

Logistiline regressioon

Mudelid I

Illustreeriv näide

Populatsioon

Inimesed Eestis:

1 331 796

	haigestub	jääb terveks	Kokku
praegu	32 822	1 298 974	1 331 796
muutes	9 940	1 321 856	1 331 796

$$OR = \frac{9\,940/1\,321\,856}{32\,822/1\,298\,974} = 0,2976$$

$$RR = \frac{9\,940/1\,331\,796}{32\,822/1\,331\,796} = 0,3028$$

$$RD = 32\,822/1\,331\,796 - 9\,940/1\,331\,796 = 0,01718$$

Illustreeriv näide

Populatsioon

Inimesed Eestis:

1 331 796

Valim
(randomiseeritud
katsesse värvatud)

20 000

kontrollid
10 000
saavad ravi
10 000

Randomiseeritud katse; esindav valim

sugu	vanus	ravim	tulemus
mees	32	ravi	-
mees	5	-	H+
naine	48	-	-
naine	12	ravi	-
naine	63	ravi	-
naine	47	ravi	-
naine	39	ravi	-
mees	0	-	-
naine	6	ravi	-
mees	62	ravi	-
naine	3	-	-
mees	35	ravi	-
...	...	...	...

$$OR = \frac{9\,940/1\,321\,856}{32\,822/1\,298\,974} = 0,2976$$

$$\widehat{OR}_{n=200\,000} = 0,296$$

$$\widehat{OR} = \frac{83/9\,917}{216/9\,784} = 0,3791$$

95%-UI
0,292...0,487

	H+	H-
kontrollid	216	9 784
ravi	83	9 917

Logistiline regressioon

model = glm(1*(tulemus=="H+") ~ ravim,
family=binomial(), data=valim)

log(*shanss*) = ...

log($P(Y = 1)/P(Y \neq 1)$) = ...

log($P(Y = 1)/P(Y \neq 1)$) = $c_0 + c_1 I(\text{"ravi"})$

> summary(model)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.813	0.069	-55.4	< 2e-16 ***
ravimravi	-0.970	0.130	-7.5	8.3e-14 ***

$$\widehat{OR} = \frac{83/9\,917}{216/9\,784} = 0,3791$$

95%-UI
0,292...0,487

Randomiseeritud katse; esindav valim

sugu	vanus	ravim	tulemus
mees	32	ravi	-
mees	5	-	H+
naine	48	-	-
naine	12	ravi	-
naine	63	ravi	-
naine	47	ravi	-
naine	39	ravi	-
mees	0	-	-
naine	6	ravi	-
mees	62	ravi	-
naine	3	-	-
mees	35	ravi	-
...	...	...	...

	H+	H-
kontrollid	216	9 784
ravi	83	9 917

Logistiline regressioon

Randomiseeritud katse; esindav valim

```
model = glm( 1*(tulemus=="H+") ~ ravim,
             family=binomial(), data=valim)
```

$\log(\text{shanss}) = \dots$

$\log(P(Y = 1)/P(Y \neq 1)) = \dots$

$\log(P(Y = 1)/P(Y \neq 1)) = c_0 + c_1 I(\text{"ravi"})$

```
> summary(model)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.813      0.069  -55.4   < 2e-16 ***
ravimravi    -0.970      0.130   -7.5   8.3e-14 ***
```

$$\widehat{OR} = \frac{83/9917}{216/9784} = 0,3791 \quad \text{95\%-UI} \quad 0,292...0,487$$

Ilma ravita

$$\log(\text{šanss}_R) = -3,813$$

Raviga

$$\log(\text{šanss}_{R+}) = -3,813 - 0,97$$

$$\begin{aligned} \log(OR) &= \log(\text{šanss}_{R+}/\text{šanss}_R) \\ &= \log(\text{šanss}_{R+}) - \log(\text{šanss}_R) \\ &= -0,97 \end{aligned}$$

$$\begin{aligned} OR &= \exp(-0,97) \\ &= 0,3791 \end{aligned}$$

	H+	H-
kontrollid	216	9 784
ravi	83	9 917

Logistiline regressioon

Randomiseeritud katse; esindav valim

Ilma ravita

$$\log(\text{šanss}_R) = -3,813$$

$$\text{šanss}_R = \exp(-3,813)$$

$$P(Y = 1|R-) = \frac{\exp(-3,813)}{\exp(-3,813) + 1}$$

$$P(Y = 1|R-) = 0,0216$$

$$\text{šanss} = p/(1-p)$$

$$\text{šanss}(1-p) = p$$

$$\text{šanss} - p \cdot \text{šanss} = p$$

$$\text{šanss} = (\text{šanss} + 1) \cdot p$$

$$p = \text{šanss}/(\text{šanss} + 1)$$

$$\widehat{OR} = \frac{83/9917}{216/9784} = 0,3791 \quad \text{95\%-UI} \quad 0,292...0,487$$

