Epidemiology VHM 8120 Course
Winter 2019, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture L10b: Analysis of clustered data (VER Ch. 20)

Index Page

Today

- Small exercise QBA 5.1.a

Wednesday

- Quiz QBA II

Clusters

- observations that share some feature(s) in common
  - not considered by the predictors of the model
- derived from data structure
- observations within a cluster “more alike” (usually)
  - due to common features
- observations within a cluster “less alike” (occasionally)
  - competition for feed

Sources of clustering

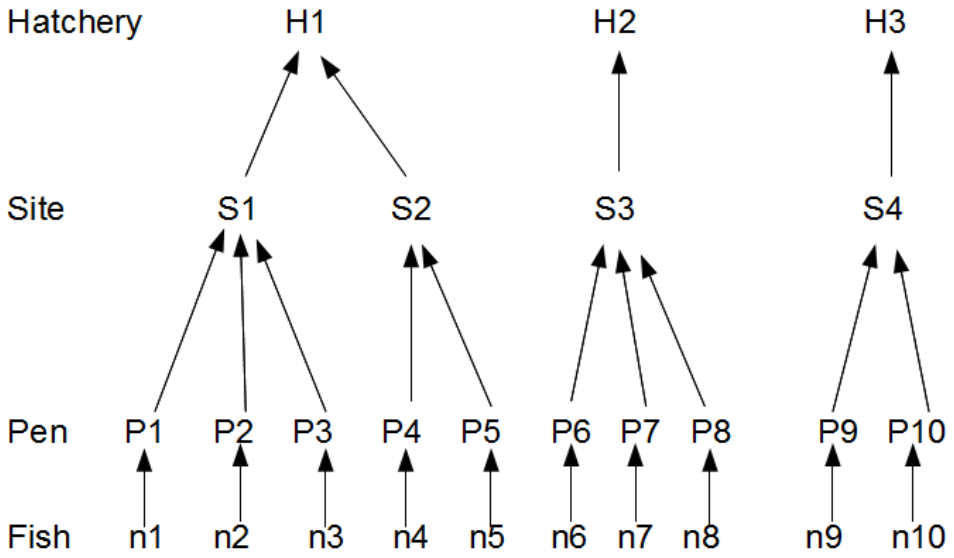
- common environment
  - eg. fish in a pen, cows in a herd, child in a school
  - same correlation among all pairs
    - Daisy and Nelly has same correlation as Daisy and Jessie from the same farm
  - multiple levels
    - region -> herd -> cow-> lactation
    - hierarchical or multilevel structure

- spatial clustering
  - dependence depends on distance among units
  - topographical features??
- repeated measurements (temporal clustering)
  - dependence depends on separation in time between observations
    - milk production more highly correlated with preceding day's production than production from 1 month ago
  - data from 3 cows - 1st 6 monthly tests of the lactation
    - dim = days in milk at time of test
    - milk = daily milk production (energy corrected)

			Test Number					
Herd	Cow		1	2	3	4	5	6
1	5	dim	11	39	67	102	130	165
		milk	25.12	19.92	19.13	20.38	18.21	14.64
1	12	dim	18	46	74	109	137	172
		milk	18.13	21.28	16.64	16.4	15.63	10.37
1	14	dim	23	58	86	121	149	177
		milk	18.84	18.81	17.1	13.47	11.29	10.46

Hierarchical vs cross-classified structures

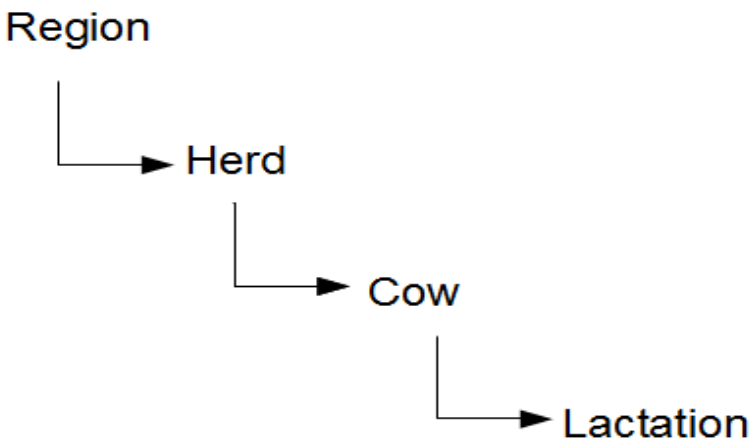
- (a) Hierarchical structure
  - requires that every site receives fish from one hatchery
  - requires that every pen is located within one site



