KLIINISEN BIOSTATISTIIKAN JATKOKURSSI

Sisällysluettelo:

Yleistä monimuuttujamenetelmistä

Monimuuttuja-analyysi: mistä siinä on kyse?

Muuttuja ”variable” on suure, jota käytetään havaintojen tekemiseen, mittaamiseen tai koeolosuhteiden määrittelemiseen. Sen saamat arvot voivat olla joko satunnaisia, kuvastaen biologista tai muuta vaihtelua yksilöstä toiseen tai ne voivat olla tutkijan määrittelemiä koodiarvoja. Kliinisissä tutkimusongelmissa tarkasteltavat satunnaismuuttujat riippuvat yleensä enemmän tai vähemmän toisistaan, eli niillä on keskenäistä yhteisvahtelua. Ongelmien ratkaisemisessa ei siten riitä, että muuttujia tarkastellaan yksitellen, vaan niiden ratkaisemiseksi tarvitaan monimuuttujamenetelmiä. Niiden avulla voidaan esim. löytää ne muuttujat, jotka parhaiten selittävät jossain tietyssä lopputulos- tai vastemuuttujassa tapahtuvia vaihteluita tai niiden avullla voidaan kätevästi kontrolloida sekoittavien tekijöiden vaikutusta.

Monimuuttujamenetelmä ”multivariate method” Tavallisessa kielenkäytössä nimitys viittaa menetelmiin, joilla analysoidaan kahden tai useamman muuttujan keskinäisiä riippuvuussuhteita samanaikaisesti. Useimmat tilastollista käsittelyä vaativat lääketieteelliset tutkimusongelmat ovat tässä mielessä luonteeltaan monimuuttujaisia ”multivariable”. Tarkan määritelmän mukaan monimuuttujamenetelmä on kuitenkin sellainen menetelmä, jonka avulla tutkitaan kahta tai useampaa lopputulos- tai vastemuuttujaa samanaikaisesti. Esim. monimuuttujavarianssianalyysi (MANOVA), erotteluanalyysi ja ryhmittelyanalyysi ovat tämän määritelmän mukaisia monimuuttujamenetelmiä. Regressiomenetelmätkin voivat olla tarkan määritelmän mukaisia monimuuttujamenetelmiä, mutta useimmiten lääketieteellisissä tutkimuksissa regressiomenetelmiä käytetään siten, että selitettäviä lopputulos- tai vastemuuttujia on mallissa vain yksi kerrallaan ja silloin tarkan määritelmän mukaisesti englanninkielisessä terminologiassa näihin menetelmiin saatetaan viitata käsitteellä ”multivariable method”. Useimmat monimuuttujamenetelmien oppikirjatkaan eivät pitäydy tarkassa määritelmässä, vaan tarkastelevat monimuuttujamenetelminä myös yhden selitettävän muuttujan menetelmiä edellyttäen, että selittäjämuuttujia on useita.

Monimuuttuja-asetelmissa ja -malleissa analysoitavat muuttujat voivat esiintyä joko tasavertaisesti (symmetrisesti) tai niiden välillä voi olla tietyntyyppinen rakenteellinen riippuvuussuhde, esim. ajallinen tai kausaalinen, jolloin toinen muuttuja edeltää toista. Tavallisin asetelma on kuitenkin se, että jonkin tietyn muuttujan (**vastemuuttuja**, **selitettävä muuttuja** “dependent variable”,) vaihteluita pyritään selittämään joukolla muita muuttujia (**selittävät muuttujat**, “independent variables”, **esim. ennustetekijä, hoitoryhmä, interventiotoimenpide**). Tällainen tilanne syntyy esim. tutkittaessa erilaisten kliinisten suureiden välisiä riippuvuussuhteita tai pyrittäessä selvittämään prognostisten tekijöiden ja/tai hoitojen vaikutusta potilaan ennusteeseen. Termiä riippumattomat muuttujat (”independent variables”) ei tulisi käyttää, koska nämä muuttujat eivät yleensä ole toisistaan riippumattomia.

Tutkimusongelmaa voidaan lähestyä yleensä monin tavoin eri menetelmiä ja malleja käyttäen. Mallilla tarkoitetaan yleensä havaitun ilmiön kuvaamiseen käytettyä matemaattista kaavaa tai kaavajoukkoa. Malli antaa aina yksinkertaistetun kuvan todellisuudesta ja kannattaa muistaa, että kaikki mallit ovat vääriä, mutta jotkut niistä ovat hyödyllisiä. Tarkoituksenmukaisimman menetelmän ja parhaimmin soveltuvan mallin valinta ei ole helppoa.

Tavallisimmin käytettyjä ovat **lineaarinen** tai sellaiseksi palautuva malli ja **epälineaariset mallit**. Näistä eniten kliinisissä tutkimuksissa käytettyjä ovat etenkin logistinen malli. Lisäksi eloonjäämisanalyysien ja regressiomallien yhdistelmiä, esim. Cox:in mallin käyttö on nykyisin yleistä, mm. syöpätutkimuksissa. Varianssi- ja kovarianssianalyysit voidaan formaalisti esittää ja ratkaista regressiomallien muodossa. Etenkin monimutkaisemmissa asetelmissa, kuten useampiulotteisten toistomittausten tai puuttuvien tietojen tapauksessa, menettelystä on selvä hyöty.

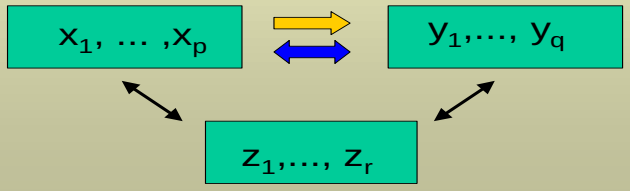
Monimuuttujamenetelmien käyttö kannattaa aina aloittaa tutkimusaineiston huolellisella kuvaamisella ja perusriippuvuuksien selvittelyllä. Mallien rakentaminen **vuorovaikutteisesti** ("interactively") on tehokas menetelmä, koska se pakottaa tutkijan miettimään kysymyksenasetteluaan. Malliin valittavien muuttujien määrä tulee suhteuttaa potilasaineiston kokoon. Erityisen tärkeää tämä on logistisessa mallissa, jossa estimoitavien parametrien määrä on useimmiten selvästi suurempi kuin muuttujien määrä. Tämä johtuu parametrien erilaisesta estimointitavasta verrattuna ns. **korrelatiivisiin** malleihin, esim. tavalliseen lineaariseen regressioon.

Mallin rakentaminen kannattaa yleensä aloittaa tutkimushypoteesin kannalta olennaisimmista muuttujista, ja muita tutkimushypoteesiin liittyviä muuttujia lisätään malliin tarpeen mukaan.

Monimuuttujamenetelmien luokittelu

Tarkastellaan lopputulos- tai vastemuuttujien **(y_1, …, y_q)**, q ≥1 riippuvuutta muuttujista **(x_1, …, x_p)**, p ≥ 1. Lisäksi **a)** halutaan mahdollisesti hallita sekoittavien tekijöiden **(z_1, …, z_r)**, r ≥ 0 vaikutusta ja **b)** y:t tai x:t tai z: t voivat olla ajasta t riippuvia.

Sekoittava tekijä, 'confounding factor'
Tekijä, joka häiritsee tutkittavan suureen (x) ja lopputulos- tai vaikutusmuuttujan (y) välisen yhteyden tutkimista. Jotta jokin suure olisi sekoittava tekijä (z), niin z:n täytyy **itsenäisesti assosioitua sekä x:ään että y:hyn.**



Monimuuttujamenetelmien valintaan vaikuttaa:

1. vastemuuttujien lukumäärä q
2. muuttujien y, x ja z asteikko; välimatka/suhde-, järjestys- tai laatuero
3. aikatekijä t

Joitakin analyysitilanteita

Seuraavissa taulukoissa on esitetty joitakin tavallisimpia tässä monisteessa tarkasteltavia analyysitilanteita ja niihin soveltuvia menetelmiä.

1. Oletetaan, että
- **a)** q = 1
- **b)** y satunnaismuuttuja välimatka-asteikolla
- **c)** x:t voivat olla myös ei satunnaisia ja kaksiarvoisia
- **d)** ei aikatekijää
- **e)** z:ja voi olla mukana

► **Menetelmä: Monimuuttujaregressioanalyysi**
”multivariate regression analysis” on menetelmä, missä pyritään selittämään muuttujassa (y) esiintyvää vaihtelua joukolla muita muuttujia (x_1, ..., x_p). Käytännössä tämä tapahtuu siten, että havaintoaineistoon sovitetaan malli, joka kuvaa mahdollisimman hyvin näiden muuttujien välisiä riippuvuussuhteita. Lääketieteellisissä sovelluksissa lineaaristen mallien ohella käytetään runsaasti myös epälineaarisia malleja esim. logistinen malli ja erilaiset annos-vaste-mallit.

Esim.

Sandvik L, Erikssen G, Thaulow E. **Long term effects of smoking on physical fitness and lung function: a longitudinal study of 1393 middle aged Norwegian men for seven years.** BMJ.1995 Sep 16; 311(7007):715-8.

Tavoitteena oli tutkia tupakoinnin vaikutusta fyysisen kunnon ja keuhkojen toiminnan heikentymiseen pitkäaikaisseurannassa. Seuranta-aika oli keskimäärin 7.2 vuotta. Tutkimuksen kohteena oli yhteensä 1393 40 – 59 vuotiasta miestä, jotka oli arvioitu terveiksi tutkimuksen alkaessa. Pääloppututulosmuuttujina olivat FEV1:n muutos ja ergometritestillä arvioitu fyysisen kunnon muutos seuranta-aikana. Selitettävät muuttujat (y) ovat siten välimatka-asteikollisia. Selittävinä muuttujina (x) oli tupakointi ja tupakointitapojen muutos. Sekoittavina tekijöinä (z) olivat mm. ikä ja y-muuttujien lähtötason arvot.

Tuloksina todettiin muun muassa., että fyysisen kunnon heikentyminen seuranta-aikana oli tupakoitsijoilla 13.6 % suurempaa kuin tupakoitsemattomilla (P<0.001). FEV1:ssä vastaava ero oli 6.0 % sekoittavien tekijöiden vakioinnin jälkeen.

Esim.

Forrester TE, Wilks RJ, Bennett FI, Simeon D, Osmond C, Allen M, Chung AP, Scott P. **Fetal growth and cardiovascular risk factors in Jamaican schoolchildren.** BMJ 1996 Jan 20; 312(7024)156-60

Tavoitteena oli tutkia kouluikäisten jamaikalaisten lasten verenpaineen, sokeroidun hemoglobiinitason ja kolesterolitason (y) yhteyttä syntymässä mitattuihin antropometrisiin mittareihin, kuten lapsen syntymäpainoon ja pituuteen (x). Sekoittavina tekijöinä (z) oli muun muassa äidin sosioekonominen status. Muuttujat y ja x ovat siten välimatka-asteikollisia ja z laatueroasteikollinen tai järjestysasteikollinen.

Tuloksina todettiin muun muassa, että lasten systolinen verenpaine oli kääntäen verrannollinen heidän syntymäpainoonsa (P<0,0001) ja suoraan verrannollinen heidän nykyiseen painoonsa. Sokeroitu hemoglobiinitaso oli korkeampi lapsilla,

joiden ihopoimun paksuus oli suuri (P<0,001) ja jotka olivat olleet lyhyitä syntyessään (P=0.003). Seerumin kolesterolitaso oli kääntäen verrannollinen lasten nykyiseen pituuteen (P=0,001) ja syntymäpituuteen (P=0,09) ja oli suoraan verrannollinen ihopoimun paksuuteen ja korkeampaan sosioekonomiseen statukseen (P<0.001).

2. Oletetaan, että
- **a)** q ≥ 1
- **b)** y satunnaismuuttuja
- **c)** x:t indikoivat vertailtavia ryhmiä

► **Menetelmä: Varianssianalyysi**

- Jos **q = 1**, niin kyseessä on yhden muuttujan varianssianalyysi ("**Anova**")
- Jos **q > 1**, ja y:t ovat saman muuttujan **toistoja** esim. eri ajankohtina, niin kyseessä on **toistomittausten varianssianalyysi** ("Anova with repeated measures")
- Jos **q > 1** ja y:t ovat eri muuttujia (niillä on eri sisältö), niin kyseessä on **monimuuttujavarianssianalyysi** ("**Manova**")
- Ryhmitteleviä / luokittelevia tekijöitä voi olla useita; varianssianalyysin suunnat/tasot
- Luokitukset voivat olla myös sisäkkäisiä ("nested analysis of variance")
- Jos mukana on myös z-muuttujia (kovariaatteja) niin kyseessä on **kovarianssianalyysi** ("**Ancova**", "**Mancova**")

Esim.

Imai K, Nakachi K.Cross sectional study of effects of drinking green tea on cardiovascular and liver diseases. BMJ. 1995 Mar 18;310(6981):693-6.

Tavoitteena oli tutkia vihreän teen käytön vaikutusta lipideihin ja maksaentsyymeihin japanilaisessa väestössä. Ajatuksena oli näyttää, että vihreän teen juonnilla voisi olla ennaltaehkäisevää vaikutusta sydän- ja verisuonisairauksiin ja maksasairauksiin. Tutkittavina oli 1371 yli 40 vuotiasta japanilaismiestä. Tutkimus tehtiin poikkileikkausasetelmaa käyttäen.

Tutkijat ilmoittivat käyttäneensä MANOVA:aa, mutta ilmeisesti kyseessä oli ANCOVA, eli kovarianssianalyysi, jossa vakioitiin mahdollisina sekoittavina tekijöinä ikä, tupakointi, alkoholin kulutus ja japanilaiseen väestöön suhteutettu painoindeksi. Teen juonti luokiteltiin kolmeen luokkaan: ≤3, 4-9 ja ≥ 10 teekupillista päivässä. Teen juonti oli analyyseissä ryhmittelevänä muuttujana.

Tutkijat totesivat päivittäisen teeannoksen ja kokonaiskolesterolin välillä tilastollisesti merkitsevän (P<0,001) trendin; mitä enemmän vihreää teetä joi, sitä alempi oli kolesterolitaso. Vastaavanlainen tulos saatiin triglyseriidin suhteen (P=0,02) ja samansuuntainen assosiaatio, joskaan ei tilastollisesti merkitsevä (P=0,06 ja P=0,07), myös maksaentsyymien (Asat, Alat) suhteen.

Koska kyseessä oli poikkileikkaustutkimus, mitään syy-seuraus-suhteita tuloksista ei kuitenkaan voi päätellä, huolimatta siitä, että tutkijat vakioivat analyyseissään ison joukon mahdollisia sekoittavia tekijöitä.

Laskennallisesti kaikki varianssianalyysin eri versiot voidaan kätevästi esittää ns. yleisenä lineaarisena mallina ("**G**eneral **L**inear **M**odel", GLM-proseruurit). Tällöin ryhmää

ilmaisevista laatueroasteikollisista muuttujista (x) rakennetaan ilmaisinmuuttujia, joiden koodaamisessa käytetään kahta eri tapaa joko **a) vaikutuksen mukaista koodausta** (**"effects coding"**) tai **b) vertailuryhmän mukaista koodausta** (**"reference cell coding"**).

Esim.

Oletetaan, että ryhmää indikoiva laatueroasteikollinen muuttuja (x) on neliluokkainen. GLM-malleissa muuttujasta x tehdään kolme joko **a)** (1,0,-1)- tai **b)** (0,1)-muuttujaa seuraavasti:

a)

Ryhmä (x)	x_1	x_2	x_3
1	1	0	0
2	0	1	0
3	0	0	1
4	-1	-1	-1

b)

Ryhmä (x)	x_1	x_2	x_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Muuttujilla x_1, x_2 ja x_3 on yhdessä sama tietosisältö kuin alkuperäisellä muuttujalla x.

Varianssianalyysin käyttö edellyttää teoriassa, että muuttuja noudattaa **likimain normaalijakaumaa** x:ien eri kombinaatioissa tai on muunnettavissa sellaiseksi. Mikäli näin ei ole, niin pitää edellä käsiteltyjä parametrittomia analyysimenetelmiä, Kruskal-Wallis, Friedman ym. Parametrittomien menetelmien puolella tilastopakettien tarjonta on hyvin rajoitettua. Esim. kahden tai useamman ryhmän toistomittausanalyysit eivät niillä onnistu, eivät myöskään kaksi tai useampisuuntaiset, eli monen ryhmittelevän tekijän varianssianalyysit. Tällöin tutkijan täytyy esim. erilaisia muunnoksia käyttäen yrittää tulla toimeen parametrisilla menetelmillä.

3. Oletetaan, että

- **a)** q >1, p > 1, r > 1
- **b)** muuttujat laatueroasteikollisia
- **c)** ei aikatekijää

► **Menetelmä: Kontingenssitaulukkoanalyysit**

- **testit, assosiaation mitat, ym.**
- **loglineaariset mallit**

4. Oletetaan, että

- **a)** q = 1
- **b)** y on dikotominen (0, 1)
- **c)** ei aikatekijää mukana
- **d)** tavallisesti z:ja mukana
- **e)** x:t voivat olla kaksiarvoisia ja/tai jatkuvia

► **Menetelmä: binäärinen logistinen malli**

Kun y on useampi luokkainen, niin menetelmä on **multinomiaalinen** (polytominen) logistinen malli tai **ordinaalinen** logistinen malli, jos y on moniluokkainen ja järjestysasteikollinen muuttuja.

Esim.

Roberts, Rosebud O. MD, MS; Jacobson, Debra J. MS; Girman, Cynthia J. DrPH; Rhodes, Thomas MS; Lieber, Michael M. MD; Jacobsen, Steven J. MD, PhD. **Population-Based Study of Daily Nonsteroidal Anti-inflammatory Drug Use and Prostate Cancer.** Mayo Clinic Proceedings 2002; 77:219-25

Tavoitteena oli tutkia ei-steroidialisten tulehduskipulääkkeiden (NSAID) päivittäisen käytön yhteyttä prostatasyöpään. Tutkittavat olivat 50 - 79 vuotiaita, jotka oli satunnaisesti valittu väestöotannalla Olmsted maakunnasta Minnesotassa vuonna 1990. Tutkittavia oli yhteensä 1362. Kyseessä oli seurantatutkimus ja seuranta-ajan mediaani oli 66 kuukautta ja maksimi 6 vuotta.

Lopputulosmuuttujana, päätetapahtumana (y) oli prostatasyövän ilmaantuminen seuranta-aikana. Tutkijat käyttivät sekä binaarista logistista regressiomallia että Coxin mallia.

Tuloksena todettiin, että 23/569 (4 %) särkylääkkeiden käyttäjistä ja 68/ 793 ei-käyttäjistä sairastui prostatasyöpään seurannan kuluessa. Ero on tilastollisesti merkitsevä (P=0,001). Riskisuhteen arvioksi tutkijat saivat OR=**0.45** (95 %:n luottamusväli: 0,28 -0,73). Riskisuhde oli riippuvainen iästä; 50 – 59 vuotiailla se oli 0,9, 60 - 69 vuotiailla 0,4 ja 70 -79 vuotiailla 0,2. Tutkijat ilmoittivat että sekä logistinen malli että Coxin malli antoi suunnilleen samat tulokset.

5. Oletetaan, että
- **a)** q = 1
- **b)** y on kaksiarvoinen (0,1) ja lisäksi on mukana aikatekijä, joka ilmaisee kuinka pitkä aika seurannan alusta tarkasteltavaan tapahtumaan (y=1) kuluu.
- **c)** x:t luokittelevia/ryhmitteleviä

► **Menetelmä: Kaplan-Meier–menetelmä**

6. Oletetaan, että
- **a)** q = 1
- **b)** y on kaksiarvoinen (0,1) ja lisäksi on mukana aikatekijä, joka ilmaisee kuinka pitkä aika seurannan alusta tarkasteltavaan tapahtumaan (y=1) kuluu.
- **c)** x:t luokittelevia/ryhmitteleviä
- **d)** aineisto ositettu, esim. jonkin sekoittavan tekijän hallitsemiseksi. z on ositetta ilmaiseva muuttuja

► **Menetelmä: Ositettu Kaplan- Meier-menetelmä**

7. Oletetaan, että
- **a)** q = 1
- **b)** aikatekijä mukana y:ssä, y kaksiarvoinen
- **c)** x:t luokittelevia/ryhmitteleviä
- **d)** z:t sekoittavia tekijöitä, joiden vaikutus halutaan hallita
- **e)** aikatekijä voi olla mukana z:ssa

► **Menetelmä: Coxin malli**

Esim.

The **HOPE** and HOPE-TOO Trial Investigators. **Effects of Long-term Vitamin E Supplementation on Cardiovascular Events and Cancer.** JAMA 2005; 293:1338-47.

Tavoitteena oli arvioida laskeeko pitkäkestoinen E-vitamiinilisuplementaatio syöpä- ja sepelvaltimotauti-ilmaantuvuutta ja kuolemia. Tutkimus oli asetelmaltaan satunnaistettu lumekontrolloitu kaksoissokkokoe. Satunnaistetut potilaat olivat vähintään 55-vuotiaita ja heillä oli todettu vaskulaarinen sairaus tai diabetes. Tulokset analysoitiin hoitoaikeen mukaista periaatetta käyttäen; ts. kaikki satunnaistetut osallistujat huomioitiin. Seuranta-ajan pituuden mediaani oli 7 vuotta. Tutkijat käyttivät Kaplan-Meier menetelmää ja log-rank-testiä. Osaryhmäanalyyseissä käytettiin Coxin mallia.

Erot pääloppututulosmuuttujissa eivät olleet merkitseviä ja Kaplan-Meier-käyrät olivat lähes yhteneviä, joten tutkijat päätyivät johtopäätökseen, että kyseisellä potilasjoukolla E-vitamiinisuplementaatiolla ei ollut vaikutusta syöpien eikä tärkeimpien sepelvaltimotautitapahtumien ennaltaehkäisyssä.

Viitteet

Imai K, Nakachi K. Cross sectional study of effects of drinking green tea on cardiovascular and liver diseases. BMJ. 1995 Mar 18;310(6981):693-6.

Forrester TE, Wilks RJ, Bennett FI, Simeon D, Osmond C, Allen M, Chung AP, Scott P. Fetal growth and cardiovascular risk factors in Jamaican schoolchildren**.** BMJ 1996 Jan 20; 312(7024)156-60

Roberts, Rosebud O. MD, MS; Jacobson, Debra J. MS; Girman, Cynthia J. DrPH; Rhodes, Thomas MS; Lieber, Michael M. MD; Jacobsen, Steven J. MD, PhD
A Population-Based Study of Daily Nonsteroidal Anti-inflammatory Drug Use and Prostate Cancer. Mayo Clinic Proceedings 2002; 77:219-25

Sandvik L, Erikssen G, Thaulow E. Long term effects of smoking on physical fitness and lung function: a longitudinal study of 1393 middle aged Norwegian men for seven years. BMJ.1995 Sep 16; 311(7007):715-8.

Kirjallisuutta

Afifi AA, **Clark** V, **May** S. Computer aided multivariate analysis. Chapman & Hall/CRC Texts, London, 2004, 4th edition. ISBN: 9781584883081. Hinta US$93,95. http://crcpress.com

Campbell MJ. Statistics at square two, Understanding modern statistical applications in medicine. BMJ Books 2001. ISBN: 0-7279-1394-8. Hinta €28. http://www.bmjbooks.com

Egret for Windows (1999) Software for the Analysis of Biomedical and Epidemiological Studies. CYTEL Software Corporation, MA, USA. http://www.cytel.com. Hinta US$ 398

Glantz SA, **Slinker** BK. Primer of applied regression and analysis of variance. McGraw-Hill, Inc., 1990. ISBN: 0-07-023407-8. Hinta £56,99

Harris EK, **Albert** A. Survivorship Analysis for Clinical Studies. Marcel Dekker, Inc., New York 1991. ISBN: 0-8247-8400-6

Hosmer DW, **Lemeshow** S. Applied logistic regression. John Wiley & Sons, New York, 2000. ISBN: 0-471-35632-8. Hinta €127,80

Hosmer DW, **Lemeshow** S. Applied survival analysis. Regression Modeling of Time to Event Data. John Wiley & Sons, New York, 1999. ISBN: 0-471-15410-5. Hinta € 107.80.

Kalbfleisch JD, **Prentice** RL. The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York, 2002. ISBN: 0-471-36357-X. Hinta €87,90.

Kleinbaum DG, **Kupper** LL, **Muller** KE. Applied regression analysis and other multivariable methods. 3rd Ed. Wadsworth, 1996. ISBN: 0534209106. Hinta £ 29.99

Kleinbaum DG. Logistic Regression, A Self-learning Text, 2nd edition. Springer-Verlag, New York, 2002. ISBN: 0-387-95397-3. Hinta €79,95

Kleinbaum DG. Survival Analysis, A Self-learning Text, 2nd edition. Springer-Verlag, New York, 2005. ISBN: 0-387-23918-9. Hinta €62,95

LogXact 6 for Windows (2005). CYTEL Software Corporation, MA, USA. http://www.cytel.com. Hinta US$494 ja yhdessä Egretin kanssa US$548

Milliken GA, **Johnson** DE. Analysis of messy data. Vol I. Chapman & Hall 1996. ISBN: 0412055414. Hinta £30.99.

Parmar MKB, **Machin** D, Survival Analysis: A practical approach. Chichester: John Wiley & Sons 1995, ISBN: 0 471 93640 5. Hinta US$99.95

Weerahandi, S. Generalized Inference in Repeated Measures, Exact methods in MANOVA and Mixed Models. John Wiley & Sons, New York, 2005, ISBN: 0-471-47017-1. Hinta €82,90.

Lineaarinen regressioanalyysi

Regression käsite

Regressioanalyysi nimitys on saanut alkunsa Francis Galtonin (1886) käyttämästä käsitteestä regressio. Tällä hän tarkoitti havaitsemaansa ilmiötä, että erittäin pitkien tai erittäin lyhyiden isien poikien aikuispituus on keskimäärin lähempänä väestön keskipituutta kuin isien pituus. Tätä ilmiötä kutsutaan **regressioksi kohti keskiarvoa.**

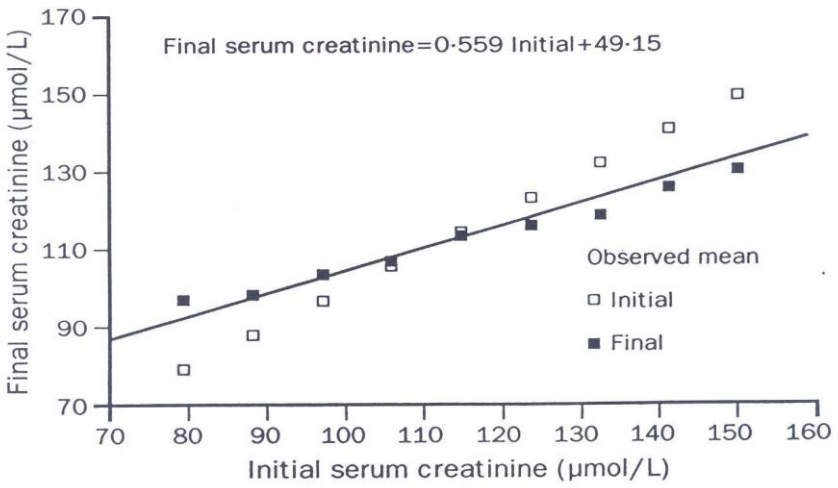
Regressio kohti keskiarvoa **"regression towards the mean"** on yleinen ilmiö seurantatutkimuksissa; Esim. jos verenpainetutkimuksissa tarkastellaan niiden henkilöiden osaryhmää, joilla on lähtötutkimuksessa korkea verenpaine, niin toistomittauskerroilla tämän ryhmän keskiarvo lähestyy perusjoukon keskiarvoa ilman mitään hoitovaikutustakin. Likimain pätee relaatio: (erotus 2. mittauskerralla) = r·(erotus 1. mittauskerralla), missä r on ensimmäisen ja toisen mittauskerran välisten mittausten korrelaatiokerroin. Yllä kuvattu ilmiö johtuu siitä, että $r < 1$ mittausvirheestä ja biologisesta vaihtelusta johtuen.

Erityisesti silloin, kun eri ajankohtina suoritetut mittaustulokset on muunnettu siten, että hajonta kunakin ajankohtana on sama, regressio kohti keskiarvoa on väistämätön ilmiö (Healy ja Goldstein 1978). Muun muassa monissa verenpaineen seurantatutkimuksissa on todettu, että jos tutkimusaineisto jaetaan seurannan alkaessa verenpainearvojen perusteella fraktiileihin (=sopiviin osiin) ja tarkastellaan jakauman ääripäitä, niin seuraavalla mittauskerralla ylimmässä ääripäässä verenpaine on keskimäärin laskenut ja alimmassa keskimäärin kohonnut. Regressio kohti keskiarvoa kuuluu väistämättä satunnaisvaihteluun ja se tulisi ottaa huomioon tuloksia analysoitaessa. Pääasiassa käsitettä regressio käytetään nykyisin kuitenkin regressiomalleista puhuttaessa.

Esim.

Madhavan S, Stockwell D, Cohen H, Alderman MH. Renal function during antihypertensive treatment. Lancet 1995 Mar 25; 345(8952):749-51.

Tavoitteena oli tutkia verenpainehoidon vaikusta munuaisten toimintaan. Munuaisten toiminnan mittarina käytettiin seerumin kreatiniinitasossa tapahtuneita muutoksia. Tutkittavina oli 2125 lievää ja keskivaikeaa verenpainetautia potevaa potilasta. Hoidon pituus oli keskimäärin vähän yli viisi vuotta. Tutkijoiden lähtökohtana oli se, että vaikka verenpaineen hoidolla tiedetään olevan vaikutusta sydänkohtausten ja aivohalvausten ehkäisyssä, niin vastaavaa ei näyttäisi tapahtuvan munuaisten toiminnanvajauden suhteen. Regressioanalyysiä käyttäen tutkijat päätyivät johtopäätökseen, että hoidon aikainen verenpainetaso ei assosioitunut itsenäisesti seerumin kreatiniinitasoon tutkimuksen lopussa. Tutkijat päättelivät, että munuaisten toiminnassa seurannan aikana tapahtunut muutos heijastelisi "regressio kohti keskiarvoa"- ilmiötä eikä tukisi sitä käsitystä, että verenpaineen hoito olisi tärkeä munuaisten toiminnanvajauden määre lievää tai keskivaikeaa verenpainetautia sairastavilla potilailla. Seerumin kreatiniinarvot laskivat keskimäärin jakauman ylimmässä ääripäässä ja kohosivat keskimäärin jakauman alimmassa ääripäässä tutkimuksen kuluessa.



Relation of initial to final serum creatinine

Yksinkertainen lineaarinen regressio

Tarkastellaan muuttujia x ja y, joista käytetään nimityksiä:

y = vastemuuttuja "dependent variable" on tutkimusmielenkiinnon kohteena oleva lopputulos- tai vastemuuttuja, jonka vaihtelua pyritään selvittämään.
x = selittävä muuttuja "independent variable" Esim. jokin interventiotoimenpide tai ennustetekijä), jonka avulla pyritään selittämään riippuvassa muuttujassa ilmenevää vaihtelua.

Ongelma: Kuinka paljon y:n keskiarvo muuttuu x:n muuttuessa yhden yksikön verran?

Tyyppi A:
x ei ole satunnaismuuttuja; x:n arvot ovat tutkijan valitsemia

Tyyppi B:
x on satunnaismuuttuja; havaintoparit (x_i, y_i) muodostavat otoksen kaksiulotteisesta Normaalijakaumasta.

Lineaarinen malli: $y = \beta_0 + \beta_1 x + \varepsilon,$

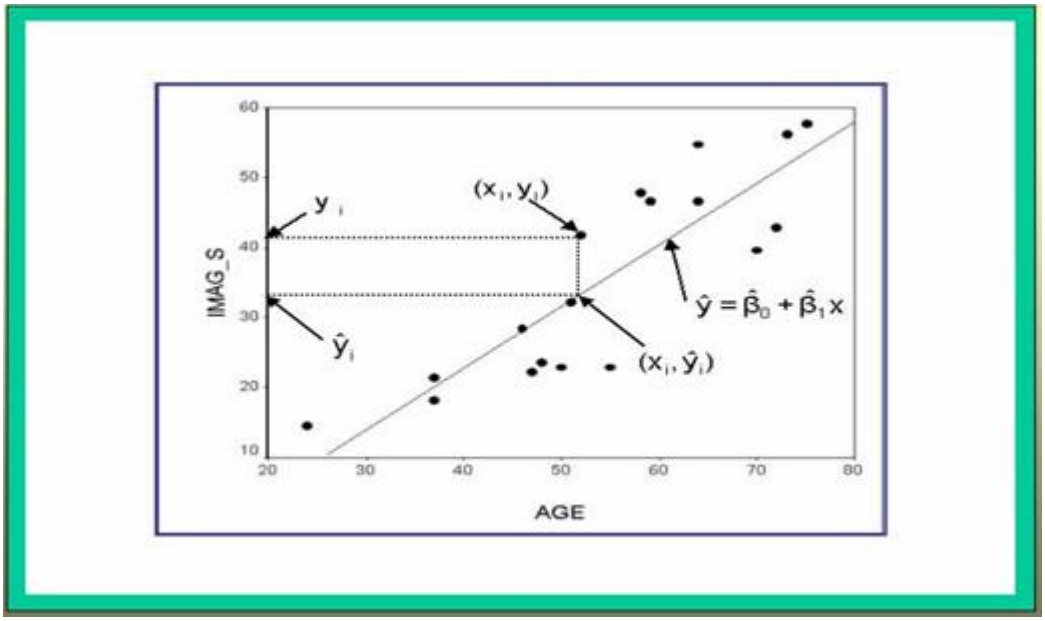
Yhtälössä β_0 on nk. vakiotermi, β_1 on regressiokerroin, joka ilmaisee kuinka paljon y keskimäärin muuttuu, kun x muuttuu yhden yksikön verran ja ε on virhevaihtelua edustava termi, jonka oletetaan noudattavan normaalijakaumaa.

Regressiosuoran kerrointen β_0 ja β_1 estimointi (arviointi) suoritetaan tavallisesti nk. pienimmän neliösumman menetelmällä minimoimalla **poikkeamaneliösumma:**

$$SS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Kaavassa **n** on havaintoparien **(x_i, y_i)** lukumäärä tutkimusaineistossa ja $\hat{y}_i$ on regressiosuoran perusteella arvioitu / ennustettu y_i:n arvo.

Kuva: Trial-tiedoston perusteella laskettu regressiosuora muuttujan Imag_S ja iän välisestä riippuvuudesta hoitoryhmässä R



Pienimmän neliösumman menetelmällä kerrointen **β_0** ja **β_1 arvioiksi** saadaan:

$$\beta_1 = b_1 = r_{xy} s_y / s_x \quad \text{ja} \quad \beta_0 = b_0 = \bar{y} - b_1 \bar{x}$$

r_{xy} on Pearsonin korrelaatiokerroin, **s_y** (=SD_y), **s_x** (=SD_x) ovat y:n ja x:n hajonnat, **$\bar{y}$ ja $\bar{x}$** niiden keskiarvot. Estimoidun regressiosuoran yhtälö on siten:
$y = b_0 + b_1 x$. Sijoittamalla tähän yhtälöön x:n paikalle eri arvoja, x = **x_i**, niin saadaan niitä vastaavat regressiosuoran perusteella ennustetut y:n arvot y= **$\hat{y}_i$**. (ks. Kuva)

Regressiokerrointen **keskivirheet** saadaan kaavoista:

$$S\hat{E}(b_1) = \frac{s_{y|x}}{s_x \sqrt{n-1}}, \qquad S\hat{E}(b_0) = s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

$s_{y|x}$ on y:n ehdollinen hajonta ja sen neliö on y:n **ehdollinen varianssi**. Sitä nimitetään myös **residuaalivarianssiksi** ja merkitään **s_{res}.**

Muuttujan x arvoa x_0 vastaavan regressiosuoran perusteella arvioidun y:n keskimääräisen arvon **keskivirhe** on:

$$SE(\hat{y}(x_0)) = s_{y|x} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$$

Regressiokertoimen b_1 ja vakiotermin b_0 tilastollinen merkitsevyys voidaan testata seuraavasti:

$$z = b_1 / SE(b_1) \quad \text{ja} \quad z = b_0 / SE(b_0)$$

Tätä testiä kutsutaan **Waldin testiksi.** P-arvot saadaan Normaalijakaumasta.

100(1- α) %:n luottamusvälit regressiokertoimille ja koko regressiosuoralle saadaan kaavoista:

$$b_1 \pm t_{1-\alpha/2} SE(b_1), \quad b_0 \pm t_{1-\alpha/2} SE(b_0), \quad \hat{y}(x_0) \pm t_{1-\alpha/2} S\hat{E}(\hat{y}(x_0))$$

Sijoittamalla viimeiseen näistä kaavoista eri arvoja $x=x_0$ yli x:n koko vaihtelualueen saadaan laskettua regressiosuoran luottamusvälikäyrät (x, y)-koordinaatistoon. Näiden käyrien väliin jäävä alue on kapeimmillaan muuttujan x keskiarvon kohdalla ja levenee kohti pienempiä ja suurempia x:n arvoja. Nämä regressiosuoran luottamusvälit määrittelevät alueen, johon perusjoukon regressiosuora sisältyy **100(1- α) %:n** varmuudella.

On usein myös hyödyllistä laskea yksittäisten y:n arvojen ennustettavuutta kuvaava niin kutsuttu ennuste (toleranssi-) väli (Huom. kyseessä ei ole luottamusväli!)
Välin laskemiseksi tarvitaan ennustetun **y:n hajonta**. Se voidaan laskea kaavasta:

$$SD(\hat{y}(x_0)) = s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$$

Ennusteväli saadaan kaavasta:

$$\hat{y}(x_0) \pm t_{1-\alpha/2} SD(\hat{y}(x_0))$$

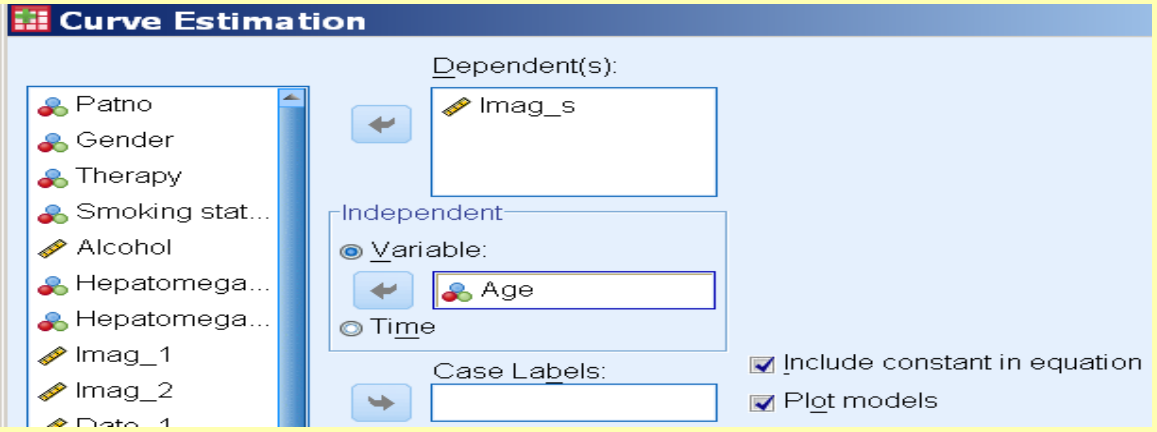
Ennustevälikäyrät voidaan piirtää (x, y)-koordinaatistoon menettelemällä samoin kuin regressiosuoran tapauksessa. Ennustevälikäyrien väliin jäävä alue on paljon laveampi kuin regressiosuorien väliin jäävä alue, koska yksittäisen arvon ennustaminen on aina epävarmempaa kuin keskimääräisen arvon.

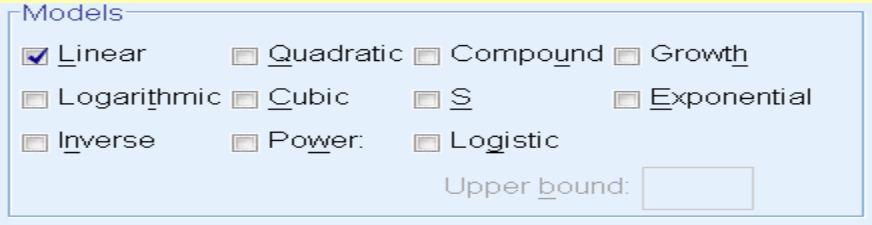
Useimmilla nykyisillä tilastopaketeilla molempien yllä kuvattujen käyrien piirtäminen samaan kuvaan regressiosuoran kanssa on mahdollista (esim. SPSS 15 ja erityisohjelma CIA).

Esim.

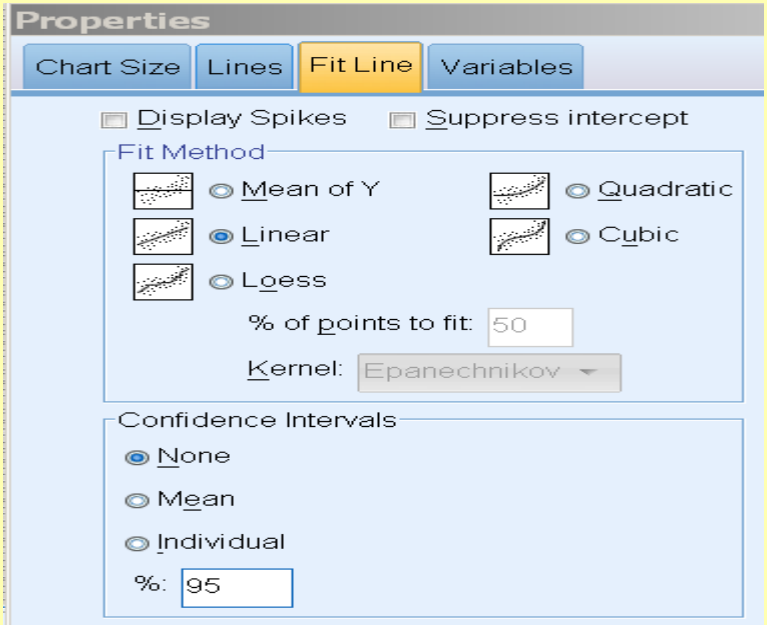
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav
Vastemuuttuja ”Imag_S” on hypoteettinen laboratoriomittaus seurannan alussa ja selittävä muuttuja on potillaan ikä seurannan alussa. Tarkoituksena on katsoa kuinka paljon ikä vaikuttaa suureeseen ”Imag_S”.

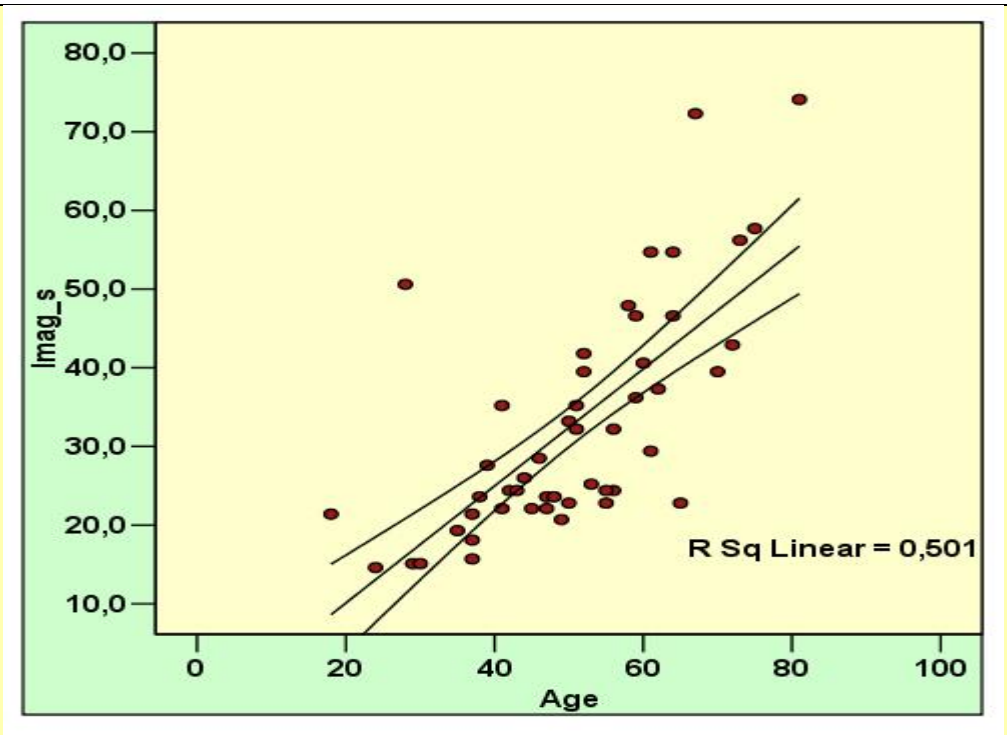
SPSS:llä kuvat saadaan valikoista: **”Analyze”►”Regression”►”Curve Estimation”**. Laitetaan kohtaan **”Dependent”** muuttuja **”Imag_S”** ja kohtaan **”Independent”** muuttuja **”Age”.** Malliksi valitaan **”Linear”** ja tuplaklikataan saatua kuvaa tulosteessa, jolloin se siirtyy editorille. Aktivoidaan käsiteltävät havaintopisteet klikkaamalla, jolloin pisteet näkyvät editorilla sinisenä.



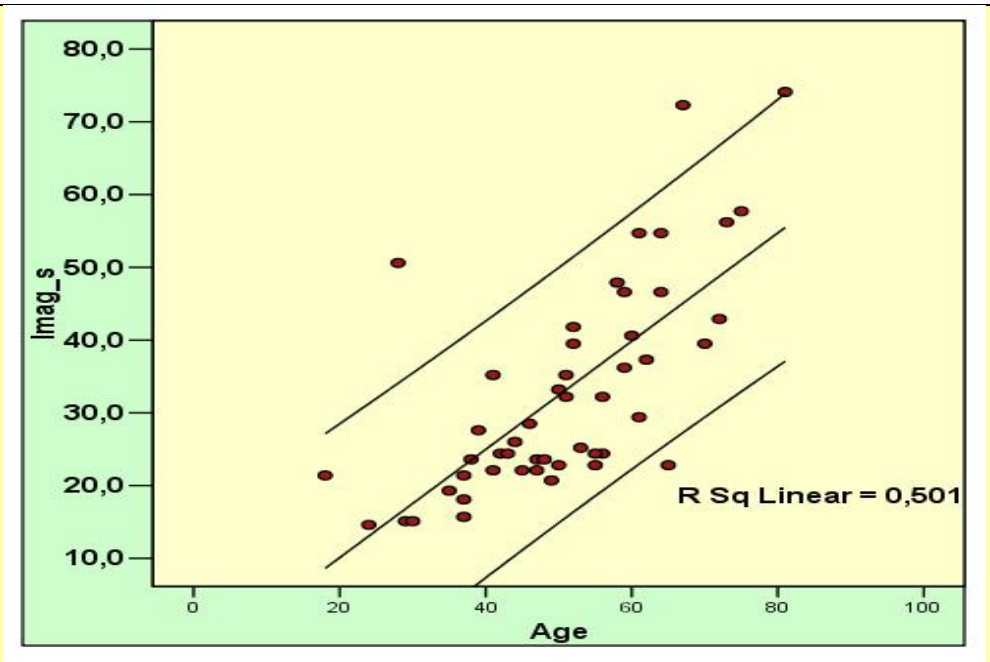


Kuvaeditorin kohdista **”Elements” ►”Fit Line at Total”** saadaan esiin valikko **”Properties”**, jossa voidaan määritellä minkä tyyppistä käyrää havaintopisteisiin halutaan sovittaa; esim. lineaarinen, neliöllinen ja kuutiollinen ja halutaanko luottamusvälit keskiarvolle vai ns. ennusteväli. Myös luottamustaso voidaan määritellä. Valitaan sekä ”Mean” että ”individual” ja luottamustaso 95 %.





Kuva: **Regressiokäyrän luottamusväli**



Kuva: **Regressiokäyrän ennusteväli**

Tulkinta:

- Regressiosuoran luottamusväli on kapeimmillaan muuttujan “Age” keskiarvon kohdalla, mikä tarkoittaa, että regressiokertoimen arviointi on sillä kohtaa luotettavimmillaan.
- Ennusteväli on paljon laveampi kuin regressiosuoran luottamusväli.

Regressiomallit ja niiden käyttö

Kirjallisuudessa nimitys regressioanalyysi viittaa useisiin erityyppisiin malleihin. Kaksi tavallisinta perusmallityyppiä ovat seuraavat:

malli A: klassinen regressiomalli

Tutkitaan satunnaismuuttujan y riippuvuutta ei-satunnaisista muuttujista $x_1, \ldots, x_k$. Muuttujiin x_i, i = 1, …, k ei liity mitään todennäköisyysjakaumaoletuksia. Niiden saamat arvot ovat tutkijan valinnan varassa.

Esim.

1) Annos-vaste tutkimukset, jolloin y on vastemuuttuja ja x:t ovat tutkimussuunnitelman mukaisia eri lääkeannoksia.

2) Säätelymalli. Kuinka y:n arvot muuttuvat keskimäärin, kun muutetaan joko yhden tai useamman selittäjämuuttujan (x) arvoja samanaikaisesti? Tällainen ongelma on kokeellisissa tutkimuksissa yleinen, mutta se ei yleensä sovellu epäkokeellisiin tutkimuksiin, kuten esim. sydän- ja verisuonitautien riskitekijätarkasteluihin.

malli B: monimuuttujaregressiomalli

Tämä malli eroaa edellisestä siten, että myös x_i:t ovat satunnaisia ja niiden oletetaan yhdessä noudattavan jotain jatkuvaa jakaumaa. Tavoitteena on rakentaa paras mahdollinen ennustaja millekä tahansa tekijälle x_i muiden tekijöiden x_j ($j \neq i$) perusteella.

Esim.

1) Tutkitaan x_1:n vaikutusta y:hyn, mutta tiedetään x_2:n, x_3:n jne. vaikuttavan myös y:hyn ja x_1:een. Esim. y = respiratorinen funktio (esim. FEV1), x_1 = altistusaika tutkittavalle altisteelle (esim. tupakointi), x_2 = ikä

2) Halutaan löytää niiden muuttujien joukko $x_1, \ldots, x_r$, joka parhaiten kuvastaa y:n vaihteluita. Esim. y = jokin nivelreuman vaikeusastetta kuvaava indeksi ja $x_1, \ldots, x_p$ joukko muuttujia, jotka assosioivat y:n kanssa ja keskenään

Mallien A ja B välillä vallitsee muodollinen samankaltaisuus. Rakenteellisesti regressioanalyysi on aivan sama molempien mallien pohjalta. Mallin tulkintaan liittyvät suureet ja testit ovat myös samoja, tarkasteltiinpa mallia A tai mallia B. Arvioinnin (estimoinnin) suhteen on joitakin eroavuuksia.

Yleinen regressiomalli voidaan esittää muodossa:

$$\mathbf{y = f(x_1, \ldots, x_p, \beta_0, \ldots, \beta_p) + \varepsilon}$$

y on selitettävä muuttuja, **f** on mallin muodon ilmaiseva funktio, $\mathbf{x_1, \ldots, x_p}$ ovat selittäjämuuttujia, $\boldsymbol{\beta_0, \ldots, \beta_p}$ ovat mallin parametrit (= regressiokertoimet) ja $\boldsymbol{\varepsilon}$ virhevaihtelua edustava satunnaissuure

Regressioanalyysin päätehtävänä on löytää mahdollisimman hyvin tutkimuksen kohteena olevaa ilmiötä kuvaava malli ja arvioida (estimoida) tämän mallin parametrit siten, että yhteensopivuus ilmiöstä kerätyn tutkimusaineiston kanssa on mahdollisimman

hyvä. Funktion f määräämisen kannalta regressiomallit voidaan jakaa kahteen päätyyppiin **lineaariset** tai sellaiseksi palautettavissa olevat tai **epälineaariset** mallit. Lineaarisessa mallissa parametrien vaikutus on lineaarinen, mutta selittävien muuttujien vaikutusten ei tarvitse olla lineaarisia.

Tavallisimmat kliinisissä sovelluksissa käytetyistä regressiomalleista ovat muotoa:

A) $f = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

B) $f = \beta_0 \cdot x_1^{\beta_1} \cdot x_2^{\beta_2} \cdots x_p^{\beta_p}$

C) $f = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p$

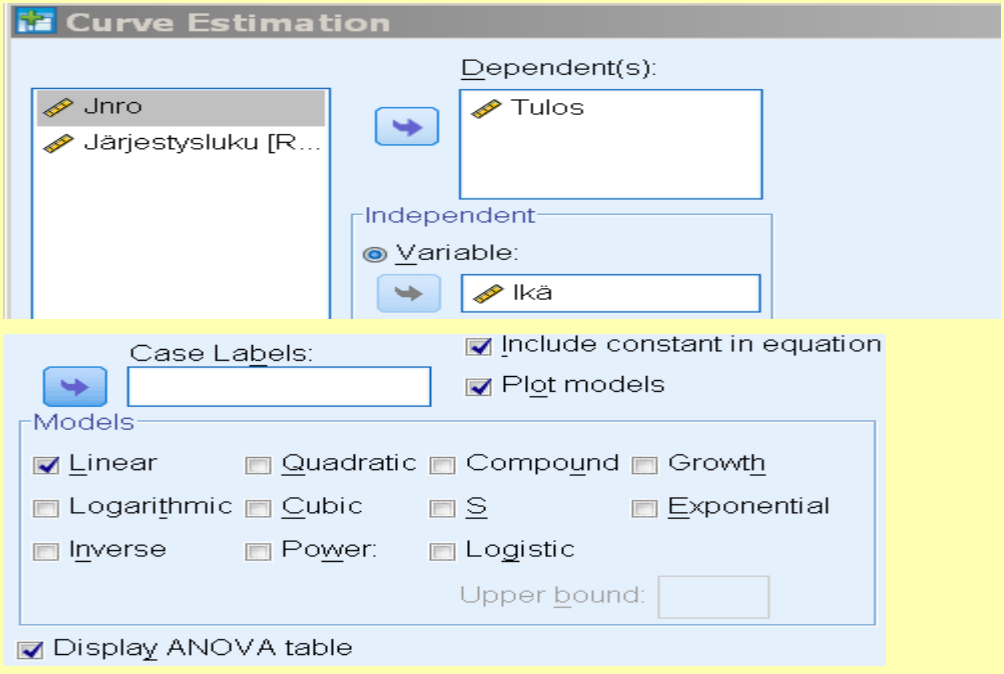
D) $f = 1/(1 + \exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)))$

Näistä A on tavallinen lineaarinen monimuuttujaregressiomalli, B on logaritmi-muunnoksella lineaariseksi palautuva malli: $\log f = \log \beta_0 + \sum \beta_i \log x_i$, C on polynomiregressiomalli ja D on logistinen regressiomalli:
$\log[f / (1 - f)] = \text{logit } f = \beta_0 + \sum \beta_i x_i$.

Käyränsovitus

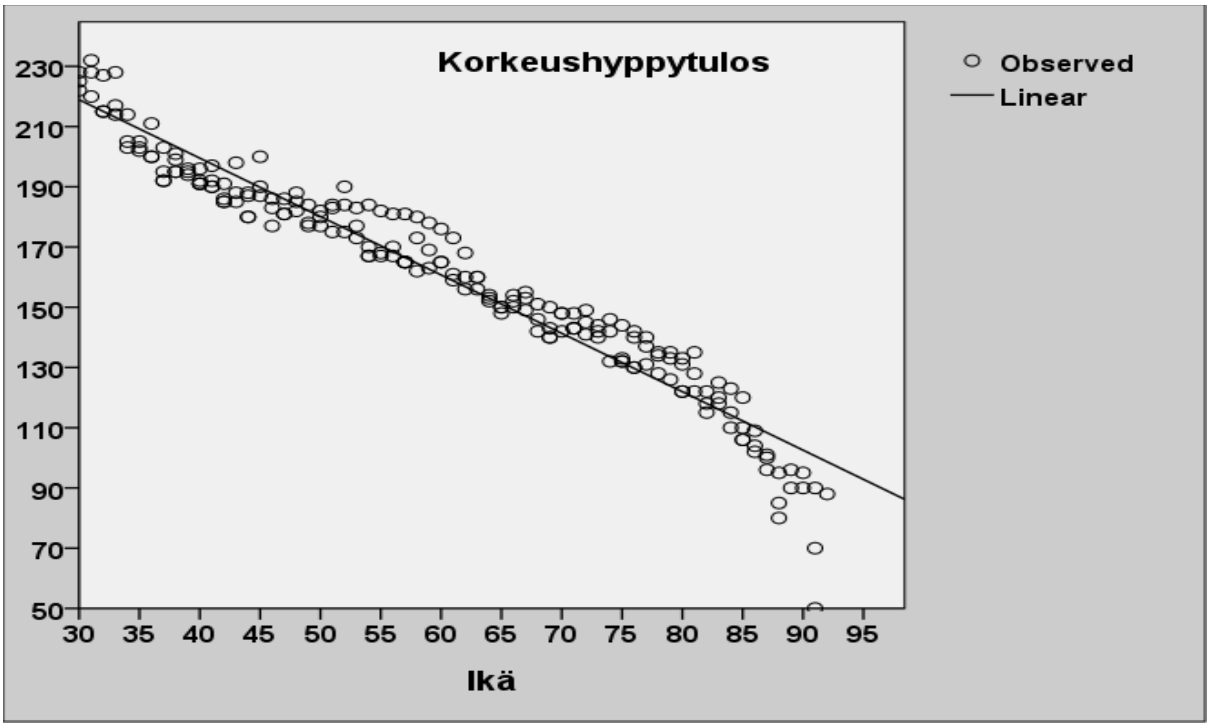
Käyränsovitus, ”curve fitting” on menetelmä, jonka avulla pyritään sovittamaan esim. (x, y) -koordinaatistossa kuvattuun havaintoaineistoon (esim. annos-vaste riippuvuus) jokin matemaattisesti määritelty käyrä, joka mahdollisimman hyvin sopii havaittuun pisteistöön. Yleisesti käyränsovitukseen käytetään ns. pienimmän neliösumman menetelmää ”least squares method”.

