HELSINGIN YLIOPISTO

Sosiaalitieteiden laitos

Kimmo Vehkalahti

Lineaaristen mallien sovellukset

Tämä moniste on toiminut oheismateriaalina Helsingin yliopiston matematiikan ja tilastotieteen laitoksen kurssilla **Lineaaristen mallien sovellukset**, jota pidin vuosina 2001–2008. (Vuosina 2001–2004 kurssin nimi oli Data-analyysi II.)

Syksyllä 2009 oli vuorossa pieni tauko tästä kurssista. Tällöin sen piti upeasti Jyrki Möttönen. Kurssi järjestettiin vaihteeksi Kumpulassa, ja moniste toimi ainakin kurssin oheismateriaalina, joten jätin sen sellaisenaan paikoilleen (lisälehdellä varustettuna).

Palaan näiden asioiden pariin lukuvuonna 2010–2011, jolloin kurssi järjestetään taas keskustassa, nyt aivan uuteen ajankohtaan: kevätlukukauden jälkimmäisellä periodilla.

Syksyllä 2008 julkaistun oppikirjani **Kyselytutkimuksen mittarit ja menetelmät** (Tammi) luku 5 pohjautuu eräiltä osin vahvasti tämän kurssin sisältöihin. Kannattaa tutustua!

Toivotan taas mukavia hetkiä lineaaristen mallien sovellusten parissa!

Kumpulassa, 14. syyskuuta 2010

Kimmo Vehkalahti
Kimmo.Vehkalahti@helsinki.fi

HELSINGIN YLIOPISTO

Matematiikan ja tilastotieteen laitos

Kimmo Vehkalahti

Lineaaristen mallien sovellukset

Tämä moniste on toiminut oheismateriaalina Helsingin yliopiston matematiikan ja tilastotieteen laitoksen kurssilla **Lineaaristen mallien sovellukset**, jota pidin vuosina 2001–2008. (Vuosina 2001–2004 kurssin nimi oli Data-analyysi II.)

Syksystä 2009 lähtien kurssin pitäjä vaihtuu, mutta monisteesta saattaa olla edelleen hyötyä ainakin kurssin oheismateriaalina tai muuten vain, joten jätän sen sellaisenaan paikoilleen (tällä lisälehdellä varustettuna).

Syksyllä 2008 julkaistun oppikirjani **Kyselytutkimuksen mittarit ja menetelmät** (Tammi) luku 5 pohjautuu eräiltä osin vahvasti tämän kurssin sisältöihin. Kannattaa tutustua!

Toivotan mukavia hetkiä lineaaristen mallien sovellusten parissa!

Kumpulassa, 19. elokuuta 2009

Kimmo Vehkalahti
Kimmo.Vehkalahti@helsinki.fi

HELSINGIN YLIOPISTO

Matematiikan ja tilastotieteen laitos

Kimmo Vehkalahti

Data-analyysi II

Tämä moniste toimii oheismateriaalina Helsingin yliopiston matematiikan ja tilastotieteen laitoksen Data-analyysi II -kurssilla. Se sisältää tiiviissä muodossa mm. kurssilla käsiteltävien aiheiden teoreettisia perusteluja. Luennoilla asioita havainnollistetaan erilaisten käytännön esimerkkien avulla, kysellen ja keskustellen yhdessä. Monisteen päätarkoitus onkin innostaa tutustumaan luennoilla käsiteltävien aiheiden taustoihin jo etukäteen sekä helpottaa omien muistiinpanojen tekemistä. Yksityiskohtaisemmin asioihin paneudutaan viikottaisissa, ohjatuissa harjoituksissa ja esitystilaisuuksissa.

Itsenäisen työskentelyn, esimerkiksi kurssin harjoitustyön tekemisen tueksi suosittelen kurssin kotisivulla mainittuja teoksia, etenkin *Juha Purasen* (1997) laajaa monistetta. Tuo moniste, kuten myös *Simo Puntasen* (1999) kaksiosainen regressioanalyysin kirja, ovat olleet tämän monisteen keskeiset lähteet. Tärkeimpänä innoittajana on puolestaan toiminut *Dennis Cookin* ja *Sanford Weisbergin* (1999) kirja, jonka tulin hankkineeksi matkoilla ollessani Stanfordin yliopiston kirjakaupasta kesällä 2001, kun aloin valmistautua kurssin pitämiseen ensimmäistä kertaa.

Kaavapitoinen alkuosa on pääosin ennestään tuttua asiaa ainakin tilastollisen päättelyn ja lineaaristen mallien kurssit käyneille. Mikäli alkuosa ei tunnu tutulta, kannattaa varmasti kerrata asioita etenkin lineaaristen mallien teorian osalta.

Sivuaineopiskelijat, jotka aikovat suorittaa tämän kurssin **3 ov laajuisena**, voivat sivuuttaa kaavat ja keskittyä niiden käytännön soveltamisen opetteluun.

Harjoitustehtävien tekemiseen, dokumentointiin ja esittämiseen sekä harjoitustyön laatimiseen voit käyttää mitä tahansa osaamiasi, tarkoitukseen soveltuvia ohjelmistoja. Regressio- ja varianssianalyysi ovat menetelmistä yleisimpiä, joten ne löytyvät useimmista alan ohjelmistoista. Niitä ovat mm. **SAS**, **SPSS** ja **R** sekä **Survo**. Useimmat ohjelmat on mahdollista saada myös omaan kotikoneeseen.

Opetusvälineinäni käytän pääasiassa Survoa sekä *Arc*-nimistä ohjelmaa. Survo on tekstin ja numeerisen tiedon käsittelyn käyttöympäristö, joka lienee jossain määrin tuttu monille Data-analyysi I -kurssilta. Minnesotan yliopistossa kehitetty Arc on puolestaan regressioanalyysia koskevien asioiden visualisointiin painottunut erikoisohjelma. Sen saa kopioitua itselleen verkosta.

Kurssin kotisivulta **www.helsinki.fi/~kvehkala/da2** löytyy luettelo lähdekirjallisuudesta sekä linkkejä kirjoittajien ja ohjelmistojen sivuille, tietoja kurssin tavoitteista, aikataulusta jne.

Kaikki materiaalia (ja yleensäkin tätä kurssia) koskeva palaute on tervetullutta!

Vuosaaressa, 23. joulukuuta 2003

Kimmo Vehkalahti
Kimmo.Vehkalahti@helsinki.fi

Sisällys

Moniste koostuu kolmesta osasta (I–III), jotka vastaavat luentojen pääteemoja. Niiden alaotsikoina on tässä lueteltu kunkin luennon aiheet. Vastaavaa jaottelua ei esiinny varsinaisessa tekstissä, vaan kaikki pääteemaan liittyvät aiheet on vain koottu yhteen ja otsikoitu tarpeen mukaan.

I Regressiomallin rakentaminen

- Lineaarinen tilastollinen malli
- Regressioanalyysi pääpiirteittäin
- Mallintamisen valintatilanteet

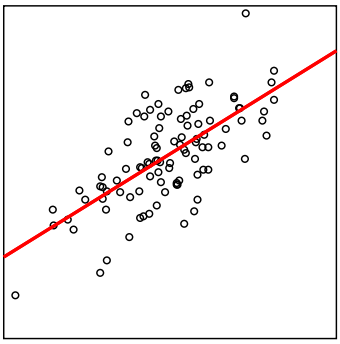
II Diagnostiikka ja muunnokset

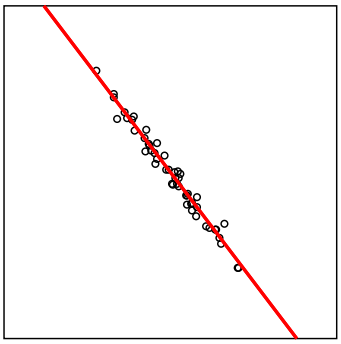
- Mallin yleinen diagnosointi
- Riippuvuuksien linearisointi
- Vaikutusvaltaiset havainnot

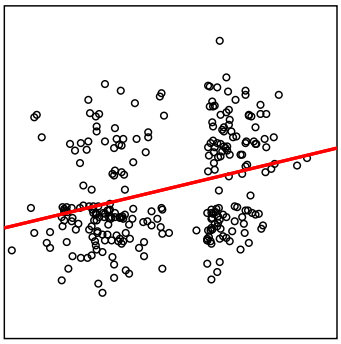
III Luokittelevat muuttujat

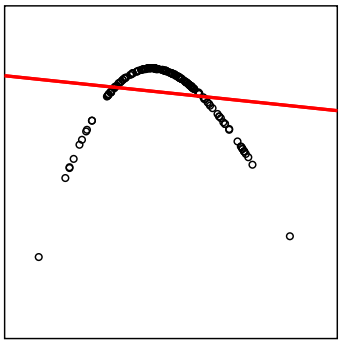
- Kategoriset selittäjät eli faktorit
- Yhdysvaikutukset eli interaktiot
- Varianssianalyysi pääpiirteittäin

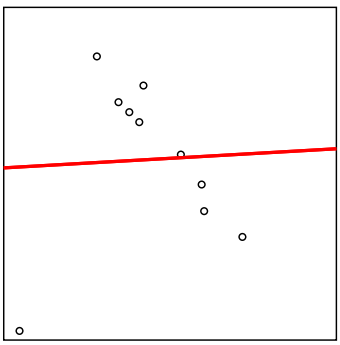
Kuvissa 1–8 on sovitettu regressiomalli simuloituihin aineistoihin:

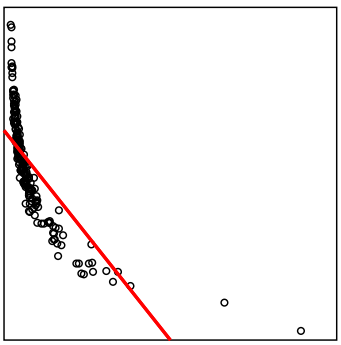

Kuva 1

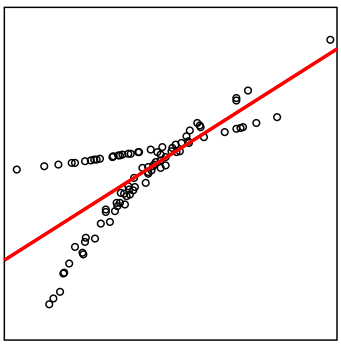

Kuva 2

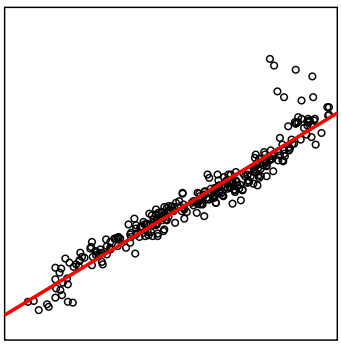

Kuva 3


Kuva 4


Kuva 5


Kuva 6


Kuva 7


Kuva 8

Osassa kuvista mallin sovite, siis regressiosuora, ei näytä kovin hyvin kuvaavan aineistoa. Mitä syitä tähän havaitset kuvia tarkastelemalla? Pohdi asiaa myös lineaarisen mallin oletuksia (ks. osa I) silmälläpitäen. Entä mitä sanoisit ao. muuttujien korrelaatioista?

I Regressiomallin rakentaminen

Tällä kurssilla tarkastellaan lineaarista tilastollista mallia

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

jossa $\mathbf{y}$ on selitettävä muuttuja ($n \times 1$ -vektori), $\mathbf{X}$ on selittävien muuttujien $n \times p$ -matriisi, $\boldsymbol{\beta}$ on regressiokertoimien kiinteä $p \times 1$ -vektori ja $\boldsymbol{\varepsilon}$ on tuntematon, satunnainen mallivirhe ($n \times 1$ -vektori). Dimensio n viittaa aineiston havaintojen ja p mallin parametrien lukumäärään. Malliin kuuluu vakiotermi $\mathbf{1} = (1, 1, ..., 1)$ sekä k varsinaista selittäjää, siis $p = k + 1$. Matriisi $\mathbf{X}$ koostuu kiinteistä (ei-satunnaisista) luvuista. Sitä kutsutaan mallimatriisiksi ja sen oletetaan olevan täysiasteinen: $\mathrm{r}(\mathbf{X}) = p$.

Mallivirheestä $\boldsymbol{\varepsilon}$ oletetaan yleensä, että

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}\ , \tag{1.2a}$$

$$\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} \text{ ja} \tag{1.2b}$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})\ . \tag{1.2c}$$

Kaksi ensimmäistä sisältävät ns. Gaussin ja Markovin ehdot eli mallivirheiden odotusarvo on nolla (mallin identifiointi oikea), niiden varianssi on vakio (homoskedastisuus) ja ne ovat keskenään korreloimattomia. Mm. hypoteesien testauksessa oletetaan lisäksi, että mallivirheet ovat normaalisti jakautuneita (1.2c).

Selitettävä muuttuja $\mathbf{y}$ oletetaan tunnetuksi. Kyseessä ovat siis vastaavan satunnaismuuttujan realisaatiot eli havaitut arvot. Näin määritelty malli tarkoittaa, että

$$\mathrm{E}(\mathbf{y}) = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + ... + \beta_k\mathbf{x}_k = \mathbf{X}\boldsymbol{\beta} \in \mathrm{C}(\mathbf{X}), \tag{1.3}$$

eli $\mathbf{y}$:n odotusarvo on jokin $\mathbf{X}$:n sarakkeiden lineaarikombinaatio.

Jakaumaoletus (1.2c) voidaan kirjoittaa myös $\mathbf{y}$:n avulla muodossa

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})\ . \tag{1.4}$$

Koska $\mathbf{X}\boldsymbol{\beta}$ on vakio, niin $\mathrm{var}(y) = \sigma^2 = \mathrm{var}(\varepsilon)$. Varianssi σ^2 on käytännössä tuntematon, ja se on estimoitava otoksesta. On tärkeää huomata, että y:n varianssi ei riipu $\mathbf{X}$:n arvoista.

Pienimmän neliösumman menetelmä

Regressiokerroinvektori $\boldsymbol{\beta}$ oletetaan siis kiinteäksi, mutta sen arvoja ei tunneta vaan ne on estimoitava havaitun otoksen perusteella. Huomaa että $\mathbf{y} \in \Re^n$ mutta $\boldsymbol{\beta} \in \Re^p$, ja yleensä $n >> p$. On siis löydettävä sellainen $\boldsymbol{\beta}$, että $\mathbf{X}\boldsymbol{\beta}$ olisi mahdollisimman lähellä $\mathbf{y}$:tä. Tarkemmin sanottuna on määritettävä se sarakeavaruuden $\mathrm{C}(\mathbf{X})$ vektori eli se mallimatriisin $\mathbf{X}$ sarakkeiden lineaarikombinaatio $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, joka on lähimpänä vektoria $\mathbf{y}$.

Graafisten tarkastelujen (ks. esim. Faraway 2002, 17–18; Patovaara 1991, 87–89; Puntanen 1999a, 9; Puranen 1997, 7) perusteella optimiratkaisu saavutetaan, kun $\mathbf{X}\hat{\boldsymbol{\beta}}$ on $\mathbf{y}$:n ortogonaaliprojektio $\mathrm{C}(\mathbf{X})$:lle. Tätä merkitään

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}\ , \tag{1.5}$$

jossa $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ on projektiomatriisi (ortogonaaliprojektori). Sille pätee

$$\mathbf{H}^2 = \mathbf{H}' = \mathbf{H} \text{ sekä } \mathrm{r}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}) = \mathrm{r}(\mathbf{X})\ . \tag{1.6}$$

Estimointiongelman ratkaisuksi saadaan näin ollen

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\ , \tag{1.7}$$

jota kutsutaan pienimmän neliösumman (PNS) estimaatiksi, sillä se minimoi mallivirheiden neliösumman

$$\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\ . \tag{1.8}$$

Minimiarvoon johtavia erotuksia

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{1.9}$$

kutsutaan residuaaleiksi (jäännöksiksi) ja minimiarvoa $\mathbf{e}'\mathbf{e}$ jäännösneliösummaksi.

Ratkaisu saadaan myös analyyttisesti kirjoittamalla lauseke (1.8) auki, derivoimalla se vektorin $\boldsymbol{\beta}$ suhteen ja merkitsemällä derivaatat nolliksi. Näin päädytään ns. normaaliyhtälöihin

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} , \tag{1.10}$$

josta ratkaisu seuraa kertomalla vasemmalta $\mathbf{X}'\mathbf{X}$:n käänteismatriisilla.

Edellä $\mathbf{X}$ oletettiin täysiasteiseksi, joten kyseinen käänteismatriisi on olemassa. Mikäli $\mathbf{X}$ kuitenkin olisi vajaa-asteinen, ratkaisu saataisiin yleistettyjen käänteismatriisien avulla, mutta se ei olisi yksikäsitteinen. Käytännössä tilastollisten ohjelmien estimointialgoritmit eivät lainkaan ratkaise normaaliyhtälöitä. Niissä sovelletaan numeerisesti turvallisempia ja tehokkaampia menetelmiä, jotka perustuvat usein $\mathbf{X}$:n sarakkeiden ortogonalisointiin, ks. esim. Patovaara (1991, 166-171).

Mikäli oletus (1.2a) pätee, niin PNS-estimaattori $\hat{\boldsymbol{\beta}}$ on $\boldsymbol{\beta}$:n harhaton estimaattori:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} . \tag{1.11}$$

Satunnaisvektorin $\hat{\boldsymbol{\beta}}$ kovarianssimatriisi on

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \mathrm{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} , \tag{1.12}$$

mikä osaltaan kuvastaa $\mathbf{X}'\mathbf{X}$:n käänteismatriisin merkitystä regressioanalyysissa.

Sovite ja residuaalit

Mallin tarkastelun kannalta tärkeässä asemassa ovat sovite ja residuaalit. Sovite tarkoittaa siis mallin antamia arvoja, kun selitettävä muuttuja $\mathbf{y}$ on projisoitu selittäjien aliavaruuteen, ja sitä merkitään

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y} . \tag{1.13}$$

Vastaavasti residuaalit ovat tällöin jäljelle jäävät erotukset, joita merkitään

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} . \tag{1.14}$$

Kun mallissa on vakio, niin sovitteen ja selitettävän muuttujan summat (ja siten myös keskiarvot) ovat samat:

$$\mathbf{1} \in \mathrm{C}(\mathbf{X}) \;\Rightarrow\; \mathbf{H}\mathbf{1} = \mathbf{1} \;\Rightarrow\; \hat{\mathbf{y}}'\mathbf{1} = \mathbf{y}'\mathbf{H}\mathbf{1} = \mathbf{y}'\mathbf{1} . \tag{1.15}$$

Kun mallissa on vakio, niin residuaalien summa on nolla:

$$\mathbf{1} \in \mathrm{C}(\mathbf{X}) \;\Rightarrow\; \mathbf{e}'\mathbf{1} = (\mathbf{y} - \hat{\mathbf{y}})'\mathbf{1} = \mathbf{y}'\mathbf{1} - \hat{\mathbf{y}}'\mathbf{1} = 0 . \tag{1.16}$$

Tämän geometrinen tulkinta on, että residuaalivektori $\mathbf{e}$ on kohtisuorassa vektoria $\mathbf{1}$ (itse asiassa jokaista mallimatriisin $\mathbf{X}$ saraketta) vastaan. Myös residuaalien odotusarvo on nolla:

$$\mathrm{E}(\mathbf{e}) = \mathrm{E}(\mathbf{y} - \mathbf{H}\mathbf{y}) = \mathrm{E}(\mathbf{y}) - \mathrm{E}(\mathbf{H}\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{0} , \tag{1.17}$$

mutta kovarianssimatriisista

$$\mathrm{cov}(\mathbf{e}) = \mathrm{cov}[(\mathbf{I} - \mathbf{H})\mathbf{y}] = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{1.18}$$

näkyy selvästi, miten residuaalit eroavat mallivirheistä (1.2b).

Sovitteen $\hat{\mathbf{y}}$ odotusarvo on

$$\mathrm{E}(\hat{\mathbf{y}}) = \mathrm{E}(\mathbf{H}\mathbf{y}) = \mathbf{H}\mathrm{E}(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} , \tag{1.19}$$

eli $\hat{\mathbf{y}}$ on $\mathbf{X}\boldsymbol{\beta}$:n harhaton estimaattori. Sen kovarianssimatriisi on puolestaan

$$\mathrm{cov}(\hat{\mathbf{y}}) = \mathrm{cov}(\mathbf{H}\mathbf{y}) = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H}' = \sigma^2\mathbf{H} . \tag{1.20}$$

Useissa edellä olevissa kaavoissa esiintyy matriisi $\mathbf{H}$ *(hat matrix)*, joka siis projisoi selitettävän muuttujan $\mathbf{y}$ ortogonaalisesti mallimatriisin $\mathbf{X}$ sarakeavaruuteen $\mathrm{C}(\mathbf{X})$. Erityisen hyödyllinen $\mathbf{H}$ on lineaaristen mallien teoriatarkasteluissa, mutta käytännössä kannattaa muistaa että kyseessä on $n \times n$ -matriisi. Myöhemmin käsiteltävän regressiodiagnostiikan yhteydessä nähdään, millaisia asioita matriisin $\mathbf{H}$ alkioista voidaan käytännössä päätellä.

