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, June 7, 2013
Time: 9:00–13:00

Number of credits: 7.5.

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

Grading finished: June 28, 2013.

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

Note that:

- 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.

Problem 1 Body fat percentage and foot treatment

The body fat percentage of a person can be measured by sending a very small electrical signal through the body to measure body impedance. The signal is conducted through the water contained in the body. Lean muscle has much more water than fat tissue, and this can be used to convert the signal information into a measurement of body fat percentage.

In a master thesis at the Obesitas Clinic at St. Olav's Hospital the aim was to investigate if pedicure treatment would affect the measurement of body fat (by the technology described above).

A total of 40 persons were enrolled in the study, all with body mass index (weight divided by squared height) above 30 kg/m^2. All persons were measured with the impendance technology twice, before and after a foot pedicure treatment. For person i let X_{1i} denote the fat body percentage measurement before the treatment and X_{2i} denote the fat body percentage measurement after the treatment, $i = 1, \ldots, 40$.

Scatter plots and normal plots of the data are presented in Figure 1.

In Figure 2 you find printouts from performing the following four statistical analyses:

A a one-sample t-test on the differences $X_{2i} - X_{1i}$,

B a two-sample t-test (not assuming equal variances) on the before and after pedicure treatment measurements,

C a Wilcoxon signed-rank test on the differences $X_{2i} - X_{1i}$, and

D a Wilcoxon rank-sum test (Mann–Whitney) on the before and after pedicure treatment measurements.

a) Are the two samples (before and after pedicure) independent samples?
Based on the normal plots in Figure 1 would you assume that any of the before sample, the after sample or the difference sample $(X_{2i} - X_{1i})$ can be seen to come from a normal population?
Based on your answers above, decide which of the four analyses A–D you think fits the research question and the data the best.
For your chosen analysis write down the null and alternative hypotheses begin tested, list the assumptions you make to perform the hypothesis test and report the result of the hypothesis test.

