

Biomeetria praks 6

Illustreeritud (mittetäielik) tööjuhend

Eeltöö

1. Avage *MS Excel*'is ankeedivastuseid sisaldav andmestik,
 2. lisage uus tööleht, nimetage see ümber leheküljeks 'Praks6' ja
 3. kopeerige kogu 'Andmed'-lehel paiknev andmetabel lehekülje 'Praks6' ülemisse vasakusse nurka.
-

Ülesanne 1.

- Illustreerige tunnuste 'HEIGHT' ja 'SHOE_SIZE' vahelist seost hajuvus- ehk punktdiagrammiga.
- Jälgige, et x-telg (horisontaalne telg) vastaks jalanumbritele ja y-telg (vertikaalne telg) pikkustele. Vajadusel kujundage joonis ümber.
- Prognoosimaks pikkust jalanumbri alusel, lisage joonisele lineaarne regressioonisirge, samuti regressioonivõrrand ja viimase alusel leitavate prognooside täpsust kirjeldav determinatsioonikordaja R^2 .
- Prognoosige leitud võrrandi alusel, keskmiselt kui pikk on jalanumbrit 40 omav tudeng.

Ülesanne 2.

- Teostage statistikaprotseduuri Regression (Data-sakk -> Data analysis...) abil lineaarne regressioonanalüüs prognoosimaks tudengite pikkust jalanumbri alusel.
 - Kirjutage protseduuri tulemuste põhjal välja lineaarne regressioonivõrrand (ehk regressioonimudel) kujul
$$Pikkus = a + b \times Jalanumber,$$
kus a ja b asemel on Exceli poolt välja arvatud kordajate väärtused.
 - Kui suur on keskmiselt pikkuste vaheline erinevus tudengitel, kelle jalanumbrid erinevad 2 võrra?
 - Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!
 - Kirjeldamiseks prognooside täpsust, sõnastage üks lause kas mitmese korrelatsioonikordaja (R), mudeli standardvea (*Standard Error*) või determinatsioonikordaja (R^2) kohta.
-

Ülesande 1 tööjuhend

1. Illustreerige tunnuste 'HEIGHT' ja 'SHOE_SIZE' vahelist seost hajuvus- ehk punktdiagrammiga.

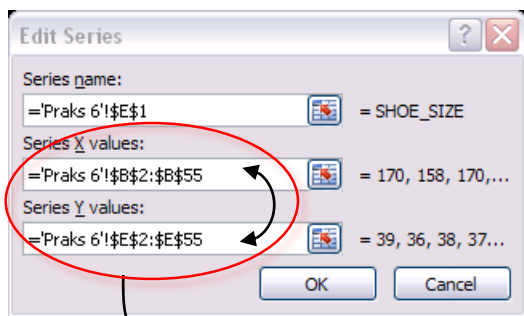
Joonisel peab x-telg vastama jalanumbritele ja y-telg pikkustele. Vajadusel kujundage joonis ümber (vt allpool toodud juhendit).

The image shows a Microsoft Excel spreadsheet with a scatter plot. The x-axis is labeled 'SHOE_SIZE' and the y-axis is labeled 'HEIGHT'. The data points are blue dots. The chart is rotated 90 degrees counter-clockwise. A text box with the following text is overlaid on the chart:

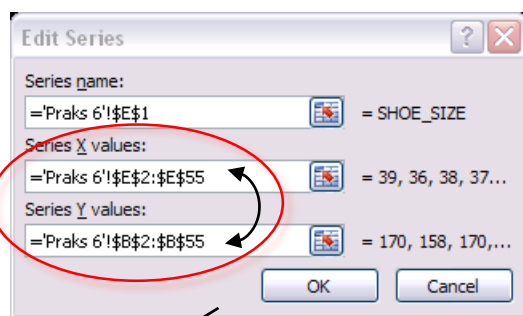
Teljed on valepidi!
 Prognosimaks pikkust jalanumbri
 alusel, peab pikkus olema y-teljel.
 Telgede vahetamiseks ...

The 'CHART TOOLS' ribbon is open, and the 'SELECT DATA SOURCE' dialog box is shown. The 'Chart data range' is set to '=Praks 6!\$B\$1:\$B\$55,Praks 6!\$E\$1:\$E\$55'. The 'Legend Entries (Series)' list contains 'SHOE_SIZE'. The 'Horizontal (Category) Axis Labels' list contains the values 170, 158, 170, 170, 179.

