

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.
-

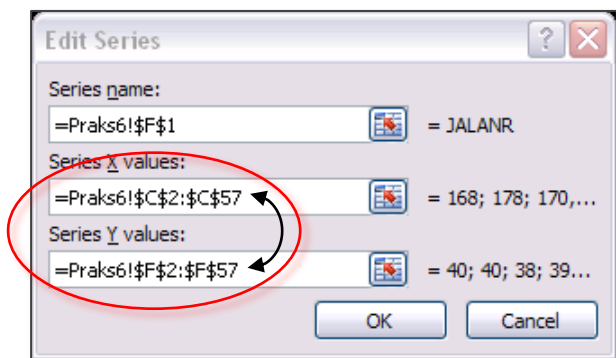
Ülesande 1 tööjuhend

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

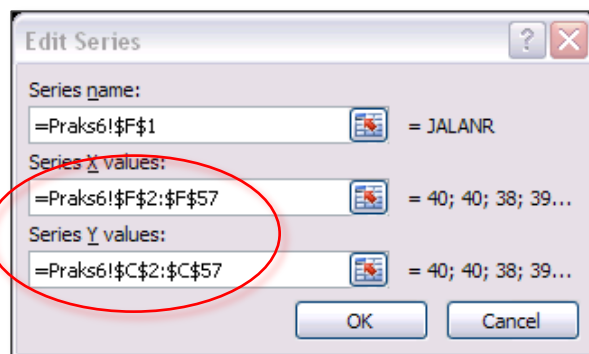
Joonisel peab x-telg vastama jalanumbritele ja y-telg pikkustele. Vajadusel kujundage joonis ümber (vt allpool toodud juhendit).

The image shows a Microsoft Excel spreadsheet with a scatter plot. The spreadsheet has columns for 'RIIK', 'SUGU', 'PIKKUS', 'MASS', 'PEA_P', 'JALANR', 'ODE_VEND', 'MAT_HINNE', 'HOMMIK', 'PUDER', and 'LEMMIK'. The scatter plot shows 'PIKKUS' on the y-axis and 'JALANR' on the x-axis. A text box points to the chart with the text: "Teljed on valepidi! Prognoosimaks pikkust jalanumbri alusel, peab pikkus olema y-teljel. Telgede vahetamiseks ...". Another text box points to the 'Select Data' button in the 'Chart Tools' ribbon. A 'Select Data Source' dialog box is open, showing the chart data range as '=Praks6!\$C\$1:\$C\$57;Praks6!\$F\$1:\$F\$57' and the legend entries as 'JALANR'. The horizontal axis labels are listed as 168, 178, 170,5, 158, and 163.

RIIK	SUGU	PIKKUS	MASS	PEA_P	JALANR
Eesti	N	168	62	53	40
Eesti	N	178	75	56	40
Eesti	N	170,5	64	54	38
Eesti	N	158	56	50	39
Eesti	N	163	60	54	38
Eesti	N	177	70	53	39
Eesti	N	170	51	52	40
Eesti	N	170	56	39	39
Soome	M	183	90	60	43
Soome	N	167	60	54	40
Soome	N	157	46	54	37
Soome	N	175	65	56	40
Soome	N	157	70	56	36
Soome	N	173	62	56	39
Soome	N	161	65	58	39
Eesti	N	169	68	55,5	39
Eesti	N	178	56	56	40
Eesti	N	164	62	55	37,5
Eesti	N	159	55	54	37
Eesti	N	164	55	54,5	37
Eesti	N	174	74	58	40
Eesti	N	171	67	66	38
Soome	N	160	56	57	38
Soome	N	170	76	56	39
Soome	M	187	75	55	44
Soome	N	165	55	56	37
Eesti	N	168	55	57	39
Eesti	N	159	48	52	34
Soome	N	166	58	57	34
Soome	N	169	59	58	34
Soome	N	160	57	58	34
Soome	N	178	70	53	44
Soome	N	169	68	58	44
Eesti	N	174	62	58	44
Eesti	N	165	59	55	37
Eesti	N	177	76	55	44
Soome	N	170	80	59	39
Eesti	N	173	75	55	40
Eesti	N	175,5	66	55	39
Eesti	N	172	66	56	40
Eesti	N	165	60	57	39
Eesti	N	170	60	58	38
Eesti	N	168	55	58	39
Soome	N	158	75	58	38
Eesti	N	170	54	53	39
Soome	N	181	72	57	42
Soome	N	167	70	51,5	40
Soome	N	166	49	55	38
Soome	N	174	51	55	38
Soome	N	160	60	57	38
Soome	N	170	65	59	39
Soome	N	170	66	58	38
Eesti	M	171	75	56	39
Eesti	M	177	69	55	43
Eesti	N	168	65	55	39
Eesti	N	172	62	53	39

