NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

# Eksamensoppgave i **TMA4267 Lineære statistiske modeller**

**Faglig kontakt under eksamen:**

**Tlf:**

**Eksamensdato:** August 2014

**Eksamenstid (fra–til):**

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

**Målform/språk:** bokmål

**Antall sider:** 7

**Antall sider vedlegg:** 0

**Kontrollert av:**

Dato Sign

## Oppgave 1 Multivariat normalfordeling

La $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ være en bivariat normalfordelt stokastisk vektor med forventningsverdi $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ og kovariansmatrise $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

**a)** La $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, hvor $Y_1 = 3X_1 - 2X_2$ and $Y_2 = X_1 + X_2$.

Hvilken fordeling har $\boldsymbol{Y}$?
La $Z = X_1 + aX_2$. Hvordan kan du velge $a$ slik at $Z$ og $Y_2$ er uavhengige?

I figur 1 finner du egenverdiene og egenvektorene til kovariansmatrisen $\boldsymbol{\Sigma}$.

**b)** La $f(\boldsymbol{x})$ være sannsynlighetstettheten (pdf) til $\boldsymbol{X}$.

Beskriv grafen for ligningen $f(\boldsymbol{x}) = d$, der $d > 0$ er en konstant.
Hvilken verdi av $d$ vil gi en graf hvor sannsynligheten for at $\mathbf{X}$ er innenfor området avgrenset av grafen er lik 95%?
Tegn en tegning av grafen, for verdien av $d$ funnet over.

```
> sigma <- matrix(c(1,0.5,0.5,2),ncol=2)
> eigen(sigma)
$values
[1] 2.2071068 0.7928932

$vectors
           [,1]       [,2]
[1,] 0.3826834 -0.9238795
[2,] 0.9238795  0.3826834
```

Figur 1: Egenverdier og egenvektorer til kovariansmatrisen i Problem 1b.

## Oppgave 2 Predikere fettinnholdet i kjøtt

Fettinnholdet i kjøtt kan måles ved hjelp av analytisk kjemi. Men, dette er en metode som det tar langt tid å utføre. Som et eksperiment brukte forskere nær infrarød transmisjon for å måle absorpsjon i et 100 kanals spektrum (bølgelengder i intervallet 850–1050 nm). Dette ble gjort for hver av 215 prøver av finhakket kjøtt. For hver prøve ble også fettinnhold målt. Målet med eksperimentet var å utvikle en prediksjonsmetode for fettinnholdet i kjøtt, basert på de 100 absorpsjonsverdiene.

**a)** En multippel lineær regresjonsmodell (MLR) ble tilpasset til datasettet av 215 prøver, med de 100 absorpsjonene som kovariater (kall disse `xV1`, `xV2`, ..., `xV100`) og logaritmen til fettinnholdet som respons. Et utdrag av resultatene finnes i figur 2, og residualplott i figur 3. I figur 4 vises de estimerte regresjonskoeffisientene grafisk.

I figur 2 er $p$-verdien for variabelen `xV100` byttet ut med et spørsmåltegn. Skriv ned null- og alternative hypotesen som her skal testes. Er den manglende $p$-verdien mindre eller større enn 0.05?

Hvordan vil du *kort* evaluere modelltilpasningen?

Forklar begrepet *overtilpasning* i MLR.
Tror du overtilpasning kan være et problem i regresjonen som er utført her? Begrunn.

**b)** En prinsipalkomponentanalyse ble utført for de 215 observasjonene av 100 absorpsjoner.

Hva er den matematiske definisjonen av prinsipalkomponentladninger (loadings) og -skårer (scores)?

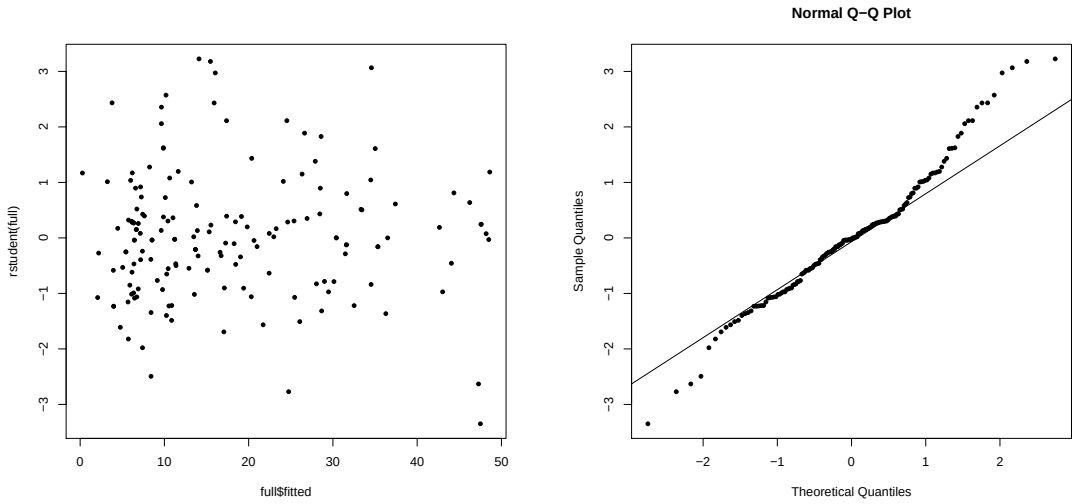
I figur 6 vises de estimerte prinsipalkomponentladingene for de tre første prinsipalkomponentene. Hvordan vil du fortolke hver av disse tre prinsipalkomponentene?

