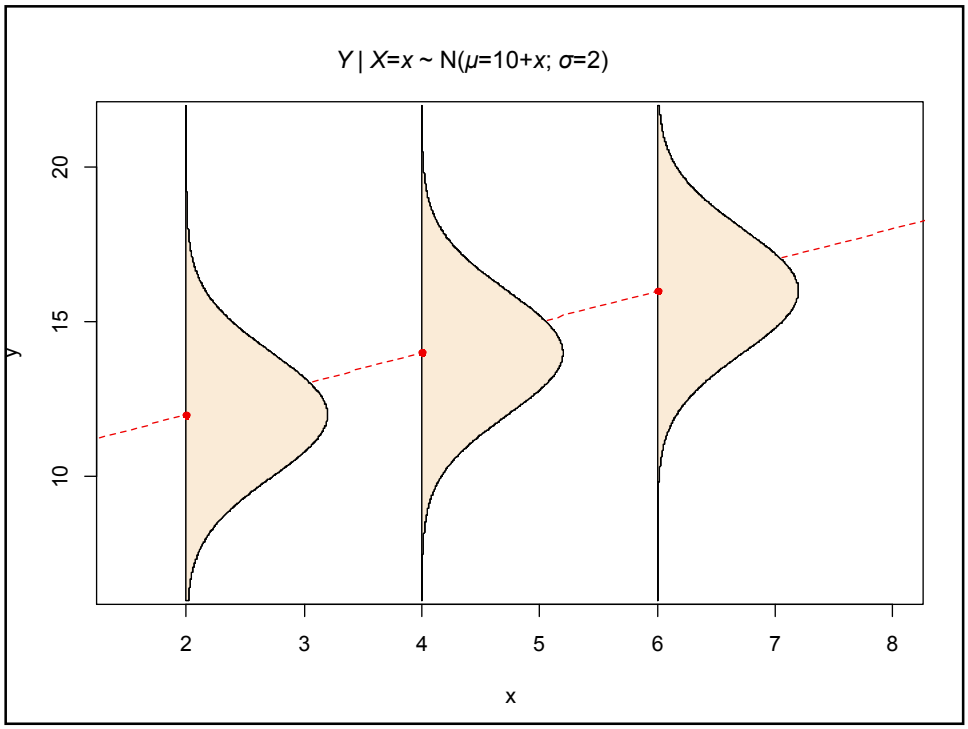
Biomeetria

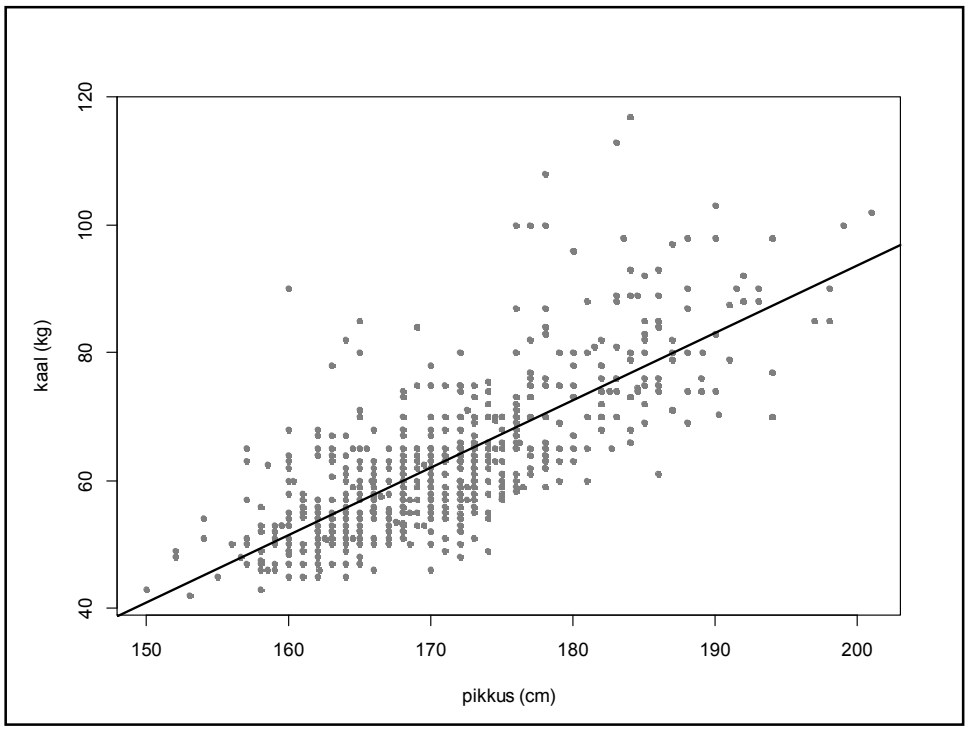
2. loeng

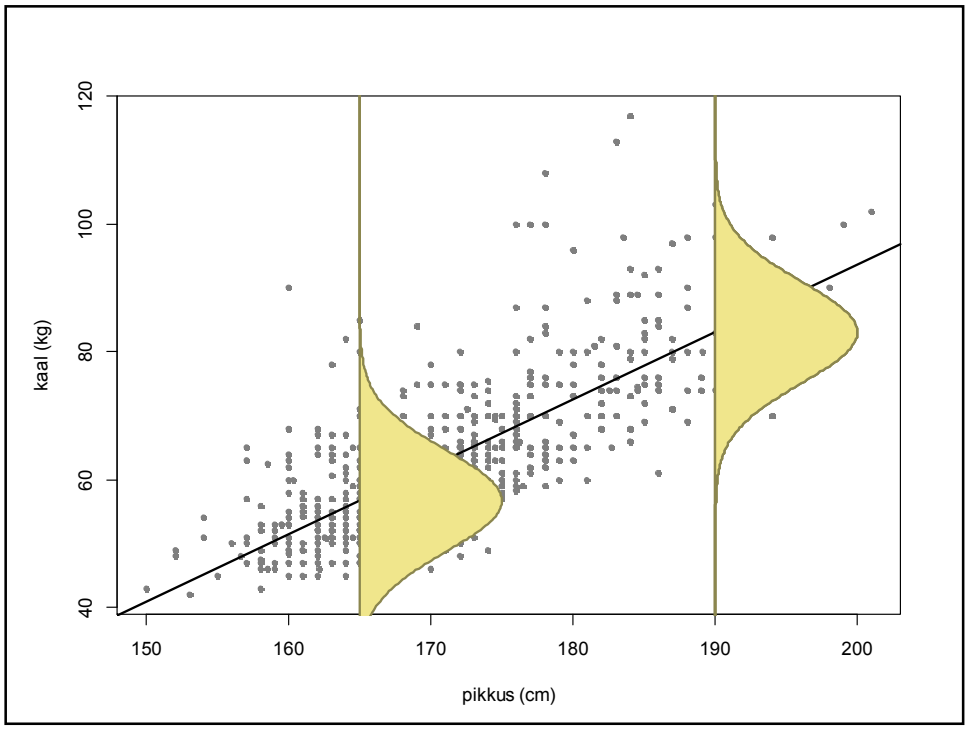
Lihtne lineaarne regressioon

mudeli hindamisest; usaldusintervall;
prognoosiintervall; determinatsioonikordaja; …

Märt Möls
martm@ut.ee




kaal (kg)
40
60
80
100
120
150
160
170
180
190
200
pikkus (cm)


kaal (kg)
