Epidemiology VHM 8120 Course
Winter 2020, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 10b: Mixed models for discrete data (VER Ch. 22.1-22.5.2)

Index Page

- Today
  - intro clustered discrete data (JS)
  - dealing with clustering / margins mixed models (HS)
- Tuesday
  - review clustered data exercises
    - continuous and binary data

Introduction – random effects logistic regression

- Animal diseases observed in several herds, then the probability "p_i" of the i^{th} animal being diseased is
  - $\text{logit}(p_i) = \beta_0 + \beta_1 * X_{1i} + ... + \beta_k * X_{ki} + u_{\text{herd(i)}}$
    - $u_{\text{herd(i)}} \sim Normal(0, \sigma^2_{herd})$
    - σ^2_{herd} = variability among herds (on the logit scale)
    - only difference from ordinary logistic regression is the herd random-effects term
  - alternative notation
    - Y_{ij} = pn status (0/1) of pig "i" in herd "j"
    - $\text{logit}(p_{ij}) = \beta_0 + \beta_1 * X_{1ij} + ... + \beta_k * X_{kij} + u_j$

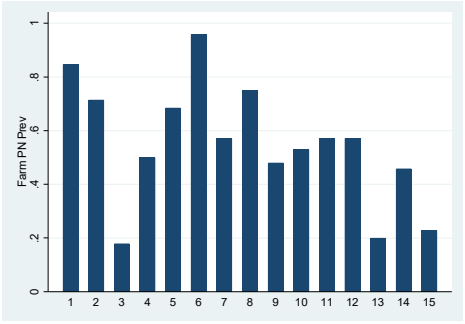
Example – pig pneumonia data - pig_adg.dta

- outcome = pn, predictor = ar_g1 (ar>1)

Farm (15) (range: 14-28)

ar_g1

Pig (341) -> outcome (pn)



Simple analysis

- 2 x 2 analysis

```
. cc pn ar_g1
                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |       109          77  |        186       0.5860
        Controls |        66          89  |        155       0.4258
-----------------+------------------------+------------------------
           Total |       175         166  |        341       0.5132
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.908894       |     1.21155    3.009556 (exact)
 Attr. frac. ex. |         .4761365       |    .1746111    .6677251 (exact)
 Attr. frac. pop |         .2790262       |
                 +-------------------------------------------------
                              chi2(1) =     8.69  Pr>chi2 = 0.0032
```

  - unconditional OR is 1.91 ($\beta = 0.647$)

Ordinary logistic regression

```
. logit pn ar_g1

Logistic regression                               Number of obs   =        341
                                                  LR chi2(1)      =       8.72
                                                  Prob > chi2     =     0.0031
Log likelihood = -230.59173                       Pseudo R2       =     0.0186

------------------------------------------------------------------------------
          pn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ar_g1 |   .6465241   .2203379     2.93   0.003     .2146697    1.078378
       _cons |  -.1448309   .1556373    -0.93   0.352    -.4498744    .1602125
------------------------------------------------------------------------------
```

  - unconditional OR = exp(0.647) = 1.91
  - exact same results

Random effects logistic regression

$$\text{logit}(p_i) = \beta_0 + \beta_1 * \text{ar\_g1} + u_{\text{herd(i)}} \qquad u_{\text{herd(i)}} \sim Normal(0, \sigma^2_{herd})$$

```
melogit pn ar_g1 || farm:

....output omitted....

Mixed-effects logistic regression               Number of obs     =        341
Group variable:            farm                 Number of groups  =         15

                                                Obs per group:
                                                              min =         14
                                                              avg =       22.7
                                                              max =         28

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(1)      =       2.86
Log likelihood = -213.51176                     Prob > chi2       =     0.0905
------------------------------------------------------------------------------
          pn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ar_g1 |   .4369322   .2581451     1.69   0.091    -.0690229    .9428872
       _cons |   .0196548   .3009262     0.07   0.948    -.5701497    .6094593
-------------+----------------------------------------------------------------
farm         |
   var(_cons)|   .8774246   .4325507                      .333877    2.305861
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 34.16       Prob >= chibar2 = 0.0000
```

