Technical Report on Household Study

Riddhiman Bhattacharya

November 29, 2020

Introduction

Consider a Phase 3 trial for SARS-CoV-2 infection prevention using the REGN-Cov2 anti-SPIKE cocktail. We shall assess the efficacy of the vaccine in a household contact study where the subjects are from a particular household with sustained exposure to SARS-CoV-2. We consider two models for our analysis- the GEE or Generalized Estimating Equations and Logistic Regression and want to assess which model works best to ascertain the efficacy of the drug. Our trial design is a two arm placebo controlled study with one active arm and the other placebo arm. Each subject in the design has equal probability to receive treatment and placebo. In this report, we shall try to examine the effect of house structure on the models mentioned above. We shall consider the households with single members and households with multiple members separately. We shall vary the size of the household from $10\%, 20\%, \ldots, 80\%$ and compare the performance of the mentioned models in such a scenario. The main idea is that GEE after a certain threshold for households with size 1 shall be a poor fit due to the fact that the effective number of samples used to calculate the intraclass correlation coefficient drops which in turn implies poor estimates of the same and hence the model should be a bad fit. We generate the households with size > 1 as realizations from a truncated Poisson with truncation at 1 and some λ. We shall also generate correlated binary endpoint data as our response using the method of Emrich and Piedmonte 1991. This we do to account for correlated structure within household. We shall compare both the models by estimating the type-1 error, power and MSE for both models.

Correlated Bernoulli Generation

To generate correlated Bernoulli data, we use the method proposed by Emrich and Piedmonte which uses Gaussian Copula for simulation. We exhibit the steps to do this:

- Step1: Generate $Z \sim N(0, \Sigma)$ where $\Sigma = \begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \ldots & \rho & 1 \end{bmatrix}$. (Generating this step is easy both with and without packages.). We can choose this ρ to generate the correlation that we want in the Bernoulli data.
- Step2: On each entry Z_i of Z calculate $\Phi(Z_i)$ where Φ is the standard Normal cdf.
- Step3: If $\Phi(Z_i) < p$, define $Y_i = 1$ else $Y_i = 0$ this vector $Y = (Y_1, Y_2, \ldots, Y_n)'$ is a correlated bernoulli vector.

See Appendix C for code.

Truncated Poisson

Before we begin to generate the data and calculate type-1 errors, we shall need some functions which shall help us in the analysis.

Firstly, we would require a truncated Poisson generator as we want Households with random sizes. We shall of course exclude Households of sizes 0 and by the nature of our problem, we shall also exclude households of sizes 1.

The idea of generating truncated Poisson is as such- generate a Poisson with mean λ if it is > 1 accept the sample, else reject it. So, this, in essence is an accept reject sampler which conditions on the event $X > 1$.
See Appendix C for code.

Problem and Theory

Our main goal here is to compare the GEE and logistic model in our simulation setting. We shall try to do this with varying number of households with size 1 and different intra-cluster correlation. The aim here is to see which model does better in which setting.

The models we consider are:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta' x_i,\ i = 1, 2, 3, \dots, n$$

$$Y_i \sim^{ind} Ber(p_i)$$

the logistic model and

$$\log(\frac{E(y_{i,j})}{1-E(y_{i,j})}) = \beta_0 + \beta' x_{i,j},\ ,i = 1, 2, \dots, n,\ j = 1, 2, 3, \dots, n_i$$

the GEE model. Where,

$$(y_{i,1}, y_{i,2}, \dots, y_{i,n_i})$$

have a correlated Bernoulli structure for each i.

We wish to compare these two models in the setting where we give patients in each household either a placebo or treatment based on the independent randomization design. Inside each cluster, we shall generate responses which are correlated Bernoulli, using the copula method mentioned previously. The intra-cluster correlation for Bernoulli is a tricky aspect as for a given set of success probabilities the pairwise correlation for the Bernoulli random variables is bounded i.e. has a range smaller than $[-1, 1]$ based on what the probabilities are. We provide this bound for a pair-(this can be found in Prentice,1988) Let us say that we have 2 correlated binary variables (Y_1, Y_2) with success probabilities p_1, p_2 with $q_1 = 1 - p_1$, $q_2 = 1 - p_2$ and let δ_{12} denote the correlation. Then-

$max\{-(\frac{p_1p_2}{q_1q_2})^{\frac{1}{2}}, -(\frac{p_1p_2}{q_1q_2})^{-\frac{1}{2}}\} \leq \delta_{12}$

and

$\delta_{12} \leq min\{(\frac{p_1q_2}{q_2p1})^{\frac{1}{2}}, (\frac{p_1q_2}{q_2p1})^{-\frac{1}{2}}\}$

We notice that if $p_1 = p_2$ then the range transforms to $[-1, 1]$ which tells us that when we estimate type-1 error in the above cases the correlation can be as high as possible. However, if we calculate power we have limitations on the intra-class correlation. Since the change in p_1, p_2 changes correlation in the responses, this affects modelling somewhat.

