

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) 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 ...

ankeet_2011k_jahendatud.xlsx - Microsoft Excel

Chart Tools: Design, Layout, Format

Switch Row/Column, Select Data

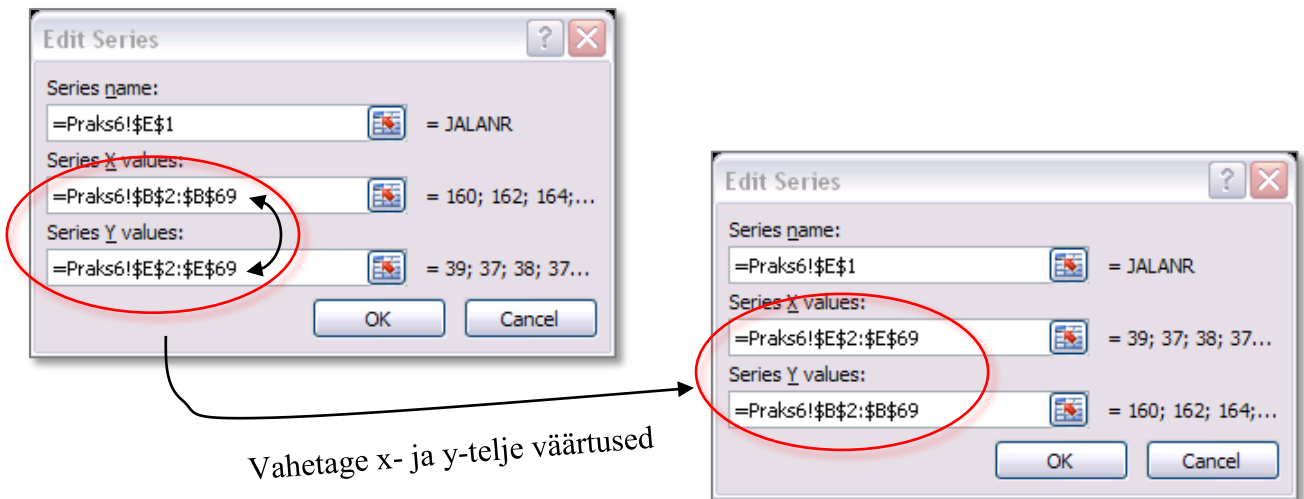
Select Data Source

Chart data range: =Praks6!\$B\$1:\$B\$69;Praks6!\$E\$1:\$E\$69

Legend Entries (Series): JALANR

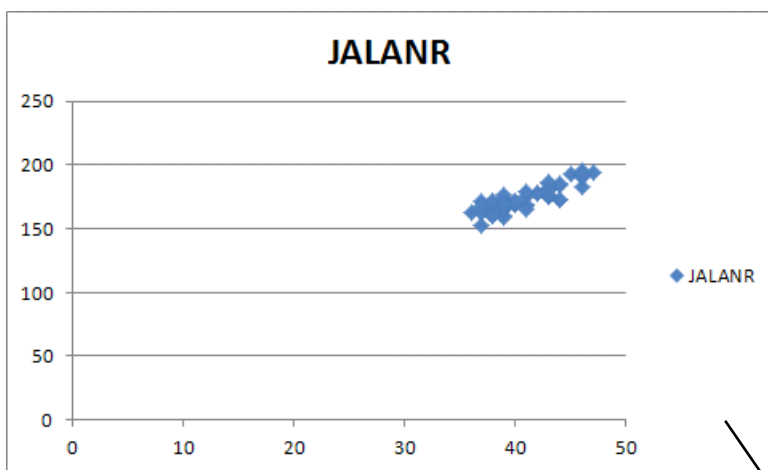
Horizontal (Category) Axis Labels: 160, 162, 164, 165, 165

