571HW4

Coco_Luo

2023-02-16

Problem 1

1.1

I analyzed this data set using a random intercept logistic mixed model by assuming a normally distributed random intercept, the model is shown below.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: wheezing ~ age + maternal_smoking + +age:maternal_smoking + (1 |
##     id)
##    Data: df
##
##      AIC      BIC   logLik deviance df.resid
##   1599.3   1627.7   -794.7   1589.3     2143
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3995 -0.1778 -0.1589 -0.1276  2.6024
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 5.502    2.346
## Number of obs: 2148, groups:  id, 537
##
## Fixed effects:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.40171    0.27885 -12.199   <2e-16 ***
## age                   -0.21704    0.08678  -2.501   0.0124 *
## maternal_smoking1      0.47824    0.29927   1.598   0.1100
## age:maternal_smoking1  0.10465    0.13912   0.752   0.4519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) age    mtrn_1
## age          0.272
## mtrnl_smkn1 -0.442 -0.193
## ag:mtrnl_s1 -0.146 -0.621  0.280
```

1.2

In GLMMs, the βs have child-specific interpretations. For example, Beta_age is conditional on the child-specific random effects. Beta_age models the evolution of each child separately. β_{age} is the log of odds ratio if the same child was to change from unexposed to exposed.

In GEE, the βs are the average interpretation over all children. For example, β_{age} is the log of odds ratio comparing the exposed children with the unexposed children.

```
##          (Intercept)                   age     maternal_smoking1 
##           -1.9008426            -0.1412531             0.3139540 
## age:maternal_smoking1 
##            0.0708441
```

```
## 
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998) 
## 
## Model:
##  Link:                      Logit 
##  Variance to Mean Relation: Binomial 
##  Correlation Structure:     Exchangeable 
## 
## Call:
## gee(formula = wheezing ~ age + maternal_smoking + age:maternal_smoking, 
##     id = id, data = df, family = "binomial", corstr = "exchangeable")
## 
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max 
## -0.1906393 -0.1654776 -0.1468831 -0.1148906  0.8851094 
## 
## 
## Coefficients:
##                          Estimate Naive S.E.     Naive z Robust S.E.
## (Intercept)           -1.90049539 0.11871090 -16.0094430  0.11908696
## age                   -0.14123592 0.05608034  -2.5184570  0.05820089
## maternal_smoking1      0.31382583 0.18719721   1.6764450  0.18784180
## age:maternal_smoking1  0.07083185 0.08917757   0.7942788  0.08827886
##                          Robust z
## (Intercept)           -15.9588874
## age                    -2.4266968
## maternal_smoking1       1.6706922
## age:maternal_smoking1   0.8023647
## 
## Estimated Scale Parameter:  1.001273
## Number of Iterations:  1
## 
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.3543843 0.3543843 0.3543843
## [2,] 0.3543843 1.0000000 0.3543843 0.3543843
## [3,] 0.3543843 0.3543843 1.0000000 0.3543843
## [4,] 0.3543843 0.3543843 0.3543843 1.0000000
```

1.3

Below is my own code implementing a random intercept logistic mixed model assuming a normally distributed random intercept.

```r
get_density_prod = function(b, sub, beta, theta) {
  p = 1 / (1 + exp(-1 * (beta[1] + sub["age"] * beta[2] + sub["maternal_smoking"]
                         * beta[3] + b)))
  prod(p ^ sub["wheezing"] * (1 - p) ^ (1 - sub["wheezing"])) * dnorm(b, sd = theta)
}

get_neg_logL = function(params, df) {
  beta = params[1:3]
  theta = params[4]
  log_prod_all = 0
  for (m_id in unique(df$id)) {
    sub = df[df$id == m_id,]
    f_integral = integrate(get_density_prod, -10, 10, sub, beta, theta)$value
    log_prod_all = log_prod_all + log(f_integral)
  }
  -1 * log_prod_all
}

df = read.table("./sixcity.dat")
colnames(df) = c("wheezing", "id", "age", "maternal_smoking")


result = optim(c(-3, 0, 0.4, 2), get_neg_logL, NULL, df)
result$par
```

	result
Fix effects: beta_intercept	-3.0911393
Fix effects: beta_age	0.4986091
Fix effects: beta_maternal_smoking	0.7728190
Random effects: b_intercept(std.dev)	0.9221002

Problem 2

GEE and logistic mixed models with random intercepts are both used to analyze clustered or longitudinal data but there are some differences in their assumptions. Logistic mixed models with random intercepts allowed modeling of random effects which account for the correlation between repeated measures within the same cluster or individual. It can handle unbalanced data where the number of measurements per individual may differ. If can model both continuous and categorical predictors and interactions. Can provide estimates of individual level effects which are useful for understanding within subject changes over time. And provides estimates of within and between cluster variability in the outcome. However, this model assumes that the random effects are normally distributed which may not be correct. It also assumes a linear relationship between the outcome and the predictors on the log odds scale which might not be suitable for every dataset. And it can be computationally intensive especially for large datasets.

On the other hand, GEE model does not require specification of a complete model for the distribution of random effects, making it more robust to misspecification of the covariance structure. It can deal with unbalanced data where the number of measurements per individual might be different. It can mdoel both continuous and categorical predictors and interactions and provide population averaged estimates, which are useful to understand the overall trends in the data. However, it assumes that the correlation structure must be correctly specified, the outcome and predictors are linearly related and it does not account for individual level effects so cannot provide estimates of within subject changes over time. It cannot provide estimates of within and between cluster variability in the outcome because it can only give us estimates of the population averaged effects.

As a result, our decision on choosing the model depends on if there is a random effect within each individual. If so, the μ_{ij} in GEE is hard to characterize. In this case, I would choose logistic (generalized) mixed model. GEE has an advantage when we know the working correlation matrix or the matrix is unstructured. When the working correlation matrix is unstructured, the covariance matrices are different across individuals so the assumptions of logistic (generalized) mixed model don't hold.

Finally, if I have more time, I will consider on measuring a continuous variable on each individual n times longitudinally. I would also try different sample sizes, and different number of measurement times to see if the results changed.

Table 2: stimulation one LMM

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.009944	0.2206938	4.576225	4.70e-06
x1	1.656165	0.3794547	4.364592	1.27e-05

Table 3: stimulation one GEE - exchangeable

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.6922711	0.0115121	60.134388	0.0146979	47.099867
x1	0.2293371	0.0162048	14.152421	0.0162479	14.114900
x2	0.0522833	0.0160754	3.252376	0.0331968	1.574949

Table 4: stimulation two LMM

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0342346	0.2952936	0.1159342	0.9077047
x1	1.3992988	0.2308933	6.0603691	0.0000000

Table 5: stimulation two GEE - independence

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.5442837	0.0264689	20.563163	0.0335352	16.230222
x1	0.2562013	0.0373887	6.852363	0.0389793	6.572757
x2	0.1642767	0.0185729	8.844979	0.0168879	9.727458