TILASTOMATEMATIIKKA

Harjoitus 7 ratkaisut, kevät 2018

1. Heitetään kolme kertaa virheetöntä kolikkoa. Olkoon X ilmestyvien kruunujen lukumäärä ja Y kahdessa ensimmäisessä heitossa esiintyvien klaavojen lukumäärä.

   a) Määrää muuttujien X ja Y yhteisjakauma.

   b) Määrää muuttujien X ja Y odotusarvo.

   c) Esitä intuitiivinen perustelu, miksi X ja Y eivät ole riippumattomia muuttujia. Totea sama asia laskemalla. Määrää muuttujien X ja Y välinen korrelaatiokerroin.

   d) Millä todennäköisyydellä kruunujen lukumäärä on suurempi kuin kahdessa ensimmäisessä heitossa esiintyvien klaavojen lukumäärä?

Ratkaisu:

a) Määrätään ensin muuttujien X ja Y määräämän satunnaisvektorin (X, Y) arvojoukko $S_{XY} = S_X \times S_Y$, missä S_X on muuttujan X arvojoukko ja S_Y on muuttujan Y arvojoukko. Koska kruunia voi olla joko $0, 1, 2$ tai 3 ja kahdessa ensimmäisessä heitossa klaavoja voi olla $0, 1$ tai 2, niin arvojoukoksi saadaan

$$S_{XY} = S_X \times S_Y = \{(0,0), (0,1), \ldots, (3,1), (3,2)\}.$$

Alkioiden lukumäärä on $\#S_{XY} = 4 \times 3 = 12$, joista jokaisen pistetodennäköisyys voidaan laskea helposti luettelemalla kullekin vaihtoehdolle suotuisat alkeistapahtumat satunnaiskokeesta

”heitetään kolikkoa 3 kertaa”,

jonka otosavaruudeksi voidaan ottaa

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\},$$

missä $H =$ ”heitto on kruuna” ja $T =$ ”heitto on klaava”.
Laaditaan yhteisjakauman pistetodennäköisyyksille taulukko, mihin on merkitty myös reunajakaumien pistetodennäköisyydet.

$X \backslash Y$	0	1	2	p_i
0	0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$
2	$\frac{1}{8}$	$\frac{1}{4}$	0	$\frac{3}{8}$
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$
q_j	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

Otetaan esimerkkinä kahden arvopisteen pistetodennäköisyyksien laskeminen. Esimerkiksi arvolle $(X, Y) = (0, 1)$ ei ole yhtään suotuisaa alkeistapahtumaa, sillä jos kruunia ei esiinny yhtään, täytyy kahdessa ensimmäisessä heitossa olla täsmälleen 2 klaavaa. Otetaan toiseksi esimerkiksi arvopiste $(X, Y) = (1, 1)$. Kruunujen lukumäärälle $X = 1$ on 3 suotuisaa alkeistapahtumaa HTT, THT, TTH, joista alkeistapahtumat HTT ja THT ovat suotuisia arvolle $Y = 1$. Tästä saadaan todennäköisyydeksi

$$P(X = 1 \,\text{ja}\, Y = 1) = \frac{\#\{HTT, THT\}}{\#S} = \frac{2}{8} = \frac{1}{4}.$$

b) Muuttujien X ja Y odotusarvot voidaan laskea yhteisjakauman taulukosta. Merkitään

$$p_{ij} = P(X = i \,\text{ja}\, Y = j).$$

Luentokalvoissa esitetyn odotusarvon laskentakaavan

$$E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) \mathrm{d}x \mathrm{d}y$$

diskreetti analogia on

$$E(X) = \sum_i \sum_j i \cdot p_{ij},$$

joka nyt meidän esimerkissämme saa arvon

$$E(X) = \sum_{i=0}^{3} \sum_{j=0}^{2} i \cdot P(X = i \,\mathrm{ja}\, Y = j) = \cdots = \frac{3}{2},$$

joka oltaisiin toki voitu laskea myös suoraan X:n jakaumasta. Odotusarvo on itse asiassa laskettu harjoituksen 3 tehtävässä 1 a).
Vastaavasti muuttujan Y odotusarvoksi saadaan

$$E(Y) = \sum_{i=0}^{3} \sum_{j=0}^{2} j \cdot P(X = i \,\mathrm{ja}\, Y = j) = \cdots = 1,$$

joka olisi saatu helposti myös suoraan.

