# Statistical Modeling 1
# Problem set 5 2024

1. The research group wanted to study the diversity of plants in the Galapagos Islands. The group measured the number of different plants in each island (total of 30 islands) $Y =$ Species and measured the values of the different geographic variables for each island: $X_1 =$ Area - Surface area of island, hectares,
$X_2 =$ Elevation - Elevation in m,
$X_3 =$ Nearest- Distance to closest island, km,
$X_4 =$ Scruz- Distance from Santa Cruz Island, km,
$X_5 =$ Adjacent- Area of closest island, hectares.

```
          name Species   Area Elevation Nearest  Scruz Adjacent
1       Baltra      58  25.09       346     0.6    0.6     1.84
2    Bartolome      31   1.24       109     0.6   26.3   572.33
3     Caldwell       3   0.21       114     2.8   58.7     0.78
4     Champion      25   0.10        46     1.9   47.4     0.18
5      Coamano       2   0.05        77     1.9    1.9   903.82
.
.
29     Tortuga      16   1.24       186     6.8   50.9    17.95
30        Wolf      21   2.85       253    34.1  254.7     2.33
```

The data set can be found at the file galapagos.txt.

(a) Let us assume that the random variable $Y_i$ follows the Poisson distribution $Y_i \sim Poi(\mu_i)$. Consider modeling the expected value $\mu_i$ by the following model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}.$$

Calculate the maximum likelihood estimate for the expected value $\mu_{i_*}$ when the explanatory variables $X_1, X_2, \ldots X_5$ has the values

```
 Area Elevation Nearest Scruz Adjacent
58.27       198     1.1  88.3     0.57
```

(1 point)

(b) Let us continue with the assumption $Y_i \sim Poi(\mu_i)$ and the model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}.$$

Calculate the 95% confidence interval estimate for the expected value $\mu_{i_*}$ when the explanatory variables $X_1, X_2, \ldots X_5$ has the values

```
 Area Elevation Nearest Scruz Adjacent
58.27       198     1.1  88.3     0.57
```

Particularly, what is your obtained lower bound of the confidence interval?

(1 point)

(c) Let us continue with the assumption $Y_i \sim Poi(\mu_i)$ and the model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}.$$

Calculate the 80% prediction interval for the new observation $Y_f$ when the explanatory variables $X_1, X_2, \dots X_5$ has the values

```
  Area Elevation Nearest Scruz Adjacent
 58.27       198     1.1  88.3     0.57
```

Particularly, what is your obtained lower bound of the prediction interval? (1 point)

(d) Assume $Y_i \sim Poi(\mu_i)$. Consider the following hypotheses

$$\begin{aligned} H_0 &: \log(\mu_i) = \beta_0 + \beta_1 x_{i1}, \\ H_1 &: \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}. \end{aligned}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic, and return it as your answer to the question. (1 point)

(e) Consider modeling the expected value $\mu_i$ by the following Quasi-Poisson model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5},$$

where we assume that for variance $\mathrm{Var}(Y_i)$ it holds $\mathrm{Var}(Y_i) = \phi\mu_i$. Calculate the unbiased estimate $\hat{\phi}$ for the dispersion parameter $\phi$. (1 point)

(f) Let us assume that the appropriate link function is square root link $g(\mu_i) = \sqrt{\mu_i}$. Further, assume that $\mathrm{Var}(Y_i) = \phi\mu_i$. Consider the following hypotheses

$$\begin{aligned} H_0 &: \sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1}, \\ H_1 &: \sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}. \end{aligned}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic, and return it as your answer to the question. (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:

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

(a) Consider modeling the expected value $\mu_i$ by the model

$$\mathcal{M}_{1|2|3|4}: \quad \text{logit}\,(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}.$$

Calculate the maximum likelihood estimate for the expected value $\mu_i$ when the explanatory variables are set on values

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

(2 points)

(b) Use the same model as at (a). Calculate the 95% confidence interval estimate for the expected value $\mu_i$ when the explanatory variables are set on values

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

(2 points)

(c) Consider the hypotheses

$$H_0 : \text{Model logit}\,(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \text{ is the true model,}$$
$$H_1 : \text{Model logit}\,(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \text{ is the true model.}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic.

(1 points)

(d) Which link function fits best to data in your opinion, if you use the main effect model

$$g(\mu_{jh}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

to model the data?

i. Logit link $g(\mu_i) = \text{logit}(\mu_i)$,
ii. Probit link $g(\mu_i) = \Phi^{-1}(\mu_i)$,
iii. Cauchy link $g(\mu_i) = F^{-1}_{\text{cauchy}}(\mu_i)$,
iv. Gumbel link (complementary log-log) $g(\mu_i) = \log\left(-\log(1-\mu_i)\right)$.

(1 point)

3. (a) In case of generalized linear model $g(\mu_i) = \beta_0 + \beta_1 x_i$, the maximum likelihood estimates for the parameters $\beta_0$ and $\beta_1$ are $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 0.5$. At the value $x_i = 5$, calculate the maximum likelihood estimate of $\mu_i$, 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 $t_i = 10$.

(2 points)

(b) In generalized linear models, the likelihood equations can written in form

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\text{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0, \qquad j = 0, 1, 2 \ldots p.$$

Consider now the most simplest Poisson model with the identity link function

$$Y_i \sim Poi(\mu_i),$$
$$\mu_i = \eta_i = \beta_0.$$

What kind of more simplified form the likelihood equations have in this case? That is, what form $\frac{\partial l(\beta_0)}{\partial \beta_0}$ has in the simplest Poisson model? By using the likelihood equations, find the maximum likelihood estimator $\hat{\beta}_0$.

(2 points)

(c) In generalized linear models, the likelihood equations can written in form

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} x_{ij} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) = 0, \qquad j = 0, 1, 2 \ldots p.$$

Consider now the simple logit model with

$$\begin{aligned} Y_i &\sim Ber(\mu_i), \\ \mathrm{logit}(\mu_i) &= \eta_i = \beta_0. \end{aligned}$$

What kind of more simplified form the likelihood equations have in this case? That is, what form $\frac{\partial l(\beta_0)}{\partial \beta_0}$ has in the simple logit model? By using the likelihood equations, find the maximum likelihood estimator $\hat{\beta}_0$.

(2 points)