MODEL AVERAGING: METHODS AND APPLICATIONS

MODEL AVERAGING: METHODS AND APPLICATIONS

By CAMILLE SIMARDONE , M.A., B.Com.

A Thesis Submitted to the School of Graduate Studies In Partial Fulfillment of the

Requirements for

the Degree Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2021) Hamilton, Ontario (Economics)

TITLE: Model Averaging: Methods and Applications

AUTHOR: Camille Simardone,

M.A. (University of Toronto); B.Com. (Univeristy of Toronto Mississauga)

ADVISOR: Jeffrey S. Racine

NUMBER OF PAGES: xxi , 134

Lay Abstract

This thesis focuses on model averaging, a leading approach for handling model uncertainty, which is the likelihood that one's econometric model is incorrectly specified. I examine the performance of model averaging compared to conventional econometric methods and to more recent machine learning algorithms in simulations and applied settings, and show how easily model averaging can be applied to empirical problems in economics. This thesis makes a number of contributions to the literature. First, I focus on frequentist model averaging instead of Bayesian model averaging, which has been studied more extensively. Second, I use model averaging in empirical problems, such as estimating the returns to education and using model averaging with COVID-19 data. Third, I compare model averaging to machine learning, which is becoming more widely used in economics. Finally, I focus attention on different approaches for constructing the set of candidate models for model averaging, an important yet often overlooked step.

Abstract

This thesis focuses on a leading approach for handling model uncertainty: model averaging. I examine the performance of model averaging compared to conventional econometric methods and to more recent machine learning algorithms, and demonstrate how model averaging can be applied to empirical problems in economics. It comprises of three chapters.

Chapter 1 evaluates the relative performance of frequentist model averaging (FMA) to individual models, model selection, and three popular machine learning algorithms – bagging, boosting, and the post-lasso – in terms of their mean squared error (MSE). I find that model averaging performs well compared to these other methods in Monte Carlo simulations in the presence of model uncertainty. Additionally, using the National Longitudinal Survey, I use each method to estimate returns to education to demonstrate how easily model averaging can be adopted by empirical economists, with a novel emphasis on the set of candidate models that are averaged. This chapter makes three contributions: focusing on FMA rather than the more popular Bayesian model averaging; examining FMA compared to machine learning algorithms; and providing an illustrative application of FMA to empirical labour economics.

Chapter 2 expands on Chapter 1 by investigating different approaches for constructing a set of candidate models to be used in model averaging – an important, yet often overlooked step. Ideally, the candidate model set should balance model complexity, breadth, and computational efficiency. Three promising approaches – model screening, recursive partitioning-based algorithms, and methods that average over nonparametric models – are discussed and their relative performance in terms of MSE is assessed via simulations. Addi-

tionally, certain heuristics necessary for empirical researchers to employ the recommended approach for constructing the candidate model set in their own work are described in detail.

Chapter 3 applies the methods discussed in depth in earlier chapters to currently timely microdata. I use model selection, model averaging, and the lasso along with data from the Canadian Labour Force Survey to determine which method is best suited for assessing the impacts of the COVID-19 pandemic on the employment of parents with young children in Canada. I compare each model and method using classification metrics, including correct classification rates and receiver operating characteristic curves. I find that the models selected by model selection and model averaging and the lasso model perform better in terms of classification compared to the simpler parametric model specifications that have recently appeared in the literature, which suggests that empirical researchers should consider statistical methods for the choice of model rather than relying on ad hoc selection. Additionally, I estimate the marginal effect of sex on the probability of being employed and find that the results differ in magnitude across models in an economically important way, as these results could affect policies for post-pandemic recovery.

Acknowledgements

First and foremost, I would like to thank my supervisor, Jeff Racine, for his guidance and feedback over the past five years. His encouragement – and humour – kept me motivated through some of the most challenging years of my life, which included the 2019-2021 COVID-19 pandemic.

I would also like to thank my two other committee members, Mike Veall and Youngki Shin, for their support and comments, which no doubt improved the quality of my work.

Thanks to the amazing professors that have contributed to my growth as a scholar and economist: Arthur Sweetman, Steve Jones, Alok Johri, and Gordon Anderson.

Thank you to my parents, Noell and Joe, my brother, Aidan, and my extended family for believing in me and reminding me that I can do anything I put my mind to.

Thanks and hugs to Sennah, Bretton, Shannon, and Laura. Thank you also to Beth Moore. Without these wonderful people, I would not have gotten through this.

Finally, thank you to my husband, Chris, for his unconditional love and support, and to my darling pets, Ouzo and Poppy, for keeping me sane.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

- AIC: Akaike Information Criterion
- AME: Average Marginal Effect
- ANOVA: Analysis of Variance
- ARM: Adaptive Regression by Mixing
- ARMS: Adaptive Regression by Mixing with Model Screening
- AUC: Area Under the Curve
- BIC: Bayesian Information Criterion
- BMA: Bayesian Model Averaging
- CART: Classification and Regression Tree
- CCR: Correct Classification Rate
- CDF: Cumulative Distribution Function
- CV: Cross Validation
- DGP: Data Generating Process
- FIC: Focused Information Criterion
- FMA: Frequentist Model Averaging
- GCV: Generalized Cross Validation
- GDP: Gross Domestic Product
- GUM: General Unrestricted Model
- JMA: Jackknife Model Averaging
- LFS: Labour Force Survey
- LOF: Lack-of-Fit

- OECD: Organization for Economic Co-operation and Development
- OLS: Ordinary Least Squares
- MA: Model Averaging
- MAB: Model Average Bagging
- MAFE: Mean Absolute Forecast Error
- MARF: Model Average Random Forests
- MARS: Multivariate Adaptive Regression Splines
- MMA: Mallows' Model Average Criterion
- MSE: Mean Squared Error
- MSFE: Mean Squared Forecast Error
- NLS: National Longitudinal Survey
- NLSY: National Longitudinal Survey of Youth
- PcGets: Automated General-to-Specific
- PSE: Predicted Squared Error
- RF: Random Forests
- ROC: Receiver operating characteristic
- SNR: Signal-to-Noise Ratio
- SSR: Sum of Squared Residuals

Declaration of Academic Achievement

The material in this dissertation is my own research. I conducted all the simulations, empirical analysis, and writing of the manuscripts from 2017 to 2021.

Chapter 1

Handling Model Uncertainty: Model Averaging and Machine Learning Methods for Empirical Problems in Economics

1.1 Introduction

Standard practice in economics typically proceeds as follows: the researcher selects a parametric model, then carries on with estimation and inference asserting that the chosen model could have plausibly generated the data; in other words, as though this model represents the true, unknown data generating process (DGP). Often, the model is chosen out of convenience or convention, that is, with no relation between the asserted model and the data. Although common practice, *model assertion* – the practice of selecting a model in an ad hoc manner – can have serious consequences stemming from the fact that it is impossible to know whether the chosen model is correctly specified or not. Estimates may hinge on the model selected and, when these estimates inform policy, the consequences of ignoring

the uncertainty in the model specification can be severe and unforgiving. Consider the basic assumption underlying simple ordinary least squares (OLS) estimation that the linear additive functional form must mimic the true (unknown) DGP. If this assumption fails, the finite sample properties of the OLS estimator, such as unbiasedness (for the DGP), do not hold and any subsequent inference falls apart. Additionally, valid inference (for the DGP) relies on having the correct model specification. Consequently, ignoring model uncertainty may lead to inference that is overly optimistic and potentially misleading.

Consider the simple simulated illustrative example in Figure 1.1. This simulation generates 1,000 observations for x drawn from the uniform distribution. The DGP is $f(x) = \sin(2\pi x)$. The outcome y is generated as $y = f(x) + \epsilon$, where $\epsilon \sim N(0, 0.5\sigma_{f(x)})$ and $\sigma_{f(x)}$ is the standard deviation of the DGP. Suppose the researcher asserts a naïve specification that is linear in regressors and additive, and uses OLS to estimate the model. It is evident that the linear specification is not a good fit to the data.[1] Suppose, instead, that the researcher uses model averaging. She does not need to make any assumptions about the underlying (unknown) functional form and instead simply averages over a set of, in this case, polynomials of differing degrees, and produces a better fit to the data. In reality, unlike in this example, we do not know the functional form of the DGP. As such, the probability that the model asserted by the researcher is correctly specified may be low and should be addressed.

[1]This is a very simple example used to illustrate, firstly, the consequences of ignoring model uncertainty and asserting a particular functional form and, secondly, the accuracy of model averaging and ease with which it can be implemented. Of course, in practice, some researchers do examine plots of data and the goodness of fit of models before asserting a particular parametric model specification. This is good practice. However, it becomes more complicated when there is more than one regressor. Additionally, given that the model space is dense, the selected parametric model may be misspecified. As such, methods like model averaging may be considered as an alternative to traditional econometric methods.

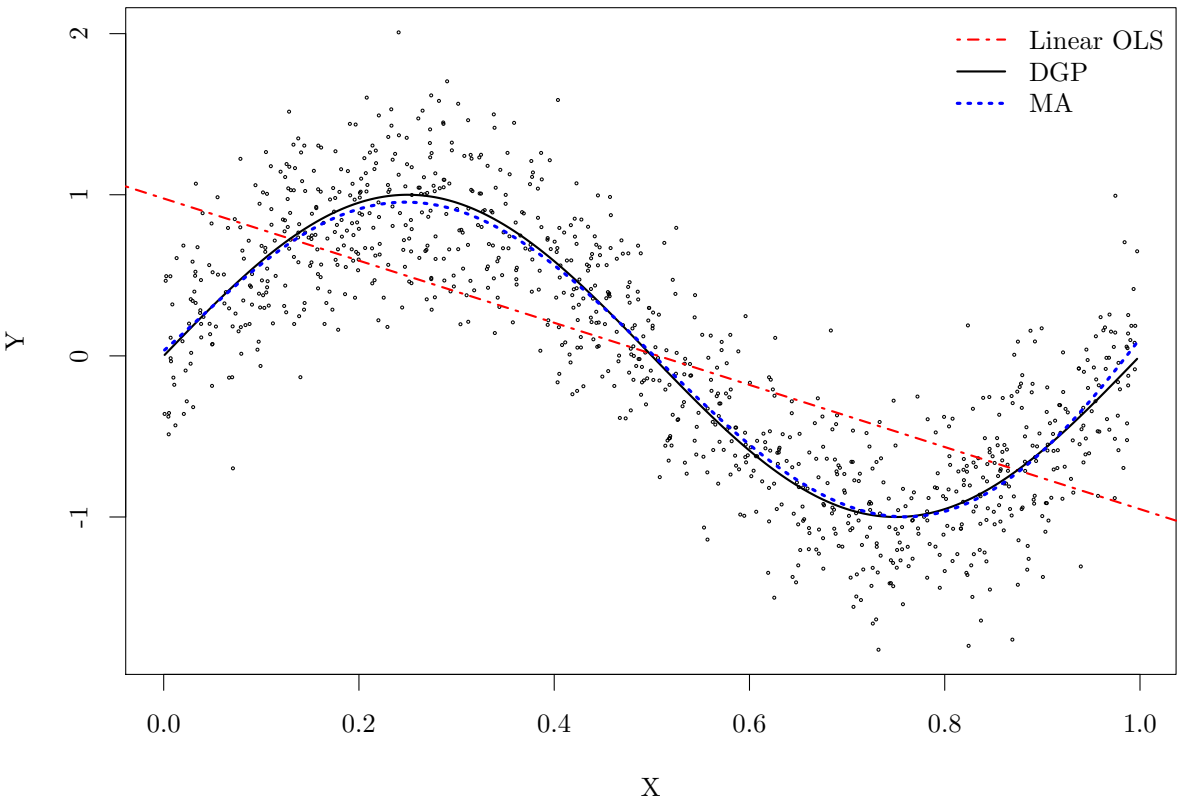

Figure 1.1: Simulation comparing model averaging (MA) and ordinary least squares (OLS) estimation.

Uncertainty in this context exists at two levels: uncertainty regarding the parameter estimates, which is addressed in nearly all economic papers, and uncertainty regarding the model specification, which is often ignored (Moral-Benito, 2015). *Model uncertainty* – the likelihood that one's model is incorrectly specified – is typically disregarded by economic researchers.[2] One approach that is sometimes adopted and acknowledges model uncertainty is *model selection*, which selects the single "best" (in other words, least misspecified) model among a set of candidate models. The model selected as the "best" model is the one that minimizes some criterion, such as the Akaike information criterion (AIC;

[2]Some researchers may attempt to acknowledge model uncertainty by reporting results from multiple model specifications that include a different set of regressors, or that have a different functional form altogether. While this practice is a step in the right direction (away from simply reporting the results of a single preferred model), there still exists the possibility that all of these models are misspecified. As stated in Xie & Lehrer (2017), "Researchers who ignore model uncertainty implicitly assume their selected model is the 'true' one that generated the data", and when this assumption fails, the reported results may not be able to tell us much. Thus, there is a benefit to considering other methods, like model averaging, that explicitly acknowledge model uncertainty.

Akaike, 1970) or the Bayesian information criterion (BIC; Schwarz, 1978). Model selection acknowledges that the selected model is, at best, an approximation to the DGP and is an improvement over model assertion. However, empirical researchers typically entertain only a small number of models and there is no guarantee that the DGP lies within this limited set, especially when the set of candidate models is chosen in an ad hoc manner. *Model averaging*, on the other hand, is a leading approach for handling model uncertainty. Model averaging constructs a weighted average over a set of candidate models. The goal is to reduce estimation variance, while controlling for misspecification bias (Hansen, 2007). While model averaging has been embraced in fields such as statistics, mathematics and biology, it has not been as widely adopted by empirical economists, despite its advantages over conventional methods such as model assertion and model selection (Steel, 2020).

Some advantages of model averaging include better predictive ability than using any single model among the set of candidate models (Hoeting, Madigan, Raftery, & Volinsky, 1999), more robust results compared to model assertion or model selection (Moral-Benito, 2015), broad applicability, fewer assumptions compared to conventional econometric methods, and standard errors that account for the bias that arises from model uncertainty (Tobias & Li, 2004). With regard to the latter, it is important to recognize that results from model assertion and model selection may be overly optimistic since these methods ignore model uncertainty. Model averaging, in contrast, can report standard errors that account for model uncertainty.[3] Some limitations of model averaging include increased computational burden, lack of closed-form solutions for some estimators, and lack of precedent for post-model-average inference. Additionally, one barrier to the adoption of model averaging methods by empirical economists may be that there is no universal standard for constructing the set of candidate models, something that I will address in Section 1.3.

In addition to model averaging, I also assess the relative performance of three machine

[3]When model uncertainty is ignored, the variance of the parameter of interest β is biased: $\text{Var}(\beta|y) = \text{E}_M[\text{Var}(\beta|M,y)] + \text{Var}_M[\text{E}(\beta|M,y)]$. Variance estimates obtained from model selection approximate the first term, but ignore the second. Thus, the variance from model averaging can be larger than those obtained from model assertion or model selection, as it accounts for model uncertainty (Tobias & Li, 2004).

learning algorithms that, like model averaging, acknowledge model uncertainty. These machine learning algorithms are boosting, bagging and the post-lasso. There are very few papers that assess the relative performance of traditional econometric methods, model averaging, and machine learning algorithms. The ones that do use a dataset that combines social media data with film industry data, providing limited external validity of their findings (see Xie & Lehrer, 2017, 2018; and Liu & Xie, 2019). This chapter adds to the literature by assessing the relative performance of a number of machine learning algorithms compared to model assertion, model selection, and model averaging in simulated Monte Carlo experiments that cover a wide range of sample sizes, signal-to-noise ratios, and DGPs.

This chapter can serve as a guide for empirical economists who wish to use model averaging in their own work. First, it highlights the gains in prediction error, as measured by MSE, that model averaging has over conventional econometric methods and machine learning algorithms. Second, it illustrates the ease with which model averaging can be adopted by researchers, with clear steps for using frequentist model averaging and a novel emphasis on how to construct the set of candidate models.

This chapter proceeds as follows. Section 1.2 gives an overview of two of the most popular model averaging methods, Bayesian model averaging and frequentist model averaging. Section 1.3 outlines two approaches for constructing the set of candidate models, a key step in implementing model averaging. Section 1.4 describes three machine learning algorithms – bagging, boosting, and the post-lasso – that, similar to other model selection-based methods, address model uncertainty and can improve predictive performance of an estimator. In Section 1.5, I run a Monte Carlo experiment and rank model averaging, model assertion, model selection, and the machine learning algorithms according to their MSE to evaluate relative performance in a simple single predictor setting. Section 1.6 demonstrates how model averaging can be applied to an empirical problem in labour economics, using the National Longitudinal Survey to estimate returns to education. The Monte Carlo experiment and empirical example show two different applications of model averaging; the

former involves a prediction, whereas the latter estimates a causal relationship. These two exercises exhibit the breadth of applicability of model averaging methods (in contrast to, for example, machine learning algorithms that can only be used for prediction problems). Section 1.7 concludes.

1.2 Model Averaging Methods

Model averaging can be approached from a Bayesian or frequentist perspective. Currently, there exists a comprehensive literature on *Bayesian model averaging* (BMA; see Raftery, Madigan, & Hoeting, 1997; Claeskens & Hjort, 2008; Hoeting et al., 1999). BMA requires priors on parameters and models in order to construct the BMA estimator. Following Hoeting et al. (1999), let $\mathbb{M} = \{M_1, ..., M_K\}$ denote the set of all models under consideration. Given some quantity of interest, Δ, its posterior distribution given observed data D is:

$$\text{Pr}(\Delta|D) = \sum_{k=1}^{K} \text{Pr}(\Delta|M_k, D)\text{Pr}(M_k|D). \tag{1.1}$$

The posterior distribution of Δ not conditioned on a particular model is a weighted average of the posterior distributions under each of the models, $\text{Pr}(\Delta|M_k, D)$, weighted by the posterior model probabilities, $\text{Pr}(M_k|D)$.

Using Bayes' theorem, the posterior probability for model M_k is:

$$\text{Pr}(M_k|D) = \frac{\text{Pr}(D|M_k)\text{Pr}(M_k)}{\sum_{j=1}^{K} \text{Pr}(D|M_j)\text{Pr}(M_j)}, \tag{1.2}$$

where $\text{Pr}(M_k)$ is the prior probability that M_k is the true model and the marginal likelihood of model M_k is given by:

$$\text{Pr}(D|M_k) = \int \text{Pr}(D|\boldsymbol{\theta}_k, M_k)\text{Pr}(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{1.3}$$

where $\boldsymbol{\theta}_k$ is the vector of parameters of model M_k, $\Pr(\boldsymbol{\theta}_k|M_k)$ is the prior density of $\boldsymbol{\theta}_k$ under model M_k, and $\Pr(D|\boldsymbol{\theta}_k, M_k)$ is the likelihood (Hoeting et al., 1999). The calculation of the posterior model probabilities is non-trivial. However, with linear regressions, the calculation can be solved analytically.

One limitation of BMA is that it can be computationally inefficient when there is a large number of models under consideration (e.g. $M = 2^q$, where q represents the number of potential independent variables to be included). Algorithms such as Occam's window (Madigan & Raftery, 1994) and Markov chain Monte Carlo model composition (Madigan, York, & Allard, 1995) can drastically reduce the number of models under consideration, which reduces computation time.

Economists – especially empirical economists – typically operate within a frequentist framework rather than a Bayesian one. For this reason, *frequentist model averaging* (FMA) is quite appealing and will be used exclusively in this chapter. FMA exploits nonparametric principles to construct a combined estimator that is a weighted average of estimators from each model in a set of candidate models. The steps for constructing the FMA estimator are as follows: First, choose a set of M candidate models. The set can be flexible to include, for example, non-linearities in regressors along with interactions. Construction of the set of candidate models is an important component because the FMA estimator will inherit properties from the candidate models. Two approaches for constructing the set of candidate models will be discussed in detail in Section 1.3. Second, solve for the model weights, ω_m, $m = 1, ..., M$, for each model in the set of candidate models using some criterion and, typically, quadratic programming, which turns out to be a straightforward exercise. Some popular criteria include AIC, BIC, the focused information criterion (FIC; Claeskens & Hjort, 2003), Mallows' model average criterion (MMA; Hansen, 2007), and the jackknife model average criterion (JMA; Hansen & Racine, 2012). The choice of criterion for model weights is important because different model weights will result in different asymptotic properties of the FMA estimator (Claeskens & Hjort, 2008). Finally, for some quantity

of interest (e.g. a coefficient estimate, β_j), construct a weighted average over the set of candidate models using the estimated model weights, $\hat{\omega}_m$, and the estimates, $\hat{\beta}_{j,m}$, from each of the candidate models. Some models may receive a weight of 0, but no one model will receive a weight of 1, except possibly if the "true" model lies in the set of candidate models.

While model selection and model assertion rely on having a good approximation to the DGP, model averaging does not require the "true" model – or a good approximation to the DGP – to be included in the set of candidate models. Thus, one would expect model averaging to outperform model selection or model assertion when the DGP is *not* within the set of candidate models. This relaxes the strong assumption on which traditional econometric methods rely, that of "correct" parametric specification.

After obtaining the model average weights, ω_m, $m = 1, ..., M$, the FMA estimator of the regression coefficient (in this case, a scalar and assumed to be common to all models), $\hat{\beta}_{j,FMA}$, is:

$$\hat{\beta}_{j,FMA} = \sum_{m=1}^{M} \omega_m \hat{\beta}_{j,m}, \tag{1.4}$$

where $j = 1, ..., q$ indexes the regression coefficient, $m = 1, ..., M$ indexes the candidate model, $0 \leq \omega_m \leq 1$ and $\sum_{m=1}^{M} \omega_m = 1$.

The construction of the FMA estimator of the coefficient estimate $\hat{\beta}_{j,FMA}$, in particular via summation, follows the common practice of assuming a linear additive model and a constant marginal effect. This common approach has the advantage of highly interpretable models. However, there exist many other functional forms that would result in a marginal effect which is not a scalar but, instead, a function of one or more regressors. For example, consider a simple function that is an additive model with one regressor and includes one additional layer of complexity: $g(x) = \beta_1 + \beta_2 x + \beta_3 x^2$. Assuming that the functional form of $g(x)$ is a linear additive model (i.e. $\beta_1 + \beta_2 x$), it produces a marginal effect that is a scalar

(i.e. β_2). In reality, the marginal effect of x is itself a function of x: $\frac{dg(x)}{dx} = \beta_2 + 2\beta_3 x$. Thus, it would be beneficial to go beyond the common approach of assuming a linear additive model with a constant marginal effect and embrace models that allow for greater model flexibility. The suggested approach would be to obtain marginal effects vectors, which will be functions of regressors, and use these in model averaging. One drawback to this improvement, however, is potentially decreased interpretability.

Buckland, Burnhamn, & Augustin (1997) propose model weights of the following form:

$$\omega_m = \frac{\exp(-I_m/2)}{\sum_{j=1}^{M} \exp(-I_j/2)}, m = 1, ..., M, \tag{1.5}$$

where I_m is the information criterion for model m and the sum in ω_m extends over every model in the set of candidate models. Thus, two models with the same information criterion score (i.e. I_m) will be given the same weight. FMA can handle a large set of candidate models. Currently, however, FMA lacks a comprehensive framework for post-model-average inference (Hansen, 2014).

1.3 Constructing the Set of Candidate Models

Both model averaging and model selection methods require a set of candidate models. For model averaging, it is important to construct the set of candidate models with care, as the model average estimator will inherit properties from the candidate models. Current practice is to write down models that have a common parameter (as detailed in the construction of the FMA estimator of the parameter vector in Section 1.2). For example, Moral-Benito (2015) uses model averaging to estimate the effect of capital punishment in the United States, where the coefficient on a particular regressor (in this case, the execution rate) is common to all models. First, the author compares three parametric models – specifications typically chosen in an ad hoc manner by economic researchers – that attempt to describe the

relationship between the death penalty and murder rate. He finds that the resulting estimates contradict each other – one estimate shows a positive effect of the death penalty on the murder rate, one shows a deterrent (negative) effect, and the final one shows no statistically significant effect – and it is not clear which is the best model among the three. Next, Moral-Benito implements model averaging – three variations of BMA (using different priors) and three alternative weighting schemes for FMA (using different criteria) – using a set of candidate models that is constructed from all possible combinations of the 16 optional control variables, resulting in 2^{16} = 65,536 candidate models. The execution rate, which is the variable of interest, and a constant are included in every model specification. The author finds similar results across all six model average estimates. Furthermore, the results are statistically insignificant. However, due to limitations – such as not accounting for reverse causality – the author states that "[o]ne should interpret these results with caution, and only as an illustration of the usefulness of model averaging techniques summarized in this paper" (Moral-Benito, 2015, p. 63). Even with this ad hoc approach, averaging cannot be expected to perform any worse than any one model from the set of candidate models.

Another approach to constructing the set of candidate models is one that could be adopted widely in the near future. It is nonparametric in nature and follows Racine (2019). Consider any transformation of a regressor as forming a basis. For example, experience and experience^2 are two bases. The candidate models should span a rich set of models in order to capture a wide range of DGPs, which requires a broad set of bases. In the univariate case, bases can be constructed using orthogonal polynomials, Bernstein polynomials, B-splines, or Bezier curves. With more than one regressor, one must consider not only higher-order polynomials and the inclusion or exclusion of regressors, but also the interactions between regressors. As such, multivariate bases could be constructed using additive models (e.g. $y = f(x_1) + f(x_2)$), multivariate Taylor approximations (e.g. $y = f(x_1, x_2) = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) + \frac{1}{2!}\Big[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(y-b)^2\Big] + ...$), or tensor products (e.g. $y = x_1 \otimes x_2$). As a practical matter,

model selection methods can be used to select the basis function type for each candidate model (e.g. additive or tensor). In small sample settings, the maximum dimension must be restricted. Thus, the complexity of the candidate models will be tied to the sample size, n, and some assumed smoothness class.

1.4 Other Approaches for Dealing with Model Uncertainty

Model averaging has been shown to have better predictive performance and can deliver more robust results than methods using a single candidate model, whether obtained through model assertion or model selection (Hansen, 2007; Hoeting et al., 1999). This begs the question: Is there another method that exists that can outperform model averaging? To answer this, I look to the machine learning literature and evaluate three machine learning algorithms that, like other model selection-based methods, acknowledge model uncertainty: bagging (Breiman, 1996), boosting (Breiman, 1996; Freund & Schapire, 1997) and the post-lasso (Belloni & Chernozhukov, 2013).

Bagging (short for *bootstrap aggregating*) is a machine learning algorithm that generates multiple versions of a predictor using bootstrap resamples with replacement, then constructs a single predictor that is an unweighted average over these multiple versions (Breiman, 1996). Note that in the machine learning literature, the term "predictor" is analogous to "fitted values", whereas in economics, "predictor" is often synonymous with "regressor" or "independent variable". Some statistical methods may have improved accuracy through this process of "perturbing and combining" to reduce variance (Breiman, 1998). Suppose we have a random sample $\mathcal{L} = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}$ where $\mathbf{x}_i$ is a p-dimensional vector of independent variables. Suppose we wish to predict y by some predictor (a function or model) denoted $f(\mathbf{x}, \mathcal{L})$. Ideally, we would have a sequence of K random samples $\{\mathcal{L}_k, k = 1, ..., K\}$ each consisting of n independent observations from the same under-

lying distribution as $\mathcal{L}$. We could then use the sequence of random samples to get a better predictor than the single learning set predictor by replacing $f(\mathbf{x}, \mathcal{L})$ by the average of $f(\mathbf{x}, \mathcal{L}_k)$ over k. However, in reality, we usually only have a single random sample. To overcome this limitation, take B bootstrap resamples of size n with replacement from $\mathcal{L}$ to form a sequence of bootstrap resamples $\{\mathcal{L}^{(b)}, b = 1, ..., B\}$. Then, for each bootstrap resample, form a set of bootstrap predictors (i.e. fitted values from each bootstrap resample) $f^{(b)}(\mathbf{x}, \mathcal{L}^{(b)}), b = 1, ..., B$. Finally, take an average over the B bootstrap resamples to create the bagged predictor:

$$f_{\text{bagg}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} f^{(b)}(\mathbf{x}, \mathcal{L}^{(b)}), b = 1, ..., B. \tag{1.6}$$

Unlike model averaging, where the data are fixed and the model changes, in bagging, the model is fixed and the data change in each resample. In order to give this algorithm the opportunity to perform as well as possible, a model selection method such as stepwise AIC or stepwise BIC can be used as a preliminary step to select the model to be bagged. In this way, bagging can be construed to be a model specification exercise.

Some advantages of this approach are that the mean squared error (MSE) of the bagged predictor may be lower than the MSE of the unbagged predictor and bagging may improve the accuracy of unstable predictors, such as subset selection, classification and regression trees (Breiman, 1996). A prediction method is said to be unstable if small perturbations in the data can result in large changes in the predictor. A limitation of bagging is that as stability increases, the bagged predictor may do worse than the unbagged predictor in terms of prediction error, and so bagging works best for unstable predictors.

