STAT/BIOST 571: Homework 8

Philip Pham

March 15, 2019

Problem 1: GEE and GLMM; interpretation of marginal parameters in logistic regression models; missing data (20 points)

Download the `fluoride.csv` *dataset from the course website. This dataset contains 3846 observations of fluoride intake for 1279 children, with follow-ups at ages 1.5, 3, 6, and 9 months, but with some observations missing for individual children. The variable id indexes unique children, age denotes age in months, income is an indicator for maternal income over 30 thousand dollars per year, $fluoride$ is total fluoride intake (mg per kg of body weight), and fl is an indicator for $fluoride > 0.05$. Our primary interest is the relationship between the binary outcome fl and the child's age, potentially including effect modification by maternal income. We will fit logistic regression models for the fl outcome with the standard mean variance relationship and either a multiplicative interaction*

$$\mu = expit(\beta_0 + \beta_1 \times age + \beta_2 \times income + \beta_3 \times age \times income) \tag{1}$$

or just an intercept and a main effect

$$\mu = expit(\beta_0 + \beta_1 \times age). \tag{2}$$

In all analyses, we account for correlation within children and assume the data from different children are independent.

(a) *Fit model (1) using GEE with independence and exchangeable working correlation models and using a standard GLMM with random intercepts. Report point estimates and standard error estimates for all four regression coefficients and all three model fits in a single table (use robust standard errors for GEE and model-based versions for GLMM).*

Correlation Structure	GEE Independent		GEE Exchangeable		Mixed Model	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Coefficient						
(Intercept), β_0	0.576457	0.117464	0.524520	0.115018	1.171506	0.208960
age, β_1	-0.048729	0.017763	-0.023578	0.017634	-0.054638	0.029439
income, β_2	-0.964447	0.148575	-0.944916	0.145484	-2.083890	0.269759
age:income, β_3	0.076834	0.022462	0.061876	0.022088	0.131501	0.036506

Table 1: Model fits of Equation 1 with different correlation structures to the data in `fluoride.csv`.

Solution: The estimates and standard errors can be found in Table 1. Robust sandwich estimates were used for standard errors in the GEE model. For the Mixed Model correlation structure with random intercepts, model-based standard errors were used.

Code to fit the models is found in the Appendix.

(b) *Discuss any differences between the estimated values of β_1 from your three fitted models.*

Solution: In all three models, β_1 is negative. For those with low income, one is more likely to have low fluoride as one ages. See the second row of Table 1 for the values.

The effect is most pronounced in the mixed model which lets the intercept term vary most freely since each subject has a subject-specific adjustment. The exchangeable model encourages within-subject correlation, so it prefers that the log odds does not vary much with age.

(c) *For each of your three fitted models, write a short paragraph summarizing your main findings. Specifically, give scientifically interpretable statements (including confidence intervals) about the relationships between fluoride intake and age in children with maternal income greater than 30 thousand dollars per year and in children with maternal income less than 30 thousand dollars per year.*

	Point Estimate	95% CI lower bound	95% CI upper bound
GEE Independent	-0.04872948	-0.08354415	-0.01391481
GEE Exchangeable	-0.02357841	-0.05814036	0.01098355
Mixed Model	-0.05463762	-0.11233786	0.00306262

Table 2: Point estimates and confidence intervals for β_1, which describes how age affects fluoride intake for low-income children.

	Point Estimate	95% CI lower bound	95% CI upper bound
GEE Independent	0.02810417	0.00115686	0.05505148
GEE Exchangeable	0.03829750	0.01222844	0.06436657
Mixed Model	0.07686298	0.03462592	0.11910004

Table 3: Point estimates and confidence intervals for $\beta_1 + \beta_3$, which describes how age affects fluoride intake for children with maternal income greater than 30 thousand dollars per year.