I figur 5 finnes utskrift fra prinsipalkomponentanalyse i R av de 100 absorpsjonsverdiene. Hvor stor prosentandel av total varians er forklart av de tre første prinsipalkomponentene?

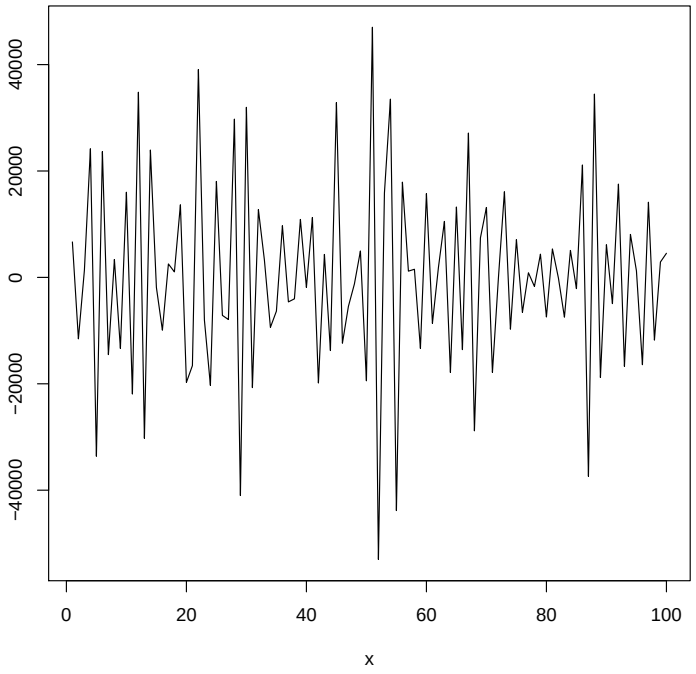
Hvordan kan resultater fra denne prinsipalkomponentananalysen brukes i en regresjonsmodell der logaritmen til fettinnholdet i kjøttprøvene er respons? Hva kan være grunner til å utføre en slik regresjon istedenfor regresjonen i a)?

```
> full <- lm(y~x)
> summary(full)
Call:
lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-0.52853 -0.11046  0.00315  0.11128  0.53530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.7503     0.3598   2.085 0.039264 *
xV1          2404.3987   576.0421   4.174 5.87e-05 ***
xV2         -3926.1439  1058.6867  -3.709 0.000324 ***

Information on xV3 to xV98 not included.

xV99         1051.4592  1517.3418   0.693 0.489743
xV100        -222.4339   688.7246  -0.323 ?
---

Residual standard error: 0.2341 on 114 degrees of freedom
Multiple R-squared:  0.9544,        Adjusted R-squared:  0.9145
F-statistic: 23.88 on 100 and 114 DF,  p-value: < 2.2e-16

> ad.test(rstudent(full))

        Anderson-Darling normality test

data:  rstudent(full)
A = 0.6021, p-value = 0.1166
```

Figur 2: Utdrag fra utskrift fra MLR av de 100 absorpsjonene mot logaritmen til fettinnhold i Problem 2a.



