Epidemiology VHM 802/812 Course
Winter 2016, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 11b: Mixed models for discrete data (VER Ch. 22)

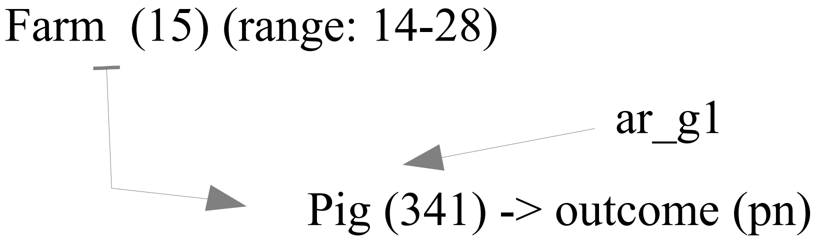
Index Page

Random effects logistic regression

- animal diseases observed in several herds, then the probability "p_i" of the i^{th} animal being diseased is
  - $\mathrm{logit}(p_i) = \beta_0 + \beta_1 * X_{1\mathrm{i}} + \ldots + \beta_k * X_{ki} + u_{\mathrm{herd(i)}}$
    - $u_{\mathrm{herd(i)}} \sim Normal(0, \sigma^2_{herd})$
    - σ^2_{herd} = variability among herds (on the logit scale)
    - only difference from ordinary logistic regression is the herd random-effects term
  - Note: probability of the predicted outcome (p_i) is conditional on the random effect ($u_{herd(i)}$)
    - predicted value for any individual is the predicted value given that it is in herd "i"

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)

- alternative notation
  - Y_{ij} = pn status (0/1) of pig "i" in herd "j"

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 (β = 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

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

```
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
  - <u>cluster-specific</u>
    - effect of Ca suppl. to a cow in a specific herd
  - <u>population-averaged</u>
    - effect of Ca suppl. across all herds
    - difference in risk of MF across all herds between tx and non-tx cows
- Breed
  - <u>cluster specific</u>
    - 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)
  - <u>population-averaged</u>
    - difference in risk of MF between 2 breeds across all herds

Pig dataset

- coefficient parameters
  - 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
      - most farms between 0.02 +/- 1.96*.94 = -1.82 to 1.86
      - prev. in ar_g1 neg. farms between 14% and 87%
  - also can be interpreted as cluster median odds ratio (MOR)
    - more details in text

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
  - $\log(\lambda_i) = \beta_0 + \beta_1 * X_{1i} + \ldots + \beta_k * X_{ki} + u_{\text{herd(i)}}$
  - $Y_i \sim Poisson(\lambda_i * par_i)$; $u_{\text{herd(i)}} \sim \text{Normal}(0, \sigma_{herd}^2)$

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-days at risk
- fixed effects and random effects models

Variable	***Poisson***		***Negative Binomial***		***Poisson rand. (Normal)***	
	β	SE	β	SE	β	SE
Type						
beef	0.442	(0.236)	0.605	(0.675)	-0.394	(0.333)
cervid	1.066	(0.233)	0.666	(0.684)	-0.238	(0.487)
other	0.438	(0.615)	0.800	(1.119)	-0.104	(0.800)
Gender						
male	-0.362	(0.195)	-0.057	(0.405)	-0.339	(0.208)
Age						
12-24	2.673	(0.722)	2.253	(0.903)	2.717	(0.747)
> 24	2.601	(0.714)	2.481	(0.882)	2.467	(0.726)
Constant	-11.690	(0.740)	-11.181	(1.061)	-11.05	-0.83
α	-		1.740		-	
σ^2_{herd}	-		-		1.698	
LL	-238.7		-157.7		-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
- also Negative binomial models with random effects
    - see VER22.4.3

Interpretation of the parameters

- fixed effect coefficients
  - no distinction between population averaged and subject specific interpretation (except for constant – rarely of interest)
  - CS -> $IRR_{>24\ mo}$ = exp(2.467) = 11.78
    - effect when comparing two groups from the same herd
  - PA -> $IRR_{>24mo}$ = exp(2.443) = 11.51
    - effect when comparing two groups from any herd
    - fitted using xtgee
- variance parameters (σ_{herd} = 1.299)
  - approximate range in estimate of incidence rate across herds (for baseline group: dairy, female, 0-12 mo) was:
    - -11.055 +/- 1.96*1.299 = -13.601 to -8.509
    - 1.2 to 202 per 1,000,000 animal-days at risk

Estimation procedures

- not straightforward, various approaches
- ML estimation is becoming the “norm”
  - numerically difficult for large datasets
- approximate methods
  - quasi-likelihood
    - uses iterative weighted least squares
    - can get cluster-specific (PQL) or population-averaged estimates (MQL)
    - eg. MLwiN
  - 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		>1 level	ICC	Comments
	β	SE			
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

```
* Mixed models for discrete data
* do-file for lecture 11b of VHM 802/812, Winter 2016
version 14
set more off
cd "c:\vhm812-data"

capture log close
log using l11b_glmm_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 ar_g1
predict pa_noclus

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

* prediction random effects - not in the notes
predict u_farm, reffects
scatter u_farm farm, xlabel(1(1)15)

*comparing predicted prob - CS vs PA - not in the notes
* cluster specific - prob. for a pig from a specific farm
predict re_cs, mu
* population averaged
xtgee pn ar_g1, fam(binomial) link(logit) i(farm) robust
predict pa, mu /*mean prob from any farm */

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

* negative binomial model no random effects
glm reactors i.type i.sex i.age, exp(par) 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(par) || farm_id:
mepoisson reactors i.type i.sex i.age, exp(par) irr || farm_id:

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

**PA Poisson model
xtgee reactors i.type i.sex i.age, exp(par) i(farm_id) family(poisson) link(log)
robust
```