

Praktikum 7

R – logistiline regressioon.

☞ Salvestage arvutisse Eesti puude andmestik R-i failina:

http://www.eau.ee/~ktanel/DK_0007/puud.rda

☞ Avage R ja käivitage lisamoodul *Rcmdr*.

☞ Võtke R-is kasutusele salvestatud puude andmestik

– *R Commanderis* käsud *Data -> Load data set ...*

– või skripti aknas käsk `load`, näiteks

```
load("C:/Documents and Settings/Tanel/DK_0007/puud.rda")
```

☞ Puude andmestikus vastab üks rida ühele puule (st, et uurimisobjektiks on puu), veergudes on vastavalt:

veerus nimega 'A' puu vanused aastates;

veerus 'D' puu diameetersentimeetrites;

veerus 'H' puu kõrgus meetrites;

veerus 'ARENGUKL' puu arenguklass väärtustega A – lage, N – noorendikud, L – latimets (noorendikust järgmine), K – keskealised, V – valmiv, Y – küps, S – selguseta, – puuduv väärtus;

veerus 'PE' puu liik (HB – haab, KS – kask, KU – kuusk, LV – hall lepp, ...);

veerus 'KKT' kasvukohatüüp (AN – angervaksa, JM – mustika, ...)¹;

...

1.1. Puu raieküpsuse prognoosimine vanuse abil (logistiline regressioon)

Legend. Algajat kasumiahnet metsameest huvitab, millal tema mets valmis saab. Paraku ei suuda ta kuidagi mõista eksperte ja nende hinnanguid metsa küpsusele. Tema tahab lihtsat, ignorandist linnamehele mõistetavat otsustuseeskirja. Siinkohal tuleb appi **logistiline regressioon**.

☞ Esmalt tuleb andmestikku tekitada uus 0-1-tüüpi tunnus, mis näitab, kas mets on küps (uue tunnuse väärtus 1, kui ARENGUKL="Y") või mitte (uue tunnuse väärtus 0, kui ARENGUKL≠"Y").

Selleks tuleb skripti-aknas sisestada käsk ('ready' saab olema uue loodava veeru/tunnuse nimi)

```
puud$ready = 1 * (puud$ARENGUKL=="Y")
```

Alternatiivina võib uue tunnuse tekitada ka *R Commanderi* menüüde abil:

Data -> Manage variables in active dataset -> Compute new variable...

Lahtrisse 'Expression to compute' trükkige (või kopeerige-kleepige) käsk

`1 * (ARENGUKL=="Y")`

¹ Kasvukohatüüp (ka metsakasvukohatüüp) on mullastikult ja taimestikult ühtlane metsaala. Nimetus tuleb enamesineva taime järgi (näiteks kasvab naadi kasvukohas väga palju naate, jänesekapsa kasvukohatüübis jänesekapsaid jne) – <http://et.wikipedia.org/wiki/Kasvukohatüüp>.

Täpsemalt vt näiteks <http://www.hot.ee/sinumets3/sinumets03-08.pdf>.

⌘) Edasiste käskude lihtsama esituse huvides võiks puude andmestiku määrata vaikimisi andmestikuks ka skripti-akna tarvis:

```
attach(puud)
```

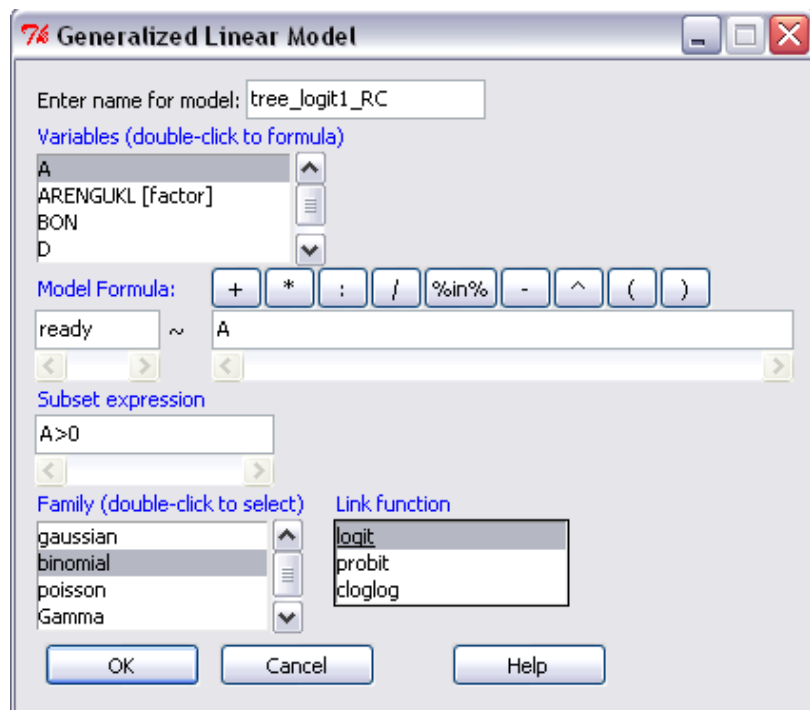
⌘) Püüame nüüd prognoosida puu raieküpsuse tõenäosust vanuse abil.

