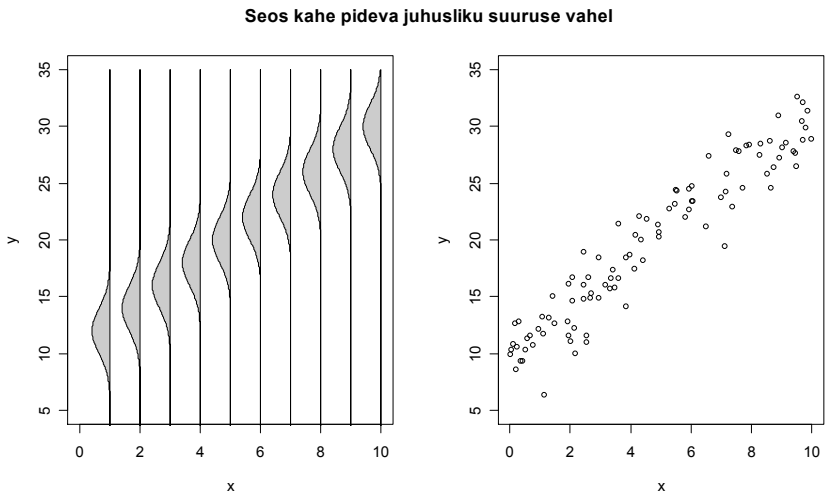
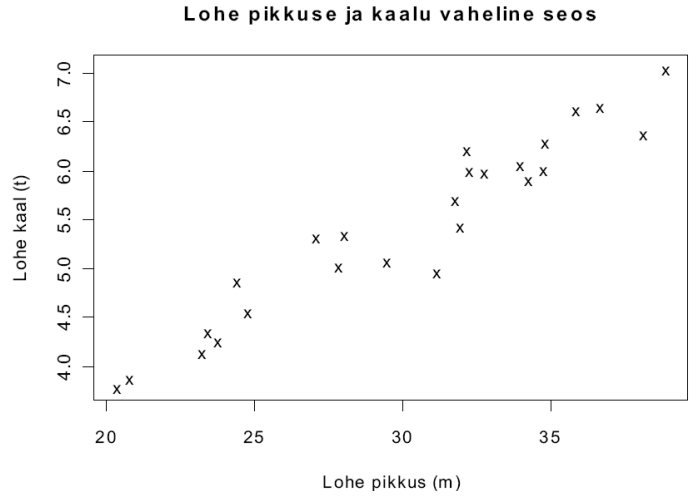
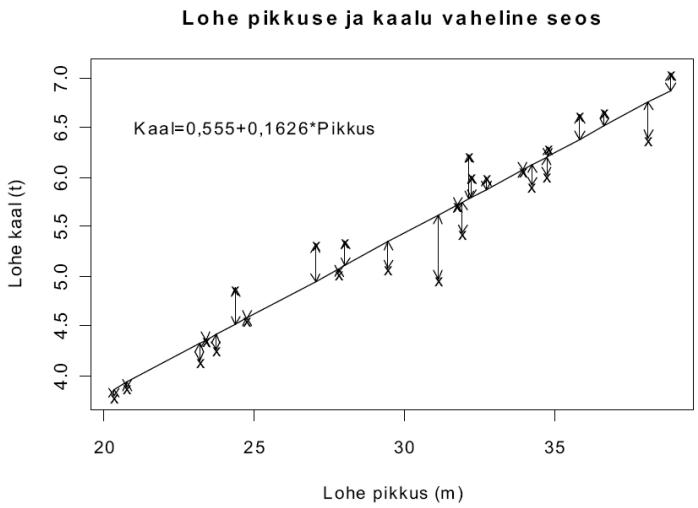
Lihtne lineaarne regressioonanalüüs

Biomeetria, 6. loeng



Juhul, kui uuritav tunnus on normaaljaotusega, piisab uuritava tunnuse jaotuse määramiseks kui teame, millised on uuritava tunnuse keskväärtus ja dispersioon. Eeldame esialgu, et uuritava tunnuse hajuvus (dispersioon) ei muutu katsetingimuste muutudes. Sellisel juhul piisab uuritava tunnuse jaotuse (muutumise) kirjeldamiseks, kui suudame kirjeldada, kuidas muutub uuritava tunnuse keskväärtus.





