Education for Statistics in Practice, DAGStat 2016

Variable selection – a review and recommendations for the practicing statistician

Updated version!

Georg Heinze & Daniela Dunkler

Medical University of Vienna

CeMSIIS – Section for Clinical Biometrics

georg.heinze@meduniwien.ac.at, daniela.dunkler@meduniwien.ac.at

Education for Statistics in Practice, DAGStat 2016

Variable selection – a review and recommendations for the practicing statistician

Georg Heinze & Daniela Dunkler
Medical University of Vienna
CeMSIIS – Section for Clinical Biometrics

georg.heinze@meduniwien.ac.at, daniela.dunkler@meduniwien.ac.at

Aims of the lecture

- To explain the need for variable selection in analyses of observational studies.
- To understand the statistical concepts that variable selection could be based on.
- To review different variable selection strategies and modeling philosophies.
- To illustrate the urgent need for background knowledge in statistical modeling.

Agenda

- Part I-1: Philosophy
- Part I-2: Prerequisites
- Part I-3: Variable selection methods and strategies

Break

- Part II-1: Consequences of variable selection
- Part II-2: Case studies
- Part II-3: Recommendations

PART I-1: PHILOSOPHY

Magritte, Ockham, Einstein

What is this?

What is this?

„This is not a pipe“
René Magritte, 1928-29

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)

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: provide predictions with acceptable accuracy.
- They should be practically useful: allow conclusions such as ‘how large is the expected change in outcome if one of the explanatory variables changes by one unit’.
- They should be robust.

What are typical components of a statistical model?

Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age:	45 years
Gender:	○ Female ◉ Male
Total Cholesterol:	180 mg/dL
HDL Cholesterol:	50 mg/dL
Smoker:	◉ No ○ Yes
Systolic Blood Pressure:	135 mm/Hg
Are you currently on any medication to treat high blood pressure.	◉ No ○ Yes

Calculate Your 10-Year Risk

What can we learn from this model?

- **Prediction**
  Risk Score = 2%.
  Means 2 of 100 people with this level of risk will have a heart attack in the next 10 years.

- **Explanation**
  240 mg/dL and above 'high' blood cholesterol. A person with this level has more than twice the risk of heart disease compared to someone whose cholesterol is below 200 mg/dL.

(from http://cvdrisk.nhlbi.nih.gov/)

Purposes of multivariable models

- Prediction of an outcome of interest
- Identification of 'important' predictors
- Understanding the effects of predictors ('explanatory')
- Adjustment for predictors uncontrollable by experimental design
- Stratification by risk

(Royston & Sauerbrei, 2008)

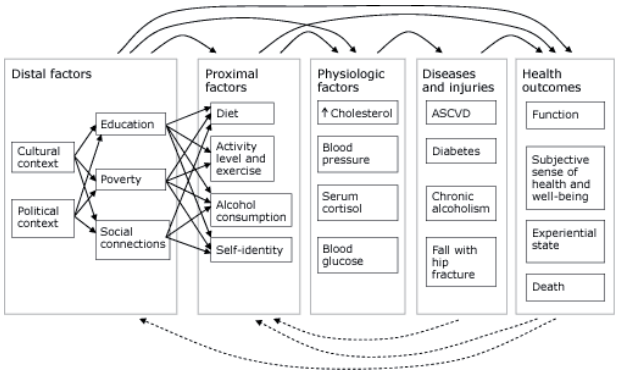
To Explain or to Predict?

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

(Shmueli, 2010)

Why multivariable modeling?

- 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)

Classes of modeling processes

1. The model is predefined. Estimate parameters and check assumptions. (Randomized trial.)
2. Develop a good predictor. Number of variables should be small.
3. Develop a good predictor. No limits in model complexity.
4. Assess the effect of a new factor of interest, adjusting for established factors.
5. Assess the effect of a new factor of interest, adjusting for confounding factors selected by data analysis.
6. Hypothesis generation of possible effects of factors in studies with many covariates.

Data-driven!

(Royston & Sauerbrei, 2008)

Is there 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.*'

Is there a true model?

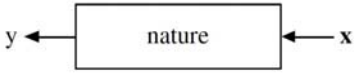
A 'true model' = a 'true data generating mechanism'.

Contra:

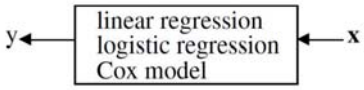
- '*We do not accept the notion that there is a simple "true model" in the biological sciences.*' (Burnham & Anderson, 2002)
- '*We recognize that true models do not exist… A model will only reflect underlying patterns, and hence should not be confused with reality.*' (Steyerberg, 2009)
- '*I started 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)

Do we need statistical models at all?

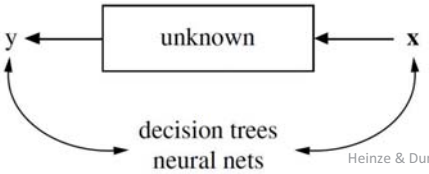
- Statistics starts with data. These data are 'generated' inside a black box by nature.



- *Statistical culture I*: Assume a stochastic data model for the inside of the box.

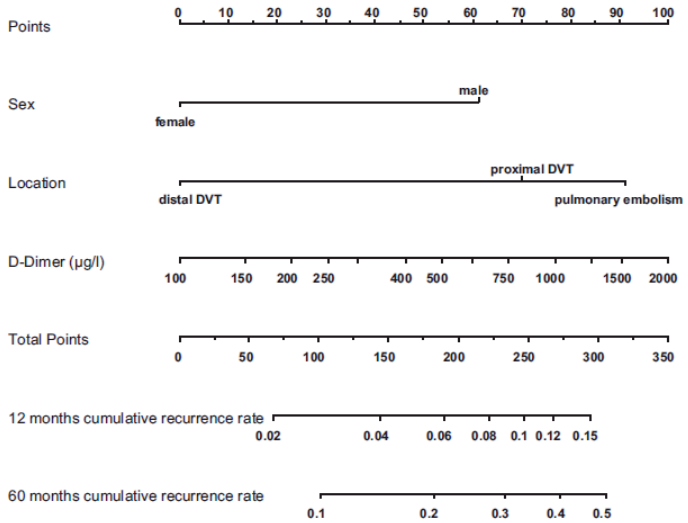


- *Statistical culture II*: The inside of the box is complex and unknown. Find a function $f(X)$ – an algorithm – that operates on X to predict the responses Y.



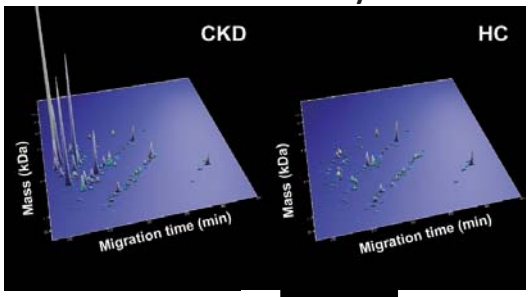
(Breiman, 2001)

Example I: Prediction of recurrence of venous thromboembolism

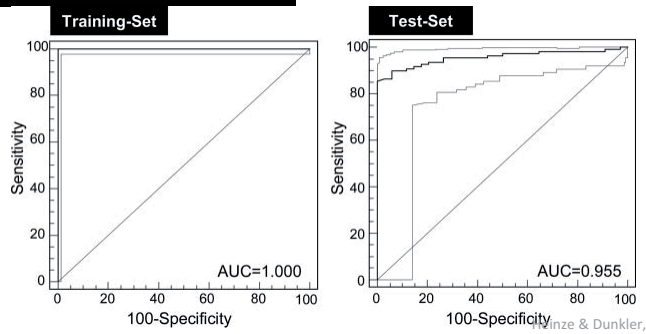


(Eichinger et al, 2010)

Example II: Urine-proteomic predictor of incidence of early chronic kidney disease



Support Vector Machine



(Good et al, 2010)

William of Ockham

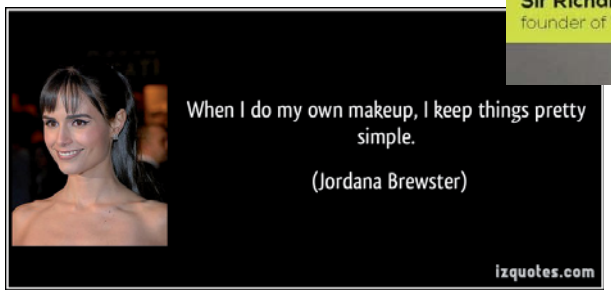
- 14th century logician and Franciscan friar: *'Pluralitas non est ponenda sine neccesitate.'* (Entities should not be multiplied unnecessarily.)
- When you have 2 competing theories that make exactly the same predictions, the simpler one is the better.
- If you have 2 equally likely solutions to a problem, choose the simplest.
- The explanation requiring the fewest assumptions is most likely to be correct.
- *'Simplicity is the ultimate sophistication.'* (Leonardo da Vinci)
- *'Everything should be made as simple as possible, but not simpler.'* (~Einstein)

Summary

- Models are not reality.
- There is no such thing as a 'true model'.
- There is not a single model that will ultimately explain data generation.
- Models can be useful: for pure prediction or for understanding multidimensional association.
- If two models have the same explanatory power, we prefer the simpler one.
- Complex models can be more accurate than simple ones, but are often less useful.

Focus of this presentation:

- Methods and consequences of variable selection



Statistical prerequisites

Types of models by distribution of error

Assumptions of models

Hypothesis tests: Likelihood ratio, Score, Wald

Model estimation: maximum likelihood

Likelihood and information-theoretic measures

Resampling techniques

Shrinkage

AIC and AICc

Penalized likelihood

Prior knowledge

Bias-variance tradeoff

Confounding

Change-in-estimate criterion

Preselection of variables

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

Discussion with non-statistical collaborator!

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

- Preselection = Bayes!

What models do we typically see?

Linear model

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

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

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

Common assumptions

Linearity: linear combination of variables

- (Relaxation: splines, fractional polynomials, GAMs)

Additivity: sum of effects

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

Interpretation of regression coefficients

- 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	-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

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' $\mathrm{EPV} = E/K$, $\mathrm{NPV} = N/K$.
- Often cited minimum $\mathrm{EPV} = 10$.

Events Per Variable (EPV)

- $\mathrm{EPV} = 10$ (Harrell 2001, p. 61)
  - Number of candidate variables, not variables in the final model.
  - Should be considered as lower bound!
- Non-linearity, interactions, etc. ➔ $\mathrm{EPV} \uparrow$.
- Prediction ➔ $\mathrm{EPV} \uparrow$ (logistic regression $\mathrm{EPV}\ 20-50$).
- Modern modeling techniques (random forests, neural networks, support vector machines) ➔ 10 times EPV compared to logistic regression ➔ $\mathrm{EPV} \uparrow\uparrow$ (van der Ploeg et al. 2014).

Likelihood and the principle of maximum likelihood

- Likelihood: probability of data given the model,
  interpreted as function of model parameters.

$$L(\beta|X,Y) = p(Y|\beta,X)$$

Fisher (aged 22):

- Maximum likelihood principle:
  find β such that $L(\beta|X,Y) \rightarrow \max$!

Ronald A. Fisher in 1913

Maximum likelihood theory

- First derivative,
- Second derivative,
- How to estimate (Newton-Raphson),
- Fisher Information,
- Variance of regression coefficients.

Hypothesis tests

Likelihood ratio test

- Compare likelihood of two hierarchically nested models M_1 and M_2 (M_2 nested in M_1)
- 'Nested' means that some β's in M_2 are forced to be 0.

$$2\log(L_1/L_2) \sim \chi^2(\Delta df)$$

- where Δdf is the difference in number of regression coefficients between the two models.
- Needs the fully fitted models M_1 and M_2.

Hypothesis tests

Scores test

- Needs only the model fit M_2, where $\beta_K = 0$.
- Evaluates if relaxing the restriction $\beta_K = 0$ would improve the model fit.
- Evaluates the first derivative of L_2 in the direction of β_K.
- If slope of L_2 is ‚steep', $\beta_K \neq 0$ should be assumed.

- = Classical 'forward' test.

Hypothesis tests

Wald test

- Needs only the model fit M_1, where $\beta_K \neq 0$.
- Evaluates if imposing the restriction $\beta_K = 0$ would not cause a significant drop in model fit.
- Evaluates the estimated variance of β_K at $\hat{\beta}_K$.

- = Classical 'backward' test.

Abraham Wald, 1902-1950

Testing models

- Likelihood ratio test is the 'state of the art' and widely considered the most precise test.
- Wald test and scores test are approximations to it, at low computational cost.

Testing between models

- What does it mean to test models?
  - OK if the test is 'prespecified' – rarely done in practice.
  - Not informative if models result from earlier testing (iterated testing: tests on 'generated' hypotheses).
- Consequence:
  - 'Tests' are interpretable if a few, pre-specified working models are compared.
  - We cannot trust the p-values from selected models!
- Modeling and hypothesis testing – two hostile brothers?

Information theory

- Suppose Likelihood = 1
- This is achieved if the data-generating mechanism is fully known.
- Expressed differently, log(likelihood) = *entropy* = 0.

Information theory

$$\text{entropy} \propto -\log(probability)$$

Ludwig Boltzmann, 1844-1906
Physicist and Philosopher

Photo by Janez Stare, http://graves.mf.uni-lj.si/

- Kullback-Leibler information happened to be the negative of Boltzmann's entropy developed 50 years earlier.

Akaike information criterion

- Akaike showed that for model selection we need to maximize the 'cross-validated' expectation of $\log L$ across several competitive models:

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

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

- This can be approximated by

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

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

- He defined $\mathrm{AIC} = -2\log L(x_{train}|\hat{\beta}_{train}) + 2K$.

K … number of parameters

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

Small-sample correction

- For small data sets:

$$\mathrm{AIC_C} = \mathrm{AIC} + \frac{2\mathrm{K}(\mathrm{K}+1)}{\mathrm{N}-\mathrm{K}-1}$$

K … number of parameters
N … sample size

- Use for $\frac{N}{K} < 40$.

The value of AIC

- We can compare two non-hierarchical models.
- We can compare several models.
- Hierarchical models: corresponding p-values

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

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

Comparing 2 models with AIC

AIC

- Interpret $\exp(-\frac{AIC}{2})$ as likelihood of the model, given data.

Evidence Ratios (ER)

$$\mathrm{ER} = \exp\left(-\frac{AIC_j}{2}\right) / \exp\left(-\frac{AIC_i}{2}\right)$$

- ER = 'How much likelier is M_j than M_i?'

Comparing R models with AIC

AIC differences

- $\Delta_i = AIC_i - AIC_{\min}$

Δ_i	1/ER	Level of empirical support for Model *i*
0-2	1 - 2.7	*Substantial*
4-7	7.4 - 33.1	*Considerably less*
> 10	>148	*Essentially none*

AIC weights

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_r \exp(-\Delta_r/2)}$$

- w_i is considered the weight of evidence in favor of M_i being the actual Kulback-Leibler best model *given* that one of the R models must be the Kulback-Leibler best model in that set.

(Burnham & Anderson, 2002)

Schwarz's Bayesian Information Criterion (BIC)

- Defined as $\mathrm{BIC} = -2\log L + \log(N)K$
- If the ‚true' model is among the candidate models, then BIC will select the true model as $N \to \infty$ (consistent model selection)
- For Cox or logistic models, , N' is the number of events, or min(events, non-events)
- More stringent selection for large N than for small N
- Compute equivalent *p*-value in R by `1-pchisq(log(N)*K, K)`
- For *K*=1, *N*=100: equivalent to $\alpha = 0.032$
- →AIC selects more variables than BIC

Resampling methods

Bootstrap

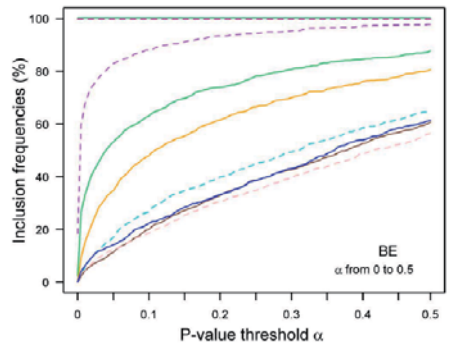
- Draw B samples with replacement from original data set.
- Perform model selection on each sample.
- Compute probability of selection of each model.

