

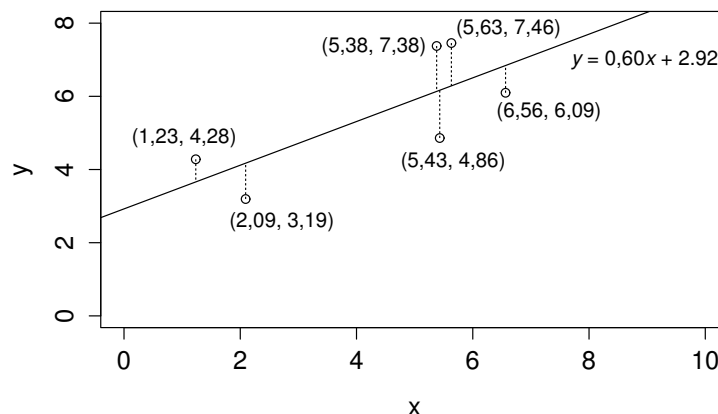
Suoran sovittaminen pistejoukkoon

Kun halutaan tutkia kahden tilastollisen muuttujan riippuvuutta toisistaan, käytetään usein *lineaarista regressiota* eli suoran sovittamista havaintojoukkoon. Siinä on annettu joukko havaintopareja (x_i, y_i) , ja tähän pistejoukkoon yritetään sovittaa suora $y = ax + b$ niin, että se kuvaisi havaintoja parhaalla mahdollisella tavalla. Kriteerinä käytetään yleensä *pienimmän neliösumman* kriteeriä, mikä tarkoittaa että suora sopii parhaiten, jos sen pisteiden y -koordinaattien ja havaintojen y -arvojen erotuksien neliöiden summa

$$\sum_i (y_i - (ax_i + b))^2$$

on mahdollisimman pieni.

Oheisessa kuvassa on esimerkiksi kuuden havaintopisteen joukkoon sovitettu suora, joka minimoi pisteistä laskettujen pystysuuntaisten etäisyyksien neliöiden summan. Esimerkiksi vasemmanpuoleisen pisteen etäisyys suorasta on $4,28 - (0,60 \cdot 1,23 + 2,92) \approx 0,62$. Muut etäisyydet ovat järjestyksessä $-0,98$, $1,24$, $-1,31$, $1,17$ ja $-0,74$. Etäisyyksien neliöiden summaksi tulee siten $6,51$.



Suoran sovittamiselle pienimmän neliösumman kriteerillä on olemassa yksinkertainen geometrinen tulkinta, johon seuraavassa tutustutaan.

Esimerkki

Tutkitaan esimerkin vuoksi kuvitteellisten opiskelijoiden koepisteiden ja tehtyjen harjoitusten välistä yhteyttä. Oletetaan, että kokeesta saatava maksimipistemäärä on 60 ja harjoitustehtäviä on kurssilla ollut 100. Jokaisesta opiskelijasta tiedetään hänen tekemiensä harjoitusten määrä sekä koetulos. Tarkoituksena on tutkia, onko harjoitusten tekemisellä vaikutusta koetulokseen.

Jos havaintojen määrä on n , on siirryttävä n -ulotteiseen avaruuteen. Esimerkin yksinkertaistamiseksi tarkastellaan aluksi 3 opiskelijan joukkoa, jolloin käytettävä avaruus on $\mathbb{R}^3$. Olkoon x_i opiskelijan i tekemien harjoitusten määrä ja y_i koetulos. Oletetaan, että opiskelijoiden tulokset ovat seuraavat:

i	x_i	y_i
1	40	35
2	80	50
3	60	55

Lineaarialgebrallinen tulkinta

Ilmaistaan esimerkin koetulokset avaruuden $\mathbb{R}^3$ vektorina $\mathbf{y} = (35, 50, 55)$. Kun suoraa sovitetaan pienimmän neliösumman kriteerillä, mitataan havaintopisteiden etäisyydet suorasta pystysuoraan (ks. ensimmäisen sivun kuva). Tällöin suorasta riittää tarkastella vain niitä pisteitä, jotka ovat täsmälleen jonkin havaintopisteen ylä- tai alapuolella eli joiden x -koordinaatit löytyvät havaintojoukosta. Jokainen suora $y = ax + b$ voidaan siis ajatella pistejoukkona