Miks minimiseerime vigade ruutude summat…

Olgu vaatlused $y_1, y_2, y_3, \ldots$.

Tahame leida sellist arvu c, nii et prognoosivead oleksid võimalikult väikesed – täpsemalt: prognoosivigade ruutude summa oleks minimaalne:

$$\min \sum_{i=1}^{n} (y_i - c)^2$$

Isegi mittematemaatikul võiks keskkoolist varuks olla piisavalt matemaatikat, et vastust leida:

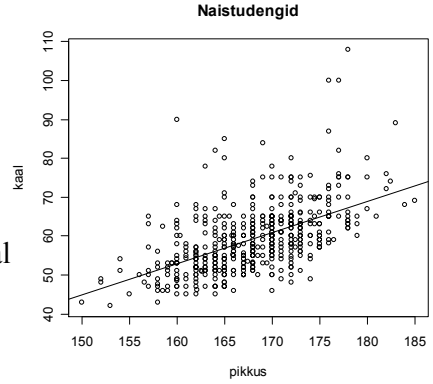
$$c = \frac{1}{n} \sum_{i=1}^{n} y_i \quad (= \bar{x})$$

Veel üks näide

Naistudengite kaalu kirjeldab ligikaudu järgmine mudel

kaal = 0.8 * pikkus – 75 + juhuslik viga

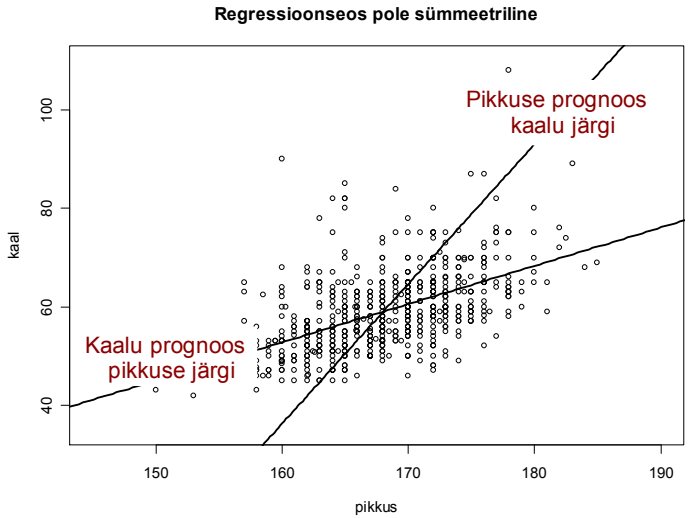
Antud võrrand on hinnanguks regressioonisirgele, mida võib graafiliselt kujutada lisatuna hajuvusgraafikule. Pikkuse kordaja ehk regressioonisirge tõus on siin 0,8. Kui kahe inimese pikkus erineb ühe ühiku võrra (1cm) siis nende kaal erineb keskmiselt 0,8 ühiku võrra. Vabaliige on -75.



Regressioonseos ei ole sümmeetriline. Seoses olevate tunnuste rollid on:

Funktsioontunnus ehk sõltuv tunnus (dependent variable), tähistus Y – tunnus, mille jaotust (keskväärtust) mudel kirjeldab (prognoosib). Eelnenud näites oli selleks tunnuseks kaal.

Argumenttunnus ehk sõltumatu tunnus (independent variable), tähistus X – tunnus, mille abil funktsioontunnuse jaotust kirjeldatakse. Eelnevas näites oli selleks pikkus.



