UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK9900 — Statistical methods and applications.

Day of examination: 9 June 2011.

Examination hours: 09.00 – 13.00.

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.

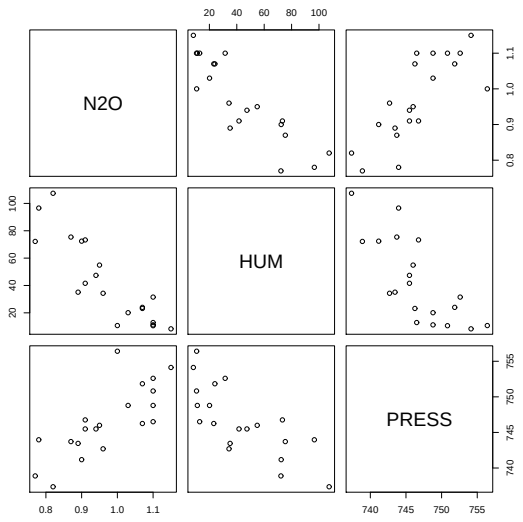
Problem 1

A study has been performed on a diesel-powered pickup truck to see if humidity and barometric pressure influence the emission of nitrous oxide (N_2O). A total of 20 emission measurements were taken on different times under varying conditions.

We will use linear regression to analyze the data. The outcome, `N2O`, is the emission of nitrous oxide (in ppm), while the predictors are:

`HUM`	Humidity (in percent)
`PRESS`	Barometric pressure (in mmHg)

The matrix scatter plot below gives an overview of the data.



(Continued on page 2.)

We start out by fitting simple linear regression models using only one predictor at a time. When we use pressure as the only predictor, we obtain the following results (the output has been edited):

Model 1

```
Call: lm(formula = N2O ~ PRESS)

             Estimate Std. Error
(Intercept) -12.558064   2.608668
PRESS         0.018122   0.003494

Residual standard error: 0.07506 on 18 degrees of freedom
Multiple R-squared: 0.5991,     Adjusted R-squared: 0.5768
```

a) Use an appropriate hypothesis test to decide if barometric pressure has a significant influence on the emission of nitrous oxide.

We then fit a model with humidity as the only predictor:

Model 2

```
Call: lm(formula = N2O ~ HUM)

             Estimate Std. Error
(Intercept)  1.1144267  0.0224412
HUM         -0.0033235  0.0004284

Residual standard error: 0.05688 on 18 degrees of freedom
Multiple R-squared: 0.7698,     Adjusted R-squared: 0.757
```

b) Consider the two simple linear regression models fitted above. Which of the two predictors, humidity or pressure, is by itself the best predictor for the emission of nitrous oxide? Give an argument for your answer.

We then fit a model with both predictors:

Model 3

```
Call: lm(formula = N2O ~ HUM + PRESS, data = emission)

             Estimate Std. Error
(Intercept) -3.4446014  2.9247707
HUM         -0.0025823  0.0006294
PRESS        0.0060639  0.0038901

Residual standard error: 0.05475
Multiple R-squared: 0.7986,     Adjusted R-squared: 0.7749
```

(Continued on page 3.)

c) Discuss why model 1 and model 3 give different estimates for the effect of pressure. Is there a significant effect of pressure when humidity is included in the model?

Problem 2

The table below gives data from a study of infant respiratory disease. The table shows the proportions of children developing a respiratory disease (bronchitis pneumonia) in their first year of life by sex and type of feeding. For example we see from the table that 458 boys got only bottle feeding, and among these 77 developed respiratory disease during their first year of life.

	Bottle only	Some breast with supplement	Breast only
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

We want to study if the sex of the child and the type of feeding influence the risk of developing respiratory disease.

a) Explain why logistic regression is an appropriate model for analyzing the data.

In order to use logistic regression, the data are given in a data frame with six lines, one for each combination of sex and type of feeding, and with the following variables:

`NODISEASE`	Number of children who develop respiratory disease
`NOTOT`	Total number of children
`SEX`	Sex (1: Boys, 2: Girls)
`FEEDING`	Type of feeding (1: Bottle only, 2: Some breast with supplement, 3: Breast only)

We first fit a model only using the categorical covariate type of feeding:

Model 4

```
Call: glm(formula = cbind(NODISEASE,NOTOT-NODISEASE)~factor(FEEDING),
                    family = binomial)

                  Estimate Std. Error
(Intercept)       -1.74676    0.09693
factor(FEEDING)2 -0.17435    0.20531
factor(FEEDING)3 -0.67645    0.15281

    Null deviance: 26.375  on 5  degrees of freedom
Residual deviance:  5.699  on 3  degrees of freedom
(edited output)
```

(Continued on page 4.)

b) Define the odds ratio for respiratory disease for a child who is breast fed relative to one who is fed by bottle. Estimate the odds ratio and derive a 95% confidence interval for it. Describe what the estimated odds ratio and the confidence interval tell you.

