Ridge Estimation and its Modifications for Linear Regression with Deterministic or Stochastic Predictors

James Younker

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

A common problem in multiple regression analysis is having to engage in a bias-variance trade-off in order to maximize the performance of a model. A number of methods have been developed to deal with this problem over the years with a variety of strengths and weaknesses. Of these approaches the ridge estimator is one of the most commonly used. This paper conducts an examination of the properties of the ridge estimator and several alternatives in both deterministic and stochastic environments. We find the ridge to be effective when the sample size is small relative to the number of predictors. However, we also identify a few cases where some of the alternative estimators can outperform the ridge estimator. Additionally, we provide examples of applications where these cases may be relevant.

Contents

List of Tables

Introduction

In this thesis we deal with a multiple regression problem. Statistical inference in such problems is often accompanied by the 'bias-variance trade-off'. For a fixed number of observations every additional explanatory variable typically causes coefficient estimates to become less certain. As confidence intervals around these estimates expand, eventually interpretation and forecasting power are jeopardized. In extreme cases, estimates can no longer be obtained; for example in a linear parametric regression with more parameters than observations. This often forces models to either omit some informative variables, or to employ constraints in the estimation process. One of the most common ways to deal with this problem is the ridge estimator. We examine the properties of the ridge estimator in both a deterministic and stochastic predictor context, and compare it against some alternatives.

In the first chapter, we review how the bias-variance trade-off has been addressed by constrained estimation techniques. We then summarize the estimation methods of ordinary least squares, the ridge regression as well as many alternatives.

In chapter two we examine the theoretical properties of the ridge estimator, and consider a modification of the ridge estimator. It should be noted that although the formulas for mean square error of OLS (ordinary least squares) and ridge estimator are well-known for the case of deterministic predictors (see Section 2.1.1), to the best of our knowledge, there are no formulas in the case of stochastic explanatory variables. Therefore, in such cases, one has to rely on simulation studies.

Consequently, in chapter three we illustrate the properties of the ridge and several competing estimators through the use of computer simulations. We find that the ridge is generally very competitive when the sample size is small relative to the number of predictors. However, there are a few special cases where the ridge is outperformed by alternatives, particularly in situations involving errors-in-variables, very small samples, or predictors with a low amount of correlation.

The final section of the paper presents examples of some applications, where these special cases are potentially meaningful.

Finally, in the appendix we include some simulation codes in `R` (statistical language) and `EVIEWS` (econometrics language).

We summarize our contribution as follows:

- We compared the ridge and several alternative estimators in situations of deterministic or stochastic predictors, including errors-in-variables specifications.
- We considered the model averaging family of estimators as an alternative to the ridge.
- We proposed a modification of ridge estimator, where instead of increasing diagonal entries in $\left(\mathbf{X}^\top\mathbf{X}\right)$ we decrease the off diagonal entries.
- We compare several methods of selecting the tuning parameter in ridge regression to an optimal selection from Monte Carlo simulation. We conclude that existing methods could be improved.
- We conducted an extensive simulation study for the ridge and alternative estimators, where we identified a few cases where the ridge is outperformed by some of the alternatives.
- We provided examples of applications where these cases are potentially meaningful.

Chapter 1

Multiple Linear Regression

1.1 Definition of a Model

Throughout the thesis we assume that we observe a vector of dependent variables

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

together with a matrix of predictors

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

The matrix is also called the design matrix. It can be deterministic or random. We consider the multiple linear regression model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1.1}$$

where $\beta_1, \ldots, \beta_p$ are parameters to be estimated and ε_i, $i = 1, \ldots, n$, are independent, identically distributed normal random variables with mean 0 and variance σ^2. The

regression model can be written as

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

where

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Finally, we assume from the beginning that all random variables are centered.

1.2 Bias-Variance Trade-off

If θ is a parameter and $\hat{\theta}$ is its estimator, then we have the following formula for the mean square error:

$$E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + \Big(E[\hat{\theta}] - \theta\Big)^2.$$

This formula is called the variance-bias decomposition. In many statistical problems, the goal is to minimize the mean square error, which amounts to finding a bias-variance trade-off. As ordinary least squares is unbiased and one of the most commonly used estimators, many of the alternative methods attempt to reduce the variance proportion of the bias-variance trade-off in order to minimize the mean squared error of a regression. This is typically done by either excluding explanatory variables from a model or employing a constraint in the estimation or a combination of the two. A traditional example of such an approach is the ridge estimator. However, there are also a vast number of alternative estimators. Our review of previous work will begin with the ordinary least squares estimator and the ridge estimator, then we will examine some of the alternatives.

1.3 Least Squares Estimator

The ordinary least squares linear regression (OLS) is one of the oldest and most commonly used approaches in multiple regression. The estimator relates the dependent variable to a set of explanatory variables. In particular, if a model is constructed from variables with mean zero, then the estimator takes the covariance between the explanatory and dependent variables $\mathbf{X}^\top\mathbf{Y}$, and scales it by the inverse of the variance-covariance matrix of the explanatory variables $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Therefore, the OLS estimator is defined as follows:

$$\hat{\beta}_{\text{OLS}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

1.3.1 Least Squares under Deterministic Predictors

When the least squares estimator is applied to deterministic predictors the estimator's characteristics are well known. The resulting estimates are consistent and unbiased. Additionally, the estimator is the most efficient among the set of linear and unbiased estimators (Greene, 2008).

The OLS estimator can be shown to be unbiased by simply taking an expectation:

$$\begin{aligned} E\left[\hat{\beta}_{\text{OLS}}\right] &= E\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top E\left[\mathbf{Y}\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top E\left[\mathbf{X}\beta + \varepsilon\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta + \left(\mathbf{X}^\top\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^\top E[\varepsilon] \\ &= \beta. \end{aligned} \tag{1.3.1}$$

As ε has zero mean and independent of $\mathbf{X}$ equation (1.3.1) establishes that the OLS estimator is unbiased. When taken together with equation (1.3.2) which shows variance declining with sample size, this establishes that the OLS estimator is also consistent.

The variance-covariance matrix of $\hat{\beta}_{\text{OLS}}$ can be established as follows.

$$\begin{aligned} cov\left[\hat{\beta}_{\text{OLS}}\right] &= cov\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top cov\left[\mathbf{Y}\right]\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right]^\top \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top cov\left[\mathbf{Y}\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-\mathbf{1}}. \end{aligned}$$

Using the fact that $cov[\mathbf{Y}] = \sigma^{\mathbf{2}}\mathbf{I}$, we have

$$\begin{aligned} cov\left[\hat{\beta}_{\text{OLS}}\right] &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-\mathbf{1}} \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}. \end{aligned} \tag{1.3.2}$$

As $\hat{\beta}_{\text{OLS}}$ is a linear function of independent Gaussian variables it also has a Gaussian distribution. From (1.3.1) and (1.3.2) we have

$$\hat{\beta}_{\text{OLS}} \sim N\left(\beta, \sigma^{\mathbf{2}}\left(\mathbf{X}^\top\mathbf{X}\right)^{-\mathbf{1}}\right).$$

Of course, if errors ε_i are not normal but are independent and identically distributed, we can apply central limit theorem to conclude the above statement asymptotically.

1.3.2 Least Squares under Stochastic Predictors

When the least squares estimator is applied to stochastic predictors the theory is much more complicated. We are not aware of explicit formulas for either $\text{Var}[\hat{\beta}_{\text{OLS}}]$ or $E(\hat{\beta}_{\text{OLS}} - \beta)$. However, below is a result on consistency and asymptotic normality from Wei (1985), Lai and Wei (1982):

Let $\lambda_{\min}(n)$ and $\lambda_{\max}(n)$ are the minimal and the maximal eigenvalues of the variance-covariance matrix $\mathbf{X}^\top\mathbf{X}$.

Theorem 1.3.1 *If* $\lambda_{\min}(n) \to \infty$ *almost surely as* $n \to \infty$ *and* $\log(\lambda_{\max}(n)) = o(\lambda_{\min}(n))$*, then* $\hat{\beta}_{\text{OLS}}$ *is consistent. Furthermore,*

$$(\mathbf{X}^\top\mathbf{X})^{1/2}(\hat{\beta}_{\text{OLS}} - \beta) \to N(0, \sigma^2\mathbf{I}). \tag{1.3.3}$$

Least Squares under Errors-in-Variables

Errors-in-variables models form a sub set of stochastic predictor models. An errors-in-variables specification occurs when the true predictor variables X_i' and the observable predictor variable X_i differ by a random noise term η_i. Common examples of errors-in-variables applications include situations where a predictor is observed with measurement error or where an observed predictor serves as proxy for an unobservable series.

When an OLS estimator is applied to an errors-in-variables situation the resulting estimate is no longer consistent. However, the OLS estimate will converge to the value that is optimal for forecasting if the distribution of the noise term is unknown (Greene, 2008).

If the true model is the following:

$$Y_i = \beta_1 X_{i1}' + \cdots + \beta_p X_{ip}' + \varepsilon_i, \qquad i = 1, \ldots, n, \varepsilon_i \sim N(0, \sigma^2),$$

where an observable predictor equals the true predictor plus a measurement error:

$$X_i = X_i' + \eta_i, \qquad \varepsilon_i \sim N(0, \sigma_\eta^2),$$

then OLS estimator of β will be inconsistent and biased toward zero. However, although the OLS estimator is inconsistent it can be optimal from the perspective of minimizing a mean squared forecast error:

$$\hat{\beta} \to \beta \frac{\sigma^2}{\sigma^2 + \sigma_\eta^2}.$$

1.4 Classical Ridge Regression

A ridge regression is an ordinary least squares estimation, with a constraint on the sum of the squared coefficients. One of the original motivations for the ridge regression

was to ensure matrix invertibility when estimating a linear regression. However, it is most frequently used to reduce the variance of parameter estimates. The constraint is applied to the regression through the value chosen for the tuning parameter λ. As λ increases the regression parameters $\beta_1, \ldots, \beta_p$ are forced to smaller values with lower variance. Some explanatory variables may be completely excluded from the model as their parameters are forced to zero. The choice of λ can be made through minimizing prediction error or cross validation. A ridge regression estimation is dependent upon the scale and intercept of the model. As a consequence, variables are typically centered and standardized prior to model estimation. We follow this approach in our example. An alternative treatment of the intercept is to include it in the model but not make it subject to the ridge constraint; see Hastie, Tibshirani, Friedman, 2001; p 59. The most common choices for standardization are non-theoretical and often take the form of equalizing variances. However, the theoretical foundations of a model can suggest approaches to standardization; for instance converting variables to common units. The original publication of the ridge regression was Hoerl and Kennard (1970), and discussions can be found in several texts (Hastie, Tibshirani and Friedman, 2001; Lzenman, 2008; Greene, 2008). There are several variations of the classical ridge regression. These often involve multiple λ's or different approaches to the standardization of the explanatory variables. In mathematical terms, the ridge regression estimator is defined as

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right). \tag{1.4.1}$$

Equation (1.4.1) is equivalent to

$$\hat{\beta}_{\text{ridge}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{Y}. \tag{1.4.2}$$

Moreover, it is equivalent to

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq s^*$. There is one-to-one correspondence between s^* and λ.

From equation (1.4.2) one can see that a ridge regression can be expressed as modifying an OLS by adding λ to variance terms in the variance-covariance matrix of the explanatory variables. This modification impacts both the variance and covariance terms in the variance-covariance matrix, but disproportionately down weights the covariance terms. To offer some intuition behind this feature an example is provided below in the case of two explanatory variables. In this case, as λ increases, the variance terms are reduced by an order of λ and the covariance terms are reduced by an order of λ^2. However, the disproportionate down weighting of the covariance terms holds in any dimension. The easiest way to establish this is to utilize the fact that the ridge regression algorithm is equivalent to a Bayesian regression when the prior on the beta's is a multivariate normal distribution with mean zero, zero covariance and with variance proportional to the product of λ and the variable standardization scalar (see Hoerl and Kennard, p. 64, 1970; Hastie, Tibshirani and Friedman, p 60; Lzenman, p 139-140, 2001). Therefore, as the ridge regression is equivalent to applying a prior to beta parameters with zero covariance and positive variance, the ridge estimation also reduces the impact of the covariance terms between explanatory variables (Lzenman, p 139). Intuition for the previous sentence can be found by considering the variance-covariance matrix for OLS parameters $\left(\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)$. A related discussion can be found in (Hoerl and Kennard, 1970).