Lihtsa regressioonimudeli võime üles kirjutada kujul

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

kus ε-juhuslik viga; β_0 - vabaliige; β_1 - argumenttunnuse kordaja. Juhusliku vea keskväärtus peab olema 0: Eε = 0. Y tunnuse keskväärtus etteantud X korral on siis:

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

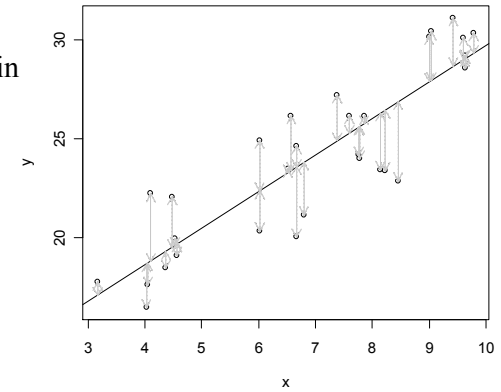
Naistudengite kaalu jaoks tehtud mudelis β_0=-75 ja β_1=0.8. Seda mudelit kasutades leiame 170cm pikkuste tudengite keskmise kaalu:

E (kaal| pikkus=170cm) = -75 + 0.8*170 = 61 (kg)

Kuidas hinnata regressioonmudeli parameetreid – vabaliiget ja argumenttunnuse kordajat?

Üks võimalus on kasutada vähimruutude meetodit – valida mudeli parameetrid selliselt, et mudeli jääkide – prognoosivigade – ruutude summa oleks minimaalne:

$$\sum_i \varepsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 \to \min$$



Regressioonsirge hindamine programmi abil (R näide)

```
> lm(kaal~pikkus)

Call:
lm(formula = kaal ~ pikkus)

Coefficients:
(Intercept)       pikkus
   -76.7846       0.8044
```

Seega on naistudengite kaalu prognoosiv mudel täpsemalt kirja pandav kui:

kaal = -76,78 + 0,80*pikkus + ε.

Regressioonmudeli abil prognoosi leidmine (R näide)

```
> m1=lm(kaal~pikkus)
> m1
Coefficients:
(Intercept)       pikkus
   -71.6788       0.7779
> predict(m1, data.frame(pikkus=167))
[1] 58.22937
```

Võrdluseks:
E(kaal | pikkus=167) = -71,68 + 0,7779*167 = 58,2293

Statistilised testid regressioonanalüüsis, usaldusintervall

Regressioonmudeli hindamise käigus saame ühtlasi kontrollida, kas vaadeldud kahe tunnuse vahel üleüldse eksisteerib lineaarset seost.

H_0: $\beta_1 = 0$
H_1: $\beta_1 <> 0$

Nende hüpoteeside testimiseks hinnatakse regressiooniparameetri β_1 hinnangu standarhälve. Regressioonparameetri hinnangu jaotuseks on normaaljaotus ja suurus $\hat{\beta}_1 / s(\hat{\beta}_1)$ peab nullhüpoteesi kehtides olema t-jaotusega juhuslik suurus. Ligikaudne otsus: kui see suhe tuleb suurem kui 2 või väiksem kui -2, siis nullhüpotees ei saa kehtida.

Märkus: Numbri 2 asemel oleks täpsem kasutada *t*-jaotuse kvantiili $t_{0,975;\ n-k}$, kus *n* on vaatluste arv andmestikus ja *k* regressioonmudeli (tundmatute) parameetrite arv (kaalu mudelis *k*=2).

Hüpoteeside kontroll regressiooniparameetrite kohta R-is:

```
> m1=lm(kaal~pikkus)
> summary(m1)

[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.67878    8.40428  -8.529   <2e-16 ***
pikkus        0.77789    0.04999  15.560   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[...]
```

Kas vabaliige võib olla 0?

Kas sirge tõus võib olla 0? Kas (lineaarset) seost pikkuse ja kaalu vahel eksisteerib?

Märkus: Soovitatav on vabaliige jätta mudelisse ka siis, kui ta pole statitiliselt oluline (mudeli vabaliige võib olla ka 0). Erandjuhuks oleks olukord, kus me tõepoolest (põhjendatult) soovime, et regressioonsirge läbiks punkti (0,0).

