

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.
-

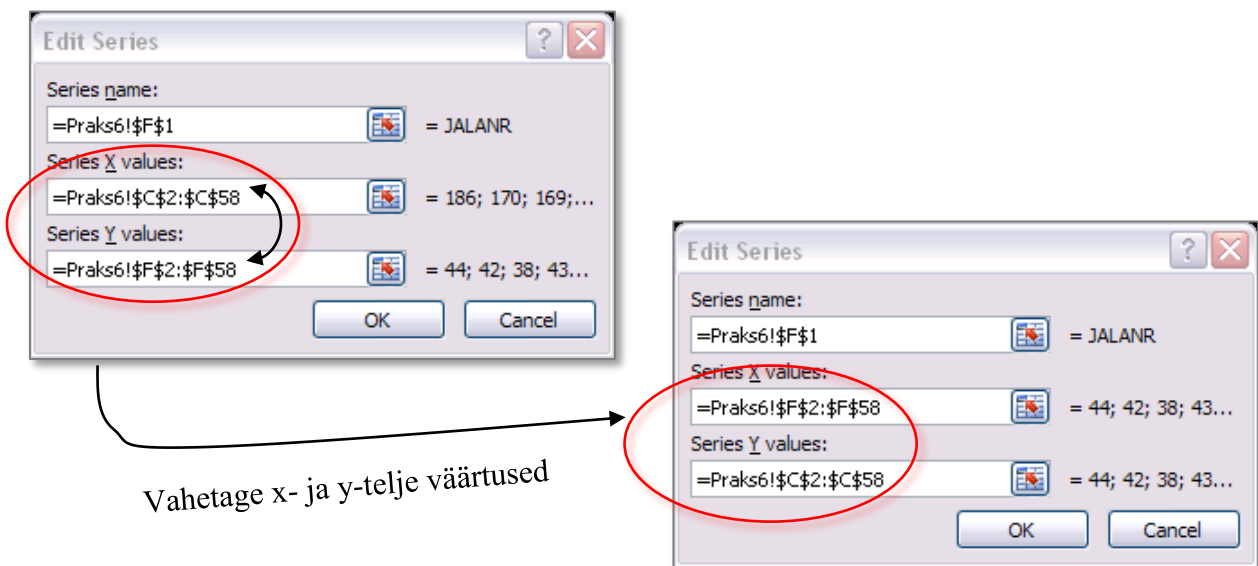
Ülesande 1 tööjuhend

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', and 'JALANR'. The scatter plot shows a positive correlation between 'PIKKUS' (height) on the y-axis and 'JALANR' (shoe size) on the x-axis. A text box points to the chart with the message: "Teljed on valepidi! Prognoosimaks pikkust jalanumbri alusel, peab pikkus olema y-teljel. Telgede vahetamiseks ...". Below the chart, the 'Select Data Source' dialog box is open, showing the chart data range as '=Praks6!\$C\$1:\$C\$58;Praks6!\$F\$1:\$F\$58'. The legend entries list 'JALANR' and the horizontal axis labels list '186', '170', '169', '180', and '179'. The 'Chart Tools' ribbon is also visible, showing the 'Design' tab with 'Chart Layouts' and 'Chart Styles' options.