c) Muuttujat X ja Y eivät intuitiivisesti voi olla riippumattomia, sillä saatujen kruunien lukumäärä luonnollisestikin vaikuttaa kahdessa ensimmäisessä heitossa esiintyvien klaavojen mahdollisiin arvoihin. Jos kruunia on esimerkiksi 3, ei kahdessa ensimmäisessä heitossa voi esiintyä yhtään klaavaa.
Sama asia voidaan todeta myös laskennallisesti käyttämällä a)-kohdassa muodostettua taulukkoa. Koska

$$P(X = 3 \,\mathrm{ja}\, Y = 1) = 0 \neq \frac{1}{8} \cdot \frac{1}{2} = P(X = 3) P(Y = 1),$$

eivät X ja Y ole riippumattomia.
Määritelmän mukaan korrelaatiokerroin on

$$\rho \overset{\text{merk.}}{=} \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

joten täytyy laskea muuttujien X ja Y hajonnat σ_X ja σ_Y sekä kovarianssi

$$\mathrm{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y). \tag{1}$$

Kaikki tunnusluvut voidaan laskea muunnoksen odotusarvon kaavalla

$$E(h(X, Y)) = \sum_i \sum_j h(X, Y) P(X = i \,\mathrm{ja}\, Y = j),$$

joka on luennoilla jatkuvalle yhteisjakaumalle esitetyn kaavan

$$E(h(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(X, Y) f_{XY}(x, y) \mathrm{d}x \mathrm{d}y$$

diskreetti analogia.

Muuttujan X varianssiksi saadaan yhteisjakauman taulukosta ja b)-kohdasta

$$\begin{aligned}\mathrm{Var}(X) = E((X-E(X))^2) &\overset{\text{b)}}{=} \sum_{i=0}^{3}\sum_{j=0}^{2}\left(i-\frac{3}{2}\right)^2 \cdot P(X=i \,\text{ja}\, Y=j)\\ &= \left(0-\frac{3}{2}\right)^2\left(0+0+\frac{1}{8}\right)+\left(1-\frac{3}{2}\right)^2\left(0+\frac{1}{4}+\frac{1}{8}\right)\\ &\quad+\left(2-\frac{3}{2}\right)^2\left(\frac{1}{8}+\frac{1}{4}+0\right)+\left(3-\frac{3}{2}\right)^2\left(\frac{1}{8}+0+0\right)\\ &= \cdots = \frac{3}{4}.\end{aligned}$$

Tämä olisi voitu laskea myös suoraan X:n jakaumasta, minkä itse asiassa teimme harjoituksessa 3, ja päädyttiin samaan lopputulokseen. Edellisestä saadaan muuttujan X hajonnaksi

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \frac{\sqrt{3}}{2}.$$

Muuttujan Y hajonta voidaan laskea samalla tavalla. Tehdään se tässä kuitenkin toisella tavalla. Koska $Y \sim \mathrm{Bin}(2, \frac{1}{2})$, niin $\mathrm{Var}(Y) = \frac{1}{2}$ ja siten

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \frac{1}{\sqrt{2}}.$$

Kovarianssin laskemista varten lasketaan ensin $E(XY)$, joka on yhteisjakauman taulukon mukaan

$$E(XY) = \sum_{i=0}^{3}\sum_{j=0}^{2} ij \cdot P(X=i \,\text{ja}\, Y=j) = \underbrace{1\cdot\frac{1}{4}+2\cdot\frac{1}{8}}_{i=1}+\underbrace{2\cdot\frac{1}{4}}_{i=2} = 1,$$

