Lectures 13-14 Categorical Data Analysis

Joshua Freeman, Luke Toomey, Liz Austin

Mixture distribution/Models

- Overview
- Beta-binomial model
- Negative binomial model
- Example

Mixture models

Mixture distribution: A probability distribution formed by a weighted combination of two or more distributions. For example:

$$Y_i \sim Binomial(n, \pi_i) for i = 1, ..., k$$
$$f_i(Y) = \left[\begin{array}{c} n \\ y \end{array} \right] \pi_i^y (1 - \pi_i)^{n-y}$$

Z is an equal mixture of all Y_i for which

$$f(Z) = \sum_{i=1} \frac{f_i(Z)}{K}$$

The marginal distribution for a whole population with subpopulations that follow specific distributions.

GLMM –>Mixed effects = both random and fixed effects. Not for mixture models.

A mixture model is a probabilistic model for representing the presence of subpopulations within the overall population, i.e. another way to account for overdispersion. Generally, data do not <u>*NOT*</u> provide identifiers of a subpopulation membership. Estimating groups and group membership. Subgroups are assured known and we want to get at estimates of characteristics for them.

GLMM as a mixture model: e.g.

$$Y_{ij} \left[\begin{array}{ll} 1 & if\ jth\ child\ to\ ith\ woman\ is\ M \\ 0 & if\ jth\ child\ to\ ith\ woman\ is\ F \end{array} \right]$$

$$Y_{ij} \sim Bernoulli(\pi_i)$$
$$\pi_i = logit^{-1}(\alpha_i)$$
$$\alpha_i \sim N(0, \sigma_x^2)$$

The population level distribution of boy/girl babies is a mixture of $Y_{ij}s$ (plural)

This is not an east close-form solution for the marginal population-level distribution of boys and girls. We can at least get at expectation and variance.

Tower Property/Conditional Expectation

==========

$$E(Y) = E[E(Y|X)]$$
$$Var(Y) = E(V(Y|X)) + V(E(Y|X))$$

For example:

$$E(Y) = E[E(Y|\alpha_i)]$$
$$E(Y) = E[logit^{-1}(\alpha_i)] = ?$$

This is not a closed form solution for this distribution.

Beta-Binomial Distribution

Conjugate mixture for binary outcome is a model where the marginal distribution has a closed form solution (c.f. "conjugate prior")

$$(a) Y|\pi \sim Binomial(n, \pi)$$
$$E(Y|\pi) = n\pi$$
$$V(Y|\pi) = n\pi(1-\pi)$$

$$(b) \pi \sim Beta(\alpha_1, \alpha_2)$$

$$\mu = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$
$$\theta = \frac{1}{\alpha_1 + \alpha_2}$$
$$E(\pi) = \mu$$
$$V(\pi) = \frac{\mu(1-\mu)\theta}{1-\theta}$$

$$E(Y) = E[E(Y|\pi)] = E[n\pi] = n\mu$$
$$Var(Y) = ...(n\mu(1-\mu))(1 + (n-1)[\frac{\theta}{1-\theta}])$$

As θ approaches 0, $Var(\pi)$ also approaches 0. Get a single expected value.

Not an exponential family model, but model with logit link.

$$logit(\mu_i) = \alpha + \beta^T X_i$$

Estimated using Newton-Raphson.

Poisson Models for Overdispersion

-Poisson GLM with $\hat{\phi}$ adjustment. -Poisson GLMM with normal, random intercepts. -Gamma Poisson mixture –>Negative Binomial, Poisson GLMM with non-Normal random intercept.

Poisson GLMM are more flexible alternatives than a Negative-Binomial model.

Negative-Binomial as a Gamma-Poisson mixture

$$(a)Y|\lambda \sim Poisson(\lambda) ==> E(Y|\lambda) = \lambda, Var(Y|\lambda) = \lambda$$
$$(b)\lambda \sim Gamma(K, \mu) ==> E(\lambda) = \mu, Var(\lambda) = \mu^2/K$$
$$\gamma = 1/K = "dispersion"$$
$$Y \sim NegativeBinomial(K, \mu)$$
$$E(Y) = E[E(Y|\lambda)] - E(\lambda) = \mu$$
$$V(Y) = E(V(Y|\lambda)) + V(E(Y|\lambda)) = E(\lambda) + V(\lambda) = \mu + \mu^2/K$$

As γ approaches 0 ==> Y $\sim$ Poisson(μ)

Compared to Poisson GLM with normal μ

1. Structural problem with the model
$$logE(Y_i|\mu_i) = X_{ij}^T\beta + \mu_i, \mu_i \sim N(0, \sigma^2)$$
2. Negative Binomial model with fixed γ is a GLMM with non-Normal random intercept if $exp\mu_i \sim$ Gamma (mean=1, Var=γ) $Y_ij \sim$ Poisson(λ_i), $log\lambda_i = \alpha_i + x_{ij}^T\beta, \alpha_i \sim N(0, \sigma^2) = \lambda_i \sim lognormal()$
3. Poisson GLMM
$$E(Y) = E[E(Y|\lambda)] = E(\lambda)$$
$$\lambda \sim logNormal = exp\frac{\sigma^2}{2}$$
$$Y_ij \sim Poisson(\lambda_i)$$
$$log\lambda_i = \alpha_i, \alpha_i \sim N(0, \sigma^2)$$

Lecture 14

Zero-Inflated Models (See countreg PDF on Google Drive)

$f_{zero}(y, Z, \gamma)$ The distribution follows the probability (π) of count data such that there is an excess number of zeroes clumping together and a Poisson distribution. What is the probability of following a distribution of Non-zero distribution vs. Zero counts?

For a graphic of the Zero-inflated model distribution, please see: https://www.researchgate.net/publication/224040413_Comparing_species_abundance_models (Figure 1)

and

https://support.sas.com/rnd/app/stat/examples/GENMODZIP/roots.htm (Figure 1)

There are two possible sources of zeroes: 1. f_{zero} 2. f_{count}

Read sections on Hurdle and Zero-inflated models.

Distribution between hurdle and zero-inflated models only source of zeroes is f_{zero}–>the zero point mass.

$f_{zero}(y, Z, \gamma)$–>Binary distribution describing whether y follows count distribution or binary distribution.

$f_{count}(y; x, \beta)$–>count distribution (e.g. Poisson, Negative Binomial)

$f_{zero-inflated}(y; x, Z, \beta, \gamma) = f_{zero}(0; Z, \gamma)$ * I$\{0\}(y) + f_{count}(y, x, \beta)[1 - f_{zero}(0; Z, \gamma)]$

Binomial GLM –> $f(0; Z, \gamma) = \pi = g^{-1}(Z^T y)$ $\mu_i = \pi_i 0 + (1 - \pi_i) exp X_i^T \beta$

Modelling expected mean of ith observation is a mixture of two components: Probability of zero count distribution and count distribution of $X^T \beta$

β, γ, θ can be estimated with MLE parameter (dispersion if Negative Binomial) The pscl package in R is useful for modelling zero-inflated data.