ADJUSTMENT UNCERTAINTY AND VARIABLE SELECTION IN A BAYESIAN CONTEXT

by

ANDREW HENREY

THESIS

Presented to the Faculty of the Graduate School of

Simon Fraser University

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

June 22 2012

Acknowledgements

This text is dedicated to my friends.

I don't have any friends.

Abstract

Bayesian Model Averaging (BMA) has previously been proposed as a solution to the variable selection problem when there is uncertainty about the true model in regression. Some recent research discusses the drawbacks; specifically, BMA can (and does) give biased parameter estimates in the presence of confounding. This is because BMA is optimized for prediction rather than parameter estimation. Though some newer research attempts to fix the issue of bias under confounding, none of the current algorithms handle either large data sets or survival outcomes. The Approximate Two-phase Bayesian Adjustment for Confounding (ATBAC) algorithm proposed in this paper does both, and we use it on a large medical cohort study called THIN (The Health Improvement Network) to estimate the effect of statins on risk of stroke. We use simulation and some analytical techniques to discuss two main topics in this paper. Firstly, we demonstrate the ability of ATBAC to perform unbiased parameter estimation on survival data while accounting for model uncertainty. Secondly, we discuss when it is, and isn't, helpful to use variable selection techniques in the first place, and find that in some large data sets variable selection for parameter estimation is unnecessary.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A common goal in biomedical observational studies is to estimate the effect of an exposure on an outcome, while controlling for confounding factors. A standard procedure is to construct a model for the outcome (based on both the exposure and some subset of the confounding factors) and to estimate the regression parameters based on the model. This procedure does not account for the variability inherent in deciding which model is correct. The phrase "adjustment uncertainty" is used to describe the uncertainty arising from the decision regarding which variables should be included in the model.

Bayesian Model Averaging (BMA) was proposed as a procedure to account for model uncertainty. The general idea behind BMA is that instead of considering a single model, where some predictors are included and some are excluded, we fit all possible models and average their parameter estimates based on a metric of predictive performance. This procedure performs competitively in out-of-sample prediction comparisons (Yeung et al. 2005).

Despite the apparent merits of the BMA routine, there is an inconvenient drawback to prediction-based metrics: there is no guarantee that BMA will give high weight to models with all the confounders. Instead, BMA chooses variables that maximizes predictive performance. This issue arises when covariates are correlated with the exposure, which is (by definition) present in studies that control for confounding factors. Crainiceanu et al. (2008) provide evidence that BMA does not necessarily provide sensible estimates of the exposure effect under confounding. They instead proposed a novel method called Structured Estimation under Adjustment Uncertainty (STEADy) that estimates the exposure effect without bias.

The STEADy routine is a frequentist approach that proceeds in two stages. First, it fits a model to identify the predictors of exposure. After identifying these confounders, a model that fits the outcome based on all the other variables is run. The second model is guaranteed to include the exposure and the confounders related to exposure (found in step 1), and variable selection is done on the remaining confounders. In the presence of confounding, regular BMA does not do as well as STEADy at estimating the exposure effect (Crainiceanu et al. 2008).

A Bayesian version of the STEADy procedure is proposed in Wang et al. (2012). Wang et al. (2012) provide two methods, called Bayesian Adjustment for Confounding (BAC) and Two-phase Bayesian Adjustment for Confounding (TBAC). Both methods involve two models: one that uses confounders to predict exposure, and one that uses both exposure and confounders to predict outcome. Their methods are truly Bayesian in the sense that they use a Markov Chain to calculate the posterior distribution.

This paper proposes a new Bayesian solution to the adjustment uncertainty problem that deals more readily with large samples than the approach used by Wang et al. (2012). The disadvantage of Wang et al.'s algorithms is that they are too slow to handle larger data sets. Instead, we use the same approximation as Raftery et al. (1997) to the posterior, specifically that for large data sets the posterior is proportional to the Bayesian Information Criterion (BIC). This approximation makes it possible to handle data sets of greater size than the previous algorithms. We call our new method ATBAC, which stands for Approximate Two-phase Bayesian Adjustment for Confounding.

We apply our new ATBAC algorithm to The Health Improvement Network (THIN) dataset, which contains information on statin users from Britain. We study the association between statin use and risk of stroke, adjusting for confounders. We observe, on a set of about 90,000 patients, several covariates including whether or not they took statins, their age, gender, alcohol usage, other drugs they take, and some additional information. It is unclear initially which variables are related to the exposure (statin usage), or the outcome (stroke), so it seems reasonable that this is a problem of model uncertainty. Since clinically

we are interested in the unconfounded effect of statin on stroke, we claim that this is a problem that fits into the adjustment uncertainty paradigm.

In Chapter 2, we will look at a technical overview of several existing variable selection methods, and see how the new algorithm fits in. In Chapter 3, we discuss the THIN data, and see how the new algorithm performs on it. In Chapter 4, we present some simulations that discuss an underdeveloped area in variable selection. We also show via simulation that the novel Approximate Two-phase Bayesian Adjustment for Confounding (ATBAC) algorithm gives reasonable results when the distribution of the response is a survival outcome instead of a normal random variable. We finish with some discussion and concluding remarks in Chapter 5.

Chapter 2

Method

For the purposes of this paper, we will consider the variable selection problem on P covariates and a single response. Without loss of generality, we exclude the interaction terms, so overall we consider 2^P models for the response. Furthermore, we assume that there is a single covariate, denoted X, whose effect on the response Y we are interested in estimating. We call this X the exposure.

One obvious method to consider is simply fitting every model to the data, and choosing whichever model provides the best predictive performance. Ignoring for the moment that this is infeasible for even moderate P, there is another potential issue: what if there are several models with very similar predictive performance, but different regression parameter estimates? Making inference on a single model alone, when the model is chosen from the data, has been argued to be risky (Draper 1995).

Using the same notation as Wang et al. (2012), we begin by considering a model for estimating the effect of an exposure, X, on an outcome Y. We begin by discussing the same approach as Wang et. al (2012), using the context of simultaneous regression models with two equations: one equation relates the covariates, denoted by U, to the exposure X, and the second relates the covariates U and X to the outcome, denoted by Y. In each equation, the potential confounders U are included or excluded according to a vector of indicators $\alpha^x \in \{0,1\}$ and $\alpha^y \in \{0,1\}$. We let $\alpha^x_m = 1$ (or $\alpha^y_m = 1$) whenever U_m is included in the exposure (or outcome) model. We have a set of P potential confounders, denoted by $U = U_1, U_2 ... U_P$, identified because they possibly affect Y. If these variables are related to both X and Y, it is important to the estimate of X that they are included in the model on Y. We write the true model as

$$E(X_i|U_i) = \sum_{m=1}^{P} {\alpha_m}^X {\delta_m}^{\alpha^X} U_{im}$$

$$E(Y_i|X_i, U_i) = \beta^{\alpha^Y} X_i + \sum_{m=1}^{P} {\alpha_m}^Y {\delta_m}^{\alpha^Y} U_{im}.$$

We denote the true value of the interesting coefficient by β, and the remaining coefficients by δ. The relationship between the confounders and X is expressed in the $\delta^{\alpha^X} s$ and the relationship between the confounders and Y is expressed in the $\delta^{\alpha^Y} s$. The interpretation of the δs changes depending on which variables are included in the model, but their estimates are not of interest.

