National and Kapodistrian University of Athens
School of Science
Department of Mathematics
Section of Statistics and Operations Research

Master Thesis

Predictive Regressions: Variable Selection and the Complete Subset Approach

Author:
Alexandros Karampateas

Supervisor:
Dr. Loukia Meligkotsidou

September 27, 2018

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ

που απονέμει το Τμήμα Μαθηματικών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών.

Εγκρίθηκε την 27/9/2018, από την εξεταστική επιτροπή:

Ονοματεπώνυμο	Βαθμίδα	Υπογραφή
Λ.Μελιγκοτσίδου (επιβλέπουσα)	Επίκουρη Καθηγήτρια	
Σ.Τρέβεζας	Λέκτορας	
Α.Μπουρνέτας	Καθηγητής	

National and Kapodistrian University of Athens
School of Science
Department of Mathematics
Section of Statistics and Operations Research

Abstract

Predictive Regressions: Variable Selection and the Complete Subset Approach

This dissertation is concerned with the problem of controlling the estimation error in forecasting, having many potential predictor variables. While having to deal with a limited number of independent variables permits any strategy that includes and analyzes all of them, when this number gets higher (or equivalently the data sample is relatively short), it is important to limit the number of parameters or in other ways reduce the effect of the parameter estimation error. Otherwise, analysis can become from time intensive to impossible.

Complete Subset Regression is a simple and powerful method/technique for combining forecasts, first introduced by Elliott et al. (2013). In particular, for a given set of potential predictor variables, forecasts from all possible linear regression models that keep the number of predictors fixed are combined. This method is akin to a complex version of shrinkage which, in general, does not reduce to shrinking the Ordinary Least Squares estimates coefficient by coefficient.

Apart from the apparent savings in terms of computational effort, combinations of subset regressions can produce accurate forecasts compared to other conventional, still very well established, approaches.

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται το θέμα του περιορισμού του σφάλματος εκτίμησης για τις περιπτώσεις προβλέψεων όπου εμπλέκονται πολλές επεξηγηματικές μεταβλητές. Ενώ στις περιπτώσεις όπου οι ανεξάρτητες μεταβλητές είναι λίγες σε αριθμό είναι δυνατή η εφαρμογή οποιασδήποτε στρατηγικής που λαμβάνει υπόψη το σύνολο των μεταβλητών, όταν αυτός ο αριθμός μεγαλώσει (ή ισοδύναμα το μέγεθος του δείγματος είναι σχετικά μικρό), καθίσταται σημαντικός ο περιορισμός του αριθμού των παραμέτρων, ώστε να μειωθεί η επίδραση του εκτιμητικού σφάλματος. Σε αντίθετη περίπτωση, η ανάλυση μπορεί να καταστεί από χρονοβόρα έως και αδύνατη.

Η μέθοδος Complete Subset Regression είναι μία απλή και ισχυρή μέθοδος/τεχνική συνδυασμού προβλέψεων, που πρώτοι παρουσίασαν οι Elliott et al. (2013). Συγκεκριμένα, διαθέτοντας ένα δεδομένο σύνολο πιθανών επεξηγηματικών μεταβλητών, συνδυάζονται οι προβλέψεις από κάθε υποσύνολο μοντέλων γραμμικής παλινδρόμησης το οποίο έχει σταθερό αριθμό μεταβλητών. Αυτή η μέθοδος προσιδιάζει σε μία περίπλοκη εκδοχή συρρίκνωσης, η οποία όμως εν γένει δεν περιορίζεται στη συρρίκνωση των εκτιμητών ελαχίστων τετραγώνων, συντελεστή προς συντελεστή.

Πέρα από την προφανή εξοικονόμηση υπολογιστικού φόρτου, οι συνδυασμοί των πλήρων υποσυνόλων παλινδρόμησης μπορούν να οδηγήσουν σε ακριβείς προβλέψεις, συγκρίσιμες με άλλες καθιερωμένες μεθόδους.

Acknowledgements

I would like to express my gratitude to my employer, for giving me the opportunity to complete this Master of Science. I sincerely hope to have made the best of it and that I'll be able to overcome all the challenges in the years to come.

Moreover, I need to thank my wife for the unconditional support and patience throughout this great period of our life. Even greater some weeks now with our newborn son!

Last but not least, I would like to thank my thesis supervisor Loukia Meligkotsidou, for her patient guidance and firm encouragement.

Alexandros Karampateas
September 27, 2018

Contents

Introduction

The Complete Subset Regression is a simple and powerful method/technique for combining forecasts. For a given set of potential predictor variables, forecasts from all possible linear regression models that keep the number of predictors fixed are combined. In cases where the number of independent variables is limited this may not seem so useful. However, when the number of potential predictors gets larger, this method offers a computationally efficient technique that can produce accurate forecasts compared to other conventional, still very well established, approaches.

This dissertation examines the main theoretical features, as well as the performance of this method, and is organized as follows.

Chapter 1 starts with a general description of statistical inference and linear regression analysis, then provides a brief outline of the multiple linear model, the ordinary least squares and maximum likelihood methods for coefficients' estimation, and concludes with the concept of prediction and the predictive distribution.

Chapter 2 describes (some of) the classical methods and criteria for model selection, such as the coefficient of determination, the residual mean square, the Akaike Information Criterion and the stepwise regression.

Chapter 3 is about the Bayesian methods for model selection. Briefly covering the basic concepts of Bayesian analysis, such as the prior and posterior distributions, it deals with the Bayesian linear model, to conclude with Bayesian Model Averaging.

Chapter 4 is where the Complete Subset Regression is presented, including its theoretical review and its aspects of implementation.

Chapter 5 starts with the evaluation of forecasting accuracy in general and then facilitates the simulation study held in order to assess the performance of the Complete Subset Regression and compare it with several other widely known and used methods.

Finally, chapter 6 presents an application to a data set from Goyal and Welch (2008), serving the very same means of evaluation as chapter 5.

Chapter 1: Statistical inference, linear regression and prediction

1.1 Statistical inference: classical and Bayesian approach

Probability and statistics can be characterized as the study of variability. Statistics is profoundly concerned with the collection of data and with their analysis and interpretation. Statistical inference, in particular, is the science of inferring properties and making conclusions about a "population" from "sample", including testing hypotheses and deriving estimates. Taking the data as given, the answer of what they have to tell us depends not only on what is being observed, but also on background knowledge of the situation, which is formalized in the assumptions with which the analysis is entered. Leaving aside the data analysis, where data are analyzed on their own terms, essentially without extraneous assumptions, there have been typically two principal lines of approach: classical inference and Bayesian analysis.

Classical inference and decision theory: the observations are postulated to be the values taken on by random variables which are assumed to follow a joint probability distribution, P, belonging to some known class ***P***. Frequently, the distributions are indexed by a parameter, say θ, taking values in a set, so that $\mathrm{P} = \{\mathrm{P}_\theta, \theta \in \Omega\}$. The aim of the analysis is then to specify a plausible value for θ (the problem of point estimation) or to determine a subset of Ω of which we can plausibly assert that it does, or does not, contain θ (estimation by confidence sets or hypothesis testing). Such a statement about θ can be viewed as a summary of the information provided by the data and may be used as a guide to action. The most fundamental point of this classical approach (actually the cornerstone of classical theory) is that the parameter θ, whilst not known, is being treated as constant rather than random. Unfortunately, this leads to problems of interpretation, since statistical procedures receive a long-term meaning, like an infinite repetition of the same experiment.

Bayesian analysis: in this approach, it is assumed (in addition) that θ is itself a random variable (though unobservable) with a known distribution. This distribution, called the prior, is specified according to the problem and is modified in light of the data to determine a posterior distribution (the conditional distribution of θ given the data), which summarizes what can be said about θ on the basis of the assumptions made and the data. In essence, inference is based on $\mathrm{f}(\theta|\mathrm{x})$ rather than $\mathrm{f}(\mathrm{x}|\theta)$; that is the probability distribution of the parameter given the data, rather than the data given the parameter. In many ways this leads to much more natural inferences (since mere use of the rules of probabilities is needed), but the specification of the prior

probability distribution is a prerequisite, which represents beliefs about the distribution of θ prior to having any information about the data. This is the core of Bayesian thinking and as so, is considered either its primary advantage over classical theory or its biggest pitfall. Putting together the aforementioned aspects of the Bayesian approach we get the following key steps:

- Specification of a likelihood model $f(x|\theta)$
- Determination of a prior distribution $f(\theta)$
- Calculation of posterior distribution $f(\theta|x)$ from Bayes' theorem. That is, $f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)d\theta'}$, stated in terms of random variables with densities denoted generically by f
- Drawing inferences from this posterior information.

1.2 Regression analysis: simple and multiple linear regression

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. In linear regression the relationships are modeled using linear predictor functions; thus, a linear regression analysis yields a mathematical equation—a linear model—that estimates a dependent variable Y from a set of predictor/explanatory variables (regressors) X. Each regressor in a linear model is given a numerical weight called its regression coefficient, regression slope, or simply its regression weight that determines how much the equation uses values on that variable to produce an estimate of Y. These regression weights are derived by an algorithm that produces a mathematical equation or model for Y that "best" fits the data, using some kind of criterion for defining "best."

The simplest linear model involves only one independent variable, the regressor X and states that the relationship with the response Y is a straight line. This simple linear regression model, for the i-th sample unit, is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where the intercept β_0 and the slope β_1 are unknown constants and ε_i is a random error component. The Y_i and X_i are paired observations, both measured on every observational unit. For the random errors we make the following assumptions:

$$E(\varepsilon_i) = 0, \forall i,$$

$$Var(\varepsilon_i) = \sigma^2, \forall i,$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j,$$

that is, they are assumed to have zero mean, unknown constant variance σ^2 (the homoscedasticity assumption) and additionally we assume that they are uncorrelated.

It is convenient to view the regressor X as controlled by the data analyst and measured with negligible error, while the response Y is a random variable. That is, there is a probability distribution for Y at each possible value for X. The mean of this distribution is:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i,$$

and the variance is:

$$Var(Y_i|X_i) = Var(\beta_0 + \beta_1 X_i + \varepsilon_i) = \sigma^2.$$

Thus, the mean of Y is a linear function of X although the variance of Y does not depend on the value of X. The aforementioned assumptions imply that the $\mathrm{Y_i}$ also have common variance and are pairwise independent. Furthermore, for purposes of making tests of significance, the random errors are assumed to be normally distributed, which again implies that the $\mathrm{Y_i}$ are also normally distributed.

Now, let $\mathrm{X_1, X_2, \dots, X_k}$ be a set of k predictors believed to be related to a response variable Y. In the multiple linear regression model, the expected value of the response variable is assumed to be a linear function of these k regressors. The linear regression model has now the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i,$$

where we now have $\mathrm{p = k + 1}$ unknown parameters (the $\mathrm{\beta_i, i = 0,1, \dots, k}$ regressor coefficients), and the rest of the assumptions are still applicable, as in the simple regression model. Under these assumptions the coefficient $\mathrm{\beta_j, j = 1, \dots, k}$ measures the change in the expected value of Y when $\mathrm{X_i}$ increases by one unit, the other regressors being held fixed.
Denoting the sample size with n, the linear model can be written in matrices form as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

or

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_\kappa \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Each column of **X** contains the values for a particular independent variable. The elements of a particular row of **X**, say row r, are the coefficients on the corresponding parameters in **β** that give $\mathrm{E}(\mathrm{Y_r})$. The vectors **Y** and **ε** are random vectors; the matrix **X** is considered to be a matrix of known constants.

The variance-covariance matrix for the random errors is: $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I_n}$, following the assumption of homoscedasticity and uncorellation, and hence for **Y** we obtain $\mathrm{E}(\mathrm{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I_n}$.

Letting $\hat{\beta}$ be an estimator of the vector β of the regression coefficients, then the predicted regression is $\widehat{\mathrm{Y}} = \mathrm{X}\hat{\beta}$, which allow us to calculate the fitted value $\widehat{\mathrm{Y}_\mathrm{i}}$ for each $\mathrm{Y_i}$.

The difference $\widehat{\varepsilon_\mathrm{i}} = \mathrm{Y_i} - \widehat{\mathrm{Y}_\mathrm{i}}, \forall \mathrm{i} = 1, \dots, \mathrm{n}$ is called residual and measures the discrepancy between the data and the fitted model. In matrix form, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\Upsilon} - \widehat{\boldsymbol{\Upsilon}}$.

1.3 Regression coefficients estimation: the least square method

As stated, a criterion must be set, for the regression weights to be estimated. The least squares estimation procedure uses the criterion that the solution must give the smallest possible sum of squared deviations of the observed $\mathrm{Y_i}$ from the estimates of their true means provided by the solution. In terms of linear regression the sum of squared residuals is:

$$SSE = \sum_{i=1}^{n} \widehat{\varepsilon_\iota^2} = \widehat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right),$$

and so, for the vector of least squares estimators $\hat{\beta}$ that minimizes the SSE can easily be shown that the least squares normal equations hold:

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}.$$

Provided that the inverse matrix $(\mathrm{X}'\mathrm{X})^{-1}$ exists, which is always the case if the regressors are linearly independent, the least squares estimator of **β** is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

This estimator is linear to Y, unbiased, since $\mathrm{E}(\hat{\beta}) = \beta$, and can be shown that has the least variance among all the linear and unbiased estimators of β; thus known us BLUE (Best Linear Unbiased Estimator), based on the Gauss-Markov theorem. Its variance is: $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$. Moreover, $\mathrm{Var}(\hat{\mathrm{Y}}) = \mathrm{Var}(\mathrm{X}\hat{\beta}) = \mathrm{X}\mathrm{Var}(\hat{\beta})\mathrm{X}' = \sigma^2\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$.

1.4 Maximum likelihood estimation

The method of least squares can be used to estimate the parameters in a linear regression model regardless of the form of the distribution of the random errors. Other statistical procedures, such as hypothesis testing and construction of confidence intervals, assume that the errors are normally distributed. If the form of the random errors' distribution is known, the maximum likelihood estimators are derived using the criterion of finding those values of the parameters that would have maximized the probability of obtaining the particular sample, called the likelihood function. As already established, if $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I_n})$, then $\mathbf{Y} \sim \mathrm{N_n}(\mathbf{X\beta}, \sigma^2\mathbf{I_n})$. The likelihood function for the unknown parameters **β** and σ^2 is:

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right\},$$

and thus the log-likelihood is:

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon},$$

Then, we obtain the MLE for **β** and σ^2 as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

$$\widehat{\sigma_{MLE}}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n}.$$

The MLE and the least squares estimator for **β**, coincide. This is not the case for $\widehat{\sigma^2}$. An unbiased estimator used along with the least squares method is $\widehat{\sigma^2} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{\mathrm{n-k-1}}$, which takes into account the degrees of freedom (the number of parameters estimated subtracted from the number of available observations). While $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ are unbiased estimators for **β** and σ^2, $\widehat{\sigma_{\mathrm{MLE}}}^2$ has Bias$(\widehat{\sigma_{\mathrm{MLE}}}^2) = \sigma^2\frac{\mathrm{n-k-1}}{\mathrm{n}}$. Obviously, $\widehat{\sigma_{\mathrm{MLE}}}^2$ is an asymptotically unbiased estimator confirming the general fact that the MLE have better statistical properties (they are unbiased, consistent and sufficient and have minimum variance when compared to all other unbiased estimators).

1.5 Prediction

Usually the purpose of formulating a statistical model is to make predictions about future values of the process. The essential point in this is that there are two sources of uncertainty: uncertainty in the parameter values which have been estimated on the basis of past data and uncertainty due to the fact that any future value is itself a random event.

Consider the problem of predicting some future observation $Y_0 = X_0\beta + \varepsilon_0$, at a specific value X_0 of X, where it is assumed that $\varepsilon_0 \sim N(0, \sigma^2 I_n)$ independent of the current observations. Since $\widehat{Y_0} = X_0\hat{\beta}$ is used as an estimate of the mean $X_0\beta$ of Y_0, and the best prediction for ε_0 is its mean zero, $\widehat{Y_0}$ is also used as the predictor of Y_0. The variance for prediction must take into account that the quantity being predicted is itself a random variable. Thus, the success of the prediction will depend on how small the difference is between $\widehat{Y_0}$ and the future observation Y_0. The difference $Y_0 - \widehat{Y_0}$ is called the prediction error. The average squared difference $E(\widehat{Y_0} - Y_0)^2$ is called the mean squared error of prediction. It can be shown that $\widehat{Y_0}$ is an unbiased estimator of Y_0 and that $Var(\widehat{Y_0} - Y_0) = \sigma^2[X_0'(X'X)^{-1}X_0 + 1]$. Thus, under the same assumptions for the random errors, the prediction error is normally distributed, with zero mean and the aforegiven variance.

Comparing the variance of the prediction error and the variance for estimation of the mean, we observe that the first is the sum of the latter plus the variance of the quantity being predicted. The derived variances are the true variances and depend on knowledge of σ^2. Estimated variances are obtained by replacing σ^2 in the variance equations with an estimate of it. The residual mean square from the analysis provides an estimate of σ^2 if the correct model has been fitted.
Under the assumption $\varepsilon_i \sim N(0, \sigma^2 I_n)$, we have $(\widehat{Y_0} - Y_0) \sim N\left(0, Var(\widehat{Y_0} - Y_0)\right)$, and thus $\frac{(\widehat{Y_0} - Y_0)}{\sqrt{Var(\widehat{Y_0} - Y_0)}} \sim N(0,1)$. Having to deal with unknown σ^2, we use the fact that $\frac{(\widehat{Y_0} - Y_0)}{\sqrt{\hat{\sigma}^2[X_0'(X'X)^{-1}X_0 + 1]}} \sim t_{n-k-1}$ and the confidence interval established at the α significance level is $\widehat{Y_0} \pm t_{\frac{\alpha}{2}, n-k-1} SE(\widehat{Y_0} - Y_0)$.

So, in classical statistics it is usual to fit a model to the past data, and then make predictions of future values on the assumption that the model is correct, the so called estimative approach. That is, only the second source of uncertainty is included in the analysis, leading to estimates which are believed to be more precise than they

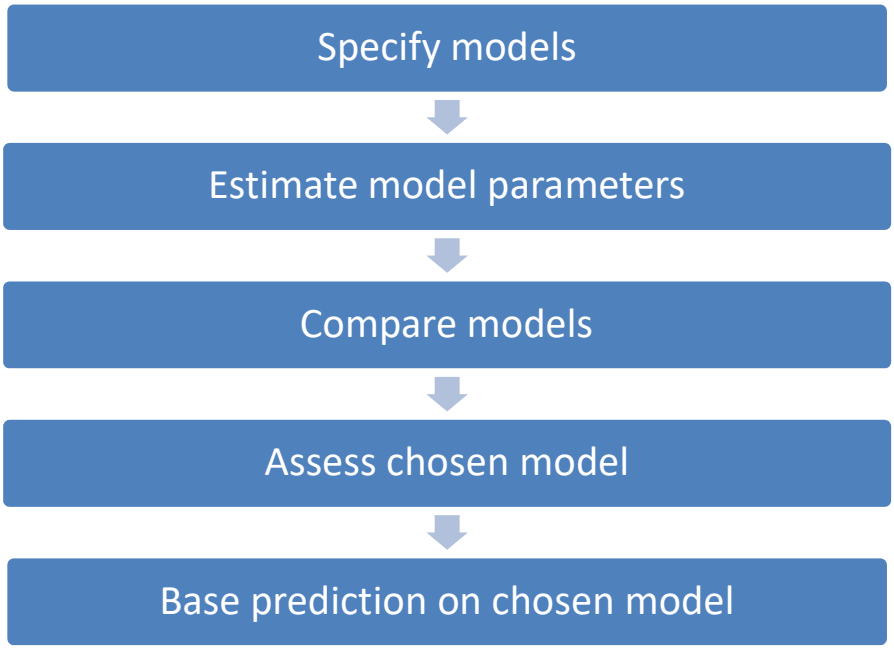
really are. Since parameters are not thought as being random, there is no efficient way to go around this problem in the classical framework. Within Bayesian inference on the other hand, it is straightforward to allow both sources of uncertainty by simply averaging over the uncertainty in the parameter estimates, the information of which is completely contained in the posterior distribution.

Chapter 2: Classical methods and criteria for model selection

2.1 Model development and variable selection

Modeling refers to the development of mathematical expressions that describe in some sense the behavior of a random variable of interest. Statistical models help us to understand the random process by which observed data have been generated, which may be of interest in itself, but also allows us to make predictions and decisions contingent on our inferences concerning the process. In order to identify "good" statistical models some principles are required on which the modeling procedures are based. In general, there are three requirements for a statistical model: plausibility, parsimony and goodness of fit. While the first one is not of statistical consideration, the latter two involve a tradeoff between optimal fit and parsimony, seeking the best model in terms of goodness of fit at the minimum cost of degrees of freedom.

The process of statistical analysis might be simplistically displayed as follows:



In Chapter 1, it has been assumed that the independent variables to be used in the model as well as the form in which they would be expressed were known. Moreover, the properties of the least squares estimators were based on the assumption that the model was correct. Most regression problems, however, require decisions about which variables must be included in the model, the form that the variables should take, and the functional form of the model. It is assumed that there is a set presumably including all relevant variables, from which a subset of variables is to be chosen for the regression equation. The candidate variables may include different

forms of the same basic variable, and the selection process may include constraints on which variables are to be included. For example, X may be forced into the model if X^2 is in the selected subset.