Cluster-specific vs population-averaged interpretation

Example – breed and Ca supplementation (tx) on milk fever (MF)

- Tx= Ca supplementation
  - cluster-specific
    - effect of Ca suppl. to a cow in a specific herd
  - population-averaged
    - effect of Ca suppl. across all herds
    - difference in risk of MF across all herds between tx and non-tx cows
- Breed
  - cluster specific
    - herd with 2 breeds (Holstein and Jersey)
      - difference in risk of MF between 2 breeds in that herd
    - herd with 1 breed
      - no meaningful interpretation (unless you replace all cows with different breed)
  - population-averaged
    - difference in risk of MF between 2 breeds across all herds

Pig dataset

- Cluster-specific
  - β = effect in a individual if changed value of X
    - within a cluster
  - mixed (random-effects) models
    - β_1 = 0.437 (SE=0.258, P=0.091)
      - reduced effect, borderline significant
    - effect of ar_g1 when comparing two pigs in the same farm
- Population-averaged
  - β = average effect of X in population
  - estimates closer to null
  - overall comparison of pigs with and without ar_g1 from any herd

$$\beta_{PA} \approx \frac{\beta_{CS}}{\sqrt{(1+0.346*\sigma^2_{herd})}} = \frac{0.437}{\sqrt{(1+0.346*0.877)}} = 0.383$$

- Variance parameter
  - σ_h^2 = 0.877 (SE = 0.432)
    - substantial variation between farms in logit of pn
    - level of logit(pn) in pigs with ar_g1=0
      - 95% farms expected within:
        - 0.02 +/- 1.96*.94 = -1.82 to 1.86
    - probability of pneumonia in ar_g1=0 farms
      - 95% farms expected within: 14% and 87%
  - also can be interpreted as cluster median odds ratio (MOR)
    - more details in text (VER example 22.2)
  - ICC
    - variance estimates at each level
    - lowest level variance is on binomial scale while higher level variances are on logistic scale
    - several methods to approx. lowest level variance
      - latent response variable - $\sigma_i^2 = \pi^2/3 = 3.29$
    - ICC - 2 level pig example
      - $\sigma_h^2 = 0.88$; total variance = 3.29+0.88 = 4.17
      - ICC = 0.88/4.17 = 0.21

Random effects Poisson regression

- Overdispersion in count data is common
  - options
    - add overdispersion parameter to the model
    - negative binomial model
      - appropriate if overdispersion not due to clustering
    - random effects models (Poisson/Neg. Bin)
- Poisson model with random effects
  - $\log(\lambda_i) = \beta_0 + \beta_1 * X_{1i} + ... + \beta_k * X_{ki} + u_{\text{herd(i)}}$
  - $Y_i \sim Poisson(\lambda_i * par_i)$; $u_{\text{herd(i)}} \sim \text{Normal}(0, \sigma^2_{herd})$

Example: tuberculosis data - tb_real.dta

- 30 herds
- 134 groups of animals defined by:
  - type: dairy (15); beef (58); cervid (52); other (9)
  - age: 0-12 mo (37); 12-24 mo (38), > 24 mo (59)
  - sex: female (74), male (60)
  - outcome: # TB reactors
  - exposure: # animal-months at risk

- Fixed effects and random effects models

Variable	*Poisson*		*Negative Binomial*		*Poisson rand. (Normal)*		*NB rand. (Normal)*	
	β	SE	β	SE	β	SE	β	SE
Type								
beef	0.442	(0.236)	0.605	(0.675)	-0.394	(0.333)	-0.394	(0.333)
cervid	1.066	(0.233)	0.666	(0.684)	-0.238	(0.487)	-0.238	(0.487)
other	0.438	(0.615)	0.800	(1.119)	-0.104	(0.800)	-0.104	(0.800)
Gender								
male	-0.362	(0.195)	-0.057	(0.405)	-0.339	(0.208)	-0.339	(0.208)
Age								
12-24	2.673	(0.722)	2.253	(0.903)	2.717	(0.747)	2.717	(0.747)
> 24	2.601	(0.714)	2.481	(0.882)	2.467	(0.726)	2.467	(0.726)
Constant	-11.69	(0.740)	-11.18	(1.061)	-11.05	-0.83	-11.05	-0.83
α	-		1.740		-		0	
σ^2_{herd}	-		-		1.698		1.698	
LL	-238.7		-157.7		-143.6		-143.6	
Dispersion	8.71		2.95					

