

Mat-1.2620 Sovellettu todennäköisyyslaskenta B**12. harjoitukset / Ratkaisut****Aiheet: Yhden selittäjän lineaarinen regressiomalli****Avainsanat:**

Ehdollinen jakauma, Ehdollinen odotusarvo, Ehdollinen varianssi, Estimointi, F -jakauma, F -testi, Homoskedastisuus, Jakaumaoletus, Jäännöseliösumma, Jäännöstermi, Jäännösvarienssi, Keskihajonta, Kaksiulotteinen normaalijakauma, Kokonaiseliösumma, Korrelaatio, Korreloimattomuus, Kovarianssi, Kriittinen arvo, Kulmakerroin, Lineaarinen regressiomalli, Luottamuserroin, Luottamustaso, Luottamusväli, Mallineliösumma, Merkitsevyystaso, Momenttimenetelmä, Nollahypoteesi, Normaalijakauma, Otos, Painopiste, Parametri, p -arvo, Pienimmän neliösumman menetelmä, Rakenneos, Regressiofunktio, Regressiokerroin, Regressiomalli, Regressiosuora, Residuaali, Riippumattomuus, Riippuvuus, Satunnainen osa, Selitettävä muuttuja, Selittäjä, Selittävä muuttuja, Selitystaste, Sovite, Suurimman uskottavuuden menetelmä, Systemaattinen osa, t -jakauma, t -testi, Testi, Vakiotermin, Vapausasteet, Varianssi-analyysihajotelma

Yhden selittäjän lineaarinen regressiomalli**Yhden selittäjän lineaarinen regressiomalli**

Tavanomaisen yhden selittäjän lineaarisen regressiomallin yleinen muoto on

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

jossa

y_i = **selitettävän muuttujan** y satunnainen ja havaittu arvo havaintoyksikössä i

x_i = **selittäjän** (selittävän muuttujan) x ei-satunnainen ja havaittu arvo havaintoyksikössä i

ε_i = **jäännös-** eli **virhetermin** ε satunnainen ja ei-havaittu arvo havaintoyksikössä i

β_0 = ei-satunnainen ja tuntematon vakio (**vakioselittäjän regressiokerroin**)

β_1 = selittäjän x ei-satunnainen ja tuntematon **regressiokerroin**

Mallin jäännöstermistä ε tehdään seuraavat *stokastiset oletukset*:

$$(1) \quad \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ ovat riippumattomia}$$

$$(2) \quad \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$$

Sanomme, että jäännöstermien $\varepsilon_i, i = 1, 2, \dots, n$ varianssi σ^2 on mallin **jäännösvarienssi**.

Huomaa, että oletuksista (1) ja (2) seuraa, että kaikilla jäännöstermeillä $\varepsilon_i, i = 1, 2, \dots, n$ on sama varianssi eli ne ovat **homoskedastisia** ja lisäksi jäännöstermit ovat **korreloimattomia**.

Jos malli

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

ja sen osat toteuttavat kaikki em. oletukset, sanomme että *malli on tavanomainen yhden selittäjän lineaarinen regressiomalli* tai, että *malli toteuttaa tavanomaiset oletukset yhden selittäjän lineaariselle regressiomallille*.

Satunnaisen selittäjän tapaus

Jos tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

selittäjän x arvot ovat *satunnaisia*, mutta jäännöstermiä ε koskeva oletus (2) yllä voidaan *korvata* esitetyissä ns. *standardioletuksissa* oletuksella

$$(2)' \quad \varepsilon_i | x_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$$

niin *jatkossa esitettävä teoria pätee sopivasti modifioituna*. Oletus (2)' tarkoittaa sitä, että satunnaismuuttujan ε_i *ehdollinen jakauma* ehdolla x_i on oletettu normaaliseksi.

Yhden selittäjän lineaarisen regressiomallin systemaattinen osa ja satunnainen osa

Olkoon

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

tavanomaiset oletukset toteuttava yhden selittäjän lineaarinen regressiomalli. Tällöin

$$E(y_i) = \beta_0 + \beta_1 x_i, i = 1, 2, \dots, n$$

$$\text{Var}(y_i) = \sigma^2, i = 1, 2, \dots, n$$

Sanomme, että odotusarvo

$$E(y_i) = \beta_0 + \beta_1 x_i, i = 1, 2, \dots, n$$

muodostaa mallin **systemaattisen osan** eli **rakenneosan** ja

$$\varepsilon_i = y_i - E(y_i) = y_i - \beta_0 - \beta_1 x_i, i = 1, 2, \dots, n$$

muodostaa mallin **satunnaisen osan**.

Regressiosuora

Tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

systemaattinen osa

$$E(y_i) = \beta_0 + \beta_1 x_i, i = 1, 2, \dots, n$$

määrittelee **regressiosuoran**

$$y = \beta_0 + \beta_1 x$$

jossa

β_0 = suoran **vakiotermin**

β_1 = suoran **kulmakerroin**

Regressiosuoran kulmakertoimen tulkinta

Oletetaan, että selittäjän x arvo kasvaa yhdellä yksiköllä:

$$x \rightarrow x + 1$$

Regressiokerroin β_1 kertoo paljonko *selitettävän muuttujan y vastaava odotettavissa oleva arvo*

$$E(y) = \beta_0 + \beta_1 x$$

muuttuu:

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x \\ &\rightarrow \beta_0 + \beta_1 (x + 1) \\ &= \beta_0 + \beta_1 x + \beta_1 \\ &= E(y) + \beta_1 \end{aligned}$$

Regressiokertoimien estimointi

Mallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

regressiokertoimien (parametrien) β_0 ja β_1 **pienimmän neliösumman (PNS-) estimaattorit** saadaan *minimoimalla neliösumma*

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

regressiokertoimien β_0 ja β_1 suhteen.

Regressiokertoimien β_0 ja β_1 *PNS-estimaattoreiksi* saadaan

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = r_{xy} \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$

PNS-estimaattoreiden kaavoissa

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

ovat x -havaintojen ja y -havaintojen *aritmeettiset keskiarvot*,

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

ovat x -havaintojen ja y -havaintojen *otosvarianssit*,

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

on x -havaintojen ja y -havaintojen *otoskovarianssi* ja lisäksi

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

on x -havaintojen ja y -havaintojen *otoskorrelaatiokerroin*.

Estimoitu regressiosuora

Tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

regressiokertoimien β_0 ja β_1 *PNS-estimaattorit* b_0 ja b_1 ja määrittelevät suoran

$$y = b_0 + b_1 x$$

Sovitteet ja residuaalit

Estimoidun mallin **sovitteet** saadaan kaavalla

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, 2, \dots, n$$

Estimoidun mallin **residuaalit** saadaan kaavalla

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, i = 1, 2, \dots, n$$

Malli selittää sitä paremmin selitettävän muuttujan y käyttäytymistä mitä lähempänä sovitteet ovat selitettävän muuttujan y havaittuja arvoja eli mitä pienempiä ovat estimoidun mallin residuaalit.

Koska mallissa on mukana *vakio*, niin sovitteiden summa yhtyy selitettävän muuttujan havaittujen arvojen summaan:

$$\begin{aligned} \sum_{i=1}^n \hat{y}_i &= \sum_{i=1}^n (b_0 + b_1 x_i) = \sum_{i=1}^n b_0 + b_1 \sum_{i=1}^n x_i \\ &= nb_0 + nb_1 \bar{x} = n(\bar{y} - b_1 \bar{x}) + nb_1 \bar{x} \\ &= n\bar{y} = \sum_{i=1}^n y_i \end{aligned}$$

Lisäksi koska mallissa on mukana *vakio*, residuaalien summa = 0:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{y}_i = 0$$

Varianssianalyysihajotelma

Olkoon

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)s_y^2$$

selitettävän muuttujan y arvojen vaihtelua kuvaava **kokonaisneliösumma** ja olkoon

$$SSE = \sum_{i=1}^n e_i^2$$

estimoidun mallin PNS-residuaalien vaihtelua kuvaava **jäännöseliösumma**.

