1. obligatoriske oppgave

i STK2120

Vår 2011

Obligatorisk oppgave 1 i STK2120, vår 2011 Innleveringfrist: Torsdag 24. februar kl 14.30 i obligkassa i 7. etasje i Niels Henrik Abels hus. Erfaringsmessig blir det lange køer rett før innleveringsfristen, så det er smart å levere tidligere. Husk å bruke den offisielle obligforsiden ved innlevering! Dersom du på grunn av sykdom eller lignende har behov for å utsette innleveringen, må du sende søknad til Robin Bjørnetun Jacobsen (rom B718, NHA, e-post: studieinfo@math.uio.no, tlf. 22 85 59 07). Husk at sykdom må dokumenteres ved legeattest! Se forøvrig http://www.math.uio.no/academics/obligregler.shtml for nærmere informasjon om reglement rundt obligatoriske oppgaver ved matematisk institutt.

Oppgaven er obligatorisk og studenter som ikke får besvarelsen godkjent, vil ikke få adgang til avsluttende eksamen. For å få besvarelsen godkjent, må man minst ha gjort et forsøk på å løse alle deloppgaver. Du kan få poeng på en oppgave selv om du ikke har kommet frem til et svar, og det er derfor viktig at du leverer inn det du har kommet frem til. Det er lov å samarbeide og bruke alle slags hjelpemidler. Den innleverte besvarelsen skal imidlertid være skrevet av deg (for hånd eller på datamaskin) og gjenspeile din forståelse av stoffet. Er vi i tvil om du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse.

Oppgave 1 (Analyse av torskefangst)
For å sette fangstkvoter for fisk, er det viktig å anslå hvor mange fisk som er fanget i ulike aldersgrupper. Aldersfastsettelse av fisk er en komplisert prosess (som i hovedsak består i å telle "åreringer" i fiskenes otolitter) og kan kun utføres på et begrenset antall fisk. Estimering av antall fangede fisk i hver aldersgruppe utføres derfor ved en komplisert statistisk prosedyre (som vi ikke vil gå nærmere inn på her). En viktig del av analysen er imidlertid å estimere hvordan lengden til fisk varierer med alder men også fra fra sesong til sesong og fra område til område.

I denne oppgaven skal vi studere et datasett av torsk. Datasettet er tilgjengelig fra kursets hjemmeside og kan leses inn med kommandoen

```
cod <- read.table("cod.dat",header=T,
                  colClasses=c("numeric","factor","factor","numeric"))
```

Datasettet inneholder 4 variable, `length`, `seas`, `area` og `age` svarende til lengde på fisken (i cm), sesong fisken er fanget i, region den er fanget i og alder (i år) på fisken, henholdsvis.

I oppgaven skal vi se på hvordan vi kan bygge opp en modell som sier noe om sammenhengen mellom lengde og de øvrige forklaringsvariable.

(a) Prøv først ut ulike `plot` kommandoer og `summary` kommandoen for å bli litt kjent med dataene. Kommenter de ting du ser.

(*b*) Vi vil starte med å se hvordan lengde på fisk varierer med region. Forklar hvordan variansanalyse kan brukes for å svare på dette. Sett opp en modell for dette og utfør en analyse på dataene. Prøv ut modellen med lengde både på opprinnelig og log-skala. Argumentér for at det er mer fornuftig å jobbe på log-skala. Kommentér resultatene.

(*c*) Vi vil så utvide modellen til også å ta hensyn til at lengde på fisk kan variere med sesong. Utfør analyse på data også i dette tilfellet. Kommentér igjen resultatene. Kommentér spesielt eventuelle forskjeller i betydningen av region i forhold til forrige deloppgave.

(*d*) Utfør nå parvise sammenlikninger av sesong-effektene. Hva slags ordninger finner du på disse. Er dette rimelig?

(*e*) Utvid så modellen ytterligere til å ta med et interaksjonsledd mellom region og sesong. Kommentér resultatene også i dette tilfellet?

La oss nå glemme for en liten stund variasjonen som skyldes sesong eller regionsforskjeller og heller fokusere på variasjoner som skyldes alder. For å se på sammenhengen mellom alder og lengde vil vi bruke regresjonsanalyse.