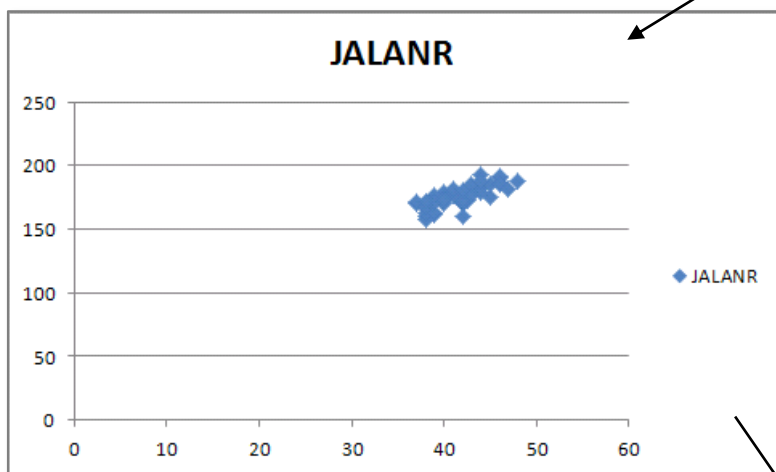
| RIIK | SUGU | PIKKUS | MASS | PEA_P | JALANR |
|-------|------|--------|------|-------|--------|
| Eesti | M | 186 | 95 | 59 | 44 |
| Eesti | N | 170 | 85 | 57 | 42 |
| Eesti | N | 169 | 50 | 54 | 38 |
| Eesti | M | 180 | 70 | 56 | 43 |
| Eesti | N | 179 | 72 | 55 | 40 |
| Eesti | N | 170 | 55 | 55 | 37 |
| Eesti | N | 160 | 58 | 55 | 38 |
| Eesti | N | 161 | 57 | 55 | 39 |
| Eesti | N | 171,5 | 59 | 57 | 38 |
| Eesti | N | 180 | 63 | 58 | 41 |
| Eesti | N | 168 | 54 | 57 | 38 |
| Eesti | N | 170 | 57 | 52 | 40 |
| Eesti | N | 163 | 61 | 57,5 | 39 |
| Eesti | M | 172 | 66 | 54 | 42 |
| Eesti | M | 183 | 73 | 54,5 | 44 |
| Eesti | M | 185 | 72 | 56 | 44 |
| Eesti | M | 187 | 94 | 59 | 46 |
| Eesti | M | 183 | 83 | 56 | 43 |
| Eesti | M | 190 | 102 | 59 | 46 |
| Eesti | M | 173 | 58 | 55,5 | 42,5 |
| Eesti | N | 157 | 63 | 55,5 | 38 |
| Eesti | M | 180 | 80 | 56 | 43 |
| Eesti | M | 180 | 84 | 60 | 43 |
| Eesti | M | 175 | 87 | 54 | 45 |
| Eesti | M | 181 | 81 | 55 | 43 |
| Eesti | M | 177 | 75 | 54 | 42 |
| Eesti | N | 175 | 60 | 53 | 39 |
| Eesti | M | 185 | 100 | 67 | 45 |
| Eesti | N | 176 | 75 | 56 | 41 |
| Eesti | N | 170 | 100 | 56 | 42 |
| Eesti | M | 179 | 59 | 56 | 44 |
| Eesti | M | 193 | 75 | 55 | 44 |
| Eesti | N | 169 | 60 | 56 | 38 |
| Eesti | N | 185 | 80 | 60 | 43 |
| Eesti | N | 163 | 64 | 57 | 38 |
| Eesti | N | 181 | 74 | 56 | 41 |
| Eesti | M | 191 | 70 | 59 | 46 |
| Eesti | M | 180 | 65 | 56 | 42 |
| Eesti | M | 173 | 67 | 55 | 42 |
| Eesti | N | 172 | 65 | 53 | 37 |
| Eesti | N | 173 | 80 | 56 | 40 |
| Eesti | N | 167 | 61 | 57 | 38 |
| Eesti | M | 185 | 100 | 60 | 46 |
| Eesti | M | 182 | 100 | 47 | |
| Eesti | N | 171 | 64 | 54,5 | 39 |
| Eesti | N | 173 | 55 | 57 | 39 |
| Eesti | N | 174 | 75 | 57 | 40 |
| Eesti | N | 162 | 55 | 57 | 39 |
| Eesti | N | 176 | 70 | 46 | 39 |
| Eesti | N | 168 | 58 | 55 | 38 |
| Eesti | N | 171 | 58 | 55 | 39 |
| Eesti | N | 172 | 55 | 53 | 38 |
| Eesti | M | 176 | 66 | 55 | 43 |
| Eesti | M | 188 | 95 | 59 | 44 |
| Eesti | M | 188 | 80 | 48 | 48 |
| Eesti | M | 180 | 74 | 56 | 42 |
| Eesti | N | 173 | 59 | 58 | 40 |



