Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Contact during exam:
Bo Lindqvist
Tel 975 89 418

ENGLISH

EXAM IN TMA4255
APPLIED STATISTICS

Wednesday May 26 2010

Time: 09:00–13:00

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

Grading: 16 June 2010

The exam consists of 8 points which are given equal weight in the grading.

Problem 1

A biologist will examine how the growth of mussels is affected by light when certain other factors are held constant. The growth (y) for 10 mussels are measured under different degrees of lighting (x). The observations (y_i, x_i) for $i = 1, 2, \ldots, 10$ are given in the table below:

i	1	2	3	4	5	6	7	8	9	10
x_i	1	1	2	2	3	3	4	4	5	5
y_i	16	18	17	20	25	21	23	20	17	19

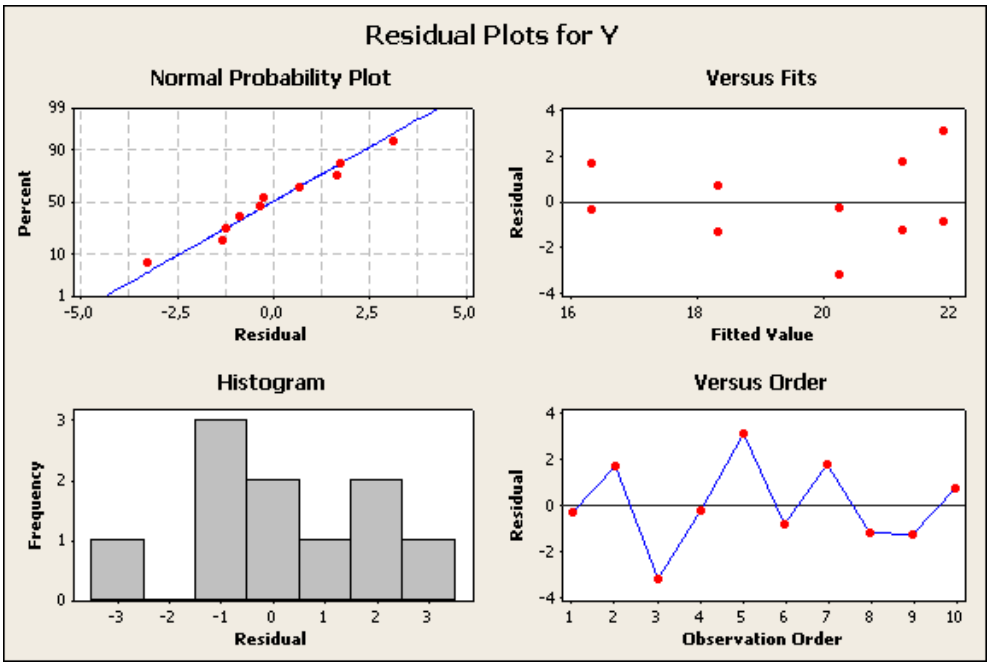
The MINITAB output and plots given on the next page show the results of fitting a multiple linear regression model where the expected growth y is a second order polynomial in lighting x. More precisely, the assumed model is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{1}$$

for $i = 1, \ldots, 10$, where $\epsilon_1, \ldots, \epsilon_{10}$ are independent and $N(0, \sigma^2)$.

In the MINITAB output is the covariate x^2 denoted as $x * x$.

```
Regression Analysis: Y versus x; x*x

The regression equation is
Y = 10,1 + 7,36 x - 1,14 x*x

Predictor      Coef  SE Coef      T      P
Constant    10,100    3,183   3,17  0,016
x            7,357    2,425   3,03  0,019
x*x        -1,1429   0,3966  -2,88  0,024

S = 2,09859   R-Sq = 57,4%   R-Sq(adj) = 45,3%


Analysis of Variance

Source           DF      SS      MS     F      P
Regression        2  41,571  20,786  4,72  0,050
Residual Error    7  30,829   4,404
Total             9  72,400

Source  DF  Seq SS
x        1   5,000
x*x      1  36,571
```



a) Use the output and plots from MINITAB to discuss briefly whether the model (1) and the corresponding assumptions provide a good description of the data.

