

*Faglig kontakt under eksamen:*

Bo Lindqvist  
Tlf. 975 89 418

EKSAMEN I FAG TMA4255  
FORSØKSPLANLEGGING OG ANVENDTE STATISTISKE METODER

Onsdag 23. mai 2007

Tid: 09:00–13:00

*Tillatte hjelpemidler:*

Alle trykte og håndskrevne hjelpemidler. Alle kalkulatorer.

Sensur: 13. juni 2007

Ved løsningen av enkelte spørsmål må du muligens utføre interpolasjoner i tabellene for å finne ønskede verdier. Dette kan godt gjøres forholdsvis grovt (“på øyemål”).

### Oppgave 1

En idrettsforsker har undersøkt motorisk ferdighet for åtte år gamle gutter fra tre skoler i henholdsvis Trondheim, Namsos og Røros. Siktemålet med undersøkelsen var å finne ut om det er forskjeller i motorisk ferdighetsnivå for elever ved de tre skolene.

Guttene som deltok i undersøkelsen gjennomgikk en grovmotorisk test bestående av fire deloppgaver. På grunnlag av resultatene på de fire oppgavene ble det regnet ut en motorisk kvotient (MQ).

Vi lar  $Y_{ij}$  være målt MQ for gutt nr.  $j$  fra skole nr.  $i$ , der  $i = 1, 2, 3$  svarer til henholdsvis skolene i Trondheim, Namsos og Røros, mens  $j$  løper fra 1 til 18 som var antallet deltakere fra hver av de tre skolene.

Vi antar at  $Y_{ij}$  er normalfordelt med forventning  $\mu_i$  og varians  $\sigma^2$  ( $i = 1, 2, 3$ ;  $j = 1, 2, \dots, 18$ ), samt at alle de 54 observasjonene er uavhengige.

Resultatene står på side 3.

- a) Forskeren var i utgangspunktet bare interessert i å undersøke om det er forskjell i motorisk ferdighetsnivå mellom elevene i Trondheim og Namsos.

Bruk de oppgitte resultatene fra Trondheim og Namsos (dvs. de to første kolonnene i utskriften på side 3) til å utføre en to-utvalgs t-test for denne situasjonen. Sett opp nullhypotese og alternativ hypotese. Sett så opp og beregn testobservatoren og angi konklusjonen når du velger signifikansnivå 5%.

- b) En kollega mente at forskeren i testingsproblemet i forrige punkt burde bruke Wilcoxon's to-utvalgs test istedenfor t-testen. Forklar kort hva som eventuelt kunne være begrunnelsen for et slikt forslag.

Det viser seg at summen av rangene for utvalg nr. 1 (Trondheim) er  $w_1 = 296$ . Forklar hvordan man kommer fram til dette tallet. Du trenger ikke gjennomføre beregningen.

Beregn p-verdien for testen ved å bruke den oppgitte verdi for  $w_1$ . Hva blir konklusjonen på testen når du velger 5% signifikansnivå? Bruk tilnærming til normalfordelingen.

I resten av oppgaven brukes resultatene fra alle de tre skolene. Forskeren ønsker å finne ut om det er forskjeller mellom de tre skolene når det gjelder motorisk ferdighetsnivå for elevene. Han vil derfor teste nullhypotesen

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

ved å utføre en enveis variansanalyse.

- c) Du skal gjennomføre analysen. Beregn først kvadratsummene SSA og SSE ved å bruke resultatene som er oppgitt til slutt i oppgaven.

Beregn så testobservatoren og finn konklusjonen når du bruker signifikansnivå 5%.

Vis at estimatet for standardavviket  $\sigma$  blir  $s = 13.2$ .

- d) Det viser seg at testen i forrige punkt fører til forkastning av  $H_0$ . Forskeren er derfor interessert i å vite hvilke, om noen, av forskjellene  $\mu_1 - \mu_2$ ,  $\mu_1 - \mu_3$  og  $\mu_2 - \mu_3$  som er signifikant forskjellige fra 0.

