571HW3

Coco_Luo

2023-02-08

Question 1

1.1

I fit a regular logistic regression ignoring within-subject correlation and calculate the naive SEs below:

Table 1: logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8837347	0.0838430	-22.467399	0.0000000
age	-0.1134128	0.0540820	-2.097052	0.0359889
maternal_smoking1	0.2721386	0.1234731	2.204032	0.0275221

1.2

Below is an analysis of the data set using GEE1 assuming working independence and exchangeable, and the results with the naive logistic regression.

Table 2: gee independence model

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-1.8837347	0.0838659	-22.461271	0.1142402	-16.489245
age	-0.1134128	0.0540967	-2.096481	0.0438777	-2.584749
maternal_smoking1	0.2721386	0.1235066	2.203432	0.1779818	1.529024

Table 3: gee exchangeable model

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-1.8804277	0.1148394	-16.374411	0.1138929	-16.510489
age	-0.1133850	0.0435414	-2.604073	0.0438553	-2.585434
maternal_smoking1	0.2650809	0.1770009	1.497625	0.1777465	1.491342

From the output of the three models,the results all look similar.

To interpret the coefficient estimates in naive logistic regression, for each year of change in `age`, the odds ratio of presence of wheezing over absence of wheezing is expected to change by a factor of $e^{-0.1134128}$. This suggests that the older children are, the less likely they are going to be examined with the presence of wheezing.

For each unit of change in `maternal_smoking`, the odds ratio of presence of wheezing over absence of wheezing is expected to change by a factor of $e^{0.2721386}$. This suggests that if the mother smoked during pregnancy, the more likely their children are going to be examined with the presence of wheezing.

To Interpret the coefficient estimates in GEE1 assuming working independence: for each year of change in `age`, the odds ratio of presence of wheezing over absence of wheezing is expected to change by a factor of $e^{-0.1134128}$. This suggests that the older children are, the less likely they are going to be examined with the presence of wheezing.

For each unit of change in `maternal_smoking`, the odds ratio of presence of wheezing over absence of wheezing is expected to change by a factor of $e^{0.2721386}$. This suggests that if the mother smoked during pregnancy, the more likely their children are going to be examined with the presence of wheezing.

To interpret the coefficient estimates in GEE1 assuming working exchangeable: for each year of change in `age`, the odds ratio of presence of wheezing over absence of wheezing is expected to change by a factor of $e^{-0.1133850}$. This suggests that the older children are, the less likely they are going to be examined with the presence of wheezing.

For one unit change in `maternal_smoking`, the odds ratio of presence of wheezing over absence of wheezing is expected to change by a factor of $e^{0.2650809}$. This suggests that if the mother smoked during pregnancy, the more likely their children are going to be examined with the presence of wheezing.

1.3

There is a strong within-subject correlation because we have a large alpha (2.038). The code details are in Appendix.

1.4

We can let the actual "id" be the combination of "children id" + "age". And our model would be like: `wheezing ~ age + maternal_smoking, id = id + age`.

Question 2

2.1

I stimulate 200 data sets and run GEE1 with exchangeable correlation to estimate the regression coefficients and their sandwich SEs, and the correlation parameter ρ. The stimulated data table gives me 200 rows, here I only included 15 rows from the stimulated results.

est.coef.beta0	est.coef.beta1	est.coef.beta2	SE.beta0	SE.beta1	SE.beta2	correlation.rho
-1.424	0.327	0.588	0.154	0.162	0.084	0.133
-1.445	0.457	0.509	0.155	0.165	0.083	0.151
-1.586	0.428	0.601	0.159	0.165	0.086	0.129
-1.662	0.488	0.526	0.163	0.168	0.088	0.125
-1.611	0.755	0.447	0.161	0.169	0.084	0.149
-1.381	0.576	0.488	0.154	0.166	0.081	0.173
-1.515	0.500	0.440	0.156	0.163	0.086	0.112
-1.624	0.407	0.581	0.159	0.161	0.089	0.086
-1.478	0.435	0.576	0.153	0.156	0.087	0.081
-1.350	0.445	0.331	0.157	0.173	0.080	0.206
-1.510	0.380	0.523	0.158	0.169	0.084	0.166
-1.382	0.397	0.428	0.155	0.168	0.082	0.170
-1.389	0.561	0.592	0.153	0.164	0.082	0.161
-1.609	0.580	0.581	0.160	0.166	0.085	0.142

est.coef.beta0	est.coef.beta1	est.coef.beta2	SE.beta0	SE.beta1	SE.beta2	correlation.rho
-1.681	0.336	0.577	0.167	0.177	0.085	0.192

