E-kursuse "Andmeanalüüs II" materjalid

Aine maht 6 EAP

Ene Käärik (Tartu Ülikool), 2013

Andmeanalüüs II. Lühitutvustus

Aine maht: 6 EAP

Ainekood: MTMS.01.007

Vastutav õppejõud: Ene Käärik (lektor, Tartu Ülikool, matemaatilise statistika instituut)

Sihtgrupp: matemaatilise statistika bakalaureuseõppe üliõpilased, finants- ja kindlustusmatemaatika magistrandid, informaatika magistrandid

Kohustuslikud eeldusained: MTMS.02.001 Matemaatiline statistika I, MTMS.01.069 Andmeanalüüs I

Kursuse lühikirjeldus: Kursusel käsitletakse tunnustevahelisi seoseid ja mitmesuguste statistiliste mudelite koostamist (regressioonanalüüsi, dispersioonianalüüsi, kovariatsioonanalüüsi, logistilise regressiooni ja Poissoni mudelit). Praktikumides töötatakse reaalsete andmestikega kasutades tarkvarapaketti SAS.

Kursuse eesmärk: Anda terviklik ülevaade andmete statistilise analüüsi läbiviimisest ja professionaalsed oskused järelduste tegemiseks analüüsi põhjal.

Õpiväljund:

Aine läbinud üliõpilane
- oskab valida andmetele sobiva analüüsimeetodi, teab vajalikke analüüsietappe
- kasutab oskuslikult sobivat tarkvara vajaliku probleemi lahendamiseks
- suudab hinnata ja interpreteerida saadud tulemusi

Lõpphindamine: eristav (A, B, C, D, E, F, mi)

Eksamile pääsemise tingimused: Positiivsele hindele sooritatud 2 kontrolltööd (+tunnikontrollid) ja kaitstud projekt.

Lõpphinde kujunemine: kontrolltööd 20% + projekt 30% + eksam 50% (+ lisapunktid tunnikontrollidest)

Lõpphinne kujuneb vastavalt kogutud punktidele järgmiselt:

- 91+ %: A
- 81 - 90.9 %: B
- 71 - 80.9 %: C
- 61 - 70.9 %: D
- 51 - 60.9 %: E
- alla 51 %: F

E-õpe: käesolev e-kursus toetab auditoorset õpet ja sobib ka iseseisvaks õppeks.

Loengukonspekt sisaldab kogu kursuse jaoks vaja minevat teoreetilist materjali. Ülesannete lahendamiseks on kursusel ette nähtud praktikumid, aga soovi korral on võimalik ülesandeid lahendada ka iseseisvalt. Ülesanded on jagatud vastavalt teooria peatükkidele ja ülesannete juures on selgitused, kuidas neid paketi SAS abil lahendada. Tekkivaid küsimusi saab esitada üldfoorumisse. Moodle'i keskkonnas on näha koondhinde kujunemine kontrolltööde, projekti hinde ja eksamitulemuste põhjal. Kontrolltööd, projekti kaitsmine ja eksam tuleb läbida auditoorselt.

Soovituslikud loengumaterjalid:

- E. Ehasalu, E.-M. Tiit (1993) Faktoranalüüs ja kanooniline analüüs SAS-süsteemis. Käsiraamat üliõpilastele II, Tartu.
- E. Käärik (1995). Kordusmõõtmistest. ESS Teabevihik N 5, lk 33-38, Tartu.
- E. Käärik, I. Jakoreva (1997). Valiidsusest ja reliaablusest. ESS Teabevihik N 9, lk 42-46, Tartu.
- A.-M. Parring, M. Vähi, E. Käärik (1997). Statistilise andmetöötluse algõpetus. Tartu.
- D.G. Kleinbaum, L.L. Kupper, K.E. Muller, A. Nizam (1998). Applied regression analysis and other multivariable methods, Duxbury Press.
- R. H. Myers (1990). Classical and modern regression with applications, Duxbury Press.

Lisainfo: Ene Käärik (ene.kaarik@ut.ee)

Tartu Ülikool
Matemaatika-informaatikateaduskond
Matemaatilise statistika instituut

Andmeanalüüs II (MTMS.01.007)

Loengukonspekt

Õppejõud: Ene Käärik

Sisukord

Peatükk 1

Sissejuhatus

1.1 Statistilise andmeanalüüsi ülesanne

Andmeanalüüsi **eesmärgiks** on teha teaduslikke järeldusi empiiriliste (vaatlustest, katsetest, mõõtmistest pärinevate) andmete põhjal. Andmeanalüüs kasutab oma meetodina matemaatilist statistikat (statistiliste hüpoteeside kontrollimise teooriat, mudelite konstrueerimist nende parameetrite hindamise teel jm.). Tavaliselt lahendatakse andmeanalüüsi ülesanded arvuti abil, kasutades standardset tarkvara - statistikaprogrammide pakette.

Andmeanalüüsi **ülesande lahenduskäik**

- Ülesande püstitamine, üldkogumi määratlemine, valimi mahu määramine, mõõdetavate tunnuste valimine.
- Katse planeerimine ja läbiviimine, eetilised probleemid. Töötlusplaani täpsustamine, töötlusvahendi (riist- ja tarkvara) valik.
- Andmete kirjeldamine, kogumine, kodeerimine, vigadest puhastamine. Andmestiku kontrollimine arvutiprotseduuride abil.
- Sisulise ülesande tõlkimine matemaatilise statistika keelde. Eelduste kontroll, tõestatavate hüpoteeside loetelu fikseerimine.
- Ülesande matemaatiline lahendamine. Lähtetunnuste jaotuste kirjeldamine, jaotusparameetrite hindamine, hüpoteeside kontrollimine, mudelite konstrueerimine.
- Tulemuste esitamine ja tõlgendamine, järelduste tegemine, publitseerimine.

Andmeanalüüsi **ülesannete tüübid**

1. Kirjeldav ülesanne. Eesmärgiks on andmeid kirjeldada.
2. Hüpoteeside kontrollimine. Otsustuste ja järelduste tegemine.
3. Mudeli koostamine. Sisaldab mudeli parameetrite hindamist, mudeli olulisuse ja adekvaatsuse kontrolli, mudeli diagnostikat, mudeli sisulist tõlgendamist.
4. Prognoosiülesanne. Leitud mudeli kasutamine.

Iga järgnev ülesande tüüp sisaldab eneses eelmisi.

1.1.1 Tunnus

Tunnus (*variable*) on mõõtmise, küsitluse, katse või vaatluse tulemusena saadud arvuline või mittearvuline näitaja, matemaatilise statistika mõttes juhuslik suurus (*random variable*), juhuslik muutuja. Tema väärtus sõltub sellest, missugust objekti mõõdetakse.

Tunnus on see, mida me mõõdame.

Mõõtmise protsessis omistatakse igale mõõdetavale objektile mõõdetava tunnuse väärtus.

Tunnuse puuduv väärtus on tühik st puuduvat väärtust ei tohi asendada!

Tunnusetüübid:

A. Arvulised tunnused ehk kvantitatiivsed (*numerical*)

- **pidevad** (*continuous*) arvtunnused, saadakse otsesel mõõtmisel. Näiteks lapse kaal ja kasv, vererõhk, palk.
- **diskreetsed** (*discrete*) arvtunnused e täisarvulised, saadakse loendamisel. Näiteks laste arv peres, munade arv pesas, laste arv klassis.

B. Mittearvulised ehk kvalitatiivsed tunnused (*categorical*)

- **järjestustunnused** (*ordinal*) (sh binaarsed kahe väärtusega, näiteks sugu) mittearvulised tunnused, mille väärtuste vahel on võimalik objektiivne järjestus (hinnangud etteantud skaalal jm). Näiteks haridus (alg-põhi-kesk-kõrgem), hinded (väga hea-hea-keskm-nõrk), haiguse raskusaste (I astme põletus-IIastme põletus-III astme põletus).
- **nominaaltunnused** (*nominal*), mittearvulised tunnused, mille väärtuste vahel ei ole sisulist järjestust. Näiteks rahvus, silmade värv, eriala.

Tunnuste kodeerimine

Mittearvuliste tunnuste puhul on otstarbekas nad enne töötlemise algust kodeerida, st asendada sõnalised vastusevariandid arvudega ehk koodidega. Kodeerimise puhul on tarvis järgida teatavaid otstarbekuse reegleid:

- järjestustunnuste kodeerimisel tuleb jälgida, et koodid säilitaksid väärtuste sisulise järjestuse;
- binaarse tunnuse kodeerimisel on eelistatav lihtsaim võimalus, näiteks 1 ja 2 (või ka 0 ja 1, kui see on sisuliselt mõistetavam).
- nominaaltunnuseid ei ole vaja arvuliseks kodeerida. Sageli on otstarbekas asendada pikad vastusevariandid kokkuleppeliste lühenditega.

Eritüüpi tunnustele on rakendatavad erinevad töötlusreeglid!

1.2 Statistiline andmestik

Statistiline andmestik on mõõtmis-, katse- või vaatlustulemuste hulk, mis mõõdetud teataval viisil määratletud objektidel (valimil).

Statistilist andmestikku iseloomustab:

(1) mõõdetud tunnuste loetelu,

(2) mõõdetud objektide loetelu.

Ideaaljuhul on kõigil objektidel kõik tunnused mõõdetud. Niisugust andmestikku nimetatakse täielikuks. Kui osa mõõtmistulemusi puudub on andmestik lünklik.

1.2.1 Objekt-tunnus-maatriks

Ülevaatlikuma pildi saamiseks andmestikust esitatakse see tavaliselt tabelina, kus igale reale vastab üks objekt ja igale veerule vastab üks tunnus. Niisugust tabelit nimetatakse objekt-tunnus-maatriksiks (tabeliks). Tunnuseid tähistatakse tavaliselt suurtähtedega X, Y, Z, lisades vajaduse korral indekseid. Tunnuste arv andmestikus tähistatakse kokkuleppeliselt tähega m. Objektide arv ehk valimi maht tähistatakse tavaliselt tähega n. Objekt-tunnus-maatriksi i-ndas reas ja j-ndas veerus olev väärtus x_{ij} on saadud

i-ndal objektil j-nda tunnuse mõõtmisel (x_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, m$).

$$\begin{pmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{im} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{nm} \end{pmatrix}$$

Praktiliselt kõik statistikapaketid eeldavad, et andmestik oleks esitatud objekt-tunnus tabelina.

MS Excel on tabelarvutussüsteem (mitte statistikapakett) ja seal kehtivad mõnevõrra erinevad reeglid.

1.2.2 Vead andmetes

Eksimine on inimlik ja seetõttu esineb andmestikus tavaliselt vähem või rohkem mitmesuguseid vigu.

Vead saab jagada järgmistesse tüüpidesse:

- juhuslikud vead (tingitud mõõtmise või protokollimise ebatäpsusest, üldreeglina muudavad tulemust vähe ja on raskesti avastatavad);
- süstemaatilised vead (tingitud enamasti instrumendi ebatäpsusest);
- jämedad vead (tunnuse väärtus on väljaspool tunnuse võimalike väärtuste piirkonda, näiteks on mõõdetud kasvu sentimeetites, aga ühel on antud meetrites);
- loogilised vead, kus erinevate tunnuste väärtused ei ole kooskõlas (näiteks tööstaaž on suurem kui vastaja vanus, jne)

Andmete kontrollimine ja parandamine

Kõige tähtsam on avastada **jämedad ja loogilised** vead, mis analüüsitulemusi väga oluliselt mõjustavad.

Enne sisulise töötlemise algust peab uurija olema veendunud, et tema andmestikus ei ole ühtki jämedat viga! Jämedad vead andmestikus põhjustavad ekslikke järeldusi!

Jämedate vigade avastamisel aitavad arvutiprogrammid, mis esitavad kas arvuliselt või graafiliselt tunnuse suurimad ja vähimad väärtused. Jämedad vead, kui neid esineb, satuvadki enamasti tunnuse ekstremaalväärtusteks. Jämeda vea kõrvaldamiseks on kaks võimalust: asendada see väärtus tühikuga,

või, kui tegemist on arvutisse sisestamise veaga, asendada mõõtmisprotokollis leiduva õige väärtusega.

Loogiliste vigade leidmine on keerukam. Nende avastamiseks võib kasutada kahe tunnuse hajuvusdiagrammi, kus teistest eraldiseisvad punktid võivad viidata loogilistele vidagele.

1.2.3 Puuduvad andmed

Tavaliselt on meil tegemist andmestikuga, kus esineb puuduvaid (*missing*) väärtusi st meil on lünklik andmestik (*incomplete data*). Lünkliku andmestikuga seotud probleemidega hakati tegelema 1980ndatel ja see kujutab endast keerulist ja kiiresti arenevat valdkonda (Little & Rubin, 1987).

Puudumiste struktuuri defineerimisel kasutatakse järgmisi tähistusi: andmemaatriks $\mathbf{X}$ jaguneb kaheks osaks $\mathbf{X} = (\mathbf{X}_{OBS}, \mathbf{X}_{MIS})$, kus $\mathbf{X}_{OBS}$ on täielik ja $\mathbf{X}_{MIS}$ sisaldab puuduvaid väärtusi, puudumisindikaatorite maatriks $\mathbf{M} = \{m_{ij}\}$, ($m_{ij} = 1$, kui väärtus on).

Puudumiste struktuur:

- Täiesti juhuslik puudumine (*Missing Completely at Random*, MCAR). $\mathbf{P}(\mathbf{M}|\mathbf{X}) = \mathbf{P}(\mathbf{M})$. Puudumine on täiesti juhuslik.
- Juhuslik puudumine (*Missing at Random*, MAR) $\mathbf{P}(\mathbf{M}|\mathbf{X}) = \mathbf{P}(\mathbf{M}, \mathbf{X}_{OBS})$. Puuduv väärtus mingis tunnuses võib sõltuda mingi teise tunnuse väärtusest.
- Mittejuhuslik puudumine (*Not Missing at Random, Nonignorable* NMAR) $\mathbf{P}(\mathbf{M}|\mathbf{X}) = \mathbf{P}(\mathbf{X}_{OBS}, \mathbf{X}_{MIS})$. Puudumine sõltub väärtusest endast st sellest, milline oleks olnud tegelik väärtus.

On olemas suur hulk erinevaid puuduvate andmete käsitlusmeetodeid, neist tuntumad on järgmised:

- jäetakse välja kõik puuduvate andmetega objektid (*listwise, casewise data deletion*);
- tunnuspaari analüüsimisel jäetakse välja puuduvate andmetega objektid (pairwise data deletion);
- puuduvad väärtused asendatakse vaadeldava tunnuse keskväärtusega;
- arvutatakse puuduvate väärtuste lineaarsed prognoosid;
- puuduv väärtus asendatakse selle tunnuse väärtusega mõnel teisel (kas juhuslikult valitud või mingis mõttes sarnasel) objektil (*Hot deck*);
- kasutatakse nn EM-algoritmi (*Expectation Maximization*, EM)
- kasutatakse mitmest asendamist (*multiple imputation*).

1.3 Valimi mahu määramisest

Valimi moodustamise protseduur määrab uuringu usaldatavuse ja valimilt populatsioonile tehtava üldistuse laadi. Valimi mahu leidmisel peab lähtuma püstitatud statistilisest hüpoteesist.

Statistiliste hüpoteeside juures räägitakse kaht liiki vigadest:

- $\alpha = \mathbf{P}$(I liiki viga)= $\mathbf{P}$(lükatakse ümber H_0 kui ta on õige)
  Type I error: "rejecting the null hypothesis when it is true"
- $\beta = \mathbf{P}$(II liiki viga)= $\mathbf{P}$(ei lükata ümber H_0 kui ta on vale)
  Type II error: äccepting the null hypothesis when it is false"

Valimi mahu määramisel kaks lähenemist:

I Minimiseeritakse I liiki viga (*precision analysis* – täpsuse analüüs)

II Minimiseeritakse II liiki viga (*power analysis* – võimsuse analüüs)

I Täpsuse analüüs

Hoitakse kontrolli all I liiki viga. Usaldusnivoo $1 - \alpha$ näitab tõenäosust, et ei lükata ümber õiget nullhüpoteesi. Leitakse valimi maht lähtudes usaldusvahemikust.

$(1 - \alpha)$ usaldusvahemik (*confidence interval*) on vahemik, mis tõenäosusega $(1 - \alpha)$ sisaldab parameetri õiget väärtust.

0,95 usaldusvahemik keskväärtusele leitakse järgmiselt

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Pool usaldusvahemiku laiusest on maksimaalne hinnangu viga d, seega

$$d = z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Siit saab avaldada valimi mahu n

$$\boxed{n = \frac{z^2_{\frac{\alpha}{2}} s^2}{d^2}}$$

Et tavaliselt $\alpha = 0,05$, siis standardnormaaljaotuse tabelist $z_{\frac{\alpha}{2}} = 1,96$.

II Võimsuse analüüs

Hoitakse kontrolli all II liiki viga β.

Valimi mahu määramiseks peab olema teada:

- Olulisuse nivoo α
- Testi võimsus P
- Hinnatav mõju suurus / minimaalne täpsus / minimaalne erinevus
- Hajuvus σ

Mida raskem võib olla esimest liiki vea poolt põhjustatud tagajärg, seda väiksem valitakse olulisuse nivoo ($\alpha = 0,05; \alpha = 0,01$).

Testi võimsus $P(power)$ on defineeritud kui tõenäosus valimi põhjal tõestada seaduspära, kui see ka tegelikult esineb $P = 1-\beta = \mathbf{P}$(lükatakse ümber H_0 kui ta on vale). Kui uuringu võimsus on madal, siis see seab kahtluse alla uuringu tulemused, testi võimsus peaks olema suurem või võrdne kui 0,8.

Mõju suurus (*effect size*) defineeritakse tavaliselt kui erinevus või standardiseeritud erinevus:

- $d = \mu_1 - \mu_2$
- $\delta = \frac{\mu_1 - \mu_2}{\sigma}$ (standardiseeritud mõju)

On antud ka empiirilised hinnangud (standardiseeritud) mõju suurusele (Cochen, 2001):

- $\delta < 0.1$ triviaalne mõju;
- $0.1 - 0.3$ väike mõju;
- $0.3 - 0.5$ keskmine mõju;
- $\delta > 0.5$ suur mõju.

NB! *Mõju statistiline olulisus ja mõju sisuline suurus on erinevad mõisted!*

Valimi mahu määramine. Rusikareeglid

Allikas: C. R. Wilson Van Voorhis, B.L. Morgan (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology*, vol 3(2), 43-50.

Probleem	Soovitatav valimi maht 80% võimsuse jaoks
Rühmakeskmiste erinevus	$n = 30$ (rühmas)
Seoste hindamine (korrelatsioon, regressioon)	$n \approx 50$
χ^2–test	$n \approx 20$
Faktoranalüüs	$n \approx 300$ on hea

Tuleb silmas pidada, et tabelis on toodud tõepoolest jämedad hinnangud ja vajatakse suuremat valimi mahtu kui hinnatav mõju/erinevus on väiksem või kui soovitakse kasutada kõrgemat olulisusenivood ja/või vajatakse suuremat võimsust.

Paketis **SAS** saab hinnata vajaliku valimi mahu suurust kasutades protseduuri POWER.

`http://www.ats.ucla.edu/stat/sas/dae/t_test_power2.htm`

On olemas ka mitmesugust vabavara valimi mahu hindamiseks, vt näiteks:

PS Power and Sample Size Calculations
http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

1.4 Normaaljaotus ja selle kontrollimine

Normaaljaotus on pideva juhusliku suuruse jaotus. Paljud juhuslikkusel põhinevad nähtused on ligikaudu normaaljaotusega ning paljud teoreetilised tulemused kehtivad normaaljaotuse eeldusel.

Jaotuse kontrollimiseks võime kasutada graafilisi meetodeid või teste.

1.4.1 Graafilised meetodid

Graafilised meetodid annavad visuaalse pildi, mille abil saame teha subjektiivse oletuse tunnuse jaotuse kohta. Kuna on võimalik uurida ainult tunnuse ligikaudset jaotust, siis pole tihti oluline mingite formaalsete testide tegemine, vaid piisab graafiku põhjal saadud informatsioonist.

- **Histogramm**. Võrreldakse valimi histogrammi normaaljaotuste tihedusfunktsiooni graafikuga. Sümmeetriline histogramm kinnitab oletust normaaljaotuse kohta.
- **Q-Q plot** (*Quantile-Quantile plot*; kvantiilide graafik) või **Normal Probability Plot** (tõenäosuspaber). Joonistatakse hajuvusgraafik, kus y-teljel on uuritava tunnuse järjestatud väärtused ja x-teljel teoreetilise jaotuse (normaaljaotuse $N(0, 1)$) kvantiilid. Punktide paiknemine enam-vähem sirgel kinnitab oletust valitud teoreetilise jaotuse (normaaljaotuse) kohta.

1.4.2 Testid

Kui me ei suuda graafiku põhjal otsustada, võib kasutada teste võrdlemaks empiirilist jaotust normaaljaotusega.

Testime nullhüpoteesi, mis väidab, et andmed on juhuslik valim normaaljaotusest.

H_0 : jaotuseks on normaaljaotus
H_1 : jaotuseks ei ole normaaljaotus

Siin ei ole me huvitatud nullhüpoteesi kummutamisest (H_1 tõestamisest), vaid me tahame jääda nullhüpoteesi juurde: andmed on juhuslik valim normaaljaotusest.

Tuntumad testid:

- **Shapiro-Wilk'i test** (Shapiro, 1965)
  Teststatistik W põhineb dispersioonihinnangute suhtel, arvutatakse kui valimi maht $n \leq 2000$, suurema valimi korral kasutatakse ligikaudset lähenemist (Roystoni meetod, 1992).
- **Kolmogorov-Smirnovi test** (Kolmogorov, 1933; Smirnov, 1936)
  Kolmogorov-Smirnovi test põhineb empiirilise ja teoreetilise jaotusfunktisooni võrdlusel. Teststatistik D arvutatakse kui maksimaalne erinevus (vt täpsemalt näiteks Kollo (2004). Monte Carlo meetodid, lk 28 jj)
- **Anderson-Darlig'i test** (Stephens, 1974)
  Teststatistik A^2 baseerub empiirilise ja teoreetilise jaotusfunktisooni kaalutud erinevuse ruudul ja on Kolmogorov-Smirnovi testi modifikatsioon, kus 'sabadel' on suurem kaal.
- **Cramer-von Mises' test** (Stephens, 1974)
  Teststatistik W^2 on analoogiline Anderson-Darlingi teststatistikuga, kus kaal on võrdne ühega.

1.5 Andmeanalüüsi ajaloost

Andmeanalüüsi isaks loetakse **John Wilder Tukey**
(16. juuni, 1915 – 26. juuli, 2000; USA)

Tukey sai bakalaureuse ja magistrikraadi keemias Brown'i ülikoolis ja PhD matemaatikas Princetoni ülikoolis 1939. Kui ülikooli juurde loodi Statistika osakond 1965, oli Tukey selle esimene juhataja.

J. Tukey oli väga laia silmaringiga teadlane ja tema tähtsamateks saavutusteks erinevates valdkondades loetakse:

- uued sõnad inglise keeles 'software', 'bit'
- arvutustehnikas – kiired Fourier teisendused
- statistikas – robustsed statistikud, mitmese võrdlemise meetodid, kõrgedimensionaalne andmeanalüüs (visualiseerimine)

J. Tukey on saanud mitmeid autasusid, medaleid ja aunimetusi, temast räägitakse kui ainulaadsest isiksusest ja teda loetakse viimase 50 aasta üheks mõjukamaks statistikuks

J. Tukey oli esimene, kes 1960-ndatel aastatel *rääkis andmeanalüüsist kui omaette distsipliinist*, siiani oli olemas ainult matemaatiline statistika.

J. Tukey tähtsamad ilmunud teosed:

1962 Future of Data Analysis
1964 The technical Tools of Statistics
1968 Data Analysis, including Statistics (kaasautor Fred Mosteller)
1976 Exploratory Data Analysis

Kokku on Tukey kirjutanud üle 500 teadusliku artikli ja ta on olnud paljude ajakirjade toimetuste kolleegiumite ja paljude teaduslike organisatsioonide liige.

John Tukey on öelnud:

Everyone thinks that the data in other people's subjects are in better shape

Try, look, and try something a little different as the typical pattern of data analysis

Finding the question is often more important than finding the answer

Models should not be true but it is important that they be applicable

1.6 Ülesanne kordamiseks

Kahtlustati, et kosmeetikafirma petab oma kliente ja paneb kreemitopsi (kirjaga 200 g) lubatust vähem kreemi.

Uuringus leiti (valimis 100 topsi), et keskmiselt sisaldab tops 193g kreemi, 95% usaldusvahemikuga (185g, 201g).

Kumb väide on õige? Miks?

1. Tõestasime, et firma petab oma kliente.
2. Ei saanud tõestada, et firma petab oma kliente.

Peatükk 2

Tunnustevahelised seosed

2.1 Üldine statistiline sõltuvus

Statistiline sõltuvus (*dependence*) tunnuste X ja Y vahel on üks statistika põhimõisteid. Kaks suurust on *sõltumatud*, kui ühe suuruse muutumine ei mõjuta teise suuruse muutumist. Vastasel juhul on tegemist sõltuvate suurustega.

Tunnused on statistiliselt sõltumatud, kui nende ühisjaotus võrdub äärejaotuste korrutisega. *Statistiline sõltuvus defineeritakse kui sõltumatuse puudumine.*

Tunnused on statistiliselt sõltuvad kui nende ühisjaotus pole määratud äärejaotustega, st tunnuste koosmõjudes lisandub veel midagi, mida äärejaotused täielikult ei kirjelda.

Statistiline sõltuvus on kõige üldisem sõltuvuse viis ja seda saab määrata ka nominaaltunnuste korral.

Üldise statistilise sõltuvuse seosekordajad baseeruvad χ^2–statistikul

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

O – vaadeldud (observed), E – oodatud, teoreetiline (*expected*)

Statistiku maksimaalne väärtus sõltub valimi mahust, seega sobib statistik seose olemasolu kontrollimiseks, aga mitte seose tugevuse hindamiseks.

Seose tugevuse hindamiseks defineeritakse χ^2–statistiku abil terve rida näitajaid:

- **Pearsoni** Φ–kordaja (*Phi coefficient*), **Pearsoni** C ehk kontingentsuse kordaja (*contingency coefficient*),
- **Crameri** V–statistik (*Cramer's V*).

Need statistikuid saab paketis SAS leida kasutades protseduuri FREQ TABLES lauses valikut CHISQ (väljastatakse p-väärtus χ^2-statistikule, sama p-väärtus kehtib ka teiste statistikute jaoks).

Protseduur arvutab lisaks tõepärasuhte χ^2 (*Likelihood Ratio* χ^2) ja Mantel-Haenszel χ^2 (NB! eeldab tunnuste järjestust!).

2.2 Monotoonne sõltuvus

Monotoonne sõltuvus on selline sõltuvus kahe tunnuse vahel, kus ühe tunnuse muutus mingis kindlas suunas toob kaasa teise tunnuse muutuse kindlas suunas.

Öeldakse, et kahe tunnuse (juhusliku suuruse) $X = (x_1, x_2, \ldots, x_n)$ ja $Y = (y_1, y_2, \ldots, y_n)$ väärtuste vahel on positiivne (samapidine) monotoonne sõltuvus, kui $x_i < x_j \Rightarrow y_i < y_j$ (tõenäosusega 0.5) ja negatiivne (vastupidine) monotoonne sõltuvus, kui $x_i < x_j \Rightarrow y_i > y_j$ (tõenäosusega 0.5).

Monotoonse sõltuvuse seosekordajate korral:

- Kordajate väärtused muutuvad piirides [-1, 1];
- Kordaja märk näitab seose suunda: kas tegu on kahaneva või kasvava seosega;
- Kordaja absoluutväärtus näitab seose tugevust: mida lähemal ühele seda tugevam sõltuvus.

Põhilisteks monotoonse sõltuvuse seosekordajateks on:

1. **Spearmanni** korrelatsioonikordaja ehk **astakkorrelatsioonikordaja** (tähistatkse ρ või r_S). Arvutamisel lähtutakse mõõtmistulemuste asemel nende astakutest.
2. **Kendalli korrelatsioonikordaja** (*Kendalli* τ). Põhineb samasuunaliste ja vastassuunaliste paaride arvu analüüsil.

Tuleb tähele panna, et Spearmanni ja Kendalli korrelatsioonikordajad mõõdavad erinevaid aspekte. Spearmanni ρ arvutamisel kasutatakse vahede astakute ruute (jälgitakse samasuunalist järjestust). Kendalli τ korral kasutatakse samasuunaliste ja vastassuunaliste paaride suhtelist sagedust (jälgitakse mõlemasuunalist järjestust, arvutatakse samasuunaliste ja vastassuunaliste paaride suhteliste sageduste vahe).

Meetodite korral, mis põhinevad korrelatsioonanalüüsil (nagu faktoranalüüs), Kendalli korrelatsioonikordaja lähtealuseks ei sobi!

On defineeritud veel terve rida monotoonse seose kordajaid, mis baseeruvad paaride analüüsil:

- **Goodman-Kruskali gamma** (γ) (*conditional gamma*). Arvutamisel ignoreeritakse võrdseid paare.
- **Kendalli tau-b** (τ_b). Analoogiline gammale, võtab arvesse ka võrdseid paare.
- **Stuarti tau-c** (τ_c). Arvestab hindamisel lisaks ka valimi mahtu.
- **Somersi D**. On τ_b asümmeetriline modifikatsioon, võrdseid paare võetakse arvesse ainult sõltuvas tunnuses. Võimaldab hinnata tunnustevahelist seost ühesuunaliselt (st määrates eelnevalt, kumb tunnustest on teisest sõltuv).
- **Guttmani** λ–kordajad. Asümmeetrilisi λ–kordajaid $\lambda[C|R]$ ($\lambda[R|C]$) kasutatakse veeru(rea)tunnuse prognoosimiseks teades infot rea(veeru)-tunnuse kohta. Sümmeetriline λ–kordaja on saadud keskmistades asümmeetriakordajad.
- **Määramatusekordajad** U (*uncertainly coefficient*). Asümmeetriline määramatusekordaja $U[C|R]$ ($U[R|C]$) hindab määramatuse osa veeru-(rea)tunnuses, mis on seletatav rea(veeru)tunnuse abil. Sümmeetriline määramatuse kordaja on keskmistatud.

Monotoonse sõltuvuse seosekordajate arvutamiseks paketis SAS kasutatakse protseduuri FREQ TABLES lauses valikut MEASURES. Protseduuri töö tulemusena väljastatakse igale statistikule vastav asümptootiline standardviga ASE (*asymtotic standard error*)

Statistiku olulisuse testimine põhineb asjaolul, et järgmine suhe on standardnormaaljaotusega

$$\frac{statistik}{ASE}$$

Kui suhe$>$ 1.96, siis statistik on oluline (seose olemasolu on tõestatud olulisuse nivool $\alpha = 0.05$), vastupidisel juhul (suhe < 1.96) statistik ei ole oluline (ei saa rääkida seosest tunnuste vahel üldkogumis).

Näide: Monotoonse sõltuvuse seosekordajate arvutamine (SAS)

Hinnatakse seost depressiooni ja hariduse vahel.

```
Statistics for Table of HARIDUS by SUMMA
Statistic                                Value       ASE
 ------------------------------------------------------
 Gamma                                  -0.0326    0.0555
 Kendall's Tau-b                        -0.0276    0.0470
 Stuart's Tau-c                         -0.0272    0.0464
 Somers' D C|R                          -0.0309    0.0526
 Somers' D R|C                          -0.0247    0.0420
 Pearson Correlation                    -0.1014    0.0527
 Spearman Correlation                   -0.0352    0.0608 <---?
 Lambda Asymmetric C|R                   0.0231    0.0234
 Lambda Asymmetric R|C                   0.1111    0.0419
 Lambda Symmetric                        0.0591    0.0255
 Uncertainty Coefficient C|R             0.0915    0.0095
 Uncertainty Coefficient R|C             0.1805    0.0183
 Uncertainty Coefficient Symmetric       0.1214    0.0123
                                      Sample Size = 294
```

HARIDUS: 1-alg, 2- lõpetamata kesk, 3-keskharidus, 4-lõpetamata kõrgem, 5-kõrgem, 6-magistri kraad, 7-doktori kraad

Isik on depressioonis kui tunnus SUMMA$>$ 15

Kas saame rääkida olulisest monotoonsest seosest tunnuste vahel?

2.3 Lineaarne sõltuvus

Hajuvusdiagramm

Kahe tunnuse vahelise lineaarse seose uurimisel alustatakse hajuvusdiagrammist. Korrelatsiooniväli ehk hajuvusdiagramm (*scatter plot*) annab tunnuste paiknemisest visuaalse pildi (näitab seose suunda ja tugevust).

Tavaliselt on jooniselt näha ka jämedad vead, mis paiknevad teistest punktidest eraldi. Punktide projektsioonid telgedele annavad mõlema tunnuse variatsioonrea.

Punktiparve kuju järgi saame hinnata seose iseloomu:

- Kui punktiparv on välja venitatud kasvavas suunas, siis on tegemist positiivse ehk samapidise seosega.

- Kui punktiparv on välja venitatud langevas suunas, on tegemist negatiivse ehk vastupidise seosega.
- Kui punktiparv on hajus, siis seos puudub.

2.3.1 Pearsoni korrelatsioonikordaja

Kahe pideva normaaljaotusega tunnuse vahelise lineaarse seose ehk nn korrelatiivse sõltuvuse tugevuse hindamiseks kasutatakse *lineaarset korrelatsioonikordajat* (öeldakse ka Pearsoni korrelatsioonikordaja või lihtsalt korrelatsioonikordaja).

Valimi põhjal arvutatud Pearsoni korrelatsioonikordaja tähistatakse tavaliselt r, st korrelatsioonikordajat tunnuste X ja Y vahelise seose kirjeldamiseks tähistatakse r_{XY} või $r(X,Y)$ ja arvutatakse järgmiselt:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Valimi korrelatsioonikordaja r on hinnanguks üldkogumi korrelatsioonikordajale ρ, mis iseloomustab lineaarse seose tugevust üldkogumis

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Korrelatsioonikordaja omadused:

- $r(X,X) = 1$;
- korrelatsioonikordaja on sümmeetriline $r(X,Y) = r(Y,X)$;
- kui $r = 1$ või $r = -1$ on tunnuste vahel täpne funktsionaalne seos;
- kui $r = 0$, siis tunnused on lineaarselt sõltumatud, võib esineda mittelineaarne sõltuvus;
- r^2 näitab kui suur osa ühe tunnuse muutlikkusest on määratud teise tunnuse poolt.

Kasutades korrelatsioonikordajat saame rääkida kolmest aspektist:

I Seose suund

- kui $r > 0$ on tegemist samapidise seosega,
- kui $r < 0$ on tegemist vastupidise seosega.

II Seose tugevus

Seose tugevust näitab korrelatsioonikordaja suurus. Tugevuse hinnangud on empiirilised.

- kui $|r| \geq 0.9$ ehk $r^2 \approx 0.8$ (80%), siis on tegemist *väga tugeva* seosega kahe tunnuse vahel (üks tunnus kirjeldab teisest ligikaudu 80%);
- kui $0.9 > |r| \geq 0.7$ ehk $r^2 \approx 0.5$ (50%), siis on tegemist *tugeva* seosega kahe tunnuse vahel (üks tunnus kirjeldab teisest ligikaudu 50%);
- kui $|r| \approx 0.5$ ehk $r^2 \approx 0.25$ (25%), siis on tegemist *keskmise* seosega kahe tunnuse vahel (üks tunnus kirjeldab teisest ligikaudu 25%).

III Seose olulisus

Korrelatsioonikordaja on oluline, kui tema väärtus üldkogumis erineb nullist st seos kehtib ka üldkogumis. Kontrollitakse hüpoteese

$$H_0 : \rho = 0; \quad H_1 : \rho \neq 0.$$

Hüpoteeside kontrollimiseks on olemas korrelatsioonikordaja kriitiliste väärtuste r^* tabelid. Korrelatsioonikordajat r loetakse oluliseks antud olulisuse nivool α ja antud valimimahu n korral, kui $r > r*$ (kummutatakse nullhüpotees).

Arvuti väljastab korrelatsioonikordaja olulisuse kontrollimiseks olulisuse tõenäosuse p ja otsus tehakse standardsel viisil:

Kui $p < \alpha$, siis loeme korrelatsioonikordaja r oluliseks (H_1) st selline seos kahe tunnuse vahel kehtib ka üldkogumis.
Vastasel juhul ($p \geq \alpha$) on meil tegemist juhusliku seosega, mis kehtib valimis, mida aga ei saa üldistada üldkogumile.

Ülesanded. Korrelatsioonikordaja analüüs

1. Korrelatsioonikordaja õpilaste 60 m jooksu ja kaugushüppe tulemuste vahel on r(jooks, kaugus)$= -0,83$ ($p = 0,042$).
Selgitada tulemust (seose suund, tugevus, olulisus, sisu ?)

2. Korrelatsioonikordaja kasvu ja kaalu vahel on r(kasv, kaal)$= 0,77$, ($p = 0,012$). Selgitada tulemust (seose suund, tugevus, olulisus, sisu ?)

3. Korrelatsioonikordaja pea ümbermõõdu ja keskmise hinde vahel on r(pea, khinne)$= -0,33$ ($p = 0,28$)
Selgitada tulemust (seose suund, tugevus, olulisus, sisu ?)

2.3.2 Korrelatsioonikordajate võrdlemine

Allikas: Kleinbaum, G., Kupper, L.L., Muller, K.E., Nizam, A. (1998). Applied regression analysis and other multivariable methods, lk 99–100.

