SOSIAALITUTKIMUKSEN TILASTOLLISET MENETELMÄT 16.1.–26.2.2014.
Regressioanalyysi. Yliopistonlehtori Pekka Pere.

1 Regressio kohti odotusarvoa

Francis Galtonin ensimmäinen regressio vuodelta 1877 on hajontakuviossa ohessa ("herneen siemen -vanhemmat" ja "herneen siemen -jälkipolvi").[1] Oleellisesti sama ilmiö on seuraavassa kuviossa — ilmeisesti toisessa koskaan tehdyssä regressiossa[2] — joka kuvaa Galtonin 1886 havaitsemaa vastaavaa yhteyttä vanhempien pituuksien painotetun keskiarvon (mid-parent; "keskivanhempi") ja heidän lastensa pituuden (children) välillä.[3] Kuviosta nähdään, että vanhempien keskipituus on ollut keskimäärin runsas 68 tuumaa. Keskivanhempi-suora kuvaa vanhempien keskipituuden poikkeamaa keskimääräisestä pituudesta (suoran kulmakerroin on yksi). Kuvion mukaan

- keskimääräistä pidempien vanhempien lapsi on myös keskimääräistä pidempi muttei yhtä paljon kuin vanhempansa (suoran "children" kulmakerroin on 0:n ja 1:n välillä).
- keskimääräistä lyhyempien vanhempien lapsi on myös keskimääräistä lyhyempi muttei yhtä paljon kuin vanhempansa.
- pituus regressoituu (palautuu, taantuu) eli pyrkii palaamaan kohti odotusarvoansa (yllä runsas 68 tuumaa). (*Regression toward the mean* tai *regression to the mean.*) Pitkien vanhempien lapset ovat keskimääräistä pidempiä ja lyhyempien vanhempien lapset keskimääräistä lyhyempiä, mutteivät yhtä paljon pidempiä tai lyhyempiä keskipituuteen nähden kuin vanhempansa.

Mieleen saattaisi tulla — kuten Galtonille aikoinaan — että regressiosta keskipituutta kohti seuraisi sukupolvi sukupolvelta pituuden vaihtelun pieneneminen niin, että lopulta kaikki olisivat keskipituisia. Niin ei käy, koska lasten pituuksissa on aina sattumanvaraisuutta, vaikka lasten pituus keskimäärin

[1] Kuvio on artikkelista Nicholas Gillham (2009): Cousins, Charles Darwin, Sir Francis Galton and the Birth of Eugenics. *Significance*, 6, 132–135. Kuvion alkuperäislähde on Karl Pearson (1920): Notes on the History of Correlation. *Biometrika*, 13, 25–45. Kuvio löytyy myös kirjasta Karl Pearson (1930): *The Life, Letters and Labours of Francis Galton, nide III A*. Cambridge University Press, Cambridge. Regressiosuora on Pearsonin uusiksi laskema ja ilmeisesti hänen apulaisensa (A. Davinin) piirtämä Galtonin muistiinpanojen vuodelta 1875 avulla. (Pearson 1920, 34 ja 1930, 4.)

[2] Pearson (1930, 13).

[3] Midparent-käsitteen määritelmä löytyy esimerkiksi Wikipediasta (http://en.wikipedia.org/wiki/Midparent; viitattu 26.1.2014). Kuvio on kirjasta (s. 16) Karl Pearson (1930): *The Life, Letters and Labours of Francis Galton, nide III A*. Cambridge University Press, Cambridge. Kuvio on julkaistu alunperin artikkelissa Francis Galton (1886): Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

regressoituukin vanhempiensa pituudesta. Galton havainnollisti asiaa oheisella kuviolla vuonna 1901.[4] Galtonin keskivanhempi-käsitettä käytetään kasvututkimuksessa edelleen (esim. Saari ym. 2012).[5]

