MTTTA1 Tilastomenetelmien perusteet
Luento 12.2.2019

4.2 Useampi selittävä muuttuja (jatkoa)

Selittäjien lukumäärä k (k-RA)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Malliin liittyvät oletukset

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

Estimointi

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

Neliösummat

SST = SSR + SSE

MSR = SSR/k, MSE = SSE/(n-k-1) = $\hat{\sigma}^2$

Selityskerroin

R^2 = SSR/SST

Testaukset

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

$$t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \sim t_{n-k-1}, kun\ H_0\ \ tosi$$

H_0: $\beta_1 = \ldots = \beta_k = 0$

H_1: ainakin jokin $\beta_i \neq 0$

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}} \sim F(k, n - k - 1), kun\ H_0\ tosi$$

Esim. CTESTI-aineisto

muuttujien kuvaukset

http://www.sis.uta.fi/tilasto/tiltp1/syksy2004/CTESTI_muuttujienkuvaus.pdf

y = cooper

x_1 = ikä

x_2 = paino

x_3 = hengitystilavuus

Regressioanalyysin tuloksia

http://www.sis.uta.fi/tilasto/mttta1/kevat2015/cooper_3_RA.pdf

Regressioanalyysin taulukko

R^2 = SSR/SST

SSR	k	MSR	F=MSR/MSE
SSE	n-k-1	MSE	~F(k, n-k-1), kun H_0 tosi
SST	n-1		H_0: $\beta_1 = \ldots = \beta_k = 0$

$\hat{\beta}_0$	$s(\hat{\beta}_0)$	$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-k-1}, kun\ H_0: \beta_0 = 0\ tosi$
$\hat{\beta}_1$	$s(\hat{\beta}_1)$	$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-k-1}, kun\ H_0: \beta_1 = 0\ tosi$
...		
$\hat{\beta}_k$	$s(\hat{\beta}_k)$	$t = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \sim t_{n-k-1}, kun\ H_0: \beta_k = 0\ tosi$

Koska

SST = SSR + SSE
1 = SSR/SST + SSE/SST
SSE/SST = 1 – SSR/SST = 1 – R^2, niin

F-testisuure voidaan esittää myös R^2:n avulla

$$F = \frac{SSR/k}{SSE/(n - k - 1)} = \frac{\frac{SSR}{SST}/k}{\frac{SSE}{SST}/(n - k - 1)} = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}}$$

Esim. y = kiinteistön myyntihinta (dollars)
x_1 = asunnon koko (square feet)
x_2 = tontin koko (square feet)
x_3 = makuuhuoneiden lukumäärä
x_4 = kylpyhuoneiden lukumäärä

(Newbold, 1991)

Regressiomalli $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

Estimoinnin tulos (kertoimet ja hajonnat)

$$\hat{y} = 1998{,}5 + 22{,}352\, x_1 + 1{,}4686\, x_2 + 6767{,}3\, x_3 + 2701{,}1\, x_4$$
$$\qquad\quad (2{,}5543) \quad (1{,}4492) \quad (1820{,}8) \quad (1996{,}2)$$

$R^2 = 0{,}9843$, $n = 20$, $k = 4$

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

$t = 22{,}352/2{,}5543 = 8{,}75 > t_{0{,}05/2{,}15} = 2{,}131$

H_0 hylätään

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

$t = 1{,}4686/1{,}4492 = 1{,}01 < t_{0{,}05/2{,}15} = 2{,}131$

H_0 hyväksytään

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

$t = 6767,3/1820,8 = 3,72 > t_{0,05/2,15} = 2,131$

H_0 hylätään

$H_0: \beta_4 = 0$
$H_1: \beta_4 \neq 0$

$t = 2701,1/1996,2 = 1,35 < t_{0,05/2,15} = 2,131$

H_0 hyväksytään

H_0: $\beta_1 = \ldots = \beta_4 = 0$

H_1: ainakin jokin $\beta_i \neq 0$

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

$$F_{Hav.} = \frac{\frac{0{,}9843}{4}}{\frac{1 - 0{,}9843}{20 - 4 - 1}} = 235{,}1 > F_{0{,}01;4,15} = 4{,}89$$

H_0 hylätään

Jos selittävät muuttujat ovat keskenään voimakkaasti korreloituneita (multikollineaarisia), saattaa käydä niin, että

$H_0: \beta_1 = ... = \beta_k = 0$

hylätään (tehdään päättely, että ainakin jokin $\beta_i \neq 0$), mutta kaikki hypoteesit

$H_0: \beta_i = 0$ hyväksytään.