Tihti tekib küsimus, kui leiame kaks korrelatsioonikordajat, mis on erineva suurusega, kas see erinevus kehtib ka üldkogumis ehk kas meil on tegelikult erineva tugevusega seosed. Seda on võimalik kontrollida testides korrelatsioonikordajate erinevust. Nii nagu t-testide puhul tehakse ka siin vahet, kas korrelatsioonikordajad, mida võrreldakse, on leitud samal valimil või erinevatel st kas on tegemist sõltumatute või sõltuvate valimitega korrelatsioonikordajate võrdlemisega.

Sõltumatud valimid

Oletame, et meil on tegemist kahe sõltumatu valimiga kahest erinevast üldkogumist mahtudega n_1 ja n_2 ja me tahame testida hüpoteese

$$H_0 : \rho_1 = \rho_2; \quad H_1 : \rho_1 \neq \rho_2$$

Testimiseks vajaliku statistiku leidmiseks kasutatakse *Fisheri z-teisendust* $z = \ln \frac{1+r}{1-r}$

Oletame, et kahe valimi korral arvutatud korrelatsioonikordajad on vastavalt r_1 ja r_2 ja vastavad Fisheri z-teisendused tähistame z_1, z_2.

Teststatistikul on kuju:

$$Z = \frac{z_1 - z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}},$$

kui kehtib H_0, siis Z on asümptootiliselt standardnormaaljaotusega.

Milline on otsustusreegel?

Sõltuvad valimid

Tahetakse hinnata, kas seos esimese ja teise tunuse vahel on samasugune kui esimese ja kolmanda vahel:

$$H_0 : \rho_{12} = \rho_{13}; \quad H_1 : \rho_{12} \neq \rho_{13}$$

Tegemist on ühe üldkogumiga, kust on võetud valim n elementi.

Vaatluse all on kolm tunnust X_1, X_2, X_3 ja arvutatakse nendevahelised korrelatsioonid r_{12}, r_{13}, r_{23}.

Selge on, et saadud korrelatsioonid pole sõltumatud, sest nende arvutamisel on kasutatud samu objekte.

Saab näidata, et suure valimimahu n korral on H_0 kehtimisel järgmine teststatistik asümptootiliselt standardnormaaljaotusega (Olkin&Siotani, 1964; Olkin, 1967)

$$Z = \frac{(r_{12} - r_{13})\sqrt{n}}{\sqrt{(1 - r_{12}^2)^2 + (1 - r_{13}^2)^2 - 2r_{23}^3 - (2r_{23} - r_{12}r_{13})(1 - r_{12}^2 - r_{13}^2 - r_{23}^2)}}$$

Korrelatsioonikordajate võrdlemiseks võib kasutada võrgulehekülgedel leiduvaid statistilisi kalkulaatoreid.

Näide. Korrelatsioonikordajate võrdlemine

Lastel on mõõdetud kasv ja kaal ning on teada ka laste vanus.

Uuringus osales 12 last. Leiti seosed järgmiste tunnuste vahel

$$r_{12} = r(kaal, kasv) = 0.814; \quad r_{13} = r(kaal, vanus) = 0.77$$

$$r_{23} = r(kasv, vanus) = 0.614$$

Kas kasv ja vanus on kaaluga ühtemoodi seotud?

Teststatistiku väärtus: $Z = 0.344$

Millise testiga on tegemist? Milline on otsus?

2.3.3 Teisi korrelatsioonikordajaid

Spearmanni korrelatsioonikordajast ehk astakkorrelatsioonikordajast oli juba juttu eespool (vt osa 2.2) ja see hindab monotoonse seose tugevust, st kasutatakse kahe tunnuse vahelise seose tugevuse hindamiseks juhul, kui meil pidevad tunnused pole normaaljaotusega või on tegemist järjestustunnustega. Spearmanni korrelatsioonikordaja arvutamisel lähtutakse mõõtmistulemuste asemel nende astakutest (järjekorranumbrid variatsioonreas) ja kasutatakse Pearsoni korrelatsioonikordaja valemit.

Et Spearmanni korrelatsioonikordaja mõõdab *monotoonse sõltuvuse tugevust*, siis ta pole nii tundlik erindite suhtes.

Spearmanni korrelatsioonikordaja omadused ja olulisuse kontroll on analoogilised Pearsoni korrelatsioonikordajale. Kahe tunnuse vahelise Spearmanni korrelatsioonikordaja väärtus on enamasti suurem kui Pearsoni korrelatsioonikordaja väärtus.

Neljavälja korrelatsioonikordaja

Korrelatsioonikordajat binaarsete tunnuste jaoks nimetatakse ka tetrahooriliseks (*tetrachoric*) ehk **neljavälja** korrelatsiooniks.

Binaarsete ehk kahe väärtusega tunnuste korral on tegemist 2x2 sagedustabeliga, millel on tavaliselt järgmine kuju:

	1	0	Summa
1	a	b	$a+b$
0	c	d	$c+d$
Summa	$a+c$	$b+d$	$n=a+b+c+d$

Korrelatsioon kahe binaarse tunnuse vahel leitakse iteratiivset algoritmi kasutades (tavaliselt Newton-Rapsoni meetodil). Lähendustest tuntuim on nn Yule'i kordaja, mis eeltoodud tabeli jaoks arvutatakse järgmiselt:

Yule'i lähend $Q=\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$; Pearsoni lähend $Q=\frac{ad-bc}{ad+bc}$.

Neljavälja korrelatsiooni üldistuseks oleks **mitmevälja** (*polychoric*) korrelatsioon, mis arvutatakse kxs tabelite jaoks st tunnustele, kus ühel on k erinevat kategooriat (taset, väärtust) ja teisel s erinevat kategooriat.

Paketis SAS saab neljaväljakorrelatsiooni arvutada protseduuriga FREQ, kasutades TABLES lauses valikut PLCORR, kusjuures vastavalt tabeli suurusele leitakse automaatselt kas neljavälja või mitmevälja korrelatsioonikordaja.

Biseriaalkorrelatsioonikordajad

Allikas: Encyclopedia of statistical sciences (1982). Ed. Kotz, S., Johnson, N.L., vol 1, lk 276–279.

Biseriaalkorrelatsioonist räägitakse siis, kui üks tunnus on binaarne (X), aga teine (Y) pidev ja normaaljaotusega esimese tunnuse klassides st eeldatakse

$$Y|_{X=1} \sim N(\mu_1, \sigma^2),\ Y|_{X=2} \sim N(\mu_2, \sigma^2).$$

Sel juhul avaldub **punkt-biseriaalkorrelatsioonikordaja** (*point-biserial*) seosega

$$r_{pb} = \sqrt{p(1-p)}\ \frac{\bar{y}_1 - \bar{y}_2}{s_y}, \quad (*)$$

kus $\bar{y}_1, \bar{y}_2$ on vastavad tinglikud keskmised (pideva tunnuse Y keskmised binaarse tunnuse X klassides), p on ühtede osakaal tunnuses X ja s_y on tunnuse Y standardhälve.

Punkt-biseriaalkorrelatsioonikordaja olulisuse kontrollimiseks st hüpoteeside paari $H_0 : \rho_{pb} = 0;\ H_1 : \rho_{pb} \neq 0$ kontrollimiseks kasutatakse t-statistikut

$$t = \sqrt{n-2}\ r_{pb}\sqrt{1-r_{pb}^2} \sim t_{n-2}$$

Näide. Punkt-biseriaalkorrelatsiooni arvutamine.

Probleem: kas ekstreemse sünnitusajaga naised kasutavad enamasti ravimit ehk kas sünnitusaja ekstreemsus on seotud ravimi tarvitamisega.

Andmestikus on 34 sünnitaja andmed, kelledest 13 olid võtnud ravimit.

Lahendus:
Leitakse ühtede osakaal $13/34 = 0.382;\ \bar{y}_1 = 5.31; \bar{y}_0 = 2.48;\ s_y = 3.86$

Kasutades valemit (*) saadakse punkt-biseraiaalkorrelatsioonikordaja väärtuseks $r_{pb} = 0.36$, vastav t-statistik $t = 2.2$ ja tabelist $t_{32} \approx 2$

Milline on otsus?

Üldise biseriaalkorrelatsiooni hindamisel lähtutakse asjaolust, et tihti on binaarse tunnuse taga mingi pidev suurus Z, mida mõõta ei saa, aga mis põhjustab tunnuse X binaarsuse. Näiteks $X = 1$, kui Z saavutab mingi läve ja teistel juhtudel $X = 0$.

Sel juhul korrelatsiooni selle mittemõõdetava pideva tunnuse Z ja mõõdetava pideva tunnuse Y vahel $r(Z, Y)$ nimetatakse **biseriaalseks korrelatsiooniks** ja arvutatakse

$$r_b = \sqrt{p(1-p)}\ \frac{\bar{y}_1 - \bar{y}_2}{s_y u},$$

kus $u = \frac{\exp(-h^2/2)}{\sqrt{2\pi}}$ ja lävi h rahuldab tingimust $P(Z \geq h) = p$, eeldusel, et mittemõõdetav pidev tunnus $Z \sim N(0, 1)$. Biseriaalse korrelatsiooni olulisuse kontrollimiseks kasutatakse statistikut, mis on asümptootiliselt normaaljaotusega.

2.4 Korrelatsioonimaatriks

Enamasti analüüsitakse korraga rohkem kui kahte tunnust ja leitakse iga kahe tunnuse vahel korrelatsioonikordaja, saadud kordajad paigutatakse tabelisse, mida nimetatakse korrelatsioonitabeliks ehk *korrelatsioonimaatriksiks*. Selles tabelis vastab igale reale ja igale veerule üks tunnus. Korrelatsioonimaatriksit tähistatakse tavaliselt $\mathbf{R} = \{r_{ij}\}, i, j = 1, \ldots, m$, kus r_{ij} tähistab maatriksi i-ndas reas ja j-ndas veerus paiknevat elementi ehk korrelatsioonikordajat i-nda ja j-nda tunnuse vahel.

Korrelatsioonimaatriksi analüüsimisel võib vaadelda mitut aspekti:

- **Üldine seoste tugevus ja olulisus**. Pakub huvi kui tugevalt on kõik tunnused omavahel seotud. Seose suund sel korral tavaliselt pole tähtis. Lähtutakse korrelatsioonikordaja suuruse empiirilistest hinnangutest. Enamasti domineerivad maatriksis nõrgad ja keskmised seosed. NB! Kogu korrelatsioonimaatriksi olulisuse hindamisel tuleks arvesse võtta, et tahame teha üheaegselt palju otsustusi (kõigi korrelatsioonikordajate jaoks) ja seega kuhjuvad iga üksiku otsustuse tegemisel võimalikud vead. Vältimaks vea kasvamist tuleks otsustamise kriteerium muuta rangemaks (st valida väiksem olulisuse nivoo) vt. A-M.Parring, M.Vähi. Korrelatsioonimaatriksi ohtlikud olulisuse tõenäosused. ESS Teabevihik. No 5. Tartu, 1995, lk 24-32

- **Üksiktunnuste seoste tugevus**. Huvi pakuvad maatriksis ainult mõned tunnused ja nende seosed. Vaadates need läbi, saame teatud pildi meid huvitavate tunnuste vaheliste seose kohta.

- **Ühe tunnuse kirjeldamine teiste järgi**. Juhul, kui meil on üks uuritav tunnus, mida tahame esitada teiste kaudu, siis korrelatsioonimaatriksi järgi võime otsustada, millised tunnused võtta mudelisse. Tunnuste valikul mudelisse tuleks jälgida:

  - mudelisse võetakse uuritava tunnusega tugevalt seotud tunnused;
  - mudelisse võetakse uuritava tunnusega nii positiivselt kui negatiivselt seotud tunnuseid;
  - kui kaks tunnust on omavahel väga tugevalt seotud ($r > 0.9$), siis neist võtta mudelisse ainult üks.

Korrelatsioonimaatriksi näide (SAS: proc Corr, $m = 6$)

Pearson Correlation Coefficients, N = 445
Prob > |r| under H0: Rho=0

	LAPSI	TAISK	TOOTAB	HARIDUS	ASULA	palk
LAPSI	1.00000	0.05415	-0.05433	-0.05724	-0.14899	-0.18312
		0.2543	0.2528	0.2282	0.0016	0.0001
TAISK	0.05415	1.00000	0.48375	-0.00127	-0.03127	0.03421
	0.2543		<.0001	0.9788	0.5105	0.4717
TOOTAB	-0.05433	0.48375	1.00000	0.08243	0.03274	0.33690
	0.2528	<.0001		0.0824	0.4909	<.0001
HARIDUS	-0.05724	-0.00127	0.08243	1.00000	0.10892	0.19285
	0.2282	0.9788	0.0824		0.0216	<.0001
ASULA	-0.14899	-0.03127	0.03274	0.10892	1.00000	-0.02761
	0.0016	0.5105	0.4909	0.0216		0.5612
palk	-0.18312	0.03421	0.33690	0.19285	-0.02761	1.00000
	0.0001	0.4717	<.0001	<.0001	0.5612	

Asula 1- linn, 0-maa; Haridus 1-põhi,2-kesk, 3-keskeri, 4-kõrg
Lapsi - laste arv peres; TAISK - täiskasvanute arv peres
TOOTAB - töötavate pereliikmete arv

Millistest statistiliselt olulistest seostest saame siin rääkida?

Märkused

- Statistiliselt oluline korrelatsioon võib olla nõrk ning seega praktiliselt mitte huvi pakkuv.
- Tugev seos ei pruugi veel olla statistiliselt oluline st ta ei kehti üldkogumis, vaid on juhuslik seos valimis.
- Statistiliselt oluline korrelatsioon ei ole samaväärne põhjusliku seosega tunnuste vahel, ta võib aga viidata põhjuslikule seosele.
- Korrelatiivne seos tähendab, et muutus ühes tunnuses *kaasneb* teise tunnuse muutumisega. Korrelatiivse seose olemasolu ei tähenda, et ühe suuruse muutus põhjustab teise muutumise.
- Põhjuslik seos ehk deterministlik seos on seos, mille korral üks nähtus on põhjus ja teine tagajärg.
  - Põhjus avaldab mõju tagajärjele, põhjuslik seos on alati kindla suunaga.
  - Erinevate, omavahel põhjuslikult mitteseotud suuruste koosmuutumine võib olla põhjustatud mingi kolmanda suuruse poolt, mis mõjutab vaadeldavat kaht suurust.
  - Kui korrelatiivne seos on tugev, vihjab see küll põhjusliku seose võimalusele, ent ei tõesta veel selle olemasolu.
  - Põhjuslikke seoseid hinnatakse struktuurivõrrandite mudelite abil.
- Kui Pearsoni korrelatsioonikordaja ja Spearmanni korrelatsioonikordaja vahel on suur erinevus, on soovitav kasutada Spearmanni korrelatsioonikordajat.
- Kuigi korrelatsioonikordaja absoluutväärtus on 1, ei tähenda see seda, et kõikide andmete korral saab maksimaalne korrelatsioonikordaja olla 1, tihti jääb ta reaalselt palju väiksemaks.
- SAS väljastab erinevad korrelatsioonikordajaid seosekordajatena kahemõõtmelise sagedustabeli järele protseduuriga FREQ. Korrelatsioonimaatriksi arvutamiseks on aga sobivam kasutada protseduuri CORR.

Peatükk 3

Valiidsus ja reliaablus

3.1 Mõisted

Valiidsuse ja reliaabluse hindamise probleemidega on tegeldud juba pikka aega ja seda eeskätt psühholoogias seoses mitmesuguste psühhomeetriliste testide kasutamisega. Esimesed sellealased artiklid on 1930. aastast. Viimasel ajal on hakatud valiidsuse ja reliaabluse hindamisele suuremat tähelepanu pöörama ka teistes valdkondades peale psühholoogia ja seda eriti meditsiinis. On välja töötatud mitmed erinevad valiidsuse ja reliaabluse kordajad ning hindamismeetodid. Valiidsusest ja reliaablusest räägitakse mingi metoodika, testi või skaala korral.

Valiidsus (*validity*) tähistab metoodika paikapidavust, kehtivust või adekvaatsust.

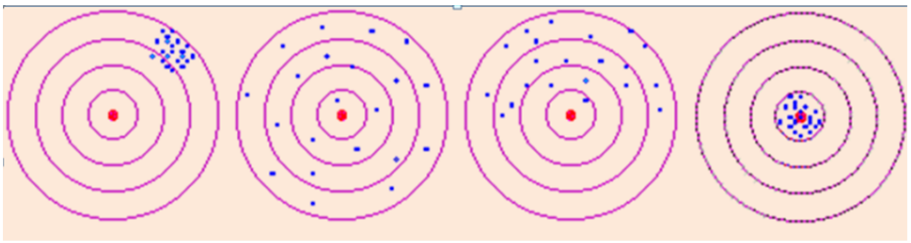
Rääkides valiidsusest mõeldake mõõtmiste korrektsust. Valiidsus näitab, missugusel määral mõõdab metoodika seda, mida ta on plaanitud mõõtma. Valiidsuse ulatust näitab süstemaatiline viga ehk nihe

Reliaabluse (*reliability*) all mõistetakse kasutatava metoodika stabiilsust, järjekindlust, kooskõla või töökindlust.

Reliaablust hinnatakse nii ühe metoodika korduval kasutamisel ühe uurija poolt kui ka ühe metoodika ühekordsel kasutamisel erinevate uurijate poolt. Mõlemal juhul tekib probleem, kas ja kuivõrd mõõtmistulemused on usaldatavad.

Valiidsuse ja reliaabluse tugevust mõõdetakse teatud seosekordajatega, mille absoluutväärtus on nulli ja ühe vahel. Mida suurem on kordaja väärtus, seda kõrgem on vastavalt valiidsus või reliaablus. Ideaalse valiidsuse ja reliaablusega uuringut pole võimalik teha. Üldiselt kui vaatame ühe metoodika valiidsuse ja reliaabluse suurust iseloomustavaid kordajaid, siis alati valiidsuse kordaja jääb arvuliselt väiksemaks kui reliaabluse kordaja.

Valiidsuse ja reliaabluse vahekorda illustreeritakse tavaliselt märklauaga. Eesmärk on tabada märklaua keskpunkt – kui tabatakse, siis on katse valiidsus kõrge ja kui tabatakse keskpunkti alati, siis on katse reliaablus kõrge.



A. Kõrge reliaablus Madal valiidsus | **B. Kõrge valiidsus madal reliaablus** | **C. Madal valiidus madal reliaablus** | **D. Kõrge reliaablus kõrge valiidus**

Valiidsus ja reliaablus on teineteist täiendavad mõisted ja omavahel seotud.

NB! Oht, et uuringut peetakse usaldusväärseks kui kordusuuringutega saadakse sama tulemus, aga võib olla tegemist ka situatsiooniga A (vt joonist) st saadakse küll sama tulemus, aga see pole õige tulemus.

3.2 Valiidsus

Kas metoodika/test/skaala mõõdab seda, mida ta on määratud mõõtma?

Kirjandusest võib leida suure hulga valiidsuse erinevaid liigitusi ja kirjeldusi.

Siinkohal vaatame kahte olulisemat võimalust:

A. Definitsioonil põhinev ehk seletav valiidsus.

- Väline valiidsus (*Face validity*) – lähtutakse väitest, et 'see näib nii olevat'
- Sisuvaliidsus (*Content validity*) – lähtutakse väitest, et 'see näib esitavat asja olemust'

Otsustamisel kasutatakse subjektiivseid hinnanguid, tavaliselt eksperthinnangud.

B. Kriteeriumitel põhinev valiidsus (*criterion validity*)

- On võimalik kasutada teisi meetodeid antud fenomeni uurimiseks. Sel juhul räägitakse valiidsusest kui korrelatsioonist nn "kuldse standardiga", selliselt on valiidsus defineeritud ka matemaatilise statistika leksikonis (Kotz, 1988). Seega, eksisteerib teatud norm ning mõõdetakse kooskõla normiga. Antud lähenemise korral tekivad teatud tunnetuslikud küsimused, millele tegelikult objektiivset vastust polegi.

1. Kui nähtuse mõõtmiseks juba leidub mingi "kuldne standard", milleks siis otsitakse uusi meetodeid? Selle küsimuse vastuseks on enamasti – uus meetod on parem ja odavam.
2. Kui uus meetod arvatakse olevat parem, milleks siis võrrelda teda vanaga ning kui seos uue ja vana meetodi vahel on nõrk, kumb meetoditest siis ikkagi on halb?

- Ei ole teisi meetodeid antud fenomeni uurimiseks. Sel juhul on valiidsus kui kaudne hüpoteetiline seos (*construct validity*). Selline olukord tekib juhul, kui mingis valdkonnas tehakse esimene uuring ja pole veel välja kujunenud mingeid standardeid.

  Näiteks 1920ndatel aastatel testiti esmakordselt biokeemiliste testidega veresuhkur. Polnud midagi, millega võrrelda, sest polnud teist veresuhkru testi. Püstitati hüpotees (mida kontrolliti empiiriliselt): haigetel, kellel on diagnoositud diabeet kliiniliste kriteeriumite järgi, on ka kõrgem veresuhkru sisaldus.

3.3 Reliaablus

Kas metoodika/test/skaala annab tema korduval rakendamisel või erinevate uurijate korral analoogilisi tulemusi?

Siin saame rääkida reliaablusest kahest aspektist:

A. Sisemise kooskõla, järjekindluse hindamine (*internal consistency*). Hinnatakse, kas tunnused/testid, mis koos peavad mõõtma mingit fenomeni, on kooskõlas. Enamlevinud on järgmised reliaabluse kui sisemise kooskõla kordajad: Cronbachi α, Kuder-Richardsoni ning Spearman-Browni ja Guttmani kordajad.

B. Stabiilsuse hindamine (*inter-oserver, intra-observer reliability, test-retest*). Hinnatakse, kas metoodika/test/skaala on püsiv erinevatel mõõtmistel. Stabiilsuse hindamisest räägitakse siis, kui üks uurija teeb sama metoodikaga uuringuid erinevatel kontingentidel või samal kontingendil erinevatel ajamomentidel või erinevad uurijad teevad sama metoodikaga uuringuid. Probleem seisneb selles, et ajas võib metoodika sisu või interpretatsioon muutuda ning saadakse võrreldamatud andmed. Taolisest olukorra vältimiseks tuleb hinnata metoodika stabiilsust.

Praktilise kasutamise jaoks on paljud erinevad autorid pakkunud välja teatud empiirilised reliaabluse kordaja suurused, mis näitavad, milline peaks olema hea metoodika reliaablus.

Reliaabluse kordaja r võimalikud väärtused on vahemikus $0 \leq r \leq 1$.

Metoodika sisemine kooskõla loetakse heaks, kui ta on vähemalt 0.7 kuni 0.8 ning metoodika stabiilsuse näitaja peaks olema vähemalt 0.5. Need ei ole kahtlemata mingid absoluutnormid, vaid teatavad orienteeruvad hinnangud.

Reliaabluse mõõtmine

Klassikaline lähenemine reliaabluse mõõtmisele on järgmine. Iga mõõtmine kajastab mingil määral õiget vastust küsimusele ja teatud osas muud ehk juhuslikku viga.

Seega $X = T + E$, kus X on suvaline mõõtmine, T on õige väärtus (latentne ehk mittemõõdetav), E on juhuslik viga. Lisaks eeldatakse, et õige vastus ja juhuslik viga on sõltumatud ning erinevatel mõõtmistel tekkinud vead on samuti sõltumatud. Reliaabluse kordajat vaadatakse sel juhul kui tundmatu õige väärtuse hajuvuse ja mõõdetud vastuse hajuvuse suhet

$$\frac{\sigma^2(T)}{\sigma^2(X)},$$

kus σ^2 tähistab dispersiooni. Tundmatu õige väärtuse hajuvust hinnatakse kaudselt erinevate meetoditega ja nii jõutakse reliaabluse mõõtmise erinevate kordajateni.

Tuntuimad reliaabluse kordajad on järgmised.

- **Cronbachi α kordaja** (1951):

  $$\alpha = \frac{k}{k-1}\left\{1 - \frac{\sum_i s_i^2}{s^2}\right\},$$

  kus k – tunnuste arv, s_i^2 – $i$nda küsimuse/tunnuse hajuvus, s^2 – kogutesti hajuvus;

- **Kuder–Richardsoni kordajad** KR20, KR21 (1937) (mis on tegelikult Cronbachi a kordaja analoogid binaarse tunnuse jaoks);

- **"Poolte/poolitatud"reliaablus** (*split–half*). Kordajate ühisel nimetusel puudub eesti keeles õige vaste, nimetus on tuletatud asjaolust, et antud juhul test jaotatakse kaheks pooleks ja hinnatakse kahe testipoole kooskõla.
  Test on kooskõlas kui tema mõlemad pooled on omavahel kõrgelt korrelleeritud.

  Spearman–Browni kordaja

  $$r_{SB} = \frac{2r_{XY}}{1 + r_{XY}},$$

  kus r_{XY} on korrelatsioon kahe poole vahel.
  Seda kordajat kasutatakse juhul, kui mõlemas testi pooles on ühesugune hajuvus

Guttmani kordaja:

$$r_G = \frac{2(s_t^2 - s_{t_1}^2 - s_{t_2}^2)}{s_t^2},$$

kus test t on jagatud kaheks pooleks t_1 ja t_2, s_t^2 on kogu testi hajuvus, $s_{t_1}^2, s_{t_2}^2$ testi esimese ja teise poole hajuvused.
Valemit kasutatakse kui testi pooltes on erinev hajuvus.

- **Korrelatsioon tunnuse ja testi vahel** kui testi homogeensuse või stabiilsuse näitaja Nunally (1978) (*item–total correlation*)

$$r_{i(t-1)} = \frac{r_{it}s_t - s_i}{\sqrt{s_i^2 + s_t^2 + 2s_i s_t r_{it}}},$$

kus $r_{i(t-1)}$ on korrelatsioon i-nda tunnuse ja testi vahel, kust i-s tunnus on välja jäetud; r_{it} on korrelatsioon i-nda tunnuse ja testi vahel; s_i^2 on i-nda tunnuse hajuvus; s_t^2 on kogutesti hajuvus.

Cronbachi α kasutamine testi sisemise kooskõla hindamiseks:

Arvutatakse Cronbachi α_i^* iga tunnuse korral selliselt, et vaadeldav i-s tunnus jäetakse välja (*Cronbach Coefficient Alpha with Deleted Variable*) ja võrreldakse seda kogutesti Cronbachi α väärtusega. Loomulik on oletada, et kui tunnus on testis vajalik (sobib testi, on kooskõlas teiste testis olevate tunnustega), siis tema väljajätmisel testi kooskõla väheneb ja kui tunnus pole testis vajalik, siis tema väljajätmisel testi kooskõla paraneb.

Otsustusreegel:

- Kui $\alpha_i^* > \alpha$, siis i-s tunnus pole testis vajalik
- Kui $\alpha_i^* < \alpha$, siis i-s tunnus on testis vajalik

Analüüs viiakse läbi nii algväärtusi (*Row*) kui ka standardiseeritud (*Standardized*) väärtusi kasutades. Kui testi küsimuste vastused on väga erineval skaalal, siis on soovitav kasutada standardiseeritud väärtustele rakendatud analüüsi (sest standardiseerimine muudab skaalad ühesuguseks).

3.4 Kooskõla hindamisest

On välja töötatud terve rida kriteeriume hindamaks teatavat klassifitseerimise kooskõla (*classification agreement*). Hinnatakse kahe erineva uurija tulemuste või ühe uurija kahe erineva hinnangu koooskõla (*agree*) või seost.

Vastavad hinnangustatistikud esitatakse kahemõõtmelise sagedustabeli erijuhu - ruut-tabelite (ridade ja veergude arv on võrdne) - analüüsi tulemusena. Kahemõõtmelises sagedustabelis on sel juhul nii veeru kui reatunnuseks sama skaala, millel on hindamine toimunud, aga hindajad on erinevad.
Seega on kxk tabelil järgmine kuju

	1	...	k	
1	n_{11}	...	n_{1k}	$n_{1\cdot}$
$\vdots$				
k	n_{k1}	...	n_{kk}	$n_{k\cdot}$
	$n_{\cdot 1}$	...	$n_{\cdot k}$	n

Uurijate vahelise kooskõla hindamiseks kasutatakse järgmist kordajat

Kapa kordaja

$$\kappa = \frac{P_O - P_e}{1 - P_e},$$

kus $P_O = \sum_i \frac{n_{ii}}{n}$ on kooskõlas olevate vastuste vaadeldud (*observed*) sagedus ja
$P_e = \sum_i \frac{n_{i\cdot} n_{\cdot i}}{n^2}$ on teoreetiline ehk oodatav (*expected*) kooskõla vastuste sõltumatuse korral.

Kui uurijate hinnangud on sõltumatud ja täielikus kooskõlas, siis kapa kordaja väärtus on +1. Kui tegelik kooskõla on madalam kui teoreetiliselt võimalik, siis kapa kordaja on negatiivne. Saab anda hinnangu ka kordaja hajuvusele.

(vt A. Agresti (1990) Categorical Data Analysis. N.Y., John Wiley)

Kaalutud kapa kordaja

Kaalutud kapa kordaja on tavalise kapa kordaja üldistus, kus kasutatakse kaale hindamaks suhtelisi erinevusi hinnangute vahel. Saab anda ka teatud hinnangu kaalude konstrueerimiseks.

Peatükk 4

Lineaarne regressioon

Statistilise mudeli üldkuju on järgmine

$$Y = f(X, \alpha, \beta, \ldots) + \varepsilon,$$

kus Y on funktsioontunnus ehk uuritav tunnus ehk sõltuv tunnus (*response, outcome, dependent variable*), mis on uurimuse seisukohalt eriti huvipakkuv ja mida soovitakse mudelina kirjeldada.

$f(\cdot)$ tähistab mingit funktsiooni. Tavaliselt alustatakse lineaarsest funktsioonist. Mudeli valik sõltub uuritava tunnuse tüübist ja jaotusest.

X on argumenttunnus ehk seletav tunnus ehk sõltumatu tunnus (*predictor, explanatory variable, independent variable*), st tunnus, millest sõltub funktsioontunnus. Argumenttunnuseid võib mudelis olla ka mitu. Enamasti *kitsendusi argumenttunnuste tüübile ja jaotusele ei ole.*

$\alpha, \beta, \ldots$ on mudeli (tundmatud) parameetrid.

ε on mudeli juhuslik viga (*random error*).

Statistilise mudeli leidmine seisneb järgmiste sammude läbimises:

- Mudeli kuju määramine: sõltuva tunnuse fikseerimine, argumentide valik mudelisse.
- Mudeli parameetritele väärtuste arvutamine statistiliste andmete põhjal.
- Mudeli ja tema parameetrite statistilise olulisuse kontroll, st kas nad üldkogumis erinevad oluliselt nullist.
  Statistiliselt oluline mudel kehtib ka üldkogumis, mitteoluline kirjeldab seoseid ainult valimi kohta. *Kui mudel pole oluline (või sisaldab mitteolulisi liikmeid), tuleb proovida leida uus mudel.*
- Mudeli headuse (ja täpsuse) hindamine - tehakse kindlaks, kui suure osa uuritava tunnuse hajuvusest mudel kirjeldab.

- Mudeli diagnostika, erindite analüüs (vaatluste väljaselgitamine, mille korral mudel töötab halvasti). Mudeli eelduste kontroll.
- Mudeli sisuline tõlgendamine.
- Mudeli kasutamine (progoosimine)

4.1 Lihtne lineaarne regressioonimudel

Vaatame ühe argumendiga lineaarset regressioonimudelit ehk lihtsat regressioonimudelit (*simple regression*) kujul

$$\boxed{Y = \alpha + \beta X + \varepsilon}$$

kus Y on funktsioontunnus, X on argumenttunnus (arvuline),
α on vabaliige (*intercept*), β on regressioonikordaja (*regression coefficient, slope*), ε on mudeli juhuslik viga.

Mudeli parameetrid hinnatakse **vähimruutude meetodil**: erinevused tegelikult mõõdetud uuritava tunnuse väärtuste ja mudeli järgi prognoositud väärtuste vahel minimiseeritakse.

Lahendatakse ekstreemumülesanne parameetrite α ja β suhtes:

$$\sum_i [y_i - (\alpha + \beta x_i)^2] \Rightarrow min.$$

Jõutakse **normaalvõrranditeni** ja saadakse parameetritele hinnangud $\hat{\alpha}, \hat{\beta}$ (hinnanguid tähistatakse ka a, b)

Vähimruutude meetodi korral pole vaja teha mingeid eeldusi mudelis olevate tunnuste jaotuste kohta. Meetod töötab alati ja on teatud mõttes parim.

Kaugusi mõõdetakse traditsiooniliselt y-telje sihis.

Vähimruutude meetodi ajaloost

Meetodi töötas välja Johann Carl Friedrich Gauss, kes oli Saksa matemaatik, astronoom ja füüsik, elas aastatel 1777–1855. Ta arvutas 1801. a. detsembris Ceres'i asteroidi orbiidi kasutades vähimruutude meetodit (suurim asteroid, avastati G. Piazzi poolt 1801 jaanuaris, aga kadus ja Gauss leidis selle uuesti). 1809. a. avaldas Gauss vähimruutude meetodi idee.

4.1.1 Regressioonimudeli olulisus

Korrelatsioonikordaja olulisusest järeldub vahetult ka regressioonikordaja olulisus ja vastupidi. Kui korrelatsioonikordaja ei ole oluline, ei ole oluline ka regressioonikordaja ning *regressioonimudel tervikuna ei ole statistiliselt oluline*.

Vabaliikme olulisust on tarvis eraldi kontrollida. Mudeli kui terviku olulisus ei sõltu sellest, kas vabaliige on oluline või mitte.

Mudel $Y = bX$, kus $a = 0$, võib pakkuda huvi. Mudel, kus $b = 0$, st $Y = a$, ei paku huvi, sest ei sisalda argumenti X.

Igasuguste järelduste tegemiseks on korrektne kasutada regressioonimudelit vaid siis, kui ta on statistiliselt oluline!

Mudeli olulisuse kontrolli saame sõnastada kui hüpoteeside paari

$H_1 : \beta \neq 0; \quad H_0 : \beta = 0.$

Hüpoteesipaari kontrollimiseks on teststatistik kujul

$$t = \frac{b}{s_b} \sim t_{n-2}, \quad s_b - \text{regressioonikordaja standardhälve.}$$

Arvutipaketid väljastavad mudeli jaoks *dispersioonanalüüsi tabeli*, kus viimases veerus on mudeli olulisuse tõenäosus ja eraldi *parameetrite hinnangute tabeli*, kus viimases veerus on parameetrite olulisuse tõenäosused.

4.1.2 Mudeli täpsus

Regressiooniseose täpsus sõltub juhusliku vea hajuvusest st oluline on juhusliku vea standardhälve.

Mudeli juhusliku vea hinnangut nimetatakse mudeli **jäägiks** (*residual*) $e_i = y_i - \hat{y}_i$, kus $\hat{y}_i = a + by_i$ on *prognoos* mudelist.

Juhusliku vea dispersiooni hinnang (*MSE – Mean Squared Error*, keskmine ruutviga) avaldub järgmiselt:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2.$$

Mudeli standardviga (*Root MSE*) on ruutjuur keskmisest ruutveast

$$s = \sqrt{s^2} = \sqrt{MSE}.$$

Mudeli standardviga (juhusliku vea standardhälve) iseloomustab uuritava tunnuse kõrvalekallet mudeliga määratud väärtusest.

4.1.3 Mudeli headus. Determinatsioonikordaja

Mudeli headuse näitajaks on **determinatsioonikordaja** R^2 (*R-squared*).

Determinatsioonikordaja näitab, *kui suure osa uuritava tunnuse Y hajuvusest (dispersioonist) kirjeldab lineaarne regressioonimudel* (öeldakse ka mitu protsenti kirjeldab mudel).

Determinatsioonikordaja arvutamisel lähtutakse koguhajuvuse lahutusest mudeli poolt kirjeldatud hajuvuseks (SSR) ja vea poolt kirjeldatud hajuvuseks (SSE): $SST = SSR + SSE$ ja definitsiooni järgi $R^2 = SSR/SST$. Koguhajuvuse avaldist silmas pidades arvutatakse determinatsioonikordaja tavaliselt järgmiselt: $R^2 = 1 - SSE/SST$ ja alati $0 \leq R^2 \leq 1$.

Ühe argumendiga mudeli headust näitab ka lineaarne korrelatsioonikordaja r. Korrelatsioonikordaja r ning regressioonikordaja b on alati samamärgilised. Kui $r > 0$ ($b > 0$, tõusev sirge), siis argumenttunnuse X suurenedes (vähenedes) keskmiselt suureneb (väheneb) ka uuritav tunnus Y. Kui $r < 0$ ($b < 0$, langev sirge), siis argumenttunnuse X suurenedes keskmiselt uuritav tunnus Y väheneb ja argumenttunnuse X vähenedes keskmiselt uuritav tunnus Y suureneb. Ühe argumendiga lineaarse regressioonimudeli korral kehtib $r^2 = R^2$.

4.1.4 Jääkide analüüs ja mudeli eelduste kontroll

Mudeli **jääk** (*residual*) e on mudeli juhusliku vea ε hinnang $e = \hat{\varepsilon}$

$$e_i = y_i - \hat{y}_i.$$

Standardiseeritud jäägid (*standardized residuals*), saadakse jäägi jagamisel tema standardhälbega

$$e_i^{stand} = \frac{e_i}{s_{e_i}}.$$

Studenti jäägid (*studentized residuals*), saadakse jäägi jagamisel tema standardhälbega, mille arvutamisel on antud vaatlus välja jäetud

$$e_i^{stud} = \frac{e_i}{s_{e_{(i)}}}.$$

Jäägi standardhälve avaldub kujul $s_{e_i} = s\sqrt{1 - h_i}$, h_i – vaatluse omapära (*leverage*).

Jääkide analüüsi eesmärk:

- Mudeli eelduste kontroll.
- Mudeli sobivuse hindamine, erindite kontrollimine.

