Andre sett med obligatoriske oppgaver
STK2120 våren 2016

Her er det andre settet med obligatoriske oppgaver i STK2120 våren 2016. Oppgavesettet består av tre oppgaver. Prøv å besvare spørsmålene kort og konsist, men likevel med gode forklaringer! Der du bruker `R`, må kommandoer og resultater innarbeides i rapporten eller legges ved.

Oppgavene er obligatoriske og studenter som ikke får besvarelsen godkjent, vil ikke få adgang til avsluttende eksamen. For å få besvarelsen godkjent, må du minst ha gjort et forsøk på å løse alle deloppgaver.

Det er helt i orden og utmerket om du samarbeider med andre og diskuterer hvordan oppgavene skal løses. Den innleverte besvarelsen skal imidlertid være skrevet av deg og gjenspeile din forståelse av stoffet. Det må gå fram av besvarelsen hvem du eventuelt har samarbeidet med. Er vi i tvil om du har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse.

Besvarelsen leveres ved Matematisk institutt, 7. etasje, Niels Henrik Abels hus.
Husk at du skal bruke Matematisk institutts forside ved innlevering. Du finner denne her:
`www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-obligforside.pdf`

Frist for innlevering er torsdag 14. april kl. 14.30.

Oppgave 1

En forsker har gjort en studie for å undersøke mulige sammenhenger mellom en score for "analytisk ferdighet" for 4 år gamle barn og en del forklaringsvariabler (beskrevet nedenfor). Analytisk ferdighet ble vurdert ved en standard psykologisk testprosedyre.

Du kan lese dataene fra undersøkelsen inn i `R` ved kommandoen:

```
skills=
read.table("http://www.uio.no/studier/emner/matnat/math/STK2120/v16/skills.txt",
header=T)
```

Dataene er organisert med én linje for hvert av de 36 barna som var med i studien og med følgende variabler i de seks kolonnene:

- `IQf`: fars IQ
- `IQm`: mors IQ
- `cntage`: alder (i måneder) da barnet først kunne telle til ti
- `read`: gjennomsnittlig antall timer per uke barnet ble lest for av mor eller far
- `edutv`: gjennomsnittlig antall timer per uke barnet har sett på læringsprogrammer på TV siste 3 måneder
- `skill`: score på den psykologiske testen

Variablene i de fem første kolonnene er forklaringsvariabler, mens variabelen i den siste kolonnen er responsvariabelen.

a) Kommandoene `summary(skills)` og `plot(skills)` gir deg beskrivende statistikk og plott for variablene. Diskuter kort hva den beskrivende statistikken og plottene forteller deg om variablene og sammenhengen mellom dem.

Vi vil bruke multippel lineær regresjon til å studere hvordan responsvariabelen `skill` avhenger av forklaringsvariablene `IQf`, `IQm`, `cntage`, `read` og `edutv`.

b) Bruk forlengs utvelgelse og Mallows C_p til å bestemme hvilke av de fem forklaringsvariablene som har betydning for å forklare responsen. Beskriv hvordan metoden fungerer og gi en fortolkning av den modellen du kommer fram til.

Vink: Se R-kommandoer til foreleningene i uke 8.

To alternativer til forlengs utvelgelse, er baklengs utvelgelse og utprøving av alle mulige regresjonsmodeller.

c) Forklar hvordan de to alternative metodene fungerer, og undersøk om de gir et annet resultat enn forlengs utvelgelse.

Oppgave 2

I denne oppgaven skal vi se nærmere på den logistiske regresjonsmodellen; jf. sidene 620–622 og 650–651 i Devore & Berk).

Vi antar at $Y_1, Y_2, \ldots, Y_n$ er uavhengige Bernoulli variabler (dvs. stokastiske variabler som kan anta verdiene 0 og 1) og at $x_1, x_2, \ldots, x_n$ er forklaringsvariabler for Y_i-ene. Vi setter

$$p(x_i) = P(Y_i = 1 \,|\, x_i) = 1 - P(Y_i = 0 \,|\, x_i) \tag{1}$$

og antar at (1) er gitt ved den logistiske regresjonsmodellen

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \tag{2}$$

Vi lar $y_1, y_2, \ldots, y_n$ være de observerte verdiene av Y_i-ene.

a) Vis at likelihood funksjonen kan skrives som

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{e^{\beta_0 y_i + \beta_1 x_i y_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

La $l(\beta_0, \beta_1) = \log L(\beta_0, \beta_1)$ være log-likelihood funksjonen. Da er vektoren av score-funksjoner gitt ved

$$\mathbf{s}(\beta_0, \beta_1) = \begin{bmatrix} s_1(\beta_0, \beta_1) \\ s_2(\beta_0, \beta_1) \end{bmatrix}$$

der

$$s_1(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_0} l(\beta_0, \beta_1) \quad \text{og} \quad s_2(\beta_0, \beta_1) = \frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1)$$

b) Vis at

$$s_1(\beta_0,\beta_1) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\,\beta_0+\beta_1 x_i}}{1+e^{\,\beta_0+\beta_1 x_i}}$$

$$s_2(\beta_0,\beta_1) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i\, e^{\,\beta_0+\beta_1 x_i}}{1+e^{\,\beta_0+\beta_1 x_i}}$$