At this point, the treatment of the models diverges depending on the algorithm chosen. What follows is an overview of variable selection algorithms in chronological order.

2.1 Review of Current Methodology for Variable Selection in Regression

2.1.1 No Model Averaging or Variable Selection Whatsoever

This algorithm is remarkably straightforward: decide on a model for the outcome, and fit *all* the covariates to it, for example using OLS. When the number of sample points N is lower than the number of covariates P, the algorithm cannot be run. Supposing that the algorithm can be run, fitting irrelevant covariates to the outcome results in over-fitting, and the result is poor out-of-sample prediction when compared to variable selection algorithms. Essentially, if a covariate is unrelated to the response and it is given a non-zero parameter estimate, then as that covariate changes, the model's estimate of the truth will be even more variable than before. Furthermore, fitting additional covariates causes a loss of degrees of freedom, which in turn creates extra variability in the estimate of the parameter of interest. The extent of these issues will be discussed later.

There are two interesting advantages to not doing model selection. The first reason is that model selection needs to evaluate up to 2^P models, which could be quite high even for moderate P. Conversely, ignoring variable selection in regression simply fits 1 model, which takes significantly less computational time. On large data sets with high N and moderate P, it might be quite a bit faster to not do model selection. The other reason one might want to not use model selection is that standard regression, using maximum likelihood, gives consistent parameter estimates for each variable. These are very desirable properties for inference. A model selection technique does not provide a guarantee on which parameter estimates are interpretable.

2.1.2 Stepwise Variable Selection

Stepwise regression does not consider the model for exposure at all. Stepwise regression uses a heuristic (like the AIC) to evaluate the quality of a model, typically imposing a penalty for extra parameters that do not add to predictive performance. Usually no special treatment is given to the exposure variable, although it can be forced into the model. In one implementation of stepwise (known as "backward selection"), the algorithm starts with $\alpha_m^y = 1 \forall m$. Then stepwise uses a greedy algorithm to maximize the AIC by removing parameters one by one from the model that do not add to predictive performance, until such a step is impossible. The algorithm then terminates.

Backward selection, and other stepwise variants, are not desirable for several reasons. Firstly, stepwise tends to under-report model uncertainty, and subsequently under-reports the estimate of the variance in parameters. Inference is made assuming the chosen model is the true model. Secondly, stepwise gives no treatment at all to the exposure model, which is necessary to control for confounding. Since determining the effect of X on Y is the whole point of doing the study, stepwise regression is not a desirable solution to this problem.

The key benefit to stepwise is that it does fairly fast model selection, and does a reasonable job of eliminating completely useless covariates. It is also fully automated. Some variable selection techniques (e.g. Raftery's BMA) do backward stepwise to eliminate the

least likely candidate variables, and do model averaging once P is more moderate.

2.1.3 Bayesian Model Averaging

Traditional BMA does not consider the exposure model. In the truest form of BMA, all 2^P models have their likelihoods evaluated; the posterior likelihood is found by multiplying the likelihood by the prior (although typically the prior is set to be uniform across the models, so that term drops out). The posterior likelihoods are then normalized into posterior probabilities.

Let the posterior probability of any model j be denoted as $P(\alpha^y_{m_j}|Y)$, and the data be denoted by D. Then, for a given covariate U_m

$$E[\beta_m|D] = \sum_{i=1}^{2^P} E(\beta_m|D, \alpha^y)P(\alpha^y|D).$$

Essentially the posterior mean of a covariate is given as the weighted average of its estimate over all possible models, weighted by the posterior probability of the model, and multiplied by an indicator that is 0 if the covariate is excluded from the model.

This approach is fine if interpreting the covariates is not of interest; that is, if high predictive performance is the only goal. However, if the actual covariate effects are of interest also, then BMA falls short.

To illustrate conceptually the problem with BMA, consider the following scenario: Two variables, X and U, are related to the outcome Y. We are interested in the effect of X on Y. Suppose also the effect of X on Y is somewhat larger than the effect of U on Y, and finally suppose that X and U are fairly heavily correlated. BMA fits 4 models to the data: one has no covariates, one includes X but not U, one includes U but not X, and one includes both X and U. The models that do not include X have relatively poor predictive performance, so BMA does not place a lot of posterior probability on them. The two remaining models have nearly equal predictive performance.

The issue with BMA lies in the model that includes X but not U. Since X and U are correlated fairly strongly, the posterior estimate of the mean of X will be biased in the

direction of the effect of U (e.g. if U was positively correlated with Y, then X would be biased upwards of the true value of X). This is an example of confounding. Clearly the model with both X and U included is unbiased for the effect of X, and averaging these two models (the one with X, and the one with both X and U) leaves us with a biased estimate of the effect of X on Y.

Various simulations have been done in recent papers (Crainiceanu et al. 2008 , Wang et al. 2012) to show that this problem exists in reality.

2.1.4 Structured Estimation under Adjustment Uncertainty (STEADy)

Crainiceanu et al. (2008) observed the following: if the algorithm choses a model that includes all the variables related to X and the variables related to Y, the the effect of X on Y will be unbiased. This eliminates confounding.

First, we define some notation: let $\alpha \in 0, 1$ be, as before, an indicator vector denoting whether the a particular variable is in the model. We also partition the 2^P models into M+1 subclasses, called orbits by Crainiceanu et al. (2008), where each class has a fixed number of covariates (e.g. the 0^{th} orbit has the single model with no covariates, the 2^{nd} orbit has $\binom{P}{2}$ models, etc.). For each orbit, define the dominant model to be the model with the highest likelihood. The set of P+1 dominant models comprises the dominant model class. As before, we let X be the covariate of interest, U be potential confounders, and Y be the outcome. Again, we wish to estimate the effect of X on Y. Finally, we define the exposure model as the regression on X by the covariates U, and the outcome model to be the regression of X and U onto Y.

The STEADy algorithm proceeds as follows:

1. Using the exposure model (i.e. the model for X), construct the dominant model class.

2. Find the point at which adding extra covariates does not significantly impact the ability to predict X (determined by looking at the differences in the deviance between

dominant models in two consecutive orbits)

3. Let the number of covariates found be denoted as L.

4. All the covariates found related to X (as well as X) are forcibly included into the future model on Y (e.g. in the notation of Wang et al., set $\alpha_m^y = 1$).

