Center for Statistik
Handelshøjskolen i København

Tue Tjur, Maj 2007
MPAS

VEJLEDNING I RAPPORTSKRIVNING

Generelt. Kursets formål er at sætte deltagerne i stand til — ved hjælp af diverse computerprogrammer og de teoretiske forudsætninger, man har erhvervet sig ved andre kurser —

(1) at gennemføre en statistisk analyse af et datamateriale

og

(2) sammenfatte resultaterne i en rapport.

Det er det sidste punkt, vi skal beskæftige os med her.

Klienten. I praksis skal en rapportforfatter naturligvis skrive på et niveau, der svarer til læserens baggrund. I nærværende undervisningssituation, hvor den virkelige læsers niveau i al beskedenhed er ganske højt, vil vi arbejde med en fiktiv læser kaldet "klienten". Klienten er den der har rekvireret og betalt den statistiske undersøgelse for at bruge den som grundlag for sine beslutninger. Man må forestille sig klienten som en person, der ved alt hvad der er værd at vide om hvordan data er fremkommet og om sit eget fagområde. Vi vil også — for ikke at gøre rapportskrivningen til en uoverkommelig opgave — gå ud fra, at klienten er i besiddelse af almen matematisk/statistisk dannelse. For at præcisere niveauet lidt nøjere, giver vi et par eksempler. Eksemplerne skal forstås sådan at en jargon, der er mere teknisk–matematisk end den der bruges her, så vidt muligt bør undgås.

Eksempel 1. Antallet af positive udfald blandt de N_g responser i gruppe g betegnes y_g. En simpel model går ud på at antage y_g binomialfordelt (svarende til at de N_g individer opfører sig uafhængigt, med samme sandsynlighed for en positiv respons) med antalsparameter N_g og sandsynlighedsparameter

$$p_g = \frac{\exp(\alpha + \beta x_g)}{1 + \exp(\alpha + \beta x_g)}.$$

Denne ligning kan også skrives

$$\text{logit}(p_g) = \alpha + \beta x_g,$$

hvor funktionen logit er defineret ved $\text{logit}(p) = \log(p/(1-p))$. I denne model (den logistiske regressionsmodel) er den afgørende parameter β, som bestemmer responsens afhængighed af eksponeringstiden x_g. Derfor vil vi først og fremmest interessere os for estimation af β og test for hypotesen $\beta = 0$ (svarende til "ingen afhængighed af eksponeringstiden x_g").

Eksempel 2. Vi antager observationerne y_{rs} normalfordelte med middelværdier

$$\mathrm{E}y_{rs} = \alpha_r + \beta_s$$

og samme varians σ^2. Med denne grundmodel (den additive model) som udgangspunkt kan vi teste hypotesen om manglende søjle– (dvs. behandlings–) virkning,

$$\mathrm{E}y_{rs} = \alpha_r.$$

Teststørrelsen

$$F(3, 12) = 3.403$$

(der kan fortolkes som et kvadratisk mål for afvigelsen mellem søjlegennemsnittene, relativt til variansen i grundmodellen) vil under hypotesen følge en F–fordeling med (3,12) frihedsgrader. Da en teststørrelse med denne fordeling er større end 3.403 med sandsynlighed lidt over 5% kan hypotesen ikke afvises på dette grundlag. Estimaternes indbyrdes ordning kan derfor ikke med sikkerhed tages som udtryk for en tilsvarende forskel mellem de fire produktionsmetoder.

Eksempel 3. ... Vi forsøger derfor med en multiplikativ Poissonmodel hvor kun faktorerne `ALDGRP`, `KØN` og `SOCGRP` indgår. Her viser det sig, at trefaktorvekselvirkningen er insignifikant (P=0.42), og tofaktorvekselvirkningerne mellem `SOCGRP` og de to andre faktorer kan fjernes ($P > 0.1$ i alle tilfælde, uanset testrækkefølge). Vekselvirkningen mellem `ALDGRP` og `KØN` er derimod signifikant (P=0.00031), og det samme gælder hovedvirkningen af `SOCGRP` (P=0.000007). Vores endelig model får således formen

$$\mathrm{E}\left(y_{aks}\right) = \exp\left(\alpha_{ak} + \beta_s\right)$$

hvor a, k og s betegner henholdsvis aldersgruppe, køn og socialgruppe. Estimaterne i denne model bliver, idet β_1 sættes lig med 0 for at undgå overparametrisering, ...

Typiske begynderfejl i statistisk rapportskrivning er

(1) Undervurdering af klientens faglige niveau i hans/hendes eget fag, i form af arrogante bemærkninger om hvordan undersøgelsen burde være udført, den manglende matematiske modelbygning etc.,

(2) Overvurdering af klientens matematiske baggrund, ofte kombineret med undervurdering af klientens evne til at forstå en statistisk model, når den forklares tilstrækkeligt omhyggeligt på det relevante niveau.

Et af kursets formål er at tage brodden af de værste brølere i den retning — bl. a. derfor den fiktive "klient".