Den observerte informasjonsmatrisen er gitt ved

$$\mathbf{J}(\beta_0,\beta_1) = \begin{bmatrix} J_{11}(\beta_0,\beta_1) & J_{12}(\beta_0,\beta_1) \\ J_{21}(\beta_0,\beta_1) & J_{22}(\beta_0,\beta_1) \end{bmatrix}$$

der

$$\begin{aligned} J_{11}(\beta_0,\beta_1) &= -\frac{\partial^2}{\partial\beta_0^2} l(\beta_0,\beta_1) \\ J_{12}(\beta_0,\beta_1) &= J_{21}(\beta_0,\beta_1) = -\frac{\partial^2}{\partial\beta_1\partial\beta_0} l(\beta_0,\beta_1) \\ J_{22}(\beta_0,\beta_1) &= -\frac{\partial^2}{\partial\beta_1^2} l(\beta_0,\beta_1) \end{aligned}$$

c) Vis at

$$\begin{aligned} J_{11}(\beta_0,\beta_1) &= \sum_{i=1}^{n} \frac{e^{\,\beta_0+\beta_1 x_i}}{(1+e^{\,\beta_0+\beta_1 x_i})^2} \\ J_{12}(\beta_0,\beta_1) &= J_{21}(\beta_0,\beta_1) = \sum_{i=1}^{n} \frac{x_i\, e^{\,\beta_0+\beta_1 x_i}}{(1+e^{\,\beta_0+\beta_1 x_i})^2} \\ J_{22}(\beta_0,\beta_1) &= \sum_{i=1}^{n} \frac{x_i^2\, e^{\,\beta_0+\beta_1 x_i}}{(1+e^{\,\beta_0+\beta_1 x_i})^2} \end{aligned}$$

Merk at den observerte informasjonsmatrisen ikke avhenger av y_i-ene. Derfor er Fishers informasjonsmatrise (den forventete informasjonsmatrisen) $\mathbf{I}(\beta_0,\beta_1)$ lik den observerte informasjonsmatrisen $\mathbf{J}(\beta_0,\beta_1)$.

En kan vise at maksimum likelihood estimatorene $\hat\beta_0$ og $\hat\beta_1$ er tilnærmet binormalt fordelt med forventningsverdier β_0 og β_1 og kovariansmatrise $\mathbf{I}(\beta_0,\beta_1)^{-1}$. Det betyr at for alle valg av konstanter c_0 og c_1 er $c_0\hat\beta_0 + c_1\hat\beta_1$ tilnærmet normalfordelt med forventningsverdi $c_0\beta_0 + c_1\beta_1$ og varians $c_0^2\, I^{11}(\beta_0,\beta_1) + c_1^2\, I^{22}(\beta_0,\beta_1) + 2c_0c_1\, I^{12}(\beta_0,\beta_1)$. Her er $I^{jk}(\beta_0,\beta_1)$ element (j,k) i den inverse informasjonsmatrisen $\mathbf{I}(\beta_0,\beta_1)^{-1}$. (Du skal ikke vise dette resultatet.)

Vi ser nå på sannsynligheten

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

svarende til en gitt verdi av forklaringsvariabelen. Denne sannsynligheten kan estimeres ved

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}}$$

d) Bestem et tilnærmet 95% konfidensintervall for $p(x)$.

Vink: Finn først et konfidensintervall for $\eta(x) = \beta_0 + \beta_1 x$.

Oppgave 3

Vi vil nå se på dataene fra eksempel 12.14 i Devore & Berk om feil ved pakningne i 23 romferge-oppskytinger før Challenger ulykken i januar 1986.

Du kan lese dataene inn i `R` ved kommandoen:

```
exmp12.14=
read.table("http://www.uio.no/studier/emner/matnat/math/STK2120/v16/exmp12-14.txt",
header=T)
```

Vi vil bruke den logistiske regresjonsmodellen (2) til å studere sammenhengen mellom temperatur og sannsynligheten for feil ved pakningene. Temperaturen er her målt i Fahrenheit grader.

a) Implementer Newton-Raphson algoritmen for å maksimere log-likelihooden $l(\beta_0, \beta_1)$. Som forklaringsvariabel x_i skal du bruke "temperatur − 70 grader". I besvarelsen din skal du gi R-kommandoene du bruker for å implemetere algoritmen.

Vink: Bruk resultatene i punktene b og c i oppgave 3. I `R`-kommandoene fra forelesningene i uke 14 og i heftet til Storvik om optimering av likelihooder er det gitt eksempler på implementering av Newton-Raphson algoritmen.

b) Hva blir maksimum likelihood estimatene for β_0 og β_1? Bestem standardfeilen til estimatene.

Da Challenger ble skutt opp i januar 1986 var temperaturen 31 grader Fahrenheit.

c) Gi et estimat for sannsynligheten for feil ved pakningene når temperaturen er 31 grader Fahrenheit. Bruk resultatet i punkt d i oppgave 2 til å gi et 95% konfidensintervall for denne sannsynligheten. Diskuter hva dette sier deg om sjansen for feil ved pakningene da Challenger ble skutt opp. Ser du noen problemer med analysen i dette punktet?

LYKKE TIL!