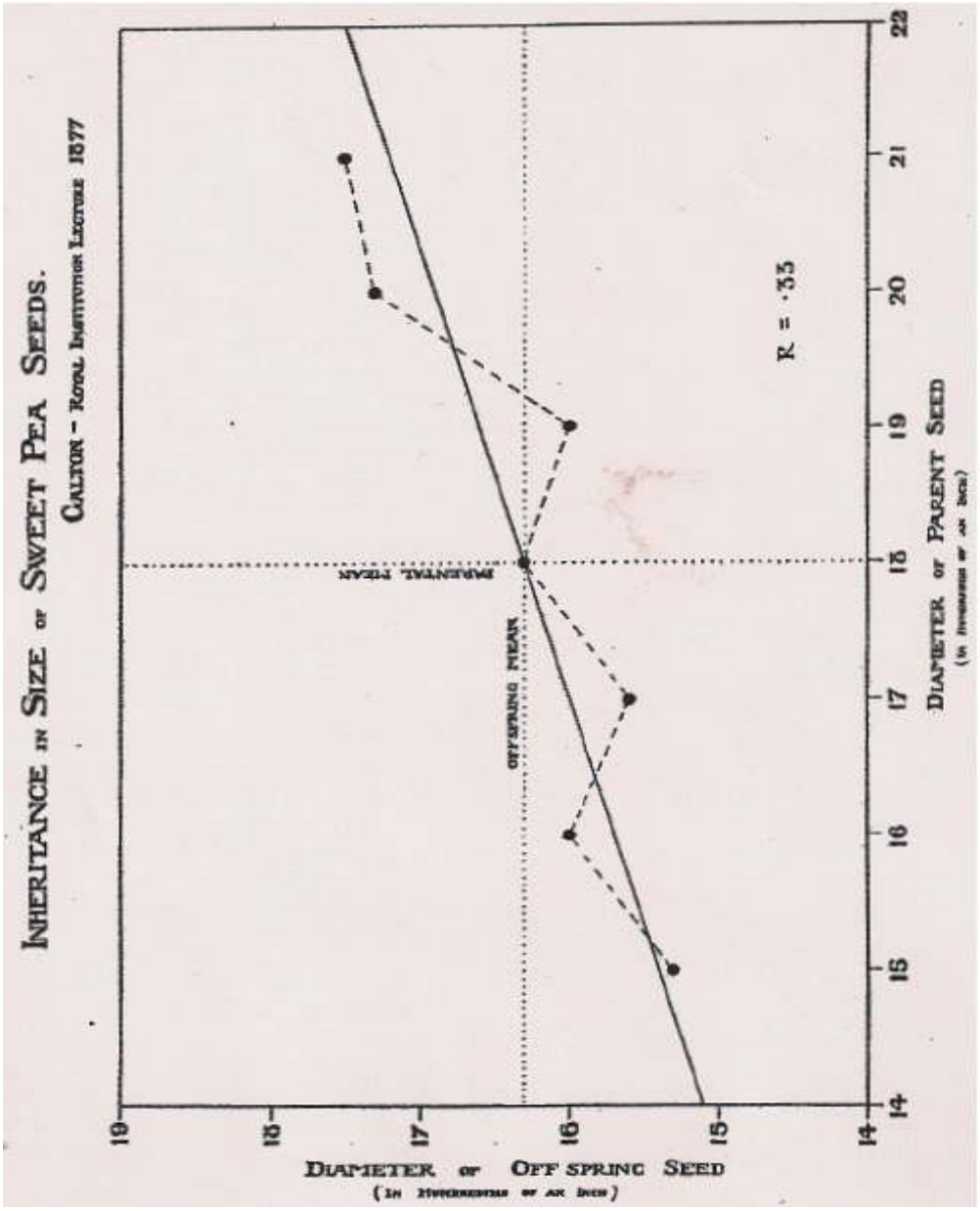
Regressiossa odotusarvoa kohti on yleisemmin kyse kahdesta samoinjakautuneesta muuttujasta, joiden yhteyden summeeraavan suoran kulmakerroin on alle yhden. Ääritilanteessa muuttujien välillä ei ole mitään yhteyttä: Kuvitteelliseen kuvioon piirretyn summeeraavan sovitteen kulmakerroin on nolla, ja poikkeamat odotusarvosta pyrkivät keskimäärin "korjaantumaan" täysin seuraavassa havainnossa.

Nykypäivään ja yhteiskuntatieteisiin liittyviä esimerkkejä on helppo keksiä. Verrataan sosiaalitukia saavien (tai rikosten, avioerojen, syntyneiden lasten jne.) lukumäärää suomalaisissa kaupungeissa vuosina 2014 (y-akseli) ja 2013 (x-akseli). Tällöin poikkeuksellisen suuri sosiaalitukea saavien määrä tietyssä kaupungissa tasoittuu lähemmäksi odotusarvoa seuraavana vuonna. Vastaavasti tavanomaista pienemmästä sosiaalitukea nauttivien lukumäärästä vuonna 2013 ilahtuneet kaupunginjohtajat joutuvat tyypillisesti pettymään, kun sosiaalitukea haetaan vuonna 2014 edellistä vuotta enemmän.