Example 1.4.1 *For $p = 2$ we have*

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{n} X_{i1}^2 + \lambda & \sum_{i=1}^{n} X_{i1}X_{i2} \\ \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i2}^2 + \lambda \end{bmatrix}$$

Thus,

$$\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1} = \frac{1}{(a+\lambda)(b+\lambda) - c^2} \begin{bmatrix} b+\lambda & -c \\ -c & a+\lambda \end{bmatrix}$$

where $a = \sum_{i=1}^{n} X_{i1}^2$, $b = \sum_{i=1}^{n} X_{i2}^2$, $c = \sum_{i=1}^{n} X_{i1}X_{i2}$. *Thus, ridge regression reduces the diagonal terms by an order of* λ *and the off-diagonal terms by an order of* λ^2, *provided* λ *is large relative to* a, b, c.

1.5 Alternative Regression Estimators

In this section we review several of the alternative to the least squares and ridge estimators. However this is by no means a complete list. The methods examined here have been grouped into two sections. The first section examines ridge-type approaches to constrained estimation described in the statistical literature. The second section covers model averaging approaches to constrained estimation described in econometrics literature. It is worth noting that although the two approaches evolved out of different branches of literature, they both function by reducing the weight of covariance terms (off-diagonal entries) in the variance-covariance matrix $\mathbf{X}^\top\mathbf{X}$ of the explanatory variables.

1.5.1 Ridge Type Constraints

This section reviews methods for constraining an estimation, which we will refer to as ridge type constraints. Although the methods are conceptually quite different from a ridge, they are computationally very similar.

Lasso

A lasso regression is very similar to a ridge regression. The difference is that the constraint is applied to the sum of absolute parameter estimates rather than the sum of their squares. However, this minor difference has significant repercussions in terms of the resulting estimates. Due to the absolute penalty, the lasso has a stronger tendency to push coefficients to zero and is less sensitive to the standardization of explanatory

variables. The close relation between the ridge and the lasso has allowed the trade-offs between the two methods to be very well documented (Hastie, Tibshirani, Friedman, 2001; Lzenman, 2008). These investigations have led to estimators that combine the ridge and lasso into what is often referred to as an elastic net (Kim and Swanson, 2010; Hastie, Tibshirani and Friedman, 2001). As with the ridge regression a lasso can be expressed as a Bayesian estimation (Hastie, Tibshirani and Friedman, p 72). The lasso estimator is defined as

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{1.5.1}$$

LARS

The Least Angle Regression algorithm is an automatic method for applying estimation constraints and variable selection to a regression problem (Efron, Hastie, Johnstone, Tibshirani , 2004). A small modification makes the algorithm a computationally efficient method for producing the path of lasso solutions. The expression LARS refers to the class of LAR, lasso and closely related algorithms (Lzenman, 2008). LARS is another example of the family of ridge type estimation constraints. These algorithms are either equivalent or slight variations of the lasso, and by extension, they are closely related to the other ridge and Bayesian regressions. Several sources offer comparative discussions for various subsets of these methods (Lzenman, Chapter 5; Hastie, Tibshirani and Friedman, Chapter 3). Although LARS produces similar results to the previously discussed methods, it is conceptually very different. The LARS algorithm initializes all coefficients at zero, then adjusts the coefficient of the series most closely correlated with the residual until there is a second variable equally correlated with the residual. At this point LARS adjusts both coefficients such that the model is moving equiangular between the two variables, until a third variable becomes as correlated with the residuals as the first two. The algorithm progresses in this fashion adding additional variables until it reaches a stopping condition (Efron,

Hastie, Johnstone, Tibshirani , 2004).

1.5.2 Model Averaging Constraints

This section reviews several model averaging approaches to constraining a regression. Model averaging is usually considered in the context of combining the benefit of several pre-existing models. Although this was the original motivation for model averaging, these techniques have also been found to be very useful at reducing estimation variance thereby adjusting the bias-variance trade-off. However, it is worth noting that with the exception of trivial cases all model averaging estimators are inconsistent. Advocates of model averaging offer a number of examples where this variance reduction successfully improves performance (Stock and Watson, 2004).

A model averaging estimator will estimate several auxiliary models in isolation from each other then weight the results to generate a single estimate. This reduces the estimation variance in two ways. Individually estimating auxiliary models is equivalent to forcing some entries in the predictor covariance matrix to zero. Secondly, the values of model weights are generally less than one which reduces variance further. This is shown in the example below where several auxiliary OLS models are being combined into a model averaging estimator with weights $\mathbf{q}$. We then review some of the more commonly used schemes to generate the weights and the motivations behind them.

Assume that we have the following models:

$$Y_i = X_{i1}\beta_1 + \ldots + X_{ip}\beta_p + \varepsilon_i \ , \qquad i = 1, \ldots, n \qquad \text{(Full model)},$$

$$Y_i = X_{i1}\beta_1 + \ldots + X_{ih}\beta_h + \delta_i \ , \qquad i = 1, \ldots, n \qquad \text{(Auxiliary model A)},$$

$$Y_i = X_{ih+1}\beta_{h+1} + \ldots + X_{ip}\beta_p + \delta_i' \ , \qquad i = 1, \ldots, n \qquad \text{(Auxiliary model B)},$$

where ε_i, δ_i, δ_i', $i = 1, \ldots, n$, are mutually independent sequences of i.i.d. random variables with mean zero and a finite variance. Set $\sigma_1^2 = var(\delta_i)$, $\sigma_2^2 = var(\delta_i')$. We consider a weighted average of two auxiliary models:

$$Y_i = q_A(X_{i1}\beta_1 + \ldots + X_{ih}\beta_h) + q_B(X_{ih+1}\beta_{h+1} + \ldots + X_{ip}\beta_p) + u_i \ , \qquad i = 1, \ldots, n,$$

where u_i, $i = 1, \ldots,$ is a sequence of independent random variables with mean zero and a finite variance, and q_A, q_B are the weights of auxiliary models respectively.

In matrix form

$$\hat{\beta}_{\text{Ave}} = \mathbf{q}\left(\left((\mathbf{X}^\top\mathbf{X})_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Y}\right),$$

where

$$(\mathbf{X}^\top\mathbf{X})_{\text{blockzero}} = \begin{bmatrix} \mathbf{X}_{1,h}^\top\mathbf{X}_{1,h} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{h+1,p}^\top\mathbf{X}_{h+1,p} \end{bmatrix}$$

with

$$\mathbf{X}_{1,h} = \begin{bmatrix} X_{11} & \ldots & X_{1h} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{nh} \end{bmatrix}, \qquad \mathbf{X}_{h+1,p} = \begin{bmatrix} X_{1h+1} & \ldots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{nh+1} & \ldots & X_{np} \end{bmatrix}$$

$$\mathbf{q} = \begin{bmatrix} q_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & q_p \end{bmatrix},$$

$$q_1, \ldots, q_h = q_A,$$

$$q_{h+1}, \ldots, q_p = q_B$$

and $\mathbf{0}$ being a zero matrix of the appropriate dimensions.

In the case of deterministic predictors, the bias of a model averaging estimator can be computed as follows:

$$\begin{aligned} E\left(\hat{\beta}_{AVE}-\beta\right) &= E\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{Y}\right)-\beta\right), \\ &= \mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{E}\left(\mathbf{Y}\right)\right)-\beta, \\ &= \mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{X}\beta\right)-\beta, \\ &= \left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{X}\right)-\mathbf{I}\right)\beta. \end{aligned}$$

From this equation one can see that model averaging estimators are not consistent except for trivial cases of perfectly correlated predictors or equal numbers of predictors and auxiliary models.

In the case of deterministic predictors the mean squared error of a model averaging estimator can be computed as follows:

$$\begin{aligned} E\left(\hat{\beta}_{AVE}-\beta\right)^{2} &= E\left(\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{Y}\right)-\beta\right)\times\right. \\ &\quad \left.\times\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{Y}\right)-\beta\right)^{\top}\right), \\ &= E\left(\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\left(\mathbf{X}\beta+\varepsilon\right)\right)-\beta\right)\times\right. \\ &\quad \left.\times\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\left(\mathbf{X}\beta+\varepsilon\right)\right)-\beta\right)^{\top}\right), \\ &= \left(\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{X}\right)-\mathbf{I}\right)\beta\right)\times \\ &\quad \times\left(\left(\mathbf{q}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{X}\right)-\mathbf{I}\right)\beta\right)^{\top} \\ &\quad +\mathbf{q}\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\left(\left(\left(\mathbf{X}^{\top}\mathbf{X}\right)_{\text{blockzero}}\right)^{-1}\right)^{\top}\mathbf{q}^{\top}\sigma^{2}. \end{aligned}$$

However, we are not aware of an analogous formula for the case of stochastic predictors.

Simple Average

One of the most common approaches to model averaging is combining models through the use of a simple average. A number of more complicated model combination methods have been proposed. However, combining models through a simple average is often very effective (Timmerman, 2005; Stock and Watson, 2004). This is especially true if there is a limited amount of data or model performance is fairly similar.

Inverse Mean Squared Error Weights

The effectiveness of combining models with a simple average has led to several weighted average variations. These weighting schemes attempt to place greater weight on the more effective models and will often try to account for correlations between models. One of the most common weighted average approaches is to weight models proportional to the inverse of their estimated mean squared errors. It is fairly simple to show that if each of the models produced mean zero independent errors then the weighted average of models that produced the smallest error would be one where the weights were inversely proportional to the true MSE's. Indeed, if δ and δ' are uncorrelated random variables such that $var(\delta) = \sigma_A^2$ and $var(\delta') = \sigma_B^2$, then

$$var(q\delta + (1-q)\delta') = q^2 var(\delta) + (1-q)^2 var(\delta')$$

which is minimized by the value of q

$$q = \frac{\frac{1}{\sigma_A^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}}.$$

Thus, the weighting scheme involves parameters q_A, $q_B = 1 - q_A$ given as

$$q_A = \frac{\frac{1}{\sigma_A^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}},$$

$$q_B = \frac{\frac{1}{\sigma_B^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}}.$$

Of course, in practice σ_A^2 and σ_B^2 have to be estimated, but this can be done using sample variance of residuals in each auxiliary model.

This can be generalized for higher dimensional cases. The approach has been mentioned in a number of papers (Bates and Granger, 1969; Timmerman, 2005; Stock and Watson, 2004).

Inverse Covariance Matrix of Errors Weighting

The inverse MSE weighting scheme can easily be generalized in order to account for the covariance of auxiliary model errors δ_i, δ_i'. This is accomplished by weighting a set of models by the inverted variance-covariance matrix of observed residuals. The approach uses the same set of assumptions as the inverse MSE weights, with the exception that the assumed error distribution has been relaxed to allow for jointly distributed errors. One of the more well known applications of this weighting scheme is the Markowitz portfolio in finance (Markowitz, 1952). The approach is also frequently discussed in a forecast combination context (Bunn, 1989). Mathematically speaking, the weights are taken as

$$\mathbf{Q} = \frac{\left(\mathbf{\Delta}^\top \mathbf{\Delta}\right)^{-1} c}{c^\top \left(\mathbf{\Delta}^\top \mathbf{\Delta}\right)^{-1} c},$$

where

$$\mathbf{c} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\mathbf{\Delta} = \begin{bmatrix} \hat{\delta}_1 & \hat{\delta}_1' \\ \vdots & \\ \hat{\delta}_n & \hat{\delta}_n' \end{bmatrix},$$

$$\mathbf{Q} = \begin{bmatrix} q_A \\ q_B \\ \vdots \end{bmatrix},$$

and $\hat{\delta}_i$, $\hat{\delta}'_i$, $i = 1, \ldots, n$, are the observed residuals in two auxiliary models. The vector $\mathbf{Q}$ can then be used to construct the vector $\mathbf{q}$.