40
60
80
100
120
150
160
170
180
190
200
pikkus (cm)

Regressioonanalüüsi mudel(id)

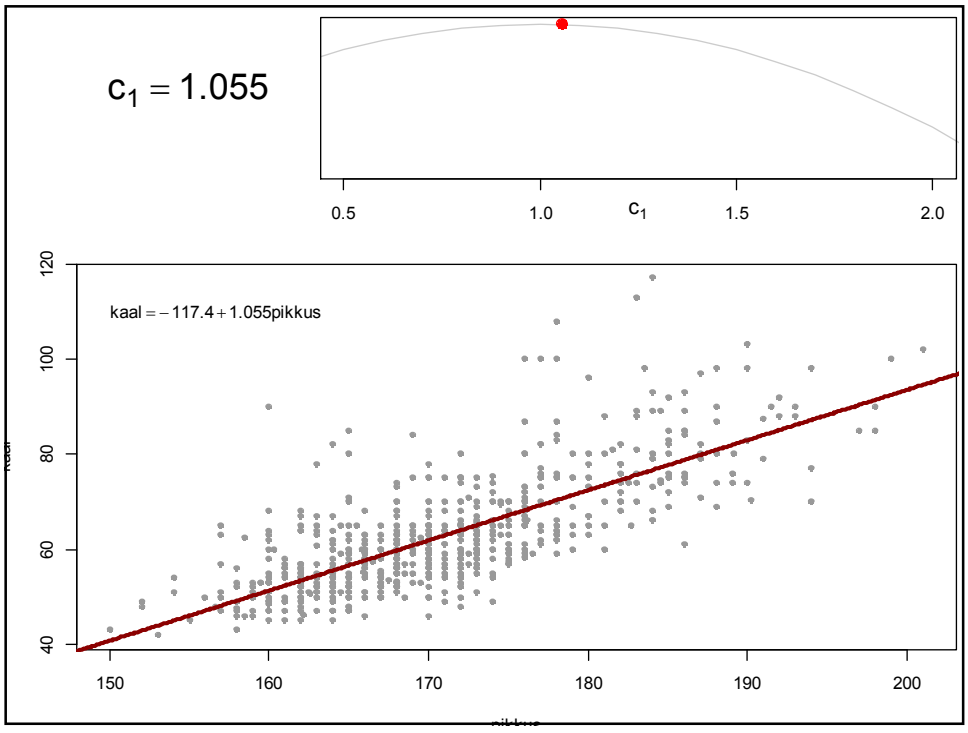
$Y \mid X=x \sim N(\mu=c_0+c_1x;\ \sigma=\sigma_\varepsilon)$

$Y \sim N(\mu=c_0+c_1x;\ \sigma=\sigma_\varepsilon)$

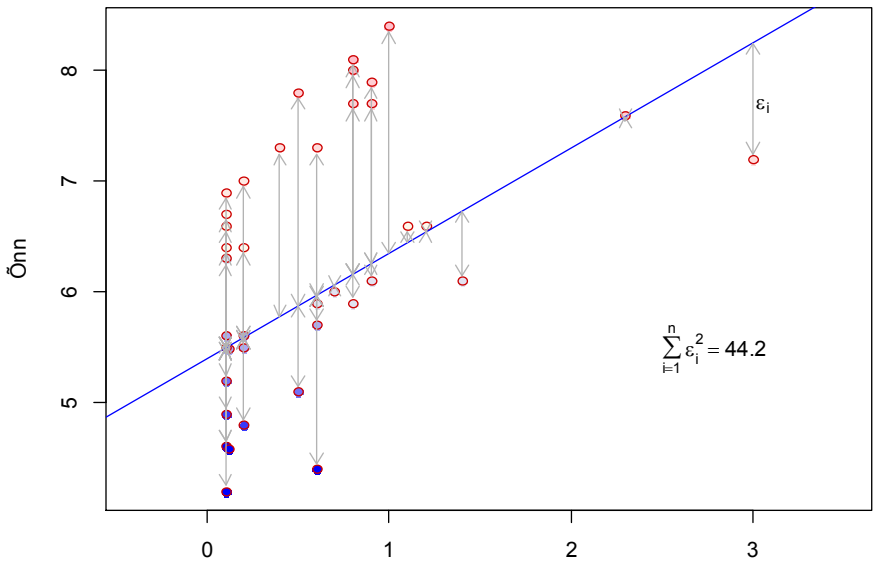
$Y = c_0+c_1x + \varepsilon; \quad \varepsilon \sim N(0;\ \sigma_\varepsilon)$

