2値データのIntraclass Correlation Coefficientの推定マクロプログラム

〇稲葉　洋介[1]、田中　紀子[1]

[1]国立国際医療研究センター　データサイエンス部
生物統計研究室

Macro program for calculating Intraclass Correlation Coefficient for binary data

Yosuke Inaba, Noriko Tanaka
Biostatistics Section, Department of Data Science
National Center for Global Health and Medicine

要旨：

アウトカムが2値の場合のIntraclass Coefficient Correlation（級内相関係数：ICC）について，幾つかの異なる推定量をSASで実装し，マクロプログラムを作成した．また条件を変化させて性能を比較し，推定量の選択の基準等を議論する。

キーワード： Intraclass Coefficient Correlation, ,ICC, Binary, Survival Cluster Randomized trial

ICCとは

ICCとは

- Intraclass Correlation Coefficient(級内相関係数:ICC)
  - 同一クラスター(定義された群、あるいは繰り返し測定された個体)内のデータの相似性を示す定量的な指標

応用

– 評価者間信頼性の指標
  - 精神医学的測定
  - 病理、画像診断
– クラスターランダム化試験
  - 検出力に影響する要因（murray 2004）
  - 生存時間をエンドポイントとする試験では、イベント有無の2値データで計算したICCを用いる場合が多い

Correlated binary data

ICC=0.1

Cluster	y
1	1 1 0 1 1 1 1 0 1 1
2	1 0 1 1 1 0 1 1 1 1
3	0 0 0 1 1 0 0 1 1 0
4	1 0 0 1 0 1 1 1 0 1
5	1 1 1 0 0 1 1 1 1 0
6	0 1 1 0 0 0 0 1 1 1
7	0 0 1 0 0 1 1 1 1 1
8	1 1 0 1 0 0 0 0 1 1
9	0 1 0 0 0 1 0 0 0 0
10	1 0 1 0 1 0 1 1 1 1

ICC=0.9

Cluster	y
1	1 1 1 1 1 1 1 1 1 1
2	0 0 0 0 0 0 0 0 0 0
3	1 1 1 1 1 1 1 1 1 1
4	1 0 0 0 0 0 0 0 0 0
5	1 1 1 1 1 1 0 1 1 1
6	0 0 1 0 0 0 0 0 0 0
7	0 0 0 0 0 0 0 0 0 0
8	0 0 0 0 0 0 0 0 0 0
9	0 0 0 0 0 0 0 0 0 0
10	1 1 1 1 1 1 1 1 1 1

定義

ICCの定義

k：クラスター数

n_i：i番目のクラスターの要素数

$X_{ij} \sim Bernulli(\pi)\ i = 1, \ldots, k\ j = 1, \ldots, n_i$: 応答（1: 成功 or 0: 失敗）

$Y_i = \sum X_{ij}$：i番目のクラスターの成功の総数

ICCの定義

- 全てのオブザベーションにおいて成功確率は共通($Pr\left(X_{ij} = 1\right) := \pi \text{ for } \forall i, j$)とし，異なるクラスター間のオブザベーションは全て独立と仮定する.
- 同一クラスター内では，任意のオブザベーションの組(X_{ij}, X_{il})の相関は共通とし，(common-correlation model)

$\mathrm{Corr}\left(X_{ij}, X_{il}\right) \coloneqq \rho$をICCと定義する

ICCの定義

- この時、$E[X_{ij}] = \pi, Var[X_{ij}] = \pi(1-\pi)$より
  $$Var[Y_i] = n_i\pi(1-\pi)\{1 + (n_i - 1)\rho\}$$

- Y_iは二項分布に従うが$\rho > 0$の時に過大分散、$\rho < 0$の時に過少分散となる。

ICCの定義（一般化線形モデル）

$\beta_i \sim N\left(0, \sigma_\beta^2\right)$：クラスターの変量効果、
$e_{ij} \sim N(0, \sigma^2)$：誤差

$$X_{ij} = \mu + \beta_i + e_{ij}$$

$$Var\left[X_{ij}\right] = Var\left[X_{ij}\right] = \sigma_\beta^2 + \sigma^2$$

ゆえに$\mathrm{Corr}\left(X_{ij}, X_{il}\right) = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma^2}$= ICC

ICCマクロについて

ICCの計算方法

- マクロ化の動機
  - 2値データのICCは出版されているものだけで10種類以上（Ridout 1996）
  - SAS/STATで直接計算できるプロシージャは無い

ICCマクロ詳細

記載例:

%Binary_ICC(

Input=library.raw2,Cluster=region,Response=y, Output=result,Method=aov aovs peq,CI=Y,iter=100)

仕様の詳細は資料集参照

Output

OBS	method	rho_hat	boot_lower	boot_upper
1	aov	.005851170	−0.016595	0.037237
2	aovs	.002868234	−0.018023	0.029335
3	peq	.	.	.