Fixed Cluster Size Simulations

In this section we shall analyze the two models mentioned previously where the number of households is fixed and the number of subjects will vary per iteration. We shall consider the attack rates as $p_1 = p_2 = .09$ for the null case and $p_1 = .05, \ , p_2 = .09$ for the alternate case. We shall consider 500 clusters with average size 4 and a thousand iterations each time for the Monte Carlo estimates of power, type-1 error and MSE.
For this particular simulation, we shall consider a γ(intra-cluster correlation) of approximately .65 ($\rho = .9$). We shall henceforth refer to intra-cluster correlation as γ

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.0420000	0.0460000	0.0001392	0.0001790
10%	0.0500000	0.0460000	0.0001433	0.0001925
20%	0.0460000	0.0550000	0.0001559	0.0002064
30%	0.0530000	0.0440000	0.0001438	0.0002024
40%	0.0690000	0.0550000	0.0001586	0.0002308
50%	0.0470000	0.0590000	0.0001752	0.0002729
60%	0.0550000	0.0560000	0.0001715	0.0002509
70%	0.0350000	0.0450000	0.0001917	0.0002977
80%	0.0600000	0.0450000	0.0011490	0.0003323

As we can see that for this particular correlation, there is an inflation in the type-1 error at 80%. This implies that the GEE model is a bad fit for the design. Logistic however exhibits more robustness than GEE and has a type-1 error which is controlled irrespective of the number of households of size 1. Also, if we observe the MSE, we can see that the MSE for GEE is better than that of Logistic, till the number of of households of size 1 is 80%.

We shall estimate the power to ascertain which model has better power. Our simulations will have $p_1 = .05, \ p_2 = .1$ and $\gamma = .65$ approximately ($\rho = .9$)

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	1.0000000	0.9500000	0.0009019	0.0009352
10%	0.9980000	0.9400000	0.0009067	0.0009498
20%	0.9990000	0.9150000	0.0009002	0.0009602
30%	0.9940000	0.8870000	0.0009224	0.0009809
40%	0.9760000	0.8580000	0.0008989	0.0009807
50%	0.9640000	0.8110000	0.0009163	0.0010121
60%	0.924000	0.752000	0.000910	0.001002
70%	0.8380000	0.6900000	0.0009138	0.0010319
80%	0.681000	0.596000	0.001577	0.001082

Since the sample size decreases, we do not get a good idea about the power comparison in this case. At an initial glance it looks as if GEE is better than Logistic.

Here we have the type-1 error estimates for GEE and Logistic Regression where $\gamma = .9$approximately ($\rho = .99$). Again, we take $p_1 = p_2 = .09$ and have the cluster size as 500 with average 4 subjects per cluster.

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.0450000	0.0670000	0.0001681	0.0002467
10%	0.0500000	0.0550000	0.0001668	0.0002437
20%	0.0570000	0.0410000	0.0001570	0.0002462
30%	0.0580000	0.0520000	0.0001720	0.0002507
40%	0.0550000	0.0510000	0.0001567	0.0002711
50%	0.0380000	0.0590000	0.0001786	0.0003101
60%	0.0580000	0.0510000	0.0001784	0.0003274
70%	0.0420000	0.0530000	0.0002107	0.0003824
80%	0.0720000	0.0470000	0.0034161	0.0003838

As we can see, the type-1 error inflates when the number of households with size 1 is at 80%. This is the same fact that we observe in the previous case with $\gamma = .65$. We observe the same thing with the MSE.

Next we shall look at the power under the same correlation with $p_1 = .05,\ ,p_2 = .09$ and $\gamma = .9$ approximately$(\rho = .99)$.

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	0.9990000	0.9200000	0.0009370	0.0009795
10%	0.9980000	0.9000000	0.0009735	0.0010157
20%	1.0000000	0.8830000	0.0009812	0.0010173
30%	0.9940000	0.8600000	0.0009789	0.0010452
40%	0.9810000	0.8200000	0.0009739	0.0010445
50%	0.928000	0.783000	0.001010	0.001067
60%	0.8390000	0.7350000	0.0009845	0.0010761
70%	0.697000	0.677000	0.001033	0.001091
80%	0.581000	0.624000	0.005490	0.001134

As we saw in the previous case, we see here that the power decrases as the sample size decreases. Also at 80% Logistic has more power than GEE.

Fixed Sample Study

Type-1 Error Estimates

In this section, we shall consider the number of subjects as fixed. The cluster counts shall vary from iteration to iteration. We shall be mainly interested in looking at the power, type-1 error and the MSE values for GEE and Logistic Regression. The issue with the previous comparison was that the decrease in sample size influenced the power to go down. Also, we did not have the same number of subjects while estimating the type-1 errors. This brings about our second stage of analysis. We shall again look at type-1 error and power at different γ values.
Our design scheme is Independent Randomization, which is providing a subject drug/placebo irrespective of which household the subject belongs to. The total number of subjects enrolled in the study are 1368 and the success probabilities are taken to be $p_1 = p_2 = .1$ for the null case and $p_1 = .05,\ ,p_2 = .1$ for the alternate case. We shall compare the two models by varying the number of households which have size 1. The percentages for the households with size 1 we consider are $5, 10, 20, 30, 40, 50, 60, 70, 80$. The total number of iterations is 2×10^3.

$$\rho = .9,\ \gamma \approx .65$$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.052000	0.047500	0.000212	0.000287
10%	0.050500	0.051500	0.000208	0.000289
20%	0.054000	0.046500	0.000182	0.000280
30%	0.049000	0.048000	0.000163	0.000242
40%	0.048500	0.051000	0.000157	0.000243
50%	0.052000	0.041500	0.000207	0.000223
60%	0.049000	0.040500	0.000171	0.000201
70%	0.056500	0.045500	0.001133	0.000184
80%	0.075500	0.043000	0.002155	0.000167

As we can see that both models perform reasonably with the GEE outperforming Logistic Regression (by looking at MSE) mildly till the percentage hits 70% and then Logistic outperforms GEE. In fact, the GEE model behaves erratically. This we suspect is due to the fact that GEE estimates the γ incorrectly when the number of households with size 1 increases. The $\gamma \approx .65$ in this case.

