EKSAMENSOPPGAVE

Eksamen i: STA-3300, Brukerkurs i statistikk 2

Dato : Onsdag 5. desember 2007

Tid : 0900 - 1300

Sted : Åsgårdv. 9

Tillatte hjelpemidler: - Alle trykte og skrevne

- Kalkulator

Oppgavesettet er på 14 sider inkl. forside

Kontaktperson: Georg Elvebakk, 77646532

DET MATEMATISK-NATURVITENSKAPELIGE FAKULTET

Universitetet i Tromsø, N-9037 Tromsø, Telefon 77 64 40 01, Telefaks 77 64 47 65

Deloppgavene vil telle likt ved vurderinga. SPSS-utskrifter er lengst bak i oppgavesettet.

Oppgave 1

Vi skal i denne oppgava se på pH-verdier fra en del innsjøer i Norge. Disse dataene blei samla inn for noen år tilbake av forskere fra nartur- og miljøforvaltning. For 26 tilfeldig valgte innsjøer er PH-verdien målt. I tillegg har en registert den såkalte fiskestatusen til innsjøen, det vil si om fiskestammen i innsjøen er blitt skadd eller fremdeles er i god forfatning. Vi kaller pH-verdien Y og fiskestatus-variabelen Z. Z kan ha to verdier.

0 Uskadd eller lettere skadd fiskestamme.
1 Sterkt skadd eller fisketom innsjø.

Dataene for de 26 tilfeldig valgte vann er lista i SPSS-utskriftene lengre bak i oppgavesettet.

Forskerne starter med å dele datasettet i to deler (16 har fiskestatus 0 og 10 har fiskestatus 1). De vil undersøke om innsjøer som er uskadd eller lettere skadd har en forventa pH-verdi som er forskjellig fra innsjøer som er sterkt skadde eller fisketomme.

a) Finn estimert forskjell i forventningsverdi av pH mellom de to gruppene?
Finn estimert felles standardavik, S_p.
Sett opp en modell for dataene og lag et 95% konfidensintervall for forskjellen i forventningsverdi. Kan du konkludere at det er en signifikant forskjell?

Nå har faktisk forskerene samla inn enda meir nøyaktige data om fiskestatusen. De to gruppene kan derfor hver splittes i to. Kaller denne fiskestatusvariabelen ZNY, den kan ha fire verdier:

1 Uskadd
2 Lettere skadd
3 Sterkt skadd
4 Fisketomt

Forskerene vil igjen undersøke om forventa pH-verdi er forskjellig for innsjøer med ulik fiskestatus (ulik verdi for ZNY).

b) Sett opp en modell for pH-verdien for de fire gruppene.
Sett opp og utfør en test for om pH-verdien er ulik i innsjøer med ulik fiskestatus. Bruk signifikansnivå 5%.

Ettersom konklusjonen i forrige punkt blei at det er forskjeller i pH-verdi mellom de ulik gruppene innsjøer, vil forskerne gjerne si noe meir om hvilke grupper som skiller seg ut eller er like.

c) Forklar hvorfor metoder for multiple sammenlikninger bør benyttes når en vil lage konfidensintervall for alle parvise sammenlikninger av forventningsverdier. Sett opp 95%-konfidensintervaller for forskjellen i pH-verdi mellom de ulike gruppene. Hvilken metode bør du velge her? Konkluder hvilke grupper du kan si har lik/ulik forventa pH-verdi.

Oppgave 2

I denne oppgava skal vi arbeide videre med dataene for pH i norske innsjøer. Vi skal fremdeles bruke de 26 tilfeldig valgte innsjøene fra forrige oppgave, men nå vil vi forsøke å finne ut hvilke kjemiske stoffer som fører til at pH-verdien i en innsjø minker eller øker. I tillegg til pH-verdien har vi derfor målt variablene

Variabelnavn	Forklaring
X_1	Innhold av SO_4 (mg/l)
X_2	Innhold av NO_3 (mg/l)
X_3	Innhold av Ca (mg/l)
X_4	Innhold av latent aluminium (μg/l)
X_5	Innhold av organiske stoffer (mg/l)
Y	pH-verdi