The problem of determining an appropriate equation based on a subset of the original set of variables contains three basic ingredients, namely, the computational technique used to provide the information for the analysis, the criterion used to analyze the variables and select a subset, if that is appropriate, and the estimation of the coefficients in the final equation. Typically, a procedure might embody all three ideas without clearly identifying them. Moreover, regression equations with fewer variables have the appeal of simplicity, as well as an economic advantage in terms of obtaining the necessary information to use the equations. In addition, there is a theoretical advantage of eliminating irrelevant variables and, in some cases, even variables that contain some predictive information about the response variable. The motivation to eliminate variables is tempered by the biases and loss of predictability that are introduced when relevant variables are eliminated. The objective is to reach a compromise where the final equation satisfies the purpose of the study.

The purpose of how the regression equation is to be used influences the manner in which the model is constructed. Simplistically, the regression might be used for one of the following purposes (or a combination of them):

- Providing a good description of the behavior of the response variable
- Prediction of future responses and estimation of mean responses
- Extrapolation, or prediction of responses outside the range of the data
- Estimation of parameters
- Control of a process by varying levels of input
- Developing realistic models of the process.

Each objective has different implications on how much emphasis is placed on eliminating variables from the model, on how important it is that the retained variables be causally related to the response variable, and on the amount of effort devoted to making the model realistic.

Of the uses of regression, prediction and estimation of mean responses are the most tolerant toward eliminating variables. At the same time, it is relatively unimportant whether the variables are causally related or the model is realistic. It is tacitly assumed that prediction and estimation are to be within the X-space of the data and that the system continues to operate as it did when the data were collected. Thus,

any variables that contain predictive information on the dependent variable, and for which information can be obtained at a reasonable cost, are useful variables. Of course, more faith could be placed in predictions and estimates based on established causal relationships, because of the protection such models provide against inadvertent extrapolations and unrecognized changes in the correlational structure of the system.

On the other hand, one should also be conservative in eliminating variables when estimation of parameters is the objective. This is to avoid the bias introduced when a relevant variable is dropped. There is an advantage in terms of reduced variance of the estimates if variables truly unrelated to the dependent variable are dropped.

By all means, the results from any variable selection procedure, and particularly those that are automated, need to be studied carefully to make sure the models suggested are consistent with the state of knowledge of the process being modeled. No variable selection procedure can substitute for the insight of the researcher.

To briefly state the implications of improper model selection, assume that the variable selection is not based on information from the current data. In addition, the correct model involves t independent variables, but only a subset of p variables (chosen randomly or on the basis of external information) is used in the regression equation. Let $\mathbf{X_p}$ and $\boldsymbol{\beta}_\mathbf{p}$ denote submatrices of **X** and **β** that relate to the p selected variables. $\widehat{\boldsymbol{\beta}_\mathbf{p}}$ denotes the least squares estimate of $\boldsymbol{\beta}_\mathbf{p}$ obtained from the p-variate subset model. Similarly, $\widehat{\mathrm{Y_{pi}}}$, $\widehat{\mathrm{Y_{pred,pi}}}$ and $\mathrm{MS(Res)_p}$ denote the estimated mean for the i-th observation, the prediction for the i-th observation and the mean squared residual, respectively, obtained from the p-variate subset model. It can be shown (Hocking, 1976) that:

- $MS(Res)_p$ is a positively biased estimate of σ^2 unless the true regression coefficients for all deleted variables are zero
- $\widehat{\boldsymbol{\beta}_\boldsymbol{p}}$ is a biased estimate of $\boldsymbol{\beta}_\boldsymbol{p}$ and $\widehat{Y_{pi}}$ is a biased estimate of $E(Y_i)$ unless the true regression coefficient for each deleted variable is zero or, in the case of $\widehat{\boldsymbol{\beta}_\boldsymbol{p}}$, each deleted variable is orthogonal to the *p* retained variables
- $\widehat{\boldsymbol{\beta}_\boldsymbol{p}}$, $\widehat{Y_{pi}}$, $\widehat{Y_{pred,pi}}$ are generally less variable than the corresponding statistics obtained from the *t*-variate model

- There are conditions under which the mean squared errors (variance plus squared bias) of $\widehat{\boldsymbol{\beta}_{\boldsymbol{p}}}$, $\widehat{Y_{p\iota}}$, $\widehat{Y_{pred,p\iota}}$ are smaller than the variances of the estimates obtained from the *t*-variate model.

Thus, a bias penalty is paid whenever relevant variables, those with $\beta_j \neq 0$, are omitted from the model. On the other hand, there is an advantage in terms of decreased variance for both estimation and prediction if variables are deleted from the model. Furthermore, there may be cases in which there is a gain in terms of mean squared error of estimation and prediction from omitting variables whose true regression coefficients are not zero.

These results provide motivation for selecting subsets of variables, but they do not apply directly to the usual case where variable selection is based on analysis of the current data. The general nature of these effects may be expected to persist, but selection of variables based on their performance in the sample data introduces another class of biases that confound these results. The process of searching through a large number of potential subset models for the one that best fits the data capitalizes on the random variation in the sample to "overfit" the data. That is to say, the chosen subset model can be expected to show a higher degree of agreement with the sample data than the true equation would show with the population data. Another problem of sample-based selection is that relative importance of variables as manifested in the sample will not necessarily reflect relative importance in the population. The best subset in the sample, by whatever criterion, need not be the best subset in the population. Important variables in the population may appear unimportant in the sample and consequently be omitted from the model, and vice versa.

Many criteria for choice of subset size have been proposed, based on the principle of parsimony which suggests selecting a model with small residual sum of squares with as few parameters as possible. Most of the criteria are monotone functions of the residual sum of squares for a given subset size and, consequently, give identical rankings of the subset models within each subset size. However, the choice of criteria may lead to different choices of subset size, and they may give different impressions of the magnitude of the differences among subset models.

Moreover, the reduction in the predictive accuracy of a sample-derived regression model when applied to new data, what is called validity shrinkage, can be managed in part by a number of variable selection methods. Methods such as forward and

backward stepwise regression and all subsets regression attempt to maximize the correlation between $\widehat{Y}$ and Y, while minimizing the number of predictors. The fewer predictors that are used in a regression model applied to future cases, the less validity shrinkage tends to be. These variable selection methods have documented problems, however, in that they tend to overfit the data. But they can be useful in some contexts if their limits are understood and it is recognized that none of them is likely to be selecting the best or correct model by some objective standard.

Some of the most commonly used criteria for evaluating and comparing subset regression models are presented briefly.

2.2 Coefficient of determination

The total variability in the response variable Y can be partitioned into variability due to change in expectations and variability due to random errors. The sum of squares decomposition is:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

that is, the corrected sum of squares of the observations, SST, which measures the total variability in the observations, equals to the amount of variability in the observations accounted for by the regression line, SSR, plus the residual variation left unexplained by the regression line, SSE. This is the fundamental analysis of variance identity for a regression model, symbolically written as:

$$SST = SSR + SSE.$$

The coefficient of determination R^2 is based on this analysis and equals to the proportion of the total (corrected) sum of squares of the dependent variable explained by the independent variables in the model. That is:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Because $0 \leq \mathrm{SSE} \leq \mathrm{SST}$, it follows that $0 \leq \mathrm{R}^2 \leq 1$. Values of R^2 that are close to 1 imply that most of the variability in the response variable is explained by the regression model. Naturally, the interpretation of R^2 reflects its significant role in model fitting.

Note that, as new explanatory variables are added into the model, the coefficient of determination increases in value. This happens even if the added variable(s) have no relation to the response variable, indicating that a model with increased R^2 is not a priori superior to the old one. In particular, unless the SSE in the new model is reduced by an amount equal to the original error mean square, the new model will have a larger error mean square than the old one because of the loss of one degree of freedom for error. Thus, the new model will actually be worse than the old one.

Moreover, the magnitude of R^2 also depends on the range of variability in the regressor variable. Generally R^2 will increase as the spread of the Xs increases, and decrease as the spread of the Xs decreases, provided the assumed model form is correct.

Further on misconceptions about R^2, it should be noted that it does not measure the appropriateness of the linear model, for R^2 will often be large even though Y and X are nonlinearly related.

So, R^2 cannot be used directly for model comparison. Furthermore, the coefficient of determination is not comparable for models where the functional form of the response variable is different.

2.3 Residual mean square

The residual mean square MS(Res) is an estimate of σ^2 if the model contains all relevant independent variables. If relevant independent variables have been omitted, the MS(Res) is biased upward. On the other hand, including an unimportant independent variable will have little impact on the MS(Res). Thus, the expected behavior is for it to decrease toward σ^2 as important independent variables are added to the model and to fluctuate around σ^2 once all relevant variables have been included. The aforementioned describe the expected behavior of MS(Res) when the variable selection is not based on sample data. As will be established below, this criterion is essentially equivalent with the ${R_{adj}}^2$.

2.4 R^2 adjusted

The adjusted R^2, which is labeled as ${R_{adj}}^2$, is a rescaling of R^2 by degrees of freedom so that it involves a ratio of mean squares rather than sums of squares and thus taking into account the number of the variables included in the model:

$$ {R_{adj}}^2 = 1 - \frac{MS(Res)}{MS(Total)} = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST} < R^2. $$

The primary attractiveness of R_{adj}^2 is that it imposes a penalty for adding additional regressors to a model. If a regressor is added to the model, then SSR increases. As a consequence R_{adj}^2 can fluctuate when a new regressor is added to the model. Its value will tend to stabilize around some upper limit as variables are added. The simplest model with R_{adj}^2 near this upper limit can be chosen as the "best" model. Moreover, it is more comparable than R^2 over models involving different number of parameters. However, its interpretation is less clear, and it can even take negative values.

R_{adj}^2 is by definition closely related to MS(Res) and these two will lead to the same conclusions. In particular, the subset regression model that minimizes MS(Res) will also maximize R_{adj}^2.

2.5 Mallows' C_p statistic

Another technique/criterion to assess fit when models, whose number of parameters differs are being compared, is the C_p statistic (Mallows, 1973), which is an estimate of the standardized total MS(Res) for the current set of data:

$$C_p = \frac{SSE_p}{\widehat{\sigma^2}} - n + 2p,$$

where *p* is the number of parameters in the model and *n* is the sample size. When the model is correct, the residual sum of squares is an unbiased estimate of $(n-p)\sigma^2$ and in this case, C_p is approximately equal to *p*. When important independent variables have been omitted from the model, the residual sum of squares is an estimate of $(n-p)\sigma^2$ plus a positive quantity reflecting the contribution of the omitted variables; in this case, C_p is expected to be greater than *p*. The most preferable model (with, say, *p* parameters) will have a C_p value which tends to be close to or smaller than *p*. Note that this is the criterion defined by Hocking (1976) when the model is primarily intended for prediction (while for parameter estimation the criterion $C_p \leq 2p - t$ has been suggested). We may consider choosing the smallest model for which this is true (to reduce intercorrelation). The C_p plot presents C_p as a function of *p* for the better subset models and provides a convenient method of selecting the subset size and judging the competitor subsets. The "best" model is the one whose coordinates (C_p, p) fall nearest the line $C_p = p$ in the plot. A C_p value that is close to the number *p* of predictors indicates that the model produces relatively precise and unbiased

estimates, while a C_p value that is greater than the number *p* of predictors indicates that the model is biased and does not fit the data well. An advantage of C_p is that it can be used to select model size (a good model can be obtained that contains as few variables as possible). However, the C_p criterion suffers from limitations too, such as that the C_p approximation is only valid for large sample sizes.

2.6 The Akaike Information Criterion (AIC)

Akaike (Akaike, 1974) proposed an information criterion, based on maximizing the expected entropy of the model. Entropy is simply a measure of the expected information, in this case the Kullback-Leibler information measure. Essentially, the AIC is a penalized log-likelihood measure. Letting L be the likelihood function for a specific model, then AIC is:

$$AIC = -2\ln(L) + 2p,$$

where *p* is the number of the parameters in the model. In the case of OLS regression:

$$AIC = nln\left(\frac{SSE}{n}\right) + 2p.$$

The key insight to the AIC is similar to R_{adj}^2 and Mallows' C_p. As we add regressors to the model, SSE cannot increase. The issue becomes whether the decrease in SSE justifies the inclusion of the extra terms. A graph of AIC against *p* will, in general, show a minimum value, and the appropriate value of the subset size is determined by the value of *p* at which AIC attains its minimum value. That is, once again the preferred model is the one with the fewest parameters that still provides an adequate fit to the data. The AIC criterion is widely used, despite the fact that it has a tendency to select models with larger subset sizes than the true model.

2.7 Bayesian extension of AIC

There are several Bayesian extensions of the AIC, Schwartz (1978) and Sawa (1978) being two of the most popular ones. The Schwartz Bayesian Information Criterion (BIC) is defined as:

$$BIC = -2lnL + pln(n).$$

This criterion uses the multiplier ln(n) (instead of 2 in AIC), thus placing a greater penalty on adding regressors as the sample size increases. More precisely, this happens for sample size greater than 7, since $2p < pln(n), \text{for } n > 7$. For OLS regression, this criterion is:

$$BIC = nln\left(\frac{SSE}{n}\right) + pln(n).$$

Again, the appropriate value of the subset size is determined by the value of *p* at which BIC attains its minimum value.

It's worth noting that:

- In the special case of linear regression AIC and BIC are not bounded (unlike R^2).
- AIC and BIC can be implied to statistical models without linearity.
- Both AIC and BIC are not relative measures as R^2 and therefore their magnitude offers no information.

2.8 Computational techniques for variable selection

Conceptually, the only way of ensuring that the best model for each subset size has been found is to compute all possible subset regressions. This is feasible when the total number of variables is relatively small, but rapidly becomes a major computing problem even for moderate numbers of independent variables. A lot of attention has been focused on identifying the best subsets within each subset size without computing all possible subsets. These methods utilize the basic least squares property that the residual sums of squares cannot decrease when a variable is dropped from a model. Thus, comparison of residual sums of squares from different subset models is used to eliminate the need to compute other subsets.

These methods are referred to as stepwise regression methods and identify good (although not necessarily the best) subset models, with considerably less computing than is required for all possible regressions. The subset models are identified sequentially by adding or deleting, depending on the method, the one variable that has the greatest impact on the residual sum of squares. These stepwise methods are not guaranteed to find the "best" subset for each subset size, and the results produced by different methods may not agree with each other.

The aforementioned methods can be classified into three broad categories, as follows.

2.9 Forward selection

Forward selection begins with the assumption that there are no regressors in the model other than the intercept. The procedure is to find an optimal subset by

inserting regressors into the model one at a time. The first regressor selected for entry into the equation is the one that has the largest simple correlation with the response variable Y. This happens to be the regressor that will produce the largest value of the F statistic for testing significance of regression. This regressor is entered if the F statistic exceeds a preselected F value, say $\mathrm{F_{in}}$ (or F to enter). Equivalently this can be performed using the p-values. The second regressor chosen for entry is the one that now has the largest correlation with Y after adjusting for the effect of the first regressor entered on Y. Thus, at each successive step, the variable in the subset of variables not already in the model that causes the largest decrease in the residual sum of squares is added to the subset. Without a termination rule, forward selection continues until all variables are in the model. The $\mathrm{F_{in}}$ criterion should be viewed as a stopping rule rather than a classical test of significance.
As an algorithm representation we have:

1. Start with a model including just the intercept
2. For each of the predictors not in the model, calculate their F statistics as if they are added to the model
3. For those F statistics that are greater than F_{in}, choose the one with the biggest value and include the corresponding predictor in the model
4. Iterate until no new predictors can be added.

2.10 Backward elimination

Forward selection has a serious drawback: each time a new regressor is added in the model, one or more already included may become non-significant. Backward elimination attempts to find a good model by working in the opposite direction. It chooses the subset models by starting with the full model (that is, the model which includes all candidate/available regressors) and then eliminating at each step the one variable whose deletion will cause the residual sum of squares to increase the least. This will be the variable in the current subset model that has the smallest partial sum of squares. At each step, this is performed by calculation of the partial F statistic for each regressor as if it were the last variable to enter the model. The smallest of these F statistics is compared with a preselected value, say $\mathrm{F_{out}}$ and if the smallest partial F value is less than $\mathrm{F_{out}}$, that regressor is removed from the model. This procedure continues by successively re-fitting reduced models and applying the same rule, until all remaining variables are statistically significant. Without a termination rule, backward elimination continues until the subset model contains only one variable. The $\mathrm{F_{out}}$ criterion should be viewed as a stopping rule rather than a classical test of significance.

Backward elimination is often a very good variable selection procedure. It is particularly favored by analysts who like to see the effect of including all the candidate regressors, just so that nothing "obvious " will be missed.
As an algorithm representation we have:

1. Start with a model including all the predictors
2. For each predictor in the model, calculate their F statistics as if each one was the last one added in the model
3. For those F statistics that are smaller than F_{out}, choose the one with the smallest value and remove the corresponding predictor from the model
4. Iterate until no new predictors can be removed.

2.11 Stepwise regression

Neither forward selection nor backward elimination takes into account the effect that the addition or deletion of a variable can have on the contributions of other variables to the model. A variable added early to the model in forward selection can become unimportant after addition of other variables (because of the relationships between it and the rest ones included), or variables previously dropped in backward elimination can become important after other variables are dropped from the model (for the very same reason). The variable selection method commonly labeled stepwise regression is a forward selection process that rechecks at each step the importance of all previously included variables. If the partial sums of squares for any previously included variables do not meet a minimum criterion to stay in the model, the selection procedure changes to backward elimination and variables are dropped one at a time until all remaining variables meet the minimum criterion. Then, forward selection resumes.

The stopping rule for stepwise selection of variables uses both the forward and backward elimination criteria. The variable selection process terminates when all variables in the model meet the criterion to stay and no variables outside the model meet the criterion to enter (except, perhaps, for the variable that was just eliminated). The criterion for a variable to enter the model need not be the same as the criterion for the variable to stay. There is some advantage in using a more relaxed criterion for entry (that is $\mathrm{F_{in}} > \mathrm{F_{out}}$) to force the selection process to consider a larger number of subsets of variables.

Stepwise selection of variables requires more computing than forward or backward selection but has an advantage in terms of the number of potential subset models checked before the model for each subset size is decided. It is reasonable to expect

stepwise selection to have a greater chance of choosing the best subsets in the sample data, but selection of the best subset for each subset size is again not guaranteed.

2.12 General comments on stepwise-type procedures

The stepwise regression algorithms have been criticized on various grounds, the most common being that none of the procedures generally guarantees that the best subset regression model of any size will be identified. Furthermore, since all the stepwise-type procedures terminate with one final equation, it may be concluded that this is the optimal, in some sense, model. Part of the problem is that it is likely, not that there is one best subset model, but that there are several equally good ones.

Moreover, the order in which the regressors enter or leave the model does not necessarily imply an order of importance to the regressors. It is not unusual to find that a regressor inserted into the model early in the procedure becomes negligible at a subsequent step. This is in fact a general problem with the forward selection procedure, since, once a regressor has been added, it cannot be removed at a later step.

Note again that forward selection, backward elimination, and stepwise regression do not necessarily lead to the same choice of final model. The intercorrelation between the regressors affects the order of entry and removal.

With any variable selection method, it is important to keep in mind that model selection cannot be separated from the underlying purpose of the investigation. Variable selection tends to amplify the statistical significance of the variables that stay in the model. Variables that are dropped can still be correlated with the response. It would be wrong to consider that these variables are unrelated to the response; it's just that they provide no additional explanatory effect beyond those variables already included in the model.

Finally, stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes. To give a simple (and extreme) example, consider the simple regression with just one predictor variable and suppose that the slope for this predictor is not quite statistically significant. We might not have enough evidence to say that it is related to Y but it still might be better to use it for predictive purposes.

Finally, it's worth noting that the various stepwise procedures, while originated for regression models, they can also be applied in settings that extend the basic linear model, such as GLMs, having the residual sum of squares replaced by deviance or other relevant measures.

Chapter 3: Bayesian methods for model selection

3.1 Motivation for the Bayesian approach to model selection

Despite the plethora of classical methods available in order to select a model, there are certain reasons for the Bayesian methods to be favored. Among the most famous ones, is the fact that their conclusions are easier to understand. The direct probability interpretation of posterior model probabilities and the interpretation of Bayes factors support this claim.

Moreover, the Bayesian approach to model selection is consistent. This means that if one of the entertained models is actually the true model, then Bayesian model selection will, under mild conditions, guarantee its selection if enough data is observed. This is not always the case with use of most classical model selection tools. At the same time, since none of the candidate models may be the true one, this advantage is tempered.

Overfitting is a continual problem, since more complex models will always provide a somewhat better fit. In classical methods a lot of effort is dedicated to choose the best penalizing criterion, while the Bayesian procedures naturally penalize model complexity through the prior, and need no introduction of a penalty term. Simpler models will be favored over more complex ones when the data provides roughly comparable fits for the models.

Moreover, the Bayesian approach is conceptually the same, regardless of the number of models under consideration, while the classical framework distinguishes when two or more models are to be compared. Additionally, nested models, standard distributions or regular asymptotics are not a prerequisite.

