Noregs teknisk–naturvitskaplege universitet
Institutt for matematiske fag

Fagleg kontakt under eksamen:
Mette Langaas (988 47 649)

NYNORSK

EKSAMEN I TMA4255 ANVENDT STATISTIKK

Måndag 6. juni 2011
Tid: 9:00–13:00

Tal på studiepoeng: 7.5

Tillatte hjelpemiddel: Alle trykte og handskrivne hjelpemiddel. Spesiell kalkulator.

Sensurfrist: 28. juni 2011.

Eksamensresultata blir annonsert frå `http://studweb.ntnu.no/`.

Merk deg følgjande:

- I utskrifta frå MINITAB er komma brukt som desimalskilleteikn.
- Bruk signifikansnivå 5%.
- Alle svar må grunngjevast.

Oppgåve 1 Body mass index

I ein finsk studie så ein på samanhengen mellom kroppsmasseindeks (body mass index, BMI) ved 31-års alder og genetiske variantar av feittmasse- og overvekt-genet (Fat Mass and Obesity, FTO). BMI er definert som vekt delt på kvadratet av høgd (kg/m^2). Eit individ har ein av tre mulige genotypar av FTO-genet, desse genotypene kallar vi 0, 1 og 2. Totalt tok 4435 individ del i studien.

a) Forklar med éi setning målet for ein ein-vegs variansanalyse (ANOVA), og skriv ned den statistiske modellen som ein ein-vegs ANOVA er basert på.

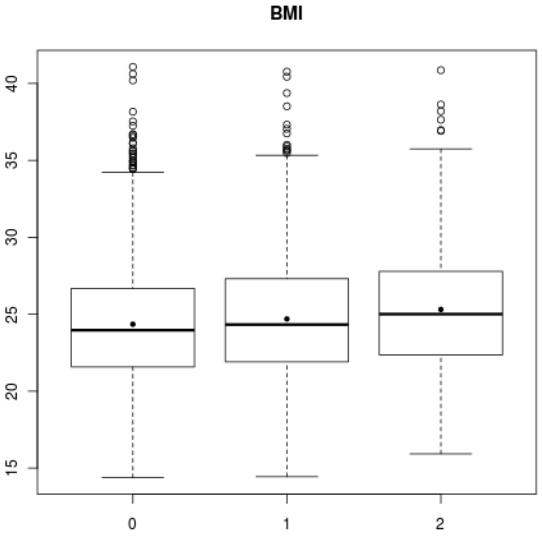
Statistiske analysar av BMI er ofte basert på transformerte data til logaritmisk skala. La ln(BMI) vere den naturlig logaritmen (grunntal e) til ei BMI-måling. Vi ser først på separate analysar av BMI vs. FTO-genotype, og ln(BMI) vs. FTO-genotype.

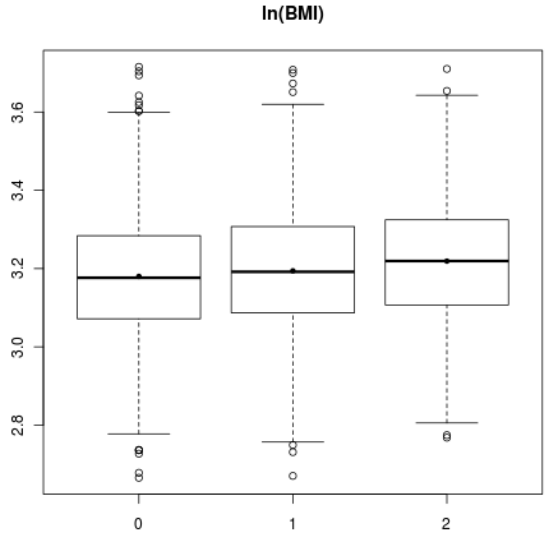
Ein ein-vegs ANOVA-modell vart tilpassa til BMI- og til ln(BMI)-dataa separat. Under finn du utskrift frå MINITAB. På neste side finn du seks grafar. I den øvre raden er boksplott av BMI og av ln(BMI) for dei tre genotypane vist. Dei fylte prikkane viser gjennomsnittsverdien for kvar genotype. I den midtarste rada har vi histogram over studentiserte residual, og i den nedre raden har vi normalkvantil-kvantil plott basert på dei studentiserte residuala. Grafane i den venstre kolonnen er basert på BMI-dataa, mens den høgre kolonnen er basert på ln(BMI)-dataa.

