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

Friday, May 25, 2012
Time: 9:00–13:00

Number of credits: 7.5.

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

Grading finished: June 18, 2012.

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

Note that:

- Significance level 5% should be used unless a different level is specified.
- All answers need to be justified.

Problem 1 Vitamin C

In a medical study on guinea pigs two different sources of intake of vitamin C, orange juice (supplement 1) and synthetic ascorbic acid (supplement 2) were used. The response measure was the length of odontoblast cells in the incisor teeth. The researchers wanted to know if there was a difference in the response measure for the two different supplements.

$X_1, X_2, \ldots, X_{n_1}$ will denote the odontoblast lengths of a random sample of n_1 guinea pigs that was given supplement 1, and $Y_1, Y_2, \ldots, Y_{n_2}$ will denote the odontoblast lengths of a random sample of n_2 guinea pigs that was given supplement 2. We assume that $\mathrm{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$, $\mathrm{E}(Y_j) = \eta$ and $\mathrm{Var}(Y_j) = \tau^2$ for $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$, and that the two random samples are independent.

In total $n_1 = 10$ guinea pigs were given supplement 1 and $n_2 = 10$ guinea pigs were given supplement 2. The (sorted) data set is presented in the table below. The lenghts are in micrometers (10^{-6} meter).

Treatment	Observations									
Supplement 1: Orange juice	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
Supplement 2: Ascorbic acid	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

Summary statistics for these data are $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 13.18$, $\bar{y} = \frac{1}{10}\sum_{j=1}^{10} y_j = 8.00$, $s_x = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2} = 4.44$, and $s_y = \sqrt{\frac{1}{9}\sum_{j=1}^{10}(y_j - \bar{y})^2} = 2.77$.

a) Write down the null and alternative hypothesis that the researchers would like to test.
What assumptions need to be made in order to perform a two-sample t-test?
Is it sensible to assume that the variances σ^2 and τ^2 are equal? Here you may use significance level $\alpha = 0.02$ if you choose to perform a test of equality of variances.
Perform a t-test.
Would you conclude that there is a difference in odontoblast cell length for the supplement 1 and supplement 2 populations?

b) An alternative to the two-sample t-test is the Wilcoxon rank-sum test (also called the Mann-Whitney test). What are the assumptions underlying the Wilcoxon rank-sum test?
When should the Wilcoxon rank-sum test be used instead of the two-sample t-test?
Perform the Wilcoxon rank-sum test based on the data in the table above.
Would you conclude that there is a difference in odontoblast cell length for the supplement 1 and supplement 2 populations?

Problem 2 Chemical yield

We will look at a chemical process, with the aim to construct a statistical model that relates the chemical yield to the temperature and pressure of the process. The following notation is used.

- y, `yield`. The yield of the chemical product produced in the process.
- x_1, `temperature`. The temperature the process was operated under.
- x_2, `pressure`. The pressure the process was operated under.

The chemical process was run for $n = 21$ different combinations of temperature and pressure, and the yield was measured.

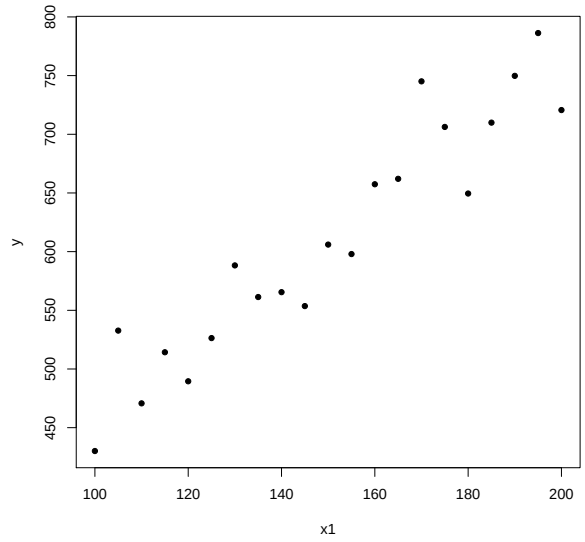
A multiple linear regression model can be used to analyse the data. Let (y_i, x_{1i}, x_{2i}) denote the measurements from the ith run of the process.

