Multivariabel statistik
beware of the wolf

Johan Lindbäck

Uppsala Clinical Research Center

Kvalitetsregisterforskningskonferens
Arlanda 26 maj 2015

UCR
Uppsala Clinical Research Center

Multivariabel statistik

Förtydligande

- Multi*variat* = flera utfallsvariabler
- Multi*variabel* = flera förklaringsvariabler/prediktorer
- Behöver inte innebära modellering men vi kommer idag uteslutande prata om (regressions)modeller

Innehåll

Frågeställning

Tre områden där vi typiskt använder multivariabla modeller är

- Hypotesprövning

  Finns det ett samband mellan graden av fysisk aktivitet och risken för hjärtsjukdom?

- Estimering

  Hur mycket minskar blodtrycket om man får en viss typ av blodtrycksmedicin?

- Prediktion

  Vad är sannolikheten att en 70-årig kvinna med förmaksflimmer drabbas av en stroke inom 3 år?

Statistisk modell

Vanliga regressionsmodeller inom medicinsk forskning

Linjär regressionsmodell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2\right)$$

Logistisk regressionsmodell

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots, \beta_k X_k}} \iff$$

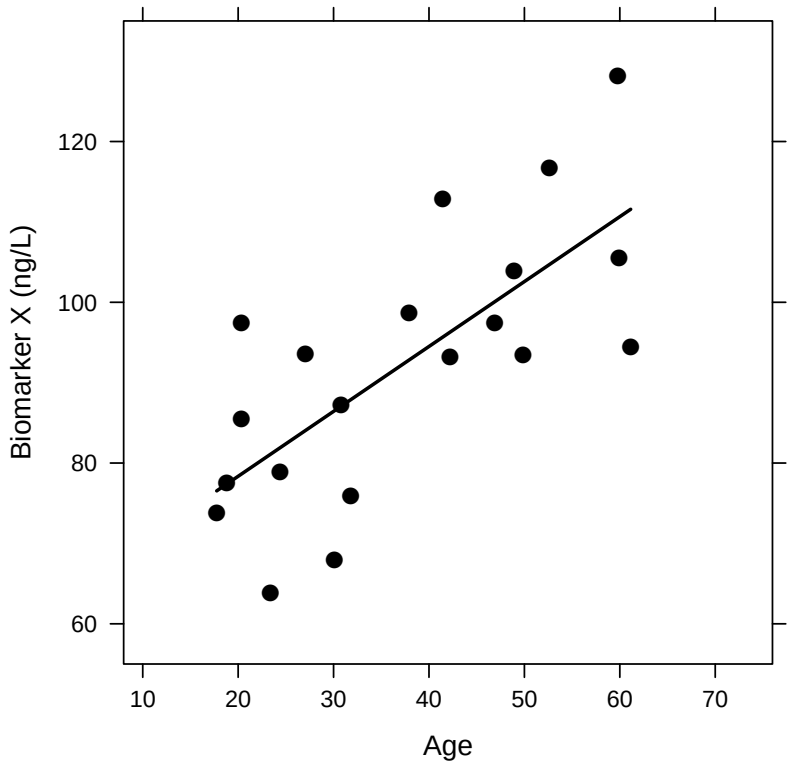
$$\ln\left[\frac{P(Y = 1)}{P(Y = 0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots, \beta_k X_k$$