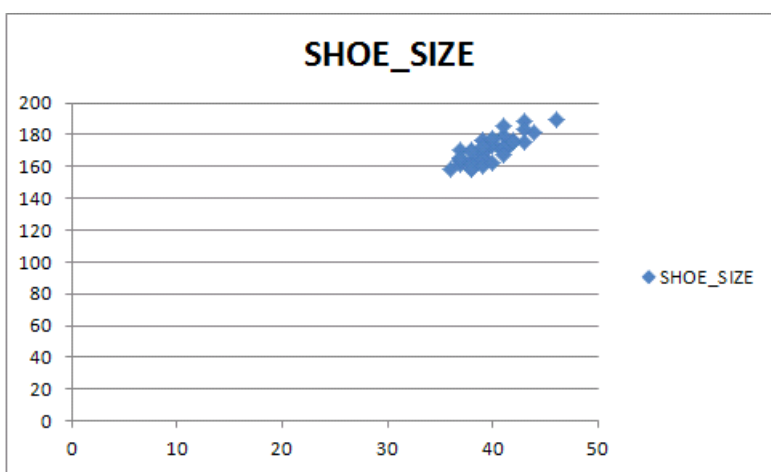
	A	B	C	D	E	F	G	H	I	J	K
	GENDER	HEIGHT	WEIGHT	HEAD	SHOE_SIZE	MATH	BREAKFAST	PORRIDGE	PET	SICK	SPORT
1											
2	W	170	70	55.5	39	3	other	yes	yes	no	yes
3	W	158	47.5	55	36	3	cereals or muesli	yes	yes	no	yes
4	W	170	60	53	38	5	cereals or muesli	yes	yes	no	yes
5	W	170	50	55	37	4	sandwich	yes	yes	no	yes
6	W	179	68	58	41	5	cereals or muesli	yes	yes	no	yes
7	W	163	56	55	37	4	sandwich	yes	yes	no	yes
8	W	177	65	55	40	3	sandwich	sometime	yes	yes	yes
9	W	162.5	53	55	38	3	porridge	yes	yes	no	yes
10	W	170	75	56	39	5	other	yes	yes	no	no
11	M	175	74	57	42	3	sandwich	y			
12	W	176	66	57	39	4	sandwich	s			
13	M	175	64	56	42	4	other	y			
14	M	190	82	58	46	4	other	y			
15	W	161	50	55	37	4	nothing	r			
16	W	170	85	57	41	4	cereals or muesli	r			
17	W	176	58	52	39	5	cereals or muesli	y			
18	W	172	90	58	41	4	porridge	y			
19	W	158	55	57	38	4	cereals or muesli	y			
20	M	189	82		43	4	cereals or muesli	r			
21	W	169	60	55.5	41	4	cereals or muesli	y			
22	W	164	52	56	37	4	other	s			
23	W	172	62	56	39	4	sandwich	s			
24	W	173	66	56	40	5	cereals or muesli	y			
25	W	169	60	55	39	3	other	y			
26	W	162	50	50	38	3	porridge	y			
27	W	165	52	50.5	37	4	sandwich	y			
28	M	170	80	56	41	4	cereals or muesli	r			
29	M	176	74	56	42	5	porridge	yes	yes	yes	yes
30	M	175	73	54	43	4	other	sometime	yes	yes	no
31	W	171	63	57	39	5	cereals or muesli	yes	yes	no	no
32	W	170	60	53	39	5	other	no	yes		
33	W	163	62	55	38	5	cereals or muesli	yes	no		
34	M	181	74	55	44	4	sandwich	yes	yes		
35	W	168	60	55	39	4	cereals or muesli	yes	yes		
36	W	174	54	55	40	5	cereals or muesli	yes	yes		
37	W	166	68	56	39	3	other	no	no		
38	W	168	63	53	39	4	sandwich	yes	yes		
39	W	165	58	56	37	5	sandwich	no	yes		
40	W	171	75	55	41	4	sandwich	yes	yes	no	yes
41	W	165	77	58	39	5	sandwich	yes	yes	yes	yes
42	W	161	55	57	38	4	sandwich	yes	yes	yes	yes
43	M	183	75		43						
44	W	169	53	55	38						
45	W	175	60	57	42						
46	W	167	80	57.5	41						
47	W	158	70	55	38						
48	M	174	87	57	40						
49	W	165	61	57	39						
50	W	164	58	57	39						
51	W	185	80	60	41						
52	W	177	63	60	40						
53	W	160	70	57	39						
54	W	162	70	55	40						
55	W	172	58	62	39						



