NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4267 Lineære statistiske modeller**

Faglig kontakt under eksamen: Mette Langaas

Tlf: 988 47 649

Eksamensdato: 19. mai 2017

Eksamenstid (fra–til): 09.00–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: C: *Tabeller og formler i statertikk* (Tapir forlag, Fagbokforlaget), *Matematisk formelsamling* (K. Rottmann), gult, stempla A5-ark med dine egne håndskrevne notater. Bestemt kalkulator.

Annen informasjon:
Alle svar skal begrunnes og besvarelsen skal inneholde naturlig mellomregning.

Målform/språk: bokmål

Antall sider: 9

Antall sider vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig ☐ **2-sidig** ☒
sort/hvit ☒ **farger** ☐
skal ha flervalgskjema ☐

Dato Sign

Oppgave 1 Stokastisk vektor

La $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ være en multivariat normalfordelt stokastisk vektor med $\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ og $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$, der ρ er et reellt tall.

Videre, la $Y_1 = \frac{1}{3}(X_1 + X_2 + X_3)$, $Y_2 = X_1 - X_2$ og $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$.

a) Finn en konstant matrise $\boldsymbol{C}$ slik at $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$.
Finn $\mathrm{E}(\boldsymbol{Y})$ og $\mathrm{Cov}(\boldsymbol{Y})$.
Hva er fordelingen til $\boldsymbol{Y}$?
Er Y_1 og Y_2 uavhengige?

La $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), (\lambda_3, \boldsymbol{e}_3)$ være egenverdi–egenvektorpar for matrisen $\boldsymbol{\Sigma}$, det vil si, $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$, for $i = 1, 2, 3$. Videre er egenverdiene $\lambda_1 = 1+2\rho$ og $\lambda_2 = \lambda_3 = 1-\rho$, og egenvektorene $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ og $\boldsymbol{e}_3$ er ortogonale og normaliserte kolonnevektorer.

b) For hvilke verdier av ρ er kovariansmatisen $\boldsymbol{\Sigma}$ positivt definitt?
Hvorfor vil vi at kovariansmatrisen skal være positivt definitt?
Hva er fordelingen til $\begin{pmatrix} \boldsymbol{e}_1^T\boldsymbol{X} \\ \boldsymbol{e}_2^T\boldsymbol{X} \\ \boldsymbol{e}_3^T\boldsymbol{X} \end{pmatrix}$?
For $\rho = 0.5$, regn ut sannsynligheten $P(\boldsymbol{e}_1^T\boldsymbol{X} + \boldsymbol{e}_2^T\boldsymbol{X} + \boldsymbol{e}_3^T\boldsymbol{X} > 4)$.

Merk: De numeriske verdiene til egenvektorene *trengs ikke* for noen av beregningene. Men, hvis det gjør det enklere å forstå – så er følgende ortogonale og normaliserte egenvektorer til $\boldsymbol{\Sigma}$ når $\rho = 0.5$:

$$\boldsymbol{e}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \; \boldsymbol{e}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \text{ og } \boldsymbol{e}_3 = \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}.$$

Oppgave 2 Modellering av systolisk blodtrykk

«The Framingham Heart Study» er en studie av etiologien (dvs. de underliggende årsakene) til hjerte-karsykdom, med deltakere fra Framingham i Massachusetts, USA[1] .

Vi skal fokusere på å modellere systolisk blodtrykk ved å bruke et datasett bestående av data fra $n = 2600$ personer. For hver person i datasettet har vi målt følgende syv variabler:

- `SYSBP`: systolisk blodtrykk (mmHg),
- `SEX`: 1=mann, 2=kvinne,
- `AGE`: alder (år) ved undersøkelsen,
- `CURSMOKE`: røykestatus (sigaretter) ved undersøkelsen: 0=røyker ikke daglig, 1= røyker daglig,
- `BMI`: kroppsmasseindeks (kg/m^2),
- `TOTCHOL`: serum total kolesterol (mg/dl), og
- `BPMEDS`: bruke av medisin mot høyt blodtrykk ved undersøkelsestidspunktet: 0=bruker ikke, 1=bruker.