What is the meaning of the F-value 4.72 in the table under "Analysis of variance" in the output? Write down the null hypothesis that is tested with this test statistic. How can you conclude from the displayed p-value?

b) One is interested in testing $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 < 0$. Write down a test statistic for this and find the critical value when the significance level is 5%. What is the conclusion of the test? What is the p-value of this test?

Similar studies have concluded that the coefficient in front of the quadratic term in the model (1) is -1.0. Is there reason to think otherwise in this case? Formulate this question as a hypothesis testing problem and perform the test with a significance level of 5%.

c) Write down the estimate for σ that can be read out from the MINITAB output.

One would also like a 95% confidence interval for σ. You are asked to find this. *(Hint: Use that SSE/σ^2 is chi-square distributed).*

How can the confidence interval be used to test the null hypothesis $H_0 : \sigma = 1$? What is in that case the alternative hypothesis, and what is the significance level?

d) Show that the estimated regression equation has a maximum at $x_0 = 3.23$.

One wants a prediction interval for the response y_0 for this x-value. First, calculate the point prediction $\hat{y}_0$. It is given that the estimated standard deviation of $\hat{y}_0$ is 1.024. Use this to calculate a 95% prediction interval for y_0.

Problem 2

A group of American biologists are studying how zooplankton live in two lakes, named Rose and Dennison. They set up twelve tanks in their laboratory with water from the two lakes, six tanks for each lake. They add one of three nutrient supplements to each tank and after 30 days they count the zooplankton in a unit volume of water from each tank. The data are given as follows, where the nutrient supplements (see the variable Supplement) are named as 1,2,3, and the variable for lake is called simply 'Lake'.

Row	Zooplankton	Supplement	Lake
1	34	1	Rose
2	43	1	Rose
3	57	1	Dennison
4	40	1	Dennison
5	85	2	Rose
6	68	2	Rose

7	67	2	Dennison
8	53	2	Dennison
9	41	3	Rose
10	24	3	Rose
11	42	3	Dennison
12	52	3	Dennison

Below is the output of a two-way ANOVA using MINITAB.

```
Two-way ANOVA: Zooplankton versus Supplement; Lake

Source       DF       SS       MS     F      P
Supplement    2  1918,50  959,250  9,25  0,015
Lake          1    21,33   21,333  0,21  0,666
Interaction   2   561,17  280,583  2,71  0,145
Error         6   622,00  103,667
Total        11  3123,00

S = 10,18   R-Sq = 80,08%   R-Sq(adj) = 63,49%
```

a) Write down the model and the model assumptions that are used in the MINITAB analysis.

How do you interepret the results regarding main effects and interaction of the two factors Supplement and Lake?

Based on the above results, the biologists decided to ignore the factor Lake and instead consider the problem as a one-way ANOVA problem with the single factor Supplement.

b) MINITAB's ANOVA-table for the one-way analysis with Supplement as the single factor is of the form:

```
Source      DF       SS   MS     F       P
Supplement   ?  1918,50    ?     ?   0,014
Error        ?        ?    ?
Total        ?  3123,00
```

Explain why the Sum of Squares (SS) corresponding to the factor Supplement, as well as the Total Sum of Squares, are unchanged from the two-way case.

Then fill in the correct numbers in each entry in the table above that are marked with a question mark (?)

What is being tested by the F-statistic and what is the conclusion in the present case?

What is the new estimate of σ?

Problem 3

In order to investigate the connection between high blood pressure and smoking, one has collected the following information from 180 randomly selected persons:

	Non-smoker	Moderate smoker	Heavy smoker
High blood pressure	20	36	32
Normal blood pressure	48	26	18

a) Test the null hypothesis that there is independence between the occurrence of high/normal blood pressure and smoking habit. What is the conclusion? Use significance level 1%.

Problem 4

A company producing large steel plates has initiated a quality control effort. The goal is to set up a control chart for the number of surface defects per plate. The data are as follows:

Plate number	1	2	3	4	5	6	7	8	9	10	11	12
Number of defects	4	2	1	3	0	4	6	3	2	2	1	2

a) Which type of control chart would you use here? Give your reasons for the answer. Calculate the control limits by using the given data.

Does the process appear to be in control?