BST 224 Final Project

Bohao Zou, 917796070
University of California, Davis

June 10, 2020

1 Question (a), Solution

In this question, we need to explore the relationship of covariates and response variable in this data set. Because the response variable in this data set is a binary variable, we can explore the relation of count of response variable with other covariates.

At first, for checking the distribution of baseline covariate, I select the two x-covariates *Baseline Age* and *Vitamin A Deficiency*. The distributions of those two variable are show below.

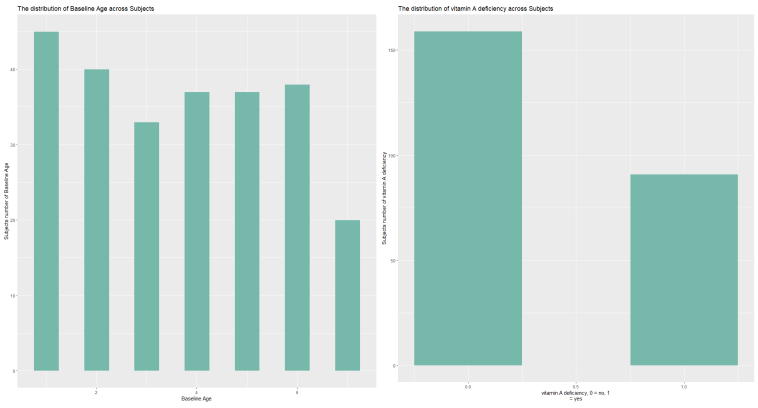

Figure 1: *The distribution of Baseline Age and Vitamin A Deficiency variables. Left is for Baseline Age and Right is for Vitamin A Deficiency*

From the left plot we can know that the number of subjects in each *Baseline Age* are approximate same except when the *Baseline Age* equals 7. This may because it is hard to collect the samples who age are 7.

From the right plot, we can know that the subjects who have no vitamin A deficiency are more than the subjects who have vitamin A deficiency. This may conduct by the limitation of circumstance that the people who have some disease will be much less than the people who have a healthy body.

In the next step, we need to explore the distribution of the response variable for different time and groups. In this project, I select four cases for exploring. The fist is *Finding*

the relation of time and infection without any group. The second is *Finding the relation of time with infection with different gender.* The third is *Finding the relation of time with infection with different baseline age.* The final exploration is *Finding the relation of time with infection with if vitamin A deficiency.* The plots are showed below.