Boosting, also known as arcing (short for *adaptive resampling and combining*), follows a similar process of perturbing and combining to generate an improved predictor (Breiman, 1998; Freund & Schapire, 1997). Given a random sample $\mathcal{L} = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}$, suppose we wish to predict y using a predictor $f(\mathbf{x}, \mathcal{L})$. The boosting algorithm maintains

an n-dimensional vector of weights $\mathbf{w}^{(b)}$ over the observations which specifies the probability that each observation will be drawn in the next resample. Weights at all times must be non-negative and all weights of the initial weight vector $\mathbf{w}^{(1)}$ must be non-negative and sum to one, i.e. $\sum_{i=1}^{n} w_i^{(1)} = 1$. The initial weights are set equally so that $w_i^{(1)} = 1/n$. For $b = 1, ..., B$, draw a bootstrap resample of size n with replacement from $\mathcal{L}$ using these weights. Each bootstrap resample $\mathcal{L}^{(b)}$ is used to generate a predictor $f^{(b)}(\mathbf{x}, \mathcal{L}^{(b)})$ which, in turn, is used to generate the next weight vector $\mathbf{w}^{(b+1)}$ by increasing the weights for observations that were poorly predicted. After B bootstrap resamples, the final boosted predictor is computed by combining each of the B bootstrap predictors using a weighted average, with higher weight given to more accurate predictors (Freund & Schapire, 1996, 1997):

$$f_{\text{boost}}(\mathbf{x}) = \sum_{b=1}^{B} \mathbf{w}^{(b)} f^{(b)}(\mathbf{x}, \mathcal{L}^{(b)}), \tag{1.7}$$

where $b = 1, ..., B$, $w_i^{(b)} \geq 0 \forall b$, and $\sum_{i=1}^{n} w_i^{(1)} = 1$.

Like bagging, model selection methods such as the stepwise AIC or stepwise BIC can be used before boosting to select the model to be boosted. One difference between boosting and bagging is that boosting resamples with replacement in a way such that the weights on the observations are increased for observations that are poorly predicted, so that these observations are more likely to be drawn in the next resample, whereas bagging simply sets weights on observations equal to $1/n$ for each resample. Another difference is that boosting combines the B bootstrap predictors using a weighted average whereas bagging constructs the final predictor using an unweighted average over the multiple predictors. Boosting has some advantages over bagging, namely that it has been shown to have better performance; both bagging and boosting decrease bias, but boosting reduces variance more than bagging does (Breiman, 1998). Boosting is simple and easy to program, may mitigate over-fitting, and, like bagging, may improve the accuracy of unstable predictors.

However, relative performance depends heavily on the data. Consequently, in the presence of insufficient data, boosting can perform poorly and is sensitive to noisy data and outliers (Freund, Schapire, & Abe, 1999). Many boosting algorithms exist, the most famous being the AdaBoost (short for adaptive boosting) (Freund & Schapire, 1996). A popular choice for boosting regressions is the gradient boost (Friedman, 2001).

The *lasso* is an acronym for *least absolute shrinkage and selection operator* (Tibshirani, 1996). It was originally developed for OLS regression models, but its applicability has been extended to generalized linear regression models, proportional hazards models, and M-estimators, to name a few. The lasso shrinks some coefficients and sets others to zero, essentially performing selection of regressors. The lasso estimator is defined as:

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij}\right)^2 \text{ subject to } \sum_j |\beta_j| \leq t, \tag{1.8}$$

where $i = 1, ..., n$, j indexes the regressor, β represents regression coefficients, and $t \geq 0$ represents the tuning (or penalty) parameter. Selection of the tuning parameter, t, is important as it controls regressor selection as well as how much shrinkage is applied to the coefficients. Cross-validation is commonly used to select the tuning parameter. The advantages of the lasso include highly interpretable models, increased stability under data perturbations, and improved prediction accuracy with relative computational efficiency. However, the non-zero coefficients resulting from the lasso tend to be biased towards zero, due in part to shrinkage (Belloni, Chernozhukov, & Hansen, 2014).

A solution to mitigate this bias is the *post-lasso* (Belloni & Chernozhukov, 2013). The post-lasso allows for the possibility that model selection is not perfect. It has two steps. First, the lasso is used for variable selection by determining which regressors can be dropped. In this way, the post-lasso is effectively a model specification exercise. Second, OLS estimation is performed on the variables that were selected in the first step (i.e. on the variables with non-zero first-step coefficients). The post-lasso is easy to implement and has

smaller bias compared to the lasso (Belloni & Chernozhukov, 2013; Belloni et al., 2014).

1.5 Monte Carlo Experiment

To assess relative performance in a simple one regressor setting, I run a Monte Carlo experiment that compares the performance of model averaging to model assertion, model selection, bagging, boosting, and the post-lasso. The true DGP is chosen to be $f(x) = 1+x^6/\sigma_{x^6}$ (see Appendix Section 1.8.1 for results from a different DGP). The response variable y is chosen to be $y = f(x) + \epsilon$ where $x \sim N(0,1)$ and $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{x^6})$, where $c \in \{0.25, 0.50, 1.0, 2.0\}$ determines the signal-to-noise ratio (SNR). The figures and tables below display the results for $c = 0.50$ (see Appendix Section 1.8.1 for results from changes in the SNR). I take $n = 100$ random draws of x from the normal distribution and conduct 1,000 Monte Carlo replications (see Appendix Section 1.8.1 for results from changes in the sample size). I consider the case where the true DGP is omitted from the set of candidate models. There are $M = 9$ candidate models, which are orthogonal polynomials of order 1 through 5 and 7 through 10. In this simulation, the exact functional form of the DGP is known, unlike in applied settings, where it is impossible to know the exact functional form of the DGP with one hundred percent certainty. Asymptotically, the model averaging and model selection criteria would select the true model from the set of candidate models if it were included in the set. Therefore, in this Monte Carlo experiment, the true or oracle model – a polynomial of order 6 – is omitted from the set of candidate models in order to evaluate the relative performance of model averaging in a setting where the true model is not within the set of candidate models.

R (version 4.0.2) is used throughout for ease of replicability. The following packages are used:

- `quadprog`, "Functions to Solve Quadratic Programming Problems" (version 1.5-8), contains functions to solve quadratic programming problems and is used to solve

for the model average weights,

- `caret`, "Classification and Regression Training" (version 6.0-86), contains functions for training and plotting classification and regression models and is used for boosting and bagging, and

- `hdm`, "High-Dimensional Metrics" (version 0.3.1), allows for the implementation of high-dimensional statistical and econometric methods for estimation and inference and is used for the post-lasso.

The mean squared error (MSE) is used to evaluate the relative performance of each method and is computed as follows:

$$\text{MSE } = \frac{1}{n}\sum_{i=1}^{n}\left(\text{DGP}_i - \hat{y}_i\right)^2, i = 1, ...n, \tag{1.9}$$

where DGP_i is the DGP, which was chosen to be $1 + x^6/\sigma_{x^6}$, and $\hat{y}_i$ are the fitted values from each method.

The criterion used to select the model average weights is Mallows' Model Average Criterion (MMA) (Hansen, 2007). The MMA criterion is defined as follows:

$$C_n(\omega) = \omega'\hat{\mathbf{E}}'\hat{\mathbf{E}}\omega + 2\hat{\sigma}^2 K'\omega, \tag{1.10}$$

where $\hat{\mathbf{E}}$ is the $n \times M$ matrix with columns containing the residual vector from the mth candidate model, $\hat{\sigma}^2$ is the estimated variance from the largest dimensional model, and K is the $M \times 1$ vector of the number of parameters in each model. The MMA criterion is used to solve for the weight vector, $\hat{\omega} = \text{argmin}_w C_n(\omega)$. As mentioned in Section 1.2, this problem can easily be solved using quadratic programming.

Mallows' C_p is used for model selection (Mallows, 1973) and is defined as follows

(Mallows, 1973):

$$C_p = \frac{SSR_p}{\hat{\sigma}^2} + 2p - n, \tag{1.11}$$

where SSR_p is the sum of squared residuals ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$) in the model with p regressors and $\hat{\sigma}^2$ is the estimated variance from the largest dimension model in the set of candidate models. A low C_p value is desirable. Therefore the model with the lowest C_p value in the set of candidate models is selected.

In this simulation, stepwise AIC is used as a preliminary step to select the model to be bagged and boosted for each Monte Carlo replication. A stepwise procedure can be used when there is a large number of nested candidate models by performing stepwise model selection using the Akaike information criterion (AIC; Akaike, 1970). The AIC criterion is defined as follows:

$$\text{AIC} = -2\ln(\hat{L}) + 2p, \tag{1.12}$$

where $\ln(\hat{L})$ is the maximum value of the log-likelihood function of a model with p regressors. The AIC balances goodness of fit (as measured by the log-likelihood) and parsimony (as measured by the penalty for the number of regressors included in the model). A low AIC value is desirable.

A bagged classification and regression tree (CART) is used for bagging, with cross-validation as the resampling method so that there are $B = n$ bootstrap resamples of size $n - 1$. The `train` function from the `caret` package is used.

A boosted linear model is used for boosting, with bootstrapping as the resampling method.The `train` function from the `caret` package is used. The tuning parameters are `mstop`, which sets the number of boosting iterations, and `nu`, which specifies the level of shrinkage. In this simulation, the values of these tuning parameters are selected using a grid search and are `nu = 0.1` and `mstop = 150`. The root mean squared error (RMSE;

$\sqrt{\text{MSE}}$) was used to select the optimal model using the smallest value.

Recall that the selection of the tuning parameter (t in $\sum_j \beta_j \leq t$) is an important step for the post-lasso because the tuning parameter determines both variable selection and the degree of shrinkage. In this simulation, the value of the tuning parameter for the post-lasso is selected using cross-validation and this optimal value is used.

1.5.1 Results

I summarize results using Tukey's box-and-whisker plot (Tukey, 1970).[4] Figure 1.2 shows a box-and-whisker plot of the MSE for each method, each candidate model, and the true model over 1,000 Monte Carlo replications. Some methods and candidate models clearly perform poorly compared to others in terms of MSE, namely bagging, $k = 1$, $k = 2$, and $k = 3$.

[4] A box-and-whisker plot is a nonparametric method of displaying data that offers a graphical overview of the data by summarizing key features such as the median and upper and lower quartiles.

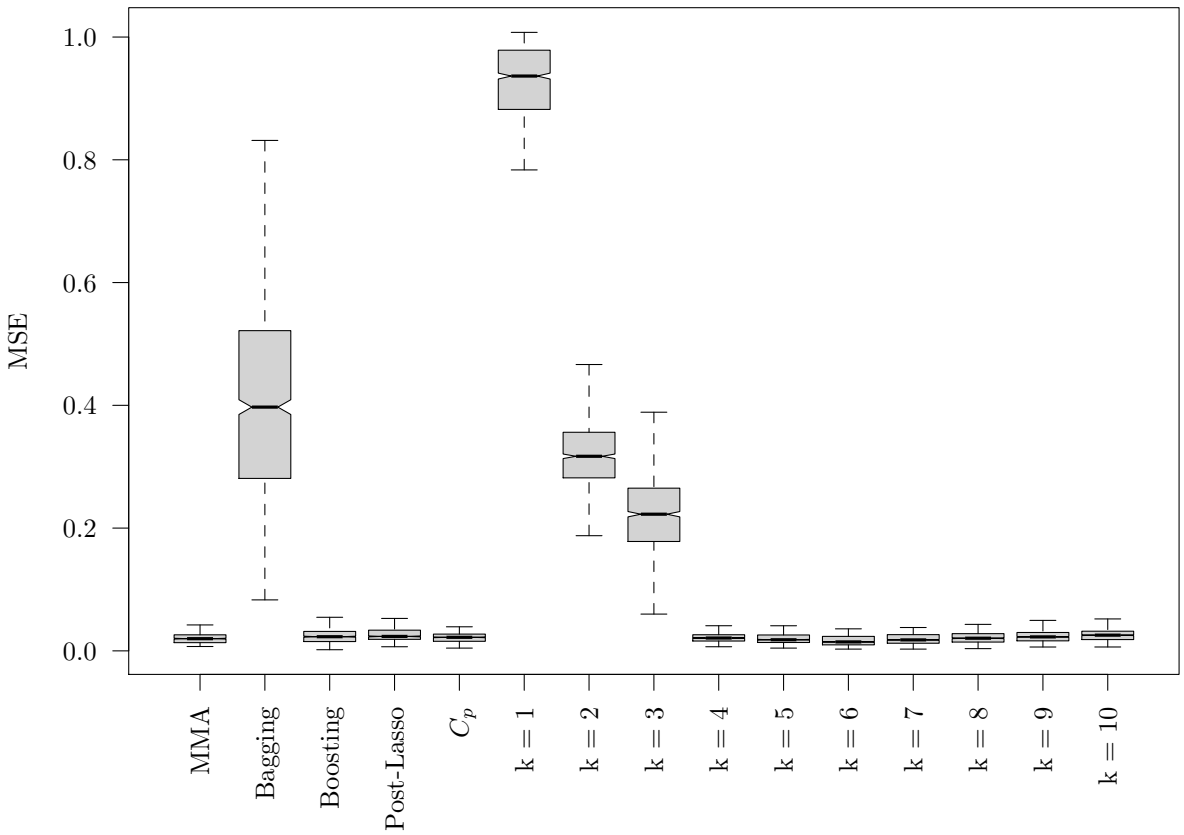

Figure 1.2: Box-and-whisker plot of MSE of each method, each candidate model, and the true model over 1,000 Monte Carlo replications ($k = 6$ is the DGP).

While there are nine candidate models in this simulation (as well as the true DGP, $k = 6$), in reality there is an infinite number of models. The MSE over 1,000 Monte Carlo replications each of model ($k = 1, ..., 10$) represents the case where a researcher always asserts, for example, a linear model (i.e. $k = 1$), even when she encounters 1,000 different datasets. This practice is deterministic in that it uses the same model from replication to replication and it would be extremely naïve to do so in applied settings. However, comparing the mean MSE of each candidate model can give us insight into some of the potential consequences of model assertion in the presence of model uncertainty. Figure 1.3 shows the MSE of each candidate model ($k = 1, ..., 5, 7, ..., 10$) and the true model ($k = 6$) over 1,000 replications. The MSE for $k < 6$ is higher than that of $k = 6$, the true model, (especially $k = 1$, $k = 2$, and $k = 3$) as these models are underspecified and have non-

zero bias. Note that the MSE can be calculated as the sum of bias squared and variance (i.e. MSE$(\hat{f}(x))$ = Bias$(\hat{f}(x))^2$ + Var$(\hat{f}(x))$). For $k > 6$, the bias is zero, which, all else being equal, lowers the MSE. However, these models are overspecified, which increases the variance, contributing to higher MSE. This is why $k = 7$ performs relatively well, as it is only slightly misspecified, and the MSE grows for $k = 8$ to $k = 10$. This can be seen more clearly in Figure 1.4, which gives a closer look at the models that perform relatively well in terms of MSE, namely $k = 4, ..., 10$.

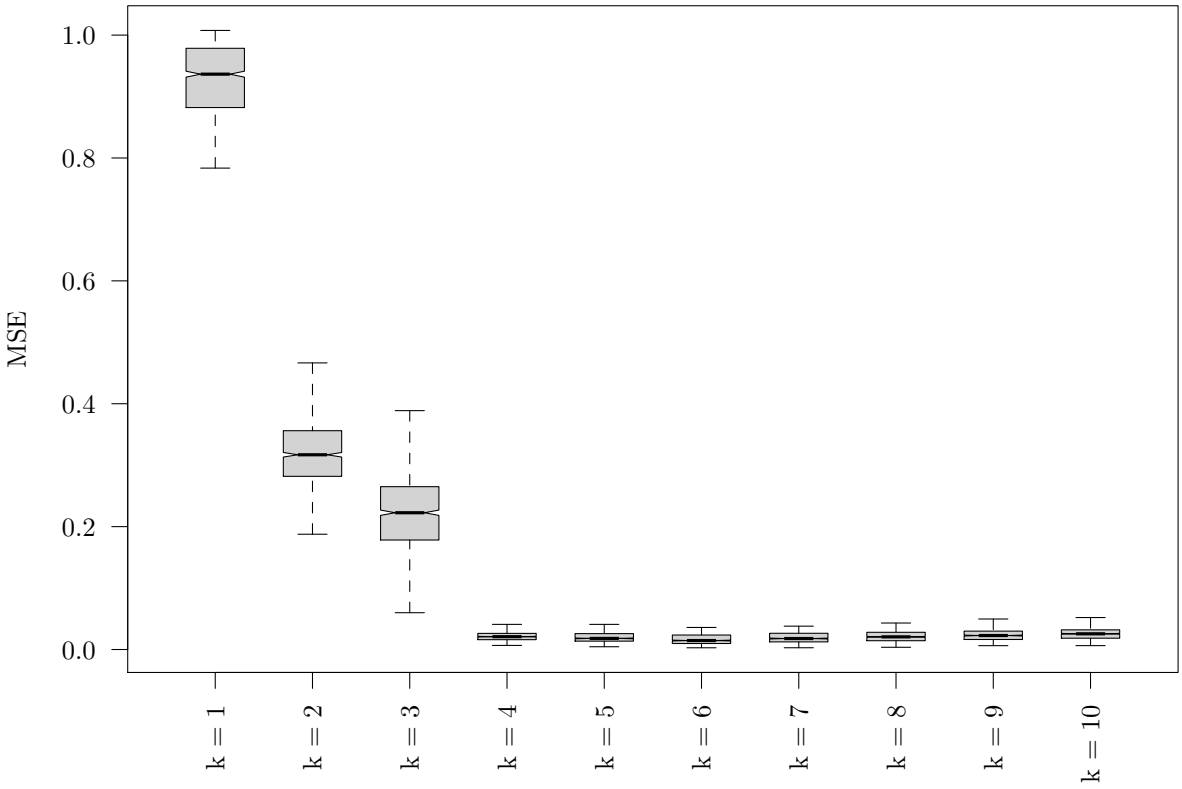

Figure 1.3: Box-and-whisker plot of MSE of each candidate model and the true model over 1,000 Monte Carlo replications ($k = 6$ is the DGP).

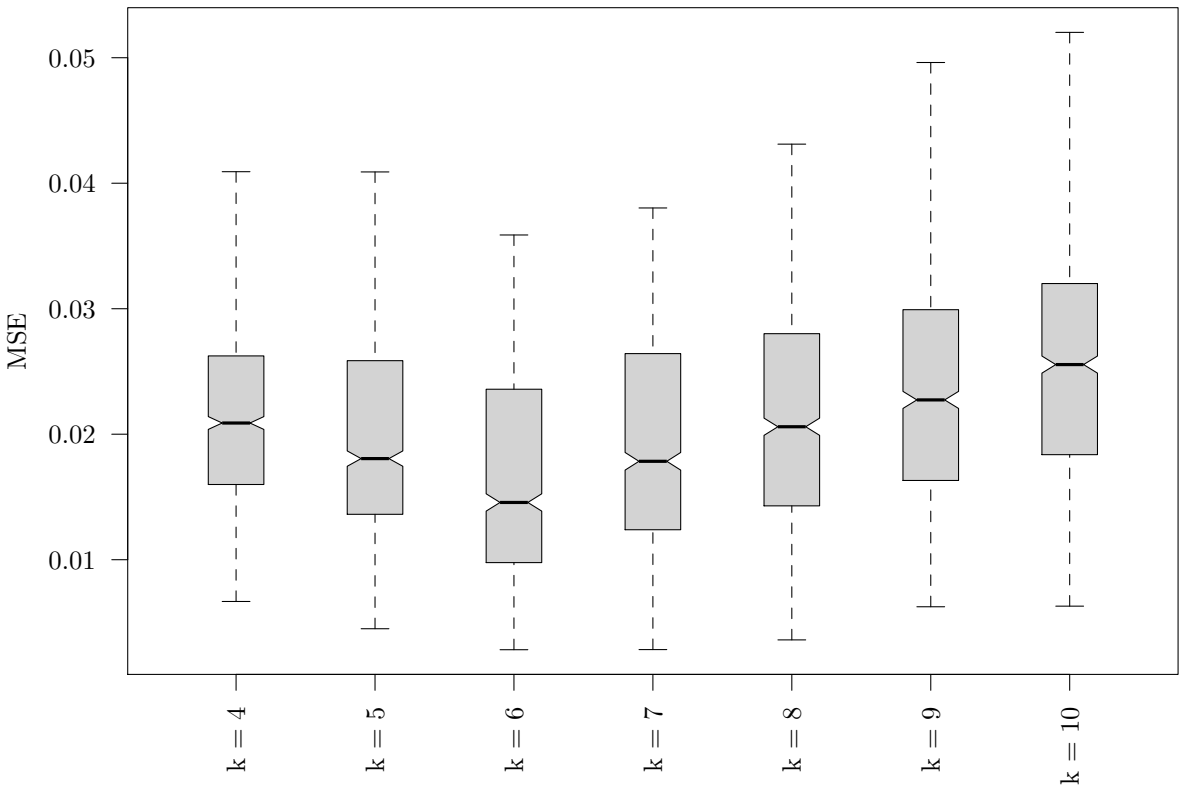

Figure 1.4: Box-and-whisker plot of MSE of the models that perform relatively well over 1,000 Monte Carlo replications ($k = 6$ is the DGP).

If models $k = 1, ..., 10$ are deterministic across 1,000 Monte Carlo replications, then model averaging, model selection, bagging, boosting, and the post-lasso can be thought of as stochastic methods, as they select a different model from replication to replication. Figure 1.5 shows the MSE for each of these methods. Model averaging using Mallows' Model Average (MMA) criterion performs the best in terms of MSE among the stochastic methods, whereas bagging performs the worst in terms of MSE. As expected, boosting outperforms bagging, because both bagging and boosting decrease bias, but boosting reduces variance by more, leading to a lower MSE. Figure 1.6 gives a closer look at the relative performance of model averaging, model selection, boosting, and the post-lasso over 1,000 Monte Carlo replications. Model selection using Mallows' C_p performs relatively well in terms of MSE, as do boosting and the post-lasso. None of these stochastic methods, however, perform better than model averaging.

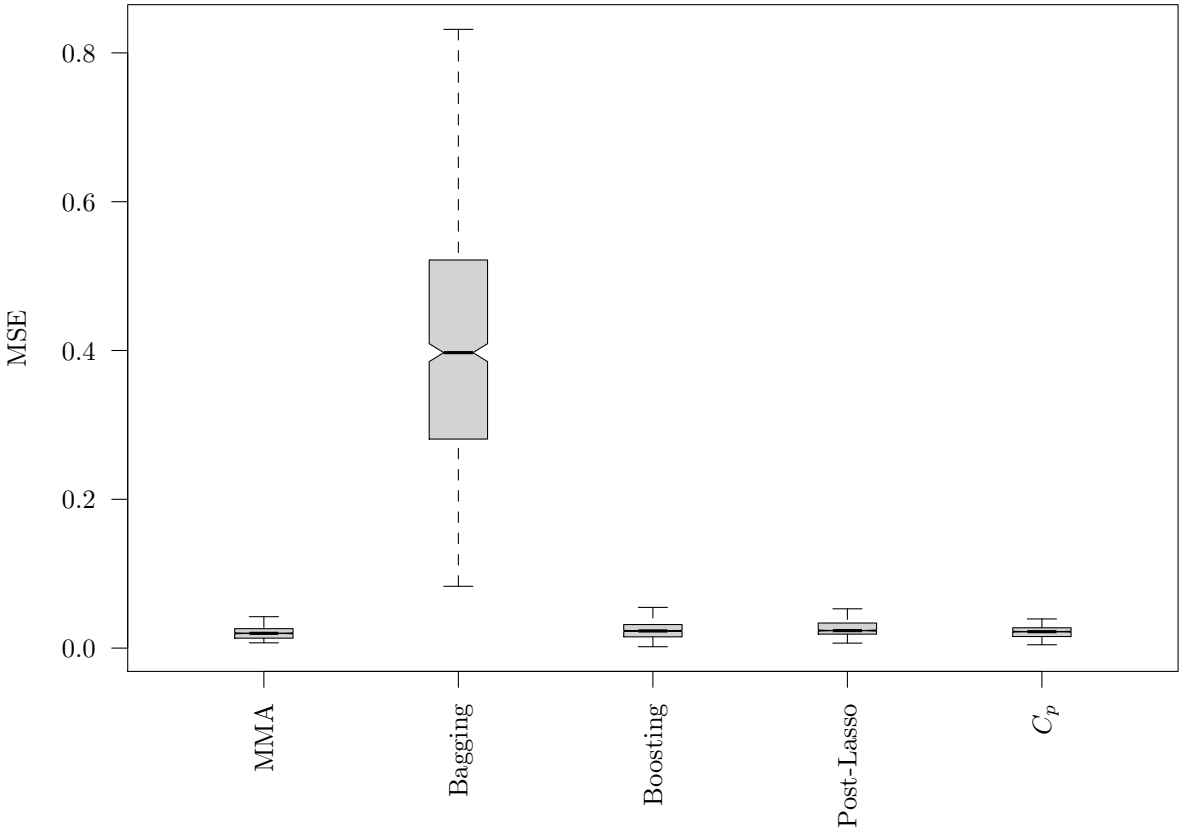

Figure 1.5: Box-and-whisker plot of MSE of each method over 1,000 Monte Carlo replications.

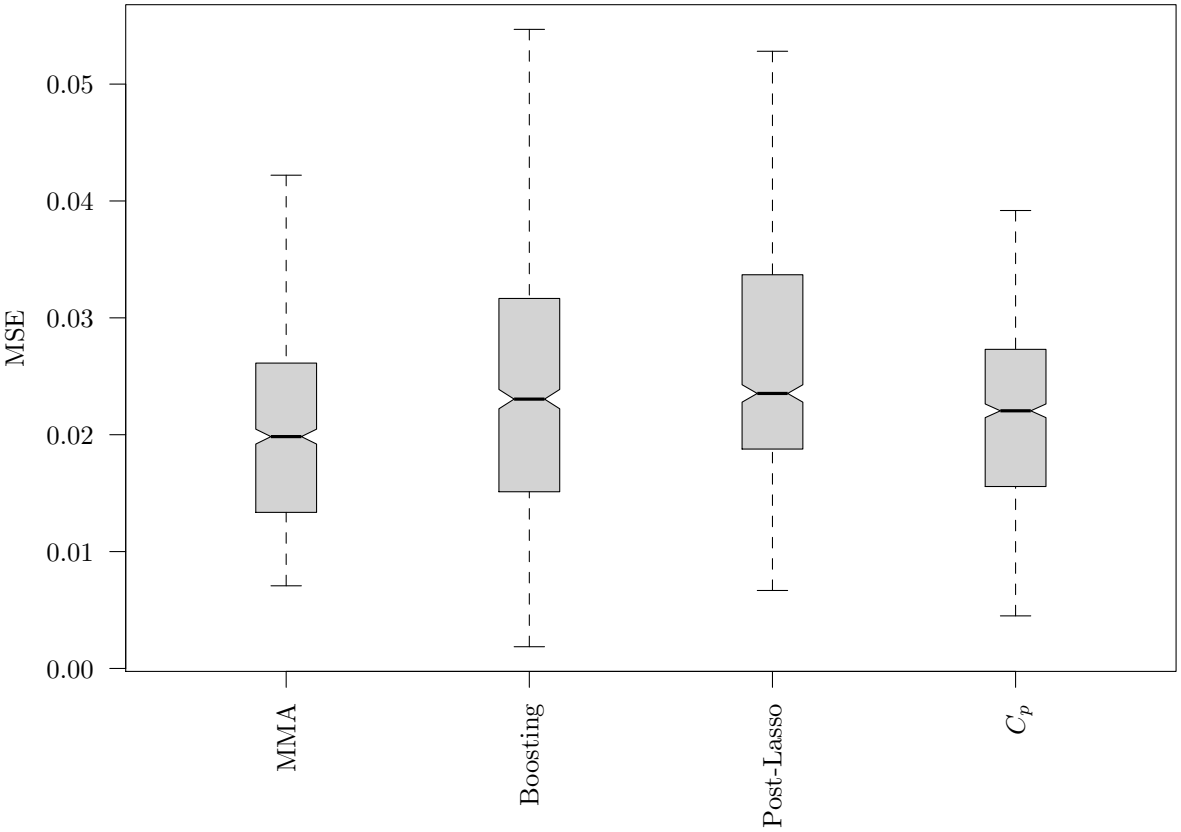

Figure 1.6: Box-and-whisker plot of MSE of the methods that perform relatively well over 1,000 Monte Carlo replications.

Table 1.1 shows the mean MSE over 1,000 replications for each method, each candidate model, and the true model. It ranks each in terms of MSE performance overall and among its competitors (i.e. among stochastic methods and among deterministic methods). As one would expect, the oracle model – a polynomial of order $k = 6$ – performs the best overall in terms of mean MSE. While we might hasten to use this as evidence for asserting a particular model, it must be stressed that this is a simulation where the functional form of the DGP is known. In reality, it is virtually impossible to know the functional form of the DGP. Thus, model assertion is strongly discouraged, as asserting the wrong model may perform much worse than model averaging. For example, the linear specification ($k = 1$) does extremely poorly in this simulation compared to the oracle model (as well as compared to every other method and model). Again, the good overall performance of the true (unknown) model as well as the polynomial of order $k = 7$ in this simulation should not be taken as evidence in

support of model assertion because, due to model uncertainty and the density of the model space, it is impossible to know with one hundred percent certainty the true functional form of the DGP when working in applied settings.

