Generalized Linear Models
Poisson data models-Overdispersion

1. The following data refer to new AIDS cases for 36 months starting from January.

0,0,3,0,1,1,1,2,2,4,2,8,0,3,4,5,2,2,2,5,4,3,15,12,7,14,6,10,14,8,19,10,7,20,10,19

a. Consider the data as independent observations Poisson(μ_i), i=1,…,36. Using the log link function and an appropriate model test the statistical significance of the following model Log(μ_i)=β0+β1*i
Provide the appropriate interpretation to coefficient β1.
b. Construct a variable which assigns data to four months season periods starting from January (winter). Fit the effect considering season as a factor. Fit also the effect considering season as a covariate (the linear effect of season). Which of the two models fits data better?
c. Consider now the following form of the data above emerging from aggregation over four-months periods
3,5,16,12,11,34,37,51,56
Plot the empirical logarithms of AIDS counts over four-months periods on the four months period. Use the four-months period as a covariate and fit its linear effect to the aggregated data. What is the difference between this fit and that made in (b.)?

2. The following data are coming from an experiment with aim to associate death and smoking in a population of British doctors. Table presents the number of deaths from coronary disease within a period of 10 years for smokers and non-smokers. The human years of observation are also shown which is the total number of years each member of this specific population was observed. The following questions should be answered. A. Is the death rate different between smokers and non-smokers? b. If yes, how different the two rates are? c. Is there an effect of age into the rate of death ignoring smoking? (age can be considered as factor) d. Does the effect of age differ between smokers and non-smokers? e. Fit a model which asserts that the effect of smoking does not depend on age. Test it goodness-of-fit f. Provide interpretation of model parameters for the model in (e.).

	Smokers		Non-smokers	
Age group	Deaths	Human years	Deaths	Human years
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

3. The table below shows the number of miles traveled by British trains (in millions) between 1970 and 1984 and the number of collisions. Fit model \mu_{i}=year_i^a*miles_i, where \mu_{i} is the expected number of collisions for year i, year_i is the year of the ith observation, and miles_i is the number of miles driven. Check goodness-of-fit of the model.

Αριθ. Περιπτ.	Ετος	Συγκρούσεις	Μίλια απόστασης
1	1970	3	281
2	1971	6	276
3	1972	4	268
4	1973	7	269
5	1974	6	281
6	1975	2	271
7	1976	2	265
8	1977	4	264
9	1978	1	267
10	1979	7	265
11	1980	3	267
12	1981	5	260
13	1982	6	231
14	1983	1	249

4. In file flowering.txt, there is data from flowering production of 5 varieties of evergreen trees. Data comes from a factorial experiment, where each tree was sprayed with one from six doses of some growth-promoting material. Six weeks later, plants were considered successful if they produced flowers and failed if they didn't produce flowers. The purpose is to relate the probability of flourishing with the dose of the material and the tree variety. Variables: flowered (number of plants that flourished), number (total number of plants), dose, variety.
A. Calculate the empirical logits and try to find the appropriate model for the data, by making an appropriate graph.
B. Start modelling with a model that fits the effect of the variety, the linear effect of the dose and the interaction of them. Comment the goodness-of-fit of the model and the possible appearance of overdispersion, the interpretation of the parameters, the adequacy of the model through residual plots.
C. Propose a better model with the purpose of restricting the overdispersion and fitting the data better. Since you have proposed a model like that, check its goodness-of-fit by using deviance and residual plots.

5. In the file germination.txt there is data of successful seed planting from two genotypes of the parasitic plant Orobanche and two other host plants that were used as exciters of the successful planting. We are interested in the probability of successful planting.

A. This is an example where overdispersion is appeared, but it cannot be solved by choosing a better model. Validate this by fitting the largest model you can (in this case it is those with the main effects and the interactions).
B. A simple way to deal with the problem of overdispersion is to assume that φ parameter does not take the value 1, but a higher value (see Chapter 16 of Crawley (2008) The R book, Wiley, where data is analyzed). Then, you have to estimate this parameter and it is suggested use an unbiased estimate from Pearson's \chi^2. However, in this case, deviance cannot be used for goodness-of-fit (but residuals should be used) and model comparisons should be accomplished

using F-tests. Find the best model, that fits the data adequately and comment the interpretation of the estimated model parameters.