Biomeetria bioloogidele
4. loeng

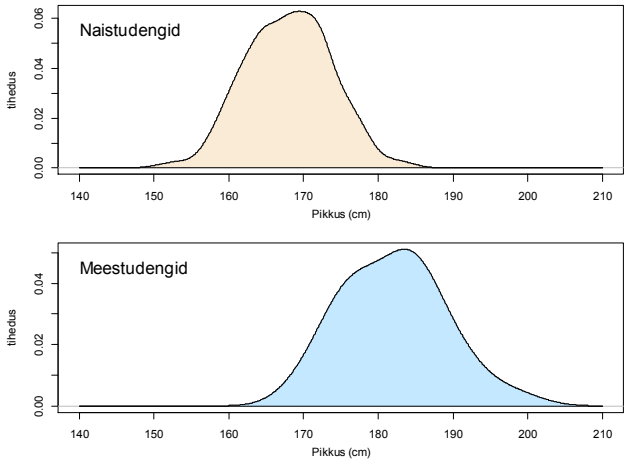
Dispersioonanalüüs

Prognoosime
pideva tunnuse Y väärtuseid
faktortunnuse abil.

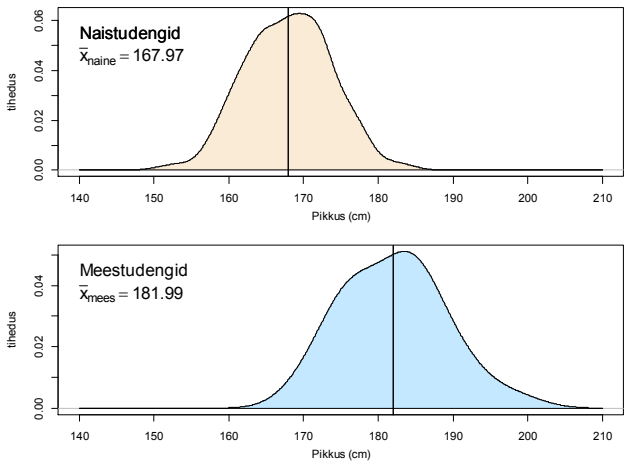
Indikaatortunnused (dummy variables).

Vahel soovime prognoosida tunnuse *Y* väärtuseid, aga tunnus, mille abil me prognoosime, *X*, pole pidev. Näiteks soovime prognoosida tudengi kaalu. Aga informatsioon, mida saame selle prognoosi tegemiseks kasutada, pole esitatav pideva tunnusena. Näiteks teame vaid inimese sugu (mees või naine).

Mida prognoosiks pakkuda?



Mida prognoosiks pakkuda?



Milline on mudel?

Variant 1: *Pikkus* = 167,97 $I_{Sugu=naine}$ + 181,99 $I_{Sugu=mees}$ + ε

$I_{Sugu=naine}$ = 1, kui tegemist naisega
$I_{Sugu=naine}$ = 0, kui ei ole tegemist naisega

Milline on mudel?

Variant 1: *Pikkus* = 167,97 $I_{Sugu=naine}$ + 181,99 $I_{Sugu=mees}$ + ε

Variant 2: *Pikkus* = 167,97 + 14,02 $I_{Sugu=mees}$ + ε

Variant 3: *Pikkus* = 181,99 - 14,02 $I_{Sugu=naine}$ + ε

Milline on mudel?

=0

Variant 1: *Pikkus* = 167,97 $I_{Sugu=naine}$ + 181,99 $I_{Sugu=mees}$ + ε

Variant 2: *Pikkus* = 167,97 + 14,02 $I_{Sugu=mees}$ + ε

Variant 3: *Pikkus* = 181,99 - 14,02 $I_{Sugu=naine}$ + ε

Prognoosid

Variant 1
Mees: 181,99
Naine:

Milline on mudel?

=0

Variant 1: *Pikkus* = 167,97 $I_{Sugu=naine}$ + 181,99 $I_{Sugu=mees}$ + ε

Variant 2: *Pikkus* = 167,97 + 14,02 $I_{Sugu=mees}$ + ε

Variant 3: *Pikkus* = 181,99 - 14,02 $I_{Sugu=naine}$ + ε

Prognoosid

Variant 1
Mees: 181,99
Naine: 167,97

Milline on mudel?