2.2

I calculate the average regression coefficient estimates and the average correlation parameters across the 200 runs, and compare them with the true values:

Average regression coefficient estimates: -1.507872 0.498703 0.5067413

Empirical bias of coefficient estimates: -0.007872412 -0.001296956 0.006741309

Average correlation parameters: 0.1468436

Eempirical bias of correlation parameters: -0.1031564

2.3

I calculate the average sandwich SEs and compare them with the "empirical SEs". We get small biases, which shows that the sandwich estimators work well in estimating the true variation of the GEE estimates $\hat{\beta}$'s.

Average estimated SEs: 0.1577237 0.1663868 0.08411087

Empirical SEs: 0.1648745 0.1696177 0.08679951

Empirical bias: -0.007150816 -0.003230895 -0.002688635

2.4

My program allow you to simulate correlated binary data with correlation 0.75. I set rho = 0.75 and run the previous code. I get the metrics again (still keep only the first 15 rows of the stimulated data).

est.coef.beta0	est.coef.beta1	est.coef.beta2	SE.beta0	SE.beta1	SE.beta2	correlation.rho
-1.252	0.195	0.537	0.166	0.203	0.062	0.520
-1.715	0.749	0.548	0.179	0.204	0.068	0.462
-1.416	0.429	0.452	0.168	0.198	0.068	0.437
-1.677	0.488	0.550	0.182	0.212	0.066	0.522
-1.447	0.262	0.522	0.173	0.208	0.062	0.542
-1.671	0.689	0.589	0.181	0.209	0.066	0.499
-1.543	0.361	0.523	0.174	0.204	0.068	0.467
-1.552	0.788	0.360	0.179	0.211	0.063	0.519
-1.524	0.330	0.563	0.170	0.196	0.072	0.403
-1.599	0.697	0.538	0.181	0.215	0.060	0.583
-1.236	0.246	0.472	0.166	0.203	0.061	0.527
-1.723	0.746	0.519	0.186	0.215	0.064	0.542
-1.555	0.388	0.488	0.174	0.201	0.070	0.431
-1.475	0.581	0.527	0.173	0.205	0.063	0.518
-1.556	0.740	0.434	0.176	0.206	0.065	0.494

The average regression coefficient estimates: -1.51857 0.5119064 0.5017869

The empirical bias of coefficient estimates: -0.01857031 0.01190643 0.001786916

The average correlation parameters: 0.4924067

The empirical bias of correlation parameters: -0.2575933

The average estimated SEs: 0.1745627 0.2053201 0.0653917

The empirical SEs: 0.180106 0.2124638 0.0643719

The empirical bias: -0.005543229 -0.007143784 0.001019797

Question 3

3.1

Linear mixed model:

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: value ~ age + gender + bmi_baseline + (year | individuals)
##    Data: data2
##
## REML criterion at convergence: 131961.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.3862 -0.5086 -0.0131  0.4944 10.0803
##
## Random effects:
##  Groups      Name        Variance Std.Dev. Corr
##  individuals (Intercept) 1334.70  36.533
##              year          13.16   3.627   -0.24
##  Residual                 444.93  21.093
## Number of obs: 13683, groups:  individuals, 2634
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)  147.05003    5.56605  26.419
## age            1.31481    0.08961  14.672
## gender         0.64452    1.46397   0.440
## bmi_baseline   0.92574    0.16555   5.592
##
## Correlation of Fixed Effects:
##             (Intr) age    gender
## age         -0.512
## gender      -0.411 -0.056
## bmi_baselin -0.621 -0.210  0.056
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00822845 (tol = 0.002, component 1)
```

GEE:

```
##  (Intercept)          age       gender bmi_baseline         year
##  140.8017237    1.1828811    2.4021498    0.9366426    2.9833249
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
```

```
##
## Call:
## gee(formula = value ~ age + gender + bmi_baseline + year, id = individuals,
##     data = data2, family = "gaussian", corstr = "exchangeable")
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -137.075102  -28.880822   -3.704493   24.411636  456.522863
##
##
## Coefficients:
##                 Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)  140.8017237 2.85907847 49.247240  2.88476133 48.808795
## age            1.1828811 0.04500613 26.282663  0.04559274 25.944507
## gender         2.4021498 0.73265858  3.278676  0.73225367  3.280489
## bmi_baseline   0.9366426 0.08399405 11.151297  0.08583172 10.912546
## year           2.9833249 0.10543957 28.294169  0.10615152 28.104401
##
## Estimated Scale Parameter:  1811.264
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    0    0    0    0    0
## [2,]    0    0    0    0    0    0
## [3,]    0    0    0    0    0    0
## [4,]    0    0    0    0    0    0
## [5,]    0    0    0    0    0    0
## [6,]    0    0    0    0    0    0
```

I use the Framingham data again to conduct our analysis. We can see from the result that LMM allows both fixed and random effects, but GEE only allows fixed effects

A LMM allows the researchers to explain the variable level variance by the predictors in the model as it partitions the variance within and between variables. However, researcher cannot describe changes in variability in GEE model because the variability is in effect treated as a nuisance factor that is adjusted for as a covariate.

3.2

It depends on the research question, the nature of the outcome variable, the availability of missing data, and the type of correlation structure in the data. An LMM might be more appropriate if the primary interest is in estimating subject-specific effects and patterns of change over time. A GEE might be more appropriate if the focus is mainly on population-level parameters. If there is a random effect within each individual, the μ_{ij} in GEE is hard to characterize. In this case, we can choose LMM.

In LMM, the outcome variable is modeled as a function of both fixed and random effects. The fixed effects are the population-level parameters, while the random effects account for the individual-level variation and dependence within each subject. LMMs are typically used when the researcher is interested in both the population-level effects and the individual-level patterns of change over time. They allow for the estimation of random subject-specific intercepts and slopes, and they also accommodate missing data and unbalanced repeated measures designs.

On the other hand, GEE focuses on the population-level parameters and accounts for the within-subject dependence using a working correlation structure. It can handle non-normal outcomes, whereas LMMs typically assume a normal distribution. GEE is useful when the researcher is mainly interested in the

population-level parameters, and the within-subject dependence structure is of secondary importance. In addition, it has advantage if 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 LMM don't hold.

Table 6: stimulation one LMM

	Estimate	Std. Error	t value
(Intercept)	1.026044	0.0219848	46.67065
x1	1.981137	0.0314216	63.05020

Table 7: stimulation one GEE

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.9976380	0.0408708	24.409563	0.0358111	27.858344
x1	2.0534510	0.0577893	35.533401	0.0577370	35.565619
x2	0.0846621	0.0313669	2.699094	0.0804535	1.052311

Table 8: stimulation two LMM

	Estimate	Std. Error	t value
(Intercept)	-0.0874674	0.0260649	-3.355757
x1	2.0093785	0.0371921	54.027028

Table 9: stimulation two GEE

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.0003658	0.0207591	-0.0176205	0.0237813	-0.0153812
x1	2.0164460	0.0293530	68.6964058	0.0328136	61.4514710
x2	1.0117128	0.0149950	67.4697994	0.0121868	83.0171523

If I have more time, I will consider on measuring a dichotomous characteristics 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.

Problem 4

Let Y_{1ij} be the fetal weight (continuous) of ith group, jth individual, and Y_{2ij} be the fetal death (binary) of ith group, jth individual. Let Y^*_{2ij} be the underlying variable of Y_{2ij}, such that $P(Y_{2ij} = 1) = P(Y^*_{2ij} > 0)$.

Assume that (Y_{1ij}, Y^*_{2ij}) follows bivariate normal: $N(y_{1ij}, y_{2ij}, \mu, \gamma, \sigma^2, 1, \rho)$, where μ and γ are population means of Y_{1ij} and Y^*_{2ij}, σ^2 and 1 are the their variances, respectively. Note that we have normalized var(Y^*_{2ij}) to 1. ρ is the correlation coefficient. Let X_{ij} be the covariate of ith group, jth individual. X_{ij} is a $p \times 1$ vector consisting of $\{X_{bij}, X_{tij}, X_{aij}, X_{sij}\}$ with dimension of $b \times 1$, $t \times 1$, $a \times 1$, and $s \times 1$.

Then we have the link functions:

$$\begin{aligned}
\mu_{ij} &= X_{aij}^t \alpha \\
\gamma_{ij} &= X_{bij}^t \beta \\
\sigma_{ij}^2 &= exp(X_{sij}^t \xi) \\
\rho_{ij} &= \frac{exp(X_{tij}^t \tau) - 1}{exp(X_{tij}^t \tau) + 1}
\end{aligned}$$

Let l_{ij} be the log-likelihood function. Then we can get independent score functions:

$$\begin{aligned}
&\sum_{i,j} \begin{bmatrix} \frac{\partial l_{ij}}{\partial \alpha} \\ \frac{\partial l_{ij}}{\partial \beta} \\ \frac{\partial l_{ij}}{\partial \xi} \\ \frac{\partial l_{ij}}{\partial \tau} \end{bmatrix} = 0 \\
\Rightarrow &\sum_{i=1}^{N} D_i^T V_i^{-1} (Y_i - \mu_i(\theta)) = 0
\end{aligned}$$

where Y_i is some function of Y_{1i} and Y_{2i}.

$\mu_i(\theta)$ is a function of α, β, ξ, τ.

$$\begin{aligned}
Var(\hat{\theta}) &= A^{-1} B A^{-1} \\
A &= D_i^T V_i^{-1} D_i \\
B &= D_i^T V_i^{-1} Var(Y_i) V_i^{-1} D_i
\end{aligned}$$

#Appendix

