RESEARCH REPORT

DIGRESSIOANALYYSI
Heterogeenisen havaintoaineiston
sovittaminen
vaihtoehtoisiin regressiomalleihin

Seppo Mustonen

No. 2 tammikuu 1976

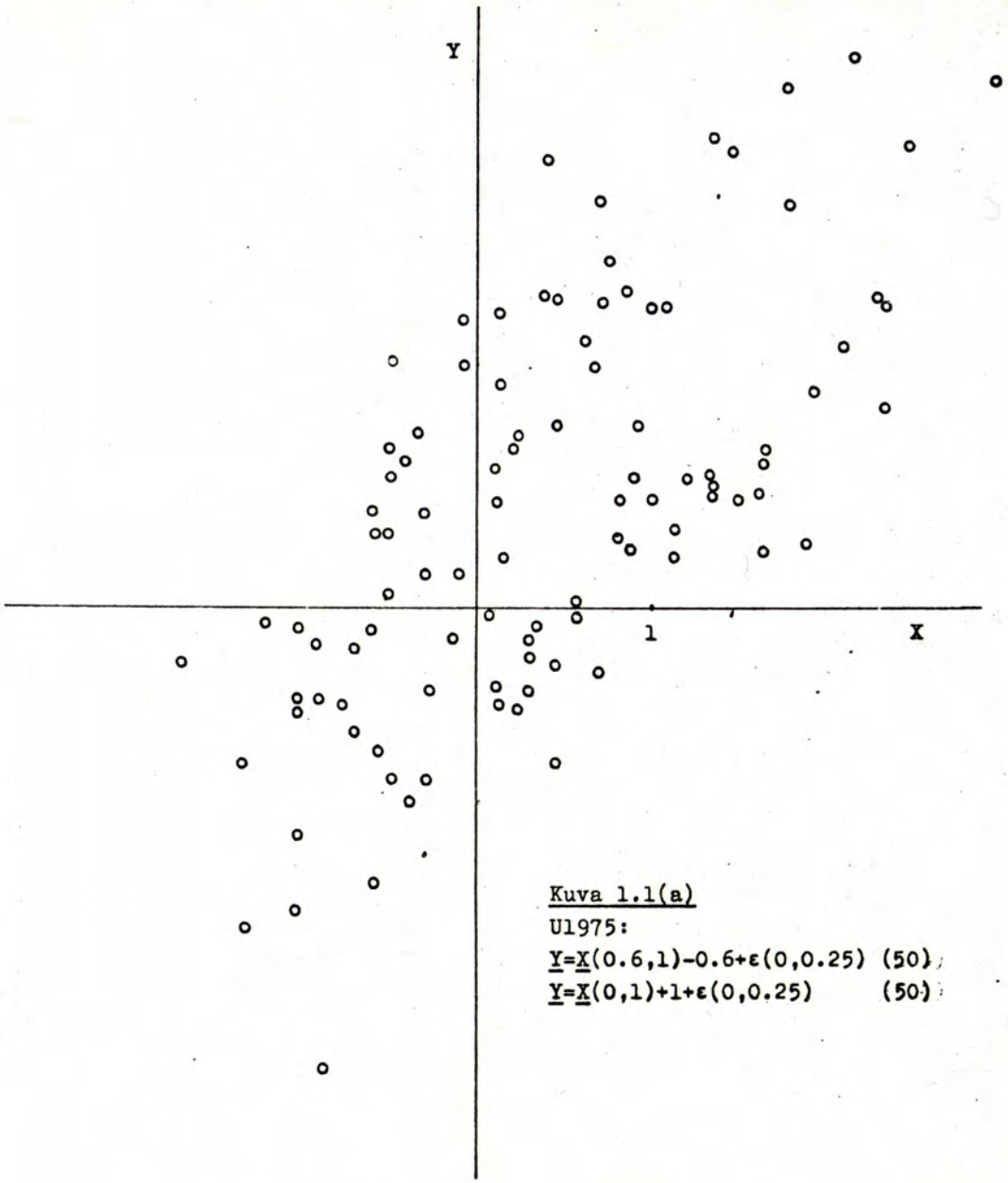
DEPARTMENT OF STATISTICS
UNIVERSITY OF HELSINKI
SF 00100 HELSINKI 10 FINLAND

1. Johdanto

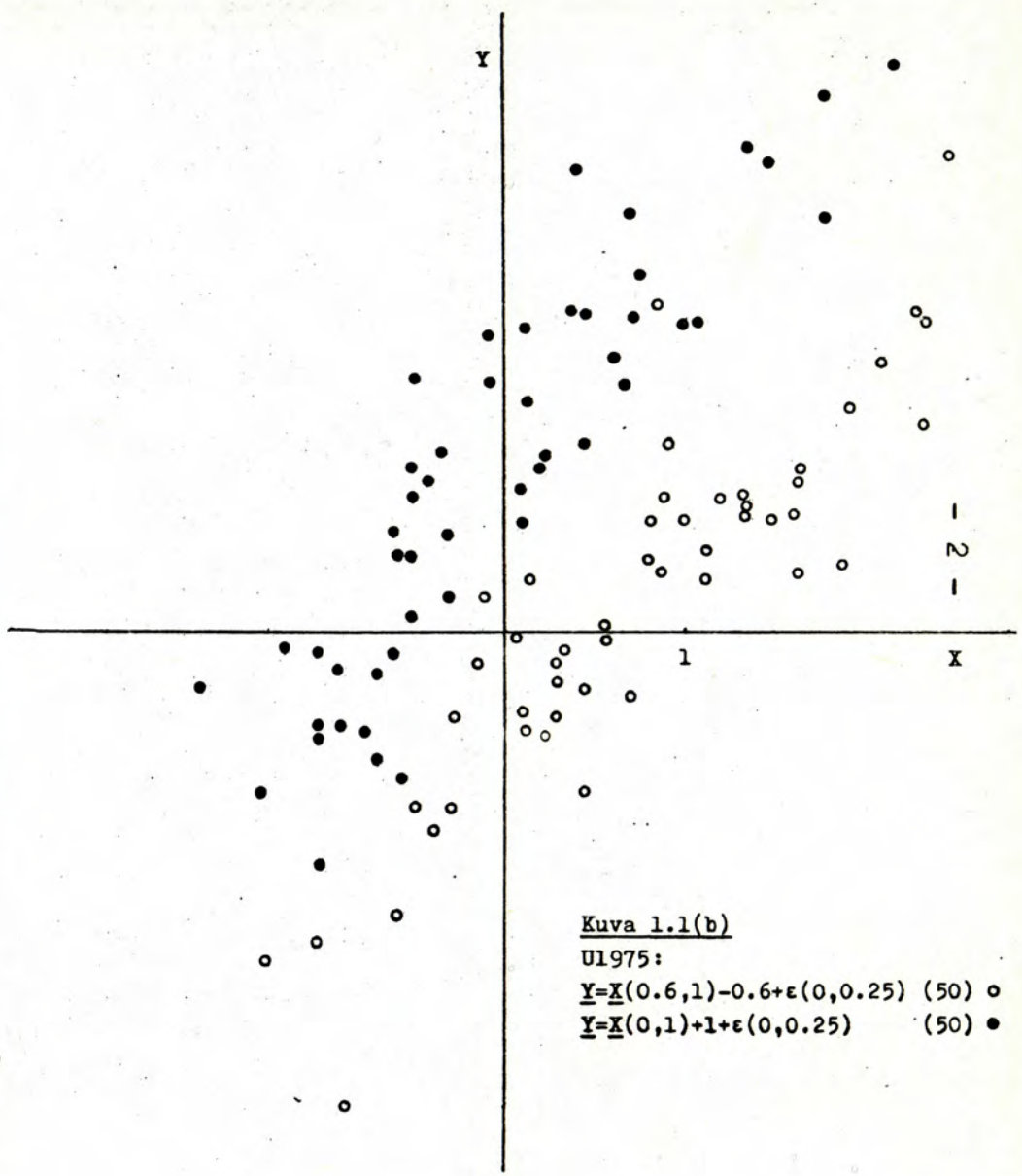
Heterogeenisen aineiston ilmaantuminen empiiriseen tutkimukseen, jossa tutkija odottaa saavansa yhteisjakaumaltaan kauniin, yksihuippuisen otoksen, ei yleensä ole ilahduttava ilmiö. Onhan havaintoarvoihin vaikuttaneet tällöin jotkin ennaltaodottamattomat ja siten mittaamatta jääneet tekijät. Jos nuo häiriötekijät voidaan jälkikäteen selvittää ja arvioida niiden vaikutukset eri havaintoihin, heterogeenisuus voidaan ehkä poistaa tai ainakin sen haitallisimpia vaikutuksia saadaan vähennetyksi. Jos sensijaan häiriötekijöitä ei pystytä jälkikäteen rekisteröimään havaintokohtaisesti, heterogeenisuus istuu ja pysyy aineistossa.

Tässä tutkimuksessa heterogeenisuuden ongelmaa tullaan pohtimaan tavanomaiseen lineaariseen tai epälineaariseen regressioanalyysiin kytkettynä. Tarkoituksena on esittää menetelmä, joka tekee mahdolliseksi regressiomallien käsittelyn heterogeenisuudesta huolimatta.

Esimerkkinä kuvitelkaamme ambivalenttia tilannetta vaikkapa sellaisessa kahden muuttujan x,y aineiston analysoinnissa, jossa tutkija haluaa sovittaa aineistonsa tavalliseen lineaariseen malliin $y=\alpha x+\beta+\varepsilon$, missä α ja β ovat mallin parametrit ja ε virhetermi. Todellisuudessa aineistoon on pujahtanutkin mukaan joukko poikkeuksellisia havaintoja, jotka noudattavat samaa mallia samalla α:n arvolla mutta eri β:n arvolla, olkoon se γ. Näillä havainnoilla on siis erilainen "lähtötaso". Kuvassa 1.1 on esitetty tällainen heterogeeninen 100 havainnon aineisto, jossa $\alpha=1$, $\beta=-0.6$, $\gamma=1$ ja $\varepsilon\sim N(0,0.25)$. Kumpaakin lajia on 50 havaintoa. (Liitteessä on selitetty tässä tutkimuksessa käytettyjen aineistojen alkuperä. Aineistot ovat kaikki pseudosatunnaislukujen avulla generoituja ja ne on tunnustensa sekä liitteessä esitettyjen tietojen avulla helppo luoda uudelleen millä tahansa tietojenkäsittelylaitteistolla.)



Kuva 1.1(a)
U1975:
$\underline{Y}=\underline{X}(0.6,1)-0.6+\varepsilon(0,0.25)$ (50) o
$\underline{Y}=\underline{X}(0,1)+1+\varepsilon(0,0.25)$ (50) •



Kuva 1.1(b)
U1975:
$\underline{Y}=\underline{X}(0.6,1)-0.6+\varepsilon(0,0.25)$ (50) o
$\underline{Y}=\underline{X}(0,1)+1+\varepsilon(0,0.25)$ (50) •

Kuvitteleminen, että aineisto olisikin homogeeninen, on kohtalokasta tärkeän α-parametrin estimoinnin kannalta. Jos aineisto sovitetaan tavallisella pns-keinolla lineaariseen malliin $y=\alpha' x+\beta'+\varepsilon$, saadaan estimaatit a'=0.736, b'=0.240.
Nyt tarkastelun kohteeksi tuleva menetelmä, josta käytän nimeä digressioanalyysi*, antaa vailla tietoa minkään havainnon todellisesta alkuperästä mutta varauduttaessa edellä mainittua tyyppiä olevaan heterogeenisuuteen parametreille estimaatit a=0.981, b=-0.632 ja c=0.968. (Estimaatteja merkitään tässä aina vastaavilla tavallisilla kirjaimilla.) Havainnot tulevat samalla analyysin yhteydessä jaetuksi kahteen ryhmään, jotka tässä tapauksessa vastaavat varsin tarkkaan havaintojen alkuperäistä jakoa. Kohdassa 9 tätäkin koetta kuvaillaan tarkemmin.

2. Digressioanalyysin periaate

Digressioanalyysissa yhdistyvät regressioanalyysin ja ryhmittelyanalyysin tehtävät ja ominaisuudet. Analyysin kohteena on muuttujien $y, x=(x_1,x_2,\ldots,x_m)$ n havainnon aineisto

$$(y_j,x^{(j)})=(y_j,x_{1j},x_{2j},\ldots,x_{mj}),\ j=1,2,\ldots,n,$$

jossa havainnot jakautuvat tuntemattomalla tavalla kahteen ryhmään G_1 ja G_2 (rajoittuminen kahteen ryhmään ei ole olennaista) siten, että ryhmässä G_i, i=1,2 havainnot noudattavat mallia

$$y=f_i(x,\alpha^{(i)})+\varepsilon_i,\quad E(\varepsilon_i)=0, D^2(\varepsilon_i)=\sigma_i^2,$$

missä $\alpha^{(i)}$ on mallin parametrivektori. Tehtävänä on estimoida kummankin mallin parametrit samanaikaisesti ilman minkäänlaista a priori-tietoutta siitä, kumpaan ryhmään G_1, G_2 mikin havainto kuuluu. Pyrkimyksenä on siis löytää jokaiselle havainnolle "oikea" ryhmä ja saadun luokittelun pohjalta estimoida mallien parametrit.

---

*digression=poikkeaminen tai tässä di(re)gressio=kaksoisregressio

Tavallisin menettely tämäntapaisissa tilanteissa lienee se, että havainnot luokitellaan aluksi jollain klusterointimenetelmällä ja parametrit estimoidaan sitten tavanomaisin regressioanalyysin keinoin. Tällöin nämä osa-analyysit toimivat toisistaan riippumatta eivätkä ilmeisesti voi käyttää kaikkea informaatiota tehokkaasti hyväksi.
Samanaikaista luokittelua ja estimointia on esiintynyt ainakin ekonometrisessa tutkimuksessa tarkasteltaessa tuntemattomaan ajankohtaan sijoittuneita epäjatkuvia parametrimuutoksia aikasarjoissa. Tietooni tulleissa tämäntyyppisissä tutkimuksissa (Goldfeld,Quandt, 1972) on kuitenkin havaintojen luokittelussa aina käytetty lisäinformaatiota aikamuuttujan tai muun vastaavanlaisen "tukimuuttujan" muodossa, mikä helpottaa olennaisesti tehtävää.
Ongelmalle on sukua myös mallinvalintatehtävä, jossa vain toinen malleista on oikea eikä mitään havaintokohtaista "digressiota" sallita (kts. esim. Hill,Hunter,Wichern, 1968). Samoin on tutkittu "toisiinsa liittyvien populaatioiden estimointia" (esim. Bradley,Gart, 1962), mutta siinä taas työtä keventää se, että eri osapopulaatioiden havainnot on tunnistettavissa ennen analyysia.

Digressioanalyysille on ominaista, että havaintojen luokittelu ja parametrien estimointi tapahtuvat samanaikaisesti vailla minkäänlaista lisätietoa yksittäisten havaintojen luonteesta. Digressioanalyysin periaate lienee sovellettavissa useihin estimointikeinoihin. Tavallista pienimmän neliösumman menetelmää käytettäessä parametrit määrätään yleistämällä pns-kriteeri

$$S(\alpha)=\sum_{j=1}^{n}(y_j-f(x^{(j)},\alpha))^2 = \min_{\alpha}$$

"valikoivaan" muotoon

$$(2.1)\quad S(\alpha^{(1)},\alpha^{(2)})=\sum_{j=1}^{n}\min\left[(y_j-f_1(x^{(j)},\alpha^{(1)}))^2,(y_j-f_2(x^{(j)},\alpha^{(2)}))^2\right]$$

$$= \min_{\alpha^{(1)},\alpha^{(2)}} .$$

Siis jokainen havainto liitetään estimoinnin yhteydessä "lähimpään" regressiokäyrään ja kummankin osamallin f_1,f_2 parametrit arvioidaan vain omien "lähipisteiden" avulla. Tässä suhteessa on siten kysymys erään tunnetun luokittelusäännön (nearest-mean classification rule, kts. esim. Fukunaga, 1972, s.332) soveltamisesta yleistetyssä muodossa. Itse estimointitapaa voitaisiin kutsua valikoivaksi pns-keinoksi.

Tuntunee aluksi oudolta, että tällainen valikoiva kriteeri saattaa toimia käytännössä, vaikka osa-aineistot eivät ole erillään, vaan peittävät osittain toisensa. Tällöinhän syntyy vakavia luokitteluvirheitä, esim. 20% havainnoista luokitellaan väärin, mikä voi häiritä parametrien estimointia. On kuitenkin pantava merkille, että virheluokitukset tulevat kohdistumaan etupäässä havaintoihin, jotka ovat osamallien "yhteisalueella" ja ne ovat estimoinnin kannalta usein melko neutraaleja. "Luokituksessa on tapahtunut virhe, mutta havainto sopii hyvin väärään malliin."