Solution: β_1 describes the expected observed change in the log odds in children with maternal income less than 30 thousand dollars. Given a year of ageing, one would expect to observe a change of β_1. As discussed in the previous section, this quantity is negative. A 95% confidence interval can be found by using the standard error. These results can be seen in Table 2. For the GEE Exchangeable model and Mixed Model, the intervals contain 0, so only in the GEE Indepedent model is the decrease in log odds statistically significant at level $\alpha = 0.0t$. Thus, I'd conclude that the effect is quite small if it exists at all.

$\beta_1 + \beta_3$ describes the change in children with maternal income greater than 30 thousand dollars per year. This quantity has variance

$$\operatorname{var}(\beta_1 + \beta_3) = \operatorname{var}(\beta_1) + \operatorname{var}(\beta_3) + 2\operatorname{cov}(\beta_1, \beta_3), \tag{3}$$

which we can use this to calculate confidence interals. Results are in Table 3. The estimates are all positive, so the probability of higher fluoride intake increases with age for children with higher maternal incomes. None of the confidence intervals contain 0, so the effect is statistically significant at level $\alpha = 0.05$ in all three models. In the models that account for the within-subject correlation (GEE Exchangeable and Mixed Model), the estimate is larger. The Mixed Model has more parameters and better accounts for the correlation and gives the largest estimate.

(d) *Now repeat part (a), but use model (2) instead of (1) (there are now only two regression coefficients to report per model).*

Correlation Structure	GEE Independent		GEE Exchangeable		Mixed Model	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Coefficient						
(Intercept), β_0	-0.024537	0.070636	-0.059908	0.068872	-0.126181	0.114182
age, β_1	-0.002281	0.010757	0.015402	0.010418	0.028707	0.015460

Table 4: Model fits of Equation 2 with different correlation structures to the data in `fluoride.csv`.

Solution: The point estimates and standard errors of fitting the model in Equation 2 can be found in Table 4. Now β_1 is close to 0 in all three models. Two of the models are slightly greater than 0, and one is slightly less than 0..

(e) *Download the dataset `fluoride.miss.csv` from the course website and repeat the calculations from part (d). Note `fluoride.miss.csv` is a subset of `fluoride.csv`, with more missing data.*

Correlation Structure	GEE Independent		GEE Exchangeable		Mixed Model	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Coefficient						
(Intercept), β_0	-0.165919	0.077793	-0.172837	0.075095	-0.362533	0.124891
age, β_1	0.008154	0.012001	0.021671	0.011511	0.041221	0.017344

Table 5: Model fits of Equation 2 with different correlation structures to the data in `fluoride.miss.csv`.

Solution: The results of fitting the model in Equation 2 to `fluoride.miss.csv` can be found in Table 5. Now all three estimates of β_1 are positive.

(f) *Discuss the differences between your results in parts (d) and (e). Speculate about the missingness mechanism that gave rise to the `fluoride.miss.csv` dataset and explain how this might account for what you observe. You might find it helpful to conduct exploratory analyses of the two datasets and to consider your findings from part (a) of this problem.*

Solution: In part (d), the estimates are close to 0. In part (e), the estimates are positive. Only the Mixed Model estimate is statistically significant, however.

Recall that children with higher maternal incomes were more likely to see an increase in fluoride intake with age. Indeed, see Figure 1. In `fluoride.miss.csv`, we are missing records of children with low maternal income, so the estimate of β_1 is closer to that of children with higher maternal income, that is, more positive.