Variant 1: *Pikkus* = 167,97 $I_{Sugu=naine}$ + 181,99 $I_{Sugu=mees}$ + ε

Variant 2: *Pikkus* = 167,97 + 14,02 $I_{Sugu=mees}$ + ε

Variant 3: *Pikkus* = 181,99 - 14,02 $I_{Sugu=naine}$ + ε

Prognoosid

Variant 1	Variant 2	Variant 3
Mees: 181,99	Mees: 181,99	Mees: 181,99
Naine: 167,97	Naine: 167,97	Naine: 167,97
		=181,99-14,02

Variant 2. Hindamine.

lm(pikkus~mees, data=tudengid)

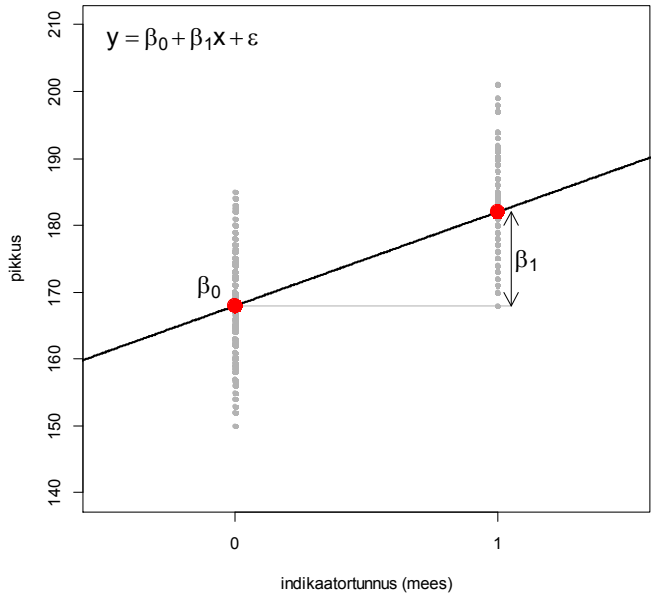
pikkus	mees
159	0
172	0
170	1
193	1
182	1
167	0
168	0
165	0
172	1
173	1
…	…

```
Call:
lm(formula = pikkus ~ mees, data = tudengid)

Coefficients:
(Intercept)         mees
     167.97        14.02
```

Variant 2: *Pikkus* = 167,97 + 14,02 $I_{Sugu=mees}$ + ε

Pikkus = 167,97 + 14,02 $I_{Sugu=mees}$ + ε



pikkus	mees	suguF
159	0	naine
172	0	naine
170	1	mees
193	1	mees
182	1	mees
167	0	naine
168	0	naine
165	0	naine
172	1	mees
173	1	mees
.....	.....	

```
> lm(pikkus~mees)

Coefficients:
(Intercept)         mees
     167.97        14.02

> lm(pikkus~factor(sugu))

Coefficients:
  (Intercept)  factor(sugu)2
       167.97          14.02
```

Tähelepanu!
Erinevad programmid valivad võrdluse aluse erinevalt! Kui analüüsime sama andmestikku erinevate statistikaprogrammide abil, võime saada vastuseks vägagi erinevaid numbreid. Tulemuste interpretatsioon jääb aga alati samaks…

SAS

```
proc glm data=vererohk;
  class ravi;
  model muutus=ravi /solution; run;
```

```
Parameter            Estimate        Error    t Value    Pr > |t|
 Intercept           -12.804 B       1.454     -8.805      <.0001
 ravi      platseebo   6.415 B       2.028      3.164      0.0021
 ravi      B           0.000 B        .          .          .
```

R

```
summary(lm(muutus~ravi))
```

```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   -6.389      1.413  -4.521  1.65e-05 ***
raviB         -6.415      2.028  -3.164  0.00205  **
```

Testimine – faktortunnusel 2 taset

Kui faktortunnusel on kõigest 2 taset, jõuame samade tulemusteni nii t-testi (võrdset hajuvust eeldava t-testi) kui ka (indikaatortunnust kasutava) regressioonanalüüsi abil:

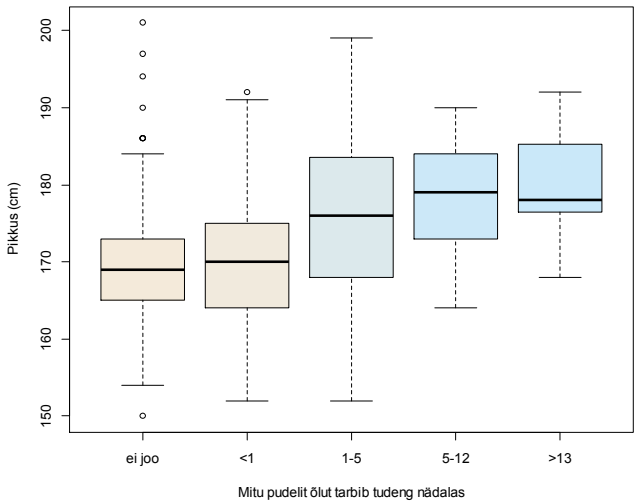
Studenti t-test

```
> t.test(muutus~ravi, var.equal=T)
t = 3.1635, df = 103, p-value = 0.002049
sample estimates:
mean in group platseebo         mean in group B
              -6.388889              -12.803922
```

```
> summary(lm(muutus~factor(ravi)))
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.389      1.413  -4.521 1.65e-05 ***
factor(ravi)B     -6.415      2.028  -3.164  0.00205 **
```

```
> summary(lm(muutus~I_raviB))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.389      1.413  -4.521 1.65e-05 ***
I_raviB        -6.415      2.028  -3.164  0.00205 **
```

Kui faktortunnusel on enam kui kaks võimalikku väärtust…



Mudel…

$$Pikkus = 169{,}3 + 0{,}8\, I_{olu=\text{„<1“}} + 6{,}8\, I_{olu=\text{„1-5“}} + 8{,}7\, I_{olu=\text{„5-12“}} + 10{,}9\, I_{olu=\text{„>12“}} + \varepsilon$$

prognoos (keskmine)

Ei joo	169,3	
<1	170,1	= 169,3+ 0,8
1-5	176,1	= 169,3+ 6,8
5-12	178,0	= 169,3+ 8,7
>12	180,2	= 169,3+10,9

$I_{olu<1}$ $I_{olu=„1\text{-}5“}$ $I_{olu=„5\text{-}12“}$ $I_{olu>12}$

pikkus	olu	I1	I2	I3	I4
172	ei joo	0	0	0	0
164	1-5	0	1	0	0
178	<1	1	0	0	0
186	>13	0	0	0	1
176	<1	1	0	0	0
172	1-5	0	1		
178	1-5	0	1		
172	ei joo	0	0		
176	<1	1	0		
165	<1	1	0		
176	ei joo	0	0		
160	ei joo	0	0		
187	5-12	0	0		
…	…	…			

Hindame mudeli:

```
lm(pikkus~I1+I2+I3+I4)
```

või

```
lm(pikkus~factor(olu))
```

```
> summary(lm(pikkus~factor(olu)))

Residuals:
    Min      1Q  Median      3Q     Max
-24.128  -5.355  -0.128   4.645  31.645

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     169.3549     0.4857 348.695  < 2e-16 ***
factor(olu)<1     0.7670     0.6875   1.116 0.264991
factor(olu)1-5    6.7734     0.9581   7.070 3.99e-12 ***
factor(olu)5-12   8.6784     1.5256   5.689 1.93e-08 ***
factor(olu)>13   10.8594     3.0331   3.580 0.000369 ***

Residual standard error: 7.921 on 655 degrees of freedom
Multiple R-squared:  0.1164,    Adjusted R-squared:  0.111
F-statistic: 21.57 on 4 and 655 DF,  p-value: < 2.2e-16
```

```
> summary(lm(pikkus~factor(olu)))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     169.3549     0.4857 348.695  < 2e-16 ***
factor(olu)<1     0.7670     0.6875   1.116 0.264991
factor(olu)1-5    6.7734     0.9581   7.070 3.99e-12 ***
factor(olu)5-12   8.6784     1.5256   5.689 1.93e-08 ***
factor(olu)>13   10.8594     3.0331   3.580 0.000369 ***

Residual standard error: 7.921 on 655 degrees of freedom
Multiple R-squared:  0.1164,    Adjusted R-squared:  0.111
F-statistic: 21.57 on 4 and 655 DF,  p-value: < 2.2e-16
```