Siirtyminen tavallisesta pns-keinosta valikoivaan pns-keinoon merkitsee lineaaristenkin osamallien tapauksessa laskentatyön olennaista lisääntymistä, sillä valikoivan pns-kriteerin minimointi on aina vaativa epälineaarinen optimointitehtävä, jonka ratkaisemisessa on turvauduttava iteratiivisiin keinoihin. Samoin digressiomalleja lienee kriteerin luonteesta johtuen sangen hankala tutkia teoreettisesti. Niinpä tässä tutkimuksessa tyydytään suurelta osalta eräänlaiseen puolikokeelliseen tarkastelutapaan.

3. Mallityyppejä

Edellä määritellyn digressiomallin rakenne kuvataan merkitsemällä

$$y=\begin{cases} f_1(x,\alpha^{(1)})+\varepsilon_1 \\ f_2(x,\alpha^{(2)})+\varepsilon_2. \end{cases}$$

Tämä asetelma ja vastaava valikoiva pns-kriteeri voidaan luonnollisesti laajentaa kahta useampaa vaihtoehtoista osamallia koskevaksi. Kuten jo johdantoesimerkistä nähtiin, osamallien parametrit voivat olla osittain samoja. Useat sovellutuskohteet ovat ilmeisesti juuri sellaisia, joissa osamallit ovat muodoltaan melko samanlaiset ja eroavat toisistaan ehkä vain yhden tai muutaman parametrin ja muuttujan suhteen. Esimerkiksi digressiomallissa

$$y=\begin{cases} \alpha_1x_1+\alpha_2x_2+\ldots+\alpha_mx_m+\beta+\varepsilon_1 \\ \alpha_1x_1+\alpha_2x_2+\ldots+\alpha_mx_m+\gamma+\varepsilon_2 \end{cases}$$

vakiotermin ambivalenssi merkitsee itse asiassa tavanomaisen dummy-muuttujan käytön yleistystä muotoon, jossa etukäteen ei tarvitse tietää, kumpi tuntemattomista arvoista β,γ kuuluu millekin havainnolle. Erikoistapaus

$$y=\begin{cases} \beta+\varepsilon_1 \\ \gamma+\varepsilon_2 \end{cases}$$

tarkoittaa pelkkää yhden muuttujan y heterogeenisen jakauman osajakaumien erottamista. Tätä kuvaillaan tarkemmin normaalijakauman osalta kohdassa 5.

Mallin

$$y=\begin{cases} \alpha_1x_1+\alpha_2x_2+\alpha_3+\varepsilon_1 \\ \alpha_1x_2+\alpha_2x_1+\alpha_3+\varepsilon_2 \end{cases}$$

voi tulkita esim. siten, että osassa havaintoja selittävien muuttujien x_1 ja x_2 roolit ovat vaihtuneet. Malli siis pyrkii paljastamaan tuollaiset "virheelliset" havainnot ja käyttämään niitäkin tehokkaasti hyväksi parametreja estimoitaessa. Tästä näkyy, että digressiomalleja saatetaan ehkä soveltaa virhehavaintojen (outliers) etsimiseen ja mikäli virheet ovat esim. edellä kuvatulla tavalla systemaattisia, pystytään ehkä tylyn poissulkemisen asemasta jopa käyttämään niitä hyväksi.

Malli

$$y=\begin{cases}\alpha x_1+\beta+\varepsilon_1\\ \alpha x_2+\beta+\varepsilon_2\\ \alpha x_3+\beta+\varepsilon_3\end{cases}$$

kuvastaa tilannetta, jossa selitettävään muuttujaan y vaikuttaa kussakin havainnossa vain yksi kolmesta toisensa poissulkevasta "impulssista" x_1, x_2, x_3. On vain jäänyt epäselväksi, mikä impulsseista on ollut "aktiivinen" kunkin havainnon kohdalla.

Tässä hahmotellut esimerkinomaiset ja pelkistetyt mallityypit eivät saa johtaa harhaan siinä suhteessa, että digressioanalyysi olisi jokin yleiskeino hankalien heterogeenisten aineistojen selvittämiseen. On aivan selvää, ettei tätä menetelmää voida käyttää pelkästään kokeiluluontoisesti vaihdellen summittaisesti osamallien rakennetta, vaan mallissa esiintyvien vaihtoehtojen tulee perustua vankkaan teoreettiseen ja kokemusperäiseen tietoon havaintojen mahdollisista syntytavoista. Ainoa, mutta samalla erittäin merkittävä epäselvyys, joka sallitaan normaalin havaintovirheen ohella, on se, ettei eri lähteistä tulleita havaintoja pystytä suoraan tunnistamaan.

4. Parametrien estimointiongelmia

Estimoitaessa digressiomallien parametreja on odotettavissa eräitä lisäongelmia tavallisten regressiomallien käsittelyyn verrattuna.
Tarkastelkaamme jälleen aikaisempaa johdattelevaa esimerkkiä. Oletetaan, että käytettävissä oleva tieto rajoittuu aineistoon U1975(A) (kuva 1.1) ja pelkkään epäilyyn siitä, että "aineisto saattaa olla heterogeenista ja noudattaa digressiomallia

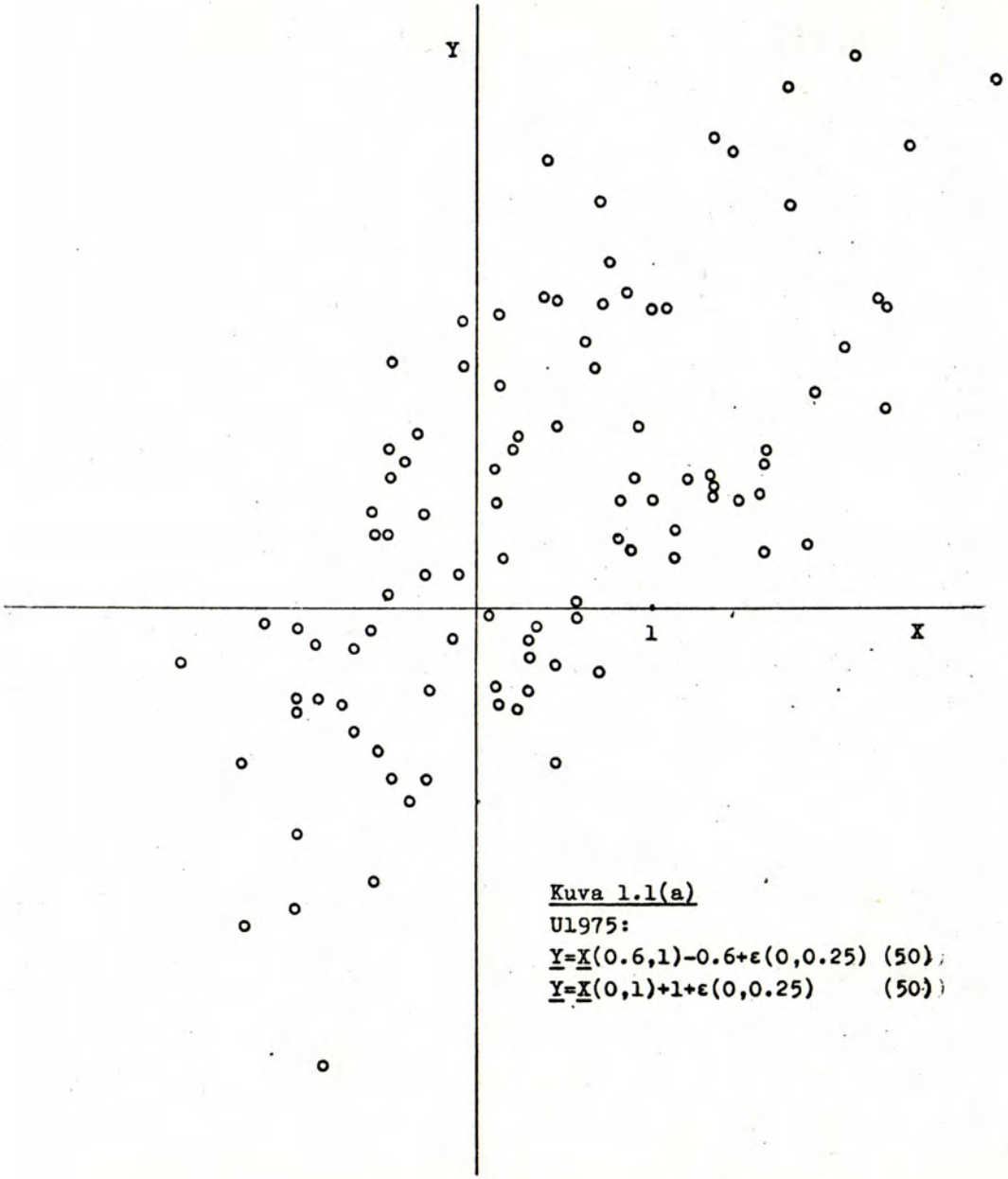
$$(4.1)\qquad y=\begin{cases}\alpha x+\beta+\varepsilon_1\\ \alpha x+\gamma+\varepsilon_2,\end{cases}$$

mutta on myös mahdollista, että kysymyksessä on suuremman satunnaisvaihtelun omaava aineisto, jota kuvaa tavallinen regressiomalli

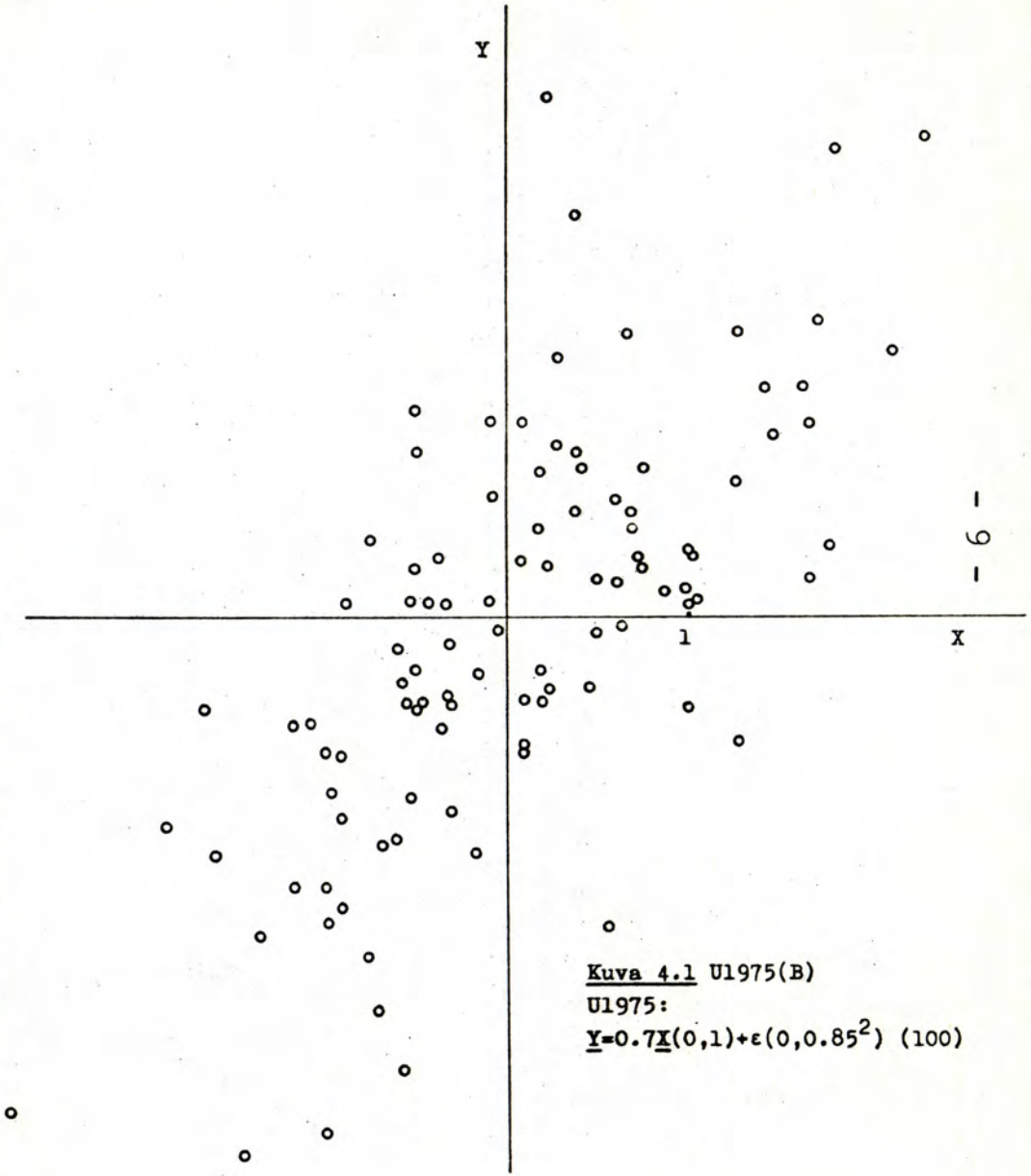
$$(4.2)\qquad y=\alpha' x+\beta'+\varepsilon \quad ".$$

Tällöin on vastassa päätösongelma, kumpaa hypoteesia käytössä oleva aineisto tukee paremmin. Ongelman voi yrittää ratkaista vertailemalla mallien (4.1) ja (4.2) selityskykyä. Digressiomallin (4.1) antama selitys kuvattuna jäännösneliösummalla on luonnollisesti aina parempi kuin regressiomallin (4.2) aineistosta riippumatta. Tässä tapauksessa nämä neliösummat ovat mallilla (4.1) S_D=18.85 (S_D on siis valikoivan pns-kriteerin minimiarvo) ja mallilla (4.2) S_R=76.14. Jonkinlaisena testikriteerinä voidaan käyttää suhdetta S_D/S_R=0.2476 tai jotain sen monotonista muunnosta esim. F-testisuureen tapaiseksi. Onko saatu S_D/S_R-arvo osoitus digressiohypoteesin erinomaisuudesta regressiohypoteesiin verrattuna eli onko jäännösneliösumman arvo siis pudonnut riittävästi? Kunnollinen vastaus edellyttää testisuureen S_D/S_R jakauman hallitsemista. Tämä ei ole tiedossa, mutta ehkä seuraava tarkastelu auttaa asiassa.

Otamme vertailun vuoksi käyttöön toisen 100 havainnon aineiston U1975(B), joka noudattaa regressiomallia (4.2) parametrein α'=0.7, β'=0.2 ja $\varepsilon \sim N(0,0.85^2)$. Sovittamalla tämä aineisto malliin (4.2) saadaan estimaatit a'=0.819, b'= 0.100 ja S_R=70.36. Aineiston ominaisuudet ovat siis regressiomallin kannalta suunnilleen samat kuin heterogeenisen aineiston U1975(A). Jos nyt tämä homogeeninen aineisto käsitellään digressiomallilla (4.1), saadaan estimaatit a=0.923, b=-0.487, c=0.832 ja S_D=28.47. Vastaavat arvot heterogeenisella aineistolla olivat a=0.981, b=-0.632, c=0.968 ja S_D=18.85. Tilanne näyttää yllättävän samanlaiselta. Digressiomalli jakaa homogeenisenkin aineiston U1975(B) tasaisesti kahdelle yhdensuuntaiselle suoralle, joiden etäisyys toisistaan on nyt tosin hieman pienempi. Merkityksellisintä on kuitenkin se, että tunnusluvut S_D=28.47 ja S_D/S_R=0.4046 ovat selvästi suuremmat. Kohdassa 5 esitettävän tarkastelun perusteella on arvioitavissa, että suhteen S_D/S_R arvo on suurilla otoskoilla tyyppiä U1975(B) olevilla homogeenisilla aineistoilla noin 0.36 ja tyyppiä U1975(A) olevilla heterogeenisilla aineistoilla noin 0.24. Suhteen hajonnasta on saatu käsitys simulointikokeiden avulla; se on suuruusluokkaa 0.03 (kts. taulukko 9.1).