$$\{(x_1, ax_1 + b), (x_2, ax_2 + b), (x_3, ax_3 + b)\}.$$

Kun suorat ajatellaan kolmen pisteen joukkoina, ne voidaan ilmaista avaruuden $\mathbb{R}^3$ vektoreina samalla tavalla kuin koetulokset. Tällä tavoin tulkittuina kaikki mahdolliset suorat muodostavat avaruuden $\mathbb{R}^3$ osajoukon

$$S = \{(ax_1 + b, ax_2 + b, ax_3 + b) \mid a, b \in \mathbb{R}\} = \{a(x_1, x_2, x_3) + b(1, 1, 1) \mid a, b \in \mathbb{R}\}.$$

Kyseessä on vektorien $\mathbf{x} = (x_1, x_2, x_3) = (40, 80, 60)$ ja $\mathbf{n} = (1, 1, 1)$ virittämä taso.

Jokaista suoraa $y = ax + b$ vastaa siis jokin vektori $\mathbf{z} \in \mathbb{R}^3$, jonka komponentit ovat muotoa $ax_i + b$. Tällaiseen suoraan liittyvä havaintopisteiden etäisyyksien neliösumma voidaan kirjoittaa avaruuden $\mathbb{R}^3$ normin avulla:

$$\sum_i (y_i - (ax_i + b))^2 = \sum_i (y_i - z_i)^2 = |\mathbf{y} - \mathbf{z}|^2.$$

Koska tämä neliösumma yritetään minimoida, suoran sovittaminen voidaan esittää kysymyksenä ”Mikä tason S vektori $\mathbf{z}$ on sellainen, että sen ja vektorin $\mathbf{y}$ erotuksen normi on pienin mahdollinen?” Lineaarialgebrasta tiedetään, että vastaus on vektorin $\mathbf{y}$ kohti-suora projektio tasolle S , eli

$$\mathbf{z} = \text{proj}_S(\mathbf{y}).$$

Pienin neliösumma on tällöin $|\mathbf{y} - \text{proj}_S(\mathbf{y})|^2$.

Projektion laskeminen

Määritetään nyt kaava pistejoukkoon sovitettavalle suoralle yleisessä tapauksessa, jossa havaintojen määrä on n . Käytetään samoja merkintöjä kuin edellä: $\mathbf{x} = (x_1, \dots, x_n)$,

$\mathbf{y} = (y_1, \dots, y_n)$ ja $\mathbf{n} = (1, \dots, 1)$ avaruudessa $\mathbb{R}^n$. Voidaan olettaa, että vektorit $\mathbf{x}$ ja $\mathbf{n}$ eivät ole yhdensuuntaiset, sillä muuten kaikkien havaintojen x -arvot olisivat identtiset, ja havainnoista olisi mahdotonta päätellä tilastollisesti mitään.

On etsittävä vektorin $\mathbf{y}$ projektio tasolle $S = \text{span}(\mathbf{x}, \mathbf{n})$. Vektorit $\mathbf{x}$ ja $\mathbf{n}$ eivät välttämättä ole kohtisuorassa toisiaan vastaan, joten projektion laskemiseksi valitaan ensin tasolle S ortogonaalinen kanta muokkaamalla vektoria $\mathbf{x}$:

$$\mathbf{x}' = \mathbf{x} - \text{proj}_{\mathbf{n}}(\mathbf{x}) = \mathbf{x} - \frac{\mathbf{n} \cdot \mathbf{x}}{\mathbf{n} \cdot \mathbf{n}} \mathbf{n} = \mathbf{x} - \frac{\mathbf{n} \cdot \mathbf{x}}{n} \mathbf{n}$$

Vektorijono $(\mathbf{x}', \mathbf{n})$ on nyt tason S ortogonaalinen kanta.

