TILASTOMATEMATIIKKA

Harjoitus 6 ratkaisut, kevät 2018

1. Hooken lain mukaan jouseen kiinnitettynä olevaan kappaleeseen kohdistuva voima F saadaan kaavasta

$$F = -kx, \tag{1}$$

missä k on jousivakio ja x on poikkeama tasapainoasemasta.

a) Jousivakion k määräämistä varten tehtiin mittauksia ja saatiin seuraavat havainnot

F	2	3	5	7
x	-0.14	-0.19	-0.32	-0.51

Määrää luennoissa esitetyllä *pienimmän neliösumman menetelmällä* havainnoista saatava estimaatti jousivakiolle k minimoimalla neliösumma

$$f(k) = \sum_{i=1}^{4}(F_i - (-kx_i))^2. \tag{2}$$

b) Mallin (1) havaintoja vastaava selitysaste R^2 määritellään kaavalla

$$R^2 = 1 - \frac{s_{\text{Res}}}{(n-1)s_{FF}}, \tag{3}$$

missä s_{FF} on muuttujan F otosvarianssi ja residuaalisummassa

$$s_{\text{Res}} = \sum_{i=1}^{n}(F_i - (-kx_i))^2$$

esiintyvä jousivakio k on (2):n minimikohta. Mitä arvoja kaavalla (3) määritelty selitysaste saa? Laske mallin (1) selitysaste.

c) Laske otoksen korrelaatiokerroin r. Onko b)-kohdassa laskettu selitysaste R^2 sama kuin korrelaatiokertoimen neliö r^2?

Ratkaisu:

a) Lasketaan kaavan (2) määrittelemän funktion derivaatan nollakohta muuttujan k suhteen:

$$f'(k) = \sum_{i=1}^{4} 2(F_i + kx_i)x_i = 0 \Leftrightarrow k = -\frac{\sum_i F_i x_i}{\sum_i x_i^2}.$$

Jousivakion pienimmän neliösumman estimaatiksi saadaan

$$\hat{k} \overset{\text{merk.}}{=} -\frac{\sum_i F_i x_i}{\sum_i x_i^2} \overset{\text{sij.}}{\approx} 14.4.$$

b) Residuaalisumman saadaan suoraan tilastollisesta ohjelmistosta tai hyödyntämällä a)-kohdassa laskettua pienimmän neliösumman estimaattia k pisteittäisten residuaalien

$$r_i = F_i + kx_i$$

laskemiseksi. Kun pisteittäiset residuaalit on laskettu, saadaan residuaalisummaksi

$$s_{\text{Res}} = \sum_{i=1}^{4} r_i^2 \overset{\text{sij.}}{\approx} 0.34194.$$

Muuttujan F otosvarianssi $s_{FF} = s_F^2$ saadaan suoraan (funktio)laskimesta, jolloin mallin (1) selitysasteeksi tulee

$$R^2 = 1 - \frac{s_{\text{Res}}}{(n-1)s_F^2} \approx 0.977.$$

Koska $s_{Res} \geq 0$, niin $R^2 \leq 1$. Huomaa, että R^2 voi (merkinnästään huolimatta) saada myös negatiivisia arvoja. Näin voi käydä, jos F-havaintojen keskiarvo $\overline{F}$ selittää paremmin muuttujan F satunnaisvaihtelua kuin malli (1). Tällöin kappaleeseen kohdistuva voima olisi likimain vakio poikkeamasta F riippumatta, mikä ei ole sopusoinnussa itse ilmiön kanssa, ellei jousi satu olemaan poikki.

