Biomeetria bioloogidele
8. loeng

Logistiline regressioon

Mudel binaarsele tunnusele

Logistiline regressioon

Uuritav tunnus binaarne, kahe võimaliku väärtusega (esines/ei esinenud; suri/ei surnud; idanes/ei idanenud; ...). Tavaliselt kodeeritakse tunnus väärtustega 1 (sündmus toimus) ja 0 (sündmus ei toimunud).

Mudel hinnatakse sellisel juhul tõenäosusele – prognoosime meid huvitava sündmuse tõenäosust (ehk tõenäosust, et uuritav tunnus omandab väärtuse 1).

Binaarse (0/1) tunnuse keskväärtus

Olgu *Y* binaarne tunnus:

$$Y = \begin{cases} 1, & \text{kui sündmus toimus} \\ 0, & \text{kui sündmust ei toimunud} \end{cases}$$

Sellise tunnuse keskväärtus on

$$EY = 0 \cdot P(Y=0) + 1 \cdot P(Y=1) = P(Y=1)$$

ja keskmine

$$\overline{Y} = \underbrace{(0+\ldots+0}_{n_0} + \underbrace{1+\ldots+1)}_{n_1} / n = \left(n_0 \cdot 0 + n_1 \cdot 1\right) / n = n_1 / n$$

näitab sündmuse toimumise suhtelist sagedust.

Seega...

Soovime oma mudeli abil hinnata meid huvitava sündmuse toimumise tõenäosust. Selleks võime aga hinnata binaarse (0/1) tunnuse keskväärtuse (sest keskväärtus ja tõenäosus on antud juhul sama asi). Keskväärtust saame aga hinnata juba tuttaval viisil – (üldistatud) vähimruutude meetodil.

Kuid... mudelisse tuleks kirja panna meile teadaolev lisainformatsioon...

Milline lisainformatsioon?

Mida teame 0/1-tunnuse keskväärtuse kohta?

- Sellise tunnuse keskväärtus ei saa olla negatiivne (väiksem nullist)
- Sellise tunnuse keskväärtus ei saa olla suurem kui 1

Vastav lisainformatsioon tuleks mudelisse kirjutada sobivat seosefunktsiooni kasutades.

Milline seosefunktsioon?

Üllataval kombel eksisteerib (ja on kasutusel) päris palju erinevaid seosefunktsioone, mis kõik tagavad keskväärtuse hinnangu püsimise vahemikus [0...1].

Logistilise regressiooni puhul kasutatakse aga seosefunktsioonina logit-funktsiooni, ehk hinnatakse mudelit kujul

$$\text{logit}(EY) = \beta_0 + \beta_1 X_1 + \ldots$$

$$\text{logit}(p_1) = \beta_0 + \beta_1 X_1 + \ldots$$

Logit ja Expit funktsioonid

logit(p_1) = log(p_1/(1-p_1)) = $\beta_0 + \beta_1 X_1 + \ldots$

$$p_1 = \text{expit}(\text{logit}(p_1))$$
$$= \text{expit}(\beta_0 + \beta_1 x_1 + \ldots)$$
$$= \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots)}$$

Interpretatsioonist: šanss

Hobuse Rukkilill võidušansid on 1/10-le

võidu võimalused | kaotuse võimalused

1/10 = (1/11) / (10/11)

võidu tõenäosus | kaotuse tõenäosus

Sündmuse „1“ toimumise šansid (*odds*) on

šanss(1) = p_1/(1- p_1)

Interpretatsioonist: šanss

Sündmuse „1“ toimumise šansid (*odds*) on

$$\text{šanss}(1) = p_1/(1- p_1)$$

Logistilise regressiooni mudel

$$\log(p_1/(1-p_1)) = \beta_0 + \beta_1 X_1 + ...$$

on mudel logaritmitud šanssidele.

