Norwegian University of Science and Technology
Department of Mathematical Sciences

Contact during the exam:
Mette Langaas (988 47 649)

ENGLISH

EXAM IN TMA4255 APPLIED STATISTICS

Monday, June 6, 2011
Time: 9:00–13:00

Number of credits: 7.5.

Permitted aids: All printed and handwritten material. Special calculator.

Grading finished: June 28, 2011.

Exam results are annonunced at `http://studweb.ntnu.no/`.

Note that:

- In the output from MINITAB comma is used as decimal separator.
- Significance level 5% should be used.
- All answers need to be justified.

Problem 1 Body mass index

In a Finnish study the association between the body mass index (BMI) at age 31 years and genetic variants of the Fat Mass and Obesity (FTO) gene was studied. BMI is defined as weight divided by squared height (kg/m^2) An individual has one out of three possible genotypes of the FTO gene, we call these 0, 1 and 2. A total of 4435 individuals participated in the study.

a) Explain in one sentence the goal of the one-way analysis of variance (ANOVA), and write down the statistical model that a one-way ANOVA is based on.

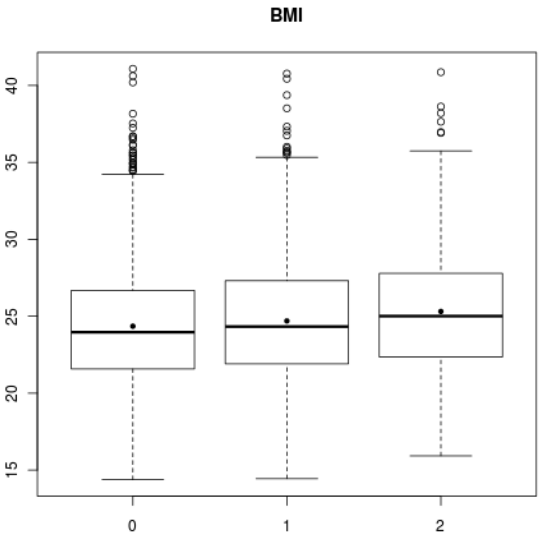
Statistical analyses of BMI are often based on transformed data to the logarithmic scale. Let ln(BMI) to be the natural logarithm (base e) of the BMI measurement. We will first consider separate analysis of BMI vs. FTO genotype, and ln(BMI) vs. FTOgenotype.

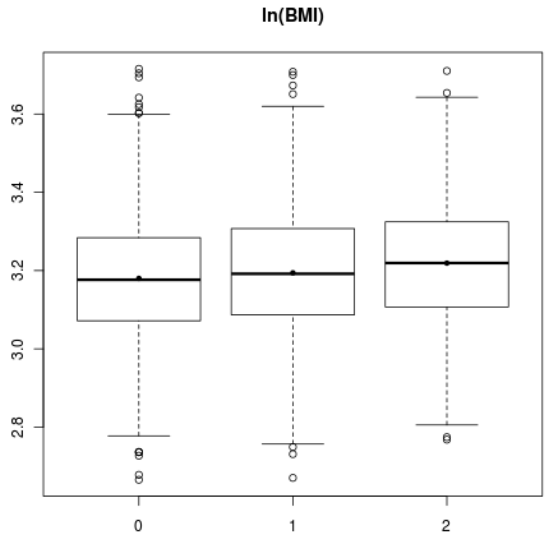
A one-way ANOVA model was fitted to the BMI and the ln(BMI) data separately. See printout from MINITAB below. On the next page you find six graphs. In the top row we have boxplots of BMI and of ln(BMI) for the three genotypes. The solid dots represent the mean value for each genotype. In the middle row we have histograms for studentized residuals, and in the bottom row normal quantile-quantile plots based on studentized residuals. The graphs in the left column are based on the BMI data, and the right column on the ln(BMI) data.