$\rho = .99$, $\gamma \approx .88$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.047000	0.044000	0.000240	0.000354
10%	0.045500	0.050000	0.000222	0.000360
20%	0.045500	0.050500	0.000177	0.000306
30%	0.045500	0.046500	0.000156	0.000298
40%	0.042000	0.051500	0.000142	0.000276
50%	0.050500	0.051500	0.000389	0.000245
60%	0.077500	0.051000	0.002765	0.000216
70%	0.101000	0.039500	0.004524	0.000204
80%	0.110000	0.056500	0.004779	0.000182

As we can see that for $\gamma \approx .88$ the GEE outperforms Logistic (even though they are comparable) except when the percentage hits 60. On and after 60 Logistic fits the design much better than GEE.

$\rho = .999, \gamma \approx .95$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.034500	0.046000	0.000235	0.000369
10%	0.028000	0.052000	0.000204	0.000342
20%	0.033000	0.050500	0.000166	0.000326
30%	0.030000	0.047500	0.000135	0.000292
40%	0.038000	0.040500	0.000351	0.000269
50%	0.047000	0.048500	0.001188	0.000257
60%	0.078500	0.049500	0.004170	0.000226
70%	0.126500	0.057000	0.006544	0.000205
80%	0.108000	0.049500	0.003395	0.000176
90%	0.077000	0.040000	0.002190	0.000149

In the case of ultra high correlation ($\gamma \approx .95$), we again have similar behaviour that is both GEE and logistic are comparable except when we hit the 60% mark, where GEE performs poorly. So, in all we can see that for moderate correlation GEE starts performing poorly from 70% onwards and at high and ultra-high correlation, GEE performs poorly from 60% onwards.

Power for Household Study with Fixed Sample Size

Next, we are going to look at how power varies with change in percentage of households of size 1. One thing to note is that, in this scenario, with the increase in the number of households of size 1, the number of clusters increase. However, the sample size is kept fixed at 1368. We take the attack rates for the active and placebo arm as $p_1 = .05$ and $p_2 = .1$ respectively with

$$\rho = .9,\ \gamma \approx .65$$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	0.998000	0.939000	0.001383	0.001464
10%	0.998000	0.941000	0.001399	0.001476
20%	0.994500	0.940000	0.001350	0.001409
30%	0.988000	0.938000	0.001329	0.001440
40%	0.970500	0.929500	0.001301	0.001405
50%	0.953500	0.938500	0.001342	0.001429
60%	0.921000	0.942000	0.001759	0.001399
70%	0.890500	0.950500	0.002304	0.001419
80%	0.880000	0.939500	0.004620	0.001372

As we can see that the power of the GEE drops below 90% from 70% onwards. The Logistic Regression has more power than the GEE in this regime.
We now do the same simulation with a higher intraclass correlation. The objective behind doing these simulations is to see how the cutoff varies (for the number of households with size 1) as one model does better than the other.

$$\rho = .99,\ \gamma \approx .88$$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	0.999000	0.912500	0.001432	0.001524
10%	0.998000	0.903500	0.001422	0.001527
20%	0.991000	0.916500	0.001382	0.001488
30%	0.961500	0.923000	0.001347	0.001448
40%	0.929500	0.921500	0.001606	0.001488
50%	0.862500	0.929500	0.001783	0.001425
60%	0.838500	0.941000	0.003944	0.001436
70%	0.820000	0.934500	0.005987	0.001409
80%	0.834000	0.942000	0.007767	0.001405

As one can see from these simulations, even at 50%, the power for GEE drops. However, at 50% the power is close to 90%. Logistic convincingly beats GEE from 60%.
Next, we do the same exercise with ultra-high correlation.

$$\rho = .999,\ \gamma = .95$$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	0.996500	0.903500	0.001414	0.001524
10%	0.995500	0.901000	0.001405	0.001499
20%	0.973000	0.915500	0.001423	0.001541
30%	0.935000	0.908500	0.001361	0.001494
40%	0.881500	0.912500	0.001378	0.001471
50%	0.826500	0.921000	0.002540	0.001454
60%	0.787000	0.922000	0.005086	0.001421
70%	0.789000	0.927500	0.007966	0.001430
80%	0.807500	0.934000	0.006698	0.001401

This exhibits that even at 50%, the power of the GEE drops and the power for Logistic Regression is more robust and remains above 90% at all levels. This we suspect is due to small size of clusters in general. Since the size of the clusters are small, each cluster under high correlation acts like a single data point repeated and logistic tackles this.

