P8157: Analysis of Longitudinal Data (Fall 2019)
Midterm Take Home Examination
October 31, 2019 - November 12, 2019

Name: Xinyi Lin

UNI: xl2836

Instructions

1. Make sure you write down your name and UNI.
2. Make sure you have all 11 pages.
3. **NO Questions to the TA**. If you are not sure of a question, provide your interpretation and proceed.
4. **Show all the necessary steps and calculations.**
5. **Please Do Not Collaborate with your classmates.**
6. *No Collaboration*

You **May** or **May Not** need this.

Some Common Distributions

	Name	$f(y)$
1	Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$
2	Binomial	$\binom{n}{y} p^y (1-p)^{n-y}$
3	Poisson	$\frac{e^{-\lambda}\lambda^y}{y!}$
4	Exponential	$\lambda e^{-\lambda y}$
5	Exponential Family	$exp\left(\frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi)\right)$
6	Multivariate Normal	$\frac{1}{(2\pi)^{K/2}\lvert\sum\rvert^{1/2}}\ exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$

Question 1

Recent studies have shown the relevance of the cerebral grey matter involvement in multiple sclerosis (MS). The number of Cortical Lesions (CLs) in a subject , detected by specific MRI sequences, are likely to become a new research outcome in MS studies. A recent research article claimed that the negative binomial model gave the best fit to the observed distribution of CLs. If Y is a Negative Binomial random variable with success probability p and a goal of r successes than the probability mass function is given by

$$f(y|p,r) = \binom{r+y-1}{y} p^r (1-p)^y \quad r > 0,\ p \in (0,1) \ .$$

~~p is the parameter~~ of interest. You may assume the parameter r as given or known.

(a) Derive the mean and variance of the random variable Y. (10)

$$f(y|p,r) = \binom{r+y-1}{y} p^r (1-p)^y \quad r>0,\ p\in(0,1)$$

$$= \exp\left\{ r\log p + y\log(1-p) + \log\binom{r+y-1}{y} \right\}$$

$$= \exp\left\{ y\log(1-p) - (-r\log p) + \log\binom{r+y-1}{y} \right\}$$

So $\theta = \log(1-p)$, $p = 1-e^{\theta}$, $b(\theta) = -r\log p = -r\log(1-e^{\theta})$

$$E[Y] = b'(\theta) = -r\frac{-e^{\theta}}{1-e^{\theta}} = \frac{re^{\theta}}{1-e^{\theta}} = \frac{r(1-p)}{p}$$

$$Var[Y] = a(\psi)\, b''(\theta) = 1 \times r\left[-\frac{-e^{\theta}}{(1-e^{\theta})^2}\right] = \frac{re^{\theta}}{(1-e^{\theta})^2} = \frac{(1-p)r}{p^2}$$

(b) A pilot study is conducted with 44 subjects with 22 each in the control and treatment group to study the efficacy of a new drug. The CLs are measured at baseline and at 6 months after the intervention. Set up the model assuming a negative binomial distribution for the outcome to study the effect of treatment on cortical lesions. (10)

Let y_{ij} the number of CLs for subject $i=1,\dots,44$ at time $j=0,1$

y_{ij} follows Negative binomial distribution

Canonical Link: $b'(\theta)=\mu=\frac{re^{\theta}}{1-e^{\theta}}$ $\quad \theta=g(\mu)=\log(\frac{\mu}{\mu+r})$

Variance Function : $b''(\theta)=\frac{re^{\theta}}{(1-e^{\theta})^2}$

Model: $\log(\frac{\mu_{ij}}{\mu_{ij}+r})=\beta_0+\beta_1 I\{Trt=1\}+\beta_2 I\{time=1\}+\beta_3 I\{Trt=1\}\{time=1\}$

$Var(Y_{ij})=\frac{(1-p_{ij})r_{ij}}{p_{ij}^2}=\frac{(\mu_{ij}+r_{ij})\mu_{ij}}{r_{ij}}$

correlation structure: $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}_{2\times 2}$

(c) Pearson Residuals are commonly used in the diagnostics of regression models involving non-gaussian outcomes. Provide an expression to compute the Pearson residual for the model constructed in part (b). (5)

Pearson Residuals : $\hat{r}_{ij} = \frac{(y_{ij} - \hat{\mu}_{ij})}{\sqrt{Var(\hat{\mu}_{ij})}}$

$$\mu = \frac{r(1-p)}{p} \Rightarrow p = \frac{r}{\mu+r}$$

$$Var(\mu) = \frac{(1-p)r}{p^2} = \frac{(1-\frac{r}{\mu+r})\,r}{\frac{r^2}{(\mu+r)^2}} = \frac{(\mu+r)\mu}{r}$$

$$\text{So } \hat{r}_{it} = \frac{(y_{ij}-\hat{\mu}_i)}{\sqrt{\frac{(\mu_{ij}+r)\mu_{ij}}{r}}} = \frac{(y_{ij} - \frac{re^{\theta_{ij}}}{1-e^{\theta_{ij}}})}{\sqrt{\frac{(\frac{re^{\theta_{ij}}}{1-e^{\theta_{ij}}}+r)\frac{re^{\theta_{ij}}}{1-e^{\theta_{ij}}}}{r}}}$$