Interpretatsioonist

Logistilise regressiooni mudel:

log(võidušanss) = $\log(p_1/(1-p_1)) = \beta_0 + \beta_1 X_1 + ...$

Olgu X=0, kui naine
X=1, kui mees

log(võidušanss) = $\log(p/(1-p)) = \beta_0 + \beta_1 X$

log(võidušanss) = $\log(p_N/(1-p_N)) = \beta_0$ naine
log(võidušanss) = $\log(p_M/(1-p_M)) = \beta_0 + \beta_1$ mees

$\beta_1 = \log(p_M/(1-p_M)) - \log(p_N/(1-p_N))$

log(a)-log(b)=log(a/b)

$$= \log\left(\frac{p_M/(1-p_M)}{p_N/(1-p_N)}\right)$$

β_1 – logaritm võidušansside (1 saamise šansside) suhtest *log-odds ratio*

Näide 1

Lõhe röövimine teiselt kotkalt õnnestus(1) / ei õnnestunud (0)	Röövija vanus noor / vana
1	vana
1	noor
0	noor
0	noor
1	vana
....	....

	noor	vana
õnnestus	38	62
Ei õnnestunud	41	19

Knight, R. L. and Skagen, S. K. (1988) Agonistic asymmetries and the foraging ecology of Bald Eagles. *Ecology* **69**, 1188–1194.

Näide 1

Õnnestumise šansid noortel

38/41 = 0,9268

	noor	vana
õnnestus	38	62
ei õnnestunud	41	19

Õnnestumise šansid vanadel

62/19 = 3,2632

	šanss	log(šanss)
Noored	0,9268	-0,0759
Vanad	3,2632	1,1827

	šanss	log(šanss)
Noored	0,9268	-0,0759
Vanad	3,2632	1,1827

```
summary(glm(onnestus~factor(vanus),
                  family=binomial()))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.07599    0.22518  -0.337 0.735781
factor(vanus)vana   1.25868    0.34564   3.642 0.000271 ***
```

Noorte lindude õnnestumise šansi logaritm: -0.07599

Ehk noorlindude õnnestumise šanss exp(-0.076) = 0.9268

Vanade lindude õnnestumise šansi logaritm: -0.07599 +1.25868 =1,1827

Ehk täiskasvanud lindude õnnestumise šanss exp(-0.07599 +1.25868) = 3.263

Vanuse kordaja interpretatsioon

```
(Intercept)        -0.07599
factor(vanus)vana   1.25868
```

täiskasvanud lindude õnnestumise šanss

exp(-0.07599 +1.25868) = exp(-0.07599)exp(1.25868)

= noorlinnu šanss * exp(1.25868)

Ehk

- exp(1.25..) korda on vanalinnu eduka röövi šansid paremad kui noorlinnul...
- exp(1.25) on hinnang šansside suhtele (odds ratio).

Seega 1.25 on logaritm šansside suhtest, *log-odds ratio*.

Šanss ja tõenäosus

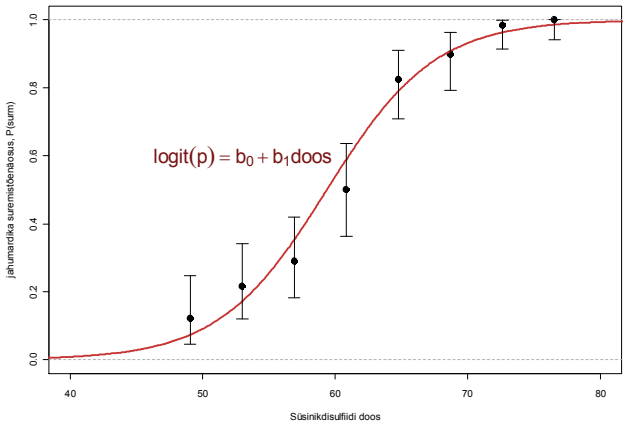
Teades ühte neist on alati võimalik leida teist…

$$Šanss = p/(1-p)$$

$$Šanss(1-p) = p$$
$$Šanss - p \cdot Šanss = p$$
$$Šanss = (Šanss+1) \cdot p$$

$$p = Šanss/(Šanss+1)$$

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
```

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.