c) Korrelaatiokertoimeksi saadaan suoraan jopa funktiolaskimesta. Jos et tiedä miten korrelaatiokertoimen saa laskimesta, laita vaikkapa Googleen hakusanoiksi *correlation using calculator* ja katso neuvoa sopivasta Youtube-videosta. Korrelaatiokerroin löytyy *regressio*-valikon, jota käytetään myös jatkossa, alta.
Korrelaatiokertoimeksi tulee

$$r \approx -0.9923,$$

jonka neliöksi saadaan

$$r^2 \approx 0.985.$$

Tämän mukaan täydellinen lineaarinen malli $F = a + bx$ sopisi paremmin havaintoaineistoon. Tämä on varsin luontevaa, sillä täydellisessä mallissa sovitettavia parametreja on kaksi, joka antaa enemmän pelivaraa suoran sovittamiseen, kun taas ”vajaassa” mallissa (1) on ainoastaan yksi sovitettava parametri. Tästä ei kuitenkaan voi vetää johtopäätöstä, että täydellinen lineaarinen malli olisi parempi, sillä nollasta poikkeava vakiotermi a ei ole ilmiön fysiikan kanssa sopusoinnussa (jos olisi $a \neq 0$, niin mallin mukaan tasapainoasemassa $x = 0$ olevaan kappaleeseen kohdistuisi nollasta poikkeava voima, mikä ei tietenkään ole järkevää).

2. Lehtori K:n pojan pituutta [cm] ja painoa [g] seurattiin hänen ensimmäisinä elinkuukausinaan ja saatiin seuraava taulukko

Pituus	53	56.3	58.7	62.8	65.5	66.5	69	71.7
Paino	4265	5030	6095	7295	8460	8990	9500	9620

a) Laske muuttujien välinen korrelaatiokerroin ja määrää selitysaste.

b) Määrää havaintoja vastaava regressiosuora. Piirrä havaintopisteet ja regressiosuora samaan koordinaatistoon.

c) Laske mallin antama painoennuste, kun pituus on 180 [cm]. Mitä voit sanoa mallin antamasta ennusteesta?

d) Olkoon pituus muuttuja X ja paino muuttuja Y. Muodostetaan regressiomalli (6) muuttujien välille. Laske kertoimien α ja β symmetrinen 95% luottamusväli.

e) Testaa riskitasolla $\alpha = 5\%$, onko muuttujien X ja Y välillä lineaarista riippuvuutta.

Ratkaisu: Olkoon pituus muuttuja X ja paino Y. Havainnoista voidaan laskea suureet

$$\begin{aligned} n &= 8 & n &= 8 \\ \overline{x} &= 62.9375 & \overline{y} &= 7406.875 \\ s_x &= 6.47610 & s_y &= 2075.33 \end{aligned}$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \approx 13284.8$$

a) Korrelaatiokerroin on

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} = \frac{s_{xy}}{s_x s_y} \approx \underline{0.988454}$$

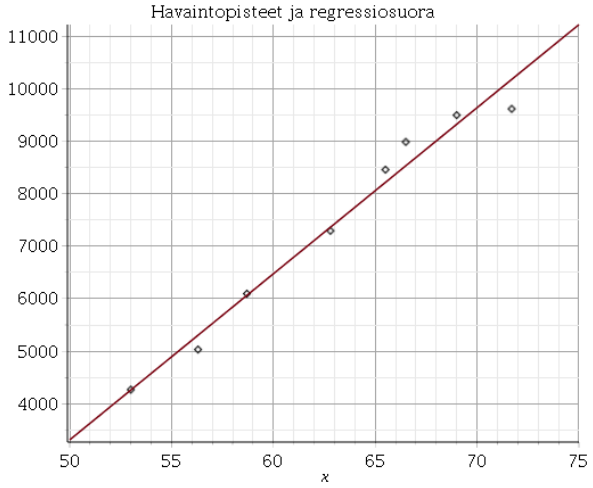
Selitysaste saadaan siitä: $r^2 = \underline{0.977042}$.

