

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

Index	Page
Random effects logistic regression.....	2
Example – pig pneumonia data - pig_adg.dta	2
Interpretation of parameters.....	5
Cluster-specific vs population-averaged interpretation	5
Random effects Poisson regression.....	8
Example: tuberculosis data - tb_real.dta	8
Fixed effects and random effects models.....	9
Interpretation of the parameters.....	10
Estimation procedures.....	11
Approaches for clustered data.....	11
Stata code.....	12

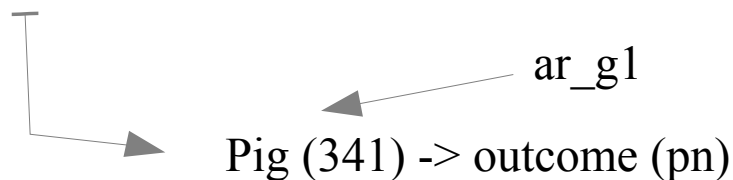
Random effects logistic regression

- animal diseases observed in several herds, then the probability “ p_i ” of the i^{th} animal being diseased is
 - ★ $\text{logit}(p_i) = \beta_0 + \beta_1 * X_{1i} + \dots + \beta_k * X_{ki} + u_{\text{herd}(i)}$
 - $u_{\text{herd}(i)} \sim \text{Normal}(0, \sigma_{\text{herd}}^2)$
 - σ_{herd}^2 = 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 ($u_{\text{herd}(i)}$)
 - 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)



- alternative notation

★ Y_{ij} = pn status (0/1) of pig “ i ” in herd “ j ”

Simple analysis

● 2 x 2 analysis

```
. cc pn ar_g1
```

	Exposed	Unexposed	Total	Proportion 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

$$\text{logit}(p_i) = \beta_0 + \beta_1 * \text{ar_g1} + u_{\text{herd}(i)} \quad u_{\text{herd}(i)} \sim \text{Normal}(0, \sigma_{\text{herd}}^2)$$

meqrlogit 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 points = 7	Wald chi2(1)	=	2.86
Log likelihood = -213.5118	Prob > chi2	=	0.0905

pn	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
ar_g1	.4369353	.2581461	1.69	0.091	-.0690217	.9428923
_cons	.0196485	.3009417	0.07	0.948	-.5701864	.6094834

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
farm: Identity				
var(_cons)	.8773258	.4324896	.3338486	2.305537

LR test vs. logistic regression: chibar2(01) = 34.16 Prob>=chibar2 = 0.0000

Interpretation of parameters

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

- coefficient parameters

- ★ cluster-specific

- β = effect in a individual if changed value of X
 - ◆ within a cluster
- mixed (random-effects) models
 - ◆ $\beta_1 = 0.437$ (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

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

- variance parameter
 - ★ $\sigma_h^2 = 0.877$ (SE = 0.432)
 - ➔ substantial variation between farms in logit of pn
 - ➔ level of logit(pn) in pigs with ar_g1=0
 - ◆ most farms between $0.02 \pm 1.96 \cdot .94 = -1.82$ 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
 - Poisson models with random effects
 - negative binomial model with random effects
- Poisson model with random effects
 - ★ $\log(\lambda_i) = \beta_0 + \beta_1 * X_{1i} + \dots + \beta_k * X_{ki} + u_{\text{herd}(i)}$
 - $Y_i \sim \text{Poisson}(\lambda_i * \text{par}_i)$
 - $u_{\text{herd}(i)} \sim \text{Normal}(0, \sigma_{\text{herd}}^2)$

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

<i>Variable</i>	<i>Poisson</i>		<i>Negative Binomial</i>		<i>Poisson rand. (Normal)</i>	
	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)	-11.055	(0.830)
α	-		1.740		-	
σ_{herd}	-		-		1.299	
LL	-238.7		-157.7		-143.7	
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 \rightarrow $IRR_{>24\text{ mo}} = \exp(2.467) = 11.78$
 - \rightarrow effect when comparing two groups from the same herd
 - ★ PA \rightarrow $IRR_{>24\text{ mo}} = \exp(2.443) = 11.51$
 - \rightarrow effect when comparing two groups from any herd
 - \rightarrow fitted using xtgee
- variance parameters ($\sigma_{herd} = 1.299$)
 - ★ approximate range in estimate of incidence rate across herds (for baseline group: dairy, female, 0-12 mo) was:
 - $\rightarrow -11.055 \pm 1.96 * 1.299 = -13.601$ to -8.509
 - \rightarrow 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 clustered data

Method	Adjust β	SE	>1 level	ICC	Comments
Mixed models	Y	Y	Y	Y	
FE	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 2015
version 13
set more off
cd "c:\vhm812-data"

capture log close
log using 11b-cluster_discrete.txt, text replace
*Random effects binary data
*Pig respiratory disease data
use pig_adg.dta, clear
* generate dichotomous atrophic rhinitis 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
meqrlogit pn ar_g1 || farm:
meqrlogit pn ar_g1 || farm:, or

**Predictions - not for the notes
* prob. for a pig from a specific farm
predict re_cs, mu

* prediction random effects
predict u_farm, reffects relevel(farm)
scatter u_farm farm, xlabel(1(1)15)

* PA logistic regression
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
gen logpar=log(par)
* Poisson model with no random effects
glm reactors i.type i.sex i.age, off(logpar) link(log) fam(poisson)
estimates store pois
* negative binomial model no random effects
glm reactors i.type i.sex i.age, off(logpar) link(log) fam(nbin ml)
estimates store nb
* Pearson dispersion parameter still large (2.95)

* Poisson model with normal distributed random effects
meqrpoisson reactors i.type i.sex i.age, off(logpar) || farm_id:
meqrpoisson reactors i.type i.sex i.age, off(logpar) 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, off(logpar) i(farm_id) family(poisson) link(log)
robust
```