Voidaan osoittaa, että

$$SSE = \sum_{i=1}^n e_i^2 = (1 - r_{xy}^2) \sum_{i=1}^n (y_i - \bar{y})^2 = (1 - r_{xy}^2) SST$$

missä

$$r_{xy} = x\text{-havaintojen ja } y\text{-havaintojen otoskorrelaatiokerroin}$$

Koska aina pätee

$$|r_{xy}| < 1$$

niin

$$SSE \leq SST$$

Määritellään estimoidun mallin **malli-** (eli **regressio-**) **neliösumma** kaavalla

$$SSM = SST - SSE$$

Voidaan osoittaa, että

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Kokonaisneliösumman SST hajotelmaa

$$SST = SSM + SSE$$

neliösummien SSM ja SSE summaksi kutsutaan **varianssianalyysihajotelmaksi**.

Varianssianalyysihajotelmassa selitettävän muuttujan y *kokonaisvaihtelua kuvaava neliösumma* SST on hajotettu kahteen osaan, joista *mallineliösumma* SSM kuvaa sitä osaa kokonaisneliösummasta, jonka estimoitu malli on selittänyt ja *jäännöseliösumma* SSE kuvaa sitä osaa kokonaisneliösummasta, jota estimoitu malli ei ole selittänyt.

Selitysaste

Varianssianalyysihajotelma motivoi määrittelemään estimoidun mallin **selitysasteen** kaavalla

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$$

jossa

SST = selitettävän muuttujan y arvojen vaihtelua kuvaava *kokonaisneliösumma*

SSE = estimoidun mallin residuaalien vaihtelua kuvaava *jäännöseliösumma*

SSM = estimoidun mallin *malli-* (eli *regressio-*) *neliösumma*

Varianssianalyysihajotelmasta seuraa, että aina pätee

$$0 \leq R^2 \leq 1$$

Selitysaste mittaa estimoidun regressiomallin *hyvyyttä*:

Mitä suurempi on selitysaste, sitä suurempi on mallineliösumman (eli estimoidun mallin selittämä) osuus selitettävän muuttujan y kokonaisvaihtelua kuvaavasta neliösummasta ja sitä pienempi on jäännöseliösumman (eli estimoidun mallin selittämättä jättämä) osuus selitettävän muuttujan y kokonaisvaihtelua kuvaavasta neliösummasta.

Voidaan osoittaa, että selitysaste yhtyy selitettävän muuttujan havaittujen arvojen ja estimoidun mallin sovitteiden otoskorrelaatiokertoimeen:

$$R^2 = [\text{Cor}(y, \hat{y})]^2$$

Huomaa, että yhden selittäjän lineaarisen regressiomallin tapauksessa pätee lisäksi

$$R^2 = r_{xy}^2$$

jossa

r_{xy} = x -havaintojen ja y -havaintojen *otoskorrelaatiokerroin*

Selitysasteen ominaisuudet

Selitysasteella R^2 on seuraavat ominaisuudet:

- (i) $0 \leq R^2 \leq 1$
- (ii) Seuraavat ehdot ovat *yhtäpitäviä*:
 - (1) $R^2 = 1$
 - (2) Kaikki residuaalit häviävät:
 $e_i = 0$, kaikille $i = 1, 2, \dots, n$
 - (3) Kaikki havaintopisteet (x_i, y_i) , $i = 1, 2, \dots, n$ asettuvat *samalle suoralle*.
 - (4) $r_{xy} = \pm 1$
 - (5) Määritelty malli *selittää täydellisesti* selitettävän muuttujan y havaittujen arvojen vaihtelun.

(iii) Seuraavat ehdot ovat *yhtäpitäviä*:

(1) $R^2 = 0$

(2) $b_1 = 0$

(3) $r_{xy} = 0$

(4) Määritelty malli *ei ollenkaan selitä* selitettävän muuttujan y havaittujen arvojen vaihtelua.

Jäännösvarianssin estimointi

Tavanomaisen yhden selittäjän lineaarisen regressiomallin jäännöstermien $\epsilon_j, j = 1, 2, \dots, n$ varianssin σ^2 *harhaton estimaattori* on

$$s^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

jossa

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, j = 1, 2, \dots, n \text{ on}$$

estimoidun mallin *residuaali*.

Laskutoimitusten järjestäminen

Jos regressiokertoimet joudutaan laskemaan *käsin* tai *laskimella*, yhden selittäjän lineaarisen regressiomallin PNS-estimoinnin vaatimat laskutoimitukset kannattaa järjestää seuraavan taulukon muotoon:

i	x_i	x_i^2	y_i	y_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e_i^2
1	x_1	x_1^2	y_1	y_1^2	$x_1 y_1$	$\hat{y}_1$	e_1	e_1^2
2	x_2	x_2^2	y_2	y_2^2	$x_2 y_2$	$\hat{y}_2$	e_2	e_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	x_n^2	y_n	y_n^2	$x_n y_n$	$\hat{y}_n$	e_n	e_n^2
Summa	$\sum_{i=1}^n x_i$	$\sum_{i=1}^n x_i^2$	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n y_i^2$	$\sum_{i=1}^n x_i y_i$	$\sum_{i=1}^n \hat{y}_i$	$\sum_{i=1}^n e_i$	$\sum_{i=1}^n e_i^2$

Jos tarkoituksena on laskea ainoastaan *PNS-estimaatit* regressiokertoimille β_0 ja β_1 , yllä olevasta taulukosta tarvitaan vain x -havaintojen summa $\sum x_i$ ja neliösumma $\sum x_i^2$, y -havaintojen summa $\sum y_i$ sekä x - ja y -havaintojen tulosumma $\sum x_i y_i$.

Jos tarkoituksena on laskea lisäksi estimoidun mallin *selityaste*, tarvitaan edellä mainittujen suureiden lisäksi myös y -havaintojen neliösumma $\sum y_i^2$ sekä estimoidun mallin residuaalien neliösumma $\sum e_i^2$.

Havaintoarvojen aritmeettiset keskiarvot $\bar{x}$ ja $\bar{y}$, otosvarianssit s_x^2 ja s_y^2 sekä otoskovarianssi s_{xy} saadaan yllä olevan taulukon sarakesummista kaavoilla

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i & \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ s_x^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) & s_y^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right) \\ s_{xy} &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right)\end{aligned}$$

joista *regressiokertoimien estimaatit* saadaan siis lasketuksi kaavoilla

$$\begin{aligned}b_1 &= \frac{s_{xy}}{s_x^2} \\ b_0 &= \bar{y} - b_1 \bar{x}\end{aligned}$$

Estimoidun mallin *sovitteet* saadaan kaavalla

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, 2, \dots, n$$

ja *residuaalit* kaavalla

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, i = 1, 2, \dots, n$$

Estimoidun mallin *selitysaste* voidaan laskea kaavalla

$$R^2 = 1 - \frac{SSE}{SST}$$

jossa

$$SSE = \sum_{i=1}^n e_i^2$$

on estimoidun mallin *jäännöseliösumma* (residuaalien neliösumma) ja

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)s_y^2$$

on selitettävän muuttujan arvojen vaihtelua kuvaava *kokonaisneliösumma*. Huomaa, että yhden selittäjän lineaarisen regressiomallin tapauksessa (koska mallissa on mukana vakio) pätee myös

$$R^2 = r_{xy}^2$$

Regressiokertoimien PNS-estimaattoreiden otosjakaumat

Jos tavanomaiset yhden selittäjän lineaarista regressiomallia koskevat oletukset pätevät, regressiokertoimien β_0 ja β_1 PNS-estimaattorit b_0 ja b_1 ovat *normaalijakautuneita*:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

$$b_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{j=1}^n x_j^2}{n(n-1)s_x^2}\right)$$

Erityisesti regressiokertoimien β_0 ja β_1 PNS-estimaattorit b_0 ja b_1 ovat *harhattomia*:

$$E(b_1) = \beta_1$$

$$E(b_0) = \beta_0$$

Regressiokertoimien luottamusvälit

Oletetaan, että tavanomaiset yhden selittäjän lineaarista regressiomallia koskevat oletukset pätevät.