Det er velkjent fra kjemi at økning av X_1, X_2 eller X_4 hver for seg vil føre til nedgang i pH-verdi, mens økning av X_3 (kalk) og X_5 vil føre til at pH-verdien stiger. Forskerne vil undersøke er hvilke av disse variablene som har betydning i praksis, de vil lage en modell som gir god prediksjon av pH-verdien. De begynner med å lage en enkel lineær regresjonsmodell.
SPSS-utskrifter er vedlagt lengre bak i oppgavesettet.

a) Forklar hvorfor X_4 er det mest naturlige valget som uavhengig variabel i en enkel lineær regresjonsmodell for Y?
Har korrelasjonsmatrisen egenskaper som kan være til ulempe når eventuelt fleire av forklaringsvariablene (X-variablene) tas med i modellen?
Skriv opp den enkle modellen med $X4$.
Modellen skal estimeres ved minste kvadrat-metoden. Forklar kort hva som meines med dette og presiser forutsetningene for modellen.
Hva blir den estimerte modellen? Forklar med ord hva den sier?

b) Vi velger oss en ny innsjø og måler innholdet av aluminium til 70 mg/l. Hva blir anslått pH-verdi i følge modellen fra forrige punkt?
Forklar forskjellen på konfidens- og prediksjonsintervallet for pH-verdien.
Konstruer et intervall som med 95% sannsynlighet vil inneholde pH-verdien i innsjøen. Her kan du bruke at $\frac{(70-\overline{X4})^2}{(26-1)\cdot S^2_{X4}} = 0.0049$.

Forskerne vurderer nå om de kan få en enda bedre modell ved å ta med alle de uavhengige variablene (X-variablene). Denne modellen kan vi kalle den maksimale modellen.

c) Sett opp den maksimale modellen med forutsetninger og skriv opp den estimerte modellen.
Tyder utskriftene på at alle variablene i denne modellen har betydning?
Forklar hva den partielle korrelasjonen

$$\rho_{Y,(X1,X2,X3,X5)|X4}$$

måler, og finn et estimat for den.
Sett opp og utfør en test for om den partielle korrelasjonen er ulik 0. Hva forteller konklusjonen av denne testen deg om den enkle kontra den maksimale modellen?

Forskerne vil nå prøve å velge ut en enda bedre modell ved hjelp av variabelutvelgingsmetoder. I utskriftene er det utført to prosedyrer, en med forlengs-utvelging og en med baklengs-eliminasjon.

d) Forklar generelt hvordan en baklengseliminering smetode utføres.
Sett opp og utfør testen som gjennomføres når den første variabelen blir eliminert fra modellen.

Fra de to utvelgingsmetodene får de to ulike modeller.

e) Skriv opp de to estimerte sluttmodellene og bruk R^2, MSE, C_p og signifikans av enkeltvariable til å velge hvilken av dem du meiner er den beste.
Utfør **en** test for om $X4$ og $X5$ er uten betyding i den maksimale modellen.

Til slutt ser de på sluttmodellen fra bakovereliminasjonen.

f) Forklar kort hvordan de tre uavhengige variablene (X-variablene) påvirker pH-verdien i følge den estimerte modellen.
En av måtene å få opp pH-verdien i en innsjø er å tilsette kalk (Ca). En tommelfingerregel sier at dersom en tilsetter kalk slik at det blir en økning på 1 mg/l Ca, når SO_4 og NO_3 ikke endres, vil pH-verdien stige med 1.0.
Hva sier den estimerte modellen om dette?
Utfør en test for om vi har grunn til å forkaste denne tommelfingerregelen.

g) Hvordan er residualene definert. Rekn ut residualet til observasjon nummer 1.
Tyder plott av residualene, oppgitte outlier- og kolinearitetsmål på at denne modellen oppfyller forutsetningene for regresjonsanalysen.