Neliösummat

Malliin (1.1) liittyvät neliösummat ja varianssit kootaan usein taulukon (1.1) muotoon:

Taulukko (1.1). Regressiomallin varianssitaulu.

SS	df	MS	F
$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$dfR = k$	$MSR = SSR / k$	$F = MSR / MSE$
$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$dfE = n-k-1$	$MSE = SSE / (n-k-1) = s^2$	
$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$dfT = n-1$	$MST = SST / (n-1) = var(y)$	

Selitykset: SS = Sum of Squares (neliösumma); df = degrees of freedom (vapausasteet); MS = Mean Square (varianssi); R = Regression (systemaattinen osa); E = Error (satunnainen osa); T = Total (koko aineisto)

Matriisimuodossa taulukon (1.1) neliösummat voidaan lausua muodoissa

$$SSR = ||\hat{\mathbf{y}} - \bar{\mathbf{y}}||^2 = ||(\mathbf{H} - \mathbf{J})\mathbf{y}||^2 = \mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y} , \quad (1.21)$$

$$SSE = ||\mathbf{y} - \hat{\mathbf{y}}||^2 = ||(\mathbf{I} - \mathbf{H})\mathbf{y}||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \text{ ja} \quad (1.22)$$

$$SST = ||\mathbf{y} - \bar{\mathbf{y}}||^2 = ||(\mathbf{I} - \mathbf{J})\mathbf{y}||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} , \quad (1.23)$$

joissa $\mathbf{J} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$ on ortogonaaliprojektori C(**1**):lle.

Kun $\mathbf{1} \in C(\mathbf{X})$, niin SSR + SSE = SST. Mallin selitysasteeksi R^2 saadaan tällöin

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} . \quad (1.24)$$

Merkintä R^2 tulee siitä, että kyseessä on myös selitettävän muuttujan ja sovitteen välisen korrelaatiokertoimen, ns. yhteiskorrelaatiokertoimen neliö.

Taulukon (1.1) yleis-F-testin *(overall-F test)* avulla testataan nollahypoteesi

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0 , \quad (1.25)$$

jonka vastahypoteesi on

$$H_1: \text{ainakin jokin } \beta_i \neq 0,\ i = 1, 2, ..., k . \quad (1.26)$$

Hypoteesit eivät siis koske vakiotermiä. Jos H_0 pätee, niin

$$F(\beta_1, \beta_2, ..., \beta_k) \sim F_{k,n-k-1} . \quad (1.27)$$

Taulukossa (1.1) esiintyy lisäksi merkintä s^2, joka tarkoittaa jäännösvarianssin σ^2 harhatonta estimaattia. Siihen liittyvät vapausasteet nähdään myös laskemalla jäännösneliösumman (SSE) odotusarvo

$$E(SSE) = E[\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}] = tr[(\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}] + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \sigma^2(n-k-1) . \quad (1.28)$$

Regressiokertoimet

Mallin yleistuloksia kuvaavien neliösummien, yleis-F-testin ja selitysasteen ohella tärkeitä ovat tietenkin estimoidut regressiokertoimet. Periaatteessa regressiokerroin ilmaisee, kuinka paljon selitettävän muuttujan arvo muuttuu, kun vastaava selittäjä muuttuu yhden yksikön verran, olettaen että mallin muut selittäjät pidetään ennallaan. Tätä yksinkertaista tulkintaa ei voida kuitenkaan sellaisenaan soveltaa, mikäli selittäjien välillä on sidoksia.

Ohjelmien tulostuksissa esiintyvät regressiokertoimien lisäksi niiden keskivirheet eli tarkemmin sanottuna regressiokertoimien estimaattoreiden hajontojen estimaatit. Ne ovat kaavassa (1.12) esiintyvän matriisin lävistäjäalkioita ja kuvaavat regressiokertoimien tarkkuutta. Yksittäisen kertoimen tilastollisen merkitsevyyden eli hypoteesin

$$H_0: \beta_i = 0 \qquad (1.29)$$

testaaminen perustuu t-testisuureeseen $t(\beta_i)$, joka saadaan jakamalla kerroin keskivirheellään. Jos H_0 pätee, niin

$$t(\beta_i) \sim t_{n-k-1} \, . \qquad (1.30)$$

Suuremmilla vapausasteilla voidaan käyttää standardoitua normaalijakaumaa, jonka perusteella saadaan hyödyllinen muistisääntö: merkitsevän selittäjän kertoimen t-arvo on itseisarvoltaan vähintään kakkosen luokkaa.

Merkitsevyyttä voidaan tarkastella myös luottamusvälien avulla. Välin päätepisteet saadaan lisäämällä ja vähentämällä kertoimesta sen keskivirhe kerrottuna t-jakauman vastaavalla kriittisellä arvolla vapausasteilla $n–k–1$. Jos nolla sisältyy luottamusväliin, katsotaan että kerroin ei poikkea merkitsevästi nollasta eli H_0 (1.29) jää voimaan.

Koska F- ja t-jakaumien välillä on yhteys

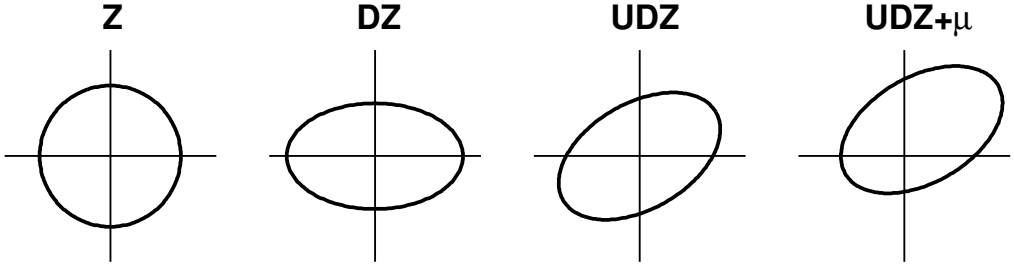
$$F_{1,df} = t^2_{df} \, , \qquad (1.31)$$

niin hypoteesi (1.29) voitaisiin testata F-testilläkin. Tällaista F-testiä kutsutaan toisinaan nimellä osittais-F-testi *(partial-F test)* erotuksena aiemmin mainitusta yleis-F-testistä.

Multinormaalijakauma

Edellä todettiin, että mm. hypoteesien testauksessa mallivirheet oletetaan normaalisti jakautuneiksi (1.2c). Oletus tarkoittaa täsmällisemmin ilmaistuna, että mallivirhe $\boldsymbol{\varepsilon}$ noudattaa n-ulotteista multinormaalijakaumaa odotusarvovektorina $\mathbf{0}$ ja kovarianssimatriisina $\sigma^2\mathbf{I}$.

Yleensä multinormaalijakauma määritellään esittämällä suoraan sen tiheysfunktio. Opettavaisempi ja samalla käsitteellisesti yksinkertaisempi tapa on määritellä se konstruktiivisesti, standardoitujen ja riippumattomien, normaalisti jakautuneiden satunnaismuuttujien lineaarikombinaatioiden avulla, käyttäen hyväksi muodostettavan kerroinmatriisin singulaariarvohajotelmaa (Mustonen 1995, 15–20):



Z : riippumattomat N(0,1)-muuttujat
DZ : venytyksiä ja kutistuksia muuttujittain
UDZ : koordinaatiston kierto
UDZ+μ : keskipisteen siirto pois origosta

Siis $\mathbf{X} = \mathbf{CZ} + \boldsymbol{\mu} \rightarrow \boldsymbol{\Sigma} = \mathbf{CC}' \rightarrow \mathbf{C} = \mathbf{UDV}' \rightarrow \mathbf{X} = \mathbf{UDV}'\mathbf{Z} + \boldsymbol{\mu}$,

mutta koska ortogonaalinen muunnos ei vaikuta N(0,1)-muuttujiin, $\mathbf{X} = \mathbf{UDZ} + \boldsymbol{\mu}$.

Keskeisiä multinormaalijakaumaan liittyviä ominaisuuksia:

- Multinormaalisuus säilyy muuttujien lineaarikuvauksissa.
- Regressiofunktiot eli ehdolliset odotusarvot $E(\mathbf{y}|\mathbf{X}=\mathbf{x})$ ovat lineaarisia.
- Kaikki riippuvuudet ovat lineaarisia, mittana korrelaatio.
- Korreloimattomuus on sama asia kuin riippumattomuus.

Jakaumaoletuksesta (1.2c) tai (1.4) seuraa täten kaavojen (1.17)–(1.20) perusteella, että

$$\hat{\mathbf{y}} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}) \text{ ja} \tag{1.32}$$

$$\mathbf{e} \sim N[\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})] \, . \tag{1.33}$$

Lineaarisuusoletus

Mallissa (1.1) *lineaarisuus* tarkoittaa, että

- Mallin systemaattinen osa $\mathbf{X}\boldsymbol{\beta}$ on parametrien $\boldsymbol{\beta}$ lineaarinen funktio.
- Mallivirhe $\boldsymbol{\varepsilon}$ lisätään systemaattiseen osaan additiivisesti.

Esimerkiksi polynomiregressiomalli

$$y = \beta_0 + \beta_1 w + \beta_2 w^2 + \varepsilon \tag{1.34}$$

on kahden selittäjän ja vakiotermin lineaarinen malli, sillä se voidaan kirjoittaa muotoon

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \, , \tag{1.35}$$

jossa $x_1 = w$ ja $x_2 = w^2$ (Puntanen 1999a, 174).

Seuraavat mallit ovat sen sijaan epälineaarisia (mts. 175):

$$y = \beta_0 x^{\beta_1} + \varepsilon \tag{1.36}$$

$$y = \beta_0 e^{-\beta_1 x} + \varepsilon \tag{1.37}$$

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} + \varepsilon \tag{1.38}$$

$$y = \beta_0 + \beta_1 e^{-\beta_2 x} + \varepsilon \tag{1.39}$$

$$y = \beta_0 x^{\beta_1} \varepsilon \tag{1.40}$$

Malli (1.36) voitaisiin linearisoida "unohtamalla" virhetermi ja logaritmoimalla puolittain:

$$\log(y) = \log(\beta_0) + \beta_1 \log(x) \, , \tag{1.41}$$

mutta vaikka tähän lisättäisiin virhetermi, ei saataisi täysin oikeaa mallia, koska alkuperäisen mallin (1.36) virhetermi oli additiivinen. Sen sijaan mallin (1.40) logaritmointi

$$\log(y) = \log(\beta_0) + \beta_1 \log(x) + \log(\varepsilon) \tag{1.42}$$

onnistuu paremmin, kunhan $\log(\varepsilon)$ toteuttaa virhetermistä tehtävät oletukset.

Mallin (1.38) systemaattista osaa kutsutaan logistiseksi funktioksi. Se voidaan linearisoida logit-muunnoksella

$$\log\left(\frac{y}{1-y}\right) = \log(e^{\beta_0 + \beta_1 x}) = \beta_0 + \beta_1 x \, . \tag{1.43}$$

Jos tähän lisättäisiin virhetermi, saataisiin ns. logistinen regressiomalli. Se soveltuu tilanteisiin, joissa selitettävä muuttuja on dikotominen. Logistinen regressiomalli on esimerkki yleistetyistä lineaarisista malleista (*Generalized Linear Models*). Ne eivät varsinaisesti kuulu tämän kurssin aihepiiriin.

Kaikkia epälineaarisia malleja ei voida linearisoida. Epälineaaristen mallien estimointi on oma "taiteenlajinsa", joka ei myöskään kuulu tämän kurssin aihepiiriin. Sen sijaan myöhemmin tarkastellaan riippuvuuksien linearisointia erilaisten muunnosten avulla.

Selittäjien erilaiset termit

Tavallisessa lineaarisessa mallissa (1.1) selitettävä muuttuja oletetaan normaalijakautuneeksi, joten sen mittaustason täytyy olla vähintään intervalliasteikko. Sen sijaan selittävät muuttujat voivat olla minkä tyyppisiä hyvänsä. Tietyillä keinoilla (joihin palataan tarkemmin myöhemmin) voidaan luontevasti käyttää jopa nominaaliasteikollisia selittäjiä. Tässä vaiheessa on hyvä hahmottaa yleisesti, että mallissa voi esiintyä useita erilaisia selittäjiin liittyviä termejä:

1˚ Vakio (useimmiten läsnä, muttei merkitä eksplisiittisesti näkyviin)

2˚ Varsinaiset selittäjät (esim. x_1)

3˚ Selittäjien potenssit (esim. x_1^2)

4˚ Selittäjien muunnokset (esim. $\log(x_1)$)

5˚ Kategoriset selittäjät (osoitinmuuttujat, dikotomiset ja useampiluokkaiset faktorit)

6˚ Interaktiot (kahden tai useamman termin tuloja, esim. x_1x_2)

Hierarkisten mallien vertailu

Hierarkisuus tarkoittaa, että kahdesta tarkasteltavasta mallista toinen sisältyy toiseen. Usein puhutaan myös sisäkkäisistä *(nested)* malleista. Tällöin voidaan jonkin selittäjän tai selittäjäryhmän tarpeellisuutta arvioida F-testillä. Esimerkiksi hypoteesien

$$H_0: \quad y = \beta_0 + \beta_1 x_1 + \varepsilon \quad \text{ja} \tag{1.44}$$

$$H_1: \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\,, \tag{1.45}$$

avulla voidaan testata, onko selittäjä x_2 tarpeellinen, eli poikkeaako β_2 merkitsevästi nollasta. Testisuureeksi saadaan

$$F = \frac{(\mathrm{SSE}_0 - \mathrm{SSE}_1) \,/\, (df_0 - df_1)}{\mathrm{SSE}_1 \,/\, df_1}\,, \tag{1.46}$$

jossa jäännösneliösummat ja vapausasteet viittaavat hypoteesien H_0 ja H_1 mukaisiin malleihin. Jos H_0 pätee, niin

$$F \sim F_{df_0 - df_1,\, df_1}\,. \tag{1.47}$$

Jos siis mallien jäännösneliösummien suhteellinen ero on suuri (F tilastollisesti merkitsevä), niin H_0 hylätään ja valitaan H_1:n mukainen malli. Toisin sanoen x_2 on tarpeellinen selittäjä mallissa. Jos taas ero on pieni, voidaan tieteellisen säästäväisyysperiaatteen mukaisesti katsoa, että H_0:n mukainen malli riittää.

Yksittäisen selittäjän kohdalla sama asia nähdään vastaavan regressiokertoimen t-testistä (1.29), mutta F-testi mahdollistaa myös selittäjäryhmien ja yleensäkin erilaisten lineaaristen hypoteesien testaamisen, kunhan vain mallit ovat hierarkisia.

Kysymys: Mitä seuraavista malleista voi vertailla toisiinsa F-testillä ja mitä ei?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon\,, \tag{1.48a}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\,, \tag{1.48b}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon\,, \tag{1.48c}$$

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon\,, \tag{1.48d}$$

$$y = \beta_0 + \beta_1 x_1 + \varepsilon\,, \tag{1.48e}$$

$$y = \beta_0 + \beta_2 x_2 + \varepsilon\,, \tag{1.48f}$$

$$y = \beta_0 + \beta_3 x_3 + \varepsilon\,, \tag{1.48g}$$

$$y = \beta_0 + \varepsilon\,, \tag{1.48h}$$

Selittäjien valitseminen malliin

Yleisenä pyrkimyksenä on löytää sellainen malli, jossa on mahdollisimman vähän selittäjiä ja joka silti kuvaa tai selittää tutkittavan ilmiön riittävän tarkasti. Ilmiön kannalta keskeisten tekijöiden on oltava joka tapauksessa mukana.

Potentiaalisten selittäjien joukkoa on voitava rajata mm. aiemman tietämyksen perusteella. Myös ilmiön ja aineiston tuntemus auttavat olennaisesti järkevien selittäjien valinnassa. Mikäli selittäjäjoukkoa ei kyetä rajaamaan etukäteen, ollaan varsin hataralla pohjalla.

Usein osa selittäjistä on substanssialan teorian kannalta välttämättömiä riippumatta siitä ovatko ne tilastollisesti merkitseviä. Onkin pidettävä mielessä ero todellisuudessa merkittävän *(remarkable, important)* ja tilastollisesti merkitsevän *(significant)* välillä. Jälkimmäinen ei yksin takaa mitään, sillä suuremmilla havaintomäärillä kaikki on helposti "merkitsevää" vaikkei todellisesta merkittävästä vaikutuksesta, erosta tms. olisi tietoakaan.

Mikäli tavoitteena on vain ennustaa jotakin monimutkaista ilmiötä mahdollisimman tarkasti, voidaan regressiomallin muuttujien valinnassa soveltaa aivan päinvastaistakin periaatetta kuin edellä kuvattu. Esimerkiksi sopii nykyaikainen sään ennustaminen: säämallit voivat sisältää tuhansia muuttujia ja niiden muunnoksia. Vastaavasti tarvitaan tietenkin runsaasti havaintoja ja normaalia järeämpää laskentakapasiteettia.

Jatkossa kuvattavat automaattiset mallinvalintamenetelmät ovat kenties hyödyllisimmillään laajojen ennustemallien yhteydessä, mutta tällä kurssilla pääpaino on ehdottomasti tieteellisessä mallin rakentamisessa, jossa jokainen mukaan valittu selittäjä pitää pystyä perustelemaan, eikä ennustaminen ole läheskään aina päätarkoitus. Tärkeämpää on ilmiön kuvaaminen tai selittäminen ja riippuvuuksien mallittaminen. Tietenkin myös ennustemalleja voi laatia näiden periaatteiden mukaisesti (esimerkiksi kansantaloudelliset ennusteet).

Automaattiset mallinvalintamenetelmät

Tietokoneiden yleistymisestä (1960-luvulta) lähtien muuttujien valintaa (kuten eräitä muitakin tutkijan työtehtäviä) on koetettu helpottaa erilaisilla automaattisilla menetelmillä. Tutkimustyö on kuitenkin luonteeltaan käsityötä, jossa jokainen työvaihe on kyettävä ajattelemaan ja perustelemaan. Niinpä tutkijat ovat yleisesti jo pitkään olleet varsin kriittisiä automaattisia mallinvalintamenetelmiä kohtaan. Osaavissa käsissä niistä saattaa olla aidosti hyötyäkin, mutta kritiikitön ja asiantuntematon käyttö osoittaa vain ettei mallittaja ole tehnyt kotitehtäviään tai ei ymmärrä tutkimaansa ilmiötä. *Juha Puranen* karrikoikin osuvasti:

"Kun dataa tarpeeksi rääkkää, niin kyllä se lopulta jotakin tunnustaa. Toisin sanoen kun aineistoon sovittaa tarpeeksi erilaisia malleja, niin kyllä niistä jokin kertoo selitettävästä muuttujasta hieman enemmän kuin muuttujan keskiarvo." (Puranen 1997, 231).

Automaattisista mallinvalintamenetelmistä yleisimpiä ovat seuraavat:

1. Täydellinen haku *(all possible regressions)*

Käydään läpi kaikki selittäjien kombinaatiot. Tutkittavien mallien lukumäärä kasvaa kakkosen potensseissa, joten työ käy melko pian käytännössä mahdottomaksi. Muodostetuista malleista etsitään muutama parempi (jollakin kriteerillä) ja tutkitaan niitä tarkemmin.

2. Lisäävä valinta *(forward selection)*

Aloitetaan mallista jossa on vain vakio. Lisätään paras, yksittäinen selittäjä (se joka korreloi eniten selitettävän kanssa). Jatketaan lisäämällä malliin yksi kerrallaan sillä hetkellä paras selittäjäehdokas. Kolme yhtäpitävää kriteeriä:

1) suurin osittaiskorrelaatio selitettävän muuttujan kanssa, kun mallissa jo mukana olevat tekijät vakioidaan
2) suurin lisäys mallin selitysasteeseen
3) suurin t- tai F-arvo niistä joita ei ole vielä lisätty

Lisääminen lopetetaan, kun

a) ennalta määrätty selittäjien lukumäärä saavutetaan tai
b) minkä tahansa ei-valitun muuttujan F-arvo on pienempi kuin ennalta valittu kynnysarvo *("F-IN", "F-to-enter")*.