En anden typisk begynderfejl går ud på at bruge et unødvendigt opstyltet sprog, hvor de statistiske termer undertiden indgår hulter til bulter. Sørg for at gøre sproget så enkelt som muligt, og sørg også for at de statistiske begreber optræder i den sammenhæng hvor de hører hjemme.

Eksempler: Der er ikke noget som hedder at *modellen er signifikant*. Signifikans er noget der kan optræde i forbindelse med et test, hvor man så siger at *teststørrelsen er signifikant* eller, til nød, at *testet er signifikant*. En estimator $\hat{\vartheta}$ kan heller ikke være signifikant. Den kan derimod godt være *signifikant forskellig fra* 0, hvilket blot er en kort måde at sige at testet for $\vartheta = 0$ fører til forkastelse.

Bemærk, at ordet *signifikans* er uløseligt forbundet med *forkastelse*, ligesom *insignifikans* er med *godkendelse*. Men hvis det drejer sig om et test af den sædvanlige slags, hvor en reduceret model testes mod en grundmodel, så er det at hypotesen forkastes bestemt ikke det samme som at grundmodellen "godkendes". Den er allerede godkendt, ellers ville vi slet ikke udføre testet. Og selvom testet fører til godkendelse, således at vi accepterer den reducerede model, skal det ikke forstås sådan at vi "forkaster" grundmodellen. Den er jo sådan set stadig gyldig, da den også omfatter den reducerede model.

En særdeles almindelig fejl består i at indflette jargon fra det program man har brugt til at udføre beregningerne. Vendinger som "output fra PROC GLM viser ... " eller "ved hjælp af en LISTPARAMETERS kommando udskrives følgende estimater" hører ikke hjemme i en statistisk rapport. Ligeledes gælder, at medens vendingerne *inddelingskriterium*, *gruppering*, *kategorisk variabel*, *kvalitativ variabel* og *klassifikation* er tilladte pseudonymer for det vi teknisk kalder en *faktor*, så er der ikke noget som hedder en "CLASS–variabel", heller ikke selv om man tilfældigvis har brugt SAS til sine udregninger. Modeller, estimationsmetoder osv. skal forklares ved hjælp af generelle matematiske, sandsynlighedsteoretiske og (især) statistiske begreber, ikke ved at referere til en statistisk programpakke, som læseren må formodes at kende meget mindre til end man selv gør. De statistiske programpakker er kun hjælpemidler, rapporten er det endelige resultat. Det bedste er faktisk, hvis man overhovedet ikke kan se på rapporten hvilken programpakke der er brugt.

Noget lignende gælder modelformelsyntax, hvis primære formål jo er at kode modeller på en form der kan bruges som input til et statistikprogram. Symbolske modelformler i en statistisk rapport bør normalt undgås. Hvis man f.eks. vil forklare modellen "to parallelle regressionslinier" for nogle personers vægte som funktion af køn og højde, kan man udmærket skrive

$$\text{vægt}_i = \alpha_{\text{køn}_i} + \beta \times \text{højde}_i + \sigma u_i$$

(hvor $u_1, \ldots, u_n$ er stokastisk uafhængige, normeret normalfordelte) eller

(hvis det er nok at specificere middelværdistrukturen)

$$\mathrm{E}(\text{vægt}_i) = \alpha_{\text{køn}_i} + \beta \times \text{højde}_i \,.$$

Det er ikke noget problem, at man på denne måde bruger variabelbetegnelser som er hele ord (vægt, køn og højde) eller (i andre tilfælde) forkortelser af ord. Tværtimod, det bliver ofte meget nemmere at læse end hvis man stereotypt bruger x, y og z til reelle tal og i, j og k til indexvariable (f.eks. faktornivieauer). Det er heller ikke noget problem hvis man udelader visse indices, f.eks. fodtegnet i til køn, meningen fremgår klart alligevel. Det giver derimod *ikke* nogen mening — udenfor de statistiske beregningsprogrammers snævre univers — at skrive

$$\text{vægt} = \text{køn} + \text{højde} \,.$$

For output fra statistiske programpakker gælder (bortset fra visse bilag til dokumentation) at man har det fulde ansvar for at forklare alt hvad der står i rapporten. Hvis man indkopierer output, der indeholder størrelser som man ikke selv forstår, så har man virkelig et problem. Hvilket ikke er urimeligt, da klienten jo i så fald nok vil have det samme problem. Læseren må naturligvis gå ud fra, at alt hvad der står (f.eks.) i en tabel har en vis betydning.