Mudeli eelduste kontroll

Mudeli eeldused:

- juhuslikud vead on erinevate vaatluste korral sõltumatud
- juhuslike vigade keskväärtus on null ($E\varepsilon_i = 0,\ \forall i$)
- juhuslike vigade hajuvus on konstantne ($D\varepsilon_i = \sigma^2,\ \forall i$)

Juhuslikud vead peavad olema sõltumatud ja normaaljaotusega $\varepsilon \sim N(0, \sigma^2)$

Graafilised meetodid: kasutatakse mitmesuguseid jääkide graafikuid

(1) Jääk *vs* prognoos või jääk *vs* argument – kui juhuslike vigade hajuvus on ühesugune, peavad jääkide graafikud kujutama endast hajusat punktiparve. Öeldakse ka püsihajuv ehk *homoskedastiline* ja muuthajuv ehk *heteroskedastiline*. Mudeli jäägid peavad olema püsihajuvad ehk homoskedastilised.

(2) *Q-Q plot* mudeli jääkidele – kontrollimaks juhuslike vigade jaotust.

Normaaljaotuse testid (vt Loeng 1, punkt 1.4).

Kui on kõrvalekaldeid normaaljaotusest, soovitatakse skaalateisendusi.

4.1.5 Mudeli diagnostika. Erindid

Mudeli erindiks (*outlier*) nimetatakse teistest mingis mõttes erinevat vaatlust. Erinditest räägitakse kolmest aspektist.

- **Erind suure jäägi mõttes** (*outlier*)

Hindamiseks leitakse standardiseeritud ja/või Studenti jäägid. Jäägi kriitiline väärtus on ligikaudu 3 ($e_i^{stand} \approx 3$), siis mudel ei tööta hästi (räägitakse erindist y-telje sihis).

- **Omapärane vaatlus** (*leverage*)

Omapära h_i määrab vaatluse i jaoks tema kauguse argumenttunnuse keskväärtusest

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}.$$

Erilised on need vaatlused, mille kaugused on suuremad kui teistel (räägitakse erindist x-telje sihis). Liiga omapärasteks loetakse need vaatlused, mille korral

$$h_i > \frac{4}{n}.$$

- **Mõjus vaatlus** (*influential*)

Vaatlus, mis mõjutab regressioonikordajat. Vaatluse mõju regressioonikordajale hinnatakse Cook'i statistikuga

$$D_i = \frac{(b - b_{(i)})^2}{s_b^2},$$

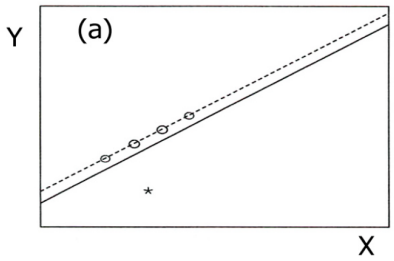
kus b on mudeli regressioonikordaja ja $b_{(i)}$ on regressioonikordaja, mis on leitud i-ndat vaatlust mitte arvestades.

Mõjusateks loetakse need vaatlused, mis vastavad F-jaotuse $F_{p,n-p}$ 50 protsentpunktile või suuremale.

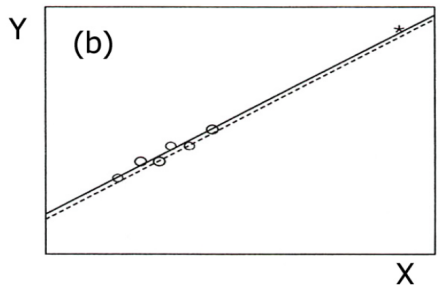
Ühel või teisel viisil leitud teistest erinevad vaatlused tuleb üle kontrollida ja välja selgitada, kas on tegemist tõepoolest mingi erilise vaatlusega või veaga andmetes.

Erind – omapärane vaatlus – mõjus vaatlus (Allikas: Fox, 1997)

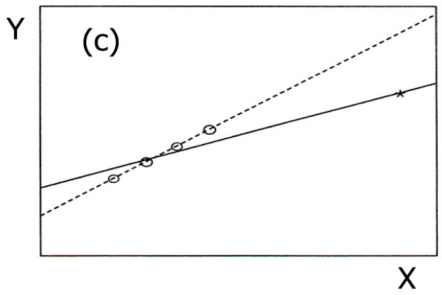
- Erind suure jäägi mõttes ei pea olema tingimata omapärane või mõjus vaatlus (vt järgmine joonis (a))



- Kõrge omapäraga vaatlus ei pea olema mõjus või erind suure jäägi mõttes (vt järgmine joonis (b))



- Vaatlus võib aga olla nii erind suure jäägi mõttes kui ka omapärane ja mõjus vaatlus (vt järgmine joonis (c))



Adapted from Figure 11.1 (Fox, 1997)

NB! *Ettevaatust vaatluste väljajätmisel! Eesmärk pole leida ilusat mudelit, vaid andmetele vastavat mudelit!*

Usaldusvahemikud

Usaldusvahemikke saab leida

1. Regressioonikordajatele
2. Prognoosile
3. Üksikväärtusele

vt Parring, Vähi, Käärik (1997). Statistilise andmetöötluse algõpetus, lk 241–245. Olemas ka e-raamat TÜ raamatukogus, kättesaadav TÜ arvutivõrgus.

4.1.6 Mudeli parameetrite interpreteerimine

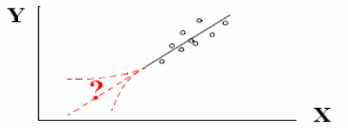
Mudeli sisuline tõlgendamine ehk mudeli parameetrite interpreteerimine on sirge võrrandi interpreteerimine. Mudeli vabaliige ehk lõikepunkt y-teljega vastab olukorrale kui $X = 0$ ja on ainult siis interpreteeritav, kui selline argumendi väärtus andmetes esineb.

Regressioonikordaja on sirge tõus, seega saame öelda:

Kui vaatame kahte objekti, kus argumendi väärtused erinevad ühiku võrra, siis vastavad uuritava tunnuse väärtused erinevad regressioonikordaja võrra.

Prognoosimine

Prognoosimine on mudeli kasutamine, et arvutada uusi väärtusi funktsioontunnusele nendele argumentidele vastavalt, mille kohta pole mõõtmisi tehtud. Prognoosimisel tuleb silmas pidada, et *prognoosida saab ainult argumentide muutumispiirkonna ulatuses* st selles piirkonnas, kus tegelikult mudel on leitud.



Me ei tea, milline on mudeli kuju väljaspool argumentide piirkonda.

Näiteks, kui leitakse mudelit kaalu jaoks kasvu ja vanuse järgi ning andmestikus on ainult 20-30 aastased naised, siis ei saa me selle mudeliga prognoosida ei 50-aastase naise kaalu ega ka ühegi mehe kaalu.

4.1.7 Regressioonimudeli ajaloost

Esmakordselt kasutas regressiooni mõistet Sir Francis Galton (1822-1911). Ta oli väga laialdaste teadmistega geograaf, meteoroloog, troopikauurija, psühholoog, näpujälgede identifitseerimise teoreetiliste aluste rajaja (hindas 1888. aastal tõenäosust, et 2 sõrmejälge on samad), regressiooni- ja korrelatsiooniteooria aluste rajaja.

1875 a. tegi ta eksperimendi suhkruhernestega. Tuntud on tema antropomeetrilised katsed (1885-1886), kus ta tegeles vanemate ja nende järglaste kasvu uurimisega. Galton tõi sisse ka korrelatsiooni mõiste, hiljem avaldas selle Carl Pearson 1930, seepärast räägime Pearsoni korrelatsioonikordajast.

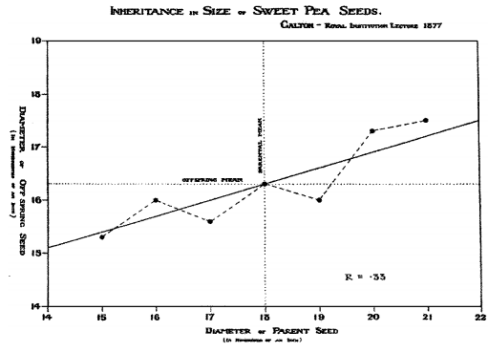


Figure 3 The first regression line. (Reprinted from K Pearson. 1930. *The Life, Labours and Letters of Francis Galton*, 3A:4, with permission from Cambridge University Press.)

Esimene regressioonisirge. Allikas: Gillham, N.W. (2001). Sir Francis Galton and the birth of eugenics, *Annu. Rev. Genet.*, 35, pp 83–101

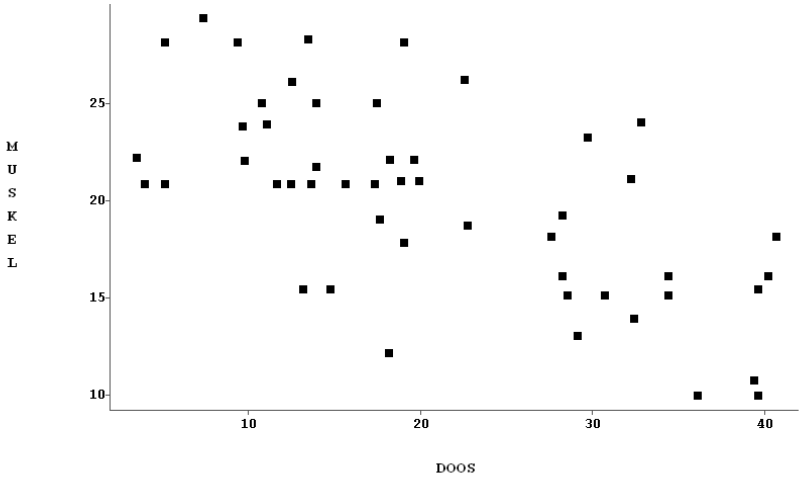
Näide. Ühe argumendiga lineaarne regressioonimudel

Uuriti alkoholismiga seotud probleeme. Valiti juhuslikult 50 meest, kelle päevane alkoholi tarbimine kõikus 118-350 g vahel. Leiti iga mehe kohta tema kogu elu jooksul tarvitatud alkoholi kogus kehakaalu kg kohta (tunnus DOOS) .

Mõõdeti ka igal mehel elektroonilise müotonomeetriga mittedominantse deltalihase jõudu (Kg) (tunnus MUSKEL).

Arvutati tarvitatud alkoholikoguse ja muskli jõu vaheline korrelatsioonikordaja $r(Muskel, Doos) = -0.64,\ p > 0.0001$, seega on tegemist vastupidise seosega, mida rohkem on alkoholi tarvitatud, seda nõrgem on muskel.

Hajuvusgraafik



Ülesande lahendus SASis kasutades proc REG

```
ods graphics on;
proc reg data=aa2.alkohol plots(label)=(CooksD RStudentByLeverage);
model muskel=doos;  run;
ods graphics off;
```

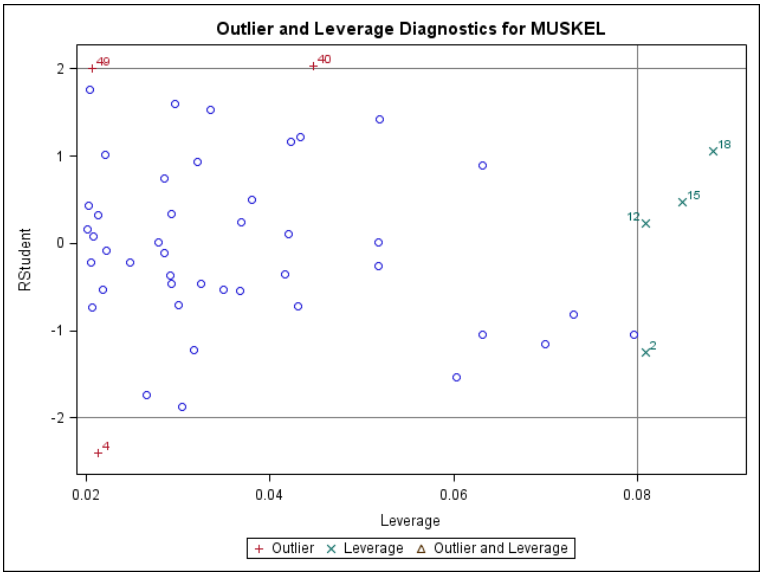
```
 Analysis of Variance
Source          DF   Sum of Squares  Mean Square    F Value  Pr > F
Model            1      504.04032     504.04032     33.59    <.0001
Error           48      720.27488      15.00573
Total           49     1224.31520


Root MSE          3.87372    R-Square     0.4117
Dependent Mean   20.16400    Adj R-Sq     0.3994   Coeff Var 19.21108
```

Parameter Estimates

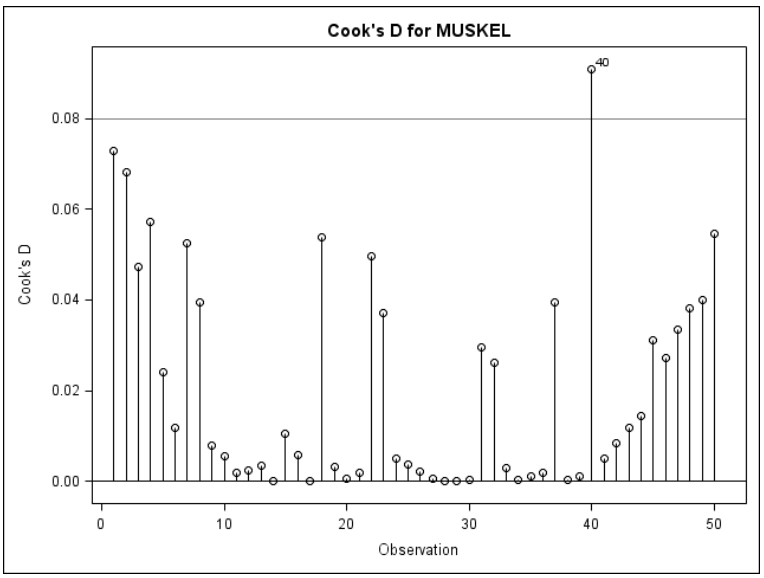
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	26.36954	1.20273	21.92	<.0001
DOOS	1	-0.29587	0.05105	-5.80	<.0001

Väljastatakse järgmised graafikud:

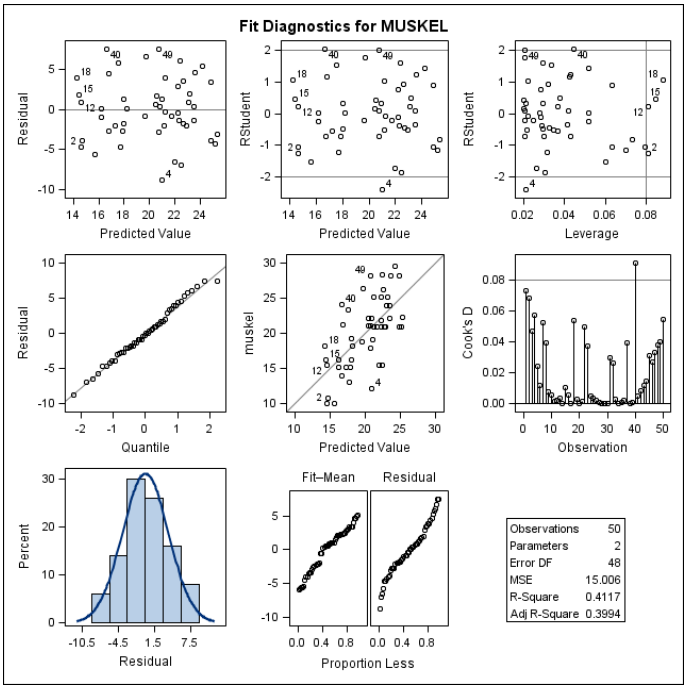
(1) Erindid ja omapärased vaatlused



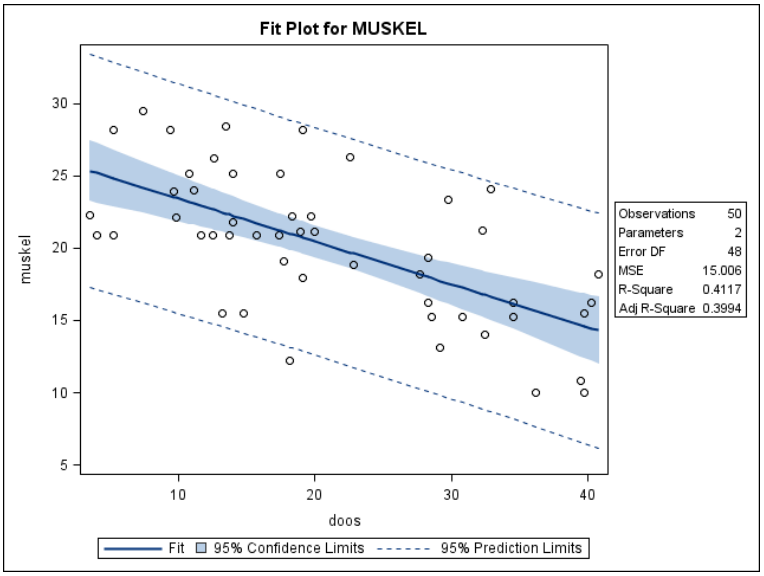
(2) Mõjusad vaatlused



(3) Diagnostika



(4) Usaldusvahemikud



4.2 Mitme argumendiga regressioonimudel

Tavaliselt pakub uurijale huvi mitme tunnuse üheaegne käitumine ja selle kirjeldamine. Soovitakse avaldada ühte (sõltuvat – *dependent, response*) tunnust Y teiste (sõltumatute, seletavate ehk argument – *independent, explonatory*) tunnuste $X_1, X_2, \ldots, X_k$ kaudu.

Kui sõltuv tunnus on pidev tunnus ja argumenttunnused arvulised, siis sobivaks mudeliks on mitme argumendiga lineaarne regressioonimudel:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i,$$

kus β_0 on vabaliige, β_j on regressioonikordajad, $j = 1, 2, \ldots, k$ (k on mudelis esinevate argumentide arv, mudeli parameetrite arv on $p = k + 1$) ja ε_i on juhuslik viga ($i = 1, 2, \ldots, n$; n on valimi maht). Juhuslike vigade kohta peavad olema täidetud järgmised eeldused:

- Juhuslike vigade keskväärtus on null $E\varepsilon_i = 0$.
- Juhuslikud vead on konstantse hajuvusega $D\varepsilon_i = \sigma^2$.
- Juhuslikud vead on sõltumatud $cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.

Regressioonimudeli parameetrite hindamiseks kasutatakse **vähimruutude meetodit** (*least square*), mis baseerub vähimruutude printsiibil:

Mudeli parameetrite väärtused valitakse selliselt, et erinevused tegelikult mõõdetud sõltuva tunnuse väärtuste ja mudeli järgi prognoositud väärtuste vahel oleks minimaalsed.

Mudeli võib lugeda konstrueerituks, kui leitud mudelis on kõik liikmed statistiliselt olulised. Ühe sõltuva tunnuse jaoks võib tihti leida mitu erinevat mudelit ning neist parima valimiseks on võimalik kasutada erinevaid kriteeriume:

- valitakse ökonoomseim (see, mis sisaldab kõige vähem argumente);
- valitakse parim näiteks determinatsioonikordaja (vms) mõttes;
- valitakse see, mida on lihtsam interpreteerida.

Erinev probleemi püstitus viib erinevate mudeliteni, seetõttu on oluline mudeli määramise juures konsulteerida vastava valdkonna spetsialistidega ja/või teostada andmete kirjeldav analüüs selgitamaks välja tunnuste omavahelisi seoseid ja käitumist.

Näide. Probleem on tõestada või ümber lükata, kas tööandja diskrimineerib naisi. Andmestikus on töötaja palk, töötaja kvalifikatsioon ja sugu. Antud juhul võib püstitada 2 erinevat küsimust, mis mõlemad on seotud antud probleemiga.

1. Kas keskmiselt saavad naised vähem palka kui sama kvalifikatsiooniga mehed? Sel juhul on sõltuvaks tunnuseks 'palk' ja seletavateks 'kvalifikatsioon' ning 'sugu'.

2. Kas keskmiselt on naistel kõrgem kvalifikatsioon, kui sama palgaga meestel? Sel juhul on sõltuvaks tunnuseks 'kvalifikatsioon' ja seletavateks 'palk' ning 'sugu'.

4.2.1 Mudeli matemaatiline esitus

Mitme argumendiga lineaarse mudeli konstrueerimise matemaatiline esitus antakse tavaliselt maatrikskujul. Maatrikskujul avaldub lineaarne regressioonimudel järgmiselt

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

kus $\mathbf{y} = (y_1, \ldots, y_n)^T$ on n–mõõtmeline funktsioontunnuse vektor, β on p–mõõtmeline tundmatute parameetrite vektor, ε on n–mõõtmeline juhuslike vigade vektor ja plaanimaatriks $\mathbf{X}$ on nxp–mõõtmeline.

Vähimruutude printsiibi realiseerimiseks minimiseeritakse vigade ruutude summad (*SSE – sum of squared errors*) $SSE(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$.

Minimiseerimisülesannet lahendades jõutakse normaalvõrrandisüsteemini $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{y}$, millel on ühene lahend parajasti siis, kui maatriksi veerud on lineaarselt sõltumatud st kui leidub pöördmaatriks $(\mathbf{X}^T\mathbf{X})^{-1}$ (korrelatsioonimaatriks peab olema positiivselt määratud).

Lahend avaldub kujul $b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Mudeli põhjal arvutatud sõltuva tunnuse väärtust nimetatakse **prognoosiks** (*prediction*) $\hat{\mathbf{y}} = \mathbf{X}b$. Saame anda ka hinnangud mudeli juhuslikele vigadele st leida **prognoosijäägid** (*residuals*) arvestades $\hat{\mathbf{y}}$ ja b avaldisi $e = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$, kus $\mathbf{I}$ on ühikmaatriks ja maatriks $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ kannab nimetust "mütsi"ehk "katuse"maatriks (*hat matrix*), mille peadiagonaali elementidel on eriline tähendus.

Hinnang juhuslike vigade hajuvusele (**keskmine ruutviga** *MSE – mean square error*) avaldub kujul $MSE = SSE/(n - k - 1)$ ja mudeli standardhälve ehk **mudeli täpsus** on ruutjuur sellest (*Root MSE*).

4.2.2 Multikollineaarsus

Olukorda, kus argumenttunnused (sõltumatud muutujad) on omavahel küllalt tugevalt seotud, nimetatakse **multikollineaarsuseks**. See tähendab, et me ei saa üheselt leida hinnangut b (ei leidu pöördmaatriksit $(\mathbf{X}^T\mathbf{X})^{-1}$), saame ebatäpsed hinnangud (võivad olla isegi vale märgiga) ja seega saame ebatäpsed prognoosid. Lisaks tekib probleeme regressioonikordajate tõlgendamisel. Regressioonikordaja näitab keskmist uuritava tunnuse muutust, mis

kaasneb vastava argumenttunnuse ühikulisele muutusele, kui teised argumenttunnused ei muutu. Argumenttunnuste sõltuvuse korral aga muutuvad argumendid üheaegselt.

Multikollineaarsusele andmestikus viitavad kõrged korrelatsioonid korrelatsioonimaatriksis ($r > 0.95$), mudeli regressioonikordajate suur hajuvus ning korrelleeritus. Multikollineaarsust esineb sageli ja see on tavaliselt valimi probleem. *Multikollineaarsust loetakse* ***suureks****, kui argumentide vaheline korrelatsioon on suurem kui samade argumentide ja uuritava tunnuse vaheline korrelatsioon.*

Multikollineaarsuse mõõtmiseks on mitu näitajat, neist tuntumad:

- **Tolerants** (*tolerance*) *TOL* on multikollineaarsuse mõõt, mis *näitab kui suur osa argumendi varieeruvusest jääb ülejäänud argumentide poolt kirjeldamata.* $TOL_i = 1 - R_i^2$, kus R_i^2 on selle mudeli determinatsioonikordaja, kus i-s tunnus on avaldatud teiste kaudu. Mida madalam tolerants, seda rohkem tunnus sõltub teistest, seda vähem ta annab uut informatsiooni. Suhteliselt madala tolerantsiga tunnuste rühm on omavahel seotud ja nende hulgast peaks mõne mudelist välja jätma.

- **Varieeruvusindeks** ehk dispersiooni mõju faktor (*variance inflation factor*) *VIF näitab argumendi mõju regressiooniparameetri hajuvusele* ja on tolerantsi pöördväärtus. Mida suurem on varieeruvusindeks, seda suurem on argumendi mõju. *VIF* on muutus regressioonikordaja hajuvuses, mis on tingitud teiste argumentide mõjust. Argumenttunnuste sõltuvuse tõttu on regressioonikordaja hajuvus *VIF* korda suurem kui oleks sõltumatuse korral.

Empiiriline kriteerium: kui $TOL < 0.15$ või $VIF > 10$ on tegemist multikollineaarsusega.

Multikollineaarsus esineb enamasti alati mitmese regressiooni mudelis ja küsimus pole seega selles, *kas ta esineb*, vaid selles, *kui tõsine ta on.*

Multikollineaarsust saab vähendada mudeli argumentide arvu vähendamisega ja omavahel tugevalt korreleeruvate argumentide mudelist väljajätmisega, kuid see ei ole hea lahendus!

Paremad võimalused multikollineaarsuse vähendamiseks:

- võtta rohkem andmeid, multikollineaarsus on enamasti väikese valimi probleem;

- defineerida uued tunnused (näiteks olemasolevate vahed jms);

- kantregressiooni kasutamine (*ridge regression*);
  et leiduks $(\mathbf{X}^T\mathbf{X})^{-1}$ lisatakse peadiagonaalile väike konstant ja saadakse normaalvõrrandid $(\mathbf{X}^T\mathbf{X} + k\mathbf{I})\beta = \mathbf{X}^T\mathbf{y}$, leitakse sobiv k, hinnang $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ väikese nihkega.

4.2.3 Mudeli olulisuse kontroll

Kontrollitakse hüpoteesidepaari parameetervektori β kohta

$H_0 : \beta = 0$, mudel ei ole oluline, kõik parameetrid on samaväärsed nulliga;
$H_1 : \beta \neq 0$, mudel on oluline, vähemalt üks parameeter erineb nullist.

Mitme argumendiga mudeli korral on mudel statistiliselt oluline siis, kui mõni argumentidest on mudelis oluline. Mudeli olulisuse hindamisel on aluseks kogu hajuvuse (SST) jaotamine mudeli poolt kirjeldatud hajuvuseks (SSR) ja jääkhajuvuseks (SSE) ning nende abil defineeritud F-statistik $F = \frac{SSR/k}{SSE/(n-k-1)}$, mis on F jaotusega $F \sim F_{k,n-k-1}$ kui juhuslikud vead on normaaljaotusega. Kogu hajuvuse analüüs esitatakse dispersioonanalüüsi tabelis (*analysis of variance*).

Regressioonikordajate olulisuse kontroll

Üksikute regressioonikordajate β_i $(i = 0, 1, \ldots, k)$ kohta püstitatakse järgmised hüpoteesid:

$H_0 : \beta_i = 0$; i-s argument ei ole oluline, ta tuleb mudelist välja jätta;
$H_1 : \beta_i \neq 0$; i-s argument on oluline.

Hüpoteese kontrollitakse t-testiga analoogiliselt ühe argumendiga mudelile.

Mitteolulised argumendid tuleb mudelist välja jätta ja mudel uuesti leida!

4.2.4 Mudeli headuse näitajad

Mitmene korrelatsioonikordaja *R on defineeritud kui korrelatsioonikordaja uuritava tunnuse tegeliku väärtuse ja prognoosi vahel* $R = r(y, \hat{y})$. Mitmene korrelatsioonikordaja on alati positiivne ($0 \leq R \leq 1$) ning mitte väiksem kui korrelatsioonikordaja sõltuva tunnuse ja mingi argumenttunnuse vahel ($R \geq max|r(y, x_i)|$).

Determinatsioonikordaja R^2 (*R-squared*) *näitab kui suure osa sõltuva tunnuse hajuvusest kirjeldab mudel.* Determinatsioonikordaja on regressioonimudeli poolt kirjeldatud hajuvuse ja koguhajuvuse suhe $R^2 = SSR/SST$ ning arvestades koguhajuvuse lahutust $SST = SSR + SSE$ esitatakse determinatsioonikordaja ka tihti kujul $R^2 = 1 - SSE/SST$.

Mitmene korrelatsioonikordaja arvutatakse kui ruutjuur determinatsioonikordajast $R = \sqrt{R^2}$.

Arvestamaks mudelis esinevate argumenttunnuste arvu, arvutatakse tihti ka parandatud determinatsioonikordaja $\bar{R}^2$ (*adjusted R-squared*):

$$\bar{R}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}.$$

Hea mudeli korral $\overline{R}^2 \approx R^2$. Kui determinatsioonikordaja ja parandatud determinatsioonikordaja on väga erinevad, siis mudel ei ole hea, st väga väikese valimiga on püütud leida suurt mudelit.

Empiiriline reegel:
iga parameetri hindamiseks on vaja vähemalt 10 vaatlust $(n \approx 10p)$.

4.2.5 Vabaliikmega mudel *vs* vabaliikmeta mudel

Tavaliselt hinnatakse vabaliikmega mudel.

Vabaliikmeta mudeli vajadus peab tulenema mingitest teoreetilistest kaalutlustest, mis puudutavad antud valdkonda. Teadmine, et $y = 0$ kui $x = 0$, pole piisav põhjendus vabaliikmeta mudeli tegemiseks.

Põhiline probleem vabaliikmega ja vabaliikmeta mudeli korral on see, et *determinatsioonikordajad ei ole võrreldavad*, kuna nad arvutatakse nende mudelite korral erinevalt.

Definitsiooni järgi on determinatsioonikordaja mudeli poolt kirjeldatud hajuvuse suhe koguhajuvusse

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Arvestades, kuidas vastavad ruutude summad avalduvad, saame vabaliikmega mudeli korral (nim ka korrigeeritud determinatsioonikordaja, *corrected*)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{\boldsymbol{y}})^2}{\sum(y_i - \bar{\boldsymbol{y}})^2} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{\boldsymbol{y}})^2}.$$

Vabaliikmeta mudeli korral avalduvad ruutude summad aga teisiti (nim ka korrigeerimata determinatsioonikordaja, *uncorrected*), sest vabaliikmeta mudel ei läbi punkti $(\bar{x}, \bar{y})$

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum y_i^2}.$$

Seega saadakse alati erinevad tulemused ja vabaliikmeta mudeli determinatsioonikordaja on suurem.

Kui on vaja võrrelda vabaliikmega ja vabaliikmeta mudeli determintasioonikordajaid, tuleks arvutada determinatsioonikordaja teisiti. Selleks tasub tähele panna, et mõlema mudeli korral on mitmene korrelatsioonikordaja definitsiooni järgi $R = r(y, \hat{y})$. Seega saamaks vabaliikmeta mudelile determinatsioonikordajat, mis oleks võrreldav sama uuritava tunnuse mudeliga, kus on vabaliige, tuleks arvutada mitmene korrelatsioonikordaja ja see ruutu tõsta, sest mitmene korrelatsioonikordaja oli ruutjuur determinatsioonikordajast, $R = \sqrt{R^2}$, seega $R^2 = (R)^2$.

4.2.6 Mudeli jääkide analüüs ja mudeli diagnostika

Mudeli jääk (*residual*) on hinnang mudeli juhuslikule veale ja ta arvutatakse kui tegeliku väärtuse ning prognoosi vahe. Jääkide analüüsil on mitu eesmärki:

- kontrollida regressioonimudeli eelduste täidetust,
- kontrollida erindite olemasolu andmestikus,
- hinnata mudeli sobivust.

Mudeli diagnostika tegeleb mudeli erindite ehk mingis mõttes iseäralike vaatluste väljaselgitamisega.

Mudeli eelduste kontroll

Regressioonimudeli eelduste täidetust kontrollitakse analoogiliselt ühe argumendiga mudelile st vaadeldakse mitmesuguseid jääkide graafikuid hindamaks jääkide normaaljaotust ja konstantset hajuvust.

Eelduse kontrollimiseks on olemas ka testid.

Normaaljaotuse testimiseks tuntumad testid on Shapiro-Wilk'i test ja Kolmogorov-Smirnovi test (vt pt 1).

Jääkide konstantse hajuvuse hindamiseks kasutatakse

(a) White test, Breusch–Pagan test (1979). Leitakse R^{*2} mudelile $e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \ldots + \alpha_k x_{ki}$. Teststatistik: $n \cdot R^{*2} \sim \chi_k^2$

Otsus: *kui* $n \cdot R^{*2} < \chi^2_{k(tabelist)} \Rightarrow H_0$ *on tegemist konstantse hajuvusega*

(b) White üldine test (*HC, heteroscedasticity-consistent estimator*)
Testitakse mudeli üldisi eeldusi, sh jääkide hajuvuse konstantsust ja jääkide sõltumatust argumentidest.

SAS: proc REG, MODEL lauses valik SPEC

Jääkide sõltumatuse testimiseks on Durbin-Watsoni test (aegridade tüüpi andmed). Durbin-Watsoni statistik d on sõltumatute jääkide korral ligikaudu 2, sest $d = 2(1 - \rho)$, kus ρ – autokorrelatsioon

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}.$$

SAS: proc REG, MODEL lauses valik DWPROB

Näide. White testi ja Durbin-Watsoni testi kasutamine

On kontrollitud mudeli eeldusi ja saadud järgmised tulemused: White üldise testi tulemused (valikuga SPEC) ja Durbin-Watsoni testi tulemused (valikuga DWPROB)

```
      Test of First and Second
       Moment Specification
     DF    Chi-Square    Pr > ChiSq
     5          3.04        0.6942

Durbin-Watson D                1.850
Pr < DW                       0.0953
Pr > DW                       0.9047
Number of Observations           288
1st Order Autocorrelation      0.072

NOTE: Pr<DW is the p-value for testing positive autocorrelation,
 and Pr>DW is the p-value for testing negative autocorrelation.
```

Mida saab nende tulemuste järgi otsustada? Kas mudeli eeldused on täidetud?

Mudeli diagnostika

Mudeli **erindite** (*outlier*) väljaselgitamiseks on mitmeid võimalusi ja erindeid vaadatakse mitmest aspektist: erind suure jäägi mõttes, omapärane vaatlus ja mõjus vaatlus.

- **Erind suure jäägi mõttes** (*outlier*)

  Seda hinnatakse kasutades standardiseeritud jääke ja/või Studenti jääke (*standardized residuals, studentized residuals*). Standardiseeritud jäägi leidmisel jagatakse jäägi väärtus tema standardhälbega. Kui standardhälve on leitud i-ndat vaatlust mitte arvestades, saadakse Studenti jääk. Mõlemal juhul on saadud jääk ligikaudu t-jaotusega.
  Empiiriline kriteerium:
  kui vastav jääk ≈ 3 *või suurem, siis on tegemist erindiga.*

- **Omapärane vaatlus** (*leverage*)

  Omapära on defineeritud kui mütsi–maatriksi peadiagonaali element h_{ii} (*hat diag*). Mida kaugemal on selle vaatluse korral argumendi väärtus argumentide keskmisest tasemest, seda suurem on selle vaatluse omapära.
  Empiiriline kriteerium: *kui* $h_{ii} > \frac{2(k+1)}{n}$ *on tegemist erindiga.*

- **Mõjus vaatlus** (*influencial*)

  - **Mõju mudeli parameetritele**
    Tuntuim statistik selle hindamiseks on Cook'i D statistik (ka

Cook'i kaugus D), mis näitab erindi mõju kõigile regressioonikordajatele. Selle suur väärtus näitab, et vaatlus mõjutab regressioonikordajaid.
Empiiriline kriteerium: *kui* $D_i > f_{k,n-k}$ *siis* i*-s punkt mõjutab mudeli kordajaid (*$f_{k,n-k}$ *on* F*-jaotuse tabelist kriitiline väärtus).*

- **Mõju konkreetsele parameetrile**
  DFBETAS näitab erindi mõju regressioonikordajale. Selle suur väärtus näitab, et vaatlus mõjutab konkreetset regressioonikordajat. Näiteks kui mudelis on mingi j-s regressioonikordaja negatiivne, mida on raske interpreteerida, siis võib olla, et ka $DFBETAS_{ji}$ on suure negatiivse väärtusega ja see tähendab, et j-nda regressioonikordaja negatiivsus on tingitud i-nda vaatluse mõjust.
  Empiiriline kriteerium: *kui* $DFBETAS > \frac{2}{\sqrt{n}}$*, siis* i*-s vaatlus omab mõju* j*-ndale regressioonikordajale.*
- **Mõju prognoosile**
  Hindamiseks kasutatakse prognoositud ("kustutatud") jääke (*predicted residuals, prediction errors, deleted residuals*), mis leitakse kui vahe tegeliku väärtuse ja ilma i-nda vaatluseta prognoositud väärtuse vahel. Prognoositud jääk on enamasti suurem kui tavaline jääk. Prognoositud jääk võimaldab hinnata, kuidas mudel prognoosib sõltuvat tunnust ilma i-ndat vaatlust arvestamata: kui jääk on negatiivne, siis mudel ülehindab ja kui jääk on positiivne, siis mudel alahindab.

Märkus mudeli jääkide kohta

SAS protseduur REG arvutab jäägid järgmiselt:

Standardiseeritud jääk (*standardized*), saadakse jäägi jagamisel tema standardhälbega, kus σ on tundmatu

$$e_i^{stand} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

Studenti jaak (*studentized*), saadakse jäägi jagamisel tema standardhälbega, kus $\hat{\sigma}$ on hinnatud

$$e_i^{stand} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$$

Kui standardhälve on hinnatud samadel andmetel, nim neid jääke ka **sisemisteks studenti jääkideks** (*internally studentized*) (STUDENT)

$$e_i^{stand} = \frac{e_i}{\sqrt{\hat{\sigma}_i^2(1-h_{ii})}}$$

Kui standardhälbe hindamisel pole antud vaatlust kaasatud, nim neid jääke **välisteks studenti jääkideks**(*externally studentized*) (RSTUDENT)

4.2.7 Tunnuseteisendused

Mida teha, kui eeldused on rikutud?

