UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK9900 — Statistical methods and applications.

Day of examination: Tuesday, 12 June, 2012.

Examination hours: 14.30 – 18.30.

This problem set consists of 6 pages.

Appendices: Tables for the standard normal distribution, the chi-square distributions, the t distributions, and the 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

In this problem we want to study how the yield of wheat depends on moisture in the ground and the variety (or type) of wheat. Each of ten varieties of wheat was planted once on each of six plots, giving a total of 60 observations.

The data set contains the variables:

yield	yield of wheat in bushels per acre
moist	moisture in the top 36 inches of soil before planting
variety	ten different types of wheat (labeled 1 to 10)
plot	six randomly chosen plots, each of size 1 acre (labeled 1 to 6)

Below is given an ANOVA-table with `yield` as the response variable and `variety` as a categorical covariate. Two numbers in the table are replaced by question marks.

Output 1:

```
anova(lm(yield~factor(variety),data=wheat))

                   Df   Sum Sq    Mean Sq    F value     Pr(>F)
factor(variety)     ?   4089.1     454.34        ?       0.00013
Residuals          50   4756.3      95.13

(edited output)
```

a) Describe the model that the ANOVA-table is based on. Are the varieties significantly different? Give the two numbers in the table that are replaced by question marks and explain how they are determined.

(Continued on page 2.)

In a second model we have also included the covariate `moist`. An edited version of this regression analysis is presented below.

Output 2:

```
summary(lm(yield~moist+factor(variety),data=wheat))

                       Estimate    Std. Error     t value       Pr(>|t|)
(Intercept)           31.995733      0.496150      64.488        < 2e-16
moist                  0.670836      0.008473      79.173        < 2e-16
factor(variety)2      -2.884687      0.515474      -5.596       9.74e-07
factor(variety)3       3.183343      0.502109       6.340       6.99e-08
factor(variety)4       1.284064      0.516489       2.486       0.016372
....
factor(variety)10      3.074284      0.506524       6.069       1.83e-07

Residual standard error: 0.8677 on 49 degrees of freedom
Multiple R-squared: 0.9958,     Adjusted R-squared: 0.995
F-statistic:  1170 on 10 and 49 DF,  p-value: < 2.2e-16

(edited output)
```

b) State the model for this analysis, and give an interpretation of the estimated coefficients.

c) Discuss the concept of R^2 and explain how this measure has been computed for the model in question b). Also calculate R^2 for the model in question a). Compare the two models with respect to their ability to predict the yield of wheat.

Problem 2

In the period from 1999 to 2001 a number of cod along the coast of Finmark in North Norway were examined for infection by the blood parasite *Trypanosoma murmanensis.* In this problem we will study how the risk of infection depends on some covariates.

The response variable is `parasite`, which is coded as 0 if a fish is not infected by the blood parasite and as 1 if a fish is infected. We will relate this response to the following covariates:

`year`	the year when the fish is caught (1: year 1999; 2: year 2000; 3: year 2001)
`weight`	weight of the fish (in kg)
`age`	age of the fish (in years)

Throughout the problem `year` is treated as a categorical covariate (with three levels).

(Continued on page 3.)

Further we center the numeric covariates weight and age by subtracting their means (the mean weight is 1.75 kg and the mean age is 4.4 years).

a) Explain why logistic regression is an appropriate model for analysing the data. Give an explicit formulation of the logistic regression model when year is the only covariate. (Remember that we treat year as a categorical covariate.)

When we fit the logistic regression model with year as the only covariate, we get the result:

Output 3:

```
Call:
glm(formula = parasite ~ factor(year), family = binomial)

              Estimate   Std. Error    z value     Pr(>|z|)
(Intercept)   -0.99119      0.19516     -5.079     3.80e-07
factor(year)2  1.28666      0.30429      4.228     2.35e-05
factor(year)3  0.07936      0.26313      0.302       0.763

    Null deviance: 467.82  on 364  degrees of freedom
Residual deviance: 445.80  on 362  degrees of freedom

(edited output)
```

b) Describe how one may ascertain that there is a significant effect of year. Discuss how the probability that a fish is infected depends on the year it is caught.

Next we fit a model with the covariates year and weight:

Output 4:

```
Call:
glm(formula = parasite ~ factor(year) + I(weight-1.75), family = binomial)

                  Estimate   Std. Error    z value     Pr(>|z|)
(Intercept)       -1.06287      0.19920     -5.336     9.52e-08
factor(year)2      1.40229      0.31186      4.497     6.91e-06
factor(year)3      0.15834      0.26622      0.595       0.552
I(weight-1.75)    -0.22404      0.09995     -2.241       0.025

(edited output)
```

c) Define the odds ratio corresponding to 1 kg increase in the weight of a fish. Estimate the odds ratio and find a 95% confidence interval for it. Describe what the estimated odds ratio and the confidence interval tell you.

Then we fit a model with all the three covariates :

(Continued on page 4.)

Output 5:

```
Call:
glm(formula=parasite~factor(year)+I(weight-1.75)+I(age-4.4), family=binomial)

                Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)      -1.1409      0.2057    -5.546     2.92e-08
factor(year)2     1.3709      0.3190     4.297     1.73e-05
factor(year)3     0.2198      0.2707     0.812     0.416923
I(weight-1.75)   -0.7821      0.1984    -3.943     8.06e-05
I(age-4.4)        0.4800      0.1332     3.604     0.000313

    Null deviance: 467.82  on 364  degrees of freedom
Residual deviance: 426.46  on 360  degrees of freedom

(edited output)
```

d) The estimated effect of weight differs for the models in output 4 and output 5. Explain the reason for this, and interpret the effect of weight for the model in output 5. Also give an interpretation of the intercept for the model in output 5. (Remember that the two numeric covariates are centered.)