SPSS-utskrifter til oppgave 1:

Case Summaries[a]

		Y	Z	ZNY
1		5,38	1,00	3,00
2		5,68	,00	1,00
3		5,04	1,00	3,00
4		4,81	1,00	4,00
5		4,92	1,00	4,00
6		5,34	1,00	3,00
7		5,74	,00	1,00
8		5,17	1,00	4,00
9		5,02	1,00	4,00
10		5,88	,00	2,00
11		5,36	,00	2,00
12		5,26	1,00	3,00
13		5,69	,00	2,00
14		5,51	1,00	3,00
15		5,25	1,00	4,00
16		6,06	,00	1,00
17		6,08	,00	1,00
18		6,08	,00	2,00
19		6,20	,00	1,00
20		5,64	,00	1,00
21		5,76	,00	1,00
22		5,43	,00	2,00
23		5,82	,00	1,00
24		5,50	,00	2,00
25		5,62	,00	2,00
26		5,41	,00	1,00
Total	N	26	26	26

a. Limited to first 100 cases.

To-utvalgs-T-test:

Group Statistics

	Z	N	Mean	Std. Deviation	Std. Error Mean
Y	,00	16	5,7469	,25927	,06482
	1,00	10	5,1700	,21975	,06949

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Y	Equal variances assumed	,270	,608	5,836	24	,000	,57688	,09884	,37287	,78088
	Equal variances not assumed			6,071	21,644	,000	,57688	,09503	,37961	,77414

Envegs-ANOVA:

Descriptives

Y

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1,00	9	5,8211	,24992	,08331	5,6290	6,0132	5,41	6,20
2,00	7	5,6514	,25667	,09701	5,4140	5,8888	5,36	6,08
3,00	5	5,3060	,17401	,07782	5,0899	5,5221	5,04	5,51
4,00	5	5,0340	,17925	,08016	4,8114	5,2566	4,81	5,25
Total	26	5,5250	,37368	,07328	5,3741	5,6759	4,81	6,20

ANOVA

Y

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2,346	3	,782	15,032	,000
Within Groups	1,145	22	,052		
Total	3,491	25			

Multiple Comparisons

Dependent Variable: Y

	(I) ZNY	(J) ZNY	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1,00	2,00	,16968	,11495	,468	-,1495	,4889
		3,00	,51511*	,12723	,003	,1618	,8684
		4,00	,78711*	,12723	,000	,4338	1,1404
	2,00	1,00	-,16968	,11495	,468	-,4889	,1495
		3,00	,34543	,13356	,074	-,0254	,7163
		4,00	,61743*	,13356	,001	,2466	,9883
	3,00	1,00	-,51511*	,12723	,003	-,8684	-,1618
		2,00	-,34543	,13356	,074	-,7163	,0254
		4,00	,27200	,14426	,263	-,1286	,6726
	4,00	1,00	-,78711*	,12723	,000	-1,1404	-,4338
		2,00	-,61743*	,13356	,001	-,9883	-,2466
		3,00	-,27200	,14426	,263	-,6726	,1286
Scheffe	1,00	2,00	,16968	,11495	,547	-,1780	,5173
		3,00	,51511*	,12723	,006	,1303	,8999
		4,00	,78711*	,12723	,000	,4023	1,1719
	2,00	1,00	-,16968	,11495	,547	-,5173	,1780
		3,00	,34543	,13356	,113	-,0585	,7494
		4,00	,61743*	,13356	,002	,2135	1,0214
	3,00	1,00	-,51511*	,12723	,006	-,8999	-,1303
		2,00	-,34543	,13356	,113	-,7494	,0585
		4,00	,27200	,14426	,338	-,1643	,7083
	4,00	1,00	-,78711*	,12723	,000	-1,1719	-,4023
		2,00	-,61743*	,13356	,002	-1,0214	-,2135
		3,00	-,27200	,14426	,338	-,7083	,1643
Bonferroni	1,00	2,00	,16968	,11495	,924	-,1635	,5029
		3,00	,51511*	,12723	,003	,1463	,8839
		4,00	,78711*	,12723	,000	,4183	1,1559
	2,00	1,00	-,16968	,11495	,924	-,5029	,1635
		3,00	,34543	,13356	,101	-,0417	,7326
		4,00	,61743*	,13356	,001	,2303	1,0046
	3,00	1,00	-,51511*	,12723	,003	-,8839	-,1463
		2,00	-,34543	,13356	,101	-,7326	,0417
		4,00	,27200	,14426	,436	-,1461	,6901
	4,00	1,00	-,78711*	,12723	,000	-1,1559	-,4183
		2,00	-,61743*	,13356	,001	-1,0046	-,2303
		3,00	-,27200	,14426	,436	-,6901	,1461

*. The mean difference is significant at the .05 level.

