

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.
-

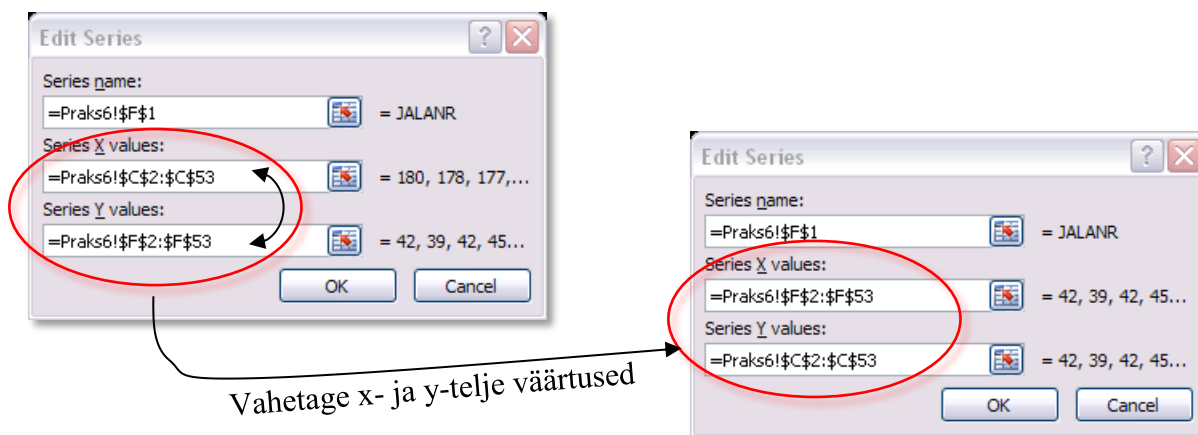
Ülesande 1 tööjuhend

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).

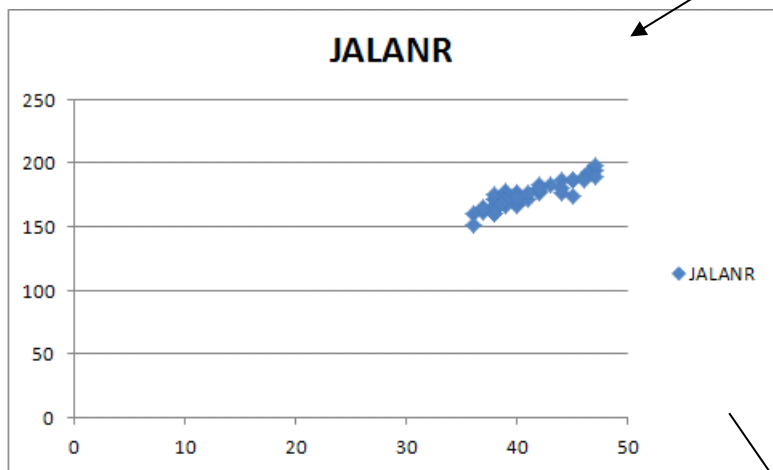
The image shows a Microsoft Excel spreadsheet with a scatter plot. The spreadsheet has columns for 'RIIK', 'SUGU', 'PIKKUS', 'MASS', 'PEA_P', 'JALANR', and various other variables. The scatter plot shows a positive correlation between 'PIKKUS' (height) on the y-axis and 'JALANR' (shoe size) on the x-axis. The plot is titled 'JALANR'. A text box with an arrow pointing to the plot contains the text: "Teljed on valepidi! Prognoosimaks pikkust jalanumbri alusel, peab pikkus olema y-teljel. Telgede vahetamiseks ...". Below the plot, the 'Select Data Source' dialog box is open, showing the chart data range as '=Praks6!\$C\$1:\$C\$53,Praks6!\$F\$1:\$F\$53'. The legend entries list 'JALANR'. The horizontal axis labels are 180, 178, 177, 187, 186. The 'Switch Row/Column' button is highlighted.

