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

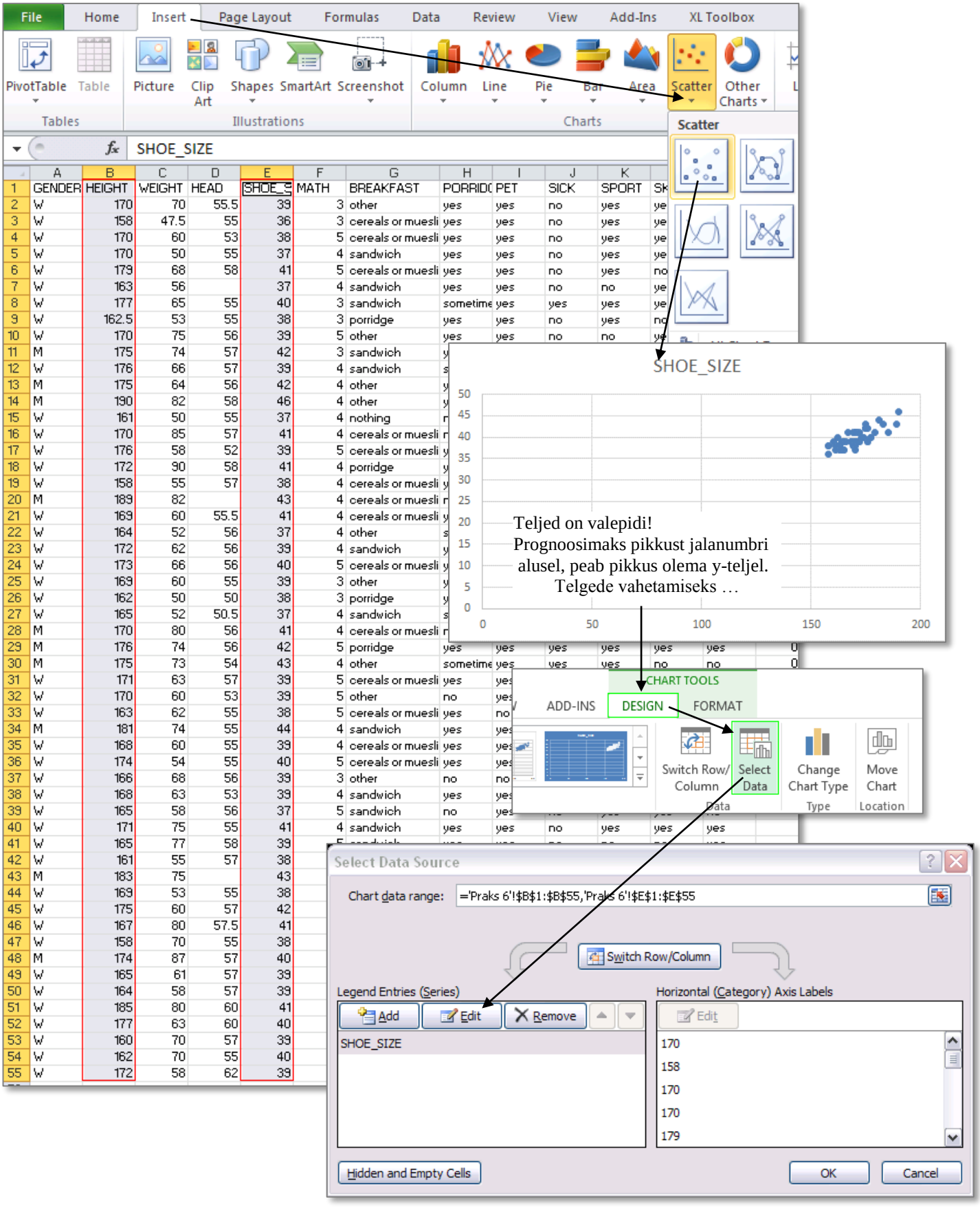
  $$Pikkus = a + b \times Jalanumber,$$

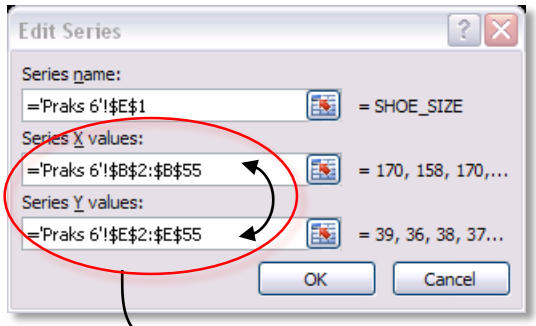
  kus a ja b asemel on Exceli poolt välja arvutatud kordajate väärtused.
- Kui suur on keskmiselt pikkuste vaheline erinevus tudengitel, kelle jalanumbrid erinevad 2 võrra?
- Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!
- Kirjeldamaks prognooside täpsust, sõnastage üks lause kas mitmese korrelatsioonikordaja (R), mudeli standardvea (*Standard Error*) või determinatsioonikordaja (R^2) kohta.

Ülesande 1 tööjuhend

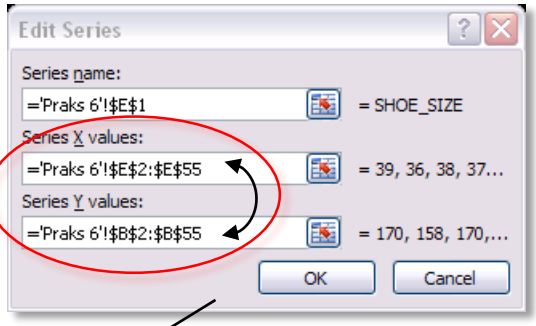
1. Illustreerige tunnuste ’HEIGHT’ ja ’SHOE_SIZE’ vahelist seost hajuvus- ehk punktdiagrammiga.

