NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4255 Applied statistics**

Academic contact during examination: Anna Marie Holand
Phone: 951 38 038

Examination date: 16 May 2015
Examination time (from–to): 09:00-13:00
Permitted examination support material: One yellow A4-sheet and special calculator.

Other information:

- In outputs from MINITAB comma is used as decimal separator.
- Significance level 5% should be used unless a different level is specified.
- All answers need to be justified.

Language: English
Number of pages: 9
Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 Manufacturer of fertilizers

A manufacturer of fertilizers wanted to study the difference between the effect of an old fertilizer (denoted as X_1) and a newly developed fertilizer (denoted as X_2) on the growth of plants. An experiment was conducted where plants were grown under identical conditions and one randomly assigned fertilizer X_1 to $n_1 = 6$ plants and fertilizer X_2 to $n_2 = 7$ plants. After 3 weeks the height of the plants were measured in cm. Data from the experiment is presented below.

i	1	2	3	4	5	6	7
x_{1i}	54.0	56.1	52.1	56.4	54.0	52.9	
x_{2i}	51.0	53.3	55.6	51.0	55.5	53.0	52.1

Descriptive measures for this dataset are $\bar{x}_1 = \frac{1}{6}\sum_{i=1}^{6} x_{1i} = 54.25$,
$s_{x1} = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_{1i} - \bar{x}_1)^2} = 1.71$, $\bar{x}_2 = \frac{1}{7}\sum_{i=1}^{7} x_{2i} = 53.07$,
$s_{x2} = \sqrt{\frac{1}{6}\sum_{i=1}^{7}(x_{2i} - \bar{x}_2)^2} = 1.91$.

a) We assume that X_{1i} and X_{2i} are normally distributed, $X_{1i} \sim N(\mu_1, \sigma^2)$, $i = 1, ..., 6$ and $X_{2i} \sim N(\mu_2, \sigma^2)$, $i = 1, .., 7$.

Based on this experiment, can the manufacturer conclude that the mean height of the plants given the two different types (X_1 and X_2) of fertilizers are different? Write down the null hypothesis and the alternative hypothesis. Choose a test statistics and perform a hypothesis test. Use significance level $\alpha = 0.05$. Specify the assumptions you make.

b) What is the difference between a non-parametric hypothesis test, and the test used in a)?

Perform a Wilcoxon rank-sum test based on the data given above. You can assume a normal approximation to the test statistics (U_1 or U_2) with mean

$$\mu = \frac{n_1 n_2}{2}$$

and variance

$$\sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Comment on your findings.

Problem 2 Cement hydration

Concrete is produced by cement mixed with sand, gravel and water. A process called cement hydration (reaction with water) plays a critical role in the micro structure of the development of the concrete. The hydration process produces a series of chemical reactions which generates heat. The heat generated depends on the cement composition, and the heat is a parameter which affects material properties and performance of the concrete.

To study the heat generated in the cement hydration process, $n = 13$ batches of concrete were investigated, and for each of these batches the heat generated and 4 possible explanatory variables were measured. The following description is given.

- y: Heat generated in calories during hardening of cement on a per gram basis
- x_1, % of tricalcium aluminate
- x_2: % of tricalcium silicate
- x_3: % of tetracalcium alumino ferrite
- x_4: % of dicalcium silicate

A pairwise scatter plot and a correlation matrix of the variables are found in Figure 1 and Figure 2, respectively.

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} and x_{4i}) denote the measurements from the ith batch, where $i = 1, ..., 13$. Define the full model (model A):

$$\text{Model A} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i, \tag{1}$$

where ϵ_i are i.i.d. $N(0, \sigma^2)$ for $i = 1, ..., n$. The MINITAB output from a statistical analysis of model A is found in Figure 3. Plots of standardized residuals are found in Figure 4.

a) Write down the fitted regression model.

A p-value is given in the row labeled x3 in the results in Figure 3. Explain what this p-value means.

Based on the plots and the result from the fit, would you say that model A is a good model for the data? You need to point out all the features of the fit and plots that you are using to come to your conclusion.

We now want to compare the full regression model (model A), with a reduced model (called model B) with only x_1 and x_2.

$$\text{Model B} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{2}$$

The results from fitting model B are found in Figure 5.

b) Comment on the most important differences between model A and model B. Model A and model B can be compared by testing the following hypotheses.

$$H_0 : \beta_3 = \beta_4 = 0 \text{ vs. } H_1 : \beta_3 \text{ and } \beta_4 \text{ are not both zero}$$

Perform the hypothesis test and conclude.

