BIOST 540 Homework 3

Ivy Zhang, Hantong Hu

05/14/2021

Responses

Part A

1. From the table and the plot, we can see that the prevalence of respiratory disease of the mother-smoking group is overall higher than the mother-non-smoking group. Furthermore, for the group of children with non-smoking mothers, the prevalence is decreasing as they grow up; for the group of children with smoking mothers, the prevalence increase from baseline to when they are 7 years old, and decrease in the following years. Estimated correlation matrix by mother smoking status are also presented below.

Table 1: Prevalence of respiratory disease and age within each of the two smoking groups

smok	age6	age7	age8	age9
0	0.160	0.149	0.143	0.106
1	0.166	0.209	0.187	0.139

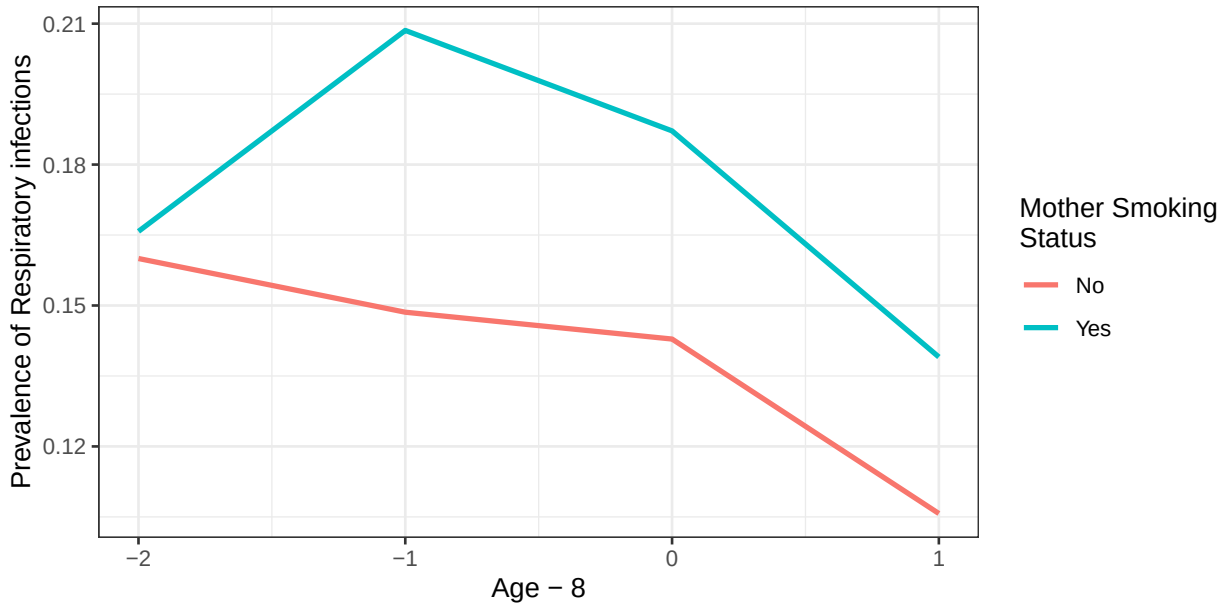

Table 2: correlation for overall

	age6.resid	age7.resid	age8.resid	age9.resid
age6.resid	1.000	0.354	0.308	0.327
age7.resid	0.354	1.000	0.441	0.327
age8.resid	0.308	0.441	1.000	0.379
age9.resid	0.327	0.327	0.379	1.000

Table 3: correlation for mother non-smoking group

	age6.resid	age7.resid	age8.resid	age9.resid
age6.resid	1.000	0.344	0.290	0.306
age7.resid	0.344	1.000	0.426	0.327
age8.resid	0.290	0.426	1.000	0.391
age9.resid	0.306	0.327	0.391	1.000

Table 4: correlation for mother smoking group

	age6.resid	age7.resid	age8.resid	age9.resid
age6.resid	1.000	0.373	0.339	0.361
age7.resid	0.373	1.000	0.462	0.326
age8.resid	0.339	0.462	1.000	0.362
age9.resid	0.361	0.326	0.362	1.000

2.

a. In the three models, we estimate that the odds of respiratory diseases is $e^{0.289} = 1.34$ in the autoregressive correlation matrix model, and $e^{0.314} = 1.37$ for the other two models, times higher among the group of children with smoking mothers compared to group with non-smoking mothers at the age of 8.

