Cocktailnålar i kemikaliehöstackar

- en statistikers irrfärder i toxikologins värld...

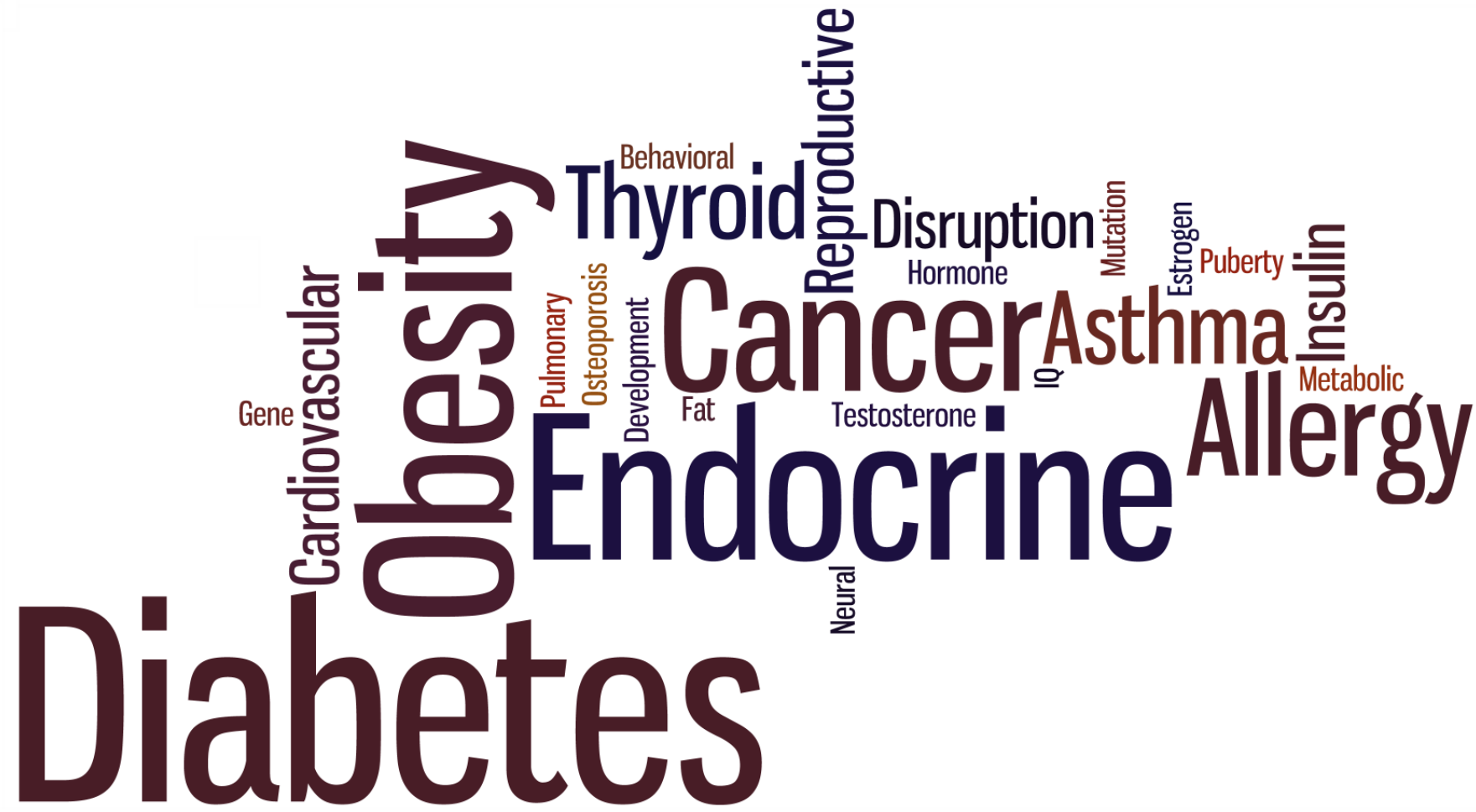
Erik Lampa

Arbets- och miljömedicin
Institutionen för medicinska vetenskaper
Uppsala Universitet

UPPSALA
UNIVERSITET

Påverkas människor av kemikalier?

En titt i PubMed...



Det traditionella synsättet

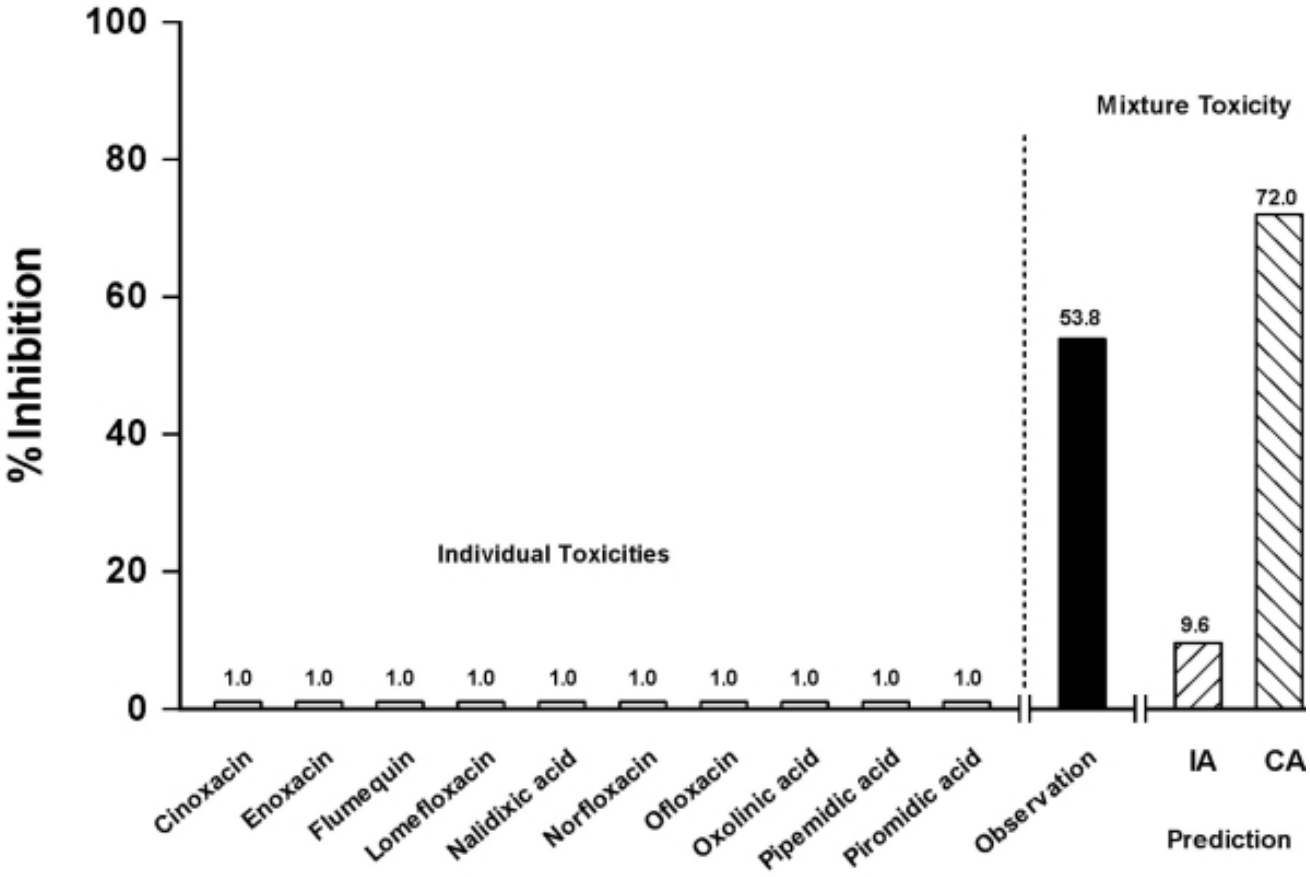
- Vi är exponerade för många ämnen samtidigt
- Traditionell riskbedömning – fokus på ett ämne i taget
- Nästan aldrig tillämpbart i verkligheten

Kemiska cocktails (mixtures)

Shaken, not stirred. . .

- Produkter som innehåller $>$ 1 kemikalie
- Kemikalier som släpps ut gemensamt, ex. avgaser
- Kemikalier som finns samtidigt i miljön

Varför bry sig?