```
## 1.1
df <- read.table("~/Desktop/571/HW3/sixcity.dat", quote="\"", comment.char="")
colnames(df) = c("wheezing", "id", "age", "maternal_smoking")
df[,"maternal_smoking"] = factor(df[,"maternal_smoking"])
df[,"wheezing"] = factor(df[,"wheezing"])
model_glm = glm(wheezing ~ age + maternal_smoking, data = df,
                family = "binomial")
knitr::kable(summary(model_glm)$coefficients ,
             caption = "logistic regression model")
```

```
## 1.2
library(gee)
model_gee_indep = gee(wheezing ~ age + maternal_smoking,
                      id = id, corstr = "independence", family = "binomial", data = df)



model_gee_exch = gee(wheezing ~ age + maternal_smoking,
                     id = id, corstr = "exchangeable", family = "binomial", data = df)
```

```
## 1.3
library(alr)
data(alrset)
df2 =  read.table("~/Desktop/571/HW3/sixcity.dat", quote="\"", comment.char="")
colnames(df2) = c("wheezing", "id", "age", "maternal_smoking")
y = as.matrix(df2[,1])
x = cbind(1, as.matrix(df2[,3:4]))
id = as.matrix(df2[,2])
model_alr = alr(y ~ x - 1, id = id, depm = "exchangeable", ainit = 0.01)
print(model_alr$alpha)
```

```
tbl_sim = cbind(coef_mat, se_mat, rho_mat)
colnames(tbl_sim) = c("est.coef.beta0", "est.coef.beta1", "est.coef.beta2",
                      "SE.beta0", "SE.beta1", "SE.beta2", "correlation.rho")
```

```
## 2.2
coef_avg = colMeans(coef_mat)
coef_bias = coef_avg - c(beta_intercept, beta_coefficients)

rho_avg = mean(rho_mat)
rho_bias = rho_avg - rho
```

```
## 2.3
se_avg = colMeans(se_mat)
se_emp = apply(coef_mat, 2, sd)
```

```
## 2.4
beta_coefficients = c(0.5, 0.5)
beta_intercept = -1.5
sample_size = 300
cluster_size = 3
rho = 0.75
latent_correlation_matrix = toeplitz(c(1, rep(rho, cluster_size - 1)))
N = 200
```