At last, one of the most significant reasons is that the Bayesian approach can account for model uncertainty. While the classical framework will complete the model selection procedure and then base predictions on the assumption that the selected model is correct, in the Bayesian framework all models may be left in the analysis with prediction being performed using a weighted average of the predictive distributions from each model, the weights being determined from the posterior probabilities of each model.

3.2 The Bayesian linear regression

Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. Following the notation of Chapter 1 concerning linear regression, which is $Y = X\beta + \varepsilon$, and

assuming that the errors are normally distributed, it has been established that $Y \sim N(X\beta, \sigma^2 I)$, while the likelihood function for the unknown parameters β and σ^2 is:

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

Under the MLE approach, estimators for β and σ^2 are obtained by maximizing this likelihood function. Under the Bayesian approach, instead of maximizing the likelihood function alone, we assume prior distributions for the parameters and use Bayes' theorem to obtain their joint posterior distribution:

$$posterior \propto prior \times likelihood$$

Thus, we first need to decide on our prior. The prior distribution reflects the information about the (unknown) parameters and combined with the probability distribution of the data yields the posterior distribution. Therefore, information that the prior should capture, as well as the properties of the resulting posterior distribution must be taken into account. We will focus on prior specification, since this piece is uniquely Bayesian.

3.3 Prior distribution of θ

Computational difficulties arise in using Bayes' theorem when it is necessary to evaluate the normalizing constant in the denominator:

$$\int f(\theta) f(x|\theta) d\theta.$$

However, judicious choices of prior can lead to posterior calculations which require no integration. In these cases we can identify a prior distribution $f(\theta)$ for which the posterior distribution $f(\theta|x)$ is in the same family of distributions as the prior. Such priors are called conjugate priors. The richness of the conjugate family is great enough for a prior to be found that is close enough to the analyst's beliefs, providing a convenient and useful mathematical technique. In particular, it emerges that conjugates can be obtained for data models within the exponential family of distributions. That is:

$$f(x|\theta) = h(x)g(\theta)\exp\{t(x)c(\theta)\},$$

for functions h, g, t and c such that:

$$\int f(x|\theta)d\theta = g(\theta)\int h(x)\exp\{t(x)c(\theta)\}dx = 1.$$

To see why this stands, given a random sample $x = (x_1, x_2, \dots, x_n)$ from this general distribution, the likelihood for ϑ is:

$$f(x|\theta) = \prod_{i=1}^{n}\{h(x_i)\}g(\theta)^n\exp\{\sum_{i=1}^{n}t(x_i)c(\theta)\} \propto g(\theta)^n\exp\{\sum_{i=1}^{n}t(x_i)c(\theta)\}.$$

Thus if a prior of the form

$$f(\theta) \propto g(\theta)^d\exp\{bc(\theta)\},$$

is chosen, we obtain:

$$f(\theta|x) \propto f(\theta)f(x|\theta) \propto g(\theta)^d\exp\{bc(\theta)\} \times g(\theta)^n\exp\{\sum_{i=1}^{n}t(x_i)c(\theta)\} =$$

$$g(\theta)^{n+d}\exp\{[b + \sum_{i=1}^{n}t(x_i)]c(\theta)\} = g(\theta)^D\exp\{Bc(\theta)\},$$

where $D = n + d$ and $B = b + \sum_{i=1}^{n}t(x_i)$. So, the result is a posterior in the same family as the prior, with its parameters modified.

The other essential point into formulating a prior distribution is the information it includes. If we have no prior information about ϑ, then a, so called, non-informative prior should be suitable. A uniform distribution could be an obvious choice, letting all possible outcomes of ϑ being equally probable. However, for non-compact parameter spaces such a flat prior ($f(\theta) = c$) is not a proper distribution, meaning it does not integrate to one. Nevertheless, it is considered to be acceptable using improper priors (to reflect vague knowledge) in the case where $\int f(x|\theta)d\theta = K < \infty$. In short, the posterior needs (and offers this way a sufficient condition) to be verified as a proper distribution.

Further on representations of ignorance, it is reasonable (and desirable) for a prior to be consistent across 1-1 parameter transformations. This can be accomplished with use of the Jeffreys' prior (Jeffreys, 1946), which is defined as:

$$J_{\Theta}(\theta) \propto |I(\theta)|^{1/2},$$

where the quantity $I(\theta)$ is the Fisher information:

$$I(\theta) = -E\left\{\frac{d^2 l(\theta)}{d\theta^2}\right\} = E\left\{\left(\frac{dl(\theta)}{d\theta}\right)^2\right\}$$

and $l(\theta) = logf(x|\theta)$, is the log likelihood.

Another substantial flexibility that the Bayesian framework offers, concerning prior elicitation, is the use of multimodal prior distribution, in order to express fluctuating prior beliefs. Let $\mathrm{f_1(\theta)}, \dots, \mathrm{f_k(\theta)}$ be conjugate distributions for θ, leading to posterior distributions $\mathrm{f_1(\theta|x)}, \dots, \mathrm{f_k(\theta|x)}$ and consider the family of mixture distributions: $\mathrm{f(\theta) = \sum_{i=1}^{k} p_i f_i(\theta)}$, where $\mathrm{0 \le p_i \le 1, i = 1, \dots, k}$ and $\mathrm{\sum_{i=1}^{k} p_i = 1}$. Then,

$$f(\theta|x) \propto f(\theta)f(x|\theta) = \sum_{i=1}^{k} p_i f_i(\theta)\, f(x|\theta) = \sum_{i=1}^{k} p_i f_i(x)\, f(\theta|x) = \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{p_i^* f_i(\theta|x)}$$

where $\mathrm{p_i^* \propto p_i f_i(x)}$, yielding that the posterior is in the same mixture-family (though with mixture proportions in the posterior generally different from those in the prior).

The aforementioned scheme of obtaining the posterior distribution by updating our beliefs on the context of observed data, can be accomplished either using the entire information the data provide at once, or sequentially. In the latter case each time we update our beliefs we use interim posterior information as our next prior, in a sequential updating mode that concludes to the very same results. To see why, consider the simple case where the independent variables X_1, X_2 have density $f(x|\theta)$. Suppose we observe x_1 and update with use of Bayes' theorem to $f(\theta|x_1) \propto f(\theta)f(x_1|\theta)$. This is considered to be our new prior before observing x_2 and we then have: $f(\theta|x_1, x_2) \propto f(\theta)f(x_1|\theta)f(x_2|\theta) = f(\theta)f(x_1, x_2|\theta)$, which equals to the result obtained by updating on the basis of the entire information (x_1, x_2) at once. Of course, as stated, the data must be conditionally independent.

3.4 The Bayesian linear model

As stated, for the Bayesian analysis we first need to specify priors for the unknown regression parameters β and σ^2. Two of the most widely used cases of prior setup will be discussed.

3.4.1 Non-informative priors

First we will use non-informative priors, the case where Bayesian analysis resembles most the classical approach. That is, selecting flat priors on β and σ^2, or equivalently $f(\beta) \propto 1, f(\sigma^2) \propto \frac{1}{\sigma^2} \Leftrightarrow f(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$. This is an improper prior, however yielding a valid posterior.

The joint posterior distribution of (β, σ^2) is given by:

$$f(\beta, \sigma^2|Y) \propto f(Y|\beta, \sigma^2) f(\beta) f(\sigma^2) \equiv L(\beta, \sigma^2) f(\beta, \sigma^2),$$

where

$$L(\beta, \sigma^2) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right\},$$

thus yielding:

$$f(\beta, \sigma^2|Y) \propto (\sigma^2)^{-\frac{n}{2}-1} exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right\}.$$

The conditional posterior distribution of β given σ^2, $f(\beta|\sigma^2, Y)$ is simply $f(\beta, \sigma^2|Y)$, seen as a function of β. Thus:

$$\beta|\sigma^2, Y \sim N((\mathrm{X'X})^{-1}\mathrm{X'Y}, \sigma^2(\mathrm{X'X})^{-1}) \equiv N\left(\hat{\beta}, \mathrm{Var}(\hat{\beta})\right).$$

That is, the posterior distribution of $\beta|\sigma^2, Y$ has mean $\hat{\beta} = (\mathrm{X'X})^{-1}\mathrm{X'Y}$ which coincides with the MLE and OLS estimator of β.

The above conditional posterior distribution of β would have been the desired distribution had σ^2 been known. Since this is not the case, we have to find the marginal posterior distribution of β, by integrating out σ^2 as:

$$f(\beta|Y) = \int f(\beta, \sigma^2|Y) f(\sigma^2|Y) d\sigma^2.$$

Thus, specification of the marginal posterior distribution of σ^2, $f(\sigma^2|Y)$ arises. Assuming β is fixed, and then the conjugate prior for σ^2 is an inverse Gamma

distribution and so we get for the marginal posterior distribution of σ^2 another inverse Gamma:

$$f(\sigma^2|Y) \propto \frac{1}{(\sigma^2)^{\frac{n-k}{2}+1}} \exp\left\{-\frac{(n-k)s^2}{2\sigma^2}\right\} \sim IG\left(\frac{n-k}{2}, \frac{(n-k)s^2}{2}\right),$$

where $s^2 = \hat{\sigma}^2 = \frac{1}{n-k}\left(Y - X\hat{\beta}\right)'(Y - X\hat{\beta})$, the unbiased estimate of σ^2 in the linear regression model.

Revisiting now the marginal posterior distribution of β, by integrating out σ^2 we get that $f(\beta|Y)$ is a non-central multivariate *t* distribution with *n-k* degrees of freedom and non-centrality parameter $\hat{\beta}$.

The above distribution is quite complicated. In order to carry out a non-informative Bayesian analysis, we use a simpler sampling based mechanism. For each $i = 1, \dots, M$ we first draw $\sigma^2{}_{(i)} \sim [\sigma^2|y]$ which is inverse Gamma, followed by $\beta_{(i)} \sim \mathrm{N}((\mathrm{X'X})^{-1}\mathrm{X'Y}, \sigma^2{}_{(i)}(\mathrm{X'X})^{-1})$. The resulting samples $\left(\beta_{(i)}, \sigma^2{}_{(i)}\right), i = 1, \dots, M$ are precisely samples from the joint marginal posterior distribution $f(\beta, \sigma^2|Y)$. Automatically the samples $\left(\beta_{(i)}\right), i = 1, \dots, M$ are samples from marginal posterior distributions $f(\beta|Y)$, while the samples $\left(\sigma^2{}_{(i)}\right), i = 1, \dots, M$ are from the marginal posterior $f(\sigma^2|Y)$.

Next we want to apply our regression analysis to a new set of data, where we have observed the new covariate matrix $\tilde{X}$ and we wish to predict the corresponding outcome $\tilde{y}$. If β and σ^2 were known, then $\tilde{y}$ would have a $N(\tilde{X}\beta, \sigma^2 I)$ distribution. In reality, where parameters are unknown, all predictors for the data must follow from the posterior predictive distribution:

$$f(\tilde{y}|y) = \iint f(\tilde{y}|\beta, \sigma^2) f(\beta, \sigma^2|Y) d\beta d\sigma^2.$$

Therefore, predictions are carried out by sampling from the posterior predictive distribution: for each posterior we draw $\left(\beta_{(i)}, \sigma^2{}_{(i)}\right), i = 1, \dots, M$, and then draw $\tilde{y}_{(i)}$ from $N(\tilde{X}\beta_{(i)}, \sigma^2{}_{(i)} I)$. The resulting sample $\tilde{y}_{(i)}, i = 1, \dots, M$ represents the predictive distribution. The theoretical mean and variance, conditional upon σ^2 is:

$$E(\tilde{y}|\sigma^2, y) = E[E(\tilde{y}|\beta, \sigma^2, y)|\sigma^2, y] = E[\tilde{X}\hat{\beta}|\beta, \sigma^2] = \tilde{X}\hat{\beta},$$

$$var(\tilde{y}|\sigma^2, y) = E[var(\tilde{y}|\beta, \sigma^2, y)|\sigma^2, y] + var[E = \tilde{y}|\beta, \sigma^2, y)|\sigma^2, y] =$$
$$E(\sigma^2 I) + var[\tilde{X}\hat{\beta}|\sigma^2, y] = \left(I + \tilde{X}(X'X)^{-1}\tilde{X}'\right)\sigma^2.$$

Thus, conditional on σ^2, the posterior predictive variance has two components: $\sigma^2 I$, representing sampling variation and $\tilde{X}(X'X)^{-1}\tilde{X}'\sigma^2$, due to uncertainty about β.

To obtain the theoretical unconditional predictive distribution $f(\tilde{y}|y)$ we marginalize over σ^2 and conclude to a multivariate t distribution, $t_{n-k}(\tilde{X}\hat{\beta}, s^2(I + \tilde{X}(X'X)^{-1}\tilde{X}')$.

3.4.2 Normal-inverse Gamma prior

One of the most commonly used priors is the Normal-inverse Gamma distribution, which is the conjugate prior for the normal linear model. To see why, we first use the fact that:

$$(Y - X\beta)'(Y - X\beta) = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) + \left(\beta - \hat{\beta}\right)'(X'X)\left(\beta - \hat{\beta}\right),$$

to rewrite the likelihood as:

$$L(\beta, \sigma^2) \propto (\sigma^2)^{-u/2} exp\left\{-\frac{us^2}{2\sigma^2}\right\}(\sigma^2)^{-\frac{n-u}{2}} exp\left\{-\frac{1}{2\sigma^2}\left(\beta - \hat{\beta}\right)'(X'X)(\beta - \hat{\beta})\right\},$$

where $u = n - k$ and $s^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{u}$.

Recognizing the kernels, this notation suggests for the prior a form of:

$$f(\beta, \sigma^2) = f(\sigma^2) f(\beta|\sigma^2),$$

where $f(\sigma^2)$ is an inverse Gamma distribution:

$$f(\sigma^2) \propto (\sigma^2)^{-\frac{\underline{u}}{2}-1} exp\left\{-\frac{\underline{us}^2}{2\sigma^2}\right\} \equiv IG\left(\frac{\underline{u}}{2}, \frac{\underline{us}^2}{2}\right).$$

The conditional prior density $f(\beta|\sigma^2)$ is a multinomial Normal distribution:

$$f(\beta|\sigma^2) \propto (\sigma^2)^{-\frac{k}{2}} exp\left\{-\frac{1}{2\sigma^2}\left(\beta - \hat{\beta}\right)'(X'X)\left(\beta - \hat{\beta}\right)\right\} \equiv N_k\left(\underline{\beta}, \sigma^2\underline{V}\right),$$

where $\underline{\beta}$ is a k-vector containing the prior means for the k regression coefficients $\beta_1, \dots, \beta_k$and $\underline{V}$ is a $k \times k$ positive definite prior covariate matrix. Note that if $\underline{V}$ is a priori large, expressing ignorance, this leads to a non-informative prior.

Based on the above the prior is:

$$f(\beta, \sigma^2) = f(\sigma^2) f(\beta|\sigma^2) = N_k\left(\underline{\beta}, \sigma^2 \underline{V}\right) \times IG\left(\frac{\underline{u}}{2}, \frac{\underline{us}^2}{2}\right) \propto$$
$$(\sigma^2)^{-\frac{k}{2}} exp\left\{-\frac{1}{2\sigma^2}\left(\beta - \hat{\beta}\right)'(X'X)\left(\beta - \hat{\beta}\right)\right\}(\sigma^2)^{-\frac{\underline{u}}{2}-1} exp\left\{-\frac{\underline{us}^2}{2\sigma^2}\right\} \propto$$
$$(\sigma^2)^{-\frac{\underline{u}}{2}-\frac{k}{2}-1} exp\left\{-\frac{1}{2\sigma^2}[\underline{us}^2 + \left(\beta - \hat{\beta}\right)'(X'X)\left(\beta - \hat{\beta}\right)]\right\} \equiv NIG\left(\underline{\beta}, \sigma^2 \underline{V}, \frac{\underline{u}}{2}, \frac{\underline{us}^2}{2}\right).$$

The *NIG* probability distribution is a joint probability distribution of a vector β and a scalar σ^2. Upon multiplication with the likelihood, we get the conjugate posterior distribution of $\beta, \sigma^2|y$ which is:

$$f(\beta, \sigma^2|y) \propto (\sigma^2)^{-\frac{\underline{u}}{2}-\frac{n+k}{2}-1} exp\left\{-\frac{1}{\sigma^2}\left[b^* + \frac{1}{2}(\beta - \mu^*)'V^{*-1}(\beta - \mu^*)\right]\right\}$$
$$\equiv NIG(\mu^*, V^*, \alpha^*, b^*),$$

where

$$\mu^* = \left(\sigma^2 \underline{V}^{-1} + X'X\right)^{-1}\left(\sigma^2 \underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right),$$
$$V^* = (\underline{V}^{-1} + X'X)^{-1},$$
$$\alpha^* = \frac{\underline{u} + n}{2},$$
$$b^* = \frac{\underline{us}^2}{2} + \frac{1}{2}\left[\underline{\beta}'\sigma^2 \underline{V}^{-1}\underline{\beta} + Y'Y - \mu^{*\prime V^{*-1}\mu^*}\right].$$

Note that the marginal posterior distribution of σ^2 is immediately recognized to be an $IG(\alpha^*, b^*)$.

The marginal posterior distribution of β is obtained by integrating out σ^2 from the *NIG* joint posterior:

$$f(\beta|y) = \int f(\beta, \sigma^2|y) d\sigma^2 = \int NIG(\mu^*, V^*, \alpha^*, b^*) d\sigma^2 \propto$$

$$\int \left(\frac{1}{\sigma^2}\right)^{a^*+1} \exp\left\{-\frac{1}{\sigma^2}\left[b^* + \frac{1}{2}(\beta-\mu^*)'V^{*-1}(\beta-\mu^*)\right]\right\} d\sigma^2 \propto$$
$$[1 + \frac{(\beta-\mu^*)'V^{*-1}(\beta-\mu^*)}{2b^*}]^{-(a^*+\frac{\kappa}{2})}.$$

This is a multivariate *t* density:

$$t_{\nu^*}(\mu^*, \Sigma^*) = \frac{\Gamma(\frac{\nu^*+k}{2})}{\Gamma(\frac{\nu^*}{2})\pi^{k/2}|\nu^*\Sigma^*|^{1/2}}[1 + \frac{(\beta-\mu^*)'\Sigma^{*-1}(\beta-\mu^*)}{\nu^*}]^{-\frac{\nu^*+k}{2}}$$

where $\nu^* = 2\alpha^*$ and $\Sigma^* = (\frac{b^*}{\alpha^*})V^*$.

Thus we have:

$$E(\beta|y) = \mu^*$$
$$var(\beta|y) = \frac{\nu^*}{\nu^*-2}\Sigma^* = \frac{b^*}{a^*-1}V^*.$$

Next we want to apply our regression analysis to a new set of data, where we have observed the new covariate matrix $\tilde{X}$ and we wish to predict the corresponding outcome $\tilde{y}$. Once again, if β and σ^2 were known, then $\tilde{y}$ would have a $N(\tilde{X}\beta, \sigma^2 I)$ distribution and would be independent of *y*. However, these parameters are unknown and summarized through their posterior samples. Therefore, all predictions for the data must follow from the posterior predictive distribution:

$$f(\tilde{y}|y) = \iint f(\tilde{y}|\beta,\sigma^2) f(\beta,\sigma^2|Y) d\beta d\sigma^2 =$$
$$= \iint N(\tilde{X}\beta, \sigma^2 I) \times NIG(\mu^*, V^*, \alpha^*, b^*) d\beta d\sigma^2 = t_{2a^*}\left(\tilde{X}\mu^*, \frac{b^*}{\alpha^*}(I + \tilde{X}V^*\tilde{X}')\right).$$

There are two sources of uncertainty in the posterior predictive distribution: the fundamental source of variability in the model due to σ^2, unaccounted for by $\tilde{X}\beta$ and the posterior uncertainty in β and σ^2 as a result of their estimation from a finite sample *y*. As the sample size gets larger ($n \to \infty$) the variance due to posterior uncertainty disappears, but the predictive uncertainty remains.

Once again, we can use a simple sampling based mechanism. For each posterior draw $(\beta_{(i)}, \sigma^2{}_{(i)}), i = 1, \dots, M$ we draw $\tilde{y}_{(i)}$ from $N(\tilde{X}\beta_{(i)}, \sigma^2{}_{(i)}I)$. The resulting sample $(\tilde{y}_{(i)}), i = 1, \dots, M$ represents the predictive distribution.

A great variety of other priors have been suggested and extensively used. Briefly stated, two of them are:

Zellner's *g*-prior (Zellner, 1986) where $\beta \sim \mathrm{N}(0, \frac{\sigma^2}{\mathrm{g}}(\mathrm{X}'\mathrm{X})^{-1})$. This prior is proper assuming *X* is full rank and yields a posterior mean of $\frac{1}{1+\mathrm{g}}\hat{\beta}_{\mathrm{OLS}}$, shrinking the least estimate towards zero and controlling the amount of shrinkage via *g*. That is, we assume a conservative prior mean of zero for the coefficients to reflect that not much is known about them, and that their variance-covariance structure is broadly in line with that of the data. The hyperparameter *g* embodies how certain we are that coefficients are indeed zero: a small g means few prior coefficient variance and therefore implies that we are quite certain that the coefficients are indeed zero.