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

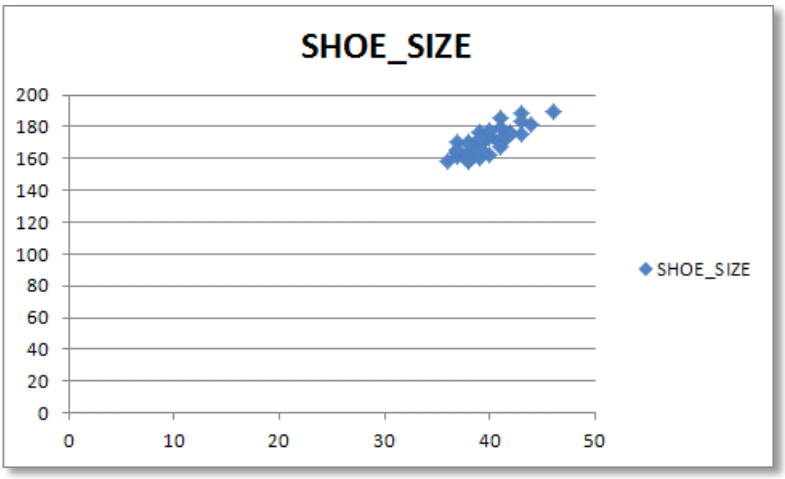




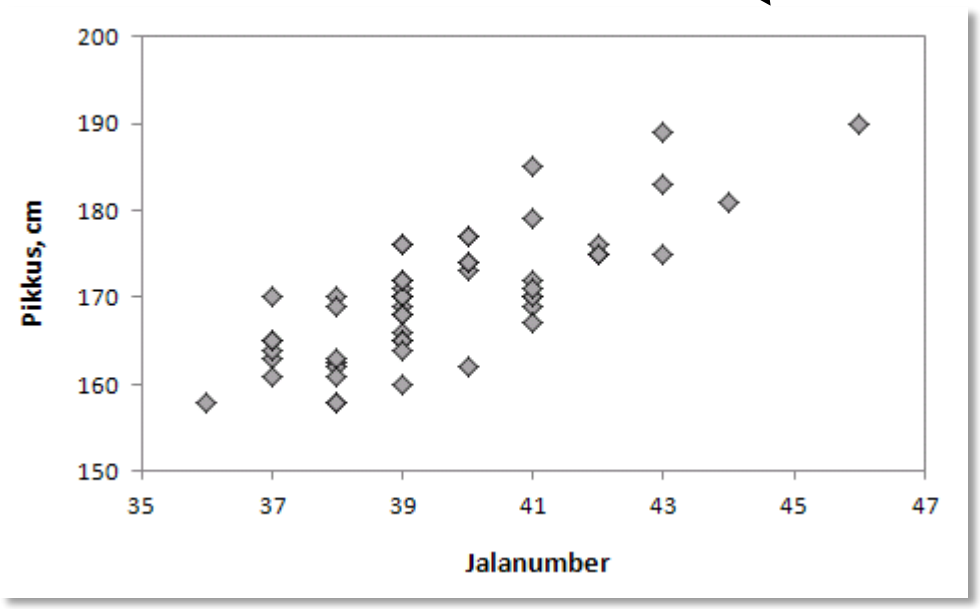