Regressiokertoimen β_1 eli regressiosuoran *kulmakertoimen luottamusväli* luottamustasolla $(1 - \alpha)$ on muotoa

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{n-1} s_x}$$

jossa $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet Studentin *t-jakaumasta*, jonka vapausasteiden luku on $(n - 2)$ ja s^2 on jäännösvarianssin σ^2 harhaton estimaattori.

Regressiokertoimen β_0 eli regressiosuoran *vakion luottamusväli* luottamustasolla $(1 - \alpha)$ on muotoa

$$b_0 \pm t_{\alpha/2} \frac{s \sqrt{\sum_{j=1}^n x_j^2}}{\sqrt{n(n-1)} s_x}$$

jossa $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet Studentin *t-jakaumasta*, jonka vapausasteiden luku on $(n - 2)$ ja s^2 on jäännösvarianssin σ^2 harhaton estimaattori.

Regressiokertoimia koskevat testit

Oletetaan, että tavanomaiset yhden selittäjän lineaarista regressiomallia koskevat oletukset pätevät.

Olkoon *nollahypoteesina*

$$H_{01} : \beta_1 = \beta_1^0$$

Määritellään *t-testisuure*

$$t_1 = \frac{b_1 - \beta_1^0}{s / (\sqrt{n-1} s_x)}$$

Jos *nollahypoteesi* H_{01} *pätee*,

$$t_1 \sim t(n-2)$$

Itseisarvoltaan *suuret* testisuureen t_1 arvot viittaavat siihen, että *nollahypoteesi* H_{01} *ei päde*.

Olkoon *nollahypoteesina*

$$H_{00} : \beta_0 = \beta_0^0$$

Määritellään *t-testisuure*

$$t_0 = \frac{b_0 - \beta_0^0}{s \sqrt{\sum x_j^2} / (\sqrt{n(n-1)} s_x)}$$

Jos *nollahypoteesi* H_{00} *pätee*,

$$t_0 \sim t(n-2)$$

Itseisarvoltaan *suuret* testisuureen t_0 arvot viittaavat siihen, että *nollahypoteesi* H_{00} *ei päde*.

Merkitsevyystasoa α vastaavan hylkäysalueen (kriittisten arvojen) tai testisuureiden havaittuja arvoja vastaavien p -arvojen määrittäminen tapahtuu täsmälleen samanlaisella tekniikalla kuin normaalijakauman odotusarvoa koskevan tavanomaisen t -testin tapauksessa.

Voidaan osoittaa, että testi *nollahypoteesille*

$$H_{01} : \beta_1 = \beta_1^0$$

voidaan perustaa myös *F-testisuureeseen*

$$F = (n-2) \frac{R^2}{1-R^2}$$

jossa R^2 on estimoidun mallin *selitysaste*. Huomaa, että koska mallissa on mukana vakio, niin

$$R^2 = r_{xy}^2$$

Jos *nollahypoteesi* H_{01} *pätee*,

$$F \sim F(1, n-2)$$

Suuret testisuureen F arvot viittaavat siihen, että *nollahypoteesi* H_{01} *ei päde*.

Tämä F -testi nollahypoteesille H_{01} on ekvivalentti edellä esitetyn t -testin kanssa:

Voidaan osoittaa, että

$$\sqrt{F} = t_1$$

Kaksiulotteisen normaalijakauman regressiofunktioiden estimointi

Kaksiulotteinen normaalijakauma ja sen tiheysfunktio

Oletetaan, että satunnaismuuttujien x ja y pari (x, y) noudattaa **kaksiulotteista normaalijakaumaa** eli

$$(x, y) \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

jossa

$$\mu_x = E(x)$$

$$\mu_y = E(y)$$

$$\sigma_x^2 = \text{Var}(x) = E[(x - \mu_x)^2]$$

$$\sigma_y^2 = \text{Var}(y) = E[(y - \mu_y)^2]$$

$$\rho_{xy} = \text{Cor}(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\sigma_{xy} = \text{Cov}(x, y) = E[(x - \mu_x)(y - \mu_y)]$$

Kaksiulotteisen normaalijakauman *tiheysfunktio* on muotoa

$$f_{xy}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left\{-\frac{1}{2(1-\rho_{xy}^2)} Q(x, y)\right\}$$

jossa

$$Q(x, y) = \left(\frac{x - \mu_x}{\sigma_x}\right)^2 + \left(\frac{y - \mu_y}{\sigma_y}\right)^2 - 2\rho_{xy} \left(\frac{x - \mu_x}{\sigma_x}\right) \left(\frac{y - \mu_y}{\sigma_y}\right)$$

Kaksiulotteisen normaalijakauman ehdolliset jakaumat

Kaksiulotteisen normaalijakauman *ehdolliset jakaumat* ovat normaalisia:

$$(y | x) \sim N(\mu_{y|x}, \sigma_{y|x}^2)$$

$$(x | y) \sim N(\mu_{x|y}, \sigma_{x|y}^2)$$

jossa

$$\mu_{y|x} = E(y | x) = \mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

$$\sigma_{y|x}^2 = \text{Var}(y | x) = (1 - \rho_{xy}^2) \sigma_y^2$$

$$\mu_{x|y} = E(x | y) = \mu_x + \rho_{xy} \frac{\sigma_x}{\sigma_y} (y - \mu_y)$$

$$\sigma_{x|y}^2 = \text{Var}(x | y) = (1 - \rho_{xy}^2) \sigma_x^2$$

Ehdollisten odotusarvojen $E(x|y)$ ja $E(y|x)$ kaavoista nähdään:

- (i) Satunnaismuuttujan x **ehdollinen odotusarvo** satunnaismuuttujan y suhteen eli satunnaismuuttujan x **regressiofunktio** satunnaismuuttujan y suhteen *riippuu lineaarisesti* ehtomuuttujan y arvoista eli on muotoa:

$$E(x|y) = \alpha_0 + \alpha_1 y$$

- (ii) Satunnaismuuttujan y **ehdollinen odotusarvo** satunnaismuuttujan x suhteen eli satunnaismuuttujan y **regressiofunktio** satunnaismuuttujan x suhteen *riippuu lineaarisesti* ehtomuuttujan x arvoista eli on muotoa:

$$E(y|x) = \beta_0 + \beta_1 x$$

Ehdollisten varianssien $\text{Var}(x|y)$ ja $\text{Var}(y|x)$ kaavoista nähdään:

- (i) Satunnaismuuttujan x **ehdollinen varianssi** satunnaismuuttujan y suhteen *ei riipu* ehtomuuttujan y arvoista.
- (ii) Satunnaismuuttujan y **ehdollinen varianssi** satunnaismuuttujan x suhteen *ei riipu* ehtomuuttujan x arvoista.

Ehdollisten odotusarvojen kaavoista nähdään edelleen, että sekä satunnaismuuttujan y *regressiofunktio* satunnaismuuttujan x suhteen että satunnaismuuttujan y *regressiofunktio* satunnaismuuttujan x suhteen kulkevat satunnaismuuttujien x ja y todennäköisyysjakauman todennäköisyysmassan *painopisteen*

$$(\mu_x, \mu_y)$$

kautta.

Otos kaksikulotteisesta normaalijakaumasta

Oletetaan, että satunnaismuuttujien x ja y pari (x, y) noudattaa *kaksikulotteista normaalijakaumaa* eli

$$(x, y) \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

Olkoot

$$y_1, y_2, \dots, y_n$$

muuttujan y havaitut arvot ja

$$x_1, x_2, \dots, x_n$$

muuttujan x havaitut arvot ja oletetaan, että havaintoarvojen x_i ja y_i parit

$$(x_i, y_i), i = 1, 2, \dots, n$$

muodostavat *yksinkertaisen satunnaisotoksen kaksikulotteista normaalijakaumasta*

$$N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

Tällöin

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \perp$$

$$(x_i, y_i) \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}), i = 1, 2, \dots, n$$

Kaksiulotteisen normaalijakauman regressiofunktioiden PNS-estimointi

Oletetaan, että havaintoarvojen x_i ja y_i parit (x_i, y_i) , $i = 1, 2, \dots, n$ muodostavat yksinkertaisen satunnaisotoksen *kaksiulotteista normaalijakaumasta* $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$.