sillä kaikki muut yhteenlaskettavat katoavat, kun $i=0$, $j=0$ tai $p_{ij}=0$.
Kovarianssin laskentakaavan (1) mukaan

$$\mathrm{Cov}(X,Y) = 1-\frac{3}{2} = -\frac{1}{2}$$

ja siten korrelaatiokerroin on

$$\rho = \frac{-\frac{1}{2}}{\frac{\sqrt{3}}{2}\cdot\frac{1}{\sqrt{2}}} = -\sqrt{\frac{2}{3}}.$$

d) Kysytään todennäköisyyttä $P(X>Y)$. Tämä todennäköisyys voidaan lukea helposti yhteisjakauman taulukosta käymällä läpi ”alakolmio-osa” ja laskemalla todennäköisyydet yhteen. Todennäköisyydeksi saadaan

$$P(X>Y) = \sum_{i>j} P(X=i \,\text{ja}\, Y=j) = \sum_{j=0}^{2}\sum_{i=j+1}^{3} p_{ij} = \frac{1}{8}+\frac{1}{8}+\frac{1}{4} = \frac{1}{2}.$$

2. Satunnaismuuttujat X ja Y kuvaavat laitteen kahden elektronisen komponentin elinikien pituuksia vuosina. Satunnaismuuttujien yhteisjakauman tiheysfunktio on

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}(xy+1)\mathrm{e}^{-x-y}, & x,y>0, \\ 0, & \text{muulloin.} \end{cases}$$

a) Millä todennäköisyydellä molemmat komponentit kestävät vähintään vuoden?

b) Ovatko X ja Y riippumattomia?

c) Mitä voit sanoa muuttujien X ja Y korrelaatiosta b)-kohdan perusteella?

Ratkaisu:

a) Kysytään todennäköisyyttä $P(\{X \geq 1\} \cap \{Y \geq 1\}) = P(X \geq 1 \,\text{ja}\, Y \geq 1)$. Kyseessä on integraali

$$I = \int_1^\infty \int_1^\infty f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y.$$

Hajotetaan integraali kahteen osaan

$$I = I_1 + I_2 = \frac{1}{2}\int_1^\infty \int_1^\infty xy\mathrm{e}^{-x-y}\mathrm{d}x\mathrm{d}y + \frac{1}{2}\int_1^\infty \int_1^\infty \mathrm{e}^{-x-y}\mathrm{d}x\mathrm{d}y$$

Lasketaan ensin I_2. Käytetään eksponenttifunktion laskusääntöä $\mathrm{e}^{-x-y} = \mathrm{e}^{-x}\mathrm{e}^{-y}$. Koska muuttujat voidaan separoida ja integroimisalue on molemmille muuttujille sama, riittää integroida vaikkapa muuttujan x suhteen. Muuttujalle x saadaan

$$\int_1^\infty \mathrm{e}^{-x}\mathrm{d}x = -\mathrm{e}^{-x}\Big|_{x=1}^\infty = \mathrm{e}^{-1},$$

joten $I_2 = \frac{1}{2}\mathrm{e}^{-1}\cdot\mathrm{e}^{-1} = \frac{1}{2}\mathrm{e}^{-2}$.
Samalla tavalla voidaan päätellä, että integraalin I_1 laskemiseksi riittää laskea

$$\int_1^\infty \underbrace{x}_{f(x)} \cdot \underbrace{\mathrm{e}^{-x}}_{g'(x)}\mathrm{d}x \overset{\text{ositt.int.}}{=} -x\mathrm{e}^{-x}\Big|_{x=1}^\infty + \int_1^\infty \mathrm{e}^{-x}\mathrm{d}x = 2\mathrm{e}^{-1}.$$

