

Institutt for matematiske fag

Eksamensoppgave i **TMA4255 Anvendt statistikk**

Faglig kontakt under eksamen: Anna Marie Holand

Tlf: 951 38 038

Eksamensdato: August 2014

Eksamenstid (fra–til):

Hjelpemiddelkode/Tillatte hjelpemidler: Alle trykte og håndskrevne hjelpemidler. 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 Reikibehandling av Fibromyalgi

Fibromyalgi er en kronisk smertetilstand med ukjent årsak, som påvirker 2 – 4% av befolkningen. Pasienter med fibromyalgi bruker ofte komplementerende og alternativ medisin, slik som Reiki. Reiki er en form for energimedisin, der healeren kanaliserer "universell livsenergi" gjennom lett berøring.

En studie ble gjennomført for å utforske nytten av Reiki som behandling mot smerte. Hundre pasienter med fibromyalgi ble rekruttert. For å tjene som placebo behandling inkluderte studien en skuespiller som ga behandling, i tillegg til en Reikimester.

Smerte ble målt ved å bruke en visuell analog skala (VAS) (0 = ingen smerte, 10 = verste smerte noensinne).

(Kilde: Nassim Assefi, M.D., Andy Bogart, M.S., Jack Goldberg, Ph.D., and Debra Buchwald, M.D. (2008). Reiki for the Treatment of Fibromyalgia: A Randomized Controlled Trial. *Journal of Alternative and Complementary Medicine*, 14(9): 1115-1122.)

- a) De 100 pasientene ble randomisert i 4 grupper, hver på størrelse $n = 25$, hvor de fikk behandling enten med direkte berøring eller uten berøring (fra avstand) fra enten en Reikimester (referert til som Reiki behandling) eller en skuespiller (referert til som placebo behandling). VAS poengskår før behandling ble målt for hver av de $n = 25$ pasientene i hver av de 4 gruppene. La gruppe A betegne pasienter som får behandling fra en Reikimester med direkte berøring, gruppe B betegne pasienter som får behandling fra en Reikimester uten berøring (fra avstand), gruppe C betegne pasienter som får behandling fra en skuespiller med direkte berøring, gruppe D betegne pasienter som får behandling fra en skuespiller uten berøring (fra avstand).

De deskriptive målene for de forskjellige gruppene i VAS poengskår før behandling er gitt i Tabell 1.

Smertegruppe	Utvalgsstørrelse	Gjennomsnitt	Standardavvik
Gruppe A	25	6.3	2.2
Gruppe B	25	6.4	2.6
Gruppe C	25	6.8	2.1
Gruppe D	25	6.1	2.4
Totalt	100	6.4	

Tabell 1: Deskriptive mål for gruppene A, B, C og D i Reiki behandlings datasettet.

For å kunne sammenligne VAS poengskår etter behandling på tvers av gruppene, bør VAS poengskårene være lik på tvers av gruppene før behandling. Kan vi fra denne studien konkludere med at poengskårene før behandling er forskjellige mellom gruppene, A, B, C og D?

Skriv ned nullhypotesen og den alternative hypotesen, utfør en hypotesetest basert på de deskriptive målene ovenfor. Bruk signifikansnivå $\alpha = 0.05$. Spesifiser hvilke antagelser du gjør og konklusjonen av testen.

- b) Forskerne var interessert i å sammenligne den direkte berørings behandlingen, gitt enten av en Reikimester eller en skuespiller.

La X_i betegne VAS poengskår etter behandling av Reikimesteren med direkte berøring, $i = 1, \dots, 25$ og Y_j betegne VAS poengskår etter behandling av skuespilleren med direkte berøring, $j = 1, \dots, 25$.

Anta at X_i og Y_j er normalfordelt, henholdsvis $X_i \sim N(\mu_X, \sigma^2)$, $i = 1, \dots, 25$ og $Y_j \sim N(\mu_Y, \sigma^2)$, $j = 1, \dots, 25$.