Edelläkuvatunkaltainen ilmiö on "voittajan kirous" (*winner's curse*): Kun suuresta joukosta esimerkiksi työpaikan tai urheilujoukkueen jäsenyyden hakijoista poimitaan suorituksiltaan paras, ei valittu yllä aivan aiempien suoritustensa mukaiseen tulokseen.

[4] Kuvio on artikkelista Warren Gilchrist (2012): Galton — A Victorian Worth Celebrating, Significance Web Exclusive (http://www.significancemagazine.org/details/webexclusive/1497449/Galton—A-Victorian-worth-celebrating.html; viitattu 26.1.2014).

[5] Antti Saari, Ulla Sankilampi, Marja-Leena Hannila, Marja-Terttu Saha, Outi Mäkitie ja Leo Dunkel (2012): Screening of Turner Syndrome with Novel Auxological Criteria Facilitates Early Diagnosis. *Journal of Clinical Endocrinology & Metabolism*, 97, 2125–2132.


INHERITANCE IN SIZE OF SWEET PEA SEEDS.
GALTON - ROYAL INSTITUTION LECTURE 1877
R = ·33
MEDIAN
OFFSPRING MEAN
DIAMETER OF OFF SPRING SEED
DIAMETER OF PARENT SEED
14
15
16
17
18
19
20
21
22

Kuva 1:

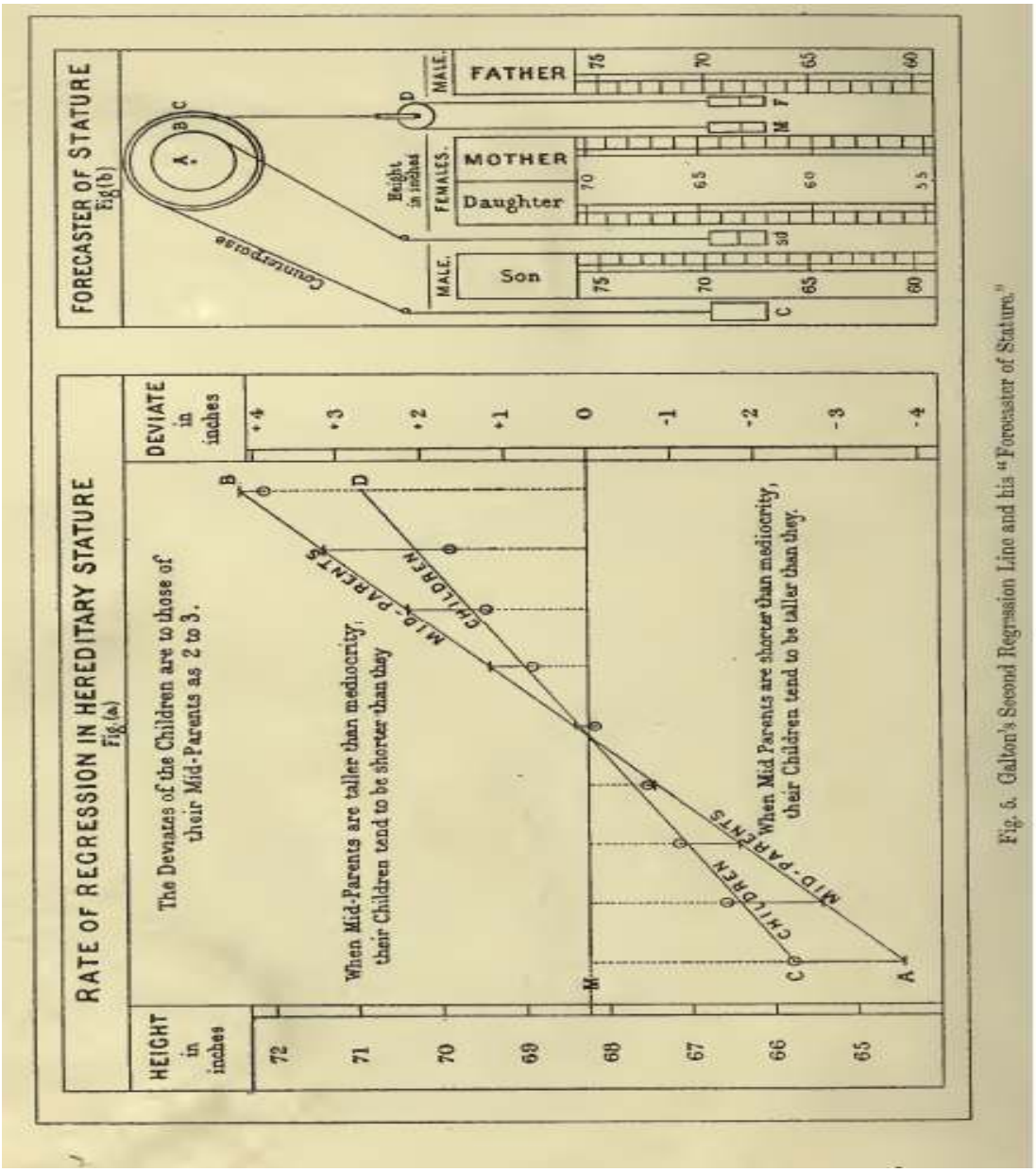
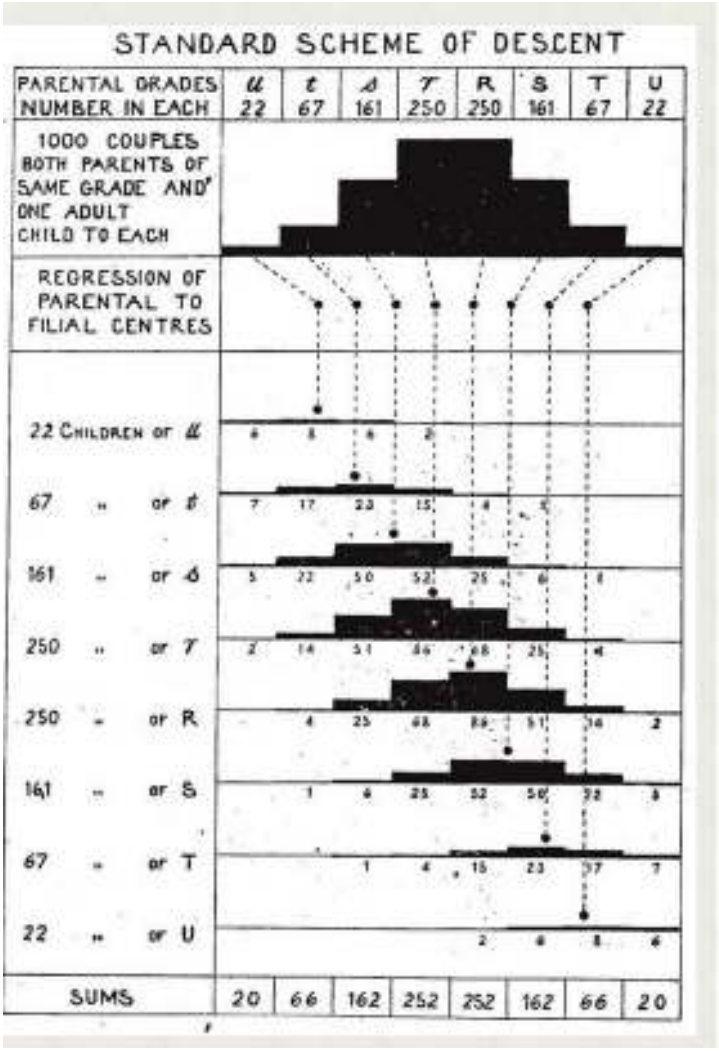


Fig. 6. Galton's Second Regression Line and his "Forecaster of Stature."