Kuva 1.1(a)
U1975:
$\underline{Y}=\underline{X}(0.6,1)-0.6+\varepsilon(0,0.25)$ (50)
$\underline{Y}=\underline{X}(0,1)+1+\varepsilon(0,0.25)$ (50)

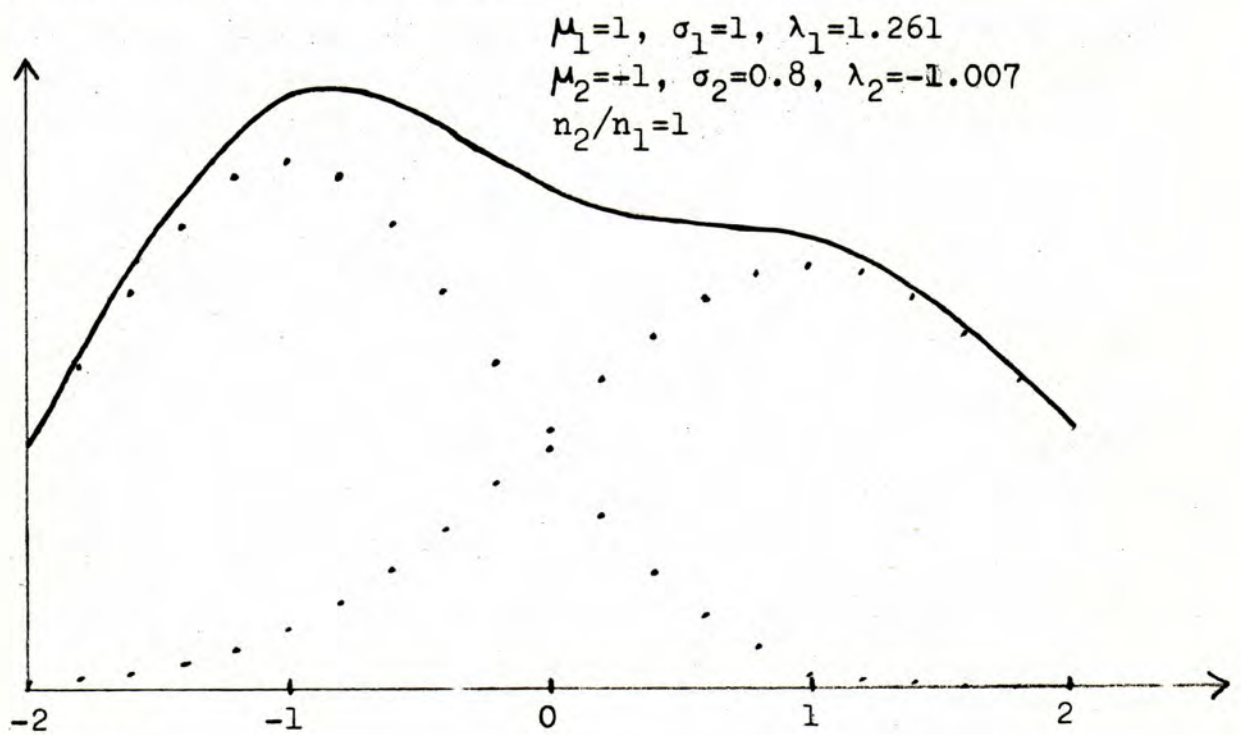


Kuva 4.1 U1975(B)
U1975:
$\underline{Y}=0.7\underline{X}(0,1)+\varepsilon(0,0.85^2)$ (100)

Toinen estimointia haittaava pulma on se, että estimaatit saattavat olla vahvasti harhaisia etenkin, jos heterogeenisuus on lievää ja osa-aineistot ovat pahasti päällekkäin. Tällöin tosin koko digressio-ongelmakin katoaa. Estimaattien harhaisuus näkyy kärjistyneesti äskeisestä esimerkistä aineiston U1975(B) kohdalla, jossa parametrien β ja γ estimaateiksi saatiin b=-0.487 ja c=0.832, kun molemmilla todellisten arvojen tulisi olla 0.2. Estimoitaessa saattaa syntyä siis liioiteltu "digressioefekti", jota tässä sanotaan digressioharhaksi. Tämän harhan suuruus riippuu monista osatekijöistä, ennen muuta aineiston heterogeenisuuden asteesta (harha lievenee heterogeenisuuden kasvaessa) ja estimoitavan parametrin luonteesta. Esim. digressiomallissa (4.1) aineiston sijaintia y-akselin suunnassa kuvaavat parametrit β, γ saattavat saada selvästi harhaiset estimaatit, mutta trendiparametri α on tässä suhteessa helpompi estimoitava.

5. Normaalijakaumien erottaminen

Edellä todetun mahdollisen digressioharhan luonnetta pyritään nyt kuvaamaan tyytymällä tarkastelemaan pelkästään yhden muuttujan heterogeenista jakaumaa, joka oletetaan kahden normaalijakauman sekoitukseksi. Olkoot tämän sekoituksen synnyttävät jakaumat $N(\mu_1,\sigma_1^2)$ ja $N(\mu_2,\sigma_2^2)$ painosuhteessa n_1/n_2.



Tutkittavana on seuraava digressiotehtävä: On määrättävä vakiot λ_1 ja λ_2 $(\lambda_1 \geq \lambda_2)$ siten, että mitan

$$(5.1) \qquad \min((x-\lambda_1)^2,(x-\lambda_2)^2)$$

odotusarvo on pienin mahdollinen. Tämä tehtävä on teoreettinen vastine digressiomallille

$$y=\begin{cases}\lambda_1+\varepsilon_1\\ \lambda_2+\varepsilon_2.\end{cases}$$

Koska tunnetusti $E(x-\alpha)^2$ saavuttaa miniminsä α:n suhteen, kun α on muuttujan x odotusarvo $E(x)$, lukupari (λ_1,λ_2) voidaan tulkita eräänlaiseksi (heterogeenisen) jakauman kaksoisodotusarvoksi. Vastaavasti mitan (5.1) odotusarvon minimiarvoa σ_D^2 sanottakoon kaksoisvarianssiksi. Parametrit λ_1,λ_2 sijoittuvat ilmeisesti sitä lähemmäksi osajakumien odotusarvoja μ_1, μ_2, mitä heterogeenisemmasta jakaumasta on kysymys eli mitä suurempi on odotusarvojen ero suhteessa hajontaan. Jos siis jakauma on täysin homogeeninen, parametrit λ_1 ja λ_2 eivät suinkaan yhdy, vaan niihin vaikuttaa erottavasti edellä kuvailtu digressioharha.

Jotta digressioharhan suuruudesta saataisiin käsitys, yritetään määrätä parametrit λ_1,λ_2. Tämä tapahtuu minimoimalla funktio

$$S(\lambda_1,\lambda_2)=\int_{\bar{\lambda}}^{\infty}(x-\lambda_1)^2 f(x)\,dx + \int_{-\infty}^{\bar{\lambda}}(x-\lambda_2)^2 f(x)\,dx,$$

missä $\bar{\lambda}=(\lambda_1+\lambda_2)/2$ ja $f(x)$ on tarkasteltavan heterogeenisen jakauman tiheysfunktio

$$f(x)=\frac{1}{n_1+n_2}\left[\frac{n_1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)+\frac{n_2}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right].$$

Käyttämällä normeeratun normaalijakauman tiheys- ja kertymäfunktiolle merkintöjä $\varphi(x)$, $\Phi(x)$ saadaan funktiolle $S(\lambda_1,\lambda_2)$ esitys

$$S(\lambda_1,\lambda_2)=\frac{1}{n_1+n_2}\sum_{i=1}^{2} n_i\Big(\sigma_i^2-2\sigma_i(\lambda_1-\lambda_2)\varphi(\delta_i)+(\mu_i-\lambda_1)^2\Phi(-\delta_i)+(\mu_i-\lambda_2)^2\Phi(\delta_i)\Big),$$

missä $\delta_i=(\bar{x}-\mu_i)/\sigma_i$, i=1,2. On ilmeistä, ettei yleisessä tapauksessa funktion $S(\lambda_1,\lambda_2)$ minimikohtaa ja minimiarvoa voi lausua suljetussa muodossa jakauman parametrien ja funktioiden φ, Φ avulla, vaan on tyydyttävä numeeriseen ratkaisuun.

Niissä tapauksissa, joissa luvut δ_1,δ_2 eivät riipu parametreista λ_1,λ_2, funktio $S(\lambda_1,\lambda_2)$ on yksinkertainen toisen asteen lauseke. Näin tapahtuu symmetrisessä tilanteessa $\sigma_1=\sigma_2=\sigma$, $n_2/n_1=1$, $\mu_1=\mu$, $\mu_2=-\mu$. Tällöin on myös $\lambda_2=-\lambda_1$ (merkitään $\lambda_1=\lambda$) ja

$$S(\lambda_1,\lambda_2)=S(\lambda,-\lambda)=\sigma^2-4\varphi(\mu/\sigma)\sigma\lambda+(\mu+\lambda)^2-4\Phi(\mu/\sigma)\mu\lambda.$$

Tämä saavuttaa miniminsä, kun

$$\lambda=(2\Phi(\mu/\sigma)-1)\mu+2\varphi(\mu/\sigma)\sigma$$

ja minimiarvo on kaksoisvarianssi

$$\sigma_D^2=\sigma^2+\mu^2-\Big((1-2\Phi(\mu/\sigma))\mu-2\varphi(\mu/\sigma)\sigma\Big)^2.$$

Erityisesti tapauksessa $\mu=0$ eli homogeenisessa normaalijakaumassa on $\lambda=\sqrt{2/\pi}\,\sigma$ eli kaksoisodotusarvo $(-\lambda,\lambda)$ sijoittuu keskipoikkeaman päähän odotusarvosta 0. Huomattakoon myös, että $\lambda\geq\mu$, $\sigma_D^2\leq\sigma^2$ ja $\lim_{\mu\to\infty}\lambda/\mu=1$, $\lim_{\mu\to\infty}\sigma_D^2=\sigma^2$.

Pyrittäessä tarkastelemaan heterogeenisuuden vaikutusta samaan tapaan kuin kohdassa 4 tarvitaan jäännösneliösummaa S_R vastaava varianssi, joka on juuri tarkastellun heterogeenisen jakauman varianssi

$$\sigma_R^2=\frac{1}{n_1+n_2}\left(n_1\sigma_1^2+n_2\sigma_2^2+4\cdot\frac{n_1n_2}{n_1+n_2}\,\mu^2\right)$$

$$=\sigma^2+\mu^2,\ \text{jos}\ \sigma_1=\sigma_2=\sigma,\ n_1=n_2,\ \mu_1=\mu,\ \mu_2=-\mu\,.$$

Tällöin suhdetta S_D/S_R vastaa suhde σ_D^2/σ_R^2. Edellä määritellyt tunnusluvut on esitetty seuraavissa taulukoissa eri parametriyhdelmillä.

Taulukko 5.1

μ	σ_2 (σ_1=1)	n_2/n_1	λ_1	λ_2	$\bar{x}$	σ_D	σ_D/σ_R
0.000	1.00	1.00	0.7978	-0.7978	0.0000	0.3633	0.36338
0.100	1.00	1.00	0.8018	-0.8018	0.0000	0.3670	0.36336
0.200	1.00	1.00	0.8137	-0.8137	0.0000	0.3777	0.36321
0.300	1.00	1.00	0.8335	-0.8335	0.0000	0.3952	0.36260
0.400	1.00	1.00	0.8608	-0.8608	0.0000	0.4188	0.36111
0.500	1.00	1.00	0.8955	-0.8955	0.0000	0.4491	0.35929
0.600	1.00	1.00	0.9373	-0.9373	0.0000	0.4813	0.35395
0.700	1.00	1.00	0.9857	-0.9857	0.0000	0.5182	0.34783
0.800	1.00	1.00	1.0404	-1.0404	-0.0000	0.5575	0.33996
0.900	1.00	1.00	1.1008	-1.1008	0.0000	0.6017	0.33246
1.000	1.00	1.00	1.1666	-1.1666	0.0000	0.6389	0.31948
1.100	1.00	1.00	1.2372	-1.2372	0.0000	0.6792	0.30734
1.200	1.00	1.00	1.3122	-1.3122	0.0000	0.7181	0.29431
1.300	1.00	1.00	1.3910	-1.3910	0.0000	0.7549	0.28065
1.400	1.00	1.00	1.4733	-1.4733	0.0000	0.7892	0.26664
1.500	1.00	1.00	1.5586	-1.5586	0.0000	0.8207	0.25253
1.600	1.00	1.00	1.6464	-1.6464	0.0000	0.8490	0.23850
1.700	1.00	1.00	1.7365	-1.7365	0.0000	0.8743	0.22475
1.800	1.00	1.00	1.8285	-1.8285	0.0000	0.8963	0.21141
1.900	1.00	1.00	1.9221	-1.9221	0.0000	0.9154	0.19858
2.000	1.00	1.00	2.0169	-2.0169	0.0000	0.9317	0.18635
2.100	1.00	1.00	2.1129	-2.1129	0.0000	0.9454	0.17476
2.200	1.00	1.00	2.2097	-2.2097	0.0000	0.9568	0.16385
2.300	1.00	1.00	2.3073	-2.3073	0.0000	0.9662	0.15361
2.400	1.00	1.00	2.4054	-2.4054	0.0000	0.9738	0.14406
2.500	1.00	1.00	2.5040	-2.5040	0.0000	0.9799	0.13516
2.600	1.00	1.00	2.6029	-2.6029	0.0000	0.9847	0.12690
2.700	1.00	1.00	2.7021	-2.7021	0.0000	0.9885	0.11924
2.800	1.00	1.00	2.8015	-2.8015	0.0000	0.9914	0.11215
2.900	1.00	1.00	2.9010	-2.9010	0.0000	0.9937	0.10560
3.000	1.00	1.00	3.0007	-3.0007	0.0000	0.9954	0.09954
3.100	1.00	1.00	3.1005	-3.1005	0.0000	0.9966	0.09393
3.200	1.00	1.00	3.2003	-3.2003	0.0000	0.9976	0.08875
3.300	1.00	1.00	3.3002	-3.3002	0.0000	0.9983	0.08396
3.400	1.00	1.00	3.4001	-3.4001	0.0000	0.9988	0.07952
4.000	1.00	1.00	4.0000	-4.0000	0.0000	0.9998	0.05881
5.000	1.00	1.00	5.0000	-5.0000	0.0000	0.9999	0.03846

μ	σ_2 (σ_1=1)	n_2/n_1	λ_1	λ_2	$\bar{x}$	σ_D	σ_D/σ_R
0.000	0.80	1.00	0.7180	-0.7180	0.0000	0.3043	0.37114
0.500	0.80	1.00	0.9308	-0.7389	0.0959	0.3821	0.35714
1.000	0.80	1.00	1.2610	-1.0074	0.1267	0.5495	0.30195
1.500	0.80	1.00	1.6267	-1.4583	0.0841	0.6975	0.22721
2.000	0.80	1.00	2.0501	-1.9714	0.0393	0.7782	0.16145
2.500	0.80	1.00	2.5163	-2.4882	0.0140	0.8088	0.11440
3.000	0.80	1.00	3.0042	-2.9965	0.0038	0.8175	0.08325
0.000	0.60	1.00	0.6383	-0.6383	0.0000	0.2725	0.40082
0.500	0.60	1.00	1.0352	-0.5840	0.2255	0.3252	0.34978
1.000	0.60	1.00	1.3538	-0.9060	0.2239	0.4533	0.26987
1.500	0.60	1.00	1.6643	-1.4143	0.1249	0.5760	0.19659
2.000	0.60	1.00	2.0606	-1.9588	0.0509	0.6434	0.13748
2.500	0.60	1.00	2.5183	-2.4858	0.0162	0.6697	0.09663
3.000	0.60	1.00	3.0044	-2.9962	0.0041	0.6776	0.07000
0.000	1.00	0.80	0.7978	-0.7978	0.0000	0.3633	0.36338
0.500	1.00	0.80	0.9417	-0.8470	0.0473	0.4470	0.35855
1.000	1.00	0.80	1.2158	-1.1107	0.0525	0.6377	0.32087
1.500	1.00	0.80	1.5843	-1.5276	0.0283	0.8202	0.25456
2.000	1.00	0.80	2.0266	-2.0050	0.0107	0.9316	0.18819
2.500	1.00	0.80	2.5069	-2.5002	0.0033	0.9799	0.13661
3.000	1.00	0.80	3.0015	-2.9998	0.0008	0.9954	0.10065
0.000	1.00	0.60	0.7978	-0.7978	-0.0000	0.3633	0.36338
0.500	1.00	0.60	0.9969	-0.7817	0.1075	0.4436	0.35943
1.000	1.00	0.60	1.2729	-1.0255	0.1237	0.6326	0.32653
1.500	1.00	0.60	1.6125	-1.4763	0.0680	0.8181	0.26313
2.000	1.00	0.60	2.0366	-1.9850	0.0257	0.9313	0.19606
2.500	1.00	0.60	2.5100	-2.4940	0.0079	0.9798	0.14285
3.000	1.00	0.60	3.0022	-2.9982	0.0020	0.9954	0.10547
0.000	0.80	0.80	0.7269	-0.7269	0.0000	0.3115	0.37086
0.500	0.80	0.80	0.9806	-0.7024	0.1390	0.3857	0.35488
1.000	0.80	0.80	1.2974	-0.9631	0.1671	0.5532	0.30269
1.500	0.80	0.80	1.6397	-1.4312	0.1042	0.7083	0.23132
2.000	0.80	0.80	2.0531	-1.9602	0.0464	0.7946	0.16587
2.500	0.80	0.80	2.5168	-2.4845	0.0161	0.8277	0.11803
3.000	0.80	0.80	3.0042	-2.9956	0.0043	0.8373	0.08606
0.000	0.80	0.60	0.7380	-0.7380	0.0000	0.3202	0.37028
0.500	0.80	0.60	1.0365	-0.6550	0.1907	0.3883	0.35326
1.000	0.80	0.60	1.3402	-0.8941	0.2230	0.5551	0.30800
1.500	0.80	0.60	1.6555	-1.3848	0.1353	0.7207	0.24232
2.000	0.80	0.60	2.0566	-1.9410	0.0578	0.8150	0.17659
2.500	0.80	0.60	2.5173	-2.4784	0.0194	0.8514	0.12661
3.000	0.80	0.60	3.0043	-2.9940	0.0051	0.8620	0.09266
0.000	0.60	0.80	0.6560	-0.6560	0.0000	0.2851	0.39852
0.500	0.60	0.80	1.0755	-0.5623	0.2565	0.3321	0.34513
1.000	0.60	0.80	1.3723	-0.8709	0.2506	0.4645	0.27276
1.500	0.60	0.80	1.6696	-1.3891	0.1402	0.5993	0.20402
2.000	0.60	0.80	2.0615	-1.9478	0.0568	0.6747	0.14459
2.500	0.60	0.80	2.5184	-2.4822	0.0180	0.7041	0.10221
3.000	0.60	0.80	3.0044	-2.9953	0.0045	0.7129	0.07423
0.000	0.60	0.60	0.6782	-0.6782	0.0000	0.3000	0.39479
0.500	0.60	0.60	1.1175	-0.5344	0.2915	0.3398	0.34181
1.000	0.60	0.60	1.3968	-0.8144	0.2911	0.4766	0.28081
1.500	0.60	0.60	1.6779	-1.3456	0.1661	0.6275	0.21870
2.000	0.60	0.60	2.0629	-1.9289	0.0669	0.7136	0.15823
2.500	0.60	0.60	2.5185	-2.4761	0.0212	0.7470	0.11285
3.000	0.60	0.60	3.0045	-2.9937	0.0053	0.7571	0.08375

6. Parametrien estimointi käytännössä

Tutkimme nyt lähemmin digressiomallin

$$y=\begin{cases} f_1(x,\alpha^{(1)})+\varepsilon_1 \\ f_2(x,\alpha^{(2)})+\varepsilon_2 \end{cases}$$

parametrien $\alpha^{(1)},\alpha^{(2)}$ estimointia valikoivaa pns-keinoa käyttäen. Tämä merkitsee käytännössä epälineaarisen optimointitehtävän (2.1) ratkaisemista. Tehtävän luonteesta saanee jonkinlaisen käsityksen seuraavan esimerkin avulla.

