Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Contact during exam:
Bo Lindqvist
Tel. 975 89 418

EXAM IN TMA4255
EXPERIMENTAL DESIGN AND APPLIED STATISTICAL METHODS

Wednesday 23 May 2007

Time: 09:00–13:00

Aids:
All printed and handwritten aids. All calculators permitted.

Grading: 13 June 2007

In your work on particular questions you may need to perform interpolations in the tables in order to find the desired values. You may do this in a rather rough manner ("by eye").

Problem 1

An athletics researcher has investigated motory skills for eight year old boys from three schools in, respectively, Trondheim, Namsos and Røros. The aim of the investigation was to find out whether there are differences in motory skills for pupils at the three schools.

The boys that participated in the investigation were given a motory test consisting of four subtasks. A motory quotient (MQ) was computed from the results of these tasks.

We let Y_{ij} be the MQ measured for boy no. j from school no. i, where $i = 1, 2, 3$ correspond to, respectively, the schools in Trondheim, Namsos and Røros, while j runs from 1 to 18 which was the number of participants from each of the three schools.

We assume that Y_{ij} is normally distributed with expected value μ_i and variance σ^2 ($i = 1, 2, 3$; $j = 1, 2, \ldots, 18$). Further, all the 54 observations are assumed to be independent.

The results are found on page 3.

a) The researcher was at the start of the study interested in investigating a possible difference in motory skills between pupils from Trondheim and Namsos only.

Use the given results for Trondheim and Namsos (i.e. the first two columns of the output on page 3) to perform a two-sample t-test for this situation. Write down the null hypothesis and the alternative hypothesis. Then write down and calculate the test statistic and give the conclusion when you choose the significance level 5%.

b) A colleague meant that the researcher should instead have used the Wilcoxon two-sample test in the testing problem of the previous point. Explain briefly what might be the reason for such a suggestion.

It turns out that the sum of the ranks of sample no. 1 (Trondheim) is $w_1 = 296$. Explain how one arrives at this number. You are not asked to perform the calculations.

Calculate the p-value of the test by using the given value of w_1. What is the conclusion of the test when you choose 5% significance level? Use the approximation to the normal distribution.

In the rest of the problem you shall use the results from all three schools. The researcher wants to find out whether there are differences between the three schools regarding motory skills of the pupils. He therefore wants to test the null hypothesis

$$H_0 : \ \mu_1 = \mu_2 = \mu_3$$

by using a one-way analysis of variance.

c) You are supposed to do this analysis. First calculate the sums of squares, SSA and SSE, by using the results given at the end of the exercise.

Then calculate the test statistic and find the conclusion when you use significance level 5%.

Show that the estimate of the standard deviation σ is $s = 13.2$.

d) It turns out that the test in the previous point leads to rejection of H_0. The researcher therefore wants to know which, if any, of the differences $\mu_1 - \mu_2$ $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$ that are significantly different from 0.

You should use either the Bonferroni method or Tukey's method for this problem, when it is required that the probability of reaching at least one wrong conclusion is at most 5% if in fact we have $\mu_1 = \mu_2 = \mu_3$.

j	y1j	y2j	y3j
1	100	90	90
2	86	75	114
3	71	111	110
4	80	80	127
5	95	105	102
6	77	91	72
7	104	69	80
8	96	105	94
9	91	108	93
10	82	79	96
11	85	80	96
12	72	83	85
13	92	119	106
14	97	97	108
15	82	105	108
16	85	111	83
17	71	79	89
18	78	76	95

$\bar{y}_{1\cdot} = 85.78, \quad \sum_{j=1}^{18}(y_{1j} - \bar{y}_{1\cdot})^2 = 1743.0$

$\bar{y}_{2\cdot} = 92.39, \quad \sum_{j=1}^{18}(y_{2j} - \bar{y}_{2\cdot})^2 = 4022.2$

$\bar{y}_{3\cdot} = 97.11, \quad \sum_{j=3}^{18}(y_{3j} - \bar{y}_{3\cdot})^2 = 3103.8$

$\bar{y}_{\cdot\cdot} = 91.76$

Problem 2

An experiment has been undertaken to investigate the resistance to corrosion of a certain type of steel plates which are exposed to sea water The experiment consisted in letting the plates be exposed to 10% hydrochloric acid (HCl) at resepctive temperatures 60, 70, 80 and 90 degrees Celsius during periods of 4, 6, 8, 10 and 12 hours, and thereafter measuring the weight loss in centigrams for each plate. (One has found by earlier experiments that 10% HCl attacks the steel in the same way as sea water, but at a much higher speed).

A single experiment was performed at each level combination of temperature and time. Let Y_i (centigram) be the weight loss, x_{1i} (degrees Celsius) be the temperature, and x_{2i} (hours) be time for the ith plate.

The data are given at the very end of the exam for your information. You will not need these raw data in order to solve the exercise.

An output from a MINITAB analysis of the data is given on the next page.

