Põhjuslikkus

Kuidas hinnata põhjuslikke mõjusid?

Mis on põhjuslik mõju?

Jaan suitsetas ja suri noorelt.

Kas suitsetamine põhjustas Jaani enneaegse surma?

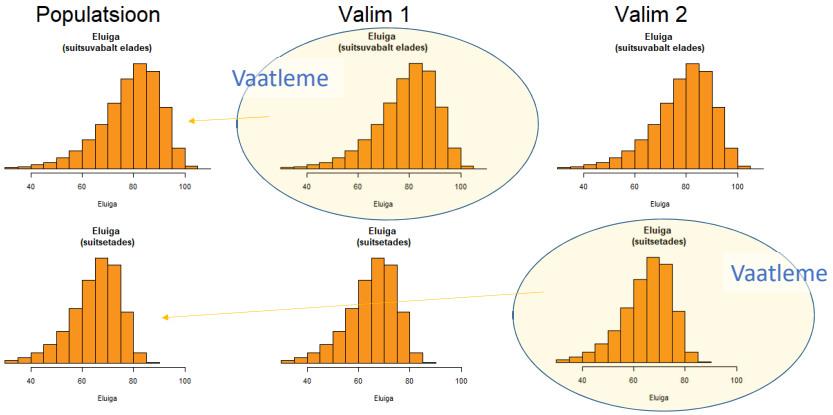
Sellele küsimusele vastamiseks vajame kontrafakte (*counterfactuals*):

- Jaan suitsetas ja suri noorena.
- Kui Jaan poleks suitsetanud, poleks ta noorena surnud.

Järeldus: suitsetamine põhjustas Jaani surma.

Jaani surma pole kunagi täie kindlusega võimalik suitsetamise süüks ajada – sest meie võimuses pole lasta tal kaks korda elada (ükskord suitsetades ja teine kord ilma suitsetamata)

Kuidas leida põhjuslikku mõju (teoreetiliselt)?



Randomiseeritud katsed

Kas geenimutatsioon kohas X mõjutab tunnust Y?

100 000 publitseeritud GWAS uuringut, igaüks uurib ~ 1 000 000 geenimutatsiooni mõju huvipakkuvale tunnusele (keskmine valimi suurus 100 000 + inimest): kokku 100 000 000 000 uurimisküsimust

Randomiseeritud katsed inimestel geenimutatsiooni tegeliku mõju kontrollimiseks mingile tunnusele:

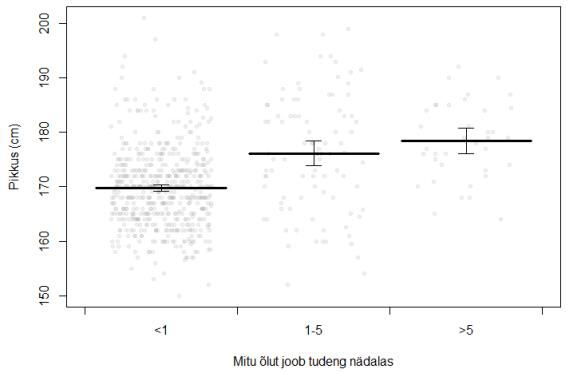
~ 84

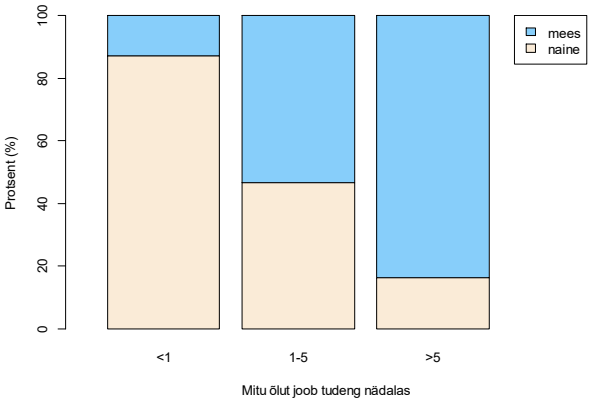
+ randomiseeritud katsed hiirtel/rottidel/merisigadel, +randomiseeritud katsed inimestel mis ei uuri otse konkreetse geenimutatsiooni mõju kuid lisavad tänu antud geenile produtseeritavat valku jne

