EKSAMEN I STA-3300

Eksamen i	:	**STA-3300 Brukerkurs i statistikk 2**
Dato	:	**Tirsdag 5. desember 2006**
Tid	:	**Kl 0900-1300**
Sted	:	**Åsgårdvegen 9**
Hjelpemiddel	:	**Alle skriftlige og alle kalkulatorer**

Oppgavesettet er på 10 sider inkl. forside

Kontaktperson under eksamen: Amanuensis Elinor Ytterstad, tlf 77 64 4015

Oppgavesettet har totalt ni delspørsmål som alle teller likt.

Oppgave 1 (delspørsmål a-h)

I denne oppgaven skal vi analysere amerikanske data fra "the Garbage Project, University of Arizona". Sortert søppel for ei hel uke fra n=62 husstander er veid og notert. Vektene er i pund. Det er også notert hvor mange personer som bodde i husstanden.

Variablene er:

Y=personer –antall personer i husstanden
metall - vekt metallavfall
glass - vekt glassavfall
papir - vekt papiravfall
plast - vekt plastavfall
mat - vekt matavfall
rest - vekt restavfall

Vi skal undersøke om vi kan si noe om antall personer i husstanden fra mengde og type søppel de produserer. Har søppeltype klar sammenheng med antall personer?

a) Hva kan en si om sammenhengen mellom type søppel og antall personer fra spredningsplottene og korrelasjonene?
Regn ut partiell korrelasjon mellom "personer" og "plast", kontrollert for "rest".
Kommenter forskjellen mellom denne og tilsvarende nulte-ordens korrelasjon.

Vi skal først undersøke om det er interaksjoner mellom "rest" og de andre uavhengige variablene.

Modell 1:

$$Y = \beta_0 + \beta_1 metall + \beta_2 glass + \beta_3 papir + \beta_4 plast + \beta_5 mat + \beta_6 rest + E$$

Modell 2:

$$Y = \beta_0 + \beta_1 metall + \beta_2 glass + \beta_3 papir + \beta_4 plast + \beta_5 mat + \beta_6 rest + \beta_7 rest \cdot metall + \beta_8 rest \cdot glass + \beta_9 rest \cdot papir + \beta_{10} rest \cdot plast + \beta_{11} rest \cdot mat + E$$

Alle analyseresultat fra SPSS er lagt inn sist i oppgaven.

b) Skriv opp estimert verdi av koeffisienten til interaksjonen mellom rest og metall.
Forklar hva denne verdien forteller om antall personer i husstander med lite og mye restavfall.
Test hypotesen: $H_0 : \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$
Hva er det som blir testet med denne hypotesen? Hva blir konklusjonen av testen?

Vi skal gå videre med modellen uten interaksjonar (modell 1)

c) Kan en ut fra spredningsplottene på side 3 si noe om svikt i forutsetningene for regresjonsmodellen eller om det er "uteligger"-observasjoner. Begrunn svaret.

d) Bruk hypotesetesting og finn variabelen (i modell 1) som er minst signifikant.
Formuler hypotesen og skriv opp P-verdien.
Hvordan stemmer dette med det du ser av spredningsplottet mellom denne variabelen og antall personer?
Hva er korrelasjonen mellom denne variabelen og antall personer?

e) Forklar hva "Forward selection"- prosedyren går ut på.
Hvilken variabel i modell 1 vil denne prosedyren legge inn først i modellen?

Resultatet av "Forward"-, Backward"- og "Stepwise" -prosedyrene med variablene i modell 1 som mulige variabler:

"Forward" og "Stepwise": plast, rest
"Backward": plast, rest, glass, metall

Her er P-verdi-grensen for å ta inn en variabel lik 0.05 og tilsvarende 0.10 for å fjerne en variabel fra modellen.