5. Using the outcome model, construct the dominant model class. X and any U_m related to X are automatically included; model selection is done on the remaining $P - L$ covariates

6. Find the final model in the same way as step 2.

The STEADy algorithm was groundbreaking in the sense that it was the first asymptotically unbiased model selection algorithm in the presence of confounding. Nonetheless, we suggest there is still some room for improvement for two reasons. Firstly, the algorithm is not Bayesian; it does not incorporate any prior information. Secondly, finding the model with the highest likelihood in the i^{th} orbit with 100% probability requires evaluation of all $\binom{P}{i}$ models. The STEADy algorithm does not evaluate all the models, rather, it runs a randomized algorithm to evaluate a small subset of them and chooses the one among those that has the highest likelihood. While this improves the runtime drastically, the algorithm remains to be fairly slow even on mid-size datasets.

2.2 New Contribution: Approximate Two-Phase Bayesian Adjustment for Confounding

Having concluded a discussion of the methods that have been published to date, we move on to the new method proposed in this paper. Similar in spirit to Wang et al.'s Two-phase Bayesian Adjustment for Confounding, this algorithm (ATBAC) provides a significant increase in speed to Wang's algorithm while maintaining similar results in simulation.

ATBAC proceeds by running BMA twice, and setting the prior probabilities on the second algorithm in accordance with the findings of the first. The first stage of ATBAC uses BMA to regress the confounders U onto the exposure X. Any confounders with non-trivial probability of being related to X ($> 5\%$) are included in the second model by setting their prior probability to 1.0. During the second stage of ATBAC, we run BMA again using both X and U to predict Y, but some of the U's (and X) are forcibly included in the models. Wang et al. (2012) mention this as a possible interpretation of TBAC; combining this idea with the BIC approximation of Raftery et al. (1997) gives a practical Bayesian solution for large data sets.

BMA provides biased posterior parameter estimates when variables that are related to both X and Y are not included in the model. The goal of ATBAC is to eliminate this possibility. ATBAC's first model regresses the covariates U onto the exposure X using regular Bayesian Model Averaging. Recall that the posterior probability of a variable being included is the sum of the posterior probability of all the models that it is included in. From the first model of ATBAC, we get for each variable a posterior probability that it is related to the exposure. We use this to construct the prior for the second model.

In the linear case, the model for X is given as above:

$$E(X_i) = \sum_{m=1}^{P} {\alpha_m}^X {\delta_m}^{\alpha^X} U_{im}$$

The key difference between the method of Wang et al. (2012) and this method is implementation; Wang uses a full Bayesian approach, whereas ours uses the same approximation as Raftery et al (1997), using BIC weights to approximate the appropriate integral instead of a MCMC routine. This approximation improves the run-time to make it applicable to larger data sets.

2.3 The Need for Model Averaging

We return to the problem of fitting a model to a data set with response Y, explanatory variable of interest X, and alternate covariates U. As the number of data points approaches infinity, and with a finite set of models, eventually one model will contain all the posterior mass. Certainly in this case averaging over the set of models weighted by posterior probability is a waste of time. It is not an unreasonable extension to claim that with very large N, the vast majority of the posterior mass might still be placed on one model, and again model averaging would be of little use. In other words, we don't need variable selection when $N >> P$.

One (computationally) inexpensive approach in this case is to assume that the true model is the full model. This obviously has some drawbacks, which will be discussed shortly. The advantage of fitting a full model (and ignoring variable selection) to the data is twofold - firstly, it produces the most unbiased estimate this data set can provide (obviously there is no accounting for unmeasured confounding), and secondly, it is computationally cheap. On large data sets, having answers in seconds instead of days is surely a desirable quality.

Fitting a full model to a data set without thought to variable selection is not a popular idea, for several reasons. Fitting unnecessary variables causes overfit, and higher variability in prediction. This is true, of course - however, this paper does not consider prediction a goal. This paper is concerned with unbiased parameter estimation. In the THIN data that we will discuss in Chapter 3, we will see that the research question discusses estimation of the effect of statin on health, which is not a prediction based goal.

A second complaint regarding fitting the full model is that the variance of the parameter estimates is higher than necessary, on account of fitting irrelevant covariates. This is correct, but what is not stated is the magnitude. For high N and a moderate number of covariates, adding a covariate with no relation to either Y or X does not actually have a particularly sizable effect. We will see a numerical assessment of this feature in the next section.

A final issue with fitting models without variable selection is that if the covariate of

interest, X, is highly related to a confounder, U, then the variance of the estimate of the coefficient on X will significantly increase. For unbiased prediction, however, including this covariate in the outcome model is essential. Without this parameter in the outcome model, we have no way of knowing whether it was related to Y, and if it was, whether it was through its correlation to X or by its own accord. Consequently, algorithms that don't include a model for both exposure and outcome provide biased estimates of the effect of exposure on outcome. An algorithm that includes both models is forced to include such correlated covariates every time, and thus fitting a full model does not differ from any other unbiased estimation procedure in this respect.

Chapter 3

Results

3.1 THIN Data

Statins are the most popular cholesterol lowering drug in the world (Rutisheiser, 2006). Several randomized experiments have confirmed their ability to prevent cardiovascular disease. In this data analysis, we study the relationship between statins and stroke risk while adjusting for confounders such as age, medical history, and prior drug use.

The Health Improvement Network database (THIN) dataset is a collection of medical records from Great Britian. These records intend to amalgamate prescribing and diagnostic information for various drugs, including statins. Information attained is anonymous. Death rates and prescription rates are similar to other sources (Smeeth et al. 2008). The source population is the 5.5 million people registered at the participating general practices between January 1995 and December 2006.

The target population in the THIN data is the same as in McCandless et al. (2010), who analyse UK patients age 65 and over who are registered at one of the 303 general practices that contributed data to THIN during the aforementioned timeframe. Following Smeeth et al. (2008), statin users are defined as anybody who starts taking statins after 1995 and is registered for at least the previous year at one of the participating general practices. McCandless et al. (2010) use a matched design where each statin user is matched up with up to 5 non-users; we use the same 90324 patients as McCandless et al (2010).

One purpose of the study was to obtain an estimate of the effect of a statin on the expected time until a person has a stroke. Ideally the analysis would show an odds ratio for statins less than 1, indicating that statins decrease a human's risk of stroke. This would

corroborate previous research.

