Epidemiology VHM 8120 Course
Winter 2021, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 2a: Model building I (VER 14.4, 14.6-7)

Index **Page**

- Learning objectives
  - understand how to model different type of variables
  - review confounding and interaction in regression models
  - interpret regression model coefficients
- Exercises
  - work on exercise 2 linear regression
- Quiz 1 – Wednesday Jan 20th

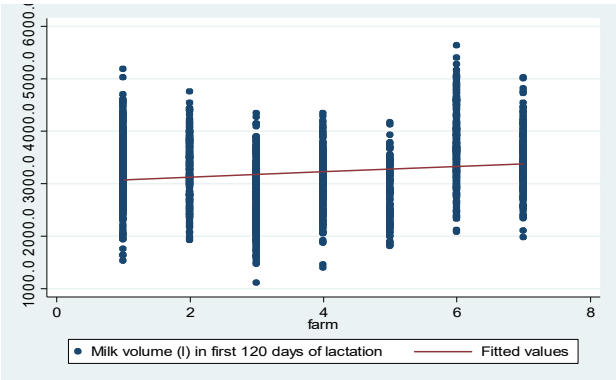
Predictors (X variables)

- Categorical
  - nominal – values represent "levels" (no numerical meaning)
  - ordinal – values represent ordered levels
  - must be recoded
    - indicator or dummy variables
- Continuous (quantitative)
  - scaling
  - assumption of linearity

Categorical predictors

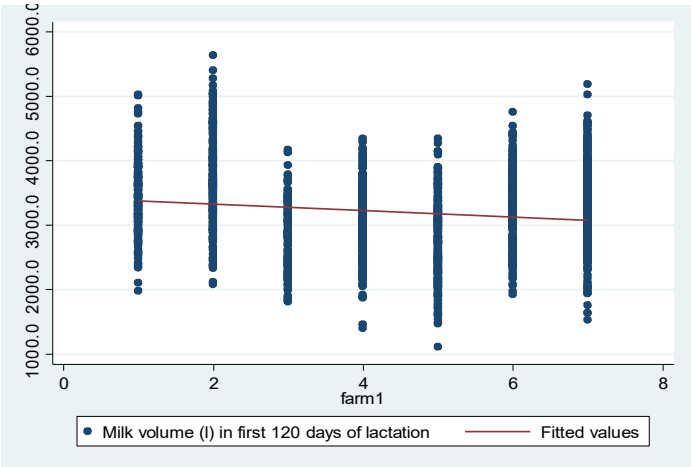
- Nominal (and sometimes ordinal) predictors with more than two levels should not be used as numeric

```
. table farm_id, c(mean milk120)

-------------------------

  farm_id | mean(milk120)
----------+--------------
        1 |        3357.8
        2 |        3279.3
        3 |        2778.6
        4 |        3032.5
        5 |        2838.2
        6 |        3730.6
        7 |        3435.8
-------------------------
```




```
. regress milk120 farm
------------------------------------------------------------------------------
     milk120 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        farm |   50.90872   8.848643     5.75   0.000       33.552     68.26543
       _cons |   3026.143   37.27521    81.18   0.000     2953.028     3099.259
------------------------------------------------------------------------------
```

```
. list farm_id milk120 farm1_id

     +-------------------------+
     | farm_jd milk120 farm1_id|
     |-------------------------|
  1. |     1    3357.8         7 |
  2. |     2    3279.3         6 |
  3. |     3    2778.6         5 |
  4. |     4    3032.5         4 |
  5. |     5    2838.2         3 |
     |-------------------------|
  6. |     6    3730.6         2 |
  7. |     7    3435.8         1 |
     +-------------------------+
```




```
------------------------------------------------------------------------------
     milk120 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       farm1 |  -50.90872   8.848643    -5.75   0.000    -68.26543     -33.552
       _cons |   3433.413   41.84204    82.06   0.000     3351.339    3515.487
```

