

## 02424 Assignment 2

This is the second of three mandatory assignments for the course 02424. It must be handed in using the Campusnet (date and time are given at campus-net). The submissions must contain one collected attached file in portable document format (pdf), other document formats will not be accepted.

### Ear infection in swimmers

The data set `earinfect.txt` contains data from an observational study in New Zealand from 1990 where ocean swimmers (some frequent swimmers other occasional swimmers) were asked to count the number of ear infections they got in 1990. The variables in the data set are:

<code>swimmer</code>	Indicates if the swimmer is a frequent or an occasional ocean swimmer
<code>location</code>	Indicates the usually chosen swimming location: beach or non-beach
<code>age</code>	The age of the swimmer: 15-19, 20-24, 25-29
<code>sex</code>	The gender of the swimmer male or female
<code>infections</code>	Number of self diagnosed ear infections in 1990
<code>persons</code>	Number persons in that group

The goal is to find whether location, age, whether the swimmer is a frequent or occasional swimmer or interactions between these have any effects on the number of ear infections. When a final model has been found write out the model and interpret the parameters of the model.

1. Describe the content of this dataset in words.
2. Explain why a linear model would be inappropriate.
3. Explain why a model with an offset might be appropriate.
4. Fit a full model. What can you say about its goodness of fit (explain).
5. Try to reduce this model by successive likelihood ratio tests. Explain how you proceed to compare two models.
6. Report your best model (formula, goodness of fit).

When a final model has been found write out the model and interpret the parameters of the model.

## Ozone

In this part you should model ozone concentration in Los Angeles, the data is uploaded to campusnet along with this assignment, but is also included in the package `gclus`, and more information on the data can be obtained from there, e.g.

```
library(gclus)
data(ozone)
head(ozone)
```

##	Ozone	Temp	InvHt	Pres	Vis	Hgt	Hum	InvTmp	Wind
## 1	3	40	2693	-25	250	5710	28	47.66	4
## 2	5	45	590	-24	100	5700	37	55.04	3
## 3	5	54	1450	25	60	5760	51	57.02	3
## 4	6	35	1568	15	60	5720	69	53.78	4
## 5	4	45	2631	-33	100	5790	19	54.14	6
## 6	4	55	554	-28	250	5790	25	64.76	3

### Part 1

In the first part you should only consider additive and linear effects

1. Make a short presentation of the data
2. Fit a general linear model, and perform a residual analysis
3. The analysis above should suggest a transformation. Use a simple transformation on the dependent variable and perform the residual analysis again
4. Fit at least two different (sensible) generalized linear models to the data (you do not have to report residual plots of all the models here), and compare these models by quantitative numbers (you can play around with the distribution assumption and the link function).
5. Compare the model under question 3 and the model chosen from question 4, which one would you prefer (if you choose a quantitative measure you will need to take the transformation into account)?
6. For the chosen generalized linear model write down explicitly the diagonal elements of the weight matrix ( $W$ ) as a function of  $\mu_i$ , check your calculation by comparing the dispersion matrix of the parameters from the R function `(summary(fit)$cov.scaled)` with your own calculation.

### Part 2:

1. Develop the model you have chosen under the previous part, you might consider both higher order polynomials and interaction terms.
2. Present the final model.