	H+	H-
kontrollid	216	9 784
ravi	83	9 917

Logistiline regressioon

Randomiseeritud katse; esindav valim

```
model = glm( 1*(tulemus=="H+") ~ ravim,
             family=binomial(), data=valim)
```

```
predict(model, data.frame(ravim="-"),
        type="response")
```

0.0216

$$\widehat{OR} = \frac{83/9917}{216/9784} = 0,3791 \quad \text{95\%-UI} \quad 0,292...0,487$$

Ilma ravita

$$\log(\text{šanss}_R) = -3,813$$

$$\text{šanss}_R = \exp(-3,813)$$

$$P(Y = 1|R-) = \frac{\exp(-3,813)}{\exp(-3,813) + 1}$$

$$P(Y = 1|R-) = 0,0216$$

Ravita Eestis, hinnanguliselt:

$$0,0216 \cdot 1\,331\,796 = 28\,767$$

Tegelik: 32 822

	H+	H-
kontrollid	216	9 784
ravi	83	9 917

Logistiline regressioon

Randomiseeritud katse; esindav valim

```
model = glm( 1*(tulemus=="H+") ~ ravim,
             family=binomial(), data=valim)
```

```
predict(model, data.frame(ravim="ravi"),
        type="response")
```

0.0083

$$\widehat{OR} = \frac{83/9917}{216/9784} = 0,3791 \quad \text{95\%-UI} \quad 0,292...0,487$$

Raviga

$$\log(\text{šanss}_{R+}) = -3,813 - 0,97$$

$$\text{šanss}_{R+} = \exp(-3,813) \cdot \exp(-0,97) = 0,00837$$

$$P(Y = 1|R+) = \frac{0,00837}{0,00837 + 1} = 0,00830$$

Ravita Eestis, hinnanguliselt:

$$0,0083 \cdot 1\,331\,796 = 11\,054$$

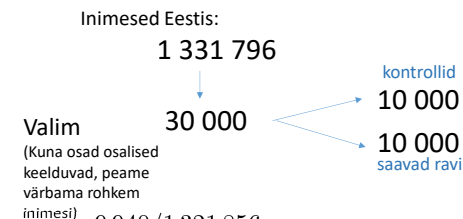
Tegelik: 9 940

	H+	H-
kontrollid	216	9 784
ravi	83	9 917

Illustreeriv näide

Randomiseeritud katse; mittevastamine

Populatsioon



Miks OR-i hinnang suure valimi korral ei lähene tegelikule OR-i väärtusele?

$$OR = \frac{9940/1\,321\,856}{32\,822/1\,298\,974} = 0,2976$$

$$\widehat{OR}_{n=300\,000} = 0,3388 \quad 95\%-UI \quad 0,314...0,365$$

$$\widehat{OR} = \frac{91/9\,909}{247/9\,753} = 0,362$$

	H+	H-
kontrollid	247	9 753
ravi	91	9 909

Illustreeriv näide

Randomiseeritud katse; mittevastamine

Populatsioon

Inimesed Eestis:

1 331 796

```
> mudel = glm( 1*(tulemus=="H+") ~ ravim+ravim*sugu,
family=binomial(), data=valimB)
> summary(mudel)
```

Coefficients:

	Estimate	Std. Error	z	Pr(> z)
(Intercept)	-3.413	0.031	-109.2	< 2e-16 ***
ravimravi	-1.446	0.071	-20.3	< 2e-16 ***
sugunaine	-0.307	0.040	-7.6	2.59e-14 ***
ravimravi:sugunaine	0.547	0.085	6.4	1.31e-10 ***

$$\widehat{OR}_{mees} = \exp(-1,446) = 0,235$$

$$\widehat{OR}_{naine} = \exp(-1,446 + 0,547) = 0,407$$

$$\widehat{shanss}_{mees,ravita} = \exp(-3,413)$$

$$\widehat{shanss}_{mees,ravi} = \exp(-3,413 - 1,446)$$

$$\widehat{OR}_{mees} = \frac{\exp(-3,413 - 1,446)}{\exp(-3,413)}$$

Illustreeriv näide

Randomiseeritud katse; mittevastamine

Populatsioon

Inimesed Eestis:

1 331 796

$$OR_{mees} = 0,229$$

$$OR_{naine} = 0,375$$

$$\widehat{OR}_{mees} = \exp(-1,446) = 0,235$$

$$\widehat{OR}_{naine} = \exp(-1,446 + 0,547) = 0,407$$

$$\widehat{shanss}_{naine,ravita} = \exp(-3,413 - 0,307)$$

$$\widehat{shanss}_{naine,ravi} = \exp(-3,413 - 1,446 - 0,307 + 0,547)$$

$$\widehat{OR}_{naine} = \frac{\exp(-3,413 - 1,446 - 0,307 + 0,547)}{\exp(-3,413 - 0,307)}$$

```
> mudel = glm( 1*(tulemus=="H+") ~ ravim+ravim*sugu,
family=binomial(), data=valimB)
> summary(mudel)
```

Coefficients:

	Estimate	Std. Error	z	Pr(> z)
(Intercept)	-3.413	0.031	-109.2	< 2e-16 ***
ravimravi	-1.446	0.071	-20.3	< 2e-16 ***
sugunaine	-0.307	0.040	-7.6	2.59e-14 ***
ravimravi:sugunaine	0.547	0.085	6.4	1.31e-10 ***

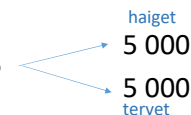
Illustreeriv näide

Juht-kontrollluuring

Populatsioon

Inimesed Eestis:

1 331 796



```
> mudel0 = glm( 1*(ravim=="ravi") ~ tulemus, family=binomial(), data=valimCC)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.58928	0.02952	19.96	<2e-16 ***
tulemusH+	-1.18292	0.04176	-28.32	<2e-16 ***

```
> predict(mudel0, data.frame(tulemus="H+"), type="response")
1
0.3558
```

Tegelikult saavad haigeks jäänutest ennetavat ravi 36%

Illustreeriv näide

Juht-kontrollluuring

Populatsioon



```
> mudel = glm( 1*(tulemus=="H+") ~ ravim, family=binomial(), data=valim)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.59083	0.02951	20.02	<2e-16 ***
ravimravi	-1.18292	0.04176	-28.32	<2e-16 ***

```
> predict(mudel, data.frame(ravim="ravi"), type="response")
1
0.3561562
```

Ravimi võtjate seas on haigestunute tegelik osakaal 0,00725...

Illustreeriv näide

Juht-kontrollluuring

Populatsioon



```
> mudel0 = glm( 1*(ravim=="ravi") ~ tulemus, family=binomial(), data=valimCC)
```

	Estimate	Std. Error	z value	Pr(> z)
tulemusH+	-1.18292	0.04176	-28.32	<2e-16 ***

```
> mudel = glm( 1*(tulemus=="H+") ~ ravim, family=binomial(), data=valim)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
ravimravi	-1.18292	0.04176	-28.32	<2e-16 ***

Illustreeriv näide

Juht-kontrollluuring

Populatsioon



```
> mudel = glm( 1*(tulemus=="H+") ~ ravim*sugu, family=binomial(), data=valim)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.75535	0.05516	13.69	< 2e-16 ***
ravimravi	-1.50784	0.06819	-22.11	< 2e-16 ***
sugunaine	-0.23322	0.06533	-3.57	0.000357 ***
ravimravi:sugunaine	0.59777	0.08842	6.76	1.38e-11 ***

$$\widehat{OR}_{mees} = \exp(-1,5078) = 0,221 \quad \widehat{OR}_{naine} = \exp(-1,5078 + 0,59777) = 0,402$$

$$OR_{mees} = 0,229 \quad OR_{naine} = 0,375$$

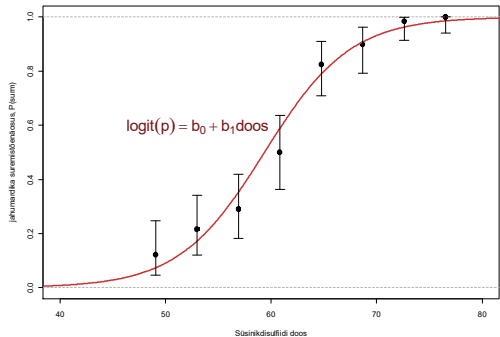
Miks armastatakse juht-kontrolluuringuid?

