Regressionsanalyse

materiale til statistikkursus

Larsen, Jørgen

Publication date:
1993

Document Version
Også kaldet Forlagets PDF

Citation for published version (APA):
Larsen, J. (1993). *Regressionsanalyse: materiale til statistikkursus*. Roskilde Universitet. Tekster fra IMFUFA Nr. 254 http://milne.ruc.dk/ImfufaTekster/

TEKST NR 254 1993

REGRESSIONSANALYSE

Materiale til et statistikkursus

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

Juli 1993

TEKSTER fra

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES
FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.

Jørgen Larsen: **REGRESSIONSANALYSE — Materiale til et statistikkursus**

IMFUFA tekst nr. 254/1993 174 sider. ISSN 0106-6242

Disse noter er en introduktion til grundideerne i den klassiske statistiske disciplin *regressionsanalyse*; som noget lidt mere avanceret handler Kapitel 10 om *generaliseret lineær regressionsanalyse*, hvilket bl.a. omfatter regressionsanalyse af Poissonfordelte og binomialfordelte observationer, logistisk regression og log-lineære modeller i Poissonfordelingen.

De enkelte kapitler er indrettet på den måde, at der først præsenteres statistisk teori og tilhørende metode, og dernæst omtales hvorledes man i praksis kan udføre det ved hjælp af statistikprogrammet ISP* som læseren bør have til sin rådighed. Hvert kapitel afsluttes med *opgaver*, bl.a. af praktisk art.

Der forudsættes kun beskedne matematiske og statistiske forkundskaber, men nok en vis fortrolighed med formelmanipulation og "bogstavregning". I øvrigt er teksten forsynet med *boxe* til mindre digressioner og *appendices* til større, således vil læsere med kendskab til lineær algebra forhåbentlig have glæde af Appendiks A.

---

*ISP står for *Interactive Scientific Processor* og er © Artemis Systems, Inc.

Indhold

Forord

Statistikkurser for ikke-statistikere kan tilrettelægges på mangfoldige måder. Ofte består de i en gennemgang af fire-fem såkaldte standardmetoder, garneret med opgaver og eksempler fra det fagområde som deltagerne i særlig grad kommer fra; resultatet heraf bliver let, at deltagerne får det fejlagtige indtryk at de nu har lært eller i det mindste fået afgrænset hvad de behøver vide om statistik, og at de får den fejlagtige opfattelse at der til enhver problemstilling hvori der indgår tal findes en statistisk metode som er den rigtige.

Det kunne imidlertid tænkes at det var andre ting man skulle bruge et kursus i statistik for ikke-statistikere til. Kursets opgave skulle måske ikke være indlæring af visse (korttidsholdbare) betingede reflekser vedr. valg af statistisk model/metode. Man kunne mene, at deltagerne overvejende skulle anvende deres tid og kræfter til at blive gode til det der nu er deres fag, og så overlade statistikken til dem der har forstand på dén. Statistikkursets opgave skulle så være at give et indtryk af, hvad det er man kan (og ikke kan) med statistik og statistiske modeller, hvad det er man kan spørge statistikeren om, og hvad der er for nogle underlige svar han/hun giver; det skulle gøre deltagerne bekendte statistikkens tosidede væsen, dels dens side som videnskabsfag hvor matematikkens krav om klarhed og eksakthed hersker, og dels dens anvendelsesside hvor den indgår i en erkendelsesproces og hvor det er meningsløst at hævde at en bestemt model/metode er den rigtige.

Nærværende noter er tænkt til brug i et statistikkursus der søger at leve op til nogle af disse ambitiøse mål. Det er nok vigtigt at understrege, at noterne netop kun er en *del*: en anden væsentlig side af sagen er naturligvis den mundtlige præsentation og samtale, ligesom det er særdeles afgørende at deltagerne regner de tilhørende opgaver.

•

Noterne er ikke overraskende bygget op på den måde, at de begynder med noget simpelt, og så bliver det sværere efterhånden. Men det er en god pointe, at det på sin vis er den samme problemstilling hele vejen igennem,

nemlig hvad man kunne kalde en regressionssituation: Der foreligger et antal sammenhørende værdier af en størrelse x og en størrelse y, hvor man ønsker at opfatte x som en baggrundsvariabel ("forklarende variabel") og y som den med usikkerhed behæftede størrelse (den "stokastiske variabel"); målet er at give en beskrivelse af dette talmateriale ved hjælp af en statistisk model.

I Kapitel 2 præsenteres den simple lineære regressionsmodel; parametrene estimeres med mindste kvadraters metode. Fra første færd præciseres nødvendigheden af modelkontrol; en af de simpleste og samtidig bedste former for modelkontrol er en tegning af de observerede punkter og den fittede kurve. Det kan hænde at man efter at have set på sådan en tegning ønsker at fitte en anden kurve end en ret linie, en andengradskurve måske. Kapitel 3 handler om generalisationen af simpel lineær regression til polynomiel regression, samt om trigonometrisk regression.

Et særligt træk ved statistiske modeller er som bekendt, at de beskriver både den systematiske og den tilfældige variation. I regressionsmodeller benytter man meget ofte normalfordelingen til at beskrive hvordan observationernes tilfældige variation; derfor er der et Kapitel 4 med en kort omtale af normalfordelingen, inden Kapitel 5 præsenterer den egentlige statistiske model for simpel lineær regression med normalfordelte fejl. Spørgsmålet om modelkontrol tages op til fornyet behandling med forskellige former for plots der kan vise om residualerne faktisk har de af modellen postulerede egenskaber.

I Kapitel 6 ser vi på noget af det som en statistisk model kan levere (efter at man har estimeret de ukendte parametre), nemlig udsagn om estimatorernes fordelinger (f.eks. i form af middelfejl eller af sikkerhedsintervaller); i den forbindelse kommer man naturligt ind på test af statistiske hypoteser og på brugen heraf i modelkontrolsammenhænge; nogle af opgaverne handler om anvendelsen af disse ting ved planlægning af forsøg.

Indholdet af Kapitel 6 er afgjort vanskeligt fordøjeligt, så det tages op på forskellig vis i alt det efterfølgende, dog ikke i Kapitel 7 hvor vi holder fri fra de statistiske begreber for at se på visualisering af flerdimensionale datasæt, blandt andet som en forberedelse til Kapitel 8 om multipel regression. I Kapitel 8 diskuteres først nogle modelleringsmæssige fordele og ulemper ved at have mange mulige forklarende variable, og derefter studeres nogle eksempler på hvordan den generelle multiple regressionsmodel kan specialiseres: til ensidet variansanalyse (Afsnit 8.6) og til en model til sammenligning af regressionslinier (Afsnit 8.7).

I de almindelige statistiske regressionsmodeller indgår en antagelse om at alle y-erne har samme varians (dvs. at der er varianshomoscedasticitet). Hvis denne antagelse ikke er opfyldt, men hvis man dog véd på hvilken måde varianserne er forskellige, kan man estimere parameterene med vægtede mindste kvadraters metode. Dette omtales i Kapitel 9. Det præciseres (i Afsnit 9.6),

at det i den forbindelse faktisk spiller en rolle om man transformerer sine y-værdier (for at opnå at den systematiske sammenhæng bliver lineær).

Det sidste og sværeste kapitel, Kapitel 10, indeholder endnu en udvidelse af regressionsmodellen, idet det nemlig introducerer modellen for generaliseret lineær regression, dog væsentligst med henblik på at kunne analysere modeller med Poissonfordelte eller binomialfordelte (eller normalfordelte) fejl, blandt andet logistisk regression.

•

Det kursus som noterne lægger op til, er ikke et kursus der indøver et vist antal standardmetoder, men derimod et kursus der skal vise at statistisk modelbygning er (eller kan være) en langvarig kreativ proces, der typisk indbefatter en del beregninger og tegninger.

Da man nutildags mest hensigtsmæssigt laver tegninger og beregninger på computer, er noterne/kurset lagt an på, at deltagerne sideløbende lærer at anvende et statistikprogram, nemlig programmet ISP[†] som er et forholdsvis let tilgængeligt statistikprogram der kører på almindelige DOS-maskiner. ISP har et relativt beskedent udvalg af "standardmetoder", men er et meget velegnet hjælpemiddel i den kreative modelbygningsproces.

•

Noterne kræver ikke de store matematiske forkundskaber i teknisk forstand, men nok en vis vant-hed til det matematiske formelsprog. Det forventes at deltagerne på forhånd (f.eks. i gymnasiet) har stiftet bekendtskab med et matematisk sandsynlighedsbegreb og med begreber som middelværdi og varians.

Det gennemgående princip til estimation af parametre er mindste kvadraters metode, der udmøntes i løsning af lineære ligninger. Mindste kvadraters metode er, i al fald når den bruges i multipel regression, et lineær algebra-problem, så derfor angribes minimaliseringsproblemet *ikke* ved at udregne partielle afledede og sætte dem lig 0. I stedet benyttes en lineær algebra-inspireret fremgangsmåde; til glæde for dem der kan lineær algebra, omtales i Appendiks A den "rigtige" måde at gøre tingene på.

[†]ISP står for *Interactive Scientific Processor* og er © Artemis Systems, Inc.

Kapitel 1

Indledning: Hvad er regressionsanalyse?

Regressionsanalyse handler om at undersøge hvordan én målt størrelse afhænger af en eller flere andre.

Antag at der foreligger et statistisk datamateriale, som er fremkommet på den måde, at man på hvert af nogle "individer" (f.eks. forsøgspersoner eller forsøgsdyr eller enkelt-laboratorieforsøg osv.) har målt værdien af et antal størrelser (variable). En af disse størrelser indtager en særstilling, idet man nemlig gerne vil "beskrive" eller "forklare" denne størrelse ved hjælp af de øvrige. Tit kalder man den variabel der skal beskrives for y, og de variable ved hjælp af hvilke man vil beskrive for $x_1, x_2, \ldots, x_p$. Andre betegnelser fremgår af følgende oversigt:

y	$x_1, x_2, \ldots, x_p$
den modellerede variabel	baggrundsvariable
den afhængige variabel	de uafhængige variable
den forklarede variabel	de forklarende variable
responsvariabel	

Her skitseres et par eksempler:

1. Lægen observerer den tid y som patienten overlever efter at være blevet behandlet for sygdommen, men lægen har også registreret en mængde baggrundsoplysninger om patienten, så som køn, alder, vægt, detaljer om sygdommen osv. Nogle af baggrundsoplysningerne kan måske indeholde information om hvor længe patienten kan forventes at overleve.

2. I en række nogenlunde ens i-lande har man bestemt mål for lungekræftforekomst, cigaretforbrug og forbrug af fossilt brændstof, altsammen pr. indbygger. Man kan da udnævne lungekræftforekomst til y-variabel og søge at "forklare" den ved hjælp af de to andre variable, der så får rollen som forklarende variable.

3. Man ønsker at undersøge et bestemt stofs giftighed. Derfor giver man det i forskellige koncentrationer til nogle grupper af forsøgsdyr og ser hvor mange af dyrene der dør. Her er koncentrationen x en uafhængig variabel hvis værdi eksperimentator bestemmer, og antallet y af døde er den afhængige variabel.

Regressionsanalyse går kort fortalt ud på at finde en statistisk model hvormed man kan beskrive en y-variabel ved hjælp af en kendt simpel funktion af nogle baggrundsvariable og nogle såkaldte *parametre*. Parametrene er de samme for alle observationssæt, hvorimod baggrundsvariablene typisk ikke er det. Parametrenes værdier bestemmes ud fra data således at man får det bedste *fit*.

Man må naturligvis ikke forvente at den statistiske model leverer en perfekt beskrivelse, et perfekt fit, dels fordi den model man måtte finde frem til næppe er fuldstændig rigtig, dels fordi en af pointerne med statistiske modeller netop er, at de kun beskriver hovedtrækkene i datamaterialet og ser stort på de finere detaljer. Der vil derfor være en vis forskel mellem den observerede værdi y og den såkaldt *fittede* værdi $\widehat{y}$, dvs. den værdi som man ifølge regressionsmodellen skulle få med de givne værdier af baggrundsvariablene. Denne forskel kaldes *residualet* og betegnes ofte ϵ. Vi har så opspaltningen

$$\begin{array}{ccccc} y & = & \widehat{y} & + & \epsilon \\ \text{observeret værdi} & = & \text{fittet værdi} & + & \text{residual} . \end{array}$$

Residualerne er det som modellen *ikke* beskriver, og derfor er det naturligt at man (eller rettere modellen) anser dem for *tilfældige*, dvs. for at være tilfældige tal fra en vis sandsynlighedsfordeling.

To væsentlige forudsætninger for at kunne benytte regressionsanalyse er

1. at det ikke er x-erne, men kun y-erne og residualerne, der er behæftede med *tilfældig variation* ("usikkerhed"),

2. at de enkelte målinger er *stokastisk uafhængige* af hinanden, hvilket vil sige at de tilfældigheder der indvirker på én bestemt y-værdi (efter at man har taget højde for baggrundsvariablene) ikke har nogen sammenhæng med de tilfældigheder der spiller ind på de øvrige y-værdier.

De simpleste eksempler på regressionsanalyse er dem hvor der kun er én enkelt baggrundsvariabel, som vi så kan betegne x. Opgaven bliver da at beskrive y-værdierne ved hjælp af en kendt simpel funktion af x. Det simpleste ikke-trivielle bud på en sådan funktion må vel være en funktion af typen $x \mapsto \beta_0 + x\beta_1$ hvor β_0 og β_1 er to parametre, dvs. man formoder at y afhænger lineært af x. Derved får man den såkaldte simple lineære regressionsmodel.

De følgende kapitler beskæftiger sig med forskellige væsentlige aspekter af regressionsmodeller og regressionsanalyse: Hvordan vælger man værdierne af β-erne så man får det bedste fit? Hvordan afgør man om en bestemt model er god *nok*? Hvis man har flere forskellige baggrundsvariable til sin rådighed, hvordan afgør man så hvilke af dem der skal med i modellen og hvilke ikke?

Kapitel 2

Den bedste rette linie

I dette og de nærmest følgende kapitler vil vi beskæftige os med såkaldt *simpel lineær regressionsanalyse*, hvor der blot er én baggrundsvariabel x (plus konstanten 1), og hvor opspaltningen af y som en sum af en fittet værdi og et residual derfor bliver af formen

$$y = \beta_0 + x\beta_1 + \varepsilon.$$

Her betegner ε det teoretiske residual, og β_0 og β_1 er ukendte parametre hvis værdier skal bestemmes således at man får den bedste tilpasning.

Mere præcist vil vi antage at det givne talmateriale består af n talpar (x, y), ét for hvert "individ", hvilket skematisk kan skrives

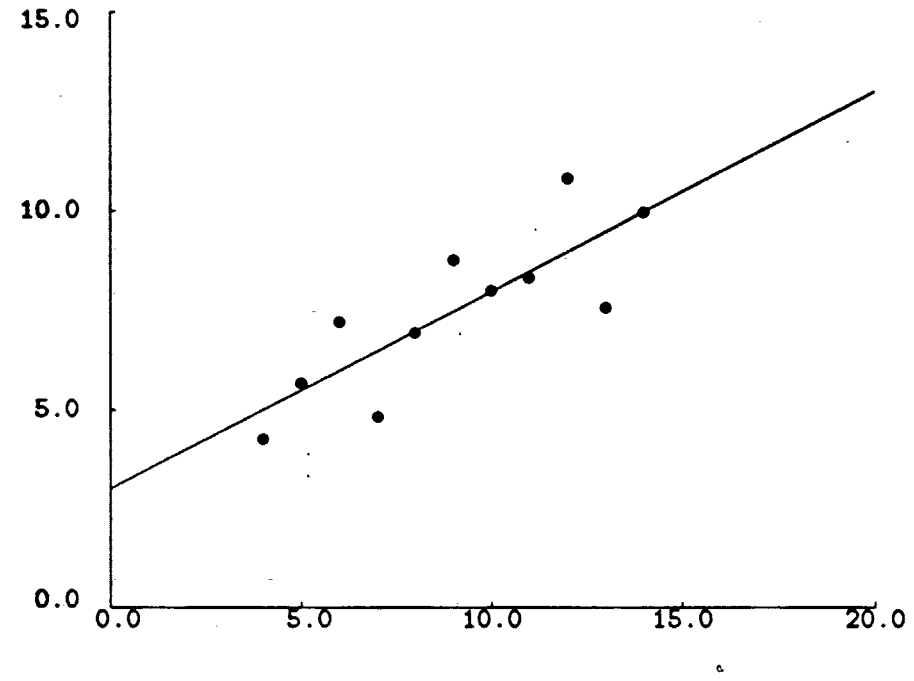
"individ"	observation	baggrundsvariabel
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Regressionsmodellen går da ud på, at for passende valg af parametrene β_0 og β_1 er

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

hvor $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er de teoretiske residualer. Ifølge denne model skal datapunkterne (x, y) ligge tilfældigt omkring den rette linie med ligning $y = \beta_0 + x\beta_1$ (denne linie kaldes *regressionslinien*), se Figur 2.1.

Symbolerne β_0 og β_1 i ligningerne (2.1) betegner de teoretisk rigtige værdier — som vi ikke kender. Vi står derfor nu over for den opgave



Figur 2.1: Et "typisk" eksempel på et sæt punkter (x, y) med den tilhørende estimerede regressionslinie.

på grundlag af observationerne at bestemme de værdier $\widehat{\beta}_0$ og $\widehat{\beta}_1$ af parametrene β_0 og β_1 der giver den linie der passer bedst muligt til datapunkterne (x_i, y_i). Man taler i denne forbindelse om at *estimere* parametrene β_0 og β_1, og man kalder $\widehat{\beta}_0$ og $\widehat{\beta}_1$ for *de estimerede parameterværdier* eller *estimaterne*. Bemærk at hvor parametrene β_0 og β_1 er nogle faste (og ukendte) teoretiske størrelser, så er estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ nogle størrelser der afhænger af observationerne.

Mindste kvadrater

Regressionslinien passer desto bedre til observationerne jo mindre residualerne er. Typisk kan man dog ikke opnå at alle residualerne er små på én gang. Derfor er man nødt til at formulere et kriterium for hvornår en linie er bedst mulig. Et meget ofte benyttet sådant kriterium er *mindste kvadraters kriteriet*, der siger at linien passer bedst muligt når *residualkvadratsummen* $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + x_i\widehat{\beta}_1)\right)^2$ er mindst mulig. Opgaven er derfor at bestemme det eller de talpar $(\widehat{\beta}_0, \widehat{\beta}_1)$ der

minimaliserer funktionen

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + x_i\beta_1)\Big)^2 . \tag{2.2}$$

Hvis talparret $(\widehat{\beta}_0, \widehat{\beta}_1)$ minimaliserer denne kvadratsum, så kalder man $\widehat{\beta}_0$ og $\widehat{\beta}_1$ for et sæt *mindste kvadraters estimater* for β_0 og β_1.

Estimationsligningerne

Man kan imidlertid også angribe estimationsproblemet på en anden måde:

Man kan et øjeblik betragte opgaven som gående ud på at bestemme et sæt fittede værdier $\widehat{y}_i = \widehat{\beta}_0 + x_i\widehat{\beta}_1$, $i = 1, 2, \ldots, n$, på en sådan måde at de "ligner" de observerede y_i-er mest muligt. Det er stærkt begrænset hvor meget lighed vi kan forlange, men vi kunne jo forsøgsvis kræve at summen af de fittede værdier skal være lig summen af de observerede værdier,

$$\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i . \tag{2.3}$$

og desuden at summen af produkterne af den forklarende variabel og den fittede værdi skal være lig med summen af produkterne af den forklarende variabel og den observerede værdi, altså

$$\sum_{i=1}^{n} x_i\widehat{y}_i = \sum_{i=1}^{n} x_i y_i . \tag{2.4}$$

Selv om det ikke uden videre springer i øjnene, så er ligningerne (2.3) og (2.4) faktisk to ligninger med de to ubekendte $\widehat{\beta}_0$ og $\widehat{\beta}_1$. For at indse at det forholder sig sådan skal man indsætte $\widehat{y}_i = \widehat{\beta}_0 + x_i\widehat{\beta}_1$ i ligningerne; derved får man

$$\sum_{i=1}^{n}(\widehat{\beta}_0 + x_i\widehat{\beta}_1) = \sum_{i=1}^{n} y_i ,$$

$$\sum_{i=1}^{n} x_i(\widehat{\beta}_0 + x_i\widehat{\beta}_1) = \sum_{i=1}^{n} x_i y_i ,$$

Box 2.1: Lidt om gennemsnit

I statistikken benytter man ofte betegnelsen $\overline{x}$ (udtales *x streg*) for gennemsnittet af talsættet $x_1, x_2, \ldots, x_n$, dvs.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

der igen kan skrives

$$n\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}_1 = \sum_{i=1}^{n} y_i\,, \tag{2.5}$$

$$\left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right)\widehat{\beta}_1 = \sum_{i=1}^{n} x_i y_i\,. \tag{2.6}$$

Bemærk at ligningerne (2.3) og (2.4) faktisk er ensbetydende med ligningerne (2.5) og (2.6). Ligningerne kaldes for *estimationsligningerne*, og de er interessante af den grund, at et talsæt $(\widehat{\beta}_0, \widehat{\beta}_1)$ er løsning til disse ligninger, hvis og kun hvis det er et minimumspunkt for kvadratsummen (2.2). Denne påstand bør naturligvis *bevises*, hvilket vi gør i Afsnit 2.1.

Løsning af estimationsligningerne

Vi vil nu vise at estimationsligningerne (2.5) og (2.6) altid kan løses, og vi vil bestemme et udtryk for denne løsning, dvs. et udtryk for estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$. Ved løsningen af ligningerne må man dele op i to tilfælde:

1. Hvis ikke alle x_i-erne er ens, så har ligningerne (2.5) og (2.6) præcis én løsning, hvilket indses således: Først divideres ligningen (2.5) igennem med n og bliver til $\widehat{\beta}_0 + \overline{x}\widehat{\beta}_1 = \overline{y}$. Denne ligning løses med hensyn til $\widehat{\beta}_0$ og man får $\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1$. Dette udtryk indsættes i (2.6), som derefter kan omformes til

$$\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)\widehat{\beta}_1 = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Box 2.2: En omskrivning af en sum af produkter af afvigelser
Hvis $a_1, a_2, \ldots, a_n$ og $c_1, c_2, \ldots, c_n$ er to talsæt og $\overline{a}$ og $\overline{c}$ betegner gennemsnittet af hhv. a-erne og c-erne (jf. Box 2.1), så er

$$\sum_{i=1}^{n}(a_i - \overline{a})(c_i - \overline{c}) = \sum_{i=1}^{n} a_i c_i - \frac{1}{n}\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} c_i\right).$$

Hvis a-erne er lig c-erne, fås specielt

$$\sum_{i=1}^{n}(a_i - \overline{a})^2 = \sum_{i=1}^{n} a_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} a_i\right)^2.$$

Påstanden vises ved at gange venstresidens parenteser ud.

der også kan skrives (jf. Box 2.2)

$$\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\widehat{\beta}_1 = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Da ikke alle x_i-erne er ens, er koefficienten til $\widehat{\beta}_1$ forskellig fra nul og vi kan derfor løse ligningen med hensyn til $\widehat{\beta}_1$. Alt i alt ender vi med følgende udtryk for den entydige løsning til ligningerne (2.5) og (2.6):

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1.$$

2. Hvis alle x_i-erne er ens, med den fælles værdi x, så er de to ligninger (2.5) og (2.6) proportionale. Den første ligning er

$$n\widehat{\beta}_0 + nx\widehat{\beta}_1 = \sum_{i=1}^{n} y_i,$$

der kan omskrives til

$$\widehat{\beta}_0 + x\widehat{\beta}_1 = \overline{y}.$$

Denne ene ligning med to ubekendte har uendelig mange løsninger $(\widehat{\beta}_0, \widehat{\beta}_1)$, men hvis vi skal udpege en enkelt af dem kan vi jo tage $\widehat{\beta}_1 = 0$ og $\widehat{\beta}_0 = \overline{y}$.

•

Når man estimerer parametrene β_0 og β_1 ved at minimalisere funktionen (2.2), siger man at man benytter *mindste kvadraters metode*, og estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ kan man så kalde for mindste kvadraters estimaterne. Som omtalt i det foregående kan man bestemme disse estimater ved at løse estimationsligningerne (2.3) og (2.4) eller (2.5) og (2.6).

Den *estimerede regressionslinie* er den linie hvis ligning er

$$y = \widehat{\beta}_0 + x\widehat{\beta}_1.$$

Man taler om at man foretager *regression af y på x*.

Figur 2.1 viser et "typisk" eksempel på et sæt punkter (x, y) med den tilhørende estimerede regressionslinie. Bemærk i øvrigt, at når man indsætter $x = \overline{x}$ i den estimerede regressionsligning så får man den tilsvarende y-værdi $\widehat{\beta}_0 + \overline{x}\widehat{\beta}_1 = (\overline{y} - \overline{x}\widehat{\beta}_1) + \overline{x}\widehat{\beta}_1 = \overline{y}$, dvs. regressionslinien må altid gå gennem "tyngdepunktet" $(\overline{x}, \overline{y})$.

2.1 Matematiske betragtninger

I dette afsnit gøres der rede for, at en løsning til estimationsligningerne faktisk også er et minimumspunkt for funktionen (2.2), og omvendt. Vi benytter en fremgangsmåde der let kan generaliseres til multipel lineær regressionsanalyse (Kapitel 8).

Det der skal vises formuleres som en sætning:

Sætning 2.1

Mindste kvadraters estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ kan bestemmes ved at løse estimationsligningerne (2.3) *og* (2.4) *eller* (2.5) *og* (2.6). *Der gælder:*

1. *Der findes altid en løsning til estimationsligningerne.*
2. *De fittede værdier $\widehat{y}_i = \widehat{\beta}_0 + x_i\widehat{\beta}_1$, $i = 1, 2, \ldots, n$ er entydigt bestemt (dvs. selv om der er flere løsninger til estimationsligningerne, så giver de de samme $\widehat{y}_i$-er).*

3. *Et talsæt $(\widehat{\beta}_0, \widehat{\beta}_1)$ er løsning til estimationsligningerne hvis og kun hvis det er et minimumspunkt for kvadratsummen (2.2).*

Bevis

Punkt 1 har vi allerede vist i det foregående. Vi kan derfor nu lade $(\widehat{\beta}_0, \widehat{\beta}_1)$ betegne en eller anden bestemt løsning og lade $\widehat{y}_i = \widehat{\beta}_0 + x_i\widehat{\beta}_1$.

Følgende opspaltning spiller en central rolle i argumentationen: For ethvert talpar (β_0, β_1) gælder at

$$\sum_{i=1}^{n} \Big(y_i - (\beta_0 + x_i\beta_1)\Big)^2 \tag{2.7}$$

$$= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} \Big(\widehat{y}_i - (\beta_0 + x_i\beta_1)\Big)^2 .$$

Dette vises ved at man først omskriver det i-te led i summen på venstre side ved brug af formlen for kvadratet på en toleddet størrelse:

$$\begin{aligned} \Big(y_i - (\beta_0 + x_i\beta_1)\Big)^2 &= \Big((y_i - \widehat{y}_i) + (\widehat{y}_i - (\beta_0 + x_i\beta_1)\Big)^2 \\ &= (y_i - \widehat{y}_i)^2 + \Big(\widehat{y}_i - (\beta_0 + x_i\beta_1)\Big)^2 \\ &\quad + 2(y_i - \widehat{y}_i)\Big(\widehat{y}_i - (\beta_0 + x_i\beta_1)\Big) . \end{aligned}$$

For at vise (2.7) er det nu nok at vise at summen af de dobbelte produkter er 0; men hvis man skriver det andet $\widehat{y}_i$ i det dobbelte produkt som $\widehat{\beta}_0 + x_i\widehat{\beta}_1$ så får man

$$\begin{aligned} &\sum_{i=1}^{n} 2(y_i - \widehat{y}_i)\Big(\widehat{y}_i - (\beta_0 + x_i\beta_1)\Big) \\ &= 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)\Big((\widehat{\beta}_0 + x_i\widehat{\beta}_1) - (\beta_0 + x_i\beta_1)\Big) \\ &= 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)\Big((\widehat{\beta}_0 - \beta_0) + x_i(\widehat{\beta}_1 - \beta_1)\Big) \\ &= 2(\widehat{\beta}_0 - \beta_0)\sum_{i=1}^{n}(y_i - \widehat{y}_i) + 2(\widehat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_iy_i - x_i\widehat{y}_i) \\ &= 0 \end{aligned}$$

ifølge estimationsligningerne (2.3) og (2.4). Dermed er opspaltningen (2.7) vist.

Det er nu klart at $(\widehat{\beta}_0, \widehat{\beta}_1)$ er et minimumspunkt for kvadratsummen g defineret i (2.2), for (2.7) viser jo at $g(\beta_0, \beta_1)$ kan skrives som en sum af $g(\widehat{\beta}_0, \widehat{\beta}_1)$ og det ikke-negative tal $\sum_{i=1}^{n} \left(\widehat{y}_i - (\beta_0 + x_i\beta_1)\right)^2$ der antager værdien 0 når $\beta_0 = \widehat{\beta}_0$ og $\beta_1 = \widehat{\beta}_1$.

Dermed er godtgjort at enhver løsning til estimationsligningerne også er et minimumspunkt for kvadratsummen. Det er endvidere klart, at hvis (β_0, β_1) er et minimumspunkt, så må det andet led på højre side af (2.7) være 0, og da dette led i sig selv er en sum af ikke-negative tal, kan det kun lade sig gøre hvis alle disse enkeltled er 0, dvs. hvis $\widehat{y}_i = \beta_0 + x_i\beta_1$ for alle i. Det viser at estimationsligningerne er opfyldt også for (β_0, β_1). □

2.2 Hvordan gør man

Regner man med håndkraft kan man bare indsætte i formlerne for $\widehat{\beta}_0$ og $\widehat{\beta}_1$; så får man det rigtige resultat, forudsat at man ikke laver afrundinger i mellemregningerne.

ISP's **regress**-kommando udregner uden videre mindste kvadraters estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$. Hvis man eksempelvis har anbragt sine x-værdier i en ISP-vektor **foder** og sine y-værdier i en ISP-vektor **udbytte**, skriver man blot `regress foder udbytte`. Dette afstedkommer en udskrift der i princippet ser ud som vist i Figur 2.2. Her kan man i **coef**-søjlen aflæse de to parameterestimater. De øvrige dele af **regress**-udskriften vil blive nærmere behandlet i kommende kapitler; der er en udførlig omtale af **regress** i *Introduktion til ISP*.

Hvadenten man benytter håndkraft eller computer bør man altid lave en *tegning* à la Figur 2.1, der på én gang viser den fittede regressionslinie og et *scatterplot* af y mod x (dvs. et plot af punkterne (x_i, y_i)). Derved får man mulighed for at *se*, om det virker rimeligt at beskrive datasættet med en ret linie.

Figur 2.2: Et eksempel på udskriften fra ISP's regress-kommando. Udskriften viser, at $\widehat{\beta}_1 = 0.6077$ og $\widehat{\beta}_0 = 1.415$.

```
ISP>>regress foder udbytte
degrees of freedom:     25 - 2 = 23
sigma      =   2.202
R-square   =    .8033
F-stat     =   93.92       (1 over 23 df)
condition  =   1.282

 var          coef          sdev
   1           .6077         .6271E-01
   const      1.415          .7100
ISP>>
```

2.3 Opgaver

En del af opgaverne til dette og de følgende kapitler indeholder mange tal. Disse tal kan indlæses med ISP's `input`-kommando og ved at man indtaster dem på tastaturet; det er dog en fordel at have tallene liggende i en fil på sin diskette eller harddisk og så indlæse dem derfra (der står hvordan i *Introduktion til ISP*).

Når man benytter den særlige RUC-udgave af ISP, har man adgang til datafiler hvor talmaterialerne til opgaverne på forhånd er indlæst. Man skal da blot benytte ISP-kommandoen `getdata` og vælge det ønskede datamateriale, så indlæses dataene;[1] i de fleste tilfælde oprettes også en tekst-makro `info` som indeholder en kort information om de indlæste data (man skriver `print info`).

Opgave 2.1

Antag at $y_1, y_2, \ldots, y_n$ er nogle kendte tal, der antages at fordele sig tilfældigt omkring et vist niveau μ, som vi ikke kender præcist. Vi ønsker at estimere μ ved mindste kvadraters metode, dvs. ønsker at finde et $\widehat{\mu}$ sådan at kvadratsummen

$$g(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2 \tag{2.8}$$

er mindst mulig.

1. En almindelig metode til at finde minimumspunkter for en funktion g er at søge dem blandt nulpunkterne for den afledede funktion g'. Find $\widehat{\mu}$ ved denne metode.

2. Man kan imidlertid også finde det $\widehat{\mu}$ uden at differentiere. Gør rede for at $(y_i - \mu)^2 = (y_i - \overline{y})^2 + (\overline{y} - \mu)^2 + 2(y_i - \overline{y})(\overline{y} - \mu)$, og omskriv derved (2.8) til

$$g(\mu) = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - \mu)^2$$

og slut heraf at $\widehat{\mu} = \overline{y}$.

[1] Man kan også gøre det hele på én linie ved at skrive f.eks. `getdata'demo'`

Opgave 2.2

Mindste kvadraters metode går ud på at finde de estimater der minimaliserer summen af *kvadratiske* afvigelser. En anden estimationsmetode er *mindste absolutte afvigelsers metode*, der går ud på at finde estimater der minimaliserer summen af de *absolutte* (dvs. numeriske) afvigelser.

Antag at der foreligger et meget lille datamateriale bestående af de tre observationer $y_1 = 1.2$, $y_2 = 1.7$ og $y_3 = 1.3$, som tænkes at fordele sig tilfældigt omkring et fælles niveau μ. Fra Opgave 2.1 ved vi at mindste kvadraters estimatet over μ er gennemsnittet $\overline{y}$. Find nu et estimat som minimaliserer summen $\sum_{i=1}^{3} |y_i - \mu|$ af absolutte afvigelser.

Vejledning: Skitsér graferne for disse tre funktioner:

$$\begin{aligned} \mu &\mapsto |1.2 - \mu|, \\ \mu &\mapsto |1.2 - \mu| + |1.7 - \mu|, \\ \mu &\mapsto |1.2 - \mu| + |1.7 - \mu| + |1.3 - \mu|. \end{aligned}$$

Af den sidste graf kan svaret let aflæses.

Hvad bliver svaret, hvis der også er en fjerde observation $y_4 = 1.8$?

Opgave 2.3: Forbes' barometriske målinger

Som bekendt aftager lufttrykket med højden over havets overflade, og derfor kan et barometer benyttes som højdemåler. Imidlertid kan man også bestemme højden ved at koge vand, fordi vands kogepunkt aftager med lufttrykket. I 1840erne og 1850erne foretog den skotske fysiker James D. Forbes på 17 forskellige lokaliteter i Alperne og i Skotland en række målinger hvor han bestemte dels vands kogepunkt, dels luftens tryk (omregnet til lufttrykket ved en standardlufttemperatur). Resultaterne er vist i Tabel 2.1.

Med ISP-kommandoen **getdata** (data-navn **forbes**) indlæses tallene til et 17 × 2-array **data** hvis første søjle indeholder kogepunkterne og anden søjle lufttrykkene. Fra dette array kan man eventuelt udtage to vektorer **kp** og **tr** indeholdende henholdsvis kogepunkter og lufttryk ved at skrive de to kommandolinier

```
ISP>>kp = data(*,1)
ISP>>tr = data(*,2)
```

Tabel 2.1: Opgave 2.3, Forbes' barometriske målinger.
Kogepunktet er angivet i °F, lufttrykket i 'inches Kviksølv'.

Kogepunkt	Lufttryk
194.5	20.79
194.3	20.79
197.9	22.40
198.4	22.67
199.4	23.15
199.9	23.35
200.9	23.89
201.1	23.99
201.4	24.02
201.3	24.01
203.6	25.14
204.6	26.57
209.5	28.49
208.6	27.76
210.7	29.04
211.9	29.88
212.2	30.06

1. Lufttrykket er angivet i 'inches Hg'. Nutildags måler man lufttryk i hPa (hektopascal = millibar). Omregn lufttrykkene til hPa. [2]

2. Kogepunkterne er angivet i °F. Omregn dem til °C. [3]

3. Meningen med eksperimentet er at undersøge *om* og *hvordan* man kan forudsige lufttrykket (og dermed højden over havet) på grundlag af en bestemmelse af vands kogepunkt. Lav et *scatterplot* for at se *om* det skulle være muligt (benyt `gscat`-kommandoen med kogepunkt som x og lufttryk som y).

4. Bestem den rette linie der fitter punkterne bedst.

5. Lav en tegning med både de observerede punkter og den estimerede linie.

   Hvordan passer linien til punkterne?

[2] 1 inch = 2.54 cm og 760 mm Hg = 1013.250 hPa.
[3] 0 °C svarer til 32 °F og 100 °C til 212 °F.

6. Fysikerne kan fortælle os, at det næppe er lufttrykket selv der afhænger lineært af kogepunktet, men snarere logaritmen til lufttrykket.[4]

   Derfor kan man forsøge sig med *logaritmen* til lufttrykkene i stedet for. Bliver det bedre af det?

Hvis man skal have nogen praktisk fornøjelse af sådanne kogepunktsbestemmelser er man nødt til at kende sammenhængen mellem højden og lufttrykket. Sålænge vi holder os til bjerghøjder aftager lufttrykket eksponentielt med højden, og der gælder at hvis lufttrykket ved havets overflade er p_0 (f.eks. 1013.25 hPa) og lufttrykket i højden h er p_h så er

$$h \approx 8150 \text{ m} \cdot \ln \frac{p_0}{p_h} .$$

Opgave 2.4: Anscombe's data

Den amerikanske statistiker Anscombe har konstrueret fire små datasæt der alle giver stort set samme numeriske resultater når man foretager regressionsanalyse på dem, specielt giver de den samme estimerede regressionslinie. Imidlertid vil *tegninger* afsløre markante forskelle og vise at det ikke er alle datasættene der beskrives lige godt ved hjælp af den estimerede linie.

Med ISP-kommandoen `getdata` (data-navn `anscombe`) indlæses disse data i otte vektorer `x1`, `y1`, `x2`, `y2` osv.

1. Benyt `regress`-kommandoen til at udføre de fire forskellige regressionsanalyser af et y på et x (altså f.eks. `regress x3 y3`).

   Udskrifterne vil være (stort set) identiske. Den estimerede regressionslinie bliver hver gang $y = \frac{1}{2}x + 3$.

2. For at se hvordan punkterne egentlig er beliggende skal man nu lave fire tegninger, en for hvert datasæt. Den estimerede regressionslinie skal tegnes ind på hver tegning.

   Tip: Det kan anbefales at tegne linien i intervallet $0 \leq x \leq 20$.

[4] Der er med god tilnærmelse en lineær sammenhæng mellem logaritmen til trykket og den reciprokke af den absolutte temperatur T. For de absolutte temperaturer som vi her har med at gøre er T^{-1} imidlertid stort set en lineær funktion af T.

