\frac{\sum_{i =1}^{G} x_i \sum_{j =1}^{R} Y_{ij}}{R \sum_{i =1}^{G} x_i }

School of Public Health
BROWN UNIVERSITY

Optimal Study Design under Hierarchical Model: A Simulation Study

Jiahao Cui[1]
[1]Brown University, RI

Abstract

In this project, we provide a theatrical insight for the simple hierarchical linear model with binary covariate. The GLS estimator of β is the same as OLS , no matter the γ^2, σ^2 is known or unknown. For unknown scenario, the empirical variance of the estimator will closer to the true variance as the number of simulation increases. And for optimal study design, we derive a closed form for the target function which is a integer non-linear programming. Instead of optimization for minimal variance of the estimator we use grid searching for the lowest empirical variance. We also derive the distribution of $\mathcal{X}$ (inflated), marginal variance of estimator Var($\hat{\beta}$). We use "ADEMP" framework to do the simulation study for non-linear case(Poisson).

Background

We consider the setting in which Y is assumed to be normally distributed. For observation j ($j = 1, ..., R$, repeated observations) in cluster i ($i = 1, ..., G$, groups), let X_i be a binary indicator of whether or not cluster i is assigned to the treatment group with probability p, (0 = control, 1 = treatment) and let Y_{ij} be the observed outcome. To estimate the treatment effect, we will assume a hierarchical model for Y_{ij}, where the fixed effect and Y_{ij} is generated as follows:

$$\mu_{i0} = \alpha + \beta X_i, \quad \mu_i \sim N(\mu_{i0}, \gamma^2) \tag{1}$$

$$Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2), \quad Y_{ij}|X_i \sim N(\mu_{i0}, \gamma^2 + \sigma^2) \tag{2}$$

Extension to Non-Linear: in some cases, Y_{ij} follows a non-normal distribution, such as a Poisson distribution: the hierarchical model introduces over dispersion:

$$\nu_{i0} = \alpha + \beta X_i, \nu_i \sim N(\nu_{i0}, \gamma^2), \quad \mu_i = \exp(\nu_i) \sim LogN(\nu_{i0}, \gamma^2) \tag{3}$$

$$Y_{ij}|\mu_i \sim \text{Poisson}(\mu_i), \quad Y_i|\mu_i = \sum_{j=1}^{R} Y_{ij}|\mu_i \sim \text{Poisson}(R\mu_i) \tag{4}$$

$$E[Y_i] = RE[\mu_i], \quad Var[Y_i] = RE[\mu_i] + R^2 Var[\mu_i] \tag{5}$$

Optimal Design: Explore relationships between the underlying data generation mechanism parameters and the relative costs $c1/c2 (c_2 << c_1)$, budget B, how these impact the optimal study design:

$$\min_{G,R} g(G, R) = MSE(\hat{\beta}) \tag{6}$$

$$Gc_1 + G(R-1)c_2 \leq B \tag{7}$$

G and R should be a number, which is a integer nonlinear programming and it is NP-hard. We can also use the grid search to find the optimal solution. Without loss of generality, we can do a normalization (re-parametrization) on c_1, c_2, B as follows:$r_1 = \frac{c_1}{c_2} >> 1$ and $r_2 = \frac{B}{c_1}$, and the restriction are as follows:

$$Gr_1 + G(R-1) \leq r_1 r_2$$

Theory

In the linear case, the correlation between two observations within the same cluster is given by the Intra-class Correlation Coefficient (ICC): $\rho = \frac{\text{Cov}(Y_{ij}, Y_{ij'})}{\text{Cov}(Y_i)} = \frac{\gamma^2}{\gamma^2+\sigma^2}$. Also we use the notation $\sigma_Y^2 = \gamma^2 + \sigma^2$ represent the total variance of the outcome. Back to the regression case: $Y = N(\alpha + X\beta, (\gamma^2 + \sigma^2)\Sigma)$ where Σ is a block diagonal matrix, for each block $\Sigma_i, (i = 1, ..., G)$, the correlation matrix is 1 in the diagonal and the other elements are ρ (Compound Symmetry or Exchangeable), and the GLS estimator is:

$$\hat{\beta} = \left(X^\top \Sigma^{-1} X\right)^{-1} X^\top \Sigma^{-1} Y = \frac{\sum_{i=1}^{G} x_i \sum_{j=1}^{R} Y_{ij}}{R \sum_{i=1}^{G} x_i}$$

which is the same as the OLS estimator$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$. And it is not related to γ^2 and σ^2 and is unbiased estimator. The variance of GLS estimator is given by:

$$\text{Var}(\hat{\beta}|X) = (\sigma^2 + \gamma^2)(X^\top \Sigma^{-1} X)^{-1} = \frac{\sigma_Y^2}{\sum_{i=1}^{G} x_i} \times (\rho + \frac{1-\rho}{R}) \tag{8}$$

conditional on X it only depends on γ^2 and σ^2 or their estimator. In order to explore the relationship between the variance of the estimator and the number of G and R, we derive the "marginal" variance of the estimator: Var($\hat{\beta}$). When X is a vector with all the same elements zero or one, it will not bring any information for the regression coefficient. Therefore, we need to consider the distribution of X has a zero probability on that which is similar as the zero-inflated distribution, we noted as $\mathcal{X}$.

$$P(\mathcal{X} = x) \propto \begin{cases} 0 & \text{if } \mathcal{X} \text{ is all zero or one,} \\ P(X = x) & \text{otherwise} \end{cases}$$

Then the marginal variance of the estimator is given below:

$$\text{Var}(\hat{\beta}) = E_{\mathcal{X}|G,R}[\text{Var}(\hat{\beta}|X)] = \sigma_Y^2 \times (\rho + \frac{1-\rho}{R}) E_{\mathcal{X}|G,R}[\frac{1}{\sum_{i=1}^{G} x_i}] \tag{9}$$

$$= \frac{\sigma_Y^2 (\rho + \frac{1-\rho}{R})}{(1 - 2p^G)} \sum_{k=1}^{G-1} \frac{1}{k} \binom{G}{k} p^k (1-p)^{G-k} \tag{10}$$

If the ρ is known, but the σ_Y^2 is unknown, we can still get the estimator of $\hat{\sigma_Y^2}$ through OLS formula. When both are unknown, the estimator of variance coefficient $\hat{\sigma^2}, \hat{\gamma^2}$ or equivalently $\hat{\sigma_Y^2}, \hat{\rho^2}$is calculated through Likelihood-Based Estimation or Restricted maximum likelihood method.

For the poisson case, the intra-class correlation coefficient (ICC) is given by:

$$\rho = \frac{\text{Cov}(Y_{ij}, Y_{ij'})}{\text{Cov}(Y_{ij})} = \frac{e^{2\nu_{i0}+\gamma^2}(e^{\gamma^2} - 1)}{e^{\nu_{i0}+\frac{\gamma^2}{2}} + e^{2\nu_{i0}+\gamma^2}\left(e^{\gamma^2} - 1\right)} \tag{11}$$

