Mixed models in categorical data analysis

Ruitao Zhang    Stanek Ed

Mixed models, characterized by inclusion of fixed and random effects, are used in many contexts in public health. Fixed effects correspond to population parameters. Random effects are used to represent parameters for clusters of similar units (such as communities, work sites, schools or classrooms, families, or clinic patient groups), or subjects (where units correspond to repeated measures). This research examines the simplest setting of mixed models in categorical data analysis for a finite population.

Suppose we are interested in the proportion of days in a month that a subject has 30+ minutes of leisure time physical activity. Each day can be categorized as no/yes. This setting underlies a mixed model with a binary response, with response represented on the logit scale. Suppose now we have a two stage simple random sample of subjects and days, and we are interested in predicting the realized random effect for the $i^{th}$ selected subject.

A finite population is defined by a listing of $M$ days, indexed by $t = 1,...,M$ for each of $N$ subjects, indexed by $s = 1,...,N$. We define $y_{st}$ as the response for subject $s$ at day $t$. Let $y_{st} = 1$ represent the response ‘yes’, and $y_{st} = 0$ represent the response ‘no’. We parameterize response for subject $s$ on day $t$ as

$$y_{st} = \pi_s + \epsilon_{st},$$

where, $\pi_s = \frac{1}{M} \sum_{t=1}^{M} y_{st}$, and $\epsilon_{st} = y_{st} - \pi_s$ is the deviation of the response for subject $s$ on day $i$ from $\pi_s$. Let the average for subject $s$ be represented by $\overline{\pi_s}$. Suppose further that we define the odds of activity as $\frac{\pi_s}{1-\pi_s}$, and

$$\ln\left(\frac{\pi_s}{1-\pi_s}\right) = \mu_s.$$  

Now, let us assume that we want to re-parameterize the $\pi_s$ in terms of a population mean, plus an increment for the subject $s$. Thus, $\mu_s = \mu + \beta_s$, where $\mu = \frac{1}{N} \sum_{s=1}^{N} \beta_s, \beta_s = \mu_s - \mu$. From the $\mu_s$ defined above, we can get

$$\pi_s = \frac{e^{\mu_s}}{1 + e^{\mu_s}} = \frac{e^{\mu + \beta_s}}{1 + e^{\mu + \beta_s}} = \frac{e^{\mu} e^{\beta_s}}{1 + e^{\mu} e^{\beta_s}}.$$
We define \( \pi^* = \frac{e^\mu}{1 + e^\mu} \) and re-parameterize model as:

\[
y_{st} = \pi^* + (\pi_s - \pi^*) + \varepsilon_{st}.
\]

We can express \( \pi_s - \pi^* \) as:

\[
\pi_s - \pi^* = \frac{e^\mu e^{\beta_s} (1 + e^\mu) - e^\mu (1 + e^\mu e^{\beta_s})}{(1 + e^\mu)(1 + e^\mu e^{\beta_s})}
= \frac{e^\mu}{(1 + e^\mu)(1 + e^\mu e^{\beta_s})} (e^{\beta_s} - 1).
= \pi^* \frac{(e^{\beta_s} - 1)}{(1 + e^\mu e^{\beta_s})}.
\]

In general, we can summarize the model as:

\[
y_{st} = f(\mu) + g_s(\mu, \beta_s) + \varepsilon_{st}.
\]

Here, \( f(\mu) = \frac{e^\mu}{1 + e^\mu} \) is non-linear function of \( \mu \) and \( g_s(\mu, \beta_s) = \frac{e^\mu}{(1 + e^\mu)(1 + e^\mu e^{\beta_s})} (e^{\beta_s} - 1) \) is a non-linear function of \( \mu, \beta_s \). To estimate \( \mu \) and \( \beta_s \) in model (3), we would approximate the non-linear model by some methods of linearization such as Taylor’s expansion. More work will be developed here.

**Appendix 1**

Taylor’s expansion at \( x_0 \) for one variable function \( f(x) \):

\[
f(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}(x_0)}{i!} (x-x_0)^i \]

Taylor’s expansion at \( (x_0, y_0) \), for two variables function \( f(x, y) \):

\[
f(x, y) = \sum_{i=0}^{\infty} \frac{1}{i!} \{ (x-x_0) \frac{\partial}{\partial x}|_{x=x_0} + (y-y_0) \frac{\partial}{\partial y}|_{y=y_0} \} f(x, y) \} \]