

## Lecture 11a: Introduction to cluster analysis (VER Ch. 20)

<b>Index</b>	<b>Page</b>
Clusters.....	2
Sources of clustering.....	2
Hierarchical vs cross-classified structures.....	4
Clustering of outcomes vs predictors.....	5
Effects of clustering.....	6
Example – effect of clustering – continuous data.....	7
Scenario 1- X herd level variable (Ex. 20.1).....	9
Scenario 2: X is a cow level variable (Ex. 20.1).....	10
Example – effect of clustering – discrete data.....	12
Variance inflation as a result of clustering.....	14
Stata code.....	15

## Clusters

- observations that share some feature(s) in common
- 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

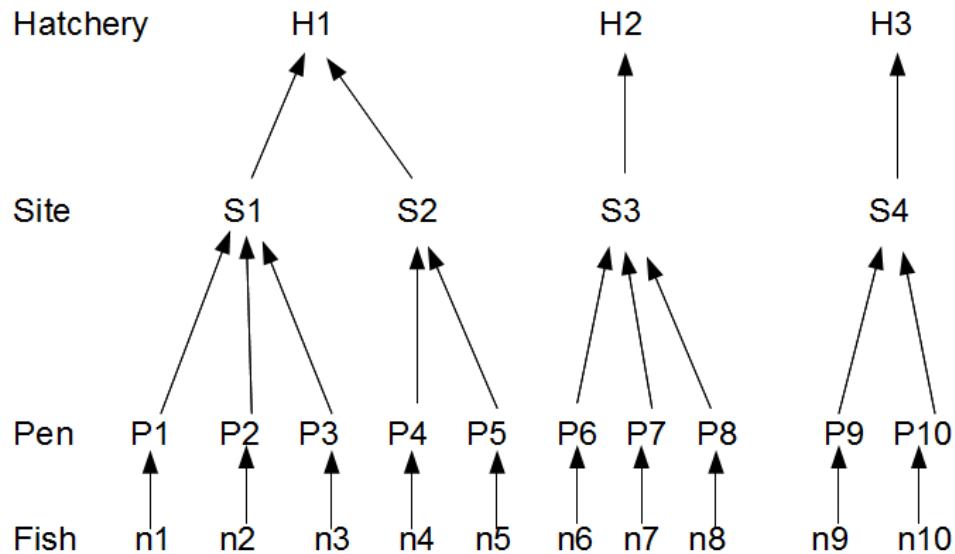
- common environment
  - ★ eg fish in a pen, cows in a herd, child in a school
  - ★ same correlation among all pairs
    - Daisy and Nelly has same correlation as Daisy and Jessie from the same farm
  - ★ multiple levels
    - region -> herd -> cow-> lactation
  
    - hierarchical or multilevel structure

- 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 - 1<sup>st</sup> 6 monthly tests of the lactation
    - dim = days in milk at time of test
    - milk = daily milk production (energy corrected)

Herd	Cow		Test Number					
			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

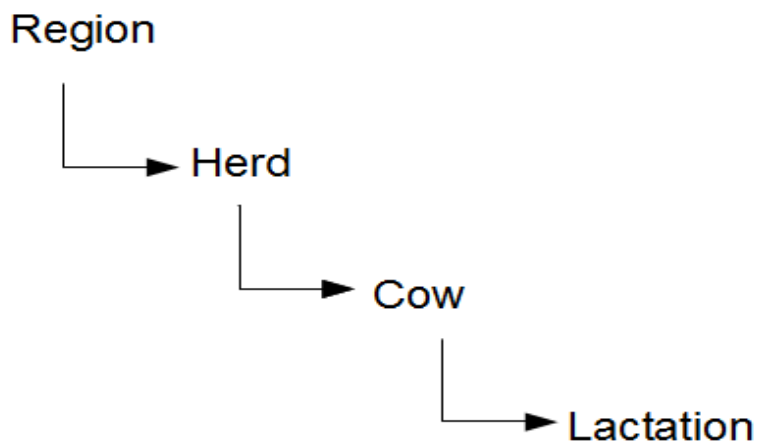
- (a) Hierarchical structure
  - ★ requires that every site receives fish from one hatchery
  - ★ requires that every pen is located within one site



- (b) Cross-classified structure
  - a pen can receive fish from different 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 data models (eg. logistic regression)
    - estimates are “marginal estimates” instead of “subject specific” estimates (described later)
    - eg. often get different parameter estimates if do/don't control for clustering