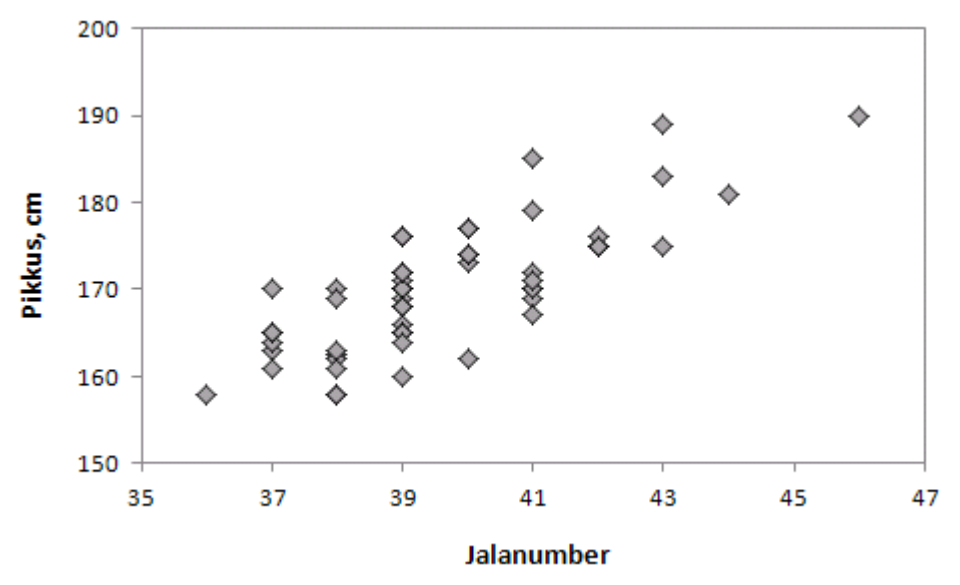
Vahetage x- ja y-telje väärtused



Tulemus:
pikkuse väärtused on y- ja
jalanumbri väärtused x-teljel.



Kujundage joonis!



2. Prognosimaks pikkust jalanumbri alusel, lisage tunnuste 'HEIGHT' ja 'SHOE_SIZE' hajuvusdiagrammile **regressioonisirge**.

Lisage joonisele ka **regressioonivõrrand** ja viimase alusel leitavate prognooside täpsust kirjeldav **determinatsioonikordaja R^2** .

Kaks variant lineaarse trendijooni lisamiseks

Regressiooni-
võrrandi ja
 R^2 lisamiseks

TRENDLINE OPTIONS

- Exponential
- Linear
- Logarithmic
- Polynomial Order
- Power
- Moving Average Period