- Yields selection probabilities which are correlated with, but not identical to, Akaike weights.
- (Akaike weights consider the full ranked list of models in a data set, bootstrap only the ‚winner model' in each resample.)

- See SAS/PROC GLMSELECT (Part II-2).

Resampling methods

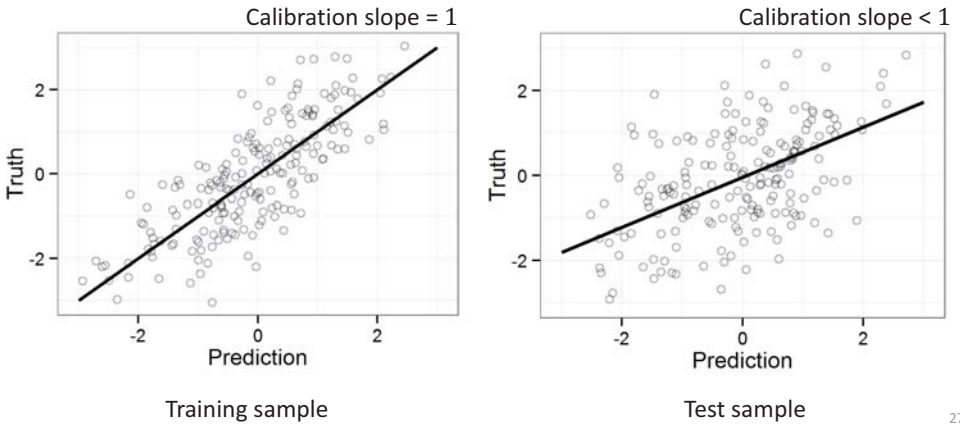
- Other uses of the bootstrap in model selection:
- **Bootstrap inclusion frequencies** (BIF) of each regression coefficient.
- **Pairwise inclusion tables**. (Sauerbrei & Schumacher, 1992)
- **Distribution of coefficients**.
- **Stability paths** (Meinshausen & Bühlmann, 2010): useful to assess dependence of inclusion on inclusion threshold.



Shrinkage

The phenomenon

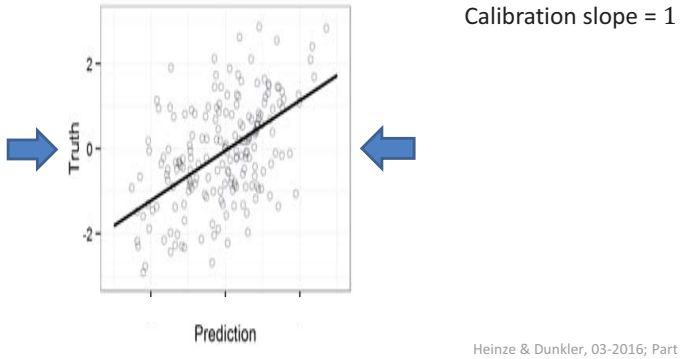
- Observed values in new samples are closer to overall mean than predicted values.



Shrinkage

The method(s)

- Anticipate shrinkage (of calibration slope) by cross-validation
- 'Shrink' regression coefficients such that a calibration slope of 1 would be expected.



Shrinkage methods

- Post-estimation shrinkage factor estimation
  - Verweij & Van Houwelingen 1993:
    global shrinkage factor c ($c < 0.8$ ➔ poor model)
  - Sauerbrei, 1999: parameterwise shrinkage factors
  - Dunkler, 2016: joint shrinkage factors, R package `shrink`
- Regularized regression
  - Ridge regression: L2 penalty on regression coefficients
  - Lasso: L1 penalty (Tibshirani, 1996 & 2011)
  - Elastic net: L2 and L1 penalty

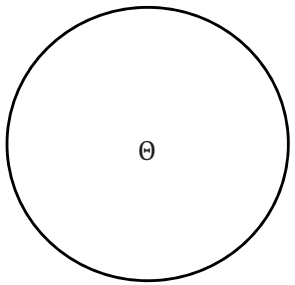
Shrinkage

- Empirical Bayes interpretation:
  penalty = data-dependent prior on regression coefficients.
- Consequences of shrinkage:
  - Controlling variance, not bias.
  - Effect estimation after shrinkage? ⚠
- Selection = extreme shrinkage!

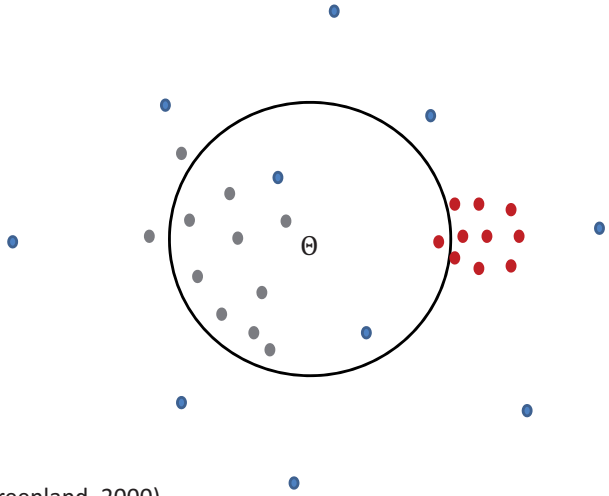
  "If it's close to 0, set it to 0."

- Not to be confused with bias correction!
  - It does not aim at unbiased regression coefficients!

Bias & efficiency



Bias & efficiency



(Figure 1 from Greenland, 2000)

Bias-variance tradeoff

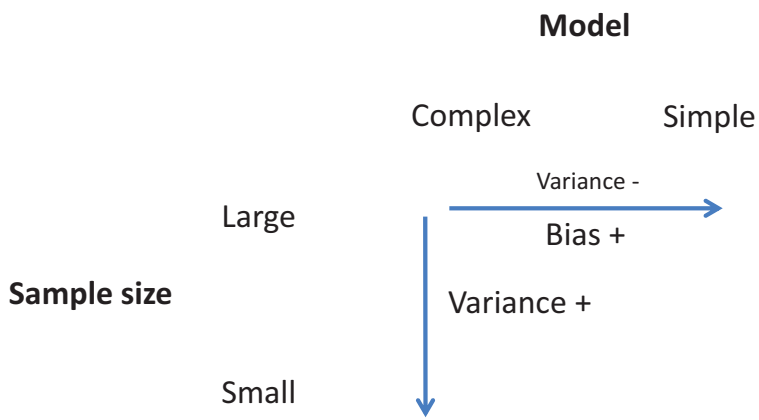
Assume $Y = f(X) + \epsilon$, with $\mathrm{E}(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$:

- Expected prediction error of a regression fit $f(X)$ at $X = x_0$:

$$\begin{aligned}\mathrm{Err}(x_0) &= E\left[\left(Y - \hat{f}(x_0)\right)^2 \middle| X = x_0\right] \\ &= \sigma_\epsilon^2 + \left[E\left(\hat{f}(x_0)\right) - f(x_0)\right]^2 + E\left[\hat{f}(x_0) - E\left(\hat{f}(x_0)\right)\right]^2 \\ &= \underbrace{\sigma_\epsilon^2}_{\textbf{Irreducible error}} + \underbrace{\mathrm{Bias}^2\left(\hat{f}(x_0)\right)}_{\mathrm{Bias}^2} + \underbrace{\mathrm{Var}(\hat{f}(x_0))}_{\mathrm{Var}}\end{aligned}$$

(Hastie, Tibshirani and Friedman, 2009, p. 223)

Bias-variance tradeoff



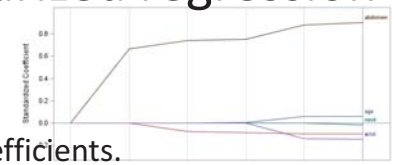
To explain or to predict?

Expected prediction error = Irreducible error + Bias^2 + Variance

Important in explanatory modeling

Important in predictive modeling

Penalized likelihood: regularized regression



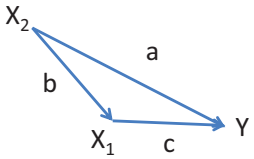
- LASSO: minimize $\sum_i (y_i - \hat{y})^2 + \lambda \sum |\beta_j|$
- Imposes a penalty on the regression coefficients.
- Prerequisite: adequate standardization of effects.

- What we obtain
  - A prediction formula with less error than ordinary least squares,
  - Variable selection.
- What we not obtain
  - Unbiased regression coefficients,
  - CI – even with bootstrap, variance of estimate is not helpful as it is not centered around true value.

Inclusion for addressing confounding

Directed acyclic graph (DAG)

- = A graph with one-way edges containing no cycles describing causal relationships.



Confounding

- Effect of X_1 on Y is confounded by X_2, if X_2 is effect of both X_1 and Y.
- ➔ X_2 must be considered to regain causal interpretation of effect of X_1 on Y.

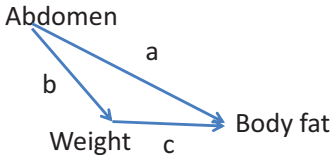
(Pearl, 1995)

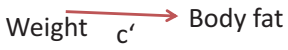
Change-in-estimate criterion

- In epidemiologic studies, it is often not clear whether adjustment for a variable X_2 is necessary or not.

Change-in-estimate criterion

- If X_2 (abdomen circumference) is a confounder (a and b exist), then its removal will change our assessment of arrow c from weight to body fat.
- So we could remove 'abdomen' and see what happens to c: $\mathrm{CIE} = \mathrm{c}' - \mathrm{c}$.





Change-in-estimate criterion

- M_1: $\beta_0 + \beta_1 X_1 + \beta_2 X_2$
- M_2: $\theta_0 + \theta_1 X_1$

- Change in estimate criterion: leave X_2 in the model
  if $\beta_1 - \theta_1 \neq 0$,
  often proxied by
  $$\text{abs}(\hat{\theta}_1 - \hat{\beta}_1)/\hat{\beta}_1 > 0.10$$

- This leads to inconsistent variable selection
  (Maldonado & Greenland, 1993)
- To get a consistent estimator, we could test for $\beta_1 \neq \theta_1$ (collapsibility of the two models).

(see also Lee, 2014)

Significance of change-in-estimate

- Tests for collapsibility by bootstrapping or
- Dunkler et al (2014) approximate the change-in-estimate and derive a simple test for $\beta_1 - \theta_1 = 0$.

They show:

- Elimination of a ‚significant' variable X_2 from a model leads to a significant change $\hat{\beta}_1 - \hat{\theta}_1$.
- Elimination of a 'non-significant' variable X_2 from a model leads to a non-significant change $\hat{\beta}_1 - \hat{\theta}_1$.

- ➔ Test of collapsibility = Test of omitted variable.

Prior knowledge: simple illustrative simulations

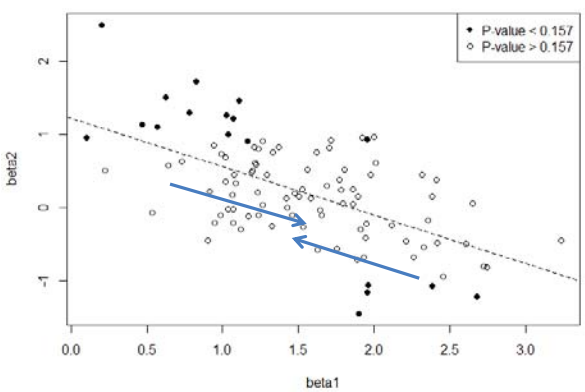
- How poor prior knowledge can result in poor results (simulation with $N = 50$).

True $\beta_1 = 1.5, \beta_2 = \mathbf{0.3}$

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 $\hat{\beta}_1$!

RMSE($\hat{\beta}_{1,FULL}$) = 0.67
RMSE($\hat{\beta}_{1,BE}$) = 0.65
Bias($\hat{\beta}_{1,FULL}$) $= -0.03$
Bias($\hat{\beta}_{1,BE}$) = +0.03



➔ **'Selection is good.'**

Prior knowledge: simple illustrative simulations

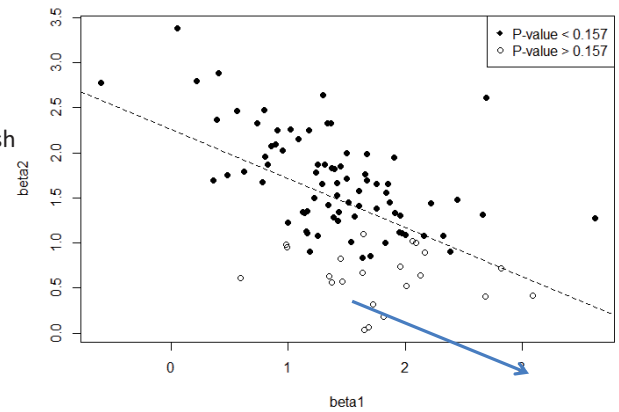
- How poor prior knowledge can result in poor results (simulation with $N = 50$).

True $\beta_1 = 1.5, \beta_2 = \mathbf{1.5}$

A strong β_2:

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

RMSE($\hat{\beta}_{1,FULL}$) = 0.68
RMSE($\hat{\beta}_{1,BE}$) = 0.67
Bias($\hat{\beta}_{1,FULL}$) $= -0.03$
Bias($\hat{\beta}_{1,BE}$) = +0.33



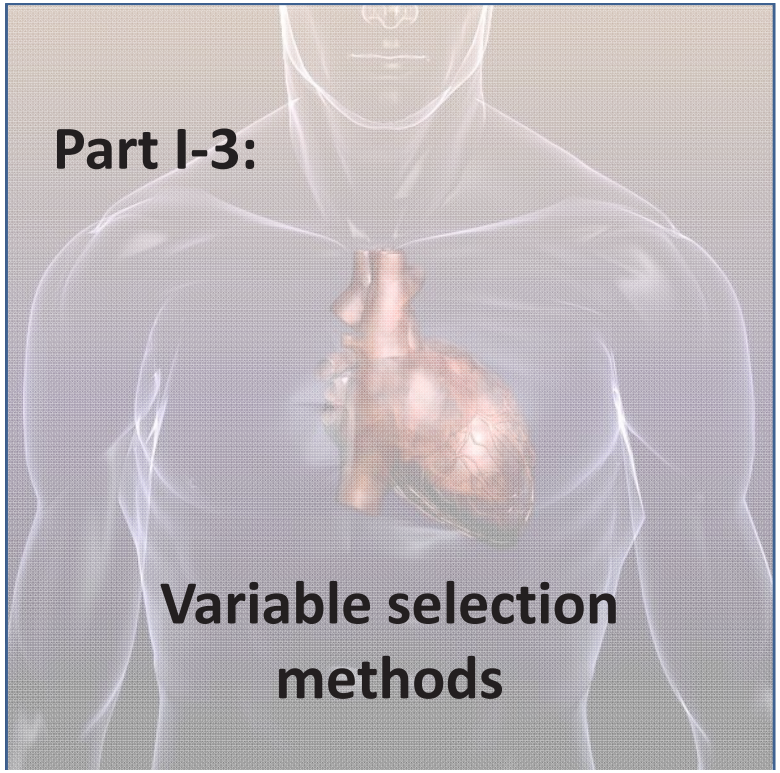
➔ **'Selection is bad.'**

Prior knowledge

We should have known the likely role of X_2 in advance:

- If it is considered a strong effect, never let it be deleted from the model!
- If it is considered a weak effect, selection can improve performance. ➔ less variance (Shmueli, 2010)
- If it is considered no effect, it should better not be used upfront ('instrumental variable').

$$X_2 \longrightarrow X_1 \longrightarrow Y$$



Basic algorithms

- 'Full' model
- Univariable filtering
- Best subset selection
- Forward selection
- Backward elimination
- Change-in-estimate: Purposeful variable selection and augmented backward selection
- Information-theoretic approach
- Directed acyclic graph (DAG)-based selection

The 'Full' model

- Means: do not perform any data-driven variable selection.
- Select, for each variable, a desired level of non-linearity (including spline transformations).
- Select some biologically plausible interactions.

