NTNU

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: 3 June 2016

Examination time (from–to): 09:00-13:00

Permitted examination support material: C: Yellow, stamped A4 sheet with your own handwritten notes, Tabeller og formler i statistikk (Tapir forlag/Fagbokforlaget). Specified 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 of pages enclosed: 0

Checked by:

Date Signature

Problem 1 Fish and parasites

In an experiment 141 fish were placed in a large tank. The fish were categorized by their level of parasitic infection, either uninfected, lightly infected, or highly infected. Some of the fish were eaten by large birds of prey. It is to the parasites advantage to be in a fish that is eaten by a bird, as this provides an opportunity to infect the bird in the parasites' next stage of life. The following cross-table was observed.

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

a) The researchers performing the experiment wanted to investigate if *being eaten or not* and *level of parasitic infection* could be seen as two dependent events?

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.

What is the conclusion based on this test?

Problem 2 Cigarettes

The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide content. Each of these substances are hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke.

In a study the following variables were measured for $n = 25$ cigarette brands,

- y: Carbon monoxide (CO) content (mg),
- x_1: Tar content (mg),
- x_2: Nicotine content (mg), and
- x_3: Weight (g).

First, three separate simple regressions were fitted to study the relationship between the CO content and each of the variables x_1, x_2 and x_3:

$$y_i = \beta_{01} + \beta_1 x_{1i} + \epsilon_i \tag{1}$$
$$y_i = \beta_{02} + \beta_2 x_{2i} + \epsilon_i \tag{2}$$
$$y_i = \beta_{03} + \beta_3 x_{3i} + \epsilon_i \tag{3}$$

where ϵ_i is i.i.d. $N(0, \sigma^2)$ for $i = 1, ..., n$.

The MINITAB output from a statistical analysis is found in Figure 1.

```
Simple regression for x1:
Predictor      Coef  SE Coef      T      P
Constant     1,4129   0,6482   2,18  0,040
x1          0,92813  0,05283  17,57  0,000

S = 1,11865   R-Sq = 93,3%   R-Sq(adj) = 93,0%
```

```
Simple regression for x2:
Predictor     Coef  SE Coef      T      P
Constant    -0,238    1,083  -0,22  0,828
x2          14,860    1,247  11,92  0,000

S = 1,58842   R-Sq = 86,6%   R-Sq(adj) = 86,0%
```

```
Simple regression for x3:
Predictor    Coef  SE Coef      T      P
Constant    -3,86    10,44  -0,37  0,715
x3          16,56    10,82   1,53  0,140

S = 4,12276   R-Sq = 9,6%   R-Sq(adj) = 5,5%
```

Figure 1: Printout from fitting simple linear regressions presented in Equations (1)-(3) for the cigarette dataset.

a) Comment on the results from the simple linear regressions in Figure 1.

We will now focus on the simple linear regression for x_2 as in Equation (2) that is fitted in the middel panel of Figure 1.

In the simple linear regression for x_2 a p-value is given in the row labeled x2. Explain what this p-value means.

Find a 90% confidence interval for β_2 in the simple linear regression for x_2.

What is an appropriate estimate for σ in the simple linear regression model for x_2?

Further, a multivariate regression with both x_1 and x_2 as covariates was performed:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{4}$$

where ϵ_i is i.i.d. $N(0, \sigma^2)$ for $i = 1, ..., n$.

The MINITAB output from fitting the multivariate regression model is found in Figure 2. Pairwise scatter plots for x_1, x_2, x_3 and y are found in the upper part of Figure 3 and pairwise Pearson correlations in the lower part of Figure 3.