Cox-regressionsmodell

$$h(t) = h_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$$

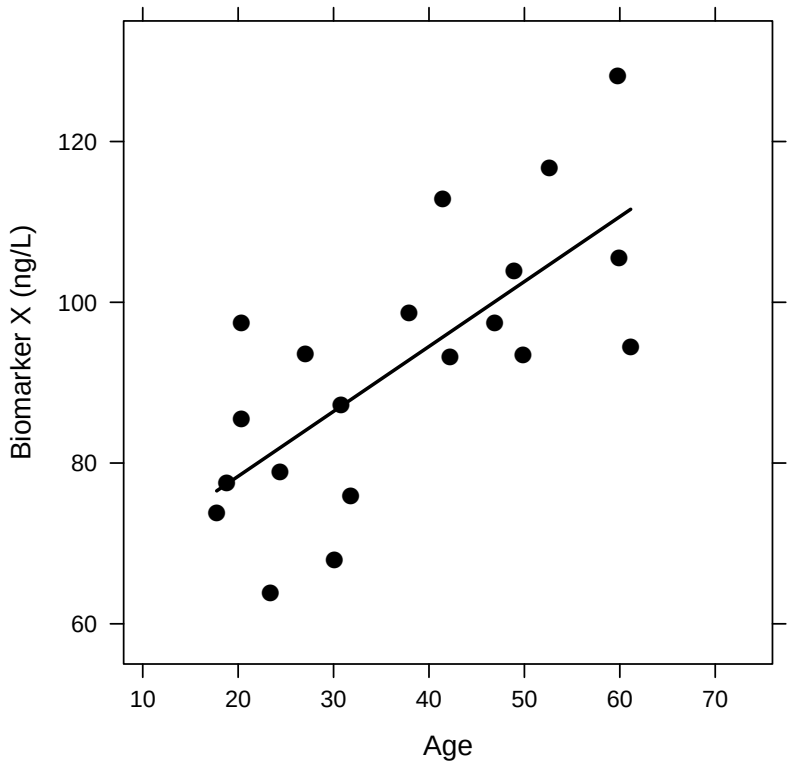
Confounding

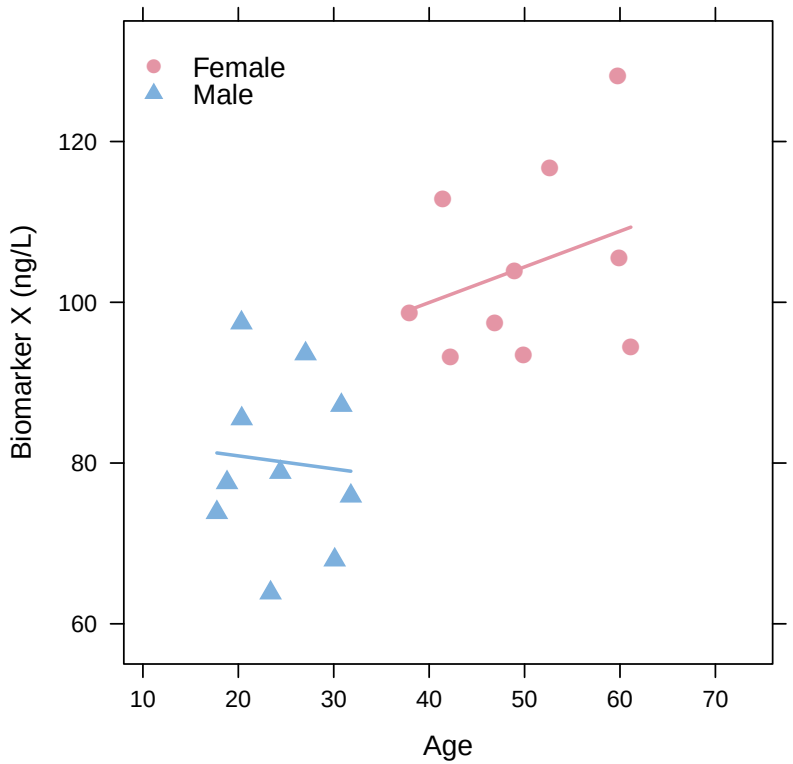
Exempel