- Variables should be pre-selected by 'expertise'.

Univariable filtering

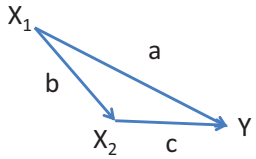
- Still by far the most often applied variable selection method in medical literature!

- Select a significance level α (e.g., α=0.20 or α=0.157)
- Perform K univariable models.
- Use all variables in multivariable model with univariable p-value < α.
- Sometimes accompanied by subsequent backward elimination.

Pros and cons of univariate selection

- Easy. (You can do that with any software.)
- Retraceable.

- Problematic:
- 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:

- 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.

In practice:

- Approximated by stepwise approaches!

(Burnham & Anderson, 2002)

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 step()`

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.

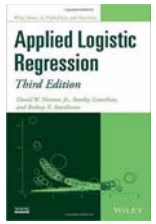
Variant: Stepwise backward

- 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 mfp:mfp()`

Purposeful selection

- Proposed by Hosmer and Lemeshow in their books on applied logistic regression and applied survival analysis.
- Starts with univariate screening.
- Then performs backward elimination, but leaves variables in the model if omission would cause a large (proportional) change-in-estimate in other variables.
- Additional forward steps.
- A bit outdated.



(Hosmer & Lemeshow, 1999 & 2000)

Augmented backward elimination

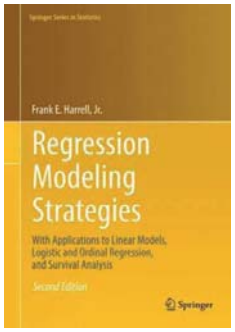
- Proposed by Dunkler et al, 2014.
- Re-investigated the change-in-estimate criterion and proposed a standardized version and a short-cut approximation to it.
- Based on backward elimination with level α_2.
- Leaves variable in a model if maximum of standardized changes-in-estimate greater than τ.
- Simulation study showed that results and performance are always close to the full model, but fewer variables are selected.

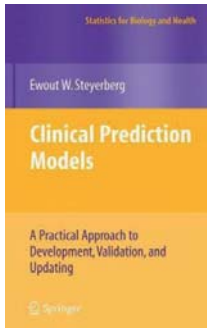
Software:
`SAS macro %ABE`

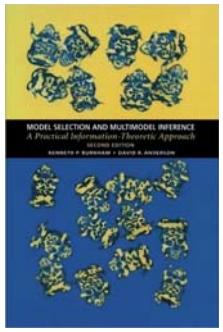
Opinions on variable selection

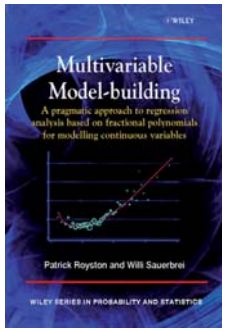
for models with focus on prediction and explanation.

Variable selection



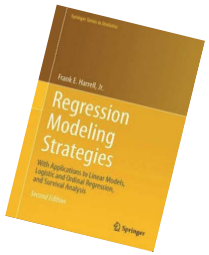






(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

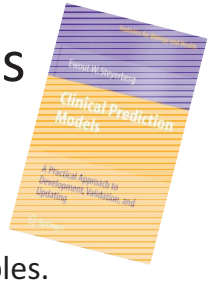
Harrell's recommendations



- Focus on prediction models.
- 'Effects cannot be assumed to be exactly 0.'
- 'Selection invalidates confidence intervals and p-values.'
- Specify a full model, including meaningful interactions and non-linear effects.
- Perform global tests for interactions or non-linear effects.
- At most: do a mild backward selection at $\alpha_2 = 0.50$.
- Model simplification using cross-validated predicted values as outcome.

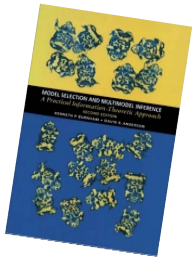
(see also Harrell, 1996)

Steyerberg's recommendations



- Focus on prediction models.
- False inclusion is better than false exclusion of variables.
- Stepwise methods may lead to
  - Instability of selection,
  - Biased estimation of coefficients,
  - Misspecification of variability (exaggerated p-values),
  - Predictions of worse quality than from a full model.

Burnham-Anderson's recommendations



- Strong focus on explanatory models.
- Select a set of models that are biologically plausible.
- These are subset models of a global model.
- Apply information-theoretic approach.
- Compute AIC weights or bootstrap weights.
- Perform multi-model inference (problem: no variable selection!).

90% confidence set

Model	Δ_i	$\mathcal{L}(M_i\|x)$	w_i
1	0	1	0.431
2	1.2	0.5488	0.237
3	1.9	0.3867	0.167
4	3.5	0.1738	0.075
5	4.1	0.1287	0.056
6	5.8	0.0550	0.024
7	7.3	0.0260	0.010

$\Delta_i = AIC_i - AIC_{\min}$

Akaike weight: $w_i = \frac{\exp(-\Delta_i/2)}{\sum_r \exp(-\Delta_r/2)}$

Model averaging

- $\bar{\hat{\beta}}_j = \frac{\sum_r \hat{\beta}_j I_{r,j} w_{j,r}}{w^+(j)}$

  $I_{r,j}$ … inclusion of β_j in model r

  $w^+(j)$ … sum of weights of models including β_j

- $\widehat{var}\left(\bar{\hat{\beta}}_j\right) = \left[\sum_r \underbrace{w_r}_{\text{weight}} \sqrt{\underbrace{\widehat{var}\left(\hat{\beta}_{j,r}|M_r\right)}_{\text{within-model variance}} + \underbrace{\left(\hat{\beta}_{j,r} - \bar{\hat{\beta}}_j\right)^2}_{\text{between-model variance}}}\right]^2$

(Buckland, 1997)

Burnham-Anderson's recommendations

For explanatory model

- If there is a dominating model with $w_i > 0.9$, just report this one unconditionally.
- Otherwise, report the best performing model, with unconditional variance based on model-averaged inference on the models of the 90% confidence set.

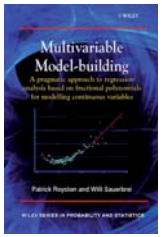
For prediction model

- Perform model-averaged inference (averaged point estimate and variance).

Bootstrap model frequencies can replace the Akaike weights.

Relative importance of a variable X_j: $w^+(j) = \sum_r w_j I_{j,r}$

Royston-Sauerbrei 's recommendations



- Focus on explanatory and descriptive models.
- Initial working set of variables.
- Coding matters.
- Backward elimination with additional forward steps.
- Function selection. (not covered here)
- 'If you have a large enough sample, you can use selection methods.'
- They propose backward elemination.
- Select α_2 according to needs; larger value means larger model.
- Emphasize importance of investigation of model stability ➔ by means of resampling.

Coding

- One interesting aspect (out of many) in the Royston-Sauerbrei book is coding of categorical variables:
- Nominal variables: choose an appropriate reference.
  - Frequent, standard group, etc.
  - Variable selection on dummys – collapse rare groups with reference
- Ordinal variables: advantages of ordinal coding
  - Variable selection can then collapse adjacent groups with similar outcome

Level	Dummy1	Dummy2
0	0	0
1	1	0
2	1	1
Etc.		

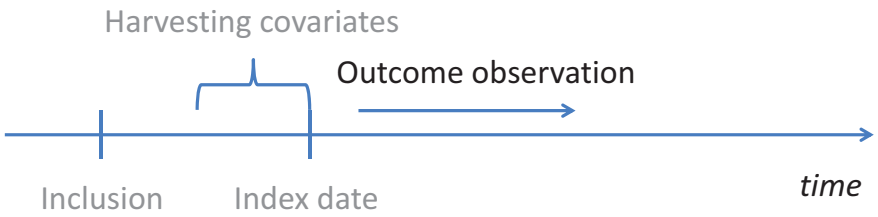
(Royston & Sauerbrei, 2008)

Differences (and similarities) in prediction and causal modeling

- Both are using maximum likelihood ➔ prediction as vehicle to find estimates.
- While prediction focusses on $\hat{Y}$, causal modeling focusses on $\hat{\beta}$.
- In prediction, important prerequisites for selecting variables are:
  - Chronology (do not use future values!, e.g. time-dependent variables in survival analysis),
  - Availability at time of prediction.
- In causal modeling, it is confounder control. ➔DAG methodology

Preselection for prediction models

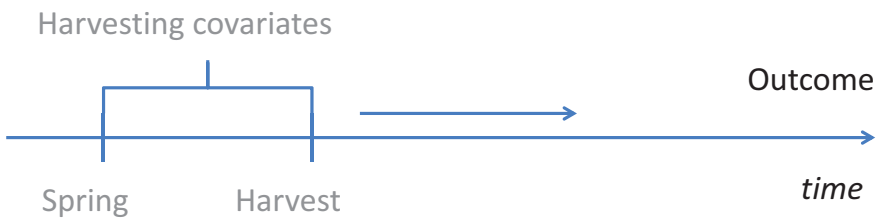
- Chronology:



- Don't use information from the future for prediction/effect estimation!
  (This is one of the most often violated conditions in practice!)
- X must be available also in prediction situation.

Example: quality of wine

- Chronology:

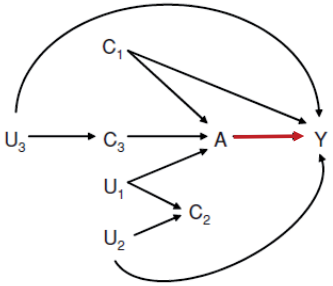


Grapes on vine

Barrel

Using causal DAGs to identify confounders

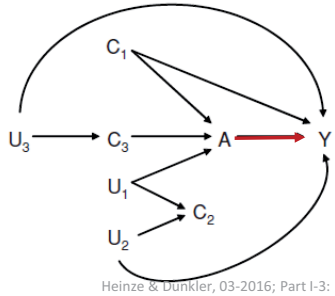
- Pearl (1995) described causal relationships by DAGs.
- We are interested in the effect of A on Y.
- Confounder adjustment should be made for:



- Confounders (parents of A and Y: C_1) (BIAS)
- Backdoor path blockers (they look like confounders: C_3) (BIAS)
- NOT for instruments (C_3 if U_3 were not there) (VARIANCE)
- NOT for colliders (C_2) (BIAS)

Implication of the DAG view on explanatory models

- **This implies that there cannot be a single model explaining Y,**
- But the choice of model depends on what we want to estimate:
- E.g., the causal effect of A on Y.
- If we were interested in the effect of C_1 on Y, we would not adjust for A (and not for any other variable).



Confounder selection criteria

In practice, true causal relationship is usually unknown.

Pretreatment criterion (Rubin, 2009)

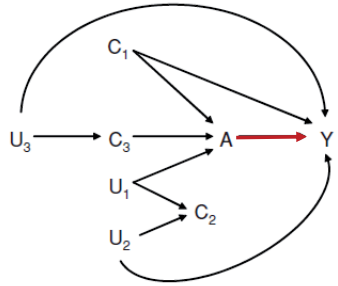
- All variables preceding A.

Common cause criterion (Glymour et al, 2008)

- Variables that are cause of A **and** of Y.

Disjunctive cause criterion (VanderWeele & Shpitser, 2011)

- Variables that are cause of A **or** of Y.



The disjunctive cause criterion (DCC)

VanderWeele & Shpitser (2011) argue that with DCC,

- No detailed knowledge about all causal relationships is needed,
- If any subset of the observed variables suffices to control confounding, those identified by DCC will also suffice.
- Further backward elimination can improve confounder control in efficiency.
- Disadvantage: the DCC can amplify bias by unmeasured confounding.

- Disadvantage: we are never completely sure about the arrows in the DAG.

DAGs

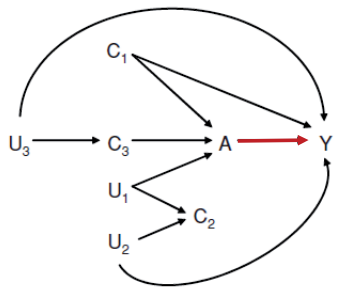
*'..., there are **known knowns**; there are things we know we know. We also know there are **known unknowns**; that is to say we know there are some things we do not know. But there are also **unknown unknowns** – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones.'*

Donald Rumsfeld, February 12, 2002 about the lack of evidence linking the government of Iraq with the supply of weapons of mass destruction to terrorist groups.

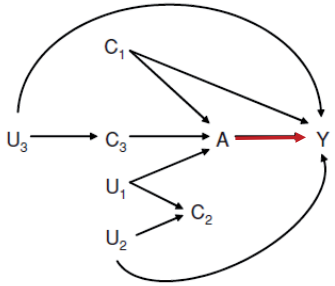
Using causal DAGs to identify confounders

Consider A 'always selected'.

- Confounding control is by adjusting for
- C_1 (a confounder of A),
- C_3 (a seeming confounder of A),
- but not for C_2 (a collider).

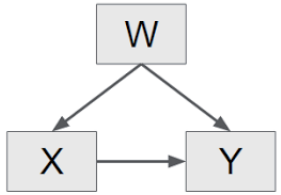


Performance of various approaches



- Pretreatment criterion: C_1, C_2, C_3
- Common cause: C_1
- DCC: C_1, C_3
- Univariate selection: (C_1), C_2, C_3
- Backward elimination, Lasso & Co: C_1, C_2, C_3
- Backward elimination after DCC: C_1, C_3

An Example – Confounder



R Code

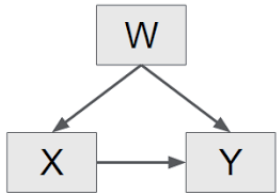
```
> N <- 100000
> w <- rnorm(N)
> x <- .5 * w + rnorm(N)
> y <- .4 * x + .3 * w + rnorm(N)

> summary(lm(y ~ x))
```

	Estimate	Std. Error	Pr(>|t|)
Intercept	-0.003	0.003	0.332
x	**0.522**	0.003	<2e-16

Adjusted R-squared: 0.2436

(http://anythingbutrbitrary.blogspot.co.at/2016/01/how-to-create-confounders-with.html)

An Example – Confounder



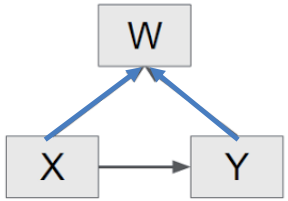
R Code

```
> N <- 100000
> w <- rnorm(N)
> x <- .5 * w + rnorm(N)
> y <- .4 * x + .3 * w + rnorm(N)

> summary(lm(y ~ x + w))
```

	Estimate	Std. Error	Pr(>|t|)
Intercept	-0.002	0.003	0.373
x	**0.403**	0.003	<2e-16
w	**0.298**	0.004	<2e-16

Adjusted R-squared: 0.294

An Example – Collider



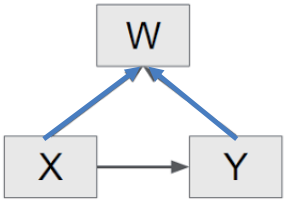
R Code

```
> N <- 100000
> x <- rnorm(N)
> y <- .7 * x + rnorm(N)
> w <- 1.2 * x + .6 * y + rnorm(N)

> summary(lm(y ~ x))
```

	Estimate	Std. Error	Pr(>|t|)
Intercept	-0.009	0.003	0.00486
x	**0.702**	0.003	<2e-16

Adjusted R-squared: **0.3285**

An Example – Collider



R Code

```
> N <- 100000
> x <- rnorm(N)
> y <- .7 * x + rnorm(N)
> w <- 1.2 * x + .6 * y + rnorm(N)

> summary(lm(y ~ x + w))
            Estimate   Std. Error   Pr(>|t|)
Intercept    -0.007       0.003      0.0135
x            -0.016       0.005      0.0008
w             0.443       0.002      <2e-16

Adjusted R-squared:  0.5075
```


DAG: summary

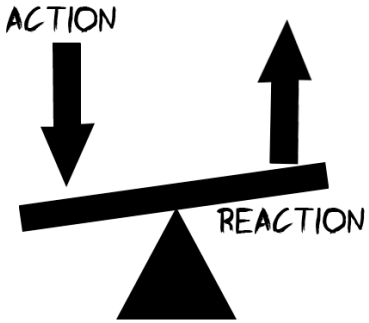
- In causal effect estimation, setting up a DAG can help to identify the set of adjustment variables.
- The DAG is 'rife with assumptions'.
- Rules like 'pretreatment', 'disjunctive cause criterion', etc. help to make the results robust against violations.