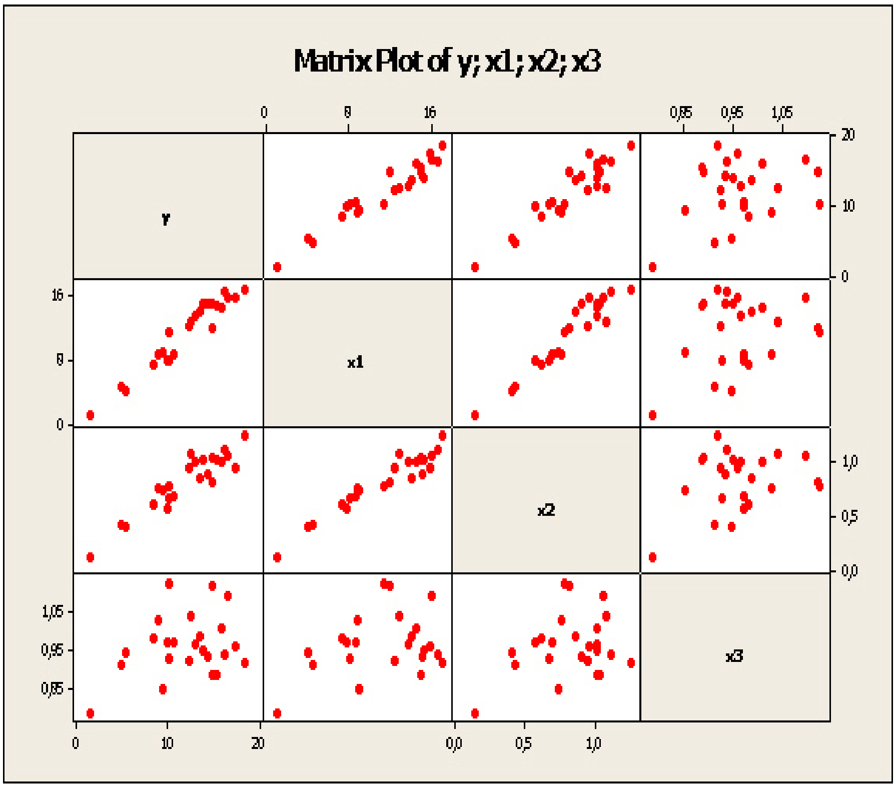
```
Predictor     Coef  SE Coef     T      P
Constant   1,3089   0,8483  1,54  0,138
x1         0,8918   0,1927  4,63  0,000
x2          0,629    3,203  0,20  0,846

S = 1,14392   R-Sq = 93,4%   R-Sq(adj) = 92,7%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  386,27  193,13  147,59  0,000
Residual Error  21   27,48    1,31
Total           23  413,75
```

Figure 2: Printout from statistical analysis of the cigarette data for the model in Equation (4).



```
          y      x1      x2
x1    0,966
x2    0,931   0,960
x3    0,310   0,284   0,286
```

Figure 3: Pairwise scatter plots (upper part) and pairwise Pearson correlation (lower part) between variables y, x_1, x_2 and x_3 in the cigarette data set.

