Medical University of Vienna
Center for Medical Statistics, Informatics and Intelligent Systems
Section for Clinical Biometrics
Section Head: Prof. M. Schemper
A-1090 Vienna, Spitalgasse 23
http://cemsiis.meduniwien.ac.at/en/kb/

Technical Report 09/2014

%ABE: A SAS® Macro for Augmented Backward Elimination

Combining Significance and Change-in-Estimate Criteria in a pragmatic and purposeful way to develop statistical models

Daniela Dunkler and Georg Heinze

e-mail: daniela.dunkler @ meduniwien.ac.at

Abstract

We present a new SAS macro **%ABE** that can be used for variable selection by combining backward elimination based on significance and the change-in-estimate criterion. By the latter criterion, non-significant variables are retained in a model if their exclusion leads to a relevant change in the parameter estimates of other variables in the model. The macro handles linear, logistic and Cox regression models. Augmented backward elimination extends the ideas of 'purposeful variable selection' by Hosmer, Lemeshow and May (1999, Chapter 5), so that the analyst can adapt the variable selection to the concrete modeling problem and use an approximative, computationally very efficient change-in-estimate quantification to decrease computational time considerably. By standardizing the change-in-estimate criterion its application is independent from the scaling of the explanatory variables, and in linear models it is also independent from the outcome variable's scale. Furthermore, we calculate the change-in-estimate criterion on the effect size estimates usually reported and interpreted in a model, i.e., odds ratios in logistic regression or hazard ratios in Cox regression.

The augmented backward elimination algorithm and its implementation in SAS is validated by a simulation in an etiological setting and its application is exemplified by means of a biomedical study. The SAS macro is freely available at http://cemsiis.meduniwien.ac.at/kb/wf/software/statistische-software/abe/.

Contents

Abstract 2
1. Overview 4
2. Augmented backward elimination 4
2.1. Purposeful selection of variables 4
2.2. Augmented backward elimination 4
3. Working with the %ABE macro 7
3.1. Syntax 7
3.2. Basic options 7
3.3. Model fitting options 8
3.4. Output options 9
3.5. Titles 9
3.6. Printed output 10
3.7. SAS Log 10
3.8. Counting process style of input 10
3.9. Selection procedures available with the %ABE macro 10
4. Simulation study 11
4.1. Setup of the simulation study 11
4.2. Results 13
4.3. Conclusions 17
5. Example: Urine osmolarity 18
6. Availability, license and disclaimer 25
References 26

1. Overview

Hosmer, Lemeshow and May attempted to semi-automatize the complex task of statistical modeling, i.e., finding parsimonious and valid statistical models which describe the dependency of an outcome on several explanatory variables [1]. They proposed to combine significance and change-in-estimate criteria for selecting explanatory variables for a final main effects model. We added a number of extensions to this procedure and propose Augmented Backward Elimination (ABE). The rationale of ABE is explained in detail in Dunkler and colleagues [2].

The remainder of the technical report is organized as follows: In Chapter 2 the Hosmer and Lemeshow proposal of purposeful variable selection is briefly revisited, followed by description of ABE. In Chapter 3 options available in the SAS macro %ABE are explained [3]. In Chapter 4 the results of a simulation study evaluating ABE and its implementation in SAS are shown and in Chapter 5 application of ABE is exemplified by a biomedical example.

2. Augmented backward elimination

2.1. Purposeful selection of variables

Hosmer and Lemeshow suggested a purposeful selection of variables [1]. This algorithm consists of three stages: Starting with an initial set of variables, at the *first stage* univariate screening with a significance level of, say 0.25, is applied to develop an initial multivariable model. At the *second stage* variables are eliminated in a stepwise manner based on significance or on the change-in-estimate criterion; a p-value lower than 0.10 or a relative change of more than 20% in any other coefficient is considered sufficient for retaining a variable in a working model. In particular, the variable with the highest p-value in the current working model will be eliminated if that p-value is greater than the significance level and its change-in-estimate criterion is lower than the respective threshold. Iteratively the two criteria are assessed for all variables until no further variable is removed. At the *third stage* variables that were not considered in the initial working model because of 'univariate' p-values larger than 0.25 are one-by-one re-entered and evaluated for significance and the change-in-estimate criterion similarly to the second stage.

2.2. Augmented backward elimination

Suppose that investigators have used subject-matter knowledge to define a set of possible candidate variables, the initial working set of variables, prior to analysis. An initial multivariable main effects

model including these variables is fitted. Then for each variable in this model the algorithm evaluates significance and the change-in-estimate criterion. For the significance criterion a mild threshold like 0.20 was proposed in the literature (see e.g. [4]). An 'active variable' X_a is excluded from the model if its p-value exceeds the significance threshold and if this exclusion will not cause changes in regression coefficients of the other 'passive variables' X_b larger than the pre-specified change-in-estimate threshold.

The change-in-estimate δ_p^{-a} related to variable X_p by elimination of variable X_a is approximated, using the current estimates $\hat{\beta}_p$ and $\hat{\beta}_a$, their covariance $\hat{\sigma}_{pa}$ and the variance of $\hat{\beta}_a$, $\hat{\sigma}_a^2$, as

$$\hat{\delta}_p^{-a} = -\frac{\hat{\beta}_a \hat{\sigma}_{pa}}{\hat{\sigma}_a^2}.$$

The change-in-estimate criterion is applied on the quantities usually reported and interpreted in a model, i.e., odds ratios in logistic regression or hazard ratios in Cox regression. Furthermore, the change-in-estimate criterion is standardized such that its application is independent on the scaling of the explanatory variables X_p (and if applicable on the outcome Y). The following definition of the change-in-estimate criterion is used in ABE:

1) For linear models, the quantity of interest is the regression coefficient itself. For a change-in-estimate independent from the scaling of the explanatory and outcome variables, an active variable X_a is not excluded from a given working model if for any passive variable X_p in the model

$$\frac{|\delta_p^{-a}| SD(X_p)}{SD(Y)} \geq \tau,$$

where $SD(X_p)$ and $SD(Y)$ are the standard deviations of the passive explanatory variable X_p and of the outcome Y, respectively.

2) For logistic or Cox models, we are interested in odds and hazard ratios, respectively. Thus, an active variable X_a is not excluded from a given model if for any passive variable X_p in the model

$$\exp[\left|\delta_p^{-a}\right| SD(X_p)] \geq 1 + \tau \text{ or equivalently, } |\delta_p^{-a}| SD(X_{p)} \geq \log(1 + \tau).$$

The threshold τ applied to the change-in-estimate criterion could be set to, say 0.05, but can be adopted to the specific modeling situation.

Furthermore, in a given modeling situation the individual explanatory variables may fill different roles with regards to their content and this may affect the selection process. We have

identified three specific roles of explanatory variables, which may require different handling when evaluating the change-in-estimate criterion:

1) Generally, explanatory variables are used as passive as well as active variables when evaluating the change-in-estimate criterion.
2) In etiologic modeling, it is necessary to always keep the exposure variable of interest in the model since this is the variable for which an appropriately adjusted effect estimate should be obtained. Furthermore, one may force the modeling process to always include some known confounders (in etiologic modeling) or predictors (in prognostic modeling). Although their inclusion is never questioned at any of the cycles of the algorithm, such exposures of interest, known confounders or predictors are always considered as passive variables in evaluating the change-in-estimate criterion for the other variables.
3) There may be explanatory variables which are deemed less important and consequently, these should only be included if their exclusion would cause changes in the estimates of explanatory variables of greater importance. Thus, such variables of minor importance are only considered as active but never as passive variables when evaluating the change-in-estimate criterion.

The augmented backward elimination algorithm ABE is briefly outlined in **Figure 1**.

