Classification Assessment - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

Model Fitness in Logistic Regression

Boston University – CS555, Data Analysis, F. Alizadeh-Shabdiz

skyhook®

Predictions of Logistic Regression

▷ **In the logistic regression setting, the observed response takes on one of two values (0 = no event and 1= event).**

▷ **The predicted probabilities take on values between 0 and 1.**

If the predicted probabilities for someone who had the event are close to or equal to 1, then we would say that the model was successful in predicting that individual's risk.

How to quantify the fitness in Logistic Regression?

For example, let's say that all predicted probabilities are **0.32 or 0.72**.

▷ What would we say about the fit of the model if all the events had a predicted probability of **0.72** and all the non-events had a predicted probability of **0.32**?
  We'd like be inclined to think that the model was successful in discriminating between those with and without the event.

▷ What if most of those with a predicted probability of **0.32** did not have the event and most of those with a predicted probability of **0.72** did?
  We'd still be inclined to say that the model fit was good.

How we define "most" may vary and it becomes a little more difficult to assess without having a measure to quantify the fit.

Cutoffs Problem - An Example

We are interested in the association between gender and risk of having a coronary event in a high-risk patient population (who have had an event in the past).

We conduct a simple logistic regression model with gender (specifically a dummy variable for male versus female) as the only explanatory variable.

From the regression model predicted risk for a male: 64% and for a female 24%

If we have a cutoff of 0% or 100% then

  ▷ we predicted that those with a predicted probabilities of bigger than 0% will have the event and assign them predicted outcomes of **1 ($\hat{y} = 1$)** and

  ▷ we predicted that those with a predicted probabilities of $\leq 100\%$ will not have the event and assign them predicted outcomes of **0 ($\hat{y} = 0$)**.

Then we could assess the fit of the predicted values versus the observed values.

The cutoff can be different.

True/False Positive and True/False Negative

▷ **True positive.** Rate of **Correct** identification of **Positive** category.
Example: Sick people correctly identified as sick

▷ **False positive.** Rate of **Incorrect** identification of **Positive** category.
Example: Healthy people incorrectly identified as sick

▷ **True negative.** Rate of **Correct** identification of **Negative** category.
Example: Healthy people correctly identified as healthy

▷ **False negative.** Rate of **Incorrect** identification of **Negative** category.
Example: Sick people incorrectly identified as healthy

In general, Positive = identified and negative = rejected.

Therefore:

▷ **True positive = correctly identified**
▷ **False positive = incorrectly identified**
▷ **True negative = correctly rejected**
▷ **False negative = incorrectly rejected**
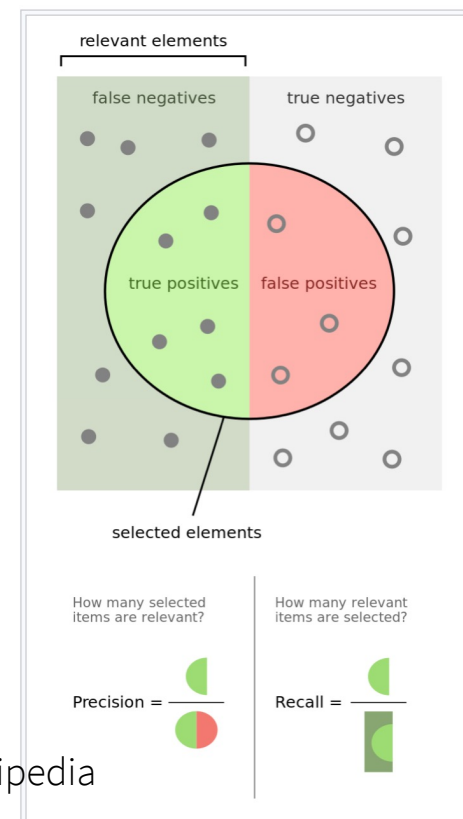
Reference: Wikipedia

Example

▷ We have identified 30 sick people **correctly identified** as sick.

▷ We have identified 4 healthy people **incorrectly identified** as sick.

▷ We have identified 25 healthy people **correctly identified** as healthy.

▷ We have identified 3 sick people **incorrectly identified** as healthy.

Example

▷ We have identified 30 sick people **correctly identified** as sick.

▷ We have identified 4 healthy people **incorrectly identified** as sick.

▷ We have identified 25 healthy people **correctly identified** as healthy.

▷ We have identified 3 sick people **incorrectly identified** as healthy.

