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
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 do we mean by a statistical model?

- A set of probability distributions on the sample space $S$.
  (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 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:**
- **Gender:**
- **Total Cholesterol:**
- **HDL Cholesterol:**
- **Smoker:**
- **Systolic Blood Pressure:**

Are you currently on any medication to treat high blood pressure.

[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.

(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**

(Predictive models)

- **Interest in accurate predictions of future observations.**
- **No concern about causality and confounding (association).**
- **Medicine: prognostic models versus predictive models.**

(Explanatory models)

- **Strong theory ➔ interest in coefficients and inference.**
- **Testing and comparing existing causal theories.**
- **Medicine: often no strong theory, etiological models**

(Descriptive models)

- **capture the data structure parsimoniously: which factors affect the outcome and how?**

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?
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.

(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.’

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 $Y$.

(William of Ockham)

---

Example I: Prediction of recurrence of venous thromboembolism

- 14th century logician and Franciscan friar: ‘Pluralitas non est ponenda sine necessitate.’ (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)

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

- Support Vector Machine

(Good et al, 2010)
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
- Model estimation: maximum likelihood
- Likelihood and information-theoretic measures
- AIC and AICc
- Penalized likelihood
- Change-in-estimate criterion
- Hypothesis tests: Likelihood ratio, Score, Wald
- Resampling techniques
- Shrinkage
- Prior knowledge
- Bias-variance tradeoff
- Assumptions of models

Heinze & Dunkler, 03-2016; Part I-2: 21
Heinze & Dunkler, 03-2016; Part I-2: 22
Preselection of variables

- Subject matter knowledge
- Chronology
- Costs of collecting measurements
- Availability at time of model use
- Quality (measurement errors)
- Confounder criteria
- Availability in data set (missing values)
- Variability (rare categories)
- Preselection = Bayes!

Discussion with non-statistical collaborator!

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
- \( \text{Pr}(Y = 1) = \expit(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_k) = \exp(X\beta) / [1 + \exp(X\beta)] \)

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.
- \( \beta_k \) measures the ‘independent’ effect of \( X_k \).
- Fundamentally different in different models!
Interpretation of regression coefficients

• Consider the following models to explain %body fat:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>11.43160</td>
<td>-3.66</td>
<td>0.0084</td>
<td></td>
</tr>
<tr>
<td>abdomen_circ</td>
<td>Abdomen_circumference</td>
<td>1</td>
<td>0.07137</td>
<td>12.75</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>height_cm</td>
<td>Height in cm</td>
<td>1</td>
<td>-0.05581</td>
<td>-9.46</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>weight_kg</td>
<td>Weight in kg</td>
<td>1</td>
<td>0.06171</td>
<td>-1.55</td>
<td>0.1213</td>
<td></td>
</tr>
</tbody>
</table>

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’ \( \text{EPV} = E/K \), \( \text{NPV} = N/K \).

• Often cited minimum \( \text{EPV} = 10 \).

Events Per Variable (EPV)

• \( \text{EPV} = 10 \) (Harrell 2001, p. 61)
  – Number of candidate variables, not variables in the final model.
  – Should be considered as lower bound!

• Non-linearity, interactions, etc. \( \Rightarrow \) \( \text{EPV} \uparrow \).

• Prediction \( \Rightarrow \) \( \text{EPV} \uparrow \) (logistic regression \( \text{EPV} \) 20–50).

• Modern modeling techniques (random forests, neural networks, support vector machines) \( \Rightarrow \) 10 times \( \text{EPV} \) compared to logistic regression \( \Rightarrow \) \( \text{EPV} \uparrow \) (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 \( \beta \) such that \( L(\beta|X, Y) \Rightarrow \text{max!} \)
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 $\beta$’s in $M_2$ are forced to be 0.

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

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

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 $\beta_K$.
- If slope of $L_2$ is ‘steep’, $\beta_K \neq 0$ should be assumed.

- = Classical ‘forward’ test.

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 $\beta_K$ 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, \( \log(\text{likelihood}) = \text{entropy} = 0 \).