Vahetage x- ja y-telje väärtused



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

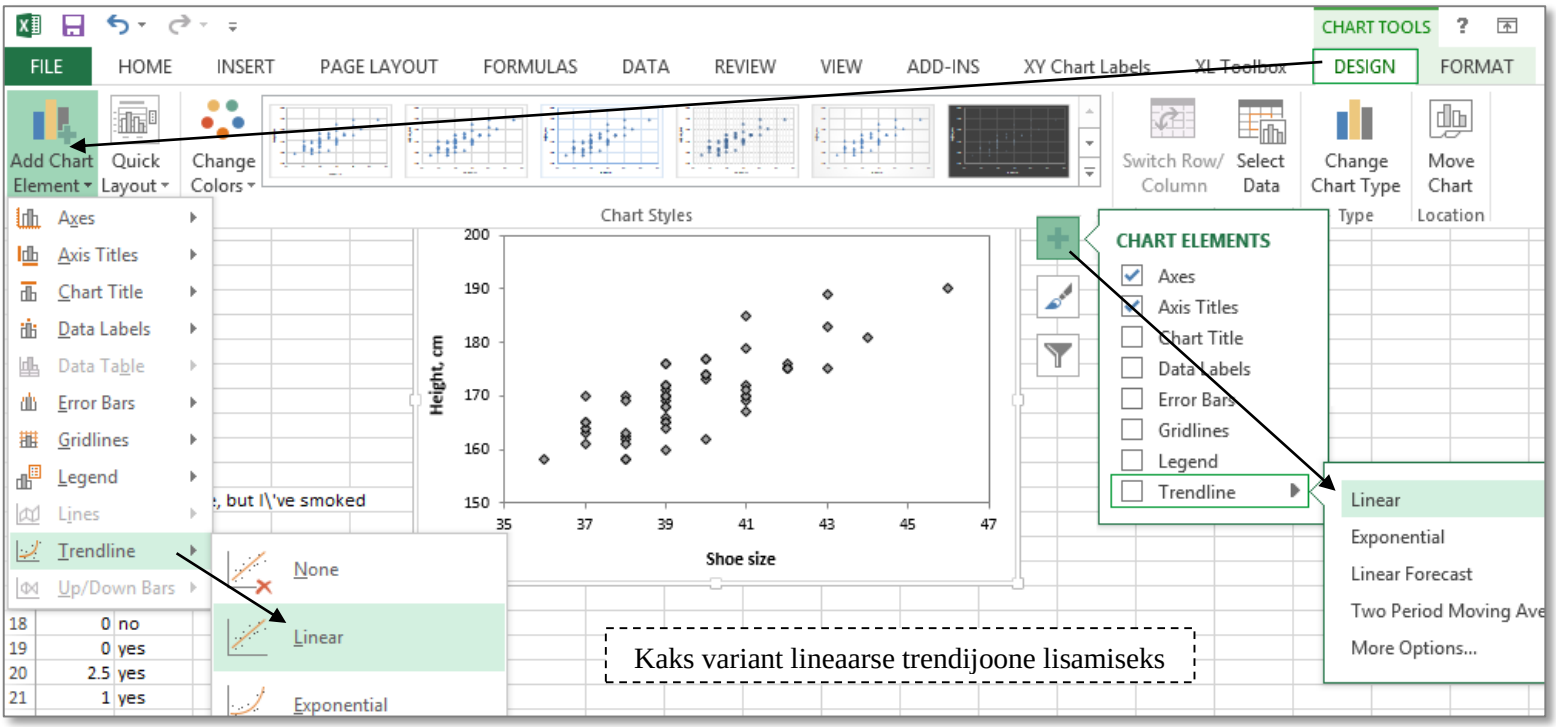


Kujundage joonis!

