Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

I^2 Statistic for One-Stage Individual Participant Data Meta-Analysis with Binary Outcomes

Andrea Benedetti
McGill University

Zhaoxue Tong
Renmin University of China

Abstract

Quantifying heterogeneity is one of the most challenging aspects of an individual participant data meta-analysis with binary outcomes. Although theories about the concept of I^2 have been established, there is no published statistical software available that can calculate the I^2 statistic. Using generalized linear mixed model (GLMM), we propose a simulation-based method to estimate the I^2 statistic. In this work, we demonstrate a SAS macro that will calculate the I^2 statistic for a GLMM that has been fitted with SAS's `GLIMMIX` procedure.

Keywords: GLMM, simulation, I^2 statistic, SAS macro.

1. Introduction

Assessing heterogeneity is an important component in any individual participant data meta-analysis. For individual participant data (IPD) with binary outcomes, meta-analysis is generally performed using a two-stage or a one-stage method. In the two-stage method, each study is summarized separately to generate aggregate data (AD) (such as odds ratios, risk differences, etc.) in the first stage, then these AD are combined across studies using standard meta-analytic techniques in the second stage. In the one-stage method, IPD from all studies are combined and analyzed in a single step by fitting a mixed model.

Several metrics for addressing the extent of heterogeneity in the two-stage method have been proposed, such as between-study variance (τ^2), R statistic and I^2 statistic (Zhou and Dendukuri 2014; Higgins and Thompson 2002). While there is no widely accepted metric for the one-stage method, the concept of I^2 can be applied and an analogue has been formulated. The goal here is to demonstrate a SAS macro that will estimate an I^2 in a one-stage IPD-MA of binary outcomes.

2. I^2 statistic

The I^2 statistic was proposed by Higgins and Thompson (2002) for indicating the extent of heterogeneity in a meta-analysis. Consider k studies for the parameter of interest θ. We denote the true value of θ in study $i\,(i = 1, \cdots, k)$ by θ_i, and its estimate by $\hat{\theta}_i$. The precision of $\hat{\theta}_i$, denoted as w_i, is defined as the reciprocal of $\hat{\theta}_i$'s variance. Higgins and Thompson assumed that the sampling variances σ_i of $\hat{\theta}_i$s from each study are known and equal ($\sigma_i = \frac{1}{w_i}, i = 1, \cdots, k$), $\mathsf{E}(\theta_i) = \mu$, $\mathsf{VAR}(\theta_i) = \tau^2$, $\mathsf{E}(\hat{\theta}_i|\theta_i) = \theta_i$ and $\mathsf{VAR}(\hat{\theta}_i|\theta_i) = \sigma^2$, where k is known, μ and τ^2 are unknown and σ^2 is assumed known. Under these assumptions, they defined the I^2 statistic of the form

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \tag{1}$$

where

$$\hat{\tau}^2 = \hat{\sigma}^2\left(\frac{\sum w_i(\hat{\theta}_i - \frac{\sum w_i\hat{\theta}_i}{\sum w_i})^2}{k-1} - 1\right) \tag{2}$$

$$\hat{\sigma}^2 = \frac{k-1}{\sum w_i - \frac{\sum \hat{w_1}^2}{\sum w_i}} \tag{3}$$

Here $\hat{\tau}^2$ and $\hat{\sigma}^2$ estimate the between-study variance and the within-study variance, respectively.

In clustered data analyses, the intraclass correlation coefficient (ICC)(Wu, Crespi, and Wong 2012) is formulated as the proportion of the between-cluster variance over the total variance in the outcome, and is similar to the I^2 statistic. Goldstein, Browne, and Rasbash (2002) estimated the ICC for binary outcomes with a simulation-based approach that partitions variation in multilevel models. They considered a 2-level variance component model with a binary response and a fixed effect exposure status:

$$\mathsf{E}(y_{ij}) = \pi_{ij} = g(\beta_0 + \beta_1 x_{ij} + u_{0i}) \tag{4}$$

$$y_{ij} \sim Bernoulli(\pi_{ij}) \tag{5}$$

$$u_{0i} \sim N(0, \sigma^2_{u0}) \tag{6}$$

where y_{ij} is the binary outcome, and $g(\cdot)$ is the logistic function. Here the level-1 variance (between-study variance) depends on $\mathsf{VAR}(y_{ij}) = \pi_{ij}(1-\pi_{ij})$ and the level-2 variance (within-study variance) is measured by σ^2_{u0}. The simulation method is as follows:

1. Fit the model to the data.
2. Using the sample estimate of the variance from the model fitted in step 1, simulate a large number of level-2 residual u_{0i} from $N(0, \sigma^2_{u0})$.
3. Estimate π_{ij} by using the fitted model in step 1, and for each of these values compute the level-1 variance $v_{1ij} = \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})$.