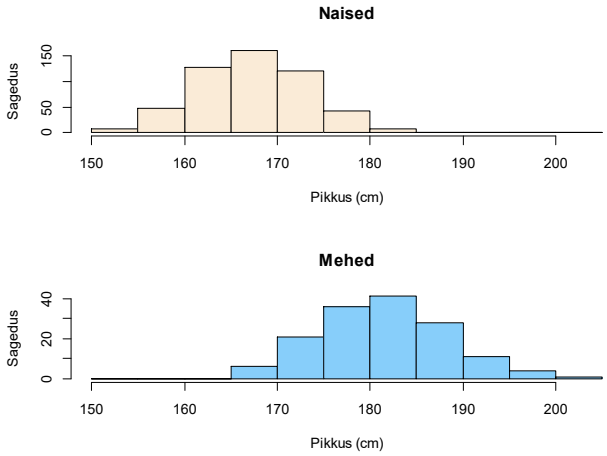
Mitterandomiseeritud uuringuid tehakse inimestel miljon korda rohkem kui randomiseeritud uuringuid

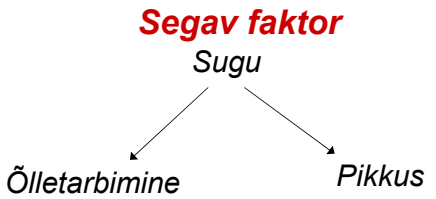
Kuidas tekib statistiline seos tunnuste *X* ja *Y* vahel?

- *X* mõjutab tunnuse *Y* väärtust
- *Y* mõjutab tunnuse *X* väärtust
- segav faktor/tunnus



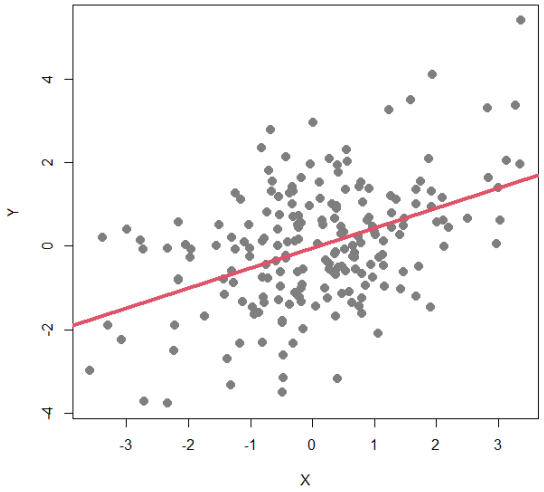






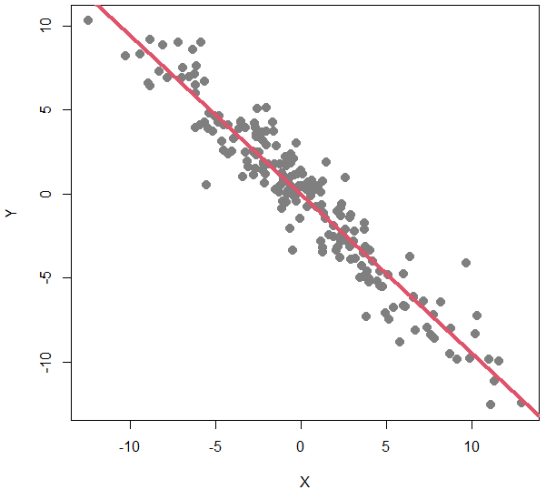
Segava faktori mõju

S <- rnorm(n, sd=**1**)
X <- S +rnorm(n)
Y <- S +rnorm(n)
plot(X,Y)



Segava faktori mõju

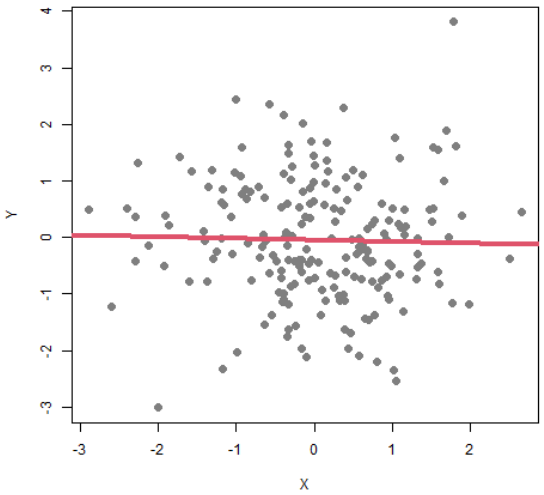
S <- rnorm(n, sd=**5**)
X <- S +rnorm(n)
Y <- (-S) +rnorm(n)
plot(X,Y)