En multippel normal lineær regresjonsmodell ble tilpasset datasettet med `SYSBP` som respons og alle de andre variablene som kovariater. Vi kaller dette `modelA`. I figur 1 finner du R-kode og utskrift fra tilpassing av `modelA`, og i figur 2 finner du residualplott.

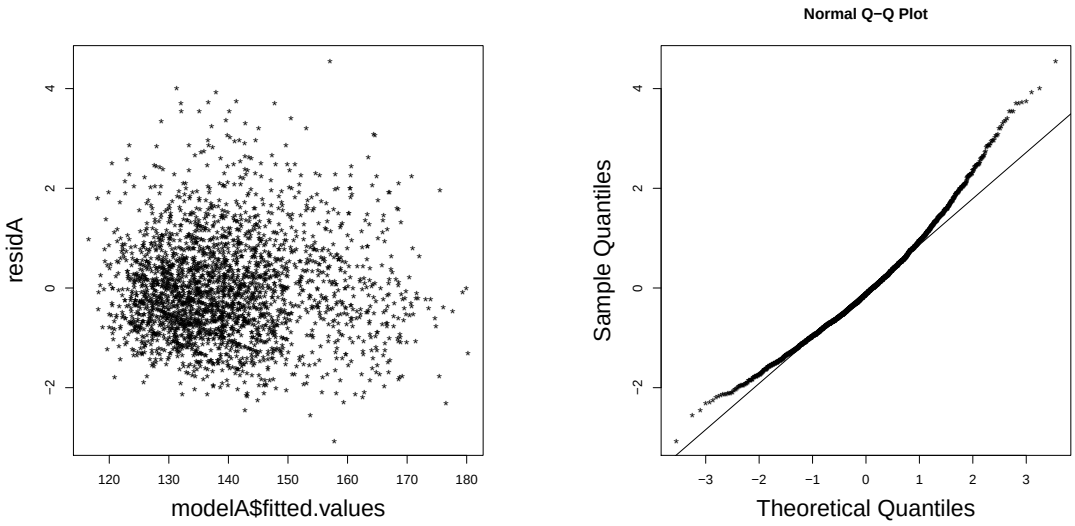
a) I utskriften fra `summary(modelA)` i figur 1 er *to* numeriske verdier byttet ut med spørsmåltegn. Skriv ned matematiske formler for hver av disse, regn ut tallverdier, og forklar hva hvert av tallene betyr.

Hvordan vil du, fra figurene 1–2, vurdere modelltilpasningen?

[1]For mer informasjon om «The Framingham Heart Study» kan du besøke https://www.framinghamheartstudy.org/. Dette datasettet er et delsett av en undervisningsversjon av Framingham-datasettet, og dette er brukt med tillatelse fra the «The Framingham Heart Study».

```
# name of dataset: thisds
> dim(thisds)
[1] 2600    7
> colnames(thisds)
[1] "SYSBP"  "SEX"  "AGE"  "CURSMOKE"  "BMI"  "TOTCHOL"  "BPMEDS"
> modelA=lm(SYSBP~SEX+AGE+CURSMOKE+BMI+TOTCHOL+BPMEDS,data=thisds)
> summary(modelA)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.505170   4.668798  12.103  < 2e-16
SEX         -0.429973   0.807048  -0.533        ?
AGE          0.795810   0.048413  16.438  < 2e-16
CURSMOKE    -0.518742   0.853190  -0.608  0.54324
BMI          1.010550          ?  10.129  < 2e-16
TOTCHOL      0.028786   0.008787   3.276  0.00107
BPMEDS      19.203706   1.102547  17.418  < 2e-16
Residual standard error: 19.65 on 2593 degrees of freedom
Multiple R-squared:  0.2508,        Adjusted R-squared:  0.249
F-statistic: 144.6 on 6 and 2593 DF,  p-value: < 2.2e-16
> residA=rstudent(modelA)
> plot(modelA$fitted.values,residA,pch=20)
> qqnorm(residA,pch=20)
> qqline(residA)
> library(nortest)
> ad.test(residA)
Anderson-Darling normality test
data:  residA
A = 13.2, p-value < 2.2e-16
```

Figur 1: Utskrift fra R-kommandoer og resultater for `modelA`. To tall er erstattet med spørsmålstegn.



