

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

Index	Page
Model building strategies.....	2
Specifying the maximum model.....	3
Develop a causal model	3
Reducing the number of predictors.....	4
Functional form of continuous predictors (linearity).....	5
Detecting and correcting non-linearity	5
Selection criteria.....	9
Selection strategies: Stepwise estimation	10
Cautions with automated selection procedures.....	12
Evaluate reliability.....	13
Presenting the results.....	14
Strobe-Vet Reporting guidelines.....	15

- Today
 - ★ lecture
 - ★ review Q6 and Q7 (interaction) linear regression 2
 - ★ questions linear regression exercise 3
- Wednesday: Quiz #2
- Exercise for Friday: Model building exercise

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)

Develop a causal model

- Identify exposure
- Identify confounders
- Identify intervening variables
- Identify exposure-independent variables
- Make a decision about what to use in model

Reducing the number of predictors

- Screening – descriptive statistics
 - ★ keep predictors with few missing values
 - ➔ complete case analysis
 - any missing value – entire observation ignored
 - ★ substantial variability
 - ★ small number observations - re-categorize predictor
- Correlation / association
 - ★ understand relationships among predictors
 - ★ collinearity/confounding
- Unconditional associations
 - ★ liberal P-value
- Multivariate analysis
 - ★ factor analysis
 - ★ principal component analysis

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 (L1a)
 - ★ lowess – commonly used
 - ★ cautions
 - ➔ potential to mask important local effects
 - ➔ behave poorly at ends

Detecting and correcting non-linearity

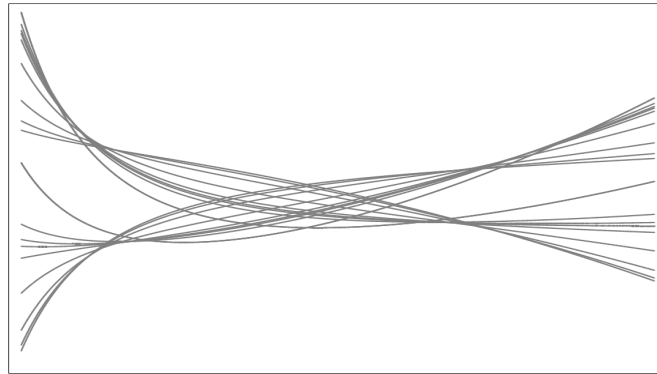
- Categorization of predictor (L2a)
 - ★ indicator dummy variable
- Compare categorical and linear variables
 - ★ linear model is a subset of a model with more categories (R)
 - ★ test of linear regression against categorical model (F)
 - ➔ F-statistic to compare R vs F

- 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: $x^{(-2,2)} \dots x^{(-2)} + x^{(2)}$

- 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%)
```

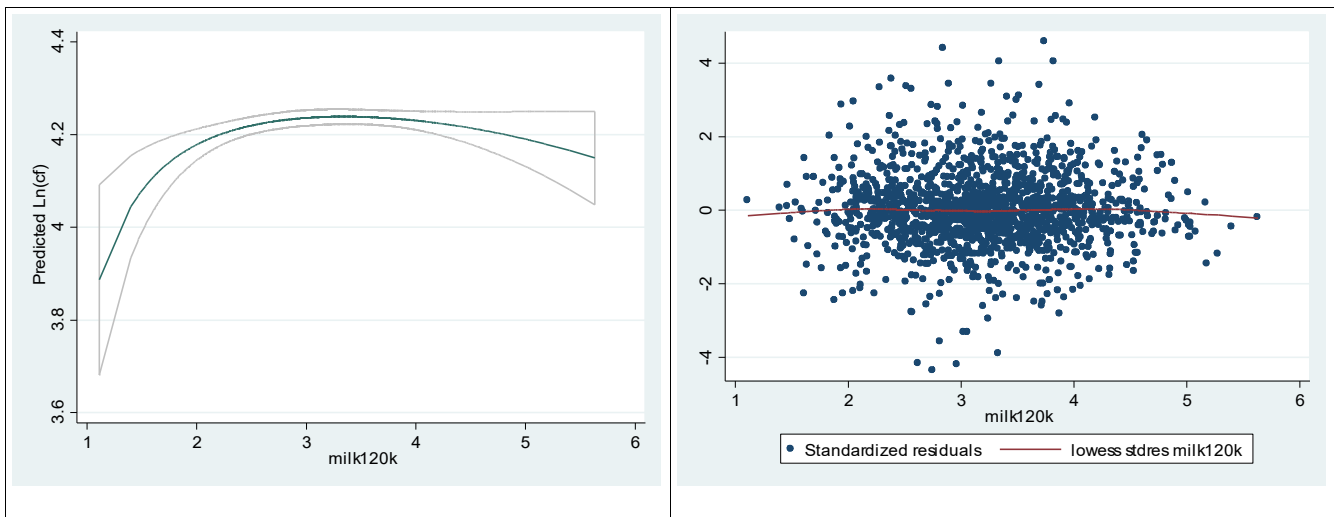
Fractional polynomial comparisons:

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
Model	.782289941	2	.391144971	F(2, 1522) =	5.27
Residual	112.870944	1522	.074159622	Prob > F =	0.0052
				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 used 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
- Statistical – non-nested models
 - ★ $\text{adjusted } R^2 = 1 - \frac{\text{MSE}}{\text{MST}}$
 - ➔ R^2 adjusted for the # predictors, highest adjusted $R^2 = \text{best}$
 - ➔ linear regression models only
- Information criteria
 - ★ AIC (Akaike's information criterion)
 - ★ BIC (Bayesian information criterion)
 - ★ not typically used in linear models (see logistic regression)

Selection strategies: Stepwise estimation

● 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
```

