NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4255 Anvendt statistikk**

Faglig kontakt under eksamen: Anna Marie Holand

Tlf: 951 38 038

Eksamensdato: 16. mai 2015

Eksamenstid (fra–til): 09:00-13:00

Hjelpemiddelkode/Tillatte hjelpemidler: Ett gult A4-ark og spesiell kalkulator.

Annen informasjon:

- I utskrift fra MINITAB er komma brukt som desimalseparator.
- Signifikansnivå 5% skal brukes hvis ikke annet er spesifisert.
- Alle svar må begrunnes.

Målform/språk: bokmål

Antall sider: 9

Antall sider vedlegg: 0

Kontrollert av:

Dato Sign

Oppgave 1 Produsent av gjødningsmiddel

En produsent av gjødningsmiddel ville studere forskjellen mellom effekten av et gammelt gjødningsmiddel (angitt som X_1) og et nyutviklet gjødningsmiddel (angitt som X_2) på veksten av planter. Et eksperiment ble utført hvor plantene ble dyrket under indentiske forhold og man tildelte tilfeldig gjødningsmiddel X_1 til $n_1 = 6$ planter og gjødningsmiddel X_2 til $n_2 = 7$ planter. Etter 3 uker ble høyden til plantene målt i cm. Data fra eksperimentet er presentert under.

i	1	2	3	4	5	6	7
x_{1i}	54.0	56.1	52.1	56.4	54.0	52.9	
x_{2i}	51.0	53.3	55.6	51.0	55.5	53.0	52.1

Deskriptive mål for dette datasettet er $\bar{x}_1 = \frac{1}{6}\sum_{i=1}^{6} x_{1i} = 54.25$,
$s_{x1} = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_{1i} - \bar{x}_1)^2} = 1.71$, $\bar{x}_2 = \frac{1}{7}\sum_{i=1}^{7} x_{2i} = 53.07$,
$s_{x2} = \sqrt{\frac{1}{6}\sum_{i=1}^{7}(x_{2i} - \bar{x}_2)^2} = 1.91$.

a) Vi antar at X_{1i} og X_{2i} er normalfordelte, $X_{1i} \sim N(\mu_1, \sigma^2)$, $i = 1, ..., 6$ og $X_{2i} \sim N(\mu_2, \sigma^2)$, $i = 1, .., 7$.

Basert på dette forsøket, kan produsenten konkludere med at gjennomsnittlig høyde til plantene som ble gitt de to forskjellige typene (X_1 og X_2) av gjødningsmiddel er forskjellige? Skriv ned nullhypotesen og den alternative hypotesen. Velg en testobservator og gjennomfør en hypotesetest. Bruk signifikansnivå $\alpha = 0.05$. Spesifiser hvilke antagelser du gjør.

b) Hva er forskjellen mellom en ikke-parametrisk hypotesetest, og testen brukt i a)?

Utfør en Wilcoxon rank-sum test basert på dataene gitt ovenfor. Du kan anta en normaltilnærming til testobservatoren (U_1 eller U_2) med forventning

$$\mu = \frac{n_1 n_2}{2}$$

og varians

$$\sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Kommenter dine resultater/funn.

Oppgave 2 Sementhydratisering

Betong er produsert ved sement blandet med sand, grus og vann. En prosess kalt sementhydratisering (en reaksjon med vann) spiller en viktig rolle for mikrostrukturen under utviklingen av betongen. Hydratiseringsprosessen produserer en serie av kjemiske reaksjoner som genererer varme. Varmen som blir generert avhenger av sementens sammensetning, og varmen er en parameter som påvirker egenskapene til materialet og styrken av betongen.

For å studere varmen som genereres i sementhydratiseringsprosessen, ble $n = 13$ partier av betong utforsket, og for hver av disse partiene ble varmen som blir generert og 4 mulige forklaringsvariabler målt. Følgende beskrivelse er gitt.

