum*, Lib., de Bary, in oilseed rape crops

- High variability of disease incidence across sites and years.
- High yield losses if disease incidence at harvest > 10%.
- Efficient chemical treatments exist, but are not always required.
Rule 1. Indicator $I_1 = \text{measured proportion of diseased plant organs}$

3. An example

Agricultural field $\rightarrow n$ plant organs are collected before treatment.

Organs are scored for the presence of the disease.

$y$ organs are diseased.

If $y/n < T$, no treatment

If $y/n \geq T$, treatment
3. An example

In this example, organs = flowers

Sowing  Flowering  Harvest

\[ n \text{ collected flowers} \]
\[ \text{Incubation in Petri dishes} \]
\[ y \text{ diseased flowers} \]

If \[ \frac{y}{n} \geq T \] treatment
else no treatment
3. An example

**Rule 2. Indicator $I_2 = \text{sum of risk points}$**

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Level</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of oil-seed crops during the last ten years</td>
<td>&gt;5</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>3-5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2-3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other host crops during the last five years</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>Level of infection in the last crop</td>
<td>High</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>0</td>
</tr>
<tr>
<td>Type of field</td>
<td>Wet</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Dry</td>
<td>0</td>
</tr>
<tr>
<td>Plant density</td>
<td>High</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>0</td>
</tr>
<tr>
<td>Rain in the last month before flowering</td>
<td>More than normal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Normal (50-60 mm)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Less than normal</td>
<td>0</td>
</tr>
</tbody>
</table>
3. An example

**Rule 3. Indicator** $I_3 = \text{output of a logistic model}$

- % diseased flowers at flowering
- Sum of risk points at flowering

**Probability that the disease incidence at harvest is higher than 10%**
3. An example

**Rule 3. Indicator \( l_3 = \text{output of a logistic model} \)**

Logistic model

\[
z = \frac{\exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}{1 + \exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}
\]
3. An example

Three rules for deciding about a chemical treatment at flowering

If $I_1 \geq S$, a treatment is recommended

If $I_2 \geq S$, a treatment is recommended

If $I_3 \geq S$, a treatment is recommended

Which rule is the best?
Two types of error

Type 1. False positive rate = 1 - Specificity

\( I \geq S \) (a treatment was recommended)

but \( \% \text{ diseased plants at harvest} < 10\% \)
(a treatment was not required)

Type 2. False negative rate = 1 – Sensitivity

\( I < S \) (a treatment was not recommended)

but \( \% \text{ diseased plants at harvest} \geq 10\% \)
(a treatment was required)
3. An example

Data from 85 experimental plots in France

Low disease incidence at harvest

High disease incidence at harvest

Fraction of diseased flowers

Risk point sum
3. An example

R code for fitting the logistic model

```r
TAB <- read.table("f:\ Exemples\Sclero0203.txt",header=T,sep="\t")
TAB <- TAB[!is.na(TAB[,1])]
Ind.1 <- TAB$KIT
Incidence.t <- 0.10
Incidence <- TAB$TxAttNT
Incidence[Incidence < Incidence.t] <- 0
Incidence[Incidence >= Incidence.t] <- 1
Fit <- glm(Incidence ~ Ind.1 + Ind.2, family = binomial)
```

# glm = R function for fitting generalized linear models (e.g logistic, Poisson)
3. An example

```r
> print(summary(Fit))

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.31581    1.20336  -3.586 0.000335 ***
Ind.1        5.27346    1.21518   4.340 1.43e-05 ***
Ind.2        0.01356    0.01800   0.753 0.451329

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
3. An example

**R code for ROC analysis for rule 1**

```r
library(ROCR)
pred<-prediction(Ind.1,Incidence)
perf<-performance(pred,"sens","spec")
spec.1<-perf@"x.values"[[1]]
sens.1<-perf@"y.values"[[1]]
plot(spec.1,sens.1, ylab="Sensibilité", xlab="Spécificité", type="l",lty=1,lwd=3)
abline(1,-1)
```
3. An example

**ROC curves for rules 1 and 2**

![Graph showing ROC curves with labeled axes: Sensitivity on the y-axis and Specificity on the x-axis. Two curves are depicted: one labeled % diseased flowers and the other labeled Risk point sum.](image)
3. An example

**R code for ROC analysis with cross validation**