▷ **True Positive:** 30

▷ **False Positive:** 4

▷ **True Negative:** 25

▷ **False Negative:** 3

Boston University – CS555, Data Analysis, F. Alizadeh-Shabdiz

Example - Confusion Matrix

▷ We have identified 30 sick people **correctly identified** as sick.

▷ We have identified 4 healthy people **incorrectly identified** as sick.

▷ We have identified 25 healthy people **correctly identified** as healthy.

▷ We have identified 3 sick people **incorrectly identified** as healthy.

▷ **True Positive (TP):** 30

▷ **False Positive (FP):** 4

▷ **True Negative (TN):** 25

▷ **False Negative (FN):** 3

		True Conditions	
		Condition Positive	Condition Negative
Predicted Class	Positive	TP=30	FP=4
	Negative	FN=3	TN=25

Boston University – CS555, Data Analysis, F. Alizadeh-Shabdiz

Sensitivity & Specificity

Sensitivity (also called Recall, Hit Rate, or true positive rate (TPR)) **measures the proportion of actual positives** that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).

$$\text{Sensitivity or Recall} = \frac{TP}{TP + FN}$$

Specificity (also called Selectivity or true negative rate (TNR)) **measures the proportion of actual negatives** that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision

Precision (also called positive predictive value) is the **fraction of relevant instances among the retrieved instances**.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Boston University – CS555, Data Analysis, F. Alizadeh-Shabdiz

F-Measure, F1 Measure

F-Measure provides a way to **combine Precision and Recall** as a measure of the overall effectiveness of a Classification algorithm.

F-Measure is calculated as a ratio of the weighted importance of Precision and Recall. The general formula for **positive real β number** is:

$$F_\beta = (1 + \beta^2) * \frac{Precision \times Recall}{(\beta \times Precision) + Recall}$$

"F Measure measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" (Rijsbergen et al.)

Using Precision and Recall Formula:

$$F_\beta = \frac{(1 + \beta^2) \times TP}{(1 + \beta^2) \times TP + \beta^2 \times FN + FP}$$

SKYHOOK®

F1 Measure, F2, F0.5

F1 Score: A measure that combines precision and recall is the **harmonic mean** of precision and recall.

$$\text{F1 Score} = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

▷ F_2 weighs recall higher than precision (by placing **more emphasis on false negatives**)

▷ $F_{0.5}$ weighs recall lower than precision (by attenuating the **influence of false negatives**)

SKYHOOK®

Example

- ▷ **True Positive:** 30
- ▷ **False Positive:** 4
- ▷ **True Negative:** 25
- ▷ **False Negative:** 3

- ▷ What is the Sensitivity or Recall for this example?
- ▷ What is the Specificity for this example?
- ▷ What is the F1-Score for this example?

Boston University – CS555, Data Analysis, F. Alizadeh-Shabdiz

Example

▷ **True Positive (TP):** 30

▷ **False Positive (FP):** 4

▷ **True Negative (TN):** 25

▷ **False Negative (FN):** 3

$$\text{Sensitivity or Recall} = \frac{TP}{TP+FN} = \frac{30}{30+3} = 0.9091 \text{ or } 90.9\%$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{25}{25+4} = 0.8621 \text{ or } 86.21\%$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{30}{30+4} = 0.8823 \text{ or } 90.9\%$$

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.88 \times 0.91}{0.88 + 0.91} = 0.89$$

SKYHOOK®

Receiver Operating Characteristic Curve (ROC Curve)

Boston University – CS555, Data Analysis, F. Alizadeh-Shabdiz

ROC (Receiver Operating Characteristic)

Sensitivity and specificity are important measure in assessing the fit of a logistic regression model.

Models with one or more continuous explanatory or independent variables have more possible values for the predicted probabilities and therefore **there are often many cutoffs that produce distinct values of sensitivity and specificity**.

The area under the **ROC (receiver operating characteristic) curve (also known as the c-statistic)** is a measure of the sensitivity and specificity across the range of all possible cutoffs.

**It is used to measure the goodness of fit of a logistic regression model.**

ROC (Receiver Operating Characteristic)

▷ The ROC curve is a plot of corresponding pairs of **sensitivity (y-axis)** and **1 minus the specificity (x-axis)** **for each possible cutoff point**.

