UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4900 — Statistical methods and applications.

Day of examination: Tuesday 7. June 2016.

Examination hours: 9.00 – 13.00.

This problem set consists of 6 pages.

Appendices: Tabels 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.

Exercise 1.

A famous data set on various mammals contains information about $n = 42$ species of animals, their body weight in kg (BodyWt), brain weight in g (BrainWt), total sleep in hrs/day (TotalSleep), and an overall danger index (1-5) (Danger) where

1 = least danger (from other animals)

5 = most danger (from other animals)

The data set contains other variables as well, but they will not be used in this simplified analysis. We will study how measures of weight and the danger index can be used as explanatory variables for the total number of hours a species typically sleeps per day.

a) By just looking to Figure 1(a) and 1(b), where TotalSleep is plotted against BrainWt and log(BrainWt), explain briefly why it is useful to log-transform the brain weight data. The log here means the natural logarithm. Then, by looking to Figure 1(c), where log transformed weight variables (log(BrainWt) and log(BodyWt)) are plotted against each other, explain in short why including both these as explanatory variables in a regression analysis might be a bad idea.

b) We first perform a univariate linear regression with TotalSleep as response variable and log(BrainWt) as covariate. The (edited) result of the analysis in R is found in (I) below. Carry out a hypothesis test in order to check if there is a significant linear association between TotalSleep and log(BrainWt). Please specify the hypotheses, test statistic, distribution and conclusion.

(Continued on page 2.)

c) If the brain weight increases with 100 g, how does the expected amount of total sleep change? How much of the variation in TotalSleep can be explained by the linear relation with log(BrainWt)?

d) We then perform an analysis where the danger index Danger is used as explanatory variable for TotalSleep. Explain why it is preferable to specify Danger as a factor, and write down the model in this situation. Use "treatment-contrast", so that your model corresponds to the automatic R choice, which resulted in (II) below.

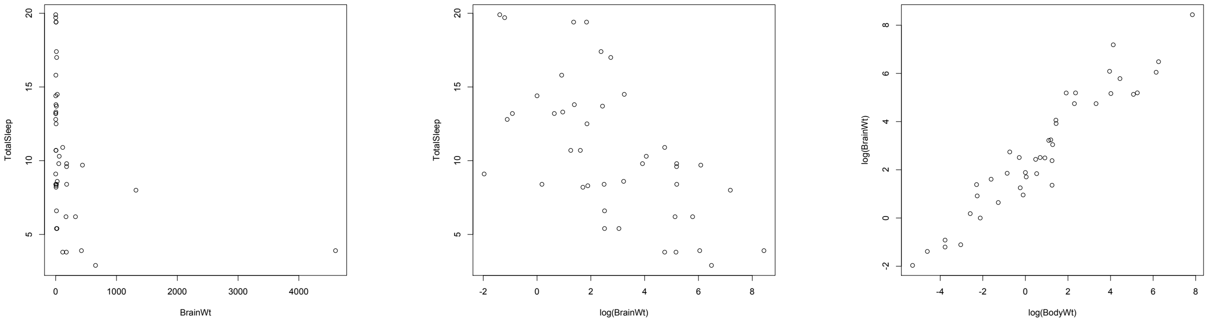
e) Interpret the fitted model that is given in (II). Especially, explain what we can say about expected amount of sleep for species that are in least and most danger from other animals.

f) In (III) below, both log(brainWt) and Danger have been used as covariates. Discuss whether this is a better model than the model used in (II). Specify the underlying assumptions in such a multiple regression model, and evaluate the validity of the assumptions and the fit in light of the figures in Figure 2.

```
(I) Call: lm(formula = TotalSleep ~ log(BrainWt), data = sleep)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.8973     0.8664  16.041  < 2e-16
log(BrainWt)  -1.1688     0.2325  -5.028 1.08e-05

Residual standard error: 3.732 on 40 degrees of freedom
Multiple R-squared:  0.3873,Adjusted R-squared:  0.3719
F-statistic: 25.28 on 1 and 40 DF,  p-value: 1.083e-05

(II) Call: lm(formula = TotalSleep ~ factor(Danger), data = sleep)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.855      1.173  11.808 4.11e-14
factor(Danger)2   -1.815      1.700  -1.067  0.29280
factor(Danger)3   -3.555      1.882  -1.889  0.06672
factor(Danger)4   -5.043      1.749  -2.883  0.00652
factor(Danger)5   -9.295      2.099  -4.428 8.12e-05

Residual standard error: 3.891 on 37 degrees of freedom
Multiple R-squared:  0.3838,Adjusted R-squared:  0.3172
F-statistic: 5.762 on 4 and 37 DF,  p-value: 0.001043
```