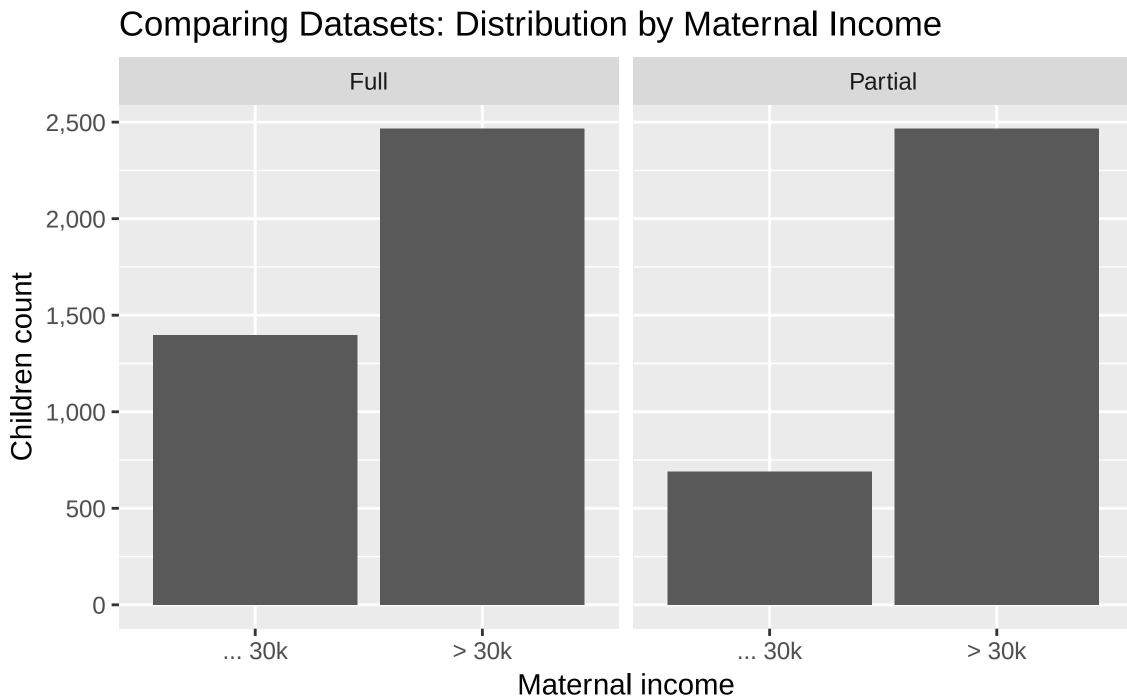

Figure 1: Number of children by maternal income in `fluoride.csv` (Full) versus `fluoride.miss.csv` (Partial).

Appendix

Code to fit the models is attached on the following pages.

GEE and GLMM; interpretation of marginal parameters in logistic regression models; missing data

We'll fit models with general estimating equations (`gee`) and general linear mixed models (`lme4`).

In [1]:
```
library(data.table)
library(gee)
library(ggplot2)
library(lme4)
library(scales)
library(tools)
library(xtable)
```

```
Loading required package: Matrix
```

Fluoride Data

In [2]:
```
head(fluoride.data <- data.table(read.csv('fluoride.csv'), key='id'))
```

id	age	income	fluoride	fl
2	3.0	1	0.00000000	FALSE
2	6.0	1	0.05063998	TRUE
2	9.0	1	0.04779446	FALSE
3	1.5	0	0.11742604	TRUE
3	3.0	0	0.08832044	TRUE
3	6.0	0	0.06216184	TRUE

In [3]:
```
summary(fluoride.data)
```

```
       id             age            income          fluoride       
 Min.   :   2   Min.   :1.500   Min.   :0.0000   Min.   :0.000000  
 1st Qu.: 444   1st Qu.:1.500   1st Qu.:0.0000   1st Qu.:0.008185  
 Median : 934   Median :3.000   Median :1.0000   Median :0.048175  
 Mean   : 929   Mean   :4.675   Mean   :0.6382   Mean   :0.067876  
 3rd Qu.:1409   3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:0.104724  
 Max.   :1886   Max.   :9.000   Max.   :1.0000   Max.   :1.794320  
     fl         
 Mode :logical  
 FALSE:1966     
 TRUE :1898     
```

Fluoride Data with Missing Entries

In [4]:
```
head(fluoride.miss.data <- data.table(read.csv('fluoride.miss.csv'), key='id'))
```

id	age	income	fluoride	fl
2	3.0	1	0.00000000	FALSE
2	6.0	1	0.05063998	TRUE
2	9.0	1	0.04779446	FALSE
3	3.0	0	0.08832044	TRUE
3	6.0	0	0.06216184	TRUE
4	1.5	1	0.03531871	FALSE