- y: Varme utviklet i kalorier under herding av sementen, målt per gram
- x_1, % av tricalcium aluminate
- x_2: % av tricalcium silicate
- x_3: % av tetracalcium alumino ferrite
- x_4: % av dicalcium silicate

Et parvis spredningsplott og en korrelasjonsmatrise av variablene finnes i henholdsvis figur 1 og figur 2.

En multippel lineær regresjonsmodell ble tilpasset til dataene med y som respons og x_1, x_2, x_3 og x_4 som forklaringsvariabler. La (y_i, x_{1i}, x_{2i}, x_{3i} og x_{4i}) angi en observasjon fra parti i, der $i = 1, ..., 13$. Definer den fulle modellen (modell A):

$$\text{Modell A} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i, \tag{1}$$

hvor ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, ..., n$. MINITAB-utskriften fra en statistisk analyse av modell A finnes i figur 3. Plott av standardiserte residualer finnes i figur 4.

a) Skriv ned den estimerte regresjonsligningen.

Det er gitt en p-verdi i raden for x3 i resultatene i figur 3. Forklar med ord hva denne p-verdien betyr.

Basert på plottene og regresjonsresultatene, vil du si at modell A er en god modell for dataene? Du må spesifisere hvilke egenskaper til plottene og regresjonsresultatene du bruker for å komme frem til svaret ditt.

Vi vil nå sammenligne den fulle regresjonsmodellen (modell A) med en redusert model (kalt modell B), som bare inneholder x_1 og x_2.

$$\text{Modell B} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{2}$$

Resultater fra tilpassing av modell B finnes i figur 5.

b) Gi en kommentar om de viktigste forskjellene mellom modell A og modell B. Modell A og modell B kan sammenlignes ved å undersøke følgende hypoteser.

$$H_0 : \beta_3 = \beta_4 = 0 \text{ vs. } H_1 : \beta_3 \text{ and } \beta_4 \text{ er ikke begge lik null}$$

Utfør hypotesetesten og konkluder.

Vi er nå interessert i å sammenligne ulike regresjonsmodeller, der vi har med ulike kombinasjoner av forklaringsvariablene x_1, x_2, x_3 og x_4. Anta at et konstantledd, β_0, er med i regresjonsmodellen.

MINITAB-utskriften fra tilpassning av ulike modeller for dataene er presentert i figur 6. Hver rad i figur 6 svarer til en modell. Antallet forklaringsvariabler inkludert i hver modell (i tillegg til konstantleddet, β_0) finner du i kolonnen med navn $Vars$. De to beste modellene for hvert antall forklaringsvariabler er rapportert. X'ene indikerer hvilke variabler som er funnet i modellen.

c) Forklar hvordan R^2, R^2_{adj}, Mallows C_p og S er definert og hvordan du kan bruke dem til å sammenligne de ulike regresjonsmodellene. Hvilken av disse 7 regresjonsmodellene mener du er den "beste" for dette datasettet?

Oppgave 3 Smerte og hårfarge

I en studie utført ved University of Melbourne var målet å sammenligne forskjellen mellom smerteterskelen av personer med blondt hår og personer med brunt hår. I denne studien ble totalt $n = 19$ menn og kvinner av forskjellige aldre delt inn i fire kategorier etter hårfarge: lys blond, mørk blond, lys brunette, eller mørk brunette. Hver person i forsøket ble gitt poengskår på sin smerteterskel basert på hennes eller hans prestasjon på en smertesensitivitetstest (jo høyere poengskår, jo høyere er personens smertetoleranse). Et boksplott av dataene er presentert i figur 7.

a) En en-veis variansanalyse (ANOVA) ble tilpasset til dataene, og MINITAB-utskriften fra ANOVA og et sammendrag av dataene er gitt i figur 8.

Seks av tallene i figur 8 er blitt erstattet med et spørsmålstegn (?). Forklar hva disse tallene betyr og regn ut numeriske verdier for de seks manglende tallene.

Hvilke antagelser ligger bak denne analysen?

