Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

8-21-2019 2:00 PM

Towards Using Model Averaging To Construct Confidence Intervals In Logistic Regression Models

Artem Uvarov
The University of Western Ontario

Supervisor
Zou, Guangyong
The University of Western Ontario

Graduate Program in Epidemiology and Biostatistics
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Recommended Citation

Uvarov, Artem, "Towards Using Model Averaging To Construct Confidence Intervals In Logistic Regression Models" (2019). *Electronic Thesis and Dissertation Repository*. 6388.
https://ir.lib.uwo.ca/etd/6388

ABSTRACT

Regression analyses in epidemiological and medical research typically begin with a model selection process, followed by inference assuming the selected model has generated the data at hand. It is well-known that this two-step procedure can yield biased estimates and invalid confidence intervals for model coefficients due to the uncertainty associated with the model selection. To account for this uncertainty, multiple models may be selected as a basis for inference. This method, commonly referred to as model-averaging, is increasingly becoming a viable approach in practice.

Previous research has demonstrated the advantage of model-averaging in reducing bias of parameter estimates. However, there is lack of methods for constructing confidence intervals around parameter estimates using model-averaging. In the context of multiple logistic regression models, we propose and evaluate new confidence interval estimation approaches for regression coefficients. Specifically, we study the properties of confidence intervals constructed by averaging tail errors arising from confidence limits obtained from all models included in model-averaging for parameter estimation. We propose model-averaging confidence intervals based on the score test. For selection of models to be averaged, we propose the bootstrap inclusion fractions method.

We evaluate the performance of our proposed methods using simulation studies, in a comparison with model-averaging interval procedures based on likelihood ratio and Wald tests, traditional stepwise procedures, the bootstrap approach, penalized regression, and the Bayesian model-averaging approach.

Methods with good performance have been implemented in the ‘mataci’ R package, and illustrated using data from a low birth weight study.

KEYWORDS: Model-averaging; Logistic regression; Confidence interval; Score function.

SUMMARY FOR LAY AUDIENCE

Data analysis in medical research often involves regression analysis that examines the associations between outcome and independent variables. Analysis consists of selection of these variables and estimation of their effects. A point estimate usually varies from sample to sample, meaning that the estimated effect has some distribution. The 95% confidence interval, a range around a point estimate within which the true effect is likely to fall, is used to quantify the uncertainty associated with the estimates. Tail errors on both sides of a valid confidence interval should be similar and close to the specified limit.

Unfortunately, using the same data to construct confidence intervals usually leads to biased results, especially in small samples. The coverage of confidence intervals obtained by such “double use” of the data is often below the specified limit. To address this problem, it was proposed to use several regression models, which results are averaged. The selection of candidate models is important for the averaging process. If done correctly it allows one to accelerate the computations and also to improve precision of results, while a insufficient set of models can negatively affect the final conclusions.

Model-averaging makes it possible to obtain more accurate point estimates, but many methods for constructing confidence intervals for such averaged estimates suffer from inaccuracy, especially if samples sizes are not large. Such intervals are often too short, and the confidence level is much lower than the specified level.

In this work, we proposed an approach for selecting candidate models that reduces the number of required models, but saves the information that can be obtained from the data. We also proposed a method that constructs valid and accurate confidence intervals for regression coefficients even for small samples. We used a method that suggests averaging the tail errors over selected candidate models. The developed methods are more accurate, but are less traditional variants of the model-averaged tail error method. We focused on building confidence intervals for logistic regression models that evaluate the effect of variables on a binary dependent variable. To demonstrate the superiority of the proposed methods, we compared them with frequently used methods.

ACKNOWLEDGMENTS

This research would have been impossible without the aid and support of many amazing people. There are no proper words to convey my deep gratitude and respect to my Ph.D. supervisor, Dr. GuangYong Zou for his patient guidance, insightful comments and invaluable experience I gained. I cannot imagine a better supervisor to learn from. In the most difficult time, he found words that supported and encouraged me.

I would like to thank Dr. Yun-Hee Choi, for a tremendous help as a supervisory committee member in editing and shaping this work. Her optimism and helpful suggestions gave me confidence that I was on the right track. I would also like to acknowledge Dr. John Koval, Dr. Wenqing He and Dr. Ricardas Zitikis for their helpful suggestions and constructive comments that allowed me to generalize and improve this study.

I am indebted to my beloved family and closest friends for their inexhaustible moral support and encouragement over these years. Despite the long distance between us, they always felt when I needed their wise counsel and sympathetic ears.

This work has been partly financially supported by the Western Graduate Research Scholarship. I would like to thank the partial financial support from Ontario Neurodegenerative Disease Research Initiative (ONDRI). I would also like to thank St. Joseph Hospital, in particular, Dr. Jeff Mahon, Dr. Tamara Spaic and Dr. Selina Liu from the endocrinology department for partial financial support and priceless experience I gained working with them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Routine data analysis in epidemiologic or health sciences research is typically conducted by first selecting a model, followed by obtaining point estimates and standard errors as well as p-values. The same data set is used for both model selection and statistical inference. When separate data sets are use separately for each purpose, results are usually valid. However, the use of the same data set repeatedly can lead to incorrect estimation of standard errors, biased p-values and invalid confidence intervals (Berk et al., 2010; Freedman et al., 1988). This is because traditional inference procedures are usually developed by assuming a given correct model. Yet the model selection process usually involves multiple comparisons of the models. Such data snooping can lead to a misspecified model and increased Type I error.

This problem has long been documented in the literature (e.g. Pötscher, 1991 and Kabaila, 2005). Leeb and Pötscher (2005) and Leeb (2006) pointed out that the distribution of post-selected variables cannot be estimated because the estimation error is not uniformly small even if the sample size goes to infinity. Nevertheless, a review by Walter and Tiemeier (2009) showed that 20% of the articles published in the four leading epidemiological journals in 2008 used either forward or backward stepwise selection methods. A recent review done by Fernández-Niño et al. (2018) found that stepwise selection based methods were used in 50% of the published articles between 2000 and 2017. These results suggest that epidemiological and medical research require more comprehensive and principled analytical procedures.

Common model selection procedures usually result in a single final model, which is then assumed to be the true model upon which the subsequent statistical inference is based. However, such inference does not reflect the uncertainty in the model selection process, so it leads to underestimation of the standard errors and consequential undercoverage of the confidence intervals (Berk et al., 2010). A possible solution to this problem may be the use of the model-averaging technique that averages over a set of candidate models instead of using a single model.

Automated techniques have been developed for model selection and sequential inference. The most popular approach is the stepwise procedure using prespecified criteria such as F-test, Akaike Information Criterion ($AIC = -2\ln L + 2k$) (Akaike, 1973) or Bayesian Information Criterion ($BIC = -2\ln L + 2k\ln(n)$) (Schwarz, 1978), where $\ln L$ is a log-likelihood function, n is a sample size and k is a number of estimated coefficients, to compare nested models at each step, and to decide whether to leave or to remove one of the variables. One widely known limitation of the stepwise selection procedure is that it yields biased regression coefficients and confidence intervals that are falsely narrow (Altman and Andersen, 1989; Hurvich and Tsai, 1990). Although it is known that stepwise methodology has performance problems and should be used cautiously, it is still the most popular model selection and inference approach among researchers, because of its simplicity. Currently, nearly all software packages have implemented this procedure.

Another method for model selection, referred to as penalized regression, maximizes a penalized likelihood function instead of the usual likelihood function. Penalized regression increases the bias and decreases the variance of the coefficient estimation by shrinking the regression coefficients. Such bias-variance tradeoff is usually beneficial, because the variance decreases faster than the increase in bias, which leads to a smaller mean square error (MSE) of the estimated model.

In general, the penalty function has a form of a sum of absolute regression coefficients raised to the γ-th power

$$\text{Penalty} = \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma},$$

where p is number of predictors, λ is the penalty multiplier that controls the trade-off between bias and variance (Frank and Friedman, 1993). The penalty function penalizes the regression coefficients whose values are far away from zero. Such shrinkage allows the less contributive parameters to be close or equal to zero. The parameter $\gamma > 0$ changes the structure of the penalty region, which also affects the properties of the regression method.

The Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani (1996), uses $\gamma = 1$ that allows one no only the shrinkage of estimators, but also the selection of variables. Further, many modifications of LASSO procedures were proposed, such as Adaptive LASSO (Zou, 2006) that uses a weighted penalty $\sum\limits_{j=1}^{p} w_j|\beta_j|$. The weights can be obtained through ordinary least squares regression (OLS) by defining $w_j = 1/|\tilde{\beta}_j|^{\delta}$, where $\tilde{\beta}_j$ are the OLS estimates for $j = 1, ..., p$, and $\delta > 0$ is often set equal to 1, but could also be estimated using cross-validation. Weighting allows an additional step of optimization and assures the selection and estimation consistency of the method.

Ridge regression is a special case of penalized regression where $\gamma = 2$, and it was developed to improve the prediction performance in the presence of multicollinearity (Hoerl and Kennard, 1970). Multicollinearity occurs when at least two predictors in the model are highly associated, such that their effects on the outcome variable cannot be distinguished. Ridge regression changes the associations between the variables, such that the MSE becomes smaller as the variance decreases, and allows more accurate estimation and interpretation of the effects.

Elastic net regularization (Zou and Hastie, 2005) is another modification of the LASSO method that adds the ridge penalty to it, which improves the performance of this method under multicollinearity. Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001) is the model selection and inference procedure that penalizes the likelihood function. The penalty function of SCAD is nonconvex, and the adaptive LASSO provides effect estimates for the original variables, while the LASSO procedure provides estimates for standardized variables.

While LASSO-type algorithms are useful for variable selection in high-dimensional data and for making predictions, it may have problems with post-selection inference, particularly accurate estimation of effects and confidence interval construction (Knight and Fu, 2000). Because of the bias-variance tradeoff implemented in the LASSO-type algorithms, it is possible to construct confidence intervals, but they would not have good properties.

Several LASSO related methods for confidence interval construction and hypothesis testing have been proposed since the introduction of LASSO for both high and low dimensional settings. Zhang and Zhang (2014) derived a low-dimensional projection estimator that uses residuals from sparse linear regression instead of a regular score vector to construct confidence intervals. Lockhart et al. (2014) and Taylor et al. (2014) proposed pathwise significance tests for predictor variables that use asymptotic or exact distributions of different pivotal quantities conditionally on the selected model. Bühlmann (2013) and Van de Geer et al. (2014) constructed confidence intervals by controlling and adjusting bias introduced by the regularization path of LASSO. Lee et al. (2016) derived a framework for post-selection inference in linear regression by conditioning on a union of polyhedrons. A similar approach was introduced by Tibshirani et al. (2016). Taylor and Tibshirani (2017) extended this methodology to generalized regression models. The above methods provide asymptotically valid confidence intervals under sparsity conditions for linear models with Gaussian additive noise and large effects; however, for logistic regression models in the finite and small sample settings, the results can be less stable.

In contrast to stepwise selection algorithms or penalized regressions that provides a single final model for inference, the model-averaging technique compromises across a set of candidate models by assigning weights to each model, and then averaging regression coefficient estimates across multiple models in order to capture the average effects of the variables (Barnard, 1963; Roberts, 1965). The method accounts for uncertainty produced by the single final model and outperforms single model approaches in terms of validity and accuracy of confidence intervals.

Leamer (1978) expanded the idea of model-averaging. This approach uses the posterior model probability as a function of a prespecified prior distribution to weight the posterior distributions of the quantity of interest under each of the considered models. The method is commonly referred to as Bayesian model-averaging (BMA).

Model-averaging can be time consuming due to the large number of possible models that need to be fitted. Moreover, the averaging of all possible models increases the risk of overfitting. To prevent overfitting, BMA accounts for uncertainty by averaging over a reduced set of models. To reduce the number of models, the Occam's window approach was proposed by Madigan and Raftery (1994). Although BMA is popular and its implementation was improved recently in terms of computational difficulties, there are still debates about what prior distribution should be used (Wasserman, 2000).

The model-averaging approach can also be applied to the frequentist framework. Frequentist model-averaging (FMA) is based upon the same idea as the Bayesian approach, but instead of using a prior distribution, a function of information criterion is used to average over a set of models. Given uninformative, uniformly distributed priors, the Bayesian posterior probability for model $m = 1,...,M$ can be approximated by the weighted function of an exponentiated BIC:

$$w_m = \frac{\exp(-BIC_m/2)}{\sum\limits_{i=1}^{M} \exp(-BIC_i/2)}.$$

Buckland et al. (1997) suggested replacing BIC with AIC as a criterion for estimation of the weight for each model. Burnham and Anderson (2002) modified the methods by replacing the information criterion by the differences in AIC with respect to the AIC of the best candidate model. This rescaling does not change the order of the models, but it facilitates subsequent calculations and comparison of the models. Hjort and Claeskens

(2003) and Claeskens and Hjort (2008) developed the focused information criterion (FIC), obtained from the estimation of MSE as a focus estimand, while AIC or BIC are based on the penalized likelihood function. The FIC method provides a ranked list of models for a prespecified parameter of interest, while AIC and BIC provide a ranked list of candidate models without considering each parameter separately. Thus, if the goal of the study is estimation for a specific variable, the FIC can be preferable over other information criteria.

Apart from the aforementioned criteria, there are many other criteria and algorithms that can be used to estimate weights. For example, Austin (2008) suggested bootstrapping the original data and applying of backward selection on each of the bootstrapped samples, the coefficients of eliminated variables are set to zero, and a point estimates are obtained by averaging over all samples. The confidence interval around the estimated effect is constructed by quantile method, that is defined by the quantiles closest to a cumulative probability of $\alpha/2$ and $1-\alpha/2$. This method can be seen as averaging over multiple models, with weights based on how often the model appears in the bootstrap procedure. The weights are seen as the estimated posterior probability.

The frequentist model-averaging faces the problem of a large number of models that need to be considered. There exist at least two frequentist approaches to reduce the number of models. One approach suggests that one construct the models only from the variables that were selected by some preceding variable selection method and ignore the eliminated variables. The other approach is to construct all possible models from the eliminated variables and add to each model the selected candidate variables (Hansen, 2007). The candidate model set based on the second approach includes the full model and has a higher risk of overfitting than the first approach. At the same time, the risk of information loss and invalid inference should be smaller for the second approach.

Methods for selecting the candidate variables can be as simple as stepwise selection or more advanced and time consuming such as a bootstrap-based inclusion fraction. The bootstrap-based approach involves selection of variables over a large number of bootstrapped samples and retains in the model only variables whose inclusion frequency exceeds some prespecified fraction (Burnham and Anderson, 2002). For example, in the analysis of patients admitted to hospital with a heart attack, Austin and Tu (2004b) showed that out of 30 predictors for mortality, eight variables that appeared in more than 60% of bootstrap samples formed a parsimonious model with great predictive ability. Although this method was used to build models for prediction, the method may also be used to estimate effects of variables.

After candidate models are selected and their weights are estimated by the bootstrap or an information criterion, confidence intervals can be constructed. Buckland et al. (1997) suggested using the bootstrap method to estimate standard errors and construct confidence intervals. Burnham and Anderson (2002) proposed an unconditional Wald-type confidence interval that uses an adjusted standard error estimator. Turek and Fletcher (2012) developed model-averaged tail area (MATA) intervals to improve model-averaged Wald intervals, and compared the performance of these intervals under different information criteria. Fletcher and Turek (2012) also proposed a method based on the profile likelihood function. Yu et al. (2014) transformed MATA with the inverse of the cumulative distribution function of standard normal and derived a method that can be applied to general parametric models, and developed the asymptotic version of transformed MATA intervals.

Model-averaging methods based on AIC and BIC are widely studied. Although model-averaging usually performs better than regular stepwise methods, it also has problems with coverage probability and coefficient estimation. There is no simple answer as to what procedure should be used, even though it is well-known that stepwise procedures usually provide confidence intervals with undercoverage. For example, if a data set contains a large number of noise variables, then penalized regressions outperform stepwise methods in terms of variable selection (Derksen and Keselman, 1992). Wang et al. (2004) and Genell et al. (2010) found that BMA has a better probability of selecting the true model than stepwise regression.

Greenland et al. (2016) compared the stepwise approaches with different criteria and Bayesian penalized regression. They concluded that standard errors based on stepwise methodology should be adjusted or corrected, otherwise penalized regression methods will outperform stepwise selection procedures in terms of construction of confidence intervals. The adjustments can be made by using bootstrap or cross-validation approaches.

Pfeiffer et al. (2017) tested the ability of choosing the true model and some inference properties of different penalized approaches on linear and logistic regressions. To test the variable selection ability, they defined the false positive (FP) rate as the percentage of times when a method estimated $\hat{\beta}_j \neq 0$ for noise variables, and false negative (FN) rate as a percentage of times when a method estimated $\hat{\beta}_j = 0$ for outcome associated predictors. The FP and FN rates then were averaged over all zero and non-zero coefficients of the β-vector, respectively. They found that for logistic regression, the LASSO approach demonstrated an increase in FP and decrease in FN with an increase of sample size and the magnitude of non-zero coefficients. For the SCAD approach, the association of FP and FN with sample size and coefficient magnitude was opposite to the LASSO method. For all settings, the coverage of the confidence intervals for irrelevant variables was close to 100% for SCAD and close to 95% for LASSO. However, the coverage of the methods for important variables was mostly far below the nominal level, that was reached only for large sample sizes and large true effects.

Each of the methods mentioned above might be useful for prediction, but all of them have a problem with model selection, especially when sample sizes are not large. Post-selection inference based on these methods is also very problematic, because the coverage of confidence intervals for non-zero predictors usually does not reach the prespecified nominal level.

The general goals of this thesis are 1) to develop algorithms for reducing the subset of models for model-averaging, and 2) to develop model-averaged confidence intervals based on the score test. We also consider the pros and cons of the replacement of Wald standard errors in model-averaged tail area confidence intervals by standard errors obtained from profile-likelihood and score confidence intervals. All methods are developed in the context of logistic regression. This is because logistic regression analysis is frequently used in epidemiological and biomedical research (Hosmer et al., 2013 and Rothman et al., 2008), recognising that approaches may also be applicable to other generalized linear models.

The specific objectives are:

1. To review common approaches for model selection, with the goal of identifying the true model;
2. To summarize procedures for post-selection inference;
3. To provide a procedures for selection of a subset of models for model-averaging;
4. To develop model-averaging score function based procedure that produces valid inference for each predictor of interest;
5. To evaluate empirically the performance of the model-averaging score-based method as compared with commonly used approaches.

This thesis is structured as follows. Chapter 2 reviews the literature on model selection and inference methods. Chapter 3 describes the proposed method for candidate model set selection. In Chapter 4, we present the proposed confidence interval construction algorithm and the improvement of the existing Wald-based model-averaging tail area confidence interval construction method. A simulation study is reported in Chapter 5. Empirical performance of the methods was assessed by changing sample size, the number of variables, correlation between variables, and probabily of outcome. Chapter 6 presents an R package that implements the recommended methods. For illustrative purposes, the data from the Baystate Medical Center Study was analysed in Chapter 7. Finally, Chapter 8 closes with a summary of the main results, discussion of the strengths and limitations of the proposed methods, and the proposed directions for future research.

Chapter 2

OVERVIEW OF VARIABLE SELECTION AND CONFIDENCE INTERVAL CONSTRUCTION METHODS

2.1 Automated model selection

Typical data analysis begins with a definition of the full model. While one may consider all collected variables as a “full” model, we define it as the model that contains all relevant variables and all potential confounders, that were selected by prior knowledge. Although the full model is valid, the confidence intervals can be too wide to be meaningful.

Thus, fitting the full model is not the best way to analyse the data, because it may not always be possible to fit the model and the results may have a lack of precision. This problem becomes more severe as the ratio of sample size to the number of predictors decreases. Moreover, if the number of predictors is large, the results of fitting the full model may be difficult to interpret due to its complexity.

Fitting a model with a large number of variables can decrease bias of point estimates but increases the variance, such that the confidence intervals become unnecessarily wide. Model selection algorithms were developed to balance the bias-variance tradeoff and obtain a smaller model that still has good estimation properties and shorter, valid confidence intervals for estimates.

2.1.1 Stepwise selection

Efroymson (1960) is among the early studies that proposed a stepwise algorithm for a linear regression model for a continuous outcome, using the partial F-test value as a criterion to compare multiple models. This strategy has been used for other outcomes as well (Harrell, 2015). Widely used algorithms include:

- Forward stepwise selection, which begins with a model with no predictors, followed by adding the most significant variable from the pool of variables, and stops when no variables meet a prespecified criterion;
- Backward stepwise selection, which starts with all variables in the model, followed by eliminating a least significant variable until no more variables need to be excluded based on the prespecified criterion;
- Bidirectional or stepwise selection approach, which combines the previous methods. It is analogous to forward selection, but each step algorithm is checking if it is possible to delete one of the selected variables.

Intuitively, it may seem that these algorithms would give the same results; however, they do not always agree (Wiegand, 2010). The agreement between these methods is very sensitive to sample size, number of predictors, the criteria for inclusion and exclusion of predictors, and correlation among the predictors. The stepwise algorithms select the same model more frequently as the sample size increases or correlation among the covariates decreases. However, even in the case of agreement, the analysis must proceed with caution, since this does not guarantee that the selected model is the correct one.

2.1.1.1 Inclusion and exclusion criteria

The literature on criteria for inclusion and exclusion in stepwise selection methods is diverse. In the context of linear regression models, Kennedy and Bancroft (1971) suggested to use significance levels $\alpha_{in} = 0.15$ and $\alpha_{out} = 0.1$ as entry and deleting criteria, respectively. Flack and Chang (1987) and Rawlings et al. (1988) suggested $\alpha_{in} = \alpha_{out} = 0.15$. These suggestions are consistent with a earlier study of Bendel and Afifi (1977), that recommended the use of a significance level between 0.15 and 0.25 for both criteria and showed that the best results for forward selection are obtained for $\alpha_{in} = 0.15$.

Aitkin (1974) pointed out that the stepwise procedure tests one variable at a time, which leads to the conclusion that the increase in inclusion or exclusion criteria will affect the Maximum Family-Wise Error Rate (MFWER), which is the probability of making at least one Type I error during the procedure. Such overall Type I error is usually unknown and greater than the Type I error of an individual test; thus usage of α_{in} and α_{out} much smaller than 0.15 to get MFWER < 0.05 was recommended. For backward selection, Aitkin (1974) proposed the use of $0.01 \leq \alpha_{out} \leq 0.10$ if a researcher's main interest is to exclude all irrelevant variables, and $0.25 \leq \alpha_{out} \leq 0.50$ if a researcher does not want to lose important variables. Instead of F-tests, AIC or BIC can be used to create stopping rules for the algorithm, but these penalty terms are still strongly related to critical values. For example, the backward procedure with AIC penalty is equivalent to $\alpha_{out} \approx 0.157$ (Sauerbrei, 1999).

To address performance of stepwise approaches and control the MFWER in logistic regression, Wang et al. (2007) and Lee and Koval (1997) found that the best choices of α vary with the number of predictors. They recommended the use of $0.2 \leq \alpha_{out} \leq 0.4$ and $0.15 \leq \alpha_{in} \leq 0.2$ for backward and forward stepwise selection methods, respectively. When the number of predictors is $5 \leq p \leq 25$, both suggested the use of $\alpha = p/100$. In a study of the performance of stepwise algorithms, Wiegand (2010) compared three inclusion/exclusion criteria: 0.50/0.05 that are default criteria for linear regression in SAS PROC REG, 0.15/0.15 criteria that were recommended by Kennedy and Bancroft (1971) and Bendel and Afifi (1977), and 0.05/0.05 that are the default settings for logistic regression in SAS in PROC LOGISTIC. Wiegand (2010) found that $\alpha_{in} = \alpha_{out} = 0.15$ have the best performance in terms of agreement on the parsimonious model, while $\alpha_{in} = \alpha_{out} = 0.05$ demonstrated the worst performance. Despite the fact that 0.15 appears in the majority of studies as the most favorable criterion, Steyerberg et al. (1999) pointed out that in small samples a less conservative criterion $\alpha_{out} = 0.5$, can be more reasonable.

Unfortunately, even $\alpha_{in} = \alpha_{out} = 0.15$ cannot guarantee that the model and inference will be valid, because stepwise selection provides a single final model and does not account for uncertainty. In general, there is still no agreement on cut-off criteria for input and output of variables. Moreover, different statistical software programs may use different criteria as defaults, and often do not rush to change them in order to comply with new findings.

2.1.1.2 The number of events per variable

Logistic regression is more sensitive to sample size than linear regression. The number of events in the rarest outcome group relative to the number of variables (EPV) was identified as the key factor that affects the performance of logistic regression models. Peduzzi et al. (1996) examined the effect of EPV on the reliability of logistic regression estimates and suggested the minimal 10 EPV rule that agrees with the recommendation of Harrell et al. (1985) to use EPV > 10. Vittinghoff and McCulloch (2007) examined a larger set of scenarios on multivariable models and concluded that the "rule of 10" was too conservative for most cases, and that even five to nine EPV can be sufficient to obtain appropriate confidence interval coverage and relatively small bias. However, they pointed out that the interpretation of the effect estimates based on five events per variable should proceed with caution, especially the interpretation of the significance of the effects. Moreover, use of EPV < 10 might be a bad strategy, if distribution of considered variables is skewed. Vittinghoff and McCulloch (2007) considered such more realistic scenarios with skewed continuous variables and unbalanced distributions.

Feinstein (1996) and Agresti (2007) suggested that 20 EPV is safer; however, Courvoisier et al. (2011) showed that even if EPV equals 20 or 25 it still might not be enough to get good logistic regression performance. They extended the Vittinghoff and McCulloch (2007) study by evaluating the effect of number of predictors, correlation between them, magnitude of their effects and the proportion of noise variables on the association between EPV and logistic regression performance. According to this study, an increase in the number of predictors, as well as in the correlation between them and in the magnitude of their effects, has a negative impact on the efficiency of logistic regression in terms of statistical power, relative bias, and convergence.

It may seem that the results of Peduzzi et al. (1996), Vittinghoff and McCulloch (2007) and Courvoisier et al. (2011) do not agree. The apparent inconsistent results are due largely to the fact that each subsequent study increased the number of factors that could affect logistic regression performance. It should be taken into account that the results of Vittinghoff and McCulloch (2007) do not deny the results of Peduzzi et al. (1996), but demonstrate that under certain conditions the choice may be in favor of a less conservative EPV. Therefore, it can be seen that as analysed data become more complex and realistic, there is more evidence we have that even a EPV that exceeds 10 may not always be sufficient for good logistic regression performance.

These three studies tested the effect of EPV on the performance of logistic regression for a prespecified model, but did not check how EPV affects logistic regression after a stepwise selection procedure. Steyerberg et al. (1999) analysed how the backward stepwise selection algorithm affects logistic regression with respect to changes in EPV. Their study demonstrated that in small samples, conventional backward selection ($\alpha_{out} = 0.05$) substantially increases the bias of estimated coefficients even for EPV=40, and that it can be used only as an exploratory tool. Steyerberg et al. (2000) pointed out that acceptable predictive performance of logistic regression can be achieved if EPV exceeds 50. Since the lower bound for sufficient EPV is unique, in this thesis we defined sample sizes, such that in each simulation block we can also analyse the effect of EPV on performance of point estimates and confidence intervals.

2.1.2 Concerns on stepwise selection

Despite the problems associated with them, stepwise selection procedures remain popular in practice. Derksen and Keselman (1992) demonstrated that stepwise selection methods provide a large number of noise variables that are not related to the outcome variable. They tested the performance of stepwise selection methods under different conditions, such as sample size, number of considered variables, and correlation between them. It was found that over all conditions, less than 50% of all important variables were included in the final model, and that the probability of choosing correct predictors reduces with the number of considered variables, while the sample size has little effect. Such poor performance was confirmed for logistic regression by Bursac et al. (2008).

Factors that affect performance of point estimates and confidence intervals produced by stepwise procedures include sample size, number of considered variables, and correlation. Austin and Tu (2004a) demonstrated this by evaluating the risk factors of acute myocardial infarction mortality. A total of 1,000 bootstrapped samples were generated from a large dataset of 4,911 patients that contained 29 preselected predictors and analysed the final models suggested by three stepwise selection methods - backward, forward, and bidirectional - were noted. Bootstrapping was used to imitate random sampling fluctuations, and it showed that even small degrees of random variation in data may highly affect the model selection process and risk factors that are included in the final model. For example, backward selection identified 940 unique subsets of risk factors of mortality, and no model appeared more than four times. Such variation of the final models means that two researchers that use a similar stepwise method to analyse two slightly different samples from the same population are likely to identify a different set of important risk factors. This also means that it is difficult to reproduce the results of any study that blindly uses a stepwise approach as a model selection method. Thus, it is recommended to use more advanced methods, such as the bootstrap, coupled with regular stepwise selection approaches, to get better understanding of the associations between an outcome and predictors, and of the strength of the evidence that a selected final model is reliable.

Overall, the final model obtained from a stepwise method is very sensitive to any changes in data and relationship anomg variables even in large samples (Austin and Tu, 2004a). Steyerberg et al. (1999) demonstrated that in small samples, stepwise selection may have substantial bias, especially if it uses the default 0.05 threshold. This makes stepwise approaches unreliable as data analysis tools, but they still can be used for exploratory analysis.

2.1.3 Bootstrapped stepwise selection and inference

Bootstrap is a well-known procedure that can improve the accuracy of point estimates. In regression modeling settings, the bootstrapinvolves resampling the original data with replacement many times and applying a prespecified regression method on each bootstrap sample. The estimates from multiple runs then are averaged to get less biased point estimates. Sauerbrei and Schumacher (1992) suggested using the bootstrap as a tool for variable selection. The frequency of a variable appearing among the results from bootstrapped samples was considered as a criterion for the importance of the variable. This criteria was set at 30% and 70% for two different case-studies, atopy and glioma studies. Austin and Tu (2004b) suggested the use of at least 60% appearance as an inclusion criterion.

Bootstrap can be combined with a stepwise procedure to get better inference. To improve the performance of stepwise procedures, Austin (2008) combined the bootstrap method with the backward selection procedure. The effects of eliminated variables are replaced by zero in each bootstrapped dataset and average over all samples. This procedure is referred to as zero-corrected backward stepwise selection and can be considered as an approximation to Bayesian model-averaging, that is described in section 2.3.3. If the real effect of some variable is small, a selection method will eliminate it more frequently, which increases the number of zeroes in the final set and shifts the final effect estimator towards zero.

The zero-corrected backward stepwise selection method was compared to a simple backward selection method and different variations of bootstrapped methods, such as (i) the conditional bootstrap model selection method that considers effects of eliminated variables as missing and calculate averaged coefficients based on a non-missing set and (ii) the naive bootstrap method that uses the original dataset to estimate a parsimonious model by applying backward selection and then estimates only this model in each bootstrapped dataset. To obtain confidence intervals for regression coefficients, the percentile method that suggests the use of specified percentiles from bootstrap estimates was used. It was shown the zero-corrected bootstrap method performs better than its competitors in terms of confidence interval coverage and smaller bias, but still cannot reach a nominal level of coverage. According to the simulations performed, the coverage accuracy decreases with the effect size. The possible explanation for a such result is a large proportion of zeros in the final set, which leads to the poor performance of the percentile confidence interval method (Austin, 2008).

2.1.4 Common approaches to confidence interval estimation

There are three conventional approaches for asymptotic confidence interval estimation:

1. Wald confidence interval:

   The Wald-based confidence interval is the well-known and commonly used type of inference (Wald, 1943). The Wald $(1-\alpha)\%$ confidence interval for a regression parameter β_j is given by

   $$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \times \widehat{\mathrm{se}}(\hat{\beta}_j), \qquad (2.1)$$

   where $z_{1-\frac{\alpha}{2}}$ is the $1-\alpha/2$ quantile of a standard normal distribution.

2. Profile-likelihood confidence interval:

   The profile-likelihood confidence interval was proposed by Wilks (1938) as a likelihood ratio based confidence interval derived from the asymptotic χ^2 distribution of the generalized likelihood ratio test (Venzon and Moolgavkar, 1988). The profile-likelihood confidence interval for β_j is defined by

$$\left\{\theta : 2[\ell(\hat{\beta}) - \max_{\gamma} \ell(\theta,\gamma)] \leq q_1(1-\alpha)\right\}, \tag{2.2}$$

where $\ell(\theta,\gamma)$ is a log-likelihood function of the parameter of interest and is maximized over the other coefficients $\gamma = \{\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p\}$, $\hat{\beta}$ is a vector of the maximum likelihood estimates, and $q_1(1-\alpha)$ is the $(1-\alpha)$th quantile of the χ^2 distribution with 1 degree of freedom. The θ has two solutions - lower and upper confidence limits for parameter of interest β_j.

3. Score confidence interval:

   The score confidence interval is based on the score test proposed by Rao (1948). It is similar to the profile-likelihood confidence interval, but optimization is done over the score function instead of likelihood function. Let us define the vector of score function as

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \left[\frac{\partial \ell(\beta)}{\partial \beta_1}, \frac{\partial \ell(\beta)}{\partial \beta_2}, \ldots, \frac{\partial \ell(\beta)}{\partial \beta_p}\right],$$

   and expected Fisher information matrix I, with j,k element given by

$$I_{jk} = -\mathrm{E}\left[\frac{\partial^2 \ell(\beta)}{\partial \beta_j \partial \beta_k}\right].$$

   In this case the score confidence interval for β_j is defined by

$$\left\{\theta : U(\theta,\gamma) I^{-1}(\theta,\gamma) U^T(\theta,\gamma) \leq q_1(1-\alpha)\right\}, \tag{2.3}$$

   where solution for this equation θ is a confidence limits of the parameter of interest β_j. Basically, for each value of θ we have to refit the model, recalculate the vector of scores such that

$$\frac{\partial \ell(\theta,\gamma)}{\partial \gamma} = 0$$

   and recalculate the Fisher information matrix.

Although the Wald method for intervals is computationally easy with a closed form, it assumes that the distribution of the estimator follows a normal distribution and produces a symmetric confidence interval around a point estimate, while profile-likelihood and score intervals are not subject to this restriction. In general, confidence intervals that do not force symmetry perform better, but may be more involved with respect to computation.

These three methods are asymptotically equivalent; however, in finite samples, the score confidence interval is usually preferable over Wald and profile-likelihood intervals in terms of coverage and length of the interval (Engle, 1984; Cox and Hinkley, 1979).

2.2 Penalized regression models

Penalized regression models are a large family of model selection and inference approaches that use penalized versions of the log-likelihood function (PLL). The penalized approach was developed to control the stability of the model. Compared to traditional stepwise variable selection methods, which are very sensitive to small changes in the data set (Sauerbrei et al., 2015), penalized regression methods are less sensitive to perturbations of the data, resulting in more stable inferences. Each member of this family has different features that we briefly summarize here:

- Ridge regression was developed to alleviate multicollinearity among regression predictor variables in a model, but it does not have the model selection property (Hoerl and Kennard, 1970). Its penalized log-likelihood function is
$$PLL(\beta)_{ridge} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1+e^{x_i\beta}) \right] - \lambda \sum_{j=1}^{p} \beta_j^2 ,$$
where λ is the shrinkage parameter that can be estimated by cross-validation, the AIC or BIC, and $\sum_{j=1}^{p} \beta_j^2$ is the penalty function. Bias-variance tradeoff is a basis of regression regularization methods. If two independent variables are highly correlated, the variance of the regression parameter estimates can be large. Ridge regression decreases MSE by significantly lowering the variance at the cost of a small increase in

bias. The bias for non-zero coefficients can vary between zero and λ and increases with the magnitude of the coefficient.

- The LASSO approach was introduced by Tibshirani (1996) and, unlike ridge regression, it shrinks some of the regression coefficients to zero by using the sum of absolute values of regression coefficients as a penalty factor instead of squared coefficients:

  $$PLL(\beta)_{LASSO} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1 + e^{x_i \beta}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| .$$

  LASSO produces sparse solutions in the case of high-dimensional data. It also can be applied to low-dimensional data.

Ridge and LASSO methods belong to a family of penalized regression methods, called Bridge regression, introduced by Frank and Friedman (1993), as given by

$$PLL(\beta)_{\ell_\gamma} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1 + e^{x_i \beta}) \right] - \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma}, \tag{2.4}$$

where $\gamma \geq 0$. By setting $\gamma = 1$ or $\gamma = 2$ we can obtain LASSO or ridge methods. Although LASSO and ridge methods are usually considered frequentist approaches, their results correspond to Bayesian estimators with Laplace or normal priors, respectively (Park and Casella, 2008).

- Adaptive LASSO (Zou, 2006) was derived to reduce bias by using a weighted penalty approach,

  $$PLL(\beta)_{adapt} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1 + e^{x_i \beta}) \right] - \lambda \sum_{j=1}^{p} w_j |\beta_j| ,$$

  where weight w_j is a data-dependent function. There are many weighting strategies that can be used, for example, $w_j = |\tilde{\beta}_j|^{-1}$, where initial estimates $\tilde{\beta}_j$ are usually obtained using ridge regression. With such weights the Adaptive LASSO penalizes more those coefficients with lower initial estimates, which reduces the estimation bias of the LASSO. Bias is not the only problem of the LASSO approach. If a data set contains a group of highly correlated variables, LASSO will select only one of the variables from a group and ignore the effects of the others, which may lead to loss of important information.

- Elastic Net (EN) was developed by Zou and Hastie (2005) to overcome the problem created by multicollinearity by combining the ridge regression penalty factor with the LASSO penalty factor:

  $$PLL(\beta)_{EN} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1 + e^{x_i \beta}) \right] - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \beta_j^2 \,.$$

- Nonnegative Garrote (NNG) introduced by Breiman (1995) is a shrinkage method that shrinks OLS estimators by minimizing

  $$PLL(\beta)_{NNG} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1 + e^{x_i \beta}) \right] - \lambda \sum_{j=1}^{p} d_j(\beta, \lambda) \,,$$

  where $d_j(\beta, \lambda) \geq 0$ for all $j = 1, \ldots, p$. It was shown that NNG outperforms subset and ridge regressions in terms of predictive accuracy if

  $$d_j(\beta, \lambda) = \max\{0, 1 - \lambda / \hat{\beta}_{j,ols}^2\} = (1 - \lambda / \hat{\beta}_{j,ols}^2)_+ \,.$$

  This means that NNG needs OLS estimates, which may lead to poor performance for small sample size and imposes an additional limitation on the dimensionality of the data ($n > p$). Later, Yuan and Lin (2007) showed that LASSO, ridge regression or EN can also be used in d_j estimation, and that if the tuning parameter, λ, is appropriately chosen then NNG is consistent in terms of coefficient estimation and variable selection.

- Smoothly Clipped Absolute Deviation (SCAD) is a member of the penalized regression family that can provide a sparse set of solutions as well as LASSO (Fan and Li, 2001). While LASSO uses a sum of the absolute value of the regression coefficients as the penalty function, SCAD uses a quadratic spline function with two knots, which makes the penalty nonconvex,

$$PLL(\beta)_{SCAD} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1+e^{x_i\beta})\right] - \lambda \sum_{j=1}^{p} \int_0^{|\beta_j|} \min\left\{1, \frac{(a-x/\lambda)_+}{a-1}\right\} dx,$$

where a can be chosen by using convexity diagnostics. It was shown that SCAD outperforms LASSO in selecting significant variables when the noise-to-signal ratio is not large, but performs poorly when the noise-to-signal ratio increases and the sample size is small. As in the LASSO case, λ can be estimated using different approaches. Zhang et al. (2010) studied properties of the SCAD method and found that a BIC-type criterion to estimate the tuning parameter identifies the true model with probability tending to 1, while an AIC-type selector is asymptotically efficient.

- Minimax concave penalty (MCP) is another penalized method with nonconvex penalties (Zhang, 2010):

$$PLL(\beta)_{MCP} = \sum_{i=1}^{n} \left[y_i x_i \beta - \log(1+e^{x_i\beta})\right] - \lambda \sum_{j=1}^{p} \int_0^{|\beta_j|} \left(1 - \frac{x}{a\lambda}\right)_+ dx.$$

It is similar to SCAD; however, MCP continuously relaxes the penalty rate down to zero as the absolute value of the coefficient of interest increases, while SCAD remains flat for a while before decreasing.

2.3 Uncertainty and post-selection inference

2.3.1 Uncertainty in model selection

Common statistical inference procedures assume the existence of a true model. In practice, since model uncertainty is associated with the selected model, the subsequent CIs usually have coverage lower than the nominal level.

Leeb and Pötscher (2005) demonstrated that the sampling distributions of parameter estimates after model selection are usually unknown. Leeb and Pötscher (2006) proved that the conditional distribution a of post-selection estimator cannot be estimated with reasonable accuracy. It was shown that even asymptotically it can be non-normal and very complex. Berk et al. (2010) came to similar conclusions regarding the estimation of the conditional distribution and showed that ignoring the selection step may lead to biased regression estimates and overoptimistic, invalid confidence intervals. To improve the undercoverage caused by the effect of model selection, Berk et al. (2010) suggested randomly splitting the data into two independent data sets, and using one as a training sample and the other as a test sample. However, splitting the data reduces the reliability and accuracy of the analysis, even if it is possible. Kabaila (2005) described a Monte Carlo method to calculate the coverage probability of the naive confidence interval and showed that coverage probability was far below the nominal coverage level if variables were selected using minimization of AIC or BIC.

2.3.2 Post-selection inference

Berk et al. (2013) proposed a procedure that corrects for model selection, but calculates confidence intervals for a non-standard coverage target, which is not a fixed parameter of the data-generating model, but depends on the selected model. This procedure simultaneously uses all possible submodels to produce valid post-selection inference and is referred to as Post-Selection Inference (PoSI). This procedure can guarantee the prespecified minimal coverage probability for the variables selected in a data-driven way.

Consider a logistic regression model given by

$$P(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}, \tag{2.5}$$

where X_i is a vector in $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^{p\times 1}$. Define $B = \{B_1, \ldots, B_M\}$ as the set of all possible models, and suppose that a model selection procedure selected model $B_m \in B$.