A short table of summary statistics of the THIN data is presented in table 3.1

Table 3.1: Summary statistics of the THIN data

Sample Size	90324
Number Exposed	19274
People with Strokes	3713
Censored Observations	86611
Number of Ties	86222
Mean of Exposed Time to Stroke (days)	1646
Mean of Non-Exposed Time to Stroke (days)	1964

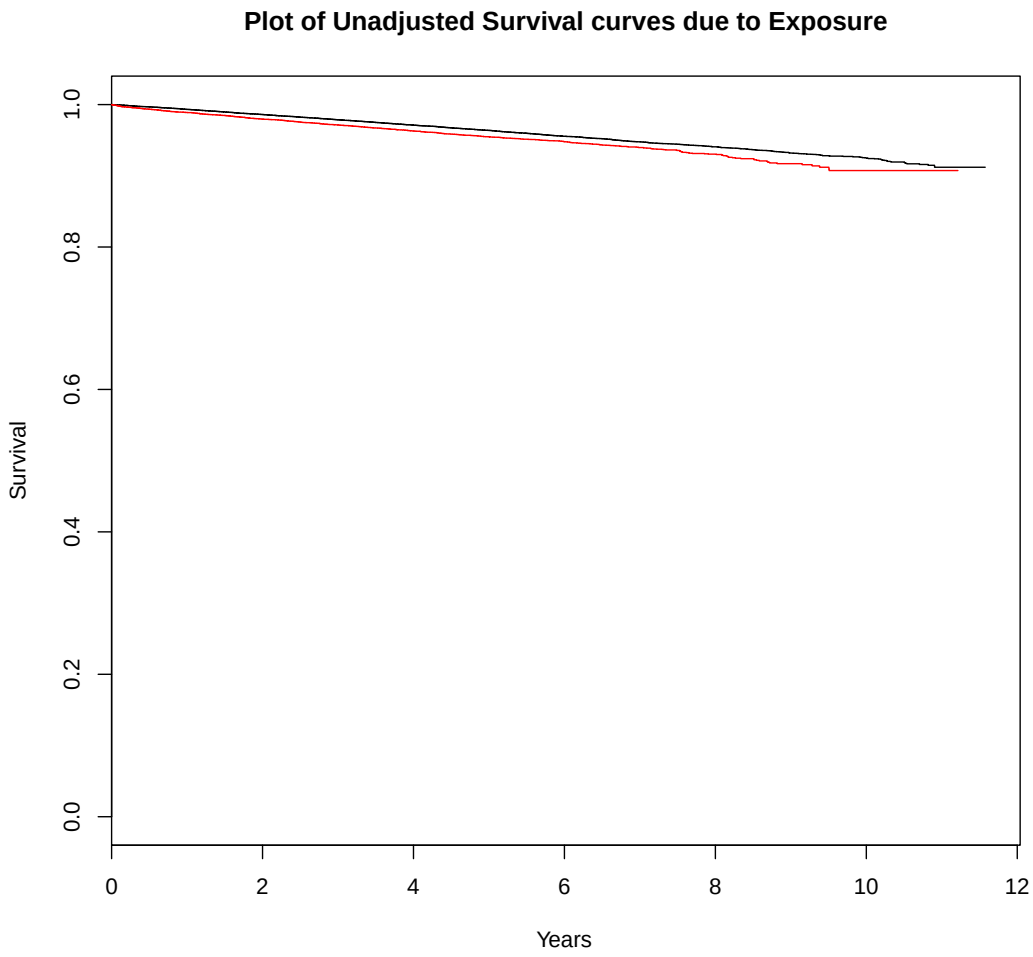
Some other characteristics of the study population are given in table 3.2 :

Table 3.2: Other characteristics of the study population

	Unexposed (Statin Non-user)	Exposed (Statin User)
Number of Patients	71050	19274
Average Age	72.57	73.09
Number of Males	30848 (43.4%)	8277 (42.9%)
Number of Females	40202 (56.6%)	10997 (57.1%)
Average BMI	26.29	26.95
Diagnosed with Diabetes	8953 (12.6%)	4867 (25.3%)
Diagnosed with Dementia	423 (0.6%)	69 (0.4%)
Diagnosed with Cancer	5479 (7.7%)	1271 (6.6%)
Prior use of Steriods	5826 (8.2%)	1492 (7.7%)
Prior use of Aspirin	12650 (17.8%)	11296 (58.6%)
Prior use of Diuretics	24287 (34.2%)	10048 (52.1%)

A plot of the unadjusted survival curves is found in Figure 3.1. We see the red line, denoting the statin users, with survival probabilities materially lower than the black line describing non-statin users. A log rank test for difference between the two curves gives p < 0.001. Interestingly, the results indicate that statins are associated with higher risk of stroke.

Figure 3.1: Unadjusted survival curves



If we fit a Cox proportional hazards model to the unadjusted data (regressing statin exposure onto survival time), we see a log-odds ratio of about 1.2. This is not a very surprising result because doctors prescribe statins to only the sickest patients (e.g. those who

have already been diagnosed with diabetes, see Table 3.2). This hazard ratio is obviously biased from confounding - we expect the true hazard ratio to be less than 1.0. It is not much of a stretch to claim that additional covariates need to be included in order to give an unbiased estimate of the effect. The question now is: which variables? For reference, a complete list of the availiable variables can be found in Table 3.3.

The classical approach to variable selection is to simply not bother—instead, add all the variables into the model. This approach can produce high variance estimates when there is high collinearity between the variables, and indeed, this data set has some strong correlations between the covariates. These correlations exist partly because a lot of the variables are measuring an idea of "overall health". For example, coronary heart disease, cardiovascular disease, peripheral vascular disease, atherosclerosis, atrial fibrilation and heart failure will all be highly correlated because they are all measures of the patient's heart. It is concerning that adding all these variables will cause an increase in variance estimates and subsequently it will be more difficult to detect a statin effect. How should we proceed with the analysis? (Raftery et al. 1997) proposed using Model Averaging to account for our uncertainty about which model is correct, and proceed using his Bayesian framework. The results of this solution (and our new, refined solution) appear below.

If we ignore model uncertainty for the time being and simply add all the covariates into the model and use a regular Cox proportional hazards model, we attain the parameter estimates in Table 3.3. For each covariate, we see the estimated log-odds ratio, the odds ratio, and the estimated standard error of the log-odds ratio.

Table 3.3: Estimates obtained from a full Cox Regression of THIN data with all covariates in the model

Covariate	Coef	Exp(Coef)	SE	Comorbidities	Coef	Exp(Coef)	SE	Drugs	Coef	Exp(Coef)	SE
Statin Exposure	-0.023	0.977	0.049	Diabetes*	0.256	1.292	0.044	Hormone Repl. Therapy	-0.097	0.908	0.118
Age*	0.053	1.054	0.003	Coronary Heart Disease	0.151	1.163	0.087	Antipsychotics	0.046	1.047	0.144
Gender*	-0.299	0.741	0.037	Cardiovascular Disease*	0.607	1.835	0.075	Antidepressants*	0.19	1.21	0.058
BMI	-0.002	0.998	0.004	Peripheral Vasc. Disease*	0.32	1.376	0.078	Steroids	0.013	1.014	0.063
Low SES*	0.137	1.147	0.041	Atherosclerosis	-0.154	0.858	0.087	Fibrates	-0.1	0.905	0.385
Rate of statin*	0.081	1.084	0.015	Dementia*	0.924	2.518	0.163	Cytochromes	-0.087	0.916	0.077
Consult Physician	0.006	1.006	0.05	Cancer	0.035	1.035	0.063	Lipid Lowering Agents	0.129	1.138	0.37
Current Smoker*	0.156	1.169	0.042	Atrial Fibrilation*	0.266	1.305	0.085	Nitrates	-0.088	0.916	0.055
Former Smoker*	-0.089	0.915	0.042	Heart Failure	-0.001	0.999	0.074	Aspirin*	0.14	1.15	0.043
High Alcohol*	0.304	1.355	0.148	Hepatic Illness	0.109	1.115	0.278	Beta Blockers*	0.101	1.106	0.042
Moderate Alcohol	0.062	1.064	0.056	Renal Disease	0.097	1.101	0.16	Calcium Channel Blockers	0.053	1.054	0.043
Former Alcohol*	0.37	1.448	0.161	Thyroid Disease	-0.1	0.905	0.074	Diuretics*	0.094	1.098	0.045
				Hyperlipidemia	0.009	1.009	0.061	Inotropes*	0.196	1.216	0.087
				Hypertension*	0.156	1.169	0.039	Anticoagulants*	-0.167	0.846	0.081
								Antihypertensives*	-0.115	0.891	0.044
								Cardiovascular Drugs	0.045	1.046	0.056