Segava faktori mõju

S <- rnorm(n, sd=**0**)
X <- S +rnorm(n)
Y <- (-S) +rnorm(n)
plot(X,Y)

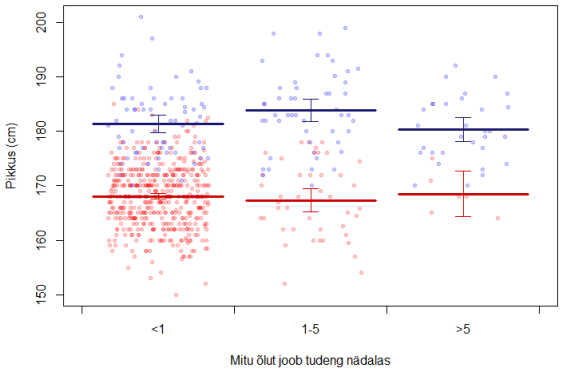
Alles siis, kui segava tunnus muutub konstantseks/mittevarieeruvaks (kõik katsetaimed kasvavad samal põllul ja saavad seega samapalju päikesevalgust) või kui tema mõju elimineeritakse (taimi kasvatatakse keldris kunstvalguse käes) muutub tegelik seos tunnuste X ja Y vahel nähtavaks (antud juhul: seose puudumine)



Lahendus segava tunnuse probleemile (1)

Uurime meid huvitavat seost fikseerides segava tunnuse väärtuse: vaatame, kas õlletarbimise ja tudengi pikkuse vahel eksisteerib seos kui võrdleme samast soost tudengeid (meestudengeid meestudengitega)

$X \perp\!\!\!\perp Y \mid S$?



Tingimustamine (kohandamine)

```
> m1=lm(pikkus~factor(olu3))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     169.7377     0.3437 493.902  < 2e-16 ***
factor(olu3)1-5   6.3906     0.8943   7.146 2.38e-12 ***
factor(olu3)>5    8.7083     1.3465   6.467 1.95e-10 ***
```

analüüs, kus sugu pole fikseeritud

sugu on fikseeritud

```
> m2=lm(pikkus~factor(olu3)+sugu)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     153.8306     0.7740 198.735   <2e-16 ***
factor(olu3)1-5   0.6859     0.7294   0.940    0.347
factor(olu3)>5   -1.3005     1.1232  -1.158    0.247
sugu             14.1013     0.6458  21.837   <2e-16 ***
```

Raskused

ei arvesta segava faktoriga

```
> summary(lm(Y~X))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.205899   0.044682   4.608 4.32e-06 ***
X           0.951924   0.006151 154.751  < 2e-16 ***
```

segav tunnus mõõdetud veaga

```
> summary(lm(Y~X+Sveaga))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.32553    0.17589  -7.536 7.31e-14 ***
X                 0.76979    0.01417  54.324  < 2e-16 ***
Sveaga(-4.2,2.86] 1.36335    0.18564   7.344 3.01e-13 ***
Sveaga(2.86,9.92] 2.68169    0.23691  11.320  < 2e-16 ***
Sveaga(9.92,17]   4.11335    0.31492  13.062  < 2e-16 ***
Sveaga(17,24.1]   5.33175    0.52312  10.192  < 2e-16 ***
```

Vale mudel

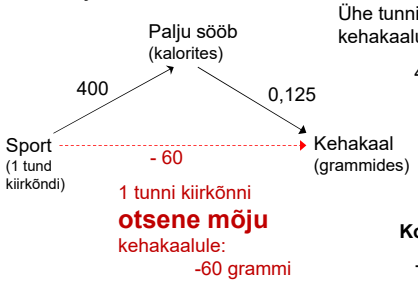
```
> summary(lm(Y~X+I(log(S))))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.10964    0.06812    1.61    0.108
X           0.80126    0.01446   55.41   <2e-16 ***
I(log(S))   0.71538    0.06350   11.27   <2e-16 ***
```

