MTTTA1 Tilastomenetelmien perusteet
Luento 5.2.2019

Regressioanalyysi

4.1 Yksi selittävä muuttuja (kertausta ja jatkoa)

Regressiomalli

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

Malliin liittyvät oletukset

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

Mallin estimointi

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

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

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

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

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

Neliösummat

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

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

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

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

$$MSE = SSE/(n-2) = \hat{\sigma}^2$$

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

Testaukset

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-2}, kun\ H_0\ \ tosi,$$

$$s(\hat{\beta}_1) = \sqrt{MSE / SS_x}$$

H_0: $\beta_0 = 0$
H_1: $\beta_0 \neq 0$

$$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-2}, kun\ H_0\ \ tosi,$$

$$s(\hat{\beta}_0) = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SS_x})}$$

Esim. 4.1.4 (jatkoa)

Malli: Satomäärä = $\beta_0 + \beta_1 \cdot$Lannoitemäärä + ε

Kertoimien testaus

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	$\hat{\beta}_0$ = 32,857	$s(\hat{\beta}_0)$ = 2,945		11,157	,000
	Lannoitemäärä	$\hat{\beta}_1$ = ,068	$s(\hat{\beta}_1)$ = ,007	,977	10,304	,000

a. Dependent Variable: Satomäärä

$H_0: \beta_0 = 0$

$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)}$

H_0 hylätään, koska $p < 0{,}001$

$H_0: \beta_1 = 0$

$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$

H_0 hylätään, koska $p < 0{,}001$

$$SSE = \sum(y_i - \hat{y}_i)^2 = 0{,}36^2 + \ldots + (-0{,}36)^2 = 60{,}7$$

$$SS_x = 1400000 - 2800^2/7 = 280000$$

$$\bar{x} = 2800/7 = 400$$

$$MSE = 60{,}7/(7-2) = 12{,}143 = \hat{\sigma}^2$$

$$s(\hat{\beta}_1) = \sqrt{12{,}143/280000} = 0{,}007$$

$$s(\hat{\beta}_0) = \sqrt{12{,}143\left(\frac{1}{7} + \frac{400^2}{280000}\right)} = 2{,}945$$

Päättelyt taulukkoarvon perusteella:

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

$t_{0,01/2,7-2} = 4,032 < 10,304$, H_0 hylätään eli lannoitemäärä pidetään mallissa

H_0: $\beta_0 = 0$
H_1: $\beta_0 \neq 0$

$t_{0,01/2,7-2} = 4,032 < 11,157$, H_0 hylätään eli vakio syytä olla mallissa

Regressiomalli ilman vakiokerrointa

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

Estimointi

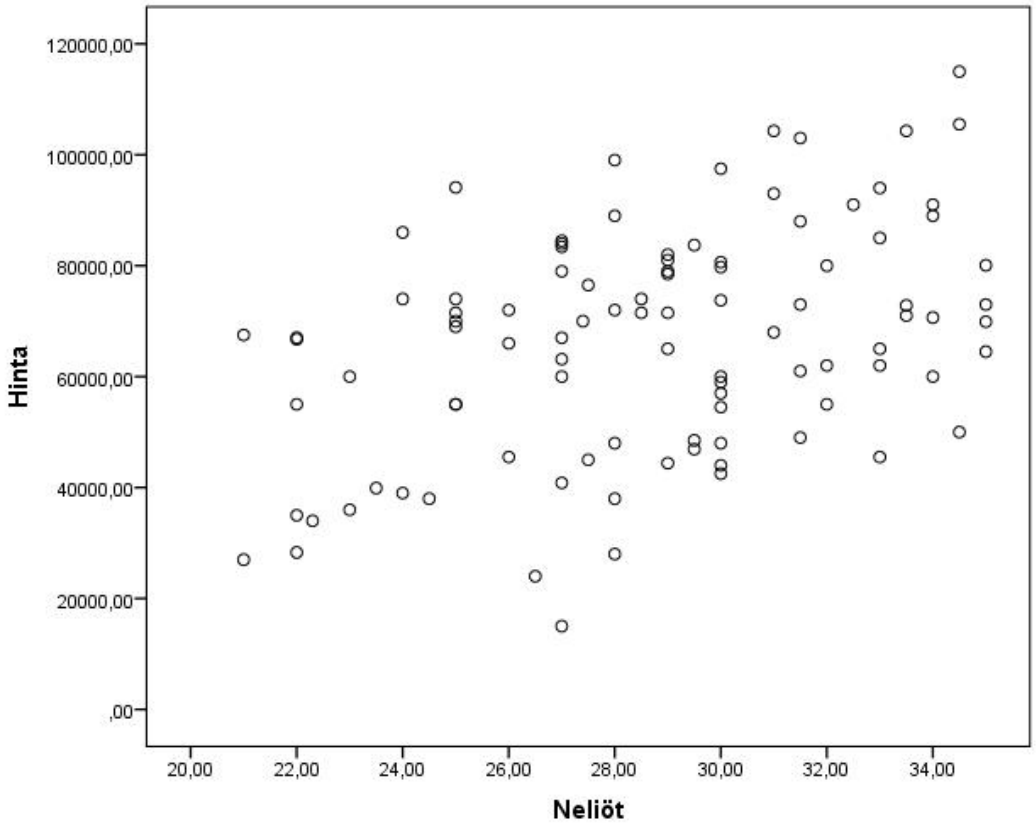
$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{y}_i = \hat{\beta} x_i,$$