The double exponential or Bayesian LASSO prior, which favors settings where there are many β_{j} near zero and a few large β_{j} (due to the distribution's thicker tales).

3.5 The (Bayesian) predictive distribution

Suppose we have past observations $(\mathrm{X}_1, \dots, \mathrm{X}_{\mathrm{n}})$ of a variable with likelihood function $\mathrm{f}(\mathrm{x}|\theta)$ and we wish to make inferences about the distribution of a future value of a random variable Y from this same model. With a prior distribution $\mathrm{f}(\theta)$, Bayes' theorem leads to a posterior distribution $\mathrm{f}(\theta|\mathrm{x})$ and we want to specify the predictive density function of y given x.

Supposing ϑ to be known, we assume that *Y* is independent of *x*. The joint density of *y, x* and ϑ is:

$$f(y, x, \theta) = \mathrm{f}(\mathrm{y}, x|\theta)\mathrm{f}(\theta) = \mathrm{f}(\mathrm{y}|\theta)\mathrm{f}(x|\theta)f(\theta).$$

Then we have:

$$\mathrm{f}(\mathrm{y}, \theta|x) = \frac{f(y, x, \theta)}{f(x)} = \frac{\mathrm{f}(\mathrm{y}|\theta)\mathrm{f}(x|\theta)f(\theta)}{f(x)} = \mathrm{f}(\mathrm{y}|\theta)\mathrm{f}(\theta|x),$$

and integrating out ϑ we get the predictive density function of y given x:

$$\mathrm{f}(\mathrm{y}|\mathrm{x}) = \int \mathrm{f}(\mathrm{y}|\theta)\mathrm{f}(\theta|\mathrm{x})\mathrm{d}\theta.$$

Thus the predictive density, evaluated at a particular value of y, is the integral of the likelihood of y times the posterior. The result can also be written as the expectation of the predictive density with respect to the posterior distribution of θ:

$$f(y|x) = E[f(y|\theta)|x].$$

Under the classical approach we would obtain a maximum likelihood estimate $\hat{\theta}$ of θ, plug it into the likelihood and base inference on the distribution $f(y|\hat{\theta})$, the estimative distribution. This makes no allowance for the variability incurred as a result of estimating ϑ, giving a false sense of precision, while the predictive density $f(y|x)$ is more variable by averaging across the posterior distribution for ϑ.

3.6 Model evidence

The marginal likelihood or evidence $f(x)$ of a given model $f(x|\theta)$ is the marginal distribution of the data under that model. It is obtained by integrating the product of the likelihood times a prior distribution $f(\theta)$ on the model parameters ϑ over ϑ:

$$f(x) = \int f(x|\theta)f(\theta)d\theta.$$

That is, $f(x)$ is the normalizing constant of the posterior distribution of ϑ, given by:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}$$

and an analytic solution is available whenever conjugate priors are used. This means that there exist formulas for posterior mean and variance, marginal likelihood, predictive. Equivalently, the marginal likelihood is defined as the expectation of the likelihood with respect to the prior distribution $f(\theta)$. In order to be able to evaluate the integral(s) involved in the calculation of $f(x)$ we need to choose carefully a proper (preferably conjugate) prior on ϑ.

Marginal likelihoods play an important role in Bayesian model comparison. While the model is defined by the likelihood function $L(\beta, \sigma^2)$ and the parameters' prior distribution $f(\beta, \sigma^2)$, the model evidence captures in a single number how well such a model explains the data. Under the Bayesian linear regression analysis, the model evidence can be used to compare competing linear models, which may differ in the number and values of the predictor variables, as well as in their priors on the model

parameters. Model complexity is already taken into account by the model evidence, since it marginalizes out the parameters by integrating $f(Y, \beta, \sigma^2)$ over all possible values of β, σ^2.

The model evidence, or marginal likelihood, $f(y|m)$ for a given model m is the probability of the data given the model m and can be expressed as:

$$f(y|m) = \int L(\beta, \sigma^2) f(\beta, \sigma^2) d\beta d\sigma^2.$$

3.7 Bayesian inference and hypothesis testing

The posterior distribution derived from Bayes' theorem is itself the inference, since it provides a complete description of the unknown parameter ϑ. In that sense, there is no real meaning into constructing a confidence interval, similarly to classical statistics. However, since point estimates give no measure of accuracy, credibility intervals come into play, having a direct interpretation which derives from the fact that parameters are regarded as random. Thus, there is a probability of 1-a, based on the posterior distribution, that ϑ lies in a region $C_\alpha(x)$, which is called the $100(1-a)\%$ credible region for ϑ, and for which:

$$\int_{C_\alpha(x)} f(\theta|x) d\theta = 1 - \alpha.$$

The credible intervals are not uniquely identified, as any region with probability 1-a will suffice. For a fixed value of a, we additionally seek for the shortest interval possible, which has a form of $C_\alpha(x) = \{\theta: f(\theta|x) \geq \gamma\}$, where γ is chosen to ensure that $\int_{C_\alpha(x)} f(\theta|x) d\theta = 1 - \alpha$. Such regions are called highest posterior density regions and in general have to be found numerically.

Hypothesis tests serve as a means of choosing between two different hypotheses. That is $H_0: \theta \in \Omega_0 \; vs \; H_1: \theta \in \Omega_1$, or in the case of simple point sets $H_0: \theta = \theta_0 \; vs \; H_1: \theta = \theta_1$. The classical approach makes use of the likelihood ratio $\lambda = \frac{f(x|\theta_1)}{f(x|\theta_0)}$ which takes large values when observed data is more likely to have occurred if θ_1 is the true value of ϑ (compared with θ_0). In the Bayesian framework, the posterior probabilities of θ_1 and θ_0 can be computed and then base the test on the relative posterior probabilities of the hypothesized values, using the posterior odds:

$$\lambda_B = \frac{f(\theta_1|x)}{f(\theta_0|x)} = \frac{f(\theta_1)f(x|\theta_1)}{f(\theta_0)f(x|\theta_0)}.$$

The concept of the test is the same, since large values of λ_B would favor H_1. Notable is the fact that there is no requirement to calculate normalizing factors, since the same factor would appear on the numerator and the denominator.

The above equation can be viewed as the product of the prior odds times the likelihood ratio, yielding the posterior odds. In this context, the likelihood ratio is termed the Bayes factor:

$$BF = \frac{f(x|\theta_1)}{f(x|\theta_0)} = \frac{f(\theta_1|x)/f(\theta_0|x)}{f(\theta_1)/f(\theta_0)}.$$

Thus we have the following scheme:

$$Posterior\ odds = Bayes\ factor \times Prior\ odds$$

The Bayes factor is a measure of the weight of information in the observed data in favor of H_1 over H_0(or the odds provided by the data for H_1 versus H_0). If it is sufficiently large, it will overcome any prior preference for H_0 so that our posterior preference may be for H_1.

Generalization of the above testing approach for more than two hypotheses is straightforward.

3.8 Bayesian model comparison

The framework of Bayesian model comparison evaluates probabilistic models based on the marginal likelihood, or the probability they assign a dataset with all the parameters marginalized out. Essentially, the Bayesian model selection rule consists of choosing the model which is a posteriori most probable for the data, as compared to another or various other alternative models.

Consider a number of competing models $M_1, \dots, M_k$ parameterized respectively by $\theta_1, \dots, \theta_k$ for an observed data set. In the presence of uncertainty about the correct model, Bayesian inference involves the evaluation of the posterior probability $f(M_j|x)$ of each model $M_j, j = 1, \dots, k$ as well as the evaluation of the posterior distribution $f(\theta_j|x, M_j)$ of the parameters θ_j of model $M_j, j = 1, \dots, k$. After specifying prior model probabilities $f(M_j)$ for all competing models and carefully choosing prior distributions for the model specific parameters $f(\theta_j|M_j), j = 1, \dots, k$,

posterior inferences can be obtained. Specifically, the posterior probability of model M_j is calculated using Bayes' theorem as:

$$f(M_j|x) = \frac{f(M_j)f(x|M_j)}{\sum_{i=1}^{k} f(M_j)f(x|M_i)}, j = 1, \dots, k,$$

where $f(x|M_j)$ is the marginal likelihood of the vector of observations *x* under model M_j. That is:

$$f(x|M_j) = \int L(x|\theta_j, M_j)\, f(\theta_j|M_j) d\theta_j.$$

It can be easily seen that this marginal likelihood is just the likelihood function integrated over the specified prior distribution for that model, provided that the integration is feasible. These marginal densities $f(x|M_j)$ are in general difficult to calculate. However, if the prior specification is conjugate to the likelihood function, some of the model parameters can be integrated out of the posterior distribution analytically.

Moreover, the posterior distribution of the parameters θ_j of model $M_j, j = 1, \dots, k$, is also given by Bayes' theorem as:

$$f(\theta_j|x, M_j) = \frac{f(\theta_j|M_j)f(x|\theta_j, M_j)}{f(x|M_j)}, j = 1, \dots, k.$$

The Bayes factor is in this case a measure of the weight of information in the observed data in favor of model $\mathrm{M_j}$ over $\mathrm{M_i}$ (or the odds provided by the data for $\mathrm{M_j}$ versus $\mathrm{M_i}$) and is given by:

$$\mathrm{B_{ji}} = \frac{\int \mathrm{f_j(x|\theta_j) f_j(\theta_j) d\theta_j}}{\int \mathrm{f_i(x|\theta_i) f_i(\theta_i) d\theta_i}} = \frac{\mathrm{f_j(x|\theta_j)}}{\mathrm{f_i(x|\theta_i)}} = \frac{f(M_j|x)}{f(M_i|x)} \frac{f(M_i)}{f(M_j)}.$$

Alternatively, $\mathrm{B_{ji}}$ is called the weighted likelihood ratio of $\mathrm{M_j}$ to $\mathrm{M_i}$, with the priors being the weighting functions.

Thus in the case we have two competing models, which are regression models with different explanatory variables, we can use the posterior odds ratio which is:

$$\mathrm{PO_{ji}} = \mathrm{B_{ji}} \frac{f(M_j)}{f(M_i)} = \frac{\mathrm{f_j}(\mathrm{x}|\theta_\mathrm{j})}{\mathrm{f_i}(\mathrm{x}|\theta_\mathrm{i})} \frac{f(M_j)}{f(M_i)}$$

With a non-informative choice of prior chances for the models ($f(M_j) = f(M_i) = 1/2$) the posterior odds ratio equals to the Bayes factor.

Despite its popularity, Bayes factor is relevant only in limited circumstances, since it is required to choose one particular model and there must be a zero-one loss on that decision. That is, if we make the wrong decision it doesn't matter how far the choice is, which is in contrast to the way statisticians think about most problems.

3.9 Bayesian model averaging

Bayesian model averaging is based on probability calculus and naturally emanates from the Bayesian paradigm by treating the model index as unknown. Generally, one important and potentially dangerous consequence of neglecting model uncertainty is that we assign more precision to our inference than is warranted by the data, and this leads to overly confident decisions and predictions. In addition, our inference can be severely biased. Standard statistical practice ignores model uncertainty. In model selection, we typically select a single "best" model from a set of candidate models and then use this model for prediction. That is, we proceed as if the selected model had generated the data, ignoring with this approach the uncertainty in model selection. In line with probability theory, the standard Bayesian response to dealing with uncertainty is to average. When dealing with parameter uncertainty, this involves averaging over parameter values with the posterior distribution of that parameter in order to get the predictive distribution. Analogously, model uncertainty is also resolved through averaging, but this time over models with the (discrete) posterior model distribution. This latter procedure is known as Bayesian model averaging and provides a coherent mechanism for accounting for this model uncertainty. So, instead of selecting a single "best" model and using it for prediction, Bayesian model averaging uses a weighted average of each model's individual prediction for the final predicted value, where the weight is the posterior probability of the model given the data.

Bayesian model averaging is best introduced by considering the concept of the predictive distribution. Assume we are interested in predicting the unobserved

quantity $\tilde{y}$ on the basis of the observations *y*. We denote the sampling model for $\tilde{y}$ and *y* jointly by $f(\tilde{y}|y,\theta_j,M_j)f(y|\theta_j,M_j)$, where M_j is the model selected from a set of *K* possible models. Moreover, we assign a prior $f(\theta_j|M_j)$ for the parameters and a discrete prior $f(M_j)$ defined on the model space. Then the predictive distribution (the sequence of one step ahead predictive densities) is:

$$f(\tilde{y}|y)=\sum_{j=1}^{K}\left[\int_{\theta_j} f(\tilde{y}|y,\theta_j,M_j)f(\theta_j|y,M_j)d\theta_j\right]f(M_j|y).$$

The quantity in the brackets is the predictive distribution given $\mathrm{M_j}$ obtained using the posterior of θ_j given $\mathrm{M_j}$, which is computed as:

$$f(\theta_j|y,M_j)=\frac{f(y|\theta_j,M_j)f(\theta_j|M_j)}{\int_{\theta_j} f(y|\theta_j,M_j)f(\theta_j|M_j)d\theta_j}\equiv\frac{f(y|\theta_j,M_j)f(\theta_j|M_j)}{f(y|M_j)},$$

with the second equality defining $f(y|M_j)$, which is used in computing the posterior probability assigned to M_j as follows:

$$f(M_j|y)=\frac{f(y|M_j)f(M_j)}{\sum_{i=1}^{K}f(y|M_i)f(M_i)}\equiv\frac{f(y|M_j)f(M_j)}{f(y)}$$

Clearly, the evaluation of the predictive distribution involves averaging at two levels: over (continuous) parameter values, given each possible model, and discrete averaging over all possible models.

The marginal likelihood of M_j is simply the likelihood integrated with the prior on parameters of M_j denoted here by $f(\theta_j|M_j)$. Thus,

$$L_y(M_j)=\int f(y|\theta_j,M_j)\,f(\theta_j|M_j)d\theta_j.$$

Formally (Hoeting et al, 1999), the posterior distribution of any quantity of interest, say Δ, which has a common interpretation across models, is a mixture of the model-specific posteriors with the posterior model probabilities as weights:

$$f_{\Delta|y} = \sum_{j=1}^{K} f_{\Delta|y,M_j} f(M_j|y).$$

In the above context, when posterior probability is spread widely among many models, using model averaging (than choosing a single model) seems natural. Empirical evidence of superior Bayesian model averaging predictive performance can be found in, e.g. Fernandez et al. (2001a) and Ley and Steel (2009).

Moreover, even within the standard Bayesian model averaging context, there are different approaches in the literature, mainly focused on different prior assumptions. Herein, we will focus on the application of standard Bayesian model averaging in the context of a linear regression model with uncertainty regarding the selection of explanatory variables. That is, the model uncertainty relates to the choice of which covariates should be included in the model.

Consider a normal linear regression model for *n* observations of some response variable, grouped in a vector *y*, using an intercept *a*, and explanatory variables from a set of *k* possible regressors in *Z*. We allow for any subset of the variables in *Z* to appear in the model, resulting in 2^k possible models, which will be characterized by the selection of regressors. Model M_j will be the model with $0 \leq k_j \leq k$ regressors grouped in Z_j, leading to:

$$y|a, \beta_j, \sigma \sim N(\alpha\iota_n + Z_j\beta_j, \sigma^2 I)$$

where ι_n is a vector of *n* ones, and $\beta_j \in \mathbb{R}^{k_j}$ groups the relevant regression coefficients.

For the parameters in a given model M_j, Fernandez et al. (2001a) propose a combination of a non-informative prior on the common intercept and scale and the aforementioned *g*-prior on the regression coefficients, leading to the prior density:

$$f(a, \beta_j, \sigma|M_j) \propto \sigma^{-1} f_N^{k_j}\left(\beta_j|0, \sigma^2\left(gZ_j'Z_j\right)^{-1}\right),$$

where $f_N^q(w|m, V)$ denotes the density function of a *q*-dimensional normal distribution on *w* with mean *m* and covariance matrix *V*. The regression coefficients not appearing in M_j are exactly zero, represented by a prior point mass at zero. The

amount of prior information requested from the analyst is limited to a single scalar *g,* which can either be fixed or assigned a hyper-prior distribution.

Consider the indicator variable γ_i which takes the value 1 if covariate i is included in the regression and 0 otherwise, $i = 1, \dots, k$. Given the probability of inclusion, say ϑ, γ_i will then have a Bernoulli distribution, and if the inclusion of each covariate is independent then the model size *W* will have a binomial distribution:

$$W = \sum_{i=1}^{k} \gamma_i \sim Bin(k, \theta).$$

This implies that if we consider ϑ fixed and prespecified, as is typically done in most of the literature, the prior model will have mean ϑk and variance $\vartheta(1-\vartheta)k$. For $\theta = 0{,}5$, which reflects complete prior ignorance, all models have equal prior probability $\frac{1}{k}$.

Making ϑ random (increasing that way the flexibility of the prior and reducing the dependence of posterior and predictive results on prior assumptions), we can choose a Beta prior for ϑ, with hyperparameters $a, b > 0$, that is $\theta \sim Be(a, b)$ and then the prior mean model size is $E(W) = \frac{a}{a+b}k$, while the prior distribution on model size is a binomial-beta distribution. In the special case where $a = b = 1$ we obtain a discrete uniform prior for model size with $P(W = w) = \frac{1}{k+1}, w = 0, \dots, k$.

It has been proposed (Ley and Steel, 2009) to facilitate prior elicitation by fixing $a = 1$. This still permits a flexible set of prior behavior and makes it attractive to elicit the prior in terms of the prior mean model size *m*. The choice of $m \in (0, k)$ will then determine *b* through $b = \frac{k-m}{m}$. Following this scheme, only a prior mean model size has to be specified, which is practically the same information one needs to specify for the case with fixed ϑ, which should then equal $\theta = \frac{m}{k}$. With this binomial-beta prior, the prior mode for *W* will be at zero for $m < \frac{k}{2}$ and will be at *k* for $m > \frac{k}{2}$. Generally, the difference between the fixed and random ϑ cases is striking. In particular, prior model size distributions for fixed θ are quite concentrated, while treating θ as random will typically imply more prior uncertainty for model size.

The posterior odds between any two models M_j, M_i are given by:

$$\frac{f(M_j|y)}{f(M_i|y)} = \frac{f(M_j)}{f(M_i)} \frac{L_y(M_j)}{L_y(M_i)},$$

where $L_y(M_j)$ is the marginal likelihood of model M_j. Thus, the prior distribution on model space only affects posterior model inference through the prior odds ratio. If we fix ϑ and express things in terms of the prior mean model size *m*, these prior odds are:

$$\frac{f(M_j)}{f(M_i)} = (\frac{m}{k-m})^{k_j - k_i}.$$

From this expression we get that if $m > \frac{k}{2}$ then the prior favors larger models. In the case of the $Be(a, b)$ prior on ϑ, with $a = 1$ and the prior elicitation in terms of *m* we obtain the prior odds:

$$\frac{f(M_j)}{f(M_i)} = \frac{\Gamma(1+k_j)}{\Gamma(1+k_i)} \frac{\Gamma(\frac{k-m}{m} + k - k_j)}{\Gamma(\frac{k-m}{m} + k - k_i)}.$$

For the aforementioned prior density scheme of:

$$f(a, \beta_j, \sigma | M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j | 0, \sigma^2 (g Z_j' Z_j)^{-1}),$$

if we consider *g* fixed and independent of the model size k_j, then the Bayes factor for any two models becomes:

$$\frac{L_y(M_j)}{L_y(M_i)} = (1+g)^{\frac{k_i - k_j}{2}} (\frac{1 + g(1 - {R_j}^2)}{1 + g(1 - {R_i}^2)})^{-\frac{n-1}{2}},$$

where ${R_j}^2$ is the coefficient of determination for model M_j, expressing the relative weight that the data assign to the corresponding models, and depends on sample size *n*, the factor *g* and the size and fit of both models.

Among the number of suggestions for the choice of fixed values for *g*, one of the most popular ones is the unit information prior (Kass and Wasserman, 1996) which corresponds to the amount of information contained in one observation. Since for regular parametric families the amount of information is defined via Fisher

information, we get $g = 1/n$, leading to log Bayes factors that behave asymptotically like the BIC (Fernandez et al., 2001b). Another one is the benchmark prior of Fernandez et al. (2001a) where after examination of various choices of *g* depending on the sample size *n* or the model dimension *k*, it concludes to $g = 1/\max\{n, k^2\}$.

Summarizing, it has been theoretically stated and empirically confirmed, that averaging over all the models in this fashion provides better predictive ability, as measured by various schemes proposed, such as a logarithmic scoring rule (Madigan et al., 1995), than using any single model $\mathrm{M_j}$ conditional on M.

Despite the attractiveness of this method as a solution to the problem of accounting for model uncertainty, its implementation presents several difficulties. Apart from the challenging prior elicitation, one of the most significant is that the number of terms can be enormous, rendering exhaustive summation infeasible. One way to tackle this problem is an algorithmic approach, drastically reducing the number of models that need to be considered in the average, based on two common sense principles: if a model predicts the data much worse than the best model it should be dropped from further consideration; and models that predict the data less well than any of their nested sub models should be discarded. Moreover, the integrals can be hard to compute which is overcome with the use of Markov Chain Monte Carlo method.

After these difficulties are overcome, choosing the class of models over which to average becomes the fundamental modeling task.