Problem 3

784 individuals aged 65-75 years participated in a blood pressure study in Bergen in the period 1966-1971. The individuals were followed until they died, emigrated or to the end of the study.

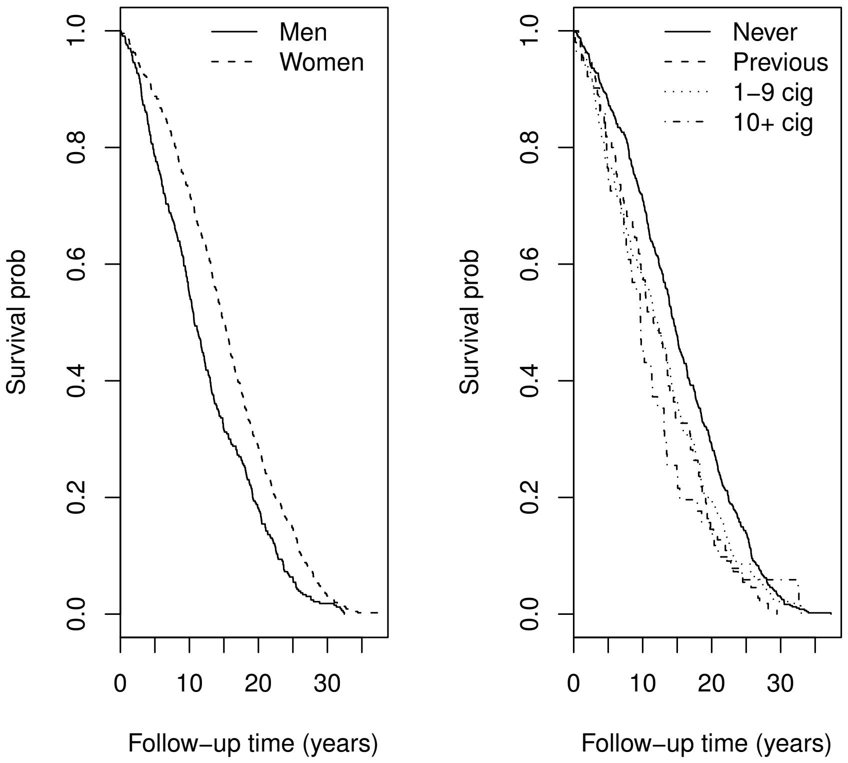
In our analyses we will use the covariates smoking and sex, in addition to the censored survival times and the indicators of death or censoring:

`time`	follow-up time from start of study to death or censoring
`death`	indicator of death/censoring (0: censoring; 1: death)
`smoking`	categorical smoking variable (1: never smoker; 2: previous smoker; 3: 1-9 cigarettes per day; 4: 10 or more cigarettes per day)
`sex`	sex (1: males; 2: females)

a) Discuss the concept of right censored survival data and explain why methods like linear or logistic regression are inappropriate for analyzing such data.

b) The plots on the next page show Kaplan-Meier estimates of the survival function according to sex and smoking categories. Discuss the observed differences in mortality based on these plots.

(Continued on page 5.)



Since we are dealing with two categorical covariates it is useful to model the mortality by regression methods, and the most common regression method for censored survival data is Cox's regression model.

c) Give a description of Cox's regression model. In particular describe how we may interpret the regression coefficients as logarithms of hazard ratios.

   Interpret the hazard ratios from the Cox-regression with sex and smoking category from the edited R-output below.

   Output 6:

```
                      coef  exp(coef)  se(coef)      z       p
sex                 -0.301       0.74    0.0865  -3.48  0.0005
factor(smoking)2     0.162       1.18    0.1201   1.35  0.1800
factor(smoking)3     0.127       1.14    0.1043   1.22  0.2200
factor(smoking)4     0.232       1.26    0.1536   1.51  0.1300
```

(Continued on page 6.)

Problem 4

Investigators have followed the growth of 27 children (16 boys, 11 girls) from age 8 years until age 14 years. Every two years (i.e. at ages 8, 10 12, and 14 years) they measured a certain distance of the skull of the children (the distance between the pituitary and the pterygomaxillary fissure). A main aim was to study the growth of the children and to assess if there is a difference between boys and girls.

The response variable is `distance` (measured in mm), and we want to relate this to the covariates:

`age`	age of the child (in years)
`sex`	sex of the child (0: boy; 1: girl)

In addition the data file contains the variable `subject`, which identifies which child a measurement belongs to.

a) We will adopt a random effects model with random effect for `subject` and with `age` and `sex` as fixed covariates. Describe the assumptions underlying this model, and discuss why the model is more appropriate than an ordinary linear regression model.

A fit of the random effects model gives the following results:

Output 8:

```
Random effects:
 Formula: ~1 | subject
          (Intercept)   Residual
StdDev:      1.807425   1.431592

Fixed effects: distance ~ age + sex
                  Value   Std.Error   DF     t-value   p-value
(Intercept)   17.706713   0.8339225   80   21.233044    0.0000
age            0.660185   0.0616059   80   10.716263    0.0000
sex           -2.321023   0.7614168   25   -3.048294    0.0054

Number of Observations: 108
Number of Groups: 27

(edited output)
```

b) Describe what we may learn from the output concerning the effects of `age` and `sex`. Discuss the interpretation of the estimated random effects.

END