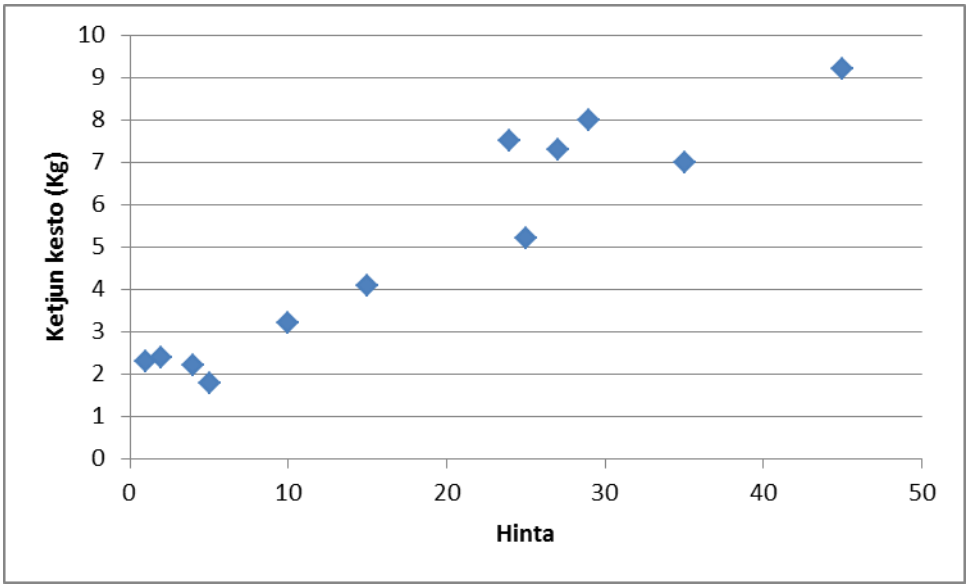
Harjoittelua:

Teollisuudessa käytettävän ketjun hinnan ja rasituskeston välillä on lineaarista riippuvutta.

Harjoittelua:

- Laske hinnan ja ketjun keston regressioyhtälö.
- Minkä hintainen ketju kannattaa ostaa jos toivoo noin 5 kg kestoa ketjulta?
- Liike ostaa nyt 15 € hintaisia ketjuja, kuinka paljon kestoa saataisiin lisää jos ostettaisiin 18 € hintaisia ketjuja.
- Laske yhtälön selitysaste ja sen tulkinta

Hinta (€)	Kesto (Kg)
1	2,3
2	2,4
4	2,2
5	1,8
10	3,2
15	4,1
24	7,5
25	5,2
27	7,3
29	8
35	7
45	9,2



	x(€)	y(kg)	x^2	y^2	xy
	1	2.3	1	5.29	2.3
	2	2.4	4	5.76	4.8
	4	2.2	16	4.84	8.8
	5	1.8	25	3.24	9
	10	3.2	100	10.24	32
	15	4.1	225	16.81	61.5
	24	7.5	576	56.25	180
	25	5.2	625	27.04	130
	27	7.3	729	53.29	197.1
	29	8	841	64	232
	35	7	1225	49	245
	45	9.2	2025	84.64	414
Summa	222	60.2	6392	380.4	1516.5

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b_0 = \frac{1}{n}\left(\sum y_i - b_0 \sum x_i\right) = \bar{y} - b_0\bar{x}$$

$\bar{y}$=60.2/12=5.02

$\bar{x}$=222/12=18.5

$b_1 = \frac{12*1516.5-222*60.2}{12*6392-222^2}$ =0.176

$b_0 = 5.02 - 0.176 * 18.5$= 1.76

Regressioyhtälö:

$\hat{y}$ = 1.76 + 0.176*x

Jos toivotaan viiden kilon kestoa:

5 = 1.76+0.176*x

<=> 5 - 1.76 = 0.176*x

<=> (5-1.76)/0.176 = x = 18.41 (€)