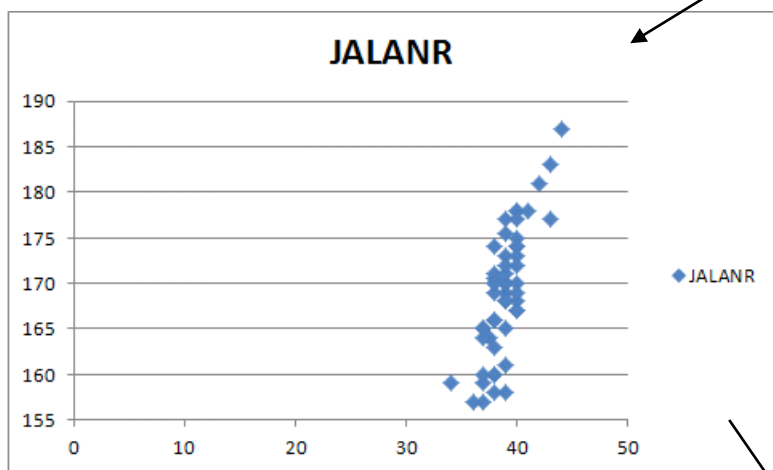


Vahetage x- ja y-telje väärtused

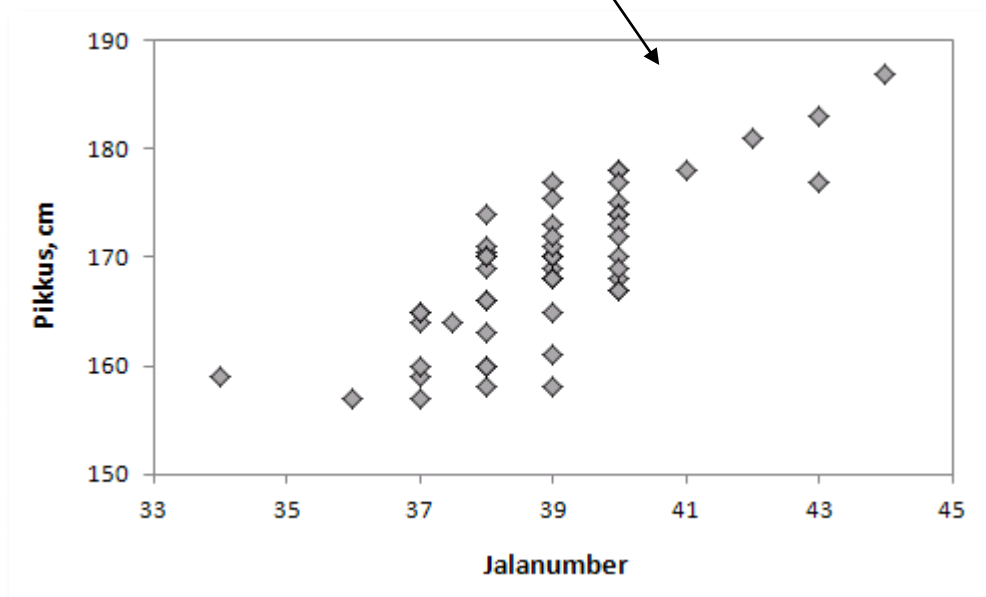


Tulemus:

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



Kujundage joonis!



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

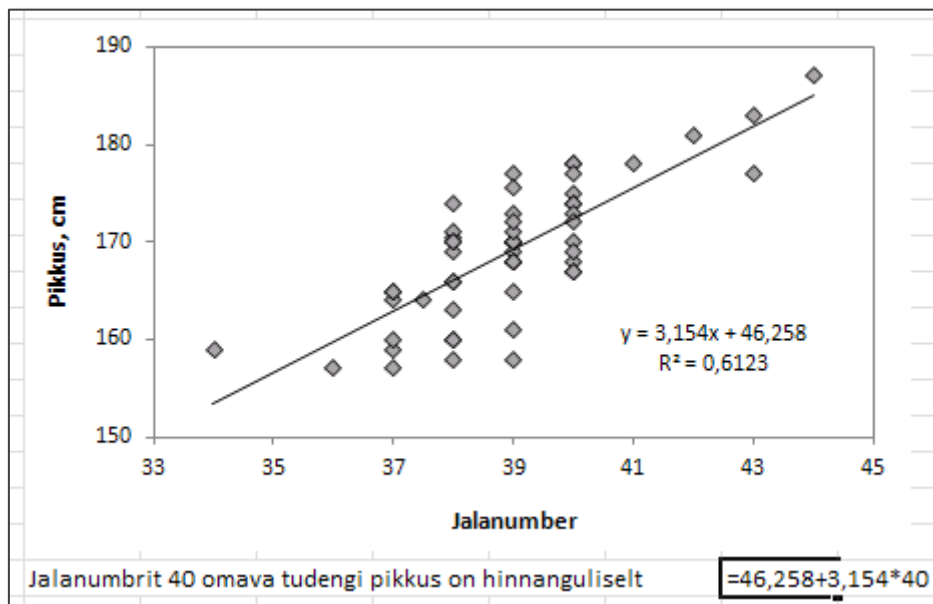
Lisage joonisele ka **regressioonivõrrand** ja viimase alusel leitavate prognooside täpsust kirjeldav **determinatsioonikordaja R^2** .