Figur 2: Residualplott for tilpasning av `modelA` for Framingham-dataene.
konstant matrise

En konkurrerende modell, `modelB`, har $-\frac{1}{\sqrt{\texttt{SYSBP}}}$ som respons og de samme kovariatene som `modelA`, se figurene 3–4.

b) Hvis du sammenligner figur 2 og 4, vil du foretrekke `modelA` eller `modelB`? Begrunn svaret.

Vi fortsetter med `modelB`.

Hvorfor kan en redusert modell være bedre å bruke enn en full modell når målet er prediksjon?
Forklar kort hva som gjøres i «best subset»-metoden, og forklar bakgrunnen for BIC-kriteriet. Husk å ta med forklaringen på hvordan man fant de 6 modellene som er presentert i bunnen av figur 3.
Velg en redusert regresjonsmodell basert på resultatene ved bruk av BIC-kriteriet i figur 3.

c) Vi kan tilpasse en redusert versjon av `modelB` ved å fjerne variablene `SEX`, `CURSMOKE` og `TOTCHOL` fra modellen. Vi kaller dette `modelC`, se figur 5. For å sammenligne `modelB` og `modelC` ønsker vi å utføre følgende hypothestest:

$$H_0\colon \beta_{\texttt{SEX}} = \beta_{\texttt{CURSMOKE}} = \beta_{\texttt{TOTCHOL}} = 0,$$
$$H_1\colon \text{minst én er ulik } 0,$$

der $\beta_{\texttt{SEX}}$, $\beta_{\texttt{CURSMOKE}}$ og $\beta_{\texttt{TOTCHOL}}$ er regresjonsparametere for henholdsvis `SEX`, `CURSMOKE` og `TOTCHOL`.

Forklar hvordan hypotesetesten kan formuleres som en lineær hypotese av formen $H_0 : \boldsymbol{C\beta} = \boldsymbol{d}$ mot $H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}$, der $\boldsymbol{\beta}$ er den fulle vektoren av regresjonsparametere, og skriv ut hva $\boldsymbol{C}$ og $\boldsymbol{d}$ er i vårt tilfelle.
Regn ut verdi for testobservatoren du bruker, velg selv signifikansnivå og utfør hypotesetesten.
Basert på resultatet av hypotesetesten vil du foretrekke `modelB` eller `modelC`?

Oppgave 3 Forsøksplanlegging

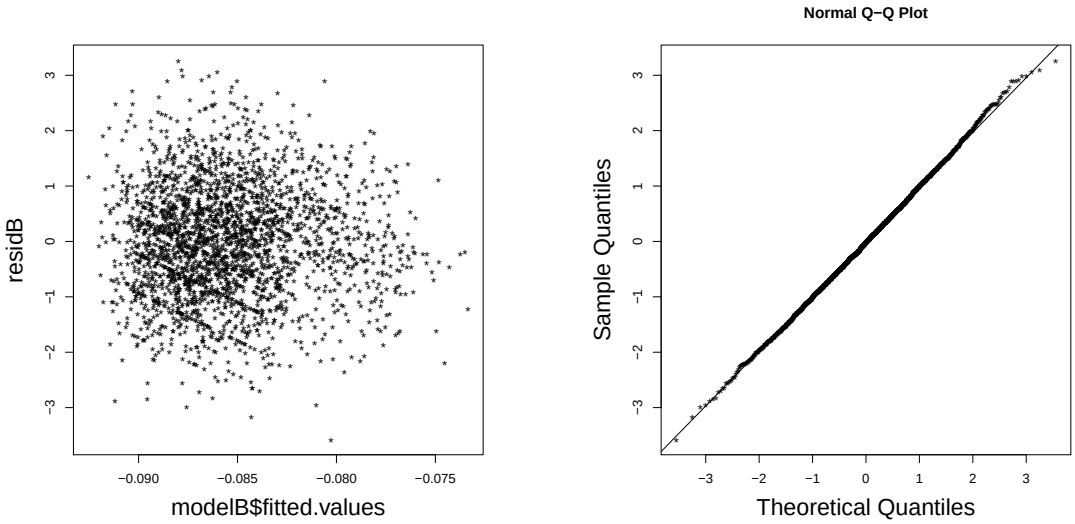
I en pilotstudie med fire faktorer A, B, C og D, ble de 8 delforsøkene under utført.

