

Variable selection – a (p)review

Georg Heinze and Daniela Dunkler for TG2

Why a (p)review

- A review:
what is the current practice of variable selection in medical research?
- A preview:
what should change?

Current practice of variable selection

Table 1 Variable selection methods used in major epidemiologic journals in 2008

Selection technique	American Journal of Epidemiology		Epidemiology		European Journal of Epidemiology		International Journal of Epidemiology	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Prior knowledge	50	29	11	28	13	30	9	20
Effect estimate change	31	18	6	15	3	7	4	9
Stepwise selection	27	16	9	23	10	23	13	29
Modern methods (shrinkage, penalized regression)	0	0	0	0	0	0	0	0
Other (e.g., principal components, propensity scores)	2	1	4	10	1	2	2	4
Not described	61	36	10	25	17	39	17	38
Total	171		40		44		45	

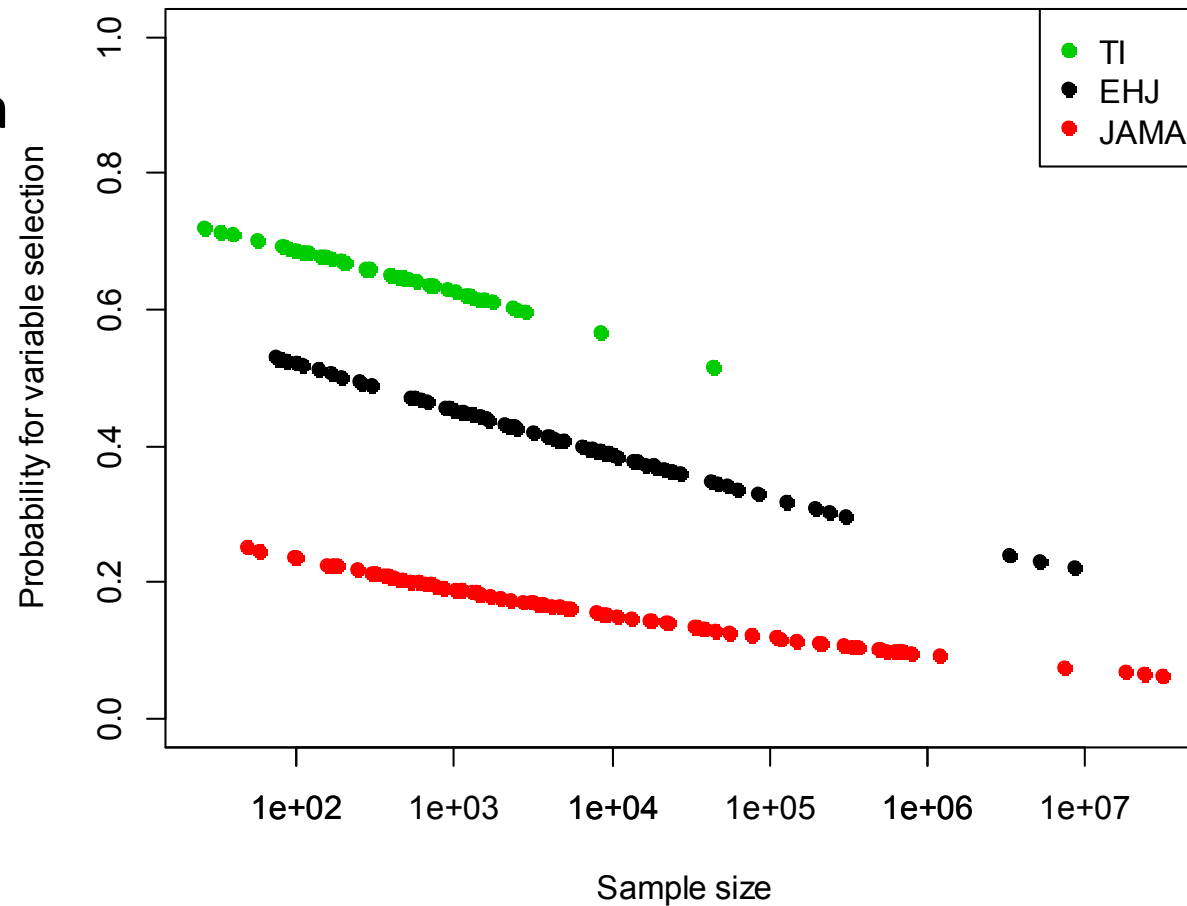
Walter & Tiemeier, EurJEpi 2009 24:733-736

Current practice of variable selection

Variable	JAMA Internal Medicine (IF=14.00)	European Heart Journal (IF=15.05)	Transplant International (IF=2.84)
A. Original articles 2015	137	132	89
B. Multivariable models	94	75	49
C. Variable selection (% of B)	17%	37%	65%
Univariate selection (% of B)	5%	21%	39%
Stepwise methods (% of B)	13%	23%	33%
Univariate filtering, then stepwise selection (% of B)	3%	8%	6%
Stability evaluation	0	0	0
Median sample size (in B)	4,396	4,319	295