Deskriptive mål for dette datasettet er $\bar{d} = \bar{x} - \bar{y} = -0.4$ og $s_p^2 = \frac{1}{48}(\sum_{i=1}^{25}(x_i - \bar{x})^2 + \sum_{i=1}^{25}(y_i - \bar{y})^2) = 1.23^2$, hvor $\bar{x} = \frac{1}{25} \sum_{i=1}^{25} x_i$ og $\bar{y} = \frac{1}{25} \sum_{i=1}^{25} y_i$.

Basert på disse dataene, har vi grunn til å tro at Reiki behandling har en større effekt på smerte, målt som VAS poengskår, enn placebo behandling gitt av en skuespiller? Skriv ned nullhypotesen og den alternative hypotesen, velg en testobservator og utfør en hypotesetest. Bruk signifikansnivå $\alpha = 0.05$.

Spesifiser hvilke antagelser du gjør.

Oppgave 2 Smaken av cheddarost

Når cheddarosten modnes skjer det forskjellige kjemiske prosesser som bestemmer smaken på sluttproduktet. Konsentrasjonen av flere kjemikalier i denne kjemiske prosessen er knyttet til smaken.

I en observasjonsstudie av smaken på cheddaroster, ble et utvalg på $n = 30$ modnede cheddarost analysert for kjemisk sammensetning og utsatt for smakstester. Den samlede poengenskjåren for testen ble funnet ved å kombinere poengene fra flere personer i et smakspanel. Følgende variabler ble målt for hver cheddarost.

- y : Den subjektive smakstest poengskåren, gjennomsnittlig smakstest poengskår varierte fra 0.7 til 57.2.
- x_1 : Den naturlige logaritmen på konsentrasjon av eddiksyre i cheddarosten.
- x_2 : Den naturlige logaritmen på konsentrasjon av hydrogensulfid i cheddarosten.
- x_3 : Konsentrasjon av melkesyre i cheddarosten.

Først ble tre separate enkle regresjonsmodeller brukt til å studere forholdet mellom smak og hver av variablene,

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad (1)$$

$$y_i = \beta_0 + \beta_2 x_{2i} + \epsilon_i, \quad (2)$$

$$y_i = \beta_0 + \beta_3 x_{3i} + \epsilon_i, \quad (3)$$

hvor ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, \dots, n$.

En MINITAB utskrift fra statistiske analyse finnes in Figur 1, 2 og 3.

Predictor	Coef	SE Coef	T	P
Constant	-61,50	24,85	-2,48	0,020
X1	15,648	4,496	3,48	0,002

S = 13,8212 R-Sq = 30,2% R-Sq(adj) = 27,7%

Figur 1: Utskrift fra statistiske analyser av cheddarostdataene for modellen in ligning (1).

Predictor	Coef	SE Coef	T	P
Constant	-9,787	5,958	-1,64	0,112
X2	5,7761	0,9458	6,11	0,000

S = 10,8334 R-Sq = 57,1% R-Sq(adj) = 55,6%

Figur 2: Utskrift fra statistiske analyser av cheddarostdataene for modellen i ligning (2).

Predictor	Coef	SE Coef	T	P
Constant	-29,86	10,58	-2,82	0,009
X3	37,720	7,186	5,25	0,000

S = 11,7450 R-Sq = 49,6% R-Sq(adj) = 47,8%

Figur 3: Utskrift fra statistiske analyser av cheddarostdataene for modellen i ligning (3).

- a) Fra Figur 1 ser vi at variabelen eddiksyre, X_1 , er signifikant ved et 5% signifikansnivå. Hva kan du konkludere med fra de gitte p-verdiene i Figur 1, 2 og 3 om de tre kjemikalienes påvirkning på smaken? Kommenter verdiene av R^2 i Figur 1, 2 og 3. Begrunn svaret.

Finn et 90% konfidensintervall for β_1 .

Hva er den predikerte poengskåret for smak for en eddiksyreverdi på $x_1^0 = 7$?

- b) Finn et passende estimat for σ , og beregn et 90% konfidensintervall for σ i regresjonsmodellen i ligning (1) (Hint: bruk at SSE/σ^2 er kjikvadrat-fordelt).

Hvordan kan vi bruke dette konfidensintervallet for å teste nullhypotesen $H_0 : \sigma = 1$? Skriv ned den alternative hypotesen som blir brukt, konklusjonen på denne testen og signifikansnivået.