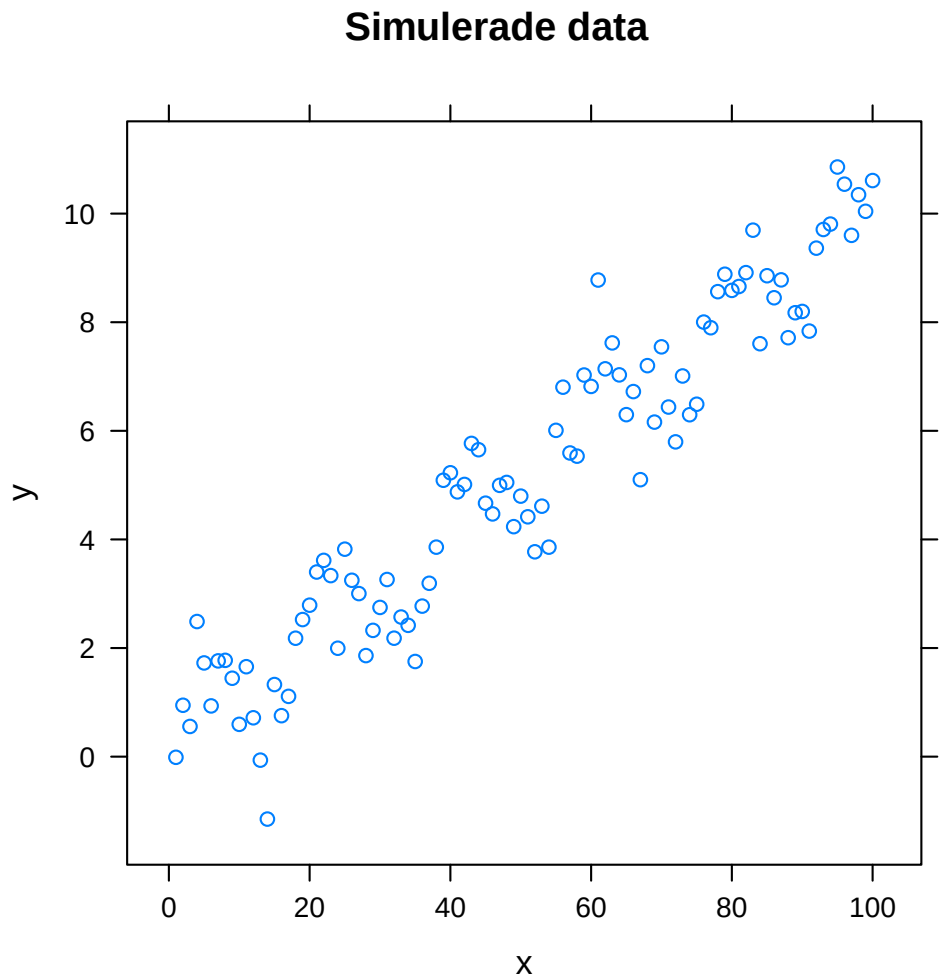
Confounding

Exempel





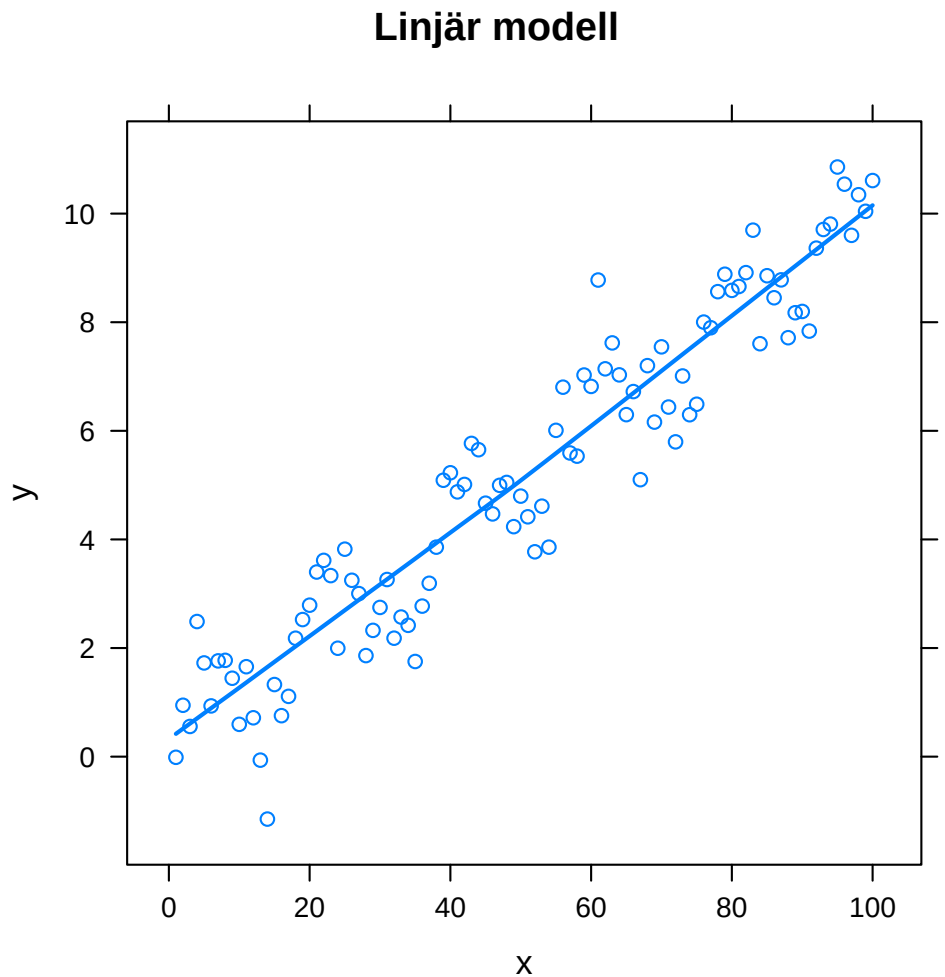
Modellanpassning



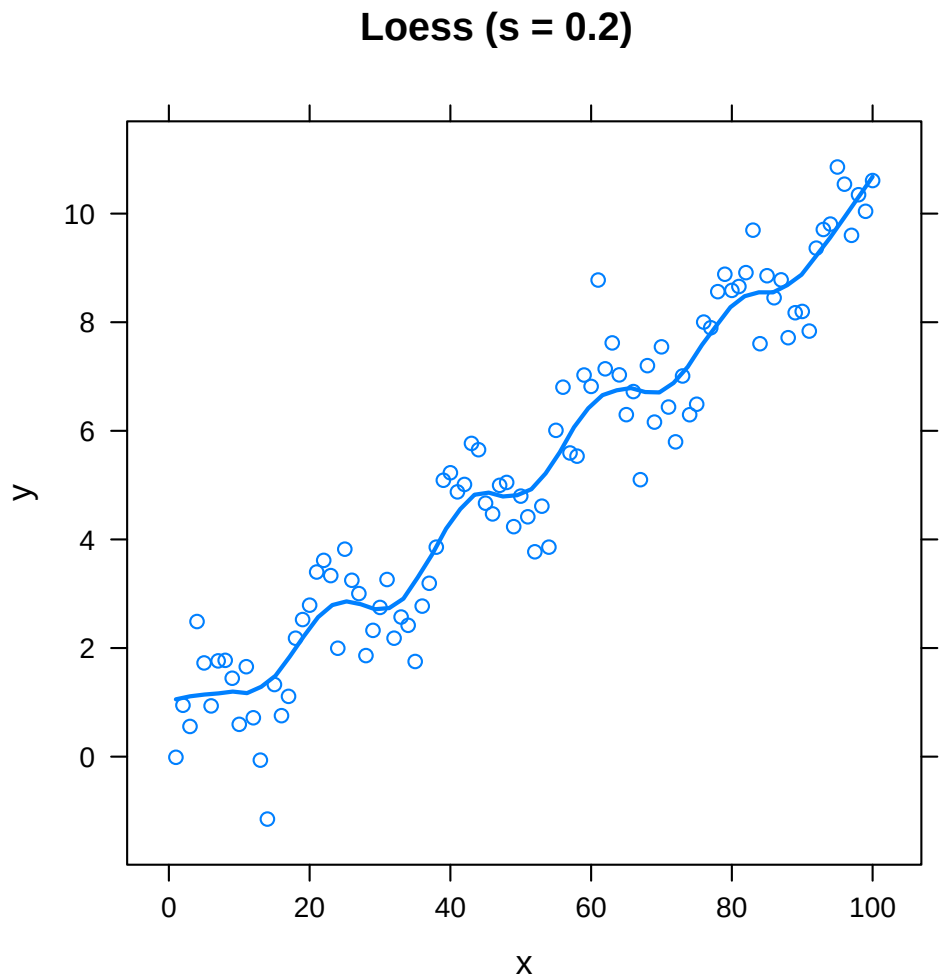
Modellanpassning

Hur bra kan vi anpassa en modell till dessa data?

