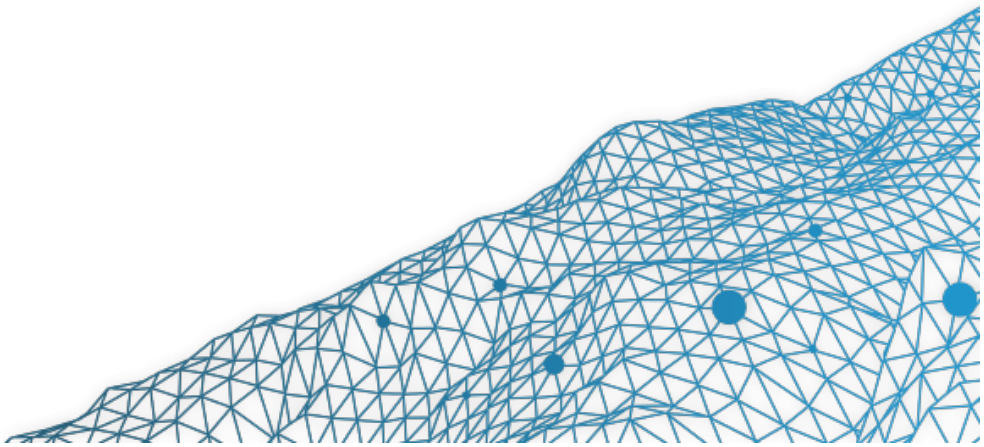
8 Classification & Goodness of Fit (Practical)

Arthur Charpentier (Université du Québec à Montréal)

Machine Learning & Econometrics

SIDE Summer School - July 2019



Test and Decision

decision \ truth	-	+
-	true negative	false negative
+	false positive	true positive

decision \ truth	-	+
-	good decision	type 2 error
+	type 1 error	true positive

We usually have a tradeoff between the two types of error, see base rate fallacy

In statistical terminology, we want to test an assumption (H_0) - which can be valid, or not - and we need to take a decision : *reject* H_0 or *accept* H_0.

Test and Decision

$$\text{Prevalence } \frac{200}{10,000} = 2\%$$

$$\text{Specificity } \frac{9,751}{9,800} = 99.5\%$$

$$\text{Sensitivity } \frac{100}{200} = 50\%$$

$$\text{Positive Predictive Value } \frac{100}{149} \sim 67\%$$

		- non-disease	+ disease	
decision	-	9,751	100	9,851
	+	49	100	149
		9,800	200	10,000

$$\text{Specificity } \frac{9,310}{9,800} = 95\%$$

$$\text{Positive Predictive Value } \frac{100}{590} \sim 17\%$$

		non-disease	disease	
decision	-	9,310	100	9,410
	+	490	100	590
		9,800	200	10,000

see Wainer & Savage (2008, Until proven guilty: False positives and war on terror).

Goodness of Fit: ROC Curve

Confusion matrix

Given a sample $(y_i, \boldsymbol{x}_i)$ and a model m, the confusion matrix is the contingency table, with dimensions *observed* $y_i \in \{0, 1\}$ and *predicted* $\widehat{y}_i \in \{0, 1\}$.

		y	
		0(-)	1(+)
$\widehat{y}$	0(-)	TN	FN
	1(+)	FP	TP
		$\frac{\text{FP}}{\text{TN+FP}}$	$\frac{\text{TP}}{\text{FN+TP}}$
		FPR	TPR

Classical measures are

true positive rate (TPR) - or sensitivity
false positive rate (FPR) - or fall-out,
true negative rate (TNR) - or specificity

TNR = 1-FPR
among others (see wikipedia)

See `ROCR::performance(prediction(Score,Y),"tpr","fpr")`

Goodness of Fit: ROC Curve

ROC (Receiver Operating Characteristic) Curve

Assume that m_t is define from a score function s, with $m_t(\boldsymbol{x}) = \mathbf{1}(s(\boldsymbol{x}) > t)$ for some threshold t. The ROC curve is the curve $(\text{FPR}_t, \text{TPR}_t)$ obtained from confusion matrices of m_t's.

$n = 100$ individuals
50 $y_i = 0$ and 50 $y_i = 1$
(well balanced)