Videre ble regresjon av smak på alle tre kjemikalier utført.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \quad (4)$$

hvor ϵ_i is u.i.f. $N(0, \sigma^2)$ for $i = 1, \dots, n$.

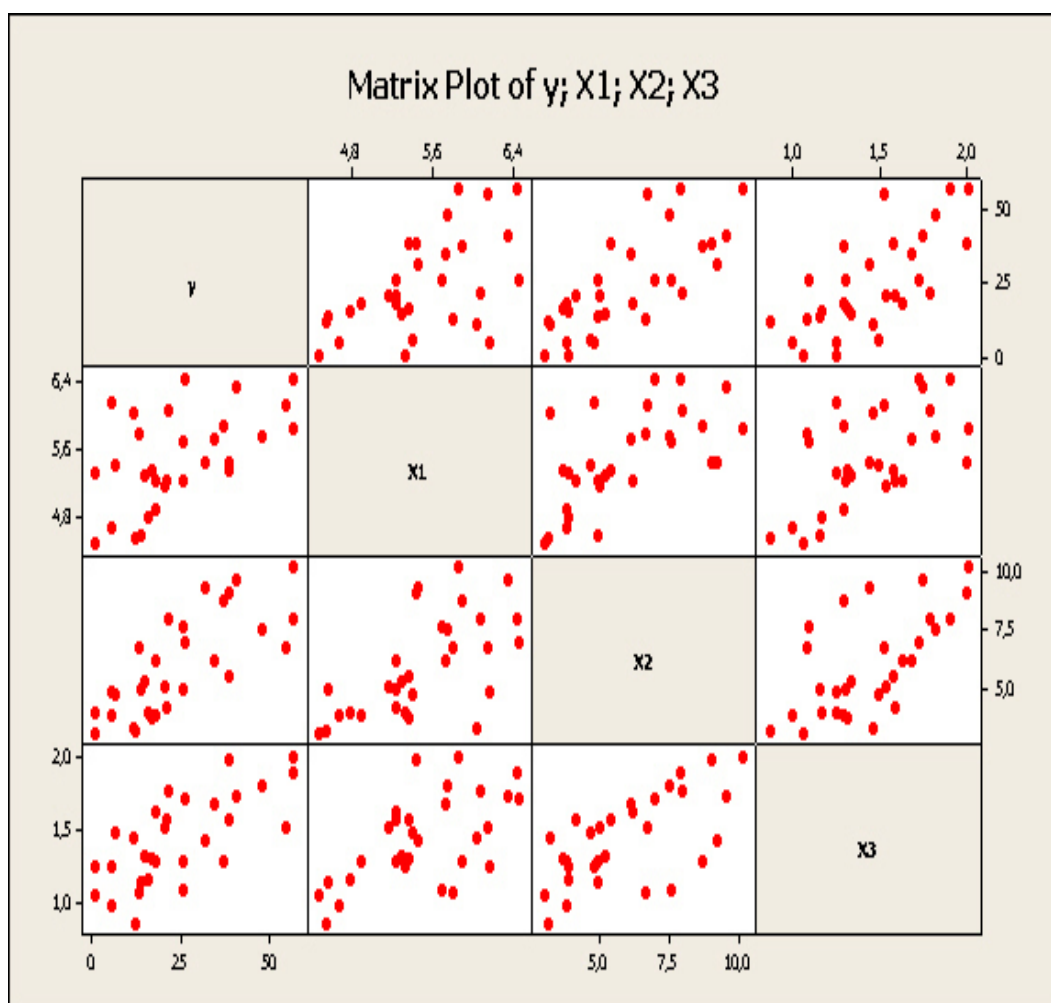
MINITAB utskriften fra statistiske analyse av denne tre-variabel modellen finnes i Figur 4. En korrelasjonsmatrise og et parvis spredningsplott finnes i henholdsvis Figur 5 og Figur 6. Residualplott finnes i Figur 7.