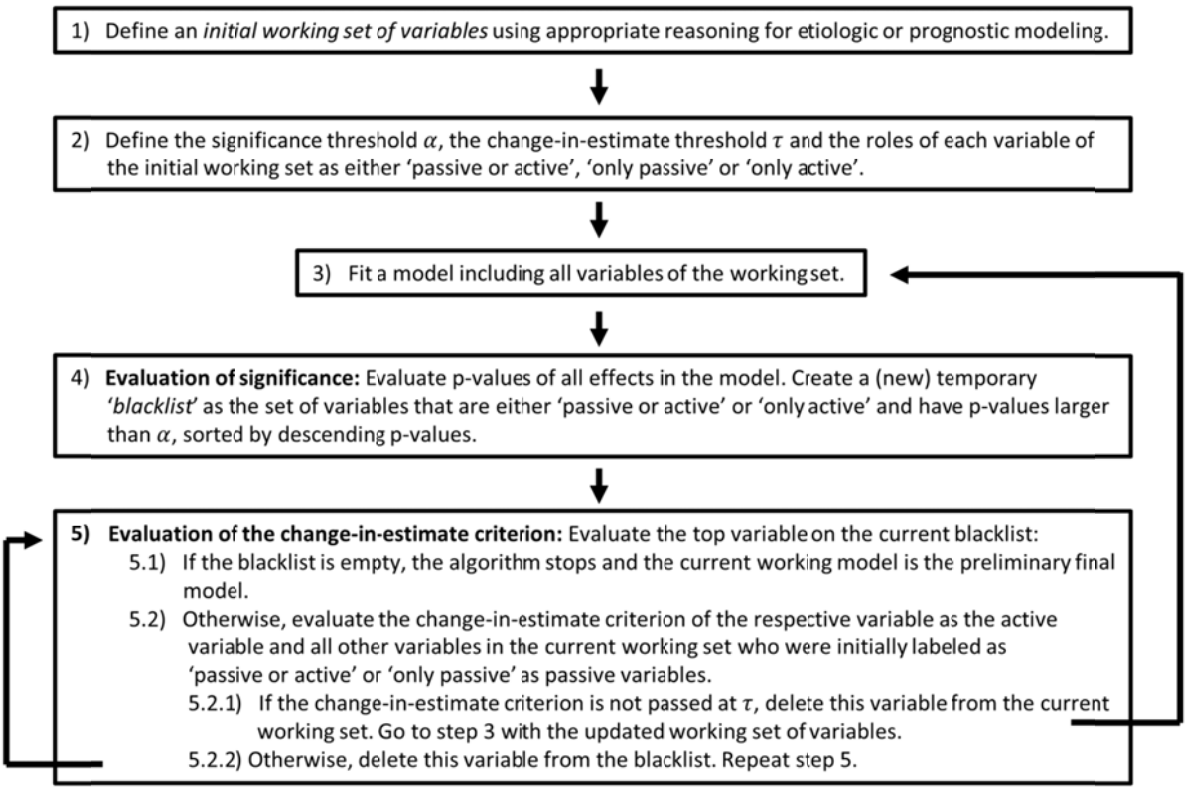

Figure 1: Brief outline of the augmented backward elimination procedure.

3. Working with the %ABE macro

3.1. Syntax

The following macro options allow the analyst to specify the initial multivariable working model and the variable selection criteria when using **%ABE**. The brackets < and > denote options that are optional.

```
%ABE( <data = SAS data set,>
<time        = variable,>
<cens        = variable,>
<censval     = value,>
<y           = variable,>
varlist      = variables,
<active      = variables,>
<include     = variable,>
proc         = string,
<first       = string,>
<puni        = value,>
<pmulti      = value,>
<pequiv      = value,>
<tau         = value,>
<equiv       = value,>
<cycles      = value,>
<options     = string,>
<print       = value,>
<notes       = value,>
<final       = value,>
<confounders = SAS data set,>
<logfile     = SAS data set >);
```

These options are described in the subsequent sections.

3.2. Basic options

- `data = SAS data set` names the input SAS data set. The default value is _LAST_.
- `y = variable` names a variable containing the dependent variable if linear or logistic regression is requested. There is no default value.

- `time = variable` names a variable containing survival times if a Cox proportional hazards model is requested. There is no default value. Survival times can also be supplied using the counting process format by specifying `time = %str((start_time, stop_time))`. For more information see chapter 'Counting process style of input'.
- `cens = variable` names a variable containing the censoring indicator for each survival time if a Cox proportional hazards model is requested. There is no default value.
- `censval = value` identifies the values of censored observations. By default, a value of 1 represents events and a value of 0 censored times.
- `proc = string` requests the modeling procedure. Three procedures are supported: linear regression '`reg`', logistic regression '`logistic`', and proportional hazards Cox regression '`phreg`'.

3.3. Model fitting options

- `varlist = variables` names a list of independent variables, separated by blanks. Variables which only appear in this option will be evaluated for both (significance and change-in-estimate) criteria. There is no default value. This option is required.
- `active = variables` names a list of independent variables, separated by blanks. Only variables which are also named in `varlist` can be used here. These less important explanatory variables will only be used as active, but not as passive variables when evaluating the change-in-estimate criterion.
- `include = variables` names a list of independent variables, separated by blanks. These variables might be exposure variables of interest or known confounders. They will never be dropped from the working model in the selection process, but they will be used passively in evaluating change-in-estimate criteria of other variables.
- `first = string` specifies the strategy to select independent variables for the initial multivariable working model. Currently supported options are univariate selection '`uni`', backward elimination '`b`', forward selection '`f`' and no preselection '`none`'. For more information see Chapter 3.9. 'Selection procedures available with the %ABE macro'. Default is no preselection (`none`).
- `puni = value` specifies the significance level of inclusion in *univariate* models used to select explanatory variables for the initial multivariable working model with univariate or forward preselection (`First = uni,` or `f`). Default is set to 0.25 as suggested by Hosmer and Lemeshow [1].

- `pmulti = value` specifies the significance level of retention applied to p-values from *multivariable* models. This is also the threshold for backward preselection (`First = b`). Default is set to 0.10.
- `tau = value` specifies the threshold of the relative change-in-estimate criterion. Default is set to 0.05.
- `equiv = value` specifies if a test of equivalence should be conducted (`equiv = 1`) or not (`equiv = 0`) [5].
- `pequiv = value` specifies the significance level for the equivalence test.
- `cycles = value` specifies the maximal number of cycles. Default is set to 10.
- `options = string` specifies further options to be passed to the model statement of the invoked procedure. Here, we recommend to make use of the SAS specific embedding of strings, `%str(…)`. For example, to print odds ratios of a logistic regression model, options can be set to `%str(expb)`; or one may close the model statement and include another statement in the procedure step, such as a weight statement, by `%str(; weight weight_variable)`.

3.4. Output options

- `print = value` if set to 0, suppresses intermediate model reporting. Default is set to 0.
- `notes = value` if set to 1, prints notes in the log-file. Default is set to 1. This option is of advantage when simulations or resampling is requested.
- `final = value` if set to 1, sends the output of the final multivariable model to the output destination. Default is set to 1.
- `confounders = SAS data set` names the output data set including the names of all independent variables which have been selected by the algorithm for the final multivariable model. The default name is `_confounders`.
- `logfile = SAS data set` names the output data which summarizes the selection process . The default name is `_logfile`.

3.5. Titles

The macro uses only title 3. All other titles can be set by the user in a statement before the macro call.

3.6. Printed output

Either only the final main effects model is printed (`final = 1`), or all working models and the final model are printed (`final = 0`). To suppress printing of any model, e.g., in a simulation, use `print = 1`.

3.7. SAS Log

In order to trace all explanatory variables through the selection process, an analyst can set `notes = 1` requesting notes in the SAS log-file for all selection decisions.

