MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi Gripenberg, Kiiski
Harjoitus 6
1–5.12.2014, viikko 49

P1. Muuttujien x ja Y havaitut arvot ovat

x	0.4	1.1	1.4	2.2	2.3	2.4
y	-0.5	1.4	1.6	3.3	4.1	4.2

Tästä otoksesta on laskettu seuraavat tunnusluvut:

s_x^2	s_y^2	s_{xy}
0.64267	3.3950	1.4640

- Määritä regressiomallin $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ kertoimien β_0 ja β_1 (pienimmän neliösumman) estimaatit.
- Piirrä pistediagrammi sekä estimoitu suora.
- Laske jäännösvarianssin estimaatti s^2 ja testaa nollahypoteesia $\beta_0 = -1$ merkitsevyystasolla 0.05 (olettaen, että satunnaismuuttujat ε_j ovat riippumattomia ja $\mathrm{N}(0,\sigma^2)$-jakautuneita).

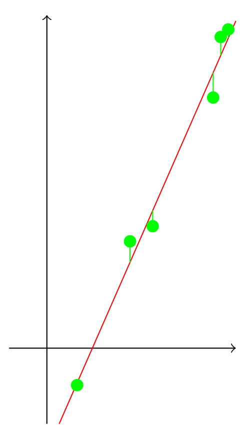
Ratkaisu: (a) Tässä tarvitaan keskiarvot $\overline{x}$ ja $\overline{y}$ ja ne ovat

$$\overline{x} = 1.6333,$$
$$\overline{y} = 2.35.$$

Parametrien β_0 ja β_1 estimaatit ovat näiden tietojen perusteella:

$$b_1 = \frac{s_{xy}}{s_x^2} \approx 2.278,$$
$$b_0 = \overline{y} - b_1\overline{x} \approx -1.3707.$$

(b)



(c) Jännösvarianssin estimaattia varten tarvitsemme korrelaatiokeroimen ja se on $r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = 0.99112$ joten jäännösvarianssin estimaatti on

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\big(b_0 + b_1 x_i - y_i\big)^2 = \frac{n-1}{n-2}s_y^2(1-r_{xy}^2) \approx 0.075.$$

Parametrin β_0 estimaattorista B_0 tiedämme, että testimuuttuja

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)}},$$

on t$(n-2)$-jakautunut. Kun nollahypoteesi on $\beta_0 = -1$ niin tämän testimuuttujan arvoksi tulee tässä tapauksessa -1.3557. Koska nollahypoteesin negaatio on kaksisuuntainen niin p-arvoksi tulee

$$p = 2F_{\mathrm{t}(4)}(-1.3557) = 0.24667,$$

joten nollahypoteesia ei hylätä.

P2. Otoksesta (x_j, y_j), $j = 1, 2 \ldots, 16$ laskettiin seuraavat tunnusluvut:

$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	r_{xy}
0.5	4.8	4.4	6.2	0.7

- Määritä regressiomallin $Y = \beta_0 + \beta_1 x + \varepsilon$ kertoimien β_0 ja β_1 (pienimmän neliösumman) estimaatit.
- Laske jäännösvarianssin estimaatti s^2 ja testaa nollahypoteesia $\beta_1 \leq 0.3$ merkitsevyystasolla 0.05 (olettaen, että satunnaismuuttujat ε_j ovat riippumattomia ja N$(0, \sigma^2)$-jakautuneita).
- Mikä on mallin selitysaste?

Ratkaisu: Tässä tehtävässä ei ole annettu otoskovarianssia s_{xy} mutta koska $r_{xy} = \dfrac{s_{xy}}{s_x \cdot s_y}$ niin $s_{xy} = r_{xy} \cdot s_x \cdot s_y$ joten saamme parametrien β_1 ja β_0 estimaateiksi

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} = 0.83094$$
$$b_0 = \overline{y} - b_1\overline{x} = 4.3845.$$

Jäännösvarianssin estimaatiksi tulee nyt

$$s^2 = \frac{n-1}{n-2}s_y^2(1 - r_{xy}^2) = 3.3879.$$

Kun testaamme nollahypoteesia $\beta_1 \leq 0.3$ niin käytämme testimuuttujaa

$$W_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}},$$

joka on t$(n-2)$-jakautunut. Tämän testimuuttujan arvoksi tulee

$$w_1 = \frac{0.83094 - 0.3}{\sqrt{\frac{3.3879}{15 \cdot 4.4}}} = 2.3434.$$

Koska nollahypoteesin negaatio on yksisuuntainen niin p-arvoksi tulee

$$p = 1 - F_{\mathrm{t}(14)}(2.3434) = 0.017197,$$

joten nollahypoteesi hylätään merkitsevyystasolla 0.05.