Variation in Type-1 Error and Power by changing Intraclass Correlation

We also inspect the behavior of the type-1 and power as we change the intraclass correlation from lower to higher values. We shall have the value of the correlation among the normals used to generate the γ intra class correlation value in the table. We only present some of our results; for the rest of the results, refer to the appendix. Again in all our simulations $p_1 = p_2 = .1$ for the null case and $p_1 = .05, \ , p_2 = .1$ for the alternate case. The number of subjects is 1368.
The following is when the number of households with size 1 is at 10%.

ρ	γ	α-GEE	α-Logistic
0.340000	.15	0.052000	0.053000
0.430000	.20	0.056500	0.058500
0.500000	.25	0.046500	0.041000
0.570000	.30	0.053000	0.048000
0.690000	.40	0.055000	0.045500
0.750000	.45	0.044500	0.046000
0.800000	.50	0.042500	0.051000
0.870000	.60	0.036500	0.042500

ρ	γ	β-GEE	β-Logistic
0.340000	.15	0.950000	0.942000
0.430000	.20	0.955000	0.943000
0.500000	.25	0.957000	0.945000
0.570000	.30	0.972000	0.947000
0.690000	.40	0.982500	0.937000
0.750000	.45	0.988000	0.945000
0.800000	.50	0.990000	0.950500
0.870000	.60	0.997000	0.938000

As one can see, that the type-1 error is controlled for both models irrespective of the correlation within cluster. As we can see that the they start from almost the same power when the correlation is low and then the power increases for GEE and decreases for Logistic.
The following is when the number of households with size 1 is at 40%

ρ	γ	α-GEE	α-Logistic
0.340000	.15	0.047000	0.053000
0.430000	.20	0.052000	0.049500
0.500000	.25	0.053500	0.058500
0.570000	.30	0.048000	0.048000
0.690000	.40	0.054000	0.056000
0.750000	.45	0.058500	0.059000
0.800000	.50	0.050500	0.053500
0.870000	.60	0.056500	0.055500

ρ	γ	β-GEE	β-Logistic
0.340000	0.150000	0.941000	0.936500
0.430000	0.200000	0.944000	0.940500
0.500000	0.250000	0.964500	0.951500
0.570000	0.300000	0.961000	0.953000
0.690000	0.400000	0.963000	0.940500
0.750000	0.450000	0.971000	0.943000
0.800000	0.500000	0.975000	0.938500
0.870000	0.600000	0.975500	0.947500

The same conclusion is valid in the above tables as was true in the 10% case. The following is when the number of households with size 1 is at 70%

ρ	γ	α-GEE	α-Logistic
0.340000	.15	0.040500	0.041500
0.430000	.20	0.042000	0.045000
0.500000	.25	0.048000	0.048500
0.570000	.30	0.054500	0.053000
0.690000	.40	0.053500	0.055000
0.750000	.45	0.046000	0.048000
0.800000	.50	0.050000	0.052500
0.870000	.60	0.059000	0.046500

ρ	γ	β-GEE	β-Logistic
0.340000	0.150000	0.943500	0.942500
0.430000	0.200000	0.941000	0.942000
0.500000	0.250000	0.947000	0.946500
0.570000	0.300000	0.934000	0.938500
0.690000	0.400000	0.935500	0.944500
0.750000	0.450000	0.927500	0.943500
0.800000	0.500000	0.914500	0.947500
0.870000	0.600000	0.911000	0.943500

Again, we see that there is not much difference between GEE and Logistic type-1 error and both maintain type-1 at 50% except the final correlation where both type-1 errors have minor inflation. The power telling in this situation is quite telling. As the correlation increases the power of the Logistic increases and that of GEE decreases with the Logistic having more power than GEE after correlation .4.

Fixed Number of People With Bounded Household Size

In this section, we shall look at fixed number of people with an upper bound bound on the household size. We shall of course have a percentage of households with size 1 as we have had previously and we shall vary this particular quantity throughout the simulation. We shall also vary two other quantities throughout this-i. The intracluster correlation and ii. the upper bound on the size of the household.

The reason for upper bounding the household size is two-fold- one practical and the other is to test the robustness of the GEE. The practical consideration is that households with sizes more than five are rare if not impossible to come up in a study and anything above the size of 5 is a practical impossibility. Hence, we shall try to bound household sizes. Also, we want to see if the cluster sizes are low, does the estimation of the correlation within the cluster suffer which in turn shall give us poor performance of the GEE. Our intuition is that GEE should be robust in this case due to the compound symmetry structure of the model(which is a model assumption we have made). However, it still might give us poor performance when we have households with quite small sizes and high correlation. We wish to perform empirical experiments to see when this occurs. One other thing to note is that we shall only take the percentage of households that have size 1 to go up to 40%. This is because till 40% we have seen GEE beat logistic regression convincingly under more or less all levels of correlation. On $50, 60\%$ both models perform comparably. From 70%, GEE starts performing poorly and Logistic fits the data well. We introduce some minor background before delving further into the simulations. The household size in this scenario is a Poisson truncated at both ends i.e. Let $X \sim Poi(\lambda)$. Then the household distribution is the same as $X|1 < X \leq b$, where b is the upper bound here. For the simulations, we consider 1368 samples and 2000 iterations with the attack rates as $p_1 = p_2 = .1$ in the null case and $p_1 = .05, \ , p_2 = .1$ in the alternate case. The intuition behind doing this simulation is that GEE performs much better when the cluster sizes increase and thus when the cluster sizes decrease, GEE should suffer somewhat. However, since we use the compound symmetry structure for the correlation it may be that the model still does well. The simulation below bounds the household size by 2 i.e. we can only have households of size 1 and 2.