Regression Weights

Another common approach is to treat the fitted values of the dependent variable from each candidate model as an explanatory variable and combine the models through an OLS linear regression. This approach offers the greatest flexibility in terms of weighting models to account for both fit and error covariance. However, as a consequence using OLS weights will not reduce estimation variance as much as the previously mentioned methods. The use of OLS weights has been mentioned in several papers, however it has often under performed some of the less parameterized approaches mentioned above (Timmerman, 2005).

Chapter 2

Properties of the Ridge Estimator

In the previous chapter, we reviewed the ordinary least squares regression and the bias-variance trade-off that accompanies most regression problems. We then presented several alternative regression estimators designed to achieve a superior trade-off under appropriate circumstances. Of these methods, the ridge is one of the oldest and most commonly used. Additionally, several of the other approaches can be shown to be a variation of the ridge regression. Due to the prominence of the ridge regression, we will begin this chapter by examining properties of the estimators. Afterwards we will consider possible modifications of the ridge regression.

2.1 Classical Ridge Estimator

2.1.1 Ridge Estimator under Deterministic Predictors

The properties of the classic ridge estimator using deterministic predictors is well established, see Lawless (1981).

Recall that $\hat{\beta}_{\text{OLS}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$. Define $\mathbf{W} = \mathbf{XQ}$ and $\alpha = \mathbf{Q}^\top \beta$, where $\mathbf{Q}$ is a matrix such that

$$\left(\mathbf{XQ}\right)^\top \left(\mathbf{XQ}\right) = \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$$

and $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_p$. The parameters $\lambda_1, \ldots, \lambda_p$ are eigenvalues of $\mathbf{X}^\top\mathbf{X}$.

In other words, we transform the multiple linear regression model into the one with the diagonal variance-covariance matrix. The transformed model typically called the canonical form can be written as

$$\mathbf{Y} = \mathbf{W}\alpha + \varepsilon,$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)^\top$. The OLS and ridge estimators in the canonical form are as follows:

$$\hat{\alpha}_{\mathrm{OLS}} = \left(\mathbf{W}^\top\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{Y} = \mathbf{\Lambda}^{-1}\mathbf{W}^\top\mathbf{Y} = \mathbf{Q}^\top\hat{\beta}_{\mathrm{OLS}}$$

and

$$\hat{\alpha}_{\mathrm{ridge}} = \left(\mathbf{W}^\top\mathbf{W} + \lambda\mathbf{I}\right)^{-1}\mathbf{W}^\top\mathbf{Y} = \left(\mathbf{\Lambda} + \lambda\mathbf{I}\right)^{-1}\mathbf{W}^\top\mathbf{Y}.$$

Write the vector $\hat{\alpha}_{\mathrm{OLS}}$ as $\hat{\alpha}_{\mathrm{OLS}} = (\hat{\alpha}_{\mathrm{OLS,1}}, \ldots, \hat{\alpha}_{\mathrm{OLS,p}})$. If the predictors are deterministic, then for each $j = 1, \ldots, p$,

$$\mathrm{E}\left(\hat{\alpha}_{\mathrm{OLS,j}}\right)^2 = \frac{\sigma^2}{\lambda_j}. \tag{2.1.1}$$

On the other hand

$$\hat{\alpha}_{\mathrm{ridge,j}} = \frac{\lambda_j}{\lambda_j + \lambda}\hat{\alpha}_{\mathrm{OLS,j}} \tag{2.1.2}$$

so that

$$\mathrm{E}\left[\hat{\alpha}_{\mathrm{ridge,j}}\right] = \frac{\lambda_j}{\lambda_j + \lambda}\alpha_j$$

and

$$\mathrm{Var}\left[\hat{\alpha}_{\mathrm{ridge,j}}\right] = \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2\frac{\sigma^2}{\alpha_j}.$$

Consequently,

$$\mathrm{E}\left(\hat{\alpha}_{\mathrm{ridge,j}} - \alpha_j\right)^2 = \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2\frac{\sigma^2}{\lambda_j} + \left(\frac{\lambda}{\lambda_j + \lambda}\right)^2\alpha_{\mathrm{j}}^2 \tag{2.1.3}$$

Formula (2.1.3) indicates that for some values of the parameter λ we may reduce mean squared error below the one for OLS estimator. The analytical formula for

the optimal λ can be written explicitly by minimizing the above expression w.r.t. λ. However, it is not of practical use, since it contains unobservable quantities.

Formula (2.1.2) indicates that the ridge estimator $\hat{\alpha}_{\text{ridge,j}}$ is only a consistent estimator of $\hat{\alpha}_j$ if the smallest eigenvalue converges to infinity (as $n \to \infty$).

2.1.2 Ridge Estimator Under Stochastic Predictors

Although the formula (2.1.2) is still valid in the random design case, we cannot confirm (2.1.3). There are two reasons for this. First, there are no easy expressions for the variance and mean of the OLS estimator in the random design case. Second, the eigenvalues λ_j are random. Furthermore, the random variables $\lambda_j/(\lambda_j + \lambda)$ and $\hat{\alpha}_{\text{OLS,j}}$ are dependent.

We are not aware of any research on ridge regression in the random design case. Formulas for mean square error seem to be hard to obtain.

2.1.3 Selection of the Tuning Parameter

Unlike OLS, the ridge estimator relies on a tuning parameter λ. The choice of this tuning parameter can pose a dilemma. Like any model parameter, the choice is an attempt to minimize a loss function, such as mean squared error which is unobservable. As a consequence there are several commonly used criteria that people use as an approximation. Examples include: minimize forecast error and various cross validation approaches; with Generalized Cross Validation being a common choice. However, beyond the standard methods for selecting a regression parameter, there have been a number of proposed approaches that have been specifically tailored to the ridge regression. Some of them are well known and can be found in papers like Hoerl and Kennard (1970), Lawless and Wang (1976), Khalaf and Shukur (2005), El-Salam (2011). These respective estimators are shown below. However, the selection of the

ridge parameter is still being investigated.

Recall the canonical form of the ridge estimator: $\mathbf{W} = \mathbf{XQ}$ and $\alpha = \mathbf{Q}^\top \beta$, where $\mathbf{Q}$ is a matrix such that

$$(\mathbf{XQ})^\top (\mathbf{XQ}) = \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$$

and $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_p$. The parameters $\lambda_1, \ldots, \lambda_p$ are eigenvalues of $\mathbf{X}^\top \mathbf{X}$. The ridge estimates are given by

$$\hat{\alpha}_{\text{ridge}} = \left(\mathbf{W}^\top \mathbf{W} + \lambda \mathbf{I}\right)^{-1} \mathbf{W}^\top \mathbf{Y} = \left(\mathbf{\Lambda} + \lambda \mathbf{I}\right)^{-1} \mathbf{W}^\top \mathbf{Y}.$$

Let $\hat{\sigma}^2$ be a consistent estimator of the variance of the noise sequence ϵ_i. Some of the commonly proposed choices of the ridge estimator are as follows:

- Hoerl and Kennard (1970):
$$\lambda = \frac{p\hat{\sigma}^2}{\hat{\alpha}_{\text{OLS}}^\top \hat{\alpha}_{\text{OLS}}},$$
- Lawless and Wang (1976):
$$\lambda = \frac{p\hat{\sigma}^2}{\sum_{i-1}^{P} \lambda_i \hat{\alpha}_{\text{OLS},i}^2},$$
- Khalaf and Shukur (2005):
$$\lambda = \frac{\lambda_{\max} \hat{\sigma}^2}{(n-p-1)\hat{\sigma}^2 + \lambda_{\max} \hat{\alpha}_{\max}^2}.$$
where $\lambda_{\max}$ is the maximal eigenvalue, and $\hat{\alpha}_{\max}$ is the largest among OLS estimates $\hat{\alpha}_{\text{OLS},i}$, $i = 1, \ldots, p$.
- El-Salam (2011):
$$\lambda = \max\left(0, \frac{p\hat{\sigma}^2}{\hat{\alpha}_{\text{OLS}}^\top \hat{\alpha}_{\text{OLS}}} - \frac{1}{n \max_{j=1,\ldots,p}(\text{VIF}_j)}\right)$$
where VIF_j is the variance inflation factor of the j regressor.

However, these methods are only applicable to the classic ridge regression and not related methods. A more general approach to selecting tuning parameters is by minimizing forecast error, which is effective under a wide variety of specifications. To be more specific, one can estimate a model on observations 1 to $n-1$, then use these parameter estimates and the predictor variables at observation n to forecast the dependent variable at observation n. After this process is repeated across several data points and several tuning parameter values, the minimum of the average squared difference between the predicted dependent variable and the actual dependent variable can be used to select an appropriate value for the tuning parameter.

2.1.4 Modified Ridge Estimator

The previous chapter discussed the ridge regression and several other alternatives to the OLS that attempt to constrain a regression in order to adjust the bias-variance trade-off and improve a regressions performance. Despite the fact that these techniques have evolved out of different veins of literature they operate in the same fashion. All of the methods discussed reduce the variance of parameter estimates by forcing the covariance estimates of explanatory variables towards zero. The ridge estimator and the other various ridge-type regressions scale down the covariance estimates in a relatively even fashion, and the model averaging regressions set blocks of covariance estimates to zero. The literature we cited offers countless examples of these techniques being more effective in some situations than ordinary least squares. Although, all of these estimators operate by reducing the off-diagonal entries of the $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ matrix, a troubling feature of these approaches is that they are very arbitrary. They all shift large numbers of covariance estimates towards zero, without much in the way of an hypothesis justifying these actions.

This is the motivation for our estimator. We wish to improve upon the effectiveness of the previous methods by modifying the ridge estimator, while maintaining the

same general approach. Instead of increasing the diagonal entries in $\left(\mathbf{X}^\top\mathbf{X}\right)$ which disproportionately decreases the off diagonal entries when the matrix is inverted, we define the modified ridge as decreasing the off diagonal entries of $\left(\mathbf{X}^\top\mathbf{X}\right)$. The idea being that the estimator will be similar but the implications of predictor correlations will be somewhat different. As with other related methods we suspect that the estimators performance will be sensitive to any stochastic properties of the predictors as this will incorporate sampling error into the variance and covariance estimates of the predictors contained in $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

We consider the model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

and we estimate the parameters β as

$$\hat{\beta}_{\text{NEW}} = \left((\mathbf{X}^\top\mathbf{X})_{\text{NEW}}\right)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

A modified variance-covariance matrix $(\mathbf{X}^\top\mathbf{X})_{\text{NEW}}$ is built as follows. We take the off-diagonal entries of the variance-covariance matrix and scale them by s, keeping the diagonal entries the same. Mathematically,

$$\left(\mathbf{X}^\top\mathbf{X}_{\text{NEW}}\right)_{ij} = s\left(\mathbf{X}^\top\mathbf{X}\right)_{ij} + (1-s)\mathbf{D}_{ij}, \qquad i, j = 1, \ldots, p,$$

where $s \in [0, 1)$ and $\mathbf{D} = (\mathbf{D}_{ij})_{i,j=1,\ldots,p}$ is the diagonal of $\mathbf{X}^\top\mathbf{X}$.

For the sake of algebraic simplicity we followed the common practice of subtracting the mean from each of the series we use. $\hat{\beta}_{\text{NEW}}$ is our proposed modification on the ridge estimator. In the following sections we will examine its consistency and mean squared error.

Consistency of the Estimator

In this section we assume that predictors are stochastic. For the purpose of clarity we present a result and establish a lack of consistency for the two dimensional case, however, this can readily be generalized to higher dimensions. We have the following lemma.

