Kursus: Mitmemõõtmeline statistika

# Seminar V: Seoste analüüs

# Logistiline regressioon

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

TALLINNA ÜLIKOOL

## Logistiline regressioon

Lineaarse regressiooni üldistus, kus:

DV – grupeeriv tunnus
IV – sisuliselt mistahes tüüpi tunnus(ed)

Eesmärk:

- välja selgitada ning kirjeldada optimaalse mudeli abil, millised tegurid ja mil määral mõjutavad/on_seotud objektide jagunemisega eri gruppidesse
- prognoosida ühte või teise gruppi kuulumist

NB! Mudel võib sisaldada IV'de koosmõjusid

TALLINNA ÜLIKOOL

## Logistiline regressioon - eeldused

Eeldused tunnuste jaotuse ning seose kuju/iseloomu kohta praktiliselt puuduvad.

Siiski on oluline pöörata tähelepanu mõnele aspektile:

- objektide arvu ja mudelisse võetud tunnuste arvu suhe - liiga palju ”hõredaid” lahtreid/gruppe võib põhjustada mitmeid probleeme
- Multikollineaarsus e IV’de omavaheline tugev seotus ei ole soovitav
- Ekstreemsed väärtused e antud juhul objektid, kes tegelikult kuluvad ühte gruppi, kuid on IV’de väärtuste poolest hoopis sarnasemad (mõne) teise grupi objektidega, mõjutavad mudelit ja selle parameetreid (vt jääkliikmeid!)

TALLINNA ÜLIKOOL

## Logistiline regressioon – mida IV’de lineaarkombinatsioon prognoosib?

Kuna DV on grupeeriv, siis ei saa otse DV väärtust prognoosida!

Kasutajale kõige lihtsamini tajutav ja tõlgendatav tulem oleks prognoosida:

valitud gruppi kuulumise tõenäosust *p*

Matemaatiliselt on parem/kasulikum koostada IV’de lineaarkombinatsioon st mudel nii, et ta prognoosiks:

suhtelise tõenäosuse (naturaal)logaritmi e **logit-suhet**

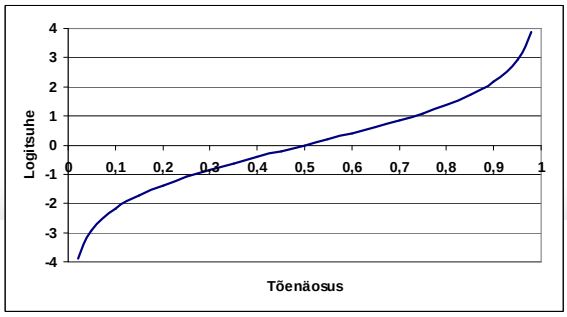
$$\ln \frac{p}{1-p}$$

TALLINNA ÜLIKOOL

## Logistiline regressioon – mida IV'de lineaarkombinatsioon prognoosib?

Suhteline tõenäosus – valitud gruppi kuulumise šansid antud gruppi mitte kuulumisega võrreldes:
kui tõenäosus on 50%, siis šanss 1/1 ehk 1
kui tõenäosus <50%, siis šanss 1/mitmele ehk 0,...
kui tõenäosus >50%, siis šanss mitu/1 ehk >1

Logitsuhte $\ln\frac{p}{1-p}$ ja (gruppi kuulumise) tõenäosuse seos:




TALLINNA ÜLIKOOL

## Logistilise regressiooni võrrand

$$\ln\frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

$b_0$ vabaliige (logitsuhte väärtus, kui kõigi IV'de väärtused =0)
$b_i$ regressioonikordaja (logitsuhte muut, kui antud IV muutub 1 ühiku võrra ja teised IV'd jäävad samaks)

Kuna tõenäosus ja logitsuhe on üks-üheselt seotud, siis saab logistilist regressioonivõrrandit teisendades välja arvutada ka gruppi kuulumise tõenäosuse prognoosi:

$$\hat{p} = \frac{e^{b_0+b_1X_1+b_2X_2+\ldots+b_mX_m}}{1+e^{b_0+b_1X_1+b_2X_2+\ldots+b_mX_m}}$$