Backhaus T *et al.* The single substance and mixture toxicity of quinolones to the bioluminescent bacterium *Vibrio fischeri*, Aquatic Toxicology, 49(1-2), 49-61, 2000

Två sidor av samma mynt

- **I**ndependent **A**ction
  - Kemikalierna verkar oberoende av varandra
  - Oftast orimligt antagande

- **C**oncentration **A**ddition
  - Kemikalierna verkar beroende av varandra
  - Interaktionseffekter

Några kunskapsluckor

- CA kräver kända dos-respons-samband
- Kunskap om mekanismer hos människor
- Verktyg för att identifiera / predicera mixtureeffekter

Mission impossible?

- Identifiera relevanta kemikalier
- Hitta interaktioner bland många kemikalier
- Identifiera icke-linjära effekter
- . . . utan några egentliga hypoteser (!!)

Några metoder...

...som inte fungerar så bra

- Vanlig regression
  - $N << p$
- Stegvis regression
  - multipla jämförelser, samt $N << p$
- Pre-conditioning med LASSO
  - Skapa modellmatrisen, icke-linjära termer, standardisering, tolkning

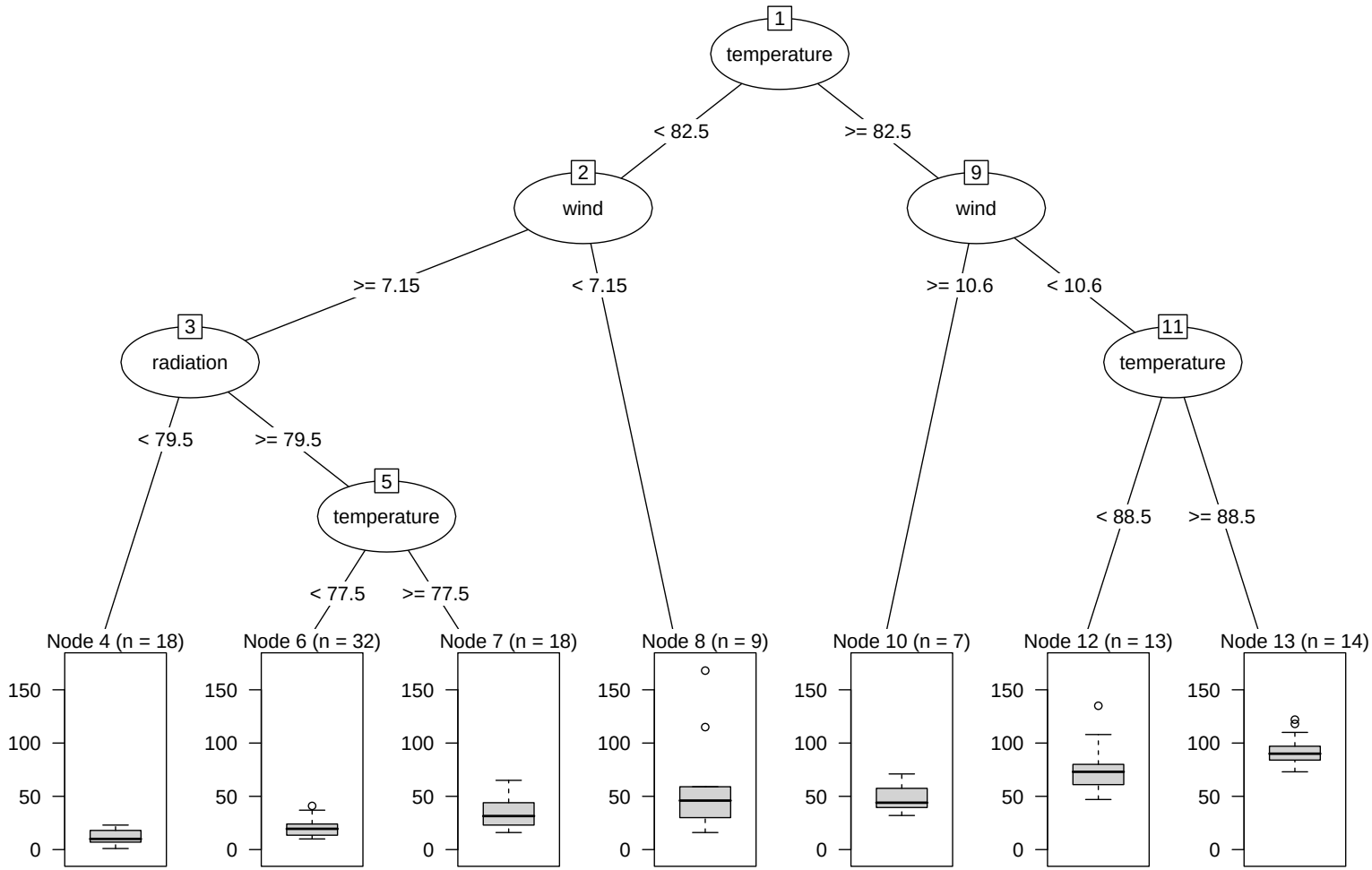
Regressionsträd

- Hanterar stökiga interaktioner
- Ej känsliga för monotona transformationer av förklarande variabler
- Ej känsliga för outliers hos förklarande variabler
- Hanterar bortfall hos förklarande variabler
- Hanterar mixade variabeltyper
- Enkla att tolka

- Dålig prediktiv förmåga
- Dålig hantering av kontinuerliga variabler
- För mycket fokus på interakioner?

Exempel

Prediktion av ozonnivåer



Stokastisk gradientboosting

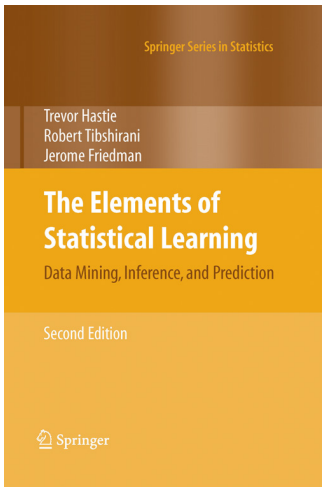
Hastie T., Tibshirani R., Friedman J. (2008) *The Elements of Statistical Learning*, kapitel 10

Anpassar en additiv modell

$$F(x) = \sum_{m=0}^{M} \beta_m b(x; \gamma_m)$$

med mål att minimera en förlustfunktion $L(y, F(x))$

$b(x; \gamma_m)$ är oftast regressionsträd



Generell boostingalgoritm

1. Välj $L[y, F(x)]$
2. Sätt $F_0(x)$ till en konstant
3. För m=1 till M
   1. Dra ett stickprov med storleken η
   2. Beräkna $r = -\frac{\partial L[y,F(x)]}{\partial F(x)}|_{F_m(x)=F_{m-1}(x)}$ och anpassa ett regressionsträd $g(x)$ till r
   3. Uppdatera $F_m(x) = F_{m-1}(x) + \epsilon\beta_m g(x)$
   4. Repetera många gånger

$0 < \epsilon \leq 1$ är en regulariseringsparameter som begränsar varje träds inflytande på $F(x)$ och reducerar överanpassning. β är steglängden längs gradienten. M kan väljas mha korsvalidering.

Minsta kvadratboosting

1. Börja med $F_0(x) = \bar{y}$ och residual $r = y - \bar{y}$, $m = 0$
2. $m \leftarrow m + 1$
3. Anpassa ett regressionsträd $g(x)$ till r
4. Uppdatera $F_m(x) = F_{m-1}(x) + \epsilon\beta_m g(x)$

   $r \leftarrow r - \epsilon\beta_m g(x)$

   och repetera steg 2 – 4 många gånger

Variabelbetydelse och partiella beroenden

- Boostade modeller är svåra att tolka
- Betydelsen av variabler är relaterade till antalet split – fler split, större betydelse
- Partiella beroendefunktioner kan ge en visuell bild av effekter och används för att utvärdera interaktioner
- P-värden, konfiensintervall?