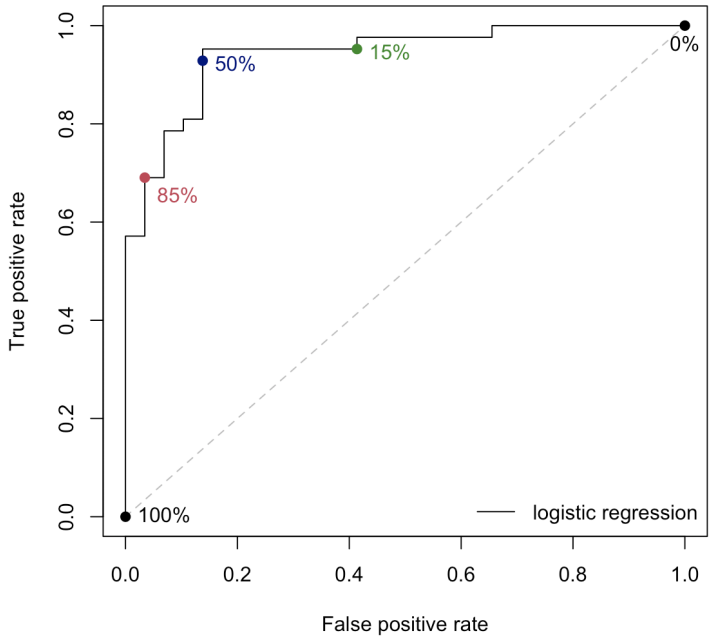
Goodness of Fit: ROC Curve

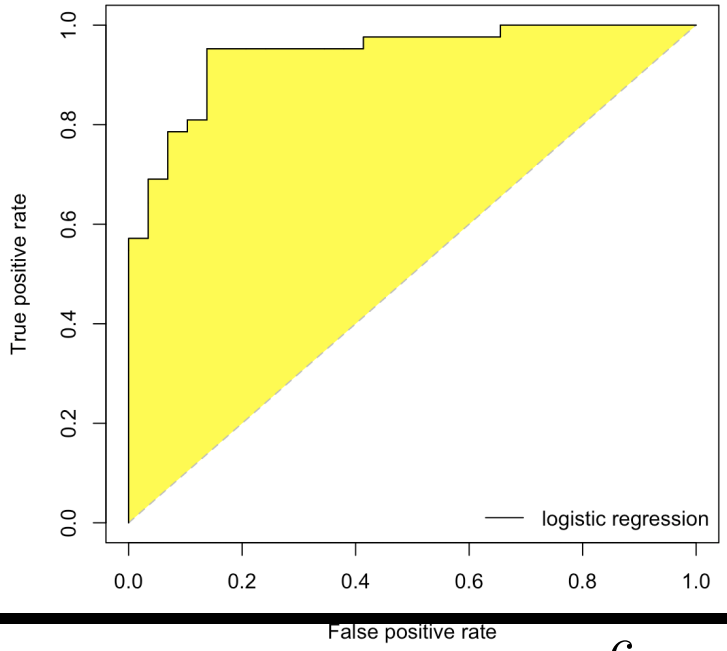
29 deaths $(y = 0)$ and 42 survivals $(y = 1)$

$$\widehat{y} = \begin{cases} 1 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] > 0\% \\ 0 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] \leq 0\% \end{cases}$$

0%

		y	
		0	1
$\widehat{y}$	0	0	0
	1	29	42
		$\frac{29}{29+0}$	$\frac{42}{42+0}$
		$= 100\%$	$= 100\%$
		FPR	TPR





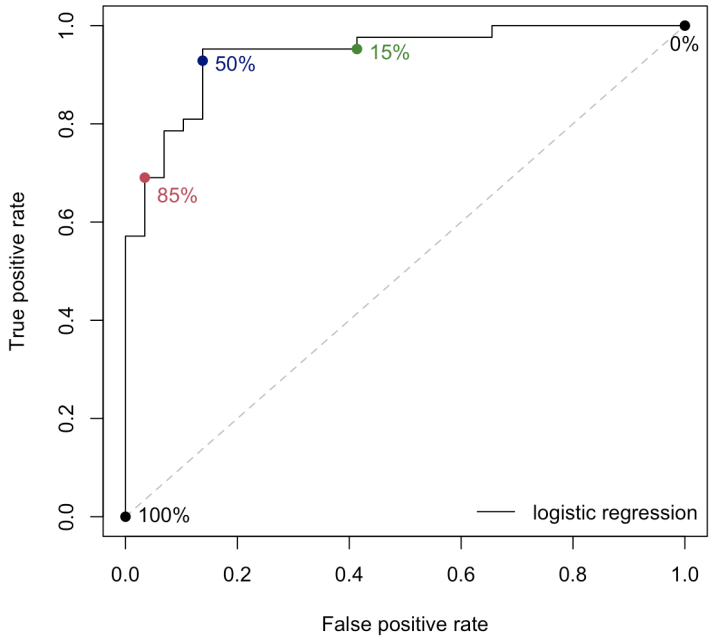
Goodness of Fit: ROC Curve

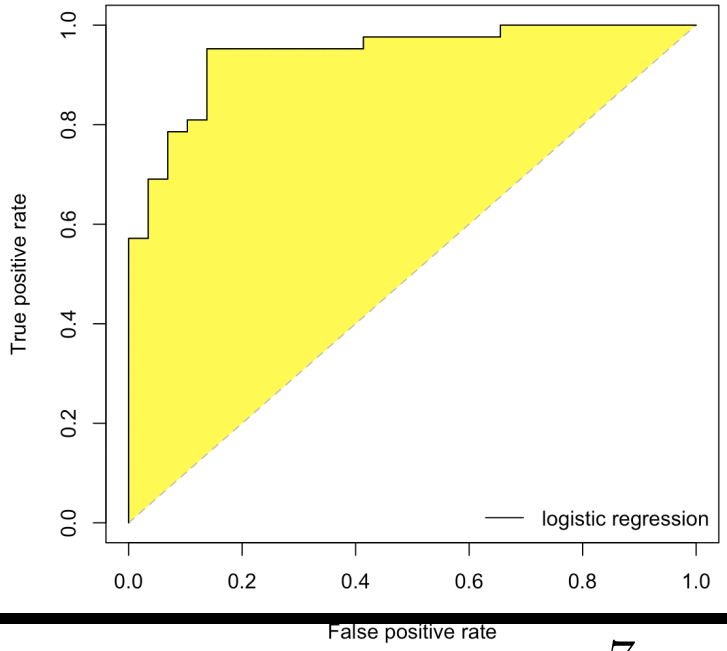
29 deaths $(y = 0)$ and 42 survivals $(y = 1)$

$$\widehat{y} = \begin{cases} 1 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] > 15\% \\ 0 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] \leq 15\% \end{cases}$$