	-	H+	$\widehat{OR} \cdot \exp\left(\pm 1,96 \sqrt{\frac{1}{1784} + \frac{1}{3221} + \frac{1}{3216} + \frac{1}{1779}}\right)$
-	1784	3221	$0,306 \cdot 0,92 = 0,282$
ravi	3216	1779	$0,306 \cdot 1,085 = 0,332$

	H+	H-	$\widehat{OR} \cdot \exp\left(\pm 1,96 \sqrt{\frac{1}{216} + \frac{1}{9784} + \frac{1}{83} + \frac{1}{9917}}\right)$
kontrollid	216	9 784	
ravi	83	9 917	$0,38 \cdot 0,77 = 0,29$
			$0,38 \cdot 1,29 = 0,49$

Vaatamata 2x suuremale valimile on usalduspiirid märksa laiemad – sest sagedustabelis esineb väikeste väärtustega lahtreid

Pidev tunnus ja logistiline regressioon



Näide 2

```
> jahumardikas
  s  n doos
1  6 49 49.06
2 13 60 52.99
3 18 62 56.91
4 28 56 60.84
5 52 63 64.76
6 53 59 68.69
7 61 62 72.61
8 60 60 76.54
```

```
logist_reg=glm(cbind(s, n-s)~doos, family=binomial())
summary(logist_reg)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.57806    1.29846  -11.23  <2e-16 ***
doos          0.24554     0.02149   11.42  <2e-16 ***
```

Doos 50, log-surmašansid:	surmašansid
-14.578+0.245*50	exp(-14.578+0.245*50)
	exp(-14.578)exp(0.245) ⁵⁰
Doos 51, log-surmašansid:	surmašansid
-14.578+0.245*51	exp(-14.578+0.245*51)
	exp(-14.578)exp(0.245) ⁵¹
Doosi suurendamisel ühe ühiku võrra suurenevad jahumardika suremise šansid exp(0.245)=1,278 korda.	

Prognoosid, tõenäosus

Logistiline regressioon arvutab sündmuse toimumise tõenäosust (ühtedes või teistes tingimustes). Tõenäosuse (uuritava 0/1 tunnuse keskvääruse) leidmiseks peame kasutama logit-funktsiooni pöördfunktsiooni expit-funktsiooni. Kui logistilise regressiooni parameetrite hinnangud olid järgmised:

```
(Intercept) -14.57806
doos          0.24554
```

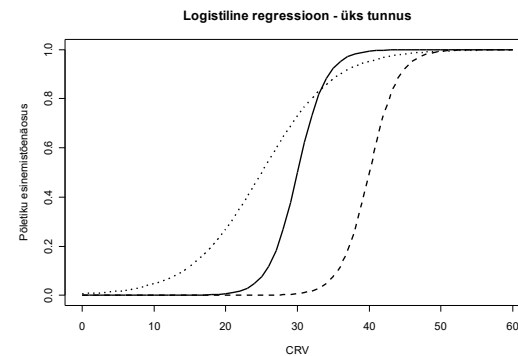
Siis jahumardika suremistõenäosus *doos*=60 korral on leitav järgmiselt:

$$P(surm) = \frac{\exp(-14.578 + 0.2455 \cdot 60)}{1 + \exp(-14.578 + 0.2455 \cdot 60)}$$
$$= \frac{1,1669}{1 + 1,1669} = 0,5385...$$

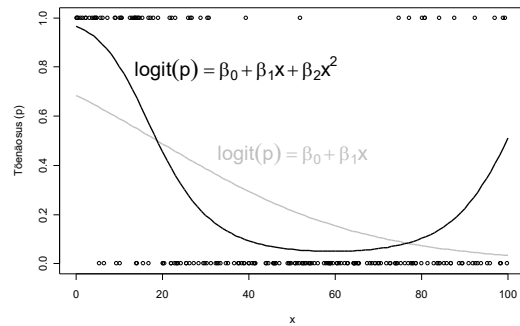
Või, predict-käsu abil:

```
> predict(logist_reg,
  data.frame(doos=60),
  type="response")
1
0.5385063
```

