

Biomeetria praks 6

Illustreeritud (mittetäielik) tööjuhend

Eeltöö

1. Avage *MS Excel*'is oma kursuse 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 'PIKKUS' ja 'JALANR' 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 *Excel*'i 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

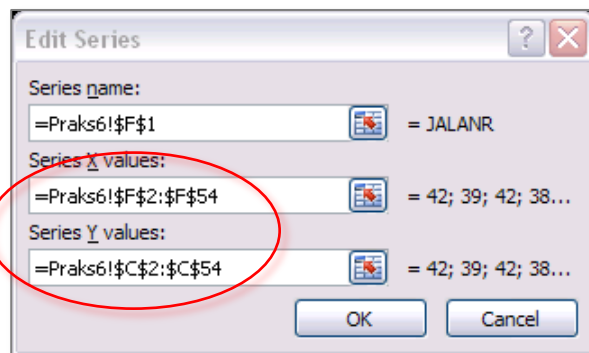
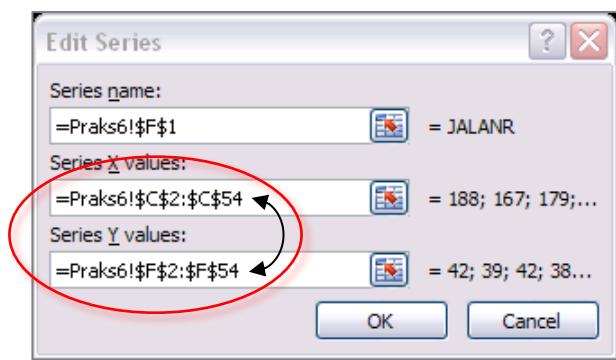
1. Illustreerige tunnuste 'PIKKUS' ja 'JALANR' vahelist seost hajuvus- ehk punktdiagrammiga.

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

JALANR

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

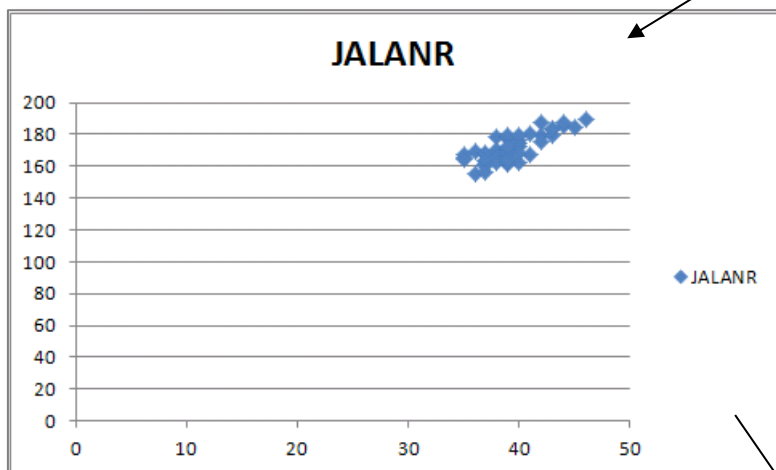
	A	B	C	D	E	F	G	H	I	J	K	L
	RIIK	SUGU	PIKKUS	MASS	PEA_P	JALANR	ODE_VENI	MAT_HINI	HOMMIK	PUDER	LEMMIK	HAIGE
2	Eesti	M	188	88	56	42	2	5	võleib	jah	jah	ei
3	Eesti	N	167	65	55	39	1	5	puder	jah	jah	ei
4	Eesti	N	179	68	54	42	2	4	võleib	jah	ei	ei
5	Eesti	N	178	60	55	38	2	4	puder	nii ja naa	jah	ei
6	Eesti	N	161	70	56,5	39	1	3	võleib	jah	ei	ei
7	Eesti	M	185	69	56	44	1	4	muu	jah	jah	ei
8	Soome	N	174	58	58	40	10	4	puder	jah	jah	ei
9	Eesti	N	171	55		39	2	4	võleib	jah	jah	ei
10	Eesti	N	165	55		35	0	3	puder	nii ja naa	jah	ei
11	Soome	N	167	65	58,5	38	1	5	puder			
12	Soome	N	163	69	58	38	2	5	muu			
13	Soome	N	165	55	58	38	1	4	võleib			
14	Soome	N	169	61	58	36	2	4	muu			
15	Soome	N	170	68	58	38	3	4	võleib			
16	Eesti	N	165	51	56	39	1	4	helbed			
17	Eesti	N	170	57	56	39	1	4	võleib			
18	Eesti	N	170	63	56	39	1	4	võleib			
19	Soome	N	162	65	58	39	1	5	võleib			
20	Eesti	N	168	80	56	37	2	4	puder			
21	Eesti	N	177	60		39	2	4	ei sööt			
22	Eesti	N	168	73	53	39	2	4	võleib			
23	Eesti	N	165	50,1	52	38	3	5	puder			
24	Eesti	M	188	80	58	44	2	4	puder			
25	Eesti	N	162	63	56	40	1	5	helbed			
26	Soome	N	162	52	55	39	3	5	puder			
27	Eesti	M	183	73	54	43	0	4	muu			
28	Soome	N	161	47	62	37	0	4	võleib			
29			163	53	58	37	2	3	puder			
30	Soome	N	179	80	62	43	3	4	võleib	nii ja naa	jah	ei
31	Soome	N	179	68	57	39	1	4	puder	jah	jah	ei
32	Soome	N	176	80	55	40						
33	Soome	N	155	52	53	36						
34	Eesti	N	162	57	53	38						
35	Eesti	M	190	85	58	46						
36	Eesti	M	184	78	55	45						
37	Eesti	N	167	66	54	41						
38	Eesti	M	175	62	56,5	42						
39	Eesti	N	180	71	56	41						
40	Eesti	N	164	52	56	39	2	5	võleib	nii ja naa	jah	ei
41	Eesti	M	183	80	56	43	1	3	võleib	ei	jah	ei
42	Eesti	N	179	75	56	40	4	5	ei söötavalise	jah	ei	jah
43	Venemaa	N	167	46,5	55	37	2	4	ei söötavalise	ei	jah	ei
44	Eesti	N	172	64	55	40						
45	Soome	N	167	55	50	35						
46	Soome	N	171	75	56	39						
47	Soome	N	160	61	54	37						
48	Soome	N	168	75	58	40						
49	Soome	N	163	50	54	37						
50		N	169	65	57	39						
51	Soome	N	172	54	55	39						
52	Eesti	N	163	52	48	39						
53	Eesti	N	164	55	55,5	35						
54	Soome	N	156	48	50	37						



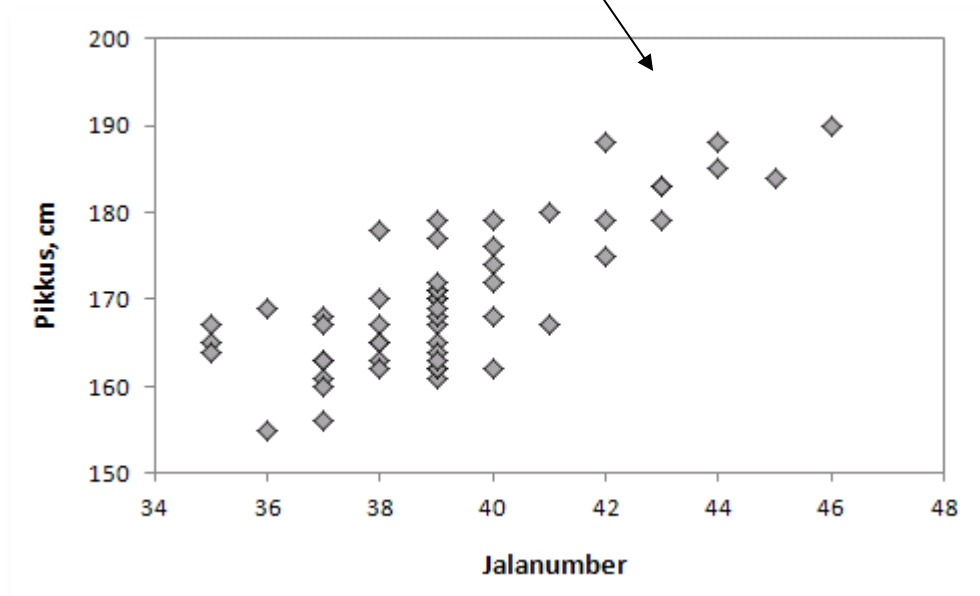
Vahetage x- ja y-telje väärtused

Tulemus:

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



Kujundage joonis!



2. Prognoosimaks pikkust jalanumbri alusel, lisage tunnuste 'PIKKUS' ja 'JALANR' hajuvusdiagrammile **regressioonisirge**.