	A	B	C	D
1	−1	−1	−1	1
2	1	−1	−1	−1
3	−1	1	−1	−1
4	1	1	−1	1
5	−1	−1	1	−1
6	1	−1	1	1
7	−1	1	1	1
8	1	1	1	−1

a) Hvilken type forsøk er dette?
Hva er generator og definerende relasjon for forsøket?
Hvilken resolusjon har forsøket?
Skriv ned alias-strukturen for forsøket.
Da de 8 delforsøkene over skal utføres, blir vi instruert å gjøre delforsøkene i tilfeldig rekkefølge. Hvorfor?

```
> modelB=lm(-1/sqrt(SYSBP)~SEX+AGE+CURSMOKE+BMI+TOTCHOL+BPMEDS,
data=thisds)
> summary(modelB)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.103e-01  1.383e-03 -79.745  < 2e-16 ***
SEX         -2.989e-04  2.390e-04  -1.251 0.211176
AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
CURSMOKE    -2.504e-04  2.527e-04  -0.991 0.321723
BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
BPMEDS       5.469e-03  3.265e-04  16.748  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.005819 on 2593 degrees of freedom
Multiple R-squared:  0.2494,        Adjusted R-squared:  0.2476
F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
> residB=rstudent(modelB)
> par(mfrow=c(1,2))
> plot(modelB$fitted.values,residB,pch=20)
> qqnorm(residB,pch=20)
> qqline(residB)
> ad.test(residB)
Anderson-Darling normality test
data:  residB
A = 0.19209, p-value = 0.8959
> library(leaps)
> x <- model.matrix(modelB)[,-1]; dim(x)
[1] 2600    6
> colnames(x)
[1] "SEX"      "AGE"      "CURSMOKE" "BMI"      "TOTCHOL"  "BPMEDS"
> y <- -1/sqrt(thisds$SYSBP)
> allfit=regsubsets(x,y,nvmax=6)
> allsummary=summary(allfit)
> allsummary
Subset selection object
6 Variables  (and intercept)
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         SEX AGE CURSMOKE BMI TOTCHOL BPMEDS
1  ( 1 ) " " " " " " "      " " " "     "*"
2  ( 1 ) " " "*" " " "      " " " "     "*"
3  ( 1 ) " " "*" " " "      "*" " "     "*"
4  ( 1 ) " " "*" " " "      "*" "*"     "*"
5  ( 1 ) "*" "*" " " "      "*" "*"     "*"
6  ( 1 ) "*" "*" "*"      "*" "*"     "*"
> allsummary$bic
[1] -332.1004 -589.2250 -700.2825 -704.0729 -697.5698 -690.6913
```

Figur 3: Utskrift fra R-kommandoer og resultater for `modelB`.



Figur 4: Residualplott for tilpasning av `modelB` for Framingham-dataene.

```
> modelC=lm(-1/sqrt(SYSBP)~AGE+BMI+BPMEDS,data=thisds)
> summary(modelC)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.093e-01  1.152e-03  -94.91   <2e-16 ***
AGE          2.449e-04  1.389e-05   17.63   <2e-16 ***
BMI          3.199e-04  2.902e-05   11.02   <2e-16 ***
BPMEDS       5.490e-03  3.254e-04   16.87   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Residual standard error: 0.005831 on 2596 degrees of freedom
Multiple R-squared:  0.2453,        Adjusted R-squared:  0.2444
F-statistic: 281.3 on 3 and 2596 DF,  p-value: < 2.2e-16
```

Figur 5: Utskrift fra R-kommandoer og resultater fra `modelC`.