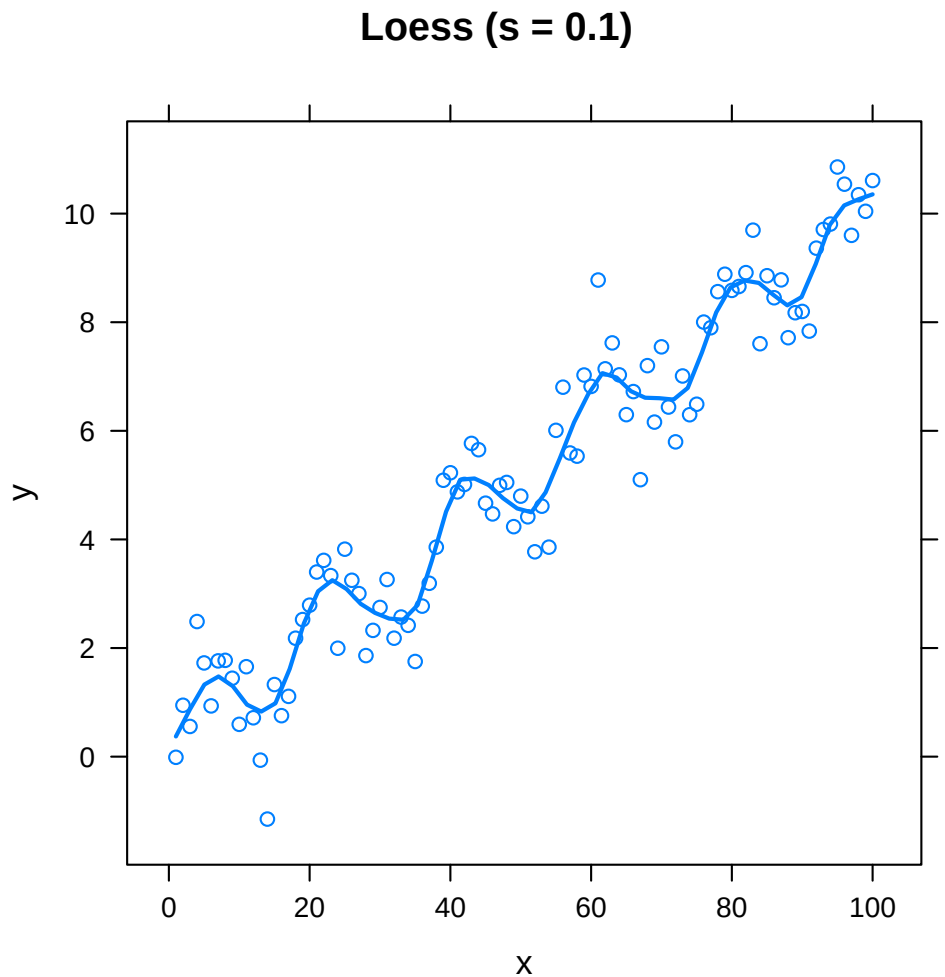
Modellanpassning



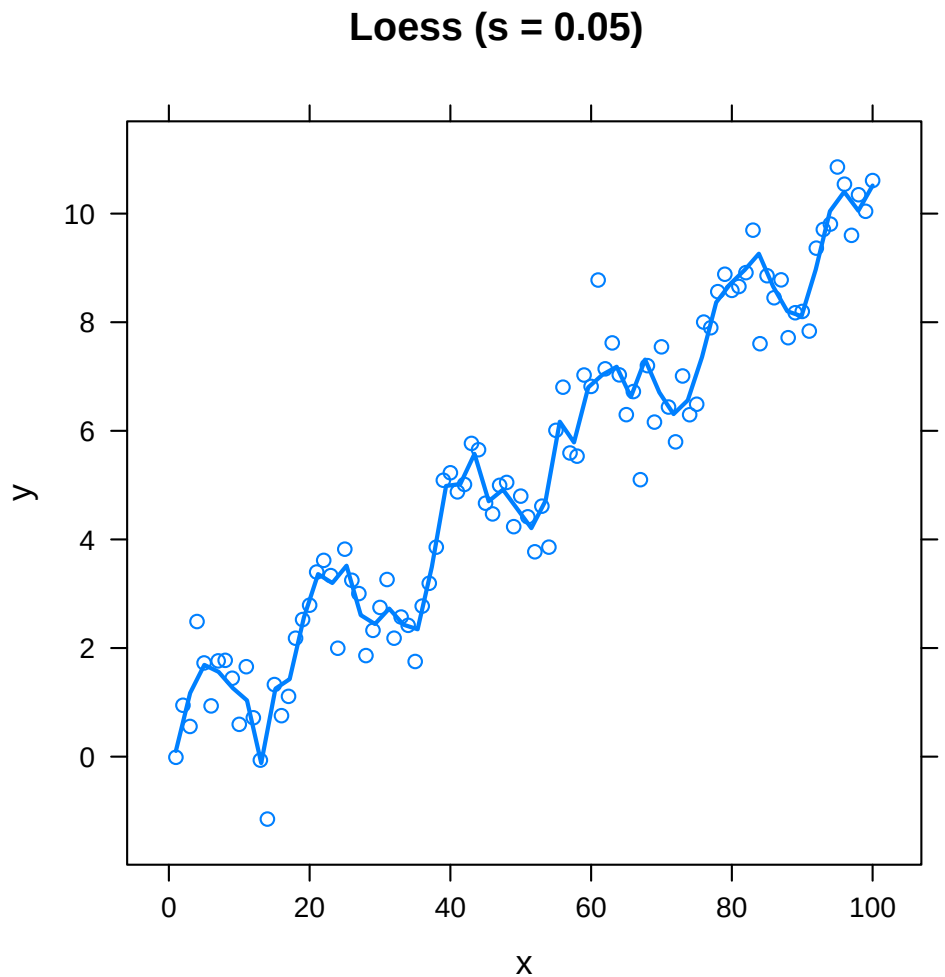
Modellanpassning



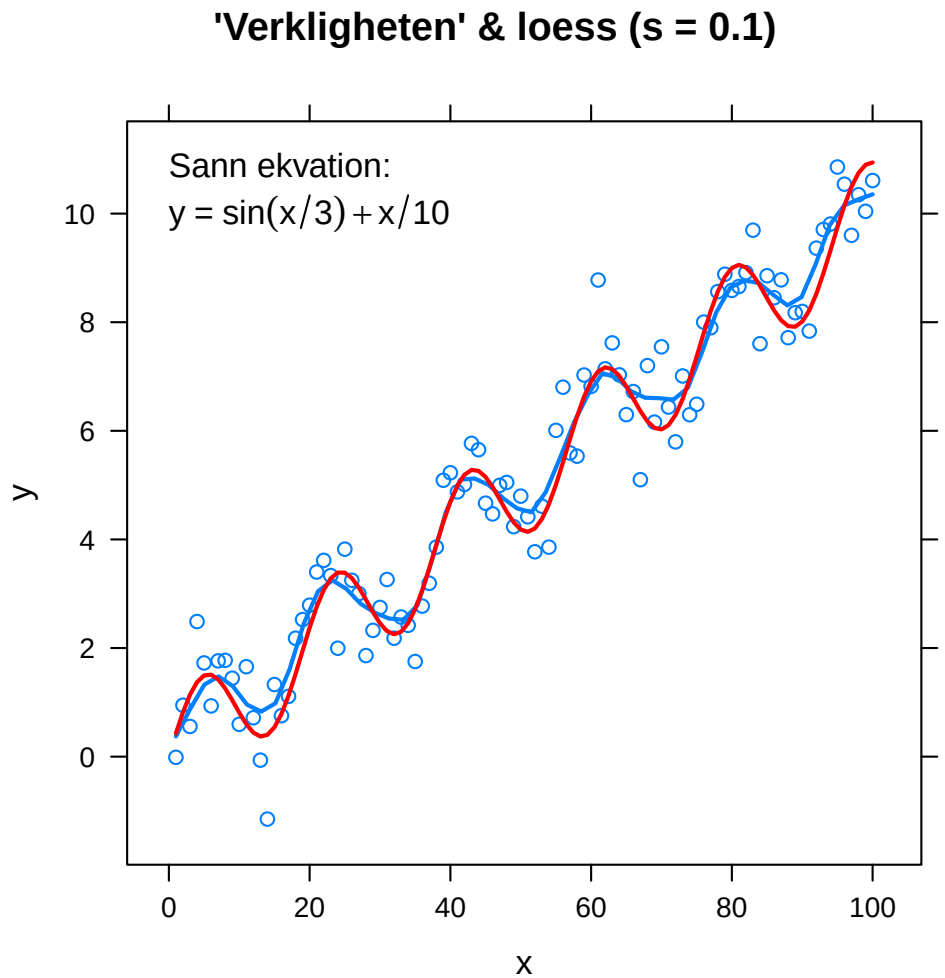
Modellanpassning



Modellanpassning



Modellanpassning



Modellanpassning

Essentially, all models are wrong, but some are useful

(George E. P. Box)