Vi skal i resten av oppgaven se på modellene:

Modell	
1, metall, glass, papir, plast, mat, rest	Full modell
3, plast, rest	"Forward" og "Stepwise"
4, plast, rest, glass	
5, plast, rest, glass, metall	"Backward"
6, plast, glass, metall	

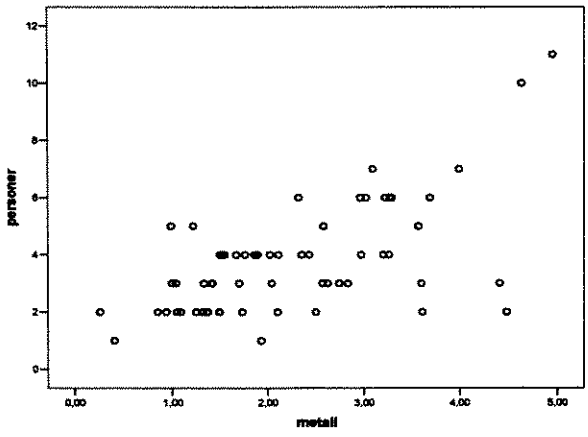
f) Fyll inn det som mangler i tabellen under.
Forklar hva R^2 måler.

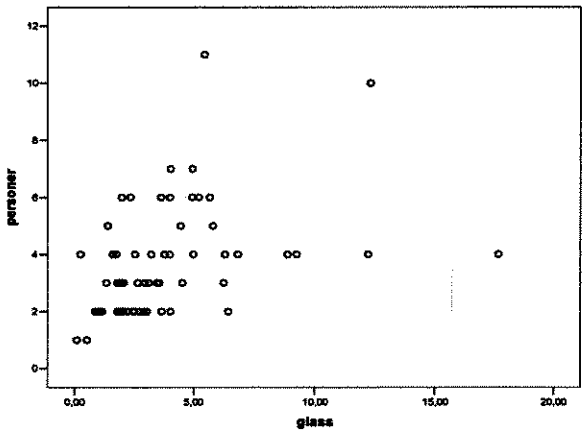
modell	SSY	SSE	R^2	MSE	Antall uavhengige variabler	Mallow's C_p i forhold til modell 1
1	232.774		0.696	1.285	6	7
3			0.658	1.350		5.97
4		75.159		1.296		4.49
5		71.327				3.51
6		85.765	0.632	1.479		

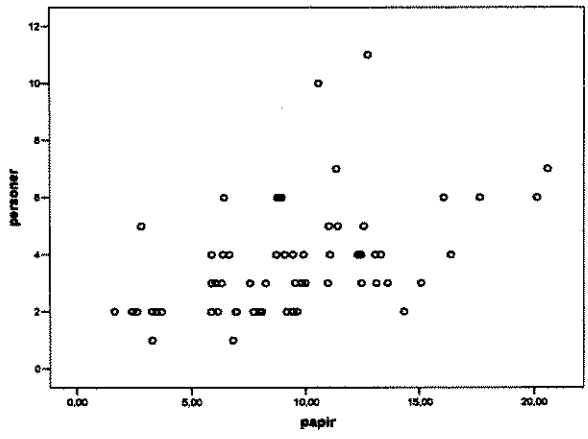
g) Gjør en "added-last-test" på 5% nivå av "glass" i modell 4.
Regn og ut et passende konfidensintervall for denne koeffisienten og forklar hvordan en kan bruke konfidensintervallet i hypotesetesten.

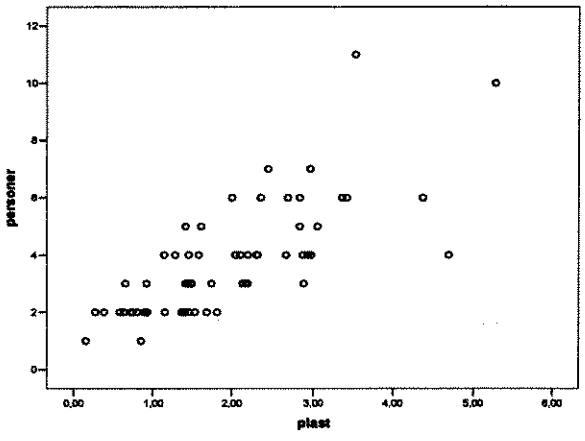
h) Hvilken modell av 2, 3, 4, 5 eller 6 er beste modell for sammenhengen mellom husstandens søppelmengde og størrelse. Begrunn svaret.