Lineaarse trendijoonse 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 following data in the spreadsheet:

A	B	C	D	E	F	G	S
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	ODE_VEND	KINO
2	N	168	62	53	40	1	viimase kuu jooksul
3	N	178	75	56	40	4	viimase kuu jooksul
4	N	170,5	64	54	38	2	viimase aasta jooksul
5	N	158	56	50	39	2	viimase 10 päeva jooksul
6	N	163	60	54	38	1	viimase aasta jooksul
7	N	177	70	53	39	3	viimase kuu jooksul
8	N	170	51	52	40	0	rohkem kui aasta tagasi
9	N	170	56	56	39	2	viimase aasta jooksul
10	M	183	90	60	43	2	viimase aasta jooksul
11	N	167	60	54	40	1	viimase aasta jooksul
12	N	157	46	54	37	2	viimase aasta jooksul
13	N	175	65	56	40	2	viimase aasta jooksul
14	N	157	70	56	36	2	viimase 10 päeva jooksul
15	N	173	62	56	39	1	viimase aasta jooksul
16	N	161	65	58	39	3	viimase aasta jooksul
17	N	169	68	55,5	39	2	viimase aasta jooksul
18	N	178	56	56	40	1	viimase aasta jooksul
19	N	164	62	55	37,5	1	viimase aasta jooksul
20	N	159	55	54	37	2	viimase kuu jooksul
21	N	164	55	54,5	37	1	viimase kuu jooksul
22	N	174	74	58	40	3	viimase 10 päeva jooksul
23	N	171	67	66	38	1	viimase 10 päeva jooksul
24	N	160	56	57	38	2	viimase aasta jooksul
25	N	170	76	56	39	2	viimase kuu jooksul
26	M	187	75	55	44	2	viimase aasta jooksul
27	N	165	55	56	37	4	viimase aasta jooksul
28	N	168	55	57	39	2	viimase aasta jooksul
29	N	159	48	52	34	1	viimase aasta jooksul
30	N	166	58	57	38	2	viimase aasta jooksul
31	N	169	59	58	38	1	viimase aasta jooksul
32	N	160	57	58	37	1	viimase aasta jooksul
33	N	178	70	53	41	2	viimase aasta jooksul
34	N	169	68	58	40	1	viimase aasta jooksul
35	N	174	62	58	40	1	viimase aasta jooksul
36	N	165	59	55	37	1	viimase aasta jooksul
37	N	177	76	55	40	3	viimase aasta jooksul
38	N	170	80	59	39	3	rohkem kui aasta tagasi
39	N	173	75	55	40	1	viimase aasta jooksul
40	N	175,5	66	55	39	1	viimase aasta jooksul
41	N	172	66	56	40	1	viimase aasta jooksul
42	N	165	60	57	39	2	viimase aasta jooksul
43	N	170	60	58	38	1	viimase aasta jooksul
44	N	168	55	58	39	0	viimase aasta jooksul
45	N	158	75	58	38	1	viimase aasta jooksul
46	N	170	54	53	39	1	viimase aasta jooksul
47	N	181	72	57	42	1	viimase aasta jooksul
48	N	167	70	51,5	40	1	viimase aasta jooksul
49	N	166	49	55	38	2	viimase aasta jooksul
50	N	174	51	55	38	3	viimase aasta jooksul
51	N	160	60	57	38	1	viimase aasta jooksul
52	N	170	65	59	39	2	viimase aasta jooksul
53	N	170	66	58	38	2	viimase aasta jooksul
54	M	171	75	56	39	1	viimase aasta jooksul
55	M	177	69	55	43	3	rohkem kui aasta tagasi
56	N	168	65	55	39	2	viimase aasta jooksul
57	N	172	62	53	39	5	rohkem kui aasta tagasi

Regressioonanalüüsi tulemus:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0,78247874					
R Square	0,61227298					
Adjusted R Square	0,60509285					
Standard Error	4,25422322					
Observations	56					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	1543,310577	1543,311	85,27324	1,06091E-12	
Residual	54	977,3144226	18,09842			
Total	55	2520,625				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	46,2582947	13,31753586	3,473488	0,00102	19,55824291	72,9583465
JALANR	3,15403873	0,341554986	9,234351	1,06E-12	2,469262215	3,83881525

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	46,2582947	13,31753586	3,473488	0,00102	19,55824291	72,9583465
JALANR	3,15403873	0,341554986	9,234351	1,06E-12	2,469262215	3,83881525

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

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

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	1543,310577	1543,311	85,27324	1,06091E-12 = <i>p</i>
Residual	54	977,3144226	18,09842		
Total	55	2520,625			

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*).

Hüpooteeside 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	0,78247874
R Square	0,61227298
Adjusted R Square	0,60509285
Standard Error	4,25422322
Observations	56

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!