Esim. Tarkastellaan jälleen yksinkertaista digressiomallia

$$y=\begin{cases} \lambda_1+\varepsilon_1 \\ \lambda_2+\varepsilon_2. \end{cases}$$

Tässä tapauksessa valikoivan kriteerin soveltaminen merkitsee havaintoaineiston $y_1,y_2,\ldots,y_n$ jakamista kahteen osajoukkoon $\{y_j, j\in J_1\}, \{y_j, j\in J_2\}$, $(J_1\cup J_2=\{1,2,\ldots,n\},\ J_1\cap J_2=\emptyset)$ ja lukujen λ_1,λ_2 valitsemista siten, että summa

$$\sum_{j\in J_1}(y_j-\lambda_1)^2+\sum_{j\in J_2}(y_j-\lambda_2)^2$$

on mahdollisimman pieni. Jokaisella annetulla J_1,J_2-osituksella minimi saavutetaan, kun

$$\lambda_i=\frac{1}{|J_i|}\sum_{j\in J_i} y_j,\quad i=1,2.$$

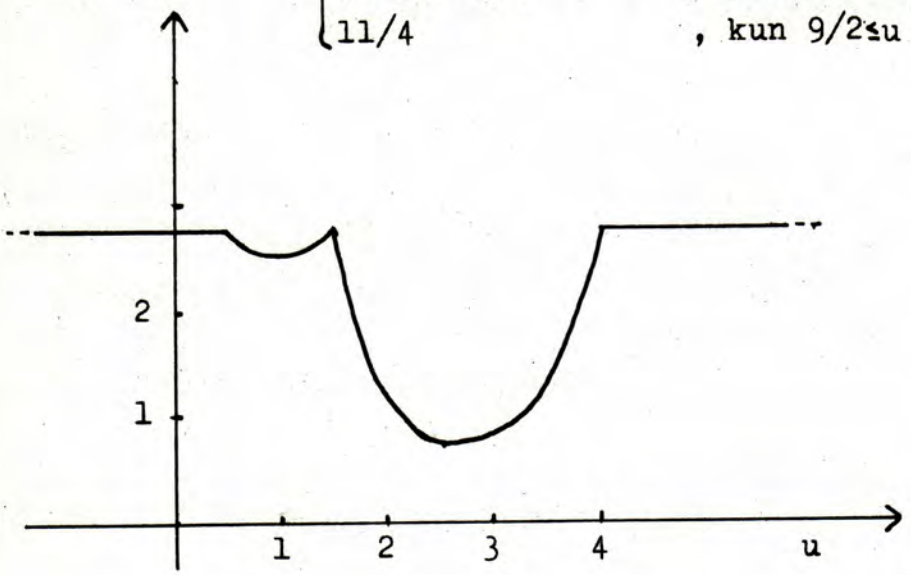
Vertaamalla siis eri osituksilla saatuja minimejä löydetään ratkaisu. Asettamalla havaintoarvot suuruusjärjestykseen joudutaan käytännössä vertaamaan vain muutamia osituksia.

Esimerkissä kuvattua yksinkertaista menettelyä ei kuitenkaan liene mahdollista soveltaa tehokkaasti kaksi- ja useampiulotteisiin tilanteisiin. Tällöin vertailtavien ositusten lukumäärä saattaa kasvaa kohtuuttoman suureksi, sillä otoksen järjestämistä vastaavaa toimitusta ei ole helposti saatavilla. Yleisessä ratkaisumenetelmässä, joka perustuu aina jonkin epälineaariseen optimointiin tarkoitetun algoritmin käyttöön, joudutaan seuraamaan

funktion $S(a^{(1)},a^{(2)})$ käyttäytymistä parametriavaruudessa lähtien liikkeelle mahdollisimman hyvistä alkuarvoista. Tehtävä ei ole aivan helppo, sillä optimoitava funktio saattaa käyttäytyä hankalasti.

Esim.(jatk.) Olkoon tutkittava otos $y_1=1, y_2=2, y_3=3$. Tällöin optimaalinen λ_1,λ_2-piste on kaksikäsitteinen. Sekä $\lambda_1=3, \lambda_2=3/2$ että $\lambda_1=5/2, \lambda_2=1$ antaa minimineliösummaksi $S_D=3/4$. Oletetaan, että toinen optimaalisista λ_2-arvoista, $\lambda_2=1$ olisi löytynyt ja katsomme, miltä neliösumma näyttää minimoidaksemme sen muuttujan $u=\lambda_1$ suhteen. Minimoitava funktio on

$$f(u)=\begin{cases} 11/4 & , \text{ kun } u\leq 1/2 \\ (u-1)^2+5/2 & , \text{ kun } 1/2\leq u\leq 3/2 \\ (u-2)^2+(u-3)^2+1/4 & , \text{ kun } 3/2\leq u\leq 5/2 \\ (u-3)^2+1/2 & , \text{ kun } 5/2\leq u\leq 9/2 \\ 11/4 & , \text{ kun } 9/2\leq u \end{cases}$$



Jos tämän funktion yrittää minimoida jollakin miniminhakumenetelmällä esim. askelpituutta 0.1 käyttäen lähtöpisteestä u=0, on melkoinen vaara kompastua pisteessä u=3/2 olevaan "ryppyyn" ja joutua lokaaliseen minimipisteeseen u=1 oikean minimipisteen $\lambda_1=u=5/2$.

"Todellisilla" aineistoilla, siis suurilla havaintomäärillä kohdefunktio S ei ehkä käyttäydy, suhteellisesti katsoen, näin pahasti, mutta silti sillä saattaa olla ryppyjen välissä lokaalisia minimikohtia, joilta minimointialgoritmin tulisi välttyä. Tällöin on tärkeätä, että algoritmi pystyy näkemään kohdefunktion suuret linjat ryppyilystä huolimatta.

Olen kokeillut tässä tehtävässä lähinnä kahta algoritmia. Toinen on ns. vaihtelevan metriikan menetelmä (variable metric method) eli Davidon-Fletcher-Powell-menetelmä; käytän siitä nimitystä DFP-menetelmä (kts. esim. Walsh, 1975). Se kuuluu gradienttimenetelmien luokkaan ja on eräs kaikkein tehokkaimpia optimointimenetelmiä edellyttäen, että optimoitava funktio on riittävän siisti. Tavallisesti edellytetään funktion toisten derivaattojen olemassaolo, mikä on tässä tapauksessa liikaa. Koska kysymyksessä on kuitenkin "suurten askelten" menetelmä, joka jatkuvasti vaihtelee askelpituutta ja jossa koko ajan ollaan valmiit hyväksymään uusi käsitys funktion senhetkisestä lokaalisesta käyttäytymisestä, algoritmi on toiminut tyydyttävästi mutta ei erehtymättömästi. Edellytyksenä DFP-menetelmän toiminnalle on, että (paikoitellen olemattomia) derivaattoja arvioitaessa numeerisesti muodossa

$$f'(x) \approx \frac{f(x+h)-f(x)}{h}$$

käytetään riittävän suurta h:ta, jolloin rypyt eivät liikaa pääse häiritsemään ja funktion yleinen käyttäytyminen hallitsee algoritmin toimintaa.

Toinen käyttämistäni menetelmistä on Hooke-Jeeves-menetelmä (esim. Walsh, 1975), joka on varsin yksinkertainen mutta tehokas hakumenetelmä. Menetelmä ei tarvitse kohdefunktion derivaattoja lainkaan. Se on näissä tehtävissä hitaampi (ajoajat tietokoneella jopa kymmenkertaisia) DFP-menetelmään verrattuna, mutta se on osoittautunut menestykselliseksi niissäkin tapauksisa, joissa DFP-algoritmi on pettänyt. Niinpä olen suorittanut useimmat digressiomallien estimoinnit Hooke-Jeeves-menetelmällä.

7. Estimaattien luotettavuudesta

Teoreettisten tulosten puuttuessa digressiomallien parametrien luotettavuuden arviointiin ei toistaiseksi ole perusteltuja keinoja. Seuraavassa tyydytään esittämään kaksi heuristista arviointitapaa.

Ensimmäinen tapa perustuu yksinkertaisesti estimoinnin yhteydessä syntyvään havaintojen luokitteluun. Kunkin osamallin parametrien luotettavuutta voidaan tällöin yrittää arvioida osamallin "omien" havaintojen avulla regressiomallien teorian mukaisesti. Tällä menettelyllä on kuitenkin selviä epäkohtia. Se tuskin toimii hyvin, jos osa-aineistot ovat pahasti päällekkäin, koska tällöin havaintojen ryhmittely ei onnistu kunnolla. Vaikuttaa nimittäin siltä, että digressioanalyysi saattaa toimia tyydyttävästi parametrien estimoinnin osalta, vaikka havaintojen ryhmittely on epäluotettavaa. Vielä vakavampi epäkohta lienee se, että osa parametreista on useimmiten eri osamallien yhteisiä, jolloin niiden estimointi on perustunut laajemmalle havaintojoukolle, kuin mitä osamallin omat vapausasteet edellyttävät. Tämä saattaa vaikuttaa ainakin tapaan, millä vapausasteet tulisi laskea virheenarvioinnissa.

Toinen parametrien luotettavuuden arviointitapa pohjautuu valikoivan pns-kriteerin $S(a_1,a_2,\ldots,a_k)$, missä $a_1,a_2,\ldots,a_k$ ovat nyt estimoidut parametrit, käyttäytymisen seuraamiseen optimipisteen $(a_1,a_2,\ldots,a_k)$ ympäristössä. Oletetaan, että funktiolle riittää kvadraattinen approksimaatio tässä ympäristössä. Tällöin estimaattoreiden $a_1,a_2,\ldots,a_k$ kovarianssimatriisi on likimain $\frac{1}{2}s^2G^{-1}$, missä $s^2=\frac{1}{n-k}S_D$ tai jokin muu jäännösvarianssin estimaattori ja

$$G=\left[\frac{\partial S(a_1,a_2,\ldots,a_k)}{\partial a_i \partial a_j}\right] .$$

Osittaisderivaatat on käytännössä korvattava vastaavilla esim. muotoa

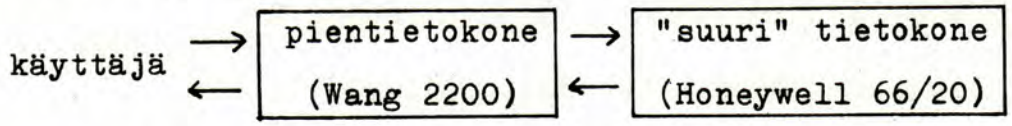
$$\frac{1}{h^2}\big(f(x+h,y+h)-f(x+h,y)-f(x,y+h)+f(x,y)\big)\approx\frac{\partial f(x,y)}{\partial x \partial y}$$

olevilla erotusosamäärillä käyttäen riittävän suurta h:ta, jotta

funktion S epäsäännöllisyydet eivät haittaisi liikaa ja jotta myös luokituksen suhteen epävarmojen havaintojen vaikutus kuvastuisi arvioissa. Tätä tapaa on käytetty myöhemmin kuvattavien simulointikokeiden yhteydessä. Se näyttää antavan parametreille hieman liian pieniä keskivirheitä, mikä johtunee liian pienistä "jäännösvarianssin" s^2 arvoista.

8. Digressiomallien laskeminen

Digressiomallien käsittely edellyttää tehokasta tietojenkäsittelyvälineistöä. Tässä tutkimuksessa analyysit on tehty laatimallani yleisellä epälineaarisella regressio-ohjelmalla, joka perustuu pientietokoneen ja "suuren" tietokoneen yhteiskäyttöön ja toimii interaktiivisella periaatteella.



Koko laskentaprosessin ohjaus on pienkoneella, jonka nopean näyttölaitteen välityksellä käyttäjä voi monipuolisesti seurata toimintaa ja vaikuttaa tarvittaessa siihen. Pienillä aineistoilla (havaintoja alle 100, parametreja alle 5) pienkone suorittaa koko laskennan. Suuremmissa tehtävissä suuri kone kannattaa ottaa "orjakoneeksi", joka hoitaa joko pelkästään kohdefunktion (neliösummien) laskemisen pienkoneen antamien parametriarvojen avulla tai suorittaa koko varsinaisen minimointitehtävän pienkoneen valvonnassa. Huolimatta alhaisesta tiedonsiirtonopeudesta koneiden välillä (tässä 300 baudia eli 30 merkkiä/s) tällainen yhteiskäyttöratkaisu on tehokas, koska siirrettävää tietoa on laskennan aikana vähän (parametrien ja kohdefunktion arvot). Käyttäjä hyötyy samanaikaisesti molempien osajärjestelmien parhaista ominaisuuksista, pienkoneen monipuolisista ja joustavista käyttömahdollisuuksista ja suuren koneen suuresta muistista ja laskentanopeudesta.

Koko ohjelmisto on laadittu BASIC-kielellä, joka on laajennetussa muodossa tässä käytetyn pienkoneen ainoa ohjelmointikieli. Regressio-ohjelmaa käytettäessä havaintoarvot annetaan joko manuaalisesti pienkoneelle tai käytetään valmista havaintotiedostoa, joka on luotu aikaisemmin. Tässä tutkimuksessa käytetyt simuloidut aineistot on generoitu tarvittaessa jopa samanaikaisesti kummallakin koneella erikseen joka ajokerralla.
Mallin määrittely tapahtuu aina laskennan alussa lisäämällä ohjelmaan mallifunktion (ja haluttaessa myös sen derivaatat parametrien suhteen) laskeva BASIC-aliohjelma. Esim. lineaarisen regressiomallin $y=\alpha_1 x_1+\alpha_2 x_2+\alpha_3+\varepsilon$ käsittelyyn, mihin löytyy tietenkin suoremmat keinot, kävisi aliohjelma

```
100 F=A(1)*X(1)+A(2)*X(2)+A(3):RETURN.
```

Mallia

$$y=\alpha_1 \exp(-\alpha_2 x)+\varepsilon$$

taas vastaa aliohjelma

```
100 F=A(1)*EXP(-A(2)*X(1)):RETURN.
```

Ohjelma muodostaa tämän funktioaliohjelman ja havaintoaineiston avulla (mahdollisesti painotetun) neliösummafunktion, jonka tilalla voi helposti käyttää myös esim. virheiden itseisarvosummafunktiota. Ohjelma minimoi tämän funktion lähtien käyttäjän antamista alkuarvoista jollakin iteratiivisella algoritmilla, joka on käyttäjän valittavissa ja jopa vaihdettavissa toiseen menetelmään tarvittaessa kesken laskennan.