TALLINNA ÜLIKOOL

## Logistiline regressioon – sõltuv muutuja

$$\ln\frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

DV –> grupeeriv tunnus:

* Binaarne tunnus (e 2 gruppi) –> binaarne logistiline regressioonimudel
* Kategoriaalne tunnus (3 või enam gruppi) –> multinomiaalne logistiline regressioonimudel

(valitakse **taustakategooria** e võrdlusgrupp ja koostatakse mitu binaarse logistilise regressioonimudeliga analoogset mudelit, millest igaüks eristab parimal võimalikul viisil üht gruppi taustagrupist)

$$\ln\frac{p}{1-p} \Rightarrow \ln\frac{p_{antud\_kategooria}}{p_{taustakategooria}}$$

NB! Ka kahe grupi korral võib teha multinomiaalse mudeli!

NB! SPSS'is on binaarse ja multinomiaalse mudeli jaoks erinevad lisavõimalused!

TALLINNA ÜLIKOOL

## Logistiline regressioon – sõltumatud muutujad

$$\ln\frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

$X_i$ ehk IV –> sisuliselt mistahes tüüpi tunnus(ed), aga ... lineaarkombinatsiooni loogika ei toimi, kui skaala vahemikud pole võrdsed või väärtused pole isegi järjestatavad!!!

Arvtunnused – võib panna mudelisse nagu on (vahest siiski tehakse grupeerivaks tunnuseks: nt. noored, keskealised, eakad)

Tunnused, mis on mõõdetud diskreetsete kategooriatena, st järjestustunnused ja nimitunnused kodeeritakse ümber binaarsete abitunnuste grupiks
X(eesti; soome; inglise; vene) ->
$X_{eesti}$(1;0) ja $X_{soome}$(1;0) ja $X_{inglise}$(1;0)

NB! Taustkategooria - regressioonikordajad näitavad prognoosi muutu taustgrupi suhtes vastavasse kategooriasse kuulumisel

NB! SPSS'is kodeeritakse tunnused ümber automaatselt!

TALLINNA ÜLIKOOL

## Logistiline regressioon – näide "õpetajad"

DV –> Kas olete viimase poolaasta jooksul kolleegidega suheldes tundnud rahulolu oma esinemise ja selle eest saadud kiituse või tänuga?
(1-pole tundnud; 2-korra olen tundnud; 3-palju kordi olen tundnud)
NB! Binaarseks tunnuseks ümberkodeerituna:
(1-pole üldse tundunud või ainult korra; 2-palju kordi olen tundnud)
**taustakategooria** NB! **"riski"rühm**

IV –> Keel (eesti, vene)
Vanus
Kooliaste (algklassid, keskaste, keskkool)
Saavutan teiste inimestega kergesti kontakti (Likerti 5-palli skaala)
Suudan lahendada konflikte (Likerti 5-palli skaala)

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| pole üldse või ainult korra tundnud | 0 |
| palju kordi olen tundnud | 1 |

**Categorical Variables Codings**

| | | Frequency | Parameter coding (1) | Parameter coding (2) |
|---|---|---|---|---|
| aste | algklassid | 150 | 1,000 | ,000 |
| | keskaste | 134 | ,000 | 1,000 |
| | keskkool | 64 | ,000 | ,000 |
| Emakeel | eesti | 200 | 1,000 | |
| | vene | 148 | ,000 | |

TALLINNA ÜLIKOOL

## Log.reg. – võrrandi kordajad ja nende stat.olulisus

$$\ln\frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2 + ... + b_mX_m$$

B regressioonikordaja

S.E. Regressioonikordaja standardviga

Wald Wald'i statistikut kasutatakse kordaja stat.olulisuse testimiseks
$(B/S.E.)^2$ => hii-ruut jaotusega

df vabadusastmete arv

Sig. olulisustõenäosus (I tüüpi vea tõenäosus)

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| [a] | keel(1=eesti) | ,580 | ,243 | 5,693 | 1 | ,017 | 1,786 |
| | vanus | ,014 | ,010 | 1,849 | 1 | ,174 | 1,014 |
| | aste | | | 9,051 | 2 | ,011 | |
| | aste(1=algkool) | ,181 | ,322 | ,315 | 1 | ,575 | 1,198 |
| | aste(2=põhikool) | -,603 | ,330 | 3,342 | 1 | ,068 | ,547 |
| | s182 (kontakt) | ,427 | ,175 | 5,912 | 1 | ,015 | 1,532 |
| | s184 (konflikt) | ,422 | ,181 | 5,444 | 1 | ,020 | 1,525 |
| | Constant | -4,455 | ,870 | 26,235 | 1 | ,000 | ,012 |

[a] Variable(s) entered on step 1: keel, vanus, aste, s182, s184.