3. Poistava valinta *(backward elimination)*

Aloitetaan täydestä mallista jossa ovat mukana kaikki selittäjät. Poistetaan selittäjiä yksi kerrallaan. Kriteerit kuten edellä mutta kääntäen: 1) pienin osittaiskorrelaatio, 2) pienin vähennys, 3) pienin t- tai F-arvo niistä joita ei ole vielä poistettu.

Lopetusehdot vastaavasti: a) sama kuin edellä, b) mallin kaikkien selittäjien F-arvo on suurempi kuin ennalta valittu kynnysarvo *("F-OUT", "F-to-remove")*.

4. Askeltava valinta *(stepwise method)*

Yhdistelmä kahdesta edellisestä: lähdetään ns. tyhjästä mallista kuten lisäävässäkin valinnassa, ja edetään lisäämällä ja poistamalla muuttujia tilanteen mukaan.

Erilaisia valintoja saadaan muuttelemalla F-testejä säätteleviä kynnysarvoja. Toiminta on siis luonteeltaan varsin heuristista, eikä testaustilanteiden osalta vastaa läheskään aina testien taustalla olevia oletuksia. On huomattava mm. seuraavaa:

- Kun mallista poistetaan selittäjä, on F-testi voimassa ehdolla että muuttujan sisältämä malli on oikea.
- Kun malliin lisätään selittäjä, sama ehto onkin epälooginen.
- Ennalta valitut F-testin kynnysarvot vaikuttavat huomattavasti valintoihin.
- Multikollineaarisuus (selittäjien keskinäinen korrelointi) muuttaa testisuureiden arvoja tilanteesta toiseen.
- Lisäävän valinnan erityinen haittapuoli on, että ns. supressiiviset muuttujat (joiden vaikutus tulee esiin vain yhdessä jonkin toisen muuttujan kanssa) jäävät helposti mallin ulkopuolelle.
- Eri valintamenetelmät johtavat yleensä eri malleihin.

Mallin valintakriteerit

Mallin selittäjien valinnan jälkeen on yleensä edessä seuraava valintatilanne: miten valitaan vaihtoehtoisista malleista "paras"? Apuna voidaan käyttää mm. seuraavia valintakriteerejä.

1. Selitysaste

Valitaan se malli jonka selitysaste R^2 (1.24) on suurin. Etenkin ilmiön kuvailussa R^2 on hyvin yleisessä käytössä: "Malli selittää 62% kokonaisvaihtelusta". Selitysasteen suosio perustuukin sen helppoon tulkittavuuteen. Sen huonoja puolia on mm. se että R^2 kasvaa, kun malliin lisätään muuttujia, oli näillä todellista selitysvoimaa tai ei.

2. Korjattu selitysaste

Kun R^2-arvoon tehdään vapausastekorjaus (siirtymällä kaavassa 1.24 neliösummista SSE ja SST vastaaviin harhattomiin varianssiestimaatteihin MSE ja MST), saadaan ns. korjattu selitysaste (*adjusted* R^2)

$$R^2_{adj} = 1 - \frac{SSE / (n-k-1)}{SST / (n-1)} \quad , \tag{1.49}$$

joka on sikäli parempi kuin R^2, että se ei välttämättä kasva kun selittäjiä lisätään.

3. Jäännösvarianssi

Valitaan malli jonka jäännösvarianssi

$$s^2 = \frac{SSE}{(n-k-1)} \tag{1.50}$$

on pienin.

4. Mallowsin C_p

Suosituin kriteereistä on pitkään ollut Mallowsin standardoitu kokonaisneliövirhe

$$C_p = \frac{SSE_p}{s^2} - (n - 2p) \quad , \tag{1.51}$$

jossa s^2 on täyden mallin jäännösvarianssi. Hyvillä malleilla C_p:n arvo on lähellä p:tä, mahdollisesti pienempikin. Mallit joilla $C_p > p$, ovat harhaisia. Käytännössä lasketaan esim. C_p kaikille malleille ja tutkitaan asiaa graafisesti *(p,C_p)* -koordinaatistossa. Kuvan perusteella valitaan harhattomista malleista tulkinnaltaan selvin, tai se jonka C_p-arvo on pienin.

5. AIC

Toiminnallisuudeltaan Mallowsin C_p:tä vastaa Akaiken informaatiokriteeri

$$\mathrm{AIC} = n \log(s_p^2) + 2p \,, \tag{1.52}$$

jossa s_p^2 on p muuttujan mallin jäännösvarianssin suurimman uskottavuuden estimaatti

$$s_p^2 = \frac{\mathrm{SSE}}{n} \,. \tag{1.53}$$

6. SBIC

Samantapainen, hieman vähäparametrisempia malleja suosiva valintakriteeri on Schwarzin bayesiläinen informaatiokriteeri

$$\mathrm{SBIC} = n \log(s_p^2) + p \log(n) \,. \tag{1.54}$$

7. MDL

Valintakriteerit 1.–6. ovat olleet käytössä jo vuosikymmeniä. Uudempia tulokkaita edustavat ns. MDL-periaatteeseen *(Minimum Description Length)* nojaavat kriteerit, joiden juuret ovat informaatioteorian puolella. *Reijo Sund* on laatinut MDL-pohjaisesta mallinvalinnasta erinomaisen esityksen, joka on saatavilla verkosta (linkki kurssin kotisivulla). Sen liitteessä on kuvattu, miten eri kriteerien arvot on laskettu Survolla. Laskelmien pitäisi olla tämän kurssin ensimmäisen harjoituksen matriisilaskentatehtävän läpikäyneille selkeää luettavaa, etenkin kun operaatiot on kommentoitu mallikkaasti.

Osittaiskorrelaatiodiagrammi

Seuraavassa esimerkissä, joka esitellään tarkemmin luennolla, käydään läpi selittäjän lisääminen regressiomalliin. Lopuksi päädytään ns. osittaiskorrelaatiodiagrammiin, jonka avulla asiaa voidaan havainnollistaa kuvallisesti. Aineistona ovat jo aiemmin esillä olleet pankkien transaktiolukumäärät `T1` ja `T2` sekä niiden suorittamiseen kulunut kokonaisaika `Time`.

Laaditaan aluksi malli, jossa selitettävänä on `Time` ja selittäjänä `T1`. Otetaan residuaalit talteen muuttujaan `eTime|T1`. [Käytän tässä esimerkissä havainnollisuuden vuoksi `VARS`-täsmennystä; yleensä `MASK` on käytännössä kätevämpi.]

```
VAR eTime|T1=MISSING TO TRANSACT
LINREG TRANSACT CUR+1 / VARS=Time(Y),T1(X),eTime|T1(R)
Linear regression analysis: Data TRANSACT, Regressand Time      N=261
Variable Regr.coeff.    Std.dev.    t     beta
T1        12.67175       0.460546  27.51  0.863
constant  3043.967       175.3228  17.36
Variance of regressand Time=14243435.22 df=260
Residual variance=3644776.086 df=259
R=0.8632 R^2=0.7451
DW=1.7097
```

Entä jos nyt lisätään malliin `T2`? Tavoitteena olisi selittää `Time`-muuttujasta *se osuus jota* `T1` *ei selitä*. Sovitetaan siis malli, jossa äskeisiä residuaaleja `eTime|T1` selitetään `T2`:lla. Otetaan residuaalit talteen muuttujaan `RES`:

```
VAR RES=MISSING TO TRANSACT
LINREG TRANSACT CUR+1 / VARS=eTime|T1(Y),T2(X),RES(R)
Linear regression analysis: Data TRANSACT, Regressand eTime|T1  N=261
Variable Regr.coeff.    Std.dev.    t     beta
T2        0.823350       0.086243  9.547  0.510
constant -1993.909       232.2720 -8.584
Variance of regressand eTime|T1=3630757.686 df=260
Residual variance=2696040.371 df=259
R=0.5102 R^2=0.2603
DW=1.6134
```

Tämä ei anna toivottua vastausta, sillä T1 ja T2 riippuvat toisistaan. Näin ollen osa T2:n sisältämästä informaatiosta on *redundanttia*, siis jo mukana mallissa.

Ratkaisu ongelmaan on sovittaa malli, jossa T2*:ta selitetään* T1*:llä*, ja ottaa siitä talteen residuaalit eT2|T1:

```
VAR eT2|T1=MISSING TO TRANSACT
LINREG TRANSACT CUR+1 / VARS=T2(Y),T1(X),eT2|T1(R)
Linear regression analysis: Data TRANSACT, Regressand T2        N=261
Variable Regr.coeff.    Std.dev.    t     beta
T1        3.543634       0.181545  19.52  0.772
constant  1425.180       69.11131  20.62
Variance of regressand T2=1394126.526 df=260
Residual variance=566359.7895 df=259
R=0.7716 R^2=0.5953
DW=1.6520
```

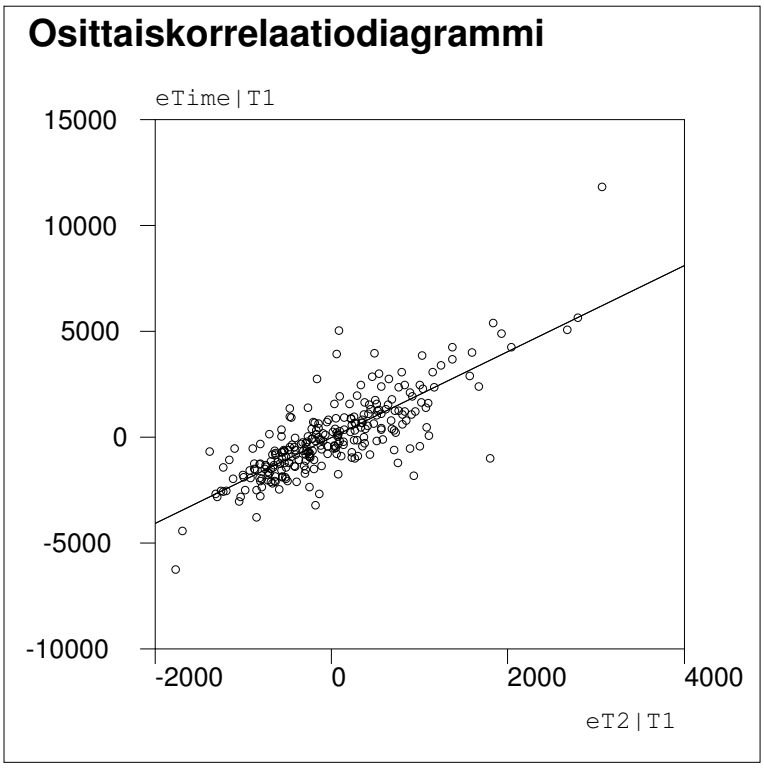
Nyt residuaaleiksi jää täsmälleen *se osa* T2*:sta joka ei selity* T1*:llä*.

Lopulta sovitetaan malli, jossa ensimmäisen mallin residuaaleja eTime|T1 selitetään toisen mallin residuaaleilla eT2|T1. Toisin sanoen *ensimmäisen* T1*-mallin selittämätöntä osaa selitetään sillä osalla* T2*:sta joka ei enää sisällä tietoa* T1*:stä*. Käytetään residuaalimuuttujana muuttujaa RES:

```
LINREG TRANSACT CUR+1 / VARS=eTime|T1(Y),eT2|T1(X),RES(R)
Linear regression analysis: Data TRANSACT, Regressand eTime|T1  N=261
Variable Regr.coeff.    Std.dev.    t     beta
eT2|T1    2.034549       0.094155  21.61  0.802
constant -0.000001       70.58575 -0.000
Variance of regressand eTime|T1=3630757.686 df=260
Residual variance=1300392.830 df=259
R=0.8020 R^2=0.6432
DW=1.7114
```

Näin saadun mallin regressiokerroin kertoo, miten paljon selitettävä muuttuja Time muuttuu, kun T2 muuttuu yhden yksikön verran ja T1 on *otettu myös huomioon mallissa ("adjusted for* T1*")*. Tämä on juuri se mitä tässä haettiinkin.

Hajontakuvaa, jossa ovat eTime|T1 pystyakselilla ja eT2|T1 vaaka-akselilla, kutsutaan **osittaiskorrelaatiodiagrammiksi** *(Added Variable Plot, AVP)*. Sen avulla tutkitaan, onko perusteltua lisätä malliin uutta selittäjää. Mikäli selittäjät korreloivat vahvasti keskenään, on AVP:n antama kuva kuitenkin harhainen ja siten hyödytön.



Suomenkielinen nimitys viittaa siihen, että tutkittaessa edellä kuvattua ongelmaa tullaan itse asiassa laskeneeksi *osittaiskorrelaatiot* kolmannen muuttujan suhteen. Se tarkoittaa ko. muuttujan *lineaarisen vaikutuksen eliminointia* eli *vakiointia*.

Usean muuttujan regressiomallissa selittäjät tulevat automaattisesti vakioitua toistensa suhteen, eli vain kunkin selittäjän itsenäinen osuus on mallissa mukana. Mitä tahansa selittäjää voidaan näin ollen tarkastella olettaen muut vakioiduiksi (vrt. "Regressiokertoimet", s. 4).

Estimoidaan lopuksi täysi malli (molemmat selittäjät mukana). Otetaan residuaalit talteen muuttujaan `RES2`:

```
VAR RES2=MISSING TO TRANSACT
LINREG TRANSACT CUR+1 / VARS=Time(Y),T1(X),T2(X),RES2(R)
Linear regression analysis: Data TRANSACT, Regressand Time      N=261
Variable Regr.coeff.    Std.dev.    t     beta
T1        5.462057       0.433268  12.61  0.372
T2        2.034549       0.094337  21.57  0.637
constant  144.3694       170.5441  0.847
Variance of regressand Time=14243435.22 df=260
Residual variance=1305433.129 df=258
R=0.9534 R^2=0.9091
DW=1.7114
```

Kysymys: Miten täyden mallin residuaalit (`RES2`) ja AVP-mallin residuaalit (`RES`) suhtautuvat toisiinsa? Miksi?

Tarkastellaan lopuksi muuttujan lineaarisen vaikutuksen eliminointia eli vakiointia eksplisiittisesti. Lasketaan ensin edelläolevan AVP-kuvan muuttujien välinen korrelaatiokerroin:

```
CORR TRANSACT / VARS=eT2|T1,eTime|T1
MAT LOAD CORR.M
MATRIX CORR.M
R(TRANSACT)
///         eT2|T1 eTime|T1
eT2|T1    1.000000 0.802008
eTime|T1  0.802008 1.000000
```

Korrelaatio on siis 0.80. Lasketaan sitten alkuperäisten muuttujien korrelaatiot:

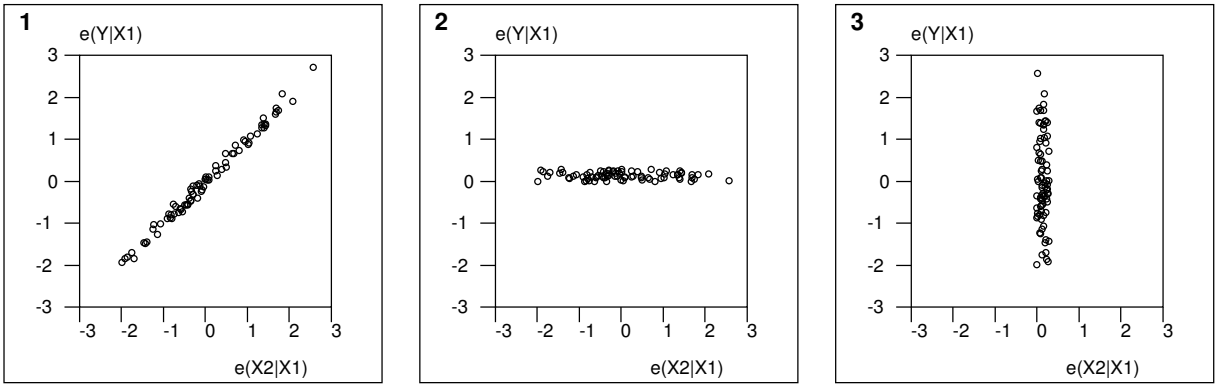
```
CORR TRANSACT / VARS=Time(A),T1(Z),T2(A) <- T1 matriisissa viimeiseksi
MAT LOAD CORR.M
MATRIX CORR.M
R(TRANSACT)
///           Time       T2       T1
Time      1.000000 0.923597 0.863187
T2        0.923597 1.000000 0.771567
T1        0.863187 0.771567 1.000000
```

Tämän avulla lasketaan lopuksi muuttujien `Time` ja `T2` osittaiskorrelaatiokerroin. Tarvittavissa matriisikaavoissa (jotka tässä on koottu ns. matriisiketjuksi `PARTCORR`) käsitellään korrelaatiomatriisia sopivasti ositettuna (ks. esim. Saikkonen 2002 tai Mustonen 1995, 22):

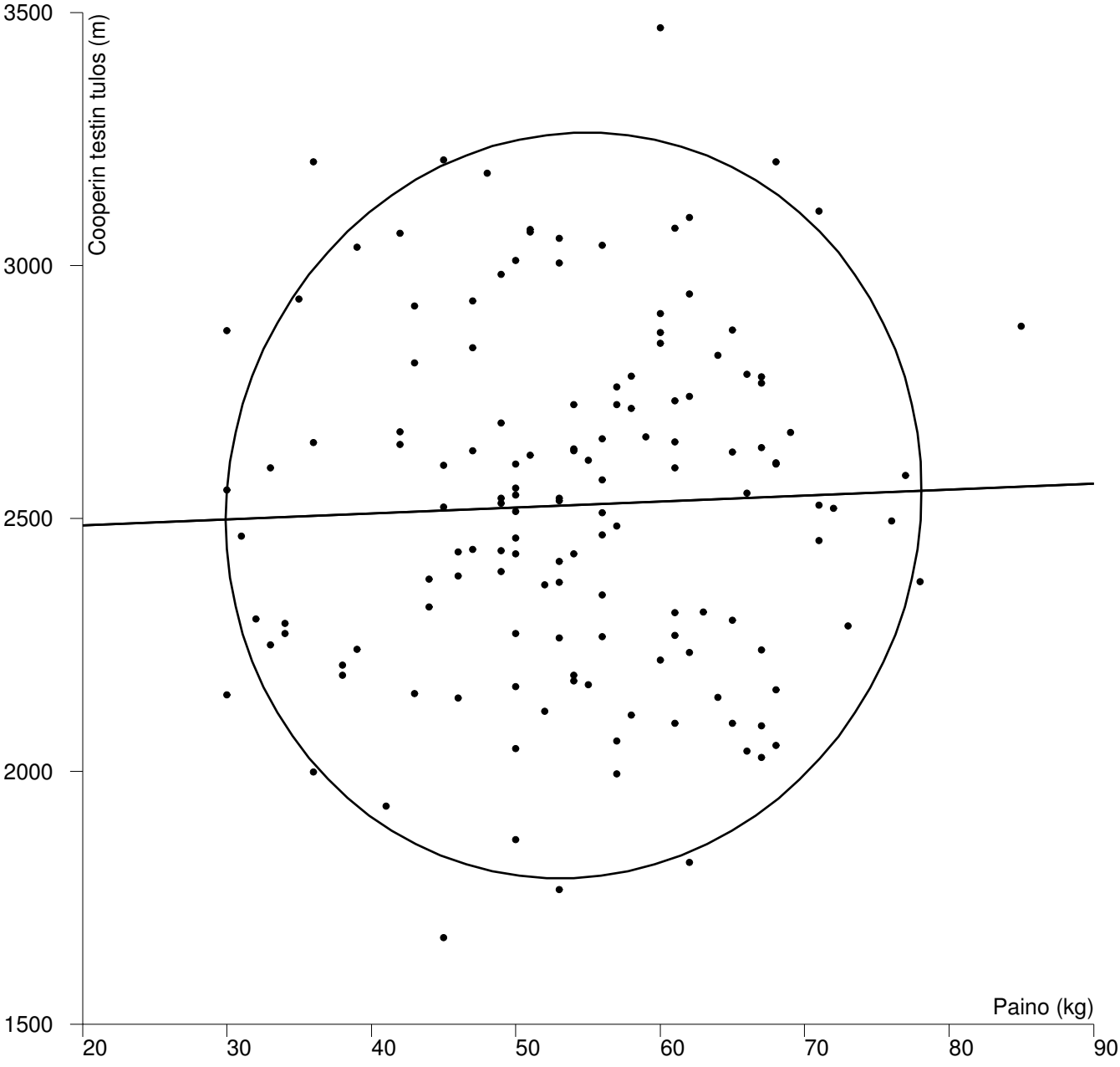
```
MATRUN PARTCORR CORR.M,1,PCORR.M / vakioidaan viimeisen muuttujan (T1) suhteen
MAT LOAD PCORR.M
MATRIX PCORR.M
Partial_correlations
///           Time       T2
Time      1.000000 0.802008
T2        0.802008 1.000000
```

Kuten nähdään, muuttujien `Time` ja `T2` osittaiskorrelaatiokerroin on täsmälleen AVP-residuaalien korrelaatiokerroin. Siis: `Time`:n ja `T2`:n korrelaatiokerroin on 0.92, mutta jos `T1`:n vaikutus eliminoidaan eli `T1` vakioidaan, se onkin vain 0.80. Mallin rakentamisen kannalta osittaiskorrelaatio ja -diagrammi antavat selvän viitteen, että `T2` kannattaa lisätä malliin.

