Introduction to clustered data

do-file for lecture 11b of VHM 8120

```
. version 18

. set more off

. cd "c:\vhm812-data"
c:\vhm812-data

.
. capture log close

. log using l11b_intro_cluster_data, replace
-------------------------------------------------------------------------------
-------------------------------------------------------------
      name:  <unnamed>
       log:  c:\vhm812-data\l11b_intro_cluster_data.smcl
  log type:  smcl
 opened on:  21 Mar 2024, 10:05:29

.
. *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

          X |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         50       50.00       50.00
          1 |         50       50.00      100.00
------------+-----------------------------------
      Total |        100      100.00

.
. *ignoring clustering
. reg milk X

      Source |       SS           df       MS      Number of obs   =
11,626
-------------+----------------------------------   F(1, 11624)     =
317.72
       Model |  36598.5078         1  36598.5078   Prob > F        =
0.0000
    Residual |     1338999    11,624  115.192618   R-squared       =
0.0266
-------------+----------------------------------   Adj R-squared   =
0.0265
       Total |  1375597.51    11,625  118.330968   Root MSE        =
10.733
```

```
------------------------------------------------------------------------------
        milk | Coefficient  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 clustering
. mixed milk X || herd:  , reml stddev

Performing EM optimization ...

Performing gradient-based optimization:
Iteration 0:  Log restricted-likelihood = -40902.479
Iteration 1:  Log restricted-likelihood = -40902.479

Computing standard errors ...

Mixed-effects REML regression                         Number of obs    =
11,626
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 | Coefficient  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 model: chibar2(01) = 6374.40       Prob >= chibar2 =
0.0000

. *herd average
. collapse (mean) milk X, by(herd)

. reg milk X

      Source |       SS           df       MS      Number of obs   =
100
-------------+----------------------------------   F(1, 98)        =
6.37
       Model |   356.97798         1   356.97798   Prob > F        =
0.0132
    Residual |  5493.56229        98   56.056758   R-squared       =
0.0610
-------------+----------------------------------   Adj R-squared   =
0.0514
       Total |  5850.54027        99  59.0963664   Root MSE        =
7.4871

------------------------------------------------------------------------------
-
        milk | Coefficient  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
------------------------------------------------------------------------------
-


.
. *Continuous data cow level predictor
. use "simcont_clustcow.dta", clear

. tab herd X, row nofreq //~50% cows treated within each herd

           |           X
      herd |         0          1 |     Total
-----------+----------------------+----------
         1 |     50.00      50.00 |    100.00
         2 |     47.62      52.38 |    100.00
         3 |     48.00      52.00 |    100.00
```

4	50.00	50.00	100.00
5	48.15	51.85	100.00
6	50.00	50.00	100.00
7	50.00	50.00	100.00
8	48.28	51.72	100.00
9	48.28	51.72	100.00
10	48.28	51.72	100.00
11	50.00	50.00	100.00
12	50.00	50.00	100.00
13	50.00	50.00	100.00
14	48.39	51.61	100.00
15	50.00	50.00	100.00
16	48.57	51.43	100.00
17	48.57	51.43	100.00
18	48.57	51.43	100.00
19	48.65	51.35	100.00
20	48.65	51.35	100.00
21	50.00	50.00	100.00
22	50.00	50.00	100.00
23	48.72	51.28	100.00
24	50.00	50.00	100.00
25	48.78	51.22	100.00
26	50.00	50.00	100.00
27	50.00	50.00	100.00
28	50.00	50.00	100.00
29	48.89	51.11	100.00
30	48.89	51.11	100.00
31	50.00	50.00	100.00
32	50.00	50.00	100.00
33	50.00	50.00	100.00
34	48.94	51.06	100.00
35	48.94	51.06	100.00
36	48.94	51.06	100.00
37	48.94	51.06	100.00
38	48.94	51.06	100.00
39	50.00	50.00	100.00
40	50.00	50.00	100.00
41	48.98	51.02	100.00
42	48.98	51.02	100.00
43	50.00	50.00	100.00
44	49.02	50.98	100.00
45	49.02	50.98	100.00
46	50.00	50.00	100.00
47	49.06	50.94	100.00
48	49.06	50.94	100.00
49	49.09	50.91	100.00
50	49.09	50.91	100.00
51	49.45	50.55	100.00
52	49.53	50.47	100.00
53	50.00	50.00	100.00
54	49.58	50.42	100.00
55	49.58	50.42	100.00
56	50.00	50.00	100.00
57	49.62	50.38	100.00
58	50.00	50.00	100.00
59	49.65	50.35	100.00
60	50.00	50.00	100.00

```
        61 |      49.66      50.34 |    100.00
        62 |      50.00      50.00 |    100.00
        63 |      50.00      50.00 |    100.00
        64 |      50.00      50.00 |    100.00
        65 |      49.68      50.32 |    100.00
        66 |      50.00      50.00 |    100.00
