BST 224 HW4

Bohao Zou, 917796070
BST 224
University of california, Davis

June 5, 2020

Question 1

(a) Solution

Please read the R code.

(b) Solution

The mixed linear model of this question is :

$$Y_{ij} = \beta_0 + \beta_1 treatment_i + \beta_2 time_{ij} + \beta_3 time_{ij} * treatment_i + U_{i1} + U_{i2} time_{ij} + Z_{ij} \quad (1)$$

where $U_{i1} \sim \mathcal{N}(0, \nu_1), U_{i2} \sim \mathcal{N}(0, \nu_2)$ and $Z_{ij} \sim \mathcal{N}(0, \tau)$. Those three coefficients are all random variables and its are all independent with each other. The coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ are all fixed variables. The random effects in this model is U_{i1} and U_{i2}. The U_{i1} is the random intercept of subject i and the U_{i2} is the random slope of the *timeday* variable of subject i.

The estimated variance of random intercept U_{i1} is $\text{Var}(U_{i1}) = 9.953$.
The estimated variance of random slope U_{i2} is $\text{Var}(U_{i2}) = 0.0343$.

(c) Solution

$$H_0 : \text{Var}(U_{i2}) = 0 \ \textit{and} \ \ \text{Cov}(U_{i1}, U_{i2}) = 0$$
$$H_1 : \text{Var}(U_{i2}) > 0 \ \textit{with a possibility that} \ \ \text{Cov}(U_{i1}, U_{i2}) \neq 0$$

The null hypothesis is on the boundary of the parameter space. The asymptotic distribution for testing this random effect is $0.5\chi_1^2 + 0.5\chi_2^2$. It is equivalent to using χ_2^2 and divide p-value by 2.

The log likelihood of full model which was calculated by ReML method is -409.2698.
The log likelihood of reduced model which was calculated by ReML method is -440.6052.

The value of $\frac{1}{2}P_r(\chi^2 \geq 31.3354; Df = 2) = 7.8445 * 10^{-8}$. It is so small that we can

treat it as zero. So, the result is significant under 0.05 level. We can reject H_0 and accept H_1.

The conclusion of this question is that we can not only use one random variable in this model. It is not defensible with only randomly varying intercept in this model.

(d) Solution

The first problem is how to deal the NA value in the observed data set.

From the data set we have found that in the *treatment* = 2, the quantile of *measures* of the 24-th subject is always the 100% (i.e. the max value) in the same time. From this we can guess he has the best athletic ability in the whole people who are participated in program 2.

For deal the NA value of this subject, we build two linear model which used all the data that the *timeday* variable is 10 and 12. The model can be expressed :

$$Y_i = \beta_0 + \beta_1 treatment_i + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This model dose not contain the influence of time because all the *measures* are in the same time.

For the *timeday* = 10 model, the $\beta_0 = 81.077$, $\beta_1 = 1.276$ and $\sigma = 3.342$. Based on the above guess, we add a value which is the 99% quantile of Normal distribution $\sim \mathcal{N}(0, \sigma = 3.342)$. The final value is $81.077 + 1.276 + 7.774655 = 90.12765$.

For the *timeday* = 12 model, the $\beta_0 = 81.0667$, $\beta_1 = 1.4667$ and $\sigma = 3.677$. Based on the above guess, we add a value which is the 99% quantile of Normal distribution $\sim \mathcal{N}(0, \sigma = 3.677)$. The final value is $81.0667 + 1.4667 + 8.553981 = 91.08738$.

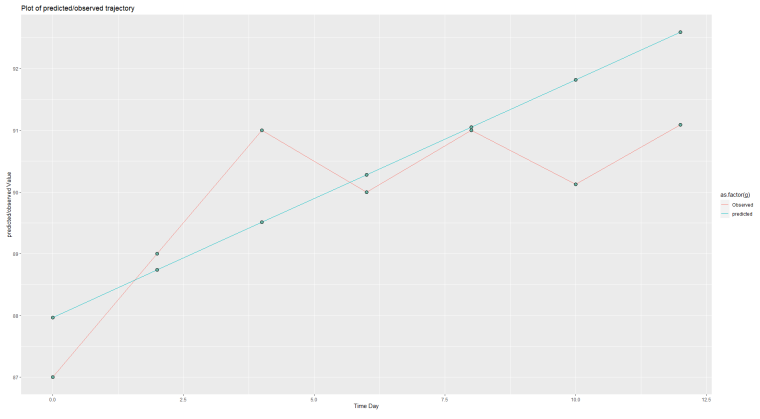

Figure 1: *The plot of predict / observed trajectory.*

Q2

(a) Solution

The sample mean and variance of the bacilli count for each treatment group for baseline and post-treatment visits are showed in the table.