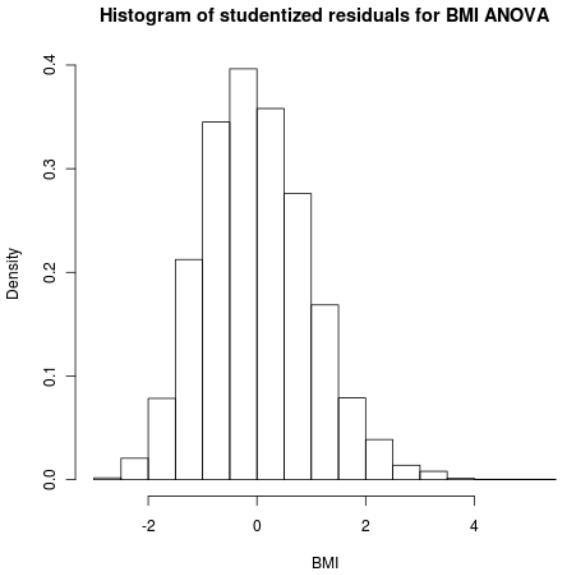
Vil du tilrå at vi bruker BMI-dataa eller ln(BMI)-dataa i vår ein-vegs ANOVA? Grunngje svaret ditt.

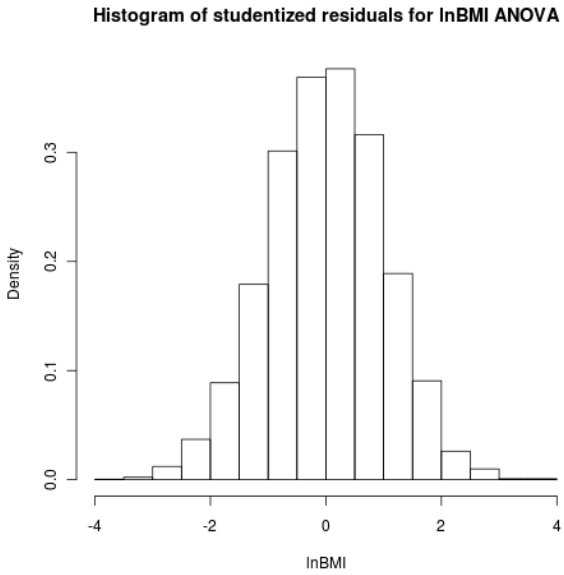
```
One-way ANOVA: lnBMI versus FTO
Source    DF        SS      MS      F          P
FTO        2    0,5727  0,2864  11,32  1,247e-05
Error   4432  112,1064  0,0253
Total   4434  112,6791
S = 0,1590   R-Sq = 0,51%   R-Sq(adj) = 0,46%
```

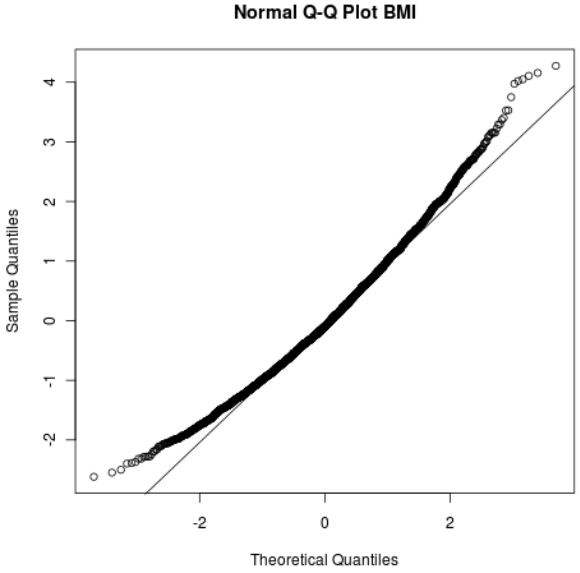
```
One-way ANOVA: BMI versus FTO
Source    DF       SS     MS      F          P
FTO        2    453,0  226,5  14,73  4.209e-07
Error   4432  68158,0   15,4
Total   4434  68611,0
S = 3,922   R-Sq = 0,66%   R-Sq(adj) = 0,62%
```