Kuva 2:


STANDARD SCHEME OF DESCENT
PARENTAL GRADES
NUMBER IN EACH
α 22
ι 67
β 161
γ 250
R 250
S 161
T 67
U 22
1000 COUPLES BOTH PARENTS OF SAME GRADE AND ONE ADULT CHILD TO EACH
REGRESSION OF PARENTAL TO FILIAL CENTRES
22 CHILDREN OF α
67 " OF ι
161 " OF β
250 " OF γ
250 " OF R
161 " OF S
67 " OF T
22 " OF U
SUMS
20
66
162
252
252
162
66
20

Kuva 3:

Tehtävä! Taulukossa on neljästä kuvitteellisesta helsinkiläisestä sosiaalitoimistosta jaetun tuen määrä euroissa viime vuonna. Sosiaalitoimistot ovat yhtäsuuria monilla muilla kriteereillä mitattuina (henkilökunnan lukumäärä, pinta-ala jne.). Erot jaettujen tukien määrässä johtuvat vaihteluista sosiaalisissa ongelmissa kaupunginosittain ja muista sattumanvaraisista tekijöistä. Tänä vuonna sosiaalitoimistojen jakaman tuen tiedetään kasvavan 10 prosentilla — esimerkiksi yleisen valtakunnallisen taloustilanteen takia. Onko järkevää ennustaa, että myös kunkin sosiaalitoimiston jakama tuki kasvaa tänä vuonna 10 prosenttia? Perustele.[6]

sosiaali-toimisto	**jaettu tuki viime vuonna (milj. e)**	**jaettu tuki tänä vuonna (milj. e)**
1	11	
2	23	
3	18	
4	29	
yhteensä	81	89,1

Edellä vertailtiin kahden satunnaismuuttujan yhteyttä, kun ne ovat samoinjakautuneita ja niiden hajontakuvioon piirretyn muuttujien välisen systemaattisen komponentin summeeraavan suoran kulmakerroin on (itseisarvoltaan) alle yksi (tyypillisesti oleellisesti sama satunnaismuuttuja jossain mielessä kahdesti mitattuna). Regressioanalyysissa (jakso 3) voidaan sallia useampia muuttujia, jotka voivat olla erilailla jakautuneita tai kiinteitä, eikä systemaattisen vaikutuksen suuruutta tarvitse rajoittaa edelliseen tapaan. (Tällöin ei voida puhua regressiosta odotusarvoa kohti aivan samassa mielessä kuin edellä.) Regressioanalyysillä pyritään selvittämään systemaattisuus yhden muuttujan ja muiden muuttujien välillä. Aina regressioanalyysissa on kuitenkin kyse pohjimmiltaan samasta ilmiöstä kuin edellä eli että osa havaintojen käyttäytymisestä on systemaattista ja osa sattumaa. Sattuman vaikutus tulisi regressoida "pois" muuttujien välistä yhteyttä arvioitaessa. Esimerkiksi Galtonin tutkimusaineistossa lasten ja vanhempien pituuksien suhteella on geneettinen (systemaattinen) selitys, mutta osin lasten pituudet johtuvat (tutkijan näkökulmasta) sattumanvaraisista seikoista kuten nimenomaisista geeneistä, jotka lapsi on perinyt, lapsen saaman ruoan ravinnepitoisuudesta tai sairastamista taudeista, kellonajasta, jolloin lapsi on mitattu (aamulla lapsi on pidempi) ja niin edelleen. On vain hieman liioiteltua sanoa, että lähes asiassa kuin asiassa on regressiota. Campbellin ja Kenny'in (1999, ix) mukaan regressio odotusarvoa kohti on yhtä väistämätön asia kuin verot tai kuolema.[7]

[6] Tehtävä on mukaelma Daniel Kahnemanin (2011, 184) kirjassa *Thinking, Fast and Slow.* Penguin Books esittämästä. Kahneman on psykologi ja taloustieteen nobelisti vuodelta 2002. Alkuperäislähde olisi Max Bazermanin kirja *Judgement in Managerial Decision Making.*

[7] Donald T. Campbell ja David A. Kenny (1999): *A Primer on Regression Artifacts.* Guilford, New York.

2 Regressiovirhepäätelmä

Regressiovirhepäätelmä (*regression fallacy* tai *Galton's fallacy*) tehdään, kun satunnaisvaihtelussa regressiota odotusarvoa kohti kuvaavan suoran ympärillä kuvitellaan kausaalisuutta kuten että regressio johtaisi jakauman typistymiseen.