	Sample Mean	Sample Variance	Ratio of Mean / Variance
Drug A, Visit 0	9.3	22.67	0.41
Drug A, Visit 1	5.3	21.57	0.24
Drug B, Visit 0	12.3	51.12	0.24
Drug B, Visit 1	6.1	37.88	0.16
Drug C, Visit 0	12.9	15.65	0.82
Drug C, Visit 1	12.3	51.12	0.24

Table 1: *The table of sample mean and sample variance of each treatment group for baseline and post-treatment.*

Poisson model assumes that the variance is the same as the mean if no overdispersion. However, from the table we can know that the ratios of mean dividing variance are not approximating with 1. Moreover, its are more smaller than 1. So, overdispersion parameter is necessary for describing this data set if a Poisson model is considered.

(b) Solution

(i)

The mean model is :

$$\mathrm{E}(Y_{ij}|\vec{x}_{ij}) = \mu_{ij} = e^{\eta_{ij}}$$

where η_{ij} is

$$\eta_{ij} = \beta_0 + \beta_1 Drug\_A_{ij} + \beta_2 Drug\_B_{ij} + \beta_3 visit_{ij} + \beta_4 Drug\_A_{ij} * visit_{ij} + \beta_5 Drug\_B_{ij} * visit_{ij}$$

The variance model is :

$$\mathrm{Var}(\vec{Y_i}|X_i) = V_i,$$

where the V_i is

$$V_i = \phi A_i^{\frac{1}{2}} C_i A_i^{\frac{1}{2}}$$

where ϕ is the dispersion parameter of variance. The matrix A_i is a diagonal matrix which the elements of diagonal are

$$(v(\mu_{i1}), v(\mu_{i2}) \dots v(\mu_{in_i}))^T$$

$n_i = 2$ at this data set. The $v(\cdot)$ means the variance function. In this possion regression model, $v(x) = x$. The matrix C_i is the variance-covariance matrix of exchangeable correlation model for i-th subject.

(ii)

The estimated $\hat{\vec{\beta}}$ is showed in the table below.

	Estimated Value
β_0	2.56
β_1	-0.33
β_2	-0.25
β_3	-0.05
β_4	-0.51
β_5	-0.45

Table 2: *The estimated $\hat{\vec{\beta}}$ of this GEE model.*

- β_0: The mean log rate ratio of baseline for subjects who will receive Drug C but do not receive any treatments at present.
- $\beta_0 + \beta_1$: The mean log rate ratio of baseline for subjects who will receive Drug A but do not receive any treatments at present.
- $\beta_0 + \beta_2$: The mean log rate ratio of baseline for subjects who will receive Drug B but do not receive any treatments at present.
- β_3: The mean of changed log rate ratio for subjects who have received Drug C.
- $\beta_3 + \beta_4$: The mean of changed log rate ratio for subjects who have received Drug A.
- $\beta_3 + \beta_5$: The mean of changed log rate ratio for subjects who have received Drug B.

(iii)

At this question, we need to test whether there is a treatment group difference at the baseline. At this case, the effect of *"Drug C"* is merged in the interception β_0. So, we need to test weather there is a different between $\beta_0, \beta_1, \beta_2$.

The null hypothesis and alternative hypothesis are showed below.

$$H_0 : \beta_0 = \beta_1 = \beta_2$$
$$H_1 : \beta_0 \neq \beta_1 \ or \ \beta_0 \neq \beta_2$$

The L matrix of this testing is showed below.

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}$$

By using the formula $T = (L\hat{\vec{\beta}})^T (L\hat{C}L)^{-1} (L\hat{\vec{\beta}}) \sim \chi^2_{df}$, where $\hat{C} = \text{Var}(\hat{\vec{\beta}})$, $DF = rank \ of \ L$, we can get the conclusion.

After calculation, the statistic $T = 177$, The $DF = 2$. So, the p-value of this test is $P\_value = P(\chi^2_2 > T) = 3.67 \times 10^{-39}$. This indicates that we need to reject H_0 at the level $\alpha = 0.05$. We can get the conclusion that there is a treatment group difference at the baseline.

(iv)

The mean model in this case is :

$$\mathrm{E}(Y_{ij}|\vec{x}_{ij}) = \mu_{ij} = e^{\beta_0+\beta_1 visit_{ij}+\beta_2 Drug\_A_{ij} visit_{ij}+\beta_3 Drug\_B_{ij} visit_{ij}}$$

The estimated value of this model is showed below.

	Estimated Value	Std.err	Wald Test	P-Value
β_0	2.373	0.080	877.10	$\leq 2e^{-16}$
β_1	-0.003	0.157	0.00	0.985
β_2	-0.563	0.222	6.42	0.011
β_3	-0.495	0.234	4.47	0.034

Table 3: *The estimated $\hat{\vec{\beta}}$ of refit GEE model.*

The model-based standard errors are estimated by the formula

$$\mathrm{Var}(\hat{\vec{\beta}}) = (\sum_i D_i^T \hat{V_i}^{-1} D_i)^{-1}$$

where $D_i = \frac{\partial \vec{\mu_i}}{\partial \vec{\beta}}$, $\hat{V_i}$ is the estimated variance-covariance matrix for subject i. It is based on the correlation model for V_i. If we do not find a correct correlation model for this data set, those variances are wrong.

The robust standard errors are estimated by the formula

$$\mathrm{Var}(\hat{\vec{\beta}}) = (\sum_i D_i^T \hat{V_i}^{-1} D_i)^{-1} (\sum_i D_i^T \hat{V_i}^{-1} (\hat{\vec{\epsilon_i}} \hat{\vec{\epsilon_i}}^T) \hat{V_i}^{-1} D_i) (\sum_i D_i^T \hat{V_i}^{-1} D_i)^{-1}$$

where $\hat{\vec{\epsilon_i}} = \vec{Y_i} - \hat{\mu_i}$. It is the residuals of fitted model. This estimated variances are not dependent on choosing a correct correlation model. It is more robust than the model-based standard errors if we choose an error correlation model. Moreover, simpler correlation model is better to estimate the truth standard error by using this method.

In my consideration, I would prefer the robust standard errors. Because in the truth situation, we can not choose a prefect correct correlation model for our model and in general we may have large samples. In those more general cases, it is better to choose the robust standard errors to estimated the variance of our coefficients.

The interpretation of β is showed below:

- β_0 : The mean log rate ratio of baseline for subjects who do not receive any treatments.
- β_1 : The mean of changed log rate ratio for subjects who have received the Drug C versus a subject who dose not receive any treatment.
- $\beta_1 + \beta_2$: The mean of changed log rate ratio for subjects who have received the Drug A versus a subject who dose not receive any treatment.
- $\beta_1 + \beta_3$: The mean of changed log rate ratio for subjects who have received the Drug B versus a subject who dose not receive any treatment.

Based on the R output of table 3, we can get the conclusion that the effect of drug C ($\hat{\beta}_1$) is not significant under $\alpha = 0.05$ level. Because the $\hat{\beta}_1$ is not significant, we can consider the truth $\beta_1 = 0$. In the next step, we need to test β_2 and β_3. From the table 3 we can know that the $\hat{\beta}_2$ and $\hat{\beta}_3$ is significant under $\alpha = 0.05$ level. So, we can get the conclusion that the effect of drug A and drug B is significant under $\alpha = 0.05$.

(v)

The estimate of the dispersion parameter $\phi = 3.21$. The estimated working correlation parameter $\alpha = 0.738$. Based on those estimated parameters, we can get the conclusion that the overdispersion parameters is necessary for this data set and it approximates equal with the mean of *Ratio of samples means and samples variances* which showed in table 1. The correlation parameter $\alpha = 0.738$ gives us confident that this can not be treated as cross-sectional data model, it has correlation between pre-treatment and post-treatment.

(c) Solution

The estimated coefficient by using generalized linear mixed effect model is showed below.

	GEE Estimated Value	GLMM Estimated Value
β_0	2.373	2.242
β_1	-0.003	0.003
β_2	-0.563	-0.606
β_3	-0.495	-0.523

Table 4: *Compare the coefficients which estimated by using different model.*

In my expectation, there do not exist disagreement between the estimators in marginal and conditional models. The U_i is a random variable $\sim \mathcal{N}(0, \nu^2)$.

For log-linear model with random intercepts, all parameters for marginal model except the intercept will have the same value and interpretation as in random intercept model. It has the relationship $\beta_0^* = \beta_0 + \frac{\nu^2}{2}$ and $\beta_k^* = \beta_k$ if $(k > 0)$.