The PoSI procedure suggests that inference should be made about the projection of $X\beta$ onto a submatrix of X used in model B_m, but not about the true parameter β. The PoSI algorithm produces a constant K such that the constructed confidence interval for each selected parameter will reach at least the nominal coverage level

$$P(\beta_{jm} \in CI_{jm}(K), \quad \text{for all} \quad j \in B_m) \geq 1 - \alpha,$$

where $K = K(X, B, \alpha, n - p)$ is the value that should be universally valid for post-selection inference for any selected model. It means that PoSI is accounting for uncertainty by calculating a multiplier for a standard error that allows a confidence intervals to achieve at least nominal coverage for any selected model. However, such a multiplier can be very high, such that it will lead to too wide confidence interval. Note that the coverage target proposed in Berk et al. (2013) is dependent on the outcome of the selection procedure and therefore it is random.

Berk et al. (2013) also proposed a simplification of PoSI focusing on a single independent variable, while other variables in the model act as confounders to adjust the effect of the primary parameter. Leeb et al. (2015) evaluated performance of the PoSI procedure and concluded that PoSI can provide adequate coverage; however, it is usually too wide to be meaningful. These methods have not been used widely because of their complexity and because they do not provide a solution for the traditional problem where the coverage target is a parameter of the data-generating model.

Bachoc et al. (2017) generalized the PoSI intervals and evaluated the prediction performance of proposed procedures. The performance of PoSI intervals for different K was compared for AIC, BIC, LASSO, SCAD, and minimax concave penalty (MCP) (Zhang, 2010) model selection procedures. The results for AIC and BIC were similar to those in Leeb et al. (2015), while LASSO, SCAD, and MCP outperformed naive intervals under all settings that were considered, but showed poor performance for small sample size and did not reach the nominal level.

2.3.3 Bayesian model-averaging

Bayesian model-averaging is an application of Bayesian theory to model selection and inference under model uncertainty. Following Hoeting et al. (1999), we consider the logistic regression model:

$$\operatorname{logit}\big[P(Y=1|X)\big] = X\beta\,, \tag{2.6}$$

where $X \in \mathbb{R}^{n\times p}$ is the non-random matrix of predictors, $\beta \in \mathbb{R}^{p\times 1}$ is the parameter vector. The number of models that should be considered is $M = 2^p$. The posterior distribution of the quantity of interest, β_j, is given by

$$p(\beta_j|D) = \sum_{m=1}^{M} p(\beta_j|D,B_m)p(B_m|D)\,,$$

where D represents the observed data and B_m denotes model indicator. The posterior probability of each model B_m is

$$p(B_m|D) = \frac{p(D|B_m)p(B_m)}{\sum_{i=1}^{M} p(D|B_i)p(B_i)}\,,$$

where

$$p(D|B_m) = \int p(D|\beta_m,B_m)p(\beta_m|B_m)d\beta_m\,,$$

where β_m denotes coefficient parameter in model m. BMA uses the posterior probability of each model to construct a weighted estimate for the coefficient:

$$\hat{\beta}_j^{BMA} = \sum_{m=1}^{M} p(B_m|D)\hat{\beta}_{jm}\,.$$

In addition to the computational difficulties of integrals and sums involved in the implementation of BMA, the specification of the prior distribution of models can be very challenging. Specification of an incorrect prior makes the analysis meaningless, because the results will be incorrect.

Problems may also arise with fitting a large number of predictors in a model. For example, if the number of potential confounders is 15, the number of possible models is $2^{15} = 32,768$. In order to overcome this problem, the Occam's window method was proposed by Madigan and Raftery (1994). This method is based on two principles. First, if a model probability is smaller than the most likely model, then this model should not be used in the averaging process. The second principle is parsimony; if a nested model performs better than the large model, then the large model should no longer be considered in model-averaging. Formally, define two sets:

$$Q = \left\{ B_k : \frac{\max_i\{p(B_i|D)\}}{p(B_k|D)} \leq C_1 \right\}$$

and

$$T = \left\{ B_k : \frac{p(B_l|D)}{p(B_k|D)} \geq C_2, \quad for \quad B_k \supset B_l \in Q \right\},$$

where the bound C_1 defines the number of models considered, and should be chosen by the data analyst. According to the two principles, the subset V that contains all the models of Q that are not in T should be considered in the BMA. Madigan and Raftery (1994) proposed the algorithm to find this subset. The algorithm compares nested models by the log posterior odds

$$\ln[p(B_0|D)/p(B_1|D)],$$

where B_0 is the smaller model. If the log posterior odds is positive, the algorithm rejects the largest model, but if it is large and negative then the algorithm rejects the smaller model with all its submodels. If the log posterior odds falls into Occam's window, then the evidence for rejecting the smallest model is not strong enough and neither model is rejected. Raftery et al. (1996) showed that using $C_1 = 20$ that emulates the popular 0.05 significance level based on p-values and $C_2 = 1/20$ as the bounds of Occam's window can improve the performance of the algorithm.

Construction of the subset V still can be very difficult. Volinsky et al. (1997) suggested the use of the "leaps-and-bounds" method proposed by Furnival and Wilson (1974), that approximates the likelihood ratio test statistic and allows one to quickly identify the models that should be included into the subset V. This method is not good for estimation, but works well for model comparison. For generalized linear models, Raftery (1996) proposed the use of a single step of the Newton-Raphson algorithm for the approximation.

The standard error of a model-averaged posterior distribution from BMA is given by

$$\text{se}(\hat{\beta}_j^{BMA}) = \sqrt{\sum_{B_m \in V} \left\{ [\text{Var}(\beta_{jm}|D,B_m) + \hat{\beta}_{jm}^2] p(B_m|D) \right\} - (\hat{\beta}_j^{BMA})^2}.$$

where V denotes the subset of models considered by the BMA procedure. Such a standard error can be used to construct a Wald based credible interval of a regression coefficient. However, such intervals cannot guarantee the nominal coverage level.

2.3.4 Frequentists model-averaging

While BMA uses posterior probabilities as weights for averaging the models, the FMA procedure uses information criteria obtained from all models to weight each model. Usually, there is no need to estimate all possible models because researchers have some set of variables that they must include and a set of variables that are under consideration, but for simplicity assume that all variables are being investigated. Selection of the models will be discussed later. Current methods of model-averaging were usually derived in a context of a prediction problem. We will describe their application for regression coefficients.

The FMA estimator of the regression coefficients can be written as

$$\hat{\beta}_j = \sum_{m=1}^{M} w_m \hat{\beta}_{jm}$$

with $\sum_{m=1}^{M} w_m = 1$, where $w_m \geq 0$ is the weight associated with $\hat{\beta}_{jm}$.

To estimate such weights, Buckland et al. (1997) proposed the use of a function of an information criterion (IC):

$$w_m = \frac{\exp(-IC_m/2)}{\sum_{i=1}^{M} \exp(-IC_i/2)}, \quad m = 1, \ldots, M. \tag{2.7}$$

This function guarantees that if penalized log-likelihood functions of two models are equal, those models are given the same weights. For numerical stability, when evaluating the exponential function, Burnham and Anderson (2002) computed the index of relative plausibility of each model as $\Delta AIC_m = AIC_m - \min AIC$ to use in Equation 2.7 instead of IC, where

$$AIC_m = -2\ln L + 2k_m, \quad m = 1,\ldots,M,$$

and k_m is the number of non-fixed coefficients in model m including the intercept. This function is also referred to as smooth AIC, and it assures that the model with the lowest AIC will get the highest weight. However, AIC is not the only criterion that can be used in model-averaging; another common criterion that can be found in the literature is the corrected AIC proposed by Hurvich and Tsai (1989),

$$AICc_m = -2\ln L_m + 2k_m + \frac{2k_m(k_m+1)}{n-k_m-1}, \quad m = 1,\ldots,M$$

that was derived for small sample sizes. The BIC,

$$BIC_m = -2\ln L_m + 2k_m \ln(n), \quad m = 1,\ldots,M$$

provides a more severe penalty for model complexity.

Another option is to use the FIC proposed by Hjort and Claeskens (2003) to estimate parameters of direct interest with good precision. Consider a logistic regression model,

$$P(Y_i = 1|X_i, Z_i) = p_i = \frac{\exp(X_i\beta + Z_i\gamma)}{1+\exp(X_i\beta + Z_i\gamma)}, \tag{2.8}$$

where X_i is a vector in $X \in \mathrm{R}^{n\times p}$ that is a set of variables of direct interest and Z_i is a vector in $Z \in \mathrm{R}^{n\times q}$ that is a set of covariates that may be of indirect interest for objects $i = 1,...,n$. The Fisher Information matrix is defined as

$$J_n = n^{-1}\sum_{i=1}^{n} p_i(1-p_i)\begin{pmatrix} X_i^T X_i & X_i^T Z_i \\ Z_i^T X_i & Z_i^T Z_i \end{pmatrix} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix}.$$

Define

$$\omega(X,Z) = \exp(X\beta + Z\gamma)(J_{n,10}J_{n,00}^{-1}X^T - Z^T)$$

$$D_n = \hat{\delta}_{full} = \sqrt{n}\hat{\gamma}_{full} \xrightarrow{d} D \sim N_q(\delta, K),$$

that has a limiting normal distribution, where $\hat{\gamma}_{full}$ is a vector of estimates based on the full model, and K is the lowest-right block of the partitioned matrix J^{-1}:

$$K = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}. \tag{2.9}$$

Then the FIC can be written as

$$FIC_m = \Big(\sum_{j:\beta_j \notin m} \hat{\omega}_j D_{n,j}\Big)^2 + 2\sum_{j:\beta_j \in m} \hat{\omega}_j^2 \hat{k}_j^2, \quad \text{for all} \quad m = \{1,\ldots,M\},$$

where $\hat{k}^2 = diag(K)$. Claeskens and Hjort (2008) derived smooth FIC weights as

$$w_m = \exp\Big(-\frac{1}{2}k\frac{FIC_m}{\hat{\omega}^T\hat{K}\hat{\omega}}\Big) / \sum_{i=1}^{M} \exp\Big(-\frac{1}{2}k\frac{FIC_i}{\hat{\omega}^T\hat{K}\hat{\omega}}\Big),$$

where $k \geq 0$ is an algorithmic parameter that moves weights from uniformly distributed for k close to zero to FIC based weidhts for large k.

2.3.5 *Confidence intervals following model-averaging*

The simplest way to calculate a confidence interval for parameters is to use the Wald approach that has a form of $\hat{\hat{\beta}}_j \pm z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{se}}(\hat{\hat{\beta}}_j)$, where $\hat{\hat{\beta}}_j$ is the model-averaged estimator of β_j. Buckland et al. (1997) proposed the unconditional confidence interval constructed by estimating a standard error as,

$$\widehat{\text{se}}_1(\hat{\hat{\beta}}_j) = \sum_{m=1}^{M} w_m \sqrt{\widehat{\text{Var}}(\hat{\beta}_{jm}|B_m) + (\hat{\beta}_{jm} - \hat{\hat{\beta}}_j)^2}.$$

This standard error estimate has two parts: the error in parameter estimation $\widehat{\text{Var}}(\hat{\beta}_{jm}|B_m)$ and a term that measures a variation in the estimates across candidate models $(\hat{\beta}_{jm} - \hat{\hat{\beta}}_j)^2$.

Such standard errors are based on the assumption that the sampling distribution of $\hat{\bar{\beta}}_j$ is asymptotically normal and that weights are known constants. However, neither of these assumptions are usually correct. This method was revised by Burnham and Anderson (2004), who proposed the use of

$$\widehat{\text{se}}_2(\hat{\bar{\beta}}_j) = \sqrt{\sum_{m=1}^{M} w_m \left[\widehat{\text{Var}}(\hat{\beta}_{jm}|B_m) + (\hat{\beta}_{jm} - \hat{\bar{\beta}}_j)^2\right]}.$$

However, they did not find any advantages in using $\widehat{\text{se}}_2(\hat{\bar{\beta}}_j)$ rather than $\widehat{\text{se}}_1(\hat{\bar{\beta}}_j)$ with respect to coverage probability.

Hjort and Claeskens (2003) and Claeskens and Hjort (2008) studied the asymptotic properties of model-averaging unconditional confidence intervals proposed by Buckland et al. (1997). They considered a local misspecification framework under which the nested structure of the candidate models depends upon the true values of underlying model parameters. Study of the limiting distributions and coverage of model-averaged confidence intervals showed that confidence intervals proposed by Buckland et al. (1997) are biased and should be corrected.

Let us consider the logistic regression model in Equation (2.8), and define vector $v = (v_1, \ldots, v_q)$ and π_m as the projection matrix needed foe maping the subsets of variables of indirect interest, such that $\pi_m v = v_m$ the vector that contains $v_j \in m$. Define

$$K_m = (\pi_m K^{-1} \pi_m^T)^{-1}$$

$$Q_m = K^{-1/2} \pi_m^T K_m \pi_m K^{-1/2}.$$

For $\hat{\theta} = \hat{\theta}(\hat{\beta}, \hat{\gamma})$, a model-average estimator of θ Hjort and Claeskens (2003) showed that the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_{true})$ is

$$\sqrt{n}(\hat{\theta} - \theta_{true}) \xrightarrow{d} \Lambda = \left(\frac{\partial \theta}{\partial \beta}\right)^T J_{00}^{-1} \Psi + \omega^T (\delta - \hat{\delta}(D)),$$

where $\Psi \sim N_p(0, J_{00})$ and $\hat{\delta}(D) = K^{1/2}\Big(\sum_m w_m Q_m\Big)K^{-1/2}D$, such that the limiting distribution Λ is non-standartized normal distribution. They proposed the use of confidence limits

$$\hat{\theta}_{L,n} = \hat{\theta} - \hat{\omega}^T\left[D_n - \hat{\delta}(D_n)\right] / \sqrt{n} - z_{1-\frac{\alpha}{2}}\hat{k}/\sqrt{n} \tag{2.10}$$

$$\hat{\theta}_{U,n} = \hat{\theta} - \hat{\omega}^T\left[D_n - \hat{\delta}(D_n)\right] / \sqrt{n} + z_{1-\frac{\alpha}{2}}\hat{k}/\sqrt{n}, \tag{2.11}$$

where $\hat{\omega}$ and $\hat{k}$ are consistent estimators of ω and $k = (X^T J_{00}^{-1} X + \omega^T K \omega)^{1/2}$.

Wang and Zhou (2013) have shown that confidence intervals proposed by Hjort and Claeskens (2003) under the local misspecification framework are asymptotically equivalent to intervals obtained from the full model. The local misspecification framework assumes that all models used in model-averaging contain all important variables, and differ only in a set of unimportant variables. However, in practice this is usually not the case.

Both unconditional Wald type intervals assume that $\hat{\bar{\beta}}_j$ has a normal sampling distribution, which is not always true. Turek and Fletcher (2012) derived MATA interval, that does not require the normality assumption for $\hat{\bar{\beta}}$, but assumes that the set of candidate models contains the true model, and that $\hat{\beta}_j$ from the true model has a normal distribution. By definition, a valid confidence limit must satisfy the following equations:

$$P(\beta_j < \beta_j^L) = \frac{\alpha}{2} \quad \text{and} \quad P(\beta_j > \beta_j^U) = \frac{\alpha}{2},$$

where β_j^L and β_j^U are lower and upper limits of the confidence interval, respectively.

Let us focus on the lower limit and assume that one of the candidate models is the true model. Thus we can define an indicator vector $\mathrm{v} = (\mathrm{v}_1, \ldots, \mathrm{v}_M)$, that gives the value one to the true model, and zero to all other models.

$$P(\beta_j < \beta_j^L) = \sum_{m=1}^{M} \mathrm{v}_m P(\beta_{jm} < \beta_j^L). \tag{2.12}$$

Suppose that model u is the true model. Using the second assumption of the normal distribution for $\hat{\beta}_{ju}$, we have

$$P(\beta_{ju} < \beta_j^L) = 1 - F_{\nu_u}(t_{ju}^L)\,, \tag{2.13}$$

where $F_{\nu_u}(\cdot)$ is the cumulative distribution function of the t-distribution with ν_u degrees of freedom associated with model u, and $t_{ju}^L = (\hat{\beta}_{ju} - \beta_j^L)/\widehat{\text{se}}(\hat{\beta}_{ju})$, where $\widehat{\text{se}}(\hat{\beta}_{ju})$ is the standard error of the maximum likelihood estimate calculated for model u. Combining Equations (2.12) and (2.13) we have

$$P(\beta_j < \beta_j^L) = \sum_{m=1}^{M} \text{v}_m\big(1 - F_{\nu_m}(t_{jm}^L)\big) = \frac{\alpha}{2}\,.$$

However, in an actual study, the true model is unknown, so that the vector v should be estimated by a vector of weights w that is based on some criterion. Thus, the MATA lower confidence limit β_j^L is the solution of

$$\sum_{m=1}^{M} w_m \left(1 - F_{\nu_m}(t_{jm}^L)\right) = \frac{\alpha}{2}\,, \tag{2.14}$$

where $t_{jm}^L = (\hat{\beta}_{jm} - \beta_j^L)/\widehat{\text{se}}(\hat{\beta}_{jm})$. The MATA upper confidence limit β_j^U is the solution of

$$\sum_{m=1}^{M} w_m F_{\nu_m}(t_{jm}^U) = \frac{\alpha}{2}\,, \tag{2.15}$$

where $t_{jm}^U = (\hat{\beta}_{jm} - \beta_j^U)/\widehat{\text{se}}(\hat{\beta}_{jm})$.

Another method of confidence interval construction is referred to as Model-Averaged Profile Likelihood proposed by Fletcher and Turek (2012). Its lower limit is the solution of

$$\sum_{m=1}^{M} w_m \Phi\big(r_m(\beta_j)\big) = \frac{\alpha}{2}\,,$$

where $\Phi(\cdot)$ is the cumulative distribution of the standardized normal distribution and

$$r_m(\beta_j) = \text{sign}(\hat{\beta}_{jm} - \beta_j)\,\sqrt{2(\log L_p(\hat{\beta}_{jm}) - \max_{\gamma} L(\beta_j, \gamma))}\,, \tag{2.16}$$

where $L_p(\cdot)$ is the profile likelihood function and γ is a set of remaining parameters. To estimate the upper limit, $\alpha/2$ should be replaced with $1-\alpha/2$.

While confidence intervals for regression coefficients can be easily constructed by estimating standard errors $\widehat{\mathrm{se}}_1(\hat{\hat{\beta}}_j)$ or $\widehat{\mathrm{se}}_2(\hat{\hat{\beta}}_j)$, the MATA intervals face the problem of zero standard errors. To see this, define the asymptotic version of the transformation-based model-averaged tail area (ATMATA) interval as in Yu et al. (2014). If we assume that with an increase in sample size the cumulative t-distribution function $F_{\nu_m}(\cdot)$ converges to the cumulative distribution function of the standard normal, we can rewrite Equations (2.14) and (2.15) as

$$\sum_{m=1}^{M} w_m \left(\frac{\hat{\beta}_{jm} - \beta_j^L}{\widehat{\mathrm{se}}(\hat{\beta}_{jm})} \right) = z_{1-\frac{\alpha}{2}}$$

$$\sum_{m=1}^{M} w_m \left(\frac{\hat{\beta}_{jm} - \beta_j^U}{\widehat{\mathrm{se}}(\hat{\beta}_{jm})} \right) = z_{\frac{\alpha}{2}}\,,$$

where z_q is a $100q\%$ quantile of a standard normal distribution. By solving these equations we obtain

$$\beta_j^L = \frac{\sum_{m=1}^{M} w_m \big[\hat{\beta}_{jm}/\widehat{\mathrm{se}}(\hat{\beta}_{jm}) - z_{1-\frac{\alpha}{2}}\big]}{\sum_{m=1}^{M} w_m/\widehat{\mathrm{se}}(\hat{\beta}_{jm})} \tag{2.17}$$

$$\beta_j^U = \frac{\sum_{m=1}^{M} w_m \big[\hat{\beta}_{jm}/\widehat{\mathrm{se}}(\hat{\beta}_{jm}) - z_{\frac{\alpha}{2}}\big]}{\sum_{m=1}^{M} w_m/\widehat{\mathrm{se}}(\hat{\beta}_{jm})}\,. \tag{2.18}$$

If we are interested in the construction of a prediction interval, the standard error will always be positive. However this is not the case for a single regression coefficient. If a regression coefficient, β_j, does not appear in model M_l, then $\widehat{\mathrm{se}}(\hat{\beta}_{jl}) \equiv 0$ as well as $\hat{\beta}_{jl} \equiv 0$. In this case, in the numerator we have $\hat{\beta}_{jl}/\widehat{\mathrm{se}}(\hat{\beta}_{jl}) = 0/0$ and in the denominator $w_l/\widehat{\mathrm{se}}(\hat{\theta}_l) = w_l/0 = \infty$. To solve the problem for some β_j, we have to calculate weights based only on the models that contain variable j. Thus, the confidence intervals for regression coefficients are given by,

$$\beta_j^L = \frac{\sum_{m=1}^{M} w_{jm} \big[\hat{\beta}_{jm}/\widehat{\mathrm{se}}(\hat{\beta}_{jm}) - z_{1-\frac{\alpha}{2}}\big] I(\widehat{\mathrm{se}}(\hat{\beta}_{jm}) \neq 0)}{\sum_{m=1}^{M} \{w_{jm}/\widehat{\mathrm{se}}(\hat{\beta}_{jm})\} I(\widehat{\mathrm{se}}(\hat{\beta}_{jm}) \neq 0)} \tag{2.19}$$

$$\beta_j^U = \frac{\sum_{m=1}^{M} w_{jm}\left[\hat{\beta}_{jm}/\widehat{\mathrm{se}}(\hat{\beta}_{jm}) - z_{\frac{\alpha}{2}}\right] I(\widehat{\mathrm{se}}(\hat{\beta}_{jm}) \neq 0)}{\sum_{m=1}^{M}\{w_{jm}/\widehat{\mathrm{se}}(\hat{\beta}_{jm})\} I(\widehat{\mathrm{se}}(\hat{\beta}_{jm}) \neq 0)}, \tag{2.20}$$

where $I(\widehat{\mathrm{se}}(\hat{\beta}_{jm}) \neq 0)$ is the indicator of $\widehat{\mathrm{se}}(\hat{\beta}_{jm})$ is non-zero, and w_{jm} is the weight given to model m and used in the estimation of confidence intervals of β_j.

Chapter 3

SELECTION OF CANDIDATE MODELS

3.1 Introduction

The model-averaging approach needs a well defined group of candidate models to be averaged. The basic requirement for the set of models is that it should include the true model or a model that is very close to it. In addition, this set should not be very large, because as the difference between sample size and number of models decreases both the coefficient estimates and the coverage of the confidence intervals may suffer, especially if there are many unnecessary variables in the data (Buckland et al., 1997).

There are different frequentist methods for defining the candidate models set proposed in the literature. Lukacs et al. (2010) proposed the model-averaging method known as full-model-averaging. This method uses all models, such that models not containing a considered variable contribute zero to the averaged estimator. This method goes against the idea of the scientific justification of each parameter and inclusion of each model into the candidate models set emphasized by Burnham and Anderson (2004), because the full-model-averaging set may contain biologically meaningless models. Their simulations show that this method helps to reduce the error created by the overly complex models chosen by the AIC.

In cases where the model with the smallest AIC has substantial weight ($w > 0.9$), the full-model-averaging approach becomes less attractive, because it averages strong models with a large set of models whose weights are relatively very low (Burnham and Anderson, 2002). Buckland et al. (1997) recommended averaging over the best model and all its submodels. The calculation of the estimate and confidence intervals for a specific variable is done only over the models that contain this variable. For proper estimation the weights have to be renormalized such that the sum of the new weights will be equal to one.

However, a dominating model does not always exist, especially when predictors are highly correlated. In this case, the selection of the best AIC model and its submodels may lead to loss of information and important variables. To avoid this, one can use the number of top AIC models for construction of the candidate models set (Buckland et al., 1997; Bolker, 2008; Richards, 2008). Buckland et al. (1997) also pointed out that in some cases there might be more efficient ways to construct the candidate set by adding models to the candidate model set from largest to smallest until their cumulative weight reaches 0.95.

Candidate models set selection methods, in one way or another, follow the principle of parsimony, coupled with biological reasoning for inclusion of predictors, and can produce good results. However, since biological reasoning is not always available and absolutely correct, these methods must be used with caution, as the inference might be unreliable.

3.2 Parsimony principle and accuracy

Consider the logistic regression model $\mathrm{logit}\big[P(Y=1|X)\big]=X\beta$, where β is a vector of unknown parameters. The vector β can be written as $\beta=(\theta,\gamma)$, where $\theta=(\theta_1,...,\theta_p)$ are parameters associated with variables of interest, and $\gamma=(\gamma_1,...,\gamma_q)$ are parameters associated with candidate variables for inclusion. If we have prior knowledge, we can define θ and γ sets. Prior knowledge is the most desirable model selection strategy (Buckland et al., 1997). It suggests selection of the variables on sound scientific principles that explain the mechanisms underlying the data. This approach is preferable for both single model selection and the candidate models set selection. However, prior knowledge is not always available, might be incomplete, and is not always applicable to different populations. This means that θ and γ cannot always be determined. In this study, we assume that prior knowledge allows us to define the variables associated with β, but the importance of each variable is unknown.

While there are many suggested methods for single model selection, methods for selecting a set of candidate models for model-averaging remain relatively unexplored. Frequentists model-averaging and Bayesian model-averaging face the problem of identifying a sufficient set of models over which the model-averaging should be processed. For the Bayesian settings, the most common approach is the Occam's window algorithm (Hoeting et al., 1999). This approach compares nested models by the log posterior odds, and for each comparison it has three options: reject the largest model, reject the smallest model with all its submodels, or reject neither model if the log posterior odds falls into the Occam's window. According to the principle of parsimony, Occam's window significantly reduces the number of models under consideration and simplifies the model-averaging process.

One of the proposed methods in the frequentist setting is stepwise selection that chooses the set of variables in the construction of the models for model-averaging. For example, if a stepwise selection procedure chooses five variables, then $2^5 = 32$ models constructed from these variables will be used in the model-averaging process. An approach based on bootstrap was proposed by Austin and Tu (2004b). The results support the use of bootstrap sampling and application of stepwise selection for each bootstrapped data set. The final model is built only from the variables that pass a 60% exclusion fraction, which also reflects the principle of parsimony. The candidate set of models is then constructed from this model and its submodels. Both methods can significantly reduce the number of models by exclusion of variables; however, it may also lead to the loss of valuable information that could be used in the model-averaging process.

There is no doubt about the importance of parsimony in the final model selection process. However, parsimony of the model is not a 'gold standard', but rather a desirable property for better generalization performance (Harrell, 2015). Regarding model-averaging, this principle is not necessarily the basis for the process since model-averaging should be carried out over a group of candidate models that include both simple and more complex models. More complex models can contain valuable information or important variables that may not be selected for the final model. However, the use of the parsimonious model as the most complex model in the model-averaging process carries a potential risk for inference.

The opposite of the parsimonious model is the full model. Assuming that the full model can be fitted, it provides valid confidence limits. However, the full model provides confidence intervals that are usually too wide, and results that are difficult to interpret if number of variables is large. Thus, there is a need of a candidate models set selection technique that prioritizes accuracy over parsimony and shrinks confidence interval length while maintaining the claimed coverage property.

3.3 Candidate models selection based on inclusion fraction

As noted previously, if the model based on the parsimony principle defines the upper bound of model complexity for model-averaging, meaning that the most parsimonious model is the most complex model in a set of candidate models, it may result in undercoverage of confidence intervals. Thus, the current candidate models set selection methods in the frequentist setting that are based on this principle should be used with caution. We propose an inclusion fraction method that suggests applying bootstrap on original sample, and apply backward selection with AIC penalty on each bootstrapped sample. Then the variables that appeared in 50% of selected models are assigned to θ that is a set of important variables and the remaining variables are assigned to γ, which presents a group of variables under consideration, or variable of indirect interest. Then we suggest permuting the parameters in the γ set, and add θ to them, so the model-averaging is done over 2^q models. The 50% inclusion fraction is analogous to the 50% posterior probability that is a conventional Bayesian threshold (Kass and Raftery, 1995; Genell et al., 2010). This method fundamentally differs from the exclusion fraction method, because parsimony defines the lower bound of model complexity. The most parsimonious model is the simplest model in the set of candidate models, while the full model is considered the most complex model used in model-averaging.

Inclusion fraction, stepwise selection, and Occam's window approaches significantly reduce the number of the models under consideration. However, unlike the exclusion fraction approach, they also allow all variables to appear in model-averaging, which means that each variable will have a corresponding point estimate and confidence interval, regardless of which final model is chosen. While a 60% cut-off point in the exclusion fraction method may be unreasonable for some data, the 50% threshold is the natural choice, because it is a non-informative point and corresponds to the initial assumption that we do not know which variables are important. Of course, if prior knowledge allows, the idea of a 50% threshold for all variables can be expanded to the unique thresholds for each variable, but we assume that there is no prior knowledge besides the one that defines the full model.

In general, the approach based on inclusion fraction should provide a larger set of models than the Occam's window approach. However, a larger set does not mean that the approach is weaker. It only means that the inclusion fraction method needs more time to obtain results. What is more important is the accuracy of the inference. If the inclusion fraction method demonstrates better performance, then it is preferable over other methods, despite the larger set of the candidate models.

Chapter 4

IMPROVING THE WALD MODEL-AVERAGING CONFIDENCE INTERVALS

4.1 Introduction

As pointed out in previous chapters, regression analysis has commonly been based on a single model that was selected by a model selection process. Model selection and subsequent inference is still a popular way to conduct statistical analysis, despite the fact that the single model ignores uncertainty due to model selection, which leads to biased inference (Berk et al., 2010). To solve this problem, model-averaging procedures were developed. Model-averaged methods account for uncertainty, but accurate CI procedures deserve further research.

Methods for constructing confidence interval by model-averaging in the frequentist setting have been studied by Buckland et al. (1997) and Burnham and Anderson (2004). These procedures assume that model-averaged estimators are approximately normal, and variances have a closed form and can be estimated, allowing symmetric confidence interval to be constructed around the point estimate. In cases where estimators do not closely follow a normal distribution, Wald intervals may not perform well. For a single model, this problem can be solved by calculating a Wald interval for a transformed parameter, and then transforming intervals back. However, even if the sampling distribution of the parameter estimates for each considered model can be assumed to be normal, it does not mean that model-averaged parameter estimates will also have the same property, because the weights used in averaging process are also estimated with uncertainty. Claeskens and Hjort (2008) and Hjort and Claeskens (2003) suggested a correction for these procedures; however, Wang and Zhou (2013) have shown that the corrected confidence interval converges to the full model in the parametric context as well as in the semi-parametric model framework.

Turek and Fletcher (2012) proposed a Wald-type confidence interval that is estimated by averaging the tail areas of the sampling distributions that does not require estimation of standard errors. This methodology provides reasonable confidence intervals for normal linear models. To account for skewness of the interval, Fletcher and Turek (2012) derived model-averaged profile likelihood confidence intervals, which performed better than Wald intervals. Kabaila et al. (2016) compared Wald based and profile-likelihood based tail area confidence interval construction methods for a simple case with only two models, and showed that the coverage of the profile-likelihood based method decreases as the ratio p/n increases, while the Wald based method demonstrates more stable coverage properties. As discussed in Fletcher and Turek (2012) and shown in Kabaila et al. (2016), the model-averaged profile likelihood confidence interval works well only if the profile confidence intervals of each model perform well. The profile-likelihood method performs poorly when correlation among variables is large or the sample size is small.

4.2 Model-averaged intervals based on score test

For single logistic regression for small samples, score based confidence intervals usually outperform the Wald or profile-likelihood intervals (Agresti, 2011). Despite the advantages of the score confidence intervals in finite samples, they are rarely applied in the context of regression analyses (Engle, 1984). Since their inception, they have not actually been applied outside of contingency table analysis. One of the possible reasons is that a score confidence interval is not accessible in statistical software. To overcome the performance issues in small samples, we propose to construct confidence intervals for model-average parameters by using a score test method. Consider the following logistic regression model:

$$\operatorname{logit}\big(P(Y=1|X)\big) = X\beta\,, \tag{4.1}$$

where $\beta^T = (\beta_1,...,\beta_p)$ and X is an $n \times p$ matrix of independent variables. The score function of this model is defined by

$$U(\beta) = \frac{\partial}{\partial\beta}\ell(\beta|Y,X) = \sum_{i=1}^{n} X_i \left(Y_i - \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \right), \tag{4.2}$$

where X_i is a vector of values of predictors for subject i, and Y_i is a binary outcome of subject i. The observed Fisher information matrix for this model can be defined by the negative second partial derivatives of the log-likelihood function

$$I(\beta)_{jk} = -\frac{\partial^2}{\partial\beta_j\partial\beta_k}\ell(\beta|Y,X) = -\sum_{i=1}^{n} \frac{\exp(X_i\beta)}{(1+\exp(X_i\beta))^2} X_{ij}X_{ik}\,, \tag{4.3}$$

where X_{ij} and X_{ik} are the values of predictors j and k for subject i. From the score function and the observed Fisher information, we can define the score test function for some β_j as

$$ST(\beta_j) = U(\beta_j)I^{-1}(\beta_j)U^T(\beta_j)\,. \tag{4.4}$$

$ST(\beta_j)$ is a function of β_j and has χ^2 distribution with one degree of freedom. For the model-averaging framework, let us define this function for regression coefficient β_j in some model m by

$$ST_m(\beta_j) = U_m(\beta_j)I_m^{-1}(\beta_j)U_m^T(\beta_j)\,. \tag{4.5}$$

To conduct just a score test, we have to fit the regression model m, estimate all relevant parameters and plug $\hat{\beta}_j$ into $S(\beta_j)$. However, if we want to calculate the confidence interval based on score test function, we have to define statistics

$$S_m(\beta_j) = \operatorname{sign}(\hat{\beta}_{jm} - \beta_j)\,\sqrt{U_m(\beta_j,\hat{\gamma}_m)I_m^{-1}(\beta_j,\hat{\gamma}_m)U_m^T(\beta_j,\hat{\gamma}_m)}\,, \tag{4.6}$$

where $\hat{\gamma}_m$ is a set of refitted coefficients of the nuisance variables. The $(1-\alpha)100\%$ score confidence interval for β_j is defined by β_j^L and β_j^U that satisfy

$$\Phi\big(S_m(\beta_j^L)\big) = \frac{\alpha}{2} \quad \text{and} \quad \Phi\big(S_m(\beta_j^U)\big) = 1 - \frac{\alpha}{2}, \tag{4.7}$$

where $\Phi(\cdot)$ is the cumulative distribution of the standardized normal distribution. By analogy with model-averaged profile-likelihood, intervals we can define the model-averaged score test interval for β_j as

$$\sum_{m=1}^{M} w_m \Phi\big(S_m(\beta_j^L)\big) = \frac{\alpha}{2} \quad \text{and} \quad \sum_{m=1}^{M} w_m \Phi\big(S_m(\beta_j^U)\big) = 1 - \frac{\alpha}{2}. \tag{4.8}$$

The score intervals as well as the model-averaged profile-likelihood confidence intervals are based on an analogy with Bayesian model-averaging, but unlike the profile-likelihood intervals, the score intervals do not require the evaluation of the maximum likelihood estimates under the alternative model.

4.3 Confidence intervals based on Wald standard errors

Turek and Fletcher (2012) proposed the model-averaging tail area confidence intervals, which are obtained by solving the equations

$$\sum_{m=1}^{M} w_m \left(1 - F_{\nu_m}\big(t_{jm}^L\big)\right) = \frac{\alpha}{2} \quad \text{and} \quad \sum_{m=1}^{M} w_m F_{\nu_m}\big(t_{jm}^U\big) = \frac{\alpha}{2}, \tag{4.9}$$

where $F_{\nu_m}(\cdot)$ is the cumulative distribution function of the t-distribution with ν_m degrees of freedom associated with model m, $t_{jm}^L = (\hat{\beta}_{jm} - \beta_j^L)/\widehat{\text{se}}(\hat{\beta}_{jm})$, $t_{jm}^U = (\hat{\beta}_{jm} - \beta_j^U)/\widehat{\text{se}}(\hat{\beta}_{jm})$, and $\widehat{\text{se}}(\hat{\beta}_{jm})$ are Wald standard errors. A key feature of this method is that standard errors used in the estimation of lower and upper limits are assumed to be equal.

While the final confidence interval obtained from Wald model-averaging tail area method is not symmetric, the standard errors equivalency for each considered model is an unnecessary restriction. It is known that for each individual model, the profile-likelihood or score methods provide asymmetric confidence intervals that usually outperform those based on the Wald method (Newcombe, 1998; Agresti, 2011). This advantage becomes even more noticeable with smaller sample sizes. Therefore, it is possible to improve the Wald model-averaging tail area confidence intervals by replacing the standard errors of the Wald confidence interval by standard errors obtained from the profile-likelihood or score confidence intervals. s

4.4 Wald MATA corrected by the profile-likelihood and score standard errors

Although the profile-likelihood and score intervals do not have closed forms, their standard errors can be estimated. Suppose $\left(\hat{\beta}^{pl}_{L,m}, \hat{\beta}^{pl}_{U,m}\right)$ and $\left(\hat{\beta}^{S}_{L,m}, \hat{\beta}^{S}_{U,m}\right)$ are confidence intervals for parameter β_j in model m obtained from profile-likelihood and score methods, respectively. For the profile-likelihood method, the lower and upper standard errors can be estimated by

$$\widehat{\text{se}}^{pl}_{L}(\hat{\beta}_{jm}) = \frac{\hat{\beta}_{jm} - \hat{\beta}^{pl}_{L,m}}{2z_{1-\frac{\alpha}{2}}} \quad \text{and} \quad \widehat{\text{se}}^{pl}_{U}(\hat{\beta}_{jm}) = \frac{\hat{\beta}^{pl}_{U,m} - \hat{\beta}_{jm}}{2z_{1-\frac{\alpha}{2}}}, \tag{4.10}$$

and for the score method they can be estimated by

$$\widehat{\text{se}}^{S}_{L}(\hat{\beta}_{jm}) = \frac{\hat{\beta}_{jm} - \hat{\beta}^{S}_{L,m}}{2z_{1-\frac{\alpha}{2}}} \quad \text{and} \quad \widehat{\text{se}}^{S}_{U}(\hat{\beta}_{jm}) = \frac{\hat{\beta}^{S}_{U,m} - \hat{\beta}_{jm}}{2z_{1-\frac{\alpha}{2}}}. \tag{4.11}$$

The standard errors recovered from profile-likelihood or score confidence intervals in Equations (4.10) and (4.11) can be used in Equation (4.9). Such replacement can improve the confidence intervals in terms of length, as well as computational time. Wald based model-averaging confidence interval calculation is much faster than the profile-likelihood or score based ones. Since the optimization algorithm is less complex, it is less likely to fail to converge.

4.5 Confidence intervals based on Bayesian model-averaging

The standard error of a model-averaged posterior distribution for $\hat{\beta}_j^{BMA}$ from BMA has the form:

$$\operatorname{se}(\hat{\beta}_j^{BMA}) = \sqrt{\sum_{B_m \in V} \left\{ [\operatorname{Var}(\beta_{jm}|D,B_m) + \hat{\beta}_{jm}^2] p(B_m|D) \right\} - (\hat{\beta}_j^{BMA})^2},$$

where V is a set of candidate models obtained by the Occam's window method (Hoeting et al., 1999).

Statistical software, such as SAS or R, allows BMA to proceed with the set of selected models and provides the point estimates and the standard errors of model-averaged posterior distributions for each coefficient. It is known that BMA can be a useful tool for model selection; however, the reliability of posterior standard errors is not very clear. One may use this information to construct the intervals based on the estimated effects as

$$\hat{\beta}_j^{BMA} \pm z_{1-\frac{\alpha}{2}} \times \widehat{\operatorname{se}}(\hat{\beta}_j^{BMA})$$

and interpret them as regular confidence intervals. For example, Fang et al. (2016) in a study of associations between air pollution and respiratory mortality, applied a Bayesian model-averaging method on the set of generalized additive mixed models and presented Wald-type confidence intervals. In a similar study, Portnov et al. (2012) calculated Wald based confidence intervals from BMA and compared these to the results from a single model.

The unconditional standard error proposed by Burnham and Anderson (2004) was motivated by the posterior standard errors. As shown in Hjort and Claeskens (2003), Wang and Zhou (2013), Turek and Fletcher (2012), and Fletcher and Turek (2012), the intervals obtained using unconditional standard errors in the frequentist setting have poor performance, although it is unclear whether a similar method using a Bayesian approach would perform better. Thus, in this research we also test the validity of the BMA Wald-type confidence interval and give our recommendations regarding its use.

Chapter 5

EVALUATION OF CONFIDENCE INTERVAL PROCEDURES

5.1 Introduction

Previous chapters presented statistical methods for model-averaging confidence interval construction. In Chapter 3, we proposed to use a 50% inclusion fraction based on the bootstrap procedure to define the set of candidate models for model-averaging. In Chapter 4, we proposed the model-averaging confidence interval construction method based on the score function and the improved Wald based model-averaging confidence intervals. In this chapter, some of the methods described in Chapter 2 are compared empirically under different parameter combinations to proposed methods.

Section 5.2 presents the confidence interval construction methods that are compared in this simulation study. Section 5.3 describes the algorithm for comparing the methods. Section 5.4 describes the parameters and their combinations used for the simulation study. The data generating process is described in Section 5.5. In Section 5.6, we list the software and tools used in this simulation study. Finally, the evaluated statistics, tabulated results of the simulation study, and the analysis of the results are presented in Section 5.7.

5.2 Confidence interval construction methods

In Chapter 2, we presented different confidence interval construction methods. In this simulation study, we evaluate the performance of our proposed methods compared to existing methods. Overall, 19 different methods are compared. The first group that contains six methods that provide confidence intervals for all independent variables is:

1. Full model regression (FULL) that fits one model with all possible variables and calculates Wald type confidence intervals,
2. Inclusion fraction (50%) based model-averaged tail area method using Wald approach (I-MATA-W),
3. Inclusion fraction (50%) based model-averaged tail area method using profile-likelihood approach (I-MATA-PL),
4. Inclusion fraction (50%) based model-averaged tail area method using score based approach (I-MATA-S),
5. Inclusion fraction (50%) based model-averaged tail area method using Wald approach with profile-likelihood standard errors (I-MATA-Wpl),
6. Inclusion fraction (50%) based model-averaged tail area method using Wald approach with score based standard errors (I-MATA-Ws).

The second group consists of 13 methods involving variable selection processes:

1. Backward stepwise selection method with AIC based penalties (STEP-AIC),
2. Backward stepwise selection method with BIC based penalties (STEP-BIC),
3. Zero-corrected backward selection method (ZERO-C) that averages over selected regression models from bootstrapped samples and uses the percentile method for confidence interval construction,
4. BMA Wald approach (BMA-W) that uses the mean and standard error of a model-averaged posterior distribution to construct Wald based confidence intervals,
5. LASSO with λ that is within 1 standard error of the minimum,
6. Stepwise AIC exclusion based model-averaged tail area method using Wald approach (E-MATA-W),
7. Stepwise AIC exclusion based model-averaged tail area method using profile-likelihood approach (E-MATA-PL),

8. Stepwise AIC exclusion based model-averaged tail area method using score based approach (E-MATA-S),
9. Occam's window based model-averaged tail area method using Wald approach (B-MATA-W),
10. Occam's window based model-averaged tail area method using profile-likelihood approach (B-MATA-PL),
11. Occam's window based model-averaged tail area method using score based approach (B-MATA-S),
12. Occam's window based model-averaged tail area method using Wald approach with profile-likelihood standard errors (B-MATA-Wpl),
13. Occam's window based model-averaged tail area method using Wald approach with score based standard errors (B-MATA-Ws).

In addition, we also fit the true model that is used in the model comparison process described in the next section. For all model-averaged tail area methods, AIC was used as the weighting criteria, because it leads to better performance for such confidence interval construction methods (Turek and Fletcher, 2012; Fletcher and Turek, 2012; Kabaila et al., 2017).

5.3 Comparison procedure

This simulation study compares different confidence interval construction methods for logistic regression coefficients in terms of:

1. Confidence interval coverage level, which reflects the validity of the methods. Valid confidence intervals should provide coverages that are close to the nominal coverage level.
2. Confidence interval tail errors, which indicate the accuracy and balance of the methods. Balanced tail errors mean that a method excludes extreme values from both sides of the interval.

3. Confidence interval width, which shows the efficiency of the methods. The shortest intervals that meet the two previous properties are considered more precise, and thus more desirable.

In addition, we present the averaged point estimates to ensure that confidence interval methods can provide consistent estimates. Methods are investigated for 95% confidence. We decided to conduct 1,000 runs only for full model, true model, zero-corrected backward selection method and five methods based on the inclusion fraction, because these methods provide confidence intervals for all variables in each run. Based on the minimal number of simulation runs, the appropriate empirical coverage range is defined by $95 \pm 1.96\sqrt{(95 \times 5)/1000}$ = (93.6%, 96.4%).

Some of the compared methods, such as the stepwise selection based approaches or LASSO method, are selecting variables before constructing the confidence intervals. If we run only 1,000 simulations for these methods, some of the variables will be selected in fewer than 1,000, which means that the empirical coverage of their coefficient effects will be based on less than 1,000 confidence intervals. Thus, if the coverage calculation for these methods is based on the same number of simulations as for the methods that do not select variables, the coverage for frequently omitted variables will not be comparable (Lukacs et al., 2010). To ensure that the methods be comparable, the number of simulation runs for the methods that select variables should be increased.

The BWald-type BMA, three model-averaging procedures based on stepwise AIC exclusion method, and five procedures based on Occam's window method include optimization algorithms that we cannot afford to run for a large number of simulations. However, these methods involve variable selection processes, thus to achieve a reasonable empirical coverage for frequently eliminated variables, we run 5,000 simulation replicates. A total of 5,000 runs is enough to ensure that the most frequently eliminated variable will be selected in at least 1,000 runs, such that empirical coverage based on these runs will be comparable to other methods. The backward stepwise selection methods and LASSO are also selecting variables; however, since the calculations for these methods are faster, we perform 10,000 runs (Lukacs et al., 2010).

Since we compare 19 methods, for convenience we grouped the results in eight tables marked by suffix letters A to H. The first three tables always present the means of point estimates. The table with suffix letter A includes means of the point estimates of the true model, full model and two backward stepwise selection methods. The second table (B) includes means of the point estimates of zero-corrected backward selection method, LASSO method and BMA-W. We combined them together since all these methods are related to the Bayesian framework (Austin, 2008; Park and Casella, 2008). The third table (C) always contains means of the model-averaged point estimates based on three different methods for selecting the candidate models - backward selection, Occam's window and 50% inclusion fraction.

Tables D, E, F, G, and H contain the performance results of the confidence intervals. Table D includes coverage, tail errors and averaged width of confidence intervals constructed by the true model, full model and two backward stepwise selection methods. Table E presents the inference results for zero-corrected backward selection method, LASSO method and BMA-W. Tables F, G, and H present coverage probability, tail errors and averaged width of confidence intervals that were constructed by three stepwise AIC exclusion based, five Occam's window based and five inclusion fraction based model-averaged tail area methods, respectively. This tabulation structure is repeated for each of the simulation parameters presented in the next section.

The results presented in tables marked by suffix letters A, B, and C can be used to calculate the bias of the point estimates. We first compare the point estimates produced by the methods. However, methods that produced large bias will not be eliminated from further comparison, since we are also interested in their ability to estimate valid confidence intervals.

The simulation framework allows us to use the information about the data generating process. Even if the true model is known and can be applied, the coverage level of some of the predictors may not fall within the (93.6%, 96.4%) range. This is possible due to small sample size or because of large correlation among the generated variables. Thus, we analyze the coverage level obtained from the true model. Only the coefficients whose empirical coverage from true model falls within the appropriate coverage interval are compared. The methods whose coverage level is below or above the empirical coverage range for a large number of regression coefficients are excluded from further analysis. The methods that demonstrate acceptable coverage level are then compared by tail errors.

Left and right tail errors indicate the percentage of times the confidence interval was above or below the true parameter, respectively. For 95% confidence intervals, the ideal tail errors both are equal to 2.5%, which means that only extreme values are excluded. After the first comparison, we compare the methods, whose sum of the tail errors varies in the 3.6% to 6.4% interval. We are interested in the difference of the tail errors $|t_u - t_l|$, thus we accept the method if this difference is smaller than 1% for a large number of regression coefficients. Methods are compared only for regression coefficients that lie within the range for the true model.

Finally, all methods that pass the two previous comparison rounds are compared by confidence interval width. The narrowest width that indicates the better precision of the method is desirable. Compasion of widths produced by the valid approaches is also presented in the Figures 5.1 to 5.4.

5.4 Choice of parameters in simulation

Estimation of regression coefficients and their confidence intervals might be sensitive to many factors. The performance of confidence interval procedures for logistic regression can be affected by sample size, probability of outcome and the number of predictors and correlation among them. In this simulation study, we compare different methods for confidence interval construction in four simulation blocks. In each block, one of the mentioned parameters, for example correlation among the predictors, changes, while other parameters are fixed. All investigated parameter combinations are summarized in Table 5.1.

Sample size is the parameter that usually can be prespecified by researchers. We are interested in checking the performance of methods when the number of observations is not large. The regression coefficient estimates in logistic regression do not have a closed form and can be estimated through the optimization of the likelihood function. Insufficient sample size in logistic regression can lead to not only estimation inaccuracy, but also divergence of the optimization algorithm, known as the separation problem (Heinze and Schemper, 2002). This problem occurs when the log-likelihood function is bounded by zero and cannot reach a maximum by increase of regression coefficient estimates of β.

In the first block of simulations, in order to test the effect of sample size on the performance of methods, we fix the number of variables at 5, correlation among variables at 0.5 and outcome probability at 0.5, and set sample sizes to be 100, 300, and 500. The probability of outcome was controlled by the effect of intercept equal to -0.15.

The number of predictors can affect the performance of the methods. In this study, we test the consequences of increasing the number of predictors on the performance of confidence interval procedures. We generated continuous variables from normal distributions and binary variables from Bernoulli distributions. To evaluate the effect of the number of predictors on the performance of confidence interval procedures we fix sample size at 500, correlation at 0.5 and outcome probability at 0.3, and set the number of predictors to be 3, 5, and 10. Such parameters were chosen with the consideration of the time required to complete the simulation and the complexity of the analysis. For example, for $p = 3$, a sample of 500 is sufficient, but for $p = 10$ the sample size will be quite small, and the difference between the methods, if any, will be more evident. We generated the data with the true effects as following,

- For the simulation model with 3 variables, we specified $\beta = (0.01, 0.5, -1)$, such that X_1 and X_3 are normally distributed, and variable X_2 associated with 0.5 effect on outcome has Bernoulli(0.5) distribution.

- For the data with 5 variables, we added one Bernoulli variable with -0.2 effect and one normally distributed variable with no effect on the outcome. The true effects $\beta = (0, 0.01, -0.2, 0.5, -1)$ are also used for simulations related to sample size, correlation, and probability of events.
- For the largest simulated data with $p = 10$, we added two redundant normally distributed variables and one from a Bernoulli distribution, one normally distributed variable with effect of -0.7 and one Bernoulli distributed variable with the largest effect of 2.5, such that $\beta = (0, 0, 0, 0, 0.01, -0.2, 0.5, -0.7, -1, 2.5)$ is the final vector of the true effects.

We expanded the simple model with three variables by adding not only important variables, but also irrelevant variables to the dataset. This should also affect the performance of some methods. The outcome probability was controlled for each model by the intercepts of -1.23, -1.15 and -1.18.

The third parameter we are interested to assess is correlations among independent variables. It is presented in the third block of simulations. We assessed how methods perform under the change in correlations from non-correlated data to moderate correlation. We fix the sample size at 300, the number of variables at 5, and the probability of outcome at 0.3, and set the correlation to be 0, 0.3 and 0.5. The intercepts of -1.42, -1.28 and -1.18 being defined to control the outcome probability across this group of simulations. For large sample size, correlation is not a big issue; however, if sample size is small relative to the number of predictors and the outcome probability is far from 0.5, it may decrease the performance of confidence intervals and increase bias.

Table 5.1: Parameter combinations used for simulation study. N - sample size, p - number of predictors, ρ - correlation among predictors, Pr - event probability.

Combination	N	p	ρ	Pr
1	100	5	0.5	0.5
2	300	5	0.5	0.5
3	500	5	0.5	0.5
4	500	3	0.5	0.3
5	500	5	0.5	0.3
6	500	10	0.5	0.3
7	300	5	0	0.3
8	300	5	0.3	0.3
9	300	5	0.5	0.3
10	500	5	0.3	0.1
11	500	5	0.3	0.3
12	500	5	0.3	0.5

The shaded area shows parameters evaluated in each simulation block.

The outcome probability, $P(Y = 1)$, is an important parameter in logistic regression, which is the last parameter of interest. We fix the sample size at 300, the number of variables at 5 and the correlation between them at 0.3, and set the probability of outcome to be 0.1, 0.3, and balanced 0.5 by defining intercept being equal to -2.7, -1.15 and -0.15, respectively. We defined these values by using the same idea as for the number of predictors. For example, 300 observations with 0.5 is enough, but it becomes more problematic with a decrease in outcome probability.

In total, there are 12 unique parameter combinations used in this simulation study. We devided them into four blocks that are indicated by horizontal lines. Each block evaluates one of the considered parameters is highlighted in Table 5.1.

5.5 Data generation

To generate datasets for the simulation study, we used the `copula` package in R that allows the generation of multivariable correlated data (Yan, 2007). In this simulation study, we are interested in testing how methods perform when data contains not only continuous, but also binary explanatory variable. For example, if we are interested in the model with five variables, then we want three of the five variables to be normally distributed $\mathrm{N}(0,1)$, and two to be from the $\mathrm{Bernoulli}(0.5)$ distribution. We first define the correlation matrix according to the final distribution of each variable with $\rho = 0.3$ as follows:

$$\begin{bmatrix} 1 & 0.3 & 0.38 & 0.38 & 0.3 \\ 0.3 & 1 & 0.38 & 0.38 & 0.3 \\ 0.38 & 0.38 & 1 & 0.45 & 0.38 \\ 0.38 & 0.38 & 0.45 & 1 & 0.38 \\ 0.3 & 0.3 & 0.38 & 0.38 & 1 \end{bmatrix}. \tag{5.1}$$

Using the matrix 5.1 we generated variables from multivariate normal distribution, such that 0.38 corresponded to the correlation between two normally distributed variables one of which will be later transformed into a Bernoulli distributed variable, and 0.45 corresponded to the correlation between two variables that both will be later transformed into Bernoulli distributed variables. After the transformation of appropriate variables, we obtain data with binary and continuous variables that have a correlation of 0.3.

Similarly, the data with $\rho = 0.5$ can be generated by using the following correlation matrix:

$$\begin{bmatrix} 1 & 0.5 & 0.63 & 0.63 & 0.5 \\ 0.5 & 1 & 0.63 & 0.63 & 0.5 \\ 0.63 & 0.63 & 1 & 0.71 & 0.63 \\ 0.63 & 0.63 & 0.71 & 1 & 0.63 \\ 0.5 & 0.5 & 0.63 & 0.63 & 1 \end{bmatrix}. \tag{5.2}$$

To simulate uncorrelated data, we use the identity matrix as a correlation matrix. The data generation process was combined in function '`copulaData`', which can be found in the Appendix A.1.

For each of the 12 parameter combinations presented in Table 5.1, we generated different numbers of data sets as follows. First, we considered the number of simulation runs $S = \{1000, 5000, \text{ or } 10000\}$. Then we generated a data matrix X that contains only predictors and then we simulate S binary output vectors of y, such that

$$y_i \sim \text{Bernoulli}(\text{Pr}_i)\,,$$

where

$$\text{Pr}_i = \text{P}(Y_i = 1) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}\,.$$

Depending on the number of runs, we apply each of the 19 methods described in Section 5.2. For inclusion fraction based methods that involve bootstrapping, we use 1000 bootstrap samples.

5.6 Software and packages

The data generating and analysis programs were written in R v3.4.4 software (R Core Team, 2013). The R packages used in the simulation study are `tictoc`, `copula`, `MASS`, `matrixStats`, `glmnet`, `parallel`, `BMA`, `MuMIn`, and `rootSolve`. The `tictoc` package provides a stopwatch timer, that allows one to measure time spent on different parts of an algorithm (Izrailev, 2014). The `copula` package was used in the data generating process (Yan, 2007). The `MASS` and `matrixStats` packages contain a large list of basic support functions (Venables and Ripley, 2002; Bengtsson, 2018). The `glmnet` package contains a list of functions for penalized regression modelling (Friedman et al., 2010). The `parallel` package allows one to run parallel computation, which significantly reduces the simulation time (R Core Team, 2013). The `BMA` and `MuMIn` packages allow one to do Bayesian and frequentist model-averaging, respectively (Raftery et al., 2005; Barton, 2009). Frequentist model-averaging from the `MuMIn` package corresponds to the method suggested by Burnham and Anderson (2002), thus we used it not as a confidence construction tool, but only as a tool for intermediate calculations. Finally, the `rootSolve` package was used in the optimization processes for profile-likelihood and score based model-averaged confidence intervals (Soetaert and Herman, 2009).

5.7 Results

The simulation results are presented in four groups, such that each group evaluates the effect of - sample size, number of predictors, correlation and probability of the outcome, respectively. Starting with the effect of sample size, the analysis was performed for each parameter of interest separately in the following order. First, we reviewed methods for acceptable point estimates, then we analyzed the empirical coverage probabilities, then discussed the tail errors balance, and finally compared the averaged widths.

5.7.1 Sample size

Tables 5.2.A to 5.2.C present the average point estimates for all considered approaches as a function of sample size. The coverage probabilities, tail errors and average widths of all methods are presented in Tables 5.2.D to 5.2.H. Comparison of the valid methods by their widths is presented in Figure 5.1. The simulations were done for the parameter combinations 1-3 in Table 5.1. We evaluated the model with five variables that have effect $\beta = (0, 0.01, -0.2, 0.5, -1)$ on binary outcome y. We considered three sample sizes of 100, 300, and 500, while correlation and probability of outcome were both fixed at 0.5.

Bias of point estimates

Overall, bias decreased with sample size. However, some of the methods provided highly biased point estimates in comparison to the point estimates produced by the true model even with the largest sample size. Among the methods presented in 5.2.A only the full model provided relatively unbiased point estimates, while estimates provided by two backward stepwise selection methods were biased away from zero (Harrell, 2015). The stepwise selection with BIC penalty tended to produce larger bias than the AIC based stepwise approach as sample size increased.

Table 5.2.A: Mean of point estimates obtained from the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different sample sizes, where $\rho = 0.5$ and outcome probability is 50%. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
N	β	TRUE	FULL	STEP-AIC	STEP-BIC
100	0	—	-0.01	0.03	-0.03
	0.01	0.01	0.01	-0.001	0.01
	-0.2	-0.19	-0.19	-0.46	-0.51
	0.5	0.54	0.55	1.11	1.42
	-1	-1.08	-1.09	-1.07	-1.03
300	0	—	0.001	0.04	0.20
	0.01	0.02	0.02	0.06	0.22
	-0.2	-0.21	-0.21	-0.46	-0.50
	0.5	0.52	0.52	0.76	0.93
	-1	-1.03	-1.03	-1.02	-0.96
500	0	—	0.001	-0.01	0.06
	0.01	0.02	0.01	0.03	0.14
	-0.2	-0.21	-0.21	-0.42	-0.63
	0.5	0.50	0.50	0.60	0.73
	-1	-1.01	-1.02	-1.01	-0.97

Among the methods presented in Table 5.2.B, only the zero-corrected backward selection method provided relatively accurate point estimates. The LASSO based approach provided estimates biased away from zero, while the Bayesian model-averaging resulted in estimates mostly biased toward zero. Due to shrinkage, LASSO usually produces estimates that are biased towards zero for non-zero coefficients; however, in our study we used LASSO only to select the variables, and the point estimates were obtained by fitting regular logistic regression with selected covariates. Since the data was used twice, first to select the variables and then to fit a regular logistic regression model, this LASSO-based approach shares the uncertainty problem with the stepwise selection procedure, which explains the bias away from zero.

Table 5.2.B: Mean of point estimates obtained from the zero-corrected backward selection, LASSO, and Wald based Bayesian model-averaging methods for different sample sizes, where $\rho = 0.5$ and outcome probability is 50%. The LASSO results are based on 10,000 simulations, the results of zero-corrected backward selection and Wald based Bayesian model-averaging are based on 5,000 simulations.

N	β	ZERO-C	LASSO	BMA-W
100	0	-0.01	-0.08	0.003
	0.01	0.01	-0.08	0.001
	-0.2	-0.19	-0.41	-0.02
	0.5	0.547	0.855	0.16
	-1	-1.16	-0.99	-0.98
300	0	0.01	-0.02	0.01
	0.01	0.02	0.01	0.01
	-0.2	-0.16	-0.33	-0.01
	0.5	0.47	0.75	0.17
	-1	-1.05	-0.99	-0.97
500	0	0.002	-0.01	0.01
	0.01	0.01	0.01	0.01
	-0.2	-0.17	-0.29	-0.04
	0.5	0.46	0.62	0.26
	-1	-1.03	-0.99	-0.98

The bias we discuss in this study is entirely a frequentist concept, and is treated differently in the Bayesian framework. In the frequentist framework, the loss function is minimized with respect to a single value that is unknown, while the Bayesian expected loss function is dependent on the prior distribution (Samaniego, 2010). The frequentist and Bayesian point estimates coincide if a non-informative prior is used. The underestimation of the BMA point estimates decreases with sample size as the distribution of β converges to a normal distribution. However, for finite and small samples, the shift toward zero of the Bayesian posterior means is common in logistic regression (Viallefont et al., 2001).

Table 5.2.C: Mean of point estimates of the backward stepwise selection (E-MATA), Occam's window (B-MATA) and inclusion fraction (I-MATA) based model-averaging tail area methods for different sample sizes, where $\rho = 0.5$ and outcome probability is 50%. The backward stepwise selection and Occam's window means are based on 5,000 simulations, the results obtained from inclusion fraction are based on 1,000 simulations.

N	β	E-MATA	B-MATA	I-MATA
100	0	0.04	0.021	-0.004
	0.01	0.01	0.01	0.01
	-0.2	-0.37	-0.05	-0.20
	0.5	1.06	0.46	0.51
	-1	-1.04	-1.05	-1.07
300	0	0.03	0.04	0.01
	0.01	0.05	0.04	0.02
	-0.2	-0.38	-0.05	-0.18
	0.5	0.71	0.45	0.50
	-1	-1.00	-1.00	-1.02
500	0	0.01	0.02	0.001
	0.01	0.02	0.03	0.01
	-0.2	-0.40	-0.19	-0.20
	0.5	0.59	0.51	0.49
	-1	-1.00	-1.00	-1.01

According to the results in Table 5.2.C, only FMA that was based on the candidate models selected by the inclusion fraction provided reliable point estimates, regardless of the sample size. Frequentist model-averaged estimates after backward elimination demonstrated smaller bias than estimates from the stepwise AIC backward elimination from Table 5.2.A; however, this bias was still significantly larger than the bias produced by the inclusion fraction method. The model-averaged estimates based on the B-MATA approach provided slightly biased estimates for moderate effect of 0.2 in the smallest sample size (100 subjects). However, this bias noticeably reduced as sample size increased.

Table 5.2.D: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different sample sizes, where $\rho = 0.5$ and outcome probability is 50%. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
N	β	TRUE	FULL	STEP-AIC	STEP-BIC
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
100	0	—	94.6 (2.7, 2.7) 1.12	68.8 (15.2, 16.1) 1.13	0.0 (48.1, 51.9) 1.11
	0.01	94.5 (3.0, 2.5) 1.03	94.8 (2.9, 2.3) 1.06	68.5 (15.4, 16.0) 1.06	0.0 (51.5, 48.5) 1.07
	-0.2	95.4 (3.0, 1.6) 2.50	95.2 (3.1, 1.7) 2.5	63.3 (23.3, 13.4) 2.41	17.2 (35.2, 47.6) 2.34
	0.5	96.1 (1.8, 2.1) 2.48	96.0 (1.8, 2.2) 2.51	86.5 (8.0, 5.5) 2.37	79.8 (17.0, 3.2) 2.37
	-1	95.5 (2.6, 1.9) 1.33	95.1 (2.7, 2.2) 1.39	93.9 (3.2, 2.9) 1.20	95.1 (2.7, 2.2) 1.10
300	0	—	95.5 (1.7, 2.8) 0.66	60.0 (24.0, 15.9) 0.63	0.0 (73.2, 26.8) 0.62
	0.01	95.0 (2.5, 2.5) 0.62	95.0 (2.7, 2.3) 0.63	63.2 (21.1, 15.8) 0.60	0.0 (77.2, 22.8) 0.59
	-0.2	94.8 (2.6, 2.6) 1.37	94.5 (2.7, 2.8) 1.38	73.0 (16.4, 10.6) 1.31	29.8 (25.1, 45.1) 1.32
	0.5	95.0 (2.2, 2.8) 1.43	95.0 (2.0, 3.0) 1.47	94.5 (5.1, 0.4) 1.31	89.8 (10.0, 0.1) 1.27
	-1	94.0 (2.5, 3.5) 0.72	94.4 (2.5, 3.1) 0.74	92.0 (4.8, 3.2) 0.67	89.0 (8.8, 2.2) 0.61
500	0	—	94.7 (2.6, 2.7) 0.51	60.7 (19.3, 20.0) 0.48	0.0 (58.7, 41.3) 0.46
	0.01	95.5 (2.7, 1.8) 0.48	95.2 (2.5, 2.3) 0.51	62.1 (19.0, 18.9) 0.48	0.0 (71.5, 28.5) 0.46
	-0.2	94.9 (2.5, 2.6) 0.92	95.0 (2.8, 2.2) 0.94	84.8 (5.5, 9.7) 0.89	51.1 (2.0, 46.8) 0.88
	0.5	96.4 (1.6, 2.0) 0.96	96.0 (1.7, 2.3) 0.98	96.1 (3.8, 0.1) 0.93	94.3 (5.7, 0.0) 0.92
	-1	95.4 (2.8, 1.8) 0.56	95.6 (2.5, 1.9) 0.57	92.6 (4.3, 3.0) 0.53	87.3 (9.8, 2.9) 0.48

Empirical coverage of confidence interval procedures

In general, not all methods showed improved performance with increasing sample size, because some of the methods yielded serious decreases in coverage probability with increasing sample size. The effect of a coefficient's magnitude on coverage probabilities also varied among the methods.

Table 5.2.E: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different sample sizes, where $\rho = 0.5$ and outcome probability is 50%. The LASSO results are based on 10,000 simulations, the results of zero-corrected and Bayesian approaches are based on 5,000 simulations.

N	β	ZERO-C	LASSO	BMA-W
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
100	0	100.0 (0.0, 0.0) 1.17	88.1 (6.2, 5.7) 1.12	99.5 (0.3, 0.2) 0.51
	0.01	74.0 (0.3, 25.7) 1.10	87.8 (6.8, 5.5) 1.06	99.3 (0.4, 0.4) 0.47
	-0.2	82.2 (17.8, 0.0) 3.03	88.3 (6.8, 4.8) 2.38	99.1 (0.4, 0.4) 1.09
	0.5	93.0 (1.1, 5.9) 3.04	86.4 (7.1, 6.5) 2.47	46.1 (0.6, 53.4) 1.27
	-1	99.7 (0.2, 0.1) 0.61	91.8 (5.9, 2.3) 1.22	92.9 (5.8, 1.3) 1.30
300	0	79.7 (0.0, 20.3) 0.58	87.3 (7.3, 5.4) 0.648	99.9 (0.1, 0.0) 0.24
	0.01	79.7 (0.0, 20.3) 0.58	87.1 (7.8, 5.1) 0.618	99.8 (0.2, 0.0) 0.23
	-0.2	89.3 (10.3, 0.4) 1.32	89.9 (4.9, 5.2) 1.317	39.9 (60.0, 0.1) 0.50
	0.5	94.9 (2.0, 3.1) 1.44	91.1 (6.0, 2.9) 1.415	40.0 (0.9, 59.1) 0.86
	-1	93.9 (1.2, 4.9) 0.78	90.0 (7.2, 2.8) 0.677	93.8 (4.9, 1.3) 0.68
500	0	99.8 (0.0, 0.2) 0.47	88.5 (6.1, 5.4) 0.50	99.6 (0.2, 0.1) 0.20
	0.01	76.2 (0.1, 23.7) 0.46	89.3 (5.6, 5.1) 0.50	99.6 (0.3, 0.1) 0.21
	-0.2	91.2 (8.1, 0.7) 0.86	92.4 (3.3, 4.3) 0.91	40.1 (59.5, 0.4) 0.44
	0.5	96.0 (1.8, 2.2) 1.00	94.9 (4.4, 0.8) 0.97	55.3 (1.6, 43.2) 0.87
	-1	95.0 (1.5, 3.5) 0.58	91.5 (5.9, 2.6) 0.54	93.0 (5.3, 1.7) 0.54

Among the methods presented in Tables 5.2.D to 5.2.E, only the full model provided valid confidence intervals. The stepwise BIC selection significantly underestimated the coverage probabilities for all sample sizes, and the AIC penalty provided higher coverage probabilities than BIC, but still far below the nominal level.

Table 5.2.F: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of three model-averaging CI construction methods for different sample sizes using set of candidate models obtained from backward AIC selection approach for 95% nominal level based on 5,000 simulations, where $\rho = 0.5$ and outcome probability is 50%; Wald based E-MATA-W, profile-likelihood based E-MATA-PL, and score function based E-MATA-S.

N	β	E-MATA-W	E-MATA-PL	E-MATA-S
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
100	0	68.8 (17.0, 14.2) 1.14	63.1 (20.0, 16.9) 1.14	64.4 (19.1, 16.5) 1.11
	0.01	70.8 (14.4, 14.9) 1.08	63.7 (18.4, 17.9) 1.08	65.1 (17.9, 17.0) 1.05
	-0.2	68.3 (21.6, 10.1) 2.45	62.4 (25.2, 12.3) 2.47	64.1 (24.2, 11.7) 2.38
	0.5	90.0 (5.6, 4.4) 2.39	87.8 (7.7, 4.6) 2.39	88.9 (6.5, 4.6) 2.33
	-1	95.2 (2.8, 2.0) 1.22	94.5 (2.4, 3.1) 1.21	94.5 (3.2, 2.3) 1.19
300	0	68.2 (21.2, 10.6) 0.64	66.6 (22.2, 11.2) 0.64	66.9 (22.1, 10.9) 0.63
	0.01	69.6 (17.6, 12.8) 0.61	68.3 (18.6, 13.1) 0.61	68.4 (18.5, 13.1) 0.60
	-0.2	75.5 (18.2, 6.3) 1.35	75.0 (18.2, 6.8) 1.35	75.1 (18.2, 6.7) 1.34
	0.5	96.2 (3.2, 0.5) 1.33	96.0 (3.5, 0.5) 1.33	96.2 (3.2, 0.5) 1.32
	-1	93.6 (4.2, 2.2) 0.68	93.5 (3.9, 2.6) 0.68	93.3 (4.4, 2.3) 0.68
500	0	66.1 (20.4, 13.5) 0.48	65.0 (21.0, 14.0) 0.48	65.4 (20.9, 13.8) 0.48
	0.01	66.2 (17.0, 16.9) 0.48	65.1 (17.4, 17.4) 0.48	65.5 (17.3, 17.2) 0.48
	-0.2	87.5 (5.8, 6.7) 0.90	87.4 (5.8, 6.8) 0.90	87.4 (5.8, 6.8) 0.90
	0.5	97.3 (2.7, 0.1) 0.94	97.0 (3.0, 0.1) 0.94	97.2 (2.8, 0.1) 0.93
	-1	93.5 (4.0, 2.6) 0.54	93.3 (3.7, 3.0) 0.54	93.3 (4.1, 2.6) 0.54

The zero-corrected backward selection method and the LASSO approach presented in Table 5.2.E also demonstrated unsatisfactory results; coverage probability improved with sample size, but did not reach the nominal level. The Wald-type BMA provided the worst coverage probabilities out of all evaluated methods. Even with the increase of sample size, BMA-W could not reach the desired range.

Model-averaging after backward selection (E-MATA) presented in Table 5.2.F also produced overoptimistic results. This indicates that model-averaging that follows the stepwise backward selection cannot improve the results and produce reliable intervals.

Table 5.2.G: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different sample sizes using set of candidate models obtained from Occam's window approach for 95% nominal level based on 5,000 simulations, where $\rho = 0.5$ and outcome probability is 50%; Wald based B-MATA-W, profile-likelihood based B-MATA-PL, score function based B-MATA-S, Wald based method corrected by the profile-likelihood B-MATA-Wpl, and Wald based method corrected by the score function B-MATA-Ws.

N	β	B-MATA-W	B-MATA-PL	B-MATA-S	B-MATA-Wpl	B-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
100	0	95.5 (2.3, 2.1) 1.09	94.7 (2.9, 2.5) 1.09	94.9 (2.7, 2.4) 1.06	94.9 (2.7, 2.4) 1.10	95.2 (2.5, 2.3) 1.07
	0.01	95.7 (2.2, 2.1) 1.02	94.5 (2.7, 2.8) 1.02	94.8 (2.7, 2.5) 0.99	94.7 (2.7, 2.6) 1.03	95.0 (2.6, 2.4) 1.00
	-0.2	94.6 (3.7, 1.6) 2.25	93.3 (4.5, 2.2) 2.25	93.6 (4.3, 2.1) 2.19	93.7 (4.2, 2.0) 2.28	94.0 (4.0, 2.0) 2.21
	0.5	94.6 (1.4, 4.0) 2.28	93.6 (2.2, 4.2) 2.21	93.8 (1.8, 4.4) 2.16	94.1 (2.1, 3.9) 2.24	94.2 (1.7, 4.2) 2.18
	-1	95.5 (2.6, 1.9) 1.29	94.6 (2.3, 3.1) 1.29	94.9 (2.9, 2.2) 1.26	94.9 (2.1, 3.0) 1.31	95.2 (2.6, 2.2) 1.27
300	0	94.5 (3.5, 1.9) 0.61	94.2 (3.7, 2.1) 0.61	94.2 (3.7, 2.0) 0.60	94.3 (3.7, 2.0) 0.61	94.3 (3.7, 2.0) 0.60
	0.01	95.0 (2.8, 2.2) 0.58	94.8 (3.0, 2.3) 0.58	94.8 (2.9, 2.3) 0.57	94.9 (2.9, 2.2) 0.58	95.0 (2.8, 2.2) 0.58
	-0.2	90.4 (8.1, 1.5) 1.22	89.9 (8.5, 1.6) 1.22	90.2 (8.3, 1.6) 1.21	90.1 (8.4, 1.5) 1.23	90.3 (8.2, 1.5) 1.22
	0.5	94.1 (1.4, 4.6) 1.26	93.9 (1.6, 4.5) 1.26	93.7 (1.4, 4.8) 1.25	94.0 (1.6, 4.4) 1.26	93.9 (1.4, 4.7) 1.25
	-1	94.9 (3.0, 2.1) 0.71	94.8 (2.6, 2.6) 0.71	94.8 (3.0, 2.2) 0.70	94.9 (2.5, 2.6) 0.71	94.8 (3.0, 2.2) 0.70
500	0	90.9 (5.4, 3.7) 0.46	90.5 (5.7, 3.7) 0.46	90.7 (5.6, 3.7) 0.45	90.7 (5.6, 3.7) 0.46	90.8 (5.5, 3.7) 0.45
	0.01	90.7 (4.5, 4.8) 0.46	90.5 (4.6, 4.9) 0.46	90.5 (4.6, 4.9) 0.46	90.5 (4.6, 4.9) 0.46	90.6 (4.6, 4.9) 0.46
	-0.2	90.4 (7.1, 2.5) 0.88	89.9 (7.5, 2.5) 0.87	90.1 (7.3, 2.5) 0.87	90.1 (7.4, 2.5) 0.88	90.2 (7.3, 2.5) 0.87
	0.5	97.3 (2.0, 0.6) 0.93	97.1 (2.3, 0.6) 0.93	97.2 (2.1, 0.6) 0.92	97.1 (2.2, 0.6) 0.93	97.3 (2.1, 0.6) 0.93
	-1	94.3 (3.2, 2.4) 0.55	94.2 (3.0, 2.9) 0.55	94.2 (3.3, 2.5) 0.55	94.2 (2.9, 2.8) 0.55	94.2 (3.3, 2.5) 0.55

Table 5.2.H: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different sample sizes using set of candidate models obtained from 50% inclusion fraction approach for 95% nominal level based on 1,000 simulations, where $\rho = 0.5$ and outcome probability is 50%; Wald based I-MATA-W, profile-likelihood based I-MATA-PL, score function based I-MATA-S, Wald based method corrected by the profile-likelihood I-MATA-Wpl, and Wald based method corrected by the score function I-MATA-Ws.

N	β	I-MATA-W	I-MATA-PL	I-MATA-S	I-MATA-Wpl	I-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
100	0	94.6 (2.9, 2.5) 1.12	93.6 (3.3, 3.1) 1.11	93.7 (3.2, 3.1) 1.08	93.7 (3.2, 3.1) 1.13	93.9 (3.1, 3.0) 1.09
	0.01	95.4 (2.6, 2.0) 1.05	94.3 (3.1, 2.6) 1.05	94.8 (3.0, 2.2) 1.01	94.7 (2.8, 2.5) 1.06	95.1 (2.8, 2.1) 1.03
	-0.2	94.3 (3.8, 1.9) 2.40	93.4 (4.2, 2.4) 2.40	93.7 (4.1, 2.2) 2.33	93.5 (4.1, 2.4) 2.42	93.8 (4.0, 2.2) 2.35
	0.5	95.8 (1.9, 2.3) 2.37	95.0 (2.6, 2.4) 2.37	95.3 (2.3, 2.4) 2.30	95.2 (2.6, 2.2) 2.40	95.5 (2.1, 2.4) 2.33
	-1	95.3 (2.7, 2.0) 1.33	93.9 (2.6, 3.5) 1.33	94.0 (3.3, 2.7) 1.29	94.2 (2.5, 3.3) 1.34	94.4 (3.0, 2.6) 1.31
300	0	95.2 (2.0, 2.8) 0.65	95.0 (2.1, 2.9) 0.65	95.1 (2.0, 2.9) 0.64	95.1 (2.0, 2.9) 0.65	95.1 (2.0, 2.9) 0.64
	0.01	94.8 (3.0, 2.2) 0.62	94.4 (3.1, 2.5) 0.62	94.4 (3.1, 2.5) 0.61	94.4 (3.1, 2.5) 0.62	94.7 (3.0, 2.3) 0.61
	-0.2	94.4 (3.1, 2.5) 1.34	94.0 (3.3, 2.7) 1.34	94.0 (3.3, 2.7) 1.32	94.0 (3.3, 2.7) 1.34	94.1 (3.3, 2.6) 1.33
	0.5	93.8 (2.4, 3.8) 1.38	93.6 (2.8, 3.6) 1.38	93.8 (2.4, 3.8) 1.37	93.6 (2.8, 3.6) 1.39	93.8 (2.4, 3.8) 1.38
	-1	94.4 (2.7, 2.9) 0.72	93.9 (2.5, 3.6) 0.72	93.8 (3.0, 3.2) 0.71	94.0 (2.4, 3.6) 0.72	94.1 (2.9, 3.0) 0.72
500	0	94.0 (3.1, 2.9) 0.49	93.7 (3.3, 3.0) 0.49	93.8 (3.3, 2.9) 0.49	93.9 (3.2, 2.9) 0.50	93.9 (3.2, 2.9) 0.49
	0.01	95.1 (2.8, 2.1) 0.49	95.0 (2.9, 2.1) 0.49	95.0 (2.9, 2.1) 0.49	95.0 (2.9, 2.1) 0.49	95.1 (2.8, 2.1) 0.49
	-0.2	94.8 (2.9, 2.3) 0.92	94.7 (2.9, 2.4) 0.92	94.7 (2.9, 2.4) 0.91	94.7 (2.9, 2.4) 0.92	94.7 (2.9, 2.4) 0.92
	0.5	95.9 (1.5, 2.6) 0.96	95.8 (1.7, 2.5) 0.96	95.8 (1.5, 2.7) 0.96	95.9 (1.6, 2.5) 0.96	95.8 (1.5, 2.7) 0.96
	-1	95.1 (2.7, 2.2) 0.56	94.9 (2.7, 2.4) 0.55	95.1 (2.7, 2.2) 0.55	95.0 (2.6, 2.4) 0.56	95.1 (2.8, 2.1) 0.55

The Occam's window based model-averaged tail area methods presented in Table 5.2.G showed quite unusual results. By looking just at the coverage performance of these methods for the smallest sample size, we may mistakenly conclude that methods provide valid coverage regardless of the magnitude of the coefficients. For the small sample, Occam's window approach was less conservative and left more models in a candidate set, but with an increase in sample size the parsimony principle used in Occam's window method is more likely to eliminate important variables or models, which may lead to undercoverage.

As sample size increases to 500, all five methods demonstrate undercoverage with around 90% coverage probability for covariates X_1 to X_3, overcoverage with around 97% coverage probability for X_4, while the desired range of (93.6%, 96.4%) was reached only for X_5 that had the largest effect among the considered variables. Overall, in up to 66.7% of the time, these methods reached the desirable range.

The methods presented in Table 5.2.H as well as the full model provided valid coverage probabilities. All methods, except for the profile-likelihood based method, provided the coverage probabilities falling within the desired range 100% of the time, regardless of the sample size and the magnitude of coefficient. The profile-likelihood based I-MATA-PL and I-MATA-Wpl failed to construct valid confidence interval for X_3, when sample size was 100.

Tail errors

Out of the 19 approaches compared, only five inclusion fraction based model-averaged tail area methods and the full model passed the first comparison stage, which means that there is no need to discuss and compare tail errors for the other 13 methods. Upper and lower tail errors from full models (Table 5.2.D) usually are closer to each other, compared to those from the model-averaged methods (Table 5.2.H). However, the proportion of unstable tail errors is the same for all considered methods, indicating that all six approaches are acceptable in terms of balanced tail errors.

Average width

For the methods that passed the two previous tests by providing accurate point estimates and valid confidence intervals, we compare their averaged widths in Figure 5.1. The method that provides valid intervals with narrowest width is preferable, since narrowest width reflects the best accuracy and efficiency. There were five MATA methods based on inclusion fraction and the full model that have passed this stage.

First, the five inclusion fraction based model-averaged approaches were compared to the full model. The full model provided larger average widths than all the model-averaged methods except for the Wald based averaging method with profile-likelihood standard errors (I-MATA-Wpl) regardless of the sample size and effect size. This implies that the full model and I-MATA-Wpl cannot be considered as the best methods and we exclude them from further analysis.

Of the four remaining methods, only the two score-based ones, I-MATA-S and I-MATA-Ws, produced tight intervals for all variables and sample sizes. The profile-likelihood based method provided slightly smaller widths than the Wald based MATA, and the difference between them reduced with sample size. The I-MATA-S approach demonstrated the best results in terms of the averaged width that was up to 8.3% smaller than the width from the full model, and up to 3.2% smaller than the Wald based MATA width.

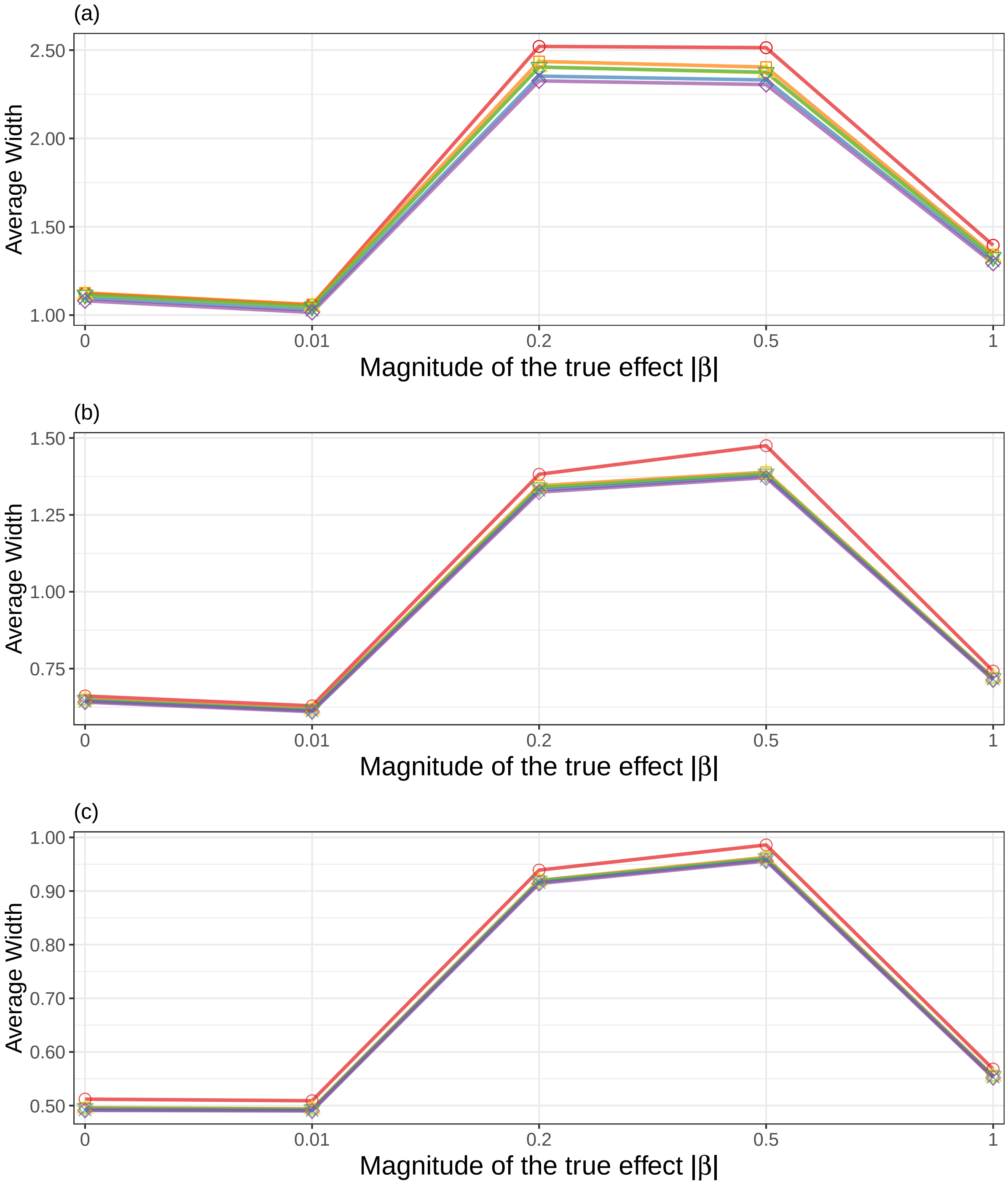


Figure 5.1: Comparison of averaged widths of the full model- ○ , I-MATA-Wpl - □ , I-MATA-PL - △ , I-MATA-W - ▽ , I-MATA-Ws - × , I-MATA-S - ◇ for sample sizes: (a) - N=100, (b) - N=300, and (c) - N=500.

5.7.2 Number of predictors

The number of predictors is partially controlled by researchers through the study objectives and prior knowledge. We evaluated the performance of the methods on the increase in the number of covariates. The average point estimates are presented in Tables 5.3.A to 5.3.C. The empirical coverage probabilities, tail errors and average width for all considered approaches are presented in Tables 5.3.D to 5.3.H that illustrate the combinations 4-6 in Table 5.1. Comparison of the valid methods by their widths is presented in Figure 5.2. The methods were compared with 3, 5, and 10 variables, while the sample size was fixed at 500, the correlation at 0.5, and the outcome probability at 0.3.

The small sample concept in logistic regression is highly related to EPV and ability of the logistic regression to converge. The sample size of 500 was selected in order to see how the methods perform under different recommended EPV values. For EPV equal to 10 and 0.3 outcome probability the smallest necessary sample sizes for a model with 10 variables is $(10 \times 10)/0.3 \approx 334$, while for EPV of 20 and 50, the sample size should be increased to 667 and 1667. This means that a sample size of 500 should be enough to get valid reliable results if we use EPV=10 as recommended by Peduzzi et al. (1996). However, according to Vittinghoff and McCulloch (2007) or Steyerberg et al. (1999), a sample size of 500 is too small and inference based on such a small sample might not be reliable.

Bias of point estimates

Comparing the means of point estimates produced for models with different complexity, we can see a picture quite similar to the one we observed in previous sections. The full model demonstrated good results, while the estimates provided by the stepwise selection methods from Table 5.3.A were mostly biased away from zero.

Table 5.3.A: Mean of point estimates obtained from the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different number of covariates, where N=500, $\rho = 0.5$ and outcome probability is 30%. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
p	β	TRUE	FULL	STEP-AIC	STEP-BIC
3	0.01	0.01	0.01	0.08	0.25
	0.5	0.50	0.50	0.62	0.78
	-1	-1.01	-1.01	-1.00	-0.95
5	0	—	-0.002	0.001	0.11
	0.01	0.01	0.01	0.03	0.11
	-0.2	-0.19	-0.19	-0.45	-0.67
	0.5	0.49	0.49	0.63	0.78
	-1	-1.01	-1.01	-1.01	-0.96
10	0	—	0.002	0.002	-0.004
	0	—	-0.003	0.000	-0.005
	0	—	0.002	0.006	0.15
	0	—	-0.01	0.001	0.07
	0.01	0.01	0.01	0.028	-0.03
	-0.2	-0.20	-0.20	-0.337	-0.46
	0.5	0.51	0.52	0.735	0.95
	-0.7	-0.72	-0.72	-0.726	-0.73
	-1	-1.03	-1.04	-1.03	-1.00
	2.5	2.57	2.60	2.57	2.54

Both LASSO and zero-corrected method from Table 5.3.B demonstrated moderate bias. While the LASSO showed quite consistent bias away from zero, the zero-corrected backward selection demonstrated bias towards zero for relatively small effects (<0.7) and bias away from zero for larger effects. The Wald-type BMA method was not able to provide accurate point estimates for most of the effects; however, for the effects with magnitude above 0.7 the bias was relatively small.

Table 5.3.B: Mean of point estimates obtained from the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different number of covariates, where N=500, $\rho = 0.5$ and outcome probability is 30%. The LASSO results are based on 10,000 simulations, the results of zero-corrected backward selection and Wald based Bayesian model-averaging are based on 5,000 simulations.

p	β	ZERO-C	LASSO	BMA-W
3	0.01	0.01	0.03	0.02
	0.5	0.46	0.57	0.29
	-1	-1.01	-1.01	-0.96
5	0	0.001	-0.01	0.01
	0.01	0.01	-0.02	0.01
	-0.2	-0.15	-0.31	-0.03
	0.5	0.45	0.64	0.23
	-1	-1.03	-0.99	-0.97
10	0	0.00	0.05	0.001
	0	-0.004	-0.05	-0.002
	0	0.004	-0.03	0.01
	0	-0.004	0.01	0.01
	0.01	0.003	-0.02	-0.001
	-0.2	-0.17	-0.26	-0.06
	0.5	0.46	0.65	0.18
	-0.7	-0.74	-0.72	-0.70
	-1	-1.07	-1.02	-1.01
	2.5	2.67	2.59	2.54

Out of all methods presented in Table 5.3.C only the model-averaging methods based on inclusion fraction and Occam's window provided point estimates with low bias. Clearly, the true model has precision that is hard to outperform; however, the inclusion fraction approach provided less biased point estimates than the full model. Despite the fact that increasing number of variables reduced the accuracy of I-MATA, its averaged point estimates are still very close to the true effects.

Table 5.3.C: Mean of point estimates of the backward stepwise selection (E-MATA), Occam's window (B-MATA) and inclusion fraction (I-MATA) based model-averaging tail area methods for different number of covariates, where N=500, $\rho = 0.5$ and outcome probability is 30%. The backward stepwise selection and Occam's window means are based on 5,000 simulations, the results obtained from inclusion fraction are based on 1,000 simulations.

p	β	E-MATA	B-MATA	I-MATA
3	0.01	0.23	0.07	0.01
	0.5	0.61	0.53	0.50
	-1	-0.99	-0.99	-1.00
5	0	-0.01	0.03	-0.001
	0.01	0.03	0.02	0.01
	-0.2	-0.42	-0.16	-0.20
	0.5	0.62	0.51	0.48
	-1	-1.00	-1.00	-1.01
10	0	-0.01	-0.002	0.002
	0	-0.01	-0.01	-0.004
	0	0.02	0.02	0.01
	0	0.02	0.02	-0.01
	0.01	0.03	0.001	0.003
	-0.2	-0.32	-0.21	-0.19
	0.5	0.71	0.52	0.50
	-0.7	-0.72	-0.72	-0.72
	-1	-1.02	-1.02	-1.03
	2.5	2.56	2.55	2.59

Confidence interval coverage

Comparing the methods, we can see that the pattern is quite similar to the one we observed in Tables 5.2.D to 5.2.H. Out of the methods presented in Tables 5.3.D to 5.3.F, the full model provided valid coverage for 94.4% of the estimated effects; it only failed to yield a proper coverage level for the effect of variable X_5 in the 10-variables model. Moreover, even the true model could not provide a valid coverage rate for this variable, which can be related to the relatively small sample size. All other methods - stepwise backward selection, zero-corrected backward selection, LASSO, and Wald-type BMA - showed unsatisfactory performance with at least 55.6% of effects whose coverage falling outside of the desired range, and usually underestimating the coverage rate.

In this group of simulations, the sample size was fixed at its highest value of 500, because otherwise the logistic model with 10 variables could have convergence problems. We have already observed in Table 5.2.G that the coverage of model-averaged methods based on the Occam's window is unacceptable for sample sizes larger than 100. The results presented in Table 5.3.G are very similar, since only up to 38.9% of the effects had the empirical coverage within the desired range, which makes this set of method unacceptable.

All the methods presented in Table 5.3.H provided very good coverage results with 94.4% of the time within the desired range of (93.6%, 96.4%). As in the full model, these methods failed to provide valid coverage rates for variable X_5, which is smallest non-zero effect (0.01) in this model. However, even with this failure, five I-MATA based methods and the full model outperformed other approaches. Thus, only these six methods are further compared.

Tail errors

We have already seen how well inclusion fraction based methods and the full model performed for different sample sizes. In this group of simulations, the validity of confidence intervals, which is the primary requirement for confidence intervals, held regardless of the increase in the number of variables. Nevertheless, the confidence intervals provided by these methods were not perfectly balanced, and we could distinguish two different patterns.

The full model, I-MATA-PL, and I-MATA-Wpl showed well-balanced errors for models with 3 and 5 variables but did not perform well for a model with 10 variables. The methods based on the Wald or score function from Table 5.3.H provided two unbalanced errors in the 3-variables model, but they did relatively well in 5 and 10 variables models. All the approaches provided six or seven unbalanced tail errors.

Table 5.3.D: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different number of covariates, where N=500, $\rho = 0.5$ and outcome probability is 30%. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
p	β	TRUE	FULL	STEP-AIC	STEP-BIC
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
3	0.01	94.9 (2.6, 2.5) 0.51	94.9 (2.6, 2.5) 0.51	64.1 (23.6, 12.3) 0.50	0.0 (85.0, 15.0) 0.50
	0.5	94.3 (2.5, 3.2) 1.02	94.3 (2.5, 3.2) 1.02	95.9 (4.0, 0.0) 0.97	91.9 (8.1, 0.0) 0.97
	-1	96.2 (1.9, 1.9) 0.55	96.2 (1.9, 1.9) 0.55	92.5 (5.0, 2.5) 0.52	85.6 (12.4, 1.9) 0.50
5	0	—	95.3 (2.8, 1.9) 0.53	64.2 (19.4, 16.4) 0.51	0.0 (66.3, 33.7) 0.49
	0.01	95.1 (2.7, 2.2) 0.53	95.0 (2.8, 2.2) 0.54	63.6 (18.4, 18.0) 0.51	0.0 (64.9, 35.1) 0.50
	-0.2	95.7 (2.1, 2.2) 1.06	96.2 (1.5, 2.3) 1.09	81.0 (8.9, 10.1) 1.02	37.6 (6.2, 56.2) 1.02
	0.5	95.2 (1.4, 3.4) 1.07	95.2 (1.9, 2.9) 1.10	96.1 (3.8, 0.1) 1.02	93.7 (6.3, 0.0) 1.00
	-1	95.5 (1.8, 2.7) 0.59	95.0 (2.5, 2.5) 0.60	92.3 (4.8, 2.8) 0.55	87.8 (9.8, 2.4) 0.50
10	0	—	95.4 (1.9, 2.7) 0.65	63.4 (18.4, 18.2) 0.63	0.0 (50.0, 50.0) 0.63
	0	—	94.6 (2.2, 3.2) 0.65	65.2 (17.3, 17.5) 0.63	0.0 (49.1, 50.9) 0.63
	0	—	95.0 (2.4, 2.6) 0.66	67.8 (16.7, 15.6) 0.65	0.0 (65.6, 34.4) 0.66
	0	—	94.8 (2.0, 3.2) 0.70	65.9 (17.4, 16.7) 0.68	0.0 (56.8, 43.2) 0.67
	0.01	92.3 (4.0, 3.7) 0.65	92.8 (3.4, 3.8) 0.67	66.7 (15.1, 18.2) 0.65	0.0 (46.6, 53.4) 0.64
	-0.2	94.7 (2.7, 2.6) 0.62	94.7 (3.0, 2.3) 0.64	92.2 (1.1, 6.7) 0.61	77.5 (0.5, 22.0) 0.61
	0.5	95.2 (2.9, 1.9) 1.25	94.4 (3.3, 2.3) 1.32	94.3 (5.3, 0.4) 1.26	88.8 (11.2, 0.0) 1.25
	-0.7	95.1 (2.7, 2.2) 0.67	94.7 (2.3, 3.0) 0.69	94.3 (2.2, 3.5) 0.66	96.3 (0.6, 3.2) 0.64
	-1	95.8 (1.6, 2.6) 0.71	95.4 (1.5, 3.1) 0.75	93.7 (2.8, 3.5) 0.70	93.5 (3.7, 2.8) 0.66
	2.5	94.8 (3.5, 1.7) 0.96	94.9 (4.1, 1.0) 1.00	94.4 (3.7, 2.0) 0.95	94.7 (2.7, 2.6) 0.92

Table 5.3.E: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different number of covariates, where N=500, $\rho = 0.5$ and outcome probability is 30%. The LASSO results are based on 10,000 simulations, the results of zero-corrected and Bayesian approaches are based on 5,000 simulations.

p	β	ZERO-C	LASSO	BMA-W
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
3	0.01	73.6 (0.2, 26.2) 0.46	91.9 (4.5, 3.6) 0.51	99.5 (0.4, 0.0) 0.25
	0.5	94.2 (3.2, 2.6) 1.02	96.2 (3.3, 0.5) 1.00	59.7 (2.5, 37.8) 0.95
	-1	95.9 (1.9, 2.2) 0.57	93.0 (4.5, 2.5) 0.54	91.2 (7.3, 1.5) 0.54
5	0	99.8 (0.1, 0.1) 0.48	89.3 (5.7, 5.0) 0.52	99.7 (0.1, 0.1) 0.22
	0.01	73.2 (0.1, 26.7) 0.48	89.2 (5.9, 4.9) 0.53	99.8 (0.2, 0.1) 0.22
	-0.2	89.5 (9.9, 0.6) 1.00	91.5 (4.0, 4.5) 1.05	41.8 (57.8, 0.4) 0.48
	0.5	95.0 (2.3, 2.7) 1.09	94.4 (4.3, 1.2) 1.07	49.8 (1.5, 48.7) 0.87
	-1	95.2 (1.3, 3.5) 0.62	91.0 (6.7, 2.4) 0.56	93.5 (5.2, 1.2) 0.56
10	0	99.9 (0.1, 0.0) 0.60	89.5 (5.9, 4.7) 0.64	99.7 (0.2, 0.1) 0.26
	0	100.0 (0.0, 0.0) 0.61	90.0 (4.9, 5.1) 0.64	99.7 (0.1, 0.2) 0.25
	0	99.9 (0.1, 0.0) 0.62	89.4 (5.4, 5.2) 0.66	99.7 (0.2, 0.1) 0.27
	0	99.7 (0.0, 0.3) 0.65	88.7 (5.9, 5.4) 0.69	99.6 (0.3, 0.1) 0.28
	0.01	73.6 (0.2, 26.2) 0.63	89.1 (5.4, 5.5) 0.66	99.6 (0.2, 0.2) 0.27
	-0.2	95.1 (3.6, 1.3) 0.62	93.6 (2.4, 4.1) 0.63	39.6 (59.5, 0.9) 0.38
	0.5	94.3 (3.8, 1.9) 1.31	93.5 (4.4, 2.1) 1.29	41.7 (1.5, 56.7) 0.89
	-0.7	93.9 (1.5, 4.6) 0.75	94.2 (2.6, 3.2) 0.67	94.3 (3.0, 2.6) 0.69
	-1	93.7 (0.6, 5.7) 0.79	93.8 (3.1, 3.1) 0.72	95.3 (2.7, 2.0) 0.69
	2.5	89.1 (10.5, 0.4) 1.03	94.5 (4.0, 1.5) 0.97	95.2 (2.2, 2.6) 0.94

Table 5.3.F: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of three model-averaging CI construction methods for different number of covariates using set of candidate models obtained from backward AIC selection approach for 95% nominal level based on 5,000 simulations, where N=500, $\rho = 0.5$ and outcome probability is 30%; Wald based E-MATA-W, profile-likelihood based E-MATA-PL, and score function based E-MATA-S.

p	β	E-MATA-W	E-MATA-PL	E-MATA-S
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
3	0.01	55.6 (42.5, 1.9) 0.48	54.2 (43.9, 1.9) 0.48	54.2 (43.9, 1.9) 0.48
	0.5	96.6 (3.4, 0.1) 0.96	96.2 (3.8, 0.1) 0.96	96.4 (3.5, 0.1) 0.95
	-1	92.8 (4.9, 2.3) 0.54	92.8 (4.6, 2.6) 0.54	92.7 (4.9, 2.4) 0.53
5	0	68.8 (17.2, 14.0) 0.51	67.9 (17.8, 14.3) 0.51	68.4 (17.7, 14.0) 0.51
	0.01	68.8 (14.9, 16.3) 0.52	67.3 (16.0, 16.8) 0.52	67.6 (16.0, 16.4) 0.52
	-0.2	84.5 (8.8, 6.6) 1.04	84.2 (8.8, 6.9) 1.04	84.2 (8.8, 6.9) 1.04
	0.5	96.8 (3.0, 0.2) 1.03	96.7 (3.1, 0.2) 1.03	96.8 (3.0, 0.2) 1.03
	-1	93.9 (4.0, 2.1) 0.56	93.9 (3.6, 2.5) 0.56	93.8 (4.0, 2.2) 0.56
10	0	67.1 (16.3, 16.6) 0.64	65.8 (16.7, 17.5) 0.64	66.0 (16.7, 17.2) 0.63
	0	71.7 (12.4, 16.0) 0.64	70.2 (13.0, 16.8) 0.64	70.4 (12.7, 16.8) 0.63
	0	67.6 (18.5, 13.9) 0.65	66.2 (19.5, 14.3) 0.66	66.9 (19.0, 14.1) 0.65
	0	69.5 (17.4, 13.1) 0.68	68.0 (18.7, 13.2) 0.68	68.8 (18.0, 13.2) 0.68
	0.01	69.6 (12.9, 17.5) 0.65	68.5 (13.5, 18.0) 0.65	68.9 (13.3, 17.7) 0.65
	-0.2	94.0 (1.0, 5.0) 0.62	93.7 (1.0, 5.3) 0.62	93.9 (1.0, 5.1) 0.62
	0.5	96.2 (3.4, 0.3) 1.27	96.0 (3.6, 0.3) 1.27	96.2 (3.5, 0.3) 1.26
	-0.7	94.8 (2.3, 3.0) 0.66	94.4 (2.2, 3.4) 0.66	94.6 (2.3, 3.1) 0.66
	-1	95.2 (2.2, 2.5) 0.70	94.9 (2.1, 3.0) 0.70	95.1 (2.3, 2.6) 0.70
	2.5	95.0 (2.8, 2.2) 0.96	94.5 (3.7, 1.7) 0.96	94.9 (2.9, 2.2) 0.95

Table 5.3.G: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different number of covariates using set of candidate models obtained from Occam's window approach for 95% nominal level based on 5,000 simulations, where N=500, $\rho = 0.5$ and outcome probability is 30%; Wald based B-MATA-W, profile-likelihood based B-MATA-PL, score function based B-MATA-S, Wald based method corrected by the profile-likelihood B-MATA-Wpl, and Wald based method corrected by the score function B-MATA-Ws.

p	β	B-MATA-W	B-MATA-PL	B-MATA-S	B-MATA-Wpl	B-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
3	0.01	91.5 (5.6, 2.9) 0.45	90.9 (6.0, 3.1) 0.50	91.0 (5.9, 3.0) 0.49	91.1 (5.9, 3.0) 0.50	91.2 (5.8, 3.0) 0.49
	0.5	97.1 (2.4, 0.6) 0.97	96.8 (2.7, 0.6) 0.97	97.0 (2.4, 0.6) 0.99	96.8 (2.6, 0.6) 0.97	97.0 (2.4, 0.6) 0.97
	-1	93.9 (3.8, 2.3) 0.55	93.9 (3.5, 2.6) 0.55	93.8 (3.9, 2.3) 0.54	94.0 (3.5, 2.5) 0.55	93.9 (3.8, 2.3) 0.55
5	0	92.0 (4.5, 3.5) 0.49	91.8 (4.6, 3.6) 0.49	91.9 (4.6, 3.5) 0.48	91.9 (4.6, 3.5) 0.49	92.0 (4.5, 3.5) 0.49
	0.01	91.5 (4.0, 4.5) 0.50	91.0 (4.2, 4.7) 0.50	91.2 (4.2, 4.6) 0.49	91.2 (4.2, 4.6) 0.50	91.2 (4.2, 4.6) 0.50
	-0.2	87.1 (10.2, 2.6) 0.98	86.6 (10.7, 2.7) 0.98	86.7 (10.6, 2.7) 0.97	86.8 (10.5, 2.7) 0.98	86.9 (10.4, 2.7) 0.98
	0.5	96.5 (2.0, 1.5) 1.01	96.4 (2.2, 1.5) 1.01	96.5 (2.0, 1.5) 1.01	96.4 (2.2, 1.5) 1.02	96.5 (2.0, 1.5) 1.01
	-1	94.9 (3.1, 2.1) 0.58	94.7 (2.9, 2.3) 0.58	94.8 (3.1, 2.1) 0.57	94.7 (2.9, 2.3) 0.58	94.8 (3.1, 2.1) 0.57
10	0	91.0 (4.4, 4.6) 0.62	90.6 (4.6, 4.8) 0.62	90.7 (4.6, 4.8) 0.61	90.7 (4.5, 4.8) 0.62	90.8 (4.5, 4.7) 0.61
	0	92.7 (3.3, 4.0) 0.62	92.4 (3.4, 4.2) 0.62	92.5 (3.3, 4.2) 0.61	92.5 (3.3, 4.1) 0.62	92.5 (3.3, 4.1) 0.61
	0	91.3 (5.2, 3.5) 0.64	90.8 (5.6, 3.7) 0.64	91.0 (5.4, 3.6) 0.63	90.9 (5.5, 3.6) 0.64	91.1 (5.3, 3.6) 0.63
	0	91.0 (5.3, 3.8) 0.65	90.4 (5.7, 3.8) 0.65	90.8 (5.4, 3.8) 0.65	90.7 (5.5, 3.8) 0.65	90.8 (5.4, 3.8) 0.65
	0.01	91.8 (3.3, 4.9) 0.63	91.5 (3.6, 4.9) 0.63	91.5 (3.6, 4.9) 0.63	91.5 (3.6, 4.9) 0.63	91.6 (3.5, 4.9) 0.63
	-0.2	90.9 (6.6, 2.5) 0.60	90.5 (6.7, 2.8) 0.60	90.6 (6.8, 2.7) 0.60	90.7 (6.6, 2.7) 0.60	90.7 (6.7, 2.6) 0.60
	0.5	93.8 (2.0, 4.1) 1.24	93.7 (2.2, 4.1) 1.24	93.8 (2.1, 4.1) 1.23	93.7 (2.2, 4.1) 1.24	93.8 (2.1, 4.1) 1.23
	-0.7	95.0 (2.3, 2.7) 0.67	94.5 (2.2, 3.3) 0.67	94.7 (2.4, 2.9) 0.66	94.6 (2.2, 3.3) 0.67	94.8 (2.3, 2.9) 0.66
	-1	95.2 (2.4, 2.3) 0.71	95.0 (2.1, 2.9) 0.71	95.2 (2.4, 2.4) 0.70	95.0 (2.1, 2.8) 0.71	95.2 (2.4, 2.4) 0.70
	2.5	95.2 (2.7, 2.1) 0.96	94.8 (3.5, 1.7) 0.96	95.1 (2.8, 2.1) 0.95	94.9 (3.5, 1.7) 0.96	95.1 (2.8, 2.1) 0.95

Table 5.3.H: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different number of covariates using set of candidate models obtained from 50% inclusion fraction approach for 95% nominal level based on 1,000 simulations, where N=500, $\rho = 0.5$ and outcome probability is 30%; Wald based I-MATA-W, profile-likelihood based I-MATA-PL, score function based I-MATA-S, Wald based method corrected by the profile-likelihood I-MATA-Wpl, and Wald based method corrected by the score function I-MATA-Ws.

p	β	I-MATA-W	I-MATA-PL	I-MATA-S	I-MATA-Wpl	I-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
3	0.01	94.5 (3.2, 2.3) 0.51	94.3 (3.3, 2.4) 0.51	94.4 (3.2, 2.4) 0.51	94.4 (3.2, 2.4) 0.51	94.5 (3.2, 2.3) 0.51
	0.5	94.8 (2.1, 3.1) 0.99	94.5 (2.4, 3.1) 0.99	94.7 (2.1, 3.2) 0.99	94.7 (2.4, 2.9) 0.99	94.7 (2.1, 3.2) 0.99
	-1	95.9 (2.7, 1.4) 0.54	95.6 (2.5, 1.9) 0.54	95.7 (2.8, 1.5) 0.54	95.6 (2.5, 1.9) 0.54	95.8 (2.8, 1.4) 0.54
5	0	94.7 (3.0, 2.3) 0.52	94.5 (3.1, 2.4) 0.52	94.6 (3.1, 2.3) 0.52	94.6 (3.1, 2.3) 0.52	94.6 (3.1, 2.3) 0.52
	0.01	94.5 (2.9, 2.6) 0.53	94.4 (2.9, 2.7) 0.52	94.4 (2.9, 2.7) 0.52	94.4 (2.9, 2.7) 0.53	94.5 (2.9, 2.6) 0.52
	-0.2	95.3 (2.2, 2.5) 1.05	95.3 (2.2, 2.5) 1.05	95.3 (2.2, 2.5) 1.04	95.3 (2.2, 2.5) 1.05	95.3 (2.2, 2.5) 1.05
	0.5	94.8 (1.6, 3.6) 1.06	94.8 (1.7, 3.5) 1.06	94.7 (1.6, 3.7) 1.05	94.8 (1.7, 3.5) 1.06	94.8 (1.6, 3.6) 1.06
	-1	94.7 (2.7, 2.6) 0.58	94.2 (2.5, 3.3) 0.58	94.4 (2.7, 2.9) 0.58	94.2 (2.5, 3.3) 0.59	94.6 (2.7, 2.7) 0.58
10	0	95.2 (2.4, 2.4) 0.64	94.9 (2.5, 2.6) 0.64	94.9 (2.5, 2.6) 0.63	95.0 (2.4, 2.6) 0.64	95.2 (2.4, 2.4) 0.64
	0	94.6 (2.1, 3.3) 0.64	94.5 (2.4, 3.1) 0.64	94.5 (2.3, 3.2) 0.64	94.5 (2.2, 3.3) 0.64	94.5 (2.3, 3.2) 0.64
	0	94.7 (2.7, 2.6) 0.66	94.2 (2.8, 3.0) 0.66	94.3 (2.8, 2.9) 0.65	94.2 (2.8, 3.0) 0.66	94.4 (2.8, 2.8) 0.65
	0	95.2 (2.0, 2.8) 0.69	95.1 (2.1, 2.8) 0.69	95.2 (2.0, 2.8) 0.68	95.1 (2.1, 2.8) 0.69	95.2 (2.0, 2.8) 0.68
	0.01	92.6 (3.5, 3.9) 0.66	92.3 (3.6, 4.1) 0.66	92.4 (3.5, 4.1) 0.65	92.4 (3.5, 4.1) 0.66	92.6 (3.5, 3.9) 0.65
	-0.2	94.8 (3.2, 2.0) 0.63	94.6 (3.2, 2.2) 0.63	94.7 (3.2, 2.1) 0.62	94.6 (3.2, 2.2) 0.63	94.8 (3.2, 2.0) 0.62
	0.5	94.3 (3.3, 2.4) 1.29	94.1 (3.5, 2.4) 1.29	94.3 (3.3, 2.4) 1.28	94.2 (3.4, 2.4) 1.29	94.3 (3.3, 2.4) 1.28
	-0.7	94.9 (2.4, 2.7) 0.68	94.2 (2.4, 3.4) 0.68	94.5 (2.6, 2.9) 0.68	94.3 (2.4, 3.3) 0.68	94.6 (2.6, 2.8) 0.68
	-1	95.3 (1.6, 3.1) 0.73	94.9 (1.6, 3.5) 0.73	95.1 (1.7, 3.2) 0.73	95.0 (1.6, 3.4) 0.73	95.1 (1.7, 3.2) 0.73
	2.5	95.4 (3.4, 1.2) 0.98	94.2 (4.8, 1.0) 0.98	95.3 (3.5, 1.2) 0.97	94.3 (4.7, 1.0) 0.98	95.4 (3.4, 1.2) 0.98

Only the full model provided symmetric confidence intervals out of all methods that were compared at this stage, and yet it did not outperform the model-average methods in terms of tail errors balance. Even though some of the methods provided slightly unbalanced intervals, we still compare them further.

Average width

The methods that provided valid and balanced intervals are presented in Figure 5.2. The full model usually provides wider confidence intervals than all other methods. By definition, provided intervals are always valid, which makes the full model preferable over most methods compared in this thesis. However, the I-MATA based methods also demonstrated reliable and balanced intervals, and outperformed the full model in terms of the average width, regardless of the number of variables in the model.

Regarding the inclusion fraction based model-averaging methods presented in Table 5.3.H, we observed the following width hierarchy that has been preserved for any model size. In terms of interval width, the order is MATA-S $<$ MATA-Ws $<$ MATA-PL $<$ MATA-W $<$ MATA-Wpl. Despite the fact that the difference between the methods is not very large, this relation is consistent for any number of variables in the models considered in this simulation. The number of variables increases the confidence interval width; however, the I-MATA-S algorithm demonstrated greater precision out of all tested methods.

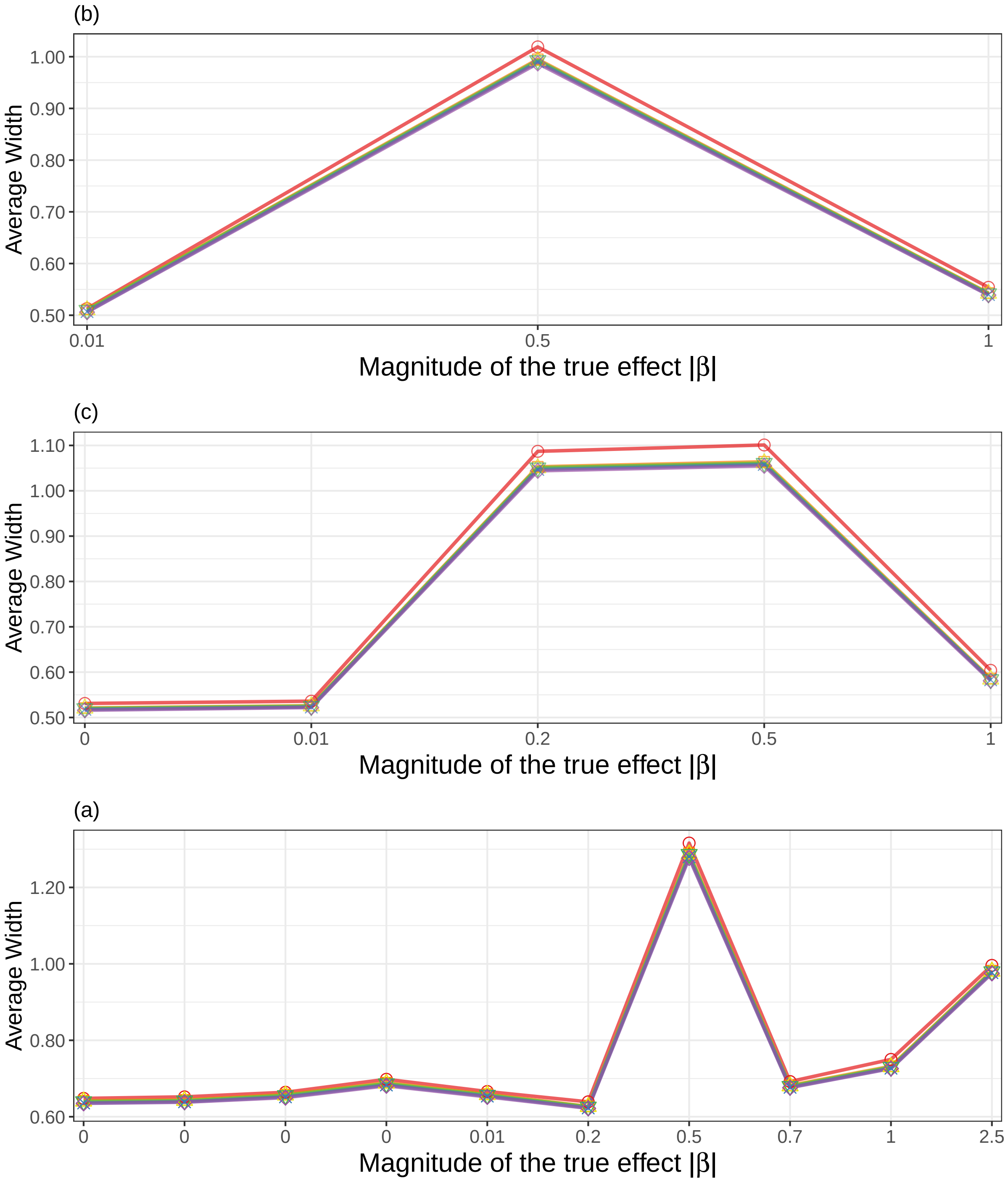


Figure 5.2: Comparison of averaged widths of the full model- ○ , I-MATA-Wpl - □ , I-MATA-PL - △ , I-MATA-W - ▽ , I-MATA-Ws - × , I-MATA-S - ◇ for number of variables: (a) - p=3, (b) - p=5, and (c) - p=10.

5.7.3 Correlation

It is known that high correlation among predictors can have negative effects on inference. We compared how methods perform under three different correlation levels among five variables: 0, 0.3 and 0.5. The probability of outcome and sample size were fixed at 0.3 and 300, respectively. The simulation results reflect combinations 7-9 from Table 5.1. Tables 5.4.A to 5.4.C present the averaged point estimates. The empirical coverage probabilities, tail errors, and average width of 19 compared approaches are presented in Tables 5.4.D to 5.4.H. Comparison of the valid methods by their widths is presented in Figure 5.3.

Bias of point estimates

As one would expect, the increase in correlation increased the bias of point estimates for the effect of any magnitude. The orientation of the bias for each method is no different from previous simulations. The stepwise selection methods from Table 5.4.A, LASSO based selection approach from Table 5.4.B and model-averaging after stepwise selection from Table 5.4.C demonstrated consistent bias away from zero for effects with magnitude below 1. These methods provided biased point estimates even for uncorrelated covariates.

The results of the Wald-type BMA from Table 5.4.B were mostly biased toward zero. This indicates that Bayesian model-averaging might not be appropriate for making frequentist inference about the effect of covariates in logistic regression if sample size is too small.

The zero-corrected method provided slightly biased point estimates; however, due to bootstrapping involved in this procedure, its point estimates are more accurate than estimates calculated by the conventional AIC backward elimination from Table 5.4.A. It can be also seen that correlation had a smaller effect on the bias of point estimates of the ZERO-

C method than on the bias of stepwise backwards selection.

Table 5.4.B: Mean of point estimates obtained from the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different correlation levels between five covariates, where N=300 and outcome probability is 30%. The LASSO results are based on 10,000 simulations, the results of zero-corrected backward selection and Wald based Bayesian model-averaging are based on 5,000 simulations.

ρ	β	ZERO-C	LASSO	BMA-W
0	0	0.002	0.02	-0.002
	0.01	0.002	0.05	-0.001
	-0.2	-0.18	-0.34	-0.05
	0.5	0.48	0.63	0.25
	-1	-1.06	-1.02	-1.01
0.3	0	0.01	0.01	0.01
	0.01	0.01	-0.003	0.01
	-0.2	-0.16	-0.32	-0.03
	0.5	0.48	0.66	0.21
	-1	-1.06	-1.02	-1.00
0.5	0	0.001	-0.01	0.01
	0.01	0.02	-0.03	0.01
	-0.2	-0.17	-0.35	-0.03
	0.5	0.46	0.73	0.20
	-1	-1.06	-0.99	-0.98

In two previous simulation groups, the averaged point estimates of the inclusion fraction MATA based methods were closer to the true effects than estimates from the full and the true models. In this set of simulations, the I-MATA based average point estimates presented in Table 5.4.C are usually closer to the true values than the true model estimates. The model-averaging based on candidate models selected by the Occam's window demonstrated good results for uncorrelated data; however, as the correlation increased, the bias became more noticable for all covariates.

Table 5.4.A: Mean of point estimates obtained from the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different correlation levels between five covariates, where N=300 and outcome probability is 30%. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
ρ	β	TRUE	FULL	STEP-AIC	STEP-BIC
0	0	—	0.002	-0.01	-0.07
	0.01	0.01	0.01	0.04	0.03
	-0.2	-0.21	-0.21	-0.51	-0.77
	0.5	0.52	0.52	0.68	0.87
	-1	-1.03	-1.04	-1.02	-1.01
0.3	0	—	0.004	0.024	0.19
	0.01	0.01	0.01	0.02	0.16
	-0.2	-0.20	-0.20	-0.49	-0.68
	0.5	0.53	0.52	0.70	0.89
	-1	-1.03	-1.03	-1.02	-0.99
0.5	0	—	-0.004	0.04	0.17
	0.01	0.02	0.02	0.03	0.20
	-0.2	-0.22	-0.22	-0.50	-0.71
	0.5	0.51	0.52	0.76	0.96
	-1	-1.03	-1.03	-1.02	-0.97

Table 5.4.C: Mean of point estimates of the backward stepwise selection (E-MATA), Occam's window (B-MATA) and inclusion fraction (I-MATA) based model-averaging tail area methods for different correlation levels between five covariates, where N=300 and outcome probability is 30%. The backward stepwise selection and Occam's window means are based on 5,000 simulations, the results obtained from inclusion fraction are based on 1,000 simulations.

ρ	β	E-MATA	B-MATA	I-MATA
0	0	-0.03	-0.01	0.002
	0.01	0.004	-0.004	0.003
	-0.2	-0.50	-0.19	-0.21
	0.5	0.69	0.51	0.52
	-1	-1.02	-1.02	-1.03
0.3	0	0.04	0.03	0.01
	0.01	0.05	0.03	0.01
	-0.2	-0.45	-0.14	-0.19
	0.5	0.69	0.49	0.52
	-1	-1.02	-1.06	-1.03
0.5	0	0.04	0.03	-0.001
	0.01	0.04	0.03	0.02
	-0.2	-0.47	-0.13	-0.20
	0.5	0.74	0.49	0.50
	-1	-1.01	-1.01	-1.03

Confidence interval coverage

The coverage probabilities were compared in Tables 5.4.D to 5.4.E. Only the full model provided stable and valid confidence interval coverages for all correlation levels. The BIC based stepwise backward selection method provided the worst results with only two acceptable coverages for the variable X_5 with the largest effect magnitude over all other variables. The STEP-AIC and ZERO-C also provided poor coverage rates, with up to six variables whose empirical coverage probability fell within the desirable range. The BMA-W and LASSO approaches provided three and four reliable coverages, respectively.

Table 5.4.D: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different correlation levels between five covariates, where N=300 and outcome probability is 30%. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
ρ	β	TRUE	FULL	STEP-AIC	STEP-BIC
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0	0	—	94.9 (2.4, 2.7) 0.57	68.2 (15.7, 16.1) 0.58	0.0 (41.8, 58.2) 0.59
	0.01	94.6 (2.5, 2.9) 0.58	94.5 (2.7, 2.8) 0.58	66.0 (17.3, 16.7) 0.59	0.0 (54.2, 45.8) 0.60
	-0.2	96.2 (1.8, 2.0) 1.16	96.1 (1.8, 2.1) 1.16	83.1 (7.0, 9.9) 1.17	42.9 (4.2, 53.0) 1.19
	0.5	94.7 (3.3, 2.0) 1.15	94.6 (3.4, 2.0) 1.16	96.1 (3.8, 0.2) 1.16	91.1 (8.9, 0.0) 1.17
	-1	95.0 (2.5, 2.5) 0.66	94.8 (2.3, 2.9) 0.67	95.3 (2.3, 2.4) 0.66	95.4 (2.7, 2.0) 0.65
0.3	0	—	95.8 (2.3, 1.9) 0.62	64.6 (20.1, 15.3) 0.60	0.0 (73.3, 26.7) 0.61
	0.01	95.7 (1.9, 2.4) 0.62	95.1 (1.8, 3.1) 0.64	64.7 (18.4, 17.0) 0.62	0.0 (69.1, 30.9) 0.61
	-0.2	94.7 (2.3, 3.0) 1.18	94.9 (2.2, 2.9) 1.18	81.3 (9.2, 9.6) 1.17	40.6 (8.6, 50.8) 1.17
	0.5	95.0 (2.9, 2.1) 1.23	94.8 (3.2, 2.0) 1.25	95.4 (4.4, 0.2) 1.20	90.6 (9.4, 0.0) 1.20
	-1	95.3 (2.2, 2.5) 0.69	94.8 (2.3, 2.9) 0.70	94.3 (3.0, 2.7) 0.67	93.6 (4.4, 2.0) 0.65
0.5	0	—	95.4 (2.2, 2.4) 0.71	60.9 (24.6, 14.6) 0.67	0.0 (70.4, 29.6) 0.65
	0.01	94.9 (3.4, 1.7) 0.70	94.2 (3.5, 2.3) 0.71	64.7 (19.3, 16.0) 0.68	0.0 (71.9, 28.1) 0.67
	-0.2	94.7 (2.8, 2.5) 1.33	94.1 (2.8, 3.1) 1.36	74.3 (13.1, 12.6) 1.30	33.2 (11.1, 55.7) 1.30
	0.5	95.0 (2.6, 2.4) 1.35	95.4 (2.6, 2.0) 1.41	94.7 (4.9, 0.4) 1.31	89.8 (10.2, 0.0) 1.30
	-1	94.8 (2.7, 2.5) 0.78	94.7 (2.5, 2.8) 0.79	92.3 (4.4, 3.4) 0.71	90.3 (7.1, 2.6) 0.65

The stepwise AIC exclusion based model-averaged method in Table 5.4.F also performed poorly with only 40% of coverage probabilities within the desired range, regardless of the methods used to estimate confidence intervals. The methods presented in Tables 5.4.D to 5.4.E, except for the full model, cannot calculate valid confidence intervals for effect sizes smaller than 0.5. The negative effect of increased correlation is observed for two stepwise selection methods and LASSO, since the number of correctly covered effects decreased for 0.5 correlation.

In general, the Occam's window based methods presented in Table 5.4.G presented good performance with no less than 86.7% of the time reaching the appropriate coverage rate. The Wald type model-averaged method demonstrated 100% success in terms of validity. The coverages for all coefficients were close to the nominal level. The coverage rate obtained from the Wald-type MATA after correction by the score standard errors provided minor underestimation for the variable X_3. Its coverage rate was slightly below the 93.6% under the largest correlation setting, while other methods struggled to achieve credible coverage for the variable X_1. Since the Occam's window selection already demonstrated poor performance in Tables 5.2.G and 5.3.G, only B-MATA-W and B-MATA-Ws techniques were accepted for further analysis.

Table 5.4.E: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different correlation levels between five covariates, where N=300 and outcome probability is 30%. The LASSO results are based on 10,000 simulations, the results of zero-corrected and Bayesian approaches are based on 5,000 simulations.

ρ	β	ZERO-C	LASSO	BMA-W
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0	0	100.0 (0.0, 0.0) 0.53	87.7 (6.1, 6.2) 0.58	99.7 (0.1, 0.2) 0.22
	0.01	71.8 (0.0, 28.2) 0.53	86.8 (6.9, 6.3) 0.58	99.8 (0.1, 0.2) 0.21
	-0.2	88.9 (10.9, 0.2) 1.08	89.4 (4.8, 5.8) 1.17	45.0 (54.5, 0.5) 0.52
	0.5	94.0 (3.7, 2.3) 1.17	95.9 (3.4, 0.8) 1.16	49.8 (2.2, 48.0) 0.94
	-1	93.8 (0.9, 5.3) 0.69	95.4 (2.2, 2.4) 0.66	94.9 (3.1, 2.0) 0.65
0.3	0	99.9 (0.1, 0.0) 0.58	87.8 (6.4, 5.7) 0.61	99.7 (0.3, 0.0) 0.23
	0.01	74.3 (0.0, 25.7) 0.59	87.5 (6.1, 6.4) 0.63	99.7 (0.3, 0.0) 0.24
	-0.2	87.3 (12.3, 0.4) 1.11	91.0 (4.5, 4.4) 1.17	40.0 (59.7, 0.3) 0.48
	0.5	94.7 (3.6, 1.7) 1.26	95.1 (3.9, 1.0) 1.23	46.1 (2.0, 51.9) 0.90
	-1	93.8 (1.4, 4.8) 0.74	93.9 (3.4, 2.7) 0.68	95.3 (3.0, 1.7) 0.67
0.5	0	100.0 (0.0, 0.0) 0.67	87.5 (7.5, 5.0) 0.70	99.8 (0.2, 0.0) 0.26
	0.01	74.9 (0.2, 24.9) 0.66	87.8 (6.7, 5.5) 0.70	99.7 (0.2, 0.0) 0.25
	-0.2	88.4 (11.3, 0.3) 1.27	91.2 (4.0, 4.8) 1.32	43.0 (56.8, 0.2) 0.52
	0.5	95.5 (2.5, 2.0) 1.38	92.2 (5.3, 2.5) 1.38	44.6 (1.8, 53.7) 0.93
	-1	93.6 (1.7, 4.7) 0.83	91.1 (6.2, 2.7) 0.73	94.6 (4.1, 1.3) 0.72

Except for two profile-likelihood related methods, the inclusion fraction based approaches in Table 5.4.H provided acceptable coverage rates. While the rest of the methods in Table 5.4.H demonstrated valid coverage 100% of the time, both I-MATA-PL and I-MATA-Wpl provided slightly underestimated coverage for the variable X_5 in highly correlated data. Valid coverage rate is the most important feature we are looking for. We still accepted the profile-likelihood based approaches for further comparison even though it slightly missed one coverage probability. Thus, all methods based on the inclusion fraction are being compared in the next section.

Tail errors

We evaluated the tail errors for eight methods that passed the first stage - the full model, Occam's window based MATA-W and MATA-Ws, and five inclusion fraction based methods. Despite the fact that BMA based methods provided good coverage rates, the balance of the confidence intervals was not achieved. Only 53.3% of confidence coverage rates provided by BMA based methods were well-balanced. The absolute difference between upper and lower tail errors for confidence intervals varied between 1 and 2.5.

The profile-likelihood based approaches and the full model demonstrated better balance of tail errors; however, the best results were obtained from the 50% inclusion fraction based MATA-W and two score based model-averaged approaches that failed to have balanced tail errors for most of variables. The full model also provided well-balanced intervals. Thus, it was also considered and compared by average width to model-averaged methods.

Table 5.4.F: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of three model-averaging CI construction methods for different correlation levels between five covariates using set of candidate models obtained from backward AIC selection approach for 95% nominal level based on 5,000 simulations, where N=300 and outcome probability is 30%; Wald based E-MATA-W, profile-likelihood based E-MATA-PL, and score function based E-MATA-S.

ρ	β	E-MATA-W	E-MATA-PL	E-MATA-S
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0	0	67.8 (13.9, 18.3) 0.58	64.9 (15.2, 19.9) 0.58	65.4 (15.0, 19.6) 0.57
	0.01	72.2 (11.7, 16.2) 0.59	70.1 (12.4, 17.5) 0.59	70.1 (12.4, 17.5) 0.58
	-0.2	83.9 (7.6, 8.5) 1.17	83.1 (7.6, 9.3) 1.17	83.4 (7.6, 9.0) 1.16
	0.5	96.4 (3.5, 0.2) 1.16	95.8 (4.0, 0.2) 1.16	96.1 (3.7, 0.2) 1.15
	-1	95.0 (2.8, 2.2) 0.66	94.5 (2.5, 3.0) 0.66	94.8 (3.0, 2.2) 0.65
0.3	0	65.7 (21.9, 12.4) 0.61	64.2 (23.0, 12.8) 0.61	64.6 (22.4, 13.0) 0.60
	0.01	67.4 (20.8, 11.8) 0.62	65.2 (21.8, 13.1) 0.62	65.6 (21.8, 12.6) 0.62
	-0.2	82.7 (10.6, 6.7) 1.18	82.3 (10.6, 7.1) 1.18	82.6 (10.6, 6.8) 1.17
	0.5	95.8 (3.9, 0.3) 1.21	95.3 (4.5, 0.3) 1.21	95.7 (4.1, 0.3) 1.20
	-1	95.3 (2.6, 2.0) 0.68	95.3 (2.2, 2.5) 0.68	95.1 (2.8, 2.1) 0.67
0.5	0	65.5 (22.0, 12.4) 0.68	63.2 (23.2, 13.5) 0.68	63.9 (23.1, 13.0) 0.67
	0.01	70.4 (16.9, 12.7) 0.69	67.7 (18.5, 13.8) 0.69	68.5 (17.8, 13.6) 0.68
	-0.2	78.9 (12.3, 8.7) 1.32	78.4 (12.3, 9.2) 1.32	78.6 (12.3, 9.1) 1.31
	0.5	95.1 (4.5, 0.4) 1.33	94.6 (5.0, 0.4) 1.33	95.0 (4.6, 0.4) 1.32
	-1	94.0 (3.8, 2.2) 0.73	93.8 (3.5, 2.7) 0.73	93.7 (4.0, 2.3) 0.72

Table 5.4.G: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different correlation levels between five covariates using set of candidate models obtained from Occam's window approach for 95% nominal level based on 5,000 simulations, where N=300 and outcome probability is 30%; Wald based B-MATA-W, profile-likelihood based B-MATA-PL, score function based B-MATA-S, Wald based method corrected by the profile-likelihood B-MATA-Wpl, and Wald based method corrected by the score function B-MATA-Ws.

ρ	β	B-MATA-W	B-MATA-PL	B-MATA-S	B-MATA-Wpl	B-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0	0	94.5 (2.4, 3.1) 0.57	94.0 (2.6, 3.4) 0.57	94.1 (2.5, 3.4) 0.56	94.1 (2.5, 3.4) 0.57	94.2 (2.5, 3.2) 0.57
	0.01	95.6 (1.9, 2.6) 0.58	95.1 (2.1, 2.8) 0.58	95.2 (2.0, 2.8) 0.57	95.2 (2.0, 2.8) 0.58	95.3 (1.9, 2.7) 0.57
	-0.2	95.1 (2.7, 2.2) 1.16	94.8 (2.9, 2.3) 1.16	94.8 (2.9, 2.3) 1.15	94.9 (2.8, 2.3) 1.16	95.0 (2.8, 2.2) 1.15
	0.5	94.8 (2.2, 3.0) 1.15	94.5 (2.5, 3.0) 1.15	94.7 (2.4, 3.0) 1.14	94.6 (2.5, 3.0) 1.16	94.8 (2.3, 3.0) 1.15
	-1	95.0 (2.9, 2.1) 0.66	94.6 (2.5, 3.0) 0.66	94.8 (2.9, 2.3) 0.65	94.7 (2.4, 2.9) 0.66	94.8 (2.9, 2.2) 0.66
0.3	0	94.3 (3.6, 2.2) 0.59	93.9 (3.8, 2.3) 0.59	94.0 (3.7, 2.3) 0.58	94.0 (3.8, 2.2) 0.59	94.2 (3.7, 2.2) 0.58
	0.01	94.3 (3.6, 2.1) 0.60	93.9 (3.8, 2.3) 0.60	94.0 (3.8, 2.2) 0.59	94.0 (3.8, 2.3) 0.60	94.1 (3.8, 2.2) 0.60
	-0.2	94.3 (4.0, 1.7) 1.15	94.0 (4.3, 1.7) 1.15	94.1 (4.2, 1.7) 1.14	94.1 (4.2, 1.7) 1.15	94.2 (4.1, 1.7) 1.14
	0.5	94.6 (2.2, 3.2) 1.19	94.3 (2.6, 3.2) 1.19	94.4 (2.3, 3.4) 1.18	94.4 (2.5, 3.1) 1.19	94.5 (2.2, 3.3) 1.18
	-1	95.6 (2.3, 2.1) 0.69	95.4 (2.0, 2.5) 0.68	95.4 (2.4, 2.2) 0.68	95.5 (2.0, 2.5) 0.69	95.5 (2.4, 2.1) 0.68
0.5	0	93.8 (3.9, 2.3) 0.65	93.3 (4.1, 2.5) 0.65	93.4 (4.1, 2.5) 0.64	93.5 (4.0, 2.5) 0.65	93.6 (4.0, 2.4) 0.64
	0.01	95.0 (2.8, 2.2) 0.66	94.6 (3.0, 2.3) 0.66	94.7 (3.0, 2.3) 0.65	94.7 (3.0, 2.3) 0.66	94.8 (2.9, 2.2) 0.65
	-0.2	93.6 (4.5, 1.9) 1.27	93.2 (4.7, 2.1) 1.26	93.3 (4.7, 2.0) 1.25	93.4 (4.6, 2.0) 1.27	93.5 (4.5, 2.0) 1.26
	0.5	94.2 (2.2, 3.6) 1.30	93.8 (2.6, 3.6) 1.29	93.9 (2.4, 3.8) 1.28	94.0 (2.5, 3.5) 1.30	94.0 (2.3, 3.7) 1.29
	-1	95.3 (2.7, 2.0) 0.75	95.0 (2.6, 2.4) 0.75	95.0 (2.8, 2.1) 0.75	95.0 (2.6, 2.4) 0.75	95.1 (2.8, 2.1) 0.75

Table 5.4.H: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different correlation levels between five covariates using set of candidate models obtained from 50% inclusion fraction approach for 95% nominal level based on 1,000 simulations, where N=300 and outcome probability is 30%; Wald based I-MATA-W, profile-likelihood based I-MATA-PL, score function based I-MATA-S, Wald based method corrected by the profile-likelihood I-MATA-Wpl, and Wald based method corrected by the score function I-MATA-Ws.

ρ	β	I-MATA-W	I-MATA-PL	I-MATA-S	I-MATA-Wpl	I-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0	0	95.2 (2.3, 2.5) 0.56	94.6 (2.7, 2.7) 0.57	94.7 (2.6, 2.7) 0.57	94.7 (2.6, 2.7) 0.58	94.9 (2.4, 2.7) 0.57
	0.01	94.5 (2.3, 3.2) 0.58	93.7 (2.8, 3.5) 0.58	94.0 (2.6, 3.4) 0.57	94.0 (2.6, 3.4) 0.58	94.2 (2.5, 3.3) 0.58
	-0.2	95.8 (1.9, 2.3) 1.17	95.5 (1.9, 2.6) 1.16	95.6 (1.9, 2.5) 1.15	95.5 (1.9, 2.6) 1.17	95.6 (1.9, 2.5) 1.16
	0.5	94.9 (3.1, 2.0) 1.16	94.3 (3.7, 2.0) 1.16	94.6 (3.4, 2.0) 1.15	94.3 (3.7, 2.0) 1.16	94.7 (3.3, 2.0) 1.15
	-1	94.9 (2.5, 2.6) 0.66	94.5 (1.9, 3.6) 0.66	94.7 (2.5, 2.8) 0.66	94.7 (1.9, 3.4) 0.67	94.8 (2.5, 2.7) 0.66
0.3	0	95.7 (2.4, 1.9) 0.61	95.2 (2.9, 1.9) 0.61	95.3 (2.7, 2.0) 0.60	95.3 (2.8, 1.9) 0.61	95.5 (2.6, 1.9) 0.61
	0.01	95.2 (2.1, 2.7) 0.63	94.9 (2.2, 2.9) 0.63	95.1 (2.2, 2.7) 0.62	95.0 (2.2, 2.8) 0.63	95.1 (2.2, 2.7) 0.62
	-0.2	94.7 (2.4, 2.9) 1.18	94.5 (2.5, 3.0) 1.17	94.6 (2.5, 2.9) 1.16	94.7 (2.4, 2.9) 1.18	94.7 (2.4, 2.9) 1.17
	0.5	94.7 (3.4, 1.9) 1.23	94.6 (3.6, 1.8) 1.22	94.5 (3.5, 2.0) 1.21	94.6 (3.6, 1.8) 1.23	94.5 (3.5, 2.0) 1.22
	-1	94.5 (2.9, 2.6) 0.69	94.1 (2.4, 3.5) 0.69	94.2 (3.0, 2.8) 0.68	94.1 (2.4, 3.5) 0.69	94.3 (3.0, 2.7) 0.69
0.5	0	94.4 (3.0, 2.6) 0.69	94.0 (3.2, 2.8) 0.69	94.2 (3.1, 2.7) 0.69	94.3 (3.0, 2.7) 0.69	94.3 (3.0, 2.7) 0.69
	0.01	94.1 (3.7, 2.2) 0.70	93.9 (3.7, 2.4) 0.70	94.0 (3.7, 2.3) 0.69	94.0 (3.7, 2.3) 0.70	94.1 (3.7, 2.2) 0.69
	-0.2	94.0 (3.3, 2.7) 1.33	93.9 (3.3, 2.8) 1.33	93.9 (3.3, 2.8) 1.32	93.9 (3.3, 2.8) 1.34	94.0 (3.3, 2.7) 1.32
	0.5	95.0 (2.5, 2.5) 1.36	94.7 (2.8, 2.5) 1.36	94.9 (2.6, 2.5) 1.35	94.8 (2.8, 2.4) 1.37	95.0 (2.5, 2.5) 1.35
	-1	94.0 (3.2, 2.8) 0.77	93.4 (3.1, 3.5) 0.77	93.9 (3.3, 2.8) 0.76	93.5 (3.1, 3.4) 0.77	93.9 (3.3, 2.8) 0.76

Average width

The full model and five inclusion fraction based MATA methods with balanced confidence intervals were compared by their average widths in Figure 5.3. All methods provided similar average widths for the uncorrelated variables, such that the maximal width difference between the methods did not exceed 1.2%. The inclusion fraction based MATA-W approach provided the largest average width, while the MATA-S method produced the shortest intervals.

Overall, the order of model-averaging methods by width was MATA-S $<$ MATA-Ws $<$ MATA-PL $<$ MATA-W $<$ MATA-Wpl. The full model outperformed the MATA-PL method for uncorrelated data; however, with the increase in correlation that expectedly inflated average width of all methods, the full model took its place at the end of this order with the highest averaged width. With the increase of correlation, the outperformance of the I-MATA-S method over all other approaches became more noticeable with differences varying in 1.6-4.7% range. The improvement may be small, but it is always in favor of the inclusion fraction based MATA-S algorithm.

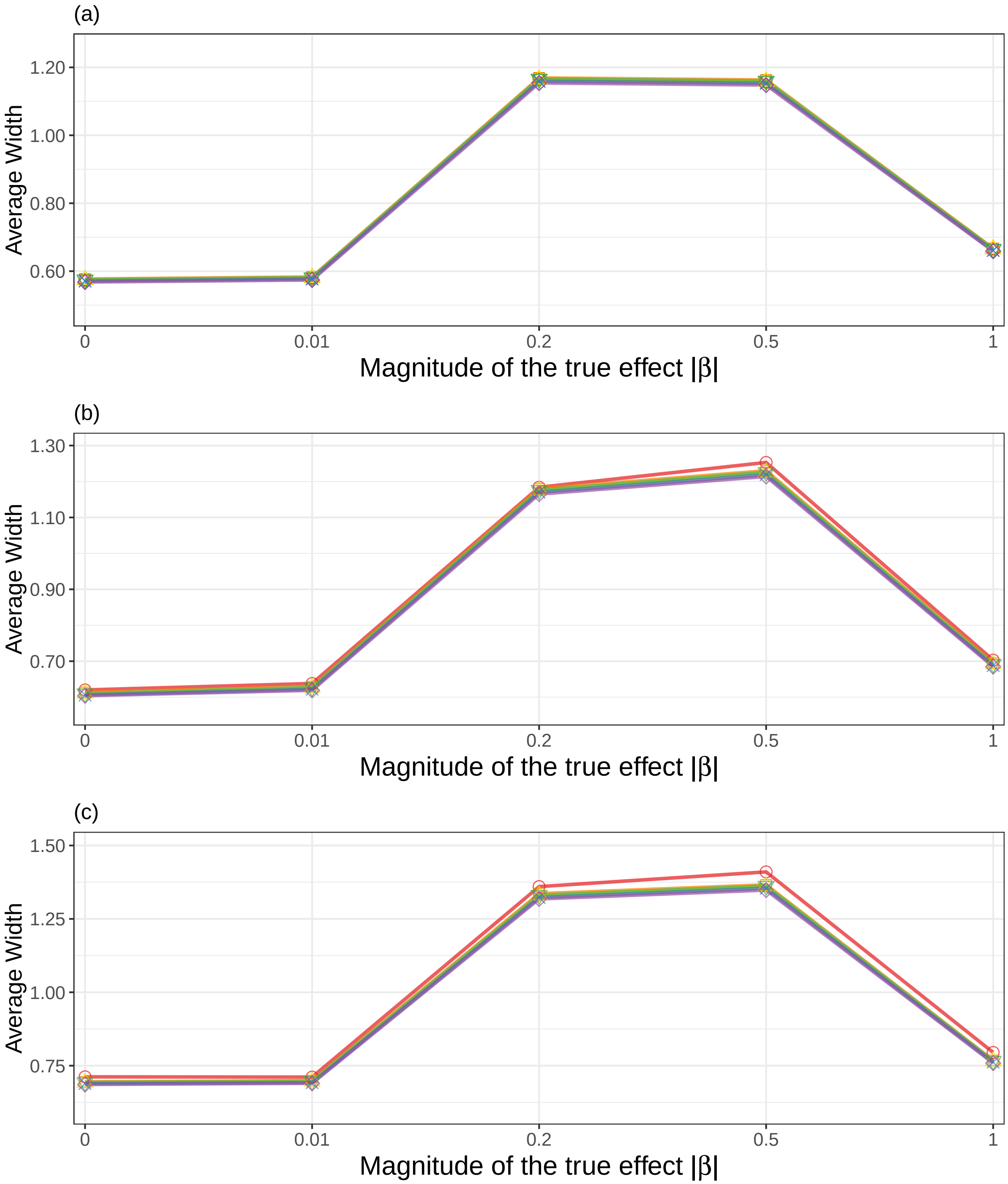


Figure 5.3: Comparison of averaged widths of the full model- ○ , I-MATA-Wpl - □ , I-MATA-PL - △ , I-MATA-W - ▽ , I-MATA-Ws - × , I-MATA-S - ◇ for correlations: (a) - ρ=0.3, (b) - ρ=0, and (c) - ρ=0.5.

5.7.4 Probability of outcome

The probability of the outcome is an important factor when evaluating logistic regression models. Performance of methods for inference improves as probability shifts toward 0.5. For the next set of simulations, the methods are compared for 0.1, 0.3 and 0.5 probability.

For evaluating the effect of outcome probability on the performance of the methods, data with five correlated variables and 500 subjects was generated, and the correlation between all variables was fixed at 0.3. The outcome probability was regulated by changing the intercept. For outcome probabilites of 0.1, 0.3, and 0.5, we used intercepts equal to -2.7, -1.15 and -0.15, respectively. The means of point estimates are presented in Tables 5.5.A to 5.5.C, while empirical coverage probabilities, tail errors and average widths for 19 methods can be found in Tables 5.5.D to 5.5.H. Comparison of the valid methods by their widths is presented in Figure 5.4.

The probability of event plays an important role in choosing the sufficient sample size. According to the different EPV suggestions - 10, 20, and 50 events per variable, for the smallest outcome probability of 0.1 the sufficient sample size should be 500, 1000, or 2500, while for the outcome probability of 0.5 the sample sizes should be 100, 200, or 500 (Peduzzi et al., 1996; Vittinghoff and McCulloch, 2007; Steyerberg et al., 1999). However, even for the smallest outcome rate the logistic regression did not have any convergence problems.

Bias of point estimates

Overall, the performance of the methods improved as the outcome probability approached 0.5. As in previous simulations, out of all methods presented in Table 5.5.A, the full model is the only one that provided the most unbiased point estimates.

Table 5.5.A: Mean of point estimates obtained from the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different outcome probabilities, where N=500 and $\rho = 0.3$. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
Prob	β	TRUE	FULL	STEP-AIC	STEP-BIC
0.1	0	—	0.000	0.022	0.1
	0.01	0.02	0.02	0.02	0.20
	-0.2	-0.21	-0.21	-0.54	-0.78
	0.5	0.52	0.52	0.74	0.99
	-1	-1.04	-1.04	-1.02	-1.00
0.3	0	—	0.004	-0.01	0.06
	0.01	0.01	0.01	0.03	0.15
	-0.2	-0.20	-0.20	-0.43	-0.62
	0.5	0.50	0.50	0.58	0.71
	-1	-1.01	-1.02	-1.02	-1.00
0.5	0	—	0.002	-0.01	0.04
	0.01	0.01	0.01	0.03	0.09
	-0.2	-0.20	-0.20	-0.42	-0.60
	0.5	0.50	0.50	0.56	0.67
	-1	-1.02	-1.02	-1.02	-1.0

The stepwise elimination methods provided significantly biased results. Since the BIC penalty penalizes additional covariates more than AIC, the estimated effect should be larger than the true effect in order to be selected for the final model, if true effect is not large. Because of this, the bias produced by the BIC based backward elimination was larger than the bias produced by the AIC penalty.

Table 5.5.B: Mean of point estimates obtained from the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different outcome probabilities, where N=500 and $\rho = 0.3$. The LASSO results are based on 10,000 simulations, the results of zero-corrected backward selection and Wald based Bayesian model-averaging are based on 5,000 simulations.

Prob	β	ZERO-C	LASSO	BMA-W
0.1	0	0.004	-0.02	0.01
	0.01	0.01	0.002	0.01
	-0.2	-0.17	-0.38	-0.04
	0.5	0.46	0.68	0.21
	-1	-1.06	-1.01	-1.01
0.3	0	0.01	-0.02	0.004
	0.01	0.01	0.02	0.004
	-0.2	-0.16	-0.30	-0.05
	0.5	0.46	0.58	0.27
	-1	-1.03	-1.01	-1.00
0.5	0	0.002	-0.02	0.004
	0.01	0.01	0.02	0.01
	-0.2	-0.16	-0.29	-0.05
	0.5	0.47	0.57	0.29
	-1	-1.04	-1.01	-1.01

All methods presented in Table 5.4.B produced relatively biased point estimates as in previous simulations. Both zero-corrected backward elimination and LASSO provided biased estimates away from zero, while most of the estimates by the Wald-type BMA were biased toward zero. The increase in outcome probability just slightly reduced the bias of these methods.

The model-averaging results from Table 5.4.C showed that Occam's window and inclusion fraction approaches selected appropriate sets of models and produced consistent point estimates. For the data with 0.1 outcome probability, the point estimates by the inclusion fraction were less biased than the estimates based on Occam's window. The results of model-averaging after backward elimination also show that even for 0.5 outcome probability, model-averaging cannot correct the bias created by poor selection of candidate models.

Table 5.5.C: Mean of point estimates of the backward stepwise selection (E-MATA), Occam's window (B-MATA) and inclusion fraction (I-MATA) based model-averaging tail area methods for different outcome probabilities, where N=500 and $\rho = 0.3$. The backward stepwise selection and Occam's window means are based on 5,000 simulations, the results obtained from inclusion fraction are based on 1,000 simulations.

Prob	β	E-MATA	B-MATA	I-MATA
0.1	0	0.03	0.04	0.004
	0.01	0.03	0.04	0.02
	-0.2	-0.51	-0.19	-0.2
	0.5	0.73	0.55	0.51
	-1	-1.02	-1.02	-1.04
0.3	0	0.004	0.02	0.01
	0.01	0.02	0.02	0.01
	-0.2	-0.41	-0.20	-0.2
	0.5	0.57	0.50	0.49
	-1	-1.01	-1.01	-1.02
0.5	0	0.01	0.01	0.001
	0.01	0.03	0.02	0.01
	-0.2	-0.41	-0.22	-0.2
	0.5	0.55	0.50	0.50
	-1	-1.02	-1.02	-1.02

Confidence interval coverage

Overall, the coverage performance of the methods was improved by the increase in outcome probability. However, the methods that showed weak results for 0.1 probability of outcome did not reach the nominal coverage level for more a balanced outcome.

Table 5.5.D: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the true model, the full model, stepwise AIC and stepwise BIC backward selection methods for different outcome probabilities, where N=500 and $\rho = 0.3$. The true and the full model results are based on 1,000 simulations, the results of backward selection methods are based on 10,000 simulations.

				Backward selection	
Prob	β	TRUE	FULL	STEP-AIC	STEP-BIC
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0.1	0	—	95.7 (2.2, 2.1) 0.70	64.9 (19.6, 15.5) 0.69	0.0 (67.1, 32.9) 0.70
	0.01	95.1 (2.5, 2.4) 0.68	94.9 (2.7, 2.4) 0.68	66.7 (18.3, 15.0) 0.67	0.0 (71.1, 28.9) 0.67
	-0.2	95.2 (3.0, 1.8) 1.39	94.8 (3.1, 2.1) 1.40	78.2 (11.9, 9.8) 1.39	21.6 (11.9, 66.5) 1.43
	0.5	95.1 (2.9, 2.0) 1.34	95.2 (2.6, 2.2) 1.36	95.2 (4.4, 0.5) 1.31	87.4 (12.6, 0.0) 1.31
	-1	94.3 (2.4, 3.3) 0.76	94.3 (2.3, 3.4) 0.77	94.8 (2.7, 2.5) 0.732	94.7 (3.5, 1.9) 0.71
0.3	0	—	95.8 (2.5, 1.7) 0.47	62.6 (17.2, 20.3) 0.46	0.0 (58.5, 41.5) 0.46
	0.01	94.0 (3.0, 3.0) 0.46	93.7 (3.1, 3.2) 0.46	64.8 (17.1, 18.1) 0.45	0.0 (74.2, 25.8) 0.45
	-0.2	94.4 (2.7, 2.9) 0.92	94.6 (2.3, 3.1) 0.93	87.0 (4.5, 8.5) 0.90	51.6 (3.5, 44.9) 0.90
	0.5	95.5 (1.9, 2.6) 0.91	95.4 (2.1, 2.5) 0.92	96.6 (3.4, 0.0) 0.89	94.5 (5.5, 0.0) 0.88
	-1	94.6 (2.8, 2.6) 0.53	94.4 (2.9, 2.7) 0.54	94.1 (2.6, 3.3) 0.52	93.5 (3.6, 2.9) 0.50
0.5	0	—	95.0 (1.9, 3.1) 0.43	67.0 (16.3, 16.6) 0.42	0.0 (57.2, 42.8) 0.42
	0.01	94.8 (2.8, 2.4) 0.42	94.9 (2.9, 2.2) 0.43	63.6 (17.0, 19.5) 0.42	0.0 (65.0, 35.0) 0.42
	-0.2	95.9 (2.2, 1.9) 0.85	96.0 (2.0, 2.0) 0.86	88.2 (3.5, 8.3) 0.84	56.0 (1.7, 42.3) 0.83
	0.5	95.9 (2.0, 2.1) 0.85	95.7 (2.2, 2.1) 0.86	97.0 (3.0, 0.0) 0.83	95.8 (4.2, 0.0) 0.82
	-1	94.5 (2.3, 3.2) 0.50	94.6 (2.2, 3.2) 0.50	94.6 (2.4, 3.0) 0.48	93.6 (3.6, 2.8) 0.47

Table 5.5.D contains results for the full model and two stepwise selection methods. While the full model provided strong results with 100% of the effects getting valid coverage, the stepwise based methods, STEP-AIC and STEP-BIC, performed poorly with acceptable coverage probability only for variable X_4 and/or X_5, such that only 26.7% of the time they reached the empirical range for nominal level.

Out of all methods presented in Table 5.5.E, the LASSO procedure provided the best results, although only one third of the empirical coverage probabilities was in the desired range. The worst performance was shown by the Wald-type Bayesian model-averaging method with 80% of the coverages outside of the desired range. The zero-corrected method also demonstrated poor performance, regardless of the probability of outcome.

Table 5.5.E: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of 95% CIs constructed by the zero-corrected backward selection, LASSO and Wald based Bayesian model-averaging methods for different outcome probabilities, where N=500 and $\rho = 0.3$. The LASSO results are based on 10,000 simulations, the results of zero-corrected and Bayesian approaches are based on 5,000 simulations.

Prob	β	ZERO-C	LASSO	BMA-W
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0.1	0	99.9 (0.0, 0.1) 0.64	87.5 (6.5, 6.0) 0.70	99.4 (0.4, 0.1) 0.30
	0.01	75.9 (0.0, 24.1) 0.63	88.2 (6.9, 5.0) 0.68	99.7 (0.3, 0.1) 0.29
	-0.2	86.1 (13.8, 0.1) 1.31	90.5 (4.7, 4.8) 1.40	61.0 (38.9, 0.1) 0.63
	0.5	95.2 (2.5, 2.3) 1.33	94.0 (4.2, 1.9) 1.34	45.4 (2.2, 52.4) 0.97
	-1	93.2 (1.0, 5.8) 0.81	94.7 (3.0, 2.3) 0.74	95.1 (2.9, 2.0) 0.73
0.3	0	99.9 (0.0, 0.1) 0.43	88.5 (5.2, 6.4) 0.47	99.7 (0.3, 0.1) 0.20
	0.01	73.5 (0.1, 26.4) 0.42	88.4 (6.1, 5.4) 0.46	99.9 (0.1, 0.0) 0.20
	-0.2	91.8 (7.6, 0.6) 0.86	92.4 (3.4, 4.2) 0.91	41.8 (57.6, 0.6) 0.46
	0.5	95.1 (2.4, 2.5) 0.95	96.2 (3.5, 0.3) 0.91	56.5 (1.9, 41.6) 0.85
	-1	93.9 (2.2, 3.9) 0.55	94.3 (2.8, 3.0) 0.52	94.6 (3.0, 2.4) 0.52
0.5	0	99.8 (0.1, 0.1) 0.39	89.9 (5.1, 5.0) 0.43	99.8 (0.1, 0.1) 0.19
	0.01	72.8 (0.1, 27.1) 0.38	88.1 (6.0, 5.9) 0.43	99.8 (0.1, 0.2) 0.18
	-0.2	92.6 (6.9, 0.5) 0.78	92.3 (3.4, 4.2) 0.85	44.0 (55.4, 0.6) 0.46
	0.5	95.3 (2.8, 1.9) 0.89	96.5 (3.3, 0.2) 0.85	62.1 (1.4, 36.4) 0.85
	-1	93.1 (1.4, 5.5) 0.51	94.7 (2.7, 2.6) 0.49	95.3 (2.4, 2.2) 0.49

The E-MATA methods in Table 5.5.F presented similar results to the stepwise AIC backward elimination from Table 5.5.D in terms of 26.7% coverage success, but they still performed better than STEP-AIC. The stepwise AIC exclusion based model-average methods mostly provided coverage intervals with underestimated variance, however its empirical coverage probabilities were slightly closer to the lower bound of 93.6% than those of the STEP-AIC approach.

Table 5.5.F: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of three model-averaging CI construction methods for different outcome probabilities using set of candidate models obtained from backward AIC selection approach for 95% nominal level based on 5,000 simulations, where N=500 and $\rho = 0.3$; Wald based E-MATA-W, profile-likelihood based E-MATA-PL, and score function based E-MATA-S.

Prob	β	E-MATA-W	E-MATA-PL	E-MATA-S
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0.1	0	67.2 (19.3, 13.5) 0.70	65.0 (20.6, 14.5) 0.70	65.5 (20.2, 14.3) 0.69
	0.01	72.0 (14.5, 13.5) 0.68	70.3 (14.8, 14.8) 0.68	70.6 (15.3, 14.1) 0.67
	-0.2	79.7 (12.7, 7.6) 1.40	78.8 (12.7, 8.4) 1.41	79.4 (12.7, 7.9) 1.39
	0.5	95.5 (4.2, 0.3) 1.32	95.0 (4.6, 0.3) 1.32	95.3 (4.4, 0.3) 1.30
	-1	94.9 (2.7, 2.3) 0.74	94.6 (2.5, 2.9) 0.74	94.7 (2.8, 2.5) 0.73
0.3	0	69.1 (16.4, 14.5) 0.47	68.1 (16.7, 15.2) 0.47	68.5 (16.5, 15.0) 0.47
	0.01	69.7 (14.6, 15.6) 0.46	68.8 (15.3, 16.0) 0.46	69.0 (15.1, 15.9) 0.46
	-0.2	87.6 (5.4, 7.0) 0.91	87.5 (5.4, 7.1) 0.91	87.5 (5.4, 7.1) 0.91
	0.5	97.3 (2.7, 0.0) 0.89	97.0 (3.0, 0.0) 0.89	97.2 (2.8, 0.0) 0.89
	-1	94.5 (2.7, 2.8) 0.52	94.3 (2.4, 3.2) 0.52	94.4 (2.7, 2.9) 0.52
0.5	0	67.1 (18.2, 14.7) 0.43	66.0 (18.9, 15.1) 0.43	66.1 (18.8, 15.1) 0.43
	0.01	69.0 (16.5, 14.5) 0.42	68.0 (17.0, 15.1) 0.42	68.1 (16.8, 15.1) 0.42
	-0.2	89.9 (3.1, 7.1) 0.85	89.8 (3.1, 7.1) 0.85	89.8 (3.1, 7.1) 0.84
	0.5	97.9 (2.1, 0.0) 0.83	97.6 (2.4, 0.0) 0.83	97.8 (2.2, 0.0) 0.83
	-1	95.1 (2.2, 2.7) 0.49	94.8 (1.8, 3.3) 0.49	95.0 (2.2, 2.8) 0.49

The Occam's window based model-averaged methods presented in Table 5.5.G also provided very disappointing results. As for the E-MATA based algorithms, this set of methods successfully provided correct coverage only 26.7% of the time. Nevertheless, out of all methods in Tables 5.5.E to 5.5.G, the BMA based model-averaged methods presented the best coverage levels with average coverage of around 91%. Unfortunately, none of the methods in Tables 5.5.D to 5.5.G could outperform the full model in terms of percentage of the time the empirical coverage probability hit the desired range.

Table 5.5.G: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different outcome probabilities using set of candidate models obtained from Occam's window approach for 95% nominal level based on 5,000 simulations, where N=500 and $\rho = 0.3$; Wald based B-MATA-W, profile-likelihood based B-MATA-PL, score function based B-MATA-S, Wald based method corrected by the profile-likelihood B-MATA-Wpl, and Wald based method corrected by the score function B-MATA-Ws.

Prob	β	B-MATA-W	B-MATA-PL	B-MATA-S	B-MATA-Wpl	B-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0.1	0	91.5 (5.2, 3.2) 0.68	91.0 (5.5, 3.5) 0.68	91.1 (5.4, 3.4) 0.67	91.2 (5.4, 3.4) 0.68	91.2 (5.4, 3.4) 0.67
	0.01	92.5 (3.9, 3.7) 0.66	92.1 (4.0, 4.0) 0.66	92.2 (4.0, 3.8) 0.65	92.1 (4.0, 4.0) 0.66	92.2 (4.0, 3.8) 0.66
	-0.2	90.6 (6.9, 2.6) 1.34	90.0 (7.0, 3.0) 1.35	89.9 (7.3, 2.8) 1.33	90.1 (6.9, 3.0) 1.35	90.0 (7.2, 2.7) 1.33
	0.5	94.0 (2.4, 3.7) 1.30	93.6 (2.7, 3.8) 1.30	93.7 (2.5, 3.8) 1.28	93.7 (2.6, 3.7) 1.30	93.8 (2.5, 3.8) 1.28
	-1	95.2 (2.4, 2.3) 0.74	94.8 (2.3, 2.9) 0.74	94.9 (2.6, 2.5) 0.74	94.9 (2.2, 2.9) 0.75	95.1 (2.5, 2.4) 0.74
0.3	0	91.9 (4.3, 3.7) 0.46	91.7 (4.5, 3.8) 0.46	91.8 (4.4, 3.8) 0.45	91.7 (4.5, 3.8) 0.46	91.8 (4.4, 3.8) 0.46
	0.01	92.2 (3.7, 4.0) 0.45	92.0 (3.8, 4.2) 0.45	92.0 (3.8, 4.2) 0.44	92.0 (3.8, 4.2) 0.45	92.1 (3.7, 4.1) 0.44
	-0.2	89.7 (7.5, 2.8) 0.89	89.3 (7.8, 2.9) 0.89	89.4 (7.7, 2.9) 0.89	89.4 (7.7, 2.9) 0.89	89.6 (7.5, 2.8) 0.89
	0.5	97.3 (2.2, 0.5) 0.89	97.1 (2.4, 0.5) 0.89	97.2 (2.3, 0.5) 0.88	97.2 (2.3, 0.5) 0.89	97.2 (2.3, 0.5) 0.88
	-1	94.7 (2.6, 2.7) 0.52	94.5 (2.4, 3.1) 0.52	94.6 (2.6, 2.8) 0.52	94.6 (2.3, 3.1) 0.53	94.7 (2.6, 2.7) 0.52
0.5	0	91.0 (4.9, 4.0) 0.42	90.9 (5.0, 4.1) 0.42	90.9 (5.0, 4.1) 0.42	90.9 (5.0, 4.0) 0.42	90.9 (5.0, 4.1) 0.42
	0.01	91.7 (4.4, 4.0) 0.41	91.4 (4.5, 4.1) 0.41	91.4 (4.5, 4.1) 0.41	91.4 (4.5, 4.1) 0.41	91.5 (4.5, 4.1) 0.41
	-0.2	90.3 (6.6, 3.1) 0.83	90.2 (6.7, 3.1) 0.83	90.2 (6.6, 3.1) 0.83	90.2 (6.7, 3.1) 0.83	90.3 (6.6, 3.1) 0.83
	0.5	97.7 (1.7, 0.6) 0.83	97.6 (1.8, 0.6) 0.83	97.6 (1.8, 0.6) 0.83	97.6 (1.8, 0.6) 0.83	97.7 (1.8, 0.6) 0.83
	-1	95.3 (2.0, 2.7) 0.49	95.0 (1.7, 3.2) 0.49	95.2 (2.1, 2.7) 0.49	95.0 (1.7, 3.2) 0.49	95.3 (2.0, 2.7) 0.49

The inclusion fraction based model-averaged methods in Table 5.5.H presented valid and stable results for all estimated effects. The decrease in probability of outcome had no effect on the performance on I-MATA based algorithms.

Tail errors

Out of all methods, only the full model and inclusion fraction based methods demonstrated consistency in the reliability of the confidence interval. These methods passed the previous test, and showed similar performance in terms of the balance between tail errors. The best results were provided by the full model and two score function based model-averaging methods. The profile-likelihood based procedures underperformed all other methods in terms of tail errors balance. However, we still consider them further, since the differences were small.

Table 5.5.H: Empirical coverage (Cov), tail errors (<, >)% and averaged width (WD) of five model-averaging CI construction methods for different outcome probabilities using set of candidate models obtained from 50% inclusion fraction approach for 95% nominal level based on 1,000 simulations, where N=500 and $\rho = 0.3$; Wald based I-MATA-W, profile-likelihood based I-MATA-PL, score function based I-MATA-S, Wald based method corrected by the profile-likelihood I-MATA-Wpl, and Wald based method corrected by the score function I-MATA-Ws.

Prob	β	I-MATA-W	I-MATA-PL	I-MATA-S	I-MATA-Wpl	I-MATA-Ws
		Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% WD
0.1	0	95.7 (2.3, 2.0) 0.69	95.6 (2.4, 2.0) 0.69	95.6 (2.4, 2.0) 0.69	95.6 (2.4, 2.0) 0.70	95.7 (2.3, 2.0) 0.69
	0.01	94.6 (3.1, 2.3) 0.68	94.4 (3.3, 2.3) 0.68	94.5 (3.2, 2.3) 0.67	94.4 (3.3, 2.3) 0.68	94.5 (3.2, 2.3) 0.67
	-0.2	94.5 (3.4, 2.1) 1.38	94.0 (3.8, 2.2) 1.39	94.0 (3.8, 2.2) 1.37	94.2 (3.6, 2.2) 1.39	94.0 (3.8, 2.2) 1.37
	0.5	95.5 (2.2, 2.3) 1.34	95.0 (2.6, 2.4) 1.34	95.1 (2.5, 2.4) 1.32	95.0 (2.6, 2.4) 1.34	95.1 (2.5, 2.4) 1.32
	-1	94.8 (2.1, 3.1) 0.76	94.2 (1.9, 3.9) 0.76	94.4 (2.3, 3.3) 0.75	94.4 (1.9, 3.7) 0.76	94.5 (2.2, 3.3) 0.75
0.3	0	95.4 (2.8, 1.8) 0.47	95.4 (2.8, 1.8) 0.47	95.3 (2.8, 1.9) 0.46	95.4 (2.8, 1.8) 0.47	95.3 (2.8, 1.9) 0.47
	0.01	94.0 (2.9, 3.1) 0.46	93.8 (3.0, 3.2) 0.46	93.9 (3.0, 3.1) 0.46	93.8 (3.0, 3.2) 0.46	93.9 (3.0, 3.1) 0.46
	-0.2	93.9 (3.1, 3.0) 0.92	93.7 (3.1, 3.2) 0.92	93.7 (3.1, 3.2) 0.91	93.7 (3.1, 3.2) 0.92	93.8 (3.1, 3.1) 0.91
	0.5	95.5 (1.8, 2.7) 0.91	95.4 (1.9, 2.7) 0.91	95.3 (1.8, 2.9) 0.90	95.4 (1.9, 2.7) 0.91	95.4 (1.8, 2.8) 0.90
	-1	94.0 (3.2, 2.8) 0.53	93.9 (2.9, 3.2) 0.53	93.8 (3.3, 2.9) 0.53	94.0 (2.9, 3.1) 0.53	93.8 (3.3, 2.9) 0.53
0.5	0	95.1 (1.9, 3.0) 0.43	94.9 (2.1, 3.0) 0.43	94.9 (2.1, 3.0) 0.43	94.9 (2.1, 3.0) 0.43	94.9 (2.1, 3.0) 0.43
	0.01	95.1 (2.8, 2.1) 0.42	94.9 (2.9, 2.2) 0.42	95.0 (2.8, 2.2) 0.42	95.0 (2.8, 2.2) 0.42	95.0 (2.8, 2.2) 0.42
	-0.2	95.8 (2.0, 2.2) 0.85	95.8 (2.0, 2.2) 0.85	95.8 (2.0, 2.2) 0.85	95.8 (2.0, 2.2) 0.85	95.8 (2.0, 2.2) 0.85
	0.5	95.7 (2.2, 2.1) 0.85	95.4 (2.5, 2.1) 0.84	95.5 (2.3, 2.2) 0.84	95.4 (2.5, 2.1) 0.85	95.7 (2.2, 2.1) 0.84
	-1	94.4 (2.3, 3.3) 0.50	93.7 (2.2, 4.1) 0.50	94.2 (2.3, 3.5) 0.49	93.7 (2.2, 4.1) 0.50	94.3 (2.3, 3.4) 0.49

Average width

The widths of the confidence intervals constructed with the full model and five I-MATA based methods are compared in Figure 5.4. The average widths provided by the full model were slightly larger than those by the model-averaged methods. The shortest average widths were produced by I-MATA-S again, however in this case the difference between I-MATA-S and the full model is small. The width of intervals constructed by I-MATA-W and two profile-likelihood based methods closer to the full model than to I-MATA-S.

The difference within the model-averaged procedures group is even smaller, but regardless of the outcome rate, the I-MATA-S outperformed all other methods. The outcome probability had a slightly positive effect on the difference in performance of score based and Wald based methods, such that the superiority of the score based procedure fades with outcome approaching the balance at 0.5.

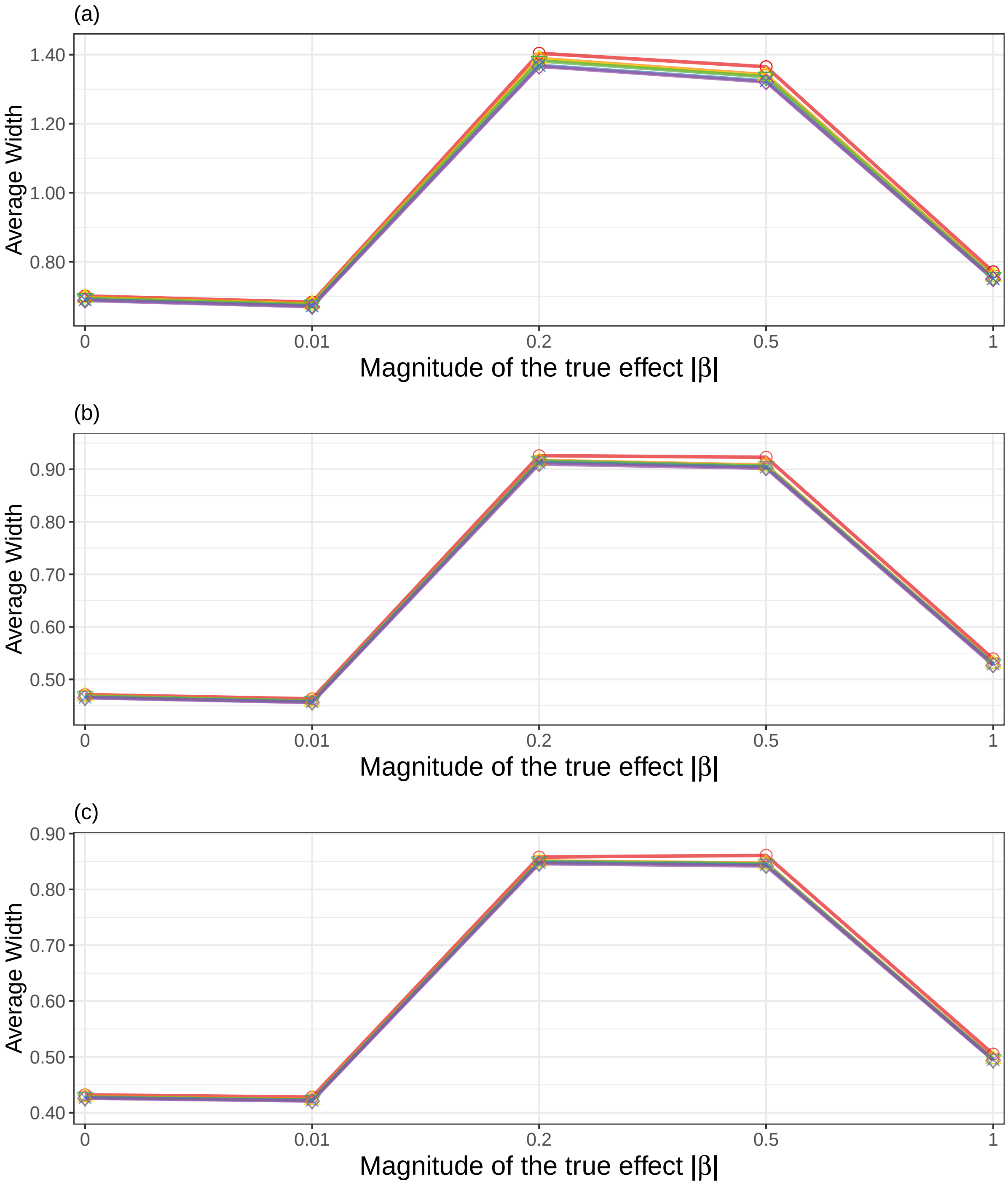


Figure 5.4: Comparison of averaged widths of the full model- ○ , I-MATA-Wpl - □ , I-MATA-PL - △ , I-MATA-W - ▽ , I-MATA-Ws - × , I-MATA-S - ◇ for outcome probabilities: (a) - Pr=0.1, (b) - Pr=0.3, and (c) - Pr=0.5.

5.8 Discussion

The performance of confidence interval procedures evaluated in this thesis varies depending on sample size, number of variables, correlation, and probability of outcome. While increase in sample size or outcome probability improved the performance of model-averaged tail area procedures in terms of tail errors balance and confidence interval width, the stepwise selection related methods demonstrated poor performance even for the less problematic combinations of parameters. Increase in correlation or number of variables adversely affected the width and balance of the compared methods. The procedures in general performed better for regression coefficients having large magnitudes. Among the compared methods, the score function based MATA method applied on a set of models provided by the inclusion fraction approach consistently demonstrated the shortest valid confidence intervals.

Bias of point estimates

Although the main objective of this work is to study and compare the confidence intervals of various methods, we first discuss their point estimation properties, since it is a necessary condition for a procedure with good performance. Comparing the averaged point estimates of the methods, it was found that the full model and the I-MATA method provided the closest results to true effects.

The Occam's window approach also provided a sufficient set of candidate models to get relatively unbiased point estimates. However, only the frequentist model-averaging based on this set provided acceptable results, while Bayesian model-averaging methods provided significant bias towards zero.

The stepwise selection approaches, LASSO, zero-corrected method and model-averaging methods based on the set of the models selected by the backward selection showed large deviations from true effects. The bias produced by these methods was mostly away from zero, because the model selection processes involved in these methods usually select variables with large coefficients, such that variables with low effects have a higher probability to be selected if their estimates are overestimated for small sample. Since the BIC penalty is more conservative than AIC, in small samples the estimates for small effects are more likely to be selected after BIC penalty. As a result, the bias produced by the BIC based stepwise selection was always larger than the bias by the AIC based method. The bootstrap involved in the zero-corrected method and the model-averaging step in E-MATA methods slightly improved the point estimates with respect to the simple stepwise AIC based backward selection; however, the bias produced by these methods was significantly larger than the bias by inclusion fraction based methods.

Confidence interval coverage

Overall, only six methods, the full model and five inclusion fraction based approaches, provided good empirical coverage probabilities, which are equivalent to those of the true model and which almost always maintained the nominal level. All other methods have shown somewhat inferior performance, usually having coverage lower than the nominal level. Since these methods failed at the first stage of comparisons, we only casually examined them. Therefore, in this section we also consider these methods more closely and discuss possible reasons.

The worst results were produced by the Wald-type BMA and the stepwise BIC methods; they provided valid confidence interval coverage less than 20% of the time. Indeed, even though STEP-BIC is a frequentist method, it was developed as an asymptotic approximation to transformation of the Bayesian posterior probability of a considered model. Since BIC penalizes a model of complexity more heavily than AIC, in the finite sample it excludes variables with small effect more often, focusing only on the greatest effects.

But even under fairly light combinations of parameters (low correlation, large sample size, etc.), these two BIC based methods could not always get valid coverage probability for large effects.

The STEP-AIC method proved to be slightly better than the STEP-BIC, but 70% of the time the empirical coverage was below the desired range. The AIC tends to overfit the model's dimension asymptotically and select a more complicated model than BIC (Shibata, 1976); however, for a finite and small sample, it is possible that this feature allowed the AIC method to handle such a task a little better (Zhang, 1993).

The zero-corrected method was proposed to improve performance of conventional stepwise selection using bootstrap algorithm. It showed a noticeable improvement of all confidence interval coverage probabilities. Nonetheless, the confidence intervals for the weakest effects were still below the lower empirical limit. Austin (2008) noted a similar pattern and suggested that this may be due to the low selection frequency of the variables with small effects.

We also assessed how the LASSO method performed as a tool for building confidence intervals for the predictor effects. Although LASSO can be explained with Bayesian theory, in this thesis we considered this method as the simplest and well-known representative of penalized methods. The LASSO method usually performed better with the choice of the model than the stepwise methods (Steyerberg et al., 2000). Its performance is noticeably better than that of the stepwise methods because the coverage probabilities provided by LASSO are much closer to the stated nominal value than those of the stepwise methods. At the same time, it provided acceptable intervals mostly for large effects in a quarter of the parameter combinations, which is worse than the stepwise AIC based method and zero-corrected approach.

The model-averaging technique, as shown by simulations, can improve confidence interval performance. Unfortunately, the presence of the true model, or a model close to it in the group of candidate models, is an important condition for this (Burnham and Anderson, 2002; Turek and Fletcher, 2012). Results from E-MATA methods suggest that averaging of models that can be constructed from variables remaining after the stepwise regression is inefficient. This method was included in the simulation to check if the model-averaged tail area approach can improve the inference after variable selection. The averaging procedure slightly improved the empirical coverage probability of the confidence intervals under all combinations of simulated parameters, but is still not enough to be recommended for practice.

The Occam's window is a better way to select models than the stepwise approaches, since it is more likely to eliminate the redundant variables (Wang et al., 2004). However, Genell et al. (2010) showed that the probability of choosing correct variables is not much different from stepwise regression. Since Occam's window allows more complex models to break into a group of candidate models, the results of B-MATA have surpassed all previous methods described in this section. Wasserman (2000) pointed out that the BMA algorithm is asymptotically consistent in choosing the correct model, but the accuracy of the method in a finite and small sample is poorly understood. Our results showed that as sample size increased, the performance of the method decreased. However, since the maximum sample size we tested was not too large, these results did not disprove the asymptotic theory, but only illustrated that the principle of parsimony applied in the BMA algorithm may have a negative effect on the validity of the confidence intervals. The B-MATA based methods also showed good coverage for low correlation cases, but in more complex settings they demonstrated chronic underestimation.

All methods based on the inclusion fraction performed better than the other methods, providing results that are competitive with those from the full and the true models. In general, the I-MATA based method showed good and stable coverage not affected by any complications of the parameter combinations. Of the five methods, only two profile-likelihood based methods showed slightly weaker results, which is consistent with Kabaila et al. (2016) results. The difference between the empirical coverage probabilities did not exceed 1.4%, indicating that the methods successfully provide valid confidence intervals.

Tail errors

The I-MATA based approaches and the full model provided relatively well-balanced tail errors. The differences between the methods are minor, especially within the group of the inclusion fraction based procedures. The full model provided slightly better balance on average but provided more tail errors with differences that were equal to or exceeded 1%.

Out of the remaining methods, we cannot single out any approach that has balanced tail errors. Even if we consider only the variables for which these methods provided valid coverage probability, the differences between two tail errors were usually higher for any of the I-MATA based methods.

Average width

Since the results of the I-MATA based and full model approaches are similar in terms of validity and balance of the tail errors, the last decisive factor is the length of the confidence intervals. In terms of interval width, throughout all simulations the order of the methods was quite stable. The full model was the worst, producing the largest average width for all parameters. Of the five I-MATA based methods, the profile-likelihood technique did not differ much from the Wald, while the substitution of the Wald standard deviations by the ones based on the profile-likelihood confidence interval construction method only worsened the results. The shortest confidence intervals were provided by two methods based on the score function. Replacing the Wald standard errors by the score based standard errors only slightly reduced the confidence intervals width, while the I-MATA-S showed the smallest intervals, which indicates greater precision of the method.

Some of the methods showed poor results of validity and balance of confidence intervals, with larger averaged confidence intervals. For example, the average length of the confidence intervals built with LASSO in 76% of cases was higher than that of the score function based MATA method, and for ZERO-C this number was 46%. Such results indicate the possibility that the underestimation of the variance was not the only reason for their failure to achieve nominal coverage level, but also the bias of these methods relative to the true point estimates.

Regarding the inadequacy in the assessment of variance, the BMA-W method can be distinguished from methods. The average widths of the confidence intervals of all methods were relatively close to each other, but the confidence intervals built by the BMA were 2 to 3 times shorter than the others. However, unsatisfactory coverage and balance results, make BMA-W the least attractive method for building confidence intervals in regression analysis.

5.9 Conclusion

In general, the 50% inclusion fraction based methods for model-averaging with the score function approach performed best in scenarios considered in the simulation study. Regardless of changes in parameters that can affect the analysis, the two inclusion fraction based MATA methods with score function, I-MATA-S and I-MATA-Ws, demonstrated consistently good coverage. For these methods, the magnitude of the effect did not affect the coverage rate, which was not true for the other methods. The predictors' effects of 0.01 or 0.2 may not be very important compared with effects larger than 1; however, model-averaging ensures that the confidence interval is always valid even for small effects. Of all the methods demonstrating valid confidence intervals, I-MATA-S built the most narrow intervals. The advantage of this method was most pronounced with decreasing sample size or increasing number of variables. The maximal difference in length did not exceed 10%; however, it is up to a researcher to decide whether such an improvement is clinically important.

Chapter 6

R-PACKAGE

6.1 Introduction

To facilitate data analysis using the proposed methods we have developed a convenient tool for constructing model-averaged confidence intervals. This should increase the adoption of model-averaging methodology for model building and for obtaining confidence intervals based on the methods proposed in this thesis.

Even though the idea of constructing model-averaged confidence intervals in the frequentist framework has been known for many years, we were able to find only two packages that could do this in R. The first package `MuMIn`, developed by Barton (2009), allows one to construct confidence intervals proposed by Burnham and Anderson (2002). It is a large package that can be applied to various types of regression, allows the use of different information criteria, and also has the option to select a set of candidate models based on the cumulative weight. Nevertheless, Burnham and Anderson (2002) stated that their method was based on the assumptions that weights were known constants. Since this assumptions often violated, the unconditional confidence interval has poor coverage properties (Hjort and Claeskens, 2003; Claeskens and Hjort, 2008; Turek and Fletcher, 2012; Fletcher and Turek, 2012).

The second package, `MATA`, was developed by Turek (2015), and allows one to calculate Wald type model-averaged tail area confidence intervals. Although this package is undoubtedly useful, we find it inconvenient to use since it neither selects the candidate models, nor estimates coefficients and weights, because all this information has to be provided by the user.

Since the `MuMIn` package provides invalid confidence intervals, and the `MATA` package is not user-friendly and limited by the Wald type intervals, we have developed an R-package `MATACI` that is easy to use that can provide confidence intervals based not only on the Wald based, but also the profile-likelihood and score function based methods. In the next sections we describe in detail the `MATACI` package and its capabilities, as well as discuss its limitations and further updates necessary in later versions of the package.

6.2 The MATACI package

We have implemented the proposed procedures into a user friendly R package, `MATACI`. The simulation study was conducted using this package. The package allows one to choose the method for candidate model selection and preferred confidence interval construction method.

6.2.1 The MATACI function

The package `MATACI` contains the main '`mataci`' function and several supporting functions which will be described in Section 6.2.2. The '`mataci`' function is applied by the statement `mataci(formula, data, nboot = 1000, selection = "Freq", cim = "Score", ci = 0.95, par = F)`. First, it selects the set of candidate models in accordance with the user's requested method, inclusion fraction or Occam's window. Then, it estimates confidence interval using a selected variant of a model-averaging tail area approach. A short description of the agruments used in the '`mataci`' function is presented in Table 6.1.

Table 6.1: Description of the ‘mataci’ function arguments.

Usage	
`mataci(formula, data, nboot = 1000, selection = "Freq",)`	
`cim = "Score", ci = 0.95, par = F)`	
Arguments	
`formula`	an object of class ‘formula’: a symbolic description of the full model to be fitted.
`data`	an object of class ‘data.frame’ (or object coercible by ‘as.data.frame’ to a data frame) containing the variables in the full model.
`nboot`	the number of bootstrap replicates.
`selection`	a description of the method for selection of candidate models to be used in the model-averaging process. If it is set to ‘Freq’ the inclusion fraction method is applied. If it is set to ‘Bayes’ the Occam’s window is used to select candidate models.
`cim`	a description of the method for construction of MATA confidence intervals to be used by the function. The possible options for this argument are ‘Wald’, ‘PL’, ‘Score’, ‘Wald-S’, and ‘Wald-PL’.
`ci`	the confidence level of the required interval.
`par`	logical string; if applied it allows one to use parallel processing for model-averaging estimation.

The first two arguments of the function ‘`formula`’ and ‘`data`’ define the full model and the dataset that are used in the analysis. The number of bootstrap samples is given by the ‘`nboot`’ argument, which is set to 1000 iterations by default.

Despite the fact that the selection of the candidate models using the Occam’s window led to unsatisfactory results, the researchers can still apply this method at their own risk by defining the ‘`selection`’ option equal to `"Bayes"`. The `"Freq"` option corresponds to the frequentists 50% inclusion fraction approach and is set as the default preference.

The '`cim`' argument is responsible for the method by which confidence intervals are constructed. There are five options in total. The `"Score"` option is the default and corresponds to the score based model-averaged tail area confidence interval construction method. In addition to this option, the user can choose `"Wald"`, `"PL"`, `"Wald-PL"`, or `"Wald-S"` corresponding to the remaining four I-MATA methods that also demonstrated good coverage properties of the confidence intervals.

The nominal level of the confidence interval is set by the argument '`ci`', and it is set to 95% by default. The '`par`' option is responsible for parallel computation used in model fitting and ordering. Parallel computing is disabled by default, but can be enabled by the user by setting it to `"TRUE"`. This option greatly reduces the computation time if the number of variables is large; however, with smaller models, it may take longer to load the functions and activate all the cores than a regular, non-parallel calculation.

6.2.2 Secondary functions

The package `MATACI` contains a total of six secondary functions. While the '`mataci`' function only brings together all the functions for model-averaging analysis, the other supporting functions estimate the confidence intervals using the method specified by the user.

The auxiliary package `rootSolve` is responsible for optimizing the final function based on either the profile-likelihood function (2.16) or the score function (4.6) (Soetaert and Herman, 2009). For its correct and fast operation, the optimization functions require starting points from which the algorithm begins to search for the optimal solution. To find such points, the secondary function '`startpoints`' is defined. This function uses the transformation based approximation to MATA-W proposed by Yu et al. (2014) . The approximation is performed automatically and does not require additional calculations from the user. This approximation greatly reduces computational time for searching for confidence intervals.

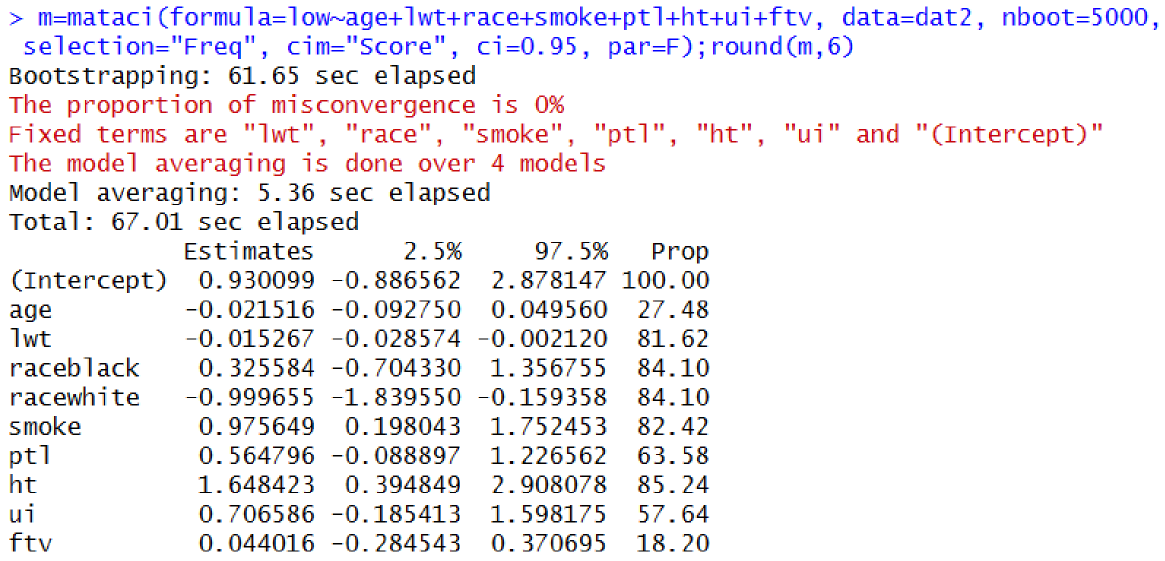
The built-in function ‘confint’ was used to estimate Wald or profile-likelihood confidence intervals (Venables and Ripley, 2002). Since, in the public domain, a working function that allows calculation of score based confidence intervals for a single regression model was not found, we used source code of ‘confint’ to write such a function. The resulting function ‘ScoreRoot’ prepares the score function to be optimized in the ‘waldFcor’ function.

The ‘waldFcor’ function is used in the package to calculate confidence intervals using the I-MATA-Ws and I-MATA-Wpl methods. To estimate confidence intervals based on the I-MATA-Ws method, this function uses the previously mentioned ‘ScoreRoot’ function on each candidate model. Then, using the obtained confidence intervals, it calculates the standard errors, which replace the Wald standard errors within the Wald type MATA confidence intervals estimation. To estimate the confidence intervals based on the I-MATA-Wpl method, it uses the built-in ‘confint’ function with the option of calculating confidence intervals based on the profile-likelihood approach. The further optimization process is no different from I-MATA-Ws, except, of course, the standard errors used.

The remaining three supporting functions ‘waldF’, ‘profLF’, and ‘scoreF’ are responsible for the model-averaged confidence intervals based on the Wald, profile-likelihood, and score functions, respectively. The default algorithm from the ‘confint’ function was also taken as the basis for these functions, but the main part of the code was rewritten so that the confidence interval was calculated not on the basis of the single model, but rather on the basis of all candidate models. Depending on the ‘selection’ argument, the main function ‘mataci’ refers to one of the listed functions to get the corresponding result. Since these three functions are very similar in structure, in the future, most likely, they will be merged into one.

6.2.3 MATACI output

The ‘`mataci`’ function generates a table that contains point estimates and confidence intervals for each variable. In addition, the table presents either the percentages of time that the variables were selected in the bootstrap process if the user chose the frequentists method, or the posterior probabilities if the Bayesian method was chosen.



```
> m=mataci(formula=low~age+lwt+race+smoke+ptl+ht+ui+ftv, data=dat2, nboot=5000,
 selection="Freq", cim="Score", ci=0.95, par=F);round(m,6)
Bootstrapping: 61.65 sec elapsed
The proportion of misconvergence is 0%
Fixed terms are "lwt", "race", "smoke", "ptl", "ht", "ui" and "(Intercept)"
The model averaging is done over 4 models
Model averaging: 5.36 sec elapsed
Total: 67.01 sec elapsed
             Estimates      2.5%     97.5%   Prop
(Intercept)   0.930099 -0.886562  2.878147 100.00
age          -0.021516 -0.092750  0.049560  27.48
lwt          -0.015267 -0.028574 -0.002120  81.62
raceblack     0.325584 -0.704330  1.356755  84.10
racewhite    -0.999655 -1.839550 -0.159358  84.10
smoke         0.975649  0.198043  1.752453  82.42
ptl           0.564796 -0.088897  1.226562  63.58
ht            1.648423  0.394849  2.908078  85.24
ui            0.706586 -0.185413  1.598175  57.64
ftv           0.044016 -0.284543  0.370695  18.20
```

Figure 6.1: Example of inclusion fraction based MATA-S results provided by the mataci function for low birth weight data.

It also informs the user about the time it took the function to estimate the confidence intervals and the size of the set of candidate models used in model-averaging. When the inclusion fraction approach is selected, the function also provides the proportion of bootstrapped models that had convergence issues and variables that were fixed by the inclusion fraction. An example of ‘`mataci`’ function output for a model with five variables is presented in Figure 6.1.

6.3 Limitations and further updates

Since in this work we compared a variety of methods, the code for the simulation was written with the practical purpose of testing the methods on a logistic model. The '`mataci`' function was extracted from the general code, and rewritten so that it was convenient to use. Unfortunately, in this version of the package, the '`mataci`' function is able to perform only tasks similar to those that we set up during the simulations.

First of all, the current version of the package can provide the MATA based confidence intervals only for logistic regressions. Since we have demonstrated that the MATA based methods are able to provide valid confidence intervals for variables in a logistic regression, one would assume that they also should perform well for the linear, Cox, or Poisson regression models. Nevertheless, these models will appear in the next version of the package, after their simulation analysis.

Since the AIC based MATA confidence intervals outperform the intervals based on other criteria, the '`mataci`' function uses the AIC criterion as the basis for the model weighting (Turek and Fletcher, 2012). The same criterion is used in the stepwise regression algorithm involved in the bootstrap process. In the next version, the pool of available criteria will be expanded, which will make the package more flexible and customizable. The BIC and AIC$_c$ criteria will be added. The LASSO method will be available to use instead of the stepwise selection approach within the bootstrap process, since LASSO is often better at eliminating redundant variables.

In the present version, the percentage of the inclusion fraction method is fixed at 50% for all variables. In the future version, we will add the ability to define the inclusion fraction for each variable manually. The ability to define a lower bound for each variable will also be added. When the option is applied, the variable that appeared less frequently than the prespecified minimum will be excluded from further calculations. In addition to these options, the possibility to protect pre-selected variables from bootstrap selection process will be added. Together these options allow the user to determine the framework in which the '`mataci`' function will work, such that in essence they represent prior knowledge.

Chapter 7

ILLUSTRATIVE EXAMPLE

7.1 Introduction

The purpose of this chapter is to illustrate the proposed methods using the R-package developed in this thesis. Birth weight is an important factor in a person's lifespan. Low birth weight increases the chances of infant mortality or birth defects, and may increase the chances of a serious disease in adulthood. A 1986 study conducted at Baystate Medical Center in Springfield, MA aimed to determine risk factors associated with delivering a low birth weight baby (Hosmer et al., 2013). Data on 189 women, 59 of whom had low birth weight babies was collected as a part of this study. In this chapter, the data on the risk factors associated with a chance of having a low birth weight baby were used to demonstrate the performance of the score based model-averaged tail area method and to compare it with other confidence interval construction approaches.

The data contained a binary outcome variable ('LOW'; 0, birth weight $\geq$ 2.5kg; 1, birth weight $<$ 2.5kg), and also included information about mother's age ('AGE'; years), mother's weight at her last menstrual period ('LWT'; pounds), mother's race ('RACE'; 0, White; 1, Black; 2, Other), whether she smoked during pregnancy ('SMOKE'; 0, No; 1, Yes), frequency of premature labour ('PTL'; 0, 1, 2, ...), history of hypertension ('HT'; 0, No; 1, Yes), presence of uterine irritability ('UI'; 0, No; 1, Yes), and the number of physician visits during the first trimester of pregnancy ('FTV'; 0,1, 2, ...). Variables of interest ‘RACE’, ‘SMOKE’, ‘HT’, and ‘UI’ are considered as categorical variables.

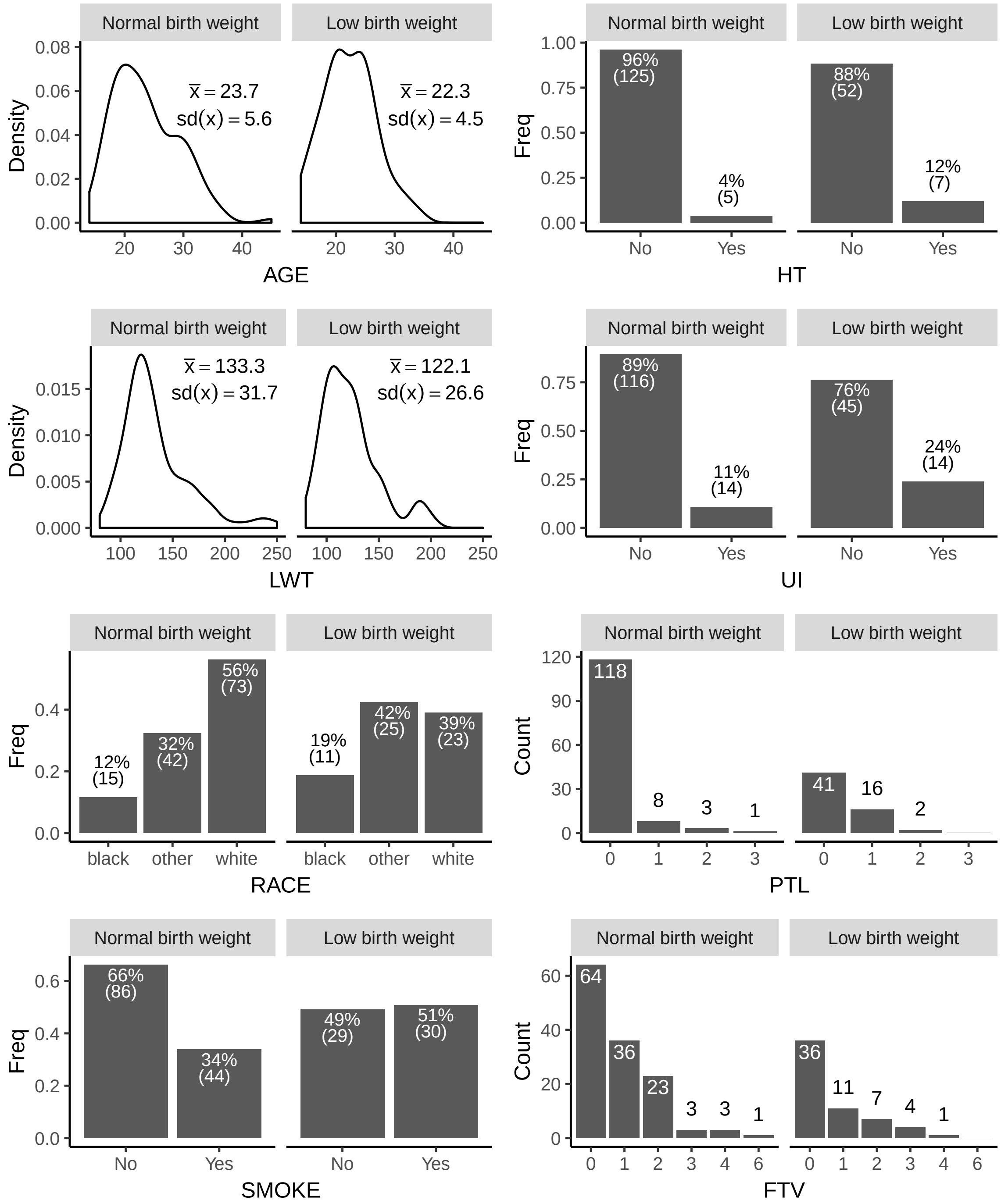


Figure 7.1: Visualization and summary statistics of all risk factors for children’s low birth weight.

7.2 Methods

The obstetric literature has shown that smoking, diet, timely visits to the doctor, and getting prenatal care are important risk factors in pregnancy. The purpose of the original study was to identify which of the possible risk factors could alter the chances of having a baby with a normal weight among patients at the Baystate Medical Center. Considering all 8 variables as potential risk factors, the total number of models for the usual averaging of models will be 2^8 =256. However, the methods for selecting a group of candidate models for model-averaging should significantly reduce the number of considered models.

The simulation results in Chapter 5 has shown that out of 19 methods only six methods, the full model and five inclusion fraction based MATA methods, can be considered as appropriate methods for construction of valid confidence intervals. Our goal is to estimate point estimates and 95% confidence intervals for the risk factors and compare the results of the proposed score based MATA methods with the full model, backward stepwise selection with AIC and BIC penalty, zero-corrected backward selection method, LASSO, the Wald type BMA method and the model-averaged tail area approaches based on the stepwise exclusio (E-MATA) and Occam's window (B-MATA) methods.

7.3 Results

The descriptive statistics for each potential risk factor is graphically visualised in Figure 7.1. From the bar charts we can see that the percentage of smoking habit, presence of hypertension, or uterine irritability is lower in the group of mothers who delivered healthy children. The physician visits during the first trimester of pregnancy and the number of premature labors are count variables, thus they are also presented by bar charts. However, since the categorization will make the model too complex and unable to fit (Hosmer et al., 2013) we considered them as continuous.

Table 7.1: Point estimates of low birth weight risk factors obtained by the full model (FULL), stepwise AIC backward selection (STEP-AIC), stepwise BIC backward selection (STEP-BIC), zero-corrected backward selection (ZERO-C), LASSO and Wald type BMA (BMA-W).

$\hat{\beta}$	FULL	STEP-AIC	STEP-BIC	ZERO-C	LASSO	BMA-W
AGE	-0.02	—	—	-0.02	-0.02	-0.01
LWT	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01
$\text{RACE}_{\text{White}}$[a]	-0.99	-1.01	—	-0.90	-0.97	-0.16
$\text{RACE}_{\text{Black}}$[a]	0.29	0.34	—	0.29	0.31	0.03
SMOKE	0.97	0.98	—	0.92	0.95	0.30
PTL	0.57	0.56	—	0.61	0.57	0.37
HT	1.68	1.64	1.57	1.70	1.66	0.81
UI	0.71	0.71	—	0.66	0.70	0.25
FTV	0.06	—	—	0.01	—	0.00

[a] ‘Other’ was used as a reference group.

The point estimates from the six methods discussed in previous chapters are presented in Tables 7.1 and 7.2. The point estimates from the full model, backward AIC based stepwise selection, LASSO, and the frequentist model-averaging based on the inclusion fraction set of candidate models were relatively close each other. ZERO-C and model-averaging over candidate models selected by the backward selection and Occam’s window demonstrated small deviations from the point estimates provided by the previous methods. The stepwise selection based on BIC method removed most of the predictors from the final model, and provided a parsimonious model with only two coefficients, the mother’s weight at her last menstrual period and history of hypertension. If based on the results of the simulations, we assume that the full model from Table 7.1 and I-MATA method from Table 7.2 are close to the true effect, then the larger bias toward zero was shown by the Wald type BMA.

Confidence intervals with their corresponding widths were calculated by 19 methods, and split into Tables 7.3 to 7.6. Methods demonstrating appropriate confidence interval performance across simulations are presented in Table 7.3. The order of the CI's widths is consistent with simulations presented earlier. Out of six presented methods, the score function based I-MATA confidence intervals (I-MATA-S) have the narrowest width for all risk factors and are followed by the I-MATA-Ws method. The full model and I-MATA-W reported very close results with moderate width, consist with the results of simulations, the profile-likelihood based methods demonstrated the largest confidence intervals out of these six approaches.

Table 7.2: Point estimates of low birth weight risk factors obtained by the frequentist model-averaging procedures based on the candidate models from backward selection (E-MATA), Occam's window (B-MATA) methods and 50% inclusion fraction (I-MATA).

$\hat{\beta}$	E-MATA	B-MATA	I-MATA
AGE	—	-0.05	-0.02
LWT	-0.02	-0.02	-0.02
$\text{RACE}_{\text{White}}$[a]	-0.99	-1.01	-1.00
$\text{RACE}_{\text{Black}}$[a]	0.30	0.29	0.33
SMOKE	1.00	1.00	0.98
PTL	0.62	0.68	0.56
HT	1.57	1.56	1.65
UI	0.77	0.84	0.71
FTV	—	-0.07	0.04

[a] 'Other' was used as a reference group.

Table 7.3: Confidence interval $\{[\text{L,U}]\}$ and width $\{\text{WD}\}$ for low birth weight risk factors obtained by the full model (FULL) and five I-MATA based methods; Wald based I-MATA-W, profile-likelihood based I-MATA-PL, score function based I-MATA-S, Wald based method corrected by the profile-likelihood I-MATA-Wpl, and Wald based method corrected by the score function I-MATA-Ws.

Coefficient	FULL	I-MATA-W	I-MATA-PL	I-MATA-S	I-MATA-Wpl	I-MATA-Ws
AGE	[-0.10, 0.05]	[-0.09, 0.05]	[-0.10, 0.05]	[-0.09, 0.05]	[-0.10, 0.05]	[-0.09, 0.05]
	0.14	0.14	0.14	0.14	0.15	0.14
LWT	[-0.03, -0.001]	[-0.03, -0.002]	[-0.03, -0.003]	[-0.03, -0.002]	[-0.03, -0.003]	[-0.03, -0.002]
	0.03	0.03	0.03	0.03	0.03	0.03
$\text{RACE}_{\text{White}}$[a]	[-1.85, -0.13]	[-1.86, -0.14]	[-1.87, -0.16]	[-1.84, -0.16]	[-1.88, -0.16]	[-1.85, -0.15]
	1.72	1.71	1.71	1.68	1.72	1.69
$\text{RACE}_{\text{Black}}$[a]	[-0.75, 1.34]	[-0.73, 1.38]	[-0.73, 1.38]	[-0.70, 1.36]	[-0.74, 1.39]	[-0.71, 1.36]
	2.10	2.11	2.11	2.06	2.12	2.07
SMOKE	[0.17, 1.76]	[0.18, 1.77]	[0.20, 1.78]	[0.20, 1.75]	[0.20, 1.79]	[0.19, 1.76]
	1.59	1.58	1.58	1.55	1.59	1.56
PTL	[-0.10, 1.24]	[-0.11, 1.24]	[-0.09, 1.26]	[-0.09, 1.23]	[-0.10, 1.27]	[-0.09, 1.23]
	1.34	1.35	1.36	1.32	1.37	1.32
HT	[0.37, 2.99]	[0.34, 2.96]	[0.37, 3.03]	[0.39, 2.91]	[0.36, 3.04]	[0.39, 2.92]
	2.62	2.62	2.66	2.51	2.68	2.53
UI	[-0.20, 1.62]	[-0.20, 1.62]	[-0.21, 1.61]	[-0.19, 1.60]	[-0.22, 1.62]	[-0.19, 1.60]
	1.81	1.82	1.82	1.78	1.84	1.80
FTV	[-0.28, 0.39]	[-0.29, 0.38]	[-0.30, 0.37]	[-0.28, 0.37]	[-0.30, 0.37]	[-0.29, 0.37]
	0.67	0.67	0.67	0.66	0.68	0.66

[a] ‘Other’ was used as a reference group.

Methods producing serious undercoverage of confidence intervals in the simulations are presented in Tables 7.4 to 7.6. We have already pointed out that the point estimates calculated by BMA-W were highly shifted towards zero relative to other methods. The same trend can be detected in the estimation of the confidence intervals in Table 7.4. Moreover, the signs of the upper limit of the confidence interval for the ‘RACE’ effect (white vs. other) and the lower limits for ‘SMOKE’ and ‘HT’ were reversed in relation to limits of other methods. Such a difference may affect the final decision on the statistical significance of the risk factor if one decides to rely on the values of the confidence intervals.

Table 7.4: Confidence interval $\{[\text{L,U}]\}$ and width $\{\text{WD}\}$ for low birth weight risk factors obtained by the stepwise AIC (STEP-AIC) based and stepwise BIC (STEP-BIC) based selection methods, zero-corrected bootstrap (ZERO-C), LASSO and Wald type Bayesian model-averaging (BMA-W).

Coefficient	STEP-AIC	STEP-BIC	ZERO-C	LASSO	BMA-W
AGE	—	—	[-0.11, 0.00]	[-0.09, 0.05]	[-0.05, 0.03]
			0.11	0.14	0.08
LWT	[-0.03, -0.002]	[-0.03, -0.004]	[-0.04, 0.00]	[-0.03, -0.001]	[-0.03, 0.01]
	0.03	0.02	0.04	0.03	0.03
$\text{RACE}_{\text{White}}$[a]	[-1.97, 0.00]	—	[-2.04, 0.00]	[-1.82, -0.12]	[-0.94, 0.63]
	1.97		2.04	1.71	1.57
$\text{RACE}_{\text{Black}}$[a]	[-0.89, 1.51]	—	[-0.94, 1.53]	[-0.74, 1.35]	[-0.42, 0.49]
	2.40		2.47	2.10	0.90
SMOKE	[0.20, 1.76]	—	[0.00, 1.94]	[0.16, 1.74]	[-0.60, 1.20]
	1.57		1.94	1.58	1.81
PTL	[-0.11, 1.23]	—	[0.00, 1.87]	[-0.10, 1.24]	[-0.50, 1.23]
	1.34		1.87	1.34	1.73
HT	[0.34, 2.94]	[0.28, 2.86]	[0.00, 3.41]	[0.36, 2.95]	[-0.95, 2.56]
	2.60	2.56	3.41	2.59	3.52
UI	[-0.20, 1.61]	—	[0.00, 1.90]	[-0.20, 1.61]	[-0.65, 1.15]
	1.81		1.90	1.81	1.80
FTV	—	—	[-0.42, 0.39]	—	[-0.05, 0.05]
			0.81		0.10

[a] ‘Other’ was used as a reference group.

As for the other methods presented in the Tables 7.4 and 7.5, STEP-AIC and based on it MATA methods exclude ‘AGE’ and ‘FTV’ from the model in a stepwise process, while LASSO proposed to eliminate the ‘FTV’ only. Unlike other approaches, the zero-corrected method based on 5,000 bootstrap iterations provided very wide intervals. Since the zero-corrected method replaces the excluded effects with zeros, and uses the percentage method for confidence interval estimation, the confidence intervals of most of the effects had zero as their lower or upper limit, which can complicate the interpretation of confidence intervals.

Table 7.5: Confidence interval $\{[\text{L,U}]\}$ and width $\{\text{WD}\}$ for low birth weight risk factors obtained by five model-averaging CI construction methods using set of candidate models obtained from backward AIC selection approach; Wald based E-MATA-W, profile-likelihood based E-MATA-PL, and score function based E-MATA-S.

Coefficient	E-MATA-W	E-MATA-PL	E-MATA-S
AGE	—	—	—
LWT	[-0.03, -0.002] 0.03	[-0.03, -0.00] 0.03	[-0.03, -0.00] 0.03
$\text{RACE}_{\text{White}}$[a]	[-1.85, -0.12] 1.73	[-1.85, -0.14] 1.74	[-1.84, -0.13] 1.71
$\text{RACE}_{\text{Black}}$[a]	[-0.76, 1.37] 2.13	[-0.77, 1.37] 2.14	[-0.74, 1.35] 2.09
SMOKE	[0.19, 1.80] 1.60	[0.21, 1.81] 1.61	[0.20, 1.78] 1.58
PTL	[-0.06, 1.30] 1.36	[-0.04, 1.33] 1.36	[-0.03, 1.29] 1.32
HT	[0.26, 2.89] 2.63	[0.29, 2.96] 2.67	[0.32, 2.84] 2.52
UI	[-0.15, 1.68] 1.82	[-0.15, 1.68] 1.83	[-0.13, 1.66] 1.78
FTV	—	—	—

[a] ‘Other’ was used as a reference group.

Comparing the widths of the confidence intervals in Table 7.3 with the widths estimated by I-MATA-S or I-MATA-Ws, the I-MATA-S intervals were mostly shorter than the intervals from other methods. The confidence intervals produced by the B-MATA methods were shorter than the analogous intervals obtained from the I-MATA methods (Table 7.3). However, based on the results of our simulation study, we would not recommed using Occam’s window for construction of confidence intervals. The B-MATA methods based on Occam’s window set of candidate models had to average 69 models and the E-MATA methods averaged 64 models to get confidence intervals, while the I-MATA method used only 4 models. The inclusion fraction estimation took 15 seconds to select a list of candidate models and another 5 seconds to estimate the score based confidence intervals. The methods that did not involve bootstrapping, such as stepwise selection methods, LASSO, or BMA-W, produced confidence intervals in less than 5 seconds.

Table 7.6: Confidence interval $\{[\text{L,U}]\}$ and width $\{\text{WD}\}$ for low birth weight risk factors obtained by three model-averaging CI construction methods using set of candidate models obtained from Occam's window approach; Wald based B-MATA-W, profile-likelihood based B-MATA-PL, score function based B-MATA-S, Wald based method corrected by the profile-likelihood B-MATA-Wpl, and Wald based method corrected by the score function B-MATA-Ws.

Coefficient	B-MATA-W	B-MATA-PL	B-MATA-S	B-MATA-Wpl	B-MATA-Ws
AGE	[-0.11, 0.02]	[-0.12, 0.02]	[-0.11, 0.02]	[-0.12, 0.02]	[-0.11, 0.02]
	0.14	0.13	0.13	0.14	0.13
LWT	[-0.03, -0.002]	[-0.03, -0.002]	[-0.03, -0.003]	[-0.03, -0.003]	[-0.03, -0.003]
	0.03	0.03	0.03	0.03	0.03
$\text{RACE}_{\text{White}}$[a]	[-1.86, -0.16]	[-1.88, -0.17]	[-1.85, -0.17]	[-1.89, -0.17]	[-1.85, -0.16]
	1.71	1.71	1.68	1.72	1.69
$\text{RACE}_{\text{Black}}$[a]	[-0.78, 1.35]	[-0.78, 1.35]	[-0.75, 1.33]	[-0.79, 1.36]	[-0.76, 1.34]
	2.13	2.13	2.08	2.14	2.10
SMOKE	[0.17, 1.81]	[0.19, 1.82]	[0.18, 1.79]	[0.18, 1.83]	[0.18, 1.80]
	1.63	1.64	1.61	1.65	1.62
PTL	[0.01, 1.35]	[0.03, 1.37]	[0.03, 1.33]	[0.02, 1.37]	[0.03, 1.34]
	1.34	1.34	1.30	1.35	1.31
HT	[0.25, 2.88]	[0.28, 2.94]	[0.31, 2.83]	[0.27, 2.95]	[0.30, 2.83]
	2.63	2.66	2.52	2.68	2.53
UI	[-0.06, 1.73]	[-0.06, 1.73]	[-0.04, 1.71]	[-0.07, 1.74]	[-0.05, 1.72]
	1.79	1.79	1.75	1.80	1.77
FTV	[-0.40, 0.26]	[-0.41, 0.26]	[-0.39, 0.26]	[-0.41, 0.26]	[-0.39, 0.26]
	0.66	0.66	0.65	0.67	0.65

[a] 'Other' was used as a reference group.

According to our simulation study, the inclusion fraction based MATA intervals obtained from optimization of score function were stably providing the narrowest reliable confidence intervals, whose narrow widths were also demonstrated in the low birth weight example. Therefore, the score function based I-MATA approaches are recommended for analysis of this data and construction of confidence intervals.

Chapter 8

SUMMARY

We presented different confidence interval construction methods and discussed how model uncertainty affects the inference validity of the existing confidence interval construction methods in Chapters 1 and 2. In Chapters 3 and 4 we proposed frequentist method for candidate model set selection and the score based model-averaged tail area confidence interval construction method. Chapter 5 compared the methods using Monte Carlo simulations for small sample sizes, demonstrating that the proposed methods provided valid and balanced confidence intervals that have greater precision across all other methods. Chapter 6 described the R-package that allows one to apply the model-averaged tail area type methods on any data with a binary outcome. In Chapter 7, the methods compared in Chapter 5 are applied to a dataset from a real study. In this chapter, we summarize the main results of our study and recommendations, discuss the limitations of this research, and the directions for future research.

8.1 Introduction

The primary objective of this thesis was to develop and evaluate score function based confidence interval construction methods for model-averaged estimators, as well as suggest appropriate methods for selecting candidate models. The focus was on logistic regression in a small sample and frequentist framework. We also discuss limitations of this study and suggest potential areas for future research.

8.2 Main findings and recommendations

In addition to five inclusion fraction based MATA methods, we also included 16 methods in the simulation study for comparison in the context of logistic regression models. Some of the methods share a similar underlining idea, but all together, except the Wald type Bayesian model-averaged method, they represent a variety of methods in the frequentist framework. The methods were compared based on empirical coverage, tail errors, and averaged width of confidence intervals that reflected the validity, balance, and accuracy of the produced intervals. In addition, we also compared averaged point estimates to ensure their acceptability and the absence of serious bias.

The results showed that out of all compared methods only the full model and five 50% inclusion fraction based methods stably produced balanced confidence intervals with empirical coverage close to the nominal 95% coverage. This means that among the three methods for candidate models set selection, the inclusion fraction with natural cut-off point of 50% was the only method that provided adequate coverage properties of MATA based methods. At the same time, it significantly reduced the set of models used in model-averaging. We demonstrated that the simple 50% threshold is sufficient to get valid results; however, if prior knowledge, hypothesis or scientific sense allows, the inclusion fraction might be selected for each variable separately, which can change the final set of the models and improve confidence interval accuracy.

Among the six methods that showed reliable confidence intervals, the inclusion fraction based MATA-S method is recommended. The greater precision demonstrated by this method, becomes more recognizable for small sample sizes, highly imbalanced outcome, or high correlation among predictors. The score based model-averaging method was the most computing intensive out of the acceptable methods but provided the best results. If one is willing to sacrifice confidence interval accuracy for the sake of time, the MATA-Ws based on inclusion fraction set of candidate models could be recommended. The substitution of Wald standard errors by the score function based standard errors improved the confidence interval precision, even though its averaged width was always slightly larger than the width of the MATA-S method. The corrected MATA intervals with the score standard errors not only performed better than the profile-likelihood based MATA method, but also was computationally less costly than the profile-likelihood variant of MATA approach. Overall, the profile-likelihood based MATA method demonstrated very close results to the Wald-type MATA, while the correction of Wald type MATA by the profile-likelihood standard errors did not paid off. The averaged width of I-MATA-Wpl method was consistently larger than the averaged width of other model-averaged tail area based methods. Although we do not recommend using the profile-likelihood methods because of their compromised accuracy, they can still be used for construction of valid confidence intervals.

The performance of Wald type intervals with the Bayesian model-averaging method demonstrated the least reliable results out of all methods. While the rest of the failed methods only provided untrustworthy confidence intervals, the BMA-W method also showed biased point estimates. Thus, it is not recommended to use this method to construct confidence intervals. As for the STEP-AIC, STEP-BIC and LASSO methods that combined model selection and sequential construction of the confidence intervals, since the intervals provided by these methods usually are overoptimistic, we advice against using them if the goal is to get valid inference for the predictors.

8.3 Study assumptions and limitations

In previous chapters we proposed the new score based model-averaged confidence interval construction method and demonstrated its performance. The I-MATA-S method does not need any assumptions about the distribution of regression coefficient estimates; however, it relies on several assumptions related to the initial data. First, we assume that all variables required for fitting the real model are available to be included into the set of the candidate models, but we do not insist on the presence of the true model in the set of candidate models. In addition, we need to assume that the full model can be fitted without any convergence issues. Both assumptions coupled with the 50% inclusion fraction method allows the MATA-S method to construct valid intervals for each variable.

The simulation results in Chapter 5 showed that changes in the sample size, the number of predictors, correlation among predictors, and outcome probability may affect the width of the confidence intervals constructed by the I-MATA-S method; however, the validity and balance of the constructed intervals and superiority in precision over other approaches remains unchanged. This indicates that the I-MATA-S method is very stable for small samples and moderate correlation among predictors, but in cases of high-dimensional modelling the 50% inclusion fraction step may not be applicable without appropriate adjustments, while MATA step might have some estimation difficulties with highly correlated data. In addition, if the outcome probability is far from 50%, at least one of the considered models may encounter the separation phenomenon, that makes model fitting problematic. This phenomenon is known to affect confidence intervals through inflated standard errors in single modelling, and we expect that it also may have a detrimental effect on the performance of model-averaging methods.

We limited evaluation on the performance of the methods in the context of logistic regression. Nevertheless, we believe that I-MATA-S and I-MATA-Ws methods can also be successfully applied to other members of this group, such a Poisson or Cox regressions, in a straightforward fashion.

8.4 Directions for future research

We showed that the 50% inclusion fraction based MATA methods outperformed the most popular confidence interval construction methods, such as stepwise regression and LASSO approach. We demonstrated that out of all evaluated model-averaging procedures, the model-averaging based on score function provides the shortest, valid confidence intervals and acceptable point estimates. We also pointed out the main limitations and parametric assumptions of this study. In this study the maximal correlation between two predictors we tested was around 0.5, and the proposed method performed well under this condition. However, we expect that averaging of regular regression estimates may lead to biased and invalid inference in presence of multicollinearity. Therefore, the development and extension of the methodology for the I-MATA methods adjustment in violation of one of parametric assumptions, such as absence of multicollinearity or low-dimensionality of the data, might be an interesting research area.

The application of the ridge regression method in model-averaging settings demonstrated very promising results in terms of accurate predictions by Yeon et al. (2010) and Zhao et al. (2018). However, the usefulness of the method for constructing of valid confidence intervals remains to be explored. The application of ridge regression models in I-MATA confidence interval construction procedure may correct the detrimental effect of multicollinearity.

The low-dimensionality of the data and absence of separation phenomenon are two limitations we mentioned in the previous section. Simulation studies of logistic regressions for small samples sizes may face the phenomenon of separation. It is desirable to remove the models, which had a separation problem from the simulation process, however this may cause informative missingness of final results (Steyerberg et al., 2011). The Firth penalized regression can overcome separation phenomenon. It corrects small-sample bias in point estimates. However, the Firth penalty affects the point estimates even if a model did not have any convergence problems, thus application of it for all simulated models may affect the reliability of simulation results. The selective application of the Firth correction requires algorithm for detection of separation problem (van Smeden et al., 2016). A guidance on the correct use of the Firth penalty in simulation and bootstrap processes is needed.

While the separation problem can be solved by the Firth penalized regression, the model-averaging for high-dimensional data has only recently been studied (Puhr et al., 2017). Ando and Li (2014) and Ando et al. (2017) proposed a two-stage model-averaging procedure for linear and generalized linear models, respectively. The basic idea is to select the candidate models by splitting the predictors into groups based on the absolute marginal correlation between the outcome and predictors, and keeping only groups with high correlation. This technique demonstrated quite accurate results for estimation of the outcome, but its performance in confidence interval construction was not studied. However, similar correlation-based approaches might be useful in defining the candidate models for MATA-based confidence intervals under high-dimensional settings. Besides this, the candidate models set might be defined by selecting appropriate range for the shrinkage parameter in penalized regressions.

We did not consider the effect of missing data on model-averaging, but we believe that this topic deserves further research. The missing observations may affect the selection of candidate models, and in turn impact performance of confidence intervals. Cavanaugh and Shumway (1998) proposed using the expectation-maximization algorithm to estimate a variant of AIC in presence of missing data. Their AIC variant allows one to select the models in case of incomplete outcome data, and it also can be used for model-averaging. Claeskens and Consentino (2008) proposed modification of the AIC for cases when the covariate data is incomplete. While these two AIC variants require estimation of the likelihood function with expectation maximization algorithm, Hens et al. (2006) proposed weighted AIC that uses inverse selection probabilities for reweighing the complete observations by analogy with the weighted Horvitz-Thompson estimator. Schomaker et al. (2010) compared the results of a single model selection based on the weighted AIC and frequentist model-averaging after multiple imputation. It was shown that model-averaging over imputed data provides slightly better estimation efficiency than the single model. Schomaker and Heumann (2014) suggested using multiple imputation inference over bootstrapped datasets to estimate confidence intervals. Schomaker and Heumann (2018) demonstrated that using multiple imputation inference after bootstrapping provides better results than doing bootstrap after multiple imputation, since the latter imposes symmetry on the estimated intervals. Brand et al. (2019) showed that even single imputation nested within the bootstrap percentile method may provide valid inferences. The AIC adjusted for data missingness and multiple imputation can be used to extend the MATA methods to data with missing observations.

The selection of candidate models is also an important part of the model-averaging procedure. We demonstrated that the AIC backward selection based 50% inclusion fraction approach is able to significantly reduce the number of considered models, while saving the prespecified coverage probability of MATA based confidence intervals. However, we think that better candidate models set selection procedures are still desirable. One of the possible ways to improve the candidate models set is to use more advanced variable selection techniques rather than simple backward stepwise selection.

This thesis focused on methods for cross-sectional data. Longitudinal studies in practice are popular. Implementation of the model-averaging approaches in these settings would be valuable. In the last decade many different model-averaging techniques for longitudinal data were developed. For example, Fan and Wang (2015) applied the BMA method to the set of longitudinal regression models with autoregressive errors and demonstrated its acceptable performance in future predictions. Zhang et al. (2014) proposed the model-averaging procedure for linear mixed-effects models under the frequentist setting to provide asymptotically optimal estimators in terms of minimization of squared errors. Additional information on model-averaging methods in mixed models can be found in Fletcher (2018). The FMA method based on the leave-subject-out cross-validation approach (Gao et al., 2016) demonstrated good prediction properties in both longitudinal and time series data; however, no corresponding methods are currently available for constructing confidence intervals in the context of longitudinal data. Yang et al. (2017) proposed FIC for the generalized estimating equation approach using the quasi-likelihood function, as well as the modified confidence intervals for focused averaged estimator developed by Hjort and Claeskens (2003). The extension and modification of MATA methodology to longitudinal settings, and comparison of its performance with existing methods both for prognostic and diagnostic purposes in finite and small samples is a promising area for future research.

BIBLIOGRAPHY

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. Wiley New York.

Agresti, A. 2011. Score and pseudo-score confidence intervals for categorical data analysis. *Statistics in Biopharmaceutical Research*, 3(2):163–172.

Aitkin, M. A. 1974. Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, 16(2):221–227.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In Petrov BN., C. F., editor, *Proceedings of the Second International Symposium on Information Theory*, page 267281. Budapest: Akademiai Kiado.

Altman, D. G. and Andersen, P. K. 1989. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8(7):771–783.

Ando, T. and Li, K.-C. 2014. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265.

Ando, T., Li, K.-C., et al. 2017. A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, 45(6):2654–2679.

Austin, P. C. 2008. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Statistics in Medicine*, 27(17):3286–3300.

Austin, P. C. and Tu, J. V. 2004a. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, 57(11):1138–1146.

Austin, P. C. and Tu, J. V. 2004b. Bootstrap methods for developing predictive models. *The American Statistician*, 58(2):131–137.

Bachoc, F., Leeb, H., and Pötscher, B. M. 2017. Valid confidence intervals for post-model-selection predictors. *arXiv: 1412.4605v3*.

Barnard, G. A. 1963. New methods of quality control. *Journal of the Royal Statistical Society. Series A (General)*, 126(2):255–258.

Barton, K. 2009. MuMIn: multi-model inference, R package version 0.12. 0. *http://r-forge. r-project. org/projects/mumin/*.

Bendel, R. B. and Afifi, A. A. 1977. Comparison of stopping rules in forward stepwise regression. *Journal of the American Statistical Association*, 72(357):46–53.

Bengtsson, H. 2018. *matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors).* R package version 0.54.0.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. 2013. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Berk, R., Brown, L., and Zhao, L. 2010. Statistical inference after model selection. *Journal of Quantitative Criminology*, 26(2):217–236.

Bolker, B. M. 2008. *Ecological Models and Data in R*. Princeton University Press.

Brand, J., van Buuren, S., le Cessie, S., and van den Hout, W. 2019. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine*, 38(2):210–220.

Breiman, L. 1995. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. 1997. Model selection an integral part of inference. *Biometrics*, 53(2):603–618.

Bühlmann, P. 2013. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Burnham, K. P. and Anderson, D. R. 2002. *Model Selection and Multimodel Inference: a practical Information-Theoretic Approach*. Springer Science & Business Media.

Burnham, K. P. and Anderson, D. R. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.

Bursac, Z., Gauss, C. H., Williams, D. K., and Hosmer, D. W. 2008. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3(1):17.

Cavanaugh, J. E. and Shumway, R. H. 1998. An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, 67(1):45–65.

Claeskens, G. and Consentino, F. 2008. Variable selection with incomplete covariate data. *Biometrics*, 64(4):1062–1069.

Claeskens, G. and Hjort, N. L. 2008. *Model Selection and Model Averaging*. Cambridge University Press; Cambridge.

Courvoisier, D. S., Combescure, C., Agoritsas, T., Gayet-Ageron, A., and Perneger, T. V. 2011. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, 64(9):993–1000.

Cox, D. R. and Hinkley, D. V. 1979. *Theoretical Statistics*. CRC Press.

Derksen, S. and Keselman, H. J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282.

Efroymson, M. A. 1960. Multiple regression analysis. In Ralston, A. and Wilf, H. S., editors, *Mathematical Methods for Digital Computers*, pages 191–203. John Wiley, New York.

Engle, R. F. 1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics*, 2:775–826.

Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, T.-H. and Wang, G.-T. 2015. Bayesian model averaging in longitudinal regression models with AR(1) errors with application to a myopia data set. *Journal of Statistical Computation and Simulation*, 85(8):1667–1678.

Fang, X., Li, R., Kan, H., Bottai, M., Fang, F., and Cao, Y. 2016. Bayesian model averaging method for evaluating associations between air pollution and respiratory mortality: a time-series study. *BMJ Open*, 6(8):e011487.

Feinstein, A. R. 1996. *Multivariable Analysis: an Introduction*. Yale University Press.

Fernández-Niño, J. A., Hernández-Montes, R. I., and Rodríguez-Villamizar, L. A. 2018. Reporting of statistical regression analyses in biomédica: A critical assessment review. *Biomédica Instituto Nacional de Salud*, 38:173–9.

Flack, V. F. and Chang, P. C. 1987. Frequency of selecting noise variables in subset regression analysis: a simulation study. *The American Statistician*, 41(1):84–86.

Fletcher, D. 2018. *Model averaging*, pages 82–83. SpringerBriefs in Statistics. Springer.

Fletcher, D. and Turek, D. 2012. Model averaged profile likelihood intervals. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(1):38–51.

Frank, L. E. and Friedman, J. H. 1993. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Freedman, D. A., Navidi, W., and Peters, S. C. 1988. On the impact of variable selection in fitting regression equations. In *On Model Uncertainty and its Statistical Implications*, pages 1–16. Springer.

Friedman, J., Hastie, T., and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Furnival, G. M. and Wilson, R. W. 1974. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

Gao, Y., Zhang, X., Wang, S., and Zou, G. 2016. Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, 192(1):139–151.

Genell, A., Nemes, S., Steineck, G., and Dickman, P. W. 2010. Model selection in medical research: a simulation study comparing Bayesian model averaging and stepwise regression. *BMC Medical Research Methodology*, 10(1):108.

Greenland, S., Daniel, R., and Pearce, N. 2016. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology*, 45(2):565–575.

Hansen, B. E. 2007. Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Harrell, F. E. 2015. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.

Harrell, F. E., Lee, K. L., Matchar, D. B., and Reichert, T. A. 1985. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69(10):1071–1077.

Heinze, G. and Schemper, M. 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419.

Hens, N., Aerts, M., and Molenberghs, G. 2006. Model selection for incomplete and design-based samples. *Statistics in Medicine*, 25(14):2502–2520.

Hjort, N. L. and Claeskens, G. 2003. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Hoerl, A. E. and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. 1999. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Hosmer, D., Lemeshow, S., and Sturdivant, R. X. 2013. *Applied Logistic Regression*. John Wiley & Sons.

Hurvich, C. M. and Tsai, C. 1990. The impact of model selection on inference in linear regression. *The American Statistician*, 44(3):214–217.

Hurvich, C. M. and Tsai, C.-L. 1989. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Izrailev, S. 2014. *tictoc: Functions for timing R scripts, as well as implementations of Stack and List structures.* R package version 1.0.

Kabaila, P. 2005. On the coverage probability of confidence intervals in regression after variable selection. *Australian & New Zealand Journal of Statistics*, 47(4):549–562.

Kabaila, P., Welsh, A., and Abeysekera, W. 2016. Model-averaged confidence intervals. *Scandinavian Journal of Statistics*, 43(1):35–48.

Kabaila, P., Welsh, A., and Mainzer, R. 2017. The performance of model averaged tail area confidence intervals. *Communications in Statistics – Theory and Methods*, 46(21):10718–10732.

Kass, R. E. and Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kennedy, W. J. and Bancroft, T. A. 1971. Model building for prediction in regression based upon repeated significance tests. *The Annals of Mathematical Statistics*, 42(4):1273–1284.

Knight, K. and Fu, W. 2000. Asymptotics for LASSO-type estimators. *Annals of Statistics*, 28(5):1356–1378.

Leamer, E. E. 1978. *Specification Searches: Ad hoc Inference with Nonexperimental Data*, volume 53. John Wiley & Sons Incorporated.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. 2016. Exact post-selection inference, with application to the LASSO. *The Annals of Statistics*, 44(3):907–927.

Lee, K. and Koval, J. J. 1997. Determination of the best significance level in forward stepwise logistic regression. *Communications in Statistics – Simulation and Computation*, 26(2):559–575.

Leeb, H. 2006. The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations. In *Optimality*, pages 291–311. Institute of Mathematical Statistics.

Leeb, H. and Pötscher, B. M. 2005. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. 2006. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Leeb, H., Pötscher, B. M., Ewald, K., et al. 2015. On various confidence intervals post-model-selection. *Statistical Science*, 30(2):216–227.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. 2014. A significance test for the LASSO. *The Annals of Statistics*, 42(2):413.

Lukacs, P. M., Burnham, K. P., and Anderson, D. R. 2010. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, 62(1):117.

Madigan, D. and Raftery, A. E. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Newcombe, R. G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, 17(8):857–872.

Park, T. and Casella, G. 2008. The Bayesian LASSO. *Journal of the American Statistical Association*, 103(482):681–686.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Pfeiffer, R. M., Redd, A., and Carroll, R. J. 2017. On the impact of model selection on predictor identification and parameter inference. *Computational Statistics*, 32(2):667–690.

Portnov, B. A., Reiser, B., et al. 2012. High prevalence of childhood asthma in northern israel is linked to air pollution by particulate matter: evidence from GIS analysis and Bayesian model averaging. *International Journal of Environmental Health Research*, 22(3):249–269.

Pötscher, B. M. 1991. Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Puhr, R., Heinze, G., Nold, M., Lusa, L., and Geroldinger, A. 2017. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*, 36(14):2302–2317.

R Core Team 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. 1996. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266.

Raftery, A. E., Madigan, D., and Volinsky, C. T. 1996. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics*, 5:323–349.

Raftery, A. E., Painter, I. S., and Volinsky, C. T. 2005. BMA: an R package for Bayesian model averaging. *The Newsletter of the R Project Volume*, 5:2.

Rao, C. R. 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge University Press.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. 1988. *Applied Regression Analysis: a Research Tool*. Wadsworth & Brooks/Cole Advanced Books & Software.

Richards, S. A. 2008. Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1):218–227.

Roberts, H. V. 1965. Probabilistic prediction. *Journal of the American Statistical Association*, 60(309):50–62.

Rothman, K. J., Greenland, S., Lash, T. L., et al. 2008. *Modern Epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Samaniego, F. J. 2010. *A comparison of the Bayesian and frequentist approaches to estimation*. Springer Science & Business Media.

Sauerbrei, W. 1999. The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):313–329.

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L., and Binder, H. 2015. On stability issues in deriving multivariable regression models. *Biometrical Journal*, 57(4):531–555.

Sauerbrei, W. and Schumacher, M. 1992. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine*, 11(16):2093–2109.

Schomaker, M. and Heumann, C. 2014. Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*, 71:758–770.

Schomaker, M. and Heumann, C. 2018. Bootstrap inference when using multiple imputation. *Statistics in Medicine*, 37(14):2252–2266.

Schomaker, M., Wan, A. T., and Heumann, C. 2010. Frequentist model averaging with missing observations. *Computational Statistics & Data Analysis*, 54(12):3336–3347.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63(1):117–126.

Soetaert, K. and Herman, P. M. 2009. *A Practical Guide to Ecological Modelling. Using R as a Simulation Platform*. Springer. ISBN 978-1-4020-8623-6.

Steyerberg, E. W., Eijkemans, M. J., and Habbema, J. D. F. 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of clinical epidemiology*, 52(10):935–942.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., and Habbema, J. D. F. 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in medicine*, 19(8):1059–1079.

Steyerberg, E. W., Schemper, M., and Harrell, F. E. 2011. Logistic regression modeling and the number of events per variable: selection bias dominates. *Journal of Clinical Epidemiology*, 64(12):1464.

Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. 2014. Post-selection adaptive inference for least angle regression and the LASSO. *arXiv: 1401.3889v2*.

Taylor, J. and Tibshirani, R. 2017. Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. 2016. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Turek, D. 2015. *MATA: Model-Averaged Tail Area Wald (MATA-Wald) Confidence Interval*. R package version 0.3.

Turek, D. and Fletcher, D. 2012. Model-averaged Wald confidence intervals. *Computational Statistics & Data Analysis*, 56(9):2809–2815.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van Smeden, M., de Groot, J. A., Moons, K. G., Collins, G. S., Altman, D. G., Eijkemans, M. J., and Reitsma, J. B. 2016. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1):163.

Venables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Venzon, D. and Moolgavkar, S. 1988. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37(1):87–94.

Viallefont, V., Raftery, A. E., and Richardson, S. 2001. Variable selection and bayesian model averaging in case-control studies. *Statistics in Medicine*, 20(21):3215–3230.

Vittinghoff, E. and McCulloch, C. E. 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6):710–718.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. 1997. Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):433–448.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482.

Walter, S. and Tiemeier, H. 2009. Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology*, 24(12):733–736.

Wang, D., Zhang, W., and Bakhai, A. 2004. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, 23(22):3451–3467.

Wang, H. and Zhou, S. Z. 2013. Interval estimation by frequentist model averaging. *Communications in Statistics – Theory and Methods*, 42(23):4342–4356.

Wang, Q., Koval, J. J., Mills, C. A., and Lee, K.-I. D. 2007. Determination of the selection statistics and best significance level in backward stepwise logistic regression. *Communications in Statistics – Simulation and Computation*, 37(1):62–72.

Wasserman, L. 2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.

Wiegand, R. E. 2010. Performance of using multiple stepwise algorithms for variable selection. *Statistics in Medicine*, 29(15):1647–1659.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Yan, J. 2007. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4):1–21.

Yang, H., Lin, P., Zou, G., and Liang, H. 2017. Variable selection and model averaging for longitudinal data incorporating gee approach. *Statistica Sinica*, pages 389–413.

Yeon, K., Song, M. S., Kim, Y., Choi, H., and Park, C. 2010. Model averaging via penalized regression for tracking concept drift. *Journal of Computational and Graphical Statistics*, 19(2):457–473.

Yu, W., Xu, W., and Zhu, L. 2014. Transformation-based model averaged tail area inference. *Computational Statistics*, 29(6):1713–1726.

Yuan, M. and Lin, Y. 2007. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.

Zhang, C.-H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, S. S. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, P. 1993. On the convergence rate of model selection criteria. *Communications in Statistics – Theory and Methods*, 22(10):2765–2775.

Zhang, X., Zou, G., and Liang, H. 2014. Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 101(1):205–218.

Zhang, Y., Li, R., and Tsai, C.-L. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhao, S., Liao, J., and Yu, D. 2018. Model averaging estimator in ridge regression and its large sample properties. *Statistical Papers*, pages 1–21.

Zou, H. 2006. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendix A

A.1 Data generation

```
# Data generation function
copulaData=function(n, p, rho){
# rho=0.5; p=3;
  if (p==3){
    if(rho==0.5){
      myCop=normalCopula(param=c(0.63, 0.5,
                        0.63), dim=3, dispstr="un");
    }else{
      myCop=normalCopula(param=c(rho), dim=3, dispstr="ex")};
    out=rCopula(n, myCop);
    out[, 1]=qnorm(out[, 1], mean=0, sd=1);
    out[, 2]=qbinom(out[, 2], size=1, prob=0.5);
    out[, 3]=qnorm(out[, 3], mean=0, sd=1);
  }else if(p==5){ if(rho==0.3){
# rho=0.3; p=5;
      myCop=normalCopula(param=c(0.38, 0.3, 0.38, 0.3,
                                      0.38, 0.45, 0.38,
                                      0.38, 0.3,
                                      0.38), dim=5, dispstr="un");
    }else if(rho==0.5){
# rho=0.5; p=5;
      myCop=normalCopula(param=c(0.63, 0.5, 0.63, 0.5,
                                      0.63, 0.71, 0.63,
                                      0.63, 0.5,
                                      0.63), dim=5, dispstr="un");
    }else{
    myCop=normalCopula(param=c(rho), dim=5, dispstr="ex")};
    out=rCopula(n, myCop);
    out[, 1]=qnorm(out[, 1], mean=0, sd=1);
    out[, 2]=qbinom(out[, 2], size=1, prob=0.5);
    out[, 3]=qnorm(out[, 3], mean=0, sd=1);
    out[, 4]=qbinom(out[, 4], size=1, prob=0.5);
    out[, 5]=qnorm(out[, 5], mean=0, sd=1);
  }else if(p==10){if(rho==0){
#rho=0; p=10;
      myCop=normalCopula(param=c(0), dim=10, dispstr="ex");
    }else if(rho==0.3){
# rho=0.3; p=10;
      myCop=normalCopula(param=
                c(0.38, 0.3, 0.38, 0.3, 0.38, 0.3, 0.38, 0.3, 0.38,
                        0.38, 0.45, 0.38, 0.45, 0.38, 0.45, 0.38, 0.45,
                        0.38, 0.3, 0.38, 0.3, 0.38, 0.3, 0.38,
                        0.38, 0.45, 0.38, 0.45, 0.38, 0.45,
                        0.38, 0.3, 0.38, 0.3, 0.38,
                        0.38, 0.45, 0.38, 0.45,
                        0.38, 0.3, 0.38,
                        0.38, 0.45,
                        0.38), dim=10, dispstr="un");
    }else if(rho==0.5){
# rho=0.5; p=10;
      myCop=normalCopula(param=
                c(0.63, 0.5, 0.63, 0.5, 0.63, 0.5, 0.63, 0.5, 0.63,
                        0.63, 0.71, 0.63, 0.71, 0.63, 0.71, 0.63, 0.71,
                        0.63, 0.5, 0.63, 0.5, 0.63, 0.5, 0.63,
                        0.63, 0.71, 0.63, 0.71, 0.63, 0.71,
                        0.63, 0.5, 0.63, 0.5, 0.63,
                        0.63, 0.71, 0.63, 0.71,
                        0.63, 0.5, 0.63,
                        0.63, 0.71,
                        0.63), dim=10, dispstr="un");
    }else{
      myCop=normalCopula(param=c(rho), dim=10, dispstr="ex");
    }
    out=rCopula(n, myCop);
    out[, 1]=qnorm(out[, 1], mean=0, sd=1) ;
    out[, 2]=qbinom(out[, 2], size=1, prob=0.5);
    out[, 3]=qnorm(out[, 3], mean=0, sd=1);
    out[, 4]=qbinom(out[, 4], size=1, prob=0.5);
    out[, 5]=qnorm(out[, 5], mean=0, sd=1);
    out[, 6]=qbinom(out[, 6], size=1, prob=0.5);
    out[, 7]=qnorm(out[, 7], mean=0, sd=1);
    out[, 8]=qbinom(out[, 8], size=1, prob=0.5);
    out[, 9]=qnorm(out[, 9], mean=0, sd=1);
    out[, 10]=qbinom(out[, 10], size=1, prob=0.5);
    out=out[, c(2, 1, 4, 3, 5, 6, 8, 7, 9, 10)];
  }else{
    myCop=normalCopula(param=c(rho), dim=p, dispstr="ex");
    out=rCopula(n, myCop)};
  out=data.frame(out);
  names(out)=sprintf("V%d", 1:p);
  return(data=out)}

### Data generation
#Different scenarios
#N (Prevalence=50%)
#n=100; rho=0.5; p=5; int=-0.15;
#n=300; rho=0.5; p=5; int=-0.15;
#n=500; rho=0.5; p=5; int=-0.15;
#Prevalence (10%, 30%, 50%)
#n=500; rho=0.3; p=5; int=-2.7;
#n=500; rho=0.3; p=5; int=-1.15;
#n=500; rho=0.3; p=5; int=-0.15;
#rho (Prevalence=30%)
#n=300; rho=0; p =5; int=-1.42;
#n=300; rho=0.3; p=5; int=-1.28;
#n=300; rho=0.5;  p= 5; int=-1.18;
#p (Prevalence=30%)
#n=500; rho=0.5; p=3; int =-1.23;
#n=500; rho=0.5; p=5; int=-1.15;
#n=500; rho=0.5; p=10; int=-1.18;
```

```
data=copulaData(n=n , p=p, rho=rho);
if(p==3){
  True=  c(int, 0.01, 0.5, -1);
  data[, c(1, 3)]=t((t(data[, c(1, 3)]) -
        apply(data[, c(1, 3)], 2, mean)) /
        apply(data[, c(1, 3)], 2, sd));
}else{if( p==5 ){
    True=c(int, 0, 0.01, -0.2, 0.5, -1);
    data=data[, c(1, 3, 2, 4, 5)];
    data[, c(1, 2, 5)]=t((t(data[, c(1, 2, 5)]) -
        apply(data[, c(1, 2, 5)], 2, mean)) /
        apply(data[, c(1, 2, 5)], 2, sd));
    names(data)=sprintf("V%d", 1:p);
  }else{if(p==10)
    True=c(int, 0, 0, 0, 0, 0.01, -0.2, 0.5, -0.7, -1, 2.5);
    data[, c(2, 4, 5, 8, 9)]=t((t(data[, c(2, 4, 5, 8, 9)]) -
        apply(data[, c(2, 4, 5, 8, 9)], 2, mean)) /
        apply(data[, c(2, 4, 5, 8, 9)], 2, sd));
    names(data)=sprintf("V%d", 1:p)}};
z=as.matrix(cbind(rep(1, n), data)) %*% True;
pr=1 / (1+exp(-z));
y=rbinom(n=n, size=1, prob=t(pr));
mean(y);
#Data set
data=data.frame((cbind(y, data)));
```

A.2 Function MATACI

```
mataci=function(formula, data, nboot=1000, selection="Freq",
        cim="Wald", ci=0.95, par=F){
  trms=terms(formula, data=data);
  variables=attr(trms, "term.labels");
  out=all.vars(formula)[1];
  n=dim(data)[1];
  rownames(data)=NULL;
  dataN=as.matrix(data[c(variables,out)]);
  alpha=(1-ci)/2;
  forname=c("(Intercept)",variables);
  if (selection=="Freq"){
    tic("Total")
    tic("Bootstrapping")
    cl=makeCluster(detectCores());
    clusterExport(cl, c("dataN", "n", "formula","variables"), envir=
       environment());
    repl1=parLapply(cl=cl, 1:nboot, function(i, dataA=dataN,
        smpl=n, ...){
      #Resampling
      dataB=dataA[sample(nrow(dataA), size=smpl, replace=TRUE), ];
      p=dim(dataB)[2];
      forname=variables;
      #Formula for fitting
      formulaZS=formula;
      mus=glm(formulaZS, family=binomial, data=data.frame(dataB));
      #Stepwise selection
      Sl.us=step(mus, direction="backward", trace=F, k=2);
      options(warn=2);
      test=try(glm(Sl.us$formula, family=binomial,
        data=data.frame(dataB)));
      options(warn=1);
      war=inherits(test, "try-error");
      Slstep.us=summary(Sl.us);
      ZCus=coef(Sl.us);
      ZC0us=setdiff(c("(Intercept)", forname), names(ZCus));
      ZCus=as.data.frame(ZCus);
      colnames(ZCus)="A";
      ZCzero.us=t(rep(0, length(ZC0us)));
      colnames(ZCzero.us)=ZC0us;
      ZCzero.us=data.frame(t(ZCzero.us));
      colnames(ZCzero.us)="A";
      Est.LS.us=rbind(ZCzero.us, ZCus);
      CEst.LS.us=cbind(Est.LS.us, rownames(Est.LS.us));
      Fest.ZS.us=CEst.LS.us[match(c("(Intercept)", forname),
        CEst.LS.us$‘rownames(Est.LS.us)‘), ][, 1];
      names(Fest.ZS.us)=c("(Intercept)", forname);
      return(list(Fest.ZS.us=Fest.ZS.us, war=war))})
    stopCluster(cl);
    p=length(variables)+1;
    beta.Step.zero=t(matrix(unlist(lapply(repl1, "[[", "Fest.ZS.us")),
        nrow=p));
    war=as.numeric(t(matrix(unlist(lapply(repl1, "[[", "war")),
        nrow=1)));
    toc()
    tic("Model␣averaging")
    message("The␣proportion␣of␣misconvergence␣is␣",
        round(mean(war),2),"%");
    b=data.frame(cbind(beta.Step.zero, war));
    beta.Step.zero=b[which(b$war==0), ];
    beta.Step.zero=beta.Step.zero[, 1:(p)];
    colnames(beta.Step.zero)=forname;
    beta.Step.zero[is.na(beta.Step.zero)]=0;
    ind.Austin=apply(ifelse(beta.Step.zero==0, 0, 1), 2, mean);
    Prob=round(ind.Austin * 100, 2);
    inc.frac=names(ind.Austin)[which(ind.Austin >= 0.5)];
    if (length(inc.frac)==length(forname)){
      minInc=names(ind.Austin)[which(ind.Austin==min(ind.Austin))];
      inc.frac=setdiff(forname, minInc)};

   #Model averaging
  if (par==F){
  options(na.action="na.fail");
  rank="AIC";
  MAdata=data.frame(dataN);
  if(length(inc.frac)<=1){
    m.inc=glm(formula, family=binomial, data=MAdata);
    MA.inc=dredge(global.model=m.inc, rank=rank);
    allModelsList=lapply(attributes(MA.inc)$model.calls, formula);
    atr.inc=lapply(allModelsList, function(x, data) glm(x,
        data=data.frame(data), family="binomial"), data=MAdata) ;
  }else{
    m.inc=glm(formula, family=binomial, data=MAdata);
    MA.inc=dredge(global.model=m.inc, rank=rank, fixed=inc.frac[-1]);
    allModelsList=lapply(attributes(MA.inc)$model.calls, formula);
```

```
   atr.inc=lapply(allModelsList, function(x, data) glm(x,
       data=data.frame(data), family="binomial"), data=MAdata)}
}else{
   MAdata=data.frame(dataN);
   m.inc=glm(formula, family=binomial, data=MAdata);
   cl=makeCluster(detectCores());
   clusterExport(cl,c("dataN","m.inc","p"),envir=environment());
   clusterExport(cl,c("n","inc.frac","MAdata"),envir=environment());
   clusterEvalQ(cl, library(MuMIn));
   rank="AIC";
   options(na.action="na.fail");
   MAdata=data.frame(dataN);
   if(length(inc.frac)<=1){
     MA.inc=pdredge(global.model=m.inc, rank=
       rank, cluster=cl);
     allModelsList=lapply(attributes(MA.inc)$model.calls, formula);
     atr.inc=parLapply(cl=cl, allModelsList, function(x, data) glm(x,
        data=data.frame(MAdata), family="binomial"));
   }else{
     MA.inc=pdredge(global.model=m.inc, rank=rank, fixed=
       inc.frac[-1], cluster=cl);
     allModelsList=lapply(attributes(MA.inc)$model.calls, formula);
     atr.inc=parLapply(cl=cl, allModelsList, function(x, data) glm(x,
       data=data.frame(MAdata), family="binomial"))}
   stopCluster(cl)};
   message("The␣model␣averaging␣is␣done␣over␣",length(allModelsList),"␣
      models");
   #####################
   #MATA after INCLUSION#
   #####################
   start.Inc=startpoints(attrib=atr.inc, modnames=forname, forname=
      forname);
   if(cim=="Wald"){
     confi=waldF(fitted=atr.inc, mma=m.inc, alpha=alpha,
       startL=start.Inc[, 1], startU=start.Inc[, 2]);
   }else if(cim=="Score") {
     confi=scoreF(fitted=atr.inc, mma=m.inc, alpha=alpha,
       startL=start.Inc[, 1], startU=start.Inc[, 2]);
   }else if(cim=="PL"){
     confi=profLF(fitted=atr.inc, mma=m.inc, alpha=alpha,
       startL=start.Inc[, 1], startU=start.Inc[, 2]);
   }else if(cim=="Wald-S"){
     confi=waldFcor(fitted=atr.inc, mma=m.inc, alpha=alpha,
       startL=start.Inc[, 1], startU=start.Inc[, 2], metci="S");
   }else if(cim=="Wald-PL"){
     confi=waldFcor(fitted=atr.inc, mma=m.inc, alpha=alpha,
       startL=start.Inc[, 1], startU=start.Inc[, 2], metci="PL");}
   ci.lower=confi[, 1];
   ci.upper=confi[, 2];
   Est=start.Inc[, 3];
   Est=Est[forname];
   toc()
   toc()
   ##########################
   #Bayesian Model Averaging#
   ##########################
 }else if(selection=="Bayes"){
   #BMA
   tic("Total")
   b=bic.glm(formula, data=data.frame(dataN), OR=20,
    glm.family="binomial");
Prob=c(100, b$probne0);
L.BMA=b$n.models;
message("The␣model␣averaging␣is␣done␣over␣",L.BMA,"␣models");
test.MRX=MRX=b$mle[1:L.BMA, ];
test.MRX.se=b$se[1:L.BMA, ];
formular=apply(data.frame(test.MRX)[-1] != 0, 1, function(x)
  as.character(paste(c(paste(c(out, "~1"), collapse=""),
    variables[x]), collapse="+")));
if (any((apply(data.frame(test.MRX) != 0, 1, sum)==0)==1)){
  zeroPred=apply(data.frame(test.MRX) !=0 , 1, sum)==0;
  formular[which(zeroPred==1)]=as.character(paste(
    c(out, "~1"), collapse=""))};
modelBMA=list();
if(L.BMA > 1){
  for (i in 1:L.BMA){
    modelBMA[[i]]=formular[i]};
}else{
  nn=names(test.MRX)[which(test.MRX != 0)];
  modelBMA=ifelse(any(nn=="(Intercept)")==T, paste(ifelse(
    length(nn)==1, paste(c(out, "~1"), collapse=""), paste(
    c(out, "~1+"),  collapse="")), paste(nn[-1], collapse="+"),
    sep=""), paste(paste(c(out, "~-1+"),
    collapse=""), paste(nn, collapse="+"), sep=""))};
allModelsResults=lapply(modelBMA, function(x, data) glm(x,
    data=data.frame(data), family="binomial"), data=dataN);
nn=names(b$postmean)[which(b$postmean != 0)];
bmodel=ifelse(any(nn=="(Intercept)")==T, paste(ifelse(length(nn)==1,
    paste(c(out, "~1"), collapse=""), paste(c(out, "~1+"),
    collapse="")), paste(nn[-1], collapse="+"), sep=""),
    paste(paste(c(out, "~-1+"), collapse=""),
    paste(nn, collapse="+"), sep=""));
bmam2=glm(bmodel, family=binomial, data=as.data.frame(dataN));
###########
start.BMA=startpoints(attrib=allModelsResults, modnames=nn, forname=
   forname);
if(cim=="Wald"){
  confi=waldF(fitted=allModelsResults, mma=bmam2, alpha=alpha,
    startL=start.BMA[, 1][nn], startU=start.BMA[, 2][nn]);
}else if(cim=="Score") {
  confi=scoreF(fitted=allModelsResults, mma=bmam2, alpha=alpha,
    startL=start.BMA[, 1][nn], startU=start.BMA[, 2][nn]);
}else if(cim=="PL"){
  confi=profLF(fitted=allModelsResults, mma=bmam2, alpha=alpha,
    startL=start.BMA[, 1][nn], startU=start.BMA[, 2][nn]);
}else if(cim=="Wald-S"){
  confi=waldFcor(fitted=allModelsResults, mma=bmam2, alpha=alpha,
    startL=start.BMA[, 1][nn], startU=start.BMA[, 2][nn],
    metci="S");
}else if(cim=="Wald-PL"){
  confi=waldFcor(fitted=allModelsResults, mma=bmam2, alpha=alpha,
    startL=start.BMA[, 1][nn], startU=start.BMA[, 2][nn],
    metci="PL")};
toc()
ci.lower=confi[, 1];
ci.lower=ci.lower[forname];
names(ci.lower)=forname;
ci.lower[is.na(ci.lower)]=NA;
ci.upper=confi[, 2];
ci.upper=ci.upper[forname];
```

```
    names(ci.upper)=forname;
    Est=start.BMA[, 3];
    Est[is.na(ci.upper)]=NA;
    ci.upper[is.na(ci.upper)]=NA};
    results=cbind(Estimates=Est, CI.lower=ci.lower,
    CI.upper=ci.upper, Prop=Prob);
    colnames(results)=c("Estimates", paste(round(100*alpha, 2), "%",
    sep=""),  paste(round(100*(1-alpha), 2), "%", sep=""),"Prop")
    return(results)};
```

A.3 Support functions

```
startpoints=function(attrib, modnames, forname){
  coefB=lapply(attrib, coef);
  aicB=lapply(attrib, function(x) x$aic);
  www=list();
  mn1=length(attrib);
  pm=length(modnames);
  for (g in 1:mn1){
    www[[g]]=rep(aicB[[g]], pm);
    names(www[[g]])=names(coefB[[g]])};
  AICe=Mrank1=SMrank=w=wu=ww=wuu=indi=list();
  for (g in 1:pm){
    AICe[[g]]=lapply(www, function(x) unlist(x[which(names(x)==
        modnames[g])]))};
  mACIe=lapply(AICe, function(x) rep(min(unlist(x)), mn1));
  for (g in 1:pm){
    Mrank1[[g]]=Map('-', AICe[[g]], mACIe[[g]])};
    SMrank=lapply(Mrank1, function(x) sum(exp(-0.5*unlist(x))));
  for (g in 1:pm){
    wu[[g]]=lapply(Mrank1[[g]], function(x){
      if(length(x)==0) {x=NA}else{x=x}})};
  for (g in 1:pm){wuu[[g]]=lapply(wu[[g]],
        function(x) exp(-0.5*((unlist(x)))))};
  for (g in 1:pm){ww[[g]]=lapply(wuu[[g]],
        function(x) unlist(x)/SMrank[[g]])};
  w=ww;;
  m=matrix(unlist(w), ncol=pm);
  m[is.na(m)]=0;
  colnames(m)=modnames;
  summ=lapply(attrib, summary);
  std.err=lapply(summ, function(x)
  as.numeric(t(x$coefficients[, "Std.␣Error", drop=FALSE])));
  MRX=as.data.frame(matrix(c(rep(0, (pm)*mn1)), ncol=pm));
  colnames(MRX)=c(modnames);
  MRX.se=MRX;
  for (i in 1:(mn1)){
    coef=coefB[[i]];
    l=length(coef);
    se1=std.err[[i]];
    nam=names(se1)=names(coef);
    coef=coef[modnames];
    names(coef)=modnames;
    coef[is.na(coef)]=0;
    coefo=coef[which(coef!=0)];
    if (length(coefo)==0){
      coefo=rep(0, l)
      names(coefo)=nam};
    MRX[i, ]=coef;
    se1=se1[modnames];
    names(se1)=modnames;
    se1[is.na(se1)]=0;
    MRX.se[i, ]=se1};
  try.lower=as.matrix(MRX/MRX.se-qnorm(0.975, 0, 1));
  try.lower[which(!is.finite(try.lower))]=0;
  try.lower=m*try.lower;
  try.upper=as.matrix(MRX/MRX.se-qnorm(0.025, 0, 1));
  try.upper[which(!is.finite(try.upper))]=0;
  try.upper=m*try.upper;
  W.se=as.matrix(m/MRX.se);
  W.se[which(!is.finite(W.se))]=0;
  MATAF.lower=apply(try.lower, 2, sum)/apply(W.se, 2, sum);
  MATAF.lower[which(!is.finite(MATAF.lower))]=0;
  MATAF.lower=MATAF.lower[forname];
  names(MATAF.lower)=forname;
  MATAF.lower[is.na(MATAF.lower)]=0;
  MATAF.upper=apply(try.upper, 2, sum)/apply(W.se, 2, sum);
  MATAF.upper[which(!is.finite(MATAF.upper))]=0;
  MATAF.upper=MATAF.upper[forname];
  names(MATAF.upper)=forname;
  MATAF.upper[is.na(MATAF.upper)]=0;
  MATAF.Est=apply(m*MRX, 2, sum);
  MATAF.Est=MATAF.Est[forname];
  names(MATAF.Est)=forname;
  MATAF.Est[is.na(MATAF.Est)]=0;
  return(cbind(MATAF.lower, MATAF.upper, MATAF.Est))};
```

```
ScoreRoot=function(t, parms){
  var=parms$var;
  fitted=parms$fitted;
  coefB=coef(fitted);
  nonA=!is.na(coefB);
  Pnames=names(coefB);
  pv0=t(as.matrix(coefB));
  mf=model.frame(fitted);
  Y=model.response(mf);
  n=NROW(Y);
  O=model.offset(mf);
  if (!length(O)) O=rep(0, n)   ;
  W=model.weights(mf);
  if (length(W)==0L) W=rep(1, n)   ;
  X=model.matrix(fitted);
  fam=family(fitted);
  B=coefB[var];
  LP= X[, nonA, drop=FALSE] %*% coefB[nonA]+O;
  a=nonA;
  a[which(names(a)==var)]=FALSE;
  Xi=X[, a, drop=FALSE];
  pi=Pnames[which(Pnames==var)];
  bi=t;
  o=O+X[, var] * bi;
  fm=glm.fit(x=Xi, y=Y, weights=W, etastart=LP,
                offset=o, family=fam, control=fitted$control);
  LP=Xi %*% fm$coefficients+o;
  ri=pv0;
  ri[, names(coef(fm))]=coef(fm);
  ri[, pi]=bi;
  d=length(ri);
  u=vector();
  IF=matrix(rep(0, d^2), d, d);
  r=as.vector(ri);
  for (k in 1:d){
    u[k]=sum((Y-(exp(X%*%r)/(1+exp(X%*%r))))*X[, k]);

    for (l in 1:d){
      In=(sum(X[, l]*X[, k]*exp(X%*%r)/(1+exp(X%*%r))^2));
      IF[k, l]=In}}
  S=u%*%solve(IF)%*%u;
  S=max(S, 0);
  z=S-qchisq(0.95, df=1);
  return(z)};
```

```
scoreF=function (fitted, mma, alpha=0.025, startL, startU){
  coefB=lapply(fitted, coef);
  nonA=lapply(coefB, function(x) !is.na(x));
  aicB=lapply(fitted, function(x) x$aic);
  www=list();
  for (g in 1:length(coefB)){
    len=length(coefB[[g]]);
    re=rep(aicB[[g]], len);
    www[[g]]=re;
    names(www[[g]])=names(coefB[[g]])};
  Pnames=lapply(coefB, names);
  pv0=lapply(coefB, function(x) t(as.matrix(x)));
  p=lapply(Pnames, length);
  which=lapply(p, function(x) 1:x);
  summ=lapply(fitted, summary);
  mf=lapply(fitted, model.frame);
  mfy=model.frame(mma);
  Fnames=names(B0 <- coef(mma));
  Fwhich=1:length(Fnames);
  Y=model.response(mfy);
  n=NROW(Y);
  O=lapply(mf, model.offset);
  O=lapply(O, function(x){
    if (!length(x)) x=rep(0, n)});
  W=lapply(mf, model.weights);
  W=lapply(W, function(x){
    if (length(x)==0L) x=rep(1, n)});
  X=lapply(fitted, model.matrix);
  fam=lapply(fitted, family);
  scor=vector("list", length=length(Fwhich));
  names(scor)=Fnames[Fwhich];
  for (i in Fwhich) {
    a=nonA;
    var=Fnames[i];
    an=length(a);
    AICe=lapply(www, function(x) x[which(names(x)==var)]);
```

```
pe=lapply(coefB, function(x) x[which(names(x)==var)]);
tl=startL[i];
tu=startU[i];
parametersu=list(an=an, var=var, AICe=AICe, nonA=nonA, coefB=coefB,
   O=O, a=a, X=X, Pnames=Pnames, W=W, fitted=fitted, fam=fam,
    pv0=pv0, alpha=alpha, errb="u", pointest=pe);
parametersl=list(an=an, var=var, AICe=AICe, nonA=nonA, coefB=coefB,
   O=O, a=a, X=X, Pnames=Pnames, W=W, fitted=fitted, fam=fam,
    pv0=pv0, alpha=alpha, errb="l", pointest=pe);
topot=function(t, parms){
  an=parms$an;
  var=parms$var;
  nonA=parms$nonA;
  errb=parms$errb;
  coefB=parms$coefB;
  O=parms$O;
  a=parms$a;
  X=parms$X;
  Pnames=parms$Pnames;
  W=parms$W;
  fitted=parms$fitted;
  fam=parms$fam;
  pv0=parms$pv0;
  alpha=parms$alpha;
  pointest=parms$pointest;
  Xi=pi=bi=d=u=IF=LP=mark=B=S=Sm=r=ri=o=fm=indi=list();
  maicB=min(unlist(AICe));
  Mrank1=Map(’-’, AICe, maicB);
  SMrank=sum(exp(-0.5*unlist(Mrank1)));
  w=lapply(Mrank1, function(x) exp(-0.5*(x))/SMrank);
  for (g in 1:an){
    if(length(pointest[[g]])==0) w[[g]]=0};
  for (j in 1:an){
    if (is.element(var, names(a[[j]]))==T){
      B[[j]]=coefB[[j]][var];
      LP[[j]]=X[[j]][, nonA[[j]], drop=FALSE] %*%
            coefB[[j]][nonA[[j]]]+O[[j]];
      a[[j]][which(names(a[[j]])==var)]=FALSE;
      Xi[[j]]=X[[j]][, a[[j]], drop=FALSE];
      pi[[j]]=Pnames[[j]][which(Pnames[[j]]==var)];
      bi[[j]]=t;
      o[[j]]=O[[j]]+X[[j]][, var] * bi[[j]];
      fm[[j]]=glm.fit(x=Xi[[j]], y=Y, weights=W[[j]],
    etastart=LP[[j]], offset=o[[j]], family=fam[[j]],
    control=fitted[[j]]$control);
      LP[[j]]=Xi[[j]] %*% fm[[j]]$coefficients+o[[j]];
      ri[[j]]=pv0[[j]];
      ri[[j]][, names(coef(fm[[j]]))]=coef(fm[[j]]);
      ri[[j]][, pi[[j]]]=bi[[j]];
      d[[j]]=length(ri[[j]]);
      u[[j]]=vector();
      IF[[j]]=matrix(rep(0, d[[j]]^2), d[[j]], d[[j]]);
      r[[j]]=as.vector(ri[[j]]);
      for (k in 1:d[[j]]){
            u[[j]][k]=sum((Y-(exp(X[[j]]%*%r[[j]])/
            (1+exp(X[[j]]%*%r[[j]]))))*X[[j]][, k]);
       for (l in 1:d[[j]]){
          In=(sum(X[[j]][, l]*X[[j]][, k]*exp(X[[j]]%*%r[[j]])/
            (1+exp(X[[j]]%*%r[[j]]))^2));
          IF[[j]][k, l]=In}}
          S[[j]]=u[[j]]%*%solve(IF[[j]])%*%u[[j]];
          S[[j]]=max(S[[j]], 0);
          Sm[[j]]=sign(B[[j]]-bi[[j]])*sqrt(S[[j]]);
          mark[[j]]=1;
        }else{
          S[[j]]=0; Sm[[j]]=0; mark[[j]]=0};
        Sm[[j]]=pnorm(Sm[[j]])};
      if (errb=="u"){
        z=sum(unlist(Sm)*unlist(w)*unlist(mark))-alpha;
      }else{
        z=sum((1-unlist(Sm))*unlist(w)*unlist(mark))-alpha};
      return(z)};
    options(warn=2);
    test=try(multiroot(topot, start=tl, maxiter=1000, useFortran=T,
         parms=parametersl));
    options(warn=1);
    war=inherits(test, "try-error");
    if (war==T){
      l=uniroot(topot, lower=tl, upper=tu, extendInt="yes",
        maxiter=1000, tol=1e-5, parms=parametersl)$root;
    }else{
      l=multiroot(topot, start=tl, maxiter=1000, useFortran=T, ctol=
        1e-5, parms=parametersl)$root};
    options(warn=2);
    test=try(multiroot(topot, start=tu, maxiter=1000, useFortran=T,
        parms=parametersu));
    options(warn=1);
    war=inherits(test, "try-error");
    if (war==T){
      u=uniroot(topot, lower=tl, upper=tu, maxiter=1000,
        extendInt="yes", tol=1e-5, parms=parametersu)$root;
    }else{
      u=multiroot(topot, start=tu, maxiter=1000, useFortran=T,
        ctol=1e-5, parms=parametersu)$root};
    med=u;
    if (l>u){u=l; l=med};
    scor[[i]]=data.frame(rbind(l, u));
    colnames(scor[[i]])=NULL};
  CI=matrix(rep(0, 2*length(Fwhich)), ncol=2);
  for(i in Fwhich){
    CI[i, ]=t(scor[[i]])};
  rownames(CI)=Fnames;
  colnames(CI)=c("lower", "upper");
  return(CI)};
```

```
profLF=function(fitted, mma, alpha=0.025, startL, startU){
  coefB=lapply(fitted, coef);
  nonA=lapply(coefB, function(x) !is.na(x));
  aicB=lapply(fitted, function(x) x$aic);
  www=list();
  for (g in 1:length(coefB)){
    len=length(coefB[[g]]);
    re=rep(aicB[[g]], len);
    www[[g]]=re;
    names(www[[g]])=names(coefB[[g]])};
  Pnames=lapply(coefB, names);
```

```
pv0=lapply(coefB, function(x) t(as.matrix(x)));
p=lapply(Pnames, length);
which=lapply(p, function(x) 1:x);
summ=lapply(fitted, summary);
std.err=lapply(summ, function(x) as.numeric(t(x$coefficients[,
      "Std.␣Error", drop=FALSE])));
for (i in 1:length(nonA)){names(std.err[[i]])=Pnames[[i]]};
mf=lapply(fitted, model.frame);
mfy=model.frame(mma);
Fnames=names(B0 <- coef(mma));
Fwhich=1:length(Fnames);
Y=model.response(mfy);
n=NROW(Y);
O=lapply(mf, model.offset);
O=lapply(O, function(x){
  if (!length(x)) x=rep(0, n)});
W=lapply(mf, model.weights);
W=lapply(W, function(x){
  if (length(x)==0L) x=rep(1, n)});
X=lapply(fitted, model.matrix);
OriginalDeviance=lapply(fitted, deviance);
DispersionParameter=lapply(summ, function(x) x$dispersion);
fam=lapply(fitted, family);
scor=vector("list", length=length(Fwhich));
names(scor)=Fnames[Fwhich];
for (i in Fwhich) {
  a=nonA;
  var=Fnames[i];
  an=length(a);
  pe=lapply(coefB, function(x) x[which(names(x)==var)]);
  AICe=lapply(www, function(x) x[which(names(x)==var)]);
  spe=lapply(std.err, function(x) x[which(names(x)==var)]);
  tl=startL[i];
  tu=startU[i]  ;
  parametersu=list(an=an, var=var, AICe=AICe, nonA=nonA, coefB=coefB,
      O=O, a=a, X=X, Pnames=Pnames, W=W, fitted=fitted, fam=fam,
      pv0=pv0, alpha=alpha, errb="u", pointest=pe);
  parametersl=list(an=an, var=var, AICe=AICe, nonA=nonA, coefB=coefB,
      O=O, a=a, X=X, Pnames=Pnames, W=W, fitted=fitted, fam=fam,
      pv0=pv0, alpha=alpha, errb="l", pointest=pe);
  topot=function(t, parms){
    an=parms$an;
    var=parms$var;
    nonA=parms$nonA;
    coefB=parms$coefB;
    O=parms$O;
    a=parms$a;
    X=parms$X;
    AICe=parms$AICe;
    errb=parms$errb;
    Pnames=parms$Pnames;
    W=parms$W;
    fitted=parms$fitted;
    fam=parms$fam;
    pv0=parms$pv0;
    alpha=parms$alpha;
    pointest=parms$pointest;
    Xi=pi=bi=d=u=IF=LP=LPm=mark=B=LPj=S=Sm=r=ri=o=fm=indi=list();
    maicB=min(unlist(AICe));
    Mrank1=Map('-', AICe, maicB);
    SMrank=sum(exp(-0.5*unlist(Mrank1)));
    w=lapply(Mrank1, function(x) exp(-0.5*(x))/SMrank);
    for (g in 1:an){
      if(length(pointest[[g]])==0) w[[g]]=0};
    for (j in 1:an){
      if (is.element(var, names(a[[j]]))==T){
        LPm[[j]]=logLik(fitted[[j]]);
        B[[j]]=coefB[[j]][var];
        LP[[j]]=X[[j]][, nonA[[j]], drop=FALSE]%*%
              coefB[[j]][nonA[[j]]]+O[[j]];
        a[[j]][which(names(a[[j]])==var)]=FALSE;
        Xi[[j]]=X[[j]][, a[[j]], drop=FALSE];
        pi[[j]]=Pnames[[j]][which(Pnames[[j]]==var)];
        bi[[j]]=t;
        o[[j]]=O[[j]]+X[[j]][, var] * bi[[j]];
        fm[[j]]=glm.fit(x=Xi[[j]], y=Y, weights=W[[j]], etastart=
              LP[[j]], offset=o[[j]], family=fam[[j]], control=
              fitted[[j]]$control);
        S[[j]]=mark[[j]]=(fm[[j]]$deviance- OriginalDeviance[[j]])/
           DispersionParameter[[j]];
        if(mark[[j]]<0) S[[j]]=0;
        Sm[[j]]=sign(B[[j]]-bi[[j]])*sqrt(S[[j]]);
        mark[[j]]=1;
      }else{
        S[[j]]=0; Sm[[j]]=0; mark[[j]]=0;};
      Sm[[j]]=pnorm(Sm[[j]])};
    if (errb=="u"){z=sum(unlist(Sm)*unlist(w)*unlist(mark))-alpha;
    }else{z=sum((1-unlist(Sm))*unlist(w)*unlist(mark))-alpha};
    return(z)};
  options(warn=2);
  test=try(multiroot(topot, start=tl, maxiter=1000, useFortran=T,
     parms=parametersl));
  options(warn=1);
  war=inherits(test, "try-error");
  if (war==T){
      l=uniroot(topot, lower=tl, upper=tu, extendInt="yes",
              maxiter=1000, tol=1e-5, parms=parametersl)$root;
  }else{
      l=multiroot(topot, start=tl, maxiter=1000, useFortran=T, ctol=
              1e-5, parms=parametersl)$root};
  options(warn=2);
  test=try(multiroot(topot, start=tu, maxiter=1000, useFortran=T,
     parms=parametersu));
  options(warn=1);
  war=inherits(test, "try-error");
  if (war==T){
      u=uniroot(topot, lower=tl, upper=tu, maxiter=1000, extendInt=
              "yes", tol=1e-5, parms=parametersu)$root;
  }else{
      u=multiroot(topot, start=tu, maxiter=1000, useFortran=T, ctol=
              1e-5, parms=parametersu)$root};
  med=u; if (l>u){u=l; l=med};
  scor[[i]]=data.frame(rbind(l, u));
  colnames(scor[[i]])=NULL;};
CI=matrix(rep(0, 2*length(Fwhich)), ncol=2);
for(i in Fwhich){
  CI[i, ]=t(scor[[i]]);};
rownames(CI)=Fnames;
colnames(CI)=c("lower", "upper");
return(CI)};
```

```
waldF=function (fitted, mma, alpha=0.025, startL, startU){
  coefB=lapply(fitted, coef);
  aicB=lapply(fitted, function(x) x$aic);
  www=list();
  for (g in 1:length(coefB)){
    len=length(coefB[[g]]);
    re=rep(aicB[[g]], len);
    www[[g]]=re;
    names(www[[g]])=names(coefB[[g]])};
  nonA=lapply(coefB, function(x) !is.na(x));
  Pnames=lapply(coefB, names);
  pv0=lapply(coefB, function(x) t(as.matrix(x)));
  p=lapply(Pnames, length);
  which=lapply(p, function(x) 1:x);
  summ=lapply(fitted, summary);
  std.err=lapply(summ, function(x) as.numeric(t(x$coefficients[,
        "Std.␣Error", drop=FALSE])));
  for (i in 1:length(nonA)){names(std.err[[i]])=Pnames[[i]]};
  mf=lapply(fitted, model.frame);
  mfy=model.frame(mma);
  Fnames=names(B0 <- coef(mma));
  Fwhich=1:length(Fnames);
  Y=model.response(mfy);
  n=NROW(Y);
  O=lapply(mf, model.offset);
  O=lapply(O, function(x){
    if (!length(x)) x=rep(0, n)});
  W=lapply(mf, model.weights);
  W=lapply(W, function(x){
    if (length(x)==0L) x=rep(1, n)});
  X=lapply(fitted, model.matrix);
  fam=lapply(fitted, family);
  scor=vector("list", length=length(Fwhich));
  names(scor)=Fnames[Fwhich];
  for (i in Fwhich) {
    a=nonA;
    var=Fnames[i];
    an=length(a);
    AICe=lapply(www, function(x) x[which(names(x)==var)]);
    pe=lapply(coefB, function(x) x[which(names(x)==var)]);
    spe=lapply(std.err, function(x) x[which(names(x)==var)]);
    tl=startL[i];
    tu=startU[i];
    parametersu=list(an=an, AICe=AICe, var=var, nonA=nonA, n=n, coefB=
        coefB, O=O, a=a, X=X, p=p, Pnames=Pnames, W=W, fitted=fitted,
        fam=fam, pv0=pv0, alpha=alpha, errb="u", pointest=pe, spe=spe);
    parametersl=list(an=an, AICe=AICe, var=var, nonA=nonA, n=n, coefB=
        coefB, O=O, a=a, X=X, p=p, Pnames=Pnames, W=W, fitted=fitted,
        fam=fam, pv0=pv0, alpha=alpha, errb="l", pointest=pe, spe=spe);
    topot=function(t, parms){
      an=parms$an;
      var=parms$var;
      nonA=parms$nonA;
      coefB=parms$coefB;
      O=parms$O;
      a=parms$a;
      X=parms$X;
      n=parms$n;
      errb=parms$errb;
    Pnames=parms$Pnames;
    AICe=parms$AICe;
    W=parms$W;
    fitted=parms$fitted;
    fam=parms$fam;
    pv0=parms$pv0;
    alpha=parms$alpha;
    pointest=parms$pointest;
    spe=parms$spe;
    p=parms$p;
    Xi=pi=bi=d=u=IF=LP=LPm=mark=B=LPj=S=Sm=SE=r=ri=o=fm=indi=list();
    maicB=min(unlist(AICe));
    Mrank1=Map('-', AICe, maicB);
    SMrank=sum(exp(-0.5*unlist(Mrank1)));
    w=lapply(Mrank1, function(x) exp(-0.5*(x))/SMrank);
    for (g in 1:an){
      if(length(pointest[[g]])==0) w[[g]]=0};
    for (j in 1:an){
      if (is.element(var, names(a[[j]]))==T){
        B[[j]]=coefB[[j]][var];
        SE[[j]]=spe[[j]];
        bi[[j]]=t;
        S[[j]]=Sm[[j]]=(B[[j]]-bi[[j]])/SE[[j]]  ;
      }else{S[[j]]=0; Sm[[j]]=0};
      Sm[[j]]=pt(Sm[[j]], df=n-length(coefB[[j]]))};
    if (errb=="u"){z=sum(unlist(Sm)*unlist(w))-alpha;
    }else{z=sum((1-unlist(Sm))*unlist(w))-alpha};
    return(z)};
  options(warn=2);
  test=try(multiroot(topot, start=tl, maxiter=1000, useFortran=T,
     parms=parametersl));
  options(warn=1);
  war=inherits(test, "try-error");
  if (war==T){
      l=uniroot(topot, lower=tl, upper=tu, extendInt="yes",
              maxiter=1000, tol=1e-5, parms=parametersl)$root;
  }else{
      l=multiroot(topot, start=tl, maxiter=1000, useFortran=T, ctol=
              1e-5, parms=parametersl)$root};
  options(warn=2);
  test=try(multiroot(topot, start=tu, maxiter=1000, useFortran=T,
     parms=parametersu));
  options(warn=1);
  war=inherits(test, "try-error");
  if (war==T){
      u=uniroot(topot, lower=tl, upper=tu, maxiter=1000, extendInt=
              "yes", tol=1e-5, parms=parametersu)$root;
  }else{
      u=multiroot(topot, start=tu, maxiter=1000, useFortran=T, ctol=
              1e-5, parms=parametersu)$root};
  med=u; if (l>u){u=l; l=med};
  scor[[i]]=data.frame(rbind(l, u));
  colnames(scor[[i]])=NULL};
CI=matrix(rep(0, 2*length(Fwhich)), ncol=2);
for(i in Fwhich){CI[i, ]=t(scor[[i]])};
rownames(CI)=Fnames;
colnames(CI)=c("lower", "upper");
return(CI)};
```

```
waldFcor=function (fitted, mma, alpha=0.025, startL, startU, metci="S"){
  if (metci=="S"){
    varN=length(coef(mma));
    Fnames=names(coef(mma));
    modN=length(fitted);
    Pnames=lapply(lapply(fitted, coef), names);
    CiL=CiU=coefB=list();
    for(j in 1:modN){
      if(length(Pnames[[j]])==0){
        CiL[[j]]=NULL;
        CiU[[j]]=NULL;
        coefB[[j]]=NULL;
      }else{
        al=au=rep(0, length(Pnames[[j]]));
        for(i in 1:length(Pnames[[j]])){
          paraml=list(fitted=fitted[[j]], errb="l", var=Pnames[[j]][i]);
          paramu=list(fitted=fitted[[j]], errb="u", var=Pnames[[j]][i]);
          tl=startL[Pnames[[j]]][i];
          tu=startU[Pnames[[j]]][i];
          al[i]=multiroot(ScoreRoot, start=tl, maxiter=1000, useFortran=
             T, parms=paraml)$root;
          au[i]=multiroot(ScoreRoot, start=tu, maxiter=1000, useFortran=
             T, parms=paramu)$root};
        CiL[[j]]=al; CiU[[j]]=au;
        coefB[[j]]=coef(fitted[[j]])}}  ;
    coefB=lapply(fitted, coef);
    sel=lapply(mapply('-', CiL, coefB, SIMPLIFY=T), function(x) abs(x)/
       qnorm(0.975));
    seu=lapply(mapply('-', CiU, coefB, SIMPLIFY=T), function(x) abs(x)/
       qnorm(0.975));
  }else{
    coefB=lapply(fitted, coef);
    ci=lapply(fitted, confint);
    se=lapply(mapply('-', ci, coefB, SIMPLIFY=T), function(x) abs(x)/
       qnorm(0.975));
    for (i in 1:length(coefB)){
      if(length(coefB[[i]])==1){
        se[[i]]=t(se[[i]]);
        rownames(se[[i]])=names(coefB[[i]])}};
    sel=lapply(se, function(x) x[, 1]);
    seu=lapply(se, function(x) x[, 2]);
    for (i in 1:length(coefB)){
      if(length(coefB[[i]])==1){
        names(seu[[i]])=names(coefB[[i]]);
        names(sel[[i]])=names(coefB[[i]])}}};
  aicB=lapply(fitted, function(x) x$aic);
  www=list();
  for (g in 1:length(coefB)){
    len=length(coefB[[g]]);
    re=rep(aicB[[g]], len);
    www[[g]]=re;
    names(www[[g]])=names(coefB[[g]])};
  nonA=lapply(coefB, function(x) !is.na(x));
  Pnames=lapply(coefB, names);
  pv0=lapply(coefB, function(x) t(as.matrix(x)));
  p=lapply(Pnames, length);
  which=lapply(p, function(x) 1:x);
  summ=lapply(fitted, summary);
```

```
std.err=lapply(summ, function(x) as.numeric(t(x$coefficients[,
      "Std.␣Error", drop=FALSE])));
for (i in 1:length(nonA)){names(std.err[[i]])=Pnames[[i]]};
mf=lapply(fitted, model.frame);
mfy=model.frame(mma);
Fnames=names(B0 <- coef(mma));
Fwhich=1:length(Fnames);
Y=model.response(mfy);
n=NROW(Y);
O=lapply(mf, model.offset);
O=lapply(O, function(x){
  if (!length(x)) x=rep(0, n)});
W=lapply(mf, model.weights);
W=lapply(W, function(x){
  if (length(x)==0L) x=rep(1, n)});
X=lapply(fitted, model.matrix);
fam=lapply(fitted, family);
scor=vector("list", length=length(Fwhich));
names(scor)=Fnames[Fwhich];
for (i in Fwhich) {
  a=nonA;
  var=Fnames[i];
  an=length(a);
  AICe=lapply(www, function(x) x[which(names(x)==var)]);
  pe=lapply(coefB, function(x) x[which(names(x)==var)]);
  spel=lapply(sel, function(x) x[which(names(x)==var)]);
  speu=lapply(seu, function(x) x[which(names(x)==var)]);
  tl=startL[i];
  tu=startU[i];
  parametersu=list(an=an, AICe=AICe, var=var, nonA=nonA, n=n, coefB=
      coefB, O=O, a=a, X=X, p=p, Pnames=Pnames, W=W, fitted=fitted,
      fam=fam, pv0=pv0, alpha=alpha, errb="u", pointest=pe, spe=seu);
  parametersl=list(an=an, AICe=AICe, var=var, nonA=nonA, n=n, coefB=
      coefB, O=O, a=a, X=X, p=p, Pnames=Pnames, W=W, fitted=fitted,
      fam=fam, pv0=pv0, alpha=alpha, errb="l", pointest=pe, spe=sel);
  topot=function(t, parms){
    an=parms$an;
    var=parms$var;
    nonA=parms$nonA;
    coefB=parms$coefB;
    O=parms$O;
    a=parms$a;
    X=parms$X;
    n=parms$n;
    errb=parms$errb;
    Pnames=parms$Pnames;
    AICe=parms$AICe;
    W=parms$W;
    fitted=parms$fitted;
    fam=parms$fam;
    pv0=parms$pv0;
    alpha=parms$alpha;
    pointest=parms$pointest;
    spe=parms$spe;
    p=parms$p;
    Xi=pi=bi=d=u=IF=LP=LPm=mark=B=LPj=S=Sm=SE=r=ri=o=fm=indi=list();
    maicB=min(unlist(AICe));
    Mrank1=Map('-', AICe, maicB);
    SMrank=sum(exp(-0.5*unlist(Mrank1)));
    w=lapply(Mrank1, function(x) exp(-0.5*(x))/SMrank);
    for (g in 1:an){
      if(length(pointest[[g]])==0) w[[g]]=0};
    for (j in 1:an){
      if (is.element(var, names(a[[j]]))==T){
        B[[j]]=coefB[[j]][var];
        SE[[j]]=spe[[j]][var];
        bi[[j]]=t;
        S[[j]]=Sm[[j]]=(B[[j]]-bi[[j]])/SE[[j]];
      }else{S[[j]]=0; Sm[[j]]=0};
      Sm[[j]]=pt(Sm[[j]], df=n-length(coefB[[j]]))} ;
    if (errb=="u"){z=sum(unlist(Sm)*unlist(w))-alpha;
    }else{z=sum((1-unlist(Sm))*unlist(w))-alpha};
    return(z)};
  options(warn=2);
  test=try(multiroot(topot, start=tl, maxiter=1000, useFortran=T,
     parms=parametersl));
  options(warn=1);
  war=inherits(test, "try-error");
  if (war==T){
      l=uniroot(topot, lower=tl, upper=tu, extendInt="yes",
              maxiter=1000, tol=1e-5, parms=parametersl)$root;
  }else{
      l=multiroot(topot, start=tl, maxiter=1000, useFortran=T, ctol=
              1e-5, parms=parametersl)$root};
  options(warn=2);
  test=try(multiroot(topot, start=tu, maxiter=1000, useFortran=T,
     parms=parametersu));
  options(warn=1);
  war=inherits(test, "try-error");
  if (war==T){
      u=uniroot(topot, lower=tl, upper=tu, maxiter=1000, extendInt=
              "yes", tol=1e-5, parms=parametersu)$root;
  }else{
      u=multiroot(topot, start=tu, maxiter=1000, useFortran=T, ctol=
              1e-5, parms=parametersu)$root};
  med=u; if (l>u){u=l; l=med};
  scor[[i]]=data.frame(rbind(l, u));
  colnames(scor[[i]])=NULL};
CI=matrix(rep(0, 2*length(Fwhich)), ncol=2);
for(i in Fwhich){CI[i, ]=t(scor[[i]])};
rownames(CI)=Fnames;
colnames(CI)=c("lower", "upper");
return(CI)};
```

CURRICULUM VITAE

Name: Artem Uvarov

Place of Birth: Tashkent, Uzbekistan, 1989

Education

Department of Statistics
Department of Economics
Hebrew University, Jerusalem
2008-2012, BA Economics and Statistics

Department of Statistics
Braun School of Public Health
Hebrew University, Jerusalem
2012-2014, MA Biotatistics

Department of Epidemiology and Biostatistics
University of Western Ontario, London, Ontario
2015-2019, Ph.D. Biotatistics

Experience:

Teaching Assistant
Department of Epidemiology and Biostatistics
University of Western Ontario
London, Ontario, 2017-2018

Research Assistant
Department of Epidemiology and Biostatistics
University of Western Ontario
London, Ontario, 2015-2019

Biostatistician
Centre for Diabetes, Endocrinology and Metabolism
St. Josephs Hospital
London, Ontario, 2018-present