In [5]:
```
summary(fluoride.miss.data)
```

```
       id              age             income          fluoride       
 Min.   :   2.0   Min.   :1.500   Min.   :0.0000   Min.   :0.000000  
 1st Qu.: 485.0   1st Qu.:3.000   1st Qu.:1.0000   1st Qu.:0.006707  
 Median : 975.0   Median :3.000   Median :1.0000   Median :0.042219  
 Mean   : 954.9   Mean   :4.709   Mean   :0.7811   Mean   :0.064560  
 3rd Qu.:1431.0   3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:0.100249  
 Max.   :1886.0   Max.   :9.000   Max.   :1.0000   Max.   :1.794320  
     fl         
 Mode :logical  
 FALSE:1679     
 TRUE :1478     
```

Models

General Estimating Equations (GEE)

In [6]:
```
gee.age.independent <- gee(fl ~ age, id=id,
                           family=binomial,
                           data=fluoride.data)
gee.age.exchangeable <- update(gee.age.independent, corstr='exchangeable')
gee.interaction.independent <- update(gee.age.independent, formula=.~. + income + age:income)
gee.interaction.exchangeable <- update(gee.interaction.independent, corstr='exchangeable')
```

```
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

 (Intercept)          age 
-0.024537225 -0.002280917 

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

 (Intercept)          age 
-0.024537225 -0.002280917 

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

(Intercept)         age      income  age:income 
 0.57645733 -0.04872948 -0.96444671  0.07683365 

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

(Intercept)         age      income  age:income 
 0.57645733 -0.04872948 -0.96444671  0.07683365 
```

General Linear Mixed Models (GLMM)

In [7]:
```
glmm.age <- glmer(fl ~ age + (1|id), family=binomial, data=fluoride.data)
glmm.interaction <- update(glmm.age, formula=.~. + income + age:income)
```

Missing Data and GEE

In [8]:
```
gee.age.independent.miss <- update(gee.age.independent, data=fluoride.miss.data)
gee.age.exchangeable.miss <- update(gee.age.exchangeable, data=fluoride.miss.data)
gee.interaction.independent.miss <- update(gee.interaction.independent, data=fluoride.miss.data
)
gee.interaction.exchangeable.miss <- update(gee.interaction.exchangeable, data=fluoride.miss.da
ta)
```

```
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

 (Intercept)          age
-0.165918518  0.008153993

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

 (Intercept)          age
-0.165918518  0.008153993

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

(Intercept)         age      income  age:income
 0.59166829 -0.05778955 -0.97965768  0.08589372

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate

(Intercept)         age      income  age:income
 0.59166829 -0.05778955 -0.97965768  0.08589372
```

Missing Data and GLMM

In [9]:
```
glmm.age.miss <- update(glmm.age, data=fluoride.miss.data)
glmm.interaction.miss <- update(glmm.interaction, data=fluoride.miss.data)
```

Estimates and Standard Errors

In [10]:
```
summarize.model <- function(model) {
    coefficients <- summary(model)$coefficients
    standard.error <- if (is(model, 'gee')) {
        coefficients[,'Robust S.E.']
    } else if (is(model, 'glmerMod')) {
        coefficients[,'Std. Error']
    }
    data.frame(coefficient=row.names(coefficients),
               estimate=coefficients[,'Estimate'],
               standard.error=standard.error,
               row.names=NULL)
}
```

In [11]:
```
key.model <- function(model) {
    data.frame(
        correlation.structure=if (is(model, 'gee')) {
            if (is.null(getCall(model)$corstr)) {
                'GEE Independent'
            } else {
                paste('GEE', toTitleCase(getCall(model)$corstr))
            }
        } else if (is(model, 'glmerMod')) {
            'Mixed Model'
        },
        has.interaction=nrow(summary(model)$coefficients) == 4,
        is.missing=getCall(model)$data == quote(fluoride.miss.data)
    )
}
```