Logistilise regressiooni teostamiseks nõ käsurealt (mõistlikuma tulemuse saamiseks jätame välja üksikud puud, kelle vanusena on kirjas 0):

```
tree_logit1 = glm(ready ~ A, family=binomial(), data=puud[A>0,])
summary(tree_logit1)
```

Sama mudeli saab tellida ka *R Commanderi* menüüdest (uut tunnust 'ready' sisaldava andmestiku *R Commanderis* kasutamiseks võib osutuda vajalikuks selle andmestiku uuesti *R Commanderi* vaikimisi andmestikuks määramine).

Statistics -> Fit models -> Generalized linear model



Väljavõtt tulemustest:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.5590724  0.0703370  -79.03  <2e-16 ***
A             0.0530644  0.0008259   64.25  <2e-16 ***
```

Vanusele vastav positiivne regressioonikordaja näitab, et puude vanuse kasvades suureneb ka nende raieküpsuse tõenäosus. Kusjuures eksponent regressioonikordajast, $\exp(0,053) = 1,054$, hindab šansside suhet *OR* (*odds ratio*; näitab raieküpsuse šansi suurenemist puu vanuse suurenemisel ühe aasta võrra).

R-i poolt hinnatud logistiline prognoosivõrrand avaldub hinnatud parameetrite kaudu kujul

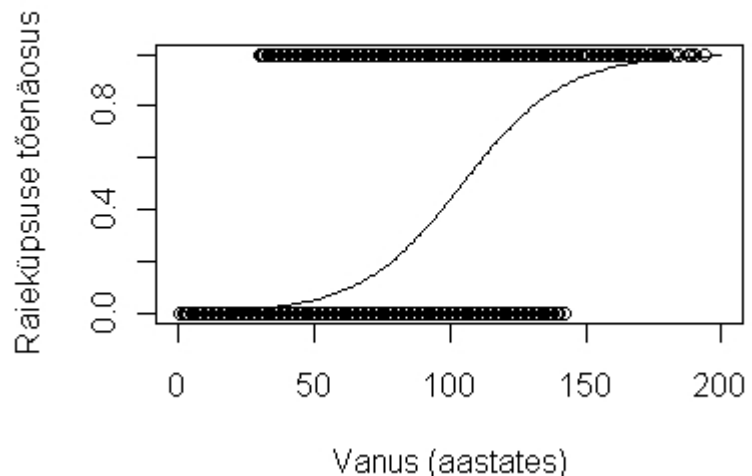
$$P(\text{ready}=1 \mid A) = \exp(-5,559 + 0,053 \times A) / (1 + \exp(-5,559 + 0,053 \times A))$$

(prognoositakse tõenäosust puu raieküpsuseks teada oleva vanuse korral).

Tulemused graafiliselt (laseme *R*-l prognoosida raieküpsuse tõenäosust vanuste 0-200 korral):

```
x = seq(0,200)
y = predict(tree_logit1, data.frame(A=x), type="response")
plot(x, y, type="l", xlab="Vanus (aastates)", ylab="Raieküpsuse tõenäosus")
points(A, ready)
```

Viimane rida eelnevas programmis lisab joonisele punktidenägemise tegelikud (empiirilised) väärtused.



Veel võimalusi illustreerimaks logistilise regressioonanalüüsi tulemusi:

```
par(mfrow=c(2,1))
hist(A[ready==1], xlim=c(0,200))
hist(A[ready==0], xlim=c(0,200))
par(mfrow=c(1,1))
```

Märkus.

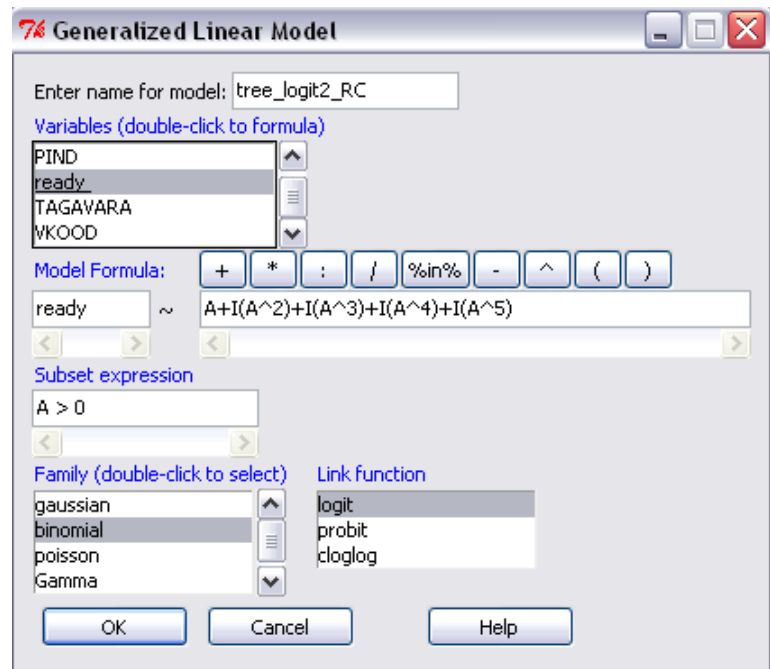
Metsa raieküpsuse saavutamise aeg on ilmselt ülehinnatud (80%-lise raieküpsuse saavutamiseks kulub prognoosi kohaselt ligikaudu 125 aastat) – põhjuseks see, et raieküpsuse saavutanud mets koheselt ka raiutakse, mistõttu nende andmed (raieküpsus teatud vanuses) andmebaasis ei kajastu. Alles jäävad vaid metsad, mis mingil põhjusel pole vaatamata suurele vanusele raieküpseks saanud ja ehk ka üksikud kaitsealused metsad.

1.2.

Kuna linnamees soovib oma metsa raieküpsust täpsemalt prognoosida, üritab ta teha oma logistilise regressiooni mudelit keerukamaks:

```
tree_logit2 = glm(ready~A+I(A^2)+I(A^3)+I(A^4)+I(A^5), family=binomial(), data=puud[A>0,])
summary(tree_logit2)
```

Sama *R Commanderis*:

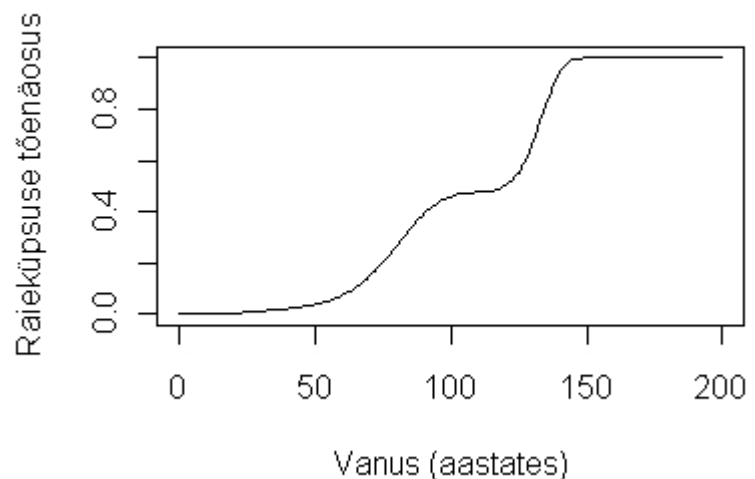


Raieküpsuse prognoos graafiliselt:

```
x = seq(0,200)
y = predict(tree_logit2, data.frame(A=x), type="response")
plot(x, y, type="l", xlab="Vanus (aastates)", ylab="Raieküpsuse tõenäosus")
```

Millega on seletatav jõnks graafikul?

Vastus – tegu on vanusevahemikuga, millal raiutakse metsi väga intensiivselt (alles on jäänud vaid mingil põhjusel raieküpsust mitte saavutanud puud).



1.3.

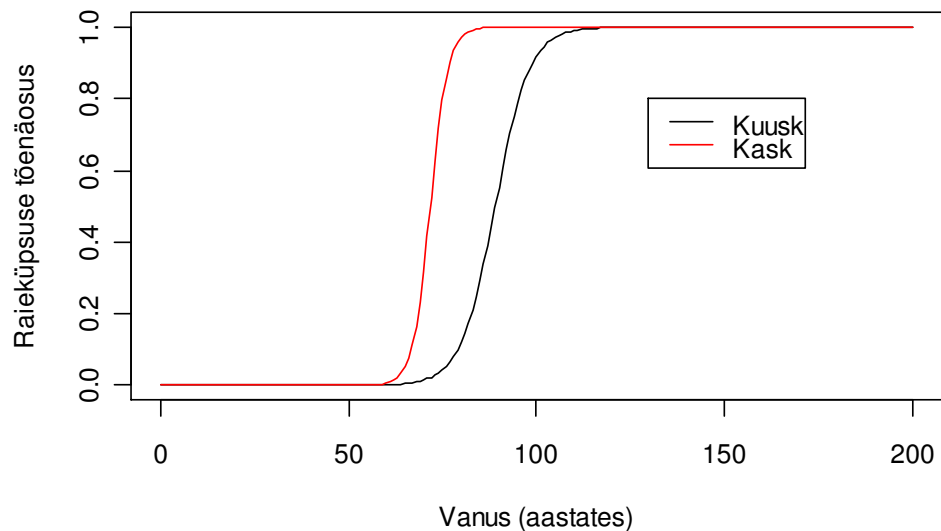
Siiani analüüsisime kõiki puuliike koos, samas on loomulik eeldada, et eri liigid saavutavad raieküpsuse erineval ajal. Seega võiks püüda hinnata erinevat mudelit iga puuliigi tarvis ...

```
tree_logit3 = glm(ready~A*factor(PE), family=binomial(), data=puud[A>0,])
```

Analüüsi tulemusena saadud hoiatus tähendab, et osade puuliikide tarvis ei olnud võimalik tellitud regressioonivõrrandit hinnata. Andmestikus suurel hulgal esindatud puuliikidega ei tohiks aga probleeme olla.

Joonistame hinnatud mudeli abil graafikud kahe puuliigi – kuuse ja kase jaoks:

```
x = seq(0,200)
yKU = predict(tree_logit3, data.frame(A=x, PE="KU"), type="response")
yKS = predict(tree_logit3, data.frame(A=x, PE="KS"), type="response")
plot(x, yKU, type="l", xlab="Vanus (aastates)", ylab="Raieküpsuse tõenäosus")
lines(x, yKS, col="red")
legend(130, 0.8, c("Kuusk", "Kask"), col=c("black", "red"), lty=1)
```



Tulemus on üsna ootuspärane – kased saavutavad raieküpsuse varem kui kuused. Seejuures on kaskede raieküpsus vanusega pisut täpsemalt määratav (logistilise kõvera tõus on järsem).

1.4.

Logistilist regressiooni võib kasutada ka vaatlusandmete lahterdamiseks: kui puu on 75 aastat vana, siis tõenäoliselt on ta raieküps, kui aga puu vanus on 60 aastat, siis arvatavasti ta veel raieküps ei ole.

Et mistahes prognoosid on harva täiesti täpsed, võib vigu esineda ka taolisel lahterdamisel.

Prognostiliste testide täpsust kirjeldatakse sageli kahe suurusega – testi tundlikkusega ja spetsiifilisusega.

Testi **tundlikkus** (*sensitivity*) näitab, kui suure tõenäosusega tuvastab test toimunud sündmuse (a'la kui suure tõenäosusega leitakse haigel indiviidil haigus).

Testi **spetsiifilisus** (*specificity*) näitab, kui suure tõenäosusega tuvastab test õigesti sündmuse mittetoimumise (a'la terve indiviid prognoositakse terveks).

Näiteks puude raieküpsuse analüüsil on loomulik otsustada, et kui mudel prognoosib raieküpsuse tõenäosuseks üle 0,5, siis loeme puu (metsa) raieküpseks.

Uurimaks sellise testi tundlikkust ja spetsiifilisust, tuleb konstrueerida tabel, mis näitaks, kui täppi taolise reegli alusel tehtud otsustused lähevad.

Järgnevas programmis defineeritud muutuja `kyps` sisaldab kõigi andmestikku kuuluvate puude raieküpsuse prognoose mudelist `tree_logit3`.

Semikoolonit kasutatakse eraldamiseks ühte ritta kirjutatud erinevaid käske