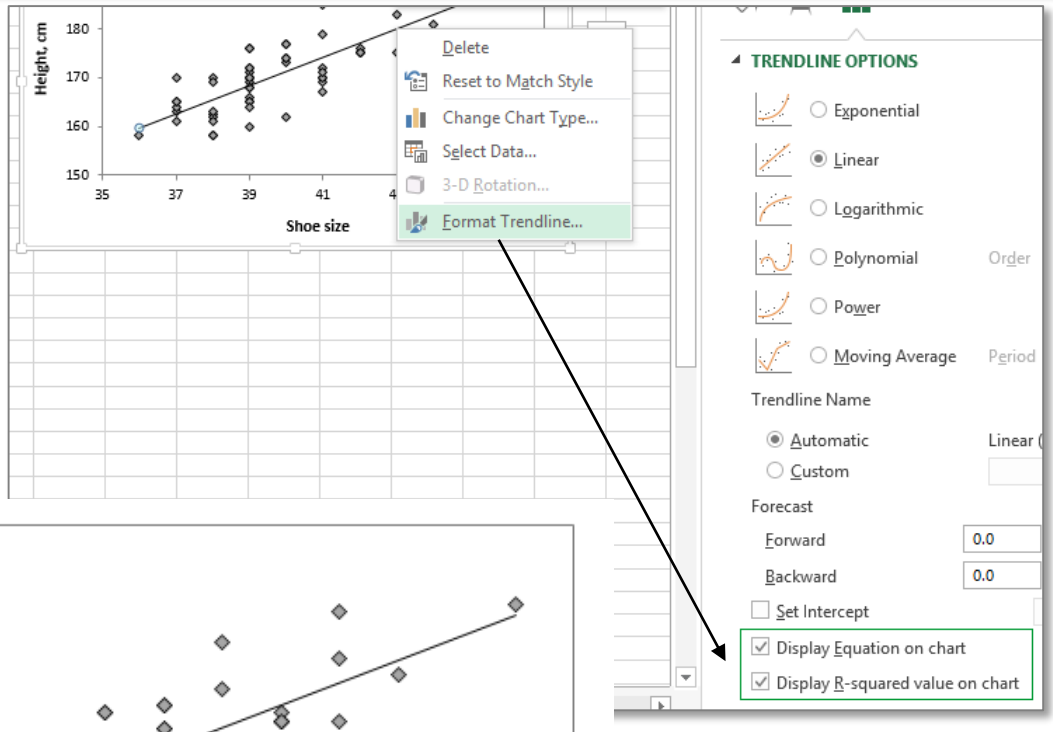


2. Prognoosimaks pikkust jalanumbri alusel, lisage tunnuste ’HEIGHT’ ja ’SHOE_SIZE’ hajuvusdiagrammile **regressioonisirge**.

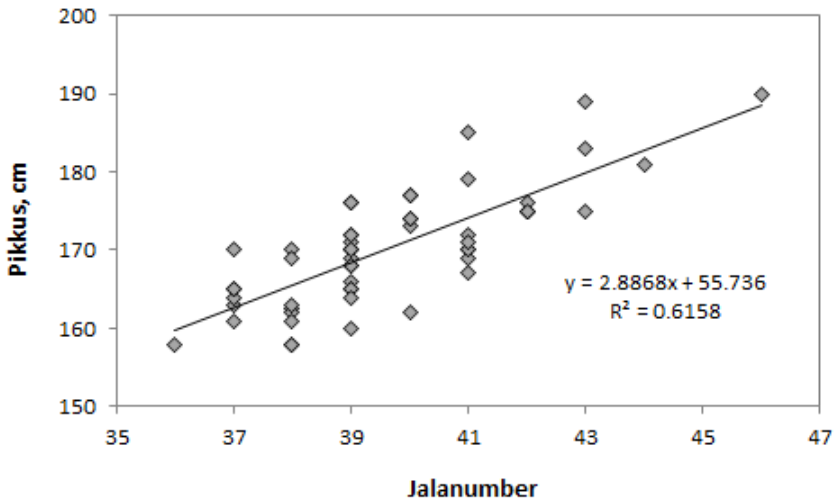
Lisage joonisele ka **regressioonivõrrand** ja viimase alusel leitavate prognooside täpsust kirjeldav **determinatsioonikordaja** R^2.