TALLINNA ÜLIKOOL

## Logistiline regressioon – riskisuhe

Exp(B) **riskisuhe** (*odds ratio*) =>

=>"riski"rühma kuulumise ja mitte kuulumise suhte e võimaluse/šansside muut vastavalt sõltumatu muutuja väärtuse kasvule ühe ühiku võrra (binaarse IV puhul antud gruppi kuulumisel võrreldes taustgruppi kuulumisega)

$$Exp(B) = \frac{N_{"riski"rühm}}{N_{taustakat.}}(IVtasemel\_j) \div \frac{N_{"riski"rühm.}}{N_{taustakat.}}(IVtasemel\_j+1/IVtaustagrupis)$$

Kuna on tegu jagatisega, siis:

1 seos puudub / samad šansid "riski"rühma kuuluda

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| a | keel(1=eesti) | ,580 | ,243 | 5,693 | 1 | ,017 | 1,786 |
| | vanus | ,014 | ,010 | 1,849 | 1 | ,174 | 1,014 |
| | aste | | | 9,051 | 2 | ,011 | |
| | aste(1=algkool) | ,181 | ,322 | ,315 | 1 | ,575 | 1,198 |
| | aste(2=põhikool) | -,603 | ,330 | 3,342 | 1 | ,068 | ,547 |
| | s182 (kontakt) | ,427 | ,175 | 5,912 | 1 | ,015 | 1,532 |
| | s184 (konflikt) | ,422 | ,181 | 5,444 | 1 | ,020 | 1,525 |
| | Constant | -4,455 | ,870 | 26,235 | 1 | ,000 | ,012 |

a. Variable(s) entered on step 1: keel, vanus, aste, s182, s184.

TALLINNA ÜL

## Log.reg. – näide "õpetajad" – riskisuhte tõlgendamine

>1 šansid suurenevad/on_suuremad

(palju kordi rahulolu tundmise ja rahulolu mitte tundmise suhe ehk "risk"/võimalus palju kordi rahulolu tunda võrreldes rahulolu mitte tundmisega on eesti õpetajatel 1,786 korda e 78,6% suurem kui vene õpetajatel)

("risk"/võimalus palju kordi rahulolu tunda võrreldes rahulolu mitte tundmisega kasvab 1,532 korda e 53,2% kui teistega hea kontakti saavutamise hinnang suureneb ühe palli võrra)

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| a | keel(1=eesti) | ,580 | ,243 | 5,693 | 1 | ,017 | 1,786 |
| | vanus | ,014 | ,010 | 1,849 | 1 | ,174 | 1,014 |
| | aste | | | 9,051 | 2 | ,011 | |
| | aste(1=algkool) | ,181 | ,322 | ,315 | 1 | ,575 | 1,198 |
| | aste(2=põhikool) | -,603 | ,330 | 3,342 | 1 | ,068 | ,547 |
| | s182 (kontakt) | ,427 | ,175 | 5,912 | 1 | ,015 | 1,532 |
| | s184 (konflikt) | ,422 | ,181 | 5,444 | 1 | ,020 | 1,525 |
| | Constant | -4,455 | ,870 | 26,235 | 1 | ,000 | ,012 |

a. Variable(s) entered on step 1: keel, vanus, aste, s182, s184.