Kapitel 3

Udvidelse af modellen

I Kapitel 2 fandt vi den bedste rette linie (ifølge mindste kvadraters metode) til et sæt datapunkter (x_i, y_i), $i = 1, 2, \ldots, n$. Selv om linien er den bedste, er det ingen garanti for at den også er *god*. Derfor bør man altid bestræbe sig på at undersøge rimeligheden af antagelsen om, at linien på fornuftig vis beskriver punkterne. I den forbindelse er det altid en god idé at lave en *tegning* hvor man indtegner både datapunkterne og den fittede linie. Ideelt skal punkterne fordele sig "tilfældigt" omkring linien — men hvad hvis de ikke gør det?

Det kan forekomme, at der er et enkelt datapunkt der falder helt uden for det almindelige mønster. Et sådant datapunkt kalder man en *outlier*. En outlier *kan* skyldes fejlskrivning af et tal eller at et deleksperiment er mislykkedes el.lgn., og i så fald bør man rette fejlen (hvis muligt) eller helt udelade det pågældende datapunkt. Men hvis man ikke har nogen klar mening om hvorfor punktet skulle indtage en særstilling, så kan man ikke tillade sig at kassere det. I stedet kan man eventuelt benytte en anden estimationsmetode end mindste kvadraters metode, f.eks. mindste absolutte afvigelsers metode, der går ud på at bestemme $\widehat{\beta}_0$ og $\widehat{\beta}_1$ således at $\sum_{i=1}^{n} |y_i - (\beta_0 + x_i\beta_1)|$ bliver mindst mulig. Denne metode er mindre følsom overfor ændringer af et enkelt datapunkt, man siger at metoden er mere *robust*.

3.1 Polynomiel regression

Det kan naturligvis også forekomme, at datapunkterne bare ganske enkelt ikke fordeler sig om en *ret linie* men om en anden slags kurve — og så må man jo fitte en kurve af denne anden slags.

Hvis for eksempel punkterne synes at ligge omkring en parabel, så kunne man prøve at beskrive y_i som en andengradsfunktion af x_i plus en tilfældig afvigelse:

$$y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

hvor der nu er de tre parametre β_0, β_1 og β_2 der skal estimeres. Denne *kvadratiske regressionsmodel* model er et eksempel på en såkaldt *polynomiel regressionsmodel*; polynomielle regressionsmodeller er, deres navn til trods, eksempler på (multiple) lineære regressionsmodeller, der omtales nærmere i Kapitel 8.

I modellen (3.1) kan man estimere parametrene β_0, β_1 og β_2 ved mindste kvadraters metode, nemlig ved at minimalisere kvadratsummen

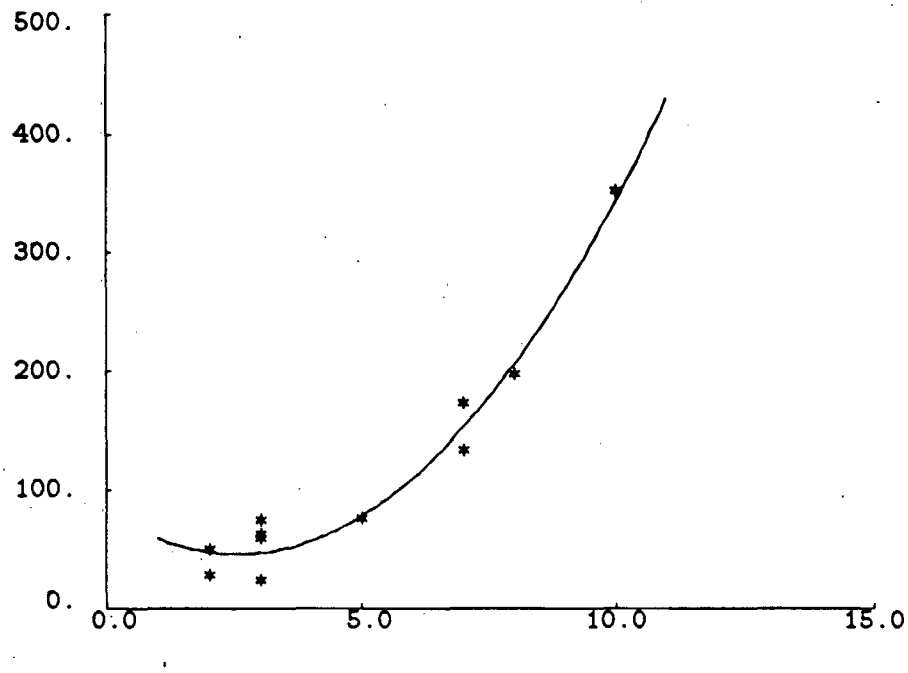
$$\sum_{i=1}^{n} \Big(y_i - (\beta_0 + x_i\beta_1 + x_i^2\beta_2)\Big)^2 . \tag{3.2}$$

Ganske som ved den simple lineære regressionsmodel i Kapitel 2 kan man finde estimaterne $\widehat{\beta}_0$, $\widehat{\beta}_1$ og $\widehat{\beta}_2$ ved at løse et sæt lineære ligninger, de såkaldte estimationsligninger. Denne gang er der tre ligninger (og tre ubekendte):

$$\begin{aligned}
\sum_{i=1}^{n} \widehat{y}_i &= \sum_{i=1}^{n} y_i , \\
\sum_{i=1}^{n} x_i\widehat{y}_i &= \sum_{i=1}^{n} x_i y_i , \\
\sum_{i=1}^{n} x_i^2\widehat{y}_i &= \sum_{i=1}^{n} x_i^2 y_i ,
\end{aligned}$$

hvor $\widehat{y}_i$ som sædvanlig betyder "den i-te fittede y-værdi", hvilket i dette tilfælde vil sige

$$\widehat{y}_i = \widehat{\beta}_0 + x_i\widehat{\beta}_1 + x_i^2\widehat{\beta}_2 .$$



Figur 3.1: Eksempel på polynomiel regression. — Den fittede kurve er $y = 82.5 - 27.9x + 5.4x^2$.

Estimationsligningerne kan derfor også skrives

$$\begin{aligned}
n\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}_1 + \left(\sum_{i=1}^{n} x_i^2\right)\widehat{\beta}_2 &= \sum_{i=1}^{n} y_i, \\
\left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right)\widehat{\beta}_1 + \left(\sum_{i=1}^{n} x_i^3\right)\widehat{\beta}_2 &= \sum_{i=1}^{n} x_i y_i, \\
\left(\sum_{i=1}^{n} x_i^2\right)\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^3\right)\widehat{\beta}_1 + \left(\sum_{i=1}^{n} x_i^4\right)\widehat{\beta}_2 &= \sum_{i=1}^{n} x_i^2 y_i.
\end{aligned}$$

Disse ligninger altid har en løsning, og mængden af løsningspunkter er lig med mængden af minimumspunkter for kvadratsummen (3.2). Vi vil ikke bevise dette, men henviser til den generelle diskussion i Kapitel 8.

Der gælder endvidere at estimationsligningerne har en entydigt bestemt løsning hvis og kun hvis der er mindst tre forskellige værdier blandt tallene $x_1, x_2, \ldots, x_n$ (det hænger sammen med at der skal tre punkter til at fastlægge en parabel).

•

Der gælder noget ganske tilsvarende for polynomier af grad $3, 4, 5, \ldots$

Hvis der er n datapunkter, kan man med et polynomium af grad $n-1$ få et perfekt fit, forstået på den måde at polynomiet går igennem alle punkterne, dvs. at alle residualerne er nul. Formentlig ville polynomiet svinge vildt og voldsomt ind imellem punkterne, så det ville ikke være nogen "pæn" kurve. Man ville heller ikke have vundet noget i retning af en simplere beskrivelse af data, eftersom man ville skulle bruge n polynomiumskoefficienter for at beskrive n y-værdier, og faktisk er et af statistikkens formål *datareduktion*!

3.2 Trigonometrisk regression

Det kan forekomme, at den forklarende variabel er et *tidspunkt* t og at y i store træk er en periodisk funktion af tiden med den kendte frekvens f. Så kunne man forsøge sig med en model af formen

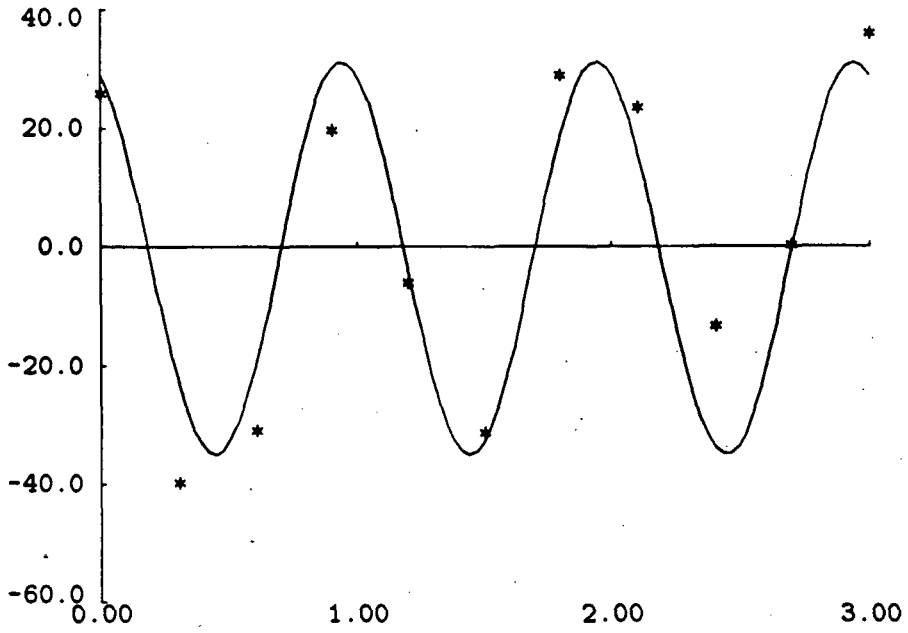
$$y_i = \beta_0 + \cos(2\pi f t_i)\beta_1 + \sin(2\pi f t_i)\beta_2 + \varepsilon_i \qquad (3.3)$$

hvor β_0, β_1 og β_2 er de ukendte parametre der skal estimeres. Dette er et eksempel på en *trigonometrisk regressionsmodel*, og den er, sit navn til trods, også en variant af multipel lineær regression. Der gælder derfor nogenlunde de samme bemærkninger som ved den polynomielle regression.

Med betegnelsen $\widehat{y}_i = \widehat{\beta}_0 + \cos(2\pi f t_i)\widehat{\beta}_1 + \sin(2\pi f t_i)\widehat{\beta}_2$ kan estimationsligningerne hørende til (3.3) skrives på formen

$$\begin{aligned}
\sum_{i=1}^{n} \widehat{y}_i &= \sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} \cos(2\pi f t_i)\, \widehat{y}_i &= \sum_{i=1}^{n} \cos(2\pi f t_i)\, y_i \\
\sum_{i=1}^{n} \sin(2\pi f t_i)\, \widehat{y}_i &= \sum_{i=1}^{n} \sin(2\pi f t_i)\, y_i
\end{aligned}$$

Estimationsligningerne kan omformes så de ubekendte optræder direkte. For at bevare overblikket indføres skrivemåden SP_{ab} for "Sum af Produkter af a og b", hvor a og b står for et af symbolerne $c, s, 1, y$



Figur 3.2: Eksempel på trigonometrisk regression. — Den fittede kurve er $y = -2.0 + 30.8\cos(2\pi t) - 11.7\sin(2\pi t)$.

svarende til hhv. $\cos(2\pi ft)$, $\sin(2\pi ft)$, konstanten 1 og y. Det betyder at f.eks.

$$SP_{cy} = \sum_{i=1}^{n} \cos(2\pi ft_i)\, y_i\,,$$

$$SP_{ss} = \sum_{i=1}^{n} \sin^2(2\pi ft_i)\,,$$

$$SP_{1c} = \sum_{i=1}^{n} \cos(2\pi ft_i)\,.$$

Med disse betegnelser kan estimationsligningerne skrives således:

$$\begin{aligned} n\widehat{\beta}_0 + SP_{1c}\widehat{\beta}_1 + SP_{1s}\widehat{\beta}_2 &= SP_{1y} \\ SP_{c1}\widehat{\beta}_0 + SP_{cc}\widehat{\beta}_1 + SP_{cs}\widehat{\beta}_2 &= SP_{cy} \\ SP_{s1}\widehat{\beta}_0 + SP_{sc}\widehat{\beta}_1 + SP_{ss}\widehat{\beta}_2 &= SP_{sy}. \end{aligned}$$

3.3 Frihedsgrader

Regressionsanalyse går — i lighed med mange andre statistiske metoder — i en vis forstand ud på at splitte den "information" der er indeholdt i observationerne $y_1, y_2, \ldots, y_n$ op i en *systematisk del* (her i form af en funktionel afhængighed af en forklarende variabel x) repræsenteret ved parameterestimaterne $\widehat{\beta}_0, \widehat{\beta}_1, \ldots$ og en *tilfældig del* repræsenteret ved residualerne $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Jo mere "information" den statistiske metode placerer i estimaterne, jo mindre bliver der tilbage til residualerne.

I den simple lineære regressionsmodel er opsplitningen af y i en sum af noget systematisk og en rest

$$y_i = \underbrace{\widehat{\beta}_0 + x_i\widehat{\beta}_1}_{\widehat{y}_i} + \underbrace{y_i - (\widehat{\beta}_0 + x_i\widehat{\beta}_1)}_{e_i},$$

hvor $e_1, e_2, \ldots, e_n$ er de *empiriske* residualer. Man kunne sige at $y_1, y_2, \ldots, y_n$ indeholder n "stykker information", forstået på den måde at de n y-er hver især kan variere frit uafhængigt af de andre — man taler om at der er n *frihedsgrader*. Disse n frihedsgrader bliver delt op i to frihedsgrader til de to estimater plus $n-2$ til de empiriske residualer. Det sker ud fra et ræsonnement om, at hvis man lægger sig fast på at estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ skal have nogle bestemte værdier, så kan y-erne ikke længere variere helt frit, idet det nemlig da vil være sådan, at når man har valgt $n-2$ af dem, så er de sidste to automatisk givet ud fra kravet om at estimaterne skal have bestemte værdier. (Se også Opgave 3.4.)

Den generelle regel er, at forskellen mellem antallet af observationer og antallet af estimerede parametre i regressionsmodellen bliver *antallet af frihedsgrader*. I den simple lineære regressionsmodel er der eksempelvis $n-2$ frihedsgrader, og i den kvadratiske regressionsmodel (3.1) er der $n-3$ frihedsgrader.

3.4 Hvordan gør man med ISP

Polynomiel regression

Det er ganske let at fitte polynomier med ISP. Her forklares som eksempel hvordan man fitter en andengradsfunktion, dvs. en model af formen

(3.1). Antag at **x** og **y** er ISP-vektorer af længde n indeholdende hhv. x-værdierne og y-værdierne. Man går så frem i to skridt:

1. Der skal oprettes et ISP-array med to søjler og n rækker. Den første søjle skal indeholde x-værdierne (den skal være lig **x**), den anden søjle skal indeholde x^2-værdierne. Et sådant array kan f.eks. oprettes med ISP-kommandolinien

   ```
   ISP>>glue/axis=2  x (x**2) > x2
   ```

   Her får det nye array navnet **x2**.

2. Dernæst benyttes **regress** med det nye array som første argument og **y** som andet:

   ```
   ISP>>regress x2 y
   ```

   I **regress**-udskriften kommer der tre **coef**-værdier. Den sidste hører til konstantleddet og de to andre hører til hhv. den første og den anden af søjlerne i **x2**-arrayet (dvs. til x og x^2 i nærværende eksempel).

Trigonometrisk regression

Det er ganske let at fitte trigonometriske funktioner med ISP. Her forklares som eksempel hvordan man fitter en model af formen (3.3). Antag at **t** og **y** er ISP-vektorer af længde n indeholdende hhv. tiderne og y-værdierne. Man går så frem i to skridt:

1. Der skal oprettes et ISP-array med to søjler og n rækker. Den første søjle skal indeholde $\cos(2\pi ft)$-værdierne, den anden søjle $\sin(2\pi ft)$-værdierne. Et sådant array kan f.eks. oprettes med ISP-kommandolinien

   ```
   ISP>>glue/axis=2 (cos(2*pi*f*t)) (sin(2*pi*f*t)) > x2
   ```

   Her får det nye array navnet **x2**.

2. Dernæst benyttes **regress** med det nye array som første argument og **y** som andet:

   ```
   ISP>>regress x2 y
   ```

I `regress`-udskriften kommer der tre `coef`-værdier. Den sidste hører til konstantleddet og de to andre hører til hhv. den første og den anden af søjlerne i `x2`-arrayet (dvs. til cosinusleddet hhv. sinusleddet i nærværende eksempel).

Outliers

Ofte kan man på en tegning klart se om et bestemt punkt falder uden for det almindelige mønster, dvs. om det er en *outlier*. Hvis der er mange punkter kan det være vanskeligt at identificere det pågældende punkt, og så kan man udnytte ISP's muligheder for interaktivt at sætte 'labels' på udpegede punkter i et plot.

3.5 Opgaver

Opgave 3.1: Pattedyrs legemsvægt og hjernevægt

Man kunne umiddelbart forestille sig at store dyr har en større hjerne end små dyr — eller måske er det de mere intelligente dyr der har de store hjerner? Tabel 3.1 opregner den gennemsnitlige legemsvægt og den gennemsnitlige hjernevægt for et antal pattedyr. Dyrene er ordnet efter legemsvægt.

Opgaven går ud på at undersøge hvordan hjernens vægt afhænger af legemsvægten.

Indlæs dataene med kommandoen **getdata** (data-navn **hjerne**); der oprettes de to datavektorer **hv** og **lv** indeholdende henholdsvis hjernevægt og legemsvægt, samt et array **dyr** som indeholder navnene[1]

1. Lav et scatterplot af hjernevægt mod legemsvægt.[2]

2. Hvordan ser det ud hvis man tager *logaritmen* til hjernevægt og til legemsvægt? Fit en ret linie og indtegn den.

   Tip: Det kan være spændende at kunne identificere de enkelte punkter. Klik på LABELS-feltet på grafik-skærmen; flyt cursoren (der er blevet til et kvadrat) hen på et datapunkt og klik på venstre museknap; derved får man dyrets *nummer* at se. Man kan få dyrets *navn* at se ved at angive makroen **dyr** som et tredie inputargument til **gscat** når man tegner punkterne. Hvis f.eks. **lnhv** og **lnlv** er logaritmen til henholdsvis **hv** og **lv**, så skal man kalde **gscat** sådan:

   ```
   ISP>>gscat lnlv lnhv dyr .
   ```

3. Nogle biologer mener at der kunne tænkes at gælde en relation af typen

$$\text{hjernevægt} = \text{konstant} \cdot \text{legemsvægt}^{2/3}. \qquad (3.4)$$

   Begrundelsen skulle være at *hjernens* størrelse og dermed vægt er proportional med dyrets overflade (der skal være nerveforbindelser ud til alle punkter på overfladen), hvorimod *legemets* vægt er proportional med dyrets rumfang. Da overflade er porportional med rumfang$^{2/3}$ når man alt i alt til (3.4).

[1] Dette array kan udskrives med **print /char dyr**.

[2] Dvs. med legemsvægt som x og hjernevægt som y.

Tabel 3.1: Legemsvægt og hjernevægt for 62 pattedyrearter.

art	legemsvægt (kg)	hjernevægt (g)
afrikansk elefant	6654.000	5712.00
asiatisk elefant	2547.000	4603.00
giraf	529.000	680.00
hest	521.000	655.00
ko	465.000	423.00
okapi	250.000	490.00
gorilla	207.000	406.00
svin	192.000	180.00
æsel	187.100	419.00
brasiliansk tapir	160.000	169.00
jaguar	100.000	157.00
gråsæl	85.000	325.00
menneske	62.000	1320.00
kæmpebæltedyr	60.000	81.00
får	55.500	175.00
chimpanse	52.160	440.00
gråulv	36.330	119.50
kænguro	35.000	56.00
ged	27.660	115.00
rådyr	14.830	98.20
bavian	10.550	179.50
husarabe	10.000	115.00
rhesusabe	6.800	179.00
vaskebjørn	4.288	39.20
rød ræv	4.235	50.40
grøn marekat	4.190	58.00
gulbuget murmeldyr	4.050	17.00
klippegrævling[3]	3.600	21.00
nibæltet bæltedyr	3.500	10.80
pungodder	3.500	3.90
polarræv	3.385	44.50
kat	3.300	25.60
myrepindsvin	3.000	25.00
kanin	2.500	12.10

(fortsættes)

(fortsat)

art	legemsvægt (kg)	hjernevægt (g)
trægrævling[4]	2.000	12.30
nordamerikansk opossum	1.700	6.30
kuskus	1.620	11.40
genette	1.410	17.50
plump-lori	1.400	12.50
bæveregern	1.350	8.10
marsvin	1.040	5.50
afrikansk kæmpepungrotte	1.000	6.60
arktisk jordegern[5]	0.920	5.70
børstesvin	0.900	2.60
pindsvin	0.785	3.50
klippegrævling[6]	0.750	12.30
ørkenpindsvin	0.550	2.40
natabe	0.480	15.50
chinchilla	0.425	6.40
rotte	0.280	1.90
galago	0.200	5.00
muldvarpegnaver	0.122	3.00
guldhamster	0.120	1.00
træspidsmus	0.104	2.50
egern	0.101	4.00
østamerikansk muldvarp	0.075	1.20
stjernemuldvarp	0.060	1.00
bisamrotte	0.048	0.33
stor brun flagermus	0.023	0.30
mus	0.023	0.40
lille brun flagermus	0.010	0.25
lille korthalet spidsmus	0.005	0.14

[3] *Procavia habessinica*
[4] *Dendrohyrax*
[5] *Citellus (Spermophilus) undulatus ablusus*
[6] *Heterohyrax brucci*

(a) Præcisér dette argument.

Tip: Hvis man havde et matematisk model-dyr som var kugleformet eller terningeformet, så kunne man let finde både dets overflade og dets rumfang.

Hvad med "rigtige" dyr?

(b) Hvis (3.4) gælder, hvilken sammenhæng er der da mellem logaritmen til hjernevægt og logaritmen til legemsvægt?

(c) Hvordan harmonerer formodningen (3.4) med de observerede data?

4. Hvordan finder man i almindelighed den bedste rette linie med en given hældning?

   Find i det konkrete eksemple den bedste rette linie (i log-log figuren) med hældning 2/3 og indtegn den.

Opgave 3.2

Betragt følgende (til lejligheden konstruerede) datasæt, der skal illudere den situation at man til bestemte velvalgte værdier af x har målt en tilhørende værdi af y:

y	x
9.8	-2
11.6	-1
14.9	0
16.2	1
17.5	2

1. Man ønsker at bestemme bedste rette linie $y = \beta_0 + x\beta_1$ til disse data. Opskriv de tilhørende estimationsligninger og løs dem (med håndkraft).

2. Man ønsker desuden den bedste parabel $y = \beta_0 + x\beta_1 + x^2\beta_2$ til disse data. Opskriv de tilhørende estimationsligninger og løs dem (med håndkraft).

3. Ville beregningerne være blevet mere eller mindre besværlige hvis x_5 havde været 3 i stedet for 2 ? Hvorfor?

4. Her er et talsæt der minder meget om det første:

y	x
7098	-2
7116	-1
7149	0
7162	1
7175	2

   Man har ganske enkelt ganget y-erne med 10 og lagt 7000 til. Hvordan kan man udnytte denne sammenhæng til let at bestemme den bedste rette linie og den bedste parabel til det nye datasæt, når man kender svarene for det første datasæt?

5. Her er endnu et talsæt der minder meget om det første:

y	x
9.8	38
11.6	44
14.9	50
16.2	56
17.5	62

   Denne gang har man ganget de oprindelige x-er med 6 og lagt 50 til. Hvordan kan man nu bestemme den bedste rette linie og den bedste parabel når man kender svarene for det første datasæt?

Hvad kan man lære heraf?

Opgave 3.3: Vands strømningsforhold i en flod

I forbindelse med en undersøgelse af vands strømningsforhold i en flod har man på et bestemt sted målt flowraten i forskellige dybder. Flowraten er den mængde vand der passerer et givet tværsnit af floden i et givet tidsrum (så den måles altså i f.eks. m^3 pr. m^2 pr. sekund). Tabel 3.2 viser sammenhørende værdier af vanddybde og flowrate.

Opgaven er at give en simpel beskrivelse af sammenhængen mellem flowrate og vanddybde.[7]

[7]Hydrologer kan sikkert opstille fornemme differentialligningsmodeller der beskriver denne sammenhæng, forudsat at flodens sider og bund ikke er alt for uregelmæssige. Det er slet ikke det vi er ude efter her. Statistikeren vil blot søge efter en simpel beskrivelse af de empiriske data.

Tabel 3.2: Opgave 3.3: Flowraten i forskellige vanddybder.

dybde	flowrate
0.34	0.636
0.29	0.319
0.28	0.734
0.42	1.327
0.29	0.487
0.41	0.924
0.76	7.350
0.73	5.890
0.46	1.979
0.40	1.124

1. Undersøg først talmaterialet *uden* at bruge ISP eller andre computerprogrammer:
   (a) Lav et scatterplot af flowrate mod dybde.
   (b) Ser punkterne ud til at ligge på en ret linie?
   Beregn den bedste rette linie og indtegn den (det er altid lettere at vurdere om punkter ligger omkring en bestemt kurve når man har kurve og punkter i samme tegning).
   (c) Man kunne forestille sig at en *andengradskurve* ville give en bedre beskrivelse af punkterne. Opstil og løs de estimationsligninger der bestemmer den bedste andengradskurve.
   (d) Indtegn den fittede andengradskurve.
2. Indlæs derefter tallene i ISP (med ISP's `input`-kommando) og foretag analysen ved hjælp af ISP:
   (a) Lav et scatterplot af flowrate mod dybde.
   (b) Beregn og indtegn den bedste rette linie.
   (c) Beregn den bedste andengradskurve og indtegn den på samme figur som punkterne og den rette linie.
3. Skal man foretrække andengradskurven fremfor den rette linie? Hvorfor?
4. Hvad er konklusionen mht. sammenhængen mellem flowrate og vanddybde?

Opgave 3.4

Denne opgave skal belyse forskellige ting om frihedsgrader, se også Afsnit 3.3.

Vi kan tage udgangspunkt i følgende lille datasæt:

y	x
9.8	1
11.6	2
14.9	3
16.2	4
17.5	5

Hvis man "regresser" y på x får man regressionsligningen $y = 8 + 2x$.

Betragt nu det "omvendte" problem: Givet de fem x-værdier og givet at $\widehat{\beta}_0$ skal være 8 og at $\widehat{\beta}_1$ skal være 2, hvordan kan man så vælge y_1, y_2, y_3, y_4, y_5 ?

1. I det foreliggende problem er de fittede værdier $\widehat{y}_i$ kendte tal. Udregn dem.

2. Estimationsligningerne (2.3) og (2.4) eller (2.5) og (2.6) skal være opfyldt — det er mest praktisk at arbejde med (2.3) og (2.4) (og de står på side 15). I denne omgang er det ikke $\widehat{\beta}_0$ og $\widehat{\beta}_1$ der er de ubekendte, men derimod y_1, y_2, y_3, y_4, y_5.

   Skriv ligningerne op så det fremgår at der er tale om to ligninger med fem ubekendte. Gør rede for at man frit kan vælge værdier for *tre* af y-erne og at de to sidste y-ers værdier så kan og skal bestemmes af de to ligninger. — Når estimaterne er givne er der $5 - 2 = 3$ frihedsgrader tilbage.

3. Man indfører det empiriske residual e_i som $e_i = y_i - \widehat{y}_i$. I denne opgave opfatter vi y_1, y_2, y_3, y_4, y_5 som ubekendte, så derfor bliver også e_1, e_2, e_3, e_4, e_5 at betragte som ubekendte.

   Indsæt $y_i = \widehat{y}_i + e_i$ i estimationsligningerne (2.3) og (2.4) (i den udgave de har i denne opgave). Derved kommer der to ligninger med de fem ubekendte e_1, e_2, e_3, e_4, e_5. Hvordan kommer de til at se ud? Sammenlign disse ligninger med ligningerne for y-erne.

   Hvor mange af residualerne kan man vælge frit? Hvor mange frihedsgrader har residualerne?

Tabel 3.3: Opgave 3.5: Antal æg pr. måned pr. høne i USA.

	1938	1939
januar	14.2	14.5
marts	17.0	16.8
maj	14.1	14.4
juli	8.9	9.0
september	6.0	5.8
november	8.5	8.8

Opgave 3.5: Høns' æglægning

Høns lægger ikke lige mange æg på alle tider af året (de lægger naturligvis flest omkring Påske!). I Tabel 3.3 er vist det gennemsnitlige antal æg lagt pr. høne i USA i de ulige måneder i årene 1938 og 1939.[8]

1. Tegn et scatterplot der viser antal lagte æg som funktion af tiden (man kan f.eks. indføre en tidsvariabel t der går fra 0 til 11, svarende til de 12 måneder).
2. Prøv at fitte en model af typen (3.3) hvor frekvensen f er 1/6.
3. Indtegn den fittede kurve i det oprindelige scatterplot.

[8] efter *Report of the Bureau of Agricultural Economics, U.S. Dept. of Agriculture on the Poultry and Egg Situation, March 1941.*

Kapitel 4

Normalfordelingen

Det er karakteristisk for en statistisk model at den beskriver både den såkaldte *systematiske variation* og den *tilfældige variation*. I de forbindelser der er omtalt i Kapitel 2 og 3, er "beskrivelsen af den systematiske variation" angivelsen af den kurve (ret linie eller parabel osv.) som datapunkterne fordeler sig om. Beskrivelsen af den tilfældige variation er der ikke hidtil gjort noget ved.

I regressionsmodeller er det tanken at alt det tilfældige skal være henlagt til residualerne, som skal kunne siges at være *tilfældige tal* fra en vis sandsynlighedsfordeling. Denne sandsynlighedsfordelings nærmere udseende er ikke uden interesse i forbindelse med forskellige vurderinger af hvor godt modellen passer og i forbindelse med test af statistiske hypoteser om modellens parametre. Da kan man nemlig være interesseret i at sammenligne sine faktiske observationer med de andre mulige observationer som man også kunne have fået, og sandsynlighedsfordelingens rolle i modellen er netop at beskrive hvilke andre værdier man også kunne have fået i stedet for de faktisk foreliggende.

Undertiden vil man søge at beskrive den tilfældige variation på den måde, at man om residualerne antager at de er stokastisk uafhængige observationer fra en og samme *normalfordeling*. I så fald bør visse betingelser være opfyldt:

1. De enkelte residualer ε skal tænkes at være tilfældige tal fra en og samme sandsynlighedsfordeling, udtrukket *stokastisk uafhængigt* af hverandre.

2. Middelværdien i residualernes fordeling skal være 0.

Box 4.1: Definition af normalfordelingen

Normalfordelingen er bestemt ved sin *sandsynlighedstæthedsfunktion* hvori der indgår to parametre: en *middelværdiparameter* (eller *positionsparameter*) der ofte betegnes μ, og en *variansparameter* der ofte betegnes σ^2. Sandsynlighedstæthedsfunktionen for normalfordelingen med parametre μ og σ^2 er

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

hvor x går fra $-\infty$ til $+\infty$.
Parameteren μ bestemmer hvor på talaksen fordelingen har sit centrum, parameteren σ^2 bestemmer hvor flad den er, se Figur 4.1.

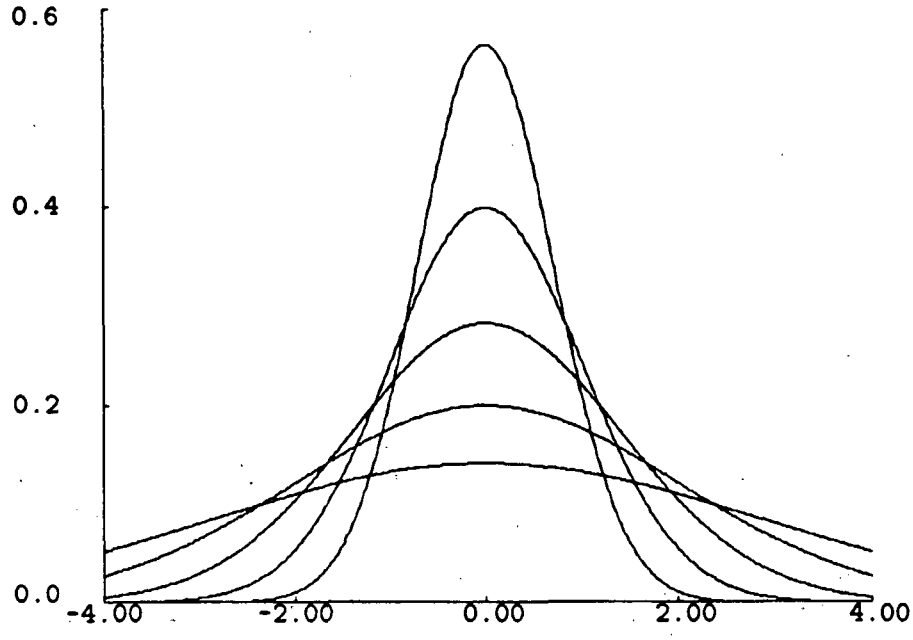
3. Residualernes fordeling skal være *symmetrisk omkring* 0.

4. De observationer y der er tale om skal være målinger på en *kontinuert skala*[1] hvis endepunkter ligger væk fra det område hvor observationerne forekommer.

Punkt 1 betyder blandt andet, at det forhold at ét y tilfældigvis har en bestemt afvigelse fra *sin* forventede værdi, ikke har noget at gøre med hvordan de øvrige y-er nu tilfældigvis måtte afvige fra deres. Meningen med Punkt 2 er, at den systematiske del af modellen (den del der specificerer "forventet værdi") netop skal tage højde for *alt* det systematiske.

Når man har et sæt virkelige observationer $y_1, y_2, \ldots, y_n$, plejer der jo ikke at stå på dem om de er normalfordelte, eller hvilke værdier parametrene i givet fald har. Man er derfor nødt til først at estimere de ukendte parametre og dernæst på en eller anden måde (eller gerne flere) vurdere rimeligheden af at mene at observationerne faktisk stammer fra en og samme normalfordeling. Dette sidste kan man blandt andet gøre med *histogrammer* og *fraktildiagrammer*.

[1] En kontinuert skala er en skala hvor i princippet enhver værdi i et vist interval kan forekomme.
Antalsobservationer er et eksempel på observationer der *ikke* er på en kontinuert skala.



Figur 4.1: Tæthedsfunktioner for normalfordelinger med middelværdi 0 og varians hhv. 0.5, 1, 2, 4 og 8.

Estimater over μ og σ^2

Hvis man har besluttet sig for at $y_1, y_2, \ldots, y_n$ er et sæt observationer fra en og samme normalfordeling, så er der stadig det problem at finde estimater for denne normalfordelings parametre. Der gælder:

- Estimatet over middelværdiparameteren μ er gennemsnittet

$$\overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

  Dette estimat er et mindste kvadraters estimat (Opgave 2.1).

- Estimatet over variansparameteren σ^2 er

$$s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2,$$

  dvs. den minimaliserede kvadratsum divideret med antallet af frihedsgrader. En grund til at dividere med antal frihedsgrader

$(n-1)$ og ikke med antal observationer (n) er, at man derved får et såkaldt *centralt* skøn over σ^2, dvs. et skøn som i middel giver den sande værdi (σ^2). [2]

Histogrammer

Et histogram over et sæt observationer $y_1, y_2, \ldots, y_n$ fås på den måde at man inddeler observationsaksen i et antal intervaller, gerne lige store, og så tegner rektangler hvis grundflader er disse intervaller og hvis arealer er lig med den brøkdel af observationerne som ligger inden for det pågældende interval. (Det samlede areal under histogrammet skal være 1.)

Histogrammet skal ligne tæthedsfunktionen for den formodede sandsynlighedsfordeling (her normalfordeling). Det er derfor en god idé at indtegne denne fordelings tæthedsfunktion i samme figur som histogrammet, se Figur 4.2

[2] Det matematiske ræsonnement ser sådan ud: Der gælder at

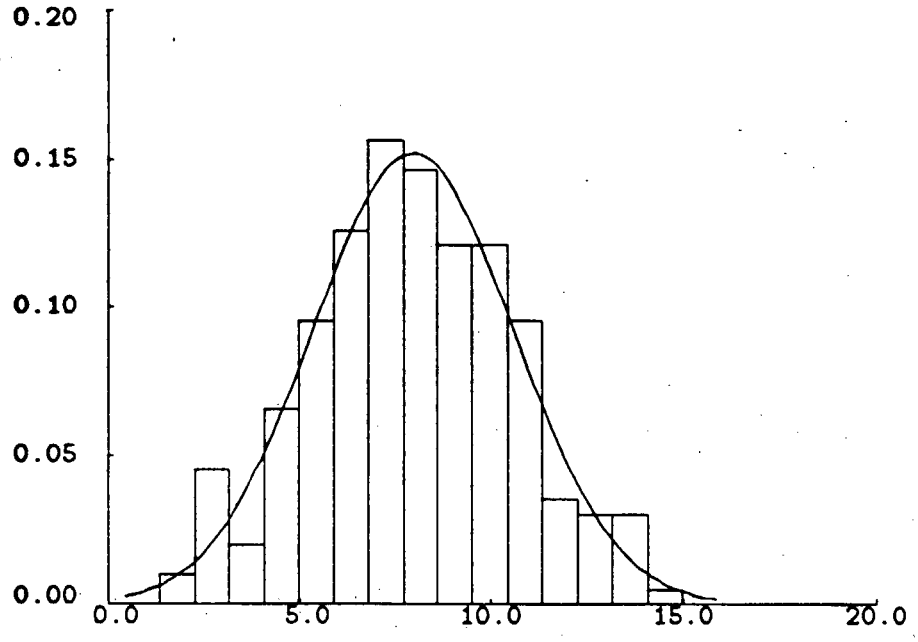
$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}\left((y_i-\mu)^2+2(y_i-\mu)(\mu-\overline{y})+(\mu-\overline{y})^2\right)\\ &= \sum_{i=1}^{n}(y_i-\mu)^2-n(\overline{y}-\mu)^2.\end{aligned}$$

Ved at tage middelværdi fås heraf

$$\begin{aligned}\mathrm{E}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}\mathrm{E}(y_i-\mu)^2-n\mathrm{E}(\overline{y}-\mu)^2\\ &= n\mathrm{Var}(y)-n\mathrm{Var}(\overline{y})\\ &= (n-1)\mathrm{Var}(y)\\ &= (n-1)\sigma^2,\end{aligned}$$

dvs.

$$\mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2\right) = \sigma^2.$$



Figur 4.2: Histogram over kvadratroden af højden (målt i 100 fod) af de 219 berømteste vulkaner.[3] Den indtegnede kurve er tæthedsfunktionen for den fittede normalfordeling.

