UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Master's thesis
(Magistrsko delo)

Developing statistical regression models by using variable selection techniques

(Izbira spremenljivk v multivariatnih modelih)

Ime in priimek: Slađana Babić
Študijski program: Matematične znanosti, 2. stopnja
Mentor: doc. dr. Rok Blagus

Koper, junij 2017

Ključna dokumentacijska informacija

Ime in PRIIMEK: Slađana BABIĆ

Naslov magistrskega dela: Izbira spremenljivk v multivariatnih modelih

Kraj: Koper

Leto: 2017

Število listov: 79 Število slik: 22 Število tabel: 18

Število referenc: 25

Mentor: doc. dr. Rok Blagus

UDK: 519.246(043.2)

Ključne besede: ABE metoda, Akaikejev informacijski kriterij, bayesovski informacijski kriterij, linearna regresija, Cox regresija, posplošeni linearni regresijski modeli.

Math. Subj. Class. (2010): 62H99

Izvleček:
Cilj naloge je posplošiti ABE metodo za izbiro spremenljivk katero so pred kratkim predlagali Dunkler et al. Metoda je dostopna samo v programskem jeziku SAS, tako da je naš cilj bil sprogramirati metodo tudi v programskem jeziku R. ABE metoda izbira spremenljivke na podlagi značilnosti in na podlagi standardizirane spremembe v oceni koeficienta. Pri tem uporablja aproksimacijo za spremembo v oceni koeficienta namesto eksaktnega izračuna. Potencialna težava metode ABE je da spremenljivke izbira na podlagi njihove značilnosti. Zaradi tega, v nalogi smo predstavili posplošitve metode ABE, katere omogočajo tudi uporabo informacijskih kriterij, Akaikejev informacijski kriterij (AIC) ali bayesovski informacijski kriterij (BIC). Pri pripravi R funkcije smo omogočili tudi uporabo eksaktnega izračuna za spremembo v oceni koeficienta.
V nalogi smo najprej predstavili problem izbire spremenljivk in obstoječe metode za izbiro spremenljivk in predstavili znane težave, ki jih imajo te metode. Glede na to da je metoda ABE sprogramirana za linearno, logistično in Cox regresijo, predstavili smo posplošene linearne regresijske modele in Coxov model. Natančno smo predstavili metodo ABE kot tudi njene izboljšave. Predstavili smo R paket, oziroma kodo in opis delovanja funkcije; kaj so argumenti in kaj funkcija izračuna. Na koncu smo s simulacijami prikazali delovanje predlaganih izboljšav metode ABE in jih primerjali z osnovno metodo ABE ter s preostalimi metodami, ki so že sprogramirane v R.

Key words documentation

Name and SURNAME: Slađana BABIĆ

Title of the thesis: Developing statistical regression models by using variable selection techniques

Place: Koper

Year: 2017

Number of pages: 79 Number of figures: 22 Number of tables: 18

Number of references: 25

Mentor: Assist. Prof. Rok Blagus, PhD

UDC: 519.246(043.2)

Keywords: ABE method, Akaike information criterion, Bayesian information criterion, linear regression, Cox regression, generalized linear models.

Math. Subj. Class. (2010): 62H99

Abstract: In this master's thesis we reviewed the most common variable selection procedures, their advantages and disadvantages. Among others we presented augmented backward elimination method recently proposed by Dunkler et al. which is a combination of backward elimination procedure and approximated change-in-estimate criterion. The method proposed by Dunkler et al. is only available in SAS, so our aim was to make an R package which will implement their method for several statistical models. Since this method chooses variables based on their significance and approximated change-in-estimated, we extended it such that information criteria AIC and BIC can also be used and that we can choose between approximated and exact change-in-estimate. Also, extended augmented backward elimination is available for all generalized linear models, not just for logistic regression. We performed extensive simulation studies.

Acknowledgement

I would like to thank my thesis advisor, assist. prof. Rok Blagus for the help, useful comments and remarks.
Posebno se želim zahvaliti mojoj porodici za bezuslovnu podršku, a naročito mom bratu Zoranu za šaljive komentare i razumijevanje tokom pisanja ovog rada.

Contents

List of Tables

List of Figures

List of Abbreviations

i.e. that is
e.g. for example
etc. et cetera.

1 Introduction

From the literature concerning model selection you can notice that statisticians are divided into two groups when it comes to variable selection methods. On one side are those who believe that variable selection is necessary for model building, while on the other side are those who consider that using variable selection only makes things worse. Often variable selection is identified with model selection but variable selection is indeed a part of it. Even though in this master thesis we will be interested in varibale selection, first we will go through what is considered as a model selection and what the good model is. Of course, as we could expect, there are many answers to these questions. The simplest definition of model selection would be:*"model selection is the task of selecting a statistical model from a set of candidate models, given data"* [23]. The question now arises, what do we mean by a statistical model and what is a good model? *Statistical models are simple mathematical rules derived from empirical data describing the association between an outcome and several explanatory variables* [6].
A statistical model represents, often in considerably idealized form, the data-generating process [24]. Herman Ader quoting Kenneth Bollen said: *"A model is a formal representation of a theory".* One of the most famous statements about statistical modelling is:*"All models are wrong"* and it is attributed to the statistician George Box. This sentence is from his paper published in the Journal of the American Statistical Association in 1976. He repeated this well-known aphorism several times more in a slightly modified form. In his 1987 book, Empirical Model-Building and Response Surfaces he says the following:*"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful".* The second edition of his book Statistics for Experiments published in 2005 also includes this aphorism:*"The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful".*
As one of the responses to this statement, we will quote the following which is stated in the Burnham and Anderson book on model selection:*"A model is a simplification or approximation of reality and hence will not reflect all of reality... Box noted that "all models are wrong, but some are useful." While a model can never be "truth", a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless."*

Also the statistician Sir David Cox made the following statement about the aphorism: *"... it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models, especially substantive ones, do not seem essentially different from other kinds of model".*
Back to the question what is a good model. Of course, in practise it may happen that sometimes a poor model may give acceptable results, while under other circumstances it can also happen that good model may fail to give the required answers. Nevertheless, there should be some criteria and any person with experience in modelling should try to adhere to them. Definitely one of the most important characteristics of a good model is simplicity. Every statistician will agree that among models with roughly equal predictive or explanatory power, the simplest one is the most desirable.
As we have already mentioned in this master thesis we are interested in variable selection. First of all we will present what are possible approaches when it comes to variable selection and we will discuss advantages and disadvantages of variable selection procedures. Recently Dunkler et al. proposed augmented backward elimination, where a standardized change-in-estimate criterion on the quantity of interest usually reported and interpreted in a model for variable selection is added to the backward elimination. We review this, and some other variable selection techniques which are the most common in medical research. The method proposed by Dunkler et al. is only available in SAS, our aim was to make an R package which will implement their method for several statistical models. In chapter three we review some of the statistical models which are the most common in medical research, i.e. linear, logistic and Cox proportional hazards regression. In chapter four we review in details augmented backward elimination method; where does the idea for it come from and how change-in-estimate criterion can be approximated. After that we propose possible generalization of augmented backward elimination. Namely, since this method chooses variables based on their significance and approximated change-in-estimated, we extended it such that information criteria Akaike information criterion (AIC) and Bayesian information criterion (BIC) can also be used and that we can choose between approximated and exact change-in-estimate. Also, extended augmented backward elimination is available for all generalized linear models, not just for logistic regression. For the end we present results of an extensive simulation study used to evaluate different variable selection techniques for several statistical models. In the simulation study we evaluated the bias of the regression coefficients and their mean squared error.

2 Variable selection

Research in variable selection started in early 1960s and despite the fact that until now extensive research has been conducted, there is no unique solution that would completely solve the problem about variable selection. Hocking said: *"One reason for the lack of resolution of the problem is the fact that it has not been well defined"* [12]. As he pointed out, there is not a single problem, but several problems for which different answers might be appropriate.

A brief overview of the content of this chapter follows. First of all we will present reasons why do we need variable selection techniques; after that we will examine various proposed solutions to the general problem of variable selection. Also, we will include a discussion of the importance of various criteria to certain goals. For the end, we will present problems and issues of different variable selection methods.

Variable selection is used when the analyst has a series of potential explanatory variables but does not have or does not use the necessary subject matter knowledge to enable him to choose "important" variables to include in the model. Very often in biomedical studies it is common to have data where the number of explanatory variables k greatly exceeds the number of observations n for which these covariates are available. Therefore the development of methods for variable selection, especially in these cases was needed. In order to explain the variability of the response as much as possible, usually we take into consideration as many explanatory variables as we can. But of course it may happen that the whole set of variables is too large for using them all in the model. Therefore, sometimes the precision of the fit is improved by the reduction of the number of explanatory variables. For example, explanatory variables for which the associated regression coefficients are not significantly different from zero may increase the variance of the estimates. Using the entire set of variables may also bring about numerical difficulties due to multicollinearity, since for large number of explanatory variables it is more likely that there are at least two highly correlated variables.

Besides these statistical reasons, there are a variety of practical and also economical reasons for reducing the number of explanatory variables. In addition it should be emphasized what are the consequences of incorrect model specification regardless of whether deleting the relevant variables or either because of retaining irrelevant ones.

However, the selection of the best subset of the explanatory variables is not trivial problem first of all because the number of subsets to be considered grows rapidly with the number of explanatory variables.

2.1 Variable Selection Approaches

Many statisticians will agree that although variable selection is very important, at the same time it is very difficult and demanding part of data analysis. Burnham and Anderson say the following:"*The so-called variable selection is arguably the most used and misused application area of model selection*" [4]. There are two main approaches towards variable selection: all possible regressions approach and sequential strategies. In case we have k explanatory variables, all $2^k - 1$ possible models are fitted and the best model is chosen according to some criteria, like information criterion or adjusted R^2. This method is useful when the number of explanatory variables is not so large, meaning that is feasible to fit all possible models. The R function $'regsubsets()'$ in the library $'leaps'$ can be used for regression subset selection.
Sequential methods are useful when the number of explanatory variables is large. In this case, it is more efficient to use a search algorithm (e.g., Forward selection, Backward elimination and Stepwise regression) to find the best model. Even though many variable selection techniques were proposed until now, we will discuss just some of the more common ones.
One of the most commonly used methods are the oldest approaches to variable selection based on different sequential tests. Their biggest advantage is the simplicity of choosing explanatory variables but on the other side their use is not well-justified theoretically. Among this type of techniques, the most commonly used are forward, backward and stepwise sequential testing, also known as stepwise methods.
Forward selection starts with a "null model" or with an "intercept only model" and sequentially adds explanatory variables. At the first step it considers all one-variable models and adds the variable with the best result based on some criterion. The criterion could be lowest p-value, lowest AIC, highest R^2, lowest Mallow's Cp, etc. We repeat this process, always adding one variable at a time, until the criterion is not met.
Backward selection starts with a full model and sequentially removes non-significant explanatory variables, one variable at a time.
Stepwise sequential testing represents a combination of forward and backward procedure, that is, it alternates between adding and removing variables. Stepwise regression term was introduced by Efroymson in 1960 in the context of linear models. This selection algorithm involves the inclusion of a single variable to the model or removal of a single variable from the model in order to improve its quality. The criterion used to

select the X_i variable to add or remove from the regression is as follows:

1 If the variance contribution of a variable in the regression is insignificant at a specified F level, the variable is removed from the regression.
If no variable is to be removed, then the following criterion is used.

2 If the variance reduction obtained by adding the variable to the regression is significant at a specified level F, this variable is entered into the regression.

R function for stepwise model selection is called "step". It can do forward or backward selection, or both, and the user can specify the smallest model to consider (so those variables are always included). It can, however, only use AIC or BIC as the selection criteria.

Even though variable selection techniques have been widely investigated in the literature, one has to be aware of the weak points of the available techniques. For instance, the stepwise techniques do not necessarily give the best subset of variables. Furthermore, these procedures induce a ranking on the explanatory variables which is often misused in practice. The order of inclusion or removal of variable can be misleading. It may, for example, happen that the first variable entered in forward selection becomes unnecessary in the presence of other variables. Also, it is possible that forward and backward procedures give the totally different best subset of variables.

As we have already mentioned, stepwise methods have no firm justification in statistical theory, but that is not their only disadvantage. These methods will never consider all possible subsets of variables so obviously there is a big chance that they will miss a good or maybe the best subset.

Another variable selection method widely used and at the same time very criticized by many statisticians is the so called screening, pretesting or bivariate analysis. Screening is a method that usually uses t-test for each explanatory variable in order to decide which are significant and to remove variables from model that are not significant. This method ignores the possibility that variables may be correlated, and therefore it can exclude important variable given that the significance of a explanatory variables depends on which other variables are in the model.

All decisions in previously mentioned methods are based on significance test but the crucial thing is that classical testing and variable selection process do not always deal with the same questions. Whichever method we choose for variable selection, we can not expect that it will work perfectly.

Nowadays, there is enormous amount of data from medicine, social science, finance, geography, geophysics, genetics, and engineering. Of course, there are variable selection techniques that are not theoretically enough justified. Also, there are so many examples in practice with misapplication of variable selection methods. It is crucial to understand

that principle "measure everything that is easy to measure and let the computer sort it out" simply does not work.
Even though variable selection is very important for many researchers in areas of applications for which datasets of huge number of variables are available. Some of these areas are biostatistics analysis, text processing of internet documents, etc. The aim of variable selection is three-fold. Namely, the objective is

1) to improve prediction performance of the explanatory variables

2) to evaluate an explanatory variable of primary interest

3) to identify the important explanatory variables of an outcome.

One of the essential things when choosing the best subset of variables is to know what we are doing. The same "rules" are not applicable for different aims. There is a big difference between for example selecting a best model in order to find the relationships between explanatory variables and the response variable, and prediction on the other side. This means we should be careful while dropping variables. For first situation we will drop exactly those variables that can not reasonably be causally related with response variable, while for the second situation we will exclude variables that can not reasonably be related to outcome for prediction. But that is not all. What we often forget is that things like sample size and subject-matter knowledge must be taken into consideration. Before choosing any variable selection technique, we should first think of possible limitation because of the sample size and how to use prior knowledge if there is any.
Many statisticians warn that it is not correct and that we can not rely on the results obtained by using automatic packages that pick a model and then use for example least squares to estimate regression coefficients using the same data. In this regard we quote Miller (2002): *"Many statisticians and other scientists have long been aware that the so-called significance levels reported by subset selection packages are totally without foundation, but far fewer are aware of the substantial biases in the (least squares or other) regression coefficients"* [15].
He also believes that variable selection is an *"unclean"* and *"distasteful"* area of statistics which can best be described using terms such as *"fishing expeditions", "torturing the data until they confess", "data mining",* and others.
Despite all the negative sides of variable selection and the fact that there is very little theory to handle this very common problem, in many situations it is unavoidable to use some of its techniques.
Maybe the problem of variable selection sounds like not so hard problem but let us look what are possible difficulties in the simple case of linear regression. The main

idea in problem of finding a subset of variables whose model fits a data set fairly well in the least-squares sense is the following. For a given set of n observations with k variables, denote observations of variables X_j with $x_{i,j}$ for $i = 1, \ldots, n$, and $j = 1, \ldots, k$. The least squares estimates for the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ denote with $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Given a set of variables $X_1, \ldots, X_k$, our aim is to find a subset of p $(p < k)$ variables for which the sum

$$S = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j} \right) \right)^2$$

is minimized or it has enough small value.

Besides common problems with variable selection techniques, additional difficulties occur since very often it happens that there are some constraints regarding the minimization of S. For example, maybe we want one or more variables to be in selected subset. Another possibility is that maybe one variable may only be included in a selected subset if another variable(s) is also included. Also it does not make sense to include a transformation of some variable if that variable is not included in the subset. Particular topic is about dummy variables which represent categorical variables. If among our variables there is one which can be represented by more than two dummy variables in such cases, it is often required that either all or none of the dummy variables should be in the model.

Nowadays, some models have their own built-in feature selection methods. Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting [19]. Other examples of embedded methods are Regularized trees, Memetic algorithm, Random multinomial logit [25], [5].

2.2 Disadvantages of Variable Selection

Whoever was reading about variable selection he could not help but notice criticism related to it by statistician Frank Harrell. In his book *Regression modelling strategies*, Harrell points on problems that stepwise variable selection brings [10]. Here is a summary of those problems.

1. "It yields R^2 values that have a high bias.
2. The ordinary F and χ^2 test statistics do not have the claimed distribution. Variable selection is based on methods (e.g., F tests for nested models) that were intended to be used to test only prespecified hypotheses.

3. The method yields standard errors of regression coefficient estimates that have a low bias and confidence intervals for effects and predicted values that are falsely narrow.

4. It yields P-values that are too small (i.e., there are severe multiple comparison problems) and that do not have the proper meaning, and the proper correction for them is a very difficult problem.

5. It provides regression coefficients that have a high bias in absolute value and need shrinkage. Even if only a single predictor was being analysed and one only reported the regression coefficient for that predictor if its association with Y was "statistically significant", the estimate of the regression coefficient β is biased (too large in absolute value). To put this in symbols for the case where we obtain a positive association $(\hat{\beta} > 0), E(\hat{\beta}|P < 0.05, \hat{\beta} > 0) > \beta$.

6. Rather than solving problems caused by collinearity, variable selection is made arbitrary by collinearity.

7. It allows us to not think about the problem."

Harrell also emphasizes observations made by Derksen and Keselman. By studying stepwise variable selection, backward elimination, and forward selection they concluded the following:

1. "The degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model.

2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.

3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.

4. The population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model."

2.3 Five Myths About Variable Selection

Dunkler and Heinze in their article *Five myths about variable selection*, discuss what are the problems related with variable selection and how some misunderstandings of crucial concepts can be misleading for it [11]. They consider the following issues:

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

   Result about the minimum number of events per variable which was derived from the simulation studies refers to a priori fixed models which are not the consequence of the variable selection. We need to bear in mind that variable selection introduce some additional insecurity in parameter estimation so in order to get reliable results we need to have much more events per variable. Based on the above mentioned result it is enough to have 5-15 events, but if we use some variable selection technique then we need at least 50 events per variable.

2) *Only variables with proven univariable model significance should be included in a model.*

   Univariable prefiltering or bivariable analysis cannot properly control for possible confounding. Using this method will not add stability to the selection process, moreover it will introduce a source of significant error thus as a consequence we can include or reject inappropriate variables.

3) *Insignificant effects should be eliminated from a model.*

   In general, the values of the regression coefficients depend on which other variables are omitted, that is which variables are in the model. So, one of the possible situations is the following. In case we would eliminate a confounder that would lead to a change of the coefficient of another variable, moving it close to zero. In other words, that variable would change from significant to insignificant meaning that on the following step we would eliminate it even though it was important predictor.

4) *The reported P-value quantifies the type I error of a variable being falsely selected.*

   We should always bear in mind that standard software report p-values from the last model, forgetting the previous steps of variable selection, so this p-values are misleading. Therefore, p-value does not quantify the type I error of a variable being falsely selected. Besides, p-value does not quantify the type I error at all.

5) *Variable selection simplifies analysis.*

   Variable selection is quite a demanding process. First of all we have to choose which variable selection technique we will use. After that to choose the values for selection parameters, for example the significance level at which we will include the variable in a model. For the end the variable selection should always be followed by sensitivity analysis on model stability in order to avoid wrong conclusions.

3 Generalized Linear Model and Cox Model

In this chapter we will present two types of statistical models of great importance for which Augmented Backward Elimination has been implemented. Namely, we will present Generalized Linear Model as a generalization of ordinary linear regression and Cox proportional hazards model, sometimes abbreviated to Cox model as a class of survival models.

3.1 Generalized Liner Model

In statistics very often there are many similar terms with different meanings. This is exactly the case with *generalized linear models*. In order not to be confused with general linear models, first of all we will go through definitions of both general and generalized linear models, even though in this section our attention will be focused on generalized linear models.
In statistics term linear model is most commonly used as a synonym for linear regression model.
The general linear model is a statistical linear model which may be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where Y is a matrix with series of multivariate measurements, X is a matrix that might be a design matrix, B is a matrix containing parameters that are usually to be estimated and U is a matrix containing errors or noise [21].
The general linear model includes different statistical models, like: ANOVA, ANCOVA, MANOVA, MANCOVA, F-test, t-test and also ordinary linear regression. Furthermore, it represents a generalization of multiple linear regression model to the case of more than one response variable.
In case the errors from general linear model do not follow a multivariate normal distribution, we use generalized linear model.
The basic model for linear regression is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i.$$

In the case $Y_i \sim N(\mu_i, \sigma^2)$, models of this form are basis of most analyses of continuous data. In the following, more general situations when response variables have distributions other than the normal distribution and in the case when the relationship between response variable and explanatory variables need not to be of the simple linear form, generalized linear model allows us to use methods analogous to those developed for linear model.
GLM represents a natural generalization of classical linear model. In GLM each outcome $\mathbf{Y}$ of the dependent variables is assumed to be generated from a particular distribution in the exponential family.
A single-parameter exponential family is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form

$$f_X(x \mid \theta) = h(x) \exp(\eta(\theta)T(x) - A(\theta))$$

where $h(x), \eta(\theta), T(x)$ and $A(\theta)$ are known functions.
An alternative, equivalent form often given is

$$f_X(x \mid \theta) = h(x)g(\theta) \exp\left(\eta(\theta)T(x)\right)$$

The value θ is called the parameter of the family [20].
Equivalently, the distribution belongs to the exponential family if it can be written in the form

$$f_X(x \mid \theta) = \exp\left[a(x)b(\theta) + c(\theta) + d(x)\right].$$

If $a(x) = x$, the distribution is said to be in canonical form.

Generalized linear model has three components:

1. Response variables $Y_1, \cdots, Y_n$, which are assumed to share the same distribution from the exponential family.

2. A set of parameters β,

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

and explanatory variables

$$\begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & & x_{nk} \end{bmatrix}$$

3. A monotone link function g such that

$$g(\mu_i) = x_i^T \beta$$

where $\mu_i = E(Y_i)$.

Link function represents the relationship between the linear predictor and the mean of the distribution function. In the initial formulation of generalized linear models by Nelder and Wedderburn (1972) link function is a simple mathematical function.
Response variables in GLM have distribution different from the Normal distribution so they may not range from $-\infty$ to $+\infty$. That is why we need the link function, because it links the expected value of the response variable Y_i to the linear term $x_i^T \beta$ in such a way that the range of non-linearly transformed mean $g(E[Y_i])$ ranges from $-\infty$ to $+\infty$. Function g must be smooth and invertible.
One of the advantages of using generalized linear models is the use of nice properties of the normal distribution shared by a wider class of distributions called the exponential family of distributions.

3.1.1 Estimation method for Generalized Linear Model

Typical estimation methods for generalized linear model are maximum likelihood estimation and Bayesian approach.
Bayesian methods involve using Laplace approximations or some type of Markov chain Monte Carlo method such as Gibbs sampling for approximation of the posterior distribution, since it can not be found in closed form. Here, we will go through maximum likelihood approach.
Scoring algorithm is a form of Newton's method used in statistics to find the maximum likelihood estimators numerically. Let us recall the Newton-Raphson approximation for numerical solving $f(x) = 0$. We start with an initial value and use the fact that the tangent line to a curve is a good approximation to the curve near the point of tangency. The slope of f at a value x_n is given by

$$f'(x_n) = \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n}.$$

If x_{n+1} is the required solution such that $f(x_{n+1})$ is zero, solving for x_{n+1} gives

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

We repeat this until the process converges.
In order to maximize the log likelihood function we require its derivative with respect

to θ to be equal to zero. Or in other words we are looking for the solution of the equation $S(\theta) = 0$ where S is a score function. The estimating equation is

$$\theta_{n+1} = \theta_n - \frac{S_n}{S_n'}$$

Given that the response variables $Y_1, \cdots, Y_N$ have the distribution belonging to the exponential family we can write their joint probability density function as

$$\begin{aligned} f(Y_1, \cdots, Y_N|\theta) &= \prod_{i=1}^{N} \exp\left[y_i b(\theta_i) + c(\theta_i) + d(y_i)\right] \\ &= \exp\left[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]. \end{aligned}$$

The log likelihood function is

$$l = \sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i).$$

We want to estimate parameters β which are related to the response variables Y_i through their expected values $E[Y_i] = \mu_i$ and $\eta_i = g(\mu_i) = x_i^T\beta$. Note, it can be shown that

$$E(Y_i) = -\frac{c'(\theta_i)}{b'(\theta_i)}$$

and

$$var(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{(b'(\theta_i))^3}.$$

Denote with l_i the log-likelihood function for each Y_i, i.e.

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i).$$

Therefore,

$$S_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}.$$

For better clearness, we will consider each term from the last expression separately. So, we have the following

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i),$$

since $\mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)}$.
On the other hand,

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}.$$

Calculating $\dfrac{\partial \mu_i}{\partial \theta_i}$ we get

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{b'(\theta_i)^2} = b'(\theta_i)\text{var}(Y_i).$$

For the last term we get

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$$

Therefore,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \text{ and } S_j = \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right].$$

For maximum likelihood estimation, it is common to approximate S_n' by its expected values. In this way, we exactly obtain an information matrix. Given that Fisher information matrix as elements has

$$\left[\mathcal{I}(\theta) \right]_{i,j} = \text{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X;\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X;\theta) \right) \middle| \theta \right]$$

in our case elements of the information matrix are terms $\mathcal{I}_{jk} = E[S_j S_k]$. Hence, the term $\mathcal{I}_{jk}$ is

$$\begin{aligned}
\mathcal{I}_{jk} &= E\left\{ \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^{N} \left[\frac{(y_l - \mu_l)}{\text{var}(Y_l)} x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\} \\
&= \sum_{i=1}^{N} \frac{E[(Y_i - \mu_i)^2] x_{ij} x_{ik}}{[\text{var}(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\
&= \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{[\text{var}(Y_i)]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.
\end{aligned}$$

Note, from the last expression, information matrix can be written as

$$\mathcal{I} = X^T V X$$

where V is the $N \times N$ diagonal matrix with elements

$$v_{ii} = \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

Finally, our estimating equation has the following form

$$b_{n+1} = b_n + \left[\mathcal{I}_n \right]^{-1} S_n$$

where b_{n+1} is the vector of the estimates of the parameters $\beta_1, \cdots \beta_p$ at the $(n+1)$st iteration and $\left[\mathcal{I}_n \right]^{-1}$ is the inverse of the information matrix. Multiplying both sides of the last expression by $\mathcal{I}_n$ we obtain

$$\mathcal{I}_n b_{n+1} = \mathcal{I}_n b_n + S_n.$$

The right-hand side of the equation represents the vector with elements

$$\sum_{k=1}^{p}\sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{\mathrm{var}(Y_i)}\Big(\frac{\partial\mu_i}{\partial\eta_i}\Big)^2 b_{nk}+\sum_{i=1}^{N}\frac{(y_i-\mu_i)}{\mathrm{var}(Y_i)}x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right).$$

Introducing a new variable z where

$$z_i=\sum_{k=1}^{p}x_{ik}b_{nk}+(y_i-\mu_i)\Big(\frac{\partial\eta_i}{\partial\mu_i}\Big)$$

the right-hand side can be written as X^TVz.
After this, estimating equation multiplied by $\mathcal{I}_n$ can be written as

$$X^TVXb_{n+1}=X^TVz.$$

We see that in general, V and z depend on b, so the estimating equation must be solved iteratively. Hence, for generalized linear models, maximum likelihood estimators are obtained by an iterative weighted least squares procedure.

3.1.2 Examples

Some of the most commonly used distributions belonging to the exponential family are: normal, exponential, gamma, chi-squared, Bernoulli, categorical, Poisson, beta, Dirichlet, Wishart and inverse Wishart. For these distributions a generalized linear model approach can be used to fit response with one of those distributions to explanatory variables.
Below is a table of several exponential-family distributions in common use along with the link functions.

Table 1: Examples of exponential family distributions

Distribution	**Support of distribution**	**Link name**	**Link function**
Normal	$(-\infty,\infty)$	Identity	$\mathbf{X}\boldsymbol{\beta}=\mu$
Exponential Gamma	$(0,+\infty)$	Inverse	$\mathbf{X}\boldsymbol{\beta}=\mu^{-1}$
Poisson	integer: $0,1,2,\ldots$	Log	$\mathbf{X}\boldsymbol{\beta}=\ln(\mu)$
Binomial	integer: $0,1,\ldots,N$	Logit	$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$
Inverse Gaussian	real: $(0,+\infty)$	Inverse squared	$\mathbf{X}\boldsymbol{\beta}=\mu^{-2}$

Given that Augmented Backward Elimination method has been implemented in a SAS macro for logistic regression and given that this type of generalized linear model is one of the most commonly used, we will present it in more detail.
Logistic regression is a type of regression model where the response variable is categorical. Thus, when we want to explain how a set of explanatory variables X is related to a categorical response variable Y, we can use logistic regression. Some of the examples where logistic regression can be very useful are many problems from medical research. It is appropriate for models involving the presence or absence of a particular disease, risk factors, drug use, death during surgery and similar. Logistic regression can be binomial, ordinal or multinomial. Typically, it is used for estimating the probability of a binary response based on one or more explanatory variables. Usually, the response Y is defined to be $Y = 0$ and $Y = 1$, where $Y = 1$ denotes the occurrence of the event of interest.
Using ordinary linear regression, we could try to model categorical response as a linear function of our explanatory variables, which means we would have common scenario where:

$$E(Y|X) = X\beta,$$

interpreting $E(Y|X)$ as $P(Y = 1)$. Linear model $E(Y|X)$ allows that $P(Y = 1)$ is outside the interval $[0, 1]$. That is why we need some other approach.
In order to introduce logistic regression as a special case of the generalized linear model, first we will define logistic function, i.e. logistic curve. A logistic function or logistic curve is a common "S" shape, with equation:

$$f(t) = \frac{L}{1 + \mathrm{e}^{-k(t-t_0)}},$$

where

e = the natural logarithm base,

t_0 = the t-value of the sigmoid's midpoint,

L = the curve's maximum value, and

k = the steepness of the curve [22].

The standard logistic function is the logistic function with parameters ($k = 1, t_0 = 0, L = 1$) which yields

$$f(t) = \frac{1}{1 + e^{-t}}.$$

We have started with the logistic function since it is useful because it can take any real input $t, (t \in \mathbb{R})$, whereas the output always takes values between zero and one and hence is interpretable as a probability. In the following figure it is shown the graph of the standard logistic function on the $t-$interval $(-6, 6)$.

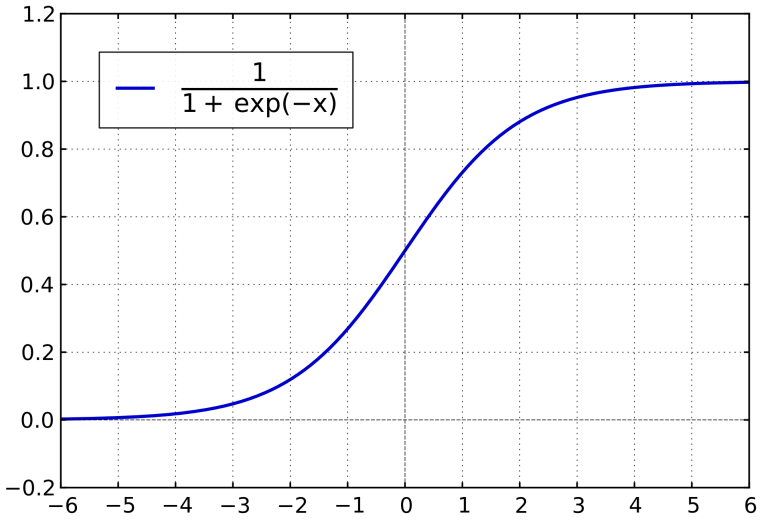

Figure 1: Logistic curve

If t represents a linear combination of multiple explanatory variables, $t = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k$, logistic function can be expressed as:

$$\mu = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}.$$

The inverse of logistic function g allows the expression to be written as a linear model structure,

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

and equivalently, after exponentiating both sides:

$$\frac{\mu}{1-\mu} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}.$$

In comparison with logistic model the major advantage of the linear model is its interpretability. In any of these cases it is not so intuitive to interpret coefficients of logistic regression, whether we are using log odds scale or odds ratios after exponentiating. Literally, the parameter β_i is then the change in the log odds per unit change in X_i if X_i represents a single factor that is linear and does not interact with other factors and if all other factors are held constant [10].

3.1.3 Generalized Linear Model in R

The R language includes a built-in function *glm* to fit Generalized Linear Models. Its main argument is the specification of a model given as a formula object. After the fitting procedure, *glm* returns parameters estimates together with a range of other indicators. Sometimes, as we have already mentioned, for different reasons we would like to drop some terms from the full fitted model. Several R packages have been created in the past years for automated variable selection. Most of them are used for variable selection in multiple regression. On the other hand, for variable selection in generalized linear models we do not have so many possibilities if we are using R.
One of the possible solutions for performing variable selection is the R function **step()**. This function is automatic method which uses a stepwise procedure based on AIC. Thus, all disadvantages of stepwise procedures hold also for step function. Using it we can perform forward, backward and stepwise elimination for all type of generalized linear models.
Second possibility is to use **glmpath** package. This package contains a path-following algorithm for L_1 regularized generalized linear models, proposed by Park and Hastie [17]. In GLM response variable Y is modelled by using a linear combination of explanatory variables, $x^{'}\beta$, where x represents explanatory variables and β coefficients to be estimated. Using likelihood function L, we estimate coefficient β using

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \ln L(y;\beta).$$

GLM path algorithm uses this criterion but modified by adding a penalty term

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}}\{-\ln L(y;\beta) + \lambda \parallel \beta \parallel_1\},$$

where $\lambda > 0$ is the regularization parameter. In comparison with forward stepwise selection, GLM path algorithm is less greedy.

3.2 Cox Model

Survival analysis is used to analyse the data where the response variable is the time until the occurrence of an event of interest. The event is often referred to as a failure time, survival time or event time. First logical or intuitive question to ask ourselves would be whether we can use linear regression to model the survival time as a function of a set of explanatory variables. Given that survival times are typically positive numbers, ordinary linear regression is not the best choice, except these times are not transformed such that this restriction is removed. But more importantly, ordinary linear regression cannot effectively handle the censoring of observations.

Censoring is present when the information about the survival time is incomplete and represents a particular type of missing data. This lack of information occurs when a variable can be measured accurately only within a specific range. Outside of that domain, the only information available is that it is greater or smaller than some given value or that it lies between two values. As a consequence of this we have three main types of censoring: right, left and interval.
Two main notions for describing the distribution of event times in survival analysis are the survival and hazard function.
The survival function is defined as

$$S(t) = P(T > t),$$

where t represents some time and T is a random variable that represents survival time. Thus, it is the probability of surviving or not experiencing the event up to the time t. The hazard function is defined as the event rate at time t conditional on survival until time t or later (that is, $T \geq t$). Usually it is denoted as $\lambda(t)$ and it is equal to:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}.$$

The hazard function is not a density or a probability, but we can interpret it as a measure of risk, the greater the hazard between the two times - the greater the risk of failure in this time interval. The hazard function gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time.
It is generally of interest to describe the relationship of a set of explanatory variables with the survival time. A number of methods are available for this and they include parametric, nonparametric and semiparametric approaches. The most frequently used regression model for survival analysis is Cox's proportional hazards model.
The Cox regression model is a semiparametric model. It assumes the nonlinear relationship between the hazard function and the set of explanatory variables. The assumption that explanatory variables do not vary over time is called the proportional hazards assumption and in that case the hazard ratio comparing any two observations is in fact constant over time.
The Cox Proportional Hazard Model is most often stated in terms of the hazard function:

$$\lambda(t \mid X) = \lambda(t)\exp(X\beta).$$

It is a semiparametric model, meaning that it makes a parametric assumption concerning the effect of the explanatory variables on the hazard function, but without assumption regarding the nature of the hazard function. Thus, we do not assume any specific shape for $\lambda(t)$, the baseline hazard function. Regardless of this, for estimating

and testing regression coefficients, the Cox model is as efficient as parametric models. Note that $\lambda(t \mid X = 0) = \lambda(t)$. The only requirement for the baseline hazard function it to be greater then 0, apart from that it is left completely unspecified. Thus, we can not use ordinary likelihood methods to estimate unknown coefficient β. Cox (1972) described in his paper the method for estimation using conditional likelihood and later in 1975 he modified it and called it partial likelihood. Intuitively described, partial likelihood estimation procedure for Cox model is the following.

The first assumption is that there are no tied failure times, that is no two subjects have the same event time. Further, denote with R_i the risk set. R_i is the set of individuals at risk of failing an instant before failure time t_i, which means that their failure or censoring time is $Y_j \geq t_i$. Conditional probability that individual i will fail at moment t_i given a risk set and that exactly one failure will occur at t_i, is

$$\frac{\lambda(t_i)\exp(X_i\beta)}{\sum_{j\in R_i}\lambda(t_i)\exp(X_j\beta)} = \frac{\exp(X_i\beta)}{\sum_{j\in R_i}\exp(X_j\beta)}.$$

For independent subject events - failures, the joint probability is partial likelihood:

$$L(\beta) = \prod_i \frac{\exp(X_i\beta)}{\sum_{j\in R_i}\exp(X_j\beta)}.$$

The log partial likelihood is

$$logL(\beta) = \sum_i \left\{ X_i\beta - log\Big[\sum_{j\in R_i}\exp(X_j\beta)\Big] \right\}.$$

Cox and others have shown that using this partial log likelihood valid maximum likelihood estimates can be derived. This function can be maximized over β using the Newton-Raphson algorithm to produce maximum partial likelihood estimates of the model parameters. Without assumption of non tied failure times calculating the true partial log likelihood can be time consuming. Some of the approaches for that situation are Breslow's method and Efron's approach [8], [1]. The Cox PH regression model is fit in R with the *coxph* function available through survival package [18].

4 Augmented Backward Elimination

Augmented backward elimination is a method for variable selection proposed by Dunkler et al [6]. It is a combination of significance and change-in-estimate criterion. Since we will follow their article the remainder of this chapter will cover brief discuss about the change-in-estimate criterion and selection by significance. We will present approximated change-in-estimate criterion proposed by Dunkler et al. After that we will present the ABE method.

4.1 The Change-In-Estimate Criterion

In statistics, a confounding variable or confounder is defined as a variable that explains some or all of the correlation between the response variable and an explanatory variable. A confounder is correlated to the response variable and to the explanatory variable at the same time. There are many strategies to identify confounders, for example, forward, backward, and stepwise variable selection. Among these strategies, simulation studies have shown that the best is the change-in-estimate criterion [14]. Using change-in-estimate we can easily identify confounders in the following way: we just have to check whether the removal of the covariate has produced an important change in the coefficients of the variables remaining in the model. That is, we have to refit the model and to check for relative or absolute changes in the coefficients. When using the change-in-estimate criterion, a cutoff of 10% is commonly used to identify confounders [2]. Lee in his paper from 2014 emphasizes that there are very few studies of the statistical properties of the change-in-estimate criterion [14]. Thus, the suitability of the cutoff of 10% should be also verified under different conditions. Using simulations he showed the following. In linear regression, larger effect size, larger sample size, and lower standard deviation of the error term led to a lower cutoff point at a 5% significance level. In contrast, larger effect size and a lower exposure-confounder correlation led to a lower cutoff point at 80% power. In logistic regression, a lower odds ratio and larger sample size led to a lower cutoff point at a 5% significance level, while a lower odds ratio, larger sample size, and lower exposure-confounder correlation yielded a lower cutoff point at

80% power [14].

Instead of refitting the model, Dunkler et al. suggest to use approximated change-in-estimate. They denote by δ_p^{-a} the change in estimate, that is, the change in a regression coefficient β_p if we remove a variable X_a from a model. With index a *active* variable is denoted and with index p *passive* variable ($p \neq a$). Approximated change-in-estimate is defined as

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

This approximation is motivated by considering estimates $\hat{\beta}_a$ and $\hat{\beta}_p$ as random variables with variances $\hat{\sigma}_a^2$ and $\hat{\sigma}_p^2$ and covariance $\hat{\sigma}_{pa}$.
It can be shown that using approximated change-in-estimate and significance-based threshold for it, is equivalent to using a significance-based selection of variables, as if the change-in-estimate criterion is not considered at all. The variance of $\hat{\delta}_p^{-a}$ is given by

$$Var(\hat{\delta}_p^{-a}) = \left(\frac{\hat{\sigma}_{pa}}{\hat{\sigma}_a^2}\right)^2 Var(\hat{\beta}_a) = \frac{\hat{\sigma}_{pa}^2}{\hat{\sigma}_a^2}.$$

Z-statistic for testing $\hat{\delta}_p^{-a} = 0$ is given by

$$z = \frac{\hat{\delta}_p^{-a}}{SE(\hat{\delta}_p^{-a})} = -\frac{\hat{\beta}_a}{\hat{\sigma}_a}.$$

We see that this z-statistics is equivalent to z-statistic for testing $\beta_a = 0$.
Instead of using a significance-based threshold for the change-in-estimate, usually some pre-specified minimum value of $\hat{\delta}_p^{-a}$ or $\hat{\delta}_p^{-a}/\hat{\beta}_p$ is used as a threshold for leaving X_a in a model. Given that this formula is not appropriate in cases where $\hat{\beta}_p$ is close to zero, Dunkler et al. propose the following criterion

$$\frac{|\hat{\delta}_p^{-a}| SD(X_p)}{SD(Y)},$$

where $SD(X_p)$ and $SD(Y)$ are standard deviations of the passive explanatory variable X_p and the response variable Y, respectively. This criterion is used for linear regression. For some threshold value τ, active variable X_a is left in a model if

$$\frac{|\hat{\delta}_p^{-a}| SD(X_p)}{SD(Y)} \geq \tau.$$

For logistic and Cox regression they propose the following standardized criterion

$$\exp\left[|\hat{\delta}_p^{-a}| SD(X_p)\right] \geq 1 + \tau.$$

Using this, we can achieve the scale-independence and at the same time we avoid problematic situations where $\hat{\beta}_p$ is close to 0.

One of the advantages of this approximation is that evaluation of the change-in-estimate is notably faster, since instead of refitting the model after some variable is removed, the only thing we have to do is to calculate the approximation using given formula. Besides that, we can easily check for the significance of the change-in-estimate. Perhaps, the most important reason for including this change-in-estimate in variable selection is to avoid selecting variables just based on its significance.

4.2 Variable selection based on significance

Numerous variable selection methods are partially or fully based on significance. Since we have already presented some of the most commonly used techniques for variable selection and problems related to them, we will just emphasize which among those are fully relying on significance.

Primarily, all stepwise selection methods are included in this group. The most frequently used among them is backward selection. Another method based on significance test that is very criticized by many statisticians is univariate screening. Even though, there are some methods for variable selection, suggesting that the first step in selection process should be univariate screening [13]. Using just significance as a criterion for including or excluding a variable from a model is not the best choice. Considering change-in-estimate in addition to significance is exactly the advantage of augmented backward elimination, since in some way that is a safety belt for insignificant variables. It is very common to use p-values from Wald test statistic,

$$\frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)},$$

to select variables in the following way: variables with larger test statistic (hence with smaller p-values) are included in a model. This brings up the next question: what is that makes the test statistic larger or smaller?

The Wald test statistic $\frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}$ can be written as

$$\frac{\hat{\beta}_i}{\frac{\hat{\sigma}}{\sqrt{n\hat{var}(X_i)}}\sqrt{VIF_i}},$$

where VIF_i is the variance inflation factor for $\hat{\beta}_i$. From this we can see that we can make any variable significant or insignificant by making its test statistic enough large or small, which can be done by reducing or increasing certain parameters from the expression while the others are not changing. For example, for larger coefficients, test statistic is larger, thus variables will be more significant; by increasing the sample size

every variable will become more significant; by increasing the variance in an explanatory variable we increase also the test statistic so that variable will become more significant, etc.
Despite that methods based only on significance have many disadvantages it seems that they are still used quite a lot.

4.3 Purposeful Selection Algorithm

The method proposed by Dunkler et al. is based on "purposeful selection algorithm" proposed by Hosmer and Lemeshow. In their book *Applied Survival Analysis*, they say the following:

"We feel that one should approach multivariable model building with patience and a keen eye for the details that differentiate a good model from one that is merely adequate for the job" [13].

Their method for variable selection consists of the following seven steps:

1) We start with univariable analysis in order to get all variables that are significant at 20-25 percent level. Those variables as well as the others that are not significant but are of clinical importance represent subset of variables that will be included in the model.

2) After this we use the p-values from the Wald test to check if some variables can be deleted from the initial model. At the same time we should be careful and using p-values of the partial likelihood ratio test we should confirm that the deleted variable is not significant.

3) The next step you need to perform is checking whether removal of some variable has caused an important change in the coefficients of variables remaining in the model.

4) At this step, we add to the model, one at the time, all variables excluded from the initial multivariable model to confirm that they are neither statistically significant nor an important confounder.

5) Authors presented this method first of all for proportional hazards regression, but as they emphasized the methods available for selecting the best subset of variables for PH regression are essentially the same as those used in any other regression model.
So, at this point we check whether the data support the hypothesis that the effect of the covariate is linear in the log hazard and, if not, what transformation of the covariate is linear in the log hazard.

6) The final step is used for checking if interactions are needed in the model.

7) Evaluation of model obtained at step 6.

4.4 Augmented Backward Elimination Procedure

Augmented backward elimination (ABE) proposed by Dunkler et al. is a combination of backward elimination procedure and change-in-estimate criterion. The ABE method is implemented in a SAS macro and it is available for linear, logistic and Cox proportional hazards regression. More about this algorithm can be found in a Technical Report [7]. Brief outline of the augmented backward elimination algorithm is presented in the following figure. Authors of the ABE method take into consideration

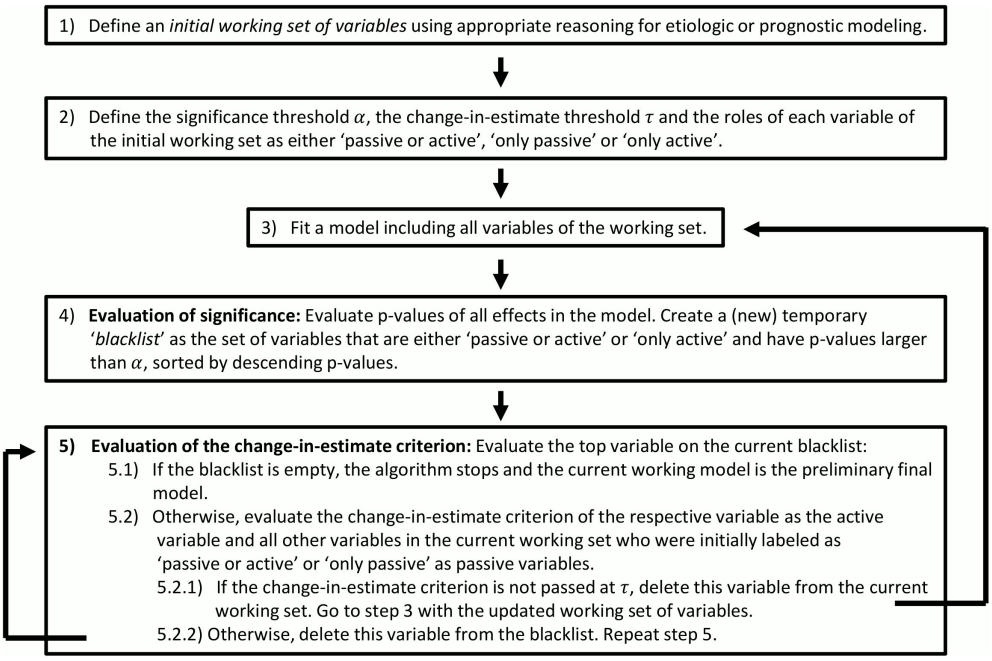

Figure 2: ABE method [6]

that different explanatory variables play different roles in the model, emphasizing that that could have important influence on variable selection process . Thus, based on subject-matter knowledge, one should identify the roles of explanatory variables before variable selection. For purposes of augmented backward elimination they have defined three types of variables. Namely, only passive, only active and passive or active.

> Only passive explanatory variables represent type of explanatory variables we want to have in the model regardless of whether they are statistically significant or not, for example. One of the reasons for keeping it in the model, is that maybe it is a variable of interest, or just based on subject-matter knowledge.

Only active explanatory variables represent less important explanatory variables which will be included in the model only if their exclusion causes changes in the estimates of more important explanatory variables. Such variables are not used for evaluating the change-in-estimate.

Passive or active explanatory variables represent type of variables considered as passive as well as active variables when evaluating change-in-estimate.

Since our aim is to write a R package for the ABE method, we will use the same definitions for three types of variables mentioned above. So, highly recommended and somehow necessary is to use *a priori* information in order to define initial working set of variables in the best possible way.

5 Generalization of ABE method

In this chapter we will present possible generalization of Augmented Backward Elimination method. Namely, we will try to use AIC and BIC as a criterion for choosing variables for the so called black list. Later, using simulations, we will check if and how good the generalization is. The idea to use AIC or BIC as a criterion arises from the results that indicate that it is better to use some information criterion for model selection in general than to use p-values. First we will present well-known AIC and BIC criteria. After that we will focus on why we should not use p-values for variable selection and why information-theoretic methods should be used.

5.1 Akaike Information Criterion

At any moment we should be aware that statistical models only approximate reality and at the same time we should try to find which model would be the best approximation given the data we have. Kullback and Leibler (1951) developed a measure called the Kullback-Leibler information which estimates the information lost. They defined it as a measure of the non-symmetric difference between two probability distributions P and Q. One way of defining the KL information or distance is by using probability density functions of P and Q:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\, \ln \frac{p(x)}{q(x)}\, \mathrm{d}x,$$

where p and q denote the densities of P and Q. If we look at this definition as a "information", intuitively $D_{\mathrm{KL}}(P\|Q)$ represents the information lost when q is used to approximate p. In the sense of the "distance", $D_{\mathrm{KL}}(P\|Q)$ is the distance between a model and the truth.
Idea of using Kullback-Leibler distance for model selection was introduced by Akaike in 1974.
For a statistical model M of some observed data x the Akaike information criterion is defined as:

$$\mathrm{AIC} = 2k - 2\ln(\hat{l}) \tag{5.1}$$

where

1. $\hat{l}$ is the maximized value of the likelihood function of the model, i.e. $\hat{l} = p(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function;

2. k is the number of free parameters to be estimated.

Akaike derived a relationship between the maximum likelihood and the Kullback-Leibler information. Soon we will see that relationship between AIC and KL information has very natural and simple concept. Main idea is to select a fitted approximating model that is estimated, on average, to be closest to the unknown data-generating model.
$D_{\mathrm{KL}}(P\|Q)$ can be written equivalently as

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\ln(p(x))\mathrm{d}x - \int_{-\infty}^{\infty} p(x)\ln(q(x))\mathrm{d}x.$$

We see that both expressions on the right side of the above expression are expectation with respect to p. Hence,

$$D_{\mathrm{KL}}(P\|Q) = \mathrm{E_p}\left[\ln(p(x))\right] - \mathrm{E_p}\left[\ln(q(x))\right].$$

The aim of Akaike information criterion is to estimate the Kullback-Leibler distance between an approximating model and the data-generating model in order to choose a model with the smallest estimated KL distance.
The first expectation is a constant, thus,

$$D_{\mathrm{KL}}(P\|Q) = Constant - \mathrm{E_p}\left[\ln(q(x))\right],$$

or

$$D_{\mathrm{KL}}(P\|Q) - Constant = -\mathrm{E_p}\left[\ln(q(x))\right].$$

So minimizing the KL distance is equivalent to maximizing the expression

$$D = \mathrm{E_p}\left[\ln(q(x))\right].$$

Hence, this expression is the quantity of interest but the problem is that it can not be estimated.
The first idea that comes to mind would be to use the following expression as an estimate

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n}\ln(q(X_i;\hat{\theta})) = \frac{\ln(\hat{l})}{n}$$

where $\ln(\hat{l})$ represents the log-likelihood function for the approximating model. Obviously this is biased since we are using the same data to obtain the maximum likelihood estimate and for the estimate of the expression of interest.

Akaike showed that an asymptotical bias is equal to $\frac{k}{n}$,where k is the number of parameters to estimate in the model. Therefore, the unbiased estimator of the $\mathrm{E_p}\left[\ln(q(x))\right]$ is

$$\frac{\ln(\hat{l})}{n} - \frac{k}{n}.$$

Finally, AIC is defined as in (5.1).

We see that there is no directed distance between two models. Instead of that we have unbiased estimator of the relative, expected Kullback-Leibler distance between the fitted model and the unknown model which generated the observed data.

Obviously, one should select the model with the smallest value of AIC since that model is the closest model to the unknown data-generating model from among the candidate models considered.

AIC has strong theoretical support since it is based on maximum likelihood and Kullback-Leibler information but at the same time supporting the idea that there is no true model. Furthermore it is easy to calculate and interpret. One of its greatest advantages is its role in variable selection.

Even though AIC attempts to choose the best approximating model of those in the set of models, everything depends on the given set, since of course it can happen that none of the models in the set are good. We will never know if a better model exists unless it is specified in the candidate set.

5.2 Bayesian Information Criterion

Closely related to the Akaike information criterion is another information criterion called Bayesian information criterion which is also used for model selection among the finite set of models. Also, it is partially based on log-likelihood but as an advantage over the AIC, BIC takes into the consideration the possibility of overfitting. This problem is caused by adding parameters in order to increase the likelihood. It is resolved by adding a penalty term for the number of parameters in the model. Formally BIC is defined as

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{l}),$$

where

1. $\hat{l}$ is the maximized value of the likelihood function of the model M, i.e. $\hat{l} = p(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function;
2. n is the number of observations;
3. k is the number of free parameters to be estimated.

The BIC was introduce by Schwarz as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model under the assumptions that the data distribution is in an exponential family. For large sample sizes, the model selected by BIC corresponds to the candidate model which is a posteriori most probable.

Let $P(M_i)$ denote a discrete prior over the models $M_1, \ldots, M_l$. Applying the Bayes theorem to calculate the posterior probability of a model given the data, we get

$$P(M_i|y_1, \ldots, y_n) = \frac{P(y_1, \ldots, y_n|M_i)P(M_i)}{P(y_1, \ldots, y_n)}.$$

Since our aim is to choose the model which is a posteriori most probable, we see that maximizing the posterior probability from the above expression is equivalent to maximizing the marginal likelihood $P(y_1, \ldots, y_n|M_i)$, because one of the assumptions is that all candidate models are equally likely. The marginal likelihoods are evaluated as

$$P(y_1, \ldots, y_n|M_i) = \int l(\theta_i|y_1, \ldots, y_n)g_i(\theta_i)\, d\theta_i,$$

where θ_i is the vector of parameters for the model M_i, l is the likelihood function and g_i is the probability density function of parameters θ_i.

Given that this is very difficult to obtain, the next step in the derivation of BIC is the approximation of this integral using Laplace method. This method uses a second order Taylor series expansion of $\ln P(Y|M_i)$ around $\tilde{\theta}_i$, the posterior mode. For large n, by keeping the terms involving n and ignoring the rest, it can be shown that

$$-2 \cdot \ln P(Y|M_i) \approx \text{BIC} = -2 \cdot \ln \hat{l} + k \cdot \ln(n).$$

We see that penalty term is larger in BIC than in AIC. Just as AIC, also BIC is widely used in model selection, first of all because of its computational simplicity. Beside that, BIC has a strong theoretical property, consistency. A consistent criterion will asymptotically select, with probability one, the candidate model having the correct structure.

In general BIC tends to choose models that are more parsimonious than those chosen by AIC. From the definitions of both AIC and BIC, we can see that scores to be minimized are quite similar, but it is very important to emphasize that AIC and BIC are not trying to answer the same questions. From informal derivation of AIC, we have seen that AIC is aimed at finding the best approximating model to the unknown data generating process (via minimizing expected estimated K-L divergence). As such, it fails to converge in probability to the true model (assuming one is present in the group evaluated), whereas BIC does converge as n tends to infinity.

5.3 To Use p-value Or Not?

"We were surprised to see a paper defending P-values and significance testing at this time in history." - This was the first sentence in paper from 2014 written by Burnham and Anderson as an answer or reaction on the paper from the same year written by Murtaugh [16], [3]. In his paper *"In defence of p-values"* Murtaugh argued that since confidence intervals, information-theoretic criteria and p-values are tools that are based on the same statistical information, the choice of which summary to present should be largely stylistic, depending on details of the application at hand. Therefore, all three tools have their places in sound statistical practice, and none of them should be excluded based on dogmatic, a priori considerations. He makes comparison between p-value and confidence interval, and explains the relationship between p-value and Akaike's information criterion.

He considered two nested linear models, i.e., two models such that one (the "reduced" model) is a special case of the other (the "full" model), with n observations and p parameters for the case of full model, while reduced model is obtained by setting the first k parameters equal to zero. Using basic definitions of p-value and AIC, he emphasizes the relationship between them as follows:

$$P = Pr(\chi^2_k > \Delta AIC + 2k),$$

where χ^2_k is a chi-square random variable with k degrees of freedom; ΔAIC represents the difference between AIC of reduced and AIC of full model, that is $AIC_R - AIC_F$ and

$$\Delta AIC = F^{-1}_{\chi^2_k}(1-p) - 2k,$$

where the $F^{-1}_{\chi^2_k}(1-p)$ is $(1-p)$ quantile of the χ^2_k distribution.

In the special case of nested linear models with Gaussian errors, it can be shown that using exact F distribution the relationship between p-value and ΔAIC is as follows:

$$P = Pr\left\{F_{k,n-p+1} > \frac{n-p+1}{k}\left[\exp\left(\frac{\Delta AIC + 2k}{n}\right) - 1\right]\right\}$$

and

$$\Delta AIC = n\log\left[\frac{k}{n-p+1}F^{-1}_{F_{k,n-p+1}}(1-P) + 1\right] - 2k$$

where $F^{-1}_{F_{k,n-p+1}}(1-P)$ is the $(1-P)$ quantile of an F distribution with k and $n-p+1$ degrees of freedom. For large n, these relationships are approximately equivalent to those based on the likelihood ratio statistic, that is based on χ^2_k distribution.

The relationship between ΔAIC and the p-value in a comparison of two models differing with respect to one parameter for different total sample sizes (n) is shown graphically in the following figure. The lines for finite n are based on the least-squares case and the

line for $n = \infty$ is based on the asymptotic distribution of the likelihood ratio statistic. From this Murtaugh concluded the following: deciding how small a p-value is needed

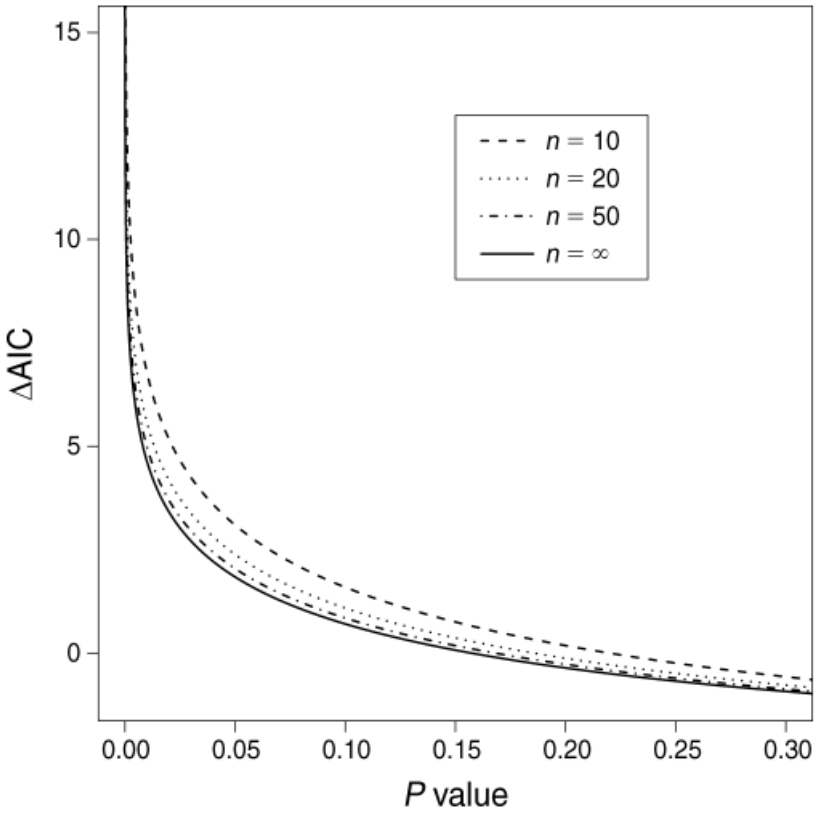

Figure 3: ΔAIC and p-value [16]

for us to prefer the more complicated model is equivalent to deciding how large a ratio of likelihoods indicates a convincing difference between models. At the same time he emphasizes that an important advantage of the information-theoretic criteria over the p-value is their ability to rank two or more models that are not nested in this way.

On the other side Burnham and Anderson believe that the subject of p-values and null hypothesis significance tests is an old one given that criticisms by statisticians began in the late 1930s and have been relentless. We will briefly repeat some of their arguments. One of the fundamental differences between p-values and information-theory methods is that after data have been collected the interest is focused on post-data probabilities, likelihood ratios, odds ratios, and likelihood intervals, while in case of p-values all theory is based on pre-data probability statements. That is, the anticipated data are being thought of as random variables, while for information criteria the theory is based on exact achieved data. Therefore, the conditioning is on the data, not the null hypothesis, and the objective is inference about unknowns parameters and models.

As limitation of p-values they mentioned the fact that by using p-values we are not able to use some very useful approaches in empirical science. Namely, we can not deal with non-nested models, we can not reduce model selection bias in high dimensional problems, difficulties with assessing the relative importance of predictor variables, problems with dealing with large systems and data sets (when number of parameters is greater than number of observations), difficulties with analyzing data from observational studies (where the distribution of the test statistic is unknown).

Furthermore, they emphasize that it is crucial to understand that using p-value means that we do not test the alternative hypothesis. Hence, if it is never tested it can not be rejected or falsified. Even greater difficulties arises when we have more than one alternative hypothesis which is the case in real-world problems.

Historical statistical approaches i.e., p-values try to find the "best" model and to make inferences from it. New methods that are based on post-data probability statements such as model probabilities or odds ratios and likelihood intervals are trying to base inference on all the models weighted by their model probabilities (model averaging).

About the Murtaugh's argument that p-values and AIC differences are closely related, they emphasize that the relationship holds only for the simplest case. That is, for comparison of two nested models differing by only one parameter. Therefore, that does not hold in general.

To draw attention to importance of the fact that new methods have taken the place of earlier ones, we will quote Burnham and Anderson.

"Statistical science has seen huge advances in the past 50-80 years, but the historical methods (e.g., t tests, ANOVA, step-wise regression, and chi-squared tests) are still being taught in applied statistics courses around the world... Students leave such classes thinking that "statistics" is no more than null hypotheses and p-values and the arbitrary ruling of statistical significance." [3]

Burnham and Anderson have already pointed out the differences between hypothesis testing and AIC in their book from 2002 [4]. They have considered set of nested models and they showed that the results can be quite different. Namely, they considered the null model with i parameters and set of alternative models with $i+j$ parameters where $j \geq 1$. They assumed that AIC value for each of the models is the same. That is, no model in the set has more support than any other model. Furthermore, they assumed that the null hypothesis is a model M_i and that it represents an adequate model for the data. Thus, M_i as a null model is tested individually against the $j \geq 1$ alternative

models from the set. So, the likelihood ratio test is used to compare the null model with any of the alternative models M_{i+j}.
Using

$$\begin{aligned} AIC_i &= -2\ln(l_i) + 2i \\ AIC_{i+j} &= -2\ln(l_{i+j}) + 2(i+j) \\ LRT &= -2(\ln(l_i) - \ln(l_{i+j})) \end{aligned}$$

we have

$$LRT = AIC_i - AIC_{i+j} + 2j.$$

In order to illustrate the difference between AIC and hypothesis testing assume that AIC value for each of the models is exactly the same. Then we have

$$LRT = 2j \text{ with j degrees of freedom.}$$

For example, if we have the following situation, M_i and M_{i+1}, that corresponds to a χ^2 value of 2 with degrees of freedom 1 and a p-value of 0.157.
For significance level $\alpha = 0.05$ the hypothesis-testing methods support the null model M_i over any of the alternative models $M_{i+1}, M_{i+2}, \ldots$ if degrees of freedom is less than about 7. This result is in contrast with AIC selection, where in this example all the models are supported equally.

Table 2: Summary of p-values

j	χ^2	p
1	2	0.157
2	4	0.135
3	6	0.112
4	8	0.092
5	10	0.075
6	12	0.062
7	14	0.051
8	16	0.042
9	18	0.035
10	20	0.029
15	30	0.012
20	40	0.005
25	50	0.005
30	60	0.001

From the above table we can see that in case there are more than 8 additional parameters, the null model M_i is rejected. For example, the likelihood ratio test of M_i versus M_{i+10} has 10 degrees of freedom, $\chi^2 = 20$ and p-value of 0.029. The conclusion is that the testing method indicates increasingly strong support of the models with many parameters and strong rejection of the simple null model.

6 Likelihood Ratio, Wald and Rao Score Tests

At the very beginning we have emphasized that augmented backward elimination is implemented in SAS and that one of our aims is to create R package for it. During testing of our R function in few examples we noticed slight differences in final models obtained with R and SAS functions. It turned out that differences were the consequences of using different tests for significance of our variables given that we had finite sample sizes. Because of that we want to stress that even though likelihood ratio, Wald, and Rao score tests used for significance testing converge to the same limiting Chi-square distribution they can give different results for small sample sizes. We will show that we can avoid conflicts that arise because of the small samples.

All three tests can be used to test the true value of the parameter based on the sample estimate. It is well-known that they are asymptotically equivalent but they differ in small samples. It can be shown that the score test statistic is always less than LR test statistic which implies it is less than Wald statistic. This means that if we use the same critical value for all three test the Wald test will reject the null hypothesis most often, while on the other side the score test will reject it least often. Even though the usage of the same critical value for all three tests is somehow indicated by the fact that these tests converge to the same limiting Chi-square distribution, this can cause different conclusions. Because of that in this chapter we will pay attention on how these tests are related in case of finite sample size.

Before we proceed, let us recall what is the convergence in probability and convergence in distribution.

Definition 6.1. A sequence X_n of random variables converges in probability towards the random variable X if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \varepsilon\right) = 0.$$

Usually, convergence in probability is denoted by adding the letter p over an arrow indicating convergence.

Definition 6.2. A sequence $X_1, X_2, \ldots$ of real-valued random variables is said to converge in distribution, or converge weakly, or converge in law to a random variable X if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every number $x \in R$ at which F is continuous. Here F_n and F are the cumulative distribution functions of random variables X_n and X, respectively.

Convergence in distribution is denoted by adding the letter d over an arrow indicating convergence.

The Wald test was introduced by Wald (1943). Assume that we want to test the hypothesis

$$H_0 : G(\theta_0) = 0,$$

where G is a function from R^p to R^r and the rank of $\frac{\partial G}{\partial \theta}$ is r.

Proposition 6.3. *For Wald test statistic W under the null hypothesis H_0, holds the following:*

$$W = nG'(\hat{\theta}) \left(\frac{\partial G(\hat{\theta})}{\partial \theta'} I^{-1}(\hat{\theta}) \frac{\partial G'(\hat{\theta})}{\partial \theta} \right)^{-1} G(\hat{\theta}) \sim \chi^2(r).$$

Proof. First order Taylor expansion of $G(\hat{\theta})$ around the true value θ_0 gives:

$$G(\hat{\theta}) = G(\theta_0) + \frac{\partial G(\theta_0)}{\partial \theta'}(\hat{\theta} - \theta_0)$$

ignoring other terms. By transforming the last expression, we get:

$$\sqrt{n}\left(G(\hat{\theta}) - G(\theta_0)\right) = \frac{\partial G(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta} - \theta_0). \tag{6.1}$$

As a consequence of asymptotic property of the MLE we know that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0)),$$

and using 6.1 we have that

$$\sqrt{n}\left(G(\hat{\theta}) - G(\theta_0)\right) \xrightarrow{d} N\left(0, \frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta}\right).$$

Under the null hypothesis $G(\theta_0)$ is equal 0. Hence,

$$\sqrt{n}G(\hat{\theta}) \xrightarrow{d} N\left(0, \frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta}\right). \tag{6.2}$$

From characteristics of normal random variables, we obtain

$$nG'(\hat{\theta})\left(\frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta}\right)^{-1} G(\hat{\theta}) \sim \chi^2(r),$$

under the null hypothesis. But this expression contains unknown parameter θ_0. Therefore, it is useless. Using consistent estimator, MLE $\hat{\theta}$, we can consistently approximate the expression

$$nG'(\hat{\theta})\left(\frac{\partial G(\hat{\theta})}{\partial \theta'} I^{-1}(\hat{\theta}) \frac{\partial G'(\hat{\theta})}{\partial \theta}\right)^{-1} G(\hat{\theta}) \sim \chi^2(r).$$

□

Before presenting score test and its asymptotic property let us recall of few notions. Let l be the likelihood function which depends on a univariate parameter θ and let x be the data. Denote with L the log-likelihood function. The score $U(\theta)$ is defined as

$$U(\theta) = \frac{\partial L(\theta \mid x)}{\partial \theta}.$$

The Fisher information is

$$I(\theta) = -\,\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} L(X;\theta)\middle|\,\theta\right].$$

Sometimes there are certain restrictions for parameter vector. In that case the MLE from constrained and unconstrained maximizations of course are not the same. We will denote MLE for unconstrained case with $\hat{\theta}$ and with $\tilde{\theta}$ for the solution of the MLE for constrained case. For constrained case we have that

$$\frac{\partial L(\tilde{\theta})}{\partial \theta} + \frac{\partial G'(\tilde{\theta})}{\partial \theta}\tilde{\lambda} = 0 \tag{6.3}$$

$$G(\tilde{\theta}) = 0 \tag{6.4}$$

where λ is the vector of Lagrange multiplier. The score test was introduced by Rao (1948). Its main advantage is that it does not require an estimate of the information under the alternative hypothesis or unconstrained maximum likelihood.

Proposition 6.4. *For score statistic S under the null hypothesis H_0, the following holds*

$$S = \frac{1}{n} U(\tilde{\theta}) I^{-1}(\tilde{\theta}) U(\tilde{\theta}) \sim \chi^2(r)$$

or

$$S = \frac{1}{n}\tilde{\lambda}' \frac{\partial G(\tilde{\theta})}{\partial \theta'} I^{-1}(\tilde{\theta}) \frac{\partial G'(\tilde{\theta})}{\partial \theta}\tilde{\lambda} \sim \chi^2(r).$$

Proof. Again, using Taylor expansion and ignoring the other terms, for both functions $G(\hat{\theta})$ and $G(\tilde{\theta})$ we have

$$\sqrt{n}G(\tilde{\theta}) = \sqrt{n}G(\theta_0) + \frac{\partial G(\theta_0)}{\partial \theta'}\sqrt{n}(\tilde{\theta} - \theta_0),$$

and

$$\sqrt{n}G(\hat{\theta}) = \sqrt{n}G(\theta_0) + \frac{\partial G(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta} - \theta_0).$$

By subtracting the two last expressions and given that $G(\tilde{\theta}) = 0$ we have

$$\sqrt{n}G(\hat{\theta}) = \frac{\partial G(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta} - \tilde{\theta}). \tag{6.5}$$

Using Taylor expansion for score functions $U(\hat{\theta})$ and $U(\tilde{\theta})$ around θ_0, we have

$$U(\hat{\theta}) = U(\theta_0) + \frac{\partial U(\theta_0)}{\partial \theta'}(\hat{\theta} - \theta_0)$$

$$\frac{1}{\sqrt{n}}U(\hat{\theta}) = \frac{1}{\sqrt{n}}U(\theta_0) + \frac{1}{n}\frac{\partial U(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta} - \theta_0).$$

Note that,

$$-\frac{1}{n}\frac{\partial U(\theta_0)}{\partial \theta'} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln l(x \mid \theta)}{\partial \theta \partial \theta'}.$$

By the weak law of large numbers this converges in probability to $I(\theta_0)$, so we have the following

$$\frac{1}{\sqrt{n}}U(\hat{\theta}) = \frac{1}{\sqrt{n}}U(\theta_0) - I(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0).$$

In the similar way we have

$$\frac{1}{\sqrt{n}}U(\tilde{\theta}) = \frac{1}{\sqrt{n}}U(\theta_0) - I(\theta_0)\sqrt{n}(\tilde{\theta} - \theta_0).$$

Subtracting the last two expressions and given that $U(\hat{\theta}) = 0$ we obtain

$$\frac{1}{\sqrt{n}}U(\tilde{\theta}) = I(\theta_0)\sqrt{n}(\hat{\theta} - \tilde{\theta}).$$

Now we can express $\sqrt{n}(\hat{\theta} - \tilde{\theta})$ as

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) = I^{-1}(\theta_0)\frac{1}{\sqrt{n}}U(\tilde{\theta}). \tag{6.6}$$

Therefore, from (6.5) and from (6.6) we have

$$\sqrt{n}G(\hat{\theta}) = \frac{\partial G(\theta_0)}{\partial \theta'}I^{-1}(\theta_0)\frac{1}{\sqrt{n}}U(\tilde{\theta}).$$

By (6.3) and using the fact that $\tilde{\theta} \xrightarrow{p} \theta_0$ we have

$$\sqrt{n}G(\hat{\theta}) \xrightarrow{p} -\frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta} \frac{\tilde{\lambda}}{\sqrt{n}}.$$

From (6.2) we know that $\sqrt{n}G(\hat{\theta}) \xrightarrow{d} N\left(0, \frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta}\right)$. Therefore,

$$\frac{\tilde{\lambda}}{\sqrt{n}} \xrightarrow{d} N\left(0, \left(\frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta}\right)^{-1}\right).$$

Again, using properties of normal random variables we have

$$\frac{1}{n}\tilde{\lambda}' \left(\frac{\partial G(\theta_0)}{\partial \theta'} I^{-1}(\theta_0) \frac{\partial G'(\theta_0)}{\partial \theta}\right) \tilde{\lambda} \sim \chi^2(r),$$

under null hypothesis. Using constrained MLE $\tilde{\theta}$ instead of true value θ_0 we can consistently approximate the last expression in order to get a usable statistic.
Similarly, from $U(\tilde{\theta}) + \frac{\partial G'(\tilde{\theta})}{\partial \theta}\tilde{\lambda} = 0$ it follows that

$$\frac{1}{n} U(\tilde{\theta}) I^{-1}(\tilde{\theta}) U(\tilde{\theta}) \sim \chi^2(r).$$

□

The likelihood ratio test was introduced by Neyman and Pearson (1928).

Proposition 6.5. *For likelihood ratio test statistic LR under the null hypothesis the following holds*

$$LR = 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) \sim \chi^2(r)$$

Proof. Second order Taylor expansions of functions $L(\hat{\theta})$ and $L(\tilde{\theta})$ around true value θ_0 give,

$$\begin{aligned} L(\hat{\theta}) &= L(\theta_0) + \frac{\partial L(\theta_0)}{\partial \theta'}(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)' \frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) \\ &= L(\theta_0) + \frac{1}{\sqrt{n}} \frac{\partial L(\theta_0)}{\partial \theta'} \sqrt{n}(\hat{\theta} - \theta_0) + \frac{1}{2}\sqrt{n}(\hat{\theta} - \theta_0)' \frac{1}{n} \frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'} \sqrt{n}(\hat{\theta} - \theta_0) \end{aligned}$$

and

$$\begin{aligned} L(\tilde{\theta}) &= L(\theta_0) + \frac{\partial L(\theta_0)}{\partial \theta'}(\tilde{\theta} - \theta_0) + \frac{1}{2}(\tilde{\theta} - \theta_0)' \frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'}(\tilde{\theta} - \theta_0) \\ &= L(\theta_0) + \frac{1}{\sqrt{n}} \frac{\partial L(\theta_0)}{\partial \theta'} \sqrt{n}(\tilde{\theta} - \theta_0) + \frac{1}{2}\sqrt{n}(\tilde{\theta} - \theta_0)' \frac{1}{n} \frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'} \sqrt{n}(\tilde{\theta} - \theta_0). \end{aligned}$$

Subtracting and multiplying by 2, we get

$$\begin{aligned} 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) = \frac{2}{\sqrt{n}} \frac{\partial L(\theta_0)}{\partial \theta'} \sqrt{n}(\hat{\theta} - \tilde{\theta}) + \sqrt{n}(\hat{\theta} - \theta_0)' \frac{1}{n} \frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'} \sqrt{n}(\tilde{\theta} - \theta_0) \\ - \sqrt{n}(\tilde{\theta} - \theta_0)' \frac{1}{n} \frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta'} \sqrt{n}(\tilde{\theta} - \theta_0). \end{aligned}$$

Given that $\frac{1}{\sqrt{n}}\frac{\partial L(\theta_0)}{\partial \theta'} = I(\theta_0)\sqrt{n}(\hat{\theta}-\theta_0)$ and since $-\frac{1}{n}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'} \xrightarrow{p} I(\theta_0)$, we have

$$\begin{aligned} 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) \to\ & 2n(\hat{\theta}-\theta_0)'I(\theta_0)(\hat{\theta}-\tilde{\theta}) - n(\hat{\theta}-\theta_0)'I(\theta_0)(\hat{\theta}-\theta_0) \\ & + n(\tilde{\theta}-\theta_0)'I(\theta_0)(\tilde{\theta}-\theta_0). \end{aligned}$$

Hence we can write

$$\begin{aligned} 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) =\ & 2n(\hat{\theta}-\theta_0)'I(\theta_0)(\hat{\theta}-\tilde{\theta}) - n(\hat{\theta}-\theta_0)'I(\theta_0)(\hat{\theta}-\theta_0) \\ & + n(\tilde{\theta}-\hat{\theta}+\hat{\theta}-\theta_0)'I(\theta_0)(\tilde{\theta}-\hat{\theta}+\hat{\theta}-\theta_0). \end{aligned}$$

By simple computation and using that $(\hat{\theta}-\theta_0)'I(\theta_0)(\hat{\theta}-\tilde{\theta}) = (\hat{\theta}-\tilde{\theta})'I(\theta_0)(\hat{\theta}-\theta_0)$, we have

$$2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) = n(\hat{\theta}-\tilde{\theta})'I(\theta_0)(\hat{\theta}-\tilde{\theta}). \tag{6.7}$$

From (6.6) and (6.7), we obtain

$$\begin{aligned} 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) &= \sqrt{n}(\hat{\theta}-\tilde{\theta})'I(\theta_0)\sqrt{n}(\hat{\theta}-\tilde{\theta}) \\ &= \frac{1}{\sqrt{n}}\frac{\partial L(\tilde{\theta})}{\partial\theta'}I^{-1}(\theta_0)I(\theta_0)I^{-1}(\theta_0)\frac{\partial L(\tilde{\theta})}{\partial\theta}\frac{1}{\sqrt{n}} \\ &= \frac{1}{n}\frac{\partial L(\tilde{\theta})}{\partial\theta'}I^{-1}(\theta_0)\frac{\partial L(\tilde{\theta})}{\partial\theta} \\ &= S \sim \chi^2(r). \end{aligned}$$

□

Asymptotically, all three test statistics are distributed according to the χ^2 distribution with r degrees of freedom.

6.1 Wald, Score and LR Test for Linear Regression

Let us look at these tests for linear regression model. Suppose that the model is of the form

$$Y = X\beta + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$. Suppose that hypothesis is of the form

$$H_0 : R\beta = b \text{ versus } H_1 : R\beta \neq b$$

where R is a known matrix of the rank r and b is a known vector. Note that this can be written as $G(\beta) = R\beta - b$. From the definitions of Wald, score and LR statistics it follows that in the case σ^2 is known, we have that

$$W = S = LR = \frac{(R\hat{\beta}-b)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-b)}{\sigma^2} \sim \chi^2(r),$$

where $\hat{\beta}$ is MLE [9].
Usually we do not know σ^2 so we use its estimate obtained by MLE, $\hat{\sigma}$. We know that F statistics can be written as

$$F = \frac{\frac{RSS_1 - RSS_2}{r}}{\frac{RSS_2}{n-k}}$$

where RSS_1 is residual sum of squares for restricted model, RSS_2 is residual sum of squares for unrestricted model, n is the number of observations, k is the number of parameters and r is the rank of the matrix R.
Using F statistics we can derive all three statistics for the case σ^2 is unknown.

$$RSS_1 = (Y - X\tilde{\beta})'(Y - X\tilde{\beta})$$
$$RSS_2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

Therefore,

$$\begin{aligned} RSS_1 - RSS_2 &= (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) - (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= (X\tilde{\beta} - X\hat{\beta})'(X\tilde{\beta} - X\hat{\beta}) \\ &= (\tilde{\beta} - \hat{\beta})'X'X(\tilde{\beta} - \hat{\beta}). \end{aligned}$$

Using that $\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - b)$, we have

$$RSS_1 - RSS_2 = (R\hat{\beta} - b)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - b).$$

From the definition of Wald statistics in case σ^2 is not known it follows that statistics is of the form

$$\begin{aligned} W &= n\left(R\hat{\beta} - b\right)'\left[[R \quad 0]I^{-1}(\hat{\theta})[R \quad 0]'\right]^{-1}\left(R\hat{\beta} - b\right) \\ &= \left(R\hat{\beta} - b\right)'\left[[R \quad 0]I_n^{-1}(\hat{\theta})[R \quad 0]'\right]^{-1}\left(R\hat{\beta} - b\right) \\ &= \frac{(R\hat{\beta} - b)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - b)}{\hat{\sigma}^2}. \end{aligned}$$

Here, we used that the inverse of Fisher information is

$$[I_n(\beta, \sigma^2)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Hence,

$$\begin{aligned} W &= \frac{(R\hat{\beta}-b)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-b)}{\frac{1}{n}(Y-X\hat{\beta})'(Y-X\hat{\beta})} \\ &= \frac{(R\hat{\beta}-b)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-b)/r}{(Y-X\hat{\beta})'(Y-X\hat{\beta})/(n-p)} \cdot \frac{nr}{n-p} \\ &= \frac{(RSS_1-RSS_2)/r}{RSS_2/(n-p)} \cdot \frac{nr}{n-p} \\ &= \frac{nr}{n-p}F \end{aligned}$$

Now we will derive the relationship between F statistic and score statistic in case σ^2 is not known. Score statistic is defined as

$$S = \frac{1}{n}\tilde{\lambda}'\frac{\partial G(\tilde{\theta})}{\partial \theta'}I^{-1}(\tilde{\theta})\frac{\partial G'(\tilde{\theta})}{\partial \theta}\tilde{\lambda}.$$

Again, the hypothesis is of the form

$$H_0 : R\beta = b \text{ versus } H_1 : R\beta \neq b$$

Therefore, $G(\beta) = R\beta - b$. In order to write our score statistic we need Lagrange multiplier for restricted linear regression, $\tilde{\lambda}$ and the inverse of Fisher information. For linear model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I)$, the log-likelihood function L is

$$L(\beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta).$$

Calculating the negative inverse of the expectation of

$$\begin{bmatrix} \frac{\partial^2 L}{\partial \beta \partial \beta'} & \frac{\partial^2 L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 L}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 L}{\partial \sigma^2 \partial \sigma^2} \end{bmatrix}$$

we get that the inverse of Fisher information

$$[I_n(\beta, \sigma^2)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Denote with $\tilde{\beta}$ and $\tilde{\sigma}^2$ solutions of restricted MLE. Then, from

$$\begin{aligned} \frac{\partial L(\tilde{\theta})}{\partial \theta} + \frac{\partial G'(\tilde{\theta})}{\partial \theta}\tilde{\lambda} &= 0 \\ G(\tilde{\theta}) &= 0 \end{aligned}$$

we can derive Lagrange multiplier. Namely, $\tilde{\lambda} = \frac{1}{\tilde{\sigma}^2}(R(X'X)^{-1}R')^{-1}(b - R\hat{\beta})$.

Using that and that $I_n = nI$ we have the following

$$\begin{aligned} S &= \frac{1}{n}\tilde{\lambda}' n(R\tilde{\sigma}^2(X'X)^{-1}R')\tilde{\lambda} \\ &= \frac{1}{\tilde{\sigma}^2}(R\hat{\beta}-b)'(R(X'X)^{-1}R')^{-1}\tilde{\sigma}^2(R(X'X)^{-1}R')\frac{1}{\tilde{\sigma}^2}(R(X'X)^{-1}R')^{-1}(R\hat{\beta}-b) \\ &= \frac{1}{\tilde{\sigma}^2}(R\hat{\beta}-b)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta}-b). \end{aligned}$$

We have already showed that

$$RSS_1 - RSS_2 = (R\hat{\beta}-b)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-b).$$

Therefore, score statistic can be expressed as

$$\begin{aligned} S &= \frac{RSS_1 - RSS_2}{\tilde{\sigma}^2} \\ &= n\frac{RSS_1 - RSS_2}{RSS_1} \\ &= \frac{n}{1 - 1 + \frac{RSS_1}{RSS_1 - RSS_2}} \\ &= \frac{n}{1 + \frac{RSS_2}{RSS_1 - RSS_2}} \\ &= \frac{n}{1 + \frac{n-p}{rF}} \end{aligned}$$

To express likelihood ratio statistic using F statistic, note that

$$\begin{aligned} L(\hat{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\cdot\frac{n}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2\hat{\sigma}^2}\hat{\sigma}^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}. \end{aligned}$$

Similarly,

$$\begin{aligned} L(\tilde{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{n}{2\tilde{\sigma}^2}\tilde{\sigma}^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tilde{\sigma}^2) - \frac{n}{2}. \end{aligned}$$

Therefore,

$$\begin{aligned} LR &= 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) \\ &= 2\left(-\frac{n}{2}\log(\hat{\sigma}^2) + \frac{n}{2}\log(\tilde{\sigma}^2)\right) \\ &= n\log\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = n\log\left(1 - 1 + \frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right) \\ &= n\log\left(1 + \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right) \\ &= n\log\left(1 + \frac{rF}{n-p}\right) \end{aligned}$$

Using the inequality

$$\frac{t}{1+t} \leq \log(1+t) \leq t \qquad \forall t > -1$$

and defining $t = \dfrac{rF}{n-p}$ it follows that,

$$S \leq LR \leq W$$

We see that in case of linear regression when σ^2 is known it holds that all three test have the same and exact χ^2 distribution. If σ^2 is not known then all three statistics are functions of F statistic meaning that we can avoid conflicts already pointed out if we use critical values that are related to each other by the same functions.

7 R Package

In this chapter we will present options available in the R package ABE.

7.1 Description

In Figure 2 is presented the brief outline of the augmented backward elimination. The R package allows the user to use exactly the same function as it is explained in the figure. Additionally, enables us to use AIC or BIC criterion instead of significance and we can choose if approximated or exact change-in-estimate criterion will be evaluated. Briefly, procedure for the augmented backward elimination is as follows.

1) At the first step of augmented backward elimination we have to define an *initial working set* of variables using appropriate reasoning for etiologic or prognostic modeling.

2) The next thing we should do is to define the values for required parameters. That is, we should set a value for significance threshold α, define the change-in-estimate threshold τ. Finally, we should classify variables as "passive", "active" and "passive or active".

3) After that, a model including all variables from the working set should be fitted.

4) In augmented backward elimination procedure proposed by Dunkler et al. the following step is evaluation of significance. That is, we have to evaluate p-values of all effects in the model and to create a (new) temporary "blacklist" as the set of the variables that are either "passive or active" or "active" and have p-values larger than α, sorted by descending p-values.

   We extended this, such that the "blacklist" could be created using information criteria, AIC and BIC. Namely, we evaluate AIC (BIC) for the full model. We drop each variable from the set of "passive or active" and "active" variables separately and evaluate AIC (BIC) for each model. After that we create temporary "blacklist". Variable will be on the blacklist if the value of AIC (BIC) for the model without that variable is less than the AIC (BIC) for the full model.

5) Next step is the evaluation of the change-in-estimate criterion. We have to evaluate the top variable on the current blacklist.

   5.1) If the blacklist is empty, the algorithm stops and the current working model is the preliminary final model.

   5.2) Otherwise, evaluate the change-in-estimate criterion of the respective variable as the active variable and all other variables in the current working set who were initially labeled as "passive or active" or "passive" as passive variables.

      5.2.1) If the change-in-estimate criterion is not passed at τ, delete this variable from the current working set. Go to step 3 with the updated working set of variables.

      5.2.2) Otherwise, delete this variable from the blacklist. Repeat step 5.

   In augmented backward elimination procedure proposed by Dunkler et al. is used approximated change-in-estimate. We extended this, such that also exact change-in-estimated can be used.

7.2 Arguments

The following arguments allow the user to specify the initial model and the variable selection criteria when using augmented backward elimination procedure.
abe(fit = *object*,
include = *variables*,
active = *variables*,
tau = *value*,
exp.beta = *logical*,
exact = *logical*,
criterion = *string*,
alpha = *value*,
type.test = *string*,
verbose = *logical*)

A brief explanation of these options follows.

- *fit* - An object of a class "lm", "glm" or "cohph" representing the fit.
- *include* - Names a vector of passive variables. 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. Note, variables which are not specified as include or active are assumed to be active or passive variables. Default is *NULL*.

- *active* - Names a vector of active variables. These less important explanatory variables will only be used as active, but not as passive variables when evaluating the change-in-estimate criterion. Note, they will be included in the model only if their exclusion causes changes in the estimates of more important explanatory variables. Default is *NULL*. Note, if we leave *include* and *active* as it is by the default then ABE will consider all variables as "active or passive".

- *tau* - Value that specifies the threshold of the relative change-in estimate criterion. Default is set to 0.05.

- *exp.beta* - Logical specifying if exponent is used in formula to standardize the criterion. Default is set to *TRUE*.

- *exact* - Logical that specifies if you will use exact change-in-estimate or approximated. Default is set to *FALSE*, which means that you will use approximation proposed by Dunkler et al. Note, setting to *TRUE* can severely slow down the algorithm, but setting to *FALSE* can in some cases lead to a poor approximation of the change-in-estimate criterion.

- *criterion* - String that specifies the strategy to select variables for the blacklist. Currently supported options are significance level "alpha", Akaike information criterion "AIC" and Bayesian information criterion "BIC". If you are using significance level, in that case you have to specify the value of "alpha". Default is set to *"alpha"*.

- *alpha* - Value that specifies the level of significance as explained above. Default is set to 0.2.

- *type.test* - String that specifies which test should be performed in case the criterion = "alpha". Possible values are "F" and "Chisq" (default) for class "lm"; "Rao", "LRT", "Chisq" (default), "F" for class "glm" and "Chisq" for class "coxph".

- *verbose* - Logical that specifies should the variable selection process be printed. Note: this can severely slow down the algorithm.

The result is an object of class "lm", "glm" or "cohph" representing the model chosen by ABE method.

Using the default settings ABE will perform augmented backward elimination based on significance. The level of significance will be set to 0.2. All variables will be treated as "passive or active". Approximated change-in-estimate will be used. Threshold of the relative change-in-estimate criterion will be 0.05. Setting τ to a very large number turns off the change-in-estimate criterion, and ABE will only perform BE. Specifying "alpha" = 0 will include variables just based on change-in-estimate criterion, since in that case variables are not safe from exclusion because of their p-values. Specifying "alpha" = 1 will always include all variables.

7.3 Example

To carry out the ABE procedure in R the following code can be submitted:

abe(`fit`, `include`="x1", `active`=NULL, `tau`=0.05, `exp.beta`=FALSE,
`exact`=FALSE, `criterion`="alpha", `alpha`=0.2,
`type.test`="Chisq", `verbose`=TRUE)

In this particular example parameter `fit` represents an object of a class "lm" representing the fit. Data were simulated as it is explained in chapter 8. Our current goal is to show how the method is used in R and what is its output. We fitted the model using all variables of the working set.

```
fit = lm(y~x1+x2+x3+x4+x5+x6+x7,x=T, y=T)
```

We defined "x1" to be "only passive variable" meaning that `include`="x1". By leaving `active` to be NULL the other six variables are defined as "passive or active". In order to see the detailed output we set `verbose` to be TRUE. The output was as follows.

```
Model under investigation:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, x = T, y = T)

Criterion for non-pasive variables: x2 : 0.0341 , x3 : 0.2233 ,
x4 : 0.3968 , x5 : 0.5151 , x6 : 0.2503 , x7 : 0
black list:  x5 : 0.5151, x4 : 0.3968, x6 : 0.2503, x3 : 0.2233
Investigating change in b or exp(b) due to omitting variable  x5;
x1 : 0.0039, x2 : 0.0072, x3 : 0.0163, x4 : 0.019, x6 : 0.0023,
x7 : 0.003

Model under investigation:
lm(formula = y ~ x1 + x2 + x3 + x4 + x6 + x7, x = T, y = T)
```

```
Criterion for non-pasive variables: x2 : 0.0268 , x3 : 0.1501 ,
x4 : 0.2733 , x6 : 0.2381 , x7 : 0
black list:  x4 : 0.2733, x6 : 0.2381
Investigating change in b or exp(b) due to omitting variable  x4;
x1 : 0.0125, x2 : 0.0203, x3 : 0.0474, x6 : 0, x7 : 0.0046

Model under investigation:
lm(formula = y ~ x1 + x2 + x3 + x6 + x7, x = T, y = T)
Criterion for non-pasive variables: x2 : 0.0128 , x3 : 0.0274 ,
x6 : 0.2406 , x7 : 0
black list:  x6 : 0.2406
Investigating change in b or exp(b) due to omitting variable  x6;
x1 : 0.0209, x2 : 0.0115, x3 : 0.0226, x7 : 0.0044

Model under investigation:
lm(formula = y ~ x1 + x2 + x3 + x7, x = T, y = T)
Criterion for non-pasive variables: x2 : 0.0087 , x3 : 0.0112 , x7 : 0
black list: empty

Final model:
lm(formula = y ~ x1 + x2 + x3 + x7, x = T, y = T)

Call:
lm(formula = y ~ x1 + x2 + x3 + x7, x = T, y = T)

Coefficients:
(Intercept)           x1           x2           x3           x7
    0.07863      1.71615      1.07317      0.96667      1.32194
```

From this we can see that at every step we see exactly what is happening. Namely, at every step the current model is presented; in this particular example p-values of non-passive variables are presented; variables that are on the blacklist sorted by descending p-values and evaluated change-in-estimate for "passive" and "passive or active" variables if we would delete the top variable from the blacklist are presented. At the end the final model is shown. If we would set `verbose` to be FALSE only the final model would be presented.

8 Simulation Study

Simulation study for evaluation of ABE methods we performed was based on the simulation study already preformed by Dunkler et al [6].

Explanatory variables $X_1, \ldots, X_7$ were drawn from a normal distribution. Namely, X_2, X_3, X_4 and X_5 were drawn from a multivariate normal distribution with a mean vector of 0, standard deviation of 1 and bivariate correlation coefficients of 0.5. X_6 and X_7 were independently drawn from a standard normal distribution. Variable of main interest X_7 was defined such that it depended on X_2, X_3 and X_6. Depending on variance inflation factor X_1 was simulated as $X_1 = 0.266(X_2 + X_3 + X_6) + 0.710\epsilon$ for case $VIF = 2$ and $X_1 = 0.337(X_2 + X_3 + X_6) + 0.4499\epsilon$ for case $VIF = 4$, where ϵ was a random number from a standard normal distribution. Outcome variable was defined as $Y^* = \beta_1 X_1 + X_2 + X_4 + X_7$. Y was generated as continuous, binary and time-to-event variable in order to simulate linear, logistic and Cox regression, respectively.

Specifically, to simulate linear regression outcome variable Y was drawn from a normal distribution with mean Y^* and standard deviation 3.6. To simulate logistic regression Y was drawn from a Bernoulli distribution with event probability $1/(1 + \exp(-Y^*)$. In order to simulate Cox regression, Weibull distributed survival times T were drawn from $[-\log(U)/0.125 \exp(Y^*)]^{1/3}$, where U was a standard uniform random variable. From a uniform distribution $U[0, 3.35]$ were drawn times U, in order to obtain approximately 55% censoring. Observable survival time was defined as $Y = min(T, U)$ and status indicator was defined as $S = I(T < U)$.

We simulated 1000 samples with sample size 120 for each scenario, for β_1 either 0 or 1, for VIF being 2 or 4 and for each type of regression (linear, logistic or Cox). Each sample was analyzed applying different methods for variable selection and variable of main interest X_1 was forced into every model. We were interested in the number of biased, correct and inflated models, as well as in the bias and root mean squared error (RMSE) of β_1 compared to the correct model. By biased model we mean model for which at least one variable from the true model was not selected. Correct model included all variables from the true model, that is X_1, X_2, X_4 and X_7. By inflated model we mean larger model, that is a model which contains all correct variables and at least one incorrect as well. Namely, we checked results for cases where we used ABE method proposed by Dunkler et al, but also for slightly modified ABE method, that

is for ABE based on BIC and AIC; all of that considering exact and approximated change in estimate. We compared those results with results obtained after using the following methods:

- Univariate model including X_1.
- Correct model including X_1, X_2, X_4 and X_7.
- Full model including $X_1, X_2, X_3, X_4, X_5, X_6$ and X_7.
- Model selected with forward selection and AIC as a criterion.
- Model selected with backward elimination and AIC as a criterion.
- Model selected with stepwise elimination and AIC as a criterion.
- Model selected with backward elimination and BIC as a criterion.
- Model selected with backward elimination and significance level 0.05 as a criterion.
- Model selected with backward elimination and significance level 0.2 as a criterion.

We were also interested in the bias and root mean squared error (RMSE) of estimated coefficients. Namely, for every $\hat{\beta}_1$ 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 were given. The results of simulation study are contained in following sections.

8.1 Simulation Results for Linear Regression

Notations for the methods we used are as follows: ABE method proposed by Dunkler et al., is denoted as "ABE", ABE method that uses AIC as a criterion for the blacklist is "abeAIC", ABE method that uses BIC is "abeBIC"; for the same methods that were using exact change-in-estimate instead of approximated the notations we used are: "eABE", "eabeAIC" and "eabeBIC"; forward selection that uses AIC as a criterion is "FE"; backward elimination and AIC as a criterion is "BE"; stepwise elimination and AIC as a criterion is "SE"; backward elimination and BIC as a criterion "BE bic"; backward elimination and significance level 0.05 as a criterion is denotead as "BE0.05"; backward elimination and significance level 0.2 as a criterion is denotead as "BE0.2". Univariate model including X_1 is denoted as "Uni"; correct model including X_1, X_2, X_4 and X_7 is denoted as "Correct"; full model including $X_1, X_2, X_3, X_4, X_5, X_6$ and X_7 is denoted as "Full". True bias and root mean squared error (RMSE) of regression coefficient $\hat{\beta}_1$ of a variable X_1 are denoted as "bias($\times$100)" and "rmse($\times$100)", while bias and RMSE of $\hat{\beta}_1$ to correct model are denoted as "bias($\hat{\beta}_1$)($\times$100)" and "rmse($\hat{\beta}_1$)($\times$100)".

Table 3: Simulation study for linear regression: $vif = 2, \beta = 1, \tau = 0.05$. VIF is variance inflation factor of X_1 conditional on $X_2, \ldots, X_7$ and τ represents the change-in-estimate threshold. Number of simulations, 1000; sample size, 120.

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	83.677	88.316	85.378	94.826
Correct	0	1,000	0	0	0	1.701	41.052
Full	0	0	1,000	0.926	21.594	2.627	46.328
ABE	324	321	355	1.548	19.328	3.249	46.256
abeAIC	354	359	287	2.056	18.785	3.757	46.138
abeBIC	477	351	172	2.491	17.788	4.192	46.138
eABE	324	321	355	1.548	19.328	3.249	46.256
eabeAIC	354	359	287	2.056	18.785	3.757	46.138
eabeBIC	477	351	172	2.491	17.788	4.192	46.138
FE	405	358	237	2.650	18.478	4.351	46.490
BE	393	348	259	2.509	18.713	4.210	46.362
SE	406	358	236	2.650	18.477	4.351	46.490
BE bic	770	207	23	10.266	18.713	11.967	46.362
BE0.2	342	321	337	1.652	19.070	3.353	46.362
BE0.05	661	290	49	7.516	17.955	9.217	47.032

Table 4: Simulation study for linear regression: $vif = 2, \beta = 0, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	85.544	90.223	86.079	95.573
Correct	0	1,000	0	0	0	0.535	41.770
Full	0	0	1,000	0.503	22.977	1.038	47.988
ABE	259	335	406	1.465	19.128	2.000	46.715
abeAIC	288	357	355	1.791	18.873	2.326	46.632
abeBIC	436	328	236	2.211	18.701	2.746	46.597
eABE	259	335	406	1.465	19.128	2.000	46.715
eabeAIC	288	357	355	1.791	18.873	2.326	46.632
eabeBIC	436	328	236	2.211	18.701	2.746	46.597
FE	353	370	277	2.503	18.081	3.038	46.413
BE	342	369	289	2.528	18.301	3.063	46.575
SE	355	371	274	2.535	18.058	3.070	46.424
BEbic	728	238	34	9.349	18.301	9.884	46.575
BE0.2	294	344	362	1.693	18.944	2.228	46.575
BE0.05	611	320	69	6.559	17.869	7.094	47.199

Table 5: Simulation study for linear regression: $vif = 4, \beta = 1, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	125.885	131.943	128.813	136.896
Correct	0	1,000	0	0	0	2.928	53.565
Full	0	0	1,000	-2.510	52.568	0.418	74.441
ABE	258	224	518	-0.539	50.475	2.389	73.475
abeAIC	286	228	486	-0.487	50.435	2.441	73.573
abeBIC	429	190	381	0.200	50.762	3.128	74.041
eABE	258	224	518	-0.539	50.475	2.389	73.475
eabeAIC	286	228	486	-0.487	50.435	2.441	73.573
eabeBIC	429	190	381	0.200	50.762	3.128	74.041
FE	421	360	219	5.881	40.354	8.808	68.699
BE	415	359	226	5.442	41.760	8.369	69.381
SE	424	361	215	6.057	40.582	8.984	68.849
BEbic	767	206	27	19.980	41.760	22.907	69.381
BE0.2	347	345	308	3.559	43.348	6.486	69.381
BE0.05	662	290	48	13.874	39.235	16.802	69.944

Table 6: Simulation study for linear regression: $vif = 4, \beta = 0, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	128.595	134.919	127.766	135.724
Correct	0	1,000	0	0	0	-0.829	53.034
Full	0	0	1,000	-2.561	53.492	-3.389	75.964
ABE	243	196	561	-1.213	51.890	-2.041	75.267
abeAIC	269	201	530	-1.046	51.859	-1.874	75.389
abeBIC	396	170	434	-0.753	52.158	-1.582	75.914
eABE	243	196	561	-1.213	51.890	-2.041	75.267
eabeAIC	269	201	530	-1.046	51.859	-1.874	75.389
eabeBIC	396	170	434	-0.753	52.158	-1.582	75.914
FE	388	374	238	5.532	42.690	4.703	71.480
BE	379	366	255	4.606	44.839	3.778	72.858
SE	391	377	232	5.601	42.859	4.772	71.619
BEbic	765	206	29	20.429	44.839	19.601	72.858
BE0.2	328	340	332	3.049	45.903	2.220	72.858
BE0.05	658	282	60	15.472	42.132	14.643	71.955

In general, for all cases in linear regression ABE methods led to less biased estimate of the variable of interest, X_1, in comparison with other methods. Among ABE methods, method based on AIC had some advantage with respect to the number of correct models but at the same time it has returned slightly larger bias. For both cases, using exact and approximated change in estimate we got the same results. For scenario $VIF = 4$ stepwise method outperformed ABE methods with regard to number of correct models, but at the same time they returned more biased than inflated models while for ABE methods the opposite was true. As we can see from the captions of the previous figures, these results are related to the default values of the parameters, that is for $\tau = 0.05$, $\alpha = 0.2$ and for the fixed sample size $n = 120$. Therefore, we will graphically present what happens if we change the values of parameters. The number of selected models (biased, correct and inflated) is presented in percentages. Abbreviations we used are similar to those from the previous tables. Namely, ABE method proposed by Dunkler et al., is denoted as "ABE", ABE method that uses AIC as a criterion for the blacklist is "aA", ABE method that uses BIC is "aB"; for the same methods that were using exact change-in-estimate instead of approximated the notations we used are: "ea", "eaA" and "eaB"; forward selection that uses AIC as a criterion is "FE"; backward elimination and AIC as a criterion is "BE"; stepwise elimination and AIC as a criterion is "SE"; backward elimination and BIC as a criterion "BEb"; backward elimination and significance level 0.05 as a criterion is denotead as "B05"; backward elimination and significance level 0.2 as a criterion is denotead as "BE2".

From the following bar-plots we can see how different were results in case of linear regression when VIF was 2, $\beta = 1$ but we had three different values of τ.

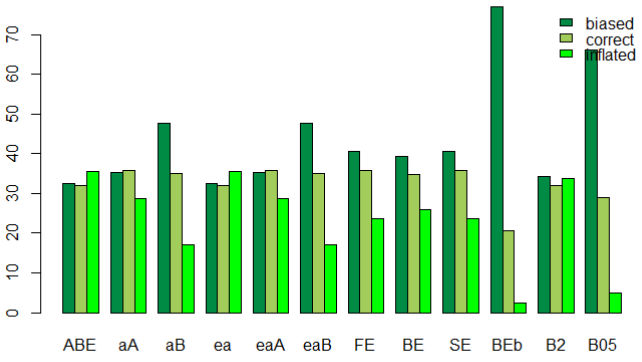

Figure 4: No. of selected models for linear regression for $\tau = 0.05$

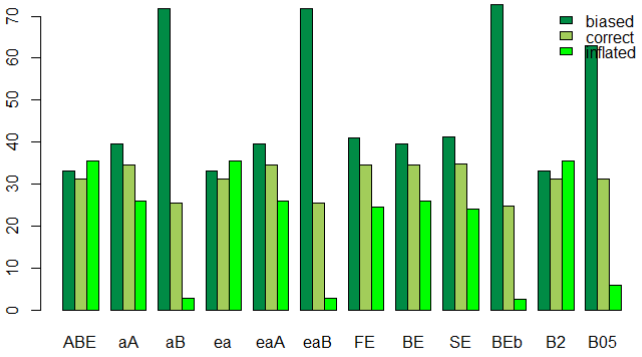

Figure 5: No. of selected models for linear regression for $\tau = 0.10$

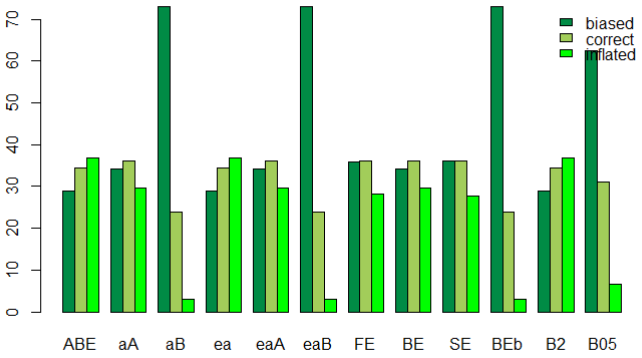

Figure 6: No. of selected models for linear regression for $\tau = 0.20$

From the above figures we see that as we increase the value of τ which represents a change-in-estimate threshold value for parameters from our initial model, the number of biased or correct models is increasing. That is obviously expected, since by increasing the value of τ we are making it difficult for variables to enter the final model. Therefore, the number of inflated models is decreased and as a consequence the number of correct model is increased; or the number of correct models is decreased so we can expect the bigger number of biased models. Let us see the results for linear regression in case VIF=2, $\beta = 1, \tau = 0.05$ if we change the sample size.

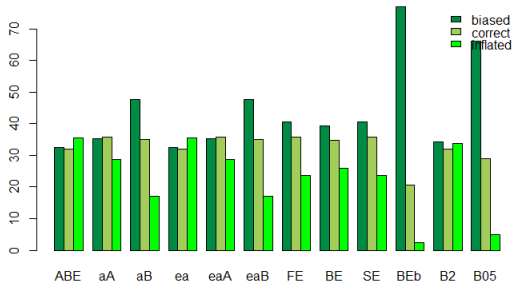

Figure 7: No. of selected models for linear regression for $n = 120$

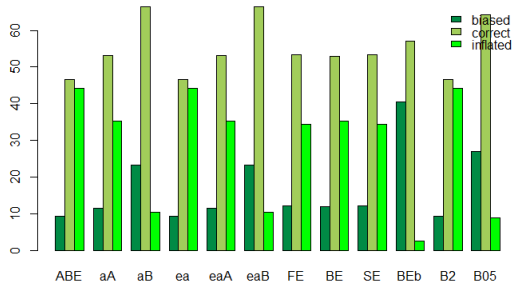

Figure 8: No. of selected models for linear regression for $n = 200$

As we can see from the above figures there is a big difference in the number of biased, correct and inflated models for different sample size. As we could expect the number of correct models is larger for all methods used for variable selection. Therefore, we can conclude that it is not only important which method will we choose but also there are other factors that can affect the selection of the true model, like sample size. In the figures below are presented results for linear regression in case VIF=2 and $\tau = 0.05$ if the sample size is 120 but we change β. From these figures we can conclude that the results in case $\beta = 1$ and $\beta = 0$ if all other parameters were fixed, are not drastically different. It turned out that also in case we changed the sample size from 120 to 200 even though the number of biased, correct and inflated models has significantly changed, the results were similar to those in case sample size was 120 in sense of the proportion for $\beta = 1$ and $\beta = 0$.

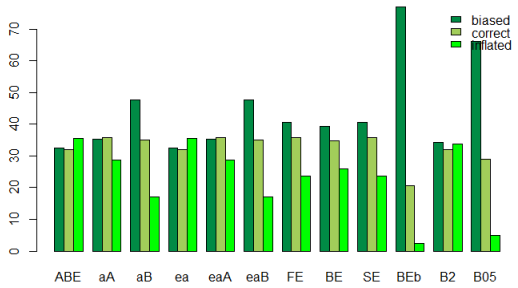

Figure 9: No. of selected models for linear regression for $\beta = 1$

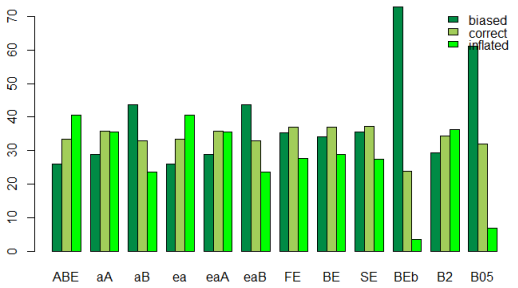

Figure 10: No. of selected models for linear regression for $\beta = 0$

The only thing we have to check more is if there is a difference in the number of correct models for cases where we used VIF=2 and VIF=4. From the below figures we see that in case VIF was 4 the number of inflated models was bigger. We could expected that the number of inflated models for bigger VIF will be bigger since VIF measures how much the variance of an estimated regression coefficient is increased because of the collinearity. If we compare ABE method with the modified ABE methods, that is

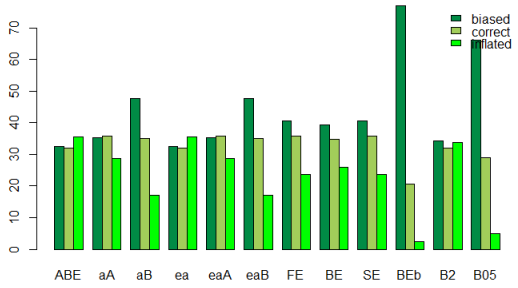

Figure 11: No. of selected models for linear regression for VIF=2

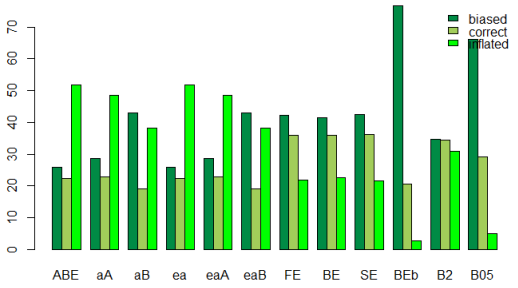

Figure 12: No. of selected models for linear regression for VIF=4

ABE using AIC or BIC as a criterion for the blacklist, we can conclude the following. It turned out that results were similar for ABE and ABE method based on AIC. Namely, the number of correct models was slightly bigger in case we used ABE-AIC method in comparison with the ABE method. For two scenarios (VIF= 2) ABE lead to less biased estimate of β_1, while for the other two (VIF= 4) the bias of $\hat{\beta}_1$ was less in case of ABE-AIC method. At the same time ABE-AIC is prone to choosing larger number of biased models than ABE method. This could be seen as an disadvantage, since maybe it is better to have inflated model than biased one, because in case of biased model we know that at least one of the "important" variables will not be in

the model. As for the ABE-BIC method, we can say that among ABE methods, the number of biased models was the largest for ABE-BIC. This was expected, since BIC is more stringent than AIC, meaning that sometimes BIC results in serious underfitting. The next we will present are results for logistic regression.

8.2 Simulation Results for Logistic Regression

Table 7: Simulation study for logistic regression: $vif = 2, \beta = 1, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	14.762	30.498	25.881	39.109
Correct	0	1,000	0	0	0	11.118	42.661
Full	0	0	1,000	5.516	23.242	16.634	52.258
ABE	21	37	942	5.537	23.139	16.655	52.196
abeAIC	21	37	942	5.538	23.138	16.657	52.195
abeBIC	21	37	942	5.538	23.138	16.657	52.195
eABE	19	37	944	5.402	23.117	16.521	52.128
eabeAIC	20	37	943	5.400	23.116	16.519	52.130
eabeBIC	21	36	943	5.397	23.116	16.515	52.133
FE	118	506	376	4.236	18.363	15.355	49.657
BE	113	492	395	4.501	19.512	15.620	50.341
SE	119	506	375	4.249	18.367	15.367	49.660
BEbic	331	603	66	3.985	19.512	15.104	50.341
BE0.2	91	433	476	4.609	20.230	15.727	50.341
BE0.05	254	622	124	4.003	15.939	15.121	48.564

Table 8: Simulation study for logistic regression: $vif = 2, \beta = 0, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	56.153	60.881	56.503	60.887
Correct	0	1,000	0	0	0	0.350	34.216
Full	0	0	1,000	0.203	18.376	0.554	40.462
ABE	10	45	945	0.401	18.157	0.751	40.333
abeAIC	10	45	945	0.401	18.157	0.751	40.333
abeBIC	11	45	944	0.400	18.159	0.750	40.334
eABE	8	45	947	0.377	18.140	0.727	40.287
eabeAIC	8	45	947	0.377	18.140	0.727	40.287
eabeBIC	10	45	945	0.378	18.139	0.728	40.289
FE	70	528	402	0.759	14.518	1.109	38.518
BE	66	522	412	0.610	14.646	0.960	38.531
SE	70	528	402	0.762	14.537	1.113	38.544
BEbic	246	680	74	2.992	14.646	3.343	38.531
BE0.2	53	461	486	0.458	15.136	0.808	38.531
BE0.05	162	710	128	1.710	12.260	2.060	37.608

Table 9: Simulation study for logistic regression: $vif = 4, \beta = 1, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	53.425	62.494	60.892	69.594
Correct	0	1,000	0	0	0	7.467	48.796
Full	0	0	1,000	3.387	48.696	10.854	70.330
ABE	7	13	980	3.471	48.728	10.938	70.344
abeAIC	8	13	979	3.478	48.732	10.945	70.342
abeBIC	8	13	979	3.478	48.732	10.945	70.342
eABE	7	12	981	3.422	48.718	10.889	70.324
eabeAIC	7	12	981	3.422	48.718	10.889	70.324
eabeBIC	9	12	979	3.425	48.719	10.892	70.328
FE	136	542	322	3.054	35.812	10.521	61.809
BE	133	530	337	2.379	37.128	9.846	62.140
SE	136	543	321	3.099	35.871	10.566	61.838
BEbic	402	542	56	9.112	37.128	16.578	62.140
BE0.2	104	453	443	3.070	39.148	10.537	62.140
BE0.05	289	606	105	6.321	30.718	13.788	61.291

Table 10: Simulation study for logistic regression: $vif = 4, \beta = 0, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	89.275	94.438	90.127	94.182
Correct	0	1,000	0	0	0	0.852	43.002
Full	0	0	1,000	0.528	44.305	1.380	62.694
ABE	10	14	976	0.583	44.246	1.435	62.712
abeAIC	10	14	976	0.583	44.246	1.435	62.712
abeBIC	10	14	976	0.583	44.246	1.435	62.712
eABE	10	15	975	0.554	44.253	1.406	62.709
eabeAIC	10	15	975	0.554	44.253	1.406	62.709
eabeBIC	10	15	975	0.554	44.253	1.406	62.709
FE	87	541	372	2.504	32.870	3.356	55.931
BE	83	531	386	2.468	34.066	3.320	56.756
SE	87	541	372	2.508	32.879	3.360	55.930
BEbic	275	652	73	7.920	34.066	8.772	56.756
BE0.2	69	459	472	2.009	36.482	2.860	56.756
BE0.05	195	673	132	5.568	27.834	6.420	53.785

Results for logistic regression were quite different from the results for linear regression. Namely, bias of $\hat{\beta}_1$ was almost the same for ABE and stepwise selection methods; only for one scenario it was considerably less in case we used ABE methods. Results were the same for exact and approximated change in estimate. For all scenarios all ABE

methods returned around 95% of inflated models.

Here, results did not change so much as for linear regression when we changed the sample size. From below figures we see that the number of correct models was slightly bigger for bigger sample size.

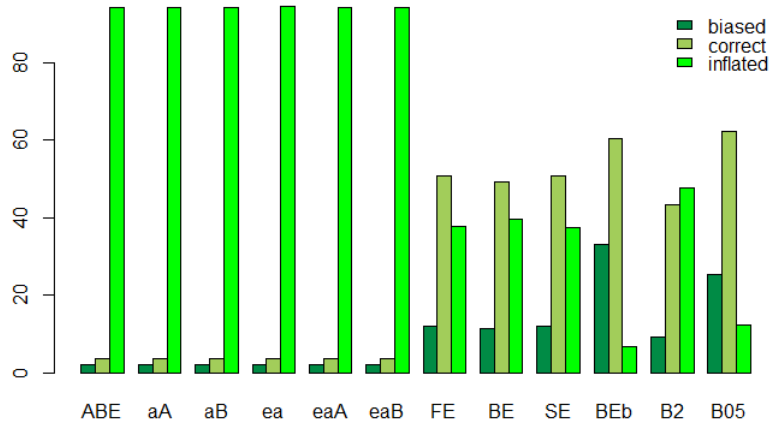

Figure 13: No. of selected models for logistic regression for $n = 120$

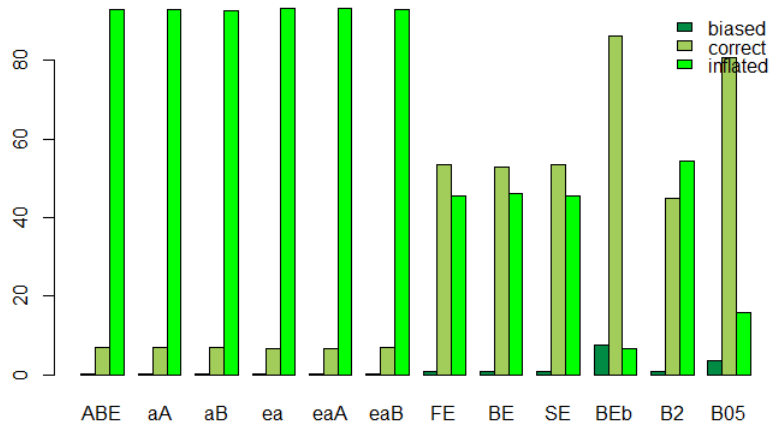

Figure 14: No. of selected models for logistic regression for $n = 200$

From the following bar-plots we can see how different were results in case of logistic regression when VIF was 2, $\beta = 1$ but we had three different values of τ.

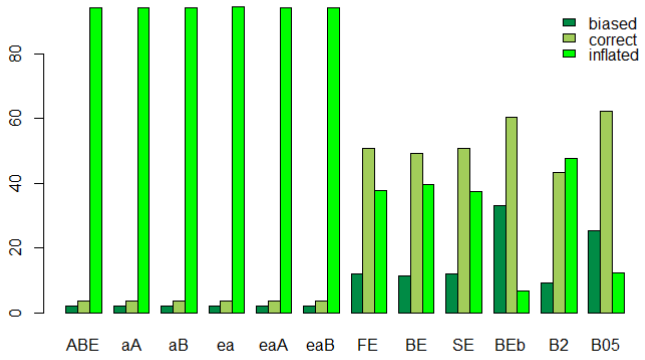

Figure 15: No. of selected models for logistic regression for $\tau = 0.05$

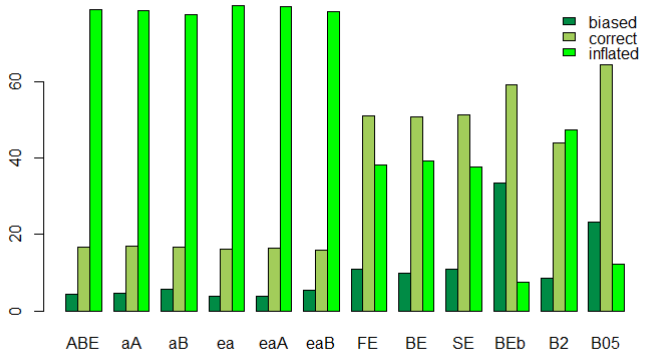

Figure 16: No. of selected models for logistic regression for $\tau = 0.10$

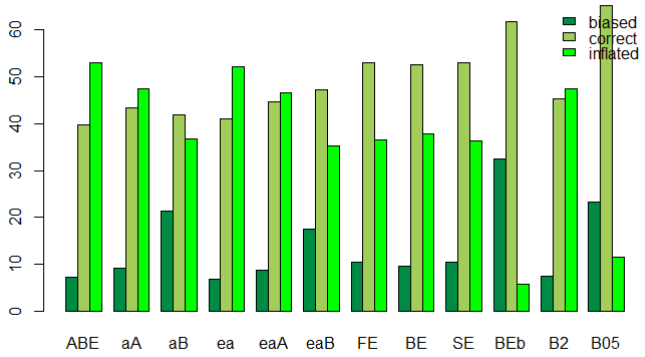

Figure 17: No. of selected models for logistic regression for $\tau = 0.20$

From this we can conclude that if we want to avoid tendency of ABE methods to

inflated models we have to set a high value for parameter τ. That is, we have to require that the threshold value of change-in-estimated is more than 0.05.

To summarize, if we compare ABE method with the modified ABE methods, that is ABE using AIC or BIC, in case of logistic regression it turned out that the results were the same for all ABE methods. By increasing the sample size results were still the same. However, by changing the threshold value for the change-in-estimate there were differences in the number of selected models among ABE methods. Namely, the number of correct models was the largest in case of ABE-AIC method.

8.3 Simulation Results for Cox Regression

Table 11: Simulation study for Cox regression: $vif = 2, \beta = 1, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	-4.696	22.365	-0.084	20.011
Correct	0	1,000	0	0	0	4.612	23.230
Full	0	0	1,000	2.734	12.697	7.346	27.657
ABE	0	116	884	2.706	12.418	7.318	27.369
abeAIC	0	117	883	2.707	12.421	7.320	27.377
abeBIC	0	119	881	2.720	12.412	7.332	27.362
eABE	0	115	885	2.678	12.411	7.291	27.351
eabeAIC	0	116	884	2.673	12.412	7.286	27.355
eabeBIC	0	118	882	2.677	12.406	7.289	27.351
FE	4	560	436	1.447	9.575	6.060	26.074
BE	4	550	446	1.540	9.835	6.152	26.260
SE	4	560	436	1.447	9.575	6.060	26.074
BEbic	8	881	111	0.971	9.835	5.583	26.260
BE0.2	1	467	532	1.870	10.396	6.482	26.260
BE0.05	5	824	171	0.916	6.809	5.529	24.712

Table 12: Simulation study for Cox regression: $vif = 2, \beta = 0, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	45.676	49.289	45.314	48.469
Correct	0	1,000	0	0	0	-0.362	19.911
Full	0	0	1,000	0.369	11.700	0.007	23.673
ABE	0	129	871	0.392	11.423	0.030	23.604
abeAIC	0	131	869	0.395	11.420	0.034	23.607
abeBIC	0	134	866	0.383	11.417	0.021	23.606
eABE	0	127	873	0.377	11.417	0.016	23.576
eabeAIC	0	128	872	0.380	11.414	0.018	23.578
eabeBIC	0	130	870	0.373	11.412	0.011	23.574
FE	0	583	417	0.059	8.792	-0.303	22.166
BE	0	569	431	0.008	9.033	-0.353	22.201
SE	0	584	416	0.056	8.791	-0.306	22.167
BEbic	5	884	111	0.225	9.033	-0.137	22.201
BE0.2	0	463	537	0.044	9.745	-0.318	22.201
BE0.05	2	816	182	0.150	6.235	-0.211	21.049

Table 13: Simulation study for Cox regression: $vif = 4, \beta = 1, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	25.390	37.294	30.010	38.730
Correct	0	1,000	0	0	0	4.620	28.737
Full	0	0	1,000	2.818	27.485	7.438	41.231
ABE	0	52	948	2.849	27.342	7.469	41.165
abeAIC	0	52	948	2.849	27.342	7.469	41.165
abeBIC	0	52	948	2.853	27.345	7.473	41.172
eABE	0	51	949	2.812	27.331	7.432	41.134
eabeAIC	0	51	949	2.812	27.331	7.432	41.134
eabeBIC	0	51	949	2.816	27.334	7.436	41.141
FE	5	578	417	1.557	19.293	6.177	35.288
BE	5	554	441	1.371	20.662	5.991	36.106
SE	5	578	417	1.587	19.312	6.207	35.266
BEbic	15	896	89	0.714	20.662	5.334	36.106
BE0.2	4	460	536	1.678	22.308	6.298	36.106
BE0.05	9	840	151	0.887	14.268	5.507	32.667

Table 14: Simulation study for Cox regression: $vif = 4, \beta = 0, \tau = 0.05$

	biased	correct	inflated	bias($\hat{\beta}_1$)($\times$100)	rmse($\hat{\beta}_1$)($\times$100)	bias($\times$100)	rmse($\times$100)
Uni	1,000	0	0	72.257	76.271	72.160	75.057
Correct	0	1,000	0	0	0	-0.097	25.431
Full	0	0	1,000	0.283	28.837	0.186	39.375
ABE	0	43	957	0.282	28.663	0.185	39.195
abeAIC	0	43	957	0.278	28.665	0.181	39.197
abeBIC	0	43	957	0.278	28.665	0.181	39.197
eABE	0	44	956	0.298	28.653	0.201	39.188
eabeAIC	0	44	956	0.298	28.653	0.201	39.188
eabeBIC	0	44	956	0.298	28.653	0.201	39.188
FE	5	550	445	0.853	21.169	0.756	33.825
BE	4	534	462	1.073	22.150	0.976	34.642
SE	5	550	445	0.853	21.169	0.756	33.825
BEbic	15	862	123	0.493	22.150	0.396	34.642
BE0.2	4	440	556	0.987	23.787	0.890	34.642
BE0.05	6	800	194	0.393	16.325	0.296	30.941

Results show that the bias in Cox regression after ABE methods for scenario $\beta = 1$ was larger in comparison with the stepwise methods but for scenario $\beta = 0$ was less. Backward elimination outperformed ABE methods with regard to the number of correct models. Results for exact and approximated change in estimate were almost the same.

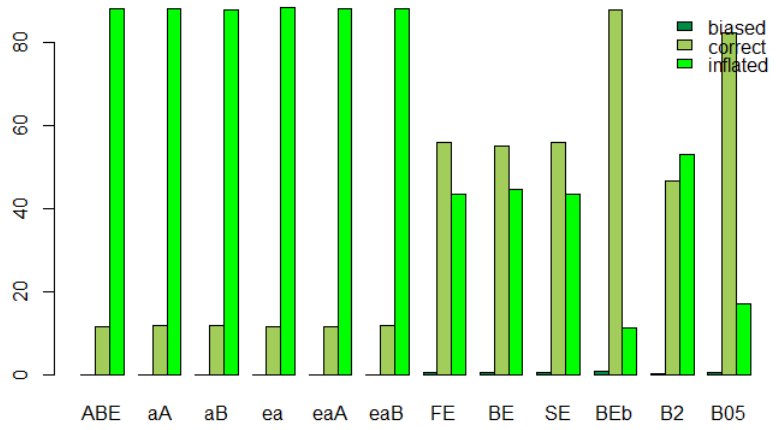

Figure 18: No. of selected models for Cox regression for $n = 120$

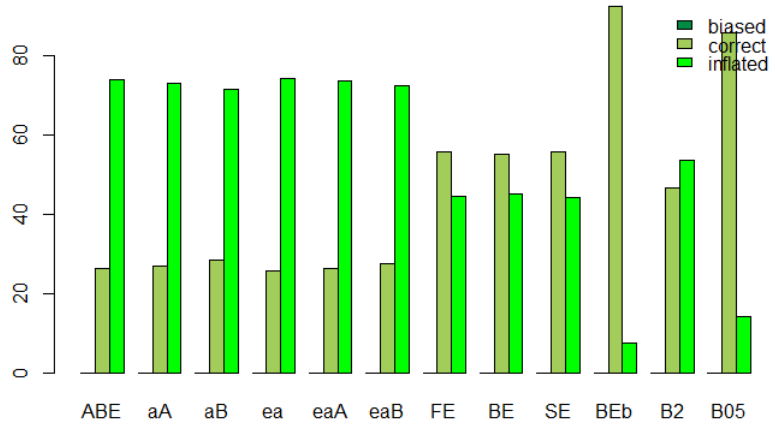

Figure 19: No. of selected models for Cox regression for $n = 200$

The results were better for bigger sample size. From the following bar-plots we can see how different were results in case of Cox regression when VIF was 2, $\beta = 1, n = 120$ but we had three different values of τ.

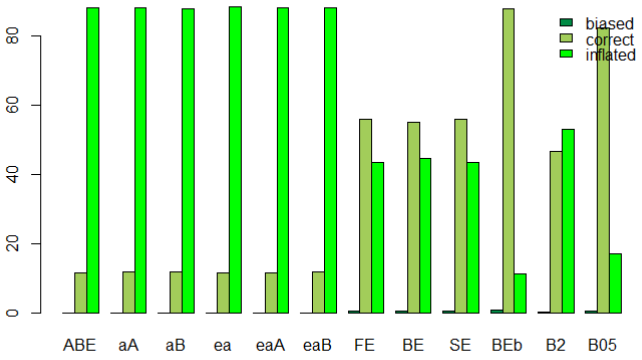

Figure 20: No. of selected models for Cox regression for $\tau = 0.05$

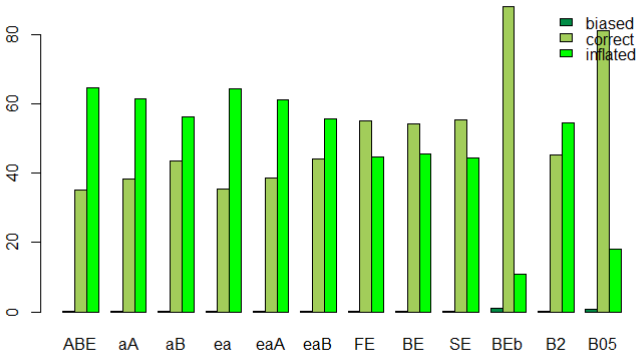

Figure 21: No. of selected models for Cox regression for $\tau = 0.10$

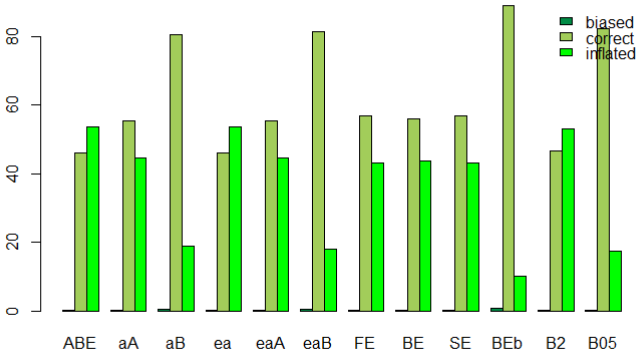

Figure 22: No. of selected models for Cox regression for $\tau = 0.20$

As in case of the logistic regression it turned out that for bigger τ results are better. This is because ABE methods with their default values of parameters ($\tau = 0.05$)

are allowing variables to enter the model. This can be seen as an advantage since confounders will be detected. It is up to us to decide what is the best cut-off percentage of change in the exposure-effect estimate.

Also in case of Cox regression the results were the same for all ABE methods. When we changed the threshold value for the change-in-estimate the number of correct models was the largest in case of ABE-BIC method. This makes sense because if ABE methods tend to choose inflated models for Cox regression and since BIC is the strictest criterion there are bigger chances that it will return more correct than inflated models. By changing the sample size from 120 to 200 results among ABE methods were the same.

8.4 Simulation study with different covariance structure of independent variables

We wanted to check what happens if the covariance structure of $X_1, \ldots, X_7$ is defined differently. We simulated seven independent standard normal variables $X_1, \ldots, X_7$. Dependent variable was defined as $Y^* = \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7$. True model form was $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7 + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. To simulate linear, logistic and Cox regression we used the same procedure as in previous simulations. We checked what is happening if $\beta_1 = \beta_2 = \beta_4 = \beta_7 = 1, 0.5$ or 0. We were using different values of σ. Namely, $\sigma = 1, 0.5$ and 2. Here, we were interested just in the number of biased, correct and inflated models, not in bias and RMSE of estimated coefficients.

From the following 6 tables we can see the results for linear regression. It turned out that results were similar for τ equal $0.05, 0.1$ and 0.2, meaning that we did not have "strong" confounders. We can notice that τ had influence on the final model if we compare the results of ABE and BE method. On the other side, for this simulation setup results among ABE methods were quite different. Namely, ABE method based on BIC criterion outperformed the other two ABE methods, but the results were the same as in case BE based on BIC criterion. Even though ABE and BE methods can give almost the same results, from here we can conclude ABE methods are more useful for certain type of data, when we expect or we know that we have confounders. If we do not have any subject matter knowledge, meaning that we can and we do not have to have confounders among our variables, we suggest to use ABE methods, since for sure they will not give worse results than backward elimination procedure.

Defining Y as $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7$ for $\beta_i, i = 1, 2, 4, 7$ all equal $1, 0.5$ or 0 had expected influence on final results. Namely, for $\beta_i = 1$ for $i = 1, 2, 4, 7$ and $\beta_i = 0.5, i = 1, 2, 4, 7$ the results were more or less the same. Small difference in the number of inflated models has been noticed. Namely, for case $\beta_i = 1$ for $i = 1, 2, 4, 7$ there was more inflated models than in case when coefficients were 0.5. When coefficients were all equal to 0 all methods for variable selection returned around 99% of biased models. Again, approximated and exact ABE methods gave the same results.

From the following tables we can see those results for linear regression in case the sample size was 120, $\tau = 0.05$, β_i for $i = 1, 2, 4, 7$ being 1 or 0.5 and for σ being $0.5, 1$ and 2 indicating how much variability is in dependent variable.

Table 15: Simulation study for linear regression: $\beta_i = 0.5$ for $i = 1, 2, 4, 7; \tau = 0.05; \sigma = 0.5, 1, 2$

	$\sigma = 0.5$			$\sigma = 1$			$\sigma = 2$		
	biased	correct	inflated	biased	correct	inflated	biased	correct	inflated
ABE	0	482	518	1	480	519	304	317	379
ABE AIC	0	570	430	1	559	440	364	343	293
ABE BIC	0	904	96	4	897	99	749	225	26
eABE	0	482	518	1	480	519	304	317	379
eABE AIC	0	570	430	1	559	440	364	343	293
eABE BIC	0	904	96	4	897	99	749	225	26
FE	0	571	429	1	559	440	367	340	293
BE	0	570	430	1	559	440	364	343	293
SE	0	571	429	1	559	440	368	341	291
BE bic	0	904	96	4	897	99	750	224	26
BE alpha 0.2	0	482	518	1	480	519	304	317	379
BE alpha 0.05	0	838	162	3	835	162	662	274	64

Table 16: Simulation study for linear regression: $\beta_i = 1$ for $i = 1, 2, 4, 7; \tau = 0.05; \sigma = 0.5, 1, 2$

	$\sigma = 0.5$			$\sigma = 1$			$\sigma = 2$		
	biased	correct	inflated	biased	correct	inflated	biased	correct	inflated
ABE	0	456	544	0	464	536	0	507	493
ABE AIC	0	539	461	0	551	449	0	590	410
ABE BIC	0	901	99	0	896	104	2	903	95
eABE	0	456	544	0	464	536	0	507	493
eABE AIC	0	539	461	0	551	449	0	590	410
eABE BIC	0	901	99	0	896	104	2	903	95
FE	0	543	457	0	553	447	0	592	408
BE	0	539	461	0	551	449	0	590	410
SE	0	543	457	0	553	447	0	592	408
BE bic	0	901	99	0	896	104	2	903	95
BE alpha 0.2	0	456	544	0	464	536	0	507	493
BE alpha 0.05	0	841	159	0	826	174	0	832	168

Results for logistic regression were quite different with regards to the previous simulation setup. In the following tables we can see results in case the sample size was 120, $\tau = 0.05$, $\beta_i = 1$ for $i = 1, 2, 4, 7$ being 1 or 0.5 and for $\sigma = 1$. Again, for $\beta_i = 1$ for $i = 1, 2, 4, 7$ being 0 we got around 99% biased models.

Table 17: Simulation study for logistic regression: $\beta_i = 0.5$ and 1 for $i = 1, 2, 4, 7; \tau = 0.05; \sigma = 1$

	$\beta = 0.5$			$\beta = 1$		
	biased	correct	inflated	biased	correct	inflated
ABE	558	195	247	25	296	679
ABE AIC	612	187	201	26	311	663
ABE BIC	738	132	130	39	319	642
eABE	569	200	231	26	339	635
eABE AIC	632	186	182	29	361	610
eABE BIC	798	109	93	66	382	552
FE	671	196	133	46	543	411
BE	669	195	136	45	542	413
SE	671	196	133	46	543	411
BE bic	957	40	3	205	726	69
BE alpha 0.2	593	209	198	38	458	504
BE alpha 0.05	906	84	10	147	726	127

If we look at the results from previous simulation where for all scenarios the number of inflated models was around 90% we can conclude that of course covariance structure of our variables can have great influence on any variable selection methods, meaning that we can not speak in general about "good" method for selection of variables since as we can see among others it depends on data set what will we get as a final model.

From the following table we see that there are differences in Cox regression for β_i for $i = 1, 2, 4, 7$ being all 0.5 or 1. Comparing with the first part of simulation study where we had $\beta_i = 1$ for $i = 1, 2, 4, 7$ all equal 1 but different covariance structure and as a result around 95% of inflated models we can conclude that covariance structure has its own influence. We checked what happens for different values of τ. Results are not presented here but it turned out that by increasing the value of τ for case $\beta_i = 1$ for $i = 1, 2, 4, 7$ equal 0.5 the number of biased models was considerably bigger and around 80% for τ equal 0.1 and 0.2 if we used ABE BIC method.

Therefore, it does not hold in general that in Cox regression if we use bigger value of τ we can expect correct model, since in can happen that the result is biased model. That means, there is no general rule, there is no "recipe" for good final model when

Table 18: Simulation study for Cox regression: $\beta_i = 0.5$ and 1 for $i = 1, 2, 4, 7; \tau = 0.05; \sigma = 1$

	$\beta = 0.5$			$\beta = 1$		
	biased	correct	inflated	biased	correct	inflated
ABE	303	263	434	1	347	652
ABE AIC	354	274	372	3	379	618
ABE BIC	559	213	228	6	427	567
eABE	304	263	433	1	353	646
eABE AIC	354	276	370	3	387	610
eABE BIC	568	204	228	7	441	552
FE	421	302	277	3	521	476
BE	415	305	280	3	515	482
SE	424	305	271	3	521	476
BE bic	794	175	31	26	842	132
BE alpha 0.2	337	285	378	1	440	559
BE alpha 0.05	699	233	68	11	796	193

it comes to variable selection. Everything depends on the data we have and on our patience and careful work we have to invest in order to expect good results.

9 Conclusion

We reviewed the most common variable selection procedures, their advantages and disadvantages. Among others we presented augmented backward elimination method recently proposed by Dunkler et al which is a combination of backward elimination procedure and approximated change-in-estimate criterion. The method proposed by Dunkler et al. is only available in SAS, so our aim was to make an R package which will implement their method for several statistical models. Since this method chooses variables based on their significance and approximated change-in-estimated, we extended it such that information criteria AIC and BIC can also be used and that we can choose between approximated and exact change-in-estimate. Also, extended augmented backward elimination is available for all generalized linear models, not just for logistic regression. We performed extensive simulation studies.

In general, for all scenarios for linear regression ABE methods lead to less biased estimate of the variable of interest in comparison with other methods but stepwise methods outperformed ABE methods in sense of the number of correct models. For bigger sample size results were better.

Results for logistic regression were quite different from the results for linear regression. Namely, bias of $\hat{\beta}_1$ was almost the same for ABE and stepwise selection methods; only for one scenario it was considerably less in case we used ABE methods. Results were the same for exact and approximated change in estimate. For all scenarios all ABE methods returned around 95% of inflated models.

If we want to avoid tendency of ABE methods to inflated models we have to set a high value for parameter τ. That is, we have to require that the change in estimated coefficient is more than 5%. This can be seen as an advantage since confounders will be detected. It is up to us to decide what is the best cut-of percentage of change in the exposure-effect estimate. Similar results were in case of Cox regression.

We checked what happens if the covariance structure of independent variables is defined differently but data are generated in the same way. In general, results were quite different, especially among ABE methods.

If we compare ABE method with the modified ABE methods, that is ABE using AIC or BIC as a criterion for the blacklist, we can conclude the following.

In case of the linear regression it turned out that results were similar for ABE and ABE

method based on AIC. Namely, the number of correct models was slightly bigger in case we used ABE-AIC method in comparison with the ABE method. For two scenarios (VIF= 2) ABE lead to less biased estimate of β_1, while for the other two (VIF= 4) the bias of $\hat{\beta}_1$ was less in case of ABE-AIC method. At the same time ABE-AIC is prone to choosing larger number of biased models than ABE method. This could be seen as an disadvantage, since maybe it is better to have inflated model than biased one, because in case of biased model we know that at least one of the "important" variables will not be in the model. As for the ABE-BIC method, we can say that among ABE methods, the number of biased models was the largest for ABE-BIC. This was expected, since BIC is more stringent than AIC, meaning that sometimes BIC results in serious underfitting.

In case of logistic regression it turned out that the results were the same for all ABE methods. For logistic regression by increasing the sample size results were still the same. However, by changing the threshold value for the change-in-estimate there were differences in the number of selected models among ABE methods. Namely, the number of correct models was the largest in case of ABE-AIC method.

Also in case of Cox regression the results were the same for all ABE methods. Later, when we changed the threshold value for the change-in-estimate the number of correct models was the largest in case of ABE-BIC method. This makes sense because ABE methods tend to choose inflated models for Cox regression and since BIC is the strictest criterion there are bigger chances that it will return more correct than inflated models. By changing the sample size from 120 to 200 results among ABE methods were the same.

To conclude we would like to emphasize that there is no general rule, there is no "recipe" for good model when it comes to variable selection. Everything depends on the data we have and on our patience and careful work we have to invest in order to expect good results. There is no single method that will give reliable results for every possible situation. Sample size, correlation structure, cut-of percentage of change-in-estimate are just some of the factors that can have influence on final model in variable selection process.

10 Povzetek naloge v slovenskem jeziku

Trenutne raziskave na medicinskem področju vključujejo veliko število potencijalnih pojasnjevalnih spremenljivk. Cilj teh študij je oceniti povezanost med pojasnjevalnimi spremenljivkami in izidom. Če imamo na voljo velik nabor potencialnih pojasnjevalnih spremenljivk, izbira najprimernejših spremenljivk modela na objektiven in praktičen način običajno ni trivialna naloga. Pred kratkim so Dunkler et al. predlagali novo metodo za izbiro spremenljivk - Augmented backward elimination [6].
Metoda je dostopna samo v programskem jeziku SAS, tako da je naš prvi cilj bil sprogramirati metodo tudi v programskem jeziku R. ABE metoda izbira spremenljivke na podlagi značilnosti in na podlagi standardizirane spremembe v oceni koeficienta. Pomembno je poudariti, da so Dunkler et al. predlagali aproksimacijo za spremembo v oceni koeficienta namesto eksaktnega izračuna. Ker nima teoretične podpore na podlagi katere bi vedli točno kdaj je aproksimacija dobra, pri pripravi R funkcije smo omogočili tudi uporabo eksaktnega izračuna. Potencialna težava metode ABE je tudi, da spremenljivke izbira na podlagi njihove značilnosti. Nedavne raziskave s tega področja so pokazale, da izbira spremenljivk na podlagi drugih kriterij, kot sta na primer Akaikejev informacijski kriterij (AIC) ali bayesovski informacijski kriterij (BIC), deluje bolje. Zaradi tega, v nalogi smo predstavili posplošitve metode ABE, katere omogočajo tudi uporabo AIC in BIC.
V nalogi smo najprej predstavili problem izbire spremenljivk in obstoječe metode za izbiro spremenljivk in predstavili znane težave, ki jih imajo te metode. Glede na to da je metoda ABE sprogramirana za linearno, logistično in Cox regresijo, predstavili smo posplošene linearne regresijske modele in Coxov model. Natančno smo predstavili metodo ABE kot tudi njene izboljšave. Predstavili smo R paket, oziroma kodo in opis delovanja funkcije; kaj so argumenti in kaj funkcija izračuna. Na koncu smo s simulacijami prikazali delovanje predlaganih izboljšav metode ABE in jih primerjali z osnovno metodo ABE ter s preostalimi metodami, ki so že sprogramirane v R.

11 Bibliography

[1] Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975. *(Cited on page 20.)*

[2] Esben Budtz-Jørgensen, Niels Keiding, Philippe Grandjean, and Pal Weihe. Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. *Annals of epidemiology*, 17(1):27–35, 2007. *(Cited on page 21.)*

[3] Ken P Burnham and DR Anderson. P values are only an index to evidence: 20th-vs. 21st-century statistical science. *Ecology*, 95(3):627–630, 2014. *(Cited on pages 31 and 33.)*

[4] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003. *(Cited on pages 4 and 33.)*

[5] Houtao Deng and George Runger. Feature selection via regularized trees. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012. *(Cited on page 7.)*

[6] Daniela Dunkler, Max Plischke, Karen Leffondré, and Georg Heinze. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PloS one*, 9(11):e113677, 2014. *(Cited on pages VIII, 1, 21, 25, 51, and 76.)*

[7] Heinze G Dunkler D. A SAS macro for augmented backward elimination. Combining significance and change-in-estimate criteria in a pragmatic and purposeful way to develop statistical models. Technical report, Medical University of Vienna, 2014. *(Cited on page 25.)*

[8] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977. *(Cited on page 20.)*

[9] Andrzej Gałecki and Tomasz Burzykowski. *Linear mixed-effects models using R: A step-by-step approach.* Springer Science & Business Media, 2013. *(Cited on page 42.)*

[10] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer, 2015. *(Cited on pages 7 and 17.)*

[11] Georg Heinze and Daniela Dunkler. Five myths about variable selection. *Transplant International*, 2016. *(Cited on page 8.)*

[12] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976. *(Cited on page 3.)*

[13] David W Hosmer, Stanley Lemeshow, and Susanne May. Applied survival analysis. 2011. *(Cited on pages 23 and 24.)*

[14] Paul H Lee. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *Journal of Epidemiology*, 24(2):161–167, 2014. *(Cited on pages 21 and 22.)*

[15] Alan Miller. *Subset selection in regression.* CRC Press, 2002. *(Cited on page 6.)*

[16] Paul A Murtaugh. In defense of p values. *Ecology*, 95(3):611–617, 2014. *(Cited on pages VIII, 31, and 32.)*

[17] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007. *(Cited on page 18.)*

[18] Terry M Therneau and Thomas Lumley. Package âĂŸsurvivalâĂŹ, 2017. *(Cited on page 20.)*

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. *(Cited on page 7.)*

[20] Wikipedia. Exponential family — wikipedia, the free encyclopedia, 2017. [Online; accessed 24-April-2017]. *(Cited on page 11.)*

[21] Wikipedia. General linear model — wikipedia, the free encyclopedia, 2017. [Online; accessed 24-April-2017]. *(Cited on page 10.)*

[22] Wikipedia. Logistic function — wikipedia, the free encyclopedia, 2017. [Online; accessed 24-April-2017]. *(Cited on page 16.)*

[23] Wikipedia. Model selection — wikipedia, the free encyclopedia, 2017. [Online; accessed 27-April-2017]. *(Cited on page 1.)*

[24] Wikipedia. Statistical model — wikipedia, the free encyclopedia, 2017. [Online; accessed 27-April-2017]. *(Cited on page 1.)*

[25] Faisal Maqbool Zahid and Gerhard Tutz. Multinomial logit models with implicit variable selection. *Advances in Data Analysis and Classification*, 7(4):393–416, 2013. *(Cited on page 7.)*