MTTTA1 Tilastomenetelmien perusteet
Luento 14.2.2019

4.2 Useampi selittävä muuttuja (kertausta)

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

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

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

Esim. Ilmansaasteille altistumisen vaikutus kuolleisuuteen suurkaupungeissa (Devore&Peck)

y = total mortality rate (deaths per 10000)
x_1 = mean suspended particle reading ($\mu g/m^3$)
x_2 = smallest sulfate reading (($\mu g/m^3$)x10)
x_3 = population density (people/mi^2)
x_4 = (percent nonwhite)x10
x_5 = (percent over 65)x10

Regressiomalli

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

Estimoinnin tuloksia (kertoimet, kertoimien estimoituja hajontoja)

$$\hat{y} = 19{,}607 + \underset{(0{,}016)}{0{,}041 x_1} + 0{,}071 x_2 + 0{,}001 x_3 + \underset{(0{,}007)}{0{,}014\, x_4} + 0{,}687 x_5$$

$R^2 = 0{,}827$, $n = 117$, $k = 5$

$H_0: \beta_1 = 0$

$H_1: \beta_1 > 0$

$t = 0{,}041/0{,}016 = 2{,}5625$

$t_{0{,}01,111} = 2{,}358$, $t_{0{,}005,111} = 2{,}617$

Siis $0{,}005 <$ p-arvo $< 0{,}01$

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

$t = 0{,}041/0{,}007 = 5{,}86 > t_{0{,}005,111} = 2{,}617$

H_0 hylätään, p-arvo $< 0{,}005$

H_0: $\beta_1 = \ldots = \beta_5 = 0$

H_1: ainakin jokin $\beta_i \neq 0$

$$F = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}} \sim F(k, n - k - 1), \text{ kun } H_0 \text{ tosi}$$

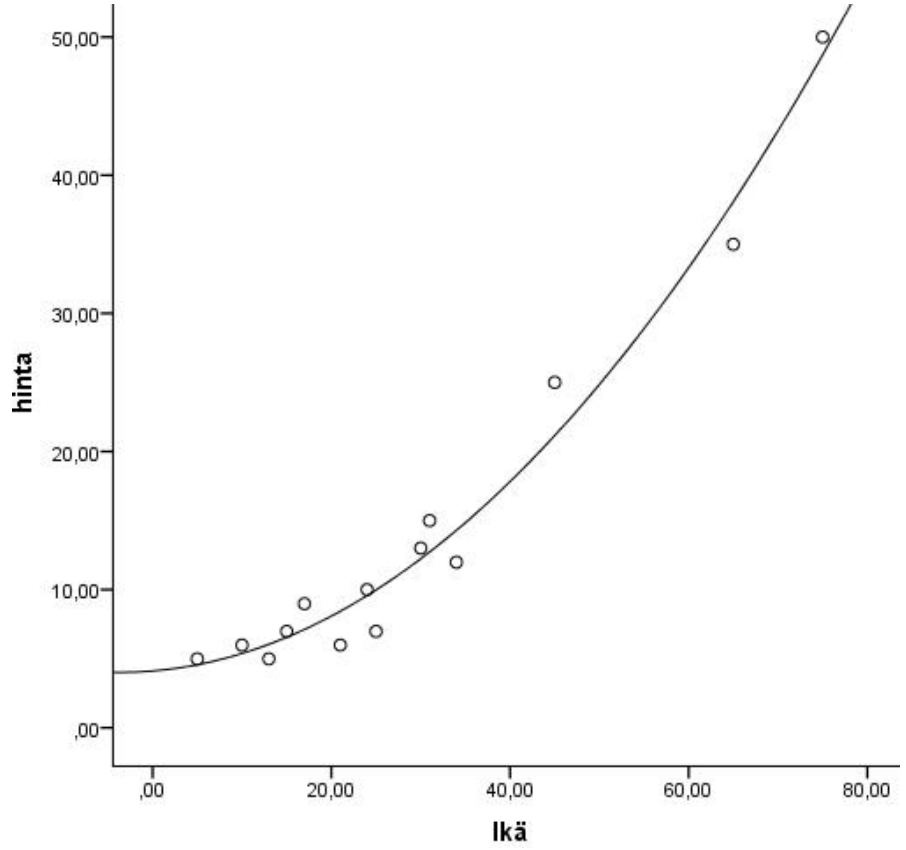
$$F = \frac{\frac{0{,}827}{5}}{\frac{1 - 0{,}827}{117 - 5 - 1}} = 106{,}124 > F_{0{,}01;5,111} = 3{,}02$$