15%

		y	
		0	1
$\widehat{y}$	0	17	2
	1	12	40
		$\frac{12}{17+12}$	$\frac{40}{42+2}$
		$\sim 41.4\%$	$\sim 95.2\%$
		FPR	TPR





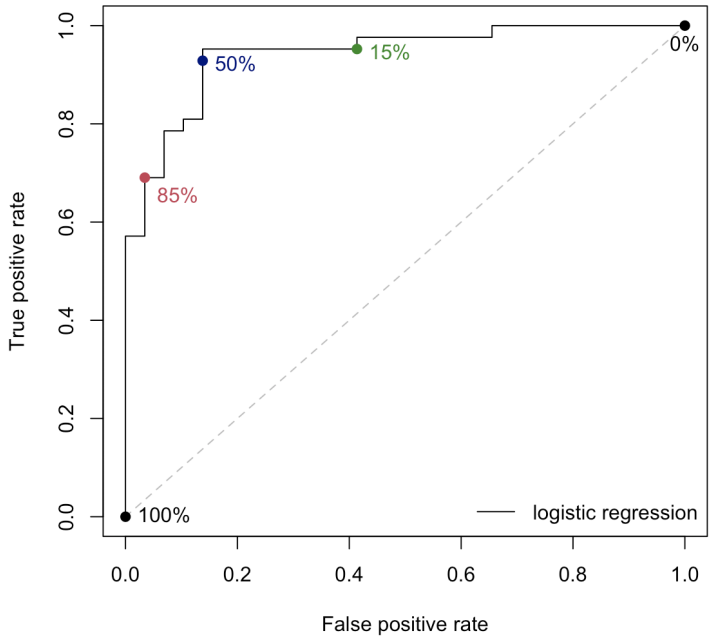
Goodness of Fit: ROC Curve

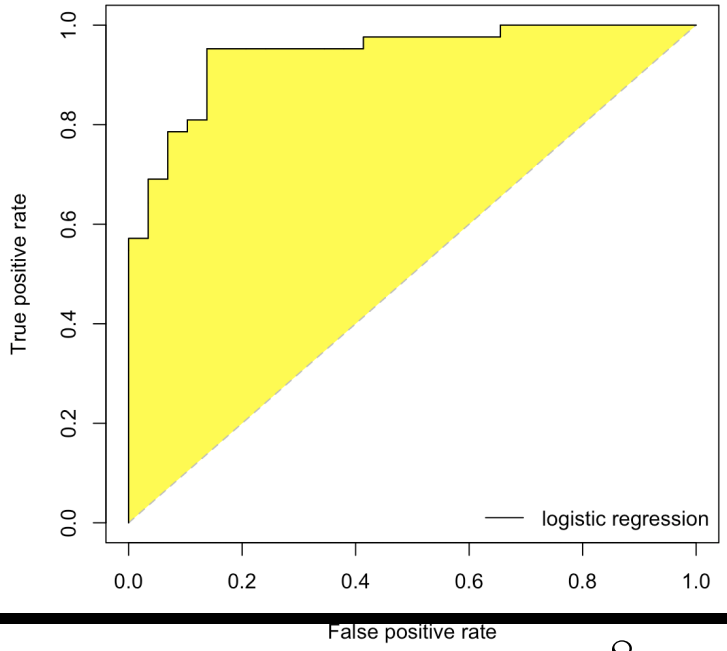
29 deaths $(y = 0)$ and 42 survivals $(y = 1)$

$$\widehat{y} = \begin{cases} 1 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] > 50\% \\ 0 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] \leq 50\% \end{cases}$$

50%

		y	
		0	1
$\widehat{y}$	0	25	3
	1	4	39
		$\frac{4}{25+4}$	$\frac{39}{39+3}$
		$\sim 13.8\%$	$\sim 92.8\%$
		FPR	TPR