Fraktildiagrammer

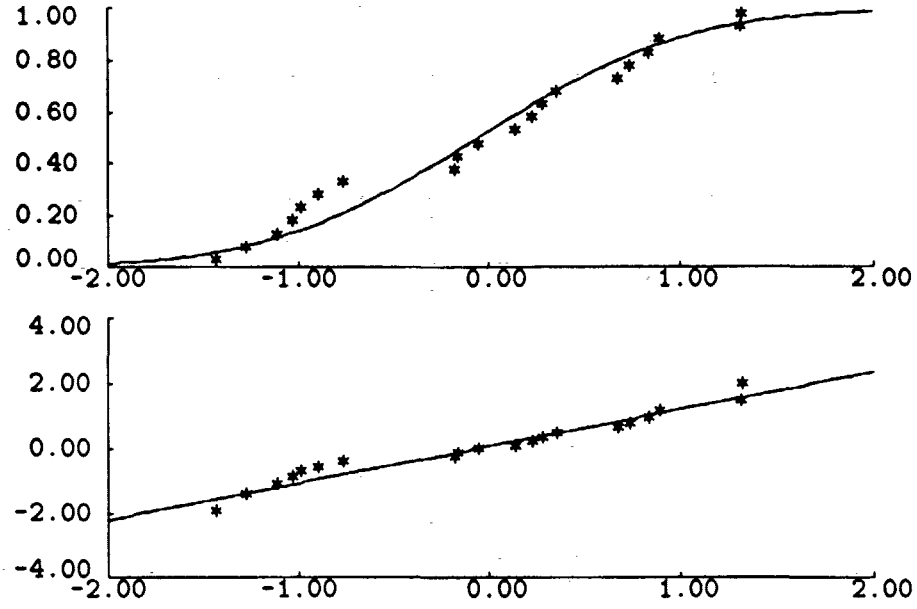
Når man har et sæt observationer $y_1, y_2, \ldots, y_n$, så plejer man at benytte betegnelsen $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ for de *ordnede observationer*, dvs. y-erne stillet op i rækkefølge (den mindste værdi først).

Nu er det sådan, at hvis alle de observerede y-er er forskellige, så er brøkdelen $(i-1)/n$ af observartionerne strengt mindre end $y_{(i)}$ og brøkdelen i/n af dem mindre end eller lig med $y_{(i)}$. Som et kompromis kan man så sige at brøkdelen $(i-0.5)/n$ af dem er mindre end $y_{(i)}$. Punkterne

$$\left(y_{(i)}, \frac{i-0.5}{n}\right), \quad i = 1, 2, \ldots, n,$$

kan derfor opfattes som punkter på den empiriske (kumulerede) fordelingsfunktion.

[3]Citeret efter P. Tukey (1977): *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.



Figur 4.3: Øverst: Punkter på en empirisk fordelingsfunktion, samt fordelingsfunktionen for den fittede normalfordeling. **Nederst:** Det tilsvarende fraktildiagram.
Abscisseaksen er i begge tilfælde "observationsaksen". Enhederne på ordinataksen er i den øverste delfigur sandsynlighed, i den nederste de såkaldte *probits* ('probability units'), dvs. antal standardafvigelser fra middeltallet.

Hvis observationerne $y_1, y_2, \ldots, y_n$ faktisk stammer fra en normalfordeling med middelværdi μ og varians σ^2, så skal den empiriske fordelingsfunktion ligne denne normalfordelings kumulerede fordelingsfunktion. (I praksis kender man sjældent de teoretiske parametre μ og σ^2, så man må erstatte dem med $\overline{y}$ og s^2.) Derfor kan man finde på at lave en tegning, der dels indeholder de n punkter på den empiriske fordelingsfunktion, dels fordelingsfunktionen for normalfordelingen med middelværdi $\overline{y}$ og varians s^2, se Figur 4.3, øverste halvdel. Imidlertid kan det være vanskeligt at vurdere hvordan nogle punkter fordeler sig omkring en *krum* kurve, så derfor gør man det at man transformerer ordinataksen på en sådan måde at den krumme kurve bliver til en ret linie, hvorved man får et såkaldt *fraktildiagram*, se Figur 4.3.

Den funktion man skal transformere ordinataksen med er funktionen Φ^{-1}, som er nærmere beskrevet i Box 4.2; funktionen er tabelleret i statistiske tabelværker og er en standardfunktion i statistikprogrammer

Box 4.2: Fraktiler i normalfordelingen. Φ^{-1}
Funktionen Φ^{-1} er den omvendte funktion til Φ, som er den kumulerede fordelingsfunktion for normalfordelingen med middelværdi 0 og varians 1, dvs. $\Phi(y) = \int_{-\infty}^{y} f_{0,1}(u)\,du$. Man har ikke noget eksplicit udtryk for Φ^{-1}. Funktionen Φ^{-1} er defineret for alle tal mellem 0 og 1. Tallet $y_p = \Phi^{-1}(p)$ er den såkaldte p-fraktil i normalfordelingen med middelværdi 0 og varians 1, dvs. y_p er den grænse til venstre for hvilken brøkdelen (fraktionen) p af sandsynlighedsmassen ligger.

til computere.

Punkterne i fraktildiagrammet er altså

$$\left(y_{(i)}, \Phi^{-1}(\tfrac{i-0.5}{n})\right), \quad i = 1, 2, \ldots, n.$$

Den rette linie som disse punkter skal ligge omkring, forudsat at y-erne er observationer fra en normalfordeling med middelværdi μ og varians σ^2, er den linie som går gennem $(\mu, 0)$ og som har hældningskoefficient $1/\sigma$ (dvs. den går gennem $(\mu - 2\sigma, -2)$ og $(\mu + 2\sigma, 2)$.

I praksis tegnes fraktildiagrammer enten ved hjælp af en computer, eller på *sandsynlighedspapir*, eller på millimeterpapir idet man benytter en tabel over funktionen Φ^{-1}.

4.1 Hvordan gør man, især med ISP

Histogrammer

Et histogram over observationer $y_1, y_2, \ldots, y_n$ fremstilles efter følgende opskrift:

1. Bestem antal delintervaller.
2. Bestem delepunkterne (det er praktisk at der ikke er sammenfald mellem observationer og delepunkter).
3. Tæl op hvor mange observationer der er i hvert interval.
4. For hvert interval udregnes det tilhørende rektangels højde. Hvis der er a observationer i et interval af længde l, så skal højden være a/nl.

5. Tegn histogrammet.

6. Hvis observationerne formodes at være normalfordelte, kan man desuden indtegne tæthedsfunktionen for normalfordelingen med middelværdi $\overline{y}$ og varians s^2.

I ISP er det let at tegne histogrammer, idet der er en makro **histo** der gør arbejdet. Man behøver blot skrive ISP-kommandolinien

```
ISP>>histo y
```

for at få et histogram over dataene i vektoren **y**. Ønsker man et histogram med udfyldte søjler, skal man tilføje **/hty=f**. Ønsker man at få indtegnet den i Punkt 6 omtalte kurve, skal man tilføje **/cur=y**. (Brugeren kan i øvrigt selv bestemme både den indtegnede normalfordelings parametre og antallet og placeringen af intervallerne, samt farver, overskrift og akselabels. Se online-hjælpen.)

Fraktildiagrammer

Et fraktildiagram over observationer $y_1, y_2, \ldots, y_n$ kan fremstilles efter følgende opskrift:

1. Dan de ordnede observationer $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$.

2. Udregn tallene $u_i = \Phi^{-1}(\frac{i-0.5}{n})$, $i = 1, 2, \ldots, n$.

3. Tegn punkterne $(y_{(i)}, u_i)$.

4. For lettere at kunne afgøre om punkterne ligger på en ret linie. indtegnes den rette linie svarende til normalfordelingen med middelværdi $\overline{y}$ og varians s^2, dvs. den linie som går gennem $(\overline{y}, 0)$ og som har hældning $1/s$.

Hvis man benytter *sandsynlighedspapir* er fremgangsmåden lidt anderledes:

1. Dan de ordnede observationer $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$.

2. Afsæt punkterne $(y_{(i)}, \frac{i-0.5}{n})$, $i = 1, 2, \ldots, n$ (man skal benytte *procentskalaen* på ordinataksen).

3. For lettere at kunne afgøre om punkterne ligger på en ret linie, indtegnes den rette linie der svarer til normalfordelingen med middelværdi $\overline{y}$ og varians s^2, dvs. den linie som går gennem punktet $(\overline{y}, 50\%)$ og har hældning $1/s$ når man benytter den *ækvidistante* ordinatskala.

Med ISP kan man danne de ordnede observationer med sorteringskommandoen **sort** (se *Introduktion til ISP*), og der er en ISP-funktion **gauqu()** der udregner Φ^{-1}, så det er ret let at få ISP til at fremstille et fraktildiagram. Der er dog også en færdig makro **fraktil**. Et fraktildiagram for observationerne **z** fås ganske enkelt ved

```
ISP>>fraktil z
```

Hvis man tilføjer **/lin=y** indtegnes den tilsvarende rette linie. (Brugeren kan i øvrigt selv vælge farver for punkter og linie, samt plottype for punkterne, se online-hjælpen.)

Middeltal og standardafvigelse

ISP har funktionerne **mean()** og **sdv()** der udregner middeltal (gennemsnit) og standardafvigelse ('standard deviation'). Hvis man f.eks. vil have beregnet og udskrevet middeltal, standardafvigelse og varians af **w**, kan det gøres med

```
ISP>>print (mean(w)) (sdv(w)) (sdv(w)**2)
```

Bemærk at hvis **w** er et array med r rækker og s søjler, så udregnes middeltal osv. søjlevis, dvs. resultaterne bliver arrays med 1 række og s søjler.

4.2 Opgaver

Opgave 4.1

I begyndelsen af kapitlet er nævnt nogle betingelser der helst skal være opfyldt, hvis man vil beskrive et observationssæt som værende en stikprøve fra normalfordeling:

- Observationerne skal være på en *kontinuert skala*.
- Observationerne skal være fremkommet af den samme "tilfældighedsmekanisme" og uafhængigt af hinanden (de skal, som man siger, være *stokastisk uafhængige* og *identisk fordelte*).
- Observationerne skal fordele sig *symmetrisk* omkring et vist niveau.

Ofte tillader man sig nu at antage at et sæt observationer stammer fra en normalfordeling, hvis der blot ikke er væsentlige argumenter *imod* denne antagelse!

Diskutér om betingelserne er opfyldt for datasæt som de nedenfor antydede. Kan man tillade sig at antage at datasættene er normalfordelte? Hvis ikke, hvad er så de væsentligste grunde dertil?

1. Bredden af kraniet på 20 toårige grønlandske sneharer fanget ved Søndre Strømfjord en bestemt sommer.
2. Vindstyrken kl. 12 på en bestemt lokalitet på 50 på hinanden følgende dage.
3. Vægten af 100 tilfældigt udvalgte sild landet i Gilleleje en bestemt dag.
4. Koncentrationen af NO_x kl. 16.30 ved Nørreport Station hver dag i november måned.
5. Høstudbyttet på hver af 10 forsøgsparceller (à 500 m^2) med en ny sort vinterbyg.
6. Vægten af leveren i 27 fem uger gamle forsøgsmus.
7. Antal nyregistrerede AIDS-tilfælde i Danmark i hver af 12 på hinanden følgende måndeder.

8. Antal nyregistrerede leukæmi-tilfælde i Danmark i hver af 12 på hinanden følgende måndeder.

9. Levetiden af 50 elektriske 40W pærer af samme fabrikat.

10. Det årlige antal trafikulykker i København og Frederiksberg kommuner hvor cyklister er indblandet, for hvert af årene 1980–1990.

Opgave 4.2: Kviksølv i sværdfisk

Sværdfisk kan være en kulinarisk oplevelse, men de er sundest når de ikke indeholder alt for mange tungmetaller. I en undersøgelse af sværdfisk på det amerikanske marked har man målt kviksølvindholdet i 115 tilfældigt udvalgte sværdfisk og fået resultaterne i Tabel 4.1.[4]

Ifølge de amerikanske sundhedsmyndigheder bør (burde) konsumfisk ikke indeholde over 1 ppm kviksølv. Den fisk der sælges via de autoriserede salgskanaler kan man kontrollere (med stikprøvekontroller), og man kan så kassere de partier der indeholder for meget kviksølv. Imidlertid sælges der også en del fisk uden om kontrolmyndighederne — i USA regner man med ca. 25%. Man er interesseret i at vide, hvordan man skal vælge kassationsgrænsen for de 75% kontrollerede fisk for at opnå, at gennemsnitsindholdet af kviksølv i de fisk der når frem til forbrugeren bliver 1 ppm (eller derunder). Hvis man skal kunne beregne denne grænse, er man nødt til at kende fordelingen af kviksølvindhold i sværdfisk.

1. Det ville være bekvemt hvis observationerne kunne beskrives ved en normalfordeling, så det skal undersøges:

   (a) Udregn estimaterne $\overline{y}$ og s^2 over μ og σ^2.

   (b) Tegn et histogram over kviksølvindholdet i de 115 sværdfisk. Indtegn (skitsemæssigt) den fittede normalfordelingstæthed (dvs. tætheden for normalfordelingen med parametre $\overline{y}$ og s^2).

   (c) Tegn et fraktildiagram (f.eks. på sandsynlighedspapir). Indtegn den rette linie der svarer til den fittede normalfordeling.

2. Løs spørgsmål 1 med ISP.

   Tip: Dataene kan indlæses med kommandoen **`getdata`** (data-navn **`svfisk`**). — Benyt ISP's online hjælp for at se hvilke parametre der kan gives til kommandoerne **`histo`** og **`fraktil`**.

[4] Lee & Krutchkoff (1980): Mean and variance of partially-truncated distributions. *Biometrics* **36**, 531-6.

Tabel 4.1: Opgave 4.2: Kviksølvindhold (ppm) i 115 sværdfisk, de ordnede observationer.

0.05	0.07	0.07	0.13	0.13	0.19	0.24	0.25	0.28	0.32
0.39	0.45	0.46	0.53	0.54	0.56	0.60	0.60	0.61	0.62
0.65	0.71	0.72	0.75	0.76	0.79	0.81	0.81	0.82	0.82
0.82	0.83	0.83	0.83	0.84	0.85	0.89	0.90	0.91	0.92
0.92	0.93	0.95	0.95	0.97	0.97	0.98	1.00	1.00	1.01
1.02	1.04	1.05	1.05	1.08	1.10	1.12	1.12	1.14	1.14
1.15	1.16	1.20	1.20	1.20	1.20	1.20	1.21	1.22	1.25
1.25	1.26	1.27	1.27	1.29	1.29	1.29	1.29	1.30	1.31
1.32	1.32	1.37	1.37	1.39	1.39	1.40	1.40	1.41	1.42
1.43	1.44	1.45	1.54	1.54	1.58	1.58	1.60	1.60	1.62
1.62	1.66	1.66	1.68	1.69	1.72	1.74	1.85	1.89	1.96
2.06	2.10	2.23	2.25	2.72					

3. I den oprindelige analyse af tallene gik man ud fra at kviksølvkoncentrationen i sværdfisk var *logaritmisk normalfordelt*, hvilket betyder at *logaritmen* til koncentrationerne er normalfordelt.

   Diskutér denne formodning.

Opgave 4.3

Histogrammer og fraktildiagrammer kan benyttes for at vurdere om et datasæt kan opfattes som observationer fra en og samme normalfordeling. Men hvor meget skal histogrammet afvige fra den "rigtige" form og hvor meget skal fraktildiagrammets punkter afvige fra en ret linie før man må forkaste normalfordelingsantagelsen? For at få en idé om det kan man prøve at lade ISP simulere stikprøver af n normalfordelte tilfældige tal og se hvordan histogrammerne og fraktildiagrammerne for de enkelte stikprøver ser ud.

Fremstil histogrammer (plus normalfordelingstætheden) og fraktildiagrammer (plus den rette linie) for stikprøver bestående af n tilfældige tal fra normalfordelingen med middelværdi 0 og varians 1, for forskellige værdier af n, f.eks. $n = 10, 50, 100$. Lav f.eks. fem af hver slags.

Tip: ISP-funktionen `gauss()` frembringer tilfældige normalfordelte tal med middelværdi 0 og varians 1. Hvis man skriver `u=gauss(array(10))` bliver `u` et array indeholdende 10 tilfældige normalfordelte tal.

I stedet for at skrive de samme kommandolinier et stort antal gange kan man eventuelt lave en lille *makro* der genererer de tilfældige tal og tegner histogram og fraktildiagram. Man kan indlæse en sådan makro kaldet **tegn** på følgende måde:

```
ISP>>input /macro > tegn
m>local u
m>getarg 'n' > n
m>u = gauss(array(n))
m>histo /cur=y u
m>fraktil /lin=y /pch=* u
m>return
m>
ISP>>
```

Makroen skal kaldes med et argument som er den ønskede værdi af n, altså f.eks. `tegn 50`.

Opgave 4.4

Når man har en stikprøve af tilfældige tal fra en normalfordeling hvis *teoretiske* middelværdi er μ, så vil gennemsnittet (den *empiriske* middelværdi) af tallene i stikprøven ligge tæt på μ, men man kan ikke gå ud fra at det er eksakt lig med μ. Dette forhold skal illustreres i denne opgave, der går ud på at beregne gennemsnittet $\overline{y}$ for hver af et antal stikprøver og undersøge hvordan disse gennemsnit fordeler sig.

1. Lav en 49×50-matrix **a** af tilfældige normalfordelte tal med middelværdi 0 og varians 1 med ISP-kommandolinien

   ```
   ISP>>a = gauss(array(49,50))
   ```

   Tænk på hver søjle i **a** som en stikprøve med 49 observationer. Funktionen `mean()` udregner gennemsnittene i hver søjle og anbringer dem i en 1×50-matrix; denne kan laves om til en vektor med funktionen `vec`, så kommandolinien

   ```
   ISP>>gns = vec(mean(a))
   ```

   vil bevirke at **gns** bliver en vektor indeholdende de 50 gennemsnit.

2. Tegn et histogram over de 50 gennemsnit i **gns**. — Hvert af tallene i **gns** er et estimat over den sande middelværdi $\mu = 0$.

3. Udregn gennemsnittet af de 50 gennemsnit i **gns**. — Dette gennemsnit bør i endnu højere grad være lig 0.

4. *Standardafvigelsen* er pr. definition kvadratroden af variansen.

   Udregn standardafvigelsen af de 50 værdier i **gns** (benyt ISP-funktionen `sdv()`).

   Man kan bevise at standardafvigelsen på gennemsnittet af n observationer er lig $1/\sqrt{n}$ gange standardafvigelsen σ på en enkeltobservation.[5] I vores tilfælde er σ lig med 1 og $n = 49$, så middelfejlen er $1/7$. Stemmer det nogenlunde overens med den beregnede standardafvigelse?

[5]Standardafvigelsen på gennemsnittet kaldes ofte for *middelfejlen* på gennemsnittet.

Kapitel 5

En *statistisk* regressionsmodel

I Kapitel 2 har vi beskæftiget os med regressionsmodellen

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

hvor $y_1, y_2, \ldots, y_n$ er observationerne, $x_1, x_2, \ldots, x_n$ er de tilsvarende værdier af den forklarende variabel, β_0 og β_1 er to parametre der beskriver den systematiske del af variationen, og $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er de teoretiske residualer som udtrykker den tilfældige variation. Modellen bliver til en *statistisk* model hvis vi udvider den så at den også modellerer den måde hvorpå residualerne varierer tilfældigt. Man benytter ofte en normalfordeling til at beskrive denne tilfældige variation.

Den *statistiske* model for simpel lineær regression med normalfordelte fejl er

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

hvor x-erne og y-erne og β-erne er som hidtil, og hvor det om residualerne $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ hævdes, at de er stokastisk uafhængige tilfældige tal fra en og samme normalfordeling med middelværdi 0 og varians σ^2. Almindeligvis er variansparameteren σ^2 ukendt, så alt i alt er der de tre ukendte parametre β_0, β_1 og σ^2.

Parametrene β_0 og β_1 estimeres som hidtil ved de værdier $\widehat{\beta}_0$ og $\widehat{\beta}_1$ der minimaliserer residualkvadratsummen $\sum_{i=1}^{n} \Big(y_i - (\beta_0 + x_i\beta_1)\Big)^2$. Variansparameteren σ^2, der skal beskrive størrelsen af y-ernes tilfældige

variation omkring regressionslinien, estimeres ved

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2, \tag{5.1}$$

hvor $\widehat{y}_i = \widehat{\beta}_0 + x_i\widehat{\beta}_1$. Variansskønnet s^2 er altså den faktiske *residualkvadratsum* divideret med antallet af frihedsgrader.

I forhold til de tidligere kapitler indeholder denne *statistiske* model en ekstra antagelse, nemlig at residualerne følger en bestemt sandsynlighedsfordeling. Til gengæld for denne ekstra antagelse kan man så også gøre flere ting med modellen: man kan blandt andet foretage forskellige former for modelkontrol, og man kan få en idé om hvor præcise estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ er.

5.1 Residualer

Den statistiske model går ud fra at de teoretiske residualer $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er stokastisk uafhængige og normalfordelte med middelværdi 0 og varians σ^2. Imidlertid kender vi ikke de teoretiske residualer, kun de *empiriske residualer* $e_i = y_i - \widehat{y}_i$, $i = 1, 2, \ldots, n$.

Som forklaret i Kapitel 2 vil de fittede værdier $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$ opfylde estimationsligningerne (2.3) og (2.4) (side 15). Hvis man i hver af disse ligninger flytter venstresiden over på den anden side af lighedstegnet og benytter definitionen på e-erne, får man

$$0 = \sum_{i=1}^{n} e_i,$$

$$0 = \sum_{i=1}^{n} x_i e_i.$$

Der er altså lagt *to bånd* på residualerne $e_1, e_2, \ldots, e_n$, og derfor får residualkvadratsummen $n-2$ *frihedsgrader*, se evt. Opgave 3.4. De to ligninger viser også, at de empiriske residualer ikke er uafhængige af hinanden.

Antagelsen om at de teoretiske residualer $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er stokastisk uafhængige tilfældige tal fra normalfordelingen med middelværdi 0 og varians σ^2 medfører desuden, at hvert e_i vil være normalfordelt med middelværdi 0 og med en bestemt varians, som er lidt mindre end σ^2.

Box 5.1: En regneregel for varianser

Hvis ε er en stokastisk variabel med varians σ^2 og a og b er konstanter, så har $a\varepsilon + b$ varians $a^2\sigma^2$:

$$\mathrm{Var}(a\varepsilon + b) = a^2\mathrm{Var}(\varepsilon).$$

Standardafvigelsen på $a\varepsilon + b$ er dermed $|a|\sigma$.

Hvis man skriver op hvordan de empiriske residualer $e_1, e_2, \ldots, e_n$ eksplicit afhænger af $y_1, y_2, \ldots, y_n$ eller af $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ og dernæst benytter regnereglerne for varianser (Box 5.1), vil man kunne udlede udtryk for e_i-ernes varianser og kovarianser. Det vil vi ikke gøre her. Vi vil blot nævne, at man plejer at skrive variansen på e_i som $(1 - h_i)\sigma^2$, hvor h_i i tilfældet *simpel lineær regression* er

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \tag{5.2}$$

og i andre tilfælde mere indviklet.

Når e_i har varians $(1 - h_i)\sigma^2$, har størrelsen $e_i/((1 - h_i)\sigma^2)^{1/2}$ varians 1. I residualundersøgelser benyttes derfor ofte de såkaldte *standardiserede residualer*

$$\begin{aligned} e_i' &= \frac{e_i}{((1 - h_i)s^2)^{1/2}} \\ &= \frac{e_i}{(1 - h_i)^{1/2}s}. \end{aligned} \tag{5.3}$$

der har middelværdi 0 og en varians der er cirka 1.

5.2 Residualundersøgelse

Residualundersøgelser er et vigtigt led i modelkontrollen ved alle slags regressionsanalyser. Man plejer at fremstille *residualplots* hvor man plotter residualerne (de rigtige eller de standardiserede) mod forskellige størrelser, f.eks.

- et *indexplot*, dvs. et plot af punkterne (i, e_i), $i = 1, 2, \ldots, n$,

- et plot af residualerne mod den forklarende variabel, dvs. et plot af punkterne (x_i, e_i), $i = 1, 2, \ldots, n$,
- et plot af residualerne mod de fittede værdier, dvs. et plot af punkterne $(\widehat{y}_i, e_i)$, $i = 1, 2, \ldots, n$. [1]

Hvis modellen giver en tilstrækkelig god beskrivelse af observationerne, så skulle alle disse plots vise punkter der fordeler sig tilfældigt omkring linien $e = 0$.

Endvidere kan man for at kontrollere normalfordelingsantagelsen lave histogrammer og fraktildiagrammer over de standardiserede residualer.

5.3 Hvordan gør man med ISP

1. Den estimerede standardafvigelse s hedder **sigma** i udskrifterne fra **regress**.
2. Residualerne får man udregnet ved på **regress**-kommandolinien at tilføje et **>** (som skal komme efter navnene på de arrays der skal "regresses") og til højre herfor skrive **re:** efterfulgt af navnet på den ISP-vektor hvor residualerne skal anbringes, f.eks.

   ```
   ISP>>regress  x y  >  re:res1
   ```

   (se også *Introduktion til ISP*).
3. De fittede værdier $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$ udregnes allerlettest ved hjælp af formlen $\widehat{y} = y - e$.
4. De ovenfor omtalte h_i-er der skal bruges i formlen for de standardiserede residualer fås ved til højre for **>** at skrive **dh:** efterfulgt af navnet på den ISP-vektor hvor de skal anbringes; desuden skal man tilføje **/bmat=y** .

   Eksempel: Kommandolinien

   ```
   ISP>>regress  x y  >  re:res1  dh:h /bmat=y
   ```

[1] Bemærk at man almindeligvis *ikke* plotter residualerne mod de *observerede* værdier, se Opgave 5.5.

bevirker at der foretages lineær regression af `y` på `x`, at residualerne anbringes i `res1` og at h-værdierne udregnes og anbringes i arrayet `h`.

Bemærk at `y` i `/bmat=y` står for *yes*; det har ikke noget med regressionsanalysens y at gøre. Bemærk også at `dh:h` ignoreres hvis man glemmer/udelader `/bmat=y`.

5.4 Opgaver

Opgave 5.1

Betragt følgende datasæt der er så lille at man kan analysere det uden brug af computer:

y	x
14.9	0
11.6	-1
16.2	1
9.8	-2
17.5	2

Hvis man "regresser" y på x får man regressionsligningen $y = 14 + 2x$.

1. Udregn de fittede værdier $\widehat{y}_1, \widehat{y}_2, \widehat{y}_3, \widehat{y}_4, \widehat{y}_5$.
2. Udregn residualerne $e_i = y_i - \widehat{y}_i$.
3. Eftervis at $\sum_{i=1}^{5} e_i$ og $\sum_{i=1}^{5} x_i e_i$ begge er 0.
4. Udregn residualkvadratsummen og variansskønnet s^2.
5. Tegn de forskellige residualplots der er omtalt i Afsnit 5.2

Opgave 5.2

I denne opgave arbejder vi videre med Forbes' data fra Opgave 2.3 (side 23). Benyt logaritmen til lufttrykket. Det er ikke nødvendigt at omregne tallene til moderne måleenheder (hvorfor?).

1. Tegn et *indexplot* af residualerne.

   Hvilke slags afvigelser fra modellen kan et indexplot afsløre?

2. Tegn et plot af residualerne e_i mod x_i.

   Hvilke slags afvigelser kan det afsløre?

   Prøv også at lave regression af lufttrykket selv på kogepunktet, og lav et residualplot.

3. Tegn et plot af residualerne e_i mod de fittede værdier $\widehat{y}_i$.

   I nogle situationer er det sådan, at størrelsen af den tilfældige variation afhænger af den forventede værdi af y (dvs. der er ikke *varianshomogenitet*). Det kan måske afsløres med denne type plot (hvordan?).

Opgave 5.3

Lav plottene i Opgave 5.2 med standardiserede residualer i stedet for "almindelige" residualer. Er der nogen synlig forskel?

Opgave 5.4

I Forbes-dataene er der (som det formentlig efterhånden er blevet klart) et mistænkeligt datapunkt. Fjern det og gentag Opgaverne 5.2 og 5.3.

Tip: I ISP kan man udelade et (eller flere) datapunkter fra analysen ved at sætte en af koordinaterne (eller dem alle) til "manglende værdi", der i ISP kan skrives ?. Her vises én måde at gøre det på (i Forbes-eksemplet), hvorved selve datavektorerne ikke ødelægges:

```
ISP>>minus = 0*kp  # en 0-vektor af samme længde som kp
ISP>>minus(12) = ? # værdi nr. 12 sættes til ?
ISP>>regress  (kp+minus) (log(tr))
```

Opgave 5.5

Denne opgave går ud på at lave residualundersøgelser i en situation hvor man véd at modellen er rigtig — fordi man selv har lavet tallene!

Lav med ISP en vektor **x** indeholdende 10 0-er, 10 4-er, 10 8-er og 10 12-er. Lav dernæst en vektor **y** efter opskriften $y = 5 + x + \varepsilon$ hvor $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{40}$ er tilfældige normalfordelte tal med middelværdi 0 og varians 4.

Tip: En ukompliceret måde at fremstille **x** på er

```
ISP>>x = array(40)
ISP>>x(11:20) = 4
ISP>>x(21:30) = 8
ISP>>x(31:40) = 12
```

Tabel 5.1: Vedr. Opgave 5.6.

y	x
y_1	-2
y_2	-1
y_3	0
y_4	1
y_5	2

De tilfældige tal fremstilles med ISP-funktionen **gauss()** der leverer normalfordelte tal med middelværdi 0 og varians 1. Ved at gange dem med 2 får man tilfældige tal med varians $2^2 = 4$. Eksempelvis kan man gøre sådan:

```
ISP>>y = 5 + x + 2*gauss(x)
```

Foretag nu regressionsanalyse af y på x og find residualerne. Lav de forskellige residualplots som er omtalt i Afsnit 5.2.

Et af de foreslåede plots er et plot af residualerne mod de *fittede* værdier. Prøv at lave et plot af residualerne mod de *observerede* værdier. Hvorfor ser det ud som det gør?

Opgave 5.6

Denne opgave går ud på at undersøge hvordan forskellige størrelser afhænger af observationerne $y_1, y_2, \ldots, y_n$, så derfor figurerer y-erne som bogstaver, medens vi bruger talværdier for x-erne (for at gøre det lidt nemmere).

Betragt det regressionsproblem med de $n = 5$ sammenhørende værdier af x og y der er givet i Tabel 5.1.

1. Find $\overline{x}$ og $\sum_{i=1}^{n}(x_i - \overline{x})^2$.

2. Vis at $\widehat{\beta}_0 = 0.2y_1 + 0.2y_2 + 0.2y_3 + 0.2y_4 + 0.2y_5$ og at $\widehat{\beta}_1 = -0.2y_1 - 0.1y_2 + 0y_3 + 0.1y_4 + 0.2y_5$.

3. Skriv op hvordan $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_5$ beregnes ud fra y_i-erne.

4. Hvilke observationer har størst indflydelse på $\widehat{y}_1$? og på $\widehat{y}_3$?

Opgave 5.7

Man kan komme ud for at enkelte datapunkter dels afviger en del fra modellen, dels har stor indflydelse på modellens parametre. Man kan overveje om sådanne betænkelige punkter helt skal udelades.

Et datapunkt af den omtalte slags kan kendes på, at hvis man forsøgsvis udelader det fra modellen, så bliver parameterestimaterne ændret en hel del. En strategi til at opdage sådanne betænkelige punkter kunne derfor være, at man for hvert enkelt datapunkt prøvede at udelade det og se hvor meget den fittede model derved ændrede sig. Man kan bede ISP's **regress**-kommando gøre dette. Hvis man på **regress**-kommandolinien tilføjer `bm:dif /bmat=y`, så vil `dif` blive en ISP-matrix med en række for hvert datapunkt og en søjle for hver parameter; den i-te række indeholder ændringerne i parameterestimaterne svarende til at man udelader det i-te datapunkt.

Gør dette med Forbes-dataene, hvor kogepunktet benyttes som x og logaritmen til lufttrykket som y.

Kapitel 6

Parameterestimaternes fordeling

I den statistiske model for lineær regression udtaler man sig om residualernes og dermed også y-ernes fordeling. Da parameterestimaterne er beregnet ud fra y-erne (og fra x-erne der opfattes som kendte konstanter), så bliver det en konsekvens af modellen, at estimaterne må betragtes som tilfældige tal fra visse sandsynlighedsfordelinger som man i princippet kan bestemme.

Lad os præcisere situationen, der er den samme som i Kapitel 5: Der er de kendte værdier $x_1, x_2, \ldots, x_n$ af baggrundsvariablen x, der er observationerne $y_1, y_2, \ldots, y_n$, der er de ukendte parametre β_0 og β_1, og endelig er der residualerne $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ som ifølge modellen er tilfældige tal fra en nærmere angivet sandsynlighedsfordeling; denne fordeling "plejer" at være en normalfordeling med middelværdi 0 og varians σ^2 hvor σ^2 er endnu en parameter. Det hele hænger sammen via relationen $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i,\ i = 1, 2, \ldots, n$.

Men spørgsmålet er nu, hvordan man rent faktisk kan bestemme fordelingen af estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ (og for den sags skyld også af variansskønnet s^2). Sagen kan angribes på to principielt forskellige måder:

1. Man udnævner nogle bestemte værdier af β_0, β_1 og σ^2 til at være de "sande" værdier (man kan f.eks. bruge estimaterne beregnet på grundlag af de faktiske observationer); derved er den sandsynlighedsmodel der frembringer y fuldstændig kendt. Man kan nu sætte computeren til at simulere et sæt *pseudoobservationer* $y_1^\star, y_2^\star, \ldots, y_n^\star$ på den måde, at den først udregner tilfældige tal

Box 6.1: Middelfejl

Middelfejlen (på engelsk 'the *standard error*') på et parameterestimat er defineret som standardafvigelsen på estimatet.
Betegnelsen skal blandt andet tjene til at hindre forveksling med begrebet "standardafvigelsen på observationerne".

$\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*$ fra en normalfordeling med middelværdi 0 og varians σ^2 og derpå sætter

$$y_i^* = \beta_0 + x_i\beta_1 + \varepsilon_i^*, \quad i = 1, 2, \ldots, n.$$

På grundlag af disse pseudoobservationer bestemmes pseudoestimater $\widehat{\beta}_0^*$ og $\widehat{\beta}_1^*$.

Denne proces kan gentages et større antal gange (f.eks. et par hundrede), og man får derved en hel stribe forskellige værdier af $\widehat{\beta}_0^*$ og $\widehat{\beta}_1^*$. Disse værdier fordeler sig på en bestemt måde som er de pågældende estimaters fordeling. (En god måde at illustrere fordelingerne på er ved at tegne histogrammer over dem.)

Den her beskrevne fremgangsmåde er et eksempel på den såkaldte *bootstrap*-metode.[1]

2. Estimaterne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ er veldefinerede og pæne funktioner af $y_1, y_2, \ldots, y_n$ og dermed også af $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, og matematikken (mere præcist sandsynlighedsregningen) er leveringsdygtig i sætninger der fortæller, at hvis $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er tilfældige tal fra en bestemt sandsynlighedsfordeling givet ved en bestemt sandsynlighedstæthedsfunktion, så vil estimaterne være at opfatte som tilfældige tal fra visse andre sandsynlighedsfordelinger som man kan angive formeludtryk for. Man kan altså løse problemet ad matematisk vej. Vi skal ikke her komme ind på hvordan, da vi i så fald først skulle stable et større matematisk apparatur på benene.

   Hvis residualerne $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er stokastisk uafhængige og normalfordelte med middelværdi 0 og varians σ^2, så får man pæne løsninger; for fuldstændighedens skyld anføres disse løsninger her i deres fulde udstrækning (i formlerne indgår symbolet SS_x som er en hyppigt anvendt forkortelse[2] for $\sum_{i=1}^{n}(x_i - \overline{x})^2$):

[1] Opgave 4.4 er et simpelt eksempel på denne metode.
[2] SS = Sum of Squares (of deviations).

Box 6.2: Sikkerhedsintervaller

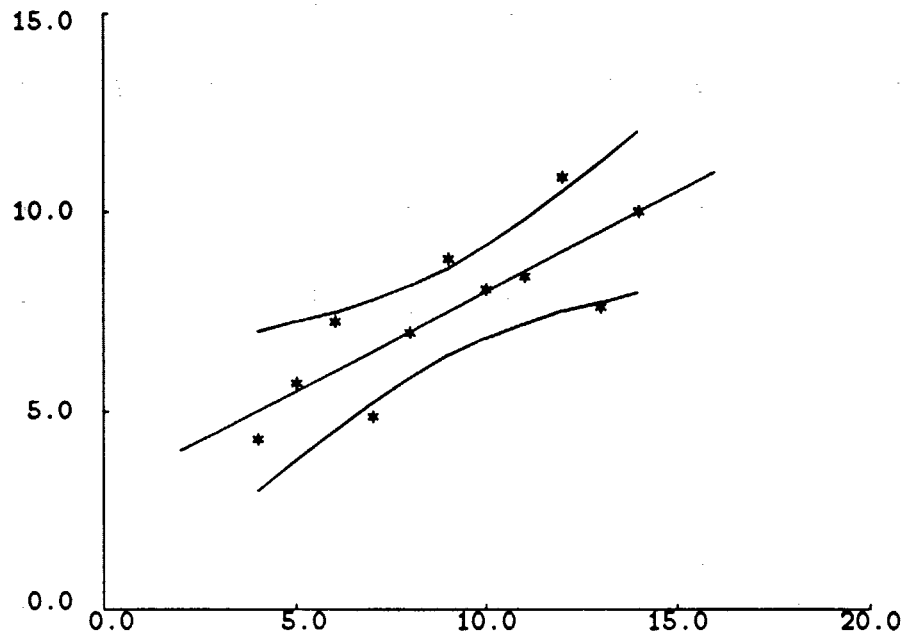
Et 95% *sikkerhedsinterval* (eller *konfidensinterval*) for en parameter β er et interval hvis endepunkter er bestemt ud fra observationerne, og som har den egenskab, at der er 95% sandsynlighed for at intervallet indeholder den sande β-værdi (der jo er ukendt).
Hvis parameterestimatet er normalfordelt omkring den sande værdi, kan man fastlægge et 95% sikkerhedsinterval som estimatet $\pm$ 1.96 gange estimatets middelfejl.