Current practice of variable selection

- Modeling the probability for variable selection by journal and sample size:



The 5 myths about variable selection

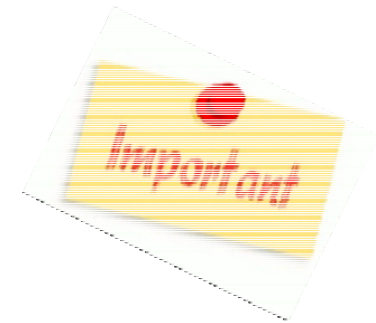
1. The number of variables in a model should be reduced until there are 10 events per variable.
2. Only variables with proven univariable-model significance should be included in a multivariable model.
3. Non-significant effects should be eliminated from a model.
4. P-value quantifies type I error.
5. Variable selection simplifies analysis.

➔ Probably because of these myths univariate selection is so popular.

Interpretation of regression coefficients

- Linear model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K + \epsilon$$



- Adjusted effect of X_k :
- Expected change in outcome, if X_k changes by 1 unit and all other X 's stay constant.
- β_k measures the 'independent' effect of X_k .
- Fundamentally different in different models!

Interpretation of regression coefficients

- Consider the following models to explain %body fat:

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	76.65092	9.97648	7.68	<.0001
height_cm	Height in cm	1	-0.58611	0.06204	-9.45	<.0001
weight_kg	Weight in kg	1	0.58177	0.03368	17.28	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-14.89166	2.76160	-5.39	<.0001
weight_kg	Weight in kg	1	0.41950	0.03371	12.44	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-30.36370	11.43150	-2.66	0.0084
abdomen	Abdomen circumference	1	0.91008	0.07137	12.75	<.0001
weight_kg	Weight in kg	1	-0.21541	0.06778	-3.18	0.0017
height_cm	Height in cm	1	-0.09593	0.06171	-1.55	0.1213

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-47.65873	2.63417	-18.09	<.0001
abdomen	Abdomen circumference	1	0.97919	0.05599	17.49	<.0001
weight_kg	Weight in kg	1	-0.29219	0.04655	-6.28	<.0001

Provided information versus desired knowledge

- Information provided by the data:
 - Number of independent observations N
 - Number of events E
(logistic: $\min(\#events, \#non-events)$, Cox: $\#events$)
- Amount of knowledge desired:
 - Number of unknown regression coefficients (K)
- Summarized by 'events per variable' $EPV = E/K$, $NPV = N/K$.
- Often cited minimum $EPV = 10$.
 - Harrell 2015, p. 72, actually recommends $EPV=15$ (with no variable selection!)
 - Schumacher et al, 2012, recommend $EPV=10$ to 25

Events Per Variable (EPV)

- But $EPV = 10$ (or $EPV = 15$) refers to
 - Number of candidate variables, not variables in the final model.
 - Should be considered as a lower bound!
- Additionally,
 - Non-linearity, interactions, etc. $\rightarrow EPV \uparrow$.
 - Prediction $\rightarrow EPV \uparrow$ (logistic regression EPV 20–50).
 - Modern modeling techniques (e.g. random forests, neural networks, support vector machines) \rightarrow 10 times EPV compared to logistic regression $\rightarrow EPV \uparrow\uparrow$
(van der Ploeg et al. 2014).

Basic variable selection algorithms

- 'Full' model
- Univariable filtering
- Best subset selection
- Forward selection
- Backward elimination
- Information-theoretic approach
- Directed acyclic graph (DAG)-based selection

The 'full' model

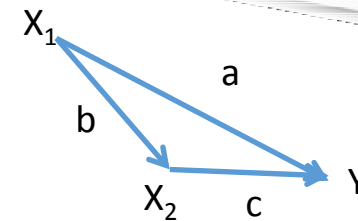
- Means: do not perform any data-driven variable selection.
- Variables should be pre-selected by 'expertise'.
- Select, for each variable, a desired level of non-linearity (including spline transformations).
- Select some biologically plausible interactions.

Univariable filtering

- Still a popular variable selection method in medical literature!
- Select a significance level α (e.g., $\alpha=0.20$ or $\alpha=0.157$)
- Perform K univariable models.
- Use all variables in multivariable model with univariable p -value $< \alpha$.
- Sometimes accompanied by subsequent backward elimination.

Pros and cons of univariate selection

- Easy. (You can do that with any software.)
- Retractable.
- Problematic (see also Sun et al, JClinEpi 1996):
- The univariate effect of X_1 on Y is $a + bc$.



a	b	c	Consequence
Pos.	Pos.	Neg.	X_1 falsely not selected (if $a = -bc$)
0	Pos./Neg.	Pos./Neg.	X_1 falsely selected.
Pos./neg	0	Pos./neg	X_1 correctly selected (only if $b = 0$ or $c = 0$).

