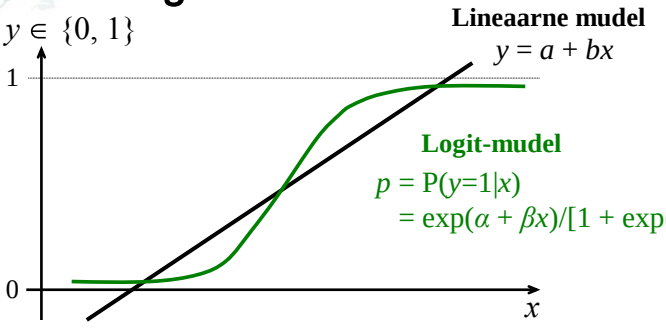
Matemaatiline statistika ja modelleerimine

Binaarsete tunnuste analüüs

EMÜ doktorikool
DK.0007

Tanel Kaart

Logistiline regressioon



- Logistilise regressiooni abil leitud tõenäosuste hinnangud jäävad alati 0 ja 1 vahele.
- Logistilise regressiooni mudeli (*logit*-mudeli) kujusid:

$$p = P(y=1|x) = \exp(\alpha + \beta x)/[1 + \exp(\alpha + \beta x)] = 1 / [1 + \exp(-\alpha - \beta x)],$$
$$\ln[p/(1-p)] = \alpha + \beta x,$$
$$\text{logit}(p) = \alpha + \beta x,$$

- p on huvi pakkuva sündmuse Y toimumise tõenäosus: $p = P(y=1)$,
- $p/(1-p)$ on šansside suhe [*odds ratio*],
- $\ln[p/(1-p)]$ on logaritmiline šansside suhe [*log odds ratio*].

Logistiline regressioon

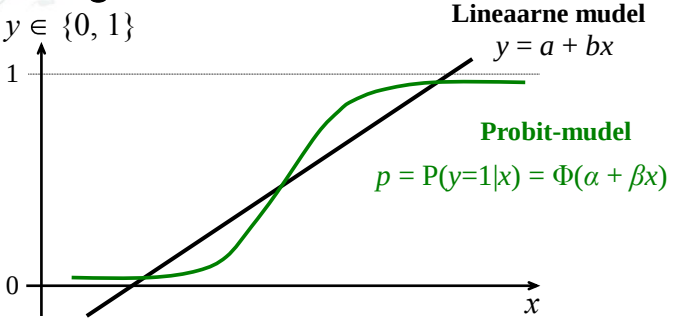
- Tõenäosuse hinnang avaldub kujul:

$$p = \exp(\alpha + \beta x)/[1 + \exp(\alpha + \beta x)] = 1 / [1 + \exp(-\alpha - \beta x)];$$

  - kui $\alpha + \beta x = 0$, siis $p = 0{,}5$
  - $\alpha + \beta x$ suurenemisel $p \to 1$,
  - $\alpha + \beta x$ vähenemisel $p \to 0$.

- Logistilise regressioonivõrrandi kordaja β eksponent, e^{β}, näitab, kui mitu korda muutub sündmuse toimumise šanss (kui mitu korda muutub šansside suhe) argumendi muutumisel ühe ühiku võrra.
  - Näiteks kui $e^{\beta} = 2$, siis kaasneb argumenttunnuse väärtuse suurenemisega 1 võrra sündmuse toimumise šansi kahekordne suurenemine.
  - Negatiivse regressioonikordaja β korral on šansside suhe ühest väiksem, $e^{\beta} < 1$, seega kaasneb argumenttunnuse suurenemisega uuritava sündmuse toimumise šansi vähenemine.

Probit-regressioon

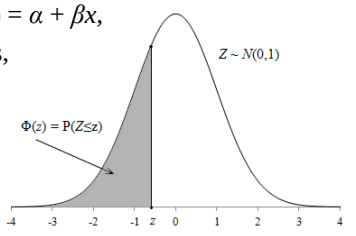


- Probit-regressiooni abil leitud tõenäosuste hinnangud jäävad alati 0 ja 1 vahele.
- Probit-regressiooni mudeli (*probit*-mudeli) kujusid:

$$p = P(y=1|x) = \Phi(\alpha + \beta x) \quad \leftrightarrow \quad \text{probit}(p) = \Phi^{-1}(p) = \alpha + \beta x,$$

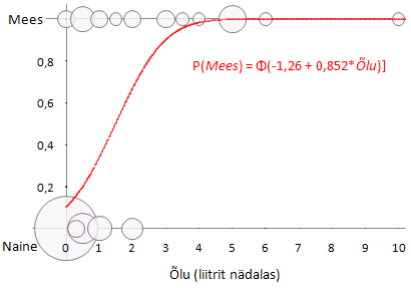
- p on huvipakkuva sündmuse Y toimumise tõenäosus,
- $\Phi(z)$ on standardse normaaljaotuse jaotusfunktsiooni väärtus kohal z:

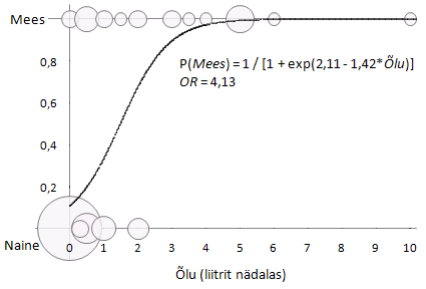
$$\Phi(z) = P(Z \le z), \quad Z \sim N(0,1).$$



Logistiline *versus* probit-regressioon

- Tõenäosuste hinnangutel vahe peaaegu puudub, kumba meetodit kasutada, sõltub enamasti mingis valdkonnas välja kujunenud traditsioonidest.
- Logistilise regressiooni täiendav tulemus on šansside suhe.





LD90, LTemp50, …

- LD90, LTemp50, … on logit- või probit-mudelil baseeruvad argumenttunnuse hinnangulised väärtused, mille korral uuritav sündmus leiab aset ette antud tõenäosusega.
  Näiteks 90%-liselt surmav doos (90% *lethal dose*, LD90)
  või 50%-liselt surmav temperatuur (50% *lethal temperature*, LTemp50).
- Logistilise regressiooni mudeli esitusest $\ln[p/(1-p)] = \alpha + \beta x$ järeldub, et
$$x = \{\ln[p/(1-p)] - \alpha\}/\beta.$$
- Probit-regressiooni mudeli esitusest $\Phi^{-1}(p) = \alpha + \beta x$ järeldub, et
$$x = [\Phi^{-1}(p) - \alpha]/\beta.$$

Supercooling ability and cold hardiness of the pollen beetle *Meligethes aeneus*

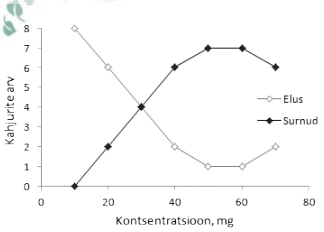
Külli Hiiesaar*, Ingrid H Williams, Marika Mänd, Anne Luik, Katrin Jõgar, Luule Metspalu, Eha Švilponis, Angela Ploomi & Irja Kivimägi
Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Kreutzwaldi St. 1, 51014 Tartu, Estonia

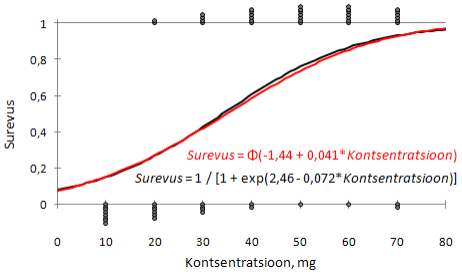
Table 1 Mean (± SE) lethal time [$Ltime_{50}$ (h); time required for 50% mortality at constant -7 °C] and lethal temperature [$Ltemp_{50}$ (°C); temperature required for 50% mortality after 24 h exposure] of *Meligethes aeneus* beetles

Ranking time	n	$Ltime_{50}$	95% fiducial limits	n	$Ltemp_{50}$	95% fiducial limits
June	333	56.2 ± 0.4a	44.3–65.6	231	−8.6 ± 0.2a	−8.7 to −8.0
August	575	80.0 ± 0.2b	73.0–86.4	294	−9.1 ± 0.9b	−9.3 to −8.8
September	406	182.8 ± 0.1c	120.9–347.8	587	−9.8 ± 1.1b	−10.5 to −8.9
January/February	376	418.1 ± 0.2c[1]	336.0–597.5	374	−9.5 ± 0.7b[2]	−10.3 to −8.8

Different letters within a column indicate significant differences between seasonal groups based on the lack of overlap of their 95% fiducial limits.

Logistiline ja probit-regressioon – näide





• Logistilise mudeli kohaselt

$$Surevus = 1 / [1 + \exp(2{,}46 - 0{,}072 * Kontsentratsioon)],$$

mistap 90%-liselt surmav kontsentratsioon

$$LC90 = \{LN[0{,}9/(1-0{,}9)] - 2{,}46\}/0{,}072 = 64{,}60 \text{ mg}.$$

• Probit-mudeli kohaselt

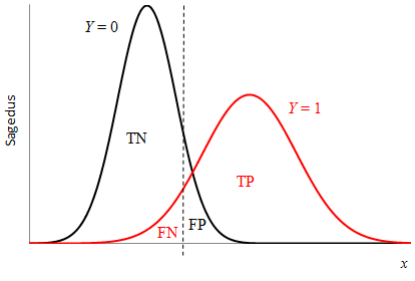
$$Surevus = \Phi(-1{,}44 + 0{,}041 * Kontsentratsioon),$$

mistap 90%-liselt surmav kontsentratsioon

$$LC90 = [\Phi^{-1}(0{,}9) + 1{,}44]/0{,}041 = (1{,}28+1{,}44)/0{,}041 = 65{,}82 \text{ mg}.$$

Testi/mudeli tundlikkus ja spetsiifilisus

- Tõeselt positiivne (*true positive*, TP).
- Valepositiivne (*false positive*, FP).
- Tõeselt negatiivne (*true negative*, TN).
- Valenegatiivne (*false negative*, FN).