- Estimatet $\widehat{\beta}_1$ er normalfordelt,
  - dets middelværdi er β_1,
  - dets varians er σ^2/SS_x,
  - dets middelfejl er $\sigma/\sqrt{SS_x}$.
- Estimatet $\widehat{\beta}_0$ er normalfordelt,
  - dets middelværdi er β_0,
  - dets varians er $\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)$,
  - dets middelfejl er $\sigma\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)^{1/2}$.
- Variansskønnet s^2 er χ^2-fordelt med $f = n-2$ frihedsgrader og med skalaparameter σ^2/f, dvs.
  - dets middelværdi er σ^2,
  - dets varians er $2\sigma^4/f$,
  - dets middelfejl er $\sigma^2\sqrt{2/f}$.
- Estimaterne $\widehat{\beta}_1$ og $\widehat{\beta}_0$ er *ikke* stokastisk uafhængige. Derimod er $\widehat{\beta}_1$ stokastisk uafhængig af estimatet $\widehat{\beta}_0 + \overline{x}\widehat{\beta}_1 = \overline{y}$ over regressionsliniens skæringspunkt med den lodrette linie med ligning $x = \overline{x}$.

  Endvidere er variansskønnet s^2 stokastisk uafhængigt af $\widehat{\beta}$-erne.

Bemærk at i alle tilfælde er estimaternes middelværdi lig med den sande værdi af den parameter som estimaterne skal estimere.

Bemærk også at i udtrykkene for estimaternes varians og middelfejl indgår de sande ukendte parameterværdier β_0, β_1 og σ^2.



Figur 6.1: Et "typisk" eksempel på et sæt punkter med den tilhørende regressionslinie samt et 95% sikkerhedsområde for linien.

I praksis må man naturligvis erstatte dem med de tilsvarende estimerede værdier $\widehat{\beta}_0$, $\widehat{\beta}_1$ og s^2.

Ved hjælp af estimaternes middelfejl kan man bestemme 95% sikkerhedsintervaller for β_0 og β_1, se Box 6.2.

Man kan i øvrigt også bestemme et "sikkerhedsinterval" for regressionslinien som sådan, dvs. et bælte inden for hvilket den sande rette linie ligger med en given sandsynlighed, se Figur 6.1. Bæltet bestemmes som den estimerede linie plus/minus en afvigelse der er et produkt af tre faktorer: 1) den estimerede standardafvigelse s, 2) noget der afhænger af sikkerhedsgraden α (f.eks. 95%), og 3) noget der afhænger af x. Man kan vise at den tredje faktor skal være

$$\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{SS_x}\right)^{1/2}$$

og at den anden faktor skal være kvadratroden af to gange α-fraktilen i F-fordelingen med 2 og $n-2$ frihedsgrader; disse ting gælder kun for simpel lineær regression.

6.1 Hypoteser om modellens parametre

En grund til at det er interessant at kende estimaternes fordelinger er, at man derigennem får information om hvilke andre værdier man også kunne have fået. På den måde kan man udtale sig om hvor præcist estimatet er (jf. sikkerhedsintervallerne), og man kan teste statistiske hypoteser om parametrene.

I den simple lineære regressionsmodel kan man undertiden finde på at teste den statistiske hypotese $H: \beta_1 = 0$ om at hældningskoefficienten er nul, dvs. at y slet ikke afhænger af x. Hvis denne hypotese er rigtig, må man formode at estimatet $\widehat{\beta}_1$ ligger temmelig tæt ved 0, i hvert fald når man måler med en målestok der er indrettet efter størrelsen af usikkerheden på $\widehat{\beta}_1$. Man bærer sig derfor således ad:

1. Som *teststørrelse* benytter man afstanden mellem $\widehat{\beta}_1$ og 0 målt med $\widehat{\beta}_1$s (estimerede) middelfejl som målestok: man udregner

$$t_{\text{obs}} = \frac{\widehat{\beta}_1 - 0}{s/\sqrt{SS_x}}.$$

2. Hvis t_{obs} er tæt på nul, så er observationerne og hypotesen H godt forenelige; hvis omvendt t_{obs} er langt fra nul, så er observationerne og hypotesen ikke forenelige, og så forkaster man hypotesen.

3. Det kan bevises, at under H er fordelingen af teststørrelsen den såkaldte *t*-fordeling med samme antal frihedsgrader som s^2, det vil her sige $n - 2$ frihedsgrader.

   Ved hjælp af tabeller over t-fordelingen kan man derfor finde *testsandsynligheden*, hvilket er sandsynligheden for at få værdier større end $|t_{\text{obs}}|$ eller mindre end $-|t_{\text{obs}}|$ (og det er det samme som to gange sandsynligheden for at få værdier større end $|t_{\text{obs}}|$).

4. Hvis testsandsynligheden er meget lille, så forkastes hypotesen. Hvis testsandsynligheden ikke er meget lille, så kan man ikke på det foreliggende grundlag afvise hypotesen.

5. Hvis hypotesen $H: \beta_1 = 0$ bliver *forkastet*, så siger man at β_1 er *signifikant forskellig fra nul*. Somme tider vil man også sige at den forklarende variabel x er signifikant eller har en signifikant virkning.

Box 6.3: Statistiske hypoteser

En *statistisk hypotese* er en påstand om, at man kan klare sig med en simplere statistisk model end den aktuelle grundmodel. Hypotesen vil ofte gå ud på, at en eller flere af modellens parametre er ens eller at de er lig med 0.

Man *tester* en hypotese ved at se på, hvor godt man kan beskrive sine observationer under hypotesen[a] i forhold til hvor godt grundmodellen beskriver dem.

[a] "Under hypotesen" er statistisk jargon for "under forudsætning af at hypotesen er rigtig".

Box 6.4: Statistiske test

Et test af en statistisk hypotese foregår i to skridt:

Først udregnes en *teststørrelse*, dvs. en bestemt velvalgt funktion af observationerne, som måler hvor meget disse afviger fra det man skulle vente under hypotesen.

Dernæst findes *testsandsynligheden*, dvs. sandsynligheden for at få en mere ekstrem teststørrelseværdi end den faktisk opnåede, forudsat at hypotesen er rigtig. Hvis testsandsynligheden er lille, er der med testet konstateret en uoverensstemmelse mellem observationer og hypotese, og man *forkaster* hypotesen. Hvis testsandsynligheden ikke er lille, er der ikke med det pågældende test konstateret uoverensstemmelse mellem observationer og hypotese, og man kan ikke afvise hypotesen.

Man siger at der er *signifikans på niveau* α, hvis testsandsynligheden er udregnet til α (og man forkaster hypotesen).

Hvis man har den strategi at forkaste hypotesen hver gang testsandsynligheden er mindre end α, så siger man at man benytter et *signifikansniveau* på α, og man kan fortolke α som sandsynligheden for at forkaste hypotesen selv om den er rigtig. — Man benytter ofte $\alpha = 0.05$.

Box 6.5: En- og tosidede t-tests

Antag at man tester en hypotese $H : \beta = 0$ med et t-test. Sædvanligvis forkaster man hypotesen når t_{obs} numerisk er stor. Testsandsynligheden er da sandsynligheden for at få værdier større end $|t_{\text{obs}}|$ plus sandsynligheden for at få værdier mindre end $-|t_{\text{obs}}|$. Dette kaldes et *tosidet* test.

Nogle gange går man ud fra at enten er $\beta = 0$ eller også er $\beta > 0$. I så fald forkaster man kun H når t_{obs} er stor og positiv. Testsandsynligheden er da sandsynligheden for at få værdier større end t_{obs}. Dette kaldes et *ensidet* test.

(Noget lignende kan gælde andre tests.)

Det er takket være den statistiske model at man kan tale om sandsynligheden for at få en t-værdi der opfylder en bestemt betingelse (så som at ligge længere fra nul end et givet tal). For t er jo en bestemt funktion af observationerne $y_1, y_2, \ldots, y_n$, der ifølge modellen og hypotesen fremkommer ud fra tilfældige tal fra nogle nærmere specificerede sandsynlighedsfordelinger. Derfor vil t være at betragte som et tilfældigt tal fra en vis sandsynlighedsfordeling. Situationen er nøjagtig den samme som for parameterestimaternes vedkommende, og man kan da også bestemme fordelingen af t på to forskellige metoder: enten ved hjælp af en *bootstrap*-metode (og en computer) eller ved hjælp af matematik. Også her gælder, at når residualerne er normalfordelte med middelværdi 0 og med samme varians, så giver matematikmetoden et "pænt" svar, nemlig at fordelingen af t er den såkaldte *t-fordeling* (også kaldet 'Student's t-fordeling) med samme antal frihedsgrader som variansestimatet s^2. Denne fordeling er en kendt fordeling der findes i statistiske tabeller, både elektroniske tabeller (på computere) og bøger.

6.2 Modelkontrol

Det altid påtrængende problem ved arbejdet med statistiske modeller er modelkontrolproblemet: hvordan vurderer man om modellerne er gode *nok*? I nogle situationer kan man inddrage test af statistiske hypoteser.

Antag eksempelvis at det almindelige scatterplot af y mod x viser punkter der så nogenlunde ligger på en ret linie, men at linien måske har en tendens til at krumme. Så kunne man forsøge sig med en polynomiel

regressionsmodel som grundmodel, f.eks.

$$y = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon,$$

og i denne model teste hypotesen $H : \beta_2 = 0$. Hvis denne hypotese bliver forkastet, er det tegn på at en ret linie ikke er så velegnet til at beskrive de pågældende data.

Mere generelt kan man vurdere brugbarheden af en formodet model ved at udvide den til en mere omfattende model hvis brugbarhed man (næsten) ikke vil drage i tvivl, og derefter teste den formodede model som en hypotese i forhold til den store model.

6.3 Hvordan gør man med ISP

ISP's `regress`-kommando beregner automatisk estimaternes middelfejl; i udskrifterne optræder de under betegnelsen `sdev`. Antallet af frihedsgrader beregnes også (`degrees of freedom`). Man må så selv sørge for at udregne t-størrelsen efter opskriften "parameterestimat divideret med middelfejl".

Testsandsynligheden kan derefter findes ved hjælp af funktionen `tpr()`, der udregner sandsynligheder i t-fordelingen. Der er to argumenter til `tpr()`, nemlig den observerede t-værdi t_{obs} og antallet af frihedsgrader. Funktionskaldet `tpr(tobs,f)` returnerer sandsynligheden for at få værdier mindre end t_{obs}. Testsandsynligheden i det tosidede t-test kan derfor f.eks. udregnes således, hvor `tobs` og `f` naturligvis på forhånd skal være sat til at indeholde de rigtige værdier:

```
ISP>>ssh = 2 * ( 1 - tpr(abs(tobs),f) )
```

Hvis testsandsynligheden er meget lille, vil denne beregningsmetode dog give et forkert resultat på grund af afrundingsfejl; det kan derfor anbefales i stedet at udregne sandsynligheden som

```
ISP>>ssh = 2 * tpr(-abs(tobs),f)
```

•

Hvis man vil have tegnet et sikkerhedsområde for regressionslinien som i Figur 6.1 og man er tilfreds med at få det gjort gennem punkter med x-koordinater $x_1, x_2, \ldots, x_n$, kan det gøres forholdsvis let, idet den x-afhængige faktor i disse punkter simpelt hen er lig med størrelsen h fra Kapitel 5 (formel (5.2) side 59), og denne størrelse beregnes af `regress` med `/bmat=y`. Man kan f.eks. gøre sådan:

```
ISP>>sort x y > x y  # sortér efter x-værdierne
ISP>>regress x y /bmat=y > re:res dh:h
degrees of freedom:     11 - 2 = 9
...
ISP>>yhat = y-res # de fittede værdier
#   fqu(0.95,2,9)  beregner 95%-fraktilen i F-fordelingen
#   med 2 og 9 frihedsgrader
ISP>>faktor2 = sqrt(2*fqu(0.95,2,9))
ISP>>faktor3 = sqrt(h)
ISP>>afv = sigma * faktor2 * faktor3
#   Bæltets grænser tegnes således:
ISP>>glue /axis=2  -1 1 > fortegn
ISP>>gscat x (yhat+fortegn*afv) /pty=1
```

Denne metode kan anvendes generelt, blot skal `faktor2` så beregnes som `sqrt(p*fqu(alfa,p,n-p))` hvor p er antallet af estimerede parametre.

6.4 Opgaver

Opgave 6.1

Georg har slået Plat eller Krone 5 gange med en almindelig mønt og fået netop én gang Krone. Gerda siger at det da må tyde på at mønten er skæv, ellers skulle man have fået 2 eller 3 gange Krone.

For at afgøre om man på denne baggrund kan sige at mønten er skæv, kan man foretage sig forskellige ting:

1. Man kan lave et *modelforsøg*: Man tager en mønt som man mener *ikke* er skæv. Denne mønt kaster man 5 gange og tæller antal Krone; dette gentages et større antal gange for at få et indtryk af hvordan "antal Krone i 5 kast" fordeler sig.

   (a) Gør dette M gange (M skal nok være mindst 10) og fremstil resultatet i form af et diagram med hyppighedspinde (dvs. tallene 0,1,2,3,4,5 (de mulige antal Krone) ud ad x-aksen og hyppighed ud ad y-aksen).

   (b) Hvor stor en brøkdel af de M gange er antallet af Krone netop lig 1 ? Hvor stor en brøkdel af de M gange er antallet 1 eller mindre?

   (c) Når man skal sammenholde noget observeret (her: 1 Krone i 5 kast) med en model (her: den ikke-skæve mønt sådan som den har manifesteret sig i de M gentagelser), plejer statistikeren at spørge om sandsynligheden for at få noget "værre" end det observerede, dvs. noget mindst lige så afvigende som det observerede. Noget "værre" er i dette eksempel noget der ligger mindst lige så langt fra det forventede ($5/2 = 2.5$) som den observerede værdi gør, dvs. det er værdierne 0, 1, 4 og 5. — Hvor stor en brøkdel af de M gange gav et udfald der var værre end udfaldet 1 ?

2. Man kan også lave en sandsynlighedsmodel for møntkastene. Der bliver tale om en *binomialfordelingsmodel*, som vi ikke på dette sted skal udlede. Man kan vise at sandsynligheden for at få netop x gange Krone i 5 kast med en ikke-skæv mønt er $p_x = \binom{5}{x} 0.5^x (1 - 0.5)^{5-x}$. Værdierne af disse sandsynligheder

er:

x	p_x
0	0.03125
1	0.1563
2	0.3125
3	0.3125
4	0.1563
5	0.03125

Hvad er ifølge denne sandsynlighedsmodel sandsynligheden for at få noget "værre" end observationen 1 Krone?

(Som tommelfingerregel/konvention siger man, at hvis der blot er 5% chance for at få noget "værre", så er der ikke nogen signifikant uoverensstemmelse mellem data og model.)

Opgave 6.2

Man kan simulere modelforsøget fra Opgave 6.1 med ISP.

ISP-funktionen `unif()` frembringer tilfældige ligefordelte tal mellem 0 og 1. Det kan vi bruge til at simulere møntkast:

1. For at få lavet en serie på 5 kast kan man f.eks. skrive `serie = unif(array(5)) < 0.5`. Derved kommer `serie` til at indeholde 5 tilfældige nuller og ettaller. Hvis vi lader 0 svare til Plat og 1 til Krone, har vi dermed fået simuleret 5 møntkast med en ikke-skæv mønt. Prøv!

2. Hvordan simulerer man kast med en skæv mønt?

3. Hvad er resultatet af `unif(array(5,3)) < 0.5` ?
   og af `sum( unif(array(5,3)) < 0.5)` ?

4. Man kan lave $M = 100$ gentagelser af det forsøg der består i at finde antal Krone i fem kast med kommandolinien
   `gentag = sum( unif(array(5,100)) < 0.5 )`.
   Gør det!

5. Arrayet `gentag` indeholder nu 100 værdier; man vil gerne vide hvor mange gange hver af de seks mulige værdier 0,1,2,3,4,5 optræder. Det kan f.eks. gøres således: Indfør `x = iota(array(6))-1`. Da `gentag` er et array med 1 række

og 100 søjler[3], er resultatet af `x == gentag` et array med 6 rækker og 100 søjler, og værdien i den i-te række og j-te søjle er 1 hvis $\mathbf{x}_i$ er lig $\mathbf{gentag}_j$ og 0 ellers. De søgte antal er derfor de seks rækkesummer i arrayet `x == gentag`, og de kan f.eks. findes med `reduce/axis=2 (x==gentag) > antal`
Gør dette!

I øvrigt kan ISP-kommandoen `pbinom` udregne binomialfordelingssandsynligheder. Tabellen i Opgave 6.1 er udregnet sådan (`x` er som ovenfor):

```
pbinom x 5 0.5 > kumssh
ssh = diff(kumssh)
print ssh
```

Opgave 6.3

Her er en udskrift af noget af en monitorfil fra løsningen af Opgave 3.3 (vands strømningsforhold i en flod):

```
ISP>>glue/axis=2 (dybde**2) dybde > dyb2
ISP>>regress dyb2 flow
degrees of freedom:     10 - 3 = 7
sigma      =  0.2794
R-square   =  0.9900
F-stat     =   346.5       (2 over 7 df)
condition =   22.42

 var          coef          sdev
   1           23.54         4.274
   2          -10.86         4.517
   const       1.683         1.059
ISP>>
```

Man har altså fittet den bedste andengradskurve.

1. (a) Afgør ved hjælp af et tosidet t-test[4] om koefficienten til andengradsleddet er signifikant forskellig fra 0 (dvs. om andengradskurven er signifikant bedre end en ret linie).

[3]Det kan man erfare ved brug af `list`-kommandoen eller ved at skrive `print (dims(gentag))`.

[4]Et uddrag af en tabel over t-fordelingen er vist i Tabel 6.1; tabellen er beregnet med ISP's `tqu()`-funktion.

Tabel 6.1: Fraktiler i t-fordelingen

	Sandsynlighed i procent				
f	90	95	97.5	99	99.5
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

(b) Afgør ved hjælp af et tosidet t-test om koefficienten til førstegradsleddet er signifikant forskellig fra 0.

(c) Afgør ved hjælp af et tosidet t-test om kurven lige så godt kunne gå gennem $(0, 0)$.

2. Løs ovenstående *uden* en tabel over fraktiler i t-fordelingen, men direkte med ISP (udregn testsandsynlighederne ved hjælp af `tpr()`-funktionen).

Opgave 6.4: Cellers overlevelse

Som led i en undersøgelse af hvordan levende celler reagerer når de bliver udsat for stråling, har man foretaget et lille eksperiment hvor man har udsat 14 plader med en vis slags celler for en bestemt type stråling i forskellige doser. Efter et givet stykke tid har man derefter opgjort, hvor stor en del af cellerne på hver enkelt plade der havde overlevet strålebehandlingen. Resultaterne fremgår af Tabel 6.2.

Biologerne mener at kunne sige følgende om sammenhængen mellem overlevelsesraten (= brøkdel overlevende) og strålingsdosis:

1. Når der ingen stråling er, er overlevelsesraten lig 1.

2. Det er hensigtsmæssigt at modellere *logaritmen* til overlevelsesraten.

Tabel 6.2: Opgave 6.4: Cellers overlevelse ved forskellige strålingsdoser.

plade nr.	dosis (10^{-4}Gy)	brøkdel overlevende
1	1.175	0.44
2	1.175	0.55
3	2.35	0.16
4	2.35	0.13
5	4.70	0.0400
6	4.70	0.0196
7	4.70	0.0612
8	7.05	0.0050
9	7.05	0.0032
10	9.40	0.00110
11	9.40	0.00015
12	9.40	0.00019
13	14.10	0.00700
14	14.10	0.00006

3. Logaritmen til overlevelsesraten formodes at afhænge enten lineært eller kvadratisk af strålingsdosis. — Man véd ikke hvilken af delene, men man er meget interesseret i at få det at vide.

Opgaven er nu at undersøge hvilken af de to skitserede modeller (jf. pkt. 3) der synes at være bedst. I den forbindelse skal man blandt andet

- give en præcis formulering af de statistiske modeller der benyttes,
- estimere modellernes parametre,
- og foretage passende former for modelkontrol.

Skal man foretrække en lineær eller en kvadratisk model?

Biologerne indrømmer at det godt kan være svært at foretage præcise bestemmelser af brøkdel overlevende celler, og der er måske en "smutter". — Hvordan ser situationen ud efter denne oplysning?

Tip: Data kan indlæses til et ISP-array `obs` med kommandoen **`getdata`** (data-navn `celler`)

Tabel 6.3: Opgave 6.5: Længden (i mm) af fortændernes odontoblaster hos to grupper marsvin, de ordnede observationer.

appelsinsaft	kunstigt
8.2	4.2
9.4	5.2
9.6	5.8
9.7	6.4
10.0	7.0
14.5	7.3
15.2	10.1
16.1	11.2
17.6	11.3
21.5	11.5

Opgave 6.5: Kunstigt og naturligt C-vitamin

C-vitamin (ascorbinsyre) findes blandt andet i appelsinsaft, men det kan også fremstilles kunstigt. Spørgsmålet er, om det "levende" C-vitamin fra appelsinsaften virker anderledes end det "døde" fra laboratoriet.

Man inddelte en gruppe på 20 marsvin i to lige store grupper, hvoraf den ene fik appelsinsaft og den anden fik en tilsvarende mængde kunstigt C-vitamin. Efter seks ugers forløb målte man længden af fortændernes odontoblaster (det tandbensdannende væv). Resultaterne er vist i Tabel 6.3. Opgaven er nu at undersøge, om der på baggrund af disse data er en signifikant forskel på de to slags C-vitamins evne til fremme væksten af tandbenet.

Ved et lille kunstgreb kan problemet omformes til et regressionsproblem:

1. Lad $y_1, y_2, \ldots, y_{20}$ betegne de 20 værdier fra Tabel 6.3 og indfør $x_1, x_2, \ldots, x_{20}$ til at identificere de to grupper, sådan at forstå at $x_i = 0$ hvis y_i hører til appelsinsaft-gruppen og $x_i = 1$ hvis y_i hører til den anden gruppe. (Lav evt. et scatterplot af y mod x.)

2. Lav regression af y på x.

3. Hældningskoefficienten $\widehat{\beta}_1$ i dette regressionsproblem er signifikant forskellig fra 0 hvis og kun hvis der er en signifikant forskel på de to grupper. Er der en signifikant forskel?

4. Hvis man i stedet for x-værdierne 0 og 1 benytter f.eks. værdierne -5 og 10, så bør man få den samme t-teststørrelse. Prøv!

— Denne opgave skulle vise hvordan man med regressionsanalyse kan løse det såkaldte "tostikprøveproblem i normalfordelingen (uparrede observationer)"; ved løsningen af dette problem plejer man at foretage et tosidet t-test (med $(n_1-1)+(n_2-1)$ frihedsgrader) på teststørrelsen

$$\frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}};$$

her betegner $\overline{y}_i$, n_i og s_i^2 henholdsvis gennemsnit, antal observationer og variansskøn i gruppe nr. i. Vis at regressionsmetodens teststørrelse kan omskrives til ovenstående udtryk.

I øvrigt findes der en særlig ISP-kommando **student2** som er beregnet til netop tostikprøveproblemer i normalfordelingen.

Opgave 6.6

Man skal måle en ny stamme forsøgsdyrs reaktion på en bestemt behandling. Fra tidligere forsøg vides at måleresultaterne er nogenlunde normalfordelt med en standardafvigelse $\sigma \approx 3$. Man ønsker at bestemme niveauet for den nye stammes reaktion med en middelfejl på 0.5. Hvor mange forsøgsdyr skal man bruge?

Tip: Middelfejlen på gennemsnittet af n observationer fra en normalfordeling med middelværdi μ og varians σ^2 er lig $\sigma/\sqrt{n}$.

Opgave 6.7

Man er ved at planlægge et forsøg der resulterer i et antal sammenhørende værdier (x, y) hvor y formodes at afhænge lineært af x. Man er interesseret i at bestemme hældningen β_1.

Man har mulighed for at foretage forsøget for x-værdierne 0.1, 0.2, 0.5 og 1, og man véd fra tidligere erfaringer at standardafvigelsen på en y-værdi er ca. 2.

Antag at man foretager 5 målinger for hver af de fire x-værdier. Hvor stor bliver middelfejlen på den estimerede hældning $\widehat{\beta}_1$?

Hvis vi siger at det er givet at der kun kan foretages 20 målinger, kunne det så bedre betale sig at bruge dem på en anden måde, f.eks. med 7 målinger for x-værdierne 0.1 og 1 og så kun 3 for hver af de midterste x-er?

Kapitel 7

Flerdimensionale datasæt

Man kan undertiden være i den situation at man for hvert af et antal "individer" (forsøgsdyr, patienter osv.) har værdier af ikke bare to variable, således som det har været tilfældet i de forrige kapitler, men af et større antal. Opgaverne til dette kapitel giver eksempler herpå.

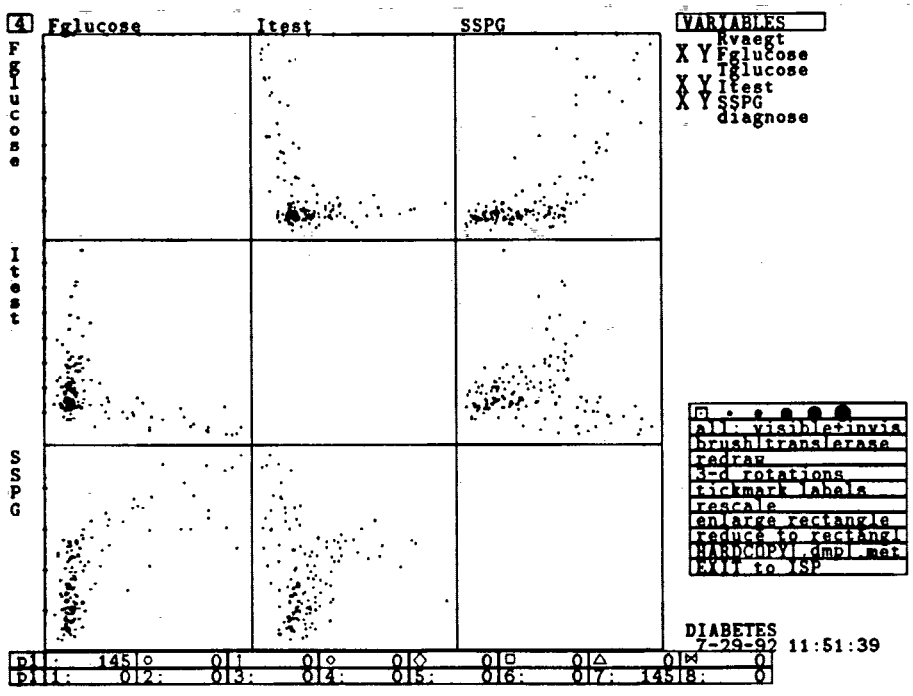
Sålænge man kun har med *to* variable at gøre kan man i et scatterplot afbilde hvert individ (eller rettere: de to oplysninger om individet) som et punkt i et koordinatsystem i planen og på den måde anskueliggøre datamaterialet. Det kan måske også være på fornuftigt at prøve at fitte en ret linie eller en anden kurve til disse punkter, sådan som det er beskrevet i de foregående kapitler.

Med tre eller flere variable er man stedt i en vanskeligere situation, fordi vi lever nu engang i en tredimensional verden og kan derfor kun lave tegninger i to dimensioner (og i én). Hvis man stadig ønsker at lave tegninger — og tegninger *er* altså gode til at formidle talmaterialer — kan man udvælge par af variable og lave et scatterplot for hvert par. Disse scatterplots kan eventuelt stilles op i en *scatterplotmatrix*.

En scatterplotmatrix er en tegning hvor man har sammenstillet et antal scatterplots i et firkantet skema, se Figur 7.1. Alle enkeltplottene i en bestemt række har den samme variabel som y-variabel, og alle enkeltplottene i en bestemt søjle har den samme variabel som x-variabel.

En mere sofistikeret mulighed er at inddrage tiden som en ekstra dimension og derved blive i stand til at visualisere en tredimensional punktsværm. Med ISP's særlige DGS-facilitet[1] kan man se hvordan en tredimensional punktsværm ser ud når den roterer. Selv om man

[1] DGS = Dynamisk GrafikSystem



Figur 7.1: En ISP-scatterplotmatrix over de tre variable **Fglucose**, **Itest** og **SSPG**, jf. Opgave 7.3.
De variable der benyttes som x-variable er mærket **X** i variabel-menuen, de variable der benyttes som y er mærket **Y**.

stadig kun ser den på et todimensionalt billede (computerskærmen), kan man, netop fordi den kan roteres og man kan se den fra forskellige synsvinkler, danne sig et indtryk af den tredimensionale struktur.

7.1 Tredimensionale punktsværme

Hvordan kan nu et sæt tredimensionale observationer se ud, og hvad er det man skal lede efter?

1. For det første kan man jo se efter, om punkterne ligger i én enkelt sværm, eller om der tilsyneladende er tale om flere adskilte punktmængder. Hvis det sidste er tilfældet, tyder det på at datasættet i virkeligheden stammer fra flere væsentligt forskellige populationer.[2]

[2]Hvis de tre betragtede variable er udvalgt blandt mange, skal man naturligvis forvisse sig om at den tilsyneladende opdeling i forskellige dele ikke modsiges af et

2. Videre kan man se efter, om punktsværmen er nogenlunde kugleformet, eller om den snarere er pølseformet. Pølseformen tyder på at der er nogle retninger der er mere interessante end andre, forstået på den måde at der er større variation mellem punkterne i nogle retninger end i andre. Det kan benyttes til at skelne mellem punkter.

3. En punktsværm i tre dimensioner kan godt være stort set todimensional, nemlig hvis den så at sige er presset sammen til noget der næsten er en (plan eller krum) flade. Det betyder altså, at selv om man har tre oplysninger pr. individ, så har man altså stort set kun to stk. forskellige informationer.[3]

Man kan godt lave scatterplotmatricer for mere end tre variable, men dog ikke for alt for mange. Den roterende tredimensionale punktsværm kan man ifølge sagens natur kun have for tre variable ad gangen, så hvis der faktisk er flere end tre variabel at vælge imellem, må man prøve sig frem med de variable man tror er mest interessante eller relevante. Man kan jo også gå systematisk frem og forsøge sig med alle forskellige tre-elementsdelmængder af de variable, men det kan godt være en omstændelig sag.

7.2 Hvordan gør man med ISP

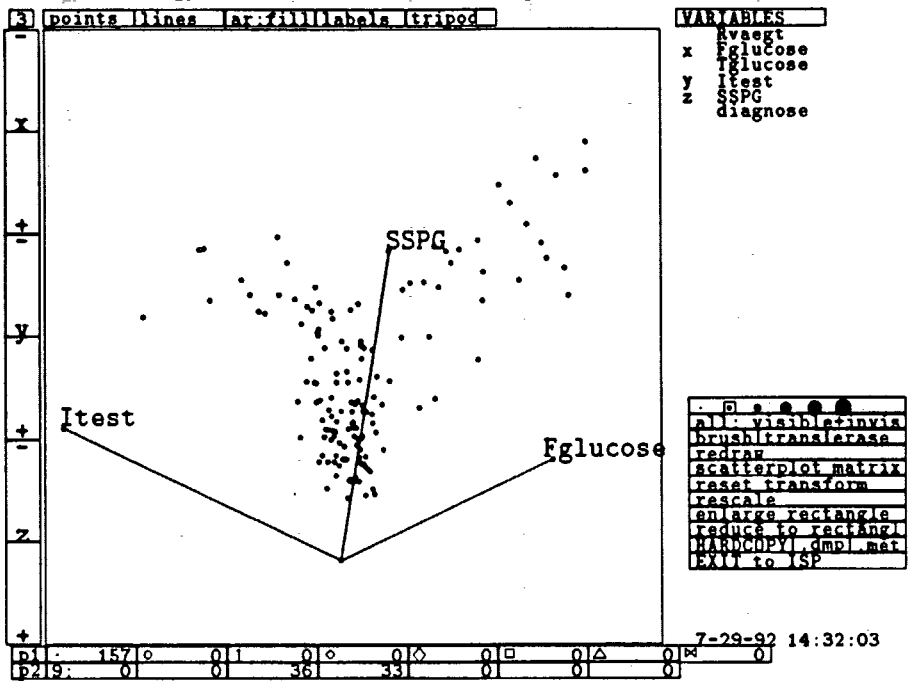
ISP kan takket være **dgs**-kommandoen fremstille scatterplotmatricer og roterende punktsværme. Efter kaldet af **dgs** kan man skifte frem og tilbage mellem de to muligheder ved at trykke **F9**.

Kommandoen **dgs** skal kaldes med mindst ét argument, nemlig et todimensionalt array indeholdende de data der skal visualiseres. Dette array skal have en søjle for hver variabel, og der *skal* være mindst tre søjler; der må højst være 96.

I **dgs**-kaldet kan man give nogle ekstra argumenter som definerer linier, flader, farver, delmængder mm. Det kan godt være lidt vanskeligt at holde styr på, og derfor har ISP's ophavsmænd skrevet en hjælpemakro

andet tripel af variable.

[3]Dette forhold går i stærkt forværret form igen i mangedimensionale datasæt. Det viser sig, at selv om man har målt f.eks. 30 variable pr. individ, så vil den 30-dimensionale punktsværm næsten altid være beliggende i en "flade" af væsentlig mindre dimension, måske 5 eller 6. Man taler om "dimensionernes forbandelse" ('the curse of dimensions').



Figur 7.2: En tredimensional punktsværm vist med dgs. De tre akser går gennem $(0,0,0)$.
Der er tale om det samme datamateriale som i Figur 7.1, se også Opgave 7.3.

`dg_show`, som, forudsat at man navngiver sine variable efter et bestemt system, gør det hele meget nemt. Denne hjælpemakro foreslåes benyttet i opgaverne.

•

Der er en del flere hjælpemakroer til brug i forbindelse med `dgs`, blandt andet en makro `dg_axes` der kan være til hjælp hvis man vil have indtegnet rigtige koordinatakser (og ikke blot den lille tripode), som f.eks. i Figur 7.2 der er fremstillet på følgende måde:

```
ISP>>getdata 'diabetes'
    ...
ISP>>dg_axes 'diab' 'akser'
origin [mean(diab_da)]>array(6)
bottom  [min(diab_da)]>origin
top  [max(diab_da)]>
```

```
blbl [diab_va]>
tlbl [diab_va]>
ISP>>dg_glueo 'diab' 'akser' 'figur' /delete=null
ISP>>dg_show 'figur'
```

Kommentarer: Som `origin` er valgt `array(6)`, dvs. et array med lige så mange nuller som der er søjler i datamatricen. Akserne går fra `bottom` til `top`, og her er `bottom` sat til at være lig `origin`. `blbl` og `tlbl` står for 'bottom labels' og 'top labels' (de labels der hører til aksernes begyndelses- og endepunkt). Man får skrevet labels ved endepunkterne af koordiatakserne (eller ved et hvilket som helst punkt) ved at pege på det med pilen og trykke l (som label); man kan fjerne labelen igen ved at trykke u.

Efter passende rotationer fremkommer Figur 7.2.

7.3 Opgaver

Opgave 7.1: Blodtryk

Ændringer i menneskers livsbetingelser kan give sig udslag i fysiologiske ændringer, eksempelvis i ændret blodtryk.

En gruppe antropologer har undersøgt hvordan blodtrykket ændrer sig hos peruvianske indianere der flyttes fra deres oprindelige primitive samfund i de høje Andesbjerge til den såkaldte civilisation, dvs. storbyen, der i øvrigt ligger i langt mindre højde over havets overflade end deres oprindelig bopæl. Antropologerne udvalgte en stikprøve på 39 mænd over 21 år der havde undergået en sådan flytning. På hver af disse måltes blodtrykket (både det systoliske og det diastoliske) samt en række baggrundsvariable, blandt andet alder, antal år siden flytningen, højde, vægt og puls. Som om det ikke kunne være nok har man udregnet endnu en baggrundsvariabel, nemlig "brøkdel af livet levet i de nye omgivelser", dvs. antal år siden flytning divideret med nuværende alder. Man forestillede sig at denne baggrundsvariabel ville have stor "forklaringsevne".

Her vil vi ikke se på hele talmaterialet, kun på de tre variable *blodtryk* (systolisk), *brøkdel af livet i de nye omgivelser* og *vægt*. Data kan indlæses til ISP med kommandoen `getdata` (data-navn `blodtryk`). Data er endvidere gengivet i Tabel 7.1, hvor man af hensyn til en senere anvendelse har benævnt de tre variable y, x_1 og x_2.

Lav en dels en "roterende punktsværm", dels en scatterplotmatrix over de tre variable. Hvad viser de to slags "afbildninger" om, i hvor høj grad blodtryk kan forudsiges/forklares ud fra vægt og brøkdel af livet i de nye omgivelser?

Opgave 7.2: Lønninger

Den schweiziske *Unionbank* har foretaget en undersøgelse af forskellige erhvervsgruppers løn i forskellige lande.[4] Data til denne opgave omfatter tre erhvervgrupper (produktionschef, folkeskolelærer og sekretær) i 22 storbyer. For hver by og hvert erhverv er angivet bruttoløn (pr. år), nettoløn (dvs. løn efter skat og sociale bidrag) samt lønnens købekraft.

Der er tale om "standardpersoner", som blandt andet ingen børn har (fordi lønnen i nogle lande afhænger af antal børn). Cheferne har en

[4] Den del af undersøgelsen der benyttes i denne opgave er omtalt i *Weekendavisen* d. 9.12.1988 i en artikel som danner baggrund for opgaven.

Tabel 7.1: Opgave 7.1: Værdier af y: systolisk blodtryk (mm Hg), x_1: brøkdel af livet i de nye omgivelser, og x_2: vægt (kg).

y	x_1	x_2	y	x_1	x_2
170	0.048	71.0	114	0.474	59.5
120	0.273	56.5	136	0.289	61.0
125	0.208	56.0	126	0.289	57.0
148	0.042	61.0	124	0.538	57.5
140	0.040	65.0	128	0.615	74.0
106	0.704	62.0	134	0.359	72.0
120	0.179	53.0	112	0.610	62.5
108	0.893	53.0	128	0.780	68.0
124	0.194	65.0	134	0.122	63.4
134	0.406	57.0	128	0.286	68.0
116	0.394	66.5	140	0.581	69.0
114	0.303	59.1	138	0.605	73.0
130	0.441	64.0	118	0.233	64.0
118	0.514	69.5	110	0.432	65.0
138	0.057	64.0	142	0.409	71.0
134	0.333	56.5	134	0.222	60.2
120	0.417	57.0	116	0.021	55.0
120	0.432	55.0	132	0.860	70.0
114	0.459	57.0	152	0.741	87.0
124	0.263	58.0			

teknisk uddannelse, er omkring 40 år og varetager ledelsen af en produktionsafdeling i en større virksomhed i metalindustrien. Lærerne er omkring 35 år og underviser i de offentlige skoler. Sekretærerne er omkring 25 år, taler et fremmedsprog flydende og stenograferer og skriver på maskine.

Der er data for disse 22 byer: Luxembourg, Hong Kong, Genève, New York, Bruxelles, København, Düsseldorff, Amsterdam, Sydney, Wien, Tokyo, Totonto, London, Singapore, Dublin, Paris, Milano, Helsinki, Stockholm, Oslo, Madrid, Lissabon.