Model averaging using the MMA criterion performs third best overall (tied with the polynomial of order $k = 5$) and performs the best among stochastic methods, suggesting that one may be better off averaging over the set of candidate models rather than performing model selection-based exercises. Bagging, boosting and the post-lasso are effectively model selection methods, and consequently suffer from the same limitations as model selection using Mallows' C_p, namely that the DGP is not guaranteed to lie among the set of candidate models. Thus, these methods simply select the best approximation to the DGP among the set of candidate models. Model selection using Mallows' C_p ranks second among the stochastic methods. As expected, boosting performs better than bagging in terms of MSE, as boosting reduces variance more than bagging does, resulting in smaller MSE (which is the sum of the bias squared and the variance). Results from changes in the SNR, different sample sizes, and with an alternate DGP lead to the same conclusions. See Appendix Section 1.8.1 for details. Table 1.2 displays the mean MMA model weights over 1,000 Monte Carlo replications. As expected, given that the true model is omitted from the set of candidate models, no one model is assigned a weight of 1 while all other models are assigned weights of 0, demonstrating the existence of model uncertainty, thus demonstrating that there is a probability greater than zero that any of the models in the set of candidate models are incorrectly specified. Table 1.3 displays how often Mallows' C_p selected each of the candidate models. Models that were never selected are not included in the table. Mallows' C_p selected the polynomial of order $k = 4$ more than any other model over 1,000 Monte Carlo replications. The polynomial of order $k = 4$ was selected by Mallows' C_p 44.2 percent of the time. However, this means that for a non-zero fraction of replications, the model selection criterion selected a different model. Thus, there is uncertainty regarding which model in the set of candidate models is the best approximation

Table 1.1: Mean MSE and ranking of MSE performance ($k = 6$ is the DGP).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0205	1.00	4	1	NA
Bagging	0.4025	19.66	14	5	NA
Boosting	0.0251	1.22	9	3	NA
Post-Lasso	0.0268	1.31	11	4	NA
C_p	0.0220	1.07	6	2	NA
k = 1	0.9247	45.18	15	NA	10
k = 2	0.3192	15.59	13	NA	9
k = 3	0.2213	10.81	12	NA	8
k = 4	0.0220	1.07	7	NA	5
k = 5	0.0205	1.00	3	NA	3
k = 6	0.0170	0.83	1	NA	1
k = 7	0.0192	0.94	2	NA	2
k = 8	0.0215	1.05	5	NA	4
k = 9	0.0240	1.17	8	NA	6
k = 10	0.0263	1.29	10	NA	7

Table 1.2: Mean MMA model weights.

	$\hat{\omega}_i$
k = 1	0.0084
k = 2	0.0253
k = 3	0.0073
k = 4	0.7338
k = 5	0.1180
k = 7	0.0818
k = 8	0.0232
k = 9	0.0006
k = 10	0.0016

of the true model.

1.6 Model Averaging using Wage Data: An Applied Illustration

Section 1.5 demonstrated the good performance of model averaging relative to model assertion, model selection, and three machine learning algorithms in the presence of model

Table 1.3: Mallows' C_p model selection proportion among the candidate models (k = 1, k = 2, and k = 3 were never selected).

Model	Selection Proportion
k = 4	0.442
k = 5	0.211
k = 7	0.222
k = 8	0.105
k = 9	0.010
k = 10	0.010

uncertainty. This exercise was concerned with prediction. In what follows, I demonstrate the broad applicability of model averaging by using it in an empirical setting.

The standard approach to estimating the returns to education is to estimate a set of parametric models and either include the results from every model or simply report the results from one preferred model specification. However, this approach ignores the possibility that each model may be misspecified, thus invalidating inference. An approach to estimating returns to education in the presence of model uncertainty is to employ model selection techniques (such as model selection using Mallows' C_p) and then report the estimates of the selected model. However, there is no guarantee that the DGP lies within the set of candidate models considered by the researcher.

Given uncertainty in model specification, Tobias & Li (2004) use BMA to estimate returns to education. In labour economics, the model specification used for estimating returns to education has been observed to vary widely across researchers. Tobias and Li reviewed 38 articles published in the years 1970-2015 that focused on returns to education and were published in general-interest economics journals, and then recorded the model specifications used in each paper. They selected variables with the highest probability of being included in the log-wage equation to construct a set of control variables. The models included the variable of interest (education), as well as experience, experience^2, and regional and urban indicators, and all possible combinations of 4 optional control variables: cognitive ability, an ability-education interaction term, and indicator variables for greater than 12

years of education and greater than 16 years of education. These variables are potentially correlated with education; consequently, whether or not variables are included could have a significant impact on the estimated returns to education. This resulted in $2^4 = 16$ model specifications. Tobias and Li found that the estimated returns to education depend crucially on the model specification. Using data from the National Longitudinal Survey of Youth (NLSY), the estimated returns to a college degree relative to a high school diploma varied in magnitude from 30 to 50 percent, depending on cognitive ability.[5] Additionally, the data do not favor any particular model, as none of the 16 candidate models receive a weight close to 1. The authors conclude that they "are not certain what is the 'correct' specification of the log wage equation [...] thus it is important to account for model uncertainty when estimating the returns to education" (Tobias & Li, 2004, p. 173).

The following empirical illustration is inspired by Tobias and Li's work. I use FMA instead of BMA to illustrate the advantages of FMA in an empirical setting. FMA is appealing to empirical economists, as they typically work within a frequentist framework. The data are from the Young Men's Cohort of the National Longitudinal Survey (NLS) for the year 1980.[6] There are 935 observations and 17 variables, including the following variables that will be used in subsequent analysis: the natural log of monthly earnings (1980 USD), education (in years), a regional indicator, an urban indicator, and cognitive ability as measured by an IQ score[7]). From the data set, I construct the following variables: experience squared, an ability-education interaction term, a indicator variable indicating that the individual has completed at least 12 years of education, and an indicator variable indicating that the individual has completed at least 16 years of education. In order to evaluate each method out-of-sample – as in-sample performance may be overstated (Mullianathan & Spiess, 2017) – this exercise first shuffles the observations to randomize their

[5]It is not stated whether this difference is statistically significant or not, but certainly the difference would be economically meaningful.

[6]This dataset is available as `wage2` through the R package `wooldridge` (sourced from Blackburn & Neumark (1992)) and was chosen for ease of replicability by other researchers.

[7]The IQ scores were collected as part of a survey administered by the respondents' schools in 1968.

order and then makes 1,000 splits of the data, separating the observations into a training set (95 percent of the data; used for estimation) and a testing set (5 percent of the data; used to evaluate out-of-sample predictive performance). I use the predicted squared error (PSE) as a measure of out-of-sample predictive performance. These results can be found in Appendix Section 1.8.2.

Following Tobias & Li (2004), I consider the following log-wage equation:

$$\ln w = \mathbf{Z}\beta^{\text{fixed}} + \mathbf{X}\beta^{\text{opt}} + \epsilon, \epsilon \sim N(0, \sigma^2\mathbf{I}_n), \tag{1.13}$$

where w represents earnings, $\mathbf{Z}$ is a $n \times p$ matrix of p "fixed" variables that will appear in every regression (education, experience, experience^2, a regional indicator, and an urban indicator), and $\mathbf{X}$ is a $n \times q$ matrix of $q = 4$ optional variables (ability, ability$\times$educ, I(educ ≥ 12), and I(educ ≥ 16)).[8] The inclusion or exclusion of the elements of $\mathbf{X}$ will define the set of $M = 2^q = 2^4 = 16$ candidate models (see Table 1.4 for reference). $\beta = [\beta^{\text{fixed}}, \beta^{\text{opt}}]$ is a $(p+q)$-dimensional vector of regression coefficients. Again, this empirical example follows current practice for constructing the set of candidate models. With this approach, model averaging cannot perform any worse than any one model. Recall that model uncertainty occurs when the probability that a model is correctly specified is less than one. When using model averaging, if no one model receives a weight of 1 (and all other models receive weights of 0), model uncertainty is expected to exist. Table 1.5 shows that model weights are non-zero for more than one model, indicating the presence of model uncertainty. Models 1, 2, 3, 5, and 10 were assigned non-zero weights, and since ev-

[8]The inclusion of indicator variables for high school completion (I(educ ≥ 12)) and college completion (I(educ ≥ 16)) changes the interpretation of β_{educ}. Excluding these indicator variables, β_{educ} represents the returns to education from an additional year of education, regardless of the level of schooling that the individual has already achieved. It implies that there is a linear relationship between (log of) wage and education. Including these indicator variables allows for the possibility that there are non-linearities in the response of (log of) wage to education. There is evidence to support the idea of sheepskin effects, where achieving a degree or diploma results in a wage premium greater than the gain from previous years of schooling (see Hungerford & Solon, 1987; Belman & Heywood, 1991; Heywood, 1994; and Jaeger & Page, 1996). The inclusion of these indicator variables allows for jumps upon degree or diploma completion and allows the returns to education to vary across the educational support. Thus, while β_{educ} represents the returns to education, it does not include potential non-linearities in education captured by the indicator variables.

Table 1.4: List of models.

	Explanatory variables
Model 1	educ, exper, exper^2, south, urban
Model 2	educ, exper, exper^2, south, urban, ability
Model 3	educ, exper, exper^2, south, urban, ability $\times$ educ
Model 4	educ, exper, exper^2, south, urban, I(educ ≥ 12)
Model 5	educ, exper, exper^2, south, urban, I(educ ≥ 16)
Model 6	educ, exper, exper^2, south, urban, ability, ability $\times$ educ
Model 7	educ, exper, exper^2, south, urban, ability, I(educ ≥ 12)
Model 8	educ, exper, exper^2, south, urban, ability, I(educ ≥ 16)
Model 9	educ, exper, exper^2, south, urban, ability $\times$ educ, I(educ ≥ 12)
Model 10	educ, exper, exper^2, south, urban, ability $\times$ educ, I(educ ≥ 16)
Model 11	educ, exper, exper^2, south, urban, I(educ ≥ 12), I(educ ≥ 16)
Model 12	educ, exper, exper^2, south, urban, ability, ability $\times$ educ, I(educ ≥ 12)
Model 13	educ, exper, exper^2, south, urban, ability, ability $\times$ educ, I(educ ≥ 16)
Model 14	educ, exper, exper^2, south, urban, ability, I(educ ≥ 12), I(educ ≥ 16)
Model 15	educ, exper, exper^2, south, urban, ability $\times$ educ, I(educ ≥ 12), I(educ ≥ 16)
Model 16	educ, exper, exper^2, south, urban, ability, ability $\times$ educ, I(educ ≥ 12), I(educ ≥ 16)

ery model specification is supported by labour economic theory, it is not clear which model is "best", that is, closest to the true unknown DGP. In the presence of model uncertainty, it may be better to explicitly acknowledge model uncertainty by adopting an approach like model averaging, rather than to make ad hoc decisions regarding the model specification (Xie & Lehrer, 2017). Examining results from the model selection criterion, Mallows' C_p

Table 1.5: Average of MMA model weights over 1,000 data splits.

	$\hat{\omega}_i$
Model 1	0.0329
Model 2	0.6922
Model 3	0.2688
Model 5	0.0009
Model 10	0.0052

selects only 3 models over 1,000 splits of the data (models 2, 3 and 10; Table 1.6). These models were also assigned non-zero weights by the MMA criterion. Thus, if one were to use model selection, there may be ambiguity regarding the model specification preferred by the data. Table 1.7 displays the estimated marginal effect of an additional year of edu-

Table 1.6: Mallows' Cp model selection proportion among the candidate models.

Model	Selection Proportion
Model 2	0.826
Model 3	0.172
Model 10	0.002

cation from model averaging and each model in the set of candidate models. The returns to education are calculated for an additional year of education for an individual with 11 years of education and the mode value for discrete variables and the median value for continuous variables. The returns to education vary widely across models, ranging from 0.0585

Table 1.7: Estimated marginal effect of 1 additional year of education at 11 years of eduation.

	Marginal Effect
MMA	0.0597
Model 1	0.0763
Model 2	0.0593
Model 3	0.0585
Model 4	0.1120
Model 5	0.0868
Model 6	0.0597
Model 7	0.0699
Model 8	0.0674
Model 9	0.0874
Model 10	0.0693
Model 11	0.1041
Model 12	0.0666
Model 13	0.0674
Model 14	0.0622
Model 15	0.0784
Model 16	0.0613

to 0.112. Labour economists argue over much smaller differences in magnitudes. Interestingly, the estimates for the models that were given the highest model weights (models 2 and 3; these models were also selected by Mallows' C_p) are similar, yet these estimates

differ in magnitude substantially from the estimates from other models. This wage data example highlights how marginal effect estimates depend greatly on the model specification chosen by the researcher. Thus, in the face of model uncertainty, it is sub-optimal to assert a particular model specification. Model averaging is strongly recommended for dealing with model uncertainty in empirical problems in economics, such as estimating returns to education.

1.7 Conclusion

This chapter demonstrates the advantages of model averaging over conventional econometric methods (model assertion and model selection) and machine learning algorithms in the presence of model uncertainty. Selecting a parametric model in an ad hoc manner and proceeding with estimation and inference asserting that this model could plausibly have generated the data can be misleading if the model does not accurately represent the DGP. Given that the model space is dense, the probability of selecting a model that represents the true, unknown DGP may be low. Model selection is an improvement over model assertion. However, selecting the least misspecified model among a finite set of candidate models does not guarantee that the DGP lies within the set of candidate models. Model averaging recognizes model uncertainty and reduces estimation variance while controlling for misspecification bias by constructing a simple weighted average over a set of candidate models. While model selection and model assertion rely on having a good approximation to the DGP, model averaging does not require a good approximation to the DGP to be included in the set of candidate models. Additionally, model averaging can be easily adopted by empirical economists, as demonstrated in this chapter. The results of Monte Carlo experiments show that model averaging performs as well as, and often better, in terms of MSE than model selection and machine learning methods (boosting, bagging and the post-lasso) in the presence of model uncertainty. While model assertion may, in some cases,

do better than model averaging when the DGP is known (as in Monte Carlo simulations), selecting a model in an ad hoc manner may do much worse than model averaging. Thus, model averaging is a feasible approach with good MSE performance for cases where model uncertainty is a concern.

I address one of the possible limitations to the adoption of model averaging methods, which is the lack of a universal standard for constructing the set of candidate models, a key step in implementing model averaging. The first approach follows the current practice for model averaging and consists of simply writing down a set of candidate models that have a common parameter (such as the coefficient on the execution rate in estimating the plausible deterrent effect of capital punishment). The second approach involves creating a set of candidate models that span a rich set of bases in order to capture a wide range of DGPs. The basis function type can be selected using a model selection criterion.

Finally, I apply model averaging to an empirical problem in labour economics and demonstrate another application of model averaging (causal estimation). Using the National Longitudinal Survey, I demonstrate how easily model averaging can be used to estimate returns to education. The non-zero weights assigned to more than one model among the 16 candidate models demonstrates the presence of model uncertainty. Additionally, the estimated returns to education vary widely across models, ranging from 0.0585 to 0.112, which is an economically meaningful range. This highlights how marginal effect estimates depend greatly on the model specification. Given the presence of model uncertainty and the variation in marginal effect estimates across models, model averaging is recommended for handling model uncertainty in empirical economic problems.

This chapter has given an introduction to model averaging methods and applications in an empirical economic environment. Further research must be done to set up best practices for post-model-average inference. Additionally, it would be compelling to reassess the performance of model averaging relative to boosting, bagging and the post-lasso in a prediction problem, as machine learning algorithms have been built for these types of

problems rather than for estimating causal relationships.

1.8 Appendix

1.8.1 Monte Carlo Experiments

Results from Different Signal-to-Noise Ratios

The tables below summarize the mean MSE for changes in the signal-to-noise ratio (SNR). Recall that in this Monte Carlo experiment, the true DGP is $f(x) = 1 + x^6/\sigma_{x^6}$ and $y = f(x) + \epsilon$, where $x \sim N(0,1)$ and $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{x^6})$. The constant c, where $c \in \{0.25, 0.50, 1.0, 2.0\}$, determines the SNR ($c = 0.25$ being the highest SNR and $c = 2.0$ being the lowest SNR). See Section 1.5 for the results for $c = 0.50$.

The results from Table 1.8 ($c = 0.25$) show that model selection using Mallows' C_p performs slightly better than model averaging using Mallows' Model Averaging criterion, but model averaging performs better than the machine learning algorithms. Tables 1.9 and 1.10 ($c = 1.0$ and $c = 2.0$ respectively) show that model averaging performs better than model selection and the machine learning algorithms across changes in the SNR. Some other models in the set of candidate models may do better than model averaging. However, this is a simulation where the DGP is known. In reality, it may be unwise to assume the functional form of the DGP in an ad hoc manner, as it is impossible to know whether the model is correctly specified and the consequences of a misspecified model can be severe. Based on these results, model averaging may be the preferred method.

Results from Different Sample Sizes

In order to assess the performance of each method under a variety of sample sizes, I run a Monte Carlo experiment (identical to that in Section 1.5). In the following simulation, the DGP is $f(x) = 1+x^6/\sigma_{x^6}$ and $y = f(x)+\epsilon$, where $x \sim N(0,1)$, $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{x^6})$, and

Table 1.8: Mean MSE and ranking of MSE performance ($k = 6$ is the oracle model; $c = 0.25$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0062	1.00	6	2	NA
Bagging	0.3869	62.61	14	5	NA
Boosting	0.0080	1.29	8	3	NA
Post-Lasso	0.0091	1.47	9	4	NA
C_p	0.0058	0.93	4	1	NA
k = 1	0.9212	149.07	15	NA	10
k = 2	0.3141	50.82	13	NA	9
k = 3	0.2136	34.56	12	NA	8
k = 4	0.0129	2.09	11	NA	7
k = 5	0.0095	1.54	10	NA	6
k = 6	0.0043	0.69	1	NA	1
k = 7	0.0049	0.79	2	NA	2
k = 8	0.0054	0.88	3	NA	3
k = 9	0.0061	0.98	5	NA	4
k = 10	0.0067	1.08	7	NA	5

Table 1.9: Mean MSE and ranking of MSE performance ($k = 6$ is the oracle model; $c = 1.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0707	1.00	4	1	NA
Bagging	0.5211	7.37	14	5	NA
Boosting	0.0857	1.21	7	3	NA
Post-Lasso	0.1915	2.71	11	4	NA
C_p	0.0712	1.01	5	2	NA
k = 1	0.9375	13.26	15	NA	10
k = 2	0.3396	4.80	13	NA	9
k = 3	0.2502	3.54	12	NA	8
k = 4	0.0589	0.83	1	NA	1
k = 5	0.0648	0.92	2	NA	2
k = 6	0.0682	0.97	3	NA	3
k = 7	0.0772	1.09	6	NA	4
k = 8	0.0865	1.22	8	NA	5
k = 9	0.0962	1.36	9	NA	6
k = 10	0.1056	1.49	10	NA	7

$c \in \{0.25, 0.50, 1.0, 2.0\}$, determines the SNR. I take $n = 500$ random draws of x from the normal distribution and conduct 1,000 Monte Carlo replications. There are $M = 9$ candidate models, which are orthogonal polynomials of order 1 through 5 and 7 through

Table 1.10: Mean MSE and ranking of MSE performance ($k = 6$ is the oracle model; $c = 2.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.2833	1.00	4	1	NA
Bagging	0.9621	3.40	14	5	NA
Boosting	0.3356	1.18	7	3	NA
Post-Lasso	0.6646	2.35	13	4	NA
C_p	0.2836	1.00	5	2	NA
k = 1	0.9906	3.50	15	NA	10
k = 2	0.4217	1.49	11	NA	8
k = 3	0.3691	1.30	9	NA	6
k = 4	0.2066	0.73	1	NA	1
k = 5	0.2419	0.85	2	NA	2
k = 6	0.2733	0.96	3	NA	3
k = 7	0.3089	1.09	6	NA	4
k = 8	0.3452	1.22	8	NA	5
k = 9	0.3848	1.36	10	NA	7
k = 10	0.4229	1.49	12	NA	9

10 (omitting the true DGP, a polynomial of order 6). Tables 1.11 to 1.14 display the mean MSE over 1,000 Monte Carlo replications. In Table 1.11, model averaging using Mallows' Model Averaging criterion ranks third after the oracle model ($k = 6$) and another model in the set of candidate models. In reality, it is almost impossible to know for certain what is the true functional form of the DGP. In the presence of model uncertainty, one may do worse by asserting the functional form than model averaging (for example, specifying a linear model in this simulation, $k = 1$, performs the worst in terms of MSE). Thus, it may be prudent to use model averaging. Model averaging ranks fourth in Table 1.12 after some models in the set of candidate models. In Table 1.13, model averaging ranks fifth best after the oracle model, some candidate models, and boosting. While boosting performs relatively well in terms of MSE in this case, it is not consistently in the top five, unlike model averaging. Finally, model averaging using the MMA criterion ranks second overall (after the oracle model) in Table 1.14. In the next simulation, the DGP is $f(x) = 1 + x^6/\sigma_{x^6}$ and $y = f(x) + \epsilon$, where $x \sim N(0, 1)$, $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{x^6})$, and $c \in \{0.25, 0.50, 1.0, 2.0\}$ determines the SNR. This time, I take $n = 1,000$ random draws of x from the normal

Table 1.11: Mean MSE and ranking of MSE performance ($k = 6$ is the oracle model; $n = 500$; $c = 0.25$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0010	1.00	3	1	NA
Bagging	0.2444	242.01	12	5	NA
Boosting	0.0020	2.03	9	4	NA
Post-Lasso	0.0011	1.06	4	2	NA
C_p	0.0011	1.11	6	3	NA
k = 1	0.9816	972.15	15	NA	10
k = 2	0.4463	442.02	14	NA	9
k = 3	0.3746	371.03	13	NA	8
k = 4	0.0220	21.79	11	NA	7
k = 5	0.0163	16.12	10	NA	6
k = 6	0.0008	0.82	1	NA	1
k = 7	0.0010	0.95	2	NA	2
k = 8	0.0011	1.07	5	NA	3
k = 9	0.0012	1.19	7	NA	4
k = 10	0.0013	1.32	8	NA	5

Table 1.12: Mean MSE and ranking of MSE performance ($n = 500$; $c = 0.50$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0042	1.00	4	1	NA
Bagging	0.2498	59.01	12	5	NA
Boosting	0.0049	1.15	7	3	NA
Post-Lasso	0.0056	1.33	9	4	NA
C_p	0.0043	1.03	5	2	NA
k = 1	0.9827	232.13	15	NA	10
k = 2	0.4479	105.79	14	NA	9
k = 3	0.3762	88.85	13	NA	8
k = 4	0.0237	5.61	11	NA	7
k = 5	0.0184	4.34	10	NA	6
k = 6	0.0032	0.76	1	NA	1
k = 7	0.0038	0.89	2	NA	2
k = 8	0.0042	0.99	3	NA	3
k = 9	0.0047	1.11	6	NA	4
k = 10	0.0052	1.23	8	NA	5

distribution and conduct 1,000 Monte Carlo replications. There are $M = 9$ candidate models and the true DGP is omitted. Tables 1.15 to 1.18 display the mean MSE over 1,000 Monte Carlo replications. Model averaging ranks third after the oracle model ($k = 6$)

Table 1.13: Mean MSE and ranking of MSE performance ($n = 500$; $c = 1.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0182	1.00	5	2	NA
Bagging	0.2792	15.31	12	5	NA
Boosting	0.0173	0.95	4	1	NA
Post-Lasso	0.0330	1.81	11	4	NA
C_p	0.0189	1.04	7	3	NA
k = 1	0.9857	54.04	15	NA	10
k = 2	0.4510	24.73	14	NA	9
k = 3	0.3817	20.93	13	NA	8
k = 4	0.0306	1.68	10	NA	7
k = 5	0.0268	1.47	9	NA	6
k = 6	0.0128	0.70	1	NA	1
k = 7	0.0150	0.82	2	NA	2
k = 8	0.0167	0.92	3	NA	3
k = 9	0.0187	1.03	6	NA	4
k = 10	0.0208	1.14	8	NA	5

Table 1.14: Mean MSE and ranking of MSE performance ($n = 500$; $c = 2.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0577	1.00	2	1	NA
Bagging	0.5107	8.85	14	5	NA
Boosting	0.0706	1.22	8	3	NA
Post-Lasso	0.1224	2.12	11	4	NA
C_p	0.0688	1.19	7	2	NA
k = 1	0.9977	17.30	15	NA	10
k = 2	0.4684	8.12	13	NA	9
k = 3	0.4040	7.00	12	NA	8
k = 4	0.0582	1.01	3	NA	2
k = 5	0.0604	1.05	5	NA	4
k = 6	0.0512	0.89	1	NA	1
k = 7	0.0602	1.04	4	NA	3
k = 8	0.0670	1.16	6	NA	5
k = 9	0.0749	1.30	9	NA	6
k = 10	0.0832	1.44	10	NA	7

and another candidate model in Tables 1.15 and 1.16. Model averaging using the MMA criterion performs fourth best in terms of MSE after the oracle model and some models in the set of candidate models in Tables 1.17 and 1.18. Some candidate models perform

well in these simulations. However, the functional form of the DGP is unknown when working with real data, and asserting a particular model specification may be worse than model averaging. For example, specifying a linear or quadratic functional form with the data from this simulation would be much worse in terms of MSE than model averaging. In

Table 1.15: Mean MSE and ranking of MSE performance ($k = 6$ is the oracle model; $n = 1,000$; $c = 0.25$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0005	1.00	3	1	NA
Bagging	0.2086	394.81	12	5	NA
Boosting	0.0017	3.18	9	4	NA
Post-Lasso	0.0005	1.04	4	2	NA
C_p	0.0006	1.12	6	3	NA
k = 1	0.9896	1873.00	15	NA	10
k = 2	0.4919	931.05	14	NA	9
k = 3	0.4311	815.95	13	NA	8
k = 4	0.0290	54.90	11	NA	7
k = 5	0.0215	40.73	10	NA	6
k = 6	0.0004	0.84	1	NA	1
k = 7	0.0005	0.96	2	NA	2
k = 8	0.0006	1.08	5	NA	3
k = 9	0.0006	1.20	7	NA	4
k = 10	0.0007	1.32	8	NA	5

general, as the sample size (n) increases, MSE decreases for all methods. However, model averaging using the MMA criterion consistently ranks among the top methods (except in one case where boosting outperforms model averaging; see Table 1.13), while there is some variability in the performance of the other methods. Thus, it is advisable to use model averaging in the presence of model uncertainty.

Results from a Different Data Generating Process

In order to assess the performance of each method under a variety of DGPs, a different DGP is selected. In the following Monte Carlo experiment, the DGP is set to $f(x) = e^x$ and $y = f(x) + \epsilon$ where $x \sim N(0, 1)$ and $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{f(x)})$. As before, the constant c, where $c \in \{0.25, 0.50, 1.0, 2.0\}$, determines the SNR. I take $n = 100$ random draws

Table 1.16: Mean MSE and ranking of MSE performance ($n = 1,000$; $c = 0.50$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0021	1.00	3	1	NA
Bagging	0.2116	98.54	12	5	NA
Boosting	0.0033	1.52	9	4	NA
Post-Lasso	0.0025	1.16	6	3	NA
C_p	0.0024	1.11	5	2	NA
k = 1	0.9892	460.76	15	NA	10
k = 2	0.4929	229.58	14	NA	9
k = 3	0.4293	199.95	13	NA	8
k = 4	0.0298	13.88	11	NA	7
k = 5	0.0221	10.30	10	NA	6
k = 6	0.0017	0.79	1	NA	1
k = 7	0.0020	0.92	2	NA	2
k = 8	0.0023	1.05	4	NA	3
k = 9	0.0025	1.18	7	NA	4
k = 10	0.0028	1.30	8	NA	5

Table 1.17: Mean MSE and ranking of MSE performance ($n = 1,000$; $c = 1.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0094	1.00	4	1	NA
Bagging	0.2213	23.54	12	5	NA
Boosting	0.0098	1.05	6	3	NA
Post-Lasso	0.0155	1.65	9	4	NA
C_p	0.0097	1.03	5	2	NA
k = 1	0.9907	105.38	15	NA	10
k = 2	0.4953	52.69	14	NA	9
k = 3	0.4322	45.97	13	NA	8
k = 4	0.0334	3.56	11	NA	7
k = 5	0.0264	2.81	10	NA	6
k = 6	0.0068	0.72	1	NA	1
k = 7	0.0079	0.84	2	NA	2
k = 8	0.0090	0.96	3	NA	3
k = 9	0.0101	1.07	7	NA	4
k = 10	0.0111	1.18	8	NA	5

of x from the normal distribution and conduct 1,000 Monte Carlo replications. There are $M = 10$ candidate models, which are orthogonal polynomials of order 1 through 10. The set of candidate models follows the current practice of selecting a set of candidate models

Table 1.18: Mean MSE and ranking of MSE performance ($n = 1,000$; $c = 2.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0362	1.00	4	1	NA
Bagging	0.2693	7.44	12	5	NA
Boosting	0.0390	1.08	5	2	NA
Post-Lasso	0.0586	1.62	11	4	NA
C_p	0.0403	1.11	6	3	NA
k = 1	0.9970	27.54	15	NA	10
k = 2	0.5042	13.93	14	NA	9
k = 3	0.4442	12.27	13	NA	8
k = 4	0.0479	1.32	10	NA	7
k = 5	0.0438	1.21	8	NA	5
k = 6	0.0271	0.75	1	NA	1
k = 7	0.0317	0.88	2	NA	2
k = 8	0.0361	1.00	3	NA	3
k = 9	0.0404	1.12	7	NA	4
k = 10	0.0445	1.23	9	NA	6

in an ad hoc manner, rather than using the approach outlined in the latter half of Section 1.3, which suggests a nonparametric approach for selecting the set of candidate models. The DGP (in this case, $f(x) = e^x$) is once again omitted from the set of candidate models.