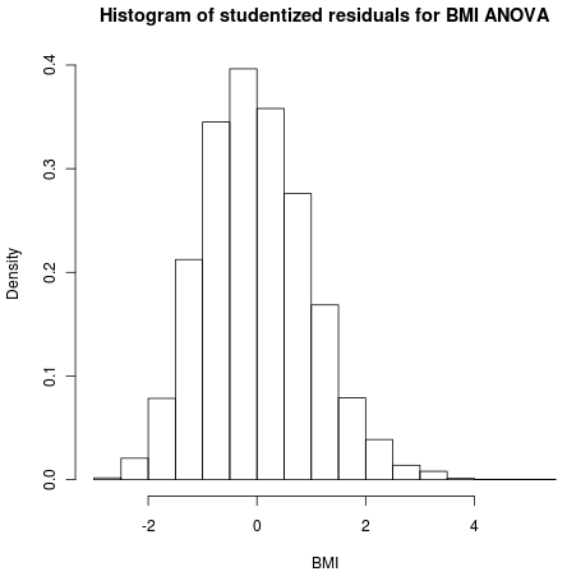
Would you recommend using the BMI or the ln(BMI) data in our one-way ANOVA? Justify your answer.

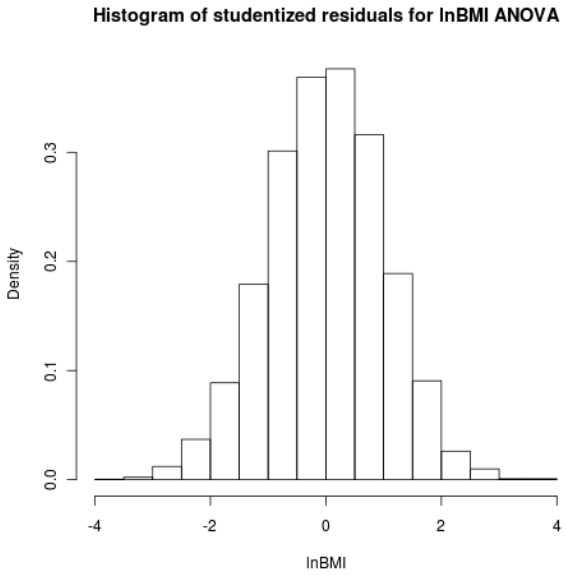
```
One-way ANOVA: lnBMI versus FTO
Source    DF        SS      MS      F          P
FTO        2    0,5727  0,2864  11,32  1,247e-05
Error   4432  112,1064  0,0253
Total   4434  112,6791
S = 0,1590   R-Sq = 0,51%   R-Sq(adj) = 0,46%
```

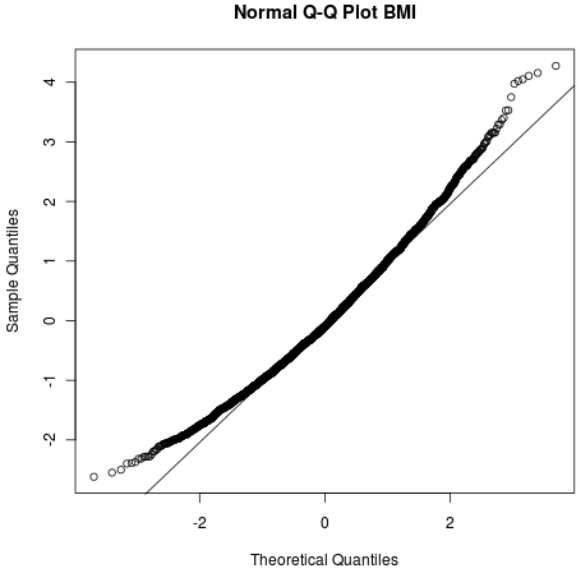
```
One-way ANOVA: BMI versus FTO
Source    DF       SS     MS      F          P
FTO        2    453,0  226,5  14,73  4.209e-07
Error   4432  68158,0   15,4
Total   4434  68611,0
S = 3,922   R-Sq = 0,66%   R-Sq(adj) = 0,62%
```

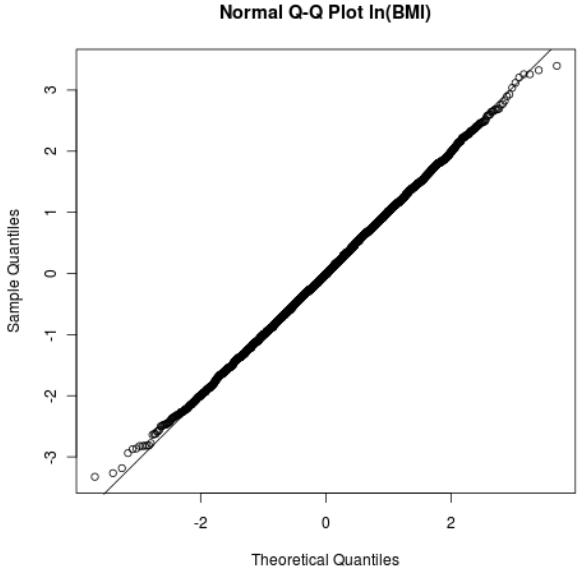

BMI
40
35
30
25
20
15
0
1
2


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) We now analyse the ln(BMI) data using a one-way ANOVA model. Refer to the printout from MINITAB («ln(BMI) vs. FTO») on page 2.