Chapter 4: Complete subset regression

4.1 Complete subset regression

Complete subset regression is a simple and powerful method/technique for combining forecasts, introduced by Elliott et al. (2013). In particular, for a given set of potential predictor variables, forecasts from all possible linear regression models that keep the number of predictors fixed are combined. For the case where we have *K* candidate predictors, there are *K* univariate models and $n_{k,K} = \frac{K!}{(K-k)!k!}$ different k-variate models for $k \leq K$. The set of models for a fixed value of *k* is referred to as a complete subset. Within these subsets indexed by *k* equal weighted combinations of the forecasts from all models are to be used. Moreover, the covariance matrix of the candidate regressors can be used to reach an optimal value of *k*.

While stating that several subset regression combinations have been proposed, such as equal weighted combinations of all possible univariate models, as well as equal weighted combinations of forecasts from all possible 2^k models, a general consideration would categorize the subject proposal into the wider group of forecast methods that try to reduce the effect of parameter estimation error (or equivalently limit the number of parameters). Other such methods, accomplishing the same task in different ways, are: shrinkage or ridge regression, model averaging, bagging and the Lasso.

Complete subset regression applies differential (as opposed to zero-one) shrinkage weights to each coefficient. Other approaches that share the same philosophy are: bagging, the adaptive Lasso, the Elastic Net and the adaptive Monte Carlo. Moreover, unlike ridge estimator and conventional Bayesian estimators, this method does not impose the same amount of shrinkage on each coefficient.

Complete subset regression is, by definition, a distinct procedure that should not be mistaken for the one known as best subset regression, sometimes found as full subset regression in the literature. The latter considers all possible models and for a given model size selects the best in terms of fit (lowest sum of squared residuals). Of the resulting set of optimal models for a given dimension, the procedure then chooses the one with the smallest value of some criterion, such as Mallows C_p, leading finally to one “best” model, which is obviously different in philosophy compared to the complete subset regression.

Naturally, if in each problem we could estimate models for all possible combinations of variables, it would be a great option. However, with only 20 variables the number of regressions would be over 1 million. The complete subset regression serves as a solution between using only one subset and all possible subsets.

4.2 Setup, symbolism and theoretical review

The symbolism herein will focus into predicting a variable y_{T+1} using a linear regression model based on a set of *K* predictors $x_T \in \mathbb{R}^K$ and a set of data, $\{(y_{t+1}, x_t), t = 0, \dots, T-1\}$. For all t, we let $E(x_t x_t') = \Sigma_X$ and moreover it can be assumed that $E(x_t) = 0$. The regressions including only a subset of predictors can be distinguished defining β to be a $K \times 1$ vector with slope coefficients in the rows representing included regressors and zeros in the rows of the excluded variables. The "true" value of β (the population value of the projection of *y* on *X*) is represented as β_0, where *y* is a $T \times 1$ vector and $X = (x_0, x_1, \dots, x_{T-1})'$ stacks the *x* observations into a $T \times K$ matrix. Moreover, S_i is a $K \times K$ matrix with zeros everywhere except for ones in the diagonal cells corresponding to the included variables, so that if its $[j, j]$ element is one, the j-th regressor is included, while if this is a zero, the j-th regressor is excluded. Apparently, sums over i are sums over all permutations of S_i.

The subject method uses equal weighted combinations of forecasts based on all possible models that include a particular subset of the predictor variables. Each subset is defined by the set of regression models that include a fixed number of regressors $k \leq K$. First, the regression of y_t on a particular subset of the regressors is run and then we average the results across all *k*-dimensional subsets of the regressors to provide an estimator, say $\hat{\beta}$, for forecasting. With *K* regressors in the full model and *k* regressors chosen for each of the "short" models, there will be subset regressions to average over. In turn, each regressor gets included a total of $n_{k-1,K-1}$ times.

For the simple case where $k = 1$, and thus there are $n_{1,K} = K$ short, single variable regressions, $\widehat{\beta}_\iota$ has all its elements zero, except for the least squares estimate of y_t on x_{it} in the i-th row. Then, the equal weighted combination of forecasts from the individual models is $\hat{y}_{T+1} = \frac{1}{K}\sum_{i=1}^{K} x'_T \widehat{\beta}_\iota$.

The main theoretical feature of this method is the fact that the subset regression coefficients can be computed as averages over least squares estimates of the subset regressions. When correlation between the covariates exists, the individual

regressions will be affected by omitted variable bias. However, the subset regression estimators are themselves approximately a weighted average of the components of the full regression OLS estimator, $\hat{\beta}_{OLS}$.

Elliott et al. (2013) formalized the above into the next theorem: assume that as the sample size gets large $\hat{\beta}_{OLS} \xrightarrow{p} \beta_0$ for some β_0 and $T^{-1}X'X \xrightarrow{p} \Sigma_X$. Then, for fixed *K*, the estimator for the complete subset regression $\hat{\beta}_{k,K}$ can be written as:

$$\hat{\beta}_{k,K} = \Lambda_{k,K}\hat{\beta}_{OLS} + O_p(1),$$

where

$$\Lambda_{k,K} = \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} (S_i'\Sigma_x S_i)'(S_i'\Sigma_x).$$

From the above equation it is obvious that $\Lambda_{k,K}$ is not diagonal in general and hence the coefficients $\hat{\beta}_{k,K}$ are not simple regressor by regressor shrinkages of the OLS estimates. Instead, they are functions (a weighted sum) of all the OLS coefficients in the regression. The weights depend not only on *k* and *K*, but on all elements Σ_{ij} in Σ_x. As an illustration, for the simple case where $K = 2, k = 1$ we have:

$$\Lambda_{1,2} = \frac{1}{2}\begin{pmatrix} 1 & \frac{\Sigma_{12}}{\Sigma_{11}} \\ \frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{pmatrix}.$$

Each row of $\Lambda_{1,2}$ reflects inclusion of a particular subset regression in the average. To see this, the first row gives the first element of $\hat{\beta}_{1,2}$ as a weighted sum of the $\hat{\beta}_{OLS}$. Its own coefficient is given a relative weight of one, while the remaining coefficient(s) are those expected from omitted variable bias formulas. Obviously, the effect of dividing by $n_{1,2} = 2$ is to shrink all coefficients towards zero.

Generally, for $k > 1$ each regressor appears with an increased frequency in the regressions. This way, its effect on $\Lambda_{k,K}$ gets larger, but at the same time tempered through the omitted variable bias. Since the first effect is greater than the latter, an increased *k* will generally reduce the amount of shrinkage, till the limit $k = K$, where no shrinkage is accomplished and the method coincides with the OLS.

For the special case where the covariates are orthonormal, $\hat{\beta}_{OLS} = X'y$ and so we have:

$$\hat{\beta}_{k,K} = \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \hat{\beta}_i = \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} S_i X'y = \left(\frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} S_i \right) \hat{\beta}_{OLS}.$$

The elements of $\sum_{i=1}^{n_{k,K}} S_i$ are zero for the off diagonal terms, and equal the number of times a regressor is included in the subset regressions for the diagonal terms. At the same time, the diagonal terms equal $n_{k,K}$ minus the number of the times a regressor is excluded. Thus, $\sum_{i=1}^{n_{k,K}} S_i = n_{k,K} - n_{k,K-1}$ and so we get, for this orthonormal special case, that:

$$\hat{\beta}_{k,K} = \frac{n_{k,K} - n_{k,K-1}}{n_{k,K}} \hat{\beta}_{OLS} = \lambda_{k,K} \hat{\beta}_{OLS},$$

where $\lambda_{k,K} = \frac{n_{k,K} - n_{k,K-1}}{n_{k,K}}$, which is a scalar.

It is obvious that $\lambda_{k,K}$ is a function of *k* and *K*. For any value of *K*, it is a linear function of *k* that increases to the value of one, corresponding to the case where we run OLS with all variables included. It can easily be shown, that for smaller *K*, the slope of $\lambda_{k,K}$ gets larger, and thus the amount of shrinkage is larger for any fixed *k*, the smaller *K* is. Essentially, the smaller *k* is relatively to *K*, the greater the amount of shrinkage.

4.3 Associated risk

Forecasting is an estimation task and risk is the expected loss as a function of the true, still unknown, model parameters. Under MSE loss, risk amounts to the expected loss. For any estimator we have:

$$E\left[\left(y_{T+1} - \hat{\beta}'_T x_T\right)^2\right] = E\left[\left(y_{T+1} - \beta'_0 x_T + (\beta_0 - \hat{\beta}_T)' x_T\right)^2\right] =$$
$$E\left[\left(\varepsilon_{T+1} + (\beta_0 - \hat{\beta}_T)' x_T\right)^2\right] = \sigma_\varepsilon{}^2 \left(1 + T^{-1} \sigma_\varepsilon{}^{-2} E\left[T\left(\hat{\beta}_T - \beta_0\right)' x_T x_T' (\hat{\beta}_T - \beta_0)\right]\right)$$

where ε_{T+1} is the residual from the population projection of y_{T+1} on x_T and $\sigma^2{}_\varepsilon$ is its variance. Since the first term does not depend on $\hat{\beta}$, the term $\sigma_\varepsilon{}^{-2} E\left[T(\hat{\beta}_T - \beta_0)' x_T x_T' (\hat{\beta}_T - \beta_0)\right]$ becomes of interest.

As in all biased methods, for values of β_0 that lie far from zero, the associated risk when shrinking the coefficients towards zero is generally large. Thus, complete subset regression is recommended when β_0 is local to zero. In order to ensure such a state, we assume that for some fixed vector b, $\beta_0 = \frac{\sigma_\varepsilon}{\sqrt{T}} b$. Elliott et al. (2013), assuming that $\{y_{t+1}, x_t\}$ are i.i.d., $E\left[\left(\hat{\beta} - \beta_0\right)^2 | x_{T+1}\right] = E\left[\left(\hat{\beta} - \beta_0\right)^2\right]$ and $\frac{\widehat{\beta}_{OLS} - \beta}{\sqrt{T}} \xrightarrow{d} N\left(0, {\Sigma_x}^{-1}\right)$, proved that in large samples:

$$ {\sigma_\varepsilon}^{-2} E\left[T\left(\hat{\beta}_T - \beta\right)' \Sigma_x \left(\hat{\beta}_T - \beta\right)\right] \approx \sum_{j=1}^{K} \zeta_j + b'\left(\Lambda_{k,K} - I\right)' \Sigma_x \left(\Lambda_{k,K} - I\right) b $$

where ζ_j are the eigenvalues of ${\Lambda'}_{k,K} \Sigma_x \Lambda_{k,K} {\Sigma_x}^{-1}$.

Based on the above, we see that the expected loss is a function of *k,K*, the elements of *b* and of the variance covariance matrix. Naturally, different trade-offs can be explored by varying these parameters.

A useful insight offers the comparison of the associated risk, against that of models estimated by OLS, which in some cases can be accomplished analytically. One example is the case where we explore combinations of univariate models, that is $k = 1$, with a Σ_x with all its diagonal elements being one and all off-diagonal elements being *ρ*. Then, *b* is a vector of ones with dimension *K*, the risk for OLS regression is also *K* and the risk of the subset regression is:

$$ E\left[\left(y_{T+1} - {\hat{\beta}'}_T x_T\right)^2\right] = \frac{1}{K}\left(1 + (K-1)\rho^2\right) + (\rho - 1)^2 \left(\frac{K-1}{K}\right)^2 (K + K(K-1)\rho) $$

From this, we can compare the risk for several (K, ρ) pairs with *K*. It emerges that for small values of *K*, such as *K*<6, $\frac{1}{\mathrm{K}}(1 + (\mathrm{K}-1)\rho^2) + (\rho - 1)^2 \left(\frac{\mathrm{K}-1}{\mathrm{K}}\right)^2 (\mathrm{K} + \mathrm{K}(\mathrm{K}-1)\rho) < K$ for nearly all possible correlations (Elliott et al., 2013). For *K*>6, it still holds that the subset regression risk is smaller than that of OLS regression, apart from a small region with small values of *ρ* and $\mathrm{k} = 1$. This means that an equal weighted average of univariate forecasts can perform better than the conventional multivariate model which includes all the predictors.

4.4 Computational aspects and variations on implementation

When *K* is large enough, $n_{k,K}$ can become too large as to allow all models in a given subset to be examined. A way to deal with this is to decrease the number of possible models considered in each subset. One approach is to randomly draw a smaller number of models and average across them, with uniform probability weighting within each subset. Alternatively, one could use some of each model's information to decide on inclusion, which of course is more complicated and computationally demanding.

Another variation rises from the weighting mode to be used. While equal weighted averages perform well and are the simplest method possible, other options are available such as Bayesian model averaging. Moreover, additional/alternative risk minimizing criteria could be used, such as Geweke and Amisano (2011) optimal prediction pool approach, which relies on maximizing the log predictive scoring rule.

4.4.1 Markov Chain Monte Carlo Samplers

With a small number of variables, it is straightforward to enumerate all potential variable combinations to obtain posterior results. For a larger number of covariates, this becomes from time intensive to impossible. In such a case, MCMC samplers gather results on the most important part of the posterior model distribution and thus approximate it as closely as possible. In particular, Bayesian Model Averaging mostly relies on the Metropolis-Hastings algorithm, which "walks" through the model space as follows: at step i, the sampler stands at a certain "current" model M_i with posterior model probability $f(M_i|y,X)$. In step $i+1$ a candidate model M_j is proposed. The sampler switches from the current model to model M_j with probability $p_{i,j} = \min\left(1, \frac{f(M_j|y,X)}{f(M_i|y,X)}\right)$. In case model M_j is rejected, the sampler moves to the next step and proposes a new model M_k against M_i. In case model M_j is accepted, it becomes the current model and has to survive against further candidate models in the next step. In this manner, the number of times each model is kept will converge to the distribution of posterior model probabilities $f(M_i|y,X)$. Naturally, the quality of an MCMC approximation to the actual posterior distribution depends on the number of draws the MCMC sampler runs for. In particular, the sampler has to start out from some model that might not be a "good" one, in terms of posterior model probability. Hence the first batch of iterations will typically not draw models with high posterior probabilities as the sampler will only after a while converge to spheres of models with the largest marginal likelihoods. Therefore, this first set of iterations (the so called burn-ins) is to be omitted from the computational results.

4.5 Bayesian selection of k

One of the most popular averaging approaches is the simple equal-weighted average (Rapach et al., 2010). Under the complete subset regression scheme this idea is extended, since the forecast combination is constructed by using equal-weighted combinations based on all possible models that keep the number of predictors fixed. That is, instead of choosing the weights, in the subset regression combinations we have to choose the number of predictors *k*. Monte Carlo shows that the predictive performance of complete subset regressions is sensitive to the choice of *k* and thus the need of a data-driven method for selection of *k* emerges. A real time algorithm of selecting *k* recursively is presented here, which is likelihood based (Bayesian).

For the one step ahead forecasts approach, at each time point in the out of sample period, indexed by $t+1$, we need to compute the posterior probabilities of all values of $k \in \{1, \dots, K\}$, based on the data up to time t. Then, the most probable value of *k* is selected and a forecast at time $t+1$ is produced, based on the selected complete subset. In the Bayesian context, uncertainty about any quantity in interest is represented by probability distributions. Apart from the value of *k*, another quantity of interest is the model specification, representing the set of predictors included in the j-th model and denoted by $m_j, j = 1, \dots, M$, where $M = \sum_{i=1}^{K} n_{i,K}$. The last quantity of interest is the totality of the model parameters denoted by θ_j. After specifying appropriate prior distribution for these three quantities, that is $f(m_j), f(k|m_j)$ and $f(\theta_j|m_j, k)$, their joint posterior distribution is given by:

$$f(m_j, k, \theta_j|y_{1:t}) \propto f(m_j)f(k|m_j)f(\theta_j|m_j, k)L(y_{1:t}|m_j, k, \theta_j),$$

where $L(y_{1:t}|m_j, k, \theta_j)$ is the likelihood of the data up to time *t*. Consequently, the marginal posterior distribution of *k* is obtained as:

$$f(k|y_{1:t}) \propto \sum_{j=1}^{M} \left[\int_{\theta_j} f(\theta_j|m_j, k)L(y_{1:t}|m_j, k, \theta_j)d\theta_j \right] f(m_j)f(k|m_j).$$

The quantity (integral) in the brackets is the marginal likelihood of the data with *k* predictors and model specification m_j, i.e. $L(y_{1:t}|m_j, k)$.

Concerning the prior specification we have the following: consider the indicator variable γ_i which takes the value 1 if covariate i is included in the regression and 0 otherwise. Given the probability of inclusion, say ϑ, γ_i will then have a Bernoulli

distribution, and if the inclusion of each covariate is independent then the model size *W* will have a binomial distribution, that is $W = \sum_{i=1}^{k} \gamma_i \sim Bin(k, \theta)$. Moreover, taking ϑ fixed and prespecified, as is typically done in most of the literature, the prior probability of the j-th models is taken to be:

$$f(m_j) = \theta^{k_j}(1-\theta)^{K-k_j},$$

where k_j is the number of predictors included in model m_j. The prior probability of *k* given the model specification m_j is apparently $f(k|m_j) = 1$, if $k_j = k$ and $f(k|m_j) = 0$, otherwise. Thus, the joint prior of (k, m_j) is:

$$f(k, m_j) = f(m_j)f(k|m_j) = \theta^{k_j}(1-\theta)^{K-k_j} I(k_j = k),$$

while the marginal prior on *k* is the $Bin(K, \theta)$. Based on the above, the marginal posterior distribution of *k* is:

$$f(k|y_{1:t}) \propto \theta^{k}(1-\theta)^{K-k} \sum_{j=1}^{M} L(y_{1:t}|m_j, k)\, I(k_j = k).$$

Chapter 5: Forecast accuracy metrics and simulation study

5.1 Evaluation of forecast accuracy

Forecasts are widely used in every field of human activity and are of great importance since good forecasts (may) lead to good decisions. The importance of forecast evaluation follows immediately. Evaluation may refer to examination of whether the forecasting method in hand gives good forecasts (absolute performance) and/or it is better than any competitor method. There are many ways of measuring the accuracy of forecasts, and the answers to these questions depend on what is being forecast, what accuracy measure is used, and what data set is analyzed. We can measure and average forecast errors in several ways. Although interpretability is a major criterion, applicability (no single measure is appropriate in most situations) and efficiency must be also taken into account.

5.1.1 Scale-dependent errors

Let y_i denote the i-th observation and $\hat{y}_i$ denote a forecast of y_i. The forecast error is simply $e_i = y_i - \hat{y}_i$ and is defined on the same scale as the data. Thus, all accuracy measures that are based directly on e_i are scale-dependent and cannot be used to make comparisons between series that are on different scales. Some of the most commonly used scale-dependent measures are based on the absolute errors or squared errors:

$$\text{Mean absolute error (MAE)} = \text{mean}(|e_i|),$$

$$\text{Mean Squared Error (MSE)} = \text{mean}({e_i}^2),$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\text{mean}({e_i}^2)}.$$

From the above, MAE is the easiest to understand and compute. Moreover, the use of absolute or squared values prevents negative and positive errors from offsetting each other. Since all these metrics are on the same scale as the data, none of them are meaningful for assessing a method's accuracy across multiple series.

5.1.2 Percentage errors

The percentage error is given by $p_i = \frac{100e_i}{y_i}$ and having the advantage of being scale independent, can be used to compare forecast performance between different data series. The most commonly used metric is:

$$\text{Mean absolute percentage error (MAPE)} = \text{mean}(|p_i|).$$

Measurements based on percentage errors have the disadvantage of being infinite or undefined if $y_i = 0$ for any *i* in the period of interest, and having extreme values (extremely skewed distribution) when any y_i is close to zero. In the case of intermittent-demand data, it is impossible to use MAPE because of the occurrences of zero periods of demand.

The MAPE also has the disadvantage of putting a heavier penalty on positive errors than on negative ones. This observation has led to the use of the symmetric MAPE (sMAPE) proposed by Armstrong (1992), which is defined by:

$$\text{Symmetric mean absolute percentage error (sMAPE)} = \text{mean}\left(\frac{200|y_i - \hat{y}_i|}{y_i + \hat{y}_i}\right).$$

However, if y_i is close to zero, then $\hat{y}_i$ is also likely to be close to zero, thus having a measurement still involving division by a number close to zero. Moreover, the value of sMAPE can be negative, giving it an ambiguous interpretation.

5.1.3 Relative errors

An alternative to percentages is to divide each error by the error obtained using some benchmark method of forecasting. Let $\mathrm{r_i} = \frac{\mathrm{e_i}}{\mathrm{e_{i*}}}$ denote the relative error where $\mathrm{e_i}^*$ is the forecast error obtained from the benchmark method (usually the so called naïve method, where each new forecast equals to the last observation). Then several measures can be defined, such as:

$$\text{Median relative absolute error (MdRAE)} = \text{median}(|\mathrm{r_i}|).$$

5.1.4 Scaled errors

Scaled errors were proposed by Hyndman and Koehler (2006) as an alternative to using percentage errors when comparing forecast accuracy across series on different scales. A scaled error is given by $q_t = \frac{e_t}{Q}$ where Q is a scaling statistic computed on the training data. For a non-seasonal time series, a useful way to define the scaling statistic is the mean absolute difference between consecutive observations:

$$Q = \frac{1}{N-1}\sum_{j=2}^{N}|y_i - y_{i-1}|.$$

That is, *Q* is the MAE for naïve forecasts computed on the training data. Because the numerator and denominator both involve values on the scale of the original data, q_t is independent of the scale of the data. A scaled error is less than one if it arises from a better forecast than the average naïve forecast computed on the training data. Conversely, it is greater than one if the forecast is worse. The most commonly used mean absolute scaled error is simply:

$$\text{Mean absolute scaled error (MASE)} = \text{mean}(|q_t|).$$

As a link to the next section of our simulation study, it is necessary to state the importance to evaluate forecast accuracy using genuine forecasts. That is, it is invalid to look at how well a model fits the historical data; the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when estimating the model. A common practice when choosing models is to use a portion of the available data for testing, and use the rest of the data for estimating (training) the model. Then the testing data can be used to measure how well the model is likely to forecast on new data.