Kuna regressiooniparameetri β_1 hinnang $\hat{\beta}_1$ on normaaljaotusega juhuslik suurus (suure valimi või normaaljaotusega vaatluste korral), siis võime talle konstrueerida usaldusintervalli täpselt samal viisil kui konstrueerisime usaldusintervalli keskväärtusele. Tulemus: 95%-usaldusintervall regressiooniparameetrile β_1 on leitav valemist:

$$\hat{\beta}_1 + t_{0,025;n-k} s(\hat{\beta}_1) ... \hat{\beta}_1 + t_{0,975;n-k} s(\hat{\beta}_1),$$

Märkus: samal viisil on võimalik leida usaldusintervall mudeli vabaliikmele ja samuti saame testida, kas mudeli vabaliige võib olla 0. Siiski ei soovita vabaliiget mudelist välja jätta ka juhul, kui mudeli vabaliige võiks olla 0. Erandjuhuks oleks olukord, kus me tõepoolest (põhjendatult) soovime, et regressioonsirge läbiks punkti (0,0).

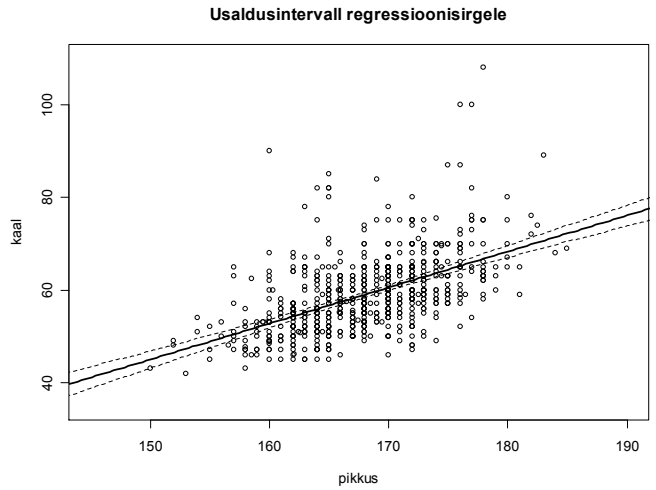
Usaldusintervall I – usaldusintervall parameetritele

```
> m1=lm(kaal~pikkus)
> summary(m1)

[...]

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.67878    8.40428  -8.529   <2e-16 ***
pikkus        0.77789    0.04999  15.560   <2e-16 ***

[...]
```

Hinnangu standardviga. Ligikaudse 95%-usaldusintervalli leidmiseks liida ja lahuta parameetri hinnangust 2 standardviga – st ligikaudne usaldusintervall sirge tõusule oleks 0,78+-2*0,05 = [0,68..0,88]

Täpsed usaldusintervallid (kasutades t-jaotuse kvantiile) on R-is leitavad käsuga confint

```
> confint(m1)
                 2.5 %      97.5 %
(Intercept) -88.1819799 -55.1755749
pikkus        0.6797251   0.8760611
```



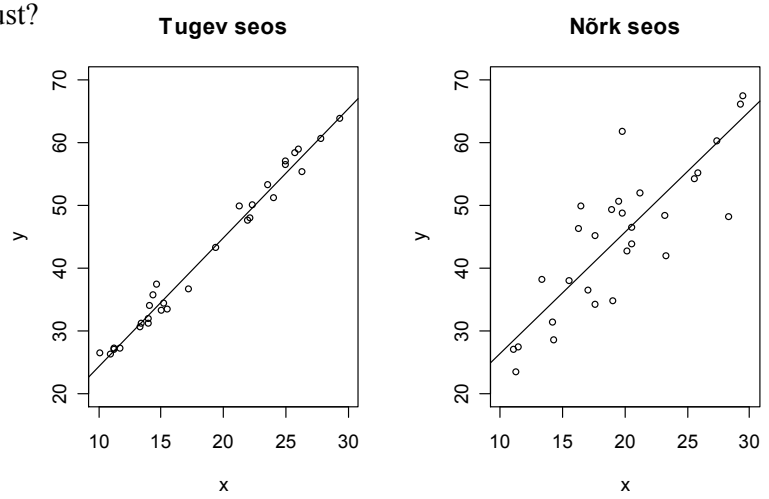
Usaldusintervall II – usaldusintervall keskväärtusele

167cm pikkuste naistudengite keskmine kaal on hinnanguliselt 58,2kg. Kui täpselt me ikkagi nende keskmist kaalu teame?

```
> m1=lm(kaal~pikkus)
> m1
Coefficients:
(Intercept)       pikkus
   -71.6788       0.7779
> predict(m1, data.frame(pikkus=167))
[1] 58.22937

> predict(m1, data.frame(pikkus=167),
                                interval="confidence")
          fit      lwr      upr
[1,] 58.22937 57.6612 58.79754
```

Seose tugevuse iseloomustamine

Mõnikord on üht tunnust teades võimalik väga suure täpsusega ära arvata teise tunnuse värtust, teinekord mitte. Kuidas mõõta seose tugevust?