Õige mudel, täpselt mõõdetud segav tunnus

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002553   0.031244  -0.082    0.935
X            0.031791   0.021601   1.472    0.141
S            0.970725   0.022123  43.879   <2e-16 ***
```

Kas arvestame kõiki tunnuseid mida oleme mõõtnud?

Probleem 1:
Kogumõju vs kaudne mõju



Ühe tunni kiirkõnni **kaudne mõju** kehakaalule (läbi söögiisu tõusu):

400*0,125 = +50 grammi

Kogumõju:

-60 grammi + 50 grammi
= -10 grammi

Fikseerides (lisades mudelile) söödud kalorid hindame spordi otsest mõju;
Ilma fikseerimata näeme kogumõju.

Kas arvestame kõiki tunnuseid mida oleme mõõtnud?

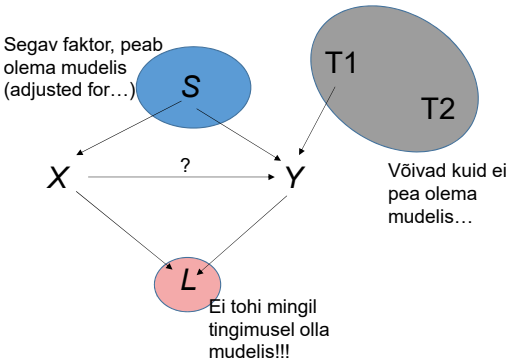
Probleem 2:
Ülesobitamine ehk
Overadjustment

```
> y <- 2*x + rnorm(100)

> lm(y~x)

Coefficients:
(Intercept)            x
    0.05976      2.05636

> lm(y~x+z)

