En statistikstuderendes bekendelser

Søren Wengel Mogensen

Om at skrive BSc-opgave i anvendt statistik.

Der findes matematikere (i hvert fald matematikstuderende), der mener, at den rene matematik er den fineste. Hvor denne vildfarelse stammer fra, er mig ubekendt, men hvis nogen læsere ligger inde med historiske kilder eller tidlige eksempler på en sådan holdning, så er redaktionen interesseret i at modtage disse.

Under alle omstændigheder gjorde jeg det kontroversielle at skrive BSc-opgave i anvendt statistik. Denne artikel er ikke et kampskrift for statistikken, men derimod en venlig håndsrækning med et tilbud om oplysning. Den giver nemlig både et indblik i forløbet omkring en BSc-opgave i anvendt statistik samt beskriver i korte træk min opgave.

Hvordan foregår det?

Forløbet omkring en BSc-opgave i anvendt statistik er strukturelt noget anderledes end ved BSc-opgaver i matematik. For det første foregår vejledningen langt hen ad vejen på gruppebasis, da alle studerende, der skriver samtidigt, har de samme vejledere. I mit tilfælde var det Helle Sørensen og Ernst Hansen. Man kan sikkert både finde styrker og svagheder ved en sådan ordning, men i hvert fald giver det en god mulighed for at ping-ponge med de andre studerende og udveksle idéer, hvilket er lærerigt.

For det andet starter hele forløbet med, at man i gruppen sammen med vejlederne løser en række opgaver, som skal sætte én i stand til lave selve sin BSc-opgave.

Hvad blev det til?

For at kunne lave anvendt statistik skal man efter de indledende opgaver have noget at anvende sine statistiske evner på. Som udgangspunkt er det vejlederne, der skaffer et datasæt til at tjene dette formål. Vi fik et datasæt omhandlende forekomsten af tre plantetyper (klokkelyng, hedelyng og blåtop) i den danske natur. Fokus for opgaven var på kløgtig vis at udtale sig om status og udvikling for klokkelyngen, som er den karakteristiske plantetype for den våde hede. Ja, så ved man det.

I bund og grund var spørgsmålene, der skulle besvares, simple: hvordan udvikler forekomsten af klokkelyng sig med tiden og hvorfor? I datasættet var der nemlig også en hel række potentielle forklarende variable.

Noget af det mest spændende ved forløbet var klart den frihed, der naturligt ligger i opgaven. Man har data, man har nogle idéer, og så skal man finde ud, hvad der giver mening, og hvad der ligger inden for ens egen rækkevidde. Samtidig er det sandsynligvis første gang, man stifter bekendskab med rigtige data. Ikke de polerede skoleeksempler fra en lærebog, men rigtige, beskidte data, som ikke selv tilbyder en åbenlys løsning.

En fyldestgørende beskrivelse af data, og hvad jeg gjorde med det, ville blive en lang smøre. Her følger derfor en ganske overfladisk beskrivelse. Jeg endte med at anvende en logistisk regressionsmodel til data. Herved opskriver man en model, der løst sagt beskriver sandsynligheden for at observere klokkelyng på et bestemt sted til en bestemt tid givet en række forklarende variable (vegetationshøjden, pH-værdien i jorden og meget andet). Vi forestiller os, at eksistensen af klokkelyng på et bestemt sted til en bestemt tid er et udfald af en binær stokastisk variabel, hvis middelværdi (og sandsynligheden for at den stokatiske variabel

er 1) afhænger af to ting: Den afhænger af nogle såkaldte faste effekter. Det kunne eksempelvis være vegetationshøjden på det givne sted og tidspunkt. Derudover afhænger den i den model, jeg opstillede, af nogle tilfældige effekter. Tilfældige effekter introduceres i STAT2, og en egentlig forklaring på, hvad det er, bliver lidt for lang til denne artikel. Dog kan man helt kort sige, at i klokkelyngstilfældet må man formode, at der er en vis korrelation mellem eksempelvis observationer, der ligger tæt på hinanden geografisk. Det kan man modellere ved at bruge tilfældige effekter.

Derudover skal det vælges, hvilke forklarende variable der skal indgå i en model. I STAT1 og STAT2 bruges en ret mekanisk fremgangsmåde, hvor man tester sig nedad i en model ved hjælp likelihoodratiotests. Denne metode kan dog let vise sig at være uheldig, når man står med mange potentielle forklarende variable. Derfor brugte jeg flere forskellige metoder for at komme frem til en model. Selve modelopbygningen er en langsommelig proces, hvor man hele tiden skal holde tungen lige i munden for at kunne argumentere sagligt for de valg, der uundgåeligt skal foretages.

