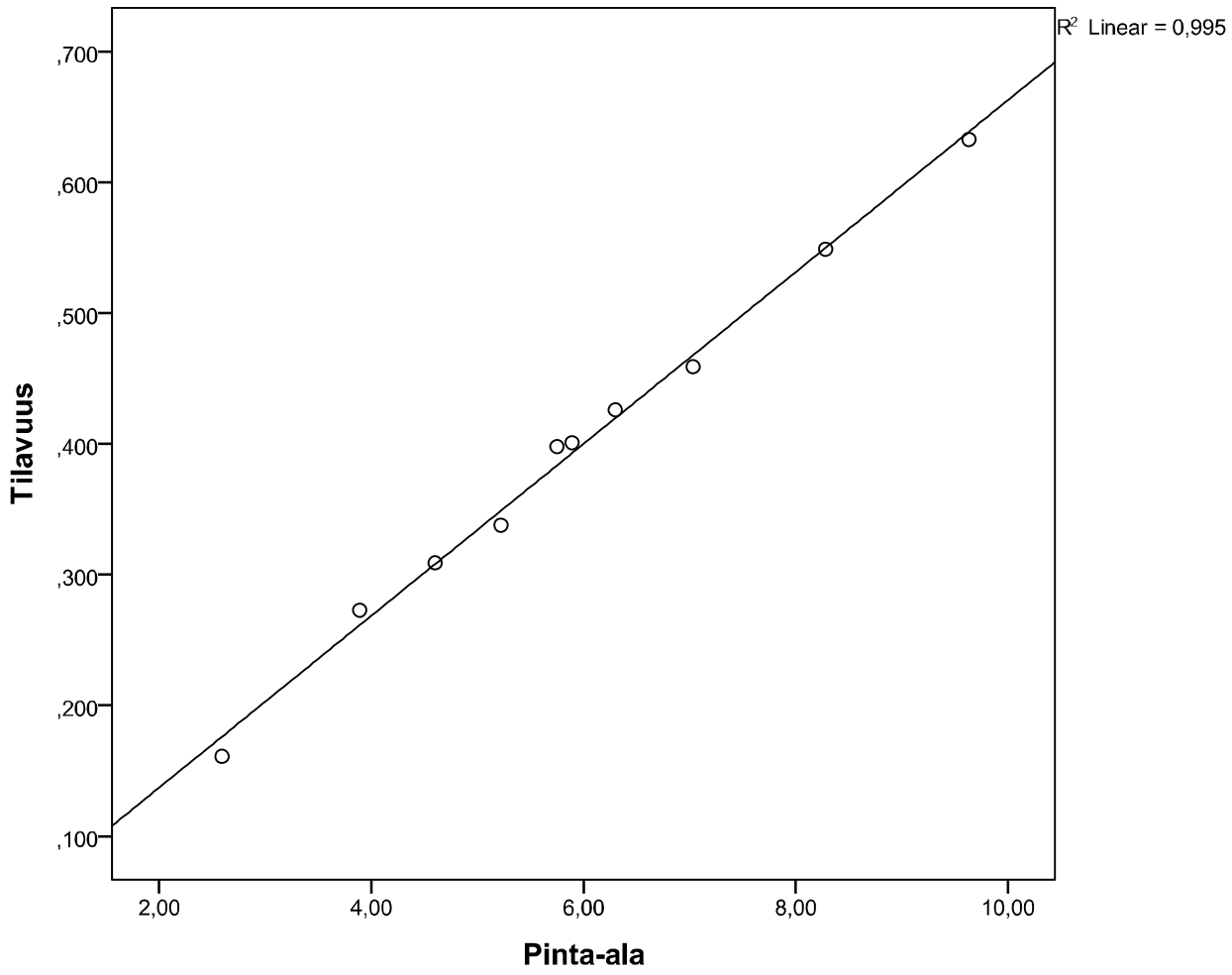
MTTTA1 Tilastomenetelmien perusteet
Luento 31.1.2019

Regressioanalyysi

4.1 Yksi selittävä muuttuja

Esim. 4.1.1 Poimittu samanikäisiä puita, mitattu poikkileikkauspinta-ala sekä puun kuutiomäärä

Pinta-ala	Tilavuus
2,59	0,161
3,89	0,273
...	
9,63	0,633


R² Linear = 0,995
,700
,600
,500
,400
,300
,200
,100
Tilavuus
2,00
4,00
6,00
8,00
10,00
Pinta-ala

Malli

$$\text{Tilavuus} = \beta_0 + \beta_1 \text{Pinta-ala} + \varepsilon$$

Estimointi

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	,006	,011		,547	,599
	Pinta-ala	,066	,002	,997	38,345	,000

a. Dependent Variable: Tilavuus

$$\widehat{\text{Tilavuus}} = \hat{\beta}_0 + \hat{\beta}_1 \text{ Pinta-ala}$$

$\hat{\beta}_0$: 0,006

$\hat{\beta}_1$: 0,066

Jos pinta-ala on 4,60, niin arvioitu tilavuus on
$0{,}006 + 0{,}066 \cdot 4{,}60 = 0{,}310$.

Jos pinta-ala on 4, niin arvioitu tilavuus on
$0{,}006 + 0{,}066 \cdot 4 = 0{,}270$.

Yhden selittäjän regressiomalli

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

missä

- Y on satunnaismuuttuja, havaittavissa oleva, selitettävä
- x on selittäjä, ei-satunnainen, havaittavissa oleva
- ε on satunnaismuuttuja, ei havaittavissa
- β_0 ja β_1 mallin parametrit, estimoidaan aineiston avulla

Malli voidaan esittää myös muodossa

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (1)$$

Malliin liittyvät oletukset ovat

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

Näistä oletuksista seuraa

- $E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i)$
  $= E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)$
  $= \beta_0 + \beta_1 x_i$

- $Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i)$
  $= Var(\varepsilon_i) = \sigma^2$

- Lisäksi $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

  Jokaista x:n arvoa kohden on olemassa Y:n todennäköisyysjakauma, joka on normaalijakauma. Havainnot näistä normaalijakaumista, graafisesti http://www.sis.uta.fi/tilasto/tiltp3/kevat2004/esim_4_1_2.pdf.

Mallin (1) parametrien estimointi

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

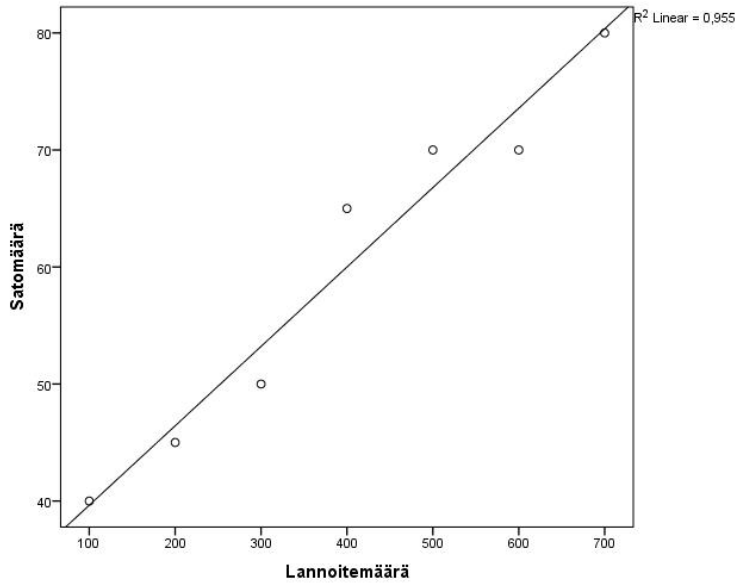
$$= \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$= \frac{SP_{xy}}{SS_x}$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Esim. 4.1.4 Lannoitemäärän vaikutus satoon



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	32,857	2,945		11,157	,000
	Lannoitemäärä	,068	,007	,977	10,304	,000

a. Dependent Variable: Satomäärä

x_i	y_i	$x_i y_i$	x_i^2
100	40	4000	10000
200	45	9000	40000
300	50	15000	90000
400	65	26000	160000
500	70	35000	250000
600	70	42000	360000
700	80	56000	490000
2800	420	187000	1400000

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{187000 - \frac{1}{7} \cdot 2800 \cdot 420}{1400000 - \frac{2800^2}{7}}$$

$$= 0{,}06786$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{420}{7} - 0{,}06786 \cdot \frac{2800}{7} = 32{,}857$$

Voidaan osoittaa, että

- $E(\hat{\beta}_1) = \beta_1$
- $E(\hat{\beta}_0) = \beta_0$

Estimoidut y:n arvot saadaan

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$

Tämä suoran on Y:n odotusarvon estimaatti

Määritellään residuaalit

- $e_i = y_i - \hat{y}_i$

Esim. 4.1.4 (jatkoa)

x_i	y_i	$\hat{y}_i = 32{,}857+0{,}06786x_i$	$e_i = y_i - \hat{y}_i$	
100	40	$32{,}857+0{,}06786\cdot 100=39{,}64$	40-39,64	= 0,36
200	45	$32{,}857+0{,}06786\cdot 200=46{,}43$	45-46,43	=-1,43
300	50	.		=-3,21
400	65	.		= 5,00
500	70	.		= 3,21
600	70	.		=-3,57
700	80	$32{,}857+0{,}06786\cdot 700=80{,}36$	80-80,36	=-0,36

Neliösummat

$$\underbrace{SST}_{Kokonaisneliösumma} = \underbrace{SSR}_{Regressioneliösumma} + \underbrace{SSE}_{Jäännösneliösumma}$$

$$SST = \sum(y_i - \bar{y})^2$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

$$SSE = \sum(y_i - \hat{y}_i)^2$$

Selityskerroin

$R^2 = SSR/SST$

Selitysaste, selitysprosentti

$100 \cdot R^2$

Korrelaatiokerroin

$$r_{xy} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Mallin (1) tilanteessa $(r_{xy})^2 = R^2$.

Esim. 4.1.4 (jatkoa)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,977[a]	,955	,946	3,485

a. Predictors: (Constant), Lannoitemäärä

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1289,286	1	1289,286	106,176	,000[b]
	Residual	60,714	5	12,143		
	Total	1350,000	6			

a. Dependent Variable: Satomäärä

b. Predictors: (Constant), Lannoitemäärä

$$\text{SST} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$= (40^2 + \cdots + 80^2) - \frac{420^2}{7} = 26550 - \frac{420^2}{7} = 1350$$

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2$$

$$= 0{,}36^2 + \ldots + (-0{,}36)^2 = 60{,}7$$

$$\text{SSR} = \text{SST} - \text{SSE} = 1350 - 60{,}7 = 1289{,}3$$

$$R^2 = \text{SSR}/\text{SST} = 0{,}955$$

$$r_{xy} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{187000 - \frac{1}{7} \cdot 2800 \cdot 420}{\sqrt{\left(1400000 - \frac{2800^2}{7}\right)\left(26550 - \frac{420^2}{7}\right)}}$$

$$= \frac{19000}{\sqrt{280000 \cdot 1350}} = 0{,}977$$

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

$$0{,}977^2 = 0{,}955$$