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

Lecture 3b: Model building II (VER Ch. 15)

Index **Page**

- Today
  - lecture
  - linear regression assignment
  - review Quiz #1 and questions linear regression exercise 3.
- Wednesday: Quiz #2
- Friday – Model building exercise – (Lab Jan 28th)

Model building strategies

- Parsimony vs fit
- Goals
  - estimates of effects
  - prediction
- Steps
  - specify maximum model
    - reduce number of predictors
    - address issues of missing values
    - functional form (linearity) of continuous predictors
  - criterion for selection
  - selection strategy
  - analysis
  - evaluate reliability
  - present results

Specifying the maximum model

- Outcome of interest
- Key predictors
- Important confounders/interactions
- Other variables of interest
  - lots / few

Causal model (DO NOT FORGET)

- Identify confounders
- Identify intervening variables
- Identify exposure-independent variables

Reducing the number of predictors

- Screening – descriptive statistics
  - few missing values
  - substantial variability
  - small number observations - re-categorize predictor
- Correlation / association
  - pairwise
- Unconditional associations
  - liberal P-value
- Multivariate analysis (eg. principal component analysis)
- Missing values
  - complete case analysis
    - any missing value – entire observation ignored

Functional form of continuous predictors (linearity)

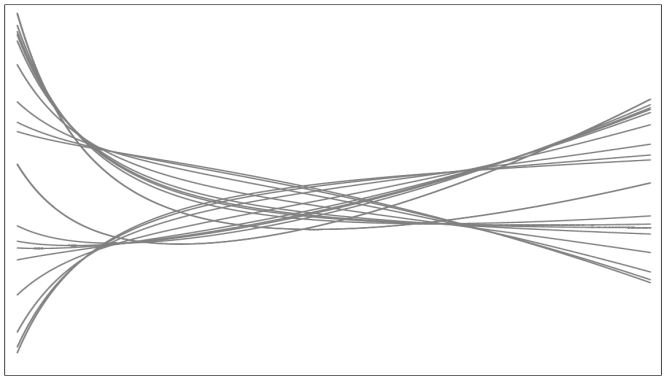
- Detecting non-linearity – in final model
  - plot residuals vs fitted values
    - simultaneous evaluation of all predictors
  - plot of residuals vs predictor
- Detecting non-linearity – before / during model building
  - scatter plot of outcome vs predictor
    - smoothing functions
- Smoothed scatter-plots
  - best fitting line through a mass of data (not confined to any specific shape)
  - local-influence property
    - position of line only affected by “neighbours”
    - # of neighbours determined by bandwidth (width of the neighbourhood)
    - adjust bandwidth to control degree of smoothing
  - different types of smoother
    - lowess – commonly used
  - cautions
    - potential to mask important local effects
    - behave poorly at ends

Detecting and correcting non-linearity

- Categorization of predictor (L2a)
  - indicator dummy variable
- Compare categorical and linear variables
  - linear model is a subset of a model with more categories (R)
  - test of linear regression against categorical model (F)
    - F-statistic to compare R vs F
- Transformation of X
  - box-cox analysis
    - same idea as for the outcome (L1a)
      - eg. boxcox ln_cf milk120k , **model(rhs)**
  - polynomial functions of X
    - quadratic, cubic, etc
    - fractional polynomials

Polynomial functions of X

- Quadratic (details L1b and L2b)
- Fractional polynomials
  - extension polynomial regression (L1a)
    - allow log, non-integer powers, repeated powers
  - select terms (usually one or two) of the form x^p
  - where "p" is from the set -2, -1, -0.5, 0, 0.5, 1, 2, 3
    - p=0 is taken to be ln(X)
    - eg $\beta_1 X^{-1}+\beta_2 X^2=\beta_1(1/X)+\beta_2(X^2)$
  - combination selected based on best fit (smallest log likelihood)
- Usually 2 power terms (2 degree) can fit most shapes
  - 2-degree FP: x(-2,2) … $x^{(-2)} + x^{(2)}$
- This graph shows some of the possibilities from a 2-degree FP