SPSS:llä käyrän sovitys tehdään valikosta: **”Analyze”►”Regression”►”Curve Estimation”**.

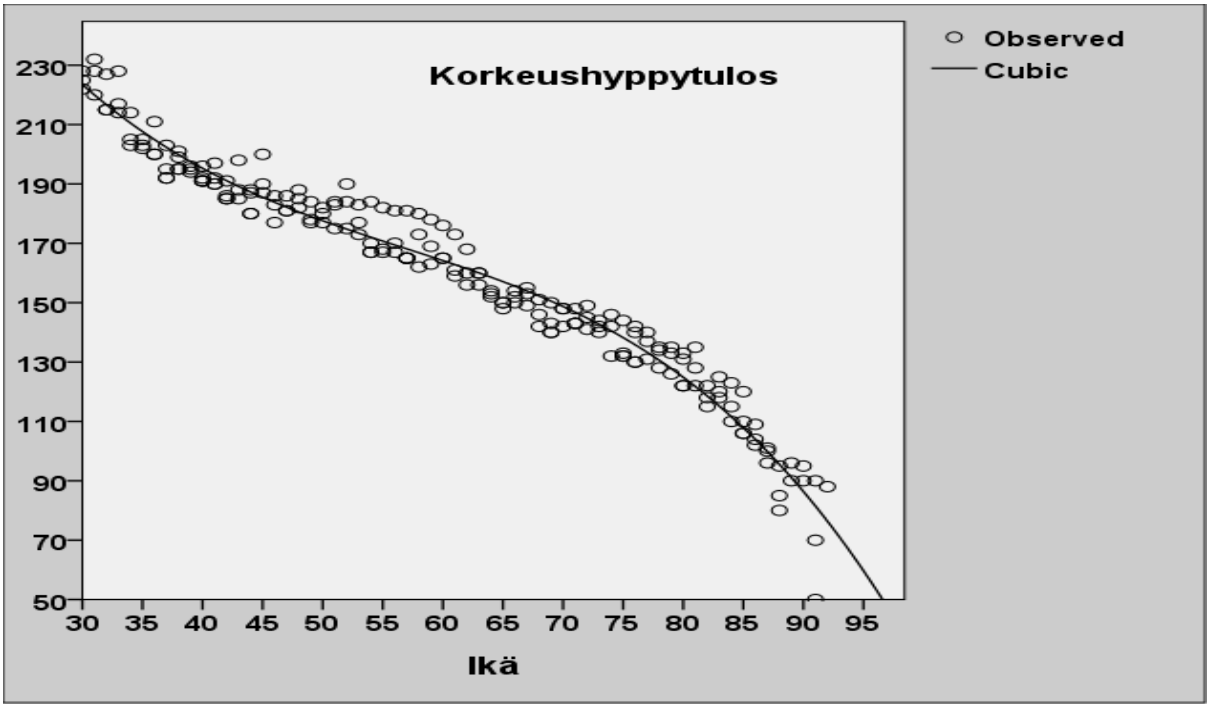


Esim.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Tulos.sav
Parhaat korkeushyppytulokset kautta aikojen yli 30 - vuotiailla Suomessa.
Aineistossa on kolme parasta tulosta kussakin ikäluokassa.



Malli: Tulos=277,07 - 1,94·(Ikä), R^2 = 0,942 (Lineaarinen regressio)



Malli: Tulos=434,4- 11,45· (Ikä)+ 0,179· $(ikä)^2$-0,001· $(Ikä)^3$, R^2=0,968, (Polynomiregressio, kuutiollinen)

Analysointitavat

Regressiomallien kuten muidenkin monimuuttujamallien analyysitapoja on useita. Tavallisimmin käytössä olevat tilasto-ohjelmistot (SPSS, Systat, SAS jne.) sallivat regressiomallien sovittamisen joko **kiinteänä** (täydellisenä) mallina, jolloin malliin tulee yhdellä kertaa kaikki tiettyyn tutkittavaan hypoteesiin liittyvät x-muuttujat, tai **askeltavana** ("stepwise") mallina jollaisella tutkija voi etsiä joko parasta mahdollista mallia tai pyrkiä valitsemaan vain kaikkein keskeisimmät muuttujat lopulliseen malliin.

Askeltavan mallin analysointitapoja on useita: **etenevä**, **takautuva** ja **"parhaan" yhtälön valinta**menettely.

Etenevässä menettelyssä malliin otetaan riippumattomista muuttujista **(x)** se, joka korreloi (Pearsonin tulomomenttikorrelaatiokerroin) parhaiten riippumattomaan muuttujaan **(y)**. Tämän jälkeen jäljellä olevista muuttujista valitaan se, joka lisää eniten mallin yhteiskorrelaatiokertomen neliötä ("R-squared"), kun huomioidaan mallissa jo oleva/olevat muuttujat. Näin jatketaan, kunnes selitysosuus ei enää oleellisesti muutu.

Takautuva menettelyssä malliin laitetaan aluksi kaikki ehdokkaana olevat x:t ja askel askeleelta pudotetaan huonoimmat muuttujat pois. On syytä muistaa, että nämä menettelyt saattavat johtaa keskenään eri malleihin.

Askeltavaa mallia voidaan myös käyttää siten, että malliin pakotetaan tietty muuttujajoukko, jonka tutkija haluaa ehdottomasti sisältyvän malliin ja sen jälkeen lopuista vaihtoehtoisista muuttujista valitaan askeltavasti parhaat.

Parhaan yhtälön mallille on luonteenomaista se, että malliin jo päässeet muuttujat eivät välttämättä siellä pysy. Ne saatetaan poistaa tarpeettomana jollain valintakierroksella, kun malliin on tullut mukaan sellaisia muuttujia, jotka sisältävät yhdessä olennaisesti saman informaation kuin jokin/jotkin jo malliin sisällytetty muuttuja/muuttujat.

Parhaan yhtälön valintatapoja ovat:

- **F-menetelmä**
- **R-menetelmä**
- **Vaihtamismenetelmä**
- **Kaikkien kombinaatioiden menetelmä**

Huom.

R-menetelmässä muuttujien sisäänotto malliin tapahtuu kuten F-menetelmässä, mutta poisto on erilainen: muuttuja poistetaan mallista, mikäli poiston seurauksena R^2-arvo kasvaa (Tämä on mahdollista muuttujien välisistä korrelaatioista johtuen!)

Epälineaarisia regressiomalleja voidaan pitää, logistista ja Coxin mallia lukuun ottamatta, siinä määrin erikoismenetelminä, että niiden käsitteleminen ei ole mahdollista tässä yhteydessä.

Toinen erikoismenetelmä on nk. **sidottu** regressioanalyysi, jolla tarkoitetaan sitä, että mallin parametreille b_0, …, b_k annetaan tiettyjä **side-ehtoja** (reunaehtoja). Sidotussa mallissa f on edelleenkin yleensä lineaarinen. Sidottuun malliin joudutaan silloin, kun halutaan regressiotason $y = b_0 + b_1x_1 + \ldots + b_px_p$ kulkevan määrättyjen pisteiden kautta. Tavallisin toivomus on, että halutaan pakottaa taso kulkemaan origon kautta eli että asetetaan vaatimus $b_0 = 0$.

Varianssi- ja kovarianssianalyysimallit voidaan muodollisesti esittää regressiomallin avulla. Kun regressioanalyysissä x:t ovat tavallisesti kvantitatiivisia (määrällisiä) suureita, niin varianssianalyysissä x:t ovat kiinteitä kokonaislukuja, esim. joko nollia tai ykkösiä, ilmaisten koetuloksiin vaikuttavien eri tekijöiden tai ryhmien olemassaoloa. Kovarianssianalyysissä esiintyy molempia em. muuttujatyyppejä; niistä käytetään tällöin nimeä **kovariaatit** (“covariates”). Esimerkiksi varianssianalyysi voidaan esittää seuraavassa muodossa:

$$y_i = x_{i0}b_0 + x_{i1}b_1 + ... + x_{ip}b_p + \varepsilon_i,\ i = 1, \ldots, n$$

jossa x_{ij}:t (j=0,1, …, p) ovat kiinteitä lukuja ilmaisten jonkin tietyn tekijän läsnäoloa tai puuttumista. Jos kyseessä on yksisuuntainen varianssianalyysi, missä tutkittavalla tekijällä on k tasoa, niin p = k.

Yleisesti r:n tekijän mallissa $\mathbf{p = \sum_{i=1}^{r} k_i}$, jossa k_i on tekijän i tasojen lukumäärä.

Vaikka varianssianalyysi voidaan muodollisesti esittää regressiomallina, varianssianalyysiä ei yleensä kannata suorittaa regressioanalyysiohjelmistolla, vaan käyttäen tähän tarkoitukseen laadittuja erityisohjelmia (esim. SPSS:ssä, Systatissa GLM, SAS:ssa GLM ja GENMOD).

Myös menetelmä nimeltä **erotteluanalyysi** (”Multiple Discriminant Analysis”) on kahden ryhmän, (esim. terveet ja sairaat) tapauksessa esitettävissä regressiomallina siten, että määritellään y muuttuja kaksiarvoiseksi (0 = terve, 1 = sairas). Tällöin regressiomallin antamat kertoimet ovat verrannollisia erotteluanalyysistä saatavien kerrointen kanssa, eli minkä tahansa kahden regressiokertoimen suhde $\mathbf{b_i^R/b_j^R}$ on sama kuin vastaavien erotteluanalyysin kerrointen suhde $\mathbf{b_i^E/b_j^E}$. Samoin kuin varianssi- ja kovarianssianalyysi kannattaa erotteluanalyysikin kuitenkin suorittaa erityisohjelmiston avulla (SPSS: **”Classify” ► ”Discriminant”**).

Erotteluanalyysin avulla voidaan tutkia esim. potilasryhmien välisiä eroja (tai erojen profiileita) tarkastelun kohteena olevan muuttujajoukon (x_1, …, x_p), p ≥ 1 suhteen ja kuinka uusia diagnosoitavia potilaita on mahdollista luokitella analyysissä muodostettavien erottelufunktioiden avulla.

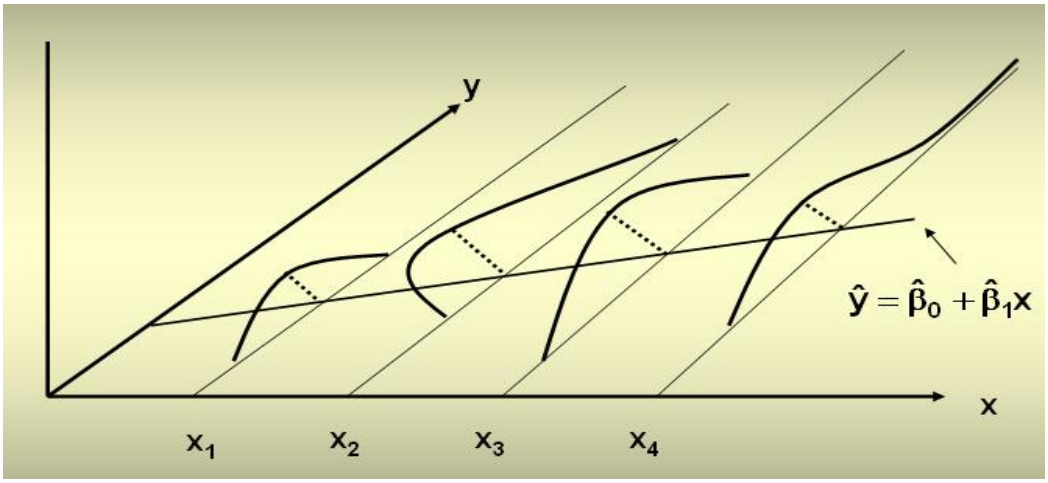
Mihin olettamuksiin mallit perustuvat?

Tarkastellaan mallia: $y = b_0 + b_1x_1 + \ldots + b_px_p + \varepsilon$, jossa b_0, b_1, …, b_p ovat mallin parametreja perusjoukossa ja ε on satunnaisvaihtelua edustava virhetermi. Regressiomallille asetetaan tavallisesti seuraavat perusolettamukset:

1. **Lineaarisuus.** y:n keskiarvo on x:ien lineaarinen funktio.
2. **Havaintojen riippumattomuus.** Kun tiedetään jonkin henkilön saama y:n arvo tutkimusaineistossa, se ei anna mitään tietoa jonkun toisen henkilön saamasta y:n arvosta.
3. **Homoskedastisuus (vakiovarianssius).** y:n varianssi (Var(ε)) on vakio mille tahansa muuttujien x_1, …, x_p yhdistelmälle.
4. **Normaalisuus.** Muuttujien x_1, …, x_p jokaisen yhdistelmän osalta y noudattaa normaalijakaumaa (Gaussin jakauma).
5. **Ei yhdysvaihtelua** (interaktiota) muuttujien x_i välillä. Mikä tarkoittaa, että minkä tahansa muuttujan x_i muutoksen vaikutus y:hyn on riippumaton muiden selittäjämuuttujien tasosta.

Olettamukset 1 ja 5 liittyvät siihen, kuinka hyvin käytetty malli sopii yhteen tutkijan havaintoaineiston kanssa. Muut olettamukset liittyvät varianssien estimointiin ja merkitsevyystesteihin.

Kuva. Olettamus 1 voimassa, mutta olettamukset 3 ja 4 eivät!



Residuaalien tulkinta

Mallin yhteensopivuutta voidaan parhaiten tutkia residuaalien (jäännösten) $\varepsilon_i = y_i - \hat{y}_i$ avulla, missä y_i on henkilön i havaittu ja $\hat{y}_i$ mallin perusteella ennustettu y:n arvo. Perusmenetelmänä jäännösten tutkimisessa on jakaa havaintoaineisto sopivasti x_i:ien tai niiden yhdistelmien mukaisiin osaryhmiin (esim. kvartiileihin)ja laskea residuaalien keskiarvot osaryhmissä. Jotta malli olisi mahdollisimman hyvin yhteensopiva havaintoaineiston kanssa, tulisi näiden residuaalien keskiarvojen olla lähelle nollaa kaikissa osaryhmissä.

Esim.

Regressiomalli, jossa on kaksi riippumatonta muuttujaa, x_1 ja x_2, ja aineisto on jaettu niiden perusteella kvartiileihin. Kussakin kvartiilissa on laskettu residuaalien keskiarvo.

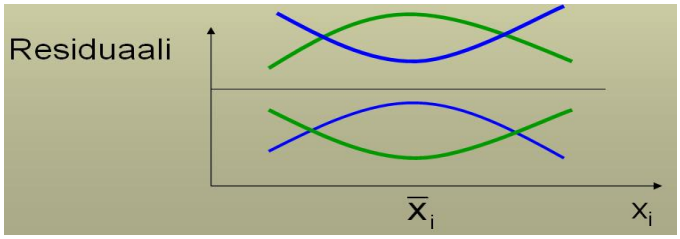
kvartiilit	x_1	x_2	Residuaalien keskiarvo
Alin	1	1	$\bar{\varepsilon}_{11}$
	1	2	$\bar{\varepsilon}_{12}$
	1	3	$\bar{\varepsilon}_{13}$
	1	4	$\bar{\varepsilon}_{14}$
	…..	…..	…..
Ylin	4	1	$\bar{\varepsilon}_{41}$
	4	2	$\bar{\varepsilon}_{42}$
	4	3	$\bar{\varepsilon}_{43}$
	4	4	$\bar{\varepsilon}_{44}$

Keskiarvojen $\bar{\varepsilon}_{ij}$ tulisi olla lähellä nollaa. Residuaalien keskiarvojen välisiä eroja kvartiilien tai niiden yhdistelmien mukaisissa osaryhmissä voidaan testata

varianssianalyysillä. Useimmat tilasto-ohjelmistot tarjoavat lisäksi tutkijalle monenlaisia keinoja residuaalien graafiseen tarkasteluun, josta on usein hyötyä mallissa esiintyvien puutteiden toteamiseksi. Tällöin saadaan esimerkiksi vastaus seuraavanlaisiin kysymyksiin:

Tuleeko muuttujan x_i vaikutus y:hyn riittävässä määrin edustetuksi lineaarisella termillä $b_i x_i$, vai pitäisikö malliin lisätä esimerkiksi neliöllinen termi $c_i x_i^2$?

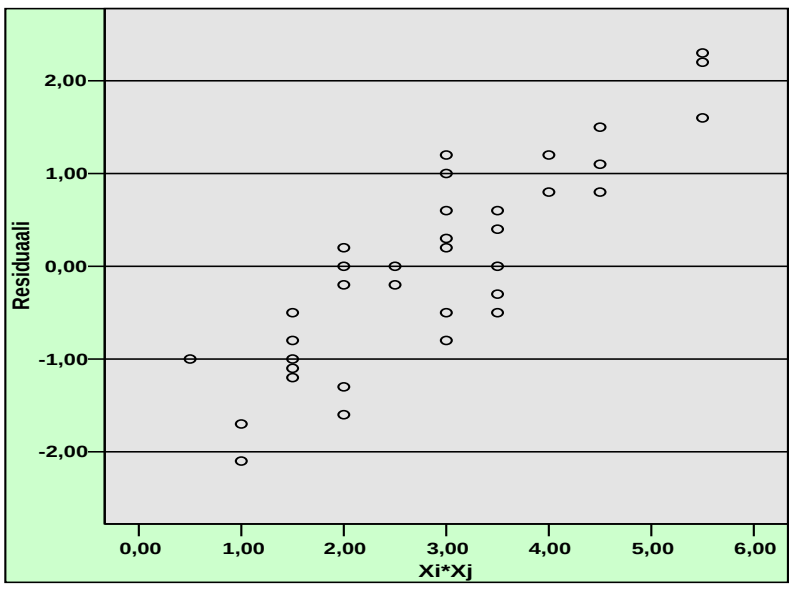
Tätä voidaan tutkia tarkastelemalla residuaaleja x_i:n funktiona. Mikäli residuaalit ovat suurimpia x_i keskiarvon kohdalla ja pienenevät lähestyttäessä x_i:n molempia ääripäitä tai päinvastoin, niin neliöllisen termin mukaanotto malliin on useimmiten aiheellista.



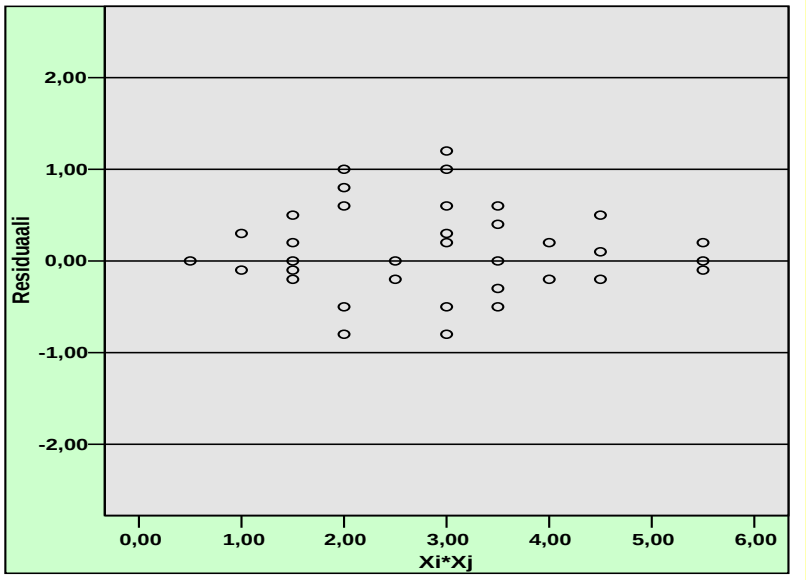
Tarvitaanko mallissa yhdysvaikutustermejä (”interaction terms”)?

yhdysvaikutus ”interaction” Kun tutkittavan tekijän (A, esim. hoito) vaikutus lopputulokseen on erilainen riippuen jostain toisesta tekijästä (B, esim. lääkkeenantotapa), niin A:n ja B:n välillä on yhdysvaikutus. Tällöin tekijöiden A ja B vaikutusta lopputulokseen ei voi välittömästi arvioida.

Tähän kysymykseen saadaan selvyyttä korreloimalla residuaalimuuttuja erilaisten tulotermien $x_i \cdot x_j$, …, $x_i \cdot x_j \cdot x_k$, jne. kanssa. Mikäli merkitsevää korrelaatiota esiintyy, tulisi kyseiset yhdysvaikutustermit sisällyttää malliin. Interaktioiden tarkastelu on usein havainnollista suorittaa siten, että tarkastelun kohteena olevat muuttujat x_i ja x_j jaetaan sopivasti luokkiin ja ristiintaulukoidaan residuaalit näiden muuttujien suhteen. Mikäli positiivisten ja negatiivisten residuaalien osuus on yhtä suuri taulukon jokaisessa solussa, yhdysvaikutusta ei esiinny.



Tulkinta: Koska interaktiotermin $x_i \cdot x_j$ mallin residuaalin välillä on selvä korrelaatio, niin interaktiotermin mukaanotto parantaa mallin yhteensopivuutta.



Tulkinta: Koska interaktiotermin $x_i \cdot x_j$ ja mallin residuaalin välillä ei ole korrelaatiota ja residuaalit ovat pieniä, niin interaktiotermin sisällyttäminen malliin ei paranna mallin yhteensopivuutta.

Pitäisikö lisämuuttuja z sisällyttää malliin?

Mikäli residuaalimuuttujan ja z:n välinen korrelaatio on merkitsevä, niin z:n lisääminen on aiheellista.

Riippumattomuusolettamuksen kanssa voi syntyä ongelmia silloin, kun tutkimusaineistossa on **toistomittauksia** samasta henkilöstä tai muulla tavoin on aiheutettu teknisiä riippuvuussuhteita havaintoyksiköiden välillä.

Mikäli **vakiovarianssiusolettamus** todetaan paikkansapitämättömäksi, kannattaa ensiksi kokeilla y:n muuntamista joko logaritmiseksi, $\log_e(y)$, tai käyttää käänteismuunnosta (1/y) tai neliöjuurimuunnosta $\sqrt{y}$. Kaikilla näillä muunnoksilla on varianssia vakioiva vaikutus. Mikäli mitkään niistä ei tehoa, niin vasta sitten kannattaa turvautua havaintojen painottamiseen, joka usein johtaa tulkinnallisiin vaikeuksiin.

Huom. Logaritmimuunnosta voi käyttää vain, mikäli y:n arvot ovat >0.

On olemassa myös kehittyneempi keino etsiä muunnosta, jolla y:hyn liittyvä **normaalisuusvaatimus** saataisiin mahdollisimman hyvin voimaan; nk. **Box-Cox-proseduuri**, joka perustuu potenssimuunnoksiin:

$$y' = (y^{\lambda} - 1)/\lambda, \text{ jos } \lambda \neq 0 \text{ tai } y' = \log_e y, \text{ jos } \lambda = 0.$$

Paras arvio λ:lle etsitään käyttäen suurimman uskottavuuden (“maximum likelihood”) menetelmää

- jos λ = 0, niin kyseessä on logaritmimuunnos
- jos λ = ½, niin kyseessä on neliöjuurimuunnos
- jos λ = -1, niin kyseessä on käänteismuunnos
- jos λ = 1, niin ei tarvitse tehdä muunnosta

On olemassa myös yhdistetty normaalisuus- ja vakiovarianssisuusmuunnos (ks. esim. Sokal ja Rohlf, s. 425)

Kuinka kertoimet tulkitaan?

Yhtälössä **$y = b_0 + b_1 \cdot x$** regressiokerroin b kuvaa lineaarista riippuvuutta x:n ja y:n välillä siten, että x:n muuttuessa yhden yksikön verran arvioitu y:n arvo muuttuu b_1:n yksikön verran. Kun x on 0, niin y:n arvioitu arvo on b_0 (=vakiotermi, regressiosuoran ja y-akselin leikkauspiste). Yleensä tutkijat ovat kiinnostuneempia b_1:stä kuin b_0:sta.

Vastaavasti usean selittäjämuuttujan tapauksessa b-kertoimien tulkinta on seuraavanlainen: Tarkastellaan yhtälöä $y = b_0 + b_1x_1 + b_2x_2$. Tällöin b_1 ilmaisee y:n keskimääräisen muutoksen kun x_1 muuttuu yhden yksikön verran ja sitä ennen x_2:n lineaarinen riippuvuus ajatellaan poistetuksi sekä x_1:stä että y:stä. Näin regressiomallin avulla voidaan vakioida mm. sekoittavia tekijöitä (“confounding factors”).

Mikäli yhden yksikön suuruinen x_1:n aiheuttaa y:ssä erilaisen muutoksen sen mukaan, onko x_2 suuri vai pieni, niin edellä esitetyn mallin yhteensopivuutta voidaan parantaa liittämällä mukaan yhdysvaikutustermi $b_3(x_1 \cdot x_2)$, kuten edellä on todettu.

Regressioanalyysissä kerrointen tulkintaa saattaa olennaisesti häiritä se, että mallissa on mukana monia keskenään voimakkaasti korreloivia muuttujia, esim. useita samaa luonteenpiirrettä kuvaavia mittareita. Tällaista tilannetta kutsutaan **multikollineaarisuusongelmaksi.** Tällöin regressiokerrointen estimaatit tulevat epävakaiksi; kerrointen keskivirheet kasvavat.

Erityisesti on syytä muistaa, että jos malliin rakennetaan (0,1)-indikaattorimuuttujia useampiluokkaisista laatueroasteikollisista muuttujista, niin indikaattoreita tulee olla yksi vähemmän kuin alkuperäisessä muuttujassa on luokkia. Yleisesti ottaen (0,1)-muuttujien käyttö regressiomallissa ei aiheuta ongelmia, mikä johtuu siitä, että b_i:t ovat tavallaan summia ja siten niiden otosjakauma pyrkii normaaliseksi huolimatta x_i:n jakaumasta.

Regressioanalyysimenetelmä on erittäin herkkä kerrointen etumerkkien suhteen, mikäli multikollineaarisuutta esiintyy. Etenkin askeltavassa regressioanalyysissä saattaa tällöin syntyä tulkinnallisesti vaikea tilanne, kun jonkin muuttujan kertoimen etumerkki vaihtuu askeleesta toiseen. Muista muuttujista riippumattomien muuttujien kertoimet ovat yleensä vakaita eivätkä aiheuta tulkintavaikeuksia.

Yksittäisten havaintoarvojen vaikutusta regressiokertoimiin voidaan tutkia käyttämällä painomuuttujia siten, että annetaan paino nolla jollekin havaintoarvolle. Etenkin **poikkeavien havaintoarvojen** (“outlier”) merkityksen arvioinnissa tällä menettelyllä on käyttöä.

Standardoidut regressiokertoimet

Regressioanalyysin tulosteissa esiintyy tavallisesti myös nk. **standardoidut kertoimet**, jotka lasketaan kaavalla: $\mathbf{b_i^* = b_i \frac{s_{x_i}}{s_y}}$

Kaavasta voidaan todeta, että kyseessä on riippumattoman muuttujan x_i ja riippuvan muuttujan y hajonnan suhteella normitettu regressiokerroin. Tällä menettelyllä pyritään saamaan selittävät muuttujat paremmin vertailukelpoisiksi keskenään, jotta niiden suhteellista osuutta y:n vaihtelun selittämisessä kyettäisiin arvioimaan. Asia ei ole kuitenkaan ollenkaan yksiselitteinen ja standardoidut kertoimet voivat olla pahasti harhaanjohtavia (ks. esim. Greenland, 1986).

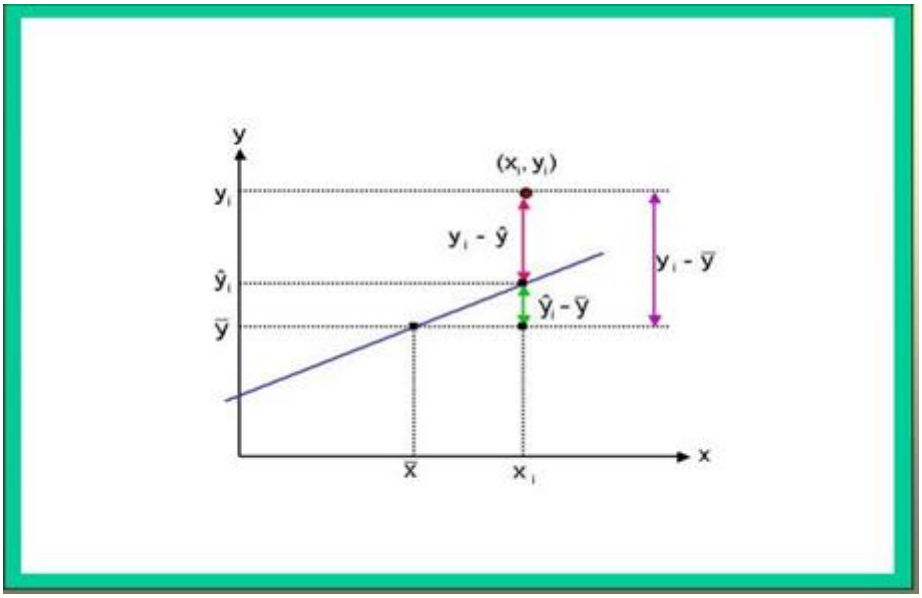
Neliösummat

Merkitään riippuvan muuttujan (y) kokonaisvaihtelua edustavaa poikkeamista $y_i - \bar{y}$ laskettava (ks. kuva) neliösummaa SS_{tot}, käytetyn regressiomallin avulla selittyvää osuutta vaihtelusta SS_{reg} ja satunnaisvaihtelun osuutta SS_{res}. Näiden neliösummien välillä pätee yhteys: $SS_{tot} = SS_{reg} + SS_{res}$, joten todetaan, että regressiomallin yhteensopivuus käytetyn havaintoaineiston kanssa on luonnollisesti sitä parempi, mitä pienemmäksi satunnaisvaihtelun osuus SS_{res} jää. Näiden neliösummien suhteellista osuutta voidaan paremmin arvioida käyttämällä nk. keskineliösummia MS_{tot}, MS_{reg} ja MS_{res}. Neliösummien laskentakaavat ovat seuraavat:

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ MS_{tot} = \frac{SS_{tot}}{n-1},\ SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y})^2,\ MS_{res} = \frac{SS_{res}}{n-p-1}$$

$$SS_{reg} = SS_{tot} - SS_{res},\ MS_{reg} = \frac{SS_{res}}{p}$$

Edellä olevissa kaavoissa n on havaintoyksiköiden määrä ja p on muuttujien määrä mallissa. Keskineliösummien (MS) nimittäjässä olevaa lukua kutsutaan **vapausasteeksi** ("degree of freedom").



Kuva: Neliösummissa esiintyvät poikkeamat

Ohjelmistopakettien antamissa tulosteissa on yleensä annettu sekä poikkeamaneliösummat että keskineliösummat. Tulosten tulkinnan ja jatkoanalyysien, kuten esimerkiksi testien kannalta keskineliösummat ovat merkityksellisemmät.

Huom.

MS:ien perusteella saadaan: $s_y^2 = MS_{tot}$ (y:n varianssi) ja $s_{y|x}^2 = MS_{res}$ (y:n ehdollinen varianssi)

Yhteiskorrelaatiokertoimen neliö

Tärkeä suure arvioitaessa regressiomallin hyvyyttä arvioitaessa on **yhteiskorrelaatiokertoimen neliö** (R^2, "multiple correlation coefficient squared"). Yleensä luku ilmaistaan prosentuaalisena $100 \cdot R^2$ ja todetaan, että tämä osuus y:n vaihtelusta kyettiin käytetyllä regressiomallilla selittämään.

Yhteiskorrelaatiokertoimen neliö (R^2), determinaatiokerroin, "R-squared", "coefficient of determination" ilmaisee regressioanalyysissä sen osuuden selitettävän eli riippuvan muuttujan ("dependent variable") vaihtelusta, joka selittyy mallissa olevilla riippumattomilla muuttujilla ("independent variables"), ts. kuinka paljon riippuvan muuttujan saamista arvoista voidaan laittaa yksinomaan mallissa olevien muuttujien tiliin.

Kun regressiomalliin lisätään muuttujia, niin teknisistä syistä johtuen R^2-arvo kasvaa. Ainoastaan numeeristen laskentamenetelmien aiheuttamat pyöristysvirheet voivat aiheuttaa toisenlaisen tilanteen. Tutkijan täytyy kuitenkin muistaa suhteuttaa malliin sisällytettävien muuttujien määrä havaintojen määrään, muuten saadut tulokset eivät ole realistisia, vaikka R^2-arvo olisikin korkea. Tavallisen R^2:n asemesta kannattaakin yleensä tarkastella ja ilmoittaa ns. **adjustoitu yhteiskorrelaatiokertoimen neliö** (R^2_{adj}), joka huomioi mallissa olevien muuttujien/parametrien määrän. Näiden suureiden laskentakaavat ovat seuraavat:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}, \text{ missä } SS_{tot} = SS_{reg} + SS_{res}$$

Adjustoitu R^2 lasketaan kaavalla:

$$R^2_{adj} = 1 - \frac{MS_{res}}{MS_{tot}}$$

Suureen R^2 positiivinen neliöjuuri R on ns. **yhteiskorrelaatiokerroin**, joka on korkein mahdollinen yksinkertainen (=Pearsonin) korrelaatio y:n ja minkä tahansa muuttujien $x_1, \ldots, x_p$ lineaarikombinaation $b_0+b_1x_1+\ldots+b_px_p$ välillä. R^2 ilmaisee, kuinka paljon y:n varianssista selittyy regressiomallin avulla

Varianssianalyysitaulukko

Tilastopaketit esittävät regressioanalyysin päätulokset yleensä seuraavanlaisena varianssianalyysitaulukkona:

Vaihtelulähteet	Neliö-summat	Vapaus-asteet	Keskineliösummat	F-testi	P-arvo
Mallin selittämä vaihtelu	SS_{reg}	p	$MS_{reg} = \frac{SS_{reg}}{p}$	$\frac{MS_{reg}}{MS_{res}}$	
Jäännösvaihtelu (residuaali / virhe-vaihtelu)	SS_{res}	n - p -1	$MS_{res} = \frac{SS_{res}}{n-p-1}$		

Testattava hypoteesi: Yhteiskorrelaatiokerroin perusjoukossa on nolla, eli malli ei selitä ollenkaan y:n vaihtelua, ts. H_0: **R = 0** . Tämä tarkoittaa samaa kuin, että kaikki regressiokertoimet ovat nollia, ts. H_0: $\beta_1=\beta_2=\ldots=\beta_p=0$.

Huom.

F-testisuure voidaan ilmaista myös yhteiskorrelaatiokertoimen R avulla muodossa:

$$F = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Multikollineaarisuus

Multikollineaarisuus tarkoittaa sitä, että regressiomallissa mukana olevat muuttujat korreloivat keskenään liian voimakkaasti ja sen seurauksena mallin parametrien arviointi häiriintyy. Multikollineaarisuutta on syytä epäillä, kun:

- regressiokertoimien keskivirheet SE(β_i) ovat poikkeuksellisen suuria
- regressiokertoimella on "väärä" etumerkki
- yhteensopivuustesti antaa mallille hyvän "fitin" vaikka minkään yksittäisen muuttujan kerroin ei ole tilastollisesti merkitsevä (Waldin testi)
- regressiokertoimet ovat herkkiä, eli epästabiileja pienille mallin rakenteellisille muutoksille tai yksittäisten "data"-pisteiden lisäämiselle tai poistamiselle

Varianssia suurentava ("inflate") tekijä x_i:lle:

$\mathbf{VIF_i} = \frac{1}{1-R_i^2}$, missä R_i = on x_i :n ja kaikkien muiden mallissa olevien muuttujien välinen yhteiskorrelaatiokerroin.

Nyrkkisääntö VIF:n tulkinnalle:

- Mikäli $VIF_i \geq 4$ (ts. mikäli $R_i \geq 0.87$), niin kyseessä on ongelma mallin kannalta.
- Mikäli $VIF_i \geq 10$ (ts. mikäli $R_i \geq 0.97$), niin kyseessä on paha ongelma mallin kannalta.

Huom.

Tilastopaketeissa esiintyy myös käsite **toleranssi**. Se on VIF:n käänteisarvo; ts. Toleranssi=1/ VIF.

Residuaalien kvantitatiiviset analysointimenetelmät

Kuten edellä, merkitään havaittuja riippuvan muuttujan y arvoja (y_1, …, y_n), missä n on aineistokoko ja residuaaleja (jäännöksiä) $\boldsymbol{\varepsilon_i = y_i - \hat{y}_i}$, missä y_i on henkilön i havaittu ja $\hat{\mathbf{y}}_\mathbf{i}$ mallin perusteella ennustettu y:n arvo. Tilastopaketeissa on useita eri tarkoitusperiä

palvelevia menetelmiä residuaalien kvantitatiiviseen arviointiin. Tavallisten, muuntamattomien, ns. “raakaresiduaalien” ε_i avulla voidaan jossain määrin paikallistaa poikkeavia havaintoja, mutta niiden ongelmana on, että niiden arvot riippuvat sekä käytetystä skaalasta että mittayksiköistä ja siksi ei voida ennalta antaa mitään nyrkkisääntöä milloin residuaali on “suuri”. Välitön ratkaisu tähän ongelmaan on normalisoida raakaresiduaalit jakamalla ne arvioidun y:n arvon keskivirheellä. Siten saadaan **standardoidut residuaalit:**

$$\varepsilon_{i,Stan} = \frac{\varepsilon_i}{s_{y|x}}, \text{missä } s_{y|x} = \sqrt{MS_{res}}$$

Huom.

$s_{y|x}$ on myös residuaalien hajonta ja siten standardoitu residuaali on mittayksiköistä riippumaton mitta. Sen keskiarvo on nolla ja hajonta yksi. Mikäli residuaalien jakauma olisi normaalinen, niin 5 %:lla havaintoarvoista standardoitu residuaali olisi itseisarvoltaan suurempi kuin 1.96.

Nyrkkisääntö:

Mikäli $\varepsilon_{i,Stan} \geq 2$, niin havaintopiste i kannattaa ottaa lähempään tarkasteluun ja mikäli $\varepsilon_{i,Stan} \geq 3$ erityistarkasteluun, mutta kummassakaan tapauksessa ei välttämättä kyseessä ole poikkeava havaintoarvo vaan normaalivaihteluun kuuluva arvo.

Huom.

Vaikka standardoiduista residuaaleista on apua poikkeavien havaintoarvojen etsinnässä, niin se, että kaikki standardoidut residuaalit ovat pieniä, ei takaa mallin hyvää yhteensopivuutta (“fittiä”) havaintoaineistoon. Edellä esitetyt graafiset tarkastelut residuaalikuvion muodosta antavat käsitystä mallin yhteensopivuudesta.

Vaikutusmitta (”leverage”):

Kuten edellä, merkitään havaittuja ja regressiomallin perusteella ennustettuja riippuvan muuttujan arvoja tutkimusaineistossa y_i ja $\hat{y}_i$, i=1,…n. Tutkijaa kiinnostaa usein onko jokaisella havaintoyksiköllä sama vaikutus regressiomalliin (ideaalinen tilanne) vai löytyykö havaintoyksiköistä sellaisia, joilla on selvästi suurempi vaikutus kuin muilla regressiomallin parametrien arviointiin. Tätä asiaa voidaan tutkia vaikutusmitan **(”leverage” = vipu tai vääntövoima)** avulla.

Vaikutusmitta saadaan siten, että esitetään ennustetut y:n arvot muodossa:
$y_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n$, missä painokertoimet h_{ij} riippuvat riippumattomien muuttujien $x_1, \ldots, x_p$ arvoista. Voidaan näyttää, että:

$$\sum_{i=1}^{n} h_{ij} = 1 \text{ ja } h_{ii} = \sum_{i=1}^{n} h_{ij}^2$$

Tutkijaa kiinnostavat arvot ovat suureet h_{ii}, i=1,…, n, joita kutsutaan vaikutuksen mitoiksi (“leverage”). Ne ovat välillä [0, 1] ja kertovat kuinka suuri vaikutus kullakin havaintoaineiston pisteellä on regressiomallin arvioinnissa.
Tavallisessa lineaarisessa regressiossa h_{ii}:t lasketaan kaavalla:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

Jos h_{ii} lähestyy arvoa 1, niin se merkitsee, että $\hat{y}_i$ lähestyy arvoa y_i, joka taas merkitsee sitä, että havainnolla i on suuri vaikutus ennustearvoon. h_{ii}:n odotusarvo on $E(h_{ii})=(p+1)/n$, missä p on riippumattomien muuttujien määrä mallissa.

Nyrkkisääntö: katkaisukohta "suurelle" vaikutukselle: **$2\cdot(p+1)/n$**

Studentisoidut residuaalit (t-jakaumaan sovitetut residuaalit):

Standardoidut residuaalit normalisoitiin jakamalla suureella $s_{y|x}$,,joka on vakio kaikilla x:ien arvoilla. Regressiosuoran (tai -tason) arviointi on kuitenkin luotettavimmillaan x:ien keskiarvon kohdalla ja heikkeni ääripäitä kohti mentäessä. Tämä oli todettavissa aiemmin tarkastelluista regressiosuoran luottamusväleistä. Niin kutsutut studentisoidut residuaalit poikkeavat standardisoiduista residuaaleista siinä suhteessa, että ne ottavat huomioon tämän luotettavuusefektin. Ne lasketaan kaavalla:

$$\varepsilon_{i,Stud} = \frac{\varepsilon_i}{s_{y|x}\sqrt{1-h_{ii}}}$$

Nyrkkisääntö:

Katkaisukohta "suurelle" vaikutukselle $|\varepsilon_{i,Stud}| \geq 2$ ($\varepsilon_{i,Stud}$ noudattaa t-jakaumaa suureeseen MS_{res} liittyvin vapausastein).

Huom.

Sekä $\varepsilon_{i,Stan}$:n että $\varepsilon_{i,Stud}$:n laskemissa käytetty hajonta $s_{y|x}$ laskettiin käyttäen kaikkia havaintopisteitä (sisäisesti standardoidut tai studentisoidut residuaalit). Toinen vaihtoehtoinen menettelytapa on jättää se havaintopiste pois $s_{y|x}$:n arvioinnissa, jonka residuaalia lasketaan, ts. piste i. Näin saatavia residuaaleja (ulkoisesti standardoidut tai studentisoidut puhdistetut **"deleted"** residuaalit). Näin saatavat residuaalit ovat herkempiä löytämään poikkeavia havaintoarvoja kuin puhdistamattomat.

Cook'in etäisyys:

Edellä tarkastellut suureet antavat mahdollisuuden paikallistaa niitä havaintopisteitä, jotka aiheuttavat poikkeamaa normaalisuusoletuksesta tai niitä, jotka dominoivat mallin estimointiprosessia. Cookin etäisyysmitta mahdollistaa tällaisten havaintopisteiden todellisen vaikutuksen arvioimisen, ts. sen kuinka paljon regressiokertoimet muuttuvat yksittäisten havaintopisteiden vaikutuksesta

Cookin etäisyys lasketaan kaavalla:

$$D_i = \frac{\varepsilon^2_{i,Stud}}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}$$

D_i riippuu sisäisesti $\varepsilon_{i,Stud}$:stä, joka kuvastaa mallin puutteellista yhteensopivuutta ("fittiä") pisteessä i, sekä pisteen i vaikutuksesta (h_{ii}). D_i noudattaa likimain F-jakaumaa vapausastein(k + 1) ja (n - k - 1).

Nyrkkisääntö: Di:tä voidaan tulkita seuraavasti:

- Mikäli $D_i > 1$, niin lisätarkastelu on aiheellista
- Mikäli $D_i > 4$, niin pisteen i kohdalla vakava poikkeavan arvon ("outlier") ongelma.

Huom.

Tilastopaketeissa esiintyy myös nk. **Mahalanobiksen etäisyys**. Sen käyttötarkoitus on hyvin samanlainen kuin Cookin mitan ja se on yhteydessä vaikutuksen mittaan seuraavasti: (Mahalanobiksen etäisyys) = (Vaikutuksen mitta)·(n-1)

Puuttuvat havaintoarvot

Lääketieteellisissä tutkimusaineistoissa ei voida välttyä siltä tilanteelta, että aineistoon jää puuttuvia tietoja; esimerkiksi näytteet saattavat pilaantua tai koe epäonnistuu laitevian takia. Monissa tilastollisissa analyysimenetelmissä tarvitaan täydelliset havainnot eli tietoja ei saa puuttua. Mikäli havaintojoukko ei ole täydellinen, valittavissa ovat seuraavat menettelytavat:

- Otetaan mukaan vain täydelliset havainnot,
- Kerätään regressiota koskeva tieto vain niistä havaintoyksiköistä, joista se on kunkin muuttujan osalta saatavilla.
- Korvataan puuttuvat tiedot keskiarvoilla,
- Ennustetaan puuttuvat tiedot muiden selittäjien avulla.
- Siirretään toistoasetelmissa viimeisin havainto eteenpäin

Mikäli puuttuvat tiedot menevät hankalasti ristiin eri muuttujien osalle eri havaintoyksiköissä, täydellisten havaintojen määrä voi supistua kovin pieneksi. Tällöin analyysimenetelmän teho heikkenee ja tuloksiin tulee virheellisyyttä.

Tietojen keräämistä vain niistä havaintoyksiköistä, joista sitä on kunkin muuttujan osalta saatavilla helppo soveltaa, mutta tämä ei takaa luotettavuutta (ks. Miettinen, 1985, s.232).

Mikäli puuttuvat tiedot korvataan keskiarvoilla, niitä ei saa olla kovin paljon ja puuttuvien tietojen tulisi jakaantua satunnaisesti aineistossa, sillä muuten erot ja riippuvuussuhteet saattavat tulla harhaisiksi.

Puuttuvien tietojen ennustaminen muiden selittäjien avulla on vaativin, mutta useimmissa tapauksissa myös luotettavin menettely.

Puuttuvien tietojen korvaamiseen on olemassa hyvä erikoisohjelma SOLAS 3.0 (www.statsol.ie/solas/solas.htm), mutta useimmissa standardeissa tilasto-ohjelmistoissa on joustavat puuttuvien tietojen käsittelymahdollisuudet, mutta kannattaa muistaa, että näennäisesti kaunis ja helposti tuotettu lopputulos ei välttämättä ole luotettava, mikäli puuttuvia tietoja on paljon.

Monissa toistomittauksia käyttävissä lääketutkimuksissa käytetään yleisesti menettelyä, jossa ennenaikaisesti keskeyttäneen potilaan viimeisin laboratoriomittausarvo siirretään edustamaan myös myöhempiä mittauskertoja. Tämä nk. LOCF-menettely ("last observation carried forward") saattaa johtaa harhaiseen tulokseen, mikäli keskeyttäneitä potilaita on paljon. Lisäksi harhan suuntaa on vaikea arvioida.

Regressioanalyysin yhteys erotteluanalyysiin

Olettakaamme, että tutkimusaineisto jakaantuu jonkin kriteerimuuttujan perusteella yksikäsitteisesti k:hon yksilöryhmään, jotka ovat tilastollisesti riippumattomia otoksia vastaavista tutkimuksen kohteena olevista perusjoukoista (esim. psykoottiset potilaat, neuroottiset potilaat ja vertailuhenkilöt). Olettakaamme, että tutkija on kiinnostunut, miten nämä ryhmät eroavat toisistaan tutkittavien asioiden suhteen, joita kuvaa muuttujajoukko $x_1, \ldots, x_p$.

Alun perin R. A. Fisherin kehittämä **erotteluanalyysi** (multiple discriminant analysis) soveltuu ratkaisumenetelmäksi seuraavan tyyppisissä ongelmissa:

1) Eroavat valittujen muuttujien ryhmäkeskiarvot toisistaan?
2) Mitkä ovat ne x_i-muuttujien lineaariset funktiot (erottelufunktiot), jotka erottelevat ryhmiä tehokkaimmin?
3) Kuinka moniulotteinen ryhmäkeskiarvojen välisten erojen ongelma on, eli kuinka moniulotteiseen avaruuteen ryhmäkeskiarvojen olennaiset erot oat projisioitavissa?
4) Löytyykö ryhmäkeskiarvojen välisiä eroja kuvaaville funktioille mielekäs tulkinta?

Erotteluanalyysin tulosten perusteella voidaan suorittaa monia jatkoanalyysejä, esimerkiksi uusien potilaiden diagnostisointia.

Ongelma 1) voidaan ratkaista **monimuuttujavarianssianalyysin** avulla edellyttäen, että perusjoukkojen kovarianssimatriisit ovat samoja. Muuttujien ryhmäkeskiarvojen välisiä eroja voidaan testata **F-testillä**. Nämä testit onkin syytä suorittaa ennen varsinaista erotteluanalyysiä. Yksittäisiin muuttujiin kohdistuvat F-testit eivät ratkaise erojen ongelmaa kokonaan, koska muuttujat korreloivat keskenään. Totaaliseksi testiksi soveltuu **Wilks'in Λ-testi** ("lambda-testi) ja sen F-approksimaatio. Λ-testiä ei kannata erikseen suorittaa, sillä se saadaan erotteluanalyysin sivutuotteena. Erotteluanalyysin edellytyksenä itse asiassa on, että ongelmaan 1 saadaan myöntävä vastaus, jonka jälkeen voidaan ryhtyä analysoimaan lähemmin havaittuja eroja.

Viitteet

Cornfield J: Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Fedr Proc 21: 58-61, 1962.

Feldstein MS: A binary variable multiple regression method of analyzing factors affecting perinatal mortality and other outcomes of pregnancy. J Roy Stat Soc Series A 129: 61-73, 1966.

Galton F: Regression towards mediocrity in hereditary statues. Journal of the Anthropological Institute 15: 246-263, 1886.

Greenland S, Schlesselman II, Criqui MH: The Fallacy of Employing Standardized Regression Coefficients and Correlations as Measures of Effect. Am J of Epid 123: 2, 1986.

Healy MJR, Golstein H: Regression to the mean. Ann Hum Biol 5: 277-280, 1978.

Miettinen OS: Theoretical Epidemiology. Principles of Occurrence Research in Medicine, John Wiley & Sons, New York, 1985.

Varianssi- ja kovarianssianalyysi

Varianssianalyysi (Anova, Manova), ”analysis of variance” on menetelmä, jonka avulla voidaan testata kvantitatiivisten suureiden (jatkuvien muuttujien) ryhmäkeskiarvojen välisiä eroja, kun vertailtavia ryhmiä on yli 2. Kahden ryhmän tapauksessa saadaan sama tulos kuin t-testillä. **Anova**-lyhennys viittaa yhden muuttujan analyyseihin ja **Manova** monimuuttuja-analyysiin, jolloin vertaillaan samanaikaisesti useiden muuttujien ryhmäkeskiarvoja. Anova:ssa testinä käytetään **F-testiä**, jolla verrataan esim. hoitoryhmien välistä vaihtelua hoitoryhmien sisällä tapahtuvaan vaihteluun, eli ns. virhevaihteluun (“error variance”, “residual variance” Anova voi olla yksisuuntainen (yksi ryhmittelevä tekijä, “one-way Anova”), kaksisuuntainen **”two-way Anova”** jne. Kliinisissä tutkimuksissa lopputulosmuuttujista suoritetaan usein myös toistomittauksia. Tällöin menetelmäksi soveltuu toistomittausten varianssianalyysi **”Anova with repeated measures”**. Anova:n tulokset voidaan tuottaa myös regressiotekniikalla käyttäen ilmaisinmuuttujia.

Huom.

Sekä varianssi- että kovarianssianalyysiin liittyvät eri mallivaihtoehdot löytyvät SPSS:n valikoista: **”Analyze”►”General Linear Model”** Molemmat menetelmät ovat periaatteessa regressioanalyysin erikoistapauksia ja voidaan mallittaa myös regressionanalyysin avulla valikosta: **”Analyze”►”Regression” ►”Linear”**.

Varianssianalyysin ongelma-asettelut

Muuttujat: Tarkastellaan muuttujia x ja y.

- Kun y on numeerinen, välimatka- tai suhdeasteikollinen ja likimain normaalijakautunut niin voidaan käyttää parametrista varianssianalyysiä.
- Kun y on järjestysasteikollinen ja ei voida tehdä mitään oletusta sen jakaumasta, niin käytetään parametritonta varianssianalyysia; Kruskal-Wallis, Friedman jne.

Käyttö: Varianssianalyysiä käytetään vastaamaan seuraaviin ongelmiin

- Onko ryhmäkeskiarvojen välillä eroa, kun ryhmiä on enemmän kuin kaksi 2 (parametrinen)?
- Eroavatko ryhmien jakaumat sijainnin suhteen toisistaan (parametriton)?

Kun varianssi- tai kovarianssianalyysiä tarkastellaan regressiomallina, niin x:t ovat ryhmiä ilmaisevia muuttujia (tekijöitä / tasoja / suuntia). Ryhmittevien tasojen perusteella varianssianalyysit ryhmitellään **yksisuuntaisiin**, **kaksisuuntaisiin** jne. analyyseihin. Mikäli tasot ovat limittäin, niin puhutaan **kytketystä** (“nested”) eli **hierarkisesta asetelmasta**.

Sisäkkäisasetelma ”nested design” on koeasetelma, missä esim. kukin tekijän B tasoista (luokista) esiintyy vain yhdellä tekijän A tasoista (luokista). Tällöin B:n sanotaan olevan “nested” A:n suhteen.

		Tekijä B						
		1	2	3	4	5	6	7
Tekijä A	1	X	X					
	2			X	X	X		
	3						X	X

Esim.

ollaan kiinnostuneita sairaalan (tekijä A) ja lääkärin (tekijä B) vaikutuksesta potilaan tyytyväisyyteen (muuttuja X). Lääkärit (1 - 7) voivat toimia vai yhdessä sairaaloista (1 -3), joten tekijä B on kytketty ("nested") tekijän A suhteen.

Nimitys varianssianalyysi johtuu siitä, että menetelmä perustuu varianssin pilkkomiseen ryhmien sisäiseksi, ryhmien väliseksi jne. varianssikomponenteiksi.

Mallityypit

1. Kiinteiden vaikutusten malli ("Fixed effects model") **malli I**

- luokittelevan tekijän luokat valittu tutkijan intressien mukaisesti
- kaikki mahdolliset luokat mukana

Esim.

a) Annos-vaste-tutkimus
(samat / eri henkilöt ja tarkastellaan vastetta eri annoksiin)

a) Meta-analyysi missä kaikki tutkittavaan aihepiiriin liittyvät ja kriteerit täyttävät tutkimukset ovat mukana

2. Satunnaisten vaikutusten malli ("Random effects model") **malli II**

- luokittelevan tekijän luokat valittu satunnaisesti
- aineisto valittu jokaiseen luokkaan satunnaisotannalla äärettömästä perusjoukosta

Esim.

a. Tutkitaan altistuksen ja taudin välinen yhteyttä ja altistustekijä luokiteltu k:hon luokkaan ja luokkarajat valittu satunnaisesti.

b. Meta-analyysi, jossa mukaan otettavat tutkimukset ovat edustava otos kaikista kriteerit täyttävistä tutkimuksista.

3. Sekamalli ("Mixed model")

- sisältää sekä kiinteitä että satunnaisia luokittelevia tekijöitä

Yksisuuntainen (parametrinen) varianssianalyysi

Käyttö:

Halutaan vertailla kvantitatiivisen suureen y keskiarvoja yhden luokittelevan tekijän mukaisissa ryhmissä, kun luokkia tai esim. hoitoryhmiä on kaksi tai enemmän, ts halutaan testata hypoteesia H_0: $\mu_1=\mu_2=\ldots=\mu_k$, $k\geq 2$. Kun k=2 yksisuuntainen varianssianalyysi ja siihen liittyvä F-testi palautuu kahden toisistaan riippumattoman ryhmän t-testiksi.

Merkitään aineistokokoa n_i:llä luokassa i; i=1, …, k ja N:llä kokonaisaineistokoa, eli

$$N = \sum_{i=1}^{k} n_i$$

.

Yksisuuntaisessa varianssianalyysissä tarkasteltavan suureen y kokonaisvaihtelu (=poikkeamaneliösumma) **SS_{tot}** pilkotaan hoitoryhmien väliseksi **SS_B** ja hoitoryhmien sisäiseksi **SS_W** vaihteluksi.

Huom.

Tämä on analoginen menettely regressioanalyysin kanssa SS_B vastaa neliösummaa SS_{reg} ja SS_W jäännösneliösummaa eli virhevaihtelua SS_{res}.

Varianssianalyysin keskeisimmät tulokset esitetään tavallisimmin seuraavanlaisena varianssitaulukkona:

Vaihtelulähteet	Neliö-summat	Vapaus-asteet	Keskineliösummat	F-testi	P-arvo
Hoitoryhmien välinen vaihtelu	SS_B	k -1	$MS_B = \frac{SS_B}{k-1}$	$\frac{MS_B}{MS_W}$	
Hoitoryhmien sisäinen vaihtelu (residuaali- / virhevaihtelu)	SS_W	N – k	$MS_W = \frac{SS_W}{N-k}$		

Suhde MS_H / MS_W noudattaa likimain F-jakaumaa vapausastein k-1 ja N – k, ja siten P-arvo taulukon viimeiseen sarakkeeseen saadaan F-jakaumasta.

F-jakauma "F-distribution" on kuuluisan matemaatikon R. A. Fisherin mukaan nimetty teoreettinen todennäköisyysjakauma, joka perustuu kahden varianssin suhteeseen. Käytetään erityisesti varianssianalyysissa.