Write down the null hypothesis tested in the one-way ANOVA.
Explain why the letter `F` is used for the test statistic, and point out the connection between the `DF` column and the test statistic.
Is the null hypothesis rejected for our ln(BMI) data?

Bartlett's test and Levene's test of equal variances were performed with the following result. What can you conclude from this?

```
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
```

We will now study the mean ln(BMI) for the three different genotypes. There are n_i individuals with FTO genotype i and x_{ij} is the ln(BMI) for individual $j = 1, \ldots, n_i$. Further, $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ is the average ln(BMI) and $s_i = \sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}$ the estimated standard deviation of ln(BMI) for FTO genotype i.

FTO genotype group	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) What is the relationship between the s_i's reported in the table and the $S = 0.1590$ from the ANOVA printout on page 2?

Construct a 95 % confidence interval for the difference in mean ln(BMI) between the 0 and 1 FTO-genotype group. Specify the assumptions that you make.

In addition to looking at differences in the mean ln(BMI) between the 0 and 1 FTO-genotype groups, we are also interesting in the difference in mean ln(BMI) among all pairs of FTO genotype groups. The following is MINITAB printout from performing «one-way multiple comparisons» using Tukey's method at familywise error rate significance level 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
```

If a 5% significance level for the familywise error rate is specified, why does the printout state that the `Individual confidence level = 98,07%`?
What can you conclude from the printout?

d) In the FTO genotype group 2 the average ln(BMI) was 3.2151 and the empirical standard deviation was 0.1656. Use approximate methods to arrive at an estimate of the mean and standard deviation for the BMI (that is, on the orginal scale, kg/m^2, and not on the logarithmic scale).

Problem 2 Normally distributed grades?

When grading an exam paper a score is given to each exam question. The scores are then summed to give a total score. Assume that the maxium score possible for an exam paper is 80. The total score on the exam paper is usually not made available to the student. Instead the total score is converted into a grade (A–F) using a conversion rule linking a score interval to a grade.

We will look at data from one course last semester. The total number of exam papers was 577. The score interval used and the number of exam papers given the different grades (grade frequency) are found in the following table.

Grades	A	B	C	D	E	F
Score interval	(68.5,80]	(58.5,68.5]	(44.5,58.5]	(36.5,44.5]	(31.5,36.5]	[0,31.5]
Grade frequency	38	80	193	131	86	49
Normal probability	0.0465	0.1323	0.3769	0.2147	0.0982	0.1314

