NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4255 Anvendt statistikk**

Fagleg kontakt under eksamen: Anna Marie Holand

Tlf: 951 38 038

Eksamensdato: 16. mai 2015

Eksamenstid (frå–til): 09:00-13:00

Hjelpemiddelkode/Tillatne hjelpemiddel: Eit gult A4-ark og spesiell kalkulator

Annan informasjon:

- I utskrifta frå MINITAB er komma brukt som desimalskilleteikn.
- Signifikansnivå 5% skal brukast om ikke anna er spesifisert.
- Alle svar må grunngjevast.

Målform/språk: nynorsk

Sidetal: 9

Sidetal vedlegg: 0

Kontrollert av:

Dato Sign

Oppgåve 1 Produsent av gjødningsmiddel

Ein produsent av gjødningsmiddel ville studera skilnaden mellom effekten av eit gammalt gjødningsmiddel (angjeve som X_1) og eit nyutvikla gjødningsmiddel (angjeve som X_2) på veksten av plantar. Eit eksperiment blei utført der plantane vart dyrka under indentiske tilhøve og ein tildelte tilfeldig gjødningsmiddel X_1 til $n_1 = 6$ plantar og gjødningsmiddel X_2 til $n_2 = 7$ plantar. Etter 3 veker blei høgda til plantane målt i cm. Dataane 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 gjennomsnittleg høgde til plantane som ble gjeve dei to forskjellige typane (X_1 og X_2) av gjødningsmiddel er forskjellige? Skriv ned nullhypotesen og den alternative hypotesen. Vel ein testobservator og gjennomfør ein hypotesetest. Bruk signifikansnivå $\alpha = 0.05$. Spesifiser kva antakingar du gjer.

b) Kva er forskjellen mellom ein ikkje-parametrisk hypotesetest, og testen brukt i a)?

Utfør ein Wilcoxon rank-sum test basert på dataane gitt ovanfor. Du kan anta ei 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 resultata/funna dine.

Oppgåve 2 Sementhydratisering

Betong er produsert ved sement blanda med sand, grus og vatn. Ein prosess kalla sementhydratisering (ein reaksjon med vatn) spelar ei viktig rolle for mikrostrukturen under utviklinga av betongen. Hydratiseringsprosessen produserer ei rekkje av kjemiske reaksjonar som genererer varme. Varmen som blir generert avheng av sammensetjinga av sementen, og varmen er ein parameter som påvirkar eigenskapane til materialet og styrken av betongen.

For å studere varmen som blir generert i sementhydratiseringsprosessen, blei $n = 13$ parti av betong utforska, og for kvar av desse partia blei generert varme og 4 moglege forklaringsvariablar målt. Følgjande beskrivelse er gitt.

- y: Varme utvikla i kaloriar under herdinga 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

Eit parvis spreiingsplott og ein korrelasjonsmatrise av variablane er gitt i henholdsvis figur 1 og figur 2.

Ein multippel lineær regresjonsmodell blei tilpassa dataane med y som respons og x_1, x_2, x_3 og x_4 som forklaringsvariablar. La (y_i, x_{1i}, x_{2i}, x_{3i} og x_{4i}) vere ein observasjon frå parti i, der $i = 1, ..., 13$. Definer den fulle modellen (modell A):

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

kor ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, ..., n$. MINITAB-utskrifta frå ein statistisk analyse av modell A er gitt i figur 3. Plott av standardiserte residual er gitt i figur 4.

a) Skriv ned den estimerte regresjonslikninga.

Det er gjeve ein p-verdi i raden for x3 i resultata i figur 3. Forklar med ord kva denne p-verdien betyr.

Basert på plotta og regresjonsresultata, vil du sei at modell A er ein god modell for dataane? Du må spesifisere kva eigenskapar til plotta og regresjonsresultata du brukar for å kome fram til svaret ditt.

Vi vil no samanlikne den fulle regresjonsmodellen (modell A) med en redusert model (kalla modell B), som berre inneheld x_1 og x_2.