F-testi on F-jakaumaan perustuva merkitsevyystesti, jota käytetään erityisesti varianssianalyysissa esim. vertaamaan ryhmien välistä vaihtelua ja ryhmien sisäiseen vaihteluun. Sen avulla voidaan esim. testata eroavatko kolmen tai useamman ryhmän keskiarvot toisistaan. Kahden ryhmän tapauksessa F-testi antaa saman tuloksen kuin
t-testi. Lineaarisessa regressioanalyysissa F-testiä käytetään vertaamaan regressiosta johtuvaa vaihtelua virhevaihteluun ja testaamaan siten regression merkitsevyyttä.

Mikäli P-arvo on pienempi kuin protokollassa ennalta määritelty tilastollisen merkitsevyyden raja (α-taso), niin johtopäätös on: ryhmäkeskiarvojen välillä on tilastollisesti merkitsevästi eroa.

Tämän jälkeen tutkija yleensä haluaa tietää tarkemmin ryhmien välisistä eroista; mitkä ryhmät eroavat keskenään toisistaan ja mitkä eivät? Tällöin yleensä menetellään niin, että tehdään ryhmien välisiä parittaisia vertailuita tai yhdistetään ryhmiä ja verrataan yhdistettyjä ryhmiä keskenään, eli muodostetaan ryhmäkeskiarvojen välille nk. **kontrasteja.** Mikäli kyseessä ei ole ns. suunnitellut vertailut, eli protokollassa ennalta määritellyt ("ad hoc") vertailut, niin joudutaan monivertailutilanteeseen, koska vertailut tehdään sen jälkeen ("post hoc") kun aineisto on kerätty ja todettu yllä olevasta taulukosta F-testin tulos. Silloin, kun F-testi antaa merkitsevän eron, niin se tarkoittaa sitä, että

ryhmäkeskiarvojen välillä on olemassa vähintään yksi kontrasti, jonka suhteen saadaan tilastollisesti merkitsevä ero.

Kaikissa post hoc vertailuissa α-taso inflatoituu ja siksi niihin liittyvissä monivertailutesteissä suoritetaan α-tason korjaus.

Esim.

α-tason inflatoituminen. Kolme hoitoryhmää P(lumehoito), A (aktiivihoito A) ja B (aktiivihoito B). Protokollassa on määritelty tilastollisen merkitsevyyden taso α=0.05. Oletetaan, että F-testillä on todettu, että ryhmien välillä on eroa ja tehdään tämän jälkeen parittaiset vertailut kaikkien ryhmien välillä. Kun tarkastellaan näitä testejä ryhmänä, niin P("ainakin yhdessä testissä ryhmien välillä on eroa") = 5 % + 5 % + 5 % = **15 %.** Jos halutaan säilyttää protokollassa määritelty merkitsevyystaso, niin parittaisissa vertailuissa pitäisi käyttää korjattua merkitsevyystasoa α*= α/3=0.0167. Tämä ajattelu edellyttää, että vertailut olisivat toisistaan riippumattomia. Näin ei kuitenkaan yleensä tämänkaltaisessa tilanteessa ole asianlaita ja siksi näin korjattu merkitsevyystaso tuottaa liian konservatiivisen tuloksen.

Yleisesti ottaen, jos ryhmiä on k kappaletta ja tehdään kaikki parittaiset vertailut, niin vertaluita tulee m=k·(k-1)/2. Bonferroni-epäyhtälöön α ≤ m·α* perustuva monivertailukorjaus on α*= α/m. Etenkin k:n ollessa suuri tämä korjaus antaa täysin epärealistisen tuloksen ja sitä ei pitäisi käyttää.

Toistettavuus, yhtäpitävyys

Kliinisissä mittauksissa saatu tulos riippuu monista eri tekijöistä, kuten mittalaitteesta, mittaajasta, olosuhteista, mittauksen kohteesta, biologisesta vaihtelusta jne. Mittausten toistettavuuden ja mittausvirheen arviointi on tärkeä osa kliinisten tutkimusten laadunarviointia ja -valvontaa.

Suomenkielessä sana toistettavuus voi tarkoittaa monia eri asioita. Sitä vastoin englanninkieli on tässä suhteessa rikkaampi, termejä on useita:

- **"repeatability"** on **samoissa** olosuhteissa suoritettujen toistettujen mittausten vaihtelua kuvaava suure
- **"reproducibility"** on **eri** olosuhteissa suoritettujen toistettujen mittausten vaihtelua kuvaava suure. Sillä mitataan esim. eri reagenssien, olosuhteiden, määrityslaitteiden ja laboratorioiden välistä vaihtelua. Hyvä toistettavuus samoissa olosuhteissa on tärkeä edellytys toistettavuudelle eri olosuhteissa.
- **"reliability"** ilmaisee kuinka konsistentisti sekä samoissa että eri olosuhteissa tehdyissä toistomittauksissa saadaan sama tulos.

Toistettavuuden arviontiin liittyvät laskut voidaan tehdä monin eri tavoin, mutta helpoin ja eri tilanteisiin parhaiten soveltuvin tapa lienee yksisuuntaisen varianssianalyysin (Anova) käyttö. Samat menetelmät soveltuvat myös eri mittalaitteiden yhtäpitävyyden arviointiin. Toistoja tai vertailtavia mittalaitteita voi olla useita eikä toistomittausten määrän tarvise olla vakio.

Mittausvirheiden hajonta **(SD_W)** voidaan laskea mm. Anovalla saaduista tuloksista kaavalla: **$SD_W = sqrt (MS_W)$**, missä MS_W on toistojen sisäistä vaihtelua kuvaava keskineliösumma.

Toistettavuutta kuvaava indeksi R voidaan laskea kaavalla:

R = 1- (havaittu epäyhtäpitävyys / sattuman perusteella odotettu epäyhtäpitävyys) =

$1 – MS_W / MS_T$

MS_W ja **MS_T** ovat keskineliösummia, jotka kuvaavat toistojen sisäistä ja kokonaisvaihtelua. MS_T mittaa havaintojen vaihtelua, kun ei huomioida toistojen kaltaistusta, eli se mittaa sattuman perusteella odotettavissa olevaa epäyhtäpitävyyttä satunnaisesti kaltaistetuille toistomittauksille.

R ilmaisee sen osuuden vaihtelusta, joka ei selity mittausvirheellä, vaan selittyy mitatuilla havaintoyksiköillä, esim. koehenkilöillä. R vaihtelee välillä (0 ”ei yhtäpitävyyttä”, 1 ”täydellinen yhtäpitävyys).

Samallakin mittalaitteella laskettu R riippuu vaihtelusta esim. kohdeväestössä. Suurella havaintoarvojen vaihtelulla (MS_T) on R:ää kasvattava vaikutus.

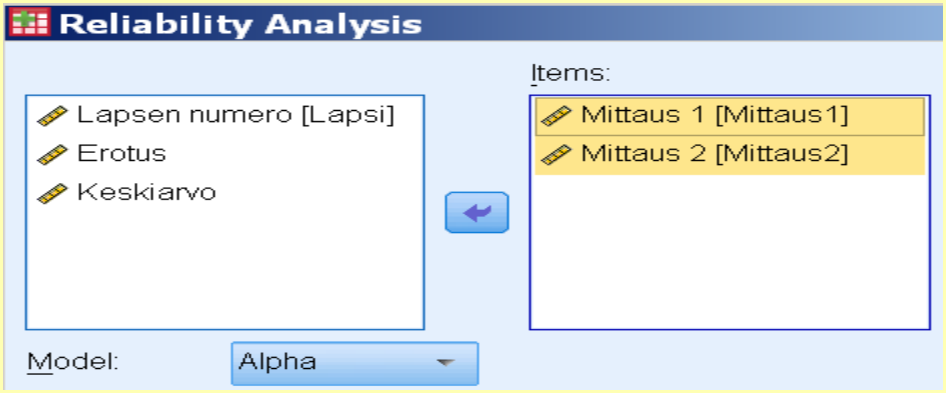
Sisäkorrelaatiokerroin ”intra-class correlation coefficient” **(ICC)** on suure, joka kuvaa luokkien/ryhmien samankaltaisuutta jonkin välimatka-asteikollisen muuttujan suhteen. Se on välimatka- ja laatueroasteikollisen muuttujan välisen riippuvuuden mitta. Kyseessä ei ole varsinaisesti korrelaatio, vaan samankaltaisuuden mitta, joka saa arvoja väliltä [0,1]. Se ilmoittaa kuinka suuri osuus kokonaisvaihtelusta johtuu luokkien välisestä vaihtelusta. Käytetään mm. toistomittausten samankaltaisuuden arvioinnissa. Suure voidaan laskea varianssianalyysin avulla. Se on erikoistapaus monen arvioitsijan reliabiliteettikertoimesta.

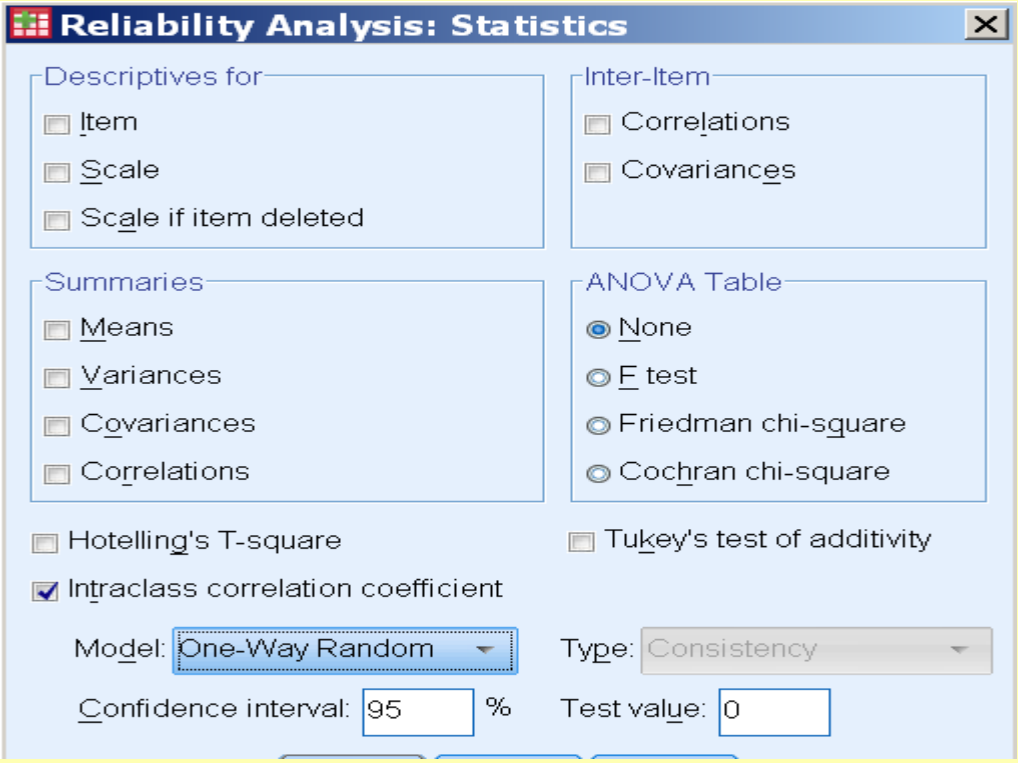
Esim. Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Syntymäpaino.sav

Toistettavuuden ja mittausvirheen arvionti lapsen syntymäpainon määrityksessä. Hypoteettinen aineisto.

	Syntymäpaino (g)		
Lapsen numero	***Mittaus 1***	***Mittaus 2***	***Erotus***
1	**3530**	**3450**	**80**
2	**3785**	**3695**	**90**
3	**3040**	**2990**	**50**
4	**2860**	**2860**	**0**
5	**3355**	**3405**	**-50**
6	**3070**	**3120**	**-50**
7	**2965**	**3040**	**-75**
8	**3750**	**3810**	**-60**
9	**3580**	**3675**	**-95**
10	**3400**	**3530**	**-130**

SPSS:llä sisäkorrelaatio ”Intra-class correlation” ja sen luottamusväli tässä yksinkertaisessa tapauksessa saadaan valikoista: **”Scale”►”Reliability Analysis”** seuraavasti:





Malliksi on valittu **”One-Way Random”**, koska ainoa merkitsevä vaihtelunlähde tassä tapauksessa on lapsi.

Tulos:

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,974	,906	,993	76,620	9	10	,000
Average Measures	,987	,951	,997	76,620	9	10	,000

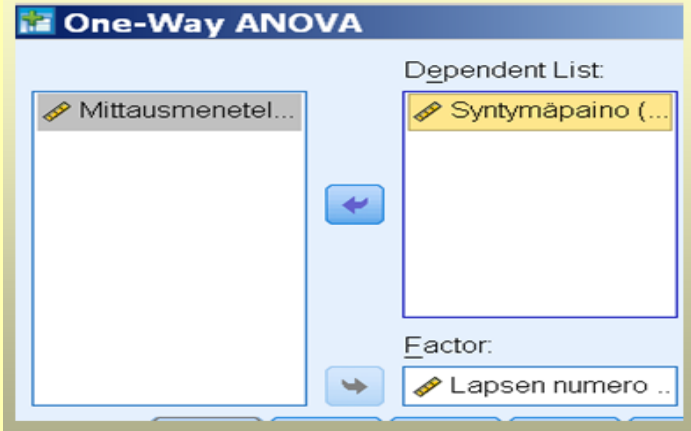
One-way random effects model where people effects are random.

Huom. Mittauksia kustakin havaintoyksiköstä voi olla myös enemmän kuin 2, mutta ei vaihteleva määrä.

Yleisempi ICC:n laskentatapa:
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Syntymäpaino2.sav

SPSS Valikot: **"Analyze"►"Compare Means" ►"One-Way ANOVA"**

Datan sisäänsyöttömuoto: ⟶



Lapsen nro	Toisto-mittaus	Syntymäpaino (g)
1	1	3530
2	1	3785
3	1	3040
4	1	2860
5	1	3355
6	1	3070
7	1	2965
8	1	3750
9	1	3580
10	1	3400
1	2	3450
2	2	3695
3	2	2990
4	2	2860
5	2	3405
6	2	3120
7	2	3040
8	2	3810
9	2	3675
10	2	3530

ANOVA

Syntymäpaino (g)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	**1971478,586**	**9**	**219053,176**	**76,620**	**,000**
Within Groups	**28589,429**	**10**	**2858,943**		
Total	**2000068,015**	**19**			

MS_W =2858,9 ja **MS_T**=2000068/19=105266,7, joten **R=0,974**

Tulkinta:

R=97,4 % sen osuuden vaihtelusta, joka ei aiheudu mittausvirheestä vaan selittyy lapsen syntymäpainon luonnollisella vaihtelulla. Mittausvirheen hajonta: **SD_W** = sqrt(2858,9)=**53,5 g**

Sisäkorrelaatio "intra-class correlation" **(ICC)** voidaan **kahden** toistomittauksen tapauksessa laskea kaavalla: **$ICC = (MS_B - MS_W)/(MS_B+MS_W)$.**

Toisin kuin Pearsonin korrelaatiokertoimessa ICC arvo pysyy muuttumattomana vaikka osa havaintoarvoista vaihdettaisiinkin mittauksesta 1 mittaukseen 2, ts. ICC on riippumaton havaintoarvojen järjestyksestä kunkin havaintoparin sisällä.

Tässä esimerkkitapauksessa **ICC**:lle saadaan arvo **0,974**, joka on kolmen desimaalin tarkkuudella kuin Pearsonin r ja lähes sama arvo kuin R:lle saatiin. Yleensä ICC:n arvo on jonkin verran pienempi kuin r:n arvo ja ICC:n hajonta on pienempi kuin r:n hajonta, eli se antaa tarkemman arvion. ICC:n likimääräinen luottamusväli voidaan laskea samoin kuin r:n luottamusväli.

		Mittaus 1
Mittaus 2	*Pearson Correlation*	,974(**)

ICC voidaan laskea myös silloin, kun havaintoyksiköissä, esim. näytteessä, on kahden toistomittauksen asemesta vaihteleva määrä toistoja. Tällöin yllä oleva kaava modifioituu muotoon:

$ICC = (MS_B - MS_W)/(MS_B+(n_0-1)\cdot MS_W)$,

missä n_0 on toistojen keskiarvo per havaintoyksikkö, joka lasketaan kaavalla:

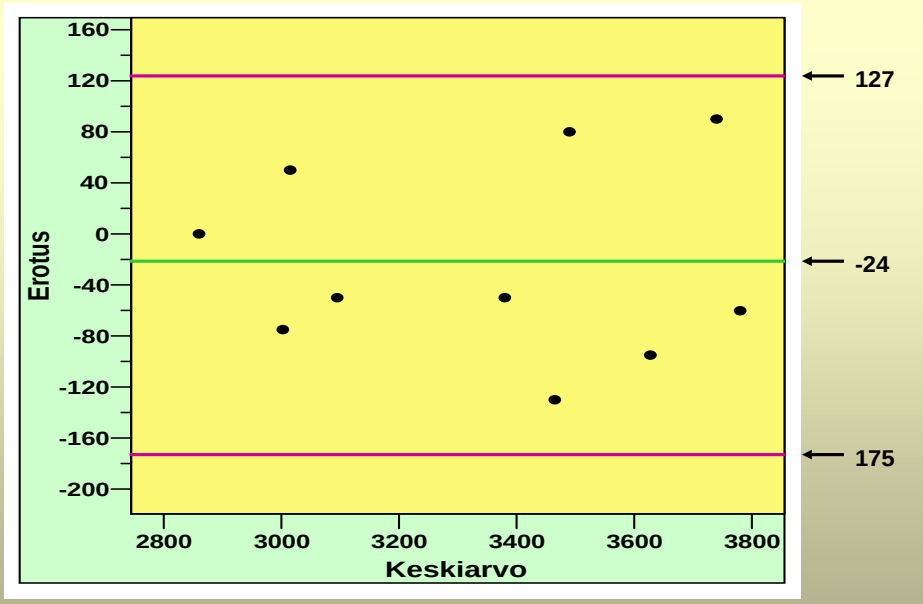
$$n_0 = \left(\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} n_i^2 \Big/ \sum_{i=}^{k} n_i\right) \Big/ (k-1)$$

missä k on havaintoyksiköiden (eli ryhmien Anovassa) lukumäärä ja n_i on i. havaintoyksikön toitomittausten määrä.

Verrattassa kahden menetelmän yhtäpitävyyttä voidaan käyttää myös **Blandin ja Altmanin ehdottamia yhtäpitävyysrajoja**. Ne lasketaan kaavalla:

(parimittausten erotuksen keskiarvo) ± 2•sqrt(2•MS_W)

Pariittaisten mittausten erotukset sijoittuvat tähän väliin 95 %:n varmuudella. Edellä olevassa esimerkissä saadaan seuraavat rajat:



Blandin ja Altmanin esittämät yhtäpitävyysrajat:
(erotuksen keskiarvo) ± 2·sqrt(2·MS_W)

11

Varianssien homogeenisuustestit

Useimmat tilastopaketit antavat varianssianalyysitaulukkoon kaksi vaihtoehtoista F-testiä: jojo **a)** oletetetaan **ryhmien varianssit homogeenisiksi** (yhtä suuriksi) tai **b) ei oleteta**. Tämän olettamuksen testaamiseksi eri paketeista löytyy monia eri tilanteisiin kehitettyjä testejä.

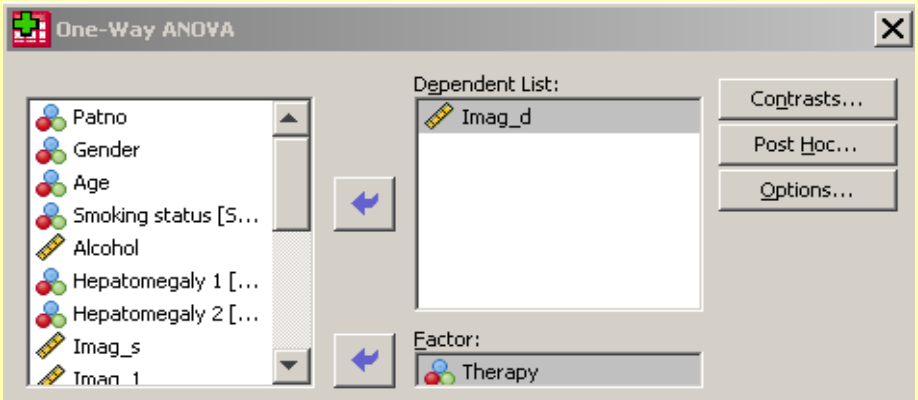
Testivalintasuositus: (Viite: Milliken & Johnson 1996)

1. Jos ollaan vakuuttuneita y:n normaalisuudesta, niin voidaan käyttää joko Bartlettin tai Hartleyn testiä. Jos ryhmäkoot kovin erisuuret, niin Bartlettin testi on suositeltavampi valinta.
2. Jos aineistokoko suuri, niin voidaan käyttää Boxin testiä, joka on melko robustinen (=ei ole herkkä poikkeamille varianssianalyysin perusoletuksista) mutta sen voima on heikko pienissä aineistoissa.
3. Kaikissa muissa tilanteissa suositellaan käytettäväksi **Levene'n testiä**. Tämä testi on hyvä yleisvalinta ja sisältyy useimpiin tilastopaketteihin, mm. SPSS:ään.

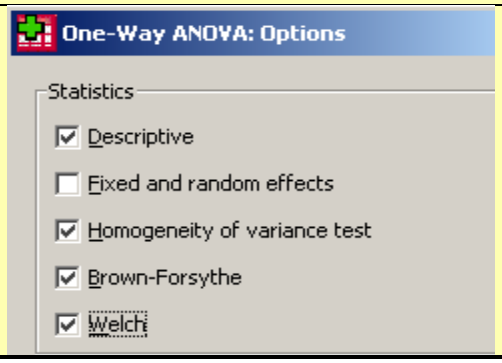
Esim.

Aineisto Trial. Kolmen hoidon **P**, **Q** ja **R** vertailu muuttujan **"Imag_D"**="Imag_2"-"Imag_S" (loppuarvo-alkuarvo) suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja.
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

SPSS: Valikot:**a)** **"Analyze"►"Compare Means"** tai valikoista **b)** **"Analyze"►"General Linear Model"►"Univariate".**
Tapa **a):** Viedään valikossa **"One-Way-Anova"** muuttuja **"Imag_d"** kohtaan **"Dependent list"** ja **"Therapy"** kohtaan **"Factor".**



Valikosta **"Options"** tehdään valinnat:



Brown–Forsythen testi ja **Welchin** testi ovat yksisuuntaisessa Anovassa käytettyjä keskiarvojen yhtäsuuruustestejä, joita tulisi käyttää F-testin asemesta silloin, kun varianssien yhtäsuuruusoletus (homogeenisuus) ei ole voimassa.

Descriptives

Imag_d

	N	Mean	Std. Deviation
Group P	15	-,0267	5,55729
Group Q	15	9,5467	6,07969
Group R	18	18,3556	16,75604
Total	48	9,8583	13,43990

Test of Homogeneity of Variances

Imag_d

Levene Statistic	df1	df2	Sig.
9,366	2	45	,0004

Tulkinta:

Vertailtavien ryhmien hajonnat poikkeavat tilastollisesti merkitsevästi toisistaan, P=0,0004 (Sig.=P-arvo)

Robust Tests of Equality of Means

Imag_d

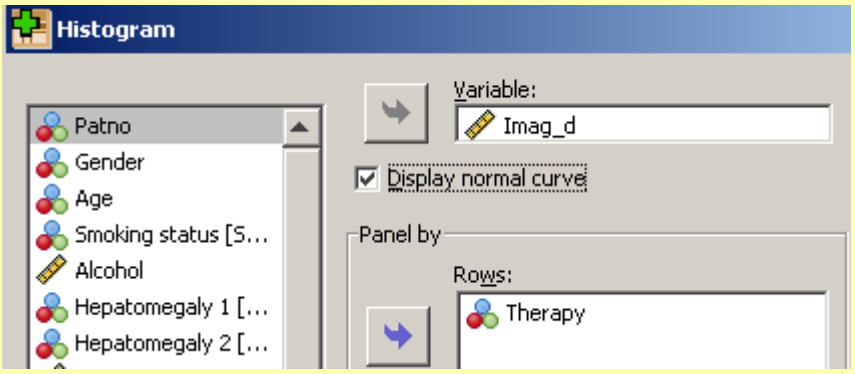
	Statistic[a]	df1	df2	Sig.
Welch	15,653	2	28,950	,00002
Brown-Forsythe	12,456	2	26,110	,00016

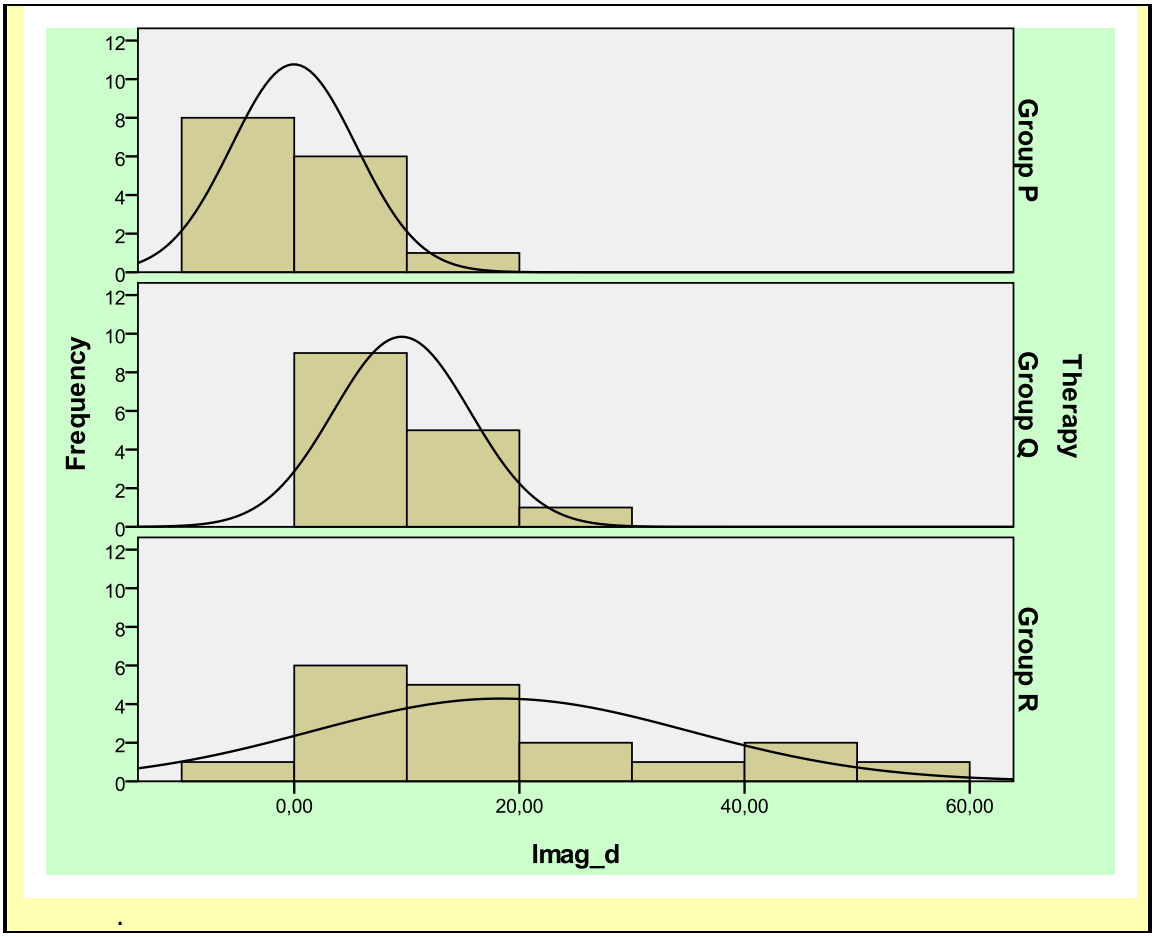
a. Asymptotically F distributed.

Tulkinta:

Varianssien homogeenisuusoletus ei ole voimassa, joten on turvallisempaa käyttää nk. robusteja testejä (Welch tai Brown-Forsythe), joissa testisuureen vapausasteita on muunnettu siten, että se paremmin noudattaisi F-jakaumaa

Jakaumakuviot saadaan valikosta **"Graphs"►"Legacy Dialogs" ►"Histogram"** seuraavin valinnoin:





.

Monivertailutestit

Monivertailutesti **”multiple comparison test”** Monivertailutestaustilanne syntyy silloin, kun pääloppututulosmuuttujan ohella vertaillaan useita muita lopputulosmuuttujia (jotka on valittu joko ad hoc tai post hoc), joiden testaamiseen ei voimalaskelmissa ole varauduttu. Toinen tavallinen monivertailutilanne syntyy, kun vertaillaan pareittain tuloksia useina ajankohtina tai samana ajankohtana useiden ryhmien välillä. Monivertailutestaustilanne kasvattaa α-virheen mahdollisuutta. Monivertailutesteissä α-virhe korjataan.

Monivertailutestien **perusidea**: Määritetään kriittiset erotukset, joita suurempia vastaavien keskiarvojen erotusten tulee olla ollakseen tilastollisesti merkitseviä.

Kriittinen arvo **”critical value”** on tietty prosenttipiste testisuureen otosjakaumassa. Se määrittelee ns. kriittisen alueen, jolle sijoittuvat testisuureen arvot johtavat nollahypoteesin hylkäämiseen prosenttipisteen ilmaisemalla todennäköisyydellä.

Ennalta suunnitellut parittaiset vertailut

Oletetaan, että F-testin antama P < 0.05 ja vertailut ovat **ennalta suunniteltuja (ad hoc).** Suositeltava menetelmä ryhmäkeskiarvojen välisiin parittaisiin vertailuihin on tällöin

Fisherin LSD-menetelmä (**L**east **S**ignificant **D**ifference), jonka testausmenettely on seuraavanlainen:

1. Vertailtavat keskiarvot $\bar{y}_1, \bar{y}_2, ..., \bar{y}_k$ lajitellaan suuruusjärjestykseen: $\bar{y}_{(1)}, \bar{y}_{(2)}, ..., \bar{y}_{(k)}$, missä $\bar{y}_{(i)}$ on jokin keskiarvoista $\bar{y}_1, \bar{y}_2, ..., \bar{y}_k$

2. Verrataan parittaisten erotusten itseisarvoja $\left|\bar{y}_{(i)} - \bar{y}_{(j)}\right|$ nk. **kriittiseen arvoon:**

$t_{\alpha/2,(N-k)} \cdot \sqrt{MS_W} \cdot \sqrt{\frac{1}{n_{(i)}} + \frac{1}{n_{(j)}}}$, missä $t_{\alpha/2,(N-k)}$ on Studentin t-jakauman merkitsevyystasoon α liittyvä ja prosenttipiste vapausastein N-k ja $\sqrt{MS_W}$ on ryhmien sisäinen hajonta, joka saadaan varianssitaulukosta.

Huom.

Tässä kohtaa LSD-menettely eroaa tavallisesta parittaisesta t-testistä, koska hajonta arvioidaan kaikkien tutkimusryhmien perusteella, eikä vain vertailtavan kahden ryhmän perusteella. on juuri

Paitsi ennalta suunniteltuja parittaisia vertailuita, protokollassa voidaan määrillä myös nk. **kontrasteja**, esim. muodostamalla yhdisttyjä ryhmiä. Merkitään perusjoukon keskiarvoja: $\mu_1, \mu_2, \ldots, \mu_k$. Kontrastit ovat muotoa:

$$L = \sum_{i=1}^{k} c_i \mu_i, \text{ missä } \sum_{i=1}^{k} c_i = 0$$

Esim.

Halutaan verrata kontrolliryhmää (perusjoukon keskiarvo μ_1) kahden aktiivihoidon yhdistettyyn ryhmään (perusjoukon keskiarvot μ_2 ja μ_3). Tällöin $L = \mu_1 - \frac{\mu_2 + \mu_3}{2}$

Esim.

Viisi hoitoa ja verrataan hoitoja 1 ja 5 keskenään. Tällöin
$L=\mu_1 + 0\cdot\mu_2 + 0\cdot\mu_3 + 0\cdot\mu_4 - \mu_5$

Esim.

Neljä hoitoa ja verrataan hoitoja 1 ja 2 hoitoihin 3 ja 4. Tällöin
$L=(\mu_1 + \mu_2) - (\mu_3 + \mu_4)$

Kontrastin luottamusvälit 100·(1-α) %:n **luottamusvälit** ($CL_{95\,\%}$) saadaan kaavasta:

$\sum c_i \bar{x}_i \pm \sqrt{(k-1)\cdot F_{\alpha/2,(k-1,N-k)}} \cdot \sqrt{MS_W} \cdot \sqrt{\sum \frac{c_i^2}{n_i}}$, missä $F_{\alpha/2,(k-1,N-k)}$ on F-jakauman prosenttipiste ja $\sqrt{MS_W}$ on residuaalihajonta.

Esim.

4 ryhmää ja N = 45, $\sqrt{MS_W}$ = 21.34 ja $F_{\alpha/2,(3,41)}$ = 2.84 (ks. StaTable). 95 %:n luottamusvälit ovat tällöin: $\sum c_i \bar{x}_i \pm 62.29 \cdot \sqrt{\sum \frac{c_i^2}{n_i}}$

Testausmenettely:

Nollahypoteesi (H_0) L= 0 paikkansa pitävyyttä testataan siten, että mikäli luku nolla ei sisälly laskettuun kontrastin 95 %:n luottamusväliin, niin johtopäätös on: P< 0.05.

Ennalta suunnittelemattomat vertailut (post hoc- vertailut)

Jälkikäteisvertailut ”a posteriori comparisons”, ”post hoc comparisons”
Vertailut, jotka ryhmien välillä suoritetaan sen jälkeen, kun yleisvaikutus lopputuloksesta on todettu. Esim. todetaan, että hoitoryhmien A, B ja C keskiarvojen välillä on eroa ja vertaillaan sen jälkeen ryhmiä pareittain keskenään.

Periaatteessa kaikki protokollassa määrittelemättömät vertailut, jotka tehdään sen jälkeen, kun aineisto on kerätty, tulokset analysoitu ja on todettu tilanne ryhmien välisten erojen suhteen yleisvertailussa ovat post hoc vertailuita. Englanninkielinen termi “data snooping” on aika kuvaava tällaiseen testausmenettelyyn.

Kaikissa ***post hoc*** -testimenetelmissä asetetaan vaatimuksia testattavalle aineistolle, kuten:

1. Varianssien homogeenisuus
2. Samat ryhmäkoot
3. Normaalisuus
4. Merkitsevä F-arvo keskiarvojen yleisvertailussa

Eri menetelmät asettavat erilaisia **vaatimuksia**, esimerkiksi:

1. **Fisherin** pienimmän merkitsevän eron testi. Olettamukset: 1, 2, 3, 4
2. **Bonferroni/Dunn**-menetelmä. Olettamukset: 1, 2, 3
3. **Tukeyn** menetelmä. Olettamukset: 1, 3
4. **Scheffen** menetelmä. Olettamukset: 4
5. **Dunnettin** menetelmä (vertailu kontrolliryhmään). Olettamukset: 3
6. **Student-Newman-Keuls.** Olettamukset: 1, 2, 3, 4(askeltava / kerroksittainen monivertailutesti)
7. **Duncanin** menetelmä. Olettamukset: 1, 3
8. **Games/Howellin** menetelmä. Olettamukset: 4, eräs uusimmista menetelmistä

Huom.

Monivertailumenetelmiä käytettäessä voi syntyä tilanne: F-testi antaa tuloksen P< 0.05, mutta mikään parittaisista vertailuista ei anna merkitsevää eroa. Tällöin on oltava vähintään yksi kontrasti, jossa saadaan tilastollisesti merkitsevä ero.

Suosituksia (Milliken & Johnson, 1984, s. 31):

1. Suorita keskiarvojen yleisvertailu F-testillä.
2. Jos F-testi antaa eron tasolla merkitsevyystasolla 5 %, niin käytä mihinkä tahansa ennalta suunniteltuihin vertailuihin, kontrastit mukaan lukien, **Fisherin LSD-menetelmää**. Mikäli kyseessä on ennalta suunnittelemattomat vertailut, niin käytä **Scheffen** menetelmää.
3. Jos yleisvertailu F-testi ei anna merkitsevää eroa (P>0.05), niin parittaisia eroja saattaa silti esiintyä. Käytä silloin **Bonferroni/Dunn**-menetelmää.

Esim.

Kliininen koe, jonka tavoitteena oli tutkia, kuinka eri työtehtävät vaikuttavat työntekijän pulssitasoon. (Lähde: Milliken & Johnson s. 37) Kokeessa oli yhteensä 78 miespuolista työntekijää, jotka oli satunnaistettu kuuteen ryhmään, joissa kussakin oli alun perin 13 työntekijää. Joitakin työntekijöitä keskeytti kokeen. Jokainen ryhmä oli koulutettu tekemään heille annettu työtehtävä. Testauspäivänä työntekijöiltä mitattiin pulssi. sen jälkeen, kun he olivat työskennelleet tunnin verran. **Tulokset** (pulssi/20 sek.):

Työtehtävä	1	2	3	4	5	6
n_i (henkilöiden määrä)	13	12	10	10	12	11
Pulssin keskiarvo	31.9	31.1	35.8	38.0	29.5	28.8

Kriittiset erotukset eri menetelmillä:

Aineistokoko		Fisherin LSD	Bonferroni	Scheffe	Tukey
13	12	4.450	6.809	7.658	6.677
13	10	4.676	7.154	8.047	7.314
13	11	4.556	6.971	7.841	6.973
12	10	4.760	7.283	8.192	7.314
12	12	4.540	6.946	7.813	6.677
12	11	4.642	7.102	7.989	6.972
10	10	4.972	7.607	8.557	7.313
10	11	4.858	7.433	8.361	7.313

Todetaan, että

1. Fisherin LSD on paras kaikista; kriittinen arvo on pienin jokaisella n_i:en kombinaatioilla.
2. Bonferroni on kaikissa kombinaatioissa parempi kuin Scheffe.
3. Bonferroni on joissakin tapauksissa parempi ja joissakin huonompi kuin Tukey.

Monivälitestit (”multiple range”-testit)

Monivälitesteissä menettelytapa verrattain monimutkainen, mutta pääpiirteissään seuraavanlainen:

1. Vertailtavat keskiarvot $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ lajitellaan aluksi suuruusjärjestykseen: $\bar{y}_{(1)}, \bar{y}_{(2)}, \ldots, \bar{y}_{(k)}$, missä $\bar{y}_{(i)}$ on jokin keskiarvoista $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$
2. Verrataan erotusta $\bar{y}_{(k)} - \bar{y}_{(1)}$ tiettyyn kriittiseen arvoon, joka vaihtelee riippuen mitä monivälitestiä käytetään.
3. Mikäli erotus $\bar{y}_{(k)} - \bar{y}_{(1)}$ ylittää kriittisen arvon, niin tarkastellaan kahta k -1:stä keskiarvosta muodostuvaa joukkoa $\bar{y}_{(1)}, \bar{y}_{(2)}, \ldots, \bar{y}_{(k-1)}$ ja $\bar{y}_{(2)}, \bar{y}_{(3)}, \ldots, \bar{y}_{(k)}$. Erotuksia $\bar{y}_{(k-1)} - \bar{y}_{(1)}$ ja $\bar{y}_{(k)} - \bar{y}_{(2)}$ verrataan kriittiseen arvoon.
4. Näin jatketaan menemällä pienempiin aina osajoukkoihin niin kauan, kun edellinen osajoukko antaa merkitsevän tuloksen. Kun testi antaa ei-merkitsevän tuloksen, niin testi ilmoittaa, että kyseiset keskiarvot eivät eroa toisistaan ja niistä

muodostetaan yhtenäinen alaryhmä ja niitä ei sisällytetä mihinkään jatkotesteihin.

5. Kun ollaan tilanteessa, että mikään väleihin liittyvä testi ei anna enää merkitsevää tulosta, niin on päästy monivälitestausprosessin loppuun.

Tilastopaketeissa on useita monivälitestausmenetelmiä, Esim. **”Hochberg’s GT2”, ”Tukey’s b”**, **Duncanin (uusi)** menetelmä ja **Student-Newman-Keulsin** menetelmä, jotka eroavat toisistaan kriittisen erotuksen määrittelytavan suhteen.

Esim. Edellä oleva pulssitasotutkimus Duncanin menetelmällä.

Työtehtävä	6	5	2	1	3	4
n_i (henkilöiden määrä)	11	12	12	13	10	10
Pulssin keskiarvo	28,2	29,5	31,1	31,9	35,8	38,0
Vaihe 1	I----------					
Vaihe 2		I----------				
Vaihe 3					I----------	

Tulkinta:

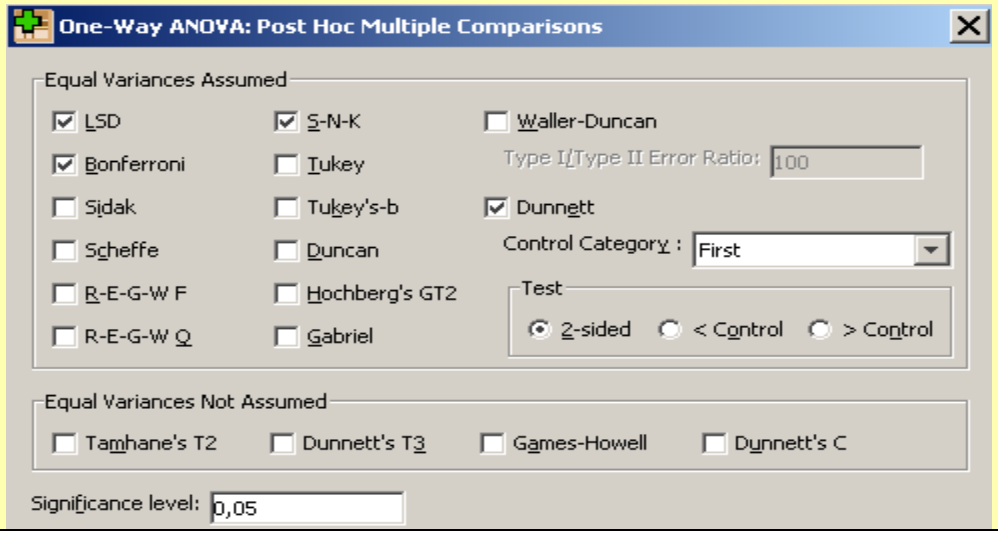
Testin perusteella todetaan, että osaryhmien (6,5,2,1), (2,1,3) ja (3,4) sisällä keskiarvot eivät poikkea tilastollisesti toisistaan.

Esim.

Aineisto Trial. Kolmen hoidon **P**, **Q** ja **R** vertailu muuttujan **”Imag_D”**=“Imag_2”-”Imag_S” (loppuarvo-alkuarvo) suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

SPSS: Valikot kuten edellä varianssien homogeenisuustesteissä joko : **a) ”Analyze” ►”Compare Means”** tai valikoista **b) ”Analyze” ►”General Linear Model”►”Univariate”.**
Tapa a): Viedään valikossa **”One-Way-Anova”** muuttuja **”Imag_d”** kohtaan **”Dependent list”** ja **“Therapy”** kohtaan **”Factor”.**
Valikosta **”Post Hoc”** valitaan halutut testivaihtoehdot:
Dunnettin testiin määritellään kontrollikategoriaksi **”First”** (eli ryhmä P)



Multiple Comparisons

Dependent Variable:Imag_d

	(I) Therapy	(J) Therapy	Mean Difference (I-J)	Std. Error	Sig.
LSD	Group P	Group Q	-9,57333*	4,11784	,025
		Group R	-18,38222*	3,94253	,000
	Group Q	Group P	9,57333*	4,11784	,025
		Group R	-8,80889*	3,94253	,030
Bonferroni	Group P	Group Q	-9,57333	4,11784	,074
		Group R	-18,38222*	3,94253	,000
	Group Q	Group P	9,57333	4,11784	,074
		Group R	-8,80889	3,94253	,091
Games-Howell	Group P	Group Q	-9,57333*	2,12675	,000
		Group R	-18,38222*	4,20202	,001
	Group Q	Group P	9,57333*	2,12675	,000
		Group R	-8,80889	4,24997	,119
Dunnett t (2-sided)[a]	Group Q	Group P	9,57333*	4,11784	,045
	Group R	Group P	18,38222*	3,94253	,000

*. The mean difference is significant at the 0.05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Tulkinta:
Näistä post-hoc-testeistä LSD on vähiten konservatiivinen ja antaa merkitsevyydet, P<0,05, kaikkien ryhmäparien välille. Bonferroni on konservatiivisin ja antaa merkitsevyydet ainoastaan ryhmäparille (P, R). Sekä Games-Howell-testin että Dunnettin testin perusteella molemmat aktiivihoitoryhmät eroavat lumeryhmästä P. Nämä testit eivät edellytä samoja samoja ryhmäkokoja eivätkä varianssien homogeenisuutta toisin kuin LSD ja Bonferroni.

Imag_d

	Therapy	N	Subset for alpha = 0.05		
			1	2	3
Student-Newman-Keuls[a,,b]	Group P	15	-,0267		
	Group Q	15		9,5467	
	Group R	18			18,3556

	Sig.		1,000	1,000	1,000

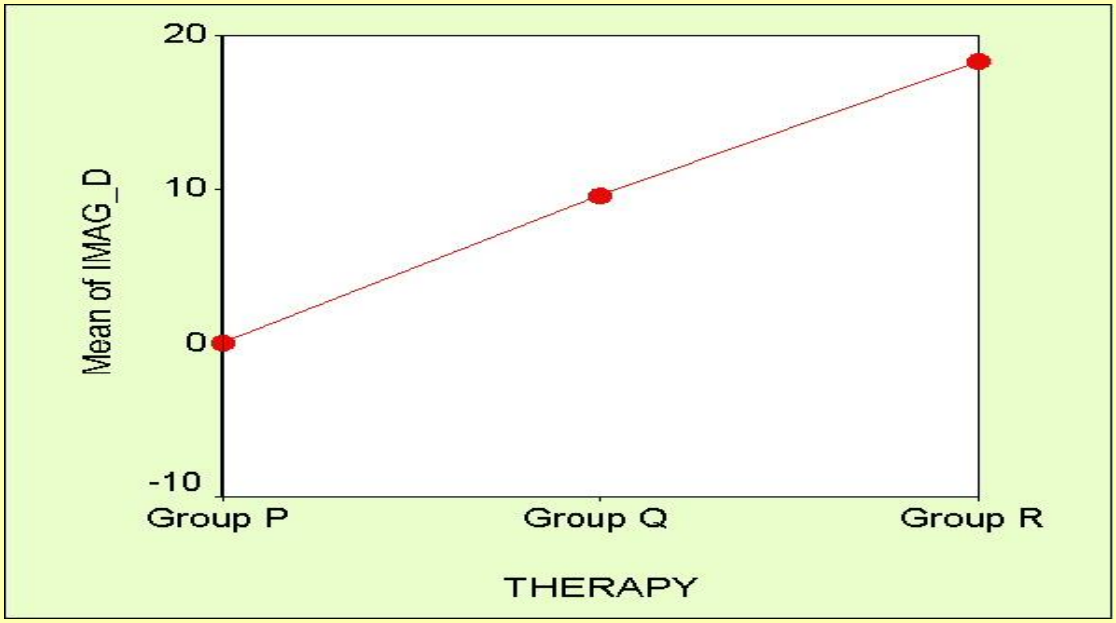
Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15,882.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Tulkinta: Aineisto on heterogeenninen muodostuen kolmesta toisistaan tilastollisesti merkitsevästi **(α=0,05)** poikkeavasta osaryhmästä

Means Plots:



Kovarianssianalyysi

Mikäli varianssimallissa on mukana sekoittavia tai adjustoivia tekijöitä eli kovariaatteja z, niin kyseessä on **kovarianssianalyysi.**

Kovariaatti **"covariate"** on muuttuja, joka liittyy tai vaihtelee yhdessä jonkin toisen muuttujan kanssa. Esim. ikää tarkastellaan usein kovariaattina tarkasteltaessa muiden muuttujien välisiä yhteyksiä.

Kovarianssianalyysi **"analysis of covariance"** **(ANCOVA)** on tilastollinen menetelmä, jolla verrataan kvantitatiivisten muuttujien keskiarvoja toisiinsa eri ryhmien välillä ottaen huomioon ryhmien väliseen vertailuun vaikuttavat kovariaatit, jotka voivat olla esim. sekoittavia tekijöitä tai lähtötason arvoja. ANCOVA:lla saadaan laskettua adjustoidut (=korjatut) ryhmäkeskiarvot. Kovarianssianalyysi suoritetaan usein regressiomallin avulla siten, että malliin sisällytetään tarvittava määrä ilmaisinmuuttujia ("dummy variable") ilmaisemaan vertailtavia ryhmiä.

Huom.

Sekä varianssi- että kovarianssianalyysissä perusolettamukset ovat samat kuin edellä tarkastellussa regressioanalyysissä.

Kliinisissä kokeissa, missä lopputulosmuuttuja on jatkuva, hoidon vaikutusta tutkitaan usein **a)** vertaamalla ryhmäkeskiarvoja lopputilanteessa toisiinsa tai **b)** laskemalla erotusmuuttuja loppuarvo-alkuarvo ja vertaamalla sen keskiarvoja toisiinsa kuten edellä tarkastellussa esimerkissä tehtiin. Mikäli vertailtavien ryhmien keskiarvot lähtötilanteessa eivät poikkea toisistaan, niin nämä kaksi menettelytapaa tuottavat lähes saman lopputuloksen. Jos hoitojen välillä on eroa, niin hoitoeron tilastollinen merkitsevyys riippuu lähtötason mittausten ja lopputulosmittauksen välisestä korrelaatiosta. Jos korrelaatio on matala, niin menettelytapa a) antaa herkemmin merkitseviä eroja kuin menettelytapa b), koska erotusmuuttujan käyttö lisää vaihtelua. Kääntäen, jos korrelaatio on korkea, niin menettelytavassa a) menetetään informaatiota ja tapa b) antaa herkemmin merkitseviä eroja hoitoryhmien välille. Valintaa käytettävästä menettelytavasta ei tietenkään valita tällä perusteella, vaan se pitää määritellä etukäteen tutkimussuunnitelmassa.

Esim.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav
Halutaan tutkia onko hoitoryhmien P, Q ja R välillä eroa Imag-muuttujassa tapahtuvan muutoksen Imag_S (alkuarvo) - Imag_2 (loppuarvo) suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja.

Correlations

		Imag_s	Imag_2
Imag_s	Pearson Correlation	1	,867**
	Sig. (2-tailed)		,000
	N	49	48

**. Correlation is significant at the 0.01 level (2-tailed).

Tulkinta:

Lähtötason muuttujan Imag_s ja lopputulosmuuttujan välillä on voimakas korrelaatio

Tapa a) *Imag_2*

ANOVA

Imag_2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4186,950	2	2093,475	4,201	,021
Within Groups	22921,708	46	498,298		
Total	27108,658	48			

Tapa b) *Imag_d*

ANOVA

Imag_d

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2766,806	2	1383,403	10,878	,00014
Within Groups	5722,851	45	127,174		
Total	8489,657	47			

Mikäli hoitoryhmien lähtötason muuttujien keskiarvoissa on eroa, niin erotusmuuttujan käyttö ei poista tätä ongelmaa, sillä ”regressio kohti keskiarvoa”-ilmiön johdosta lähtötason arvojen ja muutosten välillä voi olla negatiivinen korrelaatio (Bland & Altman 1994 ja Vickers & Altman, 2001). Esimerkiksi, jos kyseessä on mittari, missä suuret arvot merkitsevät hyvää lopputulosta, niin henkilöillä, joilla on matala lähtötaso, on suurempi mahdollisuus muutokseen kuin

niillä, joilla lähtötaso on korkea. Erotusmuuttujan käytön ongelmana on se, että sen arvo on sama riippumatta siitä missä kohtaa skaalaa erotus sijaitsee. Käytännössä usein kuitenkin on merkitystä miltä tasolta lähdetään, joten muutosta tutkittaessa lähtötaso pitäisi ottaa huomioon. Kovarianssianalyysi antaa tähän mahdollisuuden.

Kovarianssianalyysitekniikan etuna menettelytapoihin a) ja b) verrattuna on lisäksi sen parempi voima. Esim. Vickers & Altman esittävät esimerkin kliinisestä kokeesta, missä lähtötason ja lopputuloksen välinen korrelaatio oli 0,6 ja voimalaskelmien mukaisesti tarvittiin 85 per ryhmä, kun analyysit tehtiin tavalla a) 68 potilasta tavalla b) ja 54 potilasta kovarianssianalyysitekniikalla.

Kovarianssianalyysi on regressiotekniikka. Esim. jos oletetaan, että vertailtavia ryhmiä on kaksi, kovarianssianalyysiä voidaan kuvata mallilla:

Lopputulos = Vakiotermi + b_1•(Lähtötaso) + b_2•(Ryhmä),

missä muuttuja ”Ryhmä” on ryhmää ilmaiseva (0,1)-muuttuja. Jos vertailtavia ryhmiä on kolme, niin tarvitaan kaksi (0,1)-indikaattorimuuttujaa jne. Kerroin b_2 ilmaisee ryhmien välisen keskimääräisen hoitoeron. Käytännössä kovarianssianalyysi adjustoi jokaisen tutkittavan kohdalla lopputulosmuuttujan arvon lähtötason muuttujan arvolla.

Huolimatta satunnaistamisesta kliinisissä kokeissa saattaa esiintyä keskiarvoeroja vertailtavissa ryhmissä tutkittavan muuttujan suhteen. Menettelytapoihin a) ja b) verrattuna kovarianssianalyysitekniikan etuna on se, että epätasapaino lähtötason keskiarvoissa vertailtavien ryhmien välillä ei vaikuta lopputulokseen. Jos esim. hoitoryhmän lähtötaso on keskimäärin huonompi kuin lumeryhmässä, niin menettelytapa a) aliarvioi ja menettelytapa b) yliarvioi hoitoeroa toisin kuin kovarianssianalyysitekniikka.

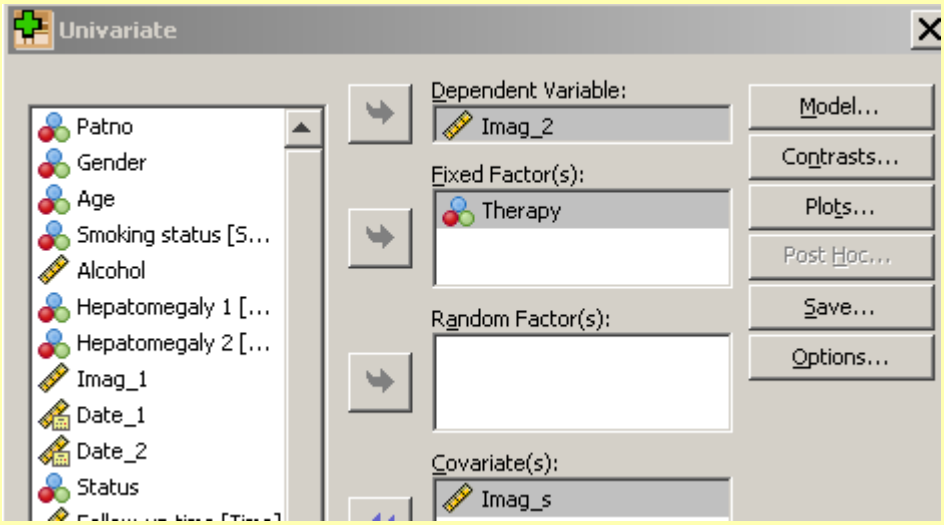
Esim.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav
Halutaan tutkia onko hoitoryhmien P, Q ja R välillä eroa Imag-muuttujassa tapahtuvan muutoksen Imag_S (alkuarvo) - Imag_2 (loppuarvo) suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja.

Edellä tätä asiaa tarkasteltiin erotusmuuttujan **”Imag_d”**=“Imag_2”-”Imag_s” (loppuarvo-alkuarvo) avulla.

Toteutus SPSS:llä voidaan tehdä kahdella eri tavalla a) Varianssianalyysivalikoista tai b) Regressioanalyysivalikoista.

Tapa a) Valikot **”Analyze”►”General Linear Model”►”Univariate”**



.

Kohdasta **“Options”** valitaan **“Descriptive Statistics”** ja kohtaan **“Display Means for:”** viedään muuttuja **“Therapy”**

Descriptive Statistics

Dependent Variable:Imag_2

Therapy	Mean	Std. Deviation	N
Group P	31,640	18,4036	15
Group Q	41,393	18,5903	15
Group R	53,911	27,7286	18
Total	43,040	23,8650	48

Tests of Between-Subjects Effects

Dependent Variable :Imag_2

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	22428,227[a]	3	7476,076	75,793	,000
Intercept	77,322	1	77,322	,784	,381
Imag_s	18310,895	1	18310,895	185,637	,0000
Therapy	2311,068	2	1155,534	11,715	,00008
Error	4340,088	44	98,638		
Total	115683,790	48			

Corrected Total	26768,315	47			

a. R Squared = ,838 (Adjusted R Squared = ,827)

Tulkinta:

Lähtötason muuttujalla ”Imag_S” on tilastollisesti merkitsevä adjustoiva vaikutus. Ryhmässä P lähtötaso on jonkin verran matalampi kuin ryhmässä R. Adjustoitujen ryhmäkeskiarvojen välillä on tilastollisesti merkitsevä ero (P<0,0001). Tässä tapauksessa kovarianssianalyysillä saatu P-arvo on hivenen pienempi kuin erotusmuuttujan perusteella saatu P-arvo.

Huom.

