# Katseandmete analüüs

Loeng 4. Kahe pideva muutuja vaheline sõltuvus - korrelatsioonid ja I tüüpi regressioon

Kirjutas Toomas Tammaru veebruaris 2002, natuke sätistatud 2003-2014, konsultant Ants Kaasik

**Korrelatsioonid**

ANOVA puhul rääkisime olukorrast, kus miski pideva muutuja väärtus sõltus miski kategoorilise muutuja väärtusest, nüüd räägime kahe pideva muutuja vahelisest seosest. Olgu meil igast objektist mõõdetud kahe erineva pideva muutuja väärtused, näiteks pikkus ka kaal. **Korrelatsioon** kahe **pideva** muutuja vahel tähendab seda, et ühe muutuja suurematele väärtustele vastavad sagedamini teise muutuja suuremad (positiivne korrelatsioon) või siis vastavalt väiksemad (negatiivne korrelatsioon) väärtused, korrelatsiooni tugevust mõõdab korrelatsioonikordaja (r). Kui ülalkirjeldatud seost üldse ei ole, on korrelatsioonikordaja väärtus null; maksimaalne on +1 ja minimaalne -1. Samamoodi nagu ANOVA puhul, võib korrelatsiooni analüüsida **parameetriliste** ja **mitteparameetriliste** meetoditega.

 Parameetriline ehk **Pearsoni** korrelatsioon eeldab, et valim on pärit kahemõõtmelisest normaaljaotusest, st mõlema muutuja jaotused on normaalsed. Peale mõlema jaotuse keskväärtuste ja standardhälvete iseloomustab sellist kahemõõtmelist jaotust niisiis veel ka (Pearsoni) korrelatsioonikordaja r, mis võib saada väärtusi miinus ühest pluss üheni.

Korrelatsioonikordaja r arvutatakse vastavalt valemile

 , kus sx ja sy on siis kummagi muutuja standardhälbed ja sxy  **kovariatsioon,** mis arvutatakse



valemi tõlgendamisel paneme tähele, et iga selline vaatlus, kus mõlema muutuja keskmisest ühele poole hälbivad väärtused esinevad koos, lisab summale suure positiivse liidetava; vaatlus, kus hälbed on eri suundades, lisab suure negatiivse liidetava ja need vaatlused, kus vähemalt ühe muutuja väärtus on nulli lähedal, ei lisa suurt midagi. Valem valemiks, aga kovariatsiooni mõistest võiks mäletada, et see on nagu “toorkorrelatsioon”, mis on andmetest otse arvutatud, kuid mille arvväärtusest on raske midagi järeldada (sest see sõltub mh muutuja ühikust). Võimaks järeldada ja võrrelda korrelatsioone omavahel, ongi kovariatsioonist r arvutatud, st ühikust lahti saamiseks on mõlema muutuja standardhälvetega läbi jagatud.

Korrelatsioonanalüüsi läbi viies **testimegi** just korrelatsioonikordaja erinevust nullist. Selle r-i absoluutväärtus mõõdab siis korrelatsiooni tugevust, p mõõdab olulisust. Peale korrelatsioonikordaja absoluutsest suurusest sõltub olulisus mõistagi ka valimi suurusest.

 **Mitteparameetriline** korrelatsioon ei eelda midagi muutujate jaotuste kohta. Testitakse, kas kahe muutuja väärtuste **järjekorranumbrid** on omavahel rohkem seotud kui see puhtjuhuslikult juhtuda võiks. Mitteparameetrilist korrelatsiooni kasutame siis juhul, kui ilmsestigi pole tegemist kahemõõtmelise normaaljaotusega. Parameetrilise korrelatsiooni seisukohalt on eriti “pahad” üksikud väga kaugel olevad punktid, mis võivad kogu korrelatsiooni pea peale keerata, mitteparameetrilist korrelatsiooni üksikute punktide absoluutne kaugus ei mõjuta - sõltub ju kõik vaid järjekorranumbritest. Levinuim mitteparameetriline korrelatsioon on **Spearmani** (astak)korrelatsioon. Miks siis jällegi mitte igas olukorras kasutada mitteparameetrilist korrelatsiooni - aga tollepärast, et mitteparameetrilise korrelatsiooni võimsus on väiksem ja tegelikult on eri korrelatsioonide sisulises interpretatsioonis ikka ka vahe sees - et kas meid huvitavad järjekorranumbrid või huvitavad ka kaugused punktide vahel.

 Mitteparameetrilist korrelatsiooni võime kasutada ka järjestustunnuste puhul (vt loeng nr. 3), ehk siis saab küsida, kas koolipoistel on muusika ja füüsika hinded omavahel positiivses või negatiivses korrelatsioonis. “Hinne” pole pidev tunnus ja seega parameetrilisi meetodeid (Pearsoni korrelatsiooni) siin kasutada ei tohi.