SPSS-utskrifter til oppgave 2:

Case Summaries[a]

		Y	X1	X2	X3	X4	X5
1		5,38	4,90	39,00	1,54	78,00	2,02
2		5,68	4,10	75,00	1,55	17,00	2,98
3		5,04	3,50	80,00	,83	157,00	3,40
4		4,81	3,80	75,00	,53	163,00	3,42
5		4,92	3,80	90,00	,82	105,00	3,41
6		5,34	2,60	49,00	,62	114,00	1,90
7		5,74	2,70	79,00	1,08	15,00	2,53
8		5,17	2,60	90,00	,67	90,00	1,89
9		5,02	2,40	64,00	,41	107,00	,97
10		5,88	2,80	27,00	1,15	12,00	3,04
11		5,36	3,40	13,00	,89	93,00	2,95
12		5,26	2,70	14,00	,74	72,00	2,75
13		5,69	3,20	13,00	1,03	99,00	3,34
14		5,51	2,50	79,00	,67	80,00	,78
15		5,25	1,50	77,00	,33	72,00	,11
16		6,06	3,70	15,00	1,94	4,00	5,53
17		6,08	1,90	16,00	1,05	,00	5,63
18		6,08	1,30	2,00	,81	,00	3,84
19		6,20	2,20	32,00	1,40	7,00	3,77
20		5,64	2,20	21,00	,75	13,00	5,62
21		5,76	1,60	7,00	,79	2,00	3,95
22		5,43	2,00	27,00	,47	16,00	2,39
23		5,82	1,80	17,00	,74	3,00	3,08
24		5,50	2,00	6,00	,49	10,00	2,82
25		5,62	1,50	3,00	,36	7,00	1,71
26		5,41	1,70	7,00	,54	9,00	6,48
Total	N	26	26	26	26	26	26

a. Limited to first 100 cases.

Correlations

		Y	X1	X2	X3	X4	X5
Y	Pearson Correlation	1	-,336	-,571**	,526**	-,813**	,406*
	Sig. (2-tailed)		,093	,002	,006	,000	,040
	N	26	26	26	26	26	26
X1	Pearson Correlation	-,336	1	,409*	,583**	,514**	-,024
	Sig. (2-tailed)	,093		,038	,002	,007	,907
	N	26	26	26	26	26	26
X2	Pearson Correlation	-,571**	,409*	1	-,059	,597**	-,478*
	Sig. (2-tailed)	,002	,038		,776	,001	,013
	N	26	26	26	26	26	26
X3	Pearson Correlation	,526**	,583**	-,059	1	-,245	,366
	Sig. (2-tailed)	,006	,002	,776		,227	,066
	N	26	26	26	26	26	26
X4	Pearson Correlation	-,813**	,514**	,597**	-,245	1	-,394*
	Sig. (2-tailed)	,000	,007	,001	,227		,046
	N	26	26	26	26	26	26
X5	Pearson Correlation	,406*	-,024	-,478*	,366	-,394*	1
	Sig. (2-tailed)	,040	,907	,013	,066	,046	
	N	26	26	26	26	26	26

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Regresjon av Y mot X4:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X4[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,662	,647	,22188

a. Predictors: (Constant), X4

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2,309	1	2,309	46,910	,000[a]
	Residual	1,182	24	,049		
	Total	3,491	25			

a. Predictors: (Constant), X4

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5,827	,062		94,107	,000
	X4	-,006	,001	-,813	-6,849	,000

a. Dependent Variable: Y

Maksimalmodellen:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X5, X1, X2, X4, X3[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,963[a]	,928	,910	,11227

a. Predictors: (Constant), X5, X1, X2, X4, X3

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,239	5	,648	51,389	,000[a]
	Residual	,252	20	,013		
	Total	3,491	25			

a. Predictors: (Constant), X5, X1, X2, X4, X3

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5,768	,088		65,746	,000
	X1	-,322	,054	-,801	-5,927	,000
	X2	-,002	,001	-,180	-2,192	,040
	X3	,930	,114	,997	8,176	,000
	X4	-,001	,001	-,089	-,746	,465
	X5	-,024	,018	-,099	-1,329	,199