Tables 1.19 to 1.22 show the mean MSE over 1,000 replications for model averaging using Mallows' Model Average (MMA), model selection using Mallows' C_p, bagging, boosting, the post-lasso, and each candidate model, and ranks each in terms of MSE performance. Across the different SNR, model averaging consistently ranks third after some candidate models. Results are similar to those from Section 1.5. Although some candidate models may do better than model averaging in this Monte Carlo experiment, many do far worse. Consequently, in empirical settings where the true DGP is unknown, model averaging may be a better approach. Additionally, once again, model averaging outperforms model selection and three machine learning algorithms. This suggests that one may prefer averaging over the set of candidate models rather than using model selection-based exercises. Even when the set of candidate models are chosen in an ad hoc manner and have no relation to the DGP, model averaging performs better than the model selection-based

methods in terms of MSE.

Table 1.19: Mean MSE and ranking of MSE performance (e^x is the DGP; $c = 0.25$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0184	1.00	3	1	NA
Bagging	0.8665	46.97	14	5	NA
Boosting	0.0237	1.29	8	3	NA
Post-Lasso	0.0239	1.29	9	4	NA
C_p	0.0194	1.05	5	2	NA
k = 1	1.5223	82.52	15	NA	10
k = 2	0.3131	16.97	13	NA	9
k = 3	0.0427	2.31	12	NA	8
k = 4	0.0158	0.85	1	NA	1
k = 5	0.0163	0.88	2	NA	2
k = 6	0.0185	1.00	4	NA	3
k = 7	0.0208	1.13	6	NA	4
k = 8	0.0233	1.26	7	NA	5
k = 9	0.0259	1.40	10	NA	6
k = 10	0.0284	1.54	11	NA	7

Table 1.20: Mean MSE and ranking of MSE performance (e^x is the DGP; $c = 0.50$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.0685	1.00	3	1	NA
Bagging	0.8829	12.89	14	5	NA
Boosting	0.0839	1.23	8	3	NA
Post-Lasso	0.0905	1.32	10	4	NA
C_p	0.0715	1.04	4	2	NA
k = 1	1.4769	21.57	15	NA	10
k = 2	0.3183	4.65	13	NA	9
k = 3	0.0738	1.08	6	NA	4
k = 4	0.0563	0.82	1	NA	1
k = 5	0.0633	0.92	2	NA	2
k = 6	0.0719	1.05	5	NA	3
k = 7	0.0801	1.17	7	NA	5
k = 8	0.0900	1.31	9	NA	6
k = 9	0.1006	1.47	11	NA	7
k = 10	0.1099	1.60	12	NA	8

Table 1.21: Mean MSE and ranking of MSE performance (e^x is the DGP; $c = 1.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.2333	1.00	3	1	NA
Bagging	1.3239	5.68	14	5	NA
Boosting	0.3258	1.40	8	3	NA
Post-Lasso	0.3630	1.56	9	4	NA
C_p	0.2627	1.13	5	2	NA
k = 1	1.5373	6.59	15	NA	10
k = 2	0.4078	1.75	12	NA	8
k = 3	0.2075	0.89	1	NA	1
k = 4	0.2211	0.95	2	NA	2
k = 5	0.2559	1.10	4	NA	3
k = 6	0.2906	1.25	6	NA	4
k = 7	0.3239	1.39	7	NA	5
k = 8	0.3641	1.56	10	NA	6
k = 9	0.4070	1.74	11	NA	7
k = 10	0.4444	1.90	13	NA	9

Table 1.22: Mean MSE and ranking of MSE performance (e^x is the DGP; $c = 2.0$).

	Mean MSE	Relative Efficiency	Overall Rank	Stochastic Rank	Det. Rank
MMA	0.7870	1.00	3	1	NA
Bagging	3.0599	3.89	15	5	NA
Boosting	1.2275	1.56	8	3	NA
Post-Lasso	2.3545	2.99	14	4	NA
C_p	0.9212	1.17	5	2	NA
k = 1	1.7578	2.23	12	NA	9
k = 2	0.7615	0.97	2	NA	2
k = 3	0.7377	0.94	1	NA	1
k = 4	0.8729	1.11	4	NA	3
k = 5	1.0158	1.29	6	NA	4
k = 6	1.1535	1.47	7	NA	5
k = 7	1.2860	1.63	9	NA	6
k = 8	1.4417	1.83	10	NA	7
k = 9	1.6159	2.05	11	NA	8
k = 10	1.7665	2.24	13	NA	10

1.8.2 Model Averaging using Wage Data: Predicted Squared Error

In Section 1.6, I illustrate the advantages of FMA in an empirical setting using data from the Young Men's Cohort of the National Longitudinal Survey (NLS) for the year 1980. The data are split into a testing set and a training set. This sample splitting procedure is described in detail in Section 1.6. Here, I compute the predicted squared error (PSE) for each method using the testing data (a 5 percent random sample of the data) to evaluate each method out-of-sample. I use the predicted squared error (PSE) as a measure of out-of-sample predictive performance. PSE is calculated as follows:

$$\text{PSE } = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2, i = 1, ...n, \tag{1.14}$$

where y_i is the observed dependent variable (in this case, the natural log of monthly earnings) and $\hat{y_i}$ are the predicted values from each respective method obtained from the testing data. PSE differs from MSE because the true DGP is unknown.

Table 1.23 displays the mean PSE over 1,000 splits of the data for each method used in the wage example, as well as for the three machine learning algorithms. Note that ranking methods by PSE is not the same as ranking by MSE (which cannot be computed in applied settings since the DGP is unknown). While model averaging does not rank highly relative to the candidate models in terms of PSE, the magnitude of its PSE is extremely close to that of the candidate models ranking above it. In principle, bootstrapping could be used to test whether the magnitude of one method's PSE is significantly different from another, but I have not done that here. Potential reasons for the relatively poor performance of model averaging could be low explanatory power of the models or a low in-sample SNR.

Table 1.23: Mean predicted squared error (PSE) and ranking of mean PSE performance.

	Mean PSE	Rank
MMA	0.142362	7
C_p	0.142413	10
Bagging	0.148186	20
Boosting	0.164819	21
Post-lasso	0.141688	1
Model 1	0.145595	16
Model 2	0.142047	2
Model 3	0.142133	3
Model 4	0.145806	17
Model 5	0.145833	18
Model 6	0.142392	9
Model 7	0.142324	5
Model 8	0.142318	4
Model 9	0.142373	8
Model 10	0.142352	6
Model 11	0.146101	19
Model 12	0.142675	14
Model 13	0.142661	13
Model 14	0.142602	11
Model 15	0.142632	12
Model 16	0.142951	15

Chapter 2

In Search of the Optimal Model Set: Methods for Generating Candidate Models for Model Averaging

2.1 Introduction

Model selection and model averaging are useful in situations where there are a number of competing models that are supported by economic theory, yet it is unclear which model is the "best" model among them. These methods use statistical approaches to deal with this inherent model uncertainty to obtain an improved estimator of the quantity of interest. Selecting the set of candidate models to be used in model selection or model averaging is an important step when dealing with model uncertainty because the resulting estimator will inherit properties from the candidate models, yet it is often overlooked. Current practice is to write down a handful of parametric models that have a common parameter of interest, estimate each model individually, and then either use a model selection criterion to select one model from the set of candidate models or obtain model average weights to compute the model average estimator, which is a weighted average of the estimates from each individual

model. A limitation of this approach is that it often relies on ad hoc decisions on the part of the researcher, for example, with regard to the functional form specification of each model and the set of regressors to include in each model.

Ideally, the set of candidate models would have these key features:

1. number of candidate models tied to the sample size and assumed smoothness class (i.e. data generating process, DGP),
2. number of parameters tied to the sample size and assumed smoothness class, and
3. a broad set of functions, or "bases", to cover a wide range of potential DGPs that are simple enough to average over.

Thus, we need a statistical procedure that balances complexity, breadth, and computational efficiency. This chapter investigates promising approaches that have the potential to produce a set of candidate models for model averaging with the key features outlined above. Additionally, I develop useful heuristics to guide practitioners in implementing the recommended methods. I consider three bodies of work from econometrics and machine learning to guide my research. First, model screening is appealing because it can shrink the set of potential candidate models prior to model averaging to improve computational efficiency. However, this approach has a number of limitations, detailed in Section 2.2. Second, computer automated algorithms such as recursive partitioning-based methods have the potential to produce an optimal set of candidate models by automatically generating a rich set of bases based on the data (see Section 2.3). Last, existing frequentist model average methods that average over nonparametric models are considered (see Section 2.4). The merits and limitations of each method are discussed in depth, especially with regard to their ability to generate a set of candidate models with the key features listed above. Heuristics are developed to guide practitioners in their research; however, a careful examination of existing methods is needed to fully understand their potential and their limitations. Section 2.5 evaluates the relative performance of the most promising approaches in a Monte Carlo experiment in order to determine whether one approach, if any, performs better than the

others in terms of mean squared error. Section 2.6 concludes.

2.2 Model Screening

Model screening encompasses model selection and variable selection methods. It can be used to shrink the set of candidate models prior to model averaging, which may be helpful in balancing complexity, breadth and computational efficiency. There exists some evidence that supports the use of model screening prior to model averaging. Zhu, Wan, Zhang, & Zou (2019) claim that "removing the poorest models before averaging can contribute to greater estimation and predictive efficiency".

In model selection, model screening can be used as a method for constructing the set of candidate models. Examples of model screening methods that result in a single model specification to be used for subsequent estimation, forecasting or inference include stepwise regression and the lasso and variations thereof.

Stepwise regression uses a sequence of t- and/or F-tests or criteria such as adjusted R^2, the Akaike information criterion (AIC; Akaike, 1970), the Bayesian information criterion (BIC; Schwarz, 1978), or Mallows' C_p (Mallows, 1973) to select a single model. AIC and BIC are defined as follows:

$$\text{AIC} = -2\ln(\hat{L}) + 2p \tag{2.1}$$

$$\text{BIC} = -2\ln(\hat{L}) + \log(n)p \tag{2.2}$$

where $\ln(\hat{L})$ is the maximum value of the log-likelihood function of a model with p regressors and n is the sample size. These information criteria balance goodness of fit (as measured by the log-likelihood) and parsimony (as measured by the penalty for the number of regressors included in the model). A low AIC or BIC value is desirable.

Mallows' C_p is defined as follows:

$$C_p = \frac{SSR_p}{\hat{\sigma}^2} + 2p - n, \tag{2.3}$$

where $SSR_p = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ is the sum of squared residuals (y_i is the response variable (the observed outcome) and $\hat{y_i}$ are the fitted values) in the model with p regressors and $\hat{\sigma}^2$ is the estimated variance from the largest dimension model in the set of candidate models. A low C_p value is desirable. Therefore the model with the lowest C_p value in the set of candidate models is selected.

The main stepwise regression approaches are forward selection and backwards elimination. Forward selection begins with the null model and iteratively adds one variable at a time (Claeskens, Croux, & VanKerckhoven, 2005). Backwards elimination begins with a full or general unrestricted model and iteratively eliminates the variable which gives the largest reduction or smallest increase to the value of the information criterion. An example of a backwards elimination procedure is automated general-to-specific (PcGets) model selection algorithm (Campos, Hendry, & Krolzig, 2003; Castle, 2006; Krolzig & Hendry, 2001, 2011). This algorithm does both model selection (through variable selection) and diagnostic tests to check the validity of the reductions. The goal of the algorithm is to select a congruent, parsimonious terminal model through a procedure that explores multiple paths, but not so many that search costs are too high, while avoiding getting stuck in a path that inadvertently deletes variables that matter. The algorithm begins with a single general unrestricted model (GUM) that undergoes an initial misspecification test. A preselection screening step eliminates highly irrelevant variables based on t- and F-tests. It then iteratively simplifies the model by eliminating insignificant variables until only relevant variables remain, using t- and/or F-tests as the simplification criteria. Encompassing tests test all distinct contending model specifications. Any model that survives these tests is retained. If multiple models survive, a new general model is formed from their union,

and the model simplification step is re-applied. This repeats until either a unique model specification emerges or the previous union is reproduced. If the previous union is reproduced (that is, multiple model specifications remain), a final selection is made using an information criterion, such as AIC or BIC. Diagnostic tests are applied at every reduction stage to test for congruency. A model is "congruent" if the model "matches evidence in all measured aspects" (Castle, 2006). Reliability scores are assigned to variables to guide the model choice of researchers. Monte Carlo experiments show that PcGets can recover the DGP from a general model, with size and power close to using the DGP itself (Castle, 2006). However, PcGets has a number of limitations. For one, an absence of an optimal sequence for simplification makes the choice of reduction path unclear. Additionally, when there is insufficient data, the algorithm may perform poorly in specifying the best approximation to the DGP. Lastly, and perhaps most importantly, the good performance of the algorithm relies crucially on the researcher specifying the GUM, consequently relying on ad hoc parametric model specification by the researcher. Thus, PcGets may not be the best strategy for selecting the set of candidate models to be used in model averaging.

The *lasso*, or *least absolute shrinkage and selection operator*, can be used for model screening (Belloni & Chernozhukov, 2013; Tibshirani, 1996). The lasso shrinks some model coefficients and sets others to zero, essentially performing selection of regressors to select a single model. The lasso estimator is defined as:

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \Big(y_i - \sum_j \beta_j x_{ij} \Big)^2 \text{ subject to } \sum_j |\beta_j| \leq t, \tag{2.4}$$

where $i = 1, ..., n$, j indexes the regressor, β represents regression coefficients in the model $y_i = x_i\beta + \epsilon_i$, and $t \geq 0$ represents the tuning (or penalty) parameter. Selection of the tuning parameter, t, is important as it controls regressor selection as well as how much shrinkage is applied to the coefficients. Cross-validation is commonly used to select the tuning parameter. The advantages of the lasso include highly interpretable models, in-

creased stability, and improved prediction accuracy with relative computational efficiency. However, the non-zero coefficients resulting from the lasso tend to be biased towards zero, due in part to shrinkage (Belloni et al., 2014).

In model averaging, model screening can be used as a preliminary step to reduce the total number of candidate models in order to improve computational efficiency and perhaps even estimation or predictive efficiency (note the two different meanings of efficiency). Some examples of model screening applied to model averaging include pre-selection and model averaging post-lasso. Pre-selection applies a backwards elimination procedure like PcGets to a large set of candidate models to reduce the size of the set prior to model averaging. Xie & Lehrer (2017) applied a version of PcGets (Campos et al., 2003) as a model screening step prior to model averaging in a forecasting exercise using film industry and social media data to predict movie success. Initially, a total of 16,777,216 potential models for open box office and 4,294,967,296 potential models for retail movie sales were screened using the estimated p-values for tests of statistical significance.[1] If the maximum of the p-values corresponding to the regression coefficients for each potential model exceeded some pre-specified benchmark – 0.1 and 0.65, for open box office and retail movie sales, respectively – the corresponding model was excluded. After this pre-selection step, they were left with 95 and 56 models respectively. Using the same data set, Xie & Lehrer (2018) use model screening before model averaging due to a large set of regressors (23 and 29 for open box office and movie unit sales respectively) resulting in hundreds of millions of potential candidate models ($2^{23} = 8,388,608$ and $2^{29} = 526,870,912$). Thus, the authors use model screening based on an automated general-to-specific (PcGets) model selection algorithm to reduce the set of models for model selection and model averaging methods. First, based on OLS results, they restrict each model to a constant and at most 7 (11) relatively signif-

[1] Xie and Lehrer assume that the DGP for outcome y_i is given as $y_i = \mu_i + u_i$, where $\mu_i = \sum_{j=1}^{\infty} \beta_j x_{ij}$, where u_i is mean-zero and homoskedastic. Their candidate models take the following form: $y_i = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} + u_i^{(m)}$ for $i = 1, ..., n$ and $m = 1, ..., M$, where $\beta_j^{(m)}$ is a coefficient in the model m and $x_{ij}^{(m)}$ is a regressor in the model m. The p-values are estimated for each coefficient within each model.

icant parameters for open box office (movie unit sales). Then, PcGets is used to control the total number of potential models by examining estimated p-values for each parameter in each potential model. If the maximum of these p-values exceeds some pre-specified benchmark, the corresponding model is excluded, eliminating models with low-t-statistic coefficients from the set of candidate models. This results in 105 and 115 potential models for open box office and retail movie unit sales respectively. The authors acknowledge that this reduction in potential models is severe, but justify it by stating that only a "handful" of models account for more than 95% of the total weight of the model average estimate. Xie & Lehrer (2018) propose a strategy for model screening with model averaging using a Mallows-type criterion (an extension of X. Zhang & Liang, 2016) to improve the selection of candidate models. This significantly increases computational efficiency with no lack of forecast accuracy. They show that the model average estimator using weights obtained from a screened model set is asymptotically optimal in the sense of achieving the lowest possible mean squared error (MSE), even compared to a model average estimator that used all potential candidate models in its set.

Model averaging post-lasso applies the lasso to remove irrelevant variables from candidate models prior to model averaging (Xie & Lehrer, 2017, 2018). One could also employ model screening using an information criterion like AIC or BIC to select the top M^* models prior to model averaging, where M^* is an integer pre-specified by the researcher. Yuan & Yang (2005) propose a model combining method, *adaptive regression by mixing with model screening* (ARMS), which adds a model screening step to *adaptive regression by mixing* (ARM) proposed by Yang (2001). Model screening shrinks the set of candidate models before combining, which can reduce the computational cost, as there will be fewer weights to calculate and to assign to models, and can improve estimation accuracy by removing very poor candidate models.

Let $y_i = f(X_i) + \epsilon_i$, $i = 1, ..., n$, where y_i is the response variable, X_i is a vector of d explanatory variables, $f(\cdot)$ is the true regression function (i.e. the data generating process

or DGP), and ϵ is the random error. Consider M candidate models, where model m is given by $y_i = f_m(X_i; \theta_m) + \epsilon_i$. Let Γ denote the set of all candidate models being considered and M be the size of Γ. For model screening, the data are split into two equal parts: $Z^{(1)} = (X_i, y_i), 1 \leq i \leq n/2$ is for estimation and screening, and $Z^{(2)} = (X_i, y_i), n/2+1 \leq i \leq n$ is for prediction. Model screening proceeds as follows. First, estimate each candidate model m using $Z^{(1)}$ to obtain $\hat{f}_m(x) = \hat{f}_m(x; \hat{\theta}_m)$. For each model m, obtain an estimate of σ_m^2, $\hat{\sigma}_m^2$. Next, calculate the model selection criterion (such as AIC or BIC) for each model m using $Z^{(2)}$. Rank the models by AIC and/or BIC value and keep only the top M^* models, where M^* is an integer chosen by the researcher. Let Γ_s denote the set of chosen candidate models and M_s be the size of Γ_s. These are the models that are used in the combining step of ARMS. Alternatively, these models could be used in model averaging. The researcher must balance the cost of screening (increasing the risk of omitting a good model) with the potential advantage of screening (reducing the negative influence of poor models). The potential advantage is greater when M is large and M_s is small. In other words, model screening should balance the probability of capturing the "best" model and the size of Γ_s.

Xie & Lehrer (2018) show that using a screened model set produces a model average estimator that is asymptotically optimal in the sense of achieving the lowest possible MSE. While model screening may improve estimation or prediction by removing poor candidate models, it has some limitations. Model screening requires the researcher to specify one or more models at the outset (for example, specifying a general unrestricted model). Therefore, model screening relies on ad hoc decisions on the part of the researcher. The aim of this chapter is to introduce statistical procedures in order to move away from this type of ad hoc decision-making. Additionally, model screening is essentially variable selection and does not provide any guidance with regard to the functional form specification of candidate models. Model screening has the potential to result in a number of candidate models tied to the sample size and assumed smoothness class. However, current practice relies on

the researcher specifying the maximum number of candidate models they wish to keep for model averaging; therefore, the number of candidate models is arbitrarily chosen. Additional research is required in order to tie this number to the sample size and assumed DGP. Thus, while model screening has some advantages and may have one of the ideal features of the set of candidate models outlined earlier, additional methods are required to generate a rich set of bases in order to capture a wide range of potential DGPs.

2.3 Recursive Partitioning-Based Algorithms

2.3.1 Machine Learning in Combination with Model Averaging

Xie & Lehrer (2018) use recursive partitioning in combination with model averaging in a forecasting exercise using film industry and social media data to predict movie success. The goal is to use this hybrid model average/learning strategy to capture richer patterns of heterogeneity that machine learning or econometric methods alone may fail to capture, as well as to improve forecast accuracy. The authors propose a hybrid strategy that uses recursive partitioning to first develop sub-groups or sub-regions, then implement model averaging within these groups to generate forecasts. Allowing for model uncertainty in the leaves of a regression tree allows for richer heterogeneity in the resulting forecasts.

Recursive partitioning strategies, such as *classification and regression trees* (CART), partition the data into sub-regions by splitting on the domain of a regressor. The split is chosen at the point where the sum of squared residuals (SSR) is minimized, resulting in two nodes. The partitioning, or splitting, continues for each node until further splits no longer contribute to the accuracy of the forecast. The final terminal nodes are called the leaf nodes or leaves. For each leaf l, the forecast is the fitted values from a regression model of the form $y_i = a + u_i, i \in l$, where a is a constant and u_i is the error term. The ordinary least squares estimate of a is $\hat{a} = \bar{y}_{i \in l}$. CART are able to capture non-linearities in the data. However, they inherently assume that any resulting heterogeneity in the outcomes within

each terminal leaf of the tree is random. CART perform well in-sample but may perform poorly out-of-sample.

Xie & Lehrer (2018) use two recursive partitioning-based ensemble methods in combination with model averaging: *random forests* and *bagging* (short for *bootstrap aggregating*). These methods create multiple bootstrapped samples of size n with replacement and multiple decision trees from a single sample, then combine the predictions from each tree using an aggregation technique with weights based on the sample proportion in each leaf of the tree. This improves predictive accuracy out-of-sample compared to CART. No structure is imposed on the data, which is not the case for parametric econometric models. However, they do assume homogeneity due to the use of the equally weighted sum of squared residuals in the algorithms.

Using random forest, the authors suggest that at each tree leaf l, there is a sequence of M linear candidate models, in which the regressors of each model m, $m = 1, ..., M$, is a subset of the regressors belonging to that tree leaf. The regressors $X_{i\in l}^m$ for each candidate model within each tree leaf are such that the number of regressors $k_l^m \ll n_l$ for all m, where n_l is the number of observations in a tree leaf l. Using these candidate models, the method then performs model average estimation and obtains

$$\underset{(K\times 1)}{\hat{\beta}_l(\omega)} = \sum_{m=1}^{M} \omega^m \underset{(K\times 1)}{\tilde{\beta}_l^m}, \tag{2.5}$$

which is a weighted average of the "stretched" estimated coefficients $\tilde{\beta}_l^m$ for each candidate model m. The $K \times 1$ sparse coefficient vector $\tilde{\beta}_l^m$ is constructed from the $k_l^m \times 1$ least squares coefficient vector $\hat{\beta}_l^m$ by filling in the extra $K - k_l^m$ elements with 0s. Therefore, the forecast for all observations is

$$\hat{y}_{t\in l} = X_{t\in l}^p \hat{\beta}_l(\omega). \tag{2.6}$$

Model average bagging (MAB) applies model averaging to each of the B samples used to

construct a bagging tree. In other words, for each bootstrap sample b, there is a sequence of M linear candidate models, in which the regressors of each model m, $m = 1, ..., M$, is a subset of the total number of regressors K. The final MAB forecast is an equally weighted average of the B model average tree forecasts. One difference between *model average random forests* (MARF) and MAB is that MARF only considers $k \leq K$ predictors for splitting at each node, so that the candidate model set for each leaf l considers only those k regressors.

The authors conduct a simulation to assess the relative prediction efficiency of different estimators with different sets of covariates. The sample consists of movies released in North America between October 2010 and June 2013. The data are from the film industry as well as social media (Twitter), with the latter containing sentiment towards a particular movie and volume of tweets regarding a particular movie as predictor variables. The set of estimation strategies evaluated consist of traditional econometric approaches (including model specification, model selection and model averaging approaches), model screening approaches, machine learning approaches (regression trees, bagging, and random forests), and the proposed model average learning methods (MARF and MAB). The exercise involves shuffling data into a training set of size n_T and an evaluation set of size $n_E = n - n_T$. The training set is used to first obtain estimates from each strategy then forecast the outcomes for the evaluation set. The evaluation set is used to evaluate each strategy in terms of mean squared forecast error (MSFE) and mean absolute forecast error (MAFE), which are defined by

$$\text{MSFE} = \frac{1}{n_E}(y_E - x_E\hat{\beta}_T)'(y_E - x_E\hat{\beta}_T) \tag{2.7}$$

$$\text{MAFE} = \frac{1}{n_E}|y_E - x_E\hat{\beta}_T|'\iota_E, \tag{2.8}$$

where ι_E is a $n_E \times 1$ vector of ones. This exercise is repeated 10,001 times for varying sizes of n_E.

A model screening step is applied before all model selection and model average exercises because, with a large regressor set (23 and 29 for open box office and movie unit sales respectively), there are millions of potential candidate models ($2^{23} = 8,388,608$ and $2^{29} = 526,870,912$ respectively). After model screening, 105 and 115 potential models remain for open box office and retail movie unit sales, respectively. Results are reported relative to the MSFE and MAFE of model selection by the heteroskedasticity-robust Mallows' C_p criterion (HRC$_p$), a model selection strategy. The authors find that recursive partitioning algorithms like bagging and random forests alone yield on average 30-40% gains in forecast accuracy relative to econometric approaches that use either a model selection criteria or model averaging. Out of all the strategies evaluated, MAB performed the best. Adding model averaging to bagging led to gains of 10%. MARF had relatively moderate performance. The proposed model average learning methods may perform better relative to pure econometric approaches because the full set of predictors is considered. Recall that model screening was used as a preliminary step prior to model averaging and model selection so that only a subset of variables were used in estimation. However, the authors rule out this explanation because they found that when MAB and MARF were restricted to the variables obtained from model screening, there still existed large gains in predictive performance of the hybrid strategies relative to econometric strategies. This suggests that these gains may come from relaxing the linearity assumption rather than from using a larger set of predictors.

A major drawback of this approach is that it still relies on a set of ad hoc parametric models. For both MARF and MAB, the final model average learning estimator is based on a sequence of M linear candidate models. The goal of this chapter is to move away from parametric model specifications, which may not be flexible enough to approximate the unknown underlying DGP. While combining machine learning (in particular, recursive partitioning-based methods) with model averaging may result in improved performance in this context, in general, a broader set of bases is more desirable in order to encompass a

wider range of potential DGPs. An alternative recursive partitioning-based algorithm is considered in the following section.

2.3.2 Multivariate Adaptive Regression Splines

Standard parametric regression requires the researcher to first determine which explanatory variables to include in the regression, then to explicitly identify and incorporate the specific degree of interaction for each explanatory variable. Today, given the rise of big data, a dataset can easily contain dozens, if not hundreds, of variables. Manually determining if and how each variable enters into the regression equation would be labour-intensive and the resulting regression equation may not fully capture the underlying DGP when the researcher is forced to make arbitrary decisions on, for example, the degree of interactions.

Multivariate adaptive regression splines (MARS) (Friedman, 1991b) offers an alternative approach to standard parametric regression that can potentially capture non-linearities present in the data. It is a method for flexible nonparametric regression modeling of high dimensional data using an expansion in product spline basis functions, where the number of basis functions and the number of parameters are automatically determined by the data. MARS is able to capture high order interactions and, unlike recursive partitioning, produces continuous models with continuous derivatives. Although it may produce complex models with high order interactions, the approximating functions are interpretable through an ANOVA decomposition and visualized by slicing. Additionally, extensions to MARS enable the algorithm to handle categorical explanatory variables, nested variables, and missing input values (Friedman, 1991a). Finally, the flexibility and adaptability of MARS allows the researcher to impose constraints on the final model based on knowledge of the system under study, such as limiting the maximum interaction order or limiting the specific variables that can participate in interactions. One limitation of MARS is that the implementation can be computationally demanding; however, updating formulae can be used to reduce the computational demands (see Friedman, 1993 for details). Another

limitation is that collinearity may lead to spurious interaction effects.

MARS is an extension of *recursive partitioning*. Recursive partitioning creates a decision tree that attempts to correctly classify observations by splitting the data into sub-regions (or partitions) based on independent variables. Partitioning is done through the recursive splitting of previous sub-regions, beginning with the entire domain and continued until a stopping criterion is met and a large number of sub-regions have been generated. Sub-regions are then re-combined in a reverse manner until an optimal set is reached based on a criterion that penalizes lack of fit and increasing the number of regions. Recursive partitioning is viewed as a geometrical procedure; however, geometrical elements (such as regions and splitting) can be replaced with arithmetic counterparts (such as adding and multiplying). In this way, MARS can be built as a stepwise regression procedure.

