2. februar 2017

STK2100

Obligatorisk oppgave 1 av 2

Innleveringsfrist

Torsdag 16. februar 2017, klokken 14:30 i obligkassen, som står i gangen utenfor ekspedisjonen i 7. etasje i Niels Henrik Abels hus.

Instruksjoner

Du velger selv om du skriver besvarelsen for hånd eller på datamaskin (for eksempel ved bruk av LaTeX). Alle besvarelser skal inkludere følgende offisielle forside:

www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-obligforside.pdf

Det forventes at man har en klar og ryddig besvarelse med tydelige begrunnelser. Husk å inkludere alle relevante plott og figurer. Studenter som ikke får sin opprinnelige besvarelse godkjent, men som har gjort et reelt forsøk på å løse oppgavene, vil få én mulighet til å levere en revidert besvarelse. Samarbeid og alle slags hjelpemidler er tillatt, men den innleverte besvarelsen skal være skrevet av deg og reflektere 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.

I oppgaver der du blir bedt om å programmere må du skrive ut programkoden og levere denne sammen med resten av besvarelsen. Det er viktig at programkoden du leverer inneholder et kjøreeksempel, slik at det er lett å se hvilket resultat programmet gir. For å skrive ut programkoden fra en av UiOs Linux-maskiner kan du gå til mappen hvor programmet ditt ligger og skrive

```
lpr -P pullprint_produsent filnavn
```

der `filnavn` er navnet på filen du ønsker å skrive ut og `pullprint_produsent` er navnet på produsenten av skriveren du ønsker å hente utskriften fra. Det er vanlig å enten bruke `pullprint_Ricoh` eller `pullprint_HP`.

Søknad om utsettelse av innleveringsfrist

Hvis du blir syk eller av andre grunner trenger å søke om utsettelse av innleveringsfristen, må du ta kontakt med studieadministrasjonen ved Matematisk institutt (7. etasje i Niels Henrik Abels hus, e-post: studieinfo@math.uio.no) i god tid før innleveringsfristen.

For å få adgang til avsluttende eksamen i dette emnet, må man bestå alle obligatoriske oppgaver i ett og samme semester.

For fullstendige retningslinjer for innlevering av obligatoriske oppgaver, se her:

www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-oblig.html

LYKKE TIL!

Spesifikke krav til denne oppgaven:

For å få obligen **godkjent** stilles det **krav** til at

- det er gjort et **reelt** forsøk på å løse **alle** enkeltoppgaver. **Dette gjelder for 1. innlevering**!
- Det er en tilfredsstillende besvarelse på minst 2/3 av deloppgavene.

Husk at det er lov å spørre om hjelp!
I de ulike oppgaver er det lagt inn kommandoer som kan brukes i **R**. Det *er* lov å bruke andre programmer, men da vil det stilles ekstra krav til god dokumentasjon av hva som er gjort og man vil ikke kunne forvente å få hjelp i det implementasjonstekniske.

Oppgave 1. Vi vil i denne oppgaven se på et dataset `Boston` og hvordan vi kan bruke regresjon for å predikere kriminalitet.

(a) Gjør data tilgjengelig og sjekk beskrivelsen av datasettet ved kommandoene

```
library(MASS)
data(Boston)
help(Boston)
```

Lag også ulike plott for å få en oversikt over datasettet.

(b) Vi vil i denne oppgaven dele datasettet opp i 2, et *treningsdatasett* og et *testdatasett*. Dette kan gjøres med kommandoene

```
set.seed(345)
ind <- sample(1:nrow(Boston),250,replace=FALSE)
Boston.train <- Boston[ind,]
Boston.test <- Boston[-ind,]
```

Vi vil i det videre bruke treningsdata for estimering og valg av modell mens testdata vil bli brukt for å validere modell.

Diskuter fordeler og ulemper med en slik inndeling av data.

(c) Vi vil først se på en modell

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i.$$

Hva er de vanlige antagelsene på støyleddene ε_i. Diskuter også hvilke antagelser som er mest viktige.

Tilpass en slik modell på treningsdata med `crim` som respons og alle forklaringsvariable innkludert. Diskuter resultatene.

(d) Fjern nå den variabel som har tilhørende størst P-verdi og tilpass den nye modell.

Hvorfor er dette en fornuftig prosedyre?

Diskuter eventuelle endringer på P-verdiene til de resterende variable. Relater det gjerne til korrelasjoner mellom forklaringsvariable.

(e) Fortsett å fjerne forklaringsvariable til alle P-verdier er mindre enn 0.05. Hva blir din endelige modell?

Lag ulike plott for å vurdere om modellen er rimelig.

(f) Bruk den endelige modell til å predikere respons i test-settet og lag et mål basert på gjennomsnittelig kvadratisk feil ($\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$) for å vurdere hvor god modellen er.

(g) Gjennomfør nå den samme modell seleksjonsprosedyre basert på *hele* datasettet. Kommenter eventuelle forskjeller du da får.