Kysymys: Miten tulkitsisit seuraavan sivun osittaiskorrelaatiodiagrammit (1,2,3)?



Oheisessa kuvassa on piirretty vastakkain 9-19-vuotiaiden poikien paino (kg) ja Cooperin testin tulos (m). Kuvaan on myös lisätty regressiosuora sekä 90 %:n tasolle piirretty *hajontaellipsi*. (Kuvan N=140 havainnosta 90%*N=126 on ellipsin kehän sisäpuolella ja loput 14 ulkopuolella.) Mitä voisit kuvan perusteella sanoa muuttujien riippuvuudesta? Miten tulkitsisit tilannetta kokonaisuudessaan?



II Diagnostiikka ja muunnokset

Mallin valinnassa huomiota kiinnitettiin pääasiassa jäännösvarianssiin ym. yleiskriteereihin. Diagnostiikka on vielä laajempi ja samalla yksityiskohtaisempi mallintamisen osa-alue, jonka keskeisinä elementteinä toimivat residuaalit sekä eräät muut diagnostiset mitat. Aivan olennaisen osan diagnostiikasta muodostavat graafiset tarkastelut.

Diagnostiikan tarkoituksena on varmistaa, että aineistoon sovitettu malli täyttää sille asetetut vaatimukset. Mahdollisten epäkohtien syyt on tutkittava, ja tehtävä tarvittavat korjaukset malliin (tai aineistoon, mikäli siitä paljastuu korjattavissa olevia virheitä). Tärkeimmät diagnostiikan kohteet ovat:

1. Mallin harhattomuus
2. Jäännösvaihtelun homoskedastisuus
3. Mallivirheiden normaalisuus
4. Mallivirheiden korreloimattomuus
5. Multikollineaarisuus
6. Vaikutusvaltaiset havainnot

1. Mallin harhattomuus

Mallin on annettava keskimäärin oikea tulos kaikilla mahdollisilla selittäjien arvojen yhdistelmillä. Harhattomassa mallissa $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathrm{E}(\mathbf{y}|\mathbf{X})$ kaikilla $\mathbf{X}$:n arvoilla. Samaa ilmaisee oletus (1.2a). Asiaa tutkitaan useimmiten graafisesti: arvioidaan silmämääräisesti, riippuvatko residuaalit jollain tavoin sovitteesta, selittäjistä tai mallista puuttuvista potentiaalisista selittäjistä. Mallin harhaisuus voi paljastua mistä tahansa kyseisistä kuvista.

Sovellusalan teorian tuntemuksella on tärkeä merkitys:

- Jos tuntee tutkittavan ilmiön, tietää aineistoon sovitettavan mallin muodon.
- Jos ei tunne tutkittavaa ilmiötä, joutuu tekemään enemmän harhattomuustarkasteluja, eikä erota, johtuvatko poikkeamat väärin spesifioidusta mallista vai tutkimusaineistosta.

Harhattomuuden sekä mallin tulkittavuuden tulisi olla etusijoilla mallintamisprosessissa.

2. Jäännösvaihtelun homoskedastisuus

Jäännösvaihtelun tulisi olla samansuuruista kunkin selittäjän koko vaihteluvälillä. Asiaa tutkitaan parhaiten graafisesti (vrt. kohta 1). Jos heteroskedastisuutta esiintyy, mallia voidaan parantaa joko havaintokohtaisella painotuksella tai sopivilla muunnoksilla (muunnoksia käsitellään myöhemmin erikseen).

Painotus tarkoittaa mallin (1.1) yleistämistä siten, että oletuksessa (1.2b) oleva yksikkömatriisi $\mathbf{I}$ korvataan yleisemmällä diagonaalimatriisilla $\mathbf{V}$. Tällä tavoin havaintojen eri suuruiset varianssit voidaan ottaa huomioon mallissa. Kaikki aiemmat tarkastelut ovat yleistettävissä tähän ns. painotetun PNS-menetelmän tapaukseen.

3. Mallivirheiden normaalisuus

Malliin liittyvien hypoteesien testaus edellyttää, että mallivirheen normaalijakaumaoletus (1.2c) on voimassa. Muussa tapauksessa testeihin perustuva päätöksenteko on hataralla pohjalla. Jos mallia ei voida tässä suhteessa parantaa, on siirryttävä yleistettyyn lineaariseen malliin, jossa mallivirheen todennäköisyysjakauma voidaan spesifioida tarkemmin.

Oletuksen (1.2c) voimassaoloa tutkitaan testaamalla residuaalien normaalisuutta mm. erilaisilla normaalisuustesteillä ja ns. todennäköisyyspaperikuvalla *(normal probability plot)*. Ks. myös Puranen (1997, 177–188).

4. Mallivirheiden korreloimattomuus

Sellaisissa malleissa joissa havaintojen järjestys on kiinteä (esimerkiksi aikasarja-aineistot ja alueelliset aineistot), residuaalien välillä saattaa esiintyä ns. autokorrelaatiota. Mallia voidaan parantaa ottamalla tämä huomioon, eli yleistämällä mallia edelleen antamalla mallivirheiden korreloida keskenään. Tällöin $\mathbf{V}$ ei ole enää diagonaalinen, vrt. kohta 2.

Residuaalien keskinäisiä korrelaatioita tutkitaan yleensä vain jos havainnot riippuvat toisistaan esimerkiksi ajallisesti tai maantieteellisesti. Autokorrelaation testaukseen käytetään mm. Durbinin ja Watsonin testisuuretta, jonka monet regressioanalyysiohjelmat tulostavat joka tapauksessa. Asiasta kertoo tarkemmin Puranen (1997, 251–257).

Aikasarja-analyysin ja ekonometrian alueilla tavallinen regressiomalli on näine yleistyksineenkin useimmiten riittämätön ilmiöihin sisältyvän dynamiikan vuoksi. Monipuolisempia mahdollisuuksia tarjoavat mm. ARIMA *(autoregressive integrated moving average)* -mallit.

5. Multikollineaarisuus

Jos selittäjien välillä on voimakkaita riippuvuuksia, malliin saattaa päätyä tulkinnan kannalta "vääriä" muuttujia. Myös muut diagnostiset tarkastelut voivat tästä syystä vaikeutua.

Eräitä keinoja multikollineaarisuuden poistamiseen ovat:

- korreloivista muuttujista muodostetut uudet muuttujat, esim. painoindeksi
- muutosten tai suhteellisten muutosten tutkiminen kasvuilmiöissä kokonaismäärien sijaan
- indeksien käyttö taloudellisissa aineistoissa
- aineiston sisältämän informaation tiivistäminen monimuuttujamenetelmillä (esimerkiksi faktorianalyysilla tai pääkomponenttianalyysilla) ja alkuperäisten selittäjien korvaaminen tällä tavoin muodostetuilla korreloimattomilla muuttujilla
- lisähavaintojen hankinta (periaatteessa hyvä, mutta usein käytännössä mahdotonta)
- valikoiva regressioanalyysi (ei hyvä, saattaa johtaa aivan vääränlaisiin malleihin)
- harjanne-estimointi *(ridge regression)* (laskennallinen keino yrittää kiertää ongelma, suhtauduttava varsin kriittisesti)

Nämäkään keinot eivät kuitenkaan välttämättä auta. Joskus voimakas riippuvuus on vain hyväksyttävä osaksi tutkittavaa ilmiötä. Lisää aiheesta esim. Puranen (1997, 257–265).

Multikollineaarisuuden voi havaita mm.

- selittäjien korrelaatiomatriisista (voimakkaita korrelaatioita)
- matriisin $\mathbf{X}'\mathbf{X}$ ominaisarvoista (pienimmät lähellä nollaa)
- mallimatriisin $\mathbf{X}$ kuntoisuusluvusta *(condition number)*

$$\kappa(\mathbf{X}) = \frac{\mu_{max}}{\mu_{min}} \quad , \tag{2.1}$$

  jossa μ_{max} ja μ_{min} ovat matriisin $\mathbf{X}$ suurin ja pienin singulaariarvo

- VIF *(variance inflation factor)* -kertoimista

$$VIF_i = \frac{1}{1 - R_i^2} \quad , \tag{2.2}$$

  missä R_i^2 on selitysaste mallista jossa muuttujaa x_i selitetään muilla mallin selittäjillä

Mitä suurempi on $\kappa(\mathbf{X})$, sitä enemmän multikollineaarisuutta ilmenee. Jos $\kappa(\mathbf{X}) > 30$, on syytä tutkia mallia näiltä osin tarkemmin. Vastaavasti mitä suurempi selitysaste R_i^2 on kaavassa (2.2), sitä suurempi on VIF-kerroin ja kyseisen regressiokertoimen estimaattorin varianssi. Selittäjien välillä on tällöin havaittavissa multikollineaarisuutta. Kannattaa huomata, että VIF-kertoimet nähdään suoraan selittäjien välisen korrelaatiomatriisin käänteismatriisin diagonaalilta (Puntanen 1999b, 373–377).

6. Vaikutusvaltaiset havainnot

Vaikutusvaltaisten havaintojen tutkimista käsitellään myöhemmin erikseen.

Riippuvuuksien linearisointi

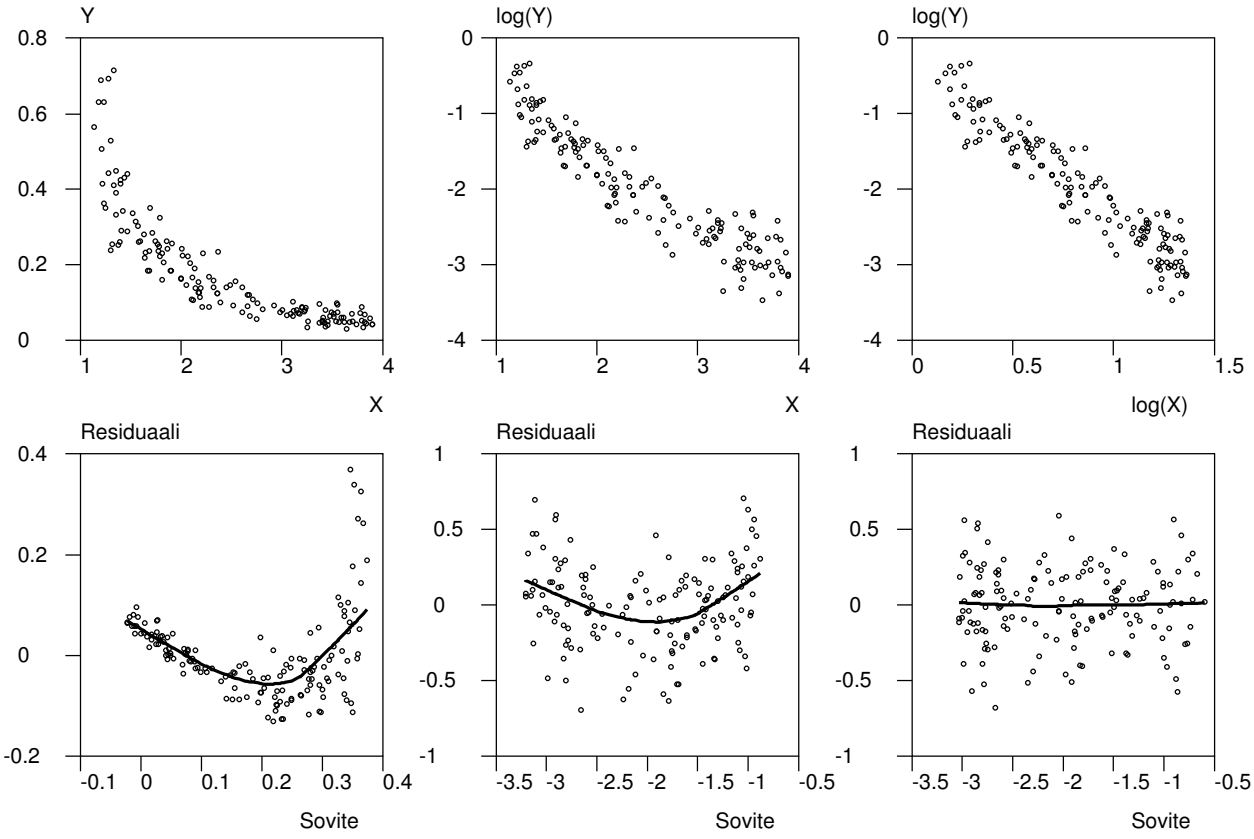
Korrelaatiodiagrammeista havaittava käyräviivainen riippuvuus on otettava jollain tavoin huomioon mallia rakennettaessa. Aineistoon voidaan kenties sovittaa epälineaarinen funktio (mikä voikin usein tuntua houkuttelevalta kuvan perusteella), mutta jos mallintaminen on yleensäkin vaikeaa, on epälineaarinen mallintaminen vielä hankalampaa. Muuttujien välistä riippuvuutta kannattaakin yrittää linearisoida sopivien muunnosten avulla. Jos linearisointi onnistuu, niin mallin analysointi selkiytyy. Esimerkiksi harhattomuustarkastelut tulevat yksinkertaisemmiksi. Lineaaristen riippuvuuksien hahmottaminen on muutenkin helpompaa.

Usein linearisointi myös normalisoi jäännösvaihtelua, jolloin malliin liittyvä jakaumaoletus (1.2c) on paremmin voimassa, ja hypoteesien testaus vankemmalla pohjalla. Tärkeintä on kuitenkin, että jäännösvaihtelun homoskedastisuusoletus (1.2b) saadaan pätemään. Sopivalla linearisoinnilla heteroskedastisuus häviää, eikä havaintokohtaista painotusta tarvita.

Muunnosten valinta voi perustua sovellusalan teoriaan, esimerkiksi fysikaalisten ilmiöiden lakeihin, kasvumalleihin, alan aiempiin tutkimuksiin tai alalla muuten vallitseviin käytäntöihin. Muunnoksia voi hakea myös puhtaasti kokeilemalla.

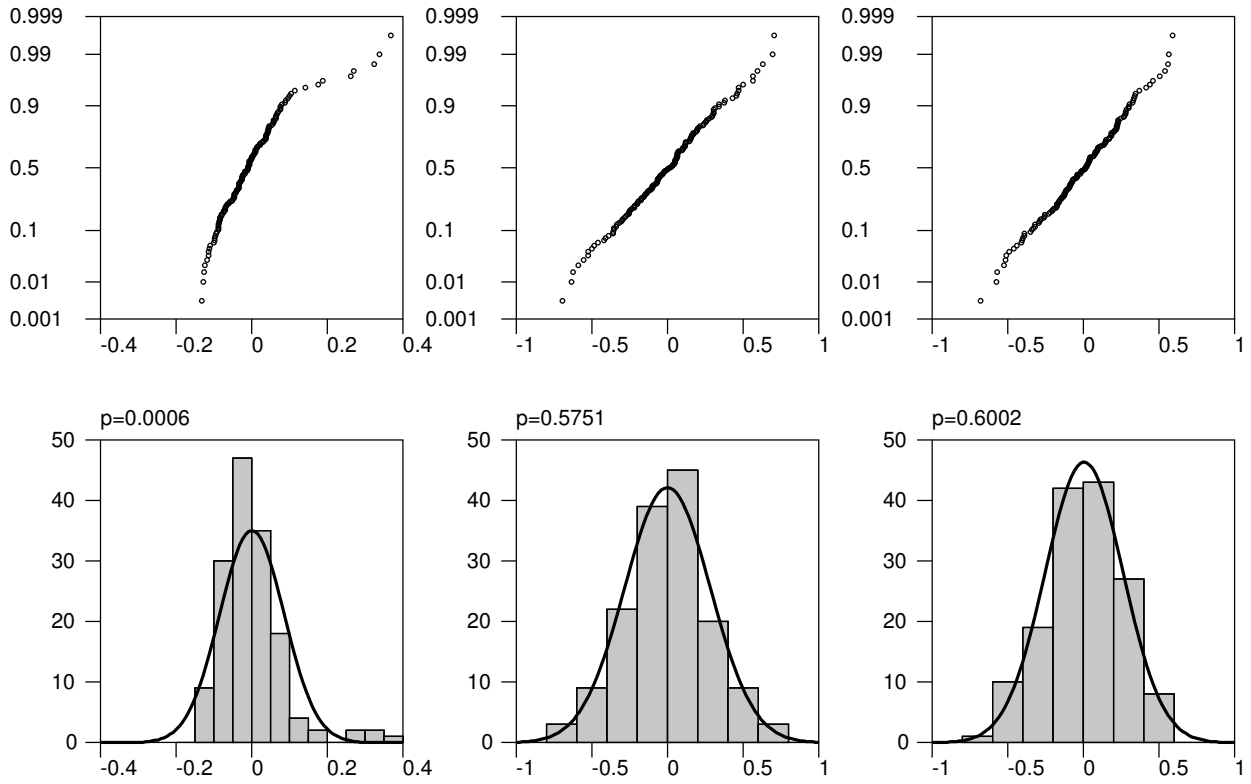
Kokeellisessa linearisoinnissa pyritään erilaisten malliin liittyvien kuvien avulla päättelemään, minkä tyyppinen muunnos linearisoisi muuttujien välisen riippuvuuden ja vakioisi jäännösvaihtelun. Tarkasteluissa on erityisesti syytä tutkia, **onko hajontakuvion leveys Y-akselin suunnassa vakio kaikilla X:n arvoilla**. Jos Y:n suuntainen hajonta riippuu X:n arvoista, on pyrittävä ensin vakioimaan jäännösvaihtelu muuntamalla Y-muuttujaa. Vasta kun Y-akselin suuntainen hajonta on likimain vakio kaikilla X:n arvoilla, tutkitaan hajontakuvion yleistä muotoa. Mikäli riippuvuus ei siinä ole lineaarista, muunnetaan X-muuttujaa.

Oheisissa kuvissa on esillä 150 havainnon simuloituun aineistoon perustuva tilanne, jossa on tarpeen muuntaa sekä Y- että X-muuttujaa. Ylemmissä kuvissa on muuttujien hajontakuvat ja alemmissa vastaavat jäännösvaihteludiagrammit eri vaiheissa. Jälkimmäisiin on lisätty myös lowess-tasoitus. Ensin on muunnettu Y:tä, sillä Y-akselin suuntainen hajonta on huomattavasti suurempaa X:n vaihteluvälin alku- kuin loppupäässä. Kokeilemalla joitakin eri muunnoksia ja piirtämällä hajontakuvia on päädytty Y:n logaritmointiin.



Kun Y on muunnettu, hajontakuva näyttää jo paremmalta, mutta jäännösvaihteludiagrammi paljastaa, että malli olisi selvästi harhainen. On siis muunnettava myös X-muuttujaa. Tässä tapauksessa myös sille näyttäisi parhaiten sopivan log-muunnos. Tämän jälkeen hajontakuva näyttää varsin hyvältä, eikä jäännösvaihtelu anna enää aihetta epäilyksille. Heteroske-

dastisuudesta ei ole tietoakaan. Lisäksi residuaalit normalisoituvat jo ensimmäisen logaritmoinnin ansiosta, mikä näkyy oheisista todennäköisyyspaperikuvista ja histogrammeista.



Parittaisten hajontakuvien lisäksi on siis syytä tarkastella myös jäännösvaihtelu- ja useamman selittäjän malleissa myös osittaiskorrelaatiodiagrammeja. Kuvissa kannattaa käyttää tasoituksia, sillä silmä valehtelee helposti. Kun sopivantuntuiset muunnokset on löydetty, muodostetaan tarvittaessa uudet muuttujat ja piirretään niihin liittyvät hajontakuvat, toisin sanoen palataan alkuun. Jos muunnokset eivät tyydytä, yritetään tehdä parempia.

Linearisoinnin kannalta on hyödyllistä tietää, miten tietyt matemaattiset funktiot käyttäytyvät, ja miten erilaiset muunnokset vaikuttavat eri tyyppisiin hajontakuviin. Näitä taitoja voi helposti kehittää simuloitujen aineistojen avulla (ks. esim. Puranen 1997, 123–133). Lisäksi voidaan käyttää erityisesti riippuvuuksien linearisointiin kehitettyjä menetelmiä, joista kertoo tarkemmin mm. Puranen (1997, 146–166).