3.8. Counting process style of input

The macro adopts the counting process formulation of Cox's model from SAS/PROC PHREG. In this formulation, the data for each subject can be represented by multiple observations, each identifying a semi-closed time interval (`time1`, `time2`], with `time1` and `time2` signifying interval entry time and (possibly censored) survival time, the values of the explanatory variables over that interval, and the event status at `time2`. The subject remains at risk during the interval (`time1`, `time2`] and an event may occur at `time2`. Values of the explanatory variables for the subject remain unchanged in the interval. The notation (a; b] means that the interval ranges from a to b, excluding a and including b.

`time1 =` *`variable`* and `time2 =` *`variable`* name variables containing the endpoints of a semi-closed interval (time1, time2] during which the subject is at risk. Specification of `time2` overrules any specification of the option `time`. Option `time1` has the default value 0.

3.9. Selection procedures available with the %ABE macro

Using the default settings %ABE will request augmented backward elimination (option `first = NONE`). A significance level `pmulti` of 0.20 and a standardized change-in-estimate criterion `tau` set to 0.05 will be applied by default.

Setting `tau` to a very large number turns off the change-in-estimate criterion, and %ABE will only perform BE. On the other hand, the specification of `pmulti` set to zero will include variables only because of the change-in-estimate criterion, as then variables are not safe from exclusion because of their p-values. Specifying `pmulti` = 1 will always include all variables.

Setting `equiv = 1` an equivalence-test of the change-in-estimate will be applied [5].

If `first = UNI` then purposeful selection of variables as suggested by Hosmer and Lemeshow will be requested [1]. However, using the %ABE macro there are some important differences with regards to the evaluation of the change-in-estimate criterion. First of all, it will be approximated reducing computation time considerably; secondly, it will be evaluated on the quantity of interest depending on the type of regression; and thirdly, it will be standardized and therefore, be independent on the scaling of the explanatory variables X_p (and if applicable the outcome Y).

4. Simulation study

4.1. Setup of the simulation study

In a simulation study, we

a) evaluated the ABE procedure and its implementation in the SAS macro %ABE,
b) investigated the robustness of the parameters of ABE for linear, logistic and Cox proportional hazards regression, and
c) compared the performance of ABE to other types of variable selection (BE and disjunctive cause criterion [6]) with regards to bias and root mean squared error (RMSE).

The structure of the true model is depicted in **Figure 2**.

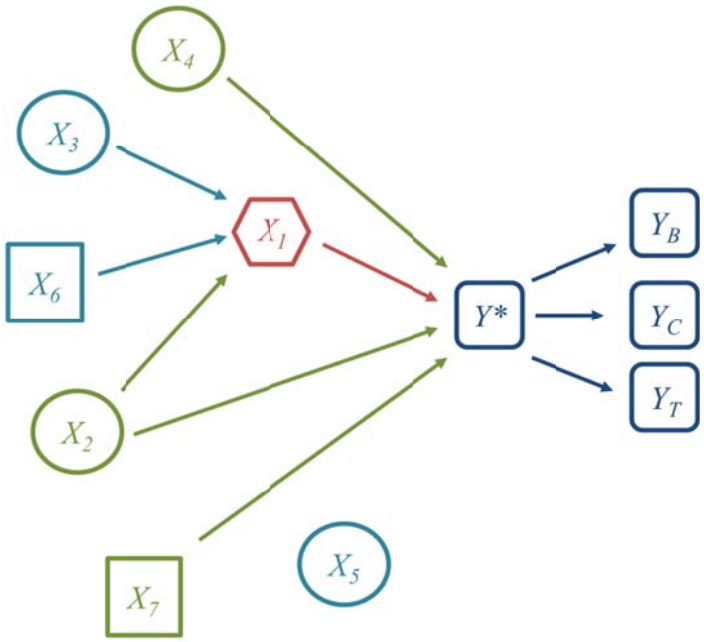

Figure 2: Simulation design. X_1 = exposure variable of interest, X_2 to X_7 additionally measured explanatory variables. Y^* = latent outcome variable, Y_B, Y_C and Y_T = binary, continuous and time-to-event outcome variables. Explanatory variables in circles (X_2, X_3, X_4, and X_5) are correlated with each other. The correct model comprises the variables X_1, X_2, X_4, and X_7.

We simulated seven normally distributed potential explanatory variables $X_1, \ldots, X_7$ as follows: X_2, X_3, X_4, and X_5 were drawn from a multivariate normal distribution with a mean vector of 0, standard deviations of 1 and bivariate correlation coefficients of 0.5. X_6 and X_7 were independently drawn from a standard normal distribution. X_1 was the exposure variable of main interest and depended on X_2, X_3 and X_6. This was achieved by simulating X_1 from the equations $X_1 = 0.266(X_2 + X_3 + X_6) + 0.710\epsilon$ (for scenarios with a variance inflation factor [VIF] of 2) and $X_1 = 0.337(X_2 + X_3 + X_6) + 0.449\epsilon$ (for a VIF of 4), where ϵ was a random number drawn from a standard normal distribution. Y^* is a latent continuous outcome variable, which was defined as $Y^* = \beta_1 X_1 + X_2 + X_4 + X_7$. Thus, false omission of X_2, X_4 or X_7 from the model, or false inclusion of X_3 could induce bias into the estimate of β_1.

Subsequently, from the latent outcome variable Y^* we generated continuous, binary and time-to-event outcome variables Y_C, Y_B and Y_T to simulate an outcome variable for linear, logistic and Cox regression, respectively. In particular, Y_C was drawn from a normal distribution with mean Y^* and standard deviation 0.36. Y_B was drawn from a Bernoulli distribution with event probability $1/(1 + \exp[-Y^*])$. Weibull distributed survival times T were drawn from $\left(-\frac{\log(U)}{0.125 \exp(Y^*)}\right)^{1/3}$, where U was a standard uniform random variable [9]. To obtain approximately 55% censoring, follow-up times F were drawn from a uniform $U[0, 3.35]$ distribution, and the observable survival time and status indicators were defined as $Y_T = \min(T, U)$ and $S_T = I(T > U)$, respectively. For Cox regression, all covariates were divided by 2. These definitions guaranteed that the sampling standard deviations of estimates of β_1 from linear, logistic and Cox regression were approximately equal when the models were specified correctly.

In a factorial design we simulated 1000 samples of 120 observations for each combination of true β_1 (either 0 or 1), VIF (either 2 or 4) and type of regression (either linear, logistic or Cox). If $\beta_1 = 1$ and $VIF = 2$, this sample size gave a power of 50% to reject the null hypothesis $\beta_1 = 0$ at a two-sided signficance level of 5% in all three types of regression, when the model was specified correctly.

Each sample was analyzed with following methods:

- Models fixed prior to analysis (e.g. based on a-priori knowledge):
  - Correct model including X_1, X_2, X_4, and X_7.
  - Univariate model including X_1.
  - Full model including X_1 to X_7. This model is used, when the pretreatment criterion is applied.
  - Model based on common cause criterion (CCC) including X_1, and X_2.

- Model based on disjunctive cause criterion (DCC) including X_1, X_2, X_3, X_4, X_6, and X_7.

- Models with data-dependent variable selection using all explanatory variables:
  - Model selected with BE and a significance level $\alpha = 0.05$ or 0.20.
  - Model selected with BE and a significance level $\alpha = 0.05$ or 0.20, but applying the disjunctive cause criterion prior to analysis (X_5 is not considered).
  - Model selected with ABE with a significance level $\alpha = 0.20$ and $\tau = 0.05$ or 0.10.
  - Model selected with ABE with a significance level $\alpha = 0.20$ and $\tau = 0.05$ or 0.10, but applying the disjunctive cause criterion prior to analysis.

The variable of main interest X_1 was forced into every model.