- (b) Cross-classified structure
  - different pens at the same site receive fish from different hatcheries

Clustering of outcomes vs predictors

- clustering of outcome
  - violates standard assumption of independence
    - ordinary linear or logistic models are invalid
  - identify levels with greatest variation
    - potential room for improvement
  - outcomes usually at lowest level
- clustering of the predictors
  - predictors at various levels
  - sample size?

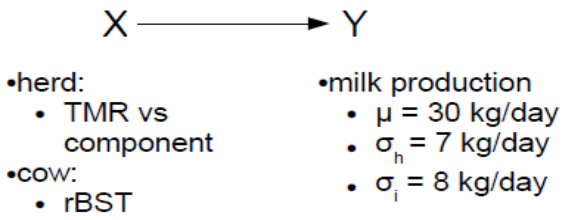


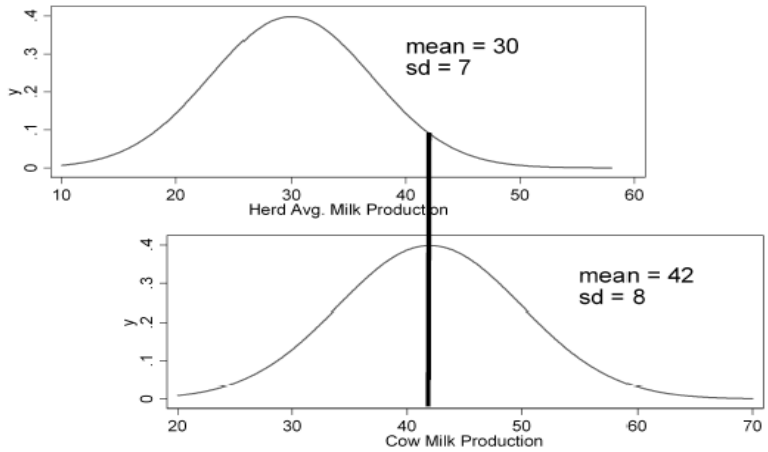
Effects of clustering

- if you ignore clustering – general
  - SE of parameter estimates usually too small
    - much too small for group levels predictors
    - may be too large for individual level predictors
  - assign unreasonable large weights to large groups
  - parameter estimates “asymptotically” unbiased
    - limited sample size estimates may be biased (discrete)
  - discrete data models (eg. logistic regression)
    - estimates are “marginal estimates” instead of “cluster specific” estimates (described later)
    - eg. often get different parameter estimates if do/don't control for clustering

Effect of clustering – continuous data (Example 20.2)

- data structure
  - 100 herds
    - 50 small herds (avg. 50 cows)
    - 50 large herds (avg. 200 cows)
  - outcome = milk production
    - varies between herds
    - herd mean = 30 kg/day, SD = 7 kg/day
    - cow level SD = 8 kg/day
  - predictor = X (herd or cow level factor)
    - true effect = +5 kg/day





- Scenario 1- X herd level variable

  - ignoring clustering

```
. reg milk X

Number of obs = 11626 - Root MSE = 10.733
  ...
------------------------------------------------------------------------------
       milk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
          X |    3.55661    .199534    17.82   0.000      3.16549     3.94773
      _cons |    30.0215   .1457715   205.95   0.000     29.73576    30.30723
------------------------------------------------------------------------------
```