Interpretatsioon: vabaliige ja tõus

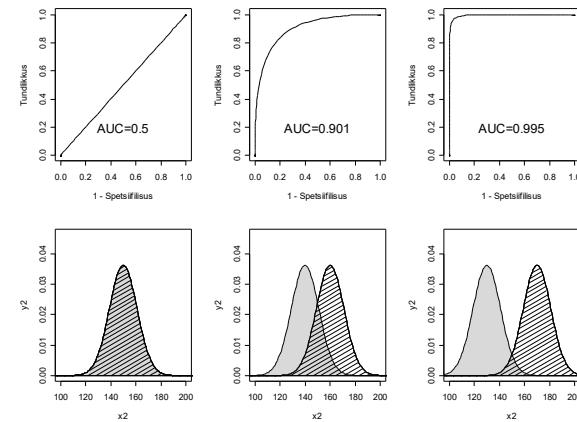
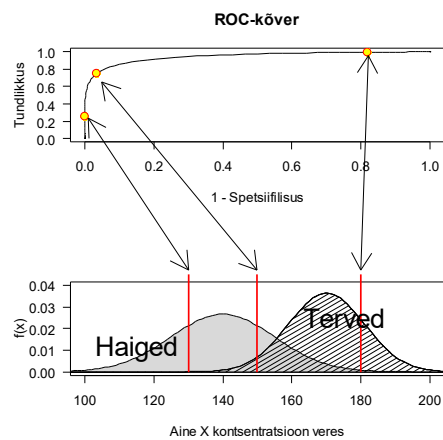


Mitte alati ei pruugi
vaikimisi seos sobida...

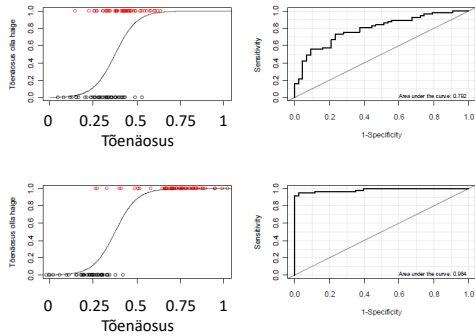


Logistilise regressioonimudeli headuse kirjeldamine.

Mudeli prognoosivõime. ROC-kõver ja AUC



Täpselt sama logistilise regressiooni mudel – kaks erinevat AUC väärtust...



Näide:

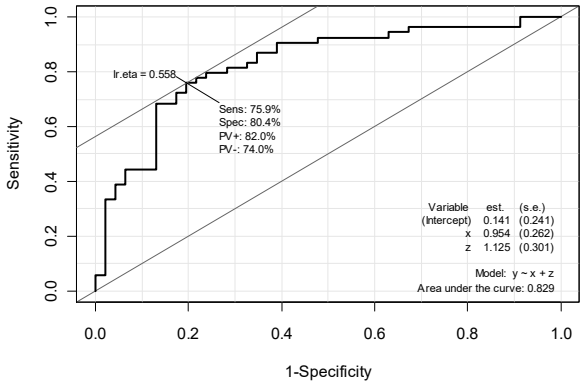
[...]

For the inductive model, we initially created eight habitat models using between two and six variables (Table 6). We used 223 sites for model training and withheld 54 additional sites for testing. The AUC values for these models ranged from 0.782 to 0.914 (Table 6). The highest AUC value (0.914) was associated with the three-variable model (land cover, distance to mesquite, and distance to permanent streams; Figure 12C). This model makes biological sense given the phainopepla's dependence on mistletoe associated with mesquite/acacia vegetation (Crampton et al. 2006), and the common occurrence of this vegetation along desert streams.

[...]