   We can see from the correlation matrices in question 1 that for all ages, the indicator of respiratory diseases are positively correlated; observations that are closer in time tend to be more correlated with each other. Therefore, it seems to be more suitable if we use the autoregressive correlation matrices since independence correlation assumes no correlation between the observations and exchangeble assumes correlation between any two measurements is the same regardless of time interval, and autoregressive assumption is more similar to this case.

b. We do not have statistically siginificant evidence in the interaction between smoking and age at the significance level of 0.05 in all three models (p>0.3). For the interaction term, we interpret this as the odds ratio of getting respiratory infection among group of children with smoking mothers increases by 0.084 for AR matrix (and 0.071 for the other two) for each additional year.

Table 5: Estimated coefficients (marginal, AR)

	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.925	0.121	254.312	0.000
smok	0.289	0.191	2.277	0.131
age	-0.148	0.060	6.101	0.014
smok:age	0.084	0.092	0.831	0.362

Table 6: Estimated coefficients (marginal, Independence)

	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.901	0.119	254.823	0.000
smok	0.314	0.188	2.794	0.095
age	-0.141	0.058	5.888	0.015
smok:age	0.071	0.088	0.644	0.422

Table 7: Estimated coefficients (marginal, Exchangeable)

	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.900	0.119	254.686	0.000
smok	0.314	0.188	2.791	0.095
age	-0.141	0.058	5.889	0.015
smok:age	0.071	0.088	0.644	0.422

3. When we are applying the conditional model, our model looks like this:

$$logit(E(resp_{ij}|smoke_i, b_i)) = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_i + \beta_3 \text{age}_{ij} \times \text{smoke}_i + b_{0i}$$

In this model, we assume that random effects (b_i) vary independently for each individual according to some distribution, and the responses for a single individual are independent observations from a distribution belonging to the exponential family, in this case, Bernoulli.

We estimate that, if two patients have the same random effect ($b_{0i} = b_{0i'}$), the estimated coefficient for mothers' smoking status (0.462) is the difference in log odds comparing an individual with smoking mother to another with non-smoking mother at age=8.

Table 8: Estimated coefficients (conditional)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.128	0.223	-14.039	0.000
smok	0.462	0.286	1.618	0.106
age	-0.216	0.087	-2.500	0.012
smok:age	0.105	0.138	0.761	0.447

4. When we are applying the transition model, our model looks like this:

$$logit(E[resp_{ij}|X_i, resp_{i,j-1}]) = \beta_0 + \beta_1 \text{age}_j + \beta_2 \text{smoke}_i + \beta_3 \text{resp}_{i,j-1}$$

We are using robust standard error since it does not require first-order Markov assumption to be valid. In this model, we assume observations of an invidual in different times are independent.

We estimate that, for children at the same age of observation and who doesn't have respiratory infection in the previous follow-up, the odds of having respiratory infection is $e^{0.296} = 1.34$ times higher among the group of children with smoking mothers compared to group with non-smoking mothers.

Table 9: Estimated coefficients (transition)

	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-2.478	0.117	445.539	0.000
age	-0.243	0.090	7.290	0.007
smok	0.296	0.155	3.661	0.056
lag1	2.211	0.187	139.878	0.000

5. The estimated coefficients for mothers' smoking status from the marginal model (for either of the working matrices) and the transition model are close, but the coefficient estimated in the conditional model is almost 1.5 times higher. Some potential problems in their interpretations are, 1) the first two models interpreted the coefficient based on comparing children at age 8, but the last model interpreted based on comparing children at the same age, 2) the first two model interpretations are different since one is population-level and one is individual-level, 3) the third model interpretation adds a condition of previous follow-up result.