Interaktioner

Friedman J.H., Popescu B. E. (2008) *Predictive learning via rule ensembles*

The Annals of Applied Statistics, Vol. 2, No. 3, 916–954

- Om x_j och x_k inte interagerar så är det partiella beroendet

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k)$$

- Definiera H som ett mått på interaktion, $0 \leq H \leq 1$

$$H = f\left(\frac{F_{jk} - F_j - F_k}{F_{jk}}\right)$$

- Generaliserar till interaktioner av högre ordning

Referensfördelning för H

- Skapa referensfördelning för H (H_0) mha boostrapvariant
- Beräkna upprepade H_0 från artificiella data $\{\tilde{y}, x\}_1^N$ genererade från riktiga data genom

$$\tilde{y} = F_A(x) + [y_p - F_A(x_p)]$$

eller

$$\Pr(\tilde{y} = 1) = [1 + \exp(-F_A(x))]^{-1}$$

- p är en permutering av $1, \ldots, N$. $F_A(x)$ är en funktion bestående av träd innehållande en variabel ("stumps").
- RuleFit - Träd + Linjära termer + LASSO
  `http://www-stat.stanford.edu/~jhf/R-RuleFit.html`

Mjukvara

- R – **gbm**, mboost, GAMboost, bst, CoxBoost, GMMBoost, . . .
- SAS – SAS Enterprise Miner
- STATA – boost
- Salford Systems – TreeNet

En enkel simulering

- Skapa $y = F(x) + \epsilon$ där

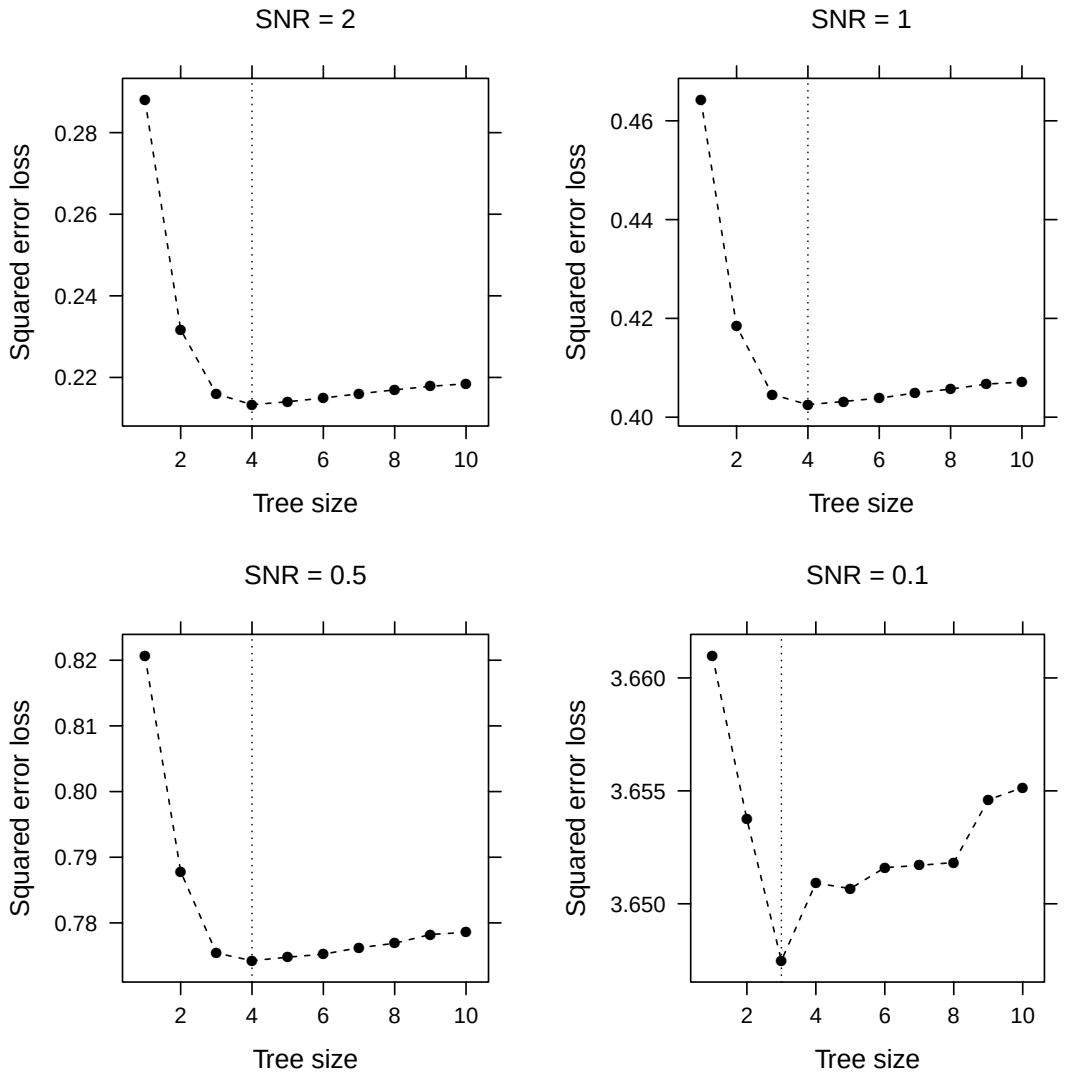
$$F(x) = 11 \prod_{i=1}^{4} \exp(-3(1 - s(x_i))^2) - 1.3\sin^2(\pi \cdot s(x_5))$$

och $\epsilon \sim N(0, \sigma^2)$ och σ väljs så att signal to noise ratio är 2, 1, 0.5 och 0.1

- x är fem olika kemikalier (pcb170, dde, mmp, cd och ocdd) simulerat från verkliga data innehållande 37 kemikalier, N = 1000.