Ohjelmaa voi sellaisenaan käyttää digressiomallien parametrien estimointiin valikoivalla pns-keinolla (tai muulla samansukuisella menetelmällä). Tällöin vain funktioaliohjelma kirjoitetaan ehdolliseen muotoon. Digressiomallin

$$y=\begin{cases} f_1(x,\alpha^{(1)})+\varepsilon_1 \\ f_2(x,\alpha^{(2)})+\varepsilon_2 \end{cases}$$

laskemiseksi tarvitaan aliohjelma, joka on muotoa

```
100 IF (y_j-f_1(x,α^(1)))^2 > (y_j-f_2(x,α^(2)))^2 THEN 102
101 F=f_1(x,α^(1)):RETURN
102 F=f_2(x,α^(2)):RETURN
```

Esim. digressiomallin

$$y=\begin{cases}\alpha_1 x+\alpha_2+\varepsilon_1\\ \alpha_1 x+\alpha_3+\varepsilon_2\end{cases}$$

tapauksessa tarvittava aliohjelma on tyyppiä

```
100 F1=A(1)*X(1)+A(2):F2=A(1)*X(1)+A(3)
101 IF ABS(Y(J)-F1)>ABS(Y(J)-F2) THEN 103
102 F=F1:RETURN
103 F=F2:RETURN
```

Käytännössä aliohjelma kannattaa hioa mahdollisimman nopeaksi, sillä se on koko laskennan ydin, joka kohdataan mallin käsittelyn aikana tuhansia kertoja.

9. Simulointikokeet

Digressiomallien estimointimahdollisuuksia on nyt tarkoitus selvittää analysoimalla keinotekoisia heterogeenisia aineistoja. Toistuvana periaatteena on se, että generoidaan valittuja regressiomalleja noudattavia havaintoaineistoja, yhdistetään niitä kaksittain ja yritetään sitten määrätä mallien parametrit tästä yhteisaineistosta, jossa havaintojen todellinen alkuperä on "unohdettu". Vain vaihtoehtoisten mallien muodot oletetaan tunnetuiksi.
Tällainen puolikokeellinen tutkimustapa lienee ainakin nopein keino päästä selville menetelmän toimintakyvystä ja sovellutuskelpoisuudesta eri tilanteissa ja se voi antaa myös viitteitä siitä, mikä on perusteellisemman tutkimisen arvoista.

Kaikissa simulointikokeissa on käytetty liitteessä esitettyä aineistojen generointitapaa, joka mm. tekee mahdolliseksi haluttaessa toistaa nämä kokeet ja jatkaa niitä täsmälleen samoilla aineistoilla millä tahansa tietojenkäsittelylaitteistolla.
Kokeet ovat vasta alustavia. Jotkin niistä on toistettu (3-25 kertaa) riippumattomilla aineistoilla menetelmän stabiilisuuden ja tarkkuuden selvittämiseksi. Iteratiiviset estimoinnit on toteutettu melkein poikkeuksetta "huonoista" alkuarvoista lähtien estimointialgoritmin toimivuuden testaamiseksi. Muutamat mallit on lisäksi estimoitu useista alkuarvoista lähtien tulosten yksikäsitteisyyden toteamiseksi.
Ei ole kuitenkaan oikeutettua tehdä tuloksista kovin pitkälle meneviä johtopäätöksiä, sillä eräitä tärkeitä koemuuttujia ei ole juuri lainkaan varioitu. Mm. osamallien virhevarianssit ovat olleet aina keskenään yhtäsuuret ja samoin ovat yleensä osaaineistojen havaintomäärät.
Otoskoko on ollut yleensä 100 tai 200, mutta myös 1000 havainnon aineistoja on käytetty tilanteissa, joissa on haluttu vähentää otantavirheen aiheuttamia häiriöitä.

9.1. Ambivalentti vakiotermi

Palaamme vielä jo johdannossa esiteltyyn esimerkkiin, joka on nyt toistettu useilla riippumattomilla aineistoilla. Tarkastelun kohteena on siis digressiomalli

$$y=\begin{cases}\alpha x+\beta+\varepsilon_1\\ \alpha x+\gamma+\varepsilon_2,\end{cases}$$

jota tutkitaan aluksi aineistoilla

U1975-U1999:

$\underline{Y}=\underline{X}(0.6,1)-0.6+\varepsilon(0,0.25)$ (50)

$\underline{Y}=\underline{X}(0,1)+1+\varepsilon(0,0.25)$ (50).

Oikeat parametrinarvot ovat siis $\alpha=1$, $\beta=-0.6$ ja $\gamma=1$. Jokainen näistä 25 aineistosta on käsitelty seuraavin tavoin:

1) On määrätty parametrien α,β,γ digressioestimaatit a,b,c ja digressiomallin jäännösneliösumma S_D. Alkuarvoina on käytetty a=b=c=0 eli regressiosuorat yhtyvät molemmat x-akseliin tässä alkutilassa.
2) Aineistot on sovitettu pns-keinolla myös tavalliseen regressiomalliin $y=\alpha' x+\beta'+\varepsilon$. Parametrien α',β' estimaatteja merkitään a,b ja jäännösneliösummaa S_R.
3) Tehtävä 1) on toistettu korvaamalla valikoiva pns-kriteeri (2.1) valikoivalla virheiden itseisarvosummakriteerillä, jonka antamia estimaatteja merkitään a",b",c".
4) Aineistot on lopuksi keskistetty origoon muotoon

   U1975-U1999:

   $\underline{Y}=\underline{X}(0,1)+\varepsilon(0,0.25)$ (100),

   jolloin ne muuttuvat homogeenisiksi. Jotta saataisiin käsitys, paljonko kussakin tapauksessa todellinen tilanne poikkeaa teoreettisesta lähtökohdasta, näillä aineistoilla on estimoitu mallit $y=\alpha''x+\beta''+\varepsilon$, jolloin on saatu estimaatit a",b".

Tulokset ovat taulukossa 9.1.

Taulukko 9.1

	a	b	c	S_D	S_D/S_R	a'	b'	S_R	a"	b"	c"	a"'	b"'	$S_{R"}$
U1975	0.981	-.632	0.968	18.85	0.248	.736	.240	76.14	1.002	-.636	0.990	1.070	-.059	24.34
U1976	0.945	-.607	1.125	23.00	0.251	.742	.264	91.46	0.941	-.518	1.065	0.995	-.019	28.52
U1977	0.928	-.552	1.112	19.89	0.230	.787	.361	86.41	1.002	-.578	1.124	0.971	-.091	26.85
U1978	0.989	-.784	0.827	16.06	0.204	.878	.141	78.65	1.093	-.832	0.836	1.056	-.104	19.31
U1979	1.018	-.521	1.120	25.75	0.285	.879	.319	90.50	1.042	-.375	1.085	1.034	.089	26.97
U1980	0.984	-.473	1.197	21.89	0.259	.771	.312	84.44	0.817	-.356	1.121	1.040	.040	25.19
U1981	0.942	-.541	1.055	24.21	0.301	.681	.372	80.51	0.932	-.438	0.982	1.000	.051	28.40
U1982	1.002	-.477	1.076	22.38	0.285	.840	.250	78.59	0.990	-.492	0.983	1.026	-.010	26.47
U1983	0.809	-.414	1.065	17.22	0.240	.793	.346	71.88	1.047	-.475	0.963	0.972	.074	20.03
U1984	0.954	-.462	1.105	23.33	0.280	.849	.326	83.34	1.076	-.619	0.911	1.071	.039	24.47
U1985	0.974	-.656	0.948	14.09	0.193	.738	.273	73.16	0.903	-.693	1.007	1.005	-.015	19.61
U1986	1.079	-.749	0.923	20.60	0.235	.891	.171	87.52	1.094	-.754	0.894	1.062	-.057	24.72
U1987	0.856	-.405	0.984	19.14	0.286	.844	.315	67.03	1.017	-.544	0.814	1.016	.081	23.34
U1988	1.005	-.631	1.080	21.32	0.233	.853	.236	91.39	0.986	-.570	1.102	1.021	-.008	26.07
U1989	1.088	-.750	0.978	24.42	0.258	.892	.209	94.67	1.129	-.603	1.066	1.083	-.008	32.21
U1990	1.025	-.757	0.990	20.06	0.226	.739	.161	88.81	1.122	-.606	1.073	1.068	-.030	21.57
U1991	0.927	-.589	0.998	21.58	0.263	.776	.275	81.98	0.908	-.605	1.024	0.970	-.004	22.76
U1992	0.980	-.565	1.104	22.00	0.241	.941	.315	91.42	0.839	-.534	1.161	1.008	.097	26.24
U1993	0.987	-.667	1.061	21.23	0.224	.929	.122	94.78	0.998	-.747	1.151	1.031	-.089	25.48
U1994	0.907	-.571	1.010	16.13	0.208	.794	.294	77.45	0.906	-.515	0.985	0.979	-.009	21.13
U1995	0.836	-.747	0.997	17.20	0.186	.738	.144	92.32	0.579	-.618	1.133	0.987	-.109	36.10
U1996	1.009	-.581	1.171	22.66	0.264	.673	.254	85.97	0.870	-.396	1.152	1.008	.010	26.10
U1997	1.029	-.651	0.944	17.82	0.241	.756	.272	74.03	1.082	-.685	0.819	1.007	.003	22.74
U1998	0.880	-.527	1.030	20.10	0.255	.748	.271	78.82	0.995	-.513	1.101	0.941	.002	24.35
U1999	0.925	-.592	0.956	19.00	0.245	.795	.222	77.44	1.027	-.536	1.035	0.961	-.021	22.06
$\bar{x}$	0.962	-.596	1.033	20.40	0.246	.803	.258	83.15	0.976	-.570	1.023	1.015	-.006	25.00
s	0.069	.108	0.085	2.90	0.030	.074	.070	7.73	0.119	.118	0.105	0.039	.059	3.81

Ehkä tärkeintä on havaita, että digressiomallin antama arvio **a** (keskiarvo 0.962) suuntaparametrille α on kaikissa kokeissa parempi kuin regressiomallilla saatu a' (keskiarvo 0.803), jolla on ymmärrettävästi erittäin selvä harha. Kuitenkin myös a lienee jonkin verran harhainen, sillä arvo 1 ei ilmeisesti pääse millekään järkeväntasoiselle a:n konfidenssivälille. Harhan olemassaolosta puhuu myös se seikka, että vain 8 tapauksessa a on suurempi kuin 1. Lisäksi vain neljässä tapauksessa a>a"'.

Tasoparametrien β,γ digressioestimaatit b,c vaikuttavat melko harhattomilta. Koska arvolla $\alpha=1$ tämä digressio-ongelma palautuu kahden normaalijakauman erottamiseksi, voidaan soveltaa likimäärin kohdan 5 tuloksia. Nyt on $\mu=1.6$ ja $\lambda=1.6464$ eli kaksoiskeskiarvon harhaksi saadaan 2.9%. Tämä on sen verran pieni, että se hukkuu tässä tapauksessa estimaattien b,c satunnaisvaihteluun. Estimaattien a,b,c likimääräiset hajonnat on määrätty joillekin aineistoille myös kohdassa 7 esitettyä heuristista laskutapaa käyttäen:

	s_a	s_b	s_c	s^2
U1975	.044	.069	.063	.194
U1976	.048	.072	.072	.237
U1977	.040	.069	.063	.205

Nämä hajonta-arviot ovat jonkin verran pienempiä kuin taulukossa 9.1 esiintyvät. Tämä johtuu suurelta osalta siitä, että jäännösvarianssin arviot $s^2=S_D/(n-k)$ ovat liian pieniä, sillä $\sigma_1^2=\sigma_2^2=0.25$. Jäännösneliösummat S_D,S_R käyttäytyvät sikäli odotusten mukaisesti, että suhteen S_D/S_R keskiarvo on 0.246, mikä vastaa melko hyvin kohdassa 5 esitettyä teoreettista σ_D^2/σ_R^2-arvoa 0.2385.
Koska tavallisen lineaarisen mallin jäännösneliösumma S_R noudattaa normaalisen selitysvirheen tapauksessa χ^2-jakaumaa siten, että $S_R/\sigma^2 \sim \chi^2(n-k)$, missä σ^2=virhevarianssi ja k=estimoitavien parametrien lukumäärä, voitaisiin ajatella, että tämä ominaisuus olisi myös ainakin sellaisilla digressiomalleilla, joiden osamallit ovat lineaarisia ja jäännösvarianssit yhtäsuuret.
Nyt on kuitenkin $\sigma^2=0.25$, $E(S_D/\sigma^2)\approx 82$ ja $D^2(S_D/\sigma^2)\approx 135$, mikä ei ole kovin hyvin sopusoinnussa χ^2-olettamuksen kanssa.

Itseisarvosummakriteerillä saadut parametrien estimaatit eivät liene parempia kuin pns-kriteerillä saadut. Tämä estimointitapa kärsii jopa täydellisen haaksirikon "vaikean" aineiston U1995 kohdalla, jossa a"=0.579.
Tässä koetilanteessa osa-aineistot ovat melko erillään toisistaan ja havaintojen luokittelu onnistuu hyvin valkoivalla pns-kriteerillä:

väärin luokiteltuja havaintoja (50:stä) osa-aineistossa

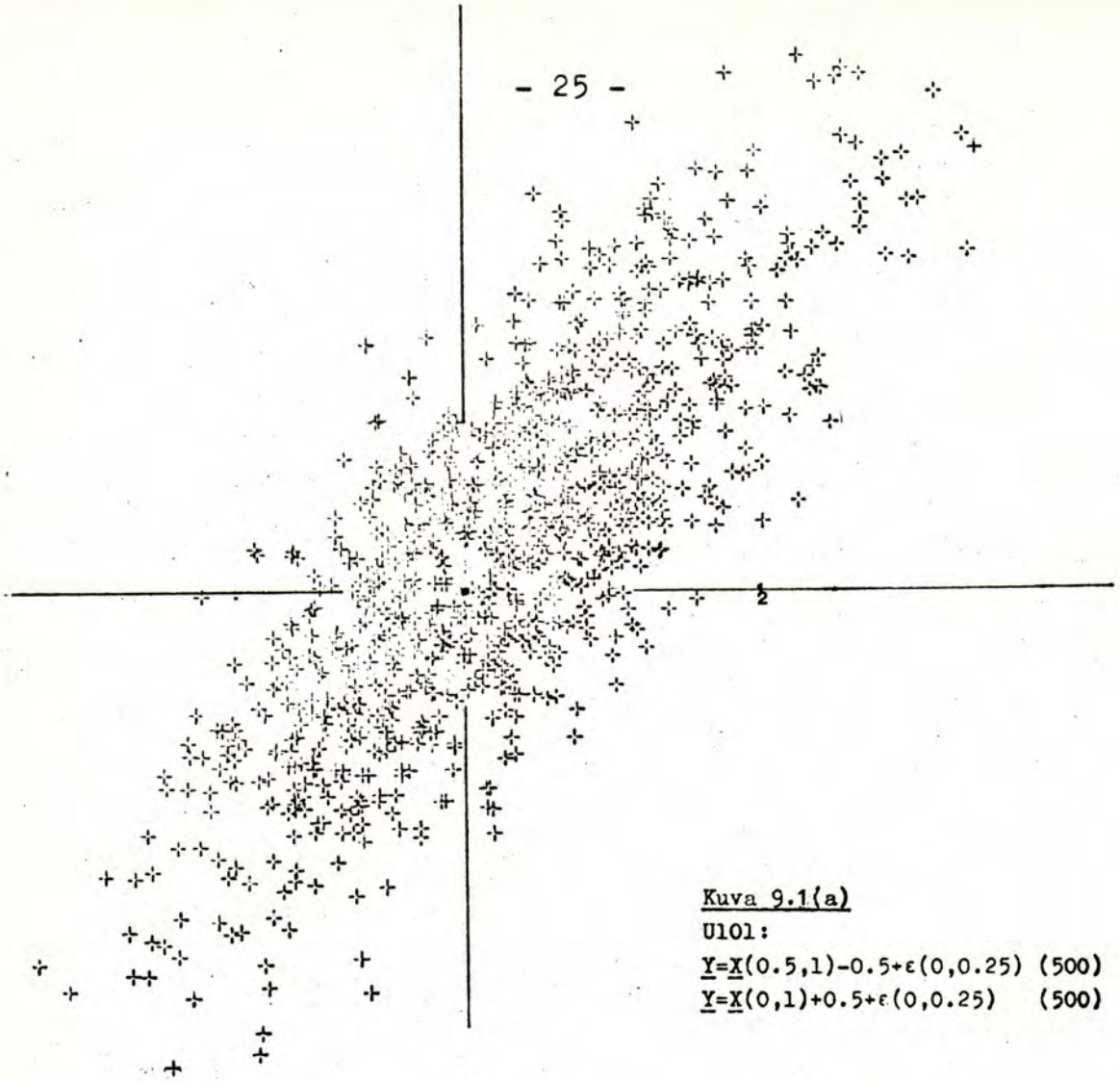
	1	2
U1975	2	3
U1976	2	6
U1977	4	2
U1983	4	3
U1995	7	7

Jotta digressioanalyysin toimivuudesta saataisiin käsitys tässä mallityypissä myös aineistojen peittäessä pahemmin toisiaan, kokeet toistettiin viidellä suuremmalla aineistolla