We then fit a model using both type of feeding and sex as categorical covariates:

Model 5

```
Call:
glm(formula=cbind(NODISEASE,NOTOT-NODISEASE)~factor(FEEDING)+factor(SEX),
            family = binomial)

                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.6127     0.1124 -14.347  < 2e-16
factor(FEEDING)2  -0.1725     0.2056  -0.839   0.4013
factor(FEEDING)3  -0.6693     0.1530  -4.374 1.22e-05
factor(SEX)2      -0.3126     0.1410

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
(edited output)
```

c) Use the results above to test in two different ways if sex has a significant effect on the risk of developing respiratory disease. What may you conclude from the tests?

d) Explain what we mean by interaction between type of feeding and sex. Use the results above to test the null hypothesis that there is no interaction between type of feeding and sex.

Problem 3

A number of studies have been performed to evaluate the influence of the magnitude and temporal pattern of low energy transfer radiation on biological systems. We will consider data from one such study. Here cultured human lymphocytes were exposed to gamma radiation (from a cesium-137 source) and the number of chromosomal abnormalities caused by the radiation was recorded. The experiment was performed with three doses of radiation (1.0, 2.5, and 5.0 Grays) and with nine dose rates. For each experiment the total number of lymphocytes were also recorded.

The results of the experiments are summarized in the table below. Here "cells" are the total number of lymphocytes in an experiment and y is the number of chromosomal abnormalities observed.

(Continued on page 5.)

Dose rate Gy/hr	Dose (Grays) 1.0 Cells	1.0 y	2.5 Cells	2.5 y	5.0 Cells	5.0 y
0.1	478	25	328	52	210	100
0.25	1907	102	185	51	138	113
0.5	2258	149	342	100	160	144
1.0	2329	160	310	100	120	106
1.5	1238	75	278	107	90	111
2.0	1491	100	259	107	100	132
2.5	1518	99	249	102	313	419
3.0	764	50	298	110	182	225
4.0	1367	100	243	107	144	206

a) Explain why it is reasonable to assume that the number of chromosomal abnormalities observed in an experiment (with a given dose and dose rate) is Poisson distributed.

The data have been analyzed using Poisson regression. To this end the data are given in a data frame with one line for each of the 27 experiments and with the columns corresponding to the following variables:

`cells`	number of lymphocytes
`ca`	number of chromosomal abnormalities observed
`dose`	total dose (Grays)
`doserate`	dose rate (Grays/hour)

First we fit a model with dose as a categorical covariate (factor) and the logarithm of the dose rate as a numerical covariate. This gives the results below (the output has been edited):

Model 6

```
Call:
glm(formula = ca~offset(log(cells))+factor(dose)+log(doserate),
              family = poisson)

                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.76958    0.03430  -80.74   <2e-16
factor(dose)2.5   1.65299    0.04857   34.03   <2e-16
factor(dose)5     2.80095    0.04251   65.89   <2e-16
log(doserate)     0.21447    0.01672   12.83   <2e-16

    Null deviance: 4753.004  on 26  degrees of freedom
Residual deviance:   42.776  on 23  degrees of freedom
```

(Continued on page 6.)

b) Give a mathematical formulation of the Poisson regression model that is fitted above. Use this formulation to define the rate ratio for dose 2.5 Grays relative to dose 1.0 Grays (when dose rate is kept constant), and estimate this rate ratio.

We then fit a model with interaction between dose and the logarithm of dose rate:

Model 7

```
Call:
glm(formula = ca ~ offset(log(cells)) + factor(dose) + log(doserate) +
    factor(dose):log(doserate), family = poisson, data = radiation)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.49101  -0.62473  -0.05078   0.76786   1.59115

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.74671    0.03426 -80.165  < 2e-16
factor(dose)2.5               1.62542    0.04946  32.863  < 2e-16
factor(dose)5                 2.76109    0.04349  63.491  < 2e-16
log(doserate)                 0.07178    0.03518   2.041 0.041299
factor(dose)2.5:log(doserate) 0.16122
factor(dose)5:log(doserate)   0.19350

    Null deviance: 4753.00
Residual deviance:   21.75
```

c) Explain what we mean by interaction between dose (as a categorical covariate) and the logarithm of dose rate (as a numerical covariate). Use the results above to test the null hypothesis that there is no interaction between dose and the logarithm of dose rate.