5.2 Simulation study

To better understand and illustrate how the complete subset combination approach works, Monte Carlo simulations were accomplished, in order to study the absolute forecast performance of the subset regression, as well as its performance relative to alternative and at the same time well established methods. The main aspects of the Monte Carlo design as well as the variations of each simulation held are briefly presented in the next paragraph.

5.2.1 Simulation setup

We first assume a multiple linear regression model:

$$Y_{t+1} = \sum_{k=1}^{K} \beta_k x_{kt} + \varepsilon_{t+1}, \varepsilon_{t+1} \sim N(0, {\sigma_\varepsilon}^2).$$

We generally assume a sample size of T=120 observations, resembling monthly observations for a 10-year period. The first 96 of them will be considered as in-sample and the latter 24 of them as out-of-sample, corresponding to a 2-year period where predictions are supposed to be made, considering one step ahead forecasts of Y_{t+1}. We generate the covariance matrix of the *X*-variables $\Sigma_X = Cov(X_1, \dots, X_k)$ and we further control the magnitude of the variances of the *X*-variables (that is the

diagonal elements of the covariance matrix) via multiplication with another diagonal matrix of our choice. Data are assumed to be normally distributed and i.i.d. which means that the explanatory variables are draw from a multivariate normal distribution, with the afore-generated covariance matrix.

Several simulation variations were accomplished concerning the number of the candidate predictors. Here will be presented the conclusions of taking into account 4, 6 and 8 predictors (that is $K = 4, 6$ and 8). As an example, in the latter case we have 255 possible models, deriving from each $1 \leq k \leq 8 = K$, (eg for $k = 1$ there are 8 univariate candidate models, for $k = 2$ there are 28 models with two variables and so on).

Concerning the regression parameter, two approaches were examined: the one is to consider a random design, where β_0 (that is the pseudo true value of β) can have values far from zero. However, in common with all biased methods, for values of β_0 far from zero, the risk (the expected loss as a function of the true, yet unknown, model parameters) is large and so it is appropriate not to shrink coefficients towards zero. To capture such a situation, where a shrinkage method is supposed to add value, the second approach assumes that β_0 is local to zero (either setting $\beta_0 = \frac{\sigma_\varepsilon}{\sqrt{T}}$ or generating β_0 by a random normal distribution having its mean systematically local to zero). We report all cases of different number of predictors for the β_0 that is local to zero and we illustrate the alternative option in the case where $K = 6$.

Further variations were studied in the case where $K = 6$. One of them examines the event of having a reduced number of (generated) observations. We assume T=20, taking the first 16 of them as in-sample and the latter 4 of them as out-of-sample. Another variation consists on changing the probability θ of inclusion for each regressor in the model, during the Bayesian analysis of the regression, as will be explained below.

For each iteration of the Monte Carlo simulation, forecasts are calculated based on each of the candidate models. Then we combine the forecasts by simple averaging for each fixed value of k. That is, we end up with K (for $k = 1, \dots, K$) proposals, so that we can first get a forecasting performance for each fixed k. For this purpose, any of the forecasting measures presented in section 5.1 could be used. Here the Mean Squared (Forecasting) Error will be presented, being one of the most widely used and easy to interpret.

In order to compare the fixed k proposals with other well established methods, we calculate, once again for each iteration, the prediction that is based on:

1. The model estimated by the OLS method.
2. The "best" models, identified by the AIC and the BIC criterion. That is, we calculate the AIC & BIC for each model (for each iteration), select the best models and then make the prediction based on them.
3. Bayesian Model Averaging. The predictions are obtained by weighting each model's forecast by its posterior probability:

$$f(\tilde{y}|y) = \sum_{j=1}^{2^k} \hat{y}_j f\left(M_j|y\right),$$

where $\hat{y}_j$ is the posterior mean and $f\left(M_j|y\right)$ is the posterior probability of the j-th model, which follows from Bayes' theorem:

$$f\left(M_j|y\right) = \frac{f\left(y\middle|M_j\right)f(M_j)}{\sum_{i=1}^{2^k} f(y|M_i)f(M_i)}.$$

We accomplish this method in accordance with the section 3.8, adopting a combination of a non informative prior on the common intercept α and scale σ_ε and a g-prior (Zellner, 1986) on the regression coefficients β_j, leading to the prior density:

$$f(a, \beta_j, \sigma_\varepsilon\middle|M_j) \propto \sigma^{-1} f_N^{k_j}(\beta_j|0, {\sigma_\varepsilon}^2\left(g Z_j' Z_j\right)^{-1}),$$

where Z_j are the demeaned regressors that are included in the j-th model. Under this specification, $\tilde{y}|y, M_j$ follows a t-distribution with location parameter $\hat{y}_j = \frac{1}{T}\sum_{i=1}^{T} y_i + \frac{{x_j}'\beta_j}{g+1}$. To sum up, we follow Fernandez et al. (2001b) and set $g = \max\{{}^1/_T, {}^1/_{K^2}\}$ and we study two variations concerning the probability θ of inclusion for each regressor in the model: first we assume that $\theta = 1/2$, reflecting complete prior ignorance about the model specification. Second, to slightly penalize models with too many predictors, we set $\theta = 1/3$, since it is known from previous studies (Goyal and Welch, 2008, Elliott et al., 2013) that the best performing forecasts arise from models including fewer predictors.

4. The most probable model, as identified by the Bayesian Model Averaging calculations. That is, we select the model with the highest posterior probability $f(M_j|y)$ for each iteration and then make the prediction based on it.
5. Bayesian Model Sampling and Averaging: with a small number of variables, it is straightforward to enumerate all potential variable combinations to obtain posterior results. For a larger number of covariates, this becomes from time intensive to impossible and MCMC samplers can be used, in order to gather results on the most important part of the posterior model distribution, as outlined in section 4.4.1. For the simulation case of the current section, complete enumeration is feasible and thus accomplished. Moreover, in this scheme we set $g = {}^{1}/_{T}$, that is the Unit Information Prior (Kass and Wasserman, 1996) for the Zellner's g-prior, as opposed to the choice of $g = \max\{{}^{1}/_{T}, {}^{1}/_{K^2}\}$ that was followed in the BMA.

Summarizing, we have a baseline configuration for the simulation consisting of a sample of T=120, a β_0 that is local to zero and $\theta = 1/2$, where we examine the cases of 4, 6 and 8 predictors and then we modify the baseline for the case of 6 predictors to get three more variations/cases.

5.2.2 Simulation results

The simulation experiment consisted of 10.000 iterations for each combination/variation. We report the averages (over all iterations) of MSE values for the OLS, the best AIC & BIC and the most probable model, the Bayesian Model Averaging and Bayesian Model Sampling, as well as for each fixed k of the Complete Subset Regression, serving as the key metric of the efficiency of the complete subset approach. Moreover, we capture the frequency of each fixed k having the lowest MSE, compared to the rest of the $k = 1, \dots, K$.

5.2.2.1 Case 1: Baseline configuration for $K = 4$

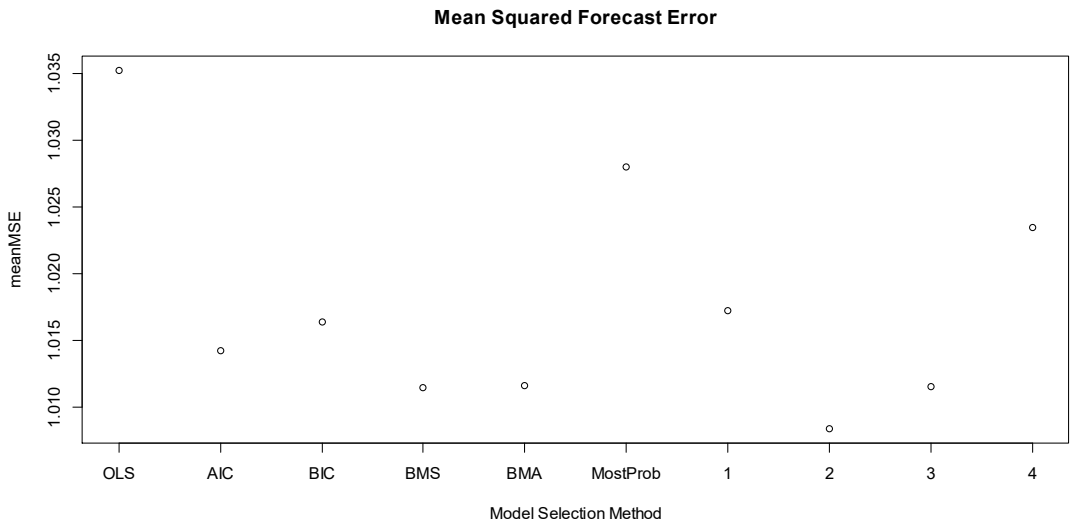
In this case, the best performing method, producing the lowest out-of-sample MSE on average, of the ones used to benchmark the subset regression, is the Bayesian Model Averaging and the Bayesian Model Sampling, followed by the best AIC and the best BIC model. As shown in table 5.1 and graph 5.1, the best subset, in terms of mean MSE, is the one with $k = 2$ and performs better than any other alternative. In particular, if we take into account the random noise in the data (that is ${\sigma_\varepsilon}^2$) and the fact that the predictive capability of a model can (should) not be better than that noise, the best subset can be said that performs 36% better than the best alternative. Moreover, the subset regressions that include $k = 3$ predictors perform

quite as well on average as the BMA and BMS. Univariate subsets perform a bit worse on average than the best BIC model, while the poorest performance corresponds to $k = 4$. Accomplishing the scheme where β_0 is local to zero, the above results are confirming the fact that shrinkage adds value.

As shown in table 5.2 the frequency of each fixed k having the lowest MSE (compared to the rest of the $k = 1, \dots, K$) is greater for $k = 1$, followed by $k = 4$. This fact, in conjunction with the MSE performance of table 5.1, indicates that the least volatile forecasts arise when we use two or three regressors. Finally, the almost identical value of MSE for the BMA and BMS, indicates that the difference in the selected values of the g-prior cannot produce a remarkable posterior differentiation.

Method	MSE
CSR, k=2	1,008395
BMS	1,011417
CSR, k=3	1,011515
BMA	1,011579
AIC	1,014205
BIC	1,016361
CSR, k=1	1,017228
CSR, k=4	1,0234592
Most Probable	1,0280151
OLS	1,0352370

Table 5.1: MSE performance in ascending order for $K = 4$



Graph 5.1: MSE performance for $K = 4$

Fixed k value	Frequency
1	0,3841
2	0,1954
3	0,1492
4	0,2713

Table 5.2: Frequency of fixed k having lowest MSE (among $k = 1, \dots, K$) for $K = 4$

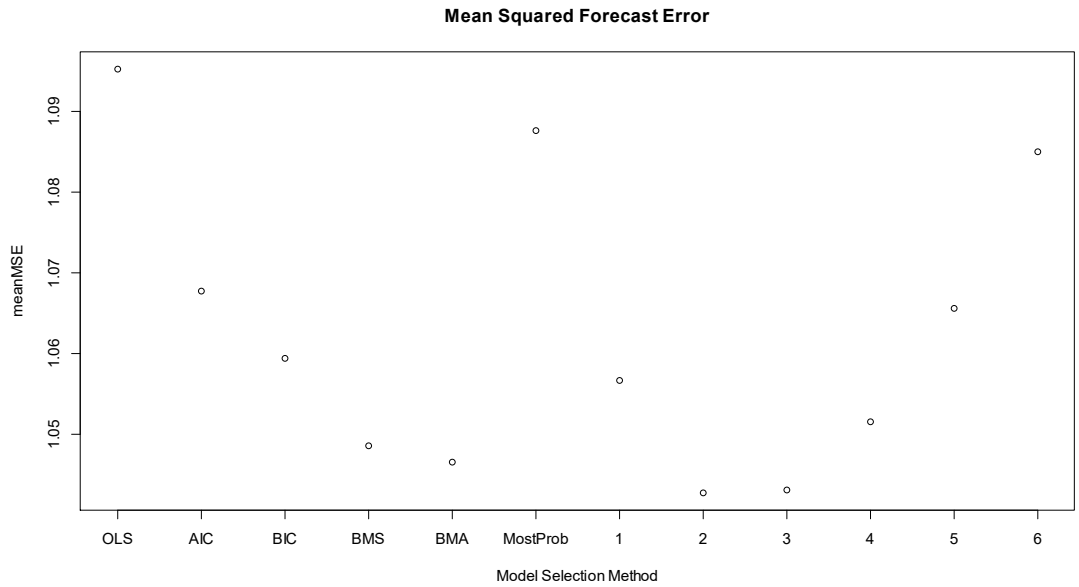
5.2.2.2 Case 2: Baseline configuration for $K = 6$

In this case, the best performing method of the ones used to benchmark the subset regression, is again the Bayesian Model Averaging and the Bayesian Model Sampling, followed by the best BIC and the best AIC model. As shown in table 5.3 and graph 5.2, the best subset, in terms of mean MSE, is the one with $k = 2$, followed by the one with $k = 3$. These subsets perform better than any other alternative. In particular, if we try to isolate the random noise in the data, the best subset performs 8,8% better than the best alternative. Moreover, the subset regressions that include $k = 4$ and $k = 1$ predictors perform better on average than the best BIC and AIC. The poorest performance once again corresponds to $k = K = 6$, which was anticipated based on β_0 that is local to zero. As shown in table 5.4 the frequency that each fixed k has the lowest MSE (compared to the rest of the $k = 1, \dots, K$) is greater for $k = 1$, followed by $k = 6$. The data of this table in conjunction with the MSE performance of table 5.3, indicates that the least volatile forecasts arise when we use two, three or four regressors. We also note that the model estimated by the OLS cannot compete the other methods in terms of out-of-sample MSE (which is also documented in cases number one and three as well), and this was expected since the poor statistical significance of the regression parameters, deriving from the fact that β_0 is local to zero, deprives the predictive precision and capability.

Method	MSE
CSR, k=2	1,042793
CSR, k=3	1,043096
BMA	1,046583
BMS	1,048614
CSR, k=4	1,051551
CSR, k=1	1,056652
BIC	1,059409
CSR, k=5	1,065599
AIC	1,067713
CSR, k=6	1,084996

Most Probable	1,087676
OLS	1,095245

Table 5.3: MSE performance in ascending order for $K = 6$



Graph 5.2: MSE performance for $K = 6$

Fixed k value	Frequency
1	0,3399
2	0,1542
3	0,1051
4	0,1053
5	0,1004
6	0,1951

Table 5.4: Frequency of fixed k having lowest MSE (among $k = 1, \dots, K$) for $K = 6$

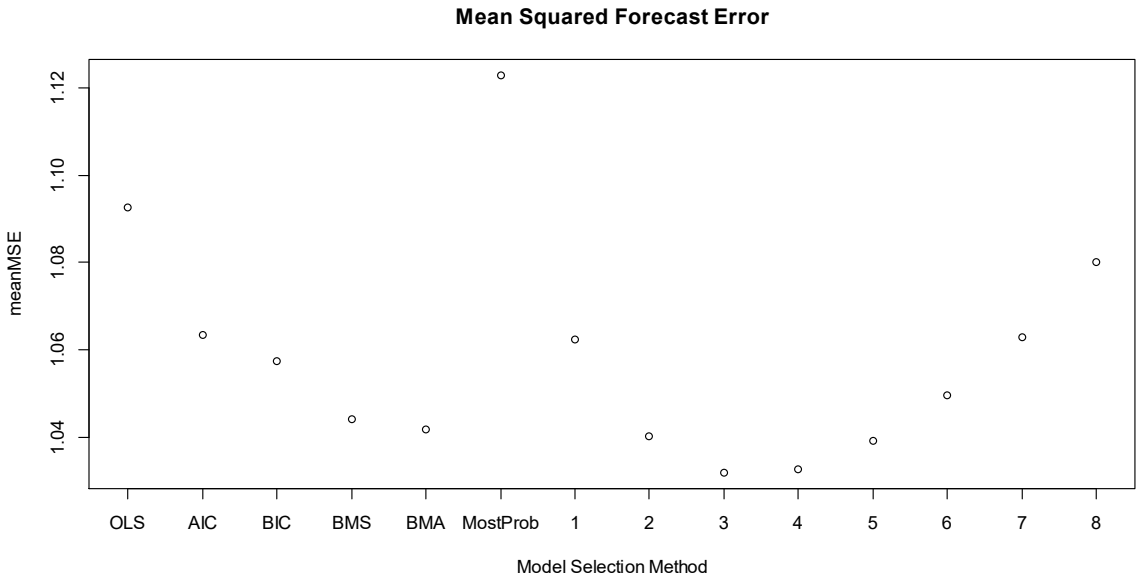
5.2.2.3 Case 3: Baseline configuration for $K = 8$

In this case, the best performing method, of the ones used to benchmark the subset regression, is the Bayesian Model Averaging, followed by the BMS, then the best BIC and the best AIC model. As shown in table 5.5 and graph 5.3, the subset regressions with $k = 2, 3, 4$ and 5 perform better than any alternative. The best subset, in terms of mean MSE, is the one with $k = 3$ and performs 30% better than the best alternative. Moreover, the subset regressions that include a large number of predictors (that is $k = 6$,7 or 8) have the poorest performance. As shown in table 5.6 the frequency that each fixed k has the lowest MSE (compared to the rest of the

$k = 1, \dots, K$) is greater for $k = 1$, followed by $k = 8$. Cross evaluation of tables 5.5 and 5.6, yields that the most volatile forecasts arise when we use one, seven or eight regressors.

Method	MSE
CSR, k=3	1,032136
CSR, k=4	1,032752
CSR, k=5	1,039215
CSR, k=2	1,040433
BMA	1,041969
BMS	1,044305
CSR, k=6	1,049589
BIC	1,057633
CSR, k=1	1,062370
CSR, k=7	1,063006
AIC	1,063568
CSR, k=8	1,080209
OLS	1,092626
Most Probable	1,122705

Table 5.5: MSE performance in ascending order for $K = 8$



Graph 5.3: MSE performance for $K = 8$

Fixed k value	Frequency
1	0,2602
2	0,1193
3	0,0981
4	0,0969
5	0,0892
6	0,0722
7	0,084
8	0,1801

Table 5.6: Frequency of fixed k having lowest MSE (among $k = 1, \dots, K$) for $K = 8$

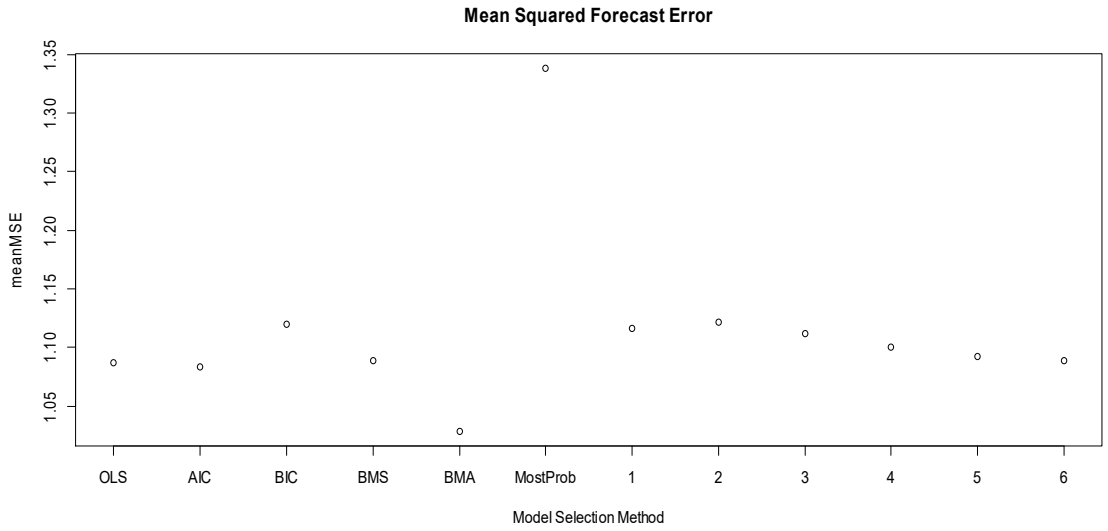
5.2.2.4 Case 4: $K = 6$ and β_0 not local to zero

In this case, we keep the baseline configuration of the simulation as in section 5.2.2.2, however modifying β_0, so that it can have values not local to zero. As shown in table 5.7 and graph 5.4, the best performing methods of the alternatives to which subset regression is being compared, are the Bayesian Model Averaging, followed by the best AIC and the OLS model. In fact, these are the best methods on average, outperforming the best subset regression. That is the subset with all the predictors, followed by the one with models that have $k = 5$ predictors, then $k = 4$ and so on. As already stated, in such a scenario where β_0 can have values far from zero, it is highly expected that no value will be added by shrinking. This prior belief is indeed confirmed by two results. The first one is that the complete subset method systematically chooses large models, in the sense that large models have better predictive performance: 35 percent of the time complete subset regression uses models with $k = K$ predictors and 22 percent of the time models with $k = K - 1$ predictors. Essentially, setting $k = K$ corresponds to simply running OLS with all variables included, which means that there is no shrinkage at all (the method includes all predictors and so does not average across multiple models at all). The second result confirming the aforementioned prior belief is the very poor performance of the complete subsets for small k and of course the poor relative performance of subset regressions on average, in comparison with the alternatives. Extending our remark concerning the OLS in case number two, we now note that the model defined by the OLS can in this scenario offer a competitive prediction performance.