Kaksiulotteisen normaalijakauman regressiofunktiot ovat muotoa

$$E(y | x) = \beta_0 + \beta_1 x$$

$$E(x | y) = \alpha_0 + \alpha_1 y$$

Estimoidaan regressiofunktiot pienimmän neliösumman menetelmällä.

Määritellään yhden selittäjän lineaariset regressiomallit

$$(1) \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

$$(2) \quad x_i = \alpha_0 + \alpha_1 y_i + \delta_i, i = 1, 2, \dots, n$$

Muuttujan y PNS-suoran yhtälö muuttujan x suhteen on

$$(3) \quad y = b_0 + b_1 x$$

jossa

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = r_{xy} \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$

Muuttujan x PNS-suoran yhtälö muuttujan y suhteen on

$$(4) \quad x = a_0 + a_1 y$$

jossa

$$a_0 = \bar{x} - a_1 \bar{y} \qquad a_1 = r_{xy} \frac{s_x}{s_y} = \frac{s_{xy}}{s_y^2}$$

Mallien (1) ja (2) regressiokertoimien $\beta_0, \beta_1, \alpha_0, \alpha_1$ PNS-estimaattoreiden b_0, b_1, a_0, a_1 lausekkeissa

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i & \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 & s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ r_{xy} &= \frac{s_{xy}}{s_x s_y} \end{aligned}$$

Muuttujan y PNS-suoran yhtälö muuttujan x suhteen voidaan kirjoittaa muotoon

$$(3)' \quad y - \bar{y} = r_{xy} \frac{s_y}{s_x} (x - \bar{x})$$

Muuttujan x PNS-suoran yhtälö muuttujan y suhteen voidaan kirjoittaa muotoon

$$(4)' \quad x - \bar{x} = r_{xy} \frac{s_x}{s_y} (y - \bar{y})$$

Yhtälöistä (3)' ja (4)' nähdään välittömästi, että molemmat PNS-suorat kulkevat havaintoaineiston *painopisteen*

$$(\bar{x}, \bar{y})$$

kautta.

Yhtälöistä (3)' ja (4)' nähdään edelleen, että PNS-suorien kulmakertoimien tulo on muuttujien y ja x *korrelaatiokertoimen neliö*:

$$a_1 b_1 = \left(r_{xy} \frac{s_x}{s_y} \right) \left(r_{xy} \frac{s_y}{s_x} \right) = r_{xy}^2$$

Voidaan osoittaa, että molempiin PNS-suoriin liittyy sama *selitysaste* R^2 ja se yhtyy muuttujien y ja x havaittujen arvojen korrelaatiokertoimen neliöön:

$$R^2 = r_{xy}^2$$

PNS-suoraan (3) liittyvä *jäännösvarianssin* (harhaton) estimaattori on

$$s_{(3)}^2 = \frac{SSE_y}{n-2}$$

jossa

$$SSE_y = \text{PNS-suoraan (3) liittyvä jäännösneliösumma}$$

Voidaan osoittaa, että

$$SSE_y = (1 - r_{xy}^2) SST_y$$

jossa

$$SST_y = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1) s_y^2$$

PNS-suoraan (4) liittyvä *jäännösvarianssin* (harhaton) estimaattori on

$$s_{(4)}^2 = \frac{SSE_x}{n-2}$$

jossa

$$SSE_x = \text{PNS-suoraan (4) liittyvä jäännösneliösumma}$$

Edelleen voidaan osoittaa, että

$$SSE_x = (1 - r_{xy}^2)SST_x$$

jossa

$$SST_x = \sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s_x^2$$

Kaksiulotteisen normaalijakauman regressiofunktioiden estimointi momenttimenetelmällä ja suurimman uskottavuuden menetelmällä

Vertaamalla edellä esitettyjä kaksiulotteisen normaalijakauman regressiofunktioiden PNS-estimaattoreiden kaavoja kaksiulotteisen normaalijakauman regressiofunktioiden lausekkeisiin nähdään välittömästi, että regressiofunktioiden PNS-estimaattorit yhtyvät niiden **momentti-estimaattoreihin**.

Edelleen voidaan osoittaa, että regressiofunktioiden PNS-estimaattorit yhtyvät myös niiden **suurimman uskottavuuden estimaattoreihin**.

Tehtävä 12.1.

Muuttujien x ja y havaitut arvot ovat:

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

- (a) Määrä tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$$

regressiokertoimien β_0 ja β_1 pienimmän neliösumman (PNS-) estimaatit.

- (b) Määrä estimoidun mallin sovitteet ja residuaalit.
 (c) Määrä estimoidun mallin jäännösvarianssin σ^2 harhaton estimaatti.
 (d) Määrä estimoidun mallin selitysaste.

Tehtävä 12.1. – Mitä opimme?

Tehtävässä tarkastellaan *yhden selittäjän lineaarisen regressiomallin estimointia*.

Tehtävä 12.1. – Ratkaisu:

Kaikki tehtävän laskutoimitukset on tehty Microsoft Excel -ohjelmalla;
 ks. Excel-taulukkoa ratkaisun lopussa.

- (a) Yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

regressiokertoimien α ja β PNS-estimaatit saadaan lasketuksi seuraavassa esitettävällä tavalla.

Määritetään ensin muuttujien x ja y havaittujen arvojen *summat*, *neliösummat* ja *tulosumma*:

$$\begin{aligned} \sum_{i=1}^n x_i &= 56 & \sum_{i=1}^n y_i &= 40 \\ \sum_{i=1}^n x_i^2 &= 524 & \sum_{i=1}^n y_i^2 &= 256 \\ \sum_{i=1}^n x_i y_i &= 364 \end{aligned}$$

Muuttujien x ja y havaittujen arvojen aritmeettiset keskiarvot $\bar{x}$ ja $\bar{y}$, otosvarianssit s_x^2 ja s_y^2 , otoskeskihajonnat s_x ja s_y , otoskovarianssi s_{xy} ja otoskorrelaatio r_{xy} saadaan muuttujien x ja y havaittujen arvojen *summista*, *neliösummista* ja *tulosummasta*:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{8} \times 56 = 7$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{8} \times 40 = 5$$

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) = \frac{1}{8-1} \left(524 - \frac{1}{8} \times 56^2 \right) = 18.857$$

$$s_y^2 = \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right) = \frac{1}{8-1} \left(256 - \frac{1}{8} \times 40^2 \right) = 8$$

$$s_x = \sqrt{s_x^2} = \sqrt{18.857} = 4.342$$

$$s_y = \sqrt{s_y^2} = \sqrt{8} = 2.828$$

$$s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right) = \frac{1}{8-1} \left(364 - \frac{1}{8} \times 56 \times 40 \right) = 12$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{12}{4.342 \times 2.828} = 0.977$$

Estimoidun PNS-suoran yhtälö on muotoa

$$y = b_0 + b_1 x$$

jossa b_0 ja b_1 ovat mallin regressiokertoimien β_0 ja β_1 PNS-estimaattorit.

Estimaattoreiden b_0 ja b_1 arvot saadaan yllä määrätyistä otostunnusluvuista:

$$b_1 = r_{xy} \frac{s_y}{s_x} = 0.977 \times \frac{2.828}{4.342} = 0.636$$

$$b_0 = \bar{y} - b_1 \bar{x} = 5 - 0.636 \times 7 = 0.545$$

Estimoidun PNS-suoran yhtälöksi saadaan siten

$$y = 0.545 + 0.636x$$

- (b) Estimoidun mallin sovitteet $\hat{y}_i$ ja residuaalit e_i saadaan seuraavilla kaavoilla:

Sovitteet:

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, 2, \dots, n$$

Residuaalit:

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$$

Sovitteet ja residuaalit on annettu alla olevassa Excel-taulukossa.

- (c) Mallin jäännösvarianssin σ^2 harhattoman estimaattorin s^2 arvoksi saadaan

$$s^2 = \frac{1}{n-2} SSE = \frac{1}{8-2} \times 2.545 = 0.424$$

jossa

$$SSE = \sum_{i=1}^n e_i^2 = 2.545$$

on estimoidun mallin residuaalien vaihtelua kuvaava jäännösneliösumma.