```
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 ***
```

Prognoosid, tõenäosus

Logistiline regressioon arvutab sündmuse toimumise tõenäosust (ühtedes või teistes tingimustes). Tõenäosuse (uuritava 0/1 tunnuse keskväärtuse) 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:

$$P(surm) = \frac{\exp(-14.578 + 0.2455*60)}{1+\exp(-14.578+0.2455*60)}$$

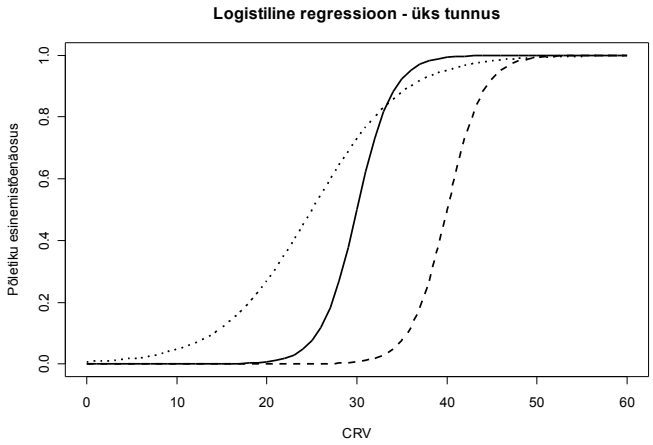
.

$$= \frac{1{,}1669}{1+1{,}1669} = 0{,}5385...$$

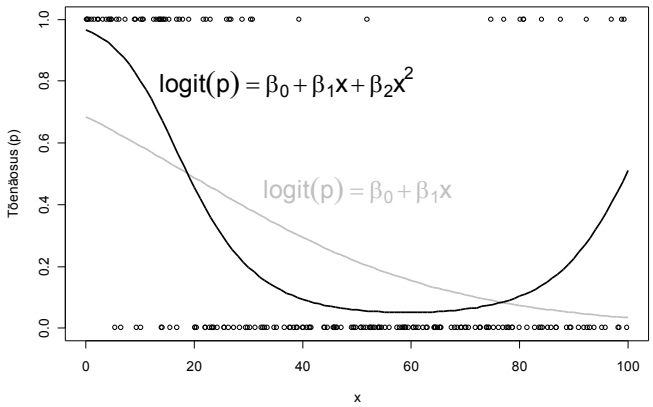
Või, `predict`-käsu abil:

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

Interpretatsioon: vabaliige ja tõus



Mitte alati ei pruugi vaikimisi seos sobida...



Logistiline regressiooni interpreteerimine erinevate katseplaanide korral

Andmete kogumise viis	tõenäosus	OR