 Filosoofilisemal tasemel võib erinevate korrelatsioonide olemasolus näha jällegi tõendit selle üldise asja kohta, et seda, mida väljendame ühe tavakeele lausega (antud juhul “on seotud” või “on korreleerunud”), võib matemaatiliselt väljendada mitmel erineval moel. Kui korrelatsioon on väga tugev, siis saab ta misiganes meetodiga paljastatud, piiripealsetel juhtudel sõltub tulemus matemaatilistest detailidest.

 Tulemuse kirjutame *“pikkuse ja laiuse vahel oli korrelatsioon (r= , N= , p= )”* või mitteparameetriliste puhul *“..... (rs= , N= , p= )”*. Vaikimisi eeldame ilmsestigi Pearsoni korrelatsiooni, seega pole seda vaja eraldi rõhutada.

 Pane tähele, et korrelatsioon on **sümmeetriline**, ta ei sõltu sellest, kumb muutuja võtta “esimeseks” ja kumb “teiseks”, ehk siis korrelatsioonanalüüsi puhul ei jaotata muutujaid sõltuvaks ja sõltumatuks muutujal.

 Ja veel pane tähele, et korrelatsioonikordaja on **ühikuta suurus**, ta ei muutu, kui muutujate ühikuid muudetakse.

**Regressioon**

 Kahe korreleerunud muutuja puhul võime tahta seost kvantitatiivselt iseloomustada ehk siis vastata küsimusele, milline **matemaatiline funktsioon** seost kõige paremini kirjeldab. Sageli huvitab meid eelkõige ühe muutuja väärtuse ennustamine teise muutuja väärtuse järgi, ehk siis kui pikkus on teada, siis milline on tõenäone kaalu väärtus.

 Ennustamise otstarbel kasutame I tüüpi (tüüp II-st järgnevates loengutes) ehk nö. tavalist **(lineaarset) regressiooni**. Selle idee on sobitada sirge - regressioonsirge - andmepilve nii, et x-teljega risti mõõdetud kauguste ruutude summa igast punktist sirgeni oleks minimaalne (vähimruutude meetod). Ehk siis nii, et **jääkide (*residual*)** ruutude summa oleks minimaalne. Miks just ruutude - no ilmselt eelkõige sellepärast, et edasi saaks minna dispersioonanalüüsist tuntud teed pidi. Sobitatud sirgel on siis kaks parameetrit nagu sirgel ikka: **algordinaat (*intercept*)** ja **tõus (*slope*, b).** Algordinaat näitab, kus sirge lõikab y-telge. Tõus on tõusunurga tangens, ehk näitab, kui püsti see sirge on. Tõus võib saada misiganes positiivseid ja negatiivseid (langeb siis see sirge) väärtusi, kui muutujate vahel seost ei ole, on tõus 0.

 **Sirge tõus näitab, kui mitme ühiku võrra muutub y-telje muutuja väärtus kui x-telje muutuja väärtus muutub ühe ühiku võrra**. Sirge tõus on 1 juhul, kui x muutumisel miski suuruse võrra muutub y muutuja sama suuruse **võrra**. Olgu siis veel kord öeldud, et seose puudumisel on sirge tõus null. Sirge tõus on **ühikuga suurus**, tema ühikuks on y-telje muutuja ühik jagatud x-telje muutuja ühikuga, näiteks siis mune kilogrammi kohta. Kui ühikut muuta, muutub ka tõus - grammi kohta on mune 1000 korda vähem ju kui kilogrammi kohta. Ühik tuleb seega alati ära mainida tulemuste esitamise puhul.

 See mainitud ANOVA teed pidi minek käib nii, et sirge tõusu erinemist nullist testitakse samamoodi F-statistiku abil, kus F=MSmodel/MSerror. Samamoodi, selline värk põhineb dispersiooni komponentideks lahutamisel ehk siis SStotal=SSmodel+SSerror, viimane on siis jääkide ruutude summa, SSmodel on ennustatud väärtuste ruutude summa. Ennustatud y väärtus loetakse siis miskile x väärtusele vastava regressioonsirge punkti järgi.

 Samamoodi nagu ANOVAs, SSmodel/SStotal=R2, mis siis kirjeldab, mitu % dispersioonist mudel ära kirjeldab (seletab) - ***accounts for ... % of variance***. Seega siis on regressioonis kaks erinevat seose tugevuse hindajat - b ja R-ruut, kumbki iseloomustab seose tugevuse eri aspekti, statistiline olulisus sõltub nii ühest kui teisest (ja muidugi ka valimi suurusest), kuid ei mõõda ega väljenda otse seose tugevust. Ka väga nõrk seos (nii b kui ka R-ruudu mõttes) võib piisava valimi korral oluline olla.

 Pane tähele, et sellise lihtsa regressiooni puhul annab tõusu nullist erinemise test täpselt sama tulemuse kui korrelatsioonikordaja nullist erinemise test ja R-ruut on selle sama (Pearsoni) korrelatsioonikordaja ruut. R2 on samamoodi ühikuta suurus ega sõltu muutujate ühikuist.