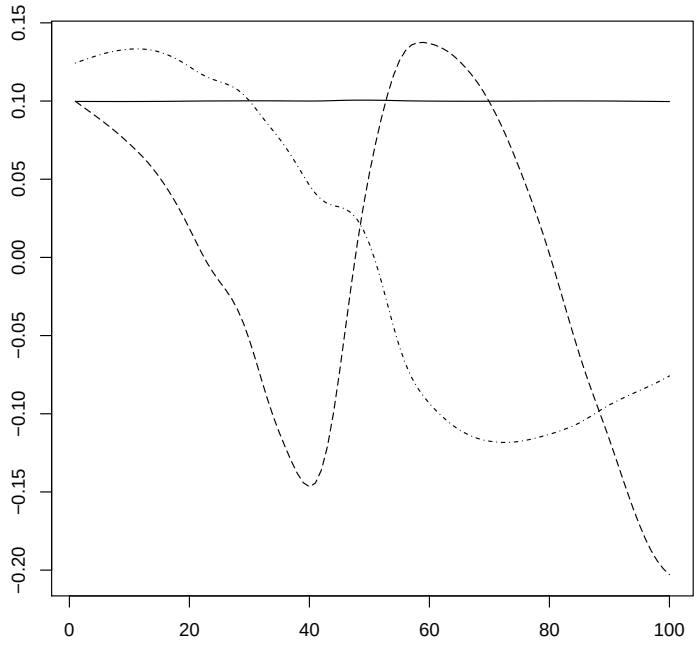

Figur 3: Residual plott (studentiserte residualer mod tilpassede verdier til venstre, normalplott basert på studentiserte residualer til høyre) for MLR av de 100 absorpsjonene mot logarimen til fettinnhold i Problem 2a.




Figur 4: Estimerte regresjonskoeffisienter (vertikal akse) for de 100 absorpsjonene (horisontal akse) i Problem 2a.

```
> res <- prcomp(x,scale=TRUE)
> summary(res)
# only the first 6 out of 100 principal
#components are presented

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     9.9311 0.9847 0.52851 0.33827 0.08038 0.05123
Proportion of Variance 0.9863 0.0097 0.00279 0.00114 0.00006 0.00003
Cumulative Proportion  0.9863 0.9960 0.99875 0.99990 0.99996 0.99999
```

Figur 5: Utdrag fra utskrift for Problem 2b.




Figur 6: De estimerte prinsipalkomponenteladningene for de 100 absorpsjonene for kjøttdataene i Problem 2b. Den horisontale aksen gir de 100 absorpsjonene og den vertikale aksen viser estimerte ladninger for de tre første prinsipalkomponentene. De estimerte ladningene for den første prinsipalkomponenten er vist som en heltrukket kurve, den andre er vekselvis prikket og streket, og den tredje er stiplet.

## Oppgave 3 Design of experiments

I en pilotstudie med fire faktorer, A, B, C og D, ble følgende 8 eksperimenter 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 generatoren og den definerende relasjonen for forsøket?

Hvilken resolusjon har forsøket?

Skriv ned alias-strukturen for forsøket.

## Oppgave 4 Multippel lineær regresjon

Den klassiske multiple regresjonsmodellen kan skrives i matrisenotasjon som følger

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

der $\boldsymbol{Y}$ er en $n$-dimensjonal tilfeldig (stokastisk) vektor, $\boldsymbol{X}$ er en gitt designmatrise med $n$ rader og $p$ kolonner, $\boldsymbol{\beta}$ er en ukjent $p$-dimensjonal vektor av regresjonskoeffisienter og $\boldsymbol{\varepsilon}$ er en $n$-dimensjonal vektor av tilfeldige feil.

Anta at $n > p$ og at $\boldsymbol{X}$ har rang $p$.

Definer matrisen $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

**a)** Hva slags matrise er $\boldsymbol{H}$? Begrunn svaret ditt.
Finn rangen til $\boldsymbol{H}$.
Hvordan kan du grafisk fortolke vektoren $\boldsymbol{HY}$?

Svar på de samme tre spørsmålene for matrisen $\boldsymbol{I} - \boldsymbol{H}$, ved å bruke det du allerede har funnet ut for $\boldsymbol{H}$. Her er $\boldsymbol{I}$ en $n \times n$ identitetsmatrise.

Videre, anta at vektoren av tilfeldige feil $\boldsymbol{\varepsilon}$ er multivariat normalfordelt med forventningsverdi $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ og kovariansmatrise $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, der $\boldsymbol{I}$ er en $n \times n$ identitetsmatrise.

**b)** La $\mathrm{SSE} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. Utled fordelingen til SSE.
Bruk dette til å foreslå en forventningsrett estimator for $\sigma^2$, og kall denne estimatoren $\hat{\sigma}^2$.
Finn variansen til $\hat{\sigma}^2$.

Definer to konstantmatriser $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ og $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{H})$.

**c)** Hva er dimensjonene til matrisene $\boldsymbol{A}$ og $\boldsymbol{B}$?
Vis at $\boldsymbol{AY}$ og $\boldsymbol{BY}$ er uavhengige stokastiske vektorer.
Bruk dette til å bevise at minste kvadratsumsestimatoren $\hat{\boldsymbol{\beta}}$ og SSE er uavhengige stokastiske variabler. Hvordan kan du bruke dette resultatet i en multippel lineær regresjonsmodell?