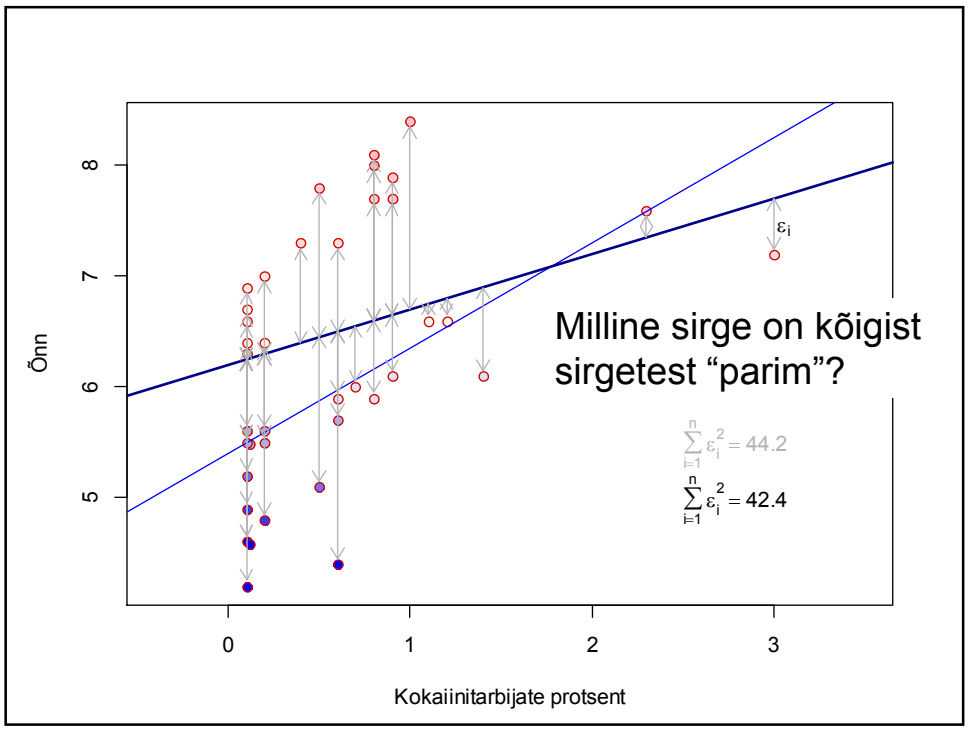
$EY = c_0+c_1x$

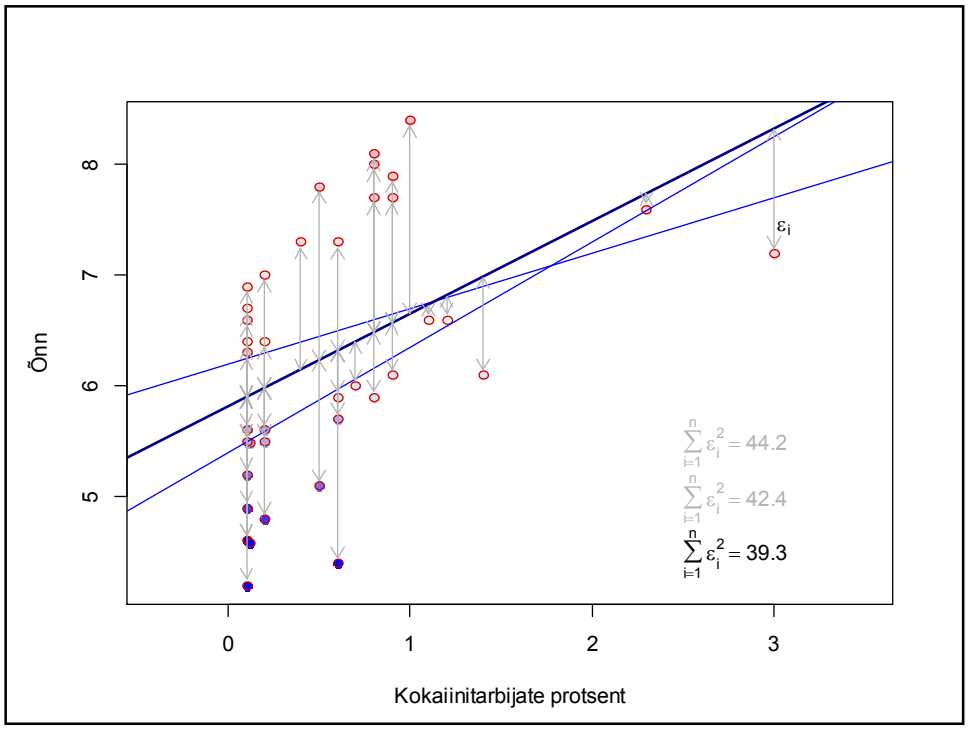
c_0, c_1 väärtused hinnatakse suurima tõepära meetodi abil



Alternatiivne interpretatsioon




Õnn
5
6
7
8
0
1
2
3
Kokaiinitarbijate protsent
ε_i
Milline sirge on kõigist sirgetest “parim”?
$\sum_{i=1}^{n} \varepsilon_i^2 = 44.2$
$\sum_{i=1}^{n} \varepsilon_i^2 = 42.4$