Information theory

\[ \text{entropy} \propto -\log(\text{probability}) \]

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

Ludwig Boltzmann, 1844-1906
Physicist and Philosopher

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

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

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

- This can be approximated by

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

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

$$K \ldots \text{number of parameters}$$

Small-sample correction

- For small data sets:

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

$$K \ldots \text{number of parameters}$$

$$N \ldots \text{sample size}$$

- 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

<table>
<thead>
<tr>
<th>Degrees of freedom difference</th>
<th>Equivalent p-value in LR test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.157</td>
</tr>
<tr>
<td>2</td>
<td>0.135</td>
</tr>
<tr>
<td>3</td>
<td>0.117</td>
</tr>
<tr>
<td>4</td>
<td>0.092</td>
</tr>
</tbody>
</table>

General: $1 - p\text{chisq}(2*df, df)$

Comparing 2 models with AIC

AIC

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

Evidence Ratios (ER)

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

- ER = ‘How much likelier is $M_j$ than $M_i$?’
Comparing R models with AIC

### AIC differences

<table>
<thead>
<tr>
<th>( \Delta_i )</th>
<th>1/ER</th>
<th>Level of empirical support for Model ( i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2</td>
<td>1-2.7</td>
<td>Substantial</td>
</tr>
<tr>
<td>4-7</td>
<td>7.4-33.1</td>
<td>Considerably less</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>&gt;148</td>
<td>Essentially none</td>
</tr>
</tbody>
</table>

### 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 \( BIC = -2 \log L + \log(N)K \)
- 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 \( \min(\text{events}, \text{non-events}) \)
- 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 \( \alpha = 0.032 \)
- \( \rightarrow \) 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.

Training sample

Test sample

Shrinkage methods

- Post-estimation shrinkage factor estimation
  - Verweij & Van Houwelingen 1993: global shrinkage factor \( c < 0.8 \) ➔ 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

The method(s)

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

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

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

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

$$Err(x_0) = E \left[ (Y - \hat{f}(x_0))^2 \mid X = x_0 \right]$$

$$= \sigma^2 + E \left[ \left( \hat{f}(x_0) - f(x_0) \right)^2 \right] + E \left[ \hat{f}(x_0) - E(\hat{f}(x_0)) \right]^2$$

$$= \sigma^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0))$$

Irreducible error

Bias

Variance

(Hastie, Tibshirani, and Friedman, 2009, p. 223)
To explain or to predict?

Expected prediction error = Irreducible error + Bias² + 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 $Y$.

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 $c$: CIE = $c' - c$.
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 $\beta_1 = 1.5, \beta_2 = 0.3$

A weak $\beta_2$:
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($\hat{\beta}_{\text{FULL}}$) = 0.67
RMSE($\hat{\beta}_{\text{BE}}$) = 0.65
Bias($\hat{\beta}_{\text{FULL}}$) = -0.03
Bias($\hat{\beta}_{\text{BE}}$) = +0.03

⇨ ‘Selection is good.’

Prior knowledge: simple illustrative simulations

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

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

A strong $\beta_2$:
Setting it to 0 will always push $\hat{\beta}_1$ away from its true value.

RMSE($\hat{\beta}_{\text{FULL}}$) = 0.68
RMSE($\hat{\beta}_{\text{BE}}$) = 0.67
Bias($\hat{\beta}_{\text{FULL}}$) = -0.33
Bias($\hat{\beta}_{\text{BE}}$) = +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’).

\[ X_2 \rightarrow X_1 \rightarrow Y \]

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

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

Pros and cons of univariate selection

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

- Problematic:
  - The univariate effect of \( X_1 \) on \( Y \) is \( a + bc \).

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos.</td>
<td>Pos.</td>
<td>Neg.</td>
<td>( X_1 ) falsely not selected (if ( a = -bc ))</td>
</tr>
<tr>
<td>0</td>
<td>Pos./Neg.</td>
<td>Pos./Neg.</td>
<td>( X_1 ) falsely selected.</td>
</tr>
<tr>
<td>Pos./neg</td>
<td>0</td>
<td>Pos./neg</td>
<td>( X_1 ) correctly selected (only if ( b = 0 ) or ( c = 0 )).</td>
</tr>
</tbody>
</table>

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 \( \text{AIC} < \text{AIC}_{\text{min}} + 2 \).
- Perform multi-model inference on the selected models.