Habitat Distribution Models for 37 Vertebrate Species in the Mojave Desert Ecoregion of Nevada, Arizona, and Utah
Kenneth G. Boykin; David F. Bradford; William G. Kepner

Veel üks näide – ROC-kõver lisamooduli Epi funktsiooniga ROC



Positive Predictive Value (PPV):

$$PPV = \frac{PR \cdot Tundlikkus}{PR \cdot Tundlikkus + (1 - PR)(1 - Spetsiifilisus)}$$

Negative Predictive Value (NPV):

$$NPV = \frac{(1 - PR) \cdot Spetsiifilisus}{(1 - PR) \cdot Spetsiifilisus + PR(1 - Tundlikkus)}$$

Mõningate testide PPV:

	Tundlikkus	Spetsiifilisus	PPV
Munasarjavähk	100%	95%	2,5%
Rinnavähk (mammograaf)			
noorem kui 50	90%	95%	alla 20%
vanem kui 50	90%	95%	üle 60%

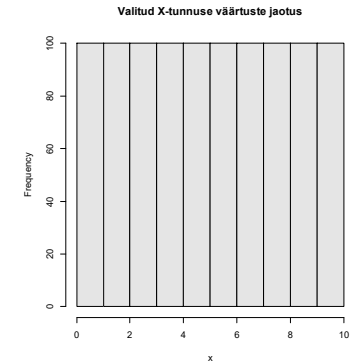
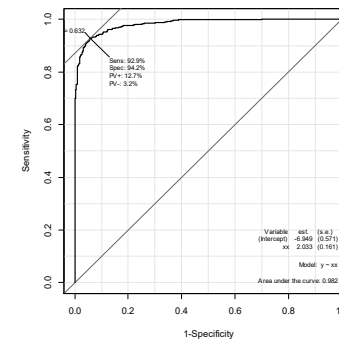
Kui X-tunnuse väärtuseid valib uurija...

ROC-kõver, spetsiifilisus ja tundlikkus, positiivne ja negatiivne prognoosiväärtus omavad tähendust siis, kui uuritavate tunnuste väärtused (nii X kui ka Y) on pärit mingist populatsioonist võetud juhuslikust valimist (ei ole toimunud „meelepäraste“ vaatluste eelnevat väljavalmist).

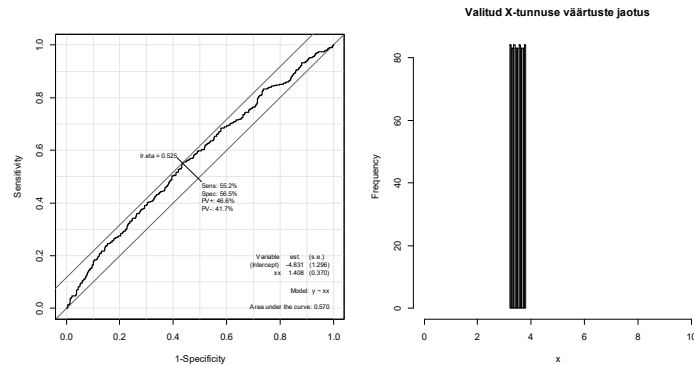
Kui aga näiteks uurija kontrollib X-tunnuse väärtuseid (otsustab keda uuringusse kaasata või millises koguses kemikaali kasutada) siis on näiteks uuringu tulemuseks saadav AUC väärtus suuresti uurija enda määrata (ja seega iseloomustab pigem uurijat kui uuritavat).

Vaata ka järgnevaid jooniseid, kõigil järgnevatel joonistel on andmeid tekitav mehhanism sama (aga uurija on eksperimendis kasutanud erinevaid X-tunnuse väärtuseid...)

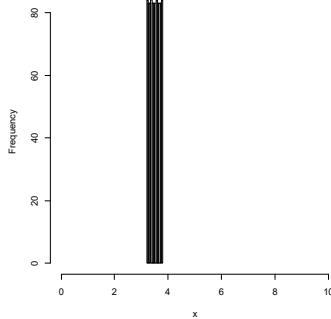
AUC=0,982



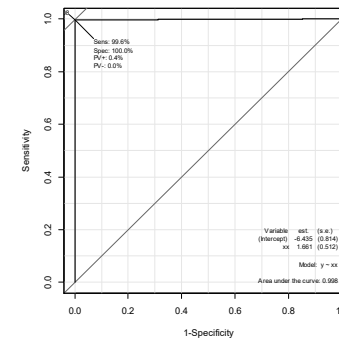
AUC=0,570



Valitud X-tunnuse väärtuste jaotus



AUC=0,998



Valitud X-tunnuse väärtuste jaotus

