UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4900 — Statistical methods and applications.

Day of examination: Wednesday June 4th 2014.

Examination hours: 14.30 – 18.30.

This problem set consists of 5 pages.

Appendices: Tables for normal, t-, χ^2- and F-distributions

Permitted aids: All printed and hand-written resources. Approved calculator.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

Measurements of ozone, temperature and wind speed has been taken on $n = 111$ occasions. In this problem we study how the level of ozone depends on temperature and wind speed using linear regression.
Specifically the variables we consider are

- `Ozone`: Mean ozone in parts per billion over a two hour period
- `Wind`: Average wind speed in miles per hour
- `Temp`: Maximum daily temperature in degrees Fahrenheit

a) In a first analysis we consider a simple linear regression with only temperature as covariate. Edited results from the analysis in `R` are given below.

Determine the correlation coefficient between ozone and temperature.

Carry out a test of whether there is an association between ozone and temperature.

```
> summary(lm(Ozone~Temp))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -147.6461    18.7553  -7.872 2.76e-12
Temp           2.4391     0.2393
---
Residual standard error: 23.92 on 109 degrees of freedom
Multiple R-squared: 0.488,      Adjusted R-squared: 0.4833
```

(Continued on page 2.)

b) In a second linear regression we also include wind as a covariate. Results are given below.

Discuss the concept of R-squared. Does including wind improve on the model?

The estimate for temperature has changed from the simple linear regression. Give an explanation for this change.

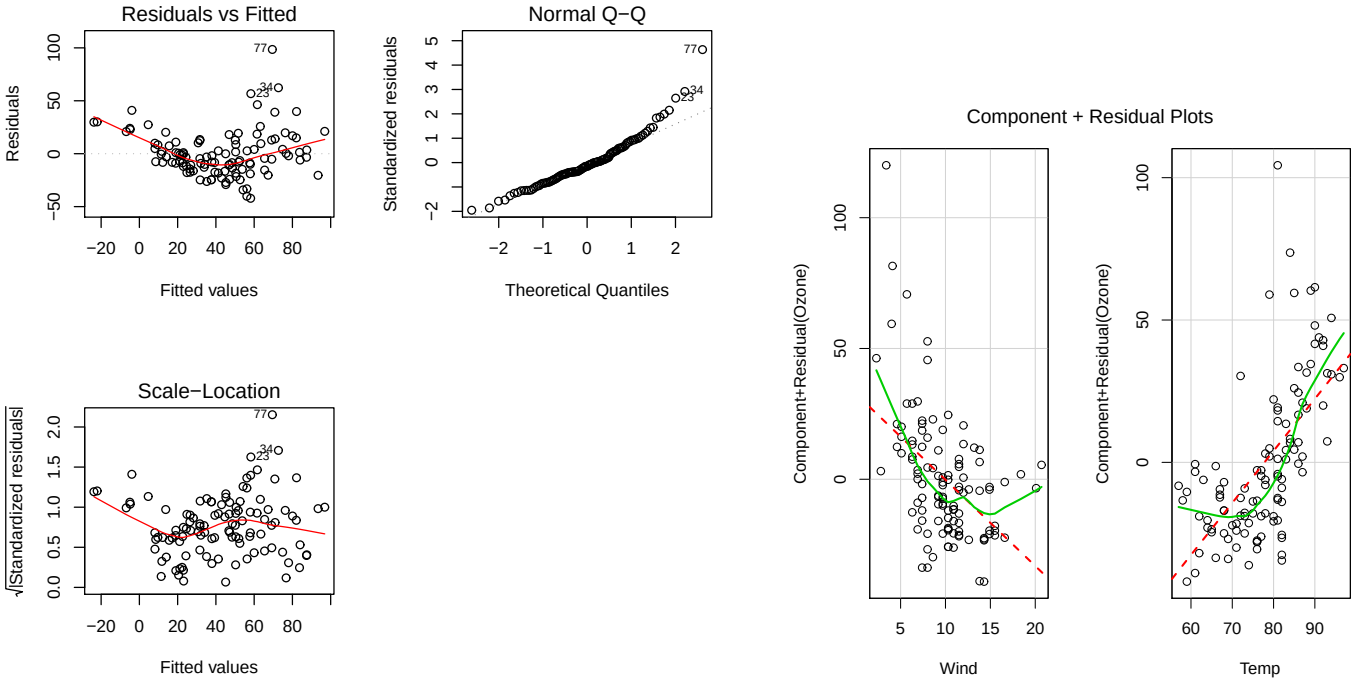
```
> summary(lm(Ozone~Wind+Temp))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -67.3220    23.6210  -2.850  0.00524 **
Wind         -3.2948     0.6711  -4.909 3.26e-06 ***
Temp          1.8276     0.2506   7.294 5.29e-11 ***
---
Residual standard error: 21.73 on 108 degrees of freedom
Multiple R-squared: 0.5814,     Adjusted R-squared: 0.5736
F-statistic: 74.99 on 2 and 108 DF,  p-value: < 2.2e-16
```

c) State explicitly the model assumptions for the linear regression in question b).

Use the plots below to discuss whether the assumptions seems to be fulfilled. Explain what each plot tells you.

Can you suggest possible improvements to the model.



(Continued on page 3.)

d) So far in this problem we have omitted the information that the ozone measurements were taken over a four month period and on the same location. Which part of the assumptions in question c) may then be violated.

Discuss what consequences such a violation of model assumptions may have. (You can disregard eventual model deviations pointed out in question c). We will in such case assume that the model has been suitably amended).

Problem 2

In a study it was registered whether lizards were staying in the sun or in the shade according to time of day and species of lizards. Time of day was divided into three groups "Early", "Midday" and "Late". We consider two species of lizards, "Grahami" and "Opalinus". Data grouped according to the different combinations of time and species are given below. The columns "Sun" and "Shade" contains the number of lizards that were found either in the sun or in the shade.

Time	Species	Sun	Shade
Early	Grahami	47	104
Early	Opalinus	5	32
Midday	Grahami	20	205
Midday	Opalinus	3	61
Late	Grahami	22	43
Late	Opalinus	8	25

In the analyses shown on the next page we use these grouped data with `Time` and `Species` as factor variables. The levels of `Time` are `Early, Midday` and `Late` with `Early` as reference. The levels of `Species` are `Grahami` (reference) and `Opalinus`. The purpose of the analyses is to study when and which lizards tend to prefer the sun to the shade.

a) Discuss why it is appropriate to analyze such data by logistic regression.

State and explain the logistic regression model used for the output on the top of the next page.

In particular discuss the concept of an odds-ratio. Sometimes an odds-ratio can be interpreted as an approximation to a relative risk. Do you think this is the case for these data?

(Continued on page 4.)

```
> summary(glm(cbind(Sun,Shade)~Time+Species,family=binomial))

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.8349     0.1697  -4.920 8.64e-07 ***
TimeLate         0.2429     0.2786   0.872   0.3833
TimeMidday      -1.4841     0.2728  -5.440 5.34e-08 ***
SpeciesOpalinus -0.7480     0.3037  -2.463   0.0138 *
```

b) Calculate and interpret the odds-ratios for time Midday versus time Early, time Late versus time Early and for Opalinus versus Grahami lizards.

Furthermore calculate a 95% confidence interval for the odds-ratio for Opalinus versus Grahami. Is there a statistically significant difference in the tendency to stay in the sun between the species?

c) Explain how inference for logistic regression models can be done using deviances.

Carry out a deviance test for whether there are significant difference between the different groups of time of the day using the output below.

```
> anova(glm(cbind(Sun,Shade)~Time+Species,family=binomial))
Analysis of Deviance Table

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev
NULL                        5     51.091
Time     2   43.624         3      7.467
Species  1    6.734         2      0.733
```

d) How can a model with interaction between species and time of the day be written up?

Determine whether an interaction is present for the current data.

Hint: What is the saturated model for these grouped data?

Problem 3

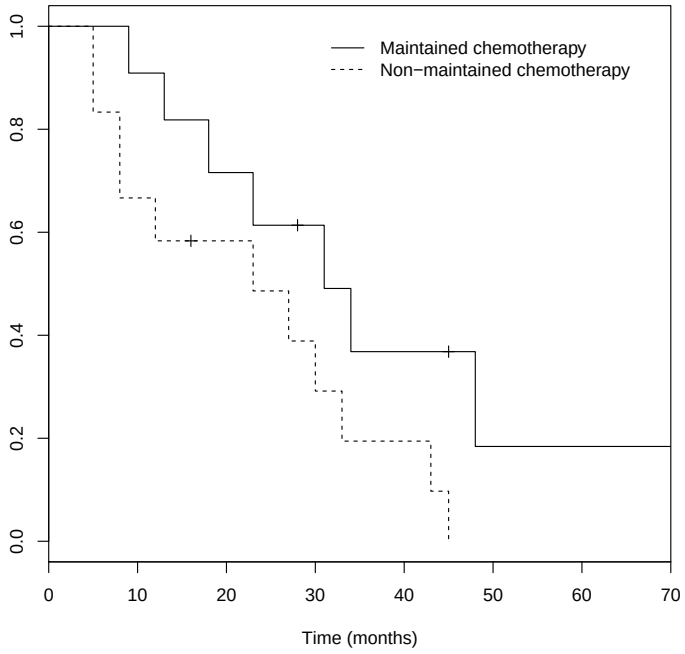
23 patients with acute myeloid leukemia were included in a study. They had first received chemotherapy and were then followed to time of remission, that is until they were free of cancer. After this they were randomized into two groups consisting of 11 patients given maintained chemotherapy and 12

(Continued on page 5.)

patients who were not given any further chemotherapy treatment. After this they were followed until time of relapse, that is reoccurrence of cancer or to a censoring time.

a) The Figure below shows the Kaplan-Meier estimators in the maintained chemotherapy and non-maintained chemotherapy groups for time until relapse. Explain what the Kaplan-Meier estimators tells about time to reoccurrence of cancer.

Find a (rough) estimate of the median time to relapse in the two groups.



b) Beneath you see the (edited) results from a Cox-regression for the same data. Here the `time` is time until relapse or censoring, `status` indicates whether it is a relapse, `status=1`, or censoring, `status=0`, and the covariate `x` equals 0 for the mainained chemotherapy group and 1 for those not receiving further chemotherapy.

Explain the model for the output. In particular interpret the value for `exp(coef)`.

Calculate a 95% confidence interval this quantity. Is there a significant difference between the two groups?

```
> coxph(Surv(time,status)~x,data=aml)

       coef exp(coef) se(coef)
group 0.916       2.5    0.512
```

END