Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative solutions TMA4255 Applied Statistics 3 June 2016

Problem 1 Fish and parasites

a) Hypothesis:

H_0 : Being eaten is independent on the level of parasitic infection vs. H_1 : not so

We will use a χ^2-test for independence, where the test statistics approximately follows a χ^2-distribution with $(c-1)(r-1)$ degrees of freedom. Here $c = 3$ and $r = 2$, yielding $2 \cdot 1 = 2$ degrees of freedom.

Expected frequencies are calculated as (columns totals)·(row totals)/(grand total). The table of observed and expected values (e) are as follows:

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1 (17)	10(15.3)	37(15.7)	48
Not eaten	49(33)	35(29.7)	9(30.3)	93
Total	50	45	46	141

Showing how the *Eaten* and *Uninfected* expected value is calculated: $\frac{48 \times 50}{141} = 17$.

The contribution from this cell to the test statistic is $\frac{(1-17)^2}{17} = 15.06$. The test statistic consists of 6 terms, and is given as $X^2 = \frac{(1-17)^2}{17} + \frac{(10-15.3)^2}{15.3} + \frac{(37-15.7)^2}{15.7} + \frac{(49-33)^2}{33} + \frac{(35-29.7)^2}{29.7} + \frac{(9-30.3)^2}{30.3} = 69.8$.

The null hypothesis is rejected if the test statistics is larger than $\chi^2_{0.05,2} = 5.991$.

Conclusion: clearly we can reject the null hypothesis and we have reason to believe that the infection status is dependent on being eaten or not by a bird for fish under these experimental conditions.

Problem 2 Cigarettes

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

Both x_1 and x_2 are significant, x_3 are not significant. x_1 and x_2 has an significant effect on the CO content in these cigarettes. x_3 do not explain a significant portion of the variation in the response.

R^2 is high for x_1 and x_2, and very low for x_3. In addition, the estimated variance, S is low for x_1 and x_2, but higher for x_3. We observe that the sum of R^2 from the 3 different models are larger than 100%. The three models together explains more than 100% of the variability in the data. This is of course not possible, and there must be some common information in the three variables x_1, x_2 and x_3. This would mean that the three covariates x_1, x_2 and x_3 are correlated (from Figure 3 we see that x_1 and x_2 are highly correlated).

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.

Hypothesis for x_2:

$$H_0 : \beta_2 = 0 \text{ vs. } H_1 : \beta_2 \neq 0$$

The p-value is found to be 0.000 (technically, p-values cannot equal 0. MINITAB is automatic rounding off or truncate to a preset number of digits after the decimal point, more correct replacing "$p = 0.000$" with "$p < 0.0001$). Given that the truth is that $\beta_2 = 0$ there is a < 0.0001 probability to observe a test statistic T which is at least as extreme ($t_{obs} \leq -11.92$ or $t_{obs} \geq 11.92$) as what we have observed.

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

CI for β_2:

$$[\text{coefficient} \pm t_{\alpha/2,(n-2)}\text{SE(coefficient)}],$$

$t_{0.05,23} = 1.714$

$$[14.860 \pm 1.714 \cdot 1.247] = [12.722, 16.997]$$

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

The estimated variance in the regression model, $s^2 = SSE/(n-2)$ is an appropriate estimator for σ^2. The estimated standard deviation in the regression model (model error variance) is found in the MINITAB output to be $S = 1.58842$, an appropriate estimate for σ.

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

Write down the fitted regression model.

$$y = 1.3089 + 0.8918x_1 + 0.629x_2 \tag{1}$$

Predicted CO content when $x_1 = 10$ and $x_2 = 0.8$ is then

$$y = 1.3089 + 0.8918 \times 10 + 0.629x_2 \times 0.8 = 10.7301 \tag{2}$$

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.