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

Resultata frå tilpassing av modell B finnest i figur 5.

b) Gje ein kommentar om de viktigaste forskjellane mellom modell A og modell B. Modell A og modell B kan samanliknast ved å undersøkje følgjande 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 no interessert i å samanlikne ulike regresjonsmodellar, der vi har med ulike kombinasjonar av forklaringsvariablane x_1, x_2, x_3 og x_4. Anta at eit konstantledd, β_0, er med i regresjonsmodellen.

MINITAB-utskrifta frå tilpasning av ulike modellar for dataane er presentert i figur 6. Kvar rad i figur 6 svarar til ein modell. Talet på forklaringsvariablar inkludert i kvar modell (i tillegg til konstantleddet, β_0) finn du i kolonna med navn $Vars$. Dei to beste modellane for kvart tal av forklaringsvariablar er rapportert. X'ane indikerer kva variablar som er funne i modellen.

c) Forklar korleis R^2, R^2_{adj}, Mallows C_p og S er definert og korleis du kan bruke dei til å samanlikne dei ulike regresjonsmodellane. Kven av desse 7 regresjonsmodellane meiner du er den "beste" for dette datasettet?

Oppgåve 3 Smerte og hårfarge

I ein studie utført ved University of Melbourne var målet å samanlikne forskjellen mellom smerteterskelen av personar med blondt hår og personar med brunt hår. I denne studien blei totalt $n = 19$ menn og kvinnar av forskjellige aldre delt inn i fire kategoriar etter hårfarge: lys blond, mørk blond, lys brunette, eller mørk brunette. Kvar person i forsøket blei gjeve poengskår på ein smerteterskel basert på hennar eller hans prestasjon på ein smertesensitivitetstest (jo høgare poengskår, jo høgare er personens smertetoleranse). Eit boksplott av dataane er presentert i figur 7.

a) Ein en-vegs variansanalyse (ANOVA) blei utført på dataane, og MINITAB-utskrifta frå ANOVA og eit samandrag av dataane er gitt i figur 8.

Seks av tala i figur 8 er blitt erstatta med eit spørsmålsteikn (?). Forklar kva desse tala betyr og rekn ut numeriske verdiar for dei seks manglande tala.

Kva antagelsar ligg bak denne analysen?

Er dei fire kategoriane hårfarge forskjellige med omsyn på smertetoleranse? Utfør ein hypotesetest for å svare på dette spørsmålet. Skriv ned nullhypotesen og den alternative hypotesen. Baser testen på kva du har funnet i figur 8. Bruk signifikansnivå $\alpha = 0.05$.

b) Tidligare studiar indikerer at personar med hårfarge *lys blond* er forskjellig fra personar med hårfargen *mørk brunette* med omsyn på smertetoleranse. Vi vil teste dette. Skriv ned nullhypotesen og den alternative hypotesen. Velg ein testobservator og gjennomfør ein hypotesetest. Bruk signifikansnivå $\alpha = 0.05$. Bruk samandraget av dataane gitt i figur 8 når du utfører hypotesetesten. Kva blir konklusjonen din?