Simulation

Table 1: Optimal Design for linear case when $p = 0.5, \sigma_Y^2 = 1$

ρ	$r1$	$r2$	G*	R*	**Var**	Var_{true},$Var_{empirical}$,$Var_{estimate}$	Rank	Difference Percentage
0.10	5	5	2	8	.2125	.2125/.2402/.2748(.2125/.2516/.2266)	1/1/2(1/1/1/[3])	0/0/23.0106(0/0/0)
0.10	5	20	10	6	.0573	.0601/.0625/.0703(.0584/.0586/.0548)	4/4/9(3/2/2/[18])	6.0335/9.7894/25.8622(6.2683/2.3677/3.1401)
0.10	5	100	50	6	.0102	.0103/.0110/.0107(.0103/.0102/.0101)	1/15/16(2/2/2/[98])	0/9.9657/18.8561(0.7249/0.7553/1.1393)
0.10	20	5	3	14	.1232	.1166/.1176/.0961(.1235/.1299/.1185)	1/1/1(1/1/1/[3])	0/0/0(0/0/0)
0.10	20	20	12	14	.0306	.0288/.0279/.0248(.0294/.0292/.0282)	1/1/2(1/1/1/[18])	0/0/12.5384(0/0/0)
0.10	20	100	62	13	.0056	.0057/.0053/.0051(.0056/.0055/.0056)	8/1/5(5/2/10/[98])	2.663/0/8.6202(0.7088/0.0683/12.9337)
0.10	100	5	4	26	.0801	.0987/.0974/.0911(.0953/.1016/.0895)	2/2/2(2/2/2/[3])	21.9816/25.6232/30.4602(11.6531/16.9766/1.3944)
0.10	100	20	16	26	.0182	.0185/.0184/.0168(.0185/.0184/.0183)	2/3/1(2/2/3/[18])	7.6515/1.6634/0(5.5075/4.4651/7.918)
0.10	100	100	78	29	.0034	.0034/.0034/.0030(.0034/.0034/.0035)	1/1/5(2/4/10/[98])	0/0/13.2487(0.2575/1.4506/5.1785)
0.50	5	5	4	2	.4464	.5162/.5703/.5732(.5214/.5206/.5306)	2/3/3(2/2/3/[3])	10.64/31.3105/53.2562(13.818/5.5996/11.6543)
0.50	5	20	16	2	.1012	.1006/.1040/.1206(.1035/.1045/.0977)	1/2/6(1/1/1/[18])	0/4.2053/38.154(0/0/0)
0.50	5	100	83	2	.0183	.0182/.0183/.0274(.0183/.0184/.0181)	2/3/59(1/1/1/[98])	.1862/1.2239/73.3736(0/0/0)
0.50	20	5	4	6	.3472	.4083/.4005/.4245(.4084/.4193/.3882)	2/2/2(2/2/2/[3])	1.8857/2.3525/10.1142(2.8394/5.7902/10.888)
0.50	20	20	16	6	.0787	.0810/.0773/.0817(.0817/.0803/.0875)	2/2/4(4/2/5/[18])	6.5552/6.548/13.9875(4.9116/2.795/15.7189)
0.50	20	100	83	5	.0146	.0148/.0148/.0132(.0148/.0148/.0139)	3/6/6(1/3/2/[98])	1.3294/1.6124/19.8726(0/0.4231/2.2043)
0.50	100	5	4	26	.3091	.3721/.3676/.3592(.3610/.3668/.3527)	1/2/1(1/1/1/[3])	0/15.9359/0(0/0/0)
0.50	100	20	18	12	.0643	.0631/.0620/.0560(.0624/.0630/.0595)	1/1/3(1/1/1/[18])	0/0/2.3479(0/0/0)
0.50	100	100	90	12	.0122	.0123/.0123/.0137(.0121/.0121/.0119)	5/6/26(2/1/6/[98])	.9471/1.1788/39.2659(0.0939/0/3.0045)
0.99	5	5	4	2	.5923	.7280/.7335/.9247(.6980/.6894/.7026)	1/2/2(1/1/1/[3])	0/0.8903/34.982(0/0/0)
0.99	5	20	19	1	.1120	.1151/.1121/.0789(.1134/.1136/.1141)	2/2/1(1/1/1/[18])	1.1942/3.3137(0/0/0/0)
0.99	5	100	99	1	.0204	.0203/.0202/.0241(.0203/.0203/.0225)	1/1/14(1/1/7/[98])	0/0/46.2597(0/0/11.7036)
0.99	20	5	4	6	.5903	.6859/.5876/.7230(.6988/.6872/.7438)	1/1/2(1/1/1/[3])	0/0/19.2385(0/0/0)
0.99	20	20	19	2	.1114	.1109/.1076/.1180(.1134/.1141/.1215)	1/1/2(1/1/1/[18])	0/0/13.5713(0/0/0)
0.99	20	100	99	1	.0204	.0200/.0201/.0179(.0203/.0202/.0214)	1/1/1(1/1/5/[98])	0/0/0(0/0/16.1999)
0.99	100	5	4	26	.5895	.7048/.6822/.5676(.7028/.6790/.7520)	1/1/1(1/1/1/[3])	0/0/0(0/0/0)
0.99	100	20	19	6	.1110	.1113/.1165/.0929(.1133/.1127/.1091)	1/2/2(1/1/1/[18])	0/3.2372/8.83(0/0/0)
0.99	100	100	99	2	.0203	.0200/.0198/.0178(.0203/.0202/.0198)	1/1/4(1/1/2/[98])	0/0/13.0957(0/0/2.5395)