juhuslik valim	sobib	sobib
juht-kontrolluuring	-	sobib

Logistilise regressioonmudeli headuse kirjeldamine.

Mudeli prognoosivõime.

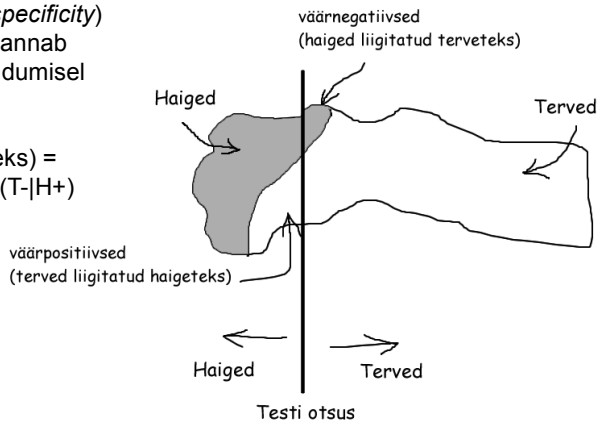
Spetsiifilisus ja tundlikkus.

Testi **tundlikkus** (*sensitivity*) on tõenäosus, et test annab testitava haiguse või seisundi olemasolul positiivse tulemuse:

tundlikkus = P(haigel leitakse haigus) = P(T+|H+)
≈ haigete arv kellel haigus leiti / haigete koguarv

Testi **Spetsiifilisus** (*specificity*) on tõenäosus, et test annab testitava seisundi puudumisel negatiivse tulemuse:

P(terve loetakse terveks) =
= P(T-|H+)

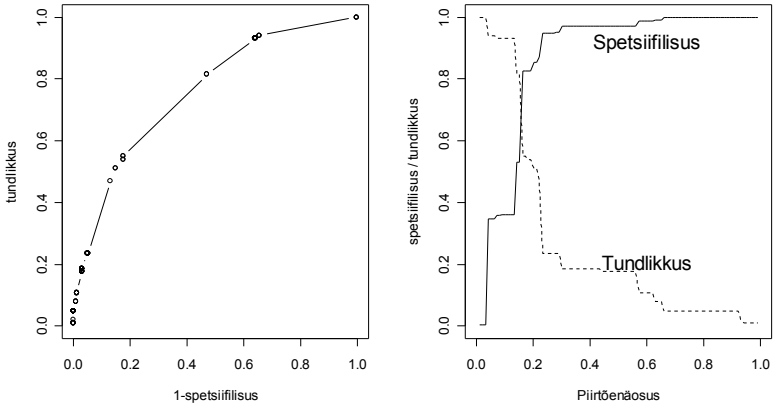


Tundlikkus ja spetsiifilisus ja logistiline regressioon

