Statistical Modeling 1
Problem set 6 2024

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 of the ratio $\mathrm{E}(Z_i) = \frac{\mu_{i*}}{t_i}$ when $x_{i_*1} = 4$, $x_{i_*2} = 0.75$.

(2 points)

(b) Create 80 % prediction interval for new observation Y_f, when $x_{f1} = 4$, $x_{f2} = 0.75$, and $t_f = 64070$. Particularly, what is your obtained lower bound of the prediction interval?

(1 point)

(c) Create 80 % prediction interval for new ratio variable $Z_f = \frac{Y_f}{t_f}$, when $x_{f1} = 4$, $x_{f2} = 0.75$, and $t_f = 64070$. Particularly, what is your obtained lower bound of the prediction interval?

(1 point)

(d) 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)

(e) 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 $Y_i \sim NegBin(\mu_i, \theta)$.

(1 point)

2. Consider the dataset betablocker.txt:

```
   Deaths Total Center Treatment
1       3    39      1   Control
2      14   116      2   Control
3      11    93      3   Control
4     127  1520      4   Control
5      27   365      5   Control
6       6    52      6   Control
.
.
22-centre clinical trial of beta-blockers for
reducing mortality after myocardial infarction.

A data frame with 44 observations on the following 4 variables.
Deaths
Number of deaths.
Total
Total number of patients.
Center
Number of clinical centre.
Treatment
A factor with levels Control and Treated.

M. Aitkin.
Meta-analysis by random effect modelling in generalized linear models.
Statistics in Medicine, 18, 2343-2351, 1999.

S. Yusuf, R. Peto, J. Lewis, R. Collins and P. Sleight.
Beta blockade during and after myocardial infarction: an overview of the randomized trials.
Progress in Cardiovascular Diseases, 27, 335-371, 1985.
```

Data is in "grouped data" form, where the response variable Y can have outcomes

$$Y = \begin{cases} 1, & \text{dead}, \\ 0, & \text{alive}. \end{cases}$$

The column Deaths measures frenquencies of the outcome dead, and the column Total measures the sum of frequencies of the outcomes dead and alive. Each row gives a unique combination of the explanatory variables, under which conditions the frequencies of the outcomes dead and alive have occurred.

Denote $P(Y_i = 1) = \mu_i$. Consider the logistic model

$$\mathcal{M}: \quad \text{logit}(\mu_i) = \beta_0 + \beta_j + \alpha_h,$$

where β_j are parameters related to the categories of X_1 = Treatment variable and α_h are parameters related to X_2 = Center variable.

(a) Under the model $\mathcal{M}$, calculate the maximum likelihood estimate $\hat{\mu}_{i_*}$ for the expected value μ_{i_*} when $x_{i_*1} = \textsf{Treated}$ and $x_{i_*2} = 10$.

(2 points)

(b) Under the model $\mathcal{M}$, calculate 95% confidence interval for the expected value μ_{i_*} when $x_{i_*1} = \textsf{Treated}$ and $x_{i_*2} = 10$.

(1 point)

(c) Under the model $\mathcal{M}$, calculate the estimate for the odds ratio

$$\psi_{\textsf{Treated,10}|\textsf{Control,10}}.$$

(1 point)

(d) Let us assume that there are 100 (new) patients with explanatory variables are set on values $x_{if1} = \textsf{Treated}$ and $x_{if2} = 10$. Create 80% prediction interval for sum $Y_S = \sum_{i=1}^{100} y_{if}$.

(2 points)

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

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

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

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

(2 points)

(b) Let the random variable Y_i be defined on ordinal scale with m distinctive possible outcomes. Let the possible outcomes have natural order "1" $<$ "2" $<$ "3". Consider cumulative proportional odds logit model

$$\log\left(\frac{P(Y_i \le k)}{1 - P(Y_i \le k)}\right) = \text{logit}\left(P(Y_i \le k)\right) = \beta_{0k} + \beta_1 x_{i1}, \qquad k = 1, 2.$$

Solve the probabilities $P(Y_i = 1), P(Y_i = 2), P(Y_i = 3)$ as functions of parameters β_{0k}, β_1.

(2 points)

(c) Let Y_i be such a random variable that for known n_i value, the product n_iY_i follows the binomial distribution $n_iY_i \sim Bin(n_i, \mu_i)$. 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)