**MINDSTE KVADRATERS METODE**

# Afsnit 1. Bedste proportionale model

I disse noter vil vi udlede formler til bestemmelse af bedste proportionale model, dvs. bestemme ligningen for den rette linje gennem , som bedst tilpasser et givet datasæt. Vi vil tillige udlede et udtryk for den såkaldte korrelationskoefficient, samt definere begrebet forklaringsgrad, som vi kender fra praktiske anvendelser af regression.

Vi tænker os givet datapunkterne , hvis førstekoordinater ikke alle er . For at bestemme ligningen for den proportionale model , der tilpasser datapunkterne bedst, betragter vi kvadratsummen

Vores opgave er at bestemme således, at bliver mindst mulig.

Vi ganger først parenteserne ud:

Dernæst deler vi sumudtrykket op i tre summer:

Der er i statistik tradition for at indføre særlige symboler for en række hyppigt forekommende sumudtryk. Således betegner man summen

med , der er en forkortelse for *Sum af Kvadrater* *på* *-erne*.

I tabellen øverst på næste side har vi for overskuelighedens skyld samlet symboler og formler for alle de sumudtryk, der anvendes i disse noter:

|  |  |  |
| --- | --- | --- |
| Symbol | Formeludtryk | Forklaring af forkortelse og eventuelle bemærkninger |
|  |  | *S*um af *K*vadrater på *-*er |
|  |  | *S*um af *P*rodukter mellem *-*er og -er |
|  |  | *S*um af *K*vadrater på *-*er  Kaldes også total variation |
|  |  | *S*um af *K*vadrater på funktionsværdier  Kaldes også forklaret variation |
|  |  | *S*um af *R*esidual*K*vadrater  Kaldes også uforklaret variation |

Dermed kan udtrykkes som

Da førstekoordinaterne ikke alle er , er kvadratsummen

er følgelig et andengradspolynomium i , hvis graf er en parabel, der vender grenene opad. Det globale minimumssted for er da førstekoordinaten for parablens toppunkt.

Da toppunktet for parablen

har førstekoordinat , får vi:

*Figur 1.1. Toppunktets førstekoordinat for grafen for .*

Vi har dermed bevist følgende sætning:

* Sætning 1.2.*Lad punkterne* *være givne og antag, at deres førstekoordinater ikke alle er . Den proportionale model*

*der tilpasser data bedst, har proportionalitetskonstant*

I eksempel 1.3 viser vi, hvorledes sætning 1.2 kan anvendes til at bestemme ligningen for den bedste proportionale model i et simpelt taleksempel. I praksis er sætningen naturligvis ikke hensigtsmæssigt at anvende; det er langt lettere at benytte et værktøjsprogram.

* ***Eksempel 1.3.*** Vi ønsker at bestemme ligningen for den proportionale model , der bedst tilpasser datapunkterne i nedenstående tabel:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 1,00 | 1,00 |
| 2 | 2,00 | 1,00 |
| 3 | 3,00 | 3,00 |
| 4 | 4,00 | 3,00 |

Det er praktisk at tilføje to søjler til tabellen med henholdsvis produkterne og kvadraterne , samt en række med disse to søjlers sum:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 1 | 1,00 | 1,00 | 1,00 | 1,00 |
| 2 | 2,00 | 1,00 | 2,00 | 4,00 |
| 3 | 3,00 | 3,00 | 9,00 | 9,00 |
| 4 | 4,00 | 3,00 | 12,0 | 16,0 |
| Sum |  |  |  |  |

Dermed kan vi beregne proportionalitetskonstanten til

Den søgte proportionalitet er da

# Afsnit 2. korrelationskoefficienten

Vi vil nu udlede udtryk for den såkaldte korrelationskoefficient for bedste proportionale model. Udledningen giver os mulighed for at opnå en forståelse af betydningen af korrelationskoefficienten og dermed give en mere præcis fortolkning af .