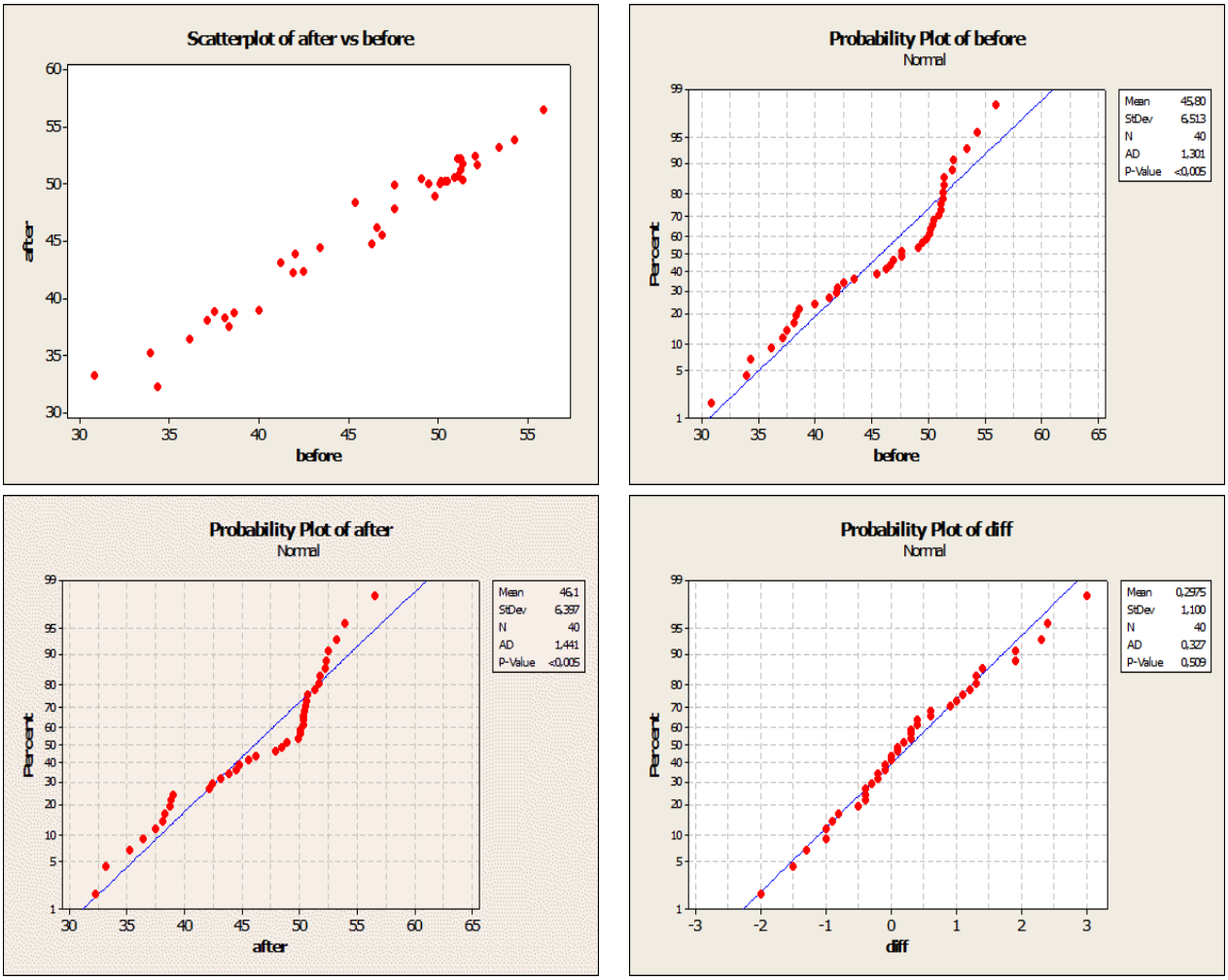

Figure 1: Scatter plot and normal plots for the body fat percentage data. The `diff` measurements are $X_{2i} - X_{1i}$.

```
A: One-Sample T: diff
--------------------
Test of mu = 0 vs not = 0

Variable   N   Mean  StDev  SE Mean       95% CI          T      P
diff        40  0,298  1,100    0,174  (-0,054; 0,649)  1,71  0,095

B: Two-sample T for after vs before
---------------------------------
             N   Mean  StDev  SE Mean
after     40  46,10   6,40      1,0
before  40  45,80   6,51      1,0

Difference = mu (after) - mu (before)
Estimate for difference:  0,30
95% CI for difference:  (-2,58; 3,17)
T-Test of difference = 0 (vs not =):
T-Value = 0,21  P-Value = 0,837  DF = 77

C: Wilcoxon Signed Rank Test: diff
-------------------------------
Test of median = 0,00 versus median not = 0,00

        N     N for Wilcoxon   Test Statistic      P         Median
diff  40     38                      470,0          0,151     0,2500

D: Mann-Whitney Test and CI: after; before
----------------------------------------
            N    Median
after     40   48,650
before  40   47,600

Point estimate for ETA1-ETA2 is 0,200
95,1 Percent CI for ETA1-ETA2 is (-1,998;2,699)
W = 1644,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,8211
The test is significant at 0,8211 (adjusted for ties)
```

Figure 2: Printout from statistical analyses of the body fat percentage data.

Problem 2 Process control with resistors

In a production plant samples of three resistors are taken every hour, and the resistance, in ohms, are measured. Let X_{ij} be the resistance measurement for resistor j, and sample i, where $j = 1, 2, 3$ and $i = 1, 2, \ldots, k$. Further, $\bar{X}_i = \frac{1}{3}\sum_{j=1}^{3} X_{ij}$, $S_i = \sqrt{\frac{1}{2}\sum_{j=1}^{3}(X_{ij} - \bar{X}_i)^2}$, $\bar{\bar{X}} = \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i$, and $\bar{S} = \frac{1}{k}\sum_{i=1}^{k} S_i$.

Based on $k = 30$ samples, assumed to be in control, we find $\bar{\bar{x}} = 5.095$ and $\bar{s} = 0.058$.

a) Construct a S-chart and a $\bar{X} - S$-chart (with 3σ limits).
A new sample is measured, with $\bar{x} = 5.15$ and $s = 0.10$. Is the process in control for this sample?

What would be the advantange of using a cusum-chart instead of a Shewhart-chart (as constructed here)?

Problem 3 Flying-bomb hits

We will examine a classical data set of flying-bomb hits in the south of London during World War II.

The city was divided into small areas of one-quarter square kilometers each. Assume that we choose one such small area at random, and let X be the number of bomb hits to this area. Based on statistical reasoning we want to investigate if X is a Poisson random variable. In our situation this means that the probability that the small area we study is hit exactly k times, $P(X = k)$, is given by

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots$$

for some intensity $\lambda > 0$.

We will use observations of the number of flying-bomb hits to each of $n = 576$ small areas in London, where we assume that the intensity, λ, is the same for all areas. Table 1 shows the observed number of small areas with 0, 1, 2, 3, 4 or more hits. There were a total of 537 hits, so the average number of hits per small area was $\hat{\lambda} = 537/576 = 0.93$.