$$\rho = .9, \ b = 2$$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.059500	0.059000	0.000145	0.000176
10%	0.049000	0.052000	0.000141	0.000172
20%	0.058000	0.045500	0.000132	0.000159
30%	0.051000	0.052000	0.000134	0.000163
40%	0.048500	0.043000	0.000125	0.000148

As we can observe from the simulations above, the type-1 error in both cases are reasonable.

$$\rho = .99, \ b = 2$$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.052000	0.054500	0.000139	0.000195
10%	0.048000	0.043000	0.000132	0.000189
20%	0.054000	0.049500	0.000122	0.000176
30%	0.049000	0.049000	0.000123	0.000180
40%	0.043000	0.044000	0.000115	0.000165

The type-1 error estimates are still good for both GEE and Logistic and GEE beats Logistic in terms of the MSE.

$\rho = .999,\ b = 2$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.021000	0.048000	0.000127	0.000192
10%	0.025000	0.049500	0.000123	0.000190
20%	0.029000	0.051000	0.000118	0.000183
30%	0.029000	0.043000	0.000108	0.000175
40%	0.030500	0.040500	0.000100	0.000163

As we can see the GEE is quite robust due to the compund symmetry structure and hence it takes ultra high correlation to make the GEE a bad fit. When $\gamma \approx .999$ the GEE performs poorly. In all the other scenarios, GEE performs well.
Next we shall do the same exercise with the upper bound as 3

$\rho = .9,\ b = 3$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.051000	0.048500	0.000159	0.000197
10%	0.049000	0.043500	0.000160	0.000197
20%	0.054000	0.041500	0.000149	0.000193
30%	0.051500	0.050500	0.000140	0.000179
40%	0.057500	0.056000	0.000145	0.000182

As we can see here, the performances of GEE and logistic are similar in terms of data fit with the GEE being mildly better in terms of MSE.

$\rho = .99,\ b = 3$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.051500	0.048000	0.000166	0.000230
10%	0.051500	0.043000	0.000151	0.000221
20%	0.048000	0.051500	0.000143	0.000218
30%	0.049000	0.044000	0.000131	0.000196
40%	0.046000	0.047500	0.000123	0.000188

We see similar behavior here.

$\rho = .999,\ b = 3$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.024500	0.051500	0.000162	0.000242
10%	0.029500	0.050000	0.000149	0.000219
20%	0.026000	0.044000	0.000132	0.000214
30%	0.035000	0.047000	0.000120	0.000204
40%	0.040500	0.054500	0.000116	0.000193

Both models seem to work decently except for GEE in ultra-high correlation setting.

$\rho = .9,\ b = 4$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.049000	0.058500	0.000186	0.000233
10%	0.049000	0.044000	0.000164	0.000210
20%	0.044000	0.056000	0.000167	0.000221
30%	0.052000	0.044500	0.000151	0.000199
40%	0.055500	0.050500	0.000136	0.000182

Here also we see both models doing well with the GEE beating logistic regression.

$$\rho = .99,\ b = 4$$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.051000	0.045000	0.000185	0.000262
10%	0.048500	0.056000	0.000175	0.000262
20%	0.052500	0.056000	0.000153	0.000240
30%	0.048000	0.045000	0.000139	0.000226
40%	0.043500	0.045000	0.000131	0.000222

We observe the same thing again for higher γ.

$$\rho = .999,\ b = 4$$

$HH\%$	α-GEE	α-Logistic	MSE-GEE	MSE-Logistic
5%	0.020000	0.051500	0.000182	0.000265
10%	0.030000	0.050000	0.000174	0.000266
20%	0.030000	0.048000	0.000142	0.000241
30%	0.034000	0.050500	0.000136	0.000240
40%	0.036000	0.052000	0.000125	0.000218

Again, both models are working well, except GEE in ultra-high dimensional settings.
We also look at some of the power estimates at different percentages of households with size 1.
Here we have the power when we have taken the bound as 2 here and $\rho = .9.p_1 = .05,\ p_2 = .1$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	0.994500	0.942000	0.001341	0.001361
10%	0.994000	0.939500	0.001331	0.001351
20%	0.990500	0.940000	0.001338	0.001371
30%	0.993500	0.943000	0.001345	0.001365
40%	0.983000	0.939500	0.001357	0.001404

Here we have the power when we have taken the bound as 2 and $\rho = .99.p_1 = .05,\ p_2 = .1$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	1.000000	0.933000	0.001342	0.001372
10%	1.000000	0.931500	0.001360	0.001389
20%	1.000000	0.935500	0.001353	0.001392
30%	0.998500	0.934000	0.001346	0.001392
40%	0.991000	0.941000	0.001595	0.001365