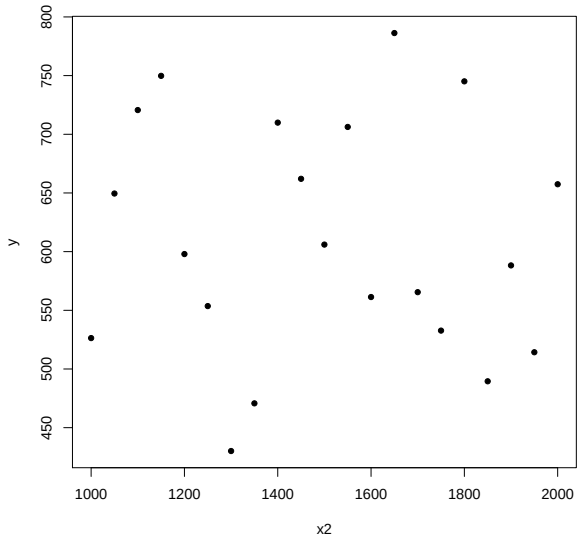
$$\text{Model A} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

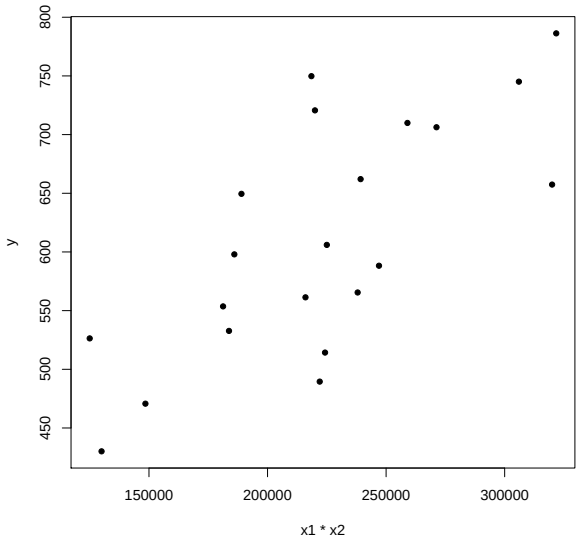
where the ε_i's are i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots, n$.

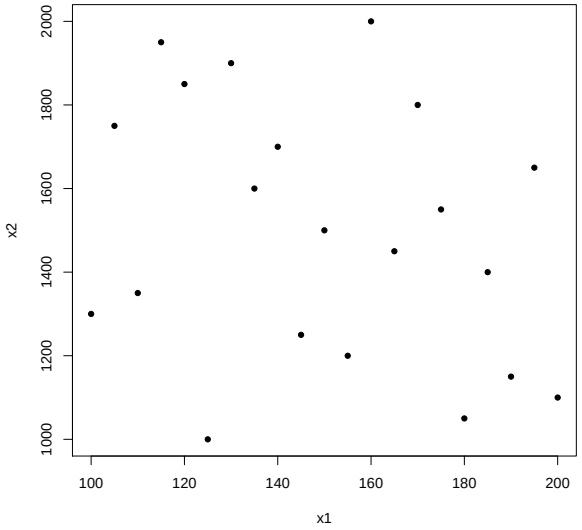
Scatter plots are found on page 4 (upper and middle rows). Model A was fitted to the data, and results from the fit are found on page 5, and plots of studentized residuals are found on page 4 (bottom row).