Olukorda on võimalik parandada, kui kasutada tunnuseteisendusi.

1. Uuritava tunnuse teisendamine muudab vigade jaotust, seega kui vigade jaotus pole normaaljaotus, võib proovida uuritava tunnuse sobivat teisendust.
2. Kui seos uuritava tunnuse ja argumentide vahel pole lineaarne on võimalik seda teatud juhtudel lineariseerida kasutades argumendi teisendusi (näiteks kasutades astmefunktsiooni). Argumendi teisendused ei muuda vigade jaotust.
3. Kui mudeli korral on suuri probleeme nii vigade jaotusega kui ka lineaarsusega võib sobida hoopis mingi alternatiivne mudel (mittelineaarne või üldistatud mudel).
4. Teisenduste kasutamisel tuleb silmas pidada, et mudelid on võrreldavad originaalskaalal, st tuleb teha tagasiteisendus.

On olemas lihtsad reeglid vastavalt sellele, milline on jääkide jaotuse kuju.

(1) Jäägid parempoolse asümmeetriaga (saba paremale)

- Kerge saba paremale: $y \to \sqrt{y}$, $y \geq 0$.
- Mõõdukas saba paremale: $y \to \log(y)$, $y > 0$.
- Tugev saba paremale: $y \to 1/y$, $y \neq 0$.

(2) Jäägid vasakpoolse asümmeetriaga (saba vasakule)

- Kerge saba vasakule: $y \to y^2$.
- Tugev saba vasakule: $y \to \exp(y)$.

Üldine lähenemine on Box-Coxi teisenduste kasutamine. Ülaltoodud reeglid on selle teisenduste pere lihtsad erijuhud.

Kui ükski neist teisendustest ei vii sihile (tulemuseks pole normaaljaotusele lähedane jaotus), siis tulebki valida mingi teine jaotus.

4.2.8 Mudeli interpretatsioon

Ütleme, et ühikulise muutusega argumendis kaasneb regressioonikordaja suurune muutus uuritavas tunnuses muude tingimuste samaks jäädes

ehk

kui vaatame kahte indiviidi, kes erinevad ühe argumendi ühiku võrra ja teised argumendid on samad, siis uuritav tunnus erineb argumendi ees oleva kordaja võrra.

NB! Tuleb silmas pidada, et tegemist ei ole põhjuslikkusega!

Võrdleme rühmi kus *argumendi väärtus erineb ühiku võrra*

Eraldi võib huvi pakkuda mudel, kus on standardiseeritud kordajad, st mudeli tegemiseks on enne andmed standardiseeritud (iga tunnuse korral on tema väärtust teisendatud $(x - \bar{x})/s_x$). Standardiseeritud kordajat saab interpreteerida kui argumendi osakaalu.

Näide. Mudeli interpretatsioon

Maja hinna ($) mudel maja üldpindala (m^2), mugavuste arvu, vanuse (a) ja kauguse keskusest (km) järgi

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	98.913	44.656	2.21	0.032
mugavusi	1	35.38445	12.43879	2.84	0.0078
vanus	1	38.78396	18.85786	2.06	0.0482
kaugus	1	-35.55632	7.23797	-4.91	<.0001
pindala	1	5.11960	0.52756	9.70	<.0001

Milline on saadud mudeli kuju ja kuidas seda interpreteeritakse?

Mudeli kontroll (*verification*)

Mudeli verifitseerimine (*verification* – kinnitamine, ehtsuse kontroll) tuleks kindlasti teostada siis, kui meil on eesmärgiks saada prognoosiv mudel.

Algandmed jagatakse juhuslikkuse alusel 2 ossa:

- 80% – mudeli loomiseks,
- 20% – mudeli töökindluse ja prognoosivõime uurimiseks.

Võimalik läbi viia, kui valim on küllalt suur.

4.2.9 Sammregressioon

Mitme argumendiga regressioonimudeli konstrueerimisel on üsna suureks probleemiks argumentide valik mudelisse ja seda eriti siis, kui on võimalik valida suurest hulgast tunnuste hulgast. Kui meil on mõõdetud k tunnust, siis kõikvõimalike argumentide kombinatsioonidega mudelite üldarv on 2^k. Argumentide valiku automatiseerimiseks on loodud terve rida meetodeid, mis töötavad teatavate mudeli headust määravate kriteeriumite alusel. Selliseid meetodeid nimetatakse **sammregressiooniks** (*stepwise regression*). *Sammregressiooni korral leitakse parim mudel samm–haaval etteantud kriteeriumi alusel suurest tunnuste hulgast argumente valides* (kas lisades argumente juurde või jättes argumente välja).

Automaatsed mudeli valikute strateegiad jagatakse enamasti kolmeks:

- **Ettepoole** ehk kasvav valik (*forward*). Sel korral lähtutakse kõige lihtsamast – ainult vabaliiget sisaldavast mudelist – ning lisatakse argument, mille lisamisel F–statistiku suurenemine oleks maksimaalne ja oluline. Protsessi jätkatakse kuni ühegi argumendi lisamine enam statistikut oluliselt ei muuda. Kord mudelisse valitud argumenti välja ei jäeta.

- **Tahapoole** ehk kahanev valik (*backward*). See on eelmisele vastupidine protsess. Alustatakse täismudelist, mis sisaldab kõiki argumente, ning jäetakse igal sammul välja argument, mille väljajätmine muudab F–statistiku suuremaks. Protsessi jätkatakse kuni F–statistik muutub oluliseks ja ühegi argumendi väljajätmine teda ei paranda. Kord mudelist väljajäetud argumenti enam mudelisse tagasi ei panda.

- **Segavalik** (*stepwise*). Sel korral toimub mõlema eelmise strateegia kasutamine. Alustatakse lihtsast mudelist ja lisatakse argumente selliselt, et mudel muutuks paremaks. Igal sammul kontrollitakse, kas juba lisatud argumentidest mõne väljajätmine ei paranda mudelit. Selliselt toimitakse kuni protsess stabiliseerub — ühegi argumendi lisamine mudelisse ega ühegi mudelis juba oleva argumendi väljajätmine ei paranda mudelit.

Ükski eelnimetatud strateegiatest ei garanteeri parimat mudelit ning igaüks neist võib anda erineva mudeli.

Paljud autorid ei soovita kasutada automaatset argumentide valiku strateegiat ehk sammregressiooni.

"The data analyst knows more tan the computer" (Henderson & Vellman (1981). *Biometrics* 37, 391-411)

"It seems unwise, to let an automatic algorithm determine the questions we do or not do ask about data"(Judd & McClelland (1989). Data Analysis: A Model Comparison Approach, p.204).

Peatükk 5

Dispersioonanalüüs

5.1 Dispersioonanalüüsi mudel

Dispersioonanalüüsi kohta on soovitav lugeda õpikust *Statistilise andmetöötluse algõpetus*, lk. 258-327 (A.-M. Parring, M. Vähi, E. Käärik (1997), Tartu; olemas e-raamatuna TÜ raamatukogus).

Räägitakse ka ANOVA-st, mis tuleneb meetodi ingliskeelsest nimetusest *Analysis of Variance.*

Klassikalises dispersioonanalüüsis on

- uuritav (sõltuv) tunnus pidev arvtunnus,
- sõltumatud tunnused diskreetsed nn ***faktor**tunnused*.

Faktortunnus jaotab uuritava tunnuse klassideks (rühmadeks), faktortunnuse erinevaid väärtusi nim **tasemeteks** (*level*), eeldatakse, et tasemete arv ei ole väga suur. Hinnatakse faktori mõju uuritavale tunnusele, mis seisneb rühmade keskväärtuste võrdlemises.

Kahe tunnuse keskväärtuse võrdlemisel kasutatakse t-testi, aga kui k ($k > 2$) on keskväärtuse võrdlemisel kasutusel dispersioonanalüüsi meetodid.

Olgu meil k üldkogumit keskväärtustega $\mu_1, \ldots, \mu_k$ (ehk meil on faktortunnus, mis jagab üldkogumi k osaks). Siis mitme keskväärtuse võrdlemise ülesanne tähendab järgmise hüpoteesipaari kontrolli:

$H_0 : \mu_1 = \ldots = \mu_k$, keskväärtused on võrdsed;
$H_1 : \exists i, j \ \mu_i \neq \mu_j$ leidub erinevaid.

Üldkogumite võrdlemiseks peab olema igast üldkogumist valim, st igal faktori tasemel peab olema tehtud sõltuva tunnuse mõõtmisi. Olgu y_{ij} uuritava sõltuva tunnuse väärtus i-ndas üldkogumis j-ndal objektil, kusjuures $i = 1, \ldots, k$ ja $j = 1, \ldots, n_i$ (rühmade suurused ei pea olema üldjuhul võrdsed).

Mõõtmistulemused saab esitada mudeliga

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

kus ε_{ij} on juhuslik mõju. Mudel esitab erineva keskväärtusega mõõtmistulemused, kuid ei kajasta keskväärtuste erinevuste põhjuseid.

Faktortunnuse mõju uurimiseks esitatakse mudel kujul

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{5.1}$$

kus y_{ij} on uuritava tunnuse väärtus ($i = 1, \ldots, k$ ja $j = 1, \ldots, n_i$), μ tähistab üldkeskmist, α_i on faktori i-nda taseme poolt põhjustatud kõrvalekalle üldkeskmisest ($\alpha_i = \mu_i - \mu$). Ühesuse tagamiseks kitsendus

$$\sum_{i=1}^{k} \alpha_i = 0.$$

Hüpoteesid:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0; \quad H_1 : \exists i,\ \alpha_i \neq 0$$

Otsuse langetamise aluseks on valimite keskväärtused: $\overline{y}_{i\cdot}$ – i-ndal tasemel tehtud mõõtmiste keskmine ehk rühma keskmine ja $\overline{y}_{\cdot\cdot}$ – kõigi vaatluste keskmine ehk üldkeskmine.

Dispersioonanalüüsi olemus

Dispersioonanalüüsi nimetuses peitub meetodi aluseks olev idee – analüüsitakse keskväärtusi kasutades dispersiooni lahutust:

Koguhajuvus = faktori poolt kirjeldatud hajuvus + vea poolt kirjeldatud hajuvus .

Seega matemaatiliselt $SST = SSW + SSB$, kus SST – koguhajuvus (*total sum of squares*), SSW – rühmade sisene ehk vea poolt kirjeldatud hajuvus *within sum of squares*, SSB – rühmade vaheline ehk faktori (mudeli) poolt kirjeldatud hajuvus (*between sum of squares*). Seose näitamiseks lähtutakse vastavatest vahedest, mille ruutude summeerimisel saadaksegi vastav dispersiooni lahutus

$$y_{ij} - \bar{y}_{\cdot\cdot} = y_{ij} - \bar{y}_{i\cdot} \ + \ \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}.$$

Arvutatakse keskruudud (*Mean squares, MS*), st ruutude summad jagatakse nende vabadusastmete arvudega:

$$\frac{SSB}{k-1} \quad \text{ja} \quad \frac{SSW}{N-k}.$$

ning F-suhe, mis on H_0 kehtides (st faktoril pole mõju) F–jaotusega

$$F = \frac{SSB/(k-1)}{SSW/(N-k)} \sim F_{k-1,N-k}.$$

Kui hajuvus rühma sees on väiksem hajuvusest rühmade vahel, siis võetakse vastu H_1 ja st, et faktoril on mõju.

Näide 5.1

Piimakombinaadis võetakse saabunud piimast piimaproove uurimaks piima bakterisisaldust, piirkonda teenindas 5 piimaautot, seega saabus kombinaati iga päev 5 partiid piima, kuuel juhuslikul päeval tehtud analüüsitulemused näitasid järgmist (tabelis on bakterite arv piimas):

Päev	*1. partii*	*2. partii*	*3. partii*	*4. partii*	*5. partii*
1	24	14	11	7	19
2	15	7	9	7	24
3	21	12	7	4	19
4	27	17	13	7	15
5	33	14	12	12	10
6	23	16	18	18	20

NB! Andmetabel ei ole objekt-tunnus maatriks!

Sõltuvaks tunnuseks on bakterite arv.
Faktortunnuseks on partii, millel on 5 väärtust ehk taset. Igal tasemel on tehtud 6 mõõtmist, seega $n_i = 6;\ i = 1, 2, 3, 4, 5$; kokku on 30 mõõtmist.
Sõltuva tunnuse väärtused $y_{ij}, (i = 1, \ldots, 5;\ j = 1, \ldots, 6)$ on tabelis ja faktortunnuse väärtused tabeli päises. Faktortunnuse väärtused kodeeritakse.

Seega on objekt-tunnus maatriksis 2 veergu: tunnused 'bakter' ja 'partii' ning 30 rida.

Keskmine bakterite arv iga partii korral (keskväärtuste μ_i hinnangud) on vastavalt $\hat{\mu}_1 = 23.8$, $\hat{\mu}_2 = 13.3$, $\hat{\mu}_3 = 11.7$, $\hat{\mu}_4 = 9.2$, $\hat{\mu}_5 = 17.8$ ja üldkeskmise hinnang $\hat{\mu} = 15.2$.

Hindamaks, kas bakterite keskmine arv erinevate partiide korral on ühesugune, on kasutatud järgmist SASi programmi

```
proc GLM data=andmed;
class partii;
model bakter=partii;
run;
quit;
```

Programmi töö tulemusena saadakse järgmine hajuvuse analüüsi tabel

```
The GLM Procedure
Dependent Variable: bakter
                        Sum of
Source       DF       Squares    Mean Square  F Value  Pr > F
Model         4     803.000000   200.750000    9.01   0.0001
Error        25     557.166667    22.286667
Total        29    1360.166667
```

Näeme, et koguhajuvus (*Total*) on jagatud faktori poolt kirjeldatud hajuvuseks (*Model*) ja vea poolt kirjeldatud hajuvuseks (*Error*):
Total = *Model* + *Error*: $1360.166667 = 803 + 557.166667$.
Otsustamiseks kasutatakse F–statistikut ning väljastatakse talle vastav olulisuse tõenäosus. Siin $p = 0.0001$, mis annab alust väita, et keskväärtused on erinevad (loeme tõestatuks H_1). Seega keskmine bakterite arv erinevate partiide korral on erinev. $\diamondsuit$

5.2 Dispersioonanalüüsi mudelite liigitus

Dispersioonanalüüsi teostamise võimalused olenevad vaadeldavast mudelist ja kasutatavast katseplaanist. Kavandades katset keskväärtuste võrdlemiseks, tuleks esmalt kindlaks määrata huvipakkuvate *faktorite arv* ja nende *kombineerimise viis*. Hästi planeeritud katsega saame vajaliku informatsiooni minimaalse katsete arvuga. Halvasti planeeritud katsega ei pruugi saada midagi, võib juhtuda, et ei saa püstitada isegi meid huvitavat hüpoteesi. Katse kavandamisega tegeleb matemaatilise statistika valdkond — *katseplaneerimise teooria*, millel siinkohal ei peatu.

Dispersioonanalüüsi mudeleid võime liigitada:

1. Sõltuvalt **faktorite arvust**. Räägitakse 1–faktorilisest, 2–faktorilisest jne dispersioonanalüüsist (*one-way Anova*, *two-way Anova* jne).

2. Sõltuvalt faktorite tasemete kombineerimise viisist.
   Räägitakse **ristmudelist** (*cross-classification*), kus katsed on läbi viidud kõigil faktori tasemete kombinatsioonidel ehk kõik katsed on lubatavad.
   Kui ühe faktori iga tase on kombineeritud ainult ühe kindla teise faktori tasemega (üks allub teisele), siis on tegemist **hierarhilise mudeliga** (*nested*).

3. Sõltuvalt faktorite tüübist.
   Kui mudelis on esindatud kõikvõimalikud faktori tasemed (või tehakse järeldusi ainult nende kohta, mis on katsesse valitud), siis on tegemist fikseeritud faktoriga (*fixed factor*) ja see on **fikseeritud mõjudega**

mudel (*fixed effects model*).
Kui faktoril on väga palju tasemeid ja katsesse on võetud ainult osa neist (mingi juhusliku valiku alusel), aga järeldusi tahetakse teha kõigi tasemete kohta, siis on tegemist juhusliku faktoriga (*random factor*) ja see on **juhuslike mõjudega mudel** (*random effects model*).
Lõpuks kui mudelis on mõlemat tüüpi faktorid, siis räägime **segamudelist** (*mixed effects model*).

4. Sõltuvalt faktori erinevatel tasemetel tehtud mõõtmiste arvust võib mudel olla **tasakaalustatud** või **tasakaalustamata** (*balanced, unbalanced*). Tasakaalustatud mudeli korral on kõikides rühmades valimimahud võrdsed.

Kõige lihtsamad mudelid on tasakaalustatud fikseeritud mõjudega mudelid.

Seega mudel (5.1) kannab fikseeritud mõjudega 1-faktorilise dispersioonanalüüsi mudeli nimetust ja kui $n_1 = \ldots = n_k = n$, siis on tegemist tasakaalus mudeliga.

Dispersioonanalüüsi klassikalised eeldused

- sõltuv tunnus on faktortunnuse tasemetel normaaljaotusega (mudeli vead on normaaljaotusega),
- sõltuva tunnuse hajuvus on faktortunnuse tasemetel ühesugune,
- mõõtmised on sõltumatud.

Dispersioonanalüüsi eelduste kontroll

Dispersioonanalüüsi eelduste kontolliks on vaja analüüsida andmeid igal faktori tasemel (igas rühmas) eraldi. Kontrollitakse kas sõltuva tunnuse hajuvus on igal fakori tasemel ühesugune, normaaljaotust kontrollitakse mudeli vigade kohta tervikuna.

Normaaljaotuse eeldus ei ole väga range, mõõdukad kõrvalekalded mõjutavad järeldusi suhteliselt vähe ja seda eriti fikseeritud mõjudega mudeli korral. Juhuslike mõjudega mudeli juures tuleb normaaljaotuse eeldust rangemalt jälgida. Suur erinevus normaaljaotusest võib viia tõsiste vigadeni järelduste tegemisel.

NB! *Eelduste kontrolli juures tahame alati jääda nullhüpoteesi juurde, st meid huvitavad suured p väärtused!*

Dispersioonide võrdsuse kontrollimiseks on terve rida teste, neist paljud eeldavad normaaljaotust, seega tuleks jaotuse test enne teha. Dispersioonide võrdsuse asemel räägitakse ka **dispersioonide homogeensusest** (*homogeneity of variance*) ehk **homoskedastilisusest** (*homoscedasticity*), mille

vastandiks on heterogeensus ehk heteroskedastilisus. Siinkohal ei peatuks testidel, mis on toodud raamatus (*Parring, Vähi, Käärik* (1997), lk 285).

Paketis SAS on võimalikud järgmised dispersioonide homogeensuse kontrolli testid:

- BARTLETT (1937) – test eeldab normaaljaotust, põhineb normaaljaotuse tõepärasuhte statistikul.
- LEVENE (1960) – hajuvust mõõdetakse kui erinevust keskväärtusest, statistikuks on erinevuse absoluutväärtus või erinevuse ruut (vaikimisi). Test on kasutusel SAS/LABis.
- O'BRIEN (1979) – Levene testi modifikatsioon, kus kasutatakse kaalufunktsiooni.
- BROWN-FORSYTHE (BF) (1974) – Levene testi modifikatsioon, kus kasutatakse erinevust mediaanist. Soovitatakse kui rühmade arv on suur või kui hajuvus rühmades on väga erinev. SAS loeb seda testi parimaks.

Kolme viimase meetodi korral, kui dispersioonid osutuvad erinevateks, pakub SAS võimaluse kasutada *Welch'i kaalutud ANOVA*[2]-t.

Tihti aga on nende testide võimsus liiga väike, et hinnata, kas Welch'i kaalutud ANOVA sobiks. Samas on tavaline dispersioonanalüüs suhteliselt robustne, kui rühmade arv on suur ja hajuvused erinevad, aga rühmade suurused ligikaudu ühesugused.

Selle kohta ütlevad klassikud järgmist:

Box (1953) "*To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!*"

Kokkuvõte

- Tasakaalustatud fikseeritud mõjudega mudeli korral pole dispersioonide võrdsuse nõue range, kerge erinevus (näiteks kui saame p=0.07) ei mõjuta üldist otsust.
- Tasakaalustamata mudeli korral, kui üks dispersioon on teistest palju suurem, ei tohi seda ignoreerida.
- Juhuslike mõjudega mudeli korral mõjutab dispersioonide erinevus isegi tasakaalus mudelit.

[2]Welch (1951) esitas kaalutud dispersioonanalüüsi lahendusalgoritmi 1-faktorilisel juhul, mis on robustne dispersioonide võrdsuse eelduse suhtes.

Näite 5.1 järg
Kontrollime, kas bakterite arvu hajuvus on kõikide partiide korral ühesugune. Kasutame Brown-Forsythe testi.

```
proc glm data=andmed;
class partii;
model bakter=partii;
means partii/hovtest=bf;
run;
quit;
```

Saame tulemuseks

```
The GLM Procedure
   Brown and Forsythe's Test for Homogeneity of bakter Variance
    ANOVA of Absolute Deviations from Group Medians
                      Sum of         Mean
 Source         DF     Squares       Square    F Value    Pr > F
 partii          4     11.5333       2.8833       0.24    0.9114
 Error          25       297.2     11.8867
```

Seega saame testi tulemusena $p = 0.9$, järelikult pole meil põhjust ümber lükata väidet, et hajuvused on ühesugused. Loeme eelduse täidetuks.

5.3 Mitteparameetrilised testid

Mitteparameetrilised testid (*nonparametric tests*) on jaotusvabad ja leiavad kasutamist kui klassikalised eeldused pole täidetud. Mitteparameetrilised testid põhinevad tavaliselt astakutel. Tuntuim on Kruskal–Wallise test.

Testi kasutamisel on nõue, et uuritav tunnus peab olema pidev või omama küllalt palju erinevaid väärtusi, sõltumatud valimid. Teststatistik baseerub astakutel (*rank*).

NB! Tuleb tähele panna, et siinkohal räägitakse sõltumatute valimite võrdlemisest. Sõltuvate valimite korral (nt kordusmõõtmised) kasutatakse Friedmanni mitteparameetrilist testi.

Testid paketis SAS
Pakett SAS teostab mitteparameetrilise (1–faktorilise) dispersioonanalüüsi protseduuriga **NPAR1WAY**, soovitud test esitatakse valikuna protseduuri päises. Pakutakse 8 erinevat testi, mis põhinevad erinevatel skooride (*score*) definitsioonidel (Wilcoxoni, mediaani, Savage, Van der Waerdeni, Siegel-Tukey, Ansari-Bradley, Klotzi, Mood'i testid; täpsemalt vt *SAS Online Help*).

Kui ei ole mingit lisainformatsiooni, sobib kasutamiseks tuntud test – valik **Wilcoxon**, mis rohkem kui 2 taseme korral teostab klassikalise Kruskal–Wallise testi ning kahe taseme korral Mann–Whitney–Wilcoxoni testi.

Protseduuriga saab arvutada eelnimetatud testide jaoks ka täpsed p–väärtused st teha *täpset testi* (lause EXACT).

Mis on täpne test?
Täpne test ehk permutatsioonitest kujutab endast teststatistiku nullhüpoteesile vastava jaotuse leidmist arvutades teststatistiku väärtused andmete kõikvõimalike ümberpaigutuste korral. Täpset testi kasutatakse juhul kui asümptootilised eeldused ei kehti ja seega asümptootilised p-väärtused ei ole õigete p-väärtuste lähenditeks. Asümptootiline meetod eeldab tavaliselt, et teststatistikul on teatud jaotus kui valimimaht on küllalt suur. Kui aga valimimaht ei ole piisavalt suur, siis asümptootika ei kehti. Asümptootilised tulemused ei pruugi kehtida ka siis, kui andmete jaotus on tugevalt asümmeetriline või pikkade sabadega (Agresti (1996); Bishop, Fienberg, Holland (1975)). Täpsed arvutused põhinevad mitmemõõtmeliste sagedustabelite analüüsil (Agresti, 1992; algoritm Mehta & Patel, 1983) või Monte-Carlo simulatsioonimeetoditel (mis nõuavad arvutusteks vähem ressursse).
Täpsete testide isaks loetakse R.A. Fisher (1890–1962).

5.4 Keskmiste mitmene võrdlemine

Dispersioonanalüüsi tulemusena saame otsustada, kas keskväärtused on erinevad või mitte ehk kas faktoril on mõju. Kui meil on tegemist rohkem kui 2 rühmaga, siis jäävad meid veel huvitama küsimused – millised keskväärtused erinevad? kas kõik omavahel? või kui faktoril on mõju, siis kuidas ta mõjutab?

Kui faktor 3 taset, kontrollitakse 3 hüpoteesipaari leidmaks erinevusi keskmiste vahel

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 = \mu_3; \quad H_1 : \mu_1 \neq \mu_3$$

$$H_0 : \mu_2 = \mu_3; \quad H_1 : \mu_2 \neq \mu_3$$

k keskväärtuse korral saab teha $m = \frac{k(k-1)}{2}$ erinevat testi. Keskmiste mitmene võrdlemine (*multiple comparison*, *comparison of means*) annab vastuse küsimusele, millised keskväärtused erinevad ehk milliste faktori tasemete mõju on erinev.

Keskmiste võrdlemisel on võimalik 2 lähenemisviisi, millega on seotud erinevad vead.

- **Võrdlusviisiline** ehk üksikute järeldustega lähenemine (*comparisonswise*), öeldakse ka paariviisiline lähenemine. Eeldatakse, et **ainult 1** võrdluspaari korral on võimalik eksida (teha I liiki viga).

- **Katseviisiline** lähenemine (*experimentwise*). Eeldatakse, et **vähemalt 1** võrdluspaari korral on võimalik eksida (teha I liiki viga).

Võrdlusviisilise meetodi näiteks on **Fisheri LSD** (*Least Significant Difference*) **test**, mis on üks vanemaid teste (Fisher, 1949).

Kui nõutakse kõikide järelduste samaaegset kehtivust, on katseviisiline lähenemine ainuvõimalik. Tuntuim katseviisiline test on **Bonferroni test**. Itaalia matemaatik Bonferroni tõestas, et kui teeme m paariviisilist võrdlust, siis katseviisiline viga $p_m \leq m\alpha$, kus α on võrdlusviisilise vea tegemise tõenäosus. Seega kui tahame, et katseviisilise vea tõenäosus ei ületaks α, tuleks võrdlusviisilise vea tõenäosuseks võtta $\frac{\alpha}{m}$.

Näide
Olgu vaja teha 3 paariviisilist võrdlust ja tahame, et katseviisilise vea tõenäosus ei ületaks 0.05. Seega tuleks paariviisilised võrdlused teostada olulisusenivool $\alpha = \frac{0.05}{3} = 0.0167$. See garanteerib, et 3 võrdluse koos vaatamisel risk eksida ei ületa 5%.

$\diamondsuit$

Suhteliselt laialdast kasutamist on leidnud **Tukey–Krameri test**, mis algselt (Tukey, 1952) eeldas tasakaalustatud mudelit. Test põhineb haarde (*range*) jaotusel. Kramer (1956) täiendas testi tasakaalustamata mudeli jaoks, esitades keskmise valimi mahu seosega $\bar{n} = \frac{k}{\frac{1}{n_1}+\ldots+\frac{1}{n_k}}$, kus k on tasemete arv ja n_i valimimaht i-nda ($i = 1, \ldots, k$) taseme korral. Tukey-Krameri testi on omakorda täiendatud ja test kannab Tukey Studentized Range (HSD) (HSD – *Honestly Significant Difference*) testi nimetust, HSD test võtab arvesse võrdluspaaride arvu, seega loetakse katseviisiliseks.

Näite 5.1 järg

Oleme tõestanud, et keskmine bakterite arv erinevate piimapartiide korral on erinev. Küsimus on, millised partiid erinevad bakterite keskmise arvu poolest? Kasutame Tukey testi.

```
proc glm data=andmed;
class partii;
model bakter=partii;
means partii/tukey;
run;
quit;
```

Saame tulemuseks

```
The GLM Procedure
  Tukey's Studentized Range (HSD) Test for bakter
NOTE: This test controls the Type I experimentwise error rate,
but it generally has a higher Type II error rate than REGWQ.
   Alpha                                      0.05
   Error Degrees of Freedom                     25
   Error Mean Square                       22.28667
   Critical Value of Studentized Range      4.15337
  Minimum Significant Difference            8.0047

Means with the same letter are not significantly different.
 Tukey Grouping          Mean      N    partii
              A        23.833      6    1
              A
         B    A        17.833      6    5
         B
         B    C        13.333      6    2
         B    C
         B    C        11.667      6    3
              C
              C         9.167      6    4
```

See on protseduuri GLM tüüpiline väljund tasakaalustatud mudeli korral, kus esitatakse **homogeensed** rühmad. Tulemustest näeme, et keskmised moodustavad 3 homogeenset rühma: A – {1. ja 5. partii}, B – {5., 2. ja 3. partii}, C – {2., 3. ja 4. partii}.

Homogeensesse rühma kuuluvad keskmised loetakse võrdseteks, neid ei saa eristada.

Seega on siit võimalik järeldada, et keskmine bakterite arv on erinev **1. partii** (kõrgeim) ja **4. partii** (madalaim) korral. Teiste partiide korral erinevust tõestada ei saa.

♢

Teistest testidest erilisemas staatuses on **Scheffe test** (1953), mis võimaldab testida suvalist keskväärtuste lineaarkombinatsiooni ehk **kontrasti**.

Kontrastiks nimetatakse keskväärtuste lineaarkombinatsiooni

$$L = \sum_{i=1}^{k} c_i \mu_i,$$

kus c_i *on kordajad, mis rahuldavad tingimust* $\sum_i c_i = 0$.

Kontrasti kordajad c_i valitakse tuginedes ülesande sisulisele tähendusele.

Hüpoteesid kontrastide kohta: $H_0 : L = 0$; $H_1 : L \neq 0$

Kontrasti hinnang saadakse kasutades keskväärtusi: $\hat{L} = c_1\overline{y}_{1\cdot} + \ldots + c_k\overline{y}_{k\cdot}$

Paketis SAS saab kontraste kasutada protseduuriga GLM, kus tuleb MEANS lauses valik SCHEFFE ja kontrasti kontrollimiseks lisada lause CONTRAST 'nimi' *faktor* c_1 c_2 ... c_k; kus 'nimi' on suvaline tekst, mille järgi on lihtne leida väljundist kontrasti kohta käivat infot
NB! kontrasti kordajad peavad olema esitatud tasemete järjekorras ja eraldatud tühikutega!

Scheffe testi võib kasutada ka lihtsalt keskväärtuste võrdlemiseks, kuid seoses tema eripäraga on tema võimsus sel juhul suhteliselt madal, st ta võib mitte avastada erinevust keskväärtuste vahel.

Näide kontrastide kohta

Kliinilises katses uuriti ravimi mõju. Katses osales 3 rühma isikuid, kahes rühmas anti ravimit: 1. rühm ravim A ja 2. rühm ravim B, kolmas rühm oli kontrollrühm, kus anti ravimitaolist ainet platseebot.

Keskväärtuste võrdlemise korral võrreldakse kõiki neid omavahel, aga sisuliselt võib huvi pakkuda näiteks see, kas ravimid üldse mõjuvad või kas nad mõjuvad erinevalt.

Huvi võiksid pakkuda järgmised kontrastid:

$$L_1 = \mu_A + \mu_B - 2\mu_K$$

$$L_2 = \mu_A - \mu_B$$

Kas on tegemist kontrastidega? Mida need kontrastid näitavad? Mida tahetakse kontrollida?

Uusi mitmese võrdlemise teste

Lisaks eelnimetatud klassikalistele testidele on palju uusi teste. Nende hulgas on püüdnud süsteemi luua Hsu (1966), kes jagas meetodid kaheks:
(1) võrreldakse kõiki keskväärtusi omavahel;
(2) võrreldakse kontrollrühmaga.

Kontrollrühmaga (ehk ühe konkreetse keskväärtusega) võrdlemise testid on välja töötanud **Dunnet** (1955), on olemas tema kahepoolne test ja ühepoolsed testid. Ühe keskväärtusega võrdlemise testideks on ka **Hsu** testid (*Hsu's test for best, Hsu's test for worst*), kus toimub võrdlemine kas suurima või vähima keskväärtusega.

Kõigi keskväärtuste omavaheliste võrdluste tegemiseks on meetodite arv väga suur. Siinkohal võiks nimetada näiteks katseviisilist testi, mille autoriks on **Sidak** (1967). Tema test sarnaneb Bonferroni testile, kuid olulisusenivooks valitakse $1 - (1 - \alpha)^{1/m}$, kus m on kõikvõimalike võrdluspaaride arv.

Enamus meetodeid põhineb usaldusvahemike konstrueerimisel, aga on ka teistsuguseid lähenemisviise (järkstatistikute kasutamine: REGWF meetod[3], Student–Neyman–Keuls (1981); teatud kaitsenivoo defineerimine: Duncan (1955); Bayesi lähenemine: Waller–Duncan (1969)).

Soovitusi mitmese võrdlemise testide kasutamiseks:

- Kui paariviisiliste võrdluste arv on väike, sobib Bonferroni test.
- Suurte võrdlusandmete korral on sobiv Tukey-Krameri test (loetakse üheks võimsamaks).
- Kui huvi pakuvad kontrastid, siis kasutada Scheffe testi.
- Kui on kontrollrühmaga võrdlus, siis kasutada Dunneti teste.
- Kui on oluline ainult edasine uuringute suund, sobib Fisher LSD test (leiab vähima erinevuse).
- Mõnikord võib faktori tasemete vahel olla loomulik järjestus, tasemed järgnevad ajas või näiteks suurenevad doosi väärtused vms. Sel juhul pakub huvi **lineaarse trendi hindamine** (*test for linear trend*), st kontrollitakse, kas keskväärtused kasvavad (kahanevad) kui liikuda ühelt faktori tasemelt teisele.

Soovitused Fisheri ja Tukey testi kasutamiseks

Allikas: G.E. Dallal (2007). *Multiple Comparison Procedures.*

1. Tukey HSD test sobib lõplike otsuste tegemiseks.
2. Üldise lähenemine võiks olla Fisher'i LSD + Tukey HSD.
    - Erinevused, mis on olulised Tukey HSD testiga on kindlalt statistiliselt olulised
    - Erinevused, mis on mitteolulised Fisher'i LSD testiga on kindlalt statistiliselt mitteolulised
    - Erinevused, mis on olulised Fisher'i LSD testiga, aga on mitteolulised Tukey HSD testiga vajavad kindlasti edasist uurimist (uusi katseid)

[3] Ryan(1959), Einot & Gabriel (1975), Welsh (1977) F-test

5.5 Kuidas interpreteerida tulemusi?

Aluseks materjal võrgulehelt *www.graphpad.com*.

Praktiliste ülesannete lahendamisel ei piisa enamasti saadud p väärtusest ja otsusest, kas keskväärtused erinevad statitiliselt oluliselt või mitte. Tavaliselt soovitakse tulemust sisuliselt lahti seletada.

Suvalise keskväärtuste võrdlemise testi tulemusena saame p väärtuse, mille põhjal otsustame, kas
(a) tulemus on statistiliselt oluline $p < \alpha$;
(b) tulemus ei ole statistiliselt oluline $p > \alpha$.

Vaatame neid situatsioone lähemalt.

(a) tulemus on statistiliselt oluline ehk erinevus rühmade vahel pole juhuslik.
Tasub tähele panna, et *kui tulemus on statistiliselt oluline, ei tähenda see veel, et ta on teaduslikus plaanis tähtis.* Analüüsime keskväärtuste erinevuse usaldusvahemikku ja teeme otsused vastavalt selle suuruse sisulisele tähtsusele.
⇒ Kui keskmise erinevuse usaldusvahemiku mõlemad otspunktid ei ole teaduslikus plaanis huvipakkuvad (st erinevus on liiga väike), siis on saame öelda, et kuigi erinevus on statistiliselt oluline, on see väike ja ei paku teaduslikus plaanis huvi.
⇒ Kui keskmise erinevuse usaldusvahemiku üks otspunkt ei ole teaduslikus plaanis huvipakkuv, aga teine on (erinevus on piisavalt suur), siis olulisi järeldusi ei saa teha, sest pole täielikku kinnitust erinevuse tähtsuse kohta, tuleks teha täiendavaid katseid.
⇒ *Kui keskmise erinevuse usaldusvahemiku mõlemad otspunktid on teaduslikus plaanis huvipakkuvad (erinevus on piisav ka sisuliselt), siis võime öelda, et statistiliselt oluline erinevus on tähtis ka teaduslikus plaanis.*

(b) tulemus ei ole statistiliselt oluline, st ei saa teha järeldust, et üldkogumite keskväärtused erinevad, erinevus on juhuslik. Samas ei saa ka öelda, et tegelikud keskväärtused on võrdsed. Jällegi pakub huvi, milline võib olla tegelik erinevus ja selleks analüüsime keskväärtuste erinevuse usaldusvahemikku.
⇒ *Kui keskmise erinevuse usaldusvahemiku mõlemad otspunktid ei ole teaduslikus plaanis huvipakkuvad, siis saame väita, et keskväärtused on tõepoolest samad või on nende erinevus nii tühine, et see ei paku huvi.*
⇒ Kui keskmise erinevuse usaldusvahemiku üks otspunkt ei ole teaduslikus plaanis huvipakkuv, aga teine on (erinevus on piisavalt suur), siis tuleks teha uusi katseid, hetkel ei saa rangelt öelda, et erinevust pole.
⇒ Kui keskmise erinevuse usaldusvahemiku mõlemad otspunktid on teaduslikus plaanis huvipakkuvad, siis ei saa me teha mingit sisulist otsust. Tuleb katset korrata.

Näite 5.1 järg

Vaatame jällegi seda piimabakterite näidet.
Oletame, et sisuline tähtsus on bakterite keskmisel arvul alates 10-st ja analüüsime statistiliselt oluliseks tunnistatud erinevuse (1. ja 4. piimapartii vahel) usaldusvahemikku.