➔ Univariate selection works only with uncorrelated variables.

Best subset selection

- Perform all 2^K regressions.
- Select the model that has the lowest AIC.

Modification (information-theoretic approach):

- Pre-specify a small number (4 – 20) of plausible models.
- Select those that have $AIC < AIC_{\min} + 2$.
- Perform multi-model inference on the selected models.
(Burnham & Anderson, 2002)

In practice:

- Approximated by stepwise approaches!

Backward elimination

- Select a significance level α_2 .
- Estimate full model.
- Repeat:
 - While least significant term has $p \geq \alpha_2$, remove it and re-estimate.

Variant: Stepwise backward

- Select α_1 and α_2 .
- Repeat:
 - While least significant term has $p \geq \alpha_2$, remove it and re-estimate.
 - If most significant excluded term has $p < \alpha_1$, add it and re-estimate.

Software: R <code>mfp:mfp()</code>

Forward selection

- Select a significance level α_1 .
- ‘Estimate’ a null model.
- Repeat:
 - While the most significant excluded term has $p < \alpha_1$, add it and re-estimate.

Variant: Stepwise forward

- Select α_1 and α_2 .
- Repeat:
 - While the most significant excluded term has $p < \alpha_1$, add it and re-estimate.
 - If least significant included term has $p \geq \alpha_2$, remove it and re-estimate.

Software: SAS/PROC GLMSELECT R <code>step()</code>
--

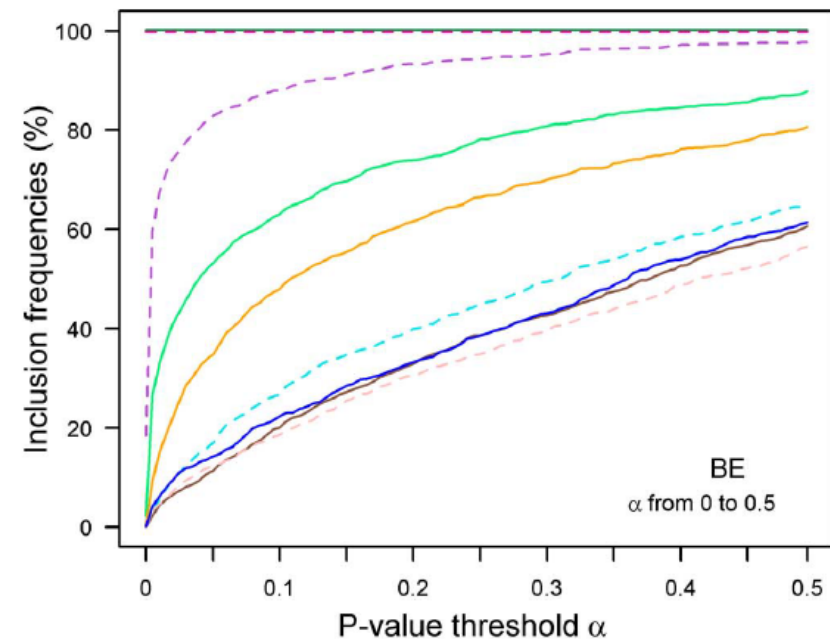
Consequences of variable selection

- Variable selection typically leads to:
 - Conditional bias away from 0
 - unconditional bias towards 0
 - Biased conditional inference (p -values too low – CI too narrow)
 - These problems vanish asymptotically (but not yet with $EPV = 10$)
 - Univariate selection: usually the worst of the algorithmic approaches, and not consistent.

- A tool is needed to check for selection stability.

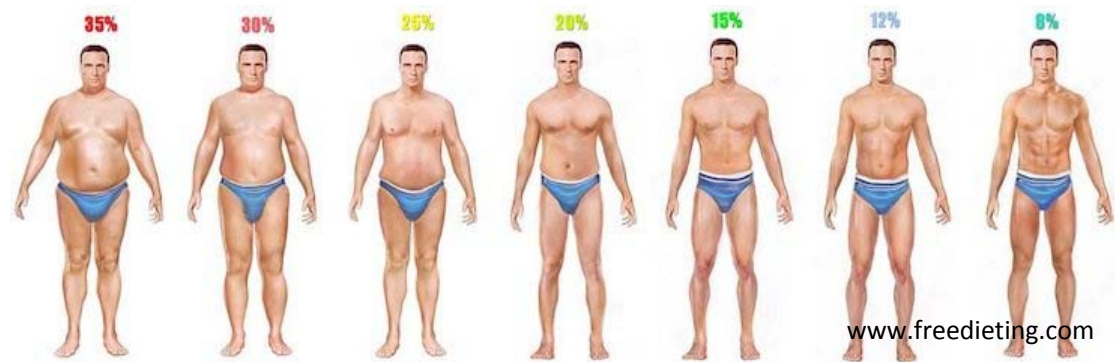
Quantification of model uncertainty

- Perform bootstrap analysis, repeating variable selection in each resample.
- Evaluate bootstrap inclusion frequencies (BIF) of variables (easy).
- Pairwise inclusion tables (easy).
(Sauerbrei & Schumacher, 1992)
- Evaluate bootstrap model selection frequencies (moderate).
- Evaluate stability paths (plot BIF vs. α) (intensive).