Again, let $y_i = f(X_i) + \epsilon_i$, $i = 1, ..., n$, where y_i is the response variable, X_i is a vector of q explanatory variables, $f(\cdot)$ is the true regression function (i.e. the data generating process or DGP), and ϵ is the random error. A challenge is selecting the appropriate functional class for $\hat{f}(X_i)$, as it should represent the unknown function $f(X_i)$ as accurately as possible while avoiding overfitting. Let the approximating function be an expansion in a set of basis functions:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} a_m B_m(\mathbf{x}). \tag{2.9}$$

where the weights a_m are estimated by minimizing the sum of square residuals ($SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ where y_i is the response variable and $\hat{y}_i$ are the fitted values) (Hastie, Tibshirani, & Friedman, 2009, p. 322).

There are M basis functions B_m that take the following form:

$$B_m(\mathbf{x}) = I[\mathbf{x} \in R_m], \tag{2.10}$$

where $I[\cdot]$ is an indicator function and R_m are disjoint sub-regions of the domain D such

that $\mathbf{x} \in R_m$. Given that the sub-regions are disjoint, only one basis function is non-zero for any point **x**. The goal is to simultaneously derive a good set of basis functions (in other words, sub-regions) based on the data and to estimate the parameters of each function in each sub-region to best fit the data.

The algorithm proceeds as follows. Setting the initial region to the entire domain D, a forward iterative splitting procedure generates a set of basis functions given a final number of regions (or basis functions), M_{max}. A basis function B_{m^*}, predictor variable x_{v^*}, and split point or knot t^* are selected by minimizing the lack-of-fit (LOF) of a model with B_{m^*} replaced by its product with the step function $H[+(x_{v^*} - t^*)]$ and the addition of a new basis function that is the product of B_{m^*} and the reflected step function $H[-(x_{v^*} - t^*)]$. This is equivalent to splitting the region R_{m^*} on variable v^* at split point t^*. This produces basis functions of the form

$$B_m(\mathbf{x}) = \prod_{k=1}^{K_m} H[s_{km} \cdot (x_{v(k,m)} - t_{km})], \tag{2.11}$$

where $H[\cdot]$ is the step function, K_m is the total number splits, or knots, that gave rise to basis function B_m, $s_{km} = \pm 1$ indicates the left/right sense of the associated step function, $v(k, m) = 1, ..., q$ labels the predictor variables, and t_{km} is the knot location on each corresponding variable. The optimal number of knots can be selected using cross-validation and the knot location can be chosen using the so-called TURBO method, a forward stepwise strategy for knot placement (Friedman & Silverman, 1989).

The forward iterative splitting procedure generates more basis functions than optimal, deliberately overfitting, to allow for the next step: backwards stepwise deletion. Given the disjoint sub-regions, removing a single basis function will leave a gap in the predictor variable space. Thus, backwards stepwise subset selection removes basis functions that no longer contribute to the accuracy of fit by deleting splits rather than regions (or basis functions) and selects the final appropriately sized basis function set (M^*) from the set of

M_{max} basis functions. Typically, $M_{\text{max}} = 2M^*$. The backwards stepwise subset selection strategy uses a modified generalized cross-validation criterion (GCV) to select the final functional estimate:

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_M(\mathbf{x}_i)]^2}{\left[1 - \frac{C(M)}{n}\right]^2}. \tag{2.12}$$

The numerator represents LOF and the denominator represents a penalty for increasing model complexity, defined as $C(M) = (d/2+1)M + 1$, where d is a smoothing parameter that can be selected by bootstrapping or cross-validation. The model that minimizes this criterion is taken to be the final model.

A major limitation of recursive partitioning is that the approximating function is discontinuous at sub-region boundaries due to the use of step functions, which severely limits the accuracy of the approximation. A modification to the aforementioned algorithm that will produce continuous functions with continuous derivatives is to replace the (discontinuous) step function $H[\cdot]$ with a (continuous) truncated power basis function of the form

$$b_q^{\pm}(x - t) = [\pm(x - t)]_+^q, \tag{2.13}$$

where q is the order of the spline, t is the knot location, and the subscript $+$ indicates the positive part of the argument. For $q > 0$, the spline approximation is continuous and has $q - 1$ continuous derivatives. The step function $H[\cdot]$ is considered to be a two-sided truncated power basis function for $q = 0$ splines. The truncated power basis automatically selects both the number of knots (K_m) (global smoothing) and their locations (t) (local smoothing) (Friedman & Roosen, 1995).

Another limitation of recursive partitioning is that it is unable to provide good approximations to some simple functions because the curse of dimensionality requires a large number of basis functions to get a good approximation to functions of low order interactions. Thus, the algorithm is modified so that parent functions $B_{m^*}(\mathbf{x})$ are simply not

removed after their split and, consequently, are eligible for further splitting. Previously, a parent function was removed after its split and replaced with its product with a truncated power spline basis function (or, in recursive partitioning, the step function $H[\cdot]$). Friedman (1991b) also restricts the product associated with each basis function to factors involving distinct predictor variables. These modification makes MARS more adaptive than recursive partitioning because it allows recursive splitting of all basis functions in the model instead of only those that are terminal and gives MARS the ability to make good approximations of simple functions, such as linear and additive ones. Finally, Friedman (1991b) suggests imposing continuity (q) of only the approximating function and its first derivative and argues that there is little to be gained by imposing continuity beyond that of the first derivative. These modifications complete the MARS algorithm. After generating $M_{\max}q$ (where, typically, $q = 1$) multivariate spline basis functions of the form

$$B_m^{(q)}(\mathbf{x}) = \prod_{k=1}^{K_m} [s_{km} \cdot (x_{v(k,m)} - t_{km})]_+^q. \tag{2.14}$$

This set of basis functions then undergoes a backwards stepwise deletion strategy to produce a final set of basis functions. Unlike recursive partitioning, the corresponding regions of each basis function overlap rather than being disjoint. Thus, removing a basis function does not produce a hole in the predictor space and so a simple one-at-a-time backward stepwise procedure, akin to regression subset selection, can be used. This constructs a sequence of $M_{\max} - 1$ models, each one having one less basis function than the previous one in the sequence. Knot locations associated with this approximation are used to derive a piecewise cubic basis with continuous derivatives. The "best" model (in terms of LOF) is returned upon termination.

The final model takes the following form:

$$\hat{f}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} [s_{km} \cdot (x_{v(k,m)} - t_{km})]_+, \tag{2.15}$$

where a_0 is the coefficient of the constant basis function $B_1 = 1$, the sum is over the basis functions B_m, $m = 1, ..., M$, that survive backwards deletion, and $s_{km} = \pm 1$. This model can be recast into what Friedman (1991b) calls the ANOVA decomposition, which allows for greater interpretability.

In simulations, the MARS algorithm performs well in terms of predicted squared error (Friedman, 1991b). The MARS algorithm is flexible, adaptable, and avoids finding structure when there is none (such as in cases of pure noise or simple additive models). The main advantage of MARS over existing methodologies is realized in high dimensional settings, but it is also competitive in low dimensions ($q \leq 2$). The algorithm is likely to favour lower order interaction terms, which has an advantage when the true underlying DGP is not dominated by high order interactions. For small sample sizes, the MARS algorithm will try to produce models involving lower order interactions, whereas for larger sample sizes, it will likely favour high order interactions as potential candidates. MARS allows the researcher to specify the maximum interaction order, mi, where $mi = 1$ is an additive model, $mi = 2$ has interactions involving at most 2 variables, and $mi = q$ has no constraints on the number variables that can enter into interactions.

MARS is a promising method for generating the set of candidate models to be used in model averaging because it can generate a rich set of basis functions which can then be combined into a single model. This is an improvement over the candidate models used in the model average learning methods proposed by Xie and Lehrer, who showed that combining machine learning and model averaging improved estimator performance. Rather than combining the bases in the way that Friedman suggests, where the weights a_m are estimated by minimizing the SSR, I propose combining the bases using a model average criterion, such as MMA or JMA.

It is known that the forward iterative splitting procedure produces more bases than optimal, after which the set is trimmed to an appropriate size by backwards stepwise deletion, resulting in M^* basis functions from an initial set of $M_{\max}$. The backwards stepwise dele-

tion strategy is akin to a model screening step. This means that MARS has the potential to tie the number of basis functions to the sample size and/or assumed smoothness class, one of the desirable features of a set of candidate models. However, just like model screening, the desired number of candidate models is chosen arbitrarily: $M_{\max} = 2M^*$, where M^* is an integer specified by the researcher .

It seems like the number of parameters is tied to the sample size and assumed smoothness class because the algorithm automatically generates bases based on the data and performs well in simulations. However, we don't know how this mechanism works exactly (it is unclear to this author, at least, how exactly the number of parameters is tied to the sample size and underlying DGP).

2.4 Averaging over Nonparametric Models

Racine, Li, & Zheng (2018) propose a fully nonparametric approach that averages over mixed-data kernel-weighted spline regressions. The set of candidate models admits both continuous and categorical predictors. The goal of the proposed approach is to not only average over a sufficiently rich set of candidate models in order to consistently estimate a large class of potential DGPs, but to also present a method that is valuable to practitioners who are to working with parametric models. Spline regressions have a number of advantages, namely that they are global in nature, computationally efficient (as they are a simple weighted least squares problem), and accessible to those who routinely use least squares and polynomials, making them a good alternative to parametric candidate models. Additionally, the B-spline basis functions offer the maximally differentiable spline basis. However, one major limitation of regression spline methods is that there is a loss in efficiency due to their inability to handle the presence of categorical predictors without resorting to sample-splitting. Ma, Racine, & Yang (2015) propose a tensor-product spline approach that overcomes the efficiency loss from sample-splitting in traditional regression

splines, is sufficiently flexible to allow for non-linearities, and is asymptotically normal, allowing for the construction of confidence intervals. Nonparametric mixed-data kernel regression methods have been proposed, but many applied economists resist their use due to certain drawbacks: they are local, rather than global, approximations; bandwidth selection is not always straightforward and can be numerically demanding; and it can be difficult to interpret results. Thus, the tensor-product kernel-weighted spline approach provides a good alternative to kernel estimators that admit both continuous and categorical predictors.

Racine et al. (2018) make use of the nonparametric approach of Ma et al. (2015) to generate a flexible set of candidate models for model averaging. Consider a nonparametric regression model with both continuous and categorical predictors:

$$\begin{aligned} Y_i &= \mu_i + \epsilon_i \\ &= g(X_i, Z_i) + \epsilon_i, \qquad\qquad i = 1, ..., n, \end{aligned} \tag{2.16}$$

where $g(\cdot, \cdot)$ is an unknown smooth function, X is a q-dimensional vector of continuous predictors, and Z is an r-dimensional vector of categorical predictors. Note that $q < \infty$ and $r < \infty$. Assume without loss of generality that for $1 \leq l \leq q$, each X_l is distributed on interval $[a_l, b_l] = [0, 1]$. Let z_s denote the sth component of z, $1 \leq s \leq r$ and assume z_s takes c_s different values in $D_s = \{0, 1, ..., c_s - 1\}$, where $s = 1, ..., r$ and c_s is a finite positive constant. To allow for heteroskedasticity, assume $E[\epsilon_i | X_i, Z_i] = 0$ and $E[\epsilon_i^2 | X_i, Z_i] = \sigma^2(X_i, Z_i) \equiv \sigma_i^2$, $i = 1, ..., n$.

The goal is to approximate μ_i using M candidate nonparametric regression models to approximate the regression equation above. For $m = 1, ..., M$, let the mth candidate model take the following form:

$$Y_i = g_{(m)}(X_{i,(m)}, Z_{i,(m)}) + e_{i,(m)}, \qquad\qquad i = 1, ..., n, \tag{2.17}$$

where $g_{(m)}(\cdot, \cdot)$ is an unknown smooth function, $X_{i,(m)}$ is a q_m-dimensional sub-vector of

X_i, $Z_{i,(m)}$ is a r_m-dimensional sub-vector of Z_i, and $e_{i,(m)}$ represents the approximation error in the mth model.

To handle the presence of both continuous and categorical predictor variables, estimate each candidate model by tensor product polynomial splines, $\mathcal{B}_{(m)}(x_{(m)})$, weighted by categorical kernel functions, $L(Z_{i,(m)}, z_{(m)}, \lambda_{(m)})$ (Ma et al., 2015).

To specify the categorical kernel function, let the univariate categorical kernel function $l(Z_{il,(m)}, z_{l,(m)}, \lambda_{l,(m)})$ be defined as follows:

$$l(Z_{il,(m)}, z_{l,(m)}, \lambda_{l,(m)}) = \begin{cases} \lambda_{l,(m)} & \text{if } Z_{il,(m)} \neq z_{l,(m)}, \\ 1 & \text{otherwise.} \end{cases} \tag{2.18}$$

Then the categorical kernel function, $L(Z_{i,(m)}, z_{(m)}, \lambda_{(m)})$, is:

$$\begin{aligned} L(Z_{i,(m)}, z_{(m)}, \lambda_{(m)}) &= \prod_{l=1}^{r_m} l(Z_{il,(m)}, z_{l,(m)}, \lambda_{l,(m)}) \\ &= \prod_{l=1}^{r_m} \lambda_{l,(m)}^{1(Z_{il,(m)} \neq z_{l,(m)})}, \end{aligned} \tag{2.19}$$

where $1(\cdot)$ is the indicator function, and $\lambda_{(m)} = (\lambda_{1,(m)}, ..., \lambda_{r_{(m)},(m)})'$ is the $r_{(m)}$-dimensional vector of bandwidths for each categorical predictor.

To specify the tensor product polynomial splines, $\mathcal{B}_{(m)}(x_{(m)})$, let $\{t_{j_l,l,(m)}\}_{j_l=1}^{N_{l,(m)}}$ be a sequence of interior knots. Let $K_{n,l,(m)} = N_{l,(m)} + d_{l,(m)}$ where $N_{l,(m)}$ is a pre-selected integer representing the number of interior knots and $d_{l,(m)}$ is the spline order. Let $B_{l,(m)}(x_{l,(m)}) = \{B_{j_l,l,(m)}(x_{l,(m)}) : 1 - d_{l,(m)} \leq j_l \leq N_{l,(m)}\}$ be a basis system of the space $G_{l,(m)} = G_{l,(m)}^{(d_{l,(m)}-2)}$ of polynomial splines of order $d_{l,(m)}$.

Define the space of tensor product polynomial splines by $\mathcal{G}_{(m)} = \otimes_{l=1}^{q_m} G_{l,(m)}$, where

$\mathcal{G}_{(m)}$ is a linear space of dimension $K_{(m)} \equiv K_{n,(m)} = \prod_{l=1}^{q_m} K_{n,l,(m)}$. Then:

$$\begin{aligned}\mathcal{B}_{(m)}(x_{(m)}) &= \left[\{\mathcal{B}_{j_1,...,j_{q_m}}(x_{(m)})\}_{j_1=1-d_{1,(m)},...,j_{q_m}=1-d_{q_m,(m)}}^{N_{1,(m)},...,N_{q_m,(m)}}\right] \\ &= B_{1,(m)}(x_{1,(m)}) \otimes ... \otimes B_{q_m,(m)}(x_{q_m,(m)}) \end{aligned} \tag{2.20}$$

is a basis system of the space $\mathcal{G}_{(m)}$, where $x_{(m)} = (x_{l,(m)})_{l=1}^{q_m}$.

Approximate the function $g_{(m)}(x_{(m)}, z_{(m)})$ by $\mathcal{B}_{(m)}(x_{(m)})'\beta_{(m)}(z_{(m)})$, where $\beta_{(m)}(z_{(m)})$ is a $K_{n,(m)}$-dimensional vector with $K_{n,(m)} \to \infty$ as $n \to \infty$. Estimate $\beta_{(m)}(z_{(m)})$ as follows:

$$\hat{\beta}_{(m)}(z_{(m)}) = \arg \min_{\beta \in \mathbb{R}^{K_{n,(m)}}} \sum_{i=1}^{n} [Y_i - \mathcal{B}_{(m)}(X_{i,(m)})'\beta]^2 L(Z_{i,(m)}, z_{(m)}, \lambda_{(m)}), \tag{2.21}$$

where $\mathcal{B}_{(m)}(\cdot)$ and $L(\cdot)$ are defined above. Thus, $\hat{g}_{(m)}(x_{(m)}, z_{(m)}) = \mathcal{B}_{(m)}(x_{(m)})'\hat{\beta}_{(m)}(z_{(m)})$.

Let $\mathbf{B}_{(m)} = [\{\mathcal{B}_{(m)}(X_{1,(m)}), ..., \mathcal{B}_{(m)}(X_{n,(m)})\}']_{n \times K_{(m)}}$ and $\mathcal{L}$ be a diagonal matrix with $L(Z_{i,(m)}, z_{(m)}, \lambda_{(m)})$, $1 \leq i \leq n$, as the diagonal entries.

Then $\hat{\beta}_{(m)}(z_{(m)})$ can be written as as a linear function of Y:

$$\hat{\beta}_{(m)}(z_{(m)}) = (\mathbf{B}'_{(m)}\mathcal{L}_{z_{(m)}}\mathbf{B}_{(m)})^{-1}\mathbf{B}'_{(m)}\mathcal{L}_{z_{(m)}}Y. \tag{2.22}$$

Thus, the estimator is simply a weighted least squares estimator where the continuous predictors have been replaced by their B-spline representations.

With this, $\mu_{i,(m)}$ can be estimated by:

$$\begin{aligned}\hat{\mu}_{i,(m)} &= \mathcal{B}_{(m)}(X_{i,(m)})'\hat{\beta}_{(m)}(Z_{i,(m)}) \\ &= \mathcal{B}_{(m)}(X_{i,(m)})'(\mathbf{B}'_{(m)}\mathcal{L}_{Z_{i,(m)}}\mathbf{B}_{(m)})^{-1}\mathcal{B}'_{(m)}\mathcal{L}_{Z_{i,(m)}}Y. \end{aligned} \tag{2.23}$$

This can be expressed as $\hat{\mu}_{(m)} = P_{(m)}Y$ where $\hat{\mu}_{(m)} = (\hat{\mu}_{1,(m)}, ..., \hat{\mu}_{n,(m)})'$ and $P_{(m)}$ is a n-dimensional square matrix with the ith row vector being $\mathcal{B}_{(m)}(X_{i,(m)})'(\mathbf{B}'_{(m)}\mathcal{L}_{Z_{i,(m)}}\mathbf{B}_{(m)})^{-1}\mathcal{B}'_{(m)}\mathcal{L}_{Z_{i,(m)}}$.

Note that the number of interior knots, $N_{l,(m)}$, as well as the bandwidths, $\lambda_{(m)}$, can be jointly selected by minimizing a cross-validation criterion:

$$\text{CV}(N, \lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - B_d(X_i)'\hat{\beta}_{-i}(Z_i))^2, \tag{2.24}$$

where $\hat{\beta}_{-i}(Z_i)$ is the leave-one-out estimate of β. Ma et al. (2015) illustrate the finite-sample behaviour of the fully data-driven cross-validation selection of N and λ by considering four simple DGPs, using the cubic B-spline basis throughout. The results show that the choices of N and λ differ depending on the DGP; larger values of λ are selected when Z_i is independent of Y_i. This method is computationally more efficient than multivariate cross-validated kernel regression.

Once each candidate model has been estimated by tensor product polynomial splines, the next step in this procedure is to select the model weights. Let $\omega = (\omega_1, ..., \omega_M)'$ be the weight vector, and let $0 \leq \omega_m \leq 1$, $m = 1, ..., M$ and $\sum_{m=1}^{M}\omega_m = 1$. Let $P(\omega) = \sum_{m=1}^{M}\omega_m P_{(m)}$. Then the model average estimator of μ is given by

$$\begin{aligned}\hat{\mu}(\omega) &= \sum_{m=1}^{M}\omega_m\hat{\mu}_{(m)} \\ &= P(\omega)Y.\end{aligned} \tag{2.25}$$

Racine et al. (2018) propose a Mallows-type criterion for selecting the model weights:

$$C_n(\omega) = \frac{1}{n}||P(\omega)Y - Y||^2 + \frac{2}{n}\text{tr}[P(\omega)\Omega], \tag{2.26}$$

where $\Omega = \text{E}[\epsilon\epsilon'] = \text{diag}(\sigma_1^2, ..., \sigma_n^2)$is the variance-covariance matrix of the highest dimension model. When Ω is known, the optimal choice of the weight vector is given by

$$\tilde{\omega} = \arg\min_{\omega} C_n(\omega). \tag{2.27}$$

In this case, the optimal model average estimator of μ is $\hat{\mu}(\tilde{\omega}) = P(\tilde{\omega})Y$. Under certain regularity conditions, the weight vector $\tilde{\omega}$ is asymptotically optimal.

When Ω is unknown, the feasible Mallows-type criterion is:

$$\hat{C}_n(\omega) = \frac{1}{n}||P(\omega)Y - Y||^2 + \frac{2}{n}\text{tr}[P(\omega)\hat{\Omega}(\omega)] \tag{2.28}$$

where $\hat{\Omega}(\omega) = \text{diag}(\hat{\epsilon}_1^2(\omega), ..., \hat{\epsilon}_n^2(\omega))$ and the new optimal weights are:

$$\hat{\omega} = \arg\min_{\omega} \hat{C}_n(\omega). \tag{2.29}$$

In this case, the optimal model average estimator of μ is $\hat{\mu}(\hat{\omega}) = P(\hat{\omega})Y$. The weight vector $\hat{\omega}$ is (still) asymptotically optimal.

2.4.1 Evidence from Monte Carlo Experiments and Empirical Examples

Racine et al. (2018) conduct two Monte Carlo experiments to assess the finite-sample performance of the proposed kernel-weight spline model average approach. The data generating process (DGP) was chosen to be:

$$y = x_1 + x_2 + x_1x_2 + x_1^2 + x_2^2 + x_3 + x_1x_3 + x_2x_3 + x_4 + x_1x_4 + x_2x_4 + \epsilon, \tag{2.30}$$

where x_1, x_2, x_4 are continuously distributed as U$[-1, 1]$ and x_3 has discrete support, generated from the binomial distribution with $n = 3$ and $p = 1/2$.

In the first experiment, there were six under-specified candidate models that include

different combinations of the independent variables:

$$\text{Model 1: } y_i = g_1(x_{1i}) + \epsilon_i \tag{2.31}$$

$$\text{Model 2: } y_i = g_2(x_{2i}) + \epsilon_i \tag{2.32}$$

$$\text{Model 3: } y_i = g_4(x_{1i}, x_{2i}) + \epsilon_i \tag{2.33}$$

$$\text{Model 4: } y_i = g_5(x_{1i}, x_{3i}) + \epsilon_i \tag{2.34}$$

$$\text{Model 5: } y_i = g_6(x_{2i}, x_{3i}) + \epsilon_i \tag{2.35}$$

$$\text{Model 6: } y_i = g(x_{1i}, x_{3i}, x_{3i}) + \epsilon_i \tag{2.36}$$

For each candidate model, cross-validation was used to select the degree of the tensor spline as well as the smoothing parameter for the discrete predictor, then each model was estimated by the nonparametric method described previously. Weights were assigned to each model using the Mallows-type criterion $\hat{C}_n(\omega)$. This exercise was repeated 1,000 times. The kernel-weight spline approach was then compared in terms of mean MSE (over 1,000 replications) to model selection using the AIC, model selection using the BIC, model selection using Mallows' C_p, and, finally, the largest model. The authors find that their proposed model averaging approach has the smallest estimation mean MSE in all cases.

In the second Monte Carlo experiment, the set of candidate models contains the true model, which coincides with the largest model. The experiment uses the same DGP and the same set of candidate models as the first experiment, with the exception that x_4 is removed from the DGP and set of candidate models. Following the same steps of the first Monte Carlo, the authors find that even when the true model is included in the set of candidate models (which would be highly unlikely in reality), model averaging can outperform model selection in small sample settings.

In an empirical setting, the authors use panel data to model the growth rates of per capita GDP using panel data for countries. The predictors are OECD status (categorical), human capital (continuous), and initial GDP (continuous). The data were shuffled into two

samples, one of size $n_1 = 600$ and the other of size $n_2 = 16$, where n_1 were used for estimation and n_2 were used to evaluate predictive performance. This exercise was repeated 1,000 times. The proposed kernel-weighted regression spline model average approach was compared in terms of mean predicted square error (PSE) to models selected by AIC, BIC, and Mallows' C_p, as well as the largest model. They use the same set of candidate models as in the Monte Carlo experiments (described above). Their results show that no model selection method does better than the proposed kernel-weighted spline regression model averaging approach, and all six candidate models receive non-zero model weights, indicating the presence of model uncertainty.

2.4.2 Heuristics

The goal of this chapter is to recommend a method for generating an ideal set of candidate models to be used in model averaging. This is an important step when dealing with model uncertainty, as the resulting model average estimator will inherit properties from the candidate models. However, the selection of the candidate models is typically done in an ad hoc manner. The nonparametric kernel-weighted spline regression model average approach described above produces a rich set of candidate models that are able to capture a wide range of potential DGPs. This method is fully data-driven and does not rely on arbitrary decisions on the part of the researcher. In this section, some heuristics are developed in order to satisfy the three features of an ideal set of candidate models, which are repeated below:

1. number of candidate models (M) tied to the sample size (n) and assumed smoothness class (DGP),
2. number of parameters tied to the sample size and assumed smoothness class, and
3. a broad set of functions ("bases") to cover a wide range of potential DGPs that are simple enough to average over.

Additionally, some suggestions are made to mitigate the curse of dimensionality that may arise when automatically generating a large set of complex candidate models with many

parameters to be estimated, as well as to improve computation time. None of these recommendations substantially change the performance of the resulting estimator.

In order for the number of candidate models (M) to be tied to the sample size and assumed smoothness class, use an information criterion, such as AIC or BIC, to automatically reduce the number of candidate models if the number of candidate models exceeds some preset maximum number of candidate models (M_{max}). This is essentially a model screening step. The suggested value for the maximum number of candidate models is $M_{\text{max}} = 2500$. This retains degrees of freedom and reduces computation time. This should not significantly affect the performance of the resulting model average estimator. Future work is needed to tie this maximum number of candidate models to the sample size and assumed smoothness class.

Similarly, it is recommended to restrict the maximum dimension for each candidate model in order to tie the number of parameters to the sample size and assumed smoothness class. The suggested value for the maximum dimension is $p_{\text{max}} = 5000$. Future work should specify the maximum dimension in such a way that it is tied to the sample size and assumed smoothness class.

The nonparametric kernel-weighted spline regression produces a rich set of candidate models – or "bases" – that are flexible enough to cover a wide range of DGPs. The researcher can use an information criterion, such as AIC or BIC, to automatically select the basis function type for each candidate model. Three basis function types are recommended:

1. Tensor product basis: the most flexible of the three options and the one used by Racine et al. (2018). However, this may limit degrees of freedom quickly due to the large number of parameters to be estimated as the dimension of the model increases.
2. Generalized Taylor polynomial (e.g. with two predictors, $y = f(x_1, x_2) = f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b) + \frac{1}{2!}\Big[f_{xx}(a, b)(x - a)^2 + 2f_{xy}(a, b)(x - a)(y - b) + f_{yy}(y - b)^2\Big] + ...$).
3. Additive basis: allows for non-linearities in the predictors, but imposes additivity be-

tween terms, making it the least flexible basis function type (e.g. with two predictors, $y = f(x_1) + f(x_2)$).

The bases are of the Bernstein polynomial class, as opposed to raw polynomials, and allow for differing degrees across multivariate predictors. When there are two or more continuous predictors, the generalized Taylor polynomial includes interaction terms up to the degree minus one. If the researcher has some preexisting knowledge about the underlying structure of the relationship between variables, the researcher can specify the desired basis function type for each candidate model; however, the recommended approach is to automatically select the basis function type based on the data. Additionally, this approach allows for all possible combinations of degrees, segments, knots, and bandwidth values of the explanatory variables to be attempted. If the researcher has some pre-existing knowledge about the underlying structure of the data, or wishes to restrict this option due to data limitations (for example, a small sample size), this can be changed so that it is restricted to combinations between only certain parameters or values in order to retain degrees of freedom.

To specify the tensor product polynomial splines, a sequence of interior knots can be included. When interior knots are included, the Bernstein polynomials become B-spline bases. The suggested increment in segments sequence is 2, the minimum number of segments is 1 (which is the number of knots minus 1; there always exist at least 2 knots, the endpoints), and the maximum number of segments is 3 by default (which is the number of knots plus 1).

If the set of explanatory variables includes categorical variables – which would be expected for many empirical applications – a kernel function is included in the construction of the set of candidate models, so that this method admits both categorical and continuous variables (Ma et al., 2015; Racine et al., 2018). These kernel weight functions have smoothing parameters or bandwidths (λ) associated with each categorical predictor. The largest value for the smoothing parameters can be specified, such that $0 \geq \lambda \leq 1$. Additionally, the maximum value for the smoothing parameter grid in each dimension can be specified.