Hidden and Empty Cells, OK, Cancel

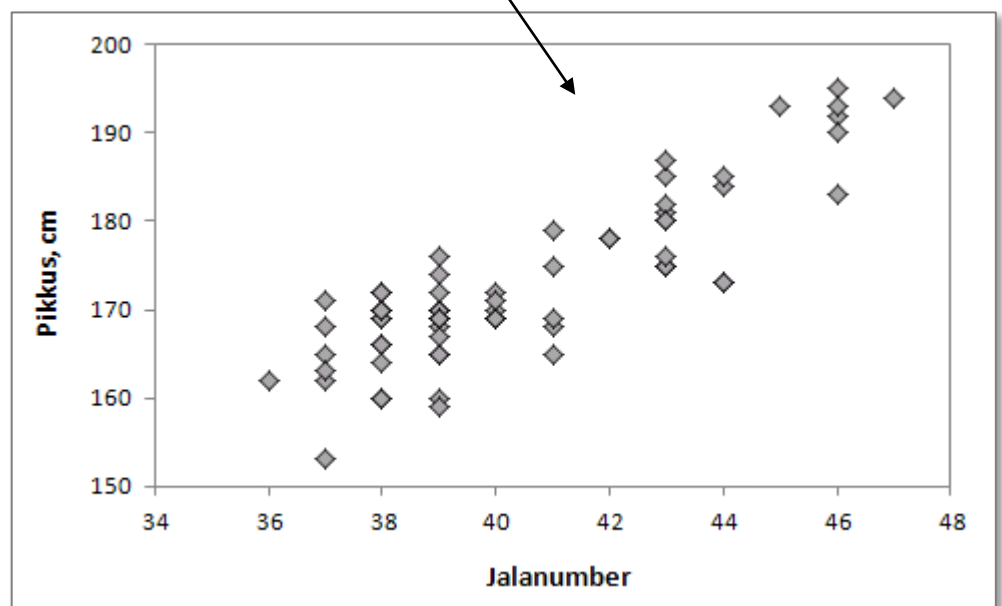


Tulemus:

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



Kujundage joonis!

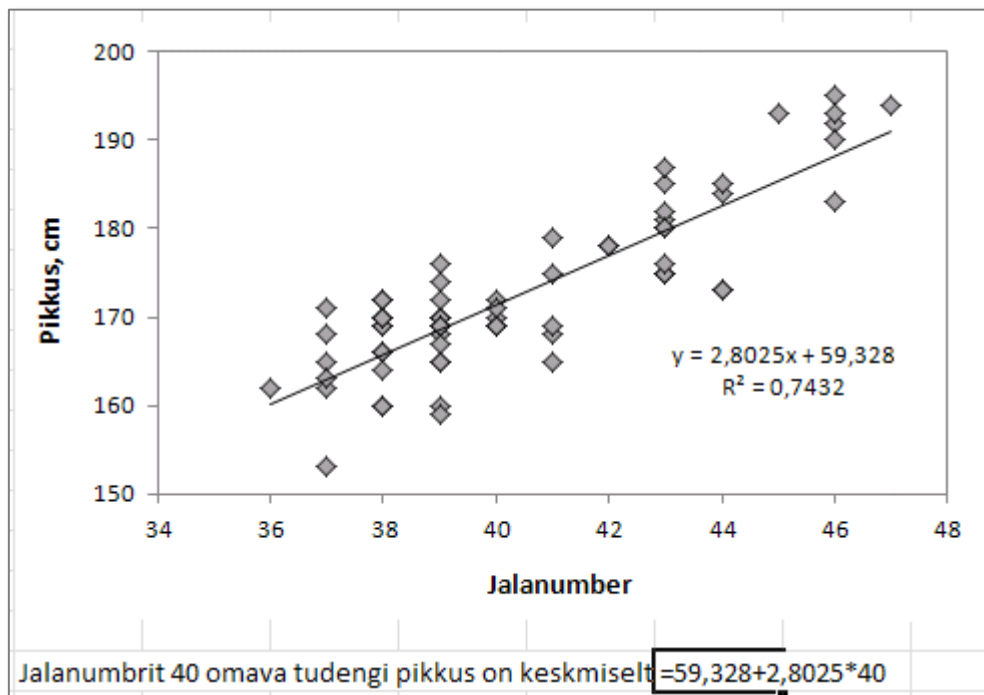


2. Prognoosimaks pikkust jalanumbri alusel, lisage tunnuste 'PIKKUS' ja 'JALANR' hajuvusdiagrammile **regressioonisirge**.

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

The image illustrates the steps to add a linear regression line to a scatter plot in Excel. It shows the 'Chart Tools' ribbon, the 'Trendline' dropdown menu, and the 'Format Trendline' dialog box. Annotations in Estonian explain the actions: 'Lineaarse trendijoonse lisamiseks' (for adding a linear trendline), 'Lisavalikute tarvis' (for the additional options), and 'Märkige, saamaks joonisele regressioonivõrrandit ja R^2 väärtust' (Mark the options to get the regression equation and R^2 value on the chart).

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 Microsoft Excel interface with the Data Analysis toolpak installed. The 'Data Analysis' dialog box is open, and the 'Regression' option is selected. The 'Input Y Range' is set to '\$B\$1:\$B\$69' and the 'Input X Range' is set to '\$E\$1:\$E\$69'. The 'Output Range' is set to '\$5\$21'. The 'Labels' checkbox is checked. The spreadsheet data is as follows:

SUGU	PIKKUS	MASS	PEA_YME	JALANR	ODE_VE	MAT_H	TEATER
N	160	36	30	38	0-1	3	rohkem kui aasta tegasi
N	162	60	33	37	0-1	3	rohkem kui aasta tegasi
N	164	51	33	38	0-1	3	rohkem kui aasta tegasi
N	165	34	34	37	0-1	4	rohkem kui aasta tegasi
N	165	68	33	39	2-3	3	rohkem kui aasta tegasi
N	168	38	36	41	>3	3	rohkem kui aasta tegasi
N	168	68	32	39	2-3	4	rohkem kui aasta tegasi
N	169	60	30	40	0-1	4	rohkem kui aasta tegasi
N	169	65	33	39	2-3	4	rohkem kui aasta tegasi
N	169	79	33	41	2-3	4	rohkem kui aasta tegasi
N	170	62	38	38	0-1	3	rohkem kui aasta tegasi
N	170	65	33	39	0-1	3	rohkem kui aasta tegasi
N	170	70	33	40	2-3	4	rohkem kui aasta tegasi
N	172	63	33	40	2-3	3	rohkem kui aasta tegasi
M	173	78	48	44	>3	3	rohkem kui aasta tegasi
M	173	80	37	44	0-1	3	rohkem kui aasta tegasi
M	175	62	36	43	0-1	3	rohkem kui aasta tegasi
M	175	74	38	43	2-3	3	rohkem kui aasta tegasi
M	175	77	36	43	0-1	3	rohkem kui aasta tegasi
M	178	110	38	42	0-1	3	rohkem kui aasta tegasi
M	180	80	36	43	0-1	3	rohkem kui aasta tegasi
M	181	88	36	43	0-1	3	rohkem kui aasta tegasi
M	182	74	34	43	>3	3	rohkem kui aasta tegasi
M	184	63	37	44	0-1	3	rohkem kui aasta tegasi
M	192	89	33,5	46		3	rohkem kui aasta tegasi
M	193	92	32	43	0-1	4	rohkem kui aasta tegasi
M	195	83	36	46	>3	3	rohkem kui aasta tegasi
N	133	47	33	37	0-1	4	viimase aasta jooksul
N	139	57	34	39	0-1	4	viimase aasta jooksul
N	160	32	33,5	38	2-3	3	viimase aasta
N	162	49	33	36	>3	3	viimase aasta
N	163	33	33	37	2-3	3	viimase aasta
N	165	70	32	39	>3	3	viimase aasta
N	166	36	34	39	>3	3	viimase aasta
N	166	60	34	38	2-3	4	viimase aasta
N	169	84	36	40	2-3	4	viimase aasta
N	169	65	38,5	38	>3	4	viimase aasta
N	169	70	33	40	2-3	3	viimase aasta
N	169	73	36	38	0-1	3	viimase aasta
N	170	33	38	38	2-3	3	viimase aasta
N	170	36	33	39	0-1	4	viimase aasta
N	170	64	36	39	2-3	4	viimase aasta
N	170	78	37	39	2-3	4	viimase aasta
N	171	70	34	40	0-1	4	viimase aasta
	172	33	34	38	>3	3	viimase aasta
N	172	75	33	39	2-3	4	viimase aasta
N	174	64	36	39	2-3	5	viimase aasta
N	175	80	32	41	0-1	4	viimase aasta
N	176	60	33	39	0-1	4	viimase aasta
M	176	61	34	43	0-1	3	viimase aasta
M	178	59	38	42	2-3	4	viimase aasta
N	179	72	37,5	41	2-3	3	viimase aasta
M	180	80	37	43	2-3	3	viimase aasta
M	183	100	63	46	0-1	3	viimase aasta
M	185	72	37	43	>3	3	viimase aasta
M	185	77	33	44	>3	3	viimase aasta
M	187	98	37	43	0-1	3	viimase aasta
M	190	83	38	46	2-3	3	viimase aasta
M	193	100	37,5	46	0-1	3	viimase aasta
M	194	88	34	47	0-1	3	viimase aasta
N	160	36	33	38	0-1	3	viimase kuu jooksul
N	171	37	34	37	2-3	3	viimase kuu jooksul
N	163	81	33	41	2-3	4	viimase kuu jooksul
N	167	60	32,5	39	0-1	4	viimase kuu jooksul
N	168	34	36	37	0-1	3	viimase kuu jooksul
N	169	30	34	39	0-1	3	viimase kuu jooksul
N	169	63	33	39	2-3	4	viimase kuu jooksul
N	172	38	37	38	0-1	4	viimase kuu jooksul

Regressioonanalüüsi tulemus:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0,86206929					
R Square	0,74316346					
Adjusted R Square	0,739272					
Standard Error	4,69044453					
Observations	68					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	4201,452774	4201,453	190,9728	3,72799E-21	
Residual	66	1452,017814	22,00027			
Total	67	5653,470588				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	59,3277985	8,23889095	7,200945	7,11E-10	42,87832203	75,7772751
JALANR	2,80250715	0,20279675	13,81929	3,73E-21	2,397610373	3,20740392

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	59,3277985	8,23889095	7,200945	7,11E-10	42,87832203	75,7772751
JALANR	2,80250715	0,20279675	13,81929	3,73E-21	2,397610373	3,20740392

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	4201,452774	4201,453	190,9728	3,72799E-21
Residual	66	1452,017814	22,00027		
Total	67	5653,470588			

= *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 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	0,86206929
R Square	0,74316346
Adjusted R Square	0,739272
Standard Error	4,69044453
Observations	68

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!