Statistical Modeling 1
Problem set 3 **2024**

---

1. Consider the data in file Alba.txt.

```
> data<-read.table("Alba.txt", sep="\t", dec=".", header=TRUE)
> head(data)
  Dose  Herbicide DryMatter
1     1 Glyphosate       4.7
2     1 Glyphosate       4.6
3     1 Glyphosate       4.1
4     1 Glyphosate       4.4

Data are from an experiment, comparing the potency of the two herbicides
glyphosate and bentazone in white mustard Sinapis alba.
Dose - a numeric vector containing the dose in g/ha.
Herbicide - a factor with levels Bentazone Glyphosate (the two herbicides applied).
DryMatter - a numeric vector containing the response (dry matter in g/pot).
Christensen, M. G. and Teicher, H. B., and Streibig, J. C. (2003)
Linking fluorescence induction curve and biomass in herbicide screening,
Pest Management Science, 59, 1303?1310.
```

Denote the variables as Y = DryMatter, X_1 = Dose, and X_2 = Herbicide.

(a) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the models

$$
\begin{aligned}
\mathcal{M}_{\text{identity}}: \quad \mu_i &= \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}, \\
\mathcal{M}_{\log}: \quad \log(\mu_i) &= \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}, \\
\mathcal{M}_{\text{inverse}}: \quad \frac{1}{\mu_i} &= \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}, \\
\mathcal{M}_{\frac{1}{\mu^2}}: \quad \frac{1}{\mu^2} &= \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1},
\end{aligned}
$$

where index j is related to the categories of X_2. Which model fits the best to the data? You may use the mean square error value $\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{n}$ as one method of choosing the model. Return the maximum likelihood estimate $\hat{\beta}_1$ of your chosen model as your solution for this question.

(2 points)

(b) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the model

$$\mathcal{M}: \quad g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}.$$

Choose the link function g based on your solution to (a). Calculate the maximum likelihood estimate for the expected value μ_{i_*} when $X_1 = 50$ and X_2 = Glyphosate. (1 point)

(c) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the model

$$\mathcal{M}: \quad g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}.$$

Choose the link function g based on your solution to (a). Calculate the 95% confidence interval estimate for the expected value μ_{i_*} when $X_1 = 50$ and $X_2 =$ Glyphosate. Particularly, what is your obtained lower bound of the confidence interval?

(1 point)

(d) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the model

$$\mathcal{M}: \quad g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}.$$

Choose the link function g based on your solution to (a). Create 80 % prediction interval for new observation Y_f, when $X_1 = 50$ and $X_2 =$ Glyphosate. Particularly, what is your obtained lower bound of the prediction interval?

(2 points)

2. Consider the following data set ratstime.txt:

```
  time poison treat
1 0.31       I     A
2 0.82       I     B
3 0.43       I     C
4 0.45       I     D
5 0.45       I     A
6 1.10       I     B
.
.

Effect of toxic agents on rats

Description

An experiment was conducted as part of an investigation
to combat the effects of certain toxic agents.

A data frame with 48 observations on the following 3 variables.

time
survival time in tens of hours

poison
the poison type - a factor with levels I II III

treat
the treatment - a factor with levels A B C D
```

The response variable is $Y =$ time and the explanatory variables are X_1=poison and X_2=treat.

(a) Let us assume $Y_i \sim N(\mu_{jh}, \sigma^2)$ and let us model the data with the main effect model

$$g(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h.$$

Based on the mean square error value $\mathrm{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_{jh})^2}{n}$, which link function fits best to the data:

i. $\mu_{jh} = \beta_0 + \beta_j + \alpha_h$,
ii. $\log(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h$,
iii. $\frac{1}{\mu_{jh}} = \beta_0 + \beta_j + \alpha_h$?

(2 points)

(b) Distributional assumption could be either $Y_i \sim N(\mu_{jh}, \sigma^2)$, $Y_i \sim Gamma(\mu_{jh}, \phi)$, or $Y_i \sim IG(\mu_{jh}, \phi)$. Choose the link function g based on your solution to (a). Based on your analysis, which one is most suitable in this case?

i. $Y_i \sim N(\mu_{jh}, \sigma^2)$,
ii. $Y_i \sim Gamma(\mu_{jh}, \phi)$,
iii. $Y_i \sim IG(\mu_{jh}, \phi)$.

(1 point)

(c) Regardless what was your solution to the question (a) and (b), let us assume $Y_i \sim IG(\mu_{jh}, \phi)$ and consider the model

$$\log(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h.$$

Create 95 % confidence interval for μ_{jh}, when

$$x_{i_*1} = II, \qquad x_{i_*2} = B.$$

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

(1 point)

(d) Regardless what was your solution to the question (a) and (b), let us assume $Y_i \sim Gamma(\mu_{jh}, \phi)$ and consider the model

$$\log(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h.$$

Create 80 % prediction interval for new observation Y_f, when

$$x_{f1} = II, \qquad x_{f2} = B.$$

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

(2 points)

3. (a) In case of generalized linear model $g(\mu_i) = \beta_0 + \beta_1 x_i$, find the inverse function g^{-1} (that is, solve what form the expected value μ_i has), when the link function g is

   i. $\sqrt{\mu_i} = \beta_0 + \beta_1 x_i$,
   ii. $\frac{1}{\mu_i^2} = \beta_0 + \beta_1 x_i$,
   iii. $\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i$.

   (2 points)

   (b) Let us assume $Y_i \sim IG(\mu_i, \phi)$. Consider the model

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_i).$$

   Let the estimates of the parameters β_0, β_1, ϕ be as $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0.5, \tilde{\phi} = 0.05$.

   i. Calculate the maximum likelihood estimate for the expected value μ_i when $x_i = 5$.
   ii. Calculate the Pearson residual

$$o_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathrm{Var}}(Y_i)}},$$

   when $x_i = 5$ and the observed value is $y_i = 12$.

   (2 points)

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

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

   Consider now the simple Gamma model with

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

   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 simple Gamma model? By using the likelihood equations, find the maximum likelihood estimator $\hat{\beta}_0$.

   (2 points)