Vektorin $\mathbf{y}$ projektio tasolle S saadaan laskemalla yhteen projektiot ortogonaalisen kannan vektorien suuntaan:

$$\begin{aligned} \text{proj}_S(\mathbf{y}) &= \text{proj}_{\mathbf{x}'}(\mathbf{y}) + \text{proj}_{\mathbf{n}}(\mathbf{y}) = \frac{\mathbf{x}' \cdot \mathbf{y}}{\mathbf{x}' \cdot \mathbf{x}'} \mathbf{x}' + \frac{\mathbf{n} \cdot \mathbf{y}}{\mathbf{n} \cdot \mathbf{n}} \mathbf{n} \\ &= \frac{\mathbf{x} \cdot \mathbf{y} - \frac{1}{n}(\mathbf{n} \cdot \mathbf{x})(\mathbf{n} \cdot \mathbf{y})}{|\mathbf{x}'|^2} \left(\mathbf{x} - \frac{\mathbf{n} \cdot \mathbf{x}}{n} \mathbf{n} \right) + \frac{\mathbf{n} \cdot \mathbf{y}}{n} \mathbf{n}. \end{aligned}$$

Tässä projektiovektori on esitetty lineaarikombinaationa vektoreista $\mathbf{x}$ ja $\mathbf{n}$. Koska nämä vektorit ovat lineaarisesti riippumattomat, lineaarikombinaation kertoimet ovat yksikäsitteiset. Kun nyt muistetaan, että kutakin sovitettavaa suoraa $y = ax + b$ vastaa vektori $a\mathbf{x} + b\mathbf{n}$, voidaan kertoimista lukea suoraan sopivimman suoran kulmakerroin ja vakiotermi:

$$a = \frac{\mathbf{x} \cdot \mathbf{y} - \frac{1}{n}(\mathbf{n} \cdot \mathbf{x})(\mathbf{n} \cdot \mathbf{y})}{|\mathbf{x}'|^2}, \quad b = \frac{\mathbf{n} \cdot \mathbf{y}}{n} - \frac{\mathbf{x} \cdot \mathbf{y} - \frac{1}{n}(\mathbf{n} \cdot \mathbf{x})(\mathbf{n} \cdot \mathbf{y})}{|\mathbf{x}'|^2} \frac{\mathbf{n} \cdot \mathbf{x}}{n}.$$

Lasketaan esimerkin vuoksi koetuloksiin sovitetun suoran kulmakerroin ja vakiotermi esitetyistä kaavoista. Aloitetaan laskemalla hieman välituloksia:

$$\begin{aligned} \mathbf{x} \cdot \mathbf{y} &= (40, 80, 60) \cdot (35, 50, 55) = 8700, \\ \mathbf{n} \cdot \mathbf{x} &= (1, 1, 1) \cdot (40, 80, 60) = 180, \\ \mathbf{n} \cdot \mathbf{y} &= (1, 1, 1) \cdot (35, 50, 55) = 140, \\ \mathbf{x}' &= (40, 80, 60) - \frac{180}{3}(1, 1, 1) = (-20, 20, 0), \\ |\mathbf{x}'|^2 &= (-20)^2 + 20^2 + 0^2 = 800. \end{aligned}$$

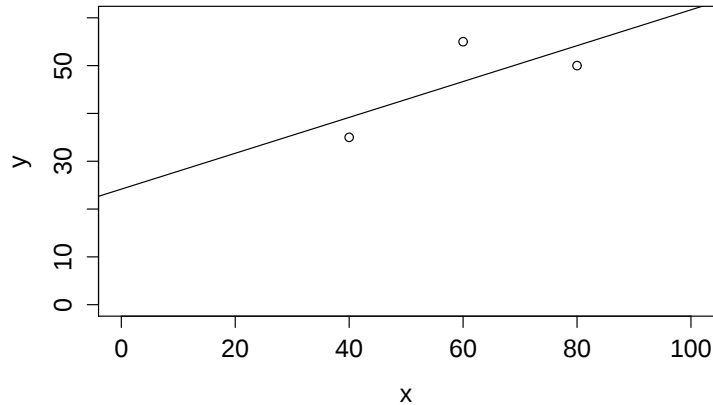
Näin saadaan

$$a = \frac{8700 - \frac{1}{3} \cdot 180 \cdot 140}{800} = 0,375$$

ja

$$b = \frac{140}{3} - 0,375 \cdot \frac{180}{3} = 24,17.$$

Oheisessa kuvassa on tuloksista piirretty pistediagrammi sekä edellä johdettu suora, joka minimoi y -arvojen poikkeamien neliösumman.