```
kyps=predict(tree_logit3, type="response")
tabel=table(kyps>0.5, ready[A>0]); tabel
spets = tabel[1,1]/sum(tabel[,1])
tund = tabel[2,2]/sum(tabel[,2])
spets; tund
```

```
> tabel
      ready
      0    1
FALSE 21926 1002
TRUE   532  3210
```

```
> specs; tund
[1] 0.9763113
[1] 0.7621083
```

Kõigist raieküpsetest puudest prognoositakse raieküpseks 3210 ja mitteraieküpseks 1002. Järelikult on puu vanuse ja liigi alusel õigesti raieküpseks määratud puude osakaal $3210/(3210+1002) = 0,762$ (76,2 %).

Saadud arv kujutab enesest raieküpsuse testi tundlikkust.

Kõigist mitteraieküpsetest puudest prognoosib logistiline mudel õigesti mitteraieküpseteks 21926 ja valesti raieküpseteks 532. Seega on raieküpsuse testi spetsiifilisus $21926/(21926+532) = 0,976$.

1.5.

Viimati konstrueeritud mudel hindas küll raieküpsuse ja vanuse vahelise seose kõigile puuliikidele eraldi, aga prognoosid ning testi tundlikkuse ja spetsiifilisuse saime ikkagi kõigile puuliikidele korraga. Kuigi ka eelneva analüüsi tulemustest on välja nopitavad tulemused puuliikide kaupa, on vahel lihtsam teha uus analüüs üksnes mingi alamandmestiku põhjal.

Hindame järgnevalt vaid kuuskede ja kaskede (eraldi) raieküpsust nende vanuse alusel ning leiame testide tundlikkuse ja spetsiifilisuse.

Esmalt moodustama puude andmestikust üksnes teadaoleva vanusega kuuski või kaski sisaldavad alamandmestikud ja teostame analüüsid nendega.

```
kuusk = subset(puud, subset = A>0 & PE=="KU")
kask = subset(puud, subset = A>0 & PE=="KS")
kuusk_logit = glm(ready~A, family=binomial(), data=kuusk)
kask_logit = glm(ready~A, family=binomial(), data=kask)
```

Tulemusi võib soovi korral illustreerida jälle ka pildiga, mis tuleb identne paar lehekülge tagasi konstrueerituga.

```
x = seq(0,200)
yKU = predict(kuusk_logit, data.frame(A=x), type="response")
yKS = predict(kask_logit, data.frame(A=x), type="response")
plot(x, yKU, type="l", xlab="Vanus (aastates)", ylab="Raieküpsuse tõenäosus")
lines(x, yKS, col="red")
legend(130, 0.8, c("Kuusk", "Kask"), col=c("black","red"), lty=1)
```

Leiame ka testide tundlikkuse ja spetsiifilisuse.

```
kyps_kuusk=predict(kuusk_logit, type="response")
tabel_kuusk=table(kyps_kuusk>0.5, kuusk$ready); tabel_kuusk
spets_kuusk = tabel_kuusk[1,1]/sum(tabel_kuusk[,1])
tund_kuusk = tabel_kuusk[2,2]/sum(tabel_kuusk[,2])
spets_kuusk; tund_kuusk
```

```
kyps_kask=predict(kask_logit, type="response")
tabel_kask=table(kyps_kask>0.5, kask$ready); tabel_kask
spets_kask = tabel_kask[1,1]/sum(tabel_kask[,1])
tund_kask = tabel_kask[2,2]/sum(tabel_kask[,2])
spets_kask; tund_kask
```