Õnn
5
6
7
8
0
1
2
3
Kokaiinitarbijate protsent
ε_i
$\sum_{i=1}^{n} \varepsilon_i^2 = 44.2$
$\sum_{i=1}^{n} \varepsilon_i^2 = 42.4$
$\sum_{i=1}^{n} \varepsilon_i^2 = 39.3$

Sama sirge…

- $Y \sim N(\mu=c_0+c_1x;\ \sigma=\sigma_\varepsilon)$, sõltumatud vaatlused. otsime c_0 ja c_1 väärtuseid mille puhul nähtud andmete saamise tõenäosus oleks kõige suurem;
- Otsime sirget $y=c_0+c_1x$, mille puhul prognoosivigade ruutude summa oleks minimaalne;
- $EY=c_0+c_1x$, $DY=\sigma_\varepsilon$, vaatlused sõltumatud. Otsime kõige täpsemat (lineaarset) nihketa hinnangut c_0+c_1x –le

Kas (lineaarne) seos Y ja X vahel eksisteerib?

Kas

$Y \sim N(\mu=c_0+c_1x;\ \sigma=\sigma_\varepsilon)$

või

$Y \sim N(\mu=c_0\ ;\ \sigma=\sigma_\varepsilon)$?

Mida prognoosime

Mille abil prognoosime

```
> mudel=lm(kaal~pikkus)
> summary(mudel)

[...]

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.36675    6.33019  -18.54   <2e-16 ***
pikkus         1.05474    0.03695   28.54   <2e-16 ***

[...]
```

Kas c_0 võib olla 0?

Kas c_1 võib olla 0?

Mida prognoosime

Mille abil prognoosime

```
> mudel=lm(kaal~pikkus)
> summary(mudel)

[...]

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.36675    6.33019  -18.54   <2e-16 ***
pikkus         1.05474    0.03695   28.54   <2e-16 ***

[...]
```

Kas c_0 võib olla 0?

Kas c_1 võib olla 0?

```
> confint(mudel)
                 2.5 %      97.5 %
(Intercept) -129.7967014 -104.936795
pikkus         0.9821798    1.127295
```

95% usaldusintervallid

```
> predict(mudel, newdata=data.frame(pikkus=170))
       1
61.93859
```

170cm pikkuste inimeste kaalude keskväärtus (hinnang)

```
> predict(mudel, newdata=data.frame(pikkus=170),
   interval="confidence")
       fit     lwr      upr
1 61.93859 61.3231 62.55408
```

Kui täpselt me teame 170cm pikkuste inimeste kaalude keskväärtust?

```
> predict(mudel, newdata=data.frame(pikkus=170),
   interval="prediction")
       fit      lwr      upr
1 61.93859 46.29655 77.58063
```

Kui täpselt me teame 170cm pikkuse inimese (Jaani) kaalu?

```
> install.packages("gmodels")
> library(gmodels)
> estimable(mudel, c(1,170))
        Estimate Std. Error t value  DF Pr(>|t|)
(1 170) 61.93859  0.3134512 197.602 654        0
> estimable(mudel, c(1,170), conf.int=0.95)
       Estimate Std. Error t value  DF Pr(>|t|) Lower.CI Upper.CI
(1 170) 61.93859  0.3134512 197.602 654        0  61.3231 62.55408
```

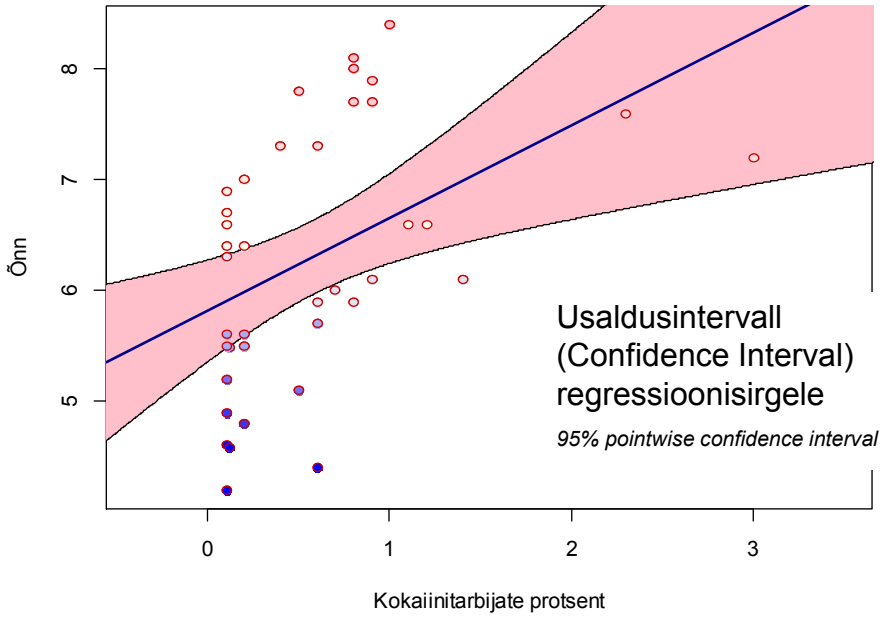
Estimable -käsust

Võimalik kasutada keerukamate mudelite juures;