- (d) Estimoidun mallin selitysaste R^2 voidaan laskea usealla eri tavalla.

Olkoot estimoidun mallin sovitteet

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, 2, \dots, n$$

ja residuaalit

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$$

Selittävän muuttujan y havaittujen arvojen vaihtelua kuvaava kokonaisneliösumma on

$$SST = (n-1)s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = 256 - \frac{1}{8} \times 40^2 = 56$$

Estimoidun mallin residuaalien vaihtelua kuvaava jäännösneliösumma on

$$SSE = \sum_{i=1}^n e_i^2 = 2.545$$

Estimoidun mallin selittämää osuutta selittävän muuttujan y havaittujen arvojen vaihtelusta kuvaava mallineliösumma on

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SST - SSE = 56 - 2.545 = 53.455$$

Selitysaste R^2 on (ks. alla olevaa Excel-tilukkoa)

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST} = 1 - \frac{2.545}{56} = \frac{53.455}{56} = 0.955$$

Yhden selittäjän lineaarisen regressiomallin tapauksessa (koska mallissa oli mukana vakio) pätee myös

$$R^2 = r_{xy}^2 = 0.977^2 = 0.955$$

Kaikki tehtävän laskutoimitukset on tehty Microsoft Excel –ohjelmalla; ks. alla olevaa taulukkoa.

<i>i</i>	<i>x</i>	<i>y</i>	<i>x</i> ²	<i>y</i> ²	<i>xy</i>	<i>yhat</i>	<i>res</i>	<i>res</i> ²
1	1	1	1	1	1	1.182	-0.182	0.033
2	3	2	9	4	6	2.455	-0.455	0.207
3	4	4	16	16	16	3.091	0.909	0.826
4	6	4	36	16	24	4.364	-0.364	0.132
5	8	5	64	25	40	5.636	-0.636	0.405
6	9	7	81	49	63	6.273	0.727	0.529
7	11	8	121	64	88	7.545	0.455	0.207
8	14	9	196	81	126	9.455	-0.455	0.207
Summa	56	40	524	256	364	40	0.000	2.545

$$\text{Mean}(x) = 7$$

$$s_x^2 = 18.857$$

$$s_x = 4.342$$

$$\text{Mean}(y) = 5$$

$$s_y^2 = 8$$

$$s_y = 2.828$$

$$s_{xy} = 12$$

$$r_{xy} = 0.977$$

$$b_1 = 0.636$$

$$b_0 = 0.545$$

$$SST = 56$$

$$SSE = 2.545$$

$$SSM = 53.455$$

$$R^2 = 0.955 \quad := 1 - SSE/SST$$

$$R^2 = 0.955 \quad := SSM/SST$$

$$R^2 = 0.955 \quad := r_{xy}^2$$

$$s^2 = 0.424$$

$$s = 0.651$$

$$t(b_1) = 11.225$$

$$(1/2)*I\text{-välin pituus} = 0.139$$

$$(I\text{-tasolla } 0.95)$$

Tehtävä 12.2.

Jatkoa tehtävälle 12.1.

Piirrä tehtävässä 12.1. estimoitu regressiosuora havaintoja

$$(x_i, y_i), i = 1, 2, \dots, n$$

esittävään pistediagrammiin. Merkitse kuvioon sovitteita vastaavat pisteet

$$(x_i, \hat{y}_i), i = 1, 2, \dots, n$$

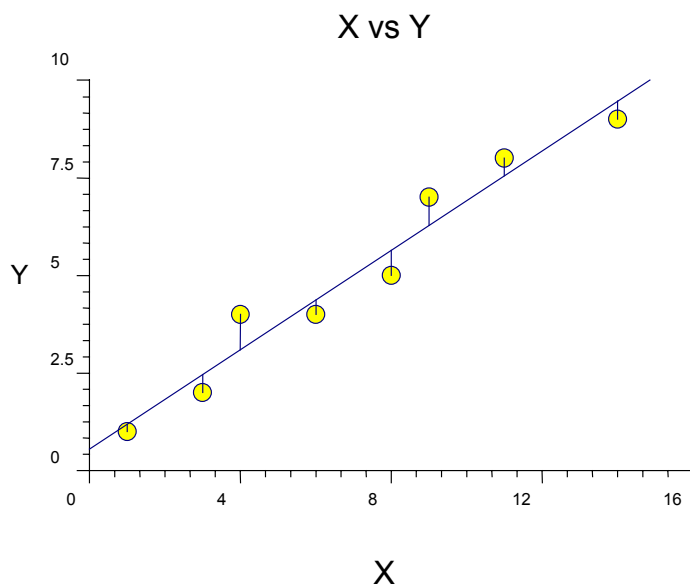
Piirrä samaan kuvioon myös residuaaleja kuvaavat janat.

Tehtävä 12.2. – Mitä opimme?

Tehtävässä havainnollistetaan *estimoidun PNS-suoran piirtämistä havaintoaineistoa kuvaavaan pistediagrammiin sekä estimoidun mallin residuaaleja.*

Tehtävä 12.2. – Ratkaisu:

Ao. kuvio on tuotettu Statistix-ohjelmalla:



Tehtävä 12.3.

Jatkoa tehtävälle 12.1.

- (a) Testaa tehtävän 12.1. regressiomallin kerrointa β_1 koskevaa nollahypoteesia

$$H_0 : \beta_1 = 0$$

Käytä kaksisuuntaista vaihtoehtoista hypoteesia ja 5 %:n merkitsevyystasoa.

- (b) Muodosta kertoimelle β_1 95%:n luottamusväli.

Tehtävä 12.3. – Mitä opimme?

Tehtävässä tarkastellaan *tilastollista päättelyä* yhden selittäjän lineaarisessa regressiomallissa.

Tehtävä 12.3. – Ratkaisu:

Kaikki tehtävän laskutoimitukset on tehty Microsoft Excel -ohjelmalla; ks. Excel-taulukkoa tehtävän 12.1. ratkaisun lopussa.

- (a) *t*-testisuure nollahypoteesille

$$H_{01} : \beta_1 = 0$$

on muotoa

$$t_1 = \frac{b_1}{s / \sqrt{n-1} s_x}$$

jossa b_1 on regressiokertoimen β_1 PNS-estimaattori, s^2 on mallin jäännösvarianssin σ^2 harhaton estimaattori ja s_x^2 on muuttujan x havaittujen arvojen otosvarianssi.

Jos nollahypoteesi H_{01} pätee, testisuure t_1 on jakautunut Studentin *t*-jakauman mukaan vapausastein $(n - 2)$:

$$t_1 \sim t(n - 2)$$

Tehtävän tapauksessa:

$$t_1 = \frac{b_1}{s / \sqrt{n-1} s_x} = \frac{0.636}{0.651 / \sqrt{8-1} \times 4.342} = 11.225$$

ja testisuureen jakauman vapausasteet ovat

$$df = n - 2 = 6$$

5 %:n merkitsevyystasoa vastaaviksi kriittisiksi arvoiksi $-t_{0.025}$ ja $+t_{0.025}$ saadaan Studentin *t*-jakauman taulukoista 2-suuntaisen vaihtoehtoisen hypoteesin tapauksessa ($df = n - 2 = 6$):

$$-t_{0.025} = -2.447$$

$$+t_{0.025} = +2.447$$

Koska

$$t_1 = 11.225 > +2.447$$

niin nollahypoteesi H_{01} hylätään.

- (b) Regressiokertoimen β_1 luottamusväli luottamustasolla $(1 - \alpha)$ on muotoa

$$b \pm t_{\alpha/2} \frac{s}{\sqrt{n-1} s_x}$$

jossa b_1 on regressiokertoimen β_1 PNS-estimaattori, s^2 on mallin jäännösvarianssin σ^2 harhaton estimaattori, s_x^2 on muuttujan x havaittujen arvojen otosvarianssi sekä $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet Studentin t -jakaumasta, jonka vapausasteiden luku on $(n - 2)$.

