High Dimensional Machine Learning Methods in Economics: A Monte Carlo Simulation

JOSEPH PAUL
HWU ID: H00269748

MA Economics

Honours Dissertation

Supervised by Prof. MARK SCHAFFER

Word Count: 10,335

HERIOT-WATT UNIVERSITY

Edinburgh Business School

Department of Economics

April 08 2021

The copyright in this dissertation is owned by the author. Any quotation from the dissertation or use of any of the information contained in it must acknowledge it as the source of the quotation or information.

Abstract

It is becoming increasingly common for economic researchers to be presented with ‘complex’ data in empirical economic work. One form of this is high-dimensional data. These data sets often require different methods than those traditionally used by economists to estimate causal effects.

In this paper, I analyse two recent estimation methods taken from the machine learning and econometric literature; the Post-double-selection (PDS) estimator and the Post-LASSO, and asses their ability to estimate average treatment effect in high dimensional data. This is done through an extensive Monte Carlo simulation that compares the two estimators’ performance in terms of root mean squared error, mean estimate, bias, and variance. The results show that in most high-sparsity experiments, neither estimator produces reliable results. However, in medium to low sparsity experiments, the PDS method provides accurate estimates and correctly selects close to the true model, whereas the Post-LASSO often misspecifies the model, resulting in higher bias.

Declaration

As the author of this work I acknowledge and understand the penalty for any found collusion or plagiarism.

Signed: *Joseph Paul*

Date: *April 08, 2021*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

This dissertation relates to three main topics, being causal inference, machine learning, and high-dimensional data. Firstly, in economics, many empirical studies aim to estimate the causal effect of a treatment on an outcome. However, when the treatment is not randomly assigned research must rely on quasi-experimental methods to estimate the causal effect (Athey and Guido W. Imbens 2017). One such approach is to take the treatment as exogenous after conditioning on a set of controls. Researchers have traditionally relied intuition and economic theory to select of controls. However this becomes a problem if too many are chosen as the model will be over-fitted, too few and the model will be under-fitted. The outcome from both events will cause the estimates to be biased.

Secondly, high dimensional are data sets in which there are a large number of parameters relative to the sample size ($p > n$). They arise naturally in situations where many characteristics per observation are collected, commonly associated with big data (Athey and G. Imbens 2019). High-dimensional models can also arise in data sets with a small number of measured characteristics where the exact functional form on the measured variables is unknown. It is, therefore, easy to reach a large number of variables when including splines, dummy variables, interactions, and polynomials. For example, 20 initial continuous variables can results in 1790 variables when including second and third-degree polynomials and interactions. Belloni and Chernozhukov (2011) argue that $p > n$ is common in economics, although it is often not explicitly acknowledged. For a long time, the field of machine learning has built methods for working with these complex data sets. A key difference between many econometric approaches and supervised machine learning approaches is that they employ data-driven model selection with the goal of predictive performance (Athey and Guido W. Imbens 2017). These methods therefore do not translate naturally for use in causal inference, as they are often biased and rarely produce asymptotically normal estimates.

Furthermore, high-dimensional data sets present new challenges to social science researchers when estimating causal effects if relying on traditional statistical/econometric tools as theory is often based on the assumption of $p < n$. In response to this, a new selection of methods have become available to the empirical researcher, based on the machine learning literature. One such tool that has been frequently used outside of the econometrics is the Least Absolute Shrinkage and Selection Operator or LASSO.

The LASSO is a penalised regression method that simultaneously performs model selection and estimation and was first introduced by Robert Tibshirani (1996). The LASSO has seen widespread popularity due to improvements in prediction accuracy over OLS and interpretability due to a model with fewer variables (ibid.). However, a fundamental assumption of the LASSO's use is that the true model has a sparse representation. A sparse model is one in which only a few coefficients have non-zero values (O'Brien 2016). However, the LASSO introduces bias into the model as all coefficients are shrunk towards zero. This means the standard LASSO does not translate well into situations that call for causal inference.

One method that address this bias is the Post-LASSO, which runs OLS on the model selected by the LASSO, providing an unbiased estimate of the coefficients. Under certain assumptions, this is sufficient to produce valid estimates for average treatment effects. However, Zhao and Yu (2006) show that the LASSO will not be able to select the 'true' model if a variable is highly correlated with the predictors but only mildly associated with the outcome.