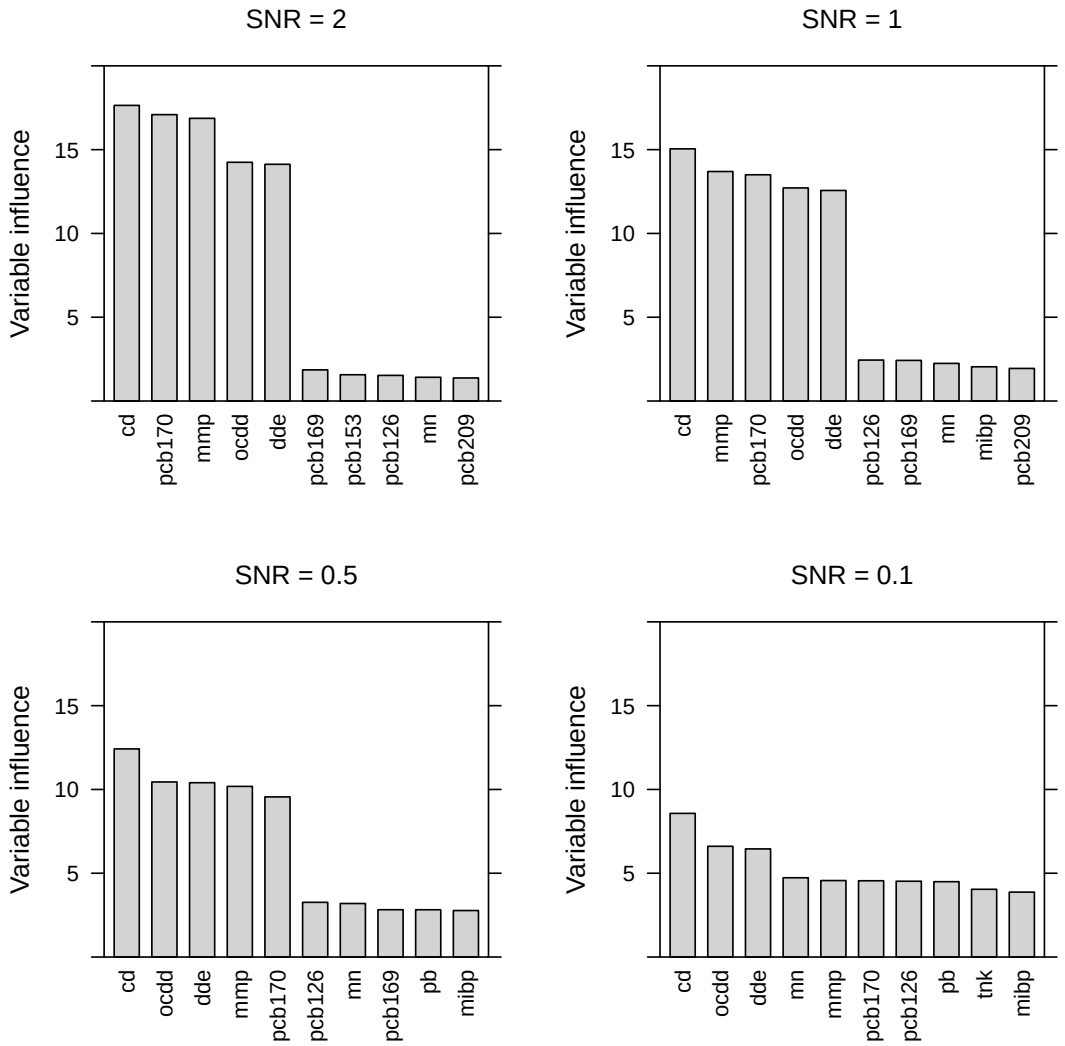
En enkel simulering

Medelvärden av 100 repitioner av 10-faldig korsvalidering i varje punkt



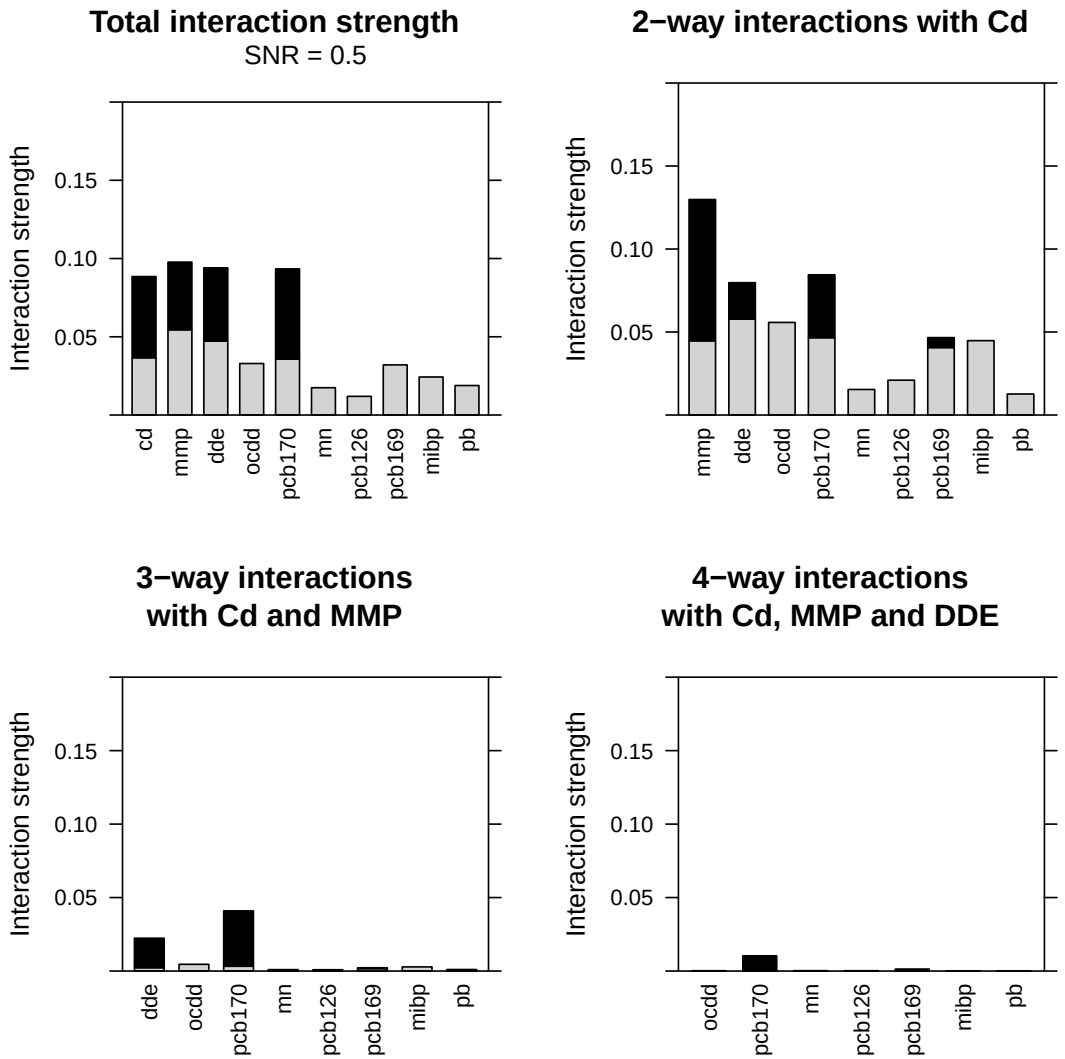
En enkel simulering

Variabelbetydelse



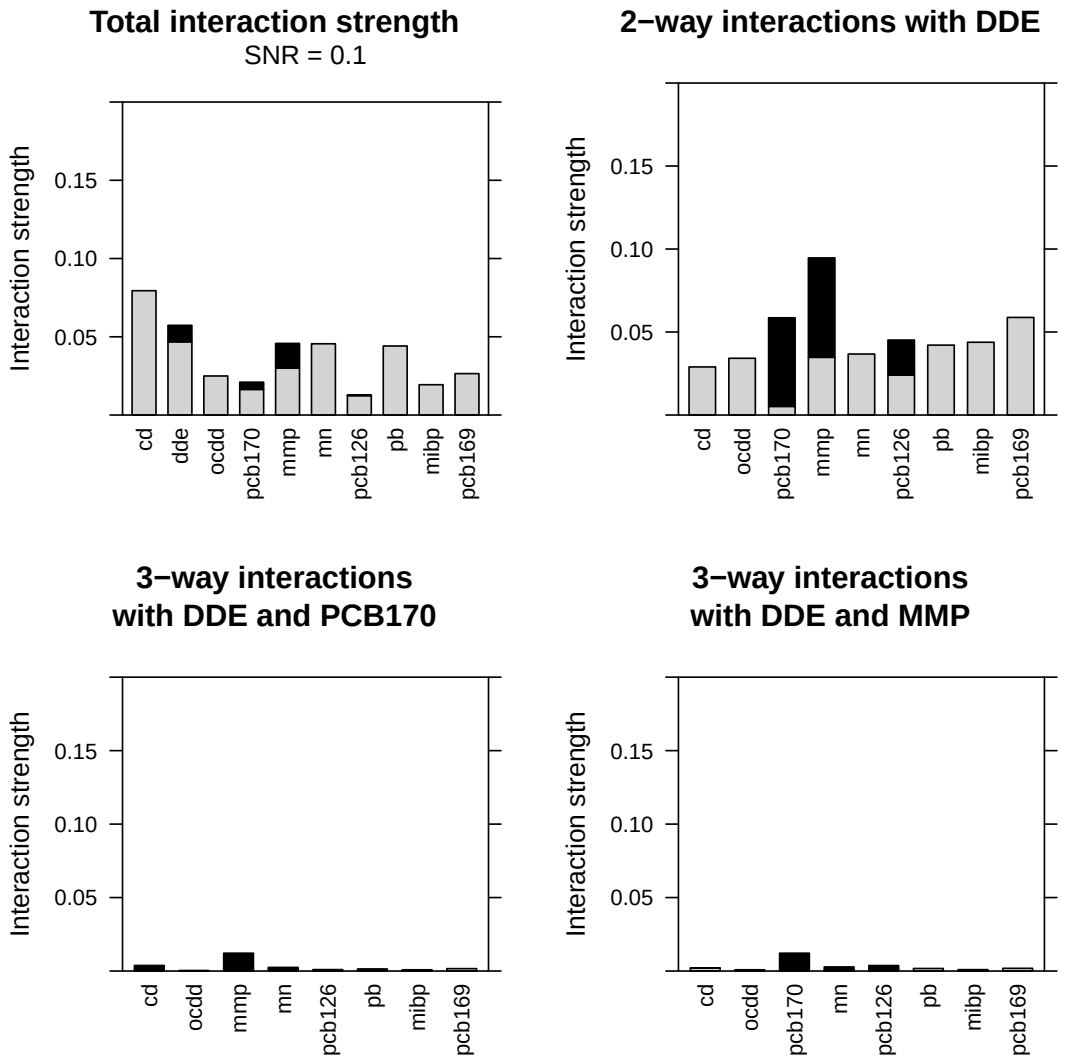
En enkel simulering

Interaktioner då SNR = 0.5



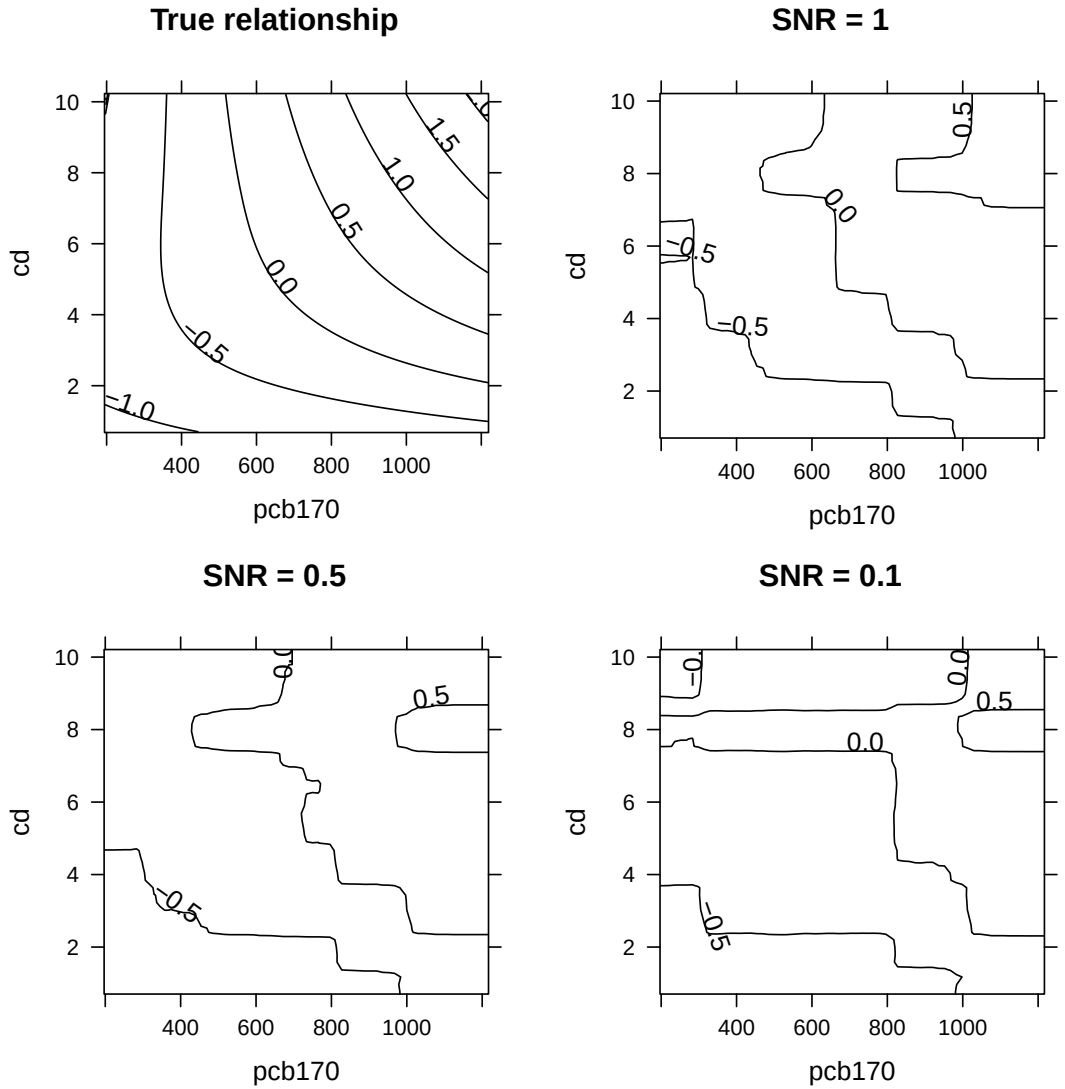
En enkel simulering

Interaktioner då SNR = 0.1



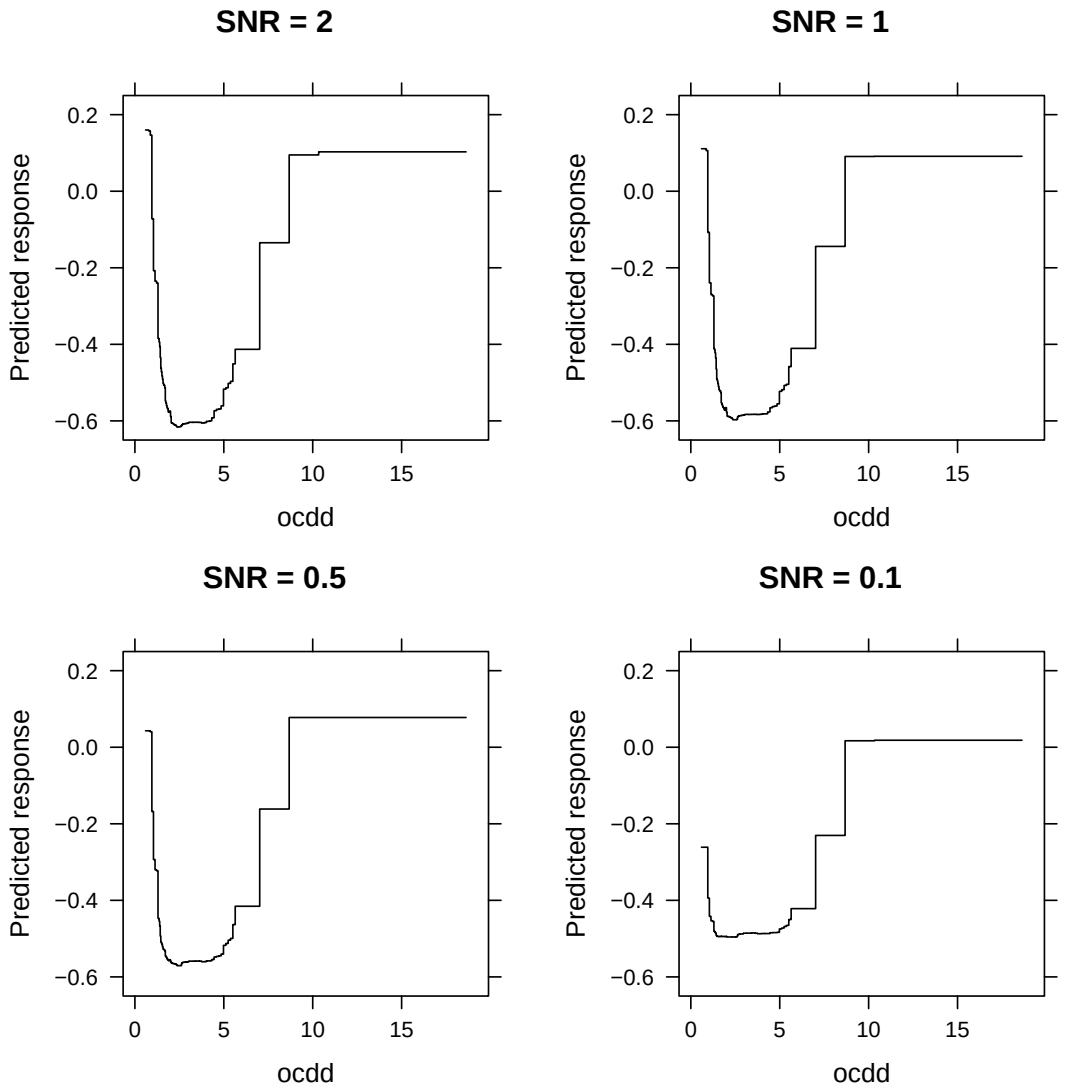
En enkel simulering

Interaktionen mellan Cd och PCB170



En enkel simulering

