Estimation of contact matrices with a new count data model and surveys

1 Modeling framework

1.1 ZIBell regression with a random intercept

Suppose we have K clusters (groups) of patients. As in Hall (2000), let Y_{ij} denote the response variable observed for individual j in cluster i, with $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$. We assume that, conditionally on a random effect b_i specific to each cluster i, the response variable Y_{ij} follows a zero-inflated Bell (ZIBell) distribution, whose probability mass function (pmf) is given by:

$$P\left(Y_{ij} = y_{ij} \mid b_i\right) = \begin{cases} \pi_{ij} + (1 - \pi_{ij}) \exp\left(1 - e^{W(\mu_{ij})}\right) & \text{if } y_{ij} = 0, \\ (1 - \pi_{ij}) \exp\left(1 - e^{W(\mu_{ij})}\right) \frac{W(\mu_{ij})^{y_{ij}} B_{y_{ij}}}{y_{ij}!} & \text{if } y_{ij} > 0. \end{cases} \tag{1}$$

where we model μ_{ij} and π_{ij} using log-linear and logistic regression models:

$$\log(\mu_{ij}) = \beta^\top X_{ij} + \sigma b_i \tag{2}$$

and

$$\operatorname{logit}(\pi_{ij}) = \gamma^\top Z_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, n_i. \tag{3}$$

Here, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top$, and σ denotes the vector of covariates associated with the response Y_{ij}. Additionally, $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^\top$ and $Z_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijq})^\top$ are covariate vectors. We assume that $b_1, \ldots, b_K$ are independent standard normal random variables.

Let $\theta = (\beta^\top, \gamma^\top, \sigma)^\top$ be the $k \times 1$ vector of all model coefficients, where $k = p + q$. Given a random sample $y = (y_1, y_2, \ldots, y_n)^\top$, the log-likelihood function of the ZIBell model with a random intercept is given by:

$$\ell(\theta) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log \left[\int_{-\infty}^{\infty} P\left(Y_{ij} = y_{ij} \mid b_i\right) \Phi(b_i)\, db_i \right].$$

By substituting $P\left(Y_{ij}=y_{ij}\mid b_i\right)$ with its expression given in (6), we obtain:

$$\ell(\theta)=\sum_{i=1}^{K}\sum_{j=1}^{n_i}\log\left[\int_{-\infty}^{\infty}\left[1(y_{ij}=0)\left(\pi_{ij}+(1-\pi_{ij})\exp\left(1-e^{W(\mu_{ij})}\right)\right)\right.\right.$$
$$\left.\left.+1(y_{ij}>0)\left((1-\pi_{ij})\exp\left(1-e^{W(\mu_{ij})}\right)\frac{W(\mu_{ij})^{y_{ij}}B_{y_{ij}}}{y_{ij}!}\right)\right]\Phi(b_i)\,db_i\right], \tag{4}$$

where The density $\Phi(b_i)$ corresponds to the normal distribution of b_i, and the integration accounts for the variability of random effects between the groups i.

Thus, the final expression of the log-likelihood, after substituting the expressions for μ_{ij} and π_{ij}, is:

$$\ell(\theta)=\sum_{i=1}^{K}\sum_{j=1}^{n_i}\log\left[\int_{-\infty}^{\infty}\left[1(y_{ij}=0)\left(\frac{e^{\gamma^\top\mathbf{Z}_{ij}}}{1+e^{\gamma^\top\mathbf{Z}_{ij}}}+\frac{\exp\left(1-e^{W(\exp(\beta^\top\mathbf{X}_{ij}+\sigma b_i))}\right)}{1+e^{\gamma^\top\mathbf{Z}_{ij}}}\right)\right.\right.$$
$$\left.\left.+1(y_{ij}>0)\left(\frac{\exp\left(1-e^{W(\exp(\beta^\top\mathbf{X}_{ij}+\sigma b_i))}\right)}{1+e^{\gamma^\top\mathbf{Z}_{ij}}}\frac{W(\exp(\beta^\top\mathbf{X}_{ij}+\sigma b_i))^{y_{ij}}B_{y_{ij}}}{y_{ij}!}\right)\right]\Phi(b_i)\,db_i\right], \tag{5}$$