Hits (k)	0	1	2	3	4 or more	Total
Observed number of small areas with k hits	229	211	93	35	8	576

Table 1: The observed number of small areas with k bomb hits.

a) Perform a hypothesis test to provide an answer to the following question: Can the observations in Table 1 be seen to be a random sample from a Poisson distribution? You may use the estimate $\hat{\lambda} = 0.93$ in your calculations.
What is your conclusion to the hypothesis test?

Problem 4 Teaching reading

In a randomized study the aim was to compare three methods for teaching reading, one method currently in use (A), and two new methods (B and C). A total of 66 pupils were randomly assigned to one of the three teaching methods, with 22 pupils for each method.

We will look at data on reading score. Reading score is a numerical value, and high value for the reading score is preferred. A box plot of the data is presented in Figure 3, and summary statistics are given in Table 2.

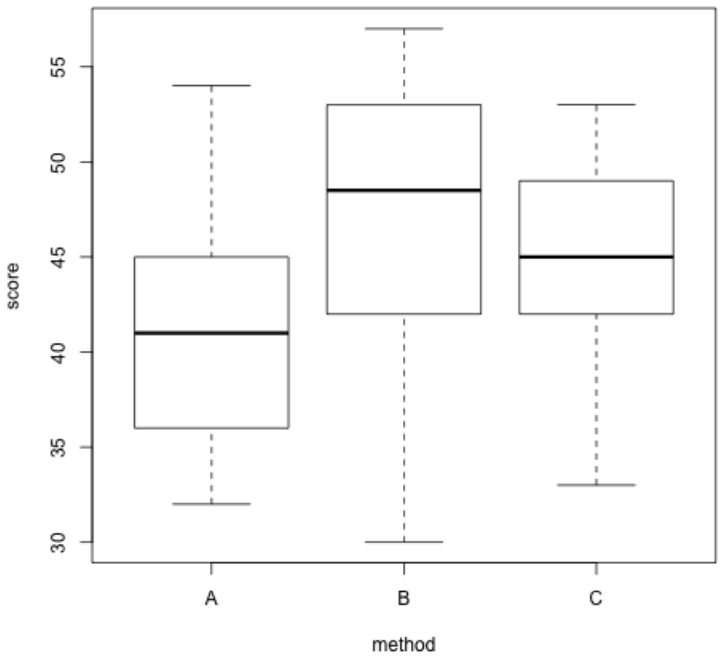

Figure 3: Box plot for reading data

Method	Sample size	Average	Standard deviation
A	22	41.05	5.636
B	22	46.73	7.388
C	22	44.27	5.767
Total	66	44.02	

Table 2: Summary statistics for the reading data

a) We would like to investigate if the expected reading score varies between the teaching methods. Write down the null and alternative hypotheses and perform a single hypothesis test based on the summary statistics in Table 2.
What are the assumptions you need to make to use this test?
What is the conclusion from the test?

We will now compare the two new teaching methods, method B and C.

b) Let μ_B and μ_C be the expected scores for teaching methods B and C. We would like to study the ratio, γ, between these two expected scores,

$$\gamma = \frac{\mu_B}{\mu_C}.$$

Suggest an estimator, $\hat{\gamma}$, for γ.

Use Taylor methods to approximate the expected value and standard deviation of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{SD}(\hat{\gamma}) = \sqrt{\mathrm{Var}(\hat{\gamma})}$.
Use the relevant data in Table 2 to calculate $\hat{\gamma}$, and the estimated approximate values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{SD}(\hat{\gamma})$ numerically.

Problem 5 Concrete

In an article in *Journal of Computing in Civil Engineering* the aim was to make a prediction model for the quantity of concrete, y, to be used in the construction of a silo complex. The prediction model will be used in the design stage of a construction job. A total of 23 possible explanatory variables were listed, and we will look at four of these. The following description is given.

- y, `concrete`. Quantity of concrete measured in m^3.
- x_1, `volume`. The volume of the silo complex.
- x_2, `perimeter`. The perimeter of the silo complex.

- x_3, `waste`. Waste percent in concrete.
- x_4, `steel`. The number of reinforcing steel crews.

Data are available from 28 construction jobs. Scatter plots are found in Figure 4.