Jos analyysi tehtäisiin siten, että kohtaan ”Dependent variable” laitetaan loppuarvon ”Imag_2” asemesta erotus **”Imag_d”**, niin tulos olisi sama hoitoeron suhteen (ks. rivi ”Therapy”)

Tests of Between-Subjects Effects

Dependent Variable:Imag_d

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4149,569[a]	3	1383,190	14,023	,000
Intercept	77,322	1	77,322	,784	,381
Imag_s	1382,763	1	1382,763	14,019	,001
Therapy	**2311,068**	**2**	**1155,534**	**11,715**	**,00008**
Error	4340,088	44	98,638		
Total	13154,620	48			
Corrected Total	8489,657	47			

a. R Squared = ,489 (Adjusted R Squared = ,454)

Imag_s

Therapy	Mean	N	Std. Deviation
Group P	30,669	16	14,4573
Group Q	31,847	15	15,3325
Group R	35,556	18	14,2304
Total	32,824	49	14,4999

Adjustoidut ryhmäkeskiarvot (“Estimated Marginal Means”)

Therapy

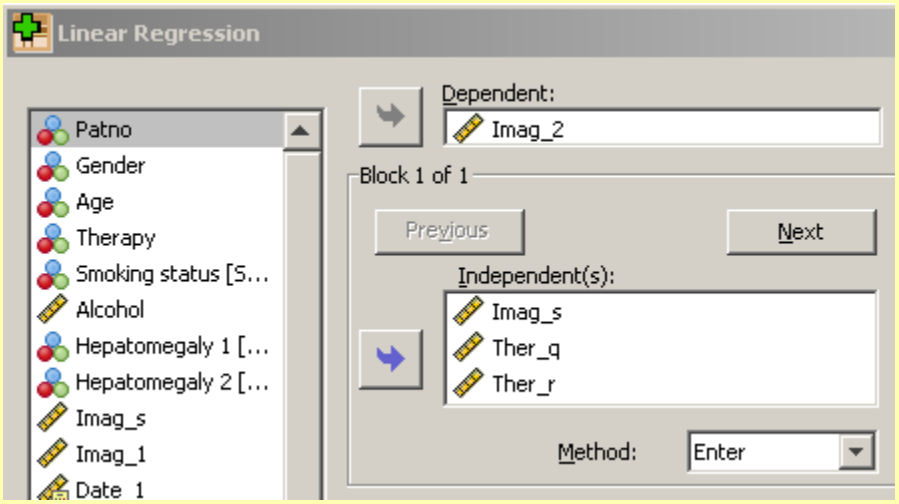
Dependent Variable:Imag_2

Therapy	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound

Group P	**33,729[a]**	**2,569**	**28,551**	**38,906**
Group Q	**43,234[a]**	**2,568**	**38,058**	**48,409**
Group R	**50,637[a]**	**2,353**	**45,894**	**55,380**

a. Covariates appearing in the model are evaluated at the following values: Imag_s = 33,181.

Tapa b) Valikot ”Analyze”►”Regression”►”Linear”



Muuttujat ”Ther_q” ja ”Ther_r” ovat (0,1)-indikaattoreita, jotka antavat ryhmäkeskiarvojen erot suhteessa lumeryhmään P ja Imag_s on kovariaatti.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**,915[a]**	**,838**	**,827**	**9,9317**

a. Predictors: (Constant), Ther_r, Imag_s, Ther_q

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	**22428,227**	**3**	**7476,076**	**75,793**	**,000[a]**
	Residual	**4340,088**	**44**	**98,638**		
	Total	**26768,315**	**47**			

a. Predictors: (Constant), Ther_r, Imag_s, Ther_q

b. Dependent Variable: Imag_2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-12,026	4,105		-2,930	,005
	Imag_s	1,379	,101	,834	13,625	,000
	Ther_q	9,505	3,627	,187	2,621	,012
	Ther_r	16,909	3,494	,347	4,839	,000016

a. Dependent Variable: Imag_2

Tulkinta:

Adjustoitu hoitoero Ryhmien Q ja P välillä on **9,5 (P=0,012)** ja vastaavasti ryhmien R ja P välillä **16,9 (P<0,0001)**. Nämä P-arvot ovat monivertailukorjaamattomia (Waldin testi).

Kaksisuuntainen varianssianalyysi

Oletetaan, että aineisto on luokiteltu kahden tekijän A ja B mukaisiin ryhmiin ja että tekijässä A on I kappaletta ja tekijässä B J kappaletta luokkia/tasoja.

Käyttötilanteet:

1. **Ristikkäisasetelma** (“crossed design”); jokaisessa lokerossa vähintään yksi havainto.
2. **Sisäkkäisasetelma** (“nested design”) eli hierarkinen asetelma; jokainen J:stä tasosta esiintyy korkeintaan yhdellä I:stä tasosta.
3. **Toistomittausasetelmat**; samoista henkilöistä on toistoja lokeron sisällä (esim. eri ajankohtina).
4. **Yhdistelykokeet** (“factorial design”); kaksi tai useampia luokittelevia tekijöitä

Silloin kun ei ole toistoja luokittelevien tekijöiden mukaisissa lokeroissa, niin varianssianalyysitaulukko on seuraavanlainen.

Vaihtelu-lähteet	Neliö-summat	Vapaus-asteet	Keskineliösummat	F-testi	P-arvo
Tekijä A	SS_A	I -1	$MS_A = \frac{SS_A}{I-1}$	$\frac{MS_A}{MS_{res}}$	
Tekijä B	SS_B	J -1	$MS_B = \frac{SS_B}{J-1}$	$\frac{MS_B}{MS_{res}}$	
Residuaali-eli virhevaihtelu	SS_{res}	$(I-1)\cdot(J-1)$	$MS_{res} = \frac{SS_{res}}{(I-1)\cdot(J-1)}$		

Kun kyseessä on **yhdistelykoe** siten, että aineisto on luokiteltu kahden tekijän mukaisiin ryhmiin, “lokeroihin” ja kussakin lokerossa on toistoja, niin varianssitaulukko on muotoa:

Vaihtelu-lähteet	Neliö-summat	Vapaus-asteet	Keskineliösummat	F-testi	P-arvo
Tekijä A	SS_A	I -1	$MS_A = \frac{SS_A}{I-1}$	$\frac{MS_A}{MS_{res}}$	
Tekijä B	SS_B	J -1	$MS_B = \frac{SS_B}{J-1}$	$\frac{MS_B}{MS_{res}}$	
Interaktio	SS_i	(I -1)·(J -1)	$MS_i = \frac{SS_i}{(I-1)\cdot(J-1)}$	$\frac{MS_i}{MS_{res}}$	
Residuaali-eli virhevaihtelu	SS_{res}	N - I·J	$MS_{res} = \frac{SS_{res}}{N-I\cdot J}$		

Huom.

N on kokonaisaineistokoko ja residuaalivaihtelu edustaa ryhmien sisäistä vaihtelua.

Esim.

2 x 2 –yhdistelykoe. (Lähde: Cutilletta AF et al., 1977, Development of left ventricularhypertrophy in young spontaneous hypertensive rats after peripheralsympathectomy. Circ Res 40: 428- 33.)
Tarkasteltava muuttuja y=munuaisten reniinipitoisuus (µg/l/h). Tutkimusasetelma: kaksi hoitoa (tekijä A) **NGFS**-hoito (“**N**erve **G**rowth **F**actor **S**erum”) ja **kontrolli**hoito (“Sham Serum”), 2 rottakantaa (tekijä B): **H**ypertensivinen ja **N**ormotensiivinen
Keskiarvot ryhmittäin:

	Hypertensiivinen	**Normotensiivinen**
Kontrolliryhmä	**2,41** (n=10)	**2,95** (n=8)
NFGS-ryhmä	**4,24** (n= 8)	**2,89** (n=6)

Kysymykset:

1. Onko hoitojen välinen ero sama molemmilla rottakannoilla vai onko interaktioita?
2. Onko hoitojen välillä eroa, kun kannat yhdistetään (painotettu keskiarvo)?

Varianssianalyysitaulukko:

Vaihtelulähde	**SS**	**Vapaus-asteet**	**MS**	**F-testi**	**P-arvo**
Hoito	5,990	1	5,990	100,3	**<0,001**
Kanta	0,619	1	0,619	10,4	**<0,001**
Interaktio	12,680	1	12,680	212,4	**<0,001**
Residuaali	1,671	28	0,060	-	

Tulkinta:

1. On interaktioita, eli hoidon vaikutus on erilainen eri kannoilla
2. Hoidon tilastollista merkitsevyyttä ei voi päätellä F-testistä, vaan pitää suorittaa kantojen sisällä monivertailutesti (Fisherin LSD):

Kanta H: $t = \sqrt{\frac{4.24 - 2.41}{0.0597 \cdot (\frac{1}{8} + \frac{1}{10})}} = 15.25, P < 0.001$

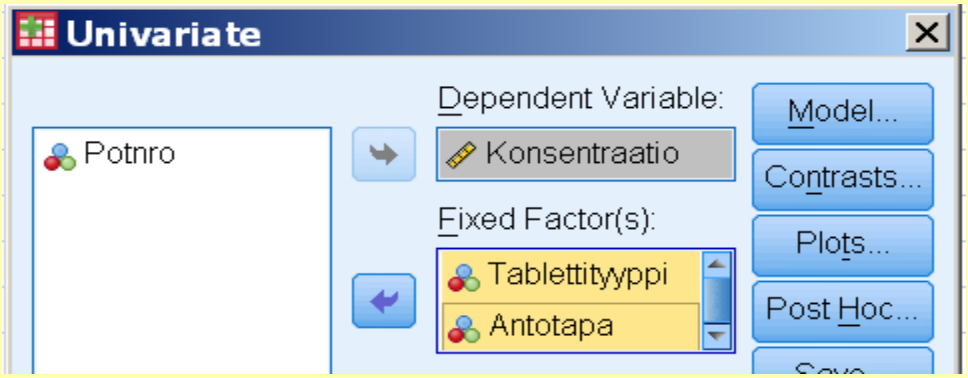
Kanta N: $t = \sqrt{\frac{2.89 - 2.95}{0.0597 \cdot (\frac{1}{6} + \frac{1}{8})}} = -0.45, P > 0.05$

Todetaan, että hoitoero saadaan vain hypertensiivisillä rotilla.

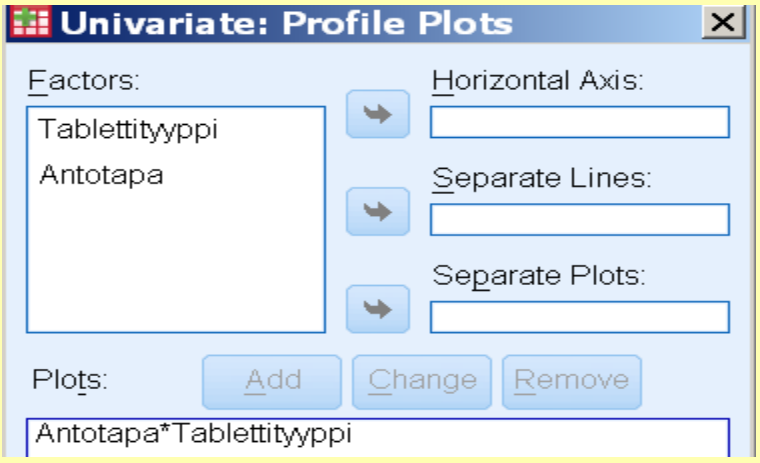
Esim.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Gastryl.sav
Yhdistelykoe 72 koehenkilöllä. Päällystetyyppejä on kolme A, B ja C. Samoin lääkkeenantotapoja: E, A ja M, missä E = tyhjään mahaan, A = antasidin kanssa ja M = aterian jälkeen. Halutaan testata onko päällystetyypin tai antotavan suhteen eroja ja onko päällystetyypin ja antotavan välillä interaktiota, ts. onko tablettityypillä eri vaikutus lopputulokseen riippuen antotavasta. Koejärjestely on seuraavanlainen: koehenkilöt on satunnaistettu siten, että kuhunkin päällystetyypin ja antotavan kombinaatioon tulee 8 koehenkilöä.

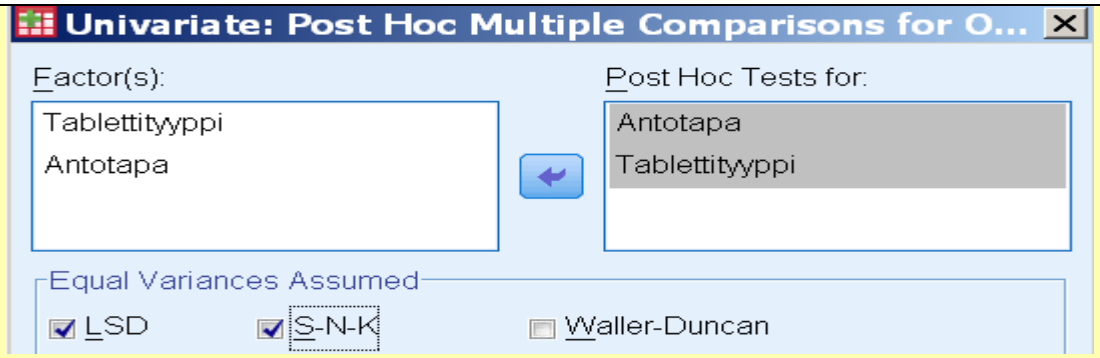
SPSS: Valikot **"Analyze" ►"General Linear Model"►"Univariate".**



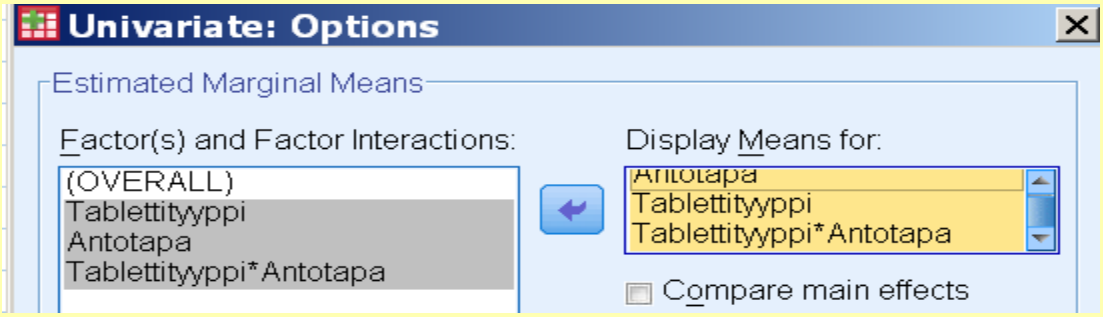
Kohdassa **"Plots"** Antotapa määritellään vaaka-akseliksi ja erilliset käyrät otetaan Tablettityypin mukaisesti ja klikataan kohtaa "Add":



Kohdassa **"Post Hoc"** tehdään valinnat:



Kohdassa **"Options"** tehdään valinnat:



Tests of Between-Subjects Effects

Dependent Variable:Konsentraatio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5074,694[a]	8	634,337	6,419	,000
Intercept	59110,681	1	59110,681	598,169	,000
Tablettityyppi	1971,694	2	985,847	9,976	,000
Antotapa	330,111	2	165,056	1,670	,196
Tablettityyppi * Antotapa	2772,889	4	693,222	7,015	,000
Error	6225,625	63	98,819		
Total	70411,000	72			
Corrected Total	11300,319	71			

a. R Squared = ,449 (Adjusted R Squared = ,379)

Tulkinta: Antotavan suhteen ei ole eroa. Tablettityypin suhteen on eroa, mutta koska interaktio on merkitsevä, niin molempien pääefektien, ts. antotavan ja tablettityypin, erojen tulkinta on ongelmallista.

1. Antotapa

Dependent Variable:Konsentraatio

Antotapa	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
E	25,625	2,029	21,570	29,680

A	30,208	2,029	26,153	34,263
M	30,125	2,029	26,070	34,180

2. Tablettityyppi

Dependent Variable:Konsentraatio

Tablettityyppi	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
A	21,583	2,029	17,528	25,638
B	30,292	2,029	26,237	34,347
C	34,083	2,029	30,028	38,138

3. Tablettityyppi * Antotapa

Dependent Variable:Konsentraatio

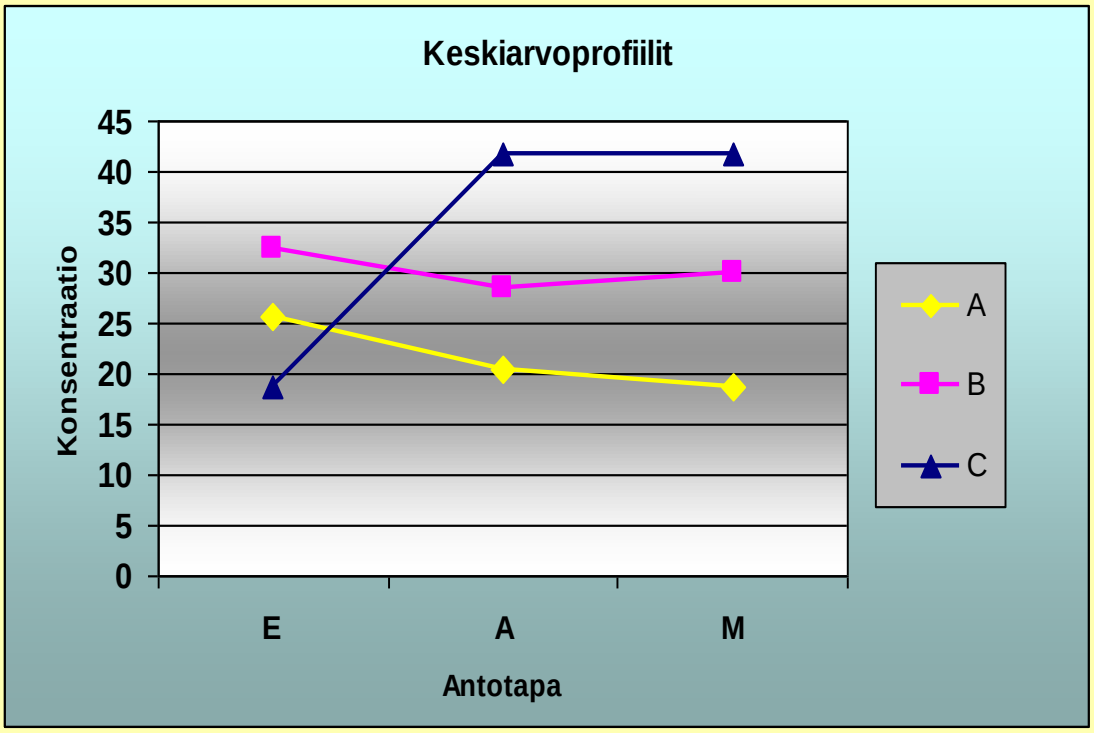
Tablettityyppi	Antotapa	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
A	E	25,750	3,515	18,727	32,773
	A	20,375	3,515	13,352	27,398
	M	18,625	3,515	11,602	25,648
B	E	32,375	3,515	25,352	39,398
	A	28,500	3,515	21,477	35,523
	M	30,000	3,515	22,977	37,023
C	E	18,750	3,515	11,727	25,773
	A	41,750	3,515	34,727	48,773
	M	41,750	3,515	34,727	48,773

Multiple Comparisons

Dependent Variable:Konsentraatio

	(I) Tablettityyppi	(J) Tablettityyppi	Mean Difference (I-J)	Std. Error	Sig.
LSD	A	B	-8,71*	2,870	,003
		C	-12,50*	2,870	,000
	B	A	8,71*	2,870	,003
		C	-3,79	2,870	,191

Based on observed means.
The error term is Mean Square(Error) = 98,819.
*. The mean difference is significant at the ,05 level.



Konsentraatio

Tablettityyppi		N	Subset 1	Subset 2
Student-Newman-Keuls[a,,b]	A	24	21,58	
	B	24		30,29
	C	24		34,08
	Sig.		1,000	,191

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 98,819.

a. Uses Harmonic Mean Sample Size = 24,000.

b. Alpha = ,05.

Tulkinta:

Tablettityypit B ja C muodostavat tablettityypistä A eroavan homogeenisen osajoukon (P=0,191>0,05). Tämä marginaalikeskiarvoihin perustuva tulkinta ei kuitenkaan ole interaktiosta johtuen kovin mielekäs (ks. kuva).

Toistomittausten varianssianalyysi

Toistomittausanalyysien (“Analysis of repeated measures”) perusideana on vähentää yksilöiden välisen vaihtelun vaikutusta itse tutkittavaan asiaan, esimerkiksi hoitojen vaikutukseen.

Toistomittausasetelmat ovat yleisesti ottaen voimakkaampia kuin tavanomaiset asetelmat, koska niissä yksilöiden välinen vaihtelu voidaan poistaa (eristää) analyysista (ks. Kuva) ja siten redusoida virhevaihtelua.

Toistoasetelmissa tullaan yleensä toimeen pienemmällä koehenkilö (tai -eläin) määrällä kuin asetelmissa, joissa ei ole toistoja. Tämä on tärkeää etenkin harvinaisten sairauksien tutkimisessa tai muuten tilanteissa, joissa koehenkilöitä on vaikea saada.

Kaaviokuva yksisuuntaisen toistomittausten varianssianalyysin vaihtelulähteistä:

Toistomittausasetelmissa tulokset kannattaa yleensä tallentaa seuraavanlaisen havaintomatriisin muotoon:

Henkilön numero	Ryhmä	Toistomittaus 1	…	Toistomittaus r
1 2 … N				

Huom. Erikoistapauksina tavalliset varianssianalyysit **ANOVA** ja **MANOVA**(monimuuttujavarianssianalyysi).

Ongelma-asettelu:

1. **Yksi ryhmä:** Tapahtuuko tarkastelun kohteena suureessa muutosta esim.ajan tai peräkkäin suoritettujen hoitojen suhteen?
2. **Useampia ryhmiä:** Ovatko muutosprofiilit erilaisia eri ryhmissä, ts. onko ryhmämuuttujan ja toistomuuttujan välillä interaktiota?

Testinä käytetään **F-testiä**, joka edellyttää teoriassa, että korrelaatiot ryhmien sisällä eri ajankohtien välillä ovat samat! Tämä on nk. **sfeerisyysoletus.** Sen paikkansa pitävyys tulisi testata ennen F-testiä esim. **Mauchlyn** testiä käyttäen. Käytännössä olettamus on harvoin voimassa ja siksi F-testin vapausasteita pitää yleensä korjata. Tilastopaketeissa on tähän tarkoitukseen mm. **Greenhouse-Geisserin** ja **Huynh-Feldtin** korjaukset. Näistä edellinen on konservatiivisempi ja siten jälkimmäinen on suositeltavampi.
Tilastopaketeissa esiintyvä käsite ”compaund symmetry” tarkoittaa käytännössä synonymi sanalle ”sphericity”.

Toistomittausten varianssianalyysin yhteydessä kannattaa suorittaa ennen analyysiä havaintoaineiston rakenteen huolellinen tarkastelu, jotta varmistuttaisiin menetelmän soveltuvuudesta, mm. ovatko kaikki edellä esitetyt varianssianalyysin yleiset käyttöedellytykset (normaalisuus, vakiovarianssius jne.) voimassa.
Tämän lisäksi kannattaa tulosten tulkinnan helpottamiseksi aina piirtää vertailtavien ryhmien keskiarvojen profiilit toistomittausten suhteen.

Mikäli aineisto on pieni, niin kannattaa piirtää myös yksittäisten havaintoyksiköiden (esim. koehenkilöiden) toistomittausten profiilit, jotta selvitettäisiin sisältyykö aineistoon profiililtaan huomattavan poikkeavia havaintoyksiköitä. Pienessä aineistossa näillä voi olla suuri vaikutus tuloksiin.

Toistomittausten varianssianalyysissä laskenta perustuu **ortogonaalisiin polynomeihin**.

Esim. Kolme toistomittausta y_1, y_2, y_3. Ortogonaaliset polynomit ovat:

$$p_0 = \frac{y_1 + y_2 + y_3}{\sqrt{3}}, p_1 = \frac{y_1 + 0 \cdot y_2 - y_3}{\sqrt{2}}, p_2 = \frac{y_1 - 2 \cdot y_2 + y_3}{\sqrt{6}}$$

Laskennassa p_0:a käytetään ryhmien välisten erojen testaamiseen, p_1:tä (lineaarinen komponentti) ja p_2:ta (neliöllinen komponentti) käytetään ryhmien sisäisten vaihtelujen ja interaktion testaamiseen. Ortogonaaliset polynomit (p_1, p_2) on konstruoitu siten, että kertoimien summa on 0 ja niiden neliöiden summa on 1.

Ortogonaalisten polynomien tulisi olla riippumattomia ja niiden tulisi olla symmetrisiä (kaikilla sama varianssi, sfeerisyysoletus). Mikäli symmetrisyysoletus ei toteudu, F-testiä voidaan käyttää pienentämällä vapausasteita korjaustekijällä ε. Esim. SPSS laskee kaksi eri arviota luvulle ε: Greenhouse-Geisserin ja Huynh-Feldtin korjaukset

Huom.

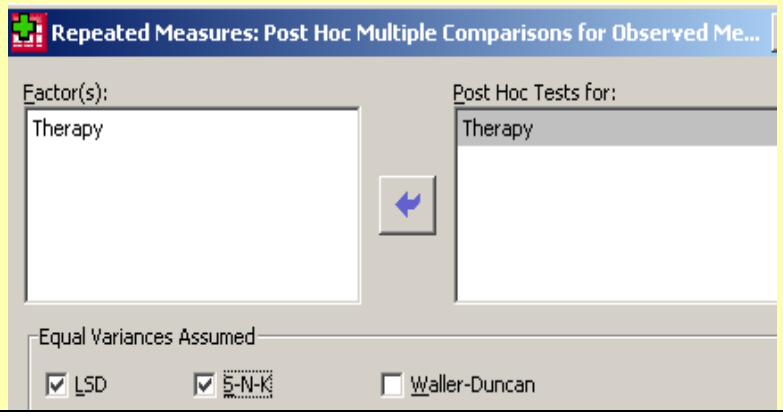
Tilastopaketeissa olevat sfeerisyystestit ovat herkkiä:
a) poikkeamille normaalisuudesta
b) poikkeaville havaintoarvoille ja
c) pienelle aineistokoolle

Esim.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

Kolmen hoidon **P**, **Q** ja **R** vertailu toistomittausten **”Imag_S”, ”Imag_1”** ja **”Imag_2”** suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja. Testattava hypoteesit ovat: **a)** onko hoitoryhmien keskiarvojen välillä eroa, **b)** onko ajan (Imag-muuttujan toistot) suhteen eroa ja **c)** onko hoitoryhmän ja ajan välinen interaktio merkitsevä, ts. onko Imag-muuttujan muutosprofiilit ryhmien välillä erilaiset.

SPSS: Valikot**"Analyze" ►"General Linear Model"►"Repeated Measures".** Määritellään toistomuuttujan nimi ja tasojen määrä: **"Repeat"►"Within Subject Factor Name", 3 ►"Number of levels"** ja klikataan kohta **"Add".** Määritellään toistomuuttujat "Imag_S", "Imag_1" ja "Imag_2" kohdassa **"Define"** ja ryhmämuuttuja: "Therapy"**►"Between Subjects Factor".** Valikossa **"Options"** kohdassa **"Factors and Factor Interactions",** valitaan vaihtoehto **"Therapy * Repeat",** joka viedään laatikkoon **"Display Means for".** Kohdassa **"Display",** klikataan vaihtoehdot **"Descriptive Statistics".** Valikossa **"Plots"** viedään "repeat"**►"Horizontal axis"** ja "therapy" **►"Separate lines".** Valikossa **"Post Hoc"** voidaan valita useita erilaisia monivertalutestejä: Tehdään tässä tapauksessa seuraavat valinnat:



Näin saadaan seuraavat tulokset:

Within-Subjects Factors

Measure:MEASURE_1

repeat	Dependent Variable
1	Imag_s
2	Imag_1
3	Imag_2

Between-Subjects Factors

		Value Label	N
Therapy	1	Group P	15
	2	Group Q	15
	3	Group R	18

Descriptive Statistics

Therapy		Mean	Std. Deviation	N
Imag_s	Group P	31,667	14,3830	15
	Group Q	31,847	15,3325	15
	Group R	35,556	14,2304	18
	Total	33,181	14,4344	48
Imag_1	Group P	31,887	19,1306	15
	Group Q	33,247	15,7997	15
	Group R	42,911	17,6594	18

	Total	36,446	17,9466	48
Imag_2	Group P	31,640	18,4036	15
	Group Q	41,393	18,5903	15
	Group R	53,911	27,7286	18
	Total	43,040	23,8650	48

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

		Within Subjects Effect
		repeat
Mauchly's W		,534
Approx. Chi-Square		27,588
df		2
Sig.		,000
Epsilon[a]	Greenhouse-Geisser	,682
	Huynh-Feldt	,728
	Lower-bound	,500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Therapy
Within Subjects Design: repeat

Tulkinta: Sfeerisyysoletus hylätään, koska P<0,05. Pitää käyttää korjattuja P-arvoja.

Tests of Within-Subjects Effects

Measure:Imag

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
repeat	Sphericity Assumed	2143,795	2	1071,898	26,988	,000
	Greenhouse-Geisser	2143,795	1,364	1571,192	26,988	,000
	Huynh-Feldt	2143,795	1,455	1473,322	26,988	,000
	Lower-bound	2143,795	1,000	2143,795	26,988	,000
repeat * Therapy	Sphericity Assumed	1448,920	4	362,230	9,120	,000
	Greenhouse-Geisser	1448,920	2,729	530,958	9,120	,000

	Huynh-Feldt	1448,920	2,910	497,885	9,120	,000
	Lower-bound	1448,920	2,000	724,460	9,120	,000
Error(repeat)	Sphericity Assumed	3574,518	90	39,717		
	Greenhouse-Geisser	3574,518	6,140E1	58,217		
	Huynh-Feldt	3574,518	6,548E1	54,591		
	Lower-bound	3574,518	4,500E1	79,434		

Tulkinta: Ajan suhteen merkitsevä muutos (kohta ”Repeat”) ja Ryhmien muutosprofiilit ajan suhteen ovat erilaiset, eli ryhmän ja ajan interaktio on tilastollisesti merkitsevä (kohta”Repeat*Therapy).

Tests of Within-Subjects Contrasts

Measure:imag

Source	repeat	Type III Sum of Squares	df	Mean Square	F	Sig.
repeat	Linear	2056,888	1	2056,888	32,347	,000
	Quadratic	86,907	1	86,907	5,484	,024
repeat * Therapy	Linear	1383,403	2	691,701	10,878	,000
	Quadratic	65,517	2	32,759	2,067	,138
Error(repeat)	Linear	2861,426	45	63,587		
	Quadratic	713,093	45	15,847		

Tulkinta:

Ryhmässä P ei tapahdu mitään muutosta, mutta ryhmissä Q ja R muutos on melko lineaarinen.

Estimated Marginal Means:

3. Therapy * repeat

Measure:imag

Therapy	repeat	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Group P	1	31,667	3,777	24,059	39,274
	2	31,887	4,541	22,740	41,033
	3	31,640	5,793	19,973	43,307
Group Q	1	31,847	3,777	24,239	39,454
	2	33,247	4,541	24,100	42,393
	3	41,393	5,793	29,726	53,061
Group R	1	35,556	3,448	28,611	42,500

	2	42,911	4,146	34,562	51,261
	3	53,911	5,288	43,260	64,562

Multiple Comparisons

Measure:imag

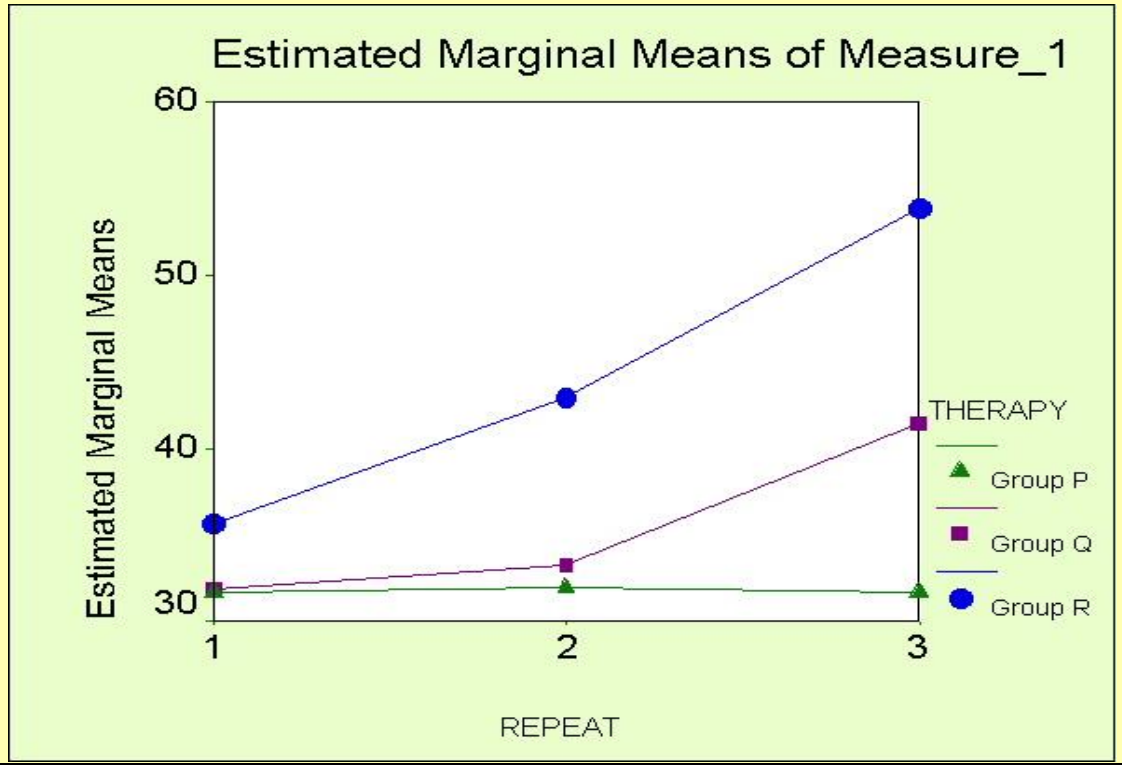
	(I) Therapy	(J) Therapy	Mean Difference (I-J)	Std. Error	Sig.
LSD	Group P	Group Q	-3,764	6,4885	,565
		Group R	-12,395	6,2122	,052
	Group Q	Group P	3,764	6,4885	,565
		Group R	-8,630	6,2122	,172
	Group R	Group P	12,395	6,2122	,052
		Group Q	8,630	6,2122	,172

Based on observed means.

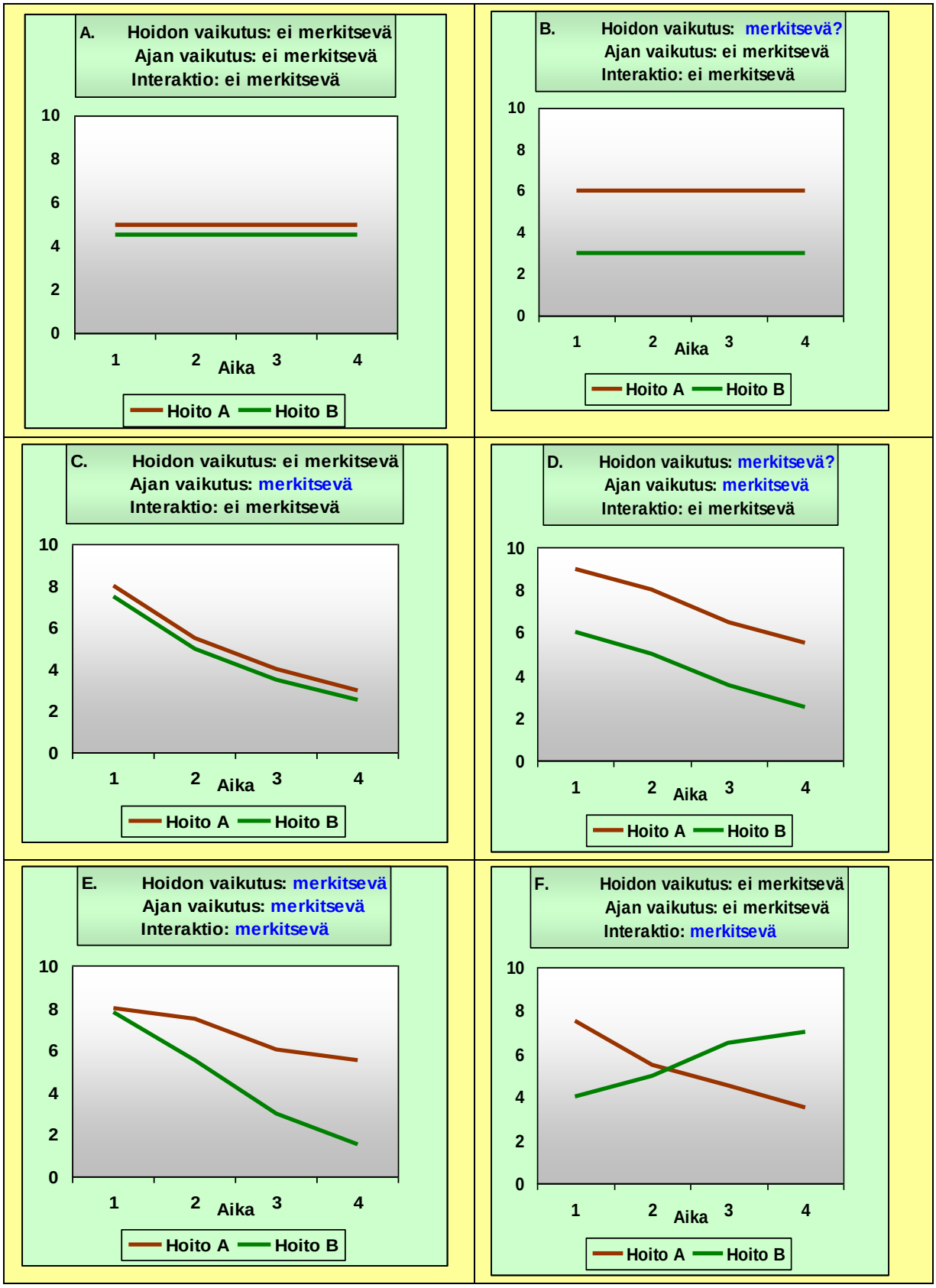
The error term is Mean Square(Error) = 315,752.

Huom. P-arvo ei ole merkitsevä, vaikka hoitoryhmän ja ajan välinen interaktio onkin erittäin merkitsevä. SPSS:llä ei voi laskea monivertailuita toistojen välillä.

Profile Plots:



Toistomittausanalyysin tulosten tulkintaa:



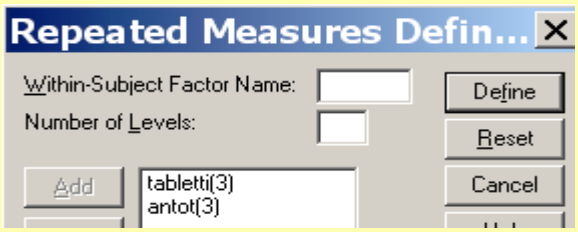
Esim.

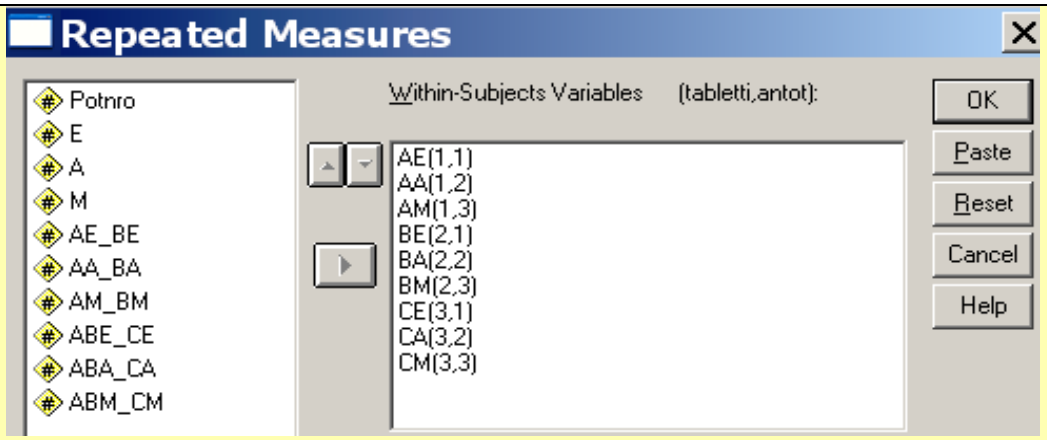
Kaksisuuntaiset toistot

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Gastryl.sav

Yhdistelykoe 8 koehenkilöllä. Jokainen koehenkilöistä testaa kaikkia kolmea päällystetyyppiä A, B ja C ja kullakin päällystetyypillä jokaista lääkkeenantotapaa E, A ja M, missä E = tyhjään mahaan, A = antasidin kanssa ja M = aterian jälkeen. Halutaan testata onko päällystetyypin tai antotavan suhteen eroja ja onko päällystetyypin ja antotavan välillä interaktiota, ts. onko tablettityypillä eri vaikutus lopputulokseen riippuen antotavasta.

SPSS: Valikot **"Analyze" ►"General Linear Model"►"Repeated Measures".**





General Linear Model

Within-Subjects Factors

Tablettityyppi	Antotapa	*Dependent Variable*
A	*E*	**AE**
	A	**AA**
	M	**AM**
B	*E*	**BE**
	A	**BA**
	M	**BM**
C	*E*	**CE**
	A	**CA**
	M	**CM**

Mauchly's Test of Sphericity(b)

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon(a) Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Tablettityyppi	,529	3,817	2	,148	,680	,787	,500
Antotapa	,705	2,096	2	,351	,772	,949	,500
Tabletti * Antot	,059	15,293	9	,096	,657	1,000	,250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.
(a) May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.
(b) Design: Intercept Within Subjects Design: tabletti+antot+tabletti*antot

Tulkinta: Kohdassa “Sig” kaikki P-arvot ovat >0,05, joten sfeerisyysoletus jää voimaan ja myöhempin testiitulosten P-arvoja ei tarvitse korjata.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Tablettityyppi	Sphericity Assumed	1971,694	2	985,847	3,471	,060
Error(tablettityyppi)	Sphericity Assumed	3976,306	14	284,022		
Antotapa	Sphericity Assumed	330,111	2	165,056	9,007	,003
Error(antotapa)	Sphericity Assumed	256,556	14	18,325		
Tablettityyppi * Antotapa	Sphericity Assumed	2772,889	4	693,222	53,066	,000
Error(tabletti*antot)	Sphericity Assumed	365,778	28	13,063		

Tests of Within-Subjects Contrasts

Source	Tablettityyppi	Antotapa	Type III Sum of Squares	df	Mean Square	F	Sig.
Tablettityyppi	Linear		1875,000	1	1875,000	11,975	,011
	Quadratic		96,694	1	96,694	,235	,643
Error(tabletti)	Linear		1096,000	7	156,571		
	Quadratic		2880,306	7	411,472		
Antotapa		Linear	243,000	1	243,000	23,958	,002
		Quadratic	87,111	1	87,111	3,286	,113
Error(Antotapa)		Linear	71,000	7	10,143		
		Quadratic	185,556	7	26,508		
Tablettityyppi * Antotapa	Linear	Linear	1815,031	1	1815,031	119,614	,000
		Quadratic	472,594	1	472,594	20,276	,003
	Quadratic						

	Linear	**283,594**	**1**	**283,594**	**27,512**	**,001**
	Quadratic	**201,670**	**1**	**201,670**	**58,222**	**,000**
*Error(tabletti*antot) Linear*	*Linear*	**106,219**	**7**	**15,174**		
	Quadratic	**163,156**	**7**	**23,308**		
	Linear	**72,156**	**7**	**10,308**		
	Quadratic	**24,247**	**7**	**3,464**		

Estimated Marginal Means

Tablettityyppi	*Mean*	*Std. Error*	*95% Confidence Interval*	
			Lower Bound	*Upper Bound*
A	**21,583**	**3,248**	**13,904**	**29,263**
B	**30,292**	**4,295**	**20,135**	**40,449**
C	**34,083**	**2,087**	**29,148**	**39,019**

Antotapa	*Mean*	*Std. Error*	*95% Confidence Interval*	
			Lower Bound	*Upper Bound*
E	**25,625**	**1,792**	**21,387**	**29,863**
A	**30,208**	**2,096**	**25,252**	**35,164**
M	**30,125**	**1,899**	**25,634**	**34,616**

Tulkinta: Lineaarinen kontrasti antaa tablettityyppien keskiarvoille merkitsevän P-arvon **0,011**, vaikka heterogeenisyystestillä kohdassa ”Within-Subjects Effects” saatu P-arvo **0,060** ei ollut merkitsevä.

Viitteet

Bland JM & **Altman** DG. Statistical methods for assessing agreement between two methods. Lancet 1986, I, 307-310.

Bland JM & **Altman** DG. Regression towards the mean. BMJ 1994; 308: 1499.

Bland JM & **Altman** DG. Measuring agreement in method comparison studies. Statistical methods in Medical Research 1999; 8: 135-160.

Milliken GA, Johnson DE. Analysis of messy data. Vol I. Chapman & Hall 1996. ISBN: 0412055414. Hinta £ 30.99 (www.crcpress.com)

Vickers AJ & **Altman** DG. Analysing controlled trials with baseline and follow-up measurements. BMJ 2001; 323: 1123- 4.

Kontingenssitaulukkoanalyysit

Tähän kappaleeseen on koottu tavallisimmat luokitellun tiedon käsittelyssä käytettävät perustestit. Monia niistä on käsitelty ja esimerkein havainnollistettu jo aiemmin tässä monisteessa.

Käyttötilanne:

Halutaan tutkia **luokiteltujen** (“categorical”) muuttujien välisiä yhteyksiä. Luokitellut muuttujat voivat olla joko laatueroasteikollisia (esim. HLA-tyyppi tai silmien väri) tai järjestysasteikollisia (esim. lääkeannos: matala, keskimääräinen, korkea tai vastaavasti esim. numeerisina arvoina ilmaistuna: 200 mg, 400 mg, 1600 mg).

Oletetaan, että muuttujat tarkasteltavat muuttujat x ja y ovat luokiteltuja siten, että x:ssä on R luokkaa ja y:ssä C luokkaa ja että muuttujien välisen yhteyden tutkimiseksi on muodostettu **R x C-kontingenssitaulukko** (“contigency table”):

		Muuttuja **y**				
		1	…	**j**	…	**C**
Muuttuja **x**	**1**					
	…					
	i			$\mathbf{f_{ij}}$		$\mathbf{m_j}$
	…					
	R	n_1	…	$\mathbf{n_j}$	…	N

Taulukossa f_{ij} tarkoittaa i. rivin ja j. sarakkeen frekvenssilukua. Tälle lokerofrekvenssille saadaan odotettu arvo kaavalla: $\mathbf{E(f_{ij}) = m_i \cdot n_j / N}$, kun oletetaan, että x:n ja y:n välillä ei olisi mitään riippuvuutta keskenään. Tämä merkitsee, että rivi-/sarakejakaumat eivät poikkea toisistaan, jolloin erot havaitussa frekvenssitaulukossa johtuvat sattumasta. Nimitys “kontingenssi” tarkoittaa sattumaa.

Huom.

Lokerokohtaiset odotusarvot ja havaittujen arvojen poikkeamat odotus-arvoista kannattaa laskea, sillä niiden perusteella voi alustavasti tarkastella x:n ja y:n välisen riippuvuuden luonnetta. Tilastopaketit tulostavat nämä suureet pyydettäessä.

Testattavat hypoteesit:

- Nollahypoteesi (H_0): “taulukon rivit ja sarakkeet eivät riipu toinen toisistaan”, ts. **x:llä ja y:llä** (riveillä ja sarakkeilla) **ei ole yhdysvaikutusta** (“interaction”) keskenään.
- Vaihtoehtoinen hypoteesi (H_A): “taulukon rivien ja sarakkeiden välillä on riippuvuutta”.

Tarkastellaan tilastopakettien tarjoamia testivaihtoehtoja seuraaviin tilanteisiin:

1. Molemmat R x C-taulukon muuttujista x ja y ovat laatueroasteikollisia,
2. Toinen muuttujista on laatueroasteikollinen ja toinen järjestysasteikollinen, jolloin kyseessä on yhteen suuntaan järjestetty taulukko,
3. Molemmat muuttujista x ja y ovat järjestysasteikollisia, eli taulukko on kahteen suuntaan järjestetty.

Tilastopaketit tarjoavat näihin testaustilanteisiin useita eri testivaihtoehtoja ja lisäksi yleensä kunkin vaihtoehdon sisällä on valittavana kolme eri tapaa P-arvon laskemiseksi: a) **asymptoottinen**, b) **Monte-Carlo** ja **c) eksakti** laskentatapa. Kaikki nämä laskentatavat tuottavat asymptoottisesti saman tuloksen.

Asymptoottinen menetelmä ”asymptotic method” on mikä tahansa menetelmä, joka perustuu approksimaatioon, eli likimäämäisarviointiin, Normaalijakaumalla tai jollain muulla todennäköisyysjakaumalla siten, että käytetty arvio tarkentuu, kun aineistokoko n kasvaa. Synonyymi asymptoottiselle menetelmälle on suurten otosten menetelmä ”large sample method” ja vaihtoehto on tarkka menetelmä ”exact method”.

Monte Carlo menetelmät ”Monte Carlo methods” ovat tietokonesimulaatiota käyttäviä ratkaisumenetelmiä matemaattisiin ja tilastollisiin ongelmiin. Simulaatiossa imitoidaan tilastollisia malleja satunnaislukujen avulla.

Laatueroasteikollinen R x C-taulukko

Käytettävissä olevat testit:

a) Pearsonin Khi^2 - heterogeenisuustesti:

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(f_{ij} - E(f_{ij}))^2}{E(f_{ij})}$$

Kaavassa f_{ij}:t ovat taulukon havaitut lukumäärät ja $E(f_{ij})$:t ovat niiden odotusarvot.

Huom.

Khi^2-testi on erittäin herkkä pienille odotusarvoille; Suhteellinen virhe tulee tällöin suureksi, koska odotusarvot ovat testisuureen nimittäjässä.

b) G^2- testi, uskottavuussuhde (”likelihood ratio”) testi:

$$G^2 = 2 \cdot \sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij} \log_e \left(\frac{f_{ij}}{E(f_{ij})}\right)$$

G^2-testi on yleisimmin käytetty testi taulukkoanalyyseissä (Esim. loglineaariset mallit). Se ei ole yhtä herkkä kuin Khi^2-testi pienille odotusarvoille log-muunnoksesta johtuen. Vapausasteiden määrä on df = (R - 1)·(C -1) – m, missä m = odotusarvojen laskemisessa tarvittavien estimoitujen parametrien määrä. Tavallisesti m = 0.

Sekä khi^2- että G^2-testiin voidaan tilastopaketeista kohdasta **”options”** pyytää myös lisäykset testisuureen arvoon lokeroittain:

$$\chi^2_{ij} = \frac{(f_{ij} - E(f_{ij}))^2}{E(f_{ij})} \quad \text{ja} \quad G^2_{ij} = 2 \cdot f_{ij} \log_e \frac{f_{ij}}{E(f_{ij})}$$

Näiden suureiden perusteella voi todeta lokerokohtaiset lisäykset testisuureen arvoon ja todeta missä lokerossa tai lokeroissa on eniten poikkeamaa riippumattomuusoletuksesta.

c) Yleistetty Fisherin eksakti testi, Freeman-Hamiltonin testi

Nimensä mukaisesti testi on yleistys Fisherin (2 x 2)- testille ja soveltuu käytettäväksi erityisesti pienillä aineistoilla.

Sekä **khi^2**- että **G^2**- testisuureet ovat aineiston koosta riippuvaisia eivätkä siten anna hyvää käsitystä riippuvuuden voimakkuudesta. Niiden perusteella voidaan kuitenkin

laskea normitettuja riippuvuuden mittoja; kontingenssikertoimia. **SPSS:**ssä polun **"Analyze"►"Descriptive statistics"►"Crosstabs"►"Statistics"** päästä löytyy mm. vaihtoehdot: **Kontingenssikerroin**, **Phi-kerroin** ja **Cramerin V**. Kaikkien näiden kerrointen vaihtelualue on: 0 – 1 (ei assosiaatiota – täydellinen assosiaatio). Niiden avulla voidaan verrata mm. assosiaation voimakkuutta eri dimensioisissa (R x C)-taulukoissa, mutta lääketieteellisissä julkaisuissa niitä käytetään erittäin harvoin.

Kun tarkastellaan x:n ja y:n riippuvuuden luonnetta lähemmin, syntyy usein tarve vertailla pareittain keskenään (R x C)-taulukon perusteella laskettuja suhdelukuja.

		Muuttuja y				
		1	…	**r**	…	**C**
Muuttuja x	**1**					
	…					
	c			$\mathbf{f_{cr}}$		$\mathbf{m_c}$
	…					
	d			$\mathbf{f_{dr}}$		m_d
	…					
	R	n_1	…	$\mathbf{n_r}$	…	N

Esim.

Tekijä x voisi olla hoitoryhmä ja y ilmaisisi eri lopputuloksia (laatueroasteikolla) Oletetaan, että halutaan verrata sarakkeen r riveillä c ja d olevia suhdelukuja:

$\hat{\mathbf{p}}_{\mathbf{cr}} = \frac{\mathbf{f_{cr}}}{\mathbf{m_{cr}}}$ ja $\hat{\mathbf{p}}_{\mathbf{dr}} = \frac{\mathbf{f_{dr}}}{\mathbf{m_{dr}}}$ keskenään ja testattavat hypoteesit ovat:

$\mathbf{H_0 : p_{cr} = p_{dr}}$ (suhdeluvut perusjoukossa ovat yhtä suuret) ja $\mathbf{H_A : p_{cr} \neq p_{dr}}$ (suhdeluvut perusjoukossa eroavat toisistaan)

Scheffen monivertailumenettelyyn perustuva erotuksen $p_{cr} - p_{dr}$ $100\cdot(1 - \alpha)$ %:n luottamusväli saadaan kaavalla:

$$(\hat{p}_{cr} - \hat{p}_{dr}) \pm \sqrt{\chi^2_{1-\alpha}(df)} \cdot SE(\hat{p}_{cr} - \hat{p}_{dr}),\ \text{missä } df=(R-1)\cdot(C-1) \text{ ja}$$

$$SE(\hat{p}_{cr} - \hat{p}_{dr}) = \sqrt{\frac{\hat{p}_{cr}(1-\hat{p}_{cr})}{m_{cr}} + \frac{\hat{p}_{dr}(1-\hat{p}_{dr})}{m_{dr}} -}$$

Huom.

$\mathbf{SE(\hat{p}_{cr} - \hat{p}_{dr})}$ on tavallinen kahden suhdeluvun erotuksen keskivirhe ja $\sqrt{\chi^2} = z$, eli standardinormaalijakauman prosenttipiste. Ero on vapausasteissa verrattuna tavalliseen suhdelukujen erotuksen luottamusväliin.

Suhdelukujen erotuksen monivertailukorjatun välin voi laskea (esim. CIA:lla) siten, että ensin lasketaan tavanomainen väli ja kasvatetaan sitten välin molempia puoliskoja korjaustekijällä: $\frac{\sqrt{\chi^2_{1-\alpha}(df)}}{z_{1-\alpha/2}}$, missä df = (R – 1) (C – 1)

Testi:

Jos nolla ei kuulu laskettuun luottamusväliin, tehdään johtopäätös: $p_{cr} \neq p_{dr}$, eli suhdeluvut ovat erisuuret ja P< α.

Yhteen suuntaan järjestetty R x C-taulukko

Oletetaan, että sarakemuuttuja (esim. lopputulosmuuttuja) y on järjestysasteikollinen, mutta rivimuuttuja on laatueroasteikollinen (esim. hoitoryhmä). Rivien ja sarakkeiden välisen yhdysvaikutuksen testaamiseksi tilastopaketeista (esim. StatXact 8) löytyy seuraavat testit:

a) Kruskal-Wallisin testi

- On yksi suosituimmista parametrittomista testeistä. Se on yleistys Wilcoxon-Mann-Whitneyn testille. Testin voima on noin 98 % verrattuna yksisuuntaiseen Anovaan, mikäli sitä käytetään normaalijakaumaa noudattavalle jatkuvalle muuttujalle.

b) Normaalijakaumaan perustuva pistemäärätesti “Normal Scores test”

- On vaihtoehto Kruskal-Wallisin testille. Testiä kannattaa käyttää mikäli muuttuja y noudattaa normaalijakaumaa ja on luokiteltu C:hen luokkaan.

c) Savagen pistemäärätesti (“Savage Scores test”)

- On vaihtoehto Kruskal-Wallisin testille. Testiä kannattaa käyttää mikäli muuttuja y noudattaa eksponentiaalijakaumaa ja on luokiteltu C:hen luokkaan.

d) ANOVA mielivaltaisin pistemäärin

- Yleistesti, ANOVAN parametriton vaihtoehto tilanteisiin, jolloin muuttujan y arvoja halutaan painottaa toisella tavalla kuin edellä mainituissa testeissä ja muuttuja on luokiteltu C:hen luokkaan.

Riippumatta painokertoimien valinnasta kaikkien näiden testien testisuure noudattaa asymptoottisesti khi^2-jakaumaa vapausastein R-1.

Kahteen suuntaan järjestetty R x C-taulukko

Oletetaan, että sekä sarakemuuttuja (esim. lopputulosmuuttuja) y ja rivimuuttuja ovat järjestysasteikollisia (esim. hoitoryhmä siten, että potilaita on hoidettu eri lääkeannoksin).

Rivien ja sarakkeiden välisen yhdysvaikutuksen testaamiseksi tilastopaketeista (esim. StatXact 8, SPSS 18) löytyy seuraavat testit:

a) Jonckheere-Terpstra testi

- Testi on Kruskal-Wallisin trenditesti.

b) "Linear by linear"-assosiaatiotesti

- On vaihtoehto Jonckheere-Terpstra-testille.