Jos ketjun hinta kasvaa 3 €:

3*0.176 = 0.528 (eli kesto kasvaa noin ½ kg)

Selitysaste: Lasketaan ensin otoskeskihajonnat

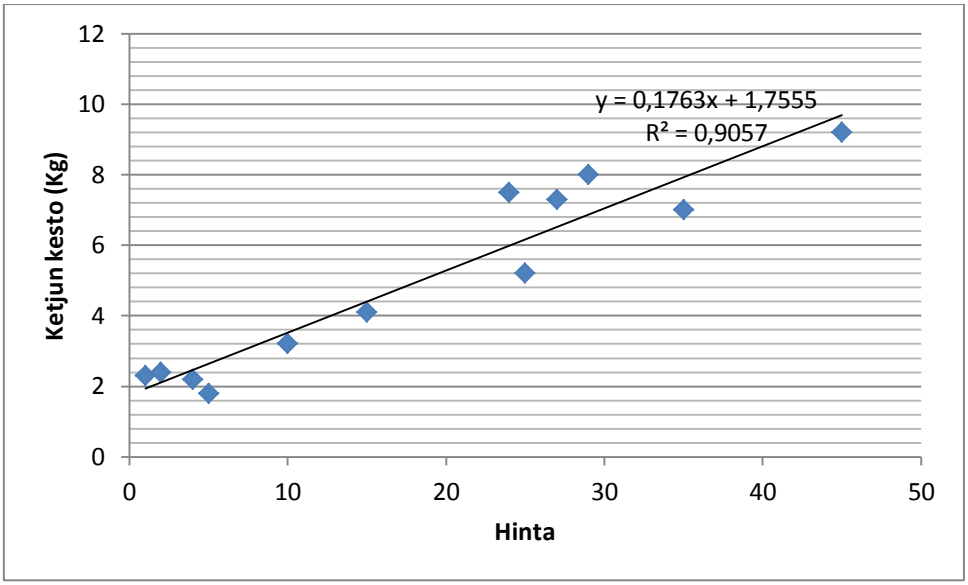
$s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x_i)^2}{n}}{n-1}} = \sqrt{\frac{6392-\frac{222^2}{12}}{12-1}} =$

$\sqrt{207.72} = 14.4$

$s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{\sum y^2 - \frac{(\sum y_i)^2}{n}}{n-1}} = \sqrt{\frac{380.4-\frac{60.2^2}{12}}{12-1}} =$

$\sqrt{7.13} = 2.67$

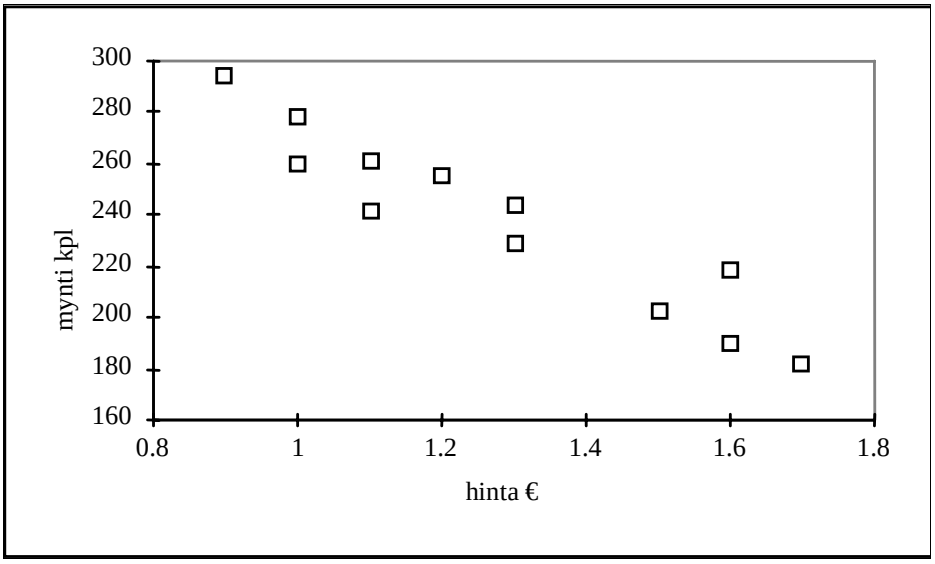
Selitysaste: $\left(\frac{s_x}{s_y} * b_1\right)^2 = \left(\frac{14.4}{2.67} * 0.176\right)^2 = 0.90$