In practice:
- Approximated by stepwise approaches!

(Burnham & Anderson, 2002)

Forward selection

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

Variant: Stepwise forward

- Select \( \alpha_1 \) and \( \alpha_2 \).
- 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 $\alpha_2$.
- Estimate full model.
- Repeat:
  - While least significant term has $p \geq \alpha_2$, remove it and re-estimate.

**Variant: Stepwise backward**
- Select $\alpha_1$ and $\alpha_2$.
- 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.

**Software:**
- SAS macro `%ABE`

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 $\alpha_2$.
- Leaves variable in a model if maximum of standardized changes-in-estimate greater than $\tau$.

- 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.

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 $\alpha_p = 0.50$.
- 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!).

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

Model averaging

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

where $w^+(j)$ is the sum of weights of models including $\beta_j$.

$$\text{var} \left( \bar{\beta}_j \right) = \left[ \sum_r w_r \text{var} \left( \beta_{j,r} | M_r \right) \right] + \left( \bar{\beta}_j - \bar{\beta}_{j} \right)^2$$

weight within-model variance between-model variance

(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 elimination.
- Select $\alpha_2$ 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 dummies – collapse rare groups with reference
- Ordinal variables: advantages of ordinal coding
  - Variable selection can then collapse adjacent groups with similar outcome

<table>
<thead>
<tr>
<th>Level</th>
<th>Dummy1</th>
<th>Dummy2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Etc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(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:
  - Harvesting covariates
  - Outcome observation

  Inclusion  Index date

  time

  - 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:
  - Harvesting covariates
  - Outcome

  Spring  Harvest

  time

  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 $Y$: $C_1$)
  - Backdoor path blockers (they look like confounders: $C_3$)
  - NOT for instruments ($C_3$ if $U_3$ were not there)
  - NOT for colliders ($C_2$)

  (BIAS)  (BIAS)  (VARIANCE) (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
```r
> 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](http://anythingbutrbitrary.blogspot.co.at/2016/01/how-to-create-confounders-with.html)

An Example – Collider

R Code
```r
> 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| -0.009     | 0.003    | 0.00486 |
| x        | 0.702      | 0.003    | <2e-16   |

Adjusted R-squared: 0.3285
An Example – Collider

R Code

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

Summary

```
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

Questions

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

Simulation setup

- Linear regression
- Sample size $N$ such that $NPV = 5,10,20,50,100$
- Methods:
  - Full model,
  - BW(AIC) ... Backward elimination 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 $R^2$:
- $X1 = 0.3\%$
- $X3 = 2.2\%$
- $X4 = 0.8\%$
- $X5 = 4.0\%$
- $X6 = 3.3\%$
- $X8 = 1.8\%$
- $X10 = 1.8\%$
- $X11 = 5.0\%$

Figure 2: Partial correlations of the variables $x_i$, $j = 1, \ldots, 15$, 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!
In these scenarios, unconditional bias of $\beta$ is towards null!
Conditional bias of $\beta$ is away from null!

Regression coefficients, conditional

RMSE of regression coefficients, unconditional

Heinzl & Dunkler, 03-2016; Part II: 13

Heinzl & Dunkler, 03-2016; Part II: 14

Heinzl & Dunkler, 03-2016; Part II: 15

Heinzl & Dunkler, 03-2016; Part II: 16
Coverage of 95% CI for $\beta$, conditional

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

Heinze & Dunkler, 09-2016; Part II: 21

Coverage of 95% CI for $\beta$, conditional

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

Of course, for BW($p = 0.05$) this happens only in 5%.

Heinze & Dunkler, 09-2016; Part II: 23

Accuracy of predictions

Heinze & Dunkler, 09-2016; Part II: 24
Accuracy of predictions

A network of dependencies...

False inclusion of variables

False exclusion

A network of dependencies...

False inclusion of variables

Bias of predictions, N=150 (10 NPV)

RMSE of predictions, N=150 (10 NPV)

Variance underestimation

Bias of \( \hat{\beta} \)

Competition!

Omission!

Bias of predictions

Undercoverage of CI for \( \beta \)

Variance underestimation (\( \hat{\beta} \))
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: 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.

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 applicable 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
Case study 1: correlation of predictors

Correlations between predictor variables are quite high:

\[ r > 0.9 \]

Case study 1: selection by backward (AIC)

Heinze & Dunkler, 03-2016; Part II-2: 4

Case study 1: selection by backward (AIC)

Heinze & Dunkler, 03-2016; Part II-2: 5

Case study 1: BIFs

Heinze & Dunkler, 03-2016; Part II-2: 6

Heinze & Dunkler, 03-2016; Part II-2: 7
Case study 1: pairwise inclusion frequencies

(Cf. Sauerbrei and Schumacher, 1992)

Heinze & Dunkler, 03-2016; Part II-2: 10

Case study 1: Distribution of regression coefficients

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

Heinze & Dunkler, 03-2016; Part II-2: 11

Case study 1: bootstrap model averaging

Extremely low selection proportions?
Very unstable selection.

Heinze & Dunkler, 03-2016; Part II-2: 11
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.