b) Regressionsuoran $y = a + bx$ kertoimet ovat

$$b = \frac{s_{xy}}{s_{xx}} = \frac{s_{xy}}{{s_x}^2} \approx \frac{13284{,}8}{6{,}47610^2} \approx 316.760$$

$$a = \overline{y} - b\overline{x} \approx 7406.875 - 316{,}760 \cdot 62{,}9375 \approx -12529.2$$

eli regressiosuoran yhtälö on $\underline{y = 316.760 \cdot x - 12529.2}$



c) Painoennuste, kun pituus on 180, saadaan sijoittamalla $x = 180$ regressiosuoran yhtälöön

$$y = b \cdot 180 + a \approx 316.760 \cdot 180 - 12529.2 \approx 44488[g] \stackrel{\wedge}{=} \underline{44{,}5[kg]}$$

d) *Kulmakertoimen β luottamusväli:* Lasketaan ensin kulmakertoimen β luottamusväli. Jos käytössä on tarpeeksi edistyksellinen graafinen laskin tai tilastollinen ohjelmisto kuten esimerkiksi Excel tai R, syötetään havaintoaineisto laskukoneeseen, joka laskee luottamusvälin. Esimerkiksi TI-Nspiresta löytyy komento *Linear Reg t Intervals*, jossa täytyy ensin kertoa mille parametrille luottamusvälin haluaa laskea. Sen jälkeen avautuvaan valikkoon annetaan tiedot, mistä x:n ja y:n arvot löytyvät ja annetaan luottamustaso, joka on $1 - \alpha = 0.95$, missä $\alpha = 0.05$ on riskitaso. Sitten vain tulkitaan laskukoneen antama tulostus.
Luottamusvälin voi määrätä myös käsin funktiolaskinta hyödyntäen, mikä tosin on hieman työläämpää. Esimerkiksi Youtubesta löytyy opetusvideoita funktiolaskimen tilastollisista ominaisuuksista. Vaikkapa hakusanat *linear regression using a calculator* antaa osumia opetusvideoihin, joiden avulla opastetaan miten data syötetään laskimeen ja miten laskimesta saadaan kaivettua tarvittavat tunnusluvut.
Kaavakokoelmasta valitaan oikeaksi testimuuttujaksi

$$T = \sqrt{n-1}s_x \frac{\frac{S_{xy}}{s_{xx}} - \beta}{S_r} \sim t_{n-2}, \tag{4}$$

missä $n - 2 = 6$ on vapausasteiden lukumäärä. Itse luottamusvälin määrääminen menee tämän jälkeen samalla tavalla kuin mille tahansa t-jakautuneelle muuttujalle. Katso mallia esimerkiksi harjoituksesta 4 tai luentoesimerkeistä.
Funktiolaskimesta saadaan suoraan muuttujan x otoskeskihajonta s_x. Laskimesta saa kaivettua myös kovarianssin suhteellisen helposti, jos käyttää kaavaa

$$s_{xy} = \frac{1}{n-1}\left(\sum_i x_i y_i - n\overline{xy}\right).$$

Huomaa, ettei otoskovarianssia välttämättä edes tarvita, sillä me tarvitaan ainoastaan osamäärä s_{xy}/s_{xx}, joka ei ole mitään muuta kuin regressiosuoran kulmakerroin b.
Jäännösvarianssile kannattaa käyttää laskukaavaa

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \frac{n-1}{n-2}(1-r^2)s_{yy}, \tag{5}$$

sillä $s_{yy} = s_y^2$ ja r^2 saadaan suoraan laskimesta. Huomaa jälleen, ettei tätäkään välttämättä tarvita, sillä meille riittää ainoastaan osamäärä

$$\frac{s_r}{\sqrt{n-1}s_x},$$

joka on estimaatin b keskivirhe (Standard Error) ja jonka (graafinen) laskin antaa suoraan. Funktiolaskimen käyttäjä sen sijaan tarvitsee kaavaa (5).
Kriittiseksi arvoksi $t_{0.05}$ saadaan t-jakauman taulukosta $t_{0.05} = 2.447$ käyttämällä vapausasteita $f = 6$ ja 2-suuntaista testiä. Välivaiheiden jälkeen kulmakertoimen luottamusväliksi I_β saadaan

$$I_\beta = [268{,}365],$$

kun luottamusvälin rajat ilmoitetaan kolmen merkitsevän numeron tarkkuudella.