        67 |      50.00      50.00 |    100.00
        68 |      50.00      50.00 |    100.00
        69 |      50.00      50.00 |    100.00
        70 |      49.72      50.28 |    100.00
        71 |      50.00      50.00 |    100.00
        72 |      49.72      50.28 |    100.00
        73 |      50.00      50.00 |    100.00
        74 |      49.73      50.27 |    100.00
        75 |      50.00      50.00 |    100.00
        76 |      49.74      50.26 |    100.00
        77 |      49.75      50.25 |    100.00
        78 |      50.00      50.00 |    100.00
        79 |      49.76      50.24 |    100.00
        80 |      50.00      50.00 |    100.00
        81 |      49.76      50.24 |    100.00
        82 |      49.77      50.23 |    100.00
        83 |      49.77      50.23 |    100.00
        84 |      49.77      50.23 |    100.00
        85 |      49.77      50.23 |    100.00
        86 |      50.00      50.00 |    100.00
        87 |      49.77      50.23 |    100.00
        88 |      49.77      50.23 |    100.00
        89 |      49.78      50.22 |    100.00
        90 |      49.78      50.22 |    100.00
        91 |      50.00      50.00 |    100.00
        92 |      49.80      50.20 |    100.00
        93 |      50.00      50.00 |    100.00
        94 |      50.00      50.00 |    100.00
        95 |      50.00      50.00 |    100.00
        96 |      50.00      50.00 |    100.00
        97 |      50.00      50.00 |    100.00
        98 |      49.82      50.18 |    100.00
        99 |      49.84      50.16 |    100.00
       100 |      49.84      50.16 |    100.00
-----------+---------------------+----------
     Total |      49.76      50.24 |    100.00

. * ignoring clustering
. reg milk X

      Source |       SS           df       MS      Number of obs   =
11,626
-------------+----------------------------------   F(1, 11624)     =
624.90
       Model |  72138.7619         1  72138.7619   Prob > F        =
0.0000
    Residual |  1341880.62    11,624  115.440522   R-squared       =
0.0510
-------------+----------------------------------   Adj R-squared   =
0.0509
```

```
       Total |  1414019.39    11,625  121.636076   Root MSE        =
10.744

------------------------------------------------------------------------------
        milk | Coefficient  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

Performing EM optimization ...

Performing gradient-based optimization:
Iteration 0:  Log restricted-likelihood = -40947.175
Iteration 1:  Log restricted-likelihood = -40947.175

Computing standard errors ...

Mixed-effects REML regression                       Number of obs    =
11,626
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 | Coefficient  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 model: chibar2(01) = 6310.00       Prob >= chibar2 =
0.0000

.
. *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

          X |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         50       50.00       50.00
          1 |         50       50.00      100.00
------------+-----------------------------------
      Total |        100      100.00

. tab Y X , col // disease level in un-exposed cows = 20%

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

           |          X
         Y |         0          1 |     Total
-----------+----------------------+----------
         0 |     4,206      4,164 |     8,370
           |     77.59      67.11 |     71.99
-----------+----------------------+----------
         1 |     1,215      2,041 |     3,256
           |     22.41      32.89 |     28.01
-----------+----------------------+----------
     Total |     5,421      6,205 |    11,626
           |    100.00     100.00 |    100.00

. cc Y X  // OR ~ 2
                                                              Proportion
                 |   Exposed   Unexposed  |      Total      exposed
```

```
-----------------+------------------------+------------------------
           Cases |      2041        1215  |       3256       0.6268
        Controls |      4164        4206  |       8370       0.4975
-----------------+------------------------+------------------------
           Total |      6205        5421  |      11626       0.5337
                 |                        |
                 |      Point estimate    |    [95% conf. interval]
                 |------------------------+------------------------
      Odds ratio |         1.696779       |    1.560346     1.84516 (exact)
 Attr. frac. ex. |          .410648       |    .3591165    .4580415 (exact)
 Attr. frac. pop |         .2574118       |
                 +-----------------------------------------------