Icke-linjärt samband

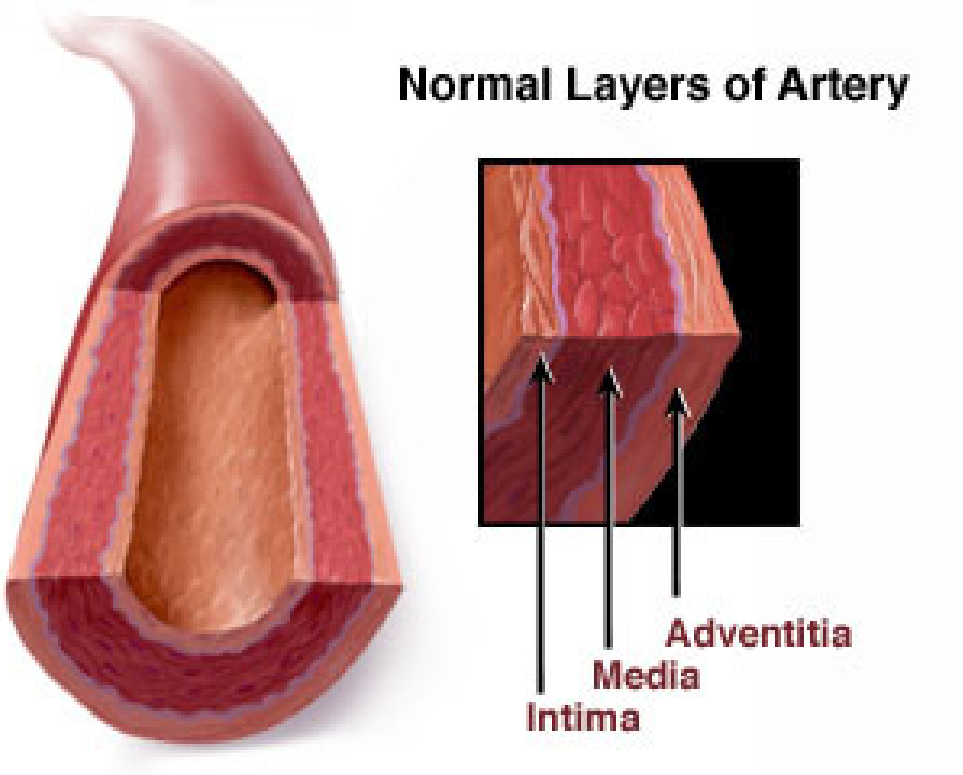


PIVUS

Prospective Investigation of the Vasculature in Uppsala Seniors

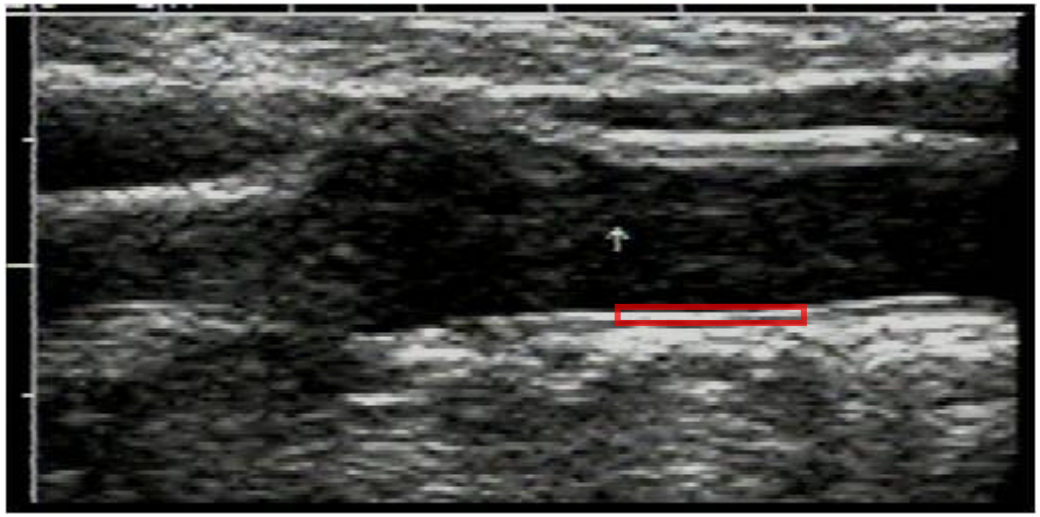
- 1016 st 70-åringar i Uppsala län
- Läkarundersökning, blodprov . . .
- 37 kemikalier uppmätta i blod
- Uppföljning 75 år och 80 år
- Finns det något samband mellan kemikalier och åderförkalkning?

Typisk artär



Källa: http://www.unc.edu/ mmlee/webproject2.html

Ultraljudsmått i en halspulsåder

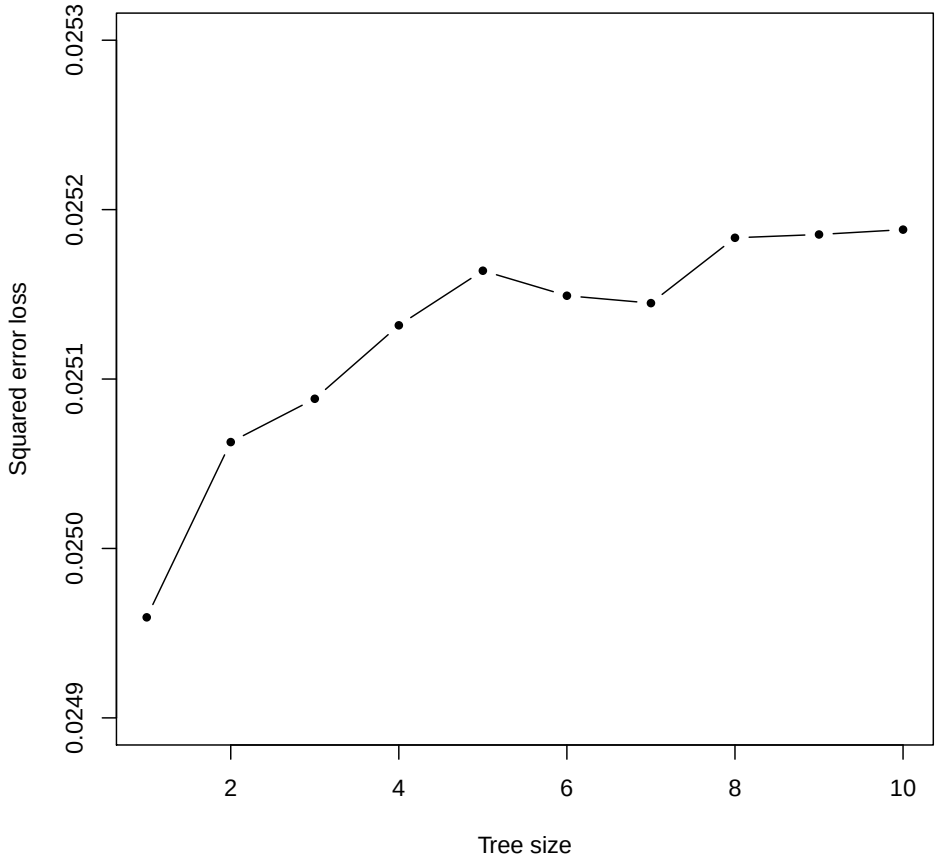