Määritelmän mukaan mallin selitysaste on $r_{xy}^2 = 0.49$.

P3. Fossiiliset hiilidioksidipäästöt (CO2-foss) olivat vuosina 2008–2012 kotitalouksien osalta seuraavat (yksikkönä miljoona tonnia)

2008	2009	2010	2011	2012
6.79	6.32	6.25	5.60	5.66

Määritä näiden lukujen perusteella 95%:n luottamusväli vuoden 2013 päästöjen odotusarvolle kun oletat että päästöja voidaan kuvata mallilla $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ missä Y_j on päästöt vuonna x_j ja satunnaismuuttujat ε_j ovat riippumattomia ja $\mathrm{N}(0, \sigma^2)$-jakautuneita.

Voit käyttää hyväksi seuraavia tunnuslukuja:

$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
2010	6.124	2.5	0.24693	-0.745

ja muista että jos vaihdat vuosiluvut luvuiksi $x_j - 2013$ ainoastaan $\overline{x}$ muuttuu ja sinun pitää laskea luottamusväli kertoimelle β_0.

Vastaus: $[4.62, 5.84]$.

Ratkaisu: Jos $z_j = x_j - 2013$ niin $\overline{z} = -3$ mutta $s_z^2 = s_x^2$ ja $s_{zy} = s_{xy}$.

Näiden lukujen perusteella voimme laskea estimaatit parametreille β_0 ja β_1 regressiomallissa $Y_j = \beta_0 + \beta_1 z_j + \varepsilon_j$ ja saamme.

$$b_1 = \frac{s_{zy}}{s_x^2} = -0.298,$$
$$b_0 = \overline{y} - b_1 \overline{z} = 5.23.$$

Mallin $Y_j = \beta_0 + \beta_1 z_j + \epsilon_j$ ennustamat vuoden 2013 päästöt saadaan sijoittamalla $z_j = 0$ jolloin päästöjen odotusarvoksi tulee β_0 jonka estimaatti on $b_0 = 5.23$.

Jotta voisimme laskea 95% luottamusvälin meidän pitää laskea jäännösvarianssin estimaatti s^2 ja sitä varten laskemme ensin otoskorrelaatiokertoimen:

$$r_{xy} = r_{zy} = \frac{s_{zy}}{s_z\, s_y} = b_1 \frac{s_z}{s_y} = -0.94820,$$
$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r_{xy}^2) = 0.03323.$$

Jos $B_0 = \overline{Y} - \dfrac{S_{xy}}{s_x^2}\overline{z}$ niin B_0 on parametrin β_0 estimaattori ja pätee

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{z}^2}{(n-1)s_z^2}\right)}} \sim \mathrm{t}(n-2).$$

Nyt $\Pr(-3.1824 \leq W_0 \leq 3.1824) = 0.95$ koska $3.1824 = F_{\mathrm{t}(3)}^{-1}(0.975)$ josta seuraa, että

$$\Pr\left(B_0 - 3.1824\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{z}^2}{(n-1)s_z^2}\right)} \leq \beta_0 \leq B_0 + 3.1824\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{z}^2}{(n-1)s_z^2}\right)}\right) = 0.95.$$

Jos nyt sijoitetaan B_0:n paikalle sen estimaatti b_0 ja S^2:n paikalle sen estimaatti s^2 samoin kuin $\overline{z}$, s_z^2 ja n niin saadaan luottamusväliksi

$$\big[4.62, 5.84\big].$$

P4. Tutkija T tutki monen eri satunnaismuuttujan välisiä riippuvuuksia ja kun hän laski 400 otoskorrelaatiokerrointa hän löysi 27 kertaa kertoimen joka poikkesi nollasta merkitsevyystasolla 5%.
Valitse ensin nollahypoteesiksi, että kaikki satunnaismuuttujat ovat riippumattomia. Tästä voit johtaa uuden nollahypoteesin joka koskee todennäköisyyttä, että löydat korrelaatiokertoimen joka poikkeaa nollasta merkitsevyystasolla 5% ja voit ottaa tämän nollahypoteesin sellaiseksi, että vaihtoehto sille on yksisuuntainen. Testaa tätä nollahypoteesia merkitsevyystasolla 5% käyttäen tutkija T:n tuloksia.