Assume that the the total score is normally distributed with mean 46.3 and standard deviation 13.2. Calculate the probability that the total score on a randomly chosen exam paper lies in the interval $(58.5, 68.5]$. Compare this result with the number found in the row labeled «Normal probability» and the column labeled «B». This and the other numbers in the «Normal probability» row may be used in the numerical calculations when answering the next question.

Test the null hypothesis H_0: «the total score is normally distributed with mean 46.3 and standard deviation 13.2» against the alternative hypothesis H_1: «the total score is not normally distributed with mean 46.3 and standard deviation 13.2» using the data from the table above.

Explain briefly how you would go about to test the null hypothesis that the total scores were normally distributed without assuming specific values for the parameters in the normal distribution. You do not need to perform any numerical calculations.

Problem 3 Happiness

We will look at data collected from 39 individuals in a Master of Business Administration class for employed students at the University of Chicago Graduate School of Business. The reason for collecting the data was to test the hypothesis that love and work are the important factors in determining an individual's happiness. As alternatives, the variables money and sex (sexual activity) were included in the study. The five variables were codes as follows.

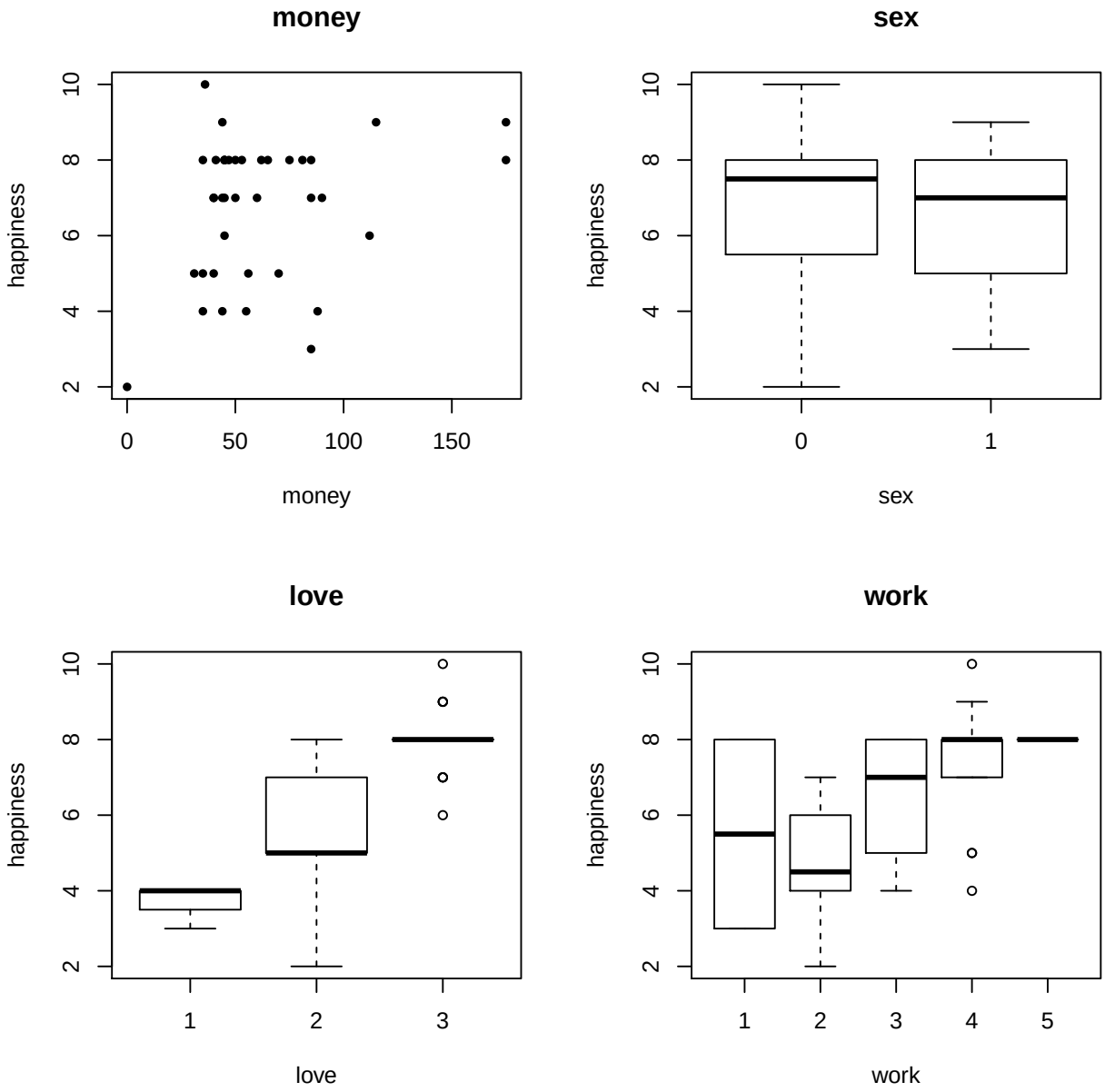
- y, `happiness`. Happiness was measured on a 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of «just muddling along», and 10 representing a euphoric state.
- x_1, `money`. Money was measured by annual family income in thousands of dollars.
- x_2, `sex`. Sex was measured as the values 0 or 1, with 1 indicating a satisfactory level of sexual activity.
- x_3, `love`. Love was measured on a 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community.
- x_4, `work`. Work was measured on a 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is OK, and 5 indicating that the job is enjoyable.

Boxplots and scatter plots are found on page 8.

A multiple linear regression was fitted to the data with y as response and x_1, x_2, x_3 and x_4 as explanatory variables. Let $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$ denote the observations from individual i, where $i = 1, \ldots, 39$. Define the full model (model A):

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

where the ε_i's are i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots, 39$. Printout from MINITAB and plots of studentized residuals are found on page 9. Three of the numerical values in the MINITAB printout have been replaced by question marks.

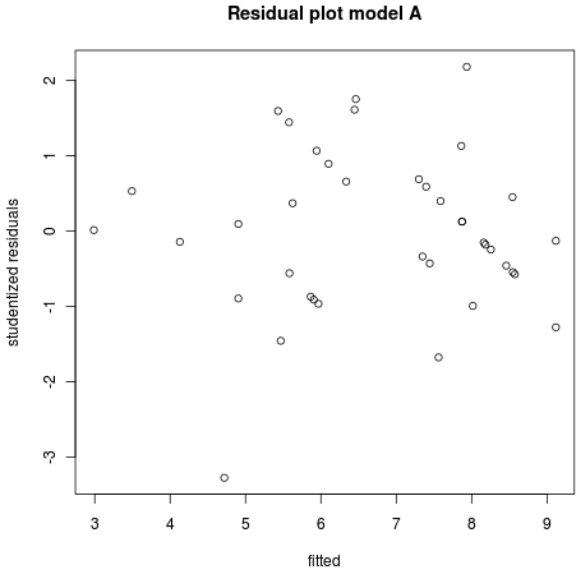

money
happiness
money
0
50
100
150
2
4
6
8
10
sex
happiness
sex
0
1
love
happiness
love
1
2
3
work
happiness
work
1
2
3
4
5

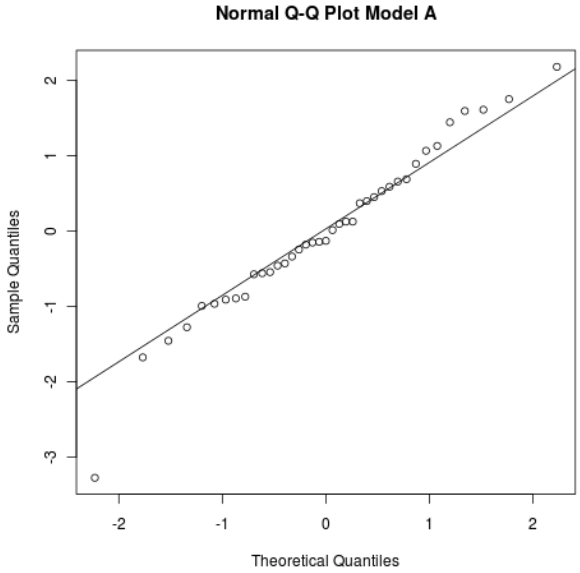
```
Regression Analysis: Happiness versus Money; Sex; Love; Work
The regression equation is
Happiness = - 0,072 + 0,00958 Money - 0,149 Sex + 1,92 Love + 0,476 Work