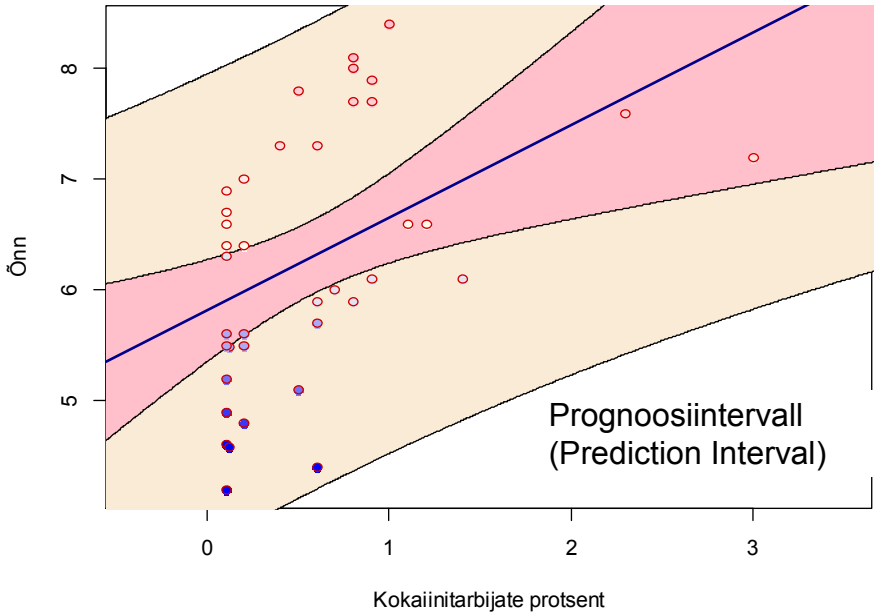
Võimalik kasutada keerukamate küsimuste jaoks (milline on 170cm ja 160cm pikkuste tudengite keskmiste kaalude erinevus?)

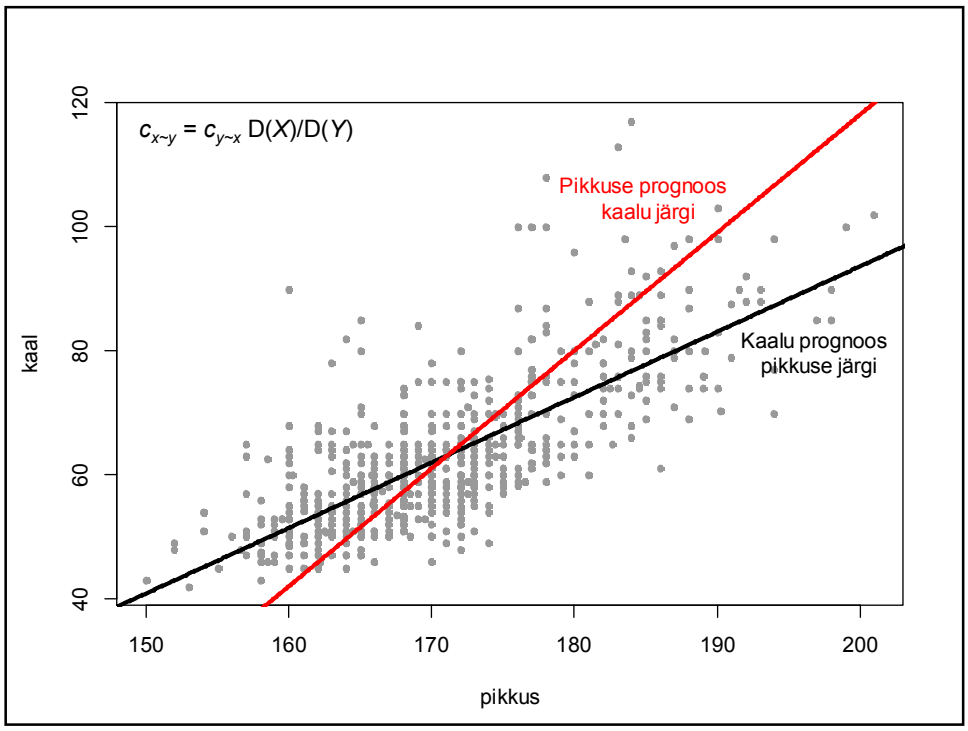
```
> estimable(mudel, c(1-1,170-160), conf.int=0.95)
       Estimate Std. Error  t value  DF Pr(>|t|) Lower.CI Upper.CI
(0 10) 10.54737   0.369513 28.54398 654        0 9.821798 11.27295
```

Kui täpselt me teame parimat sirget?

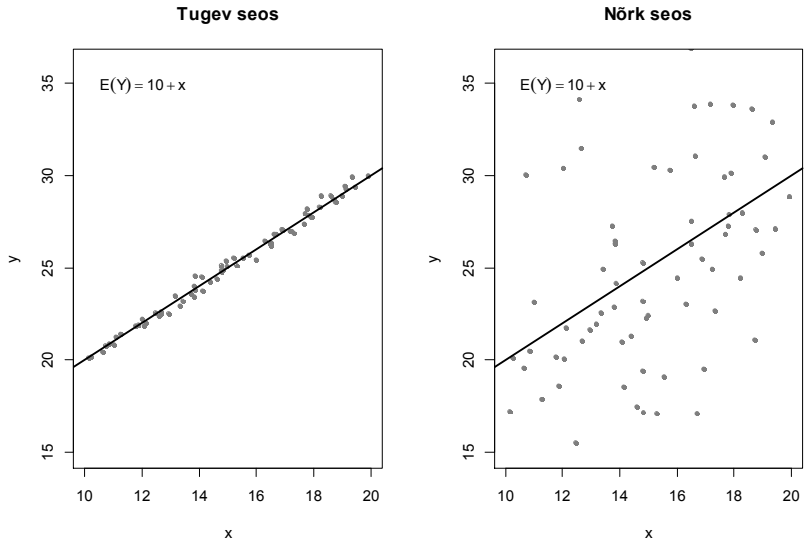


Kui täpne on meie prognoos?