Effect estimation and use of penalized likelihood methods

- The effect of interest should not be penalized to obtain an unbiased estimate.
- But: penalizing all other effects (confounders) can be harmful, as their effective degrees of freedom are reduced.
- The extreme case is that the confounder effects are shrunken such that essentially an unadjusted effect is estimated.
- It seems that an unbiased effect estimate can sometimes only be obtained at the cost of a large variance.

Summary

- There exists no single, simple 'true model'.
- Different variable selection strategies have been favored by different authors.
- Depending on the data they usually see.
- All have in common that:
  - existing knowledge should be used,
  - models should be interpretable.



Part II-1: Consequences of variable selection – some simulations

(from http://whatnextbook.com/wordpress/tag/decisions/)

Questions

- How stable is variable selection?
- Does variable selection induce bias of β?
- Does variable selection increase RMSE of β?
- Does variable selection lead to biased or inaccurate predictions?
- How does background knowledge improve results?

Simulation setup

- Binder et al (2011) setup.
- Linear regression
- Sample size N such that $NPV = 5,10,20,50,100$
- Methods:
  - Full model,
  - BW(AIC) … Backward elemination with $p = 0.157$,
  - FW(AIC) … Forward selection with $p = 0.157$,
  - Lasso(10CV) … LASSO with 10-fold cross-validation,
  - ML-after-Lasso(10CV) … Maximum-Likelihood after LASSO,
  - BW($p = 0.05$)
  - Uni … univariate selection

Correlation structure

Total R^2=46%

Partial R2:
X1 = 0.3%
X3 = 2.2%
X4 = 0.8%
X5 = 4.0%
X6 = 3.3%
X8 = 1.8%
X10 = 1.8%
X11 = 5.0%

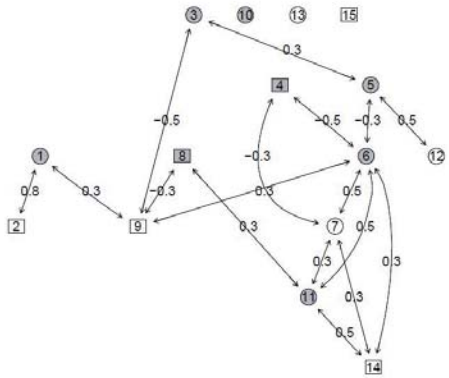


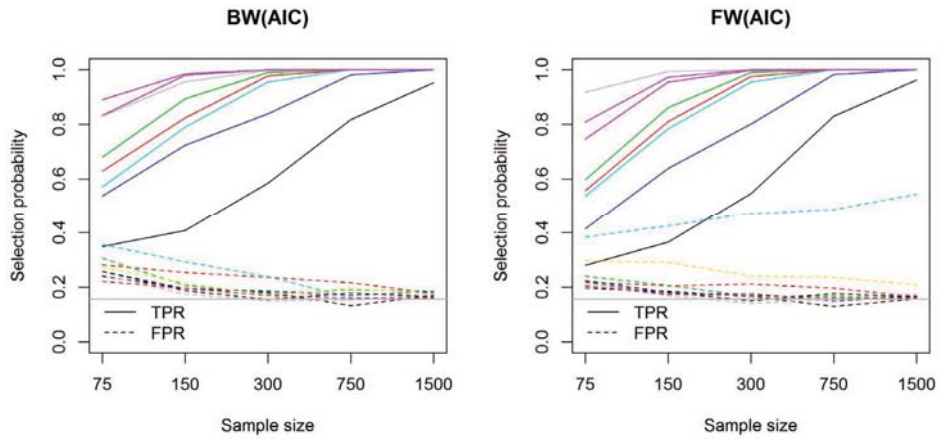
Figure 2: Partial correlations of the variables $z_{ij}, j = 1, \ldots, 15$, underlying the covariates. Variables that form the basis for continuous covariates are indicated by circles, variables that correspond to categorical covariates are indicated by rectangles. Variables corresponding to covariates that have an effect on the response are indicated by gray shading.

Simulation study: a note of caution

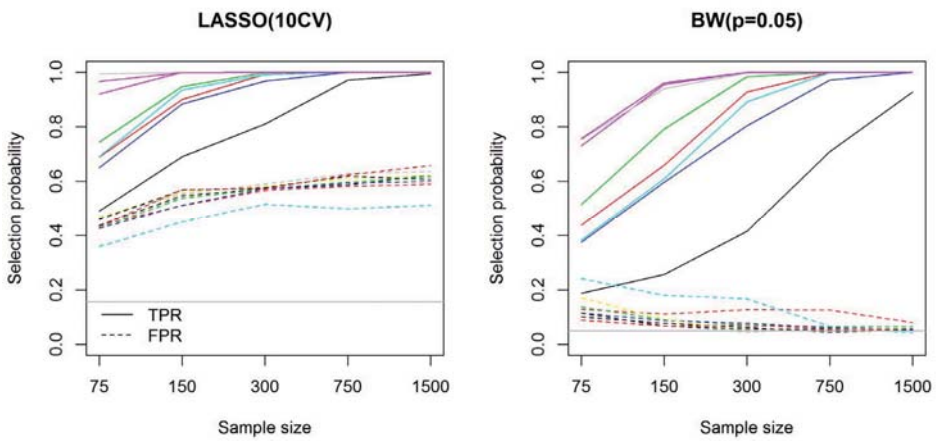
- We assume a 'true model', even if we doubted its existence in Part I.
- We assume that a variable selection method may discover that 'true model'.
- This way we can learn about the behavior of variable selection methods under known population properties.
- We can also evaluate 'explanatory performance' of the model (bias/RMSE of regression coefficients).

- Other way to compare methods: best cross-validated performance in complex data sets.
- No general properties can be derived!

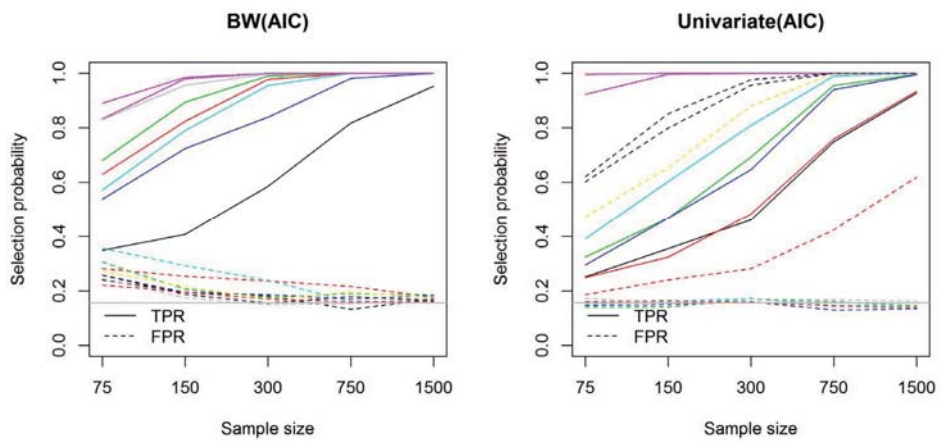
Results: selection Type I and II error



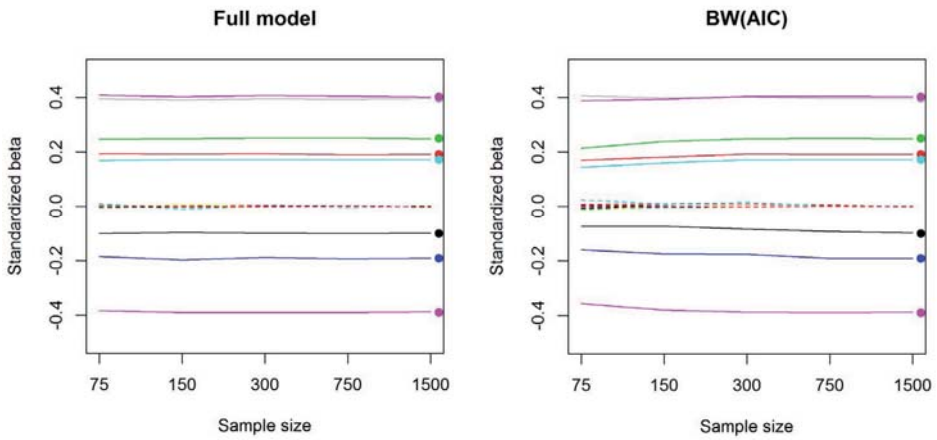
Results: selection Type I and II error



Results: selection Type I and II error

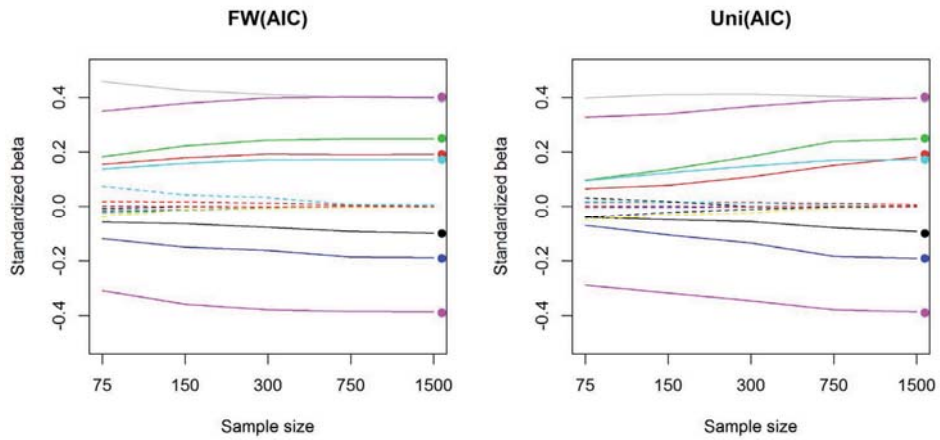


Results: regression coefficients, unconditional



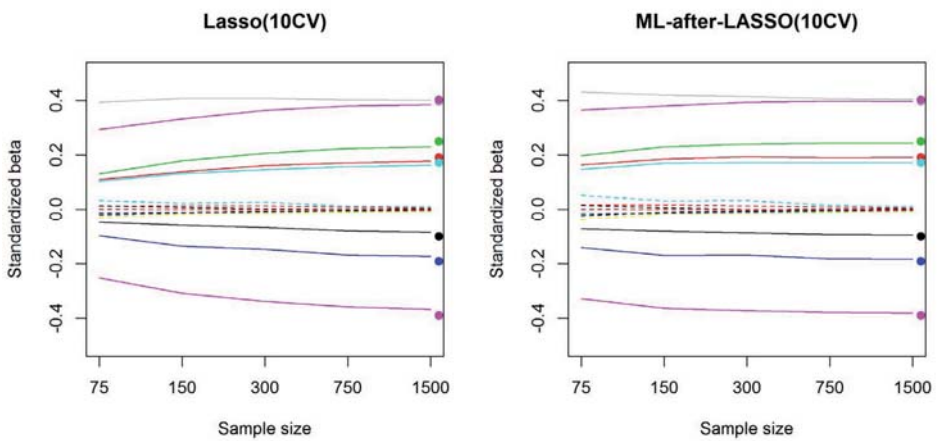
In these scenarios, unconditional bias of β is towards null!

Results: regression coefficients, unconditional



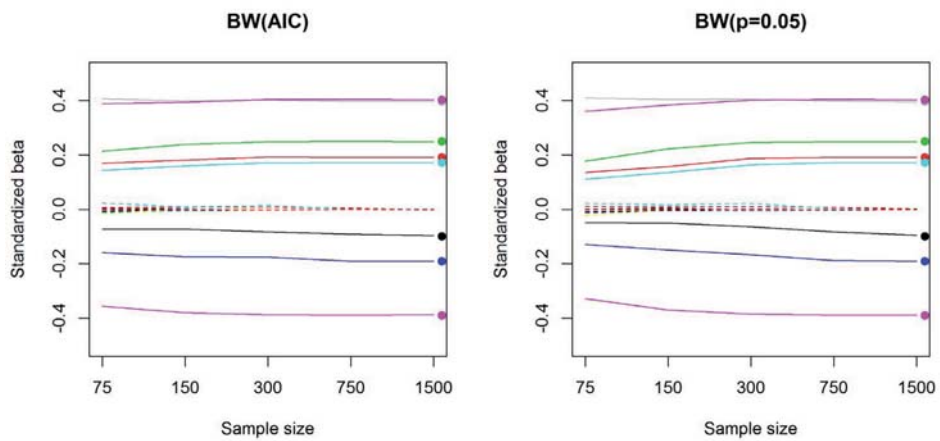
In these scenarios, unconditional bias of β is towards null!

Results: regression coefficients, unconditional



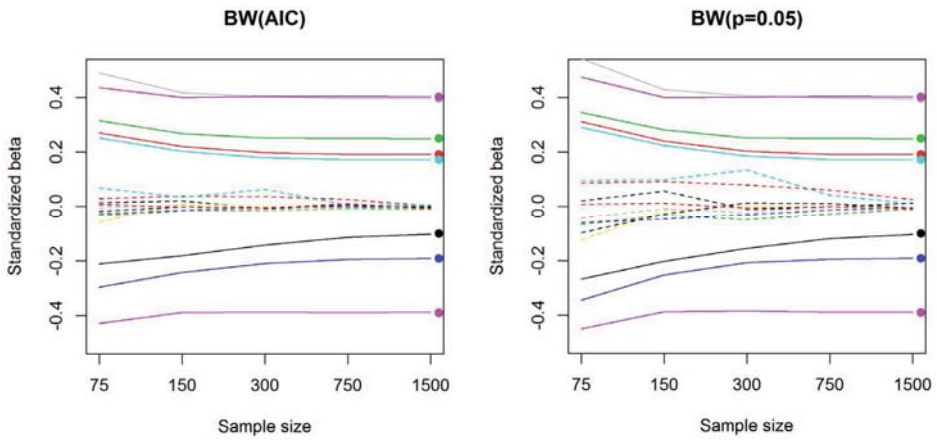
In these scenarios, unconditional bias of β is towards null!