Du skal bruke enten Bonferronis metode eller Tukeys metode på dette problemet, når det kreves at sannsynligheten er høyst 5% for å komme til minst en gal konklusjon hvis vi i virkeligheten har  $\mu_1 = \mu_2 = \mu_3$ .

j	y <sub>1j</sub>	y <sub>2j</sub>	y <sub>3j</sub>
1	100	90	90
2	86	75	114
3	71	111	110
4	80	80	127
5	95	105	102
6	77	91	72
7	104	69	80
8	96	105	94
9	91	108	93
10	82	79	96
11	85	80	96
12	72	83	85
13	92	119	106
14	97	97	108
15	82	105	108
16	85	111	83
17	71	79	89
18	78	76	95

$$\begin{aligned}\bar{y}_{1.} &= 85.78, & \sum_{j=1}^{18} (y_{1j} - \bar{y}_{1.})^2 &= 1743.0 \\ \bar{y}_{2.} &= 92.39, & \sum_{j=1}^{18} (y_{2j} - \bar{y}_{2.})^2 &= 4022.2 \\ \bar{y}_{3.} &= 97.11, & \sum_{j=3}^{18} (y_{3j} - \bar{y}_{3.})^2 &= 3103.8 \\ \bar{y}_{..} &= 91.76\end{aligned}$$

## Oppgave 2

Det er utført et forsøk for å undersøke den evne en bestemt type stålplater har til å motstå korrosjon i sjøvann. Forsøkene besto i å la platene utsettes for 10% saltsyre (HCl) ved henholdsvis 60, 70, 80 og 90 grader Celsius i perioder på 4, 6, 8, 10 og 12 timer og deretter å måle vekttapet i centigram for hver plate. (Det er ved tidligere forsøk funnet at 10% HCl angriper stålet på samme måte som sjøvann, bare mye raskere).

Det ble gjort ett forsøk ved hver nivåkombinasjon av temperatur og tid. La  $Y_i$  (centigram) være vekttap,  $x_{1i}$  (grader Celsius) være temperatur og  $x_{2i}$  (timer) være tid for  $i$ te plate.

Dataene er for ordens skyld gitt helt til slutt i oppgavesettet. Du trenger ikke disse enkelt-dataene for å løse oppgaven.

Utskrift av en MINITAB-analyse av dataene er gitt på neste side.

**Regression Analysis: y versus x1; x2**

The regression equation is

$$y = -12,8 + 0,190 x_1 + 0,439 x_2$$

Predictor	Coef	SE Coef	T	P
Constant	-12,755	3,928	-3,25	0,005
x1	0,19040	0,04782	3,98	0,001
x2	0,4391	0,1890	2,32	0,033

S = 2,39077    R-Sq = 55,6%    R-Sq(adj) = 50,3%

**Analysis of Variance**

Source	DF	SS	MS	F	P
Regression	2	121,483	60,742	10,63	0,001
Residual Error	17	97,168	5,716		
Total	19	218,652			

Source	DF	Seq SS
x1	1	90,630
x2	1	30,853

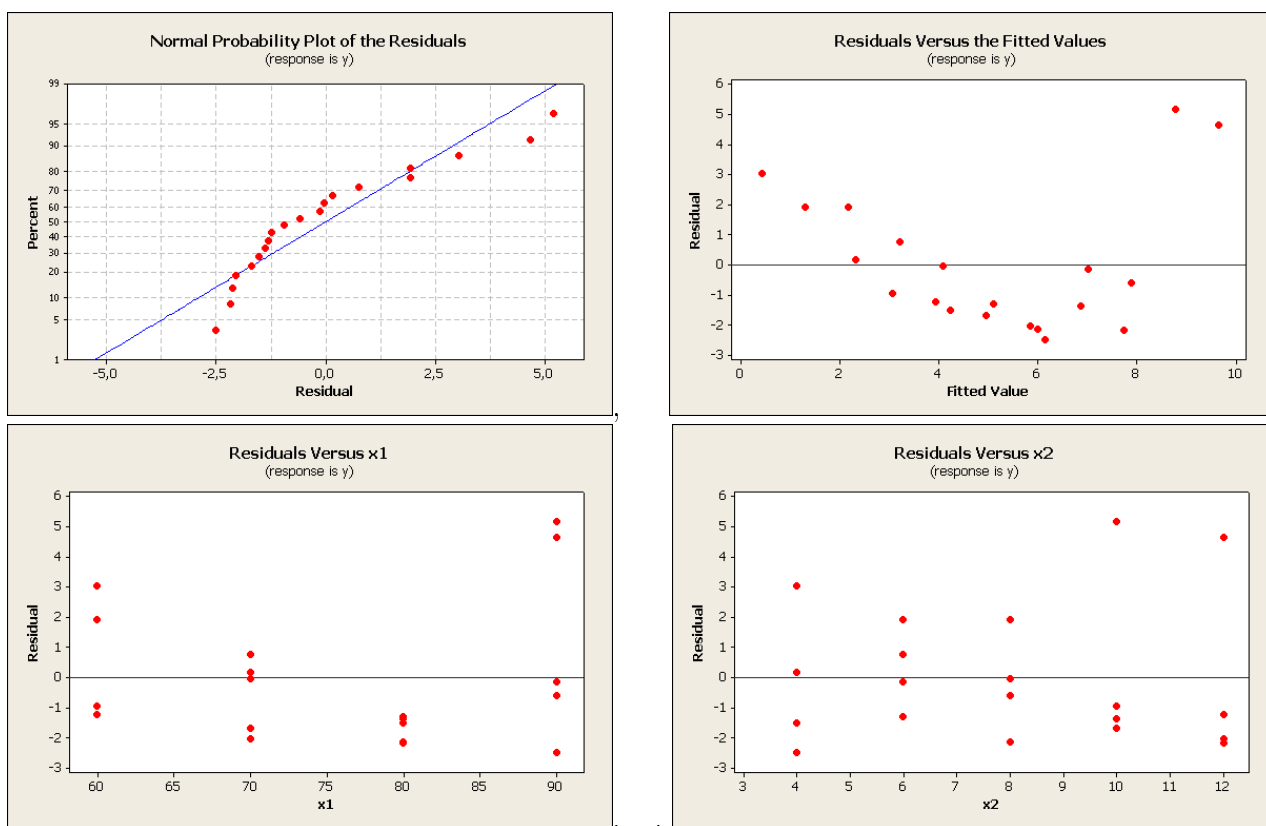
- a) Sett opp modellen og antagelsene som ligger til grunn for denne analysen.