(*f*) Anta igjen lengde på log-skala er responsen. Prøv regresjonsmodeller der alder er forklaringsvariabel enten på opprinnelig skala eller på log-skala.

Forklar hva "Multiple R-Squared" betyr og sammenlikn disse for de to modellene.

Argumentér hvorfor alder på log-skala er bedre enn på opprinnelig skala.

Så langt har vi sett separat på hvordan lengde på fisk varierer med sesong og region og hvordan lengde på fisk varierer med alder. I det videre ønsker vi imidlertid både å modellere variasjonene som skyldes ulike aldere *og* variasjoner som skyldes ulike sesonger og regioner. For å kunne utføre en slik analyse trenger vi imidlertid å kunne kombinere varianananalyse og regresjonsanalyse. Vi vil derfor ta et lite avbrekk fra torske-dataene og se mer generelt på sammenhengen mellom variansanalyse og regresjon.

(*g*) Anta en enveis variansanalysemodel, dvs

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

der e_{ij}-ene er uavhengige med forventning 0 og konstant varians σ^2.

Definer nå

$$z_{ijl} = \begin{cases} 1 & \text{hvis } l = i \\ 0 & \text{ellers} \end{cases}$$

(*h*) Vis at

$$Y_{ij} = \mu + \sum_{l=1}^{I} \alpha_l z_{ijl} + e_{ij} \qquad (*)$$

og forklar hvorfor dette kan oppfattes som en multippel regresjonsmodell.

(i) Forklar hvorfor det er mulig å redusere modellen til

$$Y_{ij} = \mu + \sum_{l=1}^{I-1} \alpha_l z_{ijl} + e_{ij}, \qquad (**)$$

dvs en modell der den siste α'en ikke er med.

(j) Gå nå tilbake til modellen i oppgave (a), dvs modellen der kun region ble brukt som forklaringsvariabel. Ved å benytte den generelle formelen

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

prøv å estimere regresjonskoeffisientene for regresjonskoeffisienter både for modell ($*$) og ($**$). Hvilke problemer får du ved modell ($*$)?

(Merk her at $\boldsymbol{X}$ er en matrise som inneholder z-ene ovenfor og at $\boldsymbol{\beta}$ inneholder både μ og α-ene.

Bruk dette til å argumentere hvorfor er dette viktig fra et regresjonssynspunkt å redusere modellen til ($**$).

(k) I praksis behøver vi ikke å lage z-variablene ovenfor. Dette klarer regresjonsrutinene våre (`lm` i **R**) å gjøre selv. Det er imidlertid viktig å si fra om en forklaringsvariabel er å bli oppfattet som en faktor (dvs definerer en gruppering som vi bruker i variansanalyse) eller om det er en numerisk variabel. Lag først et nytt datasett ved

```
cod2 <- read.table("cod.dat",header=T,
                   colClasses=c("numeric","factor","numeric","numeric"))
```

dvs region blir nå lest inn som en *numerisk* variabel.

Prøv ut følgende kommandoer

```
fit.lm <- lm(log(length)~area,data=cod)
fit2.lm <- lm(log(length)~area,data=cod2)
```

og kommentér de forskjeller som kommer frem av disse.

Prøv så kommandoen `anova(fit.lm)` og sammenlikn resultatene med det du fikk i oppgave (a).

Når vi blander sammen noen variable som definerer gruppeinndelinger og andre variable som er numeriske, pleier vi ofte å kalle den første gruppen for kategoriske variable (noen ganger blir de også kalt kvalitative variable) mens de numeriske variable blir kalt kvantitative variable.

(l) Utfør til sist en analyse der både de kategoriske variablene sesong og region samt den kvantitative variabelen alder (på log-skala) blir tatt med i modellen. Kommentér resultatene. Utfør også ulike sjekk av antagelsene som ligger bak modellen.

Hint: Du bør her fortsette å bruke `lm` kommandoen for tilpasning og `anova` kommandoen for å lage anova-tabeller.

Lykke til!