y = 0,1763x + 1,7555
R² = 0,9057
Ketjun kesto (Kg)
Hinta
0
2
4
6
8
10
12
0
10
20
30
40
50

Esimerkki: Leipomo teetti tutkimuksen siitä, miten hillomunkkien menekki riippuu niiden hinnasta. Tutkimusajanjakson aikana munkkeja myytiin vaihtelevilla tarjoushinnoilla ja saatiin seuraava aineisto

hinta €	0.90	1.00	1.00	1.10	1.10	1.20	1.30	1.30	1.50	1.60	1.60	1.70
myynti kpl	294	278	260	261	242	255	244	229	203	190	218	182

Aineiston sirontakuvio on alla.



Regressiomallin kertoimet:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - b_1 \bar{x}$$

Lasketaan esimerkin 22 aineistolle regressiosuora pienimmän neliösumman menetelmällä:

x_i	y_i	$x_i y_i$	x_i^2
0.90	294	264.6	0.81
1.00	278	278	1.00
1.00	260	260	1.00
1.10	261	287.1	1.21
1.10	242	266.2	1.21
1.20	255	306	1.44
1.30	244	317.2	1.69
1.30	229	297.7	1.69
1.50	203	304.5	2.25
1.60	190	304	2.56
1.60	218	348.8	2.56
1.70	182	309.4	2.89
Σ=15.3	Σ=2856	Σ=3543.5	Σ=20.31

Regressiosuoran kertoimet b_0 ja b_1 saadaan seuraavasti:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{12 \cdot 3543.5 - 15.3 \cdot 2856}{12 \cdot 20.31 - (15.3)^2} = -122.0$$

ja

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{15.3}{12} = 1.275,\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{2856}{12} = 238,$$

joten

$$b_0 = \bar{y} - b_1 \bar{x} = 238 + 122.0 \cdot 1.275 = 393.55$$

Regressiosuoran yhtälö on siis $\hat{y} = 393.55 - 122.0 \cdot x$.

TKMY2 Regressioanalyysi

ESIMERKKI 1.

a. **Muodosta muistitestin pistemäärää selittävä regressioyhtälö ja tulkitse se.**
b. **Määritä suoran selitysaste ja tulkitse se.**
c. **Mikä on odotettavissa oleva muistitestin pistemäärä, kun mainoksen pituus on 30 sekuntia?**
d. **Kuinka monta pistettä on muistitestin pistemäärän odotettu muutos, kun mainoksen pituuden muutos on 5 sekuntia (kasvua)?**

	Mainoksen pituus (s)	Muistitestin pistemäärä	
Henkilö	x	y	
1	20	10	
2	24	8	
3	28	10	
4	32	11	
5	36	14	
6	40	16	
7	44	12	
8	48	13	
	$\sum x = 272$	$\sum y = 94$	$\sum xy = 3320$
	$\sum x^2 = 9920$	$\sum y^2 = 1150$	

$$b_1 = \frac{8 \cdot 3320 - 272 \cdot 94}{8 \cdot 9920 - 272^2} \approx 0{,}185$$

$$b_0 = \frac{94}{8} - 0{,}185\frac{272}{8} \approx 5{,}476$$

Regressioyhtälö on siis

$$\hat{y} = 5{,}476 + 0{,}185x$$

Eli: Kun x kasvaa yhden yksikön niin y kasvaa keskimäärin 0,185 yksikköä. Tässä siis yksi lisäsekunti mainoksen pituuteen merkitsee 0,185 pistettä lisää muistitestin pistemäärään.