It is better to be vaguely right than exactly wrong

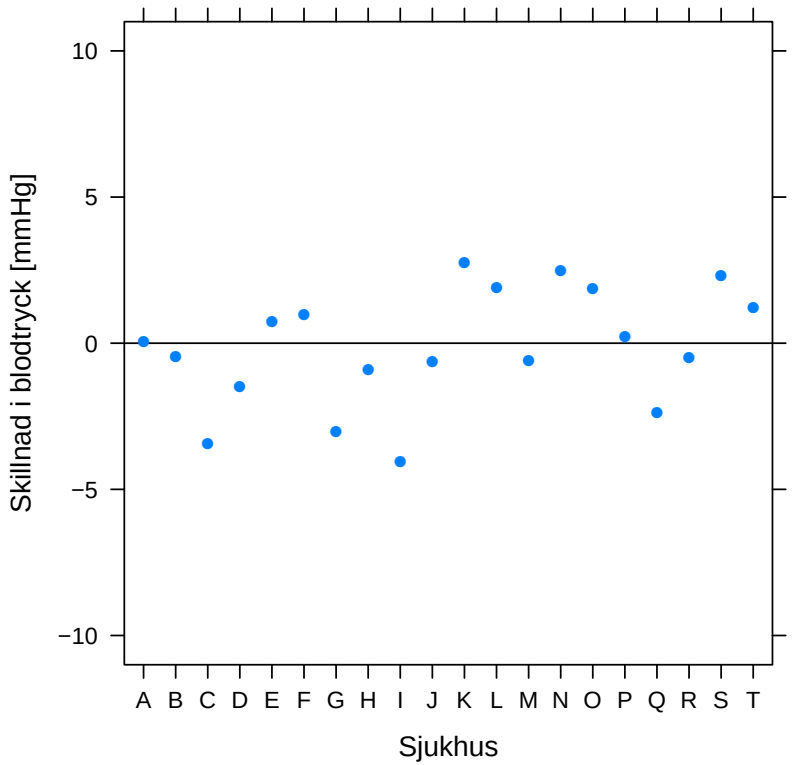
(Carveth Read)