Valimi põhjal on keskmise erinevuse hinnanguks 14.7 ja 95%–usaldusvahemikuks (6.7, 22.7). Seega võib siin tegelik erinevus bakterite keskmise arvu vahel olla alates 6.7 (mis on liiga väike) kuni 22.7 (mis on piisavalt suur).

Saame teha järelduse, et kuigi statistiline olulisus oli tõestatud, ei pruugi sisuline erinevus tegelikkuses esineda.

Kindlama otsuse tegemiseks tuleks katseid korrata, mis on üsna mõistlik järeldus, kui arvestada, et iga partii korral oli tehtud vaid 6 mõõtmist.

5.6 Mitme faktoriga dispersioonanalüüsi mudel

5.6.1 2-faktoriline dispersioonanalüüs

Mudelis on kaks argumenti (2 faktorit), faktor A, tasemed $1, 2, \ldots, a$ ja faktor B, tasemed $1, 2, \ldots, b$. Olgu tegemist ristmudeliga st iga esimese faktori taseme korral on mõõdetud ka iga teise faktori tase. Vaatame lihtsamat juhtu, kui tegemist on tasakaalus mudeliga, st kõigi faktorite tasemete kombinatsioonide korral on tehtud võrdne arv mõõtmisi n. Seega mõõtmiste koguarv on $N = abn$.

Mudeliga hindame mõlema faktori mõju ja nende koosmõju, mudelil on järgmine kuju

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

kus $i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n$; μ on üldkeskmine,
α_i on faktori A peamõju (*main effect*),
β_i on faktori B peamõju,
$(\alpha\beta)_{ij} \doteq \gamma_{ij}$ on faktorite A ja B koosmõju (*interaction*),
ε_{ijk} on mudeli juhuslik viga.

Hinnangute ühesuse tagamiseks kitsendused (mõjude summad on võrdsed nulliga)

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} = 0.$$

Eeldused ja hüpoteesid

Mudeli eeldused:

1. Uuritav tunnus on mõlema faktori igal tasemel normaaljaotusega
ehk *juhuslikud vead on sõltumatud ja normaaljaotusega*

2. Uuritava tunnuse *hajuvus on mõlema faktori korral igal tasemel sama*

Hüpoteesid mõjude kohta:

1. Faktori A peamõju olulisus:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0; \quad H_1 : \exists i,\ \alpha_i \neq 0;$$

või $\quad H_0 : \mu_{A1} = \mu_{A2} = \ldots \mu_{Aa}; \quad H_1 : \exists i, j\ \mu_{Ai} \neq \mu_{Aj}$

2. Faktori B peamõju olulisus:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_b = 0; \quad H_1 : \exists j,\ \beta_j \neq 0;$$

või $\quad H_0 : \mu_{B1} = \mu_{B2} = \ldots \mu_{Bb}; \quad H_1 : \exists i, j\ \mu_{Bi} \neq \mu_{Bj}$

3. Faktorite koosmõju olulisus:

$$H_0 : \gamma_{ij} = 0; \quad H_1 : \exists i, j\ \gamma_{ij} \neq 0;$$

või H_0: *Ühe faktori keskmine muutus teise faktori tasemetel on samasugune*
ehk
st vahed on samasugused $H_0 : \mu_{A1B1} - \mu_{A1B2} = \mu_{A2B1} - \mu_{A2B2} = \ldots$

Näide. 2-faktoriline dispersioonanalüüs (näide 6.11 õpikust)
Tarkpea koolis õpetavad matemaatikat 2 õpetajat: T ja S, kes kasutavad kahte õpetamise metoodikat K ja R. Mõlemal õpetajal on 2 juhusliku valikuga moodustatud rühma õpilasi (7 õpilast rühmas) ja mõlemad kasutavad kahte erinevat meetodit. Kevadel tehti õpilastele test ja tulemused olid järgnevad:

	Õpetaja													
Meetod	T							S						
K	72	97	33	55	64	82	81	99	84	63	77	26	38	29
R	20	75	46	35	46	26	16	24	28	30	40	19	31	41

Sisulised küsimused:
(1) Kas sõltumata õpetamise meetodist on tase ühesugune?
(2) Kas erinevate õpetajate juures õppinutel on tase ühesugune?
(3) Kas õpetajatele sobivad mõlemad meetodid ühtviisi hästi?

2 faktorit: meetod ja õpetaja, mõlemal faktoril 2 taset

Statistilised hüpoteesid:

Peamõjude olulisus
(1) faktori *Meetod* peamõju $H_0 : \mu_K = \mu_R$
(2) faktori *õpetaja* peamõju $H_0 : \mu_T = \mu_S$

Koosmõju olulisus:
(3) H_0 : õpetajal T on metoodikate K ja R keskmiste hinnete erinevus sama kui õpetajal S

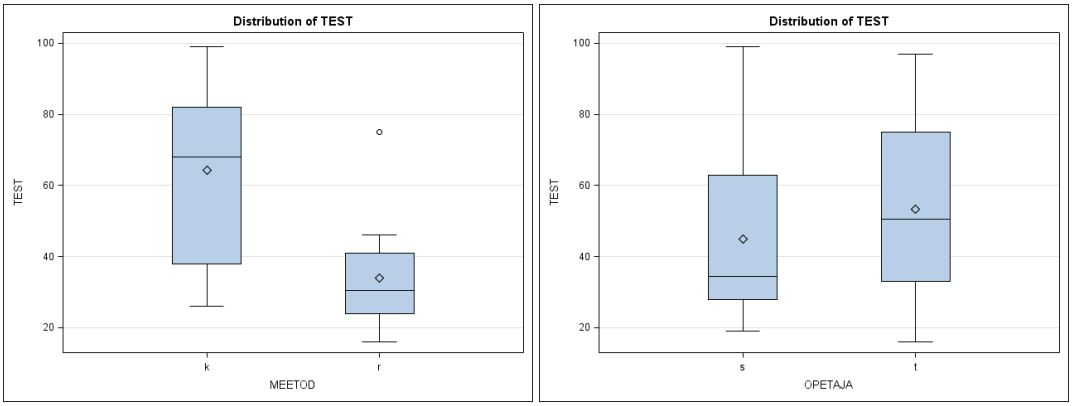
2-faktoriline dispersioonanalüüsi ülesande lahendus

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6906.39286	2302.13095	5.28	0.0061
Error	24	10455.71429	435.65476		
Total	27	17362.10714			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MEETOD	1	6390.321429	6390.321429	14.67	0.0008
OPETAJA	1	505.750000	505.750000	1.16	0.2920
MEETOD*OPETAJA	1	10.321429	10.321429	0.02	0.8790

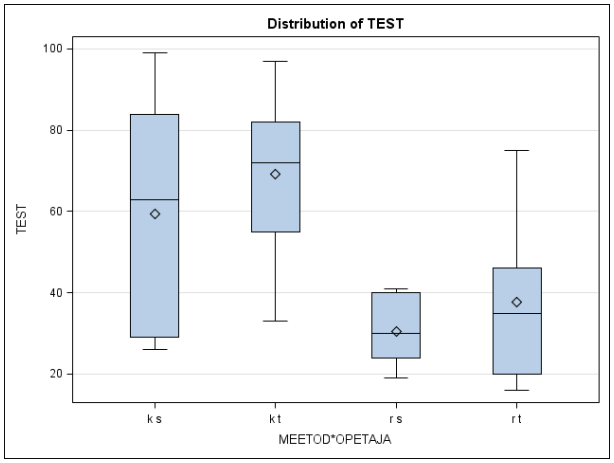
Milline on otsus?

2-faktorilise mudeli joonised, mille saame kui kasutame ODS graphics on/off proteduuri GLM juures.

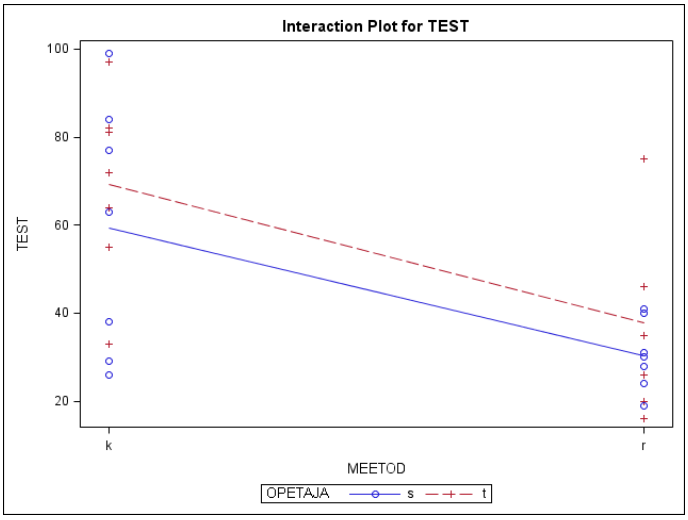


Peamõjusid iseloomustavad karpdiagrammid

Kõikvõimalike tasemete kombinatsioonidele vastavad karpdiagrammid (nn lihtmõjud):



2-faktorilise mudeli koosmõju joonis:



Paralleelsed sirged näitavad koosmõju puudumist !

Teades iga taseme keskmisi, saab viimase joonise ise visandada. Arvutatud keskmised antud näite korral on järgmised:

```
k    s    59.4
k    t    69.1
r    s    30.4
r    t    37.7
```

5.6.2 3-faktoriline dispersioonanalüüs

Kolmefaktoriline dispersioonanalüüsi korral on meil tegemist kolme argumendi ehk kolme faktoriga, olgu need tähistatud: faktor A (a taset), faktor B (b taset) ja faktor C (c taset).

Oluliseks erinevuseks 2-faktorilisest mudelist on *võimalike koosmõjude järgu suurenemine*.

3-faktorilise dispersioonanalüüsi korral on mudelis

- Peamõjud (hinnatavate parameetrite arv $a + b + c$)
- I järku koosmõjud (kahene koosmõju) ($ab + ac + bc$)
- II järku koosmõjud (kolmene koosmõju) (abc)

Hindamisel lähtutakse jällegi koguhajuvuse lahutusest.

5.6.3 Märkusi mitmefaktorilise dispersioonanalüüsi kohta

- Selgitada faktorite hierarhia (missuguse faktori tasemed olenevad mingi teise faktori tasemetest).
- Selgitada, missuguste faktorite mõju on juhuslik (kõik faktorite tasemed ei ole vaadeldud, sest neid on palju, on tehtud juhslik valik kõikvõimalike tasemete seast).
- Mõjude väärtusi on mõtet hinnata vaid fikseeritud mõjude korral.
- Selgitada kontrollitavad hüpoteesid.
- Kontrollida dispersioonanalüüsi eelduste täidetust.
- Enamasti ei lahenda ühe mudeli analüüsimine püstitatud ülesannet. Olla valmis vaatlema teisi mudeleid (erinevad koosmõjude järgud, vajadusel faktorite tasemete ühendamine jne).
- Parem on võtta mudelisse vähem kui liiga palju liikmeid, mudeli edasiarendamisel loobuda liigsetest ja teha uus mudel

5.7 Juhuslike mõjudega mudel

Juhuslike mõjudega mudel (*random effects model*) on põhjalikumalt kirjeldatud õpikus *'Statistilise andmetöötluse algõpetus'*, lk 309–320.

Juhuslik faktor on selline faktor, mille korral valitakse juhuslikult osa faktortunnuste tasemetest, aga järeldused tehakse kõigi tasemete kohta. Juhuslike mõjudega mudel on mudel juhusliku faktoriga.

Näide. Rongiliikluse uuring
Kas reisijate arv sõltub väljumisajast või on rongides ühepalju reisijaid? Vaatlused tehakse mõnedes rongides, aga otsust tahetakse teada kõigi rongide kohta. Seega keskmiste võrdlemisel pole mõtet (tahame teha järeldust ka nende rongide kohta, kus vaatlusi pole tehtud). Mõtet omab hinnata seda osa reisijate arvu hajuvusest, mis on põhjustatud erineva väljumisaja poolt.

Juhuslike mõjudega mudelis pakub huvi hinnata see osa uuritava tunnuse hajuvusest, mis on põhjustatud faktori poolt.

Mudel ja mudeli eeldused

Vaatame ühe juhusliku faktoriga mudelit, mõõdetud faktori k taset, igal n mõõtmist:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

kus μ on üldkeskmine, $\alpha_i,\ i = 1, \ldots, k$ on faktori i-nda taseme **juhuslik** mõju, ε_{ij} on mudeli juhuslik viga.

Mudeli eeldused:

1. Juhuslikud mõjud on sõltumatud ja normaaljaotusega

$$\alpha_i \sim N(0, \sigma_\alpha^2).$$

NB! Ei nõuta enam, et mõjude summa oleks null, mõjud pole fikseeritud.
2. Mudeli juhuslikud vead on sõltumatud ja normaaljaotusega $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Mudeliga seotud hüpoteesid

Lähtudes mudelist avaldub iga vaatlustulemuse dispersioon kujul

$$D(y_{ij}) = \sigma_\alpha^2 + \sigma^2,$$

kus $\sigma_\alpha^2,\ \sigma^2$ nimetatakse **dispersioonikomponentideks**.
Seega on vaatluste dispersioon jagatud kaheks: juhusliku faktori poolt põhjustatud hajuvus ja vea poolt põhjustatud hajuvus.

Kontrollitavad hüpoteesid: $H_0 : \sigma_\alpha^2 = 0; \quad H_1 : \sigma_\alpha^2 > 0$

SAS: Proc VARCOMP

Peatükk 6

Aeg mudelites. Kordusmõõtmised

6.1 Aja rollid mudelites

Aja erinevad võimalikud rollid mudelites[1]

- **Aeg kui funktsioontunnus**
  Funktsioontunnusena käsitletakse aega elukestusanalüüsis (*survival analysis*), kus uuritavaks suurusesks on aeg teatava sündmuseni (surmani, haigestumiseni, detaili riknemiseni jmt).
  Lihtsamal juhul võib aeg olla funktsioontunnuseks ka lineaarses regressioonimudelis, st funktsioontunnus on mõõdetud aja skaalal.
- **Aeg kui müratunnus**
  Aeg on müratunnuseks siis, kui vaatlused on kavandatud põhimõtteliselt ühel ajamomendil teostatutena, kuid tehnilistel põhjustel on vaatlusaeg veninud pikemaks. Selle vältimiseks tuleks aega argumenttunnuseks lugeda ja sõltuvust temast arvesse võtta.
- **Aeg kui argumenttunnus**

Aega käsitletakse argumenttunnusena aegridade mudelites ja kordusmõõtmiste mudelites.

1. **Aegridade** (*time series*) mudelites on aeg põhiline (tihti ainus) argumenttunnus. Andmestik erineb klassikalisest valimist, sest tüüpiliselt on igal ajahetkel olemas vaid üksainus mõõtmistulemus. Samas võib

[1] E.-M. Tiit (2001). Loengukonspekti põhjal refereeritud

mõõtmistulemuste hulk (mõõtmishetkede arv) olla küllaltki suur. Kui üheaegselt on mõõdetud mitut tunnust, siis on aegrida mitmemõõtmeline.

2. **Kordusmõõtmiste** (kestus- e longituudandmete; *repeated measures*, *longitudinal data*) korral on tavaliselt mõõdetud korduvalt rühma indiviide, kusjuures korduste arv ei ole väga suur. Lihtsamal juhul on mõõtmised tehtud kõigi objektide ja kõigi tunnuste puhul samadel ajahetkedel.

Aeg kui andmestiku dimensioon

Aega võib käsitleda ka kui andmestiku lisanduvat dimensiooni. Senikäsitletud andmestikud on valdavalt olnud kahemõõtmelised: ühe mõõtme määrab tunnuste loetelu, teise mõõdetavate objektide loetelu. Vastavalt sellele on ka kõik andmed identifitseeritavad kahe indeksiga - neist üks vastab tunnusele, teine objektile/ indiviidile. Siinjuures ei ole üldiselt kõneldes kummalgi dimensioonil järjestus oluline: valimis on kõik objektid samaväärsed. Ka tunnuste järjestusel puudub tavaliselt sisuline tähendus.

Vaatleme tavapärast andmestikku, mis on esitatud objekt-tunnus-tabelina, kus on n objekti ja k tunnust. Kui iga objekti iga tunnust on m korda mõõdetud, saame kolmemõõtmelise andmestiku, mida ei saa enam tavalise tabelina esitada, kuid mida võiks kujutada koosnevana m kihist, kus iga kiht on $n \times k$ tabel. Kuigi aeg lisab põhimõtteliselt ühe dimensiooni, ei tarvitse aega sisaldavad mudelid olla tingimata kõrge dimensiooniga.

Aegrida

Kõige tuntumad ajast sõltuvad andmestikud on aegread. Lihtsa ühemõõtmelise aegrea puhul vaadeldakse aega diskreetselt muutuvana, üht olulist tunnust on mõõdetud võrdsete ajavahemike järgi. Näiteks õhutemperatuuri mõõdetakse neli korda päevas, leibkondade tarbimisele antakse hinnang kord kvartalis jne. Ajahetki loetakse alates ühest (mõnikord ka nullist) ning ajamomenti näitab vaatluse indeks: $X_1, X_2, \ldots, X_n$. Aegrea mudelid on koostatud eeskätt selleks, et ajas kulgevat protsessi seletada ja prognoosida tema käitumist tulevikus. Aegrea üksikuid vaatlusi nimetatakse aegrea liikmeteks. Liikmete puhul eeldatakse, et nad koosnevad mitmest komponendist: $X_t = f_t + p_t + \varepsilon_t$, kus t tähistab ajahetke. Neist esimene f_t on fikseeritud (st mittejuhuslik), mida nimetatakse ka trendiks; teine p_t perioodiline või sesoonne, mis muutub ajas mingi fikseeritud (teadaoleva) perioodiga ja kolmas ε_t on juhuslik komponent.

Aegridade analüüsimisel kasutatakse mitmeid mudeleid, neist tuntuimad on

- **libiseva keskmise tüüpi mudelid**, mida kasutatakse eeskätt aegrea silumiseks (juhusliku komponendi elimineerimiseks), ning mille puhul iga aegrea punkt asendatakse tema lähiümbruses (ajas eelnevate ja järgnevate) punktide põhjal arvutatud (lihtsamal juhul keskmistatud) väärtustega;

- **autoregressiooni tüüpi mudelid**, mille korral eeldatakse, et aegrea käitumine on (suuremal või vähemal määral) määratud tema käitumisega lähiminevikus.

Lihtne aegrida on käsitletav ühemõõtmelise järjestatud andmestikuna, mille andmete hulk võrdub mõõdetud punktide arvuga. Seda tüüpi aegridade mudelid sõltuvad niihästi ajast kui ka aegrea olekust minevikus (olekutevahelisi sõltuvusi mõõdab autokorrelatsioonifunktsioon), kuid ei kasuta üldiselt täiendavaid argumenttunnuseid.

Mitmemõõtmeline aegrida

Kui ühe korraga on mõõdetud mitut tunnust (näiteks temperatuuri, õhurõhku, tuule kiirust ja suunda, sademete intensiivsust, õhu niiskust), siis on tegemist mitmemõõtmelise aegreaga. Selline andmestik koosneb m järjestatud punktist, millest igaühes on mõõdetud k tunnust. Erinevus võrreldes ühemõõtmelise aegreaga seisneb selles, et kõigi prognooside puhul on võimalik kasutada ka teiste aegridade andmeid. Aegridadevahelisi seoseid mõõdab vastastikune ehk *rist-korrelatsioonfunktsioon*.

6.2 Kordusmõõtmised

Selles valdkonnas on kasutusel erinevad terminid: korduvad mõõtmised, kordusmõõtmised (*repeated measures*) (lühemad uuringud) ja kestusandmed, longituudandmed (*longitudinal data*) (pikemad uuringud).

Kõikidel juhtudel on igal subjektil/objektil mingit tunnust mõõdetud rohkem kui üks kord. Aegrea puhul on tüüpiline see, et mõõdetakse ühtainust objekti. Kuid on ka ülesandeid, kus tuleb mõõta mitut objekti - näiteks jälgitakse n patsiendi seisundit, mõõtes igaühel k tunnust jälgimisperioodi vältel, so kokku m korda, sel juhul räägimegi kordusmõõtmistest.

Tavaliselt on tegemist ajas üksteisele järgnevate mõõtmistega. Teatavas mõttes toimub sellisel juhul indiviidide ökonoomsem kasutamine, igalt indiviidilt saame ühe tunnuse kohta terve rea väärtusi, samas aga tuleb arvesse võtta, et ühelt indiviidilt saadud tulemused on omavahel seotud. Täpsemalt, iga patsiendi andmestik koosneb kahest osast, osa on püsitunnused (näiteks patsiendi sugu, sotsiaalne seisund, kasv jne), teise osa moodustavad ajas muutuvad tunnused.

Tüüpiliselt esineb kordusmõõtmisi meditsiinis: ühte näitajat (tunnust) mõõdetakse teatud ajavahemike järel, uuritakse mingi ravi mõju vms. Näiteks antakse patsientidele ravimit ja mõõdetakse kehatemperatuuri, pulssi ja vererõhku samadel patsientidel iga päev nädala jooksul. Tegemist võib olla erinevate ravimeetodite võrdlemisega.

Sotsioloogias tehakse korduvaid ankeetküsitlusi samadele inimestele. Majanduse valdkonnas on kordusmõõtmistega tegemist näiteks kui vaatame panga klientide kontojääki iga kuu alguse seisuga aasta jooksul vms. Majanduses ja sotsioloogias nimetatakse kordusmõõtmistega seotud uuringuid ka paneeluuringuteks (*panel studies*).

Kordusmõõtmiste arvestamine analüüsis võimaldab eristada **subjektisisest** (*within subjects*) varieeruvust **subjektide vahelisest** (*between subjects*) varieeruvusest. Subjektide vahelist analüüsi nimetatakse ka **läbilõikeliseks** (*cross-sectional*).

Kordusmõõtmiste analüüsimisel kasutataksegi vastavalt probleemiseadele

- aegridade metoodikat;
- ristlõikelist metoodikat, mille puhul tähelepanu pööratakse ühel ajahetkel olemasolevale andmestikule;
- täielikke mudeleid, mille puhul kasutatakse kombineeritult mõlemat lähenemist, st mudelisse lülitatakse nii aeg kui ka ülejäänud faktorid.

Kordusmõõtmiste abil lahendatakse mitut tüüpi ülesandeid:

- Töötatakse välja mudel teatava funktsioontunnuse prognoosimiseks (või kirjeldamiseks) erinevatel ajahetkedel, kasutades nii püsi- kui ka muutuvate tunnuste väärtusi (regressioon- ja dispersioonanalüüsi ülesanded, kus arvestatakse oluliselt argumentide omavahelisi sõltuvusi).
- Analüüsitakse üksikindiviide, leitakse nende jaoks individuaalsed prognoosid (aegrea tüüpi ülesanded).
- Selgitatakse protsessi kulu erinevusi sõltuvalt püsi- ja muutuvtunnuste väärtustest (nn profiilide analüüs).

6.2.1 Puuduvad andmed

Kordusmõõtmiste korral võib mõõtmistes esineda puuduvaid väärtusi, mis raskendab analüüsi. Puuduvate andmete all mõeldakse siin enamasti katsest väljalangemist (*dropout*): uuritav subjekt langeb katsest välja, puudub mingist ajahetkest alates.

Puudumise esinemist mingil vahepealsel ajahetkel käsitletakse kui erineva pikkusega mõõtmisvahemikku (tasakaalustamata andmed) ja see tavaliselt ei leia eraldi tähelepanu.

Väljalangemine klassifitseeritakse analoogiliselt puudumiste struktuurile tavalise andmestiku korral (vt pt 1, lk 5) järgmiselt:

- Täiesti juhuslik väljalangemine (*CRD –completely random dropout*);
- Juhuslik väljalangemine (*RD – random dropout*);
- Mittejuhuslik (informatiivne) väljalangemine (*ID – informative dropout*).

Paljud analüüsimeetodid nõuavad täielikku andmestikku ja seega tuleks puuduvad väärtused asendada ehk imputeerida. Levinud asendusmeetod täiesti juhusliku ja juhusliku väljalangemise korral on viimase väärtuse kasutamine nn meetod *LOCF* (*Last Observation Carried Forward*), mis annab tegelikult nihkega hinnangud. See meetod on olnud kasutusel kliinilistes katsetes, kus on üsna loomulik, et jääb viimane näit. On olemas terve rida teisi ja paremaid lähenemisi nagu *Hot deck*, *direct likelihood*, kaalutud GEE jt (vt näiteks Verbeke, Molenberghs (2009). *Introduction to Longitudinal Data Analysis*).

6.2.2 Andmete visualiseerimine

Kordusmõõtmistega andmete korral on analüüsi esimeseks etapiks kindlasti andmete **visuaalne analüüs** – mitmesuguste graafikute joonistamine. Visuaalse analüüsi eesmärk:

- Näidata võimalikult suurt osa uuritavatest andmetest (kui võimalik, mitte kasutada kokkuvõtvaid statistikuid).
- Tuua esile huvipakkuvad tendentsid andmetes.
- Eristada läbilõikelisi ja ajas muutuvaid trende.

6.2.3 Katse planeerimisest ja korrelatsioonistruktuuridest

Kordusmõõtmiste kasutamine on aga tavaliselt seotud uuringu kallima hinnaga, sest igal indiviidil tehakse rohkem kui üks mõõtmine. Suuremat efektiivsust saavutame kordusmõõtmistega juhul, kui ajavahemikud mõõtmiste vahel on kõikide indiviidide jaoks võrdsed. Tihti pole see võimalik. Efektiivsus sõltub ka sellest, milline on indiviidide erinevate mõõtmiste vaheline *korrelatsioonimaatriksi struktuur*. Siin võivad esineda erinevad situatsioonid, neist lihtsaimad:

- esimene mõõtmine korreleerub viimasega sama tugevalt kui vahepealsetega st meil on tegemist võrdsete korrelatsioonidega (*uniform correlations, compound symmetry – CS*), $r_{ij} = \rho, \quad i, j = 1, \ldots, k$;
- mida rohkem aega möödub, seda nõrgemaks jääb korrelatsioon mõõtmiste vahel st meil on tegemist esimest järku autoregressiivse protsessiga (*first order autoregressive process – AR*), $r_{ij} = \rho^{|j-i|}, \quad i, j = 1, \ldots, k, i \neq j$.

Kui uuringu hind on antud, on uurijal võimalus valida, kas mõõta rohkem indiviide ja teha igal indiviidil vähem kordusmõõtmisi või mõõta vähem indiviide ja teha ühel indiviidil rohkem kordusmõõtmisi. Korrelatsioon korduvate mõõtmiste vahel mõjutab valimi mahu vajalikku suurust sõltuvalt meid huvitavast probleemist.

Kui tahame hinnata erinevate gruppide keskmisi, siis mida suurem on positiivne korrelatsioon mõõtmiste vahel, seda suurem on hajuvus ja seda suurema peame võtma valimi.

Kui aga hindame indiviidi muutusi ajas, siis samal indiviidil teostatud mõõtmiste vaheline positiivne korrelatsioon vähendab hinnangu hajuvust ja seega võib valimi maht olla väiksem (vt näiteks *Diggle, Liang, Zeger (1994). Analysis of Longitudinal Data. Oxford*).

6.2.4 Dispersioonanalüüs ja kordusmõõtmised

Teatavatel juhtudel on dispersioonanalüüs rakendatav kordusmõõtmiste korral, aga siin on teatavad piirangud. Põhiline probleem on see, et dispersioonanlüüs ei arvesta sõltuvust korduvate mõõtmiste vahel ja on lihtne vaid tasakaalus fikseeritud mõjudega mudeli korral. Dispersioonanalüüs on kasutatav lisaeeldusel, et mõõtmised toimuvad kõigil subjektidel samadel ajahetkedel. Tavaliselt pakub huvi uuritava tunnuse profiil (*mean response profile*) ehk keskmiste dünaamika ajas $\mu = (\mu_1, \mu_2 \ldots, \mu_T)$, μ_i – i-nda ajahetke keskmine.

Vaatame lähemalt mõningaid dispersioonanalüüsi kasutamise võimalusi.
Lihtsaim test kordusmõõtmiste korral on paariviisiline t-test, kus on 2 kordust.

1) Aeg-Aeg ANOVA (*Time by Time ANOVA*)

Selle lähenemise korral vaadatakse iga ajahetke eraldi ja tehakse igal ajahetkel tavaline dispersioonanalüüsi mudel.

Põhilised puudused sellise meetodi korral on:

- ei saa hinnata grupi/rühma mõju korduvate mõõtmiste suhtes,

- tehakse igal ajahetkel eraldi analüüsid, järeldused igast analüüsist pole kindlasti sõltumatud, aga pole ka selge, kuidas nad on seotud.

Selleks, et nendest puudustest vabaneda kasutatakse näiteks uuritava tunnuse eelnevaid ajamomente kui argumente mudelis (nn AD-mudelid – *ante-dependence*), mis arvestavad autokorrelatsiooni mõõtmiste vahel või defineeritakse uus tunnus, milleks on tavaliselt *uuritava tunnuse muutus*.

Näide. Aeg-Aeg ANOVA kasutamine

Uuringus on 60 vasikat, kellel on sooleparasiidid. Raviks kasutatakse 2 ravimit (A ja B). Vasikad jaotati juhuslikkuse alusel ravigruppidesse (mõlemas 30 vasikat). Mõõdeti vasikate kaal iga kahe nädala tagant (kokku 11 korda). Uuringu eesmärk on hinnata kuidas ravimid mõjutavad vasikate kaalu juurdekasvu.

Analüüs teostati kahel viisil. Kõigepealt viidi läbi t-testid, võrreldi kahe ravirühma keskmisi (Aeg-Aeg ANOVA) ja seejärel defineeriti uus tunnus – kaalu muutus (juurdekasv) ja võrreldi seda kahes rühmas. Analüüsi tulemusena t-testid erinevust ei näidanud (st vaadates iga ajahetke eraldi, ei saa rääkida kaalude erinevusest kahe ravi korral). Juurdekasvu analüüs näitas erinevust kahel ajahetkel: 8. ja 11. mõõtmisel on keskmised juurdekasvud gruppides oluliselt erinevad, kusjuures erinevused on vastasmärgilised. Ravimi A mõju on stabiilsem, mida on näha ka jooniselt.

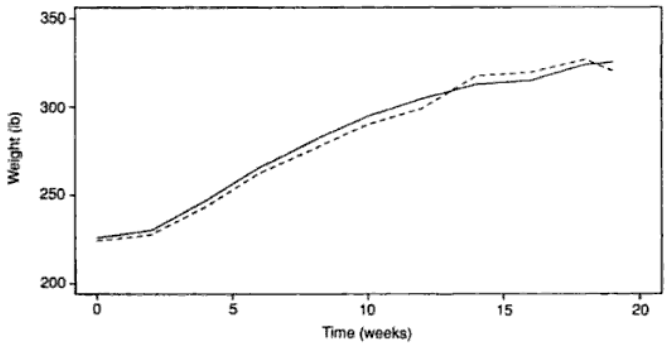


Fig. 6.1. Observed mean response profiles for data on weights of calves. ——: treatment A; – – –: treatment B.

Allikas: Diggle, Liang, Zeger (1994). Analysis of Longitudinal Data

2) Blokk-ANOVA kasutamine (1 valimi juht)

Ajas korduvat indiviidide rühma võime vaadelda ka kui blokki dispersioonanlüüsi mõttes. Lihtsaim juht on siis, kui meil on ainult üks rühm indiviide, keda on ajas korduvalt mõõdetud. Probleem taandub 2-faktorilise dispersioonanalüüsi mudeli hindamisele, sest meil on tegemist kahe faktoriga, mille

mõju peame hindama: indiviid kui faktor ja kordus kui faktor. Mõõtmistulemuse mudeli kuju on seega järgmine:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

kus μ on üldkeskmine, α_i on i-nda indiviidi juhuslik mõju, β_j on j-nda ajahetke mõju, ε_{ij} on mudeli juhuslik viga. Eeldatakse, et $\alpha_i \sim N(0, \nu^2)$ (ν^2 näitab subjektide vahelist varieeruvust) ja $\varepsilon_{ij} \sim N(0, \sigma^2)$ (σ^2 näitab subjektide sisest varieeruvust).

Lisaks on oluline eeldus, et mõõtmiste vahel on konstantsed korrelatsioonid (homogeenne korrelatsioonistruktuur). Korrelatsioonikordaja mõõtmiste vahel avaldub subjektide vahelise ja subjektide sisese varieeruvuse kaudu kujul

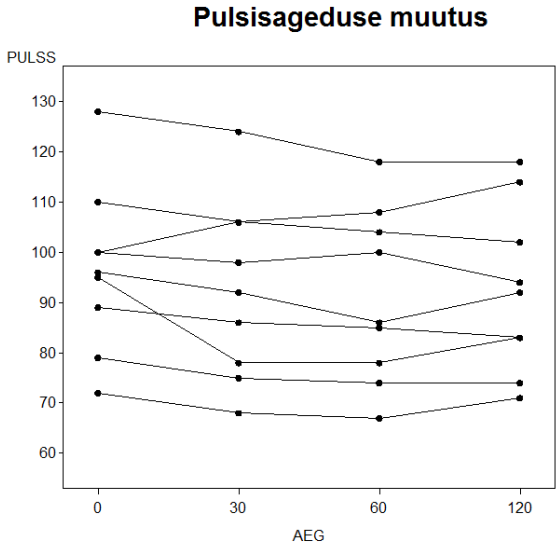
$$\rho = \frac{\nu^2}{\nu^2 + \sigma^2} \ .$$

Näide. Blokk-ANOVA kasutamine

Uuritakse uue ravimi mõju haigete pulsisagedusele. Patsientidele antakse ravimit ja mõõdetakse nende pulssi neli korda: kohe, poole tunni, tunni ja 2 tunni möödumisel. Hinnatakse ravimi mõju pulsisagedusele.
Andmetabel on järgmine:

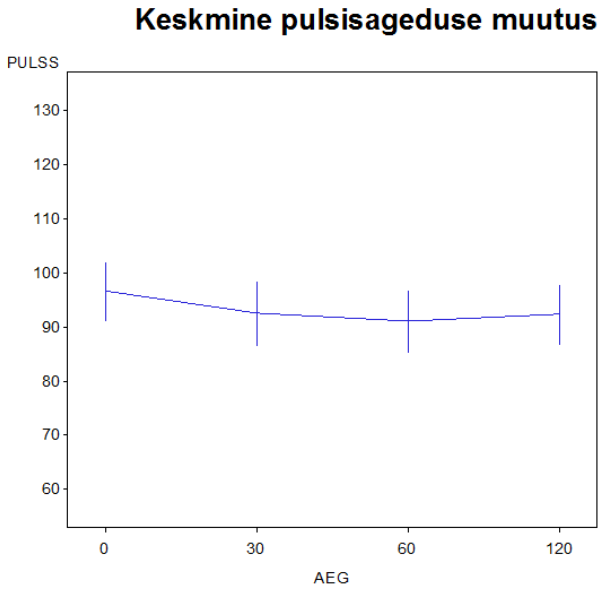
jrk	0	30	60	120	keskm
1	96	92	86	92	92
2	110	106	108	114	109
3	89	86	85	83	86
4	95	78	78	83	83
5	128	124	118	118	122
6	100	98	100	94	98
7	72	68	67	71	70
8	79	75	74	74	75
9	100	106	104	102	103
keskm	96	92	91	92	93

Andmed on kujutatud joonistel. Esimene joonis näitab, kuidas igal indiviidil pulsid aja muutuvad.



Joonis 1.

Teine joonis näitab keskmist pulsi muutust ajas.



Joonis 2.

Ülesande lahendab järgmine programm:

```
proc glm data=blokk;
  class indiv aeg;
  model pulss=indiv aeg; <-- indiv mõju eemaldatakse ja
                             siis hinnatakse aja mõju
  means aeg/tukey;
 run;   quit;
```

Programmi töö tulemusena väljastatakse järgmine informatsioon

```
Source  DF  Type I SS   Mean Square  F Value Pr > F <--NB!Type I

INDIV   8  8966.555556   1120.819444  90.64   <.0001
AEG     3   150.972222     50.324074   4.07   0.0180

 Means with the same letter are not significantly different.
       Tukey Grouping        Mean      N    AEG
                         A   96.556      9    0
                         A
                    B    A   92.556      9    30
                    B    A
                    B    A   92.333      9    120
                    B
                    B        91.111      9    60
```

Kuidas ravim mõjus?

3) Liigendatud ANOVA (mitme valimi juht). (*Split-Plot ANOVA*)

Oletame nüüd, et on tegemist mitme rühma indiviididega, keda on ajas korduvalt mõõdetud. Seega on meil indiviidi kui faktori mõju konkreetse rühma sees, ehk indiviidi mõju on allutatud rühmale (hierarhiline struktuur). Mõõtmistulemuste mudelil on järgmine kuju:

$$y_{ijk} = \mu + \alpha_{i(k)} + \beta_j + \gamma_k + (\beta\gamma)_{jk} + \varepsilon_{ijk},$$

kus $\alpha_{i(k)}$ on i-nda indiviidi juhuslik mõju k-ndas rühmas, $\alpha_{i(k)} \sim N(0, \nu^2)$, β_j on j-nda ajahetke (korduse) mõju, γ_k on k-nda rühma mõju, $(\beta\gamma)_{jk}$ on korduse ja rühma koosmõju, ε_{ijk} on k-nda rühma i-nda subjekti juhuslik viga ajahetkel j, $\varepsilon_{ijk} \sim N(0, \sigma^2)$. Nõutav on jällegi homogeenne korrelatsioonistruktuur.

