EKSAMENSOPPGAVE

Fakultet	**KBM**	
Eksamen i:	**STAT 100**	**Statistikk**
Tidspunkt	19. mai 2017	09.00-12.30. 3,5 timer
Kursansvarlig:	Trygve Almøy	

Tillatte hjelpemidler: C3. Alle typer kalkulatorer, alle andre hjelpemidler.
Oppgaveteksten er på 11 sider.

Oppgave I og II teller hver 25 % av denne eksamen, og alle 4 delspørsmål i oppgave 2 teller likt. Flervalgsspørsmål teller 50 % av denne eksamen.

Merk: Side 11: Skjema fylles ut og arket leveres inn.

Oppgave 1

Kvaliteten på frukt er ofte bedømt etter sukkerinnholdet. I et forsøk utført på Bioforsk (nå NIBIO) på Ås i 2011 ble 4 pæresorter sammenlignet, 3 av disse var av Kvede typen. En lagde juice av prøvene og målte sukkerinnholdet i juicen. Siden sukkerinnholdet kan variere innenfor en pæresort, valgte en å ta 6 gjentak for hver sort. Data finner du i *Tabell 1* i Appendix. I *Tabell 2* finner du en utskrift fra R-commander. I *Figur 1* i Appendix finner du et histogram.

Skriv en **KORT** rapport der du beskriver resultatene fra analysen. Rapporten skal gi modell med modellantagelser, informasjon om og tolkning av parametre og estimater, modellvurdering (hvor god modelltilpasningen er og om modellantagelser er oppfylt), hypotesetest(er) og resultattolkninger, inkludert mulige interessante kontraster.

```
AnovaModel.1 <- aov(Sukkerinnhold ~ Sort, data=Pear2011)
            Df Sum Sq Mean Sq      F value   Pr(>F)
Sort         3  5.602  1.8672      6.228     0.00367
Residuals   20  5.997  0.2998

               mean        sd   data:n
KvedeA     12.067        0.320      6
Kvedeadams 11.467        0.520      6
KvedeC     12.550        0.804      6
Pyrodwarf  11.350        0.423      6

                          Estimate Std.Error  t value  Pr(>|t|)    DF
Sort c=( 0.33 0.33 0.33 -1 ) 0.678  0.258      2.625    0.016       20
```

Tabell 2. Oppgave 1, utskrift fra R-commander

Oppgave 2

Noen studenter ville finne ut sammenhengen mellom bruttoinntekt i en familie, og dennes årsforbruk av mat. Begge deler er her målt i tusen kroner. La bruttoinntekten være forklaringsvariabel og årsforbruket være responsen. Anta modellen

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \text{ for } i = 1, 2, \ldots, 11.$$

Undersøkelse av $n = 11$ tilfeldig valgte tobarnsfamilier ga resultater som du finner i Tabell 3 i Appendix: Resultat fra analysen finner du i Tabell 4.

```
           mean    sd
forbruk    60.8    10.7
inntekt    609    228.9

             Estimate Std. Error      t value    Pr(>|t|)
(Intercept) 35.81     4.82           7.424      4e-05
inntekt      0.041   0.0074          5.508      0.000377
s: 5.4 on 9 degrees of freedom   Multiple R-squared: 0.77


predict_CI_PI(RegModel.1, 'inntekt'=609, level=0.95)
 fit   lwr.CI  upr.CI  lwr.PI upr.PI
 *        *       *     48.1   73.6
```

Tabell 4. Resultater fra modellen i Oppgave 2.

a) Sett opp antagelsen om feileddene (ε).
Estimer alle paramtre i modellen.
Gi en ***konkret*** tolkning av de estimerte parametre.
Skriv ned den estimerte linja.

b) Lag et 95 % konfidensintervall for regresjonskoeffisienten (β).
Hvordan kan du ut fra dette intervallet påstå at det er en signifikant sammenheng mellom inntekt og forbruk for 2 barnsfamilier?
Hvis det var slik at $\beta = 0$, hvor sannsynlig er det at du estimerer β til 0,041 eller noe som er enda større?
Hva betyr det at $R^2 = 0{,}77$ i denne undersøkelsen?

c) Hva er et residual i denne sammenhengen?
Finn residualet for den første familien i Tabell 3.
Hva betyr det for denne familien at dette residualet er negativt?

d) Anta at en 2 barnsfamilie du kjenner har en bruttoinntekt på 609 000 kroner.
Hva ville være et rimelig anslag på denne familiens forbruk av mat?
Lag et prediksjonsintervall for denne familiens forbruk og forklar dem hva dette betyr.