Lemma 2.1.1 *Consider the regression model (1.1.1) with $p = 2$. Then, in probability*

$$\lim_{n\to\infty} \hat{\beta}_{1,\mathrm{NEW}} = \frac{\sigma_2^2(\beta_1\sigma_1^2 + \beta_2\varrho) - s\varrho(\beta_1\varrho + \beta_2\sigma_2^2)}{\sigma_1^2\sigma_2^2 - s^2\varrho^2} = \beta_1 + \frac{(1-s)\varrho(\beta_2\sigma_2^2 - s\varrho\beta_1)}{\sigma_1^2\sigma_2^2 - s^2\varrho^2}. \tag{2.1.4}$$

Proof: The proof is done with respect of $\hat{\beta}_1$ however, $\hat{\beta}_2$ follows from symmetry. We have

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^n X_{i2}^2 \sum_{i=1}^n X_{i1}Y_i - \sum_{i=1}^n X_{i1}X_{i2} \sum_{i=1}^n X_{i1}Y_i}{\sum_{i=1}^n X_{i1}^2 \sum_{i=1}^n X_{i2}^2 - (\sum_{i=1}^n X_{i1}X_{i2})^2}$$

and

$$\hat{\beta}_{1,NEW} = \frac{\sum_{i=1}^n X_{i2}^2 \sum_{i=1}^n X_{i1}Y_i - s\sum_{i=1}^n X_{i1}X_{i2} \sum_{i=1}^n X_{i1}Y_i}{\sum_{i=1}^n X_{i1}^2 \sum_{i=1}^n X_{i2}^2 - s^2(\sum_{i=1}^n X_{i1}X_{i2})^2}.$$

Note that

$$\frac{1}{n}\sum_{i=1}^n X_{i1}Y_i = (1/n)\sum_{i=1}^n X_{i1}(\beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i),$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_{i1}^2 = \sigma_1^2 = \mathrm{Var}(X_{11}) \qquad \text{in probability,}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_{i2}^2 = \sigma_2^2 = \mathrm{Var}(X_{12}) \qquad \text{in probability,}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_{i1}X_{i2} = \varrho \qquad \text{in probability,}$$

where ϱ is the covariance between the first and the second predictor. Therefore, as $n \to \infty$, in probability,

$$\lim_{n\to\infty} \hat{\beta}_{1,\mathrm{OLS}} = \frac{\sigma_2^2(\beta_1\sigma_1^2 + \beta_2\varrho) - \varrho(\beta_1\varrho + \beta_2\sigma_2^2)}{\sigma_1^2\sigma_2^2 - \varrho^2} = \frac{\beta_1(\sigma_2^2\sigma_1^2 - \varrho^2)}{\sigma_1^2\sigma_2^2 - \varrho^2} = \beta_1$$

and

$$\lim_{n\to\infty}\hat{\beta}_{1,\mathrm{NEW}} = \frac{\sigma_2^2(\beta_1\sigma_1^2+\beta_2\varrho)-s\varrho(\beta_1\varrho+\beta_2\sigma_2^2)}{\sigma_1^2\sigma_2^2-s^2\varrho^2} = \beta_1+\frac{(1-s)\varrho(\beta_2\sigma_2^2-s\varrho\beta_1)}{\sigma_1^2\sigma_2^2-s^2\varrho^2}. \tag{2.1.5}$$

□

Note that the proposed estimator is not consistent as it dose not converge in probability to the true value. It shares this property with the classical ridge estimator. However, unlike the ridge estimator, it is consistent in the special case of independent predictors.

Mean Squared Error of Estimator Under Deterministic Predictors

In the case of deterministic predictors, we are able to determine the mean squared error for the proposed modification of the ridge. For simplicity, this is presented below in canonical form.

Let

$$(\mathbf{XQ})^\top(\mathbf{XQ}) = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$$

and $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_p$. The parameters $\lambda_1,\ldots,\lambda_p$ are eigenvalues of $\mathbf{X}^\top\mathbf{X}$. and $\mathbf{D}$ is the diagonal of $\mathbf{X}^\top\mathbf{X}$, with $\mathbf{Q}^\top\mathbf{DQ} = \mathbf{D} = \mathrm{diag}(d_1,\ldots,d_p)$. Then

$$\hat{\alpha}_{\mathrm{NEW}} = \left(s\mathbf{W}^\top\mathbf{W}+(1-s)\mathbf{D}\right)^{-1}\mathbf{W}^\top\mathbf{Y},$$

$$\hat{\alpha}_{\mathrm{NEW}} = \left(s\mathbf{W}^\top\mathbf{W}+(1-s)\mathbf{D}\right)^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{W}^\top\mathbf{Y},$$

$$\hat{\alpha}_{\mathrm{NEW}} = \left(s\mathbf{W}^\top\mathbf{W}+(1-s)\mathbf{D}\right)^{-1}\mathbf{\Lambda}\hat{\alpha}_{\mathrm{OLS}}.$$

Thus,

$$\hat{\alpha}_{\mathrm{NEW,j}} = \frac{\lambda_j\hat{\alpha}_{OLSj}}{s\lambda_j+(1-s)d_j}.$$

Hence,

$$\mathrm{E}\left(\hat{\alpha}_{\mathrm{NEW,j}}\right) = \frac{\lambda_j\mathrm{E}\left(\hat{\alpha}_{OLSj}\right)}{s\lambda_j+(1-s)d_j} = \frac{\lambda_j\alpha_j}{s\lambda_j+(1-s)d_j},$$

$$\mathrm{Var}\left(\hat{\alpha}_{\mathrm{NEW,j}}\right) = \left(\frac{\lambda_j}{s\lambda_j + (1-s)d_j}\right)^2 \frac{\sigma^2}{\lambda_j},$$

$$\mathrm{MSE}\left(\hat{\alpha}_{\mathrm{NEW,j}}\right) = \left(\frac{\lambda_j}{s\lambda_j + (1-s)d_j}\right)^2 \frac{\sigma^2}{\lambda_j} + \left(\frac{\lambda_j \alpha_j}{s\lambda_j + (1-s)d_j} - \alpha_j\right)^2.$$

This is a similar formula to the one for the classical ridge estimator; see (2.1.3). Thus, we can choose the tuning parameter s in a similar manner as described in section 2.1.3.

Chapter 3

Simulations

In order to better identify the properties of the ridge estimator and some of its alternatives, we have computed four sets of simulations. The exact procedures are described in the sections that follow. The simulations include several ridge and model averaging methods. We did not include the related lasso and LARS approaches, as these methods are primarily used for model selection not estimation.

In the first simulation, the ridge estimator is compared to OLS using various choices of σ^2 and correlation between predictors. This was done in order to better quantify some of the commonly cited advantages of the ridge estimator, such as it performing best when the predictors are strongly correlated.

The second simulation compares the ridge and OLS when a Monte Carlo procedure is used to select the tuning parameter. This methodology is often not feasible in applications. However, the simulation demonstrates the potential performance improvement that can be achieved with a ridge regression in various situations.

The third simulation builds upon the second by comparing the performance of the ridge and OLS using several feasible choices for the tuning parameter. We find that there are several ridge estimators that can outperform an OLS when the sample is small relative to the number of predictors. Additionally, we find that the modified

ridge estimator has some unusual characteristics relating to errors-in-variables and predictors with a low degree of correlation.

In the fourth simulation several model averaging estimators are compared to the OLS and ridge estimators. In some small sample cases, they can noticeably outperform the ridge estimators. However, they tend to under perform as the sample size increases.

3.1 Simulation 1: Preliminary Comparison of OLS and Ridge Regression

In this experiment, we compare OLS and ridge estimates for deterministic and random predictors.

First, we compare OLS and ridge for deterministic predictors. We consider the model

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where X_{i1}, X_{i2} are deterministic predictors, ε_i are independent $N(0, \sigma^2)$.

The ridge parameter λ is chosen from the grid $[0.1, 0.2, \ldots, 1.7]$. Then the best ridge estimator is chosen according to (unobservable in practice) mean square error and compared to (unobservable) mean square error for OLS estimator. We also analyze sensitivity with respect to σ^2, the variance of the noise. For simplicity we use equal betas for all of the simulations in this chapter, however, we recognize that this dose create a limitation.

Specifically, the simulation procedure runs as follows:

1. The parameters: $\beta_1 = \beta_2 = 1$; sample size $n = 100$, several choices of σ.

2. We generate X_{i1}, X_{i2}, $i = 1, \ldots, n$, from $N(0, 1)$ such that the two series have correlation 0.8.

3. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.

4. We estimate β_1 and β_2 using OLS estimator and ridge estimator with different choices of λ from a grid $[0.1, 0.2, \ldots, 1.7]$.

5. We compute residuals $Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}$ for different ridge estimates.

6. We repeat steps 3-5 $m = 1000$ times (so that the same predictors are used in each simulation, only the errors ε_i change).

7. We compute MSE(OLS) - mean squared error for OLS estimator and MSE(λ) - mean squared error for different ridge estimates.

The "true" mean squared error (based on $(\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta)^2$). The minimum is obtained and the ratio $\text{MSE}(\lambda_{\text{opt}})/\text{MSE}(\text{OLS}) = 0.8490743$ indicates that ridge estimator performs better than the OLS under these conditions. Next, we repeat the simulation procedure in stochastic predictors case. In order to do this, we repeat steps 2-5 $m = 1000$ times, so that predictors are generated at each repetition. The results for both fixed and random design are summarized in the table below.

There are several conclusions from the simulations:

- There is some improvement if σ is large.
- Improvement over OLS for "large" σ which is intuitive from the formula (2.1.3).
- Ridge estimator works similarly for deterministic and stochastic predictors.

Variance	MSE(λ_{opt})/MSE(OLS) - deterministic	MSE(λ_{opt})/MSE(OLS) - stochastic
$\sigma = 0.1$	0.9904335	0.9839907
$\sigma = 0.2$	0.9241228	0.9385038
$\sigma = 0.3$	0.840041	0.8926964
$\sigma = 0.8$	0.8552906	0.8694426
$\sigma = 0.9$	0.8721405	0.8689521
$\sigma = 1$	0.8490743	0.8675821

Table 3.1: Results for Simulation 1: ridge vs. OLS

Correlation	MSE(λ_{opt})/MSE(OLS) - dependent
$\rho = 0.5$	0.9758
$\rho = 0.8$	0.9112487
$\rho = 0.9$	0.8254

Table 3.2: Results for simulation 1: dependent predictors

Finally, we consider random design case, but predictors are correlated. We fix $\sigma = 1$ and change correlation as shown in Table 3.2.

We conclude that the improvement offered by a ridge regression relative to an OLS increases with predictor correlation.

3.2 Simulation 2: Comparison of OLS and Ridge Regression with Monte Carlo Selection of Tuning Parameter

In this simulation, we compare the mean squared forecast error (MSFE) between the OLS estimator and the ridge estimator using a Monte Carlo simulation to optimally select λ. This approach is not feasible for most applications. However, this simulation will identify how large a performance improvement can potentially be achieved by using a ridge under various situations. The mean squared forecast error is determined by estimating a model on observations 1 to $n-1$, then using these parameter estimates and the predictor variables at observation n to forecast the dependent variable at observation n. Repeating this process across several data points and comparing the resulting forecast to actual values allows one to calculate the mean squared forecast error.

Specifically, the simulation procedure runs as follows:

1. The parameters: $\beta_1 = \cdots = \beta_p = 1$; $\sigma = 1$, and several choices of n and p.
2. We generate $X_i^*, i = 1, \ldots, n$, from $N(0, 1)$.
3. We generate $\tilde{X_{i1}}, \ldots, \tilde{X_{ip}}$, $i = 1, \ldots, n$ from $N(0, 1)$.
4. We generate $X_{i1}, \ldots, X_{ip}$, $i = 1, \ldots, n$ as a linear combination of X^* and $\tilde{X}$. Such that $X_{i1} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{i1}}$ and $X_{ip} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{ip}}$. This creates

dependent predictors where the correlation between any two is 0.8.

5. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.

6. We use the sample of $n - 1$ observations and the choice of λ to compute the ridge estimate of $\beta_1, \ldots, \beta_p$.

7. We use the resulting $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and $X_{n1} \ldots X_{np}$ to compute the corresponding predictions $\hat{Y_n}$ of Y_n.

8. We compute the squared forecast errors $(\hat{Y_n} - Y_n)^2$ for the three different ridge estimators.

9. We repeat steps 2-8 $m = 100000$ times (so that new predictors are used in each simulation).

10. We compute the resulting mean squared forecast error for the ridge estimator.

11. We repeat steps 2-10 for each choice of the ridge parameter λ from the grid $[0, 0.05, 0.1, \ldots, 4]$ and present the results of the λ with the lowest mean squared forecast error.

The results of the simulation are reported in Table 3.3

The conclusions of this section are as follows:

- Although a Monte Carlo procedure to select the ridge parameter is often not feasible, this simulation shows the potential improvement that is possible through a ridge approach.

- The situations where the ridge appears to have the largest potential advantage over the OLS is when the sample size is small relative to the number of predictors.

Table 3.3: Comparison of mean squared forecast error (MSFE) for OLS and ridge regression with Monte Carlo selection of λ

MSFE for Ridge with Monte Carlo λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.867	0.729	0.574	NA
$n = 25$	0.977	0.948	0.915	0.122
$n = 50$	0.992	0.983	0.971	0.710
$n = 100$	0.997	0.995	0.991	0.920

3.3 Simulation 3: Comparison of OLS and Ridge Regression with Feasible Selection of Tuning Parameter

The simulation in the previous section established that a sizable forecasting improvement could potentially be achieved over an OLS regression by using a ridge regression. In this simulation, the parameter λ was optimized through a preliminary Monte Carlo simulation. However, in many practical applications a Monte Carlo approach is infeasible or very costly. Consequently, this section will repeat the comparison using parameter estimates that are feasible without a Monte Carlo procedure.

Over the years, a number of methods have been proposed for selecting the ridge parameter. The simulations in this section use the approaches suggested in: Hoerl and Kennard (1970), Lawless and Wang (1976), Khalaf and Shukur (2005). Each of these methods is described in section 2.1.3. The simulations in this section are repeated in section 5.1 of the appendix with an errors-in-variables specification.

The modified ridge estimator is also included in this comparison. The feasible

choice of the tuning parameter in the modified ridge estimator is provided by $S = 1 - \frac{1}{n-p-1}$. This is simply a convenient way of adjusting the tuning parameter to reflect sample size and the number of predictors. We suspect that there are better ways to estimate S.

Specifically, the simulation procedure runs as follows:

1. The parameters: $\beta_1 = \cdots = \beta_p = 1$; $\sigma = 1$, and several choices of n and p.

2. We generate $X_i^*, i = 1, \ldots, n$, from $N(0, 1)$.

3. We generate $\tilde{X_{i1}}, \ldots, \tilde{X_{ip}}$, $i = 1, \ldots, n$ from $N(0, 1)$.

4. We generate $X_{i1}, \ldots, X_{ip}$, $i = 1, \ldots, n$ as a linear combination of X^* and $\tilde{X}$. Such that $X_{i1} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{i1}}$ and $X_{ip} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{ip}}$. This creates dependent predictors where the correlation between any two is 0.8.

5. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.

6. We use the sample of $n-1$ observations to compute λ by the approaches of Hoerl and Kennard (1970), Lawless and Wang (1976), Khalaf and Shukur (2005). These equations are in section 2.1.3.

7. We use the sample of $n - 1$ observations and the three choices of λ to compute the ridge estimates of $\beta_1, \ldots, \beta_p$.

8. We use the resulting $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and $X_{n1} \ldots X_{np}$ to compute the corresponding predictions $\hat{Y_n}$ of Y_n.

9. We compute the squared forecast errors $(\hat{Y_n} - Y_n)^2$ for the different ridge estimators.

10. We repeat steps 2-9 $m = 100000$ times (so that new predictors are used in each simulation).

11. We compute the resulting mean squared forecast error for each of the ridge estimators.

The results of the simulation are reported in Tables 3.4 to 3.7.

Table 3.4: Comparison of MSFE for OLS and ridge regression with Hoerl and Kennard (1970) selection of λ

MSFE for Ridge with Hoerl and Kennard λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.946	0.865	0.786	NA
$n = 25$	0.990	0.975	0.960	0.603
$n = 50$	0.997	0.994	0.989	0.905
$n = 100$	0.999	0.998	0.997	0.979

Table 3.5: Comparison of MSFE for OLS and ridge regression with Lawless and Wang (1976) selection of λ

MSFE for Ridge with Lawless and Wang λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.977	0.950	0.917	NA
$n = 25$	0.999	0.999	0.999	0.968
$n = 50$	1.000	1.000	1.000	1.000
$n = 100$	1.000	1.000	1.000	1.000

The conclusions of this section are as follows:

- As one would expect there is a noticeable decrease in forecasting performance as Monte Carlo selection is replaced with the more feasible alternatives.

Table 3.6: Comparison of MSFE for OLS and ridge regression with Khalaf and Shukur (2005) selection of λ

MSFE for Ridge with Khalaf and Shukur λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.955	0.993	0.899	NA
$n = 25$	0.994	0.997	0.995	0.913
$n = 50$	0.998	1.00	1.000	0.996
$n = 100$	1.000	1.000	1.000	1.000

Table 3.7: Comparison of MSFE for OLS and modified ridge regression with $S = 1 - \frac{1}{n-p-1}$

MSFE for Modified Ridge with $S = 1 - \frac{1}{n-p-1}$/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.921	0.836	0.778	NA
$n = 25$	0.988	0.975	0.961	3.095
$n = 50$	0.998	0.995	0.992	1.025
$n = 100$	0.999	0.999	0.998	1.007

- Although the Hoerl and Kennard approach is the oldest of the methods to select a ridge parameter, it appears superior to the two subsequent approaches. This is possibly because we are comparing mean squared forecast errors, instead of the more common comparison of the mean squared error of the beta parameters.

- As in the previous section the classical ridge regression appears to outperform OLS the most when the sample size is small relative to the number of predictors.

- When the simulations in this section are repeated in section 5.1 of the appendix using an errors-in-variables specification, the relative performance of the classical ridge appears to slightly deteriorate, unless the number of predictors is large in which case the performance noticeably improves.

- The modified ridge performs very poorly in the case of 20 predictors, but remains competitive in many of the other cases.

- The modified ridge appears to have some unusual characteristics. In some errors-in-variables situations it can outperform the classical ridge see section5.1 of the appendix. Additionally, a simulation in section 5.3 of the appendix appears to indicate that in some respects the relative performance of the modified ridge improves as predictor correlation decreases. This is interesting as the reverse is true for the ridge.

3.4 Simulation 4: Comparison of OLS and Model Averaging Regressions

Apart from ridge type regressions, model averaging approaches also attempt to improve upon an OLS by adjusting the bias-variance trade-off. In section 1.5.2 we reviewed a number of common model averaging approaches. In this section, we will

conduct a similar simulation exercise as done above to evaluate how the model averaging techniques perform relative to an OLS. Comparing these results to those in the previous section, will show the relative forecasting performance of the model averaging techniques and ridge regression approaches.

Unlike some of the ridge regression techniques that are based on a single tuning parameter, several elements are required to define a model average. A determination must be made as to how many auxiliary models to use, how the predictors are grouped into auxiliary models and how the resulting models will be averaged together. Therefore we need to examine some of these properties before making a comparison with the ridge methods.

The number of auxiliary models is bound between 2 and p, where p is the number of predictors. Therefore, in the cases of 2 or 3 predictors, we are restricted to using exactly 2 or p auxiliary models. We have found that 2 auxiliary models tend to work noticeably better than p auxiliary models in general. Consequently, all of the results in this section are based on model averages using 2 auxiliary models. In practice, it is very important as to how predictors are grouped into auxiliary models. However, this is not particularly relevant in our comparison with the ridge approaches, as in these simulations all of the predictors have equal predictability and equal covariance. Finally, to ensure the robustness of the results, we examined three standard model averaging techniques from the least parameterized simple average to inverse mean squared error weights to the most parameterized OLS weights. All of the simulations in this section are repeated in section 5.2 of the appendix with an errors-in-variables specification.

Specifically, the simulation procedure runs as follows:

1. The parameters: $\beta_1 = \cdots = \beta_p = 1$; $\sigma = 1$, and several choices of n and p.

2. We generate $X_i^*, i = 1, \ldots, n$, from $N(0, 1)$.

3. We generate $\tilde{X_{i1}}, \ldots, \tilde{X_{ip}}$, $i = 1, \ldots, n$ from $N(0, 1)$.

4. We generate $X_{i1}, \ldots, X_{ip}$, $i = 1, \ldots, n$ as a linear combination of X^* and $\tilde{X}$. Such that $X_{i1} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{i1}}$ and $X_{ip} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{ip}}$. This creates dependent predictors where the correlation between any two is 0.8.

5. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.

6. We then use predictors $1, \ldots, [\frac{p}{2}]$ to form the first auxiliary model, and predictors $[\frac{p}{2}] + 1, \ldots, p$ to form the second auxiliary model.

7. We use the sample $n-1$ observations to compute OLS estimates of $\beta_1, \ldots, \beta_{[\frac{p}{2}]}$ in the first auxiliary model, and $\beta_{[\frac{p}{2}]+1}, \ldots, \beta_p$ in the second auxiliary model.

8. We then compute $\hat{Y}_{1,g}, \ldots, \hat{Y}_{n-1,g}$ for each of the two auxiliary models where $g = 1, \ldots, 2$.

9. The two auxiliary models are assigned weights q_g using their fitted values of $\hat{Y}_{1,g}, \ldots, \hat{Y}_{n-1,g}$, this is done using one of the methods described in section 1.5.2, specifically simple average weights, inverse mean squared error weights or OLS weights.

10. We use the resulting $\hat{\beta}_1, \ldots, \hat{\beta}_{[\frac{p}{2}]}$ and $X_{n1} \ldots X_{n[\frac{p}{2}]}$ to compute the corresponding prediction $\hat{Y}_{n1}$ of Y_n from the first auxiliary model. Then $\hat{\beta}_{[\frac{p}{2}]+1}, \ldots, \hat{\beta}_p$ and $X_{n[\frac{p}{2}]+1} \ldots X_{np}$ is used to compute the prediction $\hat{Y}_{n2}$ of Y_n from the second auxiliary model.

11. The final prediction $\hat{Y}_n$ of Y_n is computed as $\hat{Y}_n = q_1\hat{Y}_{n1} + q_2\hat{Y}_{n2}$

12. We compute the squared forecast error $(\hat{Y}_n - Y_n)^2$ for the estimator.

13. We repeat steps 2-12 $m = 100000$ times (so that new predictors are used in each simulation).

14. We repeat steps 2-13 for each of the three methods of determining model weights, to determine the mean squared forecast error for each of the three estimators.

The results of the simulation are reported in Tables 3.8 to 3.10.

Table 3.8: Comparison of MSFE for OLS and model average with simple average weights

MSFE for model average with simple average weighting/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.854	0.775	0.612	NA
$n = 25$	0.984	1.011	0.930	0.281
$n = 50$	1.015	1.061	0.993	0.970
$n = 100$	1.025	1.078	1.021	1.052

Table 3.9: Comparison of MSFE for OLS and model average with inverse mean squared error weights

MSFE for model average with inverse MSE weights/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.877	0.808	0.666	NA
$n = 25$	0.991	0.988	0.963	0.316
$n = 50$	1.015	1.019	1.009	1.025
$n = 100$	1.027	1.035	1.027	1.072

The conclusions of this section are as follows:

Table 3.10: Comparison of MSFE for OLS and model average with OLS weights

MSFE for model average with OLS weights/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	1.000	0.898	0.728	NA
$n = 25$	1.000	0.980	0.947	0.294
$n = 50$	1.000	0.991	0.978	0.978
$n = 100$	1.000	0.996	0.990	1.024

- When the model averaging estimators are compared to the feasible ridge estimators from the previous section, the model averaging estimators can perform noticeably better in a few cases where the number of observations is very small relative to the number of predictors. However, as the number of observations increases, the ridge estimators tend to outperform the model average estimators.

