Simulation Results

Based on the last time results, I made some changes on the simulation:

Overview:

What I changed for the simulation:

- 1. Fit one simulation to compare the results from package *lme4+geepack* and package *CRTgeeDR*
- 2. Use the *predict()* function to get the correct weigths from *glm* and *glmer*.
- 3. Re-run the code with varied cluster size.
- 4. Check the boxplot of weigths

Comparison of CRTgeeDR and geepack

print the data for comparison

```
head(compare_data,3)
```

```
##           x y         r cluster R arm
## 1 2.4137830 1 0.7453157       1 0   0
## 2 1.8371512 1 0.6217896       2 1   0
## 3 0.9972077 1 0.4151313       3 0   0
```

Method 1: using CRTgeeDR package

```
## using CRTgeeDR package
library(CRTgeeDR)
library(lme4)
library(geepack)
library(jomo)
result1=geeDREstimation(formula=y~x+arm,
                        id="cluster" , data = compare_data,
                        nameMISS='missing',nameY='y',
                        nameTRT='arm',
                        family=binomial("logit"),
                        corstr = "independence",
                        model.weights=I(R==1)~x)
```

Result:

```
summary(result1)
```

```
##             Estimates Model SE Robust SE   wald         p
## (Intercept)    1.1460  0.04413    0.1595  7.184 0.0000000
## x              1.0090  0.01721    0.0946 10.660 0.0000000
## arm            0.9112  0.04157    0.2437  3.739 0.0001847
##
##  Est. Correlation:  0
```

```
##  Correlation Structure:  independence
##  Est. Scale Parameter:  1.246
##
##  Number of GEE iterations: 2
##  Number of Clusters:  50    Maximum Cluster Size:  29
##  Number of observations with nonzero weight:  1254
```

Method 2, using lme4+geepack

```
## method 2
## calculate the weight
w1=glm(R ~ x , data = compare_data,
       family = binomial(link='logit'))
weight=expit(predict(w1))
compare_data$weight=round(1/weight)
compare_data=na.omit(compare_data)

## fit the gee model
result2=geese(formula=y~x+arm,data=compare_data,id=cluster,
              family = binomial(link='logit'),
              weights = weight,
              corstr = 'independence')
```

Result:

```
summary(result2)
```

```
##
## Call:
## geese(formula = y ~ x + arm, id = cluster, data = compare_data,
##     weights = weight, family = binomial(link = "logit"), corstr = "independence")
##
## Mean Model:
##  Mean Link:                 logit
##  Variance to Mean Relation: binomial
##
##  Coefficients:
##              estimate    san.se     wald            p
## (Intercept) 1.1427623 0.2013486 32.21176 1.382517e-08
## x           1.0083730 0.1142413 77.91059 0.000000e+00
## arm         0.9128119 0.2522835 13.09136 2.966607e-04
##
## Scale Model:
##  Scale Link:                identity
##
##  Estimated Scale Parameters:
##              estimate    san.se     wald           p
## (Intercept) 0.9228705 0.2923007 9.968319 0.001592568
##
## Correlation Model:
##  Correlation Structure:     independence
##
## Returned Error Value:    0
## Number of clusters:   1254   Maximum cluster size: 1
```

Table 1: The comparison result

	Estimate	Sandwich variance estimator
Method 1	0.9112	0.2437
Method 2	0.9128	0.2523

Although the two methods are a little bit different, I think the difference is negligible.

Simulation

The simulation based on Hossain's paper *"Missing binary outcomes under covariate dependent missingness in cluster randomised trials"*. The goal of the simulation is to compare the effects of inverse probability weighting (IPW) and multilevel multiple inputation (MMI)

p.s. In the last time, I used 5 imputations for the MI methods. I changed it into 15 times, which is the same number as Hossain's paper.

1. Data Generation

Assuming the true data generating model hs log link, suppose that each binary outcome Y_{ijl} is generated by:

$$\pi_{ijl} = exp(\beta_0 + \beta_1 i + f_i(x_{ijl}) + \delta_{ij})$$

The notation method consistents to Hossain's:

- ijl means the *ith* intervention group, the *jth* cluster, the *lth* individual.
- i=0 control group while i=1 intervention group.
- We have k clusters, j=1,2,...,k
- We have m individuals in one cluster. This time m can vary and is a number randomly selected from a uniform distribution UNI(m-5, m+5)
- β_0 is a constant, β_1 is the true intervention effect. $f_i(x_{ijl})$ is a function of baseline covearite X in the *ith* intervention group. We set $f_0(x_{ijl}) = f_1(x_{ijl}) = \beta_2 x_{ijl}$ Consistent with Hossain's paper, we set $\beta_0 = 1, \beta_1 = 1.36, \beta_2 = 1$
- X_{ijl} is generated by using the methods:

  $$X_{ijl} = \alpha_{ij} + u_{ijl}$$

  where α_{ij} is the (ij)th cluster effect on X and u_{ijl} is the individual-level error on X. We assumed that $\alpha_{ij} \sim N(\mu_x, \sigma_\alpha^2)$, $u_{ijl} \sim N(0, \sigma_u^2)$, where σ_α^2 and σ_u^2 are the between-cluster and within-cluster variance of X, respectively. We set $\mu_x = 0, \sigma_\alpha^2 = 0.18, \sigma_u^2 = 3.37$
- $\delta_{ij} \sim N(0, \sigma_b^2)$. We set $\sigma_b^2 = 0.2$
- Y_{ijl} is generated as Bernoulli random varaible with parameter π_{ijl}

2. Missingness generation:

We assume the missing mechanism is covariate dependent missingness (CDM).

The missingness is generated by the logistic regression model:

$$logit(R_{ijl} = 1|Y_{ij}, X_{ij}) = \psi_i + \phi_i X_{ijl}$$

For a simple example, we do not add group indicator in the model. We just let:

$\psi_0 = \psi_1 = -1.34$, $\phi_0 = \phi_1 = 1$

Table 2: Parameter Value

Parameter	value
β_0	0
β_1	1.36 (true effect)
β_2	1
α_{ij}	$N(\mu_x, \sigma_\alpha^2)$
u_{ijl}	$N(0, \sigma_u^2)$
$\psi_0 = \psi_1$	-1.34
$\phi_0 = \phi_1$	1
μ_x	0
σ_α^2	0.18
σ_u^2	3.37
δ_{ij}	$N(0, \sigma_b^2)$
σ_b^2	0.2

2. Missingness handling methods:

2.1 Complete Record Analysis (CRA)

For CRA, no imputation is performed, and only data from subjects with an observed outcome are considered for statistical analysis. Besides, we also adjusted covariates for CRA since we assume the missing mechanism is CDM. Therefore, our CRA here is adjusted CRA.

2.2 Inverse probability weigthing (IPW)

2.2.1 IPW without cluster effects

Suppose w_{ij} is the weight for y_{ij} and is defined as the inverse probability of observing y_{ij}. In other words, $w_{ij} = P(R_{ij} = 1|X_i, Y_i)^{-1}$. Suppose W_i is a $T * T$ diagonal. Consider a generalized estimating equation:

$$S(\beta) = \sum \frac{\partial \mu_i}{\partial \beta} V_i^{-1} W_i (Y - \mu_i(\beta)) = 0$$

The weights can be estimated for a logistic regrssion:

$$\hat{w}_{ijl} = expit(X_{ijl}\beta')$$

2.2.2 IPW with cluster effects (IPW_cluster)

Different with IPW without cluster effects, if we consider clusters, we need to change weights for each observed individuals and just modify the weigths equation:

$$\hat{w}_{ijl} = expit(X_{ijl}\beta' + \delta_{ij})$$

where δ_{ij} is the cluster level variable.

2.3 Multilevel Multiple Inputation (MMI)

Since many researchers believe that MMI is the best MI methods that with consideration of cluster effects. Therefore, we use MMI as a representative of MI methods to compare with IPW.

The missing data are firstly imputed based on the random logistic regression model

$$logit(\pi_{ijl} = 1|Y_{ij}, X_{ij}) = \beta_0 + \beta_1 X_{ijl}$$

After the missing values are imputated, a full data is generated. Then GEE method can be used to analyze the full data. After several times of repeats of the previous procedures, the results can be pooled according to Rubin's rule and then one pooled estimate were generated.

Analysis model:

Generalized estimated equation (GEE) is used here to analyze the results.

And choose indenpent working covariation matrix

$$logit(\pi_{ijl} = 1) = \beta_0 + \beta_1 X_{ijl} + \beta_2 * group$$

Transform

Notice that, in our simulation, for data generation, we used generalized linear mixed model, while in analysis part we applied gee to analyze the generated data. Therefore, the data generation process gives us a conditional estimate while data analysis model provides us a marginal estimate. We have to make some transformations to make them both marginal or both conditional.

Hossain faced the same issue, and he got the true value of population averaged log(OR) for GEE by empirically estimation using full data.

Also, Zeger, et al. 1988 showed another transforamtion method, that for logistic regression:

$$\beta_M \simeq [(\frac{16\sqrt{(}3)}{15\pi})^2 V + 1]^{1/2} \beta_{RE}$$

Here, we used Hossain's method to make consistency.

Results

Cluster Summary

To get the more general results, I simulated the dataset with different sizes of clusters.