Flervalg

F1
Anta at P(B) = 0,7 og P(A|B) = 0,5. Hva er P(A∩B)?
A: 0,5 B: 0,7 C: 0,35 D: 0 E: 1,2 F: 0,71

F2
Anta at 10 % av kyr lider av sjukdommen ketose.
En veterinærstudent trenger minst 2 kyr med denne sjukdommen i forbindelse med en oppgave.
Dersom hun trekker ut 10 kyr tilfeldig, hva er sannsynligheten for at hun får nok dyr.
A: 0,736 B:0,264 C: 0,625 D: 0,375 E: 0,1 F: 0,589

For Oppgavene F3 – F5
Poeng til eksamen i STAT100 følger omtrent en normalfordeling med forventning 65 og standardavvik 15.

F3
Dersom grense for A settes til 90 poeng, hvor stor andel av studentene vil da få A?
A: 0,95 B:0,90 C: 0,05 D: 1,64 E: 0,1 F: 0,12

F4
Dersom en i stedet bestemmer seg for at 12 % av studentene skal ha A, hvor mange poeng må en ha for å få A?
A: 95 B: 90 C: 0,12 D: 82,6 E: 88 F: 92

F5.
Dersom nedre grense for C settes lik 65, hvor stor prosentandel av de som får D eller dårligere stryker når strykgrense er 40 poeng?
A: 9,6 % B: 40 % C: 65 % D: 4,7 % E: 100 % F: 12 %

F6
I en forelesningssal er det 100 studenter hvorav det er 45 gutter. Dersom 10 % av guttene røyker og 15 % av jentene røyker, hva er sannsynligheten for at en vilkårlig person er en som røyker?

A: 0,45 B: 0,55 C: 0,125 D: =. 50 E:0,128 F: 0,143

For oppgavene F7 – F9:
Vi har 40 personer som alle lider av en sykdom. For hver person trekker vi lodd om vedkommende skal få behandling 1 (B1) eller behandling 2 (B2). Deretter registrer vi hvor mange som ble friske etter behandling. Dette ga følgende resultat:

```
B1     18 av 22 blir friske
B2     9 av 18 blir friske
```

Det ble brukt en Chi-kvadrat test på disse data. Da fikk vi blant annet denne utskriften
```
Chi-square components:
```

	B1	B2
Frisk	0.67	0.82
Ikke frisk	1.39	1.69

F7
Hva blir testobservatoren (kalt Q i læreboka og W på forelesning)?
A: 1,03 B: 6,37 C: 4,88 D: 3,84 E: 4,57 F: 1,70

F8
Testen fikk en p-verdi på 0,03. Hvilken konklusjon kan du trekke av dette?
A: H_0 kan forkastes på 1 % signifikansnivå.
B: H_0 kan forkastes på signifikansnivå større eller lik 3 %.
C: H_0 kan forkastes på signifikansnivå mindre eller lik 3 %.
D: H_0 kan ikke forkastes på signifikansnivå 5 %.
E: H_1 kan forkastes på signifikansnivå 1 %.
F: H_1 kan forkastes på signifikansnivå 5 %.

F9
Et 95 % konfidensintervall for sannsynligheten for at en vilkårlig pasient blir frisk ved B1 er gitt ved **(0,66; 0,98).** Hvordan tolker du dette?
A: Vi er rimelig sikre på at B1 ikke har effekt.
B: Dersom et stort antall pasienter blir behandlet med B1, vil minst 67 % bli friske.
C: Dersom et stort antall pasienter blir behandlet med B1, vil med 95 % sikkerhet færre enn 2 % av disse ikke bli friske.
D: Dersom et stort antall pasienter blir behandlet med B1, er vi 95 % sikre på at mellom 66 % og 98 % vil bli friske.
E: Vi kan ikke si noe om behandlingseffekt ut fra et konfidensintervall.
F: Konfidensintervallet sier noe om gjennomsnittlig behandlingstid ikke hvor mange som blir friske.