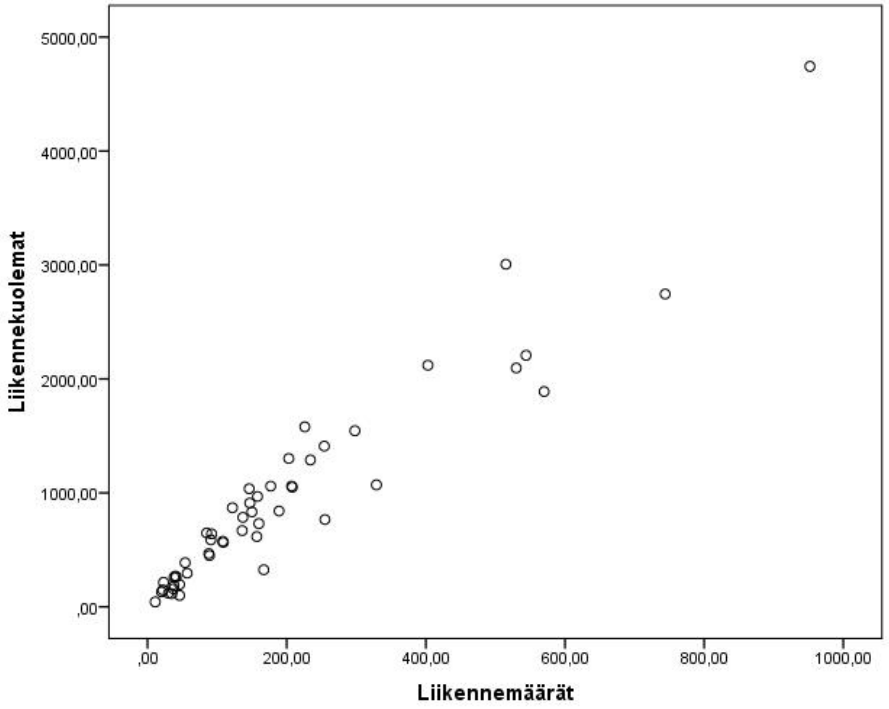
4.3 Selittävien muuttujien valinnasta ja mallin oletuksista

Mallin valinnasta

- Tarpeeksi selittäjiä, mutta käyttötarkoitukseen sopiva, tulkittavissa oleva malli.
- Tarvittaessa muunnokset, jotta mallin oletuksen voimaan.
- Automaattiset mallinvalintamenetelmät
  - etenevä valinta (Forward)
  - taaksepäin eliminointi (Backward)
  - askeltava valinta (Stepwise)

Esim. 4.3.1 Aineisto Liikennekuolemat sivulla
https://coursepages.uta.fi/mttta1/esimerkkiaineistoja/

y = liikennekuolemat
x = liikennemäärät



Malli I

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

Mallin oletukset

$\varepsilon_i \sim N(0, \sigma^2)$ ja

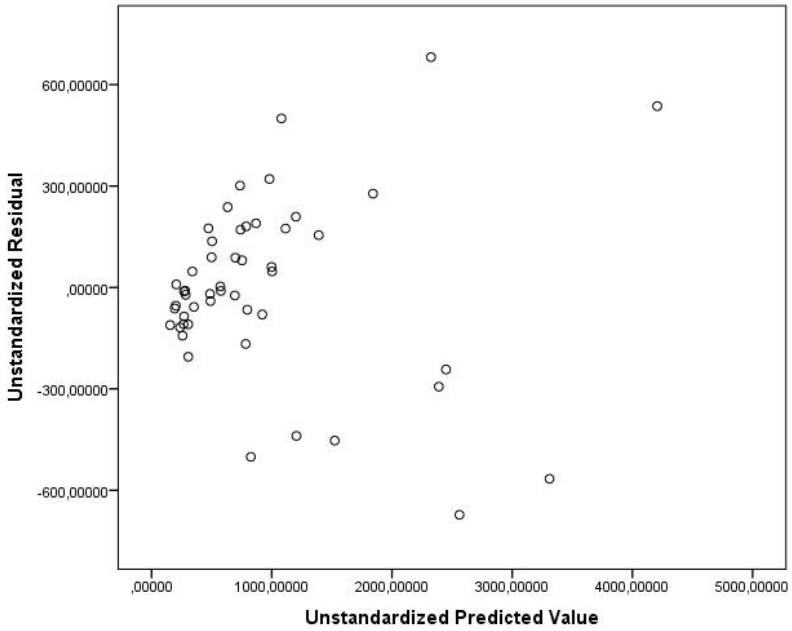
ε_i:t ovat riippumattomia

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	107,029	52,099		2,054	,045
	Liikennemäärät	4,306	,191	,956	22,549	,000