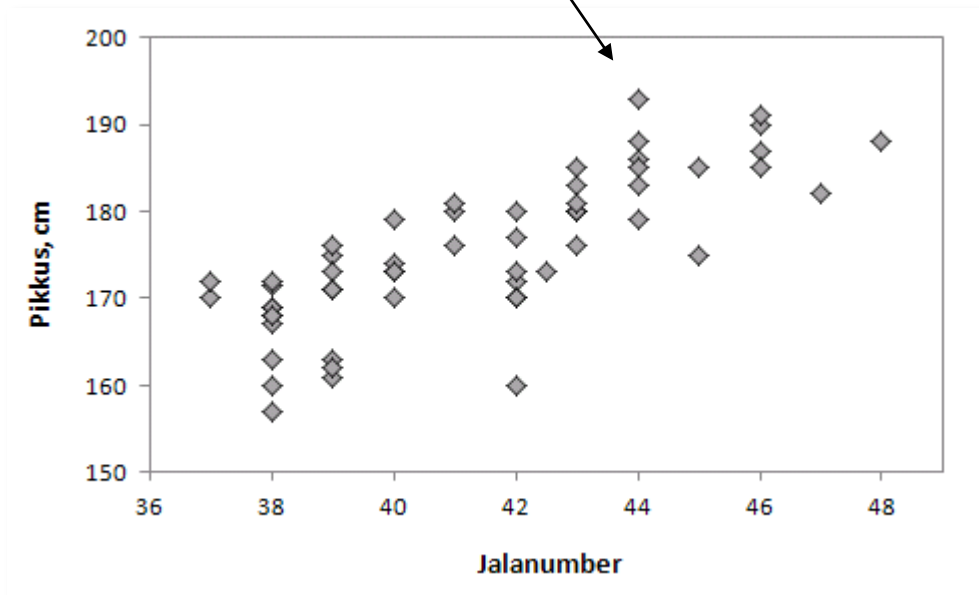
Vahetage x- ja y-telje väärtused

Tulemus:

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



Kujundage joonis!



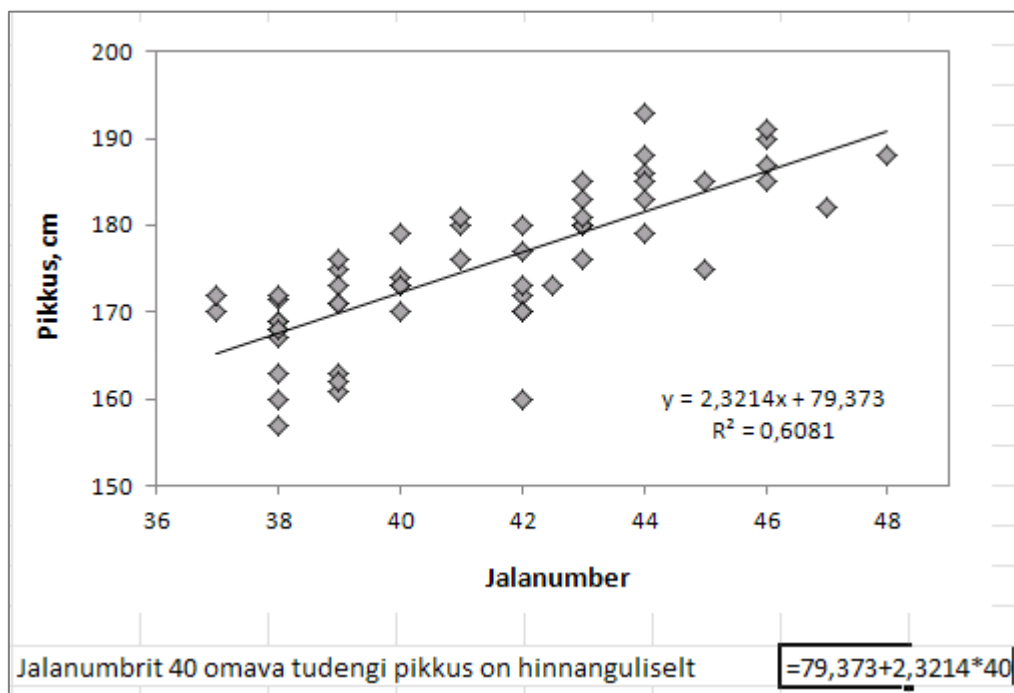
2. Prognoosimaks pikkust jalanumbri alusel, lisage tunnuste 'PIKKUS' ja 'JALANR' hajuvusdiagrammile **regressioonisirge**.