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

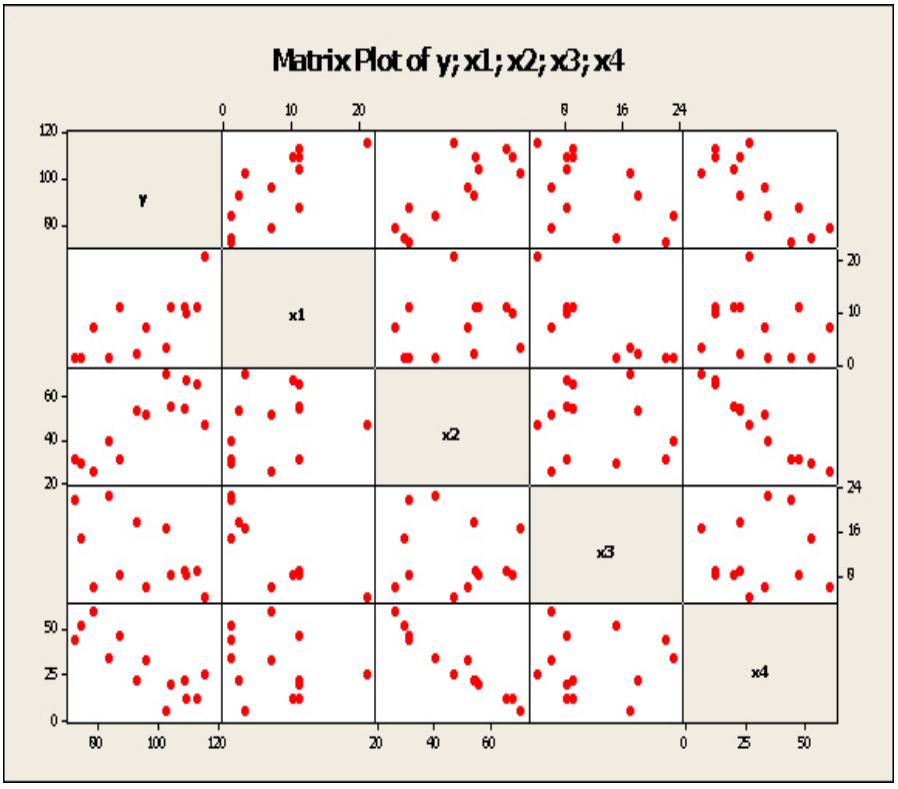
Oppgåve 4 Tilfredse kundar

Ein regional butikk-kjede utførar ein spørjeundersøking for å finne ut kor tilfredse kundane er med butikk-kjeda. Dei delte kundane i 4 geografiske regionar (Sør-Vest, Sør-Øst, Midtre og Nord). Det totale talet på personar som blei spurde i kvar av de 4 regionane blei bestemt før spørreundersøkinga blei gjennomført som ein prosentandel av populasjonsstorleiken i regionen. Kor tilfredse kundane var blei klassifisert i tre grupper og følgjande kryss-tabell blei observert.

Region/Grad av tilfreds:	Tilfreds	Veit ikkje	Ikkje 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 dataane, kan vi konkludere med at de 4 forskjellige regionane er forskjellige med omsyn på kor tilfredse kundane er? Skriv ned nullhypotesen og den alternative hypotesen og gjennomfør en hypotesetest basert på tabellen ovanfor. Bruk 5% signifikansnivå.

For å forenkle utrekningsbyrda kan du bruke at χ^2-testobservatoren blei 44.78, og du treng berre å vise korleis du reknar ut eit av de 12 ledda i summen. Kva er konklusjonen din basert på denne testen?



Figur 1: Parvis spreiingsplott av variablane 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 variablane x_1, x_2, x_3 og x_4 i sementhydratiseringsdatasettet.