Edellisen perusteella $I_1 = \frac{1}{2}(2\mathrm{e}^{-1})^2 = 2\mathrm{e}^{-2}$. Yhdistämällä edellä lasketut integraalit saadaan kysytyksi todennäköisyydeksi

$$\frac{5}{2}\mathrm{e}^{-2} \approx 0.34.$$

b) Tutkitaan, onko $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Sitä varten riittää laskea reunajakaumien tiheysfunktiot. Edellisen kohdan perusteella voidaan hyödyntää symmetriaa ja laskea pelkästään muuttujan X tiheysfunktio, sillä tiheysfunktiot $f_X(x)$ ja $f_Y(y)$ ovat muuttujan nimeä vaille samat.
Edellisessä kohdassa laskettiin oleellisesti jo kaikki sillä erotuksella, että nyt integrointi tehdään välin $[0,\infty[$ yli. Hajotetaan integraali samalla tavalla kahteen osaan

$$f_X(x) = \int_0^\infty f_{X,Y}(x,y)\mathrm{d}y = \frac{1}{2}\int_0^\infty xy\mathrm{e}^{-x-y}\mathrm{d}y + \frac{1}{2}\int_0^\infty \mathrm{e}^{-x-y}\mathrm{d}y \overset{\text{merk.}}{=} I_1 + I_2.$$

Lasketaan malliksi

$$I_2 = \frac{1}{2}\mathrm{e}^{-x}\int_0^\infty \mathrm{e}^{-y}\mathrm{d}y = \frac{1}{2}\mathrm{e}^{-x},$$

missä käytettiin eksponenttifunktion laskusääntöä $\mathrm{e}^{-x-y} = \mathrm{e}^{-x}\mathrm{e}^{-y}$. Vastaavalla tavalla voidaan laskea myös I_1, joten muuttujan X reunatiheydeksi saadaan

$$f_X(x) = \frac{1}{2}(x+1)\mathrm{e}^{-x}.$$

Koska

$$f_{X,Y}(x,y) \neq f_X(x)f_Y(y),$$

niin muuttujat X ja Y eivät ole riippumattomia.

c) Riippumattomuudesta seuraa korreloimattomuus, mutta riippuvassa tapauksessa emme voi sanoa korrelaatiosta mitään. Huomaa, että korrelaatio mittaa *lineaarisen riippuvuuden aste*tta.

3. Olkoon $\boldsymbol{X} = (X, Y) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, missä kovarianssimatriisi on

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

a) Laske muuttujien X ja Y kovarianssi.

b) Määrää kovarianssimatriisin käänteismatriisi kaavalla

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

ja esitä sitten 2-ulotteisen normaalijakauman eksponentin lauseke muodossa $Ax^2 + Bxy + Cy^2$. Mikä on normaalijakauman tiheysfunktion arvo ellipsillä $x^2 - xy + y^2 = 1$?

c) Määrää satunnaismuuttujien X ja Y jakaumat laskemalla niiden tiheysfunktiot. **Vihje:** Neliöi $\boldsymbol{X}$:n tiheysfunktion eksponentti sopivasti ja käytä integroinnissa sopivaa sijoitusta.

Ratkaisu:

a) Suoraan kovarianssimatriisista nähdään, että $\mathrm{Cov}(X, Y) = \underline{1}$.

b) $|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) = 2 \cdot 2 - 1 \cdot 1 = 3$ ja

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

2-ulotteisen normaalijakauman eksponentti, kun $\mu = (0, 0)$, on

$$\begin{aligned} -\frac{1}{2}(\mathbf{x} - \mu)\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu)' &= -\frac{1}{2}\left((x, y) - (0, 0)\right) \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \left((x, y) - (0, 0)\right)' \\ &= -\frac{1}{6}(x, y) \begin{pmatrix} 2x - y \\ -x + 2y \end{pmatrix} \\ &= -\frac{1}{6}(2x^2 - xy - yx + 2y^2) = -\frac{1}{3}x^2 + \frac{1}{3}xy - \frac{1}{3}y^2 \end{aligned}$$

Ellipsillä $x^2 - xy + y^2 = 1$ on eo. eksponentti $-\frac{1}{3}$, joten tiheysfunktion arvo on

$$f_X(X) = \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}|}} e^{-\frac{1}{3}} = \underline{\frac{1}{2\pi\sqrt{3} e^{\frac{1}{3}}}} \approx 0,065841$$

c) Lasketaan tiheysfunktiot reunajakaumille integroimalla tiheysfunktiota f_{XY}

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{3}(x^2 - xy + y^2)}}{2\pi\sqrt{3}}\, dy = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{3}[(y - \frac{x}{2})^2 + \frac{3}{4}x^2]}}{2\pi\sqrt{3}}\, dy$$