For at bestemme ligningen for den proportionale model , der bedst tilpasser datapunkterne , hvis førstekoordinater ikke alle er ens, betragtede vi i afsnit 2 kvadratsummen

Vi omskrev til formen

og så, at er et andengradspolynomium i , hvis graf er en parabel, der vender grenene opad. Den værdi af , som gør mindst mulig, er førstekoordinaten for parablens toppunkt:

Den mindste værdi af kvadratsummen er da toppunktets andenkoordinat. Denne mindsteværdi betegner vi , som er en forkortelse for *Sum af ResidualKvadrater*.

*Figur 2.1. Toppunktets beliggenhed for grafen for .*

Da toppunktet for parablen

er givet ved , hvor diskriminanten , udregner vi først diskriminanten for :

Dermed får mindsteværdien

Vi deler udtrykket for op i to brøker og reducerer:

Hvis vi yderligere forudsætter, at datapunkternes andenkoordinater ikke alle er 0, gælder . Dermed kan vi sætte udenfor parentes:

For bedste proportionalitet definerer vi korrelationskoefficienten ved ligningen

og bemærker, at

Dermed har vi følgende udtryk for :

Da er en sum af kvadrater, gælder

idet .

Da er en sum af kvadrater, er hvis og kun hvis ethvert af kvadraterne i summen er . Dette er tilfældet, hvis og kun hvis datapunkterne alle ligger på linjen . I så fald gælder

Tillige bemærker vi, at jo tættere er på enten eller på , jo tættere er på og jo tættere er på . Sagt på en anden måde: Jo tættere er på enten eller på , jo bedre tilnærmer bedste proportionalitet datapunkterne.

Endelig noterer vi os, at proportionalitetskonstanten og korrelationskoefficienten har samme fortegn, idet de har samme tæller , mens de begge har positiv nævner:

Vi har dermed bevist følgende sætning:

* Sætning 2.2.*Lad punkterne* *være givne og antag, at hverken deres første- eller andenkoordinater alle er . Korrelationskoefficienten for bedste proportionale model er givet ved*

*har samme fortegn som proportionalitetskonstanten , og opfylder . Alle datapunkter ligger på bedste proportionale model, hvis og kun hvis eller .*

* ***Eksempel 2.3.*** Vi vender tilbage til den proportionale model , der bedst tilpasser datapunkterne i tabellen fra eksempel 1.3:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 1,00 | 1,00 |
| 2 | 2,00 | 1,00 |
| 3 | 3,00 | 3,00 |
| 4 | 4,00 | 3,00 |

Vi ønsker nu at bestemme summen af residualkvadrater , samt korrelations-koefficienten .

I eksempel 1.3 fandt vi den bedste proportionale model til

Vi lader betegne proportionaliteten med forskriften

Vi kan dermed beregne *modelværdien*, *residualet* og *residualkvadratet* for ved indsættelse i forskriften for :

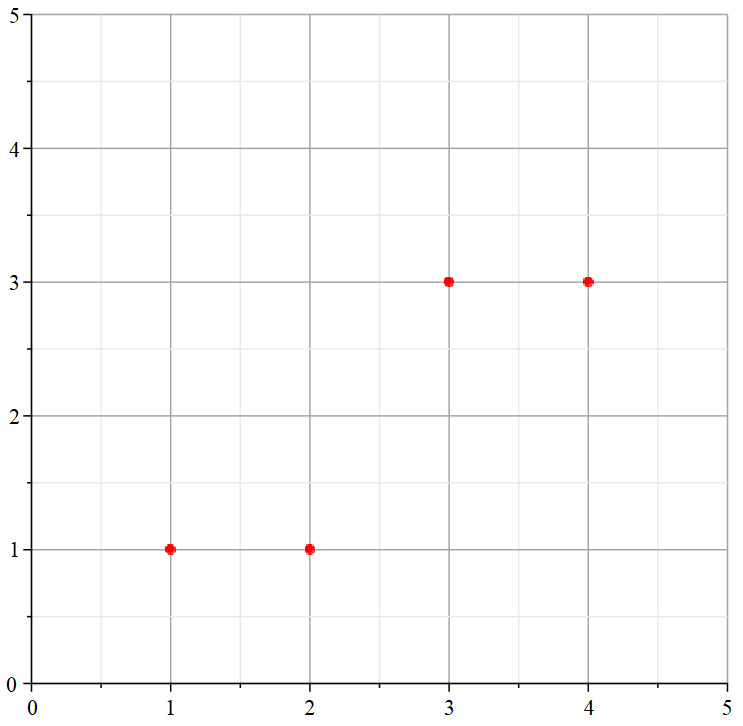
På denne måde kan vi nu tilføje tre søjler med henholdsvis modelværdier, residualer og residualkvadrater til tabellen:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 1 | 1,00 | 1,00 | 0,80 | 0,20 | 0,04 |
| 2 | 2,00 | 1,00 | 1,60 |  | 0,36 |
| 3 | 3,00 | 3,00 | 2,40 | 0,60 | 0,36 |
| 4 | 4,00 | 3,00 | 3,20 |  | 0,04 |
| Sum |  |  |  |  |  |