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

The image shows a multi-step process in Microsoft Excel:

- Top Panel:** The 'Chart Tools' ribbon is active, with the 'Layout' tab selected. The 'Trendline' button is highlighted, and its dropdown menu is open. The 'Linear Trendline' option is selected, with an arrow pointing to it from the text 'Lineaarse trendijoonse lisamiseks'.
- Middle Panel:** A scatter plot is shown with a linear trendline. A context menu is open over the trendline, with 'Format Trendline...' selected. An arrow points from the text 'Lisavalikute tarvis' to this option.
- Bottom Panel:** The 'Format Trendline' dialog box is open. Under 'Trendline Options', 'Linear' is selected. At the bottom, the checkboxes for 'Display Equation on chart' and 'Display R-squared value on chart' are checked. An arrow points from the text 'Märkige, saamaks joonisele regressioonivõrrandit ja R^2 väärtust' to these checkboxes.

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 image shows a Microsoft Excel spreadsheet with a 'Data Analysis' task pane and a 'Regression' dialog box. The spreadsheet data is as follows:

| | B | C | D | E | F | G | S |
|----|------|--------|------|-------|--------|---------|--------------------------|
| 1 | SUGU | PIKKUS | MASS | PEA_P | JALANR | ODE_VEN | KINO |
| 2 | M | 186 | 95 | 59 | 44 | 1 | viimase aasta jooksul |
| 3 | N | 170 | 85 | 57 | 42 | 6 | viimase kuu jooksul |
| 4 | N | 169 | 50 | 54 | 38 | 1 | viimase aasta jooksul |
| 5 | M | 180 | 70 | 56 | 43 | 0 | viimase 10 päeva jooksul |
| 6 | | 179 | 72 | 55 | 40 | 1 | viimase kuu jooksul |
| 7 | N | 170 | 55 | 55 | 37 | 1 | viimase 10 päeva jooksul |
| 8 | N | 160 | 58 | 55 | 38 | 1 | viimase kuu jooksul |
| 9 | N | 161 | 57 | 55 | 39 | 1 | viimase 10 päeva jooksul |
| 10 | N | 171,5 | 59 | 57 | 38 | 1 | viimase kuu jooksul |
| 11 | N | 180 | 63 | 58 | 41 | 2 | viimase aasta jooksul |
| 12 | N | 168 | 54 | 57 | 38 | 1 | viimase aasta jooksul |
| 13 | N | 170 | 57 | 52 | 40 | 2 | viimase 10 päeva jooksul |
| 14 | N | 163 | 61 | 57,5 | 39 | 0 | viimase aasta jooksul |
| 15 | M | 172 | 66 | 54 | 42 | 1 | viimase aasta jooksul |
| 16 | M | 183 | 73 | 54,5 | 44 | | rohkem kui aasta tagasi |
| 17 | M | 185 | 72 | 56 | 44 | 1 | viimase aasta jooksul |
| 18 | M | 187 | 94 | 59 | 46 | 1 | viimase aasta jooksul |
| 19 | M | 183 | 83 | 56 | 43 | 3 | viimase aasta jooksul |
| 20 | M | 190 | 102 | 59 | 46 | 1 | viimase aasta jooksul |
| 21 | M | 173 | 58 | 55,5 | 42,5 | 1 | viimase aasta jooksul |
| 22 | N | 157 | 63 | 55,5 | 38 | 1 | viimase kuu jooksul |
| 23 | M | 180 | 80 | 56 | 43 | 2 | viimase kuu jooksul |
| 24 | M | 180 | 84 | 60 | 43 | 1 | rohkem kui aasta tagasi |
| 25 | M | 175 | 87 | 54 | 45 | 3 | viimase aasta jooksul |
| 26 | M | 181 | 81 | 55 | 43 | 2 | viimase aasta jooksul |
| 27 | M | 177 | 75 | 54 | 42 | 1 | viimase kuu jooksul |
| 28 | N | 175 | 60 | 53 | 39 | | viimase kuu jooksul |
| 29 | M | 185 | 100 | 67 | 45 | | |
| 30 | N | 176 | 75 | 56 | 41 | | |
| 31 | N | 170 | 100 | 56 | 42 | | |
| 32 | M | 179 | 59 | 56 | 44 | | |
| 33 | M | 193 | 75 | 55 | 44 | | |
| 34 | N | 169 | 60 | 56 | 38 | | |
| 35 | M | 185 | 80 | 60 | 43 | | |
| 36 | N | 163 | 64 | 57 | 38 | | |
| 37 | N | 181 | 74 | 56 | 41 | | |
| 38 | M | 191 | 70 | 59 | 46 | | |
| 39 | M | 160 | 65 | 56 | 42 | | |
| 40 | M | 173 | 67 | 55 | 42 | | |
| 41 | N | 172 | 65 | 53 | 37 | | |
| 42 | N | 173 | 80 | 56 | 40 | | |
| 43 | N | 167 | 61 | 57 | 38 | | |
| 44 | M | 185 | 100 | 60 | 46 | | |
| 45 | M | 182 | 100 | | 47 | | |
| 46 | N | 171 | 64 | 54,5 | 39 | | |
| 47 | N | 173 | 55 | 57 | 39 | | |
| 48 | N | 174 | 75 | 57 | 40 | | |
| 49 | N | 162 | 55 | 57 | 39 | | |
| 50 | N | 176 | 70 | 46 | 39 | | |
| 51 | N | 168 | 58 | 55 | 38 | | |
| 52 | N | 171 | 58 | 55 | 39 | | |
| 53 | N | 172 | 55 | 53 | 38 | | |
| 54 | M | 176 | 66 | 55 | 43 | | |
| 55 | M | 188 | 95 | 59 | 44 | | |
| 56 | M | 188 | 80 | 48 | 48 | | |
| 57 | M | 180 | 74 | 56 | 42 | | |
| 58 | N | 173 | 59 | 58 | 40 | | |