Kuidas erineb 5-12 grupp õlut mittejoovatest tudengitest…

```
> õlu=relevel(olu, ref=">13")
> summary(lm(pikkus~õlu))


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  180.214      2.994  60.193  < 2e-16 ***
õluei joo    -10.859      3.033  -3.580 0.000369 ***
õlu<1        -10.092      3.033  -3.327 0.000926 ***
õlu1-5        -4.086      3.106  -1.316 0.188758
õlu5-12       -2.181      3.325  -0.656 0.512094

Residual standard error: 7.921 on 655 degrees of freedom
Multiple R-squared:  0.1164,    Adjusted R-squared:  0.111
F-statistic: 21.57 on 4 and 655 DF,  p-value: < 2.2e-16
```

Kuidas erineb 5-12 grupp enam kui 12 pudelit õlut joovatest tudengitest…

Sama mudel, kaks vaatenurka…

```
> summary(lm(pikkus~factor(sport)))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    168.9363     0.7343 230.051  < 2e-16 ***
factor(sport)2   1.6167     0.8454   1.912   0.0563 .
factor(sport)3   4.3281     1.0516   4.116 4.36e-05 ***
factor(sport)4   9.0012     1.6214   5.552 4.12e-08 ***
factor(sport)5   4.8137     4.1541   1.159   0.2470

Multiple R-squared:  0.05959,   Adjusted R-squared:  0.05384
F-statistic: 10.36 on 4 and 654 DF,  p-value: 3.851e-08
```

Referentstase Sport=1

```
> SPORT=relevel(factor(sport), ref="5")
> summary(lm(pikkus~SPORT))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 173.7500     4.0887  42.496   <2e-16 ***
SPORT1       -4.8137     4.1541  -1.159    0.247
SPORT2       -3.1970     4.1101  -0.778    0.437
SPORT3       -0.4856     4.1574  -0.117    0.907
SPORT4        4.1875     4.3367   0.966    0.335

Multiple R-squared:  0.05959,   Adjusted R-squared:  0.05384
F-statistic: 10.36 on 4 and 654 DF,  p-value: 3.851e-08
```

Referentstase Sport=5

Kas mudelist on kasu?
H_0: Kasu pole…

Testimisest…

Kas sportimisharjumuse teadmisest on kasu pikkuse prognoosimisel?

```
> mudel0=lm(pikkus~1)
> summary(mudel0)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  171.117      0.327   523.3   <2e-16 ***

Residual standard error: 8.401 on 659 degrees of freedom
```

Testimisest…

Kas sportimisharjumuse teadmisest on kasu pikkuse prognoosimisel?

```
> mudel0=lm(pikkus~1)
> mudel1=lm(pikkus~factor(sport))
> anova(mudel0, mudel1)
```

H0: Lihtsam mudel prognoosib sama hästi kui keerukam

Testimisest…

Kas sportimisharjumuse teadmisest on kasu pikkuse prognoosimisel?

Mudelid peavad olema hinnatud kasutades samu vaatluseid!!!

```
> mudel0=lm(pikkus~1,             data=tudengid[!is.na(sport),])
> mudel1=lm(pikkus~factor(sport), data=tudengid[!is.na(sport),])
> anova(mudel0, mudel1)
```

H0: Lihtsam mudel prognoosib sama hästi kui keerukam

```
Analysis of Variance Table

Model 1: pikkus ~ 1
Model 2: pikkus ~ factor(sport)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    658 46503
2    654 43732  4    2771.3 10.361 3.851e-08 ***
```