* Indicates the Frequentist p-value < 0.05

Next we fit three methods to the THIN data: full Cox, BMA, and ATBAC. We provide the estimates of the statin exposure effect from the three algorithms that can be used to fit survival data in table 3.4. The STEADy algorithm was not used here because the algorithm given by the authors does not work under the survival context.

Table 3.4: Estimates of Exposure(statin) effect on stroke from 3 different methods

Algorithm	Coef	Exp(Coef)	Estimated Standard Error
BMA	-0.020	0.980	0.045
ATBAC	-0.023	0.977	0.049
Full COX	-0.023	0.977	0.049

It is worth mentioning that ATBAC gives the same estimate for exposure that a full Cox model does. BMA, however, gives a different estimate. Admittedly here the estimates are not particularly far apart, and neither is significantly different from zero.

It is worth mentioning that a full table of regression parameter estimates from ATBAC is not presented. This is quite deliberate. ATBAC is designed to be unbiased and interpretable for exactly one parameter: the exposure effect. ATBAC does not provide unbiased or interpretable estimates of the other parameters. It is for this reason that ATBAC is sometimes more efficient than fitting the true model, and yet more useful than BMA. BMA gives zero interpretable parameters (the price paid by having such good prediction), fitting the full model gives the worst prediction but each parameter estimate is unbiased. ATBAC (in addition to STEADy and Wang et al.'s algorithms) is interpretable for just exposure.

Indeed, the conclusion of this analysis is twofold: a full Cox model appears to perform similarly to ATBAC when $N >> P$, and (according to this analysis) statins do not have a statistically significant effect at $\alpha = 0.05$.

3.2 The need for variable selection when $N >> P$

The similarity of the results for different methods in table 3.4 questions the need for variable selection when $N >> P$. Let us suppose the true model for a response Y is:

$$Y = \beta_0 X + \beta_1 U_1 + \beta_2 U_2 + \epsilon$$

$$\beta_0 = \beta_1 = \beta_2 = 0.1$$

$$\epsilon \sim N(0, 1).$$

The covariates X, U_1, and U_2 are all independently generated from N(0,1). Furthermore, suppose we add P "useless" covariates that have no contribution to the response at all and are also uncorrelated to X, for a total of P+3 columns of data. As before, the parameter estimate of X is of interest.

The best model to use for unbiased parameter estimation is the true model; no variable selection method can attain a better estimate of the coefficient on X than this model. Obviously we don't know the true model, but it makes for a convenient "gold standard"—algorithms that do nearly as well as the true model are better than ones that do not.

Let us compare two models; the first is the true model, where

$$Y = \beta_0 + \beta_1 X + \beta_2 U_1 + \beta_3 U_2 + \epsilon$$

and the second is a model with P unnecessary covariates $U = (U_3, U_2....U_{P+2})$:

$$Y = \beta_0 + \beta_1 X + \beta_2 U_1 + \beta_3 U_2 + \beta^T U + \epsilon$$

The unnecessary covariates are generated from N(0,1) with no correlation between them. We are interested in the ratio of the variances of the parameter estimates of X in these two models. When we fit both models to a set of data, their parameter estimates of β_1 are unbiased, so the expected value of the ratio of variances reduces to a ratio of the MSEs. We are interested in seeing how bad it gets when we fit the full model instead of the true model.

We begin by constructing the expected value of the variance of the true model. Without loss of generality, assume the first covariate is the one of interest; then the variance of its parameter estimate (when using Ordinary Least Squares) is given by the top left entry of

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

An alternative way to write this variance is using the Variance Inflation Factor, denoted as $\frac{1}{1-R^2}$:

$$Var(\beta) = \frac{\sigma^2}{N} \frac{1}{1 - R^2}$$

Where R^2 is the correlation coefficient of the regression of the covariates (not including X) onto X. We denote the variance estimates of the two models by $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_2)$, where model 1 is the true model and model 2 contains unnecessary covariates.

$$\begin{aligned}
\frac{Var(\hat{\beta}_1)}{Var(\hat{\beta}_2)} &= \frac{\frac{\sigma^2}{N} \frac{1}{1-{R_1}^2}}{\frac{\sigma^2}{N} \frac{1}{1-{R_2}^2}} \\
&= \frac{1 - {R_2}^2}{1 - {R_1}^2} \\
&= \frac{\frac{SSResidual_2}{SSTotal_2}}{\frac{SSResidual_1}{SSTotal_1}}
\end{aligned}$$

Since the data set is the same, and therefore the vector of X is the same in both cases, the SSTotal will be identical, and we can cancel this term out. Observe that the SSResidual is based on a regression of 2+P covariates, in addition to an intercept term.

$$\begin{aligned}
\frac{Var(\hat{\beta}_1)}{Var(\hat{\beta}_2)} &= \frac{SSResidual_2}{SSResidual_1} \\
&= \frac{{\chi^2}_{N-P-3}}{{\chi^2}_{N-3}}
\end{aligned}$$

For reasonably large N, this reduces to:

$$\frac{Var(\hat{\beta}_1)}{Var(\hat{\beta}_2)} \approx \frac{N - P - 3}{N - 3}$$

Or more generally speaking:

$$\frac{Var(\hat{\beta}_1)}{Var(\hat{\beta}_2)} \approx \frac{DF_{Model2} + 1}{DF_{Model1} + 1} \tag{3.1}$$

This result also holds if the covariates U are correlated with each other. The result does not hold if the covariates are correlated with X.

This calculation gives us an idea of the magnitude of the maximum possible effect of using variable selection. When discussing the precision of effect estimation, the result of removing K variables results in K additional degrees of freedom to estimate a parameter. In the case where N is significantly greater than P, a single additional degree of freedom provides a negligible benefit, and perhaps even P additional degrees of freedom does not provide a serious improvement in precision. If this is the case, then variable selection is not needed - fitting the full model is quite sufficient.