Oppgave 2. Vi skal i denne oppgaven se på lineær regresjon med kvalitative (kategoriske) forklaringsvariable. Anta vi har data $(c_1, y_1), ..., (c_n, y_n)$ der $c_i \in \{1, ..., K\}$. Definer for $j = 1, ..., K$

$$x_{i,j} = \begin{cases} 1 & \text{hvis } c_i = j \\ 0 & \text{ellers} \end{cases}$$

(a) Vis at de to modellene

$$Y_i = \beta_0 + \beta_2 x_{i,2} + \cdots + \beta_K x_{i,K} + \varepsilon_i \tag{1}$$

og

$$Y_i = \alpha_1 x_{i,1} + \cdots + \alpha_K x_{i,K} + \varepsilon_i \tag{2}$$

er ekvivalente. Skriv eksplisitt ned sammenhengen mellom $\boldsymbol{\beta}$ og $\boldsymbol{\alpha}$. Bruk også disse modellene til å gi en fortolkning av de ulike parametre.

Vi vil i det etterfølgende holde oss til modell (2) da denne er noe enklere å forholde seg til matematisk.

(b) La $\boldsymbol{X}$ være design matrisen for modell (2), dvs i-te rad av $\boldsymbol{X}$ inneholder verdiene $x_{i,j}, j = 1, ..., K$. Vis at $\boldsymbol{X}^T\boldsymbol{X}$ blir en diagonalmatrise med diagonalelementer n_j der n_j er antall observasjoner med $c_i = j$.

Vis også at $\boldsymbol{X}^T\boldsymbol{y}$ er en vektor der j-te element er lik $\sum_{i:c_i=j} y_j$.

Basert på dette, utled minste kvadraters estimatene for $\alpha_1, .., \alpha_K$. Diskuter om disse estimatene er rimelige.

(c) Basert på sammenhengen mellom $\boldsymbol{\beta}$ og $\boldsymbol{\alpha}$, konstruer også estimater for $\boldsymbol{\beta}$.

Argumenter for hvorfor disse estimatene også blir minste kvadraters estimater for $\boldsymbol{\beta}$.

(d) Nok en alternativ modell er

$$Y_i = \gamma_0 + \gamma_1 x_{i,1} + \cdots + \gamma_K x_{i,K} + \varepsilon_i \tag{3}$$

der $\sum_{j=1}^K \gamma_j = 0$.

Hvilke verdier må γ_j-ene ha for at også denne modellen blir ekvivalent med de foregående modellene?

Hvilken fortolkning har γ-ene i dette tilfellet.

Merk at de to foregående modeller kan skrives om til

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_K x_{i,K} + \varepsilon_i$$

der $\beta_1 = 0$ og

$$Y_i = \alpha_0 + \alpha_1 x_{i,1} + \cdots + \alpha_K x_{i,K} + \varepsilon_i$$

der $\alpha_0 = 0$. Dvs alle de 3 modellene er i utgangspunktet uttrykt på samme måte med $K + 1$ parametre men vi reduserer det til K *frie* parametre ved å legge ulike *restriksjoner* på parametrene. Vi vil se nedenfor hvordan vi kan legge inn ulike restriksjoner når vi gjør analyse i **R**.

Vi skal i resten av oppgaven se på et datasett fra Devore & Berk (2012, oppgave 11.5). Datasettet består av målinger av jerninnhold i 4 ulike jernformasjoner (1=karbonat, 2=silikat, 3=magnetitt, 4 =hematitt). Tabellen nedenfor viser dataene, med 10 observasjoner for hver type jernformasjon.

Type	1	2	3	4	5	6	7	8	9	10
1	20.50	28.10	27.80	27.00	28.00	25.20	25.30	27.10	20.50	31.30
2	26.30	24.00	26.20	20.20	23.70	34.00	17.10	26.80	23.70	24.90
3	29.50	34.00	27.50	29.40	27.90	26.20	29.90	29.50	30.00	35.60
4	36.50	44.20	34.10	30.30	31.40	33.10	34.10	32.90	36.30	25.50

Vi ønsker å teste om det er forskjell i jerninnhold mellom de ulike typene.

(e) Les inn data ved kommandoen

```
Fe <- read.table("http://www.uio.no/studier/emner/matnat/math/STK2100/v17/fe.txt",
                 header=T, sep=",")
```

Prøv så ut tilpasning ved kommandoen

```
fit1 <- lm(Fe~form+0,data=Fe)
summary(fit1)
```

Hvorfor går dette galt?

Angi så kommandoen

```
Fe$form <- as.factor(Fe$form)
```

Prøv så det samme kallet til `lm` igjen. Hvorfor får du et mer fornuftig resultat nå. Hvilken av de 3 modellene svarer denne tilpasningen til? Dvs hvilken restriksjon svarer dette til?

(f) Prøv så ut

```
options()$contrasts
fit2 <- lm(Fe~form,data=Fe)
summary(fit2)

options(contrasts=c("contr.sum","contr.sum"))
options()$contrasts
fit3 <- lm(Fe~form,data=Fe)
summary(fit3)
```

Hvilke modeller svarer disse tilpasningene til? Er det samsvar mellom resultatene du får for de ulike tilpasningene?

For alle de 3 modellene, sett opp estimatene for *alle* de $K+1$ regresjonsparametrene.

(g) Anta nå du ønsker å teste om det er forskjell mellom de ulike typer jernformasjoner. Formuler en passende hypotese for dette og forklar hvordan du kan bruke (en av) utskriftene ovenfor til å utføre en slik hypotesetest.

Hva blir konklusjonen av denne testen?

(h) Basert på de ulike utskriftene ovenfor, foreslå en mulig forenkling av modellene.