Table 3: Cluster Summary Table

		Missing Percent	Cluster Size			
k	m		min	max	mean	sd
25	25	0.3	20	30	25.01	2.87
	50	0.3	45	55	50.00	2.87
50	25	0.3	20	30	24.99	2.89
	50	0.3	45	50	50.00	2.89

- k is the cluster number iin each arm

- m is the mean value of cluster size. cluster size is from the uniform distribution UNI(m-5, m+5)

The following table showns the simulation results

Table 4: Simulation Resutls

k	m	True effect	Methods	Average Est	Average Est SD	Coverage %	Not converge time
25	25	1.325	unadj_CRA	0.886	0.134	0.06	
			adj_CRA	1.325	0.169	0.956	
			IPW1_no	1.379	0.31	0.948	57
			IPW2_no	1.369	0.311	0.944	
			IPW1_clu	1.381	0.312	0.949	63
			IPW2_clu	1.369	0.311	0.944	
			MMI	1.325	0.241	0.99	
	50	1.323	unadj_CRA	0.885	0.095	0.002	
			adj_CRA	1.323	0.12	0.963	
			IPW1_no	1.355	0.231	0.953	35
			IPW2_no	1.346	0.231	0.952	
			IPW1_clu	1.356	0.232	0.952	35
			IPW2_clu	1.346	0.231	0.952	
			MMI	1.317	0.169	0.988	
50	25	1.319	unadj_CRA	0.885	0.094	0.001	
			adj_CRA	1.319	0.119	0.95	
			IPW1_no	1.356	0.229	0.954	36
			IPW2_no	1.35	0.229	0.948	
			IPW1_clu	1.357	0.23	0.951	36
			IPW2_clu	1.35	0.229	0.948	
			MMI	1.317	0.169	0.988	
50	50	1.317	unadj_CRA	0.882	0.067	0	
			adj_CRA	1.317	0.084	0.965	
			IPW1_no	1.339	0.169	0.963	25
			IPW2_no	1.333	0.17	0.962	
			IPW1_clu	1.339	0.17	0.964	25
			IPW2_clu	1.333	0.17	0.962	
			MMI	1.319	0.119	0.988	

- unadj_CRA: unadjusted Complete Record Analysis
- adj_CRA: adjusted Complete Record Analysis
- IPW1_no: without cluster effect IPW, using glm+geese
- IPW2_no: without cluster effect IPW, using CRTgeeDR
- IPW1_clu: with cluster effect IPW, using glm+geese
- IPW2_clu: with cluster effect IPW, using CRTgeeDR
- MMI multilevle multiple imputation

Discussion

Results Comparison

For each scenario, the missing percentage is 30%, which is consistent to our generation method.

For average estimates, since we assume covariate dependent missingness (CDM), so CRA with adjusted for covariates gains unbiased effects (the estimate is the same with true value). In the MMI results, the difference between the average estimate and true effects are quite small. MMI can be considered as unbiased based on the results. However, IPW, no matter considering clusters or not, overestimates the true effects.

Compared to MMI, IPW cannot control uncertainty in missing values, and thus IPW has a larger standard deviation than MMI.

In a conclusion, the adjusted CRA and MMI have unbiased estimations. Although the differeces between IPW and true values are larger then adj CRA and MMI, I think the differences are acceptable and IPW can also be considered as unbiased.

Besides, the results of IPW with cluster effects and the results of IPW without cluster effects are very similar. This may becasue there are no cluster effects in the missingness generation model,

Non-convergence

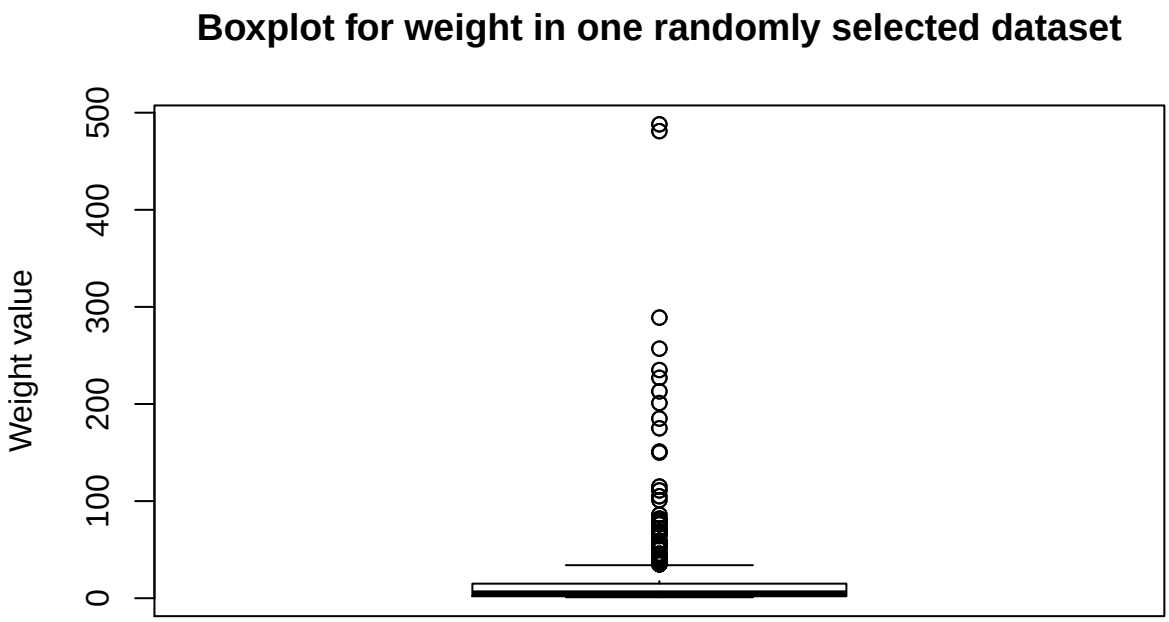
This time we still have some non-covergent results from IPW method 1 (glm/glmer + geese). As cluster size gets larger, the uncovergence time get smaller. With or Without the consideration of cluster effects when calculating the weight does not have a big effect on the time of non-covergence.