- In several cases involving larger sample sizes the model averaging techniques perform worse than OLS.

- As more parameters are used in determining model weights, the performance of a model average estimator appears to move towards that of an OLS estimator. In the extreme case, the two become identical as seen in the first column of 3.10. A corollary could be that if model averaging estimators outperform the alternatives in a given situation, greatest performance improvement is likely to be achieved by the simple average. This may explain why the simple average is often favored by practitioners.

- When the simulations in this section are repeated in section 5.2 of the appendix using an errors-in-variables specification, the outcome is similar to the ridge es-

timators. The relative performance of the model averaging estimators appears to slightly deteriorate, unless the number of predictors is large in which case the performance noticeably improves. However, the extent to which the performance improves is greater in the model averaging estimators than the ridge estimators. This would seem to indicate some additional cases where model averaging estimators can be competitive relative to a ridge.

3.5 Conclusion of Simulation Studies

The first simulation compares the ridge estimator and OLS over several choices of σ^2 and correlation between predictors. We find that the ridge regression performs well relative to the OLS when σ^2 is large and the predictors are highly correlated. These results are somewhat intuitive, as the ridge estimator was created to account for highly dependent predictors and from (2.1.3) the bias-variance trade-off appears favorable relative to the OLS when σ^2 is large.

The second simulation examines ridge and OLS using a Monte Carlo procedure to select the ridge parameter. This approach is not feasible in many applications. However, the simulation identifies the performance advantage that can potentially be obtained with a ridge estimator and shows that this is largest when the sample size is small relative to the number of predictors.

The third and fourth simulations examine the performance of the OLS, feasible ridge estimators and model averaging estimators. When the sample size is small relative to the number of predictors, the ridge estimators generally appear to perform well. However, there appear to be some cases where the best performance may be achieved by the model averaging or modified ridge estimators, particularly in situations with very small samples, errors-in-variables, or low correlation between predictors.

An overall conclusion could be that the ridge estimator often outperforms the OLS estimator when the sample size is small relative to the number of predictors.

However, within this set of situations there are a number of special cases where the ridge is outperformed by alternatives, particularly when the situation involves very small samples, errors-in-variables or predictors with low correlation.

Chapter 4

Empirical Applications

In this thesis, we have examined ridge estimators as well as several competing model averaging estimators. The scope of our comparison has included some cases of stochastic predictors such as errors-in-variables situations. In the previous chapter, we conducted a number of simulations to demonstrate the relative forecasting performance of these estimators under various situations. In this chapter, we will offer examples of applications where these various estimators may be useful.

The literature contains numerous examples of situations where the ridge estimator outperforms various alternatives. This is consistent with the strong performance of the ridge in our simulations. However, the previous chapter also suggests that there are some cases where the ridge estimator can be outperformed by related techniques; specifically in situations with errors-in-variables, very small samples or low correlation between predictors. The examples presented in this chapter will focus on some of these unusual cases that are less often examined.

Although, these situations are perhaps less standard, they do exist in several applications. Measurement error is a frequent occurrence in many fields and this always creates an errors-in-variables situation by definition. Another common errors-in-variables example is whenever one predictor is serving as a proxy for an unobserv-

able series; again this type of situation is often encountered in practice. Situations with a low correlation between predictors are perhaps less common, however there are applications in which they occur. A principle component regression where the predictors are projections onto the principle components of a data set always results in predictors with a correlation of zero. In a panel regression the predictors specific to a particular panel series often have a low correlation with the non-specific predictors. For instance, if one conducted a panel regression on house prices by municipality, using municipality specific predictors and national economic predictors, one would expect several of the municipality specific predictors to have very little correlation to the national level predictors. Beyond this one can think of other examples. For instance, actuaries will attempt to predict house insurance claims using predictors such as proximity to water, type of furnace and presence of dead bolts; where it would be difficulty to imagine these predictors as highly correlated. Situations involving extremely small sample sizes can also be found in practice whenever data is expensive or impossible to obtain. For instance, if one conducted a study where the observations represented various recessions or market crashes, the sample size would be very small. Another example of small sample size regressions would be whenever financial market participants analyze the value of a newly alliable time series. In the following sections, we consider applications that have some of these characteristics.

4.1 Gross Domestic Product Forecasting

Forecasting the growth rate of the gross domestic product (GDP) is of interest in the fields of economics and finance. Accurate forecasting models aid monetary and financial policy decisions and they can also be useful to those who speculate in financial markets. In both of these application a major focus in placed on having an accurate forecast of the current quarter and this will be the source of our example.

When forecasting the current quarters GDP growth forecasters will often make

preliminary forecasts of the components of GDP, sometimes referred to as the disaggregate approach to forecasting. The reason for this is that several of the components of GDP are easier to forecast then the total. In many countries advanced data is available for some components of GDP such as imports and exports. Other components represent expenditures that have been planed a few months in advanced and are relatively simple to forecast. Several of the investment components of GDP may fall into this category such housing construction and business investment. Once a preliminary set of forecast are made for these components, these forecasts are used to create the final GDP forecast. This is often done by using the preliminary forecasts as predictors for total GDP. Alternatively, one can make a preliminary forecast for each component and construct the corresponding GDP forecast as a weighted sum or use a combination of the two approaches.

In this example we will attempt to recreate a situation where a forecaster is using advanced data and/or preliminary forecasts of trade and investment components to forecast the current quarterly release of GDP. Out of convenience we are going to make the assumption that the advanced data/preliminary forecasts available to a forecaster can be represented by their historical values. We will apply this exercise to Canada over the entire history of the series from 1961 to present. The data series are all in real terms and provided by Statistics Canada. To achieve stationary we will follow the convention of converting all series to annualized growth rates. From the simulation results in the previous chapter it would appear that the distinctions between these estimators are most apparent when the sample size is small relative to the number of predictors. For this reason, we use a 'rolling window' methodology where by the regression is estimated over a fixed number of the most recent observations in the time series. In particular, we use observations $t-12$ to $t-1$ to estimate the regression parameters. Then these and the predictor series at time t are used to create a forecast for GDP as time t. Rolling windows of this sort are often used in practice when a forecaster is concerned about regime changes or time varying parameters.

In the tables below we show MSFE results for four different model specifications using net exports and different investment measures as predictors to forecast GDP. For each model MSFE's are presented for OLS, the ridge estimator with Hoerl and Kennard (1970) selection of λ and the modified ridge estimator with $S = 1 - \frac{1}{n-p-1}$. The modified ridge appears to be a bit more competitive in this example relative to the simulations from the previous chapter. This may be because of either lower correlations between the predictors, or errors-in-variables. The correlation matrix of the predictors is in section 5.4 of the appendix, and shows a mix of small and large values. An errors-in-variables explanation is also possible, Statistics Canada constructs GDP data in two methods and documents the statistical discrepancy.

Table 4.1: Comparison of MSFE for OLS and ridge regression with Hoerl and Kennard (1970) selection of λ in ARMA(1,1)

Sample size	MSFE($Ridge$)/MSFE(OLS) - dependent
Model 1	0.877
Model 2	0.880
Model 3	0.784
Model 4	0.878

- Model 1 predictors: net exports, housing investment, machinery and equipment investment, and non-residential construction;
- Model 2 predictors: net exports, business investment, housing investment, machinery and equipment investment;
- Model 3 predictors: net exports, non-residential construction, business investment, machinery and equipment investment;

- Model 4 predictors: net exports, non-residential construction, housing investment, business investment.

Table 4.2: Comparison of MSFE for OLS and modified ridge with $S = 1 - \frac{1}{n-p-1}$ in ARMA(1,1)

Sample size	MSFE($ModifiedRidge$)/MSFE(OLS)
Model 1	0.890
Model 2	0.857
Model 3	0.477
Model 4	0.862

- Model 1 predictors: net exports, housing investment, machinery and equipment investment, and non-residential construction;
- Model 2 predictors: net exports, business investment, housing investment, machinery and equipment investment;
- Model 3 predictors: net exports, non-residential construction, business investment, machinery and equipment investment;
- Model 4 predictors: net exports, non-residential construction, housing investment, business investment.

4.2 Combining Economic and Financial Forecasts

A common practice in finance and economics is to forecast a time series by combining several available forecasts from different sources with an attempt to account for

differences in accuracy and cross correlations. In section 1.5, we reviewed some of the more common ways that this is done. Forecast combinations, generally involve a relatively small number of observations and if several candidate forecasts are available the number of explanatory variables can become large relative to the sample size. In addition, the predictor variables contain errors by definition. Because the forecasts do not have a causal relationship with the variable they are forecasting, the information contained in these series is accompanied by some amount of error. Consequently, this could be a situation where the ridge estimator is outperformed by a model averaging approach. In fact, this would not be surprising, as this is the application that motivated the development of many of the model averaging techniques.

In this section, we will conduct a common forecast combination exercise with the OLS estimator, the model averaging estimator with simple average weighting and the ridge estimator with Hoerl and Kennard (1970) selection of λ. The exercise will take forecasts from financial institutions that are recorded in Bloomberg and treat them as explanatory variables to be combined into a composite forecast by both the three estimators. We chose dependent variables that are commonly examined with forecast combinations in this manner, particularly, commodity prices, exchange rates and gross domestic products. In order to achieve stationarity, all variables that were not expressed in growth rates were converted to a quarterly growth rate. Because these forecasts were released at different points in time, we constructed quarterly data that consisted of the most recent forecast produced by each firm that quarter. At the end of every quarter, we used these quarterly series to forecast the subsequent quarter. The forecasts were calculated in a strictly real-time environment where the estimation only utilized information that would have been available at that moment in time. We examined the series over the largest available sample. All of the samples begin between 2006Q3 and 2007Q3, and ended in 2011Q1. The dimensions of the design matrix varied with data availability. However, number of explanatory variables p ranged between two and six, and the number of observations n ranged from four

to fourteen. The predictor series were selected to maximize $n \times p$, subject to the constraints that $n > p$, and n is at least four. The first forecast of each variable was conducted with $n = 4$, and n was increased by one observation at a time to simulate a series of real-time forecasts. An unusual feature of this exercise, is the requirement to estimate a model on such an extremely small sample. However, this particular example is an estimation that is commonly conducted in practice.

Tables 4.3 and 4.4 present the results of the three estimators. The model averaging estimator with simple average weighting produces noticeably smaller MSFE for most of the series examined then either the OLS or ridge with Hoerl and Kennard (1970) selection of λ. We take this to be an example of a practical application where small sample size in conjunction with errors-in-variables creates a situation where a model averaging estimator outperforms a ridge or OLS.

Table 4.3: Forecasting results for OLS and model averaging estimator with simple average weighting

MSFE for model average with simple average weighting/MSFE for OLS					
Commodities		**Gross Domestic Product**		**Exchange Rates**	
Oil	0.001	Canada	0.478	Euro	0.286
Gold	2.270	US	0.205	Pound	0.305
Silver	0.001	UK	0.027	Yen	0.132
Aluminum	0.076	EU	0.298	Swiss Franc	0.671
Copper	0.053	Japan	0.115	NA	NA

Table 4.4: Forecasting results for OLS and ridge estimator with Hoerl and Kennard (1970) selection of λ

MSFE for ridge estimator with Hoerl and Kennard λ/MSFE for OLS					
Commodities		**Gross Domestic Product**		**Exchange Rates**	
Oil	0.964	Canada	0.871	Euro	0.885
Gold	0.661	US	0.802	Pound	0.621
Silver	1.000	UK	0.970	Yen	0.914
Aluminum	0.951	EU	0.719	Swiss Franc	0.818
Copper	0.988	Japan	0.751	NA	NA

4.3 ARMA Models

An ARMA model is a very commonly used type of time series model across several fields. Additionally, it is a model that utilizes stochastic predictors by definition. Depending on the specification of an ARMA model, the predictor variables can have a very low correlation to each other. Consequently, some ARMA models may lend themselves to being estimated by the modified ridge estimator. Although, the example below is not in itself an application, there are likely to be applications for any estimator which can improve the performance of some ARMA models.