Example – (ln) calving to first service and milk120

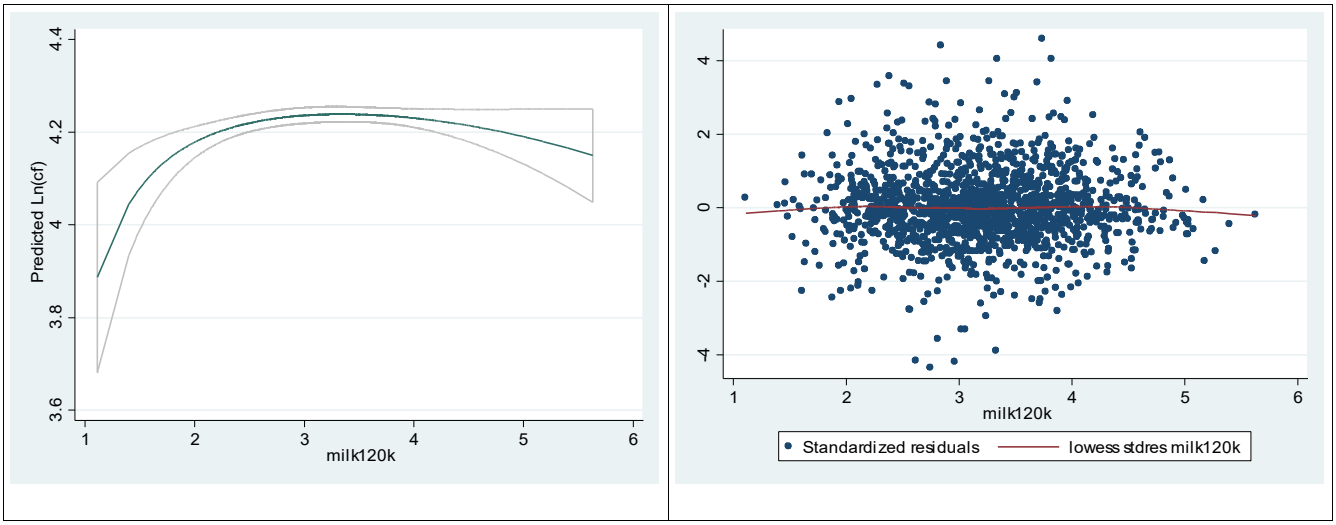
```
. fp <milk120k>,  scale center replace: reg ln_cf <milk120k>
(fitting 44 models)
(....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%)

Fractional polynomial comparisons:
------------------------------------------------------------------------------
    milk120k |   df    Deviance  Res. s.d.   Dev. dif.   P(*)   Powers
-------------+----------------------------------------------------------------
     omitted |    0    367.951      0.273     10.533    0.033
      linear |    1    366.159      0.273      8.741    0.033   1
       m = 1 |    2    361.396      0.273      3.978    0.138   -2
       m = 2 |    4    357.418      0.272      0.000       --   -2 3
------------------------------------------------------------------------------
(*) P = sig. level of model with m = 2 based on F with 1520 denominator dof.

      Source |       SS       df       MS              Number of obs =    1525
-------------+------------------------------           F(  2,  1522) =    5.27
       Model |  .782289941     2  .391144971           Prob > F      =  0.0052
    Residual |  112.870944  1522  .074159622           R-squared     =  0.0069
-------------+------------------------------           Adj R-squared =  0.0056
       Total |  113.653234  1524  .074575613           Root MSE      =  .27232

------------------------------------------------------------------------------
       ln_cf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  milk120k_1 |  -.5301266   .1654544    -3.20   0.001    -.8546694   -.2055839
  milk120k_2 |  -.0008516   .0004271    -1.99   0.046    -.0016894   -.0000138
       _cons |   4.238434    .008295   510.96   0.000     4.222163    4.254704
------------------------------------------------------------------------------
```

.