Luottamustasoa 0.95 vastaaviksi luottamuskertoimiksi $-t_{0.025}$ ja $+t_{0.025}$ saadaan Studentin t -jakauman taulukoista ($df = n - 2 = 6$):

$$-t_{0.025} = -2.447$$

$$+t_{0.025} = +2.447$$

Siten luottamusväliksi saadaan

$$\begin{aligned} b \pm t_{\alpha/2} \frac{s}{\sqrt{n-1} s_x} &= 0.636 \pm 2.447 \times \frac{0.651}{\sqrt{8-1} \times 4.342} \\ &= 0.636 \pm 0.139 \\ &= (0.497, 0.775) \end{aligned}$$

Huomautus:

Luottamustasoon 95 % liittyvät luottamuskertoimet $-t_{0.025}$ ja $+t_{0.025}$ ovat samat kuin (a)-kohdan kaksisuuntaisen testin kriittiset rajat.

Tehtävä 12.4.

Menestyminen opinnoissa saattaa vaikuttaa vastavalmistuneen alkupalkkaan.

Asiaa tutkittiin eräässä USA:n yliopistossa poimimalla vastavalmistuneiden joukosta yksinkertainen satunnaisotos, jonka koko oli 15. Otokseen poimituilta opiskelijoilta kysyttiin heidän arvosanapisteidensä keskiarvoa (muuttuja x) ja alkupalkkaa (muuttuja y ; yksikkönä 1000 \$).

Otosta kuvaavat perustunnusluvut olivat:

$$\bar{x} = 3.04 \qquad \bar{y} = 18.05$$

$$s_x^2 = 0.063 \qquad s_y^2 = 5.81$$

$$r_{xy} = 0.848$$

- (a) Määrittää regressiokertoimien estimaatit lineaarisesta regressiomallista

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

jossa alkupalkkaa y selitetään arvosanapisteiden keskiarvolla x .

- (b) Määrittää regressiokertoimien estimaatit lineaarisesta regressiomallista

$$x_i = \alpha_0 + \alpha_1 y_i + \delta_i, i = 1, 2, \dots, n$$

jossa arvosanapisteiden keskiarvoa x selitetään alkupalkalla y (ns. käänteisregressio).

- (c) Määrittää estimoitujen regressiomallien selitysasteet.
(d) Määrittää kohdissa (a) ja (b) estimoitujen regressiosuorien leikkauspiste.

Vertaa tulosta x - ja y -havaintoarvojen aritmeettisiin keskiarvoihin. Onko tulos sattuma?

Tehtävä 12.4. – Mitä opimme?

Tehtävässä tarkastellaan *kaksiulotteisen normaalijakauman regressiofunktioiden estimointia*.

Tehtävä 12.4. – Ratkaisu:

Otosta kuvaavat perustunnusluvut olivat:

$$\bar{x} = 3.04 \qquad \bar{y} = 18.05$$

$$s_x^2 = 0.063 \qquad s_y^2 = 5.81$$

$$r_{xy} = 0.848$$

(a) Mallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

regressiokertoimien β_0 ja β_1 PNS-estimaateiksi saadaan

$$b_1 = r_{xy} \frac{s_y}{s_x} = 0.848 \times \sqrt{\frac{5.81}{0.063}} = 8.14$$

$$b_0 = \bar{y} - b_1 \bar{x} = 18.05 - 8.14 \times 3.04 = -6.70$$

(b) Mallin

$$x_i = \alpha_0 + \alpha_1 y_i + \delta_i, i = 1, 2, \dots, n$$

regressiokertoimien α_0 ja α_1 PNS-estimaateiksi saadaan

$$a_1 = r_{xy} \frac{s_x}{s_y} = 0.848 \times \sqrt{\frac{0.063}{5.81}} = 0.0883$$

$$a_0 = \bar{x} - a_1 \bar{y} = 3.04 - 0.0883 \times 18.05 = 1.45$$

(c) Koska kohtien (a) ja (b) regressiomallit ovat yhden selittäjän lineaarisia regressiomalleja, niin molemmille regressiomalleille pätee:

$$R^2 = r_{XY}^2 = 0.848^2 = 0.719$$

Huomaa, että

$$r_{XY}^2 = b_1 \times a_1 = 8.14 \times 0.0883 = 0.719$$

(e) Ko. regressiosuorat leikkaavat aina (aritmeettisten keskiarvojen määräämässä) havaintoarvojen painopisteessä, jos suorissa on mukana vakiotermi.

Siten suorien leikkauspiste on

$$(\bar{x}, \bar{y}) = (3.04, 18.05)$$

Tehtävä 12.5.

Muuttujien x ja y havaitut arvot ovat:

x	2	3	4	5	8	9	11
y	10	9	7	4	3	2	0

- (a) Määrä tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$$

regressiokertoimien β_0 ja β_1 pienimmän neliösumman (PNS-) estimaatit.

- (b) Määrä estimoidun mallin sovitteet ja residuaalit.
 (c) Määrä estimoidun mallin jäännösvarianssin σ^2 harhaton estimaatti.
 (d) Määrä estimoidun mallin selitysaste.

Tehtävä 12.5. – Mitä opimme?

Tehtävässä tarkastellaan *yhden selittäjän lineaarisen regressiomallin estimointia*.

Tehtävä 12.5. – Ratkaisu:

Kaikki tehtävän laskutoimitukset on tehty Microsoft Excel -ohjelmalla;
 ks. Excel-taulukkoa ratkaisun lopussa.

- (a) Yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

regressiokertoimien α ja β PNS-estimaatit saadaan lasketuksi seuraavassa esitettävällä tavalla.

Määritään ensin muuttujien x ja y havaittujen arvojen *summat*, *neliösummat* ja *tulosumma*:

$$\begin{aligned} \sum_{i=1}^n x_i &= 42 & \sum_{i=1}^n y_i &= 35 \\ \sum_{i=1}^n x_i^2 &= 320 & \sum_{i=1}^n y_i^2 &= 259 \\ \sum_{i=1}^n x_i y_i &= 137 \end{aligned}$$

Muuttujien x ja y havaittujen arvojen aritmeettiset keskiarvot $\bar{x}$ ja $\bar{y}$, otosvarianssit s_x^2 ja s_y^2 , otoskeskihajonnat s_x ja s_y , otoskovarianssi s_{xy} ja otoskorrelaatio r_{xy} saadaan muuttujien x ja y havaittujen arvojen *summista*, *neliösummista* ja *tulosummasta*:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{7} \times 42 = 6$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{7} \times 35 = 5$$

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) = \frac{1}{7-1} \left(320 - \frac{1}{7} \times 42^2 \right) = 11.333$$

$$s_y^2 = \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right) = \frac{1}{7-1} \left(259 - \frac{1}{7} \times 35^2 \right) = 14$$

$$s_x = \sqrt{s_x^2} = \sqrt{11.333} = 3.367$$

$$s_y = \sqrt{s_y^2} = \sqrt{14} = 3.742$$

$$s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right) = \frac{1}{7-1} \left(137 - \frac{1}{7} \times 42 \times 35 \right) = -12.167$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-12.167}{3.367 \times 3.742} = -0.966$$

Estimoidun PNS-suoran yhtälö on muotoa

$$y = b_0 + b_1 x$$

jossa b_0 ja b_1 ovat mallin regressiokertoimien β_0 ja β_1 PNS-estimaattorit.

Estimaattoreiden b_0 ja b_1 arvot saadaan yllä määrätyistä otostunnusluvuista:

$$b_1 = r_{xy} \frac{s_y}{s_x} = -0.966 \times \frac{3.742}{3.367} = -1.074$$

$$b_0 = \bar{y} - b_1 \bar{x} = 5 + 1.074 \times 6 = 11.441$$

Estimoidun PNS-suoran yhtälöksi saadaan siten

$$y = 11.441 - 1.074x$$

- (b) Estimoidun mallin sovitteet $\hat{y}_i$ ja residuaalit e_i saadaan seuraavilla kaavoilla:

Sovitteet:

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, 2, \dots, n$$

Residuaalit:

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$$

Sovitteet ja residuaalit on annettu alla olevassa Excel-taulukossa.