In response to this, Belloni, Chernozhukov, and Hansen (2014b) extend the Post-LASSO and propose the post-double selection (PDS) estimator for use when trying to estimate average treatment effects. The first step of the method is to use the LASSO to select covariates correlated with the outcome as controls in the same way Post-LASSO does. Secondly, another

LASSO regression is used to select covariates that are correlated with the treatment. Finally, OLS is used to regress the outcome on the treatment and the controls selected by the LASSO in the first stages. The inclusion of variables found by regressing the treatment on potential controls guards against the exclusion of any covariates that are only mildly associated with the outcome but strongly associated with the treatment. This is meant to improve the average treatment effect estimate and minimise the risk of omitted variable bias. The work of Belloni, Chernozhukov, and Hansen (2014b) shows that the post-double-selection method can be used to produce valid causal estimates under quite plausible assumptions.

1.2 Research and Objectives

The motivation and aim of this paper is to explore the statistical properties of the post-double-selection estimator proposed by Belloni, Chernozhukov, and Hansen (ibid.) and the Post-LASSO. It also examines the relationship between the performance of these estimators and sparsity. To do this, I present a method familiar to economists but novel to the application of measuring sparsity in high dimensional data sets, the Gini coefficient. The Gini coefficient has traditionally been used as an index to measure the degree of inequality. As I show, the coefficient translates naturally to measuring the sparsity of data sets.

The methodological approach of this study takes the form of a Monte Carlo simulation that generates data sets through randomly simulated draws from a given distribution. The Monte Carlo simulation approach allows for the analysis of the estimators' performance on simulated data sets as the true population parameters are known by design. I use three distinct data generating processes and 136 parameter sets, which were chosen to test the model in different situations and sparsity levels. I compare the estimators in terms of root mean squared error, bias, variance and mean estimate. Graphical results in the form of density plots are also presented.

Overall, the results show that neither estimator was able to perform better than the other across all experiments. However, the PDS is a more robust tool for estimating the average treatment effects than the Post-LASSO. The use of the Gini coefficient value is negatively related to the performance of both estimators; however, as a measure of the performance of the estimator, its use varied depending on the functional form of the underlying model. Several other factors, such as sample size and the amount of covariance, were also important in determining the estimators' performance.

This paper adds to the rapidly growing literature at the intersection of machine learning and econometrics. Understanding the applications and limitations of any estimator is of great importance when doing applied work, and this paper will help give researchers greater insight and a better intuitive understanding of these methods.

1.3 Outline

The outline for the rest of this dissertation is as follows, Chapter 2 provides the base theory for this study and introduces various concepts related to machine learning and econometrics. Chapter 3 discusses the methodology used in this study, chapter 4 presets the results which are discussed in chapter 5 and finally chapter 6 concludes.

Chapter 2

Literature Review

This chapter introduces the relevant topics for this dissertation. We start by looking at causal inference, followed by an introduction to Machine Learning and finally, the LASSO estimator and its use in causal inference.

2.1 Causal Inference

2.1.1 Rubin Causal Model

The Rubin causal model is a powerful framework for thinking about causality and is used in most of the empirical work done in social and biomedical sciences (Guido W Imbens and Rubin 2015). It builds on the potential outcome framework first proposed by Neyman (1923) with the fundamental notion being that causality arises when an action is applied to a unit. A unit can be an object, individual or even a country at a point in time.

Holland (1986) outlines the potential outcomes model used for causal inference as follows. $Y_i(W)$ denotes the potential outcome for unit i that receives treatment W_i, where $W_i = 1$ if the unit receives the treatment and $W_i = 0$ if it does not. The unit level causal effects is the difference between the potential outcomes for unit i[1]:

$$\tau_i = Y_i(1) - Y_i(0) \tag{2.1}$$

[1] The unit-level effects can also me measured as $\frac{Y_i(1)}{Y_i(0)}$ in some settings (Rubin 2005)

We are unable to view the same unit at the same point in time while having been exposed to both the treatment and no treatment. This is what Holland (1986, p.947) has called the 'Fundamental Problem of Causal Inference'. We, therefore, have to exploit differences in the outcomes between units where some are exposed to the action (treated group), and some are not exposed (control group). To make these comparisons, however, we must make certain assumptions about our data and the assignment mechanism, the first one being the stable unit treatment value assumption.

