UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4900 — Statistical methods and applications.

Day of examination: Thursday 6. June 2013.

Examination hours: 14.30 – 18.30.

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. Several women have problems with osteoporosis after menopause. Researchers at the National Hospital (Rikshospitalet) are studying how genetic factors contribute to this condition. We will study a dataset in which bone density is measured for 84 women after menopause. For the same women the researchers have measured gene expression (a measure of how active a gene is) in bone cells for 20,000 genes. We will not analyze such high-dimensional data sets here, but restrict ourselves to study how the gene expression of four selected genes effect the bone density of women after menopause.
In the output below the variable bone density is called 'bone' and the four genes 'gene1', 'gene2', 'gene3' and 'gene4', respectively.

a) We first perform a univariate analysis with gene 1 alone as covariate and bone density as response. From the results (I) below, perform a hypothesis test to demonstrate that gene 1 alone is a significant explanatory variable for bone density.

b) We then include gene 2 in a multiple regression analysis, and get the results (II) below. Explain shortly what happens to the effect of gene 1 when also gene 2 is included in the model, and why this can happen. You can use some additional information available below the results in (II).

c) Finally, we also include gene 3 and gene 4 in the model. Two alternative fits (III) are found below. Choose the best model, and explain the reasons for your choice. Interpret the estimated effects of the various gene expressions on bone density in your chosen model.

```
(I)
Call:
lm(formula = bone ~ gene1)
```

(Continued on page 2.)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9309     1.5477    2.54  0.01298
gene1        -1.1297     0.3609

(edited output)
```

```
(II)
Call:
lm(formula = bone ~ gene1 + gene2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8276     2.0564   3.806 0.000273
gene1        -0.1102     0.5082  -0.217 0.828876
gene2        -1.9110     0.6954  -2.748 0.007387

(edited output)

> cor(gene1, gene2)
[1] 0.7300616
```

```
(III)
Call:
lm(formula = bone ~ gene1 + gene2 + gene3 + gene4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.1091     2.9757  -2.725 0.007912
gene1        -0.2141     0.3848  -0.556 0.579455
gene2        -1.8064     0.5182  -3.486 0.000804
gene3         2.4974     0.6097   4.096 0.000101
gene4         1.7865     0.2731   6.541 5.56e-09

Residual standard error: 1.088 on 79 degrees of freedom
Multiple R-squared: 0.5588,     Adjusted R-squared: 0.5365
F-statistic: 25.02 on 4 and 79 DF,  p-value: 2.114e-13

Call:
lm(formula = bone ~ gene2 + gene3 + gene4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.2321     2.9546  -2.786  0.00666
gene2        -2.0106     0.3643  -5.519 4.09e-07
gene3         2.5455     0.6009   4.236 6.05e-05
gene4         1.7658     0.2694   6.554 5.03e-09

Residual standard error: 1.083 on 80 degrees of freedom
Multiple R-squared: 0.5571,     Adjusted R-squared: 0.5405
F-statistic: 33.54 on 3 and 80 DF,  p-value: 3.863e-14
```

(Continued on page 3.)

Exercise 2. After surgery, patients might have a sore throat because of the equipment used to keep the respiratory tract open during anesthesia. We have a data set comprising 35 surgeries, where the occurrence of a sore throat after anesthesia has been registered. The variable 'sore' is 1 if a sore throat occurred, and 0 otherwise. Other variables registered for each surgery are 'duration', which is the duration of the surgery in minutes, and 'type', a factor with two levels indicating two different types of equipment used to keep the respiratory tract open. We are interested in how the risk of getting a sore throat after surgery depends on the duration of the surgery and on the type of equipment used.

a) Explain why a logistic regression model is suitable for modelling this data set.

b) Below you find the results of a logistic model fit using only the duration of the surgery as explanatory variable. What is the predicted probability of getting a sore throat for a surgery lasting 30 minutes?

```
Call:
glm(formula = sore ~ duration, family = binomial, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0964  -0.7392   0.3020   0.8711   1.3753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.21358    0.99874  -2.216  0.02667
duration     0.07038    0.02667   2.639  0.00831

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 33.651  on 33  degrees of freedom

Number of Fisher Scoring iterations: 5
```

c) For the model above with only duration as covariate, what is the odds ratio for a sore throat for surgeries lasting 40 minutes compared to surgeries lasting 30 minutes? Find a 95% confidence interval for this odds ratio. Explain what the ratio tells us.

d) Below you find the results from including also type of equipment in the model. Test in two different ways if this additional variable represents a significant improvement of the model. Use significance level 0.05.