Tilastollisia suureita

Edellä määritetty kaava vektorin $\mathbf{y}$ projektiolle tasolle S saattaa näyttää monimutkaiselta, mutta kun sen osia tarkastellaan tilastotieteen näkökulmasta, kaava saa yksinkertaisemman asun.

Aloitetaan toteamalla, että pistetulo $\mathbf{n} \cdot \mathbf{x}$ on yksinkertaisesti kaikkien x -arvojen summa, eli $\mathbf{n} \cdot \mathbf{x} = \sum_i x_i$. Tarkastellaan tämän valossa projektiota vektorin $\mathbf{n}$ suuntaiselle suoralle:

$$\text{proj}_{\mathbf{n}}(\mathbf{x}) = \frac{\mathbf{n} \cdot \mathbf{x}}{\mathbf{n} \cdot \mathbf{n}} \mathbf{n} = \frac{\sum_i x_i}{n} \mathbf{n}.$$

Skalaarikerroin $\frac{1}{n} \sum_i x_i$ on kaikkien x -arvojen keskiarvo. Kun sitä merkitään $E(X)$, voidaan projektiio kirjoittaa muodossa $\text{proj}_{\mathbf{n}}(\mathbf{x}) = (E(X), \dots, E(X))$. Samalla tavoin nähdään, että $\text{proj}_{\mathbf{n}}(\mathbf{y}) = (E(Y), \dots, E(Y))$.

Suuresta $\mathbf{x}'$ tarvittiin sen normin neliötä. Tämä tulee muotoon

$$|\mathbf{x}'|^2 = |\mathbf{x} - \text{proj}_{\mathbf{n}}(\mathbf{x})|^2 = |(x_1, \dots, x_n) - (E(X), \dots, E(X))|^2 = \sum_i (x_i - E(X))^2.$$

Kyseessä on x -arvojen keskiarvosta laskettujen poikkeamien neliöiden summa. Kun tämä jaetaan arvojen lukumäärällä, saadaan muuttujan X *varianssi*, jota merkitään $\text{Var}(X)$. Siispä $|\mathbf{x}'|^2 = n \text{Var}(X)$. Geometrisesti muuttujan X varianssi on siis verrannollinen vektorin $\mathbf{x}$ suorasta $\text{span}(\mathbf{n})$ mitatun etäisyyden neliöön. Varsinaista etäisyyttä kuvaa varianssin neliöjuuri, ns. *keskihajonta* $\sigma(X)$ (kerrottuna luvulla $\sqrt{n}$).

Tarkastellaan vielä pistetuloa $\mathbf{x} \cdot \mathbf{y} = \sum_i x_i y_i$. Jos tämä jaetaan luvulla n , saadaan x - ja y -arvojen tulojen keskiarvo, jota merkitään $E(XY)$. Näin ollen $\mathbf{x} \cdot \mathbf{y} = n E(XY)$. Edelleen voidaan kirjoittaa

$$\mathbf{x} \cdot \mathbf{y} - \frac{1}{n}(\mathbf{n} \cdot \mathbf{x})(\mathbf{n} \cdot \mathbf{y}) = n E(XY) - n E(X) E(Y) = n(E(XY) - E(X) E(Y)).$$

Suuretta $E(XY) - E(X)E(Y)$ nimitetään muuttujien X ja Y väliseksi *kovarianssiksi* ja merkitään $\text{Cov}(X, Y)$. Kovarianssin geometriseen tulkintaan palataan jäljempänä korrelaation yhteydessä.

Lopulta sovitetun suoran kulmakerroin ja vakiotermi voidaan tilastollisia merkintöjä käyttäen ilmaista muodossa

$$a = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \text{ja} \quad b = E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} E(X).$$

Korrelaatio

Kun havaintojoukkoon sopivin suora on löytynyt, voidaan kysyä, miten voimakasta riippuvuus x - ja y -arvojen välillä on. Lineaarisen riippuvuuden mittaamiseen käytetään *korrelaatiokerrointa*, joka määritellään seuraavasti:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Kaavassa $\sigma(X)$ ja $\sigma(Y)$ ovat muuttujien X ja Y keskihajonnat eli varianssin neliöjuuret.