Method glmm

- nlmixedプロシージャを使用
- データによっては収束しない事があるので注意

注意事項

- 以下の状況で、Methodによっては計算できないため欠損値が返る
  - 全ての応答が0 or 全ての応答が1
  - 全てのクラスターサイズが1
- Moment estimator（keq,keqs etc）は推定値が1を超える場合がある。
  - 本マクロでは計算値をそのまま出力している

シミュレーション

シミュレーションデータ詳細

- 各推定量の挙動を確認するため、シミュレーションを実施した。
- 以下の条件でシミュレーションデータを生成した。（Rの関数rcbinをIMLで呼び出して使用）
  - π : 0.1, 0.5, 0.9
  - Num of Cluster : 10, 30, 50
  - Cluster Size（平均、20%で変動）: 10, 30, 50
  - ρ :0.1, 0.5, 0.9
  - 各設定の反復回数は100回

データ生成方法

– 以下の混合2項分布よりシミュレーションデータを生成した。(Rのrcbin関数をSAS/IMLにより呼び出して使用)

- $Z_i, \nu_{ij} \sim bernulli(\pi), \mu_{ij} \sim bernulli(\sqrt{\rho})$

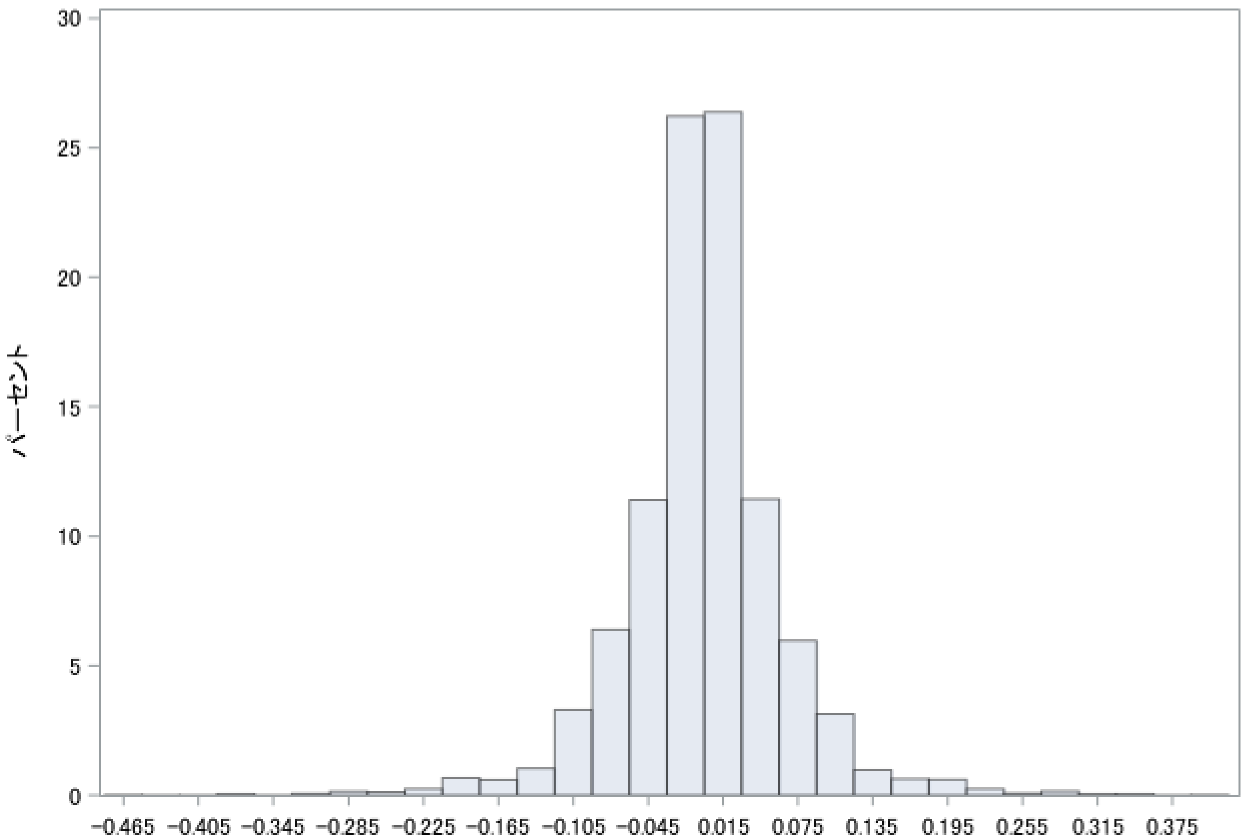
$$y_{ij} = (1 - \mu_{ij})\nu_{ij} + \mu_{ij} Z_i$$