Here we have the power when we have taken the bound as 3 here and $\rho = .9.p_1 = .05,\ p_2 = .1$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	0.994000	0.932000	0.001355	0.001373
10%	0.996000	0.933500	0.001375	0.001406
20%	0.995500	0.938500	0.001374	0.001413
30%	0.989000	0.938000	0.001329	0.001375
40%	0.988000	0.934000	0.001334	0.001374

Here we have the power when we have taken the bound as 3 and $\rho = .99.p_1 = .05,\ p_2 = .1$

$HH\%$	β-GEE	β-Logistic	MSE-GEE	MSE-Logistic
5%	1.000000	0.930500	0.001396	0.001438
10%	1.000000	0.935000	0.001365	0.001425
20%	1.000000	0.927500	0.001387	0.001432
30%	0.997000	0.944500	0.001354	0.001399
40%	0.984000	0.941000	0.001355	0.001413

Combined Estimate

In this section, we look at the performance of combined estimate of GEE and Logistic. To explain further, we divide the data up into two sections- the households with size 1 and the households with size more than 1. We shall fit the Logistic model to the subset of data which have size 1 per household and the GEE model to the subset of data having size greater than 1 per household. We then take a linear combination of the estimates of the treatment effect the two models to get our combined estimate of the treatment effect. To put it in mathematical terms-

We fit the models

$$\log(\frac{E(y_{i,j})}{1-E(y_{i,j})}) = \beta_0 + \beta_g' x_{i,j},\ ,i = 1, 2, \ldots, n,\ j = 1, 2, 3, \ldots, n_i$$

the GEE model where $(y_{i1}, y_{i2}, \ldots, y_{in_i})'$ is the response in the i-th household.

And

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_l' x_i,\ i = 1, 2, 3, \ldots, n$$

$$Y_i \sim^{ind} Ber(p_i)$$

where β_g and β_l are the true treatment effects in the two models. Let the estimated treatment effects be denoted by $\hat{\beta_g}$ for GEE and $\hat{\beta_l}$ for Logistic. Denote by w the proportion of households of size 1. Denote by

$$\hat{\beta_c} = w\hat{\beta_l} + (1-w)\hat{\beta_g}$$

the combined estimate.

We know from GEE theory and the MLE theory that $\hat{\beta_g} \rightarrow^d N(\beta_g, \sigma_g^2)$ and $\hat{\beta_l} \rightarrow^d N(\beta_l, \sigma_l^2)$. Now $\hat{\beta_g}$and$\hat{\beta_l}$ are independent as we use different sets of households to model them. Thus $\hat{\beta_c} \rightarrow^d N(w\beta_l + (1-w)\beta_g, w^2\sigma_l^2 + (1-w)^2\sigma_g^2)$

Using this fact, we shall estimate type-1 error and power values for the model. Again, the simulation is done with 1368 subjects.

Here we have the Type-1 and power estimates where $\rho = .9$.

$HH\%$	α	β
5%	0.092500	0.788000
10%	0.098000	0.963000
20%	0.105500	0.999500
30%	0.103000	0.996000
40%	0.106500	0.995000
50%	0.092500	0.993000
60%	0.103500	0.991000
70%	0.093000	0.987500
80%	0.090500	0.982000

Here we have the Type-1 and power estimates where $\rho = .99$.

$HH\%$	α	β
5%	0.086000	0.789500
10%	0.108500	0.965000
20%	0.104000	0.996000
30%	0.105000	0.995000
40%	0.105000	0.994000
50%	0.101000	0.991500
60%	0.098000	0.989000
70%	0.083500	0.983000
80%	0.109500	0.965000

Estimating Correlation from Blinded Real Data

We also provide a method to estimate the correlation within household for blinded data. We shall assume that we have exchangeable correlation structure.
We use the Pearson Correlation as mentioned in literature which is given as

$$\hat{\rho_p} = \frac{1}{\hat{\mu}(1-\hat{\mu})}[\frac{\sum Z_i(Z_i-1)}{\sum n_i(n_i-1)} - \hat{\mu}^2]$$

, where $Z_i = \sum_j y_{i,j}$ and $\hat{\mu} = \frac{\sum Z_i(n_i-1)}{\sum n_i(n_i-1)}$.
See Appendix C for code.
We can also use gee model to estimate the correlation by using ”proc genmod” in SAS. The main idea here is to regress on Age and use the household structure to generate the correlation.
See Appendix C for code.
Both methods are comparable and yield nearly same results.

Conclusion

Under high correlation the GEE performs better than Logistic Regression till the household percentage hits 40%. Between 50% and 60% both models are equivalent. From 70% onwards, the GEE performs poorly and Logistic Regression performs much better. The GEE is quite robust and even when we control the upper bound of the household size the GEE does perform better than Logistic in all cases except in the case of ultra-high correlation within group and the upper bound is 2. The GEE performs better than or equivalent to logistic when the correlation is moderate to low irrespective of what the number of households with size 1 is.

References

[1]Ross L. Prentice, Correlated Binary Regression with Covariates Specific to Each Binary Observation , BIOMETRICS 44, 1033-1048, 1988.

[2]Lawrence J. Emrich & Marion R. Piedmonte, A Method for Generating High-Dimensional Multivariate Binary Variates, The American Statistician, Volume 45.