Seose tugevuse / mudeli prognoosivõime iseloomustamine



```
> summary(mudel)

Residuals:
    Min      1Q  Median      3Q     Max
-17.814  -5.361  -1.220   3.499  40.295

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.36675    6.33019  -18.54   <2e-16 ***
pikkus         1.05474    0.03695   28.54   <2e-16 ***


Residual standard error: 7.96 on 654 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5547,    Adjusted R-squared:  0.554
F-statistic: 814.8 on 1 and 654 DF,  p-value: < 2.2e-16
```

Tegelik – Prognoos
(tegelik kaal – pikkuse järgi prognoositud kaal)

```
Residual standard error: 7.96

> predict(mudel, data.frame(pikkus=170))
       1
61.93859

> predict(mudel, data.frame(pikkus=170))-2*7.96
       1
46.01859

> predict(mudel, data.frame(pikkus=170))+2*7.96
       1
77.85859

> predict(mudel, data.frame(pikkus=170),
                                    interval="prediction")
       fit      lwr      upr
1 61.93859 46.29655 77.58063
```

Determinatsioonikordaja

Mudel 1: $Y = c_0 + \varepsilon_1$

Mudel 2: $Y = c_0 + c_1 x + \varepsilon_2$

$D\varepsilon_1$ - iseloomustab lihtsama mudeli prognooside täpsust

$D\varepsilon_2$ - iseloomustab keerukama mudeli prognoositäpsust

$D\varepsilon_1 - D\varepsilon_2$ - täpsuse suurenemine tänu x-tunnuse teadmisele

$(D\varepsilon_1 - D\varepsilon_2)/D\varepsilon_1$ - suhteline võit prognoositäpsuses

$$R^2_{tegelik} = (D\varepsilon_1 - D\varepsilon_2)/D\varepsilon_1$$

$Y = c_0 + \varepsilon_1$

$D(Y) = D(c_0 + \varepsilon_1) = D(\varepsilon_1)$

$$R^2_{tegelik} = (DY - D\varepsilon_2)/DY$$

R^2 ja $R^2_{adjusted}$ (kohandatud R^2)

$$R^2_{tegelik} = (DY - D\varepsilon_2)/DY$$

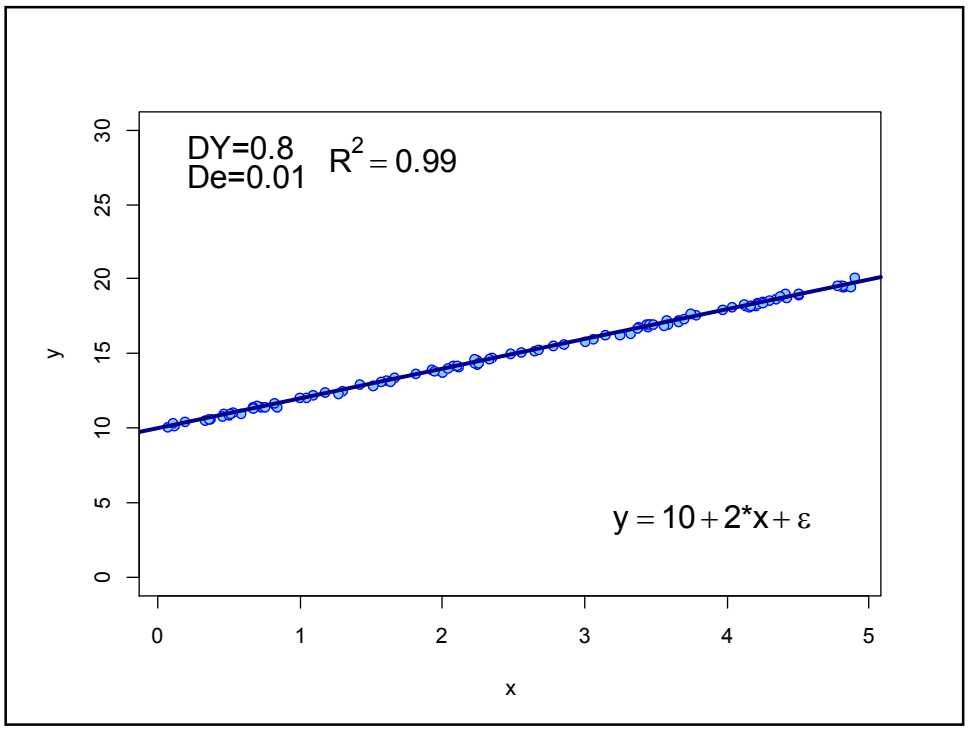
Vaja hinnata. DY hidamisega probleeme pole, $D\varepsilon_2$ hindamisega küll.

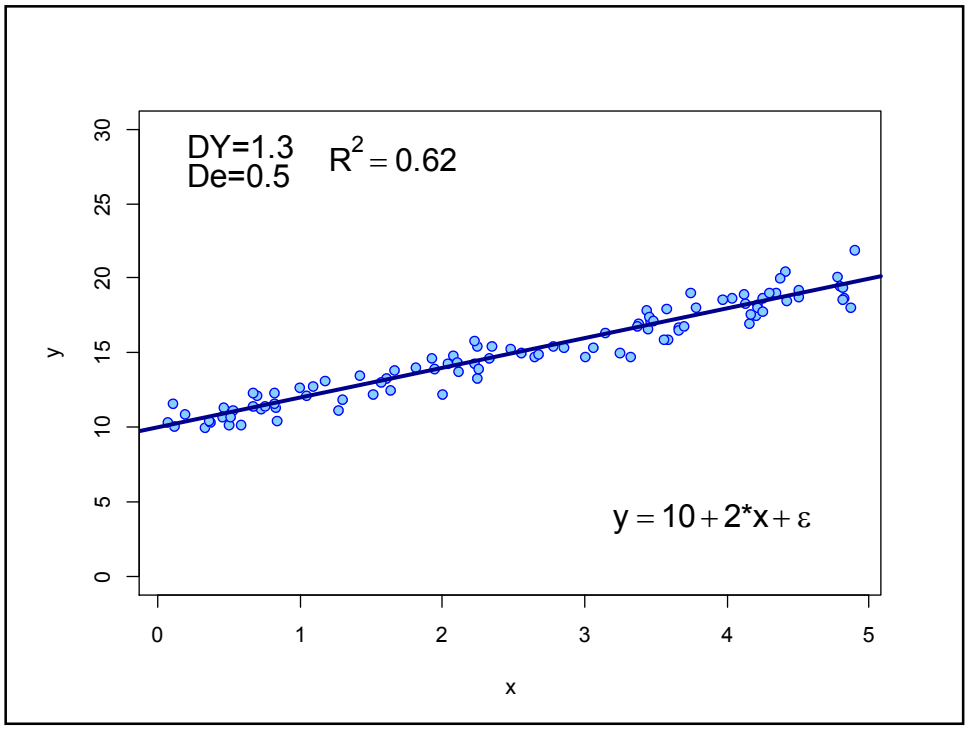
Kasutades nihkega (“vale”) hinnangut $D\varepsilon_2$–le, saame determinatsioonikordaja R^2.