Dermed har vi

Summen af arealerne af de røde kvadrater på figur 2.4 på næste side er således .

Man opfatter undertiden residualerne som den del af -erne, der *ikke* forklares af den proportionale model. Derfor kalder man somme tider for *den* *uforklarede variation*.



*Figur 2.4. Bedste proportionale model er den linje gennem ,*

*som gør summen af arealerne af de røde kvadrater mindst mulig.*

For at beregne korrelationskoefficienten er det er hensigtsmæssigt at tilføje tre søjler til den oprindelige tabel med henholdsvis kvadraterne , produkterne og kvadraterne , samt en række med de tre sidste søjlers summer:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 1 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| 2 | 2,00 | 1,00 | 4,00 | 2,00 | 1,00 |
| 3 | 3,00 | 3,00 | 9,00 | 9,00 | 9,00 |
| 4 | 4,00 | 3,00 | 16,0 | 12,0 | 9,00 |
| Sum |  |  |  |  |  |

Vi beregner nu korrelationskoefficienten ved hjælp af formlen fra sætning 2.2:

Vi bemærker, at helt i overensstemmelse med sætning 2.2, idet proportionalitetskonstanten er positiv.

# Afsnit 3. Forklaringsgraden

Vi vil nu indføre *forklaringsgraden* – der også ofte betegnes *determinationskoefficienten* – som et mål for hvor stor en del af den totale variation, der forklares af den proportionale model. Vi vil indlede med at forklare begrebet i et konkret eksempel:

* ***Eksempel 3.1.*** Vi betragter igen tabellen fra eksempel 1.3 og 2.3:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 1,00 | 1,00 |
| 2 | 2,00 | 1,00 |
| 3 | 3,00 | 3,00 |
| 4 | 4,00 | 3,00 |

I eksempel 1.3 bestemte vi forskriften for proportionalitet , der bedst tilnærmer datapunkterne, til