▷ **It ranges between 0.5 and 1.0 with larger values indicating better fit.**

▷ **When the area under the curve is equal to 0.50**, it is said that the model **does no better at classifying events than at random or by chance**.
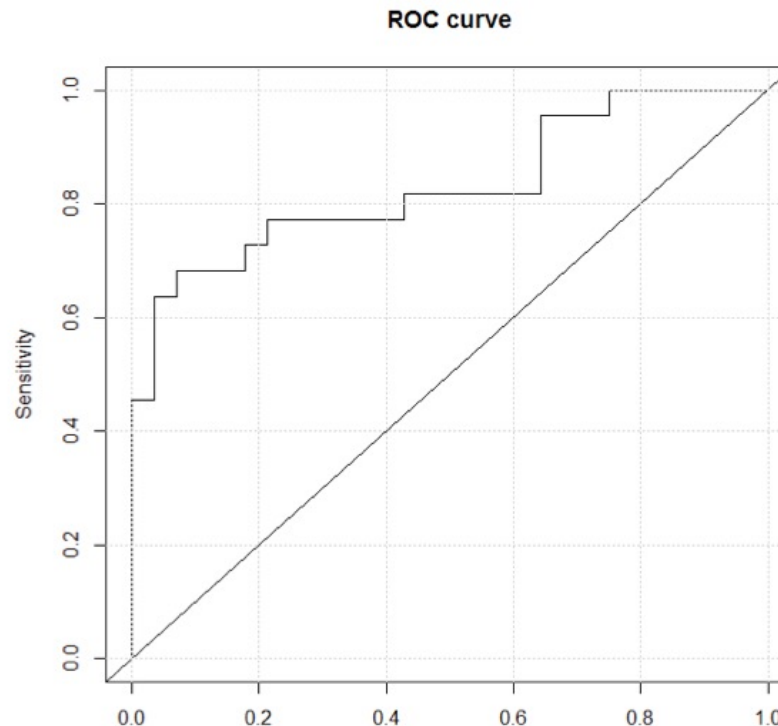
ROC curve

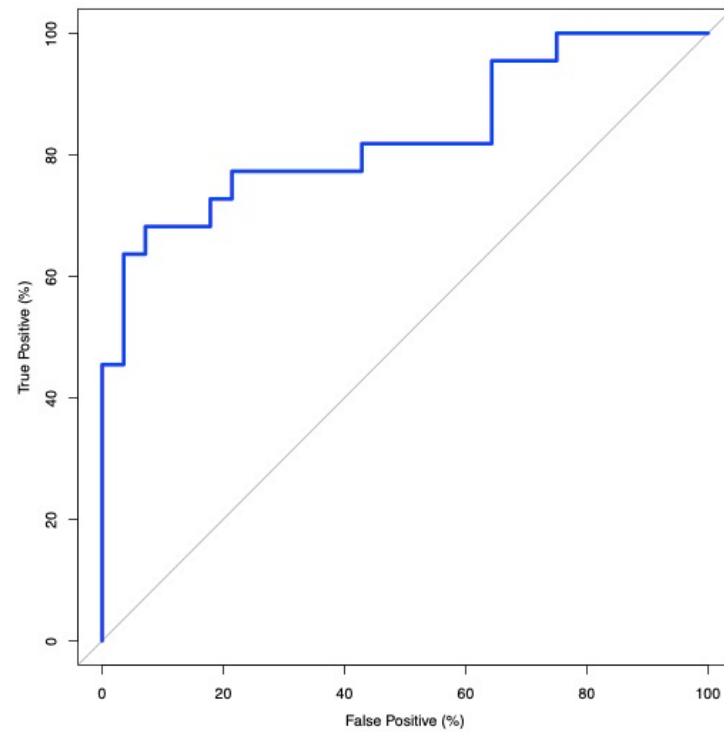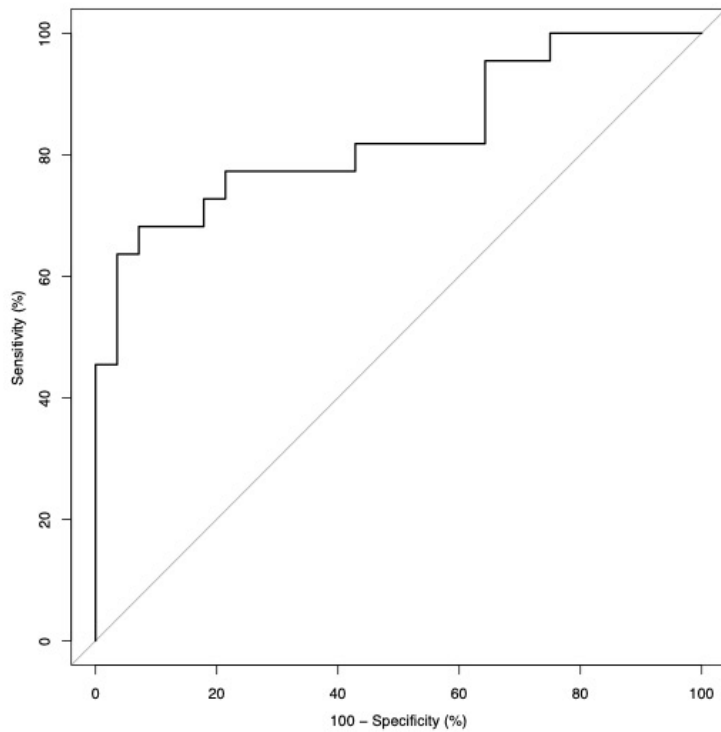# Area under the ROC curve: R commands

```
# ROC curve - Use the roc() function (from "pROC" package)
> library(pROC)
> data$prob <- predict(m, type=c("response"))
> g <- roc(data$event ~ data$prob)
> plot(g)
> plot(1-g$specificities, g$sensitivities, type="l", xlab="1-specificity
    ", ylab="Sensitivity", main="ROC  curve")
> abline(a=0. b=1)
```
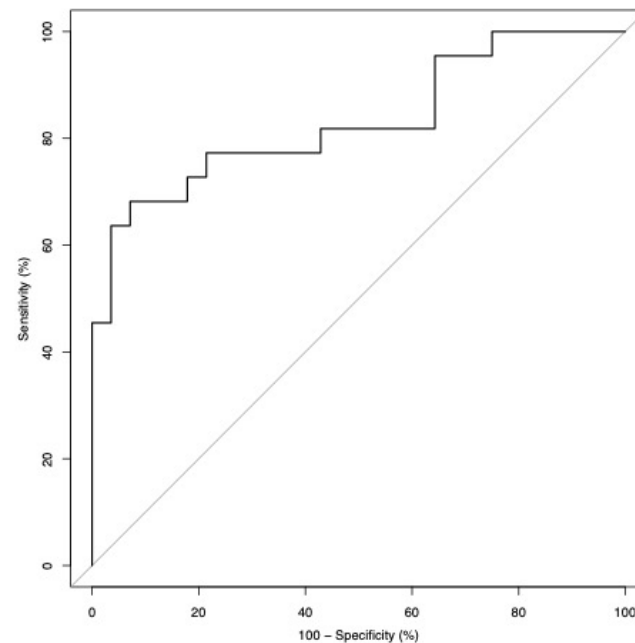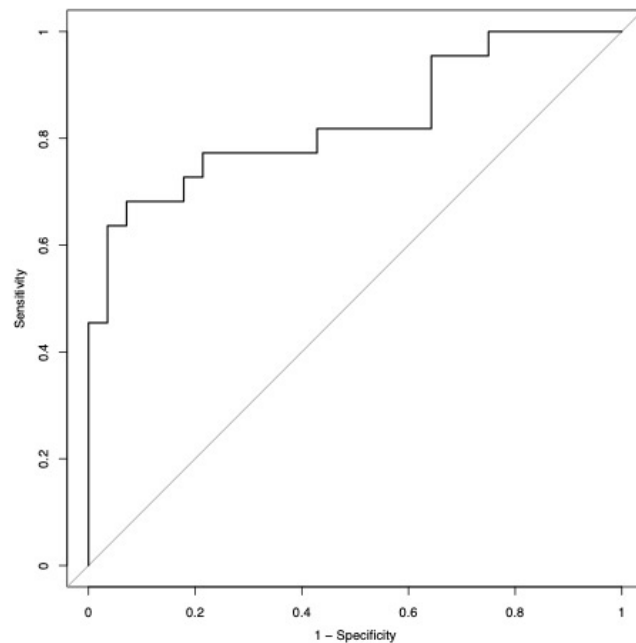
ROC - False Positive Rate, True Positive Rate Rate

Multiple Logistic Regression and Area under ROC Curve

```r
# Multiple logistic regression

> data$male <- ifelse(data$sex =="M", 1, 0)

> m2<- glm(data$event~data$chol + data$male + data$age, family=binomial)

> summary(m2)
```

Multiple Logistic Regression and Area under ROC Curve

```
> m2<- glm(data$event~data$chol + data$male + data$age, family=binomial)

# ORs per 1 unit increase
> exp(cbind(OR = coef(m2), confnt.default(m2)))

# ROC curve # install.package("pROC")
> library(pROC)

# using model with chol and sex and age
> data$prob <-predict(m2, type=c("response"))

# Build a ROC curve
> g <- roc(data$event ~ data$prob)

# see the results - c-statistics value
> print(g)

# plot the ROC Curve
> plot(g)
```