For Oppgavene F10 – F12?
Mango er en frukt som er en utmerket kilde til C-vitamin, men det er et problem at inneholdet av dette kan variere mellom hvor trærne vokser. I tillegg er det mange forskjellige sorter som også kan gi et variert innhold av C-vitamin.
En undersøkte 4 tilfeldig valgte steder i Tanzania, der en ville se på mulige sortseffekter (rød og grønn) med hensyn på C-vitamininnhold. Alle målinger er i milligram pr. 100 gram. På hvert sted ble et mangotre av hver sort plantet, og etter at trærne var utvokst ble en frukt tilfeldig plukket fra hvert tre og fikk målt C-vitamin. La X_i være C-vitaminnehold i rød mango nummer i og Y_i tilsvarende for grønn mango. Vi antar modellen $X_i \sim N(\mu_x, \sigma_x)$, og $Y_i \sim N(\mu_y, \sigma_y)$.

Sted	1	2	3	4
Rød	30	22	28	21
Grønn	28	21	27	19

Tabell 3: C-vitamin innehold for to mangosorter dyrket på 4 forskjellige steder.

F10
Estimatet for $\mu_x - \mu_y$ er?
A: 1,2 B: 0 C: 1,5 D: 6 E: 1 F: 0,58

F11 Hva er riktig?
A: Data må analyseres som parvise observasjoner fordi sted kan ha betydning.
B: Data må analyseres som parvise observasjoner fordi sort kan ha betydning.
C: Data må analyseres som parvise observasjoner fordi sted ikke har betydning.
D: Data må analyseres som ikke-parvise observasjoner fordi det ikke er naturlige par.
E: Data må analyseres som parvis observasjoner fordi rød og grønn danner naturlige par.
F: Data må analyseres som ikke-parvis observasjoner fordi det er stedseffekter vi leter etter.

F12
En ønsker å teste H_0: $\mu_x = \mu_y$, mot H: $\mu_x > \mu_y$. hva blir Hva blir T-observatoren?
A: 1,69 B: 1,28 C: 2,58 D: 10,52 E: 5,20 F: 6,81

For F13 – F15
I et annet forsøk ville en se på om jordtype hadde betydning for C-vitamininnhold dersom en hadde rød mango. En valgte ut to steder med forskjellig jordtype, på hvert sted målte en C-vitamininnholdet i 4 tilfeldig mango. La X_i være C-vitaminnehold i mango nummer i på jordtype 1 og Y_i tilsvarende for jordtype 2. Vi antar modellen $X_i \sim N(\mu_x, \sigma)$, og $Y_i \sim N(\mu_y, \sigma)$. Resultat:

Jordtype 1	30	31	28	30
Jordtype 2	28	28	30	27

Tabell 4: C-vitamin innehold for rød mangosort dyrket på 2 forskjellige steder med 4 gjentak.

I tillegg kan du bruke at
Jordtype 1 ga et gjennomsnitt på 29,75 og et (utvalgs) standardavvik på 1,26
Jordtype 2 ga et gjennomsnitt på 28,25 og et (utvalgs) standardavvik på 1,26

F13
En ønsker å teste H_0: $\mu_x = \mu_y$, mot H_1: $\mu_x > \mu_y$. Hva blir T-observatoren?
A: 1,24 B: 1,28 C: 1,69 D: 1,26 E: 2,39 F: 6,81

F14
Dersom en ønsker å se om jordtype har betydning hvordan vil du teste dette?
A: H_0: $\mu_x = \mu_y$ mot H_1: $\mu_x > \mu_y$
B: H_0: $\mu_x = \mu_y$ mot H_1: $\mu_x < \mu_y$
C: H_0: $\mu_x = \mu_y$ mot H_1: $\mu_x \neq \mu_y$
D: H_0: $\mu_x - \mu_y \neq 0$ mot H_1: $\mu_x - \mu_y = 0$
E: H_0: $\bar{X} = \bar{Y}$ mot H_1: $\bar{X} \neq \bar{Y}$
F: H_0: $\bar{X} = \mu_x$ mot H_1: $\bar{Y} = \mu_y$

F15:
Testen i F14 ga P-verdien: 0,14. Hva er konsekvensen av dette?