Assumptions of the linear regression: linearity and ϵ_i i.i.d $\sim N(0, \sigma^2)$

- Linearity: looking at the scatter plots and correlation we see a linear trend and correlation in x_1 vs. y, x_2 vs. y. In the plot of the standardized residuals vs. fitted value we see no clear trend, and thus may assume that linearity in the parameters of the model may be an adequate assumption (although we do not have the plot for standard residuals vs. x_1 and x_2 to verify these). No clear trend (maybe an indication of a "fan" like trend, but difficult to say) in the standardized residuals vs. fitted value plot also indicates equal variance of errors (homoscadasticity).
- Covariates included in the model: x_1 and x_2 were significant when testing each of the covariates. All of the covariates seems to have a high correlation with y (except x_3, that has the lowest correlation with y) and also x_1 and x_2 have a high correlation with each other. They explain a lot of the same variance. We may try to refit the model with only one of x_1 or x_2. This is shown in Figure 1, where x_1 seems to explain more of the variation of the data ($R^2 = 93.3\%$, and smallest $S = 1.11865$) than x_2.
- Standardized residuals vs. order, seems to be quit independent, but some irregularities (although we do not know if this are the order which the observations where made).
- In an overall level the regression is found to explain more than just the average yield level (p -value for the regression is 0.000).

- Normality of errors: looking at the qq-plot for the standardized residuals the assumption of normality seems plausible (Anderson Darling test gives p-value of 0.127, keeping the null hypothesis of normality). Histogram seems OK.
- Explanatory powers: the model explains 93.4% of the variability of the data (R^2), which is a high number.

Conclusion: the model seem to be adequate.

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.

The explanation if collinearity. We see from Figure 3 that x_1 and x_2 are highly correlated with each other and also with y. This means that one can be linearly predicted from the others with a substantial degree of accuracy. The estimates of the coefficients in the multiple regression may change erratically in response to small changes in the model or the data. Highly correlated independent variables are not desirable as this may affect the statistical accuracy of individual covariates. A multiple regression model with correlated predictors can indicate how well the entire bundle of covariates predicts the response, but it may not give valid results about any individual covariate, or about which covariates are redundant with respect to others. From this we conclude that there is a strong relationship between the independent covariates, and that only one of these covariates should be used as a predictor of CO.

Explain the term *multicollinearity*.

If one covariate is correlated with another covariate then we have collinearity. (Not linearity - but a tendency of linear dependence.). With several correlated covariates we call this multicollinearity.

We have that $\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, $Cov(\mathbf{B}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. When we have multicollinearity $(\mathbf{X}^T\mathbf{X})^{-1}$ may have large diagonal elements. The covariance of $\mathbf{B}$ may be large since $\mathbf{X}^T\mathbf{X}$ may be nearly singular.

This will make it difficult to know which variable to include in the model (several variables give much of the same information). The estimate of β_1 in a model with only x_1 will change if x_2 is also included into the model. This will also make prediction difficult since the prediction error will increase rapidly.

Since we have correlation between each pair of the variables in the model, multicollinearity may be a problem. So X_2 was significant in the model in Equation (1), but as we added more variables/regressors in the regression model (Equation (4)), the X_2 variables changed as these variables/regressors are dependent on each other (correlated), $\mathbf{x}_j^T\mathbf{x}_i \neq 0$.

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.

The t-tests test if each of the regression covariates are significant,

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

explains significant amount of the variation in the response, given that the other covariates are in the model.

The F-test test if the regression is significant,

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs. } H_1 : \text{at least one} \neq 0$$

or if one or more of the regression covariates explains significant amount of the variation in the response.

In a simple linear regression, for a two-sided alternative, the two tests are identical, where $t^2 = f(1, \nu)$.

Problem 3 Lifetime of batteries

a) What are the assumptions behind this analysis?

Assumptions :
The two-way ANOVA with interaction can be written as

$$Y_{ij} = \mu + \alpha_i + \beta_i + \gamma_i + \epsilon_{ij}, \tag{3}$$

where ϵ_{ij} is i.i.d. $\mathcal{N}(0, \sigma^2)$, that is the error terms are independent and normally distributed with the same variance across treatment groups. We impose the restrictions $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{i=1}^{b} \beta_i = 0$ and $\sum_{i=1}^{a} \alpha\beta_{ij} = 0$, $\sum_{i=1}^{b} \alpha\beta_{ij} = 0$

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.