We are now interested in comparing different regression models where combinations of the covariates x_1, x_2, x_3 and x_4 are present. Assume that an intercept, β_0, is present in the regression model.

The MINITAB output from fitting different regression models to the data are presented in Figure 6. Each row in Figure 6 corresponds to one model. The number of explanatory variables included in each model (in addition to the intercept, β_0) is found in the column labeled $Vars$. The two best models of each number of explanatory variables is reported. The X's indicate which variables that are found in the model.

c) Explain how R^2, R^2_{adj}, Mallows C_p and S are defined and how you can use them to compare the different regression models. Which of the 7 regression model do you rate to be the "best" for this dataset?

Problem 3 Pain and hair colour

In a study conducted at the University of Melbourne the aim was to compare the difference between the pain thresholds of blonds and brunettes. In this study, a total of $n = 19$ men and women of various ages were divided into four categories according to hair colour: light blond, dark blond, light brunette, and dark brunette. Each person in the experiment was given a pain threshold score based on his or her performance in a pain sensitivity test (the higher the score, the higher the person's pain tolerance). A box plot of the data is presented in Figure 7.

a) A one-way analysis of variance model (ANOVA) was fitted to the data, and the MINITAB output from the ANOVA and summary statistics are given in Figure 8.

Six of the entries in Figure 8 are each replaced with a question mark (?). Explain what these entries mean and calculate numerical values for these six missing entries.

What are the assumptions behind this analysis?

Are the four categories of hair colour different with respect to pain tolerance? Perform a hypothesis test to answer this question. Write down the null and alternative hypothesis. Base the test on what you have found in Figure 8. Use significance level $\alpha = 0.05$.

b) Earlier studies indicate that persons with the hair colour *light blond* are different from persons with the hair colour *dark brunette* with respect to pain tolerance. We want to test this. Write down the null hypothesis and alternative hypothesis. Choose a test statistics and perform a hypothesis test. Use a significance level $\alpha = 0.05$. Use the summary statistics in Figure 8 when performing the hypothesis test. What is your conclusion?

Can you perform the hypothesis test above using a 95% confidence interval? Calculate the 95% confidence interval and explain.

Problem 4 Satisfaction of custumers

A regional chain-store is performing a survey to determine how satisfied the customers are with their stores. They divided their customers into 4 geographical regions (South-West, South-East, Middle and North). The total of people that were asked to answer the survey in each of the 4 regions are determined before the

survey as a percentage of the population size in the region. The satisfaction of the customers were classified into three groups and the following cross-tabulation was observed.

Region/Degree of satisfaction:	Satisfied	Don't know	Discontent	Total
South-West	235	74	89	398
South-East	654	203	309	1166
Middle	366	79	244	689
North	179	54	54	287
Total	1434	410	696	2540

a) Based on the data, can we conclude that the 4 different regions differ with respect to their degree of satisfaction with the chain-store? Write down the null hypothesis and the alternative hypothesis and perform a hypothesis test on the basis of the table above. Use a 5% level of significance.

To ease the computational burden you may use that the χ^2-test statistic equals 44.78 and only show the calculation of one of the 12 terms in the sum. What is the conclusion based on this test?