Sijoitetaan tähän $y - \frac{x}{2} = \sqrt{\frac{3}{2}}\, z$, jolloin integrointirajat säilyvät ja $dy = \sqrt{\frac{3}{2}}\, dz$

$$= \int_{-\infty}^{\infty} \frac{e^{-\frac{3x^2}{12}}}{2\pi\sqrt{3}} \cdot e^{-\frac{1}{3}(\sqrt{\frac{3}{2}}z)^2} \cdot \sqrt{\frac{3}{2}}\, dz = \frac{e^{-\frac{x^2}{4}}}{2\sqrt{\pi}} \cdot \underbrace{\int_{-\infty}^{\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\, dz}_{=\Phi(\infty) - \Phi(-\infty) = 1} = \underline{\frac{e^{-\frac{x^2}{4}}}{2\sqrt{\pi}}},$$

joten $X \sim N(0, 2)$.

Vastaavasti saadaan laskettua $f_Y(y) = \underline{\frac{e^{-\frac{y^2}{4}}}{2\sqrt{\pi}}}$, joten $Y \sim N(0, 2)$.

4. Tutkittiin 11 sisarusparin pituuksia ja saatiin seuraava havaintoaineisto

sisko	175	163	165	160	165	157	165	163	168	150	157
veli	180	173	168	170	178	180	178	185	183	165	168

Oletetaan, että havainnot ovat peräisin kaksiulotteisesta normaalijakaumasta $\boldsymbol{X} = (X, Y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

a) Laske kovarianssimatriisille estimaatti

$$\boldsymbol{S} = \begin{pmatrix} s_x^2 & r s_x s_y \\ r s_x s_y & s_y^2 \end{pmatrix},$$

missä s_x on siskon pituuksien x otoshajonta, s_y on veljen pituuksien y otoshajonta ja r on muuttujien x ja y välinen korrelaatiokerroin.

b) Laske matriisin $\boldsymbol{S}$ käänteismatriisi. Onko $\boldsymbol{S}^{-1}$ välttämättä olemassa, vaikka $\boldsymbol{\Sigma}^{-1}$ on olemassa? Esitä estimoidun 2-ulotteisen normaalijakauman eksponentissa oleva termi

$$-\frac{1}{2} \cdot \begin{pmatrix} x & y \end{pmatrix} \cdot \boldsymbol{S}^{-1} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

muodossa $Ax^2 + Bxy + Cy^2$ samalla tavalla kuin tehtävässä 3.

c) Laske muuttujien x ja y välinen regressiosuora. Piirrä muuttujien x ja y sirontakuvio, regressiosuora ja ellipsi $Ax^2 + Bxy + Cy^2 = -\frac{1}{4}$ samaan koordinaatistoon. Yhtyykö regressiosuora ellipsin pääakseliin?

Ratkaisu:

a) Kaikki tunnusluvut saadaan samalla tavalla kuin edellisellä viikolla esimerkiksi suoraan laskimesta. Korrelaatiokertoimeksi saadaan $r \approx 0.558$ ja otosvariansseiksi $s_x^2 \approx 42.581$ ja $s_y^2 \approx 47.742$, joista voidaan muodostaa kovarianssimatriisin estimaatti

$$\boldsymbol{S} = \begin{pmatrix} 42.58 & 25.16 \\ 25.16 & 47.74 \end{pmatrix},$$

b) Käänteismatriisi on tällä kertaa olemassa, mutta näin ei välttämättä ole yleisesti, vaikka populaation kovarianssimatriisi olisikin säännöllinen, eli $\boldsymbol{\Sigma}^{-1}$ olisi olemassa. Käänteismatriisiksi saadaan