Assumption: SUTVA

The Stable Unit Treatment Value Assumption (SUTVA), also known as the Stability Assumption, is needed to exploit the differences between different units for estimating causal effects (Guido W Imbens and Rubin 2015). Intuitively, this means that the treated effects of other units do not affect the outcome for unit i; in other words that $Y_i(0)$ or $Y_i(1)$ is not affected by the treatment of other units. Secondly, there is only a single level of treatment and that no matter how treatment W is received, $Y_i(1)$ will not change.

From this, there are many different 'summary' causal effects we can observe, such as the Average Treatment Effect (ATE) or the Median Treatment Effect or the Conditional Average Treatment Effect (CATE) (Rubin 2005).

2.1.2 Assignment Mechanisms

The assignment mechanism is the process by which it is determined which units will receive treatments and which will not (Guido W Imbens and Rubin 2015). Rubin (2005) formally defines the assignment mechanism as follows:

Assignment Mechanism *Given a population of n units, where W_i is the treatment assignment for unit i and $W = (W_1, ..., W_n)^T$. The assignment mechanism gives the probability of vector W given X, $Y(1)$ and $Y(0)$:*

$$Pr(W|X, Y(1), Y(0)) \tag{2.2}$$

satisfying

$$\sum_{W \in \{0,1\}^n} Pr(W|X, Y(0), Y(1)) = 1$$

To claim a causal relationship, the assignment of the treatment must satisfy the following three restrictions:

Individualistic Assignment: This limits the dependence of the treatment assignment for unit *i* so that it is not dependent on the outcomes of other units:

$$p_i(X,Y(0),Y(1)) = q(X_i,Y_i(0),Y_i(1)) \tag{2.3}$$

where $q(\cdot) \in [0,1]$, and $p_i(\cdot)$ is used to denote the unit level probability of being assigned to the treatment group.

Probabilistic Assignment: This merely requires that there is a possibility of each unit being assigned to the control or treatment group:

$$0 < p_i(X,Y(0),Y(1)) < 1 \tag{2.4}$$

Unconfounded Assignment: The assignment mechanism is unconfounded if it does not depend on the potential outcomes:

$$Pr(W|X,Y(0),Y(1)) = Pr(W|X,Y'(0),Y'(1)) \tag{2.5}$$

where $Y'(0)$ and $Y'(1)$ are alternative potential outcomes. When we have unconfounded assignment, we can drop the potential outcomes and rewrite the assignment mechanism as:

$$Pr(W|X) \tag{2.6}$$

When all three restrictions are met, it is referred to as the 'strongly ignorable treatment assignment' (Rosenbaum and Rubin 1983). Classical randomised experiments satisfy all three restrictions by design. However, these experiments are often not feasible in economics, and researchers have to use experimental data. When this is the case, assumptions have to be made about the likelihood that the restrictions hold. The third assumption usually presents the biggest challenge for researchers, as discussed in section 2.3.3 (Guido W Imbens and Rubin 2015).

2.2 Machine Learning

Machine learning (ML) is a field at the intersection of statistics and computer science that uses algorithms applied to data-sets. The primary goals have traditionally been are prediction, classification, and clustering (Athey 2018). We can further divide ML into two distinct types of algorithms, supervised learning and unsupervised learning. Supervised learning takes labelled training data ($\{y_i, x_i\}, i = 1, ..., n$) and attempts to fit a model, using a 'training set' to the dependent variable. Common algorithms used include regression trees, support vector machines, regularised regression and neural nets. On the other hand, unsupervised learning has no dependent variable and attempts to find clusters or patterns within the data. Standard methods here include k-means clustering, principal component analysis, anomaly detection and much more (ibid.).

2.2.1 Prediction

Supervised learning has primarily focused in predictive out-of-sample (OOS) performance and not the underlying structural model in the way econometricians are (ibid.). Many ML algorithms, such as Regression Trees or Neural Nets, can easily perfectly fit a model to a training set; however, this usually results in poor OOS performance due to over-specification and high variance (James et al. 2013). ML's solution to this is regularisation, which can be thought of as a penalty on model complexity. The extent of regularisation is usually chosen through cross-validation to minimise the mean squared prediction error.

In economic settings, prediction has mainly been used to forecast using time series data (Ahrens, Aitken, and Schaffer 2021). Kleinberg et al. (2015) argues that economic research has often overlooked policy questions that involve prediction, despite many policy problems that are inherently prediction problems rather than causal problems. For example, ML has been used to predict which teacher will add the greatest value (Rockoff et al. 2011), to target health inspectors at restaurants most at risk of health violations (Kang et al. 2013), and to target social interventions towards the most at-risk youths (Chandler, Levitt, and List 2011). As this paper's primary focus is using ML for causal inference, we will not focus anymore on prediction policy problems.