Indicator variables

- Convert nominal or ordinal variables to a set of dichotomous variables or indicator variables
  - assign observations to one of two categories (usually 0 and 1)
  - j-1 indicator variables are required in the regression model
- One level is referent (or baseline) level

Obs. #	parity	parity_1	parity_2	parity_3
1	1	1	0	0
2	2	0	1	0
3	3	0	0	1

- Choice of referent (base) level
  - biological sense and ease of interpretation
  - reasonable sample size
  - Stata - default smallest values as base category (help *fvvarlist*)
  - Changing referent no effect on overall model fit (eg same R^2)
  - All in / all out
- Example daisy2red- parity as an ordinal variable
  - regress milk120 i.parity

```
------------------------------------------------------------------------------
     milk120 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      parity |
          2  |   708.2134   43.74086    16.19   0.000     622.4149    794.0118
          3  |   789.8435   45.94638    17.19   0.000     699.7189    879.9681
          4  |   848.5137   50.96418    16.65   0.000     748.5467    948.4808
          5  |   787.6091   56.22915    14.01   0.000     677.3147    897.9035
          6  |   878.1606   79.09305    11.10   0.000     723.0183    1033.303
          7  |   925.9548   305.0416     3.04   0.002     327.6106    1524.299
             |
       _cons |   2639.645    30.2407    87.29   0.000     2580.328    2698.963
------------------------------------------------------------------------------
```

  - compare with simple average

Parity	Avg milk120	Indicator variables
1	2639.7	2639.7*
2	3347.9	708.2
3	3429.5	789.8
4	3488.2	848.5
5	3427.3	787.6
6	3517.8	878.2
7	3562.6	925.9
*intercept (or _cons)		

Continuous predictors

Improving the "interpretation" of X variables

- Scaling X variables
  - limited range of plausible values
    - effects only the constant (not the coefficient for the variable – slope)
    - subtract min. plausible value (eg. parity)

```
. regress milk120 parity
...output omitted
------------------------------------------------------------------------------
     milk120 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      parity |    178.347   11.01266    16.19   0.000     156.7455    199.9484
       _cons |    2727.08   34.33991    79.41   0.000     2659.722    2794.438
------------------------------------------------------------------------------
```

```
. gen parity_1=parity-1

. reg milk120 parity_1
....output omitted
------------------------------------------------------------------------------
     milk120 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    parity_1 |    178.347   11.01266    16.19   0.000     156.7455    199.9484
       _cons |    2905.427   25.23474   115.14   0.000     2855.928    2954.925
------------------------------------------------------------------------------
```

  - subtract the central value (eg. mean) (centring)
    - helps to reduce collinearity with quadratic and interaction terms

```
. reg milk120 c.herd_size##c.herd_size

------------------------------------------------------------------------------
             milk120 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
           herd_size |    28.73126   1.993023    14.42   0.000     24.82192     32.6406
c.herd_size#c.herd_size |  -.0608255   .0041101   -14.80   0.000    -.0688875   -.0527634
               _cons |    66.06488   231.8877     0.28   0.776    -388.7858    520.9155
------------------------------------------------------------------------------
```

```
. estat vce, corr -> correlation = -0.9907

.summ herd_size
```

```
.gen hrdsz_ctr=herd_size - 251

. reg milk120 c.hrdsz_ctr##c.hrdsz_ctr
```

milk120	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hrdsz_ctr	-1.803116	.2847073	-6.33	0.000	-2.361573	-1.244659
c.hrdsz_ctr#c.hrdsz_ctr	-.0608255	.0041101	-14.80	0.000	-.0688875	-.0527634
_cons	3445.547	22.80494	151.09	0.000	3400.815	3490.279

```
. estat vce, corr -> correlation = 0.3115
```