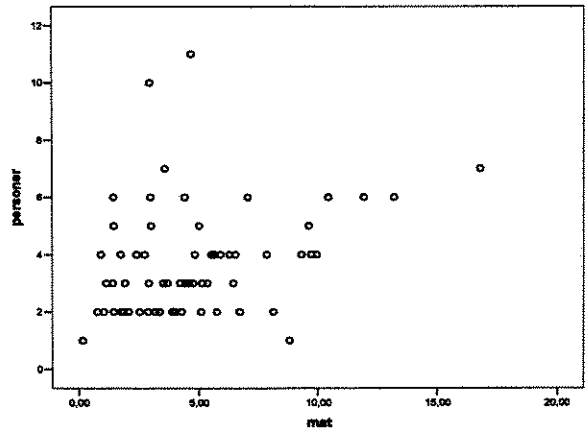
Spredningsplott av antall personer mot hver av de seks opprinnelige uavhengige variablene:

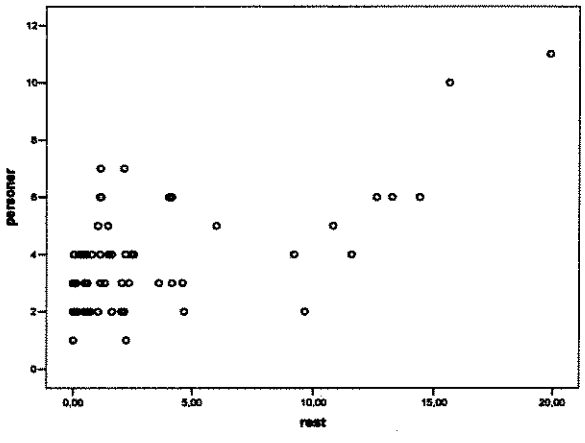












Correlations

		personer	metall	glass	papir	plast	mat	rest
personer	Pearson Correlation	1	,620**	,403**	,469**	,751**	,304*	,640**
	Sig. (2-tailed)		,000	,001	,000	,000	,016	,000
	N	62	62	62	62	62	62	62
metall	Pearson Correlation	,620**	1	,241	,493**	,586**	,313*	,510**
	Sig. (2-tailed)	,000		,059	,000	,000	,013	,000
	N	62	62	62	62	62	62	62
glass	Pearson Correlation	,403**	,241	1	,117	,354**	,006	,198
	Sig. (2-tailed)	,001	,059		,364	,005	,965	,122
	N	62	62	62	62	62	62	62
papir	Pearson Correlation	,469**	,493**	,117	1	,649**	,639**	,161
	Sig. (2-tailed)	,000	,000	,364		,000	,000	,213
	N	62	62	62	62	62	62	62
plast	Pearson Correlation	,751**	,586**	,354**	,649**	1	,458**	,498**
	Sig. (2-tailed)	,000	,000	,005	,000		,000	,000
	N	62	62	62	62	62	62	62
mat	Pearson Correlation	,304*	,313*	,006	,639**	,458**	1	,173
	Sig. (2-tailed)	,016	,013	,965	,000	,000		,180
	N	62	62	62	62	62	62	62
rest	Pearson Correlation	,640**	,510**	,198	,161	,498**	,173	1
	Sig. (2-tailed)	,000	,000	,122	,213	,000	,180	
	N	62	62	62	62	62	62	62