```r
data bodyfat;
  set casei.bodyfat;
  allometry=log(weight_kg)/log(height_cm);
  beergut=abdomen/cheat;
  heavysert=(knee/wrist*ankle)**(1/3)/height_cm;
  fleshiness=(biceps*cheat*forearm
  /(knee+wrist*ankle))**(1/3);
  age_stand=(age-44.88046)/12.62702;
  age_stand2=age_stand**2;
run;
```

```r
proc corr data=bodyfat;
  var allometry beergut heavysert 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.

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>allometry</td>
</tr>
<tr>
<td>beergut</td>
</tr>
<tr>
<td>heavysert</td>
</tr>
<tr>
<td>fleshiness</td>
</tr>
<tr>
<td>age_stand</td>
</tr>
<tr>
<td>age_stand2</td>
</tr>
</tbody>
</table>

---

Case study 1: an explanatory model

- Pearson Correlation Coefficients, N = 251

<table>
<thead>
<tr>
<th></th>
<th>allometry</th>
<th>beergut</th>
<th>heavysert</th>
<th>fleshiness</th>
<th>age_stand</th>
<th>age_stand2</th>
</tr>
</thead>
<tbody>
<tr>
<td>allometry</td>
<td>1.00000</td>
<td>0.52608</td>
<td>0.62960</td>
<td>0.47580</td>
<td>0.04177</td>
<td>0.02981</td>
</tr>
<tr>
<td>beergut</td>
<td>0.52608</td>
<td>1.00000</td>
<td>0.34205</td>
<td>0.26314</td>
<td>0.24537</td>
<td>0.48484</td>
</tr>
<tr>
<td>heavysert</td>
<td>0.62960</td>
<td>0.34205</td>
<td>1.00000</td>
<td>0.07552</td>
<td>0.22916</td>
<td>0.17979</td>
</tr>
<tr>
<td>fleshiness</td>
<td>0.47580</td>
<td>0.26314</td>
<td>0.07552</td>
<td>1.00000</td>
<td>-0.21279</td>
<td>-0.16023</td>
</tr>
<tr>
<td>age_stand</td>
<td>0.04177</td>
<td>0.24537</td>
<td>0.22916</td>
<td>-0.21279</td>
<td>1.00000</td>
<td>0.8561</td>
</tr>
<tr>
<td>age_stand2</td>
<td>0.02981</td>
<td>0.48484</td>
<td>0.17979</td>
<td>-0.16023</td>
<td>0.8561</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
Case study 1: an explanatory model

- Top two models selected in 86%
- Top three in 92.1%
- Debatable variable: heavyset (P=0.09),
- Irrelevant: age² (P=0.89)

<table>
<thead>
<tr>
<th>Model Selection Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times Selected</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>514</td>
</tr>
<tr>
<td>342</td>
</tr>
<tr>
<td>61</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Debatable variable: heavyset (P=0.09), Irrelevant: age² (P=0.89)