Results: regression coefficients, unconditional



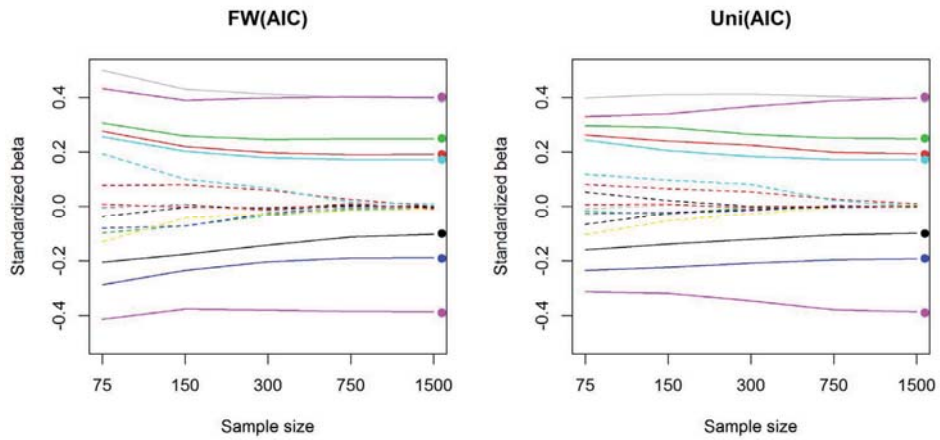
In these scenarios, unconditional bias of β is towards null!

Regression coefficients, conditional

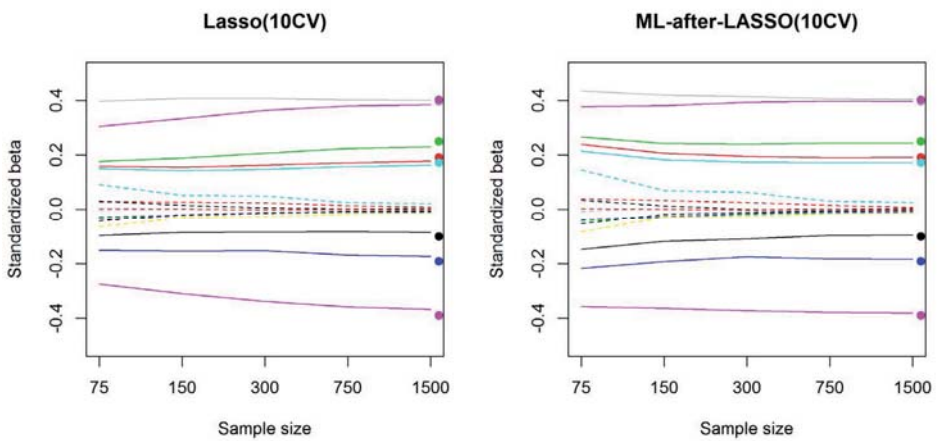


Conditional bias of β is away from null!

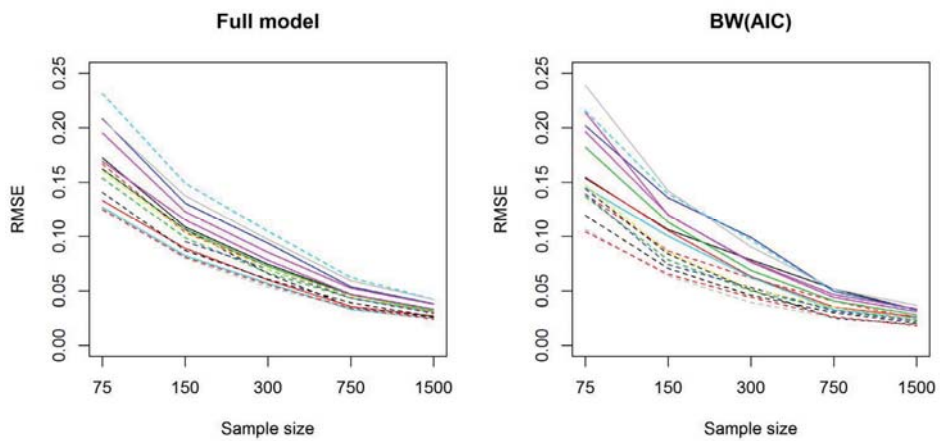
Regression coefficients, conditional



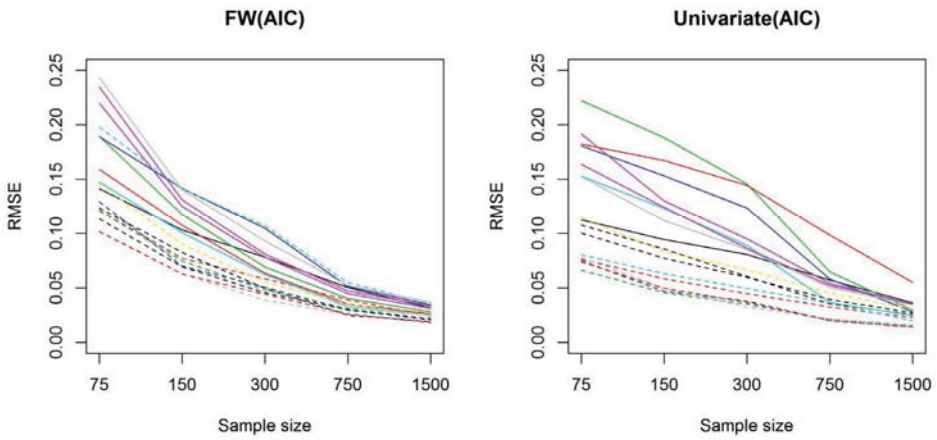
Regression coefficients, conditional



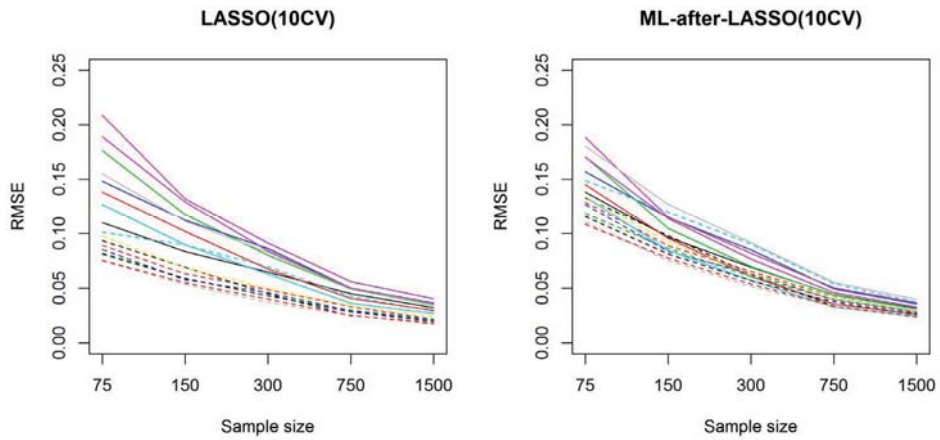
RMSE of regression coefficients, unconditional



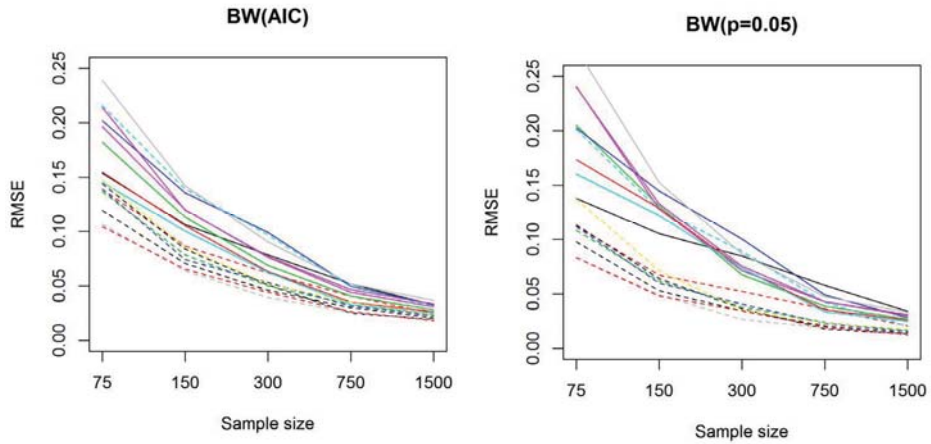
RMSE of regression coefficients, unconditional



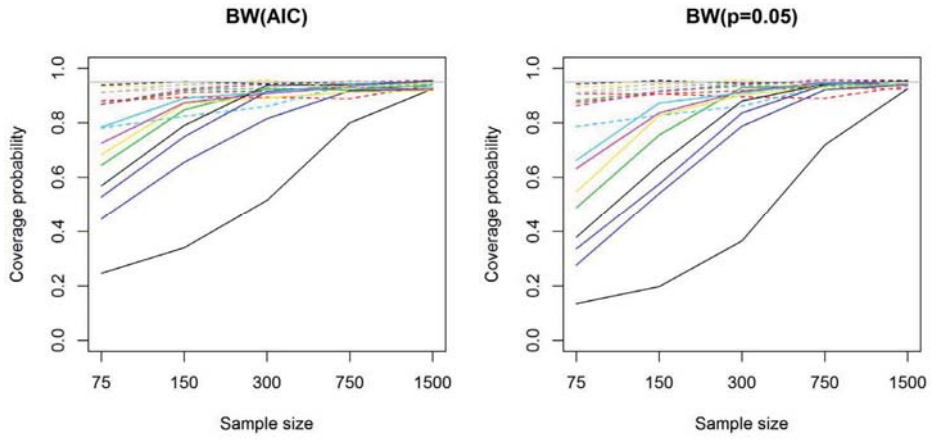
RMSE of regression coefficients, unconditional



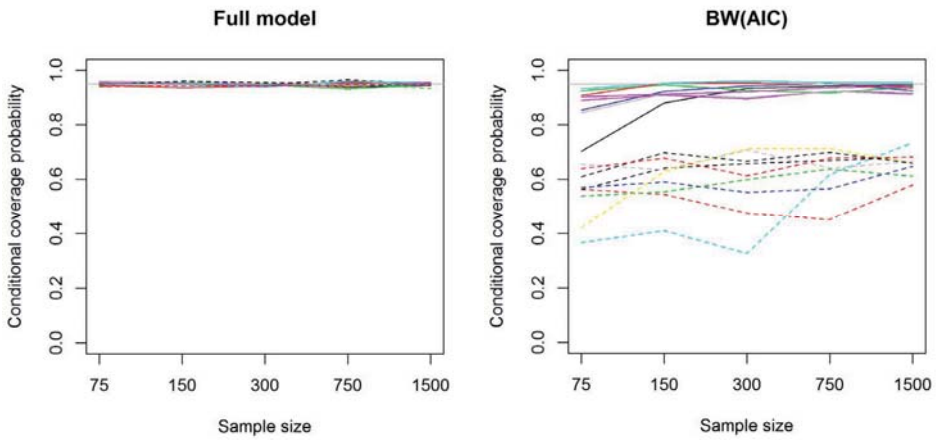
RMSE of regression coefficients, unconditional



Coverage of 95% CI for β, unconditional

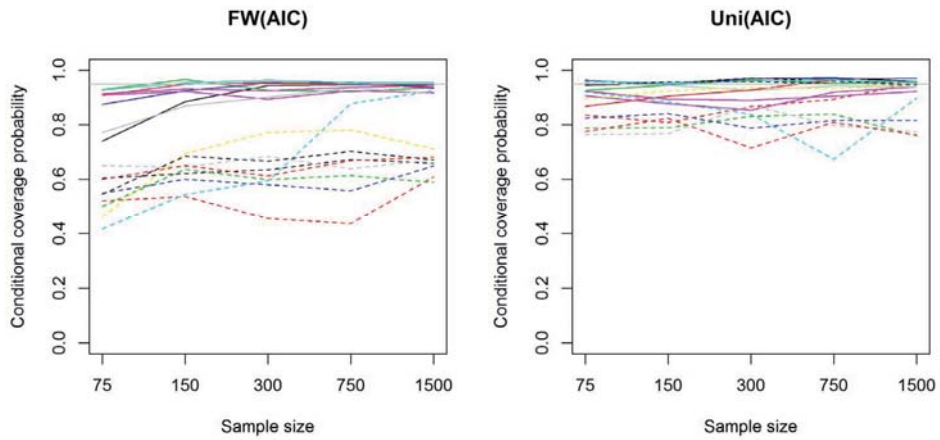


Coverage of 95% CI for β, conditional



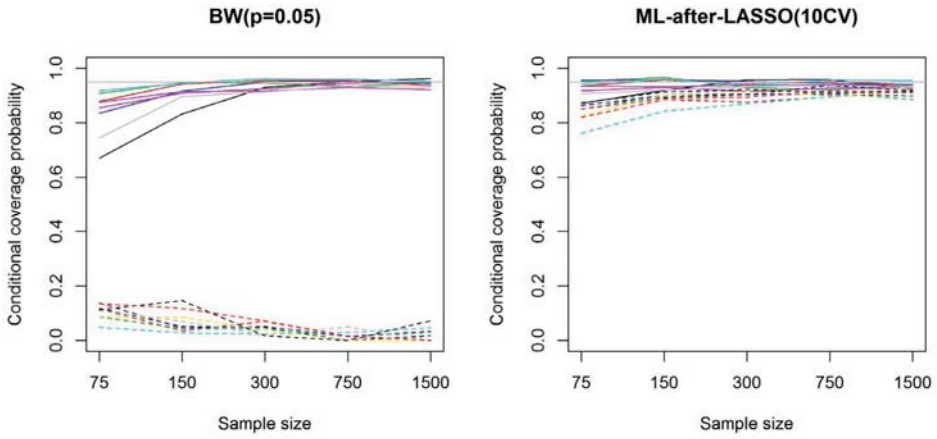
Conditional coverage for 'null' variables: how often selected and non-significant?
For BW(AIC) this happens in >50%.

Coverage of 95% CI for β, conditional



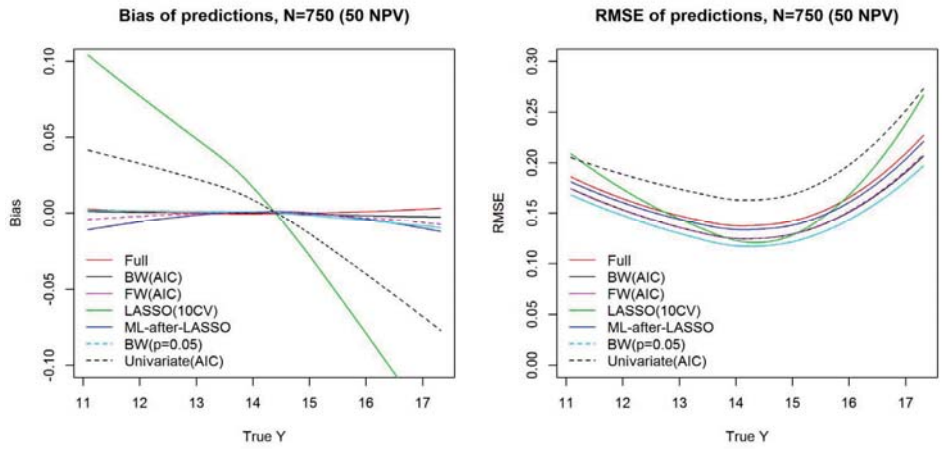
Conditional coverage for 'null' variables: how often selected and non-significant?

Coverage of 95% CI for β, conditional

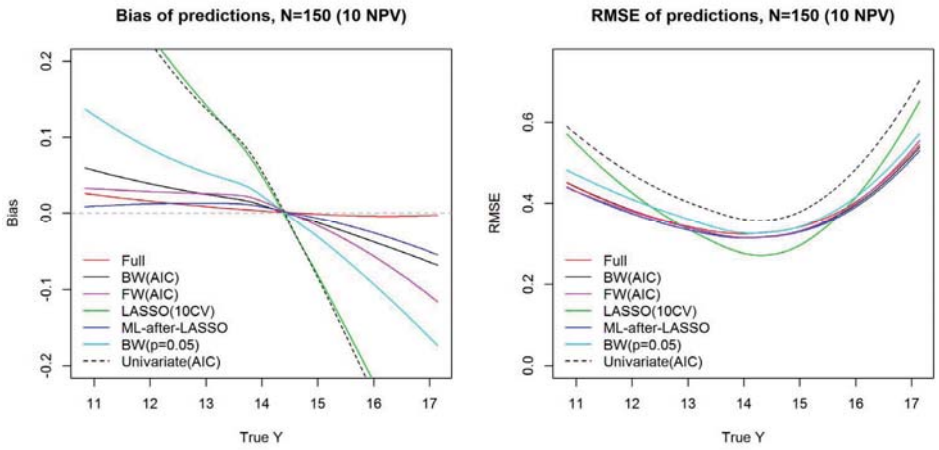


Conditional coverage for 'null' variables: how often selected and non-significant?
Of course, for BW($p = 0.05$) this happens only in 5%.