- scale of measurement (eg grams or kg)
  - look at range and IQR and establish acceptable values
  - avoid very small regression coefficients
  - example: herd size (mean=251; min=125; max=333)

```
. reg milk120 herd_size
```

milk120	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
herd_size	-.49048	.2891242	-1.70	0.090	-1.057601	.0766405
_cons	3338.164	74.69789	44.69	0.000	3191.643	3484.685

```
. gen herdsz_100=herd_size/100 /*rescale herd_size so coef are larger*/

. reg milk120 herdsz_100
```

milk120	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
herdsz_100	-49.04796	28.91242	-1.70	0.090	-105.76	7.664098
_cons	3338.164	74.69789	44.69	0.000	3191.643	3484.685

Detecting confounding (VER 13.5)

- Review
  - components: Y (outcome), E (exposure) and Z (measured or unmeasured confounder)
  - criteria
    - Z must be a risk factor of Y (in E-,because risk must not be caused by E->Y))
    - Z must be associated with E
      - cohort – start follow up period
        - if constant during follow up -> look for unconditional assoc. Z->E
      - case-control – in controls (represent source population if no selection bias)
    - Z must not be the result of E or the result of Y
  - causal model

Vaginal Discharge

Herd

WPC

- Assessment of confounding
  - when three criteria are met
  - difference between crude and adjusted effect/association changes substantially
    - (crude-adjusted)/crude
    - eg. +/- 20-30%
- Example - change in regression coefficient

```
. reg wpc i.vag_disch              /* vag_disch adds 12 days, P=0.04 */

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   vag_disch |
         yes |   11.99647   5.846716     2.05   0.040     .5282858    23.46465
       _cons |   68.17426   1.334494    51.09   0.000     65.55669    70.79184
------------------------------------------------------------------------------


. reg wpc i.herd  if vag_disch==0 /*herd is associated with WPC*/

      Source |       SS           df       MS      Number of obs   =     1,492
-------------+----------------------------------   F(6, 1485)      =     14.92
       Model |  223706.812         6  37284.4686   Prob > F        =    0.0000
    Residual |  3711781.88     1,485  2499.51642   R-squared       =    0.0568
-------------+----------------------------------   Adj R-squared   =    0.0530
       Total |  3935488.69     1,491   2639.4961   Root MSE        =    49.995

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        herd |
          2  |  -7.455598   4.572715    -1.63   0.103    -16.42527     1.51407
          3  |   12.30396   4.184586     2.94   0.003     4.095636    20.51229
          4  |   -20.0733    4.70619    -4.27   0.000    -29.30479   -10.84181
          5  |  -21.78125   5.489853    -3.97   0.000    -32.54994   -11.01256
        106  |  -15.40129   4.618509    -3.33   0.001    -24.46079   -6.341796
        119  |  -17.26021    5.05679    -3.41   0.001    -27.17942      -7.341
             |
       _cons |    75.1583   3.106548    24.19   0.000     69.06461    81.25199
------------------------------------------------------------------------------
```

```
. tab herd vag_disch, chi row /* herd is associated with vag_disc*/

           |   Vaginal discharge
      Herd |       observed
    Number |        no        yes |     Total
-----------+----------------------+----------
         3 |       318          4 |       322
           |     98.76       1.24 |    100.00
-----------+----------------------+----------
 ...output omitted
-----------+----------------------+----------
       106 |       214         39 |       253
           |     84.58      15.42 |    100.00
-----------+----------------------+----------
     Total |     1,492         82 |     1,574
           |     94.79       5.21 |    100.00

          Pearson chi2(6) =  74.2267   Pr = 0.000
```

```
. reg wpc i.vag_disch i.herd

      Source |       SS           df       MS      Number of obs   =     1,574
-------------+----------------------------------   F(7, 1566)      =     14.35
       Model |   252509.59         7  36072.7985   Prob > F        =    0.0000
    Residual |  3935580.97     1,566  2513.14238   R-squared       =    0.0603
-------------+----------------------------------   Adj R-squared   =    0.0561
       Total |  4188090.56     1,573  2662.48605   Root MSE        =    50.131

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   vag_disch |
         yes |   17.81936   5.825177     3.06   0.002     6.393393    29.24533
             |
        herd |
           2 |  -8.178615   4.509228    -1.81   0.070    -17.02338    .6661455
           3 |   11.71303   4.133611     2.83   0.005     3.605039    19.82103
           4 |  -20.74627   4.653654    -4.46   0.000    -29.87432   -11.61823
           5 |  -22.15954   5.359351    -4.13   0.000     -32.6718   -11.64728
         106 |  -18.18881   4.422295    -4.11   0.000    -26.86305   -9.514562
         119 |  -17.85657   4.920293    -3.63   0.000    -27.50763   -8.205518
             |
       _cons |   75.97555   3.052377    24.89   0.000     69.98837    81.96272
------------------------------------------------------------------------------
```