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) What is the estimated regression coefficient for x_4, `work`? How would you explain this number to the common man (that does not know linear resgression)? Is the effect of x_4, `work`, significant in this model?

Is the regression found to be significant?
Comment briefly on the residual plots on page 9.

We now want to compare the full regression model (model A), with a reduced model (called model B) where x_1 (`money`) and x_2 (`sex`) are excluded from the model:

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

The results from fitting model B are as follows.

```
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) The estimate $\hat{\beta}_3$ (`love`) is 1.919 for model A and 1.959 for model B. Explain why these two estimates differ.

Model A and model B can be compared by testing the following hypothesis.

$$H_0\colon \beta_1 = \beta_2 = 0 \text{ vs. } H_1\colon \beta_1 \text{ and } \beta_2 \text{ are not both zero}$$

Perform the hypothesis test and conclude.

c) Assume that an intercept, β_0, is present in the regression model. We will now look at variable selection. If we only consider modelling main effects (no interactions) there are 15 possible regression models to consider (4 possible models with one explanatory variable, 6 possible models with two explanatory variables, 4 possible models with three explanatory varibles and 1 model with all four explanatory variables).

Results from fitting these 15 models are presented in the table below. Each row in the table corresponds to one model. The values in the columns with the same name as the explanatory variables are the estimated regression coefficients. The number of explanatory variables included in each model is found in the column labeled N. The column labeled p gives the p-value for the regression.

Write down the definition for R^2 and R^2_{adj}, and explain how you can use these to compare the different models.
What does a negative value for R^2_{adj} mean?
Choose the «best» out of these 15 models. Justify your choice.
Which other models, graphs, and/or criteria would you have investigated if you were to analyse this data set?
Would you, based on the material presented here, conclude that love and work are the important factors in determining an individual's happiness?

	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