Näide. Liigendatud ANOVA (vasikate näide, vt lk 78)

Andmed pööratud, mõõtmised on nüüd üksteise all, veerud: *id, aeg, grupp, kaal*. Kui kasutasime t-testi, olid mõõtmised üksteise kõrval.
Mudeli lahendamiseks programm kujul:

```
proc glm data=vasikad;
class id aeg grupp;
model kaal=id(grupp) aeg grupp grupp*aeg ;
run;
quit;
```

Ja selle töö tulemusena saadud väljund:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
id(grupp)	57	126268.3177	2215.2336	35.12	<.0001
aeg	10	839271.0415	83927.1042	1330.70	<.0001
grupp	1	107.4759	107.4759	1.70	0.1923
aeg*grupp	10	2298.9182	229.8918	3.65	<.0001

Analüüs näitab, et rühmakeskmised muutuvad ajas erinevalt (koosmõju on oluline).

Kokkuvõte: dispersioonanalüüs ja kordusmõõtmised

Kokkuvõtteks võib öelda, et teatud eelduste korral on dispersioonanalüüs kasutatav kordusmõõtmiste analüüsimiseks, tema põhiliseks plussiks on teostuse lihtsus ja tulemuste hea interpreteeritavus. Miinusteks on see, et dispersioonanalüüs (nii Blokk-Anova kui ka liigendatud Anova) on siiski kasutatav vaid suhteliselt kitsal juhul, nad esitavad korrelatsioonistruktuuri kohta rangeid nõudeid (homogeenne korrelatsioonistruktuur).

Seega ei ole dispersioonanalüüs üldine lähenemine kordusmõõtmiste modelleerimiseks.

6.2.5 MANOVA mudel

MANOVA (*Multivariate ANOVA*) mudel on ANOVA mudeli üldistus, kui ühe uuritava tunnuse asemel on mitu (uuritava tunnuse korduvad mõõtmised).

Tuletame meelde, et kordusmõõtmiste korral eristatakse (a) subjektisisest hajuvust (mõjutab aeg/kordus) ja (b) subjektide vahelist hajuvust (mõjutab rühm/ravi). Seega lihtsaimal juhul, kui on tegemist mitme rühma kordusmõõtmistega hinnatakse mudelis korduse mõju ja rühma mõju:

- Aja/korduse mõju, sisefaktor (*within-subjects effect*)
- Rühma/ravi mõju, välisfaktor (*between-subjects effect*)
- Aeg*Rühm koosmõju (korduse ja rühma koosmõju)

MANOVA mudeli korral teostatakse nii mitmemõõtmeline kui ka ühemõõtmeline (*univariate*) analüüs st esitatakse nii mitmemõõtmelised statistikud kui ka ühemõõtmelised (nõuavad lisaeeldusi).

MANOVA eeldused:

(1) Mitmemõõtmeline normaaljaotus, selleks tarvilik (aga mitte piisav) tingimus: iga mõõtmine peab olema normaaljaotusest.

(2) Homogeenne dispersioonide maatriks.

(3) Homogeenne dispersioonide-kovariatsioonide maatriks (sfäärilisus).

Mardia (1971) näitas, et MANOVA mudel on robustne kergete kõrvalekallete suhtes normaaljaotusest kui kõrvalekalded on põhjustatud asümmeetriast (mitte erinditest), aga on tundlik erindite suhtes.

Homogeense dispersioonide maatriksi nõue tähendab sisuliselt homogeense korrelatsioonistruktuuri nõuet (*CS – compound symmetry*). Mitmemõõtmelised statistikud pole väga tundlikud selle eelduse rikutuse suhtes, eriti tasakaalus mudeli korral.

Homogeenne dispersioonide-kovariatsioonide maatriksi nõue kannab nimetust ka sfäärilisus (*sphericity*), st nõutakse, et uuritava tunnuse kõikvõimalike tasemetepaaride erinevuse hajuvus oleks ühesugune:

$$D(y_{ij} - y_{ik}) = Dy_{ij} + Dy_{ik} - 2cov(y_{ij}, y_{ik}) \tag{6.1}$$

on konstantne $\forall j, k$.

Miks dispersioon nii avaldub?

Eeldus (3) on üldisem (nõrgem) kui eeldus (2), sest ühtlane korrelatsioonistruktuur on üks sfäärilisuse erijuht. Praktikas on raske hinnata, millal sfäärilisuse nõude kehtimisel ei ole tegemist ühtlase korrelatsioonistruktuuriga. Ühtlase korrelatsioonistruktuuri nõue tähendab, et $Dy_{ij} = \sigma^2 + \sigma_e^2$ on konstantne ja $cov(y_{ij}, y_{ik}) = \sigma^2$ on konstantne, nii et

$$corr(y_{ij}, y_{ik}) = \frac{\sigma^2}{\sigma^2 + \sigma_e^2} \text{ on konstantne.}$$

Mitmemõõtmeline test on eelistatud, kui n küllalt suur ($n > 20$) st mitmemõõtmeline analüüs nõuab suuri andmeid.

Sfäärilisuse kontroll

Kontrollitakse hüpoteesi H_0 : 'on sfäärilisus' ehk nõue (6.1) peab olema täidetud (eelduste kontroll, seega oleme huvitatud suurtest olulisusetõenäosuse p väärtustest). Kontrollimiseks kasutatakse Mauchly kriteeriumit (John W. Mauchly, 1940). Sfäärilisuse eeldus on vajalik selleks, et F-testid töötaksid korrektselt. Box (1954) näitas, et F-suhe hindamaks korduse mõju on positiivse nihkega, kui sfäärilisuse nõe ei kehti, st F-suhe on suurem kui peaks

olema ja nullhüpotees H_0 : 'korduse mõju ei ole' lükatakse ümber, kui peaks jääma selle juurde. Kui Mauchly kriteeriumi järgi ei ole sfäärilisus, siis tuleks korrigeerida vabadusastmeid, et valida õige F-suhe.

Korrigeerimiseks kasutatakse kahte parandust

- Greenhouse–Geisseri $\hat{\varepsilon}$ (G-G) (loetakse konservatiivseks);
- Huynch–Feldti $\tilde{\varepsilon}$ (H-F) (loetakse liberaalseks, ülehindab sfäärilisust).

Mõlemad parandused $(\hat{\varepsilon}, \tilde{\varepsilon})$ mõõdavad nö kovariatsioonimaatriksi kaugust sfäärilisusest ehk astet, kui palju on sfäärilisus rikutud. Mõlemad on defineeritud selliselt, et kui $\varepsilon = 1$, siis on tegemist sfäärilisusega, mida väiksem on ε, seda rohkem on sfäärilisus rikutud.

Märkusi sfäärilisuse kontrolli juurde

Sfäärilisust testitakse Mauchly testiga, aga selle kasutamisega on praktikas teatavaid probleeme.

- Test ei tööta hästi väikeste valimite korral (on madala võimsusega, ei avasta kõrvalekallet sfäärilisusest).
- Test kipub suurte valimite korral ülehindama kõrvalekaldeid, lükkab liiga tihti ümber H_0 (sfäärilisus).
- Test on tundlik normaaljaotusest kõrvalekalletele, eriti erinditest põhjustatud kõrvalekalletele.

Erinevad empiirilised kriteeriumid testide valikuks

- Kui $\varepsilon > 0.75$ ja valimi maht n on väike, siis kasutatakse ühemõõtmelist analüüsi H-F parandusega.
- Kui sfäärilisuse nõue on rikutud ($\varepsilon < 0.7$) ja valimi maht on piisavalt suur ($n > m+10$), siis on mitmemõõtmelised teststatistikud eelistatud.
- Üldiselt soovitatakse kasutada mitmemõõtmelisi teste, kui $n - r > m$, kus r on rühmade arv, m on korduste arv, n on valimi maht.

Mitmemõõtmelised statistikud

Mitmemõõtmelise analüüsi teostamisel kasutatakse hüpoteeside kontrollimisel mitmemõõtmelisi statistikuid, mis on ligikaudu F-jaotusega

1. Wilksi lambda
2. Pillai jälg (*trace*)
3. Hotelling-Lowley jälg
4. Roy suurim omaväärtus

Iga teststatistiku korral leitakse ligikaudne F-statistik, mille abil kontrollitakse hüpoteeese

Näiteks Wilksi lambda:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = det[\mathbf{W}(\mathbf{B}+\mathbf{W})^{-1}]$$

B – rühmadevahelised, **W** - rühmasisesed mõjud (maatriksid)

Märkused:

1. Paketis SAS saab mudeli hinnata protseduuriga GLM, kasutades lauseid MANOVA ja REPEATED.
2. MANOVA kasutamisel peavad andmestikus olema kordused üksteise kõrval (mitu uuritavat tunnust), ANOVA korral kordused üksteise all (1 uuritav tunnus)!

Kasvukõverate analüüs

Kordusmõõtmiste analüüsis võib huvi pakkuda rühmade võrdlus (nagu MANOVA mudeli korral), aga võivad huvi pakkuda ka ajas kulgevad trendid. Tavaliselt on kordusmõõtmiste korral tegu ajas järgnevate mõõtmistega, seega trendi jälgimine ajas on loomulik. Huvi pakub uuritava tunnuse profiil (*mean response profile*) ehk keskmiste dünaamika ajas

$$\mu = (\mu_1, \mu_2 \ldots, \mu_m),$$

kus μ_i on i-nda ajahetke keskmine, m ajahetke, $i = 1, \ldots, m$.

Kordusmõõtmiste trendi analüüsimine on kasvukõverate analüüs (*growth curve analysis*) ehk nn profiilanalüüs. Eesmärk on kirjeldada m ajahetke mõõtmisi $q < m$ parameetriga, mitte eraldi m erineva ajaspetsiifilise keskmisega.

Kolm põhilist hüpoteesi, mis pakuvad huvi profiilanalüüsis, on järgmised

- H_0^1 : keskmiste profiilid ajas on paralleelsed (st pole aja ja rühma koosmõju);
- H_0^2 : keskmiste profiilid ei erine rühmades;
- H_0^3 : keskmiste profiilid ei erine ajas.

Tuleb tähele panna, et hüpoteesi H_0^1 : tuleb kontrollida esimesena, sest selle tulemus mõjutab teisi (millisel kujul teisi testitakse).

Manova näide

Vaatame näidet ravimi mõju kohta (kordusmõõtmiste blokk-Anova andmed lk 79). Uuritakse uue ravimi mõju haigete pulsisagedusele. Patsientidele antakse ravimit ja mõõdetakse nende pulssi neli korda: kohe, poole tunni, tunni ja 2 tunni möödumisel. Hinnatakse ravimi mõju pulsisagedusele.

Tegemist on ühe rühmaga, kus on 9 patsienti, hinnatakse ainult aja mõju, andmetabel on seega 9 rida ja 4 veergu (veergudes tunnused: pulss0, pulss30, pulss60, pulss120).

Ülesande lahendab järgmine programm

```
proc glm data=repeat;
 model pulss0--pulss120= /nouni;
  manova;
  repeated aeg 4 contrast(1)/ summary;
run; quit;
```

Antud juhul on tegemist 1 rühma patsientidega, st pole välisfaktorit (rühm), seega MODEL lauses peale võrdusmärki on tühik, hinnatakse ainult aja (korduse) mõju. Lauses REPEATED määratakse sisefaktor (korduse faktor) (antud juhul antakse kordusele nimi *aeg*) ja korduste arv (antud juhul oli 4 mõõtmist). CONTRAST(1) teostab korduva faktori 'aeg' tasemete võrdluse 1. tasemega. Võimalik valik oleks ka PROFILE, sel juhul hinnatakse järjestikuseid vahesid. MODEL lauses valik *nouni* keelab ära ühemõõtmelise analüüsi (Aeg-aeg Anova mudelid).

Programmi töö tulemusena saadakse järgmine üldine informatsioon:

```
MANOVA Test Criteria and Exact F Statistics
for the Hypothesis of no aeg Effect
Statistic      Value        F Value Num DF Den DF   Pr > F
Wilks' Lambda    0.50692425   1.95    3        6    0.2237
Pillai' Trace    0.49307575   1.95    3        6    0.2237
```

```
Hotelling-Lawley 0.97268132   1.95    3        6    0.2237
Roy's Greatest   0.97268132   1.95    3        6    0.2237

Univariate Tests of Hypotheses for Within Subject Effects
Source   DF Type IIISS Mean Square F Value Pr > F  G - G   H - F

aeg        3  150.9722   50.3240   4.07    0.0180  0.0341 0.0193
Error(aeg)24   296.7777  12.3657
Greenhouse-Geisser Epsilon    0.7065
Huynh-Feldt Epsilon           0.9680
```

Valiku CONTRAST(1) kasutamisel saadakse tulemuseks:

```
Analysis of Variance of Contrast Variables
aeg_N represents the contrast between nth level of aeg and the 1st

Contrast Variable: aeg_2

Source  DF  Type III SS  Mean Square  F Value  Pr > F
Mean    1  144.0000000   144.0000000   4.20   0.0745
Error   8  274.0000000    34.2500000

Contrast Variable: aeg_3

Source  DF Type III SS   Mean Square  F Value  Pr > F
Mean    1  266.7777778   266.7777778   6.92   0.0301
Error   8  308.2222222    38.5277778

Contrast Variable: aeg_4

Source  DF  Type III SS  Mean Square  F Value  Pr > F
Mean    1  160.4444444   160.4444444   5.90   0.0413
Error   8  217.5555556    27.1944444
```

Kuidas ravim mõjus?

6.2.6 Segamudel

MANOVA mudeli korral eeldatakse mõõtmisi samadel hetkedel, puuduvad väärtused pole lubatud ja esitatakse ranged nõuded korrelatsioonistruktuurile (nõutakse ühtlast korrelatsioonistruktuuri). Seega ei ole ka MANOVA mudel sobiv üldiseks lähenemiseks kordusmõõtmiste analüüsimisel.

Korduvate mõõtmiste korral tegelik olukord vastab tavaliselt segamudelile. Segamudel (*Mixed model*) sisaldab fikseeritud mõjusid ja juhuslikke mõjusid. Fikseeritud mõju on tavaliselt rühma/grupi/ravi mõju ja juhuslik mõju on indiviidi mõju.

Segamudel võimaldab arvesse võtta erinevaid kovariatsioonistruktuure mudeli hindamisel. Mudeli üldkuju

$$Y = \mathbf{X}\beta + \mathbf{Z}\nu + \varepsilon,$$

$\mathbf{X}$ on fikseeritud mõjudega seotud plaanimaatriks, $\mathbf{Z}$ on juhuslike mõjudega seotud plaanimaatriks, β on fikseeritud mõjudega seotud tundmatute parameetrite vektor ja ν on juhuslike mõjudega seotud tundmatute parameetrite vektor.

Segamudeli korral eeldatakse $\nu \sim N(\mathbf{0}, \mathbf{G})$, $\varepsilon \sim N(\mathbf{0}, \mathbf{R})$ ja juhuslikud mõjud on sõltumatud vigadest $cov(\nu, \varepsilon) = 0$. Uuritava tunnuse dispersioon avaldub kujul $D\boldsymbol{y} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$ ja kui $\mathbf{Z} = \mathbf{0}$, siis saame tavalise lineaarse mudeli. Segamudelite korral kasutatakse erinevaid G struktuure (sh ka juba vaadatud võrdsete korrelatsioonide ja autoregressiivne korrelatsioonistrukruur).

Statistikapaketis SAS on segamudelite hindamiseks protseduur MIXED (andmed kujul, kus kordused on üksteise all, sest tegemist on ainult ühe uuritava tunnusega)!

Peatükk 7

Üldine lineaarne mudel

7.1 Klassikaline kovariatsioonanalüüsi mudel

Seni on läbi vaadatud 2 klassikalist mudelit:

- Regressioonanalüüsi mudel, kus argumendid on üldiselt pidevad arvtunnused.
- Dispersioonanalüüsi mudel, kus argumendid on diskreetsed.

Alati võib tekkida vajadus lisada esimesse mudelisse diskreetseid argumente või teise mudelisse pidevaid argumente. Klassikaliselt nimetatakse kovariatsioonanalüüsi mudeliks sellist dispersioonanalüüsi mudelit, kuhu on lisatud pidev argument. Sellest ka nimetus ANCOVA-mudel (*Analysis of Covariance*).

Lisatavat pidevat argumenti nimetatakse **kovariandiks** (*covariate*). Pidev argument ei tohi olla mõjutatud faktori tasemete poolt. Sisuliselt tahetakse hinnata faktori mõju kui kovariandil on mingi keskmine väärtus. Klassikaline kovariatsioonanalüüsi mudel ei sisalda kovariandi ja faktori koosmõju.

Näited
(1) Uuritakse ravimite mõju vererõhule. Sel juhul vererõhu muutus on uuritav funktsioontunnus, argumentideks on patsiendi vanus, kaal, sugu, samuti ravimdoos, ravi liik – näitajad, mis kõik võivad mõjutada haiguse kulgu. Vererõhu muutuse sõltuvuse uurimiseks saame mudeli, kus on nii pidevaid kui diskreetseid argumente.
Pidevad (kovariandid): vanus, kaal, ravimdoos ;
Diskreetsed (faktorid): sugu, ravi liik.
Igas ravirühmas peaksid olema keskmiselt sama vanusega inimesed (faktori mõju hinnatakse kovariandi keskmise väärtuse korral).

(2) Ettevõttes juurutatakse uusi tehnoloogilisi protsesse (vaatluse all on meetod A ja meetod B). Uuringusse on kaasatud 14 töölist, kummagi meetodiga töötab 7 töölist. Huvi pakub, kas üks meetoditest on kiirem kui teine.

Alustatakse faktori (meetod) mõju uurimisest: uuritav tunnus on aeg mingi toote valmimiseni ja faktor on toote valmistamise meetod (2 erinevat meetodit). Hinnatakse, kas ühe meetodiga valmib toode kiiremini kui teisega. Püstitatakse hüpoteesid $H_0 : \mu_A = \mu_B$, testimise tulemuseks on $p = 0.23$, seega ei saa rääkida, et üks meetod oleks kiirem kui teine.

Aga töölistel oli erinev töökogemus, mis võis tulemust mõjutada. Arvesse peaks võtma kovariandi (kogemus) mõju. Kovariandi lisandumisel on faktori mõju oluline ($p = 0.014$) ja kovariandi mõju oluline ($p < 0.0005$).

7.1.1 Dispersioonanalüüsi mudel pideva argumendiga

Lihtsuse mõttes eeldame, et meil on 1-faktoriline mudel

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

kus $i = 1, \ldots, k$ on rühmade arv (faktori tasemete arv), $j = 1, \ldots, n_i$ on indiviidide arv rühmas (mõõtmiste arv faktori tasemetel, ei nõua tasakaalustatud mudelit). Eeldame $\varepsilon \sim N(0, \sigma^2)$.

Lisame mudelisse kovariandi X, saame kovariatsioonanalüüsi mudeli

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij},$$

Eeldused: $\varepsilon \sim N(0, \sigma^2)$ ja β ei sõltu i-st, st kovariandi ja sõltuva tunnuse vaheline seos on ühesugune kõikidel faktori tasemetel. Seega saame tulemuseks sama tõusuga regressioonisirged (*homogeneity of regression*). Kas sirged on erinevad (paralleelsed) või langevad kokku, sõltub sellest, kas faktoril on mõju ehk kas α_i on mudelis oluline.

Näiteks, kui faktoril 2 taset ($i = 1, 2$), saame mudelid

$$y_{1j} = \mu + \alpha_1 + \beta x_{1j} + \varepsilon_{1j}; \quad y_{2j} = \mu + \alpha_2 + \beta x_{2j} + \varepsilon_{2j}$$

Sirged erinevad vabaliikme poolest (faktori mõju lisandub vabaliikmele!), vabaliikmed on vastavalt $\mu + \alpha_1$ ja $\mu + \alpha_2$, sirgete vaheline kaugus (y-telje sihis) on $\alpha_2 - \alpha_1$. $\diamondsuit$

Analoogiliselt saab vaadata üldisemat olukorda, kui meil on rohkem faktoreid ja rohkem kovariante. Interpretatsioon on aga keerulisem.

7.1.2 Regressioonanalüüsi mudel diskreetse argumendiga

Diskreetse argumendi kaasamiseks regressioonimudelisse võetakse kasutusele **indikaatortunnused** ehk **libatunnused** (*dummy variable*).

Indikaatortunnuseks u_i nimetatakse tunnust, mis määrab faktori taseme

$$u_i = \begin{cases} 1, & \text{kui faktoril on tase } i \\ 0, & \text{teistel juhtudel} \end{cases}$$

ja kehtib $\sum u_i = 1$.

Kui teame indikaatortunnuste väärtusi $(k-1)$ tasemel, siis on sellega määratud ka tase k, seega *kui faktoril on* k *taset, siis vajame* $(k-1)$ *indikaatortunnust.*

Näiteid

1. Kahe väärtusega tunnus 'sugu', kaks võimalikku indikaatormuutujat

$$u_1 = \begin{cases} 1, & \text{mees} \\ 0, & \text{naine} \end{cases}$$

$$u_2 = \begin{cases} 1, & \text{naine} \\ 0, & \text{mees} \end{cases}$$

Piisab kasutada ühte neist. $\diamondsuit$

2. Oletame, et vaatluse all on 3 erinevat võimalikku ravimeetodit. Seega on tegemist kolme väärtusega tunnusega 'ravi liik', millele vastab 3 indikaatortunnust

$$u_1 = \begin{cases} 1, & \text{1. ravimeetod} \\ 0, & \text{muu} \end{cases}$$

$$u_2 = \begin{cases} 1, & \text{2. ravimeetod} \\ 0, & \text{muu} \end{cases}$$

$$u_3 = \begin{cases} 1, & \text{3. ravimeetod} \\ 0, & \text{muu} \end{cases}$$

Piisab kasutada kahte neist, sest kui $u_1 = u_2 = 0$, siis $u_3 = 1$.

$\diamondsuit$

Kasutades indikaatormuutujaid saame dispersioonanalüüsi mudeli panna kirja kui regressioonanalüüsi mudeli.

Lihtsuse mõttes vaatame ühte faktorit, millel on 3 taset ja olgu tegemist tasakaalustamata mudeliga, kus $n_1 = 2, n_2 = 2, n_3 = 3$. Sel juhul on dispersioonanalüüsi mudelil kuju

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

kus $i = 1, 2, 3;\ j = 1, \ldots n_i$.

Võtame kasutusele 3 indikaatormuutujat u_1, u_2, u_3, saame mudelile kuju

$$y = \mu + \alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3 + \varepsilon,$$

Eelmisega ekvivalentne mudel (arvestades, et $u_3 = 1 - u_1 - u_2$) on järgmine

$$y = \tau + \gamma_1 u_1 + \gamma_2 u_2 + \varepsilon,$$

kus $\tau = \mu + \alpha_3$; $\gamma_1 = \alpha_1 - \alpha_3$; $\gamma_2 = \alpha_2 - \alpha_3$ st kõik faktori mõjud on arvestatud **viimase taseme suhtes** (SASis vaikimisi).

Lisades mudelisse kovariandi X, saame mudeli

$$y = \tau + \gamma_1 u_1 + \gamma_2 u_2 + \beta X + \varepsilon,$$

Kui faktoril mõju pole, siis $\gamma_1 = \gamma_2 = 0$ (puudub erinevus 1. ja 3. ning 2. ja 3. taseme vahel) ja saame regressioonisirge $y = \tau + \beta X + \varepsilon$ (st erinevust sirgete vahel pole, sirged langevad kokku viimasega).

Kui faktoril on mõju, siis $\gamma_1 \neq 0$ ja/või $\gamma_2 \neq 0$ ja saame faktori tasemetel järgmised mudelid (paralleelsed sirged, sama tõus, erinev vabaliige)

Faktori tase	Mudel
1	$\hat{y} = \hat{\tau} + \hat{\gamma}_1 + \hat{\beta} X$
2	$\hat{y} = \hat{\tau} + \hat{\gamma}_2 + \hat{\beta} X$
3	$\hat{y} = \hat{\tau} + \hat{\beta} X$

Lihtsa ANCOVA mudeli korral (1 pidev kovariant, 1 faktor k taset) saame tulemuseks k lihtsa regressiooni mudelit, mida hinnatakse faktori viimasel tasemel oleva mudeli suhtes.

Näiteid

1. Uuriti vere kolesterooli sisalduse (tunnus HDL) (mg/dl) sõltuvust kehakaalust ja füüsilisest aktiivsusest. Uuringus osales 26 indiviidi, kes olid jagunenud füüsilise aktiivsuse (tunnus 'grupp') järgi 3 gruppi (sulgudes taseme nimetus): kontrollrühm ('contr'), sportlased ('sport'), dieetrühm ('dieet'). Mõõtmine teostati 10 nädala möödudes katse algusest, kasutusele võeti indikaatorid u_1, u_2 ja analüüsi tulemusena saadi järgmine mudel

$$HDL = 20.4 + 9.2 * u_1 + 13.4 * u_2 + 0.2 * kaal$$

SAS hindab faktori viimase taseme suhtes ja järjestab tasemed tähestiku järgi. Seega toimub siin hindamine sportlaste rühma suhtes ja kasutusel on kaks indikaatortunnust

$$u_1 = \begin{cases} 1, & \text{grupp='contr'} \\ 0, & \text{muu} \end{cases}$$

$$u_2 = \begin{cases} 1, & \text{grupp='dieet'} \\ 0, & \text{muu} \end{cases}$$

Saame rühmades järgmised mudelid:

Kontrollrühma jaoks $HDL = 20.4 + 9.2 + 0.2 * kaal = 29.6 + 0.2 * kaal$.
Dieetrühma jaoks $HDL = 20.4 + 13.4 + 0.2 * kaal = 33.8 + 0.2 * kaal$.
Sportlaste jaoks $HDL = 20.4 + 0.2 * kaal$.

Lõpliku otsuse tegemiseks tuleks loomulikult uurida kõikide liikmete olulisust mudelites.

$\diamondsuit$

2. Leibkonnauuring. Uuritav on kogutarbimine (tarbimine pereliikme kohta kuus). Teada on sissetulek (pereliikme kohta kuus) ja laste arv peres (kodeeritud 0, 1, 2, 3 ja rohkem). Tegemist on kovariatsioonanalüüsi mudeliga, kus kovariant on sissetulek ja faktor on laste arv. Tulemuseks saadakse mudel kujul

$$\text{Kogutarbimine} = 541 + 51u_1 + 68u_2 + 55u_3 + 0.65 \text{ sissetulek}$$

kus on kasutusel indikaatorid $u_1 \rightarrow$(lapsi=0); $u_2 \rightarrow$ (lapsi=1); $u_3 \rightarrow$(lapsi=2); [$u_4 \rightarrow$(lapsi=3 ja rohkem) baastase]

Mudelid faktori iga taseme jaoks:

Lasteta pered: $(u_1 = 1,\ u_2 = 0, u_3 = 0)$
Kogutarbimine=541+51 +0.65 sissetulek = 592+0.65 sissetulek

1 lapsega pered: $(u_2 = 1,\ u_1 = 0, u_3 = 0)$
Kogutarbimine=541+68 +0.65 sissetulek = 609+0.65 sissetulek

2 lapsega pered $(u_3 = 1,\ u_1 = 0, u_2 = 0)$
Kogutarbimine=541+55 +0.65 sissetulek = 596+0.65 sissetulek

3 ja enam lapsega pered: $(u_1 = 0,\ u_2 = 0, u_3 = 0)$
Kogutarbimine=541+0.65 sissetulek

Milline joonis illustreeriks tulemust?

$\diamondsuit$

7.1.3 Kovariatsioonanalüüsi mudeli üldkuju

Olgu uuritav funktsioontunnus Y normaaljaotusega ja olgu antud kovariandid $X_1, \ldots, X_s$ ja faktorid $A_1, \ldots, A_r$, seega on vaatluse all $s + r$ argumenti. Tähistame faktori A_i tasemed vastavalt $a_{i1}, \ldots, a_{ik_i}$, kus $i = 1, \ldots, r$ ja k_i on i-nda faktori A_i tasemete arv. Defineerime indikaatorfunktsiooni $I(a_{ij})$ $(i = 1, \ldots, r;\ j = 1, \ldots, k_i)$ järgmiselt

$$I(a_{ij}) = \begin{cases} 1, & \text{kui faktoril } A_i \text{ on tase } a_{ij}, \\ 0, & \text{teistel juhtudel.} \end{cases}$$

Saame mudeli järgmisel kujul

$$EY = \beta_0 + \sum_{i=1}^{s} \beta_i X_i + \sum_{h=1}^{r} \sum_{j=1}^{k_h} \gamma_{hj} I(a_{hj}).$$

Mudelis on liidetav, mis koosneb kovariantide väärtuste lineaarkombinatsioonist ja liidetav, mis kajastab faktori tasemete mõju. Mudeli parameetrite hindamine toimub vähimruutude meetodil, hinnatavate parameetrite arv on kovariantide ja faktorite mõjude arvude summa $p = s + \sum k_i + r + 1$.

Plaanimaatriksis on üks ühtede veerg, mis vastab vabaliikmele, s veergu, mis vastavad kovariantidele ja $r \cdot k_r$ veergu, mis sisaldavad 1 ja 0 ning mis vastavad faktorite tasemete indikaatorfunktsioonidele.

7.1.4 Indikaatortunnuste kasutamisest

Kasutades indikaatortunnuseid saame mudelisse lülitada mistahes tüüpi tunnuseid (ka nominaalseid). Siin peituvad aga teatud **ohud**:

- indikaatortunnustega mudelid võivad muutuda otsitavate parameetrite arvu poolest liiga suureks;
- tasakaalustatud mudeli korral võib probleem olla korrektselt lahenduv, tasakaalustamata mudeli korral võib tekkida probleeme;
- lünklikud andmed muudavad tasakaalustatuna planeeritud katse ikkagi tasakaalustamata juhuks ja liiga suure indikaatortunnuste arvu korral ei pruugi leida lahendit.

Seega tuleks nominaaltunnuseid (faktoreid) ja nende koosmõjusid kaasata mudelisse mõistlikkuse piirides.

7.2 Üldine lineaarne mudel

On läbi vaadatud 3 klassikalist mudelit:

- Regressioonanalüüsi mudel, kus argumendid on pidevad arvtunnused.
- Dispersioonanalüüsi mudel, kus argumendid on diskreetsed (kvalitatiivsed).
- Kovariatsioonanalüüsi mudel, kus on nii diskreetseid kui ka pidevaid argumente, aga puuduvad koosmõjud diskreetse ja pideva argumendi vahel.

Vaatame lineaarset mudelit maatrikskujul

$$\boldsymbol{y} = \mathbf{X}\beta + \varepsilon, \tag{7.1}$$

kus $\boldsymbol{y}$ on $n \times 1$ uuritava tunnuse vektor, $\mathbf{X}$ on $n \times p$ plaanimaatriks, $p = k+1$, β on $k \times 1$ tundmatute parameetrite vektor ja ε on $n \times 1$ juhuslike vigade vektor.

Mudel (7.1) kannab (üldise) lineaarse mudeli nimetust, mis erijuhtudel (sõltuvalt argumentide tüübist) on kas (1) regressioonanalüüsi (2) dispersioonanalüüsi või (3) kovariatsioonanalüüsi mudel.

Üldiselt sisaldab lineaarne mudel nii pidevaid argumente, diskreetseid argumente kui ka mitmesuguseid koosmõjusid. Seega võib plaanimaatriks sisaldada nii pidevaid kui ka diskreetseid argumente ja/või indikaatortunnuseid. Mudel on lineaarne parameetrite suhtes. Argumentide kohta pole mingeid kitsendusi.

Tundmatud parameetrid leitakse vähimruutude meetodil normaalvõrrandisüsteemi $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\boldsymbol{y}$ lahenditena.

Lineaarse mudeli eeldused

- Lineaarne seos uuritava tunnuse ja argumentide vahel
- Vigade sõltumatus: $cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
- Vigade normaaljaotus: $\varepsilon_i \sim N(0, \sigma^2),\ \varepsilon \sim N(0, \sigma^2\mathbf{I})$
- Vigade konstantne hajuvus: $D\varepsilon_i = \sigma^2,\ D(\varepsilon) = \sigma^2\mathbf{I}$

Inglise keeles kasutatakse eeldustest rääkides järgmist skeemi, mis aitab eeldusi hästi meeles pidada:

L – Linear relationship
I – Independent observations
N – Normally distributed
E – Equal variance

Eelduste täidetuse korral on uuritav tunnus normaaljaotusega $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

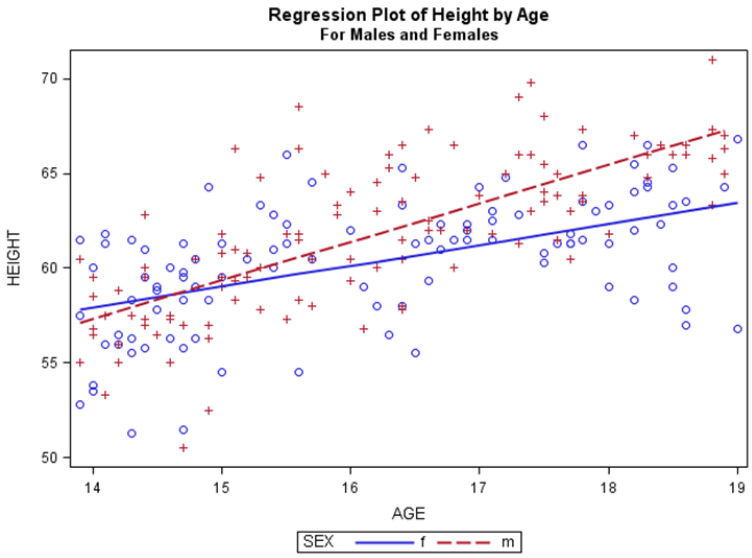
Kus võib tulla probleeme?

1. Lineaarsuse eeldus võib olla rikutud. Sel juhul tuleks mudelisse lisada kõrgemat järku liikmeid.

2. Eeldused jääkide kohta võivad olla rikutud. Andmetes võivad esineda erindid, mis võivad põhjustada mittehomogeenset hajuvust ja/või kõrvalekaldeid normaaljaotusest.

Näide.
Mõõdetud 14–25 aastaseid noori, $n = 237$. Huvi pakub kasvu mudel vanuse ja soo järgi (kasv mõõdetud tollides)

$$\text{kasv} = \beta_0 + \beta_1\ \text{sugu} + \beta_2\ \text{vanus} + \beta_3\text{sugu*vanus}$$



Vanuses alla 15 a on poisid väiksemad, nende kasv on kiirem.

Näite lahendus.

```
Parameter Estimates
                                Parameter    Standard
Variable            DF           Estimate       Error    t Value    Pr > |t|

Intercept            1           28.88281     2.87343      10.05      <.0001
sugu=n               1           13.61231     4.01916       3.39      0.0008
vanus                1            2.03130     0.17764      11.44      <.0001
vanus*(sugu=n)       1           -0.92943     0.24782      -3.75      0.0002
```

Koosmõju olulisus näitab, et vanuse mõju kasvule on tüdrukutel ja poistel erinev.

Kirjutame välja mudelite kujud.
Üldkuju: $kasv = 28.88 + 13.61(sugu = n) + 2.03vanus - 0.93vanus * (sugu = n)$

Neidude mudel: $kasv = 28.88 + 13.61 + 2.03vanus - 0.93vanus$
$= 42.5 + 1.1vanus$

Noormeeste mudel: $kasv = 28.88 + 2.03vanus$

Kokkuvõte lineaarsetest mudelitest

Vaatasime mudeleid, kus argumentideks olid kovariandid ja faktorid.

I Mudel (ANCOVA): Kovariant (X) + faktor (A)

- Faktor mõjutab vabaliiget
  tulemuseks kas paralleelsed või kokkulangevad sirged

- Mudelid faktori tasemetel (rühmades) langevad kokku või erinevad vabaliikme poolest

Miks panna faktor mudelisse? Võiks ju igal tasemel eraldi mudeli teha, tulemus on sama?

Me ei saa võrrelda üksikuid faktori tasemetel tehtud mudeleid. Lülitades faktori mudelisse, saame hinnata faktori tasemete erinevust

II Mudel (üldine lineaarne mudel):
Kovariant (X) + faktor (A) + koosmõju ($A * X$)

- Koosmõju mõjutab kovariandi kordajat
  tulemuseks kas lõikuvad või paralleelsed (kokkulangevad) sirged
- Faktori tasemetel (rühmades) võib lisaks ka kovariandil olla erinev mõju

7.3 Polünomiaalne regressioonimudel

Lineaarsete mudelite hulka võib lugeda ka polünomiaalse regressioonimudeli, mis on samuti parameetrite suhtes lineaarne.

Tavaliselt on tegemist ainult ühe argumendiga ja mudelis on ka selle *astmed*.

Üldiselt räägime $k-$**järku polünomiaalsest mudelist**, kui mudelil on kuju

$$y = \beta_0 + \beta_1 x + \ldots + \beta_k x^k + \varepsilon,$$

kus $k > 0$ mittenegatiivne täisarvuline konstant.

Parameetrid β_i leitakse vähimruutude meetodil, minimiseerides hälvete ruutude summad.

Näiteks, kui on tegemist teist järku polünomiaalse mudeliga ehk ruutmudeliga

$$y = \beta_0 + \beta_1 + \beta_2^2 + \varepsilon,$$

siis hinnatakse hälvete ruutude summat paraboolist

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2.$$

Saadud võrrandid võivad tulla suhteliselt keerulised.

Kui parameetrid on hinnatud, siis tuleb otsustada:

- Kas saadud mudel on oluline? Kas uuritava tunnuse hajuvus on paremini kirjeldatud, kui kasutame kõrgemat järku liikmeid? Mudeli olulisuse hindamiseks kasutatakse $F-$statistikut.

- Kas kõrgemat järku mudeli kasutamine on statistiliselt oluline võrreldes lineaarse mudeliga? Testitakse kõrgemat järku liikmete kordajate olulisust. Kasutatakse $F-$statistikut.
- Mudeli adekvaatsus ehk vastavus andmetele. Kui teist järku mudel on parem kui lineaarne, kas peaksime veel lisama kuupliikme jne? Kuidas saame kindlad olla, et ei vaja kõrgemat järku mudelit? Teostatakse nn *Lack of Fit* test – sobimatuse test.