Confounding and collinearity

- Confounding => collinearity
  - example: bwt – smoking (only for example, not for model building)
    - cig_2, cig_3 on bwt
  - two models
    - 1) E = cig_2
    - 2) E = cig_3
  - causal diagrams / criteria

- Model 1
  - cig_3 is not a confounder
  - cig_3 intervening variable
  - cig_3 highly correlated with cig_2 -> keep cig_2
- Model 2
  - cig_2 meets (partially) confounding criteria
  - change of coefficient?

Detecting and modeling interaction

- cross-product term
  - eg. stata = i.vag_disch##i.rp
- examples 14.9, 14.10 and 14.11

Interaction between 2 dichotomous predictors

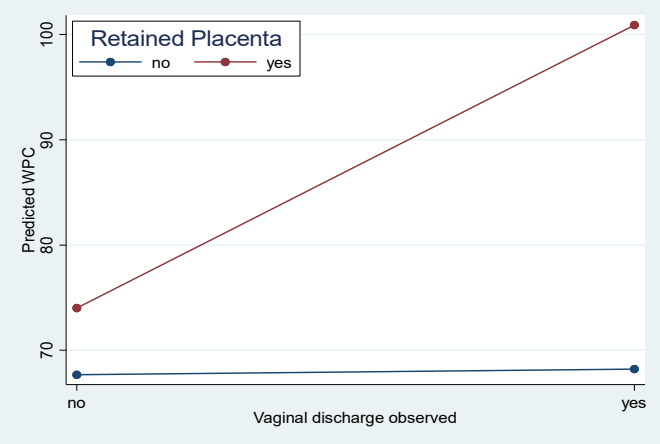
```
. reg wpc i.vag_disch##i.rp
```

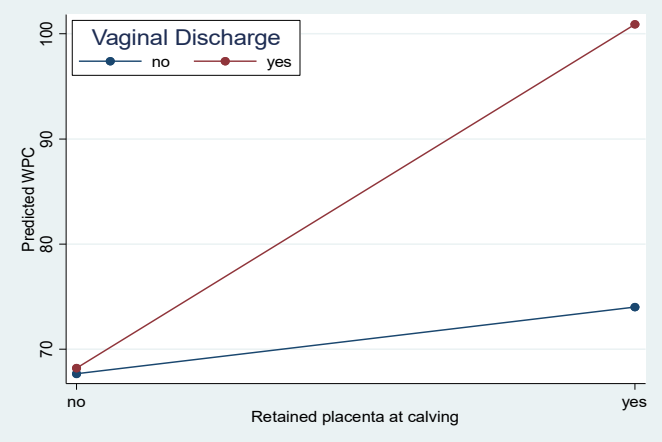
```
      Source |       SS       df       MS              Number of obs =    1574
-------------+------------------------------           F(  3,  1570) =    4.53
       Model |  35915.9774     3  11971.9925           Prob > F      =  0.0036
    Residual |  4152174.58  1570  2644.69719           R-squared     =  0.0086
-------------+------------------------------           Adj R-squared =  0.0067
       Total |  4188090.56  1573  2662.48605           Root MSE      =  51.427

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.vag_disch |   .5429296   7.265382     0.07   0.940    -13.70794     14.7938
        1.rp |   6.339794   4.914322     1.29   0.197    -3.299531    15.97912
             |
vag_disch#rp |
        1 1  |   26.34867   12.77367     2.06   0.039     1.293414    51.40392
             |
       _cons |   67.66861   1.387883    48.76   0.000     64.94631    70.39091
------------------------------------------------------------------------------
```

```
. table vag_disch rp, c(mean wpc) // display wpc means by vag_dich and rp
```