Eräs tapa hakea sopivaa muunnosta erityisesti selitettävälle muuttujalle tunnetaan nimellä *Box-Cox*-muunnos. Kyseessä on muunnosperhe

$$BC_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \log(y) & , \lambda = 0 \, , \end{cases} \qquad (2.3)$$

jossa oletetaan, että $y > 0$. Muunnos on alunperin laadittu muuttujien normalisointia ajatellen, mutta se on osoittautunut laajemminkin käyttökelpoiseksi regressiomallintamisessa, sillä varsin usein normalisointi myös linearisoi muuttujien välisen riippuvuuden.

Ideana on kokeellisesti maksimoida muunnosparametrin λ profiiliuskottavuusfunktiota (katso Tilastollisen päättelyn kurssi, esim. Saikkonen 2002) välillä, joka sisältää tulkittavissa olevia λ:n arvoja, esim. {-2, -1, -0.5, 0, 0.5, 1, 2}. Näin saadut profiiliuskottavuusfunktion arvot piirretään λ:n arvoja vastaan, jolloin voidaan arvioida, mikä muunnos olisi paras. Tarkastelua auttaa maksimiarvolle muodostettava luottamusväli: mikä tahansa luottamusväliin kuuluva, tulkittavissa oleva muunnos on mahdollinen. Yksityiskohtaisemmin muunnoksen taustalla olevasta teoriasta kertoo mm. Puranen (1997, 136–146).

Riippuvuuksien linearisoinnin jälkeen on tärkeää löytää tehdylle muunnokselle ja näin saadulle muunnetulle mallille **tulkinta**. Mikäli tulkinta käy ylivoimaiseksi, saatetaan joutua palaamaan alkuperäisiin muuttujiin, vaikka mallin diagnosointi edellyttäisikin muunnoksia.

Vaikutusvaltaiset havainnot

Diagnostiikan osalta tutkitaan lopuksi tarkemmin ns. vaikutusvaltaisia tai muuten poikkeavia havaintoja. Keskeisessä asemassa on kurssin alussa PNS-menetelmän yhteydessä esitetty ortogonaaliprojektori $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (1.5)–(1.6). Palautetaan mieliin, että **H** projisoi **y**:n sovitteeksi $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ (1.13), ja vastaavasti $(\mathbf{I} - \mathbf{H})$ projisoi **y**:n residuaaleiksi $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ (1.14).

Kuten aiemmin on mainittu, diagnostiikan yhteydessä nähdään, millaisia asioita matriisin **H** alkioista voidaan käytännössä päätellä. Diagnostiikan kannalta on oleellista, että **H** riippuu vain selittäjien **X** arvoista, ei lainkaan **y**:stä. Näin voidaan tutkia yksittäisten havaintojen vaikutusvaltaa mallissa, ts. miten voimakkaasti jokin havainto vetää mallia puoleensa.

Aiemmin on myös todettu, että normaalijakaumaoletuksen (1.2c) tai (1.4) ollessa voimassa residuaalienkin tulisi olla normaalisti jakautuneita, siis $\mathbf{e} \sim \mathrm{N}[\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})]$ (1.33). Yksittäiselle residuaalille pätee tällöin, että

$$e_i \sim \mathrm{N}[0, \sigma^2(1 - h_{ii})] , \tag{2.4}$$

jossa h_{ii} on matriisin **H** *i*. lävistäjäalkio. Tästä voi päätellä, että

$$\frac{e_i}{\sigma\sqrt{1 - h_{ii}}} \sim \mathrm{N}(0,1) , \tag{2.5}$$

mikä antaa viitteen residuaalien teoreettisesta käyttäytymisestä. On kuitenkin huomattava, että σ on käytännössä tuntematon, ja lisäksi residuaalit voivat korreloida keskenään. Kaavoissa (2.4) ja (2.5) esiintyviä **H**:n lävistäjäalkioita h_{ii} kutsutaan vetovoima-arvoiksi tai vipuarvoiksi (*leverage* = vaikutusvalta, vipuvoima).

Kun *i*. havainto jätetään pois aineistosta, voidaan regressioanalyysin kannalta tärkeä matriisi $(\mathbf{X}'\mathbf{X})^{-1}$ laskea ns. päivityskaavalla

$$(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_i\,\mathbf{x}_i'\,(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} , \tag{2.6}$$

jossa $\mathbf{x}_i'$ on *i*. havainto eli **X**-matriisin *i*. rivi, ja $\mathbf{X}_{(i)}$ on **X** josta $\mathbf{x}_i'$ on poistettu. Mikäli *i*. havainto on vaikutusvaltainen, eli sen vetovoima-arvo h_{ii} on suuri, se aiheuttaa selvän muutoksen matriisiin $(\mathbf{X}'\mathbf{X})^{-1}$. Tämä puolestaan heijastuu suoraan mm. regressiokertoimien estimaattoreiden variansseihin, vrt. (1.12). Todettakoon, että kaavan (2.6) esitti jo *C.F.Gauss* 1820-luvulla (ks. Cook & Weisberg 1999, 368).

Matriisin **H** ominaisuuksista (1.6) seuraa, että h_{ii}:n arvot ovat suuruudeltaan keskimäärin $\frac{p}{n}$. Yleensä vaikutusvaltaisina pidetään havaintoja, joilla $h_{ii} > \frac{2p}{n}$. Tällaisiin havaintoihin on syytä kiinnittää huomiota ja tutkia niiden vaikutusta malliin. On silti muistettava, ettei kyseessä ole mikään tilastollinen testaus vaan käytännön perusteella johdettu arvio, jota voi pitää lähinnä suuntaa-antavana neuvona.

Residuaalit

Regressiomallin residuaalit voidaan esittää muodossa

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} . \tag{2.7}$$

Tästä nähdään, että **H**:n lävistäjäalkioiden h_{ii} pitää olla pieniä, jotta residuaalit kuvaisivat tuntematonta mallivirhettä hyvin. Toisaalta, jotta malli olisi järkevällä pohjalla, on havaintojen (siis matriisin **X** rivien) oltava homogeenisia eli lävistäjäalkioiden h_{ii} suurinpiirtein samansuuruisia. Matriisin **H** ominaisuuksien (1.6) perusteella on joka tapauksessa selvää, että $0 \le h_{ii} \le 1$.

Tavalliset residuaalit riittävät mainiosti harhattomuus-, normaalisuus- ja homoskedastisuustarkasteluihin. Havaintokohtaisessa diagnostiikassa on parempi käyttää muita vaihtoehtoja, esimerkiksi **standardoituja residuaaleja**

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} , \; i = 1, 2, ..., n , \tag{2.8}$$

jossa s on jäännöshajonnan tavanomainen estimaatti (vrt. taulukko 1.1)

$$s = \sqrt{\mathrm{SSE} / (n–k–1)} . \tag{2.9}$$

Koska $r_i \sim N(0,1)$, kannattaa kiinnittää huomiota havaintoihin, joilla $|r_i| > 2$. Standardoitujen residuaalien r_i heikkoutena on kuitenkin, että kaavan (2.8) osoittaja ja nimittäjä riippuvat toisistaan, koska SSE on peräisin mallista, jossa *i.* havainto on mukana.

Monella tapaa vielä parempi vaihtoehto onkin ns. **(ulkoisesti) studentoitu residuaali**

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, ..., n, \tag{2.10}$$

missä $s_{(i)}$ on jäännöshajonnan estimaatti mallista, josta on poistettu *i.* havainto:

$$s_{(i)} = \sqrt{SSE_{(i)} / (n-k-2)} \quad . \tag{2.11}$$

Voidaan osoittaa, että $r_i^* \sim t_{n-k-2}$. Tähän tulokseen liittyy aikaisemmin (ks. harjoitus 2) esitetty yksittäisen havainnon osoitinmuuttujan käyttö regressiomallissa. Kyseisessä tilanteessa t-testisuure, jolla testataan ao. osoitinmuuttujan regressiokertoimeen liittyvää hypoteesia $\beta_i = 0$, on täsmälleen kyseisen havainnon studentoitu residuaali ilman osoitinmuuttujaa lasketussa mallissa. Tästä syystä *Arc*-ohjelmassa studentoidut residuaalit r_i^* (2.10) esiintyvät peräti nimellä *Outlier-T*. Joissain yhteyksissä niitä kutsutaan myös *jackknife*-residuaaleiksi. Standardoituja residuaaleja r_i (2.8) kutsutaan sen sijaan toisinaan sisäisesti studentoiduiksi residuaaleiksi.

Muita diagnostisia mittoja

Vetovoima-arvojen ja erilaisten residuaalien lisäksi on olemassa joukko muita mittoja, joilla pyritään havaitsemaan mm. havaintojen poikkeavuutta. Eräs näistä tunnetaan nimellä **Cookin etäisyys** *(Cook's distance)* tai Cookin mitta. Se saadaan kaavasta

$$C_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\, s^2} \quad , \quad i = 1, 2, ..., n, \tag{2.12}$$

jossa $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}' \mathbf{y}_{(i)}$ (vrt. edellä). C_i ilmaisee $\hat{\boldsymbol{\beta}}$:n ja $\hat{\boldsymbol{\beta}}_{(i)}$:n välisen skaalatun etäisyyden.

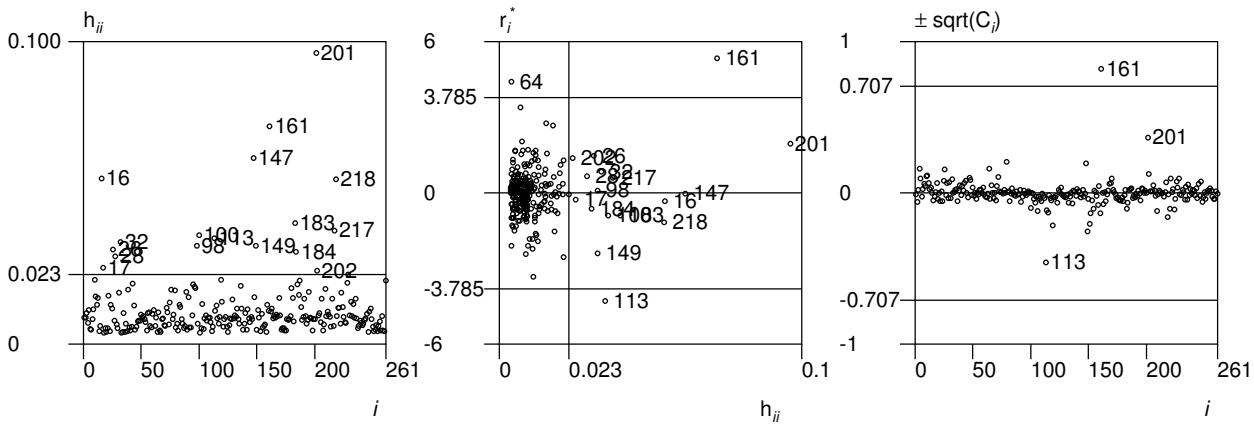
Se voidaan esittää myös muodossa

$$C_i = \frac{1}{p} \frac{h_{ii}}{1 - h_{ii}} r_i^2 \quad , \quad i = 1, 2, ..., n, \tag{2.13}$$

josta nähdään, miten siinä yhdistyvät tiedot havaintojen vaikutusvaltaisuudesta ja poikkeavuudesta. Cookin mitalle on monesti esitetty jakaumatuloksia (ks. esim. Puranen 1997, 197), mutta Cook itse on sitä mieltä, että C_i ei ole mikään tilastollinen testi vaan apuväline muiden diagnostisten mittojen ohella. On yleensä hyödyllistä tutkia havainnot, joiden $C_i > 0.5$, ja ainakin ne, joiden $C_i > 1$. (Cook & Weisberg 1999, 357–358.)

Survon `REGDIAG`-moduli laskee Cookin mitan etumerkilliset neliöjuuret, eli se säilyttää tiedon vastaavan residuaalin suunnasta ja kertoo näin hieman enemmän kuin pelkkä C_i.

Ohessa on eräitä diagnostiikkakuvia, joita käydään tarkemmin läpi luennolla. Aineistona ovat jo monesti esillä olleet pankkien transaktiolukumäärät ja niiden suorittamiseen kulunut kokonaisaika. Diagnosointia helpottamaan on kuviin piirretty kullekin mitalle ominaiset rajat. Studentoitujen residuaalien osalta t-jakaumasta saatavia rajoja on korjattu jakamalla riskitaso havaintojen lukumäärällä (ns. *Bonferroni*-korjaus). Poikkeavimpia havaintoja on merkitty niiden järjestysnumerolla.



III Luokittelevat muuttujat

Tähän asti käsitellyissä malleissa selittäjät ovat olleet pääasiassa jatkuvia muuttujia. Näin ei tarvitse tietenkään olla, vaan tietyin ehdoin myös kategorisia eli luokittelevia muuttujia voidaan käyttää, ja silti soveltaa **regressioanalyysia** edellä opituilla tavoilla. Mallimatriisista **X** oletettiin alussa vain täysiasteisuus; valtaosa muista oletuksistahan koskee joko selitettävää muuttujaa tai mallivirhettä.

Ääritapauksessa kaikki selittäjät ovat luokittelevia muuttujia. Tällöinkin malli voidaan estimoida regressioanalyysilla, mutta yleensä siihen käytetään **varianssianalyysia**, jossa parametrien tulkinta ja testaus on usein yksinkertaisempaa. Regressio- ja varianssianalyysin välimuotoa puolestaan kutsutaan **kovarianssianalyysiksi**. Kaikissa näissä on pohjimmiltaan kysymys samasta lineaarisesta mallista (1.1) sekä siihen liittyvistä oletuksista (1.2a)–(1.2c).

Toinen asia jota ei ole toistaiseksi käsitelty, ovat selittäjien väliset yhdysvaikutukset eli interaktiot, joita voi esiintyä sekä regressio- että varianssianalyysissa. Koska interaktiot liittyvät usein luokitteleviin muuttujiin, ne on luontevaa esittää samassa yhteydessä.

Kategoriset selittäjät eli faktorit

Faktorilla tarkoitetaan tässä yhteydessä kategorista eli luokittelevaa muuttujaa, jolla on kaksi tai useampia tasoja eli luokkia. Faktoreiden luokat ilmaistaan regressiomallissa indikaattori- eli osoitinmuuttujien avulla. Esimerkki osoitinmuuttujasta on jo aiemmin ollut esillä poikkeavien havaintojen tutkimisen yhteydessä, jossa sillä ilmaistiin tavallaan uusi, aineiston perusteella havaittu luokitus ("havainto sopii/ei sovi malliin").

Luokittelumuuttujien arvojen koodaamiseen osoitinmuuttujien avulla on olemassa erilaisia tapoja. Yleisin niistä on ns. *dummy*-koodaus, jossa käytetään dikotomisia (arvoja 0 ja 1 saavia) (dummy-)muuttujia. Ne ovat eräänlaisia toisensa poissulkevien tasojen ilmaisimia.

Tarkastellaan esimerkkiä, jossa selitetään työntekijän palkkaa hänen taustatiedoillaan. Olkoot nämä selittäjät sukupuoli (1=nainen, 2=mies), ikä (vuosina) ja koulutustaso (1=ylioppilas, 2=kandidaatti, 3=maisteri, 4=lisensiaatti, 5=tohtori). (Useamman tutkinnon tapauksessa vain korkein otetaan huomioon.)

Koodataan kategoriset selittäjät dummy-muuttujien S_1–S_2 ja K_1–K_5 avulla:

$S_1 = 1$, jos nainen, 0 muuten
$S_2 = 1$, jos mies, 0 muuten

$K_1 = 1$, jos ylioppilas, 0 muuten
$K_2 = 1$, jos kandidaatti, 0 muuten
$K_3 = 1$, jos maisteri, 0 muuten
$K_4 = 1$, jos lisensiaatti, 0 muuten
$K_5 = 1$, jos tohtori, 0 muuten

Jos (useimmiten kun) mallissa on vakio, kaikkia dummy-muuttujia ei voida sisällyttää malliin yhtaikaa. Jokaista kategorista selittäjää vastaavasta dummy-muuttujien joukosta pitää valita yksi edustamaan ns. vertailuryhmää. Muille dummy-muuttujille estimoitavat regressiokertoimet tulevat tällöin kuvaamaan eroja vertailuryhmiin nähden. Vertailuryhmäksi voidaan valita mikä tahansa; yleensä se johon verrattuina erot ovat mielekkäimmät tulkita.

Kaksiluokkaisten muuttujien osalta ei erillisiä dummy-muuttujia tarvitse tietenkään muodostaa, vaan alkuperäistä muuttujaa voidaan hyödyntää suoraan. On kuitenkin suositeltavaa siirtyä käyttämään arvoja 0 ja 1, jos muuttuja on alunperin koodattu toisin.

Tutkitaan esimerkkitilanteessa aluksi mallia, jossa vain sukupuoli on selittäjänä. Valitaan vertailuryhmäksi naiset, eli käytetään mallissa selittäjänä miehiä osoittavaa dummy-muuttujaa S_2. Kun merkitään palkkaa y:llä, saadaan regressiomallin yhtälö

$$y = \beta_0 + \beta_1 S_2 + \varepsilon , \tag{3.1}$$

joka on yleisen malliyhtälön (1.1) erikoistapaus siinä mielessä, että mallimatriisi **X** koostuu nyt vain nollista ja ykkösistä. Sen rivit voidaan järjestää siten, että ensimmäiset n_1 riviä kuvaavat naisia ja loput n_2 miehiä. Havaintoja on siis yhteensä $n = n_1 + n_2$. Tulosummamatriisi $\mathbf{X}'\mathbf{X}$ on tässä tapauksessa hyvin yksinkertaista tyyppiä, ja dimensioiltaankin vain 2×2, joten sen käänteismatriisikin on helppo muodostaa. Näin ollen regressiokertoimien lausekkeet voi todeta pienellä laskutoimituksella, jonka kuitenkin jätän lisätehtäväksi kaikille 5 ov:n kurssilaisille.

Osoittautuu, että tässä tilanteessa regressiokertoimien PNS-estimaatit ovat

$$\hat{\beta}_0 = \bar{y}_1 \text{ ja } \hat{\beta}_1 = \bar{y}_2 - \bar{y}_1 , \tag{3.2}$$

jossa $\bar{y}_1$ ja $\bar{y}_2$ ovat naisten ja miesten palkkojen keskiarvot.

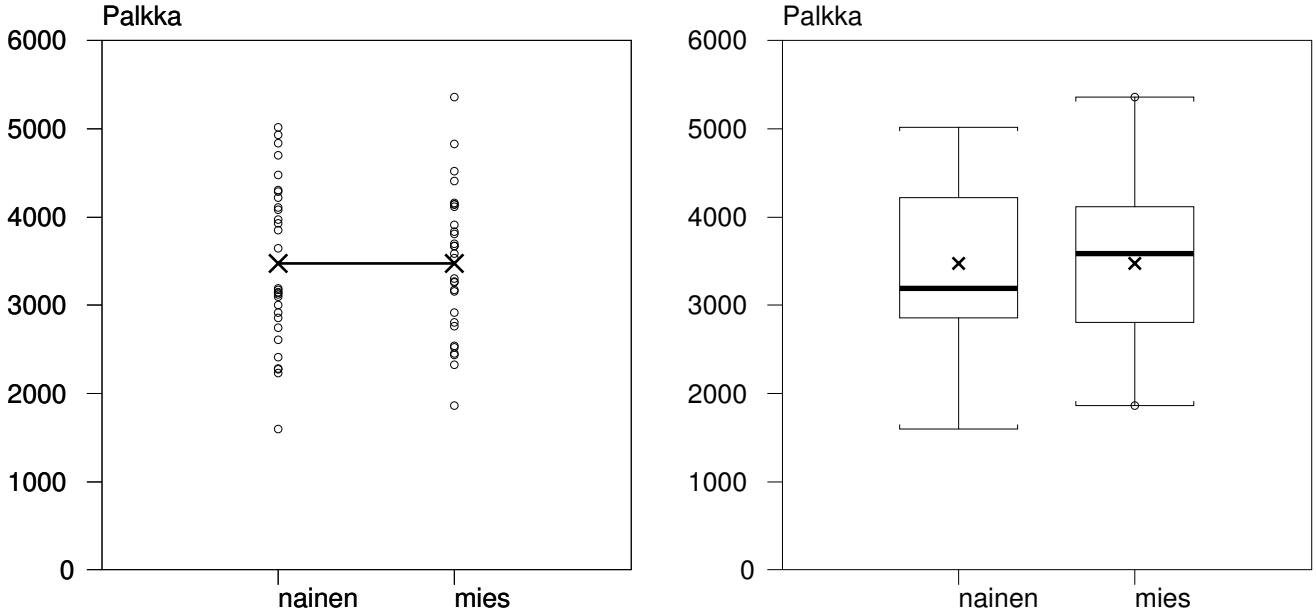
Mallista (3.1) nähdään, että niille havainnoille, joilla $S_2 = 0$ (ts. naisille) mallin systemaattiseksi osaksi jää ainoastaan vakiotermiä kuvaava kerroin β_0. Miehillä myös β_1 on mukana.