H_0 hylätään

4.3 Selittävien muuttujien valinnasta ja mallin oletuksista (jatkoa)

Esim. Polynomiregressio, y = viinin hinta, x = viinin ikä

hinta	ika	ika2
50,00	75,00	5625,00
35,00	65,00	4225,00
25,00	45,00	2025,00
11,98	34,00	1156,00
15,00	31,00	961,00
13,00	30,00	900,00
6,98	25,00	625,00
10,00	24,00	576,00
5,99	21,00	441,00
8,98	17,00	289,00
6,98	15,00	225,00
4,99	13,00	169,00
5,98	10,00	100,00
4,98	5,00	25,00



Malli $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,114	2,015		2,042	,066
	Ikä	,055	,126	,083	,432	,674
	IkäxIkä	,007	,002	,906	4,699	,001

a. Dependent Variable: hinta

Malli $Y = \beta_0 + \beta_1 x^2 + \varepsilon$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,918	,745		6,605	,000
	IkäxIkä	,008	,000	,987	21,391	,000

a. Dependent Variable: hinta

$R^2 = 0,974$

Esim. Autoregressio

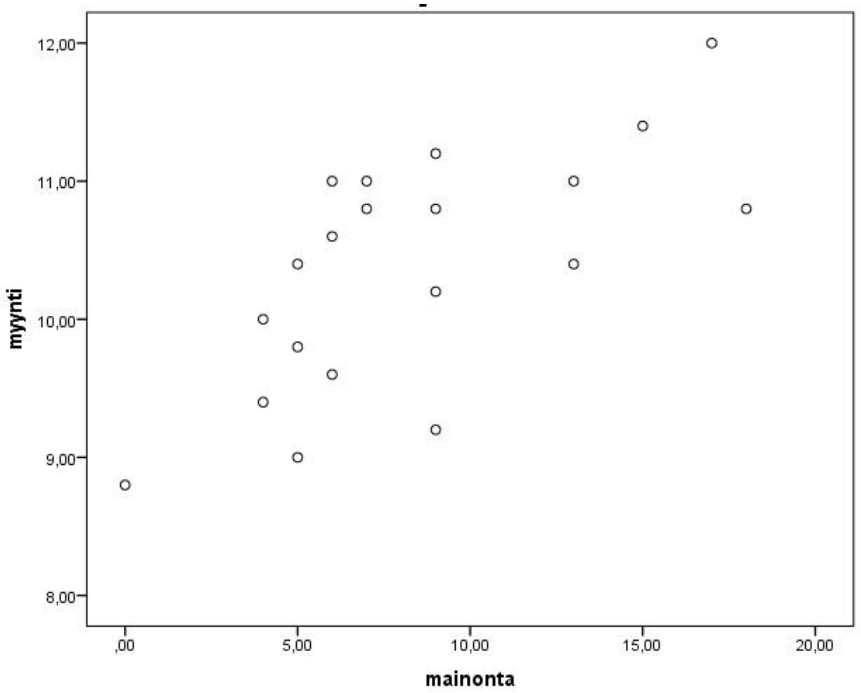
Tutkitaan vaikuttaako TV-mainonta tavaratalon myyntiin. Tarkastellaan viikoittaista myyntiä 20 viikon ajan, aineisto myynti_mainonta.sav sivulla https://coursepages.uta.fi/mttta1/esimerkkiaineistoja/

y = myynti
x = mainonta

viikko	mainonta	myynti
1	,00	8,80
2	5,00	9,00
3	9,00	9,20
4	4,00	10,00
5	6,00	9,60
6	13,00	10,40

...