Chapter 4

Simulations

4.1 The need for variable selection when the number of data points (N) greatly exceeds the number of covariates(P)

We have previously shown theoretically that variable selection might not be relevant for every choice of N (number of data points) and P (number of covariates). Specifically, we showed through (3.1) that in linear regression, when useless covariates are uncorrelated with the exposure, the number of covariates divided by the number of data points gives the expected percentage increase in MSE when adding P irrelevant covariates to the true model. Now, we turn to simulation to get an idea of the benefits of variable selection, and a better grasp on when it is, and isn't, effective.

We begin with a simulation that should demonstrate whether (3.1). We choose a particular model for Y that is taken from (Crainiceanu et al. 2008). The true model for Y is a function of 3 variables: X, U_1, U_2, with no intercept.

$$Y = \beta_0 X + \beta_1 U_1 + \beta_2 U_2 + \epsilon$$

$$\epsilon \sim N(0, 1)$$

$$\beta_0 = \beta_1 = \beta_2 = 0.1$$

The correlation between X and U_1 is 0.7. The variable of interest is X, as such, we are interested in the estimates of the parameter value β_0.

In the first simulation, we will fit two models to each generated data set. The first model will regress these 3 covariates onto Y, and will be called the "true model". The second model that is fit, called the "full model", is given an additional P covariates (denoted as U) that have no relation to X, Y, U_1, or U_2. This simulation should show how the MSE breaks down when we add numerous useless covariates to a regression model.

This first simulation uses N = 100, with P varying from 5 to 90 in increments of 5. For each P, 10,000 runs are used. We fit both models during each run. We then compute the MSE of the parameter estimates of β_0 for both the full and true model. We then take the MSE of the full model and divide by the MSE of the true model to create a summary table, shown below. Beside these values, we put the theoretical values we should obtain from the closed form of this ratio.

Table 4.1: Looking at the relationship between P, N, and the ratio of MSE : N = 100 (10,000 reps)

P	$\frac{DF_{TrueModel}+1}{DF_{FullModel}+1}$	Simulated MSE Ratio
5	1.05	1.05
10	1.12	1.11
15	1.19	1.18
20	1.26	1.27
25	1.35	1.37
30	1.45	1.47
35	1.57	1.61
40	1.71	1.73
45	1.88	1.91
50	2.09	2.11
55	2.34	2.37
60	2.67	2.72
65	3.10	3.21
70	3.69	3.67
75	4.57	4.84
80	6.00	5.98
85	8.73	9.45
90	16.00	18.95

The results given in Table 4.1 demonstrate the validity of equation 3.1. We expect these columns to be nearly identical (except for simulation error), and we indeed see that the columns are very close. This is a helpful justification that the formula is correct. It is interesting that the quantities are more similar when $\frac{P}{N}$ is small than when $\frac{P}{N}$ is large.

Now that we have compared the full model to the true model, we are interested in seeing how well ATBAC performs as a variable selection technique. This time, for each simulation run we will compute a parameter estimate from the true model and from a model that uses the ATBAC procedure. As before, we will compute the ratio of the MSE of the parameter estimates of ATBAC divided by the MSE from the true model. We would expect that these ratios would be lower than the ratios attained when fitting the full model, because ATBAC should provide some reduction in variance.

This table uses N = 100, and P from 10 to 40 in increments of 10. The theoretical values from the ratio of the full model to the true model are shown as well, for comparison. This small table was produced because the ATBAC algorithm, although fast relative to its counterparts, is still not very speedy, and 10,000 reps takes a long time.

Table 4.2: ATBAC vs Full model : N = 100 (10,000 reps)

P	ATBAC	$\frac{DF_{TrueModel}+1}{DF_{FullModel}+1}$
10	1.02	1.12
20	1.12	1.26
30	1.26	1.45
40	1.52	1.71

It is fairly apparent from this table that ATBAC provides a substantial decrease in the MSE ratio. We claim that ATBAC is worth using for N = 100 if we have a set of variables to consider that are uncorrelated with anything.

4.2 Simulations of various variable selection techniques under confounding

We have seen in the previous section that when N is small and Y is a linear function of X, ATBAC seems to provide an improvement over the full model in terms of MSE when estimating a parameter of interest. We now look at a variety of simulations that more closely mimic the THIN data. We make two changes to the previous simulations: firstly, we use N = 1000 instead of N = 100, and secondly, we will use a few additional methods to fit the models to the data. These simulations are very similar to the simulations done by Crainiceanu et al. (2008).

We present three sets of simulations in this section. First, we mimic the results of Crainiceanu et al. (2008) by constructing a simulation under a linear regression framework and estimating the parameter of interest, B_0. We find (as they did) that BMA gives a biased estimate of B_0, but STEADy and ATBAC appear unbiased. We then extend the simulation framework to more closely mimic the THIN data in the second simulation, with X as a binary variable and Y as a binary variable. Finally in the third simulation, we change Y again - this time Y is a survival time with moderate censoring, and X remains binary. We feel that the simulation framework resembles the THIN data, disregarding correlations among the U's and between the X's and U's.

It is worth mentioning that neither STEADy nor Wang's BAC/TBAC algorithms can process survival data. ATBAC, on the other hand, handles it easily because BMA software already exists for survival outcomes. One of the chief purposes of this section is to show that the ATBAC algorithm is unbiased under a survival outcome. It is unclear whether the results from linear regression are fully analogous when X is not marginally normal, or when Y has a distribution other than normal. These simulations should demonstrate the ability of the ATBAC algorithm to provide unbiased results under various distributions of Y (even when Y is not normal).

In the following sections, we will demonstrate the ability of 5 variable selection algo-

rithms to estimate the parameter of interest: BMA (Bayesian Model Averaging, with the exposure effect forced into the model), ATBAC (Approximate Two-phase Bayesian Adjustment for Confounding - the novel method proposed in this paper), STEADy (Crainiceanu et al., 2008), Stepwise Regression (Backwards stepwise), and finally, we also fit the full model (with every covariate) for comparison.