Er de fire kategoriene hårfarge forskjellige med hensyn på smertetoleranse? Utfør en hypotesetest for å svare på dette spørsmålet. Skriv ned nullhypotesen og den alternative hypotesen. Baser testen på hva du har funnet i figur 8. Bruk signifikansnivå $\alpha = 0.05$.

b) Tidligere studier indikerer at personer med hårfarge *lys blond* er forskjellig fra personer med hårfargen *mørk brunette* med hensyn på smertetoleranse. Vi vil teste dette. Skriv ned nullhypotesen og den alternative hypotesen. Velg en testobservator og gjennomfør en hypotesetest. Bruk signifikansnivå $\alpha = 0.05$. Bruk sammendraget av dataene gitt i figur 8 når du utfører hypotesetesten. Hva er din konklusjon?

Kan du utføre hypotesetesten ovenfor ved å bruke et 95% konfidensintervall? Beregn 95% konfidensintervallet og forklar.

Oppgave 4 Tilfredshet hos kunder

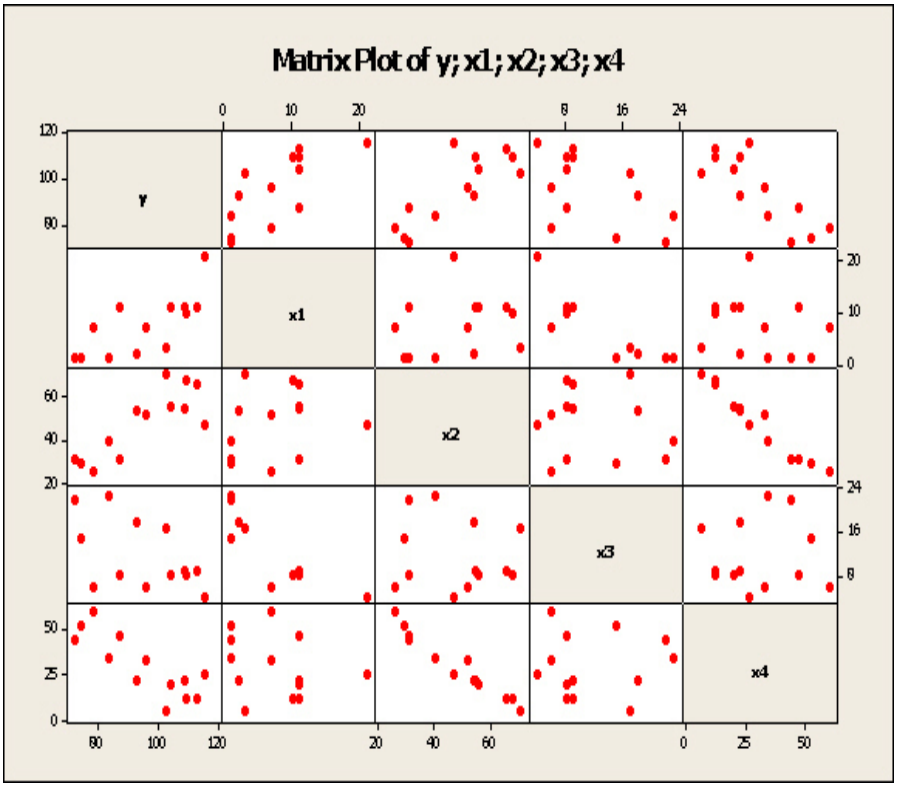
En regional butikk-kjede utfører en spørreundersøkelse for å finne ut hvor tilfreds kundene er med butikk-kjeden. De delte kundene i 4 geografiske regioner (Sør-Vest, Sør-Øst, Midtre og Nord). Det totale antall personer som ble spurt i hver av de 4 regionene ble bestemt før spørreundersøkelsen ble gjennomført som en

prosentandel av populasjonsstørrelsen i regionen. Tilfredsheten hos kundene ble klassifisert i tre grupper og følgende kryss-tabell ble observert.

