**Tilastotieteen käsikirja, Lauri Nummenmaa, lyhyt referaatti**

**Data**

* kvalitatiiviset ja kvantitatiiviset muuttujat
  + kvalitatiiviset muuttujat voivat kuvata laatua tai ominaisuutta – luokka
  + kvantitatiiviset muuttujat:
    - diskreetit eli epäjatkuvat ja jatkuvat muuttujat

1. **Laatueroasteikko** **eli nominaaliasteikko** (kvalitatiiviset muuttujat, esim. sp, ammatti) – korvaaminen numeroilla käsittelyä varten
2. **Järjestysasteikko eli ordinaaliasteikko** (laatueroasteikko, jossa järjestys, esim. likert
3. **Välimatka-asteikko eli intervalliasteikko** (jatkuva muuttuja, etäisyys voidaan ilmoittaa esim. lämpötila Celscius, ei voida laskea suhteita, kun 0 puuttuu)
4. **Suhdeasteikko** ja sen erikoistapaus absoluuttinen asteikko (jatkuva muuttuja, lisäksi absoluuttinen 0-piste, jonka vuoksi keskinäiset suhteet voidaan määritellä)

**Datan käsittely**

* puuttuvat tiedot
  + jätetään pois tilastoyksiköt, joilla puuttuvia havaintoja
  + imputoidaan puuttuvat havainnot tilastoyksiköille (keskiarvo tai mediaani)
* uusien muuttujien luominen, esim summamuuttujat jne.
* Mahdolliset transformaatiot...

**Datan kuvaaminen**

* frekvenssit
* aritmeettinen keskiarvo
* mediaani
* varianssi – mitä suurempi varianssi, sitä kauempana keskiarvosta havainnot ovat (havainnon etäisyys keskiarvosta, jokaisen muuttujan ero keskiarvosta potenssiin kaksi summattuna ja jaettuna havaintojen määrällä)
* keskihajonta std deviation eli muuttujan etäisyys keskiarvosta (varianssin neliöjuuri, joka palauttaa varianssin samalle mittayksikölle kuin alkuperäinen muuttuja)
* vinous
* huipukkuus
* muuttujien sirontakuvat

1. **Lineaarinen regressio**

* valitaan y muuttuja, jota pyritään selittämään x-muuttujien avulla
* y = a + bx , jossa a= vakiotermi (nähdään, kun x=0) ja b=kulmakerroin
* jäännöstermit kuvaavat virhekomponenttia (residual) >jäännöstermien minimiointi
* Oletukset:
  + muuttujien välillä lineaarinen yhteys > miten tsekataan -> tee sirontamatriisi malliin tulevista muuttujista
  + Selittävien muuttujien välillä ei saa olla kolineaarisuutta ->
  + Normaalijakaumaoletus (joskin siitä voidaan tinkiä, jos otoskoko on hyvin suuri)

Normaalijakaumaoletuksen testauksen askeleet:

* + - 1. normaalijakaumatestaus Shapiro-Wilk tai Kolmogorov-Smirnov (melko herkkiä arvioimaan jakauman ei-normaaliksi, joten kannattaa edelleen jatkaa tarkastelua seuraaviin kohtiin)
      2. jakauman muodon visuaalinen tarkastelu kuvaajien avulla
      3. vinous- ja huipukkuuskertoimet (skew ja kurtosi)

Jos jakauma ei ole normaali, sitä voi korjata datan muunnoksella eli transformaatiolla

|  |  |  |
| --- | --- | --- |
| JAKAUMAN MUOTO | MUUNNOS | KAAVA |
| lievästi oikealle vino | neliöjuuri | sqrt(x) |
| keskinkertaisesti oikealle vino | logaritmi | log(x) |
| voimakkaasti oikealle vino | käänteisfunktio | 1/(x) |
| lievästi vasemmalle vino | peilaus ja neilöjuuri | sqrt(k-x) |
| keskinkertaisesti vasemmalle vino | peilaus ja logaritmi | log(k-x) |
| voimakkaasti vasemmalle vino | peilaus ja käänteisfunktio | 1/(k-x) |

* Otoskoko:
  + 20-50 havaintoa/selittävä muuttuja riittää periaatteessa mallin sovittamiseen
  + jos selittäviä muuttujia muutama, havaintoaineiston koko pitäisi olla vähintään 100, mielellään 200 havaintoa >> riski, yleistyminen

Mallin tarkastelu

* (merkitsevyystaso, p-arvo)

1. Mallin sopivuus

Mallin avulla pystytään selittämään tietty osuus vaihtelusta. Jäljelle jäävä osuus muodostuu jäännöstermeistä. Mallin sopivuus näiden kahden tekijän avulla käyttämällä F-suhdetta. Fishterin F-suhde mallin selittämän varianssin ja selittämättä jääneen varianssi (jäännösten) välille:

F = selitetty vaihtelu/selittämättä jäänyt vaihtelu

* + suuri F-suhde - > malli selittää hyvin ja pieni F-suhde -> malli selittää huonosti
  + jos p<0,05 eli tulos on tilastollisesti merkitsevä

1. Mallin selitysaste

Kuinka suuri osa selitettävän muuttujan vaihtelusta voidaan kuvata mallin avulla.