[3]Jose C. Pinheiro, Douglas M. Bates, Mixed-Effects Models in S and S-PLUS.

[4]Sheng Wu,* Catherine M. Crespi, Weng Kee Wong, Comparison of Methods for Estimating the Intraclass Correlation Coefficient for Binary Responses in Cancer Prevention Cluster Randomized Trials, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426610/

Appendix A :: Chart for Binomial Correlation

Here we present the binomial correlations with probabilities of both the random variables(correlated) as .1. The table has the normal correlation and its corresponding binary correlation.

Sr. no.	ρ Normal	γ
60	0.600000	0.320299
61	0.610000	0.333390
62	0.620000	0.338511
63	0.630000	0.352630
64	0.640000	0.351589
65	0.650000	0.354911
66	0.660000	0.370997
67	0.670000	0.383210
68	0.680000	0.392231
69	0.690000	0.401673
70	0.700000	0.412078
71	0.710000	0.419381
72	0.720000	0.426947
73	0.730000	0.442895
74	0.740000	0.448367
75	0.750000	0.457945
76	0.760000	0.466962
77	0.770000	0.483625
78	0.780000	0.488610
79	0.790000	0.502653
80	0.800000	0.518877
81	0.810000	0.526040
82	0.820000	0.544271
83	0.830000	0.554954
84	0.840000	0.563967
85	0.850000	0.567436
86	0.860000	0.595492
87	0.870000	0.605483
88	0.880000	0.617673
89	0.890000	0.632926
90	0.900000	0.653914
91	0.910000	0.674368
92	0.920000	0.687391
93	0.930000	0.707300
94	0.940000	0.731254
95	0.950000	0.751957
96	0.960000	0.776379
97	0.970000	0.814849
98	0.980000	0.841794
99	0.990000	0.886386

Now, we present the same when the probabilities are different. If one looks in the previous section, we provide bounds for the correlation based on the success probabilities.

Sr. no.	ρ Normal	γ
60	0.600000	0.284847
61	0.610000	0.291763
62	0.620000	0.300240
63	0.630000	0.308596
64	0.640000	0.321704
65	0.650000	0.325146
66	0.660000	0.325280
67	0.670000	0.340619
68	0.680000	0.351497
69	0.690000	0.355081
70	0.700000	0.368469
71	0.710000	0.379101
72	0.720000	0.394272
73	0.730000	0.392886
74	0.740000	0.402678
75	0.750000	0.409903
76	0.760000	0.409657
77	0.770000	0.428568
78	0.780000	0.433954
79	0.790000	0.443254
80	0.800000	0.464362
81	0.810000	0.476067
82	0.820000	0.476523
83	0.830000	0.500828
84	0.840000	0.505416
85	0.850000	0.525144
86	0.860000	0.527976
87	0.870000	0.540429
88	0.880000	0.551817
89	0.890000	0.566491
90	0.900000	0.583511
91	0.910000	0.594248
92	0.920000	0.603194
93	0.930000	0.614792
94	0.940000	0.634234
95	0.950000	0.645085
96	0.960000	0.655615
97	0.970000	0.672118
98	0.980000	0.685853
99	0.990000	0.685830

Appendix B::Additional Tables

We provide additional tables with type-1 error and power varying w.r.t. correlation within cluster. When 20% of the households are of size 1.

ρ	γ	α-GEE	α-Logistic
0.340000	.15	0.051500	0.051500
0.430000	.20	0.056500	0.059500
0.500000	.25	0.050000	0.050500
0.570000	.30	0.052500	0.046500
0.690000	.40	0.052500	0.046500
0.750000	.45	0.051500	0.041000
0.800000	.50	0.044000	0.052500
0.870000	.60	0.053000	0.051500

ρ	γ	β-GEE	β-Logistic
0.340000	0.150000	0.948000	0.938500
0.430000	0.200000	0.955000	0.940500
0.500000	0.250000	0.954500	0.932000
0.570000	0.300000	0.967500	0.943000
0.690000	0.400000	0.977500	0.944500
0.750000	0.450000	0.981500	0.946000
0.800000	0.500000	0.985000	0.937500
0.870000	0.600000	0.992500	0.935500

When 30% of the households are of size 1.

ρ	γ	α-GEE	α-Logistic
0.340000	.15	0.051500	0.051500
0.430000	.20	0.056500	0.059500
0.500000	.25	0.050000	0.050500
0.570000	.30	0.052500	0.046500
0.690000	.40	0.052500	0.046500
0.750000	.45	0.051500	0.041000
0.800000	.50	0.044000	0.052500
0.870000	.60	0.053000	0.051500

ρ	γ	β-GEE	β-Logistic
0.340000	0.150000	0.948000	0.938500
0.430000	0.200000	0.955000	0.940500
0.500000	0.250000	0.954500	0.932000
0.570000	0.300000	0.967500	0.943000
0.690000	0.400000	0.977500	0.944500
0.750000	0.450000	0.981500	0.946000
0.800000	0.500000	0.985000	0.937500
0.870000	0.600000	0.992500	0.935500

When 60% of the households are of size 1.