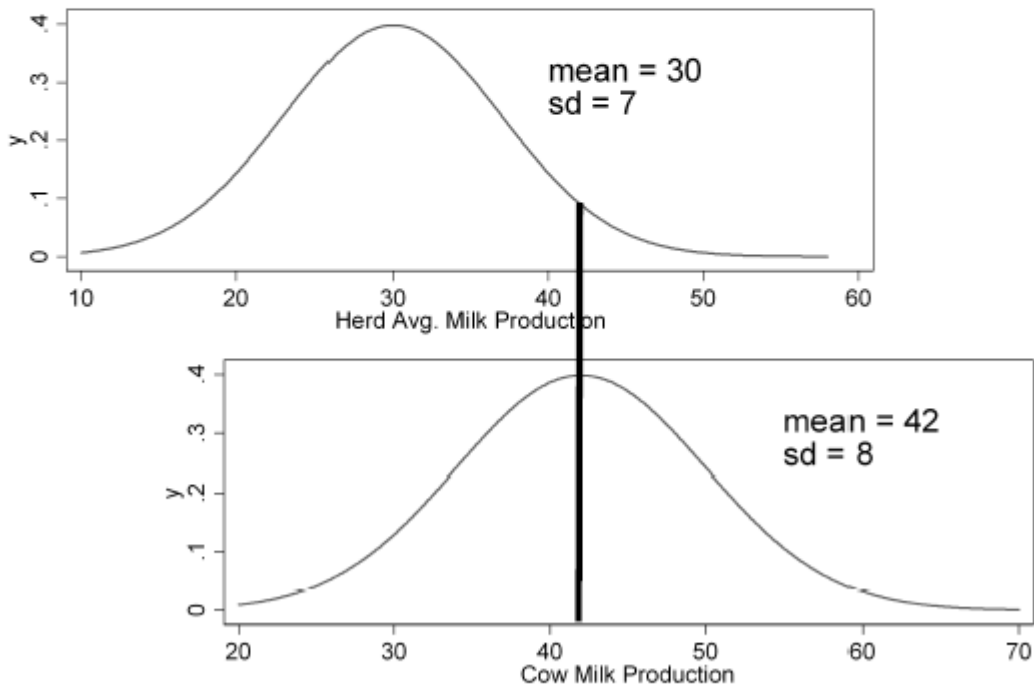
## Example – effect of clustering – continuous data

- 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



- herd:
  - TMR vs component
- cow:
  - rBST

- milk production
  - $\mu = 30$  kg/day
  - $\sigma_h = 7$  kg/day
  - $\sigma_i = 8$  kg/day



# Scenario 1- X herd level variable (Ex. 20.1)

## ● 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
```

```
Group variable: herd
```

```
Number of obs = 11626
```

```
Number of groups = 100
```

```
Obs per group: min = 20
```

```
avg = 116.3
```

```
max = 311
```

```
Log restricted-likelihood = -40902.479
```

```
Wald chi2(1) = 6.44
```

```
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 (Ex. 20.1)

★ X has a prevalence of 0.5 in all herds (eg. clinical trial)

### ● ignoring clustering

```
. reg milk X
```

Source	SS	df	MS			
Model	72138.7619	1	72138.7619	Number of obs =	11626	
Residual	1341880.62	11624	115.440522	F( 1, 11624) =	624.90	
Total	1414019.39	11625	121.636076	Prob > F =	0.0000	
				R-squared =	0.0510	
				Adj R-squared =	0.0509	
				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
Prob > chi2   =        0.0000

Log restricted-likelihood = -40947.175
```

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 scenario 1 and scenario 2

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

## Example – effect of clustering – discrete data

★ Effect of X on probability of disease (outcome)

★ cow level

→  $OR_x = 2$  (or  $\ln(2) = 0.693$ )

→ disease non-exposed  $p = 0.2$

★ herd level

→ herd level effects varied with  $SD = 1$

→ prevalence of X = 0.50

● scenario 1 – X is a herd level variable (Ex. 20.3)

★ ignoring clustering

```
. logit Y X
```

```
Logistic regression
```

```
Number of obs   =    11626  
LR chi2(1)      =    159.15  
Prob > chi2     =    0.0000  
Pseudo R2      =    0.0115
```

```
Log likelihood = -6814.7785
```

	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

. meqrlogit Y X || herd:

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

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

Integration points =      7            Wald chi2(1)      =         9.25
Log likelihood = -6065.0867          Prob > chi2       =         0.0024
```

Y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
X	.619974	.2038516	3.04	0.002	.2204322 1.019516
_cons	-1.305417	.1455518	-8.97	0.000	-1.590693 -1.020141

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
herd: Identity			
sd(_cons)	.9703703	.0769279	.8307243 1.133491

LR test vs. logistic regression: chibar2(01) = 1499.38 Prob>=chibar2 = 0.0000

### ● summary herd and cow level analysis

Dataset	Parameter	Logistic regression		Logistic mixed model	
		$\beta$	SE	$\beta$	SE
X herd level	X	0.53	0.04	0.62	0.20
	const.	-1.24	0.03	-1.31	0.15
X cow level	X	0.59	0.04	0.70	0.05
	const.	-1.25	0.03	-1.36	0.11

## 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 -  $\rho$  )
    - ➔ 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 ;  $\sigma^2 = \text{Var}(y_i)$ , and

➔  $VIF = [1 + (m - 1) * \rho]$  (Variance Inflation Factor)

ICC	m	VIF	Comments
0	20	1	no within group correlation = no VIF added to $\text{Var}(y)$
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

# Stata code

```
* do-file for lecture 11a of VHM 802/812, Winter 2014
* Introduction to clustered data
```

```
version 13
set more off
cd "c:\vhm_812\data"
```

```
*Continuous data herd level predictor
use "simcont_clustherd.dta", clear
*ignoring clustering
reg milk X
* accounting for clustering
mixed milk X || herd: , reml stddev
xtmixed milk X || herd:, reml //Version 12 only
```

```
*herd average
collapse (mean) milk X, by(herd)
reg milk X
```

```
*Continuous data cow level predictor
use "simcont_clustcow.dta", clear
* ignoring clustering
reg milk X
* accounting for clustering
mixed milk X || herd:, reml stddev
xtmixed milk X || herd:, reml //Version 12 only
```

```
*Discrete data herd level predictor
use "simbin_clustherd.dta", clear
* ignoring clustering
logit Y X
* accounting for clustering
mexrlogit Y X || herd:
xtmelogit Y X || herd: // Stata 12 only
```

```
*Discrete data cow level predictor
use "simbin_clustcow.dta", clear
* ignoring clustering
logit Y X
* accounting for clustering
mexrlogit Y X || herd:
xtmelogit Y X || herd: // Stata 12 only
```