$$= \frac{y_{ij} - \frac{re^{\theta_{ij}}}{1-e^{\theta_{ij}}}}{\sqrt{\frac{re^{\theta_{ij}}}{(1-e^{\theta_{ij}})^2}}} = \frac{y_{ij} - \frac{re^{x^T\beta}}{1-e^{x^T\beta}}}{\sqrt{\frac{re^{x^T\beta}}{1-e^{x^T\beta}}}}$$

where $x^T\beta = \beta_0 + \beta_1 I\{Trt=1\} + \beta_2 I\{time=1\} + \beta_3 I\{Trt=1\} I\{Time=1\}$

Question 2

The level of the biomarker B at four days post cardiovascular surgery is known to be a good predictor of longterm quality of life and mortality. Lower the levels better the prognostication. A Randomized Clinical Trial (RCT) to study the effects of a new drug in reducing the level of the biomarker post surgery is conducted. A total of 95 subjects were enrolled in the study (Control=45, Treatment=50). The Biomarker levels were measured for the first 4 days post surgery. All measurements are taken as soon as the patient wakes up in the morning. There are no significant measurement errors. There are no missing data. The plot of individual profiles as well as the mean at each time point is shown in figure 1

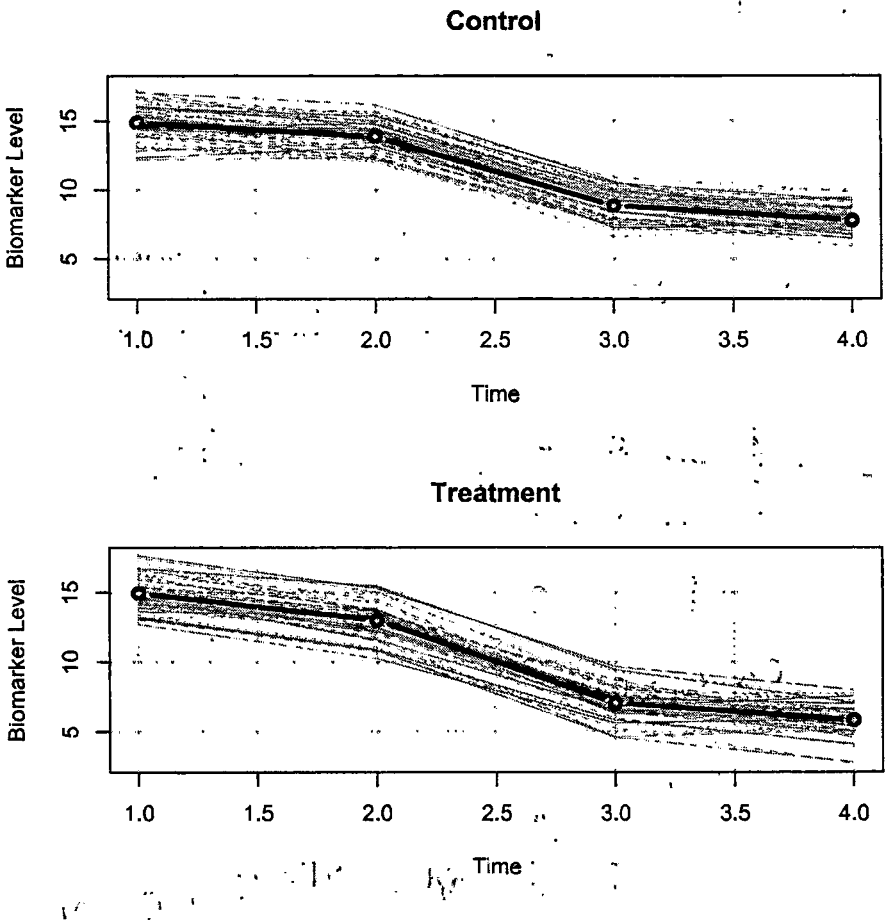

Figure 1: Plots of Biomarker levels across time for the treatment and control groups

(a) It is of interest to know at what time points prior to the fourth day are the levels indistinguishable from the fourth day measurements. Also is this difference different between the control and treatment groups. Set up a suitable *classical* multivariate model stating relevant assumptions to test this hypothesis. (8)

Let y_{ij} be the biomarker level for subject $i = 1, \dots, 95$ at time $j = 1, 2, 3, 4$

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix}, \quad y_i \sim N_4(\mu_1, \Sigma), \ y_i \text{ is independent.}$$

$$Y_{95\times4} = \begin{bmatrix} y_{11} & \cdots & y_{1,4} \\ \vdots & & \\ y_{95\,1} & \cdots & y_{95,4} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{95} \end{bmatrix} \qquad X_{95\times2} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}$$

$$B_{2\times4} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \end{bmatrix}$$

$$E_{95\times4} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} & \epsilon_{14} \\ \vdots & & & \\ \epsilon_{95\,1} & \epsilon_{95\,2} & \epsilon_{95\,3} & \epsilon_{95\,4} \end{bmatrix} \qquad Y_{95\times4} = X_{95\times2} B_{2\times4} + E_{95\times4}$$

Assumptions: ① $y_i \sim N_4(\mu_i, \Sigma)$

y_i in each group follows multivariate normal distribution with common mean μ_{group} and two group have common variance-covariance Σ.