Tillige beregnede vi i eksempel 2.3 summen af residualkvadraterne til

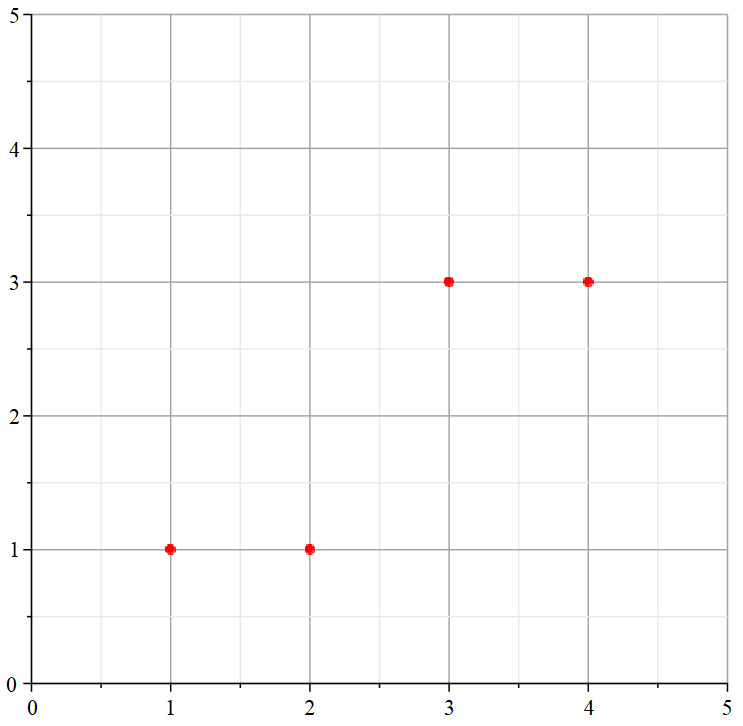
og korrelationskoefficienten til

Vi benævnte den uforklarede variation, som vi geometrisk tolkede som summen af arealerne af de røde kvadrater på figur 2.4.

I overensstemmelse med denne sprogbrug vil vi betegne summen

som *den* *totale variation* af -erne. Geometrisk kan vi tolke vi som summen af arealerne af de blå kvadrater på figur 3.2 på næste side.

Fra eksempel 2.3 ved vi, at , jævnfør tabellen side 20.



*Figur 3.2. Den totale variation er summen af arealerne af de blå*

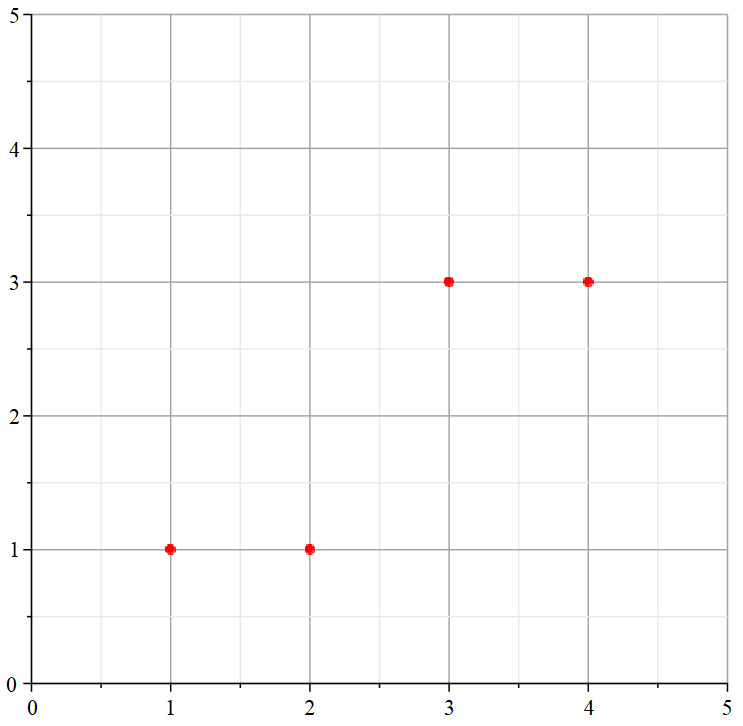
*Kvadrater. Bemærk, at de to kvadrater til højre delvist overlapper.*

Tilsvarende vil vi betegne *den forklarede variation*, altså den variation, der *kan* forklares af den proportionale model:

Geometrisk kan vi tolke vi som summen af arealerne af de grønnekvadrater på figur 3.3 nederst på næste side.

I nedenstående tabel har vi samlet de søjler fra tabellerne på side 8, der er relevante for beregning af den uforklarede variation , den totale variation og den forklarede variation :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 1 | 1,00 | 1,00 | 0,04 | 1,00 | 0,640 |
| 2 | 2,00 | 1,00 | 0,36 | 1,00 | 2,56 |
| 3 | 3,00 | 3,00 | 0,36 | 9,00 | 5,76 |
| 4 | 4,00 | 3,00 | 0,04 | 9,00 | 10,24 |
| Sum |  |  |  |  |  |



*Figur 3.3. Summen af arealerne af de grønne kvadrater*

*kaldes den forklarede variation. I eksemplet er .*

Vi lægger mærke til, at

idet

Vi har altså i dette tilfælde, at den totale variation er lig summen af den forklarede variation og den uforklarede variation.