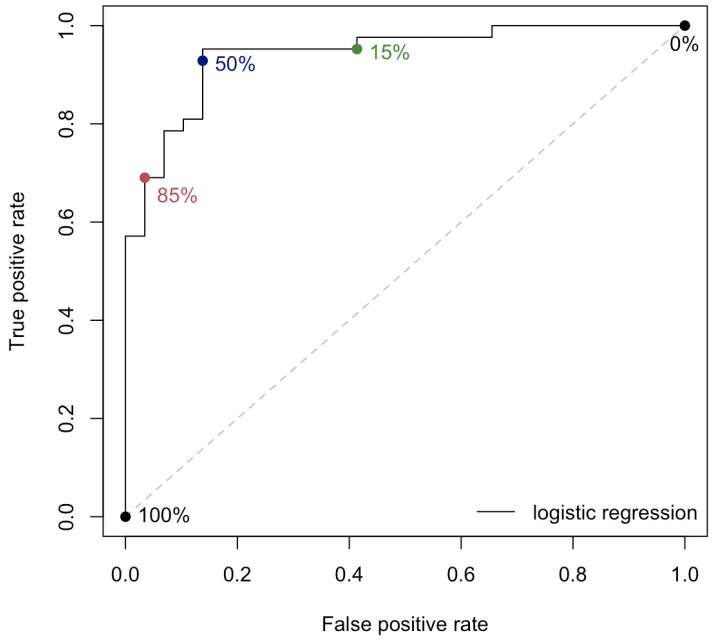
Goodness of Fit: ROC Curve

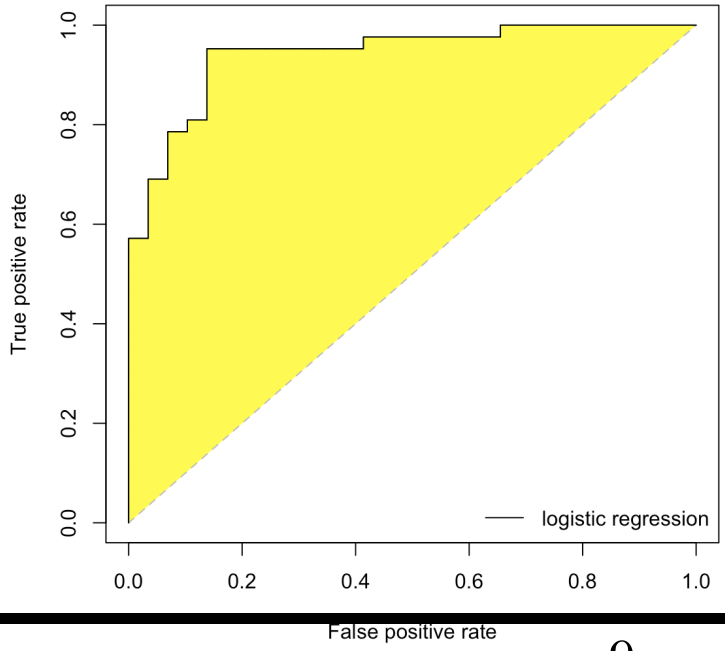
29 deaths $(y = 0)$ and 42 survivals $(y = 1)$

$$\widehat{y} = \begin{cases} 1 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] > 85\% \\ 0 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] \leq 85\% \end{cases}$$

50%

		y	
		0	1
$\widehat{y}$	0	28	13
	1	1	29
		$\frac{1}{28+1}$	$\frac{29}{29+13}$
		$\sim 3.4\%$	$\sim 69.9\%$
		FPR	TPR





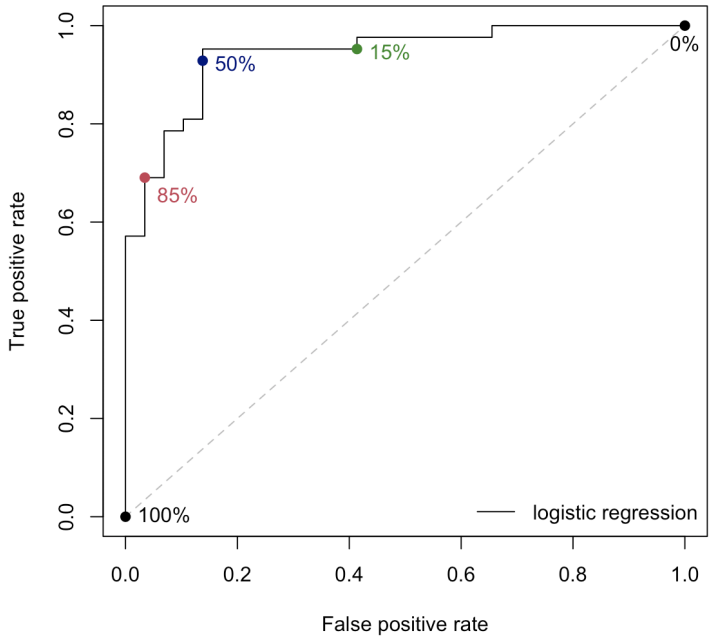
Goodness of Fit: ROC Curve

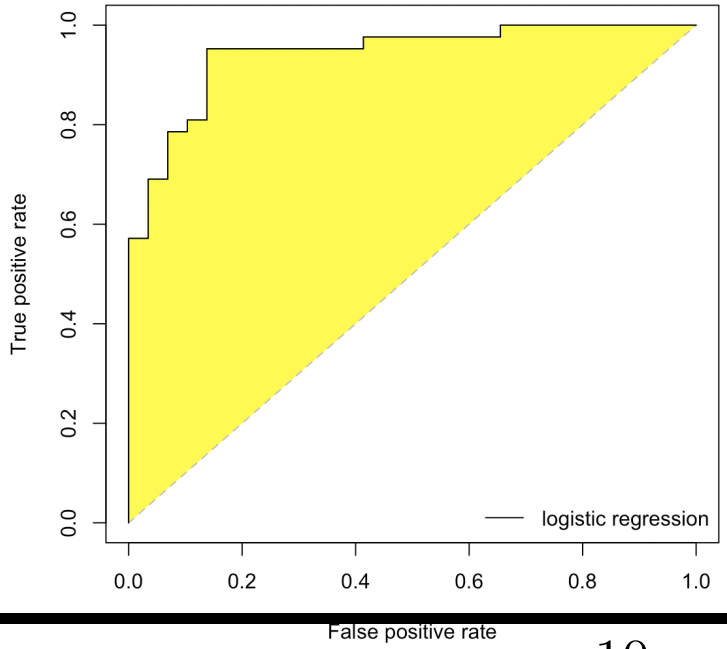
29 deaths $(y = 0)$ and 42 survivals $(y = 1)$

$$\widehat{y} = \begin{cases} 1 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] > 100\% \\ 0 \text{ if } \mathbb{P}[Y = 1|\boldsymbol{X}] \leq 100\% \end{cases}$$