```
> spets_kuusk; tund_kuusk
[1] 0.9860474
[1] 0.8386207
```

```
> spets_kask; tund_kask
[1] 0.9786216
[1] 0.8607955
```

Tulemustest võib järeldada, et kui mitteraieküpsuse prognoosimise täpsus (testi spetsiifilisus) on kuuski ja kaski eraldi vaadates enam-vähem sama suur, kui kõiki puid koos vaadates – ligikaudu 98%, siis raieküpsuse prognoosimise osas (testi tundlikkus) annab mudel nii kuuskede kui ka kaskede puhul paremad tulemused kui kõigil puudel kokku – raieküpsetest kaskedest saab vaid nende vanuse järgi õigesti tuvastada 86%, raieküpsetest kuuskedest 84%.

1.6.

Aga, kas hinnatud tõenäosus 0,5 on see kõige õigem piir klassifitseerimaks puid raieküpsseteks ja mitteraieküpsseteks?

Optimaalseima lõikepunkti (*cutpoint*, prognoositud raieküpsuse tõenäosus, millest alates puu ka tegelikult raieküpsseks lugeda) leidmiseks on kasutatavad II Maailmasõja ajal inglase poolt radarisignaalidest omade ja võõraste lennukite eristamise algoritmide testimiseks välja töötatud **ROC-kõverad** (ROC = *Receiver Operating Characteristic*).

Kuigi *R*-s on olemas mitmeid vastavate analüüside jaoks mõeldud pakette, ei sisaldu neist ühtki vaikumisi koos *R*-ga (ja *R Commander*-iga) installeeritavate pakettide hulgas. Siiski on ROC-kõverad konstrueeritavad ka ilma vastavate pakettideta, lihtsalt hüpoteeside testimine ja optimaalseima lõikepunkti headust mõõtvate karakteristikute hindamisest tuleb loobuda.

Järgnev programm teostab vajalikud arvutused ja joonistab välja ROC-kõverad kuuskede raieküpsuse tarvis.

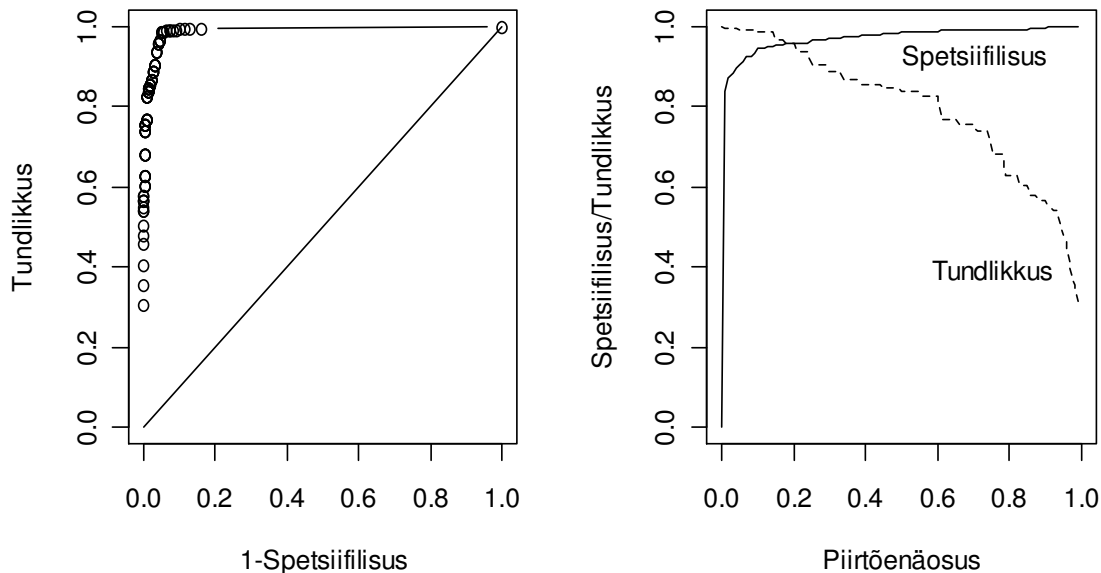
```
fitkuusk=kuusk_logit$fitted.values
kuusk_punkt=seq(min(fitkuusk), 1, by=0.01)
nn=length(kuusk_punkt)
tun_kuusk=rep(NA,nn); spe_kuusk=rep(NA,nn)
for (i in 1:nn) {
  kuusk_progn=(fitkuusk>kuusk_punkt[i])
  kuusk_tabel=table(kuusk_progn, kuusk$ready)
  tun_kuusk[i]=kuusk_tabel[2,2]/sum(kuusk_tabel[,2])
  spe_kuusk[i]=kuusk_tabel[1,1]/sum(kuusk_tabel[,1]) }