Regressioonivõrrandi ja R^2 lisamiseks

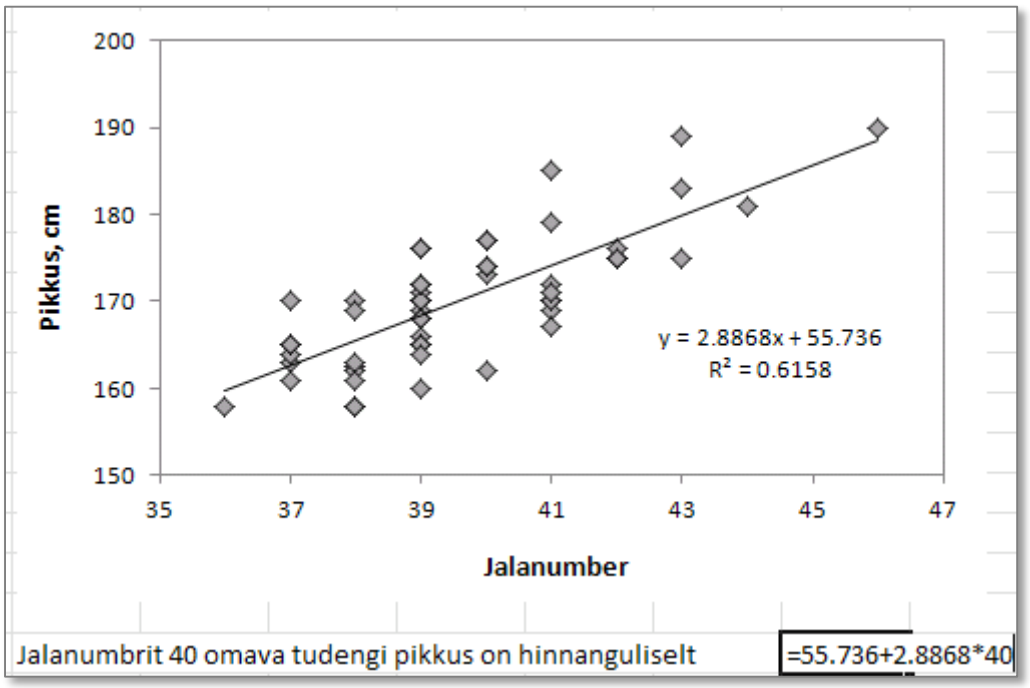


Tulemus:



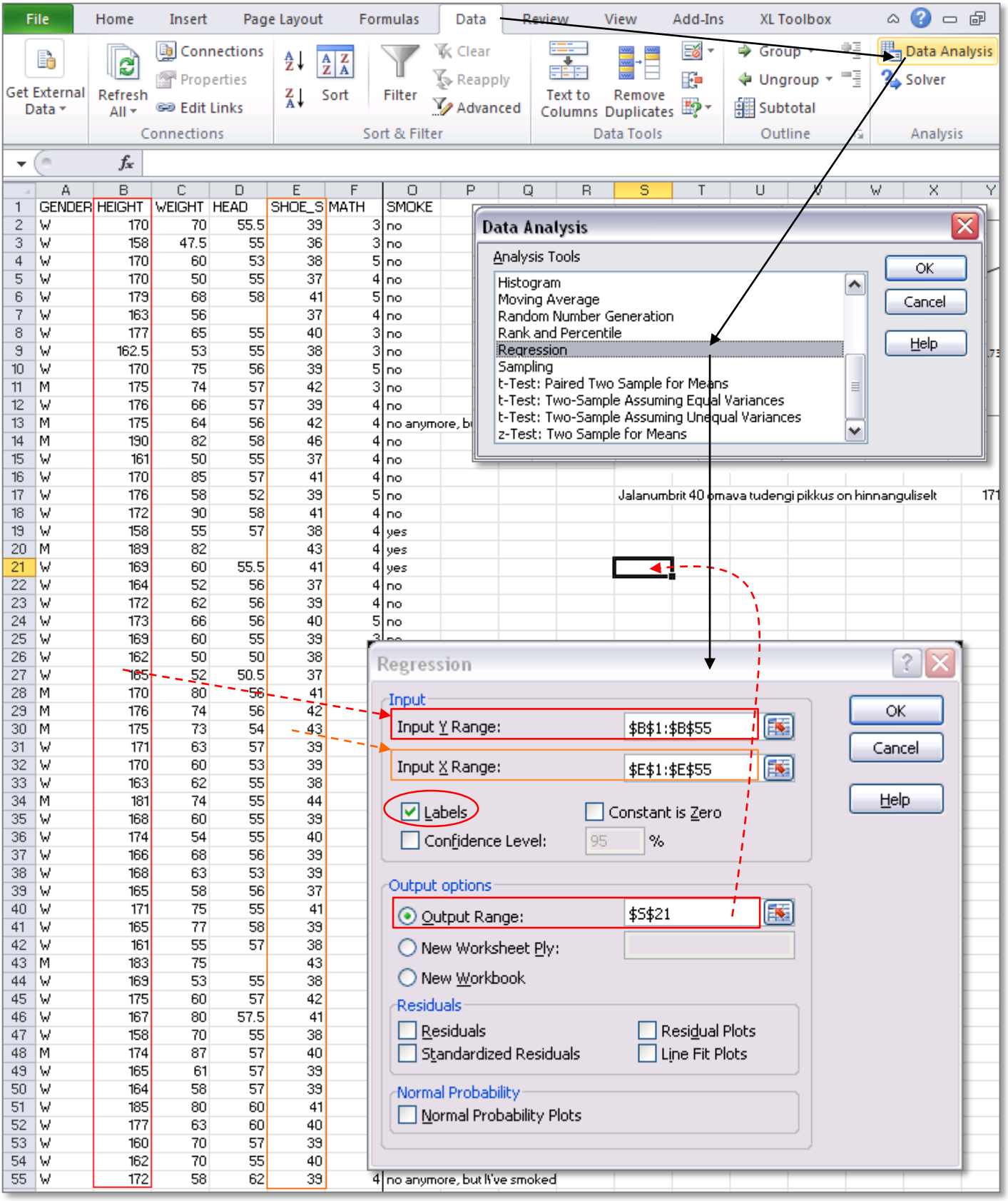
3. Prognoosige leitud võrrandi alusel, keskmiselt kui pikk on jalanumbrit 40 omav tudeng.

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.



Regressioonanalüüsi tulemus:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7847576
R Square	0.6158445
Adjusted R Square	0.6084569
Standard Error	4.6287793
Observations	54

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1786.077252	1786.077	83.36184	2.18481E-12
Residual	52	1114.131081	21.4256		
Total	53	2900.208333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
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

$$Pikkus = a + b \times Jalanumber,$$

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
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					
	df	*SS*	*MS*	*F*	*Significance F*
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

$$Pikkus = a + b \times Jalanumber$$

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 matemaatiliselt

H_0: $Pikkus = a$
H_1: $Pikkus = a + b \times Jalanumber$

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	
Regression Statistics	
Multiple R	0.7847576
R Square	0.6158445
Adjusted R Square	0.6084569
Standard Error	4.6287793
Observations	54

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, $0 \leqslant R^2 \leqslant 1$. 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 (mitmene korrelatsioonikordaja $R = 0{,}78$) ning pikkuste tegelikust varieeruvusest on leitud regressioonivõrrandi alusel ära kirjeldatav 62% ($R^2 = 0{,}62$).