On huomattava, että tässä yksinkertaisimmassa tapauksessa (vain yksi dikotominen selittäjä) saadaan ikään kuin molemmille ryhmille omat parametriestimaatit: mallin vakio kuvaa naisten keskipalkkaa, ja ainoa varsinainen regressiokerroin naisten ja miesten keskipalkkojen eroa. Yleisesti mallin vakio sisältää kaikkien luokittelevien muuttujien vertailuryhmien tiedot, ja mielenkiinto kohdistuu vain varsinaisiin tasoeroja kuvaaviin regressiokertoimiin.

Kun kerroinestimaatit (3.2) tiedetään, saadaan sovitteelle ryhmittäin arvot

$$\begin{aligned} \hat{y}_1 &= \bar{y}_1 + (\bar{y}_2 - \bar{y}_1) \cdot 0 = \bar{y}_1 \text{ (naiset) ja} \\ \hat{y}_2 &= \bar{y}_1 + (\bar{y}_2 - \bar{y}_1) \cdot 1 = \bar{y}_2 \text{ (miehet),} \end{aligned} \tag{3.3}$$

eli sovite antaa täsmälleen ryhmien keskipalkat. Graafisesti tätä voidaan havainnollistaa esimerkiksi keskiarvoprofiileilla (vas.) tai boxplotilla (oik.):



Usein keskiarvoprofiileista jätetään yksittäiset havaintopisteet kokonaan pois, etenkin silloin kun samaan kuvaan piirretään useita profiileja. Ryhmäkeskiarvot on kuvissa merkitty rasteilla. Keskipalkoissa ei eroa ole, mutta mediaanipalkoissa jonkin verran, kuten poikkiviivat boxplotissa paljastavat. Tässä käyttämäni aineisto on kuitenkin täysin keinotekoinen; näin tasan eivät palkat valitettavasti jakaannu.

Kuten yleensäkin, regressiokertoimeen β_1 liittyvä t-testi kertoo, poikkeaako kyseinen kerroin (tilastollisesti merkitsevästi) nollasta. Tässä tilanteessa se tarkoittaa, poikkeaako $\bar{y}_2 - \bar{y}_1$ nollasta, eli poikkeaako $\bar{y}_2$ merkitsevästi $\bar{y}_1$:stä.

Kysehän on itse asiassa aivan tavallisesta kahden otoksen t-testistä, jossa nollahypoteesina on H_0: $\mu_1 = \mu_2$ ja vastahypoteesina H_1: $\mu_1 \neq \mu_2$. Myös t-testissä oletetaan havainnot normaalisti jakautuneiksi ja ryhmien varianssit samoiksi. Tämä yksinkertainen tapaus vastaa siis tavallista t-testiä puettuna vain lineaarisen regressiomallin muotoon.

Tarkastelut yleistyvät johdonmukaisesti, kun selittäjä koostuu useammasta luokasta. Jos edellä kuvatussa esimerkkitilanteessa valitaan tohtorit vertailuryhmäksi ja selitetään palkkaa vain koulutuksella, tulee malli muotoon

$$y = \beta_0 + \beta_1 K_1 + \beta_2 K_2 + \beta_3 K_3 + \beta_4 K_4 + \varepsilon . \tag{3.4}$$

Tilastollisesti on samantekevää, mikä luokista valitaan vertailuryhmäksi. Tulkinta ratkaisee.

Vastaavasti kuin edellä (3.2), ovat regressiokertoimien PNS-estimaatit

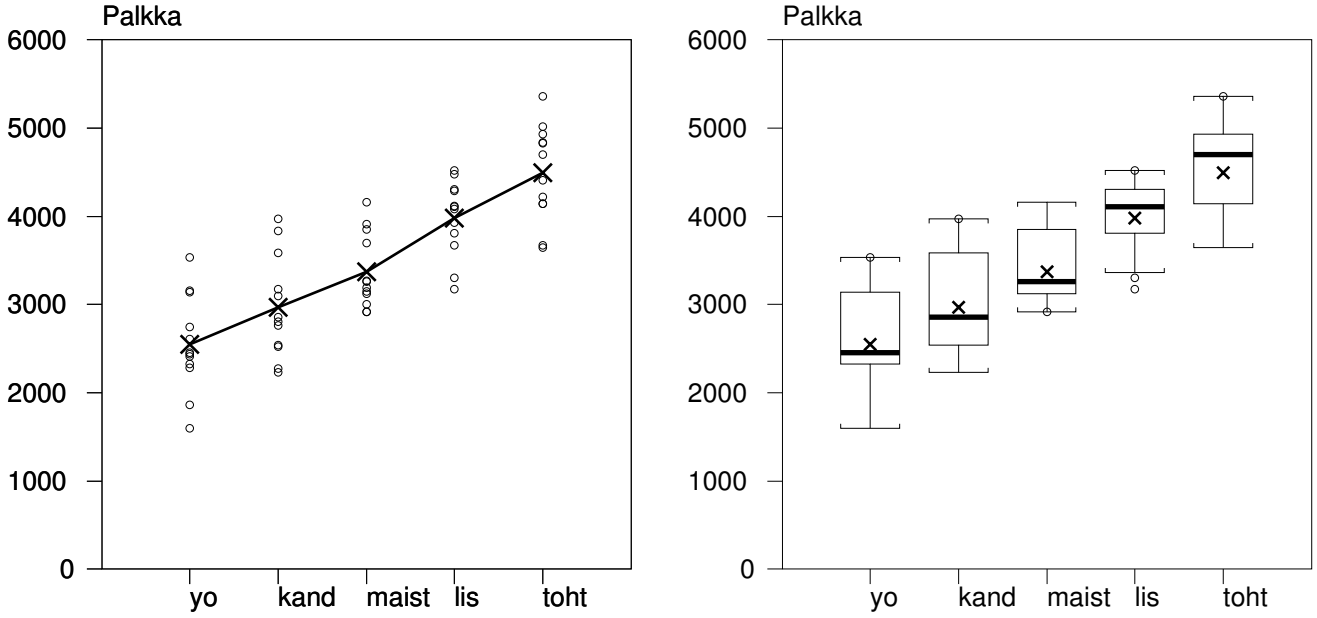
$$\hat{\beta}_0 = \bar{y}_5\,,\ \hat{\beta}_1 = \bar{y}_1 - \bar{y}_5\,,\ \hat{\beta}_2 = \bar{y}_2 - \bar{y}_5\,,\ \hat{\beta}_3 = \bar{y}_3 - \bar{y}_5 \text{ ja } \hat{\beta}_4 = \bar{y}_4 - \bar{y}_5\,, \tag{3.5}$$

joten sovitteen arvoiksi saadaan

$$\begin{aligned}
\hat{y}_1 &= \bar{y}_5 + (\bar{y}_1 - \bar{y}_5)\cdot 1 + (\bar{y}_2 - \bar{y}_5)\cdot 0 + (\bar{y}_3 - \bar{y}_5)\cdot 0 + (\bar{y}_4 - \bar{y}_5)\cdot 0 = \bar{y}_1 \text{ (ylioppilaat)},\\
\hat{y}_2 &= \bar{y}_5 + (\bar{y}_1 - \bar{y}_5)\cdot 0 + (\bar{y}_2 - \bar{y}_5)\cdot 1 + (\bar{y}_3 - \bar{y}_5)\cdot 0 + (\bar{y}_4 - \bar{y}_5)\cdot 0 = \bar{y}_2 \text{ (kandidaatit)},\\
\hat{y}_3 &= \bar{y}_5 + (\bar{y}_1 - \bar{y}_5)\cdot 0 + (\bar{y}_2 - \bar{y}_5)\cdot 0 + (\bar{y}_3 - \bar{y}_5)\cdot 1 + (\bar{y}_4 - \bar{y}_5)\cdot 0 = \bar{y}_3 \text{ (maisterit)},\\
\hat{y}_4 &= \bar{y}_5 + (\bar{y}_1 - \bar{y}_5)\cdot 0 + (\bar{y}_2 - \bar{y}_5)\cdot 0 + (\bar{y}_3 - \bar{y}_5)\cdot 0 + (\bar{y}_4 - \bar{y}_5)\cdot 1 = \bar{y}_4 \text{ (lisensiaatit) ja}\\
\hat{y}_5 &= \bar{y}_5 + (\bar{y}_1 - \bar{y}_5)\cdot 0 + (\bar{y}_2 - \bar{y}_5)\cdot 0 + (\bar{y}_3 - \bar{y}_5)\cdot 0 + (\bar{y}_4 - \bar{y}_5)\cdot 0 = \bar{y}_5 \text{ (tohtorit)}.
\end{aligned} \tag{3.6}$$

Huomaa, miten dummy-muuttujat ilmaisevat tasojen läsnäoloa. Vertailuryhmän havainnoilla kaikkien dummy-muuttujien arvot ovat nollia, jolloin jäljelle jää vain itse vertailutaso.

Visualisoidaan tilannetta jälleen kahdella kuvalla:



Kuvista on myös nähtävissä selvä riippuvuus koulutuksen ja palkan välillä, mikä osaltaan johtunee aineiston laatijan vankasta luottamuksesta koulutuksen hyödyllisyyteen...

Selittäjien merkitsevyyden testaus

Dummy-muuttujien kertoimiin liittyvät t-testit testaavat kunkin yksittäisen ryhmän eroa vertailuryhmään, mutta jos halutaan testata koko viisiluokkaisen selittäjän merkitsevyyttä mallissa, tarvitaan t-testiä yleisempi testausmenettely.

Kahden otoksen t-testin yleistys useammalle ryhmälle on yksisuuntainen varianssianalyysi. Siinä kokonaisvaihtelu hajotetaan kahteen osaan, ryhmien väliseen *(between)* ja sisäiseen *(within)* vaihteluun. Näitä vastaavien varianssien suhteesta saadaan F-testisuure hypoteesille H_0: $\mu_1 = \mu_2 = \ldots = \mu_q$, jossa q on ryhmien lukumäärä. Vastahypoteesina testissä on H_1: $\mu_i \neq \mu_j$ ainakin jollain i, j.

Kategoristen selittäjien testaus tapahtuukin F-testin avulla. Testaus noudattelee täsmälleen samoja linjoja, jotka on esitetty jo aiemmin kohdassa **Hierarkisten mallien vertailu**.

H_0:n mukaista mallia kutsutaan myös sidotuksi malliksi (arvot kiinnitetty esimerkiksi nolliksi) ja H_1:n mukaista vapaaksi malliksi (arvot estimoituvat aineiston perusteella). Sidotussa mallissa on vähemmän estimoitavia parametreja ja siten enemmän vapausasteita. Mikäli F-testisuuretta (1.46) vastaava p-arvo on riittävän pieni, niin dummy-muuttujien joukon kuvaamalla kategorisella selittäjällä on merkitystä mallissa.

Kutakin dummy-muuttujien joukkoa on syytä käsitellä kokonaisuutena, eli mallissa on pidettävä mukana joko kaikki ko. selittäjään liittyvät dummy-muuttujat tai ei yhtään. Yksittäisen dummy-muuttujan poisjättäminen tarkoittaa kyseisen luokan yhdistämistä vertailuryhmään, mikä ei useinkaan ole sisällöllisesti järkevää.

Seuraavassa on numeerisia tuloksia edellä esitetyistä tilanteista.

Vain sukupuoli selittäjänä (3.1):

```
Regression diagnostics on data PALKAT: N=60
Regressand Palkka    # of regressors=2 (Constant term included)
Condition number of scaled X: k=2.41421
Variable Regr.coeff. Std.dev.     t
Constant  3474.1452  157.61333     22.042
S2        0.3154704  222.89890     0.0014
Variance of regressand Palkka=732627.3385 df=59
Residual variance=745258.8186 df=58
R=0.0002 R^2=0.0000 Durbin-Watson=0.855
```

Tämän täysin fiktiivisen aineiston valossa naisten ja miesten palkoissa ei ole käytännössä lainkaan eroa.

Sama tarkastelu kahden riippumattoman otoksen t-testillä:

```
Independent samples           NAISET(Palkka)      MIEHET(Palkka)
Sample size                   30                  30
Mean                           3474.145            3474.461
Standard deviation             912.1604            811.4684
Student's t=-0.001 df=58 (P=0.4994 one-sided test)
Sum of ranks (R)              911                 919
Mann-Whitney (U)              446                 454
(P=0.4764 one-sided Mann-Whitney, normal approximation)
```

Vastaava 2-suuntaisen testin p-arvo on tietysti 2*0.4994=0.9988, mutta ei näin tasaisessa tilanteessa tarvita mitään tilastollista testiäkään; lopputulos näkyy suoraan keskiarvoista.

Vain koulutus selittäjänä, vertailuryhmänä tohtorit (3.4):

```
Regression diagnostics on data PALKAT: N=60
Regressand Palkka    # of regressors=5 (Constant term included)
Condition number of scaled X: k=4.23607
Variable Regr.coeff. Std.dev.     t
Constant  4494.9950  146.36526     30.711
K1       -1948.2118  206.99174    -9.4120
K2       -1521.5331  206.99174    -7.3507
K3       -1123.2783  206.99174    -5.4267
K4       -510.43711  206.99174    -2.4660
Variance of regressand Palkka=732627.3385 df=59
Residual variance=257073.4829 df=55
R=0.8203 R^2=0.6729 Durbin-Watson=1.988
```

Kaikki erot vertailuryhmään ovat merkitseviä. Jo kuvan perusteella on selvää, että jos lisensiaattien palkkaero tohtoreihin nähden on merkitsevä, niin kaikkien muidenkin on.

Sama tarkastelu yksisuuntaisella varianssianalyysilla:

```
Comparing independent samples:
Sample               N    Mean           Std.dev.
YO(Palkka)           12    2546.783       545.4778
KAND(Palkka)         12    2973.462       577.1296
MAIST(Palkka)        12    3371.717       422.0947
LIS(Palkka)          12    3984.558       429.1680
TOHT(Palkka)         12    4494.995       540.7343
Bartlett's test for equality of standard deviations=1.758
   P=0.7802 df=4 (Chi^2-approximation)
F test for equality of means=28.29
   P=0.0000 df1=4 df2=55
```

Eroja siis on, mutta tämä ei kerro tarkemmin missä (vrt. regressiomallin yleis-F-testi).

Yhdysvaikutukset eli interaktiot

Edellä käsitellyt mallit ovat yksinkertaistettuja, eivätkä tietenkään sellaisenaan riitä ilmiöiden kuvaamiseen tai selittämiseen. Hieman realistisempia malleja saadaan ottamalla useampia asian kannalta tärkeitä selittäjiä malliin yhtaikaa, kuten yleensäkin. Aina ei kuitenkaan riitä tutkia näitä ns. päävaikutuksia *(main effects)* vaan myös niiden väliset yhdysvaikutukset eli interaktiot *(interactions)* on voitava ottaa huomioon jossain laajuudessa.

Kahdella selittäjällä, x_1:llä ja x_2:lla, sanotaan olevan interaktio, mikäli x_1 riippuu selitettävästä muuttujasta eri tavoin riippuen x_2:n arvoista. Malliyhtälössä tämä ilmaistaan tulotermillä x_1x_2.

Tarkastellaan jälleen edellä esitettyä esimerkkiä, jossa palkkaa selitetään iällä, sukupuolella ja koulutuksella. Tilannetta voitaisiin mallintaa myös kovarianssianalyysilla *(analysis of covariance, ANCOVA)*. Siinä mielenkiinnon kohteena ovat pääasiassa luokittelevat muuttujat. Jatkuvat muuttujat, joita kutsutaan kovariaateiksi, ovat puolestaan haittatekijöitä tms., joiden vaikutus halutaan eliminoida.

Tällaisia malleja voidaan kuitenkin yhtä hyvin tutkia regressioanalyysilla, käyttämällä dummy-muuttujatekniikkaa. Tiivistetään aluksi mallia yhdistämällä koulutuksen luokat 2 ja 3 sekä 4 ja 5. Näin saadaan kolmiluokkainen muuttuja, joka kuvaa henkilön akateemisen tutkinnon tasoa (1=ei tutkintoa, 2=perustutkinto, 3=jatkotutkinto). Valitaan vertailuryhmäksi näistä ensimmäinen ja tehdään muille dummy-muuttujat T_2 ja T_3.

Tutkitaan malleja, joissa ovat selittäjinä ikä (jatkuva) ja tutkintotaso (3 luokkaa) ja selitettävänä palkka. Näiden avulla on mahdollista muodostaa useita malleja. Yksinkertaisimmat mallit saadaan ottamalla aiempaan tapaan huomioon vain molempien selittäjien päävaikutukset yksittäin. Tutkintotason osalta malli on tällöin

$$y = \beta_0 + \beta_{02}T_2 + \beta_{03}T_3 + \varepsilon\,, \tag{3.7}$$

jossa vain kertoimien alaindeksit on valittu merkinnällisistä syistä hieman toisin kuin aiemmin. Malli antaa joka tapauksessa tutkintotasojen mukaiset keskipalkat aivan samoin kuin aiemmin käsitelty malli (3.4). Iällä ei tietenkään ole mallissa (3.7) mitään merkitystä. Sille saadaan vastaavasti oma malli

$$y = \beta_0 + \beta_1 \mathrm{Ikä} + \varepsilon\,, \tag{3.8}$$

jossa puolestaan tutkintotasoa ei oteta lainkaan huomioon vaan iän ja palkan riippuvuus oletetaan samaksi tutkintotasosta riippumatta. Yksi regressiosuora kuvaa koko tilanteen. Sen sijaan mallissa

$$y = \beta_0 + \beta_{02}T_2 + \beta_{03}T_3 + \beta_1 \mathrm{Ikä} + \varepsilon\,, \tag{3.9}$$

jossa ovat molempien tekijöiden päävaikutukset, voi iän ja palkan riippuvuus vaihdella tutkintotasosta toiseen – kuitenkin niin, että tuo riippuvuus on rakenteeltaan samanlaista joka tasolla. Tutkintotason vaikutus ei riipu iästä vaan ne oletetaan toisistaan riippumattomiksi, ts. iällä ja tutkintotasolla ei ole yhdysvaikutusta. Tilanteen kuvaamiseen tarvitaan kolme regressiosuoraa, mutta ne ovat kaikki samansuuntaisia. Yhtä hyvin voitaisiin sovittaa jokaiselle tutkintotasolle oma malli (3.8), mutta kätevämpää on tietenkin mallintaa koko tilanne yhdellä kertaa jälkimmäisen mallin (3.9) avulla.

Erikoistapaus yhdysvaikutuksen läsnäolosta on sellainen, jossa ero muuttuu iän mukana, mutta luokittainen tutkintotasojen järjestys säilyy. Tätä kuvaisi malli

$$y = \beta_0 + \beta_1 \mathrm{Ikä} + \beta_{12}T_2\mathrm{Ikä} + \beta_{13}T_3\mathrm{Ikä} + \varepsilon\,, \tag{3.10}$$

jossa systemaattisen osan jälkimmäiset termit tarkoittavat tutkintotason ja iän välistä interaktiota. Interaktioparametreja on saman verran kuin tutkintotason dummy-muuttujia. Regressiosuoria ajatellen kyseessä on jälleen kolme suoraa, mutta nyt ne eivät enää ole samansuuntaisia vaan niillä on eri kulmakertoimet. Vakiotermi on kaikilla sama, joten ne lähtevät (aineiston rajoissa) samasta pisteestä, mutta erkanevat sitten interaktion vaikutuksesta.