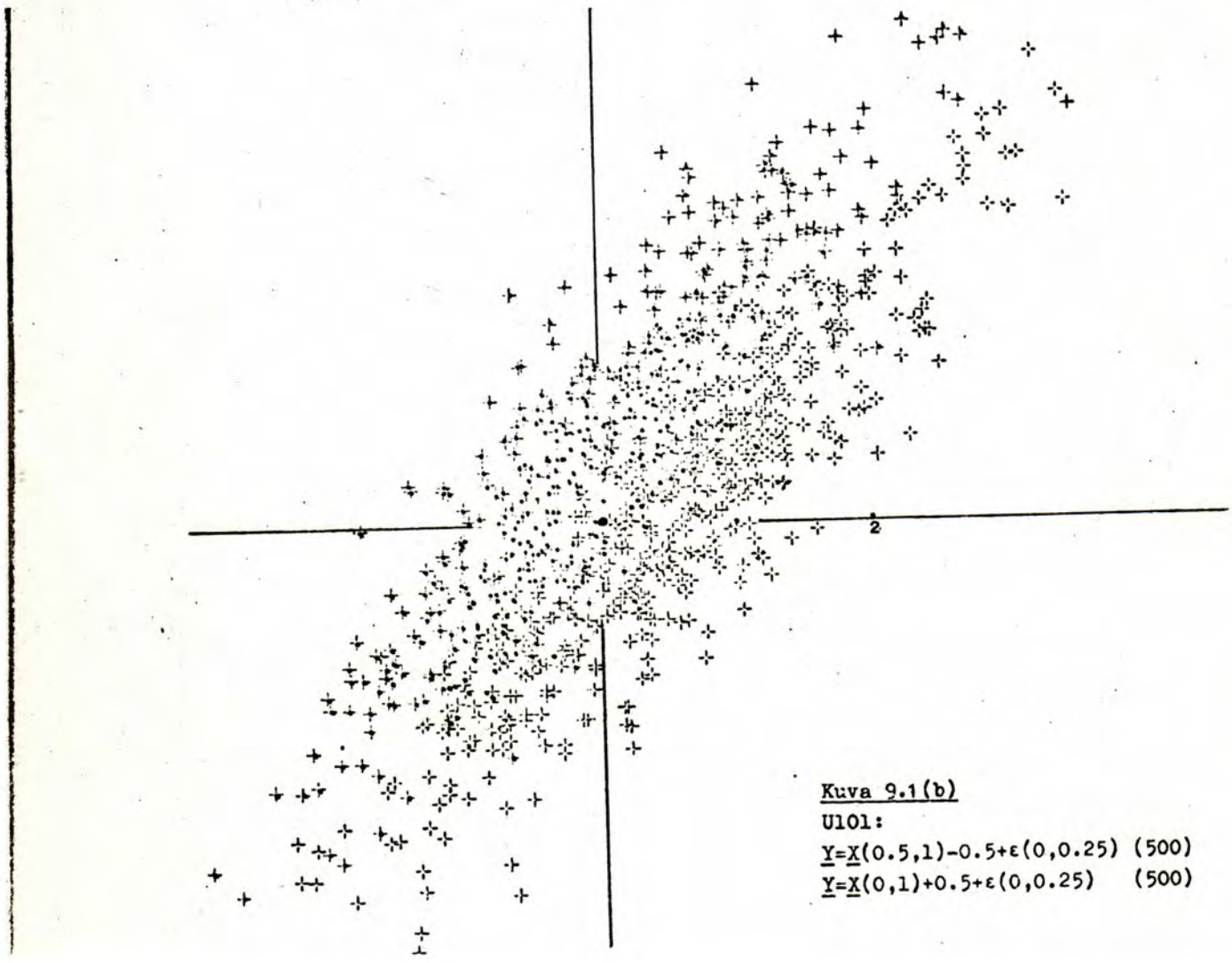
U101-U105:

$$\underline{Y}=\underline{X}(0.5,1)-0.5+\varepsilon(0,0.25)\ (500)$$
$$\underline{Y}=\underline{X}(0,\ 1)+0.5+\varepsilon(0,0.25)\ (500),$$

joiden osa-aineistot ryhmittyvät origon suhteen symmetrisesti suorien $y=x\pm 0.5$ ympärille. Koska odotusarvot ovat kahden hajonnan mitan päässä toisistaan, aineistot peittävät toisiaan yli 15%:n osuudelta.



Kuva 9.1(a)
U101:
$\underline{Y}=\underline{X}(0.5,1)-0.5+\varepsilon(0,0.25)$ (500)
$\underline{Y}=\underline{X}(0,1)+0.5+\varepsilon(0,0.25)$ (500)



Kuva 9.1(b)
U101:
$\underline{Y}=\underline{X}(0.5,1)-0.5+\varepsilon(0,0.25)$ (500)
$\underline{Y}=\underline{X}(0,1)+0.5+\varepsilon(0,0.25)$ (500)

Taulukko 9.2

	a	b	c	S_D	S_D/S_R	a'	b'	S_R
U101	0.979(.018)	-.541(.026)	.638(.027)	175.92	0.342	.903	.019	515.10
U102	1.005(.021)	-.601(.028)	.594(.032)	171.06	0.328	.924	-.008	521.03
U103	0.953(.016)	-.512(.021)	.630(.022)	169.93	0.347	.880	.067	490.22
U104	0.948(.017)	-.542(.024)	.583(.026)	150.82	0.326	.885	.036	463.12
U105	0.964(.022)	-.557(.025)	.596(.027)	157.31	0.327	.876	.027	481.18
(U101(B)	.893	-.658	.501	191.09	0.366	.923	-.009	521.62)

Tulokset ovat taulukossa 9.2. Suluissa on mainittu parametrien likimääräiset hajonnat.
Parametrin a suhteen tilanne näyttää aikaisemman kaltaiselta. Estimaatit b ja c ovat sen sijaan vahvasti harhaisia, kuten sopi odottaakin. Katsomme, vastaako harhan suuruus kohdassa 5 esitettyä tulosta. Koska nyt $\mu=1$ ja $\lambda=1.1666$ tämän harhan tulisi kummallakin olla noin 16%. Tilanne ei kuitenkaan näytä symmetriseltä, sillä b saa miltei systemaattisesti itseisarvoltaan suurempia arvoja kuin c. Tämä johtunee neljän aineiston painopisteen sijoittumisesta origon yläpuolelle, mikä näkyy b'-arvoista.
On luonnollista keskistää aineistot y-akselin suunnassa arvoon b' jokaisella aineistolla erikseen ja yrittää harhan poistamista jakamalla erotukset b-b' ja c-c' luvulla 1.1666, kuten on tehty seuraavassa taulukossa.

	b	c	b'	b-b'	c-c'	$\frac{b-b'}{\lambda/\mu}$	$\frac{c-c'}{\lambda/\mu}$
U101	-.541	.638	.019	-.560	.619	-.480	.531
U102	-.601	.594	-.008	-.593	.602	-.508	.516
U103	-.512	.630	.067	-.579	.543	-.496	.465
U104	-.542	.583	.036	-.578	.547	-.495	.469
U105	-.557	.596	.027	-.584	.569	-.501	.488
$\bar{x}$	-.551	.608		-.579	.576	-.496	.494

Näin korjatut estimaatit vaikuttavat harhattomilta. Todellisuudessa tulisi vielä näihin arvoihin lisätä b'. Käytännössä ei myöskään tunneta etukäteen heterogeenisuuden astetta eikä siis korjausvakio λ/μ ole tiedossa. Se on kuitenkin arvioitavissa suhteen S_D/S_R arvon perusteella.

Tässäkin tapauksessa havainnot on luokiteltu valikoivan pns-kriteerin avulla. Luokittelutulokset ovat:

	väärin luokiteltuja havaintoja (500:sta) osa-aineistossa 1	2	yht. %
U101	58	100	15.8
U102	79	96	17.5
U103	73	84	15.7
U104	89	90	17.9
U105	80	86	16.7
yht.	379	456	16.7

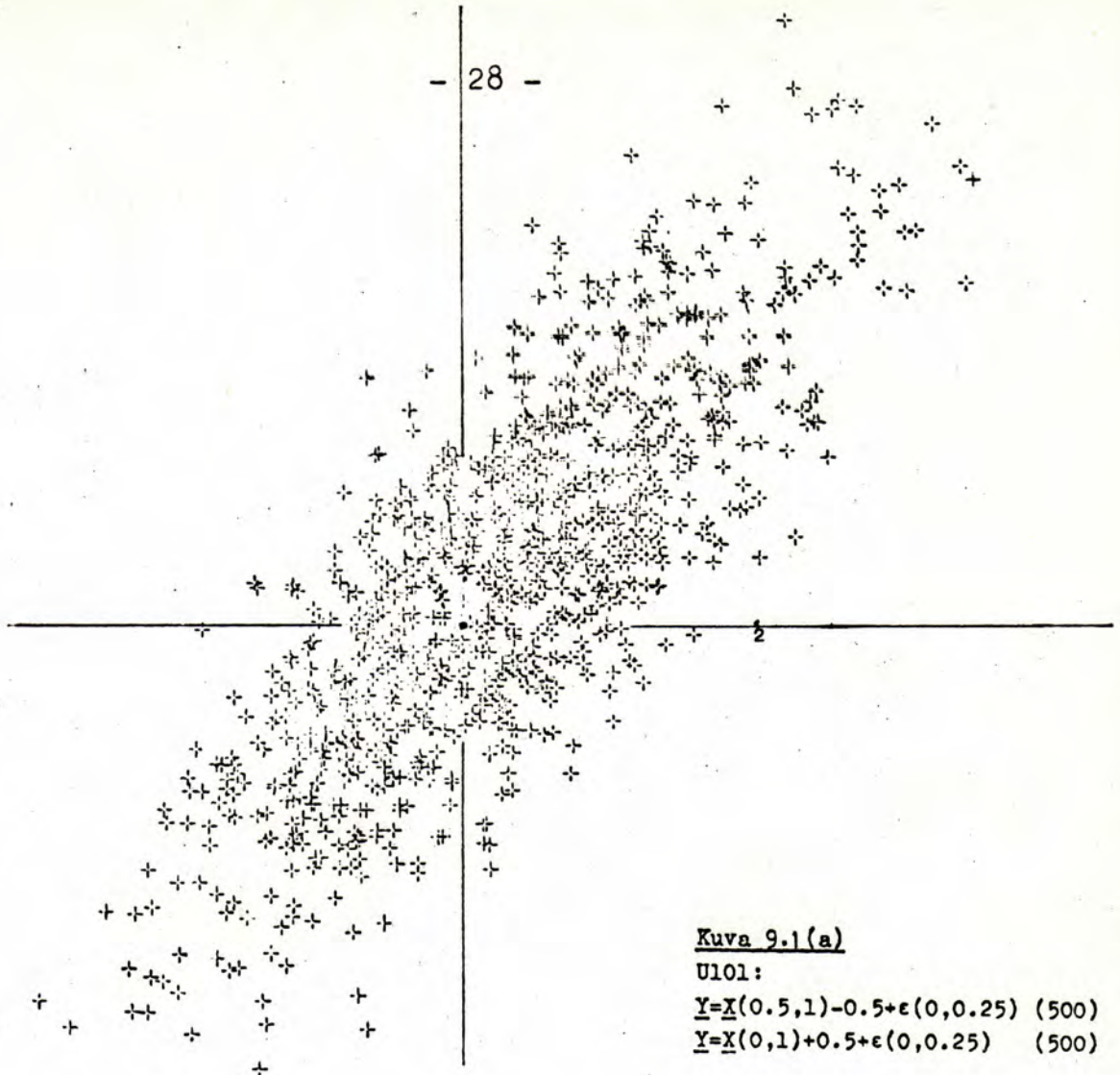
Tulokset vastaavat hyvin virheluokitusten teoreettista osuutta, joka on $100\Phi(-1)\% \approx 16.87\%$.

Taulukon 9.2 viimeinen rivi kuvaa homogeenista aineistoa U101(B), jonka rakenne on

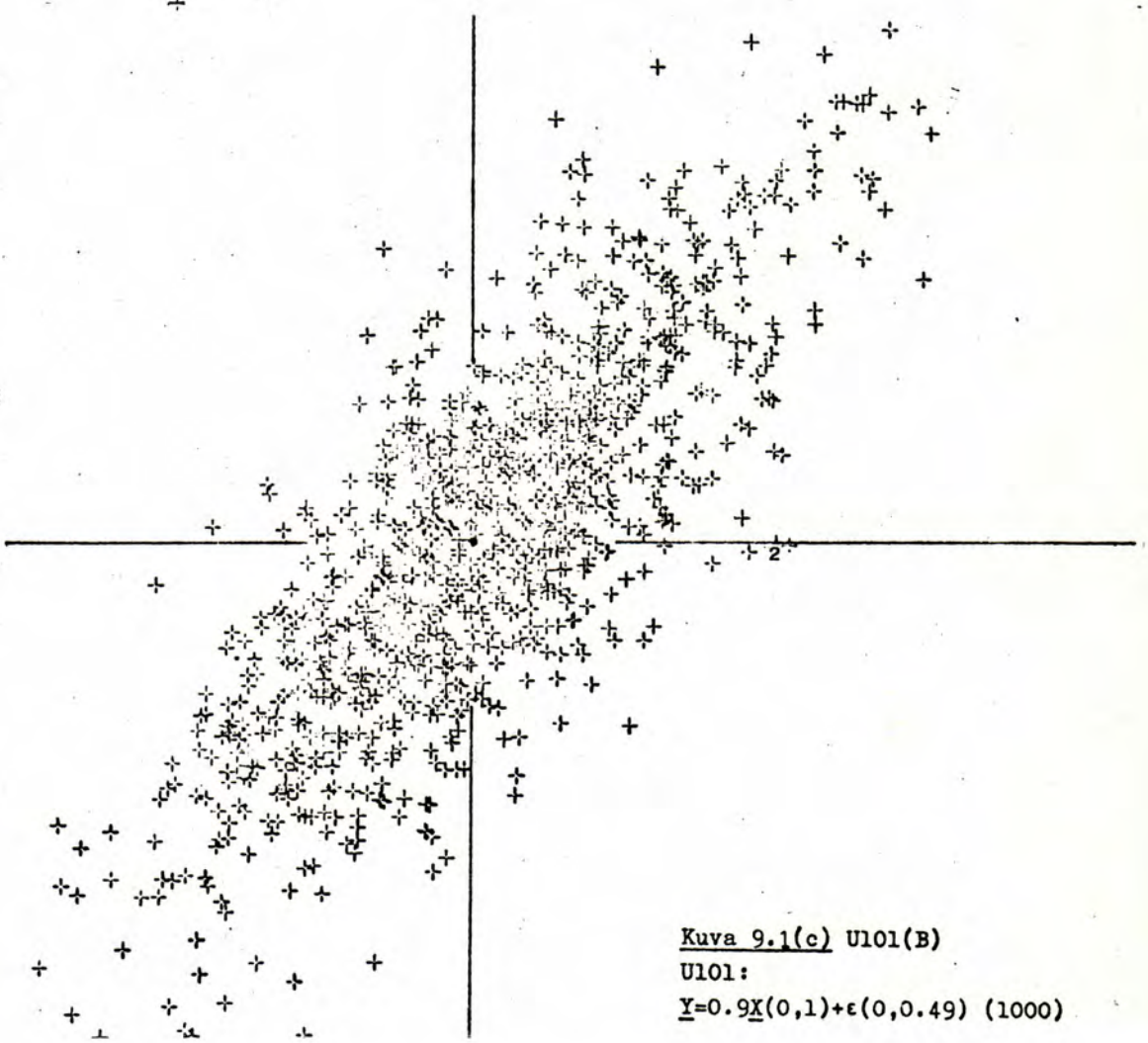
U101:

$$\underline{Y}=0.9\underline{X}(0,1)+\varepsilon(0,0.49) \quad (1000)$$

Se on siis saatavissa aikaisemmasta heterogeenisesta aineistosta U101 siirtämällä molempien osa-aineistojen havainnot suorien $y=x\pm0.5$ ympäriltä suoran $y=x$ ympärille. Samalla on kuitenkin satunnaisvirhettä kasvatettu niin paljon, että aineiston ominaisuudet ovat tavallisen lineaarisen mallin mielessä aikaisemman U101-aineiston kaltaiset. Kuvissa 9.1(a,c) nähdään molemmat aineistot. Lienee melko vaikea havaita silmämääräisesti mitään aivan vakuuttavia eroja tapausten rakenteen välillä. Aineiston U101(B) S_D/S_R-arvo 0.366 on kuitenkin suurempi kuin aineistojen U101-U105 ja vastaa hyvin homogeenisen jakauman teoreettista σ_D/σ_R-arvoa 0.363. Aineistolla U101-U105 suhteen S_D/S_R keskiarvo on 0.334 ja hajonta 0.010, mikä viittaa siihen, että jäännösneliösummasuhde toimii tässäkin tapauksessa heterogeenisuuden osoittimena. Pitäviä päätelmiä ei tietenkään voi tehdä näin rajoitetun kokeen pohjalta.