A: Dersom jordtype ikke har betydning, så er det 14 % sjanse for at vi vil observere en forskjell på 1,5 mg eller mer i våre data.

B: Dersom jordtype har betydning, så er det 14 % sjanse for at vi vil observere en forskjell på 1,5 mg eller mer i våre data.

C: Dersom jordtype ikke har betydning, så er det 86 % sjanse for at vi vil observere en forskjell på 1,5 mg eller mer i våre data.

D: Dersom jordtype har betydning, så er det 86 % sjanse for at vi vil observere en forskjell på 1,5 mg eller mer i våre data.

E: Dersom $P(|\overline{X} - \overline{Y}| > 1,5) = 0,14$, så har jordtype betydning.

F: Dersom $P(|\overline{X} - \overline{Y}| > 1,5) = 0,14$, så har jordtype ikke betydning.

For F16 og F17

Bananfluer ble selektert både for motstand (Motstand, gruppe 1) mot og ømfintlighet (Ømfintlighet, gruppe 2) for miljøgifter. I tillegg hadde vi en kontrollgruppe (Kontroll, gruppe 3) som ikke ble utsatt for seleksjon. For alle tre gruppene talte vi opp antall egg lagt av hunnfluer i løpet av 14 dager tidlig i deres livssyklus (Y). Vi ønsker å undersøker effekten av seleksjon med hensyn på fruktbarhet. Antatt modell: $Y_{ij} = \mu_i + \varepsilon_{ij}$, med vanlige antagelser på feilleddet. Dette ga:

```
            Df Sum Sq Mean Sq F value  Pr(>F)
seleksjon    2  55163   27581   7.308 0.00144
Residuals   60 226440    3774

             mean        sd    data:n
Kontroll     253.7       64.7     21
Motstand     204.8       64.9     21
Ømfintlighet 182.8       54.1     21
```

Tabell 5: Resultat fra enveis variansanalyse for spørsmål F17

F16

En student ville se på om seleksjon på motstand mot miljøgift ga signifikant reduksjon i fruktbarhet i forhold til ingen seleksjon (Kontroll). Hvilken kontrast (θ) vil fange opp dette?

A: $\theta = \overline{Y}_3 - \overline{Y}_1$ B: $\theta = (\overline{Y}_3 - \overline{Y}_1)/2$ C: $\theta = \mu_3 - \mu_1$