② $\varepsilon_{ij} \sim N_4(0, \Sigma)$

③ subjects in each group are independent.

(b) Suggest a suitable model (in the GEE framework), stating any relevant assumptions to describe the mean response of the Biomarker over time and study the effect of the treatment. (7)

Let y_{ij} the level of the biomarker B for subject $i=1 \ldots 25$ at time $j=1, 2, 3, 4$

y_{ij} follows Normal distribution $N(\mu_{ij}, \sigma^2)$.

canonical link $\theta=\mu$, variance function : σ^2

$$\text{Model}: E[Y_{ij}] = \beta_0 + \beta_1 I\{Trt=1\} + \beta_2 I\{Time=2\} + \beta_3 I\{Time=3\} + \beta_4 I\{Time=4\} + \beta_5 I\{Trt=1\} I\{Time=2\} + \beta_6 I\{Trt=1\} I\{Time=3\} + \beta_7 I\{Trt=1\} I\{Time=4\} .$$

$Var(Y_{ij}) = \sigma_j^2$

correlation structure : 4X4 matrix (could be autoregressive, compound symmetry, unstructured or other correlation structured).

① exchangeable correlation $\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$

② autoregressive correlation $corr(y_{it}, y_{it'}) = \rho^{|t-t'|}$ $\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$

③ Stationary correlation $R_{uv} = \begin{cases} \rho_{|u-v|} & \text{if } |u-v| \le k \\ 0 & , \text{else} \end{cases}$

④ non-stationary $R_{uv} = \begin{cases} 1 , & u=v \\ \rho_{uv} , & 0<|u-v| \le k \\ 0 , & \text{else} \end{cases}$

⑤ unstructured $R_{uv} = \begin{cases} 1 , & u=v \\ \rho_{uv} , & \text{else} \end{cases}$

Assumptions ① y_{ij} follows Normal distribution

② a prior correlation structure

③ the canonical link is identity link

(c) Exploratory data analysis from a prior similar study provided the following estimates for the correlation structure for the within patient association.

$$\begin{bmatrix} 1 & 0.6 & 0.36 & 0.216 \\ 0.6 & 1 & 0.6 & 0.36 \\ 0.36 & 0.6 & 1 & 0.6 \\ 0.216 & 0.36 & 0.6 & 1 \end{bmatrix}$$

Using this information, suggest a suitable correlation structure with minimal number of parameters. Justify your choice and clearly explain your correlation structure. (5)

I think autoregressive - correlation structure is suitable

Let $\alpha = 0.6$. corr$(y_{it}, y_{it'}) = \alpha^{|t'-t|}$ fit the estimated correlation structure(4x4) above with only 1 parameters.

If we want to get variance-covariance matrix, then we should estimate variance σ^2 and α two parameters.

(d) A model is constructed using treatment, time and their interaction as covariates and the estimates from the analysis are provided below. (5)

	Estimate	Std.err	P-Value
(Intercept)	15.08	0.17	< 0.01
Day-2 ①	-0.99	0.14	< 0.01
Day-3 ②	-5.97	0.19	< 0.01
Day-4 ③	-7.09	0.23	< 0.01
Group-Treatment④	-0.25	0.24	0.30
Day-2 * Group-Treatment ⑤	-1.09	0.19	< 0.01
Day-3 * Group-Treatment⑥	-2.07	0.26	< 0.01
Day-4 * Group-Treatment ⑦	-1.97	0.30	< 0.01

1. Is there is evidence to suggest that Randomization of the subjects was successful with regard to the outcome?

2. Provide your inference about the Biomarker levels with regard to time and treatment using the above table.

1. Assume measurement took on Day 1 is baseline level of two groups. The p-value of estimated coefficient for "Group-Treatment" is 0.3 which is higher than 0.05. This indicate on baseline, there is no significant difference between control and treatment group and randomization was successful.

2. β_0. the expected biomarker level of control group is 15.08 on first day.
β_1: the difference of expected biomarker levels on control group between first day and second day is -0.99.
β_2, β_3 are similar
β_4. the difference of expected biomarker levels on Day 1 between treatment and control group is -7.09
$\beta_4+\beta_5$: the difference of expected biomarker levels between control group and rtreatment group on Day 2
β_6. β_7 are similar

	Control	Treatment
Day 1	$\beta_0 = 15.08$	$\beta_0 + \beta_4 = 14.83$
Day 2	$\beta_0 + \beta_1 = 14.09$	$\beta_0+\beta_1+\beta_4+\beta_5 = 12.75$
Day 3	$\beta_0 + \beta_2 = 9.11$	$\beta_0+\beta_2+\beta_4+\beta_6 = 6.79$
Day 4	$\beta_0+\beta_3 = 7.99$	$\beta_0+\beta_3+\beta_4+\beta_7 = 5.77$