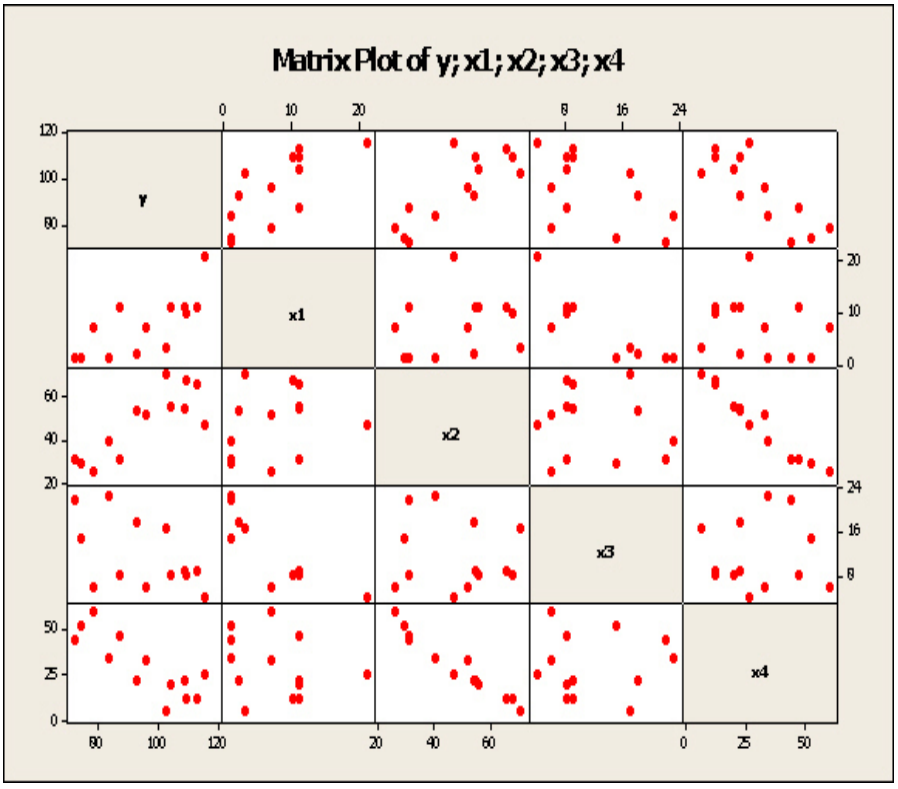

Figure 1: Pairwise scatterplot of the variables in the cement hydration data set.

```
Correlations: y; x1; x2; x3; x4

          y       x1       x2       x3
x1    0,731
x2    0,816    0,229
x3   -0,535   -0,824   -0,139
x4   -0,821   -0,245   -0,973    0,030
```

Figure 2: Pearson correlation between variables x_1, x_2, x_3 and x_4 in the cement hydration data set.

```
Predictor      Coef  SE Coef      T      P
Constant     62,41    70,07   0,89  0,399
x1           1,5511   0,7448  2,08  0,071
x2           0,5102   0,7238  0,70  0,501
x3           0,1019   0,7547  0,14  0,896
x4          -0,1441   0,7091 -0,20  0,844

S = 2,44601   R-Sq = 98,2%   R-Sq(adj) = 97,4%

Analysis of Variance
Source           DF       SS      MS       F      P
Regression        4  2667,90  666,97  111,48  0,000
Residual Error    8    47,86    5,98
Total            12  2715,76
```

Figure 3: Printout from statistical analysis of the cement hydration data set for model A.