Interactions

- 2 way
  - all possible
  - significant main effects
  - significant unconditional assoc.
  - biologically meaningful
  - with key predictor of interest
- 3 way
  - rarely in epidemiology

Selection criteria

- Non-statistical
  - predictor of interest
  - known confounder
  - evidence of being a confounder
  - component of an interaction term
- Statistical - nested models
  - F-test for the predictor (or t-test)
  - Wald test or Likelihood ratio test (LRT)
  - always use these tests if appropriate

- Other procedures – statistical – non-nested models
  - adjusted $R^2 = 1 - \frac{MSE}{MST}$
    - R^2 adjusted for the # predictors, highest adjusted R^2 = best
    - linear regression models only
  - Mallow's Cp
    - linear regression only
    - $C_p = \frac{SSE}{\sigma^2} + 2k - n$
    - k = number of parameters (excluding intercept)
    - usually a positive value – might be negative if many predictors
    - lowest Cp = best
  - information criteria
    - AIC (Akaike's information criterion)
    - BIC (Bayesian information criterion)
    - not typically used in linear models
      - discussed in logistic regression

Selection strategies

- All possible / best subset
  - look at all possible combinations of predictors
  - select best model based on some criterion (such as adjusted R^2 or Mallow's C_P)
  - best subset – computer finds "best" model with 1,2,3, etc predictors

- Example: coleman.dta – 20 schools in USA
  - outcome: test score 6th graders
  - predictors: staff salary, ses, educ mothers, etc (see 1bl)
- Stata
  - add-on command: “vselect”

```
. vselect y_test_scr x1_staff_sal x2_father_job x3_ses x4_test_teach x5_edu_mother,
best

Response :             y_test_scr
Fixed Predictors :
Selected Predictors:   x3_ses x4_test_teach x1_staff_sal
x5_edu_mother                               x2_father_job

Actual Regressions   5
Possible Regressions 32

 Optimal Models Highlighted:
   # Preds     R2ADJ         C       AIC      AICC       BIC
         1  .8518292  4.974883  90.89429  149.1518  92.88576
         2  .8740953  2.832759  88.49432  147.9185  91.48152
         3  .8820943  2.836089  87.96903  149.0123  91.95196
         4  .8756399  4.670221  89.74417  152.9633  94.72284
         5  .8728444         6  90.80893  156.8998  96.78332
Selected Predictors

1  :  x3_ses
2  :  x3_ses x4_test_teach
3  :  x3_ses x4_test_teach x1_staff_sal
4  :  x3_ses x4_test_teach x1_staff_sal x5_edu_mother
5  :  x3_ses x4_test_teach x1_staff_sal x5_edu_mother x2_father_job
```

Stepwise estimation

- Stata: stepwise command
  - old syntax (eg. needs “xi” for indicator variables)
- Forward selection
  - start with a null model
  - adds terms based on statistical significance (one at a time, always choosing the most significant predictor not yet in the model)
  - stop when no more terms are significant when added

```
. stepwise, pe(0.1): reg y_test_scr x1_staff_sal x2_father_job x3_ses x4_test_teach
x5_edu_mother
                     begin with empty model
p = 0.0000 <  0.1000  adding  x3_ses
p = 0.0566 <  0.1000  adding  x4_test_teach

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   66.95
       Model |  570.497872     2  285.248936            Prob > F      =  0.0000
    Residual |  72.4264222    17  4.26037777            R-squared     =  0.8873
-------------+------------------------------           Adj R-squared =  0.8741
       Total |  642.924294    19  33.8381207            Root MSE      =  2.0641

------------------------------------------------------------------------------
   y_test_scr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       x3_ses |   .5415607   .0500441    10.82   0.000     .4359768    .6471445
x4_test_teach |   .7498915   .3666402     2.05   0.057    -.0236517    1.523435
        _cons |    14.5827   9.175409     1.59   0.130    -4.775723    33.94112
------------------------------------------------------------------------------
```