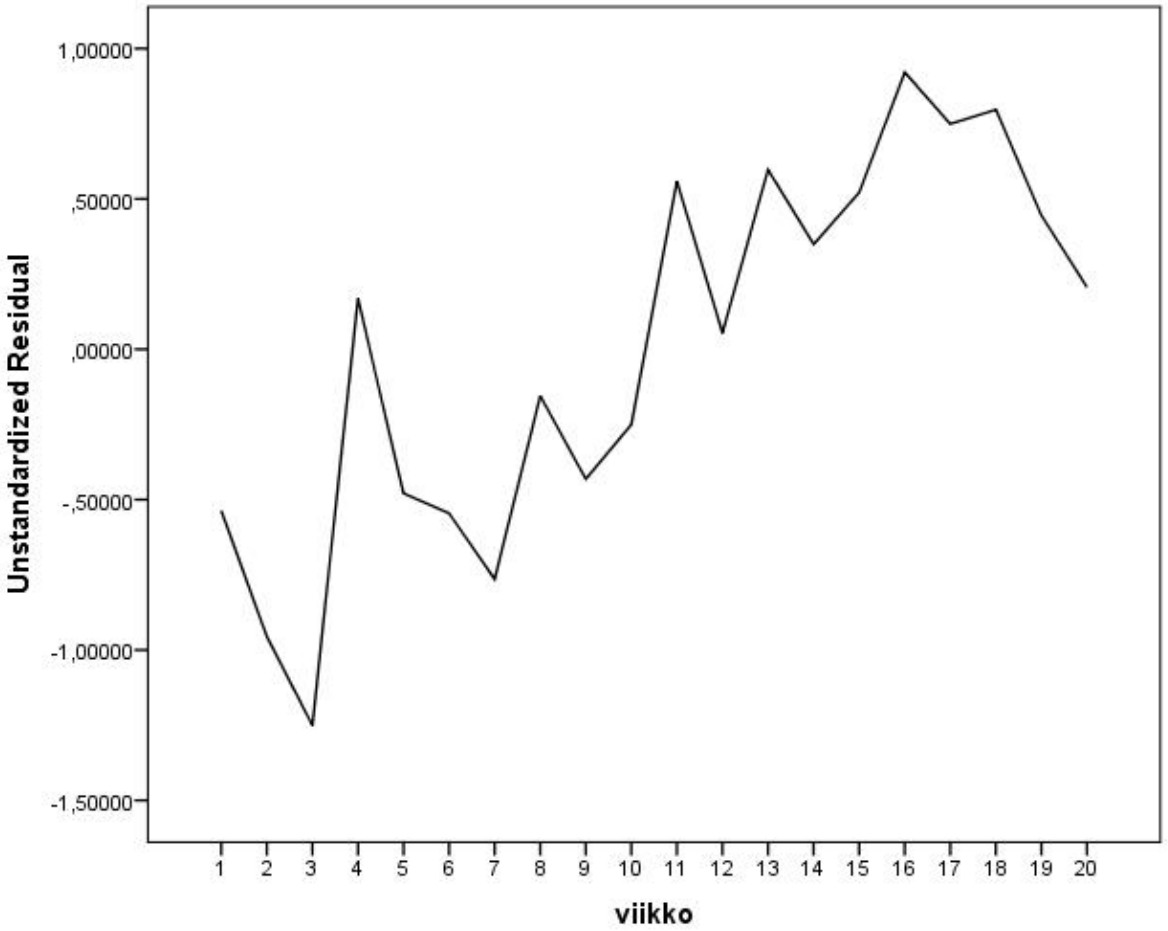
Malli I: $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9,336	,301		31,020	,000
	mainonta	,124	,032	,678	3,916	,001

a. Dependent Variable: myynti



Autokorreloituneet residuaalit

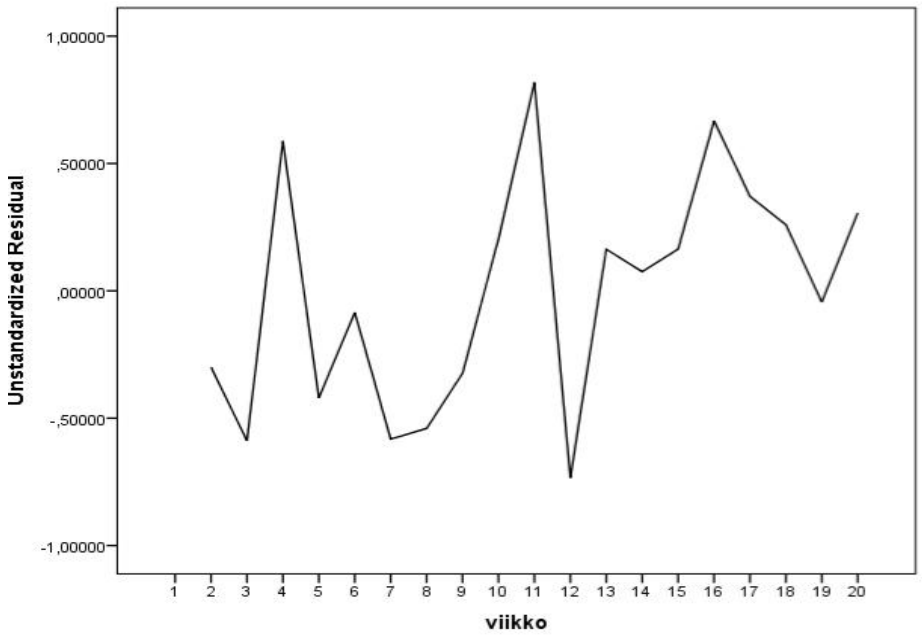
Malli II $Y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$