Vaginal discharge observed	Retained placenta at calving: no	Retained placenta at calving: yes
no	67.66861	74.0084
yes	68.21154	100.9

- Interaction plots (margins command - later)

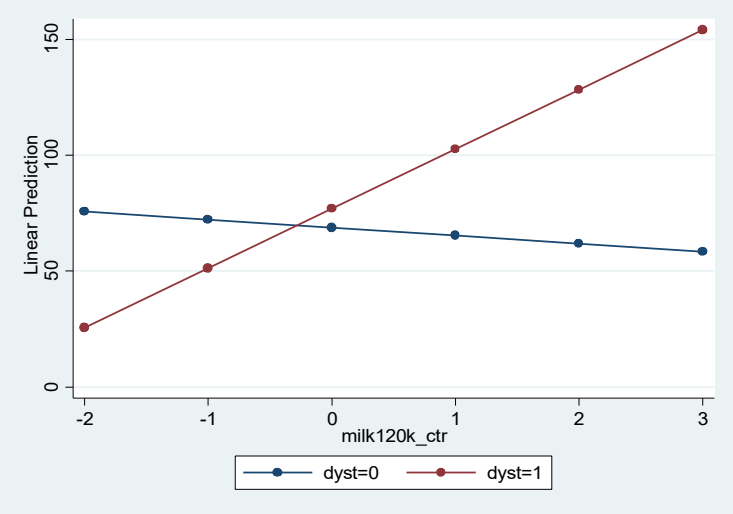




Interaction between a dichotomous and a continuous predictor

```
. reg wpc i.dyst##c.milk120k_ctr

  ...output ommitted
```

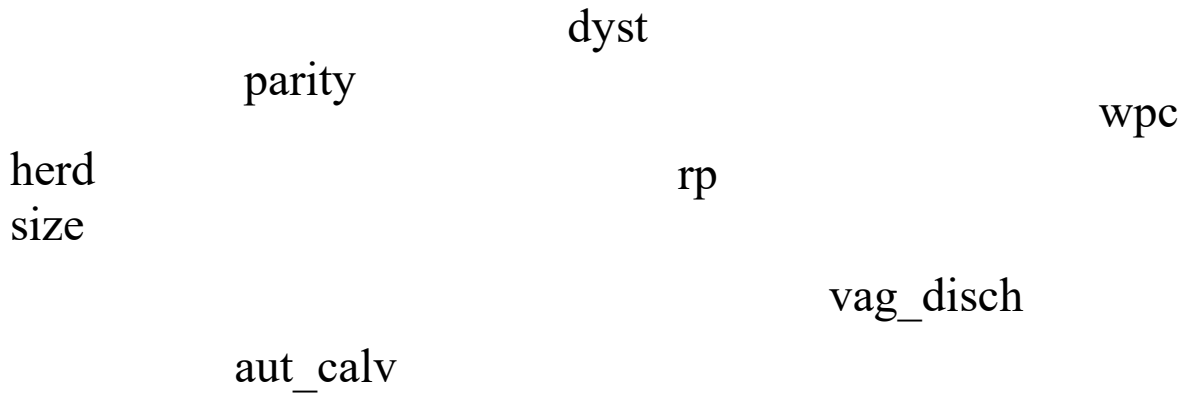
wpc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dyst						
yes	8.20714	5.718528	1.44	0.151	-3.00983	19.42411
milk120k_ctr	-3.446531	1.928535	-1.79	0.074	-7.229379	.3363161
dyst#c.milk120k_ctr						
yes	29.14238	9.468101	3.08	0.002	10.57057	47.71419
_cons	68.75682	1.357147	50.66	0.000	66.09475	71.41888