Överanpassning

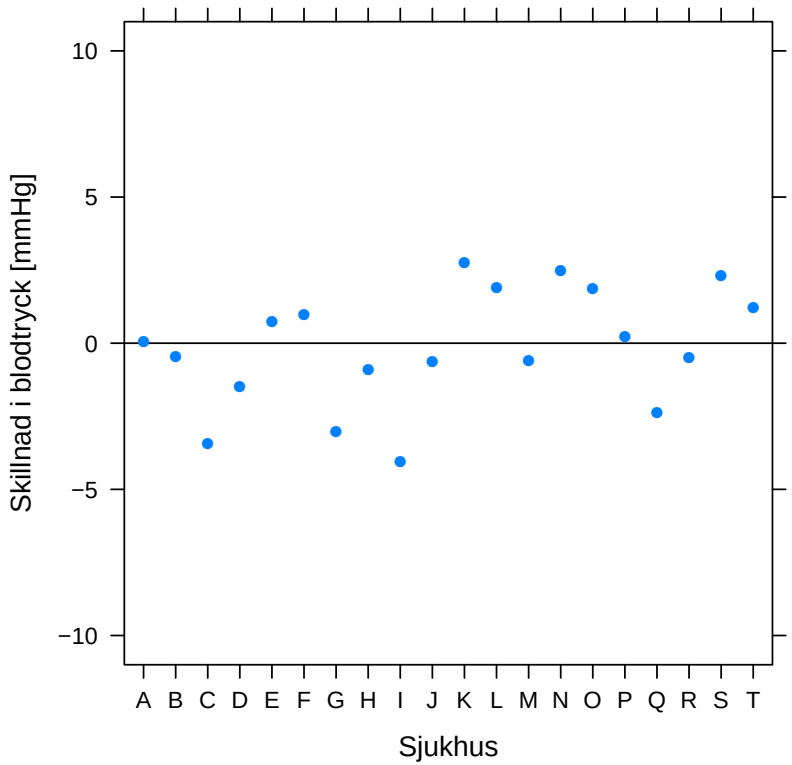
Exempel

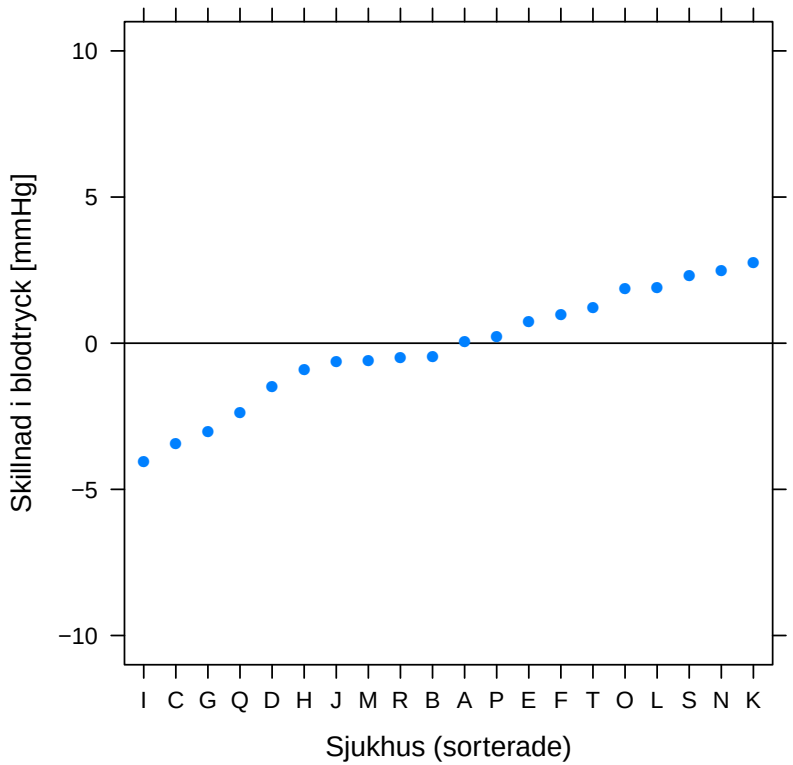
I en studie på 20 sjukhus mättes blodtrycket i både höger och vänster arm på ett antal patienter för att ta reda på om det fanns någon systematisk skillnad. Den genomsnittliga differensen mellan höger och vänster arms blodtryck plottades för alla sjukhus för att se om det fanns någon skillnad mellan sjukhusen

Överanpassning



Överanpassning





Bortfall

Statistikprogram tar typiskt bort alla observationer (rader) där *åtminstone någon* av variablerna saknar sitt värde

Table: First 6 rows of some data ...

Id	Age	Sex	BP	died
1	65	M		0
2	70	M	167	1
3	75	M	143	1
4	72	F		0
5		F	150	0
6	54	F	188	0

Bortfall

- Ju fler variabler med bortfall desto större risk att det endast är få observationer med komplett information
- Helt slumpmässigt bortfall (ovanligt!) $\rightarrow$ precisionen minskar.
- Ej slumpmässigt bortfall $\rightarrow$ systematiskt fel. Ofta kan man framgångsrikt imputera nya värden genom att fånga upp samband med övriga variabler

Variabelval

Vilka variabler?

Hur väljer man vilka variabler som ska vara med i modellen?

- Klinisk erfarenhet
- Vetenskaplig litteratur
- Tillgänglighet
- Inte alls?
- Låt datorn (statistikprogrammet) välja

Variabelval

Hur många variabler?

Hur många variabler kan (ska/bör/får) man ha med i modellen?

- Alla?
- Ingen?
- Några? Hur många?

Variabelval

Hur många variabler?

För att få tillräckligt bra precision finns några tumregler

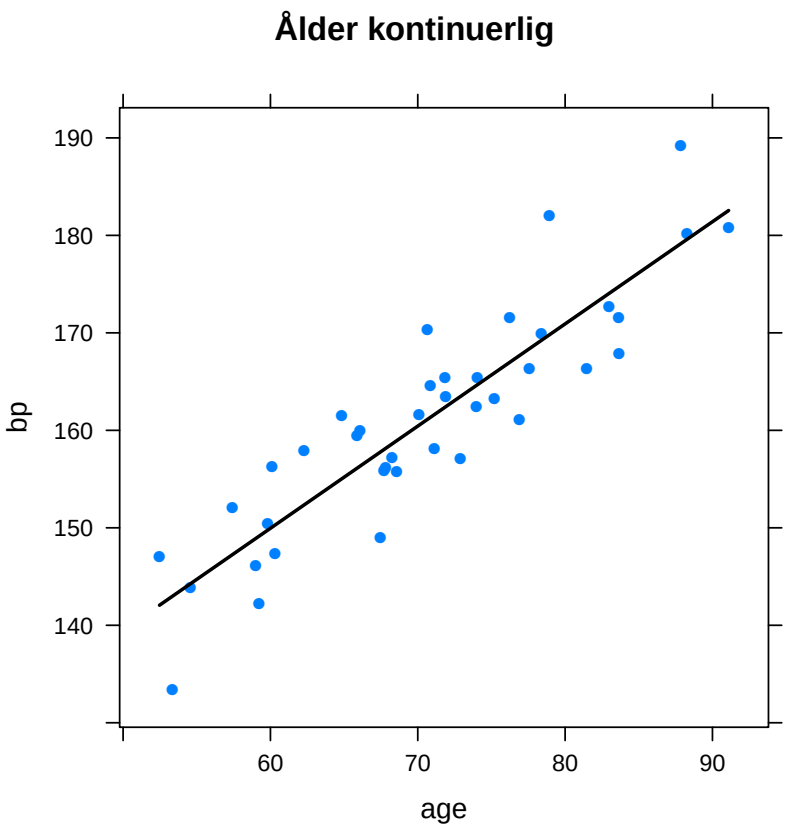
- Linjär regression: minst 10 obs per parameter i modellen
- Logistisk regression: minst 10 obs *i den minsta klassen i utfallet* per parameter i modellen
- Cox-regression: minst 10 händelser per parameter i modellen