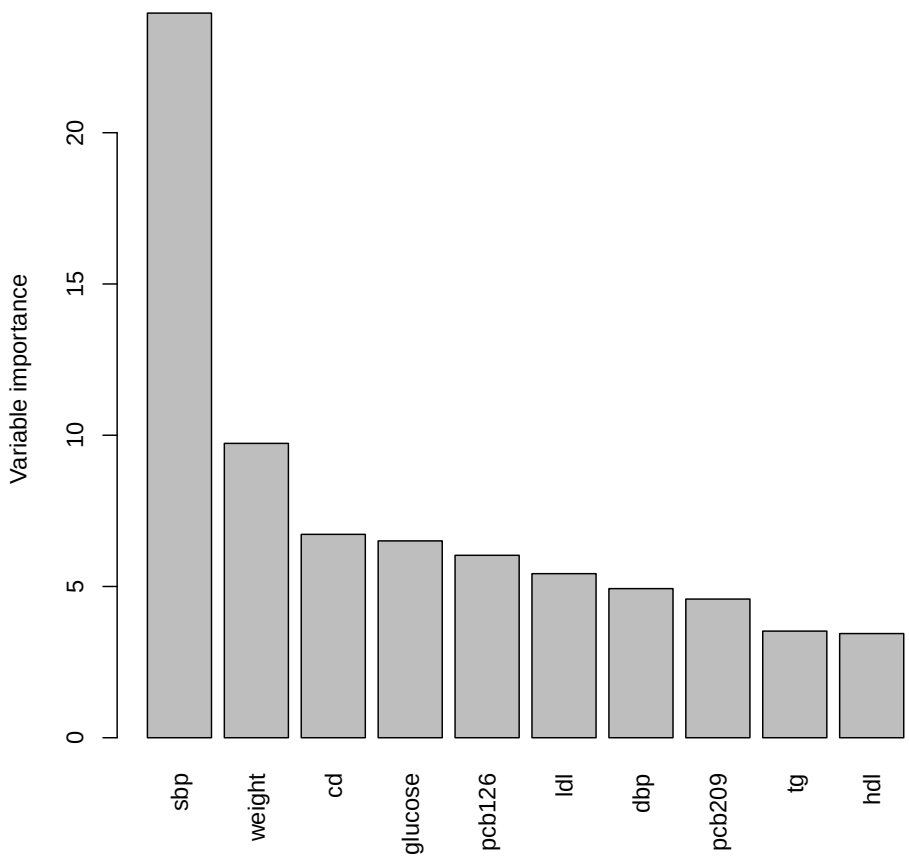
- IMT – Kärlväggens tjocklek
- IM-GSM – Kärlväggens ekogenicitet. Gråskala, relaterad till kompositionen i kärlväggen.
- "Klassiska" riskfaktorer: rökning, ↑kolesterol, ↑blodtryck, ↑blodsocker, ↑vikt, ↑triglycerider, (kön)

Strategi

- $L = \frac{1}{2}[y - F(x)]^2$
- Bestäm optimal storlek på träden samt M mha 10-faldig korsvalidering upprepad 100 gånger
- Om optimal trädstorlek > 1, Bestäm H för de 10 mest betydelsefulla variablerna och utvärdera interaktioner