ただし、以下の場合は除外

- $Y_i = 1$ for $\forall k$
- $Y_i = n_i$ for $\forall k$
- $n_i = 1$ for $\forall k$

シミュレーションエラー

πの設定値と生成されたシミュレーションデータの$\tilde{\pi}$の差の分布



シミュレーション結果

- 41サンプルについて、glmmは推定が収束せず推定値が得られなかった。
  - 41/8100なので、結果に大きな影響なし
- 結果の詳細は配布資料参照

シミュレーション考察

- aov, aovs, fc, mak, peq, pgp, ppr, stab, ub の推定値は、設定に関わらず精度が良い。
- keq, keqs, kpr, kprs, w, wsは、ρの値が大きくクラスターサイズが小さいと推定が不安定になっている。
- glmmはクラスターサイズが小さい時にややバイアスが入るが概ね安定的

まとめ

- ICCを計算するマクロを実装した。
- 各推定量の挙動をシミュレーションにより確認した。
- 今後は機能拡充を目指す
  - 推定量の追加（Extended quasi-likelihood estimator等）
  - 他の方法による信頼区間
  - Glmmでの詳細な指定

参照文献

- FISHER, R. A. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925.
- DONNER, Allan. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. International Statistical Review/Revue Internationale de Statistique, 1986, 67-82.

参照文献

- DONOVAN, A. M.; RIDOUT, M. S.; JAMES, D. J. Assessment of somaclonal variation in apple. II. Rooting ability and shoot proliferation in vitro. Journal of horticultural science, 1994, 69.1: 115-122.

参照文献

- GIBSON, G. J.; AUSTIN, E. J. Fitting and testing spatio - temporal stochastic models with application in plant epidemiology. Plant Pathology, 1996, 45.2: 172-184.
- RIDOUT, Martin S.; DEMETRIO, Clarice GB; FIRTH, David. Estimating intraclass correlation for binary data. Biometrics, 1999, 55.1: 137-148.

Back up

- $Z_i, \nu_{ij} \sim bern(\pi), \mu_{ij} \sim bern\left(\sqrt{\rho}\right)$

$$y_{ij} = \left(1 - \mu_{ij}\right)\nu_{ij} + \mu_{ij} Z_i$$

$y_{ij} \sim bernulli(\pi), ICC = \rho$の証明:

$$\Pr\left(y_{ij} = 0\right)$$
$$= \Pr\left(\mu_{ij} = 0, \nu_{ij} = 0\right) + \Pr\left(\mu_{ij} = 1, Z_i = 0\right)$$
$$= (1 - \sqrt{\rho})(1 - \pi) + \sqrt{\rho}(1 - \pi) = (1 - \pi)$$
$$\Pr\left(y_{ij} = 1\right)$$
$$= \Pr\left(\mu_{ij} = 0, \nu_{ij} = 1\right) + \Pr\left(\mu_{ij} = 1, Z_i = 1\right)$$
$$= (1 - \sqrt{\rho})\pi + \sqrt{\rho}\pi = \pi$$

Back up

- $E[y_{ij}] = 0 * (1 - \pi) + 1 * \pi = \pi$
- $Var[y_{ij}] = (1 - \pi)(0 - \pi)^2 + \pi(1 - \pi)^2 = \pi(1 - \pi)$

Back up

- $ICC = Corr\left(y_{ij}, y_{il}\right) = \frac{Cov(y_{ij}, y_{il})}{\sqrt{Var(y_{ij})Var(y_{il})}}$
- $Cov\left(y_{ij}, y_{il}\right)$
- $= Cov\left(\left(1-\mu_{ij}\right)\nu_{ij} + \mu_{ij}Z_i, (1-\mu_{il})\nu_{il} + \right.$

Back up

- 従って、$Corr\left(y_{ij}, y_{il}\right) = \frac{\pi(1-\pi)\rho}{\pi(1-\pi)} = \rho$