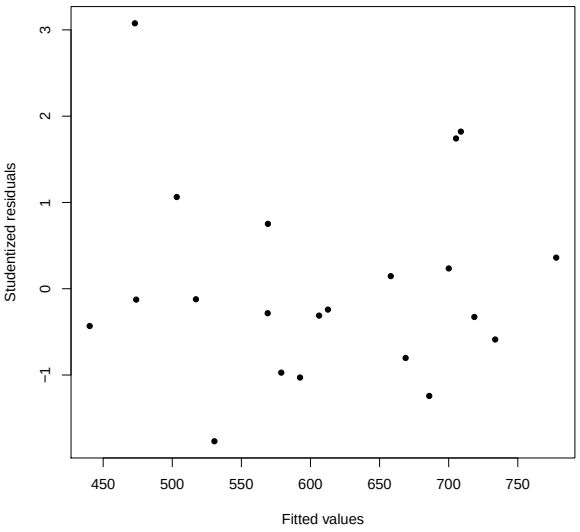
a) Write down the fitted regression model.
What is the estimate for σ^2?
A p-value is given in the row labeled `x2` in the results. Explain what this p-value means. Based on the plots and the result from the fit, would you say that model A is a good model for the data? You need to point out all the features of the fit and plots that you are using to come to your conclusion.

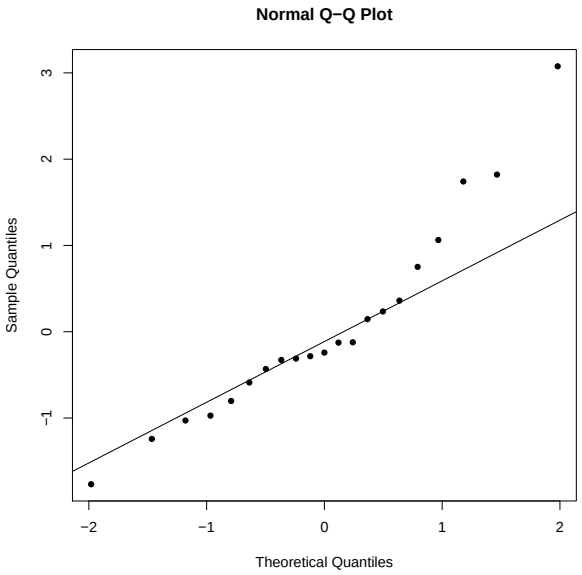

y
x1
100
120
140
160
180
200
450
500
550
600
650
700
750
800


y
x2
1000
1200
1400
1600
1800
2000
450
500
550
600
650
700
750
800


y
x1 * x2
150000
200000
250000
300000
450
500
550
600
650
700
750
800


x2
x1
100
120
140
160
180
200
1000
1200
1400
1600
1800
2000


Studentized residuals
Fitted values
450
500
550
600
650
700
750
−1
0
1
2
3


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−1
0
1
2
3

```
Regression analysis

Predictor       Coef    SE Coef      T      P
Constant       181.8      156.6   1.16  0.262
x1             2.146      1.018   2.11  0.050
x2           -0.0440     0.1043  -0.42  0.678
x1*x2      0.0007774  0.0006942   1.12  0.278

S = 27.2502   R-Sq = 93.8%   R-Sq(adj) = 92.7%

Analysis of Variance
Source          DF      SS     MS      F      P
Regression       3  190830  63610  85.66  1.825e-10
Residual Error  17   12624    743
Total           20  203454
```

Assume that an intercept, β_0, is present in the regression model. We will now look at different regression models where combinations of the covariates x_1, x_2 and x_1x_2 are present. Results from fitting different regression models to the data are presented in the table below. Each row in the table corresponds to one model. The number of explanatory variables included in each model is found in the column labeled `par` (excluding the intercept). The column labeled MSE gives the mean squared error for the regression.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	par	MSE	R2	R2.adj	C_p
1	151	3.067			1	1179	89.0	88.4	13.2
2	641		-0.020		1	10667	0.4	-4.9	255.9
3	315			0.001	1	5093	52.4	49.9	113.3
4	15	3.263	0.071		2	753	93.3	92.6	3.3
5	117	2.563		0.0005	2	709	93.7	93.0	2.2
6	505		-0.258	0.002	2	885	92.2	91.3	?
7	182	2.146	-0.044	0.001	3	743	93.8	92.7	4.0

b) The rightmost column of the above table contains the Mallow's C_p statistic. One of the numbers in this column is replaced by a question mark. Write down the definition of C_p, and calculate the C_p that is missing.
Explain how you can use C_p to compare the different regression models.
Which of the 7 regression model do you rate to be the "best" for this dataset?