The regression equation is					
y = - 28,9 + 0,33 X1 + 3,91 X2 + 19,7 X3					
Predictor	Coef	SE Coef	T	P	
Constant	-28,88	19,74	-1,46	0,155	
X1	0,328	4,460	0,07	0,942	
X2	3,912	1,248	3,13	0,004	
X3	19,671	8,629	2,28	0,031	
S = 10,1307 R-Sq = 65,2% R-Sq(adj) = 61,2%					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	3	4994,5	1664,8	16,22	0,000
Residual Error	26	2668,4	102,6		
Total	29	7662,9			

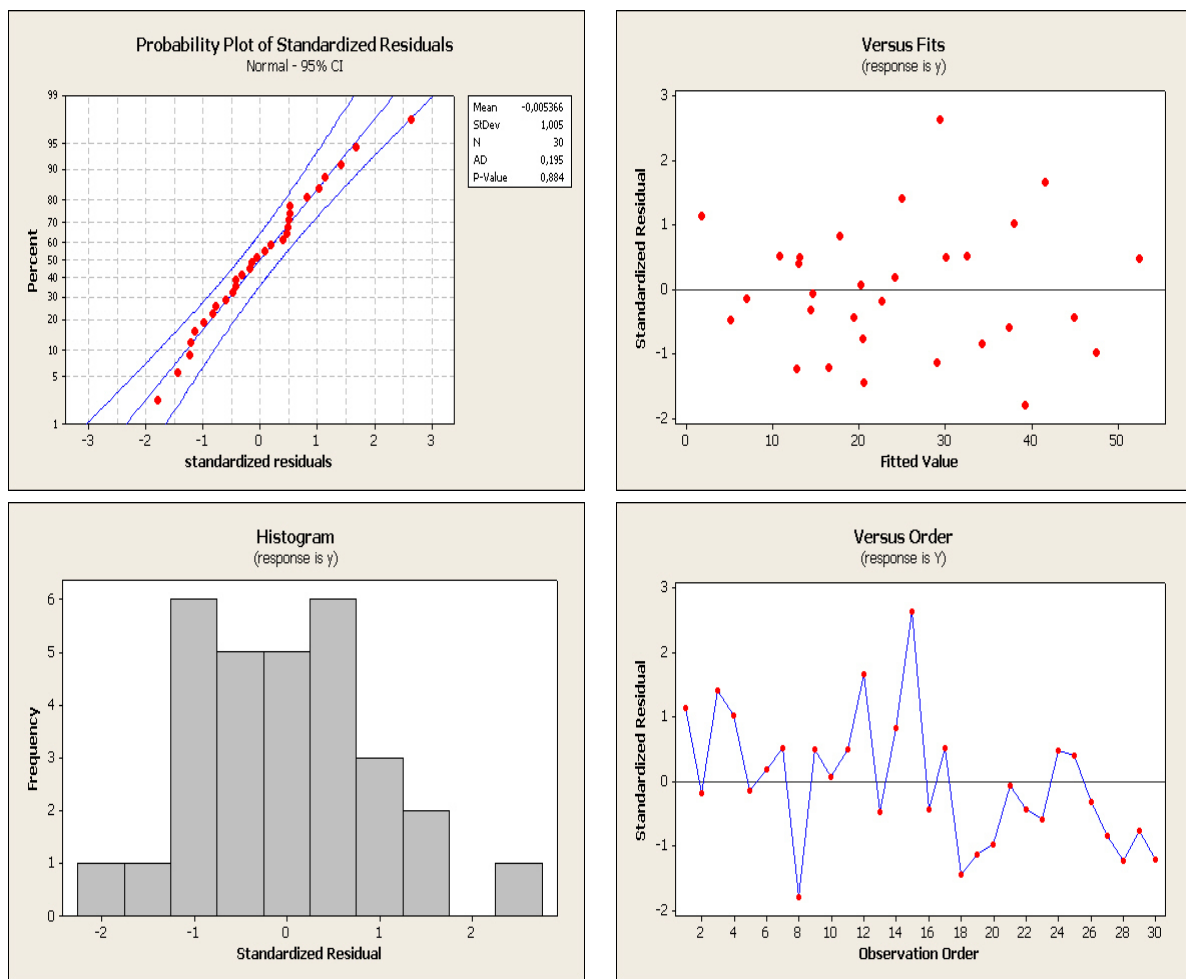
Figur 4: Utskrift fra statistiske analyser av cheddarostdataene for modellen i ligning (4).