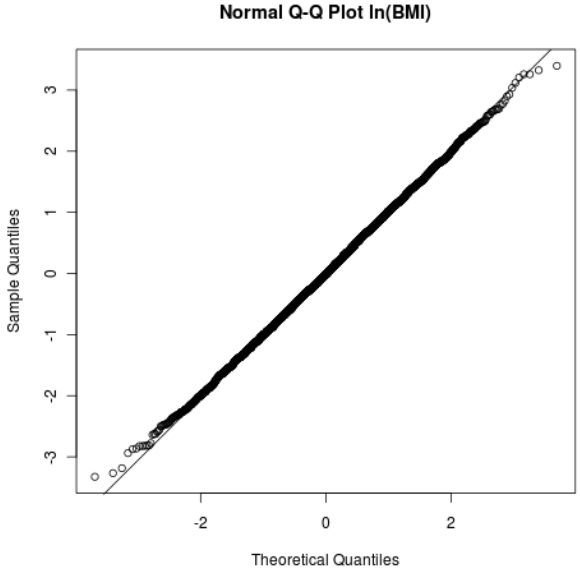

BMI
40
35
30
25
20
15
0
1
2


ln(BMI)
3.6
3.4
3.2
3.0
2.8
0
1
2


Histogram of studentized residuals for BMI ANOVA
Density
0.4
0.3
0.2
0.1
0.0
-2
0
2
4
BMI


Histogram of studentized residuals for lnBMI ANOVA
Density
0.3
0.2
0.1
0.0
-4
-2
0
2
4
lnBMI


Normal Q-Q Plot BMI
Sample Quantiles
4
3
2
1
0
-1
-2
-2
0
2
Theoretical Quantiles


Normal Q-Q Plot ln(BMI)
Sample Quantiles
3
2
1
0
-1
-2
-3
-2
0
2
Theoretical Quantiles

b) Vi vil no analysere ln(BMI)-dataa ved hjelp av ein ein-vegs ANOVA-modell. Bruk utskrifta frå MINITAB («ln(BMI) vs. FTO») på side 2.

Skriv ned nullhypothesen som blir testa i vår ein-vegs ANOVA.
Forklar kvifor bokstaven `F` er brukt for testobservatoren, og poengter kva som er samanhengen mellom `DF`-kolonnen og testobservatoren.
Vil du forkaste nullhypotesen for ln(BMI)-dataa?

Bartletts test og Levenes test for lik varians vart utført med følgjande resultat. Kva konklusjon kan du trekke frå dette?

```
Bartlett's Test (Normal Distribution)
Test statistic = 2,78; p-value = 0,250

Levene's Test (Any Continuous Distribution)
Test statistic = 0,63; p-value = 0,534
```

Vi skal no sjå på forventa ln(BMI) for dei tre ulike genotypane. Det finst n_i individ med FTO-genotype i og x_{ij} er ln(BMI) til individ $j = 1, \ldots, n_i$. Vidare er $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ gjennomsnittleg ln(BMI) og $s_i = \sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}$ eit estimat for standardavviket til ln(BMI) for FTO-genotype i.

FTO-genotypegruppe	n_i	$\bar{x}_i$	s_i
0	1678	3.1809	0.1579
1	2068	3.1920	0.1577
2	689	3.2151	0.1656

c) Kva er samanhengen mellom s_i-ane presenterte i tabellen og $S = 0.1590$ frå ANOVA-utskrifta på side 2?

Lag eit 95 % konfidensintervall for differansen i forventa ln(BMI) mellom FTO-genotypegruppene 0 og 1. Spesifiser kva antakingar du gjer.