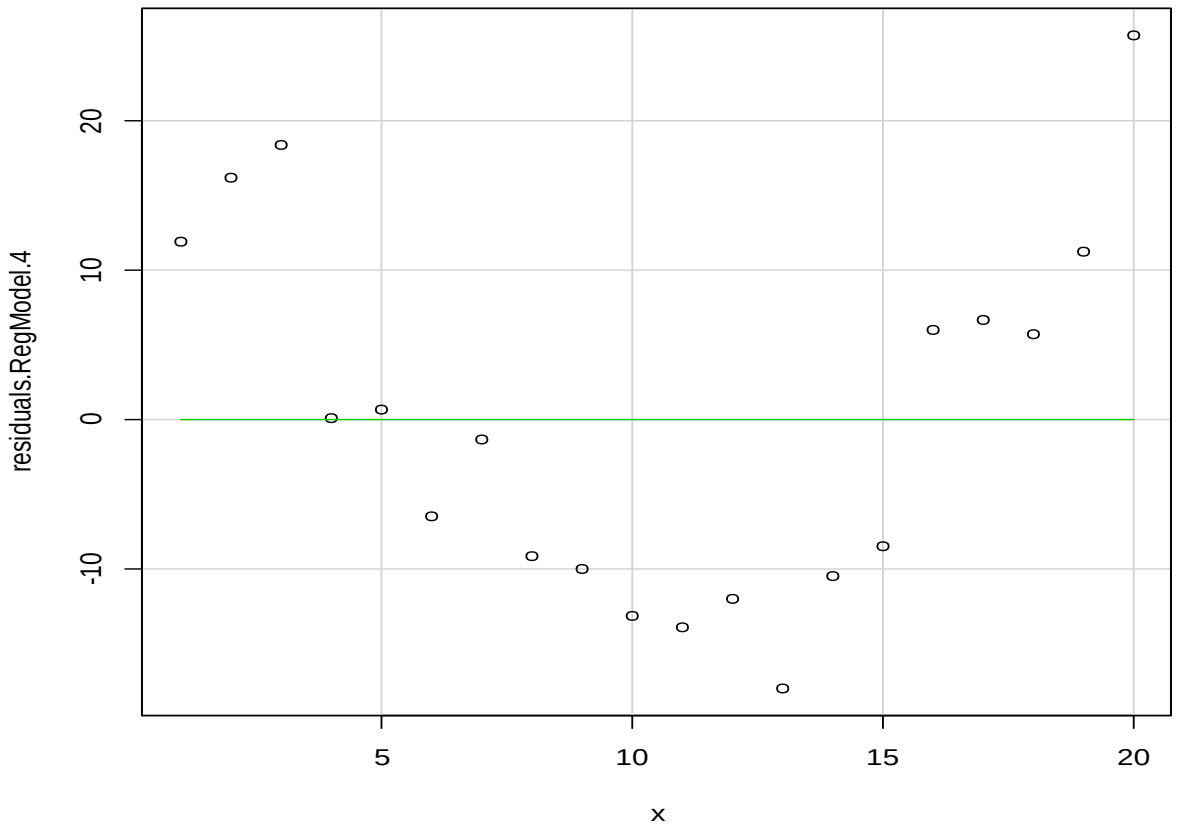
D: $\theta = \frac{\mu_3 - \mu_1}{2} - \mu_2$ E: $\theta = \mu_3 - \mu_1 - \mu_2$ F: $\theta = \frac{\mu_3 + \mu_1}{2} - \mu_2$

F17.

Standardfeilen til kontrasten som brukes for å se om det er forskjell på de to seleksjonsmetodene ($\theta = \mu_2 - \mu_1$) er?

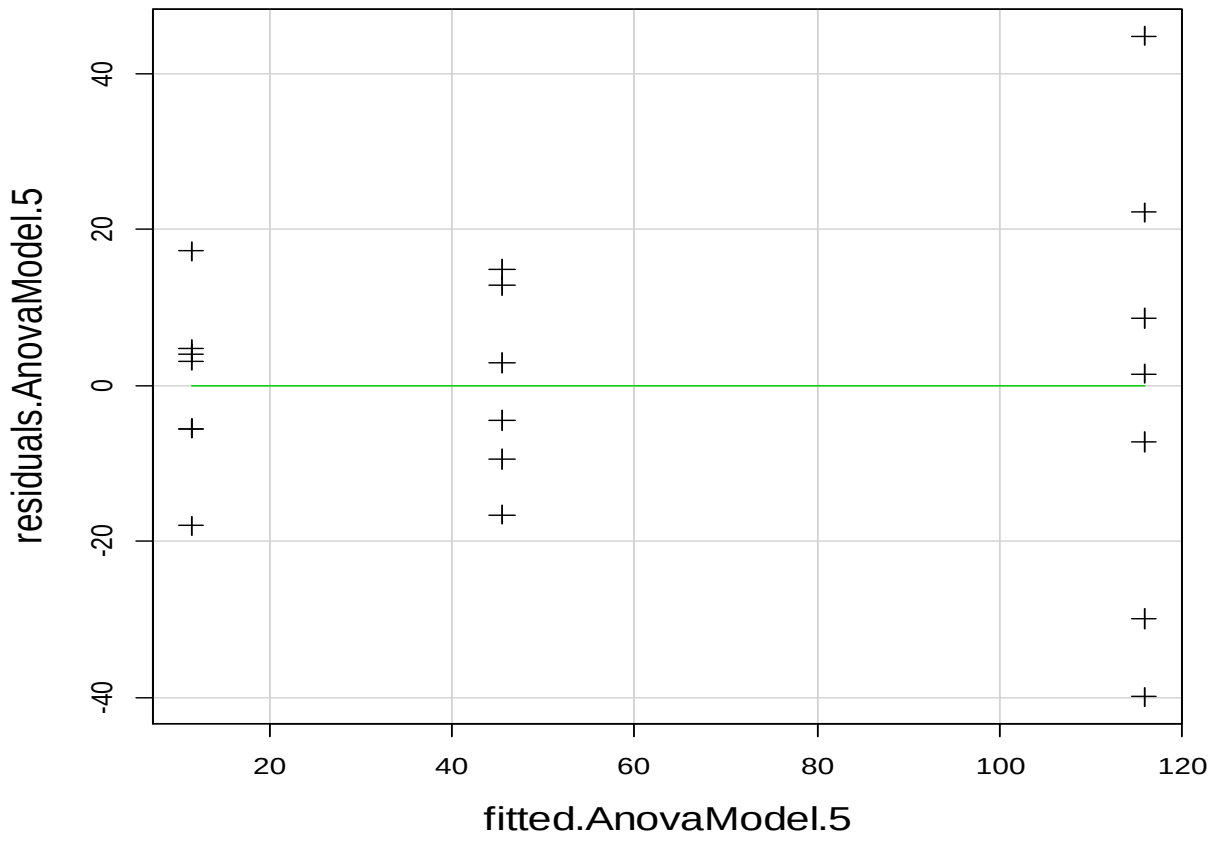
A: 22 B: 25,62 C: 61,4 D: 1,16 E: 21 F: 19