1.2 ZIBell regression with a random intercept

Suppose we have K clusters (groups) of patients. As in Hall (2000), let Y_{ij} denote the response variable observed for individual j in cluster i, with $i=1,\ldots,K$ and $j=1,\ldots,n_i$. We assume that, conditionally on a random effect b_i specific to each cluster i, the response variable Y_{ij} follows a zero-inflated Bell (ZIBell) distribution, whose probability mass function (pmf) is given by:

$$P\left(Y_{ij}=y_{ij}\mid b_i\right)=\begin{cases}\pi_{ij}+(1-\pi_{ij})\exp\left(1-e^{W(\mu_{ij})}\right) & \text{if } y_{ij}=0,\\ (1-\pi_{ij})\exp\left(1-e^{W(\mu_{ij})}\right)\frac{W(\mu_{ij})^{y_{ij}}B_{y_{ij}}}{y_{ij}!} & \text{if } y_{ij}>0.\end{cases} \tag{6}$$

where we model μ_{ij} and π_{ij} using log-linear and logistic regression models:

$$\log(\mu_{ij})=\beta^\top X_{ij}+\sigma b_i \tag{7}$$

and

$$\text{logit}(\pi_{ij}) = \gamma^\top Z_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, n_i. \tag{8}$$

Here, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top$, and σ denotes the vector of covariates associated with the response Y_{ij}. Additionally, $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^\top$ and $Z_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijq})^\top$ are covariate vectors. We assume that $b_1, \ldots, b_K$ are independent standard normal random variables.

The log-likelihood function of the ZIBell model with a random intercept is given by:

$$\ell(\theta) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log \left[\int_{-\infty}^{\infty} P\left(Y_{ij} = y_{ij} \mid b_i\right) \Phi(b_i)\, db_i \right]. \tag{9}$$

The integration in the log-likelihood expression (9) generally does not have a closed-form solution due to the presence of the term $\Phi(b_i)$, the normal density of the random effects. To make the estimation more tractable, we use a numerical approximation based on Gauss-Hermite quadrature. This method approximates the integral by a weighted sum of function evaluations at specific points, thereby reducing computational cost while maintaining high accuracy.

To approximate the integral in (9), we apply Gauss-Hermite quadrature, which approximates the integral by:

$$\int_{-\infty}^{\infty} P(Y_{ij} = y_{ij} \mid b_i)\Phi(b_i)db_i \approx \sum_{m=1}^{M} w_m P(Y_{ij} = y_{ij} \mid b_m), \tag{10}$$

where $b_m = \sqrt{2}x_m$ are the evaluation points of the random effect and w_m are the corresponding weights.

Thus, the log-likelihood function becomes:

$$\ell(\theta) \approx \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log \left[\sum_{m=1}^{M} w_m P(Y_{ij} = y_{ij} \mid b_m) \right]. \tag{11}$$

$$\begin{aligned} \ell(\theta) \approx \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log \Bigg[\sum_{m=1}^{M} w_m \Big(& 1(y_{ij} = 0) \left(\pi_{ij} + (1 - \pi_{ij}) \exp\left(1 - e^{W(\mu_{ij,m})}\right) \right) \\ & + 1(y_{ij} > 0) \left((1 - \pi_{ij}) \exp\left(1 - e^{W(\mu_{ij,m})}\right) \frac{W(\mu_{ij,m})^{y_{ij}} B_{y_{ij}}}{y_{ij}!} \right) \Big) \Bigg], \end{aligned} \tag{12}$$

$$\ell(\theta) \approx \sum_{i=1}^{K}\sum_{j=1}^{n_i} \log\left[\sum_{m=1}^{M} w_m \left(1(y_{ij}=0)\left(\frac{e^{\gamma^\top \mathbf{z}_{ij}}}{1+e^{\gamma^\top \mathbf{z}_{ij}}} + \frac{\exp\left(1-e^{W(\exp(\beta^\top \mathbf{X}_{ij}+\sigma b_m))}\right)}{1+e^{\gamma^\top \mathbf{z}_{ij}}}\right)\right.\right.$$
$$\left.\left. +1(y_{ij}>0)\left(\frac{\exp\left(1-e^{W(\exp(\beta^\top \mathbf{X}_{ij}+\sigma b_m))}\right)}{1+e^{\gamma^\top \mathbf{z}_{ij}}} \frac{W(\exp(\beta^\top \mathbf{X}ij+\sigma b_m))^{yij} B_{y_{ij}}}{y_{ij}!}\right)\right)\right]. \tag{13}$$

1.3 Minimum density power divergence estimator

To estimate the parameter vector $\theta = (\beta^\top, \gamma^\top, \sigma)^\top$ in the Zero-Inflated Bell (ZIBell) regression model with a random intercept, we propose using the Minimum Density Power Divergence Estimator (MDPDE). This estimator enhances robustness against outliers while maintaining high efficiency under correct model specifications. The presence of excess zeros and potential outliers in count data can significantly impact standard likelihood-based estimation techniques, making the MDPDE particularly relevant for ZIBell regression models.