Esimene märkus – me jagame inimesi haigeteks/terveteks kasutades prognoositud tõenäosust – mingist tõenäosusest alates „prognoosime“ inimese „haigeks“. Näiteks prognoosime inimese suitsetajaks, kui tõenäosus, et ta suitsetab, on suurem kui 0,6:

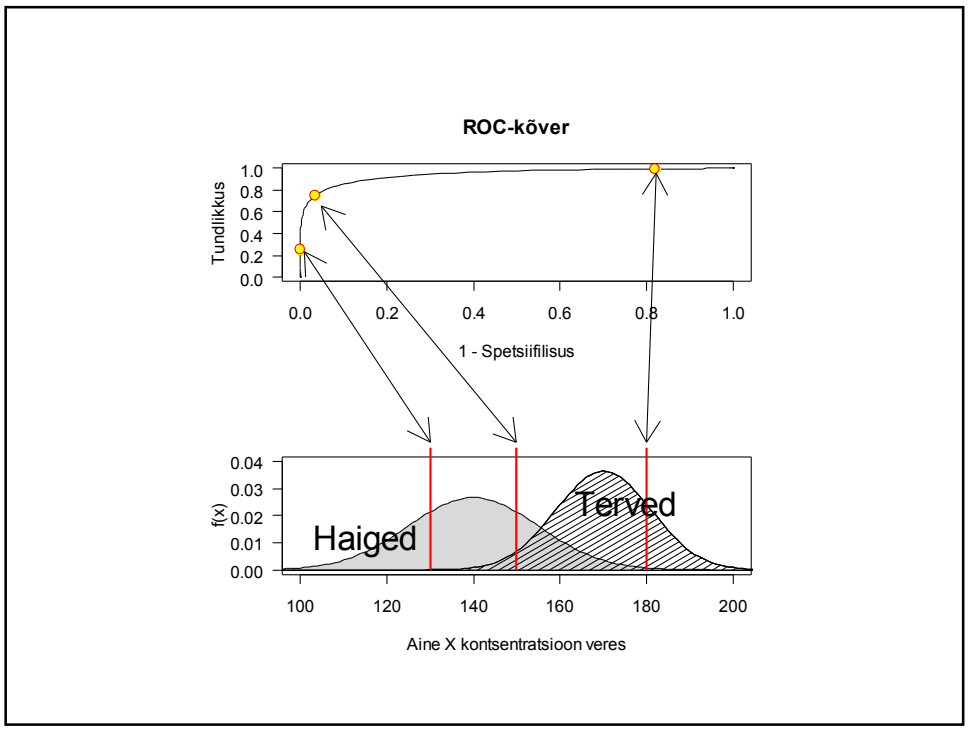
```
> m1=glm(suits2~olu2+factor(viin), family=binomial)
> risk=m1$fitted.values
> prognoos <- (risk>0.6)
> tabel <- table(prognoos,suits2); tabel
        suits2
prognoos 0   1
   FALSE 540  91
   TRUE    6  11
>
> spets = tabel[1,1]/sum(tabel[,1])
> tund = tabel[2,2]/sum(tabel[,2])
> spets; tund;
[1] 0.989011
[1] 0.1078431
```

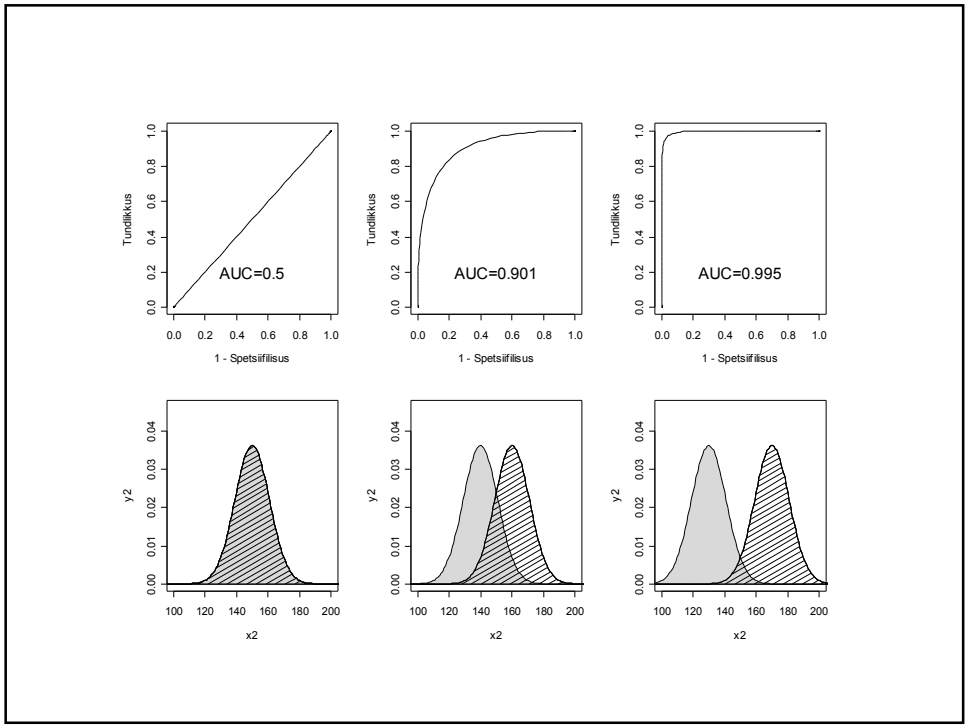
Kui lõikepunkt 0,6 meile ei meeldi, siis võime üritada valida parema otsustuskriteeriumi:



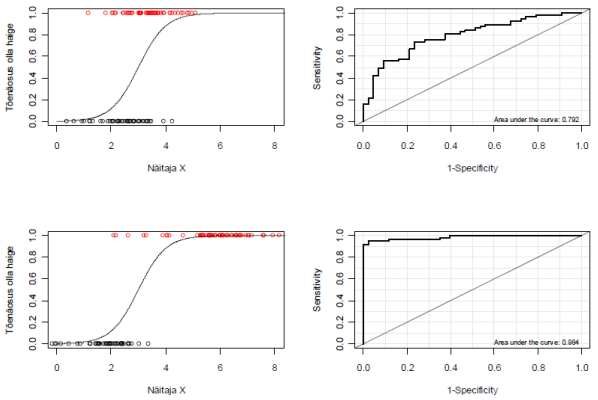
Programm, mis tegi eelmise joonise:

```
m1=glm(suits2~olu2+factor(viin), family=binomial)
risk=m1$fitted.values
loiked=seq(min(risk), 1, by = 0.01); nn=length(loiked)
sen = rep(NA, nn); spe = rep(NA, nn)
for (i in 1:nn) {
   prognoos <- (risk>loiked[i])
   tabel <- table(prognoos,suits2)
   spe[i] = tabel[1,1]/sum(tabel[,1])
   sen[i] = tabel[2,2]/sum(tabel[,2])  }
par(mfrow=c(1,2))
plot(1-spe, sen, type="b", xlab="1-spetsiifilisus",
   ylab="tundlikkus", ylim=c(0,1), xlim=c(0,1))
plot(loiked, spe, type="l", xlim=c(0,1), ylim=c(0,1),
   xlab="Piirtõenäosus", ylab="spetsiifilisus / tundlikkus")
lines(loiked, sen, lty=2)
text(0.6, 0.92, "Spetsiifilisus", cex=1.4)
text(0.6, 0.21, "Tundlikkus", cex=1.4)
```


ROC-kõver
Tundlikkus
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Spetsiifilisus
f(x)
0.04
0.03
0.02
0.01
0.00
Haiged
Terved
100
120
140
160
180
200
Aine X kontsentratsioon veres


Tundlikkus
AUC=0.5
AUC=0.901
AUC=0.995
1 - Spetsiifilisus
y2
x2

Täpselt sama logistilise regressiooni mudel – kaks erinevat AUC väärtust…



Näide:

[...]

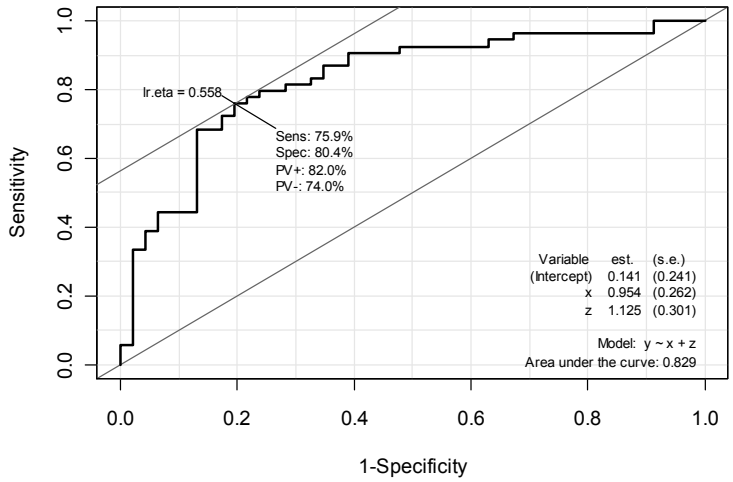
For the inductive model, we initially created eight habitat models using between two and six variables (Table 6). We used 223 sites for model training and withheld 54 additional sites for testing. The AUC values for these models ranged from 0.782 to 0.914 (Table 6). The highest AUC value (0.914) was 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



haiguse levimus uuritavas populatsioonis

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%

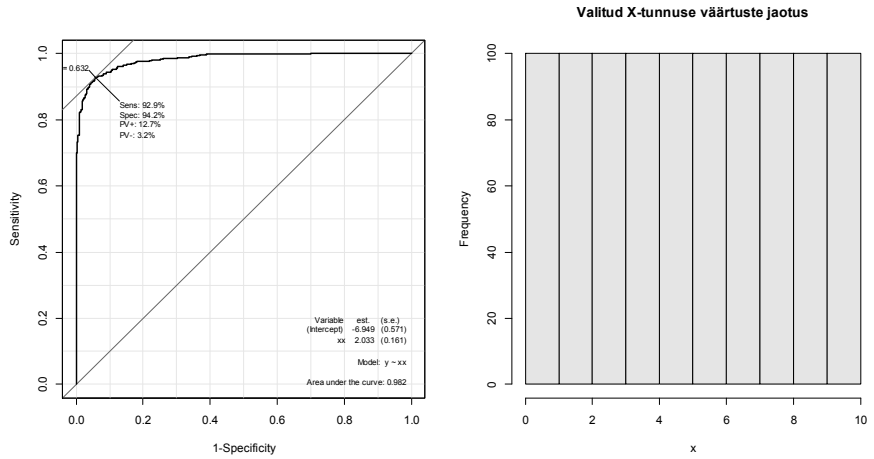
Kui *X*-tunnuse väärtuseid valib uurija…

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 väljavalimist).

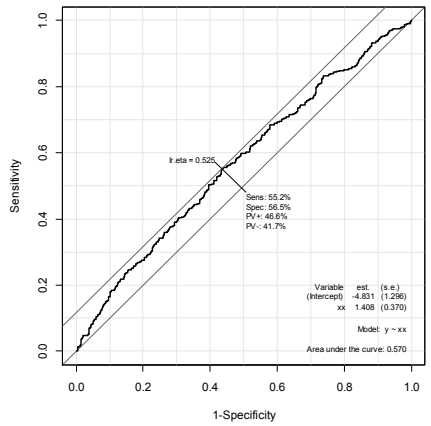
Kui aga näiteks uurija kontrollib *X*-tunnuse väärtuseid (otsustab keda uuringusse kaasata või millises koguses kemikaali kasutada) siis on näiteks uuringu tulemuseks saadav AUC väärtus suuresti uurija enda määrata (ja seega iseloomustab pigem uurijat kui uuritavat).

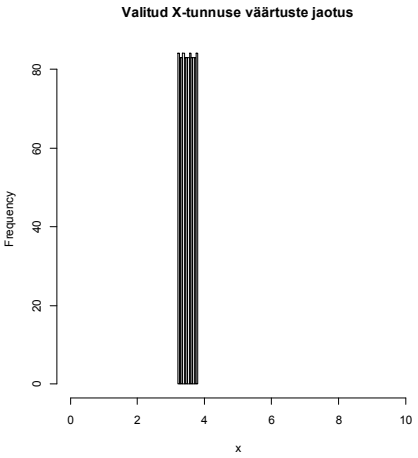