Accuracy of predictions



Accuracy of predictions

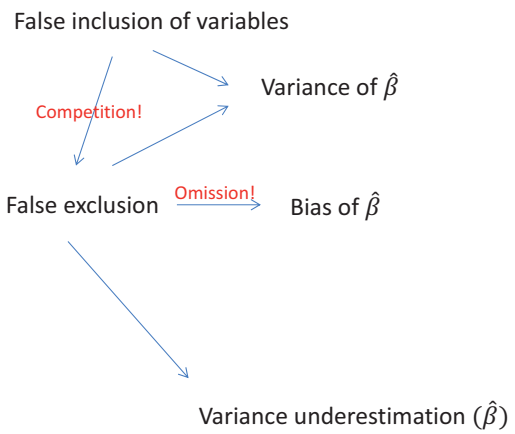


A network of dependencies…

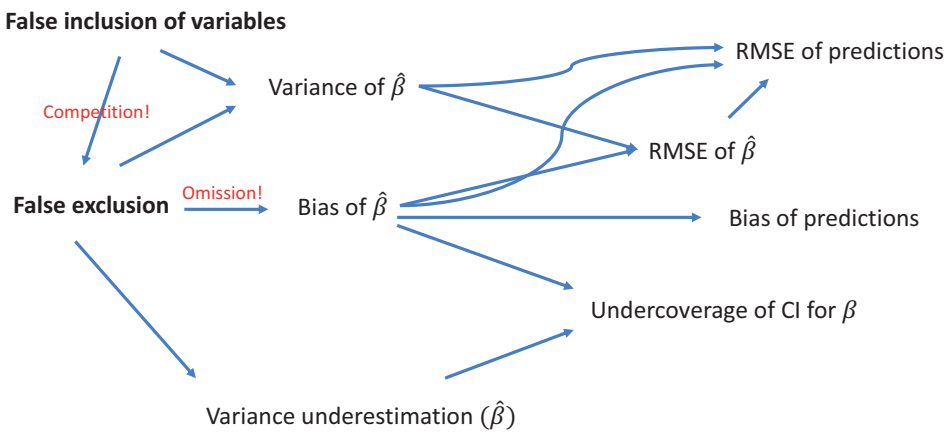
False inclusion of variables

False exclusion

A network of dependencies…



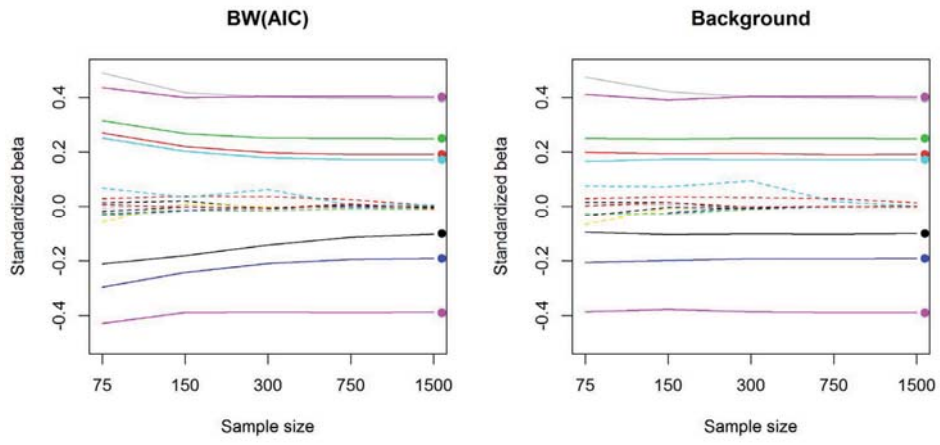
A network of dependencies…



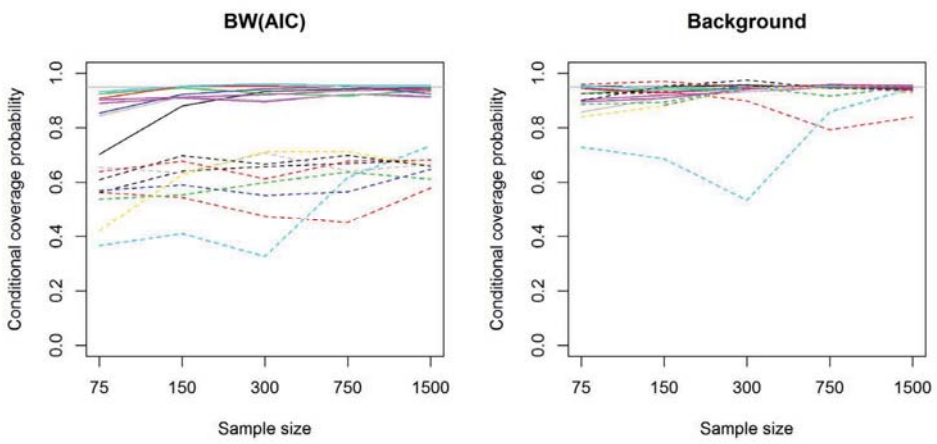
Using background knowledge

- Suppose, background knowledge is available, e.g., from a former study of equal size.
- One could simulate this background knowledge by first drawing the 'former study' to select variables, then drawing the 'actual study' to estimate effects.

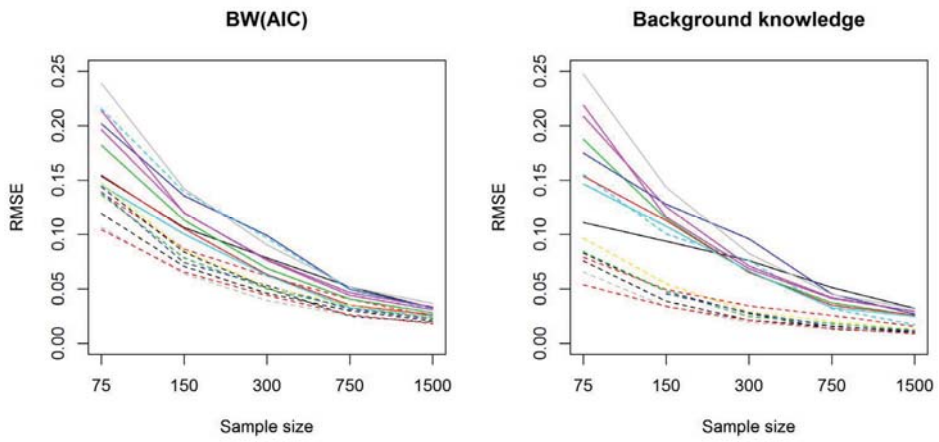
Using background knowledge: conditional regression coefficients



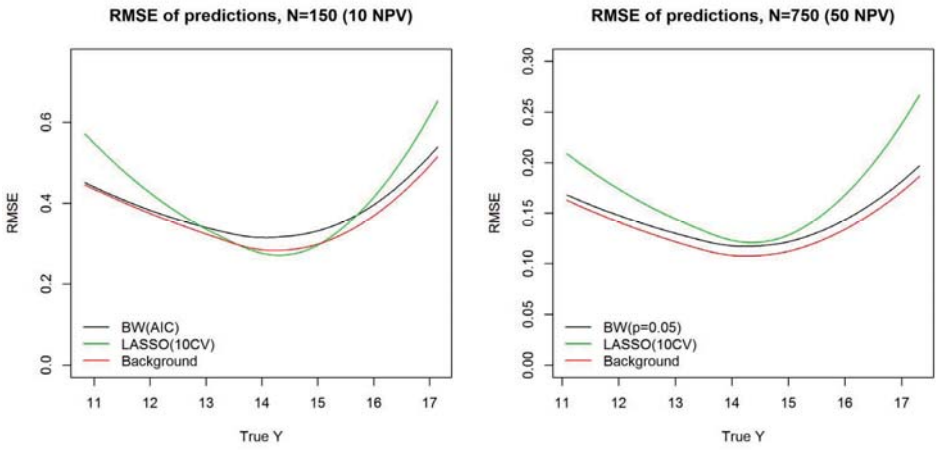
Using background knowledge: conditional coverage



Using background knowledge: unconditional RMSE



Using background knowledge: bias and RMSE of predictions



Summary from simulation study

- Careful interpretation of conditional and unconditional performance!
- E.g. conditional coverage – not meaningful for variables selected in 5%.
- Variable selection methods have been described with 'bias away from zero', but this concerns the conditional bias only.
- Unconditionally, there is bias towards 0.
- Univariate filtering results strongly depending on correlation structure!

Summary from simulation study

- For large samples (> 50 NPV), BW(0.05) dominates all other methods in predictive accuracy.
- It is close to BIC – discover the true model if it is in the scope of models evaluated.
- BW works if true positive rate (TPR) is high for 'true effects' and false positive rate (FPR) is low for 'null effects'.
- Therefore, bootstrap inclusion frequencies (BIFs) may provide a guide towards whether we can trust the best BW model:
  - BIFs should be routinely computed and reported,
  - report also performance of 'second-line' models,
  - don't trust a single model if selection is not sure.

Summary from simulation study

- Forward selection inferior to backward elimination.
- Lasso performs well in the 'center', but shrinks towards the mean (pessimistic).
- Lasso – problem with interpretability.
- Background knowledge improves conditional measures and predictive accuracy because selection and estimation are disentangled.

Summary from simulation study

- Data-driven selection is a bad idea with small samples.
- Better to work with simple, defendable, fixed models.

Part II-2: THE PROOF: CASE STUDIES

(from http://barnraisersllc.com/2015/08/10-compelling-characteristics-of-great-case-studies/)

Consulting situations

- 'We would like to approximate the proportion of body fat by simple anthropometric measures.'
- 'We want a prediction model for recurrent venous thromboembolism. Many risk factors were previously described, but the model should be clinically applicaple for making therapy decisions. Can you please develop a parsimonious model?'
- 'We want a prediction model for survival after cervical cancer diagnosis. We know our predictors. There are only few events.'

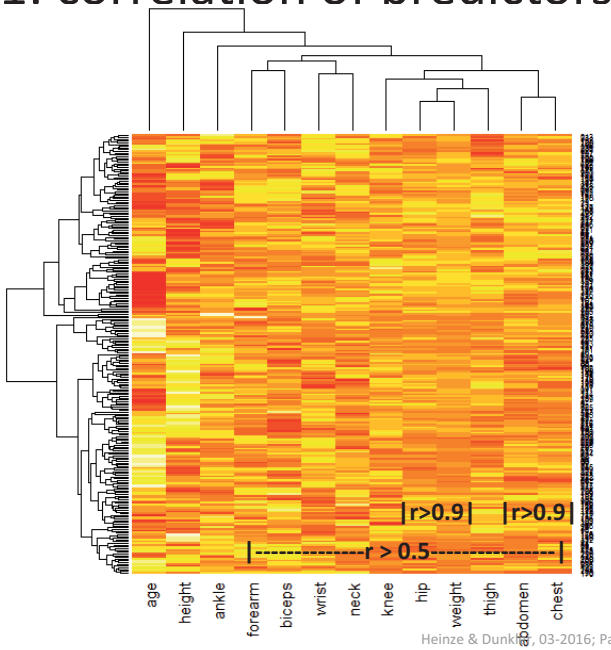
Case study 1: body fat approximation

- 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
- First goal: approximation of %body fat

www.freedieting.com

Case study 1: correlation of predictors

Correlations between predictor variables are quite high:



Case study 1: selection by backward(AIC)

```
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;
```

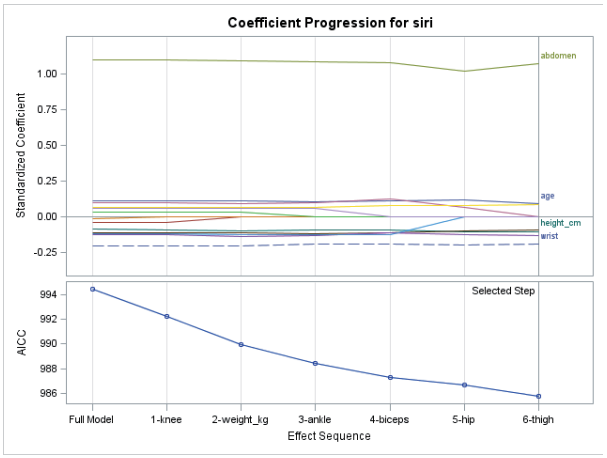

Case study 1: selection by backward(AIC)

```
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;
```

Root MSE	4.23144
Dependent Mean	19.08685
R-Square	0.7488
Adj R-Sq	0.7416
AIC	985.02609
AICC	985.77298
SBC	760.22971

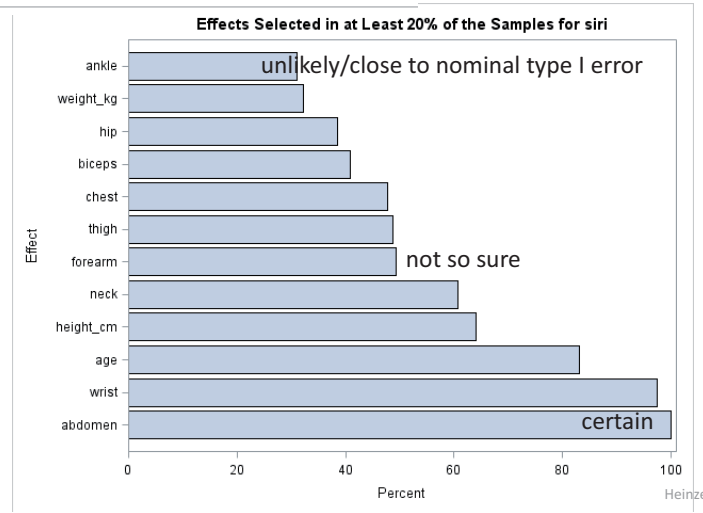
Parameter Estimates

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 1: 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 1: pairwise inclusion frequencies

```
proc surveyselect data = case1.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;
```

Table of sel_height by sel_weight (Frequency / Percent / Row Pct / Col Pct)

sel_height	sel_weight 0	1	Total
0	122 12.20 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

Table of sel_thigh by sel_biceps (Frequency / Percent / Row Pct / Col Pct)

sel_thigh	sel_biceps 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

(Cf. Sauerbrei and Schumacher, 1992)

Case study 1: pairwise inclusion frequencies

```
proc surveyselect data = case1.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;
```

Table of sel_height by sel_weight (Frequency / Percent / Row Pct / Col Pct)

sel_height	sel_weight 0	1	Total
0	122 12.20 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

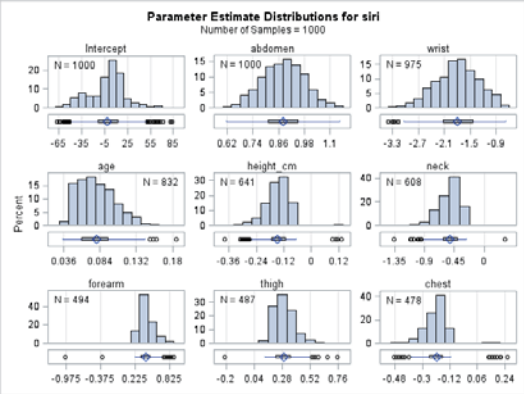
Competitive selection!

