Statistical Modeling 2
Weekly assignments 2 2021

Solutions for the following questions should be returned to Moodle's quiz platform. Platform will be opened soon. The last return time is 9 April at 12.00.

1. Consider the data set related patients having AIDS aids.txt:

```
      cd4      time drugs    age person
1     548 -0,741958     0   6,57  10002
2     893 -0,246407     1   6,57  10002
3     657  0,243669     1   6,57  10002
4     464 -2,729637     1   6,95  10005
5     845 -2,250513     1   6,95  10005
6     752 -0,221766     1   6,95  10005
7     459  0,221766     1   6,95  10005
8     181  0,774812     1   6,95  10005
9     434  1,256673     1   6,95  10005
10    846 -1,240246     1   2,64  10029
11   1102 -0,741958     1   2,64  10029
12    801 -0,251882     1   2,64  10029
13    824  0,251882     1   2,64  10029
14    866  0,769336     1   2,64  10029

.
2371  606 -0,238193     1  -5,04  41844
2372  570  0,238193     1  -5,04  41844
2373  826  0,772074     1  -5,04  41844
2374  983  1,538672     1  -5,04  41844
2375  517  2,056126     0  -5,04  41844
2376  462  3,419576     1  -5,04  41844


The aids data was a survey around 369 men who were infected with HIV.

A data frame with 2376 observations on the following 8 variables.
cd4 - number of CD4 cells
time - years since seroconversion
drugs - recreational drug use (yes=1/no=0)
age - Age centered around 30
person -Identification number
```

Denote variables as following $Y = \mathsf{cd4}$, $X_1 = t = \mathsf{time}$, $X_2 = \mathsf{drugs}$ with index j, and $X_3 = \mathsf{age}$. Consider the generalized linear mixed effects model

$$\mathcal{M}: \quad g(\mu_{it}) = \beta_0 + \beta_1 t + \alpha_j + \gamma_{1j} t + \beta_3 x_{it3} + b_{i0} + b_{i1} t,$$

where $\beta_0, \beta_1, \alpha_j, \gamma_{1j}, \beta_3$ are fixed parameters, and b_{i0}, b_{i1} are random parameters related to person i with assumptions of following normal distributions

$$b_{i0} \sim N(0, \sigma^2_{b_0}), \qquad b_{i1} \sim N(0, \sigma^2_{b_1}).$$

For each person i at the time t, the (conditional) random variable y_{it} could be considered to follow either Poisson distribution

$$y_{it} \sim Poi(\mu_{it}),$$

or Gamma distribution

$$y_{it} \sim Gamma(\mu_{it}, \phi),$$

or inverse Gaussian distribution

$$y_{it} \sim IG(\mu_{it}, \phi),$$

or just normal distribution

$$y_{it} \sim N(\mu_{it}, \sigma^2).$$

(a) Consider different distributional assumptions and try different link function g to data. Based on your analysis, choose which distributional assumption and which link function fit best to the data.

(3 points)

(b) Under the model $\mathcal{M}$, calculate the maximum likelihood estimate $\hat{\mu}_{i_*t}$ for the expected value μ_{i_*t} when

```
time drugs     age
2.02     1      13.72
```

(1 point)

(c) Under the model $\mathcal{M}$, construct 80% prediction interval for the new observation y_f, when

```
time drugs     age     person
2.02     1      13.72   10396
```

(2 points)

2. Consider the dataset locust.txt, where the aim is to analyze the effect of hunger on locomotory behaviour of locust.

```
     id move sex        time feed
1     1    0   1 0.008333333    1
2     1    0   1 0.016666667    1
3     1    0   1 0.025000000    1
4     1    0   1 0.033333333    1
.
.
3863 24    1   0 1.333333333    0
3864 24    1   0 1.341666667    0
```

```
The aim is to analyze the effect of hunger on locomotory behaviour of 24 locust
(Locusta migratoria) observed at 161 time points.
The subjects were divided in two treatment groups ("fed" and "not fed"),
and within each of the two groups, the subjects were alternatively "male" and "female".
For the purpose of this analysis the categories of the response variable
were "moving"=1 and "not moving"=0. During the observation period,
the behavior of each of the subjects was registered every thirty seconds.

id - a numeric vector that identifies de number of the individual profile.
move - a numeric vector representing the response variable.
sex - a factor with levels 1 for "male" and 0 for "female".
time - a numeric vector that identifies de number of the time points observed.
The time vector considered was obtained dividing (1:161)
by 120 (number of observed periods in 1 hour).
feed - a factor with levels 0 "no" and 1 "yes".

The response variable, move is the binary type coded as 1 for "moving" and 0 for "not moving".
The sex covariate was coded as 1 for "male" and 0 for "female".
The feed covariate indicating the treatment group, was coded as 1 for "fed" and 0 for "not fed".
```