  - accounting for cluster

```
. mixed milk X || herd: , reml stddev
Mixed-effects REML regression                   Number of obs      =     11626
Group variable: herd                            Number of groups   =       100

                                                Obs per group: min =        20
                                                               avg =     116.3
                                                               max =       311


                                                Wald chi2(1)       =      6.44
Log restricted-likelihood = -40902.479          Prob > chi2        =    0.0112

------------------------------------------------------------------------------
       milk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
          X |   3.796004   1.495943     2.54   0.011      .864009    6.727999
      _cons |   31.13696   1.058717    29.41   0.000     29.06191    33.21201
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
herd: Identity               |
                  sd(_cons) |   7.410465   .5396842      6.424728    8.547442
-----------------------------+------------------------------------------------
               sd(Residual) |   8.012545   .0527739      7.909774     8.11665
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  6374.40 Prob >= chibar2 = 0.0000
```

  - data collapsed to herd level

```
. collapse (mean) milk X, by(herd)

. reg milk X
.....
 Number of obs =     100 - Root MSE = 7.48

------------------------------------------------------------------------------
       milk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
          X |   3.778772   1.497421     2.52   0.013     .8071885    6.750356
      _cons |   31.16586   1.058837    29.43   0.000     29.06463    33.26708
------------------------------------------------------------------------------
```

- Scenario 2: X is a cow level variable
  - X has a prevalence of 0.5 in all herds (eg. clinical trial)
  - ignoring clustering

```
. reg milk X

      Source |       SS       df       MS              Number of obs =   11626
-------------+------------------------------           F(  1, 11624) =  624.90
       Model |  72138.7619     1  72138.7619           Prob > F      =  0.0000
    Residual |  1341880.62 11624  115.440522           R-squared     =  0.0510
-------------+------------------------------           Adj R-squared =  0.0509
       Total |  1414019.39 11625  121.636076           Root MSE      =  10.744

------------------------------------------------------------------------------
        milk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   4.982006   .1992962    25.00   0.000     4.591352     5.37266
       _cons |   29.25664   .1412627   207.11   0.000     28.97974    29.53354
------------------------------------------------------------------------------
```

- accounting for clustering

```
. mixed milk X || herd:, reml stddev

Mixed-effects REML regression                   Number of obs      =     11626
Group variable: herd                            Number of groups   =       100

                                                Obs per group: min =        20
                                                               avg =     116.3
                                                               max =       311


                                                Wald chi2(1)       =   1108.56
Log restricted-likelihood = -40947.175          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        milk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   4.968194   .1492174    33.30   0.000     4.675733    5.260655
       _cons |   30.64647   .7281276    42.09   0.000     29.21936    32.07357
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
herd: Identity               |
                  sd(_cons) |   7.170209   .5201795      6.219843    8.265787
-----------------------------+------------------------------------------------
               sd(Residual) |   8.044296   .0529852      7.941114    8.148818
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  6310.00 Prob >= chibar2 = 0.0000
```

- summary herd and cow level analyses

		Linear regression		Mixed model		Herd average	
Dataset	Parameter	β	SE	β	SE	β	SE
X herd level	X	3.56	0.20	3.80	1.50	3.78	1.50
	intercept	30.02	0.15	31.14	1.06	31.17	1.06
X cow level	X	4.98	0.20	4.97	0.15		
	intercept	29.26	0.14	30.65	0.73		

- scenario 1 - X is a herd level variable
  - ignoring clustering produces SE much lower that they should be
  - herd averages very similar estimates as mixed model
- scenario 2 - X is a cow level variable
  - mixed model estimate close to the true and more precise
  - intercept underestimate SE if clustering is ignored

Effect of clustering – discrete data (Example 20.3)

- Effect of X on probability of disease (outcome)
  - prevalence of X = 0.50
    - herd level: 50% herds with X=1(0) (all cows within each herd with the same level of the predictor)
    - cow level: 50% cows X=1(0) within herd
  - OR_x = 2 (or ln(2) = 0.693 on logistic scale)
  - disease non-exposed p = 0.2 (or ln(0.2/0.8) = -1.4 on logistic scale)
  - herd level effects varied (on logistic scale) with var = 1
- scenario 1 – X is a herd level variable (Ex. 20.3)
  - ignoring clustering

```
. logit Y X

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .5287317   .0423191    12.49   0.000     .4457877    .6116757
       _cons |  -1.241768   .0325699   -38.13   0.000    -1.305604   -1.177932
------------------------------------------------------------------------------
```