Given a parametric family of densities $\{f_\theta\}$, indexed by $\theta \in \Theta$, and an observed sample $y = (y_1, \ldots, y_n)^\top$, the density power divergence (DPD) between the true density g and a candidate density f_θ is defined as:

$$d_\alpha(g, f_\theta) = \int \left[f_\theta^{1+\alpha}(y) - \left(1+\frac{1}{\alpha}\right) g(y) f_\theta^\alpha(y) + \frac{1}{\alpha} g^{1+\alpha}(y)\right] dy, \quad \alpha > 0. \tag{14}$$

For $\alpha = 0$, the divergence simplifies to the Kullback-Leibler divergence, recovering the standard maximum likelihood estimation (MLE) method.

Since our model incorporates random effects, we extend the MDPDE approach to the conditional setting, where the true conditional density of Y_{ij} given the random effect b_i is denoted by $g_\theta(y_{ij} \mid b_i)$. The MDPDE for θ is obtained by minimizing:

$$\hat{\theta}_\alpha = \arg\min_{\theta\in\Theta} H_{\alpha,n}(\theta), \tag{15}$$

where the objective function $H_{\alpha,n}(\theta)$ is given by:

$$H_{\alpha,n}(\theta) = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} \int_{-\infty}^{\infty} V_\alpha(\theta; Y_{ij} \mid b_i)\Phi(b_i) db_i. \tag{16}$$

The function $V_\alpha(\theta; Y_{ij} \mid b_i)$ is defined as:

$$V_\alpha(\theta; Y_{ij} \mid b_i) = \begin{cases} \int f_\theta^{1+\alpha}(y \mid b_i)dy - \left(1 + \frac{1}{\alpha}\right) f_\theta^{\alpha}(Y_{ij} \mid b_i), & \text{if } \alpha > 0, \\ -\log f_\theta(Y_{ij} \mid b_i), & \text{if } \alpha = 0. \end{cases} \tag{17}$$

By integrating over the distribution of the random effects b_i, we account for heterogeneity across clusters. The estimator $\hat{\theta}_\alpha$ is consistent and asymptotically normal for small values of α, allowing a trade-off between robustness ($\alpha > 0$) and efficiency ($\alpha \approx 0$).

The MDPDE is obtained by modifying the likelihood function with a power divergence term. Specifically, the modified log-likelihood function for each observation j in cluster i is:

$$\tilde{\ell}_{\alpha,ij}(\theta) = \begin{cases} m_{\alpha,ij,1}(\theta) + m_{\alpha,ij,2}(\theta) - \left(1 + \frac{1}{\alpha}\right) \left[\tilde{\ell}_{\alpha,ij,1}(\theta) I(Y_{ij} = 0) + \tilde{\ell}_{\alpha,ij,2}(\theta) I(Y_{ij} > 0)\right], & \alpha > 0, \\ -\log \tilde{\ell}_{1,ij,1}(\theta) I(Y_{ij} = 0) - \log \tilde{\ell}_{1,ij,2}(\theta) I(Y_{ij} > 0), & \alpha = 0. \end{cases} \tag{18}$$

where the terms $\tilde{\ell}_{\alpha,ij,1}(\theta)$ and $\tilde{\ell}_{\alpha,ij,2}(\theta)$ represent the likelihood contributions for zero and positive counts, respectively, and $m_{\alpha,ij,1}(\theta)$ and $m_{\alpha,ij,2}(\theta)$ are the divergence terms that are adjusted by the power α. The expressions for these terms are given by:

The MDPDE estimator $\hat{\theta}_{\alpha,n}$ is then obtained by minimizing the average modified likelihood:

$$\hat{\theta}_{\alpha,n} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \tilde{\ell}_{\alpha,ij}(\theta). \tag{19}$$