Interaction between 2 continuous predictors

- eg. reg wpc c.parity_1##c.milk120k_ctr
- Two way interactions between continuous predictors are difficult to interpret, and, whenever significant, should be evaluated by fitting a range of possible values for both predictors.
  - be sure that predictions are within range of the data

Causal interpretation (VER 14.7)



- Graphical assessment confounding
  - draw causal diagram
    - nodes=variables; arrow=causal relations
    - time on horizontal axis (right most recent)
  - identify [and remove] intermediate variables
  - identify potential confounders
  - identify collider variables
    - controlling for the effect of 2 variables (eg. dyst) will create an spurious association between them (eg. between herd size and parity)

```
. reg wpc c.hs100_ctr##c.hs100_ctr parity_1 i.aut_calv i.twin i.dyst
i.rp##vag_disch, vsquish

      Source |       SS           df       MS      Number of obs   =     1,574
-------------+----------------------------------   F(9, 1564)      =     13.22
       Model |  296062.694         9  32895.8549   Prob > F        =    0.0000
    Residual |  3892027.86     1,564  2488.50886   R-squared       =    0.0707
-------------+----------------------------------   Adj R-squared   =    0.0653
       Total |  4188090.56     1,573  2662.48605   Root MSE        =    49.885

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   hs100_ctr |   19.85708   2.163397     9.18   0.000     15.61361    24.10054
 c.hs100_ctr#|
 c.hs100_ctr |   11.13827   3.111145     3.58   0.000     5.035817    17.24073
    parity_1 |    1.13721   .8583103     1.32   0.185    -.5463501     2.82077
  1.aut_calv |  -8.263839   2.537751    -3.26   0.001    -13.24159   -3.286086
        twin |
         yes |   20.68314   9.845165     2.10   0.036      1.37203    39.99425
        dyst |
         yes |   11.70041   5.462576     2.14   0.032      .985666    22.41516
          rp |
         yes |    5.98687   4.811976     1.24   0.214    -3.451734    15.42547
   vag_disch |
         yes |   1.228196   7.161395     0.17   0.864    -12.81875    15.27514
rp#vag_disch |
    yes#yes |   22.85194   12.51605     1.83   0.068    -1.698056    47.40194
       _cons |   64.33029   2.634114    24.42   0.000     59.16352    69.49705
------------------------------------------------------------------------------
```

- herd_size
- parity
- aut_calv
- twin
- dyst
- rp
- vag_disch

Assessing linearity

- Assumption about nature of relationship between X and Y
  - note: the following are all discussed in more detail under Model Building (Chapter 15)
- **Detecting non-linearity – in final model**
  - plot residuals vs fitted values (see L1a)
    - simultaneous evaluation of all predictors
    - plot of residuals vs predictor (see L1a)
- **Detecting non-linearity – before / during model building**
  - smoothed scatter plot of outcome vs predictor
  - explore polynomial functions of X
  - transformation of X
  - categorization of predictor
    - indicator dummy variable
    - compare categorical and linear variables

Stata Factor-Notation basics

- For predictors x, z and outcome y
  - i.x = categorical effect of x (x must be integer)
  - c.x = continuous effect (slope) of x (x must be numerical)
  - Stata default depends on command
    - reg y x = reg y c.x (default is c.x)
    - anova y x = anova y i.x (default is i.x)
- Combined effects
  - c.x##c.x = continuous terms for x and x^2
  - interaction
    - x#z = interaction x * z;
    - in all commands the default is x#z = i.x#i.z
    - c.x#c.z = need to use "c" for continuous variables
    - always x##z = x z x#z
- Factor terms can be used in tests
  - testparm i.x
  - testparm c.x##c.x (test c.x and c.x#c.x
  - testparm c.x#c.x (test c.x#c.x)