Prognoos \ Tegelik olek	$Y = 0$	$Y = 1$
$Y = 0$	TN	FN
$Y = 1$	FP	TP

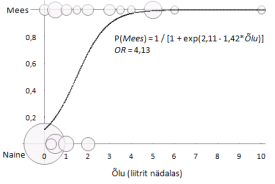
- Tundlikkus (sensitiivsus, *sensitivity*) näitab, kui suure osa (kui mitu protsenti) uuritava sündmuse toimumistest ennustab kasutatud mudel õigesti:

$$\text{Tundlikkus} = TP / (TP+FN).$$

- Spetsiifilisus (*specificity*) näitab, kui suure osa (kui mitu protsenti) uuritava sündmuse mittetoimumistest ennustab kasutatud mudel õigesti:

$$\text{Spetsiifilisus} = TN / (TN+FP).$$

Testi/mudeli tundlikkus ja spetsiifilisus



Näide: tudengi sugu *versus* nädalane õllekogus

Logistiline mudel:

$P(Mees) = 1 / [1 + \exp(2{,}11 - 1{,}42 * Õlu)]$

Õlu (l)	*p*	Mees	Naine	TP	TN	FP	FN	Tundlikkus	Spetsiifilisus
0	0,108	2	27	1	0	1	0	1	0
≤0,3	0,157	0	2	0,909	0,643	0,357	0,091	0,909	0,643
≤0,5	0,198	4	6	0,909	0,690	0,310	0,091	0,909	0,690
≤1	0,334	2	4	0,727	0,833	0,167	0,273	0,727	0,833
≤1,5	0,505	1	0	0,636	0,929	0,071	0,364	0,636	0,929
≤2	0,675	2	3	0,591	0,929	0,071	0,409	0,591	0,929
≤3	0,896	2	0	0,5	1	0	0,5	0,5	1
≤3,5	0,946	1	0	0,409	1	0	0,591	0,409	1
≤4	0,973	1	0	0,364	1	0	0,636	0,364	1
≤5	0,993	5	0	0,318	1	0	0,682	0,318	1
≤6	0,998	1	0	0,091	1	0	0,909	0,091	1
≤10	1,000	1	0	0,045	1	0	0,955	0,045	1
		22	42						

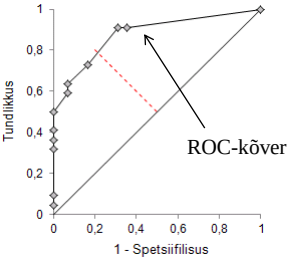
ROC-kõver

- ROC-kõver (*receiver operating characteristic curve*) või üldisemalt ROC-analüüs tähendab testi või mudeli headuse hindamist läbi tundlikkuse ja spetsiifilisuse.

Näide: tudengi sugu *versus* nädalane õllekogus.

Logistiline mudel:

$P(Mees) = 1 / [1 + \exp(2{,}11 - 1{,}42 * Õlu)]$.

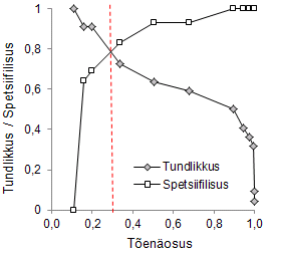


Ligikaudne optimaalne logistilisest mudelist hinnatud tõenäosus, eristamaks õlle joomise alusel naisi ja mehi, on 0,3.

Nii sellele väärtusele vastavad tundlikkus kui ka spetsiifilisus on 0,8 ja

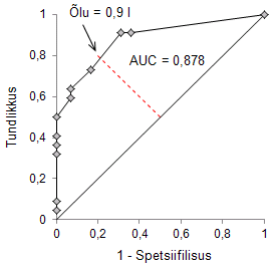
vastav õllekogus on

$õlu_{30} = \{\mathrm{LN}[0{,}3/(1-0{,}3)] - 2{,}11\}/1{,}42 = 0{,}9$ l.