- (c) Mallin jäännösvarianssin σ^2 harhattoman estimaattorin s^2 arvoksi saadaan

$$s^2 = \frac{1}{n-2} SSE = \frac{1}{7-2} \times 5.632 = 1.126$$

jossa

$$SSE = \sum_{i=1}^n e_i^2 = 5.632$$

on estimoidun mallin residuaalien vaihtelua kuvaava jäännösneliösumma.

- (d) Estimoidun mallin selitysaste R^2 voidaan laskea usealla eri tavalla.

Olkoot estimoidun mallin sovitteet

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, 2, \dots, n$$

ja residuaalit

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$$

Selitettävän muuttujan y havaittujen arvojen vaihtelua kuvaava kokonaisneliösumma on

$$SST = (n-1)s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = 259 - \frac{1}{7} \times 35^2 = 84$$

Estimoidun mallin residuaalien vaihtelua kuvaava jäännösneliösumma on

$$SSE = \sum_{i=1}^n e_i^2 = 5.632$$

Estimoidun mallin selittämää osuutta selitettävän muuttujan y havaittujen arvojen vaihtelusta kuvaava mallineliösumma on

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SST - SSE = 84 - 5.632 = 78.368$$

Selitysaste R^2 on (ks. alla olevaa Excel-taulukkoa)

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST} = 1 - \frac{5.632}{84} = \frac{78.368}{84} = 0.933$$

Yhden selittäjän lineaarisen regressiomallin tapauksessa (koska mallissa oli mukana vakio) pätee myös

$$R^2 = r_{xy}^2 = (-0.965)^2 = 0.933$$

Kaikki tehtävän laskutoimitukset on tehty Microsoft Excel –ohjelmalla; ks. alla olevaa taulukkoa.

<i>i</i>	<i>x</i>	<i>y</i>	<i>x</i> ²	<i>y</i> ²	<i>xy</i>	<i>yhat</i>	<i>res</i>	<i>res</i> ²
1	2	10	4	100	20	9.294118	0.705882	0.49827
2	3	9	9	81	27	8.220588	0.779412	0.607483
3	4	7	16	49	28	7.147059	-0.14706	0.021626
4	5	4	25	16	20	6.073529	-2.07353	4.299524
5	8	3	64	9	24	2.852941	0.147059	0.021626
6	9	2	81	4	18	1.779412	0.220588	0.048659
7	11	0	121	0	0	-0.36765	0.367647	0.135164
Summa	42	35	320	259	137	35	8.88E-16	5.632353

Laskutoimitukset Excelin taulukkolaskentaominaisuuksien avulla:

$$\text{Mean}(x) = 6 \quad s_x^2 = 11.33333 \quad s_x = 3.366502 \quad s_{xy} = -12.1667$$

$$\text{Mean}(y) = 5 \quad s_y^2 = 14 \quad s_y = 3.741657$$

$$r_{xy} = -0.96589$$

$$b_1 = -1.07353 \quad \text{Std Error}(b_1) = 0.128708 \quad t(b_1) = -8.34081$$

$$b_0 = 11.44118 \quad p\text{-value} = 0.000405$$

$$s^2 = 1.126471 \quad s = 1.061353$$

$$SST = 84$$

$$SSE = 5.632353$$

$$SSM = 78.36765$$

$$R^2 = 0.932948 \quad r_{xy}^2 = 0.932948$$

$$1 - \alpha = 0.99$$

$$\alpha/2 = 0.005$$

$$df = 5$$

$$t(\alpha/2) = 4.032143$$

$$lb(b_1) = -1.5925$$

$$ub(b_1) = -0.55456$$

Laskutoimitukset Excelin Data Analysis Tools -työkalujen avulla:**Descriptice Statistics -työkalu:**

<i>x</i>		<i>y</i>	
Mean	6	Mean	5
Std Error	1.272418	Std Error	1.414214
Std Dev	3.366502	Std Dev	3.741657
Var	11.33333	Var	14

Correlation -työkalu:

	<i>x</i>	<i>y</i>
<i>x</i>	1	
<i>y</i>	-0.96589	1

Regression -työkalu:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.965892
R Square	0.932948
Std Error	1.061353
Count	7

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>
Regression	1	78.36765	78.36765	69.56919	0.000405
Residual	5	5.632353	1.126471		
Total	6	84			

	<i>Coeffs</i>	<i>Std Error</i>	<i>t Stat</i>	<i>p-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 99%</i>	<i>Upper 99%</i>
Intercept	11.44118	0.870225	13.14738	4.55E-05	9.204192	13.67816	7.932306	14.95005
<i>x</i>	-1.07353	0.128708	-8.34081	0.000405	-1.40438	-0.74268	-1.5925	-0.55456

RESIDUAL OUTPUT

<i>Obs</i>	<i>Pred y</i>	<i>Res</i>
1	9.294118	0.705882
2	8.220588	0.779412
3	7.147059	-0.14706
4	6.073529	-2.07353
5	2.852941	0.147059
6	1.779412	0.220588
7	-0.36765	0.367647

Tehtävä 12.6.

Jatkoa tehtävälle 12.5.

Piirrä tehtävässä 12.5. estimoitu regressiosuora havaintoja

$$(x_i, y_i), i = 1, 2, \dots, n$$

esittävään pistediagrammiin. Merkitse kuvioon sovitteita vastaavat pisteet

$$(x_i, \hat{y}_i), i = 1, 2, \dots, n$$

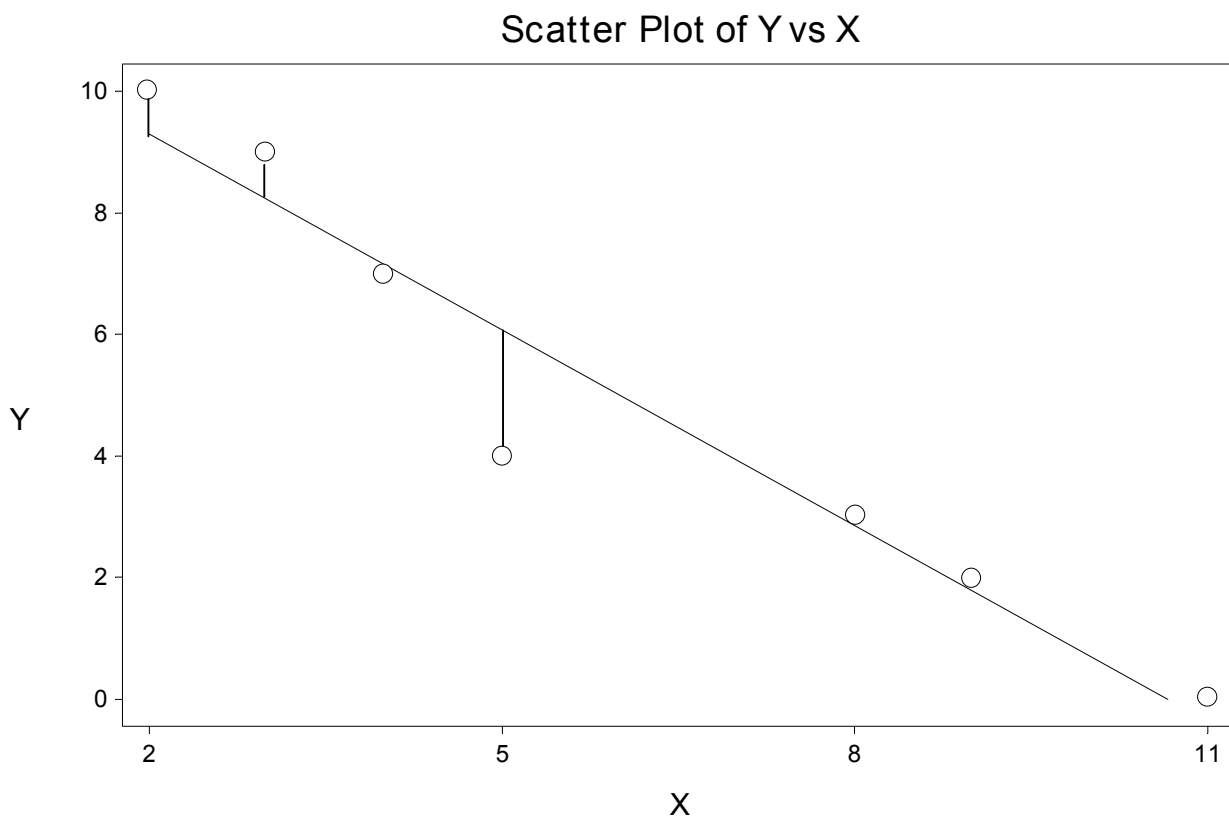
Piirrä samaan kuvioon myös residuaaleja kuvaavat janat.