Polünomiaalsete mudelite kasutamine sõltub probleemist ja andmetest

Uuringute korral, kus eeldatakse, et seos uuritava ja argumentide vahel on monotoonse iseloomuga, ei pruugi astmefunktsioon sobida

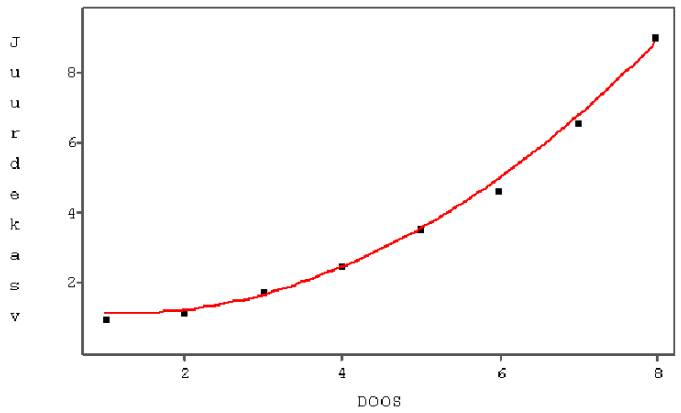
Näide (teist järku mudeli kohta).

Uuritakse kaalu juurdekasvu vastavalt mingile doosile. Kaheksa laborilooma on sama soo, vanuse ja suurusega ning on jagatud juhuslikult nii, et igaüks saab ühe kaheksast doosist. Kaalu juurdekasvu mõõdetakse pärast kahe nädala möödumist.

Andmed on tabelis

Doos X	1	2	3	4	5	6	7	8
Kaalu juurdekasv Y	1	1,2	1,8	2,5	3,6	4,7	6,6	9,1

Andmete hajuvusdiagrammilt on näha, et parabool sobib andmetega paremini kui sirgjoon.



Mudeliks saame $y = 1.35 - 0.41x + 0.17x^2$.

Peatükk 8

Mittelineaarne mudel

Mittelineaarses mudelis[1] vähemalt üks parameeter esineb mittelineaarsel kujul, lihtsad näited

$$y = \alpha e^{\beta x} + \varepsilon,$$
$$y = \alpha + \beta_1 x_1^{\gamma_1} + \beta_2 x_2^{\gamma_2}.$$

On valdkondi (füüsikas, keemias, bioloogias, inseneriteadustes), kus eksperimendist või teooriast on eelnevalt teada, et sobivaim on mingi mittelineaarne mudel.

Mittelineaarse mudeli üldkuju on järgmine

$$y_i = f(x_i, \theta) + \varepsilon_i,$$

kus $\theta = (\theta_1, \ldots, \theta_p)^T$ on parameetervektor, mis sisaldab p parameetrit (kusjuures $n > p$) ja $f(\cdot)$ on mingi mittelineaarne funktsioon parameetri θ suhtes. Sel juhul kasutatakse parameetrite hindamiseks mittelineaarset vähimruutude meetodit, millede hulgast kõige levinum on Gauss-Newtoni meetod

$$SSE = \sum [y_i - f(x_i, \hat{\theta})]^2,$$

kus $\hat{\theta}$ on parameetervektor, mis minimiseerib hälvete ruutude summa. Tegemist on iteratiivse lahendusmeetodiga, mis nõuab algväärtustamist. Olgu algväärtusvektor $\theta_0 = (\theta_{10}, \ldots, \theta_{p0})^T$. Mittelineaarne funktsioon $f(x_i, \theta)$ arendatakse Taylori ritta kohal $\theta = \theta_0$. Seega

$$f(x_i, \theta) \approx f(x_i, \theta_0) + (\theta_1 - \theta_{10}) \frac{\partial f(x_i, \theta)}{\partial \theta_1} |_{\theta=\theta_0} + \ldots$$

Esitatud avaldis on mittelineaarse funktsiooni lineariseerimine. Saadud avaldise saab ümber kirjutada järgmiselt

$$f(x_i, \theta) - f(x_i, \theta_0) \approx \gamma_1 w_{1i} + \ldots + \gamma_p w_{pi}, \; i = 1, 2, \ldots, n \qquad (8.1)$$

[1] Peatükk on refereeritud raamatust Myers (1990). Classical and modern regression with applications. Duxbury Press, pt 9.

kus

$$w_{ji} = \frac{\partial f(x_i, \theta)}{\partial \theta_j}|_{\theta=\theta_0}$$

ja $\gamma_j = \theta_j - \theta_{j0}$.

Viimase mudeli (8.1) vasakul pool on tegelikult jäägid $y - f(x_i, \theta_0)$ algväärtusest, seega on meil lineaarne mudel

$$y - f(x_i, \theta_0) = \gamma_1 w_{1i} + \ldots + \gamma_p w_{pi} + \varepsilon_i, \ i = 1, 2, \ldots, n.$$

Hindamist alustatakse algväärtusest, saadakse hinnangud kordajatele γ_j, $j = 1, \ldots, p$ ja uueks algväärtuseks on $\theta_1 = \theta_0 + \hat{\gamma}_1$, kus $\hat{\gamma}_1$ on esimesel iteratsioonisammul saadud hinnang. Protsess kordub kuni leiab aset koondumine.

8.1 Tuntud mittelineaarsete mudelite klassid

On terve rida rakendusvaldkondi, kus mittelineaarse mudeleid kasutatakse, samas tuleb tähele panna, et mõistmata probleemi teoreetilist tagapõhja, ei saa mittelineaarse mudeli kasutamine olla edukas. Osa mittelineaarseid mudeleid on jagatud teatud kategooriatesse vastavalt kasutusvaldkonnale, neist tuntumad on kasvumudelid. Kasvumudelid kirjeldavad millegi kasvu ja neid kasutatakse bioloogias, metsanduses, zooloogias, kus organismid ja/või taimed kasvavad ajas.

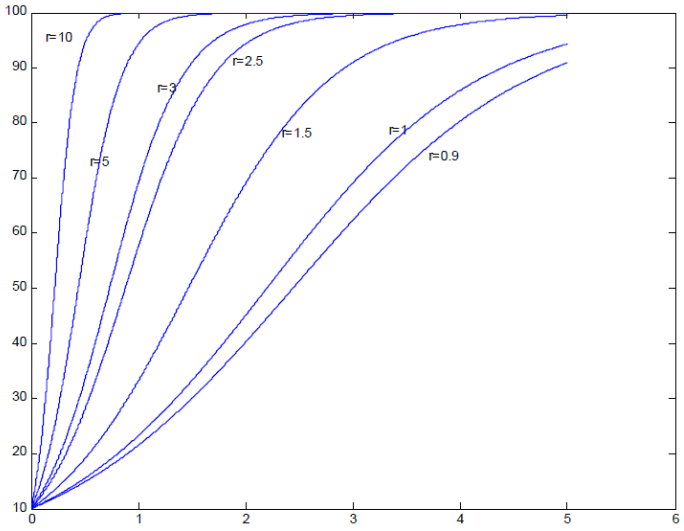
Mõned tuntumad kasvumudelite klassid on järgmised:

- Logistiline kasvumudel

  $$y = \frac{\alpha}{1 + \beta \exp(-kx)} + \varepsilon,$$

  kus parameetrid $\beta, k > 0$ ja kui $x \to \infty$, siis $y \to \alpha$, seega parameeter α kannab kasvu piiri nimetust ja parameeter k on kasvu kiirus.

  Logistilise kasvumudeli kujud erinevate kasvukiiruste korral (kasutatud tähistust r) on järgmisel joonisel.



Joonis. Logistiline kasvumudel

- Richardsi kasvumudel (iseloomustb 'S'-kujulist kasvu), on logistilise kasvumudeli edasiarendus, lisatud veel üks parameeter

$$y = \frac{\alpha}{[1 + \beta \exp(-kx)]^{\frac{1}{\delta}}} + \varepsilon.$$

- Mitcherlichi seadus (*Mitcherlich Law*), mida kasutatakse eelnevast mõnevõrra teistsuguse kasvu iseloomustamiseks, argument määrab teatava tõuke kasvu suurenemiseks (näiteks väetis). Mudelil on kuju

$$y = \alpha - \beta \exp(-\gamma x) + \varepsilon$$

  või selle erinevad parametrisatsioonid, nagu näiteks

$$y = \exp(\alpha) - \beta\gamma^x + \varepsilon, \quad y = \alpha - \exp(-\beta + \gamma x) + \varepsilon.$$

  Selle mudelite klassi üheks näiteks on MacArtur-Wilsoni kasvumudel

$$y = \alpha[1 - \exp(-\beta x)] + \varepsilon.$$

  Tegemist on mudelite klassiga, kus uuritaval tunnusel (saagikus, keemiline reaktsioon) on teatav kasvav iseloom, on sarnasus kasvumudelitele, aga pole käänupunkte nagu logistiliste kasvumudelite korral.

8.2 Mittelineaarsete mudelite näited

Näide: MacArtur-Wilsoni kasvumudel

Probleemiks oli pinnase vajumine kaevanduse kohal. Mõõdetud kaevanduse sügavus (d), kaevanduse laius (w) ja nurk, mille all kaevandamine toimub (uuritav tunnus y), $n = 16$.

Mudeli üldkuju on järgmine

$$y_i = \alpha[1 - \exp(-\beta\frac{w_i}{d_i})],$$

argumendiks on laiuse ja sügavuse suhe.

Algväärtus parameetrile α saadakse järgmiselt: kui suhe $\frac{w_i}{d_i}$ kasvab, siis kasv saavutab kasvu piiri, järelikult $y \to \alpha$, maksimaalne kasv annab parameetri algväärtuseks $\alpha_0 = 35$. Teine alglähend β_0 hinnatakse graafikult. Mudelit teisendatakse, kuni jõutakse kujuni

$$\ln(1 - \frac{y}{\alpha}) = -\beta\frac{w}{d},$$

mida võib vaadelda kui mudelit $y^* = -\beta x^*$, vaadates sellele vastavat graafikut, on näha et $\hat{\beta} = 1$, mis ongi alglähendiks β_0.

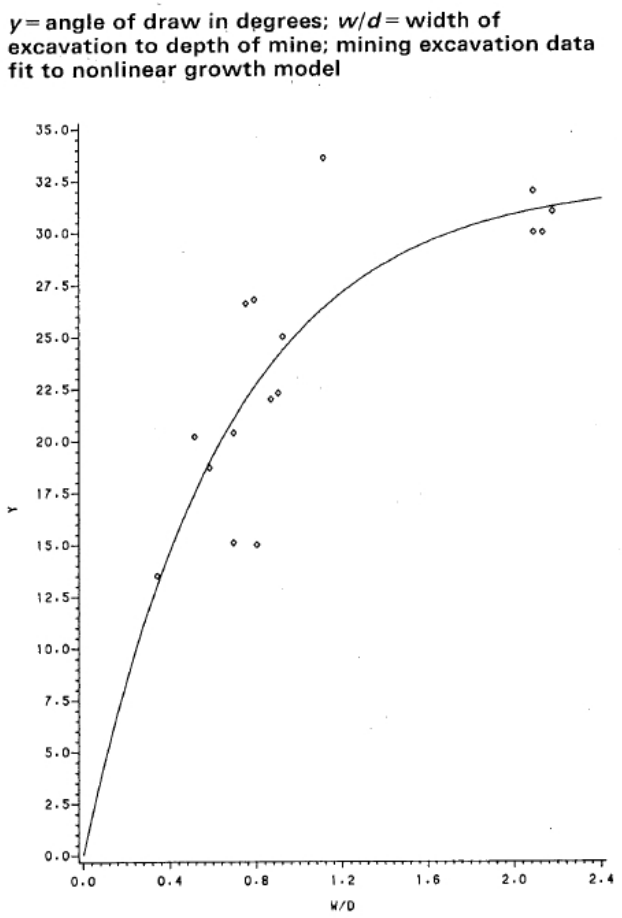
Lahendamiseks kasutatakse SAS protseduuri NLIN, kus PARMS lauses antakse algväärtused ja MODEL lauses mudeli täpne kuju

```
proc nlin data=kaevandus;
parms a=35 b=1.0;
model nurk=a*(1-exp(-b*suhe)); run;
```

Teostatakse 4 iteratsioonisammu ja tulemuseks saadakse järgmised parameetrite hinnangud (koos usaldusvahemikega):

$$\hat{\alpha} = 32.5 \quad (26.9; 38.1), \quad \hat{\beta} = 1.5 \quad (0.9; 2.1)$$

Andmed koos mudelile vastava joonega on kujutatud joonisel järgmisel leheküljel.



Joonis. Kaevandamise andmed. MacArtur-Wilsoni kasvumudel

Allikas: Myers (1990). Classical and modern regression with applications. Duxbury Press, pt 9, näide 9.1, lk 430.

SAS. Proc NLIN näiteülesanne

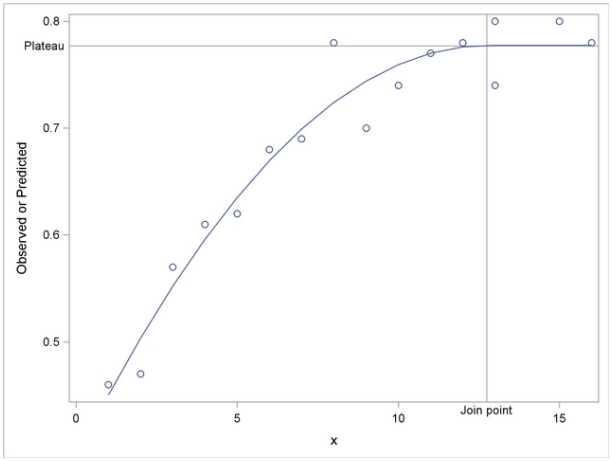
(*Example 60.1. Segmented Model*)

Oletame, et mingi protsess teatud kohani käitub kui ruutfunktsioon ja teatud kohast alates saavutab mingi platoo (nivoo). Otsitakse mudelit kujul

$$y = \begin{cases} \alpha + \beta x + \gamma x^2, & x < x_0 \\ c, & x \geq x_0 \end{cases}$$

Hindamise tulemuseks saadakse käänupunkt $x_0 = 12.75$ ja platoo väärtus $c = 0.78$

Tulemusi illustreerib järgmine joonis.



Joonis. Vaadeldud väärtused ja hinnatud mudel

8.3 Lineariseerimine

Mittelineaarse mudeli saab teatud teisendustega muuta lineaarseks. Tekib küsimus, miks siis mitte lahendadagi lineaarne mudel mittelineaarse asemel, sest lineaarse mudeli lahendamine on ju lihtsam? Vaatame kahte näidet.

Näide 1

$$y = \alpha e^{\beta x} + \varepsilon \tag{8.2}$$

Logaritmides, saame lineaarse mudeli $\ln y$ jaoks

$$\ln y = \ln(\alpha) + \beta x + \varepsilon_* \tag{8.3}$$

Näide 2

$$y = \frac{V}{k + x} + \varepsilon \tag{8.4}$$

Pöördteisendusega saame lineaarse mudeli $1/y$ suhtes

$$1/y = \frac{k}{V} + \frac{x}{V} + \varepsilon_{**} \tag{8.5}$$

Kas mudeli (8.2) võiks asendada mudeliga (8.3)? Mudeli (8.4) mudeliga (8.5)? Saaksime lineaarsed mudelid ja lahendamine oleks lihtne.

Tegelikult aga vähimruutude hinnangud on põhimõtteliselt erinevad mudelite (8.2) ja (8.3) (või (8.4) ja (8.5)) jaoks. Seega lineariseerimine ei vii ekvivalentsete mudeliteni vaid me võime seda kasutada ainult algväärtuste leidmiseks.

Märkusi mittelineaarsete mudelite juurde

Mudeli kuju peab olema teada (teooriast).

Vajalik on parameetrite algväärtustamine

- algväärtus võib olla teada eksperimentaalselt,
- algväärtus võib omada teatud sisu,
- algväärtuse võib saada lineariseerimisel.

Gauss-Newtoni meetodi modifikatsioonid on maksimaalselt sõltumatud algväärtusest, kuid siiski ei tohi ignoreerida algväärtustamise tähtsust. Halb algväärtus võib viia lokaalse miinimumini SSE minimiseerimisel ja ei pruugi seega anda õiget tulemust.

Peatükk 9

Üldistatud lineaarne mudel

Klassikaliste mudelite ühe eeldusena nõutakse, et uuritav tunnus oleks normaaljaotusega. Eelduste mittekehtimisel saame enamasti mitteefektiivse mudeli – tegelikkuses kehtivad seosed võivad mudelis tulla mitteolulised ja vastupidi, mitteolulised seosed tulevad mudelisse olulistena.

Probleemi lahendamiseks on kaks võimalust:

- Teisendame uuritavat tunnust nii, et selle jaotus muutuks võimalikult lähedaseks normaaljaotusele (st teeme skaalateisenduse, kasutame y asemel $\log y$ või $1/y$ jne). Teisenduste kasutamine ei pruugi alati olla tulemuslik, sest alati ei õnnestu skaalateisendusega asja parandada ja/või saadud mudel võib olla raskesti interpreteeritav. On terve rida jaotusi, mida me ei saa lähendada normaaljaotusele, nagu tugevasti asümmeetrilised jaotused või diskreetsed jaotused.
- Kasutame üldistatud lineaarseid mudeleid (*generalized linear models*), kus kasutame andmetele sobivat jaotust eksponentsiaalsest jaotuste perest ega proovigi seda teisendada normaaljaotusele lähedaseks.

9.1 Üldistatud lineaarsete mudelite klass

Üldistatud lineaarsete mudelite korral eeldatakse, et uuritava tunnuse jaotus on eksponentsiaalsest jaotuste perest ja kasutatakse teatud funktsiooni sidumaks jaotuse keskväärtust argumentide lineaarkombinatsiooniga nn **seosefunktsioon** (*link function*).

Klassikalise lineaarse regressioonimudeli kuju on

$$\boldsymbol{y} = \mathbf{X}\beta + \epsilon \quad \text{ehk keskmistades, } \mu = \mathbf{X}\beta,$$

kus $\mu = E\boldsymbol{y}$.

Üldistatud lineaarse mudeli korral võetakse kasutusele teatav seosefunktsioon, mida tavaliselt tähistatakse $\eta = g(\mu)$ ja mudelil on kuju:

$$g(\mu) = \mathbf{X}\beta.$$

Seosefunktsioon on mingi funktsioon uuritava suuruse keskväärtusest.

Tuntumad seosefunktsioonid:

- **Logit** seosefunktsioon

$$\eta = \ln \frac{\pi}{1-\pi}, \quad \text{siis} \quad \pi = \frac{e^\eta}{1+e^\eta}.$$

  Kasutatakse juhul, kui uuritav tunnus on binoomjaotusega $Y \sim B(n, \pi)$, kus n on katsete arv ja π on meid huvitava sündmuse tõenäosus.

- **Log** seosefunktsioon

$$\eta = \ln \mu, \quad \text{siis} \quad \mu = e^\eta.$$

  Kasutatakse juhul, kui uuritav tunnus on Poissoni jaotusega $Y \sim Po(\mu)$, kus μ on uuritava tunnuse keskväärtus (tavaliselt tähistatakse Poissoni jaotuse keskväärtus (jaotuse parameeter) tähega λ).

Klassikaline lineaarne regressioon vastab erijuhule, kui meil on tegemist **identsusseosega** (*link identity*) $\eta = \mu$.

Seega on üldistatud lineaarsel mudelil kuju

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

kus $\beta_0, \beta_1, \ldots, \beta_k$ on mudeli tundmatud parameetrid, $x_1, x_2, \ldots, x_k$ on seletavad tunnused ehk argumenttunnused, μ on uuritava tunnuse keskväärtus ja g on seosefunktsioon, mille kuju sõltub uuritava tunnuse jaotusest.

Mudeli parameetrite hindamine

Tundmatute parameetrite β_i leidmine ehk nende hindamine valimi põhjal toimub **suurima tõepära meetodil** (*maximum likelihood method*). Kasutatakse valimi tõepärafunktsiooni, mille kuju sõltub uuritava tunnuse jaotusest ja mida tähistatakse $L(x, \theta)$, kus x on valim ja θ on otsitav parameeter.

Suurima tõepära hinnangu korral leitakse selline parameetri väärtus, mille korral tõepärafunktsioon saavutab maksimumi.

Tehnilise töö lihtsustamiseks on võetud kasutusele *logaritmiline tõepärafunktsioon* $l(x, \theta) = \ln L(x, \theta)$, mis saavutab maksimumi samas kohas.

Paketis SAS kasutatakse selle maksimumi leidmiseks tavaliselt kas Newton-Raphsoni meetodit või Fisheri skoorimeetodit.

Üle- ja alahajuvus

Vaadates funktsioontunnuse jaotusi, mille korral dispersioon ja keskväärtus on seotud (nagu näiteks binoomjaotus või Poissoni jaotus), tekib parameetrite hindamisel probleeme. Empiiriliste andmete järgi saame arvutada keskväärtuse hinnangu ja dispersiooni hinnangu, aga võib juhtuda, et nende hinnangute vahel ei kehti seos, mis teoreetiliselt peaks kehtima.

Mudelit tuleks täpsustada – lisada mudelisse parameeter, mis taandaks tekkinud vastuolu – skaalaparameeter φ.

Eksponentsiaalses jaotuste peres on φ nn liigne parameeter (*nuisance*). Eeldatakse, et ta on *a'priori* teada ja kui tundmatu parameeter on hinnatud, siis avaldatakse liigne parameeter hinnatud parameetri kaudu. Parameetrit φ interpreteeritakse kui tõepärafunktsiooni *skaalaparameetrit*.

Öeldakse, et uuritava tunnuse dispersioon Dy avaldub teoreetilise dispersiooni (tähistame näiteks γ) ja skaalaparameetri korrutisena: $Dy = \gamma\varphi$.

Räägitakse enamasti **ülehajuvusest** (*overdispersion*) st valdaval osal juhtudest on dispersiooni hinnang liiga suur ja seega $\varphi > 1$, harvem on tegemist **alahajuvusega** $\varphi < 1$.

Ülehajuvuse võimalikud põhjused:

- Andmetes on täiendav hajuvuse allikas, juhuslikkus, kus mudeliga kirjeldatud hajuvusele lisandub veel mingi hajuvus (näiteks mõõtmisviga). Teatud osa hajuvusest on kirjeldatud tunnuste kaudu ja mingi osa on kirjeldamata – ülehajuvus.

- Mudeli eeldused ei ole täidetud (y_i sõltumatuse nõue on rikutud). Väga tugevalt positiivselt seotud vaatluste korral dispersioon oluliselt suureneb. Negatiivne sõltuvus toob kaasa alahajuvuse.

- Andmestikus on erindid, mis suurendavad hajuvust.

Üle- või alahajuvuse korral kasutatakse **kvaasi-tõepära** meetodit (*quasi-likelihood*). Kvaasi-tõepära arvestab dispersiooni kuju kasutades skaalaparameetrit.

Kuna β hinnang $\hat{\beta}$ ei sõltu skaalaparameetrist φ, siis saame kvaasitõepära meetodil samad hinnangud $\hat{\beta}$ nagu ülehajuvust arvestamata, aga kordajate hajuvuse hinnang muutub ning seega võib muutuda kordajate olulisus.

Mudeli headuse näitajad

Mudeli headuse määrab see, kui hästi mudel sobib, seega on mudeli headuse näitajatega samaväärne rääkida mudeli sobitusastmest (*Goodness of Fit – GOF*).

- Hälbimus (*deviance*), mis näitab hälbimust ideaalsest ehk küllastunud mudelist. Mida väiksem on hälbimus, seda parem on mudel. Hälbimus avaldub logaritmiliste tõepärafunktsioonide kaudu järgmiselt:

  $$D = 2(l(y, y) - l(y, \mu)),$$

  kus $l(y, y)$ vastab küllastunud mudelile ja $l(y, \mu)$ uuritavale mudelile. Kui mudeli jääkhajuvus on täiesti juhuslik, siis on hälbimus χ^2-jaotusega vabadusastmete arvuga $n-p$, kus n on valimi maht ja p on parameetrite arv mudelis.

- Pearsoni-χ^2 statistik, mis on samaväärne hälbimusega;

- Tõepära statistik $-2 \ln L$;

- AIC (*Akaike Information Criterion*) Akaiki informatsioonikriteerium, mis saadakse hälbimusele teatud parandusliikme lisamisel;

- SC (*Schwarz Criterion*) Svarzi kriteerium.

Toodud statistikute korral, mida väiksem on statistiku väärtus, seda parem on mudel.

Parameetrite olulisus

Mudeli kui terviku olulisuse testimiseks kasutatakse tõepärasuhte statistikut, Waldi statistikut ja skooristatistikut. Waldi teststatistiku lihtne erijuht on t-statistik. Kõik nimetatud statistikud on asümptootiliselt χ^2-jaotusega

Parameetrite β_i olulisust hinnatakse χ^2-statistikuga, mis näitab, kui palju suureneks mudeli hälbimus, kui vastav argument mudelist välja jätta. Kui χ^2-statistikule vastav olulisuse tõenäosus on väike ($p < 0.05$), siis kirjeldab see argument uuritava tunnuse hajuvusest olulise osa ja ta tuleb jätta mudelisse.

Mudeli jäägid

Nagu klassikaliste mudelite korral nii ka üldistatud lineaarsete mudelite korral on oluline tähtsus jääkide analüüsil. Üldistatud mudelite korral pole tavalise jäägi jaotust lihtne hinnata ja püütakse defineerida jäägid, mille jaotus oleks normaaljaotusele lähedane, et oleks võimalik kasutada klassikalise mudeli teooriat. Siin on kasutusel **hälbimuse jäägid** (*deviance residual*) ja standardiseeritud hälbimuse jäägid (*standardized deviance residual*)

ning **Pearsoni jäägid** ja standardiseeritud Pearsoni jäägid. Üksikud suured jäägid viitavad erinditele andmestikus.

Erinevate mudelite korral ei pruugi üldistatud jäägi jaotus olla normaaljaotusele piisavalt lähedal, aga suur standardiseeritud või Studenti üldistatud jääk (> 3) näitab siiski erindit suure jäägi mõttes.

Omapärased punktid baseeruvad üldistatud mütsimaatriksil, aga see võib sõltuda parameetrite hinnangutest, seega ei pruugi hästi töötada. Mõjusad vaatlused baseeruvad hinnangutel, mis saadakse, kui antud vaatlus jäetakse välja, mis üldistatud mudelite kontekstis on töömahukas. Lihtsuse mõttes kasutatakse tavaliselt ainult esimest iteratsioonisammu, mis ei pruugi samuti õigesti töötada. Seega omapäraste ja mõjusate vaatluste väljaselgitamiseks puuduvad empiirilised kriteeriumid, kasutatakse erinevaid graafikuid, kus teistest eemal asetsevad punktid vajaksid täiendavat uurimist.

9.2 Logistiline regressioonimudel

Paljude mudelite seas pakub tihti erilist huvi selline, kus funktsioontunnusel on ainult kaks võimalikku väärtust: on/ei ole, jah/ei, esineb/ei esine. Väärtused kodeeritakse tavaliselt 1/0, selliselt, et 1 tähistab sündmuse esinemist ja $\mathbf{P}(Y = 1) = \pi$; $\mathbf{P}(Y = 0) = 1 - \pi$.

Tegemist on Bernoulli $Y \sim B(1, \pi)$ või binoomjaotusega $Y \sim B(n, \pi)$, kus n on katsete arv, π on meid huvitava sündmuse tõenäosus.

Bernoulli jaotuse korral $EY = \pi$, $DY = \pi(1 - \pi)$ ja binoomjaotuse korral $EY = n\pi$, $DY = n\pi(1 - \pi)$. Seega mudel keskväärtusele hindab sündmuse toimumise tõenäosust. Huvi pakub seos uuritava tunnuse esinemise tõenäosuse π ja mõõdetud seletavate tunnuste vahel.

Binaarse uuritava tunnuse korral on probleemiks asjaolu, et prognoositakse tõenäosust, mis on tõkestatud lõigul $[0, 1]$. Tuleb leida teisendus (üksühene, pidev, diferentseeruv) kogu reaalteljele.

Binaarse uuritava tunnuse korral on kasutusel *Logit* seosefunktsioon

$$\eta = logit(\pi) = \ln \frac{\pi}{1 - \pi} \; ; \quad \text{kus } \frac{\pi}{1 - \pi} \text{ on sündmuse esinemise šanss}$$

Logistilise mudeliga hinnatakse seega šansi logaritmi

$$\ln \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

kus $\pi = \mathbf{P}(Y = 1)$ on sündmuse esinemise tõenäosus.

Logit seosest saame avaldada sündmuse esinemise tõenäosuse ehk *prognoosi tõenäosusele*

$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}.$$

Logistiline mudel on üks võimalik mudel binaarsele tunnusele, peale *Logit* seosefunktsiooni on võimalikud ka teised (nagu näiteks *Probit* või *CLog-Log*), aga logistiline mudel on kõige paremini interpreteeritav.

9.2.1 Logistilise mudeli interpretatsioon

Šanss ja šansside suhe

Sündmuse šanss (*odd*) on defineeritud kui sündmuse esinemise tõenäosuse ja sündmuse mitteesinemise tõenäosuse suhe

$$\Pi = \frac{\pi}{1 - \pi}.$$

Seega on *Logit* on šansi logaritm $Logit(\pi) = \ln \frac{\pi}{1-\pi}$.

NB! Kui pole näidatud logaritmi alust, siis log ja ln tähistavad mõlemad naturaallogaritmi!

Mudeli parameetrite interpretatsioon on seotud šansside suhte muutusega. Šansside suhe (*odds ratio*) on defineeritud kui kahe isiku (i-nda ja j-nda) šansside suhe

$$OR = \frac{\Pi_i}{\Pi_j} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}}.$$

Parameetri ees oleva märgi interpretatsioon on järgmine: pluss märk näitab samapidist seost argumendi ja uuritava tunnuse (tõenäosuse) vahel ja miinus vastupidist seost.

Parameetri β interpretatsioon:
Saab näidata, et *ühikulise argumendi muutusega kaasneb šansside suhte muutus* $e^{\hat{\beta}}$ *korda* ja *kui argument muutub* c *ühikut, siis kaasneb šansside suhte muutus* $e^{c\hat{\beta}}$ *korda*. Mitme argumendiga mudeli interpreteerimisel tuleb lisada, et teised argumendid ei muutu (ehk muude tingimuste samaks jäädes).

Vabaliikme interpretatsioon:
st $x = 0$, võimalik juhul, kui see on argumendi võimalik väärtus, siis saab positiivse vabaliikme korral näidata, et sündmuse esinemise tõenäosus on suurem kui pool, st kui $x = 0$ ja $\hat{\alpha} > 0$ siis $\hat{\pi} > 0.5$. Negatiivne vabaliige pole interpreteeritav.

9.2.2 Rühmitatud ja rühmitamata andmed

Rühmitamata andmed on tavaline olukord, kus meil uuritav tunnus on 1/0 tunnus, tegemist on Bernoulli jaotusega $Y \sim B(1, \pi)$, mille korral keskväärtus on $EY = \pi$. Teeme mudeli keskväärtusele (seega tõenäosusele) kasutades seosefunktsiooni.

Rühmitatud andmetega on tegemist, kui uuritav tunnus on positiivsete vastuste arv (ehk 1-de summa), tegemist on binoomjaotusega $Y \sim B(n, \pi)$, mille keskväärtus on $EY = n\pi$. Mudeli teeme tõenäosusele, aga arvesse tuleb nüüd võtta ka katsete koguarv.

Rühmitamata andmeid võib vaadelda kui rühmitatud andmete erijuhtu, kui igas rühmas tehtud 1 vaatlus.

Logistiline mudel. SAS

Paketis SAS on logistilise mudeli hindamiseks protseduur LOGISTIC. Protseduur hindab mudeli 1/2 tunnuse jaoks, vajadusel kodeerib 1/0 tunnuse ise ringi. Aga tuleb silmas pidada, et hinnatakse mudel madalama taseme jaoks, st 1/0 tunnuse korral hindab vaikimisi 0 esinemise tõenäosust, hindamaks 1 esinemise tõenäosust, tuleb protseduuri päisesse lisada valik DESCENDING.

Rühmitatud andmete korral tuleb anda MODEL lauses nii sündmuse esinemise arv kui ka koguarv, st Y-tunnus on kujul *event/trial* st antakse uuritav tunnus kahe tunnuse suhtena kujul: ühtede arv/koguarv.

Diskreetsete argumentide (faktorite) kasutamisel tuleb määrata baastase, mille suhtes hinnatakse (vaikimisi toimub hindamine tasemete keskväärtuse suhtes). Baastaseme saab määrata näiteks järgmiselt:
CLASS *faktor* param=ref ref=last; toimub hindamine viimase taseme suhtes, võimalik ka valik ref=first;

Näide. Logistiline mudel

Suur osa lahutatud naistest ei saa oma endistelt meestelt mingit toetust. Tahetakse välja selgitada, miks see nii on. Kas meeste palk ei võimalda maksta? Või arvavad mehed, et naised teenivad isegi hästi? Kas laste arv mõjutab meeste otsust maksta alimente?

Uuritav tunnus: 1 – naine saab alimente, 0 – ei saa
(valimi põhjal hinnatud mudel, tavaliselt tähistatakse $p = \hat{\pi}$)
Mudel: $logit(p) = -1.73 + 0.01$mehe palk $+ 0.31$laste arv $- 0.02$naise palk.
Tunnus 'naise palk' (viimase abieluaasta jooksul) osutus mudelis ebaoluliseks. Uus mudel: $logit(p) = -1.87 + 0.01$mehe palk $+ 0.33$laste arv.

Kui suur on tõenäosus, et mees, kellel on 3 last ja palk 10 tuhat $ maksab naisele alimente?

$$logit(p) = -1.87 + 0.01 \times 10 + 0.33 \times 3 = -0.76; \quad \text{seega} \quad \ln \frac{p}{1-p} = -0.76,$$

saame avaldada tõenäosuse, tulemuseks $p = 0.32$. Seega umbes kolmandik sellistest meestest maksab naistele alimente.

NB! Siin p on sündmuse esinemise tõenäosus, mitte olulisusetõenäosus!

9.2.3 Usaldusvahemik parameetritele ja šansside suhtele

Mudeli parameetritele usaldusvahemiku leidmiseks on 2 võimalust:

- Tõepärafunktsioonil põhinev lähenemine, kasutatakse asümptootilist χ^2-jaotust (LR usaldusvahemik);
- Parameetrite hinnangute asümptootilisel normaaljaotusel põhinev lähenemine, mis sobib suurte valimite korral (Wald'i usaldusvahemik).

Kui (a_j, b_j) on parameetri β_j usaldusvahemik, siis (e^{a_j}, e^{b_j}) on vastav šansside suhte usaldusvahemik. Kui usaldusvahemik sisaldab väärtuse 1, siis šansside suhe pole oluline (šansid on võrdsed sõltumata argumendi muutusest). Öeldakse ka, et kui usaldusvahemiku mõlemad otspunktid on ühest suuremad, siis on tegemist riskiteguriga

Näide. Šansside suhte usaldusvahemik

Probleem: laste olemasolu sõltuvus naise vanusest. Kui naise vanus on teada, siis kui suur on tõenäosus, et tal on laps?

Hinnatakse mudel:

$$Logit(p) = -5.3 + 0.11 vanus$$

Waldi usaldusvahemik argumendi ees olevale parameetrile on $(0.01; 0.21)$

Parameetri interpretatsioon: kui vaatame kahte naist, kellel vanuse vahe on 1 aasta, siis vanema naise korral on šanss, et tal on lapsi $e^{0.11} = 1.12$ korda suurem ehk 12% suurem. Usaldusvahemikuks saame $(e^{0.01}; e^{0.21}) = (1.01; 1.23)$. Seega võivad šansid olla ka 23% suuremad

Näide. Mitme argumendiga logistiline mudel

Uuriti südameataki korduvat esinemist. Uuritav tunnus on kodeeritud järgmiselt: $y = 1$, kui patsiendil oli aasta jooksul veel teine südameatakk ja $y = 0$, kui ei olnud. Tahetakse hinnata, kuidas on teistkordse südameataki esinemine seotud ravimite tarvitamisega ja patsiendi rahutuse tasemega.

Ravimite tarvitamine, st kas patsient on saanud rahusteid, on kodeeritud 1-jah, 0-ei. Patsiendi rahutuse taset mõõdetakse testiga, mille suurem testitulemus näitab rahutumat seisundit.

Saadud mudel on kujul:
$Logit(p) = \ln \frac{p}{1-p} = -6.36 - 1.02(ravi = 1) + 0.12 rahutus.$

Vastavad šansside suhted koos 0.95-usaldusvahemikuga:

```
Ravi 1 vs 0    0.36    (0.03; 3.6)
Rahutus        1.12    (1.03; 1.29)
```

Interpreteerida tulemust!

9.2.4 Üldistatud determinatsioonikordaja

Logistiline mudel on esimene üldistatud mudelitest, kus on kasutusel üldistatud determinatsioonikordaja ehk pseudo-R^2.

Cox ja Snell (1989) esitasid üldistatud determinatsioonikordaja kujul

$$R^2 = 1 - \left\{ \frac{L(0)}{L(\hat{\beta})} \right\}^{2/n},$$

kus $L(0)$ on tõepära konstantse (ainult vabaliikmega) mudeli jaoks, $L(\hat{\beta})$ on tõepära konkreetse vaadeldava mudeli jaoks, n on valimi maht. Definitsioonist on näha, et determinatsioonikordajat ei saa interpreteerida nii nagu klassikalise mudeli korral, siin on tegemist tõepärade suhtega. Samuti on selge, et selliselt defineeritud $R^2 < 1$ ja $R^2_{max} = 1 - \{L(0)\}^{2/n}$.