Correlations: X1; X2; X3		
	X1	X2
X2	0,618	
X3	0,604	0,645

Figur 5: Pearson korrelasjon mellom variablene X_1 , X_2 og X_3 i cheddarostdatasettet.



Figur 6: Parvis spredningsplott av variablene i cheddarostdatasettet.



Figur 7: Residualplott (normalplott basert på standardiserte residualer i øvre venstre panel, standardiserte residualer mot tilpassede verdier i øvre høyre panel, histogram basert på standardiserte residualer i nedre venstre panel og standardiserte residualer mot rekkefølgen på observasjonene i nedre høyre panel) for regresjonsmodellen i ligning (4) for cheddarostdatasettet.

- c) Forklar begrepet multikollinearitet. Kan dette være et problem i regresjonsmodellen i ligning (4)? Kommenter korrelasjonsmatrisen i Figur 5 og det parvise spredningsplottet i Figur 6.

Fra tre-variabel regresjonmodellen i Figur 4 kan vi se at eddiksyre, X_1 , ikke er signifikant, men i Figur 1 (ligning 1) er X_1 signifikant. Hva kan være årsaken til dette? Begrunn svaret.

Dette er en observasjonsstudie. Vil det være mulig å designe et eksperiment (design of experiment) for å utforske problemet vi studerer? Utdyp.

Oppgave 3 Lekeplastklosser

En fabrikk produserer lekeplastklosser som kan settes sammen. For å være i stand til å sette sammen klossene må de ikke ha defekter. Fabrikken tar prøver for å forsikre seg at klossene er av god kvalitet. Hver dag blir det tatt et utvalg på $n = 250$ tilfeldige klosser og hver kloss blir klassifisert som enten OK eller defekt. Resultatet etter to uker er gitt under. Prosessen antas å være i kontroll i løpet av disse to ukene utvalgene ble tatt.

Dag i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Antall defekte	3	3	1	3	4	6	4	6	4	4	5	2	3	1

- a) Lag et passende kontrolldiagram for å kontrollere sannsynligheten av defekter, p . Bruk alle dataene i tabellen ovenfor for å beregne kontrollgrensene (bruk 3σ grenser).

Kan vi anta at antallet defekter i hvert utvalg er tilnærmet normalfordelt?

- b) Fabrikken er interessert i å oppdage om prosessen er ute av kontroll ved en økning i sannsynligheten av defekter. Hvor mange observasjoner, n , i hver stikkprøve trengs for å oppdage en endring fra $p = 0.2$ til $p_1 = 0.21$?

Oppgave 4 Fedme og alkoholinntak

I en studie som så på sammenhengen mellom fedme og alkoholinntak i Australia, ble et utvalg på $n_{\text{Lav}} = 165$ personer med lav grad av fedme, $n_{\text{Gjennomsnittlig}} = 161$ personer med gjennomsnittlig grad av fedme og $n_{\text{Høy}} = 165$ personer med høy grad av fedme valgt, og deres alkoholinntak i antall drikker per dag ble registrert. Alkoholinntaket ble klassifisert i fire grupper og følgende krysstabell ble observert.

	Alkoholinntak				
Fedme	0	1-2	3-5	6+	Total
Lav	45	45	41	34	165
Gjennomsnittlig	39	32	46	44	161
Høy	33	37	47	48	165
Total	117	114	134	126	491

- a) Basert på disse dataene, kan vi konkludere med at Lav, Gjennomsnittlig og Høy fedme populasjoner er forskjellige med hensyn på alkoholinntak? Skriv ned nullhypotesen og den alternative hypotesen og utfør en hypotesetest basert på tabellen ovenfor. Bruk et 5% signifikansnivå. Det er oppgitt at χ^2 -testobservatoren er lik 6.952. Du trenger bare vise beregningen av et av de 12 leddene i summen. Hva er konklusjonen fra testen?