Tehtävä 12.6. – Mitä opimme?

Tehtävässä havainnollistetaan *estimoidun PNS-suoran piirtämistä havaintoaineistoa kuvaavaan pistediagrammiin sekä estimoidun mallin residuaaleja.*

Tehtävä 12.6. – Ratkaisu:

Ao. kuvio on tuotettu Statistix-ohjelmalla ja siihen on lisätty Wordin piirtotyökaluja käyttäen residuaaleja kuvaavista janoista ne kolme, jotka on mahdollista erottaa kuviosta (neljä havaintopisteistä on niin lähellä regressiosuoraa, että niitä vastaavat residuaalit ovat liian pieniä erottuakseen kuviosta):



Tehtävä 12.7.

Jatkoa tehtävälle 12.5.

- (a) Testaa tehtävän 12.5. regressiomallin kerrointa β_1 koskevaa nollahypoteesia

$$H_0 : \beta_1 = 0$$

Käytä kaksisuuntaista vaihtoehtoista hypoteesia ja 1 %:n merkitsevyystasoa.

- (b) Muodosta kertoimelle β_1 99%:n luottamusväli.

Tehtävä 12.7. – Mitä opimme?

Tehtävässä tarkastellaan *tilastollista päättelyä* yhden selittäjän lineaarisessa regressiomallissa.

Tehtävä 12.7. – Ratkaisu:

Kaikki tehtävän laskutoimitukset on tehty Microsoft Excel -ohjelmalla; ks. Excel-taulukkoa tehtävän 12.5. ratkaisun lopussa.

- (a) *t*-testisuure nollahypoteesille

$$H_{01} : \beta_1 = 0$$

on muotoa

$$t_1 = \frac{b_1}{s / \sqrt{n-1} s_x}$$

jossa b_1 on regressiokertoimen β_1 PNS-estimaattori, s^2 on mallin jäännösvarianssin σ^2 harhaton estimaattori ja s_x^2 on muuttujan x havaittujen arvojen otosvarianssi.

Jos nollahypoteesi H_{01} pätee, testisuure t_1 on jakautunut Studentin *t*-jakauman mukaan vapausastein $(n - 2)$:

$$t_1 \sim t(n - 2)$$

Tehtävän tapauksessa:

$$t_1 = \frac{b_1}{s / \sqrt{n-1} s_x} = \frac{-1.074}{1.061 / \sqrt{7-1} \times 3.367} = -8.341$$

ja testisuureen jakauman vapausasteet ovat

$$df = n - 2 = 5$$

1 %:n merkitsevyystasoa vastaaviksi kriittisiksi arvoiksi $-t_{0.005}$ ja $+t_{0.005}$ saadaan Studentin *t*-jakauman taulukoista 2-suuntaisen vaihtoehtoisen hypoteesin tapauksessa ($df = n - 2 = 5$):

$$-t_{0.005} = -4.032$$

$$+t_{0.025} = +4.032$$

Koska

$$t_1 = -8.341 < -4.032$$

niin nollahypoteesi H_{01} hylätään.

- (b) Regressiokertoimen β_1 luottamusväli luottamustasolla $(1 - \alpha)$ on muotoa

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{n-1} s_x}$$

jossa b_1 on regressiokertoimen β_1 PNS-estimaattori, s^2 on mallin jäännösvarianssin σ^2 harhaton estimaattori, s_x^2 on muuttujan x havaittujen arvojen otosvarianssi sekä $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet Studentin t -jakaumasta, jonka vapausasteiden luku on $(n - 2)$.

Luottamustasoa 0.99 vastaaviksi luottamuskertoimiksi $-t_{0.005}$ ja $+t_{0.005}$ saadaan Studentin t -jakauman taulukoista ($df = n - 2 = 5$):

$$-t_{0.005} = -4.032$$

$$+t_{0.025} = +4.032$$

Siten luottamusväliksi saadaan

$$\begin{aligned} b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{n-1} s_x} &= -1.074 \pm 4.032 \times \frac{1.061}{\sqrt{7-1} \times 3.367} \\ &= -1.074 \pm 0.519 \\ &= (-1.593, -0.555) \end{aligned}$$

Huomautus:

Luottamustasoon 99 % liittyvät luottamuskertoimet $-t_{0.005}$ ja $+t_{0.005}$ ovat samat kuin (a)-kohdan kaksisuuntaisen testin kriittiset rajat.

Tehtävä 12.8.

Olkoon

$$25x = -9y + 59$$

muuttujan x (estimoidun) regressiosuoran yhtälö muuttujan y suhteen

$$4y = -9x + 22$$

ja muuttujan y (estimoidun) regressiosuoran yhtälö muuttujan x suhteen.

- (a) Määrää muuttujien x ja y aritmeettiset keskiarvot.
- (b) Määrää muuttujien x ja y välinen korrelaatio.
- (c) Määrää molempien mallien selitysasteet.

Tehtävä 12.8. – Mitä opimme?

Tehtävässä nähdään, että *muuttujan x regressiosuoran yhtälö muuttujan y suhteen ja muuttujan y regressiosuoran yhtälö muuttujan x suhteen* yhdessä sisältävät tiedon sekä muuttujien x ja y aritmeettisistä keskiarvoista että niiden välisestä korrelaatiosta ja siten myös molempien mallien selitysasteista.

Tehtävä 12.8. – Ratkaisu:

Muuttujan y PNS-suoran yhtälö muuttujan x suhteen on

$$(1) \quad y - \bar{y} = r_{xy} \frac{s_y}{s_x} (x - \bar{x})$$

ja muuttujan x PNS-suoran yhtälö muuttujan y suhteen on

$$(2) \quad x - \bar{x} = r_{xy} \frac{s_x}{s_y} (y - \bar{y})$$

Siten PNS-suorat leikkaavat aina havaintoaineiston painopisteessä

$$(\bar{x}, \bar{y})$$

Edelleen yhtälöistä (1) ja (2) nähdään, että muuttujien x ja y havaittujen arvojen välisen korrelaatiokertoimen neliö saadaan suorien kulmakertoimien tulona:

$$\left(r_{xy} \frac{s_y}{s_x} \right) \times \left(r_{xy} \frac{s_x}{s_y} \right) = r_{xy}^2$$

Lisäksi voidaan osoittaa, että molemmilla regressiomalleilla on sama selitysaste, joka yhtyy muuttujien x ja y korrelaatiokertoimen neliöön:

$$R^2 = r_{xy}^2$$

(a) Suorien

$$9y = -25x + 59$$

ja

$$4y = -9x + 22$$

leikkauspiste on

$$(\bar{x}, \bar{y}) = (2, 1)$$

joten muuttujien x ja y havaittujen arvojen *aritmeettiset keskiarvot* ovat

$$\bar{x} = 2$$

$$\bar{y} = 1$$

(b) Suorien

$$x = -\frac{9}{25}y + \frac{59}{25}$$

ja

$$y = -\frac{9}{4}x + \frac{22}{4}$$

Kulmakertoimien tulo on

$$r_{xy}^2 = \left(-\frac{9}{4}\right) \times \left(-\frac{9}{25}\right) = \frac{81}{100} = 0.81$$

Siten muuttujien x ja y havaittujen arvojen *korrelaatio* on

$$r_{xy} = -0.9$$

koska korrelaatiokertoimella ja molempien regressiosuorien kulmakertoimilla on aina sama merkki.

(c) Molempien regressiomallien *selitysaste* on

$$R^2 = r_{xy}^2 = 0.81$$