Ei ole käytettävissä tutkimustuloksia siitä, kumpi testeistä olisi parempi tai huonompi eri tilanteissa. Molempien testisuureet noudattavat asymptoottisesti khii^2-jakaumaa vapausastein yksi. Testivaihtoehto b) on joustavampi painokerrointen antamisen suhteen, rivi- ja sarakemuuttujien arvoille voidaan antaa mielivaltaiset painokertoimet.

c) Gamma (Goodman ja Kruskal)

- Gamma on kahden järjestysasteikollisen suureen välinen assosiaatiomitta, joka kertoo kuinka paljon todennäköisempää on saada sama kuin eri tulos. Se on vaihtoehto Kendallin taulle ja Somerin D:lle.

Esim.

Tupakointi- ja keuhkosyöpätutkimus (Azen et al. 1977).
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/FEV1_DeltaN.sav

Tutkimuksessa vertailtiin kahdeksaa eri keuhkofunktioindeksiä keskenään; ts. missä määrin ne mittaavat samaa asiaa. FEV_1:n ja ΔN_2:n väliselle saatiin seuraava tulos:

Tulos		ΔN2			
		Huono	Kohtalainen	Hyvä	Erinomainen
FEV1	Huono	8	5	3	3
	Kohtalainen	0	8	1	0
	Hyvä/Erinomainen	0	4	14	4

Gamma-suureeksi tulee **0.6122** ja 95 % luottamusväliksi **(0.3162, 0.9082)** sekä eksaktiksi P-arvoksi 0.0004 (StatXact 6). Jonckheere-Terpstra ja linear by linear -testit antavat P-arvoiksi vastaavasti **0.0002** ja **0.0007.**

Viitteet

Azen SP, **Linn** W et al. A Comparison of eight lung function indices in smoking and non-smoking officeworkers. Lung 1977;154:213-221,.

Loglineaariset mallit

Käyttötilanne:

Loglineaarisia malleja käytetään silloin, kun halutaan tutkia riippuvuussuhteita (yhdysvaikutuksia) useiden muuttujien x, y, z… välillä moniulotteisissa taulukoissa. Tavoitteena on konstruoida malli, joka kuvaa näitä riippuvuussuhteita mahdollisimman hyvin ja testata malliin liittyviä hypoteeseja.

Käsitteitä

Interaktio:

eli yhdysvaikutus on riippuvuus kahden tai useamman tekijän välillä.
1. asteen, 2. asteen, jne. interaktio on sama kuin kahden tekijän, kolmen tekijän, jne. välinen interaktio.

Synergismi (antagonismi):

Eri tekijöiden samanaikainen vaikutus, joka on suurempi(pienempi) kuin niiden yksittäisten vaikutusten summa

Loglineaarinen malli:

Malli, joka olettaa, että moniulotteisessa taulukossa odotettujen frekvenssien $E(f_{ijk})$ logaritmi voidaan esittää useiden parametrien lineaarisena yhdistelynä (kombinaationa). Tästä johtuu nimitys loglineaarinen malli.

Esim.

Luokiteltujen muuttujien A, B ja C määrittelemä kolmiulotteinen taulukko, jonka indeksit ovat i, j, k. Merkitään taulukon lokerofrekvenssien odotusarvoja: $F_{ijk} = E(f_{ijk})$. Tähän taulukkoon liittyvä loglineaarinen malli on muotoa:

$$\log_e(F_{ijk}) = \theta + \underbrace{\lambda_i^A + \lambda_i^B + \lambda_i^C}_{(1)} + \underbrace{\lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}}_{(2)} + \underbrace{\lambda_{ijk}^{ABC}}_{(3)}$$

θ on yleiskeskiarvo, λ: t ovat nk. efektejä; (1) = pääefektit, (2) = 2. kertaluvun efektit, (3) = 3. kertaluvun efektit

Esim.

λ^A on pelkästään A:sta johtuva efekti (A:n vaikutus ennustettavaan muuttujaan) ja λ^{AB} on A:n ja B:n yhdysvaikutus (interaktio).

Kyllästetty malli, eli saturoitu malli:

Malli, joka sisältää kaikki mahdolliset efektit (edellä oleva esimerkki).

Hierarkinen malli:

Malli, jossa ei voi esiintyä korkeamman kertaluvun vaikutuksia ilman, että vastaavat alemman kertaluvun vaikutukset ovat mukana mallissa.

Esim.

Jos λ^{ABC} on mallissa ja on nollasta poikkeava, niin myös λ^{AB}, λ^{AC}, λ^{BC}, λ^A, λ^B, λ^C ja θ ovat mukana mallissa. Jos λ^{AB} on mallissa, niin λ^A, λ^B ja θ ovat mukana mallissa, muttei λ^C. (ellei mallina ole $\lambda^{AB} + \lambda^C$)

Mallin parametrin skaalaus

Jotta loglineaarisen mallin parametrit voitaisiin yksikäsitteisesti arvioida, täytyy kunkin tekijän parametreista yhtä pitää laskennassa ns. **häiriöparametrina**, “nuisance parameter”. Tämä merkitsee, että parametreille asetetaan **reunaehtoja**. Kunkin tekijän parametrien summa asetetaan joko nollaksi tai ykköseksi (skaalausparametri):

$$\sum_i \lambda_i^A = 0 \text{ tai } 1, \quad \sum_{i,j} \lambda_{ij}^{AB} = 0 \text{ tai } 1, \quad \sum_{i,j,k} \lambda_{ijk}^{ABC} = 0 \text{ tai } 1$$

Yllä olevissa ehdoissa summaus suoritetaan yli kaikkien summan termien.
Monet tilastopaketeista (esim. SPSS, Systat) käyttävät skaalausparametrina nollaa. Useimmat epidemiologiset tilastopaketit (esim. GLIM) suorittavat skaalauksen ykköseen. Tästä on etuna se, että mallin kerroinestimaateista on helppo laskea arvio ristitulosuhteelle (OR). Mikäli skaalausparametri on nolla, niin OR:ien laskenta on erittäin työlästä loglineaarisen analyysin tulosten perusteella. Lievänä haittana siitä, että skaalausparametri on 1, seuraa että kuhunkin tekijään tai interaktioon liittyvistä parametreista yksi jää tulostumatta. GLIM jättää ensimmäisen parametrin pois. Se voidaan tarvittaessa kuitenkin helposti laskea skaalausehdon perusteella.
Pääosa tässä luvussa käsiteltävien esimerkkitapausten tuloksista on laskettu GLIM-ohjelmalla.

Yhteensopivuustestit

Yhteensopivuustesti mittaa havaittujen frekvenssien ja mallin perusteella laskettujen odotettujen frekvenssien välistä yhtäpitävyyttä. Loglineaaristen mallien yhteydessä tavallisimmin käytettyjä yhteensopivuustestejä ovat **Pearsonin khi^2-testi** ja **G^2-testi** (ks. edellinen luku). Näistä jälkimmäinen perustuu uskottavuussuhteeseen (“likelihood ratio”) ja sen laskentakaava on muotoa:

$$G^2 = 2 \cdot \sum O \cdot \log_e\left(\frac{O}{E}\right)$$

Kaavassa O ja E ovat havaittuja ja odotettuja frekvenssejä.

Testisuureet kolmiulotteisessa taulukossa:

$$\underbrace{\chi^2 = \sum_{i,j,k} \frac{(f_{ijk} - F_{ijk})^2}{F_{ijk}}}_{(1)} \text{ ja } \underbrace{G^2 = 2 \cdot \sum_{i,j,k} f_{ijk} \log_e\left(\frac{f_{ijk}}{F_{ijk}}\right)}_{(2)}$$

Kummatkin näistä suureista noudattavat likimain khi^2 -jakaumaa vapausastein n – p, missä n on solujen lukumäärä ja p on estimoitavien riippumattomien parametrien määrä.

Huom. Useampiulotteisessa taulukossa indeksien i, j ja k lukumäärä vain lisääntyy.

Malleja **M_1** ja **M_2** sanotaan **kytketyiksi** (”nested”), jos kaikki malliin M_1 sisältyvät λ:t ovat osajoukko malliin M_2 sisältyvistä λ:oista.

Poikkeama D (“Deviance D”) on mitta, jonka avulla loglineaarisissa malleissa voidaan arvioida kuinka paljon tietty malli **M** poikkeaa havaintoaineistoon sovitetusta saturoidusta mallista M_S. Se lasketaan näihin malleihin liittyvien uskottavuussuhteiden **L** ja L_S perusteella kaavalla: $D = -2 \cdot (\log_e L - \log_e L_S)$.

D saa suuren arvon, kun L on pieni suhteessa L_S:ään. Tämä merkitsee, että malli M on huono. Mikäli malli on hyvä, D saa pienen arvon. Paitsi mallin vertaamista saturoituun malliin, D:n avulla voidaan kätevästi verrata mitä tahansa kahta mallia keskenään, sillä D noudattaa asymptoottisesti khi^2-jakaumaa siten, että vapausasteiden määrä on vertailtavien mallien parametrien erotus.

Yhteensopivuuden (vaikutusten), **muutosta** mallien M_1 ja M_2 välillä voidaan testata suureella: $\mathbf{D = G^2_{M_1} - G^2_{M_2}}$, joka noudattaa likimain khi^2- jakaumaa vapausastein $\mathbf{df_{M_1} - df_{M_2}}$.

Huom. Testisuureella (1), ei ole tätä ominaisuutta.

Analysointivaiheet

Loglineaarinen analyysi sisältää vaiheet:

1. Sopivan mallin etsintä
2. Erilaisten hypoteesien testaaminen
3. Taulukon solujen ja ositteiden lähempi tarkastelu havaituissa ja odotetuissa frekvensseissä esiintyvien poikkeavuuksien suhteen (esim. trendit)

Standardoitujen poikkeamien (residuaalien) tutkiminen

Kolmiulotteisessa taulukossa standardoidut poikkeamat (residuaalit) lasketaan kaavalla:

$$\chi_{ijk} = \frac{f_{ijk} - E(f_{ijk})}{\sqrt{E(f_{ijk})}}$$

Jos mallin yhteensopivuus on hyvä, niin standardoidut poikkeamat ovat pieniä. Koska suure χ_{ijk} noudattaa likimain normaalijakaumaa siten, että odotusarvo on nolla ja hajonta yksi, niin residuaalien tarkasteluun saadaan

nyrkkisääntö: Ehdon $|\chi_{ijk}| \leq 2$ tulisi olla voimassa jokaisella arvolla i,j,k, jotta mallin yhteensopivuus havaintoaineiston kanssa olisi hyvä.

Esimerkkejä

Esim. 1:

(2 x 2)-taulukon loglineaarinen analyysi. **Alkoholin käytön ja ruokatorven syövän välinen yhteys.** Breslow & Day, 1980, s. 124.
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/BDEsim1.sav

Kyseessä oli tapaus-verrokkitutkimus. Aineistossa alkoholin keskimääräisen *kulutuksen mediaani oli 80g /vrk per henkilö. Käyttäen tätä arvoa* katkaisukohtana tulokseksi saatiin taulukko:

	Tekijä B: Keskimääräinen alkoholin kulutus vrk:ssa	
Tekijä A:	>80g	≤80g
Tapaukset	96	104
Verrokit	109	666

Malli 1: $\log_e(F_{ij}) = \theta + \lambda_i^A + \lambda_j^B$

Tämä malli testaa hypoteesia: $E(f_{ij}) = F_{ij} = \lambda_i^A \cdot \lambda_i^A$ eli, että tekijät A ja B eivät riipu toisistaan. Asetetaan vaatimus: $\lambda_1^A = \lambda_1^B = 0$ estimoitavien parametrien yksikäsitteisyyden vuoksi. Sovitetuiksi lukumääriksi tulee tällöin:

	Tekijä B: Keskimääräinen alkoholin kulutus vrk:ssa	
Tekijä A:	>80g	≤80g
Tapaukset	$\exp(\hat{\theta})$	$\exp(\hat{\theta} + \hat{\lambda}_2^B)$
Verrokit	$\exp(\hat{\theta} + \hat{\lambda}_2^A)$	$\exp(\hat{\theta} + \hat{\lambda}_2^A + \hat{\lambda}_2^B)$

Huom.

Molemmille riveille sovitettujen lukumäärien logaritmien erotus on $-\lambda_2^B$.

Vastaavasti molemmille sarakkeille erotus on $-\lambda_2^A$. Voidaan helposti todeta, että ristitulosuhde/suhteellinen riski (OR) on siten 1.

Yhteensopivuustestit mallille 1: $\chi_1^2 = 110.3, P < 0.0001, G_{M_1}^2 = 96.4, df_{M_1} = 1$

Tulkinta: Yhteensopivuus on huono, koska khi^2-arvo on suuri.

Mallin 1 parametrien arvioidut kertoimet (estimaatit) ja niiden keskivirheet ovat:

Estimaatti	Keskivirhe SE(estimaatti)	Parametri
3.739	0.0941	θ
1.355	0.0793	λ_2^A
1.323	0.0786	λ_2^B

Havaitut ja mallin 1 perusteella sovitetut frekvenssit ja residuaalit ovat:

Havaittu frekvenssi	Sovitettu frekvenssi	Residuaali
96	42.0	8.319
104	157.9	- 4.293
109	162.9	- 4.226

666	612.1	2.181

Tulkinta: Yhteensopivuus on huono, koska residuaalit ovat suuria

Malli 2: $\log_e(F_{ij}) = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

Asetetaan vaatimus: $\lambda_1^A = \lambda_1^B = \lambda_{11}^{AB} = \lambda_{12}^{AB} = \lambda_{21}^{AB} = 0$. Mallin 2 mukaisesti sovitetuiksi lukumääriksi tulee tällöin:

	Tekijä B: Keskimääräinen alkoholin kulutus vrk:ssa	
Tekijä A: Tapaukset Verrokit	>80g	≤80g
	$\exp(\hat{\theta})$	$\exp(\hat{\theta} + \hat{\lambda}_2^B)$
	$\exp(\hat{\theta} + \hat{\lambda}_2^A)$	$\exp(\hat{\theta} + \hat{\lambda}_2^A + \hat{\lambda}_2^B + \hat{\lambda}_{22}^{AB})$

Voidaan helposti todeta, että OR:n arvioksi tämän mallin 2 perusteella saadaan: $\exp(\lambda_{22}^{AB})$. Yhteensopivuustestit mallille 2: $G^2_{M_2} = 0, df_{M_2} = 0$, eli kyseessä on saturoitu malli.

Mallin 2 parametrien arvioidut kertoimet (estimaatit) ja niiden keskivirheet:

Estimaatti	SE(estimaatti)	Parametri
4.564	0.1021	θ
0.127	0.1400	λ_2^A
0.080	0.1415	λ_2^B
1.730	0.1752	λ_{22}^{AB}

Koska taulukon luvut on laskettu ohjelmalla GLIM, missä skaalausparametri on yksi, niin AB-tekijän estimaatin **1.730** perusteella saadaan OR:lle arvio **exp(1.730)=5.640.** OR:n normaalijakaumaan perustuva 95 %:n luottamusväli ($CI_{95\,\%}$) saadaan kaavalla: exp(estimaatti-1.96·SE(estimaatti), estimaatti +1.96·SE(estimaatti)), eli tässä tapauksessa $CI_{95\,\%}$ = (exp(1.730-1.96·0.1752), exp(1.730+1.96·0.1752) = **(4.001,7.951).**

Esim. 2:

Loglineaarinen analyysi useista (2 x 2)-taulukoista, sekoittavan tekijän kontrollointi osittamalla. Edellä oleva aineisto iän (sekoittava tekijä) mukaan ositettuna.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/BDEsim2.sav

Ikäryhmät	**Tapaukset**		**Verrokit**		**OR**
	> 80 g	**≤ 80 g**	**> 80 g**	**≤ 80 g**	
25-34	1	0	9	106	0.000
35-44	4	5	26	164	5.046
45-54	25	21	29	138	5.665
55-64	42	34	27	139	6.006
65-74	19	36	18	88	2.580
75+	5	8	0	31	0.000

Huom.

Esimerkkiaineistossa tapausten ja verrokkien ikäjakaumat ovat erilaiset. Jos ikä korreloi myös päivittäiseen alkoholin käyttöön, niin kyseessä on sekoittava tekijä ko. aineistossa.

Taulukkoon on laskettu ikäryhmäkohtaiset OR:t. Kaksi OR:ää tulee nollaksi, koska yksi frekvensseistä on kummassakin nolla. OR:ien homogeenisuus kannattaa testata. Tulokseksi saadaan (StatXact 6) P=0.0992 > 0.05, joten homogeenisuusoletus jää voimaan.

Merkitään:

- **A** = ikäryhmä ("**A**ge group"), koodaus 1-6,
- **E** = altistus ("**E**xposure"), koodaus: 1 = "> 80 g", 2 = "≤ 80 g"
- **D** = ryhmä ("**D**isease"), koodaus: 1 = "tapaus", 2 = "verrokki"

Sovitetaan havaintoaineistoon aluksi malli, josta puuttuu taudin ja altisteen välinen interaktio.

Malli 1: $\log_e(F_{ij}) = \theta + \lambda_i^D + \lambda_j^E + \lambda_k^A + \lambda_{ik}^{DA} + \lambda_{jk}^{EA}$

Merkitään tätä hierarkista mallia: **M1 = (DA, EA)**. Tämä malli testaa otantakehikon rajoituksia ts. epätasapainoa reunasummissa. Yhteensopivuustestin tulokseksi saadaan:

$$G^2_{M_1} = 90.56, df_{M_1} = 6, P < 0.0001$$

Huom.

Jos lasketaan standardoidut poikkeamat nelikenttien soluille (1,1) (=vasen yläkulma tai sarake 2 yllä olevassa taulukossa), niin ne ovat tässä tapauksessa suuria (3.11, 2.27, 3.91, 4.25, 1.79, 2.90).

Tulkinta:

Otantaharha ei selitä vaihtelua, koska yhteensopivuus on huono.

Seuraavaksi sovitetaan malli **M2 = (DE, DA, EA)**, joka sisältää kaikki toisen kertaluvun interaktiotermit.

Malli 2: $\log_e(F_{ij}) = \theta + \lambda_i^D + \lambda_j^E + \lambda_k^A + \lambda_{ik}^{DA} + \lambda_{jk}^{EA} + \lambda_{ij}^{DE}$

Tulokset: $G^2_{M_2} = 11.04, df_{M_2} = 5, P < 0.0506$

Standardoidut poikkeamat ovat tässä mallissa pieniä (1.190, -0.031, 0.134, 0.238, -0.949 ja 1.019).

Mallien M1 ja M2 ero voidaan testata seuraavasti:

$$G^2_{M_1} - G^2_{M_2} = 90.56 - 11.04 = 80.52, \quad df_{M_1} - df_{M_2} = 6 - 5 = 1, \quad P < 0.0001$$

Tulkinta:

Mallien M1 ja M2 ero on tilastollisesti merkitsevä, mikä merkitsee, että taudin ja altisteen välinen yhdysvaikutus DE on merkitsevä.

Tilastopaketit tarjoavat G^2-suureen ohella monia muitakin kriteereitä, joilla mallin hyvyyttä ja yhteensopivuutta havaintoaineiston kanssa voidaan arvioida, esim. SYSTAT-ohjelmassa on **"Raftery's BIC"** (**B**ayesian **I**nformation **C**riterion), **"Dissimilarity"-indeksi** ja **poikkeavien lokerofrekvenssien askeltava etsintämenettely.**

Jos BIC antaa negatiivisen arvon, niin se merkitsee, että malli on suositeltavampi kuin saturoitu malli ja kun vertaillaan vaihtoehtoisia malleja, niin kannattaa valita se, jonka BIC on alhaisin.

Esim. Mallille M2 saadaan BIC= -23.57, joten saturoitua mallia (DEA) ei tarvita.

"Dissimilarity"-indeksi kuvastaa, kuinka paljon taulukoidussa aineistossa tarvitsisi tehdä frekvenssilukujen uudelleen sijoitteluja, jotta havaitut frekvenssiluvut saataisiin yhteensopiviksi odotettujen kanssa.

Esim.

Mallille M2 indeksiksi tulee 1.9, mikä merkitsee, että tarvittavien uudelleensijoittelujen määrä on 1.9 %.

Poikkeavien lokerofrekvenssien askeltava etsintä suoritetaan siten, että ensin etsitään eniten poikkeava lokero ja se tulkitaan ns. rakenteelliseksi nollaksi. Tämä tarkoittaa sitä, että seuraavassa askeleessa tämä lokero jätetään huomioimatta ja malli sovitetaan jäljellä oleviin lokeroihin. Jälleen etsitään eniten poikkeava lokero jne.

Esim.

Sovitetusta mallista M2 kolme eniten poikkeavaa tekijöiden A ("Age"), D ("Disease") ja E ("Exposure") koodiarvo/indeksikombinaatiota ovat:

"Age"	"Disease"	"Exposure"	Frekvenssi	LR-khi^2	P-arvo
6	2	1	0	5.112	0.024
5	2	1	18	8.270	0.004
1	1	2	0	2.151	0.142

Malli M2 voidaan tässä tapauksessa tulkita lopulliseksi ja sen eri efektien estimaateiksi GLIM-ohjelmalla saadaan (skaalausparametri = 1):

Para-metri (1)	Estimaatti $\hat{\lambda}$ (2)	Keski-virhe **SE($\hat{\lambda}$)** (3)	Testi-suure $\frac{\hat{\lambda}}{\mathbf{SE}(\hat{\lambda})}$ (4)	Multiplikatii-vinen efekti **exp($\hat{\lambda}$)** (5)	Efektit (6)
θ	**-1.115**	1.028	-1.805	0.328	Yleiskeskiarvo
λ_2^A	2.527	1.080	2.340	12.526	Ikäryhmä
λ_3^A	4.314	1.036	4.164	74.739	
λ_4^A	4.807	1.032	4.658	**122.36**	
λ_5^A	4.282	1.038	4.266	72.385	
λ_6^A	**2.279**	1.123	2.029	9.767	
λ_2^D	**3.384**	1.012	3.343	29.489	Ryhmä
λ_2^E	**0.7179**	0.3799	1.890	2.050	Altistus
λ_{22}^{DA}	-1.542	1.062	-1.425	0.214	Toisen kerta-luvun interaktiot ryhmän ja iän välillä
λ_{23}^{DA}	-3.199	1.019	-3.139	0.041	
λ_{24}^{DA}	-3.713	1.014	-3.662	0.024	

λ_{25}^{DA}	-3.967	1.019	-3.893	0.019	
λ_{26}^{DA}	**-3.962**	1.061	-3.734	0.019	
λ_{22}^{EA}	-0.5414	0.3885	-1.394	0.582	Toisen kerta-luvun interaktiot altisteen ja iän välillä
λ_{23}^{EA}	-0.8486	0.3758	-2.258	0.428	
λ_{24}^{EA}	-0.8299	0.3739	-2.220	0.436	
λ_{25}^{EA}	-0.4428	0.3993	-1.109	0.642	
λ_{26}^{EA}	**0.4002**	0.6041	0.6625	1.492	
λ_{22}^{DE}	**1.670**	0.1896	8.808	**5.312**	**Ryhmän ja altisteen välinen interaktio**

Taulukon sarakkeessa 2 olevien **loglineaaristen efektien** estimaattien perusteella voidaan havaituille frekvenssille laskea **odotetut frekvenssit** seuraavan esimerkin mukaisesti:

Esim.

Lokerossa ikäryhmä="+75", altistus="verrokki", alkoholi="80 g tai alle" havaittu frekvenssi on **31**. Tämän lokeron odotettu frekvenssi saadaan summaamalla ensin mallin 2 mukaiset kertoimet -1.115 + 2.279 + 3.384 + 0.7179 - 3.962 + 0.4002 + 1.670 (lihavoitu taulukossa). Summaksi tulee 3.374. Odotettu frekvenssi saadaan ottamalla eksponenttifunktio (logaritmifunktion ($\log_e$) käänteisfunktio) tästä luvusta, eli odotettu frekvenssi on exp(3.374) = **29.20**.

Kun mallin parametrin estimaatti (sarake 2) jaetaan sen keskivirheellä (sarake 3), niin saadaan suure (sarake 4), joka noudattaa likimain normaalijakaumaa N(0,1). Näitä **standardoituja loglineaarisia efektejä** tarkastelemalla saadaan käsitys mallin eri tekijöiden välisistä riippuvuussuhteista.

Nyrkkisääntö:

Jos standardoidun efektin itseisarvo > 2, niin se on tilastollisesti merkitsevä.

Esim.

Iän ja tautiryhmän väliseen interaktioon liittyvät standardoidut efektit ovat selvästi suurempia kuin 2. Tämä merkitsee, että tautiryhmällä ja iällä on vahva riippuvuus keskenään. Negatiivinen etumerkki johtuu koodaustavasta (tapaus=1, verrokki=2 ja ikä on koodattu 1-6). Alkoholin ja iän väliset efektit ovat merkittäviä vain ikäluokissa 35- 44 ja 45- 54. Taudin ja altisteen välisen interaktion efekti on merkittävin 8.808.

Taulukon sarakkeeseen 5 on laskettu suure exp(estimaatti). Tämä on ns. **multiplikatiivinen efekti**. Niistä voidaan päätellä ristiintaulukoitujen tekijöiden eri koodiarvokombinaatioiden (taulukon indeksikombinaatioiden) lisääntynyttä tai vähentynyttä esiintymistodennäköisyyttä.

Esim.

Kertoimet $\exp(\lambda_i^A)$ kuvastavat iän marginaalijakaumaa. Havaintoaineiston perusteella todetaan, että ikäluokka 55 -64 on suurin ja siksi $\exp(\lambda_4^A) = 122.36$ on ikään liittyvistä arvoista suurin. Jos tarkastellaan alkoholin ja iän interaktioon liittyviä termejä $\exp(\lambda_{ij}^{EA})$, todetaan että vähän alkoholia käyttäviä on eniten

vanhimmassa ikäryhmässä (huom. koodaus), koska $\exp(\lambda_{26}^{EA}) = 1.492$ on selvästi näistä termeistä suurin.

Taulukon viimeisellä rivillä sarakkeessa 5 on taudin ja altisteen välisen interaktion multiplikatiivisena efektinä **5.312**. Tämä on yhtä kuin OR ja keskivirheen avulla sille saadaan normaalijakaumaan perustuvaksi luottamusväliksi: $CI_{95\,\%}$ = **(3.663, 7.703).**

Huom.

StatXact 6 antaa Mantel-Haenszel arvioiksi: OR=**5.158** ja $CI_{95\,\%}$ = (**3.562, 7.468).** Näitä luottamusvälejä voidaan laskea monilla eri varianssiestimaateilla StatXact 6 käyttää ns. **RBG-varianssia** (**R**obins, **B**reslow ja **G**reeland 1986), jonka on todettu toimivan hyvin monissa käytännön tilanteissa.

Tutkimustuloksen perusteella voidaan päätellä, että runsaalla alkoholin käytöllä, yli pullollinen punaviiniä päivittäin, ja ruokatorven syöpään sairastumisella on vahva assosiaatio keskenään. Mikäli OR:ää tulkitaan riskisuhteena, niin runsaasti käyttävillä on viisinkertainen riski sairastua verrattuna vähän alkoholia käyttäviin. Taudilla ja alkoholin käytöllä on merkittävät (**P<0.0001 ja P=0.0192**) interaktiot myös iän kanssa (Systat: mallin termit DA ja EA).

Ikätrendin testaaminen, eli muuttuuko taudin ja altisteen välinen interaktio lineaarisesti iän mukana, voidaan suorittaa liittämällä malliin järjestysasteikollinen muuttuja (A*), jonka arvoina ovat ikäluokkien indeksit 1, 2, …, 6.

Malli 3: M3 = (DE, DA, EA, DEA*)

Tulokset: $G^2_{M_3} = 10.61, df_{M_3} = 4, \quad \hat{\lambda}_{11}^{DEA^*} = -0.1246, SE(\hat{\lambda}_{11}^{DEA^*}) = 0.1890$

Interaktiotermin DEA^* merkitsevyyttä voidaan testata laskemalla G^2-testien erotus mallien M2 ja M3 välillä. Tulokseksi saadaan:

$$G^2_{M_2} - G^2_{M_3} = 11.04 - 10.61 = 0.659, df_{M_2} - df_{M_3} = 5 - 4 = 1, \; P = 0.4169$$

Tulkinta:

Termin DEA^* lisääminen ei paranna mallia tilastollisesti merkitsevästi, koska P>0.05.

Esim. 3:

Loglineaarinen analyysi (2 x k_1 x k_2)- taulukosta. Alkoholin ja tupakan samanaikainen vaikutus ruokatorven syöpään. Merkitään: **T** = tupakka, **A** = alkoholi, **D** = ryhmä (tapaukset, verrokit)
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/BDEsim3.sav

Alkoholi (g/vrk)	**Tupakka (kpl/vrk)**							
	0-9		**10-19**		**20-29**		**30+**	
	Tapaus	Verrokki	Tapaus	Verrokki	Tapaus	Verrokki	Tapaus	Verrokki
0-39	9	252	10	74	5	35	5	23
40-79	34	145	17	68	15	47	9	20
80-119	19	42	19	30	6	10	7	5
120+	16	8	12	6	7	5	10	3

Malli 1: $\log_e(F_{ij}) = \theta + \lambda_i^D + \lambda_j^T + \lambda_k^A + \lambda_{jk}^{TA}$ **M_1 = (D, TA)**

Tulos: $G^2_{M_1} = 166.9, df_{M_1} = 15, \quad P < 0.001$

Tulkinta:

Malli 1 testaa otantakehikon “rajoituksia”, eli malli testaa riskitekijöiden (tupakka, alkoholi) välisiä interaktioita, muttei niiden interaktioita tapaus-verrokkimuuttujaan. Malli on huonosti yhteensopiva aineiston kanssa, koska166.9 > 37.70 = $\chi^2_{15,0.001}$

Malli 2: $\log_e(F_{ij}) = \theta + \lambda^D_i + \lambda^T_j + \lambda^A_k + \lambda^{TA}_{jk} + \lambda^{DA}_{ik}$ **M_2 = (DA, TA)**

Tulos: $G^2_{M_2} = 20.40, df_{M_2} = 12, \quad P = 0.0561$

$G^2_{M_1} - G^2_{M_2} = 166.9 - 20.4 = 146.5, df_{M_1} - df_{M_2} = 15 - 12 = 3, \; P < 0.0001$

Tulkinta:

Ruokatorven syövän suhteellinen riski on erisuuri eri alkoholiryhmissä. Mallin yhteensopivuus aineiston kanssa on jo kohtalainen **(P = 0.0561)**, muttei kovin hyvä.

Malli 3: $\log_e(F_{ij}) = \theta + \lambda^D_i + \lambda^T_j + \lambda^A_k + \lambda^{TA}_{jk} + \lambda^{DA}_{ik} + \lambda^{DT}_{ij}$ **M_3 = (DT, DA, TA)**

Tulos: $G^2_{M_3} = 7.76, df_{M_3} = 9, \quad P = 0.5580$

$G^2_{M_2} - G^2_{M_3} = 20.4 - 7.76 = 12.6, df_{M_2} - df_{M_3} = 12 - 9 = 3, \; P = 0.0056$

Tulkinta:

Tässä mallissa alkoholin mahdollinen sekoittava vaikutus tupakan ja ruokatorven syövän riippuvuuteen on kontrolloitu, joten tupakalla on itsenäinen vaikutus, koska mallit M2 ja M3 eroavat tilastollisesti merkitsevästi toisistaan (P=0.0056) ja mallin yhteensopivuus on hyvä (P=0.5580).

Synergismin testaaminen:

- Muodostetaan tupakan ja alkoholin järjestysasteikolliset muuttujat luokkaindeksejä käyttäen; arvot 1, 2, 3, 4. Merk. T* ja A*
- Liitetään malliin interaktiotermi $\lambda^{DT^*A^*}_{ijk}$ (malli 4)

Tulos: $G^2_{M_4} = 5.460, df = 8$ **ja** $G^2_{M_3} - G^2_{M_4} = 7.765 - 5.460 = 2.305, \quad P = 0.1290$

Tulkinta: Ei tilastollisesti merkitsevää evidenssiä synergismistä (**P=0.1290>0.05**).

Viitteet

Agresti A. Categorical data analysis. Wiley-Interscience, New York 1990

Bishop YMM, **Fienberg** SE & **Holland** PW. Discrete multivariate analysis: Theory and Practice. McGraw-Hill, Cambridge Mass.1975.

Breslow NE & **Day** NE. Statistical methods in cancer research - The analysis of case-control studies, IARC Scientific publications no. 32, 1980

Poisson-regressio

Mallin määrittely

Käyttö:

Poisson-regressiota käytetään silloin, kun kyseessä on melko suuri aineisto, lopputulosmuuttuja on lukumäärän tai suhteellinen osuuden muodossa ja yksittäisen tapahtuman todennäköisyys on pieni, esim. harvinaisten tautitapausten ilmaantumisen riippuvuus erilaisista ennustetekijöistä.

Esim.

Ikäluokkakohtaiset astmakuolleisuustrendit Englannissa ja Walesissa 1983-1995. (Campbell ym. 1997)
Tutkimuksessa rekisteröitiin yhteensä 23311 astmakuolemaa aikavälillä 1.1.1983 ja 31.12.1995 kyseisissä maissa. Tarkasteltava kovariaatti, ikä, oli luokiteltu 0-4, 5-14,15- 44, 45- 64, 65- 74 ja 75- 84 vuotta. Tämä ikäluokkajako määräytyi sillä perusteella, että väestömäärät olivat käytettävissä näistä ikäluokista. Tutkimuksessa tarkasteltiin vuosittaisia kuolemantapauksia, joten lukumääräkertoimina (”rate multiplier”) ovat elossa olevien määrät kyseisissä ikäluokissa kunakin seurantavuotena.

Poisson-regressio kuuluu nk. **yleistettyjen lineaaristen mallien** (”generalized linear models”) joukkoon, johon edellä tarkasteltu tavallinen pienimmän neliösumman lineaarinen regressiokin kuuluu. Perusoletus klassisessa Poisson-regressiossa on, että tapahtumien määrä tietyssä aikayksikössä noudattaa Poisson-jakaumaa, jonka odotusarvo on λ, joka on tapahtumien odotettu lukumäärä E_i jaettuna henkilöiden tai henkilövuosien määrällä n, (”person-years”), eli kuinka pitkään kukin tutkimuksen kohteena oleva henkilö on ollut alttiina tutkittavalle tapahtumalle. Poisson-regressiossa tätä suuretta nimitetään lukumääräkertoimeksi (”rate multiplier”). Malli on muotoa:

$$\log(\lambda_i) = \log(E_i/n_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Muuttujat $\mathbf{x_1,\ldots, x_p}$ ovat kovariaatteja, selittäviä muuttujia. Todetaan, että tapahtumien ilmaantuvuuden logaritmi $\log(\lambda_i)$ on tässä mallissa kovariaattien lineaarinen funktio.

Mallinnettava aineisto on voi olla yksilötasoinen tai taulukkomuotoon tiivistetty siten, että kaikki tarkasteltavat kovariaattikombinaatiot muodostavat havaintoyksiköt, eli rivit ja sarakkeina ovat niiden ilmaisinmuuttujat, tapahtumien määrät sekä kullekin kovariaattikombinaatiolle lasketut henkilö- tai henkilövuosimäärä n_i.

Yllä oleva yhtälö voidaan kirjoittaa muotoon

$$\log(E_i) = \log(n_i) + \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Termin $\log(n_i)$ regressiokerroin on kiinteä arvo 1 ja siitä käytetään nimitystä mallin ”offset”. Havaitut lukumäärät $\mathbf{y_i}$ kytketään odotettuihin lukumääriin Poisson-jakauman perusteella, eli tehdään oletus, että $\mathbf{y_i}$ noudattaa Poisson-jakaumaa, jonka parametri, odotusarvo, on $\mathbf{n_i\,\lambda_i}$. Tällöin

$$\log(y) = \log(n_i) + b_0 + b_1 x_1 + \ldots + b_p x_p + \varepsilon$$

ε on virhetermi ja ovat $b_0, b_1, \ldots, b_p$ ovat avaintoaineiston perusteella arvioituja regressiokertoimia.

Poisson-jakaumalle on tyypillistä:

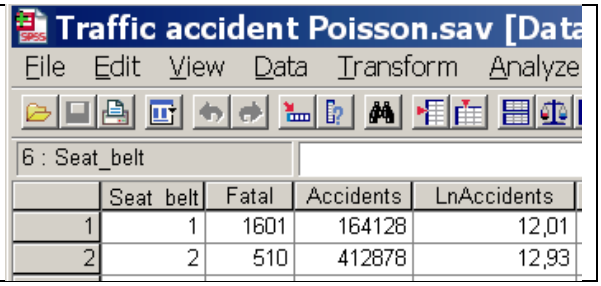
a) Poisson-jakauma on vino, joten **virhetermin ε jakauma on epäsymmetrinen** toisin kuin lineaarisessa regressiossa.
b) Poisson- jakaumassa **varianssi kasvaa, kun odotusarvo kasvaa** toisin kuin lineaarisessa regressiossa, missä perusoletuksiin kuului vakiovarianssius, homoskedastisuus. Poisson-jakaumassa varianssi on sama kuin odotusarvo.
c) Poisson-jakauma on **ei-negatiivinen**, joten ennustetut arvot ovat positiivisia toisin kuin lineaarisessa regressiossa, missä myös negatiiviset ennustetut arvot ovat mahdollisia.

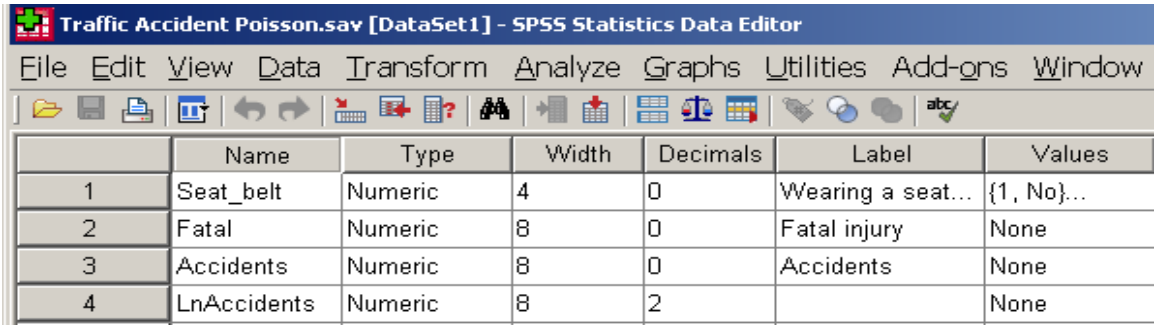
Poisson-regressiossa käytetään logaritmimuunnosta, joten siitä syystä vältytään negatiivisilta ennustetuilta arvoilta. Mikäli olettamus b) ei ole realistinen, niin hyvä vaihtoehto Poisson-regressiolle on **negatiiviseen binomiaaliseen jakaumaan** perustuva malli, missä varianssi on suurempi kuin keskiarvo. Klassisesta Poisson-regressiosta on olemassa myös lukuisia yleistyksiä, kuten malli, missä arvolle 0 annetaan suurempi todennäköisyys kuin, mitä Poisson-jakauma antaisi ja muille arvoille annetaan Poisson-jakauman mukaiset todennäköisyydet. Poisson mallia voidaan käyttää myös siten, että henkilö- tai henkilövuosimäärien asemesta käytettään, esim. standardoiduissa kuolleisuussuhteissa (”standardized mortality ratio, SMR) käytettyjä odotettuja tapausmääriä, jotka on laskettu jonkin ulkopuolisen aineiston, esim. jonkin alueen syöpärekisterin perusteella.

Esimerkkejä

Liikenneonnettomuudet ja turvavöiden käyttö

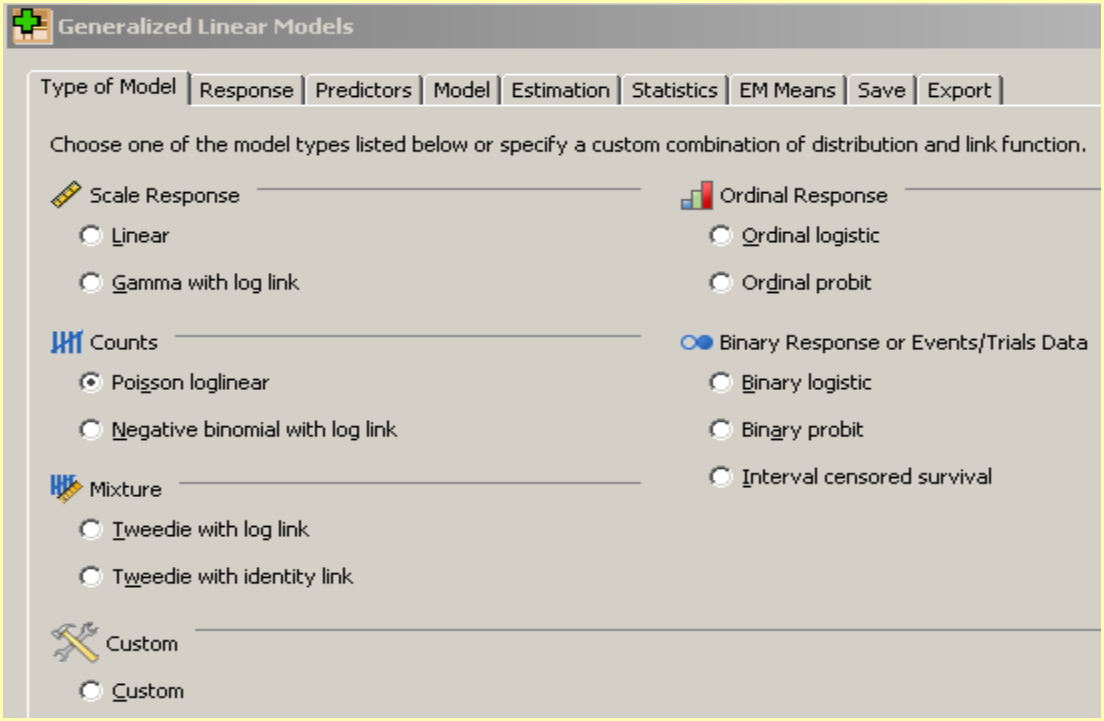
1988 Florida automobile accident data		Vammatyyppi	
		Fataali	Ei-fataali
Turvavöiden käyttö	Ei	1601	162527
	Kyllä	510	412368

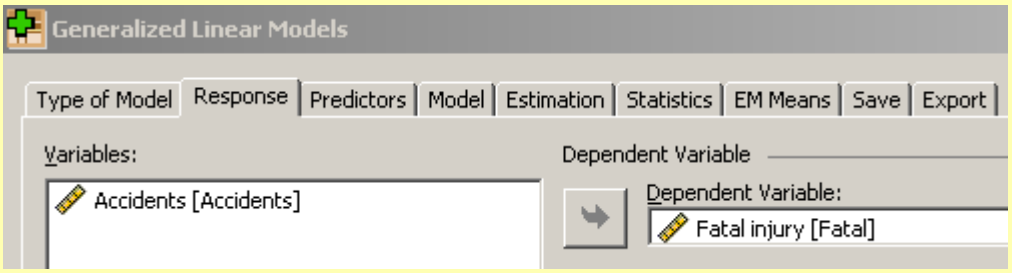


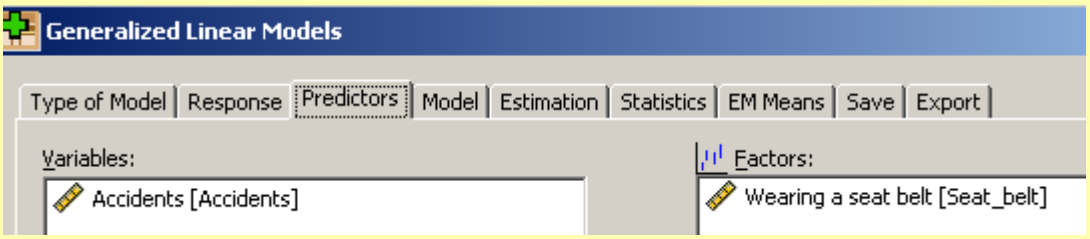


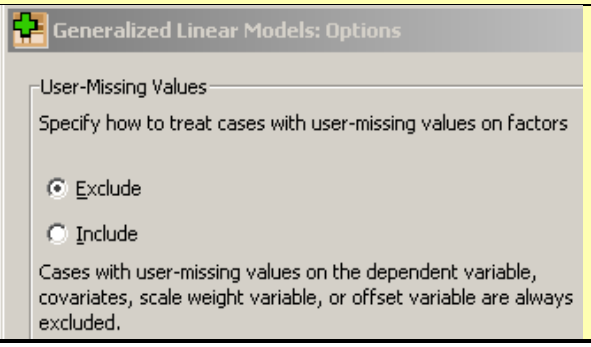
Tiedosto:
http://www.mv.helsinki.fi/home/sarna/Data/TrafficAccidentPoisson.sav

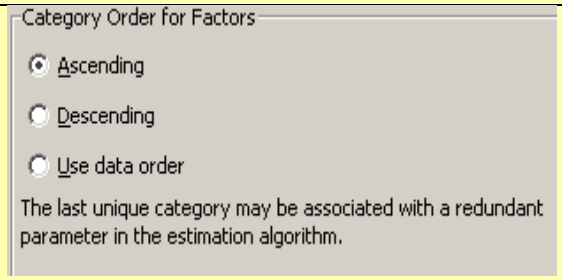
SPSS: Valikot **”Analyze” ►”Generalized Linear Models”► ”Generalized Linear Models”**

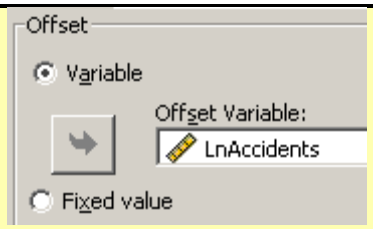




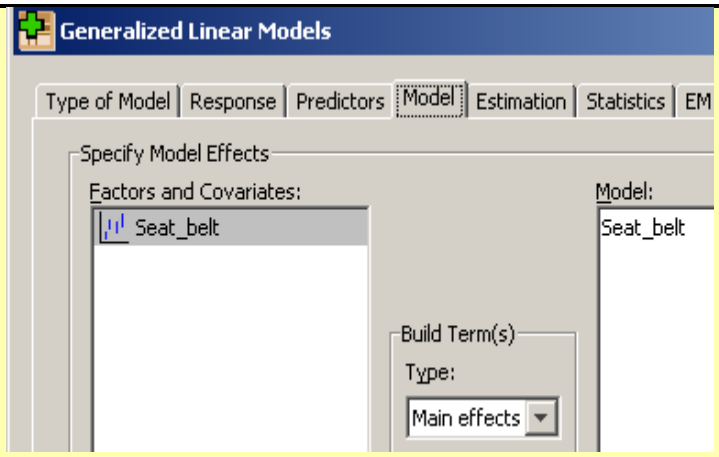


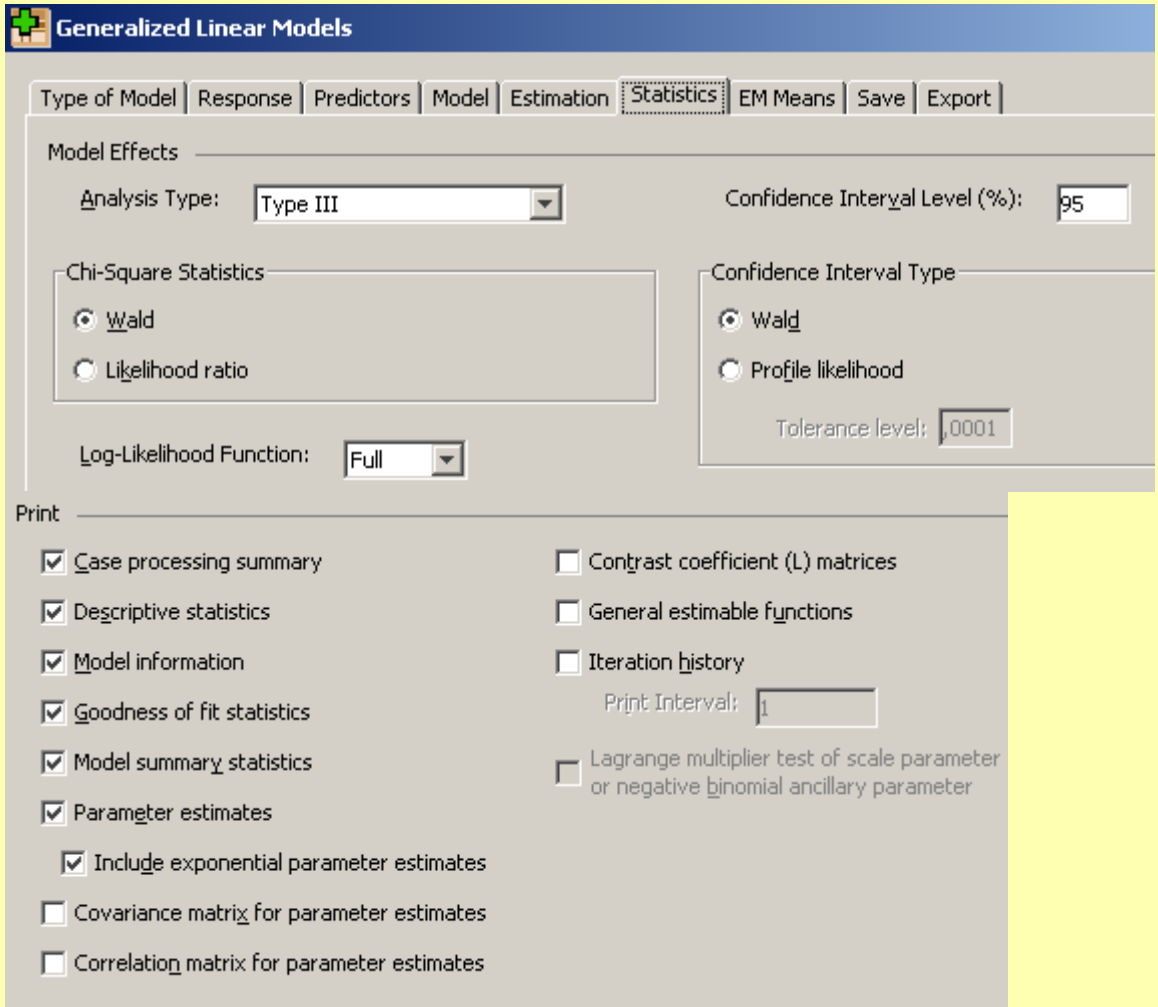






"Offset" termi pitää sisällyttää malliin. Se on vakio, jota ei estimoida. Tässä tapauksessa se on luonnollinen logaritmi onnettomuuksien kokonaismäärästä määrästä turvavöitä käyttäneillä ja käyttämättömillä.





Model Information

Dependent Variable	**Fatal injury**
Probability Distribution	**Poisson**
Link Function	**Log**

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
2032,592	**1**	**,000**

Offset Variable	LnAccidents

Dependent Variable: Fatal injury
Model: (Intercept), Seat_belt, offset = LnAccidents
a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	49621,041	1	,000
Seat_belt	1651,715	1	,000

Dependent Variable: Fatal injury
Model: (Intercept), Seat_belt, offset = LnAccidents

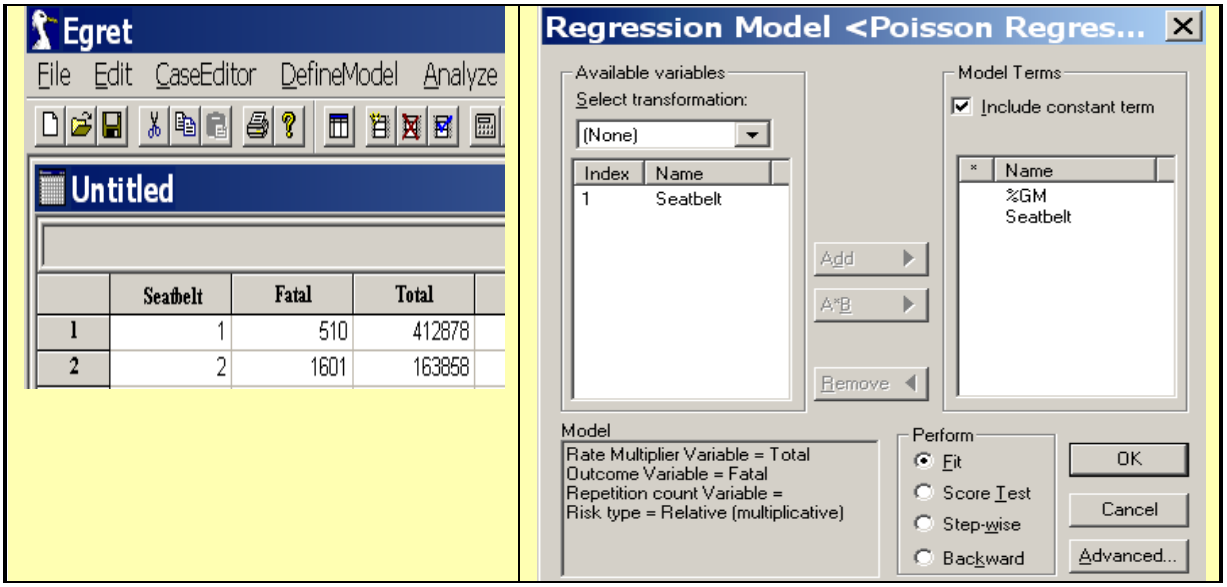
Parameter Estimates

		Parameter			
		(Intercept)	[Seat_belt=1]	[Seat_belt=2]	(Scale)
B		-6,696	2,066	0[a]	1[b]
Std. Error		,0443	,0508	.	
95% Wald Confidence Interval	Lower	-6,783	1,967	.	
	Upper	-6,610	2,166	.	
Hypothesis Test	Wald Chi-Square	22869,965	1651,715	.	
	df	1	1	.	
	Sig.	,000	,000	.	
Exp(B)		,001	7,897	1	
95% Wald Confidence Interval for Exp(B)	Lower	,001	7,148	.	
	Upper	,001	8,725	.	

Dependent Variable: Fatal injury Model: (Intercept), Seat_belt, offset = LnAccidents
a. Set to zero because this parameter is redundant. b. Fixed at the displayed value.

Tulkinta: RR = exp(2,066)=**7,897**, CI95 %: (exp(1,967)=7,148, exp(2,166)=8,725, P<0,001. Turvavöiden käyttämättä jättäminen lisää fataliteettiriskin noin **8-kertaiseksi**. Tässä yksinkertaisessa esimerkkitapauksessa RR voitaisiin laskea myös suoraan taulukon luvuista: RR = 1601/162527)/(510/412368)=7,96, mutta mikäli mukana olisi esim. sekoittavia tekijöitä, niin silloin Poisson-regressiomallin käyttö olisi perusteltua.