- Backward elimination
  - starts with a full model
  - eliminates terms that are not significant (one at a time, starting with the “least significant”)
  - stop once all terms remaining in the model are significant

```
. stepwise, pr(0.1): reg y_test_scr x1_staff_sal x2_father_job x3_ses x4_test_teach
x5_edu_mother
                      begin with full model
p = 0.4267 >= 0.1000  removing x2_father_job
p = 0.6863 >= 0.1000  removing x5_edu_mother
p = 0.1616 >= 0.1000  removing x1_staff_sal

 ....outpit ommitted

------------------------------------------------------------------------------
   y_test_scr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       x3_ses |   .5415607   .0500441    10.82   0.000     .4359768    .6471445
x4_test_teach |   .7498915   .3666402     2.05   0.057    -.0236517    1.523435
        _cons |    14.5827   9.175409     1.59   0.130    -4.775723    33.94112
------------------------------------------------------------------------------
```

- Stepwise
  - combines forward and backward
  - generally preferred approach is stepwise backward
    - starts with a full model and works backward using a stepwise approach

```
. stepwise, pe(0.1) pr(0.11): reg y_test_scr x1_staff_sal x2_father_job x3_ses
x4_test_teach x5_edu_mother
                      begin with full model
p = 0.4267 >= 0.1100  removing x2_father_job
p = 0.6863 >= 0.1100  removing x5_edu_mother
p = 0.1616 >= 0.1100  removing x1_staff_sal

...outut omitted
------------------------------------------------------------------------------
   y_test_scr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       x3_ses |   .5415607   .0500441    10.82   0.000     .4359768    .6471445
x4_test_teach |   .7498915   .3666402     2.05   0.057    -.0236517    1.523435
        _cons |    14.5827   9.175409     1.59   0.130    -4.775723    33.94112
```

- daisy2red dataset
  - create interaction terms for stepwise command
    - gen twdy=twin*dyst
    - gen rpvd=rp*vag_disch
  - reg wpc_sqrt parity1 aut_clv herd_size hs_sq dyst twin twdy rp vag_disch rpvd

```
*stepwise backward
stepwise, pe(0.05) pr(0.051) lockterm1: reg sqrt_wpc  parity1 aut_clv (hs_ct hs_sq)
      (dyst twin twdy) (rp vag_disch rpvd)
estimates store sw_1

stepwise, pe(0.05) pr(0.051) lockterm1: reg sqrt_wpc  parity1 aut_clv (hs_ct hs_sq)
      dyst twin twdy rp vag_disch rpvd
estimates store sw_2

stepwise, pe(0.05) pr(0.051) lockterm1: reg wpc_sqrt  parity1 (hs_ct hs_sq) aut_clv
      dyst twin twdy (rp vag_disch rpvd)
estimates store sw_3
```

```
. estimates table sw_1 sw_2 sw_3, star(0.10 0.05 0.001)
```

Variable	sw_1	sw_2	sw_3
parity1	.05586593	.04388077	.04720258
hs_ct	-.02346682**	-.02161068**	-.02240242**
hs_sq	.0000713***	.00006694***	.00006864***
aut_clv	-.51283075***	-.52440137***	-.51422168***
dyst	.62771805**		
twin	1.6982381**	1.4787911**	1.4063402**
twdy	-2.7727951		
rp	.39042354		.42079714
vag_disch	-.04220258		.07113917
rpvd	1.4760578**	1.8317287***	1.3840389**
_cons	3.0230207**	3.4048174**	3.2468497**

legend: * p<.1; ** p<.05; *** p<.001

Cautions with automated selection procedures

- Don't let your computer decide what variables go into your model
- R^2 and adjusted R^2 too high
- F-tests too large
- Severe problems if collinearity
- Ignores non-statistical considerations
  - exposures, confounders and intervening var.
- You need to take care of
  - dummy variables
  - interaction terms
  - polynomial terms
  - missing data
  - non significant confounders
- Useful when faced with large number of predictor variables
  - help to identify predictors that potentially are statistically significant associated with the outcome
- Perform residual and influential analysis of selected models