These uncovergences may be caused by the the weight estimate. We can choose one generated dataset (from the 4000 datasets) and draw a boxplot of the weight:

The randomly selected dataset:

```
##            x y         r cluster R arm weight
## 1  2.4137830 1 0.7453157       1 0   0      1
## 2  1.8371512 1 0.6217896       2 1   0      2
## 3  0.9972077 1 0.4151313       3 0   0      3
## 4  2.4150924 1 0.7455641       4 1   0      1
## 5  2.9833314 1 0.8379877       5 1   0      1
## 6 -0.6907425 0 0.1160128       6 0   0      9
```

```
## calculate the weight
w1=glm(R ~ x , data = data_weight,
       family = binomial(link='logit'))
weight=expit(predict(w1))
data_weight$weight=round(1/weight)
data_weight=na.omit(compare_data)
boxplot(data_weight$weight,ylab='Weight value',
        main='Boxplot for weight in one randomly selected dataset')
```



We can see that there are large oulier values in the boxplot. These maybe the unconvergence reason.

We then try to replace these large outliers with acceptable values. Here I chose 50 as the max weight:

- If weight < 50, then do not change.
- If weigth >=50, then weight = 50

After this transformation, the new results:

Table 5: The results after dealing with weight outliers

k	m	Methods	Estimate	Estimated_SD	Coverage
25	25	IPW_no: glmer+geese	1.321	0.235	0.960
		IPW_Cluster: glmer+geese	1.321	0.311	0.961
	50	IPW_no: glmer+geese	1.322	0.167	0.948
		IPW_Cluster: glmer+geese	1.323	0.231	0.949
50	25	IPW_no: glmer+geese	1.317	0.166	0.945
		IPW_Cluster: glmer+geese	1.317	0.229	0.944
	50	IPW_no: glmer+geese	1.313	0.118	0.949
		IPW_Cluster: glmer+geese	1.313	0.17	0.951

When deleted the large outliers, the results of IPW get closer to the true values.

Appendix

The following table are the results in the last time.

Table 6: Previous Resutls

k	m	True effect	Methods	Average Est	Average Est SD	Coverage %	Not Converge times
25	25	1.326	CRA_un	0.929	0.177	36.3	0
			CRR	1.323	0.214	97.5	0
			IPW	1.383	0.327	90.6	112
			IPW_cluster	1.387	0.335	89.7	118
			IPW-GEE (CRTgeeDR)	1.323	0.169	93.2	0
			MMI	1.325	0.296	99.9	0
	50	1.319	CRA_un	0.929	0.152	42.9	0
			CRA	1.317	0.174	98.1	0
			IPW	1.355	0.258	93.6	66
			IPW_cluster	1.359	0.262	93.5	70
			IPW-GEE (CRTgeeDR)	1.319	0.119	90.8	0
			MMI	1.327	0.232	99.8	0
50	25	1.320	CRA_un	0.928	0.150	56.7	0
			CRA	1.319	0.153	98.1	0
			IPW	1.362	0.242	92.0	70
			IPW_cluster	1.364	0.248	91.4	74
			IPW-GEE (CRTgeeDR)	1.322	0.119	92.8	0
			MMI	1.321	0.211	99.9	0
	50	1.319	CRA_un	0.928	0.138	57.1	0
			CRA	1.317	0.124	98.7	0
			IPW	1.343	0.190	94.5	52
			IPW_cluster	1.344	0.193	94.6	48
			IPW-GEE (CRTgeeDR)	1.320	0.084	92.6	0
			MMI	1.328	0.165	99.8	0