Disposition af rapporten. Det er farligt at sige noget generelt om dette, da inddelingen i afsnit normalt vil (og bør) afhænge af den konkrete problemstilling. Men en overordnet ramme, som i det mindste kan bruges i simple tilfælde, er følgende:

1. Beskrivelse af forsøget og datamaterialet. Dette afsnit er især vigtigt i situationer, hvor rapporten skal læses af andre end klienten (med mindre klienten skriver det selv), eller hvor data er indsamlet og videregivet til statistikeren på en så uoverskuelig måde, at klienten må formodes at have brug for en dokumentation af, at statistikeren har opfattet situationen korrekt. I mange tilfælde er der for eksempel brug for en præcis redegørelse for, hvordan datamaterialet er håndteret i henseende til manglende eller modstridende oplysninger, fjernelse af ekstreme observationer og lignende. Afsnittet kan desuden benyttes til indførelse af terminologi (f. eks. navne på variable og faktorer), som vil blive benyttet i resten af rapporten. Ved kurset MPAS vil en del af dette være indeholdt i opgaveteksten, og det er ikke nødvendigt at gentage oplysninger fra denne.

2. Grafik, tabeller, oversigter m.v. I heldige tilfælde kan væsentlige aspekter af et datamateriale opsummeres i nogle få simple tegninger (punktplots, "parallelle" histogrammer etc.) eller tabeller. I så fald er det en grov fejl at undlade det, og det er ofte mest naturligt at gøre det før man går igang med mere raffinerede modelopstillinger og

analyser. I det ofte forekommende tilfælde, hvor en enkelt (eller nogle få) responsvariables afhængighed af adskillige forklarende variable skal beskrives, kan det være en god ide her at indlede med tegninger, der beskriver responsernes afhængighed af de forklarende variable taget enkeltvis.

3. Modelopstilling og kontrol. Her diskuteres, hvilke modeller eller metoder man naturligt kan benytte til analyse af datamaterialet. Visse former for modelkontrol kan også foretages her, sammen med evt. diskussion af transformationer af data til fjernelse af fordelingsskævheder og variansinhomogenitet, reduktion til aggregerede data etc.

4. Estimation og testning. Tests for relevante modelreduktioner foretages og diskuteres. Estimater (med standardafvigelse) angives for de parametre som er af interesse i de modeller der er af interesse (typisk "slutmodellen", hvis man har sådan en).

5. Konklusion. Sammenfatning af (hovedsageligt) afsnit 4.

6. Bilag. Her kan man placere mere omfattende computeroutput (plots, programudskrifter m.v.) som er nødvendige til dokumentation og ikke med rimelighed kan "klippes ind" i teksten. Men en hovedregel er, at tabeller og grafik bør placeres det sted i rapporten, hvor læseren forventes at have brug for dem.

I forbindelse med en større statistisk analyse skal man *altid* — ikke mindst for sin egen skyld — sørge for at dokumentere udregningerne så de kan reproduceres. I nogle tilfælde vil det fremgå direkte af rapportens ordlyd hvad man har gjort, og i så fald er yderligere dokumentation naturligvis ikke nødvendig. Men hvis valget af model og afgrænsningen af datamaterialet er mere kompliceret må man sørge for at gemme programmer eller lignende, der gør det muligt at gentage hele foretagendet hvis der skulle blive brug for det. Denne dokumentation vil man ude i det virkelige liv ikke nødvendigvis vedlægge, men bare sørge for at kunne fremskaffe på forlangende. Men i forbindelse med dette kursus bedes man, af praktiske grunde, vedlægge tilstrækkelig dokumentation til at alle udregninger kan reproduceres. Altså, i de tilfælde hvor modellen og afgrænsningen af datamaterialet ikke følger entydigt af teksten, vedlæg udskrift af program (og evt. de allermest relevante dele af outputfil).

Praktiske bemærkninger. Alt — også bilag — skal være i almindeligt A4 format med skrift kun på den ene side af papiret. Siderne skal nummereres fortløbende, bortset fra at bilagenes sider eventuelt kan nummereres separat. Af forsiden skal fremgå hvem der har skrevet rapporten, hvilken opgave det drejer sig om, og hvilket semester. Håndskrevne rapporter accepteres ikke, bortset fra at græske bogstaver og lignende evt. kan indsættes i hånden.

Gem for en sikkerheds skyld den endelige rapport på en fil eller i form af en ekstra papirkopi.

Vurdering af rapporten. Den første rapport skal godkendes af læreren, for den anden gives (med ekstern censur) en karakter, som er karakteren for kurset.

Vurderingen af rapporterne er måske lidt anderledes end man er vant til. Klarhed i fremstillingen og korrekthed prioriteres højt. Dårlig formulering, mange stavefejl, dårlig tegnsætning, mange stereotype gentagelser osv. trækker ned. Fejl, der ved eksaminer i de teoretiske fag kan bortforklares som sjuskefejl, kan være ret katastrofale i en statistisk rapport. Hvis man for eksempel bytter om på forkastelse og godkendelse i et test er det en utilgivelig fejl, eftersom det i praksis kan gøre rapporten totalt vildledende. Hvis man — på et mere eller mindre underforstået 5% niveau — søvnigt rapporterer "signifikans" for et test med en P–værdi $< 10^{-6}$ er det også en meget grov fejl. Ligeså hvis den endelige konklusion (som ude i det virkelige liv er det første klienten læser, undertiden det eneste) er forvirrende, utilstrækkelig eller dårligt formuleret.