Sama esimerkki Egretillä:



Poisson Regression Model

Data file name	
Model	Fatal ~ %GM + Seatbelt
Risk Type	Relative (multiplicative)
Rate multiplier	Total
Analysis Type	Fit using Newton Raphson algorithm

Basic Information

Number of terms	2
Total Number of Observations	2
Rejected as Invalid	0
Number of valid Observations	2

Model Fit Results

Summary Statistics

	Value	DF	p-value
Deviance	0	0	
Likelihood ratio test	25766,6836	2	< 0.001

Parameter Estimates

Terms	Coefficient	Std.Error	p-value	Rate Ratio	95% C.I. Lower	95% C.I. Upper
%GM	-8,7646	0,0920	< 0.001	0,0002	0,0001	0,0002
Seatbelt	2,0681	0,0508	< 0.001	7,9100	7,1597	8,7389

Englantilaisten miespuolisten lääkärien sepelvaltimotautikuolemat ja tupakointi

Lähde: Doll R and Hill AB 1986

Ikäryhmä: 1=’35 - 44’, 2=’45 - 54’, 3=’55 - 64’, 4=’65 - 74’, 5=’75 - 84’
Tupakoitsija 0=’ei’, 1=’kyllä’
Tapaukset: sepelvaltimotautikuolemat
n: henkilövuodet (=seuranta-aika)

British male doctors Poisson.sav [DataS

File Edit View Data Transform Analyze Graphs

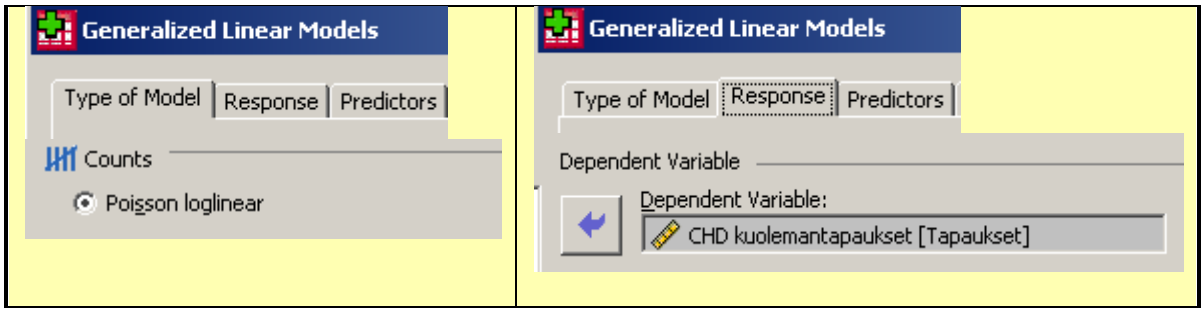
14 :

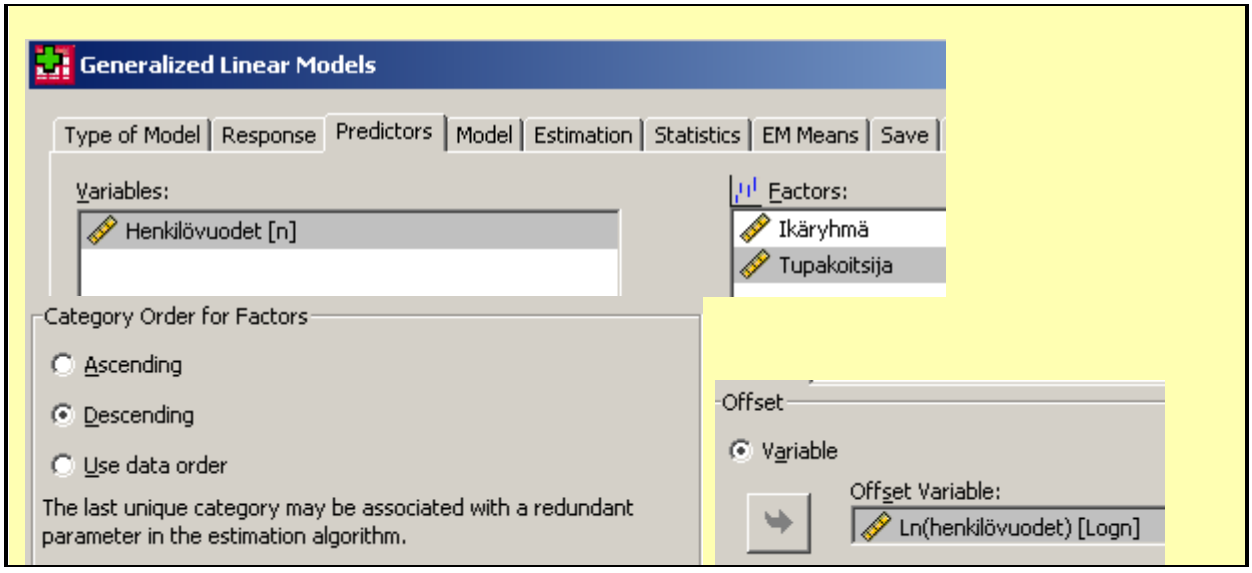
	Ikäryhmä	Tupakoitsija	Tapaukset	n	Logn
1	35-44	Kyllä	32	52407	10,87
2	35-44	Ei	2	18790	9,84
3	45-54	Kyllä	104	43248	10,67
4	45-54	Ei	12	10673	9,28
5	55-64	Kyllä	206	28612	10,26
6	55-64	Ei	28	5712	8,65
7	65-74	Kyllä	186	12663	9,45
8	65-74	Ei	28	2585	7,86
9	75-84	Kyllä	102	5317	8,58
10	75-84	Ei	31	1462	7,29

Tiedosto:
http://www.mv.helsinki.fi/home/sarna/Data/British male doctors Poisson.sav

Tutkittava ongelma:

Kuinka suuri on tupakoitsijoiden CHD-kuoleman riski verrattuna tupakoitsemattomiin, kun huomioidaan ikä mahdollisena sekoittavana tekijänä?

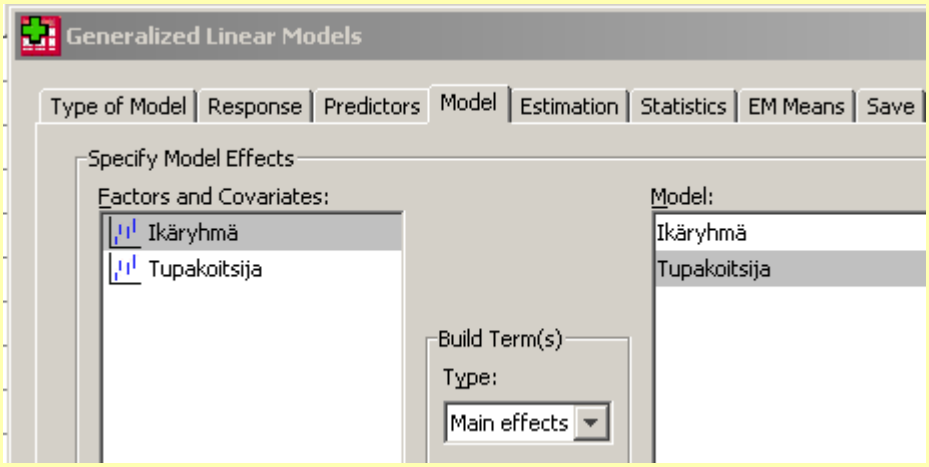




Huom.

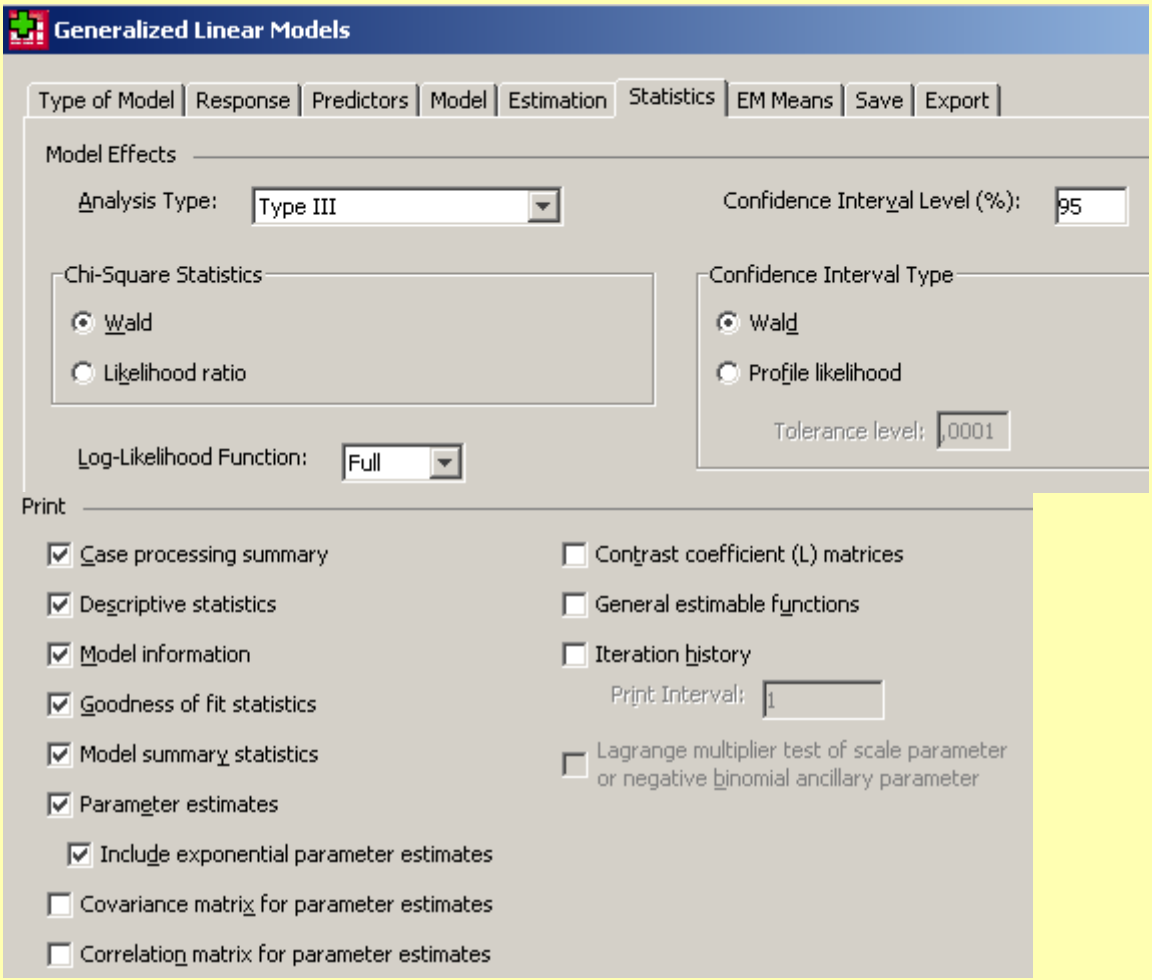
Koska tupakoitsija on kaksiarvoinen muuttuja, niin se yhdentekevää laitetaanko se lokeroon ”Factor” vai ”Covariate” Referenssikategoriaksi tässä tapauksessa tulee alin kategoria, eli tupakoimaton

Huom. Nuorin ikäryhmä otetaan referenssiksi.



Huom.

Mallissa oletetaan, että CHD kuoleman riskisuhde RR on vakio ikäryhmän sisällä.



Tulos:

Model Information

Dependent Variable	CHD kuolemantapaukset
Probability Distribution	Poisson
Link Function	Log
Offset Variable	Ln(henkilövuodet)

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
922,931	5	,000

Dependent Variable: CHD kuolemantapaukset

Model: (Intercept), Ikäryhmä, Tupakoitsija, offset = Logn

a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	8482,117	1	,000
Ikäryhmä	643,157	4	,000
Tupakoitsija	10,909	1	,001

Dependent Variable: CHD kuolemantapaukset

Model: (Intercept), Ikäryhmä, Tupakoitsija, offset = Logn

Tulkinta.

LR- testi antaa erittäin merkitsevän P-arvon koko mallille, mutta se johtuu suurelta osin ikäryhmästä, josta CHD riski luonnollisesti on vahvasti riippuvainen.

Parameter Estimates

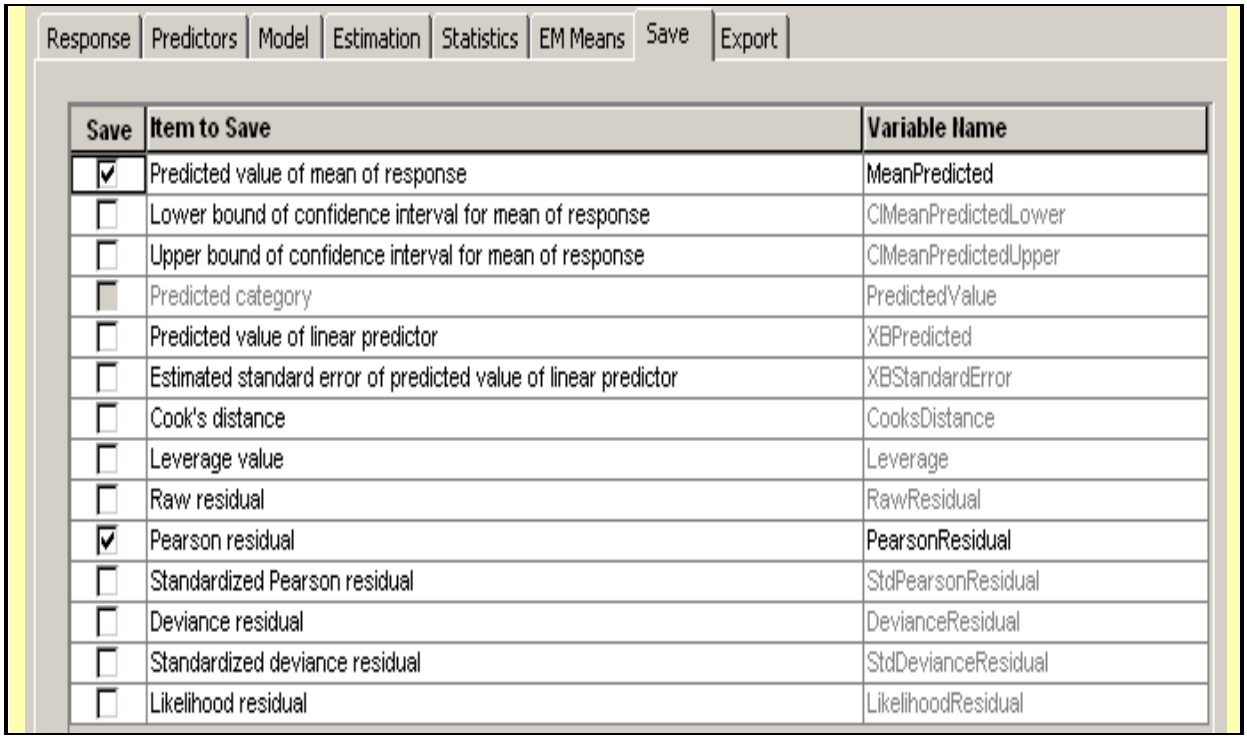
Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp (B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-7,919	,1918	-8,295	-7,544	1,706E3	1	,000	,000	,000	,001
[Ikäryhmä=5]	3,700	,1922	3,323	4,077	370,536	1	,000	40,451	27,753	58,959
[Ikäryhmä=4]	3,350	,1848	2,988	3,713	328,712	1	,000	28,517	19,852	40,964
[Ikäryhmä=3]	2,627	,1837	2,267	2,988	204,514	1	,000	13,838	9,654	19,837
[Ikäryhmä=2]	1,484	,1951	1,102	1,866	57,855	1	,000	4,411	3,009	6,465
[Ikäryhmä=1]	0[a]	.	.	.	.	.	.	1	.	.
[Tupakoitsija=1]	,355	,1074	,144	,565	10,909	1	,001	1,426	1,155	1,760
[Tupakoitsija=0]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	1[b]									

Tulkinta:

Tupakoitsijoiden ikävakioitu RR=**1,426** ja sen 95 %:n luottamusväli on (**1,155, 1,760**) ja **P=0,001**, eli tilastollisesti erittäin merkitsevä. Näin ollen englantilaisilla tupakoitsevilla mieslääkäreillä on **42,6 %** suurempi riski kuolla sepelvaltimotautiin kuin tupakoitsemattomilla. Huom. Koska vapausasteita on 1, niin Waldin khi^2-testiarvon 10,909 neliöjuuri 3,303 on z, eli standardin normaalijakauman prosenttipiste.

Mallin hyvyyden tarkistaminen

Mallin hyvyyttä voidaan testata monin eri menetelmin. SPSS:ssä on käytettävissä seuraavat suureet:



Yksinkertaisin tapa on laskea kullekin kymmenestä havaintoyksiköstä ennustetut arvot (E) ja niiden perusteella standardoidut residuaalit (E-D)/sqrt(E), missä D on havaitut kuolemantapaukset. Tällöin saadaan:

Ikäryhmä	Tupakka	Tapaukset (D)	Odotusarvo (E)	Residuaali (D-E)/Sqrt(E)	$(D-E)^2/E$
35-44	Kyllä	32	27	0,93	0,86
35-44	Ei	2	7	-1,85	3,42
45-54	Kyllä	104	99	0,52	0,27
45-54	Ei	12	17	-1,24	1,53
55-64	Kyllä	206	205	0,05	0,00
55-64	Ei	28	29	-0,14	0,02
65-74	Kyllä	186	187	-0,09	0,01
65-74	Ei	28	27	0,23	0,05
75-84	Kyllä	102	111	-0,90	0,81
75-84	Ei	31	22	2,05	4,19
				$Khii^2$ =	11,16
				df=10-(5-1)*(2-1)=4	
				P=	**0,0248**

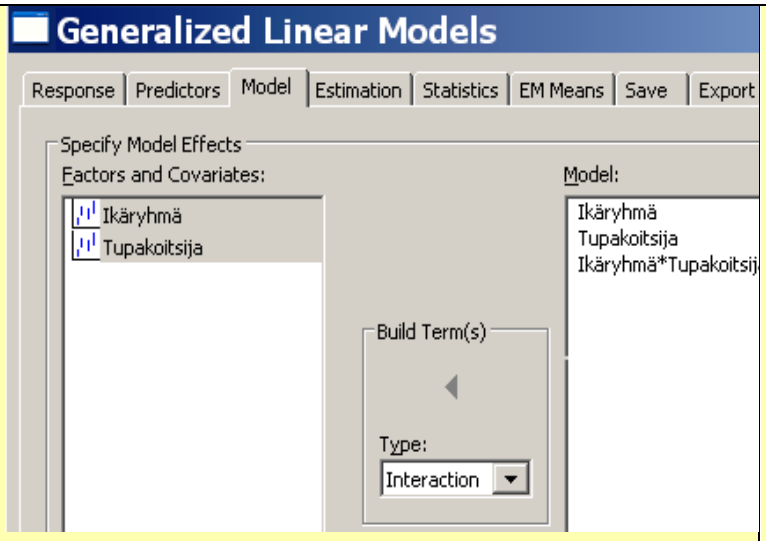
Tulkinta:

Havaittujen ja odotettujen arvojen yhteensopivuutta voidaan testata X^2-testillä. Vapausasteiden eli vapaiden solujen määrä on tässä (ikäluokkien määrä-1) *(tupakointiluokkien määrä -1), eli 4. Mallin hyvyys ei ole kovin hyvä, koska

P=0,024; alimmassa ja ylimmässä Ikäluokassa tupakoimattomien residuaalit ovat melko suuret.

Interaktion testaaminen:

Pitääkö malliin ottaa mukaan interaktiotermi ts. onko ikäryhmän ja tupakoinnin välillä interaktiota?



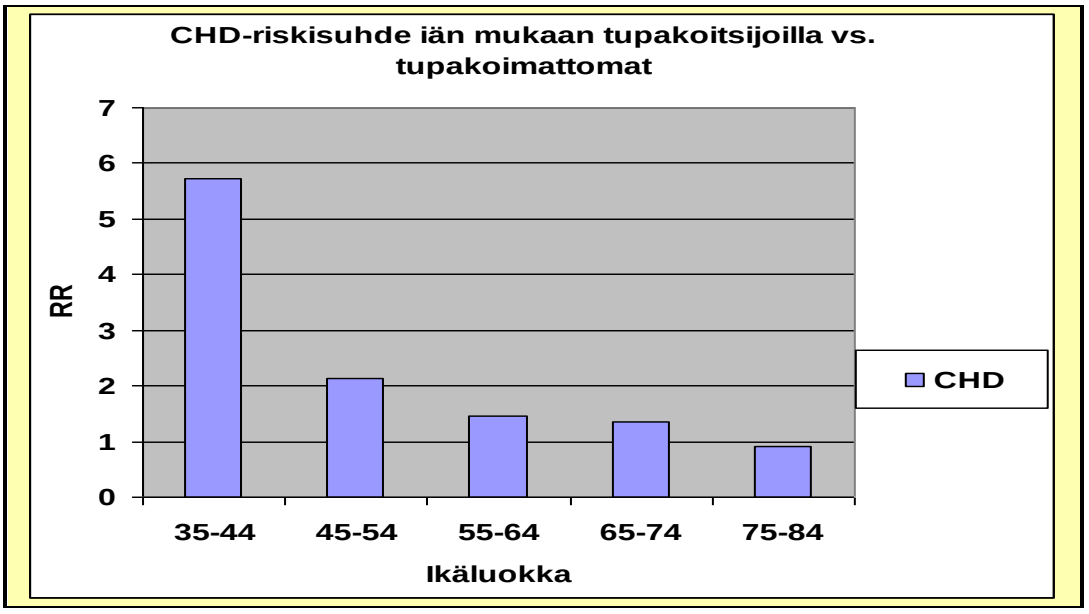
Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	**4219,199**	**1**	**,000**
Ikäryhmä	**307,137**	**4**	**,000**
Tupakoitsija	**12,816**	**1**	**,000**
*Ikäryhmä * Tupakoitsija*	**10,200**	**4**	**,037**

Dependent Variable: CHD kuolemantapaukset
Model: (Intercept), Ikäryhmä, Tupakoitsija, Ikäryhmä * Tupakoitsija, offset = Logn

Tulkinta:

Ikäryhmän ja tupakoinnin välinen interaktio on tilastollisesti merkitsevä (P=0,037), joten tupakointi vaikuttaa jonkin verran eri tavoin sepelvaltimotautiriskiin eri ikäryhmissä.



Viitteet

Campbell MJ, **Cogman** GR, **Holgate** ST, **Johnston** SL. Age specific trends in asthma mortality in England and Wales, 1983-95: results of an observational study. BMJ 1997;314:1439-41 (17 May)

Doll R and **Hill** AB. Mortality of British male doctors in relation to smoking: Observations on coronary trombosis. National Cancer Institute Monograph 1966; 19:205-68

Florida State Department of Highway Safety and Motor Vehicles, 1988

Logistiset regressioanalyysit

Johdanto

Selitettävän muuttujan y ollessa kaksiluokkainen, dikotominen (esim. 0 = “henkilö on terve”, 1 = “henkilö sairastaa sepelvaltimotautia”) voidaan regressio- tai erotteluanalyysiä käytettäessä saada epäloogisia tuloksia. On mahdollista, että joillekin aineistoon kuuluville potilaille ennustetut y-arvot (sairastumisen todennäköisyydet) menevät välin (0 - 1) ulkopuolelle. Tämä epäsuotava tilanne voidaan välttää käyttämällä regressio- ja erotteluanalyysien asemesta logistista mallia. Cornfield esitti ensimmäisenä tämän mallin käyttöä sepelvaltimotautiriskin arvioimiseksi vuonna 1962. Logistisella analyysillä on erotteluanalyysiin verrattuna sairastumisriskin mielekkään tulkinnan ohella toinenkin merkittävä etu: arvioiduista kertoimista voidaan suoraan arvioida suhteellisen sairastumisriskin muutoksia kullekin mallissa olevalle muuttujalle erikseen.

Logistinen regressio “logistic regression” on regressiomenetelmä, jota käytetään mallittamaan logaritmimuunnettua suhdelukutyyppistä ennustettua lopputulosta logit = **$\log_e(\pi/(1-\pi)) = \log_e(\text{“odds”})$,** missä π on tutkittavan tapahtuman todennäköisyys, π voisi olla esim. todennäköisyys sairastua sepelvaltimotautiin tietyllä aikavälillä. Lopputulosmuuttuja nk. **binäärisessä logistisessa** regressiossa on kaksiarvoinen, esim. sairastuu tai ei sairastu. Logistisen mallin avulla voidaan tutkia mitkä tutkimusongelman kannalta relevanteista tekijöistä (x_1, …, x_k) assosioituvat sairastumisriskian ja kuinka vahvasti. Tulokset esitetään tavallisimmin ristitulosuhteina (OR) ja niiden luottamusväleinä**. Mikäli π on lähellä 0.5:ttä, niin OR liioittelee vahvasti sairastumisriskiä**. Kaltaistetuilla aineistoilla käytetään nk. **ehdollista logistista regressiota**. Perusmenetelmän yleistyksiä ovat **multinomiaalinen** (polytominen) **logistinen regressio** laatueroasteikolliselle moniluokkaiselle lopputulosmuuttujalle ja järjestysasteikollisen muuttujan **ordinaalinen logistinen regressio**

Mikäli selitettävä muuttuja y on kaksiluokkainen, niin silloin voidaan käyttää binääristä logistista mallia. **SPSS**: **”Analyze” ►”Regression” ►“Binary Logistic”**. Jos sen sijaan y on moniluokkainen ilmaisten esim. useita eri sairaustiloja, niin silloin voidaan käyttää multinomiaalista logistista mallia. **SPSS**: **”Analyze” ►”Regression” ►”Multinomial Logistic”**. Joissakin tilastopaketeissa ja oppikirjoissa jälkimmäisestä regressiosta käytetään myös **”polytomous logistic regression”**. Seuraavassa tarkastellaan binääristä logistista regressiota. Multinomiaalisen logistisen regressiomallin tulosten tulkinta on hyvin samanlaista kuin binäärisen mallin.

Muuttujat:

- **Vastemuuttuja (y)**, josta käytetään myös nimityksiä **selitettävä muuttuja tai lopputulosmuuttuja.**

Huom.

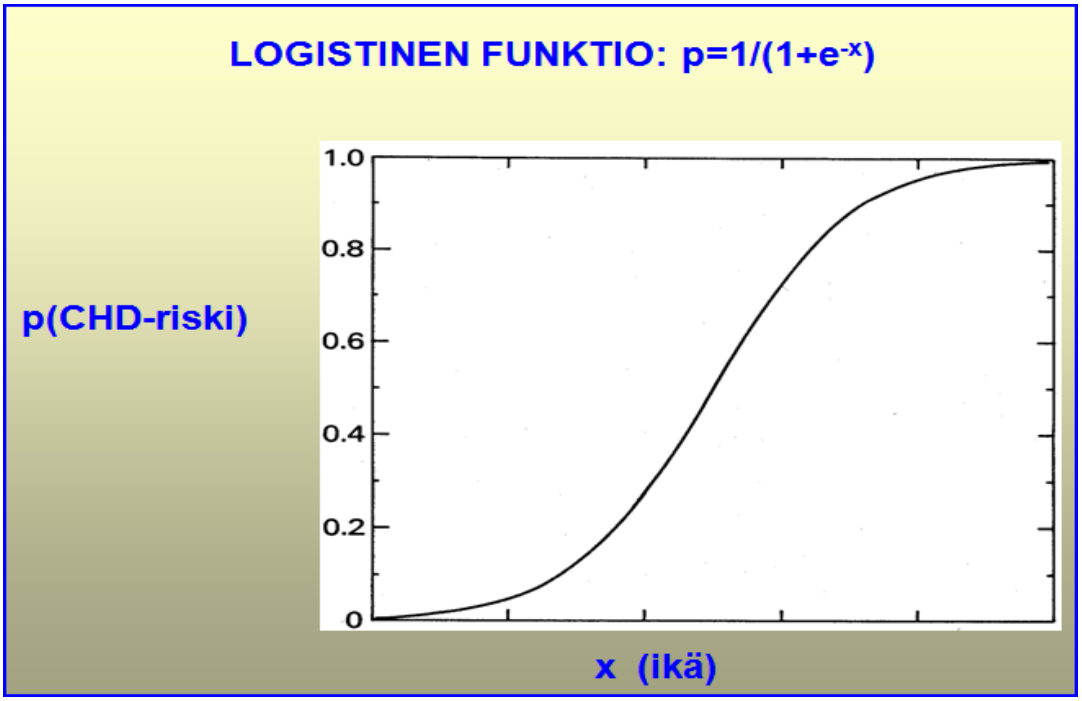
Muuttuja y kannattaa koodata siten, että 0=ei tapahtumaa ja 1=tapahtuma.

- **Selittävät muuttujat $x=(x_1, \ldots, x_p)$.** Yleisesti käytettyjä nimityksiä ovat myös **ennustavat muuttujat** ja **kovariaatit.** **x_{ij} :**t voivat olla a) jatkuvia, b) luokiteltuja tai c) järjestysasteikollisia.

Peruserona lineaarisen regression ja logistisen regression välillä on, että y on dikotominen eikä jatkuva. Tämä ero näkyy mm. mallin valinnassa ja olettamuksissa, mutta

yleisperiaatteiltaan ja analysointimenetelmiltään molemmat menetelmät noudattelevat hyvin samoja suuntaviivoja. Pääosa lineaarisen regression yhteydessä tarkasteltuja asioita pätee myös logistiseen regressioon.

Lineaarisessa regressiossa oletettiin, että y on jatkuva ja mallinnettiin **y:n ehdollista keskiarvoa**, odotusarvoa **E(y | x)**, ehdolla x on annettu. Mallina oli $E(y \mid x)=\beta_0+\beta_1 \cdot x$ Tämän mallin mukaisesti E(y | x) voi periaatteessa saada miinus äärettömästä plus äärettömään. Kun y on dikotominen, niin E(y | x):n täytyy olla välillä [0, 1]. Monia eri funktioita on ehdotettu kuvaamaan x:n ja E(y | x):n välistä yhteyttä. Logistisen funktion valinnalle tähän tarkoitukseen on monia perusteluita, joista tärkeimpiä ovat: 1) logistisella funktiolla on matemaattisesti monia miellyttäviä piirteitä; joustava, helppokäyttöinen, jne. 2) se sopii hyvin moniin käytännön tilanteisiin ja antaa biologisesti mielekkään tulkinnan.



Kuva:

Teoreettinen malli Iän ja sepelvaltimotautiriskin välisestä yhteydestä, logistinen funktio,

Tarkastellaan aluksi tilannetta, missä on vain yksi kovariaatti x, eli ikä ja merkitään yksinkertaisuuden vuoksi $\mathbf{E(y \mid x) = \pi(x)}$, **logistinen malli** on tällöin muotoa:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x)}}$$

π(x) on tarkasteltavan tapahtuman todennäköisyys. Tämä malli on parametrien β_0 (=vakiotermi) ja β_1 (=regressiokerroin) suhteen epälineaarinen funktio. Se saadaan parametrien suhteen lineaariseksi käyttämällä logistista eli **logit-muunnosta (”log-odds”):**

$$g(E(y \mid x)) = \log_e(\pi(x)/(1 - \pi(x)) = \beta_0 + \beta_1 x$$

Logaritmifunktion argumenttina oleva lauseke on tarkasteltavan tapahtuman mahdollisuus (”odds”). Se on tulkinnaltaan kuten vedonlyöntisuhde, eli tietyn tapahtuman

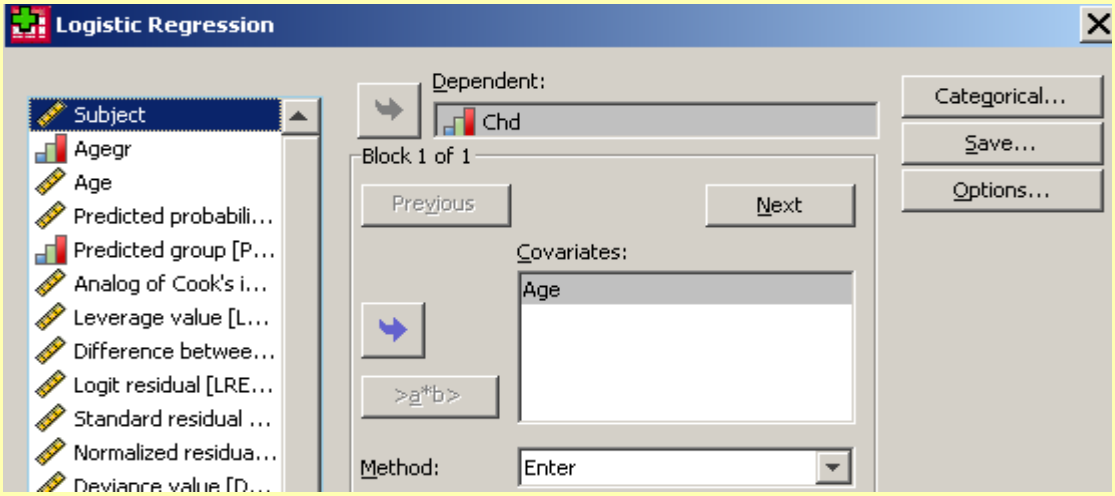
(esim. CHD) todennäköisyys jaettuna sen komplementtitapahtuman (ei CHD:tä) todennäköisyydellä.

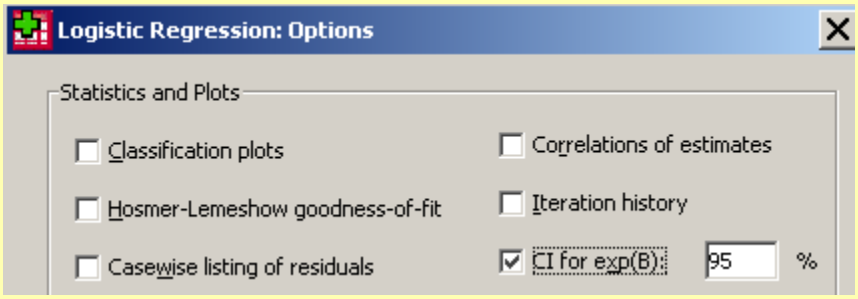
Logistisen mallin kertoimet β_0 ja β_1 arvioidaan tavallisimmin **suurimman uskottavuuden** (”maximum likelihood”) menetelmällä, mikä tässä yhteydessä tarkoittaa, että menetelmä tuottaa näille parametreille sellaiset arviot, jotka maksimoivat todennäköisyyden saada aineiston havaitut arvot, eli että poikkeama mallin perusteella odotettujen arvojen ja havaittujen arvojen välillä on mahdollisimman pieni. Logistisen mallin kertoimien arviointi on laskennallisesti hankalampaa kuin lineaarisen regressio mallin kertoimien. Laskenta perustuu nk. iteratiivisiin menetelmiin.

Esim.

Ikä ja sepelvaltimotautiriski. Hypoteettinen aineisto, 100 miestä ikäväliltä 20 - 69 vuotta. (Lähdeaineisto: Hosmer & Lemeshow, 2000 s.3 lievästi muokattuna)
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Chd_data.sav

SPSS: Valikot **”Analyze” ►”Regression”► ”Binary Logistic”**





Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Age	26,555	1	,000
	Overall Statistics		26,555	1	,000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	29,504	1	,000
	Block	29,504	1	,000
	Model	29,504	1	,000

Tulkinta:

Malliin sisältyvien muuttujien tilastollisen merkitsevyyden testaamiseksi on käytettävissä kolme testiä 1) **G-testi** (”likelihood ratio”- eli uskottavuussuhdetesti) 2) **”Score”-testi** ja 3) **Waldin testi**. Nämä testisuureet noudattavat likimain ***χ^2*-jakaumaa**. Tässä tapauksessa kaikki testit antavat muuttujalle ”Age” P-arvon (”Sig.”) 0,000, mikä tarkoittaa P<0,001, joten ikä selittää tilastollisesti merkitsevästi tässä aineistossa sepelvaltimotautitapauksia.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107,159[a]	,255	,343

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than 0,001.

Tulkinta:

Suuretta ”-2 Log likelihood” käytetään G-testin laskemisessa. R^2 suureet (Cox & Snell sekä Nagelkerke) ovat kuten vastaava suure lineaarisessa regressiossa, mutta käytännössä niiden tulkinta on logistisen mallin yhteydessä selvästi hankalampaa, koska selitettävä muuttuja on kaksiarvoinen. Ne mittaavat, kuinka hyvin malli selittää y-muuttujan vaihteluita. Tässä tapaukessa malli selittää näistä vaihteluista **34,3 %.** Tämän vaihteluosuuden tulisi käytännössä kuitenkin olla paljon suurempi, mikäli mallin perusteella pyrittäisiin ennustamaan yksittäisen henkilön sepelvaltimotautitapahtumia.

Classification Table[a]

Observed			Predicted		
			Chd		Percentage Correct
			0	1	
Step 1	Chd	0	45	12	78,9
		1	14	29	67,4
	Overall Percentage				74,0

a. The cut value is ,500

Tulkinta:

Luokittelun onnistumisprosentiksi saadaan **74.** Tämä tulos antaa kuitenkin käytännössä liian optimistisen kuvan, koska siinä on luokiteltu sama aineisto, jonka perusteella mallin kertoimet on arvioitu.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	Age	,112	,024	21,349	1	,000	1,118	1,066	1,172
	Constant	-5,339	1,138	22,027	1	,000	,005		

a. Variable(s) entered on step 1: Age.

Tulkinta:

Sarakkeessa **B** on mallin suurimman uskottavuuden menetelmällä arvioidut regressiokertoimet parametreille β_0 ja β_1. Kertoimen arvo 0,112 ilmaisee CHD:n logaritmisen mahdollisuuden (”odds”) keskimääräisen muutoksen ikävuotta kohden.
Sarakkeessa **S.E.** on kertoimien arvioidut keskivirheet.
Sarakkeessa **Wald** on kertoimien tilastollista merkitsevyyttä mittaava Waldin testisuure. Se noudattaa likimain $\chi 2$-jakaumaa vapausastein 1 (sarake d). Huom. Sqrt(Wald)=z, eli standardia normaalijakaumaa noudattava suure. Esim. iälle saadaan sqrt(21,349) = 4,62>>1,96, joka vastaa merkitsevyystasoa 0,05.
Sarakkeessa **Sig**. on kerrointen tilastollista merkitsevyyttä ilmaisevat **P-arvot**. Tässä tapauksessa molemmat ovat tilastollisesti erittäin merkitseviä. Ohjelman antama arvo 0,000 tulkitaan P<0,001.
Sarakkeessa **Exp(B)** olevat arvot **OR**:iä (ristitulosuhteita), joita yleisesti käytetään myös suhteellisen riskin (RR) arvioina. Iälle saatu arvo OR=1.118 tarkoittaa, että OR lisääntyy 11,8 %, kun ikä lisääntyy yhden yksikön verran, eli tässä yhdellä vuodella. Huom. Käytännössä OR yliarvioi aina RR:ää.
Kahdessa viimeisessä sarakkeessa on Waldin testiin perustuvat **OR:n 95 % luottamusvälit.** Iälle saatu OR:n luottamusväli (1,066, 1,172) tarkoittaa, että todellinen OR:n lisäys ikävuotta kohden on 95 %:n varmuudella välillä 6,6 % - 17,2 %.

Logistinen malli

Mikäli kovariaatteja on p kappaletta $x_1,\ldots, x_p$, niin logistinen malli on muotoa:

$$E(y \mid x) = \pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}}$$

Kaavassa $\beta_0, \beta_1,\ldots, \beta_p$ ovat mallin parametreja.

Logistisessa regressiossa näyttelee keskeistä roolia **logit-muunnos**, joka π(x):n avulla ilmaistuna on:

$$g(x) = \log_e \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Kun oletetaan, että x on annettu, niin logistisen mallin mukaisesti lopputulos y saadaan yhtälöstä: **y = π(x) + ε**, missä **ε** saa jommankumman **kahdesta** mahdollisesta arvosta:

1) Jos y = 1, ε =1 - π(x) todennäköisyydellä π(x)
2) Jos y = 0, ε = - π(x) todennäköisyydellä 1 - π(x)

Tämä tarkoittaa sitä, että virhetermin ε keskiarvo on nolla ja varianssi π(x)·(1 - π(x)), ja ε noudattaa **binomijakaumaa** toisin kuin lineaarisessa regressiossa, missä virhetermin oletettiin noudattavan standardia normaalijakaumaa. Tässä onkin yksi merkittävimmistä peruseroista näiden kahden regression välillä.

Olettamuksista

Cornfield lähti logististen funktioiden johtamisessa olettamuksesta, että muuttujat x_1, x_2, …, x_p yhdessä ovat multinormaalisia ja että kovarianssisrakenteet ovat samat kummassakin ryhmässä. Tällöin logistisessa funktiossa kertoimien β_0, β_1,…, β_p estimaatit b_0, b_1, …, b_p ovat samat kuin erotteluanalyysistä saatavat.

Logistisessa mallissa ei tarvitse kuitenkaan tehdä kovarianssirakenteiden yhtäläisyysoletusta, sillä tavallisesti logistisen mallin kertoimet arvioidaan käyttäen nk. suurimman uskottavuuden menetelmää (“maximum likelihood”) joka ei perustu ollenkaan kovarianssien laskemiseen.

Myöskään oletus multinormaalisuudesta ei ole tarpeellinen ja varsin yleisesti lääketieteellisissä tutkimuksissa käytetäänkin mallissa määrällisten muuttujien ohella luokiteltuja ja kaksiarvoisia muuttujia. Havaintojen riippumattomuusoletus on logistisessa mallissakin välttämätön.

Sairastumisriskin arviointi

Lopputuloksen y = 1 (esim. sairastumisriski) todennäköisyys, jos x=(x_1, …, x_p) on annettu, saadaan logistisen mallin mukaisesti kaavasta:

$$P_x = P(y = 1 | x) = \frac{1}{1 + \exp(-g(x))}$$

Vaihtoehtoisia malleja:

1. Lineaarinen regressio: $P_x = g(x)$
2. Probit-malli $P_x = \int_{-\infty}^{g(x)} \exp(-\frac{u^2}{2}) du$

   Probit-mallia käytetään yleisesti mm. annos-vaste tutkimuksissa (“bioassay”)

3. Coxin malli: $P_x = \lambda(t | x) = \lambda_0(t) \exp(g(x))$

Coxin mallia käytetään elossaolotutkimuksissa, kun halutaan huomioida kovariaattien vaikutus tutkittavan tapahtuman ilmaantumiseen.

Esim.

Kohorttitutkimus, 12 vuoden sepelvaltimotauti-ilmaantuvuus (Lähde: Truett et al. J Chron Dis 20: 511 -524, 1967). Framingham-tutkimus.

Aineisto: Yhteensä 742 40- 49 vuotiasta miestä, joilla alkututkimuksessa ei ollut todettu sepelvaltimotautia.

Muuttuja (x)	Parametri (β)	Estimaatti $\hat{\beta}$	Keskivirhe $SE(\hat{\beta})$	Standardoitu kerroin [4]
Vakiotermi (x_0)	β_0	-13.26		
Ikä (vuosia) (x_1)	β_1	0.1216	0.0437	0.337
Kolesteroli (mg/dl) (x_2)	β_2	0.0070	0.0025	0.303
Systolinen verenpaine (mmHg) (x_3)	β_3	0.0068	0.0060	0.132
Suhteellinen paino[1] (x_4)	β_4	0.0257	0.0091	0.345
Hemoglobiini (g/dl) (x_5)	β_5	-0.0010	0.0098	-0.001
Tupakointi[2] (x_6)	β_6	0.4223	0.1031	0.495
EKG löydös[3] (x_1)	β_7	0.7206	0.4009	0.175

[1] 100· paino / (“sukupuoli-pituus”-ryhmän mediaani)

[2] 1 = alle “toppa” päivässä, 2 = “toppa” päivässä, 3 = yli “toppa” päivässä

[3] 0 = normaali, 1 = epänormaali

[4] Standardoitu kerroin: $\hat{\beta}_i \cdot SD(x_i)$ kuvastaa muuttujien suhteellista merkitystä.

Esim.

40- 49-vuotiaan miehen, jolla on riskitekijät x = =(x_1, …, x_7) = (45, 210, 130, 100, 120, 0, 0) seurannan alussa, todennäköisyys sairastua sepelvaltimotautiin 12 vuoden kuluessa.

$$P(y=1|x) = \frac{1}{1+\exp\left[-(-13.26+(0.1216)\cdot(45)+(0.0070)\cdot(210)+\ldots+(0.7206)\cdot(0))\right]} = 0.048$$

Tulkinta:

Keskimäärin kyseisen riskitekijäprofiilin omaavalla tutkimuksen kohteena olevaan perusjoukkoon kuuluvalla henkilöllä on **4.8 %**:n todennäköisyys sairastua sepelvaltimotautiin 12 vuoden kuluessa. Luottamusvälin laskeminen tälle arviolle on laskennallisesti melko työlästä. Siihen tarvitaan arviot paitsi kertoimien keskivirheistä myös kovariansseista.

Mallin parametrien tulkinta

Logistisessa mallissa saadaan suoraan arvioitua yksittäisen muuttujan x_i aiheuttama riskin lisäys, kuten edellä todettiin. Jos riskitekijä x_i on dikotominen siten, että 0 ilmaisee riskin puuttumista ja 1 sen olemassaoloa, niin tällöin $\exp(\beta_i)$ ilmaisee suhteellista sairastumisriskiä niillä, joilla on ko. riskitekijä verrattuna niihin, joilla sitä ei ole (ks. esim. Schlesselman s. 239).

Mikäli riskitekijä on k-luokkainen, voidaan menetellä siten, että muodostetaan (osa tilasto-ohjelmistoista suorittaa tämän automaattisesti) k-1 kappaletta (0, 1)-indikaattorimuuttujia valitsemalla jokin luokka vertailuluokaksi (esim. tupakoimattomat). Tällöin $\exp(\beta_I)$, missä I on jokin näistä indikaattoreista, ilmaisee kyseisen luokan suhteellisen sairastumisriskin vertailuluokkaan nähden.

Jos x_i on jatkuva, niin tulkinta on samankaltainen kuin regressioanalyysissäkin, ts. tällöin $\exp(\beta_i)$ ilmaisee yhden yksikön muutoksen vaikutuksen suhteelliseen riskiin ja vastaavasti jos esim. riski lisääntyy k:n yksikön verran, suhteellinen sairastumisriski lisääntyy $\exp(k\cdot\beta_i)$:n verran.

Mikäli tarkastellaan kahden dikotomisen riskitekijän x_i ja x_j samanaikaista muutosta, niin suhteellinen riskin muutos on $\exp(\beta_i+ \beta_j)$.

Logistisen analyysin suorittaminen ja kerrointen estimointi voidaan suorittaa tapaus-verrokkiasetelmassa (sairauslähtöinen asetelma) samoin kuin seurantatutkimuksessakin (Prentice, 1976) ja β-kerrointen tulkinta on samankaltaista. Seurantatutkimusasetelman pohjalta tehdyssä logistisessa analyysissä $\exp(\beta_0)$ ilmaisee eräänlaisen perusriskitason. Sen sijaan tapaus-verrokkitutkimusasetelmassa kertoimella β_0 ei ole mitään mielekästä tulkintaa, sillä se on täysin keinotekoinen riippuen tapauksien ja verrokkien suhteesta. Se voidaan tosin muuntaa sopivalla muunnoskaavalla sellaiseksi, että sille saadaan mielekäs tulkinta keskimääräisenä riskinä (ks. esim. Afifi ja Clark s. 301).

Esimerkkejä

a) Yksittäisen muuttujan vaikutus sairastumisriskiin P(y = 1 | x):

Esim. Henkilöt A ja B, joilla riskitekijät:

- x_A = (45, 210, 130, 100, 120, 0, 0),
- x_B = (45, 210, 130, 100, 120, 3, 0)

Logistisen mallin perusteella: $P(y = 1 \mid x_A)$ = **0.0483**, $P (y = 1 \mid x_B)$ = **0.1526,** joten **riskisuhde** (RR) on: RR = 0.1526 / 0.0483 = **3.16**

Tulkinta:

Tupakointi (enemmän kuin “toppa” päivässä) lisää henkilön sairastumisriskin sepelvaltimotapahtumien suhteen 12 vuoden ajalla noin kolminkertaiseksi verrattuna henkilöön A, joka ei tupakoi, mutta muut riskitekijät ovat samat.

b) Suhteellinen sairastumisriski OR (“relative odds of disease”)

$$\frac{P(y=1\mid x)}{P(y=0\mid x)} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

Huom.

Yhden yksikön muutos muuttujassa x_i muuttaa suhteellista riskiä $\exp(\beta_i)$ verran.

Esim.

Iän lisäys vuodella edellä olevassa esimerkissä lisää riskiä exp(0.1216) = 1.13 eli **13 %.** Mallin kertoimien pereusteella todetaan, että vastaava riskin lisäys saadaan kasvattamalla kolesterolitasoa 0.1216 /0.0070 = 17.4 (mg/dl)

Huom.

Tupakoiminen enemmän kuin “toppa”/pv lisää suhteellista riskiä
exp(0.4223· 3) = **3.55** (vrt. RR = **3.16**)

c) Kahden riskitekijän samanaikainen muutos

Oletetaan, että molemmissa riskitekijöissä x_i ja x_j tapahtuu **yhden yksikön muutos**, eli: x_i muuttuu arvoksi $x_i + 1$ ja x_j arvoksi $x_j + 1$. Nämä riskitekijämuutokset aiheuttavat vastaavasti OR:n muutoksen: $\exp(\beta_i + \beta_j) = \exp(\beta_i) \cdot \exp(\beta_j)$

Esim.

Ikä lisääntyy vuodella ja henkilö aloittaa tupakoinnin, ts. riskitekijöissä tapahtuu muutokset: x_1: 45→ 46 (esim.) ja x_6: 0 → 1, jolloin riskin lisäys on
exp(0.1216 + 0.4223) = **1.72**

Yleisesti:

Henkilön A, jolla on riskitekijät $\mathbf{x}^A = (x_1^A, ..., x_p^A)$ OR (”odds ratio”) verrattuna henkilöön B, jolla on riskitekijät $\mathbf{x}^B = (x_1^B, ..., x_p^B)$ on:

$$OR = \frac{P(y=1 \mid \mathbf{x}^A) \cdot P(y=0 \mid \mathbf{x}^B)}{P(y=0 \mid \mathbf{x}^A) \cdot P(y=1 \mid \mathbf{x}^B)} = \exp\left(\sum_{i=1}^{p} \beta_i \cdot (x_i^A - x_i^B)\right)$$

Huom.

Jos ero on vain muuttujassa x_k, niin $\mathbf{OR = \exp(\beta_k \cdot (x_k^A - x_k^B))}$
tai jos x_k on dikotominen (0, 1), niin OR = $\exp(\beta_k)$

Huom.

Jos kaikki riskitekijät ovat dikotomisia, niin $\exp(\beta_0)$ on sellaisen henkilön sairastumisriski, jolla kaikki x:t ovat nollia. Tapaus-verrokkitutkimuksessa vakiotermillä ei ole kovin mielekästä tulkintaa; silloin β_0:n tilalla on $\beta_0 + \log_e(\pi_1/\pi_0)$, missä π_1 on todennäköisyys, että sairas henkilö tulee mukaan otantaan ja π_0 on vastaava todennäköisyys terveelle henkilölle.

d) Muuttujien suhteellinen merkitys

Muuttujien suhteellista merkitystä voidaan, samoin kuin lineaarisessa regressiossakin, arvioida standardoitujen kertoimien $\hat{\beta}_i \cdot \mathbf{SD}(x_i)$ avulla. Tämä standardoitu kerroin kuvastaa kuinka paljon sairastumisriskin “logit”, eli g(x), muuttuu, kun x_i:n arvoa muutetaan yhden SD(x_i):n verran.

Esim.

Edellä olevassa esimerkissä sekä iän (x_1) että kolesterolin (x_2) standardoitu kerroin on luokkaa 0.3, eli niiden merkitys logistisessa mallissa on samansuuruinen.

Toinen, etenkin preventiotutkimuksissa käytetty, tapa arvioida muuttujien suhteellista merkitystä on seuraava: Arvioidaan ensin kuinka paljon preventiotoimenpitein olisi realistista muuttaa muuttujien arvoja. Oletetaan, että nämä muutokset olisivat: Δx_1, Δx_2,…, Δx_p. Muuttujien suhteellinen merkitys preventio-ohjelmassa saadaan lajittelemalla ne $\hat{\beta}_i \cdot \Delta x_i$:n mukaiseen suuruusjärjestykseen.

Kumpikin näistä menettelytavoista saattaa kuitenkin olla käytännössä hyvinkin ongelmallista ja harhaanjohtavaa siksi, että muuttujat x_i ovat tavallisesti korreloituneita keskenään.

e) Osite-efektin arviointi

Jos aineisto on jaettu ositteisiin, niin ositteen efekti saadaan huomioiduksi mallilla:

$$\mathbf{logit(P_k)} = \beta_{0,k} + \sum \beta_i x_i \ ,$$

jolloin ositteeseen k liittyvä osite-efekti on: $\mathbf{RR_k = exp(\beta_{0,k})}$

Kerrointen merkitsevyyden testaaminen ja luottamusvälit

Logistisen mallin kertoimien arviointi perustuu uskottavuusfunktion logaritmin (LL, "log-likelihood") $\mathbf{\log_e(L) = \sum [y \cdot \log_e \hat{y} + (1-y) \cdot \log_e (1-\hat{y})]}$ maksimointiin. Kaavassa y on havaittu ja $\hat{\mathbf{y}}$ mallin perusteella ennustettu selitettävän muuttujan arvo.

Ongelma:

Kertooko malli, jossa on mukana tietty muuttuja x_i, enemmän lopputuloksesta kuin malli jossa x_i:tä ei ole

Huom.

Tämä on eri asia kuin mallin hyvyyden testaaminen ("goodness of fit")

Merkitään malliin M_k liittyvää uskottavuusfunktion logaritmia **$LL(M_k)$**:llä. Mikäli k=0, niin kyseessä on perusmalli, missä ei ole muuta kuin vakiotermi.

Yllä olevaan testausongelmaan soveltuu parhaiten uskottavuussuhdetesti ("likelihood ratio test"), **G-testi.** Määritellään suure:

Poikkeama D "deviance D" on mitta, joka ilmaisee kuinka paljon tietty malli (M) poikkeaa havaintoaineistoon sovitetusta saturoidusta mallista (M_S). Se lasketaan näihin malleihin liittyvien uskottavuussuhteiden L ja L_S perusteella kaavalla: $D= -2 \cdot (\log_e L - \log_e L_S)$. D saa suuren arvon, kun L on pieni suhteessa L_S:ään. Tämä merkitsee, että malli M on huono. Mikäli malli on hyvä, D saa pienen arvon. D noudattaa asymptoottisesti χ^2-jakaumaa siten, että vapausasteiden määrä on vertailtavien kahden mallin M ja M_s parametrien erotus.

Merkitään M_k:lla mallia, missä muuttuja x_k on mukana ja M_{-k}:llä mallia, mistä x_k on poistettu. G-testi lasketaan seuraavasti:

Testisuure: $G=D(M_k) - D(M_{-k})$, eli

$$\mathbf{G = [-2 \cdot LL(M_{-k})] - [-2 \cdot LL(M_k)] = 2 \cdot [LL(M_k) - LL(M_{-k})]}$$

Yleisesti, jos vertaillaan kahta mallia M_1 ja M_2 siten, että malliin M_1 sisältyvät muuttujat muodostavat osajoukon malliin M_2 sisältyvistä muuttujista, niin mallien eroa testaa suure G

= 2· [LL(M_2)-LL(M_1)]. G-testisuure noudattaa khi^2-jakaumaa parametrein $\mathbf{df_{M_2} - df_{M_1}}$. SPSS:ssä tästä testisuureesta käytetään nimitystä: **"-2 log LR"**. Pieni P-arvo tarkoittaa, että mallien M_1 ja M_2 välinen ero on tilastollisesti merkitsevä, joten mallin M_2 ennustekyky on parempi kuin mallin M_1.

Logistisen mallin yksittäisen kertoimen tilastollista merkitsevyyttä, ts. hypoteesia H_0: $\beta_i = 0$, voidaan testata **Waldin testisuureella:**

$$z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Nollahypoteesin vallitessa z noudattaa standardoitua normaalijakaumaa N(0,1).

Huom.

Waldin testi ei huomioi muita mallissa mukana olevia muuttujia ja antaa siten yleensä liian optimistisen kuvan testattavan kertoimen todellisesta tilastollisesta merkitsevyydestä.

Yksittäiseen muuttujaan x_i liittyvä ristitulosuhde, OR ("odds ratio") ja sen luottamusväli:

Oletetaan, että halutaan verrata OR: n muutosta kahden $\mathbf{x_i}$:n arvon välillä. Merkitään näitä arvoja $\mathbf{x_i^A}$ **ja** $\mathbf{x_i^B}$. Muutosta kuvaava OR ja sen 100· (1-α) %:n luottamusväli on:

$$OR = \exp(\hat{\beta}_i \cdot (x_i^A - x_i^B)), CI_{100(1-\alpha)\%}(OR) = \exp(\hat{\beta}_i \cdot (x_i^A - x_i^B) \pm z_\alpha \cdot SE(\hat{\beta}_i) \cdot (x_i^A - x_i^B))$$

Mikäli muuttuja x_i on dikotominen (0, 1), esim. indikaattorimuuttuja, niin yllä oleva kaava pelkistyy muotoon:

$$OR = \exp(\hat{\beta}_i), CI_{100(1-\alpha)\%}(OR) = \exp(\hat{\beta}_i \cdot (x_i^A - x_i^B) \pm z_\alpha \cdot SE(\hat{\beta}_i) \cdot (x_i^A - x_i^B))$$

Yksittäisen muuttujan merkitystä voidaan arvioida myös tämän luottamusvälin perusteella tarkastelemalla sisältyykö luku 1 väliin vai ei;

Mallin hyvyyden arviointi

Mallin hyvyyttä, ts. yhteensopivuutta havaintoaineiston kanssa voidaan arvioida monella eri tavalla. Kaksi perusmenettelyä on: 1) havaintoaineiston luokittelu logistisen mallin perusteella 2) yhteensopivuustestit.

Luokittelu

Lasketaan logistisen mallin perusteella jokaiselle havaintoyksikölle (esim. henkilölle) todennäköisyys tarkasteltavaan tapahtumaan (esim. sairastumiseen). Luokittelusääntöjä käyttäen muodostetaan nelikenttä:

	Ennustettu y:n arvo 1	0
Todellinen y:n arvo 1	Oikein	Väärin
0	Väärin	Oikein

Mikäli mallin perusteella pyritään ennustamaan yksittäisten henkilöiden sairastumista, niin tulosten "oikein" osuus pitää olla suuri (ainakin >80 %). Mikäli yllä oleva luokittelu tehdään samalla aineistolla kuin millä logistisen funktion kertoimet on arvioitu, niin saadaan yleensä liian optimistinen kuva todellisesta tilanteesta. Realistisemman arvion saa, kun käyttää nk. **"Jackknife"**- menetelmää (sisältyy SPSS:n optioihin) Jacknife-menetelmässä simuloidaan todellista luokittelutilannetta menettelemällä seuraavasti: Jätetään aina vuorollaan kukin luokiteltava henkilö pois, lasketaan logistisen funktion kertoimet jäljelle jäävien n-1:n henkilön perusteella ja suoritetaan poisjätetyn henkilön luokittelu näin saaduilla kertoimilla.

Yhteensopivuustestit

1) z^2-testi yhteensopivuustesti (sisältyy SPSS-pakettiin)

$$z^2 = \sum \frac{(y - \hat{y})^2}{\hat{y}(1 - \hat{y})},$$ missä y viittaa havaittuun arvoon ja $\hat{y}$ ennustettuun arvoon.

Huom.

Osoittajassa oleva suure $(y - \hat{y})^2$ on sama kuin residuaalin neliö ja kaavassa oleva summaus tehdään yli kaikkien havaintoyksiköiden.

2) χ^2-yhteensopivuustesti, uskottavuussuhdetesti:

$$G = 2 \cdot [LL(M_k) - LL(M_0)],$$

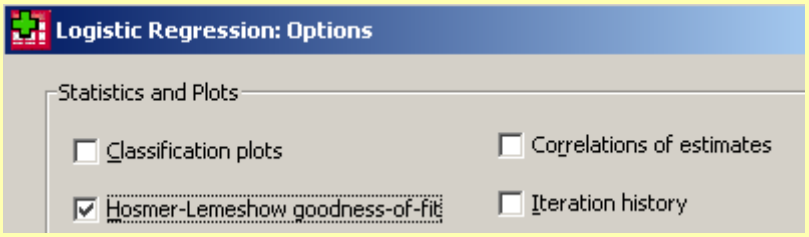
Kaavassa $LL(M_k)$ ja $LL(M_0)$ viittaavat koko mallin ja nollamallin (vain vakiotermi mukana) uskottavuussuhteisiin.