Data kan indlæses til ISP med kommandoen **getdata** (data-navn `l0nning`) (bemærk at tegnet `0` mellem `l` og `n` er et nul). Lønningerne er angivet i £. Data bliver anbragt i en 22×9-matrix, hvor søjlerne er

1. chefers bruttoløn,
2. chefers nettoløn,
3. chefers købekraft,
4. læreres bruttoløn,
5. læreres nettoløn,
6. læreres købekraft,
7. sekretærers bruttoløn,
8. sekretærers nettoløn,
9. sekretærers købekraft.

Benyt DGS til at få et indtryk af disse data.

- Se f.eks. på de tre sæt bruttolønninger. Hvilke byer falder uden for det almindelige mønster, og hvordan?

  Hvordan ser det ud med nettolønningerne? og med købekraften?

- Se f.eks. på de tre sæt "lønninger" for et bestemt af de tre erhverv. Hvilke byer falder her uden for det almindelige mønster?

Det "almindelige mønster" bestemmes som den centrale del af punktsværmen. Drej punktsværmen for at finde ud af om den er "flad" set fra visse retninger.

Opgave 7.3: Diabetes

Sukkersyge kan optræde i forskellige former. I en undersøgelse[5] har man foretaget et forsøg på 145 voksne ikke-overvægtige personer hvoraf en del har sukkersyge. Personerne skal først faste en bestemt periode, derefter foretager man en glukosetolerancetest (dvs. de får lov at spise noget sukker, en konstant mængde pr. minut, og så ser man hvad der sker).

For hver person foreligger følgende oplysninger:

1. Relativ vægt (i forhold til en "normalperson").
2. Glukoseindhold i blodet i en periode umiddelbart forud for glukosetolerancetesten.
3. Glukoseindhold i blodet i testperioden.
4. Insulinindhold i blodet i testperioden.
5. SSPG-niveau (Steady State Plasma Glucose), som er et mål for insulin-resistansen og som bestemmes efter kemisk undertrykkelse af den endogene insulinsekretion.
6. Den kliniske klassifikation af personen som værende
   (a) åbenbart nonketotisk diabetiker,
   (b) kemisk subklinisk diabetiker,
   (c) normal (i denne sammenhæng).

Data (der er for omfattende til at gengives her) kan indlæses til ISP med kommandoen `getdata` (data-navn `diabetes`).

Benyt DGS til at anskueliggøre og undersøge dette datasæt. Undersøg blandt andet

- om der er stor sammenhæng mellem nogle af variablene,
- om der er nogle variabelsæt (på tre) der giver anledning til punktkonfigurationer hvor de tre persongrupper ligger nogenlunde afgrænset fra hinanden?

Tip: Giv f.eks. de tre persongrupper hver sin farve.

[5] Reaven & Miller (1979): An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia* **16**, 17–24.

Opgave 7.4: Galápagos-øernes flora

Galápagos-øerne er et eldorado for de biologer der undersøger arternes oprindelse, udvikling og overlevelse og de faktorer der spiller ind herpå. I undersøgelser af hvilke geografiske faktorer der især bestemmer antallet af plantearter på de enkelte øer, har man[6] for hver af de 30 øer indsamlet følgende oplysninger:

1. Antal plantearter.
2. Antal plantearter som er specifikke for Galápagos-området.
3. Øens areal. — En stor ø vil typisk have flere forskelligartede områder end en lille ø, og derfor må man forvente at finde flere arter på den store ø.
4. Øens højde over havets overflade. — På en høj ø kan der være flere forskelligartede områder end på en lav ø, og derfor flere arter; de høje af Galápagos-øerne får betydelig mere nedbør end de lave.
5. Afstanden til nærmeste naboø. — Planterne kan lettere spredes til en nærliggende ø end til en fjern.
6. Afstanden til Santa Cruz, der regnes for centrum i arkipelaget.
7. Arealet af den nærmeste naboø.

Data er vist i Tabel 7.2, og de kan indlæses til ISP med kommandoen `getdata` (data-navn `galapago`).

Opgaven er nu at undersøge, hvilke faktorer der især er bestemmende for artsantallet. (Egentlig er der jo to forskellige artsantal, som man måske burde undersøge hver for sig.)

Som det ses varierer tallene for hver enkelt variabel mange størrelsesordener (det gælder især Areal, der varierer fra 0.01 km^2 til 4669.32 km^2, dvs. fem størrelsesordener). I sådanne situationer vil man ofte tage logaritmen til tallene, fordi man derved får data der ikke varierer over så mange størrelsesordener. Man skal især føle sig fristet til tage logaritmen, hvis de rå data udviser sammenhænge som ikke synes at være lineære.

[6] Johnson & Raven (1973): Species number and endemism. The Galápagos Archipelago revisited. *Science* **179**, 893–5. Grand, Price & Snell (1980): The exploration of Isla Daphne Minor. *Noticias de Galápagos* **31**, 22–7.

Tabel 7.2: Opgave 7.4: Galápagos-øernes planteartsrigdom og geografi. A: antal arter, B: antal specifikke arter, C: areal (km^2), D: højde (m), E: afstand (km) til nærmeste naboø, F: afstand (km) til Santa Cruz, G: areal (km^2) af nærmeste naboø.

navn	A	B	C	D	E	F	G
Baltra	58	23	25.09	–	0.6	0.6	1.84
Bartolomé	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamaño	2	1	0.05	–	1.9	1.9	903.82
Daphne Major	18	11	0.34	119	8.0	8.0	1.84
Daphne Minor	24	–	0.08	93	6.0	12.0	0.34
Darwin	10	7	2.33	168	34.1	290.2	2.85
Eden	8	4	0.03	–	0.4	0.4	17.95
Enderby	2	2	0.18	112	2.6	50.2	0.10
Española	97	26	58.27	198	1.1	88.3	0.57
Fernandina	93	35	634.49	1494	4.3	95.3	4669.32
Gardner øst	58	17	0.57	49	1.1	93.1	58.27
Gardner vest	5	4	0.78	227	4.6	62.2	0.21
Genovesa	40	19	17.35	76	47.4	92.2	129.49
Isabela	347	89	4669.32	1707	0.7	28.1	634.49
Marchena	51	23	129.49	343	29.1	85.9	59.56
Onslow	2	2	0.01	25	3.3	45.9	0.10
Pinta	104	37	59.56	777	29.1	119.6	129.49
Pinzón	108	33	17.95	458	10.7	10.7	0.03
Las Plazas	12	9	0.23	–	0.5	0.6	25.09
Rábida	70	30	4.89	367	4.4	24.4	572.33
San Cristóbal	280	65	551.62	716	45.2	66.6	0.57
San Salvador	237	81	572.33	906	0.2	19.8	4.89
Santa Cruz	444	95	903.82	864	0.6	0.0	0.52
Santa Fé	62	28	24.08	259	16.5	16.5	0.52
Santa María	285	73	170.92	640	2.6	49.2	0.10
Seymour	44	16	1.84	–	0.6	9.6	25.09
Tortuga	16	8	1.24	186	6.8	50.9	17.95
Wolf	21	12	2.85	253	34.1	254.7	2.33

Kapitel 8

Multipel lineær regression

Som omtalt i tidligere kapitler går regressionsanalyse ud på at beskrive en y-variabel ved hjælp af nogle baggrundsvariable efter devisen

$$y = \widehat{y} + \text{residual}$$

hvor den fittede værdi $\widehat{y}$ afhænger af baggrundsvariablene.

Hidtil har vi behandlet modeller hvor der var én baggrundsvariabel x og hvor det var muligt at *tegne* dels punkterne (x, y), dels den fittede kurve. Men man kan naturligvis også komme ud for at der er mere end én baggrundsvariabel inde i billedet. I den generelle behandling i dette kapitel er der p baggrundsvariable.

I *multipel lineær regressionsanalyse* søger man at beskrive y ud fra baggrundsvariablene $x_1, x_2, \ldots, x_p$ ved et udtryk af formen

$$\beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p \tag{8.1}$$

hvor β-erne er ukendte parametre der skal bestemmes således at modellen passer bedst muligt. Undertiden indfører man en ekstra baggrundsvariabel x_0 som altid har værdien 1, for så kan man skrive (8.1) som

$x_0\beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p$ eller $\sum_{j=0}^{p} x_j\beta_j$.

Multipel lineær regression omfatter som specialtilfælde de regressionsmodeller der har været omtalt i tidligere kapitler:

- Simpel lineær regression svarer til at $p = 1$ og at baggrundsvariablen x_1 blot hedder x.

- Kvadratisk regression (modellen (3.1)) svarer til at $p = 2$ og at baggrundsvariablen x_1 hedder x og at baggrundsvariablen x_2 er lig med x^2.

- Trigonometrisk regression (modellen (3.3)) svarer til at $p = 2$ og at baggrundsvariablene x_1 og x_2 beregnes ud fra den "rigtige" baggrundsvaribel t ved relationerne $x_1 = \cos(2\pi ft)$ og $x_2 = \sin(2\pi ft)$.

8.1 Modellen

Vi vil gå ud fra, at talmaterialet udgøres af n talsæt som hver især består af én y-værdi og p x-værdier. Med y_i betegnes y-værdien hørende til det i-te individ, og med x_{ij} betegnes den j-te baggrundsvariabels værdi hos det i-te individ. Skematisk ser det sådan ud:

individ	observation	baggrundsvariable			
1	y_1	x_{11}	x_{12}	$\dots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\dots$	x_{np}

Derudover er der den underforståede baggrundsvariabel x_0 som har værdien 1 for alle i. Den multiple regressionsmodel går da ud på at skrive y-erne på formen

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

eller

$$y_i = \sum_{j=0}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{8.2}$$

Påstanden er altså, at der findes parameterværdier $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ således at (8.2) er opfyldt — når det samtidig om residualerne $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ kræves at de skal kunne opfattes som tilfældige tal fra en fordeling der er nogenlunde koncentreret omkring 0. I den *statistiske* model for multipel lineær regressionsanalyse plejer man at gå ud fra at $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er tilfældige tal fra en og samme normalfordeling med middelværdi 0 og varians σ^2.

8.2 Estimation af parametrene

Som estimationsmetode benyttes stadig mindste kvadraters metode. Det betyder at estimaterne $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ skal bestemmes således at summen af kvadratiske afvigelser

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2$$

bliver mindst mulig. Man kan bevise (se Afsnit 8.8 side 112), at man kan finde disse estimater ved at løse $p+1$ lineære ligninger, de såkaldte *estimationsligninger*, med $p+1$ ubekendte. Med betegnelsen

$$\widehat{y}_i = \sum_{j=0}^{p} x_{ij}\widehat{\beta}_j$$

kan disse ligninger skrives

$$\left.\begin{aligned}
\sum_{i=1}^{n} x_{i0}\widehat{y}_i &= \sum_{i=1}^{n} x_{i0}y_i \\
\sum_{i=1}^{n} x_{i1}\widehat{y}_i &= \sum_{i=1}^{n} x_{i1}y_i \\
\sum_{i=1}^{n} x_{i2}\widehat{y}_i &= \sum_{i=1}^{n} x_{i2}y_i \\
&\vdots \\
\sum_{i=1}^{n} x_{ip}\widehat{y}_i &= \sum_{i=1}^{n} x_{ip}y_i,
\end{aligned}\right\} \qquad (8.3)$$

der også (med en lidt forenklet notation) kan skrives som

$$\begin{aligned}
\left(\sum x_0x_0\right)\widehat{\beta}_0 + \left(\sum x_0x_1\right)\widehat{\beta}_1 + \ldots + \left(\sum x_0x_p\right)\widehat{\beta}_p &= \sum x_0y \\
\left(\sum x_1x_0\right)\widehat{\beta}_0 + \left(\sum x_1x_1\right)\widehat{\beta}_1 + \ldots + \left(\sum x_1x_p\right)\widehat{\beta}_p &= \sum x_1y \\
\left(\sum x_2x_0\right)\widehat{\beta}_0 + \left(\sum x_2x_1\right)\widehat{\beta}_1 + \ldots + \left(\sum x_2x_p\right)\widehat{\beta}_p &= \sum x_2y \\
&\vdots \\
\left(\sum x_px_0\right)\widehat{\beta}_0 + \left(\sum x_px_1\right)\widehat{\beta}_1 + \ldots + \left(\sum x_px_p\right)\widehat{\beta}_p &= \sum x_py
\end{aligned}$$

(Her er alle x_{i0}-erne som nævnt lig 1.) Ligningerne har "som oftest" præcis én løsning. Undertiden er der uendelig mange løsninger; det er tilfældet hvis en af baggrundsvariablene er overflødig i den forstand at den ikke indeholder anden information end hvad der allerede er indeholdt i de øvrige.[1] I sådanne situationer plejer man at fjerne den eller de overflødige variable.

Variansparameteren σ^2 estimeres ved s^2, som er residualkvadratsummen divideret med antallet af frihedsgrader:

$$s^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 . \tag{8.4}$$

Bemærkning

I nogle situationer ønsker man ikke at have konstantleddet β_0 med i regressionsligningen. I så fald skal man udelade den første af estimationsligningerne samt omdefinere $\widehat{y}_i$ til $\sum_{j=1}^{p} x_{ij}\widehat{\beta}_j$. Endvidere bliver antallet af frihedsgrader for s^2 ikke $n-(p+1)$ men $n-p$.

8.3 Udvælgelse af baggrundsvariable

I første omgang kunne man måske fristes til at tro at jo flere baggrundsvariable man inddrager, jo bedre. Det er selvfølgelig rigtigt, at jo flere baggrundsvariable man medtager, jo nøjagtigere et fit kan man få, men det er ikke nødvendigvis det der er meningen med at benytte en statistisk model. Formålet med at benytte statistiske modeller er at få en *reduktion* af data, og det vil blandt andet sige at man skal stræbe efter en statistisk model med væsentlig færre baggrundsvariable (og dermed parametre) end antallet af observationer. I det hele taget skal man holde sig det princip efterretteligt som går under navnet *Occam's rageknv* og som siger, at man ikke skal antage eksistensen af flere ting end nødvendigt.

Undertiden har man mange flere baggrundsvariable end man med rimelighed kunne tage med i modellen, og så er man stillet over for den

[1] Ligningerne har en entydig løsning hvis og kun hvis det ikke er muligt at udtrykke nogen af de forklarende variable som en linearkombination af de øvrige.

opgave at udvælge en passende delmængde af dem. Det første kriterium må da være, at man kun bør medtage variable der kan tænkes at have noget at gøre med den y-variabel der er tale om. Derudover skal man have fat i et sæt baggrundsvariable der gør s^2 forholdsvis lille. Bemærk i denne forbindelse, at man i udtrykket for s^2 tager hensyn til antallet af baggrundsvariable (formel (8.4)).

Når man skal afgøre hvilke baggrundsvariable der måske kan undværes, kan man benytte sig af, at man med et t-test for hver enkelt variabel kan vurdere om den tilsvarende parameter er signifikant forskellig fra 0, dvs. om variablen har en signifikant virkning. Antag f.eks. at man har en model med p baggrundsvariable plus en konstant, og at man ønsker at undersøge om variabel nr. k behøver være med i modellen. Så udregner man

$$t = \frac{\widehat{\beta}_k}{\text{estimeret middelfejl på } \widehat{\beta}_k}$$

og sammenholder resultatet med t-fordelingen med $n-(p+1)$ frihedsgrader (= antal frihedsgrader for s^2). Hvis t er tæt på nul vil man acceptere hypotesen om at β_k er nul, og det betyder at man kan se bort fra baggrundsvariabel nr. k og altså gå videre med en reduceret model med kun $p-1$ baggrundsvariable; hvis t er langt fra nul er $\widehat{\beta}_k$ signifikant forskellig fra 0, dvs. baggrundsvariabel nr. k har en signifikant virkning og skal derfor forblive i modellen.

8.4 Modelkontrol

Når der er mere end én baggrundsvariabel, er man beklageligvis afskåret fra at kunne foretage den udmærkede modelkontrol der består i at lave en *tegning* som indeholder dels de punkter (x, y) der repræsenterer de enkelte individer, dels den fittede kurve (rette linie), for så ud fra tegningen at vurdere om de observerede punkter fordeler sig nogenlunde tilfældigt omkring den fittede linie.

Hvis der er *to* baggrundsvariable x_1 og x_2 skal de enkelte individer repræsenteres som punkter (x_1, x_2, y) i det tredimensionale rum, og man har nu ikke længere en regressions*linie*, men en regressions*plan* som punkterne skal fordele sig tilfældigt om. Den fittede regressionsplans ligning er $y = \widehat{\beta}_0 + x_1\widehat{\beta}_1 + x_2\widehat{\beta}_2$. Med ISP/DGS er det muligt at tegne regressionsplanen og de observerede punkter og måske derudfra afgøre

om regressionsplanen giver en rimelig beskrivelse af data. Men det er ikke så let.

Der kommer endnu større problemer når der er flere baggrundsvariable. Hvis man har p baggrundsvariable, så skal de enkelte individer repræsenteres som punkter i det $p+1$-dimensionale rum!

Derfor baserer man i høj grad modelkontrol på undersøgelser af de empiriske residualer $e_i = y_i - \widehat{y}_i$. Man kan udføre grafiske undersøgelser af residualerne blandt andet efter de samme principper som i tilfældet simpel lineær regressionsanalyse, se Afsnit 5.2 side 59 og Opgave 5.2:

1. man kan tegne et indexplot, dvs. punkterne (i, e_i),
2. man kan tegne et plot af residualerne mod de fittede værdier, dvs. punkterne $(\widehat{y}_i, e_i)$,
3. man kan plotte residualerne mod hver af de forklarende variable, dvs. for hvert j tegne punkterne (x_{ij}, e_i),
4. man kan plotte residualerne mod nogle af de baggrundsvariable som man kender, men som man egentlig ikke havde tænkt sig at tage med,
5. man kan tegne et histogram over residualerne,
6. man kan tegne et fraktildiagram over residualerne.

Hver gang kan man undersøge enten de almindelige residualer e_i eller de standardiserede residualer e_i' (defineret i formel (5.3) på side 59).

Variansskønnet s^2

Variansskønnet s^2 fortæller ikke noget om hvor godt modellen passer, kun noget om hvor meget punkterne varierer omkring regressionsfladen, og en sådan variation kan udmærket godt skyldes at der simpelthen er stor tilfældig variation på y-målinger af den slags som man nu har med at gøre.

Derimod kan det undertiden være fornuftigt at benytte størrelsen af s^2 som kriterium når man skal udvælge baggrundsvariable. Hvis der f.eks. er 20 baggrundsvariable at vælge imellem og man har besluttet sig for højst at ville have tre med i sin model, så kan det være fornuftigt at vælge de tre der giver den mindste s^2. Man bør dog også skele til om de tre der derved bliver udvalgt, virker som fornuftige baggrundsvariable i den givne sammenhæng.

Box 8.1: Korrelationskoefficienten

Hvis man har nogle sammenhørende værdier (a_i, c_i), $i = 1, 2, \ldots, n$, af to størrelser a og c, så er den empiriske korrelationskoefficient mellem dem defineret som tallet

$$r = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(c_i - \overline{c})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2}\sqrt{\sum_{i=1}^{n}(c_i - \overline{c})^2}}.$$

Der gælder at $-1 \leq r \leq 1$.

Determinationskoefficienten R^2

Nogle brugere af regressionsanalyse er meget begejstrede for den såkaldte *determinationskoefficient* R^2 eller *kvadratet på den multiple korrelationskoefficient*. NB! Den benyttes kun når der er et konstantled med i regressionen.

Man kan udregne R^2 efter en af følgende to formler[2]:

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{8.5}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2 \cdot \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}. \tag{8.6}$$

Formel (8.6) fortæller at R^2 er kvadratet på korrelationskoefficienten mellem de observerede og de fittede værdier.

Formel (8.5) fortæller at R^2 er et udtryk for, hvor stor en del af den samlede variation omkring totalgennemsnittet der beskrives af model-

[2]Det er ikke umiddelbart indlysende, men dog rigtigt, at de to udtryk giver samme resultat.

len. — Der er dem der mener, at R^2 derfor også er et udtryk for hvor godt modellen passer, men prøv så at vende tilbage til Opgave 2.4!

8.5 Hvordan gør man med ISP

ISP's **regress**-kommando kan uden videre udføre multipel lineær regression. Kommandoen benyttes på samme måde som hidtil, dvs. det typiske kald ser sådan ud

```
ISP>>regress x y
```

hvor y er en vektor af længde n og x er en $n \times p$-matrix.[3] Derudover kan man tilføje output-argumenter til blandt andet residualer, parameterestimater og middelfejl, se *Introduktion til ISP* og/eller online-hjælpen.

Her er en omtale af de forskellige dele af udskriften fra **regress**:

- Parameterestimaterne står i søjlen **coef** og deres estimerede middelfejl i søjlen **sdev**. Rækkefølgen af variablene i udskriften er den samme som rækkefølgen af søjlerne i x-matricen (konstantleddet kommer altid sidst).

- Den estimerede standardafvigelse s, dvs. kvadratroden af variansskønnet s^2 (formel (8.4)), udskrives under betegnelsen **sigma**.

- Hvis der er et konstantled med i regressionen, beregnes R^2 automatisk og udskrives under betegnelsen **R-square**.

- Størrelsen **F-stat** i **regress**-udskriften er den såkaldte F-teststørrelse for test af hypotesen om at alle de forklarende variable kan udelades, dvs. hypotesen $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. Store F-værdier er signifikante; der gælder altid at $F > 0$.

  For at afgøre om en F-værdi er signifikant stor udregnes testsandsynligheden, dvs. sandsynligheden for at få en større F-værdi end den observerede, under forudsætning af at hypotesen er rigtig. Hertil benyttes den såkaldte F-fordeling; denne har to frihedsgradsantal, et fra tælleren og et fra nævneren; frihedsgradsantallene står anført i **regress**-udskriften (og er $p-1$ og $n-p$).

  ISP-funktionen **fpr()** udregner testsandsynligheden, typisk

[3] Det er i øvrigt også tilladt at nøjes med ét inputargument til **regress**, f.eks. **regress data**; så opfatter ISP den sidste søjle i **data** som y og de øvrige som x.

```
ISP>>testssh = 1-fpr( F, tællerfrgr, nævnerfrgr )
```

Det er meget let at tage baggrundsvariable ind og ud af modellen. Hvis man f.eks. kun ønsker at have variablene nr. 1, 3, 5 og 7 med, kan man gøre det med `regress (x(*,1 3 5 7)) y` .

Bemærk at konstantleddet udelades fra analysen hvis man tilføjer `/const=n` til `regress`-kommandoen.

Hvis en eller flere af baggrundsvariablene er overflødige, således at estimationsligningerne ikke har en entydig løsning (jf. side 96), kommer der en advarsel fra `regress` (`"matrix is numerically singular!"`).

8.6 Ensidet variansanalyse

Indtil nu har baggrundsvariablene for det meste fremstået som kvantitative størrelser, men der er intet til hinder for at de kan være 01-størrelser der angiver om den tilsvarende y-værdi tilhører en bestemt gruppe eller ej. Hvis $y_1, y_2, \ldots, y_n$ er observationer der hver især tilhører netop en af k forskellige grupper, så kan man "kode" dette på den måde at man indfører baggrundsvariable $x_1, x_2, \ldots, x_p$ således at $x_{ij} = 1$ hvis y_i tilhører gruppe nr. j og 0 ellers.

Vi vil illustrere metoden med et eksempel som handler om "dækningsgrader for Fuglegræs".[4] Fuglegræs er en plante der i kornmarker betragtes som ukrudt; man har behandlet et antal marker på forskellig måde (idet man har bortluget forskellige andre ukrudtsarter) og så registreret dækningsgraden for Fuglegræs; dækningsgraden fortæller noget om antal planteindivider pr. arealenhed. Der er fire forskellige markbehandlinger, dvs. fire grupper, og fire observationer pr. gruppe. Resultaterne er vist i Tabel 8.1.

Vi kalder dækningsgraden for y og indfører fire forklarende variable x_1, x_2, x_3 og x_4 der skal angive medlemsskab i hver af de fire grupper. Derved fås Tabel 8.2.

På grund af valget af x-erne siger den generelle regressionsligning

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 \qquad (8.7)$$

[4] Datamaterialet stammer fra A. Greenfort, C.S.F. Jensen & S. Jeppesen (1987): *Planter og planter imellem.* Projektrapport på BIO-OB, RUC. (Eksemplet er også behandlet i Kapitel 11 i J. Larsen (1989): *Basisstatistik*. IMFUFA-tekst 167, RUC.)

Tabel 8.1: Dækningsgrader for Fuglegræs, 1.

gruppe	dækningsgrad			
1	17	38	23	26
2	19	16	16	14
3	25	33	29	33
4	27	16	30	20

Tabel 8.2: Dækningsgrader for fuglegræs, 2.

gruppe	y	x_1	x_2	x_3	x_4
1	17	1	0	0	0
1	38	1	0	0	0
1	23	1	0	0	0
1	26	1	0	0	0
2	19	0	1	0	0
2	16	0	1	0	0
2	16	0	1	0	0
2	14	0	1	0	0
3	25	0	0	1	0
3	33	0	0	1	0
3	29	0	0	1	0
3	33	0	0	1	0
4	27	0	0	0	1
4	16	0	0	0	1
4	30	0	0	0	1
4	20	0	0	0	1

da blot, at $y = \beta_j$ når y tilhører gruppe nr. j. Ved at foretage multipel lineær regression (uden konstantled) af y på x_1, x_2, x_3, x_4 får man derfor beregnet $\widehat{\beta}_j$-erne, dvs. gruppegennemsnittene.

Vi kan imidlertid formulere den samme model på andre måder, f.eks.

$$y = \beta_0 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 \tag{8.8}$$

hvor vi nu har udeladt den første forklarende variabel og til gengæld tilføjet et konstantled β_0. Denne gang gælder, at i gruppe 1 er $y = \beta_0$, i gruppe 2 $y = \beta_0 + \beta_2$, i gruppe 3 $y = \beta_0 + \beta_3$ og i gruppe 4 $y = \beta_0 + \beta_4$. Grunden til at denne formulering af modellen kan være interessant er, at hypotesen om at alle grupper er ens nu kan formuleres på den måde, at alle de (tre) forklarende variable kan undværes, og det kan testes ved brug af den F-størrelse som **regress** alligevel regner ud.

Med ISP kan man bære sig ad på følgende måde. Først indlæses observationerne (y-værdierne) og gruppenumrene:

```
ISP>>input/dims=16 > dgrad
row 1,1:16>17 38 23 26 19 16 16 14
row 1,9:16>25 33 29 33 27 16 30 20
ISP>>glue 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 > gruppe
```

I stedet for at indtaste de $16 \times 4 = 64$ x-værdier kan man få ISP til at udregne dem:

```
ISP>>x = gruppe == iota(array(1,4))
ISP>>print x
  1         0         0         0
  1         0         0         0
  1         0         0         0
  1         0         0         0
  0         1         0         0
  0         1         0         0
  0         1         0         0
  0         1         0         0
  0         0         1         0
  0         0         1         0
  0         0         1         0
  0         0         1         0
  0         0         0         1
  0         0         0         1
  0         0         0         1
  0         0         0         1
```

Vi kan nu estimere modellen (8.7):

```
ISP>>regress /const=n x dgrad
degrees of freedom:     16 - 4 = 12
sigma     =   5.870
condition =   1.000

 var          coef          sdev
   1           26.00         2.935
   2           16.25         2.935
   3           30.00         2.935
   4           23.25         2.935
```

De fundne koefficienter er de fire gruppegennemsnit. Størrelsen `sigma` er et estimat over y-ernes fælles standardafvigelse; den tilsvarende estimerede varians `sigma`^{2} ($= 5.87^2$) siges at beskrive *variationen inden for grupper*, fordi den beregnes ud fra y-ernes afvigelser fra deres gruppegennemsnit.

For at estimere modellen (8.8) kan vi gøre sådan:

```
ISP>>regress (x(*,2:4)) dgrad
degrees of freedom:     16 - 4 = 12
sigma     =   5.870
R-square  =  0.4931
F-stat    =   3.891       (3 over 12 df)
condition =   3.732

 var          coef          sdev
   1          -9.750         4.151
   2           4.000         4.151
   3          -2.750         4.151
   const       26.00         2.935
```

Da der er tale om den samme model i en anden parametrisering, er `sigma` og antal frihedsgrader som før. Konstantleddet er gennemsnittet i gruppe 1, gennemsnittet i gruppe 2 er $26 - 9.75 = 16.25$, i gruppe 3 $26 + 4 = 30$ og i gruppe 4 $26 - 2.75 = 23.25$. Den væsentligste grund til at estimere modellen på denne måde er, at vi får udregnet F-teststørrelsen for at teste hypotesen om at alle de *tre* koefficienter er 0, svarende til at de fire grupper har samme middelværdi. F-størrelsen er 3.891 med 3 og 12 frihedsgrader, og den tilsvarende testsandsynlighed findes sådan:

```
ISP>>print (1-fpr( 3.891, 3, 12))
 0.3736E-01
```

Testsandsynligheden er altså ca. 3.7%, og det vil man som regel tage som tegn på at koefficienterne *er* signifikant forskellige fra 0, altså at grupperne *er* signifikant forskellige.

•

Det turde være klart, at den fremgangsmåde der blev benyttet i fuglegræseksemplet kan benyttes generelt i situationer hvor man ønsker at sammenligne et antal grupper (af normalfordelte observationer) for at afgøre om grupperne er ens eller ej.[5]

Antag at der er k grupper, og lad os et øjeblik lave lidt om på notationen: Vi indicerer y-erne på den måde at y_{ij} skal betegne observation nr. j i gruppe nr. i. Lad $\overline{y}$ betegne gennemsnittet af alle y-erne, lad $\overline{y}_i$ betegne gennemsnittet af y-erne i gruppe i, og lad n_i betegne antal observationer i gruppe i. Så gælder at den ovenfor omtalte F-størrelse kan udregnes som

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}.$$

Tælleren i denne brøk siges at beskrive *variationen mellem grupper* og nævneren siges at beskrive *variationen inden for grupper*. F-testet sammenligner altså disse to variationer, og det er grunden til at metoden kaldes *variansanalyse*. At der er tale om *ensidet* variansanalyse kommer af at observationerne er klassificeret efter ét inddelingskriterium (gruppenummer).

Det kan tilføjes, at der er en særlig ISP-kommando `anova1` der udfører ensidet variansanalyse næsten af sig selv.

[5] I øvrigt handler Opgave 6.5 om det specialtilfælde hvor der er to grupper der skal sammenlignes.

Tabel 8.3: Standardisering af hormonpræparat: Sammenhørende værdier af logaritmen u til dosis og restlevetiden y, dels ved produktionens start, dels efter halvandet års forløb ("slut").

Start		Slut	
u	y	u	y
3.20	8.70	3.08	6.14
3.35	6.20	3.27	4.80
3.46	8.22	3.56	4.82
3.72	2.94		
3.84	3.88		

8.7 Sammenligning af regressionslinier

Multipel lineær regression er er et generelt redskab der kan specialiseres til mange forskellige slags situationer. I Afsnit 8.6 så vi et simpelt eksempel herpå, idet det drejede sig om at sammenligne nogle grupper af observationer for at undersøge om grupperne har samme middelværdier.

I dette afsnit vil vi studere en lidt mere avanceret situation, nemlig den hvor man først estimerer regressionslinier til et antal forskellige datasæt og hvor man derefter gerne vil sammenligne disse regressionslinier, mere præcist kan man være interesseret i at undersøge om regressionslinierne er parallelle og/eller om de er sammenfaldende. Vi vil gøre det ved hjælp af et eksempel med *to* regressionslinier, men metoden kan uden videre benyttes også for mere end to linier.

Eksemplet handler om standardisering af et vist hormonpræparat. Præparatets virkning bestemmes ved at man giver nogle mus det i forskellige koncentrationer og så måler musens "reaktion", nemlig tiden indtil musen dør. Da produktionen af præparatet indledtes, foretog man fem sådanne forsøg for at fastlægge en standard. Efter halvandet års forløb foretog man tre forsøg for at undersøge om standarden havde ændret sig. Resultaterne er vist i Tabel 8.3.

Erfaringen viser, at i den slags situationer er y-værdierne normalfordelte med en middelværdi der afhænger lineært af log-dosis u og med samme varians. Man kunne så foretage simpel lineær regression to gange, en for hvert datasæt, men det gør vi ikke. Af hensyn til det videre modelarbejde er vi nødt til at putte alle observationerne ind i

Tabel 8.4: Standardisering af hormonpræparat: Sammenhørende værdier af gruppenummer (0 ~ start, 1 ~ slut), restlevetid y og forklarende variable x_1, x_2, x_3 og x_4 svarende til to forskellige regressionslinier.

gruppe	y	x_1	x_2	x_3	x_4
0	8.70	1	0	3.20	0
0	6.20	1	0	3.35	0
0	8.22	1	0	3.46	0
0	2.94	1	0	3.72	0
0	3.88	1	0	3.84	0
1	6.14	0	1	0	3.08
1	4.80	0	1	0	3.27
1	4.82	0	1	0	3.56

én stor model med i første omgang fire forklarende variable som vist i Tabel 8.4.

På grund af x-matricens indretning er det sådan at den generelle regressionsligning

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 \tag{8.9}$$

har udseendet $y = \beta_1 + x_3\beta_3$, dvs. $y = \beta_1 + u\beta_3$, når vi befinder os i gruppe 0 (start), og udseendet $y = \beta_2 + x_4\beta_4$, dvs. $y = \beta_2 + u\beta_4$, når vi befinder os i gruppe 1 (slut). Ved at estimere modellen (8.9) får vi derfor på én gang skæring og hældning i startgruppen (β_1 og β_3) og i slutgruppen (β_2 og β_4).

Med ISP kan man bære sig ad på følgende måde:

```
ISP>>input/dims=8,3 > data
row 1,1:3>0 3.20 8.70
row 2,1:3>0 3.35 6.20
row 3,1:3>0 3.46 8.22
row 4,1:3>0 3.72 2.94
row 5,1:3>0 3.84 3.88
row 6,1:3>1 3.08 6.14
row 7,1:3>1 3.27 4.80
row 8,1:3>1 3.56 4.82
ISP>>g=data(*,1)      # gruppe
ISP>>u=data(*,2)      # log dosis
ISP>>y=data(*,3)      # y-værdien
```

```
ISP>># Model: to forskellige rette linier
ISP>># lav en x-matrix:
ISP>>glue/axis=2  0 1 > gv  #  de mulige værdier af G
ISP>>x = g==gv
ISP>>print x
  1          0
  1          0
  1          0
  1          0
  1          0
  0          1
  0          1
  0          1
ISP>>glue/axis=2 x (x*u) > xx # dette er den ønskede matrix!
ISP>>print xx
  1          0          3.200      0
  1          0          3.350      0
  1          0          3.460      0
  1          0          3.720      0
  1          0          3.840      0
  0          1          0          3.080
  0          1          0          3.270
  0          1          0          3.560
ISP>>regress/const=n xx y
degrees of freedom:     8 - 4 = 4
sigma      =    1.332
condition =    43.10

 var          coef          sdev
   1           35.43         8.913
   2           13.57         12.90
   3          -8.379         2.531
   4          -2.517         3.897
```

Det ses at den estimerede linie ved starten af produktionen er $y = 35.43 - 8.379u$ og efter halvandet års forløb $y = 13.57 - 2.517u$. Den estimerede standardafvigelse på y er 1.332 med 4 frihedsgrader.

De to linier synes at være temmelig forskellige, men på den anden side har de estimerede koefficienter temmelig store middelfejl, så vi fortsætter med at undersøge om linierne kan antages at være parallelle, dvs. vi tester hypotesen $\beta_3 = \beta_4$.

Tabel 8.5: Standardisering af hormonpræparat: Sammenhørende værdier af gruppenummer (0 ~ start, 1 ~ slut), restlevetid y og forklarende variable x_1, x_2 og x_3 svarende til to parallelle regressionslinier.

gruppe	y	x_1	x_2	x_3
0	8.70	1	0	3.20
0	6.20	1	0	3.35
0	8.22	1	0	3.46
0	2.94	1	0	3.72
0	3.88	1	0	3.84
1	6.14	0	1	3.08
1	4.80	0	1	3.27
1	4.82	0	1	3.56

For at estimere den fælles hældning og de to forskellige skæringer udfører vi multipel regression på det datasæt der er vist i Tabel 8.5.

Denne gang får den generelle regressionsligning, som nu er

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3, \tag{8.10}$$

udseendet $y = \beta_1 + x_3\beta_3$, dvs. $y = \beta_1 + u\beta_3$, når vi befinder os i gruppe 0 (start), og udseendet $y = \beta_2 + x_4\beta_3$, dvs. $y = \beta_2 + u\beta_3$, når vi befinder os i gruppe 1 (slut). Parameteren β_3 er således den fælles hældning og β_1 og β_2 de to skæringer.

Med ISP kan vi fortsætte således:

```
ISP>># Modellen med parallelle linier:
ISP>>glue/axis=2 x u > xxx
ISP>>print xxx
  1          0          3.200
  1          0          3.350
  1          0          3.460
  1          0          3.720
  1          0          3.840
  0          1          3.080
  0          1          3.270
  0          1          3.560
ISP>>regress/const=n xxx y
degrees of freedom:     8 - 3 = 5
sigma     =   1.409
```

```
condition =   32.39

 var          coef          sdev
   1          29.32         7.912
   2          27.19         7.459
   3         -6.640         2.245
```

Den fælles hældning estimeres altså til -6.640 (med en middelfejl på 2.245) og de to skæringer til 29.32 og 27.19. Den estimerede standardafvigelse på y er nu 1.409 med 5 frihedsgrader.