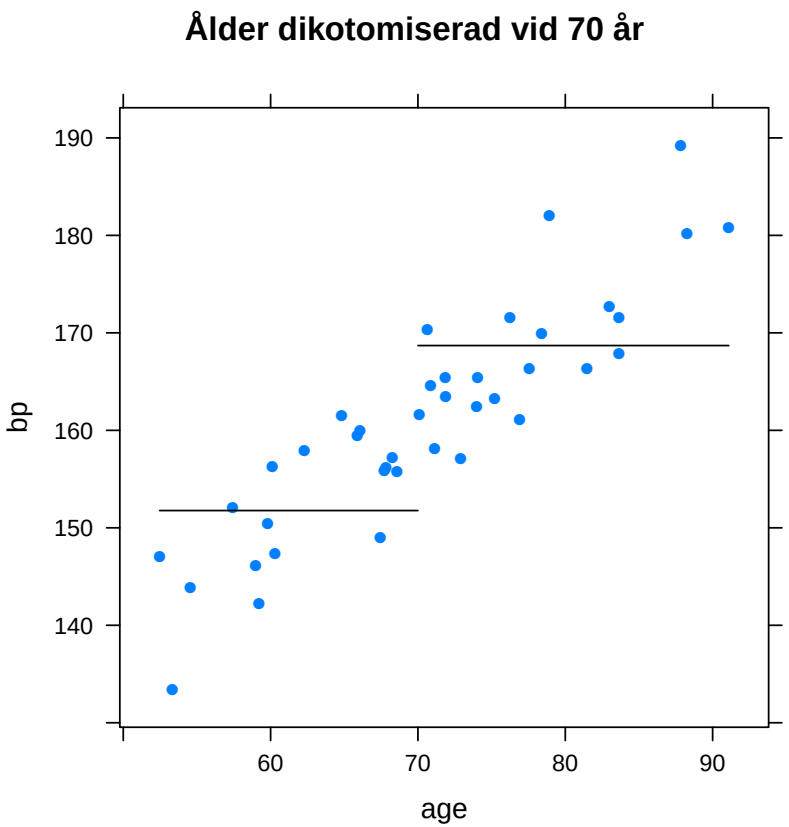
OBS! parameter $\neq$ variabel

Kodning av variabler

- För kategoriska variabler med flera klasser kan det ibland vara vettigt att slå ihop (små) klasser
- Undvik att kategorisera kontinuerliga variabler. Samband är väldigt sällan "stegformade"
- Bättre: anpassa "smarta" splinefunktioner med få parametrar
- Transformationer: log-transformation gör att vi går från additiv till multiplikativ skala

Kategorisering av kontinuerlig variabel





Stegvis regression

- "Forward"
  - Börja utan variabler (medelvärdet)
  - Lägg till en variabel i taget baserat på vilken som är "bäst"
  - Sluta när modellen inte förbättras längre
- "Backward" (bättre)
  - Börja med alla variabler i modellen.
  - Ta bort den variabel som minst påverkar modellen
  - Sluta då modellen signifikant försämras om ytterligare någon variabel tas bort

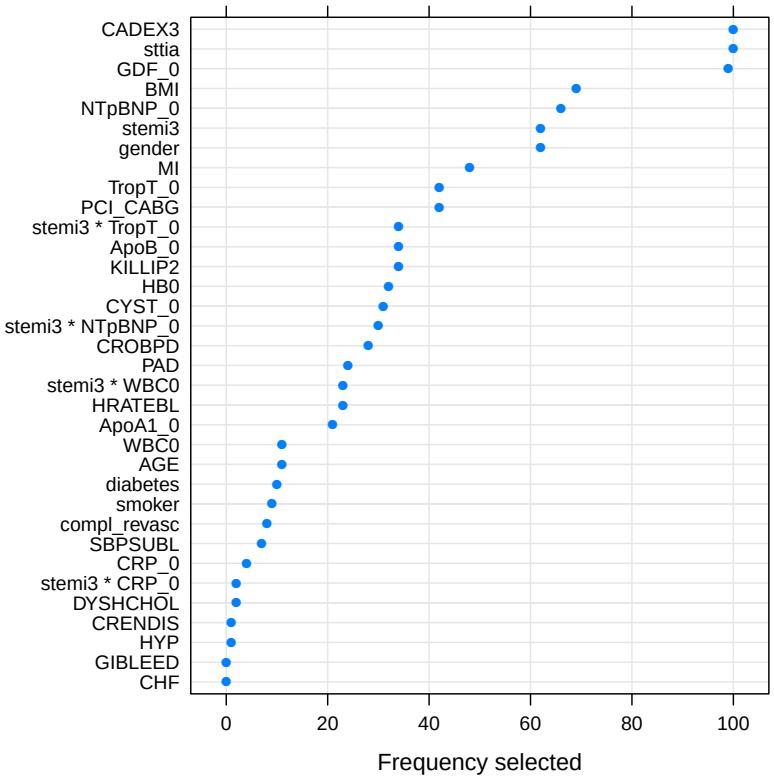
Stegvis regression (exempel)

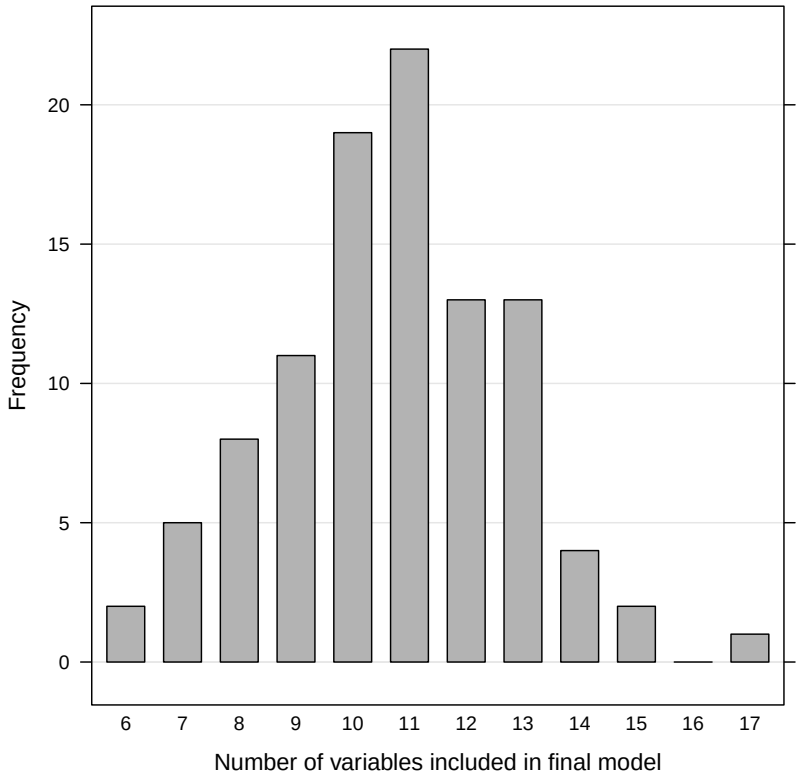
I en studiepopulation bestående av patienter med akut kranskärlssjukdom som genomgått någon form av revaskularisering (ballongvidgning med eller utan stent eller bypass-operation) vill vi ta fram en prediktionsmodell för att bestämma risken för en ny allvarlig händelse (stroke, kardiovaskulär död eller spontan hjärtinfarkt)