The 'Data Analysis' task pane shows the following options:

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression**
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

The 'Regression' dialog box has the following settings:

- Input Y Range: $\$C\$1:\$C\58
- Input X Range: $\$F\$1:\$F\58
- Labels
- Constant is Zero
- Confidence Level: 95 %
- Output Range: $\$U\21
- New Worksheet Ply:
- New Workbook
- Residuals: Residuals, Residual Plots, Standardized Residuals, Line Fit Plots
- Normal Probability: Normal Probability Plots

Regressioonanalüüsi tulemus:

| SUMMARY OUTPUT | | | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|-----------------------|------------------|
| <i>Regression Statistics</i> | | | | | | |
| Multiple R | 0,77980172 | | | | | |
| R Square | 0,60809072 | | | | | |
| Adjusted R Square | 0,60096509 | | | | | |
| Standard Error | 5,37994256 | | | | | |
| Observations | 57 | | | | | |
| <i>ANOVA</i> | | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> | |
| Regression | 1 | 2470,02182 | 2470,022 | 85,3386 | 8,82306E-13 | |
| Residual | 55 | 1591,908005 | 28,94378 | | | |
| Total | 56 | 4061,929825 | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
| Intercept | 79,3727035 | 10,42634534 | 7,612706 | 3,74E-10 | 58,47784053 | 100,26757 |
| JALANR | 2,32136296 | 0,251287157 | 9,237889 | 8,82E-13 | 1,817772241 | 2,8249537 |

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 | 79,3727035 | 10,42634534 | 7,612706 | 3,74E-10 | 58,47784053 | 100,26757 |
| JALANR | 2,32136296 | 0,251287157 | 9,237889 | 8,82E-13 | 1,817772241 | 2,8249537 |

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 | 2470,02182 | 2470,022 | 85,3386 | 8,82306E-13 = <i>p</i> |
| Residual | 55 | 1591,908005 | 28,94378 | | |
| Total | 56 | 4061,929825 | | | |

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 | 0,77980172 |
| R Square | 0,60809072 |
| Adjusted R Square | 0,60096509 |
| Standard Error | 5,37994256 |
| Observations | 57 |

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!