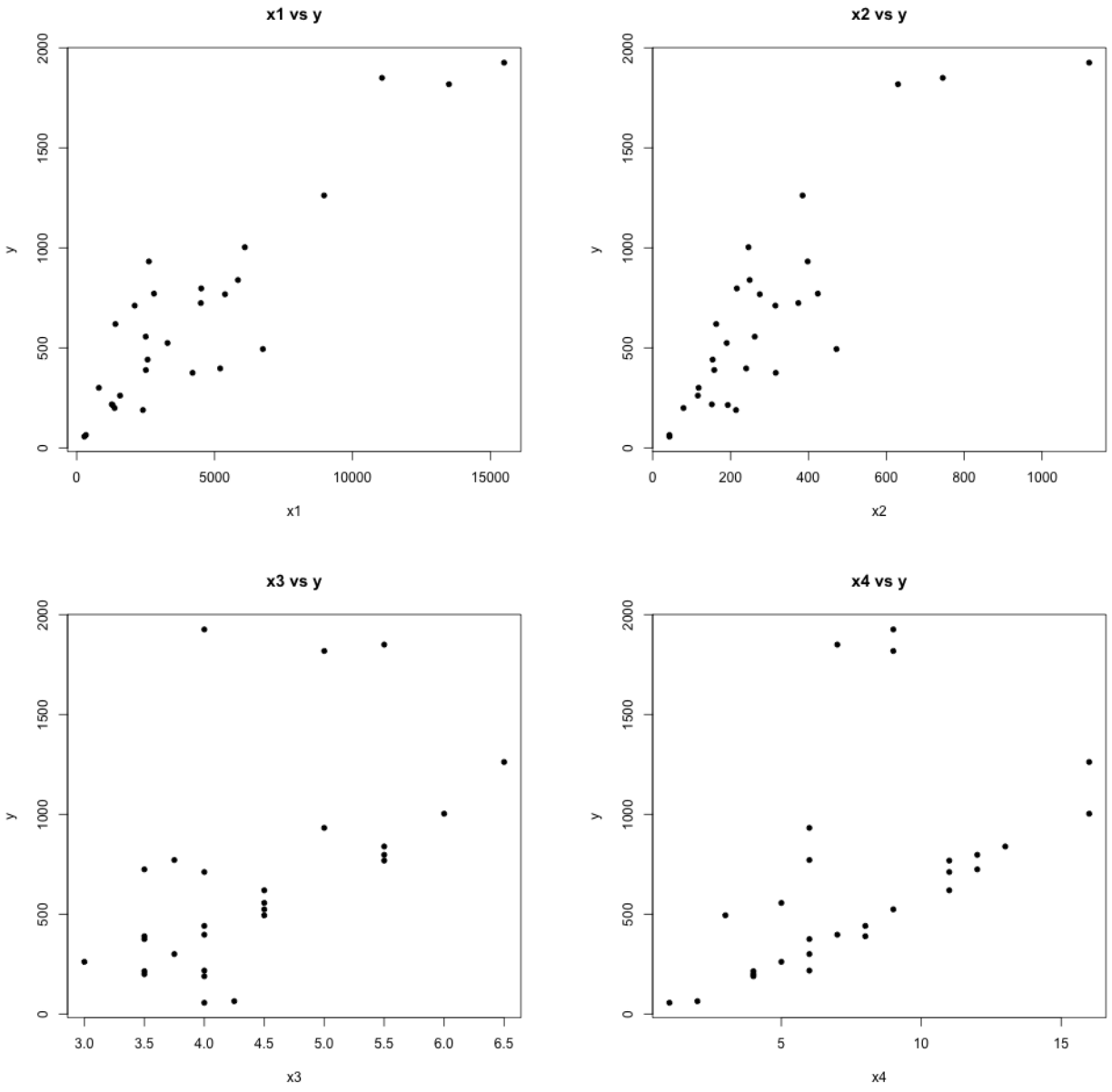

Figure 4: Scatter plots for the concrete quantities data set.

```
Predictor      Coef  SE Coef      T      P
Constant    -574,2    190,1  -3,02  0,006
x1         0,02670  0,02142   1,25  0,225
x2          1,3612   0,3241   4,20  0,0003
x3          124,08    49,39   2,51  ?
x4           23,22    10,28   2,26  0,034

S = 158,231   R-Sq = ? %   R-Sq(adj) =  90,56 %

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       4  6586535  1646634  65,77  0,000
Residual Error  23   575852    25037
Total           27  7162388
```

Figure 5: Printout from statistical analyses for Model A of the concrete data set.

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 job i, where $i = 1, \ldots, 28$. 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 $i = 1, \ldots, 28$. Printout from a statistical analysis is found in Figure 5 and plots of studentized residuals are found in Figure 6. Two of the numerical values in the printout have been replaced by question marks.

a) Write down the estimated regression equation.