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 applicable 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%)

```r
library(survival)
fit.full <- coxph(Surv(time, status) ~ sex + loc + log2ddm + durther + fvlmode + fvitum + age + bmi, data = deepvein, x = TRUE)
summary(fit.full)
```

Case study 2: risk factors and global model

```r
summary(fit.full)
```

- Parameters: sex, age, bmi, loc, proximal, log2ddm, durther, fvlmode, fvitum, male, proximal, log2ddm, durther, fvlmode, fvitum, age, bmi
- Significance 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:
ocph(formula = Surv(time, status) ~ sex + loc + log2ddim, data = deep vein, k = 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  0.00329 **
loc.proximal -0.20503  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

\[\text{this was the final model.}\]

[For selection with } \alpha = 0.05, \text{ use } k=qchisq(1-0.05,1).\]
```

Case study 2: bootstrapped coefficients

<table>
<thead>
<tr>
<th>Beta of sex.male</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of loc.distal</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>0.1</td>
</tr>
<tr>
<td>-1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of loc.proximal</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>-0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0.0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of log2ddim</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of durther</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.05</td>
<td>0.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of fvleid.present</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of fiimut.present</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-15.0</td>
<td>0.0</td>
</tr>
<tr>
<td>-5.0</td>
<td>0.2</td>
</tr>
<tr>
<td>0.0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of age</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.03</td>
<td>0.0</td>
</tr>
<tr>
<td>-0.01</td>
<td>0.4</td>
</tr>
<tr>
<td>0.01</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beta of bmi</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.06</td>
<td>0.0</td>
</tr>
<tr>
<td>0.00</td>
<td>0.4</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2</td>
</tr>
</tbody>
</table>

including 0's!

Case study 2: Model selection frequencies

Top 10 models:
9 of them contain sex, loc, log2ddim

<table>
<thead>
<tr>
<th>Model</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>1.0</td>
</tr>
<tr>
<td>loc</td>
<td>0.8</td>
</tr>
<tr>
<td>log2ddim</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Heinze & Dunkler, 03-2016; Part II-2: 20
Case study 2: further refinement

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

Case study 2: shrinkage factor estimation

Global shrinkage factor
Parameterwise shr. factors

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 $\rightarrow$ EPV=7.7

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:

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

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

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

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 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 elimination, \( \alpha \) should be set according to the sample size/events per variable.
- Stability investigation based on the bootstrap is helpful.
- Background knowledge \( \neq \) univariate selection.

Implementations: SAS and SPSS

<table>
<thead>
<tr>
<th>What</th>
<th>PROC GLMSELECT</th>
<th>PROC REG</th>
<th>PROC LOGISTIC PROC PHREG</th>
<th>%ABE macro</th>
<th>SPSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Forward</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Stepwise forward</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Stepwise backward</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Augmented backward</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>LASSO</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multi-model inference</td>
<td>(Yes)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Bootstrap stability</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>(No)</td>
<td>No(1)</td>
</tr>
<tr>
<td>investigation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Logistic</td>
<td>No</td>
<td>No</td>
<td>Yes (LOGISTIC)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cox</td>
<td>No</td>
<td>No</td>
<td>Yes (PHREG)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Implementations: R

<table>
<thead>
<tr>
<th>What</th>
<th>lm(), glm(), survival</th>
<th>step()</th>
<th>mfp</th>
<th>glmulti</th>
<th>glmnet, penalized</th>
<th>rms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Forward</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stepwise forward</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stepwise backward</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>All subsets/other</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>LASSO</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multi-model inference</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No (Yes)</td>
<td></td>
</tr>
<tr>
<td>Bootstrap stability</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No(1)</td>
</tr>
<tr>
<td>investigation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>lm()</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Logistic</td>
<td>glm()</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cox</td>
<td>coxph()</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Heinze & Dunkler, 03-2016; Part II-3
Software implementations

- Background knowledge is not implemented in any standard software.

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 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**

doi:DOI 10.1214/ss/1009213726

**Part 1-2: Prerequisites**

Part I-3: Variable selection methods


Part II-1: Consequences


Part II-2: Case studies


Part II-3: Recommendations