```
Pred.cv<-NA
for (i in (1:length(TAB[,1]))) {
    TAB.est.i<-data.frame(Ind.1[-i],Ind.2[-i],Incidence[-i])
    TAB.pred.i<-c(Ind.1[i],Ind.2[i])
    Fit.cv<-glm(TAB.est.i$Incidence~TAB.est.i[,1]+TAB.est.i[,2],family=binomial,data=TAB.est.i)
    Para<-as.vector(Fit.cv$coefficients)
    Pred.i<-exp(Para[1]+Para[2]*TAB.pred.i[1]+Para[3]*TAB.pred.i[2])/(1+
    Pred.cv<-c(Pred.cv, Pred.i)
}
pred<-prediction(Pred.cv[-1],Incidence)
perf<-performance(pred,"sens","spec")
```
3. An example

ROC curves for rule 3

without cross validation
with cross validation
3. An example

**Area under the ROC curves**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>% diseased flowers</td>
<td>0.88</td>
</tr>
<tr>
<td>Point sum</td>
<td>0.62</td>
</tr>
<tr>
<td>Logistic (without cross validation)</td>
<td>0.87</td>
</tr>
<tr>
<td>Logistic (with cross validation)</td>
<td>0.85</td>
</tr>
</tbody>
</table>
5. Exercice with R:

Assessment of models for categorizing soft wheat fields according to their grain protein content

Four candidate indicators:

i. Transmitance
ii. Nitrogen nutrition index
iii. Model 1 (dynamic crop model)
iv. Model 2 (static crop model including two input variables)

Objective: Identify plots with high grain protein content (>11.5%)

Which indicator is the best? What is its optimal decision threshold?
library(ROCR)

# Read an external data file
TAB<-read.table("f:\David\Enseignements\FormationPologne\dataAgralys.txt",header=T,sep="\t")
print(TAB)

# Grain protein threshold
GPC.t<-11.5

# Variable of reference (binary variable Y)
GPC<-TAB$Protein
GPC[GPC<GPC.t]<-0
GPC[GPC>=GPC.t]<-1

# Indicators
Ind.1<-TAB$SPAD
Ind.2<-TAB$NNI
Ind.3<-TAB$Model_1
Ind.4<-TAB$Model_2

# Some graphs
par(mfrow=c(2,2))
hist(Ind.1[GPC==0], xlab="SPAD", main="Low grain protein content")
hist(Ind.1[GPC==1], xlab="SPAD", main="High grain protein content")
hist(Ind.2[GPC==0], xlab="NNI", main="Low grain protein content ")
hist(Ind.2[GPC==1], xlab="NNI", main="High grain protein content ")
### ROC analysis for Ind.1###

```r
pred <- prediction(Ind.1, GPC)
perf <- performance(pred, "auc")

# Area under the ROC curve
auc.1 <- perf@"y.values"
print("AUC for Indicator 1")
print(auc.1)

# Sensitivity and specificity
perf <- performance(pred, "sens", "spec")
print(perf)
spec.1 <- perf@"x.values"[[1]]
sens.1 <- perf@"y.values"[[1]]

# ROC curve
plot(spec.1, sens.1, ylab = "Sensitivity", xlab = "Specificity", type = "l", lty = 1, lwd = 3)
abline(1, -1)

# Threshold
print(perf@"alpha.values"[[1]][spec.1 > 0.65 & sens.1 > 0.65])
```
#Logistic regressions
#Combination of Ind.1 and Ind.2
Fit<-glm(GPC~Ind.1+Ind.2,family=binomial)

print("Combination of Ind.1 and Ind.2")
print(summary(Fit))

print("ROC analysis for the combinations of indicators without cross-validation")

#ROC analysis for Ind.1 + Ind.2

pred<-prediction(Fit$fitted.values,GPC)
perf<-performance(pred,"auc")
auc<-perf@"y.values"
print("AUC for Ind.1 + Ind.2")
print(auc)
perf<-performance(pred,"sens","spec")
spec.1<-perf@"x.values"[[1]]
sens.1<-perf@"y.values"[[1]]
plot(spec.1,sens.1, ylab="Sensitivity", xlab="Specificity", type="l",lty=1,lwd=3)
abline(1,-1)
print("ROC analysis for the combinations of indicators with cross-validation")

#Initialization

Pred.cv<-NA

for (i in (1:length(TAB[,1]))) {

#New tables
TAB.est.i<-data.frame(Ind.1[-i],Ind.2[-i],GPC[-i])
TAB.pred.i<-c(Ind.1[i],Ind.2[i])

#Combination of Ind.1 and Ind.2
Fit.cv<-glm(TAB.est.i$GPC~TAB.est.i[,1]+TAB.est.i[,2],family=binomial,data=TAB.est.i)
Para<-as.vector(Fit.cv$coefficients)
Pred.cv<-c(Pred.cv, Pred.i)
}

Specificity
Sensitivity

Specificity
Sensitivity

Specificity
Sensitivity

Specificity
Sensitivity

Indicator | AUC
---|---
SPAD | 0.77
NNI | 0.84
Model 1 | 0.46
Model 2 | 0.62
SPAD + NNI | 0.86 (0.90)