The suggested value is $\max(2, \text{ceiling}(\log(n) - S\log(1+p)))$, where p is the number of categorical predictors and n is the number of observations. The suggested value for S is 2. Categorical predictors can enter a model either additively and linearly (in which case, only the intercept would be allowed to shift) or in a semi-parametric varying coefficient structure (in which case, all parameters would be allowed to shift). It is recommended to allow categorical variables to enter a model with a varying coefficient specification to allow for greater flexibility in the model specification.

It is recommended to specify a maximum value for the basis degree in each dimension. The suggested value is $\max(2, \text{ceiling}(\log(n) - S\log(1+k)))$, where k is the number of continuous predictors and n is the number of observations. The suggested value for S is 1. The minimum value for the basis degree in each dimension can also be specified, with the suggested value being 0. Additionally, the increment in degree sequence can be specified, with the suggested value being 2.

In frequentist model averaging (FMA), a number of criteria can be used to estimate the model average weights, including, but not limited to, AIC, BIC, the focused information criterion (FIC; Claeskens & Hjort, 2003), Mallows' model average criterion (MMA; Hansen, 2007), and the jackknife model average criterion (JMA; Hansen & Racine, 2012). It is recommended that the researcher specify a cutoff below which a model weight is essentially zero. The suggested value for this cutoff is $10^{-4} = 0.0001$. Typically, the sum of the model average model weights is restricted such that $\sum_{m=1}^{M} \omega_m = 1$, though this is not, strictly speaking, necessary.

When derivatives are required, one must specify the order thereof. For most applications, the order of the derivative is typically set to 1, but can be set to any value, if required. Since this method averages over models that are nonlinear in the predictors, the derivatives will be vectors, rather than constants, functions that potentially depend on the values of all predictors.

This approach frees the user from using either model assertion or selection methods

and thereby attenuates bias arising from model misspecification. Simulations reveal that this approach is competitive with some semi- and nonparametric approaches. Because it uses only least squares fits, it can be more computationally efficient than its nonparametric counterparts. The goal here is consistent estimations, hence the emphasis on basis function complexity, sample size, and number of predictors.

2.5 Monte Carlo Experiment

The approaches discussed in this chapter hold promise for being able to generate a set of candidate models based on the data that have the three key features previously mentioned (number of candidate models tied to the sample size and assumed smoothness class, number of parameters tied to the sample size and assumed smoothness class, and a broad set of bases to cover a wide range of potential DGPs that are simple enough to average over). While more work needs to be done to fulfill these criteria, the three approaches studied here – model screening, recursive partitioning, and model averaging over nonparametric models – are an improvement to the current practice of writing down a handful of parametric models in an ad hoc manner as the candidate models used in model averaging.

In this section, I conduct a Monte Carlo experiment to evaluate the relative performance of the approaches discussed in this chapter. The goal of this exercise is to determine whether one method – if any – performs better than the others. I compare model averaging over a set of parametric models (with no model screening), the nonparametric approach that automatically selects the basis function type for each basis used in model averaging (Racine et al., 2018), and the recursive-partitioning-based MARS algorithm (Friedman, 1991b). Following the heuristics described in Section 2.4.2, a model screening step is included in the nonparametric model average approach; however, the default value for the maximum number of candidate models is $M_{\max} = 2500$, which does not bind in this Monte Carlo experiment. Therefore, this Monte Carlo experiment does not include a model screening

step. See Section 2.5.2 for the results when a model screening step is included.

The data generating process (DGP) is set to be:

$$f(x) = 1 + \frac{x_1 + x_2 + x_1x_2 + x_1^2 + x_2^2 + x_3 + x_1x_3 + x_2x_3 + x_4 + x_1x_4 + x_2x_4}{\sigma_{f(x)}}, \quad (2.37)$$

where x_1, x_2, x_4 are continuously distributed as $N(0,1)$ and x_3 has discrete support, generated from the binomial distribution with $n = 3$ (number of trials) and $p = 1/2$ (probability of success). The response variable y is chosen to be $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{f(x)})$ and $c \in \{0.25, 0.50, 1.0, 2.0\}$ determines the signal-to-noise ratio (SNR). I take $n = 100$ random draws for each variable from their respective distributions.

R (version 4.0.2) is used throughout for ease of replicability. The following packages are used:

- `quadprog`, "Functions to Solve Quadratic Programming Problems" (version 1.5-8), contains functions to solve quadratic programming problems and is used to solve for the model average weights,
- `ma`, "Model Averaging" (version 1.0-8), contains functions to implement the nonparametric model average approach using a variety of multivariate bases, and
- `earth`, "Multivariate Adaptive Regression Splines" (version 5.3.0), contains functions to build a regression model using MARS (Friedman, 1991b).

The default heuristics described in Section 2.4.2 are used for the nonparametric model average approach. See Section 2.5.2 for the results when these heuristics are altered. Both the nonparametric model average approach and model averaging over parametric candidate models use Mallows' Model Average (MMA) criterion (Hansen, 2007) to select the model average weights. The MMA criterion is defined as follows:

$$C_n(\omega) = \omega' \hat{\mathbf{E}}' \hat{\mathbf{E}} \omega + 2\hat{\sigma}^2 K' \omega, \quad (2.38)$$

where $\hat{\mathbf{E}}$ is the $n \times M$ matrix with columns containing the residual vector from the mth candidate model, $\hat{\sigma}^2$ is the estimated variance from the largest dimensional model, and K is the $M \times 1$ vector of the number of parameters in each model. The MMA criterion is used to solve for the weight vector, $\hat{\omega} = \text{argmin}_w C_n(\omega)$, which can be solved using quadratic programming. Note that M may differ between approaches as well as from replication to replication when using the nonparametric model average approach.

The following heuristics are followed for the MARS algorithm, which were chosen based on the default parameter values specified by Friedman (1991b). For the forward iterative splitting procedure, the maximum degree of the interaction can be specified, with the default value being 1, which specifies an additive model with no interaction terms. The GCV penalty per knot has a default value of 3 when the maximum degree of the interaction is greater than 1, and 2 otherwise. Simulation studies suggest a range of between 2 and 4. The maximum number of model terms (including the intercept) before the backwards stepwise deletion strategy is employed is calculated from the number of predictors and the maximum degree of interactions permitted. The default value for the forward stepping threshold is 0.001. This determines the stopping rule: the forward iterative splitting procedure ends when the addition of a term changes the R^2 value by less than 0.001.

The default value for the maximum number of terms (including the intercept) included in the backwards stepwise deletion strategy is all terms created by the forward procedure. Note that this is different from the final terms in the model after backwards elimination. This can be adjusted to enforce an upper bound on the final model size.

There are $M = 6$ parametric candidate models, which do not include the true DGP and

are estimated by ordinary least squares (OLS). They are:

$$\text{Model 1: } y = x_1 + \epsilon_1 \tag{2.39}$$

$$\text{Model 2: } y = x_1 + x_2 + \epsilon_2 \tag{2.40}$$

$$\text{Model 3: } y = x_1 + x_3 + \epsilon_3 \tag{2.41}$$

$$\text{Model 4: } y = x_1 + x_4 + \epsilon_4 \tag{2.42}$$

$$\text{Model 5: } y = x_1 + x_2 + x_3 + \epsilon_5 \tag{2.43}$$

$$\text{Model 6: } y = x_1 + x_2 + x_3 + x_4 + \epsilon_6 \tag{2.44}$$

This exercise is repeated 1000 times. The average mean squared error (MSE) over 1000 replications is used as the performance metric to compare these approaches. The MSE is computed as follows:

$$\text{MSE } = \frac{1}{n}\sum_{i=1}^{n}\left(\text{DGP}_i - \hat{y_i}\right)^2, i = 1, ...n, \tag{2.45}$$

where DGP_i is the DGP (defined above) and $\hat{y_i}$ are the fitted values from each method.

2.5.1 Results

Figure 2.1 shows box-and-whisker plots of the MSE for each method across variations in the SNR over 1,000 Monte Carlo replications (Tukey, 1970).[2] Recall that the response variable y is chosen to be $y = f(x) + \epsilon$ where $f(x) = 1 + (x_1 + x_2 + x_1x_2 + x_1^2 + x_2^2 + x_3 + x_1x_3 + x_2x_3 + x_4 + x_1x_4 + x_2x_4)/\sigma_{f(x)}$, $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{f(x)})$ and $c \in \{0.25, 0.50, 1.0, 2.0\}$ determines the SNR. From the top left plot to the bottom right plot, the SNR decreases. The plots show that as the SNR decreases, the MSE of each method increases, as one would expect since adding more noise increases MSE ceteris paribus. It is clear from the plots that in all cases except for the bottom right plot, where the SNR is very low, the nonparametric

[2] A box-and-whisker plot is a nonparametric method of displaying data that offers a graphical overview of the data by summarizing key features such as the median and upper and lower quartiles.

model average method performs the best in terms of MSE relative to model averaging over parametric models and the MARS algorithm. The MARS algorithm performs second best, while model averaging over ad hoc parametric models performs the worst in almost every case. This suggests that the nonparametric model averaging method discussed in this chapter may perform the best relative to other methods, unless the SNR is low, in which case it does not appear to offer any significant gains in MSE over model averaging over parametric models nor the MARS algorithm.

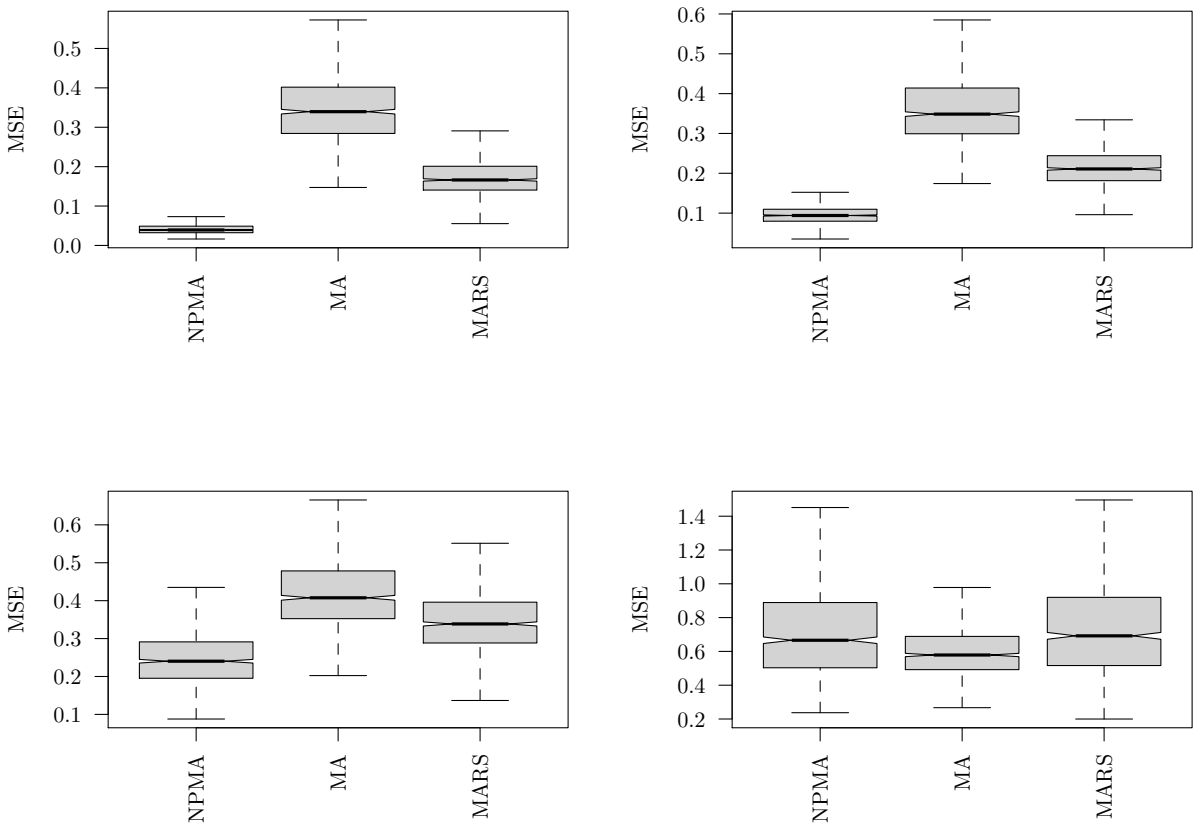

Figure 2.1: Box-and-whisker plot of MSE of each approach over 1,000 Monte Carlo replications across variations in the signal-to-noise ratio (SNR). The approaches are nonparametric model averaging (NPMA), model averaging over parametric models (MA), and multivariate adaptive regression spline (MARS). The response variable y is chosen to be $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma_\epsilon = c\sigma_{f(x)})$ and $c \in \{0.25, 0.50, 1.0, 2.0\}$ determines the signal-to-noise ratio. (Top left: 0.25σ. Top right: 0.50σ. Bottom left: 1.0σ. Bottom right: 2.0σ.)

Table 2.1 shows the mean MSE over 1,000 replications for each method across varia-

tions in the SNR. Overall, the nonparametric model average method performs the best in terms of mean MSE compared to the other two methods, except when the SNR is very low (fourth column). For example, when the SNR is very high (column 1), the nonparametric model average method performs extremely well, with a mean MSE of 0.0423248, compared to the MARS algorithm (0.1735796) and model averaging over parametric models (0.3503931). The MARS algorithm performs second best in every case except when the SNR is very low. When the SNR is very low, model averaging over parametric models outperforms the nonparametric model averaging method as well as the MARS algorithm with a mean MSE of 0.6021255; however, given its extremely poor performance relative to the other methods in other cases, this method is not recommended. From this Monte Carlo

Table 2.1: Mean MSE for each approach over 1,000 Monte Carlo replications (no model screening).

	0.25σ	0.50σ	1.0σ	2.0σ
NPMA	0.0423	0.0963	0.2520	0.7334
MA	0.3504	0.3631	0.4216	0.6021
MARS	0.1736	0.2155	0.3471	0.7421

experiment, the one method that performs better than the others in terms of MSE in almost every case is the nonparametric model averaging approach that automatically selects the basis function type for each basis. As previously discussed, this approach improves upon the standard practice of selecting a set of candidate models in an ad hoc manner by taking this arbitrary decision-making out of the hands of the researcher and, instead, using a data-driven method to select the candidate models (or bases). This Monte Carlo experiment shows that even compared to other promising methods, such as recursive-partitioning-based approaches like the MARS algorithm, the nonparametric approach to model averaging has the best performance overall.

2.5.2 Results with Model Screening

The previous exercise compared the performance of model averaging over parametric models, the nonparametric approach that automatically selects the basis function type for each basis used in model averaging (Racine et al., 2018), and the MARS algorithm (Friedman, 1991b) with no model screening. I repeat this Monte Carlo experiment but, this time, I include a model screening step prior to the nonparametric model average approach to see if and how performance in terms of MSE and mean MSE changes. Model screening is not applied to the set of parametric models, as the set is already very small. Model screening is also not applied to the MARS algorithm, as it has its own model screening step through the use of a backwards elimination procedure.

The suggested value for the maximum number of candidate models for the nonparametric model average approach is $M_{\text{max}} = 2500$. This value is not binding for this Monte Carlo experiment. As such, I change the maximum number of candidate models to be $M_{\text{max}} = 100$. This maximum is binding, as the number of candidate models generated in this Monte Carlo experiment typically exceeds 100.

Figure 2.2 shows a box-and-whisker plot of the MSE for each method in the Monte Carlo experiment from Section 2.5 as well as the nonparametric model average approach with model screening across variations in the SNR over 1,000 Monte Carlo replications.

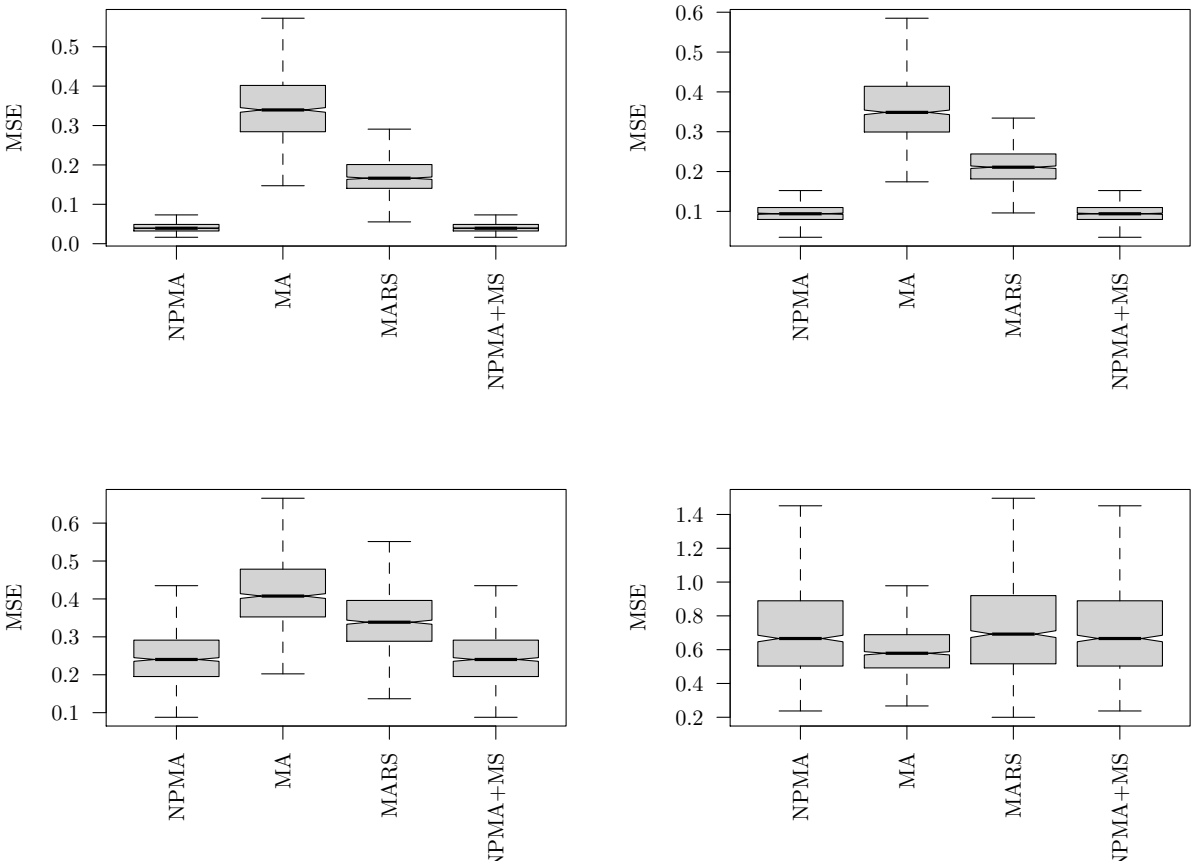

Figure 2.2: Box-and-whisker plot of MSE of each method, including NPMA with model screening (NPMA+MS), over 1,000 Monte Carlo replications across variations in SNR. (Top left: 0.25σ. Top right: 0.50σ. Bottom left: 1.0σ. Bottom right: 2.0σ.)

Table 2.2 shows the mean MSE over 1,000 Monte Carlo replications. The MSE of the nonparametric model average approach with model screening is slightly higher than that of the nonparametric model average approach without model screening. However, the difference is not substantial. Therefore, based on the results of this experiment, model screening appears to be a tool for reducing computation time rather than one to improve performance. It is important to note that this simulation had very few variables. Model screening would likely demonstrate its value in cases where there are many possible candidate models or many possible variables by significantly reducing computation time without significantly deteriorating performance.

Table 2.2: Mean MSE over 1,000 Monte Carlo replications (including model screening).

	0.25σ	0.50σ	1.0σ	2.0σ
NPMA	0.0423	0.0963	0.2520	0.7334
MA	0.3504	0.3631	0.4216	0.6021
MARS	0.1736	0.2155	0.3471	0.7421
NPMA with model screening	0.0423	0.0963	0.2520	0.7334

2.6 Conclusion

This chapter reviewed three promising approaches for building an improved set of candidate models for model averaging. Selecting the candidate models is an important, but often overlooked, step, as the resulting model average estimator will inherit its properties from these candidate models. The standard practice of simply writing down a handful of parametric candidate models is not sufficient in building a rich set of candidate models, as it relies on ad hoc decisions on the part of the researcher. The three ideal features of a set of candidate models were identified to be:

1. number of candidate models tied to the sample size and assumed smoothness class,
2. number of parameters tied to the sample size and assumed smoothness class, and
3. a broad set of functions, or "bases", to cover a wide range of potential DGPs that are simple enough to average over.

I studied model screening, recursive partitioning-based algorithms such as multivariate adaptive regression splines (Friedman, 1991b), and a nonparametric approach to model averaging that automatically selects the basis function type based on the data (Racine et al., 2018). None of these approaches alone can build a set of candidate models that satisfies all the criteria listed above; however, these approaches are an improvement to model averaging over a set of arbitrarily chosen parametric candidate models. The Monte Carlo experiment demonstrates that the nonparametric model averaging approach performs the best in terms of MSE compared to model averaging over parametric models and the MARS algorithm in

almost every case. Model screening can be added as a means of decreasing computation time without deteriorating performance. With more work, these approaches can be adapted to produce a set of candidate models that fulfill the desirable criteria for a set of candidate models to be used in model averaging.

Chapter 3

Model Averaging and Machine Learning Analysis of Employment Among Parents in Canada during the COVID-19 Pandemic

3.1 Introduction

It is crucial to have robust estimates and predictions, especially when these results influence policy. Economists almost universally report uncertainty in parameter estimates by reporting, for example, confidence intervals or standard errors. However, it is unusual for empirical economists to acknowledge uncertainty in the selected model. Sometimes, empiricists will report the results from more than one model, but then it is unclear which results should be used for reports and policy making, especially when results across models differ. At most, researchers will sometimes conduct a model misspecification test to acknowledge model uncertainty, i.e., the probability that one's model is incorrectly specified. But when a model is rejected by the data, it is unclear how to proceed.

In this chapter, I use robust statistical methods such as model selection, model averaging, and machine learning algorithms to assess the uncertainty inherent in model choice in an applied setting. This chapter applies the methods discussed in depth in chapter 1 to currently timely microdata. I use model selection, model averaging, and the lasso along with data from the Canadian Labour Force Survey to determine which model or combination of models is best for assessing the impacts of the COVID-19 pandemic on the employment of parents with young children in Canada during the first six months of the COVID-19 pandemic. Model selection methods acknowledge the uncertainty in the model chosen by using a criterion to select the "best" or least misspecified model among a finite set of candidate models. Model averaging acknowledges model uncertainty by constructing a weighted average over a set of candidate models. Model averaging has been shown to produce more robust results while requiring fewer assumptions than standard econometric approaches that use parametric models (Hansen, 2007; Hoeting et al., 1999). Additionally, model averaging cannot be expected to do any worse than any one model in the set of candidate models in the presence of model uncertainty. The lasso is itself a model selection exercise that takes a large, unrestricted model and performs selection on variables to select a final model.

I find that model selection and model averaging converge to select one model from the finite set of parametric models considered in this analysis. The largest and second largest models are chosen using different model selection criteria and across different subsamples with different dependent variables. The largest model is assigned a weight of 1 by model averaging across all subsamples and with different dependent variables. I compare each model and method using correct classification rates (CCR) and receiver operating characteristic (ROC) curves. I find that the models selected by model selection and model averaging as well as the lasso model perform better in terms of classification compared to the simpler parametric model specifications, which suggests that empirical researchers should consider statistical methods for the choice of model rather than relying on ad hoc decision

making. Additionally, I find that the choice of model matters. I estimate the marginal effect of sex on the probability of being employed over six months and find that the results differ in magnitude across models in an economically important way, as these results could affect policies for post-pandemic recovery.

This chapter proceeds as follows. Section 3.2 describes the data used. Section 3.3 gives an overview of the impact of the COVID-19 pandemic at the time of writing this chapter. Section 3.4 describes the methods used. Section 3.5 presents and discusses the results from using each method. Section 3.6 concludes.

3.2 Data

I use data from Statistics Canada's Labour Force Survey (LFS) public use microdata files. The LFS is a monthly survey with a cross-sectional design, sampling approximately 54,000 Canadian households every month. It is nationally representative data on the Canadian working-age population when weighted. Responses are recorded for each month with the outcomes for a single week, typically the 15th of the month. I use data from February 2020 to August 2020 which covers the period of school closures for the 2019-2020 academic school year caused by the COVID-19 pandemic. The sample is weighted using the survey weights included in the LFS public use microdata. I restrict my sample to adults aged 20-64 years with a strong attachment to the labour force (that is, individuals that were either currently employed or employed within the last year at the time of the survey) and with a youngest child aged 0-12 years, to focus on parents with the greatest childcare responsibilities. This brings my sample size to 95436. I then split the sample by the age of the youngest child, so that I have two subsamples. The first subsample includes parents whose youngest child is under 6 years (preschool-aged) and the second subsample includes parents whose youngest child is 6-12 years (school-aged). Evaluating these two subsamples will provide insights regarding whether parents with school-aged or preschool-aged chil-

dren are more heavily impacted by the pandemic, or if both were impacted equally. The subsample sizes are 51244 for parents of preschool-aged children and 44192 for parents of school-aged children.

Table 3.1 displays summary statistics for the dependent and explanatory variables used in the candidate model set for model selection and model averaging, as well as in the lasso regression. My main dependent variable is an indicator of employment (1 = employed, 0 = otherwise), which includes individuals who are employed and at work *or* employed and absent from work. Because employment is likely over-stated during the early months of the COVID-19 pandemic due to individuals retaining their employment status but being absent from work or reporting reduced pay or hours of work, I consider an alternative measure of employment for my dependent variable: an indicator for being employed *and* at work (1 = employed and at work, 0 = otherwise), which excludes individuals who are absent from work. This measure can better capture the changes in the labour market activity in Canada that would be masked by the traditional measure of employment, which is important because individuals who are employed but absent from work are at greater risk of being separated from their employer.

The explanatory variables included in my analysis are:

- survey month of LFS,
- sex of respondent,
- highest education level of respondent,
- age group of respondent,
- marital status of respondent,
- immigration status of respondent,
- province of residence of respondent,
- occupation of main job (40 categories),
- industry of main job (21 categories),
- category of main job,

- full-time or part-time status of main job[1],
- job tenure with current employer (in months)[2], and
- type of economic family (8 categories).

The average job tenure with the current employer is 74 months for parents with a youngest child under 6 years and 101 months for parents with a youngest child 6-12 years. The proportions of individuals in the sample in various occupations and industries can be found in the Appendix (Section 3.7).

A limitation of the LFS is that there are no data on important demographic characteristics that would likely affect employment, such as race or total number of children in the household, nor data on important market variables, such as vacancies. Additionally, there are no data for the territories, thus any results from the LFS cannot be extrapolated to the populations of the Northwest Territories, Nunavut, or Yukon.

Table 3.1: Summary statistics for Labour Force Survey subsamples of individuals aged 20-64 years, currently employed or employed within the last year, and with a youngest child aged under 6 years (preschool subsample) or 6-12 years (school subsample).

Description	Preschool (Percent)	School (Percent)
Employment		
Not employed	10.69	10.09
Employed	89.31	89.91
Employed and at work		
Not employed or employed but absent from work	29.76	23.46

(continued . . .)

[1]For individuals who were not currently employed but were employed within the past year, full- or part-time status of current employment, which had missing observations, was replaced with full- or part-time status of previous employment.

[2]For individuals who were not currently employed but were employed within the past year, tenure with current employer, which had missing observations, was replaced with tenure with previous employer.

Table 3.1: Summary statistics for Labour Force Survey subsamples.

Description	Preschool (Percent)	School (Percent)
Employed and at work	70.24	76.54
Sex		
Male	52.07	48.25
Female	47.93	51.75
Education		
0-8 years	0.97	0.77
Some high school	3.57	3.73
High school	13.47	14.52
Some postsecondary	3.84	3.75
Postsecondary	40.43	41.27
Bachelor's	25.18	23.77
Above bachelor's	12.55	12.18
Age		
20-24	2.11	0.13
25-29	12.34	1.43
30-34	30.64	7.23
35-39	33.20	23.70
40-44	16.24	34.05
45-49	4.13	23.17
50-54	0.95	7.89
55-59	0.39	1.85
60-64	0.01	0.55
Marital status		

(continued . . .)

Table 3.1: Summary statistics for Labour Force Survey subsamples.

Description	Preschool (Percent)	School (Percent)
Married	68.84	70.33
Common-law	25.23	16.23
Widowed	0.09	0.35
Separated	1.49	3.96
Divorced	0.47	3.07
Single	3.89	6.06
Immigration status		
Immigrant, <10 years	11.76	7.37
Immigrant, >10 years	9.16	14.03
Non-immigrant	79.07	78.60
Province		
NL	2.47	2.72
PE	2.77	2.84
NS	4.03	4.13
NB	3.79	4.99
QC	19.89	18.62
ON	26.82	28.05
MB	9.42	8.59
SK	7.97	7.72
AB	12.32	11.33
BC	10.53	11.00
Job category		
Public	27.63	29.40

(continued . . .)

Table 3.1: Summary statistics for Labour Force Survey subsamples.