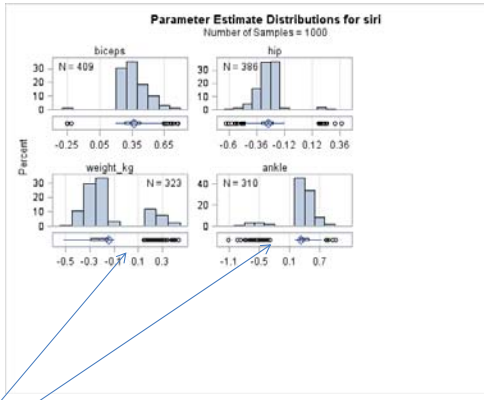
Table of sel_thigh by sel_biceps (Frequency / Percent / Row Pct / Col Pct)

sel_thigh	sel_biceps 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

(Cf. Sauerbrei and Schumacher, 1992)

Case study 1: Distribution of regression coefficients





Interesting: variables with 'negative' and 'positive' parts.
These are very unstable predictors.

Case study 1: bootstrap model averaging

Extremely low selection proportions:
Very unstable selection!

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

Case study 1: bootstrap model averaging

- Since many models are equally plausible, reporting a single model is problematic.
- Instead, report model-averaged predictors.
- SAS offers a 'refit' option to repeat the bootstrap with a reduced set of predictors (e.g. with BIF>0.2).
- In the refitting bootstrap, no selection is performed.
- The refitting-bootstrap standard errors are very close to refitting the original data with the selected variables.

Case study 1: an explanatory model

- In the textbook by Burnham & Anderson (2002), an interesting alternative model is developed based on 6 derived explanatory variables.

```
data bodyfat;
  set case1.bodyfat;
  allometry=log(weight_kg)/log(height_cm);
  beergut=abdomen/chest;
  heavyset=(knee*wrist*ankle)**(1/3)/height_cm;
  fleshiness=(biceps*thigh*forearm
            /(knee*wrist*ankle))**(1/3);
  age_stand=(age-44.88048)/12.62702;
  age_stand2=age_stand**2;
  run;

proc corr data=bodyfat;
  var allometry beergut heavyset fleshiness age_stand age_stand2 siri;
  run;
```


Case study 1: an explanatory model

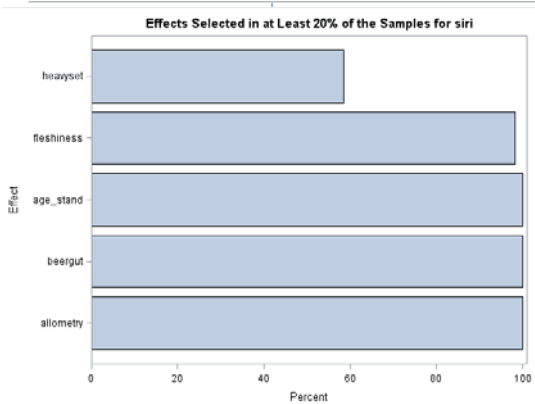
- In the textbook by Burnham & Anderson (2002), an interesting alternative model is developed based on 6 derived explanatory variables.

Pearson Correlation Coefficients, N = 251
Prob > |r| under H0: Rho=0

	allometry	beergut	heavyset	fleshiness	age_stand	age_stand2
allometry	1.00000	0.52908 <.0001	0.62691 <.0001	0.47580 <.0001	0.04177 0.5100	0.02851 0.6531
beergut	0.52908 <.0001	1.00000	0.34206 <.0001	0.20314 0.0012	0.24357 <.0001	0.04844 0.4448
heavyset	0.62691 <.0001	0.34206 <.0001	1.00000	0.07552 0.2332	0.22936 0.0002	0.17797 0.0047
fleshiness	0.47580 <.0001	0.20314 0.0012	0.07552 0.2332	1.00000	-0.21279 0.0007	-0.16023 0.0110
age_stand	0.04177 0.5100	0.24357 <.0001	0.22936 0.0002	-0.21279 0.0007	1.00000	0.22617 0.0003
age_stand2	0.02851 0.6531	0.04844 0.4448	0.17797 0.0047	-0.16023 0.0110	0.22617 0.0003	1.00000

Case study 1: an explanatory model

```
proc reg data=bodyfat;
  model siri=allometry beergut heavyset fleshiness age_stand age_stand2;
  run;

proc glmselect data=bodyfat plots=all;
  model siri=allometry beergut heavyset
        fleshiness age_stand age_stand2
       /selection=backward select=aicc details=all;
  modelaverage nsamples=1000;
  run;
```



Root MSE	4.81838	R-Square	0.6729
Dependent Mean	19.08685	Adj R-Sq	0.6649
Coeff Var	25.24450		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-182.35105	10.11316	-18.03	<.0001
allometry		1	109.22704	19.34826	5.65	<.0001
beergut		1	71.80432	8.15596	8.80	<.0001
heavyset		1	111.39222	65.59560	1.70	0.0908
fleshiness		1	18.40643	5.24758	3.51	0.0005
age_stand		1	1.64269	0.33788	4.86	<.0001
age_stand2		1	-0.03457	0.25583	-0.14	0.8926

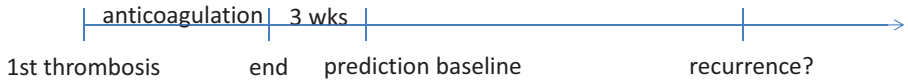
Case study 1: an explanatory model

- Top two models selected in 86%
- Top three in 92.1%
- Debatable variable: heavyset (P=0.09),
- Irrelevant: age^2 (P=0.89)

Model Selection Frequency				
Times Selected	Selection Percentage	Number of Effects	Frequency Score	Effects in Model
514	51.40	6	514.9	Intercept allometry beergut heavyset fleshiness age_stand
342	34.20	5	343.0	Intercept allometry beergut fleshiness age_stand
65	6.50	7	65.81	Intercept allometry beergut heavyset fleshiness age_stand age_stand2
61	6.10	6	61.85	Intercept allometry beergut fleshiness age_stand age_stand2
10	1.00	4	11.00	Intercept allometry beergut age_stand
5	0.50	5	5.92	Intercept allometry beergut heavyset age_stand
2	0.20	5	2.83	Intercept allometry beergut age_stand age_stand2
1	0.10	6	1.78	Intercept allometry beergut heavyset fleshiness age_stand2

Case study 2: Prediction of recurrence of venous thromboembolism

- The question: 'We want a prediction model for **recurrent venous thromboembolism**. Many risk factors were previously described, but the model should be clinically applicaple for making therapy decisions. Can you please develop a parsimonious model?'



- Patients at high risk for recurrence should continuously receive anticoagulation therapy,
- In patients at low risk for recurrence, no therapy should be given because of increased bleeding risk.
- The strategy: selection by AIC, shrinkage correction.

Case study 2: Prediction of recurrence of venous thromboembolism

- **The data set: AUREC, a prospective observational study.**
  - 929 patients included 3 weeks after end of anticoagulation therapy after first thrombosis
  - median follow-up for 30.5 months
  - 147 recurrence events
  - 8 risk factors (9DF, EPV=16.3)
- **Risk factors:**
  - **Sex** (males [60%] are at higher risk)
  - **D-Dimer** (363, 232-568) → log2
  - **Location of first thrombosis** (distal 18%/proximal 35%/pulmonary embolism 47%)
  - **BMI** (24-30), **Age** (44-63)
  - **Duration of anticoagulation therapy** (7wk, 5-9)
  - **Factor V Leiden** (23%), **Factor II mutation** (4.8%)

public-domain version contained in `shrink` R package (`deepvein`)

Case study 2: risk factors and global model

```
> library(survival)
> fitfull <- coxph(Surv(time, status) ~  sex + loc + log2ddim +  durther + fvleid +
+                  + fiimut + age + bmi, data = deepvein, x = TRUE)
> summary(fitfull)
Call:
coxph(formula = Surv(time, status) ~ sex + loc + log2ddim + durther +
    fvleid + +fiimut + age + bmi, data = deepvein, x = TRUE)

  n= 929, number of events= 147

                    coef exp(coef)  se(coef)      z Pr(>|z|)
sex.male        0.495927  1.642019  0.189718  2.614  0.00895 **
loc.distal     -0.905095  0.404504  0.311078 -2.910  0.00362 **
loc.proximal   -0.179351  0.835813  0.180336 -0.995  0.31996
log2ddim        0.219517  1.245475  0.085739  2.560  0.01046 *
durther         0.021881  1.022122  0.023681  0.924  0.35550
fvleid.present -0.108886  0.896833  0.194228 -0.561  0.57506
fiimut.present -0.162573  0.849954  0.390499 -0.416  0.67718
age            -0.003973  0.996035  0.006583 -0.603  0.54622
bmi             0.005865  1.005883  0.019051  0.308  0.75817
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Case study 2: backward (AIC)

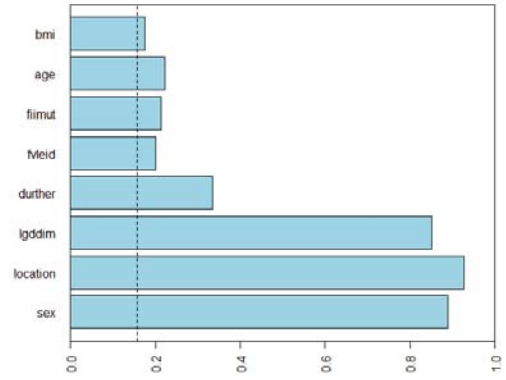
```
> bw.aic<-step(fitfull, direction="backward", k=2, trace=0)
> summary(bw.aic)
Call:
coxph(formula = Surv(time, status) ~ sex + loc + log2ddim, data = deepvein,
    x = TRUE)

  n= 929, number of events= 147

                coef exp(coef) se(coef)      z Pr(>|z|)
sex.male      0.49091   1.63380  0.18473  2.657  0.00787 **
loc.distal   -0.92237   0.39758  0.31007 -2.975  0.00293 **
loc.proximal -0.20505   0.81461  0.17867 -1.148  0.25112
log2ddim      0.21879   1.24457  0.08543  2.561  0.01043 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ this was the final model.

[For selection with $\alpha_2 = 0.05$, use `k=qchisq(1-0.05,1)`.]

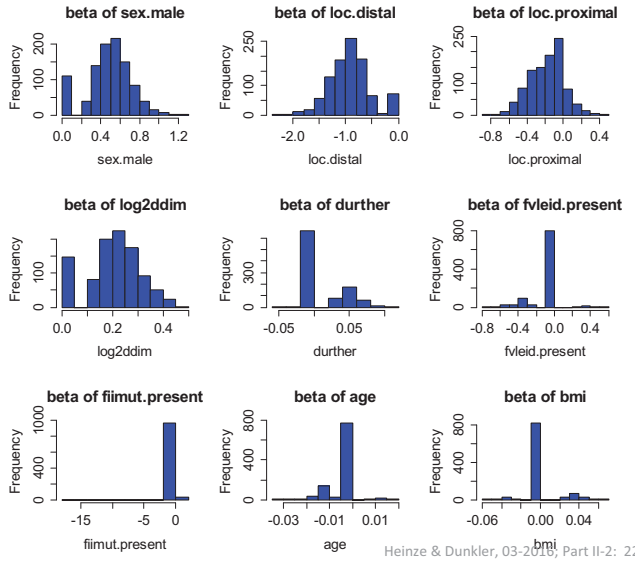
Case study 2: BIFs

```
# selection stability

set.seed(7123981)
B<-5000
beta.boot<-matrix(0,B,length(coef(fitfull)))
colnames(beta.boot)<-names(coef(fitfull))
terms.boot<-matrix(FALSE,B,length(attr(fitfull$terms,"term.labels")))
colnames(terms.boot)<-attr(fitfull$terms,"term.labels")

for(i in 1:B) {
  ind<-sample(1:nrow(deepvein), repl=TRUE)  # draw a bootstrap sample
  fitb<-coxph(Surv(time, status) ~ sex + loc + log2ddim +  durther + fvleid +
                + fiimut + age + bmi, data = deepvein[ind,], x = TRUE)
  fitb.aic<-step(fitb, direction="backward", k=2, trace=0)
  beta.boot[i,names(coef(fitb.bw05))]<-coef(fitb.bw05)   # memorize coefficients
  terms.boot[i,attr(fitb.bw05$terms,"term.labels")]<-TRUE  # record selection
}

BIF<-apply(terms.boot,2,function(X) mean(X))
```



Case study 2: bootstrapped coefficients

```
par(mfrow=c(3,3))
for(i in 1:ncol(beta.boot)) {
    hist(beta.boot[,i], breaks=11,
         main=paste("beta of", colnames(beta.boot)[i]),
         col="blue",
         xlab=colnames(beta.boot)[i])
}
par(mfrow=c(1,1))
```



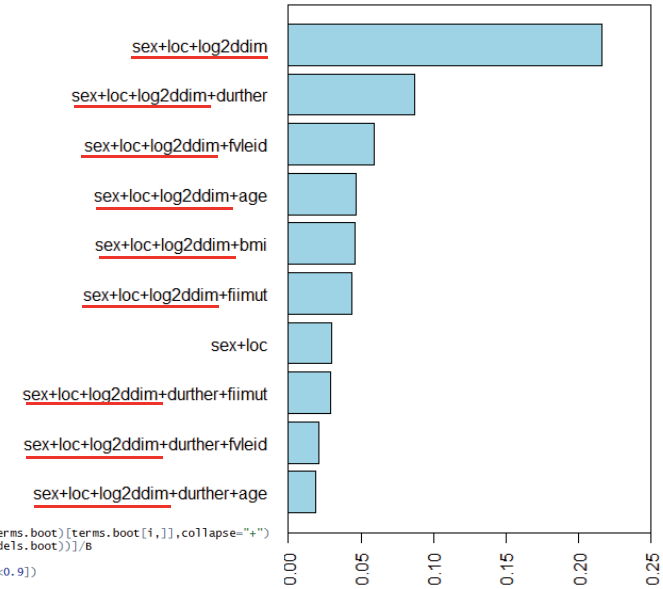
including 0's!