Yhteiskuntatieteilijät ovat joskus hahmottavinaan kausaalisuutta tilanteista, joissa sitä ei ole. Kuuluisa esimerkki on tilastotieteen(!) professori Horace Secrist. Hän julkaisi 1933 massiivisen empiirisen tutkimuksen amerikkalaisten yritysten liikevoittojen kehityksestä 1920–1930. Hän havaitsi, että yritysten, jotka pärjäsivät parhaimmin tai huonoimmin 1920, liikevoitot olivat lähestyneet 1930 kaikkien yritysten liikevoittojen keskiarvoa. Secrist päätteli, että taloudellinen kilpailu pakotti yritykset "keskiarvoistumaan" ajan myötä. Löytönsä korostamiseksi Secrist antoi kirjalleen nimeksi *The Triumph of Mediocrity in Business*. Todellisuudessa yritysten liikevoittojen jakauma ei ollut muuttunut, ja Secristin havainnot selittyvät regressiolla odotusarvoa kohti.[8]

Vaikka ongelma on tunnettu, edelleen julkaistaan vastaavia tutkimuksia. Kahneman (2011, 204–208) kritisoi business-kirjallisuutta, jossa perehdytään menestyneiden yhtiöiden strategioihin, yrityskulttuureihin ja johtamistapoihin. Esimerkkinä hän mainitsee Collinsin ja Porras'aan (2000) kirjan.[9] Kirjan viesti on, että jokaisen toimitusjohtajan, johtajan ja yrittäjän tulisi lukea se, jotta muutkin yritykset osaisivat noudattaa menestyneiden yritysten toimintamalleja ja pärjäisivät. Kahnemanin mukaan Collinsin ja Porras'aan ylistämät yritykset eivät pian tutkimuksen julkaisemisen jälkeen enää pärjänneet juurikaan kilpailijoitansa paremmin. Kahneman viittaa muihin vastaaviin tapauksiin, joissa tutkimuksessa hehkutettujen yritysten kukoistus lopahtaa sen julkaisemisen jälkeen. Regressio odotusarvoa kohti on luonteva tulkinta tällaisille tapahtumille. Ihaillut yritykset olivat erityisen menestyviä tutkimushetkellä lähinnä sattumasta johtuen.

Kahneman (mts. 174) kertoo konkreettisen ja opastavaisen esimerkin, kuinka ihmiset voivat kuvitella kausaalisuutta siellä, missä on pelkkää sattumaa (lyhennetty käännös luennoitsijan):

[8] Stephen Stigler (1999) kertoo Secristin tutkimuksesta tarkemmin kirjassaan *Statistics on the Table. The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge.

[9] Jim Collins ja Jerry I. Porras (2000): *Built to Last: Succesful Habits of Visionary Companies*. Random House Business Books, London.

Sain yhden elämäni tyydyttävimmistä eureka-kokemuksistani opettaessani Israelin ilmavoimien lentokouluttajille tehokkaan opettamisen psykologiaa. Olin kertonut kouluttajille, kuinka hyvän suorituksen palkitseminen toimii paremmin kuin virheistä rangaitseminen. Yksi vanhemmista kouluttajista arveli, että hyvästä suorituksesta palkitseminen sopii ehkä linnuille muttei hävittäjälentäjäkadeteille: "Olen monesti kehunut kadetteja puhtaasta suorituksesta vaikeassa lentoliikkeessä. Seuraavalla kerralla he järjestään suoriutuvat samasta liikkeestä huonommin. Toisaalta olen monesti huutanut kadetin korvakuulokkeeseen haukkuen häntä huonosta suorituksesta. Ylipäänsä haukkumani kadetit pärjäävät seuraavalla yrityksellä paremmin. Olkaa siis hyvä, älkääkä kertoko meille, että kehuminen toimii ja rangaistus ei, koska asia on juuri päinvastoin."

Edellä opitun perusteella on helppo hahmottaa, että vanhemman kouluttajan kokemukset selittyvät sattumalla: Erityisen hyvin pärjänneen kadetin suoritus regressoitui seuraavalla lennolla kohti odotusarvosuoritustaan ja erityisen heikosti suoriutuneen kadetin suoritus samoin. Kouluttaja virheellisesti liitti muutoksiin kuvittelemansa syy-seuraus -suhteen kehuistaan ja karjumisistaan.[10]

[10] Lisää esimerkkejä on Howard Wainerin (2005) kirjassa *Graphic Discovery*, Princeton University Press, Princeton (luku 10) ja W. Allen Wallisin ja Harry V. Robertsin (1956) kirjassa *Statistics. A New Approach*, Free Press, New York (luku 8). Ks. myös Milton Friedman (1992): Do Old Fallacies Ever Die? *Journal of Economic Literature*, 30, 2129–2132.