Description	Preschool (Percent)	School (Percent)
Private	59.02	55.30
Self-employed	13.30	15.25
Unpaid family worker	0.04	0.05
Full- or part-time status		
Full-time	86.84	86.52
Part-time	13.16	13.48
Type of economic family		
Dual-earner couple, youngest child 0-17 years	68.09	65.53
Single-earner couple, male employed, youngest child 0-17 years	16.58	12.77
Single-earner couple, female employed, youngest child 0-17 years	6.26	5.84
Non-earner couple, youngest child 0-17 years	2.76	1.86
Lone-parent family, parent employed, youngest child 0-17 years	5.15	11.97
Lone-parent family, parent not employed, youngest child 0-17 years	1.15	2.03
Other families	0.01	0.01

3.3 Initial Impact of the COVID-19 Pandemic on Canadians

The COVID-19 pandemic caused unprecedented declines in employment and aggregate hours worked across Canada. Between February 2020 and April 2020, the employment rate dropped sharply by 15 percent, after adjusting for usual changes in employment during those months, which largely came from a decrease in employment in jobs that are not amenable to working from home and jobs in non-essential industries (Lemieux, Milligan, Schirle, & Skuterud, 2020). This resulted in both an increase in the unemployment rate and a decrease in the labour force participation rate. Jones, Lange, Ridell, & Warman (2020) estimate that approximately 45 percent of job losers transitioned to unemployment while 55 percent of job losers exited the labour force entirely. The unemployment rate nearly doubled in April 2020 compared to the pre-pandemic early-2020 rate due to an increase in temporary layoffs, and then declined slightly in May 2020. Most unemployed individuals were waiting to be recalled to former jobs and, consequently, not searching for work. However, search unemployment has been increasing, likely due to the decline in labour demand making it even more difficult to find employment.

The decline in employment understates the decline in actual work performed in April and May 2020, as approximately 8 to 9 percent of the population reported full-week absences from work during that time period. Of those individuals who were employed but absent from work for a full week, approximately half were not being paid. By May 2020, paid absences had returned to pre-pandemic levels; however, unpaid absences continued to be unusually high. Individuals who were not being paid nor being productive but still reported an attachment to their employer may be vulnerable to separation from their employer, which may be contributing to the rise in search unemployment. Additionally, data on vacancies, provided by Employment and Social Development Canada, show that there was a 50 percent drop in labour demand from March to April 2020, with a small recovery

in May and June. The increase in search unemployment and decrease in labour demand makes it more difficult for individuals to find employment. Thus, standard measures of unemployment or employment are not useful for evaluating the changes in work during the COVID-19 pandemic because many workers were either laid off but waiting to be recalled, transitioned from employment to non-participation while waiting to be recalled, or retained as an employee but absent from work, marked by a substantial decline in hours (Jones et al., 2020). It is for this reason that I use two different dependent variables in my analysis: employment (including individuals both absent and not absent from work) and being employed and not absent from work.

Lemieux et al. (2020) use the LFS to measure changes in aggregate weekly work hours instead of using traditional measures such as labour force status. They found that from February 2020 to April 2020, after adjusting for typical changes during these months, there was a 32 percent decline in aggregate weekly work hours for individuals aged 20-64. This was a huge, unprecedented loss, even compared to previous recessions such as the Great Recession. This change in aggregate hours included both job losses on the extensive margin as well as declines in hours worked on the intensive margin. The most affected were workers in public-facing jobs in industries such as accommodation and food services, as well as workers aged 20-29, hourly workers, non-unionized workers, and women.

Another major impact of the pandemic was on the availability of childcare, which could impact the labour force participation and productivity of parents. In Canada, provincial governments announced the closures of schools, day care centres, and childcare centres in response to the COVID-19 pandemic, beginning with Ontario making an announcement on March 12, 2020 and other provinces following suit, with British Columbia being the last to announce the closures of public schools on March 17, 2020. While many provinces announced – and anticipated – that closures would last for only two weeks following spring break in March, every province extended the closures for many more weeks, and many for the remainder of the school year. Lemieux et al. (2020) find that women with children

aged younger than 12 were hit harder by the pandemic than women with children aged 13-17. Additionally, both hours and employment of mothers with preschool-aged and younger children dropped substantially between February 2020 and April 2020. Given that many schools and childcare centres remained closed until the end of the academic year, the lack of regular childcare and schooling likely made it harder for families – and particularly mothers, who typically bear the responsibility of childcare in dual-earning, heterosexual couples – to supply labour. This will be explored using model selection, model averaging, and machine learning in section 3.5.

Qian & Fuller (2020) estimate the gender employment gap among parents of young children in Canada at the beginning of the pandemic using data from the Labour Force Survey (LFS). Their sample consists of adults aged 25-54 years with children aged 12 years or younger and attached to the labour market (individuals currently employed or who have been employed in the past year). Using a weighted logistic regression, Qian & Fuller (2020) find that the gender employment gap widened between February 2020 and May 2020. The gender gap in employment widened more for parents with school-aged children (6-12 years) than for parents of preschool-aged children (under 6 years). The gap also widened substantially more for less educated parents (high school diploma or less) than for more educated parents (university graduates). This suggests that the COVID-19 pandemic exacerbates existing inequalities in employment. As the economy continues to recover from the pandemic, without policies to address childcare needs such as flexible leave policies, mothers will be left behind. The subsequent effects on the work experience and human capital of women with young children, if left behind, can have severe long-term impacts on their careers and future earnings. This would also cause the gender wage gap and the gender employment gap to continue to widen, disrupting a trend of narrowing gender gaps over the previous decades.

3.4 Methods

3.4.1 Model Selection

Model selection methods use a criterion to select the single "best" or least misspecified model among a finite set of candidate models. The selected model is, at best, an approximation to the data generating process (DGP) and is an improvement over the practice of selecting a model in an ad hoc manner. I use two model selection criteria in my analysis: the Akaike information criterion (AIC; Akaike, 1970) and the Bayesian information criterion (BIC; Schwarz, 1978). AIC is defined as:

$$\text{AIC} = -2\ln(\hat{L}) + 2k, \tag{3.1}$$

where $\ln(\hat{L})$ is the maximum value of the log-likelihood function of a model with k being the number of estimated parameters in the model. AIC balances goodness of fit (as measured by the log-likelihood) and parsimony (as measured by the penalty for the number of parameters included in the model). A low AIC value is desirable.

BIC is defined as:

$$\text{BIC} = -2\ln(\hat{L}) + \ln(n)k. \tag{3.2}$$

A low BIC value is desirable. When the sample size is large, the penalty in BIC is larger and so BIC tends to select smaller models in these cases relative to AIC.

Table 3.2 describes the set of candidate models used in this analysis. AIC and BIC are computed for each of the 5 models and the model with the lowest value of AIC and BIC is chosen.

3.4.2 Model Averaging

Model averaging is useful in situations where more than one model is supported by economic theory, yet it is unclear which is the "best" – in other words, least misspecified – model among the set of models under consideration. Model uncertainty – that is, the probability that one's model is incorrectly specified – can have unintended consequences, such as inference that is overly optimistic or, at worst, completely invalid. Model averaging is the leading approach for handling the issue of model uncertainty. Frequentist model averaging, henceforth referred to simply as model averaging, constructs a combined estimator that is a weighted average of estimators from a set of candidate models. Some advantages of model averaging include better predictive ability than using any single model among the set of candidate models (Hoeting et al., 1999), more robust results compared to any single model among the set of candidate models (Moral-Benito, 2015), broad applicability, fewer assumptions compared to conventional econometric methods, and standard errors that account for the bias that arises from model uncertainty (Tobias & Li, 2004). Some limitations of model averaging include increased computational burden, although this is less of a concern with the advance of technology; lack of closed-form solutions for some estimators; and lack of precedent for post-model-average inference.

I use Mallows' Model Average Criterion (MMA) (Hansen, 2007) in this chapter. The MMA criterion is defined as follows:

$$C_n(\omega) = \omega' \hat{\mathbf{E}}' \hat{\mathbf{E}} \omega + 2\hat{\sigma}^2 K' \omega, \tag{3.3}$$

where ω is the M-dimensional vector of model weights, $\hat{\mathbf{E}}$ is the $n \times M$ matrix with columns containing the residual vector from the mth candidate model, $\hat{\sigma}^2$ is the estimated variance from the largest dimensional model, and K is the $M \times 1$ vector of the number of parameters in each model. Typically, more than one model is assigned some non-zero model weight, and some models may be assigned a weight of zero. If the "true" model – that

is, a model that accurately represents the underlying, unknown data generating process – lies within the set of candidate models, that model would be assigned a weight of 1. The model average estimator is constructed using the estimated model weights obtained from the MMA criterion, $\hat{\omega}_m$, and the estimates from each of the candidate models. For example, the model average estimator of the regression coefficient (in this case, a scalar and assumed to be common to all models), $\hat{\beta}_{j,MA}$, is:

$$\hat{\beta}_{j,MA} = \sum_{m=1}^{M} \omega_m \hat{\beta}_{j,m}, \tag{3.4}$$

where $j = 1, ..., q$ indexes the regression coefficient, $m = 1, ..., M$ indexes the candidate model, $\hat{\beta}_{j,m}$ represent the coefficients from each candidate model, $0 \leq \omega_m \leq 1$ and $\sum_{m=1}^{M} \omega_m = 1$.

In this chapter, I follow the common practice of writing down a set of parametric models with a common variable of interest. The candidate model set is built from the following general logistic regression:

$$\Pr(\mathrm{E} = 1|\mathbf{X}, \mathbf{Z}) = F(\mathbf{X}\beta, \mathbf{Z}\alpha), \tag{3.5}$$

where $\Pr(\mathrm{E} = 1|\mathbf{X}, \mathbf{Z})$ represents the conditional probability of being employed (i.e. employment = 1) or, using the alternative dependent variable, the conditional probability of being employed and at work (i.e. employed and at work = 1); F is the cumulative distribution function (CDF) of the logistic distribution[3]; $\mathbf{Z}$ is a matrix of explanatory variables that are included in every model, which are sex, dummy variables for survey month (where the number of dummy variables is the number of months minus 1), and interactions between these variables; and $\mathbf{X}$ is a matrix of optional explanatory variables, which includes education, age, marital status, immigration status, province of residence, type of economic family, occupation, industry, category, full- or part-time status, and tenure of main job.

[3]The CDF of the logistic distribution is: $F(\mathbf{X}\beta) = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}}$.

Table 3.2: List of candidate models. Note that every candidate model includes the variables sex, dummy variables for survey month, and interactions between these variables.

	Optional RHS Variables
Model 1	NA
Model 2	age, education, marital status, immigration status, province, economic family
Model 3	occupation, industry, category, full- or part-time status, tenure
Model 4	age, education, marital status, immigration status, province, economic family, occupation, industry, category, full- or part-time status, tenure
Model 5	age, education, marital status, immigration status, province, economic family, occupation, industry, category, full- or part-time status, tenure, tenure squared

Note that the number of dummy variables for each categorical variable is the number of categories for that variable minus 1. The inclusion or exclusion of the elements of $\mathbf{X}$ will differentiate one candidate model from another. Table 3.2 shows the optional control variables that are included in each candidate model. The first model includes only those variables that are included in every model: sex, dummy variables for survey month, and interactions between those variables. Then, in addition, I include so-called demographic variables: age, education, marital status, immigration status, province of residence, and type of economic family. Then, instead of demographic variables, I include job-specific variables: occupation, industry, category, full- or part-time status, and job tenure of the main job of the respondent. Next, I include both groups of control variables. Finally, I include tenure squared as an additional control variable. This creates five candidate models to be used in model averaging. Each candidate model is estimated using the two subsamples described in Section 3.2: the subsample of parents whose youngest child is preschool-aged and the subsample of parents whose youngest child is school-aged.

This set of candidate models was selected in an ad hoc manner. While there exist statistical methods to generate the set of candidate models automatically based on the data, model averaging should not perform any worse – and often, better – than any one model in

the set of candidate models, regardless of how that set of models was chosen. An alternative method for selecting the set of candidate models is one that is nonparametric in nature and uses kernel-weighted regression splines (Racine, 2019) and is explored in Chapter 2.

3.4.3 Lasso

Machine learning algorithms are becoming more widely used in economics, despite the fact that they are often used without a deep technical understanding of how they work, leading people to refer to these algorithms as "black boxes". In this chapter, I use the least absolute shrinkage and selection operator, or "lasso" (Tibshirani, 1996). The lasso performs selection on regressors by shrinking some coefficients towards zero while setting others equal to zero. The lasso estimator is defined as:

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \Big(y_i - \sum_j \beta_j x_{ij}\Big)^2 \text{ subject to } \sum_j |\beta_j| \leq t, \tag{3.6}$$

where $i = 1, ..., n$, j indexes the regressor, β represents the regression coefficients, and $t \geq 0$ represents the tuning parameter. Selection of the tuning parameter, t, is important as it controls regressor selection as well as how much shrinkage is applied to the coefficients. Cross-validation is commonly used to select the tuning parameter and is used in this chapter. Some advantages of the lasso are that it produces highly interpretable models, increased stability, and improved prediction accuracy with relative computational efficiency. The lasso and other machine learning algorithms are explored in depth in Chapter 1.

For my analysis, the model that is fed to the lasso algorithm is one that is complex compared to the set of candidate models used in model selection and model averaging. It includes sex, dummy variables for survey month, and interactions between these variables, as well as all possible interactions between the explanatory variables included in the set of candidate models used with model selection and model averaging: age, education, marital status, immigration status, province, type of economic family, occupation, indus-

try, category of main job , full- or part-time status, and tenure. Thus, the initial regressor set includes age; an interaction between age and education; an interaction between age, education, and marital status; an interaction between age, education, marital status, and immigration status; etc. When this large, complex model was estimated simply using the logistic regression, the model did not converge, as almost every variable is a factor, which – along with interactions – created a large number of explanatory variables. Thus, this large, complex model was used with the lasso to perform selection on variables, while maintaining a model with potentially greater complexity than those used in model selection and model averaging.

3.5 Results

R (version 4.0.2) is used throughout for ease of replicability. The following packages are used:

- `quadprog`, "Functions to Solve Quadratic Programming Problems" (version 1.5-8), contains functions to solve quadratic programming problems and is used to solve for the model average weights,
- `glmnet`, "Lasso and Elastic-Net Regularized Generalized Linear Models" (version 4.1-2), is used for the lasso, and
- `pROC`, "Display and Analyze ROC Curves" (version 1.16.2), is used to build the ROC curves.

Table 3.3 shows the results from model selection using AIC. These results are consistent across dependent variables and subsamples. The largest model (model 5) is selected in every case, as it has the minimum AIC value across all models. Table 3.4 shows the results from model selection using BIC. These results differ from those using AIC as the model selection criterion. The model that has the minimum BIC value for the subsample

of parents with preschool-aged children using employment as the dependent variable is the second-largest model (model 4). However, for this subsample but using employed and at work as the dependent variable, it is the largest model (model 5) that has the minimum BIC value. For the subsample of parents with school-aged children, the results are reversed; using employment as the dependent variable, model 5 is selected and using employed and at work as the dependent variable, model 4 is selected. BIC tends to select more parsimonious models compared to AIC because it more heavily penalizes increasing number of parameters, which could explain this observed difference between model selection using AIC versus BIC. However, the difference in the BIC values of models 4 and 5 when model 4 is selected is very small. Given that one or two models were selected from a set of five, if a researcher chose one of the other models, they would be worse off than if they had used model selection to select a model from a model selection perspective. Table 3.5

Table 3.3: Model selection using Akaike's information criterion (AIC) for two different dependent variables across 5 candidate models. The minimum AIC for each column is in bold.

	Preschool subsample		School subsample	
	Employed	Employed and at work	Employed	Employed and at work
Model 1	34355	59182	28322	46807
Model 2	12419	51044	9553	38879
Model 3	30011	57098	24074	43525
Model 4	11127	50018	8217	36809
Model 5	**11115**	**49936**	**8182**	**36798**

reports the model average weights using the MMA criterion for the subsamples of parents with preschool-aged children (under 6 years) and school-aged children (6-12 years) for two different dependent variables (employed and employed and at work) across five candidate models. In all cases, the MMA criterion assigns a weight of 1 to the largest model (model 5), effectively resulting in a corner solution and collapsing model averaging to model selection. Given that model 5 was assigned a weight of 1, if a researcher chose one of the other 4 models among this finite set of parametric candidate models, they would do worse

Table 3.4: Model selection using Bayesian information criterion (BIC) for two different dependent variables across 5 candidate models. The minimum BIC for each column is in bold.

	Preschool subsample		School subsample	
	Employed	Employed and at work	Employed	Employed and at work
Model 1	34569	59394	28532	47017
Model 2	13038	51713	10156	39561
Model 3	31226	58323	25280	44717
Model 4	**12702**	51669	9733	**38452**
Model 5	12704	**51604**	**9713**	38455

than if they had used model averaging. While the final combined model average estimator corresponds to the largest model, this is not always the case. Model averaging performs the same or better than any one model in the set of candidate models under consideration based on the model average criterion employed. In this particular case, one model was preferred over all the others, but this does not mean that the model that was given a weight of 1 is the "true" model (although if the "true" model were in the candidate model set, it would be assigned a weight of 1). Thus, using model averaging yields results that are better or the same as any one of the individual candidate models under consideration.

Given model uncertainty, one would expect more than one model to be assigned some non-zero model weight. However, this exercise in using model averaging to construct a combined estimator over a set of candidate models is still expected to result in an improvement to selecting a model in an ad hoc manner, and there is a reasonable explanation for seeing this vector of model weights. The set of candidate models is limited in a number of ways. First, the set is finite and small at only five models. Including more models may change the results. Second, the set of models are exclusively parametric, and so they are limited in terms of functional form. Additionally, the set of candidate models are inflexible in terms of interactions among variables and order of polynomial terms. Including more interactions and higher order polynomials in the set of candidate models could change the weight vector. However, when the model that was initially fed to the lasso was, instead, es-

timated using the logistic regression, the model did not converge. Given that every variable except for tenure is a factor, including just a few interactions significantly increases the rank of each model. As such, the scope of this chapter is limited due to computational considerations arising from an ill-conditioned design matrix. Additionally, repeating this analysis using a nonparametric method for generating the set of candidate models, as explored in Chapter 2, might be warranted, but lies beyond the scope of this chapter.

Table 3.5: Model average weights using Mallows' model average (MMA) criterion for two different dependent variables across 5 candidate models. Weights are shown to the sixth digit.

	Preschool subsample		School subsample	
	Employed	Employed and at work	Employed	Employed and at work
Model 1	0	0	0	0
Model 2	0	0	0	0
Model 3	0	0	0	0
Model 4	0	0	0	0
Model 5	1	1	1	1

3.5.1 Relative Classification Performance

To compare the relative performance of each model and method, I use a number of different classification metrics and methods to assess model accuracy and overall model performance.

The *correct classification rate* (CCR) evaluates how well a model does at predicting the outcome for a binary variable (in this case, employment or employed and at work). It represents the overall classification accuracy for a model (Alboukadel, 2018). It is defined as:

$$\text{CCR} = (\text{number of true positives} + \text{number of true negatives})/n, \tag{3.7}$$

where the number of true positives is the number of observations that a model correctly

predicts as employed or employed and at work (depending on which dependent variable is used), i.e. is true; the number of true negatives is the number of observations that a model correctly predicts as not employed or not employed and at work, i.e. is false; and n is the sample size. The number of true positives (true negatives) is based on a comparison of the estimated $P(Y = 1|X = x)$ ($P(Y = 0|X = x)$ with a cutoff or threshold probability τ set by the researcher. Higher CCR values are more desirable. Table 3.6 shows the CCR for each model and method using a standard 0.5 cutoff probability. A cutoff probability of $\tau = 0.5$ means that any observation with a predicted probability of being employed or being employed and at work (depending on which dependent variable is used) greater than 0.5 is considered positive, i.e. employed, and below 0.5 is considered a negative, i.e. not employed. If we were concerned about incorrectly predicting the status for individuals who are truly not employed, we could adjust the cutoff to be higher (e.g. 0.8). Recall that model selection using BIC selected models 4 and 5, model selection using AIC selected model 5, and model averaging selected model 5. The CCR for the lasso is the largest, suggesting that the lasso performs the best in terms of CCR. The CCR of models 4 and 5 are similar in magnitude across dependent variables and subsamples and these models perform second best in terms of CCR after the lasso. Model 1, which only includes survey month and sex and interactions between those variables, has the smallest CCR across dependent variables and subsamples, and yet this ad hoc model has been used by empirical researchers to evaluate the impact of sex on labour force participation of parents in Canada using data identical to that used herein (Qian & Fuller, 2020)[4]. The relative performance of each method and model in terms of CCR suggests that using model selection, model averaging, or machine learning over a simple model chosen in an ad hoc manner is better in terms of classification accuracy. As mentioned above, a classification exercise relies on a cutoff probability τ, which is selected by the researcher, e.g. $\tau = 0.5$. The *receiver operating characteristic* (ROC) curve is a graphical summary of the overall performance of a model, including the

[4]Qian and Fuller use data from the Labour Force Survey public use microdata files for February to May 2020

Table 3.6: Correct classification rate (CCR) of 5 models and the lasso model across two different dependent variables and two subsamples. The maximum CCR for each column is in bold.

	Preschool subsample		School subsample	
	Employed	Employed and at work	Employed	Employed and at work
Model 1	0.8931	0.7024	0.8991	0.7654
Model 2	0.9294	0.768	0.9513	0.8164
Model 3	0.8938	0.7251	0.9007	0.7761
Model 4	0.9445	0.7731	0.9563	0.8234
Model 5	0.9448	0.773	0.9565	0.824
Lasso	**0.9733**	**0.804**	**0.9737**	**0.8517**

proportion of true positive rates and false positive rates at all possible cutoff probabilities $\tau \in (0, 1)$ (Fawcett, 2006). Sensitivity (on the y-axis) represents the true positive rate, which is the proportion of correctly identified positives – in this case, employed or employed and at work – among the population of individuals who are employed or employed and at work. Specificity (on the x-axis) represents the false negative rate, which is the proportion of incorrectly identified negatives – in this case, a negative is not employed or not employed and at work – among the population of individuals that are, in fact, employed or employed and at work. In general, a model is a good classifier when its ROC curve has a high true positive rate and/or a high false negative rate. Thus, an ROC curve above the 45 degree line is preferred, and a model is better than another if its ROC curve lies above and to the left of the ROC curve for the other model. An ROC that falls along the 45 degree line indicates that the model is no better at making predictions than random guessing.

Figure 3.1 shows the ROC curves for each model and method when employment is used as the dependent variable for the subsample of parents with preschool-aged children. Models 1 and 3 performed the worst based on the ranking of their ROC curves. The other models, including the models selected by model selection and model averaging as well as the lasso, all perform similarly, with the lasso having the highest ROC curve, followed second by model 5.

Figure 3.2 shows the ROC curves for each model and method when employed and at work is used as the dependent variable for the subsample of parents with preschool-aged children. Once again, models 1 and 3 have the worst overall performance based on their ROC curves. Models 2, 4 and 5 perform better and their ROC curves overlap with one another, suggesting that these models have roughly the same overall performance. The lasso performs the best overall, given that its ROC curve lies above those of the other models.

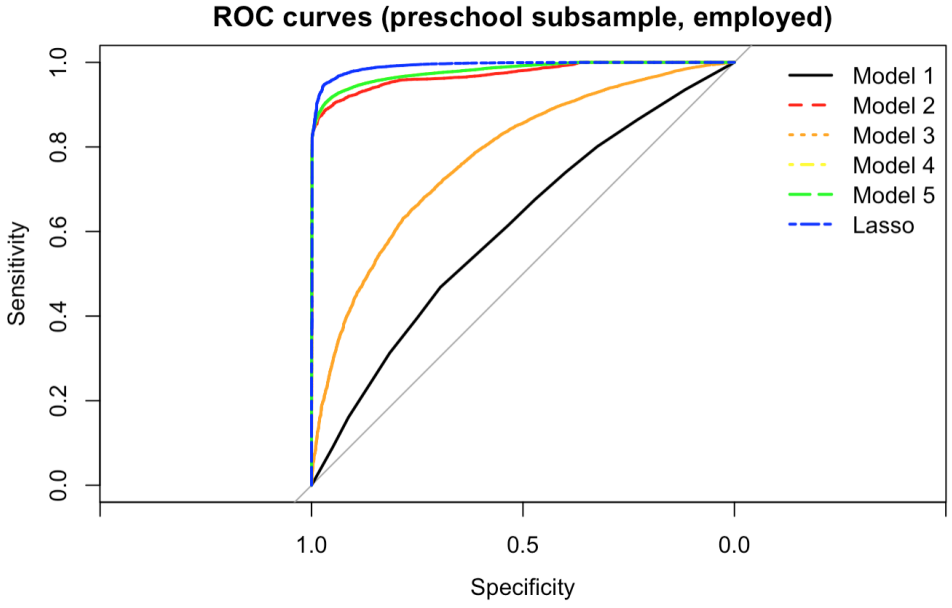

Figure 3.1: ROC curves for each method and model using employment as the dependent variable for the subsample of parents with preschool-aged children.

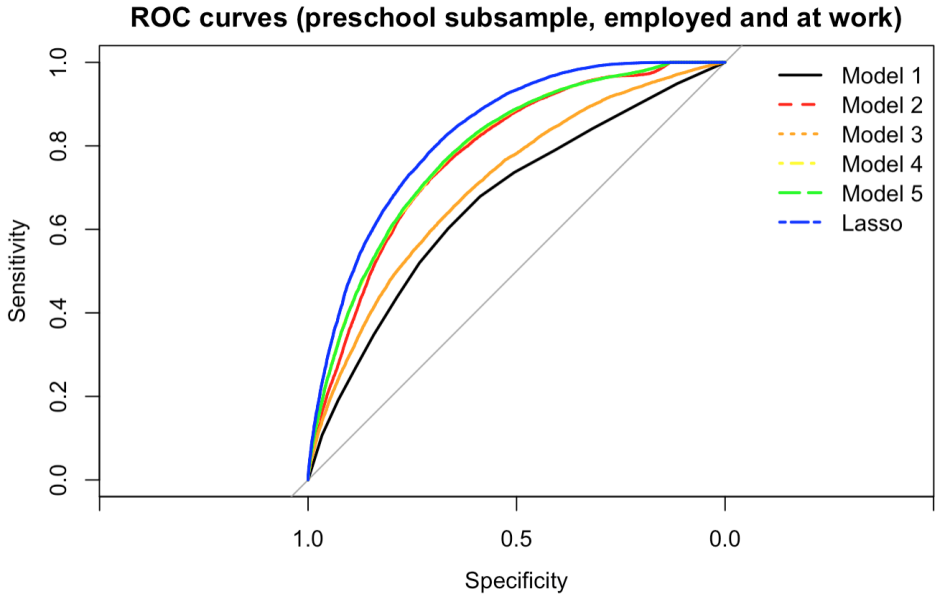

Figure 3.2: ROC curves for each method and model using employed and at work as the dependent variable for the subsample of parents with preschool-aged children.

Figure 3.3 shows the ROC curves for each model and method when employment is used as the dependent variable for the subsample of parents with school-aged children. Much like with the subsample of parents with preschool-aged children, models 1 and 3 perform the worst in terms of classification. Models 2, 4, 5 and the lasso model perform about the same, given that their ROC curves overlap.

Figure 3.4 shows the ROC curves for each model and method when employed and at work is used as the dependent variable for the subsample of parents with school-aged children. The ranking of each model based on their ROC curves is more clear on this graph compared to the other cases because there is not as much overlap of the ROC curves. Model 1 performs the worst, followed by model 3 and model 2. The ROC curves of models 4 and 5 overlap and the lasso's ROC curve dominates, indicating that the lasso model has the best overall performance in terms of classification. The results from the ROC curves for each case suggest that one should use a statistical method to select or combine models because the models that were selected by model selection, model averaging and machine learning

are the models that perform the best in terms of classification.

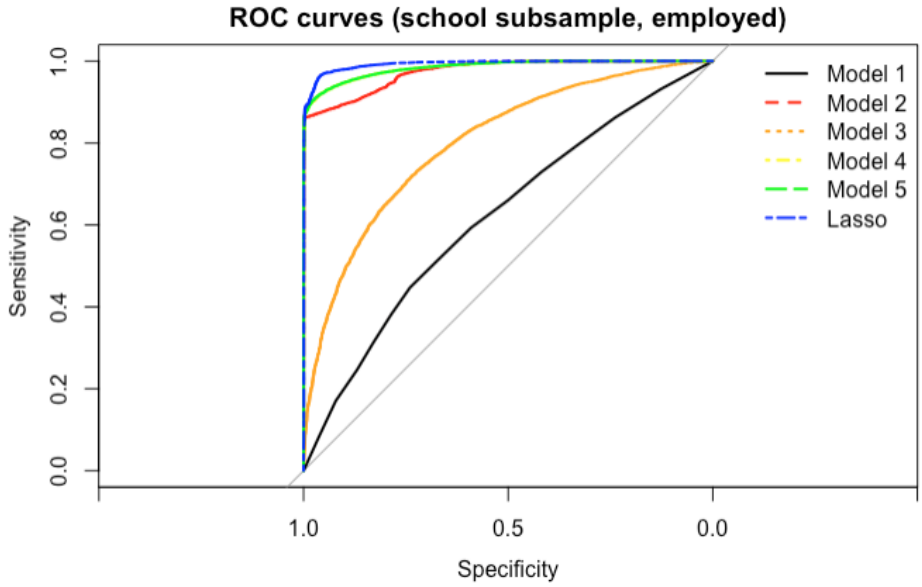

Figure 3.3: ROC curves for each method and model using employment as the dependent variable for the subsample of parents with school-aged children.