Pane veel hästi tähele, et **regressioon ei ole sümmeetriline** - sirge sõltub vägagi sellest, kumb muutuja on x-teljel (sõltumatu muutuja) ja kumb muutuja on y-teljel (sõltuv muutuja). Mida tugevam on korrelatsioon muutujate vahel, seda vähem need kahte moodi saadud sirged erinevad. Asi võib tunduda esmapilgul imelikuna, kuid kujuta ette olukorda, kus seos on väga nõrk - ühtepidi vaadatuna väga väikese tõusuga sirge (ühest muutujast on teise väärtuse ennustamisel vähe abi) oleks teistpidi vaadatuna väga suure tõusuga sirge, mis siis tähendaks, et teise muutuja väärtuse järgi esimese ennustamine oleks vägagi efektiivne - oleks vist naljakas, kui nii oleks. Tegelt on siis muidugi nii, et teistpidi tehtud regressiooni puhul hakkab regressioonsirge lähenema paralleelsusele hoopis teise teljega, st uue rõhtteljega ehk siis teljega, millel on sõltumatu muutuja.

**Tulemused esitame** “kaal sõltus pikkusest (b=... (**ühikuga**!), R2= ....., df=....., F= ..., p<0.001)” või siis vastavalt ka vähem statistikuid, kui vaid osa neist sisuliselt huvitab.

 Kui oluline, siis seose võib esitada võrrandi kujul

pikkus = 3.78\*laius + 45.6

Analüüsi tulemustes esitub peale mainitud suuruste ka veel tõusu standardviga, mis siis iseloomustab tõusu määramise täpsust.

Statistikaprogrammid väljastavad ka algordinaadi nullist erinemise testi. Algordinaadi nullist erinemise testimise mõte on järgmine - kui algordinaat on null, on seos **võrdeline** - ehk siis kui üks muutuja muutub miski arv **kordi**, siis teine muutub samapalju **kordi**. Kui algordinaat ei ole null, seos võrdeline ei ole, mis siis et on endiselt lineaarne. Kui vaatame vaid positiivseid väärtusi, x ja y on mõlemal vaid positiivseid väärtusi, siis järeldub positiivsest algordinaadist, et x muutudes k **korda** (mitte võrra!) muutub y vähem kui k korda negatiivsest algordinaadist oleks järeldus vastupidine (vt siiski juttu II tüüpi regressioonist allpool enne sellist tüüpi järelduste tegemist!).

 Olgu terminoloogia veelkord: nagu kategooriliste muutujatega ANOVA puhul, nii räägitakse ka regressioonanalüüsi puhul mudelist (see on siis see sirge, mida sinna andmetesse sobitatakse), ennustatud väärtusest (miskile x-muutujale vastav y-väärtus, mis on täpselt regressioonsirgel). Jääk on siis vaadeldud väärtus miinus ennustatud väärtus.

Regressioonanalüüsi **eeldused** on järgmised:

 - *jäägid* olgu normaaljaotusega (mitte enam algmuutujad ise nagu korrelatsioonis!);

 - jääkide dispersioon peab olema sõltumatu (ehk siis x-telje) muutuja väärtusest sõltumatu (ehk siis olukord on homoskedastiline). Vastasel korral on olukord muuthajuv ehk heteroskedastiline (***heteroscedastic).***

 - jäägid ei tohi ilmutada süstemaatilist sõltuvust sõltumatu muutuja väärtustest. Vastasel korral on põhjust kahelda selles, et just lineaarne regressioon on sobiv sõltuvust kirjeldama ja peaks mõtlema mittelineaarse regressiooni (või ka muutuja teisendamise) peale (vt hilisemates loengutes).

 Seejuures pole sõltumatu muutuja väärtuste jaotus oluline: x-teljel võivad asjad olla nagu parasjagu on.

 Kõigi nende hädade vastu võib (kuid ei pea!) aidata (sõltuva muutuja) sobiv teisendus, näiteks logaritmteisendusest võib olla abi, kui dispersioon kipub suurematel väärtustel suurem olema. Teisendatud muutujatega regressiooni tõlgendamisel tuleb muidugi hoolega tähele panna, ka võrrandisse tuleb teisendus sisse kirjutada, ehk siis nt uuritava asja *logaritm* sõltub lineaarselt pikkusest näiteks.

 Üks regressiooni variante on selline, kus sirge ‘sunnitakse’ läbima koordinaatide nullpunkti. Sellist võtet kasutame juhul, kui on kaalukaid sisulisi argumente, et sirge seda punkti läbima peab, ehk teisisõnu, kui me kindlalt teame, et seos on võrdeline. Nii on lugu näiteks siis, kui uurime seost katselapile langenud seemnete arvu ja tärganud taimekeste arvu vahel. Fikseeritud punktiks võib muidugi olla mistahes muu punkt peale koordinaatide alguspunkti, sellisel puhul tuleb uuritavate muutuja väärtustele liita sobivaid arve nii, et see eriline punkt liiguks nullpunkti.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* jutu lõpp \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*