En af de største udfordringer var at gennemføre en hæderlig modelkontrol. Den beskrevne model er en generaliseret lineær model med tilfældige effekter. I litteraturen kan man finde forskellige bud på, hvordan man kan lave modelkontrol i sådan en model, men det viser sig, at der ikke er nogen hverken autoritativ eller skudsikker metode. Jeg brugte nogle idéer fra en artikel om såkaldte kumulative residualer (D.Y. Lin et al., Model-Checking Techniques Based on Cumulative Residuals, 2002). Denne artikel beskæftiger sig dog kun med generaliserede lineære modeller uden tilfældige effekter, så jeg valgte derfor også at undersøge ved simulation, hvorvidt metoden kunne bruges i pågældende tilfælde.

Kort sagt så går idéen ud på, at man definerer de kumulative

residualer som følgende funktion

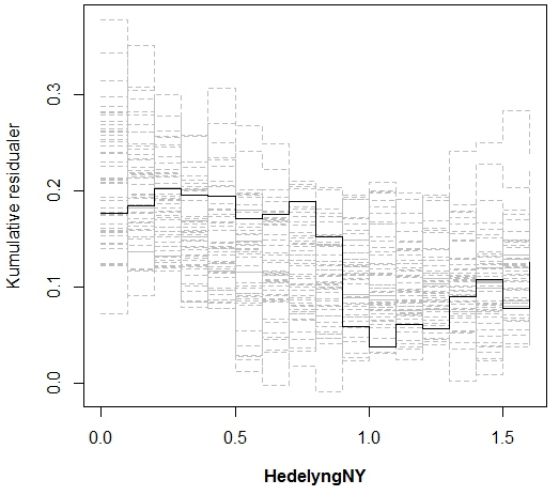
$$W(x) = n^{-1/2} \sum_{i=1}^{n} I(z_i \leq x) \cdot r_i,$$

hvor n er antal observationer, I er indikatorfunktionen, z_i er den i'te værdi af en given kovariat (forklarende variabel), og r_i er residualet hørende til den i'te observation. $W(x_0)$ for en given kovariat er altså en skaleret sum af de residualer, der hører til observationer, for hvilke den givne kovariatværdi er mindre end eller lig med x_0. Når man så vil foretage modelkontrol på en given model, gøres det ved at simulere en række datasæt fra denne model og grafisk sammenligne de kumulative residualer hørende til de simulerede datasæt med de kumulative residualer hørende til det originale data. I figur 1 ses kumulative residualer hørende til kovariaten med det mundrette navn **HedelyngNY**. Håbet er i dette tilfælde, at hvis denne kovariat indgår på en uheldig måde i modellen, vil der være væsentlig forskel på den sorte og de grå grafer.

Som tidligere bemærket er dette dog heller ikke nogen skudsikker metode, hvorfor jeg også lavede en hurtig kontrol af modelkontrollen. Her simulerede jeg data fra to forskellige modeller, fittede begge datasæt til én af modellerne og sammenlignede derefter de kumulative residualer. Her vidste jeg jo, at én af modellerne var fejlspecificeret, og det interessante var jo naturligvis, om metoden kunne afsløre det. Det kunne den til en vis grad, men det siger sig selv, at træerne ikke vokser helt ind i himlen.

Hvad endte det med?

En af statistikerens fornemmeste opgaver er efter min ydmyge opfattelse at anskueliggøre ikke kun sine konklusioner, men også i



Figur 1 Kumulative residualer for det originale data fittet til en model (sort) og data simuleret fra modellen og derefter fittet til samme (grå).

særlig grad den usikkerhed, der er forbundet med konklusionerne. Det betyder dog også, at de konklusioner, man når frem til, som oftest vil være pakket ind i indtil flere lag af modererende ord og udtryk. Således kunne jeg i min konklusion blandt andet pege på, at forekomsten af klokkelyng viste en ikke-signifikant faldende tendens. Så pas på klokkelyngen, når du støder på den. Måske er den på retræte. Som tidligere nævnt hører klokkelyngen til den våde hede, men data kunne ikke påvise, at den faldende tendens (hvis den findes) skyldtes afvanding af de våde hedeområder. Så

omvendt er der altså ikke noget argument for at give sig til at vande klokkelyngen.

På det mere personlige plan endte hele forløbet med at være en sand ilddåb udi statistikken. Man stifter bekendtskab med mange nye områder og arbejder med statistik på en helt ny måde. Det, man har lavet i STAT1 og STAT2, svarer lidt til at have fået styr på et forhåndsslag og et baghåndsslag i tennis, hvorimod BSc-opgaven er ens første rigtige kamp. I kampen får man nemlig brug for både sin forhånd og baghånd, men man får også brug for en hel bunke andre ting, som man må lære hen ad vejen.