Method	MSE
BMA	1,028824
AIC	1,083663

OLS	1,087291
CSR, k=6	1,088433
BMS	1,089098
CSR, k=5	1,092738
CSR, k=4	1,100500
CSR, k=3	1,112250
CSR, k=1	1,116449
BIC	1,120162
CSR, k=2	1,121686
Most Probable	1,338101

Table 5.7: MSE performance in ascending order for $K = 6$, β_0 not local to zero



Graph 5.4: MSE performance for $K = 6$, β_0 not local to zero

Fixed k value	Frequency
1	0,1542
2	0,0621
3	0,0723
4	0,1337
5	0,2234
6	0,3543

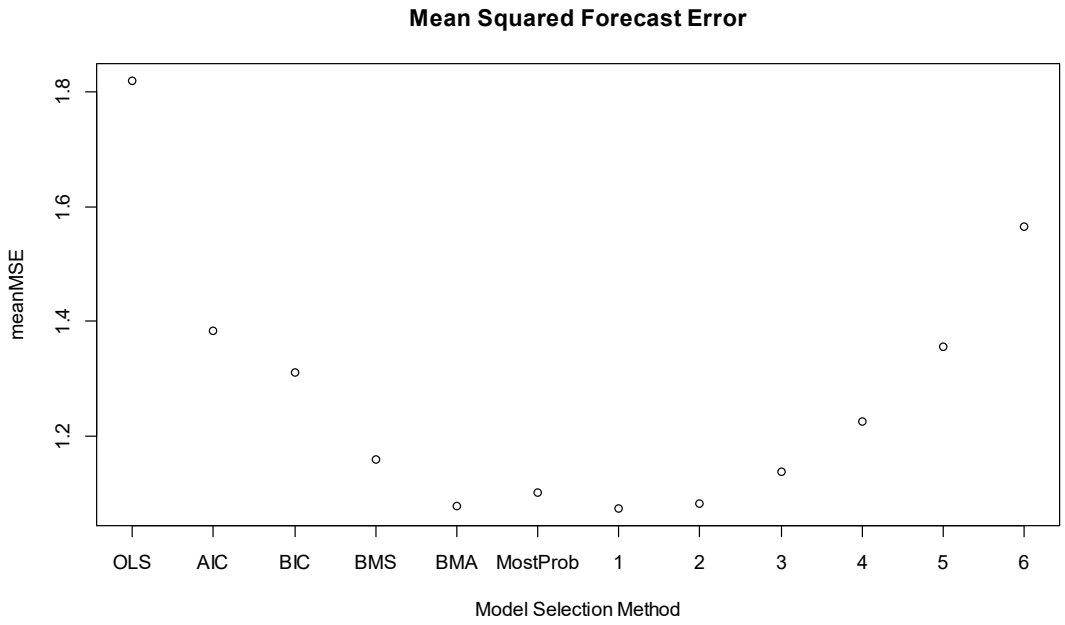
Table 5.8: Frequency of fixed k having lowest MSE (among $k = 1, \dots, K$) for $K = 6$, β_0 not local to zero

5.2.2.5 Case 5: $K = 6$ and $T = 20$

In this case, we keep the baseline configuration of the simulation as in section 5.2.2.2, however modifying the number of observations, so that we retain 16 in-sample and 4 out-of-sample observations, in an effort to catch a scheme with a limited amount of available observations. Such a case involves a high risk of overfitting, in the sense that the fitted model may extract some of the residual variation as if it represented underlying model structure. This way, the analysis corresponds too closely to the in-sample data, thus failing to built a model that has approximately equal in and out-of-sample error. As shown in table 5.9 and graph 5.5, the best performing methods of the alternatives to which subset regression are being compared for this scenario, are the Bayesian methods (BMA, most probable model, BMS). The best subset, in terms of mean MSE, is the one with the univariate model, outperforming all the alternatives, followed by the one with $k = 2$, then $k = 3$ and so on. It is obvious that the predictive performance of subset regressions deteriorates as k gets larger. This is a natural result, since models with fewer variables avoid the aforementioned overfitting pitfall and therefore can produce better forecasts, due to restricted estimation error.

Method	MSE
CSR, k=1	1,074185
BMA	1,078756
CSR, k=2	1,082716
Most Probable	1,101380
CSR, k=3	1,137679
BMS	1,158955
CSR, k=4	1,225797
BIC	1,311514
CSR, k=5	1,356598
AIC	1,384287
CSR, k=6	1,564902
OLS	1,819851

Table 5.9: MSE performance in ascending order for $K = 6, T = 20$



Graph 5.5: MSE performance for $K = 6, T = 20$

Fixed k value	Frequency
1	0,4451
2	0,0912
3	0,0721
4	0,1133
5	0,0759
6	0,2024

Table 5.10: Frequency of fixed k having lowest MSE (among $k = 1, \dots, K$) for $K = 6$, $T = 20$

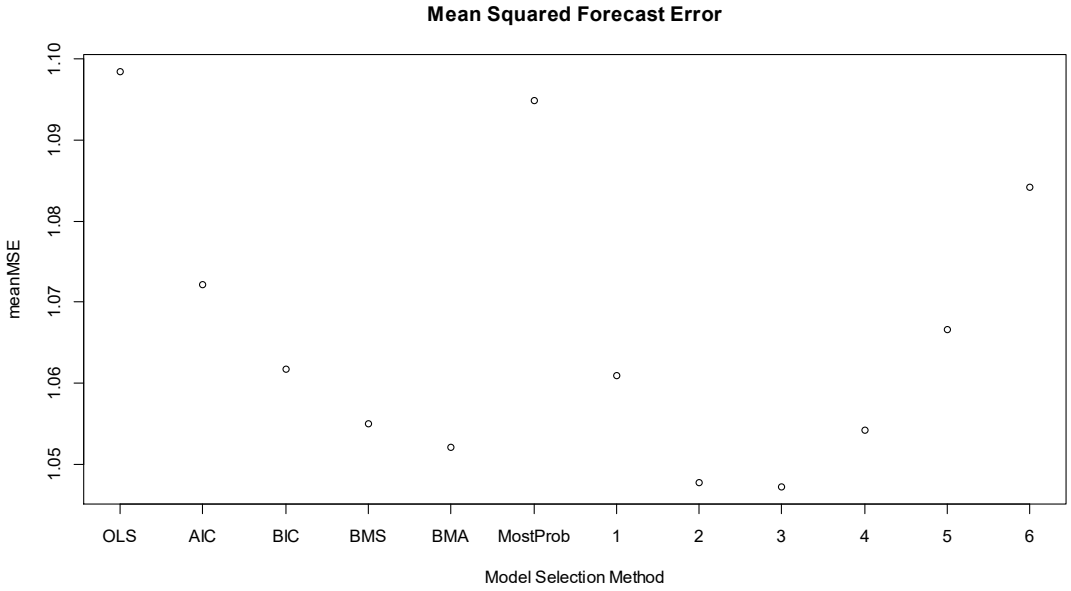
5.2.2.6 Case 6: $K = 6$ and $\theta = 1/3$

In this case, we keep the baseline configuration of the simulation as in section 5.2.2.2, however modifying the probability θ of inclusion for each regressor in the model: to slightly penalize models with too many predictors, we set $\theta = 1/3$, as opposed to $\theta = 1/2$, that was adopted in all other variations, reflecting complete prior ignorance about the model specification. As shown in table 5.11 and graph 5.6, the best subset, in terms of mean MSE, is the one with $k = 3$, followed by the one with $k = 2$. These subsets perform better than any other alternative, the best of which is the BMA. Comparing the results of this scenario to the one of section 5.2.2.2, it is obvious that the results are quite similar, with a small, though observable, increase in the frequency that smaller $k's$ produce lower MSE

predictions (as taken from comparison of tables 5.4 and 5.12), thus confirming the fact that $\theta = 1/3$ slightly favors smaller models.

Method	MSE
CSR, k=3	1,047178
CSR, k=2	1,047769
BMA	1,052092
CSR, k=4	1,054230
BMS	1,055040
CSR, k=1	1,060947
BIC	1,061724
CSR, k=5	1,066647
AIC	1,072180
CSR, k=6	1,084179
Most Probable	1,094852
OLS	1,098507

Table 5.11: MSE performance in ascending order for $K = 6, \theta = 1/3$



Graph 5.6: MSE performance for $K = 6, \theta = 1/3$

Fixed k value	Frequency
1	0,3534
2	0,1784
3	0,1359
4	0,1018
5	0,0721
6	0,1584

Table 5.12: Frequency of fixed k having lowest MSE (among $k = 1, \dots, K$) for $K = 6$, $\theta = 1/3$

Chapter 6: Application to real data

6.1 Data set and main considerations

In order to illustrate the complete subset regression approach to forecast combination and to compare its performance against that of other approaches, this chapter houses an empirical application to US stock returns. This application is well suited for our analysis because there is a great amount of uncertainty about which, if any, predictors help forecast stock returns. The data we consider are taken from Goyal and Welch (2008), updated to 2010, and are recorded at the quarterly horizon over the period 1947:1 to 2010:4 (Goyal and Welch provide a detailed description of transformations and data sources). The dependent variable is always the equity premium, that is, the total rate of return on the stock market (S&P 500 index returns, including dividends) minus the prevailing short-term interest rate (the Treasury-bill rate).

The candidate independent variables can be divided into three sets. The first one involves primarily stock characteristic variables:

- Dividend Price Ratio (D/P) is the difference between the log of dividends (paid on the S&P 500 index) and the log of stock prices, where dividends are measured using a one-year moving sum.
- Dividend Yield (D/Y) is the difference between the log of dividends and the log of lagged stock prices.
- Earnings Price Ratio (E/P) is the difference between the log of earnings and the log of stock prices, where earnings are measured using a one-year moving sum.
- Dividend Payout Ratio (D/E) is the difference between the log of dividends and the log of earnings.
- Stock Variance (SVAR) is the sum of squared daily returns.
- Book-to-Market Ratio (B/M) is the ratio of book value to market value for the Dow Jones Industrial Average.
- Net Equity Expansion (NTIS) is the ratio of twelve-month moving sums of net issues by NYSE-listed stocks to total end-of-year market capitalization of NYSE stocks.

The second set of candidate independent variables is interest-rate related:

- Treasury Bill Rate (TBL) is the interest rate on a three-month Treasury bill (secondary market).

- Long-term Yield (LTY) is the long-term government bond yield.
- Long-term Return (LTR) is the return on long-term government bonds.
- Term Spread (TMS) is the difference between the long-term yield and the Treasury bill rate.
- Default Yield Spread (DFY) is the difference between BAA- and AAA- rated corporate bond yields.
- Default Return Spread (DFR) is the difference between long-term corporate bond and long-term government bond returns.

The last set of candidate independent variables involves features of the overall macroeconomic environment:

- Inflation (INFL) is the Consumer Price Index (all urban consumers) from the Bureau of Labor Statistics.
- Investment-to-Capital Ratio (I/K) is the ratio of aggregate (private nonresidential fixed) investment to aggregate capital for the entire economy.

Our forecasting experiment is conducted on a quarterly basis having available a data span from 1947:1 to 2010:4, that is, 256 observations for the dependent and the candidate independent variables. For any scenario examined, the in-sample as well as the out-of-sample data span is always taken to be an integer multiple of four quarters (an annual period). We group the data this way, treating the quarter of the year as a blocking factor, accounting for the potential variability between the four quarters of each year. Thus, we may reduce any unexplained variability that is not of our primary interest.

An essential point of interest is the estimation period: it is not always clear how to choose the periods over which a regression model is estimated and subsequently evaluated. Although any choice is necessarily ad-hoc in the end, the criteria are clear. It is important to have enough initial data to get a reliable regression estimate at the start of the evaluation period, and it is important to have an evaluation period that is long enough to be representative. Using different periods reflects different trade-offs between the desire to obtain statistical power and the desire to obtain results that remain relevant today. The main results herein present an experimental basis of an 18-year in-sample period, consisting of 72 observations, and out-of sample predictions for a whole year (that is 4 observations/quarters), using a moving window scheme. As an alternative scheme, we briefly give the conclusions based on an in-sample period of 28 years. It is worth to note that the available dataset has already ignored periods where some variables did not have complete data

(occurrences documented before 1947), finally providing a set where no such peculiarity exists.

As a first step, we examine our initial pool of covariates thoroughly. Checking Pearson's correlation between all pairs of independent variables, and for a variety of data spans (mainly focusing in the 18 to 28 year range, which covers our experimental basis) we find that there are several cases where the correlation is higher than 0,9, and sometimes even closer to 1. In particular, we come across the fact that the variables D/P and D/Y is such a pair and, moreover, that D/E and TMS have great linear dependence (they are strongly correlated) with other variables. For this reason, the information given by these variables (D/E, TMS and one of D/P or D/Y) is already contained in the other twelve variables and is thus redundant, so that their coefficients in the regression model cannot be defined. Essentially, we reject D/P, D/E and TMS and we proceed with the remaining twelve independent variables. Moreover, in the results we include the case where D/Y is rejected in place of D/P.

Furthermore, we take into account Goyal and Welch (2008) who summarize that models can be significant only for certain periods of time and that the few models that still are, may usually fail simple regression diagnostics. For this reason, we explore the (in-sample) statistical significance of the regression covariates, as well as the proportion of the variance in the dependent variable that is predictable from the independent variables, using for this purpose the ${R^2}_{adj}$. We use a representative sample of ten overlapping periods, each one having a range of eighteen years. For each period, we fit the full model, that is the model with all thirteen variables, however excluding D/E and TMS (the design matrix is not invertible and therefore cannot be used to develop a regression model, resulting from linearly dependency that these two strongly correlated variables have). We also fit a model with twelve variables, further excluding D/P. We then report ${R^2}_{adj}$ for these two cases, as well as for the model that the stepwise model selection method suggests.

6.2 Results

In line with the simulation study of chapter 5, we launch the complete subset regression method, getting twelve different values of forecasting performance (for $k = 1, \dots, K = 12$). Once again, the Mean Squared (Forecasting) Error is used to represent this performance. In order to compare the fixed k proposals we once again, as in the simulation, calculate the predictive performance of:

1. The model estimated by the OLS method.

2. The “best” models, identified by the AIC and the BIC criterion. That is, we calculate the AIC & BIC for each model, select the best models and then make the prediction based on them.
3. Bayesian Model Averaging: we use the same setup as in chapter 5 and the predictions are obtained by weighting each model’s forecast by its posterior probability.
4. The most probable model, as identified by the Bayesian Model Averaging calculations. That is, we select the model with the highest posterior probability $f(M_j|y)$ and then make the prediction based on it solely.
5. Bayesian Model Sampling and Averaging: we use the same setup as in chapter 5.

The results are summarized in table 6.1 and graph 6.1 and have been calculated as follows: for each method (OLS, AIC, BIC, BMA, BMS, Most Probable and for each $k = 1, \dots, K = 12$) we first get an annual forecast for the year 1955, based on the in-sample span of 1947:1 to 1954:4 and we calculate the average MSE of the forecasts (of its four quarters). We then use the moving window scheme and each time we get another annual prediction (and thus a MSE value for each method). We finally average across all these years that we have made predictions for, that is from 1955 to 2010.

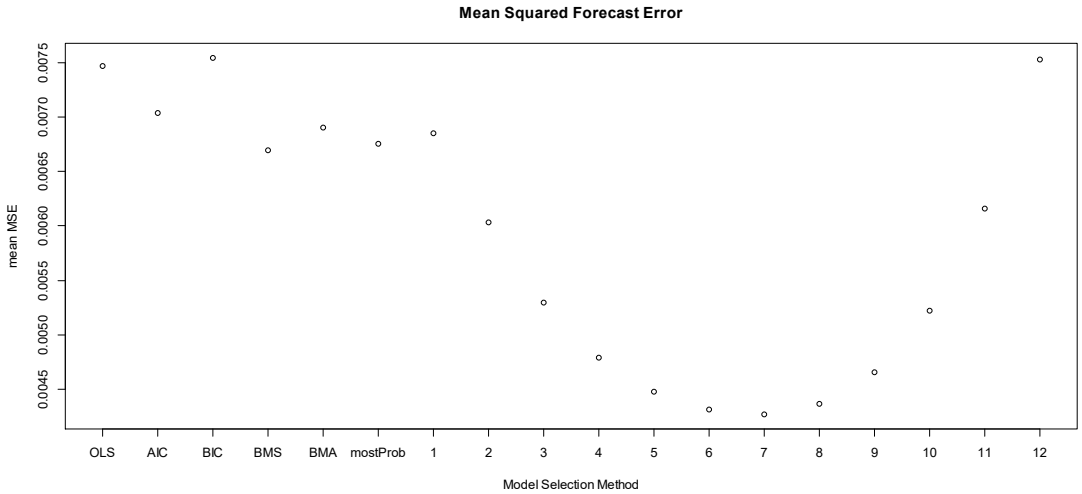
As shown in table 6.1 and graph 6.1 the best performing method, producing the lowest out-of-sample MSE on average, of the ones used to benchmark the subset regression, is the Bayesian Model Sampling, followed by the Most Probable model and the Bayesian Model Averaging. The best subset, in terms of mean MSE, is the one with $k = 7$ and performs better than any other alternative. In particular, for $k = 2, \dots, 11$ the subset regression produces lower out-of-sample MSE on average, than any other alternative. The superiority of the subset regression can be summarized in the following facts:

- For most of the fixed k values, it provides better forecasts (with lower MSE).
- The best $k = 7$, provides a substantially smaller MSE (on average) than the best alternative.
- Using the best fixed value, $k = 7$, we only need to combine (average over) $\binom{12}{7}$=792 models, which is a very small fraction compared to the totality of the 4.096 models. This is a significant saving in computational terms, comparing with any method that would average over all models.

Moreover, for $k \leq 11$ the k-variate combinations produce better results than models selected by recursively applying information criteria such as the AIC or the BIC. This happens despite the fact that these subset combinations may contain the same or even larger number of predictors, since, on average, the AIC and the BIC criteria select 7,82 and 6,58 predictors, respectively.

Method	MSE
CSR, k=7	0,004269
CSR, k=6	0,004309
CSR, k=8	0,004366
CSR, k=5	0,004476
CSR, k=9	0,004657
CSR, k=4	0,004791
CSR, k=10	0,005224
CSR, k=3	0,005297
CSR, k=2	0,006029
CSR, k=11	0,006157
BMS	0,006692
Most Probable	0,006752
CSR, k=1	0,006853
BMA	0,006904
AIC	0,007037
OLS	0,007472
CSR, k=12	0,007528
BIC	0,007545

Table 6.1: MSE performance in ascending order for an 18-year in-sample period

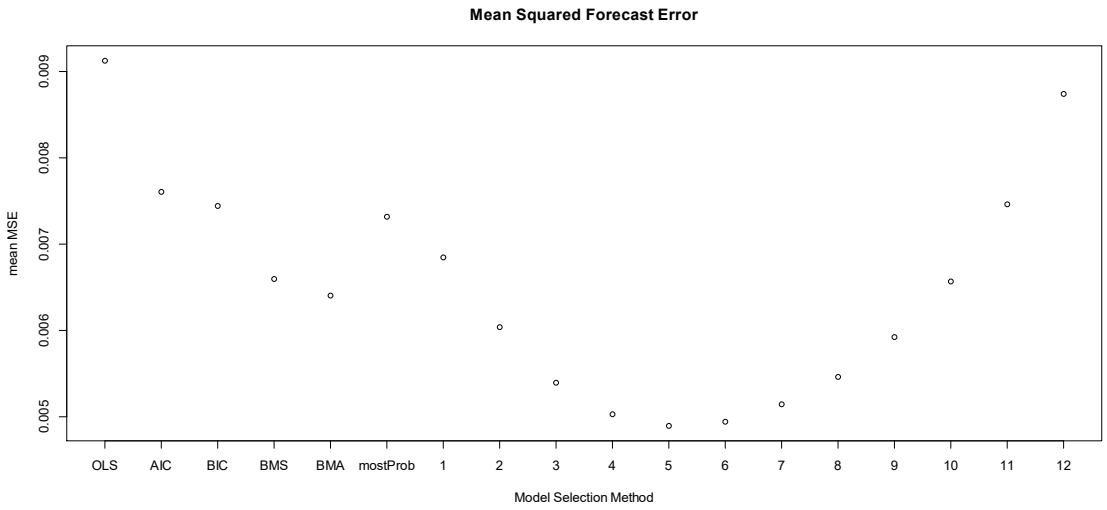


Graph 6.1: MSE performance for an 18-year in-sample period

Table 6.2 and graph 6.2 summarize the results of the variation where twelve variables are included in the model, however having retained D/P instead of D/Y. In this case, for $k = 5,6,4,7,3,8,9,2$ (and with this order) the subset regression is superior in forecasting compared to any other alternative, the best ones being the BMA, followed by the BMS. The similarity of these results, compared to the ones table 6.1 and graph 6.1, concerning the superiority of subset regressions against the alternatives, is obvious.