Region/Grad av tilfredshet:	Tilfreds	Vet ikke	Ikke tilfreds	Totalt
Sør-Vest	235	74	89	398
Sør-Øst	654	203	309	1166
Midtre	366	79	244	689
Nord	179	54	54	287
Totalt	1434	410	696	2540

a) Basert på dataene, kan vi konkludere med at de 4 forskjellige regionene er forskjellige med hensyn på grad av tilfredshet? Skriv ned nullhypotesen og den alternative hypotesen og gjennomfør en hypotesetest basert på tabellen ovenfor. Bruk 5% signifikansnivå.

For å forenkle beregningsbyrden kan du bruke at χ^2-testobservatoren blir 44.78, og du trenger bare å vise hvordan du regner ut ett av de 12 leddene i summen. Hva er konklusjon din basert på denne testen?



Figur 1: Parvis spredningsplott av variablene i sementhydratiseringsdatasettet.

```
Correlations: y; x1; x2; x3; x4

          y       x1       x2       x3
x1    0,731
x2    0,816    0,229
x3   -0,535   -0,824   -0,139
x4   -0,821   -0,245   -0,973    0,030
```

Figur 2: Pearson korrelasjon mellom variablene x_1, x_2, x_3 og x_4 i sementhydratiseringsdatasettet.

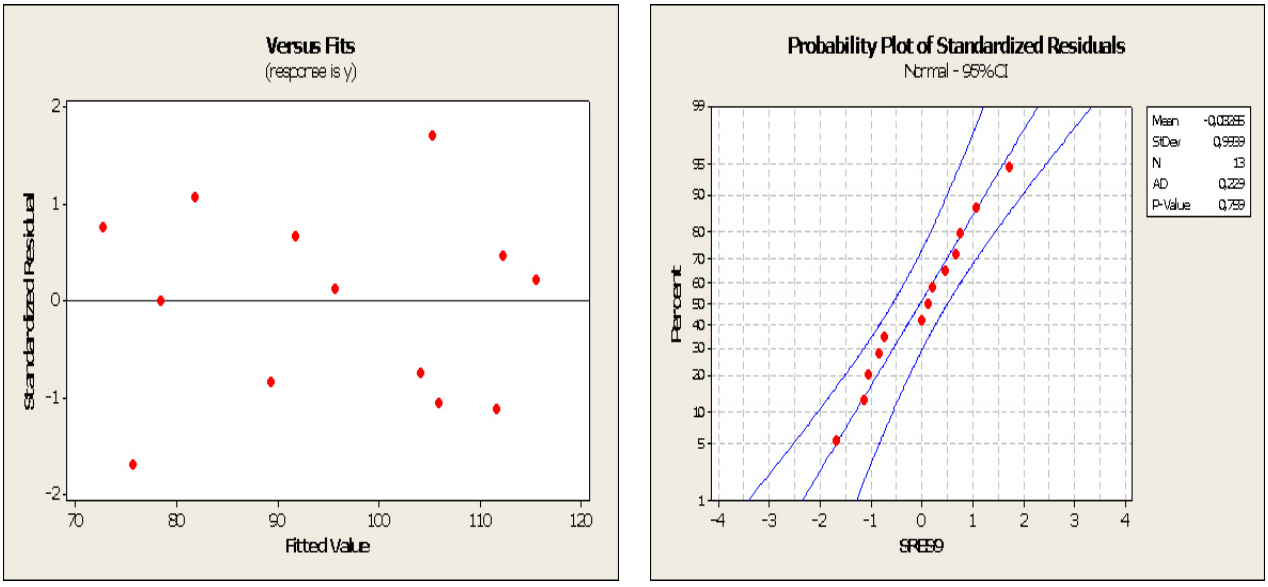
Predictor	Coef	SE Coef	T	P
Constant	62,41	70,07	0,89	0,399
x1	1,5511	0,7448	2,08	0,071
x2	0,5102	0,7238	0,70	0,501
x3	0,1019	0,7547	0,14	0,896
x4	-0,1441	0,7091	-0,20	0,844

S = 2,44601 R-Sq = 98,2% R-Sq(adj) = 97,4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2667,90	666,97	111,48	0,000
Residual Error	8	47,86	5,98		
Total	12	2715,76			

Figur 3: Utskrift fra statistisk analyse av sementhydratiseringsdataene for modell A.