```
Call:
glm(formula = sore ~ duration + factor(type), family = binomial,
    data = dat)
```

(Continued on page 4.)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3802  -0.5358   0.3047   0.7308   1.7821

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.41734    1.09457  -1.295  0.19536
duration       0.06868    0.02641   2.600  0.00931
factor(type)1 -1.65895    0.92285  -1.798  0.07224

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 30.138  on 32  degrees of freedom

Number of Fisher Scoring iterations: 5
```

Exercise 3. In order to study the incidents of damage to cargo ships caused by waves for different types of ships, constructed in different years and operating in different periods, a shipping company has collected the following data. For each of the observed 34 combinations of type of ship, year of construction and period of operation, we have information on five variables as follows:

ship type	types A, B, C, D and E
year of construction	periods 1960-64, 1965-69, 1970-74, and 1975-79
period of operation	periods 1960-74 and 1975-79
months in service	ranging from 45 to 44882 months
damage incidents	ranging from 0 to 58

Note that there were no ships of type E built in 1960-64, and that ships built in 1975-79 could not have operated in 1960-74. These combinations are not in the data file. The damage incident counts can be modelled as occurring at a Poisson rate given the values of the predictors. Months in service is the number of months ships of a certain type, built in a certain period and operated in a certain period, have been in service, and hence under exposure for risk of damage. We therefore use log(months in service) as offset.

a) Using only the year of construction (in four categories) as explanatory variable for the damage rate, we arrive at the results below. Find the rate ratio for the ships built in period 1970-74 compared to the ships built in the first period 1960-64. Explain how this rate ratio should be interpreted.

(Continued on page 5.)

```
Call:
glm(formula = damage ~ offset(log(months)) + factor(construction),
    family = poisson, data = shipdamage)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9924  -0.6915  -0.2859   0.7141   3.7998

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.8241     0.1195 -57.094  < 2e-16
factor(construction)1965-69   0.8115     0.1477   5.496 3.89e-08
factor(construction)1970-74   1.2016     0.1509   7.965 1.65e-15
factor(construction)1975-79   0.9679     0.2070   4.676 2.93e-06

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  73.225  on 30  degrees of freedom

Number of Fisher Scoring iterations: 4
```

b) In the last analysis, below, we have included type of ship and period operated as covariates. Write down the formula for this 'full' model. Use a deviance test to test if these new variables significantly improve the model. Remember to write down clearly the hypothesis you are testing.

```
Call:
glm(formula = damage ~ offset(log(months)) + factor(type) + factor(construction) +
    factor(operation), family = poisson, data = shipdamage)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6768  -0.8293  -0.4370   0.5058   2.7912

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.40590    0.21744 -29.460  < 2e-16
factor(type)B               -0.54334    0.17759  -3.060  0.00222
factor(type)C               -0.68740    0.32904  -2.089  0.03670
factor(type)D               -0.07596    0.29058  -0.261  0.79377
factor(type)E                0.32558    0.23588   1.380  0.16750
factor(construction)1965-69  0.69714    0.14964   4.659 3.18e-06
factor(construction)1970-74  0.81843    0.16977   4.821 1.43e-06
factor(construction)1975-79  0.45343    0.23317   1.945  0.05182
factor(operation)1975-79     0.38447    0.11827   3.251  0.00115

(Dispersion parameter for poisson family taken to be 1)
```

(Continued on page 6.)

```
    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.695  on 25  degrees of freedom

Number of Fisher Scoring iterations: 5
```

c) Discuss which ship type is the safest (A, B, C, D or E), that is, least prone to damages from waves, when the other covariates are unchanged.
Also, find the rate ratio for the last period of operation compared to the first period of operation, and write an interpretation of your findings in simple words.

Data problem 3:

```
   type construction operation months damage
1     A      1960-64   1960-74    127      0
2     A      1960-64   1975-79     63      0
3     A      1965-69   1960-74   1095      3
4     A      1965-69   1975-79   1095      4
5     A      1970-74   1960-74   1512      6
6     A      1970-74   1975-79   3353     18
7     A      1975-79   1975-79   2244     11
8     B      1960-64   1960-74  44882     39
9     B      1960-64   1975-79  17176     29
10    B      1965-69   1960-74  28609     58
11    B      1965-69   1975-79  20370     53
12    B      1970-74   1960-74   7064     12
13    B      1970-74   1975-79  13099     44
14    B      1975-79   1975-79   7117     18
15    C      1960-64   1960-74   1179      1
16    C      1960-64   1975-79    552      1
17    C      1965-69   1960-74    781      0
18    C      1965-69   1975-79    676      1
19    C      1970-74   1960-74    783      6
20    C      1970-74   1975-79   1948      2
21    C      1975-79   1975-79    274      1
22    D      1960-64   1960-74    251      0
23    D      1960-64   1975-79    105      0
24    D      1965-69   1960-74    288      0
25    D      1965-69   1975-79    192      0
26    D      1970-74   1960-74    349      2
27    D      1970-74   1975-79   1208     11
28    D      1975-79   1975-79   2051      4
29    E      1960-64   1960-74     45      0
30    E      1965-69   1960-74    789      7
31    E      1965-69   1975-79    437      7
32    E      1970-74   1960-74   1157      5
33    E      1970-74   1975-79   2161     12
34    E      1975-79   1975-79    542      1
```

The end