Leikkauspisteen α luottamusväli: Tässä voidaan menetellä kuten yllä. Jos käytössä on ainoastaan funktiolaskin, joudutaan temppuilemaan hieman enemmän. Tilanteeseen sopiva testimuuttuja on

$$T = \frac{1}{\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}} \frac{\overline{Y} - \overline{x}\frac{S_{xy}}{s_{xx}} - \alpha}{S_r} \sim t_{n-2}.$$

Kriittinen arvo $t_{0.05} = 2.447$ on sama kuin kulmakertoimelle. Testimuuttujan otoksessa saama arvo voidaan laskea samaan tapaan kuin kulmakertoimelle. Otostunnusluvut $\overline{x}$,$\overline{y}$,s_x ja $s_{xx} = (n-1)s_x^2$ saadaan näppärästi laskimesta. Jäännösvarianssille s_r^2 kannattaa hyödyntää laskukaavaa (5) tai ottaa se jo kertaalleen laskettuna edeltä.
Leikkauspisteen α luottamusväliksi I_α saadaan

$$I_\alpha = [-15600, -9460],$$

jos käytetään kolmen merkitsevän numeron tarkkuutta.

e) Lineaarista riippuvuutta tutkitaan korrelaatiokertoimen ρ avulla. Sopivat hypoteesit ovat $H_0 : \rho = 0$ vastaan $H_1 : \rho \neq 0$. Luentojen mukaan yhtäpitävät hypoteesit ovat $H_0 : \beta = 0$ vastaan $H_1 : \beta \neq 0$. Esimerkiksi TI-Nspiresta löytyy komento *Linear Reg t Test*, jolla kulmakerroin β voidaan testata normaalin t-testin tapaan. Tähän kohtaan sopii sama opastus kuin luottamusvälien laskemisessa.
Toinen ja helpompi tapa on hyödyntää jo laskettua luottamusväliä. Tällöin meillä riittää ainoastaan tutkia, kuuluuko 0 luottamusvälille vai ei.
Koska $0 \notin I_\beta$, täytyy H_0 hylätä, joten johtopäätös on $H_1 : \rho \neq 0$, eli *muuttujat X ja Y ovat tilastollisesti lineaarisesti riippuvia.*

3. Tutkittiin erään kemiallisen yhdisteen liukenemista veteen. Oheisessa taulukossa on ilmoitettu 100 grammaan vettä liuenneen yhdisteen määrä y [g] eri lämpötiloissa x [$°C$].

x	y		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

a) Piirrä havaintoja vastaava sirontakuvio.

b) Laske muuttujien välinen korrelaatiokerroin ja määrää selitysaste.

c) Määrää havaintoja vastaava regressiosuora ja piirrä havaintoja vastaavat pisteittäiset residuaalit.

d) Muodostetaan regressiomalli

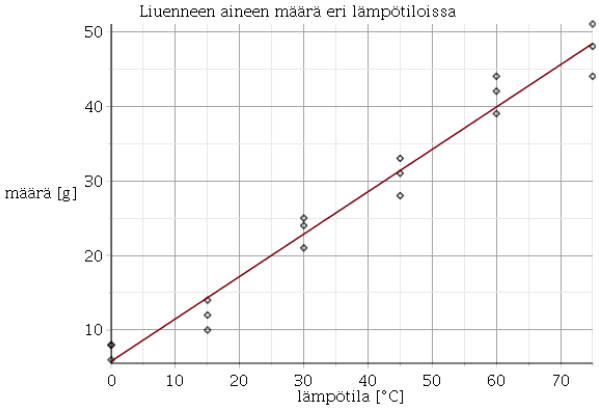
$$Y_i = \alpha + \beta x_i + \epsilon_i \tag{6}$$

muuttujien välille. Laske kulmakertoimen β symmetrinen 95% luottamusväli.

e) Testaa riskitasolla 5% hypoteesi $H_0 : \beta = 0$ vastaan $H_1 : \beta \neq 0$.