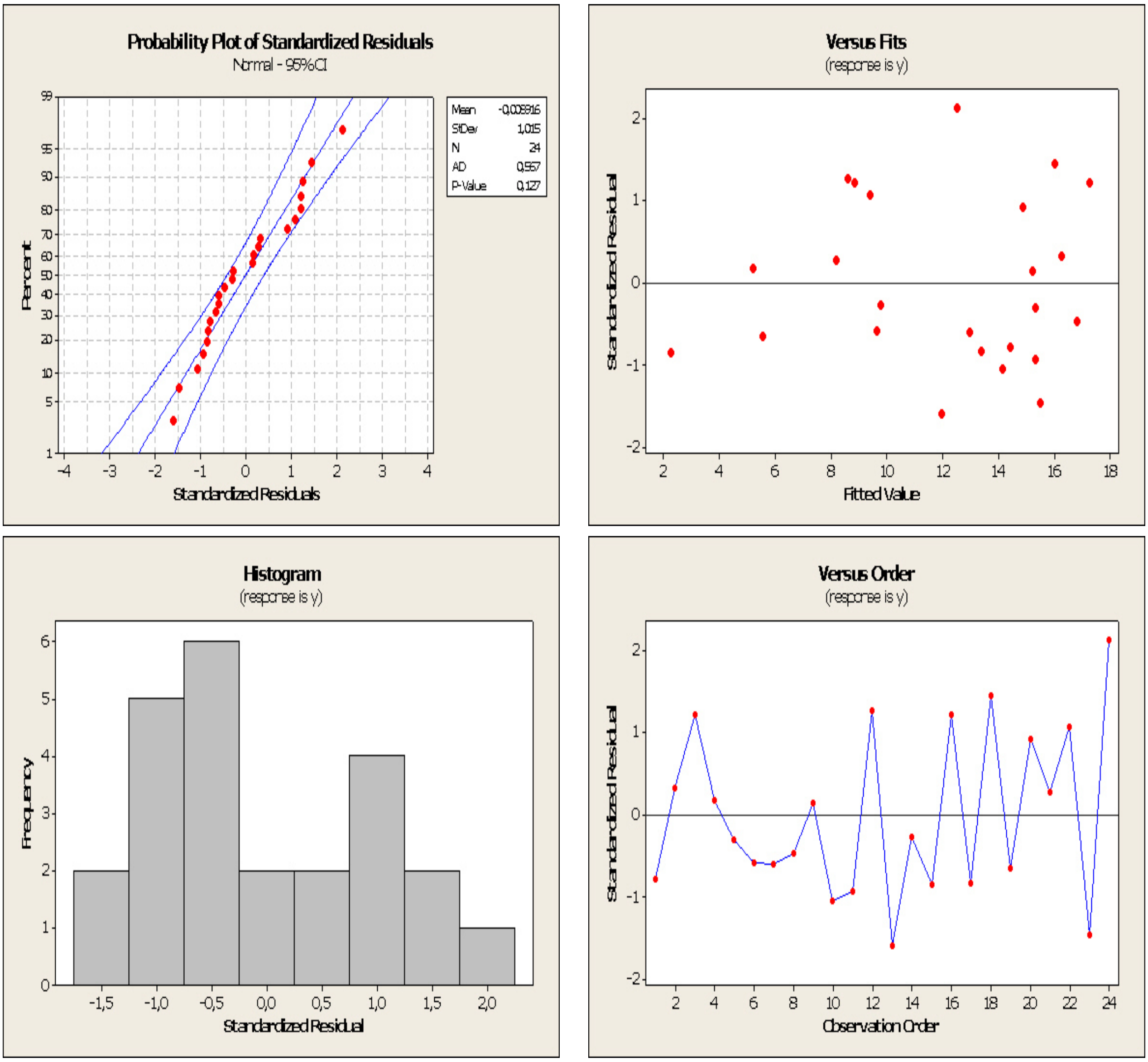


Figure 4: Residual plots (normal plot based on standardized residuals in the upper left panel, standardized residual versus fitted values in the upper right panel, histogram based on standardized residuals in lower left panel and standardized residual versus observation order in the lower right panel) for the regression model in Equation (4) for the cigarette data set.

b) Based on the model in Equation (4) what is the predicted CO content when $x_1 = 10$ and $x_2 = 0.8$?

Based on the plots in Figure 4 and statistical results of the model fit in Figure 2, would you say that the model in Equation (4) 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 arrive at your conclusion.

c) When fitting a simple linear regression with only nicotine (x_2) as covariate, middel panel of Figure 1, we found that the effect of nicotine was significant at significance level 5%, but in the multiple linear regression with tar (x_1) and nicotine (x_2), Figure 2, nicotine is not significant. What could be the reason for this? Justify your answer.

Explain the term *multicollinearity*.

From the MINITAB output from fitting the multivariate regression model found in Figure 2 there are three t-tests and one F-test made. Explain the difference of these t-tests and F-test.