Consider the ARMA(1,1) shown below:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 \varepsilon_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, n,$$

where β_1, β_2 are parameters to be estimated and ε_t, $i = 1, \ldots, n$, are independent, identically distributed normal random variables with mean 0 and variance σ^2.

We conduct a simulation similar to simulation 3 in the previous chapter, where we forecast an ARMA(1,1) series with $\beta_1 = 0.95, \beta_2 = -0.1$. The choice of $\beta_1 = 0.95, \beta_2 = -0.1$ induces a low amount of correlation between predictors while similar to many values used in practice. The results below show that the modified ridge $S = 1 - \frac{1}{n-p-1}$ slightly outperforms the ridge with the Hoerl and Kennard (1970) selection of λ, or an OLS. Although, admittedly this is an extremely marginal improvement over the previous methods.

Table 4.5: Comparison of MSFE for OLS and ridge regression with Hoerl and Kennard (1970) selection of λ in ARMA(1,1)

Sample size	MSFE($Ridge$)/MSFE(OLS) - dependent
$n = 10$	0.976
$n = 25$	1.014
$n = 50$	1.004
$n = 100$	1.001

Table 4.6: Comparison of MSFE for OLS and modified ridge with $S = 1 - \frac{1}{n-p-1}$ in ARMA(1,1)

Sample size	MSFE($ModifiedRidge$)/MSFE(OLS)
$n = 10$	0.962
$n = 25$	0.998
$n = 50$	1.000
$n = 100$	1.000

Chapter 5

Conclusion

Statistical models are often faced with a trade-off between bias and variance, in order to maximize their performance. This problem is often addressed with estimation constraints in the form of ridge type estimators or related methods; such as model averaging estimators. In this thesis, we examined a number of the properties of these estimators, including some less frequently considered situations using stochastic predictors, such as errors-in-variables specifications.

We conducted a number of simulation studies comparing the relative performance of these estimators. We found that the ridge estimator often outperforms the OLS when the sample size is small relative to the number of predictors. However, within this set of circumstances, there are some cases where the ridge estimator is outperformed by alternatives. These cases often involve very small samples, or errors-in-variables. Finally, we offer some examples of applications where these special cases are potentially meaningful.

Appendix

5.1 Appendix: Simulation 3 with Errors-In-Variables

In this section we repeat simulation 3 with an errors in variables specification.

Specifically, the simulation procedure runs as follows:

1. The parameters: $\beta_1 = \cdots = \beta_p = 1$; $\sigma = 1$, and several choices of n and p.

2. We generate $X_i^*, i = 1, \ldots, n$, from $N(0,1)$, and $\eta_{i1}, \ldots, \eta_{ip}$ from $N(0,1)$.

3. We generate $\tilde{X_{i1}}, \ldots, \tilde{X_{ip}}$, $i = 1, \ldots, n$ from $N(0,1)$.

4. We generate $X_{i1}, \ldots, X_{ip}$, $i = 1, \ldots, n$ as a linear combination of X^* and $\tilde{X}$. Such that $X_{i1} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{i1}} + \eta_{i1}$ and $X_{ip} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{ip}} + \eta_{ip}$.

5. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.

6. We use the sample $n-1$ observations to compute λ by the approaches of Hoerl and Kennard (1970), Lawless and Wang (1976), Khalaf and Shukur (2005). These equations are in section 2.1.3.

7. We use the sample $n-1$ observations and the three choices of λ to compute the ridge estimates of $\beta_1, \ldots, \beta_p$.

8. We use the resulting $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and $X_{n1} \ldots X_{np}$ to compute the corresponding predictions $\hat{Y_n}$ of Y_n.

9. Compute the squared forecast errors $(\hat{Y}_n - Y_n)^2$ for the different ridge estimators.

10. Repeat steps 2-9 $m = 100000$ times (so that new predictors are used in each simulation).

11. Compute the resulting mean squared forecast error for each of the ridge estimators.

The results of the simulation are reported in Tables 5.1 to 5.4.

Table 5.1: Comparison of MSFE for OLS and ridge regression with Hoerl and Kennard (1970) selection of λ under an errors-in-variables specification

MSFE for Ridge with Hoerl and Kennard λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.945	0.866	0.785	NA
$n = 25$	0.997	0.983	0.965	0.573
$n = 50$	0.999	0.996	0.991	0.825
$n = 100$	1.000	0.999	0.997	0.946

5.2 Appendix: Simulation 4 with Errors-In-Variables

In this section we repeat simulation 4 with an errors in variables specification.

Specifically, the simulation procedure runs as follows:

1. The parameters: $\beta_1 = \cdots = \beta_p = 1$; $\sigma = 1$, and several choices of n and p.

2. We generate $X_i^*, i = 1, \ldots, n$, from $N(0, 1)$, and $\eta_{i1}, \ldots, \eta_{ip}$ from $N(0, 1)$.

3. We generate $\tilde{X_{i1}}, \ldots, \tilde{X_{ip}}$, $i = 1, \ldots, n$ from $N(0, 1)$.

Table 5.2: Comparison of MSFE for OLS and ridge regression with Lawless and Wang (1976) selection of λ under an errors-in-variables specification

MSFE for Ridge with Lawless and Wang λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.984	0.964	0.931	NA
$n = 25$	1.000	0.999	0.999	0.911
$n = 50$	1.000	1.000	1.000	0.999
$n = 100$	1.000	1.000	1.000	1.000

Table 5.3: Comparison of MSFE for OLS and ridge regression with Khalaf and Shukur (2005) selection of λ under an errors-in-variables specification

MSFE for Ridge with Khalaf and Shukur λ/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.953	0.970	0.952	NA
$n = 25$	0.995	0.999	0.998	0.962
$n = 50$	0.999	1.00	1.000	0.996
$n = 100$	1.000	1.000	1.000	1.000

Table 5.4: Comparison of MSFE for OLS and modified ridge regression with $S = 1 - \frac{1}{n-p-1}$ under an errors-in-variables specification

MSFE for Modified Ridge with $S = 1 - \frac{1}{n-p-1}$/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.958	0.865	0.725	NA
$n = 25$	0.998	0.993	0.987	0.257
$n = 50$	1.000	0.999	0.998	0.924
$n = 100$	1.000	1.000	1.000	0.994

4. We generate $X_{i1}, \ldots, X_{ip}$, $i = 1, \ldots, n$ as a linear combination of X^* and $\tilde{X}$. Such that $X_{i1} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{i1}} + \eta_{i1}$ and $X_{ip} = \sqrt{0.8}X_i^* + \sqrt{0.2}\tilde{X_{ip}} + \eta_{ip}$.

5. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.

6. We then use predictors $1, \ldots, \frac{p}{2}$ to form the first auxiliary model, and predictors $\frac{p}{2} + 1, \ldots, p$ to form the second auxiliary model.

7. We use the sample $n - 1$ observations to compute OLS estimates of $\beta_1, \ldots, \beta_{\frac{p}{2}}$ in the first auxiliary model, and $\beta_{\frac{p}{2}+1}, \ldots, \beta_p$ in the second auxiliary model.

8. We then compute $\hat{Y_{1,g}}, \ldots, \hat{Y_{n-1,g}}$ for each of the two auxiliary models where $g = 1, \ldots, 2$.

9. The two auxiliary models are assigned weights q_g using their fitted values of $\hat{Y_{1,g}}, \ldots, \hat{Y_{n-1,g}}$, this is done using one of the methods described in section 1.5.2, specifically simple average weights, inverse mean squared error weights or OLS weights.

10. We use the resulting $\hat{\beta}_1, \ldots, \hat{\beta}_{\frac{p}{2}}$ and $X_{n1} \ldots X_{n\frac{p}{2}}$ to compute the corresponding

prediction $\hat{Y_{n1}}$ of Y_n from the first auxiliary model. Then $\hat{\beta}_{\frac{p}{2}+1}, \ldots, \hat{\beta}_p$ and $X_{n\frac{p}{2}+1} \ldots X_{np}$ is used to compute the prediction $\hat{Y_{n2}}$ of Y_n from the second auxiliary model.

11. The final prediction $\hat{Y_n}$ of Y_n is computed as $\hat{Y_n} = q_1\hat{Y_{n1}} + q_2\hat{Y_{n2}}$

12. Compute the squared forecast error $(\hat{Y_n} - Y_n)^2$ for the estimator.

13. Repeat steps 2-12 $m = 100000$ times (so that new predictors are used in each simulation).

14. Repeat steps 2-13 for each of the three methods of determining model weights, to determine the mean squared forecast error for each of the three estimators.

The results of the simulation are reported in Tables 5.5 to 5.7.

Table 5.5: Comparison of MSFE for OLS and model average with simple average weights under an errors-in-variables specification

MSFE for model average with simple average weighting/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.865	0.778	0.624	NA
$n = 25$	1.025	1.037	0.986	0.172
$n = 50$	1.058	1.093	1.058	0.771
$n = 100$	1.078	1.119	1.089	0.940

Table 5.6: Comparison of MSFE for OLS and model average with inverse mean squared error weights under an errors-in-variables specification

MSFE for model average with inverse MSE weights/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.889	0.806	0.670	NA
$n = 25$	1.030	1.027	1.003	0.189
$n = 50$	1.066	1.078	1.063	0.788
$n = 100$	1.079	1.093	1.097	0.947

Table 5.7: Comparison of MSFE for OLS and model average with OLS weights under an errors-in-variables specification

MSFE for model average with OLS weights/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	1.000	0.906	0.736	NA
$n = 25$	1.000	0.986	0.955	0.186
$n = 50$	1.000	0.994	0.982	0.768
$n = 100$	1.000	0.997	0.992	0.907

5.3 Appendix: Modified Ridge under Low Predictor Correlation

In this section of the appendix simulation 3 is repeated for the modified ridge with a lower amount of correlation between predictors, specifically 0.2.

The simulation procedure runs as follows:

1. The parameters: $\beta_1 = \cdots = \beta_p = 1$; $\sigma = 1$, and several choices of n and p.
2. We generate $X_i^*, i = 1, \ldots, n$, from $N(0, 1)$.
3. We generate $\tilde{X_{i1}}, \ldots, \tilde{X_{ip}}$, $i = 1, \ldots, n$ from $N(0, 1)$.
4. We generate $X_{i1}, \ldots, X_{ip}$, $i = 1, \ldots, n$ as a linear combination of X^* and $\tilde{X}$. Such that $X_{i1} = \sqrt{0.2}X_i^* + \sqrt{0.8}\tilde{X_{i1}}$ and $X_{ip} = \sqrt{0.2}X_i^* + \sqrt{0.8}\tilde{X_{ip}}$. This creates dependent predictors where the correlation between any two is 0.2.
5. We generate ε_i, $i = 1, \ldots, n$. We compute Y_i.
6. We compute the tuning parameter as $S = 1 - \frac{1}{n-p-1}$.
7. We use the sample $n - 1$ observations and S to compute the modified ridge estimates of $\beta_1, \ldots, \beta_p$.
8. We use the resulting $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and $X_{n1} \ldots X_{np}$ to compute the corresponding predictions $\hat{Y_n}$ of Y_n.
9. We compute the squared forecast errors $(\hat{Y_n} - Y_n)^2$ for the modified ridge estimator.
10. We repeat steps 2-9 $m = 100000$ times (so that new predictors are used in each simulation).