Nagelkerke (1991) esitas parandatud determinatsioonikordaja (SAS: *max-rescaled R-squared*). Tuntud ka kui Nagelkerke-R^2, mille maksimaalne väärtus on 1

$$\tilde{R}^2 = \frac{R^2}{R^2_{max}}.$$

SAS: Proc LOGISTIC korral MODEL lauses valik RSQUARE

9.2.5 Väärtuste järjestamisest

Tavaliselt hinnatakse *'edu' tõenäosust* st väärtuse 1 esinemise tõenäosust

$$Logit(\pi) = \ln \frac{\pi}{1-\pi}, \quad \pi = \mathbf{P}(Y = 1).$$

On lihtne näha, et kehtib

$$Logit(\pi) = -Logit(1 - \pi).$$

Järelikult mudeli hindamine $y = 0$ jaoks, võrreldes mudeliga $y = 1$ jaoks, tähendab kordajate märkide muutust.

Protseduur Logistic hindab vaikimisi uuritava tunnuse madalamat taset, vaatleb tasemeid kasvavas järjekorras ja hindab esimest. Mudeli parameetrite interpreteerimisel tuleb arvestada, mida on hinnata tahetud ja mida on tegelikult hinnatud. Seega, kui tahame hinnata 1 tõenäosust, siis mudeli lihtsama interpreteeritavuse huvides on soovitav uuritava tunnuse väärtused ümber järjestada (valik DESCENDING protseduuri lauses).

9.2.6 Uuritaval tunnusel rohkem kui 2 taset

Vaatasime mudeleid, kui uuritaval tunnusel on 2 taset 1/0 või 1/2.

Uuritaval tunnusel võib aga olla rohkem kui 2 väärtust $1, 2, \ldots, K$, kus K ei ole väga suur (tihti $K < 5$).

Näiteks uuritakse valu esinemist: huvi võib pakkuda, kas valu esineb või mitte (1/0) või kas on tugev valu, keskmine valu, nõrk valu või valu pole st kodeeritakse $1, 2, 3, 4$ – uuritaval on 4 taset (väärtust).

Erinevad väärtused võivad olla järjestatud või mitte. Enne mudeli hindamist tuleb välja selgitada, kas uuritava tunnuse hindamine rohkem kui 2 väärtusega skaalal on tõepoolest õigustatud, st kas iga uuritava tunnuse taseme jaoks on piisavalt andmeid. Tihti esineb olukord, kus uuritavat tunnust püütakse hinnata rohkem kui 2 tasemega, aga tegelikkuses taandub probleem ikkagi 2 taseme hindamiseks. Olenevalt probleemist, võib näiteks 4 tasemega uuritava tunnuse korral leida mudelid {1 *vs* 2, 3, 4}, {1, 2 *vs* 3, 4}, {1, 2, 3 *vs* 4}.

Kui uuritaval on rohkem kui 2 taset pole tegemist binoomjaotusega vaid selle mitmemõõtmelise üldistusega – multinomiaaljaotusega. Mitme tasemega uuritavale tunnusele hinnatakse kumulatiivne mudel

$$g(\mathbf{P}(Y \leq j|x)) = \alpha_j + \beta x, \quad 1 \leq j \leq K,$$

kus $\alpha_1, \alpha_2, \ldots, \alpha_K$ on erinevad vabaliikmed iga taseme jaoks, β on ühine tõus.

Kui tasemed on järjestatud sobib sellisele uuritavale tunnusele logistiline mudel. Protseduur Logistic hindab kumulatiivse mudeli kui tasemete arv $K > 2$. Logistilist kumulatiivset mudelit nim ka võrdeliste šansside mudeliks (*proportional odds*), sest šansside suhe $Y \leq j$ jaoks ei sõltu tasemest j.

Mudelid iga taseme jaoks saab välja kirjutada järgmiselt

$$Logit(p_1) = \alpha_1 + \beta x \quad Logit(p_1 + p_2) = \alpha_2 + \beta x \quad \ldots$$
$$Logit(p_1 + \ldots + p_k) = \alpha_k + \beta x.$$

9.3 Loendusandmete mudelid

Loendusandmed kujutavad endast mingi sündmuse esinemise arvu teatud ajavahemikus. Klassikaline näide on Geigeri loendaja, mis mõõdab radioaktiivsust. Tuntud näited on kindlustuse ja töökindluse valdkonnast: kindlustussumma saajate arv, õnnetusjuhtumite arv, tööpingi tõrgete arv jne. Teatud tingimustel saab selliseid andmeid lähendada ka normaaljaotusega, kuid üldiselt on tegemist Poissoni jaotusega.

Poissoni jaotusega on andmed, mis tekivad millegi loendamisel, nii et suurte arvude esinemine on väikese tõenäosusega. Poissoni jaotuse tüüpiline histogramm on ebasümmeetriline ja mida väiksem on jaotuse parameeter, seda ebasümmeetrilisem on jaotuse histogramm.

Jaotuse parameetrit interpreteeritakse kui harvaesinevate sündmuste keskmist arvu ajaühikus. Poissoni jaotust nimetatakse ka harvaesinevate sündmuste jaotusseaduseks ehk väikeste arvude seaduseks.

Poissoni jaotuse korral on *keskväärtus ja dispersioon võrdsed*, see on jaotuse tähtis omadus.

Teine oluline Poissoni jaotuse omadus on aditiivsus:

Kui $Y_1 \sim Po(\mu_1)$ *ja* $Y_2 \sim Po(\mu_2)$ *on sõltumatud, siis* $Y_1 + Y_2 \sim Po(\mu_1 + \mu_2)$.

Viimane omadus annab võimaluse käsitleda sarnaselt nii rühmitatud kui ka rühmitamata andmeid. Olgu Y_{ij} sündmuste arv i-ndas rühmas j-ndal vaatlusel ja Y_i kogu sündmuste arv rühmas i. Siis sõltumatuse eeldusel, kui üksikvaatlused $Y_{ij} \sim Po(\mu_i), \quad j = 1, \ldots, n_i$, siis $Y_i \sim Po(n_i\mu_i)$ – nii üksikvaatluste kui ka rühmitatud vaatluste korral kasutatakse *sama tõepärafunktsiooni*.

9.3.1 Võimalikud mudelid

Loendusandmetele on võimalik sobitada sõltuvalt andmestiku iseloomust järgmisi mudeleid

- Lineaarne ja/või multiplikatiivne Poissoni mudel.
- Negatiivse binoomjaotuse mudel (*NB*-mudel).
- Nulliprobleemidega mudelid (nulle on kas liiga palju või liiga vähe)
  – ZIP mudel, nullidega mõjutatud mudel (*Zero Inflated*),
  – ZAP mudel (*Zero Alternated*) ehk tõkestatud (*Hurdle*) regressioon,
  – ZTP mudel, nullidega lõigatud mudel (*Zero Truncated*).

Nulliprobleemid võivad esineda ka negatiivse binoomjaotuse korral, sel korral räägitakse ZINB, ZTNB, ZANB mudelitest.

Märkus. Klassikaliselt on log-lineaarne regressioonimudel mudel sagedustabelile.

Poissoni mudeli kujud

Oletame, et meil on valim n vaatlust $y_1, \ldots, y_n$, mida võime vaadata kui Poissoni jaotusega juhusliku suuruse Y_i realisatsioone $Y_i \sim Po(\mu_i)$ ja oletame, et keskväärtus μ_i (ja seega ka dispersioon!) sõltub seletavatest muutujatest.

Lineaarne (aditiivne) Poissoni mudel $\mu = \mathbf{X}\beta$
Seda mudelit võib kasutada, kui on eeldatud, et kovariantide mõju on aditiivne. Mudeli puuduseks on asjaolu, et mudeli parem pool võib omandada mistahes väärtusi reaalteljel, kuid mudeli vasak pool, mis vastab prognoositud sündmuste arvule, saab olla ainult positiivne.

Sellest puudusest vabanemiseks võetakse kasutusele *Log-teisendus*.

Log-lineaarne Poissoni mudel $\eta = \ln(\mu), \quad \eta = \mathbf{X}\beta$.
Selles mudelis regressioonikordaja β_j vastab muudatusele uuritava keskväärtuse logaritmis kui kovariant x_j muutub 1 ühiku võrra.

Seega saame **multiplikatiivse mudeli** $\mu = \exp(\beta_0 + \beta_1 x_1 \ldots + \beta_k x_k)$, kus e^{β_j} vastab x_j muutuse multiplikatiivsele mõjule: *kovariandi* x_j *muutus 1 ühiku võrra tekitab keskväärtuses* μ_i *muutuse* e^{β_j} *korda.*

Logaritmilise seosefunktsiooni eeliseid on kinnitanud ka empiirilised uuringud. Loendusandmete korral on kovariantide mõju pigem multiplikatiivne kui aditiivne, st tüüpiliselt on kovariandil väike mõju väikesele sündmuste esinemise arvule ja suur mõju suurele. Seega mõjud on proportsionaalsed sündmuse esinemise arvuga ja sel juhul *Log-skaala* kasutamine viib lihtsamale mudelile ja õigustab end.

9.3.2 Ülehajuvus

Ülehajuvus on loendusandmete korral tõsine probleem. Poissoni jaotuse korral peab kehtima range vastavus $E(Y) = D(Y)$. Tavaliselt eeldatakse $D(Y) = \varphi E(Y)$, kus φ on skaalaparameeter, mis on konstantne üle kõigi andmete. Kui $\varphi = 1$, siis probleemi pole, kui $\varphi > 1$ ehk $D(Y) > E(Y)$, siis on tegemist ülehajuvusega ja kui $\varphi < 1$ ehk $D(Y) < E(Y)$, siis on tegemist alahajuvusega. Öeldakse, et alahajuvus ei tekita probleeme mudeli parameetrite hindamisel ja seda on vähem uuritud.

Skaalaparameeter hinnatakse hälbimuse D või Pearsoni χ^2-statistiku abil järgmiselt

$$\hat{\varphi} = \frac{D}{df}, \quad \hat{\varphi} = \frac{\chi^2}{df}.$$

Otsustusreegel on järgmine: skaalaparameetri hinnang peab olema ≈ 1, siis ei ole probleeme üle- ega alahajuvusega.

Ülehajuvuse põhjused

Räägitakse nn näilisest ülehajuvusest ja tegelikust ülehajuvusest.

Ülehajuvuse põhjused:

- mudeli süstemaatiline osa on valesti määratud, kas puudub mõni oluline argument või argumentide koosmõju;
- argumentide skaala pole hea, tuleks teha skaalateisendus (näiteks *log*);
- andmetes on erindid;
- andmetes on lisajuhuslikkus: näiteks tegemist Poissoni protsessiga vahemikus, mille pikkus on juhuslik;
- andmetes on probleemid nullidega (nulle liiga palju või liiga vähe).

Empiiriline reegel:

- Kui ülehajuvus on suur (> 5), siis midagi valesti, tegemist tegeliku ülehajuvusega, vali teine mudel.
- Kui ülehajuvus on väike (< 5), siis näiline ülehajuvus, võta see arvesse.

Ülehajuvuse probleemi lahendamisest

Näilise ülehajuvuse korral tuleks kontrollida erindite olemasolu ja mudeli süstemaatilist osa (kas on vajalikud argumendid kaasatud). Ülehajuvuse arvesse võtmiseks hinnatakse skaalaparameeter ja kasutatakse kvaasi tõepära (*quasi-likelihood*).

Tegeliku ülehajuvuse korral kasutatakse teisi jaotusi (hinnatakse *NB*-mudel või nulliprobleeme arvesse võtvad mudelid).

9.3.3 Mudel rühmitatud andmetele

Olgu meil rühmitatud andmed. Oletame, et y_{ij} on sündmuste arv i-ndas rühmas j-ndal vaatlusel ja $y_i = \sum_j y_{ij}$. Iga vaatlus rühmas on realisatsioon sõltumatust Poissoni jaotusega juhuslikust suurusest $Y_{ij} \sim Po(\mu_i), i = 1, \ldots, k; j = 1, \ldots, n_i$. Kogu rühm on realisatsioon Poissoni jaotusega juhuslikust suurusest $Y_i \sim Po(n_i\mu_i)$.

Log-lineaarne mudel individuaalse keskväärtuse jaoks avaldub kujul

$$\ln E(Y_{ij}) = \ln(\mu_i) = \beta_0 + \beta_1 x_1 \ldots + \beta_k x_k,$$

siis

$$\ln E(Y_i) = \ln(n_i\mu_i) = \ln n_i + \ln \mu_i = \ln n_i + \beta_0 + \beta_1 x_1 \ldots + \beta_k x_k,$$

Seega rühmitatud andmete korral on log-lineaarsel mudelil samad kordajad β kui rühmitamata andmete korral, ainult rühmitatud andmete korral lisandub liige $\boxed{\ln n_i}$, mida nimetatakse **OFFSET**.

Võime hinnata mudeli nii rühmitatud kui ka rühmitamata andmete korral. Parameetrite hinnangud ja standardvead tulevad samad, hälbimused loomulikult erinevad.

Peatükk 10

Faktoranalüüsi mudel

Faktoranalüüs[1] on mitmemõõtmelise analüüsi mudel.

Faktoranalüüsi mudeli korral asendatakse esialgsed tunnused väiksema arvu tunnustega nn faktoritega, mis kirjeldavad võimalikult suure osa lähtetunnuste hajuvusest.

Meetod töötati välja 1920ndate aastate alguses psühholoogias, kus kasutati mitmesuguseid teste mõõtmaks inimese vaimseid võimeid. Püsitati hüpotees, et leiduvad teatavad mittemõõdetavad (latentsed) tunnused (olulised iseloomujooned), mille kohta saab informatsiooni testide kaudu. Latentseid tunnuseid on väike arv ja need kirjeldavad inimese isiksuse struktuuri.

Latentne tunnus kannab siin nimetust *faktor*. Testide tulemused on kirjeldatavad faktorite kombinatsioonidena.

NB! Mitte segi ajada faktortunnuse kui diskreetse argumendiga!

Faktoranalüüsi ülesanne

Faktoranalüüsi ülesandeks on leida võimalikult väike arv uusi tunnuseid, mille lineaarkombinatsioonidena saaks uuritavaid lähtetunnuseid küllalt hästi kirjeldada. Faktoranalüüsi korral ei räägita funktsioontunnusest ja argumenttunnustest, kõik tunnused on käsitletavad samamoodi.

Faktoranalüüsi eesmärgid:

- info kokkusurumine (mõõdetud tunnuste asendamine väiksema arvu mittemõõdetavatega),
- tunnustevahelise sõltuvusstruktuuri analüüs,

[1] Peatükk põhineb õpikul Ehasalu, Tiit (1993). *Faktoranalüüs ja konooniline analüüs SAS-süsteemis.* Käsiraamat üliõpilastele II, TÜ.

- mudel ise (latentsete tunustega mudeli kirjeldamine).

Faktoranalüüs on enamasti hüpoteeside genereerimise metoodika st ei kontrollita statistilisi hüpoteese (sõltub eelduste rangusest) ja ei väljastata olulisusetõenäosusi. Seega pole faktoranalüüsis saadud tulemused tõestatud väited, vaid hüpoteetilised.

Faktoranalüüsi subjektiivsusest

Faktoranalüüsi läbiviimisel peab uurija ise otsustama paljude asjade üle ja seetõttu võib samade andmete korral saada mitu faktormudelit, mis võivad sobida ühtviisi hästi.

Uurija peab ise otsustama

- milliste tunnuste alusel faktoranalüüs teostatakse,
- kui palju faktoreid mudelis on,
- millist meetodit faktorite leidmisel kasutada,
- kas ja kuidas faktormaatriksit pöörata,
- kuidas saadud faktoreid interpreteerida.

Tulemused on seega tihti subjektiivsed, sõltuvad uurija teadmistest, oskustest ja kogemustest (väheste kogemuste korral on oht saada triviaalseid mudeleid).

Faktoranalüüsi eeldused

Faktoranalüüsi aluseks on korrelatsioonanalüüs või kovariatsioonanalüüs. Korrelatiivne sõltuvus on lineaarne sõltuvus kahe arvtunnuse vahel, järelikult saab faktoranalüüsi kasutada arvtunnuste (või järjestustunnuste) korral. Faktoranalüüsi aluseks sobib Pearsoni või astakutel põhinev Spearmanni korrelatsioonikordaja. Kui andmetes on erindeid, soovitatakse faktoranalüüsi mudeli tegemisel kasutada Spearmanni korrelatsioonikordajaid.

Andmete kohta võib teha erinevaid eeldusi, alates sellest, et iga üksik tunnus on normaaljaotusega. Range eeldus on mitmemõõtmeline normaaljaotus. See eeldus nõutav vaid siis, kui tahame kontrollida hüpoteese. Praktiliste ülesannete korral on see eeldus harva täidetud, sest faktoranalüüsi teostame pigem järjestustunnuste korral ja seega ei saa kogu tunnuste vektor olla mitmemõõtmelise normaaljaotusega. Tavaline nõue on, et tegemist on arvtunnustega või järjestustunnustega, eeldamata midagi nende jaotuse kohta. Kindlasti ei sobi faktoranalüüsi mudelisse nominaaltunnused.

10.1 Faktoranalüüsi matemaatiline mudel

Vaatame m tunnust, $X_1, \ldots, X_m$, tunnusvektor $\mathbf{X} = (X_1, \ldots, X_m)^T$. Eeldame andmete standardiseeritud kuju: $EX_i = 0; \quad DX_i = 1; \; i = 1, \ldots, m$ $E\mathbf{X}\mathbf{X}^T = \mathbf{R}$ (korrelatsioonimaatriks). Oletame, et lähtetunnused avalduvad teatud faktorite lineaarkombinatsioonidena.

Faktoranalüüsi mudeli kuju:

$$\mathbf{X} = \mathbf{A}F + U$$

$\mathbf{A} = \{a_{ij}\}, \; i = 1, \ldots, m; \; j = 1, \ldots, k$ on faktorlaadungite/faktorkaalude maatriks (*factor loading*), $F = (F_1, \ldots, F_k)^T$ on faktorite vektor, kusjuures faktorite arv peaks olema palju väiksem esialgsete tunnuste arvust ($k << m$). Faktorid on standardiseeritud ja sõltumatud: $EF_j = 0; \quad DF_j = 1; \; EF_jF_s = 0, \; j, s = 1, \ldots, k; \; j \neq s.$

$U = (U_1, \ldots, U_m)^T$ on omapärade vektor. Omapärad on tsentreeritud, sõltumatud omavahel ja sõltumatud faktoritest: $EU_i = 0; \quad DU_i = d_i^2; \quad EU_iU_v = 0; \quad EU_iF_j = 0, \; i, v = 1, \ldots, m; \; i \neq v; \; j = 1, \ldots, k.$ Omapära on see osa tunnuse hajuvusest, mis jääb faktorite poolt kirjeldamata, mudeli juhuslik viga. Omapärade dispersioonid rahuldavad nõuet $0 \leq d_i^2 \leq 1$.

Faktoranalüüsi olemus

Kasutades toodud eeldusi, saab lähtudes esitatud faktormudelist jõuda algtunnuste korrelatsioonimaatriksi uue esitusviisini

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{D},$$

kus $\mathbf{D}$ on omapärade dispersioonide diagonaalmaatriks $\mathbf{D} = diag(d_1^2, \ldots, d_k^2)$. Sisuliselt püütakse leida selline maatriks $\mathbf{A}$, et tunnustevahelised seosed oleks kõige paremini esitatud (otsitakse parimat esitusviisi seoste struktuurile). Defineeritakse nn redutseeritud korrelatsioonimaatriks

$$\bar{\mathbf{R}} = \mathbf{R} - \mathbf{D} = \mathbf{A}\mathbf{A}^T.$$

Redutseeritud korrelatsioonimaatriksi peadiagonaalil on kommunaliteedid (*communality*)

$$h_i^2 = 1 - d_i^2.$$

Kommunaliteedid näitavad seda osa tunnuse hajuvusest, mis on kirjeldatud faktormudeli poolt ehk tunnuse hajuvuse süstemaatilist osa ja on määratud iga tunnuse jaoks.

Seega jaguneb tunnuse dispersioon kaheks: süstemaatiline osa (faktorite poolt kirjeldatud h_i^2) + omapära (kirjeldamata osa d_i^2).

Faktoranalüüsi etapid

Vaatame andmemaatriksit (valim) $\mathbf{X} = \{x_{ij}\}, i = 1, \ldots, n;\ j = 1, \ldots, m$. Lihtsuse mõttes eeldatakse, et tunnused on standardiseeritud. Valimi põhjal hinnatakse faktormaatriks $\mathbf{A}$, tulemusi kasutatakse lähtetunnuste struktuuri kirjeldamiseks. Faktorid ehk uued (latentsed) tunnused $F_1, \ldots, F_k$ arvutatakse iga indiviidi jaoks ja seega saame maatriksi $\mathbf{F} = \{f_{ig}\}, i = 1, \ldots, n;\ g = 1, \ldots, k$. Faktorite väärtused igal indiviidil f_{ig} on nn individuaalsed faktorkaalud (*factor scores*).

Etapid:

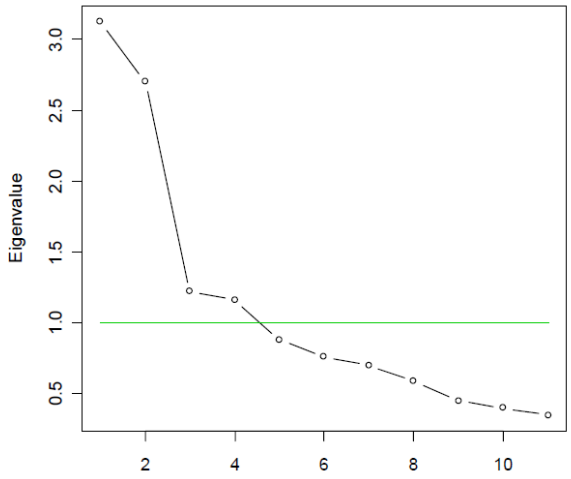
- Määrata faktorite arv k;
- Hinnata $\mathbf{A} = \{a_{ij}\}$ (faktorkaalud);
- Hinnata mudeli headus;
- Interpreteerida faktoreid (selleks tihti tuleb faktoreid enne pöörata).

Faktorite arvu määramine

Faktorite arvu määramiseks kasutatakse mitmesuguseid kriteeriume

- Omaväärtuste kriteerium (*eigenvalue criterion*) ehk Kaiser'i reegel. Faktorite arvu määrab ühest suuremate korrelatsioonimaatriksi omaväärtuste arv (vaikimisi arvutipakettides). See on loomulik reegel, sest ühega võrdumise korral kirjeldavad faktorid samapalju kui tunnused ja see mudel ei oma mõtet.
  Lisaks kasutatakse mitmesuguseid täpsustavaid reegleid, millal Kaiser'i reegel on sobiv (näiteks kasutatakse kui $n < 30$ ja kommunaliteedid > 0.7).
- Nõlvakudiagramm (*scree test*), joonistatkse graafik, kus y-teljel on omaväärtused, x-teljel faktori number. Hinnatakse joonisel suuri vahesid, faktorite arv enne suuremat vahet loetakse mõistlikuks.
- Interpretatsiooni kriteerium, hinnatakse tagantjärgi, kas valitud faktorite arv on sobilik, erinevad empiirilised kriteeriumid: igas faktoris peaks olema vähemalt 3 tunnust, faktoritel peaks olema lihtne struktuur jms.

Nõlvakudiagramm (*Scree plot*)



Nõlvakudiagrammi järgi oleks valik 2 faktorit, Kaiser'i reegli järgi 4

10.2 Faktoranalüüsi mudeli headus ja interpretatsioon

Kommunaliteet näitab tunnuse kirjeldatust faktorite poolt, $h_i^2 = 1 - d_i^2$, kus d_i^2 on omapära (kirjeldamata osa), $0 \leq h_i^2 \leq 1$. Kommunaliteet arvutatakse kui faktormaatriksi elementide ruutude summa piki rida (üle faktorite)

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2.$$

Arvutipaketid võivad väljastada $h_i^2 > 1$ st on tegemist väga halva mudeliga, kas on palju/vähe faktoreid või vähe andmeid.

Faktori kirjeldusvõime ehk kirjeldusmäära arvutamiseks summeeritakse faktormaatriksi elementide ruudud piki veergu (üle tunnuste)

$$g_j^2 = \sum_{i=1}^{m} a_{ij}^2.$$

Saab näidata, et kirjeldusvõime on redutseeritud korrelatsioonimaatriksi omaväärtus $g_j^2 = \lambda_j$, siit järeldub ka Kaiser'i reegel.

Faktori suhteline kirjeldusvõime, esitatakse tavaliselt protsentides (*Percent of Variation*) g_j^2/m.

Faktorite summaarne kirjeldusvõime on kõigi faktorite kirjeldusvõime kokku

$$g^2 = \sum_{j=1}^{k} g_j^2 = \sum_{i=1}^{m} h_i^2.$$

Faktoranalüüsi tulemuste interpretatsioon põhineb asjaolul, et faktormaatriksi element on korrelatsioonikordaja tunnuse ja faktori vahel

$$a_{ij} = r(X_i, F_j).$$

Reegel: *tunnus kuulub sellesse faktorisse, millega tal on maksimaalne korrelatsioon.*

Loomulik algus on selline, et vaadatakse läbi faktormaatriksi kõik read ja märgitakse ära igas reas kõige suurem (st iga tunnuse jaoks märgitakse, kuhu faktorisse ta kuulub). Seejärel vaadatakse faktoreid ükshaaval ja analüüsitakse tunnuseid, mis sinna kuuluvad, mis on neis ühist, mida nad kõik koos kirjeldavad jne, et leida faktorile sisu ja nimi. Tihti on esimene faktor teatav 'üldfaktor', mis kirjeldab kogu tunnustehulga mingeid ühiseid jooni. Faktorite interpretatsioon vajab probleemi sügavat tundmist, et tulemused ei oleks triviaalsed.

10.3 Faktormudeli hindamismeetodid

Faktormudeli hindamine tähendab maatriksite **F** ja **A** leidmist. On olemas terve rida hindamismeetodeid, neist tuntuimad:

- Peakomponentide meetod (tavaliselt vaikimisi) (*principal components*)
  Teostatakse teatav koordinaatteisendus lähtetunnuste ruumis, leitakse uued tunnused nn peakomponendid, kus esimene telg on tunnuste maksimaalse hajuvuse suunas ja järgmine sellega risti, edasi vaadatakse jääkhajuvust.
- Klassikaline faktoranalüüs
  Eelnevalt peavad olema teada kommunaliteedid, protsess on iteratiivne.
- Maksimaalse tõepära meetod (ranged eeldused, nõutakse tunnuste mitmemõõtmelist normaaljaotust). Selle meetodiga hinnatud faktormudeli korral saab kontrollida hüpoteese.
- Alfa-faktoranalüüs
  Eeldatakse, et valim on nii objektide kui ka tunnuste hulgast.
- Kaalutud/kaalumata faktoranalüüs.

Paketis SAS on võimalik kasutada 9 erinevat hindamismeetodit.

Faktorite pööramine

Esialgne faktormaatriks on tavaliselt halvasti interpreteeritav. Parema interpretatsiooni huvides pööratakse faktoreid, st teisendatakse esialgset faktormaatriksit

$$\mathbf{A}^* = \mathbf{AT},$$

kus $\mathbf{A}^*$ on pööratud faktormaatriks, $\mathbf{T}$ on teisendusmaatriks. Tavaliselt nõutakse, et $\mathbf{TT}^T = \mathbf{I}$, st teostatakse ortogonaalne pööramine.

Tuntuim ortogonaalse pööramise meetod on VARIMAX.

On võimalik teostada ka kaldpööramine (*oblique*), sel juhul $\mathbf{TT}^T \neq \mathbf{I}$. Kaldpööramise korral ei kehti enam seos, et faktormaatriksi element a_{ij} on korrelatsioonikordaja. Faktorite ja tunnuste korrelatsioonimaatriks tuleb eraldi leida, seda nimetatakse struktuurimaatriksiks. Tuntuim kaldpööramise meetod on Proscrustese pööramine. Kui kaldpööramisel saadakse faktorid, mis on sõltumatud siis järelikult ortogonaalne pööramine on õigem.

Üldiselt on pööratud faktormaatriksit lihtsam interpreteerida. Kui valitakse pööre nii, et faktormaatriksi elemendid oleksid võimalikult kas 0 või 1 lähedal (VARIMAX), siis on selge, et nullilähedasi võib interpreteerimisel eirata ja teisi tuleb kindlasti arvestada (interpreteerimine lihtsustub). Iga tunnuse kirjeldatuse tase (kommunaliteet) jääb pööramise tulemusena muutumatuks.

Ortogonaalselt pööratud faktormaatriksi elemendid on samuti korrelatsioonikordajad tunnuse ja faktori vahel. Ortogonaalselt pööratud faktormaatriksi puhul jääb faktorite summaarne kirjeldatuse tase samaks, esimese faktori kirjeldatus kindlasti väheneb, viimaste faktorite kirjeldatuse tase üldiselt suureneb.

10.4 Individuaalsed faktorkaalud

Faktoranalüüsi tulemusena saadakse uued tunnused st faktorid $F_1, \ldots, F_k$. Faktorid avaldatakse lähtetunnuste kaudu, faktorlaadungite maatriks seob lähtetunnused faktoritega $X_i \xrightarrow{\mathbf{A}} F_j$. Seega saab leida iga objekti/indiviidi jaoks uute tunnuste (faktorite) väärtused f_{ig} $(i = 1, \ldots, n;\ g = 1, \ldots, k)$, mida nimetatakse individuaalseteks faktorkaaludeks.

Individuaalsed faktorkaalud on standardnormaaljaotusega $N(0, 1)$, mistõttu neid on neid suuruse järgi lihtne interpreteerida

- keskmisi faktori väärtusi kirjeldavad individuaalsed faktorkaalud vahemikus $(-1; 1)$;
- arvestatavad hälbed negatiivses või positiivses suunas on vahemikes $(-2; -1), (1; 2)$;

- harvemini esineb väärtuseid üle 2 (või alla -2).

Näide. Oletame, et faktor on mingi isiksuse omadus (näiteks loovus) ja interpretatsiooni järgi on selge, mida suurem väärtus seda suurem loovus. Vaatame kahte isikut, faktori väärtustega $f_{Tiit} = 2.5$ ja $f_{Peep} = -1.5$.

Mida võime öelda nende isikute loovuse kohta?

Faktoranalüüsi näited

Näide 1.

Uuriti 220 meesüliõpilast, uuritavateks tunnusteks oli 6 eksamitulemust (prantsuse ja inglise keel, ajalugu, aritmeetika, algebra ja geomeetria). Soovitakse teada, millised üliõpilaste omadused määravad nende edukuse erinevates ainetes.

Esialgne hinnatud faktormaatriks

Tunnus	Faktor 1	Faktor 2	Kommunaliteedid
Prantsuse keel	0.553	0.429	0.490
Inglise keel	0.568	0.288	0.406
Ajalugu	0.392	0.450	0.356
Aritmeetika	0.740	-0.273	0.623
Algebra	0.742	-0.211	0.569
Geomeetria	0.595	-0.132	0.372

Faktormaatriks pärast pööramist

Tunnus	Faktor 1	Faktor 2
Prantsuse keel	0.369	0.594
Inglise keel	0.433	0.467
Ajalugu	0.211	0.558
Aritmeetika	0.789	0.001
Algebra	0.752	0.054
Geomeetria	0.604	0.083

Kuidas interpreteerite tulemusi?

Näide 2. Faktoranalüüs paketiga SAS

Uuritud on 160 olümpia kümnevõistlejat (alates II Maailmasõjast). Kümnevõistlejate kohta on kogutud kümnevõistluse iga ala punktid.

Tunnused on tähistatud järgmiselt:

V1 - 100 m jooks, V2 - kaugushüpe
V3 - kuulitõuge, V4 - kõrgushüpe
V5 - 400 m jooks, V6 - 100 m tõkkejooks
V7 - kettaheide, V8 - teivashüpe
V9 - odavise, V10 - 1500 m jooks

Ülesande lahendab järgmine programm:

```
proc factor data=sport method=prin   rotate=varimax ;
var v1-v10;
run; quit;
```

Valikud:
method=prin faktormaatriks hinnatakse peakomponentide meetodil
rotate=varimax pööratakse kasutades Varimax meetodit

Hinnati 3 faktoriga mudel. Pööratud faktormaatriks on järgmine

```
               Rotated Factor Pattern
       Factor1       Factor2       Factor3
v1     >0.87784       0.07145      -0.13731 100m jooks
v2     >0.77807       0.26635       0.12753  kaugus
v3      0.30920      >0.82130      -0.10258  kuul
v4     >0.50533       0.34840       0.33807  kõrgus
v5     >0.73822      -0.06424       0.38160 400m jooks
v6     >0.62798       0.32032       0.05057 100m tõkke
v7      0.22112      >0.78520      -0.06346  ketas
v8      0.24724       0.39623      >0.49824  teivas
v9     -0.02151      >0.70476       0.14939  oda
v10     0.02201      -0.12204      >0.89139 1500m jooks
     Variance Explained by Each Factor
Factor1    Factor2     Factor3
 2.777      2.263       1.377  <--- iga faktori kirjeldusvõime
 Final Communality Estimates: Total = 6.418 <--- kogu kirjeldusvõime
 v1    v2     v3     v4     v5     v6     v7      v8   v9     v10
0.794  0.692  0.780 0.491  0.694  0.499  0.669  0.466  0.519  0.809
                          <---  v1 - v10 on iga tunnuse kommunaliteedid
```

Interpreteerida tulemust!

10.5 Selgitav/kirjeldav ja kinnitav faktoranalüüs

Siiani vaatasime selgitavat ehk kirjeldavat faktoranalüüsi. Selgitav faktoranalüüs (*explanatory*) on hüpoteeside genereerimise meetod, hüpoteese ei kontrollita, erinevatel andmestikel saadud mudelid pole võrreldavad.

Kinnitav faktoranalüüs (*confirmatory*) võimaldab leida etteantud struktuuriga faktormaatrikseid, võrrelda sama struktuuriga mudeleid.

Paketis SAS teostab selgitavat faktoranalüüsi protseduur FACTOR, kinnitavat faktoranalüüsi protseduurid CALIS ja TCALIS (viimane on uuem ja nõuab mudeli kirjeldamisel vähem informatsiooni).

Kinnitava faktoranalüüsi korral hinnatakse, kas antud faktormudel sobib andmetega, väljastatakse terve rida statistikuid ja erinevaid sobivuse indekseid.

Mõned näited sobivuse kriteeriumitest:

- χ^2-kriteerium (mudel sobib, kui $p > 0.05$),
- suhe χ^2/df (mudel sobib, kui suhe < 2)
- suhtelise sobivuse indeks CFI (*Comparative Fit Index*) (peaks olema > 0.9, mida lähema ühele, seda parem mudel).

Soovitusi faktoranalüüsi tegemise juurde

1. Kui suur peaks olema valim?

   Tuntuim on nn 10-ne reegel: vähemalt 10 vaatlust iga tunnuse kohta

2. Kui palju peab olema tunnuseid?

   Kinnitavas faktoranalüüsis pole kitsendusi tunnuste arvu kohta (struktuurivõrrandite kontekstis 2-3 tunnust faktori kohta). Suurem tunnuste arv tõstab valiidsust.

   Kirjeldavas faktoranalüüsis peaks olema vähemalt 3 tunnust faktori kohta, kusjuures 'mida rohkem, seda parem' ei pruugi kehtida (võib tekkida nn *suboptimaalsus*).

   Antakse ka empiirilised reeglid, mitu tunnust faktori kohta peaks olema: 2 on miinimum, 3 on parem, 4 on kindlam, 5 on rohkem kui vaja.

Faktoranalüüs ja teised mitmemõõtmelised meetodid

Faktoranalüüs ja Cronbachi α
Faktoranalüüsi mudeli korral teostatakse mudeli sobivuse kontrollimiseks tihti ka reliaabluse analüüs, st arvutatakse iga faktori korral reliaabluse kordaja Cronbachi α hindamaks faktorisse kuuluvate tunnuste kooskõla. Kõrged kordaja väärtused (> 0.8) näitavad, et faktorisse kuuluvad tunnused on hästi kooskõlas.

Faktoranalüüs ja regressioonimudel
Üks võimalus faktoranalüüsi edasiseks rakendamiseks on saadud faktorite kaasamine regressioonanalüüsi mudelisse. Käsitletakse faktoreid kui uusi tunnuseid ja leitakse tavalisel viisil regressioonimudel.

Korrespondentsanalüüs
Faktoranalüüs ei sobinud nominaaltunnustele, aga kui tahame analüüsi kaasta ka nominaalseid tunnuseid, siis sobiv mudel on korrespondentsanalüüsi mudel. Korrespondentsanalüüsi võib vaadata kui faktoranalüüsile analoogilist analüüsimeetodit, mida rakendatakse nominaaltunnustele. Korrespondentsanalüüs lähtub sagedustabelist ja põhineb χ^2-statistikul.

Struktuurivõrrandite mudelid (*SEM – Structural Equation Modeling*)
Struktuurivõrrandite mudelid on põhjuslike seoste testimiseks andmetel st saame oma teooriat kontrollida (mudeliga ei saa otsida põhjuslikke seoseid), latentsed tunnused pannakse regressioonimudelisse.

Kinnitav faktoranalüüs on struktuurivõrrandite mudeli erijuht.

Kirjandus

1. Ehasalu, E., Tiit, E.-M. (1993). Faktoranalüüs ja kanooniline analüüs SAS-süsteemis. Käsiraamat üliõpilastele II, Tartu.

2. Käärik, E. (1995). Kordusmõõtmistest. ESS Teabevihik N 5, lk 33-38, Tartu.

3. Käärik, E., Jakoreva, I. (1997). Valiidsusest ja reliaablusest. ESS Teabevihik N 9, lk 42-46, Tartu.

4. Parring, A.-M., Vähi, M., Käärik, E. (1997). Statistilise andmetöötluse algõpetus. Tartu.

5. Myers, R. H. (1990). Classical and modern regression with applications, Duxbury Press.

6. Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A. (1998). Applied regression analysis and other multivariable methods, Duxbury Press.