Kuidas mõõta meie ennustuste headust - seose tugevust? Juhul, kui me poleks mõõtnud sõltumatu tunnuse (X) väärtuseid, oleks parim prognoos, mida me sõltuvale tunnusele (Y) anda suudaksime, sõltuva tunnuse väärtuste keskmine. Sellise "prognoosi" täpsust saab iseloomustada "ennustusvigade" dispersiooni abil ehk tunnuse Y dispersiooni kasutades (tähistame tunnuse Y dispersiooni tähega $D(Y)$). Kasutades sõltuva tunnuse (Y) prognoosimiseks sõltumatu tunnuse (X) väärtuseid, saame teha täpsemaid prognoose. Võime taas mõõta oma prognoosi täpsust, kasutades selleks mõõtmisvigade dispersiooni ($D(\varepsilon)$). Järelikult teades tunnuse X väärtust, on meil võimalik prognoosida tunnuse Y väärtust $D(Y)$-$D(\varepsilon)$ võrra täpsemalt (väiksema dispersiooniga). Prognoosi täpsuse suurenemise jagatist prognoositava tunnuse dispersiooniga kutsutakse determinatsioonikordajaks (R^2) ja seda kasutatakse sageli tunnustevahelise seose tugevuse mõõtmiseks:

$$R^2 = [\, D(Y) - D(\varepsilon) \,] / D(Y)$$

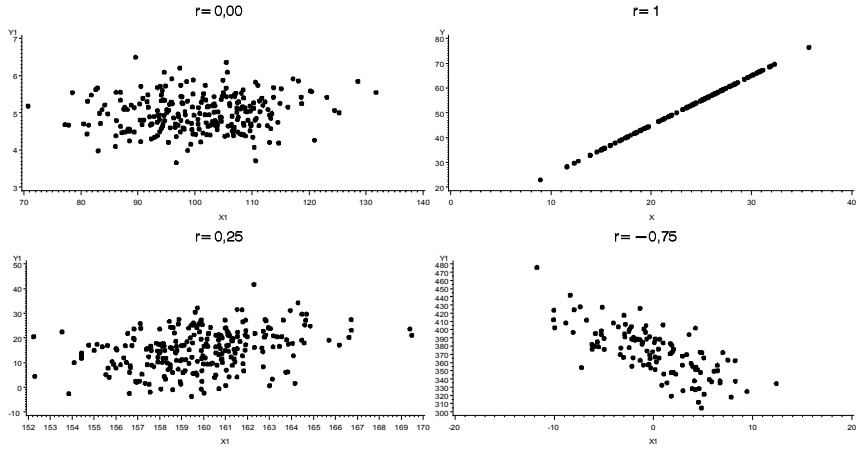
Determinatsioonikordajat esitatakse vahel ka protsentides – näidates, mitu protsenti õnnestus tänu kasutatavale regressioonimudelile tõsta prognoosi täpsust.

