COMPARISON OF VARIABLE SELECTION METHODS

Elizabeth Koehler Rowley

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2019

Approved by:

Annie Green Howard

Chirayath Suchindran

Matthew A. Psioda

Penny Gordon-Larsen

Amy Herring

ABSTRACT

Elizabeth Koehler Rowley: Comparison of Variable Selection Methods
(Under the direction of Annie Green Howard and Chirayath Suchindran)

Use of classic variable selection methods in public health research is quite common. Many criteria, and various strategies for applying them, now exist including forward selection, backward elimination, stepwise selection, best-subset selection and so on, but all suffer from similar drawbacks. Chief among them is a failure to account for the uncertainty contained in the model selection process. Ignoring model uncertainty can cause several serious problems. Variance estimates are generally underestimated, p-values are generally inflated, prediction ability is overestimated, and results are not reproducible in another dataset.

Modern variable selection methods have become increasingly popular, especially in applications of high-dimensional or sparse data. Some of these methods were developed to address the short-comings of classic variable selection methods, such as backward elimination and stepwise selection methods. However, it remains unclear how modern variable selection methods behave in a classical, meaning non-high-dimensional, setting.

A simulation study investigates the estimation, predictive performance and variable selection capabilities of three representative modern variable selection methods: Bayesian model averaging (BMA), stochastic search variable selection (SSVS), and the adaptive lasso. These three methods are considered in the setting of linear regression with a single variable of interest which is always included in the model.

A second simulation study compares BMA to classical variable selection methods, including backward elimination, two-stage method, and change-in-effect method in the setting of logistic regression. Additionally, the data generated in both simulation studies closely mimic a real study and reflect a realistic correlation structure between potential covariates. Sample sizes ranging from 150 to 20000 are investigated. BMA is demonstrated in an example building a predictive model using data from the China Health and Nutrition Survey.

ACKNOWLEDGMENTS

Thank you to my wonderful family, your love and support have carried me to this finish line. Saying yes to you, Neil Rowley, is far and away the smartest thing I have ever done. Your inspiring determination, patience, gentle encouragement, and unfailing love are essential to me in this, and every, endeavor. My sweet Helen, you have made this the most joyful dissertation experience a mother could imagine. My parents, Mary and Tom Koehler, your example always emboldened me to try my very best and have faith that God will take care of the rest. Thank you for endlessly listening to all my rambles, giving excellent advice, and for providing a concrete example of unconditional love. Thank you to the best siblings, Joey, Christine, and Margaret whose loving support is so appreciated.

I would like to thank Dr. Annie Green Howard and Dr. Chirayath Suchindran for their expertise and enthusiasm while completing this project. This dissertation had many twists and turns and I am grateful that you were both eager to encourage me to the finish. I would also like to thank Dr. Matt Psioda. His generous contribution of ideas and time have been invaluable to this work. I have benefited greatly by getting to observe Dr. Amy Herring and Dr. Penny Gordon-Larsen in action. Their example of women passionately using their abilities to improve public health has been inspiring.

I am particularly indebted to my friends who have brightened my entire experience at UNC. Far from being competitors, you were always quick to smile, eager to offer any assistance, and genuine in concern. This dissertation would, quite frankly, not have even been attempted without your friendship. I am very grateful for each of you!

This work was supported by a training grant from NIEHS, T32 ES007018. This research uses data from China Health and Nutrition Survey (CHNS) and is funded by NIH

R01DK104371. I thank the National Institute for Nutrition and Health, China Center for Disease Control and Prevention, Carolina Population Center (P2C HD050924, T32 HD007168), the University of North Carolina at Chapel Hill, the NIH (R01-HD30880, DK056350, R24 HD050924, and R01-HD38700) and the NIH Fogarty International Center (D43 TW009077, D43 TW007709) for financial support for the CHNS data collection and analyses from 1989 to 2015 and future surveys, and the China-Japan Friendship Hospital, Ministry of Health for support for CHNS 2009, Chinese National Human Genome Center at Shanghai since 2009, and Beijing Municipal Center for Disease Prevention and Control since 2011.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Motivation

Statistical models are a fundamental tool for the public health researcher. Successful analysis certainly depends on studying the appropriate group of patients, selecting the correct model family, and accurately interpreting the results. However, selecting the variables to include in the model is crucial not only for answering the scientific question, but variable selection is also critical in understanding the replicability of the conclusions. This dissertation explores the effects of modern and classical variable selection techniques relating to three central goals of statistical models: effect estimation, outcome prediction, and understanding variable relationships.

In the 1970s, several statistics were introduced for the purpose of selecting between competing models including Akaike Information Criterion (AIC)(Akaike 1974), Mallow's C_p (Mallows 1973), and Bayesian Information Criterion (BIC) (Schwarz et al. 1978). Many more criteria, and various strategies for applying them now exist including forward selection, backward elimination, stepwise selection, best-subset selection and so on, but all suffer from similar drawbacks. Chief among them is a failure to account for the uncertainty contained in the model selection process. Considering multiple models and then proceeding with the selected model as if it were known to be the correct model can cause several serious problems. Variance estimates are generally underestimated, p-values are generally inflated, prediction ability is overestimated, and results are not reproducible in another dataset (Harrell 2001, Viallefont et al. 2001, Sun et al. 1996, Hurvich and Tsai 1990).

Model uncertainty can be appropriately represented if estimates from every model considered are somehow accounted for (Buckland et al. 1997). Bayesian model averaging (BMA) provides an opportunity for exploring many possible models while appropriately accounting for the uncertainty surrounding variable selection (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). BMA is only one of several statistical methods developed for appropriately accounting for model uncertainty but has the advantage of providing the best predictive qualities (George 2000). Stochastic search variable selection (SSVS) is a Bayesian variable selection method initially proposed as a clever way to focus on promising subsets of variables (George and McCulloch 1993). Although both methods have been applied in other fields, methods which account for model uncertainty have yet to see much use in public health (Walter and Tiemeier 2009).

Large population studies typically contain many related variables and selecting between these related variables can be quite challenging. Failure to acknowledge model uncertainty can hinder the public health researcher's ability to coalesce seemingly disparate results from multiple studies of the same scientific question into knowledge. According to a study by Walter and Tiemeier (2009) of the articles presented in *American Journal of Epidemiology*, *Epidemiology*, *European Journal of Epidemiology*, and *International Journal of Epidemiology* in 2008, 27.7% of authors used only prior knowledge to build their model, 19.7% used a form of stepwise methods, 14.7% used change in effect methods, and 3% used other methods such as propensity scores or principal components. None of the 300 articles reviewed used ridge regression or shrinkage methods, while a 35% did not disclose their model building strategy, rendering their results uninterpretable. This report suggests methods which directly account for model uncertainty, such as BMA, are either unknown or not easily accessed by those publishing in public health journals.

This dissertation presents a set of simulations to compare BMA with other variable selection methods both modern (SSVS and adaptive lasso) and classic (backward elimination, two-stage method, and change-in-effect). Variable selection methods are compared on their abilities to estimate a coefficient, to predict an outcome, and to select variables belonging in the true model. These methods are applied to linear and logistic regression in a study where a single variable of interest exists in the presence of several correlated potential covariates. BMA is also used in an applied setting to demonstrate its use and its advantages. In these comparisons and example, this work aims to make BMA better known and more accessible to public health researchers.

1.2 Some Modeling Challenges in the China Health and Nutrition Survey

1.2.1 About CHNS

The China Health and Nutrition Survey (CHNS) collected health data in 361 communities (15 provinces and autonomous cities/districts of Beijing, Chongqing, Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning, Shaanxi, Shandong, Shanghai, Yunnan, and Zhejiang) throughout China in ten survey rounds from 1989 to 2015. Using a multistage, random cluster design, a stratified probability sample was used to select counties and cities stratified by income and urbanicity. Communities and households were then randomly selected from these strata. Survey procedures have been described elsewhere (Popkin 2010). The study was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill, the China-Japan Friendship Hospital, the Ministry of Health and China, and the Institute of Nutrition and Food Safety, China Centers for Disease Control. Participants gave informed consent.

The most obvious challenge of using CHNS in modeling is the shear number of variables available. CHNS records very detailed information about a participant's diet, health, socioeconomic situation, physical activity and community which leave the investigator with tens of thousands of variables available for study. It would not be hard to develop a model which suffered from collinearity or over-parameterization.

1.2.2 Predicting Visceral Adipose Tissue

A more specific challenge arises from a particular scientific interest. There is heterogeneity in the metabolic risk of obesity, some obese individuals are at very high metabolic risk, while others are not and being able to predict people who fall in this category is critical for targeting intervention and for understanding the health of a population. While there is debate about which depot of fat may be causally responsible for metabolic complications of obesity (Fabbrini et al. 2009, Klein 2004), visceral fat has been shown to be associated with metabolically abnormal obesity (Pouliot et al. 1992, Banerji et al. 1995, Gastaldelli et al. 2002). Visceral fat has stronger associations with cardio-metabolic diseases than BMI (Wajchenberg 2000, Fontana et al. 2007, Saito et al. 2012, Beaumont et al. 2016), the standard measure of obesity.

Visceral adipose tissue (VAT) can be expensive to measure and may not be historically available in large population studies. Computed tomography (CT) and magnetic resonance imaging (MRI) are considered the gold standard of VAT measurement (Rankinen et al. 1999, Seidell et al. 1990, Koester et al. 1992, Ross et al. 1992, Van der Kooy et al. 1993). Dual-energy x-ray absorptiometry (DXA) whole body scans have been suggested as an alternative (Snijder et al. 2002, Bertin et al. 2000, Direk et al. 2013). None of these measuring techniques are feasible in large population studies. Instead, a variety of anthropometric measures have been suggested as indices of VAT. Waist circumference (Pouliot et al. 1994, Grundy et al. 2013, Ross et al. 1996) and waist-to-hip ratio (Ashwell et al. 1985, Rankinen et al. 1999) have been found to correlate with visceral fat. Body mass index (BMI) is used to define obesity and is commonly used in clinical and epidemiological studies (Smalley et al. 1990, Spiegelman et al. 1992). Other measures considered with varying success in multivariable models have included BMI (Janssen et al. 2002, Goel et al. 2008), waist-to-height ratio (Swainson et al. 2017), hip circumference (Goel et al. 2008), conicity index (Pinho et al. 2017), sagittal diameter (Pinho et al. 2017), neck circumference (Pinho et al. 2017). Investigating the predictive ability of more readily accessible anthropometric and demographic measures in a multivariable model is an important step in exploiting the richness of existing population studies to better understand the role of visceral fat in the development of metabolically abnormal obesity. Further, a predictive model could help establish better identification of metabolically abnormal obesity in a clinical setting.

1.3 Notation

1.3.1 Linear Model Review

To better understand the competing methods of variable selection, we need to first review linear models. Linear models are devised to investigate the relationship between the response, Y_i and the explanatory variables $x_{i0}, \ldots, x_{ip}$. Assume there are n subjects, $i = 1 \ldots, n$, with p variables recording their traits. A linear model is of the form

$$Y_i = \beta_0 + \beta_1 \phi_1(X_{i1}) + \ldots + \beta_p \phi_p(X_{ip}) + \epsilon_i \tag{1.1}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{1.2}$$

Note, $\phi_1, \ldots \phi_M$, can be nonlinear functions. The predicted values of Y are denoted and are defined as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \phi_1(X_{i1}) + \ldots + \hat{\beta}_p \phi_p(X_{ip}) \tag{1.3}$$

where $\hat{\beta}_j$ are the estimates which minimize the squared error, also called the residual sum of squares, RSS.

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1\phi_1(X_{i1}) - \ldots - \hat{\beta}_p\phi_p(X_{ip}))^2 \tag{1.4}$$

For convenience, we will use matrix notation as shown below.

$$\mathbf{Y} = (Y_1, \ldots, Y_n)' \tag{1.5}$$

$$X_i = (X_{i1}, \ldots, X_{ip}) \tag{1.6}$$

$$\mathbf{X} = (X_1, \ldots, X_n)' \tag{1.7}$$

$$\beta = (\beta_0, \ldots, \beta_m)' \tag{1.8}$$

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \tag{1.9}$$

Continuing with the matrix notation, the estimates of β, predicted values and RSS are shown.

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.10}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \tag{1.11}$$

$$RSS = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \tag{1.12}$$

Ideally, we would use every available variable in our model, but we are often limited by n or run into problems where two or more variables are trying to represent the same concept, called collinearity. In very large studies, where sample size poses no real limitation on modeling, we are still limited by our ability to understand and draw conclusions from a very large and complicated model and can benefit by introducing some parsimony to the model.

CHAPTER 2: LITERATURE REVIEW

It is helpful to review types of modeling error. We the describe the various variable selection strategies used in this dissertation. Lastly, we examine existing comparisons and critiques of these methods.

2.1 Defining Modeling Error

In order to be able to determine an appropriate parsimonious model, it is useful to consider ways to quantify model error beyond the RSS. Here we define modeling error, ME, and prediction error, PE. Consider two independent samples drawn from a single population, sample A and sample B. Suppose we use sample A to build a model and sample B to examine the model's prediction ability (Seber and Lee (2003)).

$$\mathbf{Y_A}' = (Y_{A1}, \ldots, Y_{An}) \tag{2.1}$$

$$\mathbf{Y_B}' = (Y_{B1}, \ldots, Y_{Bm}) \tag{2.2}$$

Again, Y represents the response and we assume $\mathbf{Y_A}$ and $\mathbf{Y_B}$ have covariance matrices of $\sigma^2 I_n$ and $\sigma^2 I_m$ respectively. Let $\mathbf{X}$ represent the covariates with X'_{ai} and X'_{bi} as $(p+1) \times 1$ vectors with the first entry equal to 1.

$$\mathbf{X_A} = \begin{bmatrix} X'_{a1} \\ \vdots \\ X'_{an} \end{bmatrix} \quad \mathbf{X_B} = \begin{bmatrix} X'_{b1} \\ \vdots \\ X'_{bm} \end{bmatrix} \tag{2.3}$$

The least squares estimate of model based on sample A is $\hat{\beta}_{\mathbf{A}} = (\mathbf{X_A}'\mathbf{X_A})^{-1}\mathbf{X_A}'\mathbf{Y_A}$. Using these estimates, we can determine predicted values for $\mathbf{Y_B}$, $\mathbf{Y_B}^* = \mathbf{X_B}\hat{\beta}_{\mathbf{A}}$. We define prediction error, PE, in the way Seber and Lee do, as the expectation with respect to sample B of the sum of the squared errors.

$$\begin{aligned} PE &= E_{\mathbf{Y_B}}\left[\sum_{i=1}^{n}(Y_{Bi} - X'_{bi}\hat{\beta}_{\mathbf{A}})^2\right] \\ &= n\sigma^2 + \sum_{i=1}^{n}(E_{\mathbf{Y_B}}[Y_{Bi}] - X'_{bi}\hat{\beta}_{\mathbf{A}})^2 \end{aligned} \tag{2.4}$$

In this way, the prediction error is seen as a function of the underlying variability of the data, $m\sigma^2$, and a term which describes how well the model fits the data, which we define as model error or ME. Because sample A and sample B are from the same population and further have identical probability structures, we can further simplify ME. For example, let $\mathbf{X_A} = \mathbf{X_B}$ which implies $n = m$ and $E[\mathbf{Y_A}] = \mu_A = \mu_B$. Define $\epsilon = \mathbf{Y_A} - \mu_A$ and $\mathbf{P} = \mathbf{X_A}(\mathbf{X}'_{\mathbf{A}}\mathbf{X_A})^{-1}\mathbf{X}'_{\mathbf{A}}$.

$$\begin{aligned} ME &= \sum_{i=1}^{n}(\mu_A - X'_{ai}\hat{\beta}_A)^2 \\ &= \mu'(I_n - P)\mu + \epsilon' P \epsilon \end{aligned} \tag{2.5}$$

The expected value for ME is

$$
\begin{aligned}
E\left[ME\right] &= \mu'(I_n - P)\mu + \sigma^2 k & (2.6)\\
&= \sum_{i=1}^{n}(\mu - E\left[\mathbf{X_A}\hat{\beta}_{\mathbf{A}}\right])^2 + \sigma^2 tr(P) & (2.7)\\
&= TotalBias^2 + TotalVariance & (2.8)
\end{aligned}
$$

Further, when $\mathbf{X_A} = \mathbf{X_B}$ the expected prediction error is a sum of $E\left[n\sigma^2\right]$ and $E\left[ME\right]$.

$$E\left[PE\right] = (n + p + 1)\sigma^2 + TotalBias^2 \tag{2.9}$$

From these we see that adding variables to a model will affect our expected errors in a variety of ways. As we add more variables, our total variance will increase as p increases, while total bias should decrease unless the new variables linearly depend on the old. The model with the lowest expected ME will also have the lowest expected PE. Including all possible variables can indeed lead to poor prediction qualities. Deciding weather unbiased estimates or better prediction is desired will be key to determining which variable selection method to use.

2.2 Sequential Testing Techniques: Forward, Backward, and Stepwise

Sequential testing techniques in variable selection are data-driven methods of developing a model. As the name suggests, this method consists of steps beginning with either the full model, one which includes all possible variables and steps backward to a concise reduced model, or of steps beginning with a null model, one which only includes the most essential variables perhaps only the intercept term and steps forward to an adequate model. At each step, a model criterion is assessed to determine when the steps are complete. There are a variety of established criterion including the Akaike information criterion (AIC), Bayesian information criterion (BIC), as well as individual variable p-values. Additionally, forward and backward selection can be combined in an algorithm that can move in either direction called stepwise.

2.2.1 Variable Selection Criteria Used in Comparing Models

When faced with many possible variable combinations, we would usually like to find the model that most closely fits with reality. Unsurprisingly, there are a variety of ways to quantify how closely any model fits the unobservable truth. Perhaps it could be sensible to select the model with the lowest prediction error, or the model with the best fit to the observed data, or the model with the least overall differences in the distribution of $\mathbf{Y}$ and the modeled distribution of $\mathbf{Y}$,$\hat{\mathbf{Y}}$, or lastly, maybe we can estimate a probability that each model is the true model and select the model which is deemed most likely. These concepts are the driving force behind statistics such as Mallows' C_p, R^2, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). We focus on the selection criteria which are most commonly used.

Akaike Information Criterion, AIC

Akaike Information Criterion, AIC, attempts to quantify the difference between the distribution of the data, Y, and the distribution specified by the model in question. Specifically, the AIC estimates the Kullback-Leibler discrepancy between the true model, noted here as f, and the estimates from competing models, g (Seber and Lee (2003)).

$$AIC = -2logg(Y, \hat{\theta}(Y)) + 2r \tag{2.10}$$

where r is the dimension of the parameter vector θ (Akaike (1998)). Essentially AIC is a measure of the log-likelihood with a penalty based on the number of parameters estimated.

While many modifications to the AIC exist, this is the formulation we will be using throughout this work.

For the purposes of elucidating model comparison criteria, one useful modification to AIC arises if we assume σ^2 is known. Then there are only p parameters to estimate and

$$-2logg(Y, \hat{\theta}(Y)) = nlog(2\pi\sigma^2) + \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{\sigma^2}. \tag{2.11}$$

In this way, up to a constant, $nlog(2\pi\sigma^2)$, which does not depend on the model, AIC can be defined as

$$AIC = \frac{RSS_p}{\sigma^2} + 2p. \tag{2.12}$$

Note, if we replace σ^2 above with an estimate, $\hat{\sigma}^2$, and add n we arrive at Mallows' C_p. Additionally, if we replace the penalty term of $2p$ with a penalty of $nlogp$, which depends on n, we arrive at the Bayesian Information Criterion discussed below.

Bayesian Information Criterion, BIC

Introduced by Schwartz in 1978, BIC is based on an approximation to the posterior probability of a model. Although Bayesian is in the name, and the criterion is motivated by Bayesian ideas, the priors do not form part of the criterion. In general, Bayesian methods take an investigator's prior beliefs about parameters and update these beliefs based on the observed data. In the case of BIC, the prior belief is in regard to the probability that a model is correct and the criterion selects the model with the highest probability after updating the priors with the observed data, also called the posterior probability. Seber and Lee (2003) show how the log of the posterior probability can be approximated by

$$-\frac{RSS_p}{2\sigma^2} - \frac{1}{2}plogn + O(1). \tag{2.13}$$

Thus, selecting the model with the highest posterior probability is asymptotically equivalent to selecting the model which minimizes

$$BIC = \frac{RSS_p}{\sigma^2} + plogn. \tag{2.14}$$

When there are 8 or more observations, BIC imposes a stronger penalty for adding a new parameter than AIC, and as n increases the two criterion continue to diverge (Schwarz et al. (1978)).

2.2.2 Backward Elimination

Backward elimination begins with a model which is as full as possible, in our notation, it would contain p variables. Then p models are fit each missing a single variable from the full model. The best model is chosen from among these and the process is repeated fitting next $p - 1$ models. The eliminations are deemed complete when a chosen criteria is met. Backward elimination is particularly useful in a study suffering from collinearity. If two variables are equally explicatory of the outcome and are correlated with each other, then included together neither may appear significant. However, backward elimination can only remove one at a time, not both and the stronger one would be retained in the model (Mantel (1970)). This will not be the case in a combined forward and backward stepwise procedure.

2.2.3 Forward Selection

Forward select begins with a null model. Each of p models are fit each containing a single new variable and the best among these is chosen. This process repeatedly adds the best variable until stopping criteria are met. Mantel (1970) also cautions against forward selection in favor of backward elimination. He claims that when two variables only jointly effect the outcome and have no effect individually, the forward selection procedure will exclude them, while the backward procedure will retain them.

2.2.4 Two-Stage Selection

Another form of forward selection consists of pre-screening explanatory variables in models which include each variable alone. Then, only variables that are determined to be individually important are included in a multivariable model. Note, for the two-stage method to function adequately, a higher p-value cut-off, such as 0.2, should be considered (Mickey and Greenland 1989). This two-stage method has been shown to grossly underestimate the final p-value and will miss confounders which may be insignificant alone, but important to include in the model (Sun et al. 1996) (Viallefont et al. 2001).

2.3 Change in Effect, CIE

As presented in Kleinbaum et al. (1998), methods that eliminate variables based on whether their coefficients are different from zero do not address confounding. Consider a very simple situation with only three variables. We are interested in whether T is related to Y and have a possible confounder, C. We fit two models, one we call crude and the other adjusted.

$$Y = \beta_{0,crude} + \beta_{1,crude}T + \epsilon \tag{2.15}$$

$$Y = \beta_{0,adj} + \beta_{1,adj}T + \beta_2 C + \epsilon \tag{2.16}$$

Confounding is present when $\beta_{1,crude} \neq \beta_{1,adj}$. However, Kleinbaum et al. (1998) states that this inequivalence is a subjective decision and not meant to be based on a statistical test. This idea was previously presented by Kleinbaum et al. (1982). Statistical tests and other non-subjective criteria applied to this assessment of confounding proved too tempting in practice and it did not take long for rules of thumb to appear. Mickey and Greenland (1989) investigate using percent change cut-offs of 5, 10, 15, 20, or 25% in simulation studies, ultimately suggesting 10%.

CIE (also called change in estimate) is a procedure that begins with a full model. Then models with one variable removed are fit for each variable and the change in the variable of interest is recorded. The variable that changed the variable of interest the least is dropped and the procedure is repeated until dropping any variable results in a percent change greater than some threshold. The commonly recommended threshold is 10%. Lee (2014) suggests a procedure to determine a threshold specific to study with the following steps:

1. Simulate a random variable that follows a standard normal distribution.
2. Fit a model of the standardized outcome by the standardized exposure.
3. Compute the percentage difference of the regression slope, with and without adjusting for the random variable.
4. Repeat the above procedure r times.
5. Obtain the 95^{th} percentile and use this as the threshold.

Statistical properties of CIE are less explored than stepwise techniques.

2.4 Bayesian Model Averaging

Rather than settling on just one of the many possible models, Bayesian model averaging techniques take a weighted average of all the models. Consider the setting where an investigator has a proposed (finite) class of models M for a univariate outcome. When prediction is the primary interest, it can be shown that Bayesian Model Averaging (BMA) provides better average predictive ability than using any single model under the logarithmic scoring ruleMadigan and Raftery (1994). We define BMA in the following paragraph.

Suppose that Δ is a quantity of interest such as a future observation, and that the class of proposed models M contains elements $M_1, \ldots, M_K$. Then the posterior distribution given data D is given by

$$p(\Delta \,|\, D) = \sum_{k=1}^{K} p(\Delta \,|\, M_k, D)\, p(M_k \,|\, D) \tag{2.17}$$

by the law of total probability. Next, an application of Bayes' rule gives us

$$p(M_k \,|\, D) = \frac{p(D \,|\, M_k)\, p(M_k)}{\sum_{h=1}^{K} p(D \,|\, M_h)\, p(M_h)} \tag{2.18}$$

where

$$p(D \,|\, M_k) = \int p(D \,|\, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k \,|\, M_k) d\boldsymbol{\theta}_k \tag{2.19}$$

for $\boldsymbol{\theta}$ a vector of parameters of model M_k. In our case this would be a vector of the form $(\boldsymbol{\beta}', \sigma^2)'$. We will often refer to $p(M_k \,|\, D)$ as the Posterior Model Probability (PMP), as is convention. Then BMA refers to the process of constructing a weighted-average model through the use of the PMP as the weights. Another very useful posterior probability from this process is the posterior inclusion probability, or PIP. PIP is the probability of a variable appearing in a model and can be found by summing the PMP's of all model's which include the variable. PIP can also be thought of as $P(\beta_j \neq 0 | D)$.

BMA requires the user to define $p(M_j)$, also called the prior model probability. This allows the investigator to place a heavier weight on models which are deemed more likely. If there is not information available to inform this decision, each model can be assumed to be equivalently likely, or each variable could be assumed to have a 50% chance of being included.

While some settings allow for a completely Bayesian approach to model averaging, the BMA approach taken in this paper is one that is more widely applicable across settings which differ in model type and number of parameters considered.

One difficulty with BMA is calculating the integral in 4.7 above. As suggested in Yeung et al. (2005) we apply the BIC approximation presented in Raftery (Yeung et al. (2005), Raftery (1995)). Raftery suggests

$$p(D \,|\, M_k) \propto exp(-\frac{1}{2} BIC') \tag{2.20}$$

Using a Taylor series expansion Raftery shows

$$BIC'_k = \chi^2_{k0} + p_k logn \qquad (2.21)$$

where χ^2_{ko} is that likelihood ratio test statistic reported from a model.

See Hoeting et al. (1999)for a thorough history of BMA and a practical guide to BMA implementation.

Instead of enumerating every possible model, it has been argued that it more closely mirrors the scientific process to restrict the model set (Madigan and Raftery 1994, Raftery et al. 1997). Also, in settings with routinely large datasets it is not always possible to examine every possible model. Note, with K variables, this would result in 2^K models. With 15 variables, we would need to examine 32,768 models. The Occam's window idea described in detail in Madigan and Raftery (1994) provides a systematic method for wading through the large number of possible models resulting in a computationally efficient searching algorithm while accounting for model uncertainty. The Occam's window approach used in this work excludes models from consideration that are significantly less likely than the most likely model and/or contain sub-models which are dramatically more likely (Hoeting et al. 1999).

2.5 Stochastic Search Variable Selection (SSVS)

SSVS was presented in great detail by George and McCulloch in 1993. SSVS is a hierarchical fully Bayesian model which uses latent variables to model variable inclusion. For example, in the setting of variable selection, β_j can be modeled as a mixture of two normal distributions with different variances (George and McCulloch 1993; 1997). Consider a latent variable, γ_j, where $P(\gamma_j = 1) = \pi_j$. The mixture distribution for β_j is

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2). \qquad (2.22)$$

Selecting τ_j to be a small positive number and c_j a large number greater than 1, ensures that when $\gamma_j = 0$, β_j is likely to be zero and when $\gamma_j = 1$, β_j is unlikely to be zero. In this way, π_j can be thought of as the prior probability that variable X_j should be included in the model (George and McCulloch 1993). The above can be achieved with a multivariate normal prior

$$\beta|\gamma \sim N_p(0, \boldsymbol{D}_\gamma \boldsymbol{R} \boldsymbol{D}_\gamma) \tag{2.23}$$

where $\gamma = (\gamma_1, ...\gamma_p)$, $\boldsymbol{R}$ is the prior correlation matrix, and

$$\boldsymbol{D}_\gamma = diag[a_1\tau_1, ..., a_p\tau_p] \tag{2.24}$$

with $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$. An inverse gamma prior is used for $\sigma^2|\gamma$,

$$\sigma^2|\gamma \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma\lambda_\gamma}{2}). \tag{2.25}$$

Using the interpretation that ν_γ is the number of observations and $\frac{\nu_\gamma}{(\nu_\gamma-2)}\lambda_\gamma$ is the estimate of σ^2 in an imaginary previous experiment can be helpful in selecting the hyper-parameters. For a detailed discussion about all the prior specifications, see George and McCulloch (1993).

In standard Bayesian form, these prior distributions are combined with the observed data to to form posterior distributions. SSVS uses a Gibbs sampler to characterize the posterior distribution (George and McCulloch 1993; 1997). The estimated posterior mean value of γ_j approximates PIP and the estimated posterior mean of the vector γ approximates PMP. This technique is not ideal but running the chain until stationarity is reached helps (George and McCulloch 1993).

When introduced, the authors cautioned that SSVS may be slow to converge if multiple models had high posterior probability. This can frequently happen when variables are collinear. They suggest eliminating collinearities before using SSVS, or performing two rounds of selection this first using SSVS and the second using SSVS again or backward elimination (BE) (George and McCulloch 1993).

2.6 Adaptive Lasso

Ridge regression, lasso and adaptive lasso all seek a model which minimizes a penalized residual sum of squares (RSS) by simultaneous estimation and variable selection. Ridge regression penalizes with a tuning parameter, λ, multiplied by the sum of the squared coefficients and the lasso penalizes with a tuning parameter times the sum of the absolute value of the coefficients (Tibshirani 1996). While ridge regression can shrink estimates, it cannot eliminate them from the model by forcing the estimates to be zero thereby simplifying the results. Lasso was developed as a way to combine the shrinkage idea in ridge regression and the selection idea of best subsets to both improve predictive performance and simplify the model(Tibshirani 1996). The adaptive lasso generalizes the lasso by using a penalty which allows weights to be applied to the absolute value of the coefficients as follows (Zou 2006).

$$RSS(\beta) + \lambda \sum_{j=1}^{p} \hat{w}_j \left|\beta_j\right| \tag{2.26}$$

Unlike the lasso, adaptive lasso has what is known as an oracle property, meaning, under certain conditions, and as n increases the adaptive lasso will find the correct model (Zou 2006). Adaptive lasso is selected in this study to allow investigation of the setting where there is a particular variable of interest that must be included in every model. By setting $w_j = 0$ for the variable of interest, inclusion is guaranteed (Friedman et al. 2010). See Hastie and Qian for an introduction to running these computationally intensive procedures with **`glmnet`** package in RHastie and Qian (2016), Simon et al. (2011).

Since introduction, the lasso has provided a computationally efficient approach for exploring large sparse data (Tibshirani 2011). However, Tibshirani acknowledges that it is difficult to obtain standard error estimates from the lasso (Tibshirani 1996). Zou suggests an approximation for standard errors with adaptive lasso. Zou's approximation is not easily implemented. Chatterjee et al. (2013) suggest using residual bootstrap methods. To mimic use by a non-expert user, neither are included in this study.

2.7 Previous Work Comparing Variable Selection Criteria and Methods

Model building sits squarely on the line between science and art. Almost every investigator has their own strategy for selecting variables and assessing model fit. This goes against the urge to have scientific uniformity and, therefore, much has been written about methods for variable selection. Presented here are the works most relevant to our endeavor of comparing popularly used methods in epidemiology.

2.7.1 Popularity of Variable Selection Methods

Walter and Tiemeier (2009) presents a survey of the articles presented in *American Journal of Epidemiology*, *Epidemiology*, *European Journal of Epidemiology*, and *International Journal of Epidemiology* in 2008. He shows 27.7% of authors used only prior knowledge to build their model, 19.7% used a form of stepwise methods, 14.7% used change-in-effect methods, and 3% used other methods such as propensity scores or principal components. None of the 300 articles reviewed used ridge regression or shrinkage methods, while a 35% did not disclose their model building strategy. This report suggests BMA methods are either unknown or not easily accessed by those publishing in epidemiological journals.

37.3% reported using a method of variable selection beyond relying solely on prior knowledge (Walter and Tiemeier 2009). To put 37.3% in perspective, consider a separate study of commonly used statistical methods in published public health research journals which included the above four in addition to *American Journal of Public Health*, *Bulletin of World Health Organization*, and *American Journal of Preventive Medicine*. Hayat et al.

(2017) reports that 25.9% of studies report a Chi-squared test or a Fishers exact test, 38.4% report using logistic regression, and 40.7% report an odds ratio (Hayat et al. 2017). Variable selection methods are reported more often in public health research than linear models and Cox proportional hazards models combined, 19.4% and 15.3% respectively (Hayat et al. 2017). While many articles have been published using results from a variable selection method, these methods are not without potentially alarming flaws.

Model uncertainty can be appropriately represented if estimates from every model considered are somehow accounted for (Buckland et al. 1997). Model averaging and Bayesian methods have different approaches for accounting for the various models considered (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999, George and McCulloch 1993). Bayesian variable selection methods were initially proposed as a clever way to focus on promising subsets of variables without enumerating 2^p models (George and McCulloch 1993). Shrinkage techniques were developed to build predictive models more stable and with less variance than those built by the discrete process of classic variable selection methods (Tibshirani 1996). All of these modern methods have seen rapid increase in use. Figure 2.1 shows citation frequency for three popular representatives of these modern method categories: Bayesian model averaging (BMA), Bayesian variable selection with stochastic search variable selection (SSVS), and least absolute shrinkage and selection operator (lasso). As modern methods grow in popularity, they spur adaptations for specific problems and become more computationally accessible to researchers. Application of these modern methods will only become more wide-spread. It is vital to understand how these modern methods behave in a real-data setting (George 2000).

2.7.2 Debates about Best Criterion in Sequential Analysis

Both the variable selection method and the specific criterion used within these methods has been debated. George (2000) provides a concise, helpful history and summary of the methods

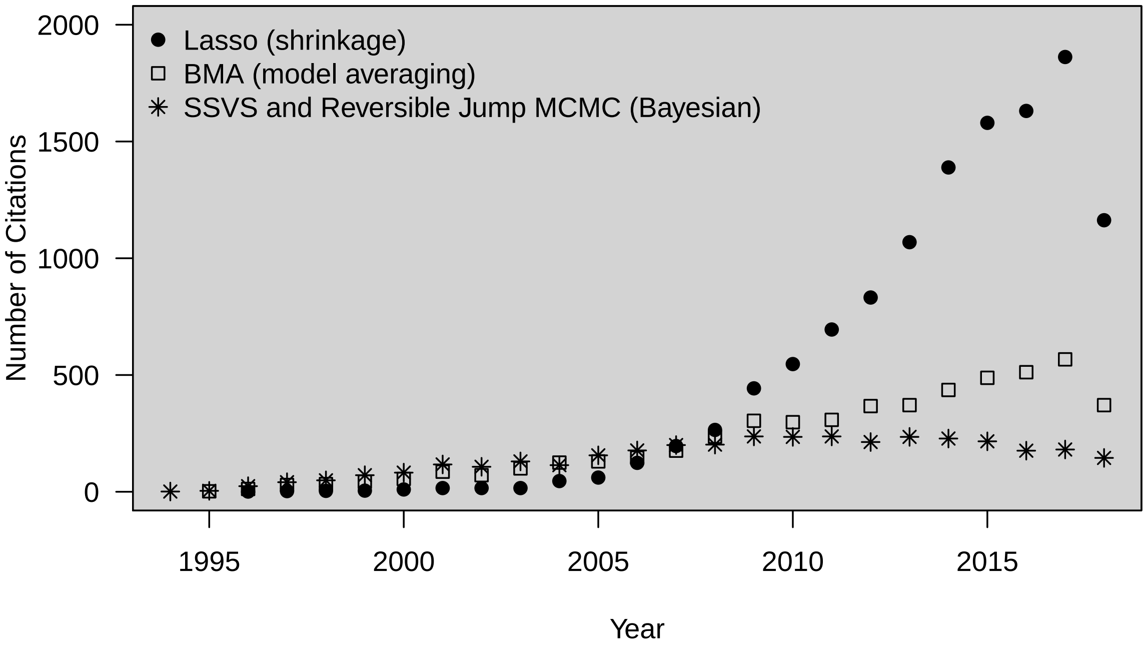

Figure 2.1: Citations of Modern Variable Selection Methods.
All included citations were counted on until September 10, 2018 from ISI Web of Knowledge. BMA includes citations from Raftery (1995), Raftery et al. (1997), Hoeting et al. (1999). Bayesian includes citations from George and McCulloch (1993), Green (1995). Lasso includes citations fromTibshirani (1996).

described here. First we consider the differences in AIC and BIC. A good deal of discussion exists about which criteria AIC or BIC will select the best model (Burnham and Anderson 2004). Recall AIC was derived as an approximation to the Kullback-Leibler distance between the distribution of Y as modeled in the γ^{th} model and the true distribution of Y. (Stone 1977) showed model selection with AIC is asymptotically equivalent to model selection with cross-validation. BIC was derived as an approximation to the log of the posterior probability of a model. (Schwarz et al. 1978) shows this is asymptotically equivalent to choosing a model based on Bayes factors.

According to Yang (2005), AIC, and other estimators of this type such as Mallows' C_p, usually yield minimax-rate optimal estimators of the regression function under a squared-error loss. Yang (2005) describes BIC as a consistent model selector, meaning that if the true model is among the models considered, as n increases, the probability that BIC will identify the true model converges to 1. The same cannot be said of AIC. However, unlike AIC, BIC cannot be said to be a minimax-rate optimal estimator. This would suggest that the common practice of first selecting a best model and then drawing inference from this "best model", could be a deeply flawed procedure. If an investigator uses AIC as the model building criterion, they would yield asymptotically good estimates of the coefficients included but could quite easily be good estimates from a seriously flawed model. Similarly, relying on BIC is asymptotically likely to land on the correct variables to include in the model, if the true model is under consideration, but could yield seriously biased estimates of coefficients. Yang (2005) nicely summarizes their investigation by stating "... when model selection instability is high, combining the models can substantially improve the accuracy of estimation/prediction. On the other hand, when the best model can be easily identified, combining the models usually loses out to model selection."

Using p-values as criteria also has drawbacks particularly in the presence of confounders. As stated by Kleinbaum et al. (1998), testing $H : \beta_{2,adj} = 0$ does not address confounding, but precision. In other words, such a test evaluates whether significant additional variation in Y is explained by adding C to the model. For questions of etiology, confounding likely takes precedence over precision. Also, $\beta_2 \neq 0$ does not imply that $\beta_{1,crude} \neq \beta_{1,adj}$. Although CIE directly assesses whether $\beta_{1,crude} \neq \beta_{1,adj}$, it suffers in its common use with a percent cutoff. Specifically, Lee (2014) investigates how study traits such as sample size, effect size, variance, and exposure correlation with the confounder affect what percentage cutoff would yield 80% power and 5% type I error in linear and logistic regression. Perhaps unsurprisingly, these study traits greatly affect how CIE performs demonstrating that a general rule-of-thumb cutoff should be avoided. Instead, careful examination of the traits of the study should be undertaken to better understand CIE's operational characteristics in specific scenarios. Indeed, Lee (2014) proposes an interesting procedure to determine a situation specific cut-off.

Also reliant on the p-value criterion, so-called two-stage analyses which first individually test the relationship between explanatory variables and the outcome and proceed to a multivariable model only with variables found significant in the first stage have been shown to grossly underestimate the final p-value (Viallefont et al. (2001)). Sun et al. (1996) caution strongly that this method will miss confounders which may be insignificant alone, but important to include in the model. Mickey and Greenland (1989) compare a two-stage method in the setting of a logistic regression model in a case-control study to other methods to identify confounders, in particular CIE. The two-stage methods they explore are not quite as straightforward as what we have described so far, but the conclusion drawn is the same. When confounding is present, the two-stage method too easily dismisses important variables. For the two-stage method to function adequately a higher p-value cut-off, such as 0.2, should be considered. Also, a lower percentage cut-off for CIE could also be preferable (Mickey and Greenland (1989)).

A strikingly similar conclusion is drawn in a set of simulations investigating confounder selection in Poisson models (Maldonado and Greenland (1993)). Maldonado and Greenland (1993) find that CIE performs best when the cut-point is set to a "low" 0.10, while the two-stage methods required a higher cut-off of 0.20. They further suggest the CIE estimator "...does not start to adjust for the confounder until the magnitude of confounding is about half of the cut-point value; at this degree of confounding and below, this estimator has about the same amount of bias as the crude estimator. This bias occurs even when the sample size is large, but setting the cut-point to a tolerable level of bias seems to ensure that bias will be held well below that level. For example, using a 20 percent cut-point yields a point estimator with an average bias of about 10 percent when confounding is weak."

2.7.3 Model Uncertainty

AIC, BIC, p-values and CIE have so far been discussed in the context of comparing many models, selecting a single "best" model and proceeding with our analysis. This model selection process changes the meaning of the p-values in the "best" model and leads to underestimation

(Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). A proper accounting for the uncertainty in the actual model selection procedure needs to be incorporated to correct for this underestimation. An alternative to building the "best" single model is to fit multiple models and combine them in a sensible way. According to Hoeting et al. (1999), the idea to combine models first saw a rush of interest in the 1960s in economics journals and flourished again in the 1990s when new advancements were made and computing power was sufficient. Hoeting et al. (1999) provides a thorough history of BMA and a practical guide to BMA implementation. Madigan and Raftery (1994) note that averaging over all the models as BMA does provides better average predictive ability, as measured by a logarithmic scoring rule, than using any single model. BMA may not be a simple method to implement in every statistical analysis program, however Raftery (1995) show how to use BIC to approximate the posterior probability of a model making BMA more accessible to the analyst.

2.7.4 Previous Variable Selection Comparisons

Others have previously used simulation to compare some of the above methods. Viallefont et al. (2001) compares the two-stage analysis of individually testing possible explanatory variables before including then in the final model, backward stepwise regression and BMA in the setting of a case-control study. This study focuses on the difference in the p-values and posterior inclusion probabilities but doesn't explore how these methods differ in their estimates of the main effect or their predictive abilities. Genell et al. (2010) also compares variables selected by stepwise (forward with AIC criterion) to BMA. Oddly though, they require variables to have p-value<0.05 after stepwise and a posterior inclusion probability greater than 50% to be considered "selected" variables. Wang et al. (2004) compare backward elimination, forward inclusion, backward stepwise and forward stepwise to BMA in the context of logistic regression, claiming that BMA accounts for model uncertainty unlike the step-wise procedures. They find that BMA tends to outperform the other methods at selecting the correct model and at predicting the outcome (using the prediction score by Good). However, their simulation is limited to only 10 replications in each of two settings. All of the explanatory variables used in the simulation are independent and therefor this simulation does not address confounding. Wiegand (2010) compare backward elimination, forward selection and stepwise methods in logistic and proportional hazards setting. They particularly investigate the agreement between these methods. While many comparisons of various variable selection methods exist, comparisons of the change in effect method and Bayesian model averaging do not. Further, comparisons of stepwise regression and BMA focus heavily on which variables are included and less on how the method performs in terms of estimation and prediction.

Comparisons between modern variable selection methods have already been made (Srivastava and Chen 2009, Xu 2007, Yazdani and Dunson 2015, Blattenberger et al. 2014, Rockova et al. 2012). While interesting and well-tailored to the high-dimensional analysis setting, these previous studies all lack three key traits necessary for understanding behavior in a non-high-dimensional setting. First, all of these studies only report variable selection method performance in terms of prediction abilities and/or variable selection characteristics and do not report estimation capabilities. Second, most of these studies assume predictor variables are either uncorrelated or have a very simple correlation structure. Independent variables or minimal correlation is unlikely to appear in non-high-dimensional practice. Third, all of these studies report simulations based on 100 or fewer subjects. A few other studies show comparisons of some modern methods highlighted here in the non-high-dimensional setting (Swartz et al. 2008, Viallefont et al. 2001, Genell et al. 2010). In addition to lacking comparisons between all the methods of interest here, SSVS, BMA and adaptive lasso, these also fail to report on the quality of estimation (e.g., bias, coverage, etc.) of the variable selection methods. Finally, previous simulation studies have examined effects of various priors, tuning parameters, cut-offs and computational specification to understand how to maximize the capabilities of selection methods (Rockova et al. 2012, O'Hara et al. 2009). A naive user may not understand the effects of these selections on results and rely on the suggested defaults. Non-expert's use of variable selection methods needs careful examination to understand the actual impact of variable selection methods to applied research.

2.8 Summary

Many of the methods above have been shown to have excellent qualities. AIC and BIC are asymptotically consistent as p and n go to infinity respectively. BMA strong predictive performance, asymptotically. Sadly, our models do not live in the land of asymptotia. These models are squarely rooted in a finite reality, a reality in which we cannot know if the true model is in our group of examined models. We must instead turn to simulations to help us understand the finite sample behavior of these techniques. There are other variable selection techniques not discussed here. Methods such classification and regression trees, and bootstrapping the entire model selection process can be useful for fitting predictive models and each have their own merit. We focus our investigations here to mimic the use of these variable selection methods in public health practice. In chapter 3 we compare SSVS, adaptive lasso and BMA in a non-high-dimensional linear regression setting. In chapter 4 we rigorously compare backward selection based on AIC, BIC and p-values to CIE and BMA in the setting of logistic regression models. Finally, chapter 5 applies BMA to the CHNS example and acts as an introduction of BMA to the field of obesity epidemiology.

CHAPTER 3: APPLYING MODERN VARIABLE SELECTION TECHNIQUES TO A CLASSIC LINEAR REGRESSION SETTING

3.1 Introduction

Several variable selection techniques were introduced at the end of the 20th century. Computing advancements continue to make these methods more accessible, resulting in booming popularity for these modern methods. Simultaneous advancements in data collection, particularly in the "-omics" fields, continue to encourage further development and refinement of modern variable selection techniques. Many of these modern variable selection methods were developed specifically for high-dimensional data, where the number of variables considered is significantly larger than the number of subjects. As the popularity of modern methods grows, it remains unknown how these methods behave in a classical regression model.

Classical regression models were first introduced in the early 19th century (Stigler 1986) and have been used to predict outcomes, estimate specific effects, and understand the relationship between many variables in a model. Unlike the high-dimensional setting, the classic regression setting builds a model with fewer parameters than subjects. In the 1970s, several statistics were introduced for the purpose of selecting between competing models including Akaike Information Criterion (AIC)(Akaike 1974), Mallow's C_p (Mallows 1973), and Bayesian Information Criterion (BIC) (Schwarz et al. 1978). Many more criteria, and various strategies for applying them now exist including forward selection, backward elimination, stepwise selection, best-subset selection and so on, but all suffer from similar drawbacks. Chief among them is a failure to account for the uncertainty contained in the model selection process. Considering multiple models and then proceeding with the selected model as if it were known to be the correct model can cause several serious problems. Variance estimates are generally underestimated, p-values are generally inflated, prediction ability is overestimated, and results are not reproducible in another dataset (Harrell 2001, Viallefont et al. 2001, Sun et al. 1996, Hurvich and Tsai 1990). In addition to causing problems with estimation and prediction, these classic variable selection methods can also lead to a final model that is not a good representation of the relationships between the variables. For example, if two variables only jointly affect the outcome, forward selection may exclude them both (Mantel 1970). Similarly, if two variables are equally explicatory of the outcome and are correlated with each other, backward elimination would only retain one of them (Mantel 1970). Lastly, classic variable selection methods may not perform well or be impossible to utilize when the number of variables, p, is large. To completely examine every model, 2^p models would need to be examined. Even though these short-comings were well known as long ago as the 1980s (Miller 1984, Freedman and Freedman 1983, Flack and Chang 1987, Freedman et al. 1988), classic variable selection techniques continue to be widely used (Walter and Tiemeier 2009).

Several modern variable selection methods were developed as a direct response to the short-comings of classic variable selection. The modern methods discussed in this paper represent three broad categories of methods: model averaging, Bayesian models, and shrinkage techniques. Model uncertainty can be appropriately represented if estimates from every model considered are somehow accounted for (Buckland et al. 1997). Model averaging and Bayesian methods have different approaches for accounting for the various models considered (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999, George and McCulloch 1993). Bayesian variable selection methods were initially proposed as a clever way to focus on promising subsets of variables without enumerating 2^p models (George and McCulloch 1993). Shrinkage techniques were developed to build predictive models more stable and with less variance than those built by the discrete process of classic variable selection methods

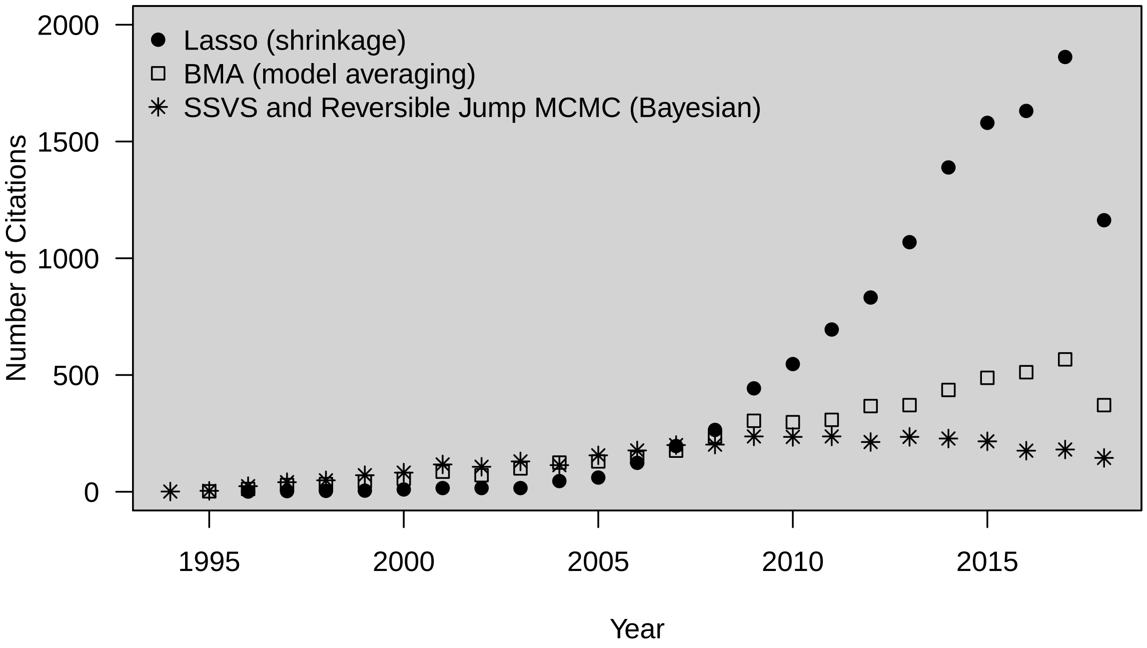

Figure 3.1: Citations of Modern Variable Selection Methods.
All included citations were counted on until September 10, 2018 from ISI Web of Knowledge. BMA includes citations from Raftery (1995), Raftery et al. (1997), Hoeting et al. (1999). Bayesian includes citations from George and McCulloch (1993), Green (1995). Lasso includes citations fromTibshirani (1996).

(Tibshirani 1996). All of these modern methods have seen rapid increase in use. Figure 3.1 shows citation frequency for three popular representatives of these modern method categories: Bayesian model averaging (BMA), Bayesian variable selection with stochastic search variable selection (SSVS), and least absolute shrinkage and selection operator (lasso). As modern methods grow in popularity, they spur adaptations for specific problems and become more computationally accessible to researchers. Application of these modern methods will only become more wide-spread. It is vital to understand how these modern methods behave in a real-data setting (George 2000).

Comparisons between modern variable selection methods have already been made (Srivastava and Chen 2009, Xu 2007, Yazdani and Dunson 2015, Blattenberger et al. 2014, Rockova et al. 2012). While interesting and well-tailored to the high-dimensional analysis setting, these previous studies all lack three key traits necessary for understanding behavior in a non-high-dimensional setting. First, all of these studies only report variable selection method performance in terms of prediction abilities and/or variable selection characteristics and do not report estimation capabilities. Second, most of these studies assume predictor variables are either uncorrelated or have a very simple correlation structure. Independent variables or minimal correlation is unlikely to appear in non-high-dimensional practice. Third, all of these studies report simulations based on 100 or fewer subjects. A few other studies show comparisons of some modern methods highlighted here in the non-high-dimensional setting (Swartz et al. 2008, Viallefont et al. 2001, Genell et al. 2010). In addition to lacking comparisons between all the methods of interest here, SSVS, BMA and adaptive lasso, these also fail to report on the quality of estimation (e.g., bias, coverage, etc.) of the variable selection methods. Finally, previous simulation studies have examined effects of various priors, tuning parameters, cut-offs and computational specification to understand how to maximize the capabilities of selection methods (Rockova et al. 2012, O'Hara et al. 2009). A naive user may not understand the effects of these selections on results and rely on the suggested defaults. Non-expert's use of variable selection methods needs careful examination to understand the actual impact of variable selection methods to applied research.

The study presented here compares estimation, prediction, and variable selection performance among modern variable selection methods, specifically BMA, SSVS and adaptive lasso. This study also includes comparisons to classic variable selection techniques. These methods are applied to linear regression in a study where a single variable of interest exists in the presence of possible confounding. A variety of sample sizes ranging from 150 to 20,000 are investigated. Freely available and easily executable methods with default tuning parameters and simple priors are used to mimic use by non-expert users.

Before exploring the merits of these methods more carefully via simulation, a brief background and description of the variable selection methods compared in this study are presented in Section 2. Section 3 presents the motivating example for this simulation study.

The simulation design follows in Section 4 with simulation results described in Section 5. Section 6 summarizes the detailed results presented in Section 4. Lastly, summary conclusions are made in the final section.

3.2 Background and Review of Representative Modern Variable Selection Methods

This simulation study focuses on the application of variable selection in a linear model. Linear models are employed to investigate the relationship between the response, Y_i and the p possible explanatory variables $x_{i0}, \dots, x_{ip}$. Assume there are n subjects, $i = 1 \dots, n$. A linear model is of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i \tag{3.1}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{3.2}$$

Note, $X_{i1}, \dots X_{ip}$, could also be nonlinear functions. The predicted values of Y are denoted by and are defined as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_p X_{ip} \tag{3.3}$$

In classical linear regression the $\hat{\beta}_j$ minimize the squared error, also called the residual sum of squares, RSS.

Many modern variable selection methods exist. Three representative methods are studied in this paper and are described below along with a brief description of the classic methods.

3.2.1 Bayesian Model Averaging

Regardless of the method used to select a best model, most researchers will acknowledge that it is difficult to ascertain whether the correct model has in fact been chosen, but will proceed with inference and prediction as if the model they use is indeed correct. Bayesian model averaging (BMA) provides an opportunity for exploring many possible models while appropriately accounting for the uncertainty surrounding variable selection (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). BMA is only one of several statistical methods developed for appropriately accounting for model uncertainty but has the advantage of providing optimal predictive performance (George 2000, Madigan and Raftery 1994). Madigan and Raftery (1994) measured predictive performance in a logistic regression model using a logarithmic scoring rule. However, given their reported lack of use, BMA methods are either unknown or not easily accessed by many researchers (Walter and Tiemeier 2009).

Bayesian model averaging performs inference using a weighed average of model-specific results from all the models considered. Consider the setting where a researcher has a proposed (finite) class of models M. Suppose that Δ is a quantity of interest, such as a predicted value of a future observation or an effect estimate, and that the set of proposed models M contains as elements the models $M_1, \ldots, M_K$. Then the posterior distribution given data D is defined by the following weighted average

$$p(\Delta \,|\, D) = \sum_{k=1}^{K} p(\Delta \,|\, M_k, D)\, p(M_k \,|\, D). \tag{3.4}$$

If the model building goal includes understanding the relationships between variables, it can be helpful to know the probability that each model is correct. It is convention to refer to $p(M_k \,|\, D)$, or the probability model M_k is correct given the observed data, as the Posterior Model Probability (PMP).

$$p(M_k \,|\, D) = \frac{p(D \,|\, M_k)\, p(M_k)}{\sum_{h=1}^{K} p(D \,|\, M_h)\, p(M_h)} \tag{3.5}$$

where

$$p(D \,|\, M_k) = \int p(D \,|\, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k \,|\, M_k) d\boldsymbol{\theta}_k \tag{3.6}$$

for $\boldsymbol{\theta}$ a vector of parameters of model M_k. In the case of the linear model, $\boldsymbol{\theta}$ would be a vector of the form $(\boldsymbol{\beta}', \sigma^2)'$. BMA constructs a weighted-average model by using PMP as the weights. Another very useful posterior probability from this process is the posterior inclusion probability, or PIP. PIP is the probability of a variable appearing the true model and can be found by summing the PMP's of all models which include the variable. PIP can also be thought of as $P(\beta_j \neq 0|D)$.

BMA requires the user to define $p(M_j)$, also called the prior model probability. This allows the researcher to place heavier weight on models which are deemed more likely. If there is not information available to inform this decision, each model can be assumed to be equivalently likely, or each variable could be assumed to have a 50% chance of being included.

While some settings allow for a completely Bayesian approach to model averaging, the BMA approach taken in this paper is one that is more widely applicable across settings which differ in model type and number of parameters considered. This study applies the BIC approximation presented in Raftery for the integral in (4.7) (Yeung et al. 2005, Raftery 1995). Instead of enumerating every possible model, it has been argued that it more closely mirrors the scientific process to restrict the model set (Madigan and Raftery 1994, Raftery et al. 1997). The Occam's window approach used in this paper excludes models from consideration that are significantly less likely than the most likely model and/or contain sub-models which are dramatically more likely (Hoeting et al. 1999). In settings with many variables, Occam's window is a less computationally intensive option. See Hoeting et al. (1999)for a thorough history of BMA and a practical guide to BMA implementation.

3.2.2 Stochastic Search Variable Selection (SSVS)

SSVS was presented in great detail by George and McCulloch in 1993. SSVS is a hierarchical fully Bayesian model which uses latent variables to model variable inclusion. For example, in the setting of variable selection, β_j can be modeled as a mixture of two normal distributions with different variances (George and McCulloch 1993; 1997). Consider a latent variable, γ_j, where $P(\gamma_j = 1) = \pi_j$. The mixture distribution for β_j is

$$\beta_j|\gamma_j \sim (1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2). \tag{3.7}$$

Selecting τ_j to be a small positive number and c_j a large number greater than 1, ensures that when $\gamma_j = 0$, β_j is likely to be zero and when $\gamma_j = 1$, β_j is unlikely to be zero. In this way, π_j can be thought of as the prior probability that variable X_j should be included in the model (George and McCulloch 1993). The above can be achieved with a multivariate normal prior

$$\beta|\gamma \sim N_p(0, \boldsymbol{D_\gamma R D_\gamma}) \tag{3.8}$$

where $\gamma = (\gamma_1, ...\gamma_p)$, $\boldsymbol{R}$ is the prior correlation matrix, and

$$\boldsymbol{D_\gamma} = diag[a_1\tau_1, ..., a_p\tau_p] \tag{3.9}$$

with $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$. An inverse gamma prior is used for $\sigma^2|\gamma$,

$$\sigma^2|\gamma \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma\lambda_\gamma}{2}). \tag{3.10}$$

Using the interpretation that ν_γ is the number of observations and $\frac{\nu_\gamma}{(\nu_\gamma - 2)}\lambda_\gamma$ is the estimate of σ^2 in an imaginary previous experiment can be helpful in selecting the hyper-parameters. For a detailed discussion about all the prior specifications, see George and McCulloch (1993).

In standard Bayesian form, these prior distributions are combined with the observed data to to form posterior distributions. SSVS uses a Gibbs sampler to characterize the posterior distribution (George and McCulloch 1993; 1997). The estimated posterior mean value of γ_j approximates PIP and the estimated posterior mean of the vector γ approximates PMP. This technique is not ideal but running the chain until stationarity is reached helps (George and McCulloch 1993).

When introduced, the authors cautioned that SSVS may be slow to converge if multiple models had high posterior probability. This can frequently happen when variables are collinear. They suggest eliminating collinearities before using SSVS, or performing two rounds of selection this first using SSVS and the second using SSVS again or backward elimination (BE) (George and McCulloch 1993). The two-round selection is not employed here.

3.2.3 Adaptive Lasso

Ridge regression, lasso and adaptive lasso all seek a model which minimizes a penalized residual sum of squares (RSS) by simultaneous estimation and variable selection. Ridge regression penalizes with a tuning parameter, λ, multiplied by the sum of the squared coefficients and the lasso penalizes with a tuning parameter times the sum of the absolute value of the coefficients (Tibshirani 1996). While ridge regression can shrink estimates, it cannot eliminate them from the model by forcing the estimates to be zero thereby simplifying the results. Lasso was developed as a way to combine the shrinkage idea in ridge regression and the selection idea of best subsets to both improve predictive performance and simplify the model(Tibshirani 1996). The adaptive lasso generalizes the lasso by using a penalty which allows weights to be applied to the absolute value of the coefficients as follows (Zou 2006).

$$RSS(\beta) + \lambda \sum_{j=1}^{p} \hat{w}_j \, |\beta_j| \tag{3.11}$$

Unlike the lasso, adaptive lasso has what is known as an oracle property, meaning, under certain conditions, and as n increases the adaptive lasso will find the correct model (Zou 2006). Adaptive lasso is selected in this study to allow investigation of the setting where there is a particular variable of interest that must be included in every model. By setting $w_j = 0$ for the variable of interest, inclusion is guaranteed (Friedman et al. 2010). See Hastie and Qian for an introduction to running these computationally intensive procedures with **`glmnet`** package in RHastie and Qian (2016), Simon et al. (2011).

Since introduction, the lasso has provided a computationally efficient approach for exploring large sparse data (Tibshirani 2011). However, Tibshirani acknowledges that it is difficult to obtain standard error estimates from the lasso (Tibshirani 1996). Zou suggests an approximation for standard errors with adaptive lasso. Zou's approximation is not easily implemented. Chatterjee et al. (2013) suggest using residual bootstrap methods. To mimic use by a non-expert user, neither are included in this study.

3.2.4 Classical Variable Selection Methods

For the purposes of understanding how modern variable selection techniques behave in a non-high-dimensional regression setting, classic variable selection techniques are also included in the simulation. Popular classic variable selection methods include forward selection, backward elimination, stepwise, best-subset, and two-phase (all significant univariate relationships included in a multivariate model) (Walter and Tiemeier 2009). Classical variable selection techniques have been described and compared in detail elsewhere (Mantel 1970, Miller 2002, Mickey and Greenland 1989, Sun et al. 1996, Viallefont et al. 2001). It is known that backward elimination is more appropriate than forward or step-wise in settings with collinearity (Mantel 1970). This study evaluates BE with selection criteria of AIC, BIC, p=0.05 and p=0.20, however for simplicity, only BE with BIC is shown. BE with BIC was found to be the least likely of these to suffer from an underestimate of model uncertainty in our full simulation and behaved similarly to BMA in regards to variable selection.

3.3 Motivating Example

The China Health and Nutrition Survey (CHNS) collected health data in 361 communities (15 provinces and autonomous cities/districts of Beijing, Chongqing, Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning, Shaanxi, Shandong, Shanghai, Yunnan, and Zhejiang) throughout China in ten survey rounds from 1989 to 2015. Using a multistage, random cluster design, a stratified probability sample was used to select counties and cities stratified by income and urbanization. Communities and households were then randomly selected from these strata. Survey procedures have been described elsewhere (Popkin 2010). The study was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill, the China-Japan Friendship Hospital, the Ministry of Health and China, and the Institute of Nutrition and Food Safety, China Centers for Disease Control. Participants gave informed consent.

For the purposes of demonstrating the traits of variable selection methods, this paper considers the specific goal of modeling whether waist-to-height ratio greater is than 0.5 among participants who were surveyed in at least two rounds while they were between the ages of 18 and 30 and have their waist circumference and height recorded. After excluding all missing data, N=1195. The specific variable of interest is a measure of urbanization and potential explanatory variables include the following: age, sex, income, years of education, physical activity, caloric intake, sodium, potassium, whether the participant smokes, whether the participant consumes alcohol, and whether the participant drinks coffee. Coffee drinking was included because it was highly correlated to urbanization, but was not believed to be related to waist-to-height ratios. Although perhaps unlikely to be included in an actual research setting, coffee drinking was included to investigate how the model selection methods would handle this type of relationship between variables. The simulation study generates data meant to mimic this example. All explanatory variables were standardized. The correlation between the explanatory variables, the effect sizes, and the error variance found in the example data guide the simulation parameters.

3.4 Design of the Simulation Study

All simulations were performed in R version 3.4.3 and all results presented are based on 5000 replications(R Core Team 2013). Note, 5000 replications results in Monte Carlo Error less than 0.7% for a proportion, such as the probability of including a variable (Koehler et al. 2009).

3.4.1 Data Generation

Table 3.1 shows the correlations between the explanatory variables. Explanatory variables are labeled with an X if continuous and B if binary. $X_{INTEREST}$ represents the variable of interest which is forced to be included in every model. Among the remaining explanatory variables, a subscript of E represents a variable with a true non-zero effect, and C represents a variable which is correlated with the variable of interest. Two variables are constructed with both an effect on the outcome as well as a correlation with the variable of interest ($X_{1,EC}$, $B_{11,EC}$). Also included are four variables completely independent of the others. Two of these are binary (B_{12}, $B_{13,E}$), and two are normally distributed (X_{14}, $X_{15,E}$). This correlation matrix, with the additional rows for the uncorrelated variables, is used to generate data from a multivariate normal of size $2n$ with all means equal to zero. The first n observations are used to build the model (i.e. the training set) and the second n are treated as an external sample for assessing predictive ability of the models (i.e. the testing set).

Table 3.1: Explanatory Variable Correlations for Data Generation.

	$X_{INTEREST}$	$X_{1,EC}$	$X_{2,C}$	$X_{3,C}$	$X_{4,C}$	X_5	X_6	$X_{7,C}$	B_8	$B_{9,E}$	B_{10}	$B_{11,EC}$
$X_{INTEREST}$	1	0.1	0.2	-0.5	-0.2			0.5				0.2
$X_{1,EC}$	0.1	1		0.2	-0.1						0.2	
$X_{2,C}$	0.2		1	-0.1		-0.1		0.2				0.1
$X_{3,C}$	-0.5	0.2	-0.1	1	0.1			-0.4				-0.1
$X_{4,C}$	-0.2	-0.1		0.1	1	0.4	0.6	-0.1	0.2	0.2	-0.3	-0.1
X_5			-0.1		0.4	1	0.3					
X_6					0.6	0.3	1		0.1	0.1	-0.1	
$X_{7,C}$	0.5		0.2	-0.4	-0.1			1		0.1	-0.1	0.2
$B_8{}^*$					0.2		0.1		1	0.5	-0.5	
$B_{9,E}{}^*$					0.2		0.1	0.1	0.5	1	-0.4	
$B_{10}{}^*$		0.2			-0.3		-0.1	-0.1	-0.5	-0.4	1	0.1
$B_{11,EC}{}^*$	0.2		0.1	-0.1	-0.1			0.2			0.1	1

[*] Some of the variables are transformed from normally distributed variables into binary variables with a cut point determined at a random quantile. For example, it is desired that on average 25% of observations have B_8 = 1. At each replication of the simulation a cut point varied uniformly between 20% and 30% to achieve an average of 25%. Desired observed proportions for B_8 through B_{13} are 0.25, 0.25, 0.5,0.05, 0.7, and 0.3 respectively.

Four variables, B_{12}, $B_{13,E}$, X_{14}, $X_{15,E}$ are uncorrelated with all other variables and are not included here.

Lastly, the outcome variable is generated in the following manner:

$$Y = 0.45 + 0.01X_{INTEREST} + 0.01X_{1,EC} - 0.01B_{9,E} - 0.01B_{11,EC} + 0.005B_{13,E} + \\ 0.005X_{15,E} + \epsilon \tag{3.12}$$

Sample sizes of $n = 150, 250, 500, 1000, 2500, 5000, 10000,$ and 20000 are investigated. Error is distributed normally with mean of zero and variance of 0.0016 for all n, resulting in an $r^2 \sim 0.13$. When $n = 1000$, error variances of .0625, 0.01 and 0.0004 are also considered in order to investigate models with higher and lower predictive abilities than what was observed in the CHNS example. These error variances result in r^2 of approximately 0.40, 0.03, and 0.01 respectively. Also for the setting of $n = 1000$, the situation where the variable of interest has no effect is considered. Although this data is more complicated than is generally reported in simulations comparing variable selection methods, it is a more accurate depiction of the conditions of a large observational study (Wang et al. 2004, Wiegand 2010, Lee 2014). A table outlining the full design of the simulation study is available in the supplemental material, Table A.1.

3.4.2 Variable Selection

After the data are generated, Bayesian model averaging (BMA), stochastic search variable selection (SSVS), adaptive lasso, and classic backward elimination are performed using the first n observations. Adaptive lasso and classic variable selection yield final models. BMA and SSVS do not reduce the number of variables but instead report the average estimates for all variables. In practice a researcher may still want to report a final model after using averaging techniques. This study considers the model with the highest PMP (top model), and the model resulting from including all variables with PIP¿0.5 (median model) as likely candidates for selecting a reduced model and are therefore also recorded. The variables included in these final models, the coefficient estimates, the standard error estimates and predicted values of the outcome for these selected models are recorded. BMA and SSVS provide averaged estimates of coefficients, standard errors and predicted values of the outcome which are also recorded. For BMA and SSVS, PMP and PIP values for each replication are recorded. Lastly, coefficient estimates from the final model or the averaged model from each method are applied to the last n of the $2n$ observations and predicted outcomes are recorded.

BMA uses the `bic.glm` function in the **`BMA`** package (Raftery et al. 2017). Occam's window with a maximum exclusion ratio of 20 as suggested in Madigan and Raftery is used(Madigan and Raftery 1994). SSVS is implemented with the **`BoomSpikeSlab`** package (Scott 2017). For both BMA and SSVS, prior probabilities for variable inclusion are 1 for the variable of interest, guaranteeing inclusion, and 0.5 for all remaining variables. A prior probability of inclusion of 0.5 indicates that, prior to collecting data, a researcher believes a variable is related to the outcome just as much as they believe the same variable is not related to the outcome. This is sometimes called an indifference prior. Adaptive lasso is implemented with **`glmnet`** package (Simon et al. 2011). For adaptive lasso, lambda was selected automatically using 10-fold cross-validation, penalty factors for the intercept and variable of interest were set to 0 to force their inclusion in every model, all other variables were set at the default of 1 (which **`glmnet`** rescales). Both adaptive lasso and SSVS allow for several other user-defined specifications. For example, other user-specified parameters for adaptive lasso allowed every variable to be included and no limits were set on what values the coefficients could be. In order to most most closely mimic performance of these methods in research applications, the default settings are used. BE to minimize BIC is performed using the `step` function provided in base R.

3.4.3 Quantities of Interest

Model Selection and Variable Inclusion Probabilities

The first of four key modeling aims this simulation will investigate is selecting the correct model. The model selected by each procedure is recorded, and the probability of selecting any particular model is the number of replications selecting a specific model and dividing by r, here 1000. Because so many variables are examined, the variety of models chosen by the selection methods considered here are well over 1000. For this reason, the models selected are summarized by determining if they select all, some or none of the variables with $\beta \neq 0$ ($X_{1,EC}$, $B_{9,E}$, $B_{11,EC}$, $B_{13,E}$, and $X_{15,E}$),and all, some, or none of the variables related to $X_{INTEREST}$ but with $\beta = 0$ ($X_{2,C}$, $X_{3,C}$, $X_{4,C}$, and $X_{7,C}$). These two groups are referred to as the E group, for effect, and the C group, for correlated. Note, the true model includes only the E group. Ideally the methods will select the true model, but if that cannot be achieved, the model should at least include all of the E group.

Additionally, the variable inclusion probabilities (IP) are presented. For all methods except BMA and SSVS, IP is the percent of simulation replicates selecting each variable in the final model. Average PIP across replications is reported for BMA and SSVS. The individual variable's IP are essential for understanding how these selection methods function with variables that differ in effect size, correlation, and type (binary vs. continuous). If the true model is not selected, these IP can help determine why.

Estimation of Main Effect

The second key modeling aim investigated is the estimation of the effect of $X_{INTEREST}$. If a researcher is mostly interested in the size of the association between the variable of interest and the outcome, it matters less which specific variables are included, and instead the accuracy and precision of $\hat{\beta}_{INTEREST}$ matter much more. Much like in practice, this study considered one variable to be the variable of interest and required all the models in consideration to include this variable. The bias, standard error of the coefficient estimates, ratio of the standard deviation of the coefficient estimates and the mean estimated standard error, coverage, and the probability of type I error or power, depending on the scenario, are recorded. The ratio of the standard deviation of the coefficients and the mean estimated standard deviation is a measure of how well methods account for model uncertainty. The standard deviation of the coefficients reflects the variance of the coefficient estimate including the variance attributed to the uncertainty in the model selection process. The mean of the standard error estimates represents the value likely to be reported by a researcher using a variable selection technique. If this ratio is greater than 1, the reported standard deviations are too small which in turn results in overly optimistic p-values.

Predictive Performance

The third modeling goal is predicting the outcome. Predictive performance is measured by the Pearson correlation between the predicted values of Y and the actual values of Y,$\rho_{Y_{orig},\hat{Y}}$, in the simulated original and external datasets. The difference between these two correlations reveals a degree of over-fitting, ω (Harrell 2001).

$$\hat{Y}_{orig} = X_{orig}\hat{\beta}_{orig} \tag{3.13}$$

$$\hat{Y}_{external} = X_{external}\hat{\beta}_{orig} \tag{3.14}$$

$$\omega = \rho_{Y_{orig},\hat{Y}_{orig}} - \rho_{Y_{external},\hat{Y}_{external}} \tag{3.15}$$

The ideal model has the highest $\rho_{Y_{external},\hat{Y}_{external}}$ and the lowest ω. Negative ω would indicate the model performed better in the external dataset than it did in the original.

Agreement

Agreement is not really a modeling goal but aids our understanding of how these methods relate to each other. The fourth and final quantity investigated is the probability that these variable selection strategies agree with each other. With so many variable selection methods available, one suggested technique is to perform multiple methods and declaring a model final if it is chosen by both methods (Wiegand 2010). Investigating method agreement will also highlight the similarities between the selection methods.

3.5 Simulation Results

3.5.1 Model Probabilities

Variable selection methods are often used with the expectation that they can uncover the true model. Figure 3.2 shows the probability of selecting the true model by n. Reliable selection of the true model does not occur for any method until $n \geq 5000$. BMA and BE BIC largely outperform SSVS and adaptive lasso until $n = 10000$, when SSVS catches up and remains similar for all larger sample sizes. Even when $n = 20000$, adaptive lasso only selects the true model in 5.3% of the replications. This suggests the oracle property of adaptive lasso may require an extremely large sample size, or the default settings are not adequate in this example. Investigating specific variable IP helps explain why the methods fall short of selecting the true model.

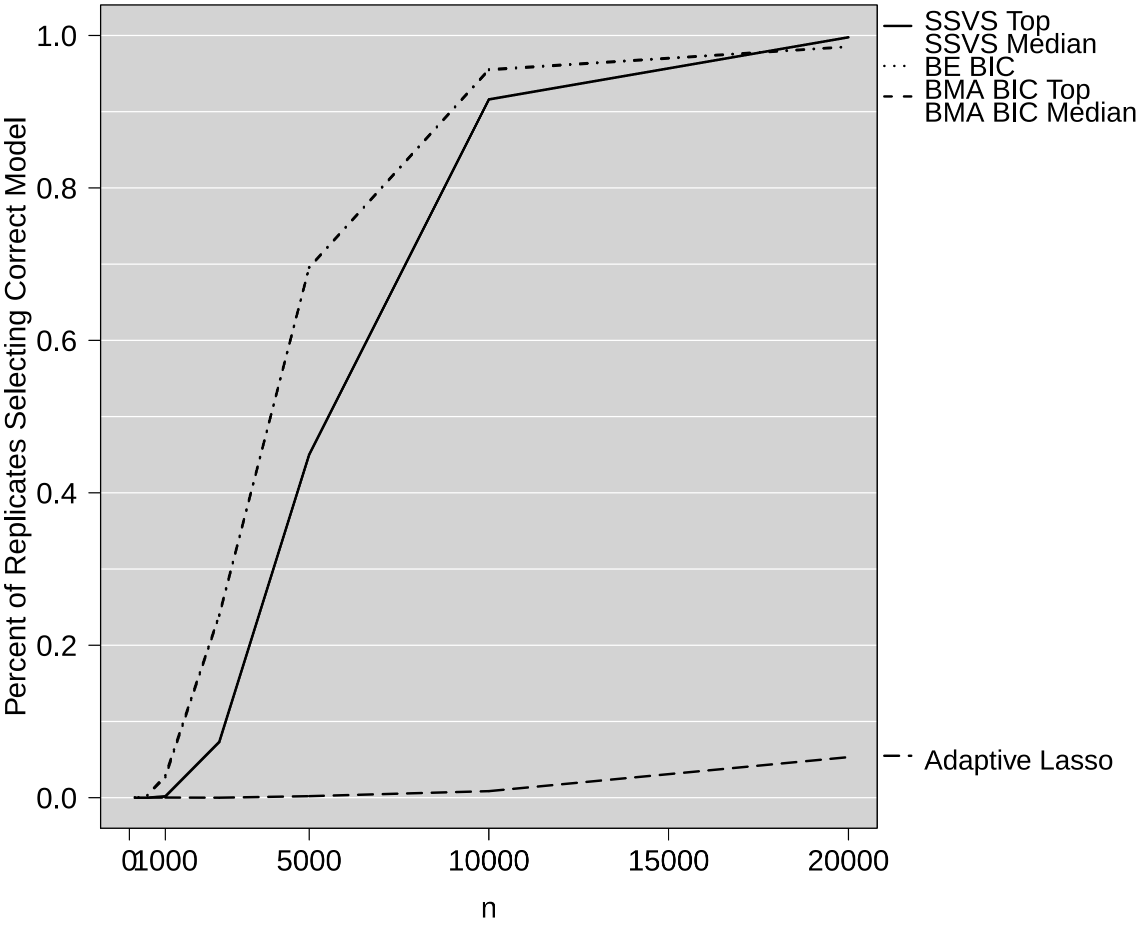

Figure 3.2: Probability of Selecting the True Model.

BMA Top and BMA median were almost always identical to each other and are both represented with the short dashed line. BE BIC also matched these two and is represented with a dotted line, giving the appearance of a single dot-dash line.

Figure 3.2 shows the probability that each method reports the model that contains all variables with $\beta \neq 0$ with no other variables. Adaptive lasso struggles to find $B_{11,EC}$, the only variable with an effect in the opposite direction of its correlation with $X_{INTEREST}$.

3.5.2 Variable Probabilities

Variable IP for a subset of the sample sizes is shown in Figure 3.3. The variables have been arranged with the most important group at the top, variables with both an effect and a correlation with $X_{INTEREST}$ and have EC as a subscript. These are followed by variables with an effect but no correlation (E subscript), variables with correlation but no effect (C subscript), and lastly variables with neither an effect nor a correlation (noise, no subscript). Other noise variables are not shown in the plot. The noise variables shown here had no correlation with the variable of interest and no correlation with any other variables considered. Also not shown are the IP from SSVS and BMA top and median models. The top and median models closely match their corresponding averages for IP. Variable IPs vary considerably not only by method but also by the variable effect size, potential correlation with $X_{INTEREST}$, variable type (continuous or binary), and sample size. Complete tables of inclusion probabilities for $n = 1000$ and $n = 20000$ can be found in the appendix.

All selection methods included our strongest variable, $X_{1,EC}$, which had a true effect and was positively correlated with the variable of interest, in 100% of the replications except when $n = 250$. $B_{11,EC}$ and $B_{9,E}$ have the same negative effect size, but $B_{11,EC}$ has a positive correlation with $X_{INTEREST}$ while $B_{9,E}$ in uncorrelated with all other variables. This correlation makes it harder for all the methods to include this variable, but it is particularly hard for adaptive lasso. Other models eventually find $B_{11,EC}$ as n increases, but adaptive lasso lags far behind. The difference in IP for $B_{11,EC}$ and $B_{9,E}$ when $n = 20000$ is particularly large. $B_{11,EC}$, a variable with an effect in the opposite direction of its correlation with the variable of interest, prevents adaptive lasso from selecting the true model, even when $n = 20000$.

Binary variables were more difficult to find than continuous variables. This is particularly striking for adaptive lasso. When $n \leq 1000$, BMA was approximately 1.5 times more likely to include $B_{9,E}$, $B_{11,EC}$, and $B_{13,E}$ than SSVS. All methods did well at excluding the variables with no effect. The probability of including these noise variables when $n = 250$ is 2% with BE BIC, and 6% with BMA. The selection rate of noise variables is the same regardless of correlation with $X_{INTEREST}$. This suggests that, unlike confounding, collinearity alone is not problematic for selecting variable with a true effect. It also suggests these methods alone cannot be used to gain understanding of the underlying relationships between all the variables.

3.5.3 Estimation of the Main Effect

The estimates of percent bias in $\beta_{INTEREST}$ by n are shown in Table 3.2 along with the 95% Monte Carlo confidence intervals. All of the methods have some bias when $n \leq 5000$, but this bias tends to shrink. However, adaptive lasso never reaches unbiasedness. Also, the averaged coefficients from BMA and SSVS are less biased than the coefficients from the top or median models. The average standard errors did not differ much by method and all decreased with increasing n and are not presented here.

Table 3.2: Estimation of Variable of Interest: Percent Bias.

	Percent Bias % [*]			
Selection Method	**n=250**	**n=1000**	**n=5000**	**n=20000**
BMA	-0.652	-1.099	-0.31	0.068
	(-1.414,0.11)	(-1.465,-0.733)	(-0.47,-0.151)	(-0.012,0.148)
BMA Top	-1.134	-1.264	-0.274	0.076
	(-1.909,-0.36)	(-1.634,-0.895)	(-0.435,-0.114)	(-0.005,0.156)
BMA Median	-1.133	-1.264	-0.275	0.076
	(-1.904,-0.361)	(-1.633,-0.895)	(-0.435,-0.114)	(-0.005,0.156)
SSVS	0.487	-1.762	-0.653	0.066
	(-0.253,1.227)	(-2.123,-1.402)	(-0.812,-0.493)	(-0.013,0.146)
SSVS Top	-0.142	-1.904	-0.639	0.067
	(-0.889,0.606)	(-2.265,-1.543)	(-0.799,-0.479)	(-0.013,0.146)
SSVS Median	-0.144	-1.903	-0.635	0.067
	(-0.887,0.6)	(-2.265,-1.542)	(-0.795,-0.475)	(-0.013,0.146)
Adaptive Lasso	-2.287	-2.179	-2.292	-2.133
	(-3.008,-1.565)	(-2.541,-1.816)	(-2.45,-2.134)	(-2.214,-2.052)
BE BIC	-1.152	-1.258	-0.273	0.075
	(-1.929,-0.375)	(-1.628,-0.887)	(-0.434,-0.112)	(-0.005,0.156)

[*] The interval shown is the Monte Carlo confidence interval calculated using the standard deviation of the estimates of the coefficients.

Variable selection methods are frequently criticized for underestimating the variance resulting in p-values which are too small and confidence intervals which are too narrow. The standard deviation of the beta coefficients can be compared to the mean estimates of the standard error in a simulation setting. Table 3.3 shows the ratio of these two. Values greater than 1 indicate the model selection method reports an under-estimate of the standard error of the coefficient. BMA and SSVS report ratios less than one indicating these methods do not suffer from under-estimation of the standard error of $\hat{\beta}_{INTEREST}$. BMA top, BMA median, SSVS top, SSVS median and BE BIC all have ratio's close to one. Adaptive lasso does not as readily provide standard error estimates as the other methods do. A combination of bias and standard error, coverage, is also shown in Table 3.3. Only BMA and SSVS had Monte Carlo intervals for coverage which included 0.95 for n=250, 1000, 5000 and 20000 (intervals not shown). By the time n = 20000 all methods have intervals which include 0.95.

An ideal method has low bias and standard error ratios close to or less than 1. Figure 3.4 shows these traits simultaneously. BMA and SSVS have conservative standard error ratios. As n increases, all methods except for adaptive lasso move closer to the ideal of no bias and standard error ratio of one.

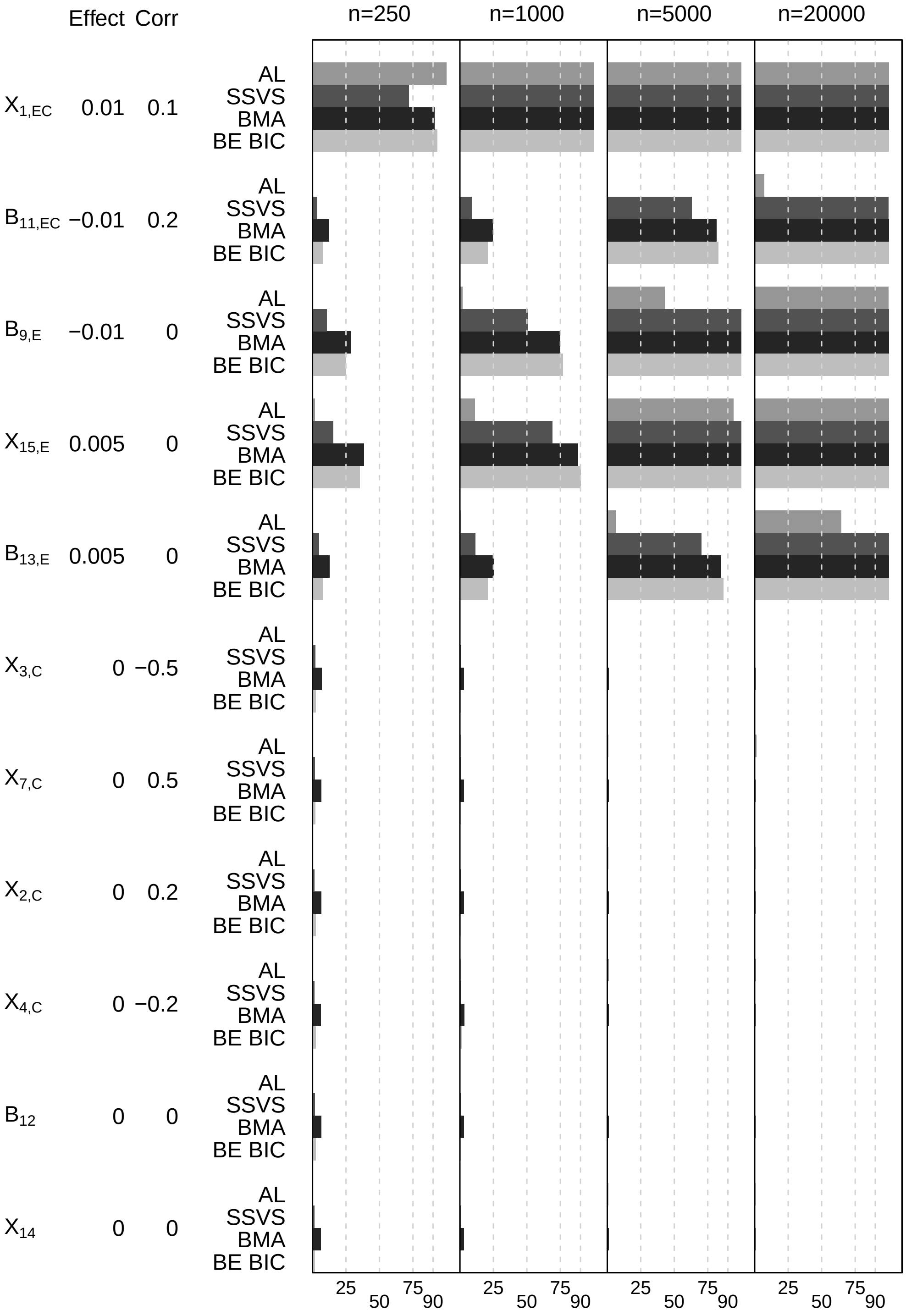

Figure 3.3: Variable Inclusion Probabilities.

All the methods do well at excluding variables with no effect. All the methods have a hard time identifying $B_{11,EC}$ when $n < 5000$, but adaptive lasso (AL) struggles to find it even when $n = 20000$.

Table 3.3: Estimation of Variable of Interest: Variance and Coverage.

	Ratio of $sd(\hat{\beta}_{INT.})$ and $mean(\hat{se}(\beta_{INT.}))$ [*]				**Coverage** [†]			
Selection Method	**n=250**	**n=1000**	**n=5000**	**n=20000**	**n=250**	**n=1000**	**n=5000**	**n=20000**
BMA	0.995	0.989	0.983	1.000	0.949	0.951	0.949	0.952
BMA Top	1.078	1.040	1.013	1.011	0.931	0.942	0.945	0.947
BMA Median	1.074	1.039	1.013	1.011	0.934	0.943	0.945	0.947
SSVS	0.990	0.998	0.994	1.004	0.952	0.949	0.947	0.949
SSVS Top	1.039	1.017	1.011	1.005	0.941	0.944	0.945	0.947
SSVS Median	1.034	1.018	1.011	1.005	0.942	0.944	0.945	0.947
Adaptive Lasso	NA	NA	NA	NA	NA	NA	NA	NA
BE BIC	1.081	1.043	1.013	1.011	0.932	0.942	0.945	0.947

[*] Ratio

[†] Coverage is the percent of replications whose 95% confidence interval for $\hat{\beta}_{INTEREST}$ includes the true value, 0.01. Coverage less than 0.95 suggest the interval is either too narrow or too biased. Since bias has been shown to be small, any problems with coverage are likely to be caused by an underestimate of the standard error of $\beta_{INTEREST}$.

3.5.4 Predictive Performance

A selection method with high $\rho^{Y_{external}, \hat{Y}_{external}}$. Ideal models also have a lower difference,ω, between this external correlation and the correlation between the observed and predicted outcomes in the original dataset, which was used to build the model. Figure 3.5 presents these values as n increases. Increasing n improves prediction in all methods, both in increasing correlation and decreasing ω. Adaptive lasso had the lowest ω values of all the methods for all n. For small studies, n < 500, adaptive lasso does quite well having one of the highest external correlations as well as the lowest ω. The SSVS average model had the most noticeable problem with over-fitting, while the SSVS top model and the SSVS median model performed much better. BE BIC only did slightly better than the SSVS average model in ω. By n = 1000, BMA, BMA top model and BMA median model all converge to having low ω with high external correlation and are joined by SSVS top and median models when n > 1000.

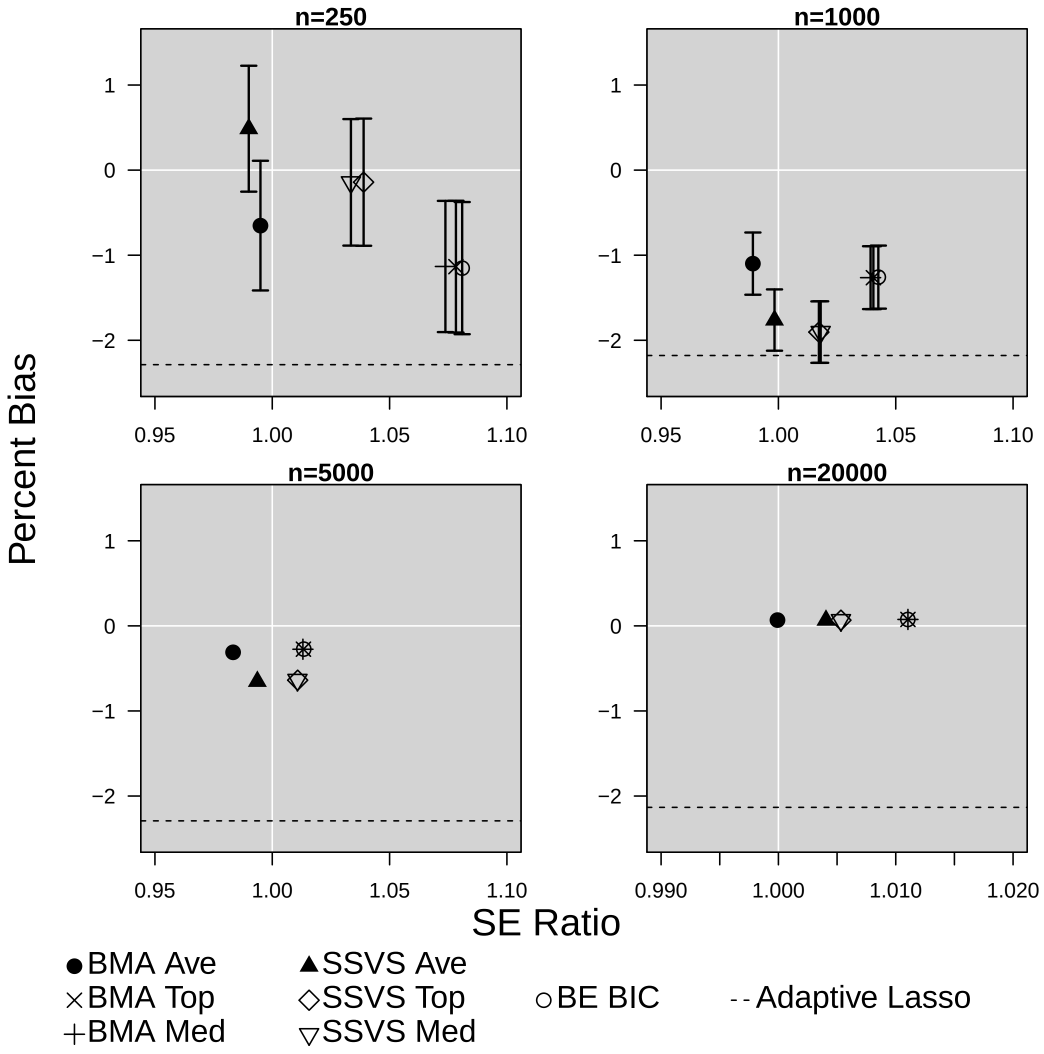

Figure 3.4: Percent Bias by SE Ratio.

Ideal estimates have bias of zero and an SE ratio close to one. This target is indicated by the crossing white lines. If the target is not achieved, conservative methods have negative bias and ratio less than one. Adaptive lasso is represented with a line because it did not have standard error estimates. All the methods have negative bias. BMA and SSVS average have SE ratios less than one for all n. Percent bias and SE ratio for all the methods except adaptive lasso become more alike as n increases. Bias intervals are quite small and not shown for n=5000 or 20000.

Table 3.4: Agreement.

Selection Method	n=250	n=1000	n=5000	n=20000
BMA Top and BMA Median	0.901	0.958	0.994	1.000
SSVS Top and SSVS Median	0.968	0.977	0.996	1.000
SSVS Top and BMA Top	0.371	0.404	0.661	0.986
SSVS Median and BMA Median	0.380	0.405	0.664	0.987
Adaptive Lasso and SSVS Top	0.591	0.151	0.027	0.053
Adaptive Lasso and SSVS Median	0.591	0.151	0.027	0.053
Adaptive Lasso and BMA Top	0.316	0.019	0.008	0.054
Adaptive Lasso and BMA Median	0.324	0.020	0.008	0.054
BE BIC and BMA Top	0.956	0.989	0.998	1.000
BE BIC and BMA Median	0.885	0.954	0.994	1.000

3.5.5 Agreement

Agreement between the methods is shown Table 3.4 as the percent of simulation replicates were the methods selected the same final model. Because they do not result in a final reduced model BMA and SSVS average models are excluded. The model from BMA and SSVS with the highest PMP (top) very often matched the model which included all variables with PIP¿0.5 (median). Recall the BMA method employed here uses a BIC approximation for (4.7) (Yeung et al. 2005, Raftery 1995), and indeed the models selected by BMA and BE BIC usually agree, with agreement being higher between BE BIC and BMA Top. BMA and SSVS do not see strong agreement until $n > 5000$. Adaptive lasso had shows its strongest agreement with SSVS when $n \leq 250$, roughly 60%.

3.6 Summary and Limitations

This study aims to illustrate the capabilities of modern variable selection methods applied to a non-high-dimensional setting of linear regression. A brief overview of the results is shown in Table 3.5. First, consider estimation as the modeling goal. Based on three measures, bias, underestimation of the standard error of $\hat{\beta}_{INTEREST}$, and coverage, BMA and SSVS are the only methods that succeed on all three at most n. Although no methods were technically unbiased for n=1000 and 5000, bias was quite small. Adaptive lasso is biased at all n and does not easily provide standard errors. Next, consider prediction the modeling goal. When n < 500 adaptive lasso is the clear choice, indeed this was the only bright spot for adaptive lasso in this study. Once n = 1000 adaptive lasso is overtaken by all three BMA models and SSVS top and median. SSVS average and BE BIC both have a high level of overfitting, ω, and should be avoided. Lastly, consider the goal of understanding the relationship between the variables by studying the selected model. All methods neglect to find the true model for smaller samples of n ≤ 1000. This emphasizes the difficulty of variable selection. It is not until n = 5000 that the true model is selected with any reliability, with BMA top and median models and BE BIC all finding it roughly 70% of the time. Even when n = 20000 adaptive lasso only selects the true model in 5.3% of the replications. This suggests the conditions for the attractive oracle property of adaptive lasso have not been met. At small n, the methods are missing variables with small effects and often miss the confounding variable. Finally as n increases to 20000, all methods except adaptive lasso settle on the truth, and are unbiased.

This simulation study does have limitations. It could be argued that criteria used to judge a good model in Table 3.5 are too stringent. For example, with bias rarely over 5% almost all the methods perform reasonably well. Modern variable selection methods have been applied to other regression models including logistic, Cox proportional hazards, and multinomial models. Linear regression was chosen to demonstrate the ”best-case scenario”. All the selection methods are expected to perform more poorly in more complex models.

Additional parameters of the simulation set-up could have been explored, such as a greater variety of effect sizes, strength of correlation, types of variables and number of variables under consideration. However, this study does glean valuable information about how these variable selection methods behave in a realistic setting and it does not seem that modifying the simulation parameters would drastically change the overall conclusions.

This study cannot claim to exhaustively include all of the many variable selection methods, both modern and classic, employed by researchers. Methods presented have existing packages in R allowing for relative ease of use. While SSVS and adaptive lasso in particular have a variety of user-controllable options, this study assumes naive use by keeping all defaults in place. Champions of SSVS and adaptive lasso could argue both of these methods may perform better in the hands of an expert practitioner than what is shown here. While many previous studies have focused on these methods' ability to discover the true model, this study acknowledges that recovering the correct model in practice not typically of the highest importance in a non-high-dimensional setting, rather effect estimation and outcome prediction are often more valuable.

3.7 Discussion

Modern variable selection techniques are seeing a rapidly increasing rate of use. Computing advances have allowed these methods to be applied to high-dimensional and sparse data settings. In particular, the lasso family of methods are extremely computationally efficient for these large data settings (Tibshirani 2011). As the excitement around modern methods builds, researchers will be more likely to apply modern methods to the classic regression setting, or in other words, to a non-high-dimensional setting where classic variable selection techniques have more commonly been employed. Classic variable selection techniques have known deficiencies in estimation, prediction and model selection. Modern methods make bold claims to solve some of these shortcomings. It is imperative to understand the behavior of these modern methods in a classic setting.

The bold claims made by modern variable selection deserve close examination. Both BMA and Bayesian variable selection claim to account for model uncertainty (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999, George and McCulloch 1993). In this study, BMA was found to be superior to SSVS in not underestimating the standard error or in over-fitting the predicted values. BMA also better accounted for uncertainty than the classic variable selection method, BE BIC. BMA claims to provide better average predictive ability than using any single model under the logarithmic scoring rule(Madigan and Raftery 1994). BMA was found to result in a high correlation between observed and predicted values in an external dataset with minimal amounts of over-fitting. However, BMA was outperformed by adaptive lasso when $n \leq 250$. Lasso was designed to increase prediction ability at the cost of bias. Increased prediction was only seen at the small sample sizes and bias did not go away as n increased. Adaptive lasso in particular claims to have the oracle property in some settings (Zou 2006). The oracle property was certainly not demonstrated in this study.

Beyond addressing the claims of modern variable selection method capabilities, the study provides a few unexpected results. SSVS resulted in a higher level of over-fitting than anticipated. SSVS over-fitting was more dramatic in the averaged model than in the SSVS top or median model. However, the SSVS average model did a better job of accounting for model uncertainty in estimating the coefficient of interest. When introduced, it was suggested that SSVS could serve as a first step for eliminating variables and could be followed by an additional step of SSVS or other variable selection method on the reduced set (George and McCulloch 1993). However, in this study SSVS did not select unnecessary variables. Although cautioned that SSVS and lasso will select only one of a group of collinear variables (Tibshirani 2011, George and McCulloch 1993), the difficulty all methods exhibited in finding $B_{11,EC}$ was surprising. Recall, $B_{11,EC}$ was a binary variable with a positive effect on the outcome and a negative correlation with the effect of interest. It was the only EC variable to have effects in opposing directions. While neglecting to include this confounder had seemingly small effects of prediction or estimation, failing to include a confounder can have serious consequence for model interpretation.

Of the modern methods included in this study, BMA performed best overall in modeling goals of estimation, prediction and variable selection. BMA also had fewer user-specified parameters to set, making it slightly easier for a non-expert to employ. In addition to being superior in estimation, prediction, and variable selection, BMA along with SSVS provide the researcher with helpful PIP without requiring a simulation. Often, the p-value is misinterpreted as the probability that a variable belongs in a model(Greenland et al. 2016). This is in fact the interpretation of PIP, while a p-value is the probability of observing a coefficient of this magnitude or greater when in fact there is no relationship between the outcome and predictor variable, assuming the correct model has been used. In spite of these appealing characteristics, BMA has yet to see wide-spread use in many areas of application. In public health, for example, a survey of papers in 2009 revealed no authors had chosen any modern variable selection methods (Walter and Tiemeier 2009). The ultimate aim of this article is to provide researchers with information to guide one of the most critical parts of their choice in data analysis. Just like every other analysis tool, no single variable selection method can be recommended for use in every study. By presenting direct comparisons of these varied modern variable selection methods in a simulation setting where the truth is both well-defined and complex enough to represent a real analysis, this study allows the researcher to judge their own study's traits and goals and select the method that fits best.

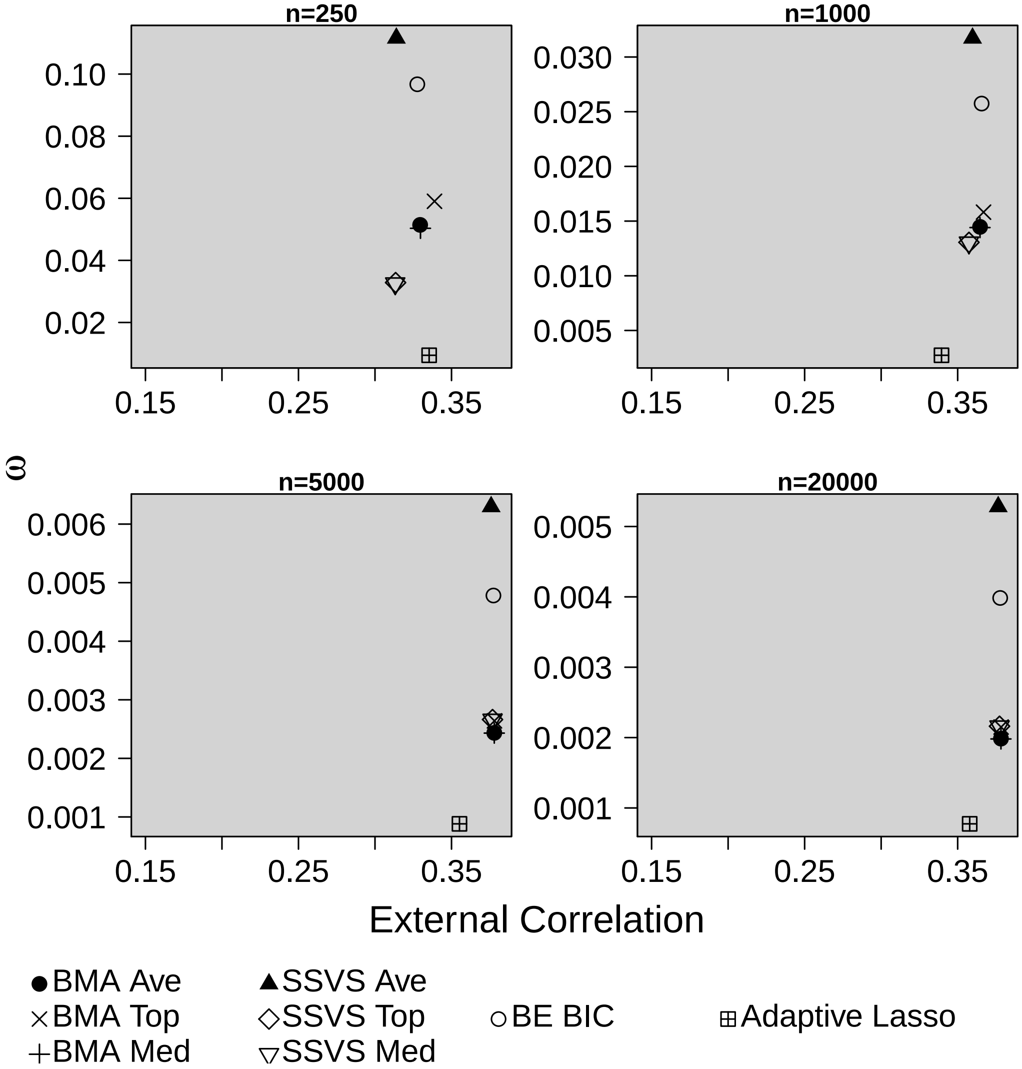

Figure 3.5: Predictive Performance: $\rho_{Y_{external},\hat{Y}_{external}}$ and ω.

Ideal models result in high correlation between the predicted and observed outcomes in an external dataset, which was not used to build the model. Ideal models also have a lower difference,ω, between this external correlation and the correlation between the observed and predicted outcomes in the original dataset, which was used to build the model.

Table 3.5: Summary.

	n=250	**n=1000**	**n=5000**	**n=20000**
Estimation				
No Underestimate	BMA	BMA	BMA	BMA
of $se(\hat{\beta})$				
Ratio ¡1	SSVS	SSVS	SSVS	
Unbiased				
MC Interval	BMA			BMA
Includes 0				BMA Med,Top
	SSVS			SSVS
	SSVS Med,Top			SSVS Med,Top
				BE BIC
Coverage				
MC Interval	BMA	BMA	BMA	BMA
Includes 0.95			BMA Med,Top	BMA Med,Top
	SSVS	SSVS	SSVS	SSVS
		SSVS Med,Top	SSVS Med,Top	SSVS Med,Top
			BE BIC	BE BIC
Predictive	BEST:	BEST:	BEST:	BEST:
Performance	Adaptive Lasso			
Highest $\rho_{Y_{external},\hat{Y}_{external}}$		BMA	BMA	BMA
and lowest ω		BMA Med,Top	BMA Med,Top	BMA Med,Top
		SSVS Med,Top	SSVS Med,Top	SSVS Med,Top
	AVOID:	AVOID:	AVOID:	AVOID:
		Adaptive Lasso	Adaptive Lasso	Adaptive Lasso
	SSVS	SSVS	SSVS	SSVS
	SSVS Med,Top			
	BE BIC	BE BIC	BE BIC	BE BIC
Variable Selection				
Selected True Model			BE BIC (69.6%)	BE BIC (98.5%)
			BMA Med,Top (69.6%)	BMA Med,Top (98.5%)
				SSVS Med,Top(>99%)

CHAPTER 4: A COMPARISON OF TRADITIONAL VARIABLE SELECTION METHODS WITH BAYESIAN MODEL AVERAGING IN LOGISTIC REGRESSION MODELS

4.1 Introduction

Public health researchers use statistical models to estimate effect sizes, predict outcomes and to understand underlying relationships between variables. A good data analyst is taught there is no substitute for subject area expertise, and the ideal situation is to rely on the body of knowledge to guide model building. However, the ideal is not always attainable. Perhaps a small sample size hinders the complexity of the desired model, or a new scientific area is being explored that does not yet have the benefit of pre-existing knowledge. Data-driven variable selection techniques have existed for at least fifty years. These variable selection techniques have been a popular method for public health researchers and many varieties and strategies exist. A fair number of warnings about the dangers of relying on these popular methods also exist. The lack of accounting for uncertainty that stems from selecting a single model among many possibilities lies at the root of many of these complaints. As data collecting and storage capabilities advance, public health researchers are likely to become more reliant on data-driven techniques to build statistical models. This paper compares the most popular data-driven variable selection techniques to a variable selection method that was designed specifically to account for model uncertainty, called Bayesian model averaging. The models' ability to estimate effects, predict outcomes and accurately include important variables is investigated in a realistic logistic regression setting, with correlated variables and a single variable of interest which is always included.

Data-driven variable selection is frequently used in public health research. In a survey of the 300 articles presented in *American Journal of Epidemiology*, *Epidemiology*, *European Journal of Epidemiology*, and *International Journal of Epidemiology* in 2008, 37.3% reported using a method of variable selection beyond relying solely on prior knowledge (Walter and Tiemeier 2009). To put 37.3% in perspective, consider a separate study of commonly used statistical methods in published public health research journals which included the above four in addition to *American Journal of Public Health*, *Bulletin of World Health Organization*, and *American Journal of Preventive Medicine*. Hayat et al. (2017) reports that 25.9% of studies report a Chi-squared test or a Fishers exact test, 38.4% report using logistic regression, and 40.7% report an odds ratio (Hayat et al. 2017). Variable selection methods are reported more often in public health research than linear models and Cox proportional hazards models combined, 19.4% and 15.3% respectively (Hayat et al. 2017). While many articles have been published using results from a variable selection method, these methods are not without potentially alarming flaws.

All varieties of variable selection methods have been criticized. Sequential methods, and therefore results from sequential methods, rely heavily on the cut-off chosen. Various recommendations when using a p-value cut-off exist and depend on the desired balance between including important variables and including noise (Harrell 2001). For example, a high p-value cut-off will be less likely to exclude important variables but more likely to include noise variables. Other criterion such as Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC) are often used and were developed with different modeling goals. Theoretically, BIC is a consistent model selector, meaning that if the true model is among the models considered, as n increases, the probability that BIC will identify the true model converges to 1 (Yang 2005). Unlike BIC, AIC does not possess the property of consistency. Variable selection with AIC is asymptotically equivalent to model selection with cross-validation (Stone 1977). The AIC criterion usually yields minimax-rate optimal estimators of the regression function under a squared-error loss(Yang 2005). The probability of selecting the true model using the AIC criterion converges to 1 as the number of variables considered increases. In a setting with many more variables considered, AIC may surpass BIC in selection ability. A good deal of discussion exists about which criteria AIC or BIC will select the best model (Burnham and Anderson 2004). The traits of change-in-effect methods are also highly dependent on the cut-point selected (Maldonado and Greenland 1993). Beyond the challenge of selecting an appropriate cut-off, any variable selection method which examines several models only to select a single model fails to account for model uncertainty. For this reason, sequential methods, such as stepwise regression, are known to provide R^2 which are deceptively high, estimates which are more impressive than they should be, and p-values which are too small (Harrell 2001). Model averaging was developed to directly combat this understatement of variability.

Even with these flaws, variable selection is seen as a useful tool in approaching a large set of variables to build a concise model both when the researcher cannot reasonably be expected to identify all relevant variables and when variables are collinear (Harrell 2001). Even though these short-comings were well known as long ago as the 1980s (Miller 1984, Freedman and Freedman 1983, Flack and Chang 1987, Freedman et al. 1988), use of sequential selection methods is still common (Walter and Tiemeier 2009). Further, because data collection, storage and analysis techniques have advanced, public health researchers will only be faced with more variables to choose from and will need to rely on "automated" methods more. Instead of abandoning such a frequently used technique, we aim to finesse its use by addressing problems of model uncertainty. Model uncertainty can be appropriately represented if estimates from every model considered are somehow accounted for (Buckland et al. 1997). One modern and simple way of addressing model uncertainty is through Bayesian model averaging (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999).

This study compares sequential methods, change-in-effect methods and BMA. Others have previously used simulation to compare some of the above methods in a variety of settings (Viallefont et al. 2001, Genell et al. 2010, Wang et al. 2004, Wiegand 2010, Mickey and Greenland 1989). Change-in-effect (CIE) has only been compared to the two-phase inclusion method of sequential selection(Mickey and Greenland 1989, Maldonado and Greenland 1993). Bayesian model averaging (BMA) is compared to a few sequential methods (i.e. backward elimination, stepwise, forward inclusion) in three papers, but in two of these, only variable inclusion probabilities and p-values are reported with no indication of how the models performed at effect estimation or outcome prediction(Viallefont et al. 2001, Genell et al. 2010). In the third, none of the explanatory variables are correlated and the simulation includes only 10 replications (Wang et al. 2004). The study we present is the first not only to compare all these methods simultaneously, but also to assess performance in all three goals of model building: estimation, prediction and variable inclusion. This comparison is made in the common setting of logistic regression where there is a single variable of interest, with correlated explanatory variables acting as possible confounders. Before exploring the criticisms as well as the merits of these methods more carefully via simulation, a brief background and description of the variable selection methods compared in this study are presented in Section 2. Section 3 presents the motivating example for this simulation study. The simulation design follows in Section 4 with simulation results described in Section 5. Section 6 summarizes the detailed results presented in Section 4. Lastly, summary conclusions are made in the final section.

4.2 Background and Review of Commonly Used Variable Selection Methods

The application of variable selection in a logistic models is the focus of the what is presented here. Logistic models are devised to investigate the relationship between the binary response, Y_i and the p possible explanatory variables $X_{i0}, \ldots, X_{ip}$. Assume there are n

subjects, $i = 1 \ldots, n$. A logistic model is of the form

$$log\frac{P(Y_i = 1)}{P(Y_i = 0)} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} \tag{4.1}$$

Note, $X_{i1}, \ldots X_{ip}$, could also be nonlinear functions. The predicted values of $P(Y_i = 1)$ are denoted and are defined as

$$\hat{P}(Y_i = 1) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_p X_{ip})}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_p X_{ip})} \tag{4.2}$$

where $\hat{\beta}_j$ are the estimates.

4.2.1 Sequential Testing Techniques: Forward, Backward, and Stepwise

Sequential testing techniques in variable selection are data-driven methods of developing a model. Roughly 20% of public health articles use some form of sequential variable selection techniques(Walter and Tiemeier 2009). As the name suggests, this method consists of steps beginning with either the full model, one which includes all possible variables and steps backward to a concise reduced model, or of steps beginning with a null model, one which only includes the most essential variables, perhaps only the intercept term, and steps forward to an adequate model. At each step, a model criterion is assessed to determine when the steps are complete. There are a variety of established criterion including the Akaike information criterion (AIC), Bayesian information criterion (BIC), as well as individual variable p-values. Additionally, forward selection and backward elimination (BE) can be combined in an algorithm that can move in either direction called stepwise.

Backward elimination is particularly useful in a study suffering from collinearity. Further, when two variables only jointly effect the outcome and have no effect individually, the forward selection procedure will exclude them, while the backward procedure will retain them (Mantel 1970). If two variables are equally explicatory of the outcome and are correlated with each other, then included together neither may appear significant. However, backward elimination can only remove one at a time, not both, and the stronger one would be retained in the model (Mantel 1970). Because public health data often includes collinear variables, this study focuses on BE with selection criteria of AIC, BIC, and p=0.05. Another common criteria is p=0.20, however the AIC criteria is equivalent to using a p-value of 0.157, and for the sake of brevity results using p=0.20 are shown in the appendix.

Another form of forward selection consists of pre-screening explanatory variables in models which include each variable alone. Then, only variables that are determined to be individually important are included in a multivariable model. Note, for the two-stage method to function adequately, a higher p-value cut-off, such as 0.2, should be considered (Mickey and Greenland 1989). This two-stage method has been shown to grossly underestimate the final p-value and will miss confounders which may be insignificant alone, but important to include in the model (Sun et al. 1996) (Viallefont et al. 2001).

4.2.2 Change-in-Effect Methods, CIE

Methods that eliminate variables based on whether their coefficients are different from zero do not address confounding (Kleinbaum et al. 1998). Consider a very simple situation with only three variables. It is of interest to know whether E is related to Y and there is a possible confounder, C. Two models can be estimated, one called crude and the other adjusted.

$$log\frac{P(Y=1)}{P(Y=0)} = \beta_{0,crude} + \beta_{1,crude}E \tag{4.3}$$

$$log\frac{P(Y=1)}{P(Y=0)} = \beta_{0,adj} + \beta_{1,adj}E + \beta_2 C \tag{4.4}$$

Confounding is present when $\beta_{1,crude} \neq \beta_{1,adj}$. However, this inequivalence is a subjective decision and not meant to be based on a statistical test (Kleinbaum et al. 1998; 1982). Statistical tests and other non-subjective criteria applied to this assessment of confounding proved too tempting in practice and many rules of thumb exist. Mickey investigates using percent change cut-offs of 5, 10, 15, 20, or 25% in simulation studies, ultimately suggesting 10% (Mickey and Greenland 1989).

CIE (also called change-in-estimate) is a procedure that begins with a full model. Models with one variable removed are fit for each variable and the change in the variable of interest is recorded. The variable that changed the variable of interest the least is dropped and the procedure is repeated until dropping any variable results in a percent change greater than some threshold, commonly 10%. A data-dependent threshold which depends on both β_1 and correlation between E and C has been suggested (Lee 2014). Roughly 15% of public health research articles employ change-in-effect methods of variable selection (Walter and Tiemeier 2009).

4.2.3 Bayesian Model Averaging

Regardless of the method used to select a best model, most researchers will acknowledge that it is difficult to ascertain whether the correct model has in fact been chosen, but will proceed with inference and prediction as if the model they use is indeed correct. Bayesian model averaging (BMA) provides an opportunity for exploring many possible models while appropriately accounting for the uncertainty surrounding variable selection (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). BMA is only one of several statistical methods developed for appropriately accounting for model uncertainty but has the advantage of providing the best predictive qualities (George 2000, Madigan and Raftery 1994). However, given their reported lack of use, BMA methods are either unknown or not easily accessed by many researchers (Walter and Tiemeier 2009).

Bayesian model averaging takes a weighted average of all the models examined. Consider the setting where a researcher has a proposed (finite) class of models M. Suppose that Δ is a quantity of interest, such as a predicted value of a future observation or an effect estimate, and that the class of proposed models M contains elements $M_1, \dots, M_K$. Then the posterior distribution given data D is defined by the following weighted average

$$p(\Delta \mid D) = \sum_{k=1}^{K} p(\Delta \mid M_k, D)\, p(M_k \mid D). \tag{4.5}$$

If the model building goal includes understanding the relationships between variables, it can be helpful to know the probability that each model is correct. It is convention to refer to $p(M_k \mid D)$, or the probability model M_k is correct given the observed data, as the Posterior Model Probability (PMP). PMP is used as the weight in the weighted average.

$$p(M_k \mid D) = \frac{p(D \mid M_k)\, p(M_k)}{\sum_{h=1}^{K} p(D \mid M_h)\, p(M_h)} \tag{4.6}$$

where

$$p(D \mid M_k) = \int p(D \mid \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k \mid M_k) d\boldsymbol{\theta}_k \tag{4.7}$$

for $\boldsymbol{\theta}$ a vector of parameters of model M_k. In the case of the logistic model, $\boldsymbol{\theta}$ would be a vector of $\boldsymbol{\beta}$. BMA constructs a weighted-average model by using PMP as the weights. Another very useful posterior probability from this process is the posterior inclusion probability, or PIP. PIP is the probability of a variable appearing the true model and can be found by summing the PMP's of all models which include the variable. PIP can also be thought of as $P(\beta_j \neq 0|D)$.

BMA requires the user to define $p(M_j)$, also called the prior model probability. This allows the researcher to place heavier weight on models which are deemed more likely. If there is not information available to inform this decision, each model can be assumed to be equivalently likely, or each variable could be assumed to have a 50% chance of being included. This study applies the BIC approximation presented in Raftery for the integral in (4.7) (Yeung et al. 2005, Raftery 1995) . In settings with many variables it is computationally prohibitive to explore every possible model. Instead, the model space can be explored with methods beyond enumerating every possible model. One of these methods is using the Occam's window approach (Madigan and Raftery 1994, Raftery 1995, Hoeting et al. 1999). In this approach, models are excluded from consideration if they are significantly less likely and/or contain sub-models which are dramatically more likely. See hoeting1999bayesian for a thorough history of BMA .

4.3 Motivating Example

The China Health and Nutrition Survey (CHNS) collected health data in 228 communities (nine diverse provinces: Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning and Shandong) throughout China in ten survey rounds from 1989 to 2015. Using a multistage, random cluster design, a stratified probability sample was used to select counties and cities stratified by income and urbanization. Communities and households were then randomly selected from these strata. Survey procedures have been described elsewhere (Popkin 2010). The study was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill, the China-Japan Friendship Hospital, the Ministry of Health and China, and the Institute of Nutrition and Food Safety, China Centers for Disease Control. Participants gave informed consent.

One modeling challenge for investigations using CHNS is the relationships between variables. For example, consider including alcohol consumption as a variable. CHNS contains as either survey responses or derived variables the frequency of alcohol consumption, number of bottles of beer per week, whether beer was consumed in the last week, whether the person drinks alcohol, whether liquor was consumed in the last year, three day average of alcohol by volume, a three day average of alcohol by weight, amount of wine consumed in the last week, whether the person drinks wine, whether the person drinks liquor, and possibly others. CHNS also records very detailed information about a participant's diet, health, socioeconomic situation, physical activity and community which leaves the public health researcher with thousands of variables available for study. Given the related nature of variables available, it is not hard to develop a model which suffers from collinearity, over-parameterization, or overlooking a useful variable.

For the purposes of demonstrating the traits of variable selection methods, this paper considers the specific goal of modeling waist-to-height ratio among participants who were surveyed in at least two rounds while they were between the ages of 18 and 30 and have their waist circumference and height recorded. After excluding all missing data, N=1195. The specific variable of interest is a measure of urbanization and potential explanatory variables include the following: age, sex, income, years of education, physical activity, caloric intake, sodium, potassium, whether the participant smokes, whether the participant consumes alcohol, and whether the participant drinks coffee. Coffee drinking was included because it was highly correlated to urbanization, but was not believed to be related to waist-to-height ratios. The simulation study generates data meant to mimic this example. All explanatory variables were standardized. The correlation between the explanatory variables, the effect sizes, and the error variance found in the example data guide the simulation parameters.

4.4 Design of the Simulation Study

All simulations were performed in R version 3.4.3 and all results presented are based on 5000 replications(R Core Team 2013). Note, 5000 replications results in Monte Carlo Error less than 0.7% for a binary result, such as the probability of including a variable (Koehler et al. 2009).

4.4.1 Data Generation

Table 4.1 shows the correlations between the explanatory variables. Explanatory variables are labeled with an X if continuous and B if binary. $X_{INTEREST}$ represents the variable of interest. Among the remaining explanatory variables, a subscript of E represents a variable with a true non-zero effect, and C represents a variable which is correlated with the variable of interest. Four variables are constructed which had both an effect on the outcome as well as a correlation with the variable of interest ($X_{1,EC}$, $X_{2,EC}$, $X_{7,EC}$, $B_{11,EC}$). Of these four, two have correlations in the opposite direction of the effect ($X_{7,EC}$, $B_{11,EC}$). Also included are four variables completely unrelated to the others. Two of these are binary (B_{12}, $B_{13,E}$), two are normally distributed (X_{14}, $X_{15,E}$). Variables with no letter subscript, such as X_5 and X_6, are noise variables with no direct effect on the outcome and no correlation with the variable of interest. The correlation matrix shown in Table 4.1, with the additional rows for the uncorrelated variables, is used to generate data from a multivariate normal of size $2n$ with all means equal to zero. The first n observations are used to build the model and the second n are treated as an external sample for assessing predictive ability of the models.

Table 4.1: Explanatory Variable Correlations for Data Generation.

	$X_{INTEREST}$	$X_{1,EC}$	$X_{2,C}$	$X_{3,C}$	$X_{4,C}$	X_5	X_6	$X_{7,C}$	B_8	$B_{9,E}$	B_{10}	$B_{11,EC}$
$X_{INTEREST}$	1	0.1	0.2	-0.5	-0.2			0.5				0.2
$X_{1,EC}$	0.1	1		0.2	-0.1						0.2	
$X_{2,C}$	0.2		1	-0.1		-0.1		0.2				0.1
$X_{3,C}$	-0.5	0.2	-0.1	1	0.1			-0.4				-0.1
$X_{4,C}$	-0.2	-0.1		0.1	1	0.4	0.6	-0.1	0.2	0.2	-0.3	-0.1
X_5			-0.1		0.4	1	0.3					
X_6					0.6	0.3	1		0.1	0.1	-0.1	
$X_{7,C}$	0.5		0.2	-0.4	-0.1			1		0.1	-0.1	0.2
$B_8{}^*$					0.2		0.1		1	0.5	-0.5	
$B_{9,E}{}^*$					0.2		0.1	0.1	0.5	1	-0.4	
$B_{10}{}^*$		0.2			-0.3		-0.1	-0.1	-0.5	-0.4	1	0.1
$B_{11,EC}{}^*$	0.2		0.1	-0.1	-0.1			0.2			0.1	1

Blanks indicate no correlation.

[*] Some of the variables are transformed from normally distributed variables into binary variables with a cut point determined at a random quantile. For example, it is desired that on average 25% of observations have B_8 = 1. At each replication of the simulation a cut point varied uniformly between 20% and 30% to achieve an average of 25%. Desired observed proportions for B_8 through B_{13} are 0.25, 0.25, 0.5,0.05, 0.7, and 0.3 respectively.

Four variables, B_{12}, $B_{13,E}$, X_{14}, $X_{15,E}$ are uncorrelated with all other variables and are not included here.

Lastly, the outcome variable is generated in the following manner:

$$Y \sim Binomial(P(Waist - to - Height\ Ratio\ > 0.5)) \tag{4.8}$$

$$log\frac{P(Waist - to - Height\ Ratio > 0.5)}{1 - P(Waist - to - Height\ Ratio > 0.5)} = -1.4 + 0.3X_{INTEREST} + 0.4X_{1,EC}$$
$$+ 0.2X_{2,EC} - 0.25X_{7,EC} - 0.75B_{9,E} - 0.3B_{10,E} - 1.1B_{11,EC} + 0.2B_{13,E} - 0.2X_{15,E} \tag{4.9}$$

Sample sizes of $n = 150, 250, 500, 1000, 2500, 5000, 10000$, and 20000 are investigated. The coefficients in (4.9) result in approximately 18.5% of the sample having $Y = 1$. Also for the setting of $n = 1000$, the situation where the variable of interest had no effect was considered. Although this data is more complicated than is generally reported in simulations comparing variable selection methods, it is a more accurate depiction of the conditions of a large observational study (Wang et al. 2004, Wiegand 2010, Lee 2014).

4.4.2 Variable Selection

After the data are generated, backward elimination (BE), two-stage analysis, change-in-effect procedures (CIE), and Bayesian model averaging (BMA), are performed using the first n observations. BE, two-stage, and CIE procedures yield final models and the variables included in these final models, coefficient estimates, standard error estimates and predicted values of the outcome for these selected models are recorded. BMA does not reduce the number of variables but instead reports the average estimates for all variables. In practice a researcher may still want to report a final model after using averaging techniques. This study considers the model with the highest PMP (top model), and the model resulting from including all variables with PIP¿0.5 (median model) as likely candidates for selecting a reduced model and are therefore also recorded. BMA provides averaged estimates of coefficients, standard errors and predicted values of the outcome. PMP and PIP values for each replication are recorded. Lastly, the final model or the averaged model from each method is applied to the last n of the $2n$ observations and predicted outcomes, $\hat{P}(Y_i = 1)$, are recorded.

Three BE strategies are considered, one based on a p-value cut-offs of 0.05, one on minimizing AIC, and another on minimizing BIC. The p-value cut-off BE is performed with the `fastbw` function in the **`rms`** package in R (Harrell Jr 2018). BE to minimize AIC is performed using the `stepAIC` function in the **`MASS`** package (Venables and Ripley 2002) and BE to minimize BIC is performed using the `step` function. The functions `fastbw`,`stepAIC` and `step` are only used to determine the final model. Estimates of the coefficients are made using the variables identified by these functions in a logistic regression model fit with `glm`. Two-stage analysis did not use an existing function. A cut-off of p=0.05 is used. CIE is not available in a mainstream R package and a function was created to perform this selection strategy which is shown in the appendix. Thresholds of 5%, 10%, and 20% were applied. CIE with 5% performed best of these three and is presented here, while results for 10% and 20% can be found in the appendix. Results from a p-value cut-off of 0.05 are shown here, and the p-value cut-off of 0.2 is shown in the appendix. Note, when variables have only one degree of freedom as the ones in this simulation study, the AIC cut-off is equivalent to the p-value cut-off of 0.157 (Steyerberg et al. 2000).

BMA uses the `bic.glm` function in the **`BMA`** package (Raftery et al. 2017). Occam's window with a maximum exclusion ratio of 20 as suggested in Madigan and Raftery is used(Madigan and Raftery 1994). Prior probabilities for variable inclusion are 1 for the variable of interest, guaranteeing inclusion, and 0.5 for all remaining variables. A prior probability of inclusion of 0.5 indicates that, prior to collecting data, a researcher believes a variable is related to the outcome just as much as they believe the same variable is not related to the outcome. This is sometimes called an indifference prior. In order to most most closely mimic performance of these methods in public health applications, the default settings are used.

4.4.3 Quantities of Interest

Model Selection and Variable Inclusion Probabilities

The first of four key modeling aims this simulation will investigate is selecting the correct model. The model selected by each procedure is recorded, and the probability of selecting any particular model is the number of replications selecting a specific model and dividing by r, here 1000. Because so many variables are examined, the variety of models chosen by the selection methods considered here are well over 1000. For this reason, the models selected are summarized by determining if they select all, some or none of the variables with $\beta \neq 0$ ($X_{1,EC}$, $X_{2,EC}$,$X_{7,EC}$,$B_{9,E}$, $B_{10,E}$,$B_{11,EC}$, $B_{13,E}$, and $X_{15,E}$),and all, some, or none of the variables related to $X_{INTEREST}$ but with $\beta = 0$ ($X_{3,C}$ and $X_{4,C}$). These two groups are referred to as the E group, for effect, and the C group, for correlated. Ideally the methods will select the true model, but if that cannot be achieved, the model should at least include all of the E group.

Additionally, the variable inclusion probabilities (IP) are presented. For all methods except BMA, IP is the percent of simulation replicates selecting each variable in the final model. Average PIP across replications is reported for BMA. The individual variable's IP are essential for understanding how these selection methods function with variables that differ in effect size, correlation, and type (binary vs. continuous). If the true model is not selected, these IP can help determine why.

Estimation of Main Effect

The second key modeling aim investigated is the estimation of $X_{INTEREST}$. If a researcher is mostly interested in the size of the association between the variable of interest and the outcome, it matters less which specific variables are included, and instead the accuracy and precision of $\hat{\beta}_{INTEREST}$ matter much more. Much like in practice, this study considers one variable to be the variable of interest and requires all the models a variable selection method considers to include this variable. The bias, standard error of the coefficient estimates, ratio of the standard deviation of the coefficient estimates and the mean estimated standard error, coverage, and the probability of type I error or power, depending on the scenario, are recorded. Percent bias is calculated as $100 * \frac{\hat{\beta}_{INTEREST}-0.3}{0.3}$. The ratio of the standard deviation of the coefficients and the mean estimated standard deviation is a measure of how well methods account for model uncertainty. If this ratio is greater than 1, the reported standard deviations are too small which in turn results in overly optimistic p-values.

Predictive Performance

The third quantity addresses how well each model predicts the outcome. Prediction is measured by Somers' D rank correlation of $\hat{P}(Y = 1)$ and the actual values of Y in the simulated original and external datasets. The difference between these two correlations reveals a degree of over-fitting (Harrell 2001).

$$over - fit = D_{Y_{orig},\hat{P}(Y=1)_{orig}} - D_{Y_{external},\hat{P}(Y=1)_{external}} \tag{4.10}$$

The ideal model has the highest $D_{Y_{external},\hat{P}(Y=1)_{external}}$ and the lowest over-fit. Note Somers' D ranges from -1 to 1 where 1 indicates the highest concordance between the Y and $\hat{P}(Y = 1)$. Also, $\frac{D+1}{2}$ is the area under the curve, or AUC.

Agreement

Agreement is not really a modeling goal but aids our understanding of how these methods relate to each other. The fourth and final quantity investigated is the probability that these variable selection strategies agree with each other. With so many variable selection methods available, one suggested technique is to perform multiple methods, declaring a model final if it is chosen by both methods (Wiegand 2010). Investigating method agreement will also highlight the similarities between the selection methods.

4.5 Simulation Results

4.5.1 Model Probabilities

Variable selection methods are often used with the expectation that they can uncover the true model. Figure 4.1 shows the probability of selecting the true model by n. Reliable selection of the true model does not occur for any method until $n \geq 5000$. At $n = 5000$, BE p=0.05 selects the true model most frequently with 45% of the replications finding the true model. Once $n \geq 10000$, BE p=0.05 is no longer the top performer with BE BIC and BMA more frequently finding the true model. CIE and the Two-Stage method very rarely select the true model, even when $n = 20000$. Investigating specific variable IP helps explain why the methods fall short of selecting the true model.

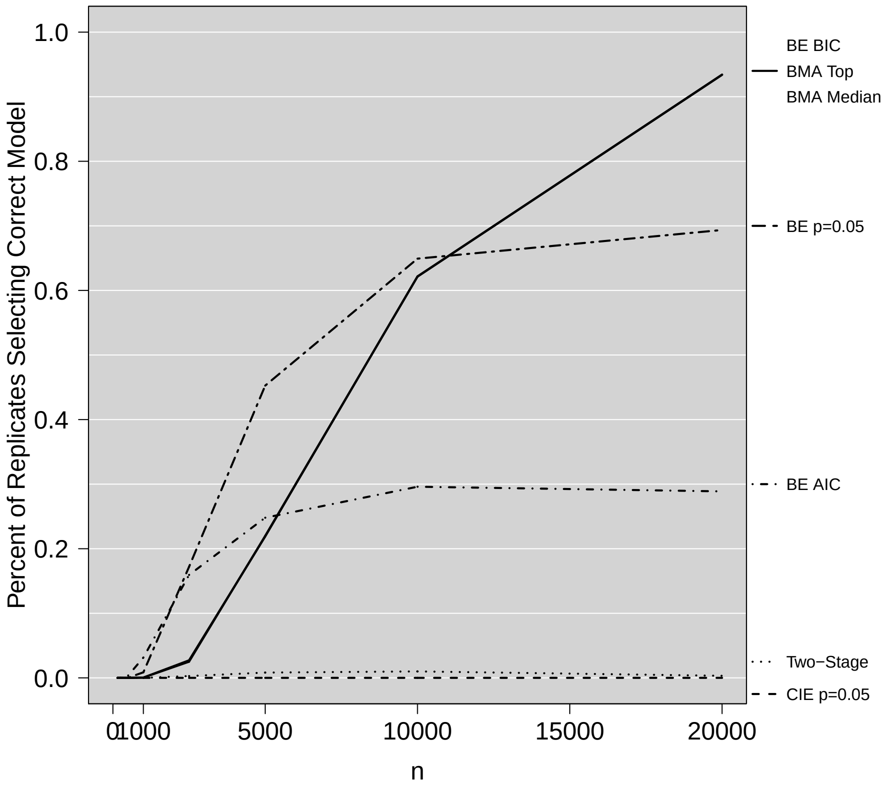

Figure 4.1: Probability of Selecting the True Model.

Both CIE and Two-Stage variable selection methods fail dramatically at selecting the correct model.

4.5.2 Variable Probabilities

Variable inclusion probabilities (IP) are helpful for investigating how variable selection models operate. In practice, unless using BMA, sequential variable selection methods and CIE do not report an IP but instead report a final model. Figure 4.2 shows the IP for $n = 250,\ 1000,\ 5000$, and 20000. Variables appearing at the top of the plot have both an effect and a correlation with the variable of interest and have EC as a subscript, next appear the variables with only an effect, then those with only a correlation, and lastly appear a few of the noise variables. Other noise variables are not shown in the plot. The noise variables shown here had no correlation with the variable of interest and no correlation with any other variables considered. Only IP from the BMA average model are shown here, not the IP from BMA top and median models. The top and median models closely match their corresponding averages for IP.

First, consider CIE. CIE most stands out from the other methods for variables which are correlated with the variable of interest. CIE erroneously includes $X_{4,C}$ and $X_{3,C}$ very often when $n \leq 1000$ although this decreases with n. However, as n increases CIE fails to include variables with an effect and no correlation with $X_{INTEREST}$. This suggests CIE is very sensitive to the correlation of variables with the variable of interest, but is not really finding variables with a relationship to the outcome.

Next, consider the Two-Stage method. The Two-Stage methods has IP similar to that of the BE methods and BMA with two striking exceptions: $B_{10,E}$ and $X_{4,C}$. The Two-Stage method is failing to include $B_{10,E}$. Recall the correlation matrix of the variables shown in Table 4.1. $B_{10,E}$ has some of the strongest correlations present with B_8 and $B_{9,E}$. $B_{9,E}$ has a much larger effect size than $B_{10,E}$. The other variable in Table 4.1 with strong correlations is $X_{4,C}$. However, it has strong correlations with X_5 and X_6, which have no effect. Unlike $B_{10,E}$, Two-Stage methods are including $X_{4,C}$ more than it should. In this way, the Two-Stage method is quite susceptible to collinearity not only with the variable of interest, but with all the variables.

BE methods perform similarly to each other, in that, as n increases, they tend to include variables with an effect and exclude those without. BE AIC does include variables with an effect at a higher rate than BE BIC and BE p=0.05, however, BE AIC also includes noise variables at a higher rate. The rate of inclusion of noise variables for BE p=0.05 and BE AIC does not change with n. These constant rates of including noise of 5% and 15.7% explain why Figure 4.1 displays a striking leveling off of the probability of BE AIC and BE p=0.05 selecting the true model. BE BIC does not have a constant error rate and is therefore able to improve as n increases. Lastly, BMA selects variables very similarly to BE BIC, with inclusion of noise decreasing with n, inclusion of variables with an effect increasing with n. Ignoring CIE, all the methods had a harder time including variables having a correlation with the variable of interest in the opposite direction of the effect like $B_{11,EC}$ and $X_{7,EC}$. The methods also had an easier time identifying continuous rather than binary variables.

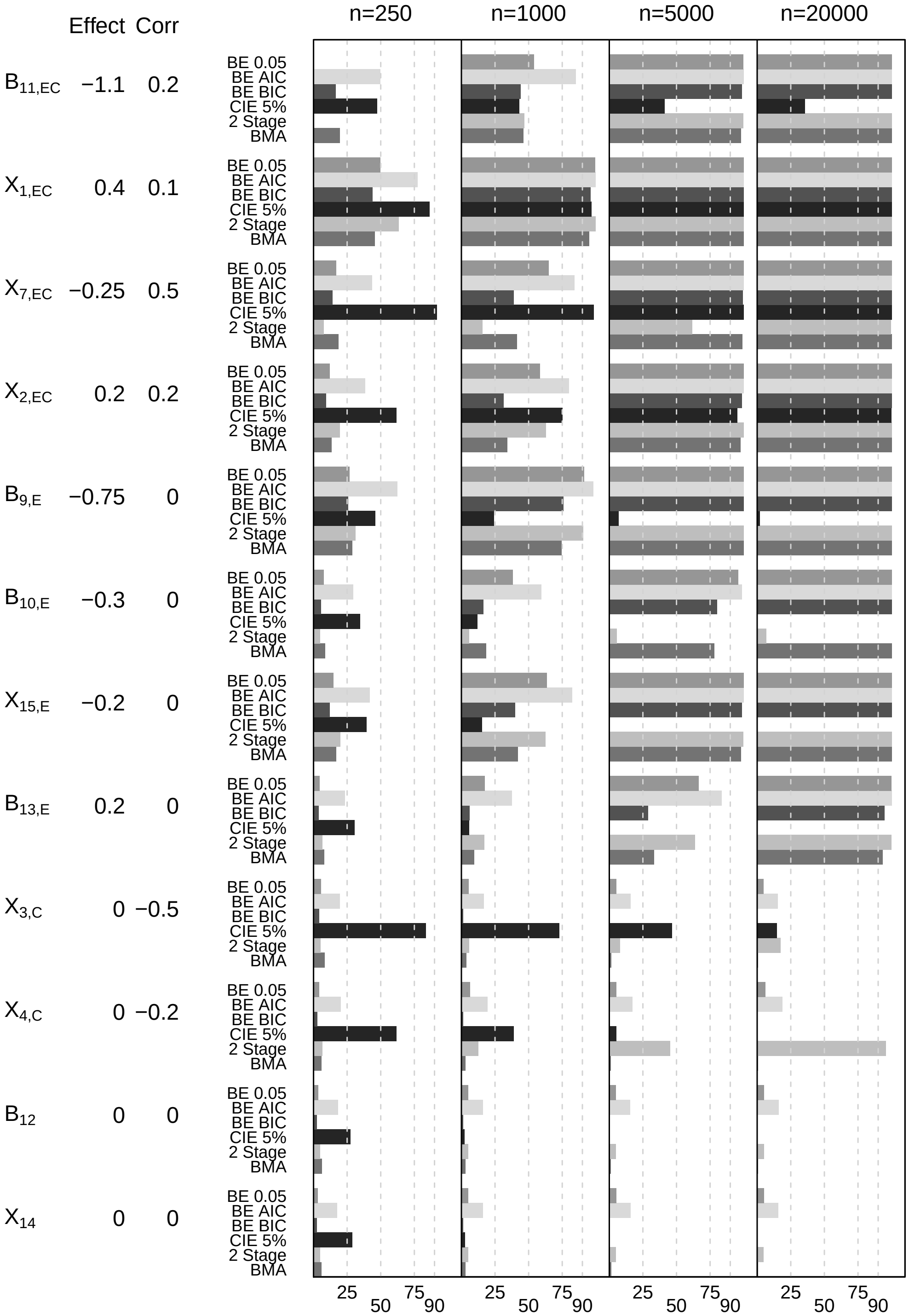

Figure 4.2: Variable Inclusion Probabilities.
From top to bottom, variables have both an effect and a correlation with the variable of interest (EC subscript), an effect but no correlation (E subscript), a correlation but no effect (C subscript), and neither an effect nor a correlation (noise, no subscript).

4.5.3 Estimation of the Main Effect

The estimates of percent bias in $\beta_{INTEREST}$ by n are shown in Table 4.2 along with the 95% Monte Carlo confidence intervals. Unlike any of the other methods, CIE has an odd property of increasing bias as n increases. Most of the methods have some bias when $n \leq 5000$, but this bias tends to shrink and the Monte Carlo interval includes zero. However, the Two-Stage model and CIE 5% models both show bias even when $n = 20000$. The Two-Stage methods shows an alarming amount of bias for $n < 20000$, particularly when compared to the other methods. Also, the averaged coefficients from BMA are less biased than the coefficients from the BMA top or median model.

Table 4.2: Estimation of Variable of Interest: Percent Bias.

	Percent Bias % *			
Selection Method	**N=250**	**N=1000**	**N=5000**	**N=20000**
BE p=0.05	-10.932	-5.785	0.550	0.050
	(-12.967,-8.897)	(-6.863,-4.706)	(0.121,0.98)	(-0.166,0.266)
BE AIC	6.259	-0.590	0.625	0.083
	(3.993,8.525)	(-1.633,0.452)	(0.180,1.070)	(-0.141,0.308)
BE BIC	-9.912	-14.069	0.286	0.055
	(-11.922,-7.902)	(-15.179,-12.960)	(-0.138,0.711)	(-0.154,0.265)
CIE 5%	12.689	0.785	-0.225	-1.098
	(10.356,15.021)	(-0.250,1.819)	(-0.672,0.222)	(-1.313,-0.883)
Two-Stage	-28.150	-30.569	-12.084	1.537
	(-29.740,-26.560)	(-31.451,-29.688)	(-12.685,-11.482)	(1.315,1.759)
BMA Average	-7.766	-13.315	0.209	0.053
	(-9.667,-5.865)	(-14.346,-12.284)	(-0.217,0.635)	(-0.157,0.262)
BMA Top	-9.991	-14.436	0.286	0.055
	(-11.995,-7.987)	(-15.542,-13.33)	(-0.138,0.711)	(-0.154,0.265)
BMA Median	-10.998	-14.404	0.279	0.056
	(-12.998,-8.999)	(-15.52,-13.287)	(-0.145,0.703)	(-0.153,0.265)

* Percent Bias is calculated as $100 * \frac{\hat{\beta}_{INTEREST} - 0.3}{0.3}$. The interval shown is the Monte Carlo confidence interval calculated using Monte Carlo error (MCE).

Variable selection methods are frequently criticized for underestimating the variance of coefficient estimates which results in p-values which are too small and confidence intervals which are too narrow. In a simulation setting, the standard deviation across simulation replicates of the estimates of the coefficients can be compared to the mean estimates of the standard error. Table 4.3 shows the ratio of these two. Values greater than 1 indicate the model underestimates the standard error of the coefficient. With n < 20000, all methods have a ratio greater than one, with the exception of the Two-Stage method at $n = 250$. Although not less than 1, BMA tends to provide the lowest ratio of the methods. CIE 5% and the Two-Stage method also occasionally provide lower ratios but this is not as consistent across n as the BMA trend is. Coverage is the probability that the true effect, 0.3, is found in the confidence interval for $\hat{\beta}_{INTEREST}$. Coverage less than 0.95 suggests either the interval is too narrow (standard error too small), or $\hat{\beta}_{INTEREST}$ is too biased. When n is small, $n \leq 250$, only Monte Carlo intervals for BMA Average include 0.95(intervals not shown). When $n = 20000$, BE BIC and all three BMA estimates include 95% coverage in the Monte Carlo interval(intervals not shown).

An ideal method has low bias and standard error ratios close to or less than 1. Figure 4.3 shows these traits simultaneously. BMA is closest to this ideal when $n = 250$. As n increases, BE BIC and all three BMA methods move closer to the ideal of no bias and standard error ratio of one. BE p=0.05 and BE AIC move towards unbiasedness, but fail to achieve a ratio of 1 or less. CIE and Two-Stage achieve neither goal.

Table 4.3: Estimation of Variable of Interest: Variance and Coverage.

	Ratio of $sd(\hat{\beta}_{INT.})$ **and** $mean(\hat{se}(\beta_{INT.}))$ [*]				**Coverage** [†]			
Method	**N=250**	**N=1000**	**N=5000**	**N=20000**	**N=250**	**N=1000**	**N=5000**	**N=20000**
BE p=0.05	1.206	1.197	1.024	1.033	0.901	0.884	0.948	0.941
BE AIC	1.210	1.103	1.047	1.059	0.904	0.925	0.943	0.937
BE BIC	1.201	1.298	1.019	1.006	0.902	0.830	0.949	0.947
CIE 5%	1.081	1.031	1.038	1.035	0.932	0.943	0.945	0.941
Two-Stage	0.958	1.052	1.494	1.037	0.933	0.803	0.698	0.942
BMA Average	1.031	1.080	1.009	1.003	0.946	0.896	0.952	0.947
BMA Top	1.200	1.297	1.019	1.006	0.903	0.829	0.949	0.947
BMA Median	1.199	1.308	1.018	1.006	0.901	0.826	0.949	0.947

[*] Ratio

[†] Coverage is the percent of replications whose 95% confidence interval for $\hat{\beta}_{INTEREST}$ includes the true value, 0.3. Coverage less than 0.95 suggest the interval is either too narrow or too biased.

4.5.4 Predictive Performance

A selection method with high $D_{Y_{external},P(\hat{Y}=1)_{external}}$ and small over-fit is ideal. Figure 4.4 presents these values as n increases. Increasing n improves predictive performance in all methods, both in increasing Somers' D and decreasing the over-fit. For small studies, n = 250, all methods have a low D and high level of over-fit. Most methods, except BMA Average, BE AIC and CIE 5%, result in D close to 0.18 and an over-fit around 0.14. This is equivalent to an AUC of 0.59 in an external sample, and an overestimate of AUC of 0.07. The other three methods, BMA Average, BE AIC and CIE 5%, are better at external prediction, but this comes with the price of worse over-estimation. As n increases, CIE 5% has the lowest over-fit, but is the worst method for predictive performance in the external sample. Conversely, BE AIC had close to the best $D_{Y_{external},P(\hat{Y}=1)_{external}}$, but always had the most over-fit. BE

BIC, BMA Average, BMA Top, and BMA Median predict very similarly as n increases and approach the ideal represented by the bottom right of the plots of lowest over-fit and highest $D_{Y_{external},P(\hat{Y}=1)_{external}}$.

4.5.5 Agreement

Although every pairwise agreement was recorded in the simulation, only the interesting ones are shown in Table 4.4. BE BIC often matched the BMA model with the highest PMP. BMA Top also very often matched the model which included all variables with PIP¿0.5 (BMA Median). Recall the BMA method employed here uses a BIC approximation for (4.7) (Yeung et al. 2005, Raftery 1995), and indeed the models selected by BMA and BE BIC are quite similar. BE BIC and BE AIC show little to no agreement. BE p=0.05 shows its highest agreement with BE BIC. This agreement shows an odd trend of first decreasing as n goes from 250 to 5000 then increasing as n goes to 20000. Two-Stage method showed highest similarity to BMA and BE p=0.05 at small n, approximately 32 and 38%, but this decreases to zero as n increases. CIE never agreed with the other methods. Table 4.5 shows the probability that given two sequential methods agree on the model, the model they are agreeing on is the truth. Agreement rates between sequential methods are never greater than 67%. Further, only when BE p=0.05 and BE BIC, or BE AIC and BE BIC agree do they have a good chance, greater than 95%, of agreeing on the true model, but the chance of these agreeing is only 67% and 28% respectively. In fact, randomly selecting any pair of these sequential methods, when n = 20000, yields only 20.4% chance of the two sequential methods agreeing on the true model. Only considering BE p=0.05, BE AIC and BE BIC yields only 40.6% chance that two of these methods will agree and agree on the true model. At the same time, using BE BIC alone finds the correct model 94% of the time. As others have suggested, using agreement between sequential variable selection techniques as a selection method is unwise (Wiegand 2010).

Table 4.4: Agreement.

Selection Method	**n=250**	**n=1000**	**n=5000**	**n=20000**
BE p=0.05 and BE BIC	0.621	0.163	0.366	0.668
BE p=0.05 and BMA Top	0.604	0.155	0.365	0.667
BE p=0.05 and BE AIC	0.036	0.102	0.369	0.449
BE AIC and BE BIC	0.027	0.011	0.110	0.275
BE p=0.05 and Two-Stage	0.381	0.129	0.015	0.003
Two-Stage and BMA Top	0.316	0.133	0.027	0.003
Two-Stage and BMA Median	0.335	0.140	0.027	0.003
CIE 5% and Two-Stage	0	0	0	0
BE BIC and BMA Top	0.956	0.967	0.998	1.000
BE BIC and BMA Median	0.820	0.851	0.983	0.999
BMA Top and BMA Median	0.836	0.865	0.984	0.999

Table 4.5: Probability that Agreement is on the True Model.

Sequential Method	**n=250**	**n=1000**	**n=5000**	**n=20000**
BE p=0.05 and BE BIC	No Agreement	0.002	0.442	0.986
BE p=0.05 and BMA Top	No Agreement	0.003	0.442	0.986
BE p=0.05 and BE AIC	No Agreement	0.033	0.525	0.642
BE AIC and BE BIC	No Agreement	0.018	0.631	0.988
BE p=0.05 and Two-Stage	No Agreement	0	0.425	0.625
Two-Stage and BMA Top	No Agreement	No Agreement	0.119	1
Two-Stage and BMA Median	No Agreement	No Agreement	0.117	1
CIE 5% and Two-Stage	No Agreement	No Agreement	No Agreement	No Agreement
BE BIC and BMA Top	No Agreement	0	0.219	0.934
BE BIC and BMA Median	No Agreement	0	0.221	0.935
BMA Top and BMA Median	No Agreement	0	0.221	0.935

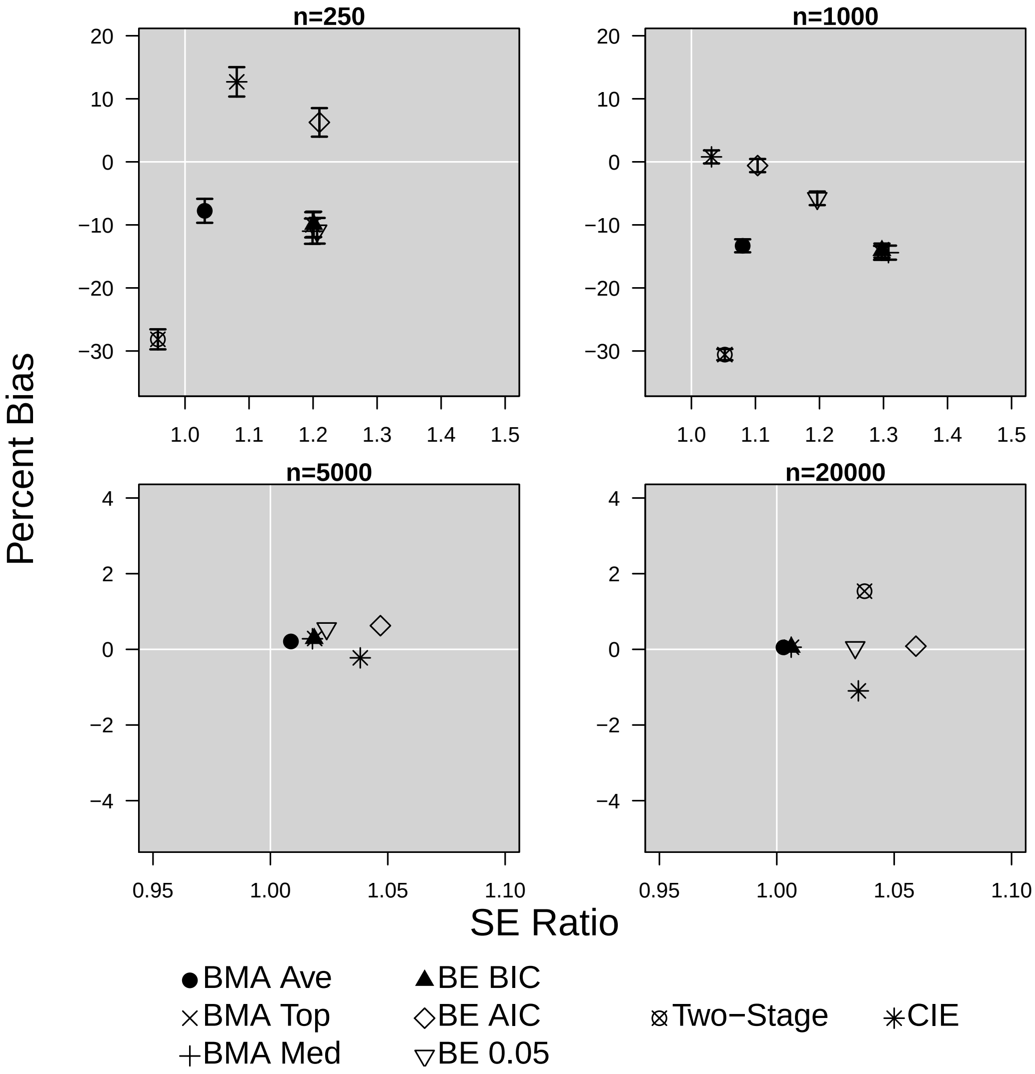

Figure 4.3: Percent Bias by SE Ratio.

Ideal estimates have bias of zero and an SE ratio close to one. This target is indicated by the crossing white lines. If the target is not achieved, conservative methods have negative bias and ratio less than one. Not until n = 20000 do any of these methods come close to the target of no bias and SE ratio equal to one. Note for n = 5000 the Two-Stage method with bias -12.08 and SE ratio 1.49 is not shown. Also, for $n \geq 20000$ the intervals are not shown because they were all approximately the same size as the points.

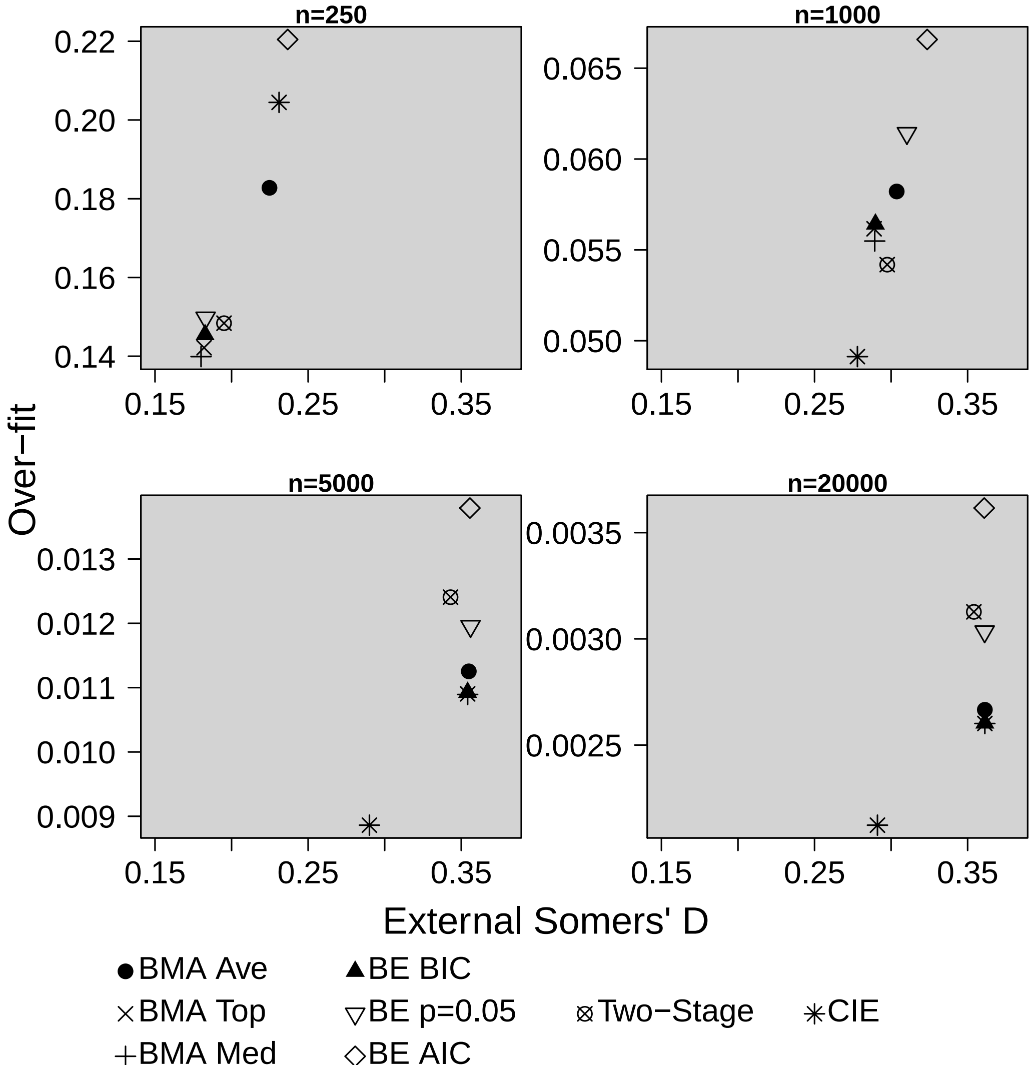

Figure 4.4: Prediction Capabilities: $D_{Y_{external},P(\hat{Y}=1)_{external}}$ and $over-fit$.

4.6 Summary and Limitations

In order to simultaneously consider variable selection method performance in variable selection, effect estimation and prediction, findings are condensed into one last table, Table 4.6. To begin, consider selecting a method when estimation of the variable of interest is the primary goal of a study. Unless n is quite large, ≥ 20000, no method provides an estimate without bias, without underestimation of the standard error of $\hat{\beta}_{INTEREST}$, and with at least 95% coverage. When n = 20000 BE BIC, BMA average, BMA top, and BMA median models achieve all three of these ideals. Although there is still technically underestimation of the standard error, BMA has the least underestimation and is unbiased with good coverage when n = 5000. Likewise, CIE 5% has the least underestimation and is unbiased with good coverage when n = 1000.

Next, consider prediction the modeling goal. Although BMA methods were developed to possess some of the best predictive qualities, there was not an overwhelming indication that this is indeed the case. When n = 250, BMA had the lowest over-fit of the models with better external prediction. As n increases, BMA approaches the ideal of high external prediction and low over-fit, but it is accompanied by BE BIC. BE AIC always had the highest over-fit, and usually CIE 5% had the worst external prediction ability.

Finally, consider the goal of understanding the relationship between the variables by studying the selected model. All methods neglect to find the true model for smaller samples of n = $150, 250, 500$ and 1000. Even at n = 5000, the best model selector, BE p=0.05, only selects the true model 45.3% of the time. This emphasizes the difficulty of variable selection. It is not until n = 10000 that the true model is selected with any reliability, with the BMA Top, BMA Median, BE BIC, and BE 0.05 all finding it roughly 62% of the time. At small n, the methods are missing variables with small effects (in Group E) and too often including correlated variables with no effect (Group C). Then as n increases, the methods still miss variables with small effects but no longer include any variables from Group C. Lastly as n increases to 20000, many methods settle on the truth, but methods that by design allow a high level of noise, like BE AIC, or that fail to eliminate collinear variables, like Two-Stage, never achieve the true model.

This simulation study does have limitations. It could be argued that criteria used to judge a good model in Table 4.6 are too stringent. This study cannot claim to exhaustively include all of the many variable selection methods, both ad-hoc and data driven, employed by public health researchers. However, the most commonly used techniques are represented. Similarly, logistic regression is one of the most frequently used types of regression in public health studies. A simulation with these variable selection methods used in linear regression was performed, but is not presented here. All the methods performed better in the linear setting, but conclusions were not substantively different. Additional parameters of the simulation set-up could have been explored, such as strength of the coefficient of interest, strength of correlation, types of variables and number of variables under consideration. However, this study does glean valuable information about how these variable selection methods behave in a realistic public health setting and it does not seem that these modifying the simulation parameters would drastically change the overall conclusions.

Table 4.6: Summary.

	n=250	**n=1000**	**n=5000**	**n=20000**
Estimation				
No Underestimate	Two-Stage			BE BIC
of $se(\hat{\beta})$				BMA Average
Ratio ¡1				BMA Med,Top
Smallest Ratio ¿1	BMA Average	CIE 5%	BMA Average	
	(1.06)	(1.02)	(1.03)	
Unbiased				
		BE AIC	BE AIC	BE AIC
MC Interval for Bias		CIE 5%	CIE 5%	
Overlaps with 0			BE BIC	BE BIC
			BE p=0.05	BE p=0.05
			BMA Average	BMA Average
			BMA Med,Top	BMA Med,Top
Coverage				
	BMA Average		BMA Average	BMA Average
MC Interval		CIE 5%	CIE 5%	
Includes 0.95			BMA Med,Top	BMA Med,Top
			BE BIC	BE BIC
			BE AIC	
				Two-Stage
Predictive	BEST:	BEST:	BEST:	BEST:
Performance				
Highest $D_{Y_{ext.},P(\hat{Y}=1)_{ext.}}$	BMA Average	BMA Average	BMA Average	BMA Average
and lowest $over-fit$			BMA Med,Top	BMA Med,Top
		BE BIC	BE BIC	BE BIC
		Two-Stage		
	AVOID:	AVOID:	AVOID:	AVOID:
	CIE 5%	CIE 5%	CIE 5%	CIE 5%
	BE AIC	BE AIC	BE AIC	BE AIC
				BE p=0.05
				Two-Stage
Variable Selection				
Selected True Model			BE p=0.05 (48%)	BE p=0.05 (69%)
			BMA Med,Top	BMA Med,Top
			(22%)	(94%)
			BE BIC (22%)	BE BIC (94%)
			BE AIC (23%)	BE AIC (31%)

4.7 Discussion

This study attempts to provide insight into one of the most commonly used analysis tool, variable selection. While sometimes there are enough subjects in a study that a researcher could include every variable available to them and not over-extend the model, there are, however, studies where variable selection is necessary. Often variable selection is not formally reported in an article, but many combinations of variables were tried until the final model is decided. This simulation study shows the dramatic effects variable selection can have on goals of estimation, prediction and model recovery in logistic regression setting with a variable of interest and other possible confounders and leaves the us with three key thoughts. First, although quite popular, change-in-effect methods fail to effectively select variables with an effect on the outcome, fail to estimate the effect without bias or overstating the p-value, and fail to predict the outcome well. Second, success of sequential methods relies heavily on the cut-point or criteria chosen. Third, BMA promises to solve issues related to model uncertainty, and succeeds with diminishing returns.

CIE, designed to discover confounders, never reliably found the potential confounders, but instead focused on variables correlated with the variable of interest. It is also not surprising that CIE fails to estimate well. CIE is known to allow bias up to one half of the cut-off (Maldonado and Greenland 1993). For example, CIE 20% can result in up to 10% bias. However, this study found bias rates greater than 2.5% for CIE 5%. One of the early studies on CIE which is frequently cited (over 1600 times in fact) only examines CIE in a small case-control study with one variable of interest and one confounder cautions the reader about using CIE in cohort studies and suggests instead a variety of the Two-Stage approach (Mickey and Greenland 1989). Interestingly, in this study the Two-Stage approach fared little better than CIE, and resulted in greater bias and over-fit. Although popular and somewhat easily implemented, both CIE and Two-Stage variable selection should be avoided. Other easily implemented techniques are far more effective.

Effectiveness in estimation, prediction and variables selection varied dramatically among the other sequential variable selection methods. In samples of n = 10000 or fewer BE p=0.05 was the most effective at discovering the true model. However, this model struggles with bias and over-fitting the predicted values. For $n \geq 10000$ BE BIC performed very well in terms of bias, overstating the significance, and over-fitting the predicted values. BIC was derived as an approximation to the log of the posterior probability of a model where each model has an equal prior probability and is asymptotically equivalent to choosing a model based on Bayes factors (Schwarz et al. 1978). Theoretically, BIC is a consistent model selector, meaning that if the true model is among the models considered, as n increases, the probability that BIC will identify the true model converges to 1 (Yang 2005). This property is reflected in the successful variable selection of BE BIC Figure 4.1 demonstrates. In the data-scenario investigated by our simulation, this consistency required a much larger n than expected. One rule of thumb suggests good model fit can be achieved in logistic regression when the number of subjects in the smaller of the outcome groups is 10 times the number of variables included in the model (Peduzzi et al. 1996). When n = 10000, our simulation averaged 1850 subjects with Y = 1 and used a maximum of 16 variables, or 115 events-per-variable. Variable selection with AIC is asymptotically equivalent to model selection with cross-validation (Stone 1977). The probability of selecting the true model using the AIC criterion converges to 1 as the number of variables considered increases. In a setting with many more variables considered, AIC may surpass BIC in selection ability. However, none of these sequential methods directly address model uncertainty. Sequential testing techniques have been said to over-state the variable magnitude and significance, be unstable (a small change in the data can lead to a different model), and the p-value cut-off is considered quite arbitrary (Simon and Altman 1994, Harrell et al. 1996, Harrell 2001).

BMA was designed to directly account for model uncertainty, and therefor it was expected not to underestimate the variance of the coefficients and not over-fit the predicted values. This trait has been theoretically shown and claimed before (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). In this simulation we find that BMA does quite well at predicting the outcome with the least amount of over-fitting. However, BMA is beaten slightly by BE BIC. In this study we relied on the easily implementable BMA which relies on a BIC approximation. BMA proponents may claim it is possible that in it's purest form BMA would surpass BE BIC. However, the goal of this study was to investigate how these variable selection methods perform in the hands of public health researchers. In addition to accounting for model uncertainty, BMA provides the researcher with the helpful PIP without needing a simulation. Often, the p-value is misinterpreted as the probability that a variable belongs in a model(Greenland et al. 2016). This is in fact the interpretation of PIP, while a p-value is the probability of observing a coefficient of this magnitude or greater when in fact there is no relationship between the outcome and predictor variable, assuming the correct model has been used. This interpretation is often what researchers are wanting a p-value to be when they misinterpret a p-value (Greenland et al. 2016).

Given the relative success BE p=0.05 has at selecting the correct model for moderately sized n, it would be interesting to investigate other corrections to account for model uncertainty. Applying the bootstrap to account for model uncertainty is one option that may address these concerns (Harrell 2001, Draper 1995, Buckland et al. 1997, Harrell et al. 1996). Harrell 1996 provides a tutorial for using the bootstrap to account for the optimism in multivariable models (Harrell et al. 1996). A shrinkage penalty can also be used after sequential selection to improve the final model (Harrell et al. 1996, Steyerberg et al. 2000). If BMA is unavailable, it may that BE performs quite similarly, however, adjustment needs to be made to account for model uncertainty. BE with a specific p-value could be chosen based on how much noise one is willing to tolerate in the model, or BIC could be chosen.

The ultimate aim of this article is to provide the public health researcher with information to guide one of the most critical parts of their choice in data analysis. Just like every other analysis tool, no single variable selection method can be recommended for use in every study. By presenting direct comparisons of these varied variable selection methods in a simulation setting where the truth is both well-defined and complex enough to represent a real analysis, this study allows the researcher to judge their own study's traits and goals and select the method that fits best. These results should also serve as a reminder that popularity of a method can have little to do with it's objective abilities.

CHAPTER 5: PREDICTING VISCERAL ADIPOSE TISSUE WITH BAYESIAN MODEL AVERAGING IN THE CHINA HEALTH AND NUTRITION SURVEY

5.1 Introduction

There is heterogeneity in the metabolic risk of obesity, some obese individuals are at very high metabolic risk, while others are not and being able to predict people who fall in this category is critical for targeting intervention and for understanding the health of a population. While there is debate about which depot of fat may be causally responsible for metabolic complications of obesity (Fabbrini et al. 2009, Klein 2004), visceral fat has been shown to be associated with metabolically abnormal obesity (Pouliot et al. 1992, Banerji et al. 1995, Gastaldelli et al. 2002). Visceral fat has stronger associations with cardio-metabolic diseases than BMI (Wajchenberg 2000, Fontana et al. 2007, Saito et al. 2012, Beaumont et al. 2016), the standard measure of obesity.

Visceral adipose tissue (VAT) can be expensive to measure and may not be historically available in large population studies. Computed tomography (CT) and magnetic resonance imaging (MRI) are considered the gold standard of VAT measurement (Rankinen et al. 1999, Seidell et al. 1990, Koester et al. 1992, Ross et al. 1992, Van der Kooy et al. 1993). Dual-energy x-ray absorptiometry (DXA) whole body scans have been suggested as an alternative (Snijder et al. 2002, Bertin et al. 2000, Direk et al. 2013). None of these measuring techniques are feasible in large population studies. Instead, a variety of anthropometric measures have been suggested as indices of VAT. Waist circumference (Pouliot et al. 1994, Grundy et al. 2013, Ross et al. 1996) and waist-to-hip ratio (Ashwell et al. 1985, Rankinen et al. 1999) have been found to correlate with visceral fat. Body mass index (BMI) is used to define obesity and is commonly used in clinical and epidemiological studies (Smalley et al. 1990, Spiegelman et al. 1992). Investigating the predictive ability of more readily accessible anthropometric and demographic measures in a multivariable model is an important step in exploiting the richness of existing population studies to better understand the role of visceral fat in the development of metabolically abnormal obesity. Further, a predictive model could help establish better identification of metabolically abnormal obesity in a clinical setting.

Other studies have investigated the predictive ability of a variety of anthropometric measures of visceral fat, but these studies have small sample sizes (Pinho et al. 2017, Goel et al. 2008, Swainson et al. 2017), are limited to only including overweight patients (Pinho et al. 2017), have poor predictive performance (Pinho et al. 2017), or only examine one anthropometric measure at a time (Swainson et al. 2017). The China Health and Nutrition Survey (Popkin 2010) offers a large sample with a wide range of BMI. Using a single anthropometric measure at time avoid problems of collinearity, but there is always debate about which anthropometric measure is the best to include. Although stepwise procedures can help choose between variables to include, they are prone to over-estimating the predictive ability of a model, rendering a seemingly good model unable to replicate its predictive abilities in another study. Bayesian model averaging (BMA) provides an opportunity for exploring many possible models while appropriately accounting for the uncertainty surrounding variable selection (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). BMA is only one of several statistical methods developed for appropriately accounting for model uncertainty but has the advantage of providing the best predictive qualities (George 2000, Madigan and Raftery 1994). By averaging over many possible models, BMA de-emphasizes the importance of selecting a perfect model and instead harnesses the usefulness of a variety of plausible variable combinations. BMA also provides a useful estimate of the probability of any explored model being the true model given the observed data. We employ BMA to investigate a predictive model of VAT using anthropometric measures in CHNS to help understand which combinations of anthropometric measures are most predictive.

5.2 Methods

5.2.1 The China Health and Nutrition Survey

The China Health and Nutrition Survey (CHNS) collected health data in 361 communities (15 provinces and autonomous cities/districts of Beijing, Chongqing, Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning, Shaanxi, Shandong, Shanghai, Yunnan, and Zhejiang) throughout China in ten survey rounds from 1989 to 2015 using a multistage, random cluster design. Survey procedures have been described elsewhere (Popkin 2010). The study was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill, the China-Japan Friendship Hospital, the Ministry of Health and China, and the Institute of Nutrition and Food Safety, China Centers for Disease Control. Participants gave informed consent.

This analysis includes adults at least 18 and no more than 70 years old in the 2015 survey, with VAT measured (n=12,272) who were not pregnant (n=12,258). Participants with VF score values within five standard deviations from the mean were eligible for inclusion (n=12,219).

5.2.2 Visceral Adipose Tissue

VAT was measured as a Visceral fat (VF) score was measured in light clothing without shoes and socks to the nearest 0.1 kg on Tanita BC601 Bioelectric Impedance Analysis (BIA) scale (Tokyo, Japan), which measures segmental body composition based on the electrical impedance value of a current passing through the body (Lee and Gallagher 2008).

5.2.3 Anthropometry

Weight was measured without shoes and in light clothing to the nearest 0.1 kg on a calibrated beam scale. Height was measured without shoes to the nearest 0.2 cm using a portable SECA stadiometer. BMI was calculated as kg/m^2. Waist circumference was measured midway between lowest rib and iliac crest using SECA tape.

5.2.4 Demographic Variables

Ever smoked (Y/N) and whether beer or alcohol was consumed in the last year (Y/N) were reported at each exam.

5.2.5 Statistical Analyses

Analyses were done in R version 3.5 (R Core Team 2013) using the **BMA** package (Raftery et al. 2018). Anthropometric and demographic variables are summarized across VF quartiles with percentages for categorical variables and with the median and 25th and 75th percentiles for continuous variables. A model to predict VF from anthropometric and demographic measures was built using Bayesian model averaging with the BIC approximation (Draper 1995, Raftery 1995, Raftery et al. 1997, Hoeting et al. 1999). In order to allow this prediction to be useful outside of CHNS, we limited our model to commonly collected variables: age, sex, weight (kg), height (cm), waist circumference (cm), hip circumference (cm), smoking history (Y/N), and alcohol consumption in the past year (Y/N). Additionally, transformations of these basic anthropometric variables are also included: BMI (kg/m^2), waist-to-height ratio and waist-to-hip ratio. All of these anthropometric measures are correlated with each other and a model which includes them all may make little sense. Using BMA allows us to use the strengths of all of these measures and gauge which is the most important.

Stepwise procedures are frequently used to determine a single best model. BMA does not select a single model, but instead investigates all possible models. Each model then contributes to a weighted average of models, in this case, the weight depends on the Bayesian Information Criterion, or BIC. Both the estimates of the coefficients, and estimates of the predicted outcome are averaged over all the models. This averaging scheme provides better predictions than any single model could (George 2000, Madigan and Raftery 1994). Further, BMA provides a measure of the likelihood of the data being supported by each model, given the observed data, relative to the other models considered (posterior model probability, PMP) and a probability that each coefficient is not equal to zero (posterior inclusion probability, PIP). PMP are the weights for the weighted average and the sum of PMP across all models is one. For a more thorough introduction to BMA see Hoeting et al. (1999) and Raftery (1995).

Separate predictive models are built for men and women. Because smoking prevalence was so low (1.7%), ever smoked was only included in the men's model. We specify the prior as our belief that each variable is included in the model. In this case, we selected our prior inclusion probabilities to be 0.5 indicating that we believe each variable is as likely to be included in the model predicting VAT as it is to be excluded. We used Occam's window approach to examining all the possible models with an OR=20 (Madigan and Raftery 1994). Example code to run the model and summarize the results can be found in the appendix.

5.3 Results

5.3.1 Participant Characteristics

Compared with the overall eligible CHNS population (n=12,219), the group included in the study (n=12,143) was (older, of lower weight, shorter and from less-urbanised areas, and had smoked more TBD) than the excluded group (n= 73).

Table 5.1 presents the characteristics of the potential predictor variables by sex. Among both men and women, weight, waist circumference, hip circumference, waist-to-weight ratio,

Table 5.1: Sample Characteristics by Sex.

	MEN n=5593	WOMEN n=6626	Missing
Age (years)	49.7 (40.9, 60.1)	48.9 (39.7, 59.5)	0
Ever Smoked	50.6% (2818)	1.7% (109)	25
Consumed Alcohol	55.8% (3123)	6.6% (437)	0
Weight (kg)	68.4 (60.0, 75.8)	59.1 (52.5, 65.0)	7
Height (cm)	167.6 (163.0, 172.0)	156.8 (153.0, 160.8)	4
Waist Circumference (cm)	87.0 (80.0, 94.0)	82.3 (75.0, 89.0)	66
Hip Circumference (cm)	95.4 (91.0, 100.5)	94.1 (89.7, 99.7)	15
Waist-to-Height Ratio	0.52 (0.48, 0.56)	0.53 (0.48, 0.57)	68
Waist-to-Hip Ratio	0.93 (0.86, 0.95)	0.90 (0.82, 0.91)	71
BMI (kg/m^2)	24.3 (21.8, 26.6)	24.0 (21.5, 26.2)	7

Consumed alcohol indicates reported beer or alcohol consumption in previous year.

waist-to-hip ratio and BMI increased with increasing quartile of VF score (Table 5.2). Only height did not increase with VF score quartile.

Table 5.2: Anthropometry by Visceral Fat Score.

	Quartile 1 [0, 6)	Quartile 2 [6, 8)	Quartile 3 [8,12)	Quartile 4 [12,31]	Missing
MEN					
	N=719	N=542	N=1503	N=2829	
Weight (kg)	56.3 (51.4, 60.4)	61.3 (55.6, 66.9)	65.2 (59.5, 70.2)	74.6 (67.5, 80.7)	4
Height (cm)	167.3 (163.0, 172.0)	166.7 (162.0, 172.0)	167.5 (163.0, 172.0)	167.9 (163.5, 172.0)	3
Waist Circ. (cm)	75.7 (70.0, 80.0)	81.1 (75.0, 86.0)	84.1 (79.0, 88.4)	92.5 (87.0, 97.8)	32
Hip Circ.(cm)	88.9 (85.0, 92.3)	91.6 (87.7, 96.0)	93.5 (90.0, 98.0)	98.9 (95.0, 103.0)	9
Waist-to-Ht Ratio	0.45 (0.42, 0.48)	0.49 (0.45, 0.51)	0.50 (0.47, 0.53)	0.55 (0.52, 0.58)	34
Waist-to-Hip Ratio	0.86 (0.80, 0.89)	0.90 (0.83, 0.92)	0.91 (0.85, 0.93)	0.96 (0.89, 0.96)	34
BMI (kg/m^2)	20.1 (18.6, 21.1)	22.0 (20.5, 23.1)	23.2 (21.5, 24.5)	26.4 (24.4, 28.1)	4
WOMEN					
	N=2142	N=2092	N=2157	N=235	
Weight (kg)	51.3 (47.2, 55.3)	58.4 (54.3, 62.0)	66.3 (61.0, 71.6)	70.8 (60.4, 79.8)	3
Height (cm)	157.3 (153.3, 161.0)	156.6 (152.5, 160.5)	156.4 (152.5, 160.2)	156.7 (152.5, 161.8)	1
Waist Circ. (cm)	74.1 (69.0, 78.6)	81.6 (77.3, 85.0)	90.0 (85.0, 95.0)	93.9 (85.0, 95.0)	34
Hip Circ. (cm)	88.7 (85.8, 93.0)	93.4 (90.6, 97.4)	99.4 (96.0, 104.0)	101.9 (95.2, 111.0)	6
Waist-to-Ht. Ratio	0.47 (0.44, 0.50)	0.52 (0.49, 0.55)	0.58 (0.54, 0.61)	0.60 (0.54, 0.67)	34
Waist-to-Hip Ratio	0.85 (0.78, 0.87)	0.90 (0.82, 0.90)	0.93 (0.86, 0.94)	0.98 (0.86, 0.95)	37
BMI (kg/m^2)	20.7 (19.4, 22.0)	23.8 (22.7, 24.8)	27.1 (25.5, 28.7)	28.9 (24.5, 33.2)	3

For continuous variables we present Mean(25th percentile, 75th percentile), and for categorical Percent (N).

5.3.2 Prediction of VF with Anthropometrics

The model predicting VF score for men included 10 potential predictor variables, leading to 2^{10}, or 1024, potential sets of predictors. Table 5.3 shows the top 5 most likely models of the 1024 for men and the top 5 of the 512 for women (2^9, smoking variable not included). For men and women, no top model attains more than 47% posterior model probability, signaling uncertainty among the models. There are, however, striking similarities across the ten models. For example, age and weight are always included while alcohol consumption and waist-to-hip ratio are never included.

For men, the top model includes waist circumference, weight, height, BMI and hip circumference with posterior model probability (PMP) of 0.346. The next most likely model includes waist-to-height ratio instead of waist circumference and height (PMP=0.246). Between the second and third most likely models, height is added (PMP=0.237). In all the top 5 models for men, if waist circumference was not included, waist-to-height ratio was. In general, men have a core model of age, weight, BMI and hip circumference and then some combination of waist-to-height ratio, waist circumference and height.

The top model two models for women account for 76.5% of the posterior model probability, compared to only 59.2% of the top two models for the men. Unlike men, hip circumference is not included in the top 5 models for women. The top model for women includes age, weight, BMI and waist-to-height ratio (PMP=0.474). The second most likely model includes waist circumference instead of BMI (PMP=0.291). The third most likely model is like the first, but substitutes waist circumference for waist-to-height ratio (PMP=0.172). In general, the women's models always include age and weight and includes no more than two of waist circumference, waist-to-height ratio and BMI.

Table 5.4 provides the posterior inclusion probabilities (PIP) determined by summing the posterior model probabilities of the models including each variable, with the interpretation for these PIP (Kass and Raftery 1995). Average effect estimates and standard error estimates weighted by PMP are also provided. Predicted VF scores for men from the averaged model have a correlation with the observed VF scores of 0.828, or $R^2 = 0.685$. Predicted VF scores for women from the averaged model have a correlation with the observed VF scores of 0.781, or $R^2 = 0.610$. Figure 5.1 shows the predicted VF score by the observed VF score. The predicted scores are higher than the observed scores for men and women with low observed VF score. In women especially, predicted VF scored is lower than the observed VF score for those with a high observed VF score.

Table 5.3: Summary of Top 5 Models

	MEN					**WOMEN**				
	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
Model Probability	0.346	0.246	0.237	0.075	0.031	0.474	0.291	0.172	0.035	0.029
Age (years)	•	•	•	•	•	•	•	•	•	•
Alcohol (yes)										
Waist (cm)	•			•			•	•	•	
Height (cm)	•		•		•					•
Weight (kg)	•	•	•	•	•	•	•	•	•	•
Waist-to-Height		•	•		•	•	•		•	•
Waist-to-Hip										
BMI (kg/m^2)	•	•	•	•		•		•	•	•
Hip	•	•	•	•	•					
Ever Smoke (in men)										

The • indicates which variables are included in each model. When the prior probability of each model is equal, like we have here, the model probability is determined as the BIC for a particular model divided by the sum of the BIC for all models investigated.

Table 5.4: Estimation of Variable of Interest.

Variable	**Inclusion** **Probability**[*]		**Average** **Effect**	**Average** **SE**
MEN				
R^2=0.685^{dagger}				
Intercept	100		-10.976	7.773
Age (10yr)	100	very strong	1.361	0.029
Alcohol	0	evi. against	0	
Smoke	0	evi. against	0	
Height (cm)	65.5	weak	-0.055	0.046
Weight (kg)	100	very strong	0.143	0.052
Waist (cm)	44.5	evi. against	0.019	0.022
Hip (cm)	93.4	positive	0.015	0.006
Waist-to-Height	55.5	weak	4.075	3.714
Waist-to-Hip	0	evi. against	0	
BMI	96.9	strong	0.381	0.144
WOMEN				
R^2=0.610				
Intercept	100		-9.604	1.050
Age (10yr)	100	very strong	0.567	0.019
Alcohol	0	evi. against	0	
Height (cm)	2.9	evi. against	-0.001	0.007
Weight (kg)	100	very strong	0.098	0.066
Waist (cm)	49.7	evi. against	-0.058	0.095
Hip (cm)	0	evi. against	0	
Waist-to-Height	82.8	positive	12.873	14.911
Waist-to-Hip	0	evi. against	0	
BMI	70.9	positive	0.243	0.162

[*] Inclusion probability means given the data we observe, what is the probability that a particular variable is included in the model. This probability is determined as the sum of the posterior model probabilities of the models which include the variable. Evidences against includion is abbreviated evi. against. R^2 is the Pearson correlation between the predicted VF score and the observed VF score squared.

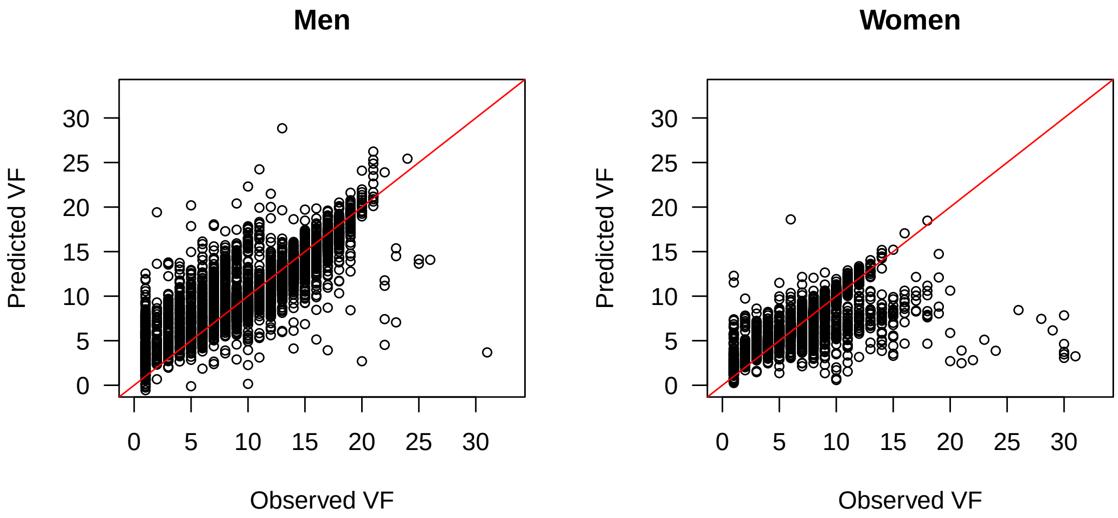

Figure 5.1: Predicted VF Score by Observed VF Score

5.4 Discussion

This study develops predictive models of VAT using commonly collected anthropometric measures in the China Health and Nutrition Survey. Employing Bayesian Model Averaging techniques accounts for uncertainty in predictor selection and ensures good predictive qualities. VAT has been shown to be associated with metabolically abnormal obesity, and determining an inexpensive and widely available measure of VAT is an essential step to investigating the role of VAT in large epidemiological studies.

Previous studies of VAT and anthropometrics have suggested a variety of anthropometric predictors. Waist-to-hip ratio has been suggested over waist or BMI for men ¡40 years of age (Rankinen et al. 1999) and more generally for all ages of men (Ross et al. 1992). Waist-to-hip ratio was suggested to be a useful predictor of VAT in women in a small study (n=28) as well (Ashwell et al. 1985). Waist circumference has also been selected as the preferred predictor of VAT (Ross et al. 1996, Pouliot et al. 1994). Waist-to-height ratio has been touted as the best predictor of VAT (Swainson et al. 2017). Other measures considered with varying success in multivariable models have included BMI (Janssen et al. 2002, Goel et al. 2008), hip circumference (Goel et al. 2008), conicity index (Pinho et al. 2017), sagittal diameter pinho2017predictive, neck circumference pinho2017predictive. By exploring all possible models and quantifying the uncertainty related to selecting a single "best" model, BMA is ideally suited to this scenario and can help us understand the variety of sometimes conflicting previous results.

We found the models for men almost always included age, weight, BMI, hip circumference, and a representation of waist circumference as either the circumference, or waist-to-height ratio. The models for women always included age, weight, and typically two of the following: BMI, waist circumference or waist-to-height ratio. While hip circumference had a high PIP for men, waist-to-hip ratio was never included for men or women. The predictive ability of our models (R^2 = 0.685 in men and r^2 = 0.610 in women) is similar to what has previously been observed (Janssen et al. 2002, Swainson et al. 2017), and in some cases much better (Pinho et al. 2017, Goel et al. 2008). Models including BMI and waist circumference to predict VAT with n=341 reported an R^2 = 0.76 in men and R^2 = 0.57 in women (Janssen et al. 2002). Another study individually examined waist-to-height ratio, waist circumference, BMI, waist-to-hip ratio and the ratio of waist to the square root of height in n=41 men and n=32 women and found waist-to-height to be the best predictor (R^2 = $.71$ in men, R^2 = 0.65 in women)(Swainson et al. 2017). An un-stratified model for VAT which included age, BMI, waist and hip circumference in n=124 was less successful than our model R^2 = 0.521 (Goel et al. 2008). Lastly, a model predicting VAT which included abdominal circumference, waist-to-hip ratio, and conicity index in n=28 men had an R^2 = 0.669 and a model using age, sagittal diameter, conicity index and neck circumference in n=81 women had an R^2 = 0.462 (Pinho et al. 2017).

Although BMA did not provide a markedly higher R^2, it provides two significant advantages. First, because BMA explicitly accounts for the variability attributed to variable selection, our estimate of predictive ability is more likely to be replicable in other studies which examine the relationship between anthropometric measures and VF scores among Chinese individuals. By incorporating all the possible models and not selecting just one “best” model, we are less likely to have stumbled upon a statistical fluke. Harnessing the uncertainty in the variable selection also provides our second advantage. In predicting VAT, there have been a variety of combinations of related variables investigated. We learn more about the relationship between these correlated variables and VAT by investigating every possible model and seeing which 5 are the most supported by the data than we would by just selecting one. For example, backward stepwise regression with BIC would have likely selected the model in men that included age, waist circumference, height, BMI and hip circumference. However, by examining the next 4 most likely models we see that waist circumference is somewhat inter-changeable with waist-to-height ratio as far as which model the data most support, which is a substantively different conclusion than saying waist circumference is more important than waist-to-height ratio.

BMA’s ability to account for variable selection uncertainty make it useful in any modeling situation where there is a question about which variables to include that goes beyond prior knowledge. BMA is particularly useful in situations where there are related variables to select from, as there were here with the anthropometric measures. BMA can be easily applied to linear models, logistic regression, poisson regression, and survival models (Raftery et al. 2018).

The strengths of this study were that the sample size (n=12,219) was large and representative of the current Chinese population (Popkin et al. 2009). The main limitation in our study is that BIA was the only measurement tool for VAT. BIA has been compared to MRI for measuring VF in a smaller study of Chinese individuals (n=200) where BIA tended to overestimate at lower levels of VF and underestimate at higher levels of VF (Xu et al. 2011). BIA scale is inexpensive, non-invasive, portable, and a practical screening technique(Lee and Gallagher 2008).

CHAPTER 6: CONCLUSION

Variable selection techniques continue to be commonly employed in public health research. We have provided a comprehensive comparison of both modern and classical variable selection techniques in a setting with realistic variable relationships. We compare the variable selection techniques in terms of their abilities not only to correctly include and exclude variables, but also their abilities to estimate the effects and to predict the outcome. We first compare modern methods in a linear regression setting. The best method from among these, Bayesian model averaging, is then compared to classical variable selection techniques in the logistic regression setting. A predictive model is built using the China Health and Nutrition Survey which showcases the strengths of Bayesian model averaging in practice.

As computing capabilities continue to improve, modern variable selection methods have seen increasing use particularly in high-dimensional data analysis. As excitement around methods such as least absolute shrinkage and selection operator (lasso), stochastic search variable selection (SSVS) and Bayesian model averaging (BMA) grows, it is increasingly likely that we will see these methods applied to non-high-dimensional data. It is imperative that we understand how these methods behave in a non-high-dimensional setting. We have found that BMA did a better job than adaptive lasso or SSVS at accounting for model uncertainty. BMA performed prediction better than the other methods when n was greater than 250, for smaller studies adaptive lasso had better predictive performance. Lastly, BMA tended to have less bias than the other methods, with all methods, except for adaptive lasso, becoming less biased as n increased. BMA performance was next examined is a logistic regression setting. This time, the comparison was not to more modern selection methods, but instead was with the methods most popular to public health practitioners: sequential methods, change-in-effect methods and two-stage methods. We found the very popular change-in-effect methods performed exceptionally poorly at variable selection, effect estimation and prediction. Surprisingly, 15% of public health research articles use this sub-optimal method. Although it performs better than the CIE methods, the two-stage method of using a univariate test to pre-screen variable to include in a multivariable model also performs poorly. BMA was best at estimation, and prediction. However, backward elimination with $p<0.05$ and AIC criteria were best at selecting the true model when $n<10000$. This suggests that another method for addressing model uncertainty might have useful gains for these two methods in particular.

BMA is ideally suited to a particular prediction problem in the China Health and Nutrition Survey. Metabolically abnormal obesity is an important health problem and being able to predict people who fall in this category is critical for targeting intervention and for understanding the health of a population. The standard methods for measuring VAT include MRI and computed tomography, neither of which are typically available in large population studies which were designed to examine obesity trends. Predicting VAT from typically available anthropometric measures is of great importance. BMA provides not only strong predictive performance qualities, but provides useful statistics about plausible models which are particularly helpful given correlations between the anthropometric variables.

BMA is one way of dealing with model uncertainty. Future work could include more careful examination of other methods, particularly bootstrapping methods. This work also highlights the importance of healthy collaboration between statisticians and researchers. Although there are several short-comings, if not failures, of sequential, two-stage, and change-in-effect methods, they are used with considerable frequency. Efforts need to be made to allow better methods to be more accessible and better-known to the public health researcher.

APPENDIX A: SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Table A.1: Simulation Design.

Model Type	n	Error Variance	Effect Size
Linear	150	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	250	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	500	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	1000	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	1000	σ^2 = 0.0016, r^2 ~ 0.13	0
Linear	1000	σ^2 = 0.0625, r^2 ~ 0.40	0.01
Linear	1000	σ^2 = 0.01, r^2 ~ 0.03	0.01
Linear	1000	σ^2 = 0.0004, r^2 ~ 0.01	0.01
Linear	2500	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	5000	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	10000	σ^2 = 0.0016, r^2 ~ 0.13	0.01
Linear	20000	σ^2 = 0.0016, r^2 ~ 0.13	0.01

Table A.2: Variable Inclusion Probabilities n = 1000.

Variable Name	True Effect	Corr†	BMA* PIP	BMA Top	BMA Median	SSVS* PIP	SSVS Top	SSVS Median	Adaptive Lasso	BE BIC
$X_{INT.}$	0.01	1	100	100	100	100	100	100	100	100
$X_{1,EC}$	0.01	0.1	100	100	100	100	100	100	100	100
$X_{2,C}$	0	0.2	2.94	0.68	0.66	0.82	0.06	0.06	0.30	0.68
$X_{3,C}$	0	-0.5	2.88	0.76	0.78	0.92	0	0	0.20	0.82
$X_{4,C}$	0	-0.2	3.08	1	0.88	1	0.18	0.18	0.42	1.26
X_5	0	0	2.76	0.66	0.70	0.76	0.02	0.02	0.28	0.66
X_6	0	0	3.04	1	0.94	0.88	0.06	0.08	0.30	1.10
$X_{7,C}$	0	0.5	3.04	0.78	0.84	0.98	0.04	0.06	0.36	0.88
B_8	0	0	3.74	1.34	1.1	1.52	0.36	0.3	0.02	1.32
$B_{9,E}$	-0.01	0	74.74	76.76	76.30	50.96	50.08	49.62	1.96	76.86
B_{10}	0	0	3.38	1.18	1.08	1.38	0.16	0.14	0	1.26
$B_{11,EC}$	-0.01	0.20	24.62	20.36	20.30	8.88	5.48	5.46	0.02	20.54
B_{12}	0	0	2.92	0.70	0.72	0.92	0.04	0.04	0.06	0.7
$B_{13,E}$	0.005	0	25.16	20.58	20.36	11.52	7.46	7.40	0.42	20.66
X_{14}	0	0	3.04	0.72	0.70	0.82	0.02	0.02	0.34	0.72
$X_{15,E}$	0.005	0	88.24	90.06	90.00	68.82	69.94	70.08	11.06	90.02

† Corr. here refers to the correlation between the variable and $X_{INTEREST}$.

* The BMA PIP and SSVS PIP columns provide the average PIP across the replications instead of the proportion of models including each variable as the other columns present.

Table A.3: Variable Inclusion Probabilities n = 20000.

Variable Name	True Effect	Corr[†]	BMA[*] PIP	BMA Top	BMA Median	SSVS[*] PIP	SSVS Top	SSVS Median	Adaptive Lasso	BE BIC
$X_{INT.}$	0.01	1	100	100	100	100	100	100	100	100
$X_{1,EC}$	0.01	0.1	100	100	100	100	100	100	100	100
$X_{2,C}$	0	0.2	0.46	0.08	0.08	0.20	0.02	0.02	0.48	0.08
$X_{3,C}$	0	-0.5	0.52	0.1	0.1	0.22	0	0	0.18	0.10
$X_{4,C}$	0	-0.2	0.54	0.14	0.14	0.22	0.02	0.02	1	0.14
X_5	0	0	0.66	0.20	0.20	0.26	0.04	0.04	0.52	0.20
X_6	0	0	0.64	0.20	0.20	0.20	0.02	0.02	0.82	0.20
$X_{7,C}$	0	0.5	0.72	0.24	0.24	0.28	0.04	0.04	1.18	0.24
B_8	0	0	0.68	0.12	0.10	0.24	0.02	0.02	0.02	0.14
$B_{9,E}$	-0.01	0	100	100	100	100	100	100	99.84	100
B_{10}	0	0	0.54	0.10	0.10	0.24	0	0	0	0.12
$B_{11,EC}$	-0.01	0.20	99.98	99.98	99.98	99.84	99.92	99.92	7.08	99.98
B_{12}	0	0	0.52	0.10	0.10	0.20	0	0	0	0.1
$B_{13,E}$	0.005	0	100	100	100	100	100	100	64.52	100
X_{14}	0	0	0.66	0.18	0.18	0.18	0	0	0.68	0.18
$X_{15,E}$	0.005	0	100	100	100	100	100	100	100	100

[†] Corr. here refers to the correlation between the variable and $X_{INTEREST}$.

[*] The BMA PIP and SSVS PIP columns provide the average PIP across the replications instead of the proportion of models including each variable as the other columns present.

APPENDIX B: SUPPLEMENTAL MATERIALS FOR CHAPTER 4

Table B.1: Selection of the True Model

	Probability of Selecting True Model %							
Selection Method	**n=150**	**n=250**	**n=500**	**n=1000**	**n=2500**	**n=5000**	**n=10000**	**n=20000**
BE AIC	0	0	0.0036	0.0314	0.1594	0.248	0.296	0.2888
BE p=0.20	0	0	0.004	0.0318	0.124	0.1786	0.2098	0.2016
CIE 5%	0	0	0	0	0	0	0	0
CIE 10%	0	0	0	0	0	0	0	0
CIE 20%	0	0	0	0	0	0	0	0

Table B.2: Estimation of Variable of Interest: Percent Bias.

	Percent Bias % [*]			
Selection Method	**n=250**	**n=1000**	**n=5000**	**n=20000**
	N=250	N=1000	N=5000	N=20000
BE AIC	6.259	-0.590	0.625	0.083
	(-12.967,-8.897)	(-6.863,-4.706)	(0.121,0.980)	(-0.166,0.266)
BE p=0.20	3.479	-0.1	0.59	0.118
	(3.993,8.525)	(-1.633,0.452)	(0.18,1.07)	(-0.141,0.308)
CIE 5%	12.689	0.785	-0.225	-1.098
	(-11.922,-7.902)	(-15.179,-12.96)	(-0.138,0.711)	(-0.154,0.265)
CIE 10%	10.81	2.303	3.581	3.903
	(10.356,15.021)	(-0.250,1.819)	(-0.672,0.222)	(-1.313,-0.883)
CIE 20%	9.945	9.816	17.859	18.471
	(-29.740,-26.560)	(-31.451,-29.688)	(-12.685,-11.482)	(1.315,1.759)

[*] The interval shown is the Monte Carlo confidence interval calculated using the standard deviation of the estimates of the coefficients.

Table B.3: Estimation of Variable of Interest: Variance and Coverage.

	Ratio of $sd(\hat{\beta}_{INT.})$ and $mean(\hat{se}(\beta_{INT.}))$ *				Coverage †			
Selection Method	**n=250**	**n=1000**	**n=5000**	**n=20000**	**n=250**	**n=1000**	**n=5000**	**n=20000**
BE AIC	1.21	1.103	1.047	1.059	0.904	0.925	0.943	0.937
BE p=0.20	1.21	1.09	1.051	1.063	0.904	0.928	0.943	0.936
CIE 5%	1.081	1.031	1.038	1.035	0.932	0.943	0.945	0.941
CIE 10%	1.105	1.119	1.297	1.242	0.928	0.92	0.863	0.863
CIE 20%	1.139	1.218	1.109	1.021	0.923	0.89	0.737	0.277

[*] Ratio

[†] Coverage is the percent of replications whose 95% confidence interval for $\hat{\beta}_{INTEREST}$ includes the true value, 0.01. Coverage less than 0.95 suggest the interval is either too narrow or too biased. Since bias has been shown to be small, any problems with coverage are likely to be caused by an underestimate of the standard error of $\beta_{INTEREST}$.

Table B.4: Prediction: External Somers' D and Overfit.

	External Somers' D				Over-fit †			
Selection Method	**n=250**	**n=1000**	**n=5000**	**n=20000**	**n=250**	**n=1000**	**n=5000**	**n=20000**
BE AIC	0.2367	0.3236	0.3557	0.3608	0.2204	0.0666	0.0138	0.0036
BE p=0.2	0.2293	0.3240	0.3555	0.3608	0.2193	0.0675	0.0143	0.0038
CIE 5%	0.2310	0.2780	0.2900	0.2911	0.2045	0.0491	0.0089	0.0021
CIE 10%	0.2124	0.2470	0.2567	0.2641	0.1725	0.0384	0.0085	0.0018
CIE 20%	0.1893	0.1996	0.1791	0.1768	0.1387	0.0296	0.0042	0.0007

[*] Ratio

[†] Over-fit is the difference in Somers' D in the original and external sample. Positive values indicate Somers' D is greater in the original sample than the external sample.

```
cie<-function(plim=0.1, init.model=lm(y˜xbin+x1+x2),
var.interest="xbin")
{
  plimit<-plim
  full<-init.model
  mainx<-coefficients(full)[var.interest]

  d<-full$model[,c(-1)]
  nuisance<-dim(d)[2]-1
  nuisance2<-nuisance
  pchange<-rep(NA,nuisance)

  for(m in 1:nuisance){

    for(k in 2:(nuisance-m+2)){
      if(dim(d)[2]>2) newm<-lm(y˜., d[,-k])
      if(dim(d)[2]==2) newm<-lm(as.formula(paste("y˜",
var.interest)),d)

      pchange[k-1]<-abs((mainx-coefficients(newm)
[var.interest])/mainx)

    }
    if(min(pchange, na.rm=TRUE)>plimit) return(names(
full$model)[-1])
    if(min(pchange, na.rm=TRUE)>plimit) break
    dropvar<-names(full$model)[
   which(pchange==min(pchange, na.rm=TRUE))+2]

    d<-d[,-which(names(full$model)[-1] %in% dropvar)]
    if(m<nuisance) full<-lm(y˜., d)

    pchange<-rep(NA,nuisance-m)



}
    return(names(full$model[-1]))
}
```

APPENDIX C: SUPPLEMENTAL MATERIALS FOR CHAPTER 5

```
library(BMA)

# read in data
dex<-read.csv("file path")

# create datasets for men and women
dexF<-dex[which(dex$sex==2),]
dexM<-dex[which(dex$sex==1),]

# linear models
modAll<-lm(VATBIA˜age+sex+smoke+drink+
   waist+height+weight+waisttoheight+waisttohip+BMI+waist,
 data=dex)
modF<-lm(VATBIA˜age+drink+
   waist+height+weight+waisttoheight+waisttohip+BMI+waist,
 data=dexF)
modM<-lm(VATBIA˜age.x+smoke+drink+
   waist+height+weight+waisttoheight+waisttohip+BMI+waist,
 data=dexM)

# use data matrices created in linear models to run BMA
bmaall<-bicreg(x=modAll$model[,-1], y=modAll$model[,1],
drop.factor.levels=FALSE)
bmaF<-bicreg(x=modF$model[,-1], y=modF$model[,1],
drop.factor.levels=FALSE)
bmaM<-bicreg(x=modM$model[,-1], y=modM$model[,1],
drop.factor.levels=FALSE)

# summary of BMA models, includes both estimates and
# top models
summary(bmaF)
summary(bmaM)

# plot predicted values by observed values for men
#and women
predictionF<-predict(bmaF, modF$model[,-1], quantiles=
c(.05, .95))
plot(predictionF[[1]], modF$model[,1], xlim=c(0,30),
ylim=c(0,30), xlab="Observed VAT", ylab="Predicted VAT",
main="Women")
abline(0,1, col="red")

predictionM<-predict(bmaM, modM$model[,-1], quantiles=
c(.05, .95))
plot(predictionM[[1]], modM$model[,1], xlim=c(0,30),
ylim=c(0,30), xlab="Observed VAT", ylab="Predicted VAT",
main="Men")
abline(0,1, col="red")
```

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Ashwell, M., Cole, T. J., and Dixon, A. K. (1985). Obesity: new insight into the anthropometric classification of fat distribution shown by computed tomography. *Br Med J (Clin Res Ed)*, 290(6483):1692–1694.

Banerji, M., Buckley, M., Chaiken, R., Gordon, D., Lebovitz, H., and Kral, J. (1995). Liver fat, serum triglycerides and visceral adipose tissue in insulin-sensitive and insulin-resistant black men with niddm. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 19(12):846–850.

Beaumont, M., Goodrich, J. K., Jackson, M. A., Yet, I., Davenport, E. R., Vieira-Silva, S., Debelius, J., Pallister, T., Mangino, M., Raes, J., et al. (2016). Heritable components of the human fecal microbiome are associated with visceral fat. *Genome biology*, 17(1):189.

Bertin, E., Marcus, C., Ruiz, J., Eschard, J., and Leutenegger, M. (2000). Measurement of visceral adipose tissue by dxa combined with anthropometry in obese humans. *International Journal of Obesity*, 24(3):263.

Blattenberger, G., Fowles, R., and Loeb, P. D. (2014). Variable selection in bayesian models: Using parameter estimation and non parameter estimation methods. In *Bayesian Model Comparison*, pages 249–278. Emerald Group Publishing Limited.

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

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Chatterjee, A., Lahiri, S., et al. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259.

Direk, K., Cecelja, M., Astle, W., Chowienczyk, P., Spector, T. D., Falchi, M., and Andrew, T. (2013). The relationship between dxa-based and anthropometric measures of visceral fat and morbidity in women. *BMC cardiovascular disorders*, 13(1):25.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97.

Fabbrini, E., Magkos, F., Mohammed, B. S., Pietka, T., Abumrad, N. A., Patterson, B. W., Okunade, A., and Klein, S. (2009). Intrahepatic fat, not visceral fat, is linked with metabolic complications of obesity. *Proceedings of the National Academy of Sciences*, 106(36):15430–15435.

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.

Fontana, L., Eagon, J. C., Trujillo, M. E., Scherer, P. E., and Klein, S. (2007). Visceral fat adipokine secretion is associated with systemic inflammation in obese humans. *Diabetes*, 56(4):1010–1013.

Freedman, D. A. and Freedman, D. A. (1983). A note on screening regression equations. *the american statistician*, 37(2):152–155.

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.

Gastaldelli, A., Miyazaki, Y., Pettiti, M., Matsuda, M., Mahankali, S., Santini, E., DeFronzo, R. A., and Ferrannini, E. (2002). Metabolic effects of visceral fat accumulation in type 2 diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 87(11):5098–5103.

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.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Goel, K., Gupta, N., Misra, A., Poddar, P., Pandey, R. M., Vikram, N. K., and Wasir, J. S. (2008). Predictive equations for body fat and abdominal fat with dxa and mri as reference in asian indians. *Obesity*, 16(2):451–456.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350.

Grundy, S. M., Neeland, I. J., Turer, A. T., and Vega, G. L. (2013). Waist circumference as measure of abdominal fat compartments. *Journal of obesity*, 2013.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387.

Harrell Jr, F. E. (2018). *rms: Regression Modeling Strategies*. R package version 5.1-2.

Hastie, T. and Qian, J. (2016). Glmnet vignette. https://web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf.

Hayat, M. J., Powell, A., Johnson, T., and Cadwell, B. L. (2017). Statistical methods used in the public health literature and implications for training of public health professionals. *PloS one*, 12(6):e0179032.

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

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

Janssen, I., Heymsfield, S. B., Allison, D. B., Kotler, D. P., and Ross, R. (2002). Body mass index and waist circumference independently contribute to the prediction of nonabdominal, abdominal subcutaneous, and visceral fat. *The American journal of clinical nutrition*, 75(4):683–688.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Klein, S. (2004). The case of visceral fat: argument for the defense. *The Journal of clinical investigation*, 113(11):1530–1532.

Kleinbaum, D., , Kupper, L., Muller, K., and Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Duxbury Pr, Pacific Grove.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic research: principles and quantitative methods*. John Wiley & Sons.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162.

Koester, R., Hunter, G., Snyder, S., Khaled, M., and Berland, L. (1992). Estimation of computerized tomography derived abdominal fat distribution. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 16(8):543–554.

Lee, P. H. (2014). Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *Journal of epidemiology*, 24(2):161–167.

Lee, S. Y. and Gallagher, D. (2008). Assessment methods in human body composition. *Current opinion in clinical nutrition and metabolic care*, 11(5):566.

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.

Maldonado, G. and Greenland, S. (1993). Simulation study of confounder-selection strategies. *American journal of epidemiology*, 138(11):923–936.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, 12(3):621–625.

Mickey, R. M. and Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American journal of epidemiology*, 129(1):125–137.

Miller, A. (2002). *Subset selection in regression*. Chapman and Hall/CRC.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, pages 389–425.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

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.

Pinho, C. P. S., da Silva Diniz, A., de Arruda, I. K. G., Leite, A. P. D. L., Petribú, M. d. M. V., and Rodrigues, I. G. (2017). Predictive models for estimating visceral fat: The contribution from anthropometric parameters. *PloS one*, 12(7):e0178958.

Popkin, B. M. (2010). The implications of the nutrition transition for obesity in the developing world. *Obesity epidemiology: from aetiology to public health*, pages 136–142.

Popkin, B. M., Du, S., Zhai, F., and Zhang, B. (2009). Cohort profile: The china health and nutrition survey-monitoring and understanding socio-economic and health change in china, 1989–2011. *International journal of epidemiology*, 39(6):1435–1440.

Pouliot, M.-C., Després, J.-P., Lemieux, S., Moorjani, S., Bouchard, C., Tremblay, A., Nadeau, A., and Lupien, P. J. (1994). Waist circumference and abdominal sagittal diameter: best simple anthropometric indexes of abdominal visceral adipose tissue accumulation and related cardiovascular risk in men and women. *The American journal of cardiology*, 73(7):460–468.

Pouliot, M.-C., Després, J.-P., Nadeau, A., Moorjani, S., Prud'Homme, D., Lupien, P. J., Tremblay, A., and Bouchard, C. (1992). Visceral obesity in men: associations with glucose tolerance, plasma insulin, and lipoprotein levels. *Diabetes*, 41(7):826–834.

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

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2017). *BMA: Bayesian Model Averaging*. R package version 3.18.7.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2018). *BMA: Bayesian Model Averaging*. R package version 3.18.9.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 25:111–164.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Rankinen, T., Kim, S., Perusse, L., Despres, J., and Bouchard, C. (1999). The prediction of abdominal visceral fat level from body composition and anthropometry: Roc analysis. *International journal of obesity*, 23(8):801.

Rockova, V., Lesaffre, E., Luime, J., and Löwenberg, B. (2012). Hierarchical bayesian formulations for selecting variables in regression models. *Statistics in medicine*, 31(11-12):1221–1237.

Ross, R., Leger, L., Morris, D., de Guise, J., and Guardo, R. (1992). Quantification of adipose tissue by mri: relationship with anthropometric variables. *Journal of applied physiology*, 72(2):787–795.

Ross, R., Rissanen, J., and Hudson, R. (1996). Sensitivity associated with the identification of visceral adipose tissue levels using waist circumference in men and women: effects of weight loss. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 20(6):533–538.

Saito, T., Murata, M., Otani, T., Tamemoto, H., Kawakami, M., and Ishikawa, S.-e. (2012). Association of subcutaneous and visceral fat mass with serum concentrations of adipokines in subjects with type 2 diabetes mellitus. *Endocrine journal*, 59(1):39–45.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, S. L. (2017). *BoomSpikeSlab: MCMC for Spike and Slab Regression*. R package version 0.9.0.

Seber, G. and Lee, A. (2003). *Linear regression analysis*. Wiley, New Jersey.

Seidell, J. C., Bakker, C., and van der Kooy, K. (1990). Imaging techniques for measuring adipose-tissue distribution–a comparison between computed tomography and 1.5-t magnetic resonance. *The American journal of clinical nutrition*, 51(6):953–957.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Simon, R. and Altman, D. G. (1994). Statistical aspects of prognostic factor studies in oncology. *British journal of cancer*, 69(6):979.

Smalley, K. J., Knerr, A. N., Kendrick, Z. V., Colliver, J. A., and Owen, O. E. (1990). Reassessment of body mass indices. *The American journal of clinical nutrition*, 52(3):405–408.

Snijder, M., Visser, M., Dekker, J., Seidell, J., Fuerst, T., Tylavsky, F., Cauley, J., Lang, T., Nevitt, M., and Harris, T. B. (2002). The prediction of visceral fat by dual-energy x-ray absorptiometry in the elderly: a comparison with computed tomography and anthropometry. *International journal of obesity*, 26(7):984.

Spiegelman, D., Israel, R. G., Bouchard, C., and Willett, W. C. (1992). Absolute fat mass, percent body fat, and body-fat distribution: which is the real determinant of blood pressure and serum glucose? *The American Journal of Clinical Nutrition*, 55(6):1033–1044.

Srivastava, S. and Chen, L. (2009). Comparison between the stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis. In *BMC proceedings*, volume 3, page S21. BioMed Central.

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.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 44–47.

Sun, G.-W., Shook, T. L., and Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of clinical epidemiology*, 49(8):907–916.

Swainson, M. G., Batterham, A. M., Tsakirides, C., Rutherford, Z. H., and Hind, K. (2017). Prediction of whole-body fat percentage and visceral adipose tissue mass from five anthropometric variables. *PloS one*, 12(5):e0177175.

Swartz, M. D., Robert, K. Y., and Shete, S. (2008). Finding factors influencing risk: comparing variable selection methods applied to logistic regression models of cases and controls. *Statistics in medicine*, 27(29):6158.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Van der Kooy, K., Seidell, J. C., et al. (1993). Techniques for the measurement of visceral fat: a practical guide. *International journal of obesity*, 17:187–187.

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

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.

Wajchenberg, B. L. (2000). Subcutaneous and visceral adipose tissue: their relation to the metabolic syndrome. *Endocrine reviews*, 21(6):697–738.

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

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.

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

Xu, L., Cheng, X., Wang, J., Cao, Q., Sato, T., Wang, M., Zhao, X., and Liang, W. (2011). Comparisons of body-composition prediction accuracy: a study of 2 bioelectric impedance consumer devices in healthy chinese persons using dxa and mri as criteria methods. *Journal of Clinical Densitometry*, 14(4):458–464.

Xu, S. (2007). An empirical bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, 63(2):513–521.

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yazdani, A. and Dunson, D. B. (2015). A hybrid bayesian approach for genome-wide association studies on related individuals. *Bioinformatics*, 31(24):3890–3896.

Yeung, K. Y., Bumgarner, R. E., and Raftery, A. E. (2005). Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10):2394–2402.

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