a. Dependent Variable: Liikennekuolemat

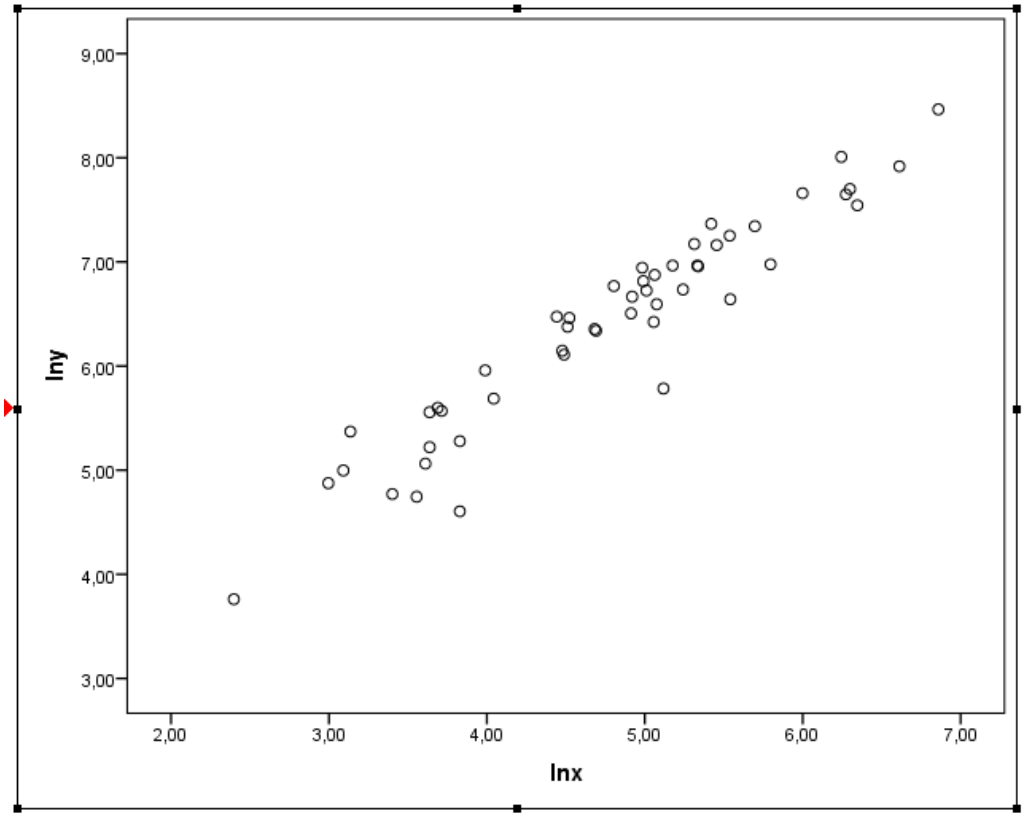
$R^2 = 0{,}914$



Residuaalitarkastelut: ei voi olettaa, että $Var(\varepsilon_i) = \sigma^2$, kun i = 1, 2, …, n.

Malli II

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

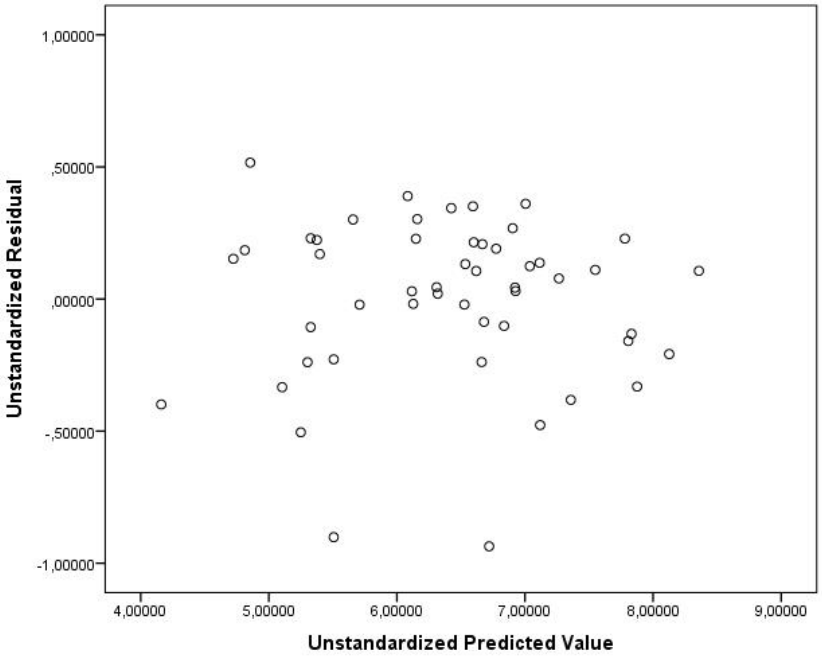


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,904	,209		9,101	,000
	lnx	,941	,043	,954	21,965	,000

a. Dependent Variable: lny

R^2 = 0,910, residuaalitarkastelut OK



Esim. Aineisto Audi_A6 sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = auton hinta
x = vuosimalli
z = ajetut kilometrit
v = moottorin tilavuus

Malleja:

- $Y = \beta_0 + \beta_1 x + \varepsilon$
- $\ln(Y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$
- $Y = \beta_0 + \beta_1 z + \varepsilon$
- $Y = \beta_0 + \beta_1 z + \beta_2 z^2 + \varepsilon$
- $Y = \beta_0 + \beta_1 x + \beta_2 v + \varepsilon$
- $Y = \beta_0 + \beta_1 x + \beta_2 v + \beta_3 z + \varepsilon$