4.2.1 Simulation 1: Normal Y, Normal X

The first simulation gives Y a normal distribution as follows:

$$Y = \beta_0 X + \beta_1 U_1 + \beta_2 U_2 + \epsilon$$

$$\epsilon \sim N(0, 1)$$

$$\beta_0 = \beta_1 = \beta_2 = 0.1$$

and the true model for X is

$$X = 0.5U_1 + 0.5\epsilon$$

$$\epsilon \sim N(0, 1).$$

In keeping with Crainiceanu et al. (2008), there is about 70% correlation between X and U_1. In addition to these variables, there are an additional 49 variables that are unrelated to X or U (in other words, completely uncorrelated with everything). As mentioned previously, $N = 1000$. We use 500 trials to generate the following results:

From left to right, the columns in Table 4.3 are the average simulated bias, the average reported standard error (from the algorithm used to fit the data), the simulated average standard error, the simulated mean squared error, and the percentage of the 95% confidence intervals that actually cover the true value. The CI's were generated from each run using the parameter estimate $+/-$ twice the standard error estimate. We draw the attention of the reader to a couple key points: firstly, neither BMA nor Stepwise regression have coverage

Table 4.3: Five different algorithms fitting linear models with correlation; N = 1000

	Bias	SEE	SSE	MSE	95% coverage probability
BMA	0.0387	0.0354	0.0447	0.0035	0.674
ATBAC	0.0005	0.0446	0.0433	0.0019	0.940
STEADy	0.0024	0.0445	0.0465	0.0022	0.948
Stepwise	0.0499	0.0390	0.0530	0.0052	0.640
Full	-0.0003	0.0455	0.0470	0.0022	0.942

probabilities anywhere close to 95%, and secondly, both BMA and stepwise regression appear to give biased prediction. These are unsurprising conclusions that corroborate Crainiceanu et al. (2008). ATBAC, STEADy, and the full model are all unbiased, and the difference in MSE between the 3 algorithms does not appear materially different for N = 1000, P = 50 in the linear setting.

4.2.2 Simulation 2: Binary Y, Binary X

We now consider a model for a binomial outcome. The covariate of interest and the response are both binomial, to simulate an interpretation of the statin data (e.g. let $Y = 1$ denote a stroke within a given timeframe). The model for X is given by

$$\begin{aligned} X &\sim Bernoulli(p) \\ p &= \tfrac{\exp(U_1)}{1+\exp(U_1)}. \end{aligned}$$

The model for Y is given by

$$\begin{aligned} Y &\sim Bernoulli(p) \\ p &= \tfrac{\exp(\beta_0 X+\beta_1 U_1+\beta_2 U_2)}{1+\exp(\beta_0 X+\beta_1 U_1+\beta_2 U_2)} \\ \beta_0 &= \beta_1 = \beta_2 = 0.1 \end{aligned}$$

As before we ran all 5 algorithms on the simulated data to estimate B_0. Because these models take longer to run than in the linear setting, we used only 100 simulations. Since

the correlation between X and U_1 is somewhat lower than in the linear regression case, we would expect BMA and stepwise to perform slightly better than in simulation 1, but they should still give biased parameter estimates. These results are given in 4.4.

Table 4.4: Five different algorithms fitting binomial models with correlation; N = 1000, 100 simulations

	Bias	SEE	SSE	MSE	95% coverage probability
BMA	0.0395	0.1323	0.1647	0.0292	0.880
ATBAC	0.0010	0.1426	0.1351	0.0179	0.940
STEADy	0.0208	0.1418	0.1451	0.0213	0.970
Stepwise	0.0881	0.0945	0.1152	0.0208	0.840
Full	-0.0110	0.1472	0.1504	0.0227	0.940

It is pleasing to note that we see the same conclusions as in simulation 1. When Y and X are both binary variables, BMA and Backwards stepwise provide biased prediction with confidence intervals that are too narrow. ATBAC, STEADy, and fitting the full model do not appear to be very different in terms of overall performance.

4.2.3 Simulation 3: Survival Y, Binary X

This final simulation intends to convince the reader that the ATBAC algorithm is unbiased for parameter estimation when Y is a survival time (perhaps with censoring). This simulation is of particular interest because it closely mimics the THIN data, as the THIN data also has binary X and survival Y. It is worth mentioning again that neither STEADy nor BAC/TBAC are able to handle survival data. Consequently, it is necessary to provide some simulated evidence that this style of algorithm is able to provide unbiased parameter estimation. We intend to show that ATBAC is the first algorithm to use both an exposure model and an outcome model to accurately estimate paramters under the survival framework.

For this simulation we omit STEADy, as it cannot conveniently handle survival data. We fit the remaining algorithm to the survival outcome.

The model for Y is given by

$$\begin{aligned} Y &\sim Exponential(\mu) \\ \mu &= 0.1X + 0.1U_1 + 0.1U_2. \end{aligned}$$

We have censoring times C according to a similar distribution:

$$\begin{aligned} C &\sim Exponential(\mu) \\ \mu &= 2 + 0.1X + 0.1U_1 + 0.1U_2. \end{aligned}$$

The censoring is fairly moderate; about 12% of the data are censored.

Table 4.5: Two different algorithms fitting survival models with correlation; N = 1000, 250 simulations

	Bias	SEE	SSE	MSE	95% coverage probability
BMA	0.0124	0.0754	0.0884	0.0083	0.896
ATBAC	-0.0046	0.0752	0.0804	0.0064	0.944
Full	0.0224	0.0783	0.0876	0.0077	0.924
Stepwise	-0.0212	0.0778	0.0933	0.0089	0.933
STEADy	—	—	—	—	—

— Indicates that this algorithm does not apply to the problem

The results for this simulation are given in Table 4.5. Although we see that BMA exhibits lower coverage probabilities than we would like, we observe that it is not nearly as bad as the linear regression simuation. It is worth mentioning, however, that this is mostly due to a poor model for X. Since the correlation between X and U_1 is not quite as strong under this setup as it was under linear regression, BMA gives comparable performance to the other algorithms. When the correlation between X and U_1 is increased, some informal simulations indicate that BMA performs much worse than its counterparts. The full

model is, as usual, not very different from ATBAC - we expect the difference in coverage probability between these two algorithms to be mostly due to chance. With that said, the simulation results suggest that the full model gives estimates that are somewhat too high. A further simulation of 100,000 runs was run to determine the magnitude of the bias of the full model; it turns out that a better estimate is 5% bias.

It is worth noting that ATBAC, STEADy, and the full model display similar results in all 3 simulations. Even though we see that the full model can give biased parameter estimation under the non-linear scenario, the magnitude of the bias is not large (and it decreases with N, unlike error due to confounding). All three algorithms are consistent for the true parameter and provide the best coverage probabilities, and additionally have the lowest MSE. It certainly seems that in the linear case, when $N >> P$, N is of reasonable magnitude, and the correlations between covariates are not extreme, that variable selection is not worth using. These simulations do not present here a material difference in MSE, Coverage probability, or Bias between a full model and model that used variable selection in the linear case, and we see only slight bias in the non-linear case. Indeed, we have shown analytically that the full model in the linear regression case, with N = 1000 and P = 50, has only 5% lower precision than the true model. No variable selection technique that provides unbiased parameter estimation could hope to improve the MSE in parameter estimation by more than 5%.

Chapter 5

Discussion

We have demonstrated a Bayesian variable selection and model averaging technique that also gives unbiased parameter estimates. We will now discuss further where variable selection is worthwhile in the first place.

First, it is worth recalling the purpose of variable selection and model averaging. Variable selection is used to "free up" degrees of freedom from the data by eliminating variables that are worthless. While a more traditional method like BMA used predictive performance of the final model to judge a variable's "worthiness", and STEADy/BAC/ATBAC use two simultaneous models to determine this, the idea remains the same. Model averaging is used when it is unclear from the data as to whether a variable makes a positive impact; averaging over both possibilities weighted by posterior probability is an excellent choice.