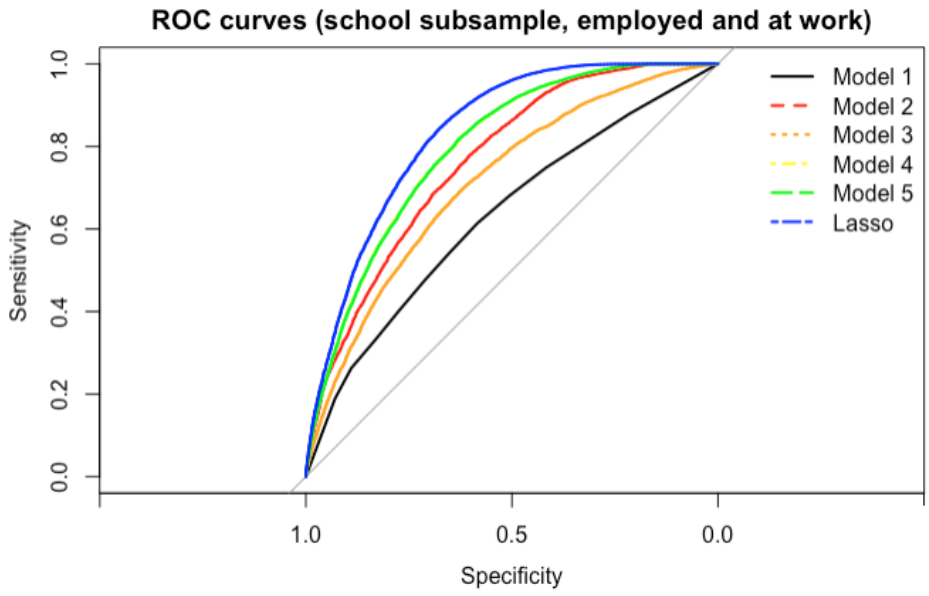

Figure 3.4: ROC curves for each method and model using employed and at work as the dependent variable for the subsample of parents with school-aged children.

We can also calculate the *area under the curve* (AUC) for each ROC curve. The AUC is another measure of overall model performance. The AUC is a value between 0 and 1, with no reasonable classifier having an AUC value less than 0.5, as this would correspond to an ROC along the 45 degree line and thus be equivalent to random guessing (Fawcett, 2006). A higher AUC value is preferred. Table 3.7 shows the AUC for each method and model across different dependent variables and subsamples. The results from the AUC confirm the results from the ROC curves: the lasso model has the best overall classification performance, followed by the models selected by model selection and model averaging (models 4 and 5). Finally, we can obtain the "optimal" cutoff probability $\tau*$ from each

Table 3.7: Area under the curve (AUC) for two different dependent variables across 5 candidate models. The maximum AUC for each column is in bold. A higher AUC value is preferred.

	Preschool subsample		School subsample	
	Employed	Employed and at work	Employed	Employed and at work
Model 1	0.607	0.665	0.62	0.631
Model 2	0.971	0.779	0.974	0.765
Model 3	0.776	0.709	0.801	0.713
Model 4	0.979	0.788	0.984	0.797
Model 5	0.979	0.789	0.984	0.796
Lasso	**0.992**	**0.828**	**0.993**	**0.835**

ROC curve using a criterion and recalculate the CCR at this cutoff instead of the standard $\tau = 0.5$ cutoff probability. *Youden's J statistic*, or Youden's index, can be used as the optimal cutoff probability and is defined as:

$$J = \text{sensitivity} + \text{specificity} - 1. \tag{3.8}$$

The value of Youden's J statistic ranges from 0 to 1 (Youden, 1950). Youden's J statistic is defined for all points along an ROC curve and its maximum value can be used as the optimal cutoff probability. Table 3.8 shows the CCR for each model and method using the optimal cutoff probability obtained for each model across different dependent variables

and subsamples using Youden's J statistic to select the optimal cutoff for each model. Compared to Table 3.6, models 1 and 3 perform the worst. The lasso has the highest CCR for all cases, followed by models 4 and 5, which perform relatively the same in terms of CCR with optimal cutoff probabilities. This provides further evidence that model uncertainty must be addressed by using a method like model selection, model averaging, or machine learning, as one can do much worse by choosing a model in an ad hoc manner compared to using these methods.

Table 3.8: Correct classification rate (CCR) of 5 models and the lasso model across two different dependent variables and two subsamples using optimal cutoff points. The maximum CCR for each column is in bold.

	Preschool subsample		School subsample	
	Employed	Employed and at work	Employed	Employed and at work
Model 1	0.4924	0.6515	0.4769	0.6066
Model 2	0.8967	0.7165	0.8755	0.7373
Model 3	0.6649	0.6537	0.7226	0.6676
Model 4	0.9115	0.7403	0.9052	0.7798
Model 5	0.9113	0.736	0.9153	0.7728
Lasso	**0.9511**	**0.7688**	**0.9633**	**0.8044**

3.5.2 Marginal Effects

Comparing the methods and models in the previous section using classification metrics such as CCR, ROC curves, and AUC showed that some models perform markedly better than others, and those models are the ones selected by model selection or model averaging as well as the lasso model. Another important consideration is whether the estimates from each model differ in an economically important way. If estimates differ in an economically meaningful way, this further emphasizes the importance of using a leading statistical method to choose a model or combine models rather than choosing a model in an ad hoc manner.

In this section, I calculate the marginal effect of sex on employment or being employed

and at work by survey month for parents with young children over the first six months of the COVID-19 pandemic. While this is but an illustrative exercise to see if and how estimates from each method and model differ, the marginal effect of sex on employment – otherwise known as the gender employment gap – is an interesting and timely effect to measure, as the COVID-19 pandemic has disproportionately affected women. As Canadian provincial governments closed schools, childcare facilities, and recreational programs to mitigate the spread of the novel coronavirus, childcare and homeschooling responsibilities fell on the mother in most heterosexual couples. Many women were forced to leave their jobs to take on household and childcare work. Access to safe and affordable childcare may be one of the biggest determinants of the speed of economic recovery of women in Canada. Some work has already been done to estimate the early impact of the pandemic on the labour force participation of women in Canada (Lemieux et al., 2020; Qian & Fuller, 2020). However, the papers cited above appear, without exception, to be based on a parametric model chosen by the researcher from the universe of models available. That model may have been selected by an arbitrary and ad hoc selection procedure or even by some model selection criterion. However, this reliance on a single model, if misspecified, can have serious consequences, such as inference that is overly optimistic or possibly misleading. Additionally, the estimated effects and predictions may vary depending on the model selected and, when these estimates are meant to be used by policymakers to respond to the recession caused by the pandemic, the consequences of ignoring the uncertainty in the model specification can be serious. Therefore, the comparison of methods and models above as well as the exercise below, which compares the difference in magnitude of estimates resulting from these methods and models, is timely and beneficial.

Due to the non-linear nature of the logistic model, the marginal effect of the change in an explanatory variable, such as sex, on the conditional probability that employment = 1

(or that employed and at work = 1) is:

$$\frac{d\Pr(\mathrm{E}=1|\mathbf{X},\mathbf{Z})}{dx_j} = F'(\mathbf{X}\beta,\mathbf{Z}\alpha)\beta_j$$
$$= F'(\mathbf{X}\beta,\mathbf{Z}\alpha)(1-F'(\mathbf{X}\beta,\mathbf{Z}\alpha))\beta_j. \quad (3.9)$$

The marginal effect depends on the values of $\mathbf{X}$ and $\mathbf{Z}$. In my analysis, I calculate the marginal effects at the mode for discrete variables and the median for continuous variables (in this case, tenure is the only continuous variable). Table 3.9 shows the mode and median values for each explanatory variable across each subsample. The coefficients β and α indicate the sign of the marginal effects and can be thought of as placing an upper bound on the marginal effect. My variable of interest is sex, and so the marginal effect of sex on employment represents the gender employment gap. Table 3.10 shows the marginal effect

Table 3.9: Mode (for discrete variables) and median (for continuous variables) values of explanatory variables across 2 subsamples.

	Preschool subsample	School subsample
Province	ON	ON
Age	35-39 years	40-44 years
Marital status	Married	Married
Education	Postsecondary	Postsecondary
Immigration status	Non-immigrant	Non-immigrant
Occupation	Industrial, electrical and construction trades	Professional occupations in education services
Industry	Health care and social assistance	Health care and social assistance
Job category	Private	Private
Full- or part-time	Full-time	Full-time
Tenure	58 months	83 months
Type of economic family	Dual-earner couple, youngest child 0-17 years	Dual-earner couple, youngest child 0-17 years

of sex by survey month for parents with preschool-aged children estimated in percentage points (e.g. 6.0 is six percentage points). Using employment as the dependent variable, we see that there is a difference in the magnitude of estimated marginal effects across the

models. The estimates from models 2, 4 (which was selected in some cases by BIC as the model selection criterion), 5 (which was selected by model selection and model averaging) and the lasso (which performed very well based on classification metrics) show no effect of sex on the probability of being employed across these months for an individual with the characteristics described in Table 3.9. The estimates from models 1 and 3, on the other hand, show a widening gender employment gap from February to August 2020. Given that model 4, model 5 and the lasso have controls for demographic characteristics, job-related characteristics, and/or interactions among the explanatory variables, this means that some of the gender gap observed using the simplest or most "naïve" model (model 1) can be attributed to these explanatory variables. Using employed and at work as the dependent variable, there is a widening gender gap across all models from February to August 2020, with a partial recovery in April. This indicates that mothers were less likely to be employed and at work during the pandemic compared to fathers, all else being equal. The magnitude of the marginal effect of sex differs across models. Table 3.11 shows the magnitude of the marginal effects relative to model 1, the "naïve" model. A value less than 100 shows that the estimated marginal effect of sex from a model is smaller than that of model 1; a value greater than 100 shows that the estimated marginal effect of sex is larger than that of model 1. These results show that model 1 tends to overstate the magnitude of the marginal effect of sex on the probability of being employed and not absent from work compared to the models that performed the best in terms of CCR, AUC, and ROC curves. Thus, different models produce different estimates of the gender gap in employment and in being employed and at work among parents with preschool-aged children over the first six months of the pandemic.

Table 3.12 shows the marginal effect of sex by survey month for parents with school-aged children estimated in percentage points. Using employment as the dependent variable, there is a difference in the estimated gender employment gap across models. Like with the subsample of parents with preschool-aged children, the estimates from models 2, 4, 5 and

the lasso show zero effect of sex on the probability of being employed across these months for an individual with the characteristics described in Table 3.9, whereas the estimates from models 1 and 3 show a widening gap from February to August 2020. Again, this suggests that the observed gender gap in employment using the naïve model (model 1) can be explained by demographic and job-related variables. Using being employed and at work as the dependent variable, there exists a widening gender gap among parents between February and August 2020, with some partial recovery in the intervening months. The gender gap for parents with school-aged children is smaller in magnitude than the gap for parents of preschool-aged children, possibly because preschool-aged children need more time and attention from their parents compared to older children. This shows that while there may not have been a gender employment gap among parents during the pandemic, mothers were less likely to be employed and at work during the pandemic compared to fathers, all else being equal. Additionally, the choice of model matters. Table 3.13 shows the estimated marginal effect of sex for each model relative to the naïve model (model 1). These results show that model 1 overstates the marginal effect of sex on the probability of being employed and at work compared to the models that performed the best in terms of CCR, AUC, and ROC curves (model 4, model 5 and the lasso). Interestingly, the lasso yields a larger gender gap in being employed and at work for these parents in February 2020 compared to model 1. This exercise shows that different models will produce different estimates of the gender gap in employment and in being employed and at work among parents with school-aged children over the first six months of the pandemic.

Policymakers who are concerned about model uncertainty should consider adopting one of the methodologies used in this chapter. Model selection, model averaging, and the lasso acknowledge model uncertainty by taking the decisions with regard to model specification out of the hands of the researcher and adopting a statistical approach to select or combine models or regressors. Additionally, these methods are accessible and straightforward to use, especially for those who routinely use parametric models; flexible, in that they can be

applied to many different kinds of models; and produce results that, as demonstrated in this chapter, can perform at least as well as or better than, in terms of classification metrics, any other model in the set of candidate models.

Table 3.10: Marginal effect of sex by survey month for parents with preschool-aged children (under 6 years) for two different dependent variables across candidate models and the lasso model (percentage points).

	Employed						Employed and at work					
	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso
Feb	-1.11	0.00	0.08	0.00	0.00	0.42	-14.81	-13.54	-15.15	-12.63	-13.39	-20.16
Mar	-3.80	-0.01	-2.80	-0.01	-0.01	0.42	-25.05	-25.08	-25.18	-23.85	-24.84	-23.40
Apr	-0.82	0.00	1.41	0.00	0.00	0.42	-17.85	-19.06	-16.73	-17.48	-18.07	-13.02
May	-2.65	0.00	-1.80	0.00	0.00	0.42	-20.25	-20.00	-20.22	-18.92	-19.68	-26.10
Jun	-3.87	-0.01	-3.47	-0.01	-0.01	0.42	-21.84	-20.63	-22.69	-19.85	-20.72	-20.43
Jul	-6.29	-0.01	-6.27	-0.01	-0.01	0.42	-25.05	-24.33	-25.81	-23.46	-24.45	-20.40
Aug	-6.00	-0.01	-6.05	-0.01	-0.01	0.42	-28.17	-27.27	-29.35	-26.54	-27.52	-23.97

Table 3.11: Magnitude of the marginal effect of sex by survey month for parents with preschool-aged children relative to model 1 (percentage points).

	Employed						Employed and at work					
	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso
Feb	100	0.00	7.21	0.00	0.00	37.84	100	91.42	102.30	85.28	90.41	136.12
Mar	100	0.26	73.68	0.26	0.26	11.05	100	100.12	100.52	95.21	99.16	93.41
Apr	100	0.00	171.95	0.00	0.00	51.22	100	106.78	93.73	97.93	101.23	72.94
May	100	0.00	67.92	0.00	0.00	15.85	100	98.77	99.85	93.43	97.19	128.89
Jun	100	0.26	89.66	0.26	0.26	10.85	100	94.46	103.89	90.89	94.87	93.54
Jul	100	0.16	99.68	0.16	0.16	6.68	100	97.13	103.03	93.65	97.60	81.44
Aug	100	0.17	100.83	0.17	0.17	7.00	100	96.81	104.19	94.21	97.69	85.09

Table 3.12: Marginal effect of sex by survey month for parents with school-aged children (6-12 years) for two different dependent variables across candidate models and the lasso model (percentage points).

	Employed						Employed and at work					
	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso
Feb	-0.78	0.00	0.45	0	0	0.45	-0.76	-1.81	1.90	0.23	0.25	-5.85
Mar	-4.39	-0.01	-1.63	0	0	0.45	-13.09	-15.78	-9.35	-12.27	-11.77	-8.77
Apr	-5.51	-0.01	-1.42	0	0	0.45	-8.97	-10.92	-4.40	-7.32	-6.98	-0.68
May	-7.08	-0.01	-3.10	0	0	0.45	-9.67	-9.76	-5.51	-6.28	-5.98	-9.73
Jun	-6.44	-0.01	-3.39	0	0	0.45	-10.56	-10.24	-7.89	-8.03	-7.66	-4.83
Jul	-6.59	-0.01	-3.27	0	0	0.45	-12.76	-13.21	-9.77	-10.57	-10.15	-4.51
Aug	-7.63	-0.01	-4.36	0	0	0.45	-13.37	-14.55	-10.26	-11.69	-11.24	-10.58

Table 3.13: Magnitude of the marginal effect of sex by survey month for parents with school-aged children relative to model 1 (percentage points).

	Employed						Employed and at work					
	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso	Model 1	Model 2	Model 3	Model 4	Model 5	Lasso
Feb	100	0.00	57.69	0	0	57.69	100	238.16	250.00	30.26	32.89	769.74
Mar	100	0.23	37.13	0	0	10.25	100	120.55	71.43	93.74	89.92	67.00
Apr	100	0.18	25.77	0	0	8.17	100	121.74	49.05	81.61	77.81	7.58
May	100	0.14	43.79	0	0	6.36	100	100.93	56.98	64.94	61.84	100.62
Jun	100	0.16	52.64	0	0	6.99	100	96.97	74.72	76.04	72.54	45.74
Jul	100	0.15	49.62	0	0	6.83	100	103.53	76.57	82.84	79.55	35.34
Aug	100	0.13	57.14	0	0	5.90	100	108.83	76.74	87.43	84.07	79.13

3.6 Conclusion

This chapter uses model selection, model averaging, and machine learning to estimate binary-choice regressions of employment among parents with preschool- and school-aged children in Canada during the first six months of the COVID-19 pandemic. Using 5 parametric binary-choice models, I find that model selection using BIC selects the largest or second largest model; model selection using AIC selects the largest model; and model averaging selects the largest model. I use classification metrics to evaluate the relative performance of each method and model – including a model obtained through the lasso – and find that the models selected by model selection and model averaging as well as the lasso model perform better than simpler parametric model specifications, which have been used in practice based on identical data (Qian & Fuller, 2020). The lasso model performs the best overall in terms of classification metrics. This demonstrates that there are methods that perform better than others from a classification perspective and suggests that methods like model selection, model averaging, or machine learning should be used instead of selecting a model in an ad hoc manner. Finally, to further support the use of these methods, I show that the marginal effect of sex on the probability of employment or being employed at work varies across models in an economically meaningful way. Thus, the choice of model matters, especially when the estimates from these models could inform, for example, policies for post-pandemic economic recovery. Future research should focus on evaluating the performance of these methods across a larger set of candidate models that include interactions among explanatory variables as well as higher order polynomials in explanatory variables.

3.7 Appendix

3.7.1 Descriptive Statistics

Table 3.14 shows the proportion of individuals in each occupational category for the Labour Force Survey subsample of individuals aged 20-64 years, currently employed or employed within the last year, and with a youngest child aged 0-12 years.

Table 3.14: Summary statistics of occupation for Labour Force Survey subsamples of individuals aged 20-64 years, currently employed or employed within the last year, and with a youngest child aged under 6 years (preschool subsample) or 6-12 years (school subsample).

Occupational category	Preschool (Percent)	School (Percent)
Senior management occupations	0.13	0.35
Specialized middle management occupations	2.71	4.31
Middle management occupations in retail and wholesale trade and customer services	2.11	2.98
Middle management occupations in trades, transportation, production and utilities	3.29	3.53
Professional occupations in business and finance	4.44	4.53
Administrative and financial supervisors and administrative occupations	5.25	5.86
Finance, insurance and related business administrative occupations	1.30	1.42
Office support occupations	2.95	3.27

(continued . . .)

Table 3.14: Summary statistics of occupation for Labour Force Survey sub-samples.

Occupational category	Preschool (Percent)	School (Percent)
Distribution, tracking and scheduling co-ordination occupations	1.19	1.34
Professional occupations in natural and applied sciences	5.67	4.53
Technical occupations related to natural and applied sciences	3.88	3.30
Professional occupations in nursing	2.95	2.06
Professional occupations in health (except nursing)	2.48	2.39
Technical occupations in health	3.21	2.52
Assisting occupations in support of health services	2.31	2.11
Professional occupations in education services	6.22	6.89
Professional occupations in law and social, community and government services	3.43	3.27
Paraprofessional occupations in legal, social, community and education services	3.18	3.16
Occupations in front-line public protection services	1.01	1.21

(continued . . .)

Table 3.14: Summary statistics of occupation for Labour Force Survey sub-samples.

Occupational category	Preschool (Percent)	School (Percent)
Care providers and educational, legal and public protection support occupations	1.61	2.09
Professional occupations in art and culture	0.65	0.74
Technical occupations in art, culture, recreation and sport	1.32	1.13
Retail sales supervisors and specialized sales occupations	2.96	3.30
Service supervisors and specialized service occupations	2.63	2.65
Sales representatives and salespersons - wholesale and retail trade	2.54	2.70
Service representatives and other customer and personal services occupations	3.09	2.91
Sales support occupations	1.33	1.41
Service support and other service occupations, n.e.c.	2.69	3.05
Industrial, electrical and construction trades	6.81	5.16
Maintenance and equipment operation trades	3.99	3.75

(continued . . .)

Table 3.14: Summary statistics of occupation for Labour Force Survey subsamples.

Occupational category	Preschool (Percent)	School (Percent)
Other installers, repairers and servicers and material handlers	1.11	1.04
Transport and heavy equipment operation and related maintenance occupations	3.51	3.51
Trades helpers, construction labourers and related occupations	0.71	0.67
Supervisors and technical occupations in natural resources, agriculture and related production	1.73	1.55
Workers in natural resources, agriculture and related production	1.02	0.86
Harvesting, landscaping and natural resources labourers	0.49	0.43
Processing, manufacturing and utilities supervisors and central control operators	1.24	1.25
Processing and manufacturing machine operators and related production workers	1.52	1.46
Assemblers in manufacturing	0.77	0.64
Labourers in processing, manufacturing and utilities	0.59	0.63

Table 3.15 shows the proportion of individuals in each industry for the Labour Force Survey subsample of individuals aged 20-64 years, currently employed or employed within

the last year, and with a youngest child aged 0-12 years.

Table 3.15: Summary statistics of industry for Labour Force Survey subsamples of individuals aged 20-64 years, currently employed or employed within the last year, and with a youngest child aged under 6 years (preschool subsample) or 6-12 years (school subsample).

Industry category	Preschool (Percent)	School (Percent)
Agriculture	2.17	2.13
Forestry and logging	0.45	0.40
Fishing, hunting and trapping	0.38	0.39
Mining, quarrying, and oil and gas extraction	2.99	2.44
Utilities	1.17	1.03
Construction	9.65	7.51
Manufacturing - durable goods	4.69	4.93
Manufacturing - non-durable goods	3.87	3.88
Wholesale trade	2.74	3.21
Retail trade	7.21	7.79
Transportation and warehousing	4.62	4.85
Finance and insurance	4.31	4.66
Real estate and rental and leasing	1.26	1.53
Professional, scientific and technical services	7.22	6.47
Business, building and other support services	3.12	3.01
Education services	9.33	11.46
Health care and social assistance	17.07	15.97
Information, culture and recreation	3.15	3.22
Accommodation and food services	3.79	3.80
Other services (except public administration)	3.79	3.60
Public administration	7.01	7.71

Conclusion

My research focuses on model averaging, a leading approach for handling the issue of model uncertainty – that is, the likelihood that one's model is misspecified. I evaluate the performance of model averaging compared to conventional econometric approaches as well as machine learning algorithms, and show how model averaging can be applied to an empirical problem in labour economics.

Chapter 1 evaluates the performance of frequentist model averaging (FMA) compared to models in the set of candidate models as well as model selection and machine learning algorithms. Results from Monte Carlo simulations show that model averaging does relatively well compared to individual models and other methods in terms of mean squared error (MSE) in the presence of model uncertainty. Additionally, using the National Longitudinal Survey, I estimate the returns to education to demonstrate how easily model averaging can be adopted by empirical economists. I also include a novel emphasis on the set of candidate models that are averaged, highlighting this step of model averaging which is further studied in Chapter 2.

Chapter 2 investigates three promising approaches for constructing a set of candidate models to be used in model averaging: model screening, recursive partitioning-based algorithms, and methods that average over nonparametric models. The candidate model set should balance model complexity, breadth, and computational efficiency. The Monte Carlo experiments demonstrate that the nonparametric model averaging approach performs the best in terms of MSE compared to model averaging over parametric models and the MARS

algorithm in most cases. None of these approaches alone can build a set of candidate models that satisfies all the desired criteria; however, these approaches are an improvement to model averaging over a set of arbitrarily chosen parametric candidate models.

Chapter 3 uses the methods examined in Chapter 1 – model selection, model averaging, and the lasso – to assess the impacts of the COVID-19 pandemic on the employment of parents with young children in Canada. I use data from the Canadian Labour Force Survey public use microdata files. I find that the models selected by model selection and model averaging and the lasso model perform better in terms of classification compared to the simpler parametric model specifications. Additionally, I estimate the marginal effect of sex on the probability of being employed and find that the results differ in magnitude across models in an economically important way, as these results could affect policies for post-pandemic recovery.

My work shows how important model choice is to statistical and economic analysis, as final estimates and predictions may rely on the model chosen. If one is concerned about model uncertainty, model averaging, machine learning and model selection methods can be used. These methods take ad hoc decision-making out of the hands of the researcher and instead select or combine models based on the data.

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, *22*(1), 203–217.

Alboukadel, K. (2018). *Machine learning essentials: Practical guide in r*. Sthda.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Belman, D., & Heywood, J. S. (1991). Sheepskin effects in the returns to education: An examination of women and minorities. *The Review of Economics and Statistics*, *73*(4), 720–724.

Blackburn, M., & Neumark, D. (1992). Unobserved ability, efficiency wages, and interindustry wage differentials. *Quarterly Journal of Economics*, *107*(4), 1412–1436.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, *26*(3), 801–849.

Buckland, S. T., Burnhamn, K. P., & Augustin, N. H. (1997). Model selection: An integral

part of inference. *Biometrics*, *53*, 603–618.

Campos, J., Hendry, D. F., & Krolzig, H.-M. (2003). Consistent model selection by an automatic gets approach. *Oxford Bulletin of Economics and Statistics*, *65*, 803–819.

Castle, J. L. (2006). Automatic econometric model selection using pcgets. *Medium Econometrische Toepassingen*, *14*(3), 16–19.

Claeskens, G., Croux, C., & VanKerckhoven, J. (2005). *Variable selection for logistic regression using a prediction focussed information criterion*. Belgian Statistical Society.

Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*(464), 900–916.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. New York: Cambridge University Press.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*, *96*, 148–156.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Freund, Y., Schapire, R. E., & Abe, N. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*(5), 771–780.

Friedman, J. (1991a). Estimating functions of mixed ordinal and categorical variables using adaptive splines. *Stanford University Technical Report*, (108), 1–49.

Friedman, J. (1991b). Multivariate adaptive regression splines. *The Annals of Statistics*,

19(1), 1–141.

Friedman, J. (1993). Fast mars.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J., & Roosen, C. (1995). An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research*, *4*, 197–217.

Friedman, J., & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, *31*(1), 3–21.

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

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, *5*(3), 495–530.

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*(1), 38–46.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). Springer.

Heywood, J. S. (1994). How widespread are sheepskin returns to education in the u.s.? *Economics of Education Review*, *13*(3), 227–234.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Hungerford, T., & Solon, G. (1987). Sheepskin effects in the returns to education. *The Review of Economics and Statistics*, *69*(1), 175–177.

Jaeger, D. A., & Page, M. E. (1996). Degrees matter: New evidence on sheepskin effects in the returns to education. *The Review of Economics and Statistics*, *78*(4), 733–740.

Jones, S. R. G., Lange, F., Ridell, W. C., & Warman, C. (2020). *Waiting for recovery: The canadian labour market in june 2020*.

Krolzig, H.-M., & Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, *25*(6-7), 831–866.

Krolzig, H.-M., & Hendry, D. F. (2011). New developments in automatic general-to-specific modeling.

Lemieux, T., Milligan, K., Schirle, T., & Skuterud, M. (2020). Initial impacts of the covid-19 pandemic on the canadian labour market. *Canadian Public Policy*, *46*(S1), S55–S65.

Liu, Y., & Xie, T. (2019). Machine learning versus econometrics: Prediction of box office. *Applied Economics Letters*, *26*(2), 124–130.

Ma, S., Racine, J. S., & Yang, L. (2015). Spline regression in the presence of categorical predictors. *Journal of Applied Econometrics*, *30*, 705–717.

Madigan, D., & 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.

Madigan, D., York, J., & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 215–232.

Mallows, C. L. (1973). *Technometrics*, *15*(4), 661–675.

Moral-Benito, E. (2015). Model averaging: An overview. *Journal of Economic Surveys*, *29*(1), 46–75.

Mullianathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Qian, Y., & Fuller, S. (2020). COVID-19 and the gender employment gap among parents

of young children. *Canadian Public Policy*, *46*(S2), S89–S101.

Racine, J. S. (2019). *Reproducible econometrics using r* (pp. 191–195). Oxford University Press.

Racine, J. S., Li, Q., & Zheng, L. (2018). Optimal model averaging of mixed-data kernel-weighted spline regressions. *McMaster Department of Economics Working Paper Series*.

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

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

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719.

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

Tobias, J. L., & Li, M. (2004). Returns to schooling and bayesian model averaging: A union of two literatures. *Journal of Economic Surveys*, *18*(2), 153–180.

Tukey, J. (1970). *Exploratory data analysis*. Addison-Wesley.

Xie, T., & Lehrer, S. (2017). Box office buzz: Does social media data steal the show from model uncertainty when forecasting for hollywood? *The Review of Economics and Statistics*, *99*(5), 749–755.

Xie, T., & Lehrer, S. (2018). The bigger picture: Combining econometrics with analytics improve forecasts of movie success. *NBER Working Series*.

X. Zhang, G. Z., D. Yu, & Liang, H. (2016). Optimal model averaging estimation for

generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, *111*(516), 1775–1790.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, *100*(472), 1202–1214.

Zhu, R., Wan, A. T. K., Zhang, X., & Zou, G. (2019). A mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, *114*(526), 882–892.