Testimisest…

```
> anova(mudel0, mudel1)

Analysis of Variance Table

Model 1: pikkus ~ 1
Model 2: pikkus ~ factor(sport)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    658 46503
2    654 43732  4    2771.3 10.361 3.851e-08 ***
```

Alternatiivne interpretatsioon:

H_0: E(pikkus|sport=1) = E(pikkus|sport=2)
= E(pikkus|sport=3)
= E(pikkus|sport=4)
= E(pikkus|sport=5)

Sama test

```
> summary(mudel1)

[…]

Residual standard error: 8.177 on 654 degrees of freedom
Multiple R-squared:  0.05959,   Adjusted R-squared:  0.05384
F-statistic: 10.36 on 4 and 654 DF,  p-value: 3.851e-08
```

Kus peituvad erinevused?

(Mitmese testimise probleemi arvestav lahendus)

```
> mudel1=lm(pikkus~olu)
> summary(mudel1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 169.3549     0.4857 348.695  < 2e-16 ***
olu<1         0.7670     0.6875   1.116 0.264991
olu1-5        6.7734     0.9581   7.070 3.99e-12 ***
olu5-12       8.6784     1.5256   5.689 1.93e-08 ***
olu>13       10.8594     3.0331   3.580 0.000369 ***
```

Kus peituvad erinevused?

(Mitmese testimise probleemi arvestav lahendus)

```
> mudel1=lm(pikkus~relevel(olu, ref="<1"))
> summary(mudel1)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   170.1219     0.4866 349.616  < 2e-16 ***
relevel(olu, ref = "<1")ei joo  -0.7670     0.6875  -1.116 0.264991
relevel(olu, ref = "<1")1-5      6.0064     0.9585   6.266 6.71e-10 ***
relevel(olu, ref = "<1")5-12     7.9114     1.5259   5.185 2.88e-07 ***
relevel(olu, ref = "<1")>13     10.0924     3.0332   3.327 0.000926 ***
```

Kus peituvad erinevused?

(Mitmese testimise probleemi arvestav lahendus)

```
> mudel1=lm(pikkus~relevel(olu, ref="1-5"))
> summary(mudel1)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   176.1283     0.8258 213.270  < 2e-16 ***
relevel(olu, ref = "1-5")ei joo  -6.7734     0.9581  -7.070 3.99e-12 ***
relevel(olu, ref = "1-5")<1      -6.0064     0.9585  -6.266 6.71e-10 ***
relevel(olu, ref = "1-5")5-12     1.9051     1.6654   1.144    0.253
relevel(olu, ref = "1-5")>13      4.0860     3.1058   1.316    0.189
```

Kus peituvad erinevused?

(Mitmese testimise probleemi arvestav lahendus)

	ei joo	<1	1-5	5-12	>13
ei joo	NA	0,26499	0,00000	0,00000	0,00037
<1	0,00000	NA	0,00000	0,00000	0,00093
1-5	0,00000	0,00000	NA	0,25308	0,18876
5-12	0,00000	0,00000	0,00000	NA	0,51209
>13	0,00000	0,00037	0,00093	0,18876	NA

Korrigeerimata p-väärtused

	ei joo	<1	1-5	5-12	>13
ei joo	NA	1,00000	0,00000	0,00000	0,00369
<1	0,00000				
1-5	0,00000				
5-12	0,00000	0,00000	0,00000	NA	1,00000
>13	0,00000	0,00369	0,00926	1,00000	NA

SAAB PAREMINI!!!

Bonferroni meetodil korrigeeritud p-väärtused

```
> install.packages("multcomp")
> library(multcomp)
> mudel1=lm(pikkus~olu)
> abi=glht(aov(mudel1), linfct=mcp(olu="Tukey"))
> summary(abi)

         Simultaneous Tests for General Linear Hypotheses

Linear Hypotheses:
                    Estimate Std. Error t value Pr(>|t|)
<1 - ei joo == 0     0.7670     0.6875   1.116  0.77154
1-5 - ei joo == 0    6.7734     0.9581   7.070  < 1e-04 ***
5-12 - ei joo == 0   8.6784     1.5256   5.689  < 1e-04 ***
>13 - ei joo == 0   10.8594     3.0331   3.580  0.00263 **
1-5 - <1 == 0        6.0064     0.9585   6.266  < 1e-04 ***
5-12 - <1 == 0       7.9114     1.5259   5.185  < 1e-04 ***
>13 - <1 == 0       10.0924     3.0332   3.327  0.00647 **
5-12 - 1-5 == 0      1.9051     1.6654   1.144  0.75490
>13 - 1-5 == 0       4.0860     3.1058   1.316  0.64623
>13 - 5-12 == 0      2.1810     3.3249   0.656  0.95939
```

