連続量とその離散化を考慮した回帰モデルのメタアナリシス

総合研究大学院大学　米岡大輔,　統計数理研究所　逸見昌之

1　背景

医学研究において, メタアナリシスは主に複数の臨床研究の結果を統合するための手法として用いられ, その結果は Evidence Based Medicine において最も質の高い根拠とされている. これまで, 多くのメタアナリシス研究は一つの効果指標の統合（死亡率等）に注目してきたが, 近年, 回帰モデルの係数統合にも注目が集まってきた. しかし, 回帰係数のメタアナリシスにおいては, 未解決の種々の問題点が挙げられる. その 1 つが, 研究間のカテゴリー変数の閾値の相違である. 同じ共変量（血圧等）であっても, 研究によっては連続量として処理されることもあれば, 異なる閾値でカテゴリー化されていることもあるのが現状である. したがって本研究では, この差異を考慮した回帰係数の多変量メタアナリシスを提案する.

2　提案手法

以下の様に, ある一つの変数のみがカテゴリー化された状態を考える. C_1 は $X_1 \sim N(\mu_X, \sigma_X^2)$ をある閾値 c_1 で 2 値にしたもので, 研究によってその閾値は異なる.

$$\text{(連続モデル)}\ \ Y = \alpha_0 + \alpha_1 X_1 + \boldsymbol{X}_2\boldsymbol{\alpha}_2 + \varepsilon,\ \ \varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$$
$$\text{(離散化モデル)}\ \ Y = \beta_0 + \beta_1 C_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \tau,\ \ \tau \sim N(\mu_\tau, \sigma_\tau^2)$$

従来のメタアナリシスでは，連続モデルのみを集めるか, もしくは X_1 を c_1 で 2 値化し, 同じ共変量にした上で多変量のメタアナリシスを行っていた. しかし, これは情報量の損失が起こっている. これを解決するために, 以下の 2 段階の統合方法を提案する.

- Step 1【sampling 分布の最尤推定】: 観測データ $(\hat{\mu}_X^i, \hat{\sigma}_X^{2i}, n_0^i, n_1^i, c_1^i)(i = 1, \ldots, N)$ から最尤法で未知パラメーター μ_X, σ_X^2 を推定する. (ただし, n_0^i, n_1^i は各研究で二値化された場合の各カテゴリーごとのサンプル数で, c_1^i は各研究ごとの閾値.)

- Step 2【脱落変数バイアスを考慮した GLS】: 漸近的なバイアスを導出するために, スコア関数の普遍性の条件を考える. つまり,

$$E\left[\{Y - (\beta_0^* + \beta_1^* C_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2^*)\}\begin{pmatrix} 1 \\ C_1 \\ \boldsymbol{X}_2^T \end{pmatrix}\right] = \mathbf{0}$$

  を満たす β_0^*, β_1^*, $\boldsymbol{\beta}_2^*$ を真の分布のパラメーターの関数 $\beta_j^* = f_j(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2, p_{X_1\boldsymbol{X}_2})(j = 0, 1, 2)$ で表現することを考える (ただし, $p_{X_1\boldsymbol{X}_2}$ は共変量の同時分布). 以上の脱落変数バイアスの式を用い, GLS によって $\alpha_0, \alpha_1, \boldsymbol{\alpha}_2$ を推定する.

$$\hat{\alpha}_{i0} = \alpha_0 + \epsilon_{i0},\ \ \hat{\alpha}_{i1} = \alpha_1 + \epsilon_{i1},\ \ \hat{\boldsymbol{\alpha}}_{i2} = \boldsymbol{\alpha}_2 + \boldsymbol{\epsilon}_{i2}\ \ \text{(連続の場合)}$$
$$\hat{\beta}_{i0} = f_0(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2, p_{X_1\boldsymbol{X}_2}) + \epsilon_{i0},\ \hat{\beta}_1 = f_1(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2, p_{X_1\boldsymbol{X}_2}) + \epsilon_{i1}, \hat{\boldsymbol{\beta}}_2 = f_2(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2, p_{X_1\boldsymbol{X}_2}) + \boldsymbol{\epsilon}_{i2}\ \ \text{(離散化の場合)}$$

  ただし, $(\epsilon_{10}, \epsilon_{11}, \boldsymbol{\epsilon}_{12}^T, \ldots \epsilon_{N0}, \epsilon_{N1}, \boldsymbol{\epsilon}_{N2}^T)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ で $\boldsymbol{\Sigma}$ は各研究の係数分散共分散行列を Block diagonal に並べたもの.