Kuva 9.1(a)
U101:
$\underline{Y}=\underline{X}(0.5,1)-0.5+\varepsilon(0,0.25)$ (500)
$\underline{Y}=\underline{X}(0,1)+0.5+\varepsilon(0,0.25)$ (500)



Kuva 9.1(c) U101(B)
U101:
$\underline{Y}=0.9\underline{X}(0,1)+\varepsilon(0,0.49)$ (1000)

9.2. Kahden lineaarisen mallin yhdistelmä

Huomion kohteena on nyt yleisempi digressiomalli

$$y=\begin{cases}\alpha_1 x+\beta_1+\varepsilon_1\\ \alpha_2 x+\beta_2+\varepsilon_2\end{cases}$$

eli kahden täysin erillisen lineaarisen mallin yhdistelmä. Tätä on tutkittu arvoilla $\alpha_1=1, \beta_1=0, \alpha_2=0.7, \beta_2=0.42$ käyttäen kymmentä 200 havainnon aineistoa

U501-U510:

$\underline{Y}=\underline{X}(0,1)+\varepsilon(0,0.01)$ (100)

$\underline{Y}=0.7\underline{X}(1.4,1)+0.42+\varepsilon(0,0.01)$ (100),

millä valinnalla on haluttu saada aikaan lievä epälineaarisuuden vaikutelma, mikäli uskoo, että aineisto on homogeeninen. Kuvassa 9.2 näkyy aineisto U501, jolle digressioanalyysilla saatiin tulos

$$y=\begin{cases}1.002x-0.0111\\ 0.706x+0.4074\end{cases},\quad S_D=1.477.$$

Jos aineisto sovitetaan tavalliseen lineaariseen malliin, saadaan

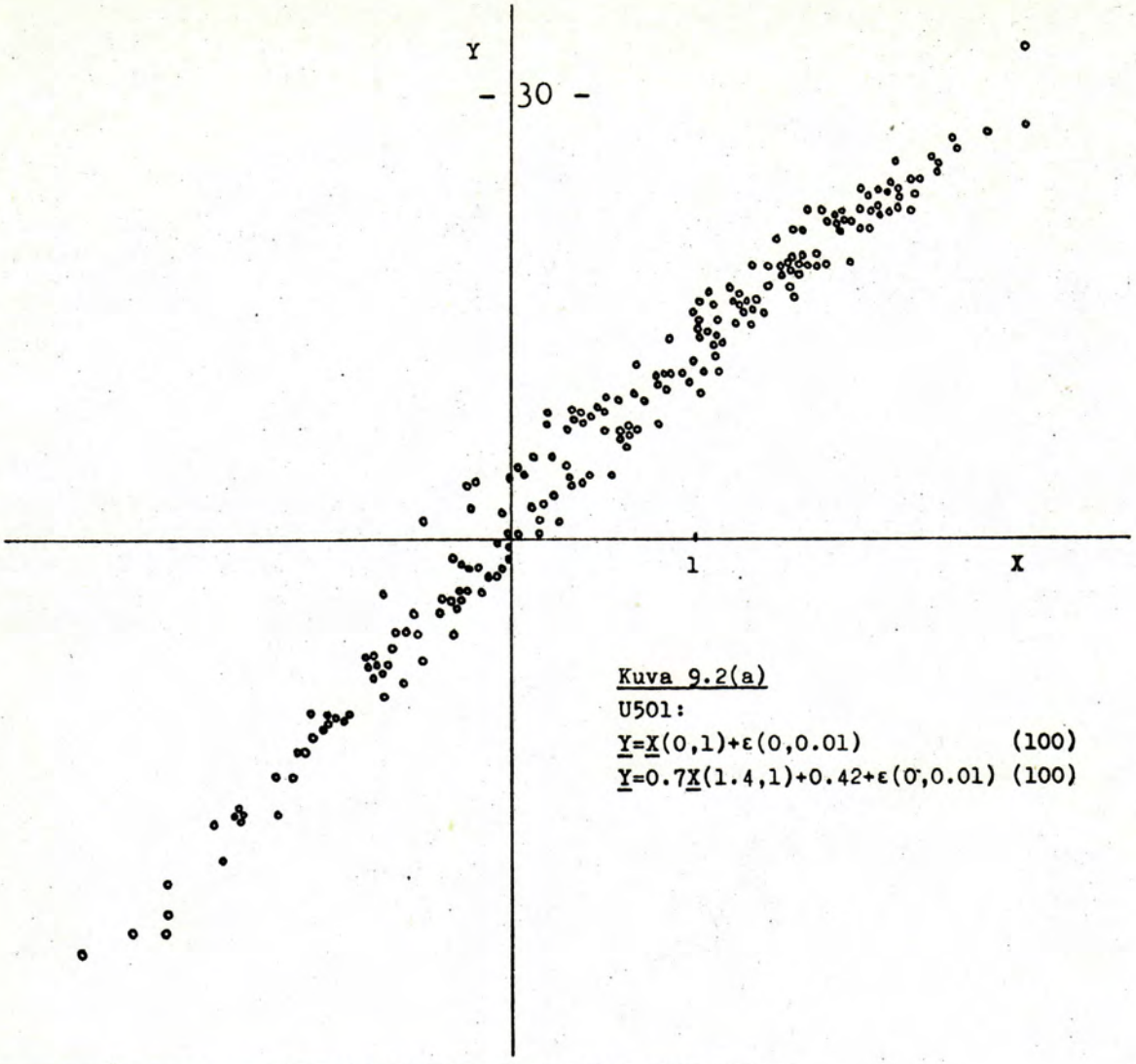
$$y=0.952x+0.0418,\quad S_R=5.787.$$

Lineaarisen mallin selitysastetta voidaan hieman parantaa käyttämällä vaikka kolmannen asteen polynomimallia, jolla tulos on

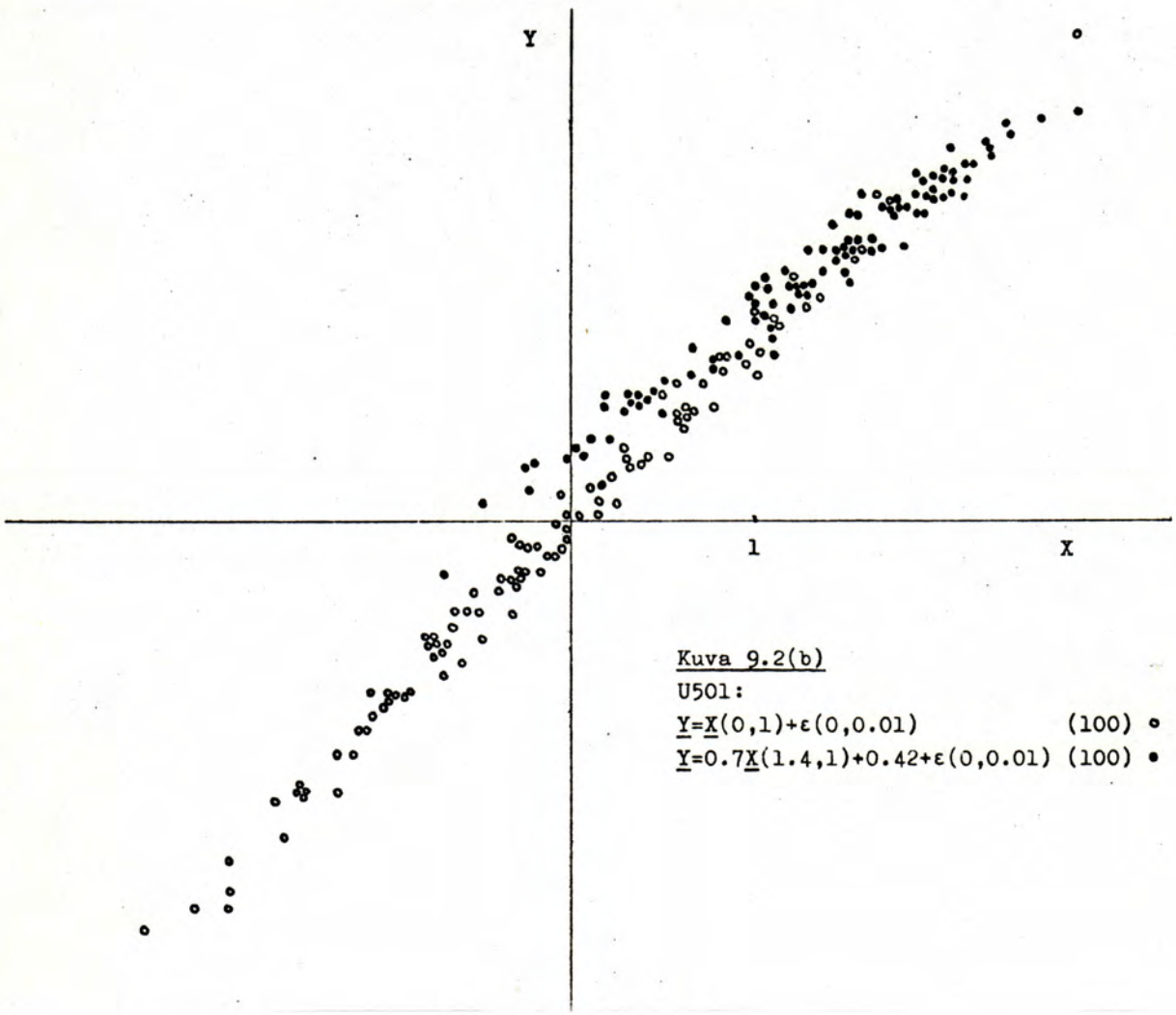
$$y=0.097+1.031x-0.0476x^2-0.0121x^3,\quad S_R=4.256,$$

mutta sekään ei yllä digressiomallin tasolle.

	Digressiomalli:					Regressiomalli:			
	a_1	b_1	a_2	b_2	S_D	a'	b'	S_R	S_D/S_R
U501	1.002	-.0111	.706	.407	1.477	.952	.042	5.787	0.255
U502	0.987	.0015	.691	.427	1.431	.906	.070	7.319	0.196
U503	0.992	-.0160	.696	.432	1.351	.897	.060	8.043	0.168
U504	0.994	.0092	.683	.438	1.344	.881	.088	7.052	0.191
U505	0.982	-.0371	.710	.394	1.566	.915	.046	7.359	0.213
U506	0.963	-.0223	.678	.448	1.579	.895	.058	7.562	0.209
U507	0.987	-.0039	.695	.433	1.790	.889	.073	8.155	0.219
U508	1.004	-.0037	.692	.422	1.494	.883	.093	8.968	0.167
U509	1.000	-.0100	.699	.422	1.625	.901	.075	7.514	0.216
U510	1.012	-.0057	.687	.434	1.878	.913	.032	7.089	0.265
$\bar{x}$	0.992	-.0099	.694	.426	1.554	.903	.064	7.485	0.210
s	0.014	.0130	.010	.016	0.175	.021	.020	0.832	0.032



Kuva 9.2(a)
U501:
$\underline{Y}=\underline{X}(0,1)+\varepsilon(0,0.01)$ (100)
$\underline{Y}=0.7\underline{X}(1.4,1)+0.42+\varepsilon(0,0.01)$ (100)



Kuva 9.2(b)
U501:
$\underline{Y}=\underline{X}(0,1)+\varepsilon(0,0.01)$ (100) ∘
$\underline{Y}=0.7\underline{X}(1.4,1)+0.42+\varepsilon(0,0.01)$ (100) •

Nyt ei ole lainkaan havaittavissa harhaa digressioestimaattien kohdalla. Analyysi siis näyttää toimivan moitteettomasti huolimatta siitä, että luokittelutulokset ovat melko vaatimattomat varsinkin toisen osa-aineiston kohdalla, jonka painopiste on osutettu ensimmäisen aineiston regressiosuoralle.

	väärin luokiteltuja havaintoja (100:sta) osa-aineistossa		yht.
	1	2	%
U501	9	26	17.5
U502	11	14	12.5
U503	4	18	11.0
U504	5	27	16.0
U505	10	15	12.5
U506	13	26	19.5
U507	7	19	13.0
U508	7	20	13.5
U509	7	20	13.5
U510	6	21	13.5
yht.	79	206	14.3

9.3. Vaihtuneet selittäjät

Pelkistettynä esimerkkinä usean muuttujan digressioanalyysista esitetään malli

$$y=\begin{cases}\alpha x_1+\beta x_2+\gamma+\varepsilon_1\\ \beta x_1+\alpha x_2+\gamma+\varepsilon_2,\end{cases}$$

joka antaa mahdollisuuden siihen, että osassa havaintoaineistoa selittäjien x_1 ja x_2 tehtävät ovat vaihtuneet. Tätä mallia on tutkittu vain aineistolla

U601:

$$\underline{Y}=\underline{X1}(0,1)+2\underline{X2}(=X1+z(0,1))+3+\varepsilon(0,0.7^2)\quad(50)$$
$$\underline{Y}=2\underline{X1}(0,1)+\underline{X2}(=X1+z(0,1))+3+\varepsilon(0,0.7^2)\quad(50)$$

eli siis arvoilla $\alpha=1,\beta=2,\gamma=3$ ja tehtävää on hieman vaikeutettu sillä, että selittäjät on tehty korreloituneeksi (korrelaatiokerroin noin $1/\sqrt{2}\approx 0.7$) valitsemalla $x_2=x_1+z$, missä x_1 ja z ovat riippumattomia 0,1-normaalisia muuttujia.
Lisäksi toista osa-aineistoa, havainnot 51-100, on vaiheittain supistettu, jotta saataisiin käsitys menetelmän toiminnasta erisuurilla osa-aineistoilla. Tulos eri havaintomäärillä on:

havainnot	1-100	1- 75	1- 60
a	0.974	0.968	0.829
b	2.127	2.116	2.141
c	2.866	2.883	2.943
S_D	23.53	17.68	12.54

Tarkastelemme tilannetta vielä lähemmin täydellä havaintomäärällä 100. Osa-aineistojen pahasta päällekkäisyydestä johtuen havaintojen luokittelu ei voi onnistua hyvin. Virheluokitusten osuus onkin 37%. Huomattakoon kuitenkin, että vaikka parametrien oikeat arvot tunnettaisiin, virheluokitusten määrä olisi silti 34%. Näinkään suuri luokittelun epävarmuus ei tässä tapauksessa näytä pahemmin häirinneen parametrien estimointia.
Jos vertailun vuoksi vielä koko aineisto sovitetaan tavalliseen lineaariseen malliin, eikä siis lainkaan oteta huomioon mahdollisia selittäjien roolinvaihdoksia, saadaan tulos

$$y=1.586x_1+1.592x_2+2.890,$$

joten vain vakiotermi saa oikeantuntuisen arvon.

9.4. Epälineaariset osamallit

Viimeisenä käsitellään epälineaarista digressiomallia

$$y=\begin{cases}\alpha x e^{-\beta x}+\varepsilon_1\\ \alpha x e^{-\delta x}+\varepsilon_2.\end{cases}$$

Tässä kokeessa satunnaisuuden astetta vaihdeltu käyttäen lähtökohtana aineistoa ($\alpha=1, \beta=1, \delta=1.1$)

U701:
$\underline{X}$(tas.(0,3)) (100)
$\underline{Y}=Xe^{-X}+\varepsilon(0,\sigma^2)$ (50)
$\underline{Y}=Xe^{-1.1X}+\varepsilon(0,\sigma^2)$ (50).

Selittäjän x arvot on siis saatu välin (0,3) tasaisesta jakaumasta ja selitettävän y arvoja rasittaa normaalinen satunnaisvirhe, jonka hajonta on σ. Tälle hajonnalle on käytetty arvoja $\sigma=0.02, 0.05, 0.1$. Kuva 9.3 esittää tapausta $\sigma=0.02$. Kuvassa 9.3(b) näkyvät myös regressiokäyrät. Tulokset ovat:

σ	a	b	c	S_D
0.02	0.993(.032)	0.989(.021)	1.104(.021)	0.0241
0.05	1.037	0.974	1.175	0.1142
0.1	1.085	0.927	1.302	0.4352

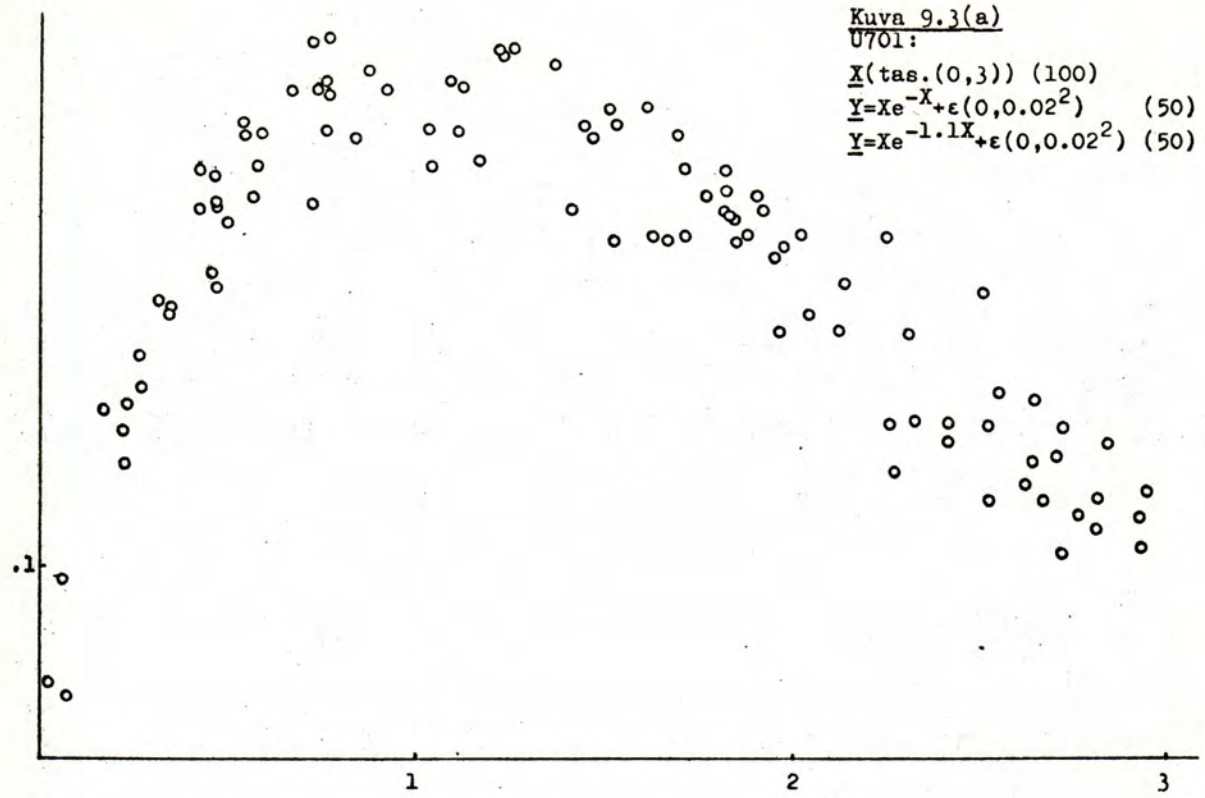
Tarkastelemme pienimmän satunnaisvaihtelun $\sigma=0.02$ omaavaa aineistoa hieman tarkemmin. Parametrien hajontojen ohella on laskettu myös niiden korrelaatiot

	a	b	c
a	1	0.97	0.94
b	0.97	1	0.91
c	0.94	0.91	1

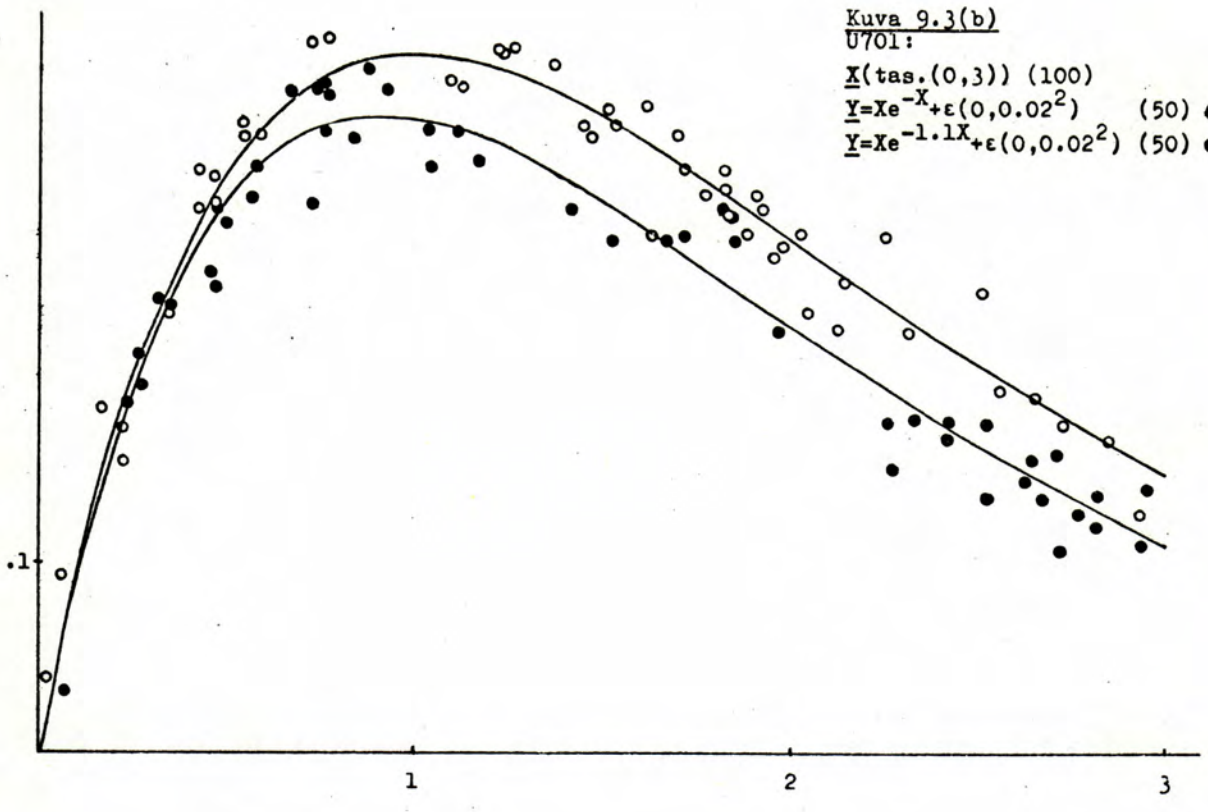
joiden suuruudesta huolimatta malli näyttää löytäneen suhteellisen tarkat parametrien arvot. Havaintojen luokittelutarkkuus on noin 20% kummassakin osa-aineistossa. Aineisto on sovitettu myös regressiomalliin

$$y=\alpha' x e^{-\beta' x}+\varepsilon$$

ja tulos on $a'=1.003, b'=1.041, S_R=0.0702$.



Kuva 9.3(a)
U701:
X(tas.(0,3)) (100)
$Y=Xe^{-X}+\varepsilon(0,0.02^2)$ (50)
$Y=Xe^{-1.1X}+\varepsilon(0,0.02^2)$ (50)



Kuva 9.3(b)
U701:
X(tas.(0,3)) (100)
$Y=Xe^{-X}+\varepsilon(0,0.02^2)$ (50) o
$Y=Xe^{-1.1X}+\varepsilon(0,0.02^2)$ (50) •

Katsomme vielä, mitä tapahtuu vastaavanlaiselle homogeenista alkuperää olevalle aineistolle samassa analyysissa. Aineisto

U701:

$\underline{X}$(tas.(0,3)) (100)

$\underline{Y}=Xe^{-1.05X}+\varepsilon(0,0.03^2)$ (100)

tuottaa malliin $y=\alpha' xe^{-\beta' x}+\varepsilon$ sovitettuna tuloksen a'=1.013, b'=1.052, S_R=0.0791 ja digressiomalliin sovitettuna a=1.007, b=0.997, c=1.102, S_D=0.0396.

Odotetusti digressiomalli jakaa tämänkin aineiston kahtia jopa samannäköisin parametrinarvoin kuin heterogeenisessa U701-aineistossa. Kuitenkin jälleen jäännösneliösumman S_D suuruus paljastanee eron; S_D/S_R-arvo 0.501 lienee merkitsevästi suurempi kuin heterogeenisella aineistolla saatu S_D/S_R=0.343.

10. Loppupäätelmiä

Edellä kuvatut kokeet antanevat melko rohkaisevan käsityksen digressioanalyysin toimivuudesta. Huolimatta kohtalaisen suurista, aineistojen päällekkäisyydestä johtuvista havaintojen luokitteluvirheistä analyysi näyttää toimivan tyydyttävästi parametrien estimoinnissa. Luonnollisena selityksenä tälle ilmiölle tarjottiin jo alussa sitä, että virheluokitukset kohdistuvat etupäässä osamallien väliin jääviin neutraaleihin havaintoihin, jotka vaikuttavat regressiomallin parametrien estimaatteihin vähemmän kuin vaihtelualueen reunoilla olevat havainnot. Olen kokeellisesti todennut, että tavallisessa kahden muuttujan regressioanalyysissa saattaa poistaa selittävän muuttujan vaihtelualueen keskeltä vaikkapa puolet koko havaintomäärästä ilman, että estimaattien tarkkuus kärsii olennaisesti.

Luokitusvirheet saattavat kuitenkin vaikeuttaa digressioanalyysia aiheuttamalla harhaa joihinkin estimaatteihin. Joissakin tapauksissa tuo harha on arvioitavissa ja ehkä poistettavissa jäännösneliösummavertailujen avulla, mutta näistä mahdollisuuksista ei ole toistaiseksi riittävää näyttöä.
Hankalin ongelma käytännössä lienee se, että digressioanalyysi saattaa nähdä heterogeenisuutta sellaisissakin aineistoissa, joissa sitä ei taatusti ole. Eräs tämän liioitellun digressioefektin ilmentymähän on juuri eräiden estimaattien harhaisuus. Tätä heikkoutta on vaikea poistaa. On sen tähden tärkeätä kehittää luotettava menettely valeheterogeenisuuden erottamiseksi oikeasta. Eräät edellä esitetyt koetulokset osoittavat, että jäännösneliösummasuhde S_D/S_R vaikuttaa hyvältä kriteeriltä. Tulisi kuitenkin hallita sen jakauma eri digressioprobleemoissa. Ainoa toimiva keino on toistaiseksi tämän jakauman simulointi.

Monet mielenkiintoiset kysymykset ovat tässä vaiheessa jääneet käsittelemättä.
Joissakin sovellutuksissa osa havainnoista on ehkä tunnistettavissa varmasti tai tietyllä todennäköisyydellä tunnetaan niiden alkuperä. Olisi paikallaan selvittää, miten tällainen lisätietous saataisiin käyttöön. Eräs mahdollisuus on painottaa havaintoja valikoivaa pns-kriteeriä käytettäessä tämän tunnistusvarmuuden

mukaisesti.
Itse digressiokriteeriä, valikoivaapns-keinoa, voi monella tapaa muuntaa. Olen kokeillut jonkin verran itseisarvosummakriteeriä, mutta tulokset ovat olleet ristiriitaisia menettelyjen vertailun kannalta. Koska poikkeukselliset havainnot saattavat pahasti vääristää tuloksia, kriteeriä on joskus ehkä syytä täydentää pahimmat mittausvirheet paljastavilla lisäehdoilla. Virheluokitusten aiheuttamia ongelmia saatetaan vähentää painottamalla havaintoja siten, että paino riippuu havainnon sijainnista eri osamallien suhteen. Paino on esim. sitä pienempi, mitä voimakkaammin osamallit kilpailevat havainnosta.
Laskennallisella puolella on hyödyllistä kehittää algoritmeja, jotka paremmin kuin yleiset epälineaariset optimointimenetelmät ottavat huomioon ongelman erityispiirteet. En pidä mahdottomana, etteikö kohdassa 6 esitetyssä esimerkissä mainittua ratkaisutapaa voitaisi yleistää.
Tässä tutkimuksessa digressioanalyysia on käsitelty lähinnä regressioanalyysin yleistyksenä ja parametrien estimointimenetelmänä. Sen käyttöön klusterointikeinona ei ole paljon puututtu. Tuntuu kuitenkin mahdolliselta, että regressiomallien avulla tapahtuva muuttujien riippuvuuden tavallista tarkempi huomioonottaminen saattaa olla hyödyksi myös puhtaassa havaintojen luokittelussa. Yleensähän luokittelumenetelmät toimivat tässä suhteessa melko löysien sääntöjen mukaan. Esim. aineistoa U101 (kuva 9.1) tuskin mikään tavanomainen klusterointimenettely uskaltaisi pilkkoa kahteen osaan.
Näissä alustavissa tarkasteluissa on tarkoituksellisesti tyydytty käyttämään vain keinotekoisia aineistoja, jolloin tositilanteen tuottamat lisäongelmat on saatu pysymään loitolla ja tutkimus on voinut keskittyä analyysin perusominaisuuksiin. On selvää, että tällaisen analyysitekniikan hiominen edellyttää kuitenkin myös kokeilua tosiaineistoilla.

Lähdeluettelo

Bradley, Ralph A., Gart, John J.(1962): The asymptotic properties of ML estimators when sampling from associated populations. Biometrika 49, 205-213.

Fukunaga, Keinosuke (1972): Introduction to statistical pattern recognition. Academic Press, New York.

Goldfeld, Stephen M., Quandt, Richard E.(1972): Nonlinear methods in econometrics. North Holland, Amsterdam.

Hill, William J., Hunter, William G., Wichern, Dean W.(1968): A joint design criterion for the dual problem of model discrimination and parameter estimation. Technometrics 10, 145-160.

Mihram, G.Arthur (1972): Simulation. Statistical foundations and methodology. Academic Press, New York.

Walsh, G.R.(1975): Methods of optimization. Wiley, New York.

Liite Koeaineistojen generointi

Jotta tässä tutkimuksessa suoritetut kokeet olisivat helposti toistettavissa ja jatkettavissa, kaikki simuloidut aineistot luotiin seuraavalla yhtenäisellä menettelyllä. Lähtökohtana on "sekakongruenssiperiaatteella" toimiva pseudosatunnaisluku-generaattori (kts. esim. Mihram, 1972)

$$u_n = au_{n-1} + c \pmod{m},$$

joka tuottaa täyden jakson m mittaisen kokonaislukujonon $u_0=0, u_1, u_2, \ldots, u_{m-1}$, mikäli

1) c ja m ovat keskenään jaottomia,
2) $a \equiv 1 \pmod{p}$ jokaiselle luvun m alkutekijälle p,
3) $a \equiv 1 \pmod{4}$.

Lukuja $u_n/m, n=1,2,\ldots$ voidaan tällöin pitää riippumattomina ja tasaisesti jakautuneina välillä (0,1).
Jotta aineistot voitaisiin luoda helposti hyvin rajoitetunkin sananpituuden omaavalla koneella, generaattorin parametreiksi valittiin $m=10^5$, jolloin ehdot 2-3) pelkistyvät muotoon $a \equiv 1 \pmod{20}$, $c=9^5=59049$, $a=a_j=20(1000+j)+1, j=1,2,\ldots,3000$.
Generaattorista

$$u_n = a_j u_{n-1} + 9^5 \pmod{10^5}$$

käytetään tunnusta Uj. Generaattoriperhettä U1,U2,...,U3000 ei ole erikseen perusteellisesti testattu.

Generaattorien U1-U6 "tuotantoa":

	U1	U2	U3	U4	U5	U6
1	59049	59049	59049	59049	59049	59049
2	79078	60058	41038	22018	2998	83978
3	79687	81427	22367	2507	21847	80387
4	72476	37556	63436	2116	5596	25876
5	1045	18845	48645	50445	44245	10045
6	80994	31694	26394	45094	27794	74494
7	39923	38503	49083	91663	46243	52823
8	57432	97672	13112	43752	89592	10632
9	5121	3601	98881	42961	47841	85521
10	86590	26690	10790	58890	10990	27090

Normaalisti jakautuneet muuttujat generoitiin Mullerin log-trig-muunnoksella

$$x=(-2\log u)^{1/2}\cos(2\pi v)$$
$$y=(-2\log u)^{1/2}\sin(2\pi v),$$

joka muuntaa riippumattomat, tasaisesti yli välin (0,1) jakautuneet muuttujat u,v riippumattomiksi 0,1-normaalisiksi muuttujiksi. Näin luotuja normaalisia muuttujia ja niistä lineaarimuunnosten avulla rakennettuja johdettuja muuttujia merkitään $X(\mu,\sigma^2)$, missä μ=odotusarvo ja σ^2=varianssi.
Koeaineistot tehtiin yhdistelemällä näin syntyneitä muuttujia käytössä olevan mallin mukaisesti. Aineistoista käytetään seuraavanlaisia koodimerkintöjä:
Esim.

U5:
$\underline{Y}=2\underline{X}(0,1)+3+\varepsilon(0,0.25)$ (100)

tarkoittaa generaattorin U5 avulla luotua 100 havainnon ja kahden muuttujan y,x aineistoa, joka jäljittelee mallia

$$y=2x+3+\varepsilon,$$

missä $x\sim N(0,1)$ ja $\varepsilon\sim N(0,0.25)$. Osoituksena siitä, että x ja y ovat havaittuja muuttujia, niitä vastaavat symbolit X ja Y on alleviivattu aineiston koodimerkinnässä.
Aineistot generoidaan havainnoittain ja havaintojen sisällä selittäjien ja virhetermin arvot valitaan siinä järjestyksessä, jossa ne esiintyvät mallin merkinnässä.
Edellä U5-generaattorilla määrätyn esimerkkiaineiston ensimmäiset havainnot ovat:

	y	x
1	3.8886	0.1922
2	5.0202	0.6007
3	5.4033	1.2574
4	1.9818	-0.7555
5	5.0150	0.7734
6	4.4916	0.5888
7	5.7582	1.1057
8	4.7961	1.0359
9	2.4695	-0.5273
10	5.0445	0.8069