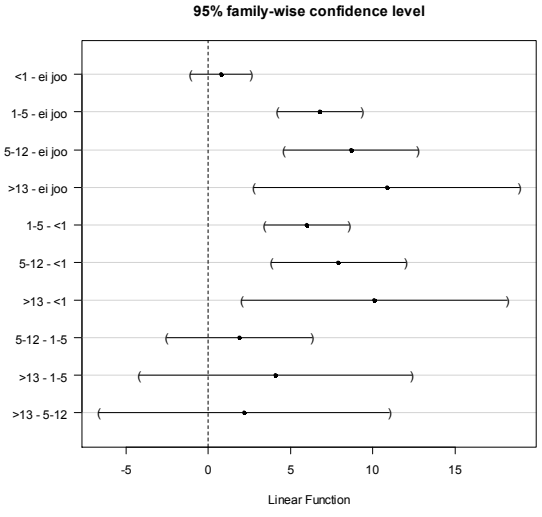
Bonferroni: 0,00369

```
> install.packages("multcomp")
> library(multcomp)
> mudel1=lm(pikkus~olu)
> abi=glht(aov(mudel1), linfct=mcp(olu="Tukey"))
> confint(abi)

95% family-wise confidence level

Linear Hypotheses:
                    Estimate lwr     upr
<1 - ei joo == 0     0.7670  -1.0626  2.5966
1-5 - ei joo == 0    6.7734   4.2237  9.3231
5-12 - ei joo == 0   8.6784   4.6185 12.7384
>13 - ei joo == 0   10.8594   2.7876 18.9312
1-5 - <1 == 0        6.0064   3.4555  8.5573
5-12 - <1 == 0       7.9114   3.8507 11.9722
>13 - <1 == 0       10.0924   2.0202 18.1646
5-12 - 1-5 == 0      1.9051  -2.5270  6.3371
>13 - 1-5 == 0       4.0860  -4.1792 12.3513
>13 - 5-12 == 0      2.1810  -6.6676 11.0295
```

```
> install.packages("multcomp")
> library(multcomp)
> mudel1=lm(pikkus~olu)
> abi=glht(aov(mudel1), linfct=mcp(olu="Tukey"))
> plot(abi)
```



Prognoosimine, usaldusintervallid ja prognoosiintervallid leitavad analoogselt regressioonanalüüsi mudelitega…

```
> predict(mudel1, newdata=data.frame(olu="<1"))
       1
170.1219


> predict(mudel1, newdata=data.frame(olu="<1"),
                                    interval="confidence")
       fit      lwr      upr
1 170.1219 169.1664 171.0774


> predict(mudel1, newdata=data.frame(olu="<1"),
                                    interval="prediction")
       fit      lwr      upr
1 170.1219 154.5385 185.7053
```

Dispersioonanalüüsi eeldused

Dispersioonanalüüsi mudeli hindamisel ja hüpoteeside testimisel tehakse samu eelduseid, mis regressioonmudeli hindamisel ja testimisel.

Mudeli jäägid peavad olema normaaljaotusega (muidu arvutab arvuti **prognoosiintervallid**, olulisustõenäosused ja usaldusintervallid valesti välja);

Uuritava tunnuse hajuvus peab iga faktortunnuse taseme korral olema samasuur (prognoosiintervallid, testid, usaldusintervallid muidu valed)

Valim esindav, sisestusvigu pole.

Kui eeldused pole täidetud, võib proovida samu lahendusteid, mida regressioonanalüüsigi korral – näiteks uuritava tunnuse transformeerimist (logaritmimist).