1. jos yksi selittävä muuttua: Selittävän ja selitettävän muuttujan välisen tulosmomenttikorrelaatiokertoimen neliö r2
2. jos useampi: selitettävän muuttujan ja selittävien muuttujien multippelikorrelaation neliö R2

vaiheluväli 0-1, 1 selittää kaiken

1. Selittäjien sopivuus

Regressiokertoimet ilmaisevat, kuika paljon kukinx-muuttuja selittää y-muuttujan vaihtelusta

positiivinen tai negatiivinen kerroin

* lisäksi kollineaarisuustoleranssit

1. Jäännöstermit

Jäännöstermit itseisarvoltaan mahd pieniä

Jäännöstermien jakauman täytyy olla satunnainen ja normaalijakautunut (ks. kirjan toka vika sivu lineaarinen regressio ja residuals kuvaajat)

1. **Logistinen regressio**

* vähemmän oletuksia kuin lineaarinen regressio
* muuttujien jakaumaan ei liity oletuksia
* muuttujien välisistä yhteyksistä ei oleteta mitään
* otoskoko keskimäärin huomattavasti suurempi kuin lineaarinen malli eli vähintään 40-100 havaintoa / selittävä muuttuja (n= 100 + 50i)
* selitettävä muuttuja on kaksi- tai useampiluokkainen
* jos muuttuja on useampiluokkainen tai jatkuva, se voidaan dikotomisoida eli muuttaa kaksiluokkaiseksi
* selittävät muuttujat voivat olla millaisia hyvänsä
* Huom! selittävien muuttujien välillä ei saa olla kolineaarisuutta (sama kuin lineaarinen regressio)

Malli ei perustu lineaarisin yhteyksiin, eteneminen on erilaista.

1. Mallin muodostaminen
2. Mallin sopivuus

Vasta kun malli sopiva malli on kyetty muodostamaan, mennään tarkastelussa eteenpäin.

Logistisessa regressiossa ennustetaan havaintojen jakautumista luokkiin (y)-

Jos luokkia on kaksi, toinen valitaan ennustekategoriaksi, toinen on referenssikategoria.

Fisherin F-suhde mallin sopivuuden arvioimiseksi. Varianssianalyysia ei voi käyttää.

X2-testi toimii.

* pienet X2-testisuureeseen liittyvät p-arvot tarkoittavat sopivaa mallia

X2-testiin perustuva sopivuuskerroin (Goodness of Fit eli GOF).

Näiden avulla voidaan tarkastella, kyetäänkö mallin avulla tuottamaan alkuperäinen havaittu datamatriisi. X2-testin avulla verrataan poikkeavatko havaittu datamatriisi ja mallilla tuotettu datamatriisi toisistaan. Tässä tapauksessa suuret X2-testisuureeseen liittyvät arvot tarkoittavat sopivaa mallia.

1. Selitysaste

Pseudo R2 -kerroin, kun mukana kategorisia muutujia. Huom! ei suoraan ilmaise mitään….?

1. Ennustetarkkuus

Arvioidaan, kuinka tarkasti malli jakaa havainnot todellisiin luokkiin.

1. Ennustetarkkuus kaikille luokille, paljonko kaikista oikein
2. Ennustetarkkuus erikseen jokaisen selitettävän muuttujan luokalle.

Ennusteiden hyvyys tarkastelussa ennusteen herkkyys ja tarkkuus. Nämä saadaan laskettua sekaannusmatriisista, joka kuvaa erilaisten luokittelun lopputulosten frekvenssejä.

>>>todellisten positiivisten tulosten osuus > herkkyys = todelliset positiiviset /(todelliset positiiviset + väärät negatiiviset)

>>>todellisten negatiivisten tulosten osuus > tarkkuus = todelllisten negatiiviset /(väärät positiiviset+ todelliset negatiiviset)

Vaihteluväli 0-1. Mitä lähempänä 1, sitä paremmin malli toimii.

1. Selittäjien merkitys (selitysasteeseen)

Jos johonkin muuttujaan liitettävä regressiokerroin (Likelihood ratio) poikkeaaa tilastostollisesti merkitsevästi nollasta, kyseinen selittäjä on tilastollisesti merkitsevä.

Jos siis ko.selittäjään liittyvä p-arvo on pienempi kuin .05, muuttujan voidaan katsoa sopivan malliin.

Regressiokertoimet(betat) ovat logistisessa regressiossa logiteja. Mutta yhteys sama, negatiivinen tai positiivinen.

Kun regressiokertoimet muutetaan Exp(b)-kertoimiksi, niiden mittayksiköksi tulee odds eli riski.

* ykköstä pienemmät Exp(b) -kertoimet tarkoittavat negatiivista ja suuremmat positiivista yhteyttä.
* ykkönen tarkoittaa, että yhteyttä ei ole
* ykköstä pienemmät Exp(b)-kertoimet tarkoittavat, että kyseisen muuttujan arvon kasvattaminen pienentää riskiä
* ykköstä suuremmat Exp(b)-kertoimet tarkoittavat, että kyseisen muuttujan arvon kasvattaminen kasvattaa riskiä

Luokitteluratkaisun tulosten luotettavuus

* Ristiinvalidointi uudella aineistolla
* K-ositettu ristiinvalidointi. Otos jaetaan K:hon yhtä suureen osaan. Luokittelija muodostetaan K-1 osion perusteella ja testataan jäljelle jäävillä niin kauan kuin niitä on. Ongelma on sopivan K-arvon määrittäminen.
* Leave-one-out-ristiinvalidointi. Sama menetelmä kuin K, mutta siinä aineisto jaetaan kaikkiin mahdollisiin n-1 suuruisiin osiin. Malli muodostetaan n-1 havainnolla ja testataan aina vuorotellen jäljelle jäävään havaintoon.