Ryhdytään etsimään korrelaatiokertoimelle geometrista tulkintaa. Palautetaan mieleen, että $n \text{Cov}(X, Y) = \mathbf{x}' \cdot \mathbf{y}$ ja $\sqrt{n} \sigma(X) = |\mathbf{x}'|$. Merkitään lisäksi $\mathbf{y}' = \mathbf{y} - \text{proj}_{\mathbf{n}}(\mathbf{y})$, jolloin $|\mathbf{y}'| = \sqrt{n} \sigma(Y)$. Lisäksi

$$\mathbf{x}' \cdot \mathbf{y}' = \mathbf{x}' \cdot \left(\mathbf{y} - \frac{\mathbf{n} \cdot \mathbf{y}}{\mathbf{n} \cdot \mathbf{n}} \mathbf{n} \right) = \mathbf{x}' \cdot \mathbf{y} - \frac{\mathbf{n} \cdot \mathbf{y}}{\mathbf{n} \cdot \mathbf{n}} \underbrace{\mathbf{x}' \cdot \mathbf{n}}_{=0} = \mathbf{x}' \cdot \mathbf{y}.$$

Nyt saadaan

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)} = \frac{\frac{1}{n} \mathbf{x}' \cdot \mathbf{y}}{\frac{1}{\sqrt{n}} |\mathbf{x}'| \frac{1}{\sqrt{n}} |\mathbf{y}'|} = \frac{\mathbf{x}' \cdot \mathbf{y}'}{|\mathbf{x}'| |\mathbf{y}'|}.$$

Tästä nähdään, että muuttujien X ja Y välinen korrelaatio on vektorien $\mathbf{x}'$ ja $\mathbf{y}'$ välisen kulman kosini.

Entä millä tavoin korrelaatiokerroin liittyy suoran sovittamiseen? Kuten aiemmin on nähty, projektiovektori $\text{proj}_{\mathbf{n}}(\mathbf{y})$ vastaa vaakasuuntaista suoraa $y = E(Y)$, sillä

$$\text{proj}_{\mathbf{n}}(\mathbf{y}) = \frac{\mathbf{n} \cdot \mathbf{y}}{\mathbf{n} \cdot \mathbf{n}} \mathbf{n} = \left(\frac{1}{n} \sum_i y_i \right) \mathbf{n} = E(Y) \mathbf{n}.$$

Erotusvektori $\mathbf{y} - \text{proj}_{\mathbf{n}}(\mathbf{y})$ kuvaa siis y -havaintojen poikkeamaa keskiarvosta. Osa tästä poikkeamasta selittyy lineaarisella riippuvuudella x -havainnoista. Lineaarista riippuvuutta kuvaa sovitettu suora, ja tätä puolestaan vastaa $\text{proj}_S(\mathbf{y})$. Siispä lineaarisen riippuvuuden selittämää osaa poikkeamasta kuvaa erotus

$$\text{proj}_S(\mathbf{y}) - \text{proj}_{\mathbf{n}}(\mathbf{y}).$$

Jäljelle jäävä, selittämätön tai satunnainen osa on tällöin

$$(\mathbf{y} - \text{proj}_{\mathbf{n}}(\mathbf{y})) - (\text{proj}_S(\mathbf{y}) - \text{proj}_{\mathbf{n}}(\mathbf{y})) = \mathbf{y} - \text{proj}_S(\mathbf{y}).$$

Vertailemalla edellä mainittujen vektoreiden pituuksien neliöitä voidaan määritellä ns. *selityskerroin*

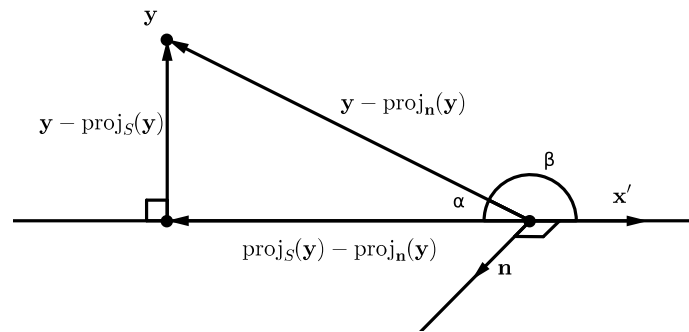
$$R^2 = \frac{|\text{proj}_S(\mathbf{y}) - \text{proj}_n(\mathbf{y})|^2}{|\mathbf{y} - \text{proj}_n(\mathbf{y})|^2}.$$

Selityskerroin kuvaa tilastollisesti sitä, miten suuri osa muuttujan Y varianssista, eli vektorin $\mathbf{y}$ neliöllisestä etäisyydestä keskiarvovektorista $\text{proj}_n(\mathbf{y})$, selittyy lineaarisella riippuvuudella.