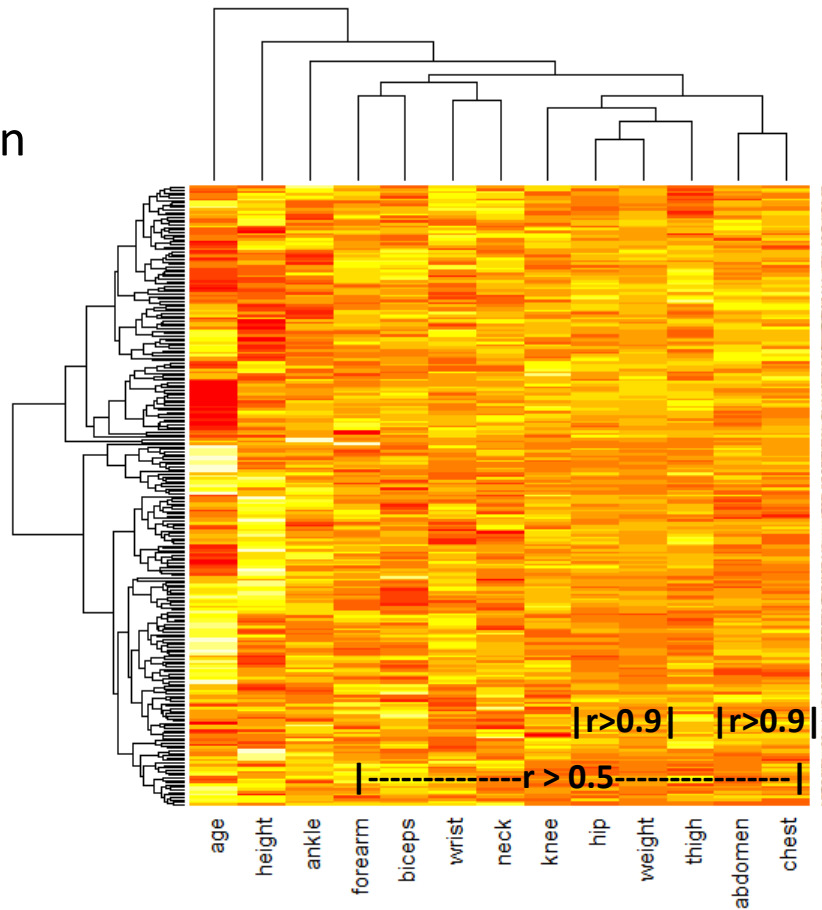
Case study: body fat approximation

- Johnson's (1996) body fat data example
- Publicly available: <http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html>
- 251 males aged 21 to 81
- Response variable: %body fat (Siri formula), based on costly underwater density measurement
- Predictors: age, height, weight, +10 circumference measures
- First goal: approximation of %body fat



Case study: correlation of predictors

Correlations between predictor variables are quite high:



Case study: selection by backward(AIC) - SAS code

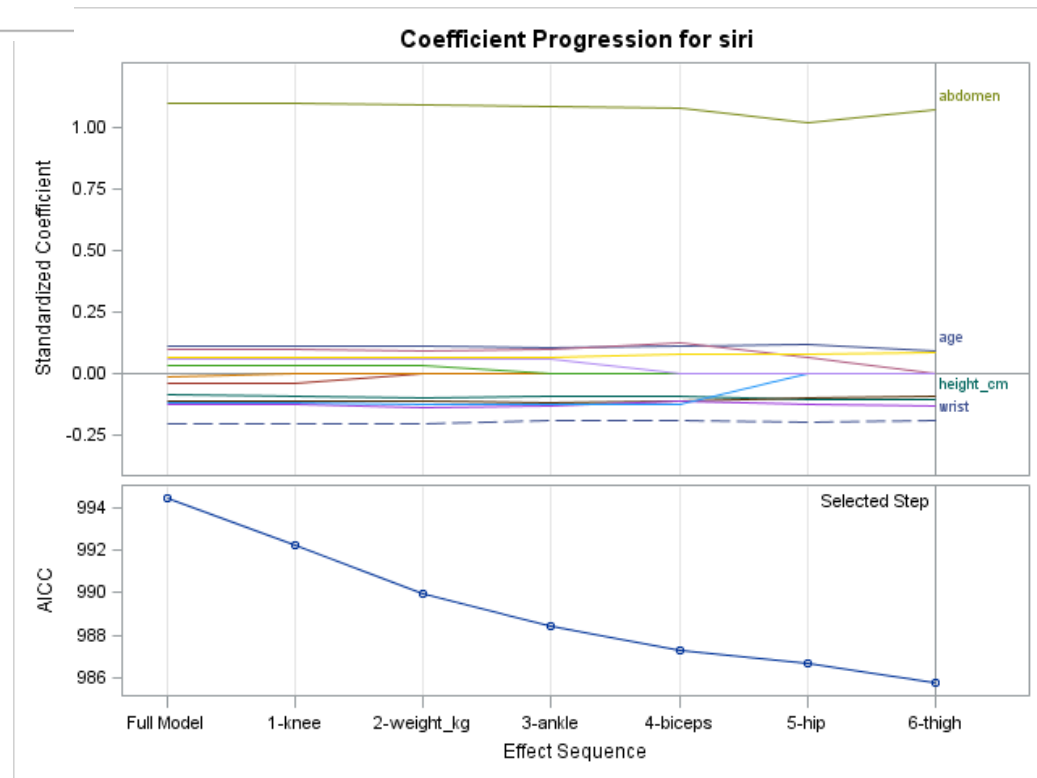
```
proc glmselect data=cas1.bodyfat plots=all;  
  model siri=age weight_kg height_cm neck chest  
        abdomen hip thigh knee ankle biceps forearm wrist  
        /selection=backward select=aicc details=step;  
run;
```