...output 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

● 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  
....output 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

● 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

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

```
stepwise, pe(0.05) pr(0.051) lockterm1: reg wpc_sqrt parity1 aut_clv  
(c.hs_ct##c.hs_ct) (i.dyst##i.twin) (i.rp##i.vag_disch)  
estimates store sw_1
```

```
stepwise, pe(0.05) pr(0.051) lockterm1: reg wpc_sqrt parity1 aut_clv  
(hs_ct hs_sq) i.dyst##i.twin i.rp##i.vag_disch  
estimates store sw_2
```

```
stepwise, pe(0.05) pr(0.051) lockterm1: reg wpc_sqrt parity1 (hs_ct  
hs_sq) aut_clv i.dys##i.twin (i.rp##i.vag_disch)  
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
aut_clv	-.51283075***	-.52440137***	-.51422168***
hs_ct	.01232803***	-.02161068**	-.02240242**
c.hs_ct#			
c.hs_ct	.0000713***		
dyst			
yes	.62771805**		
twin			
yes	1.6982381**	1.4787911**	1.4063402**
dyst#twin			
yes#yes	-2.7727951		
l.rp	.39042354		.42079714
vag_disch			
yes	-.04220258		.07113917
rp#vag_disch			
l#yes	1.4760578**	1.8317287***	1.3840389**
hs_sq		.00006694***	.00006864***
_cons	7.5152743***	3.4048174**	3.2468497**

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

Cautions with automated selection procedures

- R² and adjusted R² too high
- F-tests too large
- Ignores non-statistical considerations
 - ★ exposures, confounders and intervening var.
- You need to take care of
 - ★ dummy variables, interaction terms, polynomial terms
 - ★ 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
 - ★ 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
 - ➔ $\text{PRESS} = \sum \{(\text{residual}_i / (1 - \text{leverage}_i))^2\}$
 - ➔ $R^2(\text{prediction}) = 1 - (\text{PRESS} / \text{Total Sum of Squares})$
 - compare $R^2(\text{prediction})$ vs $R^2(\text{full model})$
 - 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
...output omitted...

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 "R2(pred) =" "1-(PRESS/ (e(mss) + e(rss)))
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 range (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. \neq 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

Strobe-Vet Reporting guidelines

- Strengthening the Reporting of Observational Studies in Epidemiology statement checklist for Veterinary medicine (the STROBE-Vet statement)
 - ★ <https://strobevvet-statement.org/strobe-vet-statement/>
- 22 items checklist for assessing reporting quality
- Items related to model building
 - ★ (7) Variables
 - ★ (8) Data source measurements
 - ★ (9) Bias
 - ★ (11) Quantitative variables
 - ★ (12) Statistical methods
- Explanation document
 - ★ expanded document describing examples of reporting