4. Using the results in step 3, compute the level-1 variance $v_1 = \mathsf{E}(v_{1ij})$ and the level-2 variance $v_2 = \mathsf{VAR}(\hat{\pi}_{ij})$. Then ICC is estimated as

$$ICC = \frac{v_2}{v_2 + v_1} \tag{7}$$

3. Estimating I^2 statistic

Chen and Benedetti (2017) proposed using the simulation based ICC estimator of Goldsten as an estimate of the I^2. The following algortihm was proposed:

$$\mathsf{E}(y_{ij}) = \pi_{ij} = g(\beta_0 + u_{0i} + (\beta_1 + u_{1i})x_{ij}) \tag{8}$$

$$y_{ij} \sim Bernoulli(\pi_{ij}) \tag{9}$$

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim MVN(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{00} & \sigma^2_{01} \\ \sigma^2_{01} & \sigma^2_{11} \end{bmatrix}) \tag{10}$$

This generalized linear mixed model (GLMM) allows both the intercept and the slope to vary across studies by adding the random effects u_{0i} and u_{1i}, which are assumed to have a multivariate normal distribution with parameters shown in (10). π_{ij} denotes the expectation of the binary outcome y_{ij} of the jth individual in the ith study, and is defined as the probability of having $y = 1$. x_{ij} is the exposure status and β_1 is the parameter representing the pooled log odds ratio.

To obtain the I^2 the following algorithmn is followed:

1. Fit a GLMM to the data.
2. Using the sample estimate of the covariance matrix from the model fitted in step 1, simulate a large number of random effects u_{0i} and u_{1i} from (10). We denote these as u_{0ij} and u_{1ij}.
3. Estimate the log odds ratio for each subject ($v_{1ij} = \hat{\beta}_1 + u_{1ij}$) in the dataset by using the model fitted in step 1. The between-study variance is estimated as $v_1 = \mathsf{VAR}(v_{1ij})$.
4. Estimate π_{ij} by using the fitted model in step 1 and the simulated random effect values in step 2.
   (a) replacing (u_{0i}, u_{1i}) by (u_{0ij}, u_{1ij}) for each subject.
   (b) plugging in the fixed effect estimator from the fitted model and the covariates from the dataset.
   (c) compute $\hat{\pi}_{ij}$ for each individual from (8).
5. Compute the variance of the estimated log odds ratio by using the results in Step 4 and the Delta method, $v_{2ij} = \frac{1}{n\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}$, where n is the average number of subjects among all the studies. The within-study variance is estimated as $v_2 = \mathsf{E}(v_{2ij})$.

Argument	Description
`dataset`	Required; the input dataset
`study_id`	Required; the variable in the dataset identifying each study
`exposure`	Required; the variable in the dataset indicating the exposure status
`paraest_mat`	Required; the name given to the '`ParameterEstimates=`' data set in the `ODS OUTPUT` statement; stores the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$
`g_mat`	Required; the name given to the '`G=`' data set in the `ODS OUTPUT` statement; stores the estimated covariance matrix of u_{0i} and u_{1i}

Table 1: Arguments for the macro `I_Squared`

6. The I^2 statistic is estimated as

$$I^2 = \frac{v_1}{v_1 + v_2} \tag{11}$$

Chen and Benedetti (2017) compared several ways to quantify heterogeneity in the context of IPD-MA of binary data. Their results demonstrated that when there was no effect modification, the estimated I^2 statistic calculated above was overestimated. When there was effect modification, the estimated I^2 from the one-stage IPD-MA performed better than that from the two-stage model.