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

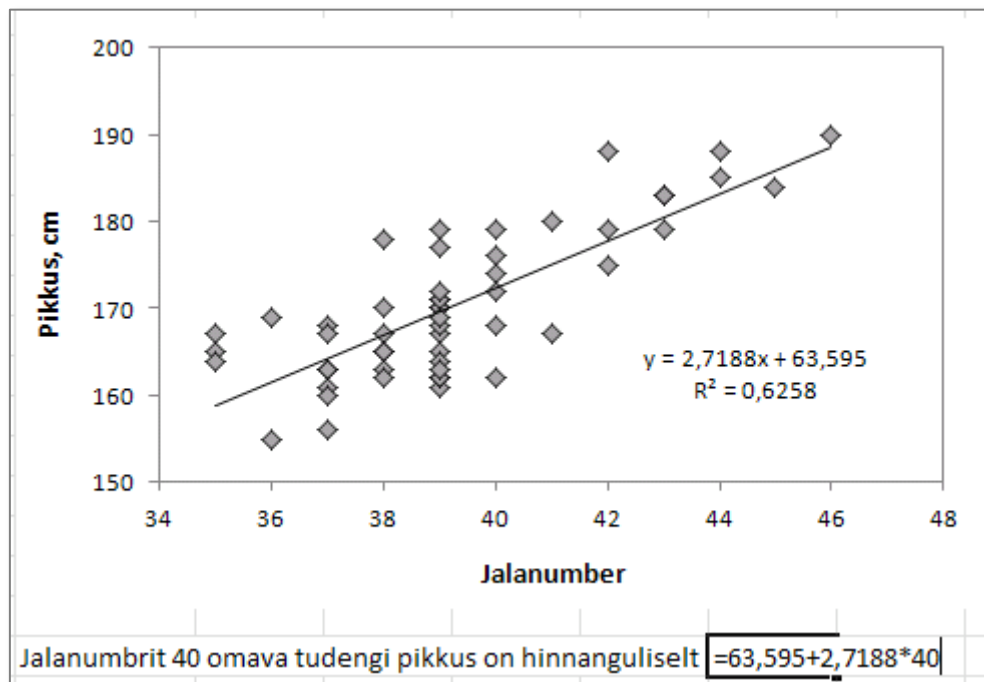
The image shows a step-by-step process in Microsoft Excel to add a linear regression line to a scatter plot. The chart displays 'Pikkus, cm' (Height, cm) on the vertical axis and 'Jalanumber' (Shoe number) on the horizontal axis. A linear trendline is applied to the data points. The 'Format Trendline' task pane is open, showing 'Trendline Options' with 'Linear' selected. The 'Display Equation on chart' and 'Display R-squared value on chart' options are checked. The 'Trendline Name' is set to 'Linear (JALANR)'. The 'Forecast' section shows 'Forward' and 'Backward' values of 0,0 periods. The 'Set Intercept' is set to 0,0. The 'Display Equation on chart' and 'Display R-squared value on chart' options are checked.

Lineaarse trendijoonelise lisamiseks

Lisavalikute tarvis

Märkige, saamiseks joonisele regressioonivõrrandi ja R^2 väärtust

3. Prognoosige leitud võrrandi alusel, keskmiselt kui pikk on jalanumbrit 40 omav tudeng. Selleks pange joonise alla kirja *Excel*'i 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 displays the Microsoft Excel interface with the 'Data' tab selected. The 'Data Analysis' toolpak is visible in the ribbon. The 'Data Analysis' dialog box is open, showing the 'Regression' option selected. The 'Regression' dialog box is also open, showing the following configuration:

- Input Y Range:** \$C\$1:\$C\$54
- Input X Range:** \$F\$1:\$F\$54
- Labels**
- Constant is Zero**
- Confidence Level:** 95 %
- Output options:**
 - Output Range:** \$U\$21
 - New Worksheet Ply:
 - New Workbook
- Residuals:**
 - Residuals
 - Standardized Residuals
 - Residual Plots
 - Line Fit Plots
- Normal Probability:**
 - Normal Probability Plots

The spreadsheet data is as follows:

	A	B	C	D	E	F	G	K
	RIIK	SUGU	PIKKUS	MASS	PEA_P	JALANR	ODE_VENI	KINDO
2	Eesti	M	188	88	56	42	2	viimase 10 päeva jooksul
3	Eesti	N	167	65	55	39	1	viimase kuu jooksul
4	Eesti	N	179	68	54	42	2	viimase aasta jooksul
5	Eesti	N	178	60	55	38	2	viimase kuu jooksul
6	Eesti	N	161	70	56,5	39	1	viimase 10 päeva jooksul
7	Eesti	M	185	69	56	44	1	rohkem kui aasta tagasi
8	Soome	N	174	58	58	40	10	rohkem kui aasta tagasi
9	Eesti	N	171	55	55	39	2	rohkem kui aasta tagasi
10	Eesti	N	165	55	55	35	0	viimase kuu jooksul
11	Soome	N	167	65	58,5	38	1	viimase 10 päeva jooksul
12	Soome	N	163	69	58	38	2	viimase 10 päeva jooksul
13	Soome	N	165	55	58	38	1	viimase kuu jooksul
14	Soome	N	169	61	58	36	2	viimase aasta jooksul
15	Soome	N	170	68	58	38	3	viimase aasta jooksul
16	Eesti	N	165	51	56	39	1	viimase aasta jooksul
17	Eesti	N	170	57	56	39	1	viimase 10 päeva jooksul
18	Eesti	N	170	63	56	39	1	viimase 10 päeva jooksul
19	Soome	N	162	65	58	39	1	viimase 10 päeva jooksul
20	Eesti	N	168	80	56	37	2	viimase 10 päeva jooksul
21	Eesti	N	177	60	56	39	2	viimase aasta jooksul
22	Eesti	N	168	73	53	39	2	viimase kuu jooksul
23	Eesti	N	165	50,1	52	38	3	viimase kuu jooksul
24	Eesti	M	188	80	58	44	2	viimase kuu jooksul
25	Eesti	N	162	63	56	40	1	viimase kuu jooksul
26	Soome	N	162	52	55	39	3	viimase 10 päeva jooksul
27	Eesti	M	183	73	54	43	0	viimase kuu jooksul
28	Soome	N	161	47	62	37	0	viimase kuu jooksul
29	Eesti	N	165	53	58	37	2	viimase 10 päeva jooksul
30	Soome	N	179	80	62	43	3	viimase 10 päeva jooksul
31	Soome	N	179	68	57	39	1	viimase 10 päeva jooksul
32	Soome	N	176	80	55	40	5	viimase 10 päeva jooksul
33	Soome	N	155	52	53	36	2	viimase 10 päeva jooksul
34	Eesti	N	162	57	53	38	2	viimase 10 päeva jooksul
35	Eesti	M	190	85	58	46	0	viimase 10 päeva jooksul
36	Eesti	M	184	78	55	45	2	viimase 10 päeva jooksul
37	Eesti	N	167	66	54	41	2	viimase 10 päeva jooksul
38	Eesti	M	175	62	56,5	42	2	viimase 10 päeva jooksul
39	Eesti	N	180	71	56	41	1	viimase 10 päeva jooksul
40	Eesti	N	164	52	56	39	2	viimase 10 päeva jooksul
41	Eesti	M	183	80	56	43	2	viimase 10 päeva jooksul
42	Eesti	N	179	75	56	40	4	viimase 10 päeva jooksul
43	Venemaa	N	167	46,5	55	37	2	viimase 10 päeva jooksul
44	Eesti	N	172	64	55	40	1	viimase 10 päeva jooksul
45	Soome	N	167	55	50	35	2	viimase 10 päeva jooksul
46	Soome	N	171	75	56	39	1	viimase 10 päeva jooksul
47	Soome	N	160	61	54	37	0	viimase 10 päeva jooksul
48	Soome	N	168	75	58	40	4	viimase 10 päeva jooksul
49	Soome	N	163	50	54	37	0	viimase 10 päeva jooksul
50	Soome	N	169	65	57	39	1	viimase 10 päeva jooksul
51	Soome	N	172	54	55	39	2	viimase 10 päeva jooksul
52	Eesti	N	163	52	48	39	1	viimase 10 päeva jooksul
53	Eesti	N	164	55	55,5	35	1	viimase 10 päeva jooksul
54	Soome	N	156	48	50	37	1	viimase 10 päeva jooksul

Regressioonanalüüsi tulemus:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0,7910688					
R Square	0,6257898					
Adjusted R Square	0,6184523					
Standard Error	5,2725009					
Observations	53					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	2370,916716	2370,917	85,28703	1,8092E-12	
Residual	51	1417,762529	27,79927			
Total	52	3788,679245				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	63,594941	11,58737161	5,488297	1,28E-06	40,33232199	86,85756
JALANR	2,7187647	0,294394799	9,235098	1,81E-12	2,127742485	3,3097869

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

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

kus a ja b asemel on *Excel*'i 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	63,594941	11,58737161	5,488297	1,28E-06	40,33232199	86,85756
JALANR	2,7187647	0,294394799	9,235098	1,81E-12	2,127742485	3,3097869

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	2370,916716	2370,917	85,28703	1,8092E-12 = <i>p</i>
Residual	51	1417,762529	27,79927		
Total	52	3788,679245			

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

Reaalselt 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	0,7910688
R Square	0,6257898
Adjusted R Square	0,6184523
Standard Error	5,2725009
Observations	53

Mitmene korrelatsioonikordaja *R* mõõdab 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 5,3 cm võrra. Samas on seos prognoositud ja tegelike pikkuste vahel tugev (mitmese korrelatsioonikordaja $R = 0,79$) ning pikkuste tegelikust varieeruvusest on leitud regressioonivõrrandi alusel ära kirjeldatav 63%.