F18
Dersom en antar en lineær regresjonsmodell med de vanlige antagelsene, og får et residualplot som ser ut som figuren nedenfor. Hvilket problem kan da dukke opp i forhold til modellen?



A: Variansen til feilleddet øker med økende nivå på forklaringsvariabelen.
B: Det er avhengighet mellom feilleddene.
C: Feilledd er ikke normalfordelt.
D: Residualene har ikke forventning null.
E: Sammenhengen mellom forklaringsvariabel og forventet respons er ikke-lineær.
F: R^2 blir nær null.

F19.
Det ble kjørt en variansanalysemodell med de vanlige antagelsene. Et plot av residualer mot tilpassede verdier ser du på neste side. Hvilken modellantagelse er problematisk?



A: Variansen er avhengig av gruppe.
B: Feileddene er avhengige
C: Feiledden har ikke forventning nnull.
D: Residualene har ikke forventning null.
E: Sammenhengen mellom forventet respons og gruppe er ikke lineær.
F: R^2 blir nær null.

F20

Anta at X_1 og X_2 kommer fra en normalfordeling med ukjent forventning (μ), men kjent standardavvik (σ). Et 95 % konfidensintervall for μ ser slik ut: $\overline{X} \pm 1{,}96 * sd(\overline{X})$, der sd betyr standardavviket. Dersom $\sigma = 1$ og X_1 og X_2 er korrelert med korrelasjon $\rho = 0{,}8$ hva blir riktig konfidensintervall.

A: $\mu \pm 1{,}86$ B: $\overline{X} \pm 2{,}71$ C: $\overline{X} \pm 1{,}96$ D $\overline{X} \pm 1{,}32$ E: $\overline{X} \pm 3{,}92$ F: $\overline{X} \pm 1{,}86$