Selityskerroin on myös helppo laskea käyttäen hyväksi aiempia tietoja:

$$\begin{aligned} R^2 &= \frac{|\text{proj}_S(\mathbf{y}) - \text{proj}_n(\mathbf{y})|^2}{|\mathbf{y} - \text{proj}_n(\mathbf{y})|^2} = \frac{|\text{proj}_{\mathbf{x}'}(\mathbf{y})|^2}{|\mathbf{y}'|^2} = \frac{(\text{Cov}(X, Y) / \text{Var}(X))^2 |\mathbf{x}'|^2}{|\mathbf{y}'|^2} \\ &= \frac{(\text{Cov}(X, Y) / \text{Var}(X))^2 \cdot n \text{Var}(X)}{n \text{Var}(Y)} = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X) \text{Var}(Y)} = \rho(X, Y)^2. \end{aligned}$$

Nähdään, että selityskerroin on yksinkertaisesti korrelaatiokertoimen neliö. Tämä yhteys voidaan selittää oheisella kuvalla.



Kuvassa on kolmio, joka muodostuu vektoreista $\mathbf{y} - \text{proj}_n(\mathbf{y})$, $\text{proj}_S(\mathbf{y}) - \text{proj}_n(\mathbf{y})$ ja $\mathbf{y} - \text{proj}_S(\mathbf{y})$. Selityskerroin R^2 on kuvan kulman α viereisen kateetin ja hypotenuusan suhteen neliö, eli kyseisen kulman kosinin neliö. Korrelaatiokerroin ρ on puolestaan vektorien $\mathbf{x}'$ ja $\mathbf{y}'$ välisen kulman kosini. Koska $\mathbf{y} - \text{proj}_n(\mathbf{y}) = \mathbf{y}'$, kyseessä on kuvan kulma β . Kun huomataan, että vektori $\text{proj}_S(\mathbf{y}) - \text{proj}_n(\mathbf{y}) = \text{proj}_{\mathbf{x}'}(\mathbf{y})$ on yhdensuuntainen vektorin $\mathbf{x}'$ kanssa, nähdään, että joko $\beta = \alpha$ tai $\beta = 90^\circ - \alpha$. Koska $\cos(90^\circ - \alpha) = -\cos \alpha$, voidaan lopulta päätellä, että $\cos^2 \beta = \cos^2 \alpha$. Näin ollen korrelaation neliö on yhtä kuin selityskerroin.

Myös korrelaatiokertoimen lukuarvosta voidaan tehdä johtopäätöksiä kuvan avulla. Kulma β on nollakulman ja oikokulman välillä, joten korrelaatio $\cos \beta$ on arvojen -1 ja 1 välillä. Korrelaatio on 1 silloin, kun β on nollakulma, ja tämä toteutuu, jos ja vain jos $\mathbf{y} = \text{proj}_S(\mathbf{y})$ ja vektorit $\mathbf{y}$ ja $\mathbf{x}'$ ovat yhdensuuntaiset. Tällöin sopivin suora kulkee kaikkien havaintopisteiden kautta ja sen kulmakerroin on positiivinen (sillä kulmakerroimen merkki määräytyy pistetulosta $\mathbf{x}' \cdot \mathbf{y}$). Vastaavasti korrelaatio on -1 , kun β on oikokulma, ja tämä tarkoittaa, että havaintopisteet ovat kaikki samalla suoralla, jonka

kulmakerroin on negatiivinen. Lopulta korrelaatio on 0, kun kulma β on suora, ja tämä on mahdollista ainoastaan, jos $\text{proj}_S(\mathbf{y}) = \text{proj}_{\mathbf{n}}(\mathbf{y})$. Tässä tilanteessa pistejoukkoa kuvaa parhaiten suora $y = E(Y)$, eli minkäänlaista lineaarista riippuvuutta ei ole.

Jokke Häsä, syksyllä 2015