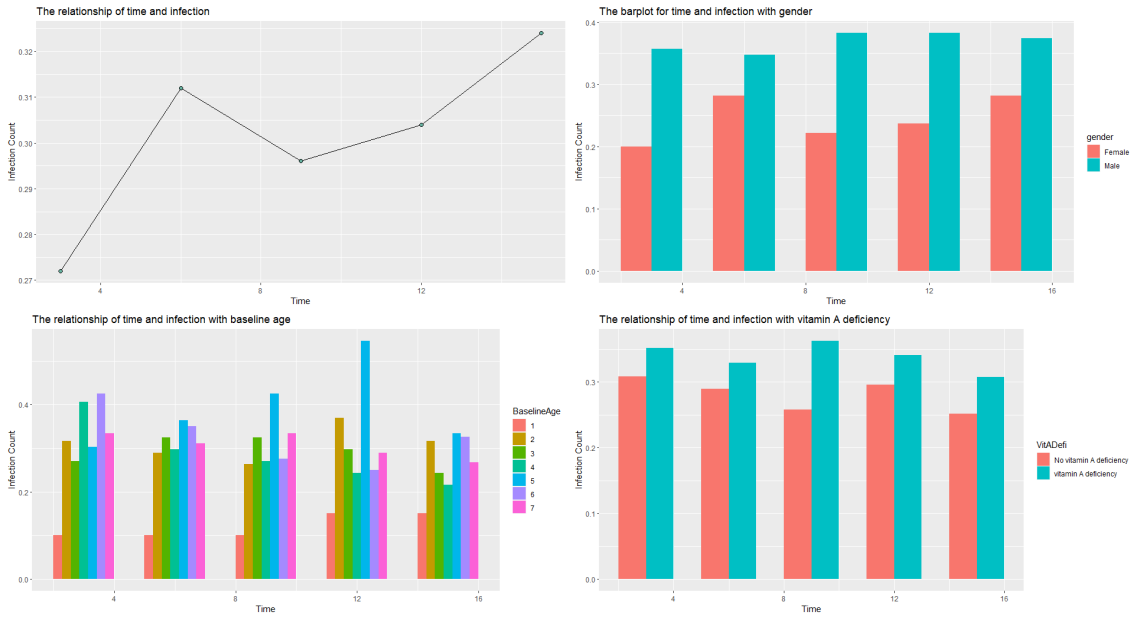

Figure 2: ***Top-Left****: The relation of time and infection without any group.* ***Top-Right****: The relation of time with infection with different gender.* ***Bottom-Left****: The relation of time with infection with different baseline age.* ***Bottom-Right****: The relation of time with infection with if vitamin A deficiency*

From the top-left plot we can rough assert that with the time passed, the risk of respiratory infection is increasing. From the top-right plot we can say that the male subjects have higher probability to infect the disease than female subjects. From the bottom-left plot, we can know that the subjects who baseline age equal 1 has the the smallest probability to infect this disease. However, the subjects who baseline age equal 5 has the highest probability to infect this disease. The infect probability in the remaining subjects are roughly same. From the bottom-right plot we can clear see that if suffering vitamin A deficiency have high linking with if a child infect this disease. The subjects who suffer vitamin A deficiency have high probability to infect this disease.

From those plot and conclusion, we can initially assert that children with Vitamin A deficiency were at increased risk for developing respiratory infection.

2 Question (b), Solution

The **Linear Predictor** of this model is

$$\begin{aligned}\eta_{ij} = \vec{x}_{ij}^T\vec{\beta} =& \beta_0 + \beta_1 time_{ij} + \beta_2 gender_i + \beta_3 vita_{ij} \\ &+\beta_4 time_{ij} * gender_i + \beta_5 time_{ij} * bage_i + \beta_6 time_{ij} * vita_{ij}\end{aligned}$$

The **Link function** of this model is

$$"log\ \ odds" = logit(\mu) = log(\frac{\mu}{1-\mu})$$

This means that

$$\mathrm{E}(Y_{ij}|\vec{x}_{ij}) = \mu_{ij} = \frac{1}{1+e^{-\eta_{ij}}}$$

The **Variance Function** of this model is

$$\begin{aligned}\phi &= 1 \\ v(\mu) &= \mu(1-\mu)\end{aligned}$$

The correlation matrix of this model in this data set is

	Time.0	Time.3	Time.6	Time.9	Time.12	Time.15
Time.0	1.00	0.57	0.44	0.44	0.51	0.44
Time.3	0.57	1.00	0.57	0.59	0.54	0.57
Time.6	0.44	0.57	1.00	0.56	0.43	0.49
Time.9	0.44	0.59	0.56	1.00	0.45	0.53
Time.12	0.51	0.54	0.43	0.45	1.00	0.52
Time.15	0.44	0.57	0.49	0.53	0.52	1.00

Table 1: *The correlation matrix*

We can also draw the correlogram plot for this model and this data set. The correlogram plot is showed below.
From the correlation matrix and correlogram plot we can know that there is a correlation between different times under the significant $\alpha = 0.05$ level. By this observation, we can not consider those data set as an cross-sectional data set. We must consider the correlation in this data set.

The criterion of how to use the ACF correlogram plot to choose a roughly correct correlation model is that

1. If ACF declines strongly with lag, then: exponential correlation model.
2. If ACF does not appear to $\to 0$ as $lag \to \infty$, then: exchangeable correlation model.
3. If ACF does not appear to $\to 1$ as $lag \to 0$, then: measurement error correlation model.