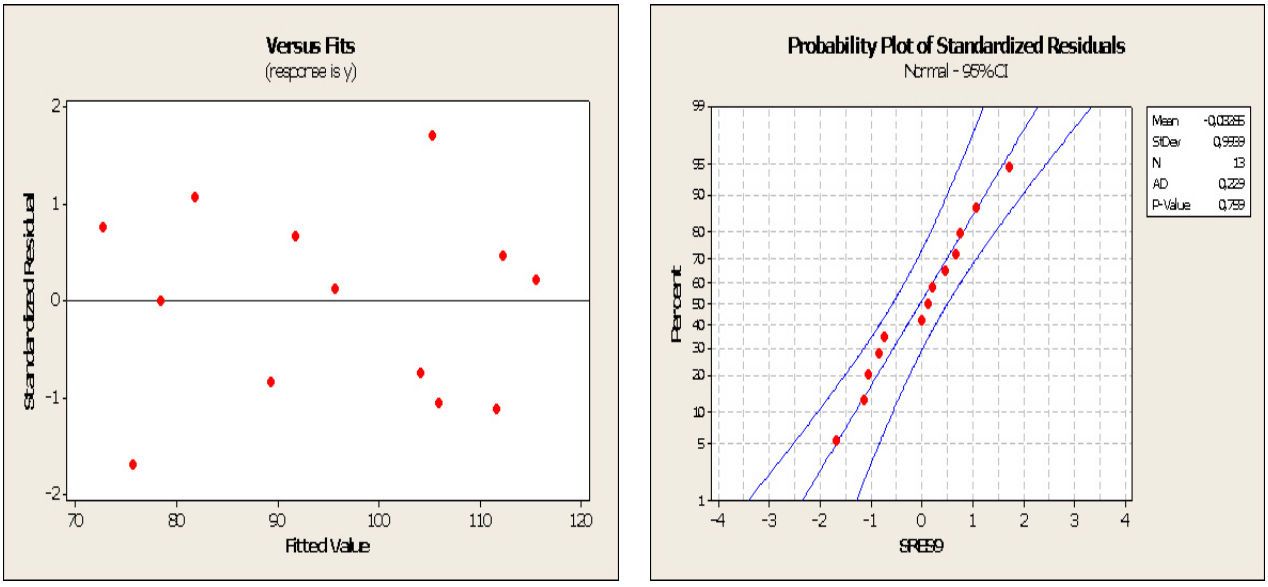

Figure 4: Residual plots (standardized residual versus fitted values in the left panel and normal plot based on standardized residuals in the right panel) for regression model A for the cement hydration data set.

Predictor	Coef	SE Coef	T	P
Constant	52,577	2,286	23,00	0,000
x1	1,4683	0,1213	12,10	0,000
x2	0,66225	0,04585	14,44	0,000

S = 2,40634 R-Sq = 97,9% R-Sq(adj) = 97,4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	2657,9	1328,9	229,50	0,000
Residual Error	10	57,9	5,8		
Total	12	2715,8			

Figure 5: Printout from statistical analysis of the cement hydration data set for model B.

			Mallows		x	x	x	x
Vars	R-Sq	R-Sq(adj)	Cp	S	1	2	3	4
1	67,5	64,5	138,7	8,9639				X
1	66,6	63,6	142,5	9,0771		X		
2	97,9	97,4	2,7	2,4063	X	X		
2	97,2	96,7	5,5	2,7343	X			X
3	98,2	97,6	3,0	2,3087	X	X		X
3	98,2	97,6	3,0	2,3121	X	X	X	
4	98,2	97,4	5,0	2,4460	X	X	X	X

Figure 6: Printout from statistical analysis of cement hydration data set.

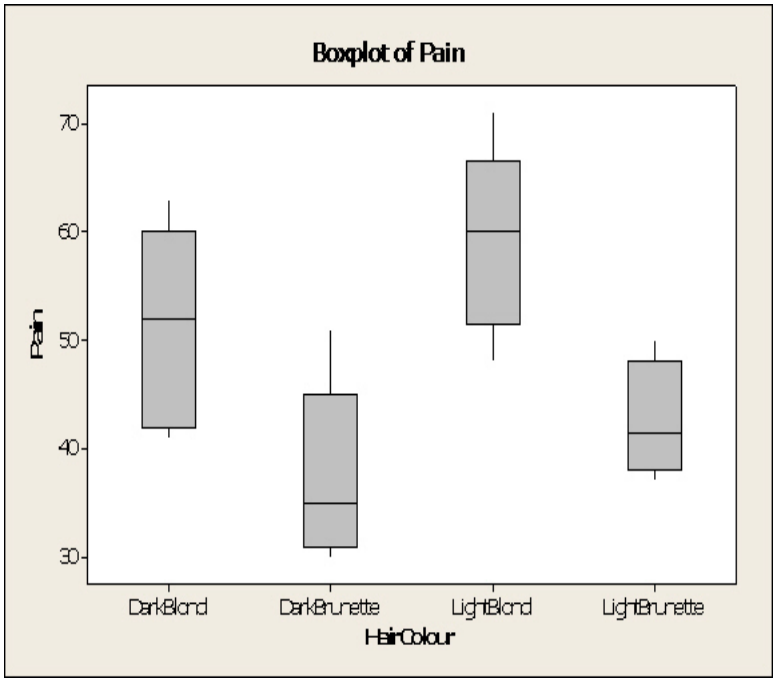

Figure 7: Boxplot of pain and haircolour data.

```
One-way ANOVA: Pain versus HairColour

Source      DF      SS     MS     F      P
HairColour  3       ?      453,6  ?   0,004
Error       ?       ?      ?
Total       18  2362,5

S = ?   R-Sq = 57,60%   R-Sq(adj) = 49,12%

Level          N    Mean  StDev
DarkBlond      5  51,200  9,284
DarkBrunette   5  37,400  8,325
LightBlond     5  59,200  8,526
LightBrunette  4  42,500  5,447
```

Figure 8: Printout from statistical analysis of the pain and hair colour data set.