                              chi2(1) =   157.60  Pr>chi2 = 0.0000

. * ignoring clustering
. logit Y X

Iteration 0:  Log likelihood = -6894.3552
Iteration 1:  Log likelihood = -6815.0583
Iteration 2:  Log likelihood = -6814.7785
Iteration 3:  Log likelihood = -6814.7785

Logistic regression                                     Number of obs =
11,626
                                                        LR chi2(1)    =
159.15
                                                        Prob > chi2   =
0.0000
Log likelihood = -6814.7785                             Pseudo R2     =
0.0115

------------------------------------------------------------------------------
-
           Y | Coefficient  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:

Fitting fixed-effects model:

Iteration 0:  Log likelihood = -6828.9777
Iteration 1:  Log likelihood = -6814.7876
Iteration 2:  Log likelihood = -6814.7785
Iteration 3:  Log likelihood = -6814.7785

Refining starting values:

Grid node 0:  Log likelihood =  -6065.269
```

```
Fitting full model:

Iteration 0:  Log likelihood =  -6065.269
Iteration 1:  Log likelihood =  -6065.089
Iteration 2:  Log likelihood = -6065.0864
Iteration 3:  Log likelihood = -6065.0864

Mixed-effects logistic regression               Number of obs     =
11,626
Group variable: herd                            Number of groups  =
100

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

Integration method: mvaghermite                 Integration pts.  =
7

                                                Wald chi2(1)      =
9.26
Log likelihood = -6065.0864                     Prob > chi2       =
0.0023
------------------------------------------------------------------------------
           Y | Coefficient  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

.
. *Discrete data cow level predictor
. use "simbin_clustcow.dta", clear

. tab herd X, row nofreq // ~50% cows treated within each herd

           |           X
      herd |         0          1 |     Total
-----------+----------------------+----------
```

1	50.00	50.00	100.00
2	47.62	52.38	100.00
3	48.00	52.00	100.00
4	50.00	50.00	100.00
5	48.15	51.85	100.00
6	50.00	50.00	100.00
7	50.00	50.00	100.00
8	48.28	51.72	100.00
9	48.28	51.72	100.00
10	48.28	51.72	100.00
11	50.00	50.00	100.00
12	50.00	50.00	100.00
13	50.00	50.00	100.00
14	48.39	51.61	100.00
15	50.00	50.00	100.00
16	48.57	51.43	100.00
17	48.57	51.43	100.00
18	48.57	51.43	100.00
19	48.65	51.35	100.00
20	48.65	51.35	100.00
21	50.00	50.00	100.00
22	50.00	50.00	100.00
23	48.72	51.28	100.00
24	50.00	50.00	100.00
25	48.78	51.22	100.00
26	50.00	50.00	100.00
27	50.00	50.00	100.00
28	50.00	50.00	100.00
29	48.89	51.11	100.00
30	48.89	51.11	100.00
31	50.00	50.00	100.00
32	50.00	50.00	100.00
33	50.00	50.00	100.00
34	48.94	51.06	100.00
35	48.94	51.06	100.00
36	48.94	51.06	100.00
37	48.94	51.06	100.00
38	48.94	51.06	100.00
39	50.00	50.00	100.00
40	50.00	50.00	100.00
41	48.98	51.02	100.00
42	48.98	51.02	100.00
43	50.00	50.00	100.00
44	49.02	50.98	100.00
45	49.02	50.98	100.00
46	50.00	50.00	100.00
47	49.06	50.94	100.00
48	49.06	50.94	100.00
49	49.09	50.91	100.00
50	49.09	50.91	100.00
51	49.45	50.55	100.00
52	49.53	50.47	100.00
53	50.00	50.00	100.00
54	49.58	50.42	100.00
55	49.58	50.42	100.00
56	50.00	50.00	100.00
57	49.62	50.38	100.00

```
        58 |      50.00      50.00 |    100.00
        59 |      49.65      50.35 |    100.00
        60 |      50.00      50.00 |    100.00
        61 |      49.66      50.34 |    100.00
        62 |      50.00      50.00 |    100.00
        63 |      50.00      50.00 |    100.00
        64 |      50.00      50.00 |    100.00
        65 |      49.68      50.32 |    100.00
        66 |      50.00      50.00 |    100.00
        67 |      50.00      50.00 |    100.00
        68 |      50.00      50.00 |    100.00
        69 |      50.00      50.00 |    100.00
        70 |      49.72      50.28 |    100.00
        71 |      50.00      50.00 |    100.00
        72 |      49.72      50.28 |    100.00
        73 |      50.00      50.00 |    100.00
        74 |      49.73      50.27 |    100.00
        75 |      50.00      50.00 |    100.00
        76 |      49.74      50.26 |    100.00
        77 |      49.75      50.25 |    100.00
        78 |      50.00      50.00 |    100.00
        79 |      49.76      50.24 |    100.00
        80 |      50.00      50.00 |    100.00
        81 |      49.76      50.24 |    100.00
        82 |      49.77      50.23 |    100.00
        83 |      49.77      50.23 |    100.00
        84 |      49.77      50.23 |    100.00
        85 |      49.77      50.23 |    100.00
        86 |      50.00      50.00 |    100.00