Har modellen signifikant forklaringsgrad? Du skal besvare dette spørsmålet ved å sette opp den relevante nullhypotesen, sette opp testobservatoren, velge et signifikansnivå og konkludere ved hjelp av utskriften.

Hvor mye av variasjonen i dataene er forklart ved hjelp av modellen?

- b) Definer hva som menes med residualene til observasjonene i en regresjonsanalyse. Hvilke egenskaper har residualene? Forklar kort hvordan ulike plott av residualer kan brukes til å evaluere en foreslått regresjonsmodell.

Figur 1 på neste side viser noen residualplott for modellen som ble studert i punkt (a). Tyder disse på at modellen gir en tilfredsstillende beskrivelse av dataene? Besvar spørsmålet ved å gi korte kommentarer til hvert av de fire plottene i figuren. Hvordan burde de ideelt ha sett ut hvis modellen var korrekt?



Figur 1: Residualplott for modellen i punkt (a).

Det ble besluttet å utvide modellen fra den første analysen ved også å ta med 2. ordens ledd av  $x_1$  og  $x_2$ . Dette ble gjort ved å innføre nye prediktorer:  $x_1 * x_2$  med verdier  $x_{1i} \cdot x_{2i}$  for *ite* stålplate,  $x_1 * x_1$  med verdier  $x_{1i}^2$  og  $x_2 * x_2$  med verdier  $x_{2i}^2$ .

I det følgende er vist en MINITAB-utskrift for en regresjonsanalyse av de gitte dataene når prediktorene er  $x_1, x_2, x_1 * x_2, x_1 * x_1, x_2 * x_2$ .

### Regression Analysis: y versus x1; x2; x1\*x2; x1\*x1; x2\*x2

The regression equation is

$$y = 78,3 - 1,91 x_1 - 2,98 x_2 + 0,0494 x_1 * x_2 + 0,0114 x_1 * x_1 - 0,0176 x_2 * x_2$$

Predictor	Coef	SE Coef	T	P
Constant	78,30	16,76	4,67	0,000
x1	-1,9069	0,4311	-4,42	0,001
x2	-2,9804	0,9578	-3,11	0,008
x1*x2	0,049355	0,008946	5,52	0,000

```
x1*x1      0,011350  0,002829  4,01  0,001
x2*x2      -0,01763  0,04226  -0,42  0,683
```

S = 1,26511    R-Sq = 89,8%    R-Sq(adj) = 86,1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	196,245	39,249	24,52	0,000
Residual Error	14	22,407	1,600		
Total	19	218,652			

Source	DF	Seq SS
x1	1	90,630
x2	1	30,853
x1*x2	1	48,718
x1*x1	1	25,764
x2*x2	1	0,279

Videre er vist en MINITAB-utskrift av “best subsets” regresjon.