Method	MSE
CSR, k=5	0,00489
CSR, k=6	0,004941
CSR, k=4	0,005022
CSR, k=7	0,005136
CSR, k=3	0,005389
CSR, k=8	0,005459
CSR, k=9	0,00592
CSR, k=2	0,00604
BMA	0,006407
CSR, k=10	0,006563
BMS	0,006594
CSR, k=1	0,006847
Most Probable	0,007322
BIC	0,007444
CSR, k=11	0,007466
AIC	0,007610
OLS	0,007931
CSR, k=12	0,008742

Table 6.2: MSE performance in ascending order for an 18-year in-sample period (twelve variables, D/P included)



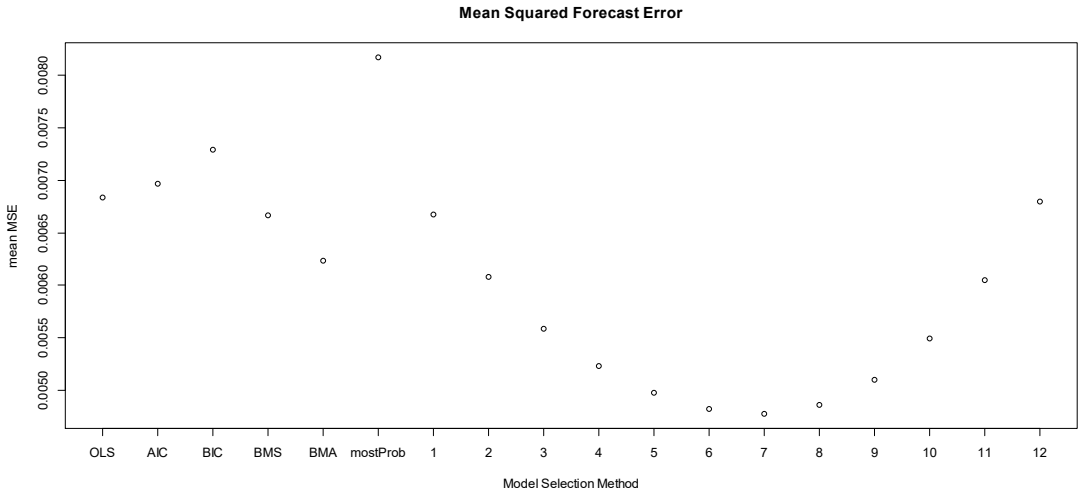
Graph 6.2: MSE performance for an 18-year in-sample period (twelve variables, D/P included)

Table 6.3 and graph 6.3 summarize the results of the case that we use a 28-year in-sample period, instead of an 18-year one that was adopted in the main scenario. In this case the best performing method, producing the lowest out-of-sample MSE on average, of the ones used to benchmark the subset regression, is the Bayesian Model Averaging. The best subset, in terms of mean MSE, is the one with $k = 7$ and performs better than any other alternative. In particular, for $k = 2, \dots, 11$ the subset regression produces lower out-of-sample MSE on average, than any other alternative. The superiority of the subset regression can be summarized into the fact that for most of the fixed k values it provides better forecasts (with lower MSE) as well as to the promising reduction of computational needs, since using the best fixed value, $k = 7$, we only need to combine (average over) $\binom{12}{7}$=792 models, which is a very small fraction compared to the totality of the 4.096 models. The similarity of these results, compared to the ones presented in table 6.1 and graph 6.1, concerning the superiority of subset regressions against the alternatives, is obvious.

Method	MSE
CSR, k=7	0,004776
CSR, k=6	0,004823
CSR, k=8	0,004861
CSR, k=5	0,004980
CSR, k=9	0,005097
CSR, k=4	0,005233
CSR, k=10	0,005493
CSR, k=3	0,005589

CSR, k=11	0,00605
CSR, k=2	0,006079
BMA	0,006234
BMS	0,006667
CSR, k=1	0,006673
CSR, k=12	0,006797
OLS	0,006839
AIC	0,006965
BIC	0,007294
Most Probable	0,008173

Table 6.3: MSE performance in ascending order for a 28-year in-sample period



Graph 6.3: MSE performance for a 28-year in-sample period

Table 6.4 houses the results of the considerations stated in section 6.1, concerning the (in-sample) statistical significance of the regression covariates, as well as the proportion of the variance in the dependent variable that is predictable from the independent variables, using for this purpose the $R^2{}_{adj}$. Summarizing, the main results are the following two: the proportion of the variance explained by the regression is very high, for all the aforementioned cases. At the same time, in-sample statistical significance of each covariate fluctuates across the period that the data set covers, thus amplifying the value of taking into account all these twelve variables, as already stated.

	Full Model (thirteen variables)	Model selected by stepwise	Final model (twelve variables)	Independent variables (first row for each in-sample period is the p-value when fitting the full model and the second row indicates selection of this variable by the stepwise method). Blanks indicate insignificance and lack of selection, respectively												
In-sample Period	R^2 adj	R^2 adj	R^2 adj	D/P	D/Y	E/P	SVAR	B/M	NTIS	TBL	LTY	LTR	DFY	DFR	INFL	I/K
1947:1 - 1964:4	99,69	99,72	71,09	0	0			0,02	0,126	0,084						
				x	x			x	x	x						
1952:1 - 1969:4	99,9	99,9	77,08	0	0					0,075		0,027		0,028	0,009	0,733
				x	x					x		x		x	x	x
1957:1 - 1974:4	99,96	99,97	77,69	0	0					0		0,01		0,037	0,182	0,201
				x	x					x		x		x		x
1962:1 - 1979:4	99,98	99,98	78,47	0	0	0,465				0		0,205	0,088	0,062	0,03	
				x	x	x				x		x		x	x	
1967:1 - 1984:4	99,97	99,97	72,02	0	0					0		0		0,068	0,01	
				x	x					x		x		x		
1972:1 - 1989:4	99,97	99,97	82,54	0	0					0		0,002		0,06	0,039	
				x	x					x		x		x		
1977:1 - 1994:4	99,97	99,96	79,88	0	0		0,076			0		0,001		0,207	0,035	
				x	x		x			x		x		x		
1982:1 - 1999:4	99,97	99,97	63,67	0	0				0,074	0	0,144					
				x	x				x	x	x					
1987:1 - 2004:4	99,97	99,97	57,01	0	0	0,007			0,003	0			0,061		0,163	0,305
				x	x	x			x	x			x			x
1992:1 - 2009:4	99,96	99,96	57,32	0	0			0,037		0	0,137		0,119	0,001		
				x	x			x		x	x		x	x		

Table 6.4: Statistical significance and R^2 adj for representative in-sample periods

Chapter 7: Conclusion

This dissertation has explored the main theoretical features, as well as the performance of the complete subset regression, a forecast combination approach that averages forecasts across complete subset regressions with the same number of predictor variables and thus the same degree of model complexity.

Based on the simulation, as well as the empirical study, we find that the subset regression appears to perform quite well when compared to competing approaches such as Bayesian Model Averaging, or the models selected based on the OLS, the AIC and the BIC criterion. This is more evident in cases where a shrinkage method is supposed to add value, that is, where β_0 (the pseudo true value of the regression coefficients, β) has values local to zero. In many cases subset regression combinations amount to a form of shrinkage, but one that is more general than the conventional variable-by-variable shrinkage.

Especially for the data set of chapter 6, where most models are unstable or even spurious (Goyal and Welch, 2008) and, thus, relying on a single model is not a solid option, a method that is based on averaging, with the additional benefit of being able to confine the total number of models to average over, seems really promising.

This method could also be adopted in order to reduce the computational effort, or even to make an infeasible analysis feasible, in cases where the number of covariates prohibits analyzing all of them.

Appendix A: Main distributions

A.1 Normal Distribution

The Normal (or Gaussian) Distribution in a variate X is a statistic distribution with density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}},$$

on the domain $x \in (-\infty, +\infty)$, where μ is the mean (as well as the median and the mode) of the distribution, σ is the standard deviation and σ^2 is the variance. A graph of the probability density function of the Normal distribution for several values of μ and σ^2 is given in the Figure A.1.

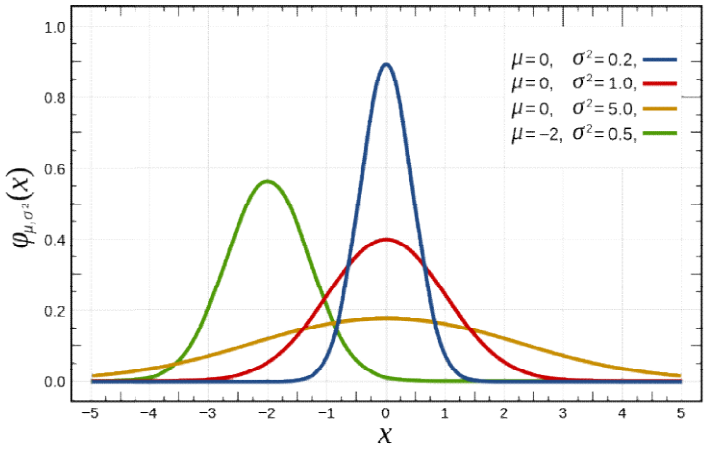


Figure A.1: Normal Probability Density Function

A.2 Multivariate Normal Distribution

The Multivariate Normal Distribution is a generalization of the one dimension (univariate) Normal distribution to higher dimensions. A random vector $X = (X_1, \dots, X_k)$ is said to de k −variate normally distributed if it has density function:

$$f(x; \mu, \Sigma) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{(x-\mu)'\Sigma^{-1}(x-\mu)\},$$

where $\mu = (\mu_1, \dots, \mu_k)$ is the mean vector and $\Sigma = E[(x-\mu)(x-\mu)']$ is the covariance matrix. A graph of the probability density function of a bivariate Normal distribution (that is $k = 2$) is given in the Figure A.2.

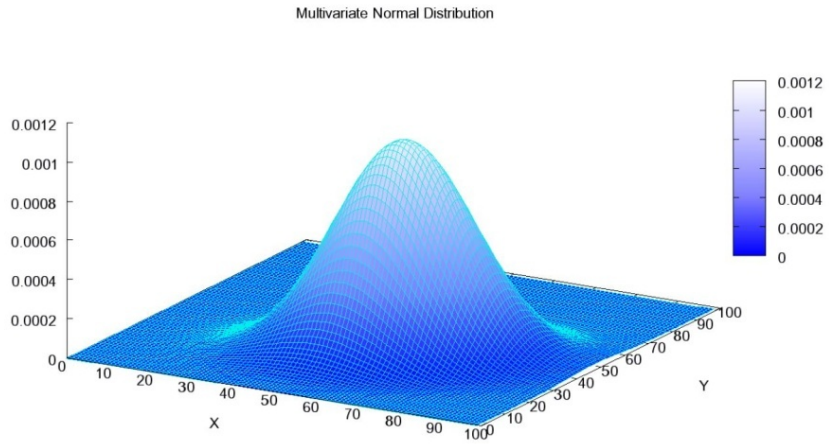


Figure A.2: Multi(bi)variate Normal Probability Density Function

A.3 Gamma Distribution

The Gamma Distribution is a two-parameter family of continuous distributions, with three different parameterizations in common use:

1. With a shape parameter $k > 0$ and a scale parameter $\theta > 0$.
2. With a shape parameter $\alpha = k$ and an inverse scale parameter $\beta = \frac{1}{\theta}$, called a rate parameter.
3. With a shape parameter k and a mean parameter $\mu = \frac{k}{\beta}$.

For the second case, the Gamma distribution in a variate X is a statistic distribution with density function:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)},$$

on the domain $x \in [0, +\infty)$, where $\Gamma(x)$ is a complete gamma function. The mean of the distribution is $E[x] = \frac{\alpha}{\beta}$ and the variance is $var(x) = \frac{\alpha}{\beta^2}$. The exponential distribution, the Erlang distribution and the chi-squared distribution, are special cases of the Gamma distribution. A graph of the probability density function of the first parameterization case of a Gamma distribution for several values of k and θ is given in the Figure A.3.

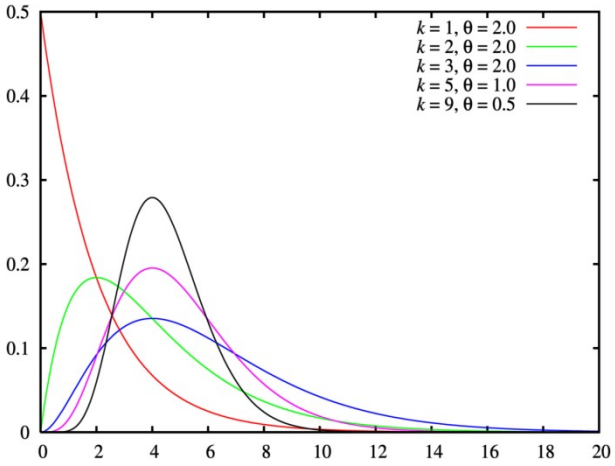


Figure A.3: Gamma Probability Density Function

A.4 Inverse Gamma Distribution

The Inverse Gamma Distribution is a two parameter family of continuous probability distributions commonly used in Bayesian statistics. It is the distribution of the reciprocal of a variable distributed according to the Gamma distribution. The Inverse Gamma distribution in a variate X, with a shape parameter α and a scale parameter β is a statistic distribution with density function:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{-\alpha-1} e^{-\frac{\beta}{x}}}{\Gamma(\alpha)},$$

on the domain $x \in [0, +\infty)$, where $\Gamma(x)$ is a complete gamma function. The mean of the distribution is $E[x] = \frac{\beta}{\alpha-1}$, for $\alpha > 1$ and the variance is $var(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, for $\alpha > 2$. Note that if $X \sim Gamma(\alpha,\beta)$ then $X^{-1} \sim IG(\alpha,\beta)$. A graph of the probability density function of an Inverse Gamma distribution for several values of α and β is given in the Figure A.4.

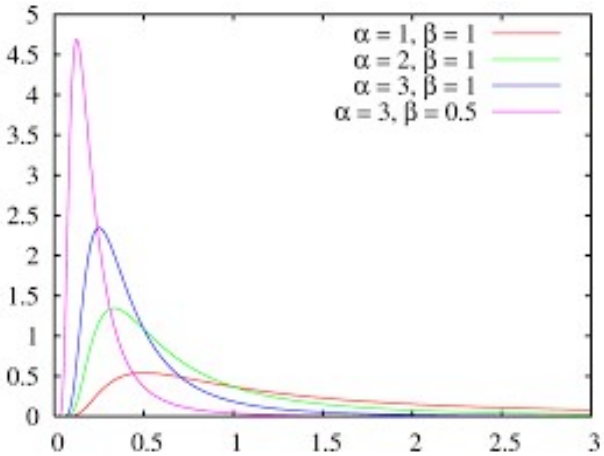


Figure A.4: Inverse Gamma Probability Density Function

A.5 Student's *t* Distribution

The student's t distribution is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population, for cases where the standard deviation is unknown. The *t* distribution in a variate X, has a density function:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

on the domain $x \in (-\infty, +\infty)$, where ν is the number of degrees of freedom and $\Gamma(x)$ is a complete gamma function. The mean of the distribution is zero if $\nu > 1$ (otherwise undefined) and the variance is $var(x) = \frac{\nu}{(\nu-2)}$, for $\nu > 2$ and ∞ for $1 < \nu \leq 2$ (otherwise undefined). The t distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tales, being more prone to producing values that fall far from its mean and moreover varies based on the number of degrees of freedom. A graph of the probability density function of a *t* distribution for two different values of degrees of freedom, compared to the graph of a normal distribution is given in the Figure A.5.

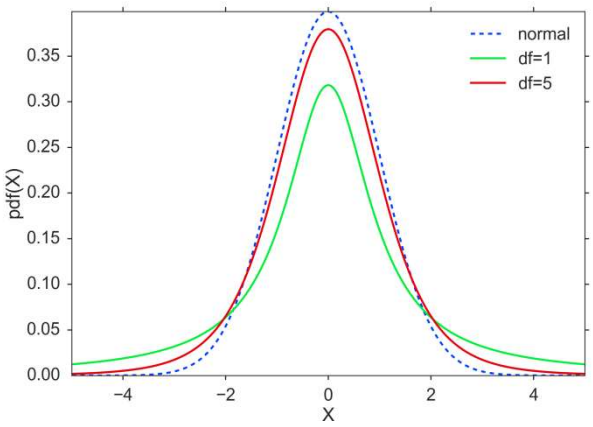


Figure A.5: Student's *t* Probability Density Function

A.6 Multivariate *t* Distribution

The Multivariate t Distribution is a generalization of the one dimension (univariate) student's t distribution to higher dimensions. A random vector $X = (X_1, \ldots, X_k)$ is said to de k −variate *t* distributed if it has density function:

$$f(x;\mu,\Sigma) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}\left[1+\frac{1}{\nu}(x-\mu)'\Sigma^{-1}(x-\mu)\right]^{-\frac{\nu+k}{2}},$$

where $\mu = (\mu_1, \dots, \mu_k)$ is the location parameter, Σ is the shape matrix and ν is the number of degrees of freedom. The mean of the distribution (as well as the median and the mode) is μ if $\nu > 1$ (otherwise undefined) and the variance is $var(x) = \frac{\nu}{(\nu-2)}\Sigma$, for $\nu > 2$ (otherwise undefined).

Appendix B: Statistical Software

The statistical analysis conducted in this thesis, including the simulation study, was based on free licensed software. Free software means that users of a program have the following four essential freedoms:

1. The freedom to run the program as they wish, for any purpose.
2. The freedom to study how the program works, and adapt it to their needs. Access to the source code is a precondition for this.
3. The freedom to redistribute copies.
4. The freedom to improve the program, and release their improvements to the public, so that the whole community benefits. Access to the source code is again a prerequisite.

In particular, the R statistical programming language has been used, for data manipulation, calculation and graphical display, including some available intermediate tools for data analysis, as well packages that extend the main features.

B.1 The R statistical programming language

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. R is not solely a statistics system. It is an environment within which statistical techniques are implemented.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. Moreover, it can be extended easily via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of internet sites covering a very wide range of modern statistics. All necessary information can be found in https://www.r-project.org/.

Appendix C: References

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. Proceedings of the 2nd International Symposium on Information Theory, pp. 267-281.

Elliott, G., Gargano, A. & Timmermann, A. (2013). *Complete Subset Regressions*. Journal of Econometrics, Vol. 177, pp. 357-373.

Fernandez, C., Ley, E. & Steel, M.F.J. (2001a). *Benchmark Priors for Bayesian Model Averaging*. Journal of Econometrics, Vol. 100, pp. 381-427.

Fernandez, C., Ley, E. & Steel, M.F.J. (2001b).*Model Uncertainty in Cross-Country Growth Regressions*. Journal of Econometrics, Vol. 16, pp. 381-427.

Geweke, J. & Amisano, G. (2011). *Optimal Prediction Pools.* Journal of Econometric, Vol 164, pp. 130-141.

Goyal, A. & Welch, I. (2008). *A Comprehensive Look at the Empirical Performance of Equity Premium Prediction*. Review of Financial Studies, Vol. 21, pp. 1455-1508.

Hocking, R. (1976). *A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression*. Biometrics, Vol. 32, pp. 1-49.

Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. (1999). *Bayesian Model Averaging: a Tutorial.* Statistical Science, Vol. 14, No. 4, pp. 382-417.

Jeffreys, H. (1946). *An Invariant Form for the Prior Probability in Estimation Problems.* Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 186, No 1007, pp. 453-461.

Kass, R. & Wasserman, L. (1996). *The Selection of Prior Distributions by Formal Rules.* Journal of the American Statistical Association, Vol. 91, No. 435, pp. 1343-1370.

Ley, E. & Steel, M.F.J. (2009). *On the Effect of Prior Assumptions in Bayesian Model Averaging With Applications to Growth Regression*. Journal of Applied Econometrics, Vol. 24, pp. 651-674.

Madigan, D., Gavrin, J. & Raftery, A.E. (1995). *Eliciting Prior Information to Enhance the Predictive Performance of Bayesian Graphical Models*. Journal of the American Statistical Association, Vol. 24, Issue 9, pp. 2271-2292.

Mallows, C.L. (1973). *Some Comments on Cp*. Technometrics, Vol. 15, No. 4, pp. 661-675.

Rapach, D., Neely, C., Tu, J., Zhou, G. (2010). *Out-of-sample Equity Premium Prediction: Economic Fundamentals vs. Moving-Average Rules.* Review of Financial Studies, Vol. 23, pp. 821-862.

Sawa, T. (1978). *Information Criteria for Discriminating Among Alternative Regression Models*. Econometrica, Vol. 46, Issue 6, pp. 1273-1291.

Schwarz, G. (1978). *Estimating the Dimension of a Model*. The Annals of Statistics, Vol. 6, No 2, pp. 461-4644.