I tillegg til å sjå på differansen i forventa ln(BMI) mellom FTO-genotypegruppene 0 og 1 er vi òg interesserte i å sjå på differansen i forventa ln(BMI) for alle par av FTO-genotyper. Det følgjande er utskrift frå MINITAB, der «one-way multiple comparisons» ved hjelp av Tukeys metode er utført med familywise feilrate-signifikansnivå 5 %.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of FTO
Individual confidence level = 98,07%

FTO = 0 subtracted from:
FTO    Lower  Center   Upper  ------+---------+---------+---------+---
1    -0,0011  0,0111  0,0234                (---*----)
2     0,0174  0,0342  0,0510                       (------*-----)
                             ------+---------+---------+---------+---
                                -0,025     0,000     0,025     0,050
FTO = 1 subtracted from:
FTO   Lower  Center   Upper  ------+---------+---------+---------+---
2    0,0067  0,0231  0,0394                   (-----*------)
                             ------+---------+---------+---------+---
                                -0,025     0,000     0,025     0,050
```

Kvifor står det `Individual confidence level = 98,07%` i utskrifta når vi har spesifisert eit 5 % signifikansnivå for familywise feilrate?
Kva konklusjonar kan du trekke frå utskrifta?

d) Gjennomsnittleg ln(BMI) for FTO-genotypegruppe 2 er funne å vere 3.2151 og det empiriske standardavviket er 0.1656. Bruke tilnærma (approksimerte) metodar til å finne eit estimat for forventningsverdi og standardavvik for BMI (dvs. på originalskala, kg/m^2, ikkje på logaritmisk skala).

Oppgåve 2 Normalfordelte karakterer?

Når ein evaluerer ei eksamensløysing gir ein poeng på kvart eksamensspørsmål. Desse poenga blir så summerte til ein total poengsum. Gå ut frå at maksimal poengsum for ei eksamensløysing er 80 poeng. Denne poengsummen er vanlegvis ikke gjort kjend for studenten. I staden blir poengsummen konvertert til ein karakter (A–F) ved hjelp av ein konverteringsregel. Kvar karakter tilsvarer eit poengsum-intervall.