Geometrisk kan vi tolke dette således, at summen af arealerne af de blå kvadrater er i lig med summerne af arealerne af de røde og de grønne kvadrater til sammen, jævnfør figur 3.4 nedenfor.

|  |  |
| --- | --- |
|  | |
|  |  |

*Figur 3.4. Summen af arealerne af de blå kvadrater er lig summerne*

*af arealerne af de røde og de grønne kvadrater til sammen.*

*Forklaringsgraden* defineres som forholdet mellem den forklarede variation og den totale variation:

Vi har dermed

Ved konkret udregning får vi i dette tilfælde

Vi vender os igen mod igen det almene tilfælde. Vi lader altså betegne datapunkter, som tillige opfylder at hverken punkternes førstekoordinater eller andenkoordinater alle er .

Vi lader betegne den proportionalitet, hvis graf netop er den rette linje gennem , der tilpasser datapunkterne bedst muligt:

I afsnit 2 beviste vi, at i så fald er kvadratsummen

mindst mulig, samt at

Som nævnt i eksempel 1.3 omtaler man

som *den totale variation af -erne*, dvs. den samlede kvadratsum. Geometrisk kan tolkes som summen af kvadraterne på de blå afstande på figur 3.5 på næste side.

Vi betegner som tillige kvadratsummen

*den forklarede variation*, da er den del af , der forklares af den proportionale model, jævnfør figur 3.5.

Endelig kalder vi summen af residualkvadrater

for *den uforklarede variation*, da residualerne ikke forklares af den proportionale model, jævnfør figur 3.5.

*Figur 3.5. Den proportionale model forklarer ,*

*men ikke , af dataværdien .*

Vi vil nu bevise sætning 3.6, som siger, at resultatet

fra eksempel 3.1 er et almengyldigt resultat:

* Sætning 3.6.*Lad punkterne* *være givne og antag, at hverken deres første- eller andenkoordinater alle er . Lad betegne den proportionalitet, hvis graf er den rette linje gennem , der tilnærmer punkterne bedst. Så er den totale variation lig summen af den forklarede variation og den uforklarede variation:*

Med samme farvesymbolik som i eksempel 3.1 kan vi illustrere resultatet i sætningen således:

Forklaret variation

Uforklaret variation

Total variation

*Bevis:*

I afsnit 2 (se midt på side 6) beviste vi at kvadratsummens mindsteværdi er givet ved formlen

Forlænger vi brøken med , fås

idet vi ved sidste lighedstegn udnytter, at

Men da kvadraterne på funktionsværdierne kan omskrives

har vi, at

Følgelig gælder

Dermed er sætningen bevist.

⎕

I eksempel 3.1 så vi, at kvadratet på korrelationskoefficienten var lig med forklaringsgraden:

Vi vil nu bevise, at også dette også er en almengyldig regel:

* Sætning 3.7.*Lad punkterne* *være givne og antag, at hverken deres første- eller andenkoordinater alle er . Lad betegne den proportionalitet, hvis graf er den rette linje gennem , der tilnærmer punkterne bedst. Korrelationskoefficienten og forklaringsgraden model defineres ved*

*Kvadratet på korrelationskoefficienten er lig forklaringsgraden:*

*Bevis:*

Kvadratet på korrelationskoefficienten er givet ved

Forlænger vi brøken med , fås

idet vi ved sidste lighedstegn igen udnytter, at .

Af beviset for sætning 3.6 fremgår, at

Følgelig gælder

Dermed er sætningen bevist.

⎕

Med samme farvesymbolik som i eksempel 3.1 kan vi illustrere resultatet i sætning 3.7 således:

Forklaret variation

Total variation