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

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,641	,408		1,571	,122
	metall	,265	,181	,148	1,465	,149
	glass	,092	,051	,146	1,797	,078
	papir	,040	,057	,085	,698	,488
	plast	,763	,226	,416	3,384	,001
	mat	-,026	,058	-,044	-,446	,657
	rest	,139	,041	,323	3,372	,001
2	(Constant)	1,025	,480		2,134	,038
	metall	,039	,204	,022	,189	,851
	glass	,089	,057	,142	1,559	,125
	papir	,068	,074	,145	,917	,363
	plast	,624	,276	,340	2,259	,028
	mat	,022	,077	,036	,280	,781
	rest	-,043	,155	-,099	-,278	,782
	rest*metall	,083	,047	,770	1,751	,086
	rest*glass	-,010	,024	-,151	-,390	,698
	rest*papir	-,008	,018	-,192	-,414	,681
	rest*plast	,026	,075	,212	,347	,730
	rest*mat	-,009	,014	-,132	-,619	,539

a. Dependent Variable: personer

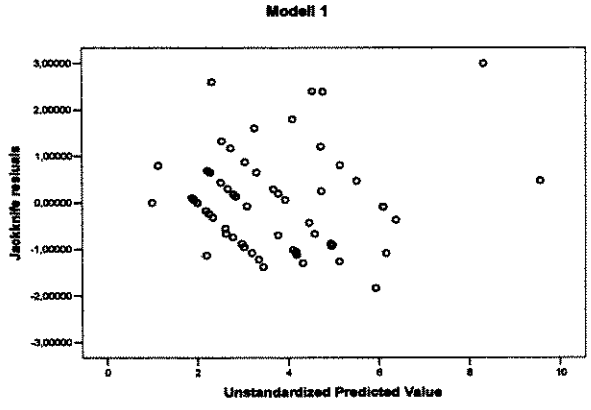
ANOVA[c]

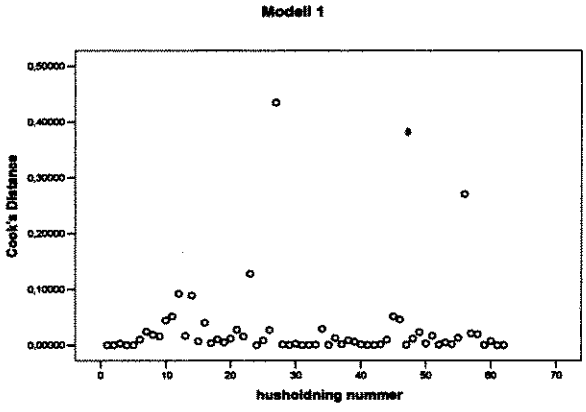
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	162,092	6	27,015	21,021	,000[a]
	Residual	70,682	55	1,285		
	Total	232,774	61			
2	Regression	172,175	11	15,652	12,915	,000[b]
	Residual	60,599	50	1,212		
	Total	232,774	61			

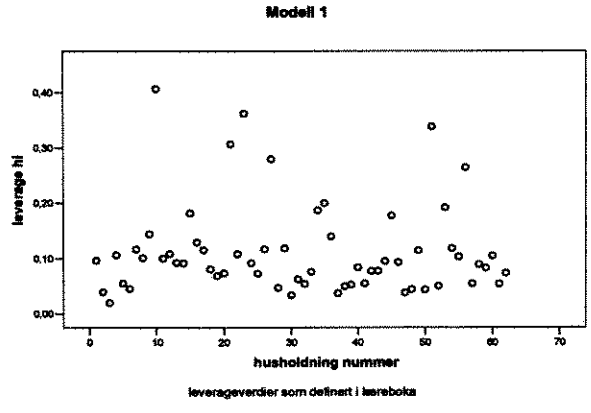
a. Predictors: (Constant), rest, papir, glass, mat, metall, plast

b. Predictors: (Constant), rest, papir, glass, mat, metall, plast, rest*glass, rest*mat, rest*papir, rest*metall, rest*plast

c. Dependent Variable: personer







ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
3	Regression	153,150	2	76,575	56,740	,000[a]
	Residual	79,625	59	1,350		
	Total	232,774	61			
4	Regression	157,615	3	52,538	40,544	,000[b]
	Residual	75,159	58	1,296		
	Total	232,774	61			
5	Regression	161,447	4	40,362	32,254	,000[c]
	Residual	71,327	57	1,251		
	Total	232,774	61			