11. We compute the resulting mean squared forecast error for the modified ridge estimator.

Table 5.8: Comparison of MSFE for OLS and modified ridge regression with $S = 1 - \frac{1}{n-p-1}$ under predictor correlation of 0.2

MSFE for Modified Ridge with $S = 1 - \frac{1}{n-p-1}$/MSFE for OLS				
Number of Predictors	2	3	4	20
$n = 10$	0.972	0.898	0.785	NA
$n = 25$	0.999	0.997	0.994	0.587
$n = 50$	1.000	1.000	0.999	0.976
$n = 100$	1.000	1.000	1.000	1.003

5.4 Appendix: Correlation of Predictors from 4.1

Table 5.9: Table of predictor correlation from example in 4.1

Predictor	housing	business	non-residential	machinery	net exports
housing	1.00	0.20	0.13	0.21	0.10
business	0.20	1.00	0.80	0.89	0.39
non-residential	0.13	0.80	1.00	0.44	0.19
machinery	0.21	0.89	0.44	1.00	0.43
net exports	0.10	0.39	0.19	0.43	1.00

5.5 Appendix: Simulation Codes

This section of the appendix contains codes used for our numerical experiments. They are written in R and EVIEWS language. EVIEWS is a regression software used in regression and time series analysis.

R code for simulation one:

```
n=100;
m=1000;
sigma=2;
lambda=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1);

# Fixed design OLS+ridge

Design=matrix(c(rep(0,2*n)),2,n);
X1=rnorm(n);
X2=rnorm(n);
Design[1,]=X1;
Design[2,]=X2;
Design=t(Design);

OLS=t(Design)%*%Design;

mse=0;
total.mse=0;
total.mse.ridge=c(rep(0,length(lambda)));
total.mse.ridge.residuals=c(rep(0,length(lambda)));
```

```
for(i in 1:m){
errors=sigma*rnorm(n);
beta1=1;
beta2=1;
Y=beta1*X1+beta2*X2+errors;
#beta1.estim=lsfit(t(Design),Y)$coefficients[2];
#beta2.estim=lsfit(t(Design),Y)$coefficients[3];

for(j in 1:length(lambda))
{
Lambda=matrix(c(lambda[j],0,0,lambda[j]),nrow=2,ncol=2);
Ridge=OLS+Lambda;
beta.ridge=solve(Ridge)%*%t(Design)%*%Y;
beta1.estim.ridge=beta.ridge[1]; beta2.estim.ridge=beta.ridge[2];
residuals.ridge=Y-beta1.estim.ridge*X1-beta2.estim.ridge*X2;

total.mse.ridge[j]=total.mse.ridge[j]+(beta1.estim.ridge-beta1)^2
    + (beta2.estim.ridge-beta2)^2
total.mse.ridge.residuals[j]=total.mse.ridge.residuals[j]+sum(residuals.ridge^2);
}

beta=solve(OLS)%*%t(Design)%*%Y;
beta1.estim=beta[1]; beta2.estim=beta[2];
total.mse=total.mse+(beta1.estim-beta1)^2 + (beta2.estim-beta2)^2

}
mse=total.mse/m;
mse.ridge=total.mse.ridge/m;
```

```
mse.ridge.residuals=total.mse.ridge.residuals/m

par(mfrow=c(1,2))
plot(lambda,mse.ridge,ylab="MSE",main="Fixed design");
plot(lambda,mse.ridge,ylab="MSE",main="Fixed design - residuals");

print(min(mse.ridge)/mse);

# Random design OLS+ridge

mse=0;
total.mse=0;
total.mse.ridge=c(rep(0,length(lambda)));

for(i in 1:m){

Design=matrix(c(rep(0,2*n)),2,n);
X1=rnorm(n);
X2=rnorm(n);
Design[1,]=X1;
Design[2,]=X2;
Design=t(Design);

OLS=t(Design)%*%Design;

errors=sigma*rnorm(n);
beta1=1;
beta2=1;
```

```
Y=beta1*X1+beta2*X2+errors;
#beta1.estim=lsfit(t(Design),Y)$coefficients[2];
#beta2.estim=lsfit(t(Design),Y)$coefficients[3];

for(j in 1:length(lambda))
{
Lambda=matrix(c(lambda[j],0,0,lambda[j]),nrow=2,ncol=2);
Ridge=OLS+Lambda;
beta.ridge=solve(Ridge)%*%t(Design)%*%Y;
beta1.estim.ridge=beta.ridge[1]; beta2.estim.ridge=beta.ridge[2];
total.mse.ridge[j]=total.mse.ridge[j]+(beta1.estim.ridge-beta1)^2
    + (beta2.estim.ridge-beta2)^2
}

beta=solve(OLS)%*%t(Design)%*%Y;
beta1.estim=beta[1]; beta2.estim=beta[2];
total.mse=total.mse+(beta1.estim-beta1)^2 + (beta2.estim-beta2)^2
}

mse=total.mse/m;
mse.ridge=total.mse.ridge/m;

print(mse);
print(mse.ridge);
plot(lambda,mse.ridge,ylab="MSE",main="Random design");
```

EVIEWS Code for the two predictor trials in simulation three:

```
'Step one create an unstructured workfile with
```

```
    the desired number of observations you want to use.

scalar alpha_level=0.00 'this is always zero in the thesis

scalar shrink=0.75  'insert scalar

scalar obs=100      'insert number of observations
scalar obs_l1=obs-1

scalar totalsqols=0
scalar totalsqnew=0
vector(100000) errors_new
vector(100000) errors_new_squared
vector(100000) errors_ols
vector(100000) errors_ols_squared
vector(100000) b1_new
vector(100000) b2_new
vector(100000) b1_ols
vector(100000) b2_ols
vector(100000) covar_ols

for !i=1 to 100000

'Create data
series common=nrnd

series x1diff=nrnd
series x1info=(0.8^0.5)*common+(0.2^0.5)*x1diff
```

```
series x1err=nrnd  'for error-in-variable simulations
vector x1full=x1info+x1err
vector(obs_l1) x1= @subextract(x1full,1,1,obs_l1,1)
scalar x1mean=@mean(x1)
x1=x1-x1mean
x1full=x1full-x1mean

series x2diff=nrnd
series x1info=(0.8^0.5)*common+(0.2^0.5)*x2diff
series x2err=nrnd    'for error-in-variable simulations
vector x2full=x2info+x2err
vector(obs_l1) x2= @subextract(x2full,1,1,obs_l1,1)
scalar x2mean=@mean(x2)
x2=x2-x2mean
x2full=x2full-x2mean

series yerr=nrnd
vector yfull=x1info+x2info+yerr
vector(obs_l1) y= @subextract(yfull,1,1,obs_l1,1)
scalar ymean=@mean(y)
y=y-ymean
yfull=yfull-ymean

matrix(obs_l1,2) x
matplace(x,x1,1,1)
matplace(x,x2,1,2)

'ols betas
```

```
vector bols= @inverse(@transpose(x)*x)*@transpose(x)*y

'new betas

matrix covarmat=@transpose(x)*x
covar_ols(!i)=covarmat(1,2)

't test - t test not used in the thesis, alpha=0

matrix var_temp=@emult(x1, x2)
scalar mean_temp=@mean(var_temp)
scalar std_temp=@stdev(var_temp)

scalar teststat_temp= (mean_temp-0)/(std_temp/(obs_l1^0.5))
scalar pvaule_temp=@ctdist(teststat_temp,obs_l1-1)

    if pvaule_temp<1-alpha_level then
        covarmat(1,2)=shrink*covarmat(1,2)
        covarmat(2,1)=shrink*covarmat(2,1)
        endif

delete var_temp
delete mean_temp
delete std_temp
delete teststat_temp
delete pvaule_temp
```

```
vector bnew= @inverse(covarmat)*@transpose(x)*y

scalar errornew=yfull(obs)-(x1full(obs)*bnew(1)+x2full(obs)*bnew(2))

scalar errorols=yfull(obs)-(x1full(obs)*bols(1)+x2full(obs)*bols(2))

b1_new(!i)=bnew(1)
b2_new(!i)=bnew(2)
b1_ols(!i)=bols(1)
b2_ols(!i)=bols(2)

totalsqnew=totalsqnew+errornew*errornew
totalsqols=totalsqols+errorols*errorols

errors_new(!i)=errornew
errors_new_squared(!i)=errornew*errornew
errors_ols(!i)=errorols
errors_ols_squared(!i)=errorols*errorols

next

series z_totalsqnew=totalsqnew
series z_totalsqols=totalsqols
```

Bibliography

[1] Bates, J. M.; Granger, C. The Combination of Forecasts. OR Vol. 20 (1969), no. 4, 451-468, Operational Research Society.

[2] Brown, P.; Zidek, J. Adaptive multivariate ridge regression. The Annals of Statistics, Vol. 8 (1980), no. 1, 64-74.

[3] Bunn, Derek; Forecasting with more than one model. Journal of Forecasting, Vol. 8 (1989), 161-166.

[4] Efron, Bradley; Hastie, Trevor; Johnstone, Iain; Tibshirani, Robert Least Angle Regression. Annals of Statistics, Vol. 32 (2004), no. 2, 407-499.

[5] El-Salam, Moawad El-Fallah Abd; An Efficient Estimation Procedure For Determining Ridge Regression Parameter. Asian Journal of Mathematics and Statistics, 4 (2011), no. 2, 90-97.

[6] Greene, William H. Econometric Analysis. 6th edition. Pearson, New Jersey, 2008.

[7] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome The Elements of Statistical Learning. Springer, New York, 2001.

[8] Hoerl, A.; Kennard, Robert Biased estimation for nonorthogonal problems. Technometrics, Vol. 12 (1970), no. 1, 55-67.

[9] Khalaf, G; Shukur, G; Choosing Ridge Parameter for Regression Problem. Communications in Statistics - Theory and Methods, 34 (2005), 1177-1182.

[10] Kim, Hyun; Swanson, Norman Forecasting financial and macroeconomic variables using data reduction methods. Unpublished. (2010).

[11] Lai, Tze Leung; Wei, Ching Zong. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Ann. Statist. 10 (1982), no. 1, 154166.

[12] Lawless, J. F. Mean squared error properties of generalized ridge estimators. J. Amer. Statist. Assoc. 76 (1981), no. 374, 462466.

[13] Lawless, J. F.; Wang, P; A Simulation Study of Ridge and Other Regression Estimators. Communications in Statistics - Theory and Methods, 14 (1976), 1589-1604.

[14] Lindley, D.; Smith, A. Bayes estimates for the linear model. Journal of the Royal Statistical Society, Vol. 34 (1972), no. 1, 1-41.

[15] Little, Roderick; Rubin, Donald Statistical Analysis with Missing Data. Wiley, New York, 1987.

[16] Lzenman, Alan J. Modern Multivariate Statistical Techniques. Springer, New York, 2008.

[17] Markowitz, H. Portfolio selection. The Journal of Finance, Vol. 7 (1952), no 1. 77-91.

[18] Stock, James H. and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, John Wiley and Sons, Ltd., vol. 23 (2004), no. 6, 405-430.

[19] Timmermann, Allan G. Forecast combinations. CEPR Discussion Paper 5361 (2005), C.E.P.R. Discussion Papers.

[20] Vinod, Hrishikesh; A survey of ridge regression and related techniques for improvements over ordinary least squares. The Review of Economics and Statistics, Vol. 60 (1978), no. 1, 121-131.

[21] Wei, C. Z. Asymptotic properties of least-squares estimates in stochastic regression models. Ann. Statist. 13 (1985), no. 4, 14981508.

[22] Wright, Jonathan H. Forecasting US inflation by Bayesian model averaging. Journal of Forecasting. John Wiley and Sons, Ltd., vol. 28 (2009), no. 2, 131-144.