Now turn to the estimated regression coefficient for x_3, `waste`, in this model. How would you explain this number to the common man (that does not know linear regression)? Is the effect of x_3, `waste`, significant in this model?

Calculate the R^2 and explain how you can interpret this value.

Do you think Model A is a good model for the data? To answer this you need to comment on important features in the printout from the statistical analysis and the residual plots in Figure 6.

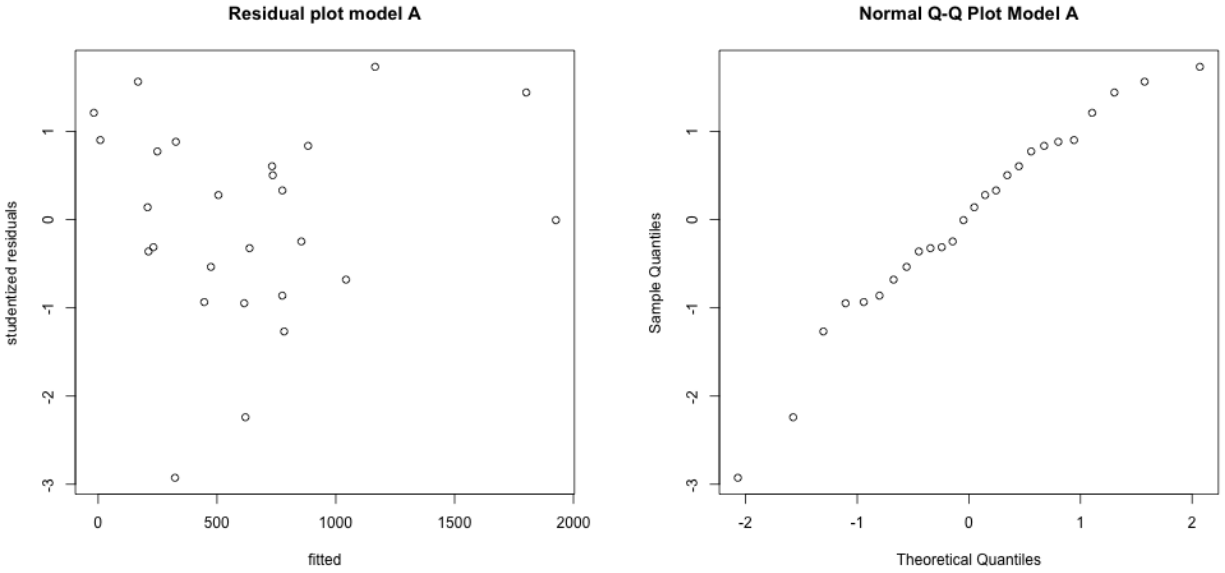

Figure 6: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for Model A for the concrete quantities data set.

We now want to compare the full regression model (model A), with a reduced model (called model B) with only x_2 (`perimeter`) and x_3 (`waste`).

$$\text{Model B: } y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

The results from fitting model B are found in Figure 7.

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.

$$H_0\colon \beta_1 = \beta_4 = 0 \text{ vs. } H_1\colon \beta_1 \text{ and } \beta_4 \text{ are not both zero}$$

Perform the hypothesis test and conclude.

Would you prefer to use Model A or B?

```
Predictor         Coef  SE Coef         T            P
Constant    -815,6       174,1    -4,68       8,45e-05
x2            1,7575     0,1521    11,55      1,62e-11
x3            219,65      40,09      5,48      1,09e-05

S = 176,717   R-Sq = 89,1%   R-Sq(adj) = 88,2%

Analysis of Variance
Source           DF       SS       MS       F      P
Regression        2  6381667  3190833  102,18  0,000
Residual Error   25   780721    31229
Total            27  7162388
```

Figure 7: Printout from statistical analyses for Model B of the concrete data set.

We will now use Model B.

c) Assume that a new construction job is planned and that we expect that $x_2 = 300$ and $x_3 = 4.4$ for this new job. What would be the best prediction for the quantity of concrete to be used?

Let X denote the design matrix used to fit Model B, and let $\boldsymbol{x}_0$ be the covariate vector we want to use in the prediction, $\boldsymbol{x}_0^T = [1\ 300\ 4.4]$. Is is given that then $\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0 = 0.0357$. Use this information to construct a 95% prediction interval for the quantity of concrete to be used when $x_2 = 300$ and $x_3 = 4.4$.
What is the interpretation of this interval?