Case study: selection by backward(AIC) - results

```
proc glmselect data=case1.bodyfat plots=all;
  model siri=age weight_kg height_cm neck chest
    abdomen hip thigh knee ankle biceps forearm wrist
    /selection=backward select=aicc details=step;
run;
```

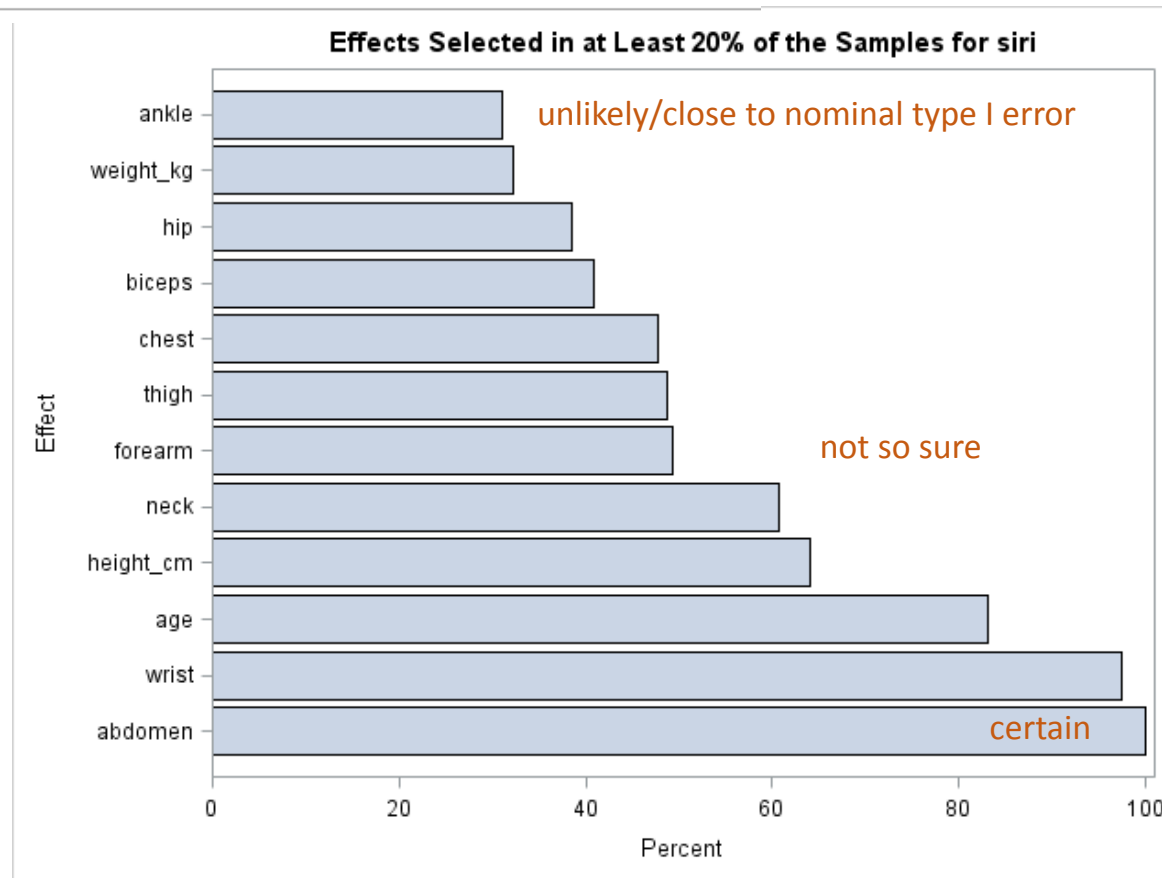
R-Square	0.7488
Adj R-Sq	0.7416
AIC	985.02609
AICC	985.77298
SBC	760.22971

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	5.945152	8.149537	0.73
age	1	0.060301	0.024738	2.44
height_cm	1	-0.129879	0.047052	-2.76
neck	1	-0.329725	0.218693	-1.51
chest	1	-0.135123	0.087549	-1.54
abdomen	1	0.874948	0.064762	13.51
forearm	1	0.364969	0.191709	1.90
wrist	1	-1.729208	0.482605	-3.58



Case study: bootstrap inclusion frequencies (BIFs)

```
proc glmselect data=case1.bodyfat plots=all;  
  model siri=age weight_kg height_cm neck chest  
    abdomen hip thigh knee ankle biceps forearm wrist  
    /selection=backward select=aicc ;  
  modelaverage nsamples=1000 ;  
run;
```



Effects Selected in at Least 20% of the Samples	
Effect	Selection Percentage
age	83.20
weight_kg	32.30
height_cm	64.10
neck	60.80
chest	47.80
abdomen	100.0
hip	38.60
thigh	48.70
ankle	31.00
biceps	40.90
forearm	49.40
wrist	97.50

Case study: pairwise inclusion frequencies

```

proc surveysselect data = casel.bodyfat
  out = bootfat seed = 7123981
  method = urs samprate = 1 outhits rep = 1000;
run;

proc reg data=bootfat noprint outest=estboot;
  by replicate;
  model siri=age weight_kg height_cm neck chest
    abdomen hip thigh knee ankle biceps forearm wrist
    /selection=backward slstay=0.157;
run;

data estboot;
  set estboot;
  sel_age=age ne .;
  sel_weight=weight_kg ne .;
  sel_height=height_cm ne .;
  sel_neck=neck ne .;
  sel_chest=chest ne .;
  sel_abdomen=abdomen ne .;
  sel_hip=hip ne .;
  sel_thigh=thigh ne .;
  sel_knee=knee ne .;
  sel_ankle=ankle ne .;
  sel_biceps=biceps ne .;
  sel_forearm=forearm ne .;
  sel_wrist=wrist ne .;
run;

proc freq data=estboot;
  tables sel_height*sel_weight sel_thigh*sel_biceps;
run;

```

Competitive selection!

Frequency Percent Row Pct Col Pct		Table of sel_height by sel_weight		
		sel_weight		
sel_height		0	1	Total
0		122 12.28 34.76 18.37	229 22.90 65.24 68.15	351 35.10
1		542 54.20 83.51 81.63	107 10.70 16.49 31.85	649 64.90
Total		664 66.40	336 33.60	1000 100.00

Frequency Percent Row Pct Col Pct		Table of sel_thigh by sel_biceps		
		sel_biceps		
sel_thigh		0	1	Total
0		218 21.80 41.44 37.91	308 30.80 58.56 72.47	526 52.60
1		357 35.70 75.32 62.09	117 11.70 24.68 27.53	474 47.40
Total		575 57.50	425 42.50	1000 100.00

Case study: bootstrap model selection frequencies

Model Selection Frequency				
Times Selected	Selection Percentage	Number of Effects	Frequency Score	Effects in Model
23	2.30	7	23.76	Intercept age height_cm chest abdomen biceps wrist
19	1.90	7	19.79	Intercept age height_cm neck abdomen forearm wrist
18	1.80	7	18.78	Intercept age height_cm neck abdomen biceps wrist
15	1.50	8	15.74	Intercept age height_cm neck chest abdomen biceps wrist
14	1.40	9	14.71	Intercept age height_cm neck abdomen hip thigh forearm wrist
14	1.40	10	14.69	Intercept age height_cm neck chest abdomen hip thigh forearm wrist
13	1.30	7	13.77	Intercept age height_cm chest abdomen forearm wrist
12	1.20	7	12.73	Intercept age weight_kg abdomen thigh forearm wrist
12	1.20	9	12.70	Intercept age height_cm neck chest abdomen ankle forearm wrist
11	1.10	8	11.75	Intercept age height_cm neck abdomen thigh forearm wrist
11	1.10	9	11.70	Intercept age height_cm neck abdomen hip thigh biceps wrist
10	1.00	8	10.72	Intercept age neck abdomen hip thigh forearm wrist
9	0.90	8	9.75	Intercept age height_cm neck chest abdomen forearm wrist
9	0.90	8	9.74	Intercept age height_cm neck abdomen hip thigh wrist
9	0.90	9	9.72	Intercept age height_cm neck chest abdomen biceps forearm wrist
9	0.90	8	9.71	Intercept age weight_kg neck abdomen thigh forearm wrist
9	0.90	8	9.71	Intercept age neck abdomen hip thigh biceps wrist
9	0.90	8	9.71	Intercept age height_cm chest abdomen ankle biceps wrist
9	0.90	10	9.67	Intercept age height_cm neck chest abdomen ankle biceps forearm wrist
8	0.80	6	8.84	Intercept age height_cm neck abdomen wrist


Extremely low selection proportions:
Very unstable selection!

Preselection of variables

- **Prior subject matter knowledge**
- Chronology
- Confounder criteria
- Availability at time of model use
- Quality (measurement errors)
- Costs of collecting measurements

- Availability in data set (missing values)
- Variability (rare categories)

- Preselection = Bayes!



Discussion
between
researcher and
statistician!

Prior knowledge: simple illustrative simulations

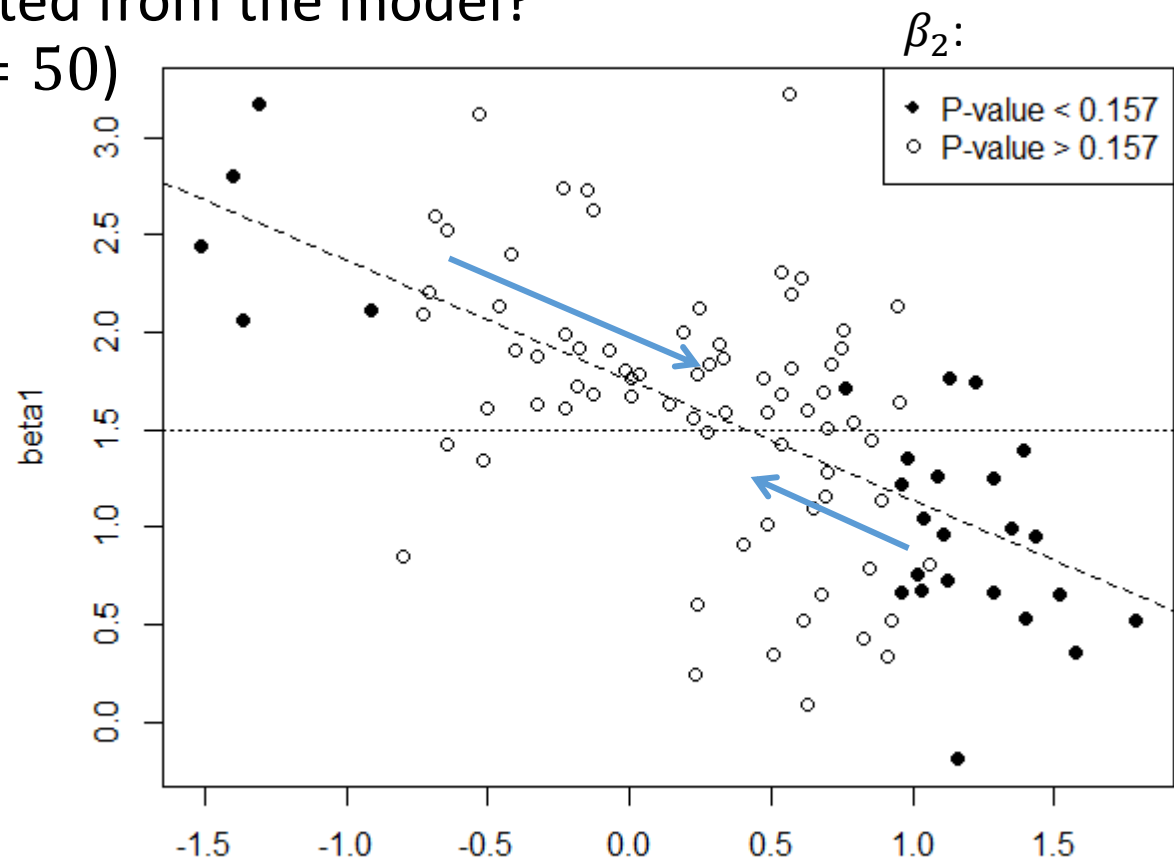
- Should X_2 be eliminated from the model?
(simulation with $N = 50$)

True $\beta_1 = 1.5, \beta_2 = 0.3$

A weak β_2 :

Setting it to 0 will more often push $\hat{\beta}_1$ towards its true value than away from it.

→ Shrinkage effect on $\hat{\beta}_1$!



→ 'Selection is good.'