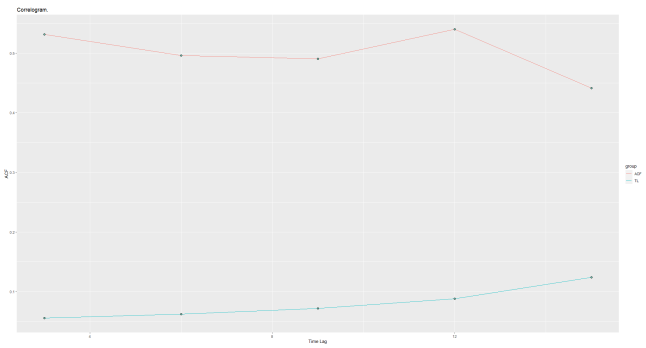

Figure 3: *The correlogram of this model. The read line represents the ACF, the blue line represent the 95% confidence interval of ACF*

From figure 3, we can know that the criterion 2 and 3 is roughly fitful for this data set. However, we can not guarantee our selected model is 100% correct for this data set and we will use GEE model in the next analysis, the GEE model should choose simpler correlation model. We can also know that the correlation between two response Y in the same subject should be roughly same base on the correlation matrix which is showed in Table 1. At the end, I select the exchangeable correlation model for next analysis.

3 Question (c), Solution

At the first stage, we need to find a basic model and modify this model in the next steps. For modifying the basic model, we need to use a criterion to decide which part should exclude an which term should be included. We can see the logistic regression as an predictive model. By using this intuition, the criterion which I used is

1. Finding the item which has the biggest P-Value and drop it. The hypothesis test which I used is χ^2 test which derives from the formula $(L\hat{\vec{\beta}})^T(L\hat{C}L^T)^{-1}(L\hat{\vec{\beta}}) \sim \chi^2_{df=1}$.
2. Calculate the F1-Score of this model with judgement threshold = 0.5.
3. If the F1-Score of this model with less parameter decreased, stop (Note: Do not stop at the first time, this is because we need to drop at least one item from basic model), Otherwise, back to the first step.
4. The best model is the last model before the stop.

This best model which we will be used is the model has high F1-Score and less items. The reason that why we should choose F1-Score but not accuracy is that this is an unbalanced data set. If we use an constant model which can only predict 0 to calculate the accuracy of this data, the accuracy of this constant model is 70.4%. So, we need use another evaluation tool to assess the result like F1-Score.

The basic models contains all single item and all interaction between those single item. It will show below.

$$\begin{aligned}\eta_{ij} = \vec{x}_{ij}^T\vec{\beta} = \beta_0 + \beta_1 time_{ij} + \beta_2 gender_i + \beta_3 vita_{ij} + \beta_4 bage \\ +\beta_5 time_{ij} * gender_i + \beta_6 time_{ij} * bage_i + \beta_7 time_{ij} * vita_{ij} \\ +\beta_8 gender_i * bage_i + \beta_9 gender_i * vita_{ij} + \beta_{10} bage_i * vita_{ij}\end{aligned}$$

Based on this basic model, I will display which item will I drop in the next series steps and present the P-Values, F1-Score of those models in a table. Beside of F1-Score, we can also use QIC to evaluate the goodness of one model. The smaller of QIC means that the model is better.

	Drop Item	P-Value	F1-Score	QIC
Basic Model	gender*vita (Drop)	0.957	0.515	1778.20
Model 2	time*gender (Drop)	0.861	0.515	1776.20
Model 3	gender (Drop)	0.775	0.515	1774.04
Model 4	time*vita (Drop)	0.618	0.515	1772.14
Model 5	bage (Drop)	0.5328	0.515	1770.25
Model 6	vita (Conserve)	0.2386	0.515	1769.55
Model 7	time*bage (Stop)	0.238	0.500	1772.76

Table 2: *The sequence of model fitting*

From this table we can get the best model. The best model is Model 6, it has the least parameters but contains the high F1-Score. The **Linear Predictor** is showed below.

$$\eta_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 vita_{ij} + \beta_3 time_{ij} * bage_i + \beta_4 gender_i * bage_i + \beta_5 bage_i * vita_{ij} \quad (1)$$

This is the final model which I selected from the basic model. The estimated coefficients and the confidence interval will be showed below. We know that the Wald test is based on the MLE and the asymptotic distribution is $\mathcal{N}(0,1)$. So, we can use the $\mathcal{N}(0,1)$ as the distribution of one coefficient.

