

## Lecture 3a: Model building II

<b>Index</b>	<b>Page</b>
Model building strategies.....	2
Specifying the maximum model.....	2
Reducing the number of predictors.....	3
Functional form of continuous predictors (linearity).....	4
Detecting and correcting non-linearity .....	5
Interactions.....	8
Selection criteria.....	9
Selection strategies.....	10
Stepwise estimation .....	12
Cautions with automated selection procedures.....	16
Presenting the results.....	20
A Structured Approach to Data Analysis (VER 30).....	21

- Today
  - ★ review Quiz 2
  - ★ L3a and Margins command (HS)
- Datasets
  - ★ daisy2red.dta
  - ★ coleman.dta

## Model building strategies

- parsimony vs fit
- goals
  - ★ prediction
  - ★ estimates of effects
- **steps**
  - ★ specify maximum model
  - ★ 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
- 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

- 10 obs. per parameter (excl. intercept)
- screening - descriptive statistics
  - ★ few missing values
  - ★ substantial variability
  - ★ small categories
- correlation / association
  - ★ pairwise
- unconditional associations
  - ★ liberal P-value
- multivariate analysis (eg princ. comp.)
- missing values
  - ★ complete case analysis
    - ➔ any missing value - entire observation ignored
  - ★ more on missing data - discussed later

## Functional form of continuous predictors (linearity)

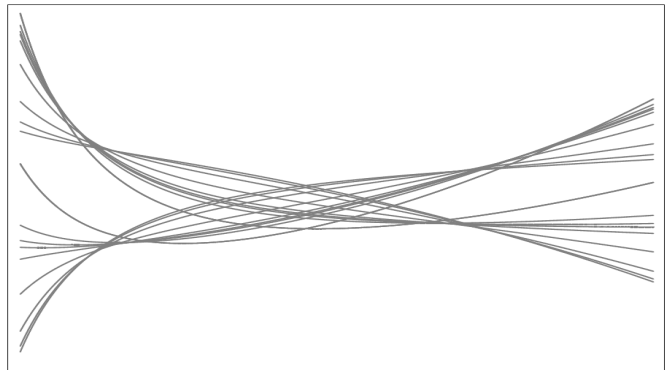
- detecting non-linearity - in final model
  - ★ plot residuals vs fitted values
    - simultaneous evaluation of all predictors
  - ★ plot of residuals vs predictor
- detecting non-linearity - before / during model building
  - ★ scatter plot of outcome vs predictor
    - smoothing functions
- smoothed 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 neighborhood)
    - 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
  - ★ eg. parity (L1b)
- 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
    - ➔ 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%)
```

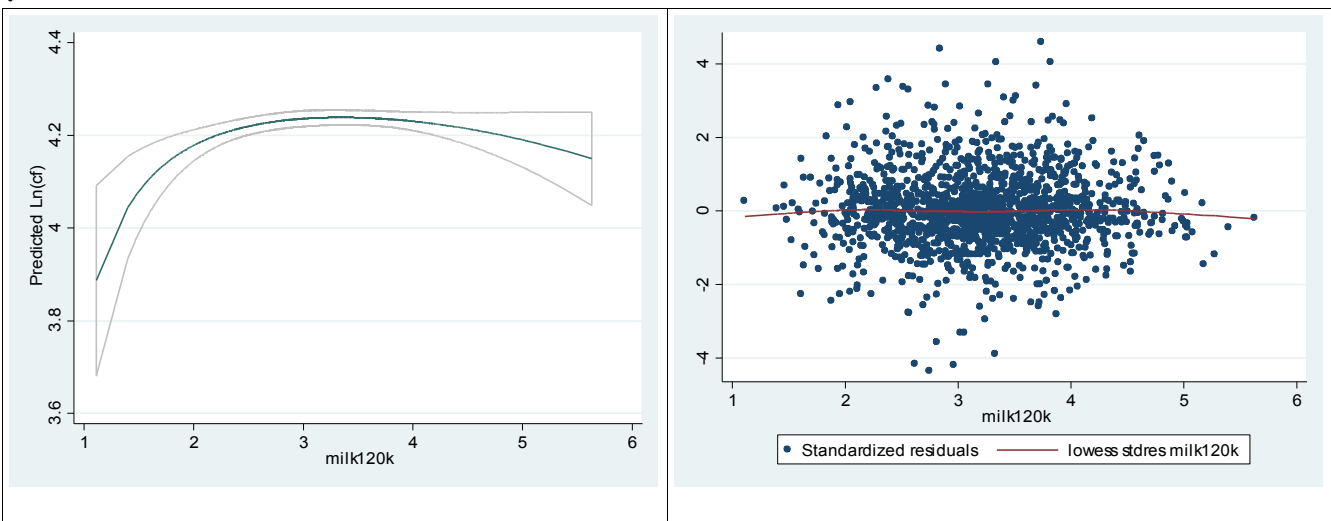
Fractional polynomial comparisons:

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

## 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
  - ★ 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
    - linear regression models only

★ Mallow's Cp

→ linear regression only

$$\rightarrow C_p = \frac{\text{RSS}}{\sigma^2} + 2k - n$$

→ 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 CP)

★ best subset - computer finds "best" model with 1, 2, 3, etc predictors

- Example: coleman.dta - 20 schools in USA
  - ★ outcome: test score 6<sup>th</sup> graders
  - ★ predictors: staff salary, ses, educ mothers, etc (see 1b1)
- 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	<b>2.832759</b>	88.49432	<b>147.9185</b>	<b>91.48152</b>
3	<b>.8820943</b>	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

- ★ stepwise command in Stata
- ★ 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
Model	570.497872	2	285.248936	F( 2, 17) =	66.95
Residual	72.4264222	17	4.26037777	Prob > F =	0.0000
				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
```