In [12]:
```
model.summaries <- do.call(rbind, lapply(list(
    gee.age.independent, gee.age.exchangeable, glmm.age,
    gee.interaction.independent, gee.interaction.exchangeable, glmm.interaction,
    gee.age.independent.miss, gee.age.exchangeable.miss, glmm.age.miss,
    gee.interaction.independent.miss, gee.interaction.exchangeable.miss, glmm.interaction.miss
),
       function(model) {
           cbind(key.model(model), summarize.model(model))
       }))

write.csv(model.summaries, file='model_summaries.csv', row.names=FALSE)
```

In [13]:
```
data.frame(list(a=c(p=1), b=c(p=2), c=c(p=3)))
```

	a	b	c
p	1	2	3

Confidence Intervals

In [14]:
```
make.intervals <- function(model, indicator, confidence=0.95) {
    beta <- if (is(model, 'gee')) {
        coef(model)
    } else if (is(model, 'glmerMod')) {
        fixef(model)
    }
    sigma <- if (is(model, 'gee')) {
        model$robust.variance
    } else if (is(model, 'glmerMod')) {
        vcov(model)
    }
    sigma <- sqrt(as.numeric(t(indicator) %*% sigma %*% indicator))
    z <- qnorm((1 - confidence)/2)
    estimate <- beta %*% indicator
    c(`Point Estimate`=estimate,
      `95\\% CI lower bound`=z*sigma + estimate,
      `95\\% CI upper bound`=-z*sigma + estimate)
}
```

In [15]:
```
interaction.models <- list(
    `GEE Independent`=gee.interaction.independent,
    `GEE Exchangeable`=gee.interaction.exchangeable,
    `Mixed Model`=glmm.interaction)

(beta.1.intervals <- t(data.frame(
    lapply(interaction.models, make.intervals, indicator=c(0, 1, 0, 0)), check.names=FALSE)))
(beta.1.3.intervals <- t(data.frame(
    lapply(interaction.models, make.intervals, indicator=c(0, 1, 0, 1)), check.names=FALSE)))
```

	Point Estimate	95\% CI lower bound	95\% CI upper bound
GEE Independent	-0.04872948	-0.08354415	-0.013914806
GEE Exchangeable	-0.02357841	-0.05814036	0.010983553
Mixed Model	-0.05463762	-0.11233786	0.003062619

	Point Estimate	95\% CI lower bound	95\% CI upper bound
GEE Independent	0.02810417	0.001156858	0.05505148
GEE Exchangeable	0.03829750	0.012228442	0.06436657
Mixed Model	0.07686298	0.034625919	0.11910004

In [16]:
```
print(xtable(beta.1.intervals,
             caption=paste(
                 'Point estimates and confidence intervals for $\\beta_1$,',
                 'which describes how age affects fluoride intake for low-income',
                 'children.'),
             label='tab:beta_1_intervals',
             digits=c(0, 8, 8, 8)), booktabs=TRUE,
      sanitize.colnames.function=identity,
      sanitize.rownames.function=identity,
      size='small',
      file='beta_1_intervals.tex')
```

In [17]:
```
print(xtable(beta.1.3.intervals,
             caption=paste(
                 'Point estimates and confidence intervals for $\\beta_1 + \\beta_3$,',
                 'which describes how age affects fluoride intake for children',
                 'with maternal income greater than 30 thousand dollars per year.'),
             label='tab:beta_1_3_intervals',
             digits=c(0, 8, 8, 8)), booktabs=TRUE,
      sanitize.colnames.function=identity,
      sanitize.rownames.function=identity,
      size='small',
      file='beta_1_3_intervals.tex')
```

In [19]:
```
options(warn=-1)
pdf('dataset_comparison.pdf', width=6, height=3.75)
ggplot(rbind(cbind(Dataset='Full', fluoride.data),
             cbind(Dataset='Partial', fluoride.miss.data)) +
  geom_bar(aes(x=factor(income, labels = c('\u2264 30k', '> 30k')))) +
  facet_wrap(~Dataset) +
  scale_y_continuous('Children count', label=comma) +
  scale_x_discrete('Maternal income') +
  ggtitle('Comparing Datasets: Distribution by Maternal Income')
dev.off()
options(warn=0)
```

png: 2