**Best Subsets Regression: y versus x1; x2; x1\*x2; x1\*x1; x2\*x2**

Response is y

Vars	R-Sq	R-Sq(adj)	Mallows		x x x x x					
			C-p	S	1	2	2	1	2	
1	44,0	40,9	60,5	2,6088						X
1	42,9	39,7	62,0	2,6337						X
2	72,5	69,3	23,5	1,8796		X	X			
2	70,4	66,9	26,5	1,9527			X		X	
3	77,8	73,7	18,3	1,7402	X	X	X			
3	75,3	70,7	21,7	1,8372		X	X	X		
4	89,6	86,9	4,2	1,2298	X	X	X	X		
4	82,7	78,0	13,7	1,5896	X		X	X	X	
5	89,8	86,1	6,0	1,2651	X	X	X	X	X	

- c) Hvilken modell vil du velge for dataene? Begrunn svaret både ut fra resultatet av regresjonsanalysen med den utvidede modellen, og ved å bruke “best subsets”-utskriften.

Hvor stor del av variasjonen i dataene er forklart med den valgte modellen? Sammenlign med modellen i punkt (a).

Beregn Mallows C-p også for modellen i punkt (a) og kommenter.

Hvilken matematisk motivasjon kan gis for å utvide modellen ved å ta inn de nye prediktorene?

- d) I dette punktet skal du bruke modellen der alle de fem prediktorene  $x_1, x_2, x_1 * x_2, x_1 * x_1, x_2 * x_2$  er med.

Test nullhypotesen at koeffisientene til de to annengradsleddene  $x_1 * x_1$  og  $x_2 * x_2$  begge er 0. Bruk signifikansnivå 5%.

La  $\hat{y}_0$  være den predikerte verdien for  $y_0$ , som er en tenkt respons for en observasjon ved temperatur  $x_1^0 = 80$  (grader) og tid  $x_2^0 = 10$  (timer). Regn ut verdien for  $\hat{y}_0$ .

Det kan vises at estimert standardavvik for  $\hat{y}_0$  er 0.503. Bestem ut fra dette et 95% konfidensintervall for den forventede verdi av  $y_0$ , samt et 95% prediksjonsintervall for  $y_0$  selv.

Hva er forskjellen i tolkning av de to intervallene?

### Oppgave 3

En plastprodusent har gjennomført en test av bruddstyrke for et bestemt produkt. Hver av 280 plaststaver ble støpt under samme betingelser og ble testet ved fem posisjoner. En antagelse om at hver stav har uniform sammensetning fører til at antall brudd,  $X$ , for en gitt stav er binomisk med  $n = 5$  forsøk og en ukjent sannsynlighet  $p$ . Hvis alle stavene har samme uniforme styrke, vil  $p$  også være den samme for alle de 280 stavene. Hvis stavene derimot har ulik styrke, vil  $p$ -ene variere fra stav til stav.

Man ønsker å teste nullhypotesen at alle stavene har den samme  $p$ , dvs. at forsøket gir oss  $X_1, X_2, \dots, X_{280}$  som er uavhengige og binomiske(5,  $p$ ) for en ukjent  $p$ .

Dataene er gitt ved følgende tabell:

$x =$ antall brudd for en stav	Antall staver med $X = x$
0	157
1	69
2	35
3	17
4	1
5	1

a) Vis at  $\hat{p} = 0.142$  er et rimelig estimat for  $p$  hvis nullhypotesen gjelder.

Hva blir sannsynlighetene for henholdsvis 0,1,2,3,4,5 brudd for en stav med denne verdien for  $p$ ?

Bruk disse sannsynlighetene til å gjennomføre en kji kvadrat-test for den gitte nullhypotesen. Velg signifikansnivå 1%. Gi en begrunnelse for antallet frihetsgrader for testobservatoren.

Hvorfor bør de tre siste cellene, svarende til  $X = 3, 4, 5$ , slås sammen til en celle ved testingen?

Dataene fra Oppgave 2:

Row	x1	x2	y
1	60	4	3,47
2	70	4	2,49
3	80	4	2,71
4	90	4	3,64
5	60	6	3,24
6	70	6	3,97
7	80	6	3,80
8	90	6	6,88
9	60	8	4,11
10	70	8	4,05
11	80	8	3,86
12	90	8	7,30
13	60	10	2,11
14	70	10	3,28
15	80	10	5,49
16	90	10	13,96
17	60	12	2,71
18	70	12	3,80
19	80	12	5,58
20	90	12	14,31