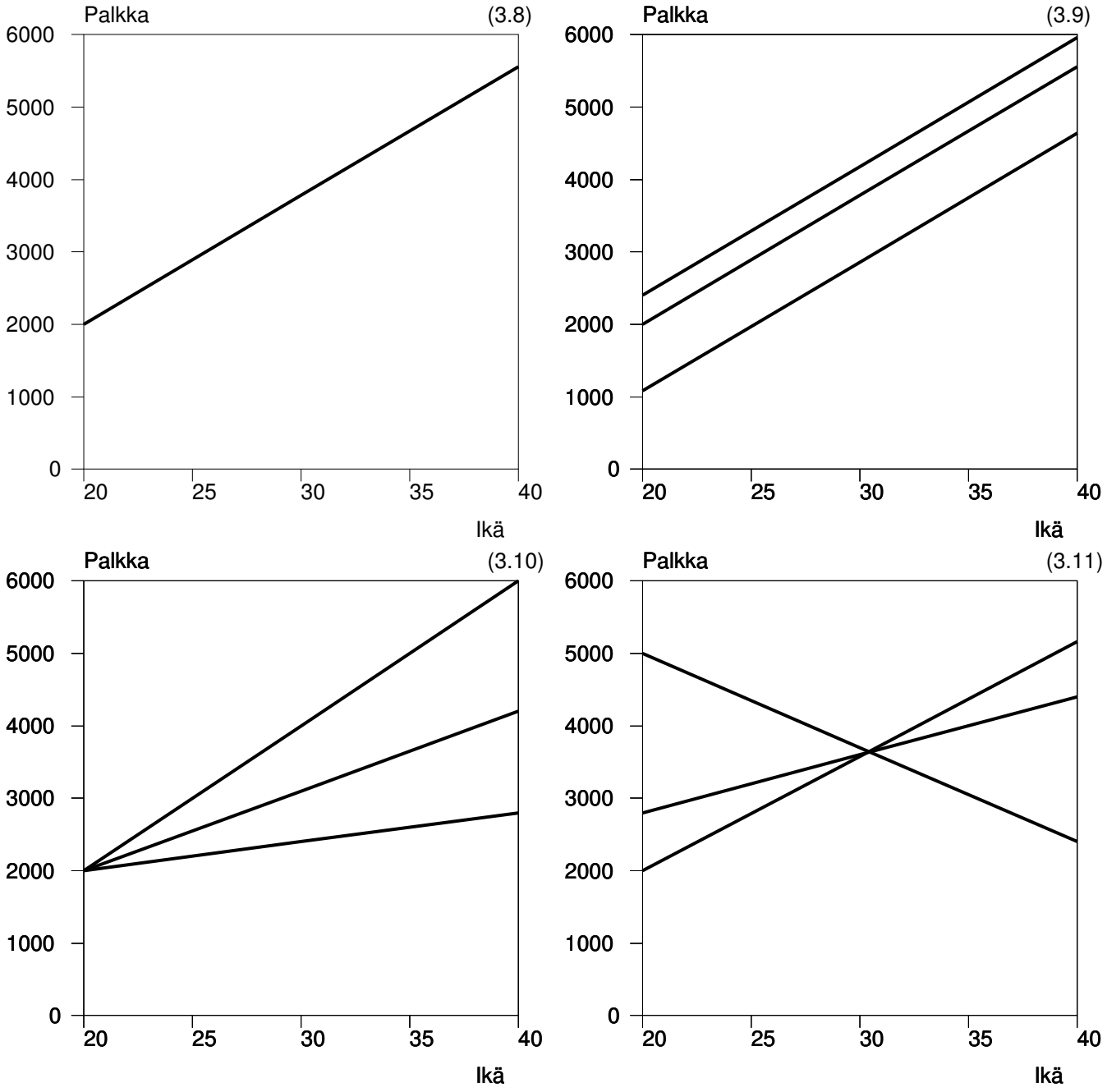
Interaktiomuuttujia kutsutaan toisinaan myös kulmakerroin-dummeiksi *(slope shifters)*, sillä ne kertovat kuinka paljon vertailuryhmään liittyvän mallin kulmakerrointa on muutettava, jotta saataisiin ao. ryhmään sovitetun mallin kulmakerroin. Mallin vakiotermiä samalla tavalla sääteleviä dummy-muuttujia kutsutaan vastaavasti taso-dummeiksi.

Lisäämällä edelliseen malliin (3.10) taso-dummyt saadaan yleisin malli

$$y = \beta_0 + \beta_{02}T_2 + \beta_{03}T_3 + \beta_1 \text{Ikä} + \beta_{12}T_2\text{Ikä} + \beta_{13}T_3\text{Ikä} + \varepsilon\,, \qquad (3.11)$$

jossa iän ja palkan riippuvuus voi vaihdella sekä rakenteeltaan että tasoltaan tutkintotasosta toiseen. Tällöin sanotaan, että iällä ja tutkintotasolla on interaktio. Kuvallisesti se ilmenee siten, että tutkintotason mukaisten ryhmien regressiosuorat leikkaavat toisensa yhdessä tai useammassa pisteessä.

Malleissa (3.8)–(3.11) on siis kysymys jatkuvan selittäjän (ikä) ja kategorisen selittäjän (tutkintotaso) päävaikutusten ja näiden mahdollisten yhdysvaikutusten ottamisesta huomioon eri tavoin. Malleja voidaan havainnollistaa graafisesti seuraavantyyppisillä kuvilla:



Kuvat esittävät selkeästi toisistaan poikkeavia tilanteita. Käytännössä erot voivat olla verraten pieniä, ja onkin hyödyllistä tutkia, onko esimerkiksi interaktiotermien sisällyttäminen malliin välttämätöntä. Tyypillisiä testaustilanteita ovat esim. seuraavat:

- Onko interaktiota vai ei, eli ovatko kulmakertoimet samat (3.9 vs. 3.11)?
- Ovatko interaktiomallien vakiot samat (3.10 vs. 3.11)?
- Ovatko vakiot samat, kun interaktiota ei ole (3.8 vs. 3.9)?

Testaaminen tapahtuu F-testillä (vrt. **Selittäjien merkitsevyyden testaus** edellä). Huomaa, että malleja (3.9) ja (3.10) ei voi verrata F-testillä, koska ko. mallit eivät ole sisäkkäisiä!

Seuraavassa on numeerisia esimerkkejä edellä esitetyistä malleista.

Vain ikä selittäjänä (3.8):

```
Regression diagnostics on data PALKAT: N=60
Regressand Palkka    # of regressors=2 (Constant term included)
Condition number of scaled X: k=20.2819
Variable Regr.coeff. Std.dev.     t
Constant -4995.1676  186.36517    -26.803
Ikä       292.38678  6.4025752     45.667
Variance of regressand Palkka=732627.3385 df=59
Residual variance=20165.80818 df=58
R=0.9864 R^2=0.9729 Durbin-Watson=1.290
```

Ikä ja tutkintotaso, vain päävaikutukset (3.9):

```
Regression diagnostics on data PALKAT: N=60
Regressand Palkka    # of regressors=4 (Constant term included)
Condition number of scaled X: k=32.8401
Variable Regr.coeff. Std.dev.     t
Constant -4138.9202  158.05749    -26.186
Ikä       255.50459  5.9514120     42.932
T2        146.73498  34.938155     4.1998
T3        383.53222  45.016263     8.5199
Variance of regressand Palkka=732627.3385 df=59
Residual variance=8769.229658 df=56
R=0.9943 R^2=0.9886 Durbin-Watson=1.488
```

Ikä ja sen interaktiot tutkintotason kanssa (3.10):

```
Regression diagnostics on data PALKAT: N=60
Regressand Palkka    # of regressors=4 (Constant term included)
Condition number of scaled X: k=35.855
Variable Regr.coeff. Std.dev.     t
Constant -3933.5009  164.56906    -23.902
Ikä       247.51305  6.3210784     39.157
IkäT2     5.9257631  1.2578276     4.7111
IkäT3     13.718379  1.5303990     8.9639
Variance of regressand Palkka=732627.3385 df=59
Residual variance=8181.358812 df=56
R=0.9947 R^2=0.9894 Durbin-Watson=1.497
```

Ikä ja tutkintotaso, päävaikutukset ja interaktiot (3.11):

```
Regression diagnostics on data PALKAT: N=60
Regressand Palkka    # of regressors=6 (Constant term included)
Condition number of scaled X: k=105.838
Variable Regr.coeff. Std.dev.     t
Constant -3437.8269  297.08754    -11.572
Ikä       228.71122  11.311295     20.220
T2       -646.81008  392.33318    -1.6486
T3       -778.27039  416.56381    -1.8683
IkäT2     30.090317  14.527014     2.0713
IkäT3     41.516441  14.652237     2.8335
Variance of regressand Palkka=732627.3385 df=59
Residual variance=7889.966499 df=54
R=0.9951 R^2=0.9901 Durbin-Watson=1.565
```

Luokittelevien selittäjien yhdysvaikutukset

Regressio- ym. mallien tulkinnan kannalta ratkaisevaa on muuttujien luokittelu. Erilaisissa tilanteissa tarvitaan erilaisia luokitteluja. Olisi tärkeää, ettei informaatiota hukattaisi liian karkeilla luokitteluilla, muttei toisaalta ajauduttaisi liian monimutkaisiinkaan malleihin.

Luokitellaan nyt esimerkkiaineiston havainnot vielä kerran uudelleen koulutuksen mukaan. Muodostetaan vain kaksi luokkaa sen perusteella, onko henkilöllä ylempi korkeakoulututkinto vai ei. Yhdistetään siis koulutusmuuttujan luokat 1 ja 2 sekä 3–5, ja tehdään suoraan dummy-muuttuja Yk (ei ole/on ylempi korkeakoulututkinto). Tämän lisäksi otetaan mukaan sukupuoli, vrt. malli (3.1) ja ikä (jatkuvana kuten edellä). Yleisin malli on nyt

$$y = \beta_0 + \beta_1 \text{Ikä} + \beta_2 S_2 + \beta_3 \text{Yk} + \beta_{12} \text{Ikä} S_2 + \beta_{13} \text{Ikä Yk} + \beta_{23} S_2 \text{Yk} + \beta_{123} \text{Ikä} S_2 \text{Yk} + \varepsilon , \qquad (3.12)$$

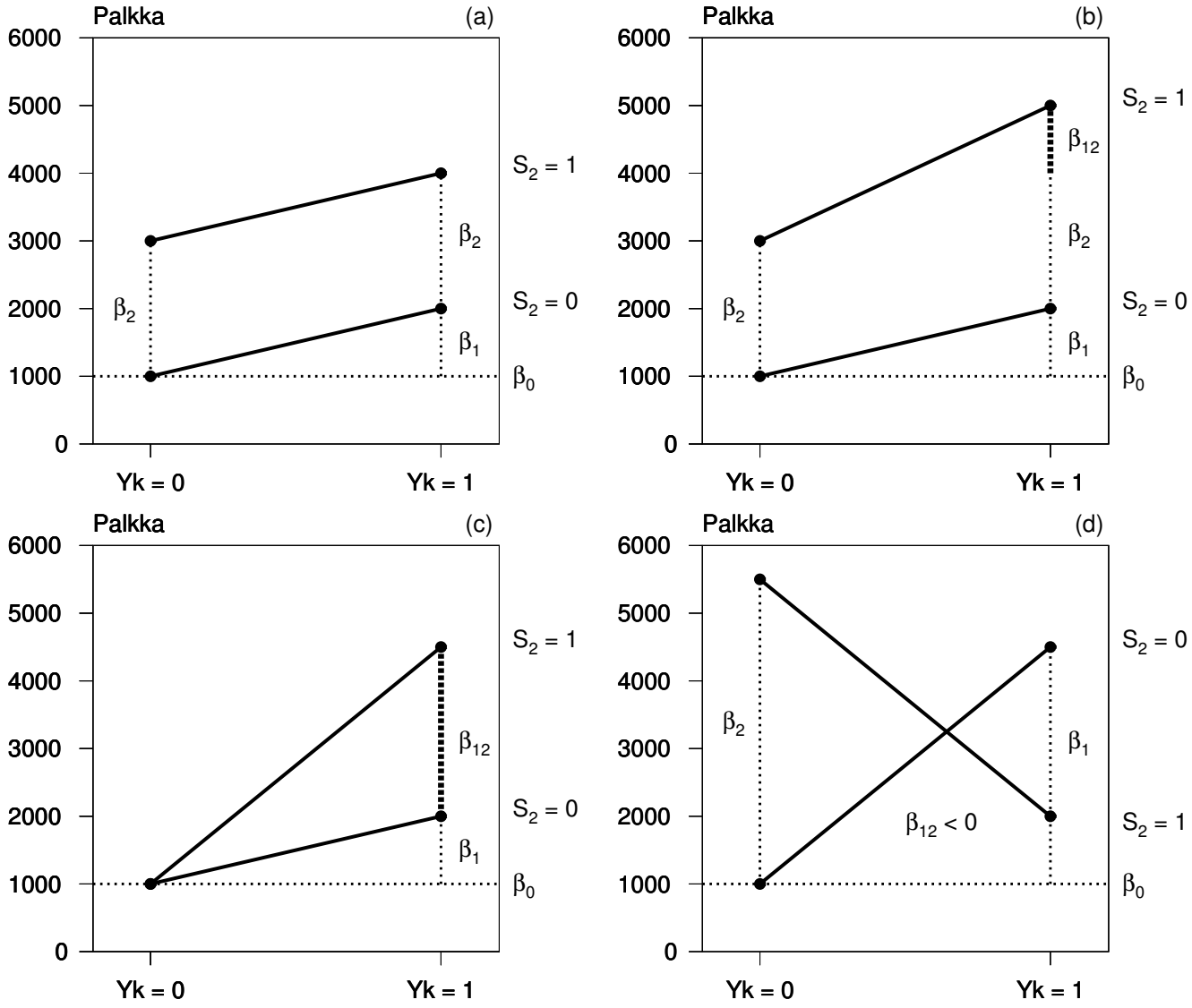
jossa systemaattisen osan viimeinen termi tarkoittaa kolmen selittäjän yhtäaikaista yhdysvaikutusta. Mallin parametreista puolet liittyy interaktiotermeihin. On selvää, ettei laajempien mallien yhteydessä ole mahdollista hallita kaikkia mahdollisia interaktioita, sillä parametrien määrä kasvaa äkkiä hyvin suureksi. Tutkittavan ilmiön tuntemus onkin tärkeää, jotta voi vähentää interaktiotarkastelujen määrää.

Interaktiotermejä sisältävien mallien testauksissa on huomattava, että päävaikutustekijöitä ei saa poistaa mallista, jos niitä sisältäviä interaktioita on mukana. Toisin sanoen testaus aloitetaan korkeimmantasoisista interaktiotermeistä ja edetään hierarkisesti kohti päävaikutuksia. Esimerkkimallin (3.12) tapauksessa testattaisiin siis ensin, onko iän, sukupuolen ja tutkinnon suorittamisen interaktio (IkäS_2Yk) merkitsevä. Testaukseen käy tässä t-testi, koska on kyse vain yhdestä parametrista. Yleisesti tarvitaan tietenkin F-testiä (vrt. edellä). Mikäli ylin interaktio ei ole merkitsevä, testataan alemman tason interaktiot (jossakin järjestyksessä). Usein toivotaan, että ylin interaktio ei olisi merkitsevä, koska jos se on, ei mallista voida poistaa mitään. Etenkin useammanasteisille interaktioille on myös tyypillisesti vaikea löytää sovellusalan kannalta mielekkäitä tulkintoja.

Tarkastellaan lopuksi pelkästään luokittelevia selittäjiä ja näiden yhdysvaikutuksia. Jättämällä äskeisestä mallista jatkuva selittäjä Ikä pois saadaan malli

$$y = \beta_0 + \beta_1 Yk + \beta_2 S_2 + \beta_{12} Yk S_2 + \varepsilon \, . \qquad (3.13)$$

Jos interaktiotermin (YkS_2) kerroin β_{12} on nolla, niin voidaan sanoa, että tutkinnon suorittaminen (Yk) ja sukupuoli (S_2) ovat *additiivisia*. Kertoimen testaaminen on mielekästä, koska additiivisia vaikutuksia on helpompi ymmärtää ja tulkita. Kuvallisesti esitettynä on kysymys seuraavanlaisista tilanteista:



Kaikki kuvissa (a)–(d) näkyvät viivat ovat vain apuviivoja. Paksummat yhtenäiset viivat kuvaavat keskiarvoprofiileja samaan tapaan kuin aikaisemminkin, kun taas katkoviivoilla on osoitettu mallin (3.13) kertoimien tehtävät. Kuvassa (a) vaikutukset ovat additiivisia, eli interaktiota ei ole. Muissa kuvissa vaikutukset ovat ei-additiivisia, eli tutkinnon suorittamisella ja sukupuolella on interaktio. Kuvassa (c) vaikutusta on vain Yk:n toisessa luokassa. Kuvan (d) ääritapauksessa interaktio on negatiivinen: tutkinnon suorittamisella on päinvastainen vaikutus palkkaan naisilla ja miehillä (aineisto on todellakin täysin keinotekoinen!).

Varianssianalyysi

Kurssin viimeisenä aiheena on varianssianalyysi *(analysis of variance, ANOVA)*, josta yhdessä koesuunnittelun *(experimental design)* kanssa saisi helposti kokonaisen kurssin. Tässä yhteydessä on tavoitteena tutustua varianssianalyysin peruskäsitteistöön ja koesuunnittelun periaatteisiin sekä tuoda esiin yhteyksiä regressio- ja varianssianalyysien välillä.

Varianssianalyysissa kaikki selittäjät ovat kategorisia eli luokittelevia muuttujia. Selitettävää muuttujaa koskevat samat oletukset kuin aikaisemminkin; kysehän on jatkuvasti lineaarisesta mallista (1.1). Menetelmän nimi tulee siitä, että johtopäätöksiä tehdään erilaisista varianssihajotelmista. Kysymyksessä on itse asiassa laaja valikoima menetelmiä, joista oikean mallin valinta määräytyy ensisijaisesti taustalla olevan koejärjestelyn perusteella.

Koesuunnittelu ja koejärjestelyt

Koesuunnittelun pyrkimyksenä on saada käytettävissä olevilla resursseilla tietoa tutkittavasta ilmiöstä mahdollisimman tehokkaasti. Koesuunnittelun ytimen muodostavat koeyksiköt *(experimental units)* ja käsittelyt *(treatments)*. Ideaalimaailmassa voitaisiin ajatella, että olisi olemassa loputtomasti homogeenisia koeyksiköitä, joiden avulla voitaisiin kaikista käsittelyjen yhdistelmistä mitata kiinnostavan muuttujan arvot ns. täydellisesti satunnaistetussa kokeessa *(completely randomized design)*. Edelleen ideaalimaailmassa faktoreiden lukumäärästä riippumatta voitaisiin toistaa mittaukset useita kertoja eri käsittelyjen yhdistelmille ja näin tarkentaa saatuja estimaatteja. Monimutkaisempiin koejärjestelyihin ei olisi tarvetta. Reaalimaailmassa asiat ovat toisin. Niinpä tehokkaat koejärjestelyt ovat tarpeen.

Onnistunut koesuunnittelu edellyttää tutkimusalan ja tutkittavan ilmiön vankkaa tuntemusta. Lisäksi vaaditaan tietoutta käytettävissä olevista menetelmistä, erityisesti varianssianalyysin eri muodoista. Tämän vuoksi kokeellisessa tutkimuksessa onkin erittäin tärkeää, että tilastotieteilijä osallistuu tutkimusprojekteihin jo koesuunnittelu- eikä vasta analyysivaiheessa.

Yksisuuntainen varianssianalyysi

Yksisuuntaisella varianssianalyysilla *(one-way ANOVA)* testataan kahden tai useamman ryhmän keskiarvojen yhtäsuuruuksia. Ryhmät muodostuvat yhden luokittelevan muuttujan eli faktorin perusteella. Joskus käytetäänkin nimitystä yhden faktorin varianssianalyysi. Malli voidaan kirjoittaa ns. vaikutusten mallina *(effects model)* muodossa

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,, \tag{3.14}$$

jossa y_{ij} on selitettävän muuttujan arvo ryhmässä i havainnolla j, μ on yleiskeskiarvo (populaation keskiarvo, jota estimoidaan otoksen perusteella), α_i on luokittelevan muuttujan i:nnen luokan vaikutus y:n keskiarvoon ja ε_{ij} on mallivirhe ryhmässä i havainnolla j. Testattavat hypoteesit ovat

$$\begin{aligned} &H_0:\ \alpha_i = 0 \ \text{kaikilla } i \text{ ja} \\ &H_1:\ \alpha_i \neq 0 \ \text{ainakin jollain } i\,. \end{aligned} \tag{3.15}$$

Toinen tapa formuloida malli on ns. keskiarvojen malli *(cell means model)*

$$y_{ij} = \mu_i + \varepsilon_{ij}\,, \tag{3.16}$$

jossa μ_i on i:nnen ryhmän keskiarvo. Nyt hypoteesit ovat

$$\begin{aligned} &H_0:\ \mu_1 = \mu_2 = \ldots = \mu_k \ \text{ja} \\ &H_1:\ \mu_h \neq \mu_i \ \text{ainakin joillain } h, i\,, \end{aligned} \tag{3.17}$$

jossa k on luokittelevan muuttujan luokkien eli vertailtavien ryhmien lukumäärä.