Oppgave 4 Undertilpasning

Den normale multiple regresjonsmodellen kan skrives i matrisenotasjon som

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

der $\boldsymbol{Y}$ er en n-dimensjonal stokastisk kolonnevektor, $\boldsymbol{X}$ er en gitt designmatrise med n rader og p kolonner, $\boldsymbol{\beta}$ er en ukjent p-dimensjonal kolonnevektor av regresjonsparametre og $\boldsymbol{\varepsilon}$ er en n-dimensjonal kolonnevektor av normalfordelte tilfeldige feil med forventningsverdi $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ og kovariansmatrise $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, der $\boldsymbol{I}$ er $n \times n$-identitetsmatrisen. Anta at $n > p$ og at $\boldsymbol{X}$ har rang p.

Anta videre at vi deler modellen inn i to ledd:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

der $\boldsymbol{X}_1$ er en $n \times k$-matrise med rang k, $\boldsymbol{X}_2$ er en $n \times (p-k)$-matrise med rang $p-k$, $\boldsymbol{\beta}_1$ er en k-dimensjonal kolonnevektor av regresjonsparametre og $\boldsymbol{\beta}_2$ er en $p-k$-dimensjonal kolonnevektor av regresjonsparametre.

Anta at vi feilaktig undertilpasser regresjonsmodellen, dvs. at vi bare tilpasser en del av den sanne regresjonsmodellen,

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}^*.$$

Minste kvadrasums estimator for $\boldsymbol{\alpha}_1$ i den undertilpassede modellen er gitt som $\hat{\boldsymbol{\alpha}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X_1})^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$. Kvadratsummen for feil («sums of squares of errors») for den undertilpassede modellen er $\mathrm{SSE}_1 = (\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\alpha}}_1)^T(\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\alpha}}_1)$. La videre $H_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

a) Vis først at $\boldsymbol{H}_1$ er idempotent og finn trasen til $\boldsymbol{H}_1$.
Vi så at $\mathrm{SSE}_1 = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}$.

Finn til slutt forventningsverdien til estimatoren for varians

$$S^2 = \frac{\mathrm{SSE}_1}{n-k}.$$

Svaret skal være en funksjon av n, k, σ^2, $\boldsymbol{\beta}_2$, $\boldsymbol{X}_1$ (eller $\boldsymbol{H}_1$) og $\boldsymbol{X}_2$.

Hint: den såkalte traseformelen er gitt som: $\mathrm{E}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{Y}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\mu}$, der $\boldsymbol{Y}$ er en stokastisk vektor med forventningsverdi $\boldsymbol{\mu}$ og kovariansmatrise $\boldsymbol{\Sigma}$, og $\boldsymbol{C}$ er en passende konstant matrise.

Oppgave 5 Uavhengighet av lineærkombinasjoner

La $\boldsymbol{X}$ være en p-dimensjonal normalfordelt stokastisk vektor med forventningsverdi $\boldsymbol{\mu}$ og positivt definitt kovariansmatrise $\boldsymbol{\Sigma}$. Den momentgenererende funksjonen til $\boldsymbol{X}$ er
$M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$, der $\boldsymbol{t}$ er en p-dimensjonal reell kolonnevektor.

La videre $\boldsymbol{A}$ være en $q \times p$ konstant matrise og $\boldsymbol{B}$ en $r \times p$-konstant matrise.

a) Bruk momentgenererende funksjon til å vise at $\begin{pmatrix}\boldsymbol{AX}\\ \boldsymbol{BX}\end{pmatrix}$ er multivariat normalfordelt.
Utled deretter en betingelse for når $\boldsymbol{AX}$ og $\boldsymbol{BX}$ er uavhengige.