Määritä suoran selitysaste ja tulkitse se.

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

$$= \left(\frac{s_X}{s_Y} b_1\right)^2 = \left(\frac{9.798}{2.550} 0.185\right)^2 = r_{XY}^2 = 0.71^2 \approx 0.50$$

Muistitestin pistemäärän vaihtelusta voitiin selittää 50% radiomainoksen pituuden avulla.

Mikä on odotettavissa oleva muistitestin pistemäärä, kun mainoksen pituus on 30 sekuntia?

$$\hat{y}_{X=30} = 5{,}476 + 0{,}185 * 30 \approx 11.0$$

Keskimääräinen muistitestin pistemäärä on 11.0, kun mainoksen pituus on 30 sekuntia.

Kuinka monta pistettä on muistitestin pistemäärän odotettu muutos, kun mainoksen pituuden muutos on 5 sekuntia (kasvua)?

$$\Delta y_{\Delta X=5} = 0{,}185 * 5 \approx 0.925$$

Muistitestin muutos keskimäärin on 0.925 pistettä kun mainoksen pituus on 5 sekuntia suurempi.

ESIMERKKI 2.

Kolmen selittäjän regressiomallin yksittäiset kertoimet,

Muuttuja	Kerroin
vakiotermi	-16,058
X1	4,146
X2	-0,236
X3	4,831

Selitysasteeksi saadaan 0,916. Malli toimii siis hyvin.

Huom! Yksi kolmen selittäjän regressio ei ole sama asia kuin kolme yhden selittäjän regressiota. Kertoimia laskettaessa otetaan myös muut selittäjät huomioon.

ESIMERKKI 3.

Dummy- eli kaksiarvoisen (arvot 0 ja 1) muuttujan käyttö regressioanalyysissä: Dummy-muuttujia käytetään kategorista selittäjää kuvattaessa.

Dummy-muuttujan kulmakertoimen tulkinta: Vakiotermin b_0 ero eli suorien välimatka kahdessa ryhmässä. Seuraavassa kategorisella muuttujalla on kolme erilaista arvoa, joten dummy-muuttujia tarvitaan kaksi:

Regressioyhtälöksi saadaan

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

jossa selitettävä (Y) muuttuja on asunnon hinta, selittävä muuttuja (X1) pinta-ala ja dummy-muuttujat (X2 ja X3) lämmitysmuodosta kertova muuttuja. Mahdollisia lämmitysmuotoja ovat sähkö, öljy ja kaasu. Dummy-muuttujia tarvitaan kaksi, koska mahdollisia lämmitysmuotoja on kolme. Dummy-muuttuja –parin arvo (0,0) merkitsee sähköä, (1,0) öljyä ja (0,1) kaasua.

Regressioyhtälöksi saadaan

$$\hat{y} = 21{,}4 + 3{,}69 x_1 - 12{,}4 x_2 + 13{,}7 x_3$$

eli kaavasta nähdään erikseen regressio-yhtälöt kolmelle eri lämmitysmuodolle. Sähkölle se on

$$\hat{y} = 21{,}4 + 3{,}69x_1$$

öljylle

$$\hat{y} = 21{,}4 + 3{,}69x_1 - 12{,}4 \cdot 1 = 9{,}0 + 3{,}69x_1$$

ja kaasulle

$$\hat{y} = 21{,}4 + 3{,}69x_1 + 13{,}7 \cdot 1 = 35{,}1 + 3{,}69x_1$$

Jos kategorisella selittäjällä on K kpl erilaista arvoa, tarvitaan K-1 kpl dummy-muuttujia niiden arvojen esittämiseksi.