        87 |      49.77      50.23 |    100.00
        88 |      49.77      50.23 |    100.00
        89 |      49.78      50.22 |    100.00
        90 |      49.78      50.22 |    100.00
        91 |      50.00      50.00 |    100.00
        92 |      49.80      50.20 |    100.00
        93 |      50.00      50.00 |    100.00
        94 |      50.00      50.00 |    100.00
        95 |      50.00      50.00 |    100.00
        96 |      50.00      50.00 |    100.00
        97 |      50.00      50.00 |    100.00
        98 |      49.82      50.18 |    100.00
        99 |      49.84      50.16 |    100.00
       100 |      49.84      50.16 |    100.00
-----------+----------------------+----------
     Total |      49.76      50.24 |    100.00

. cc Y X  // OR ~ 2

                                                      Proportion
                 |   Exposed   Unexposed  |      Total      exposed
-----------------+------------------------+------------------------
           Cases |      1985        1288  |       3273       0.6065
        Controls |      3856        4497  |       8353       0.4616
-----------------+------------------------+------------------------
           Total |      5841        5785  |      11626       0.5024
                 |                        |
                 |      Point estimate    |    [95% conf. interval]
                 |------------------------+------------------------
```

```
      Odds ratio |         1.797341       |     1.65396    1.953169 (exact)
 Attr. frac. ex. |         .4436226       |    .3953904    .4880115 (exact)
 Attr. frac. pop |          .269047       |
                 +------------------------------------------------
                                   chi2(1) =   197.35  Pr>chi2 = 0.0000

. * ignoring clustering
. logit Y X

Iteration 0:  Log likelihood = -6910.3442
Iteration 1:  Log likelihood =   -6811.48
Iteration 2:  Log likelihood = -6811.0741
Iteration 3:  Log likelihood = -6811.0741

Logistic regression                                     Number of obs =
11,626
                                                        LR chi2(1)    =
198.54
                                                        Prob > chi2   =
0.0000
Log likelihood = -6811.0741                             Pseudo R2     =
0.0144

------------------------------------------------------------------------------
-
           Y | Coefficient  Std. err.      z    P>|z|     [95% conf.
interval]
-------------+----------------------------------------------------------------
-
           X |   .5863084   .0419748    13.97   0.000     .5040393
.6685775
       _cons |   -1.25032   .0316033   -39.56   0.000    -1.312261   -
1.188379
------------------------------------------------------------------------------
-

. * accounting for clustering
. melogit Y X || herd:

Fitting fixed-effects model:

Iteration 0:  Log likelihood =  -6824.417
Iteration 1:  Log likelihood = -6811.0819
Iteration 2:  Log likelihood = -6811.0741
Iteration 3:  Log likelihood = -6811.0741

Refining starting values:

Grid node 0:  Log likelihood = -5999.0535

Fitting full model:

Iteration 0:  Log likelihood = -5999.0535
Iteration 1:  Log likelihood = -5995.9716
Iteration 2:  Log likelihood = -5995.9694
Iteration 3:  Log likelihood = -5995.9694
```

```
Mixed-effects logistic regression               Number of obs     =
11,626
Group variable: herd                            Number of groups  =
100

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

Integration method: mvaghermite                 Integration pts.  =
7

                                                Wald chi2(1)      =
229.28
Log likelihood = -5995.9694                     Prob > chi2       =
0.0000
------------------------------------------------------------------------------
           Y | Coefficient  Std. err.      z    P>|z|     [95% conf.
interval]
-------------+----------------------------------------------------------------
           X |   .6974798    .046063    15.14   0.000     .6071979
.7877616
       _cons |  -1.361196   .1111563   -12.25   0.000    -1.579059   -
1.143334
-------------+----------------------------------------------------------------
herd         |
   var(_cons)|   1.068314   .1682536                      .7845836
1.45465
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 1630.21     Prob >= chibar2 =
0.0000

.
. log close
      name:  <unnamed>
       log:  c:\vhm812-data\l11b_intro_cluster_data.smcl
  log type:  smcl
 closed on:  21 Mar 2024, 10:05:30
------------------------------------------------------------------------------
-------------------------------------------------------------
```