50%

		y	
		0	1
$\widehat{y}$	0	29	42
	0	0	29
		$\frac{0}{29+0}$	$\frac{0}{42+0}$
		$= 0.0\%$	$= 0.0\%$
		FPR	TPR





Goodness of Fit: ROC Curve

See Fawcett (2006, An introduction to ROC analysis)

AUC (Area Under the Curve) for classification

The AUC is the area enclosed by the ROC curve

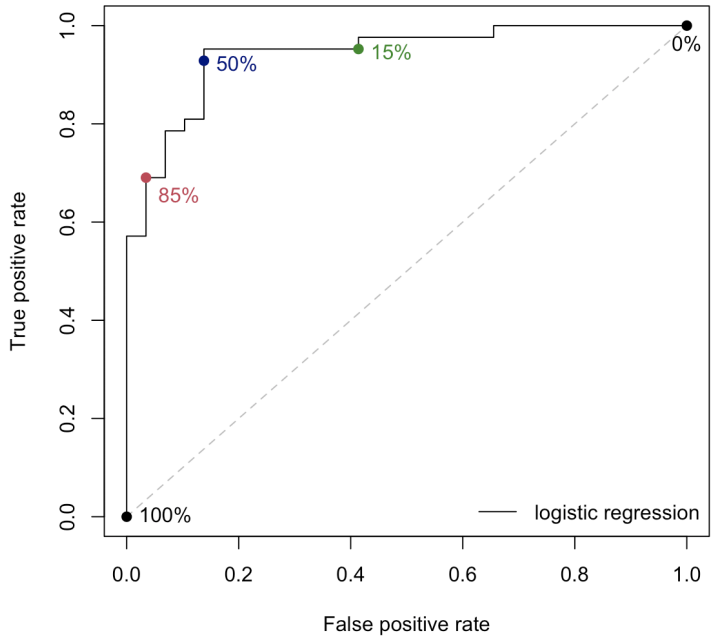
Gini's γ for classification

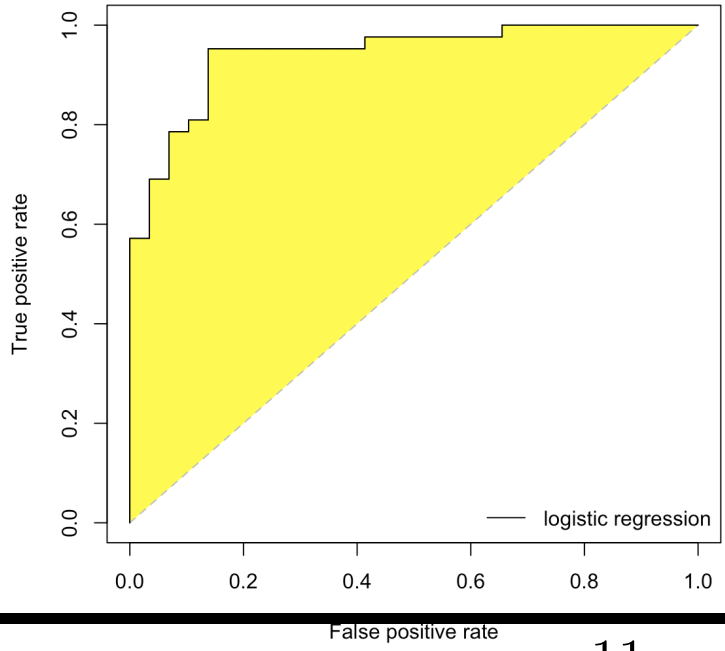
$$\gamma = 2\text{AUC} - 1$$

AUC= $\gamma = 1$ for a perfect classifier

AUC= $1/2$ and $\gamma = 0$ for a random classifier*

* see chi-square independence test

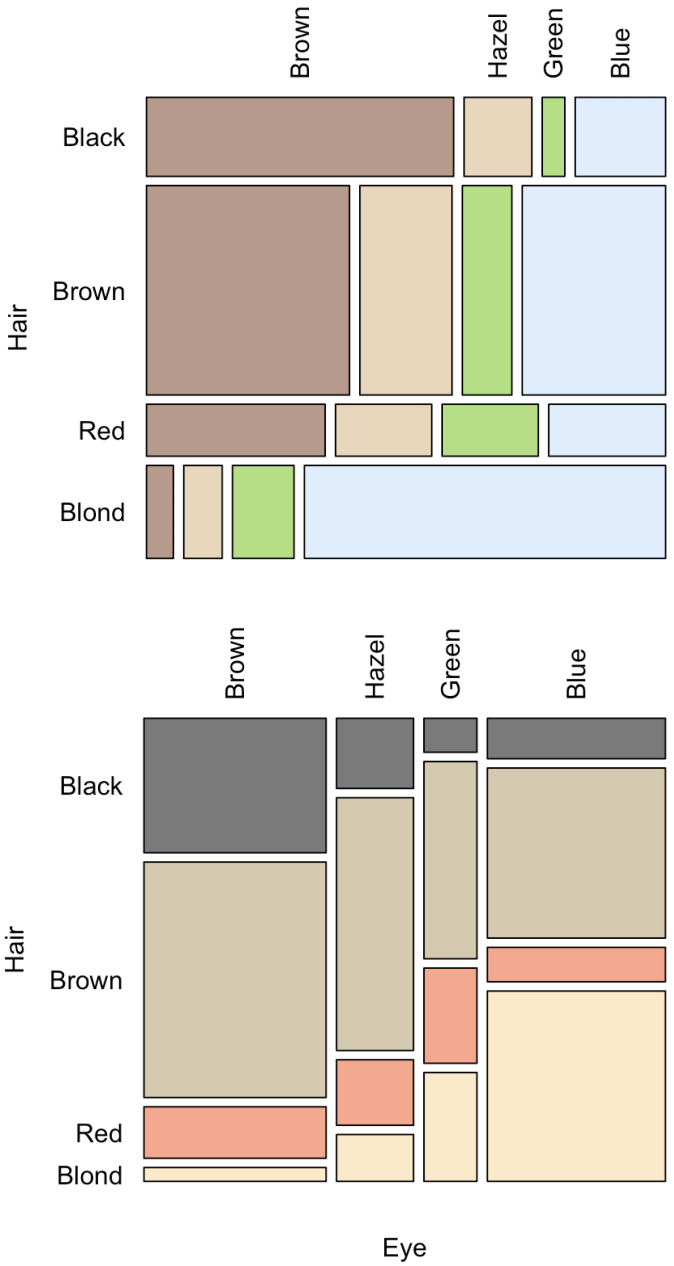




Chi-Square Test for Contingency Tables

	brown	hazel	green	blue	
black	63.0%	13.9%	4.6%	18.5%	100.0%
brown	41.6%	18.9%	10.1%	29.4%	100.0%
red	36.6%	19.7%	19.7%	23.9%	100.0%
blond	5.5%	7.9%	12.6%	74.0%	100.0%
	37.2%	15.7%	10.8%	36.3%	

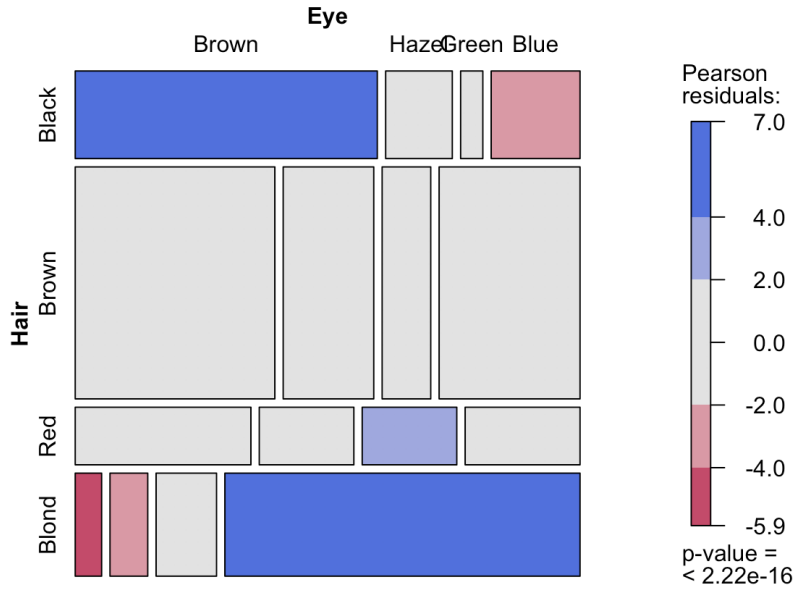
	brown	hazel	green	blue	
black	30.9%	16.1%	7.8%	9.3%	18.2%
brown	54.1%	58.1%	45.3%	39.1%	48.3%
red	11.8%	15.1%	21.9%	7.9%	12.0%
blond	3.2%	10.8%	25.0%	43.7%	21.5%
	100.0%	100.0%	100.0%	100.0%	



Chi-Square Test for Contingency Tables

	brown	hazel	green	blue	
black	68	15	5	20	108
brown	119	54	29	84	286
red	26	14	14	17	71
blond	7	10	16	94	127
	220	93	64	215	

	brown	hazel	green	blue	
black	40	17	12	39	108
brown	106	45	31	104	286
red	26	11	8	26	71
blond	47	20	14	46	127
	220	93	64	215	



Compare $n_{i,j}$ and $n_{i,j}^{\perp}$

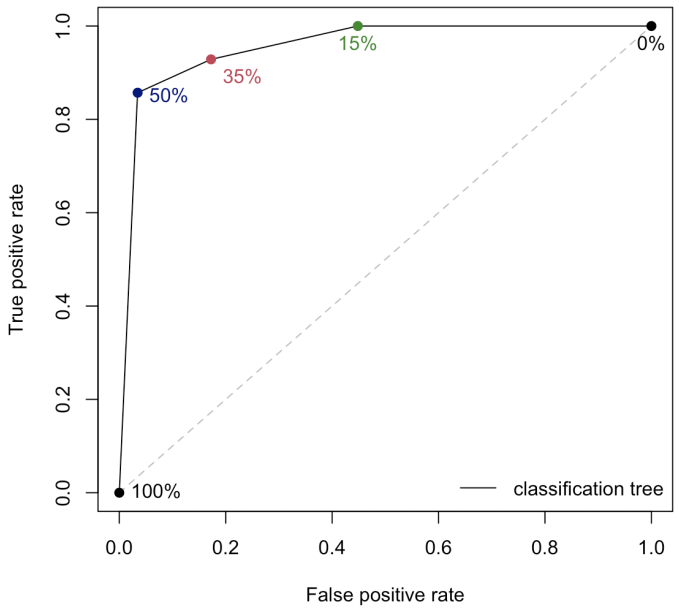
$$n_{i,j}^{\perp} = \frac{n_{i,\cdot} \times n_{\cdot,j}}{n}$$

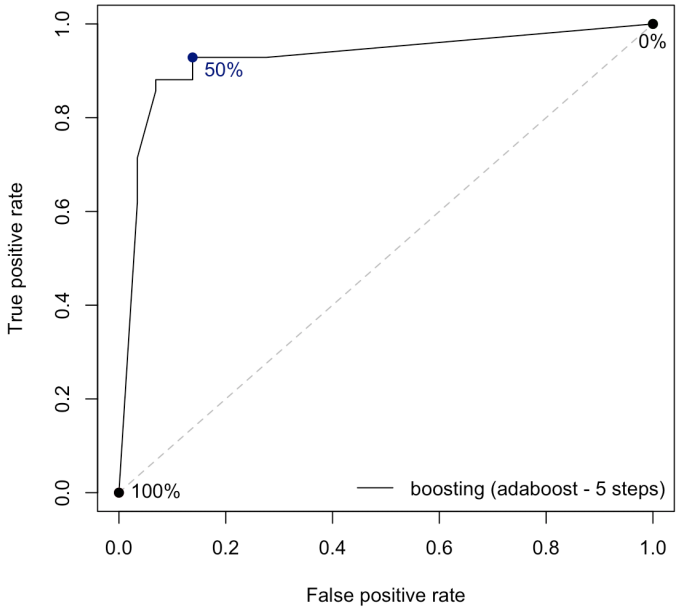
Goodness of Fit

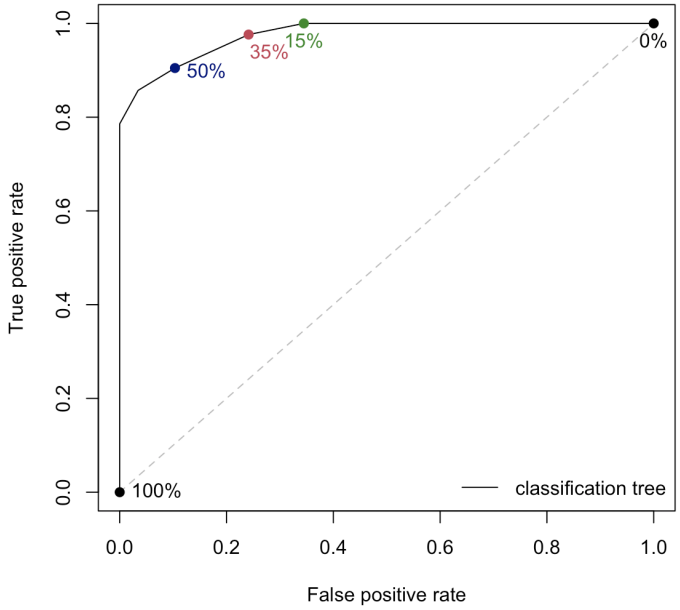
Be carefull of overfit...

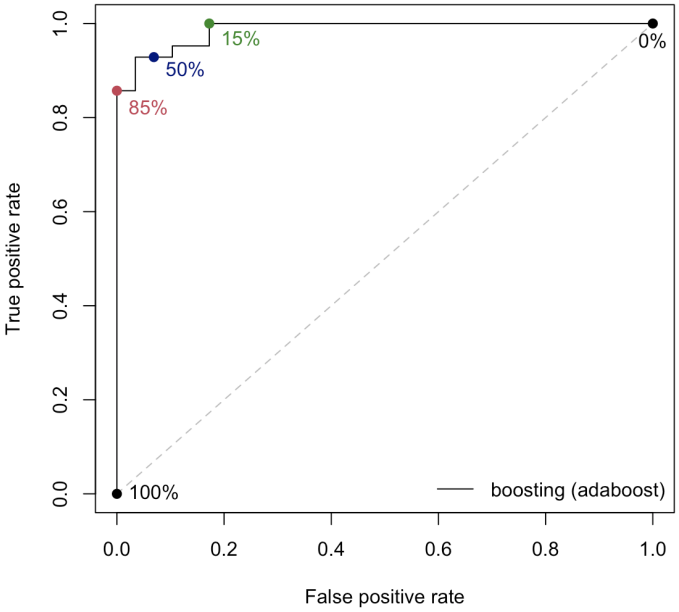
Important to use training / validation

- classification tree
- boosting classifier









Goodness of Fit: ROC Curve

ROC (Receiver Operating Characteristic) Curve

Assume that m_t is define from a score function s, with $m_t(\boldsymbol{x}) = \mathbf{1}(s(\boldsymbol{x}) > t)$ for some threshold t. The ROC curve is the curve $(\text{FPR}_t, \text{TPR}_t$ obtained from confusion matrices of m_t's.