4.2. Results

Table 1 gives some of the results of the simulation study. For every $\widehat{\beta_1}$ in all samples the $bias \times 100$ and the $RMSE \times 100$, as well as, the $bias \times 100$ and the $RMSE \times 100$ compared to the correct model are given. Furthermore, it is stated how often the correct model was selected by either BE or ABE and how often either a smaller, 'biased' model (i.e., at least one of the correct variables X_2, X_4 or X_7 was not selected into the final model) or a larger, 'inflated' model (i.e., all correct variables X_2, X_4 and X_7 were selected into the final model, but at least one incorrect variable X_3, X_5 or X_6, as well).

For the linear model results of the models fixed before analysis or defined based on a-priori knowledge and models selected by BE are independent of the true value of β_1.

Table 1: Main results of the simulation study for linear, logistic and Cox proportional hazards regression.

	Type	Model	Bias * 100	RMSE * 100	Bias * 100 of $\widehat{\beta_1}$ to correct model	RMSE * 100 of $\widehat{\beta_1}$ to correct model	No. models selected (%) biased	correct	inflated
Linear regression (VIF = 2, $\beta_1 = 0$)	**Fixed**	**correct model**	**-1.867**	**42.037**				**100**	
	Fixed	univariate model	92.481	99.852	94.347	98.970	100		
	Fixed	full model	-1.116	48.058	0.751	24.423			100
	Fixed	CCC model	12.179	45.047	14.045	22.038	100		
	Fixed	DCC model	-0.764	48.041	1.103	24.083			100
	BE	$\alpha = 0.05$	7.962	48.881	9.828	21.366	66.2	28.8	5.0
	BE	DCC, $\alpha = 0.05$	8.041	49.023	9.907	21.456	65.4	31.3	3.3
	BE	$\boldsymbol{\alpha = 0.20}$	**0.903**	**48.250**	**2.770**	**21.221**	**32.7**	**35.2**	**32.1**
	BE	DCC, $\alpha = 0.20$	1.334	48.482	3.201	20.993	32.0	44.1	23.9
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**-0.022**	**47.994**	**1.845**	**21.726**	**27.7**	**34.2**	**38.1**
	ABE	$\alpha = 0.20, \tau = 0.10$	0.961	48.272	2.827	21.304	32.2	35.1	32.7
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	0.484	48.163	2.350	21.548	26.5	43.8	29.7
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	1.408	48.526	3.275	21.087	31.5	44.1	24.4
Linear regression (VIF = 2, $\beta_1 = 1$)	**Fixed**	**correct model**	**-1.867**	**42.037**				**100**	
	Fixed	univariate model	92.481	99.852	94.347	98.970	100		
	Fixed	full model	-1.116	48.058	0.751	24.423			100
	Fixed	CCC model	12.179	45.047	14.045	22.038	100		
	Fixed	DCC model	-0.764	48.041	1.103	24.083			100
	BE	$\alpha = 0.05$	7.962	48.881	9.828	21.366	66.2	28.8	5.0
	BE	DCC, $\alpha = 0.05$	8.041	49.023	9.907	21.456	65.4	31.3	3.3
	BE	$\boldsymbol{\alpha = 0.20}$	**0.903**	**48.250**	**2.770**	**21.221**	**32.7**	**35.2**	**32.1**
	BE	DCC, $\alpha = 0.20$	1.334	48.482	3.201	20.993	32.0	44.1	23.9
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**0.318**	**47.921**	**2.184**	**21.353**	**29.2**	**35.2**	**35.6**
	ABE	$\alpha = 0.20, \tau = 0.10$	0.961	48.272	2.827	21.304	32.2	35.1	32.7
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	0.808	48.106	2.675	21.111	28.1	44.8	27.1
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	1.408	48.526	3.275	21.087	31.5	44.1	24.4
Linear regression (VIF = 4, $\beta_1 = 0$)	**Fixed**	**correct model**	**-2.705**	**50.346**				**100**	
	Fixed	univariate model	117.707	123.514	120.411	126.918	100		
	Fixed	full model	-1.764	75.977	0.940	56.871			100
	Fixed	CCC model	21.999	55.372	24.704	32.618	100		
	Fixed	DCC model	-1.208	75.951	1.497	56.594			100
	BE	$\alpha = 0.05$	17.704	69.700	20.408	45.466	71.7	24.1	4.2
	BE	DCC, $\alpha = 0.05$	17.781	69.850	20.486	45.686	70.9	26.2	2.9
	BE	$\boldsymbol{\alpha = 0.20}$	**3.014**	**72.357**	**5.718**	**49.812**	**39.5**	**33.7**	**26.8**
	BE	DCC, $\alpha = 0.20$	3.504	72.060	6.209	49.358	38.5	42.1	19.4
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**-0.598**	**75.993**	**2.106**	**55.970**	**25.1**	**14.4**	**60.5**
	ABE	$\alpha = 0.20, \tau = 0.10$	1.034	73.758	3.739	52.215	36.3	30.4	33.3
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	-0.518	75.952	2.186	55.784	23.9	16.6	59.5
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	0.908	73.315	3.612	51.809	34.8	38.1	27.1
Linear regression (VIF = 4, $\beta_1 = 1$)	**Fixed**	**correct model**	**-2.705**	**50.346**				**100**	
	Fixed	univariate model	117.707	123.514	120.411	126.918	100		
	Fixed	full model	-1.764	75.977	0.940	56.871			100
	Fixed	CCC model	21.999	55.372	24.704	32.618	100		
	Fixed	DCC model	-1.208	75.951	1.497	56.594			100
	BE	$\alpha = 0.05$	17.704	69.700	20.408	45.466	71.7	24.1	4.2
	BE	DCC, $\alpha = 0.05$	17.781	69.850	20.486	45.686	70.9	26.2	2.9
	BE	$\boldsymbol{\alpha = 0.20}$	**3.014**	**72.357**	**5.718**	**49.812**	**39.5**	**33.7**	**26.8**
	BE	DCC, $\alpha = 0.20$	3.504	72.060	6.209	49.358	38.5	42.1	19.4
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**-0.745**	**75.932**	**1.960**	**55.679**	**27.1**	**16.6**	**56.3**
	ABE	$\alpha = 0.20, \tau = 0.10$	0.926	73.773	3.631	51.867	37.1	31.3	31.6
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	-0.483	75.813	2.221	55.548	25.9	19.2	54.9
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	1.186	73.429	3.890	51.578	36.2	39.1	24.7