Problem 3 Treatment of tennis elbow

The term *tennis elbow* is used to describe a state of inflammation in the elbow, causing pain. This injury is common in people who play racquet sports, however, any activity that involves repetitive twisting of the wrist (like using a screwdriver) can lead to this condition. The condition may also be due to constant computer keyboard and mouse use.

In a randomized clinical study the aim was to compare three different methods for treatment of tennis elbow, A: physiotherapy intervention, B: corticosteroid injections and C: wait-and-see (the patients in the wait-and-see group did not get any treatment but was told to use the elbow as little as possible).

We will look at the short-term effect of treatment by studying measurements at 6 weeks. All patients participating in the study only had one affected arm. There were several outcomes measured for the study, and we will in the following problems look at two different outcomes.

a) The treatment was considered a success if the patient rated him- or herself as either much improved or completely recovered (on a standardized scale of improvement). The number of successes and failures after 6 weeks of treatment are presented in the table below.

Treatment	Failure	Success	Total
A (physiotherapy)	22	41	63
B (injection)	14	51	65
C (wait-and-see)	44	16	60
Total	80	108	188

We would like to investigate if the rate of success differs between the treatments. Write down the null- and alternative hypothesis and perform one hypothesis test using the information given in the table above. To easen the computational burden you may use that the χ^2-test statistic equals 36.6 and only show the calculation of one of the 6 terms in the sum.
What are the assumptions you need to make to use this test?
What is the conclusion from the test?

We now turn to the outcome measure called *pain-free grip force*. This was measured by a digital grip dynamometer and normalized to the grip force of the unaffected arm. A pain-free grip force of 100 would mean that the affected and the unaffected arm performed equally good. Summary statistics for each of the treatment groups are presented in the table on the top of page 7.

Treatment	Sample size	Average	Standard deviation
A (physiotherapy)	63	70.2	25.4
B (injection)	65	83.6	22.9
C (wait-and-see)	60	51.8	23.0
Total	188	69.0	

b) We would like to investigate if the expected pain-free grip force varies between the treatment groups. Write down the null- and alternative hypothesis and perform one hypothesis test using the summary statistics in the table above.
What are the assumptions you need to make to use this test?
What is the conclusion from the test?

c) We would like to compare the expected pain-free grip force for all treatment groups pairwisely. Let us now assume that the sample size for each treatment group is $n_A = n_B = n_C = 63$ and that the average and standard deviation for each treatment group is as given in the table above.

Perform the comparisons by constructing simultanous confidence intervals for the expected difference between the pain-free grip force for all pairs of treatments using Tukey's method. Use an overall confidence level of 95% for all comparisons.
What assumptions do you need to make?
What is the common individual confidence level used for each of the comparisons?

d) Let μ_A be the expected pain-free grip force for a population where the physiotherapy intervention treatment is used to treat tennis elbow, and μ_C be the expected pain-free grip force for a population where the wait-and-see treatment is used. Define the relative difference between these two expected values as

$$\gamma = \frac{\mu_A - \mu_C}{\mu_C}.$$

This can be interpreted as the expected relative gain by using physiotherapy instead of wait-and-see. Based on two independent random samples of size n_A and n_C from the physiotherapy and wait-and-see treatment groups, respectively, suggest an estimator, $\hat{\gamma}$, for γ.

Use approximate methods to find the expected value and variance of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.
Use the sample sizes, averages and empirical standard deviations for the physiotherapy and wait-and-see treatment groups presented on the top of this page to calculate $\hat{\gamma}$ numerically and estimated numerical values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.

If you were told that the two samples were not independent, how would that affect your approximation of $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$?