3) Hosmer–Lemeshovin testi

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i \cdot (1 - \frac{E_i}{n_i})}, \quad df = g - 2$$

Suuri testisuureen arvo merkitsee huonoa yhteensopivuutta. Testi perustuu ryhmiteltyyn aineistoon. Kaavassa n_i on ryhmän i koko ja g on ryhmien lukumäärä. Tavallisesti g=10 (desiilijako), joten summauksessa i=1,…, 10. Mikäli ryhmät perustuvat desiilijakoon, niin ryhmä i=1, muodostuu niistä henkilöistä, joilla tapahtuman y=1 todennäköisyys on pienin ja ryhmä i=10 niistä joilla todennäköisyys on suurin, esim. ensimmäiseen ryhmään kuuluvat ne, joilla ennustetodennäköisyys on ≤ 0.1 ja viimeiseen ryhmään ne, joilla ennustetodennäköisyys ≥0.9. Kunkin desiilin havaituille tapahtumien määrille (O_i) lasketaan logistisen mallin perustuvat odotetut tapahtumat (E_i).

Esim. Ikä ja sepelvaltimotautiaineisto.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	,902	8	,999

Tulkinta:

Yhteensopivuus logistisen mallin kanssa on erittäin hyvä; testisuureen arvo on pieni, joten P-arvo on suuri.

Contingency Table for Hosmer and Lemeshow Test

		Chd = 0		Chd = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	9	9,222	1	,778	10
	2	9	8,656	1	1,344	10
	3	8	8,106	2	1,894	10
	4	8	8,048	3	2,952	11
	5	7	6,953	4	4,047	11
	6	5	5,323	5	4,677	10
	7	5	4,194	5	5,806	10
	8	3	3,722	10	9,278	13
	9	2	2,121	8	7,879	10
	10	1	,655	4	4,345	5

4) Mallin hyvyyttä voidaan arvioida myös kahdella uskottavuusfunktiota hyödyntävällä suureella, jotka vastaavat lineaarisen regression R^2-suuretta, mallin selittämän varianssin osuutta.

Coxin ja Snellin R^2

$$R^2 = 1 - \exp\left(\frac{2}{N}\left[LL(M_k) - LL(M_0)\right]\right)$$

Kaavassa $LL(M_k)$ ja $LL(M_0)$ viittaavat koko mallin ja nollamallin (vain vakiotermi mukana) uskottavuussuhteisiin ja N on aineistokoko. Mikäli malli selittäisi täysin y:ssä esiintyvän

vaihtelun, niin R^2 olisi yksi. Coxin ja Snellin suure ei kuitenkaan saavuta koskaan tätä maksimiarvoa.

Nagelkerken adjustoidulla R^2-suureella tätä ongelmaa ei ole. Suure lasketaan kaavalla:

$$R^2_{adj} = \frac{R^2}{1 - \exp(\frac{2}{N} \cdot LL(M_0))}$$

Esim.

Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

Kolmen hoidon P, Q ja R vertailu elossaolostatuksen suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja. Muuttuja **"Status"** ilmoittaa potilaan tilan seurannan päättyessä: 0=ei tapahtumaa, 1=tapahtuma, 2=kadotettu seurannasta (tulkitaan tässä analyysissä puuttuvaksi tiedoksi). Tapahtumana on siten arvo 1. SPSS:llä tehtävä suoritetaan seuraavasti:

SPSS: "Analyze"►"Regression"►"Binary logistic". Asetetaan kohtaan **"Dependent"** muuttuja **"Status2"** Kohtaan **"Covariates"** asetetaan muuttuja **"Therapy".** Kohdassa **"Caterorical"** tehdään muuttujasta "**Therapy"** indikaattoreita valiten referenssikategoriaksi 1, eli lumehoitoryhmä P. Kohdasta **"Options"►"Statistics and Plots"** valitaan **"CI for exp(B)"**.

Malli 1: Mukana on vain hoitoryhmä ("Therapy")

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	49	98,0
	Missing Cases	1	2,0
	Total	50	100,0
Unselected Cases		0	,0
Total		50	100,0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
Censored	0
Dead	1

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
Therapy	Group P	15	,000	,000
	Group Q	16	1,000	,000
	Group R	18	,000	1,000

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Therapy	2,654	2	,265
		Therapy(1)	,352	1	,553
		Therapy(2)	2,563	1	,109
	Overall Statistics		2,654	2	,265

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	57,737[a]	,052	,074

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Therapy			2,565	2	,277	
	Therapy(1)	-,288	,866	,110	1	,740	,750
	Therapy(2)	-1,163	,801	2,108	1	,146	,313
	Constant	1,386	,645	4,612	1	,032	4,000

a. Variable(s) entered on step 1: Therapy.

Tulkinta:

Hoitoryhmissä Q ja R OR:t, eli exp(B):t, ovat 0,750 ja 0,313, eli kummassakin hoitoryhmässä on lievä suojaava vaikutus, koska OR:t ovat alle 1, mutta se ei ole tilastollisesti merkitsevä. Nagelkarken R^2 on huono, eli 7,4 %

Malli 2: Taustamuuttujien vaikutus loppustatukseen

Taustamuuttujat: ikä (**"Age"**), sukupuoli (**"Gender"**) koodaus 1=mies, 2=nainen, tupakointi **"Smoker"**) koodaus: 0=ei tupakoi, 1=tupakoi, alkoholin kulutus/viikko (yksikkö=ravintola-annos) (**"Alcohol"**) koodaus: 0 = ei käytä, 1=1-7 annosta/vk, 2=8-21 annosta/vk, 3= yli 21 annosta/vk

Muuttujien väliset riippuvuudet:

		Smoking status		
		Non-smoker	*Smoker*	*Total*
Gender	*Male*	6	18	24
	Female	11	15	26
Total		17	33	50

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,666(b)	1	,197		
Likelihood Ratio	1,686	1	,194		
Fisher's Exact Test				,242	,161
N of Valid Cases	50				

a) Computed only for a 2x2 table
b) 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,16.

Tulkinta: Sukupuolella ja tupakoinnilla ei ole tilastollisesti merkitsevää yhteyttä

		Alcohol				Total
		No	A little	Moderately	Much	
Gender	Male	1	5	7	11	24
	Female	6	7	9	4	26
Total		7	12	16	15	50

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,353(a)	3	,061
Likelihood Ratio	7,865	3	,049
Linear-by-Linear Association	6,431	1	,011

a 2 cells (25,0%) have expected count less than 5. The minimum expected count is 3,36.

Tulkinta: Alkoholin käytön määrällä ja on yhteys sukupuoleen; miehet käyttävät enemmän.

		Alcohol				
		No	A little	Moderately	Much	Total
Smoking status	Non-smoker	6	4	4	3	17
	Smoker	1	8	12	12	33
Total		7	12	16	15	50

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,233(a)	3	,017
Continuity Correction			
Likelihood Ratio	10,079	3	,018
Linear-by-Linear Association	7,122	1	,008
N of Valid Cases	50		

a) 3 cells (37,5%) have expected count less than 5. The minimum expected count is 2,38.

Tulkinta: Alkoholin käytön määrällä ja tupakoinnilla on vahva yhteys; tupakoivat käyttävät enemmän.

Correlations:

		Gender	Age	Smoking status	Alcohol
Gender	Pearson Correlation	1	,097	-,183	-,362(**)
	Sig. (2-tailed)		,501	,205	,010
	N	50	50	50	50
Age	Pearson Correlation	,097	1	-,206	-,171
	Sig. (2-tailed)	,501		,152	,234
	N	50	50	50	50
Smoking status	Pearson Correlation	-,183	-,206	1	,381(**)
	Sig. (2-tailed)	,205	,152		,006
	N	50	50	50	50
Alcohol	Pearson Correlation	-,362(**)	-,171	,381(**)	1
	Sig. (2-tailed)	,010	,234	,006	
	N	50	50	50	50

** Correlation is significant at the 0.01 level (2-tailed).

		Status		Total
		No event	Event	
Gender	Male	6	18	24
	Female	9	16	25
Total		15	34	49

Tulkinta: Sukupuolella ei ole vaikutusta loppustatukseen.

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,698(b)	1	,404		
Likelihood Ratio	,701	1	,402		
Fisher's Exact Test				,538	,300
Linear-by-Linear Association	,683	1	,408		
N of Valid Cases	49				

b 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,35.

		Status		Total
		No event	Event	
Smoking status	Non-smoker	10	6	16
	Smoker	5	28	33
Total		15	34	49

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	11,373(b)	1	,001		
Likelihood Ratio	11,123	1	,001		
Fisher's Exact Test				,002	,001
N of Valid Cases	49				

a) Computed only for a 2x2 table
b) 1 cells (25,0%) have expected count less than 5. The minimum expected count is 4,90.

Tulkinta: Tupakoinnilla on selvä vakutus loppustatukseen; tupakoitsijoilla on enemmän tapahtumia.

		Status		Total
		No event	Event	
Alcohol Unit=1 beer	0	5	1	6
	1-7	5	7	12
	8-21	3	13	16
	>21	2	13	15
Total		15	34	49

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11,710(a)	3	,008
Likelihood Ratio	11,434	3	,010
Linear-by-Linear Association	9,809	1	,002
N of Valid Cases	49		

a 5 cells (62,5%) have expected count less than 5. The minimum expected count is 1,84.

Tulkinta: Alkoholin käytön määrällä on selvä vaikutus loppustatukseen; runsaasti käyttävillä on enemmän tapahtumia.

SPSS: Valikot:**"Analyze"►"Regression"►"Binary logistic".** Asetetaan kohtaan **"Dependent"** muuttuja **"Status2"** Kohtaan **"Covariates"** asetetaan tarkasteltavat muuttujat. Kohdassa tehdään muuttujista tarvittaessa indikaattoreita valiten referenssikategoriaksi joko ensimmäinen tai viimeinen. Kohdasta **"Options"►"Statistics and Plots"** valitaan **"CI for exp(B)"**. Mallissa 2 on valittu lisäksi **"Hosmer-Lemeshow goodness of fit"**

Malli 2a: Sukupuoli ja ikä mukana

Variables not in the Equation					
			Score	df	Sig.
Step 0	Variables	Gender	,698	1	,404
		Age	,033	1	,855
	Overall Statistics		,705	2	,703

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	,709	2	,702
	Block	,709	2	,702
	Model	,709	2	,702

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	59,656(a)	,014	,020

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	Gender	-,517	,633	,666	1	,414	,596	,172	2,063
	Age	-,002	,023	,007	1	,932	,998	,954	1,044
	Constant	1,712	1,482	1,334	1	,248	5,541		

Tulkinta:

Sukupuolella ja iällä ei tilastollisesti merkitsevää yhteyttä loppustatukseen.

Malli 2b: Tupakointi

Variables not in the Equation

		Score	df	Sig.
Step 0	Variables **Smoker**	11,373	1	,001
	Overall Statistics	11,373	1	,001

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11,123	1	,001
	Block	11,123	1	,001
	Model	11,123	1	,001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square

1	49,242(a)	,203	,287

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1(a)	Smoker	2,234	,709	9,931	1	,002	9,333	2,327	37,442
	Constant	-,511	,516	,979	1	,323	,600		

a) Variable(s) entered on step 1: Smoker.

Tulkinta:

Tupakointi lisää tapahtumien riskiä yli yhdeksän kertaiseksi. OR=9,33.

Malli 2c: Alkoholi

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Alcohol	10,013	1	,002
	Overall Statistics		10,013	1	,002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	49,913(a)	,192	,271

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	,978	2	,613

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1(a)	Alcohol	1,096	,381	8,292	1	,004	2,992	1,419	6,309
	Constant	-,975	,668	2,132	1	,144	,377		

Tulkinta:

Alkoholin käytön määrällä on tilastollisestoi merkitsevä yhteys lopputulokseen; yhden kulutusluokan muutos kasvattaa OR:n noin kolminkertaiseksi.

Malli 2d: Tupakointi ja alkoholi jatkuvana muuttujana

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Smoker	11,373	1	,001
		Alcohol	10,013	1	,002
	Overall Statistics		15,864	2	,000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	16,654	2	,000
	Block	16,654	2	,000
	Model	16,654	2	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	43,710(a)	,288	,407

a) Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5,041	5	,411

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	Smoker	1,868	,765	5,956	1	,015	6,473	1,444	29,004
	Alcohol	,897	,410	4,797	1	,029	2,453	1,099	5,474
	Constant	-1,795	,850	4,464	1	,035	,166		

a Variable(s) entered on step 1: Smoker, Alcohol.

Tulkinta:

Sekä tupakoinnilla että alkoholin määrällä on itsenäinen tapahtumien riskiä lisäävä vaikutus; molemmat ovat tilastollisesti merkitseviä.

Malli 2e: Tupakointi ja alkoholi kategorisena muuttujana

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)
Alcohol	No	6	,000	,000	,000	
	A little	12	1,000	,000	,000	
	Moderately	16	,000	1,000	,000	
	Much	15	,000	,000	1,000	

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Smoker	11,373	1	,001
		Alcohol	11,710	3	,008
		Alcohol(1)	,914	1	,339
		Alcohol(2)	1,574	1	,210
		Alcohol(3)	3,039	1	,081
	Overall Statistics		16,524	4	,002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	43,353(a)	,293	,414

a) Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3,789	5	,580

Classification Table(a)

Observed			Predicted		
			Status2		Percentage Correct
			No event	Event	
Step 1	Status2	No event	8	7	53,3
		Event	2	32	94,1
	Overall Percentage				81,6

a) The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	Smoker	1,825	,784	5,416	1	,020	6,201	1,334	28,831
	Alcohol			5,136	3	,162			
	Alcohol(1)	1,216	1,349	,813	1	,367	3,375	,240	47,481
	Alcohol(2)	2,361	1,357	3,029	1	,082	10,600	,742	151,355
	Alcohol(3)	2,694	1,421	3,595	1	,058	14,785	,913	239,359
	Constant	-2,069	1,188	3,033	1	,082	,126		

a Variable(s) entered on step 1: Smoker, Alcohol.

Tulkinta:

Kun alkoholi on kategorisena muuttujana, niin Waldin testi ei anna merkisevää tulosta, vaikka ”Score”-testi antaa! Tämä on tyypillistä Waldin testille. Syynä on aineiston pienuus ja tapahtumien vähäinen määrä. Huomaa luottamusvälien laveus!

Malli 2e: Askeltava menetelmä. Tupakointi ja alkoholi jatkuvana muuttujana

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11,123	1	,001
	Block	11,123	1	,001
	Model	11,123	1	,001
Step 2	Step	5,532	1	,019
	Block	16,654	2	,000
	Model	16,654	2	,000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	**Smoker**	**2,234**	**,709**	**9,931**	**1**	**,002**	**9,333**	**2,327**	**37,442**
	Constant	**-,511**	**,516**	**,979**	**1**	**,323**	**,600**		
Step 2(b)	**Smoker**	**1,868**	**,765**	**5,956**	**1**	**,015**	**6,473**	**1,444**	**29,004**
	Alcohol	**,897**	**,410**	**4,797**	**1**	**,029**	**2,453**	**1,099**	**5,474**
	Constant	**-1,795**	**,850**	**4,464**	**1**	**,035**	**,166**		

a) Variable(s) entered on step 1: Smoker.
b) Variable(s) entered on step 2: Alcohol.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	Smoker	**-30,182**	**11,123**	**1**	**,001**
Step 2	Smoker	**-24,956**	**6,203**	**1**	**,013**
	Alcohol	**-24,621**	**5,532**	**1**	**,019**

Variables not in the Equation

		Score	df	Sig.
Step 1	Variables Alcohol	**5,419**	**1**	**,020**
	Overall Statistics	**5,419**	**1**	**,020**

Mallin diagnostiikkaa

Logistisen mallin hyvyyden ja toimivuuden testaamiseksi on käytettävissä yhteensopivuustestien ohella monia muitakin suureita samoin kuin lineaaristen regressiomallien yhteydessä. On erilaisia residuaalien kvantitatiivisia arviointikeinoja, vaikuttavuusmittoja jne.

Esim. Ikä ja sepelvaltimotautiriski

Subject	Age	Chd	Predicted probability	Predicted group
1	20	0	,04278	0
2	23	0	,05879	0
3	24	0	,06527	0
4	25	0	,07242	0
5	25	1	,07242	0
6	26	0	,08028	0
7	26	0	,08028	0
8	28	0	,09836	0
9	28	0	,09836	0
10	29	0	,10871	0
11	30	0	,12000	0
12	30	0	,12000	0
13	30	0	,12000	0
14	30	0	,12000	0
15	30	0	,12000	0
16	31	1	,13228	0
17	32	0	,14562	0
18	32	0	,14562	0
19	33	0	,16005	0
20	33	0	,16005	0
...				

Subject	Age	Chd	Predicted probability	Predicted group
...				
83	57	1	,73480	1
84	57	1	,73480	1
85	57	1	,73480	1
86	58	0	,75596	1
87	58	1	,75596	1
88	58	1	,75596	1
89	59	1	,77595	1
90	59	1	,77595	1
91	60	0	,79474	1
92	60	1	,79474	1
93	61	1	,81234	1
94	62	1	,82875	1
95	62	1	,82875	1
96	63	1	,84401	1
97	64	0	,85814	1
98	64	1	,85814	1
99	65	1	,87118	1
100	69	1	,91354	1

Tulkinta:

Esim. henkilöllä 5 on todettu CHD, vaikka logistisen mallin antama CHD tapahtuman todennäköisyys on pieni; eli **0,07242**. Tämän tapauksen malli luokittelee väärin.

Subject	Difference between observed and predicted probabilities	Logit residual	Standard residual	Normalized residual	Deviance value
1	-,04278	-1,04470	-,29850	-,21142	-,29573
2	-,05879	-1,06246	-,35168	-,24992	-,34810
3	-,06527	-1,06983	-,37130	-,26425	-,36742
4	-,07242	-1,07807	-,39191	-,27941	-,38774
5	,92758	13,80910	2,31608	3,57898	2,29143
6	-,08028	-1,08728	-,41357	-,29543	-,40910
7	-,08028	-1,08728	-,41357	-,29543	-,40910
8	-,09836	-1,10909	-,46013	-,33030	-,45507
9	-,09836	-1,10909	-,46013	-,33030	-,45507
10	-,10871	-1,12197	-,48510	-,34924	-,47976
11	-,12000	-1,13636	-,51124	-,36927	-,50563
12	-,12000	-1,13636	-,51124	-,36927	-,50563
13	-,12000	-1,13636	-,51124	-,36927	-,50563

14	-,12000	-1,13636	-,51124	-,36927	-,50563
15	-,12000	-1,13636	-,51124	-,36927	-,50563
16	,86772	7,55952	2,03351	2,56116	2,01137
17	-,14562	-1,17044	-,56712	-,41284	-,56103
18	-,14562	-1,17044	-,56712	-,41284	-,56103
19	-,16005	-1,19055	-,59690	-,43652	-,59062
20	-,16005	-1,19055	-,59690	-,43652	-,59062
…					

Standardoimaton residuaali (Res) on kunkin havaintoyksikön havaitun ja ennustetun arvon erotus.

Tulkinta:

Esim. henkilön 5 CHD tapahtuman todennäköisyys oli **0,07242**, joten residuaali on 1- 0,07242 **= 0,92758.**

Logit-residuaali = Res/(P•(1-P)), missä P on ennustettu todennäköisyys. Sitä käytetään, kun ennustamiseen käytetään logit-skaalaa.

Standardoitu residuaali (ZRes) on residuaali jaettuna sen standardipoikkeamalla, eli **ZRes = Res/(P•(1-P)).**

Normalisoitua residuaalia käytetään poikkeavien havaintojen tunnistamiseen.

Poikkeama (”Deviance”) = sqrt(-2•Log$_e$(P)). Suuretta käytetään mallin hyvyyden arvioimiseksi kunkin havaintoyksikön kohdalla. Se noudattaa likimain standardia Normaalijakaumaa.

Nyrkkisääntö: Suuri arvo (>2) tarkoittaa, että malli toimii huonosti kyseisen havaintoyksikön kohdalla.

Tulkinta:

Henkilö 5:lle saadaan sqrt(-2•Log$_e$(0,072) = 2,29143 > 2, joten malli toimii huonosti tämän henkilön kohdalla.

Subject	*Analog of Cook's influence statistics*	*Leverage value*	*DFBETA for constant*	*DFBETA for Age*
1	,00084	,01854	-,03299	,00066
2	,00129	,02029	-,04057	,00081
3	,00148	,02076	-,04328	,00086
4	,00169	,02117	-,04605	,00091
5	,27702	,02117	,58984	-,01165
6	,00192	,02150	-,04886	,00096
7	,00192	,02150	-,04886	,00096
8	,00244	,02187	-,05449	,00106
9	,00244	,02187	-,05449	,00106
10	,00273	,02191	-,05724	,00111
11	,00304	,02184	-,05988	,00115
12	,00304	,02184	-,05988	,00115

13	,00304	,02184	-,05988	,00115
14	,00304	,02184	-,05988	,00115
15	,00304	,02184	-,05988	,00115
16	,14517	,02165	,40910	-,00780
17	,00372	,02135	-,06464	,00122
18	,00372	,02135	-,06464	,00122
19	,00408	,02094	-,06662	,00125
20	,00408	,02094	-,06662	,00125
…				

Cookin etäisyys mittaa kunkin havainnon vaikutusta omaan ja muiden havaintojen residuaaleihin.

Esim.
Henkilöillä 5 ja 16 on selvästi muita suurempi arvo Cookin etäisyysmitassa.

Leverage-mitta on sekä käyttötarkoitukseltaan että tulkinnaltaan logistisessa analyysissä hyvin samankaltainen kuin vastaava mitta lineaarisessa regressioanalyysissä Sen avulla etsitään niitä havaintoarvoja, joilla on suuri vaikutus ennustettuihin arvoihin. Se saa arvoja väliltä 0 - 1

DFBeta suureet kertovat kuinka paljon kukin havainto vaikuttaa logistisen mallin kertoimiin.

Interaktion hallinta

Menettelytavat:
- Tulotermien käyttö
- Osittaminen ja erillisten mallien sovitus kuhunkin ositteeseen

Esim. Tulotermin käyttö.
Muuttujien x1 ja x2 interaktion huomioiminen mallissa tapahtuu siten, että malliin otetaan mukaan tulotermi (x1·x2), joten malli on muotoa:

$$\mathbf{logit(P) = g(x) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \gamma \cdot (x_1 \cdot x_2)}$$

Testit:
Interaktiotermin tarpeellisuutta mallissa voidaan testata kahdella tavalla:
1) G-testisuureen avulla vertaamalla mallia M_1, jossa ei ole interaktiotermiä mukana malliin M_2, jossa se on mukana. Testaaminen tapahtuu samalla tapaa kuin edellä lisämuuttujan merkityksen arviointi.

2) Waldin testi: $\mathbf{z = \dfrac{\hat{\gamma}}{SE(\hat{\gamma})}}$

Esim. Ositetut analyysit
Truett, Cornfield, Kannel (1967)

Ikä	$\hat{\beta}$	$SE(\hat{\beta})$
30-39	0.0231	0.0040
40-49	0.0074	0.0027

Interaktion testi:

$$z = \frac{(0.0231 - 0.0074)}{\sqrt{0.0040^2 + 0.0027^2}} = 4.2895,\ P < 0.001 \text{(kaksipuolinen)}$$

Tulos perustui ositettuun analyysiin siten, että kussakin ositteessa tehtiin erilliset logistiset mallit ja lisäksi näissä malleissa adjustointiin muita riskitekijöitä.

Sekoittavien tekijöiden hallinta

Esim. Ositetut analyysit
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/BDEsim2.sav

lle-et-Vilaine- tutkimus: Alkoholin osuus ruokatorvensyövässä, tapaus-verrokki-tutkimus, ikäefektin testaaminen (Breslow & Day, 1980, Vol. 1, s. 211).
Muuttuja: **x** =alkoholialtistus, koodit. x = 1 jos yli 80 g/vrk, muuten x = 0.

Ikäosite Ikäväli	**Koodi (i)**	**Alkoholi (x)**	**Tapaukset (lukumäärä)**	**Yhteensä** (=tapaukset +verrokit)
25-34	1	1	1	10
	1	0	0	106
35-44	2	1	4	30
	2	0	5	169
45-54	3	1	25	54
	3	0	21	159
55-64	4	1	42	69
	4	0	34	173
65-74	5	1	19	37
	5	0	36	124
75+	6	1	5	5
	6	0	8	39
Yhteensä			200	775

Ilmaisinmuuttujat ositteille:

I	z_2	z_3	z_4	z_5	z_6
1 (ref. kategoria)	0	0	0	0	0
2	**1**	0	0	0	0
3	0	**1**	0	0	0
4	0	0	**1**	0	0
5	0	0	0	**1**	0
6	0	0	0	0	**1**

Mallit:

1. $\mathbf{logit(P) = g(x) = \beta_{0,i}}$, missä $\beta_{0,i} = \beta_0 + \sum_{j=2}^{6} \beta_{z_j} z_j$
2. $\mathbf{logit(P) = g(x) = \beta_{0,i} + \beta \cdot x}$
3. $\mathbf{logit(P) = g(x) = \beta_{0,i} + \beta \cdot x + \gamma \cdot x \cdot (i - 3.5)}$

Huom.

exp($\beta_{0,i}$) on ositteeseen i liittyvä riski, exp(β) "tyypilliseen" ikään, 50 vuotta (vastaa koodiarvoa i = 3.5) liittyvä riski ja **exp(γ)** on ikään liittyvä lineaarisen trendin osuus

Tuloksia:

Malli	Parametrien määrä	Vapaus-asteet	Testisuure χ^2	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\gamma}$	SE($\hat{\gamma}$)
1	6	6	101.8	-	-	-	-
2	7	5	9.32	1.67	0.190	-	-
3	8	4	8.50	1.71	0.201	0.125	0.189

Huom.

Mallissa 1 on mukana ainoastaan ositekohtaiset parametrit β_0 ja $\beta_{0,i}$, i=2, …, 6

Tulkinta:

Yhteensopivuus on huono, koska khi^2-testisuurreen arvo on suuri.

Malli 2:

OR = exp(1.67) = 5.31 ja sen 95 %:n luottamusväli: exp(1.67 ± 1.96·0.190) = (3.66, 7.71)

Tulkinta:

Testisuureen arvo paranee olennaisesti, kun x otetaan malliin mukaan.

Vertaamalla malleja 2 ja 3 keskenään todetaan, ettei ikätrendi ole lineaarinen, koska muutos testisuureen arvossa, $\chi^2_1 = 9.32 - 8.50 = 0.82$ ei ole tilastollisesti merkitsevä.

Mallin yleistyksiä:

- alkoholimuuttuja 4-luokkaisena (kolme indikaattoria x_1, x_2, x_3)
- tupakka 4-luokkaisena (kolme indikaattoria x_4, x_5, x_6)
- ikä 6-luokkaisena (ositteet)

Näiden luokiteltujen muuttujien yhdistelynä tulee siten 96 ositetta, joten logistiseen malliin syötettävään tiivistettyyn havaintoaineistoon tulee 96 riviä seuraavasti:

Koodit				Alkoholi (A)			Tupakka (T)			Interaktio A· T		
Hav.	Ikä	Alk.	Tup.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	…	x_{15}
1	1	1	1	0	0	0	0	0	0	0		0
2	1	1	2	0	0	0	1	0	0	0		0
3	1	1	3	0	0	0	0	1	0	0		0
4	1	1	4	0	0	0	0	0	1	0		0
5	1	2	1	1	0	0	0	0	0	0		0
jne.	…	…	…									

Mallin sovitus: Tulokset:

Malli	Muuttujat	Parametrien määrä	Vapausasteet (df.)	Testisuure G
1	Ikä	6	82	246.9
2	Ikä ja alkoholi	9	79	105.9
3	Ikä ja tupakka	9	79	210.3
4	Ikä, alkoholi ja tupakka	12	76	82.3

Malleihin liittyvät hypoteesit:

- Malli 1: Tupakalla ja alkoholilla ei ole vaikutusta sairastumisriskiin.
- Malli 2: Pelkästään alkoholin vaikutus, adjustoituna iän suhteen.
- Malli 3: Pelkästään tupakan vaikutus, adjustoituna iän suhteen.
- Malli 4: Alkoholin ja tupakan efektit (multiplikatiivinen hypoteesi), adjustoituna iän suhteen.

Mallin valinta

Mallin valinnassa huomioonotettavia asioita:

- funktionaalinen muoto
- muuttujien valinta
- interaktiotermien mukaanotto
- muuttujien muunnokset (esim. luokitukset)

Muuttujien valintaongelmia

- askeltava menetelmä on sokea assosiaation luonteelle
- aineistokoko korostuu
- kausaalisia polkuja ei huomioida

Viitteet

Cornfield J: Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Fedr Proc 21: 58-61, 1962.

Hosmer DW, **Lemeshow** S. Applied logistic regression. John Wiley & Sons, New York, 2000. ISBN: 0-471-35632-8. Hinta €127,80

Kleinbaum DG. Logistic regression, A self-learning text, 2nd edition. Springer-Verlag, New York, 2002. ISBN: 0-387-95397-3. Hinta €79,95

Prentice R: Use of logistic model in retrospective studies. Biometrics 32: 599-606, 1976.

Schlesselman JJ: Case-control studies. Design, conduct, analysis. Oxford Univ. Press, 1982.

Elossaoloanalyysit

Väestö- ja kohorttielinaikataulut

Elinaikataulu ”life table” on etenkin aktuaarisissa, kuten. vakuutusmatemaattisissa ja eläkelaskentasovelluksissa, käytetty taulu tai taulukko, jonka avulla kuvataan henkilöjoukon tai väestön todennäköisyyksiä päätyä johonkin tarkastelun kohteena olevaan tilaan (päätetapahtumaan) seuranta-ajan kuluessa. Tarkasteltu päätetapahtuma on tavallisimmin kielteinen (”failure”), esim. kuolema tai jonkin taudin ilmaantuminen, mutta se voi olla myös myönteinen, esim. työhön palaaminen kuntoutuksen jälkeen. Seuranta-aika on taulukossa jaettu sopiviin väleihin, joiden pituuksien ei tarvitse olla yhtä pitkiä. Kunkin välin kohdalla taulukossa ilmoitetaan seurannassa vielä mukana olevien määrä, välin aikana päätetapahtumaan päätyneiden määrä, välillä seurannasta poistuneiden määrä, päätetapahtuman todennäköisyys kyseisellä välillä, kumulatiivinen todennäköisyys ja sen keskivirhe. Lisäksi taulukossa esitetään johdettuja suureita kuten esim. kumulatiivinen elossaolotodennäköisyys ja elinajan odote sekä niiden keskivirheet ja/tai luottamusvälit.

Käyttö:

- eliniän odotteen laskeminen
- muut aktuaariset menetelmät

Väestöelinaikatauluissa oletetaan, että väestön kuolleisuus noudattaisi laskentahetkellä vallitsevaa ikä-sukupuoli -kohtaista kuolleisuutta myös tulevaisuudessa. Niillä on siten käyttöarvoa eräänlaisena yhteenvetona tietyllä hetkellä tietyssä väestössä vallitsevasta kuolleisuudesta.

Käytetään seuraavia merkintöjä:

D_i = kuolleiden määrä ikävälillä i
P_i = keskiväkiluku ikävälillä i
M_i = kuolleisuus = D_i/P_i
a_i = keskimääräinen eletyn ajan osuus ikävälistä, jolloin kuolemat ovat tapahtuneet
$\hat{q}_i$ = kuolintodennäköisyysarvio
SE_{qi} = kuolintodennäköisyyden keskivirhe
$CI_L(q_i)$ = kuolintodennäköisyyden luottamusvälin alaraja
$CI_U(q_i)$ = kuolintodennäköisyyden luottamusvälin yläraja
l_i = keinotekoisen kohortin elossaolevien määrä
d_i = kuolleiden määrä keinotekoisessa kohortissa = $q_i \cdot l_i$
L_i = elettyjen vuosien määrä keinotekoisessa kohortissa kullakin ikävälillä
T_i = kumulattivinen elettyjen vuosien määrä keinotekoisessa kohortissa ikävälistä i alkaen
e_i = jäljellä olavan eliniän odote
SE_{ei} = jäljellä olavan eliniän odotteen keskivirhe
$CI_L(e_i)$ = jäljellä olavan eliniän odotteen luottamusvälin alaraja
$CI_U(e_i)$ = jäljellä olavan eliniän odotteen luottamusvälin yläraja
p_i = elossaolotodennäköisyys kullakin ikävälillä
P_i = kumulatiivinen elossaolotodennäköisyys

Kun henkilö kuolee jollain tietyllä ikävälillä, niin hän on elänyt siitä vain osan. Näiden osien keskiarvoa on merkitty a_i:llä. Vastasyntyneillä a_i on suuruusluokkaa 0,1 ja

muissa ikäluokissa 0,5:n molemmin puolin. Oletetaan, että tarkastelun kohteena on esimerkiksi 10000 hengen keinotekoinen väestöryhmä. Jäljellä oleva keskimääräinen elinaika e_i voidaan laskea seuraavasti (Chiang 1968):

$$M_i = D_i / P_i \qquad \hat{q}_i = \frac{n_i M_i}{1 + (1 - a_i) n_i M_i}$$

$$d_i = l_i \hat{q}_i,\ i = 0,1, \ldots, w, \qquad l_0 = 10000, \qquad l_{i+1} = l_i - d_i,\ i = 0,1, \ldots, w-1$$

$$L_i = n_i (l_i - d_i) + a_i n_i d_i,\ i = 0,1, \ldots, w-1, \qquad L_w = l_w / M_w, \qquad T_i = \sum_{i}^{w} L_i$$

$$\hat{e}_i = T_i / l_i$$

Kaavoissa w on ikävälien lukumäärä, d_i on kuolleiden määrä ikävälillä keinotekoisessa väestöryhmässä, T_i on niiden elinvuosien kokonaismäärä niille henkilöille, jotka ovat olleet elossa ikävälin i alkaessa ja e_i on jäljellä olevien elinvuosien odote ikävälin i alkaessa.

Suureille q_i ja e_i voidaan laskea **keskivirheet** ja luottamusvälit seuraavasti (Chiang 1968):

$$SE(\hat{q}_i) = \sqrt{\hat{q}_i^2 (1 - \hat{q}_i) / D_i} \qquad SE(\hat{e}_\alpha) = \sqrt{\sum_{i=\alpha}^{w-1} \hat{p}_{\alpha i}^2 \left[\hat{e}_{i+1} + (1 - a_i) n_i\right]^2 SE(\hat{p}_i)^2}$$, missä

$$\hat{p}_{\alpha i} = l_i / l_\alpha \text{ ja } SE(\hat{p}_i) = SE(\hat{q}_i), \text{ koska } \hat{p}_i = 1 - \hat{q}_i$$

Normaalijakaumaan perustuvat likimääräiset **95 % luottamusvälit** suureille q_i ja e_i voidaan laskea kaavoilla:

$$\hat{q}_i \pm 1{,}96 \cdot SE(\hat{q}_i) \qquad \hat{e}_i \pm 1{,}96 \cdot SE(\hat{e}_i)$$

Jäljellä olevan eliniän odotteet ja niiden luottamusvälit voidaan laskea mm. osoitteessa: www.mv.helsinki.fi/home/sarna/stats/Väestöelinaikataulu.xls

Esim. Suomen väestötilasto vuodelta 2006 (Tilastokeskus)

Naiset

Ikä	D_i	P_i	e_i	SE_{ei}	$CI_L(e_i)$	$CI_U(e_i)$
0	71	28875	82,67	0,0762	82,53	82,82
1-4	86	140984	81,88	0,0724	81,74	82,02
5-9	18	141563	78,07	0,0695	77,94	78,21
10-14	14	158679	73,12	0,0686	72,99	73,26
15-19	39	159204	68,15	0,0681	68,02	68,29
20-24	43	162144	63,23	0,0670	63,10	63,36
25-29	46	161773	58,31	0,0659	58,18	58,44
30-34	65	153383	53,39	0,0650	53,27	53,52
35-39	110	165214	48,50	0,0637	48,38	48,63

40-44	205	185085	43,65	0,0622	43,53	43,78
45-49	331	186172	38,88	0,0605	38,76	39,00
50-54	590	196010	34,20	0,0584	34,09	34,32
55-59	839	205649	29,68	0,0558	29,57	29,79
60-64	887	159818	25,24	0,0536	25,14	25,35
65-69	1148	137477	20,88	0,0506	20,78	20,98
70-74	1653	114577	16,66	0,0472	16,57	16,75
75-79	2978	109787	12,72	0,0431	12,63	12,80
80-84	4729	84986	9,20	0,0399	9,12	9,27
85-89	4902	47628	6,38	0,0380	6,30	6,45
90-94	3662	18589	4,14	0,0377	4,07	4,22
95-99	1522	4958	2,60	0,0367	2,53	2,67

Miehet

Ikä	D_i	P_i	e_i	SE_{ei}	$CI_L(e_i)$	$CI_U(e_i)$
0	96	30033	75,75	0,0861	75,58	75,92
1-4	110	147143	74,99	0,0828	74,83	75,15
5-9	15	147656	71,21	0,0803	71,05	71,37
10-14	30	165156	66,24	0,0798	66,09	66,40
15-19	124	166143	61,30	0,0792	61,15	61,46
20-24	194	169860	56,52	0,0771	56,37	56,67
25-29	164	169194	51,83	0,0742	51,68	51,97
30-34	193	162205	47,07	0,0722	46,93	47,21
35-39	284	171224	42,33	0,0700	42,20	42,47
40-44	461	190918	37,66	0,0678	37,53	37,80
45-49	791	189147	33,09	0,0657	32,96	33,22
50-54	1251	196272	28,73	0,0630	28,61	28,86
55-59	2005	204428	24,58	0,0603	24,46	24,70
60-64	1996	153214	20,68	0,0581	20,57	20,80
65-69	2302	121620	16,90	0,0552	16,80	17,01
70-74	2870	91351	13,33	0,0521	13,23	13,43
75-79	3796	72900	10,17	0,0484	10,07	10,26
80-84	3701	42275	7,45	0,0465	7,36	7,55
85-89	2428	17038	5,26	0,0440	5,17	5,34
90-94	1252	5065	3,41	0,0902	3,23	3,58
95-99	33	56	1,81	0,3161	1,19	2,43

Eri väestöryhmien välisiä, esim. kuntien A ja B väestöjen, välisiä jäljellä olevan eliniän odotteiden $\hat{e}_{i,A}$ ja $\hat{e}_{i,B}$ eroja ikäluokkakohtaisesti voidaan testata z-testillä, esim. osoitteessa: www.mv.helsinki.fi/home/sarna/stats/z-testiKahdelleSuureelle.xls

$$z = \frac{\hat{e}_{i,A} - \hat{e}_{i,B}}{\sqrt{SE(\hat{e}_{i,A})^2 + SE(\hat{e}_{i,B})^2}}$$

Kohorttielinaikatauluissa elossaolevien määrän laskenta on jonkin verran monimutkaisempaa kuin väestöeliaikatauluissa. Se perustuu samana vuonna syntyneiden henkilöiden ryhmiin (kohortteihin), ja siinä huomioidaan kunakin vuonna vallitseva todellinen kuolleisuus. Koska kuolleisuus on viime vuosikymmeninä pienentynyt, niin edellä lasketut arviot jäljellä olevan eliniän odotteelle ovat jossain määri aliarvioita.

Seurantaelinaikataulut

Käyttö:

Menetelmää käytetään, kun aineisto on ryhmitelty elossaoloaikojen perusteella.

Käsitteitä:

Päätetapahtuma "end-point" on kliinisissä seurantatutkimuksissa tutkimuksen kohteena oleva tapahtuma; esim. potilaan kuolema seuranta-aikavälillä. Päätetapahtuman tulee olla kullekin tutkittavalle selkeästi ja yksiselitteisesti määritelty.

Laskentaedellytykset:

1) Selvästi ja tarkkaan määritelty seurannan alkamis- ja päättymisajanhetki
2) Päätepiste kaksiarvoinen
3) Alkuajankohta voi vaihdella kronologisesti
4) Seuranta-aika voi olla vaihtelevan pituinen
5) Sellaisten seuranta-aikojen osuus, jotka ovat lyhyitä ja joissa tutkittava ei ole päätynyt päätepistetapahtumaan (rajatut, so. elävät ja kadotetut), ei saa olla suuri. Muuten mm. eliniän odotteen laskenta-algoritmi tuottaa aliestimaatin
6) Seuranta-ajoista muodostuvien jakaumien tulisi olla suunnilleen samanlaiset vertailtavilla ryhmillä, ja rajattujen osuudet likimain yhtä suuret

Laskentakaavat:

Riskille alttiina olevien määrä välillä [t, t +1]: $n'_t = n_t - \frac{c_t}{2}$, missä n'_t on adjustoitu määrä, n_t on määrä ajanhetkellä t ja c_t on rajattujen lukumäärä ("censored") eli elävinä poistuneet tai kadotetut. Kaava olettaa, että poistumiset ovat tapahtuneet tasaisesti koko välillä, eli laskennallisesti voidaa ajatella, että välin keskikohdalla ja siitä syystä vähennys on $c_t/2$.

Kuolintodennäköisyys: $q_t = \frac{d_t}{n'_t}$, **Elossaolotodennäköisyys:** $p_t = 1 - q_t$

Kumulatiivinen elossaolotodennäköisyys ja sen keskivirhe:

$$P(t) = p_{t-1} \cdot P(t-1), \text{missä } P(0) = 1, \quad SE(P(t)) = P(t) \cdot \left(\sum_{j=0}^{t-1} \frac{q_j}{n'_j \cdot p_j}\right)^{1/2}$$

Elossa olevien määrä ajanhetkellä t:

$$S(t) = S(0)\cdot p_0\cdot p_1\cdots p_{t-1}$$, missä S(0) on mielivaltainen vakio, esim. S(0) =1000.

$$SE(S(t)) = S(t)\cdot \left(\sum_{j=0}^{t-1} \frac{d_j}{n_j'\cdot (n_j' - d_j)}\right)^{1/2}$$

Yllä olevat keskivirheiden kaavat (Greenwood 1926) antavat liian pieniä keskivirheen arvoja, jos rajattujen osuus (c_t) on suuri. Tämän aliestimoinnin korjaamiseksi Peto et al. ehdottivat vuonna 1977 vähän konservatiivisemman laskennallisesti yksinkertaisemman kaavan:

$$SE(S(t)) = S(t)\cdot \left(\frac{1 - S(t-1)}{n_t'}\right)^{1/2}$$

95 %:n luottamusvälit P(t):lle ja S(t):lle:

$$P(t) \pm 1.96\cdot SE(P(t)), \quad S(t) \pm 1.96\cdot SE(S(t))$$

Näiden likimääräisten normaalijakaumaan perustuvien välien käyttö saattaa johtaa S(t):n mahdottomiin arvoihin; ovat välin [0 %, 100 %] ulkopuolella. Kalbfleischin ja Prenticen 1980 ehdottamalla modifikaatiolla tätä ongelmaa ei ole:

$$S(t)^{\exp(\pm 1.96\cdot s)}, \text{missä } s = \frac{SE(S(t))}{-S(t)\log_e S(t)}$$

Kohortti (seuranta-) elossaoloanalyysit

Yleistä

Käyttö:

Halutaan tutkia aikaa, joka kuluu seurannan aloittamisesta johonkin kriittiseen päätetapahtumaan (“end-point”). Seurannan aloitusajankohta voi olla esim. hoidon aloittaminen ja päätetapahtuma mikä tahansa tarkkaan määritelty tila esim. relapsi tai kuolema. Näitä menetelmiä voidaan käyttää myös silloin, kun ei ole varsinaisesti mistään elossaolosta vaan halutaan tutkia ja mallittaa aikaa, joka kuluu kahdesta tarkkaan määritellystä tapahtumasta toiseen.

Elossaolomenetelmiä on kahta päätyyppiä: 1) **parametrittomat** ja 2) **parametriset** Parametrittomia ovat erityyppiset elossaolotaulut ja **Kaplan–Meierin menetelmä**. Näistä jälkimmäinen on erittäin yleisesti käytetty menetelmä lääketieteellisissä tutkimuksissa.

Esim.

Epästabiilin angiina pectoriksen hoitokokeilu kalsiumin estäjällä. Viite: Gobble E et al. Randomized, double-blind trial of intravenous diltiazem versus glyseryl nitrate for unstable angina pectoris. Lancet 1995;346:1653- 57.

Tutkimuksessa oli 129 potilasta, joilla oli epästabiili angiina pectoris. Potilaat oli satunnaistettu kahteen vertailtavaan ryhmään: A) suonensisäisesti annettu diltiazem ja B) glyseryylitrinitraatti. Päätetapahtumina oli refraktorinen angiina pectoris ja sydäninfarkti sekä niiden yhdistelmä. Tuloksena todettiin, että hoito A vähensi merkitsevästi sepelvaltimotautitapahtumia (P<0.05). Analyysi tehtiin Kaplan-Meier-menetelmällä.

Merkitään T:llä satunnaismuuttujaa, joka kliinisissä kokeissa kuvastaa aikaa seurannan alusta tarkasteltavaan tapahtumaan ja t:llä jotain tiettyä T:n arvoa. Elossa olevien osuuden ajanhetkellä t määrittelee **elossaolofunktio S(t) (”survival function”).** S(t) = P(T≥ t), eli todennäköisyys, että satunnaisesti valittu aineistoon kuuluva henkilöä elää kauemmin kuin t aikayksikköä (esim. vuotta). Mikäli aineistossa ei ole **rajattuja, sensuroituja** (“censored”) havaintoarvoja, niin S(t) on yksinkertaisesti elossa olevien osuus ajanhetkellä t, mikäli rajattuja havaintoja on, niin S(t) arvioidaan tavallisimmin Kaplan–Meierin elossaolokäyrän perusteella.

Rajattu havaintoarvo ”censored observation” on sellainen havaintoarvo, jonka arvoa ei tiedetä, mutta sen olemassaolo tiedetään. Esim. seurantatutkimuksissa tiedetään tutkimuksen päättyessä elossa olevien potilaiden seuranta-aika, mutta ei tiedetä kuinka kauan potilas eläisi, jos tutkimuksen seuranta-aikaa jatkettaisiin. Näin syntyy oikealle rajattu havaintoarvo. Laboratoriotutkimuksissa mittauslaitteen tarkkuuden alittava arvo on esimerkki vasemmalle rajatusta havaintoarvosta.

Mikäli kliinisissä tutkimuksissa päätetapahtumana jokin tietty kuolemansyy, niin rajattu havainto voi syntyä kolmella eri tavalla:

1) Potilas on elossa seurannan päättyessä
2) Potilas katoaa seurannasta, esim. ulkomaille muuton johdosta
3) Potilas kuolee seuranta aikana johonkin muuhun kuin tutkittavaan kuolemansyyhyn.

Tiettynä ajanhetkenä riskiä päätyä tarkasteltavaan tilaan kuvaa funktio **h(t) = P(T=t) / P(T≥t**), josta käytetään nimityksiä **hasardifunktio** tai **riskitiheysfunktio (”hazard function”).**

Hasardifunktio h(t) ”hazard function”, ”hazard rate” on elossaolotutkimuksissa käytetty funktio. Se ilmaisee todennäköisyyden ajan t funktiona sille tapahtumalle, että henkilö, joka on ollut elossa ajanhetkellä t, kuolee seuraavalla “differentiaalisen” pienellä aikavälillä tämän ajanhetken jälkeen.

Parametriset menetelmät perustuvat erilaisiin elossaoloajan todennäköisyysjakaumiin. Niihin sisältyy tuntemattomia parametreja, jotka malleja sovitettaessa arvioidaan havaintoaineiston perusteella. Funktio h(t) kuvastaa tarkastelun kohteena olevaan tilaan päätymisnopeutta; ts. elossaolofunktion derivaattaa (kulmakerrointa):

$$h(t) = -\frac{d}{dt}\left[\log_e S(t)\right]$$

Esim.

- **Eksponentiaalinen malli**: $S(t) = \exp(-\lambda t)$, $h(t) = \lambda$ **(vakio) ,**

- Tähän malliin sisältyvä oletus, että h(t) ei riipu ajasta on usein epärealistinen; esim. välittömästi leikkauksen jälkeen riski kuolla voi aluksi suurempi ja sitten tasaantua. Tällöin sopivampi malli voisi olla **Weibullin malli**:

  $$\mathbf{S(t) = \exp(-(\lambda t)^{\kappa}),\quad h(t) = \kappa\lambda(\lambda t)^{\kappa-1}\ \ (monotoninen\ tai\ vakio),}$$

- On myös tilanteita, jolloin sairastumis- tai kuoleman riski ei käyttäydy monotonisesti. Kyseeseen voisi tulla silloin **log-normaalinen malli,** missä h(t) kasvaa ensin kohti maksimiarvoa ja sitten laskee.

Kaplan–Meierin menetelmä

Kliinisissä tutkimuksissa käytetään paljon elossaolomenetelmiä (”survival methods”), esim. lonkkaproteesien pysyvyys- ja syöpätutkimukset.

Kaplan–Meierin menetelmä on yksi tavallisimmin käytetyistä elossaolokäyrien parametrittomista laskentamenetelmistä. Sen avulla voidaan arvioida kumulatiivinen elossaolotodennäköisyys seurannan alusta tarkastelun kohteena olevaan tapahtumaan, esim. kuolemaan. Kaplan–Meierin menetelmää käytetään erityisesti silloin, kun esiintyy rajattuja havaintoja, mikä tarkoittaa, että kaikille seurattaville ei ole tullut tarkasteltavaa tapahtumaa tutkimuksen päättyessä. Kaplan–Meierin käyrät piirretään tavallisesti porrasfunktioina ja rajatut havainnot merkitään käyriin. Kahden tai useamman ryhmän käyriä voidaan verrata esim. logrank-testillä.

Kaplan–Meierin menetelmän laskenta perustuu seuraavaan yleisperiaatteeseen: Aina, kun missä tahansa seurattavista ryhmistä tulee tutkimuksen kohteena oleva tapahtuma (“event”), esim. kuolema, reinfarkti jne., niin jokaisen tapahtuman jälkeen lasketaan nk. **elossaolotodennäköisyys:**

$$\hat{S}(t) = \prod_{t_i < t} \frac{N - i + 1 - \delta_i}{N - i + 1},$$

jossa N=aineistokoko ja δ_i=1 jos kyseessä on tapahtuma ja δ_i=0, jos kyseessä on rajattu tapaus.

Mikäli jonain ajanhetkenä t tulee vain rajattu tapahtuma, niin $\hat{S}(t)$ ei muutu, joten sitä ei tarvitse laskea uudelleen. Periaatteessa jos tapahtuma-aika voitaisiin määrittää tarkasti, niin jokaisessa aikapisteessä olisi vain yksi tapahtuma. Käytännössä tapahtuma-ajan yksikkönä on esim. päivä tai kuukausi, joten yhteen ajankohtaan voi tulla useampia tapahtumia.

$\hat{S}_N(t) : n$ keskivirhe saadaan kaavasta:

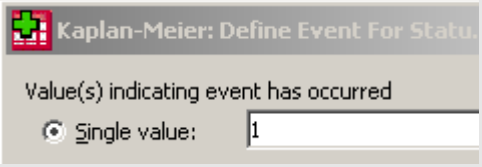
$$SE(\hat{S}(t)) = \hat{S}(t) \cdot \sqrt{\sum_{t_i < t} \frac{\delta_i}{(N - i) \cdot (N - i + 1)}}$$

jossa t_i on i. henkilön seuranta-aika. Elossaolokäyrälle saadaan Normaalijakaumaa perustuvat likimääräiset 95 %:n luottamusvälit seuraavasti: $\hat{S}(t) \pm 1{,}96 \cdot SE(\hat{S}(t))$

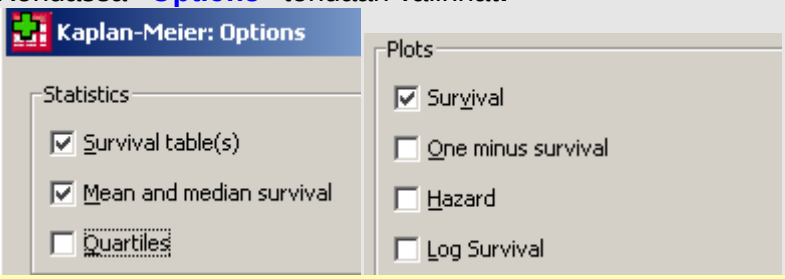
Esim.

Aineisto Trial. Muuttuja **"Time"** on seuranta-aika. Muuttuja **"Status"** ilmoittaa potilaan tilan seurannan päättyessä: 0=ei tapahtumaa, 1=tapahtuma, 2=kadotettu seurannasta. Tapahtumana on siten arvo 1.
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

SPSS": Valikot: **Analyze"►"Survival"►"Kaplan-Meier",** Asetetaan: **"Time"** kohtaan **"Time", "Status"** kohtaan **"Status"** ja kohdassa **"Define Event"** määritellään:



Kohdassa **"Options"** tehdään valinnat**:**



Huom.

Mikäli on kyseessä harvinainen päätetapahtuma, ts. $\hat{S}(t)$ jää koko seurantavälillä pieneksi, niin kannattaa valita optio "One minus survival", jolloin saadaan käyrä $1-\hat{S}(t)$. Optiosta "Hazard" saadaan tapahtumien kumulatiivista kerääntymisnopeutta kuvaava käyrä $-\ln(\hat{S}(t))$

Tulos:

Case Processing Summary

Total N	N of Events	Censored	
		N	Percent
49	34	15	30,6%

Survival Table

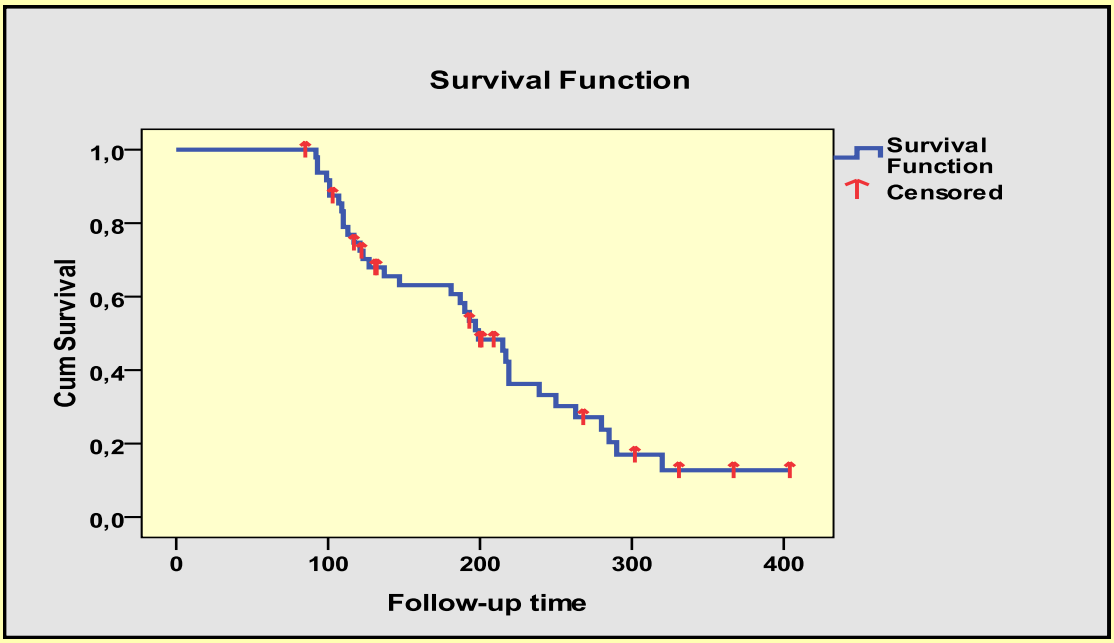
	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate $\hat{S}(t)$	Std. Error $SE(\hat{S}(t))$		
1	92	Dead	,979	,021	1	47
2	93	Dead	.	.	2	46
3	93	Dead	,938	,035	3	45
4	99	Dead	,917	,040	4	44

5	101	Dead	.	.	5	43
6	101	Dead	,875	,048	6	42
7	103	Lossed	.	.	6	41
...						
43	290	Dead	,170	,065	33	5
44	302	Alive	.	.	33	4
45	320	Dead	,127	,061	34	3
46	331	Alive	.	.	34	2
47	367	Lossed	.	.	34	1
48	404	Alive	.	.	34	0

Means and Medians for Survival Time

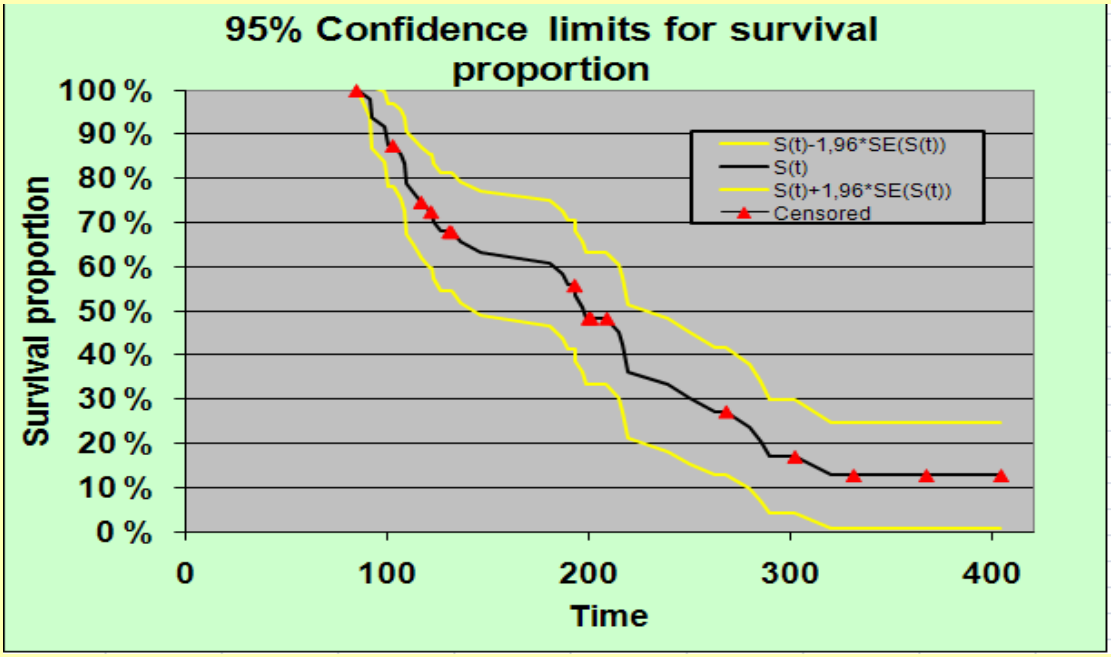
Mean[a]				Median			
		95% Confidence Interval				95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
212,877	15,508	182,482	243,272	199,000	15,184	169,239	228,761

a. Estimation is limited to the largest survival time if it is censored.