Autoregressio

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	4,240	1,401		3,027
	mainonta	,096	,026	,530	3,651
	myynti_1	,520	,137	,553	3,808

a. Dependent Variable: myynti

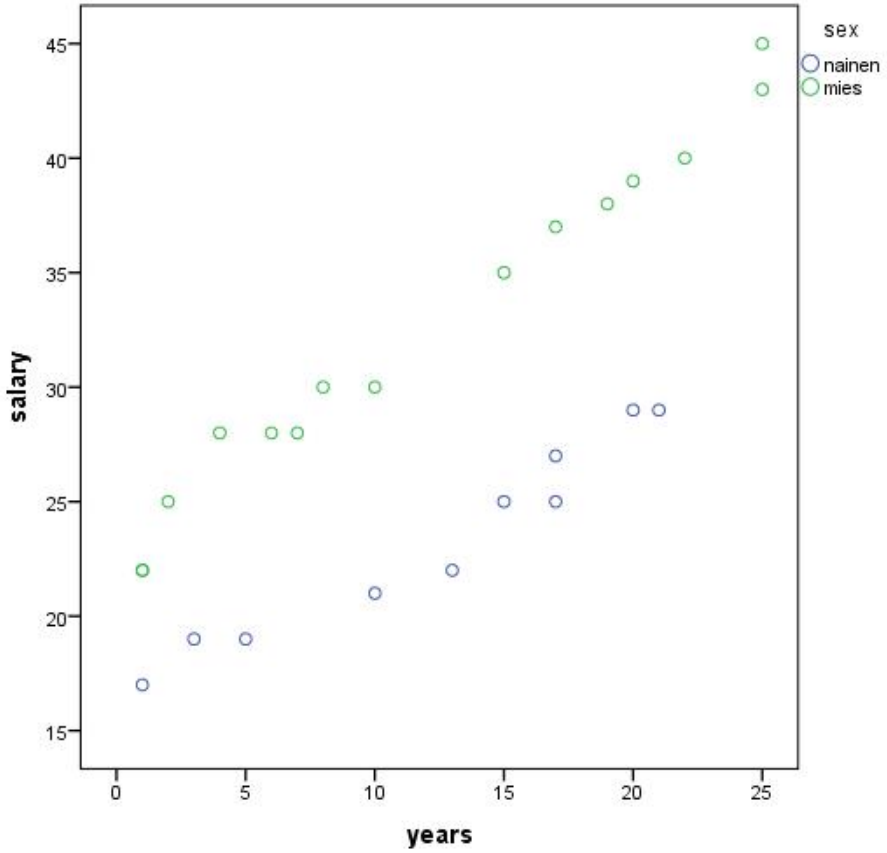


viikko	myynti	mainonta	myynti_1
1	8,80	,00	.
2	9,00	5,00	8,80
3	9,20	9,00	9,00
4	10,00	4,00	9,20
5	9,60	6,00	10,00
6	10,40	13,00	9,60
7	10,80	18,00	10,40

...

Esim. 4.3.3 Dummy-muuttuja selittäjänä mallissa

y = Salary
x = Years
z = Sex (0 = nainen, 1 = mies)



Salary	Years	Sex*
35	15	1
27	17	0
45	25	1
22	13	0
25	2	1
30	10	1
37	17	1
25	17	0
17	1	0
28	4	1
43	25	1
25	15	0
22	1	1

Salary	Years	Sex*
28	6	1
29	20	0
19	3	0
29	21	0
38	19	1
19	5	0
22	1	1
39	20	1
40	22	1
21	10	0
28	7	1
30	8	1

*1 = mies

Malli $Y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$

$E(Y) = \beta_0 + \beta_1 x$, kun Sex=0 (naiset)
$E(Y) = \beta_0 + \beta_1 x + \beta_2$, kun Sex = 1 (miehet)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13,970	,627		22,277	,000
	years	,765	,036	,782	21,191	,000
	sex	9,418	,577	,603	16,335	,000

a. Dependent Variable: salary

Naisilla $\widehat{Salary} = 13{,}970 + 0{,}765 \cdot Years$

Miehillä $\widehat{Salary} = 13{,}970 + 0{,}765 \cdot Years + 9{,}418$
$= 23{,}388 + 0{,}765 \cdot Years$