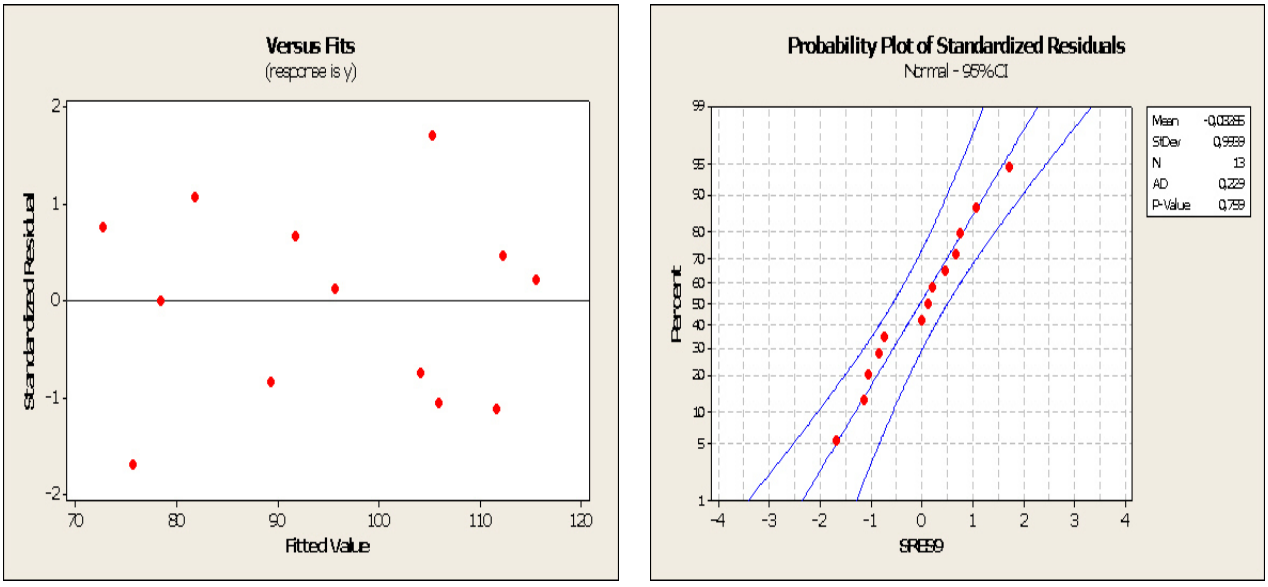
Predictor	Coef	SE Coef	T	P
Constant	62,41	70,07	0,89	0,399
x1	1,5511	0,7448	2,08	0,071
x2	0,5102	0,7238	0,70	0,501
x3	0,1019	0,7547	0,14	0,896
x4	-0,1441	0,7091	-0,20	0,844

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 frå statistisk analyse av sementhydratiseringsdataene for modell A.



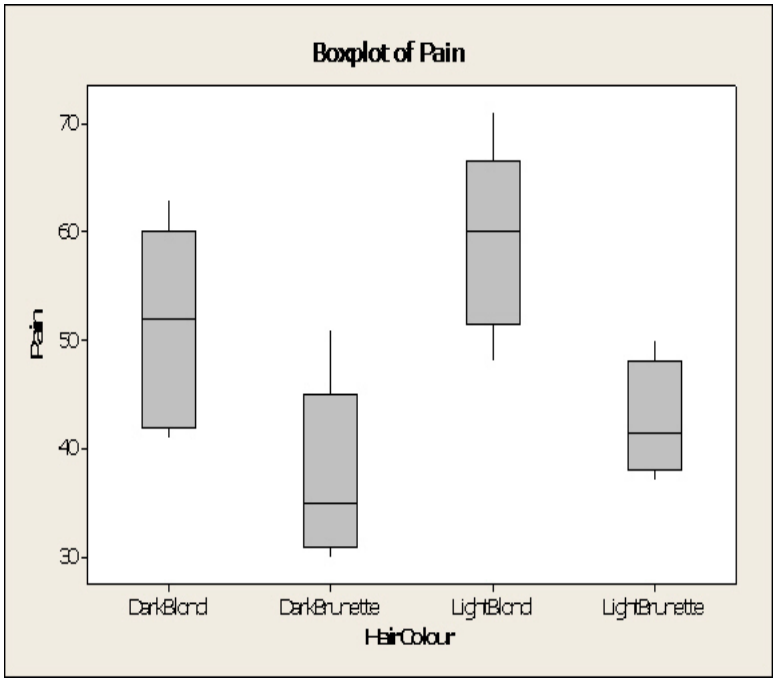
Figur 4: Residualplott (standardiserte residual mot tilpassa verdiar i venstre panel og normalplott basert på standardiserte residual i høgre panel) for regresjonsmodell A for sementhydratiseringsdatasettet.

```
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 frå 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 frå statistisk analyse av sementhydratiseringsdatasettet.



Figur 7: Boksplott av smerte og hårfargedataane.

```
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 frå statistisk analyse av smerte og hårfargedatasettet.