Regression Analysis: y versus x1; x2

```
The regression equation is
y = - 12,8 + 0,190 x1 + 0,439 x2


Predictor     Coef  SE Coef      T      P
Constant   -12,755    3,928  -3,25  0,005
x1         0,19040  0,04782   3,98  0,001
x2          0,4391   0,1890   2,32  0,033


S = 2,39077   R-Sq = 55,6%   R-Sq(adj) = 50,3%


Analysis of Variance

Source           DF       SS      MS      F      P
Regression        2  121,483  60,742  10,63  0,001
Residual Error   17   97,168   5,716
Total            19  218,652


Source  DF  Seq SS
x1       1  90,630
x2       1  30,853
```

a) Write down the model and the assumptions on which this analysis is based.

Is a significant amount of variation explained by the model? You are supposed to answer this question by writing down the relevant null hypothesis, write down the test statistic, choose a significance level and conclude by means of the MINITAB output.

What proportion of the variability in the data is explained by the model?

b) Define what is meant by the residuals of the observations in a regression analysis. What are the properties of the residuals? Explain briefly how various plots of residuals can be used to evaluate a suggested regression model.

Figure 1 on the next page shows some residual plots for the model that was studied in point (a). Do these indicate that the model gives a sufficient description of the data? Answer the question by commenting briefly on each of the four plots in the figure. How should they ideally look if the model was correct?

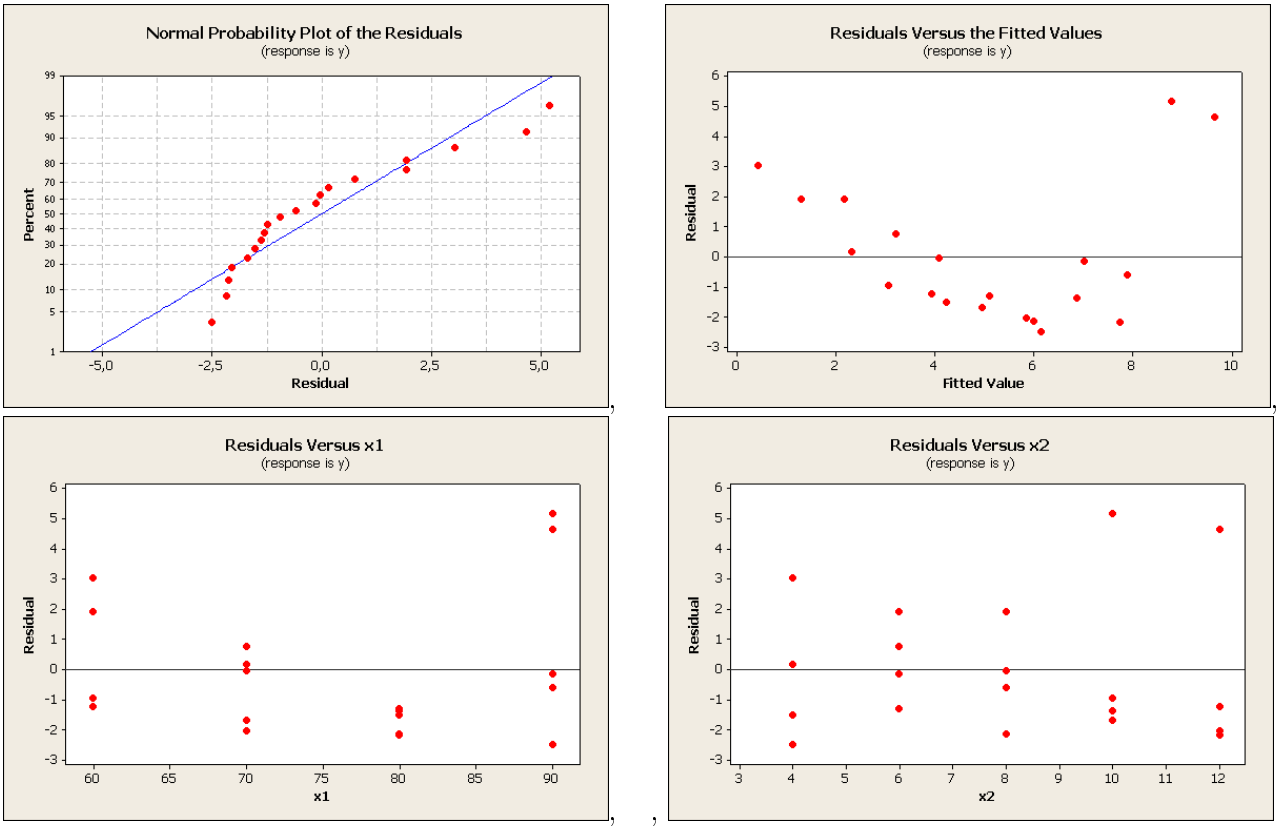

Figure 1: Residual plots for the model in point (a).

It was decided to extend the model from the first analysis by also including 2nd order terms of x_1 and x_2. This was done by introducing new predictors: $x_1 * x_2$ with values $x_{1i} \cdot x_{2i}$ for the ith steel plate, $x_1 * x_1$ with values x_{1i}^2, and $x_2 * x_2$ with values x_{2i}^2.