par(mfrow=c(1,2))
plot(1-spe_kuusk, tun_kuusk, type="b", xlab="1-Spetsiifilisus",
     ylab="Tundlikkus", ylim=c(0,1), xlim=c(0,1))
lines(c(0,1), c(0,1))
plot(kuusk_punkt, spe_kuusk, type="l", xlab="Piirtõenäosus",
     ylab="Spetsiifilisus/Tundlikkus", ylim=c(0,1), xlim=c(0,1))
lines(kuusk_punkt,tun_kuusk, lty=2)
text(0.7, 0.93, "Spetsiifilisus")
text(0.75, 0.4, "Tundlikkus")
par(mfrow=c(1,1))
```

Tulemuseks on kaks joonist (vt järgmine lk), millest vasakpoolne illustreerib tundlikkuse ja spetsiifilisuse vahetõenäosuse kõigi võimalike lõikepunktide korral (minimaalsest hinnatud raieküpsuse tõenäosusest kuni üheni sammuga 0,01). Diagonaalil paiknev joon märgib olukorda, mille korral test mõtet ei oma, reaalseste lõikepunktide alusel välja joonistatud kõvera kaugus diagonaalist näitab testi headust. Antud juhul paiknevad punktid peaaegu täisnurgana (peaaegu maksimaalselt eemal

diagonaalist), mis tähendab testi väga head sobivust – kuuskede vanus võimaldab suurepäraselt prognoosida kuuskede raieküpsust.

Parempoolsel joonisel on vertikaalteljel tundlikkus ja spetsiifilisus ning horisontaalteljel võimalikud raieküpsuse tõenäosused. Punkt, kus tundlikkuse ja spetsiifilisuse graafikud lõikuvad, märgib optimaalseimat testi tulemust (logistilise regressiooniga hinnatud tõenäosust, millest alates lugeda puu raieküpseks). Ligikaudu on vastav tõenäosus 0,2.



Kontrollime, kas piirtõenäosus 0,2 annab paremad tundlikkuse ja spetsiifilisuse väärtused, kui eelnevalt kasutatud väärtus 0,5.

```
kyps_kuusk=predict(kuusk_logit, type="response")
tabel_kuusk=table(kyps_kuusk>0.2, kuusk$ready); tabel_kuusk
spets_kuusk = tabel_kuusk[1,1]/sum(tabel_kuusk[,1])
tund_kuusk = tabel_kuusk[2,2]/sum(tabel_kuusk[,2])
spets_kuusk; tund_kuusk
```

```
> tabel_kuusk=table(kyps_kuusk>0.2, kuusk$ready); tabel_kuusk
```

	0	1
FALSE	5222	31
TRUE	225	694

Tõepoolest, spetsiifilisuse väärtus küll 2,8% võrra langes, seevastu tundlikkus suurenes peaaegu 12% (0,839-lt 0,957-ni).

Seega oleks mõistlik lugeda kuusk raieküpseks juba siis, kui

logistilise regressiooni võrrandist vanuse alusel leitud prognoos raieküpsuse tõenäosusele on >0,2.

```
> spets_kuusk; tund_kuusk
```

[1]	0.9586929
[1]	0.9572414

PS. Kui te tahate teada, milline see kuuskede raieküpsuse tõenäosusele hinnatud logistilise regressiooni võrrand ikka on, siis tuleb rakendada käsku `summary` kuuskedele sobitatud mudelile:

`summary(kuusk_logit)`

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-19.64508	0.87718	-22.40	<2e-16 ***
A	0.22066	0.01005	21.96	<2e-16 ***

Seega

$$P(\text{kuusk vanuses } A \text{ on raieküps}) = \exp(-19,645 + 0,221 \times A) / (1 + \exp(-19,645 + 0,221 \times A))$$

ja kuuse vanuse suurenemisel 1 aasta võrra suureneb šanss osutada raieküpseks $\exp(0,221) = 1,25$ korda.