1. Consider the dataset in the file chromoabnormal.txt:

```
  cells  ca doseamt doserate
1   47800  25       1     0.10
2  190700 102       1     0.25
3  225800 149       1     0.50
4  232900 160       1     1.00
5  123800  75       1     1.50
6  149100 100       1     2.00
.
27   14400 206     5.0     4.00

An experiment was conducted to determine the effect of gamma radiation
on the numbers of chromosomal abnormalities observed

A data frame with 27 observations on the following 4 variables.
cells - Number of cells
ca - Number of chromosomal abnormalities
doseamt - amount of dose in Grays
doserate - rate of dose in Grays/hour

Purott R. and Reeder E. (1976)
The effect of changes in dose rate on the yield of chromosome
aberrations in human lymphocytes exposed to gamma radiation.
Mutation Research. 35, 437-444.
```

Focus in the study is to model how the ratio between variables Y=ca and t=cells

$$Z = \frac{Y}{t} = \frac{\text{ca}}{\text{cells}}$$

depends on the explanatory variables X_1=doseamt and X_2=doserate. Let us also first assume that $Y_i \sim Poi(\mu_i)$.

(a) Consider the log link model with interaction term

$$\mathcal{M}_{12}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}.$$

Calculate the maximum likelihood estimate for the expected value μ_{i_*} when $x_{i_*1} = 4$, $x_{i_*2} = 0.75$, and $t_{i_*} = 64070$.

(2 points)

(b) Calculate the maximum likelihood prediction for the ratio

$$\frac{Y_f}{t_f}$$

when $x_{f1} = 4$, $x_{f2} = 0.75$. Also, create suitable prediction intervals for the ratio $\frac{Y_f}{t_f}$.

(2 points)

(c) Assume that $\mathrm{Var}(Y_i) = \phi\mu_i$. Test at 5% significance level, is the explanatory variable X_2=doserate statistically significant variable in the model

$$\mathcal{M}_{12}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2},$$

Calculate the value of the test statistic.

(1 point)

(d) Consider the model

$$\mathcal{M}_{12}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}.$$

Under which distribution, the model $\mathcal{M}_{12}$ fits best on data in your opinion?

i. Y_i follows Poisson distribution with the variance $\mathrm{Var}(Y_i) = \mu_i$,
ii. Y_i follows quasi-Poisson distribution with the variance $\mathrm{Var}(Y_i) = \phi\mu_i$,
iii. Y_i follows negative binomial distribution $Y_i \sim NegBin(\mu_i, \theta)$.

Report to which findings you have based your decision. (1 point)

2. Consider the data set appleCRA7152.txt, where it has been studied how the probability of bacterial spores of Alicyclobacillus Acidoterrestris CRA7152 growing in apple juice depends on the properties of the apple juice.

```
    pH Nisin Temperature Brix Growth
1  5.5    70          50   11      0
2  5.5    70          43   19      0
3  5.5    50          43   13      1
4  5.5    50          35   15      1
5  5.5    30          35   13      1
.
73 5.5    70          50   19      0
74 3.5     0          25   11      0

Presence/Absence of growth of CRA7152 in apple juice
as a function of pH (3.5-5.5), Brix (11-19), temperature (25-50C),
and Nisin concentration (0-70)

X1=pH
X2=Nisin concentration
X3=Temperature
X4=Brix Concentration
Y=Growth    (1=Yes, 0=No)

Source: W.E.L. Pena, P.R. De Massaguer, A.D.G. Zuniga, and S.H. Saraiva (2011).
"Modeling the Growth Limit of Alicyclobacillus Acidoterrestris CRA7152
in Apple Juice: Effect of pH, Brix, Temperature, and Nisin Concentration,"
Journal of Food Processing and Preservation, Vol. 35, pp. 509-517.
```

Denote the variables as following:

$$Y = \text{Growth}, \quad X_1 = \text{pH}, \quad X_2 = \text{Nisin}, \quad X_3 = \text{Temperature}, \quad X_4 = \text{Brix}.$$

(a) Consider modeling the expected value of the response variable $Y = \mathsf{Growth}$ by the explanatory variables X_1, X_2, X_3, X_4. Select the appropriate default distribution for the response variable Y, and consider several competing models. Choose the model which you feel is the most suitable one for modeling the expected value of the response variable $Y = \mathsf{Growth}$. Not all explanatory variables X_1, X_2, X_3, X_4 need to be included into your final model. Which link function $g(\mu_i)$ you chose for you model?