Trendline Name

- Automatic Linear (y = 2.8868x + 55.736)
- Custom

Forecast

Forward: 0.0

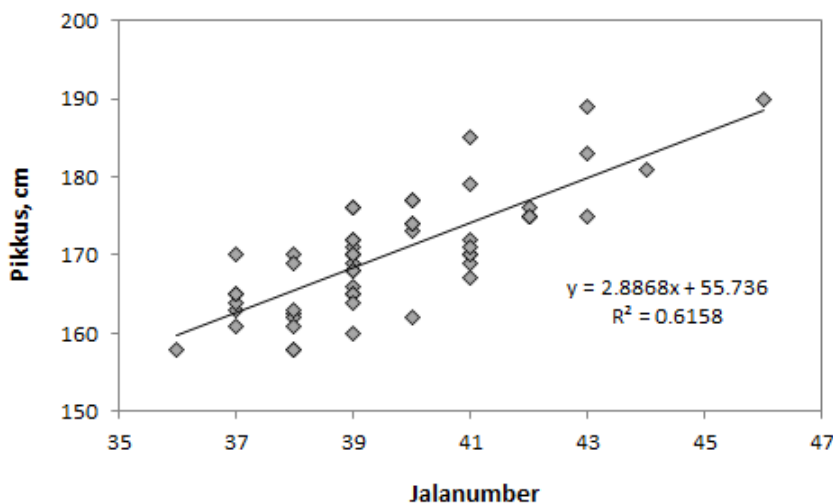
Backward: 0.0

Set Intercept

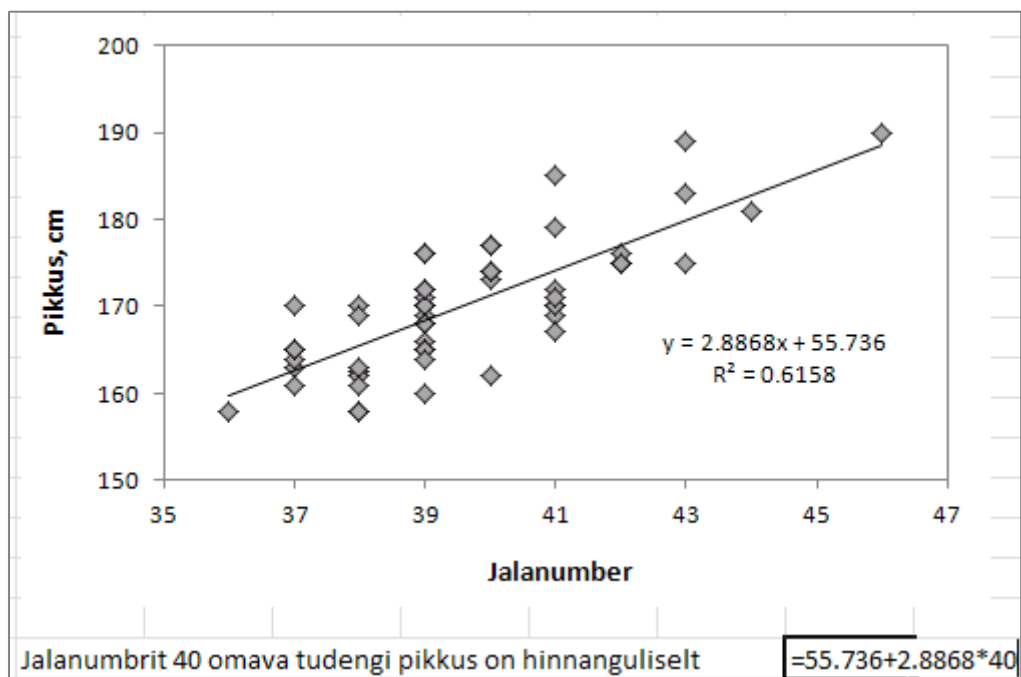
Display Equation on chart

Display R-squared value on chart

Tulemus:



3. Prognoosige leitud võrrandi alusel, keskmiselt kui pikk on jalanumbrit 40 omav tudeng. Selleks pange joonise alla kirja Exceli poolt välja arvutatud regressioonivõrrand, asendades lihtsalt suuruse x arvuga 40. ☺



Ülesande 2 tööjuhend.

1. Teostage statistikaprotseduuri Regression (Data-sakk -> Data analysis...) abil lineaarne regressioonanalüüs prognoosimaks tudengite pikkust jalanumbri alusel.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
1	GENDER	HEIGHT	WEIGHT	HEAD	SHOE_S	MATH	SMOKE																			
2	W	170	70	55.5	39	3	no																			
3	W	158	47.5	55	36	3	no																			
4	W	170	60	53	38	5	no																			
5	W	170	50	55	37	4	no																			
6	W	179	68	58	41	5	no																			
7	W	163	56	55	37	4	no																			
8	W	177	65	55	40	3	no																			
9	W	162.5	53	55	38	3	no																			
10	W	170	75	56	39	5	no																			
11	M	175	74	57	42	3	no																			
12	W	176	66	57	39	4	no																			
13	M	175	64	56	42	4	no anymore, b																			
14	M	190	82	58	46	4	no																			
15	W	161	50	55	37	4	no																			
16	W	170	85	57	41	4	no																			
17	W	176	58	52	39	5	no																			
18	W	172	90	58	41	4	no																			
19	W	158	55	57	38	4	yes																			
20	M	189	82	43	43	4	yes																			
21	W	169	60	55.5	41	4	yes																			
22	W	164	52	56	37	4	no																			
23	W	172	62	56	39	4	no																			
24	W	173	66	56	40	5	no																			
25	W	169	60	55	39	3	no																			
26	W	162	50	50	38																					
27	W	165	52	50.5	37																					
28	M	170	80	56	41																					
29	M	176	74	56	42																					
30	M	175	73	54	43																					
31	W	171	63	57	39																					
32	W	170	60	53	39																					
33	W	163	62	55	38																					
34	M	181	74	55	44																					
35	W	168	60	55	39																					
36	W	174	54	55	40																					
37	W	166	68	56	39																					
38	W	168	63	53	39																					
39	W	165	58	56	37																					
40	W	171	75	55	41																					
41	W	165	77	58	39																					
42	W	161	55	57	38																					
43	M	183	75	55	43																					
44	W	169	53	55	38																					
45	W	175	60	57	42																					
46	W	167	80	57.5	41																					
47	W	158	70	55	38																					
48	M	174	87	57	40																					
49	W	165	61	57	39																					
50	W	164	58	57	39																					
51	W	185	80	60	41																					
52	W	177	63	60	40																					
53	W	160	70	57	39																					
54	W	162	70	55	40																					
55	W	172	58	62	39																					