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = TRUE)
### ------------------------------------------------------------
### Reading in the data.
library(tidyverse)
sixcity <- read.csv("C:/Users/second/Desktop/BIOST 540/Data Set/sixcity.csv")
### ------------------------------------------------------------
### Q1
library(geepack)
library(reshape2)
library(lme4)
library(dplyr)

sixcity$X <- NULL
sixcity$aXs <- NULL
sixcity_wide = reshape(sixcity, direction = "wide", idvar = c("id","smok"), timevar = "age")

prev_table <- sixcity_wide %>% group_by(smok) %>%
  summarise(age6 = mean(`resp.-2`), age7=mean(`resp.-1`),
            age8=mean(resp.0), age9=mean(resp.1))
kable(round(prev_table,3),
      caption = "Prevalence of respiratory disease and age within each of the two smoking groups")

p <- ggplot(data = sixcity, aes(x = age, y = resp, group = id, col = factor(smok)))
p + geom_line(data = sixcity %>% group_by(smok, age) %>% summarise(value=mean(resp)),
            aes(x = age, y = value, group = smok), size=1) +
  theme_bw() + ylab("Prevalence of Respiratory infections") + xlab("Age - 8") +
  scale_color_discrete(name="Mother Smoking\nStatus", labels=c("No", "Yes"))

sixcity_cor <- merge(sixcity_wide, prev_table, by="smok")

sixcity_cor$age6.resid <- sixcity_cor$`resp.-2` - sixcity_cor$age6
sixcity_cor$age7.resid <- sixcity_cor$`resp.-1` - sixcity_cor$age7
sixcity_cor$age8.resid <- sixcity_cor$resp.0 - sixcity_cor$age8
sixcity_cor$age9.resid <- sixcity_cor$resp.1 - sixcity_cor$age9

kable(round(cor(sixcity_cor[,c("age6.resid","age7.resid", "age8.resid", "age9.resid")]),3),
      caption = "correlation for overall")
corr_summary = by(sixcity_cor[,c("age6.resid", "age7.resid", "age8.resid", "age9.resid")],
                  INDICES = sixcity_cor$smok, FUN=cor)
kable(round(corr_summary[[1]],3), caption = "correlation for mother non-smoking group")
kable(round(corr_summary[[2]],3), caption = "correlation for mother smoking group")
### ------------------------------------------------------------
### Q2
#---Marginal model using autoregression correlation construction--------
auto_model = geeglm(resp~smok*age, data = sixcity, id=id,
                    family = binomial("logit"), corstr = "ar1")
#---Marginal model using independence correlation construction---------
ind_model = geeglm(resp~smok*age, data = sixcity, id=id,
                   family = binomial("logit"), corstr = "independence")
```

```r
#---Marginal model using exchangble correlation construction-------------
exc_model = geeglm(resp~smok*age, data = sixcity, id=id,
                   family = binomial("logit"), corstr = "exchangeable")

#------------------------Display Results-------------------------------------
kable(round(coef(summary(auto_model)),3), caption = "Estimated coefficients (marginal, AR)")
kable(round(coef(summary(ind_model)),3), caption = "Estimated coefficients (marginal, Independence)")
kable(round(coef(summary(exc_model)),3), caption = "Estimated coefficients (marginal, Exchangeable)")

### ------------------------------------------------------------
### Q3
#---------------------------------------Conditional Model-----------------------------------------
condi_model_ind = glmer(resp~smok*age+(1|id), data = sixcity,
                        family = binomial(link = "logit"), nAGQ = 20)
kable(round(coef(summary(condi_model_ind)),3), caption = "Estimated coefficients (conditional)")
### ------------------------------------------------------------
### Q4
sixcity_all = sixcity
sixcity_all = sixcity_all %>% group_by(id) %>% mutate(lag1 = lag(resp, default =  NA))
sixcity_subset = sixcity_all[complete.cases(sixcity_all),]
transit.fit_gee = geeglm(resp~age+smok + lag1, data = sixcity_subset, id = id,
                         family = binomial(link = "logit"), corstr = "independence")
kable(round(coef(summary(transit.fit_gee)),3), caption = "Estimated coefficients (transition)")
### ------------------------------------------------------------
### Q5
```