4. Software, syntax and parameters

The SAS macro for computing the I^2 statistic is called `I_Squared`. It takes as input several outputs that `PROC GLIMMIX` produces while fitting a GLMM and parameters specifying information about the data. Then it uses the information and the input data sets to compute the I^2 statistic following the algorithm discussed earlier, and displays the results. The user should follow the procedure listed below:

1. Run the `PROC GLIMMIX` step to fit a GLMM.
2. Modify step 1 so that it will produce the output datasets that the macro requires as input.
3. Run the modified `PROC GLIMMIX` step.
4. Invoke the `I_Squared` macro.

There are five parameters that need to be entered. See Table 1 for more details about the arguments. To invoke the macro, one must issue the following code with the parameters filled in.

```
%I_Squared(dataset=, study_id=, exposure=, paraest_mat=, g_mat=);
```

To use the macro, the user should either copy the macro code into his/her SAS program or use the "%include" statement at the beginning of a SAS program to load the macro `I_Squared`. The output of the macro is the estimated I^2 statistic.

Variable	Description
`age`	Age; non-negative integer
`gender`	Gender; binary, 1 for female and 0 for male
`HIV`	HIV infection status; binary, 1 for positive and 0 for negative
`extent`	Extent of disease; binary, 1 for extensive and 0 for non-extensive
`Group5_tx`	Use of group 5 drugs; binary, 1 for use and 0 for non-use
`study`	Identifies the study to which each patient belongs to; positive integer
`label`	Treatment outcome; binary, 1 for success and 0 for unsuccess (treatment success is defined as treatment cure or completion, and success is treatment failure or relapse)

Table 2: Variables in `g_five`

The macro is able to check whether the input estimated covariance matrix of u_{0i} and u_{1i} is positive definite. Note that as assumed in (8), u_{0i} and u_{1i} are jointly Gaussian, which implies that the covariance matrix of u_{0i} and u_{1i} should be positive definite. If the input matrix fails to meet this requirement, the macro will print an error message and stop execution.

5. Application

Through this example, we demonstrate the input, output and the functionality of the `I_Squared` macro.

The data set used was an individual patient data meta-analysis of factors contributing to multidrug resistant tuberculosis (MDR-TB) treatment success (Ahuja, Ashkin, Avendano, Banerjee, Bauer, Bayona, Becerra, Benedetti, Burgos, Centis *et al.* 2012). It contains MDR-TB treatment information and outcomes from 30 studies, as well as patient level information including age, sex, HIV infection, extent of disease, etc. Patients were considered as clustered within studies, and intercepts and slopes of the main drug-exposure variables were allowed to vary across studies in order to account for inter-study differences in patient populations. Note that the estimate of effect of the treatment parameter was adjusted for four covariates: age, gender, HIV co-infection, and extent of disease. Missing values of these four clinical covariates were imputed using the modes of patients at the same center with non-missing values.

In this example, we consider the association of use of group 5 drugs (see Ahuja *et al.* (2012) for detailed definition) with treatment success. The data set used in this analysis is named `g_five` and the variables are listed in Table 2. All observations in the input dataset should be used to fit the GLMM model, since `I_Squared` takes the number of observations in the dataset as the number of subjects among all the studies. If the input dataset contains observations with missing values, and is used in `PROC GLIMMIX` without cleaning, then the records with missing values should be removed before `I_Squared` takes it as input.

At the beginning of the SAS Program, a statement "%include 'c:\i2\I_squared'; " (assume that the macro is saved in the directory c:\i2) is issued to invoke the macro procedure to avoid copying the macro code as part of the program.

The first step is estimating the GLMM via `PROC GLIMMIX` as usual. Note that we need to fit

Covariance Parameter Estimates	
Covariance	Estimate
Variance of the random intercept	1.07
Covariance between random intercept and random slope	0.04
Variance of the random slope	0.54

Solutions for Fixed Effects[1]		
Effect	aOR	95%CI
`Intercept`	0.02	(0.01-0.05)
`Group5_tx`	1.75	(1.09-2.81)
`age`	1.01	(1.00-1.01)
`gender`	1.02	(0.86-1.22)
`HIV`	0.79	(0.55-1.13)
`extent`	2.73	(1.80-4.14)