	Estimated Value	95% Confidence Interval	P-Value
β_0	-0.78370	[-1.12, -0.444]	6.1×10^{-6}
β_1	0.036690	[0.00586, 0.0675]	0.01971
β_2	-0.53650	[-1.43, 0.356]	0.23859
β_3	-0.00553	[-0.0132, 0.00211]	0.15656
β_4	-0.17615	[-0.273, -0.0798]	0.00034
β_5	0.229440	[0.0239, 0.435]	0.02872

Table 3: *The estimated coefficients and the corresponding confidence interval.*

The estimate of exchangeable correlation model parameters fitted in my model is $\rho = 0.499$, the standard error of this parameter is 0.0588.

4 Question(d), Solution

In the question, we need to confirm that if the mean response is well-captured by the fitted mean model. By using the method that compare the final fitted model to the empirical proportion of subjects with respiratory infections. In this question, I will check linearity of *time* firstly and *baseline age* secondly. The method of checking is by plotting the empirical and predicted proportions by time or baseline age, where predicted proportion is the average of my fitted values.

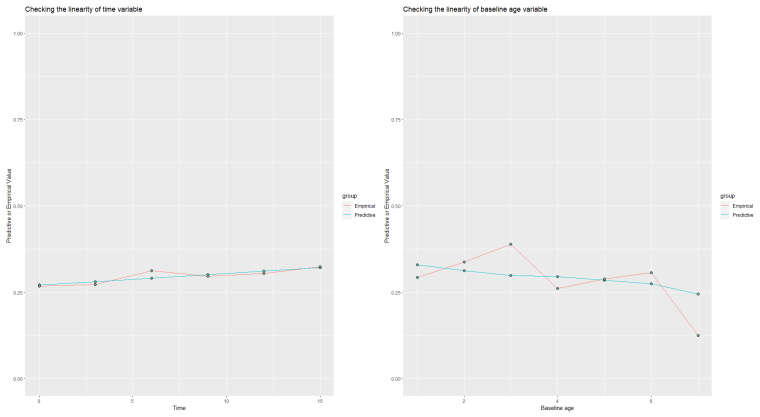

Figure 4: *The plot for checking linearity of time variable and baseline age variable. The left is for checking time. The right is for checking baseline age.*

From the left plot we can see that the probability which predicted by fitted mean model is nearly completely coincide with the empirical proportion of subjects with respiratory infections. This indicates that he mean response Y is well-captured by the fitted mean model on the *time* x-variable.

From the right plot we can see that the empirical proportion of subjects with respiratory infections is fluctuated around the probability which predicted by fitted mean model. The tendency of empirical proportion and predictive probability are all going decrease. This shows that our fitted mean model captured most of information of mean response variable Y on the *baseline age* x-variable.

5 Question (e), Solution

The **Linear Predictor** of final model is showed below.

$$\eta_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 vita_{ij} + \beta_3 time_{ij} * bage_i + \beta_4 gender_i * bage_i + \beta_5 bage_i * vita_{ij}$$

The interpretation of $\vec{\beta}$

- logistic(β_0): the mean probability of infection for the subjects who are man without vitamin A deficiency and at baseline time.

- logistic($\beta_0 + \beta_4$): the mean probability of infection for the subjects who are woman, without vitamin A deficiency at baseline time and the baseline age is 1.
- logistic($\beta_0 + 3\beta_1 + 3\beta_3$): the mean probability of infection for the subjects who are man, without vitamin A deficiency at 3 month passed and the baseline age is 1.
- logistic($\beta_0 + \beta_2 + \beta_4 + \beta_5$): the mean probability of infection for the subjects who are woman, with vitamin A deficiency at baseline time and the baseline age is 1.

The estimate of correlation model parameter fitted in my model is $\rho = 0.499$, the standard error of this parameter is 0.0588. This correlation parameter gives the correlation of two response Y on the same subject. This means that the correlation of response Y on the same subject between any time is 0.499.

From the table 3 we can know that under the significant level $\alpha = 0.05$, the coefficients which relate with vitamin A deficiency $\beta_2(vita)$ is not significant but $\beta_5(bage * vita)$ is significant. This indicates that we need to accept the hypothesis that $\beta_2 = 0$. So, we only need to care about the β_5. From the model, we know that the effect of Vitamin A deficiency on risk for respiratory infection will be determined by $\beta_5 \times bage$. This indicates that in different *baseline age*, it will have different effect on the risk for respiratory infection. The effects will be calculated by $\beta_5 \times bage$ and its will show on the table.

	Effect	Raise/Decrease
Baseline Age = 1	0.22944	Raise
Baseline Age = 2	0.45888	Raise
Baseline Age = 3	0.68832	Raise
Baseline Age = 4	0.91776	Raise
Baseline Age = 5	1.1472	Raise
Baseline Age = 6	1.37664	Raise
Baseline Age = 7	1.60608	Raise

Table 4: *The effect of Vitamin A deficiency on risk for respiratory infection*

From the table we can know that, Vitamin A deficiency will raise the risk for respiratory infection. Furthermore, under the assumption that a child without Vitamin A deficiency will have same probability to infect this disease, the child who has larger *baseline age* will have larger risk for infecting this respiratory disease.

6 Question (f), Solution

The **Linear Predictor** of conditional model is showed below.

$$\eta_{ij} = (\beta_0 + U_{i0}) + \beta_1 time_{ij} + \beta_2 vita_{ij} + \beta_3 time_{ij} * bage_i + \beta_4 gender_i * bage_i + \beta_5 bage_i * vita_{ij}$$

The $U_{i0} \sim \mathcal{N}_p(0, G^2)$ and U_{i0} is independent of X_i. The fitted conditional model (GLMM with a random intercept) with previous covariates is showed below.

	Estimated	Std Error	P-Value
β_0	-1.60977	0.34909	4×10^{-6}
β_1	0.07795	0.03227	0.01571
β_2	-1.34938	0.87766	0.12417
β_3	-0.012	0.00767	0.11426
β_4	-0.36249	0.09615	0.00016
β_5	0.53565	0.20678	0.00959
U_{i0}	Var 7.28	2.7	NA

Table 5: *The estimated coefficients and the corresponding P-Value of fitted conditional model with a random intercept.*

The interpretation of $\vec{\beta}$

- logistic$(\beta_0 + U_{i0})$: the subject i-th specific probability of infection for the i-th subject who is man without vitamin A deficiency and at baseline time.
- logistic$(\beta_0 + U_{i0} + \beta_4)$: the subject i-th specific probability of infection for the i-th subject who is woman, without vitamin A deficiency at baseline time and the baseline age is 1.
- logistic$(\beta_0 + U_{i0} + 3\beta_1 + 3\beta_3)$: the subject i-th specific probability of infection for the i-th subject who is man, without vitamin A deficiency at 3 month passed and the baseline age is 1.
- logistic$(\beta_0 + U_{i0} + \beta_2 + \beta_4 + \beta_5)$: the subject i-th specific probability of infection for the i-th subject who is woman, with vitamin A deficiency at baseline time and the baseline age is 1.
- U_{i0}: The random intercept for subject i. This includes the effect factors which are influenced with risk for respiratory infection by the subject i him/her self.

By comparing the coefficients in the GEE model and GLMM model. We can obviously find that the estimated values are different in those two models. This is because in marginal model(GEE), we construct the correlation Corr(Y_{ij}, Y_{ik}). It is from the correlation model which we selected. However, in the conditional model(GLMM), the correlation arises from U_{i0}s. Because of this reason, the estimated coefficients is different in those two models. In the end, the conclusion is that their differences in agreement with what the theory predicts.