Ratkaisu:

a) Nyt datan syöttämisessä täytyy huomioida, että kutakin x:n arvoa kohti löytyy 3 mittausta muuttujasta y. Näin ollen kukin x:n arvo syötetään kolmeen kertaan, jolloin xy-vastinpisteitä tulee yhteensä $n = 18$ kappaletta. Alla olevaan kuvaan on piirretty havaintoja vastaava sirontakuvio ja regressiosuora.



b) Lasketaan ensin havaintoaineistosta otostunnusluvut

$$\begin{aligned} n &= 18 & n &= 18 \\ \mu_x &= 37{,}5000 & \mu_y &= 27{,}1111 \\ s_x &= 26{,}3600 & s_y &= 15{,}1691 \end{aligned}$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \approx 394{,}412$$

Korrelaatiokerroin on

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} = \frac{s_{xy}}{s_x s_y} \approx \underline{0,986372}$$

Selitysaste saadaan siitä

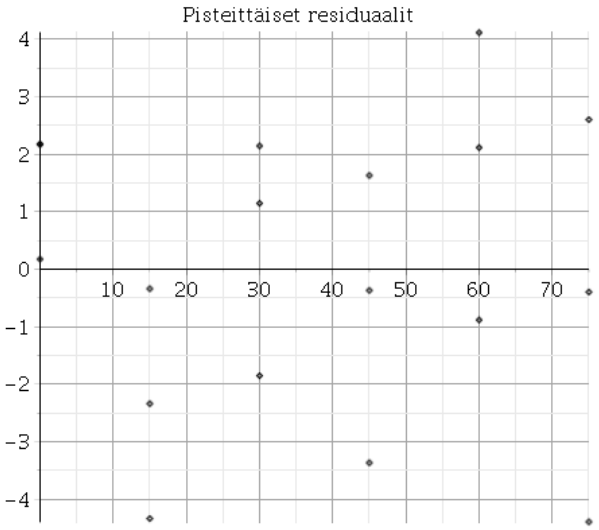
$$r^2 = \underline{0{,}972930}$$

c) Regressionsuoran $y = a + bx$ kertoimet ovat

$$b = \frac{s_{xy}}{s_{xx}} = \frac{s_{xy}}{{s_x}^2} \approx \frac{394{,}412}{26{,}3600^2} \approx 0.567619$$
$$a = \overline{y} - b\overline{x} \approx 27{,}1111 + 0.567619 \cdot 37{,}5000 \approx 5.82540$$

eli regressiosuoran yhtälö on $\underline{y = 0.567619 \cdot x + 5.8253}$

Havaintoja vastaavat pisteittäiset residuaalit $(x_i, y_i - \hat{y}_i)$:



d) Tähän käy samat neuvot kuin tehtävän 2 vastaavassa kohdassa. Nyt havaintoja on yhteensä $n = 18$ kappaletta, joten funktiolaskimen käyttö ei enää ole tarkoituksenmukaista.
Tehtävän 2 opastusta noudattaen kulmakertoimen luottamusväliksi saadaan

$$I_\beta = [0.517, 0.618],$$

kun käytetään kolmen merkitsevän numeron tarkkuutta.

e) Myös tähän kohtaan käyvät samat neuvot kuin tehtävän 2 vastaavaan kohtaan sillä erotuksella, että kohtaa ei ole järkevää laskea funktiolaskimella. Helpoin tapa testata hypoteesit $H_0 : \beta = 0$ vastaan $H_1 : \beta \neq 0$ on jo lasketun luottamusvälin hyödyntäminen.
Koska $0 \notin I_\beta$, hyväksytään vaihtoehtoinen hypoteesi $H_1 : \beta \neq 0$, joten *tilastollisesti kulmakerroin poikkeaa nollasta riskitasolla* $\alpha = 0.05$.

Huomautus! Vaikka edellä olevissa kohdissa d) ja e) laskimesta on suuri hyöty ja sellaista ilman muuta kannattaa hyödyntää, **on luottamusvälin määräämisen ja hypoteesien testauksen idea ja mekanismi oltava näkyvissä**.