Vi skal sjå på data frå eit kurs førre semesteret. Det vart levert inn 577 eksamensløysingar. Poengsum-intervalla og talet på løysingar som fekk dei ulike karakterane (karakterfrekvens) finn du i følgjande tabell.

Karakter	A	B	C	D	E	F
Poengsum-intervall	(68.5,80]	(58.5,68.5]	(44.5,58.5]	(36.5,44.5]	(31.5,36.5]	[0,31.5]
Karakterfrekvens	38	80	193	131	86	49
Normalsannsyn	0.0465	0.1323	0.3769	0.2147	0.0982	0.1314

Gå ut frå at poensummen til ei løysing er normalfordelt med forventningsverdi 46.3 og standardavvik 13.2. Rekn ut sannsynet for at poengsummen for ei tilfeldig vald løysing ligg i intervallet $(58.5, 68.5]$. Sammanlikn dette resultatet med talet du finn i rada med namn «Normalsannsyn» og kolonnen med namn «B». Dette talet og dei andre tala i «Normalsannsyn»-raden kan du bruke i dei numeriske utrekningane når du svarer på det neste spørsmålet.

Test nullhypotesen H_0: «poengsummen er normalfordelt med forventningsverdi 46.3 og standardavvik 13.2» mot den alternative hypotesen H_1: «poengsummen er ikkje normalfordelt med forventningsverdi 46.3 og standardavvik 13.2» ved å bruke data i tabellen over.

Forklar kort korleis du vil gå fram for å teste nullhypotesen at poengsummen er normalfordelt utan å gå ut frå konkrete verdiar for parametrane i normalfordelinga. Du skal ikkje utføre nokon numeriske utrekningar.

Oppgåve 3 Lukke

Vi skal no sjå på eit datasett som består av data frå 39 studentar, som òg hadde lønna arbeid, ved studiet for Master of Business Administration ved University of Chicago Graduate School of Business. Dataa er samla inn for å teste hypotesen om at kjærleik og arbeid er dei viktigaste faktorane for å forklare kor lukkeleg eit individ er. Som alternative forklaringsvariablar vart det òg samla inn data om inntekt og nivå av seksuell aktivitet. Desse fem variablane er i datasettet kodet på følgjande måte.