Dernæst skal vi teste om vi kan tillade os at godtage den nye model, altså om vi kan tillade os at godtage hypotesen om at regressionslinierne er parallelle. Det gøres på denne måde:

1. Udregn variansestimatet s_G^2 i grundmodellen, og lad f_G være dets frihedsgradsantal.
2. Udregn variansestimatet s_H^2 under hypotesen, og lad f_H være dets frihedsgradsantal.
3. Den teststørrelse der skal benyttes er da

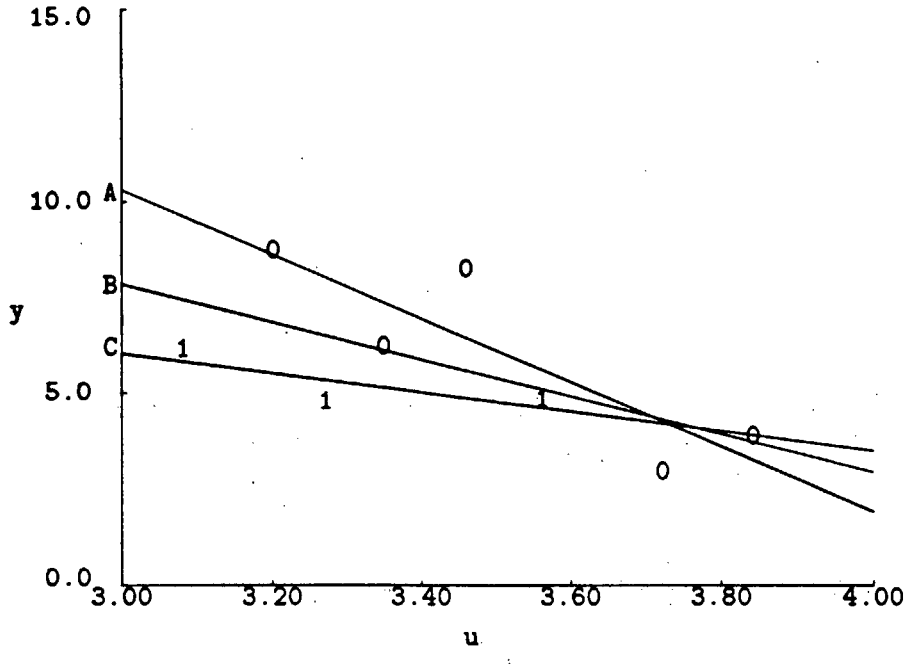
$$F = \frac{(f_H s_H^2 - f_G s_G^2)/(f_H - f_G)}{s_G^2}$$

der altid vil blive et positivt tal. Hvis observationerne beskrives næsten lige så godt af hypotesen som af grundmodellen, så er tælleren tæt på 0, og hvis observationerne beskrives væsentlig dårligere af hypotesen end af grundmodellen så er tælleren temmelig stor (i forhold til nævneren). Med andre ord: store værdier af F fører til at man forkaster hypotesen. Om en observeret værdi af F er stor eller ej afgøres ved at sammenligne den med F-fordelingen med frihedsgradantal $f_H - f_G$ og f_G.

I eksemplet forløber det således ved brug af tallene fra ISP-udskrifterne:

1. $s_G^2 = 1.332^2 = 1.774$ og $f_G = 4$.
2. $s_H^2 = 1.409^2 = 1.985$ og $f_H = 5$.
3. $F = \dfrac{(5 \times 1.985 - 4 \times 1.774)/(5-4)}{1.774} = 2.829/1.774 = 1.59.$

   Med ISP bestemmes testsandsynligheden sådan:



Figur 8.1: Standardisering af hormonpræparat.
De observerede punkter fra hhv. gruppe 0 og 1. Linien mærket A er regressionslinien for gruppe 0, linien mærket C er regressionslinien for gruppe 1, og linien mærket B er den fælles regressionslinie.

```
ISP>>print (1-fpr(1.59, 5-4, 4))
 0.2759
```

Det ses at der er over 27% chance for at få F-værdier som er større end den observerede værdi 1.59, der altså ikke er signifikant. Vi kan således godt tillade os at antage at linierne er parallelle.

Herefter kan vi gå over til at undersøge om de to parallelle linier er sammenfaldende. Hvis de er det, er der tale om en ganske almindelig simpel lineær regressionsmodel $y = \beta_0 + u\beta_1$ der estimeres på sædvanlig måde:

```
ISP>># Modellen med én linie:
ISP>>regress u y
degrees of freedom:     8 - 2 = 6
sigma     =   1.682
R-square  =  0.4005
```

```
F-stat     =   4.008        (1 over 6 df)
condition  =   1.567

 var           coef          sdev
   1          -4.874         2.435
   const       22.45         8.384
```

Denne model skal nu testes som en hypotese i forhold til modellen med parallelle linier. Den nye s_H^2 er $1.682^2 = 2.829$, så teststørrelsen er $F = \frac{(6 \times 2.829 - 5 \times 1.985)/(6-5)}{1.985} = 3.55$ med 1 og 5 frihedsgrader, svarende til en testsandsynlighed på ca. 12%. Det betyder at vi kan tillade os at antage at de to linier er sammenfaldende. — Figur 8.1 viser de observerede punkter og de forskellige estimerede linier.

Man kan naturligvis også teste om modellen med to forskellige linier direkte kan reduceres til at linierne er sammenfaldende. Så bliver F-størrelsen $F = \frac{(6 \times 2.829 - 4 \times 1.774)/(6-4)}{1.774} = 2.78$ med 2 og 4 frihedsgrader, og testsandsynligheden bliver 17.5%.

8.8 Matematiske betragtninger

Dette afsnit uddyber matematikken bag den multiple regressionsanalyse. Blandt andet bevises påstanden om at estimationsligningerne faktisk har mindste kvadraters estimaterne som løsning. Afsnittet er en generalisering af Afsnit 2.1, hvor specialtilfældet $p = 1$ betragtedes.

Sætning 8.1

Mindste kvadraters estimaterne $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ kan bestemmes ved at løse estimationsligningerne (8.3). (Hvis der ikke er noget konstantled med i regressionen, så udelades den første af ligningerne i (8.3) og β_0 og x_{i0} sættes til 0.) Der gælder:

1. *Der findes altid mindst én løsning til estimationsligningerne.*
2. *De fittede værdier $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$ er entydigt bestemt (dvs. selv om der er flere løsninger til estimationsligningerne, så giver de de samme $\widehat{y}_i$-er).*

3. *Et talsæt* $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$ *er løsning til estimationsligningerne, hvis og kun hvis det minimaliserer kvadratsummen*
$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 .$$

Det er ingen selvfølgelighed at et sæt lineære ligninger altid kan løses. Estimationsligningerne har en ganske bestemt struktur, og det er takket være den at de altid kan løses. Det vil vi først se på.

Notationsmæssigt er det noget besværligt, så man kan begynde i det små med tilfældet $p = 2$, svarende til at der er tre ligninger med tre ubekendte:

$$\left(\sum x_0x_0\right)\widehat{\beta}_0 + \left(\sum x_0x_1\right)\widehat{\beta}_1 + \left(\sum x_0x_2\right)\widehat{\beta}_2 = \sum x_0y \quad (8.11)$$

$$\left(\sum x_1x_0\right)\widehat{\beta}_0 + \left(\sum x_1x_1\right)\widehat{\beta}_1 + \left(\sum x_1x_2\right)\widehat{\beta}_2 = \sum x_1y \quad (8.12)$$

$$\left(\sum x_2x_0\right)\widehat{\beta}_0 + \left(\sum x_2x_1\right)\widehat{\beta}_1 + \left(\sum x_2x_2\right)\widehat{\beta}_2 = \sum x_2y \quad (8.13)$$

I mangel af bedre ideer kan man løse den sidste af disse ligninger med hensyn til $\widehat{\beta}_2$ og indsætte resultatet i de to andre. Ligning (8.13) kan omskrives til

$$\widehat{\beta}_2 = \frac{\sum x_2y}{\sum x_2^2} - \frac{\sum x_2x_0}{\sum x_2^2}\widehat{\beta}_0 - \frac{\sum x_2x_1}{\sum x_2^2}\widehat{\beta}_1$$

under forudsætning af at $\sum x_2^2$ ikke er 0, dvs. under forudsætning af at ikke alle x_2-værdierne er 0. Dette udtryk for $\widehat{\beta}_2$ indsættes i ligningerne (8.11) og (8.12), som derved bliver to ligninger med to ubekendte. Ligningernes koefficienter kommer til at se noget voldsomme ud, i den første ligning bliver højresiden

$$\sum x_0y - \frac{(\sum x_0x_2)(\sum x_2y)}{\sum x_2^2}, \quad (8.14)$$

koefficienten til $\widehat{\beta}_0$ bliver (8.14) med y erstattet af x_0, og koefficienten til $\widehat{\beta}_1$ bliver (8.14) med y erstattet af x_1. De tilsvarende udtryk for den anden ligning får man ved at skrive x_1 i stedet for x_0 i (8.14).

Næste skridt er at omskrive alle koefficienterne i de to nye ligninger ved hjælp af formlen i Box 8.2. Eksempelvis omskrives (8.14) på den måde at man som a bruger x_0, som c bruger y og som v tallene $v_i =$

Box 8.2: Endnu en omskrivning af en sum af produkter

Lad $a_1, a_2, \ldots, a_n$ og $c_1, c_2, \ldots, c_n$ være to vilkårlige talsæt, og lad $v_1, v_2, \ldots, v_n$ være et talsæt med den egenskab at $\sum_{i=1}^{n} v_i^2 = 1$.

Sæt $\alpha = \sum_{i=1}^{n} v_i a_i$ og $\gamma = \sum_{i=1}^{n} v_i c_i$. Så gælder at

$$\sum_{i=1}^{n} a_i c_i - \alpha\gamma = \sum_{i=1}^{n} (a_i - \alpha v_i)(c_i - \gamma v_i).$$

(I specialtilfældet $v_i = 1/\sqrt{n}$ får man Box 2.2.)
Det vises ved at gange højresidens parenteser ud.

$x_{i2}/\sqrt{\sum x_2^2}$. Derved kan (8.14) skrives som $\sum x_0^* y_0^*$, dvs. $\sum_{i=1}^{n} x_{i0}^* y_{i0}^*$, hvor

$$x_{i0}^* = x_{i0} - \frac{\sum x_2 x_0}{\sum x_2^2} x_{i2}$$

og

$$y_i^* = y_i - \frac{\sum x_2 y}{\sum x_2^2} x_{i2} .$$

Man kan indføre x_1^* på samme måde som x_0^*, og de to ligninger med to ubekendte kan derefter skrives som

$$\left(\sum x_0^* x_0^*\right) \widehat{\beta}_0 + \left(\sum x_0^* x_1^*\right) \widehat{\beta}_1 = \sum x_0^* y^*$$
$$\left(\sum x_1^* x_0^*\right) \widehat{\beta}_0 + \left(\sum x_1^* x_1^*\right) \widehat{\beta}_1 = \sum x_1^* y^* .$$

Pointen er nu at disse ligninger er opbygget på nøjagtig samme måde som de oprindelige tre, blot er der nu kun to af dem.[6] Det oprindelige problem at løse de tre ligninger kan derfor omformes til et tilsvarende men reduceret problem med kun to ligninger. Det reducerede problem kan nu angribes efter samme metode, og reduceres til én ligning, og den

[6] Nogle vil måske bemærke, at i de oprindelige ligninger var alle x_{i0}-erne lig med 1, hvilket ikke gælder for x_{i0}^*-erne. Dette er rigtigt, men det har ingen betydning eftersom vi ikke i argumentationen har benyttet at x_{i0}-erne havde en speciel værdi.

har en løsning. Altså har de to ligninger også en løsning, og altså har de tre ligninger en løsning.

På ganske tilsvarende måde kan man argumentere for, at et problem med $p+1$ ligninger med $p+1$ ubekendte kan reduceres til et tilsvarende problem med p ligninger med p ubekendte. Alt i alt vil den beskrevne metode altså føre til en løsning.

•

Vi vil nu lade $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$ betegne en eller anden bestemt løsning til estimationsligningerne (8.3). Vi skal benytte følgende opspaltning: For ethvert talsæt $(\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ gælder at

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 \tag{8.15}$$

$$= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} \left(\widehat{y}_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 .$$

Denne formel vises ved at man først omskriver det i-te led i summen på venstre side:

$$\begin{aligned} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 &= \left((y_i - \widehat{y}_i) + \left(\widehat{y}_i - \sum_{j=0}^{p} x_{ij}\beta_j \right) \right)^2 \\ &= (y_i - \widehat{y}_i)^2 + \left(\widehat{y}_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 \\ &\quad + 2(y_i - \widehat{y}_i) \left(\widehat{y}_i - \sum_{j=0}^{p} x_{ij}\beta_j \right) . \end{aligned}$$

For at vise (8.15) er det derfor nok at vise at summen af de dobbelte produkter er 0; men hvis man skriver det andet $\widehat{y}_i$ i det dobbelte produkt som $\sum_{j=0}^{p} x_{ij}\widehat{\beta}_j$, så fås

$$\sum_{i=1}^{n} 2(y_i - \widehat{y}_i) \left(\widehat{y}_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)$$

$$
\begin{aligned}
&= 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)\left(\sum_{j=0}^{p} x_{ij}\widehat{\beta}_j - \sum_{j=0}^{p} x_{ij}\beta_j\right) \\
&= 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)\sum_{j=0}^{p} x_{ij}(\widehat{\beta}_j - \beta_j) \\
&= 2\sum_{j=0}^{p}(\widehat{\beta}_j - \beta_j)\sum_{i=1}^{n} x_{ij}(y_i - \widehat{y}_i) \\
&= 0
\end{aligned}
$$

ifølge estimationsligningerne (8.3). Dermed er opspaltningen (8.15) vist.

På tilsvarende måde som i Afsnit 2.1 følger heraf, at enhver løsning til estimationsligningerne også er et minimumspunkt for kvadratsummen, og at ethvert andet minimumspunkt også opfylder estimationsligningerne.

Om R^2

Vi vil her gøre rede for at de to udtryk (8.5) og (8.6) for R^2 giver samme resultat. Det er nu nødvendigt at der er et konstantled med i regressionen.

Vi omskriver udtrykket fra tælleren af (8.6):

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{y}) &= \sum_{i=1}^{n}\Big((y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})\Big)(\widehat{y}_i - \overline{y}) \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2.
\end{aligned}
$$

Hvis vi kan vise at $\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = 0$, er det klart at de to R^2-udtryk er ens.

I det foregående har vi vist at $\sum_{i=1}^{n}(y_i - \widehat{y}_i)\left(\widehat{y}_i - \sum_{j=0}^{p} x_{ij}\beta_j\right) = 0$ for ethvert valg af β-erne. Hvis vi vælger $\beta_0 = \overline{y}$ og $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, fås heraf det ønskede.

8.9 Opgaver

Opgave 8.1

Betragt følgende lille, til lejligheden konstruerede datasæt:

y	x_1	x_2
−11	−4	7
9	3	4
−4	−1	7
16	7	4
−6	−5	1

Man ønsker at udtrykke y som en linearkombination af de forklarende variable (baggrundsvariablene) x_1 og x_2, dvs. finde en sammenhæng af formen $y = \beta_0 + x_1\beta_1 + x_2\beta_2$.

- Opskriv estimationsligningerne.

  Hvorfor er de forholdsvis nemme at løse?

- Her er et uddrag af en ISP-monitorfil fra en løsning af denne regressionsopgave.

```
ISP>>glue/axis=2 x1 x2 > x
ISP>>regress x y > cm:korrel
degrees of freedom:     5 - 3 = 2
sigma      =  0.8972
R-square   =  0.9968
F-stat     =   313.8       (2 over 2 df)
condition =   1.517

 var          coef          sdev
   1           2.170        0.8972E-01
   2          -1.167        0.1787
   const       6.167        0.9148
ISP>># korrel indeholder korrelationsmatricen for
ISP>># parameterestimaterne beta1, beta2, beta0.
ISP>>print korrel /format=(3f10.4)
    1.0000    0.0000    0.0000
    0.0000    1.0000   -0.8987
    0.0000   -0.8987    1.0000
ISP>>
```

Hvad er løsningen til estimationsligningerne?

- Rent faktisk blev y-værdierne frembragt ved hjælp af den statistiske model $y = 5 + 2x_1 - x_2 + \varepsilon$, hvor ε er normalfordelt med middelværdi 0 og varians 1.

  Hvordan harmonerer de foreliggende y-værdier med denne model?

- Ved hjælp af tilføjelsen **cm:korrel** på **regress**-kommandolinien blev korrelationsmatricen for parameterestimaterne anbragt i arrayet **korrel**.

  Korrelationer er størrelser der altid vil ligge mellem -1 og 1, og hvis to parameterestimater har en korrelation som er meget tæt på -1 eller 1, er det et tegn på at de tilhørende baggrundsvariable indeholder stort set samme information (det er altså ønskeligt med korrelationer tæt på 0).

  Er der nogen sammenhæng mellem estimationsligningernes udseende og korrelationsmatricens udseende?

Opgave 8.2

Fortsættelse af Opgave 7.1:

Man har en formodning om at civilisationen har en stressende og dermed blodtryks-øgende virkning. Prøv derfor at lave regression af 'blodtryk' (y) på 'brøkdel af livet i de nye omgivelser' (x_1). Er der en signifikant afhængighed af x_1?

Der foreligger endnu en baggrundsvariabel, nemlig 'vægt' (x_2). Undersøg om et plot af residualerne fra den første regression mod x_2 viser noget interessant. Prøv at lave multipel regression af 'blodtryk' (y) på 'brøkdel af livet i de nye omgivelser' (x_1) og 'vægt' (x_2). Er der nu en signifikant afhængighed af x_1?

Forklar resultaterne.

Opgave 8.3

Hvad kan man få ud af at benytte (simpel eller multipel) regressionsanalyse i Opgave 7.4?

Opgave 8.4: Træers rumfang

Inden for skovbruget er man interesseret i at kunne vurdere et træs indhold af tømmer, dvs. dets *rumfang*, uden alt for stort besvær. Nogle størrelser der er nemme at bestemme er *diameter* og *højde*, og det ville

Tabel 8.6: Opgave 8.4: Diameter d (i *inches*), højde h (i *feet*) og rumfang v (i kubik*feet*) for 31 sortkirsebærtræer.

d	h	v	d	h	v
8.3	70	10.3	12.9	85	33.8
8.6	65	10.3	13.3	86	27.4
8.8	63	10.2	13.7	71	25.7
10.5	72	16.4	13.8	64	24.9
10.7	81	18.8	14.0	78	34.5
10.8	83	19.7	14.2	80	31.7
11.0	66	15.6	14.5	74	36.3
11.0	75	18.2	16.0	72	38.3
11.1	80	22.6	16.3	77	42.6
11.2	75	19.9	17.3	81	55.4
11.3	79	24.2	17.5	82	55.7
11.4	76	21.0	17.9	80	57.3
11.4	76	21.4	18.0	80	51.5
11.7	69	21.3	18.0	80	51.0
12.0	75	19.1	20.6	87	77.0
12.9	74	22.2			

være praktisk hvis man kunne forudsige et træs rumfang så nogenlunde ud fra disse to størrelser.

Man har derfor målt diameteren d (i en højde af 4.5 feet over jorden), højden h og rumfanget (volumenet) v for 31 træer af en bestemt slags (sortkirsebærtræer i Allegheny National Forest, Pennsylvania). Resultaterne er vist i Tabel 8.6.

Opgaven er nu at undersøge, *om* man med en simpel statistisk model kan bestemme v ud fra kendskab til d og h, og i givet fald *hvordan* og *hvor godt*.

Tip: Der er mulighed for forskellige regressionsanalyser. Man kan også prøve at udnytte, at rumfang er noget med højde gange tværsnitsareal.

Data kan indlæses med kommandoen `getdata` (data-navn `tree`).

Kapitel 9

Vægtede mindste kvadrater

I de hidtidige kapitler har vi som estimationsmetode benyttet mindste kvadraters metode, der, som det er fremgået, går ud på at estimere parametrene således at summen af de kvadratiske afvigelser bliver mindst mulig. I den forbindelse indgår alle observationer med samme vægt, og det er begrundet i at alle observationer har samme varians σ^2, svarende til at de tilfældige afvigelser antages at være tilfældige tal fra en og samme (normal)fordeling med middelværdi 0 og varians σ^2.

Man kan imidlertid sagtens komme ud for situationer hvor en sådan fremgangsmåde ikke er rimelig. Nogle eksempler kan være:

- Observationerne $y_1, y_2, \ldots, y_n$ er fremkommet som middeltal af *forskellige* antal målinger. — Hvis y_i er et middeltal af n_i målinger og hvis de enkelte målinger har samme varians σ^2, så har y_i variansen σ^2/n_i.

  Se Afsnit 9.5 og Opgave 9.2.

- Der er opgivet en kendt usikkerhed (standardafvigelse) på hver enkelt af observationerne $y_1, y_2, \ldots, y_n$ (se f.eks. Opgave 9.3).

- Man kender ad teoretisk vej en funktionel sammenhæng mellem y-ernes middelværdi og varians (Kapitel 10 behandler eksempler herpå).

9.1 Estimation af parametrene

Sagen kan gribes an på følgende måde. For nemheds skyld ser vi på den simple lineære regressionsmodel, men metoden kan uden videre generaliseres til multipel regression.

Der foreligger datapunkter (x_i, y_i), $i = 1, 2, \ldots, n$, og vi går som modelantagelse ud fra, at der findes parametre β_0 og β_1 således at

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i \tag{9.1}$$

for alle i. Observationerne $y_1, y_2, \ldots, y_n$ har nu ikke længere nødvendigvis *samme* varians, men det antages at y_i har variansen σ^2/w_i, hvor $w_1, w_2, \ldots, w_n$ er *kendte* tal og σ^2 (i nogle situationer) er ukendt.

Når man her skal opskrive den kvadratsum der skal minimaliseres, skal den i-te kvadratiske afvigelse indgå med en vægt som er omvendt proportional med den i-te varians, mere præcist med vægten w_i, så kvadratsummen skal være

$$\sum_{i=1}^{n} w_i \Big(y_i - (\beta_0 + x_i\beta_1)\Big)^2. \tag{9.2}$$

Nu er

$$\begin{aligned} w_i \Big(y_i - (\beta_0 + x_i\beta_1)\Big)^2 &= \Big(\underbrace{w_i^{1/2}}_{y_i^*} y_i - (\underbrace{w_i^{1/2}}_{x_{i1}^*}\beta_0 + \underbrace{w_i^{1/2}x_i}_{x_{i2}^*}\beta_1) \Big)^2 \\ &= \Big(y_i^* - (x_{i1}^*\beta_0 + x_{i2}^*\beta_1)\Big)^2, \end{aligned}$$

så den vægtede kvadratsum (9.2) kan opfattes som en sædvanlig kvadratsum i det multiple regressionsproblem

$$y_i^* = x_{i1}^*\beta_0 + x_{i2}^*\beta_1. \tag{9.3}$$

Der gælder at størrelserne $y_1^*, y_2^*, \ldots, y_n^*$ har samme varians σ^2,[1] og derfor er det korrekt at løse det "stjernede" multiple regressionsproblem (9.3) med sædvanlig uvægtet mindste kvadraters metode. Pointen i det hele er, at det oprindelige og det "stjernede" regressionsproblem har

[1] Ifølge regnereglerne for varianser (se f.eks. Box 5.1 på side 59) er variansen på $y_i^* = w_i^{1/2}y_i$ lig med variansen på y_i ganget med $(w_i^{1/2})^2 = w_i$, dvs. variansen er $(\sigma^2/w_i) \times w_i = \sigma^2$.

samme β_0, β_1 og σ^2, og at man derfor kan finde løsningen til det oprindelige vægtede regressionsproblem ved almindelig mindste kvadraters metode på det nye problem.

Hvis man vil løse regressionsproblemet (9.3) med håndkraft skal man som sædvanlig opstille og løse et sæt estimationsligninger. Ligningerne bliver (efter at der er pyntet lidt på formlerne)

$$\begin{aligned}\left(\sum_{i=1}^{n} w_i\right)\widehat{\beta}_0 + \left(\sum_{i=1}^{n} w_i x_i\right)\widehat{\beta}_1 &= \sum_{i=1}^{n} w_i y_i \\ \left(\sum_{i=1}^{n} w_i x_i\right)\widehat{\beta}_0 + \left(\sum_{i=1}^{n} w_i x_i^2\right)\widehat{\beta}_1 &= \sum_{i=1}^{n} w_i x_i y_i\,.\end{aligned}$$

Løsningerne $\widehat{\beta}_0$ og $\widehat{\beta}_1$ er vægtet mindste kvadraters estimaterne i det oprindelige problem (9.1). Estimatet over variansparameteren σ^2 er

$$\begin{aligned}s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i^* - (x_{i1}^*\widehat{\beta}_0 + x_{i2}^*\widehat{\beta}_1)\right)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n} w_i\left(y_i - (\widehat{\beta}_0 + x_i\widehat{\beta}_1)\right)^2\end{aligned}$$

med $n-2$ frihedsgrader.

9.2 Vægtet regression med ISP

Det er forholdsvis let at foretage vægtet regression med ISP, man skal blot lave det vægtede problem om til det tilsvarende uvægtede problem som så kan løses med **regress** (med **/const=n**). Det vil typisk foregå sådan, idet **x**, **y** og **w** er arrays indeholdende værdierne af x, y og w:

```
ISP>>glue/axis=2 (sqrt(w)) (sqrt(w)*x) > xx
ISP>>regress/const=n xx (sqrt(w)*y)
```

Det er imidlertid endnu lettere at benytte kommandoen **rg_glim**; dette er en kommando til generaliseret lineær regression (der omtales nærmere i Kapitel 10), men den kan også benyttes til almindelig vægtet lineær regression. Hvis **x**, **y** og **w** er som ovenfor, skal man blot skrive

```
ISP>>rg_glim x y w
```

Med **regress** kan man få gemt *residualerne* i et output-array. Det kan man ikke med **rg_glim**, til gengæld kan man få gemt de *fittede værdier* $\widehat{y}$, typisk sådan:

```
ISP>>rg_glim x y w > fi:yhat
```

Udskriften fra **rg_glim** minder en del om udskriften fra **regress**. De estimerede koefficienter hedder dog nu **beta** (i stedet for **coef**), og de står i en anden rækkefølge. Den estimerede standardafvigelse s hedder **s0** (i stedet for **sigma**). Af interesse er også størrelserne **deviance** og **Pearson X2**; de er ens når der er tale om normalfordelingsmodeller, og deres værdi er værdien af residualkvadratsummen.

9.3 Modelkontrol

Hvis der kun er én forklarende variabel x, er det naturligt som led i modelkontrollen at lave en tegning der dels indeholder de rigtige datapunkter (x_i, y_i), dels den estimerede kurve.

Under alle omstændigheder bør man lave residualundersøgelser (residualplots, histogrammer osv., se f.eks. Afsnit 8.4), og det skal være af de såkaldt *vægtede residualer* $(y_i - \widehat{y}_i)/w_i^{1/2}$. — Grunden til at det skal være de vægtede residualer er, at det er dem der bør have samme varians.

I visse tilfælde kan man foretage modelkontrol baseret på variansskønnet s^2, nemlig i situationer hvor man enten ad anden vej kan bestemme et skøn s_1^2 over σ^2, eller kender den teoretiske varians σ^2. Modelkontrollen går ud på at teste om s^2 afviger signifikant fra denne anden værdi; den nærmere udformning af testet afhænger af om "den anden værdi" selv er et estimat eller om den er eksakt:

1. **σ^2 ukendt.**
   Hvis s^2 er variansestimatet fra regressionen og s_1^2 er et andet estimat (med f_1 frihedsgrader) over σ^2, og hvis s^2 og s_1^2 er uafhængige af hinanden, og hvis observationerne er normalfordelte, så kan man benytte

   $$F_{\text{obs}} = \frac{s^2}{s_1^2}$$

   som teststørrelse. Hvis regressionsmodellen er rigtig så er denne størrelse F-fordelt med frihedsgradsantal f og f_1, og den bør ikke

være alt for meget større end 1. Det betyder at man almindeligvis vil forkaste modellen hvis der er meget lille sandsynlighed for at få en F-værdi større end F_{obs}.

Se Afsnit 9.5 for et eksempel.

2. **σ^2 kendt.**
Hvis observationerne $y_1, y_2, \ldots, y_n$ har de *kendte* varianser σ^2/w_i, $i = 1, 2, \ldots, n$, kan vi gå ud fra at w-erne er indrettet sådan at $\sigma^2 = 1$. I så fald bør estimatet s^2 være ca. 1.

Det er hensigtsmæssigt at benytte residualkvadratsummen

$$\begin{aligned} D &= \sum_{i=1}^{n} w_i (y_i - \widehat{y}_i)^2 \\ &= f\, s^2 \end{aligned}$$

som teststørrelse (her er f antallet af frihedsgrader). Hvis modellen er rigtig, og hvis observationerne er normalfordelte, så er denne størrelse χ^2-fordelt med f frihedsgrader. Det betyder at man vil forkaste modellen hvis der er en meget lille sandsynlighed for at få værdier større end den observerede værdi D.

9.4 Estimaternes middelfejl

På samme måde som i Kapitel 6 og i Afsnit 8.3 kan man teste hypoteser om modellens parametre idet man sammenligner parameterestimaterne med deres (estimerede) middelfejl:

- Hvis σ^2 er ukendt, altså hvis observationerne har en varians som er "kendt pånær en konstant faktor", så kan middelfejlene direkte aflæses som `sdev` i ISP-udskriften.

- Hvis σ^2 er kendt og lig med 1, så skal parameterestimaternes middelfejl udregnes som ISP-udskriftens `sdev` divideret med den estimerede standardafvigelse `sigma` (eller `s0`).

Box 9.1: Om tests i normalfordelingsmodeller

I forbindelse med regressionsanalyse af normalfordelte observationer optræder der blandt andet t-tests, F-tests og χ^2-tests.

- t-testet benyttes til at teste hypoteser vedrørende en af regressionsligningens koefficienter. Det har typisk formen

$$t = \frac{\widehat{\beta}_k - c}{\text{est. middelfejl på } \widehat{\beta}_k}.$$

  Antal frihedsgrader for t-størrelsen er lig antal frihedsgrader for det variansskøn som den estimerede middelfejl i nævneren beregnes ud fra.

- F-testet benyttes til at sammenligne to forskellige og uafhængige estimater over den samme varians σ^2. [a] Det har typisk formen

$$F = \frac{s_1^2}{s_2^2}.$$

  Der er to frihedsgradsantal for en F-teststørrelse, nemlig tællervariansskønnets og nævnervariansskønnets.

- χ^2-testet benyttes til at sammenligne et variansskøn med en formodet sand værdi af variansen. Det har typisk formen

$$\frac{fs^2}{\sigma_0^2}$$

  hvor f er antal frihedsgrader for s^2 og σ_0^2 er den formodede sande værdi.

[a] Det har man blandt andet brug for når man vil teste en hypotese der vedrører mere end én af koefficienterne i regressionsligningen.

Tabel 9.1: Kvælning af hunde: Målinger af hypoxantinkoncentration til de fire forskellige tidspunkter. I hver gruppe er observationerne ordnet efter størrelse.

varighed (min)	koncentration (μmol/l)						
0	0.0	0.0	1.2	1.8	2.1	2.1	3.0
6	3.0	4.9	5.1	5.1	7.0	7.9	
12	4.9	6.0	6.5	8.0	12.0		
18	9.5	10.1	12.0	12.0	13.0	16.0	17.1

9.5 Et eksempel

Dette afsnit består af et eksempel der viser, hvordan man kan foretage vægtet lineær regression i en situation hvor y-erne er gennemsnit af forskellige antal målinger der hver især kan antages at have samme varians.

Syv hunde er (under bedøvelse) blevet udsat for iltmangel ved sammenpresning af luftrøret, og hypoxantinkoncentrationen i cerebrospinalvæsken måltes efter 0, 6, 12 og 18 minutters forløb. Det var af forskellige grunde ikke muligt at foretage målinger på alle syv hunde til alle fire tidspunkter, og det kan heller ikke afgøres hvordan målinger og hunde hører sammen. Resultaterne af forsøget er vist i Tabel 9.1. Man ønsker at beskrive sammenhængen mellem varighed (x) og koncentration (y).

Hvis vi går ud fra at sammenhængen er *lineær*, er det eneste rigtige i denne situation naturligvis at foretage almindelig lineær regression på de 25 sammenhørende værdier af tid og koncentration, hvilket giver følgende resultat:

```
ISP>>regress tid konc
degrees of freedom:     25 - 2 = 23
sigma      =   2.202
R-square   =  0.8033
F-stat     =   93.92       (1 over 23 df)
condition  =   1.282

 var          coef          sdev
   1          0.6077        0.6271E-01
   const       1.415        0.7100
```

Tabel 9.2: Kvælning af hunde: det reducerede datamateriale.

varighed (min)	gnsnt-konc (μmol/l)	antal målinger
0	1.457	7
6	5.500	6
12	7.480	5
18	12.81	7

Antag nu at en eller anden entreprenant person havde besluttet, at der ikke var nogen grund til at opgive alle koncentrationsmålingerne, det var nok at opgive gennemsnittene. Så ville datamaterialet se ud som i Tabel 9.2. I så fald burde man foretage *vægtet* lineær regression med antallene som vægte. Det kan gøres på følgende måde, idet `mtid`, `mkonc` og `mn` indeholder tider, antal og koncentrationer:

```
ISP>>print mtid mkonc mn

  MTID
  0          6          12         18

  MKONC
  1.457      5.500      7.480      12.81

  MN
  7          6          5          7

ISP>>rg_glim mtid mkonc mn
GENERALIZED LINEAR REGRESSION   Ver. 2.0
  Error distribution: normal
  Link function:      identity
End after 1 iterations.
Number of good obs.: 4,   degrees of freedom: 2

 no.         beta       sdev       sdev/s0
  0          1.415     0.7277     0.3224
  1         0.6077     0.6428E-01 0.2848E-01

         s0 = 2.257
 Pearson X2 = 10.19
   deviance = 10.19
```

Det ses at vi på denne måde får den samme regressionlinie som da vi korrekt brugte alle observationerne, nemlig $y = 1.415 + 0.6077x$. Derimod bliver den estimerede standardafvigelse og de estimerede middelfejl anderledes.[2]

Hvis vi foretager *uvægtet* regression, bliver resultatet

```
ISP>>regress mt mkonc
degrees of freedom:     4 - 2 = 2
sigma      =  0.9706
R-square   =  0.9718
F-stat     =   68.98       (1 over 2 df)
condition  =   1.342

 var           coef          sdev
   1           0.6009        0.7235E-01
   const        1.405        0.8121
```

•

Da eksemplet er sådan indrettet at der til hver x-værdi er flere y-værdier, er det muligt at foretage et F-test for om den lineære regressionsmodel giver en tilstrækkelig god beskrivelse af observationerne (det svarer til Punkt 1 på side 124).

1. Den vægtede regression (`rg_glim mtid mkonc mn`) gav os variansestimatet $s^2 = 2.257^2$ med 2 frihedsgrader; denne varians fortæller hvor meget gennemsnittene varierer tilfældigt omkring regressionslinien.

2. Hvis vi tænker på observationsmaterialet på den måde, at der er 25 y-værdier der er delt op i fire grupper (svarende til de fire tider), så er der tale om en *ensidet variansanalyse*-situation (se Afsnit 8.6), og vi kan udregne et skøn s_1^2 over variationen inden for grupper:

```
ISP>>x = tid==trn(mtid)
ISP>>regress/const=n x konc
degrees of freedom:     25 - 4 = 21
sigma      =   2.196
```

[2]Middelfejlen divideret med standardafvigelse (`sdev/s0`) bliver den samme ved de to metoder.

```
condition =   1.183

 var          coef          sdev
   1          1.457        0.8302
   2          5.500        0.8967
   3          7.480        0.9823
   4          12.81        0.8302
```

Det ses at variationen inden for grupper er $s_1^2 = 2.196^2$ med 21 frihedsgrader.

3. Hvis gennemsnittenes variation omkring regressionslinien er meget større end variationen inden for grupper, så er det tegn på at regressionslinien ikke beskriver observationerne godt, dvs. tegn på at man må forkaste regressionsmodellen. Vi vil altså forkaste modellen hvis teststørrelsen

$$F_{\text{obs}} = \frac{s^2}{s_1^2}$$

er meget større end 1, dvs. hvis der er lille chance for at få en værdi større end F_{obs}.

Den fordeling man skal sammenligne F_{obs} med, er F-fordelingen med 2 og 21 frihedsgrader.

4. Vi har at $F_{\text{obs}} = 2.257^2/2.196^2 = 1.056$. Sandsynligheden for at få en værdi større end 1.056 i F-fordelingen med 2 og 21 frihedsgrader kan findes i statistiske tabeller; man kan også bruge ISP's `fpr()`-funktion, f.eks. således

```
ISP>>print (1-fpr(1.056, 2, 21))
 0.3656
```

Da der er over 36% chance for at få en større F-værdi, slutter vi at regressionsmodellen beskriver observationerne udmærket.

9.6 Endnu et eksempel

Dette afsnit handler om, at et ønske om at transformere y-værdierne med (f.eks.) logaritmefunktionen har betydning for antagelsen om at y-erne har samme varians.

Betragt eksemplet om vands strømningsforhold i en flod, Opgave 3.3. Som det vil erindres, tyder tegninger på at flowraten y ikke beskrives særlig godt med en lineær funktion af vanddybden x. Vi vil nu diskutere hvordan man kan fitte en *eksponentialkurve* af formen

$$y = \exp(\beta_0 + x\beta_1) \tag{9.4}$$

til de observerede punkter.

En nærliggende løsningsmetode ville være at foretage almindelig regression af $\ln(y)$ på x, fordi (9.4) jo er ensbetydende med

$$\ln(y) = \beta_0 + x\beta_1.$$

Denne fremgangsmåde er imidlertid bestemt ikke uproblematisk. For den almindelige regression forudsætter som bekendt, at de størrelser der bruges som afhængig variabel (y) har samme varians. Men hvis flowværdierne har samme varians for alle dybder, så har logaritmen til flowværdierne *ikke* denne egenskab, og hvis det omvendt var sådan at logaritmen til flowværdierne havde samme varians for alle dybder, så ville flowværdierne selv ikke have denne egenskab; dette kommer af at logaritmefunktionens hældningskoefficient ikke er konstant.[3]

Lad os nu gå ud fra at det er flowværdierne der har samme varians, så den korrekte metode til at fitte en ret linie er almindelig simpel lineær regression af y på x. Hvis vi ønsker at fitte eksponentialkurven (9.4), vil det så være mest korrekt at foretage vægtet regression af $\ln(y)$ på x. Sagen er bare, at for at kende vægtene præcist, skal vi kende den fittede kurve, og for at kende den skal vi kende vægtene .... Problemet løses ved en iterativ metode: man begynder med et forslag til fittet kurve, derudfra beregnes vægte, ved hjælp af vægtene beregnes en ny fittet kurve, udfra den beregnes nye vægte, de benyttes til en ny fittet kurve osv.

Forsynet med passende oplysninger kan ISP-kommandoen `rg_glim` foretage en sådan iteration. Det gøres med et kald af formen

[3] Hvis y har middelværdi μ og varians σ^2, så har $\ln(y)$ en varians der ca. er σ^2/μ^2. Mere generelt har $f(y)$ en varians der ca. er $f'(\mu)^2\sigma^2$. For at indse dette benyttes rækkeudviklingen $f(y) \approx f(\mu) + f'(\mu)(y - \mu)$.

```
ISP>>rg_glim x y /link=1
```

hvor det afgørende er tilføjelsen `/link=1` der fortæller at den såkaldte *linkfunktion* skal være logaritmefunktionen[4]. Linkfunktionen er den funktion der transformerer y-erne (eller snarere y-ernes middelværdier) over på en skala hvor der bliver tale om en lineær afhængighed af de forklarende variable.

I eksemplet giver det følgende resultat (bemærk at man ikke skal angive noget array med vægte):

```
ISP>>rg_glim/link=1 dybde flow
GENERALIZED LINEAR REGRESSION   Ver. 2.0
  Error distribution: normal
  Link function:      log              log(y)
End after 4 iterations.
Number of good obs.: 10,   degrees of freedom: 8

 no.         beta       sdev       sdev/s0
  0         -2.000     0.1970     0.7566
  1          5.226     0.2715      1.043

        s0 = 0.2604
 Pearson X2 = 0.5424
   deviance = 0.5424
```

Den estimerede eksponentialkurve er således

$$y = \exp(-2 + 5.226x)$$

og den estimerede varians på flowværdierne er 0.2604^2 med 8 frihedsgrader.