In the simulation study, we can get four kinds of variance for $\hat{\beta}$. One is the theoretical variance Var_{true}, which is calculated by the true γ^2, σ^2 in equation [8] , the second is theoretical marginal variance **Var** which is taking expectation on distribution of $\mathcal{X}$ from equation [10], the third is the empirical variance $Var_{emperical}$, which is calculated through the process of simulation, and the last is the variance $Var_{estimate}$ calculated by plugging in the other estimator.

The optimal $G^*, R^* = argmin\textbf{Var}(\hat{\beta})$. The "Rank" is the optimal G^*, R^* from marginal variance Var, represent the rank in other variance type. Difference Percentage = $[Var(G^*, R^*) - min(Var)]/\textbf{Var}$. We get the results from 100 simulation(1000 simulation), when $\alpha = 0, \beta = 1$. Without considering the situation $R = 1, G = 1$, the range is $2 \leq G \leq r2 - 1$.

For Poisson case, $G^*, R^* = argmin Var_{emperical}(\hat{\beta})$. $\rho_0 and \rho_1$ is calculated through equation [11] when $X = 0, 1$.

Table 2: Optimal Design for Poisson case when $r1 = r2 = 20$

α	β	γ^2	P	G*	R*	Bias	SE_{est}	Rank	Diff	SE_{emp}	MSE	ρ_0	ρ_1	Ratio
0.1	0.1	0.1	0.3	8	31	-0.041	0.23	5	0.132	0.254	0.066	0.109	0.119	-6.261
0.1	0.1	0.1	0.7	13	11	-0.004	0.24	4	0.14	0.247	0.061	0.109	0.119	-66.423
0.1	0.1	1	0.3	14	9	-0.106	0.584	5	0.468	0.614	0.384	0.758	0.776	-5.766
0.1	0.1	1	0.7	15	7	-0.021	0.618	7	0.507	0.636	0.401	0.758	0.776	-29.645
0.1	1	0.1	0.3	13	11	-0.043	0.227	10	0.143	0.23	0.054	0.109	0.249	-5.348
0.1	1	0.1	0.7	13	11	-0.009	0.237	3	0.152	0.226	0.051	0.109	0.249	-24.325
0.1	1	1	0.3	15	7	-0.039	0.597	7	0.495	0.579	0.333	0.758	0.895	-14.779
0.1	1	1	0.7	18	3	0.008	0.557	3	0.456	0.545	0.294	0.758	0.895	71.335
1	0.1	0.1	0.3	16	6	0.008	0.211	7	0.147	0.209	0.043	0.231	0.249	25.595
1	0.1	0.1	0.7	18	3	0.001	0.244	16	0.18	0.226	0.05	0.231	0.249	301.057
1	0.1	1	0.3	19	2	0.028	0.551	5	0.475	0.56	0.31	0.885	0.895	20.036
1	0.1	1	0.7	19	2	0.078	0.505	3	0.433	0.576	0.335	0.885	0.895	7.437
1	1	0.1	0.3	17	4	0.002	0.194	3	0.141	0.186	0.034	0.231	0.45	107.36
1	1	0.1	0.7	14	9	-0.026	0.211	5	0.158	0.216	0.047	0.231	0.45	-8.472
1	1	1	0.3	19	2	-0.006	0.518	4	0.456	0.557	0.308	0.885	0.954	-86.129
1	1	1	0.7	19	2	0.021	0.554	5	0.489	0.53	0.278	0.885	0.954	24.875

Research was supported by NIH R01 MH75639. ClinicalTrials.gov Identifier: NCT00644150