The Regression dialog box is configured as follows:

- Input Y Range: \$B\$1:\$B\$55
- Input X Range: \$E\$1:\$E\$55
- Labels
- Constant is Zero
- Confidence Level: 95 %
- Output Range: \$5\$21
- Residuals
- Standardized Residuals
- Residual Plots
- Line Fit Plots
- Normal Probability Plots

Regressioonanalüüsi tulemus:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.7847576					
R Square	0.6158445					
Adjusted R Square	0.6084569					
Standard Error	4.6287793					
Observations	54					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	1786.077252	1786.077	83.36184	2.18481E-12	
Residual	52	1114.131081	21.4256			
Total	53	2900.208333				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	55.73624	12.55193924	4.440448	4.71E-05	30.54893075	80.923548
SHOE_SIZE	2.8868487	0.316184355	9.130271	2.18E-12	2.252378374	3.521319

2. Kirjutage protseduuri tulemuste põhjal välja lineaarne regressioonivõrrand (ehk regressioonimudel) kujul

$$Pikkus = a + b \times \text{Jalanumber},$$

kus a ja b asemel on Exceli poolt välja arvatud kordajate väärtused.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	55.73624	12.55193924	4.440448	4.71E-05	30.54893075	80.923548
SHOE_SIZE	2.8868487	0.316184355	9.130271	2.18E-12	2.252378374	3.521319

3. Kui suur on keskmiselt pikkuste vaheline erinevus tudengitel, kelle jalanumbrid erinevad 2 võrra?

Vastus: $2 \times b$ (aga arvuliselt?). Pange arvuline vastus kirja täislausega.

4. Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1786.077252	1786.077	83.36184	2.18481E-12
Residual	52	1114.131081	21.4256		
Total	53	2900.208333			

= *p*

Märkus. Regressioonivõrrandi statistiline olulisus tähendab seda, et leitud regressioonivõrrand kujul

$$Pikkus = a + b \times \text{Jalanumber}$$

võimaldab pikkust täpsemalt prognoosida võrreldes konstantse võrrandiga

$$Pikkus = a.$$

Ehk siis, statistiliselt olulise regressioonivõrrandi korral võimaldab jalanumbri arvestamine pikkust täpsemalt prognoosida võrreldes konstateeringuga, et kõigi tudengite pikkused on ühesugused (ja võrdsed suurusega *a*).

Hüpoteeside paar, mille testimiseks vajaliku *p*-väärtuse väljastab *Excel* tabelisse nimega ANOVA, on kujul:

H_0 : regressioonivõrrand ei ole statistiliselt oluline

H_1 : regressioonivõrrand on statistiliselt oluline

ehk

H_0 : leitud võrrand ei ole parem võrreldes konstantse võrrandiga

H_1 : leitud võrrand on parem võrreldes konstantse võrrandiga

ehk matemaatilisel

H_0 : $Pikkus = a$

H_1 : $Pikkus = a + b \times \text{Jalanumber}$

Praktikas rakendada on põhjust vaid statistiliselt olulist regressioonivõrrandit.

5. Sõnastage üks lause regressioonivõrrandist saadavate prognooside täpsuse kohta kas mitmese korrelatsioonikordaja (*R*), determinatsioonikordaja (R^2) või mudeli standardvea baasil.

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.7847576
R Square	0.6158445
Adjusted R Square	0.6084569
Standard Error	4.6287793
Observations	54

Mitmene korrelatsioonikordaja *R* mõeldab uuritava tunnuse ja tema prognoositud väärtuste vahelist korrelatsiooni. Mida suurem, seda parem!

Determinatsioonikordaja R^2 näitab, kui suure osa uuritava tunnuse varieeruvusest võrrandist saadud prognoosid ära kirjeldavad, $0 \leq R^2 \leq 1$. Esitatakse enamasti protsentides. Mida suurem, seda parem!

Mudeli standardviga *SE* on prognoosijääkide standardhälve. Näitab tegelike ja prognoositud väärtuste vahelist keskmist erinevust (mudeli keskmist viga). Mida väiksem, seda parem!

Antud juhul saaks seega järeldada, et prognoosides tudengi pikkust tema jalanumbri alusel, erineb prognoositud pikkus tegelikust keskmiselt 4,6 cm võrra. Samas on seos prognoositud ja tegelike pikkuste vahel tugev (mitmese korrelatsioonikordaja $R = 0,78$) ning pikkuste tegelikust varieeruvusest on leitud regressioonivõrrandi alusel ära kirjeldatav 62% ($R^2 = 0,62$).