Den bedste rette linie (fundet på sædvanlig vis) er $y = 13.83x - 3.982$, resulterende i en estimeret varians på y-erne på 0.6035^2, ligeledes med 8 frihedsgrader, og den bedste parabel er $y = 23.54x^2 - 10.86x + 1.683$, resulterende i en estimeret varians på 0.2794^2 med 7 frihedsgrader. Eksponentialmodellen og andengradsmodellen er altså stort set lige gode, forstået på den måde at de giver nogenlunde samme variansestimat.

[4]Læs mere om linkfunktioner i Kapitel 10.

9.7 Opgaver

Opgave 9.1

Denne opgave går ud på at fitte en ret linie $y = \beta_0 + x\beta_1$ med *vægtede* mindste kvadraters metode. De sammenhørende værdier af x, y og vægt w er

y	x	w
10.5	1	4
12.1	2	4
13.3	3	1
14.7	4	1
16.2	5	1
16.8	6	1
17.1	7	4
17.9	8	4

1. (a) Som forklaret i teksten kan man finde løsningen til det vægtede problem på den måde, at man formulerer et helt andet, uvægtet problem, der er sådan indrettet at det har den samme løsning som det vægtede problem. Hvordan ser dette uvægtede problem ud?

   (b) Opskriv estimationsligningerne og find $\widehat{\beta}_0$ og $\widehat{\beta}_1$.
   (Hjælp: $\sum wy = 291.4$ og $\sum wxy = 1470.9$.)

   (c) Lav en tegning (skitsemæssigt) af punkter plus fittet linie.

   (d) Hvordan skal man bære sig ad for at finde residualerne?

2. (a) Løs Punkt 1 med ISP.
   Lav også en tegning af punkter plus fittet linie.

   (b) Hvad bliver resultatet hvis man laver sædvanlig uvægtet regression?

Opgave 9.2: McIntosh-æbler

Æbletræer (og mange andre træer) har to typer skud: lange skud, ledeskud, der kan vokse op til 15-20 cm i løbet af en vækstsæson, og korte skud, sporer, der ofte ikke bliver mere end 1 cm lange. Det kan forekomme at det ene års lange skud bliver til sporer det næste år og omvendt. — Det er almindeligvis på sporerne at frugten kommer.

I en undersøgelse af forskellen mellem de to slags skud på McIntosh-æbler har man[5] i 1971 indsamlet data fra et antal podede sunde træer fra 1933 og 1934 på den måde, at man vækstsæsonen igennem med få dages mellemrum har udvalgt nogle tilfældige skud og så for hvert skud registreret

1. om det var et ledeskud eller en spore,
2. hvor mange stilk-ansatser der var.

(De udvalgte skud blev skåret af og bragt til laboratoriet.)

Tabel 9.3 viser nogle af forsøgsresultaterne. Som det ses har man ikke gengivet data for hvert enkelt skud, man har kun noteret middeltal, standardafvigelse og antal observationer for hver dag.

Man ønsker at finde en ligning der beskriver hvordan antal stilk-ansatser afhænger af antal vækstdage (dvs. antal dage fra vækstperiodens start). Der foreligger ingen hortonomiske teorier der udtaler sig herom, så man må bare prøve med almindelige simple statistiske modeller, f.eks. simpel lineær regression.

Gør det!

Da $\overline{y}$-erne er gennemsnit af *forskellige* antal observationer, er det ikke rimeligt at benytte almindelig mindste kvadraters metode. Man bør derimod benytte vægtede mindste kvadraters metode hvor vægtene afhænger af antal observationer (jf. Afsnit 9.5).

Standardafvigelserne $s_1, s_2, \ldots$ benyttes ikke til noget i forbindelse med bestemmelsen af den bedste rette linie. Til gengæld kan man benytte dem ved en vurdering af, om den fittede linie er god nok: Ud fra de individuelle standardafvigelser udregnes den såkaldte *varians inden for grupper* s_0^2 som et vægtet gennemsnit af de individuelle varianser, med frihedsgradsantallene som vægte:

$$s_0^2 = \frac{\sum(n_i - 1)s_i^2}{\sum(n_i - 1)}.$$

Dette variansskøn, der har $\sum(n_i - 1)$ frihedsgrader, kan benyttes som nævner i et F-test for modellens rigtighed, jf. punkt 1 på side 124.

Tip: ISP-kommandoen **getdata** (data-navn **mcintosh**) indlæser data til arrays **tl**, **nl**, **yl**, **sl** og **ts**, **ns**, **ys**, **ss** (**l**-et står for ledeskud, **s**-et for sporer).

[5] J. Bland (1978): A comparison of certain aspects of ontogeny in the long and short shoots of McIntosh apple during one annual growth cycle. Ph.D.-afhandling fra University of Minnesota. Her citeret efter S. Weisberg (1980): *Applied Linear Regression*. New York: Wiley.

Tabel 9.3: Opgave 9.2: Stilk-ansatser på McIntosh-træer, bestemt på forskellige dage i løbet af vækstperioden.
t er tidspunktet (antal dage fra vækstperiodens start), n er antal undersøgte skud på den pågældende dag, $\overline{y}$ er det gennemsnitlige antal stilk-ansatser pr. skud, og s er standardafvigelsen på antal stilk-ansatser pr. skud.

ledeskud				sporer			
t	n	$\overline{y}$	s	t	n	$\overline{y}$	s
0	5	10.20	0.83	0	5	10.00	0.00
3	5	10.40	0.54	6	5	11.00	0.72
7	5	10.60	0.54	9	5	10.00	0.72
13	6	12.50	0.83	19	11	13.36	1.03
18	5	12.00	1.41	27	7	14.29	0.95
24	4	15.00	0.82	30	8	14.50	1.19
25	6	15.17	0.76	32	8	15.38	0.51
32	5	17.00	0.72	34	5	16.60	0.89
38	7	18.71	0.74	36	6	15.50	0.54
42	9	19.22	0.84	38	7	16.86	1.35
44	10	20.00	1.26	40	4	17.50	0.58
49	19	20.32	1.00	42	3	17.33	1.52
52	14	22.07	1.20	44	8	18.00	0.76
55	11	22.64	1.76	48	22	18.46	0.75
58	9	22.78	0.84	50	7	17.71	0.95
61	14	23.93	1.16	55	24	19.42	0.78
69	10	25.50	0.98	58	15	20.60	0.62
73	12	25.08	1.94	61	12	21.00	0.73
76	9	26.67	1.23	64	15	22.33	0.89
88	7	28.00	1.01	67	10	22.20	0.79
100	10	31.67	1.42	75	14	23.86	1.09
106	7	32.14	2.28	79	12	24.42	1.00
				82	19	24.79	0.52
				85	5	25.00	1.01
				88	27	26.04	0.99
				91	5	26.60	0.54
				94	16	27.12	1.16
				97	12	26.83	0.59
				100	10	28.70	0.47
				106	15	29.13	1.74

Opgave 9.3: π^--mesoner

I kernefysik er man interesseret i at undersøge den såkaldte stærke vekselvirkningskraft, der blandt andet er ansvarlig for at holde sammen på atomkernens partikler, selv om mange af disse har positiv elektrisk ladning og derfor skulle frastøde hinanden. Fysikerne kan få et indtryk af nogle sider af denne kraft gennem eksperimenter der går ud på at sende visse elementarpartikler med stor hastighed ind mod andre elementarpartikler; når partiklerne kolliderer sker der forskellige omdannelser, og man stiller sig op og måler hvad det er der kommer ud af kollisionen.

I ét sådant forsøg[6] bombarderer man protoner med nogle partikler der hedder π^--mesoner. Eksperimentator kan selv bestemme partiklernes impuls p (masse gange hastighed); kvadratet på den totale energi er givet som $E^2 = 2mp$ hvor m er protonens masse. Ved sammenstødet dannes forskellige partikler, og man tæller blandt andet hvor mange π^--mesoner der sendes ud i en bestemt retning; derved fås det såkaldte spredningstværsnit $\Delta\sigma$, som stort set er antal partikler registreret af måleapparatet pr. sekund divideret med antal partikler udsendt pr. sekund. Forsøgsresultaterne er vist i Tabel 9.4. De enkelte målinger er gentaget så mange gange, at fysikerne hævder at de kan give en præcis værdi for standardafvigelsen på $\Delta\sigma$-værdien.

Visse fysiske teorier forudsiger at sammenhængen mellem $\Delta\sigma$ og E er af formen

$$\Delta\sigma = \beta_0 + \beta_1 E^{-1} + \text{noget småt}; \tag{9.5}$$

muligvis skal man dog have et andengradsled med, altså

$$\Delta\sigma = \beta_0 + \beta_1 E^{-1} + \beta_2 E^{-2} + \text{noget småt}. \tag{9.6}$$

De indgående β-er har en fortolkning og betydning i den fysiske model.

Den statistiske opgave er nu

- at undersøge hvilken af de to modeller (9.5) og (9.6) der er bedst;
- at afgøre om den bedste af de to modeller også er god *nok*;
- at angive estimater (inklusive standardafvigelser) over β-erne.

[6] H. Weisberg m.fl. (1978): s-dependence of proton fragmentation by hadrons. II. Incident laboratory momenta 30-250 GeV/c. *Phys. Rev. D*, **17**, 2875-2887.

Tabel 9.4: Opgave 9.3: Målinger på π^--mesoner.
Impulsen p er opgivet i GeV (giga-elektronvolt = 10^9eV), størrelsen E^{-1} er opgivet i GeV^{-1}, og spredningstværsnittet $\Delta\sigma$ samt dets standardafvigelse er opgivet i $10^{-34}\mathrm{m}^2$.

p	E^{-1}	$\Delta\sigma$	st.afv.
4	0.345	367	17
6	0.287	311	9
8	0.251	295	9
10	0.225	268	7
12	0.207	253	7
15	0.186	239	6
20	0.161	220	6
30	0.132	213	6
75	0.084	193	5
150	0.060	192	5

Gør dette!

— Man kan foretage et test for modellens brugbarhed ved hjælp af metoden beskrevet i Punkt 2 på side 125.

Tip: Data kan indlæses til ISP med kommandoen **getdata** (data-navn **meson**).

Kapitel 10

Generaliseret lineær regression

De forrige kapitlers regressionsmodeller (med undtagelse af den i eksemplet i Afsnit 9.6) bygger på bestemte modelantagelser om observationerne $y_1, y_2, \ldots, y_n$:

1. Den *systematiske* del af observationernes variation beskrives ved hjælp af et antal baggrundsvariable $x_1, x_2, \ldots, x_p$ og tilhørende parametre $\beta_0, \beta_1, \ldots, \beta_p$ på den måde, at den *forventede* y-værdi μ hørende til et bestemt sæt værdier $x_1, x_2, \ldots, x_n$ er givet som

$$\mu = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p . \qquad (10.1)$$

2. Den *tilfældige* del af variationen beskrives på den måde at

   (a) observationerne er *stokastisk uafhængige*, dvs. dé tilfældige faktorer der indvirker på én af observationerne, virker uafhængigt af dem der indvirker på de øvrige observationer,

   (b) der er ikke nogen funktionel sammenhæng mellem den systematiske og den tilfældige del af modellen, dvs. β-erne har ingen indflydelse på den tilfældige variation,

   (c) observationerne er *normalfordelte* med samme (ukendte) varians σ^2 (eller med en varians der er kendt pånær en konstant faktor).

Spørgsmålet er nu, hvad man kan stille op hvis man ikke er parat til at gøre disse antagelser, og svaret er at det kommer an på hvad der så gælder i stedet. I *generaliseret lineær regression* svækkes modelantagelserne i to retninger:

1. Det behøver ikke være selve den forventede værdi μ der kan skrives på formen (10.1), det er nok at der for en eller anden kendt funktion g gælder

$$g(\mu) = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p$$

eller ensbetydende hermed

$$\mu = g^{-1}(\beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p),$$

hvor g^{-1} betegner den omvendte funktion til g.

(Det så vi et eksempel på i Afsnit 9.6, hvor funktionen g var logaritmefunktionen.)

Denne funktion g der således transformerer middelværdien μ over på en skala hvor der bliver en lineær afhængighed af de forklarende variable, kaldes for *link-funktionen*. — Om en linkfunktion kræves, at den skal være defineret på det interval hvori y-værdierne kan ligge, og at den skal være monoton og kontinuert differentiabel.

I Tabel 10.1 ses nogle ofte anvendte linkfunktioner.

2. Hvad den tilfældige del af variationen angår, så skal observationerne stadig være stokastisk uafhængige. Derimod behøver de ikke længere være normalfordelte, men kan f.eks. være binomialfordelte eller Poissonfordelte, og derved er der mulighed for at modellere *antalsobservationer*.

I generaliseret lineær regression skal der være en vis veldefineret sammenhæng mellem observationernes middelværdi og deres varians (altså mellem den systematiske og den tilfældige side af modellen), nemlig på den måde at y-s varians $\mathrm{Var}(y)$ er proportional med en kendt funktion V af den forventede værdi μ,

$$\mathrm{Var}(y_i) = \sigma^2 \cdot V(\mu_i)/w_i,$$

hvor σ^2 enten er en ukendt parameter eller konstanten 1, og hvor $w_1, w_2, \ldots, w_n$ er kendte positive tal.

Tabel 10.1: Nogle ofte anvendte linkfunktioner og deres omvendte funktioner.

navn	$g(y)$	def.mængde	$g^{-1}(z)$
identiteten	y	$y \in]-\infty, +\infty[$	z
log	$\ln y$	$y \in]0, +\infty[$	$\exp(z)$
reciprok	$1/y$	$y \neq 0$	$1/z$
logit	$\ln \frac{y}{1-y}$	$y \in]0,1[$	$\frac{\exp(z)}{1+\exp(z)}$
probit	$\Phi^{-1}(y)$	$y \in]0,1[$	$\Phi(z)$
c-log-log	$\ln(-\ln(1-y))$	$y \in]0,1[$	$1-\exp(-\exp(z))$

Den generaliserede lineære regressionsmodel kommer derfor i sin generelle udformning til at se sådan ud: For hvert i ($i = 1, 2, \ldots, n$) er der en observation y_i, værdier $x_{i1}, x_{i2}, \ldots, x_{ip}$ af de p baggrundsvariable[1], samt et positivt tal w_i. Desuden er der en kendt linkfunktion g og en kendt variansfunktion V. Endelig er der parametrene $\beta_0, \beta_1, \ldots, \beta_p$ og σ^2. Modellen siger da, at den forventede værdi μ_i af y_i er

$$\mu_i = g^{-1}\left(\sum_{j=0}^{p} x_{ij}\beta_j\right)$$

og at variansen på y_i er

$$\mathrm{Var}(y_i) = \sigma^2 \cdot V(\mu_i)/w_i .$$

Denne modelformulering er generel nok til at omfatte en god del interessante statistiske modeller, samtidig med at den indeholder tilstrækkelig meget matematisk struktur til at det er muligt at studere matematisk-statistiske egenskaber ved generaliserede lineære modeller som sådan.[2]

[1] Som i Kapitel 8 er der desuden en underforstået baggrundsvariabel x_0 der konstant er lig 1.

[2] Standardreferencen om generaliseret lineær regression mm. er: P. McCullagh & J.A. Nelder (1983): *Generalized Linear Models*. London: Chapman and Hall.

10.1 Estimation

Hvis man vil prøve at tillempe mindste kvadraters metode til brug i generaliserede lineære regressionsmodeller, støder man på to slags komplikationer:

1. Tilstedeværelsen af linkfunktionen ødelægger den lineære struktur, sådan at estimaterne ikke mere kan findes blot ved at løse nogle lineære ligninger. — Den kvadratsum der skal minimaliseres er ikke længere som i Kapitel 8 af formen $\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2$ men derimod af formen $\sum_{i=1}^{n}\left(y_i - g^{-1}\left(\sum_{j=0}^{p} x_{ij}\beta_j\right)\right)^2$ hvor g er en eller anden ikke-lineær funktion, f.eks. ln.

2. Observationerne har ikke mere (nødvendigvis) samme varians; nu har variansen lov til at afhænge af middelværdien efter nærmere fastsatte regler. Derfor bør man benytte *vægtede* mindste kvadraters metode.

   Det besværlige er blot, at vægtene afhænger af β-erne. Man er derfor i en situation, hvor man for at estimere β-erne korrekt skal kende vægtene, og for at estimere vægtene korrekt skal kende β-erne.

Det er muligt at finde estimaterne med en såkaldt *iterativ metode*, dvs. en metode der består i at man først vælger et sæt begyndelsesværdier og dernæst bliver ved med at anvende en "forbedringsmetode" på disse værdier, indtil "forbedringen" ikke har nogen synlig virkning.

Her følger en beskrivelse af den iterative estimationsmetode man ofte benytter. Først lidt notation:

- $y_1, y_2, \ldots, y_n$ er observationerne,
- $\beta_0, \beta_1, \ldots, \beta_p$ er de parametre der skal estimeres,
- $\beta_0^k, \beta_1^k, \ldots, \beta_p^k$ er de approksimationer til β_j-erne som vi er nået frem til efter k skridt,
- $\mu_1, \mu_2, \ldots, \mu_n$ er de teoretiske forventede værdier (middelværdier), $\mu_i = g^{-1}\left(\sum_{j=0}^{p} x_{ij}\beta_j\right)$,

- $\mu_1^k, \mu_2^k, \ldots, \mu_n^k$ er de approksimationer til $\mu_1, \mu_2, \ldots, \mu_n$ som vi er nået frem til efter k skridt, $\mu_i^k = g^{-1}\left(\sum_{j=0}^{p} x_{ij}\beta_j^k\right)$.

Metoden er opbygget på følgende måde:

1. Som startværdier $\mu_1^0, \mu_2^0, \ldots, \mu_n^0$ benyttes $\mu_i^0 = y_i$, $i = 1, 2, \ldots, n$.
2. Antag at $\mu_1^k, \mu_2^k, \ldots, \mu_n^k$ er de i øjeblikket bedste approksimationer til $\mu_1, \mu_2, \ldots, \mu_n$.
3. Udfør en snedigt valgt vægtet lineær regression der leverer de forbedrede β-værdier $\beta_0^{k+1}, \beta_1^{k+1}, \ldots, \beta_p^{k+1}$:
   - som afhængig variabel ("y") bruges tallene $z_1^k, z_2^k, \ldots, z_n^k$ givet ved
   $$z_i^k = g(\mu_i^k) + (y_i - \mu_i^k)g'(\mu_i^k),$$
   - som forklarende variable bruges x-erne fra det oprindelige problem,
   - som vægte bruges tallene $w_i / (g'(\mu_i^k)^2 V(\mu_i^k))$.
4. Udregn de forbedrede μ-værdier $\mu_i^{k+1} = g^{-1}\left(\sum_{j=0}^{p} x_{ij}\beta_j^{k+1}\right)$.
5. Hvis forskellen mellem β^{k+1}-værdierne og β^k-værdierne er tilstrækkelig lille så slut, ellers gå tilbage til Punkt 2.

Når iterationen er slut, har man fået et sæt estimerede parametre $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ med tilsvarende fittede y-værdier $\widehat{y}_i = \widehat{\mu}_i$, $i = 1, 2, \ldots, n$.

Som begrundelse for i Punkt 3 at foretage netop den beskrevne regression dette: Hvis man tillader sig at opfatte $\mu_1^k, \mu_2^k, \ldots, \mu_n^k$ som konstanter, kan man finde udtryk for dels den forventede værdi, dels variansen af z_i^k. Den forventede værdi af z_i^k bliver $g(\mu_i^k) + (\mu_i - \mu_i^k)g'(\mu_i^k)$, hvilket omtrent er lig med $g(\mu_i) = \sum_{j=0}^{p} x_{ij}\beta_j$. Ifølge regneregler for varianser[3] er variansen af z_i^k lig med $\left(g'(\mu_i^k)\right)^2 \cdot \mathrm{Var}(y_i)$, altså $\sigma^2 \left(g'(\mu_i^k)\right)^2 \cdot V(\mu_i)/w_i$.

[3] se f.eks. Box 5.1

Den pågældende regression er derfor (næsten) den korrekte vægtede regression for at estimere β-erne ud fra z^k-erne.

Den kvadratsum der minimaliseres i regressionen i Punkt 3 er

$$\sum_{i=1}^{n} w_i \frac{\left(z_i^k - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2}{g'(\mu_i^k)^2 V(\mu_i^k)}$$

eller

$$\sum_{i=1}^{n} w_i \frac{\left(g(\mu_i^k) + (y_i - \mu_i^k)g'(\mu_i^k) - g(\mu_i)\right)^2}{g'(\mu_i^k)^2 V(\mu_i^k)}.$$

Når vi her indsætter $\widehat{\mu}$ i stedet for både μ og μ^k får vi en slags residualkvadratsum kaldet *den generaliserede Pearson X^2 størrelse*:

$$X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

Man kan bruge X^2 divideret med antal frihedsgrader $(n-(p+1))$ som et estimat over variansparameteren σ^2.

Man plejer desuden at udregne en såkaldt *deviance* D, hvis systematiske definition vi ikke skal komme ind på her. Der vil som oftest gælde at $D \approx X^2$, og den ene eller den anden af de to størrelser kan benyttes til tests af hvor godt modellen beskriver observationerne.

10.2 ISP-kommandoen `rg_glim`

Generaliseret lineær regression udføres af ISP-kommandoen **`rg_glim`**. Syntaksen for **`rg_glim`** minder en hel del om syntaksen for **`regress`**.

Inputargumenter: **`rg_glim`** skal have mindst to inputargumenter, nemlig et x-array og et y-array, ganske som ved **`regress`**.

Ved binomialfordelte observationer *skal* der være et tredie inputargument indeholdende binomialfordelingens antalsparameter (ofte kaldet n), se Afsnit 10.2.2.

Tabel 10.2: Liste over de fordelinger der kan vælges i `rg_glim` med `/error=` , samt tilhørende kanoniske linkfunktioner.

Fordeling		kanonisk link
navn	`/error=`	
normal	`n`	identiteten
Poisson	`p`	log
binomial	`b`	logit
gamma	`g`	reciprok
inv. Gauss	`i`	inv. squared

Desuden kan der være et array med vægte og et array med offsetværdier[4]. Et "typisk" kald af `rg_glim` med både vægte og offsetværdier se sådan ud:

```
ISP>>rg_glim x y w:vægte of:offset
```

hvor `vægte` og `offset` er to arrays med henholdsvis vægte og offsetværdier.

Fordelinger: Man vælger fordelingstype med parameteren `/error` (default er `/error=n` svarende til normalfordelte observationer).

Tabel 10.2 viser alle de p.t. mulige fordelinger.

Linkfunktioner: Man vælger linkfunktion med parameteren `/link` (default er den såkaldte kanoniske linkfunktion, se Tabel 10.2).

Tabel 10.3 viser alle de p.t. mulige linkfunktioner. Bemærk dog at ikke enhver kombination af linkfunktion og sandsynlighedsfordeling er meningsfuld.

Andre parametre: Som ved `regress` kan man tilføje `/const=n` for at undgå konstantleddet.

[4]Hvis modellens middelværdistruktur er givet ved et udtryk af formen

$$g(\mu_i) \;=\; \alpha_i + \sum_{j=0}^{p} x_{ij}\beta_j$$

hvor α_i-erne er *kendte* tal (og β_j-erne er ukendte parametre som altid), siges $\alpha_1, \alpha_2, \ldots, \alpha_n$ at være *offsetværdier*.

Tabel 10.3: Liste over de linkfunktioner der kan vælges i `rg_glim` med `/link=` .

navn	/link=	tilsv. ISP-udtryk
identiteten	i	y
log	l	log(y)
reciprocal	r	1/y
inverse squared	s	1/y**2
logit	g	log(y/(1-y))
c-log-log	c	log(-log(1-y))
probit	p	gauqu(y)

Derudover er der parametre `/maxits` og `/toler` der kontrollerer iterationen.

Outputargumenter: Man kan angive outputargumenter (efter en `>`) som vil komme til at indeholde de *fittede værdier* og/eller de estimerede parametre og/eller parameterestimaternes *korrelationsmatrix*, f.eks.

```
ISP>>rg_glim x y > fi:yhat co:beta cm:kormat
```

I eksemplet i Afsnit 10.3 ses eksempler på udskrifter fra `rg_glim`, se f.eks. side 152. Bemærk blandt andet, at de estimerede koefficienter findes i søjlen `beta` (og konstantleddet har nummer 0), og at estimatet over σ hedder `s0`. Om brugen af `sdev`, `sdev/s0`, `Pearson X2` og `deviance` se de følgende specialafsnit for de enkelte fordelinger.

10.2.1 Normalfordelte observationer

Hvis man ikke angiver nogen `/error` til `rg_glim`, antages normalfordelte observationer. Standardlinkfunktionen er identiteten. Det vil sige, at et kald af formen

```
ISP>>rg_glim x y
```

fitter den samme model som

```
ISP>>regress x y
```

Hvis man angiver et tredie inputargument, benyttes det som vægte; hvis **vægte** er et sådant vægtarray, kan man skrive

```
ISP>>rg_glim x y vægte
```

eller (med mærkning)

```
ISP>>rg_glim x y w:vægte
```

(Se også Kapitel 9.)

Med `/link=l` kan man fitte normalfordelingsmodeller hvor *logaritmen* til y-erne afhænger lineært af en (eller flere) forklarende variabel x, typisk

```
ISP>>rg_glim /link=l  x y
```

(Man kan godt kombinere brugen af vægte og brugen af linkfunktioner.)

•

I normalfordelingsmodeller er deviance og X^2 det samme, nemlig residualkvadratsummen.

I de fleste tilfælde er variansparameteren σ^2 *ukendt*. I så fald gælder, at σ^2 estimeres ved kvadratet på `s0` og at β-ernes middelfejl er det der står i `rg_glim`-udskriften under **sdev**. **Men** hvis σ^2 er kendt og hvis man vægter med de reciprokke σ^2-er, så vil `s0` være lig 1 pånær tilfældige afvigelser, og β-ernes middelfejl aflæses under `sdev/s0`; endvidere kan man i dette tilfælde benytte deviancen D til et 'goodness-of-fit' test idet nemlig D ikke må ligge for langt ude i χ^2-fordelingen med det pågældende antal frihedsgrader.

10.2.2 Binomialfordelte observationer

Binomialfordelte observationer vælges med `/error=b` . Ved binomialfordelte observationer *skal* der være tre argumenter til `rg_glim`, nemlig arrays med værdierne af x, y og n, hvor meningen altså er, at y_i er binomialfordelt med antalsparameter n_i. Et typisk kald kan derfor se sådan ud:

```
ISP>>rg_glim /error=b  x y n
```

Box 10.1: Binomialfordelingen

En stokastisk variabel Y siges at være *binomialfordelt* med antalsparameter n og sandsynlighedsparameter p, hvis Y har de $n+1$ mulige udfald 0, 1, 2, ..., n og hvis

$$\mathrm{P}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

for $y = 0, 1, 2, \ldots, n$.
Der gælder at $\mathrm{E}Y = np$ og $\mathrm{Var}(Y) = np(1-p)$.
Hvis Y er en sum af n stokastisk uafhængige 01-fordelte variable med samme parameter p, så er Y binomialfordelt med parametre n og p.

Box 10.2: 01-fordelingen

En stokastisk variabel Y siges at være *01-fordelt* med parameter $\theta \in [0,1]$, hvis Y kun har de to mulige udfald 0 og 1, og hvis sandsynligheden for at Y antager værdien 1 er θ, dvs. $\mathrm{P}(Y = 1) = \theta$ og $\mathrm{P}(Y = 0) = 1 - \theta$.

Hvis man skriver

```
ISP>>rg_glim /error=b  x y n > fi:yhat
```

bliver de fittede y-værdier anbragt i **yhat**.

Ved binomialfordelingen er det rent faktisk ikke y_i-erne og deres middelværdier μ_i der bliver modelleret med diverse regressioner, men derimod de relative hyppigheder y_i/n_i og deres middelværdier, nemlig binomialfordelingsparametrene p_i. De linkfunktioner der anvendes ved binomialfordelingen er derfor nogen der afbilder intervallet $]0,1[$ på den reelle akse.

Standardlinkfunktionen er *logit*, dvs. $p \mapsto \ln \frac{p}{1-p}$. Når man benytter den, siger man undertiden at man foretager *logistisk regression*. Den simpleste form for logistisk regression, nemlig med én forklarende variabel x, består således i at man modellerer p's afhængighed af x ved relationen

$$\ln \frac{p}{1-p} = \beta_0 + x\beta_1$$

eller ensbetydende hermed

$$p = \frac{\exp(\beta_0 + x\beta_1)}{1 + \exp(\beta_0 + x\beta_1)}.$$

Andre linkfunktioner der undertiden anvendes er *probitfunktionen*, og den *komplementære log-log-funktion*, se Tabel 10.1 og 10.3.

•

Når alle n_i-erne er store gælder

- Middelfejlen (eller rettere den omtrentlige middelfejl) på β-erne aflæses under `sdev/s0`.
- Hvis modellen er rigtig, er deviancen D og Pearson's X^2 omtrent ens.

  Man kan benytte D (eller X^2) til et 'goodness-of-fit' test, idet D skal ligge centralt i χ^2-fordelingen med det pågældende antal frihedsgrader for at man kan sige at modellen er god nok.
- Man kan endvidere benytte deviancen ved test af statistiske hypoteser: Hvis man har en grundmodel med en deviance D_0 med f_0 frihedsgrader, og hvis man har en hypotese der giver en deviance D_1 med f_1 frihedsgrader, så kan man teste denne hypotese med teststørrelsen $D_1 - D_0$, som ikke må være for stor sammenlignet med χ^2-fordelingen med $f_1 - f_0$ frihedsgrader. Afsnit 10.3 indeholder eksempler herpå.

10.2.3 Poissonfordelte observationer

Poissonfordelte observationer vælges med `/error=p` . Standardlinkfunktionen er log (naturlig logaritme).

Hvis man f.eks. har Poissonfordelte observationer $y_1, y_2, \ldots, y_n$ og ønsker at fitte en model hvor disses middelværdier på en logaritmisk skala afhænger lineært af x, kan det gøres med et kald af formen

```
ISP>>rg_glim /error=p  x y
```

Hvis det derimod er middelværdierne selv der afhænger lineært af x, må man udtrykkeligt angive linkfunktionen:

```
ISP>>rg_glim /error=p  /link=i  x y
```

Box 10.3: Poissonfordelingen

En stokastisk variabel Y siges at være *Poissonfordelt* med parameter μ, hvis de mulige udfald af Y er 0, 1, 2, 3, ... og hvis

$$P(Y = y) = \frac{\mu^y}{y!} \exp(-\mu)$$

for $y = 0, 1, 2, \ldots$.
Der gælder at $EY = \mu$ og $\mathrm{Var}(Y) = \mu$.

Undertiden har man brug for at analysere Poissonfordelingsmodeller hvor den forventede værdi μ_i af y_i er af formen $\mu_i = \lambda_i n_i$, hvor n_i er et kendt tal og hvor det er λ_i der skal modelleres ved hjælp af de forklarende variable. Det kan man gøre ved at tilføje en vektor af n_i-er som et tredie inputargument til `rg_glim`, f.eks.

```
ISP>>rg_glim /error=p  x y n
```

I alle tilfælde kan man få anbragt de fittede y-værdier i et array `yhat` ved at tilføje `> fi:yhat`.

•

Middelfejlen (eller rettere den omtrentlige middelfejl) på β-erne aflæses under `sdev/s0`.

Man kan benytte deviancen D (eller Pearson's X^2) på nøjagtig samme måde som ved binomialfordelingen, se Afsnit 10.2.2.

10.3 Et eksempel

Dette afsnit består af et eksempel på logistisk regression af binomialfordelte observationer. Der er tale om et datasæt der består af to dele, som i første omgang har hver sin logistiske kurve; senere undersøges om kurverne er "parallelle" og om de er sammenfaldende (se eventuelt Afsnit 8.7 for et tilsvarende eksempel i forbindelse med almindelig regression).

I en undersøgelse[5] af insekters reaktion over for insektgiften pyrethrum har man udsat nogle rismelsbiller (*Tribolium castaneum*) for forskellige

[5]Her citeret efter Pack and Morgan (1990): A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **42**, 749–757.

Tabel 10.4: Rismelsbillers overlevelse. Tabellen viser *antal døde / totalantal* for hvert køn og for fire forskellige doser (mg/cm^2).

dosis	M	F
0.20	43 / 144	26 / 152
0.32	50 / 69	34 / 81
0.50	47 / 54	27 / 44
0.80	48 / 50	43 / 47

mængder gift og derpå set hvor mange der var døde efter 13 dages forløb. Resultaterne (i reduceret form) ses i Tabel 10.4.

Lad n_{dk} og y_{dk} betegne henholdsvis totalantal biller og antal døde biller i den gruppe der har fået dosis d og har køn k (= M,F). En simpel modelantagelse er, at der i hver gruppe er tale om en binomialfordelingssituation, således at biller af samme køn og med samme gift-"behandling" har samme chance p_{dk} for at dø og således at de dør uafhængigt af hinanden. Antallet y_{dk} af døde skulle derfor være en observation fra en binomialfordeling med antalsparameter n_{dk} og sandsynlighedsparameter p_{dk}.

Sandsynlighedsparameteren afhænger formodentlig på en eller anden måde af den kvantitative baggrundsvariabel d, og erfaringsmæssigt kan man ofte beskrive sådanne afhængigheder ved en regressionsmodel af formen

$$\text{logit}\, p_{dk} \;=\; \alpha_k + \beta_k \ln d\,, \tag{10.2}$$

dvs. på den logistiske skala afhænger p lineært af logaritmen til dosis. Man kan nu først fitte denne model til data, og derefter kan man undersøge om den kan forsimples. Eksempelvis kan det være fornuftigt at teste om linierne er parallelle, dvs. om $\beta_M = \beta_F$. Hvis det kan accepteres, kan man derefter teste om linierne er sammenfaldende, dvs. om $\alpha_M = \alpha_F$.

Vi vil nu vise hvordan den skitserede undersøgelse kan udføres med ISP og **rg_glim**. Data er allerede tastet ind således at de kan indlæses med **getdata**. Dernæst konstrueres vektorer **dead**, **total** og **lldose** der hver især indeholder otte værdier, først de fire for $k = M$, så de fire for $k = F$. Desuden konstrueres et indikator-array **sex** der angiver om billerne i den pågældende gruppe er hanner eller hunner:

```
ISP>>getdata 'tribol'
Antal døde biller 13 dage efter sprøjtning med pyrethrum.
  Dosis i  DOSE
  Antal døde M hhv. F i  DEADM  hhv.  DEADF
  Totalantal M hhv. F i  TOTALM  hhv.  TOTALF
ISP>>ldose = log(dose)
ISP>>glue deadm deadf > dead    # én lang vektor med DØDE
ISP>>glue totalm totalf > total # do. med TOTALANTAL
ISP>>glue ldose ldose > lldose  # én lang vektor med LOGDOSIS
ISP>>glue 1 1 1 1 0 0 0 0 > sex # indikator for KØN (1=M,0=F)
```

For at fitte modellen (10.2) skal vi bruge den tilsvarende x-matrix (også kaldet designmatrix), her kaldet **xx**; den konstrueres på samme måde som i Afsnit 8.7:

```
ISP>># 1. To forskellige kurver:
ISP>>glue/axis=2  1 0  > sexvalues
ISP>>x = sex==sexvalues
ISP>>glue/axis=2 x (x*lldose) > xx # den ønskede X-matrix
ISP>>print xx
  1          0         -1.609       0
  1          0         -1.139       0
  1          0        -0.6931       0
  1          0        -0.2231       0
  0          1           0         -1.609
  0          1           0         -1.139
  0          1           0        -0.6931
  0          1           0        -0.2231
```

Herefter kaldes `rg_glim` med **xx** som baggrundsvariable. Husk at man skal sætte `/const=n`.

```
ISP>>rg_glim /error=b /const=n xx dead total
GENERALIZED LINEAR REGRESSION   Ver. 2.0
  Error distribution: binomial
  Link function:      logit              log(y/(1-y))
End after 3 iterations.
Number of good obs.: 8,   degrees of freedom: 4

 no.         beta        sdev       sdev/s0
  1          4.270      0.4917      0.5370
  2          2.562      0.3466      0.3785
  3          3.138      0.3528      0.3853
  4          2.582      0.2790      0.3047
```

```
        s0 = 0.9156
Pearson X2 = 3.353
  deviance = 3.364
```

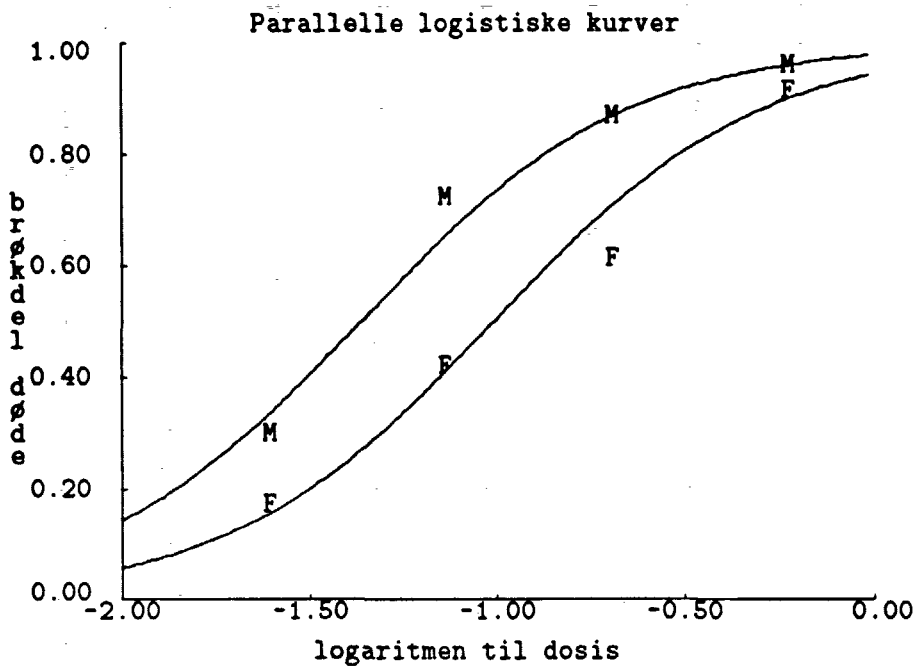
Rækkefølgen af parameterestimaterne er bestemt af rækkefølgen af søjlerne i **xx**, så derfor er

$$\begin{aligned}\widehat{\alpha}_M &= 4.270 \text{ (middelfejl } 0.5372)\\ \widehat{\alpha}_F &= 2.562 \text{ (middelfejl } 0.3785)\\ \widehat{\beta}_M &= 3.138 \text{ (middelfejl } 0.3853)\\ \widehat{\beta}_F &= 2.582 \text{ (middelfejl } 0.3047)\end{aligned}$$

Deviancen $D_0 = 3.364$ er et mål for afvigelsen mellem de oprindelige data og den fittede lineære logistiske model. Der gælder, at hvis modellen er rigtig, så må D_0 ikke være for stor målt i forhold til χ^2-fordelingen med $8 - 4 = 4$ frihedsgrader. Man finder at i χ^2-fordelingen med 4 frihedsgrader er der ca. 50% chance for at få en større D_0-værdi end den faktisk opnåede på 3.364, og det tyder på at der god overensstemmelse mellem den fittede model og de faktiske observationer.