	A	B	C	D	E	F	G	H	I	J	K	L
1	RIIK	SUGU	PIKKUS	MASS	PEA_P	JALANR	ODE_VENI	MAT_HINI	HOMMIK	PUDER	LEMMIK	HAIGE
2	Eesti	N	180	76	56	42	2	3	muu	jah	ei	ei
3	Eesti	N	178	65	56	39	4	4	ei söö tava	jah	jah	ei
4	Eesti	M	177	70	57	42	3	3	võleib	nii ja naa	ei	jah
5	Eesti	M	187	75	45	45	3	3	ei söö tava	jah	jah	ei
6	Eesti	M	186	74	43	44	1	3	helbed või	jah	jah	ei
7	Eesti	N	165	62	42	37	3	3	puder	jah	ei	jah
8	Eesti	N	170	67	55,5	40	4	3	puder	jah	jah	jah
9	Eesti	N	177	59	42	39	1	4	võleib	nii ja naa	ei	ei
10	Eesti	N	166	47	55	38	2	4	võleib	jah	jah	ei
11	Eesti	N	165	55	42	38	1	3	võleib			
12	Eesti	M	180	68	51	44	0	3	võleib			
13	Eesti	N	161	49	56	37	2	3	võleib			
14	Eesti	N	168	54	50	40	0	5	võleib			
15	Eesti	N	167	67	56	40	2	3	võleib			
16	Eesti	N	160	50	55	38	0	4	võleib			
17	Eesti	N	164	53	59	38	1	4	võleib			
18	Eesti	M	194	105	57	47	4	4	ei söö ta			
19	Eesti	M	178	65	53	42	3	3	võleib			
20	Eesti	M	177	90	57	44	1	3	muu			
21	Eesti	N	171	57	50	38	2	4	puder			
22	Eesti	M	187	99	58	45	1	3	ei söö ta			
23	Eesti	M	189	81	54	46	1	3	ei söö ta			
24	Eesti	M	186	98	56	45	0	3	võleib			
25	Eesti	N	171	65	55	38	1	3	võleib			
26	Eesti	M	183	110	57	43	0	4	ei söö ta			
27	Eesti	M	193	100	58	46,5	1	3	puder			
28	Eesti	N	183	79	54	42	4	5	muu			
29	Eesti	N	177	75	55	40	1	5	helbed v			
30	Eesti	N	164	54	52	37	1	4	puder	nii ja naa	ei	jah
31	Eesti	N	170	60	55	39						
32	Eesti	N	175	65	57	38						
33	Eesti	N	168	69	55	39						
34	Eesti	M	186	94	58	46						
35	Eesti	N	160	55	55	36						
36	Eesti	N	175	69	53	39						
37	Eesti	N	151	53	52	36						
38	Eesti	M	198	110	60	47						
39	Eesti	M	174	120	56	45						
40	Eesti	N	180	80	58	42						
41	Eesti	N	171	72	41	41						
42	Eesti	N	169	80	57	40						
43	Eesti	N	160	64	54	38						
44	Eesti	N	170	65	55	39						
45	Eesti	N	176	58	54	40						
46	Eesti	N	167	55	54	39						
47	Eesti	N	160	59	56	38						
48	Eesti	N	172	80	54	39						
49	Eesti	N	170	60	39	39						
50	Eesti	N	176	57	56	41						
51	Eesti	N	168	62	56	38,5						
52	muu	M	189	78	57	47						
53	Eesti	N	173	68	45	40						

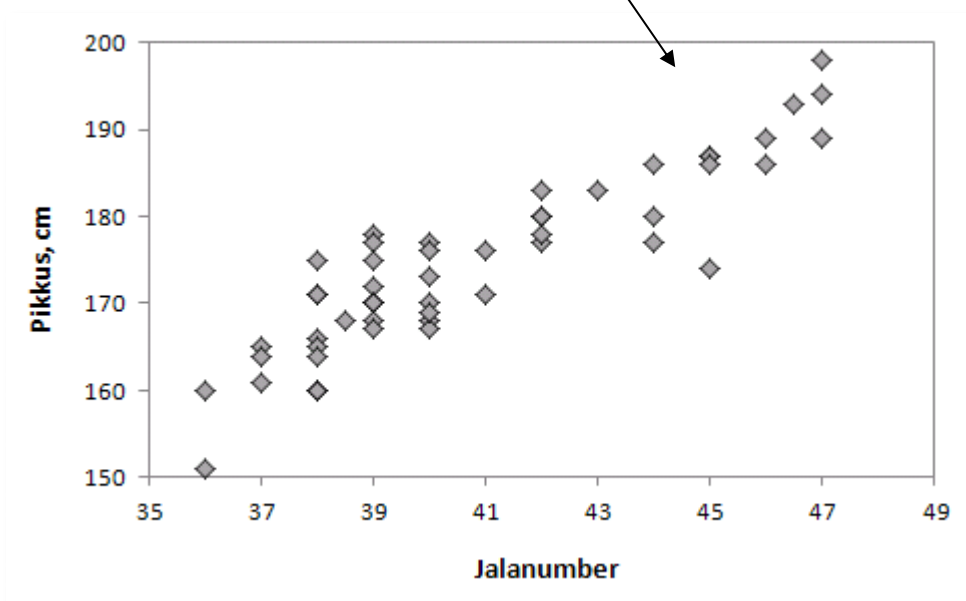


Tulemus:

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



Kujundage joonis!