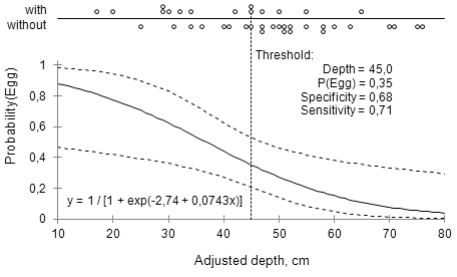
ROC-kõvera alune pindala

- ROC-kõvera alune pindala (*area under the curve*, AUC) on üks testi /mudeli headuse mõõte.
- Kui uuritava sündmuse toimumist ennustada näiteks kulli ja kirja viskamise teel, on ennustuse täpsus 0,5 (50%) – seega vastab AUC = 0,5 juhule, kus mudeli argument mingit rolli ei mängi, ehk mingit seost ei ole. Mida enam erineb AUC 0,5-st, seda täpsemini antud mudel ennustab ehk prognoosib (seda parem on mudel).
- Kokkuleppelised piirid, hindamaks testi/mudeli headust:
  - kui AUC≥0,9, siis on testi/mudeli täpsus suurepärane (*excellent*),
  - AUC≥0,8 puhul hea (*good*),
  - AUC≥0,7 puhul rahuldav (*fair*),
  - AUC≥0,6 puhul kasin (*poor*) ja alla selle ei ole erilist mõtet ennustuse/prognoosi täpsusest rääkida.

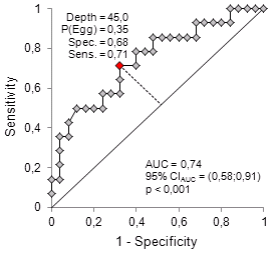


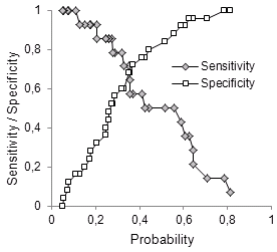
Logistiline regressioon ja ROC-kõver

Näide. Kui sügavad veekogud sobivad mudakonnale (*bufo calamita*) kudemiseks?



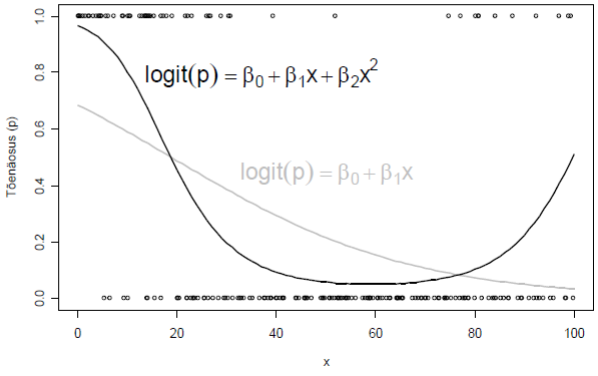
- Logistilise regressiooni ja ROC-kõvera analüüsi alusel on sobivaimaks piiriks eristamaks mudakonnale kudemiseks sobivaid ja mittesobivaid veekogusid veekogu maksimaalne sügavus 45 cm,
- seejuures on 45 cm-st madalamad 71% kudeveekogusid ja 45 cm-st sügavamad 68% kudemiseks mittevalitud veekogusid ning
- kudemiseks sobivaks võib veekogu lugeda juba siis, kui logistilisest mudelist hinnatud kudemiseks sobivuse tõenäosus on üle 0,35;
- $OR = e^{-0,0743} = 0,928$, seega väheneb veekogu sügavuse suurenemisega 1 cm võrra šanss sobida mudakonnale kudemiseks 0,928 korda.





Logistiline regressioon

Mis teha, kui pakutud seos ei sobi?



Joonis Märt Mölsi loengukonspektist

Logistiline mudel kui üldistatud lineaarne mudel (*generalized linear model*)

Mudel hindamaks soo mõju ravi tulemusele

$$\text{logit}(p_{ij}) = \mu + S_i + \varepsilon_{ij},$$

millest

$$p_{ij} = \exp(\mu + S_i)/[1 + \exp(\mu + S_i)].$$

Sugu	Ravi tulemus		
	Terve	Haige	Kokku
Isane	10	2	12
Emane	4	9	13
Kokku	14	11	25

Effect	sugu	Estimate	Standard Error
Intercept		-1.6094	0.7746
sugu	E	2.4204	0.9804
sugu	I	0	.

Seega

$$P(\text{haige}|\text{sugu=I}) = \exp(-1{,}61 + 2{,}42)/[1 + \exp(-1{,}61 + 2{,}42)],$$

$$P(\text{haige}|\text{sugu=E}) = \exp(-1{,}61 + 0)/[1 + \exp(-1{,}61 + 0)]$$

ja

$$OR = \exp(2{,}42) = 11{,}25,$$

st et šanss mitte terveks saada on emaste hulgas 11,25 korda kõrgem.