SPSS 17 ei laske eikä piirrä elossaolokäyrälle luottamusvälejä, mutta sen tulostamien

suureiden **”Time”**, **”$\hat{S}(t)$”** ja **”SE($\hat{S}(t)$)”** perusteella ne voidaan laskea ja piirtää Excelillä, esim. osoitteessa:
http://www.mv.helsinki.fi/home/sarna/stats/ElossaolokäyränLuottamusvälit.xls



Elossaolokäyrien vertailu

Tavallinen ongelma kliinisissä hoitotutkimuksissa on: saadaanko hoidolla A parempi lopputulos kuin hoidolla B, kun lopputulosta mitataan ajalla, joka kuluu hoidon aloittamisesta johonkin tiettyyn tapahtumaan (esim. kuolemaan). Ryhmien välinen vertailu voidaan suorittaa **logrank-testillä.**

Logrank -testi ”logrank test” on merkitsevyystesti, jolla verrataan kahden tai useamman ryhmän elossaolokäyrien välisiä eroja keskenään. Se on erityissovellus Mantel-Haenszelin khi^2-testistä. Testi voidaan laskea myös ositetusta aineistosta ja siitä on olemassa myös trendiversio. Synonyymi on Mantel-Cox:n testi ja se painottaa kaikkia tapahtumia seurantajakson aikana samalla tavalla, toisin kuin esim. Breslow:n testi.

Logrank-testi soveltuu useammankin kuin kahden ryhmän tapaukseen. Oletetaan yksinkertaisuuden vuoksi, että vertailtavia ryhmiä on vain kaksi, A ja B. Merkitään ajanhetkellä t_i näissä ryhmissä seurannassa mukana olevien määriä (=efektiivinen ryhmäkoko) $n'_{i,A}$ ja $n'_{i,B}$ ja vastaavasti tapahtumien määriä $d_{i,A}$ ja $d_{i,B}$. Merkitään $n'_i = n'_{i,A} + n'_{i,B}$ ja $d_i = d_{i,A} + d_{i,B}$. Mikäli samana ajankohtana ei ole useampia tapahtumia (tasapelejä ”ties”), niin $d_i = 1$ ja $d_{i,A}$ ja $d_{i,B}$ = 0 tai 1.

Logrank-testin laskemiseksi seuranta-ajat yhdistetyssä aineistossa lajitellaan ensin nousevaan järjestykseen. Kun jompaankumpaan tai kumpaankin ryhmään tulee

ajankohtana t_i vähintään yksi tapahtuma, niin konstruoidaan seuraavanlainen nelikenttä

Ryhmä	Tapahtuma	Ei tapahtumaa	Ryhmän koko
A	$d'_{i,A}$	$n'_{i,A} - d_{i,A}$	$n'_{i,A}$
B	$d'_{i,B}$	$n'_{i,B} - d_B$	$n'_{i,B}$
	d_i	$n'_i - d_i$	n'_i

Oletetaan, että halutaan testata nollahypoteesia (H_0): ”tapahtumavaarat ovat kummassakin ryhmässä yhtä suuret”. Mikäli H_0 on voimassa, niin $d'_{i,A}$:n odotusarvo ja **hypergeometriseen** jakaumaan perustuva varianssi ovat:

$$E(d'_{i,A}) = n'_{i,A} \cdot d_i / n'_i \text{ ja } Var(d'_{i,A}) = \frac{d_i(n'_i - d_i)n'_{i,A}n'_{i,B}}{n'^2_i(n'_i - 1)}$$

Kun näin lasketut arvot summataan kaikista yllä olevan kaltaisista nelikentistä, niin saadaan:

$$O_A = \sum_i d_{i,A} \text{ , } E_A = \sum_i E(d_{i,A}) \text{ ja } V_A = \sum_i Var(d_{i,A})$$

H_0:n paikkansapitävyyttä voidaan testata **Mantel–Haenszelin** versiolla **logrank-**testisuureesta (käytössä mm SPSS:ssä):

$$\chi^2_{MH} = \frac{(O_A - E_A)^2}{V_A}$$

Tämä logrank-testisuure noudattaa khii^2-jakaumaa vapausastein yksi. Laskenta voidaan helposti yleistää useammallekin ryhmälle ja silloin vapausasteiden määrä on ryhmien lukumäärä-1. **Mantel–Coxin testi** on synonyymi tälle logrank-testin versiolle.

Ryhmän A ja B havaittujen ja odotettujen tapahtumien välillä vallitsee relaatio: $(O_B - E_B) = -(O_A - E_A)$. Koska neliöön korotus hävittää etumerkin, niin on samantekevää lasketaanko testisuure ryhmän A vai B suhteen. Tästä relaatiosta seuraa:

$$O_B = \sum_i d_{i,B} \text{ ja } E_B = O_A + O_B - E_A$$

Hasardisuhde (HR) ”Hazard ratio” (HR) on elossaolotutkimuksissa käytetty suhteellisen riskin mitta. HR = 1 tarkoittaa siten, että riski päätyä tarkasteltavaan tapahtumaan on sama kummassakin ryhmässä. On olemassa monta tapaa arvioida HR:ää.

a) $HR = \frac{O_A/E_A}{O_B/E_B}$

b) $HR_{MH} = \exp\left(\frac{O_A - E_A}{V}\right)$

Tapa b) toimii luotettavammin kuin tapa a) silloin, kun tapahtumien määrä on pieni.

Hasardisuhteelle voidaan laskea likimääräinen 100(1-α) % luottamusväli kaavoilla (Simon 1986):

Alaraja: $\exp\left(\frac{O_A - E_A}{V_A} - z_{1-\alpha/2} / \sqrt{V_A}\right)$, **Yläraja:** $\exp\left(\frac{O_A - E_A}{V_A} + z_{1-\alpha/2} / \sqrt{V_A}\right)$

Esim.

Diffuusi histiosyyttinen lymfooma (McKelvey et al., 1976) Asteiden 3 ja 4 vertailu. Päätetapahtumana on kuolema.

Aika (pv)	n_3'	n_4'	d_3	d_4	$E(d_3)$	$Var(d_3)$
4	19	61	0	1	0.2375	0.1811
6	19	60	1	1	0.4810	0.3606
10	18	59	0	1	0.2338	0.5278
...	...	...	...	...	...	...
253	7	8	1	0	0.4667	0.2489
Yht.			**O_3=8**	**O_4=46**	**E_3=16,6870**	**11,2471**

Havaittujen ja odotettujen frekvenssien välillä pätee relaatio: $O_3 + O_4 = E_3 + E_4$, joten $E_4 = 8+46 -16.6870=37.3130$ ja $V_A=V_B=11,2471$ sekä $z_{1-\alpha/2} = 1,96$. Logrank-testi antaa tuloksen:

$$\chi^2_{MH} = \frac{(46 - 37,3130)^2}{11,2471} = 6,71,\ P = 0,0096$$

Hasardisuhteeksi ja sen 95 % luottamusväliksi saadaan:

$$HR_{MH} = \exp\left(\frac{46 - 37,3130}{11,2471}\right) = 2,16$$ ja **(1,20, 3,88)**

Tulkinta:

Asteen 4 potilailla on asteen 3 potilaisiin verrattuna tilastollisesti merkitsevästi suurempi, 2,16-kertainen, kuolemanvaara. 95 % varmuudella hasardisuhde on välillä (1,20, 3,88).

Tavallisimmat testit elossaolokäyrien vertaamiseksi

- Mantel–Cox
- Tarone–Ware
- Breslow
- Peto–Perentice

Huom.

Kaikki nämä testit ovat parametrittomia testejä ja niiden toimivuuden kannalta on olennaista, että hasardisuhde pysyy muuttumattomana koko seuranta-ajan, mikä tarkoittaa, että Kaplan–Meier-käyrät eivät saa mennä ristiin.

Mantel–Cox (= logrank testi) antaa **saman** painon kaikille tapahtumille seuranta-aikana.

Breslowin testi on analoginen Kruskal-Wallisin (eli yleistetyn Wilcoxonin) -testin kanssa. Se antaa **enemmän** painoa **aikaisemmille** tapahtumille seuranta-aikana kuin Mantel–Cox. Se ei siten ole yhtä herkkä seurannan lopussa tuleville tapahtumille, jolloin suhteelliset virheet pyrkivät kasvamaan, kun seurattavien määrä pienenee.

Tarone–Ware testisuure on kompromissi Mantel–Coxin ja Breslowin välillä. Peto–Perentice on myös analoginen Kruskal-Wallisin testin kanssa. Tämä testi painottaa havaittujen ja odotettujen tapahtumien erotuksia tietyllä ajanhetkellä käyttäen arvioitua olemassaolofunktiota eikä seurannassa jäljellä olevien määrää kuten Breslow ja Tarone-Ware

Yllä olevista testeistä on olemassa myös **trendin** testaamiseen soveltuvat versiot. Trenditesteillä on käyttöä silloin, kun vertailtavia ryhmiä indikoiva muuttuja on järjestysasteikollinen, ilmaisten esimerkiksi annosmääriä tai ikäryhmiä. Trenditestit ovat samanlaisia kuin (2 x k)-taulukoiden yhteydessä tarkastellut trenditestit. Testisuureet noudattavat khii^2-jakaumaa vapausastein yksi riippumatta vertailtavien ryhmien määrästä. Trenditesteihin voidaan määritellä testattavan trendin muotoa määritteleviä kotrasteja. Lineaarisen trendin yhteydessä kontrasti on -1,0,1. Tämä on oletusarvo esim. SPSS:ssä.

Testit voidaan laskea myös **ositettuna** ja kontrolloida siten sekoittavien tekijöiden vaikutusta samoin kuin **Mantel–Haenszelin** testin yhteydessä tehtiin. Ositteisiin jaon perusteena voi olla esim. ikä tai taudin vaikeusaste. Mikäli sekoittavia tekijöitä on useita, niin kannattaa käyttää Coxin mallia.

Eliniän odotteen laskeminen

Oletetaan, että aineistossa tulee seuranta-aikana yhteensä d tapahtumaa, joihin kuluvat ajat, eliniät, ovat aikajärjestyksessä: $t_1 \leq t_2 \leq \ldots \leq t_d$, ja $\hat{S}(t_1), \hat{S}(t_2), \ldots, \hat{S}(t_d)$ ovat vastaavat Kaplan-Meier menetelmällä tuotetut elossaolotodennäköisyydet. Tällöin **eliniän odote** (**keskimääräinen** odotettavissa oleva elinikä) aineistossa saadaan kaavalla:

$$\hat{\mu} = t_1 + \sum_{i=1}^{d-1} \hat{S}(t_i) \cdot (t_{i+1} - t_i)$$

Huom.

Jos viimeinen tapahtuma aineistossa olisi rajattu havainto (aika t_C), esim. elävänä seurannan päättymisenvuoksi poistettu henkilö, niin kaavaan pitää lisätä termi $\hat{S}(t_d) \cdot (t_c - t_d)$

Eliniän **mediaani** (M) voidaan laskea siten, että etsitään Kaplan-Meier-käyrän se ajankohta x-akselilta, joka vastaa y-akselilla kohtaa 0.5 (50 %), eli toteuttaa yhtälön $\hat{S}(M) = 0{,}5$. Kahden ryhmän välistä hasardisuhdetta voidaan arvioida paitsi edellä

kuvatuilla tavoilla, niin myös mediaanien avulla. Olkoon ryhmien A ja B eliniän mediaanit M_A ja M_B. Tällöin hasardi suhteelle saadaan arvio:

$$HR_{A/B} = M_A/M_B$$

Tämä arvio olettaa, että elinajat noudattavat eksponentiaalista jakaumaa.

Samoin kuin edellä HR:n luottamusväli kannattaa laskea $\log_e$HR:n avulla, sillä tämä logaritminen suure noudattaa likimain Normaalijakaumaa toisin kuin HR. Likimääräinen 100(1-α) % luottamusväli voidaan laskea kaavoilla:

Alaraja: $\exp\left[\log_e HR_{A/B} - z_{1-\alpha/2} \cdot SE(\log_e HR_{A(B)})\right]$,

Yläraja: $\exp\left[\log_e HR_{A/B} + z_{1-\alpha/2} \cdot SE(\log_e HR_{A/B})\right]$, **missä**

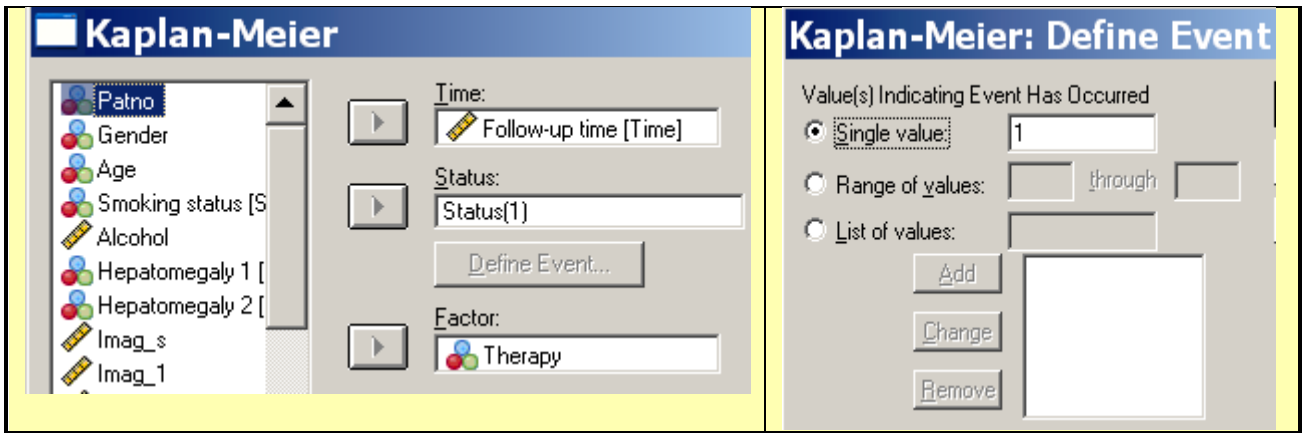
$$SE(\log_e HR_{A/B}) = \sqrt{\frac{1}{O_A} + \frac{1}{O_B}}$$

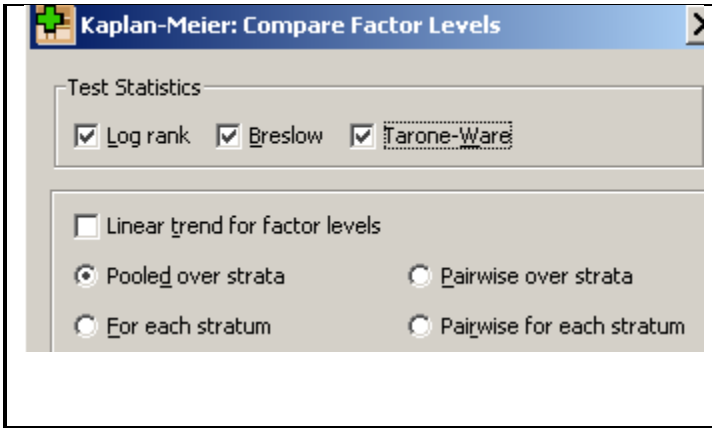
, missä O_A ja O_B ovat havaitut tapausmäärät ryhmissä A ja B

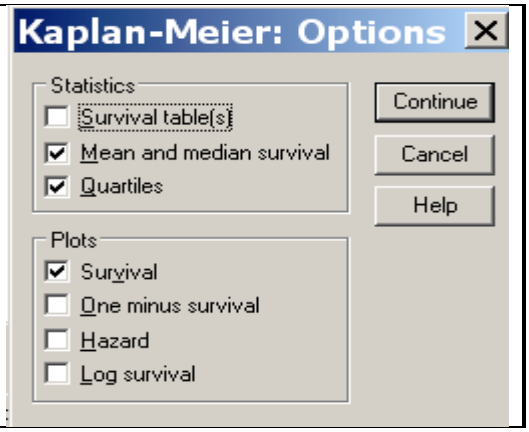
Esim.

Aineisto Trial. Kolmen hoidon P, Q ja R vertailu elossaoloaikojen suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja. Muuttuja **"Time"** on seuranta-aika. Muuttuja **"Status"** ilmoittaa potilaan tilan seurannan päättyessä: 0=ei tapahtumaa, 1=tapahtuma, 2=kadotettu seurannasta. Tapahtumana on siten arvo 1. SPSS:llä tehtävä suoritetaan seuraavasti:
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

SPSS: Valikot: **"Analyze"►"Survival"►"Kaplan-Meier",** Asetetaan: "Time"►**"Time",** "Status"►**"Status"** ja määritellään **Status=1.** Kohtaan **"Factor"** asetetaan muuttuja **"Therapy".** Valikosta **"Compare Factor Levels"** valitaan kaikki testivaihtoehdot ja lisäksi trenditestin laskemiseksi klikataan kohta **"Linear trend for factor levels"**. kohdasta **"Options"►"Statistics"** valitaan **"Mean and median survival"** ja valikosta **"Plots":** klikataan **"Survival".** Näin saadaan seuraavat tulokset:







Case Processing Summary

Therapy	Total N	N of Events	Censored	
			N	Percent
Group P	15	12	3	20,0%
Group Q	16	12	4	25,0%
Group R	18	10	8	44,4%
Overall	49	34	15	30,6%

Means and Medians for Survival Time

Therapy	Mean(a)				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Group P	141,545	12,251	117,533	165,557	123,000	12,241	99,008	146,992
Group Q	194,156	20,281	154,405	233,907	219,000	69,678	82,431	355,569
Group R	277,043	25,643	226,783	327,304	239,000	60,319	120,774	357,226
Overall	212,877	15,508	182,482	243,272	199,000	15,184	169,239	228,761

(a) Estimation is limited to the largest survival time if it is censored.

Huom. Yleensä elossaoloaikojen jakaumat ovat niin vinoja, että kannattaa mieluummin käyttää mediaanielossaoloaikoja kuin eliniän keskiarvoja!

Percentiles

Therapy	25,0%		50,0%		75,0%	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Group P	190,000	33,253	123,000	12,241	101,000	8,693
Group Q	263,000	17,884	219,000	69,678	110,000	8,660

Group R	.	.	239,000	60,319	215,000	20,509
Overall	280,000	21,762	199,000	15,184	117,000	7,452

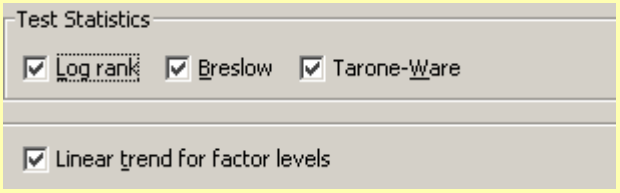
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	19,949	2	,000
Breslow (Generalized Wilcoxon)	13,977	2	,001
Tarone-Ware	16,551	2	,000

Test of equality of survival distributions for the different levels of Therapy.

Tulkinta: Kaikki käytetyt testit antavat tilastollisesti erittäin merkitsevän eron ryhmien välillä elossaoloaikojen suhteen

Elossaolokäyrien välisen **trendin testaaminen** voidaan tehdä SPPS:llä siten, että kohdasta **”Compare Factor Levels”** klikataan kohta **”Linear trend for factor levels”**



Tulos:

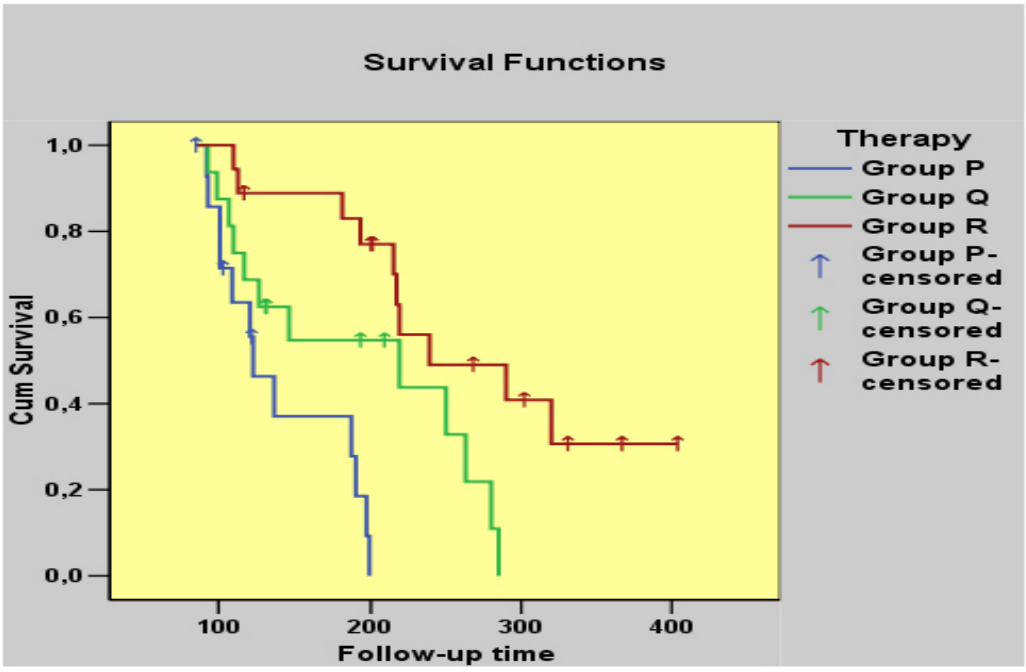
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	18,942	1	,000
Breslow (Generalized Wilcoxon)	13,727	1	,000
Tarone-Ware	16,073	1	,000

The vector of trend weights is -1, 0, 1. This is the default.

Huom. Tämä on nk. lineaarinen kontrasti

Tulkinta: Ryhmien elossaolokäyrien välillä on tilastollisesti erittäin merkitsevän lineaarinen trendi



Coxin regressio

Yleistä

Coxin regressio
”Cox regression” on regressiomenetelmä, jonka avulla mallitetaan elossaoloaikoja. Sitä kutsutaan myös nimellä **verrannollisten riskien malli** “proportional hazards model”, koska se perustuu olettamukseen, että tutkittavan kaksiarvoisen tapahtuman riskitiheyksien suhde (**HR**, “hazard ratio”) vertailtavien ryhmien välillä on **vakio** missä tahansa seurannan aikapisteessä. Muita oletuksia elossaoloaikojen jakaumasta ei tehdä, joten menetelmä on tässä suhteessa puolittain parametrinen (”semi-parametric”). Coxin mallia käytetään runsaasti esimerkiksi syöpätutkimuksissa. Sen avulla voidaan tutkia lopputulokseen vaikuttavia ennustetekijöitä ja kontrolloida sekoittavia tekijöitä.

Hasardifunktio (“hazard function”) h(t; x) mittaa sitä nopeutta, jolla aineistoon kuuluvat päätyvät päätetapahtumaan. Oletetaan, että $x=(x_1, \ldots, x_p)$ ovat kovariaatteja, joiden vaikutus tarkastelun kohteena olevaan tapahtumaan kuluvaan aikaan halutaan huomioida. Coxin regressiossa h(t; z) on muotoa:

$$h(t; x) = h_0(t) \cdot \exp(\beta_1 x_1 + \ldots + \beta_p x_p),$$

Perusriski $h_0(t)$ vastaa lineaarisen regressiomallin vakiotermiä kuitenkin sillä erotuksella, että tässä mallissa $h_0(t)$ on ajan funktio eikä vakio kuten lineaarisessa

mallissa. Mikäli kaikki kovariaatit $x_1, \ldots, x_p$ ovat dikotomisia (0,1), niin $h_0(t)$ on sellaisen henkilön riski päätyä tarkastelun kohteena olevaan tilaan, jolla kaikki kovariaatit ovat nollia. Coxin mallin yhtälö voidaan kirjoittaa myös muotoon:

$$\log_e [(h(t; x)/h_0(t)] = \beta_1 x_1 + \ldots + \beta_p x_p.$$

Logistisessa mallissa yhtäsuuruusmerkin vasemmalla puolella oli logit-muunnos, mistä johtuen mallin perusteella saadut suureet $\exp(\beta_i)$:t olivat OR:iä. Tässä mallissa yhtäsuuruusmerkin vasemmalla puolella on hasrardisuhteen logaritmi, joten mallin tuottamat $\exp(\beta_i)$:t ovat hasardisuhteita (HR). Muuten kaikki kerrointen tulkinnat, testit ja residuaalien tarkistamiset ovat hyvin samankaltaisia kuin logistisen mallin yhteydessä.

Esim.

Olkoon x muuttuja, joka saa arvon 1 tai 0 riippuen siitä kuuluuko henkilö aktiivi- vai lumehoitoryhmään ja että mallissa joko ei ole muita kovariaatteja tai että ryhmät ovat muiden kovariaattien suhteen täysin samanlaisia.
Merkitään $h_A(t; x)$:llä ja $h_C(t; x)$:llä näiden ryhmien hasardifunktioita. Coxin mallin perusteella saadaan hasardisuhteeksi: **$HR = h_A(t; x) / h_C(t; x) = \exp(\beta)$.**

Perusriski **$h_0(t)$** on yleensä analyysin kannalta epäolennainen nk. **häiriöparametri**, jota ei tarvitse määritellä. Menetelmä ei edellytä siten mitään tiettyä elossaoloaikojen jakaumafunktiota, joten se on parametriton ajan suhteen, mutta se on parametrinen kovariaattien suhteen ja siitä johtuen Coxin regressiosta käytetään nimitystä **semi-parametrinen menetelmä**. Verrattuna täysin parametrisiin malleihin tästä on huomattava etu, koska olettamuksia tulee vähemmän ja siten menetelmän käytön joustavuus lisääntyy.

Verrannollisuusoletus

Oletetaan, että $S_0(t)$ on perusriskiin $h_0(t)$ liittyvä elossaolofunktio, niin kovariaatteihin $x=(x_1, \ldots, x_p)$ liittyvä elossaolofunktio on:

$$S(t; x) = S_0(t)^{\exp(\beta_1 x_1 + \ldots + \beta_p x_p)}$$

Tästä relaatiosta on hyötyä, mm. kun tarkistetaan **verrannollisuusoletuksen** paikkansapitävyyttä. Tämä oletus on Coxin-mallin toimivuuden kannalta keskeinen. Graafisesti verrannollisuusoletuksen paikkansapitävyyttä voidaan tarkistaa seuraavilla tavoilla:

Tapa 1) Mikäli mallissa on vain yksi luokiteltu muuttuja x, joka ilmaisee vertailtavia ryhmiä, niin verrannollisuusoletuksen tarkistus voidaan yksinkertaisesti tehdä Kaplan-Meier käyrien avulla. Mikäli ne menevät ristiin, niin verrannollisuusoletus ei ole voimassa. On syytä kuitenkin huomata, että pienissä aineistoissa käyriin S(t) voi sisältyä niin paljon virhettä, että käyrät voivat jossain kohtaa aika-akselia mennä jonkin verran ristiin vaikka verrannollisuusoletus olisi voimassakin.

Tapa 2) Komplementaarinen log-log-piirros ”Complementary log-log-plot”-menetelmä. Oletetaan, että $S_1(t)$ ja $S_2(t)$ ovat kahden vertailtavan ryhmän elossaolokäyrät. Jos piirretään sekä $\log_e(-\log_e(S_1(t))$ että $\log_e(-\log_e(S_2(t))$ samaan (x, y)-koordinaatistoon siten, että x-akselina on $\log_e(t)$, niin käyrien tulisi olla yhdensuuntaiset, mikäli verrannollisuusoletus on voimassa.

Tapa 3) Regressiomallin avulla. Lasketaan ensin käytetyn mallin perusteella nk. Schoenfeldin osittaiset residuaalit kaikkien malliin sisältyvien kovariaattien suhteen. Sen jälkeen tehdään lineaarinen regressioanalyysi, missä kukin residuaali on vuorollaan selitettävänä muuttujana (y) ja selittäjänä (x) on seuranta-ajan logaritmi $\log_e(t)$. Mikäli x:n kerroin on tilastollisesti merkitsevä, niin verrannollisuusoletus ei ole voimassa. Kannattaa myös tehdä sirontakuvio, missä $\log_e(t)$ on x-akselina ja residuaali y-akselina. Tästä kuviosta voi visuaalisesti päätellä muuttuuko hasardisuhteen logaritmi **log(HR)** ajan mukaan. Regressiotekniikalla voidaan tutkia oletuksen paikkansapitävyyttä myös jatkuvien muuttujien suhteen toisin kuin graafisella tekniikalla. Lisäksi saadaan kvantitatiivine tulos, P-arvo, oletuksen paikkansapitävyydestä.

Mikäli verrannollisuusoletus ei ole voimassa, niin voidaan rakentaa malli, missä käytetään ajasta riippuvia kovariaatteja. Malleista tulee silloin helposti monimutkaisia ja vaikeasti hallittavia. Yksinkertaisempi tapa on tehdä ositettu malli, missä verrannollisuusoletuksen paikkansapitävyyttä sekoittava tekijä laitetaan ositteeksi (SPSS:ssä valikkoon ”Strata”). Menettely on samankaltainen kuin ositetuissa Kaplan–Meier-analyyseissäkin.

Mallin parametrien estimointi

Coxin mallin parametrien $\beta_1, \ldots, \beta_p$ estimointi suoritetaan siten, että ensin jokaiselle seuranta-aikana tapahtuneelle tapahtumalle (esim. kuolemantapaus) lasketaan todennäköisyys, nk. **osittainen uskottavuus (”partial likelihood”)** ja niiden perusteella muodostetaan tulotermi, jonka avulla arviot kertoimille saadaan käyttäen suurimman uskottavuuden menetelmää (“maximum likelihood”) samoin kuin muissakin regressiotekniikoissa.

Oletetaan, että mallissa on vain yksi kaksiarvoinen muuttuja x, esim. kliinisessä kokeessa hoito, ja tarkastellaan aineistossa ajanhetkellä t_j tapahtuvaa kuolemantapausta. Olkoon tämän henkilön x:n arvo $\mathbf{x}_j^*$. Oletetaan lisäksi, että J on samaan kuolemanvaaraan kuuluvien henkilöiden joukko. Osittainen uskottavuus “partial likelihood” saadaan kaavalla:

$$p_j = \frac{\exp(\beta \cdot x_j^*)}{\sum_{i \in J} \exp(\beta \cdot x_i)}$$

Todennäköisyydet p_j lasketaan kaikille niille ajankohdille t_j, joissa tulee vähintään yksi tapahtuma. Olkoon näiden j:n arvojen joukko K. Seuraavaksi lasketaan tulo: $\prod_{j \in K} p_j$,

Estimoinnin helpottamiseksi tästä tulosta otetaan logaritmi, jolloin saadaan

logaritminen uskottavuus “log-likelihood”: $\mathbf{L = \sum_{j \in K} \log_e p_j}$. Kerroinarvio $\hat{\beta}$ saadaan maksimoimalla suure L β:n suhteen.

Esim.

Diffuusi histiosyyttinen lymfooma. (McKelvey et al., 1976) Asteiden 3 ja 4 vertailu. Määritellään kovariaatti x siten, että: x=0, jos kyseessä on aste=3 ja x=1, jos aste=4. Mallin parametrin β:n estimointi:

Aika (vrk)	n_3'	n_4'	d_3	d_4
4	**19**	**61**	0	**1**
6	**19**	**60**	**1**	**1**
10	18	59	0	1
...	...	...	...	...

Hasardifunktiot: $\mathbf{h(t, x = 0) = h_0(t) \cdot exp(\beta \cdot 0) = h_0(t)}$, mikäli aste=3 ja $\mathbf{h(t, x = 1) = h_0(t) \cdot exp(\beta \cdot 1) = h_0(t)\, exp(\beta)}$, mikäli aste on 4.

Ensimmäinen kuolemantapaus tulee ryhmään 4 ajanhetkellä 4 vrk. Samaan kuolemanvaaraan kuuluvien joukkoon **J** kuuluu 4 vrk:n kohdalla 19 henkilöä ryhmästä 3 ja 61 ryhmästä 4, joten

$$\mathbf{p_1 = \frac{h_0(t) \cdot exp(\beta)}{19 \cdot (h_0(t)) + 61 \cdot (h_0(t) \cdot exp(\beta))} = \frac{exp(\beta)}{19 + 61 \cdot exp(\beta)}}$$

Seuraavat kuolemantapaukset tulevat ajanhetkellä 6 vrk, yksi kummassakin ryhmässä. Jos oletetaan, että nämä kuolemantapaukset tapahtuvat samanaikaisesti ja toisistaan riippumatta, niin tämän yhdistetyn tapahtuman todennäköisyydeksi tulee todennäköisyyksien kertolaskusäännön perusteella:

$$\mathbf{p_2 = \frac{h_0(t)}{19 \cdot (h_0(t)) + 61 \cdot (h_0(t) \cdot exp(\beta))} \cdot \frac{h_0(t) \cdot exp(\beta)}{19 \cdot (h_0(t)) + 61 \cdot (h_0(t) \cdot exp(\beta))}}$$
$$\mathbf{= \frac{exp(\beta)}{(19 + 61 \cdot exp(\beta))^2}}$$

Näin jatkaen saadaan kaikki todennäköisyydet p_j missä j kuuluu joukkoon **K**, oka sisältää kaikki ne j:n arvot, joihin liittyviin ajankohtiin $\mathbf{t_j}$ ajoittuu vähintään yksi tapahtuma. Sijoittamalla nämä arvot edellä esitettyyn L:n kaavaan ja suorittamalla maksimointi saadaan tulokset:

Hasardisuhde: $\mathbf{HR_{4/3} = exp(\hat{\beta}) = 2{,}61,\ SE(\hat{\beta}) = 0{,}3856}$

Testi: H_0: β = 0, eli exp(β)=1, **z**=0,9610/0.3856=2,49, P=0,0128

95 %:n luottamusväli: $\mathbf{exp(\hat{\beta}) \pm 1{,}96 \cdot SE(\hat{\beta}) = (1{,}85, 3{,}37)}$

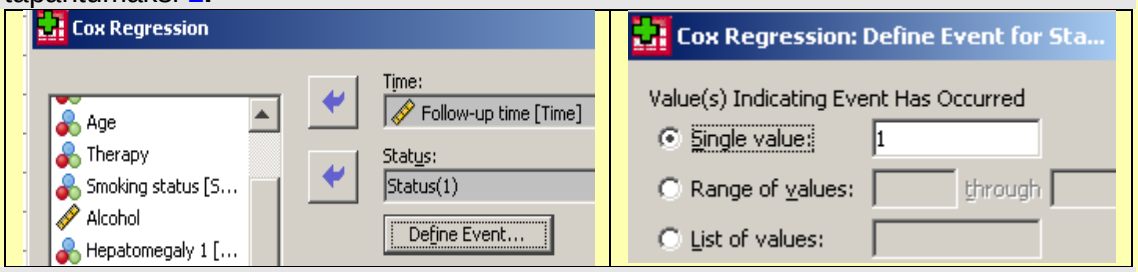
Esimerkki

Esim.

Aineisto Trial. Kolmen hoidon P, Q ja R vertailu elossaoloaikojen suhteen. P on lumehoito ja Q sekä R ovat aktiivihoitoja. Muuttuja **"Time"** on seuranta-aika. Muuttuja **"Status"** ilmoittaa potilaan tilan seurannan päättyessä: 0=ei tapahtumaa, 1=tapahtuma, 2=kadotettu seurannasta. Tapahtumana on siten arvo **1**.

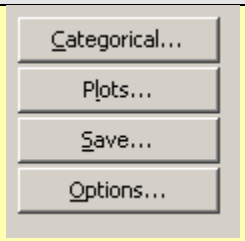
Tiedosto: http://www.mv.helsinki.fi/home/sarna/Data/Trial.sav

SPSS: Valikot: **"Analyze"►"Survival"►"Cox Regression"**. Asetetaan: **"Time"** kohtaan **"Time"**, "Status" kohtaan **"Status"** ja kohdassa **"Define Event"** määritellään tapahtumaksi **1.**



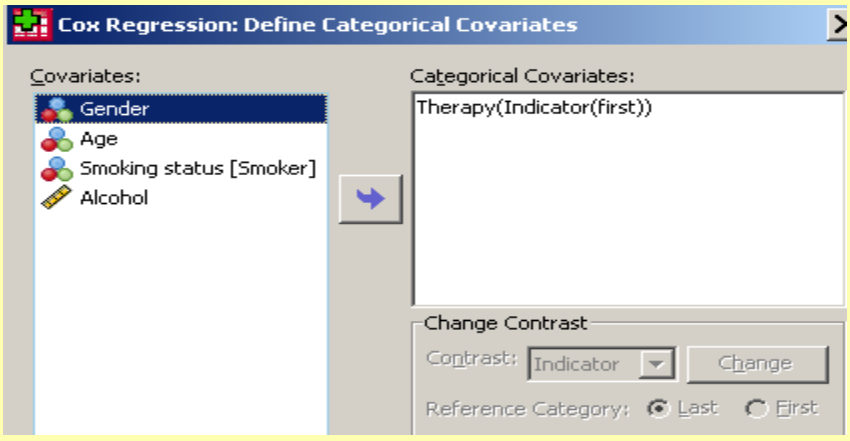
Kohtaan **"Covariates"** viedään muuttjat **"Therapy", "Gender", "Age", "Smoker"** ja **"Alcohol".**

Menetelmä "Enter" tarkoittaa kiinteää mallia, ts. kaikki kovariaatti-kohtaan viedyt muuttujat tulevat malliin.

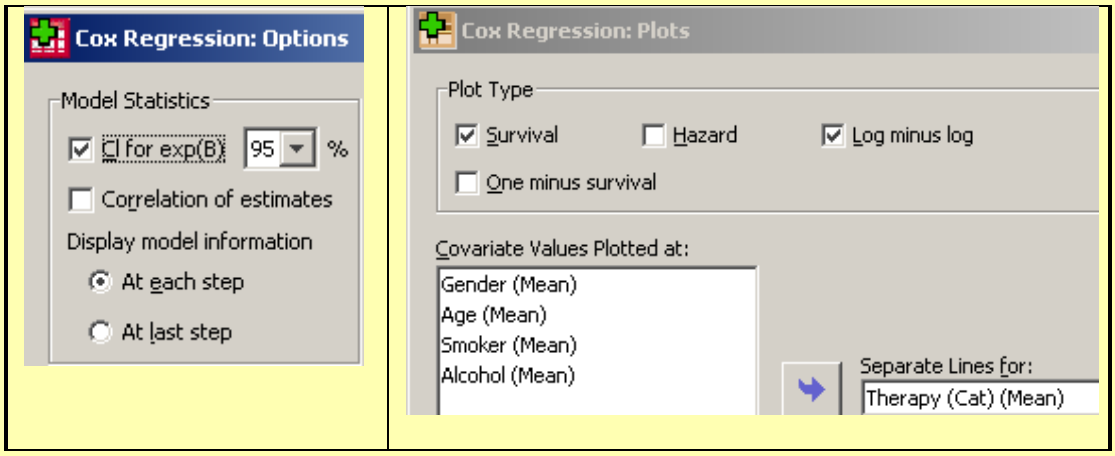


Valikossa **"Categorical"** määritellään ryhmiä Q ja R indikoivat muuttujat

Valikossa **"Categorical"** määritellään ryhmiä Q ja R indikoivat muuttujat siten, että kohdassa **"Change Contrast"** valitaan vaihtoehto **"Indicator"** ja kohdassa **"Reference category"** valitaan vaihtoehto **"First"** ja klikataan kohtaa **"Change"** Näin lumehoitoryhmästä P, jonka koodiarvo on 1, tulee referenssiryhmä. Kohdasta



”Options” klikataan kohta **”CI for exp(B)”** ja kohdasta **”Plots”** valitaan **”Survival”** ja kohtaan **”Separate Lines for”** viedään **Therapy”.**



Cox Regression

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	34	68,0%
	Censored	14	28,0%
	Total	48	96,0%
Cases dropped	Cases with missing values	1	2,0%
	Cases with negative time	0	,0%
	Censored cases before the earliest event in a stratum	1	2,0%
	Total	2	4,0%
Total		50	100,0%

a. Dependent Variable: Follow-up time

Categorical Variable Codings[b]

		Frequency	(1)	(2)
Therapy[a]	1=Group P	15	0	0
	2=Group Q	16	1	0
	3=Group R	18	0	1

a. Indicator Parameter Coding b. Category variable: Therapy

Omnibus Tests of Model Coefficients

-2 Log Likelihood
211,806

Block 1: Method = Enter:

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step		
	Chi-square	df	Sig.	Chi-square	df	Sig.
187,136	25,288	6	,000	24,670	6	,0004

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 211,806

b. Beginning Block Number 1. Method = Enter

Tulkinta:

Testisuureen -2LL muutos mallien M_0 (=Block 0) ja M_1 (=Block 1) välillä on tilastollisesti erittäin merkitsevä, **P=0.0004 (=Sig.)**, ts. Malliin sisällytetyt muuttujat Therapy, Age, Gender ja Smoker selittävät tilastollisesti merkitsevästi elossaoloaikojen eroja.

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
Therapy			11,698	2	,003			
Therapy(1)	-,854	,488	3,062	1	,080	,426	,164	1,108
Therapy(2)	-1,825	,534	11,681	1	,001	,161	,057	,459
Gender	,199	,395	,253	1	,615	1,220	,562	2,646
Age	,000	,014	,001	1	,979	1,000	,973	1,027
Smoker	,469	,499	,885	1	,347	1,599	,602	4,248
Alcohol	,425	,216	3,868	1	,049	1,529	1,001	2,335

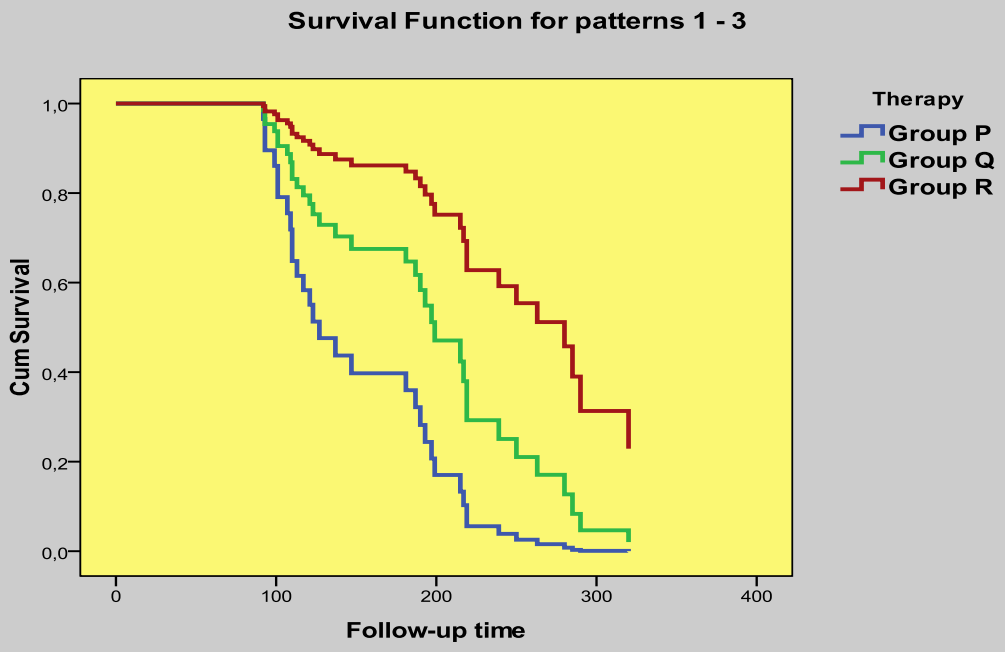
Tulkinta:

Hoitoryhmiä Q ja R indikoiviin muuttujiin Therapy(1) ja Therapy(2) liittyvät P-arvot ovat **0,080** ja **0,001**. Se merkitsee, että ryhmä R eroaa tilastollisesti merkitsevästi referenssiryhmästä P elossaoloaikojen suhteen, kun ikä, sukupuoli ja tupakointistatus ja alkoholinkulutus on huomioitu mahdollisina sekoittavina tekijöinä. Näistä tosin vain alkoholi yltää tilastollisesti merkitsevälle tasolle Waldin testissä, joten ainakaan kovin merkittävää sekoittavaa vaikutusta näillä tekijöillä ei ole. Alkoholin ja tupakan HR on suuruusluokkaa 1,5. Iällä ja sukupuolella ei tässä aineistossa ole mitään vaikutusta tuloksiin.

Adjustoidut HR:t ryhmissä Q ja R suhteessa P:hen ovat **0,426** ja **0.161** ja näistä jälkimmäisen luottamusvälin yläraja on paljon pienempi kuin 1, joten hoidolla R on erittäin selkeä suojaava vaikutus kuolemantapausten suhteen verrattuna lumehoitoon P.

Covariate Means and Pattern Values

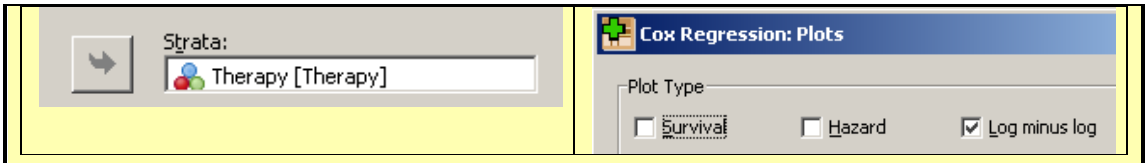
	Mean	Pattern		
		1	2	3
Therapy(1)	,333	,000	1,000	,000
Therapy(2)	,375	,000	,000	1,000
Gender	1,521	1,521	1,521	1,521
Age	50,729	50,729	50,729	50,729
Smoker	,667	,667	,667	,667
Alcohol	1,833	1,833	1,833	1,833

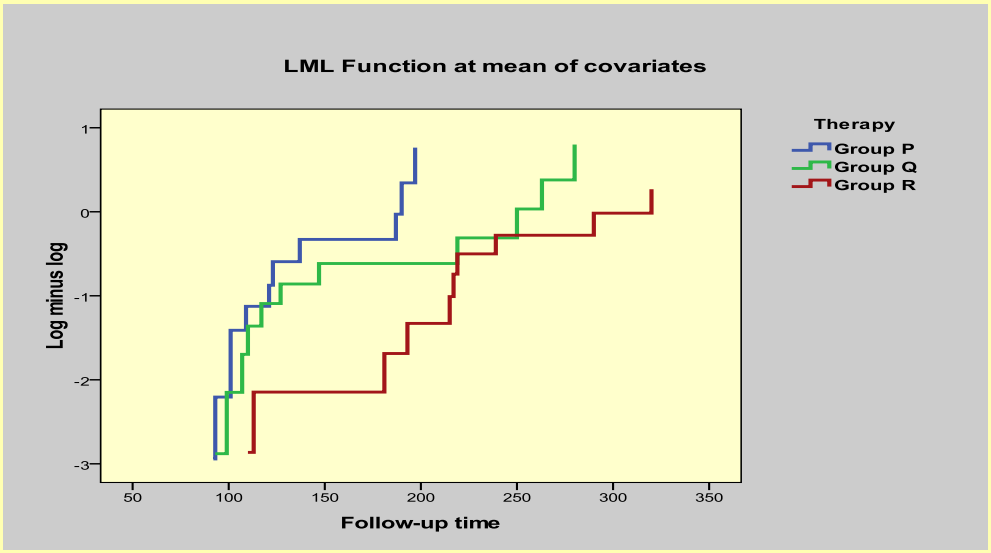


Verrannollisuusoletuksen tarkistaminen:

Onko verrannollisuusoletus voimassa muuttujan “Therapy” suhteen?

Tapa 2:
Valikossa: **“Analyze”►”Survival”►”Cox Regression”** luokiteltu muuttuja **”Therapy”** viedään kohtaan **”Strata”** ja kohdassa **”Plots”** klikataan **”Log minus log”**





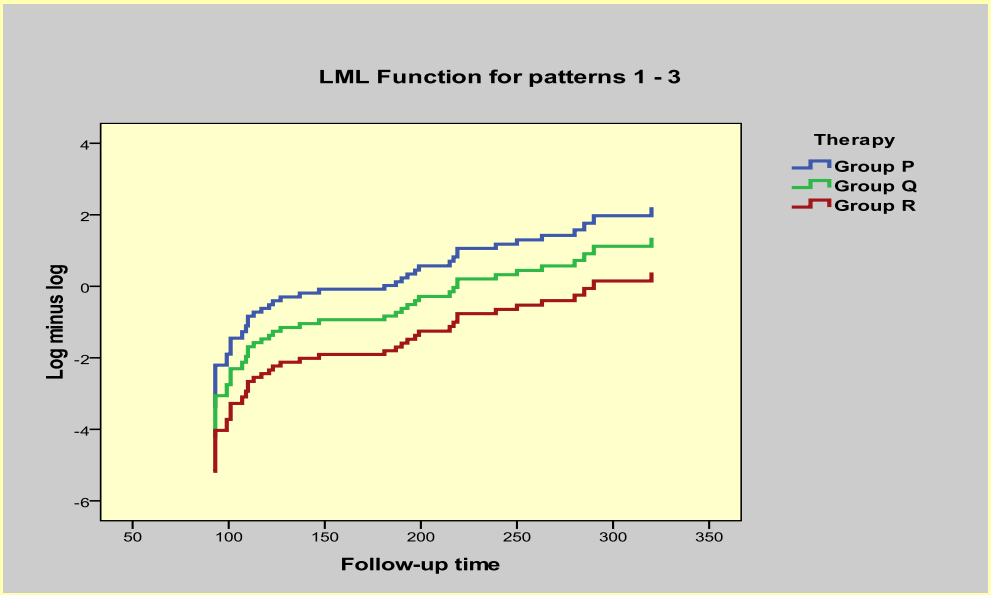
Tulkinta:

Mahdollisesti aineiston pienuudesta johtuen ryhmien Q ja R käyrät menevät lievästi päällekkäin

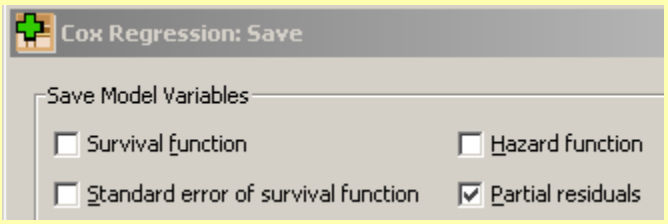
Varoitus:

Mikäli muuttuja **"Therapy"** vietäisiin kohdassa **"Plots"** lokeroon **"Separate Lines for"** kohdan **"Strata"** asemesta, niin tuloksena syntyvät käyrät eivät ole oikein. Ne ovat pakostakin lähes yhdensuuntaisia, koska ne ovat adjustoituja käyriä!

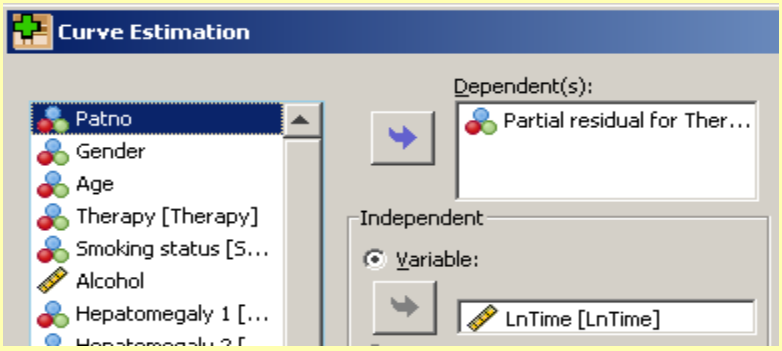
Virheelliset käyrät:



Tapa 3 Onko verrannollisuusoletus voimassa muuttujan **"Therapy"** suhteen? Lasketaan ja tallennetaan Schoenfeldin osittaiset residuaalit havaintoaineistoon ja tehdään uusi muuttuja **Ln(Time)** kohdassa **"Transform"**



SPSS: Valikot **"Analyze"►"Regression"►"Curve Estimation"**



Partial residual for Therapy(1)

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
LnTime	-,137	,213	-,113	-,645	,524
(Constant)	,695	1,081		,643	,525

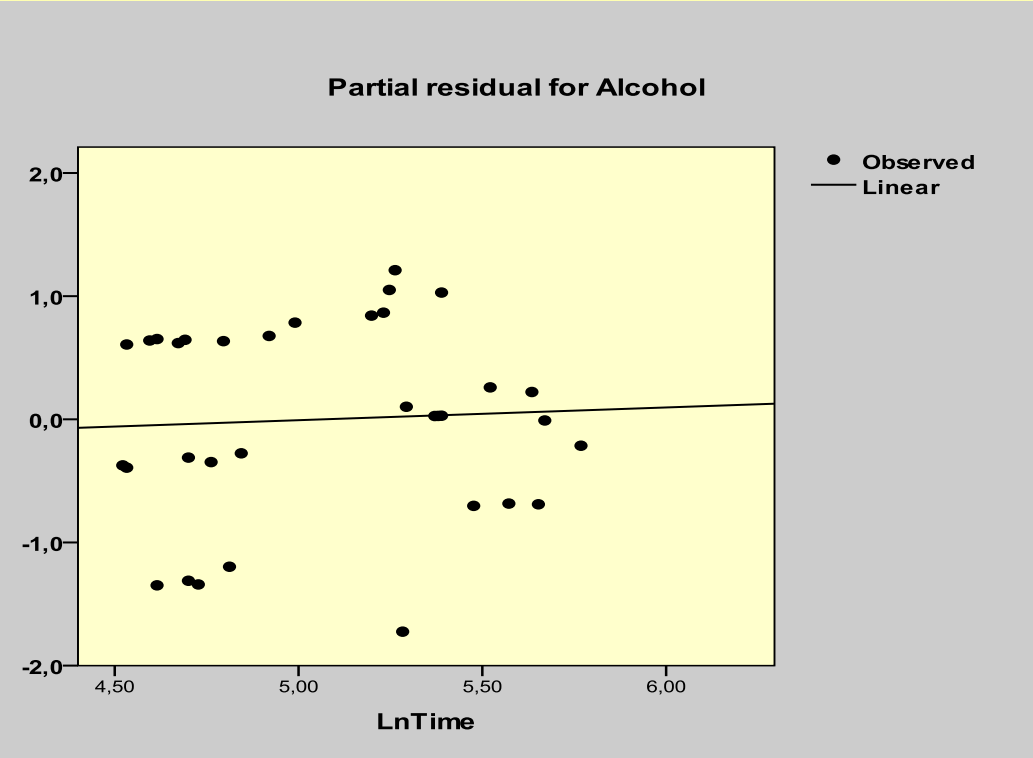
Partial residual for Therapy(2)

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
LnTime	-,057	,185	-,054	-,306	,762
(Constant)	,288	,943		,305	,762

Tulkinta:

Tässä on verrattu ryhmien Q ja R käyriä ryhmän P käyrään indikaattorimuuttujien Therapy(1) ja Therapy(2) perusteella. Waldin testien P-arvot eivät kummankaan kohdalla ole merkitseviä, joten verrannollisuusoletus pitää paikkansa muuttujan "Therapy" suhteen.

Onko verrannollisuusoletus voimassa muuttujan **"Alcohol"** suhteen?



Partial residual for Alcohol

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
LnTime	,103	,349	,052	,296	,769
(Constant)	-,524	1,775		-,295	,770

Tulkinta:

Koska muuttujan **"LnTime"** regressiokerroin ei poikkea tilastollisesti merkitsevästi nollasta (P=0,769) niin verrannollisuusoletus on voimassa myös muuttujan **"Alcohol"** suhteen.

Viitteet

Armitage P & Berry G. Statistical methods in medical research, 2nd edition. Blackwell Scientific Publications 1987, s. 423.

Chiang CL Introduction to Stochastic Processes in Biostatistics, John Wiley & Sons, Inc. New York, 1968.

Greenwood M. A report on the natural duration of cancer. Reports on Public Health and Medical Subjects 1926; 33: 1-26.

Göbel E et al. Randomised, double-blind trial of intravenous diltiazem versus glyseryl nitrate for unstable angina pectoris. Lancet 1995;346:1653-57.

Harris EK, Albert A. Survivorship Analysis for Clinical Studies. Marcel Dekker, Inc., 1991. ISBN: 0-8247-8400-6

Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York 1980.

McKelvey M, Gottlieb Ja, Wilson HE, Haut A, Talley R, Stephens R, Lane M, Gamble J, Jones SE, Grozea P, Gutterman J, Coltman C, Moon TE. Hydroxyldaunomucin (Adriamycin) combination chemotherapy in malignant lymphoma. Cancer 1976; 38:1484-93.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J and Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Br. J. Cancer 1977; 35: 1-39.

Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982 69(1):239-241.

Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986 105: 429-435.