Huom! Tällöin R^2 ei ole käytettävissä.

Esim. Aineisto Tre_myydyt_asunnot_2009, sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Malli: Hinta = $\beta_0 + \beta_1 \cdot$Neliöt + ε



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,404[a]	,163	,155	18874,53878

a. Predictors: (Constant), Neliöt

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3076,025	14759,533		,208	,835
	Neliöt	2205,035	509,399	,404	4,329	,000

a. Dependent Variable: Hinta

Hypoteesi H_0: $\beta_0 = 0$ hyväksytään, vakiokerroin voidaan jättää pois mallista.

Estimoidaan uusi malli

$$\text{Hinta} = \beta \cdot \text{Neliöt} + \varepsilon$$

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,963[a]	,928	,927	18781,24256

a. Predictors: Neliöt

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

Nyt ei voida laskea selitysprosenttia!

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	Neliöt	2310,309	65,478	,963	35,284	,000

a. Dependent Variable: Hinta

b. Linear Regression through the Origin

Estimoinnin tulos origon kautta kulkeva suora

$$\widehat{Hinta} = 2310{,}309 \cdot \text{Neliöt}$$

Korrelaatiokertoimen testaus

Populaatiossa muuttujien X ja Y välinen korrelaatiokerroin

$$\rho = \mathrm{Cov}(X, Y)/\sigma_X\sigma_Y.$$

Tätä estimoidaan otoskorrelaatiokertoimella

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

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

Testaus

H_0: $\rho = 0$

H_1: $\rho \neq 0$

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \sim t_{n-2}, kun\ H_0\ \ tosi$$

Esim. 4.1.9 Esimerkin 4.1.4 muuttujat

y = satomäärä
x = lannoitemäärä

r = 0,977, n = 7

H_0: $\rho = 0$
H_1: $\rho \neq 0$

$$t = \frac{0{,}977}{\sqrt{\frac{1 - 0{,}977^2}{7 - 2}}} = 10{,}304 > t_{0{,}01/2;5} = 4{,}032$$

H_0 hylätään 1 %:n riskitasolla. Päätellään lineaarista riippuvuutta olevan.

Esim. Aineisto Jalkapalloilijat_2006 sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = paino
x = pituus

r_{xy} = 0,823679, n = 154

Correlations

		Pelaajan paino	Pelaajan pituus
Pelaajan paino	Pearson Correlation	1	,824**
	Sig. (2-tailed)		,000
	N	154	154
Pelaajan pituus	Pearson Correlation	,824**	1
	Sig. (2-tailed)	,000	
	N	154	154

**. Correlation is significant at the 0.01 level (2-tailed).

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$$t = \frac{0{,}823679}{\sqrt{\frac{1 - 0{,}823679^2}{154 - 2}}} = 17{,}908 > t_{0{,}005;152} = 2{,}617$$

H_0 hylätään 1 %:n riskitasolla. Päätellään lineaarista riippuvuutta olevan.

Regressiomalli: Paino = β_0 + β_1 Pituus + ε

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-87,476	9,227		-9,480	,000
	Pelaajan pituus	,907	,051	,824	17,908	,000

a. Dependent Variable: Pelaajan paino

$t_{hav.}$ = 17,908

Siis korrelaatiokertoimen testaus on sama kuin regressiomallissa (1) β_1:n testaus!

Esim. Aineisto Jalkapalloilijat_2006 sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Regressioanalyysin tuloksia
http://www.sis.uta.fi/tilasto/mttta1/kevat2015/RA_jalkapalloilijat.pdf