Goodness of Fit: ROC Curve

With categorical variables, we have a collection of points

Need to interpolate between those points to have a *curve*, see convexification of ROC curves, e.g. Flach (2012, Machine Learning)

Goodness of Fit

One can derive a confidence interval of the ROC curve using boostrap techniques,

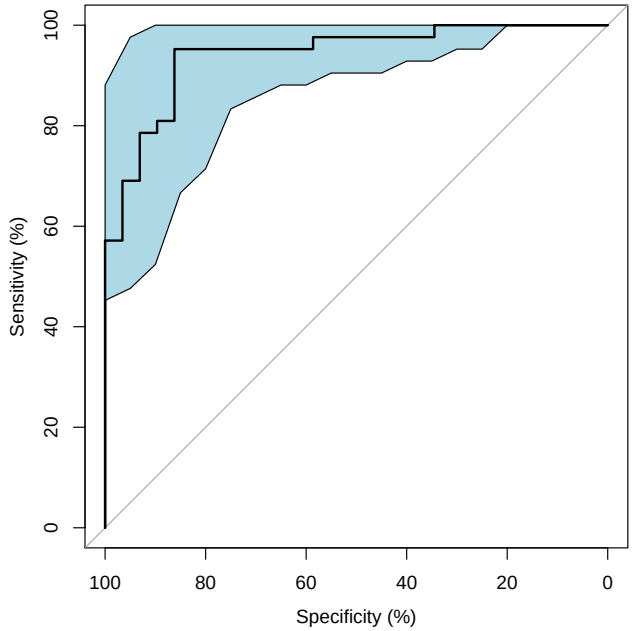
`pROC::ci.se()`

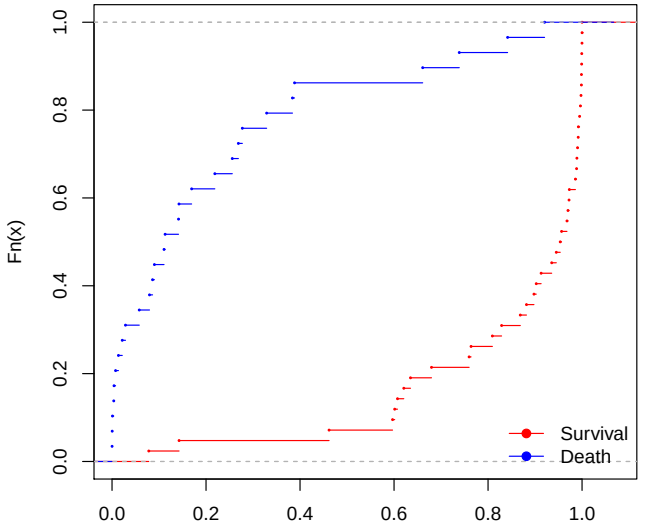
Various measures can be used see library `hmeasures`, with Gini index, the Area Under the curve, or

Kolmogorov-Smirnov for classification

$$ks = \sup_{t\in\mathbb{R}} \left\{ \left| \widehat{F}_1(t) - \widehat{F}_0(t) \right| \right\}$$

$$ks = \sup_{t\in\mathbb{R}} \left\{ \left| \frac{1}{n_1} \sum_{i:y_i=1} \mathbf{1}_{s(\boldsymbol{x}_i)\leq t} - \frac{1}{n_0} \sum_{i:y_i=0} \mathbf{1}_{s(\boldsymbol{x}_i)\leq t} \right| \right\}$$





Goodness of Fit

The Area Under the Curve, AUC, can be interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, see Swets, Dawes & Monahan (2000, Psychological Science Can Improve Diagnostic Decisions)

Kappa statistic κ compares an Observed Accuracy with an Expected Accuracy (random chance), see Landis & Koch (1977, The Measurement of Observer Agreement for Categorical Data)

	$Y=0$	$Y=1$	
$\widehat{Y}=0$	TN	FN	TN+FN
$\widehat{Y}=1$	FP	TP	FP+TP
	TN+FP	FN+TP	n

See also Obsersed and Random Confusion Tables

	$Y=0$	$Y=1$	
$\widehat{Y}=0$	25	3	28
$\widehat{Y}=1$	4	39	43
	29	42	71

	$Y=0$	$Y=1$	
$\widehat{Y}=0$	11.44	16.56	28
$\widehat{Y}=1$	17.56	25.44	43
	29	42	71

Goodness of Fit

Accuracy for classification

$$\text{(total) accuracy} = \frac{TP+TN}{n}$$

$$\text{total accuracy} = \frac{TP+TN}{n} \sim 90.14\%$$

$$\text{random accuracy} = \frac{[TN+FP]\cdot[TP+FN]+[TP+FP]\cdot[TN+FN]}{n^2} \sim 51.93\%$$

Cohen's κ for classification

$$\kappa = \frac{\text{(total) accuracy} - \text{random accuracy}}{1 - \text{random accuracy}}$$

from Cohen, Jacob (1960, A coefficient of agreement for nominal scales). Here $\kappa \sim 79.48\%$.