a. Predictors: (Constant), rest, plast

b. Predictors: (Constant), rest, plast, glass

c. Predictors: (Constant), rest, plast, glass, metall

d. Dependent Variable: personer

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3	(Constant)	1,229	,307		4,004	,000
	plast	1,052	,161	,574	6,538	,000
	rest	,153	,038	,355	4,039	,000
4	(Constant)	1,062	,314		3,382	,001
	plast	,960	,165	,524	5,805	,000
	rest	,151	,037	,350	4,070	,000
	glass	,093	,050	,148	1,856	,068
5	(Constant)	,728	,363		2,006	,050
	plast	,828	,179	,452	4,624	,000
	rest	,131	,038	,302	3,397	,001
	glass	,090	,049	,143	1,821	,074
	metall	,298	,171	,167	1,750	,086

a. Dependent Variable: personer

Datamaterialet:

husholdning	personer	metall	glass	papir	plast	mat	rest
1	2	1,09	0,86	2,41	0,27	1,04	4,66
2	3	1,04	3,46	7,57	1,41	3,68	2,34
3	3	2,57	4,52	9,55	2,19	4,43	3,6
4	6	3,02	4,92	8,82	2,83	2,98	12,65
5	4	1,5	6,31	8,72	2,19	6,3	2,18
6	2	2,1	2,49	6,96	1,81	1,46	2,14
7	1	1,93	0,51	6,83	0,85	8,82	2,22
8	5	3,57	5,81	11,42	3,05	9,62	10,83
9	6	2,32	1,96	16,08	3,42	4,41	4,14
10	4	1,89	17,67	6,38	2,1	2,73	0,25
11	4	3,26	3,21	13,05	2,93	9,31	11,61
12	7	3,99	4,94	11,36	2,44	3,59	1,15
13	3	2,04	3,1	15,09	2,17	5,36	4,15
14	5	0,99	1,39	2,8	1,41	1,47	1,03
15	6	2,96	5,21	6,44	2	7,06	14,43
16	2	1,5	2,03	5,86	0,93	2,52	9,65
17	4	2,43	1,74	11,08	2,97	1,75	2,54
18	4	2,97	3,99	12,43	2,04	5,64	9,2
19	3	1,42	6,26	6,05	0,65	1,93	0
20	3	3,6	3,52	13,61	2,13	6,46	1,32
21	2	4,48	2,01	6,98	0,63	6,72	0,18
22	2	1,36	2,21	14,33	1,53	5,76	1,62
23	4	2,11	0,25	13,31	4,69	9,72	0,4
24	1	0,41	0,09	3,27	0,15	0,16	0
25	4	2,02	6,85	6,67	1,45	5,52	0,03
26	6	3,27	2,33	17,65	2,68	11,92	4,03
27	11	4,95	5,45	12,73	3,53	4,68	19,89
28	3	1	2,04	9,83	1,49	4,76	0,12
29	4	1,55	4,98	16,39	2,31	7,85	1,48
30	3	1,41	3,54	6,33	0,92	2,9	0,04
31	2	1,05	1,06	9,19	0,89	2,87	0,03
32	2	1,31	2,7	9,41	0,8	5,09	0,71
33	2	2,5	1,14	9,45	0,72	3,17	0,01
34	4	2,35	12,24	12,32	2,66	2,4	0,78
35	6	3,69	5,67	20,12	4,37	13,2	1,17
36	2	3,61	2,43	7,72	0,92	2,07	0
37	2	1,49	4,02	6,16	1,4	4	0
38	2	1,36	6,45	7,98	1,45	4,27	2,02
39	2	1,73	1,89	9,64	1,68	1,87	0,58
40	2	0,94	1,78	8,08	1,53	8,13	0,05
41	3	1,33	2,93	10,99	1,44	3,51	0,59
42	3	2,62	1,82	13,11	1,44	4,21	0,49
43	2	1,25	2,89	3,26	1,36	3,34	0,16
44	2	0,26	0,99	1,65	0,38	0,77	0
45	3	4,41	1,93	10	1,74	1,14	4,6
46	6	3,22	3,61	8,96	2,35	1,45	1,12
47	4	1,86	2,53	9,46	2,3	6,54	2,45
48	4	1,76	3,76	5,88	1,14	0,92	0,04
49	3	2,83	1,32	8,26	2,88	5,14	2,03
50	3	2,74	2,64	12,45	2,13	4,59	1,14
51	10	4,63	12,33	10,58	5,28	2,94	15,65
52	3	1,7	1,79	5,87	1,48	1,42	0,59
53	6	3,29	3,99	8,78	3,36	10,44	13,3
54	5	1,22	4,44	11,03	2,83	3	6,02
55	4	3,2	9,25	12,29	2,87	5,91	0,55
56	7	3,09	4,02	20,58	2,96	16,81	2,13
57	5	2,58	1,38	12,56	1,61	5,01	1,46
58	4	1,67	1,59	9,92	1,58	9,96	1,13
59	2	0,85	0,85	3,45	1,15	3,89	1,04
60	4	1,52	8,87	9,09	1,28	4,83	1,61
61	2	1,37	3,64	3,69	0,58	1,78	0
62	2	1,32	3,03	2,61	0,74	3,37	0,46