i. Identity link $g(\mu_i) = \mu_i$,
ii. log link $g(\mu_i) = \log(\mu_i)$,
iii. Inverse link $g(\mu_i) = \frac{1}{\mu_i}$,
iv. logit link $g(\mu_i) = \mathrm{logit}(\mu_i)$,
v. Probit link $g(\mu_i) = \Phi^{-1}(\mu_i)$,
vi. Cauchy link $g(\mu_i) = F^{-1}_{\mathsf{cauchy}}(\mu_i)$,
vii. Gumbel link $g(\mu_i) = \log\left(-\log(1-\mu_i)\right)$,

(2 points)

(b) Based on your chosen model, calculate the maximum likelihood estimate for the expected value μ_i when the explanatory variables are set on values

$$X_1 = 4.5, \quad X_2 = 20, \quad X_3 = 30, \quad X_4 = 17.$$

(1 point)

(c) Based on your chosen model, calculate the 95% confidence interval estimate for the expected value μ_i when the explanatory variables are set on values

$$X_1 = 4.5, \quad X_2 = 20, \quad X_3 = 30, \quad X_4 = 17.$$

(1 point)

(d) Extra question! If you solve this one, you get points, if you don't, you don't loose any points.
Let us assume that there are 100 apple juices with explanatory variables are set on values

$$X_1 = 4.5, \quad X_2 = 20, \quad X_3 = 30, \quad X_4 = 17.$$

How many of these 100 juices are such that bacterial spores of Alicyclobacillus Acidoterrestris CRA7152 are occuring in them? Create 80% prediction interval for the number of apple juices affected by Alicyclobacillus Acidoterrestris CRA7152 bacteria.

(2 points)

3. (a) In case of generalized linear model $g(\mu_i) = \beta_0 + \beta_1 x_i$, the maximum likelihood estimates for the parameters β_0 and β_1 are $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 0.5$. At the value $x_i = 5$, calculate the maximum likelihood estimate of μ_i, when the model is

   i. $Y_i \sim Poi(\mu_i)$ and $\log(\mu_i) = \beta_0 + \beta_1 x_i$,

   ii. $Y_i \sim Poi(\mu_i)$ and $\sqrt{\mu_i} = \beta_0 + \beta_1 x_i$,

   iii. $Y_i \sim Poi(\mu_i)$ and $\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_i$, where $t_i = 10$.

   iv.

$$
\begin{aligned}
P(Y_i = 0) &= \theta_i + (1 - \theta_i)e^{-\mu_i}, \\
P(Y_i = y_i) &= (1 - \theta_i)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \quad y_i = 1, 2, 3, \ldots \\
\log(\mu_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},
\end{aligned}
$$

   where the maximum likelihood estimate for the parameter θ_i is $\hat{\theta}_i = 0.25$.

(2 points)

(b) Let Y_i be such a random variable that for known n_i value, the product $n_i Y_i$ follows the binomial distribution $n_i Y_i \sim Bin(n_i, \mu_i)$. Derive with help of the properties of the binomial distribution what are the expected value $\mathrm{E}(Y_i)$ and the variance $\mathrm{Var}(Y_i)$ of the random variable Y_i.

(2 points)

(c) Let $Y_i \sim Cat(\theta_{i1}, \theta_{i2}, \theta_{i3})$, and consider the multinomial logit models

$$
\begin{aligned}
\log\left(\frac{\theta_{i2}}{\theta_{i1}}\right) &= \mathbf{x}_i'\boldsymbol{\beta}_2, \\
\log\left(\frac{\theta_{i3}}{\theta_{i1}}\right) &= \mathbf{x}_i'\boldsymbol{\beta}_3,
\end{aligned}
$$

   where $\theta_{i1} + \theta_{i2} + \theta_{i3} = 1$. Show that

$$
\begin{aligned}
\theta_{i1} &= \frac{1}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_2} + e^{\mathbf{x}_i'\boldsymbol{\beta}_3}}, \\
\theta_{i2} &= \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_2}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_2} + e^{\mathbf{x}_i'\boldsymbol{\beta}_3}}, \\
\theta_{i3} &= \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_3}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_2} + e^{\mathbf{x}_i'\boldsymbol{\beta}_3}}.
\end{aligned}
$$

(2 points)