UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK4900/STK9900 — Statistical methods and applications.
Day of examination:	Friday May 29th - Monday June 8th 2020.
Examination hours:	Deadline – June 8th, 2:30PM

This problem set consists of 4 pages.

Appendices:	None
Permitted aids:	All printed, hand-written and internet based resources. Computer with R.

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

Problem 1

Data have been collected to check whether the presence of urea formaldehyde foam insulation (UFFI) has an effect on the formaldehyde (CH_2O) concentration inside a house. Twelve houses with and twelve houses without UFFI were studied, and the average weekly CH_2O concentration was measured. It was thought that the CH_2O concentration could also be influenced by the amount of air that can move through the house via windows, cracks, chimneys, etc. For this purpose, a measure of "air tightness" was determined for each house.

The variables in the data set are coded as follows:

- `CH2O`: Average weekly CH_2O concentration in parts per billion (note capital O in CH2O)
- AIR: Measure of air tightness on a scale from 0 to 10 (high values correspond to a tight house)
- UFFI : Indicator of formaldehyde foam insulation (0=absent, 1=present)

You read the data into R using commands

```
path="https://www.uio.no/studier/emner/matnat/math/STK4900/data/uffi.txt"
uffi=read.table(path,header=T)
```

(Continued on page 2.)

a) Do a t-test comparing the levels of CH_2O between the UFFI groups - assuming equal variance in both groups. Explain the model that the test is based on. State a conclusion to the test.

   Also run a simple linear regression with CH_2O as outcome and UFFI as explanatory variable. Relate the results to the t-test and explain the correspondence between the t-test and this simple linear regression.

b) Make a scatter plot of AIR versus CH_2O and calculate the (Pearson) correlation coefficient between these two variable.

   Also run a simple linear regression with CH_2O as outcome and AIR as explanatory variable and explain the results.

   In particular explain the concept of (multiple) R^2 and relate it to the correlation coefficient.

c) State the multiple linear regression model with CH_2O as outcome and both UFFI and AIR as explanatory variables.

   Run this model and interpret the results. Discuss similarities and differences between these results and what you found in questions a) and b).

Problem 2

The data for this problem stems from an investigation of whether a health reform in Germany in 1997 led to reduced number of doctoral visits. Some individuals were interviewed in 1996 (before the reform) and others in 1998 (after the reform). The data for this problem are restricted to women working full time and are given in the file drvisits.txt found at the course webpage. It has the following variables

`numvisits`	number of doctoral visits during the three months prior to the interview
age	age in years
educ	education in years
married	indicator variable for being married
badh	indicator for self-reported current health being classified as very poor or poor versus very good, good and fair
loginc	logarithm of household income (in 1995 Gemarn Marks, based on OECD weights for household members)
reform	indicator for the interview being undertaken in the year following the reform, coded 1, compared to the year preceding the reform, coded 0

You read the data into `R` using commands

```
path="https://www.uio.no/studier/emner/matnat/math/STK4900/data/drvisits.txt"
drv=read.table(path,header=T)
```

(Continued on page 3.)

a) Explain why Poisson regression may be reasonable for analyzing how the number of doctoral visits depends the covariates.

In particular explain the concept of a rate ratio.

Carry out a Poisson regression using only the health reform variable as covariate and use the results to find an estimate of the rate ratio. Give an interpretation of this estimate. Also calculate a 95% confidence interval for the rate ratio.

b) Consider a model with all covariates included. Investigate how the different covariates affect the tendency to visit a doctor. (For this question ignore possible non-linearities and interactions)

In particular note which covariates have a significant association with the tendency to visit the doctor.

Discuss and compare two different ways of carrying out such significance tests. For these comparisons you need only consider the two covariates `reform` and `badh`.

Also discuss why the effect of the health reform variable is only modestly changed after including the other covariates.

c) Study whether there are non-linear effects of the numerical covariates `age, educ` and `loginc` for instance by including quadratic or log-terms in the models.

d) Discuss the concept of over-dispersion relative to the Poisson assumption.

How can the analyses be corrected for such over-dispersion.

Carry out such an analysis using only main effects of the covariates. Compare with the previous analysis in question b). Comment on the differences.

Problem 3

In this problem we will consider the risk of wheezing (a whistling sound produced in the respiratory airways during breathing) among children according to whether their mother smokes and the childrens age. The file `wheezing.txt` found at the course website contains $n = 2148$ such responses. It has the following variables

(Continued on page 4.)

`smoking`	smoking status of the mother (1=yes, 0=no)
`age`	age of the child in years
`wheezing`	wheezing status of the child (1=yes, 0=no)

You may read the data into `R` using commands

```
path="https://www.uio.no/studier/emner/matnat/math/STK4900/data/wheezing.txt"
wheezing=read.table(path,header=T)
```

a) Calculate the proportions of children with wheezing both among children with smoking and non-smoking mothers.

Test whether the probabilities of wheezing are significantly different in these two groups.

Also calculate and compare the observed relative risk and observed odds-ratio of wheezing among children with smoking mothers compared to children of non-smoking mothers. Comment.

b) Analyze the wheezing data using logistic regression. In particular, demonstrate how the odds-ratio from question a) can be obtained from logistic regression. This demonstration should be done both theoretically and numerically from `R`-output.

Furthermore investigate whether age alone has an influence on the outcome wheezing and then whether the association with smoking is changed when taking age into account.

c) In the two first questions in this problem we have omitted the fact that the wheezing information is obtained on the same children at ages 7, 8, 9 and 10 years. In an extended file `wheezingb.txt` a variable `id` is included which identifies each pair of child and mother.

Discuss in general terms an issue that should handled with such longitudinal data.

Carry out an analysis that is appropriate for the data.

Comment on the differences with the analysis in question b). (You may note that each mother is recorded as a smoker or a non-smoker at every age).

END