2.2.2 Re-sampling Methods

Re-sampling is a key part of the ML process that involves repeatedly drawing samples from a training set and refitting a model to obtain more information about the fitted model (James et al. 2013). The two most common methods employed in ML are Cross-Validation and the Bootstrap.

Cross-Validation (CV) involves randomly splitting the data set into two or more parts, usually a training set and a validation set. The training set is used to fit the model, and the validation set to gauge its effectiveness, usually measured with the Mean Squared Error (MSE). With CV techniques, we can tune our 'nuisance' parameters to get the best OOS predictions. A few different methods are discussed below.

Leave-One-Out Cross Validation

Leave-one-out (LOO) CV uses just one observation (x_1, y_1) as the validation set, and trains the model on the remaining data $[(x_i, y_i),\ i = \{2, ..., n\}]$ (ibid.). This process is repeated with every observation used once as the validation set to get the $MSE_1, ..., MSE_n$ for every observation. The leave-one-out CV estimate of the MSE is thus the average of all MSE:

$$CV_{LOO} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

As the validation set was not used to train the model, it provides an approximately unbiased estimate of the error. However, this method can be computationally expensive as the model has to be refit n times (Friedman, Hastie, and Robert Tibshirani 2001).

K-Fold Cross Validation

A less computationally expensive method is K-Fold CV. This method involves splitting the data-set into K number of distinct 'folds', typically 5 or 10, of approximately equal size. We treat the first fold K_1 as the cross-validation set and train the model of the remaining $K - 1$

folds. We then get the MSE for K_1 and repeat the process using each fold as the validation set and training on the remaining folds. It is less computationally expensive as the model needs only be refit K times. The the CV MSE estimate is calculated by:

$$CV_{kfold} = \frac{1}{K}\sum_{i=1}^{K} MSE_i$$

As we can see, LOO is just a particular case of K-fold where $K = n$. The advantage of K-fold over LOO is that it is less computationally expensive and that the MSE has a lower variance as there is less overlap in the fitted models (Friedman, Hastie, and Robert Tibshirani 2001).

The Bootstrap

The Bootstrap has slightly different aims to cross-validation (ibid.). It is a powerful method that can be applied to a wide range of algorithms to measure an estimator's uncertainty. To produce a bootstrapped data-set Z, we randomly select n' observations, where $n' < n$, from the original data-set. Sampling is done with replacement, meaning the same observation can be in a bootstrap data-set more than once. The data-set is then used to produce an estimate of the parameter of interest $\hat{\alpha}$. We repeat this process B times with $Z^1, ..., Z^B$ bootstrapped data-sets which corresponds to $\hat{\alpha}_1, ..., \hat{\alpha}_B$ estimates. From this, we can produce the standard error of the bootstrap estimates with:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^r - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{r'})^2}$$

2.3 The LASSO

The Least Absolute Shrinkage and Selection Operator, or LASSO, first proposed by Robert Tibshirani (1996) is a penalised linear regression method that shrinks coefficients, setting some to zero, by imposing a penalty on their absolute size. The LASSO objective function is:

$$\min_{\beta_0,\beta}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad s.t. \quad ||\beta||_1 \leq t \tag{2.7}$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 norm of β, p is the number of variables, and t is the user-specified parameter controlling the extent of regularisation.

Alternatively the LASSO objective function can be written it its Lagrangian form:

$$\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2.8}$$

The first part of Eq.(2.8) is the same as the OLS objective function, and λ is the regularisation parameter and controls the extent of regularisation. There is a 1-2-1 comparison of λ and t that produces the same result. When λ is equal to zero, the LASSO has the same objective fucntion as OLS. The extent of regularisation (size of λ) is usually chosen to minimise the expected prediction error and is typically done through cross-validation techniques as discussed in2.2.2.

In traditional machine learning applications, the LASSO has been used as an alternative to OLS due to improvements in:

- *Prediction Accuracy:* OLS often has low bias but large variance. We can sometimes improve prediction accuracy by shrinking some of the coefficients to zeros; however, this introduces some bias into the model.
- *Interpretability:* It can identify the subset of predictors with the strongest predictive power (O'Brien 2016).