TALLINNA ÜLIK

## Log.reg. – näide "õpetajad" – riskisuhte tõlgendamine

0,... šansid vähenevad/on_väiksemad

("risk"/võimalus palju kordi rahulolu tunda võrreldes rahulolu mitte tundmisega on põhikooli õpetajatel võrreldes keskkooli õpetajatega väiksem
=> võib ümber pöörata: 1:0,547=1,828
"risk"/võimalus palju kordi rahulolu tunda võrreldes rahulolu mitte tundmisega on keskkooli õpetajatel 1,828 korda e 82,8% suurem kui põhikooli õpetajatel)

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| a | keel(1=eesti) | ,580 | ,243 | 5,693 | 1 | ,017 | 1,786 |
| | vanus | ,014 | ,010 | 1,849 | 1 | ,174 | 1,014 |
| | aste | | | 9,051 | 2 | ,011 | |
| | aste(1=algkool) | ,181 | ,322 | ,315 | 1 | ,575 | 1,198 |
| | aste(2=põhikool) | -,603 | ,330 | 3,342 | 1 | ,068 | ,547 |
| | s182 (kontakt) | ,427 | ,175 | 5,912 | 1 | ,015 | 1,532 |
| | s184 (konflikt) | ,422 | ,181 | 5,444 | 1 | ,020 | 1,525 |
| | Constant | -4,455 | ,870 | 26,235 | 1 | ,000 | ,012 |

a. Variable(s) entered on step 1: keel, vanus, aste, s182, s184.

TALLINNA ÜLI

## Logistiline regressioon – mudeli "headus"

Hindamise aluseks **tõepärafunktsioon** –
lähtub sõltuva tunnuse väärtuste tõenäosusest valimis => kõige optimaalsemad kordajad annavad suurima tõepärafunktsiooni väärtuse (lähim tõenäosusele 1); sisuliselt sellise tulemi, mis on parimal viisil kooskõlas tegelikult toimunuga.

Matemaatilistel kaalutlustel esitatakse tõepärafunktsioon logaritmituna ja -2'ga läbi korrutatuna (*-2 Log Likelihood*)

=> seetõttu tõepärafunktsiooni maksimeerimine võrdub funktsiooni *-2 Log Likelihood* minimeerimisega e mida väiksem *-2 Log Likelihood* näitaja, seda parem mudel!

**Model Fitting Information**

| Model | Model Fitting Criteria -2 Log Likelihood | Likelihood Ratio Tests Chi-Square | df | Sig. |
|---|---|---|---|---|
| Intercept Only | 440,851 | | | |
| Final | 398,011 | 42,840 | 6 | ,000 |

TALLINNA ÜLIKOOL

## Logistiline regressioon – mudeli "headus"

Mudeli statistiline olulisus e usaldusväärsus

=> Võrreldakse sõltumatute muutujatega mudelit sellise mudeliga, kus on ainult vabaliige e sisuliselt sellise mudeliga, mille puhul kõigile objektidele prognoositakse üht ja sama (st valimi üldist) tõenäosust

=> kui tõepärafunktsiooni või täpsemini *-2 Log Likelihood* muut(umine) (vt hii-ruut näitaja) on nullist erinev statistiliselt olulisel määral, loetakse mudel tervikuna statistiliselt oluliseks e usaldusväärseks

**Model Fitting Information**

| Model | Model Fitting Criteria | Likelihood Ratio Tests | | |
|---|---|---|---|---|
| | -2 Log Likelihood | Chi-Square | df | Sig. |
| Intercept Only | 440,851 | | | |
| Final | 398,011 | 42,840 | 6 | ,000 |

TALLINNA ÜLIKOOL

## Logistiline regressioon – mudeli "headus"

Mudeli kirjeldusvõime

Mudeli kirjeldusastet väljendab see, kui palju erineb tõepärafunktsioon konstantse mudeli tõepärafunktsioonist (tõepärafunktsiooni muut).

Et saada näitajale lihtsamini tajutavaid ja tõlgendatavaid ühikuid:

=> kasutatakse lineaarsest regressioonist tuttava determinatsioonikordaja analooge e pseudo-determinatsioonikordajaid "**pseudo-$R^2$**"

=> muutuvad reeglina 0 ja 1 vahel (on suhteliselt madalad) ning peegeldavad argumentide kirjeldusvõimet

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 424,586[a] | ,116 | ,157 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

TALLINNA ÜLIKOOL

# Logistiline regressioon – mudeli "headus"

Mudelite omavaheline võrdlemine – erinevate prediktorite mõju mudelis

Selle asemel, et võrrelda üht "valmismudelit" konstantse nn 0-mudeliga kasutatakse vahel sammuviisilist meetodit ja võrreldakse järjest lisatavate või eemaldatavate prediktoritega (e IV'dega) mudeleid omavahel või "valmismudelit" mudeliga, kus üks prediktor puudub.