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

- 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
```

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

● daisy2red dataset

★ 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
estimates table sw_1 sw_2
```

```
. estimates table sw_1 sw_2
```

Variable	sw_1	sw_2
parity1	.05586593	.04388077
aut_clv	-.51283075	-.52440137
hs_ct	-.02346682	-.02161068
hs_sq	.0000713	.00006694
dyst	.62771805	
twin	1.6982381	1.4787911
twdy	-2.7727951	
rp	.39042354	
vag_disch	-.04220258	
rpvd	1.4760578	1.8317287
_cons	3.0230207	3.4048174

## 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
  - ★ missing data
- 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
    - ➔ split data into two parts (eg. 50:50)
    - ➔ build model using one part (1<sup>st</sup> model), generate predictions and compare them with observed values for the other part (2<sup>nd</sup> model)
    - ➔ cross-validation correlation
      - corr. predicted and obs. values 2<sup>nd</sup> mod.
    - ➔ shrinkage on cross-validation
      - difference  $R^2$  from the 1<sup>st</sup> model and the square of the cross-validation correlation
      - guideline:  $<0.1$  is Ok
        - but will depend on the scale

★ 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
- ➔  $PRESS = \sum\{(\text{residual}_i / (1 - \text{leverage}_i))^2\}$
- ➔ compare  $R^2$  (prediction) vs  $R^2$  (full model)
  - eg.  $R^2(\text{prediction}) = (1 - PRESS) / \text{Total Error}$

● Example: daisy2red dataset

★ split-sample analysis

```
. gen rand=uniform()
```

```
. reg wpc_sqrt c.hs_ct##c.hs_ct parity1 aut_calv twin i.dyst##vag_disch if
rand<0.6, vsquish
```

Source	SS	df	MS	Number of obs = 920			
Model	807.033629	8	100.879204	F( 8, 911)	=	12.41	
Residual	7405.23534	911	8.12868863	Prob > F	=	0.0000	
				R-squared	=	0.0983	
				Adj R-squared	=	0.0904	
Total	8212.26897	919	8.93609246	Root MSE	=	2.8511	

wpc_sqrt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hs_ct	.0134746	.0016229	8.30	0.000	.0102897	.0166596
c.hs_ct#c.hs_ct	.0000853	.0000237	3.60	0.000	.0000389	.0001318
parity1	.064454	.0638652	1.01	0.313	-.0608859	.189794
aut_calv	-.5758069	.1899446	-3.03	0.003	-.9485868	-.2030269
twin	1.597395	.8017033	1.99	0.047	.0239953	3.170795
dyst						
yes	1.512701	.4427911	3.42	0.001	.6436914	2.38171
vag_disch						
yes	1.358874	.4449721	3.05	0.002	.4855847	2.232164
dyst#vag_disch						
yes#yes	-4.092153	1.13334	-3.61	0.000	-6.316413	-1.867892
_cons	7.454233	.1914211	38.94	0.000	7.078556	7.829911

```

. predict pv
(option xb assumed; fitted values)

. *cross validation correlation
. corr pv wpc_sqrt if rand>=0.6
(obs=654)

-----+-----
          |          pv wpc_sqrt
-----+-----
          |          1.0000
pv        |          0.2409   1.0000
wpc_sqrt |
-----+-----

. *shrinkage on cross-validation
. scalar cv_sq = r(rho)^2 /* this computes r2 - r(rho) is a "saved result" from
                        -corr- */

. di "R2 first model= " r2_1 " and CV2= " cv_sq
R2 first model= .0982717 and CV2= .05804935

. di "shrinkage on cross-validation = " r2_1-cv_sq
shrinkage on cross-validation = .04022235

```

## ● leave-one-out analysis (PRESS analysis)

```

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

. di "R2(pred) = "1-(r(sum)/ (`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.
$$\beta^* = \beta(\sigma_x / \sigma_y) \text{ or } \beta / \sigma_x$$
  - ★ 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 75<sup>th</sup> and 25<sup>th</sup> 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