	Type	Model	Bias * 100	RMSE * 100	Bias * 100 of $\widehat{\beta_1}$ to correct model	RMSE * 100	No. models selected (%) biased	correct	inflated
Logistic reg. (VIF = 2, $\beta_1 = 0$)	**Fixed**	**correct model**	**0.12**	**32.328**				**100**	
	Fixed	univariate model	63.711	67.083	63.591	67.668	100		
	Fixed	full model	0.971	39.016	0.851	20.35			100
	Fixed	CCC model	11.272	28.353	11.151	20.606	100		
	Fixed	DCC model	0.890	38.102	0.769	19.441			100
	BE	$\alpha = 0.05$	3.374	37.002	3.253	13.971	20.1	69.4	10.5
	BE	DCC, $\alpha = 0.05$	3.414	36.756	3.293	13.433	19.5	73.2	7.3
	BE	$\boldsymbol{\alpha = 0.20}$	**1.14**	**37.371**	**1.019**	**16.630**	**7.2**	**47.5**	**45.3**
	BE	DCC, $\alpha = 0.20$	1.166	36.674	1.045	15.754	6.4	58.8	34.8
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**0.958**	**39.018**	**0.838**	**20.165**	**1.0**	**4.3**	**94.7**
	ABE	$\alpha = 0.20, \tau = 0.10$	1.013	38.763	0.893	19.630	3.0	19.2	77.8
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	0.994	38.107	0.873	19.289	0.4	11.3	88.3
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	1.116	37.973	0.996	18.883	2.6	32.8	64.6
Logistic reg. (VIF = 2, $\beta_1 = 1$)	**Fixed**	**correct model**	**9.066**	**41.246**				**100**	
	Fixed	univariate model	35.477	45.642	26.412	37.04	100		
	Fixed	full model	15.209	51.168	6.143	24.764			100
	Fixed	CCC model	-7.144	33.064	-16.210	27.309	100		
	Fixed	DCC model	13.123	49.269	4.058	22.690			100
	BE	$\alpha = 0.05$	14.707	47.312	5.641	17.578	34.6	55.7	9.7
	BE	DCC, $\alpha = 0.05$	14.082	46.726	5.016	16.399	33.9	59.7	6.4
	BE	$\boldsymbol{\alpha = 0.20}$	**13.763**	**48.837**	**4.698**	**20.989**	**11.1**	**42.6**	**46.3**
	BE	DCC, $\alpha = 0.20$	12.038	47.529	2.973	19.086	9.4	56.7	33.9
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**15.376**	**51.184**	**6.310**	**24.709**	**1.6**	**2.8**	**95.6**
	ABE	$\alpha = 0.20, \tau = 0.10$	15.846	51.532	6.781	24.317	5.1	14.5	80.4
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	13.235	49.311	4.170	22.650	1.4	9.7	88.9
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	13.858	49.829	4.792	22.218	3.9	28.0	68.1
Logistic reg. (VIF = 4, $\beta_1 = 0$)	**Fixed**	**correct model**	**-0.305**	**38.912**				**100**	
	Fixed	univariate model	85.455	88.455	85.760	90.879	100		
	Fixed	full model	1.536	61.683	1.840	46.620			100
	Fixed	CCC model	19.571	36.587	19.876	29.075	100		
	Fixed	DCC model	1.407	60.238	1.711	45.247			100
	BE	$\alpha = 0.05$	10.012	55.57	10.317	33.474	30.7	59.9	9.4
	BE	DCC, $\alpha = 0.05$	9.447	54.571	9.752	31.708	29.8	64.0	6.2
	BE	$\boldsymbol{\alpha = 0.20}$	**3.510**	**56.305**	**3.815**	**38.345**	**12.4**	**45.9**	**41.7**
	BE	DCC, $\alpha = 0.20$	3.854	55.221	4.159	36.734	11.6	57.3	31.1
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**1.603**	**61.733**	**1.908**	**46.627**	**0.9**	**1.6**	**97.5**
	ABE	$\alpha = 0.20, \tau = 0.10$	1.886	61.526	2.191	46.380	3.0	5.4	91.6
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	1.476	60.306	1.780	45.265	1.2	2.9	95.9
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	1.603	60.085	1.907	45.105	2.5	9.5	88.0
Logistic reg. (VIF = 4, $\beta_1 = 1$)	**Fixed**	**correct model**	**9.312**	**49.328**				**100**	
	Fixed	univariate	62.078	69.965	52.766	61.559	100		
	Fixed	full	15.433	76.319	6.121	55.332			100
	Fixed	CCC	2.002	38.998	-7.310	25.828	100		
	Fixed	DCC	13.092	73.187	3.780	52.591			100
	BE	$\alpha = 0.05$	23.426	68.263	14.114	38.651	45.0	46.2	8.8
	BE	DCC, $\alpha = 0.05$	22.179	66.959	12.867	37.431	44.1	50.1	5.8
	BE	$\boldsymbol{\alpha = 0.20}$	**16.601**	**71.492**	**7.289**	**46.489**	**16.5**	**40.6**	**42.9**
	BE	DCC, $\alpha = 0.20$	15.205	69.238	5.892	44.234	15.6	52.2	32.2
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**15.536**	**76.343**	**6.224**	**55.321**	**1.2**	**1.0**	**97.8**
	ABE	$\alpha = 0.20, \tau = 0.10$	15.686	76.169	6.374	55.169	4.0	4.6	91.4
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	13.167	73.225	3.855	52.576	1.1	1.9	97.0
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	13.386	73.107	4.073	52.432	3.0	7.0	90.0

	Type	Model	Bias * 100	RMSE * 100	Bias * 100 of $\widehat{\beta_1}$ to correct model	RMSE * 100	No. models selected (%) biased	correct	inflated
Cox regression (VIF = 2, $\beta_1 = 0$)	**Fixed**	**correct model**	**-1.410**	**39.896**				**100**	
	Fixed	univariate model	77.319	83.515	78.728	84.400	100		
	Fixed	full model	-2.326	46.355	-0.916	22.163			100
	Fixed	CCC model	11.515	40.382	12.925	27.034	100		
	Fixed	DCC model	-2.466	45.753	-1.056	21.087			100
	BE	$\alpha = 0.05$	4.696	46.287	6.105	19.357	44.7	46.3	9.0
	BE	DCC, $\alpha = 0.05$	4.661	46.348	6.071	18.761	43.5	49.7	6.8
	BE	$\boldsymbol{\alpha = 0.20}$	**-0.066**	**45.413**	**1.344**	**18.893**	**19.7**	**39.5**	**40.8**
	BE	DCC, $\alpha = 0.20$	0.002	45.286	1.411	17.853	17.6	50.5	31.9
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**-2.034**	**46.222**	**-0.624**	**21.618**	**7.0**	**13.2**	**79.8**
	ABE	$\alpha = 0.20, \tau = 0.10$	-0.958	45.672	0.452	20.118	14.8	33.9	51.3
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	-2.042	45.561	-0.633	20.580	5.9	26.9	67.2
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	-1.134	45.411	0.276	19.407	13.4	45.1	41.5
Cox regression (VIF = 2, $\beta_1 = 1$)	**Fixed**	**correct model**	**5.283**	**43.147**				**100**	
	Fixed	univariate model	58.731	68.951	53.448	61.482	100		
	Fixed	full model	7.082	49.968	1.798	22.998			100
	Fixed	CCC model	0.939	41.661	-4.344	25.919	100		
	Fixed	DCC model	5.907	49.087	0.623	21.801			100
	BE	$\alpha = 0.05$	11.674	50.258	6.391	19.355	47.1	44.0	8.9
	BE	DCC, $\alpha = 0.05$	11.108	50.047	5.825	18.824	46.5	47.2	6.3
	BE	$\boldsymbol{\alpha = 0.20}$	**8.097**	**48.417**	**2.813**	**19.919**	**19.4**	**39.8**	**40.8**
	BE	DCC, $\alpha = 0.20$	7.231	47.950	1.948	18.902	18.3	50.9	30.8
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**7.229**	**49.889**	**1.945**	**22.559**	**7.0**	**12.7**	**80.3**
	ABE	$\alpha = 0.20, \tau = 0.10$	7.828	49.039	2.545	21.079	14.8	32.6	52.6
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	6.104	49.086	0.821	21.308	5.7	25.1	69.2
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	6.879	48.513	1.596	20.196	13.7	45.1	41.2
Cox regression (VIF = 4, $\beta_1 = 0$)	**Fixed**	**correct model**	**-1.482**	**50.392**				**100**	
	Fixed	univariate model	106.792	111.689	108.274	115.963	100		
	Fixed	full model	-3.677	73.285	-2.195	51.942			100
	Fixed	CCC model	23.938	53.312	25.421	39.992	100		
	Fixed	DCC model	-3.898	72.334	-2.416	50.719			100
	BE	$\alpha = 0.05$	14.759	70.336	16.241	43.793	54.6	38.1	7.3
	BE	DCC, $\alpha = 0.05$	13.853	69.54	15.335	42.689	53.8	40.6	5.6
	BE	$\boldsymbol{\alpha = 0.20}$	**2.720**	**70.917**	**4.202**	**46.579**	**26.4**	**36.2**	**37.4**
	BE	DCC, $\alpha = 0.20$	2.351	70.101	3.833	44.688	24.8	46.2	29
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**-3.512**	**73.141**	**-2.03**	**51.726**	**5.9**	**4.6**	**89.5**
	ABE	$\alpha = 0.20, \tau = 0.10$	-2.902	72.789	-1.420	50.464	12.9	18.7	68.4
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	-3.984	72.26	-2.502	50.567	5.4	9.9	84.7
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	-3.34	71.662	-1.858	49.638	12.1	24.5	63.4
Cox regression (VIF = 4, $\beta_1 = 1$)	**Fixed**	**correct model**	**4.782**	**52.057**				**100**	
	Fixed	univariate model	91.183	98.380	86.401	95.436	100		
	Fixed	full model	5.855	75.059	1.073	52.150			100
	Fixed	CCC model	13.098	51.253	8.316	33.017	100		
	Fixed	DCC model	4.431	73.610	-0.351	50.691			100
	BE	$\alpha = 0.05$	22.370	74.188	17.588	44.318	58.4	35.3	6.3
	BE	DCC, $\alpha = 0.05$	21.734	72.861	16.953	42.438	57.1	38.0	4.9
	BE	$\boldsymbol{\alpha = 0.20}$	**10.861**	**73.030**	**6.079**	**46.393**	**27.6**	**36.7**	**35.7**
	BE	DCC, $\alpha = 0.20$	9.654	71.266	4.872	44.823	26.6	46.5	26.9
	ABE	$\boldsymbol{\alpha = 0.20, \tau = 0.05}$	**6.012**	**74.908**	**1.230**	**51.968**	**7.4**	**3.4**	**89.2**
	ABE	$\alpha = 0.20, \tau = 0.10$	6.741	74.435	1.959	51.018	14.5	16.7	68.8
	ABE	DCC, $\alpha = 0.20, \tau = 0.05$	4.591	73.409	-0.191	50.573	6.1	8.7	85.2
	ABE	DCC, $\alpha = 0.20, \tau = 0.10$	5.489	72.826	0.707	49.720	13.2	22.7	64.1