  - accounting for clustering

```
. melogit Y X || herd:

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .6199967   .2037578     3.04   0.002     .2206389    1.019355
       _cons |  -1.305448   .1454551    -8.97   0.000    -1.590534   -1.020361
-------------+----------------------------------------------------------------
herd         |
   var(_cons)|   .9417563   .1493109                      .6902154    1.284968
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 1499.38     Prob >= chibar2 = 0.0000
```

- summary herd and cow level analyses

Dataset	Parameter	Logistic regression β	Logistic regression SE	Logistic mixed model β	Logistic mixed model SE
X herd level	X	0.53	0.04	0.62	0.20
	intercept	-1.24	0.03	-1.31	0.15
X cow level	X	0.59	0.04	0.70	0.05
	intercept	-1.25	0.03	-1.36	0.11

- in general, ignoring clustering underestimates SE
- can't compare estimates between log. reg. and mixed model (next lecture)

Variance inflation as a result of clustering

- group level predictor (outcome of interest is mean value for groups)
- variance of group mean affected by:
  - ICC (intra-class correlation coefficient - ρ)
    - measure of similarity between observations within a cluster
  - group size
- Variance of group means is:

$$Var(\bar{y}) = \frac{\sigma^2}{m} * VIF$$

  - where
    - m = group size ;
    - σ^2 = within-group var(y_i), and
    - $VIF = [1 + (m-1) * \rho]$ = Variance Inflation Factor

- ICC (ρ) = 0 -> $Var(\bar{y}) = \frac{\sigma^2}{m}$
  - independent observations – all observations
- ICC (ρ) = 1 -> $Var(\bar{y}) = \frac{\sigma^2}{m} * m$
  - complete correlation – one observation

ICC	m	VIF	Comments
0	20	1	no within group correlation = no VIF added to Var(y)
1	20	20	complete within group correlation => VIF = m
0.1	6	1.5	low ICC and moderate group size had similar impact as
0.5	2	1.5	high ICC and small group
0.1	101	11	very large group size, even with low ICC has a very big impact

- VIF can be used to adjust sample size estimates for clustering (VER Ch. 2.11.6)
  - $n` = n * [1 + (m - 1) * \rho]$
  - effective sample size will be = m/VIF
    - 20 cows per herd and high ICC (e.g. 1)
    - effective sample size = 20/20 =1

Stata code

```
* do-file for lecture 10b of VHM 8120
* Introduction to clustered data

version 15
set more off
cd "c:\vhm812-data"

capture log close
log using l10b_intro_cluster_data.txt, text replace

*Continuous data herd level predictor
use "simcont_clustherd.dta", clear
bysort herd: gen w=_n
tab X if w==1 // factor present in 50% herds
*ignoring clustering
reg milk X
* accounting for clustering
mixed milk X || herd:  , reml stddev
*herd average
collapse (mean) milk X, by(herd)
reg milk X

*Continuous data cow level predictor
use "simcont_clustcow.dta", clear
tab herd X, row nofreq //~50% cows treated within each herd
* ignoring clustering
reg milk X
* accounting for clustering
mixed milk X || herd:, reml stddev

*Discrete data herd level predictor
use "simbin_clustherd.dta", clear
bysort herd: gen w=_n
tab X if w==1 // factor present in 50% herds
tab Y X , col // disease level in un-exposed cows = 20%
cc Y X  // OR ~ 2
* ignoring clustering
logit Y X
* accounting for clustering
melogit Y X || herd:

*Discrete data cow level predictor
use "simbin_clustcow.dta", clear
tab herd X, row nofreq // ~50% cows treated within each herd
cc Y X  // OR ~ 2
* ignoring clustering
logit Y X
* accounting for clustering
melogit Y X || herd:
```