4.4 Varianssianalyysimalli

Oletukset yksisuuntaisessa varianssianalyysissä:

$$
\begin{aligned}
&Y_{11}, Y_{12}, \ldots, Y_{1n_1} && \text{satunnaisotos } N(\mu_1, \sigma^2)\text{:sta,}\\
&Y_{21}, Y_{22}, \ldots, Y_{2n_2} && \text{satunnaisotos } N(\mu_2, \sigma^2)\text{:sta,}\\
&\quad\vdots\\
&Y_{I1}, Y_{I2}, \ldots, Y_{In_I} && \text{satunnaisotos } N(\mu_I, \sigma^2)\text{:sta.}
\end{aligned}
$$

Halutaan tutkia ovatko jakaumien odotusarvot yhtä suuret, jolloin

$$
\begin{aligned}
&H_0\colon \mu_1 = \mu_2 = \cdots = \mu_I,\\
&H_1\colon \text{kaikki odotusarvot eivät ole samoja.}
\end{aligned}
$$

Oletuksista seuraa, että varianssianalyysi voidaan ajatella mallina

$$
Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \text{missä } \varepsilon_{ij} \sim N(0, \sigma^2).
$$

$\mu_1, \mu_2, \ldots, \mu_I$ ovat mallin parametrit. Vaihtoehtoisesti myös $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$.

Luku 5

Epäparametrisista menetelmistä (ei tenttiin)

Ei oletuksia populaatiosta, esim. normaalijakaumaoletusta.

- Mann-Witneyn testi
  Kahden riippumattoman otoksen t-testin epäparametrinen vastine (normaalijakaumaoletus ei voimassa)
- Kruskal-Wallisin testi
  Epäparametrinen vastine yksisuuntaiselle varianssianalyysille (normaalijakaumaoletusta ei

tehdä, selitettävä muuttuja voi olla järjestysasteikollinen)

- Welchin tai Brown-Forsythen testi
  Yksisuuntainen varianssianalyysi, kun oletus varianssien yhtäsuuruudesta ei voimassa

Ks. luentorunko s. 51,
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=52

Tentit

- ti 26.2.2019 klo 12.15-15.00 ls. A1, voi osallistua, jos on tehnyt vähintään 30 % harjoituksista, ilmoittaudu viimeistään 24.2.
- pe 5.4.2019
- pe 3.5.2019
- to 6.6.2019

Osaamistavoitteet

http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luento_10_1_2019.pdf

Mitä jatkoksi?

Matematiikan ja tilastotieteen tutkinto-ohjelman opiskelijat

Matematiikan ja tilastotieteen perusopinnot (tilastotieteen opintopolku)
https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14974&lang=fi&uiLang=fi&lvv=2018

- MTTTP4 Todennäköisyyslaskenta (S2019)

Tilastotieteen aineopinnot (pakolliset)

https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14600&lang=fi&uiLang=fi&lvv=2018

- MTTTA2 Matemaattisen tilastotieteen perusteet, (S2019)
- MTTTA4 Tilastollinen päättely 1, (K2020)
- MTTTA14 Tilastotieteen matriisilaskenta ja laskennalliset menetelmät, (S2019)

Tilastotieteen aineopinnot (muut)
https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14611&lang=fi&uiLang=fi&lvv=2018

- MTTTA13 Empiirinen projekti

Tilastotieteen perusopintokokonaisuus valinnaisina opintoina

https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14644&lang=fi&uiLang=fi&lvv=2018

- MTTTA13Empiirinen projekti

Tilastotieteen aineopintokokonaisuus valinnaisina opintoina

https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14909&lang=fi&uiLang=fi&lvv=2018

Pakolliset

- MTTTP4 Todennäköisyyslaskenta, (S2019)
- MTTTA2 Matemaattisen tilastotieteen perusteet (S2019)
- MTTTA4 Tilastollinen päättely 1, (K2020)
- MTTTA14 Tilastotieteen matriisilaskenta ja laskennalliset menetelmät, (S2019)

Muut (valitaan 3)

- MTTTA5 Monimuuttujamenetelmät
- MTTTA6 Regressioanalyysi
- MTTTA7 Yleistetyt lineaariset mallit 1
- MTTTA9 Tilastollinen ennustaminen
- MTTTA10 Sekamallit
- MTTTA11 Tilastolliset ohjelmistot (esitietona MTTTA1)
- MTTA2 Muu erikseen sovittava opintojakso