The following is a MINITAB output from a regression analysis of the given data when the predictors are $x_1, x_2, x_1 * x_2, x_1 * x_1, x_2 * x_2$.

Regression Analysis: y versus x1; x2; x1*x2; x1*x1; x2*x2

```
The regression equation is
y = 78,3 - 1,91 x1 - 2,98 x2 + 0,0494 x1*x2 + 0,0114 x1*x1 - 0,0176 x2*x2
```

Predictor	Coef	SE Coef	T	P
Constant	78,30	16,76	4,67	0,000
x1	-1,9069	0,4311	-4,42	0,001
x2	-2,9804	0,9578	-3,11	0,008
x1*x2	0,049355	0,008946	5,52	0,000

```
x1*x1       0,011350  0,002829   4,01  0,001
x2*x2       -0,01763   0,04226  -0,42  0,683


S = 1,26511   R-Sq = 89,8%   R-Sq(adj) = 86,1%


Analysis of Variance

Source           DF       SS      MS      F      P
Regression        5  196,245  39,249  24,52  0,000
Residual Error   14   22,407   1,600
Total            19  218,652


Source  DF  Seq SS
x1       1  90,630
x2       1  30,853
x1*x2    1  48,718
x1*x1    1  25,764
x2*x2    1   0,279
```

Next is shown a MINITAB output for "best subsets" regression.

Best Subsets Regression: y versus x1; x2; x1*x2; x1*x1; x2*x2

```
Response is y

                                                  x x x
                                                  1 1 2
                                                  * * *
                         Mallows                  x x x x x
Vars  R-Sq  R-Sq(adj)      C-p       S            1 2 2 1 2
   1  44,0       40,9     60,5  2,6088              X
   1  42,9       39,7     62,0  2,6337                X
   2  72,5       69,3     23,5  1,8796              X X
   2  70,4       66,9     26,5  1,9527                X   X
   3  77,8       73,7     18,3  1,7402            X X X
   3  75,3       70,7     21,7  1,8372              X X X
   4  89,6       86,9      4,2  1,2298            X X X X
   4  82,7       78,0     13,7  1,5896            X   X X X
   5  89,8       86,1      6,0  1,2651            X X X X X
```

c) Which model would you choose for the data? Give reasons for your answer, both by using the result of the regression analysis with the extended model, and by using the "best subsets" output.

What proportion of the variability in the data is explained by the chosen model? Compare with the model in point (a).

Calculate Mallow's C-p also for the model in point (a) and comment.

What mathematical motivation can be given for the extension of the model by including the new predictors?

d) In this point you shall use the model with all the five predictors $x_1, x_2, x_1{*}x_2, x_1{*}x_1, x_2{*}x_2$.

Test the null hypothesis that the coefficients of the two 2nd order terms $x_1 * x_1$ og $x_2 * x_2$ both are 0. Use significance level 5%.

Let $\hat{y}_0$ be the predicted value for the potential response y_0 of an observation made at temperature $x_1^0 = 80$ (degrees) and time $x_2^0 = 10$ (hours). Calculate the value of $\hat{y}_0$.

It may be shown that the estimated standard deviation of $\hat{y}_0$ is 0.503. Determine from this a 95% confidence interval for the expected value of y_0, and in addition a 95% prediction interval for y_0 itself.

What is the difference in interpretation of the two intervals?

Problem 3

A manufacturer of plastic equipment has performed a test of breaking strength for a certain product. Each of 280 plastic bars were molded under the same conditions and were tested at five positions. An assumption that each bar has a uniform composition implies that the number of breaks, X, for a given bar is binomial with $n = 5$ experiments and an unknown probability p. If all the bars have the same uniform strength, then p should be the same for all the 280 bars. If the bars, on the other hand, have different strength, then the p will vary from bar to bar.

One wishes to test the null hypothesis that all the bars have the same p, i.e. that the experiment gives us $X_1, X_2, \ldots, X_{280}$ which are independent and binomial$(5, p)$ for an unknown p.

The data are given by the following table:

x = number of breaks for a bar	Number of bars with $X = x$
0	157
1	69
2	35
3	17
4	1
5	1

a) Show that $\hat{p} = 0.142$ is a reasonable estimate of p if the null hypothesis is valid.

What is the probability of, respectively, 0,1,2,3,4,5 breaks for a bar with this value for p?

Use these probabilities to perform a chi-square test for the given null hypothesis. Choose significance level 1%. Give a reason for the number of degrees of freedom for the test statistic.

Why should the last three cells, corresponding to $X = 3, 4, 5$, be pooled to one cell in the testing?

The data from Problem 2:

Row	x1	x2	y
1	60	4	3,47
2	70	4	2,49
3	80	4	2,71
4	90	4	3,64
5	60	6	3,24
6	70	6	3,97
7	80	6	3,80
8	90	6	6,88
9	60	8	4,11
10	70	8	4,05
11	80	8	3,86
12	90	8	7,30
13	60	10	2,11
14	70	10	3,28
15	80	10	5,49
16	90	10	13,96
17	60	12	2,71
18	70	12	3,80
19	80	12	5,58
20	90	12	14,31