Yksisuuntaiseen varianssianalyysiin liittyvää koejärjestelyä kutsutaan täydellisesti satunnaistetuksi kokeeksi. Satunnaistaminen tarkoittaa, että koeyksiköt jaetaan eri käsittelyihin täysin satunnaisesti. Tilastollisten testien jakaumaoletusten kannalta satunnaistaminen on ensiarvoisen tärkeää. Ryhmiä, joita koesuunnittelussa kutsutaan usein käsittelyiksi, on k kpl ja havaintoja n kpl. Yhteensä koeyksiköitä on kn kpl. Ne voidaan esittää $k \times n$ -taulukkona, jonka alkioina ovat kuhunkin koeyksikköön liittyvät mittaukset. Lineaarisen mallin (1.1) kannalta mittaukset ovat selitettävän muuttujan y arvoja, ja mallimatriisi **X** (jota koesuunnittelun ja varianssianalyysin yhteydessä kutsutaan usein *design*-matriisiksi) koostuu käsittelyjä ilmaisevista indikaattorimuuttujista.

Tulokset esitetään varianssitauluna, joka on oleellisesti sama kuin regressioanalyysin yhteydessä esitetty taulukko (1.1). Nyt ryhmien välistä *(between)* vaihtelua verrataan niiden sisäiseen *(within)* vaihteluun ja saadaan näin F-testi hypoteesin (3.15) tai (3.17) testaamiseksi. Jos nollahypoteesi hylätään, eli ryhmäkeskiarvoissa on eroa, halutaan yleensä myös selvittää, missä erot tarkemmin sanottuna ovat. Tähän on useita mahdollisuuksia.

Kontrastit

Eräs tapa testata erilaisia osahypoteeseja perustuu ns. kontrasteihin *(contrasts)*. Keskiarvojen kontrasteilla tarkoitetaan niiden lineaarista yhdistelmää

$$c_1\mu_1 + c_2\mu_2 + ... + c_k\mu_k \,, \tag{3.18}$$

jossa kertoimien c_i summa on nolla. Esimerkiksi hypoteesia

$$H: \mu_1, \mu_2, \mu_3, \mu_4 = \mu_5, \mu_6 \tag{3.19}$$

voidaan testata kontrastilla

$$\frac{(\mu_1 + \mu_2 + \mu_3 + \mu_4)}{4} - \frac{(\mu_5 + \mu_6)}{2} \,. \tag{3.20}$$

Kontrastit ovat esimerkki tavallisia dummy-muuttujia tehokkaammasta koodaustavasta, jolla testattavat parametrit saadaan toisiinsa nähden ortogonaalisiksi. Näin niitä voidaan testata toisistaan riippumatta. Osaavissa käsissä kontrastit ovat joustava työkalu monenlaisten vertailujen suorittamiseen varsinkin ns. *a priori* -testaustilanteissa, joissa vertailtavat keskiarvoparit on päätetty jo ennalta. Koodaukset voidaan tällöin sisällyttää suoraan design-matriisiin. Toisin kuin F-testi, joka ottaa huomioon kaikki mahdolliset erot, *a priori* -tyyppinen testi keskittyy vain mielenkiinnon kohteena oleviin testauksiin.

Yhteisvertailumenetelmät

Yksittäisten erojen selvittäminen parittaisilla vertailuilla on joka tapauksessa ongelmallista. Jos jokaista keskiarvoparia vertaillaan erikseen esimerkiksi parittaisilla t-testeillä, kunkin yksittäisen vertailun merkitsevyystasoa kutsutaan vertailukohtaiseksi *(comparisonwise)* riskitasoksi. Menettely johtaa siihen, että ns. I lajin virheen (hylätään H_0 vaikka se on tosi) riski koko vertailussa on huomattavasti suurempi kuin vertailukohtainen riskitaso. Tätä vertailtaviin pareihin liittyvää kokonaisriskiä kutsutaan koekohtaiseksi *(experimentwise)* riskitasoksi. Sitä pyritään kontrolloimaan käyttämällä mahdollisimman hyvin tilanteeseen soveltuvia vertailumenetelmiä.

Edellä kuvattuun ongelmaan on esitetty lukuisia ratkaisukeinoja, joita kutsutaan yhteisvertailumenetelmiksi *(multiple comparison procedures)*. Esimerkkejä näistä ovat mm. *Tukeyn* ja *Scheffén* menetelmät. Molemmat niistä ovat luonteeltaan konservatiivisia eli niitä voi melko huoletta käyttää kaikissa mahdollisissa *post hoc* -testauksissa. Sopivan menetelmän valinta riippuu tilanteesta, mm. siitä ovatko ryhmät yhtäsuuret ja ovatko niiden varianssit samat. Tavoitteena useimmissa näistä menetelmistä on esittää sekä yhtaikaisia testejä että yhtaikaisia luottamusvälejä pitäen samalla koekohtainen riskitaso hallinnassa. Ongelmia voi silti tulla, on esimerkiksi mahdollista että F-testin mukaan eroja on, mutta parittaisten vertailujen perusteella niitä ei löydy.

Kaksisuuntainen varianssianalyysi

Kun malliin lisätään toinen tekijä, tullaan kaksisuuntaiseen varianssianalyysiin *(two-way ANOVA)*. Tarkastelut yleistyvät siitä edelleen useampisuuntaisiin malleihin, mutta idean ymmärtämiseksi riittää tarkastella eräisiin koejärjestelyihin liittyviä kaksisuuntaisia malleja.

Täydellisesti satunnaistetulla kokeella voidaan hallita vain käsittelyjen vaikutus (muu vaihtelu sisältyy mallivirheeseen). Jos koeyksiköt eivät ole homogeenisia, ne kannattaa ryhmitellä keskenään homogeenisiksi lohkoiksi *(blocks)*, jolloin osa virhevaihtelusta saadaan mallinnettua lohkojen välisenä vaihteluna. Näin saadaan satunnaistettujen lohkojen koe *(randomized block design)*, joka on itse asiassa parittaisten (riippuvien) otosten t-testin yleistys (vrt. yksisuuntainen varianssianalyysi ja tavallinen t-testi). Malli on muotoa

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \,, \tag{3.21}$$

jossa α_i kuvaa käsittelyjen *i* ja β_j lohkojen *j* vaikutuksia. Käsittelyillä ja lohkoilla ei ole yh-

dysvaikutusta. Jos sitä epäillään olevan, on käytettävä yleisempää koejärjestelyä, jossa se voidaan ottaa huomioon.

Hieman yleisempi koejärjestely on nimeltään latinalaisen neliön koe *(Latin squares design)*, jossa tehokkuutta on mahdollista parantaa järjestämällä koeyksiköt lohkotekijöiden mukaan neliön muotoon riveiksi ja sarakkeiksi siten, että jokainen käsittely esiintyy täsmälleen kerran kullakin rivillä ja sarakkeella. Jos käsittelyjä on esimerkiksi viisi, ja niitä merkitään (latinalaisilla) kirjaimilla A, B, C, D ja E, niin mahdollisia koejärjestelyjä ovat mm.

A	B	C	D	E			A	E	C	B	D			
B	C	D	E	A			C	B	A	D	E			
C	D	E	A	B		ja	E	A	D	C	B		.	
D	E	A	B	C			D	C	B	E	A			
E	A	B	C	D			B	D	E	A	C			

Latinalainen neliö tällaisenaan on melko rajallinen, koska yhdysvaikutuksia ei voi olla (kaikkia käsittelyjen yhdistelmiäkään ei esiinny) ja käsittelyitä on oltava yhtä monta riveittäin ja sarakkeittain. Toisaalta pienissäkin kokeissa voidaan ottaa huomioon useita tekijöitä.

Faktorikokeet

Yhdysvaikutusten tutkimiseksi on siirryttävä ns. faktorikokeisiin *(factorial design)*. Kun faktoreita (siis kategorisia eli luokittelevia muuttujia) on kaksi, vertailtavat ryhmät muodostuvat niiden luokkien yhdistelmistä. Samalla tulee mukaan myös ao. faktoreiden yhdysvaikutus eli interaktio aivan vastaavasti kuin aiemmin regressiomallin yhteydessä. Faktorikokeilla ei varsinaisesti tarkoiteta uutta koejärjestelyä, vaan niitä voidaan soveltaa täydellisesti satunnaistettuna, satunnaistettujen lohkojen tai latinalaisen neliön tyyppisinä. Myös erilaiset hierarkiset koejärjestelyt kuten osaruutukokeet *(split plot design)* voivat tulla kyseeseen.

Täydellisesti satunnaistetussa kokeessa kaksisuuntainen varianssianalyysimalli on muotoa

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}\,, \tag{3.22}$$

jossa α_i ja β_j kuvaavat faktoreiden vaikutuksia luokissa i, j ja γ_{ij} on näiden yhdysvaikutus vastaavissa luokissa. Keskiarvojen mallina tämä voidaan kirjoittaa muodossa

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}\,, \tag{3.23}$$

jossa μ_{ij} on ensimmäisen faktorin i:nnen luokan keskiarvo toisen faktorin j:nnessä luokassa. Toisistaan riippumattomia faktoreita sanotaan additiivisiksi. Tällöin yhdysvaikutukset ovat nollia (vrt. regressiomallin interaktiot ja profiilikuvat).

Kiinteät ja satunnaiset vaikutukset

Mallissa (3.22) α_i , β_j ja γ_{ij} voivat olla joko kiinteitä vakioita tai satunnaismuuttujia. Sen mukaan mallia kutsutaan joko kiinteiden vaikutusten *(fixed effects)* tai satunnaisvaikutusten *(random effects)* malliksi. Myös sekamallit *(mixed models)* ovat mahdollisia (ja yleisiä).

Vaikutusten ja näin ollen mallin tyyppi määräytyy aineiston sisällön, keruutavan ja käyttötarkoituksen mukaan. Kiinteiden vaikutusten mallissa vain mallivirhe ja selitettävä muuttuja ovat satunnaismuuttujia, eli malliyhtälön (3.22) molemmilla puolin on vain yksi satunnaismuuttuja. Kun siirrytään sekamalliin tai satunnaisvaikutusten malliin, yhtälön oikealle puolelle tulee muitakin satunnaismuuttujia. Selitettävän muuttujan varianssin sanotaan tällöin sisältävän varianssikomponentteja. Satunnaisvaikutusten mallit ja sekamallit ovatkin erikoistapauksia ns. varianssikomponenttimalleista *(variance component models)*.

Muita yleistyksiä

Toistettujen mittausten kokeissa *(repeated measures design)* samasta koeyksiköstä voi olla useita mittauksia esimerkiksi eri ajankohtina. Kovarianssianalyysissa on mukana jatkuvia muuttujia (vrt. regressioanalyysi luokittelevilla ja jatkuvilla selittäjillä). Moniulotteisessa varianssianalyysissa *(MANOVA)* selitettäviä muuttujia on samanaikaisesti useita. Tällöin voidaan myös soveltaa monimuuttujamenetelmiin lukeutuvaa erotteluanalyysia *(discriminant analysis)*. Kanonista analyysia *(canonical analysis)* voidaan puolestaan kutsua regressioanalyysin yleistykseksi usealle selitettävälle muuttujalle.

Seuraavassa on yksisuuntaisen varianssianalyysin tuloksia esimerkkiaineistolla:

Results for dependent variable Palkka:

Means and deviations

	yo	kand	maist	lis	toht	Total
Means	2546.78	2973.46	3371.72	3984.56	4495.00	3474.30
Deviations	545.48	577.13	422.09	429.17	540.73	855.94
N of obs.	12	12	12	12	12	60

Jackknife test for equality of variances: F value = 0.55161
It equals the 30.14% point of the F(4, 55) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.698631

Levene test for equality of variances: F value = 0.52362
It equals the 28.13% point of the F(4, 55) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.718747

One-way fixed effects analysis of variance

Source	sum of squares	df	mean squares
Between groups	29085971.4099	4	7271492.85248
Within groups	14139041.5622	55	257073.482949
Total	43225012.9721	59	

F test for equality of means: F-value = 28.2857
It equals the 100.00% point of the F(4, 55) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.000000

Test for equality of means without assuming equal group variances:
Brown-Forsythe statistic = 28.2857 with df 4 and 51.83
Appr. risk of rejecting the null hypothesis, when true, is 0.000000

t tests for pairs of means

		yo	kand	maist	lis
kand	pooled t	1.861			
	df	22			
	prob.	0.076			
maist	pooled t	4.143	1.929		
	df	22	22		
	prob.	0.000	0.067		
lis	pooled t	7.176	4.870	3.527	
	df	22	22	22	
	prob.	0.000	0.000	0.002	
toht	pooled t	8.787	6.665	5.672	2.561
	df	22	22	22	22
	prob.	0.000	0.000	0.000	0.018

Multiple comparisons of means by the Scheffe's method

Pairwise mean differences

Degrees of freedoms for each test statistic are 4 and 55

For each test statistic the lowest experimentwise significance level
at which the null hypothesis can be rejected are given

The mean of the first set of group means will be compared with
the mean of the second set of group means

First set	Second set	contrast	test st.	prob.
yo	kand	-426.679	1.06227	0.3840
yo	maist	-824.933	3.97074	0.0067
yo	lis	-1437.77	12.0619	0.0000
yo	toht	-1948.21	22.1466	0.0000
kand	maist	-398.255	0.92546	0.4559
kand	lis	-1011.10	5.96511	0.0005
kand	toht	-1521.53	13.5082	0.0000
maist	lis	-612.841	2.19144	0.0819
maist	toht	-1123.28	7.36222	0.0001
lis	toht	-510.437	1.52026	0.2090

Vertaa näitä aiempiin regressioanalyysin yhteydessä esitettyihin tuloksiin. Huomaa myös, miten herkästi parittaiset t-testit havaitsevat merkitseviä eroja parempiin yhteisvertailumenetelmiin (tässä *Scheffén* menetelmään) verrattuna.

Kaksisuuntainen varianssianalyysi samasta aineistosta (koulutus ja sukupuoli):

Results for dependent variable Palkka:

Means and deviations

Column variable: Koulutus
Row variable: Sukup

		yo	kand	maist	lis	toht	Total
nainen	Means	2463.64	2936.74	3207.32	4201.33	4561.69	3474.15
	Deviations	518.97	646.01	331.51	195.60	528.97	912.16
	N of obs.	6	6	6	6	6	30
mies	Means	2629.92	3010.18	3536.12	3767.78	4428.30	3474.46
	Deviations	607.20	558.76	466.04	504.15	593.96	811.47
	N of obs.	6	6	6	6	6	30
Total	Means	2546.78	2973.46	3371.72	3984.56	4494.00	3474.30
	Deviations	545.48	577.13	422.09	429.17	540.73	855.94
	N of obs.	12	12	12	12	12	60

Analysis of variance for fitting factors and interactions

Source	Sum of squares	Df	Mean square
K	29085971	4	7271492.9
S	1.4928234	1	1.4928234
KS	1040729.0	4	260182.25
ERROR	13098311	50	261966.22

Variance components (or their counterparts in fixed eff. models):
K 584127 S -8732.2 KS -297.33 ERROR 261966
WARNING: Negative estimates of variance components

Exact F test for H(K) is MS(K)/MS(ERROR): F value = 27.7574
It equals the 100.00% point of the F(4, 50) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.0000

Exact F test for H(S) is MS(S)/MS(ERROR): F value = 0.00001
It equals the 0.190% point of the F(1, 50) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.9981

Exact F test for H(KS) is MS(KS)/MS(ERROR): F value = 0.99319
It equals the 58.00% point of the F(4, 50) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.4200

Koulutuksen (K) ja sukupuolen (S) interaktio ei ole merkitsevä (p=0.42), ei myöskään sukupuolen päävaikutus. Näin ollen edellä esitetty yksisuuntainen analyysi riittää.

Lopuksi kovarianssianalyysi, jossa kovariaattina on ikä (vrt. interaktiomallit regressioanalyysin yhteydessä):

Results for dependent variable Palkka:

Estimates of effects in analysis of variance model
corresponding to the given constraints

Koulutus	Sukup	estimate of the effect
yo		-927.5197
kand		-500.8411
maist		-102.5862
lis		510.2550
toht		1020.6921
	nainen	-0.1577
	mies	0.1577
yo	nainen	-82.9813
kand	nainen	-36.5602
maist	nainen	-164.2433
lis	nainen	216.9327
toht	nainen	66.8521
yo	mies	82.9813
kand	mies	36.5602
maist	mies	164.2433
lis	mies	-216.9327
toht	mies	-66.8521
general mean		3474.3029

(Analyysi jatkuu seuraavalla sivulla...)

Analysis of variance for fitting factors and interactions

Source	Sum of squares	Df	Mean square
Factors and interactions	30126701.9202	9	3347411.32447
Residual	13098311.0519	50	261966.221039
Total	43225012.9721	59	

F-test for the analysis of variance model: F-value is 12.77803
It equals the 100.00000% point of the F(9, 50) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.000000

Estimates of effects in analysis of covariance model
corresponding to the given constraints

Koulutus	Sukup	estimate of the effect
yo		325.3955
kand		-413.7062
maist		99.6623
lis		690.7386
toht		-702.0901
	nainen	16.7783
	mies	-16.7783
yo	nainen	-23.8328
kand	nainen	33.0300
maist	nainen	3.1056
lis	nainen	7.5085
toht	nainen	-19.8113
yo	mies	23.8328
kand	mies	-33.0300
maist	mies	-3.1056
lis	mies	-7.5085
toht	mies	19.8113
general mean		-3751.2488

Multiple correlation squared for the covariance model 0.9951032

Analysis of variance for the covariance model

Source	Sum of squares	Df	Mean square
Factors, interactions and covariates	43013347.8294	14	3072381.98781
Residual	211665.142715	45	4703.66983810
Total	43225012.9721	59	

F-test for the analysis of covariance model: F-value is 653.18828
It equals the 100.00000% point of the F(14, 45) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.000000

Within-class regression coefficients

corresponding to the levels yo of factor K

Variable	Regression coefficient	standard deviation	t-value	loss in MCS if vble deleted	prob.
Ikä	228.253625	8.829574	25.85103	0.072721	0.00000

corresponding to the levels kand of factor K

Variable	Regression coefficient	standard deviation	t-value	loss in MCS if vble deleted	prob.
Ikä	259.578794	9.361945	27.72701	0.083658	0.00000

corresponding to the levels maist of factor K

Variable	Regression coefficient	standard deviation	t-value	loss in MCS if vble deleted	prob.
Ikä	245.713233	13.407389	18.32670	0.036549	0.00000

corresponding to the levels lis of factor K

Variable	Regression coefficient	standard deviation	t-value	loss in MCS if vble deleted	prob.
Ikä	230.985840	13.281100	17.39207	0.032916	0.00000

corresponding to the levels toht of factor K

Variable	Regression coefficient	standard deviation	t-value	loss in MCS if vble deleted	prob.
Ikä	278.909110	10.821452	25.77372	0.072286	0.00000

Test for equality of within-class regression coefficients

Loss in MCS	Mean square	Df		F value	prob.
0.001777	19207.5	4	45	4.08352	0.006586

Tests for hypotheses among effects

Source under the Null hypotheses	loss in MCS	Df		Mean square	F value	Prob.
KS & equal slopes	0.00216	8	45	11660.202	2.4790	0.0255
S	0.00037	1	45	16048.733	3.4120	0.0713
K	0.01948	4	45	210557.02	44.764	0.0000

Lähteet

Cook, D. R., & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* John Wiley & Sons, New York.

Faraway, J. J. (2002). *Practical Regression and Anova using R.* (ps/pdf, 212 p.) http://www.stat.lsa.umich.edu/~faraway/book/

Mustonen, S. (1995). *Tilastolliset monimuuttujamenetelmät.* Survo Systems, Helsinki.

Patovaara, T. (1991). *Vektori- ja matriisilaskenta.* Helsingin yliopisto, tilastotieteen laitos.

Puntanen, S. (1999a). *Regressioanalyysi I.* Matematiikan, tilastotieteen ja filosofian laitos, Tampereen yliopisto, B48.

Puntanen, S. (1999b). *Regressioanalyysi II.* Matematiikan, tilastotieteen ja filosofian laitos, Tampereen yliopisto, B49.

Puranen, J. (1997). *Data-analyysi.* Luentomoniste, Helsingin yliopisto, tilastotieteen laitos.

Puranen, J., Virtanen, M., Lahdenkari, M., Hyhkö, H., & Vehkalahti, K. (2001). *Data-analyysi I - Survo-kurssi.* Harjoitusmoniste (pdf, 32 s.), http://www.helsinki.fi/%7ekvehkala/da1/moniste.pdf Helsingin yliopisto, tilastotieteen laitos.

Saikkonen, P. (2002). *Tilastollinen päättely ja lineaariset mallit.* Luentomoniste, Helsingin yliopisto, tilastotieteen laitos.

Sund, R. (2001). *Minimum Description Length based model selection in linear regression.* http://www.helsinki.fi/~sund/pdf/sund_mdl.pdf Department of Statistics, University of Helsinki.

Lisää kirjallisuutta, linkkejä ohjelmistojen sivuille yms. löytyy kurssin kotisivulta.