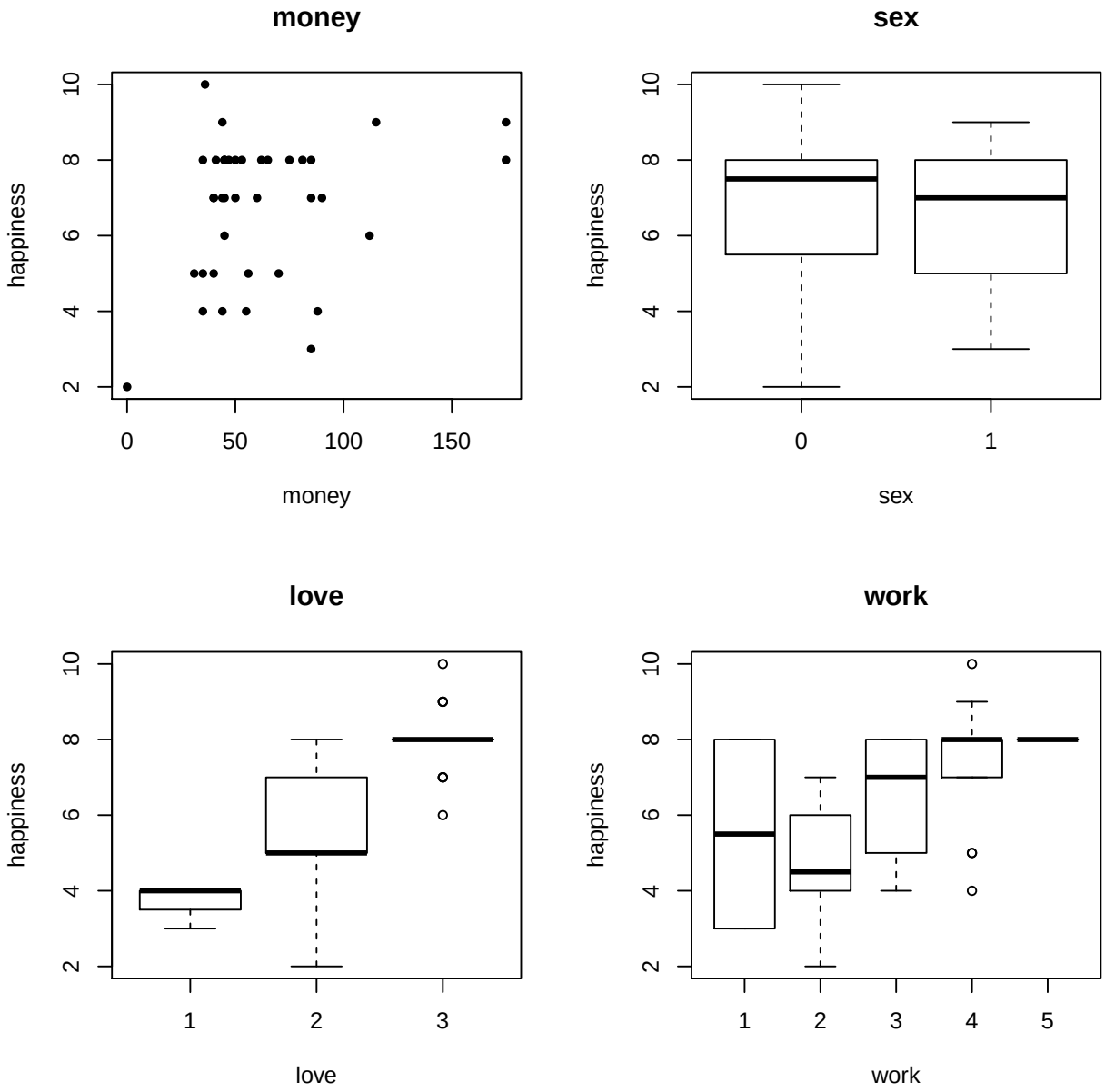
- y, `happiness`. Lukke vart målt på ein skala med 10 nivå, der 1 står for ein suicidal tilstand, 5 står for ei kjensle av å halde det gåande på eit vis og 10 står for ein euforisk tilstand.
- x_1, `money`. Står for kor mange tusen dollar ein tener pr. år i hushaldet til individet.
- x_2, `sex`. Verdiane 0 og 1 vart brukte, der 1 stod for eit tilfredsstillande nivå av seksuell aktivitet.
- x_3, `love`. Kjærleik vart målt på ein skala med 3 nivå, der 1 vart skildra som einsemd og isolasjon, 2 som eksistens av trygge relasjonar, og 3 representerte ei djup kjensle av tilhøyrsle og omsorg i ein familie eller eit samfunn.
- x_4, `work`. Arbeid vart koda på ein skala med 5 nivå. Koden 1 vart brukt for individ som forsøkte å finne ein annan jobb, 3 tydde at jobben var OK, og 5 at jobben var fornøyeleg.

Boksplott og spreiingsplott (scatter plots) finn du på side 8.

Ein multippel lineær regresjonsmodell vart tilpassa til dataa med y som respons og x_1, x_2, x_3 og x_4 som forklaringsvariablar. La $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$ stå for observasjonar frå individ i, der $i = 1, \ldots, 39$. Definer den fulle modellen (modell A):

$$\text{Modell A: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i,$$

der ε_i-ane er u.i.f. $N(0, \sigma^2)$ for $i = 1, \ldots, 39$. Utskrift frå MINITAB og plott av studentiserte residual finst på side 9. Tre av dei numeriske verdiane i MINITAB-utskrifta er bytta ut med spørjeteikn.

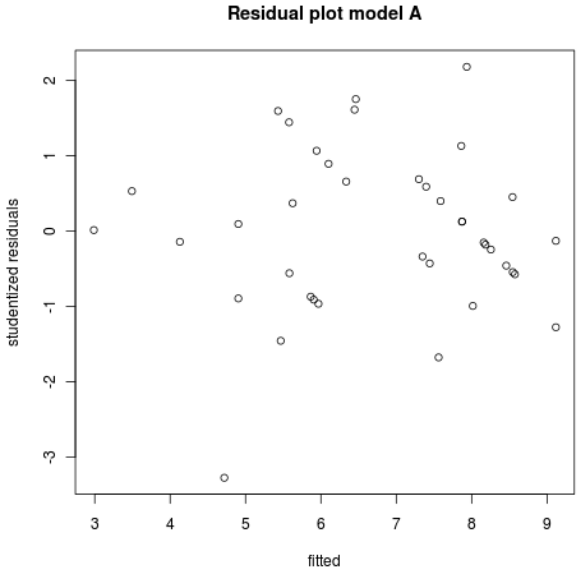

money
happiness
money
sex
happiness
sex
love
happiness
love
work
happiness
work

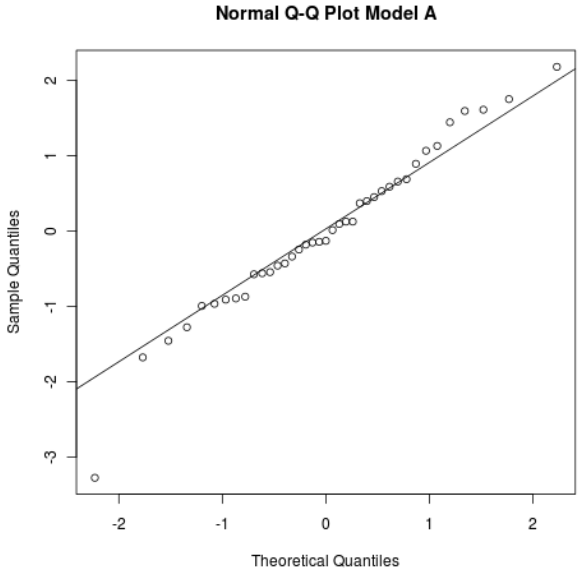
```
Regression Analysis: Happiness versus Money; Sex; Love; Work
The regression equation is
Happiness = - 0,072 + 0,00958 Money - 0,149 Sex + 1,92 Love + 0,476 Work