Table 3: Part of the result in the `PROC GLIMMIX` step

a GLMM with a random intercept, a random slope for the main exposure variable `Group5_tx` and a fixed slope for the four covariates. The SAS code which would fit such a model is below:

```
proc glimmix
     data=g_five;
     model label=Group5_tx age gender HIV extent / solution cl dist=binary;
     random int Group5_tx / sub=study_id type=un g;
     ods output ParameterEstimates=para_est G=g_matrix;
run;
```

The highlighted parts request the output required by the macro:

- The option `s` or `solution` is added to the `model` statement after a forward slash so that the output data set containing the fixed-effect solutions will be produced.
- The option `g` is added to the random statement so that the output data set containing the estimated random-effects covariance matrix will be produced.
- The `ods output` statement is added to name the output data sets that will be used as input to the macro.

Part of the result is listed in Table 3. The adjusted odds ratio of the variable `Group5_tx` indicates the effect of the drug. In this example, both the aOR and the 95%CI suggest that the group 5 drugs are protective, that is, increase the odds of treatment success.

If the fit is successful, and the estimated random-effects covariance matrix is positive definite, then the final step is to invoke the macro `I_Squared`, supplying to it the names of the two data

[1] aOR: adjusted odds ratio. CI: confidence interval. The original output is transformed by the natural exponencial function (i.e. $x \mapsto e^x$) to produce aOR and CI.

Estimated I^2 statistic	
	i_squared_est
Group5_tx	0.60

Table 4: Estimated I^2 statistic for the example mode

sets that were created in the third step. In our case, the macro call should be the following one:

```
%I_Squared( dataset=g_five,
            study_id=study,
            exposure=Group5_tx,
            paraest_mat=para_est,
            g_mat=g_matrix);
```

The output of the macro is the estimated I^2 statistic. It is reported in Table 4. There are several widely used benchmarks for I^2, for instance, I^2 values of 25%, 50%, and 75% have been interpreted as small, moderate and high levels of heterogeneity. The value of I^2 in this case suggests that there is moderate level of heterogeneity, in other words, the impact of using group 5 drugs indicated by aOR varies from study to study though not that widely.

6. Conclusion

Here we have explained the computational methods necessary to estimate a simulation-based I^2 statistic, which serves to quantify the heterogeneity in an individual participant data meta-analysis with binary outcomes. The macro takes the output of SAS's GLMMIX procedure, estimates a GLMM from the data, and calculates the I^2 statistic following the algorithm proposed by Chen and Benedetti (2017).

Computational details

The results in this paper were obtained using SAS 9.4 for Windows. The most recent version can be found at https://github.com/Hepdrey/SAS.git.

References

Ahuja SD, Ashkin D, Avendano M, Banerjee R, Bauer M, Bayona JN, Becerra MC, Benedetti A, Burgos M, Centis R, *et al.* (2012). "Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: an individual patient data meta-analysis of 9,153 patients." *PLoS medicine*, **9**(8), e1001300.

Chen B, Benedetti A (2017). "Quantifying heterogeneity in individual participant data meta-analysis with binary outcomes." *Systematic reviews*, **6**(1), 243.

Goldstein H, Browne W, Rasbash J (2002). "Partitioning variation in multilevel models." *Un-*

derstanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, **1**(4), 223–231.

Higgins J, Thompson SG (2002). "Quantifying heterogeneity in a meta-analysis." *Statistics in medicine*, **21**(11), 1539–1558.

Wu S, Crespi CM, Wong WK (2012). "Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials." *Contemporary clinical trials*, **33**(5), 869–880.

Zhou Y, Dendukuri N (2014). "Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: The case of meta-analyses of diagnostic accuracy." *Statistics in medicine*, **33**(16), 2701–2717.

Affiliation:

Andrea Benedetti
Department of Medicine
and
Department of Epidemiology, Biostatistics and Occupational Health
McGill University
5252 boul de Maisonneuve, Montreal, QC, Canada
E-mail: andrea.benedetti@mcgill.ca
URL: https://www.mcgill.ca/epi-biostat-occh/andrea-benedetti

Zhaoxue Tong
Department of Mathematics
School of Information
Renmin University of China
No.59 ZhongGuanCun Street, Haidian District, Beijing, China
E-mail: tzx.ruc@hotmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd