Georg HEINZE

Medical University of Vienna Center for Medical Statistics, Informatics and Intelligent Systems. Section for Clinical Biometrics Vienna, Austria

Variable selection a review and recommendations for the practicing statistician

janvier 2019

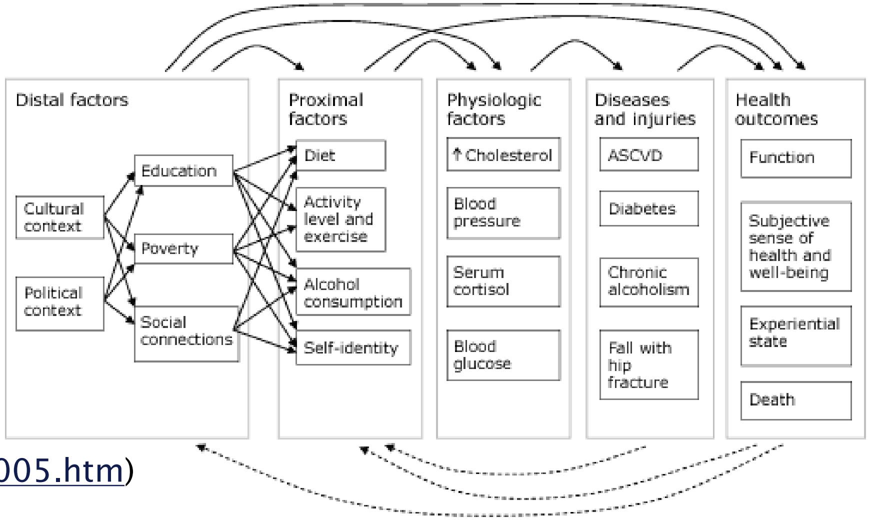
Variable selection – a review and recommendations for the practicing statistician

Georg Heinze, Christine Wallisch, Daniela Dunkler

Medical University of Vienna

Why multivariable modeling?

- Statistical models are useful tools…
- Disease causation is usually multifactorial.
- Influential variables can only be identified in a multivariable context.



(from http://www.cdc.gov/pcd/issues/2010/jul/10_0005.htm)

What do we mean by a statistical model?

- *A set of probability distributions on the sample space $\mathcal{S}$.*
  (e.g. Cox and Hinkley, 1974)

- *Statistical models summarize patterns of the data available for analysis.*
  (Steyerberg, 2009)

- *A powerful tool for developing and testing theories by way of causal explanation, prediction, and description.*
  (Shmueli, 2010)

- *A simplification or approximation of reality.*
  (Burnham, Anderson, 2002)

- *A model represents, often in considerably idealized form, the data-generating process.* (Wikipedia)

Is there such thing as a true model?

A ‘true model’ = a ‘true data generating mechanism’.

Pro:

- Aristotle: ‘*Nature operates in the shortest way possible.*’
- Newton: ‘*We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.*’

Contra:

- ‘*We do not accept the notion that there is a simple “true model” in the biological sciences.’* (Burnham & Anderson, 2002)
- ‘*We recognize that true models do not exist… A model will only reflect underlying patterns, and hence should not be confused with reality.*’ (Steyerberg, 2009)
- ‘*I started reading Annals of Statistics, and was bemused: Every article started with „Assume that the data are generated by the following model: …“ followed by mathematics exploring inference, hypothesis testing and asymptotics.*‘ (Breiman, 2001)
- ‘*All models are wrong, but some are useful.’* (Box)

What do **we** mean by a statistical model?

- *Statistical models are simple mathematical rules derived from empirical data describing the association between an outcome and several explanatory variables.*
  (Dunkler et al, 2014)
- They should be valid, practically useful, robust.
- *‘Simplicity is the ultimate sophistication.’* (Leonardo da Vinci)
- *‘Everything should be made as simple as possible, but not simpler.’*
  (~Einstein)

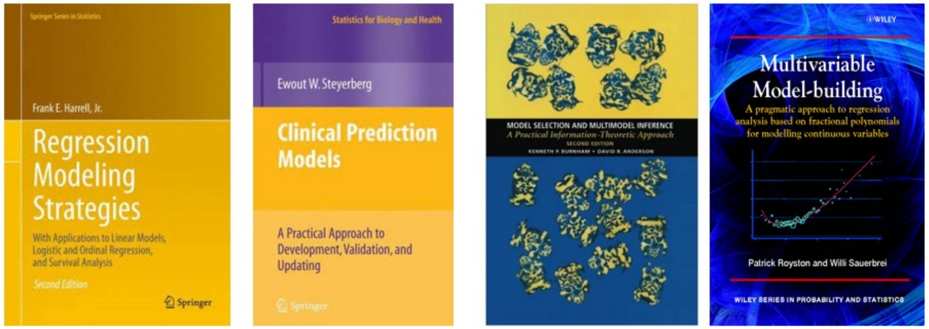
Complexity is your enemy. Any fool can make something complicated.
It is hard to keep hings simple.

Sir Richard Branson
founder of Virgin Group

Ockham? yes but it‘s hard to be simple

- Ockham‘s razor is often used to justify ‚simpler models‘
- However, in search of simpler models, statistical analysis gets more complex!
  - Model instability
  - Multiple equally likely competing models
  - Post-selection inference …

Variable selection

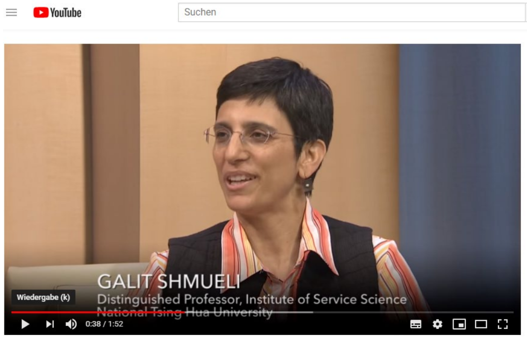


Springer Series in Statistics
Frank E. Harrell, Jr.
Regression Modeling Strategies
With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis
Second Edition
Springer
Statistics for Biology and Health
Ewout W. Steyerberg
Clinical Prediction Models
A Practical Approach to Development, Validation, and Updating
Springer
MODEL SELECTION AND MULTIMODEL INFERENCE
A Practical Information-Theoretic Approach
SECOND EDITION
KENNETH P. BURNHAM • DAVID R. ANDERSON
WILEY
Multivariable Model-building
A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables
Patrick Royston and Willi Sauerbrei
WILEY SERIES IN PROBABILITY AND STATISTICS

(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

To Explain or to Predict?

Shmueli, 2010


YouTube
Suchen
GALIT SHMUELI
Distinguished Professor, Institute of Service Science
National Tsing Hua University
Wiedergabe (k)
0:38 / 1:52

Galit Shmueli discusses the distinction between explaining and predicting (Preview)

- **Explanatory models**
  - Strong theory ➔ interest in coefficients and inference.
  - Testing and comparing existing causal theories.
- **Predictive models**
  - Interest in accurate predictions of future observations.
  - No concern about causality and confounding (association).
- **Descriptive models**
  - capture the data structure parsimoniously: which factors affect the outcome and how?

- $\text{expected prediction error} = \text{irreducible error} + \text{bias}^2 + \text{variance}$

Hastie et al 2009, p.223

What models do we typically see?

Linear model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_k + \epsilon = X\beta + \qquad \epsilon \sim N(0, \sigma)$

Logistic model $\Pr(Y = 1) \quad = \text{expit}(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_k)$

$$= \exp(X\beta)\,/[1 + \exp(X\beta)]$$

Cox model $h(X, t) = h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_K X_k) = h_0(t)\exp(X\beta)$

Linearity: linear combination of variables

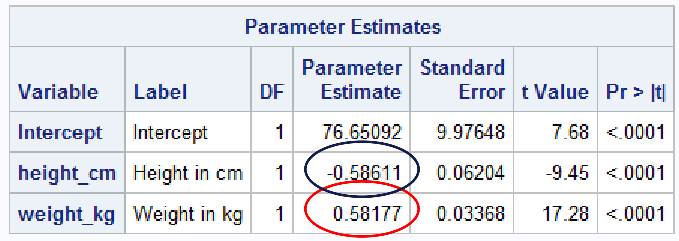
- (Relaxation: splines, fractional polynomials, GAMs)

Additivity: sum of effects

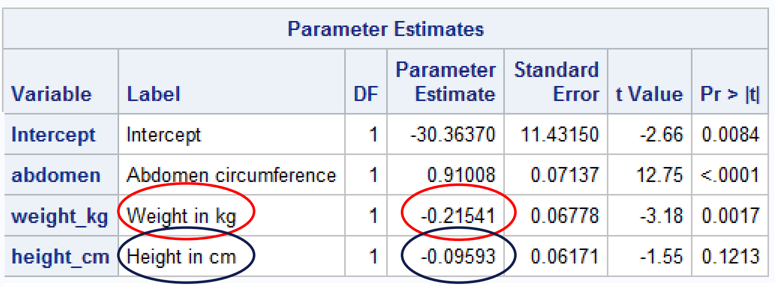
- (Relaxation: include interactions, power functions, etc.)

Interpretation of regression coefficients

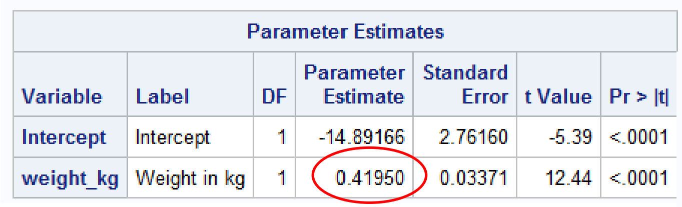
- Consider the following models to explain %body fat:



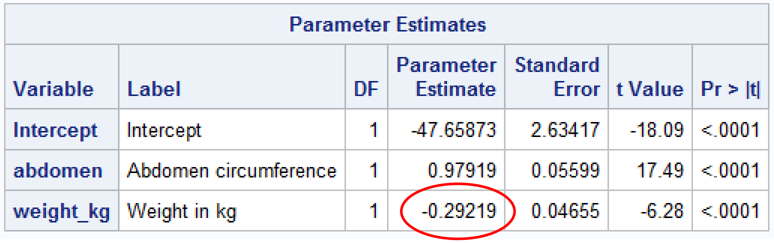
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	76.65092	9.97648	7.68	<.0001
height_cm	Height in cm	1	-0.58611	0.06204	-9.45	<.0001
weight_kg	Weight in kg	1	0.58177	0.03368	17.28	<.0001



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-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	-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	-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

Sample size and events per variable (EPV)

- EPV = effective sample size / number of variables
- Logistic, Cox regression:
  effective sample size = number of less frequent outcomes, events
- EPV ≥ 15 (Harrell 2015, p. 72)
  - Number of candidate variables, not variables in the final model.
  - Should be considered as lower bound!
- Rough guide, but many other quantities important
  - Courvoisier et al 2011, van Smeden et al 2016
- When considering variable selection:
  EPV = effective sample size / number of candidate variables !!!

Significance criteria and stepwise procedures

- Consider the nested models:

$$M_1: \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$
$$M_2: \quad Y = \gamma_0 + \gamma_1 X_1 \qquad\qquad + \epsilon$$

- Null hypothesis $\beta_2 = 0$ implies that $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1$

Likelihood ratio test	fit both M_1 and M_2	Model comparison
Step-up approximation: score test	fit only M_2	Forward selection
Step-down approximation: Wald test	fit only M_1	Backward elimination

- With many X_j‘s, iterated testing could lead to stepwise selection of variables
- Are these iterated tests reliable?
  - Unaccounted multiple testing!
  - Testing if β_j is relevant *given the current set of adjustment variables*

Information criteria

- Approximate the ‘cross-validated’ expectation of $\log L$

$$E_{test}E_{train}[\log L(x_{test}|\hat{\beta}_{train})]$$

- by

$$\text{AIC} = \log L(x_{train}|\hat{\beta}_{train}) - K$$

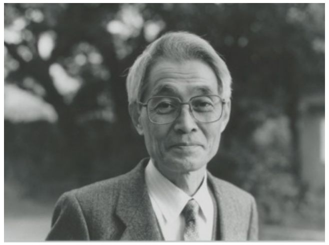
Model developed on x_{train},
Evaluated on x_{test}.

Model developed on x_{train},
Evaluated on x_{train}.

K … number of parameters

Degrees of freedom difference	AIC-equivalent p-value in LR test
1	0.157
2	0.135
3	0.117
4	0.092

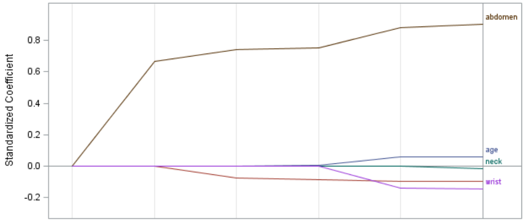
General: `1-pchisq(2*df, df)`



Hirotumi Akaike, 1927-2009,
(from http://andrewgelman.com)

- $\text{BIC} = \log L(x_{train}|\hat{\beta}_{train}) - \log(n)K/2$ =>more stringent!

Penalized likelihood: regularized regression



- LASSO: minimize $\log\mathrm{L}(\beta) - \lambda\sum|\beta_j|$
- Imposes a penalty on the regression coefficients.
- Prerequisite: adequate standardization of effects.

- What we obtain
  - A prediction formula with less error than ordinary least squares,
  - Variable selection.
- What we do not obtain
  - Unbiased regression coefficients,

  New developments for inference: Taylor and Tibshirani, 2015
  - Independence from transformations in X

Variable selection algorithms

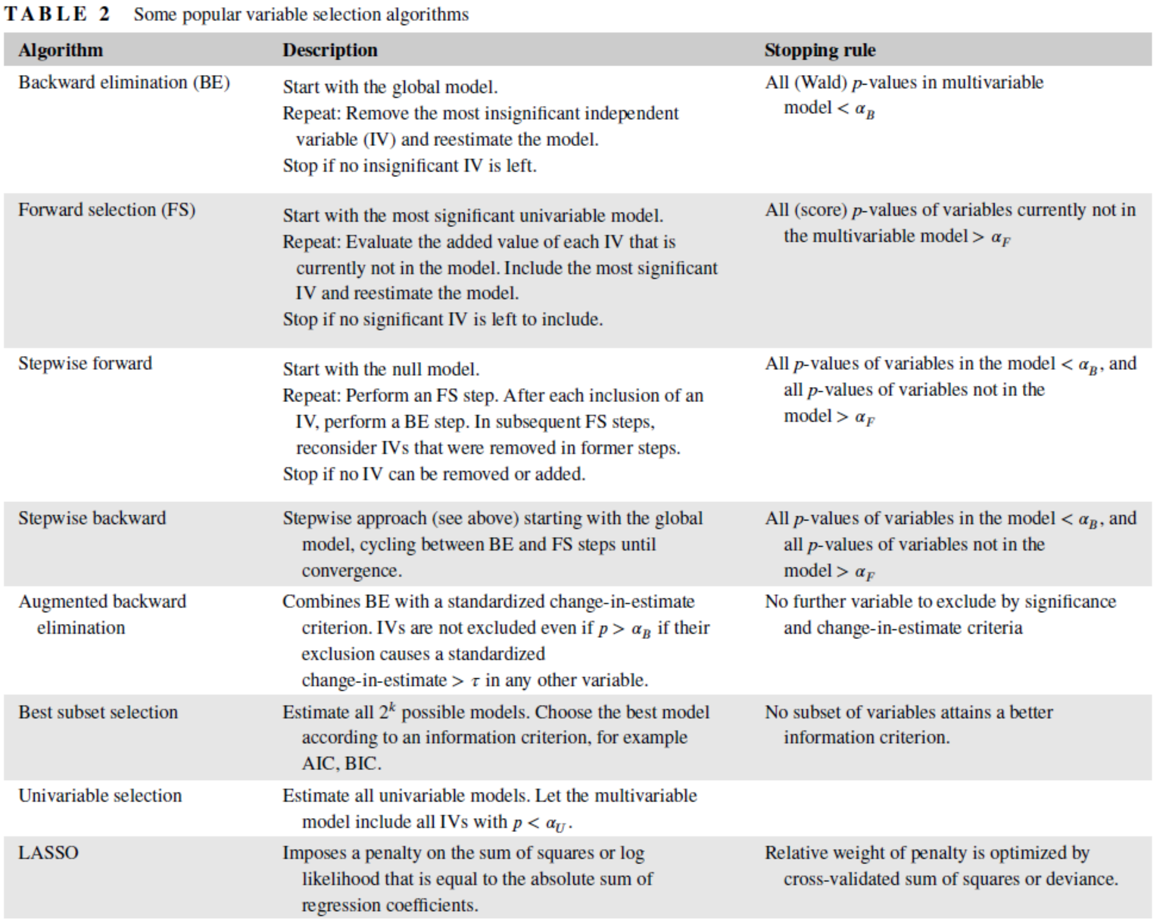
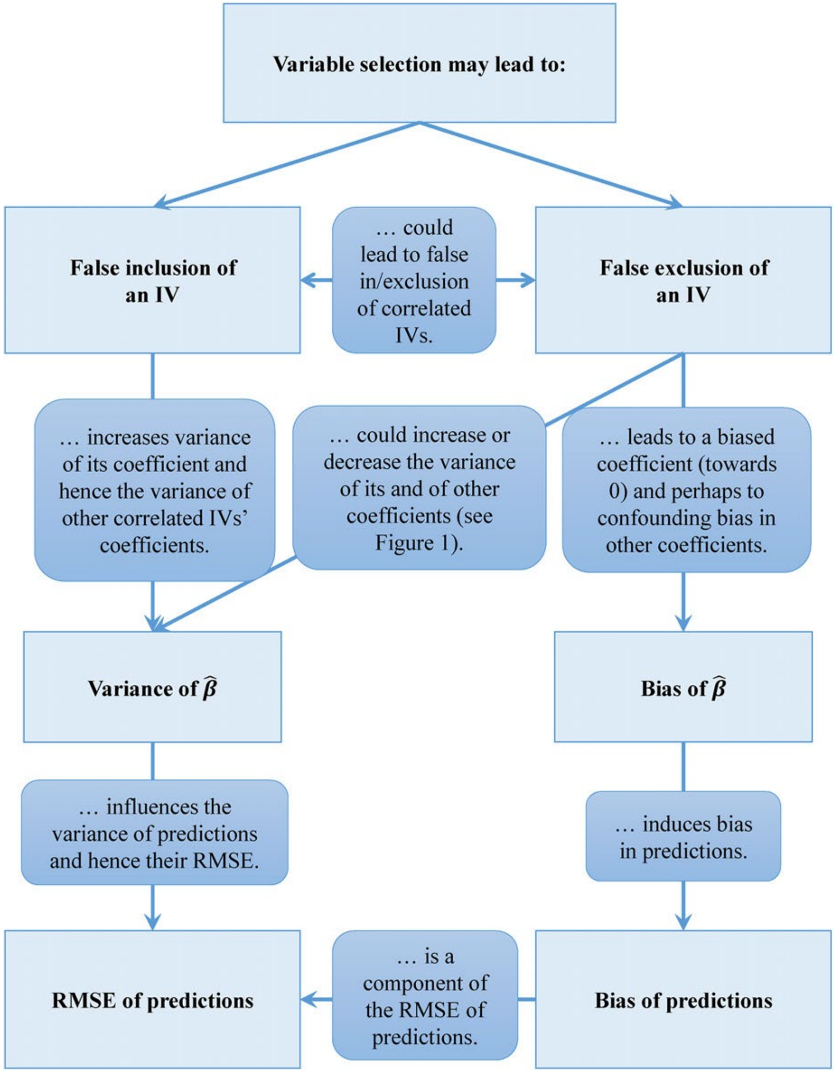


TABLE 2 Some popular variable selection algorithms

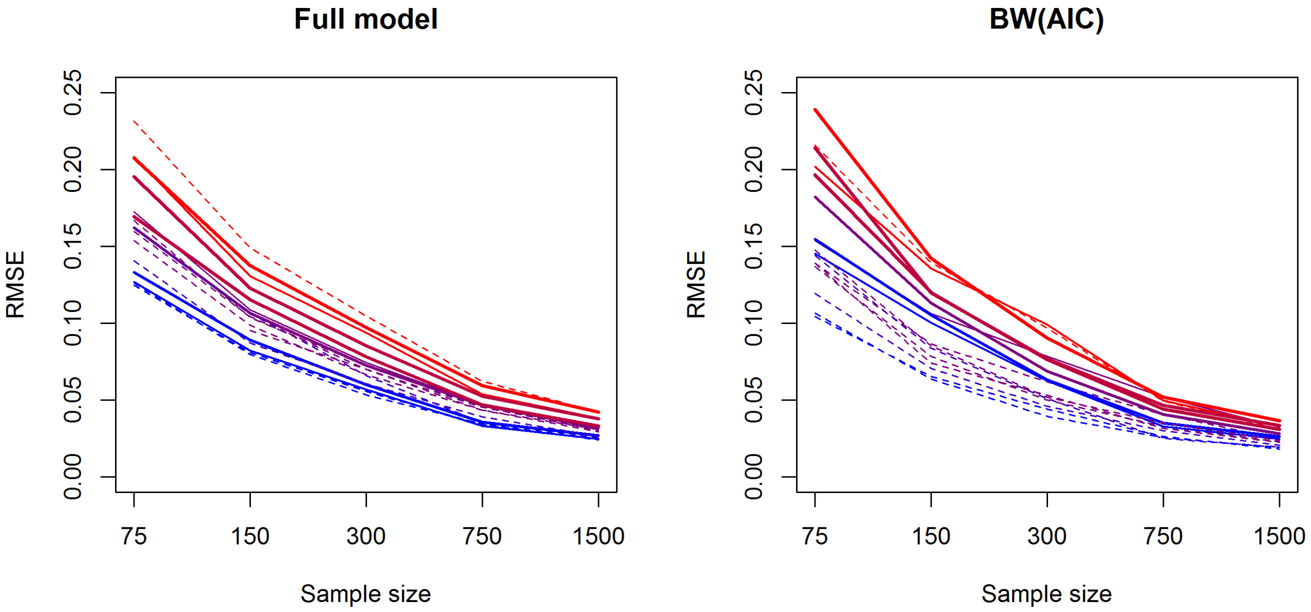
Algorithm	Description	Stopping rule
Backward elimination (BE)	Start with the global model. Repeat: Remove the most insignificant independent variable (IV) and reestimate the model. Stop if no insignificant IV is left.	All (Wald) p-values in multivariable model $< \alpha_B$
Forward selection (FS)	Start with the most significant univariable model. Repeat: Evaluate the added value of each IV that is currently not in the model. Include the most significant IV and reestimate the model. Stop if no significant IV is left to include.	All (score) p-values of variables currently not in the multivariable model $> \alpha_F$
Stepwise forward	Start with the null model. Repeat: Perform an FS step. After each inclusion of an IV, perform a BE step. In subsequent FS steps, reconsider IVs that were removed in former steps. Stop if no IV can be removed or added.	All p-values of variables in the model $< \alpha_B$, and all p-values of variables not in the model $> \alpha_F$
Stepwise backward	Stepwise approach (see above) starting with the global model, cycling between BE and FS steps until convergence.	All p-values of variables in the model $< \alpha_B$, and all p-values of variables not in the model $> \alpha_F$
Augmented backward elimination	Combines BE with a standardized change-in-estimate criterion. IVs are not excluded even if $p > \alpha_B$ if their exclusion causes a standardized change-in-estimate $> \tau$ in any other variable.	No further variable to exclude by significance and change-in-estimate criteria
Best subset selection	Estimate all 2^k possible models. Choose the best model according to an information criterion, for example AIC, BIC.	No subset of variables attains a better information criterion.
Univariable selection	Estimate all univariable models. Let the multivariable model include all IVs with $p < \alpha_U$.	
LASSO	Imposes a penalty on the sum of squares or log likelihood that is equal to the absolute sum of regression coefficients.	Relative weight of penalty is optimized by cross-validated sum of squares or deviance.

Consequences of variable selection

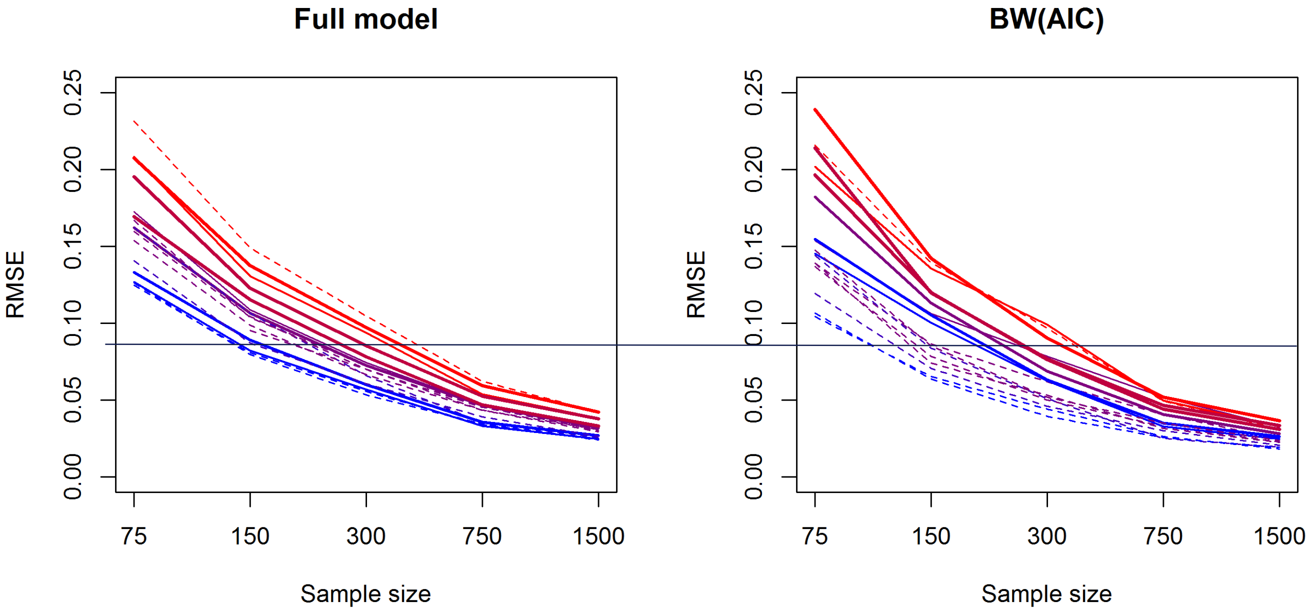
- **FIGURE 2** A schematic network of dependencies arising from variable selection. β, regression coefficient; IV, independent variable; RMSE, root mean squared error


Variable selection may lead to:
False inclusion of an IV
… could lead to false in/exclusion of correlated IVs.
False exclusion of an IV
… increases variance of its coefficient and hence the variance of other correlated IVs' coefficients.
… could increase or decrease the variance of its and of other coefficients (see Figure 1).
… leads to a biased coefficient (towards 0) and perhaps to confounding bias in other coefficients.
Variance of $\hat{\beta}$
Bias of $\hat{\beta}$
… influences the variance of predictions and hence their RMSE.
… induces bias in predictions.
RMSE of predictions
… is a component of the RMSE of predictions.
Bias of predictions

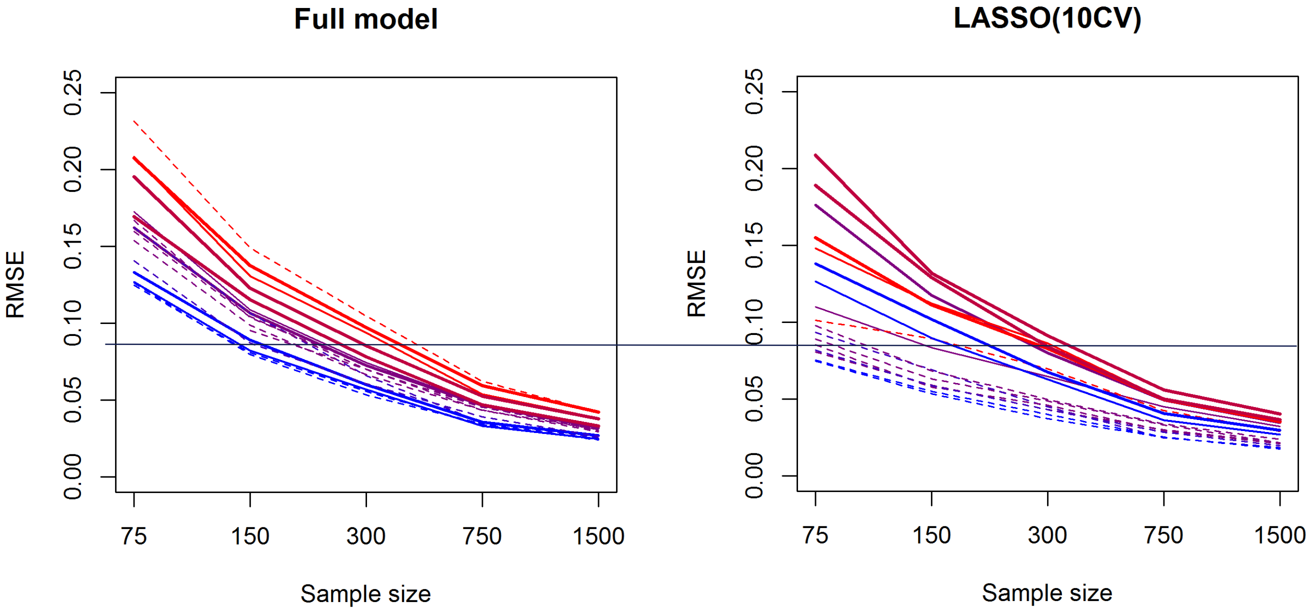
RMSE of regression coefficients, unconditional
simulation with 15 covariates



RMSE of regression coefficients, unconditional
simulation with 15 covariates

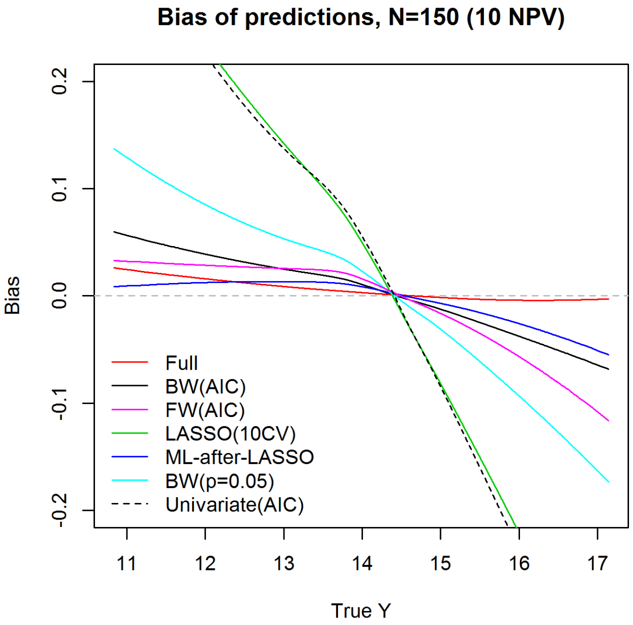


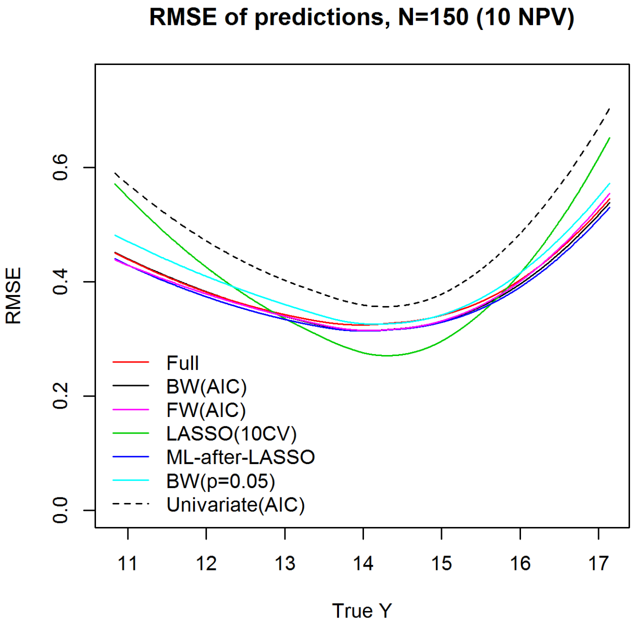
RMSE of regression coefficients, unconditional
simulation with 15 covariates



Accuracy of predictions
simulation with 15 covariates

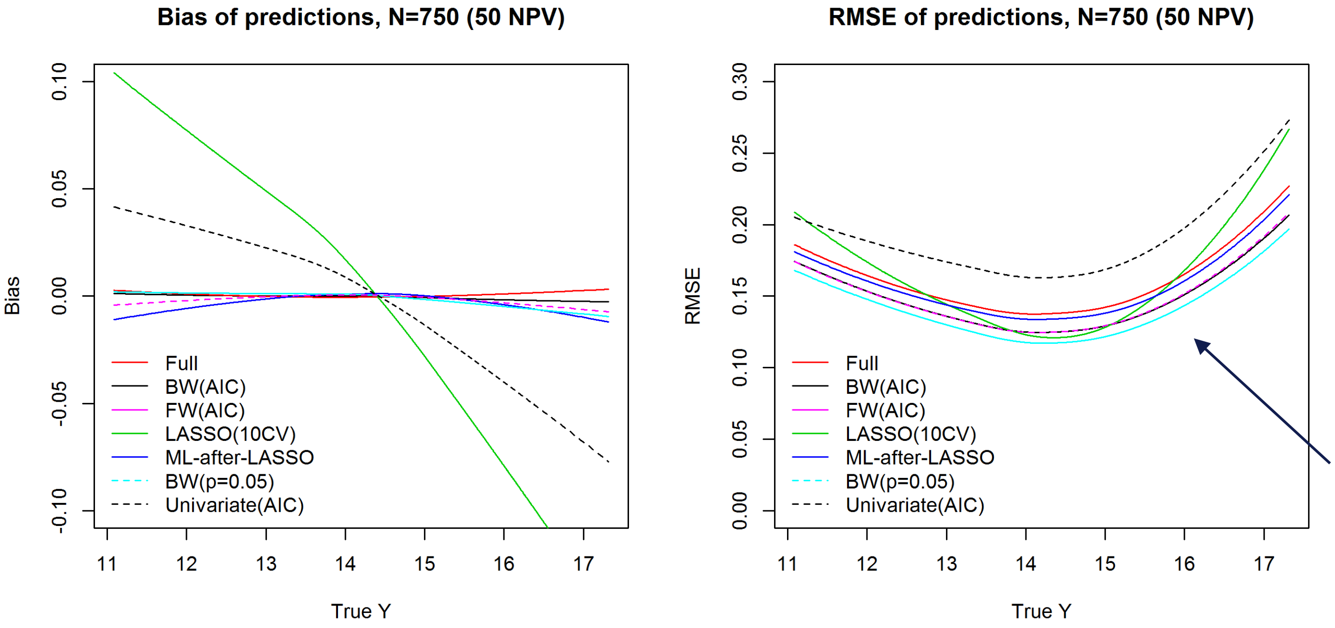
N=150 (10 EPV)


Bias of predictions, N=150 (10 NPV)


RMSE of predictions, N=150 (10 NPV)

Accuracy of predictions
simulation with 15 covariates

N=750 (50 EPV)


Bias of predictions, N=750 (50 NPV)

RMSE of predictions, N=750 (50 NPV)

Backward with $\alpha = 0.05$ was the best method!

Model (in)stability

- Variable selection generally introduces additional uncertainty
  - Instability of selection
  - Additional variance of regression coefficients
- Quantify this uncertainty using stability investigations:
  - Repeat selection algorithm in B bootstrap resamples
  - Compute (and report):
    - Variable inclusion frequencies (VIF) of each covariate
    - Model selection frequencies
    - Assess bias: *relative conditional bias (RCB)*
    - Assess variance inflation: *root mean squared difference ratio (RMSDR)*

Shrinkage

- *Phenomenon*: predictions from a model are too optimistic (too extreme)
  - Caused by overfit (too many parameters) in too small samples
- *Technique*: anticipate the shrinkage by adjusting estimates
  - Adjusted estimates of β are shrunken towards 0
  - Regularized regression: LASSO, ridge, …
  - Post-estimation shrinkage: Sauerbrei 1999, Dunkler et al 2016
    - Global shrinkage factor, equal for all β‘s
    - Parameterwise shrinkage factors: shrinkage according to strength

Recommendations: Generate initial working set

- Defendable assumptions on the role of covariates from background knowledge:
  - Previous studies in the same field
  - **Expert knowledge** (from PI, the domain expert)
  - Common sense

This defines the global model.

- Assumed relationships between covariates may be summarized in a DAG
  - Some covariates not needed?
  - Some effect estimates not interpretable?
- Background knowledge-based assessment of the effect strength
  - ‚strong‘: covariate should be in the model
  - ‚unclear‘: inclusion of a covariate debatable

This is where VS may be applied!

Recommendations: to select or not to select, and how

- No variable selection on ‚strong‘ covariates!
- Variable selection on ‚unclear‘ covariates: if sample size allows

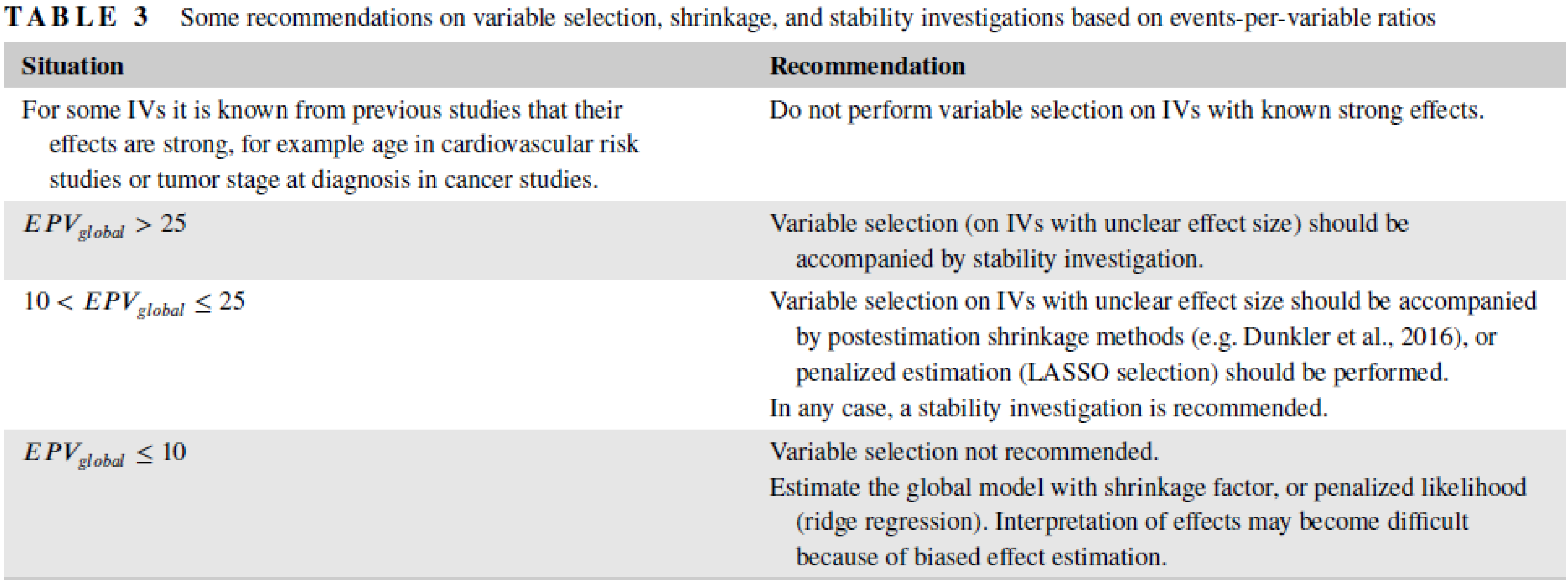
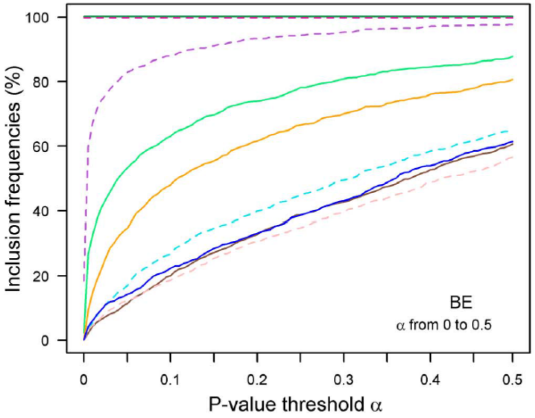


TABLE 3 Some recommendations on variable selection, shrinkage, and stability investigations based on events-per-variable ratios

Situation	Recommendation
For some IVs it is known from previous studies that their effects are strong, for example age in cardiovascular risk studies or tumor stage at diagnosis in cancer studies.	Do not perform variable selection on IVs with known strong effects.
$EPV_{global} > 25$	Variable selection (on IVs with unclear effect size) should be accompanied by stability investigation.
$10 < EPV_{global} \leq 25$	Variable selection on IVs with unclear effect size should be accompanied by postestimation shrinkage methods (e.g. Dunkler et al., 2016), or penalized estimation (LASSO selection) should be performed. In any case, a stability investigation is recommended.
$EPV_{global} \leq 10$	Variable selection not recommended. Estimate the global model with shrinkage factor, or penalized likelihood (ridge regression). Interpretation of effects may become difficult because of biased effect estimation.

Recommendations: what to do afterwards

- Compute (and report) stability measures:
  - Variable inclusion frequencies (VIF) of each covariate
  - Model selection frequencies
  - Assess bias: *relative conditional bias (RCB)*
  - Assess variance inflation: *root mean squared difference ratio (RMSDR)*
- Sensitivity analysis:
  - Changing decisions made in previous steps
  - Initial set of covariates?
  - Selection criterion?

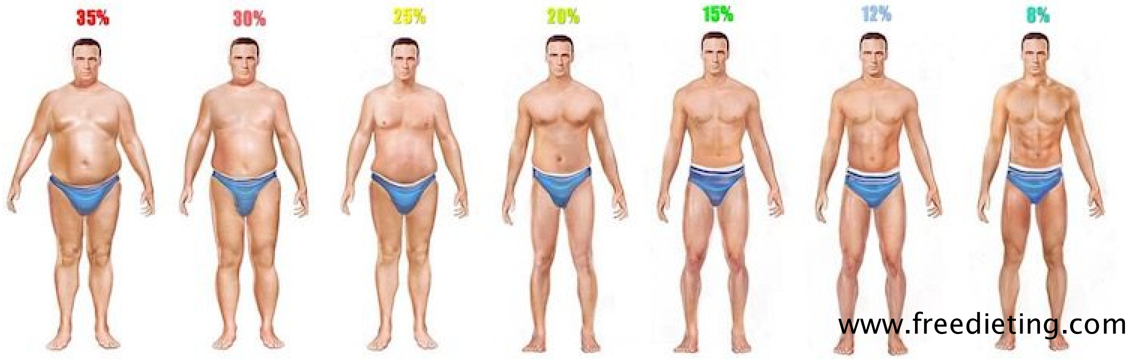


Recommendations: post-selection inference

1. The effect of a covariate should be formally tested, but no theory exists which variables should be included in the model
   - Solution: Perform inference in the global model.
2. Strong theory supporting only a small number of models
   - Solution: Perform multi-model inference with AIC (see Burnham Anderson 2002)
3. No strong theory for model building, but global model is implausible
   - Solution: Multi-model inference with resampled β's
   - Caveat: does not give formally valid confidence intervals (bias)
   - Overestimation bias may be corrected by shrinkage

Case study: body fat approximation

- Johnson‘s (1996) body fat data example
- Publicly available
- 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 (highly correlated)
- First goal: approximation of %body fat



www.freedieting.com

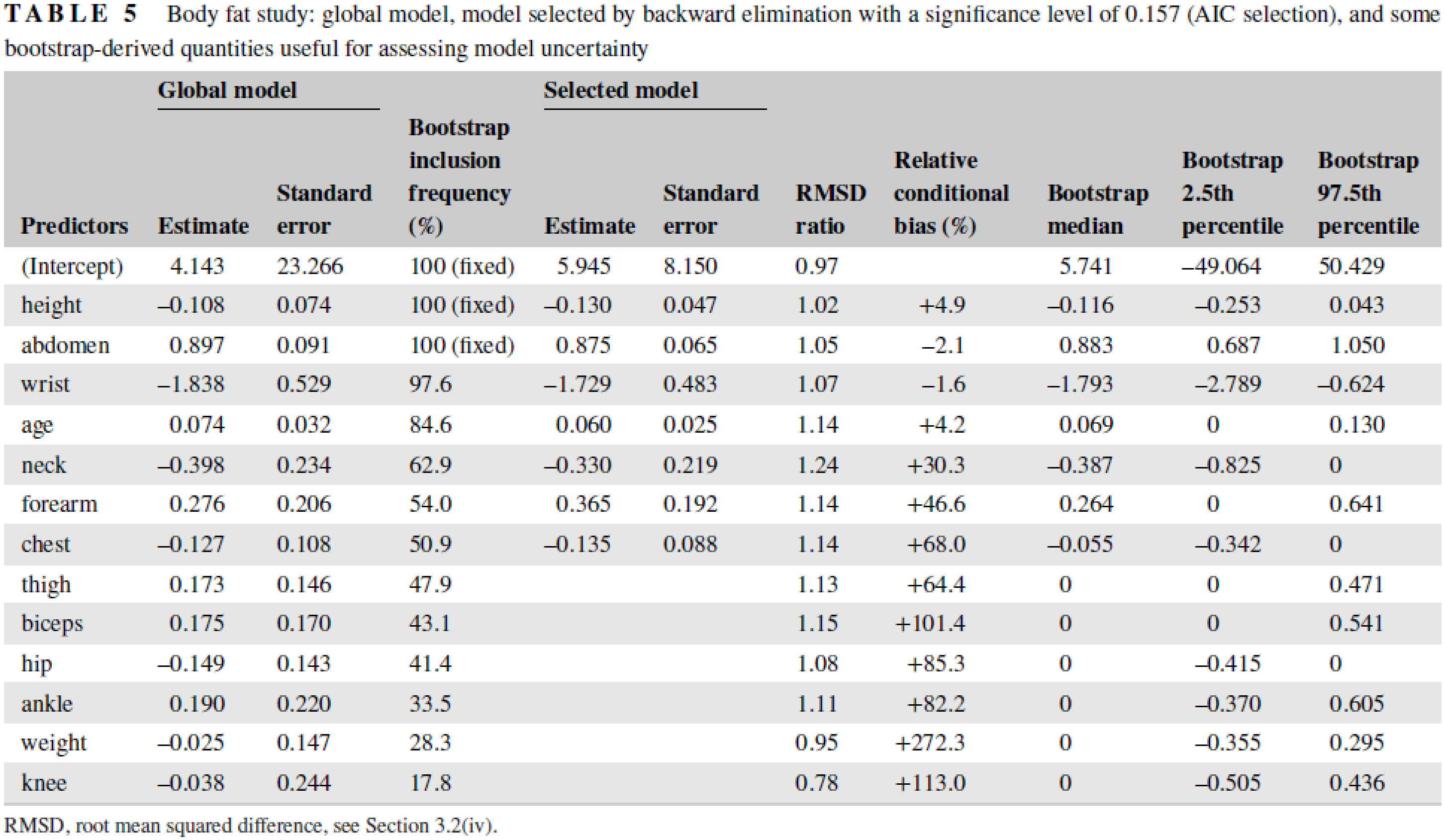


TABLE 5 Body fat study: global model, model selected by backward elimination with a significance level of 0.157 (AIC selection), and some bootstrap-derived quantities useful for assessing model uncertainty

	Global model			Selected model						
Predictors	Estimate	Standard error	Bootstrap inclusion frequency (%)	Estimate	Standard error	RMSD ratio	Relative conditional bias (%)	Bootstrap median	Bootstrap 2.5th percentile	Bootstrap 97.5th percentile
(Intercept)	4.143	23.266	100 (fixed)	5.945	8.150	0.97		5.741	–49.064	50.429
height	–0.108	0.074	100 (fixed)	–0.130	0.047	1.02	+4.9	–0.116	–0.253	0.043
abdomen	0.897	0.091	100 (fixed)	0.875	0.065	1.05	–2.1	0.883	0.687	1.050
wrist	–1.838	0.529	97.6	–1.729	0.483	1.07	–1.6	–1.793	–2.789	–0.624
age	0.074	0.032	84.6	0.060	0.025	1.14	+4.2	0.069	0	0.130
neck	–0.398	0.234	62.9	–0.330	0.219	1.24	+30.3	–0.387	–0.825	0
forearm	0.276	0.206	54.0	0.365	0.192	1.14	+46.6	0.264	0	0.641
chest	–0.127	0.108	50.9	–0.135	0.088	1.14	+68.0	–0.055	–0.342	0
thigh	0.173	0.146	47.9			1.13	+64.4	0	0	0.471
biceps	0.175	0.170	43.1			1.15	+101.4	0	0	0.541
hip	–0.149	0.143	41.4			1.08	+85.3	0	–0.415	0
ankle	0.190	0.220	33.5			1.11	+82.2	0	–0.370	0.605
weight	–0.025	0.147	28.3			0.95	+272.3	0	–0.355	0.295
knee	–0.038	0.244	17.8			0.78	+113.0	0	–0.505	0.436

RMSD, root mean squared difference, see Section 3.2(iv).

Conclusions

- VS methods have always been seen controversially
- VS methods can incur instabilities
- Software needed to assess model instability – repeat model building process in resamples
- In large samples, VS may reduce MSE and separate irrelevant information from the model
- In small samples, VS may have disastrous effects on precision and inference; this may go unnoticed in standard software!

- Recommended reading:
  Heinze, Wallisch, Dunkler (2018) Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal* 60:431-449. DOI: 10.1002/bimj.201700067
- Recommended R package abe (Blagus, 2017)

e-mail: Georg.heinze@meduniwien.ac.at Twitter: @Georg__Heinze