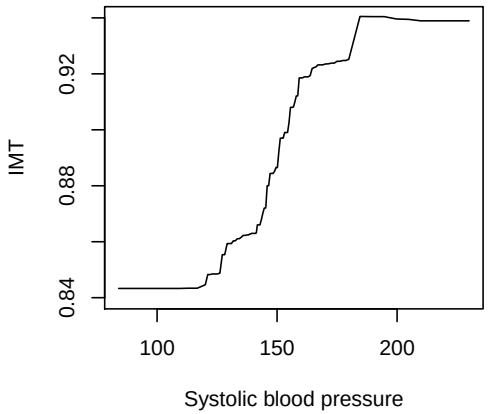
IMT

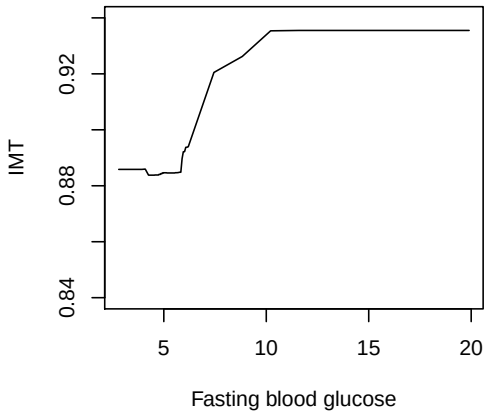

Squared error loss
0.0249
0.0250
0.0251
0.0252
0.0253
2
4
6
8
10
Tree size

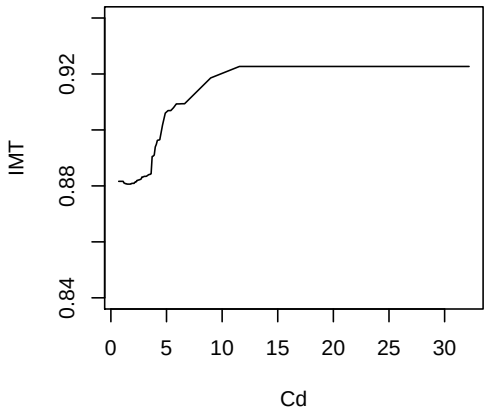

Variable importance
0
5
10
15
20
sbp
weight
cd
glucose
pcb126
ldl
dbp
pcb209
tg
hdl

