Expressing GMDe Coefficient Matrix $\mathbf{A}_y$ using Logistic Regression Coefficients

The Generalized Multivariate Difference Estimator (GMDe) is inherently a linear estimator, defined by the transformation:

$$\hat{\mathbf{y}}^{+} = \hat{\mathbf{y}}^{-} + \mathbf{A}_y \mathbf{r} \tag{1.1}$$

When the study variables in vector $\mathbf{y}$ are counts (or proportions derived from counts) and modeled linearly, this approach is analogous to a **Linear Probability Model (LPM)**.

If we instead assume the underlying data generating process follows a **Logistic Regression** model (which restricts estimates to the feasible $\{0,1\}$ range, unlike the LPM), the linear coefficient matrix $\mathbf{A}_y$ can be expressed as a linearization of the logistic coefficients.

1. The Relationship via Linearization

In GMDe, the matrix $\mathbf{A}_y$ represents the rate of change (slope) of the study variables with respect to the auxiliary residuals. In a regression context, this is the "Marginal Effect."

For a Logistic Regression model, the expected value p (proportion) is related to the auxiliary variables x via the logistic function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{1.2}$$

While the logistic coefficient β_1 represents the change in the **log-odds** for a unit change in x, the GMDe matrix $\mathbf{A}_y$ requires the change in the **raw probability/count** for a unit change in x.

We use a first-order Taylor series approximation (linearization) to express $\mathbf{A}_y$ in terms of β.

2. Derivation

The derivative of the logistic function with respect to x is:

$$\frac{\partial p}{\partial x} = \beta_1 \cdot p \cdot (1 - p) \tag{1.3}$$

Therefore, the linear coefficient A (which approximates this derivative about the mean) is:

$$A \approx \beta_{logistic} \times \mu_p (1 - \mu_p) \tag{1.4}$$

3. The Matrix Expression for GMDe

Let **B** be the matrix of logistic regression coefficients where B_{mj} links the j^{th} auxiliary variable to the m^{th} study variable.

Let $\hat{\mathbf{y}}$ be the vector of prior estimates (proportions or counts) for the M study variables.

The GMDe coefficient matrix $\mathbf{A}_y$ can be expressed as:

$$\mathbf{A}_y \approx \mathbf{B} \odot \mathbf{V}_{var} \tag{1.5}$$

where:

- $\odot$ denotes the Hadamard (element-wise) product.

$\mathbf{V}_{var}$ is a scaling matrix derived from the variance of the binomial distribution.

If **y** contains proportions:

$$\mathbf{A}_{y_{mj}} = \beta_{mj} \left[\hat{y}_m (1 - \hat{y}_m) \right] \tag{1.6}$$

If **y** contains total population counts $\left(N\hat{p} \right)$, the derivative must be scaled by the population size N:

$$\mathbf{A}_{y_{mj}} = \beta_{mj} \cdot N \cdot \left[\hat{p}_m (1 - \hat{p}_m) \right] \tag{1.7}$$

4. Computational Formula for Matrix **B**

Unlike the optimal linear coefficient matrix $\mathbf{A}_{opt}$, which has a closed-form solution based on population covariance matrices (*i.e.,* $\mathbf{A} \propto \Sigma_{yr} \Sigma_{rr}^{-1}$), the logistic coefficient matrix **B** typically requires an iterative solution such as **Maximum Likelihood Estimation (MLE)**.

The most common computational method is the **Iteratively Reweighted Least Squares (IRLS)** algorithm. Since the M study variables are typically modeled as independent logistic regressions conditional on the auxiliary variables, the matrix **B** is constructed row-by-row.

For the m^{th} study variable (corresponding to the m^{th} row of **B**, denoted $\boldsymbol{\beta}_m$), the computational formula at iteration k+1 is:

$$\boldsymbol{\beta}_m^{(k+1)} = \boldsymbol{\beta}_m^{(k)} + \left(\mathbf{X}^T \mathbf{W}_m^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^T \left(\mathbf{y}_m - \mathbf{p}_m^{(k)}\right) \tag{1.8}$$

where:

- **X** is the N×(J+1) design matrix of auxiliary variables (including the intercept).
- $\mathbf{y}_m$ is the N×1 vector of observed binary outcomes (or success counts) for the m^{th} study variable.
- $\mathbf{p}_m^{(k)}$ is the vector of predicted probabilities at iteration k, calculated using the logistic function and $\boldsymbol{\beta}_m^{(k)}$.
- $\mathbf{W}_m^{(k)}$ is a diagonal weight matrix where the i^{th} diagonal element is $p_{m,i}^{(k)}(1-p_{m,i}^{(k)})$.

The final M×J matrix **B** is the concatenation of the converged coefficient vectors for all M study variables (excluding the intercept if centering is handled separately, or including it if the auxiliary vector $\hat{\mathbf{r}}$ accounts for it):

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_M^T \end{bmatrix} \tag{1.9}$$

Summary

While GMDe uses a linear adjustment structure (Equation 1.1), specifying it with logistic assumptions implies that the "arbitrary constants" in matrix $\mathbf{A}_y$ are **variable**. They depend not just on the logistic slope β, but also on the current value of the estimate $\hat{y}$ itself:

$\mathbf{A}_y$ (logistic) = logistic coefficients × binomial variance factor