In the ANOVA output the sums of squares (SS) of the total variability in the data is decomposed into variability between the Material types (Material), between the operating temperatures (Temperature), and due to interaction between material and temperature (Material·

Temperature), SST=SS(Material)+ SS(Temperature) + SS(Material· Temperature) + SSE. The sum of squares give these partitioned numerical values. The degree of freedom (DF) is associated with each sum, reflecting the amount of information in the sum (- and technically associating the scaled sum with a χ^2-distribution with this number of degrees of freedom). The Mean squares (MS) are the Sum of squares (SS) divided by the degrees of freedom. The F-value is the ratio between the MS for the the given factor (Material, Temperature, Material· Temperature) and the MS for the Error/residuals. The p-value is related to the F-value and the F-distribution. The null hypotheses testes are wrt the parameters means, μ_i (or α_i) begin equal (or the same for the other parameters).

Missing entries:
df for Total: number of observations -1 = 36-1=35, or add all df above= 2+2+4+27=35.
MS for Material:SS for Material/ DF for Material=10684/2=5342.
SS for Material· Temperature: MS for Material· Temperature · DF for Material· Temperature=2403.4· 4=9613.6.
p-value for Material· Temperature: Tail in the Fisher distr with 4 and 27 df for F=3.5595. From the table for 0.05 this critical value is 2.73, which means that the p-value is below 0.05. For 0.025 table the critical value is 3.31, which means that the p-value is below 0.025. For 0.01 table the critical value is 4.11, which means that the p-value is above 0.01. Computer software (which is not available at the exam) would give 0.018611 as the p-value.
SS for Total: SS for Material + SS for Temperature + SS for Material· Temperature + SS for Error = 10684+ 39119 +9613.6 +18231=77647.6.

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 $\alpha = 0.05$.

Hypothesis:

H_0 : that the effects of interaction effects (Material·Temperature$_i$) are all equal to zero vs.

H_1 : at least one of the interaction effects(Material·Temperature$_i$) are not equal to zero

F-test for interaction: $\frac{SS(Material \cdot Temperature)/(a-1)(b-1)}{SS(Error)/ab(n-1)} = \frac{MS(Material \cdot Temperature)}{MS(Error)} = \frac{s^2_{Material \cdot Temperature}}{s^2} = \frac{2403.4}{675.2} = 3.5595$ (where $a = 3$, $b = 3$ and $n = 4$). We found the p-value to be below 0.05. So we can reject the null hypothesis, and we have a significant interaction between material type and operating temperature.

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

The interaction term was significant. This does not allow for broad conclusions on main

effects on Material and Temperature, as different operating temperature may give different results for the same Material type. We are not able to separate the main effects of Material and Temperature as there are significant interaction, and strict cautions about the interpretation of main effects in the presence of interaction. We could proceed to look at difference between different operating temperatures on the same material type, etc.

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.

$$H_0 : \mu_{Medium} - \mu_{High} = 0 \text{ vs. } H_1 : \text{not equal}$$

We perform a t-test to compare the two operating temperatures for material type 3. We borrow an estimate of σ^2 from the overall analysis, using $s^2 = 675.2$ with 27 df. $s = \sqrt{675.2} = 25.98461$. The t-test is based in the t-statistics

$$T = \frac{\bar{X}_{Medium} - \bar{X}_{High}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}s}$$

which we calculate to be $t_{obs} = \frac{145.75-85.50}{\sqrt{\frac{1}{4}+\frac{1}{4}}\cdot 25.98461} = 3.279109$

Two-sided t-test: reject H_0 when $t_{obs} > t_{\alpha/2,n_1+n_2-2}$ or when $t_{obs} < t_{1-\alpha/2,n_1+n_2-2}$. From the table we find that the critical values are $t_{0.025,6} = 2.447$ and $t_{0.975,6} = -2.447$. We have observed a value more extreme than the critical values and we reject the null hypothesis.

Conclusion: We have reason to believe that there is a difference in the operating temperatures Medium and High for Material type 3.

A t-test can be performed with the given mean and standard deviation in the text. We can use a pooled estimate of variance. $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, $S_p = \sqrt{\frac{(4-1)22.54^2+(4-1)19.28^2}{4+4-2}} = 20.97344$

$$T = \frac{\bar{X}_{Medium} - \bar{X}_{High}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}S_p} = \frac{145.75 - 85.5}{\sqrt{\frac{1}{4} + \frac{1}{4}}20.97} = 4.062584$$

We here also reject H_0 as above.

List the assumptions you need to make to perform the test.

Assumptions:
We assume that the data are normally distributed, that is, $X_{Medium,i} \sim N(\mu_{Medium}, \sigma^2)$ for Material type 3 and operating temperature Medium and $X_{High,j} \sim N(\mu_{High}, \sigma^2)$ for Material type 3 and operating temperature High, $i = 1, ..., n_{Medium}$ and $j = 1, ..., n_{High}$, and that the

two samples are independent.

$$H_0 : \mu_{Medium} - \mu_{High} = 0 \text{ vs. } H_1 : \text{not equal}$$

Assumed equal variance from the ANOVA (assuming data are normally distributed we may test the equality by performing an F-test).

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.

Let $\bar{X}_{Medium}$ be the mean of the Medium operating temperature sample. A natural estimator for γ is

$$\hat{\gamma} = ln(\bar{X}_{Medium}) = ln(145.75) = 4.981893$$

Use approximate methods to find the expected value and variance of this estimator, that is, E($\hat{\gamma}$) and Var($\hat{\gamma}$). Use the summary statistics given in the text to calculate $\hat{\gamma}$ numerically and give estimated numerical values for E($\hat{\gamma}$) and Var($\hat{\gamma}$).

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

We turn to first order Taylor approximations with

$$g(\bar{X}_{Medium}) = ln(\bar{X}_{Medium})$$
$$\frac{\partial g(\bar{X}_{Medium})}{\partial \bar{X}_{Medium}} = \frac{1}{\bar{X}_{Medium}}$$

where the random variable $\bar{X}_{Medium}$ has $E(\bar{X}_{Medium}) = \mu_{Medium}$ and $Var(\bar{X}_{Medium}) = \sigma^2/n_{Medium}$.

Define

$$g'(\mu_{Medium}) = \frac{\partial g(\bar{X}_{Medium})}{\partial \bar{X}_{Medium}} \mid_{\bar{X}_{Medium}=\mu_{Medium}} = \frac{1}{\mu_{Medium}}$$

The first order Taylor approximation for one sample:

$$E(g(\bar{X}_{Medium})) \approx g(\mu_{Medium}) = ln(\mu_{Medium})$$
$$Var(g(\bar{X}_{Medium})) \approx g'(\mu_{Medium})^2 Var(\bar{X}_{Medium}) = (\frac{1}{\mu_{Medium}})^2 \sigma^2/n_{Medium}$$

Estimates using numerical values $n_{Medium} = 4$, $\mu_{Medium} = \bar{X}_{Medium} = 145.75$ and $\hat{\sigma}^2 = s^2_{Medium} = 22.54^2$ are as follows

$$\hat{E}(g(\bar{X}_{Medium})) \approx ln(\mu_{Medium}) = ln(145.75) = 4.981893$$
$$\hat{Var}(g(\bar{X}_{Medium})) \approx (\frac{1}{145.75})^2 22.54^2/4 = 0.00597903$$

Problem 4 **Vaccine efficiency** Construct a S-chart and a $\bar{X}$-S-chart (with 3σ limits).

S-chart has limits

$$\bar{S} \pm 3\frac{\bar{S}}{c_4}\sqrt{1-c_4^2}$$
$$[B_3\bar{S}, B_4\bar{S}]$$

According to table A22, and for rational subgroup size $n = 3$, we have $c_4 = 0.8862$, $B_3 = 0$ and $B_4 = 2.568$. Thus, the S-chart has lower limit equal to 0 and upper limit equal to $2.568 \cdot 0.168 = 0.431424$.

$\bar{X}$-S-chart has limits

$$\bar{\bar{X}} \pm 3\frac{\bar{S}}{c_4\sqrt{n}} = \bar{\bar{X}} \pm A_3\bar{S}$$

According to table A22, and for rational subgroup size $n = 3$, we have $c_4 = 0.8862$ and $A_3 = 1.954$. Thus, the chart has lower limit equal to $1.012 - 1.954 \cdot 0.168 = 0.683728$ and upper limit equal to $1.012 + 1.954 \cdot 0.168 = 1.340272$

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

The new samples are within the control limits for the $\bar{X}$-S-chart, but the new samples are outside the control limits for the S-chart (variance is not in control). The process variability is not in control.