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

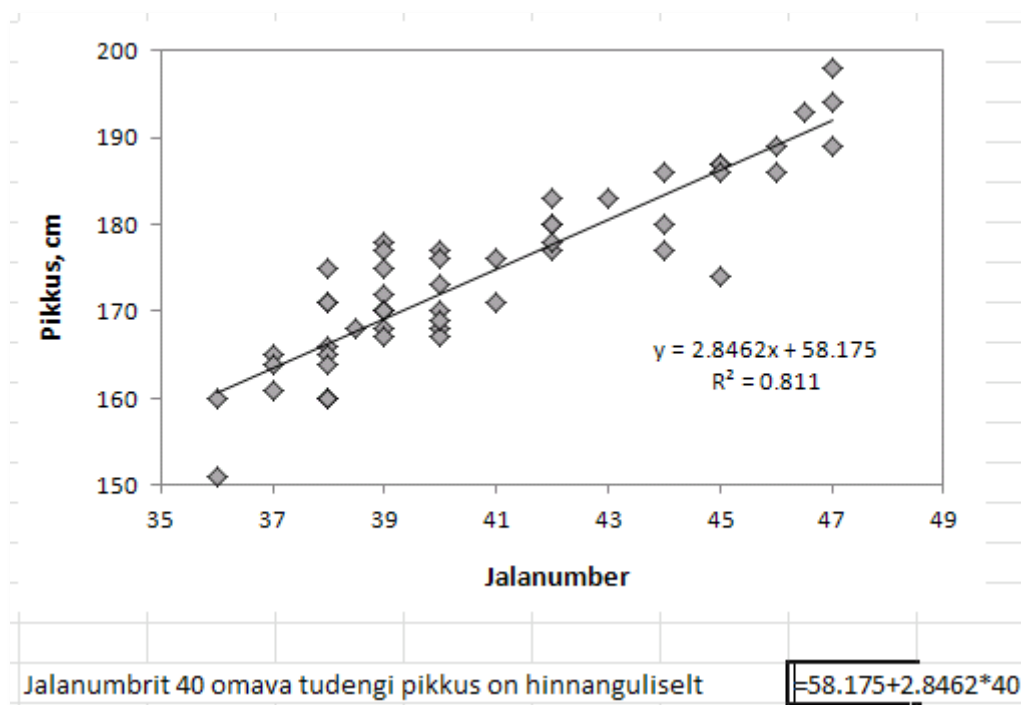
Lisage joonisele ka **regressioonivõrrand** ja viimase alusel leitavate prognooside täpsust kirjeldav **determinatsioonikordaja R^2** .

Lineaarse trendijoonelise lisamiseks

Lisavalikute tarvis

Märkige, saamaks joonisele regressioonivõrrandit 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 shows the Excel interface with the 'Data Analysis' task pane open on the right. The 'Data Analysis' dialog box is open, and 'Regression' is selected in the list. The 'Regression' dialog box is also open, showing the following settings:

- Input Y Range:** \$C\$1:\$C\$53
- Input X Range:** \$F\$1:\$F\$53
- Labels**
- Constant is Zero**
- Confidence Level:** 95 %
- Output options:** **Output Range:** \$U\$21
- New Worksheet Ply:**
- New Workbook**
- Residuals:** **Residuals**, **Residual Plots**, **Standardized Residuals**, **Line Fit Plots**
- Normal Probability:** **Normal Probability Plots**

The spreadsheet data is as follows:

	A	B	C	D	E	F	G	S	T	U	V	W	X	Y	Z	AA	AB
	RIIK	SUGU	PIKKUS	MASS	PEA_P	JALANR	ODE_VENI	KINO									
2	Eesti	N	180	76	56	42	2	viimase kuu jooksul									
3	Eesti	N	178	65	56	39	4	viimase aasta jooksul									
4	Eesti	M	177	70	57	42	3	viimase aasta jooksul									
5	Eesti	M	187	75	45	45	3	viimase aasta jooksul									
6	Eesti	M	186	74	43	44	1	viimase kuu jooksul									
7	Eesti	N	165	62	42	37	3	viimase kuu jooksul									
8	Eesti	N	170	67	55.5	40	4	viimase aasta jooksul									
9	Eesti	N	177	59	42	39	1	viimase kuu jooksul									
10	Eesti	N	166	47	55	38	2	viimase kuu jooksul									
11	Eesti	N	165	55	42	38	1	viimase kuu jooksul									
12	Eesti	M	180	68	51	44	0	rohkem kui aasta tagasi									
13	Eesti	N	161	49	56	37	2	viimase 10 päeva jooksul									
14	Eesti	N	168	54	50	40	0	rohkem kui aasta tagasi									
15	Eesti	N	167	67	56	40	2	viimase 10 päeva jooksul									
16	Eesti	N	160	50	55	38	0	viimase 10 päeva jooksul									
17	Eesti	N	164	53	59	38	1	viimase aasta jooksul									
18	Eesti	M	194	105	57	47	4	viimase kuu jooksul									
19	Eesti	M	178	65	53	42	3	rohkem kui aasta tagasi									
20	Eesti	M	177	90	57	44	1	viimase kuu jooksul									
21	Eesti	N	171	57	50	38	2	rohkem kui aasta tagasi									
22	Eesti	M	187	99	58	45	1	rohkem kui aasta tagasi									
23	Eesti	M	189	81	54	46	1	viimase aasta jooksul									
24	Eesti	M	186	98	56	45	0	viimase aasta jooksul									
25	Eesti	N	171	65	55	38	1	viimase 10 päeva jooksul									
26	Eesti	M	183	110	57	43	0	viimase aasta jooksul									
27	Eesti	M	193	100	58	46.5		viimase kuu jooksul									
28	Eesti	N	183	79	54	42											
29	Eesti	N	177	75	55	40											
30	Eesti	N	164	54	52	37											
31	Eesti	N	170	60	55	39											
32	Eesti	N	175	65	57	38											
33	Eesti	N	168	69	55	39											
34	Eesti	M	186	94	58	46											
35	Eesti	N	160	55	55	36											
36	Eesti	N	175	69	53	39											
37	Eesti	N	151	53	52	36											
38	Eesti	M	198	110	60	47											
39	Eesti	M	174	120	56	45											
40	Eesti	N	180	80	58	42											
41	Eesti	N	171	72		41											
42	Eesti	N	169	80	57	40											
43	Eesti	N	160	64	54	38											
44	Eesti	N	170	65	55	39											
45	Eesti	N	176	58	54	40											
46	Eesti	N	167	55	54	39											
47	Eesti	N	160	59	56	38											
48	Eesti	N	172	80	54	39											
49	Eesti	N	170	60		39											
50	Eesti	N	176	57	56	41											
51	Eesti	N	168	62	56	38.5											
52	muu	M	189	78	57	47											
53	Eesti	N	173	68	45	40											

Regressioonanalüüsi tulemus:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.90054285					
R Square	0.81097743					
Adjusted R Square	0.80719697					
Standard Error	4.40650328					
Observations	52					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	4165.367209	4165.367	214.5187	1.01554E-19	
Residual	50	970.8635599	19.41727			
Total	51	5136.230769				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	58.1753562	7.949933922	7.317716	1.91E-09	42.20744403	74.143268
JALANR	2.84624303	0.194329813	14.64646	1.02E-19	2.455920118	3.2365659

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 arvutatud 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	58.1753562	7.949933922	7.317716	1.91E-09	42.20744403	74.143268
JALANR	2.84624303	0.194329813	14.64646	1.02E-19	2.455920118	3.2365659

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	4165.367209	4165.367	214.5187	1.01554E-19
Residual	50	970.8635599	19.41727		
Total	51	5136.230769			

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	0.90054285
R Square	0.81097743
Adjusted R Square	0.80719697
Standard Error	4.40650328
Observations	52

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 4,4 cm võrra. Samas on seos prognoositud ja tegelike pikkuste vahel tugev (mitmese korrelatsioonikordaja $R = 0,90$) ning pikkuste tegelikust varieeruvusest on leitud regressioonivõrrandi alusel ära kirjeldatav 81%.