Problem 3 Lifetime of batteries

A producer of batteries wanted to investigate if battery lifetime is dependent on the two factors *material type* and *operating temperature*. Three material types, named type 1, 2 and 3, and three operating temperatures, Low (-10°C), Medium (20°C) and High (45°C), were investigated, and the measured response was the effective lifetime of a battery (hours, measured on a continuous scale). The experiment was conducted by randomly selecting 12 batteries of each material type, and then randomly allocating the batteries to each of the three temperature levels. In total 36 measurements were made.

A two-way analysis of variance model (ANOVA) with interaction was fitted to the data, and the results are given in Table 1.

Source	DF	SS	MS	F value	p-value
Material	2	10684	?	7.9114	0.001976
Temperature	2	39119	19559.4	28.9677	$1.909{\cdot}10^{-7}$
Material·Temperature	4	?	2403.4	3.5595	?
Error	27	18231	675.2		
Total	?	?			

Table 1: Result from two-way ANOVA with interaction on the batteries data.

a) What are the assumptions behind this analysis?

Five of the entries in Table 1 are replaced by question marks (?). Calculate numerical values for each of these, and explain what each of the values means.

Is there a significant effect of the interaction term Material·Temperature? Perform a hypothesis test to answer this question. Write down the null and alternative hypothesis. Use significance level $\alpha = 0.05$.

Explain important features about the results from this two-way ANOVA and how you would proceed with further analyses?

The producer was interested in comparing the battery lifetime for the operating temperatures Medium and High for material type 3.

There were $n_{Medium} = 4$ observations for operating temperature Medium for material type 3. The average observed battery lifetime was $\bar{x}_{Medium} = 145.75$ and the empirical standard deviation was $s_{Medium} = 22.54$. Further, there were $n_{High} = 4$ observations for operating temperature High for material type 3. The average observed battery lifetime was $\bar{x}_{High} = 85.50$ and the empirical standard deviation was $s_{High} = 19.28$.

b) Perform an hypothesis test to investigate if the expected battery lifetime for the two operating temperatures Medium and High differ for material type 3. List the assumptions you need to make to perform the test.

Let μ_{Medium} be the expected battery lifetime for operating temperature Medium for material type 3. The producer was interested in the expected battery lifetime on the natural logarithmic scale, that is

$$\gamma = \ln(\mu_{Medium}).$$

c) Based on the independent random sample of size $n_{Medium} = 4$ from the operating temperature Medium for material type 3 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 summary statistics given in the text to calculate $\hat{\gamma}$ numerically and give estimated numerical values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.

Hint: You may use that $\frac{d}{dx}(\ln x) = \frac{1}{x}$.

Problem 4 Vaccine efficiency

A typhoid vaccine was tested, with the aim to control the efficiency of the vaccine. The test of efficiency is measured as the biological activity of the vaccine. For a period of 13 weeks, each week three lots of vaccine was tested for the efficiency.

Let X_{ij} be the measure of the vaccine efficiency for lot j, in week i, where $j = 1, 2, 3$ and $i = 1, 2, ..., 13$. Further, $\bar{X}_i{=}\frac{1}{3}\sum_{j=1}^{3} X_{ij}$, $S_i{=}\sqrt{\frac{1}{2}\sum_{j=1}^{3}(X_{ij} - \bar{X}_i)^2}$, $\bar{\bar{X}} = \frac{1}{13}\sum_{i=1}^{13}\bar{X}_i$, and $\bar{S} = \frac{1}{12}\sum_{i=1}^{13} S_i$.

Based on these 13 samples, assumed to be in control, we find $\bar{\bar{x}} = 1.012$ and $\bar{s} = 0.168$.

a) Construct a S-chart and a $\bar{X}$-S-chart (with 3σ limits).

A new sample is measured, with $\bar{x} = 0.93$ and $s = 0.65$. Is the process in control for this sample?