4.3. Conclusions

In contrast to the true model the $bias$ (of $\widehat{\beta_1}$) of ABE with τ set to 0.05 was in almost all scenarios smaller than the $bias$ (of $\widehat{\beta_1}$) of BE. The median (minimum, maximum) absolute $bias \times 100$ (of $\widehat{\beta_1}$) in contrast to the true model over all scenarios and types of regression analyzed was

- for BE with $\alpha = 0.20$: $4.008\ (1.019,\ 7.289)$,
- for ABE with $\alpha = 0.20$ and $\tau = 0.05$: $1.953\ (0.624,\ 6.310)$, and
- for the full model: $1.007\ (0.751, 6.143)$.

The $RMSE$ (of $\widehat{\beta_1}$) of effect estimates by ABE was between the $RMSE$ (of $\widehat{\beta_1}$) of BE and $RMSE$ (of $\widehat{\beta_1}$) of the full model. The median (minimum-maximum) $RMSE \times 100$ (of $\widehat{\beta_1}$) in contrast to the true model over all scenarios and types of regression analyzed was

- for BE with $\alpha = 0.20$: $29.78\ (16.63,\ 49.81)$,
- for ABE with $\alpha = 0.20$ and $\tau = 0.05$: $35.67\ (20.16,\ 55.97)$, and
- for the full model: $35.69\ (20.35,\ 56.87)$.

The tendency of BE to select too few variables can be limited by setting the significance level α to 0.20. For small samples one might use even higher significance levels, like 0.50, as suggested by Steyerberg et al. [4]. In contrast ABE tends to select too many variables and approximates the 'full' model up to negligible differences in point estimates of the regression coefficients.

Incorporating (correct) prior knowledge about the causal relationship between explanatory variables to reduce the initial set of variables generally improves the performance of all methods.

These conclusions were similar in linear, logistic and Cox regression analyses. Differences between BE and ABE were even more pronounced when the sample size was reduced (results not shown). Differences between BE and ABE are expected to be more pronounced if correlated groups of explanatory variables are in the initial set.

Thus, we conclude that in the scenarios studied, application of ABE with the proposed settings for α and τ is at least as safe as application of BE, and is at least as good as, but often better than, including all available variables from the initial set for adjustment.

5. Example: Urine osmolarity

Using data of 295 patients with chronic kidney disease who attended the nephrology outpatient clinic at the Medical University of Vienna, Plischke and colleagues investigated the etiologic effect on the cause-specific hazard of urine osmolarity on progression to end stage renal disease defined as admission to dialysis [10]. During a median follow-up time of 7.93 years 95 patients (35.78%) required dialysis. The outcome of interest is time from start of the study until patients require their first dialysis or until end of the study. Patients who died during follow-up before they required dialysis are censored at the date of their death (variable `time`). The event of interest is dialysis with 0 and 1 representing no dialysis and dialysis, respectively. The exposure of main interest urine osmolarity was $\log_2$-transformed prior to analysis, because of its skewed distribution (`log2Uosm`). Based on a-priori knowledge nine potential confounders measured at baseline were identified: $\log_2$ of creatinine clearance (ml/min; `log2CCL`), $\log_2$ of proteinuria (g/L; `log2Prot`), presence of polycystic kidney disease (`pkd`), whether or not beta-blockers (`bblock`), diuretics (`diur`), or angiotensin-converting enzyme inhibitors and Angiotensin II type 1 receptor blockers (ACEI/ARBs; `acei`) were used, age in decades (`age10`), gender (`gender`), and mean arterial pressure (mmHg, `map`).

To estimate the exposure-outcome relationship using ABE the following SAS code can be submitted:

```
%abe(data = example, time = time, cens = dialysis, include = log2Uosm,
     varlist = log2CCL log2Prot bblock pkd diur age10 acei map gender,
     first = none, pmulti = 0.20, tau = 0.05, proc = phreg,
     options = %str(rl), logfile = _logfile, confounders = _confounders,
     print = 0, notes = 1, final = 1);
```

The exposure of interest, `log2Uosm`, will be forced into every working model (option `include`). The other nine explanatory variables will only be included into the final main effects model, if they are significant at a significance threshold `pmulti` of 0.20 or if their inclusion into the model changes at least one other standardized hazard ratio by at least 5% (option `tau`). Only the final model (option `print` and `final`) will be printed. As requested in the `options` risk limits will be added to the final model. The selection process can be closely traced by inspecting the notes in the SAS log-file (option `notes`) and it is also summarized in the `_logfile`. The names of variables selected into the final model are saved in the `_confounders` file.

Except `log2Uosm` seven of the nine potential confounders were selected into the final main effects model. Six of them were selected because they reached significance (`pmulti`), only `acei` was selected because it changes the hazard ratio of `log2Prot` by more than 5% (**Figure 3**).

```
Final model selected by ABE with first = NONE, puni = 0.20, pmulti = 0.20, tau = 0.05

The PHREG Procedure

                Model Information

Data Set                  EXAMPLE
Dependent Variable        time
Censoring Variable        status
Censoring Value(s)        0
Ties Handling             BRESLOW

Number of Observations Read         245
Number of Observations Used         245

Summary of the Number of Event and Censored Values

                                   Percent
   Total       Event    Censored    Censored

     245          95         150       61.22

         Model Fit Statistics

                 Without           With
Criterion     Covariates     Covariates

-2 LOG L         987.191        791.686
AIC              987.191        805.686
SBC              987.191        823.563

        Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio       195.5050        7         <.0001
Score                  167.9333        7         <.0001
Wald                   140.0343        7         <.0001

                              Analysis of Maximum Likelihood Estimates

              Parameter    Standard                                 Hazard      95% Hazard Ratio
Parameter DF  Estimate     Error       Chi-Square    Pr > ChiSq     Ratio       Confidence Limits

log2Uosm  1   0.71702      0.30485        5.5323        0.0187      2.048       1.127       3.723
log2CCL   1  -1.99816      0.23168       74.3850        <.0001      0.136       0.086       0.214
log2Prot  1   0.66026      0.08484       60.5710        <.0001      1.935       1.639       2.285
bblock    1   0.45731      0.22300        4.2055        0.0403      1.580       1.020       2.446
pkd       1   1.09230      0.34629        9.9496        0.0016      2.981       1.512       5.877
diur      1   0.37068      0.22286        2.7664        0.0963      1.449       0.936       2.242
acei      1  -0.34352      0.36314        0.8949        0.3442      0.709       0.348       1.445
```

Figure 3: Final model selected by ABE.

The analyst can closely trace the variables through the selection process by inspecting the SAS log (Figure 4).

```
16135  %abe(data = example, time = time, cens = status,
16136       varlist = log2CCL log2Prot bblock pkd diur age10 acei map gender,
16137       include = log2Uosm, proc = phreg, first = none,
16138       pmulti = 0.20, tau = 0.05, cycles = 10, options = %str(rl),
16139       confounders = _confounders, logfile = _logfile, print = 1, notes = 1,final = 1);

Variable log2Uosm is INCLUDE.

Variable log2CCL
Variable log2Prot
Variable bblock
Variable pkd
Variable diur
Variable age10
Variable acei
Variable map
Variable gender

...

NOTE: ***********************************************
NOTE: ***************** Stage 2 CYCLE 1 ******************
NOTE: ***********************************************

...

NOTE: Processing Variable 1, log2Uosm
NOTE: is include.
NOTE: Variable         1 will be kept.

NOTE: Processing Variable 2, log2CCL
NOTE: p-value =         0 < 0.20
NOTE: Variable         2 will be kept.

NOTE: Processing Variable 3, log2Prot
NOTE: p-value = 1.742E-13 < 0.20
NOTE: Variable         3 will be kept.

NOTE: Processing Variable 4, bblock
NOTE: p-value = 0.0569955 < 0.20
NOTE: Variable         4 will be kept.

NOTE: Processing Variable 5, pkd
NOTE: p-value = 0.0021386 < 0.20
NOTE: Variable         5 will be kept.

NOTE: Processing Variable 6, diur
NOTE: p-value = 0.0914799 < 0.20
NOTE: Variable         6 will be kept.

NOTE: Processing Variable 7, age10
NOTE: p-value = 0.5927991
NOTE: ...Passive variable log2Uosm: stand. delta= |-0.014723|
NOTE: ...Passive variable log2CCL: stand. delta= |0.0130174|
```

```
NOTE: ...Passive variable log2Prot: stand. delta= |0.0074404|
NOTE: ...Passive variable bblock: stand. delta= |-0.005547|
NOTE: ...Passive variable pkd: stand. delta= |-0.000916|
NOTE: ...Passive variable diur: stand. delta= |-0.021088|
NOTE: ...Passive variable acei: stand. delta= | 0.001047|
NOTE: ...Passive variable map: stand. delta= |-0.004984|
NOTE: ...Passive variable gender: stand. delta= | 0.018736|
NOTE: Variable         7 is neither a confounder nor significant and might be dropped.

NOTE: Processing Variable 8, acei
NOTE: p-value = 0.3101402
NOTE: ...Passive variable log2Uosm: stand. delta= |0.0017327|
NOTE: ...Passive variable log2CCL: stand. delta= |-0.002522|
NOTE: ...Passive variable log2Prot: stand. delta= |-0.054686| >= 0.0487902
NOTE: ...Passive variable bblock: stand. delta= |0.0000527|
NOTE: ...Passive variable pkd: stand. delta= |-0.000641|
NOTE: ...Passive variable diur: stand. delta= |-0.017804|
NOTE: ...Passive variable age10: stand. delta= |0.0016863|
NOTE: ...Passive variable map: stand. delta= |-0.018567|
NOTE: ...Passive variable gender: stand. delta= |-0.003173|
NOTE: Variable         8 is a confounder and will be kept.

NOTE: Processing Variable 9, map
NOTE: p-value =  0.728678
NOTE: ...Passive variable log2Uosm: stand. delta= |0.0106589|
NOTE: ...Passive variable log2CCL: stand. delta= | -0.00133|
NOTE: ...Passive variable log2Prot: stand. delta= |0.0052343|
NOTE: ...Passive variable bblock: stand. delta= |0.0071485|
NOTE: ...Passive variable pkd: stand. delta= |0.0050525|
NOTE: ...Passive variable diur: stand. delta= | -0.00384|
NOTE: ...Passive variable age10: stand. delta= |0.0028433|
NOTE: ...Passive variable acei: stand. delta= |0.0065761|
NOTE: ...Passive variable gender: stand. delta= |0.0021673|
NOTE: Variable         9 is neither a confounder nor significant and might be dropped.

NOTE: Processing Variable 10, gender
NOTE: p-value = 0.5764517
NOTE: ...Passive variable log2Uosm: stand. delta= | 0.024175|
NOTE: ...Passive variable log2CCL: stand. delta= |-0.001295|
NOTE: ...Passive variable log2Prot: stand. delta= |0.0108175|
NOTE: ...Passive variable bblock: stand. delta= |0.0075574|
NOTE: ...Passive variable pkd: stand. delta= |-0.000231|
NOTE: ...Passive variable diur: stand. delta= |0.0118738|
NOTE: ...Passive variable age10: stand. delta= |-0.018882|
NOTE: ...Passive variable acei: stand. delta= |0.0019852|
NOTE: ...Passive variable map: stand. delta= | 0.003829|
NOTE: Variable        10 is neither a confounder nor significant and might be dropped.

NOTE: ***********************************************
NOTE: ***************** Stage 2 CYCLE 2 ******************
NOTE: ***********************************************

...

NOTE: ***********************************************
NOTE: ***************** Stage 2 CYCLE 3 ******************
NOTE: ***********************************************

...

NOTE: ***********************************************
NOTE: ***************** Stage 2 CYCLE 4 ******************
```

```
NOTE: ********************************************

NOTE: Variable list after cycle 3 :

      log2Uosm log2CCL log2Prot bblock pkd diur acei

...

NOTE: Processing Variable 1, log2Uosm
NOTE: is include.
NOTE: Variable         1 will be kept.

NOTE: Processing Variable 2, log2CCL
NOTE: p-value =         0 < 0.20
NOTE: Variable         2 will be kept.

NOTE: Processing Variable 3, log2Prot
NOTE: p-value = 7.105E-15 < 0.20
NOTE: Variable         3 will be kept.

NOTE: Processing Variable 4, bblock
NOTE: p-value = 0.0402939 < 0.20
NOTE: Variable         4 will be kept.

NOTE: Processing Variable 5, pkd
NOTE: p-value = 0.0016089 < 0.20
NOTE: Variable         5 will be kept.

NOTE: Processing Variable 6, diur
NOTE: p-value = 0.0962626 < 0.20
NOTE: Variable         6 will be kept.

NOTE: Processing Variable 7, age10
NOTE: not in model.

NOTE: Processing Variable 8, acei
NOTE: p-value = 0.3441553
NOTE: ...Passive variable log2Uosm: stand. delta= |-0.003548|
NOTE: ...Passive variable log2CCL: stand. delta= |-0.001023|
NOTE: ...Passive variable log2Prot: stand. delta= | -0.05403| >= 0.0487902
NOTE: ...Passive variable bblock: stand. delta= |-0.002307|
NOTE: ...Passive variable pkd: stand. delta= |-0.003618|
NOTE: ...Passive variable diur: stand. delta= |-0.016625|
NOTE: Variable         8 is a confounder and will be kept.

NOTE: Processing Variable 9, map
NOTE: not in model.

NOTE: Processing Variable 10, gender
NOTE: not in model.
NOTE: The data set WORK._KEEPS has 1 observations and 10 variables.
NOTE: Exiting IML.
NOTE: PROCEDURE IML used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds

...

NOTE: Variable list after cycle 4 :

      log2Uosm log2CCL log2Prot bblock pkd diur acei
```

```
NOTE: Variable list did not change, iteration stops.

NOTE: Adding variables originally not selected to the variable list

NOTE: Variable 1 keep=            1 keepuni=1
NOTE: Variable 2 keep=            1 keepuni=1
NOTE: Variable 3 keep=            1 keepuni=1
NOTE: Variable 4 keep=            1 keepuni=1
NOTE: Variable 5 keep=            1 keepuni=1
NOTE: Variable 6 keep=            1 keepuni=1
NOTE: Variable 7 keep=           -1 keepuni=1
NOTE: Variable 8 keep=            1 keepuni=1
NOTE: Variable 9 keep=           -1 keepuni=1
NOTE: Variable 10 keep=           -1 keepuni=1

NOTE: No nonsignificant candidates. Macro stops.
```

Figure 4: A part of the SAS log-file noting all selection decisions after submitting the call to the ABE macro.

Per default a data set denoted as `_logfile` is saved, which contains a short summery of the selection process (**Figure 5**), and a second data set denoted as `_confounders` with the names of the variables selected in the final main effects model (**Figure 6**).

Obs	cycle	stage	vars
2	1	Stage 2 full	log2Uosm log2CCL log2Prot bblock pkd diur age10 acei map gender
3	2	Stage 2	log2Uosm log2CCL log2Prot bblock pkd diur age10 acei gender
4	3	Stage 2	log2Uosm log2CCL log2Prot bblock pkd diur acei gender
5	4	Stage 2	log2Uosm log2CCL log2Prot bblock pkd diur acei
6	4	Final model	log2Uosm log2CCL log2Prot bblock pkd diur acei

Figure 5: Data set `_logfile` when submitting the call to the %ABE macro.

Obs	variables
1	log2Uosm
2	log2CCL
3	log2Prot
4	bblock
5	pkd
6	diur
7	acei

Figure 6: Data set `_confounders`.

Standard errors and 95% confidence limits given in **Figure 3** do not account for uncertainty of model selection and consequently may be too small or narrow, respectively. However, bootstrapping methods can be applied to derive at more appropriate estimates for inference. Using the

%ABE_BOOTSTRAP macro, at first B bootstrap resamples are drawn using PROC SURVEYSELECT, secondly for each bootstrap resample %ABE is executed and thirdly the results are summarized in two data sets. In the first data set _bootstrap_variables for each variable (selected in at least one bootstrap sample) inclusion or exclusion in each bootstrap resample is recorded and in the second data set _bootstrap_regcoeffs regression coefficients estimated in each bootstrap resample are saved. The following SAS code can be submitted to apply ABE to 1000 bootstrap resamples:

```
%abe_bootstrap(seed = 12345, B = 1000,
     bootstrap_variables = _bootstrap_variables,
     bootstrap_regcoeffs = _bootstrap_regcoeffs,
     data = example, time = time, cens = dialysis, include = log2Uosm,
     varlist = log2CCL log2Prot bblock pkd diur age10 acei map gender,
     first = none, pmulti = 0.20, tau = 0.05, proc = phreg);
```

Note, that variable selection and estimation of 1000 bootstrap resamples is time-consuming. On a contemporary computer it took approximately 30 minutes to apply ABE to 1000 bootstrap resamples of the urine osmorality data.

Figure 7 and Figure 8 show the first rows of the _bootstrap_variables and _bootstrap_regcoeffs data sets.

Obs	_NAME_	acei	age10	bblock	diur	log2Uosm	log2 CCL	log2 Prot	map	pkd	gender
1	boot1	1	.	1	1	1	1	1	.	1	.
2	boot2	1	.	1	1	1	1	1	.	1	.
3	boot3	.	.	1	1	1	1	1	.	1	.
4	boot4	.	.	.	1	1	1	1	.	1	.
5	boot5	.	1	1	.	1	1	1	.	1	.
6	boot6	1	1	1	1	1	1	1	.	1	1
7	boot7	.	1	.	1	1	1	1	.	1	.
8	boot8	1	.	1	1	1	1	1	1	1	.
9	boot9	1	.	1	1	1	1	1	.	1	.
10	boot10	.	1	1	1	1	1	1	1	1	1

Figure 7: The first ten rows of the data set _bootstrap_variables showing which explanatory variables where selected into the final models of the first ten bootstrap resamples. 1 denotes inclusion into the respective final model. (Note that log2Uosm was forced into every model.)

Obs	log2 Uosm	log2 CCL	log2 Prot	bblock	pkd	diur	age10	acei	map	gender	b
1	0.3208	-1.5085	0.7296	0.7944	0.6146	0.5631	0.0000	-0.5721	0.00000	0.0000	1
2	0.4335	-1.7620	0.7723	0.7273	1.5696	0.4399	0.0000	-0.9542	0.00000	0.0000	2
3	0.8031	-2.3458	0.6800	0.4360	1.3224	0.3929	0.0000	0.0000	0.00000	0.0000	3
4	0.4643	-2.0030	0.7546	0.0000	1.1650	0.4021	0.0000	0.0000	0.00000	0.0000	4
5	1.1195	-2.2272	0.6724	0.4129	1.8188	0.0000	0.2373	0.0000	0.00000	0.0000	5
6	0.6540	-2.6411	0.8268	0.7697	1.1421	0.7635	-0.2623	-0.3349	0.00000	0.4659	6
7	-0.0070	-2.2868	0.6295	0.0000	0.5994	0.5550	-0.3086	0.0000	0.00000	0.0000	7
8	0.0846	-1.7488	0.6585	0.3880	1.2440	0.8001	0.0000	-0.9625	0.02712	0.0000	8
9	0.5041	-2.7827	0.8554	0.7153	1.3656	0.3561	0.0000	-0.4386	0.00000	0.0000	9
10	-0.1469	-1.9398	0.6239	0.6781	0.8166	0.5141	-0.2207	0.0000	0.02999	0.3978	10

Figure 8: The first ten rows of the data set `_bootstrap_regcoeffs` stating the regression coefficients in the final models of the first ten bootstrap resamples. Explantory variables not selected into the respective final model have a regression coefficient of 0.

Furthermore, as good statistical practice suggests the final main effects model should be tested and checked thoroughly with regards to the linearity assumption, inclusion of pairwise product term interactions, goodness-of-fit, influential observations, model stability, and so on, leading to a final multivariable model.

6. Availability, license and disclaimer

The macro is available under a GNU GPL license, version 2 (http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html) at http://cemsiis.meduniwien.ac.at/kb/wf/software/statistische-software/abe/.

The license text can be accessed at http://www.gnu.org/licenses/gpl-2.0.txt.

References

1. Hosmer, DW Jr., Lemeshow S, and May S (1999) Applied survival analysis: Regression modeling of time to event data. New York: John Wiley & Sons, Chapter 5 - Model Development.

2. Dunkler D, Plischke M, Leffondré K, Heinze G (2014) Augmented backward elimination: A pragmatic and purposeful way to develop statistical models (submitted).

3. SAS Institute Inc (2010) SAS/STAT Software, version 9.3. Cary, NC.

4. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JD (2000) Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 19: 1059-1079.

5. Maldonado G, Greenland S (1993) Simulation study of confounder-selection strategies. Am J Epidemiol 138: 923-936.

6. Van der Weele TJ, Shpitser I (2011) A new criterion for confounder selection. Biometrics 67: 1406-1413.

7. Rubin DB (2009) Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? Stat Med 28: 1420-1423.

8. Glymour MM, Weuve J, Chen J (2008) Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: Measurement, selection, and bias. Neuropsychol Rev 18: 194-213.

9. Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate Cox proportional hazards models. Stat Med 24: 1713-1723.

10. Plischke M, Kohl M, Bankir L, Handisurya A, Heinze G, Haas M (2014) Urine osmolarity and risk of dialysis initiation in a chronic kidney disease cohort - a possible titration target? PLoS ONE 9: e93226.