Coefficients:
(Intercept)            x            z
     0.0443      -0.1766       0.3629
```

Ole ettevaatlik tunnustega mis ei ole segavad faktorid!!!

Antud juhul *z* sõltub *y*-tunnusest: sellise tunnuse lisamine mudelisse (tema järgi tingimustamine) rikub põhjusliku analüüsi!

Kuidas leida põhjuslikku mõju?

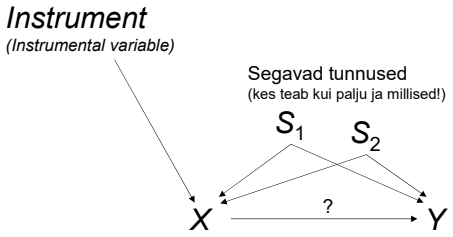


Vahel raske otsustada, mida tuleks mudelisse lisada ja mida mitte.

Nõuab modelleeritava süsteemi väga head mõistmist.

Märksõna:
Struktuurivõrrandid
(Structural equations)

Kuidas leida põhjuslikku mõju?

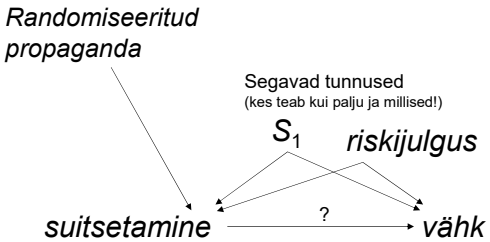


Randomiseerida ei saa; kõiki segavaid tunnuseid mõõta ei oska või ei lubata?

Mis saab siis?

Siis vajad instrumenttunnust!

Kuidas leida põhjuslikku mõju?

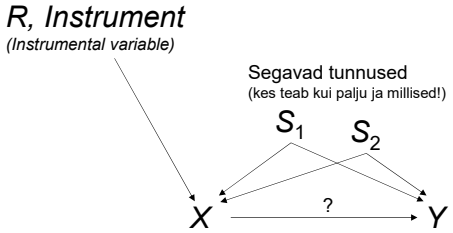


Instrumenttunnus:

- Mõjutab (põhjuslikult) *X*-tunnuse väärtust
- Instrumenttunnuse väärtust ei mõjuta segavad tunnused
- Instrumenttunnusel puudub muu mõju *Y*-le (mõjutab *Y*-tunnuse väärtust vaid seeläbi, et omab mõju *X*-tunnusele)

Instrumenttunnuse olemasolu korral on võimalik hinnata X-i põhjuslikku mõju Y-le ka siis, kui me ei tea mis võiks olla segavaks faktoriks (või segavate faktorite mõõtmine pole võimalik)

Kuidas leida põhjuslikku mõju?



Lineaarne juhtum:

$$Y = \beta_1 + \beta_{SY} S + \beta_{XY} X + \varepsilon_Y$$
$$X = \beta_2 + \beta_{SX} S + \beta_{RX} R + \varepsilon_X$$

$$\begin{aligned}\operatorname{cov}(R;Y) &= \operatorname{cov}(R;\beta_{XY}\beta_{RX}R) \\ &= \beta_{XY}\beta_{RX}\mathrm{D}R\end{aligned}$$

$$\begin{aligned}\operatorname{cov}(R;X) &= \operatorname{cov}(R;\beta_{RX}R) \\ &= \beta_{RX}\mathrm{D}R\end{aligned}$$

$$\beta_{XY} = \operatorname{cov}(R;Y)/\operatorname{cov}(R;X)$$

$$\hat{\beta}_{XY} = \widehat{\operatorname{cov}}(R;Y)/\widehat{\operatorname{cov}}(R;X)$$

Põhjuslik mudel ei ole hea mudel prognoosimiseks!

Tegelik andmeid genereeriv mudel:

$$Y = 1 \cdot X + \varepsilon$$

Ise tegin:
y = x + epsilon

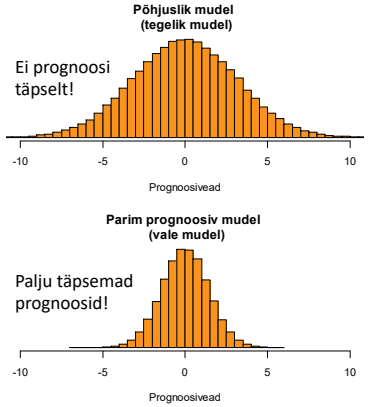
Parim prognoosiv mudel:

$$Y = 2 \cdot X + \varepsilon$$

Kumba mudelit kasutada uute vaatluste *y*-tunnuse väärtuste prognoosimiseks?

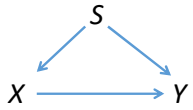
Ise hindasin:

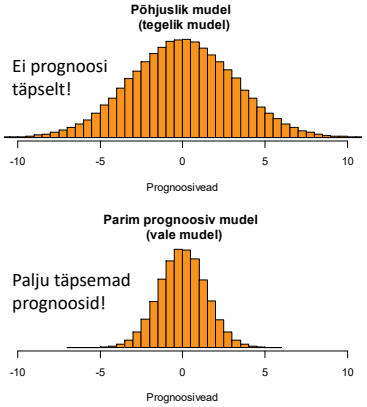
```
> mudel=lm(y~x)
> summary(mudel)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.007855   0.017239   0.456    0.649
x           2.002038   0.012200 164.105   <2e-16 ***
Residual standard error: 1.724 on 9998 degrees of freedom
Multiple R-squared:  0.7293,    Adjusted R-squared:  0.7292
```



Põhjuslik mudel ei ole hea mudel prognoosimiseks!

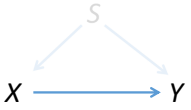
Miks?



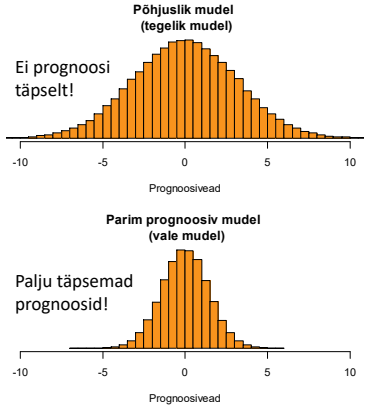


Põhjuslik mudel ei ole hea mudel prognoosimiseks!

Miks?

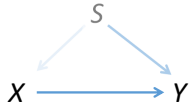


Põhjuslik mudel vaatab vaid *X*-i mõju *Y*-le



Põhjuslik mudel ei ole hea mudel prognoosimiseks!

Miks?



Prognoosiv mudel üritab välja mõistatada nii *X*-i kui ka *S*-i mõju *Y*-le (kuigi segava faktori *S* väärtused pole teada...)