Case study 2: Model selection frequencies

Top 10 models:

9 of them contain

sex, loc, log2ddim




```
models.boot<-character(B)
for(i in 1:B) models.boot[i]<-paste(colnames(terms.boot)[terms.boot[i,]],collapse="+")
tab.models<-table(models.boot)[order(-table(models.boot))]/B
cumweight<-cumsum(tab.models)
topn.mod<-max((1:length(tab.models))[cumweight<0.9])
par(mar=c(5.1,15.1,1.1,2.1),las=2)
barplot(tab.models[10:1], horiz=TRUE, xlim=c(0,0.25), col="lightblue")
box()
```

Case study 2: further refinement

```
> ## recoding of loc
> deepvein$loc_proximal<-(deepvein$loc==".proximal")
> deepvein$loc_distal<-(deepvein$loc==".distal")
>
> bw2.aic<-step(coxph(data=deepvein, Surv(time,status)~sex + loc_proximal + loc_distal +
+                 + log2ddim +  durther + fvleid +
+                 + fiimut + age + bmi, x=TRUE), direction="backward", k=2, trace=0)
> summary(bw2.aic)
Call:
coxph(formula = Surv(time, status) ~ sex + loc_distal + log2ddim,
    data = deepvein, x = TRUE)

  n= 929, number of events= 147

                   coef exp(coef) se(coef)      z Pr(>|z|)
sex.male        0.49535   1.64107  0.18496  2.678   0.0074 **
loc_distalTRUE -0.84053   0.43148  0.30277 -2.776   0.0055 **
log2ddim        0.20392   1.22621  0.08483  2.404   0.0162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ Use dummies as 'standard' variables, collapse categories by selection.
→ Locations 'proximal' and 'pulmonary embolism' are collapsed.

Case study 2: shrinkage factor estimation

```
> library(shrink)
> shrink(bw.aic, type="global")
Shrinkage Factors (type=global, method=jackknife):
[1] 0.8076362

Shrunken Regression Coefficients:
    sex.male   loc.distal loc.proximal     log2ddim
   0.3964779   -0.7449390   -0.1656066    0.1767045
>
> shrink(bw.aic, type="parameterwise")
Shrinkage Factors (type=parameterwise, method=jackknife):
    sex.male   loc.distal loc.proximal     log2ddim
   0.8351074    0.8393993    0.1321006    0.7321036

Shrunken Regression Coefficients:
    sex.male   loc.distal loc.proximal     log2ddim
  0.40996379  -0.77423621  -0.02708736   0.16017851
>
>
> shrink(bw2.aic, type="parameterwise")
Shrinkage Factors (type=parameterwise, method=jackknife):
      sex.male loc_distalTRUE       log2ddim
     0.8317218      0.8975722      0.7700839

Shrunken Regression Coefficients:
      sex.male loc_distalTRUE       log2ddim
     0.4119946     -0.7544400      0.1570393
```

← Global shrinkage factor

← Parameterwise shr. factors

← Parameterwise shr. factors

(Dunkler et al., 2016)

Case study 2: further aspects

- Global shrinkage factor was used.
- Clinical practicability: 3 simple, easily available clinical parameters.
- Study was published in *Circulation.* (Eichinger et al, 2010)
- Presented as nomogram and as web calculator.

- First prediction model for recurrent thromboembolism.

- External validation of the model suggested age as additional predictor.
  In our study, age was an 'explanatory', but not a 'predictor'.

- Follow-up paper on dynamic prediction. (Eichinger et al, 2014)

Case study 3: cervical cancer prognosis

- The question: 'We want a prediction model for survival after cervical cancer diagnosis. We know our predictors. There are only few events.'

- The data set: baseline and follow-up data from 692 consecutive patients diagnosed with cervical cancer from two centers (Vienna, Innsbruck)

- Follow-up: median 46 months

Case study 3: cervical cancer prognosis

- Risk factors:
  - FIGO stage (I, II, III, IV) (3df)
  - Tumour size (<2cm, >2cm)
  - Age
  - Histologic subtype (squamous cell carcinoma, adenocarcinoma, other) (2df)
  - Proportion positive lymph nodes (2df)
  - Parametrical involvement (yes/no)

- 528 patients had all these variables available
- 77 deaths → EPV=7.7

Case study 3: cervical cancer prognosis

- Because of the critical EPV (7.7), we did not attempt to perform any variable selection.
- Instead, L2-penalization (ridge regression) was used.

- Clinical collaborators asked for dividing the data into 'training' and 'validation' sets.
- I said: 'No way!'

- Bootstrap validation revealed a decent performance of the model:

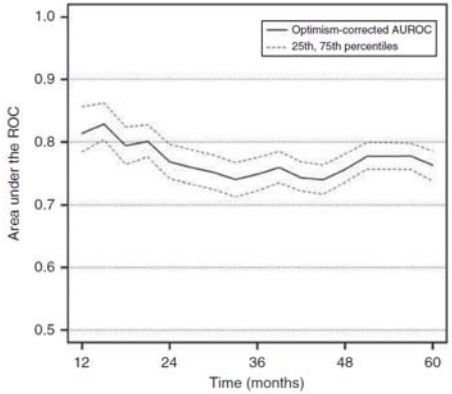
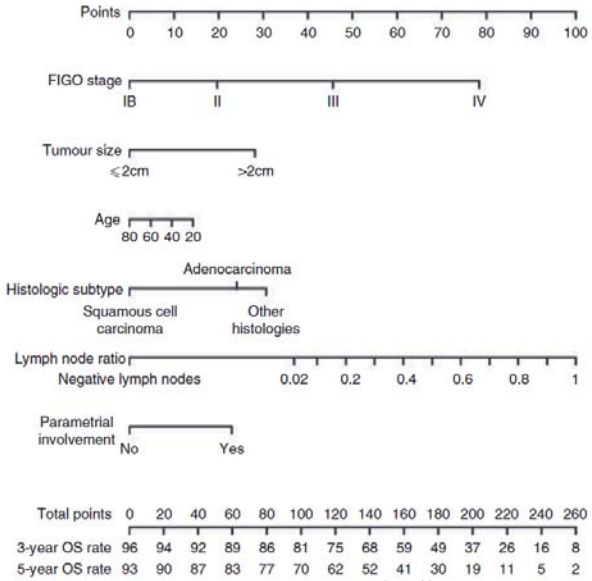


Figure 3 Time-dependent discrimination curves. Optimism-corrected area under the ROC (AUROC): median over 1000 bootstrap replicates shown as solid line, dashed lines denote 25th and 75th percentiles.

(Polterauer et al, Br J Cancer 2012)

Case study 3: cervical cancer prognosis

- The model was implemented as nomogram and web calculator.
- Nomogram nicely shows the relative importance of the prognostic factors.



(Polterauer et al, Br J Cancer 2012)

Case study 3: cervical cancer prognosis

- Recently, the prognosis model was validated using data from an Australian center.

- Confirms the performance estimate (c-index) presented in paper.

- → good idea to use penalized model

Summary of case studies

- Variable selection may sometimes be an option, sometimes not.
- Variable selection should always be accompanied by stability investigation.
- While AIC selection provides a useful point of reference, size of models can be accommodated to practical needs.
- 'Significance' level for selection can be used to control size of models.
- For good explanatory models, use substance matter knowledge (or brains).

Part II-3:

The 'best' procedure

- Depends on information provided and knowledge desired:
- Small data set – large data set?
- Many unknowns or few?

Go for a good enough model?

- No or mild selection (AIC) with small to moderate data sets.
- AIC provides the best approximating model among a candidate set of models.

Go for the 'true' model?

- More stringent (BW/p-value) selection in large samples.

Importance of background knowledge

- Incorporating background knowledge is like increasing the sample size.
- Can be seen (or even implemented) as a Bayesian procedure.
- Select in one data set – estimate in another.
- Avoids the overestimation bias conditional on selection.
- Background knowledge is also important for preselecting variables, for specifying their coding, interactions, transformations, …

Some recommendations: after selection

- Regression coefficients, confidence intervals and p-values conditional on the selected model often biased/too optimistic.
- Important: is there one dominating model?
- Stability investigation by bootstrap!
- In large samples, the optimism is often not too severe (simulation).

Estimation/correction of optimism:

- Shrinkage methods (Sauerbrei, 1999; Dunkler et al, 2016)
- Model averaging (Buckland et al, 1997)
- Bootstrap resampling (Sauerbrei et al, 2014)
- Unfortunately these methods are still missing in standard packages (SPSS)!

Our own strategy

- In our environment, we work a lot with **real-life data sets**.
- We try to get as much information from of our clinical collaborators as possible to determine a **working set of variables**.
- We **do not select** variables in **small** samples.
- Otherwise, we recommend **backward elimination**.
- In backward elemination, α should be set according to the sample size/events per variable.
- **Stability investigation** based on the bootstrap is helpful.

- Background knowledge ≠ univariate selection.

Implementations: SAS and SPSS

What	PROC GLMSELECT	PROC REG	PROC LOGISTIC PROC PHREG	%ABE macro	SPSS
Backward	Yes	Yes	Yes	Yes	Yes
Forward	Yes	Yes	Yes	No	Yes
Stepwise forward	Yes	Yes	Yes	No	Yes
Stepwise backward	No	No	No	No	No
Augmented backward	No	No	No	Yes	No
LASSO	Yes	No	No	No	No
Multi-model inference	(Yes)	No	No	No	No
Bootstrap stability investigation	Yes	No	No	(No)	No(!)
Linear	Yes	Yes	No	Yes	Yes
Logistic	No	No	Yes (LOGISTIC)	Yes	Yes
Cox	No	No	Yes (PHREG)	Yes	Yes

Implementations: R

What	lm(), glm(), survival	step()	mfp	glmulti	glmnet, penalized	rms
Backward	No	Yes	No	No	No	Yes
Forward	No	Yes	No	No	No	No
Stepwise forward	No	Yes	No	No	No	No
Stepwise backward	No	No	Yes	No	No	No
All subsets/other	No	No	No	Yes	No	No
LASSO	No	No	No	No	Yes	No
Multi-model inference	No	No	No	Yes	No	(Yes)
Bootstrap stability investigation	No	No	No	No	No	(Yes)
Linear	lm()	Yes	Yes	Yes	Yes	Yes
Logistic	glm()	Yes	Yes	Yes	Yes	Yes
Cox	coxph()	Yes	Yes	?	Yes	Yes

Software implementations

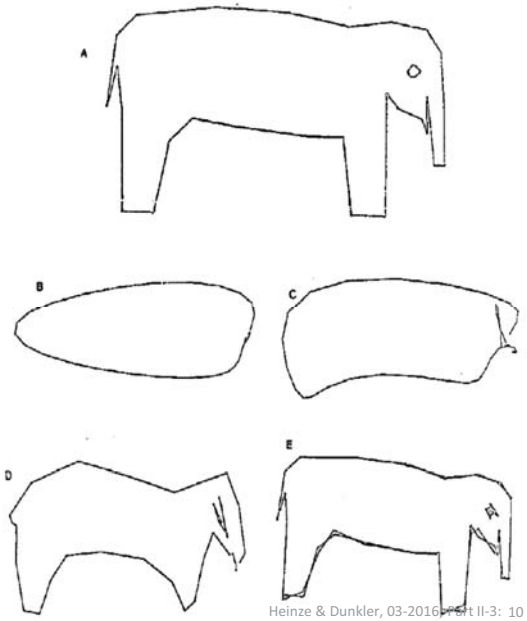
- Background knowledge is not implemented in any standard software.

Principle of Parsimony

(https://www.zoovienna.at/news/elefantenbaby/)

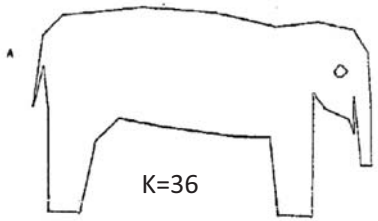
Principle of parsimony

- Avoid overfitting to achieve a good model fit.
- Wel 1975: ‘How many parameters does it take to fit an elephant?’
- ‘E may not satisfy the third-grade art teacher, but would carry most chemical engineers into preliminary design.”



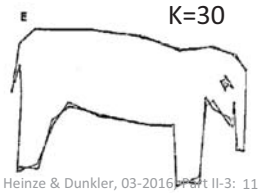
Principle of parsimony

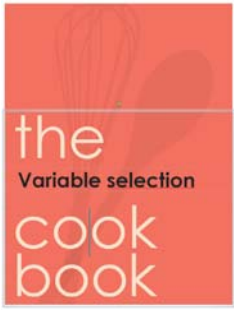
- Avoid overfitting to achieve a good model fit.



➔ Principle of parsimony should lead to the smallest possible number of parameters for adequate representation of the data.

➔ This number of parameters might be different for explanatory and predictive purposes.





Recipe for disaster

- Prepare a long list of poorly conceived predictors.
- Add only small n.
- Mix together in an extensive iterative data dredging.
- Select the model with the smallest p-values.
- Present this final model without further considerations.

Bon appétit!

References -
‘Variable selection – a review and recommendations for the practicing statistician’
by Georg Heinze and Daniela Dunkler, March 2016

All references are only stated at their first appearance.

Part I-1: Philosophy

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199-231. doi:DOI 10.1214/ss/1009213726

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*: Springer.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical Statistics* (1 ed.). Boca Raton: Chapman and Hall/CRC.

Dunkler, D., Plischke, M., Leffondré, K., & Heinze, G. (2014). Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models. *PLoS ONE, 9*(11), e113677. doi:doi:10.1371/journal.pone.0113677

SAS %ABE macro available at: http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/abe/

Eichinger, S., Heinze, G., Jandeck, L., M., & Kyrle, P., A. (2010). Risk Assessment of Recurrence in Patients With Unprovoked Deep Vein Thrombosis or Pulmonary Embolism. The Vienna Prediction Model. *Circulation, 121*(14), 1630-1636.

Good, D. M., Zurbig, P., Argiles, A., Bauer, H. W., Behrens, G., Coon, J. J., . . . Schmitt-Kopplin, P. (2010). Naturally Occurring Human Urinary Peptides for Use in Diagnosis of Chronic Kidney Disease. *Molecular & Cellular Proteomics, 9*(11), 2424-2437. doi:10.1074/mcp.M110.001917

Royston, P., & Sauerbrei, W. (2008). *Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester: John Wiley & Sons, Ltd.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289-310.

Steyerberg, E., W. (2009). *Clinical Prediction Models. A practical approach to developement, validation, and updating*: Springer.

Part 1-2: Prerequisites

Dunkler, D., Sauerbrei, W., & Heinze, G. (2016). Global, parameterwise and joint shrinkage factor estimation. *Journal of Statistical Software, 69*(8), 1-19. doi:doi:10.18637/jss.v069.i08

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology, 29*(1), 158-167. doi:10.1093/ije/29.1.158

Harrell Jr., F. E. (2001). *Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis*. New York, Berlin, Heidelberg: Springer Verlag.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
Lee, P. H. (2014). Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *Journal of Epidemiology, 24*(2), 161-167.
Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology, 138*(11), 923-936.
Meinshausen, N., & Buhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 72*, 417-473.
Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82*(4), 669-688. doi:10.1093/biomet/82.4.669
Sauerbrei, W. (1999). The Use of Resampling Methods to Simplify Regression Models in Medical Statistics. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 48*(3), 313-329. doi:10.1111/1467-9876.00155
Sauerbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the cox regression model. *Statistics in Medicine, 11*(16), 2093-2109. doi:10.1002/sim.4780111607
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, 58*(1), 267-288.
Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(3), 273-282.
van der Ploeg, T., Austin, P., C., & Steyerberg, E., W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology, 4*(137).
Verweij, P., J.M., & Houwelingen Hans C., v. (1993). Cross-validation in survival analysis. Statistics in Medicine, 12(24), 2305-2314.

Part I-3: Variable selection methods

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. *Biometrics, 53*(2), 603-618.
Glymour, M. M., Weuve, J., & Chen, J. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: Measurement, selection, and bias. *Neuropsychology Review, 18*(3), 194-213.
Harrell Jr., F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*(4), 361-387.
Hosmer, D., W.Jr., Lemeshow, S., & May, S. (1999). Applied Survival Analysis: Regression Modeling of Time to Event Data: Chapter 5 Model Developement *Applied Survival Analysis: Regression Modeling of Time to Event Data* (pp. 1-416): John Wiley & Sons.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2 ed.). New York: John Wiley & Sons, Inc.
Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine, 28*(9), 1420-1423.
VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics, 67*(4), 1406-1413. doi:10.1111/j.1541-0420.2011.01619.x

Part II-1: Consequences

Binder, H., Sauerbrei, W., & Royston, P. (2011). *Multivariable model-building with continuous covariates: 1. Performance measures and simulation design*. Technical Report FDM-Preprint 105, University of Freiburg, Germany.

Part II-2: Case studies

Eichinger, S., Heinze, G., & Kyrle, P. A. (2014). d-Dimer Levels Over Time and the Risk of Recurrent Venous Thromboembolism: An Update of the Vienna Prediction Model. *Journal of the American Heart Association, 3:e000467*. doi:doi:10.1161/JAHA.113.000467

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education, 4*(1).

Polterauer, S., Grimm, C., Hofstetter, G., Concin, N., Nattern, C., Sturdza, A., . . . Heinze, G. (2012). Nomogram prediction for overall survival of patients diagnosed with cervical cancer. *British Journal of Cancer, 107*, 918-824.

Part II-3: Recommendations

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., & Binder, H. (2015). On stability issues in deriving multivariable regression models. *Biometrical Journal, 57*(4), 531-555. doi:10.1002/bimj.201300222

Wel, J. (1975). Least squares fitting of an elephant. *Chemtech, Feb.*, 128-129.