Evaluate reliability

- Validity
  - regression diagnostics
- Reliability
  - ability to predict future observations
  - split sample analysis (see VER)
  - leave-one-out analysis
    - fit the model using all data except one observation
    - generate prediction for the left-out obs.
    - compare the **P**redicted **R**esidual **S**um of **S**quares (PRESS) from predicted points to prediction error from full model
    - PRESS = sum{(residual$_i$ / (1-leverage$_i$))^2}
    - R^2(prediction) = 1-(PRESS/Total Sum of Squares)
      - compare R^2 (prediction) vs R^2 (full model)
      - difference should not be “too large”

- Example: daisy2red dataset (PRESS analysis)

```
. reg  wpc_sqrt c.hs_ct##c.hs_ct parity1 i.aut_clv i.twin i.dyst##i.vag_disch

      Source |       SS           df       MS      Number of obs   =     1,574
-------------+----------------------------------   F(8, 1565)      =     17.67
       Model |  1102.29212         8  137.786515   Prob > F        =    0.0000
    Residual |  12203.9747     1,565  7.79806691   R-squared       =    0.0828
-------------+----------------------------------   Adj R-squared   =    0.0782
       Total |  13306.2668     1,573  8.45916519   Root MSE        =    2.7925

------------------------------------------------------------------------------
    wpc_sqrt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hs_ct |   .0124288   .0012117    10.26   0.000      .010052    .0148056
             |
     c.hs_ct#|
     c.hs_ct |   .0000717   .0000174     4.13   0.000     .0000377    .0001058
             |
     parity1 |   .0624408     .04795     1.30   0.193    -.0316122    .1564939
   1.aut_clv |  -.5313957   .1419088    -3.74   0.000    -.8097471   -.2530443
             |
        twin |
        yes  |   1.484421   .5487805     2.70   0.007      .407999    2.560844
             |
        dyst |
        yes  |   .8159669   .3267812     2.50   0.013     .1749918    1.456942
             |
   vag_disch |
        yes  |   .9922199   .3503534     2.83   0.005     .3050084    1.679431
             |
        dyst#|
   vag_disch |
    yes#yes  |  -2.257598   .8832373    -2.56   0.011    -3.990051   -.5251447
             |
       _cons |   7.529392   .1445102    52.10   0.000     7.245938    7.812846
------------------------------------------------------------------------------
predict res, res

predict lev ,lev

gen eq1=(res/(1-lev))^2

. summ eq1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         eq1 |      1574    7.851331     10.9765   1.85e-07   83.92817

. di "PRESS =" r(sum) //error sum square from predicted points
PRESS =12357.995

. di "RSS full= "e(rss) //residuals sum square from full model
RSS full= 12203.975


. di "R2(pred) = "1-(PRESS/ TSS)
R2(pred) = .07126506

. di "R2 full model = " e(r2)
R2 full model = .08284007
```

Presenting the results

- Standardized coefficients
  - scales all coefficients so that they represent a change of 1 SD.

$$\beta^{*} = \beta(\sigma_x / \sigma_y)$$

  - compare the relative magnitude of several predictors
    - use with caution to compare coeff. from different studies
- Interquartile ranges (IQR)
  - how big is the change in the outcome if the predictor changes through out the IQR
    - IQR = diff. between 75th and 25th percentile
      - parity IQR = 3, coef. = 0.06
      - effect = 0.06*3 = 0.18 units
    - avoid impact of outlying observations
- Predictors eliminated from the model
  - don't ignore predictors just because they have been eliminated from the model
    - not statistically sig. ≠ no effect
    - unconditional associations?
    - force back in to the final model?
- Scale of results
  - dealing with transformed data
  - compute some expected effects of key predictors on the original scale at various levels of the other factors

A Structured Approach to Data Analysis (VER 30)

- "Jump to the finish"
  - start over
- Data collections sheets
- Data coding
- Data entry
- Keeping track of files
- Keeping track of variables
- Analyses
  - data editing
  - data verification
  - data processing - outcome variable
  - data processing - predictor variables
  - data processing - multilevel
  - unconditional associations
- Keeping track of analyses