When N is large, and P is not terribly large compared to N, neither variable selection nor model averaging is necessary. Picking up K degrees of freedom improves the precision of the effect estimates by a multiplicative factor of K/N, which is not large (since P is not large relative to N, and K is surely no larger than P). Furthermore, model averaging is best used when it is unclear from the data whether a parameter has a material effect on the predictive performance of the final model. When N is nearly infinite, it is unlikely that there are any parameters that we are uncertain about - either they will have clear effects on the predictive performance, or the magnitude of their effects is sufficiently negligible that their additional impact on the predictive performance is marginal, and excluding them is not a problem.

Crainiceanu et al. (2008) propose a simulation study that demonstrates why BMA does not give unbiased parameter estimation. What they neglect to discuss is that the STEADy

algorithm does not materially outperform a regular OLS regression (that simply puts all the terms into the model) in terms of effect estimation in the same simulation. Our additional simulations in the non-linear case demonstrate the need to

We have already seen that the maximum possible gain from a variable selection routine in the linear setting in terms of precision is roughly P/N. We conclude with the recommendation that for large data sets with moderate P, say, $N > 1000$ and $N >= 20P$, that variable selection and model averaging is not worth using. There are, actually, two good reasons not to do variable selection or model averaging: firstly, using OLS allows for interpretation of ALL the covariates, instead of just one, and secondly, OLS is a significantly faster algorithm.

5.1 Clinical Variable Selection

It may be uncomfortable for a researcher to simply put all the available variables into a model. It is generally considered poor practice to put a slew of variables that measure approximately the same thing into a model and assume that the results will work out well. We now discuss the context of this problem in terms of parameter estimation for $N >> P$.

Firstly, we will make some assumptions and definitions to make the problem easier to discuss. As before, we consider the regression of an interesting variable X onto the response Y, in the presence of unknown confounding. We measure some alternative covariates U, which may or may not be related to each other, X, and Y. We are concerned that adding a subset of the covariates U may be undesirable for the performance of the algorithm.

One interpretation of this subset might be a set of factors that measure socio-economic status, such as the household income, maximum education level of either parent, or the cost of the family's most expensive car. We assume that these are correlated because they are all measures of the same thing, but we don't know how well they estimate socio-economic status. Furthermore, we make the assumption that socio-economic status is truly related to X and Y, and these auxiliary variables (such as household income) are related to X and

Y only through socio-economic status. In other words, the model that we fit to the data is incorrect.

The important part of the argument, however, is from a practical standpoint. The researcher cannot measure socio-economic status directly, so he is never given the choice of including that variable into the model. His choice, instead, is to choose which, if any, of the covariates measuring socio-economic status should be included.

Let us suppose that a covariate was supposed to measure socio-economic status, but instead it had no relationship to it whatsoever. The cost of including this variable into the model would be 1 degree of freedom. The claim is that when doing variable selection with $N >> P$, this is very cheap. Since this is essentially the worst that can happen if we include a variable, we conclude that adding the variables is not expensive. The advantage of adding variables is that they will either reduce bias in the exposure effect estimate, or at worst keep it the same. The conclusion of adding every covariate when $N >> P$ seems logical.

5.2 A brief comment on N

We have provided a proof that in the case where both X and Y are marginally normal random variables, N is not too small, and N is at least twenty times what P, then from equation 3.1 removing every covariate P gives at most a 5% increase in efficiency. However, if Y is not normal, then the number of rows in the dataset divided by P will not give similar results.

The results generated from the analytical procedure do not strongly resemble simulated results when the response follows a distribution other than normal. If the response is, for example, a survival time, then even without censoring we would expect that the analytical formula would underestimate the "cost" of adding extra unnecessary covariates. Once we add in censoring, it is even more difficult to determine how much an extra covariate costs in terms of additional standard error. We can, however, provide a couple general comments

on the approximate cost under a response that follows a survival curve. Firstly, we would expect the cost to be higher than in the linear regression case, and secondly, censored points contribute less information than uncensored points. Data sets with large amounts of censoring but a large number of rows may give a misleading impression of data strength. As an example, the THIN dataset has 90344 observations, but 86611 are censored.

Future work, via simulation study, could be done on this problem. It would be of practical use to be able to estimate the ability of a data set to absorb irrelevant parameters without too much additional standard error. Finding an approximate formula for the ratio of the MSE of the parameter estimates between the true model and the full model in the non-normal case would be helpful indeed.

5.3 Concluding Remarks

Traditional model averaging falls short when unbiased parameter estimation of regression coefficients is of interest. We have implemented a Bayesian routine that gives unbiased parameter estimates of a single covariate, and the routine is fast enough to handle the THIN dataset (STEADy, BAC, and TBAC as given could not run it in a reasonable amount of time). We further observe that Model Averaging (or any variable selection technique) is only helpful towards the estimation of an interesting regression parameter when N and P are of about the same magnitude; when N is more than 20 times P, the model for the response is linear, and N is in any way large, model averaging provides no material advantage over OLS.

When we apply both the improved BMA routine (ATBAC) and OLS to the THIN dataset, the results are the same. Using this algorithm, the THIN dataset detects no significant impact of statin on stroke. It is reasonable that other unbiased methods (e.g. using Propensity scores) will estimate the exposure effect differently. Additionally, there are probably further unmeasured confounders that may still bias our estimate of exposure.

References

[1] Crainiceanu, C.M., Dominici, F. and Parmigiani, G. (2008) "Adjustment uncertainty in effect estimation," *Biometrika*, 95 , 635–651.

[2] Draper, D. (1995) "Assessment and Propagation of Model Uncertainty" *Journal of the Royal Statistical Society*, Ser. B, 57, 45-97.

[3] Madigan, D.M. 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*, 85, 1335–1346.

[4] McCandless, L.C., Douglas, I.J., Evans, S.J., and Smeeth, L. (2010) "Cutting Feedback in Bayesian Regression Adjustment for the Propensity Score," *The International Journal of Biostatistics*, Vol 6: Issue 2, Article 16.

[5] Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997) "Bayesian model averaging for regression models," *Journal of the American Statistical Association*, 92, 179-191.

[6] Rutishauser, J. (2006) "The role of statins in clinical medicine LDL-cholesterol lowering and beyond," *Swiss medical weekly* , 136, 41–49.

[7] Smeeth, L., Douglas, I., Hall, A.J.,Hubbard, R., Evans, S. (2008) "Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials," *British Journal of Clinical Pharmacology*, 67, 99–109.

[8] Wang, C., Parmigiani, G. and Dominici, F. (2012) "Bayesian Effect Estimation Accounting for Adjustment Uncertainty," *Biometrics*.

[9] Yeung, K., Bumgarner, R., Raftery, A. (2005) "Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, 21, 2394–2402.