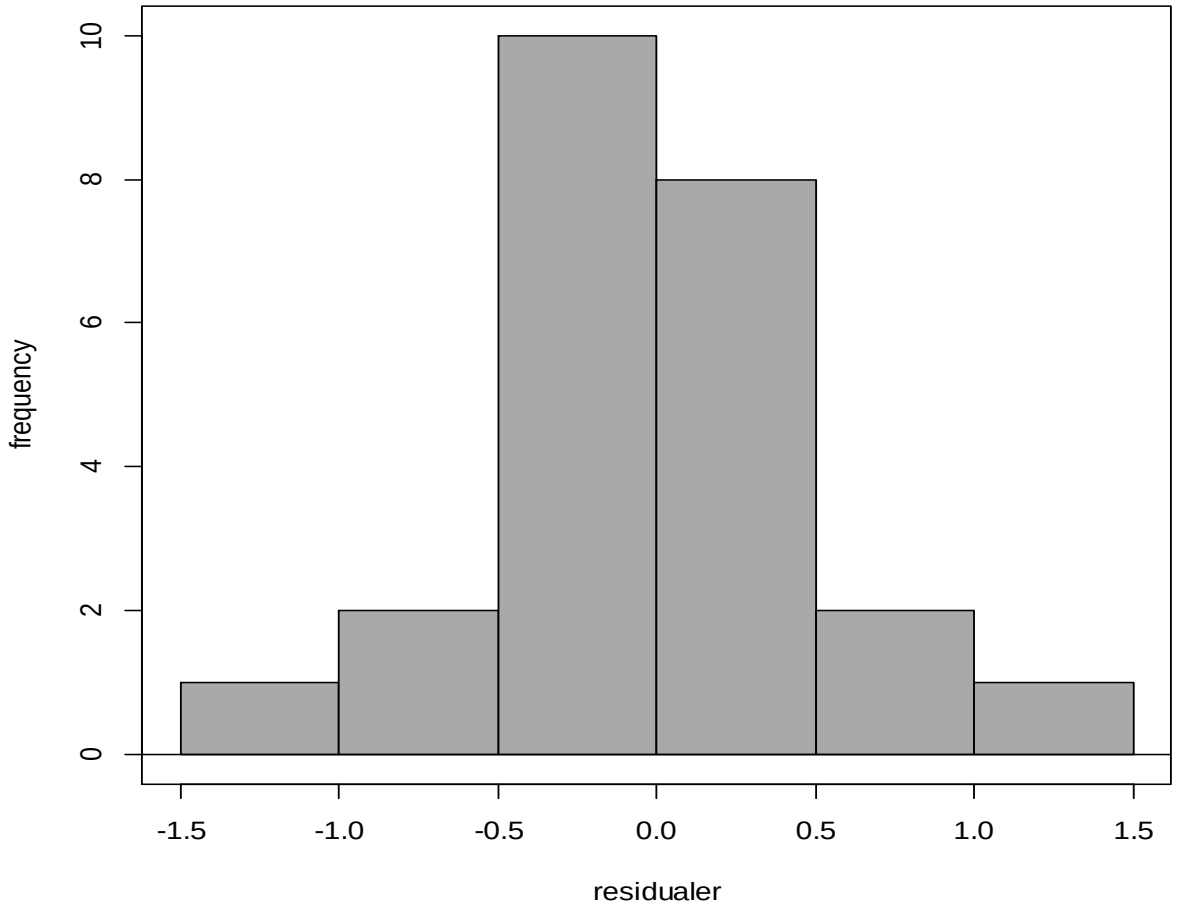
Appendix

Sukker	Sort
12.1	KvedeA
12.1	KvedeA
12.4	KvedeA
11.8	KvedeA
11.6	KvedeA
12.4	KvedeA
12.7	KvedeC
12.7	KvedeC
12.3	KvedeC
13.8	KvedeC
12.5	KvedeC
11.3	KvedeC
11.3	Pyrodwarf
11.2	Pyrodwarf
12.1	Pyrodwarf
11.3	Pyrodwarf
10.8	Pyrodwarf
11.4	Pyrodwarf
12.4	kvedeadams
11.5	kvedeadams
11.4	kvedeadams
11.4	kvedeadams
10.8	kvedeadams
11.3	kvedeadams

Tabell 1: Data for oppgave 1.

Inntekt	Forbruk
350	49
350	52
450	44
450	50
500	58
500	67
600	63
750	68
850	71
900	73
1000	74

Tabell3. Data for oppgave 2



Figur 1: Et histogram over residualer oppgave 1

Riv ut arket og levere dette sammen med besvarelsen.
Bare ett kryss i hver rute.

Oppgave	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						