a. Dependent Variable: Y

Forlengsutvelging:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	X4	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
2	X3	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
3	X1	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)

a. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,662	,647	,22188
2	,880[b]	,775	,755	,18487
3	,954[c]	,909	,897	,11991

a. Predictors: (Constant), X4

b. Predictors: (Constant), X4, X3

c. Predictors: (Constant), X4, X3, X1

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2,309	1	2,309	46,910	,000[a]
	Residual	1,182	24	,049		
	Total	3,491	25			
2	Regression	2,705	2	1,352	39,570	,000[b]
	Residual	,786	23	,034		
	Total	3,491	25			
3	Regression	3,175	3	1,058	73,599	,000[c]
	Residual	,316	22	,014		
	Total	3,491	25			

a. Predictors: (Constant), X4

b. Predictors: (Constant), X4, X3

c. Predictors: (Constant), X4, X3, X1

d. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5,827	,062		94,107	,000
	X4	-,006	,001	-,813	-6,849	,000
2	(Constant)	5,518	,104		52,921	,000
	X4	-,005	,001	-,728	-7,134	,000
	X3	,324	,095	,347	3,401	,002
3	(Constant)	5,679	,073		77,518	,000
	X4	-,001	,001	-,154	-1,282	,213
	X3	,902	,118	,966	7,612	,000
	X1	-,330	,058	-,820	-5,716	,000

a. Dependent Variable: Y

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	X1	,112[a]	,801	,431	,165	,735
	X2	-,133[a]	-,891	,382	-,183	,643
	X3	,347[a]	3,401	,002	,579	,940
	X5	,101[a]	,772	,448	,159	,844
2	X1	-,820[b]	-5,716	,000	-,773	,200
	X2	-,182[b]	-1,509	,146	-,306	,635
	X5	-,011[b]	-,095	,925	-,020	,767
3	X2	-,138[c]	-1,792	,088	-,364	,629
	X5	-,037[c]	-,493	,627	-,107	,765

a. Predictors in the Model: (Constant), X4

b. Predictors in the Model: (Constant), X4, X3

c. Predictors in the Model: (Constant), X4, X3, X1

d. Dependent Variable: Y

Baklengseliminasjon:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X5, X1, X2, X4, X3[a]	.	Enter
2	.	X4	Backward (criterion: Probability of F-to-remove >= ,100).
3	.	X5	Backward (criterion: Probability of F-to-remove >= ,100).

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,963[a]	,928	,910	,11227
2	,962[b]	,926	,912	,11108
3	,959[c]	,920	,909	,11300

a. Predictors: (Constant), X5, X1, X2, X4, X3

b. Predictors: (Constant), X5, X1, X2, X3

c. Predictors: (Constant), X1, X2, X3

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,239	5	,648	51,389	,000[a]
	Residual	,252	20	,013		
	Total	3,491	25			
2	Regression	3,232	4	,808	65,482	,000[b]
	Residual	,259	21	,012		
	Total	3,491	25			
3	Regression	3,210	3	1,070	83,800	,000[c]
	Residual	,281	22	,013		
	Total	3,491	25			

a. Predictors: (Constant), X5, X1, X2, X4, X3

b. Predictors: (Constant), X5, X1, X2, X3

c. Predictors: (Constant), X1, X2, X3

d. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	5,768	,088		65,746	,000		
	X1	-,322	,054	-,801	-5,927	,000	,198	5,053
	X2	-,002	,001	-,180	-2,192	,040	,538	1,859
	X3	,930	,114	,997	8,176	,000	,243	4,114
	X4	-,001	,001	-,089	-,746	,465	,256	3,907
	X5	-,024	,018	-,099	-1,329	,199	,654	1,529
2	(Constant)	5,772	,087		66,578	,000		
	X1	-,352	,035	-,877	-10,015	,000	,461	2,168
	X2	-,002	,001	-,197	-2,538	,019	,586	1,707
	X3	,990	,079	1,061	12,529	,000	,493	2,030
	X5	-,024	,018	-,098	-1,329	,198	,654	1,528
3	(Constant)	5,701	,069		82,063	,000		
	X1	-,349	,036	-,870	-9,783	,000	,463	2,160
	X2	-,002	,001	-,156	-2,150	,043	,699	1,430
	X3	,955	,076	1,024	12,600	,000	,554	1,805

a. Dependent Variable: Y

Excluded Variables[c]