Predictor       Coef   SE Coef      T         P
Constant    -0,0721    0,8525  -0,08     0,933
Money      0,009578  0,005213   1,84     0,075
Sex         -0,1490    0,4185  -0,36     0,724
Love         1,9193    0,2955   6,50  1,97e-07
Work         0,4761    0,1994      ?         ?

S = 1,05840   R-Sq = 71,0%   R-Sq(adj) = 67,6%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       4   93,349  23,337  20,83      ?
Residual Error  34   38,087   1,120
Total           38  131,436
```





a) Kva verdi har den estimerte regresjonskoeffisienten for x_4, `work`? Korleis vil du forklare denne verdien til «mannen i gata» (som ikkje kjenner til lineær regresjon)? Er effekten av x_4, `work`, signifikant i denne modellen?

Er regresjonen signifikant?
Gje òg ein kort kommentar til residualplotta på side 9.

Vi ønskjer å samanlikne den fulle modellen (modell A), med ein redusert modell (kalla modell B), der x_1 (`money`) og x_2 (`sex`) er fjerna frå modellen:

$$\text{Modell B: } y_i = \beta_0 + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$$

Resultata av å tilpasse modell B er som følgjer.

```
Regression Analysis: Happiness versus Work; Love
The regression equation is
Happiness = 0,206 + 0,511 Work + 1,96 Love

Predictor    Coef  SE Coef     T      P
Constant   0,2057   0,7757  0,27  0,792
Work       0,5106   0,1874  2,72  0,010
Love       1,9592   0,2954  6,63  0,000

S = 1,07951   R-Sq = 68,1%   R-Sq(adj) = 66,3%

Analysis of Variance
Source          DF       SS      MS      F           P
Regression       2   89,484  44,742  38,39  1,182e-09
Residual Error  36   41,952   1,165
Total           38  131,436
```

b) Estimatet $\hat{\beta}_3$ (`love`) er 1.919 for modell A og 1.959 for modell B. Forklar kvifor dette estimatet er ulikt for dei to modellane.

Modell A og modell B kan samanliknast ved å teste følgjande hypotese.

$$H_0 : \beta_1 = \beta_2 = 0 \text{vs. } H_1\text{: } \beta_1 \text{ og } \beta_2 \text{ er ikkje båe lik null}$$

Utfør hypotesetesten og konkluder.

c) Gå ut frå at eit konstantledd, β_0, er med i regresjonsmodellen. Vi vil no sjå på variabelseleksjon. Dersom vi berre ser på hovedeffektar (og ikkje tek med samspel), finst det 15 moglege regresjonsmodellar (4 modellar med éin forklaringsvariabel, 6 modellar med to forklaringsvariablar, 4 modellar med tre forklaringsvariablar og 1 modell med fire forklaringsvariablar).

Resultat frå tilpassinga av desse 15 modellene er presenterte i tabellen under. Kvar rad i tabellen svarer til ein modell. Verdiane i kolonnane med same namn som forklaringsvariablane er estimerte regresjonskoeffisientar. Talet på forklaringsvariablar inkludert i kvar modell finn du i kolonnen med namn N. I kolonnen med namn p finn du p-verdien til regresjonen.

Skriv ned definisjonen av R^2 og R^2_{adj}, og forklar korleis du kan bruke desse til å samanlikne modellane.
Kva tyder det at R^2_{adj} tek ein negativ verdi?
Velj den «beste» av desse 15 modellane. Grunngje valet ditt.
Kva andre modellar, grafar og/eller kriterium ville du ha undersøkt dersom du skulle ha analysert desse dataa?
Vil du, basert på det som er presentert i denne oppgåva, konkludere med at kjærleik og arbeid er dei viktigaste faktorane for lukka til eit individ?

	money	sex	love	work	N	p	R^2	R^2_{adj}
1	0.014				1	0.000747	7.3	4.8
2		−0.130			1	1	0.1	−2.6
3			2.270		1	8.35e-24	61.5	60.5
4				0.990	1	1.36e-13	29.1	27.2
5	0.016	−0.508			2	0.0504	8.8	3.8
6	0.009		2.206		2	8.77e-19	64.5	62.5
7	0.012			0.961	2	3.68e-10	34.6	31.0
8		−0.277	2.279		2	5.55e-18	62.0	59.9
9		0.610		1.079	2	3.48e-09	31.2	27.4
10			1.959	0.511	2	5.75e-20	68.1	66.3
11	0.011	−0.536	2.209		3	9.49e-16	66.2	63.3
12	0.011	0.305		1.009	3	1.84e-07	35.1	29.5
13	0.009		1.902	0.504	3	2.63e-17	70.9	68.4
14		0.108	1.944	0.530	3	2.22e-16	68.1	65.4
15	0.010	−0.149	1.919	0.476	4	9.89e-15	71.0	67.6