=> prediktori e IV mõju väljendab funktsiooni *-2 Log Likelihood* muut võrreldes mudeliga, kus antud IV'd ei ole.

PS! Tõepärafunktsiooni muut on usaldusväärsem näitaja kui Wald'i statistik prediktori stat.olulisuse testimiseks!

**Step Summary**

| Model | | Action | Effect(s) | Model Fitting Criteria: -2 Log Likelihood | Effect Selection Tests: Chi-Square[a,b] | df | Sig. |
|---|---|---|---|---|---|---|---|
| Step 0 | 0 | Entered | Intercept | 440,851 | . | | |
| Step 1 | 1 | Entered | s184 | 422,024 | 18,826 | 1 | ,000 |
| Step 2 | 2 | Entered | keel | 415,326 | 6,698 | 1 | ,010 |
| Step 3 | 3 | Entered | s182 | 408,647 | 6,679 | 1 | ,010 |
| Step 4 | 4 | Entered | aste | 399,863 | 8,785 | 2 | ,012 |

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the likelihood ratio test.

b. The chi-square for removal is based on the likelihood ratio test.

**Model if Term Removed[a]**

| Variable | | Model Log Likelihood | Change in -2 Log Likelihood | df | Sig. of the Change |
|---|---|---|---|---|---|
| Step 1 | s184 | -233,741 | 18,882 | 1 | ,000 |
| Step 2 | keel | -224,306 | 6,711 | 1 | ,010 |
| | s184 | -229,748 | 17,595 | 1 | ,000 |
| Step 3 | keel | -221,425 | 7,628 | 1 | ,006 |
| | s182 | -220,963 | 6,703 | 1 | ,010 |
| | s184 | -220,537 | 5,851 | 1 | ,016 |

a. Based on conditional parameter estimates

# Logistiline regressioon – mudeli "headus"

Mudelite omavaheline võrdlemine – erinevate "valmismudelite" võrdlemine

NB! Sammuviisilised meetodid lähtuvad ainult statistilisest poolest

=> IV'de valikul tähtsam IV'de sisuline relevantsus ja tõlgendatavus

=> Vali ise erinevaid IV'de "komplekte" ja võrdle nende pseudo-R statistikuid ning tõepärafunktsioonide muute konstantse mudeli suhtes (st hii-ruut näitajaid PS! Suuruse tõlgendamine sõltub vabadusastmete arvust, mis tuleneb prediktorite arvust mudelis!)

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 424,586[a] | ,116 | ,157 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

**Model Fitting Information**

| Model | Model Fitting Criteria: -2 Log Likelihood | Likelihood Ratio Tests: Chi-Square | df | Sig. |
|---|---|---|---|---|
| Intercept Only | 440,851 | | | |
| Final | 398,011 | 42,840 | 6 | ,000 |

TALLINNA ÜLIKOOL

## Logistiline regressioon – mudeli "headus"

Mudeli testimise täiendavad võimalused: *goodness-of-fit test*

Binaarne mudel: Hosmer & Lemeshow test

Mulitinomiaalne mudel: Pearson'i test ja Deviance test

$H_0$:Mudel sobib antud andmetele (andmed vastavad mudeli eeldustele)

=> analoogselt eelduste täidetuse testidega on antud testi puhul hea, kui olulisustõenäosuse näitaja *Sig* on suur ja saab jääda $H_0$ juurde!

NB! Need testid on tundlikud "tühjade lahtrite" e tühjade alamgruppide suhtes!
=> seetõttu ei tööta, kui arvtunnuste põhjal alamgrupid tehakse

**Hosmer and Lemeshow Test**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 5,228 | 8 | ,733 |

**Goodness-of-Fit**

| | Chi-Square | df | Sig. |
|---|---|---|---|
| Pearson | 147,696 | 124 | ,072 |
| Deviance | 157,982 | 124 | ,021 |

TALLINNA ÜLIKOOL