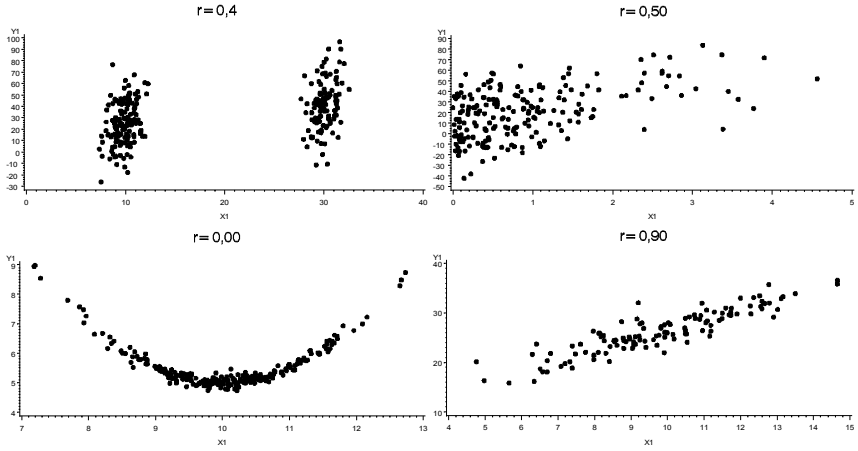
Lineaarne korrelatsioonikordaja

Arvatavasti kõige sagedamini kasutatakse kahe tunnuse vahelise seose tugevuse iseloomustamiseks lineaarset korrelatsioonikordajat (Pearsoni korrelatsioonikordajat). Korrelatsioonikordaja ruut on determinatsioonikordaja, kusjuures lineaarne korrelatsioonikordaja on positiivne, kui ühe tunnuse väärtuste kasvades teise tunnuse väärtused kipuvad samuti kasvama. Kui aga ühe tunnuse väärtuste kasvades teise tunnuse väärtused kipuvad kahanema siis on korrelatsioonikordaja negatiivne. Lineaarset korrelatsioonikordajat tähistatakse tähega r.

Pearsoni korrelatsioonikordaja omadused:

- Kui tunnuste X ja Y vahel on lineaarne funktsionaalne seos $Y=a+bX$ (ehk täpne lineaarne seos), siis on korrelatsioonikordaja väärtus kas 1 või -1 vastavalt kordaja b märgile.
- Kui $r>0$, siis ühe tunnuse suurenedes keskmiselt teine tunnus kasvab ja vastupidi - ühe vähenedes väheneb ka teine.
- Kui $r<0$, siis ühe tunnuse väärtuste suurenedes keskmiselt teise tunnuse väärtused kahanevad ja vastupidi - ühe kahanedes teine kasvab.
- Kui tunnused on lineaarselt sõltumatud (tunnuste vahel võib aga olla mittelineaarne sõltuvus), siis on korrelatsioonikordaja null $r = 0$.
- Korrelatsioonikordaja ruut r^2 ehk determinatsioonikordaja näitab, kui suur osa ühe tunnuse hajuvusest (dispersioonist) on kirjeldatud teise poolt.
- Mida suurem on korrelatsioonikordaja absoluutväärtus, seda tugevam on korrelatiivne seos tunnuste vahel.
- Mõõtühiku (lineaarne) vahetus ei muuda korrelatsioonikordaja suurust (Korrelatsioonikordaja ei muutu, kui mõõdame temperatuuri Celsiuse kraadide Co asemel Farenheitides Fo, samuti võime pikkust mõõta sentimeetrites või meetrites- korrelatsioonikordaja jääb samaks).





Prognoosimine regressioonimudeli abil

Hinnatud regressioonimudel võimaldab prognoosida keskmist funktsioontunnuse väärtust etteantud argumenttunnuse väärtuse korral. Vahel on keskmise teadmisest vähe – soovime täpsemalt teada, millised tulemused on mõeldavad ja millised mitte.

Varem kirjeldasime uuritava tunnuse võimalikke väärtuseid prognoosiintervalli abil. Sarnast lähenemist saab kasutada ka nüüd – leida mingi etteantud argumenttunnuse väärtuse korral vahemik, kuhu järgmine uuritav tunnuse väärtus sattuks suure tõenäosusega, näiteks tõenäosusega 0,95. Vastavat vahemikku nimetatakse (ka) prognoosi-intervalliks.

Usaldusintervalli ja prognoosiintervall - näide

```
> mudel=lm(kaal~pikkus)
> predict(mudel, data.frame(pikkus=170),
interval="confidence")
         fit      lwr      upr
[1,] 59.9666 58.63289 61.30031
> predict(mudel, data.frame(pikkus=170),
interval="prediction")
         fit      lwr      upr
[1,] 59.9666 45.91387 74.01933
```

95%-usaldusintervall 170cm pikkuste tudengite kaalu keskväärtusele on (58,6...61,3); 95%-prognoosiintervall 170cm pikkuste tudengite kaalule on 45,9...74,0 kg.