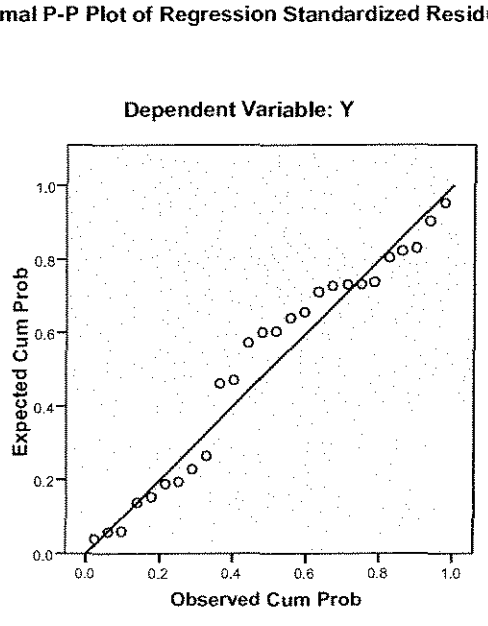
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
2	X4	-,089[a]	-,746	,465	-,164	,256	3,907	,198
3	X4	-,085[b]	-,707	,487	-,152	,256	3,905	,198
	X5	-,098[b]	-1,329	,198	-,279	,654	1,528	,461

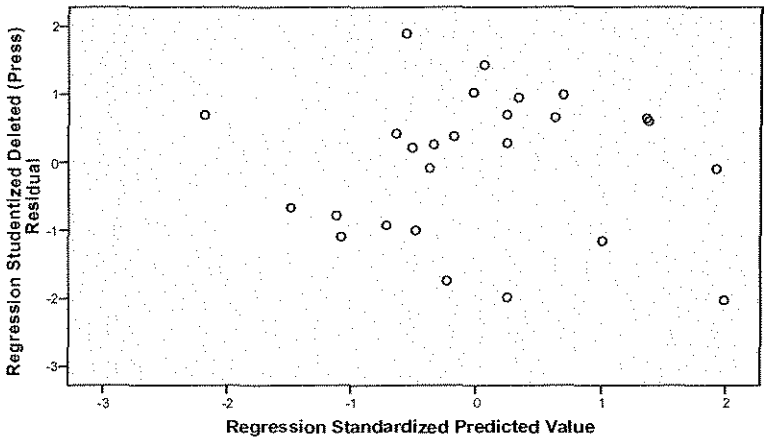
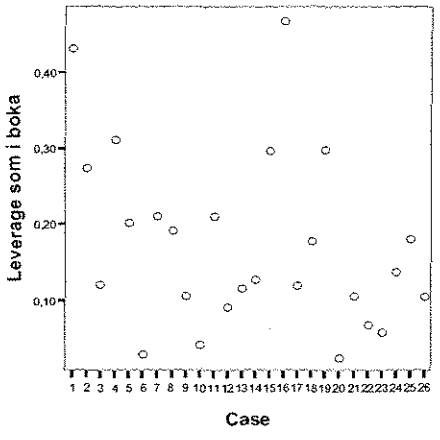
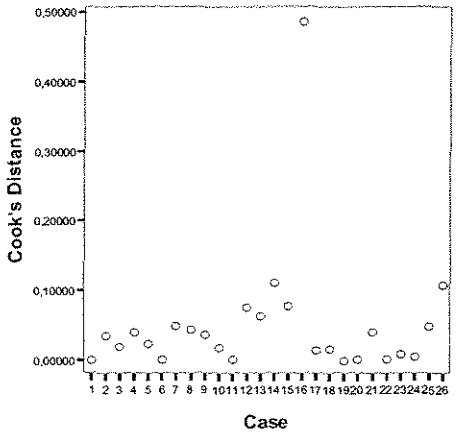
a. Predictors in the Model: (Constant), X5, X1, X2, X3

b. Predictors in the Model: (Constant), X1, X2, X3

c. Dependent Variable: Y

Ulike plott for sluttmodellen fra baklengseliminasjonen:



Scatterplot

Dependent Variable: Y