Denote the variables as following

$$Y = \mathsf{move}, \quad X_1 = \mathsf{sex}, \quad X_2 = \mathsf{feed}, \quad T = \mathsf{time}.$$

Note that the variable id identifies the locust i which behaviour is observed repeated times. Assume $y_{it} \sim Ber(\mu_{it})$. Consider the mixed effect model

$$\mathcal{M}: \quad \text{logit}(\mu_{it}) = \beta_0 + \beta_1 t + \alpha_j + \gamma_h + b_{i0} + b_{i1} t,$$

where the indexes j, h are related to variables X_1, X_2. For each subject i, the random effects $\mathbf{b}_i = (b_{i0}, b_{i1})'$ are assumed to follow normal distribution $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$.

(a) Under the model $\mathcal{M}$, find the estimate for the parameter β_1.

(2 points)

(b) Under the model $\mathcal{M}$, calculate the prediction $\tilde{\mu}_{i_*t}$ for the expected value μ_{i_*t} when

id	sex	time	feed
24	0	1.35	0

(1 point)

(c) Test at 5% significance level, is the explanatory variable $X_2 = \mathsf{feed}$ statistically significant variable in the model. Calculate the value of the test statistic.

(1 point)

(d) Find the estimate for the covariance matrix

$$\mathrm{Cov}(\mathbf{b}_i) = \mathrm{Cov}\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_0,b_1} \\ \sigma_{b_0,b_1} & \sigma^2_{b_1} \end{pmatrix}.$$

(1 point)

(e) Suppose that there are extra 100 locusts outside the data with all having the features

sex	feed
0	0

Predict how many of these extra 100 locusts are moving at the time point $\mathsf{time} = 1.35$. That is, create 80% prediction interval for the number of locusts of these extra 100 locusts which are moving at the time point $\mathsf{time} = 1.35$.

(1 point)

3. (a) Consider the simple generalized linear mixed effects model

$$g(\mu_i) = \beta_0 + b_h,$$

with the random effect b_h following the normal distribution $b_h \sim N(0, \sigma_z^2)$.

i. Calculate the expected value $\mathrm{E}(y_i) = \mathrm{E}(\mu_i) = \mathrm{E}(\mathrm{E}(y_i|b_h))$ when the link function g is the identity link

$$\mu_i = \beta_0 + b_h.$$

ii. Calculate the expected value $\mathrm{E}(y_i) = \mathrm{E}(\mu_i) = \mathrm{E}(\mathrm{E}(y_i|b_h))$ when the link function g is the log link

$$\mu_i = \exp(\beta_0 + b_h).$$

(3 points)

(b) Consider the generalized linear mixed effects model

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + b_h,$$

with random effects $\mathbf{b} = \{b_h\} = (b_1, b_2, \ldots b_q)'$ following the normal distribution

$$\mathbf{b} = N(\mathbf{0}, \sigma_z^2\mathbf{I}).$$

Let $l(\mathbf{b}) = \log(f(\mathbf{b}))$ be log-likelihood function of the marginal distribution $\mathbf{b} = N(\mathbf{0}, \sigma_z^2\mathbf{I})$. Write the log-likelihood function $l(\mathbf{b})$ in simplest form you can.

(2 points)

(c) Consider the simple logistic mixed effect model

$$\begin{aligned} y_i &\sim Ber(\mu_i), \\ \mathrm{logit}(\mu_i) &= \beta_0 + b_h, \\ b_h &\sim N(0, \sigma_z^2). \end{aligned}$$

What form the joint density function $f(y_i, b_h) = f(y_i|b_h) \cdot f(b_h)$ has?

(1 point)