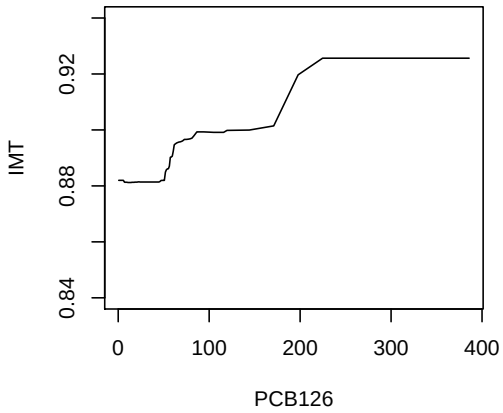
IMT

Partiella beroenden

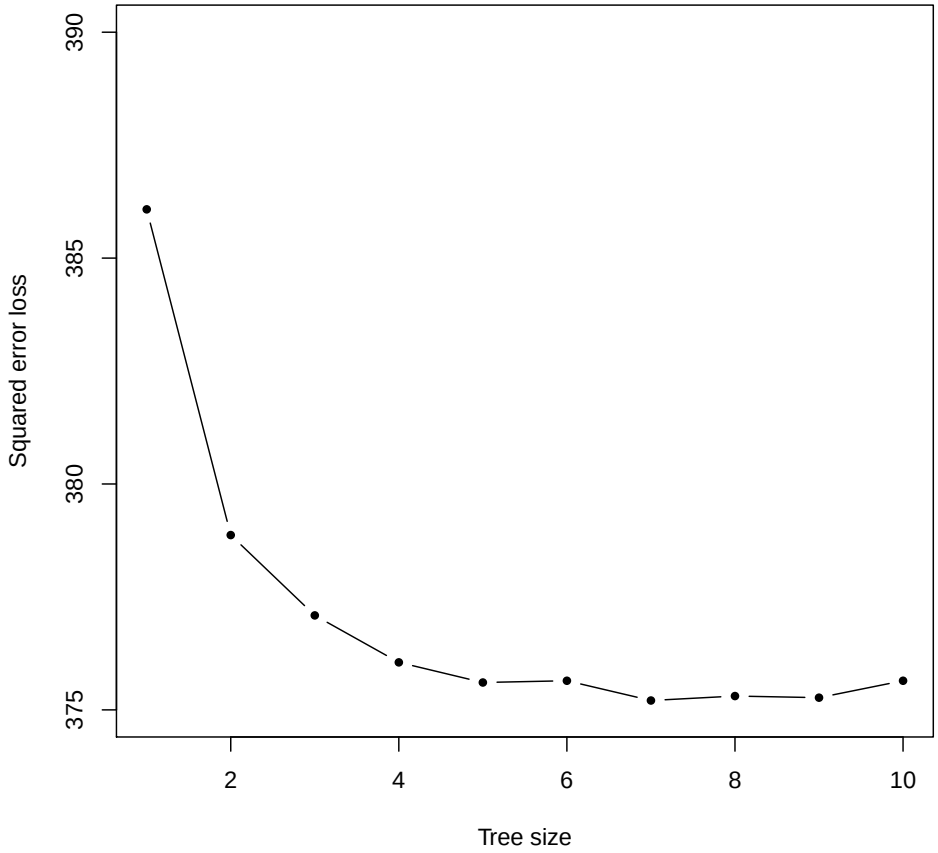


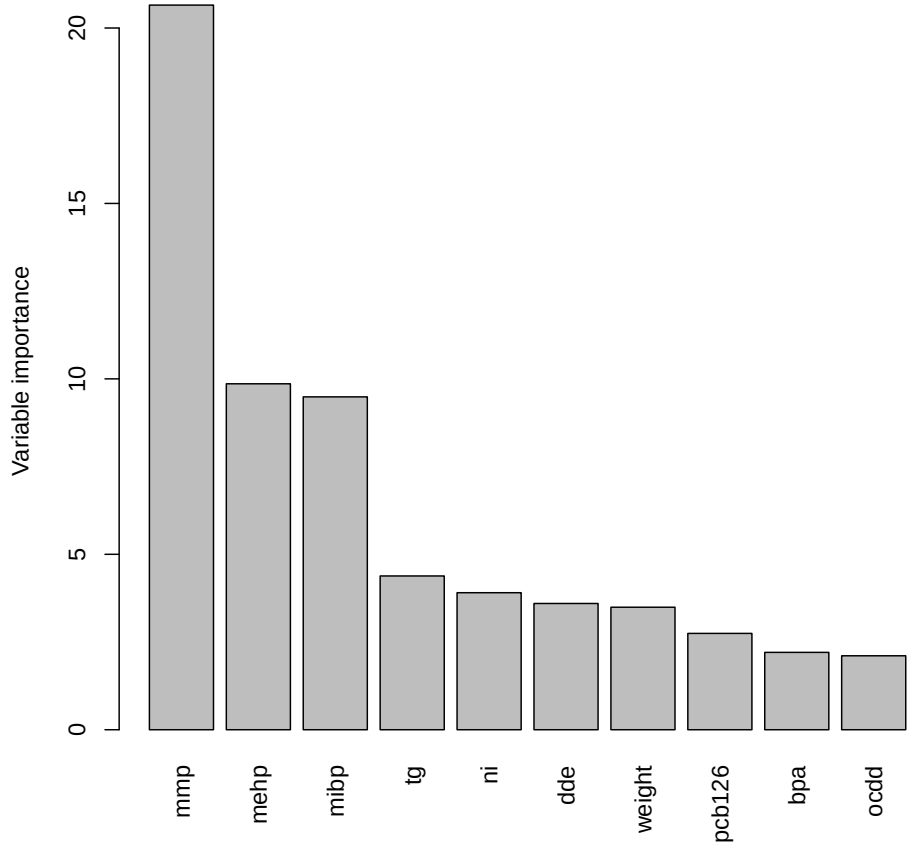






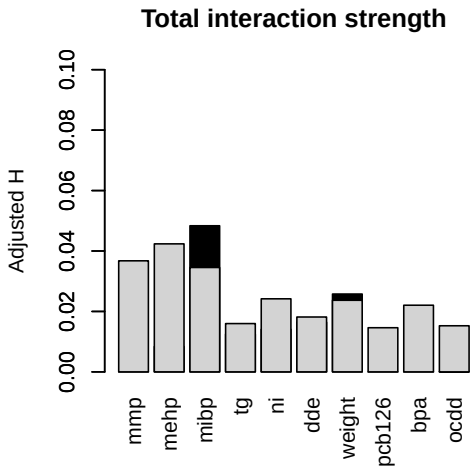
IM-GSM

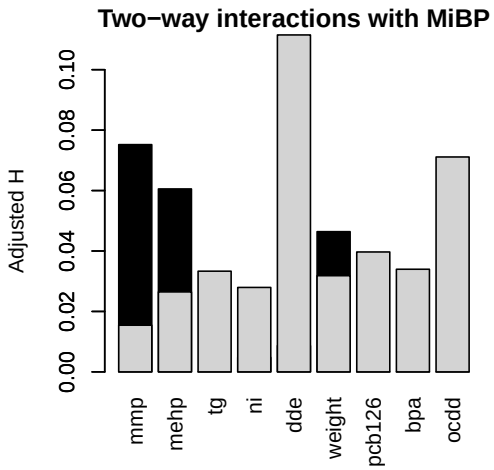

Squared error loss
375
380
385
390
2
4
6
8
10
Tree size

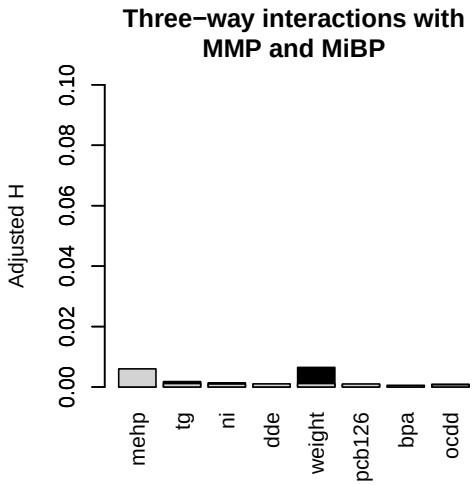

Variable importance
0
5
10
15
20
mmp
mehp
mibp
tg
ni
dde
weight
pcb126
bpa
ocdd

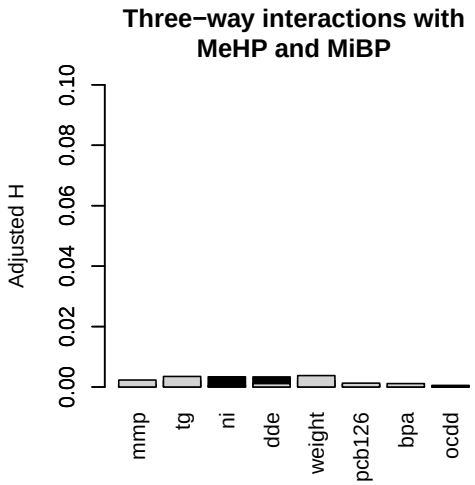
IM-GSM

Interaktioner



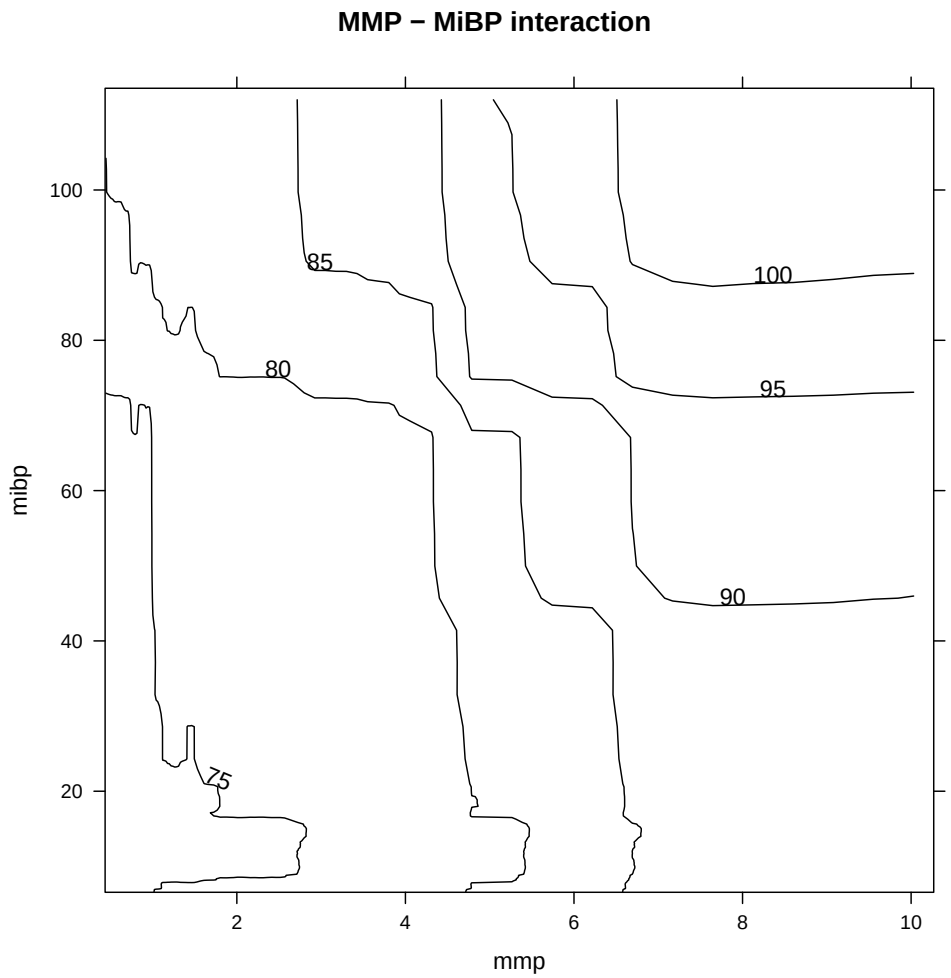


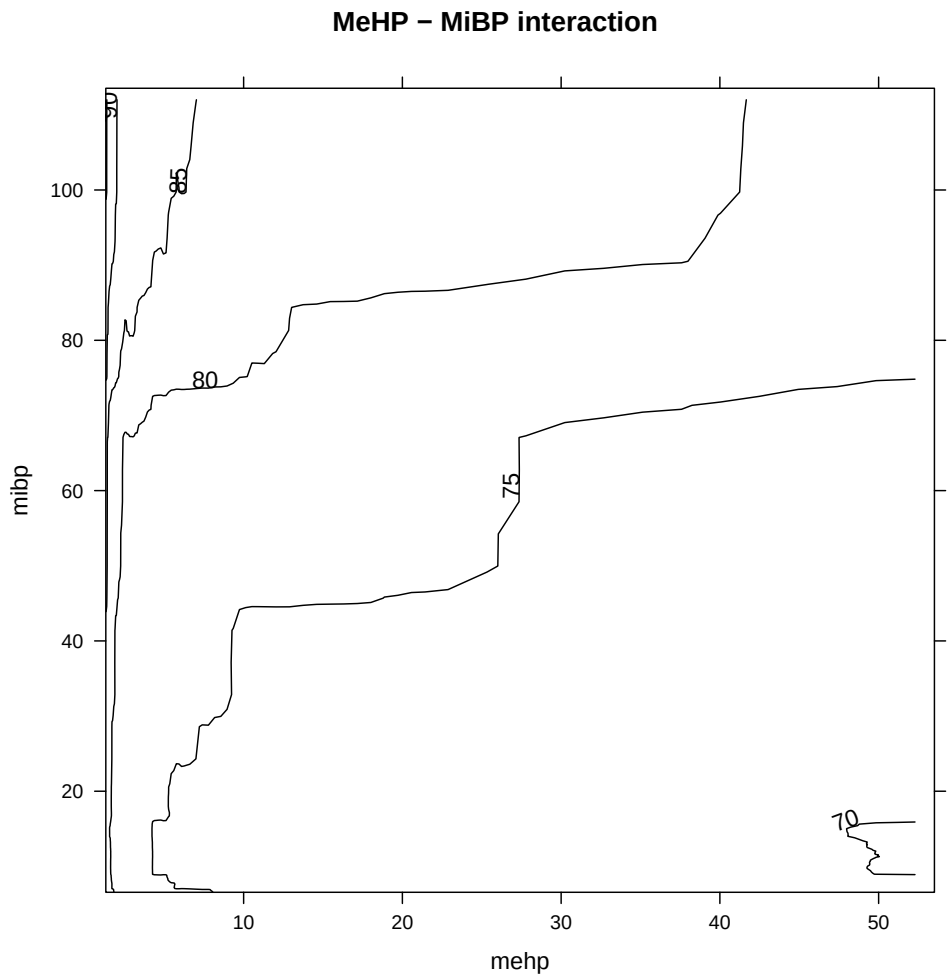




IM-GSM

Partiella beroenden





Summering

Hittar vi nålarna?

- Mixtureeffekter = interaktioner mellan kemikalier
- Boostade regressionsträd kan hitta komplexa interaktioner
- Andra (enklare) metoder?
- Biologisk relevans?

Tack till. . .

- Monica Lind, Uppsala Universitet
- Lars Lind, Uppsala Universitet
- Anna Bornefalk Hermansson, UCR