Dernæst fitter vi den model hvor linierne er parallelle, og til det formål skal vi have en ny designmatrix, her kaldet **xxx**:

```
ISP>># 2. Parallelle kurver:
ISP>>glue/axis=2 x lldose > xxx
ISP>>print xxx
  1          0         -1.609
  1          0         -1.139
  1          0        -0.6931
  1          0        -0.2231
  0          1         -1.609
  0          1         -1.139
  0          1        -0.6931
  0          1        -0.2231
ISP>>rg_glim /error=b /const=n xxx dead total
GENERALIZED LINEAR REGRESSION   Ver. 2.0
  Error distribution: binomial
  Link function:      logit              log(y/(1-y))
End after 3 iterations.
Number of good obs.: 8,   degrees of freedom: 5
```



Figur 10.1: Tribolium-eksemplet, den endelige model.

no.	beta	sdev	sdev/s0
1	3.835	0.3335	0.3443
2	2.831	0.3008	0.3105
3	2.813	0.2311	0.2386

s0 = 0.9688
Pearson X2 = 4.693
deviance = 4.670

Det ses at den fælles "hældning" er $\widehat{\beta} = 2.813$ (med en middelfejl på 0.2386) og de to "skæringer" 3.835 og 2.831 (med middelfejl hhv. 0.3443 og 0.3105).

Som teststørrelse for hypotesen $H_1 : \beta_M = \beta_F$ om at kurverne faktisk er parallelle kan man benytte forskellen mellem den nye deviance $D_1 = 4.670$ og grundmodellens deviance D_0, altså $D_1 - D_0 = 1.3$. Hvis hypotesen er rigtig, må $D_1 - D_0$ ikke være for stor sammenholdt med χ^2-fordelingen med $5 - 4 = 1$ frihedsgrader (forskellen mellem de to deviancers frihedsgrader). Man finder at der er omkring 25% chance for at få en større værdi end 1.3, så vi kan sagtens godtage hypotesen om parallelle kurver.

Endelig kan man fitte den model hvor der ikke er forskel på de to køn.

```
ISP>># 3. Ens kurver:
ISP>>rg_glim /error=b lldose dead total
GENERALIZED LINEAR REGRESSION   Ver. 2.0
  Error distribution: binomial
  Link function:      logit              log(y/(1-y))
End after 4 iterations.
Number of good obs.: 8,   degrees of freedom: 6

 no.          beta        sdev        sdev/s0
  0           3.204      0.6989      0.3010
  1           2.709      0.5321      0.2291

         s0 = 2.322
 Pearson X2 = 32.35
   deviance = 32.17
```

Man får da en deviance på 32.17, svarende til en devianceændring på $32.17 - 4.670 = 27.5$. Med én frihedsgrad er sandsynligheden praktisk taget 0 for at få værdier større end 27.5, og man vil derfor forkaste hypotesen om sammenfaldende linier.

Konklusionen bliver at vi kan beskrive sammenhængen mellem dosis d og sandsynligheden p for at dø på den måde, at for hvert køn afhænger logit p lineært af $\ln d$; de to linier er parallelle men ikke sammenfaldende. Den fælles hældningskoefficient estimeres til 2.81 med en standardafvigelse på 0.23, skæringspunktet med linien $\ln d = 0$ estimeres for hanner til 3.84 med en standardafvigelse på 0.34, for hunner til 2.83 med en standardafvigelse på 0.31. Figur 10.1 illustrerer situationen. Figuren er fremstillet på følgende måde:

```
ISP>>gscat ldose (deadm/totalm) /pch=M > figmp
ISP>>gscat ldose (deadf/totalf) /pch=F > figfp
ISP>>x=-iota(array(200))/100
ISP>>zm=3.835+2.813*x
ISP>>gscat x (exp(zm)/(1+exp(zm))) /pty=l > figml
ISP>>zf=2.831+2.813*x
ISP>>gscat x (exp(zf)/(1+exp(zf))) /pty=l > figfl
ISP>>set xlab="logaritmen til dosis"
ISP>>set ylab="brøkdel døde"
ISP>>set title="Parallelle logistiske kurver"
ISP>>gplot figmp figml figfp figfl /xyax=cl
```

10.4 Opgaver

Opgave 10.1: Poissonfordelte data

En bestemt mutation af nogle *Salmonella typhimurium* bakterier har en defekt cellevæg, og det medfører at kemiske stoffer fra cellens omgivelser let kan trænge ind i cellen. Endvidere har mutationen den egenskab at en række kemiske stoffer efter indtrængning meget let kan fremkalde nye mutationer i cellen — sådanne kemiske stoffer kaldes *mutagener*. De nye mutationer er oftest tilbagemutationer, *reversioner*, til vildtypen af bakterien.

Man foretager forsøg (kaldet Ames' test), som består i at fordele bakterier plus kemisk stof i en skål med næringsopløsning og så efter et passende stykke tid at tælle op, hvor mange bakteriekolonier af vildtypen (dvs. af revertanterne) der er.

Der er den særlig interessante omstændighed at kemiske stoffer som er *karcinogene* (dvs. kræftfremkaldende) ofte også er mutagene i Ames' test.

For at undersøge den mutagene effekt af asfaltrøg har man[6] udsat skåle med testbakterier for forskellige doser asfaltrøg. Hver dosis er givet til tre skåle, og på hver skål har man optalt antallet af revertanter. Resultatet heraf ses i Tabel 10.5.

Spørgsmålet er nu, hvordan man kan beskrive "antal revertanter"s afhængighed af dosis.

Det er meget almindeligt at søge at opfatte tælletal som dem der her er tale om som Poissonfordelte. Det er derfor nærliggende at forsøge sig med noget generaliseret lineær regression med Poissonfordelte data. Gør det!

Man kan hævde, at forsøgene med dosis = 0 er væsensforskellige fra de øvrige forsøg og det kunne man bruge som argument for at se bort fra 0-forsøgene. — Det giver i hvert fald pænere data set fra statistikerens synspunkt.

Tip: Data kan indlæses med `getdata` (data-navn `ames`).

[6] A. Elkjær, P. Ingvardsen og T.D. Nielsen (1978): *Undersøgelse af asfaltrøgs mutagenicitet i Salmonella levermikrosom testen.* Projektrapport fra RUC, BIO-OB.

Tabel 10.5: Opgave 10.1: Antal revertanter pr. skål.

dosis (μg pr. skål)	antal revertanter		
0	316	289	295
100	391	403	380
200	409	398	417
400	468	418	410
600	484	474	440
800	489	535	505
1000	590	568	565
2000	706	671	632

Opgave 10.2: Logistisk regression

I biologiske og medicinske sammenhænge udfører man ofte eksperimenter (såkaldte 'bioassays') der skal vise hvordan forskellige koncentrationer af et bestemt stof virker på nogle forsøgsdyr. Eksperimenterne er indrettet på den måde, at der er nogle "ens" forsøgsdyr som er inddelt i et antal grupper. Dyr i samme gruppe får samme behandling, dvs. samme koncentration af stoffet, og forskellige grupper får forskellige koncentrationer. Et vist stykke tid efter at dyrene har fået deres *stimulus* måler man deres *respons*, som meget tit er *død* eller *levende* (altså et binært respons).

I ét sådant eksperiment har man taget nogle mus, der som led i forsøget var smittet med en bestemt slags bakterier, og behandlet dem med forskellige koncentrationer af et antibiotikum, og man har derpå registreret om musene døde eller ej. Der er benyttet syv forskellige koncentrationer, og der er 10 mus i hver gruppe. Man har lavet et antal gentagelser af dette forsøg, og nogle af resultaterne er vist i Tabel 10.6.[7]

Biologerne er især interesseret i at kende den såkaldte LD50, dvs. den koncentration for hvilken der netop er 50% chance for at musen dør. De vil formentlig også gerne vide noget om, hvor præcise de udregnede LD50-værdier er.

Forsøgets opbygning gør det nærliggende at påstå, at antallet af døde mus i en bestemt gruppe er *binomialfordelt* med en antalsparameter n, som er lig antallet af mus i gruppen, og en (ukendt) sandsynlighedsparameter p, som er en bestemt funktion af den dosis d som dyrene har

[7]Data er hentet fra Sanathanan, Gade & Shipkowitz (1987): Trimmed logit method for estimating the ED50 in quantal bioassay. *Biometrics* **43**, 825–832.

Tabel 10.6: Opgave 10.2, bioassay-data: antal døde ud af 10

	Koncentration (mg/kg)						
datasæt	150	75	37.5	18.75	9.38	4.7	2.3
D1	0	0	0	10	9	10	10
D2	0	0	0	3	9	10	9
D3	6	10	10	10	10	9	10
D5	0	0	0	8	10	10	10
D7	2	9	8	10	10	10	10
D10	0	0	0	0	1	4	4

været udsat for. Selve modelbygningsarbejdet består derefter i at finde ud af *hvordan* p afhænger af d.

For det første plejer man ikke at modellere afhængigheden af d men af $u = \ln d$. Man plejer endvidere at benytte en af de to konkurrende modeltyper *probit-modeller* og *logit-modeller* — her vil vi benytte den sidste.[8] Logitmodellen går ud på, at $\text{logit}\, p$ afhænger lineært af u:

$$\text{logit}\, p \;=\; \beta_0 + u\beta_1.$$

Det ses at der er tale om en situation hvor man kan forsøge sig med generaliseret lineær regression af binomialfordelte data, og hvor linkfunktionen i så fald skal være logit.

Gør det:

- Estimér β_0 og β_1 i hvert datasæt for sig.
- Man har en formodning om at parametrene β_0 og β_1 er de samme i alle datasættene, så derfor vil man gerne estimere den fælles værdi. Det kunne man forestille sig at gøre på to forskellige måder:

  1. man kunne sige at der er 42 sammenhørende værdier af logkoncentration og "antal døde ud af 10" og så lave logistisk regression på det.

[8] Logitmodeller kan man læse om f.eks. i Cox (1970): *The Analysis of Binary Data*. London: Methuen. En klassisk reference til probitmodeller er Finney (1971): *Probit Analysis*. 3. udgave. Cambridge: Cambridge University Press.

2. man kunne sige at der er 7 sammenhørende værdier af log-koncentration og "antal døde ud af 60" (nemlig samtlige 60 dyr der har fået den pågældende koncentration), og så lave logistisk regression på det.

Prøv begge metoder.

Metoderne giver *ikke* samme udskrift fra `rg_glim`; hvad er forskellen? Hvad er grunden til forskellen?

- Lav tegninger indeholdende datapunkterne plus de(n) fittede kurve(r).
- Udregn LD50-værdierne.
- Hvad kan man sige om den præcision (f.eks. middelfejl) hvormed de forskellige estimater er bestemt? [9]

Tip: Data kan indlæses til ISP med kommandoen **getdata** (data-navn **assay**).

Opgave 10.3

Hvis man vil finde ud af hvor smitsom en bestemt sygdom er, kan man ideelt set gøre det ved at undersøge en række personer der i forskellig grad er blevet udsat for smitten og se hvor mange af dem der rent faktisk har fået sygdommen. Det byder på en række praktiske problemer, eksempelvis er det ikke altid så let at få et realistisk skøn over det antal gange en person er blevet udsat for den pågældende smitte.

I en undersøgelse[10] af HIVs smitsomhed ved seksuel kontakt har man data om 159 heteroseksuelle par, hvor det er blevet konstateret at manden er blevet smittet med HIV; man har så undersøgt om virus også er blevet overført til hans kvindelige partner, og man har et skøn over antallet af seksuelle kontakter mellem de to fra det formodede smittetidspunkt indtil undersøgelsestidspunktet, se Tabel 10.7.

Man kan opstille en simpel og måske også lidt naiv model som følger: Antag at der ved hver kontakt mellem den smittede mand og hans usmittede partner er sandsynligheden λ for at virus overføres. Så er der altså sandsynligheden $1-\lambda$ for at smitten ikke overføres ved én kontakt, og hvis de enkelte kontakter er uafhængige af hinanden i smittemæssig

[9] Benyt evt. Appendiks B i forbindelse med beregning af middelfejlen på LD50.

[10] *California Partners' Study*, her citeret efter Jewell and Shiboski (1990): Statistical analysis of HIV infectivity based on partner studies. *Biometrics* **46**, 1133-1150.

Tabel 10.7: Opgave 10.3: Antal HIV-smittede y og antal undersøgte personer n i forskellige grupper bestemt ved antal seksuelle kontakter k.

k	y	n
0 - 9	2	24
10 - 49	6	26
50 - 99	2	20
100 - 199	3	21
200 - 299	8	21
300 - 399	3	10
400 - 599	8	14
600 - 799	2	11
800 - 1499	2	8
1500 - 2170	2	4

henseende, så er sandsynligheden for at kvinden efter k kontakter ikke er smittet lig med $(1-\lambda)^k$, dvs. sandsynligheden for at en kvinde der har haft k kontakter med sin smittede partner selv er blevet smittet er

$$p_k = 1-(1-\lambda)^k. \tag{10.3}$$

For hver "kontakt-gruppe" skulle antallet y af smittede derfor være binomialfordelt med antalsparameter n og en sandsynlighedsparameter $p_k = 1-(1-\lambda)^k$ der afhænger af baggrundsvariablen k og smitsomhedsparameteren λ. Anvend nu den såkaldte komplementære log-log funktion på begge sider af (10.3) og få

$$\begin{aligned} \ln(-\ln(1-p_k)) &= \ln(-\ln(1-\lambda)) + \ln k \\ &= \beta_0 + \ln k, \end{aligned}$$

hvor $\beta_0 = \ln(-\ln(1-\lambda))$. Det viser at der er tale om en generaliseret lineær model med binomialfordelte fejl, den komplementære log-log funktion som link og $\ln k$ som forklarende variabel.

1. Fit den mere generelle model

$$\ln(-\ln(1-p_k)) = \beta_0 + \beta_1 \ln k,$$

og lav passende tegninger af observerede og fittede værdier (f.eks. med $\ln k$ ud ad den vandrette akse og $\ln(-\ln(1-p))$ ud ad den lodrette).

2. Den oprindelige model svarer til at $\beta_1 = 1$. Man kunne derfor være interesseret i at teste hypotesen $\beta_1 = 1$. Hvordan kan man gøre det? Hvordan kan man estimere β_0 hvis $\beta_1 = 1$?

   Hvordan kan man forklare resultatet?

Tip: Data kan indlæses til ISP med kommandoen **getdata** (data-navn **aids**).

Appendiks A

Regressionsanalyse i lineær algebra-sprog

Man kan formulere en stor klasse af statistiske modeller for normalfordelte observationer i lineær algebra-sprog. Derved kan man "forstå" de statistiske begreber ved hjælp af begreber fra lineær algebra (eller omvendt), og meget bliver lettere at overskue.

Den grundlæggende nye tanke er, at i stedet for at tænke på observationerne som en stribe tal $y_1, y_2, \ldots, y_n$, tænker man på dem som et talsæt, dvs. en vektor $\boldsymbol{y}$ i $\mathbb{R}^n$.

Grundmodellen

Den generelle multiple lineære regressionsanalysemodel (8.2) på side 94 kan skrives

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A.1}$$

hvor $\boldsymbol{y}$ er en vektor (søjlematrix) i $\mathbb{R}^n$ indeholdende y-værdierne, parametervektoren $\boldsymbol{\beta}$ er en vektor (søjlematrix) i $\mathbb{R}^p$ indeholdende β-værdierne, designmatricen $\boldsymbol{X}$ er en $n \times p$-matrix hvis j-te søjle indeholder værdierne af den j-te forklarende variabel, og $\boldsymbol{\varepsilon}$ er en vektor i $\mathbb{R}^n$ indeholdende residualerne. Bemærk at konstantleddet (der sædvanligvis hedder β_0) her kommer ind i billedet ved at den første søjle i $\boldsymbol{X}$ bliver sat til at indeholde lutter ettaller (og i nærværende formulering kommer konstantleddet så til at hedde β_1).

Estimation af β

Den kvadratsum der skal minimaliseres er $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$. Når $\boldsymbol{\beta}$ gennemløber $\mathbb{R}^p$, vil $\boldsymbol{X}\boldsymbol{\beta}$ gennemløbe underrummet $L = \{\boldsymbol{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}$ af $\mathbb{R}^p$. (Dimensionen af L er lig med rangen af $\boldsymbol{X}$.) Opgaven at minimalisere kvadratsummen $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ er derfor identisk med opgaven at bestemme de(t) punkt(er) $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ i L der ligger tættest på $\boldsymbol{y}$. Fra den lineære algebra vides at der altid findes præcis et punkt $\widehat{\boldsymbol{y}}$ med denne egenskab og at dette punkt kan findes ved at projicere $\boldsymbol{y}$ ortogonalt ned på L, dvs. det er entydigt bestemt af betingelserne $\widehat{\boldsymbol{y}} \in L$ og $\boldsymbol{y} - \widehat{\boldsymbol{y}} \perp L$.

Nu er $\boldsymbol{y} - \widehat{\boldsymbol{y}} \perp L$ ensbetydende med at $(\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = 0$ for ethvert $\boldsymbol{\beta}$, og det er igen ensbetydende med at $\boldsymbol{X}'(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = 0$, dvs. med at

$$\boldsymbol{X}'\widehat{\boldsymbol{y}} = \boldsymbol{X}'\boldsymbol{y}. \tag{A.2}$$

Dette er estimationsligningerne (8.3) fra side 95; de kaldes ofte også for *normalligningerne*, eftersom de altså udtrykker at $\boldsymbol{y} - \widehat{\boldsymbol{y}} \perp L$.

Bemærk i øvrigt at omskrivningen (8.15) på side 115, der i vektornotation lyder

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 + \|\widehat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

blot er Pythagoras' formel, idet $\boldsymbol{y} - \widehat{\boldsymbol{y}} \perp L$ og $\widehat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} \in L$.

Om mængden af løsninger $\widehat{\boldsymbol{\beta}}$ gælder

1. Hvis $\boldsymbol{X}$ har fuld rang (dvs. rang p) bliver L parametriseret bijektivt ved $\boldsymbol{\beta} \mapsto \boldsymbol{X}\boldsymbol{\beta}$, og så er der et entydigt bestemt $\widehat{\boldsymbol{\beta}}$ således at $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

   Dette $\widehat{\boldsymbol{\beta}}$ bestemmes således: Indsæt $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ i (A.2) og få $(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$. Da $\boldsymbol{X}$ har fuld rang er $\boldsymbol{X}'\boldsymbol{X}$ regulær, så

   $$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

   Der gælder at $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ og $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$.

2. Hvis $\boldsymbol{X}$ ikke har fuld rang skyldes det at en af søjlerne er en linearkombination af de øvrige, det vil sige der er en af de forklarende variable der ikke indeholder information der ikke allerede er indeholdt i de øvrige (man siger så at der er tale om *multicollinearitet*). I dette tilfælde bliver L ikke parametriseret injektivt, og der er uendelig mange løsninger $\widehat{\boldsymbol{\beta}}$ til normalligningerne.

Box A.1: Middelværdi og varians af stokastiske vektorer
Hvis $\boldsymbol{y}$ er en stokastisk vektor med værdier i $\mathbb{R}^n$, så er dens *middelværdivektor* den vektor $\mathrm{E}\boldsymbol{y}$ i $\mathbb{R}^n$ hvis i-te koordinat er middelværdien $\mathrm{E}(y_i)$ af den i-te koordinat y_i af $\boldsymbol{y}$. Endvidere er *variansmatricen* for $\boldsymbol{y}$ den $n \times n$-matrix $\mathrm{Var}(\boldsymbol{y})$ hvis (i,j)-te element er kovariansen mellem y_i og y_j, i særdeleshed er det i-te diagonalelement i variansmatricen lig med variansen af den i-te koordinat y_i.
Hvis $\boldsymbol{A}$ er en konstant matrix med n søjler, så gælder følgende regneregler:

$$\begin{aligned} \mathrm{E}(\boldsymbol{A}\boldsymbol{y}) &= \boldsymbol{A}\,\mathrm{E}\boldsymbol{y} \\ \mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) &= \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{y})\,\boldsymbol{A}'. \end{aligned}$$

Man kan så fjerne det fornødne antal søjler fra $\boldsymbol{X}$ eller pålægge parametervektoren nogle lineære bånd, således at man kan estimere entydigt (hvis man er interesseret i det).

Estimation af σ^2

Variansen σ^2 estimeres (under alle omstændigheder) ved

$$s^2 = \frac{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2}{n - \dim L}.$$

Hvis $\boldsymbol{X}$ har fuld rang, er $\dim L$ lig med p.

Residualer

Man kalder tit den projektionsmatrix der afbilder $\boldsymbol{y}$ over i $\widehat{\boldsymbol{y}}$ for $\boldsymbol{H}$ ("hat-matricen"): $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. Når $\boldsymbol{X}$ har fuld rang er

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

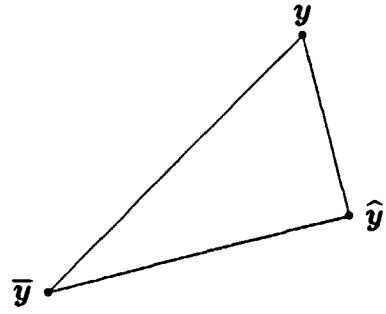
Da $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ er residualvektoren $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, hvor $\boldsymbol{I}$ er $n \times n$-enhedsmatricen.

Determinationskoefficienten R^2

Antag at vektoren $\mathbf{1}$ tilhører L (vektoren $\mathbf{1}$ er den vektor der består af lutter ettaller). Lad $\overline{\boldsymbol{y}}$ betegne vektoren $\overline{y}\mathbf{1}$. Så er

$$\boldsymbol{y} - \overline{\boldsymbol{y}} = (\boldsymbol{y} - \widehat{\boldsymbol{y}}) + (\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}})$$

en opspaltning af $\boldsymbol{y} - \overline{\boldsymbol{y}}$ i en sum af to ortogonale vektorer.



Kvadratet på cosinus til vinklen mellem $\boldsymbol{y} - \overline{\boldsymbol{y}}$ og $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$ kaldes for R^2. Denne cosinus er dels lig med skalarproduktet af de to vektorer divideret med deres længder, dels er den længden af den hosliggende katete divideret med længden af hypotenusen i den retvinklede trekant. Det giver to forskellige udtryk for R^2:

$$R^2 = \frac{\left[(\boldsymbol{y} - \overline{\boldsymbol{y}})'(\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}})\right]^2}{\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2 \, \|\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2},$$

og

$$R^2 = \frac{\|\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2}.$$

Dette er de to formler (8.5) og (8.6) på side 99.

Vægtede mindste kvadrater

Ved vægtede mindste kvadraters metode minimaliserer man en kvadratisk form

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

hvor $\boldsymbol{W}$ er en positiv definit symmetrisk $n \times n$-matrix, f.eks. en diagonalmatrix ($\boldsymbol{W} = \text{diag}(w_1, w_2, \ldots, w_n)$). Der findes en entydigt bestemt symmetrisk positiv definit matrix $\boldsymbol{W}^{1/2}$ således at $\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2} = \boldsymbol{W}$. Derved får man omskrivningen

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= \left[\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]' \left[\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] \\ &= \|\boldsymbol{W}^{1/2}\boldsymbol{y} - (\boldsymbol{W}^{1/2}\boldsymbol{X})\boldsymbol{\beta}\|^2.\end{aligned}$$

Herved er problemet blevet omformet til et problem der består i at minimalisere en kvadratsum af formen $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, hvor $\boldsymbol{y}$ er erstattet af $\boldsymbol{W}^{1/2}\boldsymbol{y}$ og $\boldsymbol{X}$ er erstattet af $\boldsymbol{W}^{1/2}\boldsymbol{X}$. Vi kan derfor uden videre sige at løsningerne skal findes som løsningerne til estimationsligningen (normalligningen) $\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{y}} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$.

Hvis $\boldsymbol{X}$ har fuld rang er $\hat{\boldsymbol{\beta}}$ entydigt bestemt og givet ved

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left[(\boldsymbol{W}^{1/2}\boldsymbol{X})'(\boldsymbol{W}^{1/2}\boldsymbol{X})\right]^{-1} (\boldsymbol{W}^{1/2}\boldsymbol{X})'\boldsymbol{W}^{1/2}\boldsymbol{y} \\ &= [\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}.\end{aligned}$$

•

Vægtede mindste kvadrater benyttes blandt andet når observationerne $y_1, y_2, \ldots, y_n$ har kendte varianser $v_1, v_2, \ldots, v_n > 0$ eller mere generelt når variansmatricen (også kaldet dispersionsmatricen) for $\boldsymbol{y}$ er en kendt matrix $\boldsymbol{D}$. Så benyttes vægtmatricen

$$\boldsymbol{W} = \text{diag}\left(\frac{1}{v_1}, \frac{1}{v_2}, \ldots, \frac{1}{v_n}\right)$$

og i det generelle tilfælde $\boldsymbol{W} = \boldsymbol{D}^{-1}$. Derved opnås nemlig at

$$\text{Var}(\boldsymbol{W}^{1/2}\boldsymbol{y}) = \boldsymbol{I},$$

dvs. de enkelte komponenter i $\boldsymbol{W}^{1/2}\boldsymbol{y}$ bliver ukorrelerede og får varians 1. Under den yderligere antagelse at observationerne er normalfordelte gælder da, at den minimaliserede residualkvadratsum $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ er χ^2-fordelt med $f = n - p$ frihedsgrader. Dette kan udnyttes til et test for modellens brugbarhed som i Afsnit 9.3, punkt 2.

Vægtede mindste kvadraters metode benyttes også når variansmatricen for $\boldsymbol{y}$ er kendt pånær en konstant faktor, dvs.

$$\mathrm{Var}(\boldsymbol{y}) = \sigma^2 \boldsymbol{D},$$

hvor $\boldsymbol{D}$ er en kendt matrix og σ^2 er en ukendt positiv parameter. Også her benyttes $\boldsymbol{W} = \boldsymbol{D}^{-1}$, og nu er $\mathrm{Var}(\boldsymbol{W}^{1/2}\boldsymbol{y}) = \sigma^2 \boldsymbol{I}$. Som skøn over σ^2 bruges

$$s^2 = \frac{1}{f}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

hvor $f = n - p$ er antallet af frihedsgrader.

A.1 Opgaver

Opgave A.1: Simpel lineær regression

Gør rede for at den simple lineære regressionsmodel fra Kapitel 2 fremkommer når designmatricen $\boldsymbol{X}$ er en $n \times 2$-matrix hvis første søjle er lutter ettaller og hvis anden søjle indeholder værdierne $x_1, x_2, \ldots, x_n$ af den forklarende variabel.

Eftervis at ligningen $\boldsymbol{X}'\widehat{\boldsymbol{y}} = \boldsymbol{X}'\boldsymbol{y}$ er det samme som estimationsligningerne (2.5) og (2.6) på side 16 (eller (2.3) og (2.4) på side 15).

Den generelle formel for $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ er $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$ (side 164) . Benyt den til at eftervise formlerne på side 67 for variansen på hhv. $\widehat{\beta}_0$ og $\widehat{\beta}_1$.

Lad $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ være residualvektoren (som altså er projektionen på det ortogonale komplement til underrummet udspændt af søjlerne i $\boldsymbol{X}$). Opskriv et udtryk for $\mathrm{Var}(\boldsymbol{e})$, og eftervis påstanden i Afsnit 5.1 om at variansen på e_i er $(1 - h_i)\sigma^2$ hvor h_i er givet ved formel (5.3).

Appendiks B

Delta-metoden til vurdering af usikkerheden på en funktion af stokastiske variable

De forskellige regressionsmetoder beregner automatisk den estimerede middelfejl på de beregnede parametre. Imidlertid er man ofte ude for at det ikke er selve de beregnede parametre der er interessante i den praktiske problemstilling, men bestemte funktioner af dem, f.eks. en af dem divideret med en anden, og hvad er så middelfejlen på den størrelse?

Antag at man har to parameterskøn $\widehat{\alpha}$ og $\widehat{\beta}$ og at man er interesseret i middelfejlen på $f(\widehat{\alpha},\widehat{\beta})$ hvor f er en vis funktion. Man kan få et overslag over middelfejlen på følgende måde.

Først Taylor-udvikles funktionen f omkring de sande parameterværdier, vi tager dog kun førsteordensleddene med:

$$f(\widehat{\alpha},\widehat{\beta}) \approx f(\alpha,\beta) + (\widehat{\alpha}-\alpha)\frac{\partial f}{\partial \alpha} + (\widehat{\beta}-\beta)\frac{\partial f}{\partial \beta}.$$

Heraf følger ved brug af de almindelige regneregler for varianser (se f.eks. Box 5.1) at

$$\mathrm{Var} f(\widehat{\alpha},\widehat{\beta}) \approx \left(\frac{\partial f}{\partial \alpha}\right)^2 \mathrm{Var}(\widehat{\alpha}) + \left(\frac{\partial f}{\partial \beta}\right)^2 \mathrm{Var}(\widehat{\beta}) + 2\frac{\partial f}{\partial \alpha}\frac{\partial f}{\partial \beta}\mathrm{Cov}(\widehat{\alpha},\widehat{\beta}),$$

og ved at tage kvadratroden heraf får man et skøn over middelfejlen på $f(\widehat{\alpha}, \widehat{\beta})$.

Størrelserne $\mathrm{Var}(\widehat{\alpha})$ og $\mathrm{Var}(\widehat{\beta})$ er ganske enkelt kvadraterne på middelfejlene på $\widehat{\alpha}$ og $\widehat{\beta}$, og kovariansen $\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta})$ kan udregnes som produktet af middelfejlen på $\widehat{\alpha}$, middelfejlen på $\widehat{\beta}$ og korrelationen mellem $\widehat{\alpha}$ og $\widehat{\beta}$. — ISP's `regress` kan bringes til at returnere en korrelationsmatrix indeholdende korrelationerne mellem de forskellige parameterskøn, se *Introduktion til ISP*.

Eksempel: Hvis $f(\alpha, \beta) = \alpha/\beta$, så er $\partial f/\partial\alpha = 1/\beta$ og $\partial f/\partial\beta = -\alpha/\beta^2$, og dermed er kvadratet på middelfejlen på $\widehat{\alpha}/\widehat{\beta}$

$$\frac{1}{\beta^2}\mathrm{Var}(\widehat{\alpha}) + \frac{\alpha^2}{\beta^4}\mathrm{Var}(\widehat{\beta}) - \frac{2\alpha}{\beta^3}\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta})$$

eller

$$\frac{\beta^2\mathrm{Var}(\widehat{\alpha}) + \alpha^2\mathrm{Var}(\widehat{\beta}) - 2\alpha\beta\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta})}{\beta^4}.$$

Liste over boxe

Stikord

Liste over tidligere udkomne tekster tilsendes gerne. Henvendelse herom kan ske til IMFUFA's sekretariat tlf. 46 75 77 11 lokal 2263

217/92 "Two papers on APPLICATIONS AND MODELLING IN THE MATHEMATICS CURRICULUM"
by: Mogens Niss

218/92 "A Three-Square Theorem"
by: Lars Kadison

219/92 "RUPNOK - stationær strømning i elastiske rør"
af: Anja Boisen, Karen Birkelund, Mette Olufsen
Vejleder: Jesper Larsen

220/92 "Automatisk diagnosticering i digitale kredsløb"
af: Bjørn Christensen, Ole Møller Nielsen
Vejleder: Stig Andur Pedersen

221/92 "A BUNDLE VALUED RADON TRANSFORM, WITH APPLICATIONS TO INVARIANT WAVE EQUATIONS"
by: Thomas P. Branson, Gestur Olafsson and Henrik Schlichtkrull

222/92 On the Representations of some Infinite Dimensional Groups and Algebras Related to Quantum Physics
by: Johnny T. Ottesen

223/92 *THE FUNCTIONAL DETERMINANT*
by: Thomas P. Branson

224/92 *UNIVERSAL AC CONDUCTIVITY OF NON-METALLIC SOLIDS AT LOW TEMPERATURES*
by: Jeppe C. Dyre

225/92 *"HATMODELLEN" Impedansspektroskopi i ultrarent en-krystallinsk silicium*
af: Anja Boisen, Anders Gorm Larsen, Jesper Varmer, Johannes K. Nielsen, Kit R. Hansen, Peter Bøggild og Thomas Hougaard
Vejleder: Petr Viscor

226/92 *"METHODS AND MODELS FOR ESTIMATING THE GLOBAL CIRCULATION OF SELECTED EMISSIONS FROM ENERGY CONVERSION"*
by: Bent Sørensen

227/92 "Computersimulering og fysik"
af: Per M.Hansen, Steffen Holm, Peter Maibom, Mads K. Dall Petersen, Pernille Postgaard, Thomas B.Schrøder, Ivar P. Zeck
Vejleder: Peder Voetmann Christiansen

228/92 "Teknologi og historie"
Fire artikler af:
Mogens Niss, Jens Høyrup, Ib Thiersen, Hans Hedal

229/92 "Masser af information uden betydning"
En diskussion af informationsteorien i Tor Nørretranders' "Mærk Verden" og en skitse til et alternativ basseret på andenordens kybernetik og semiotik.
af: Søren Brier

230/92 "Vinklens tredeling - et klassisk problem"
et matematisk projekt af
Karen Birkelund, Bjørn Christensen
Vejleder: Johnny Ottesen

231A/92 "Elektrondiffusion i silicium - en matematisk model"
af: Jesper Voetmann, Karen Birkelund, Mette Olufsen, Ole Møller Nielsen
Vejledere: Johnny Ottesen, H.B.Hansen

231B/92 "Elektrondiffusion i silicium - en matematisk model" Kildetekster
af: Jesper Voetmann, Karen Birkelund, Mette Olufsen, Ole Møller Nielsen
Vejledere: Johnny Ottesen, H.B.Hansen

232/92 "Undersøgelse om den simultane opdagelse af energiens bevarelse og især om de af Mayer, Colding, Joule og Helmholtz udførte arbejder"
af: L.Arleth, G.I.Dybkjær, M.T.Østergård
Vejleder: Dorthe Posselt

233/92 "The effect of age-dependent host mortality on the dynamics of an endemic disease and
Instability in an SIR-model with age-dependent susceptibility
by: Viggo Andreasen

234/92 "THE FUNCTIONAL DETERMINANT OF A FOUR-DIMENSIONAL BOUNDARY VALUE PROBLEM"
by: Thomas P. Branson and Peter B. Gilkey

235/92 OVERFLADESTRUKTUR OG POREUDVIKLING AF KOKS
- Modul 3 fysik projekt -
af: Thomas Jessen

236a/93 INTRODUKTION TIL KVANTE HALL EFFEKTEN

af: Ania Boisen, Peter Bøggild

Vejleder: Peder Voetmann Christiansen
Erland Brun Hansen

236b/93 STRØMSSAMMENBRUD AF KVANTE HALL EFFEKTEN

af: Anja Boisen, Peter Bøggild

Vejleder: Peder Voetmann Christiansen
Erland Brun Hansen

237/93 The Wedderburn principal theorem and Shukla cohomology

af: Lars Kadison

238/93 SEMIOTIK OG SYSTEMEGENSKABER (2)

Vektorbånd og tensorer

af: Peder Voetmann Christiansen

239/93 Valgsystemer - Modelbygning og analyse

Matematik 2. modul

af: Charlotte Gjerrild, Jane Hansen, Maria Hermannsson, Allan Jørgensen, Ragna Clauson-Kaas, Poul Lützen

Vejleder: Mogens Niss

240/93 Patologiske eksempler.
Om sære matematiske fisks betydning for den matematiske udvikling

af: Claus Dræby, Jørn Skov Hansen, Runa Ulsøe Johansen, Peter Meibom, Johannes Kristoffer Nielsen

Vejleder: Mogens Niss

241/93 FOTOVOLTAISK STATUSNOTAT 1

af: Bent Sørensen

242/93 Brovedligeholdelse - bevar mig vel

Analyse af Vejdirektoratets model for optimering af broreparationer

af: Linda Kyndlev, Kare Fundal, Kamma Tulinius, Ivar Zeck

Vejleder: Jesper Larsen

243/93 TANKEEKSPERIMENTER I FYSIKKEN

Et 1.modul fysikprojekt

af: Karen Birkelund, Stine Sofia Korremann

Vejleder: Dorthe Posselt

244/93 RADONTRANSFORMATIONEN og dens anvendelse i CT-scanning

Projektrapport

af: Trine Andreasen, Tine Guldager Christiansen, Nina Skov Hansen og Christine Iversen

Vejledere: Gestur Olafsson og Jesper Larsen

245a+b /93 Time-Of-Flight målinger på krystallinske halvledere
Specialerapport

af: Linda Szkotak Jensen og Lise Odgaard Gade

Vejledere: Petr Viscor og Niels Boye Olsen

246/93 HVERDAGSVIDEN OG MATEMATIK
- LÆREPROCESSER I SKOLEN

af: Lena Lindenskov, Statens Humanistiske Forskningsråd, RUC, IMFUFA

247/93 UNIVERSAL LOW TEMPERATURE AC CONDUCTIVITY OF MACROSCOPICALLY DISORDERED NON-METALS

by: Jeppe C. Dyre

248/93 DIRAC OPERATORS AND MANIFOLDS WITH BOUNDARY

by: B. Booss-Bavnbek, K.P.Wojciechowski

249/93 Perspectives on Teichmüller and the Jahresbericht Addendum to Schappacher, Scholz, et al.

by: B. Booss-Bavnbek

With comments by W.Abikoff, L.Ahlfors, J.Cerf, P.J.Davis, W.Fuchs, F.P.Gardiner, J.Jost, J.-P.Kahane, R.Lohan, L.Lorch, J.Radkau and T.Söderqvist

250/93 EULER OG BOLZANO - MATEMATISK ANALYSE SET I ET VIDENSKABSTEORETISK PERSPEKTIV

Projektrapport af: Anja Juul, Lone Michelsen, Tomas Højgård Jensen

Vejleder: Stig Andur Pedersen

251/93 Genotypic Proportions in Hybrid Zones

by: Freddy Bugge Christiansen, Viggo Andreasen and Ebbe Thue Poulsen

252/93 MODELLERING AF TILFÆLDIGE FÆNOMENER

Projektrapport af: Birthe Friis, Lisbeth Helmgaa Kristina Charlotte Jakobsen, Marina Mosbæk Johannesssen, Lotte Ludvigsen, Mette Haas Nielsen