```
coef_mat = matrix(0, nrow = N, ncol = 3)
se_mat = matrix(0, nrow = N, ncol = 3)
rho_mat = matrix(0, nrow = N, ncol = 1)

for (i in 1:200) {
  x1 = rep(c(1, 0), rep(sample_size * cluster_size / 2, 2))
  x2 = rep(0:(cluster_size-1), sample_size)
  simulated_binary_dataset = rbin(clsize = cluster_size, intercepts = beta_intercept,
                                  betas = beta_coefficients, xformula = ~x1 + x2,
                                  cor.matrix = latent_correlation_matrix, link = "logit")
  binary_gee_model = quiet(gee(y ~ x1 + x2, family = "binomial", id = id, corstr = "exchangeable",
                         data = simulated_binary_dataset$simdata))
  summary_gee = summary(binary_gee_model)
  coef_mat[i,] = summary_gee$coefficients[,1]
  se_mat[i,] = summary_gee$coefficients[,2]
  rho_mat[i,] = summary_gee$working.correlation[2,1]
}

tbl_sim = cbind(coef_mat, se_mat, rho_mat)
colnames(tbl_sim) = c("est.coef.beta0", "est.coef.beta1", "est.coef.beta2",
                      "SE.beta0", "SE.beta1", "SE.beta2", "correlation.rho")


coef_avg = colMeans(coef_mat)
coef_bias = coef_avg - c(beta_intercept, beta_coefficients)

rho_avg = mean(rho_mat)
rho_bias = rho_avg - rho

se_avg = colMeans(se_mat)
se_emp = apply(coef_mat, 2, sd)
```

```
## 3.1
library(reshape2)
library(dplyr)
library(lme4)


data =  read.table("~/Desktop/571/HW3/framingham.dat", quote="\"", comment.char="")
colnames(data) = c("age", "gender", "bmi_baseline","bmi_year10", "cigarattes",
                   "cholesterol_year0",  "cholesterol_year2",  "cholesterol_year4",
                   "cholesterol_year6",  "cholesterol_year8", "cholesterol_year10",
                   "dead")
data["individuals"] = 1:nrow(data)
data[6:11][data[,6:11] == -9] = NA
data2 = melt(data, id=c("age", "gender", "bmi_baseline","bmi_year10",
                        "cigarattes", "dead", "individuals"))
data2 = data2[complete.cases(data2),]

data2["year"] = recode(data2$variable, cholesterol_year0=0, cholesterol_year2=2,
       cholesterol_year4=4, cholesterol_year6=6, cholesterol_year8=8,
       cholesterol_year10=10)
```

```r
# LMM
model_lmm = lmer(value ~ age + gender + bmi_baseline + (year | individuals),
              data = data2)
```

```r
# GEE
model_gee = gee(value ~ age + gender + bmi_baseline + year, id = individuals,
                famil = "gaussian",
                corstr = "exchangeable", data = data2)
```

```r
## 3.2
library(MASS)

m = 200
n = 6
x1 = rep(c(1, 0), rep(m * n / 2, 2))
x2 = rnorm(m * n)
id = factor(rep(1:200, rep(6, 200)))
rho = 0.5
sig_err = 0.5
sig_effect = 1

# Simulation 1, LMM setup
beta1 = 2
beta2 = sig_effect * rep(rnorm(200), rep(6, 200))
y = 1 + x1 * beta1 + x2 * beta2 + sig_err * rnorm(m * n)

# lmm
model_lmm = lmer(y ~ x1 + (x2 | id))
# gee
model_gee = gee(y ~ x1 + x2, id = id, famil = "gaussian", corstr = "exchangeable")
```

```r
# Simulation 2, GEE setup
beta2 = 1
y = beta1 * x1 + beta2 * x2 + sig_err * rnorm(m * n)
model_lmm = lmer(y ~ x1 + (x2 | id))

# gee
model_gee = gee(y ~ x1 + x2, id = id, famil = "gaussian", corstr = "independence")
```