  - only age was stat. sig., estimates for age reasonably consistent
  - Poisson with random effects appears to fit better
  - negative binomial models with random effects
    - same as Poisson with random effects
    - alpha = 0, overdispersion due to clustering

Interpretation of the parameters

- Fixed effect coefficients
  - no distinction between population averaged and subject specific interpretation
  - $IRR_{>24\ mo} = \exp(2.467) = 11.78$
    - effect when comparing two groups from the same herd
    - effect when comparing two groups from any herd
- Variance parameters ($\sigma_{herd} = 1.299$)
  - approximate range in estimate of incidence rate across herds (for baseline group: dairy, female, 0-12 mo) was:
    - 95% farms expected within:
      - log(I) = -11.055 +/- 1.96*1.299 = -13.601 to -8.509
      - I = 1.24 to 202 per 1,000,000 animal-months at risk
- ICC
  - no simple way to compute it
  - variance depends on the mean
  - many different ICCs can be obtained
    - see more in VER 22.3.2

Estimation procedures for discrete data

- Not straightforward, various approaches
- ML estimation is becoming the "norm"
  - numerically challenging for large datasets
- Approximate methods
  - quasi-likelihood
    - uses iterative weighted least squares
    - can get cluster-specific (PQL) or population-averaged estimates (MQL) (eg. MLwiN software)
  - Laplace approximation
  - some disadvantages
    - no likelihood based statistics
    - biases estimates (particularly of variance estimates - biased towards the null)

Approaches for dealing with clustered data

Method	Adjust β	Adjust SE	>1 level	ICC	Comments
Mixed Ef. model	Y	Y	Y	Y	
Fixed Ef. model	Y	Y	N	N	no cluster-level predictors
Stratified	Y	Y	N	N	binary data
Dispersion	N	Y	N	N	no-within cluster predictors and not for continuous data
Robust SE	N	Y	N	N	adjust for other model violations (continuous data)
General Est. Eq (GEE)	Y	Y	(N)	(Y)	population average parameters (discrete data

Stata code

```
version 16
set more off
cd "c:\vhm812-data"

capture log close
log using l10b-intro_cluster_discrete.txt, text replace

*Random effects binary data
*Pig respiratory disease data
use pig_adg.dta, clear
* generate dichotomous atrophic rhinits variable
egen ar_g1=cut(ar), at(0, 1.5, 99) icodes

* 2x2 table analysis
cc pn ar_g1

* ordinary logistic regression
logit pn i.ar_g1
logit pn i.ar_g1,or

* random effect logistic regression
melogit pn ar_g1 || farm:
melogit pn i.ar_g1 || farm:, or

*ICC // approximation to real ICC by pi^2/3 = 3.29
estat icc

* PA logistic regression
xtgee pn ar_g1, fam(binomial) link(logit) i(farm) robust

* Random effects models for count data
* open the tb_real dataset
use tb_real, clear
rename par tar
* Poisson model with no random effects
glm reactors i.type i.sex i.age, exp(tar) link(log) fam(poisson)
estimates store pois

* negative binomial model no random effects
glm reactors i.type i.sex i.age, exp(tar) link(log) fam(nbin ml)
estimates store nb
* Pearson dispersion parameter still large (2.95)

* Poisson model with normal distributed random effects
mepoisson reactors i.type i.sex i.age, exp(tar) || farm_id:
mepoisson reactors i.type i.sex i.age, exp(tar) irr || farm_id:
**same with meglm
meglm reactors i.type i.sex i.age, exp(tar) || farm_id:, fam(poisson)

* Negative binomial with Normal random effects
meglm reactors i.type i.sex i.age, exp(tar) || farm_id:, fam(nbin)
* the overdispersion lnalpha is non signfincant which suggests that the
* overdispersion is due to clustering since this model is exactly
* the same as the Poisson with random effects

estimates store pois_norm
estimates table pois nb pois_norm, se(%4.3f) b(%4.3f)
estimates stats pois nb pois_norm
```