(Continued on page 3.)

```
(III) Call: lm(formula = TotalSleep ~ log(BrainWt) + factor(Danger), data = sleep)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      16.6548     1.0993  15.150  < 2e-16
log(BrainWt)     -1.0480     0.2192  -4.781 2.93e-05
factor(Danger)2  -2.4324     1.3543  -1.796 0.080887
factor(Danger)3  -5.2402     1.5329  -3.419 0.001579
factor(Danger)4  -3.8441     1.4093  -2.728 0.009799
factor(Danger)5  -6.8684     1.7398  -3.948 0.000351

Residual standard error: 3.085 on 36 degrees of freedom
Multiple R-squared:  0.6231,Adjusted R-squared:  0.5708
F-statistic:  11.9 on 5 and 36 DF,  p-value: 7.714e-07
```

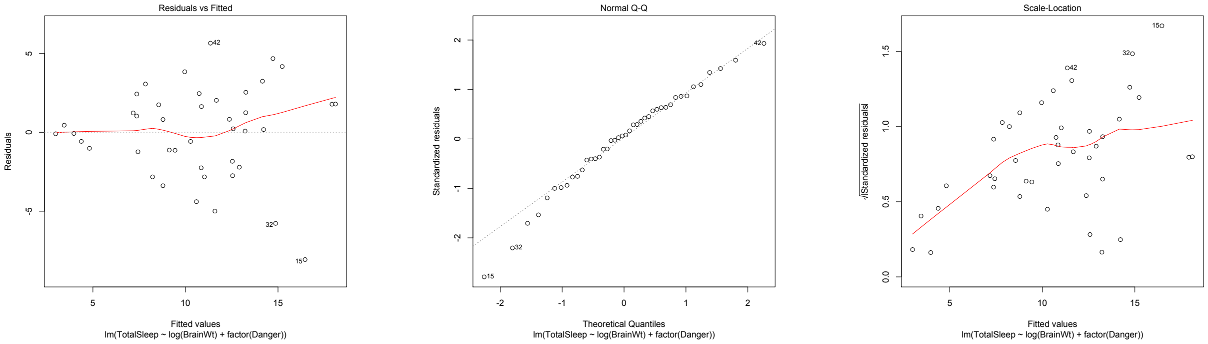


(a) TotalSleep vs. BrainWt

(b) TotalSleep vs. log(BrainWt)

(c) log(BrainWt)vs. log(BodyWt)

Figure 1



(a) Residuals vs Fitted

(b) Normal Q-Q

(c) Scale-Location

Figure 2

(Continued on page 4.)

Exercise 2.

For $n = 32$ randomly chosen cars in the US, we know for each of them if they are equipped with an automatic or manual transmitter, their weight (in 1000 lbs) and their horsepower. The variable indicating type of transmitter is called am, and am=1 if the transmitter is manual and 0 if it is automatic. The weight is called wt and the horsepower hp in the (edited) R analyses in the following. We would like to model the probability of a car having a manual transmitter, given a certain horsepower and weight.

a) In the fitted model (I) below, only weight (wt) is used as a covariate. In the fitted model (II) below, also horsepower (hp) is included. Use the deviance test to test if this additional variable represents a significant improvement of the model. Use significance level 0.05.

b) Based on model (II) below, find the estimated probability that a car weighting 2800 lbs and with an engine of 120 hp comes with a manual transmitter. Explain in short the concept of an odds ratio. Find the odds ratio for an increase of 10 units in horsepower, keeping the weight fixed.

```
(I) Call: glm(formula = am ~ wt, family = binomial, data = mtcars)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   12.040      4.510   2.670  0.00759
wt            -4.024      1.436  -2.801  0.00509

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 19.176  on 30  degrees of freedom
AIC: 23.176

Number of Fisher Scoring iterations: 6

(II) Call: glm(formula = am ~ hp + wt, family = binomial, data = mtcars)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.86630    7.44356   2.535  0.01126
hp           0.03626    0.01773   2.044  0.04091
wt          -8.08348    3.06868  -2.634  0.00843

     Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 10.059  on 29  degrees of freedom
AIC: 16.059

Number of Fisher Scoring iterations: 8
```

(Continued on page 5.)

Exercise 3.

Doctors have followed 42 leukaemia patients for a certain period. Half of them were randomly assigned to be treated with the drug 6-mercaptopurine (6MP) and the rest are controls, who received placebo treatment (the variable **treatment**, 1=6MP, 0=placebo). We are supposed to study the effect of treatment with 6MP, also adjusting for gender (the variable **female**, 1=female, 0=male) and white blood cells (log-transformed, the variabel **logWBC**). When entering the study, patients are in remission (no active leukaemia), and the time registered is relapse free survival in weeks (the variable **time**), possibly right censored (the variable **event**, 1=observed, 0=censored).

a) We first compare two Kaplan-Meier curves, for the group that received 6MP and the control group, respectively. The plot is in Figure 3. Find the median survival in both groups from the plot. The R function survdiff performs a so called logrank test. Explain what a logrank test is testing, and use the result below to write a conclusion.

```
Call:
survdiff(formula = Surv(time, event) ~ treatment)

             N Observed Expected (O-E)^2/E (O-E)^2/V
treatment=0 21       21     10.7      9.77      16.8
treatment=1 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

b) In order to further study the effect of treatment, we would like to use a Cox regression model where we can adjust for gender and white blood cells counts. Specify the model we use for the hazard rate in this case, and use the results from R below to find estimates for the regression coefficients in your model.

c) Find the hazard ratio for treatment versus placebo, and its 95% confidence interval. Interpret this hazard ratio and write a conclusion regarding the effect of 6MP.

(Continued on page 6.)

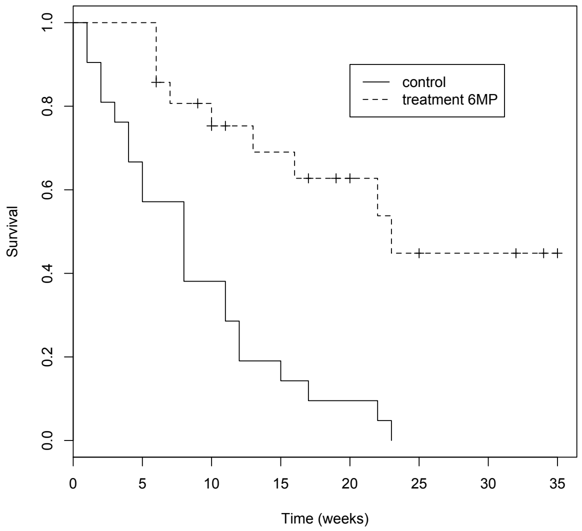


Figure 3

```
Call:
coxph(formula = Surv(time, event) ~ treatment + female + logWBC,
    data = remission)

  n= 42, number of events= 30

             coef exp(coef) se(coef)      z Pr(>|z|)
treatment -1.5036    0.2223   0.4615 -3.258  0.00112 **
female     0.3147    1.3698   0.4545  0.692  0.48872
logWBC     1.6819    5.3760   0.3366  4.997 5.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
treatment    0.2223      4.498   0.08998    0.5493
female       1.3698      0.730   0.56206    3.3384
logWBC       5.3760      0.186   2.77944   10.3982

Concordance= 0.851  (se = 0.062 )
Rsquare= 0.675   (max possible= 0.988 )
Likelihood ratio test= 47.19  on 3 df,   p=3.171e-10
Wald test            = 33.54  on 3 df,   p=2.475e-07
Score (logrank) test = 48.01  on 3 df,   p=2.114e-10
```

The end