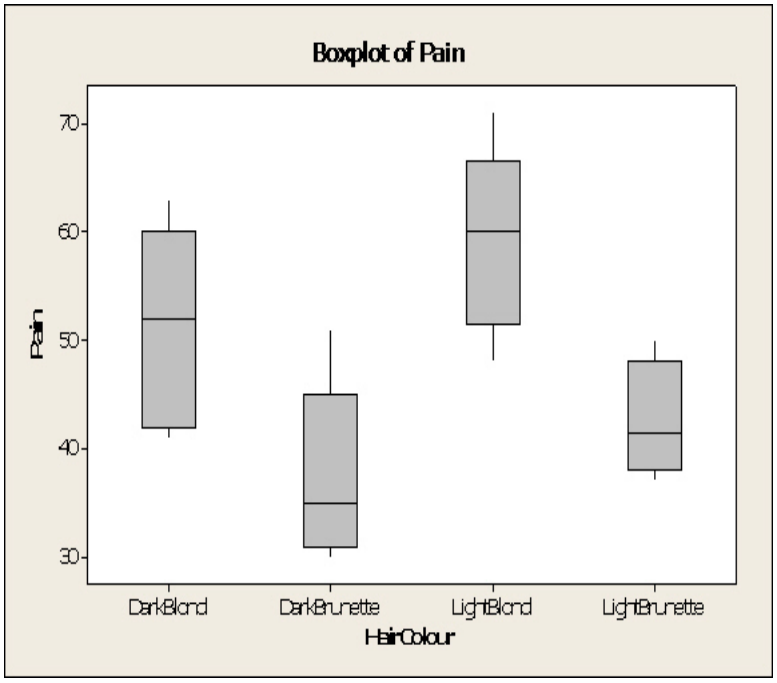
Figur 4: Residualplott (standardiserte residualer mot tilpassede verdier i venstre panel og normalplott basert på standardiserte residualer i høyre panel) for regresjonsmodell A for sementhydratiseringsdatasettet.

```
Predictor     Coef  SE Coef      T      P
Constant    52,577    2,286  23,00  0,000
x1           1,4683   0,1213  12,10  0,000
x2          0,66225  0,04585  14,44  0,000

S = 2,40634   R-Sq = 97,9%   R-Sq(adj) = 97,4%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       2  2657,9  1328,9  229,50  0,000
Residual Error  10    57,9     5,8
Total           12  2715,8
```

Figur 5: Utskrift fra statistisk analyse av sementhydratiseringsdataene for modell B.

```
                        Mallows          x x x x
 Vars  R-Sq  R-Sq(adj)       Cp       S  1 2 3 4
    1  67,5       64,5    138,7  8,9639        X
    1  66,6       63,6    142,5  9,0771    X
    2  97,9       97,4      2,7  2,4063  X X
    2  97,2       96,7      5,5  2,7343  X     X
    3  98,2       97,6      3,0  2,3087  X X   X
    3  98,2       97,6      3,0  2,3121  X X X
    4  98,2       97,4      5,0  2,4460  X X X X
```

Figur 6: Utskrift fra statistisk analyse av sementhydratiseringsdatasettet.



Figur 7: Boksplott av smerte og hårfargedataene.

```
One-way ANOVA: Smerte versus Hårfarge

Source      DF      SS     MS     F      P
Hårfarge    3       ?      453,6  ?   0,004
Error       ?       ?      ?
Total       18  2362,5

S = ?   R-Sq = 57,60%   R-Sq(adj) = 49,12%

Level          N    Mean  StDev
MørkBlond      5  51,200  9,284
MørkBrunette   5  37,400  8,325
LysBlond       5  59,200  8,526
LysBrunette    4  42,500  5,447
```

Figur 8: Utskrift fra statistisk analyse av smerte og hårfargedatasettet.