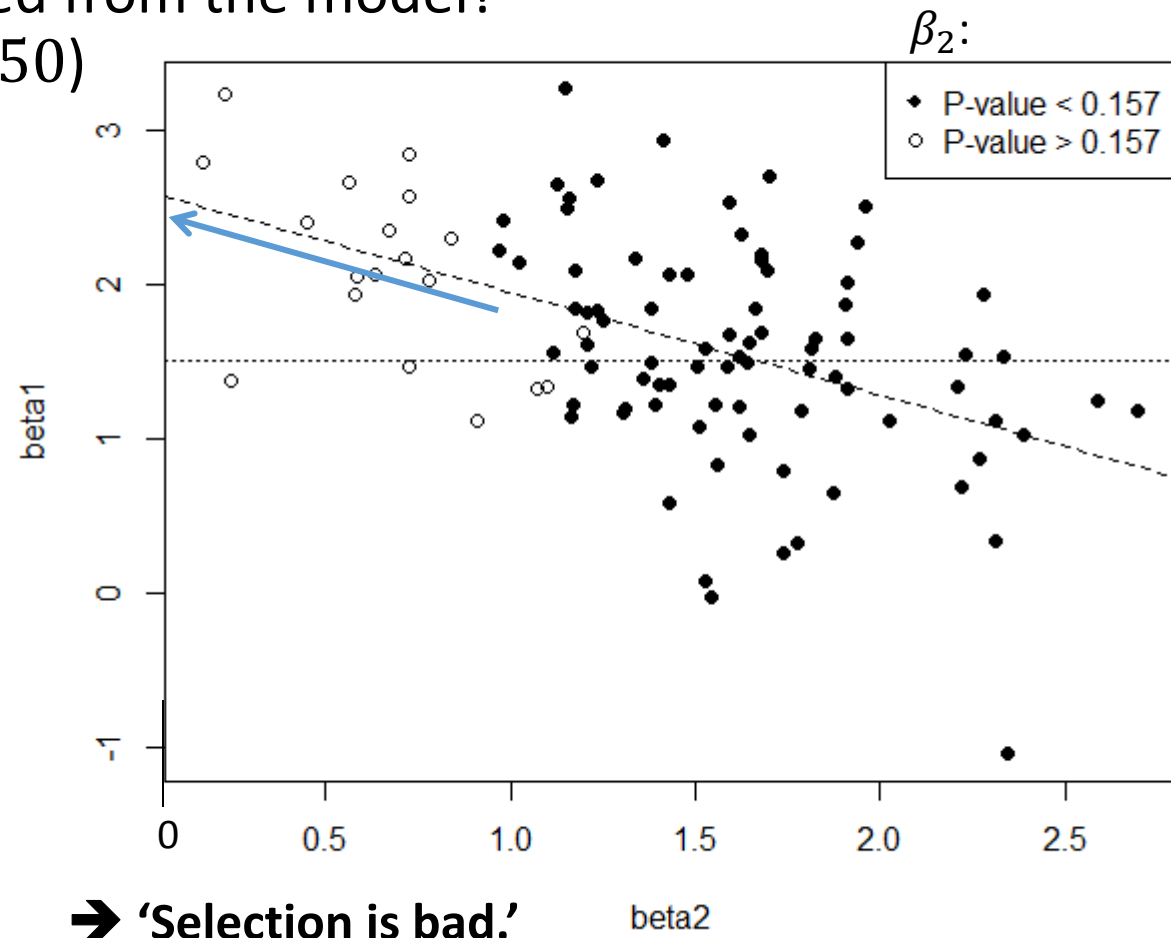
Prior knowledge: simple illustrative simulations

- Should X_2 be eliminated from the model?
(simulation with $N = 50$)

True $\beta_1 = 1.5, \beta_2 = 1.5$

A strong β_2 :

Setting it to 0 will
always push $\hat{\beta}_1$ away
from its true value.



The 5 myths: and what should change

1. The number of variables in a model should be reduced until there are 10 events per variable.

Resp: No, there should be $\gg 10$ events per candidate variable.

2. Only variables with proven univariable-model significance should be included in a multivariable model.

Resp: No, univariable-model significance can be strongly misleading as criterion for inclusion in a multivariable model.

3. Non-significant effects should be eliminated from a model.

Resp: No, non-significant effects do not harm a model.

4. P-value quantifies type I error.

Resp: No, P-values after model selection are almost impossible to estimate.

5. Variable selection simplifies analysis.

Resp: No, stability investigations are needed and must become part of routine software output.

References

- **Full tutorial ‘Variable selection for statistical models: a review and recommendations for the practicing statistician’ with additional references:**
<http://tinyurl.com/variable-selection-talk>
- Harrell Jr. FE. *Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. Second edition.* Springer: New York, 2015.
- van der Ploeg T, Austin P, C., Steyerberg E, W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* 2014; **14**: 137.
- Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer, 2002.
- Johnson RW. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education* 1996; **4**. <http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html>
- Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine* 1992; **11**: 2093-2109
- Schumacher M, Holländer N, Schwarzer G, Binder H, Sauerbrei W. Prognostic Factor Studies. In: Crowley J, Hoering A (eds.), *Handbook of Statistics in Clinical Oncology*, 3rd ed., CRC press: Boca Raton, 2012.
- Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology* 1996; **49**: 907-916.