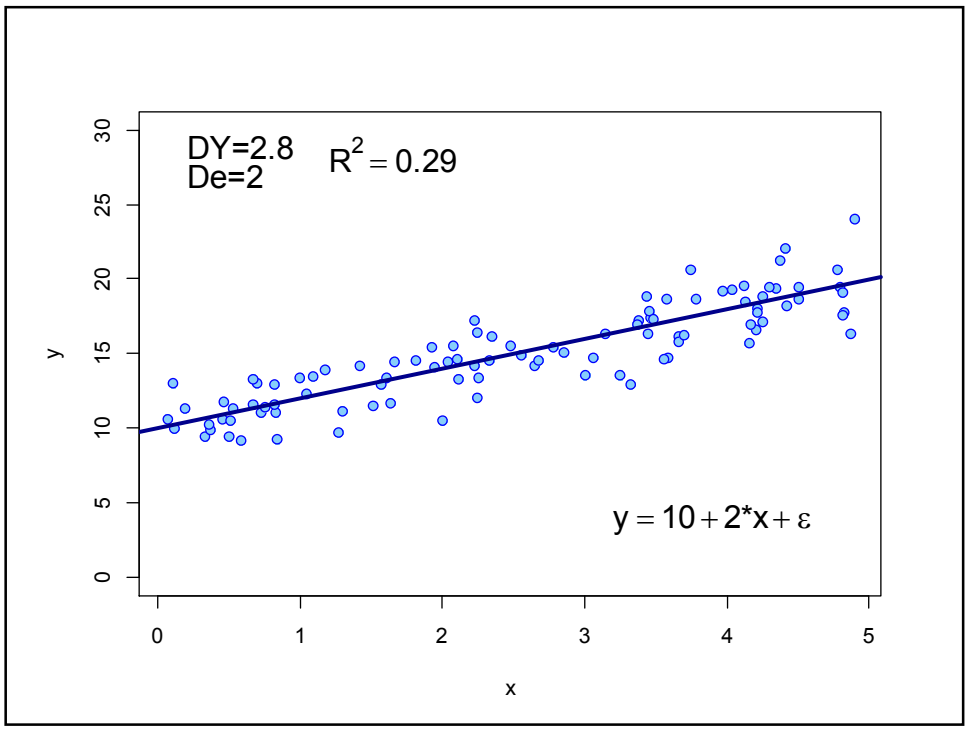
Kasutades nihketa (“õiget”) hinnangut $D\varepsilon_2$–le, saame kohandatud determinatsioonikordaja $R^2_{adjusted}$.

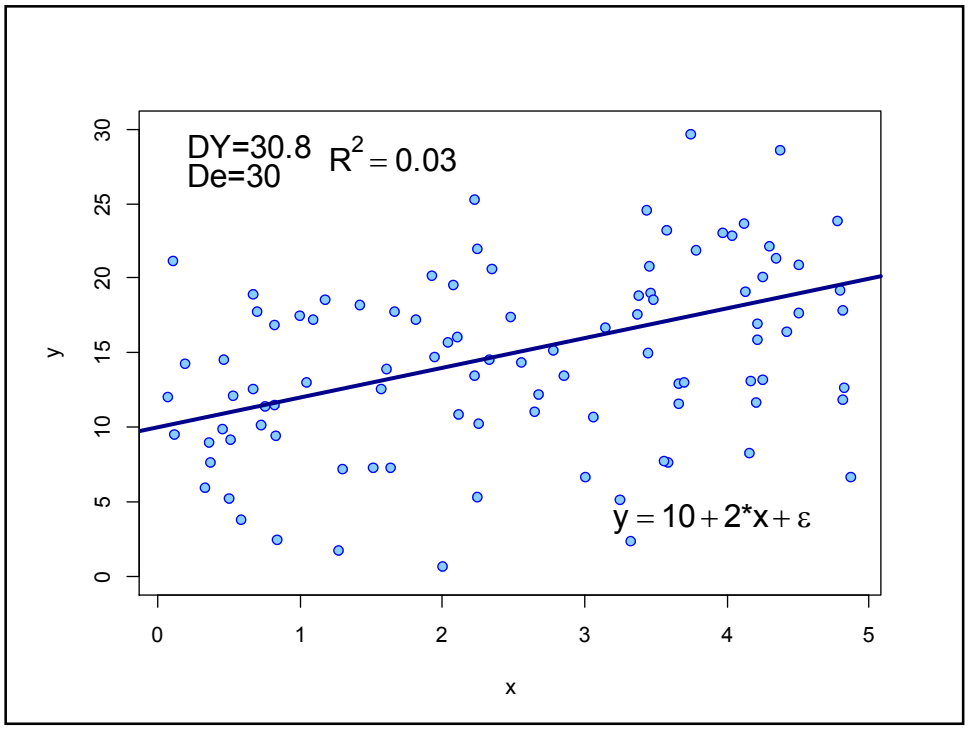
```
> summary(lm(kaal~pikkus))
[...]
Residual standard error: 7.96 on 654 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5547,    Adjusted R-squared:  0.554

> Dkaal=var(kaal, na.rm=TRUE)
> (Dkaal-7.96**2)/Dkaal
[1] 0.5540274
```