Oppgave 2 (delspørsmål a)

Datasettet under inneholder vekt (i kg) av 20 poppeltrær som er gjødslet og/eller vannet eller ingen av delene. Alle trær har vokst i samme jordsmonn, rikt og fuktig nær en bekk. I de to nederste linjene finner du gjennomsnitt og standardavvik for hver av behandlingene.

	ingen behandling	gjødslet	vannet	gjødslet og vannet
	0,15	1,34	0,23	2,03
	0,02	0,14	0,04	0,27
	0,16	0,02	0,34	0,92
	0,37	0,08	0,16	1,07
	0,22	0,08	0,05	2,38
$\overline{Y}_{i.}$	0,184	0,332	0,164	1,334
S_i	0,127	0,565	0,126	0,859

a) Test hypotesen: Ingen forskjell på de ulike behandlingene med hensyn til vekt av poppeltrær.
Hva ble konklusjonen av testen?
Gå videre med Tukeys metode for å finne hvilke behandlinger som eventuelt gir ulik vekst.
Hvorfor er Tukeys metode den beste metoden å benytte her?

Resultatet av en enveis variansanalyse:

Tests of Between-Subjects Effects

Dependent Variable: weight

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4,682[a]	3	1,561	5,731	,007
Intercept	5,070	1	5,070	18,618	,001
trearment	4,682	3	1,561	5,731	,007
Error	4,357	16	,272		
Total	14,110	20			
Corrected Total	9,040	19			

a. R Squared = ,518 (Adjusted R Squared = ,428)

Multiple Comparisons

Dependent Variable: weight

Tukey HSD

(I) trearment	(J) trearment	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
gjø_vann	gjødsel	1,0020*	,33005	,036	,0577	1,9463
	ingen	1,1500*	,33005	,015	,2057	2,0943
	vann	1,1700*	,33005	,013	,2257	2,1143
gjødsel	gjø_vann	-1,0020*	,33005	,036	-1,9463	-,0577
	ingen	,1480	,33005	,969	-,7963	1,0923
	vann	,1680	,33005	,956	-,7763	1,1123
ingen	gjø_vann	-1,1500*	,33005	,015	-2,0943	-,2057
	gjødsel	-,1480	,33005	,969	-1,0923	,7963
	vann	,0200	,33005	1,000	-,9243	,9643
vann	gjø_vann	-1,1700*	,33005	,013	-2,1143	-,2257
	gjødsel	-,1680	,33005	,956	-1,1123	,7763
	ingen	-,0200	,33005	1,000	-,9643	,9243

Based on observed means.

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