Vaata ka järgnevaid jooniseid, kõigil järgnevatel joonistel on andmeid tekitav mehhanism sama (aga uurija on eksperimendis kasutanud erinevaid *X*-tunnuse väärtuseid…)

AUC=0,982

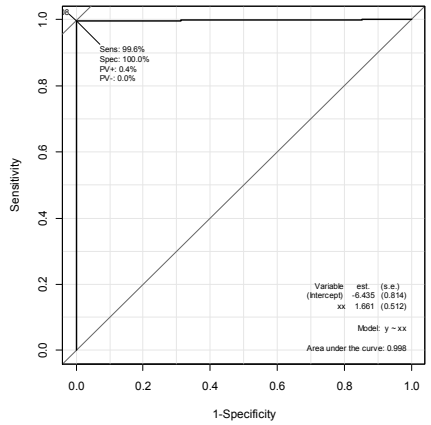


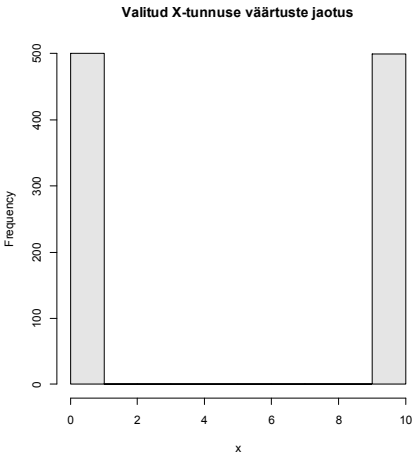
AUC=0,570





AUC=0,998





logistilise regressiooni (ja šansside suhte) populaarsuse tagamaadest

- Enamasti on väga rumal mõte selekteerida valimisse vaatluseid uuritava tunnuse väärtuste põhjal (kui võtaksime valimisse ainult kõrgepalgalisi inimesi, kuidas saaksime siis loota, et valimi keskmine kirjeldaks adekvaatselt inimeste keskmist palka?)

logistilise regressiooni (ja šansside suhte) populaarsuse tagamaadest

- Kui otsustame ise, mitut õnnelikku (1) ja õnnetut (0) uuritavat soovime uuringusse kaasata siis pole võimalik saadud valimi abil kirjeldada õnnelike osakaalu (õnnelikuks olemise tõenäosust) uuritavas populatsioonis; AUC, spetsiifilisus, tundlikkus jne ei kirjelda enam uuritavaid vaid pigem uurijat jne. Aga ühte hinnatud näitajat on siiski võimalik adekvaatselt interpreteerida…
- Nimelt hindab šansside suhe juht-kontrolluuringus (case-control study) sedasama näitajat mis tavalise juhusliku valimi korra.

Mudeli valikust

```
> m1=glm(suits2~olu2+factor(viin)+kaal+factor(sugu),
   family=binomial)
> drop1(m1, test="Chisq")
Single term deletions
Model:
suits2 ~ olu2 + factor(viin) + kaal + factor(sugu)
             Df Deviance    AIC    LRT   Pr(Chi)
<none>            471.54 489.54
olu2          1   493.73 509.73  22.19 2.473e-06 ***
factor(viin)  5   498.64 506.64  27.10 5.463e-05 ***
kaal          1   480.47 496.47   8.92  0.002814 **
factor(sugu)  1   472.75 488.75   1.21  0.270701
```

Kui suudad teadustöö tegemist usaldada arutule masinale, siis:

```
m2=step(m1)
```

Töötab ja kasutatav on ka AIC-käsk

Üle- ja alahajuvusest

Logistilise regressiooni korral eeldatakse, et

$$D(Y) = E(Y) \cdot (1-E(Y)) = p \cdot (1-p)$$

Enamasti on vastava eelduse kontrollimine keeruline.

Vastava eelduse kontroll võib aga osutuda vajalikuks/otstarbekaks siis, kui logistilise regressiooni abil modelleeritakse tunnust, mis pole binaarne.