Ratkaisu: Jos kaikki tutkitut satunnaissuureet ovat riippumattomia niin kaikki korrelaatiokertoimet ovat nollia, eli korrelaatiokertoimien testauksessa käytetty nollahypoteesi on voimassa. Koska tutkija T testasi tämän nollahypoteesin merkitsevyystasolla 5% niin merkitsevyystason määritelmästä seuraa, että todennäköisyys että nollahypoteesi hylätään, kun se on voimassa, on 5%. Näin ollen otamme nollahypoteesiksi, että todennäköisyys, että tutkija T löytää merkitsevästi nollasta poikkeavan korrelattiokertoimen on korkeintaan 0.05. (Epäyhtälö otetaan tässä toisaalta yksinkertaistamaan laskuja ja toisaalta koska ei tunnu järkevältä että hylkäisimme nollahypoteesin riippumattomuudesta jos olisi löytynyt vain vähän nollasta poikkeavia korrelattiokertoimia.)

Jos X on merkitsevästi nollasta poikkeavien otoskorrelaatiokertoimien lukumäärä niin nollahypoteesin (ääritapauksen) nojalla X on $\mathrm{Bin}(400, 0.05)$-jakautunut satunnaismuuttuja. Näin ollen

$$p = \Pr(X \geq 27) = 1 - \Pr(X \leq 26) = 1 - F_{\mathrm{Bin}(400,0.05)}(26) \approx 0.0726.$$

Jos käytämme normaaliapproksimaatiota niin vastaava lasku on seuraavanlainen:

$$\begin{aligned} p = \Pr(X \geq 27) &= \Pr\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \geq \frac{27 - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}\right) \\ &= \Pr\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \geq \frac{27 - 0.05 \cdot 400}{\sqrt{0.05 \cdot 0.95 \cdot 400}}\right) \\ &= \Pr\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \geq 1.60591\right) \approx 1 - F_{\mathrm{N}(0,1)}(1.60591) = 0.0541. \end{aligned}$$

Joka tapauksessa nollahypoteesia ei hylätä merkitsevyystasolla 0.05 eli ei ole erityisen epätodennäköistä että tutkija T löytää näin monta merkitsevästi nollasta poikkeavia otoskorrelaatiokertoimia.

P5. Havaintoaineistosta (x_i, y_i), $i = 1, \ldots, n$ on laskettu regressiosuora $y = b_0 + b_1 x$ kun x on "selittävä" muuttuja (eli kun oletetaan, että $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) ja (käänteisregressiosuora) $x = a_0 + a_1 y$ kun y on "selittävä" muuttuja (eli kun oletetaan, että $X_i = \alpha_0 + \alpha_1 y_i + \epsilon_i$). Lisäksi on laskettu otoskorrelaatiokerroin r_{xy} ja on todettu, että nollahypoteesi $\mathrm{H}_0 : \beta_1 = 0$ voidaan hylätä merkitsevyystasolla 0.01. Mitä voidaan sanoa seuraavista väitteistä (olettaen, että $s_x > 0$ ja $s_y > 0$):

(a) Suorat $y = b_0 + b_1 x$ ja $x = a_0 + a_1 y$ ovat samat jos ja vain jos $|r_{xy}| = 1$.
(b) Jos testataan hypoteesia $\mathrm{H}_0 : \alpha_1 = 0$ ne sekin voidaan hylätä merkitsevyystasolla 0.01.

Vihje: Regressiosuorat voidaan kirjoittaa muodossa $y - \overline{y} = b_1(x - \overline{x})$ *ja* $x - \overline{x} = a_1(y - \overline{y})$.

Ratkaisu: (a) Molemmat suorat kulkevat pisteen $(\overline{x}, \overline{y})$ kautta joten suorat ovat samat jos kulmakertoimet ovat samat. Nyt $b_1 = r_{xy}\frac{s_y}{s_x}$ ja $a_1 = r_{xy}\frac{s_x}{s_y}$ ja kulmakertoimet ovat b_1 ja $\frac{1}{a_1}$ joten nämä ovat samat täsmälleen silloin kun $r_{xy} = \pm 1$ eli väite pätee.

(b) Testisuureen arvo voidaan molemmissa tapauksissa (koska nimenomaan testataan hypoteeseja $\beta_1 = 0$ ja $\alpha_1 = 0$) esittää muodossa $\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ joten kyseessä on molemmissa tapauksessa sama testi, eli väite pätee.