$$\boldsymbol{S}^{-1} = \begin{pmatrix} 0.0341 & -0.0180 \\ -0.0180 & 0.0304 \end{pmatrix},$$

ja kysytyksi termiksi

$$-\frac{1}{2}\left(0.0341x^2 - 0.0360xy + 0.0304y^2\right).$$

c) Regressiosuora

$$y = 0.598x + 78.08$$

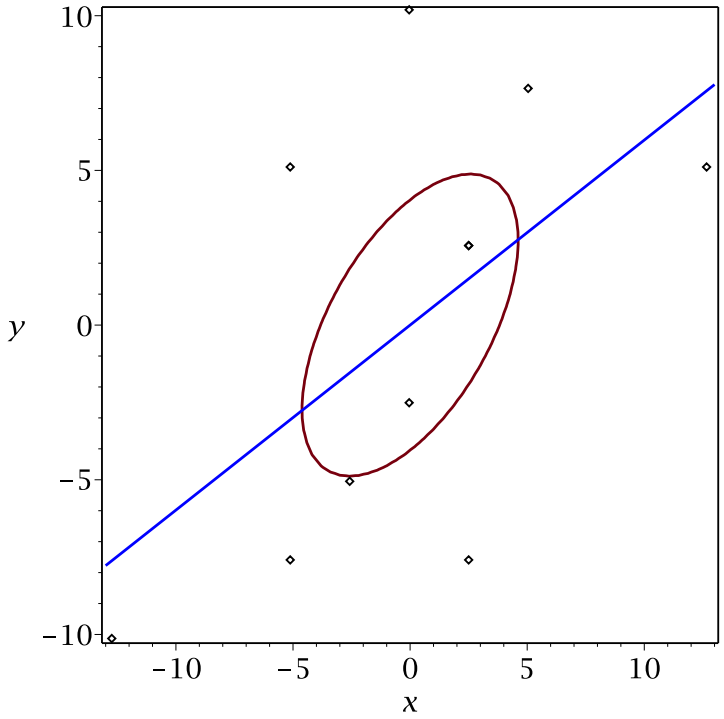
saadaan laskimesta samalla tavalla kuin edellisellä viikolla.
Koska havainnot ovat keskittyneet keskiarvopisteen $(\overline{x}, \overline{y})$ ympäristöön, voidaan havainnot halutessa siirtää origon ympäristöön muuttujanvaihdolla $(x', y') = (x - \overline{x}, y - \overline{y})$. Samoin, koska ainoastaan regressiosuoran suunnalla on nyt merkitystä, voidaan myös regressiosuora siirtää origon ympäristöön samalla muuttujanvaihdolla, jolloin saadaan origon kautta kulkeva suora

$$y = 0.598x.$$

Toinen vaihtoehto on tarkastella ellipsiä keskiarvopisteen ympäristössä, jolloin tarkastellaan kaksiulotteisen normaalijakauman aidon eksponentin estimaattia

$$-\frac{1}{2}\cdot\begin{pmatrix}x-\overline{x} & y-\overline{y}\end{pmatrix}\cdot\boldsymbol{S}^{-1}\cdot\begin{pmatrix}x-\overline{x}\\ y-\overline{y}\end{pmatrix}$$

Havainnollistamisen kannalta ei ole merkitystä kumpaa edellisistä vaihtoehtoisista tavoista käytetään. Piirretään kaikki origon ympäristöön, jolloin saadaan oheinen kuva.



Kuva 1: Origoon siirretty regressiosuora ja sirontakuvio sekä ellipsi $0.0341x^2 - 0.0360xy + 0.0304y^2 = -\frac{1}{4}$.

Kuten kuvasta näkyy, regressiosuora ei yhdy ellipsin pääakseliin. Näin on myös yleisesti. Jos $|\rho| < 1$, niin regressiosuora ei yhdy pääakseliin. Huomaa, että kuvan perusteella regressiosuoran kulmakerroin on (itseisarvoltaan) pienempi kuin pääakselin kulmakerroin eli on tapahtunut *regressio keskiarvoa kohti*.