DY=0.8
De=0.01
$R^2 = 0.99$
$y = 10 + 2*x + \varepsilon$
y
x
0
5
10
15
20
25
30
0
1
2
3
4
5


DY=1.3
De=0.5
$R^2 = 0.62$
$y = 10 + 2*x + \varepsilon$
y
x
0
5
10
15
20
25
30
0
1
2
3
4
5

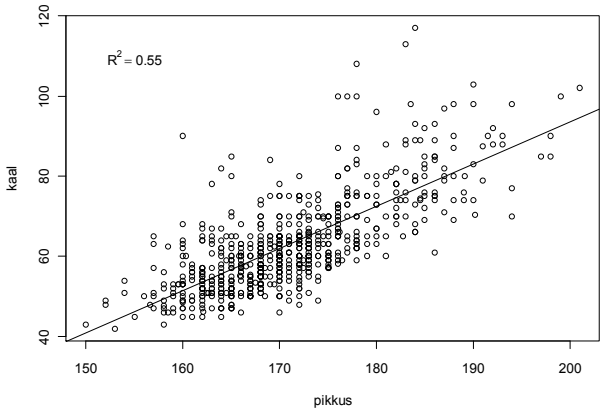

DY=2.8
De=2
$R^2 = 0.29$
$y = 10 + 2*x + \varepsilon$
y
x
0
5
10
15
20
25
30
0
1
2
3
4
5


DY=30.8
De=30
$R^2 = 0.03$
$y = 10 + 2*x + \varepsilon$
y
x
0
5
10
15
20
25
30
0
1
2
3
4
5

Determinatsioonikordajaga manipuleerimine

Determinatsioonikordaja on eelkõige interpreteeritav siis, kui uuritav valim on tõepoolest juhuslik valim mingist populatsioonist. Kui uuritavad andmed on kogutud eksperimenteerides (olukorras, kus me ise otsustame, millised saavad olema *X*-tunnuse väärtused) on determinatsioonikordaja R^2 teatavates piirides eksperimentaatori enda valida/otsustada.

Kaalu prognoosimine pikkuse järgi. Juhuslik valim: R^2=0,55



Kuidas suurendada või vähendada determinatsioonikordajat?

$$KAAL = c0 + c1\ PIKKUS + e.$$

Determinatsioonikordaja

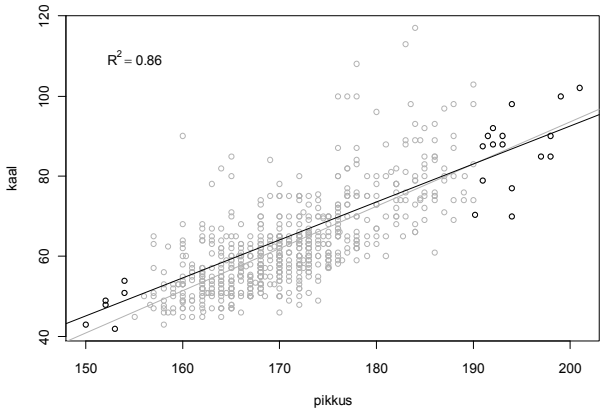
$$R^2 = 1 - D(e)/D(KAAL).$$

Vaja oleks kas muuta D(*e*) või D(*KAAL*) väärtust. Antud näites on lihtsam muuta D(*KAAL*) väärtust:

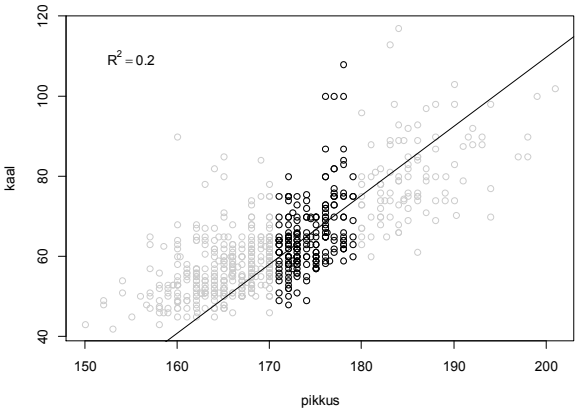
$$D(KAAL) = c_1^2\ D(PIKKUS) + D(e)$$

Valides uuringusse väga erinevate pikkustega tudengeid saame suure R^2 väärtuse, valides sarnase pikkusega tudengeid saame väikese R^2 väärtuse.

Pikkus<155cm või pikkus >190cm R^2=0,86



Pikkus>170cm ja pikkus <180cm
R^2=0,2




```
> mudelMees=lm(kaal~pikkus, data=kokku2[sugu==2,])
> summary(mudelMees)
Multiple R-squared:  0.3704,    Adjusted R-squared:  0.3661
```

```
> mudelNaine=lm(kaal~pikkus, data=kokku2[sugu==1,])
> summary(mudelNaine)
Multiple R-squared:  0.2968,    Adjusted R-squared:  0.2954
```

```
> mudel=lm(kaal~pikkus, data=kokku2)
> summary(mudel)
Multiple R-squared:  0.5547,    Adjusted R-squared:  0.554
```