- 9448 patienter
- 841 händelser
- 30 kandidatvariabler
- 4 prespecificerade interaktioner
- Icke-linjära transformationer för samtliga kontinuerliga variabler
- Totalt 79 parametrar i modellen

Stegvis regression

100 bootstrapstickprov





Problem med stegvis regression

- Risk för överanpassning
- R^2 systematiskt för hög
- Regressionskoefficienter systematiskt för stora (i absoluta tal)
- Medelfel systematiskt för små
- p-värden systematiskt för låga
- Konfidensintervall systematiskt för smala
- Dålig på att hantera multikolinjäritet (godtycklighet)
- "It allows us to not think about the problem" (F.E. Harrell)
- Automatiska procedurer är sällan implementerade så att de hanterar interaktioner på "rätt" sätt

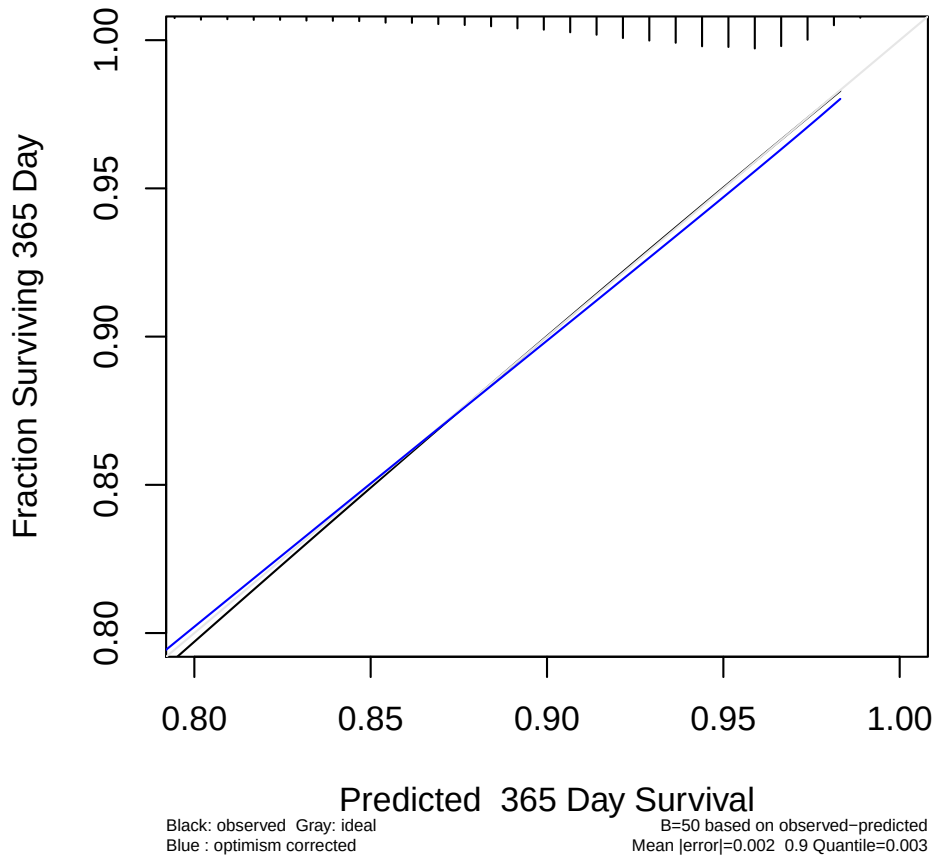
Diskriminering

En modell har god diskriminerande förmåga om den klarar av att skilja på individer med "höga" och "låga" utfall. D.v.s. om modellen förutsäger att individ A har högre risk än individ B och individ A dör innan individ B så har modellen god diskriminerande förmåga

Kalibrering

- Om det predicerade utfallet är lika som det observerade utfallet så är modellen välkalibrerad. Plotta observerat mot predicerat utfall
- Övergripande kalibrering (interceptet på kalibreringskurvan). Om alla predicerade sannolikheter är för höga eller för låga. Det skulle t.ex. kunna bero på att det totala risken har minskats (p.g.a. förändringar i samhället?)
- Lutningen på kalibreringskurvan. Om lutningen < 1 innebär det att låga prediktioner är för låga och höga prediktioner är för höga. Kan vara ett tecken på överanpassning

Kalibrering
Exempel



Validering

- Vi vill veta hur bra modellen generaliserar till nya data
- Vi skiljer på intern och extern validitet

Extern validitet

- Extern validitet är "gold standard"
- Fungerar modellen i den kliniska verkligheten den var framtagen att användas i?
- Helst en annan *region* under en annan *tid* av andra *tillämpare*
- Beräkna prediktioner baserat på den framtagna modellen och jämför med observerat utfall i valideringsdatat (OBS! inte samma sak som att ta fram en ny modell på de nya data)

Intern validitet

- Modellen utvärderas på samma data som den framtagits på
- Ser ofta bra ut p.g.a. överanpassning
- Försök efterlikna extern validitet genom att dela upp data.

Intern validitet

Dela i två delar

- Dela upp i modelleringsdata och ett valideringsdata.
- Vanligen 2/3 för framtagning av modell och 1/3 för validering.
- Slöseri med data. Minskar möjligheten till flexibel modellering.
- Annan indelning ger troligen annat resultat
- Rekommenderas ej!

Intern validitet

Bättre: korsvalidering

- Dela upp data i k st delar.
- Ta fram modellen på samma sätt som för alla data men använd bara $k-1$ av delarna och utvärdera på den del som hölls ute.
- Upprepa k ggr så att alla observationer utvärderats en gång.
- Fungerar ofta ok
- k väljs vanligen till 5 eller 10.
- Kan upprepas (upprepad korsvalidering) med nya indelningar

Intern validitet

Bäst: bootstrap

- Dra nya stickprov (med återläggning) av samma storlek som ursprungsdata
- Ta fram modellen på samma sätt som för alla data
- Utvärdera modellen på orginaldatat
- Beräkna mått för diskrininering, kalibrering, etc. ta genomsnitt över alla bootstrapstickprov.
- Skillnaden mellan dessa och de ursprungliga måtten mäter överoptimismen.
- Genom att ta hänsyn till optimismen kan man få en uppskattning om hur modellen kommer att fungera på "nya" data och justera sin modell utifrån det.