ρ	γ	α-GEE	α-Logistic
0.340000	0.150000	0.049000	0.047000
0.430000	0.200000	0.051000	0.052000
0.500000	0.250000	0.050500	0.054000
0.570000	0.300000	0.056500	0.057000
0.690000	0.400000	0.054500	0.053000
0.750000	0.450000	0.059500	0.048000
0.800000	0.500000	0.053500	0.049500
0.870000	0.600000	0.055500	0.050500

ρ	γ	α-GEE	α-Logistic
0.340000	0.150000	0.948500	0.943500
0.430000	0.200000	0.949000	0.934000
0.500000	0.250000	0.959000	0.937000
0.570000	0.300000	0.963000	0.946000
0.690000	0.400000	0.971500	0.944000
0.750000	0.450000	0.972500	0.945000
0.800000	0.500000	0.982500	0.941000
0.870000	0.600000	0.989000	0.934500

When 80% of the households are of size 1.

ρ	γ	α-GEE	α-Logistic
0.340000	.15	0.051500	0.048500
0.430000	.20	0.051500	0.051500
0.500000	.25	0.042500	0.044000
0.570000	.30	0.053500	0.057000
0.690000	.40	0.055500	0.048500
0.750000	.45	0.065000	0.053000
0.800000	.50	0.058500	0.049500
0.870000	.60	0.070500	0.054500

ρ	γ	β-GEE	β-Logistic
0.340000	0.150000	0.934000	0.932500
0.430000	0.200000	0.944000	0.944500
0.500000	0.250000	0.942000	0.947500
0.570000	0.300000	0.941000	0.952000
0.690000	0.400000	0.914500	0.937500
0.750000	0.450000	0.918500	0.942500
0.800000	0.500000	0.901000	0.941500
0.870000	0.600000	0.887000	0.940500

As we can see that at 80%, we have inflation of type-1 error as the values for correlation increases.

Appendix C::Important Code Snippets

Below, we give the function for generating correlated Bernoulli.

```
> ###function to generate correlated bernoulli
> cor.bin=function(p.vec,rho,reps,k){
+   n=length(p.vec)
+   ber.out.mat=matrix(0,n,reps)#initialization
+   Sig.1=matrix(rho,n,n)
+   diag(Sig.1)=rep(1,n)#the variance covariance matrix of the Gaussian as compound symmetry
+   out=svd(Sig.1) #SVD
+   out1=list()
+
+   for(i in 1:reps){
+     z=rnorm(n) #Random std normal generation
+     x=out$u%*%diag(out$d^{.5})%*%z #correlated normal generation
+     u.vec=pnorm(x) #transformation to uniform
+     ber.out=1*(u.vec<p.vec) #the output vector
+     ber.out.mat[,i]=ber.out #storage
+   }
+   out1[[1]]=ber.out.mat[,1:k]
+   out1[[2]]=cor(t(ber.out.mat))
+   return(out1)
+
+ }
> ###Example
>
> cor.bin(c(.09,.09),.9,1e3,4)

[[1]]
     [,1] [,2] [,3] [,4]
[1,]    0    1    0    0
[2,]    0    1    0    0

[[2]]
          [,1]      [,2]
[1,] 1.0000000 0.6606015
[2,] 0.6606015 1.0000000

> cor.bin(c(.09,.09),.99,1e3,4)

[[1]]
     [,1] [,2] [,3] [,4]
[1,]    0    1    1    0
[2,]    0    1    1    0

[[2]]
          [,1]      [,2]
[1,] 1.0000000 0.9103024
[2,] 0.9103024 1.0000000

> cor.bin(c(.09,.09),.999,1e3,4)

[[1]]
     [,1] [,2] [,3] [,4]
```

```
[1,]    0    0    0    0
[2,]    0    0    0    0

[[2]]
          [,1]      [,2]
[1,] 1.0000000 0.9892685
[2,] 0.9892685 1.0000000
```

Below, we give the function for generating truncated Poisson.

```
> new.truncated.poisson=function(n,lambda){
+   k=1
+   out=c()
+   while(k<=n){
+     y=rpois(1,lambda)
+     if(y>1){
+       out=c(out,y)
+       k=k+1
+     }else{k=k}
+   }
+   return(out)
+ }
> #Example
> new.truncated.poisson(10,4)

 [1] 4 2 4 3 4 3 4 3 5 2

>
```

Below, we give the code for estimating the intraclass correlation. We provide both the R and SAS code for this. The R code:

```
##Function to estimate Correlation
##Pearson Method
pearson.cor.fn=function(y,hh.tag){
  if(length(y)==length(hh.tag)){
    Z=aggregate(y,by=list(Category=hh.tag),FUN=sum)[,2]
    one=rep(1,length(y))
    n=aggregate(one,by=list(Category=hh.tag),FUN=sum)[,2]
    mu=sum(Z*(n-1))/sum(n*(n-1))
    rho=(1/(mu*(1-mu)))*((sum(Z*(Z-1))/sum(n*(n-1)))-mu^2)
    return(rho)
  }else{
    return("Error")
  }
}
```

The SAS code.

```
proc genmod data=df1;
class Age(ref='0') Household_Number ;/*age=0 if age<50; otherwise, age=1*/
```

```
model Visit_1(event='1') = Age/dist=bin link=logit;
repeated subject=Household_Number/type=cs;
ods output geeexchcorr=$\gamma$;
run;
```