Lineær regression

Et datasæt , hvor og søges forklaret vha. en lineær sammenhæng:

Her er , mens står for *hypotese*. Vi leder altså efter det der giver den hypotese der – i en eller anden forstand - passer bedst med observationerne.

# Vektorisering

Mange formler kan simplificeres ved at skrives dem i vektor/matrixform. Med vektor af dimension menes her en søjlevektor, altså et element i , med mindre andet er nævnt.

Prikproduktet mellem to vektorer af dimension , og , kan f.eks. skrives:

Dette kan benyttes til at finde afledte:

Eller på vektorform:

Hvis er en matrix i stedet for en vektor holder formlen stadig, da relationen gælder søjle for søjle.

En kvadratisk form kan skrives:

Her er en symmetrisk matrix. Hvis man afleder får man:

Eller på vektorform:

## Lineær hypotese på vektorform

Den lineære hypotese kan nu skrives:

Hvis redefineres, så kan dette kort skrives:

## Designmatrix

Til et givent observationssæt defineres *designmatricen* ved:

Ved at gange med kan denne matrix benyttes til at lave *prædiktion*, altså forudsige værdierne af under den givne hypotese:

Hvis man ønsker hypoteser fra en anden familie af funktioner kan designmatricen ændres derefter.

# Mindste kvadraters metode: Normalligningen

Givet en hypotese kan man vurdere hvor godt denne beskriver de observerede data. Dette kan gøres på flere måder, men her vil vi benytte *mindste kvadraters metode*. Vi søger her den hypotese der gør følgende *omkostnings-funktion* (en såkaldt *cost function*) mindst mulig:

Summen kan tænkes på som kvadraten på længden af en vektor, der iflg. ovenstående må være givet ved . Da kan vi igen bruge en formel fra ovenstående afsnit og få:

Vi ønsker at finde den der gør mindst mulig. Derfor differentierer vi efter ved at bruge resultaterne fra sidste afsnit:

I minimum skal gradienten være nul:

Dette kaldes for *normalligningen*. Hvis ikke er invertibel taler man om *perfekt multikollinearitet*.

# Eksempler på brug af normalligningen

## Kun konstantled

Her har hypotesefunktionen den simple form . Designmatricen er dermed blot en søjlevektoren af dimension fyldt med 1-taller. En sådan vektor vil vi benævne (ikke at forveksle med kostfunktionen). Dermed bliver simpelthen lig med en sum over 1-taller, dvs:

Og dermed:

Tilsvarende bliver simpelthen en sum over alle -værdier, altså . Så i alt . Altså simpelthen gennemsnittet af y-værdierne.

## Et lineært led, ingen konstantled – Et datapunkt

Her har hypotesefunktionen formen , altså en ligefrem proportionalitet. Datapunktet vil vi blot kalde . I dette tilfælde er der ikke nogen søjle med 1-taller i designmatricen . I stedet er , så og dermed er:

Derfor er som forventet. Her opstår der multikollinearitet hvis .

## Et lineært led – to datapunkter

Her er hypotesefunktionen . De to punkter vil vi betegne og . Her bliver designmatricen:

Dermed er:

For at invertere denne matrix beregnes determinanten:

Med den sædvanlige notation betyder det . Der er altså kollinearitet når . Nu kan den inverse skrives op:

Vi regner:

Dermed bliver givet med:

Heraf ses umiddelbart den kendte formel . Det er sværere at genkende formlen for . Vi forventer:

Regn på første parentes:

Udtrykket stemmer altså overens med de sædvanlige formler for to punkter.