However, the LASSO does not just shrink some coefficients to 0, but biases all $\hat{\beta}$ towards 0. This is a problem when trying to do statistical inference as the estimates are not asymptotically normally distributed (Jankova, Van De Geer, et al. 2018). Fortunately, this bias can be accounted for with the approach of Post-LASSO which usually results in better predictive performance (O'Brien 2016). The steps to implementing Post-LASSO is to first let the LASSO select the parameters, shrinking some coefficients to 0, and re-regress Y on the non-zero coefficients using OLS to remove the bias as outlined below:

$$\hat{\beta}_{post} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x'\beta)^2 \quad s.t. \quad \beta_j = 0 \iff \tilde{\beta}_j = 0 \tag{2.9}$$

where $\tilde{\beta}_j$ is the first stage LASSO β estimation.

2.3.1 LASSO and other penalty forms

Different penalty forms can be given by different values in q:

$$\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right]$$

Where $q = 1$ is the LASSO, $q = 2$ is Ridge, and $q = 0$ is Best Subset Selection.

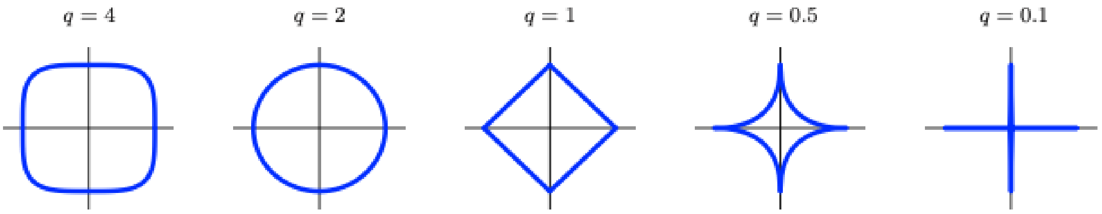


Figure 2.1: *Constraint regions for different penalty forms(O'Brien 2016)*

Ridge regression is also a commonly used method developed by Hoerl and Kennard (1970) that uses ℓ^2 regularisation, which applies a penalty to the squared values of the coefficients. As there is no kink at 0, Ridge will rarely shrink coefficient values to 0 and will therefore produce a less sparse model than LASSO regression. The best sub-set selection is a 'hard thresholding operator' that penalises the number of non-zero coefficients and by varying the size of λ, we can select the number of variables that best describe the outcome. When $q > 1$, then there is no longer a convex solution and so $\hat{\beta}$ rarely gets set to 0.

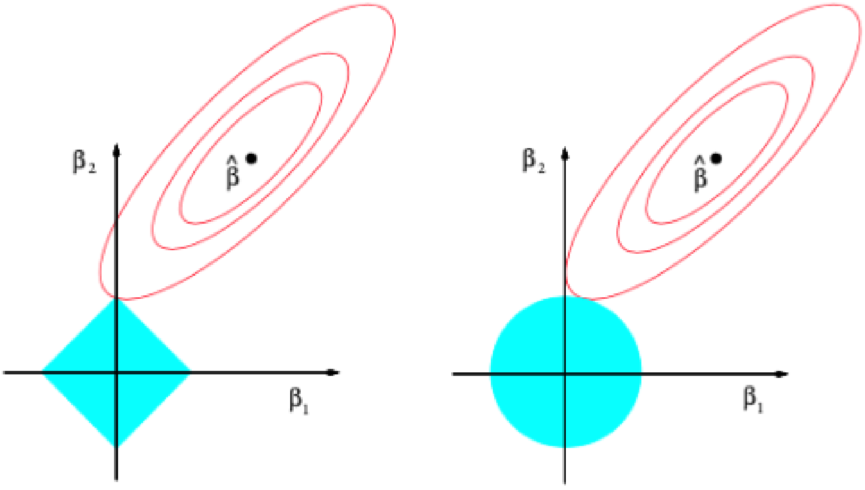


Figure 2.2: *Constraint regions for LASSO (left) and Ridge regression (right) (O'Brien 2016)*

A comparison of LASSO and Ridge is shown in Figure 2.2. The blue diamond and circle are the constraint regions for $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$ respectively, when $p = 2$. The residual sum of squares (RSS) is depicted by the red contours and correspond to the first term in Eq 2.8. $\hat{\beta}$ is the OLS solution. Both algorithms attempt to minimise the RSS by setting β_1 and β_2 equal to where the contours touch the constraint regions (Friedman, Hastie, and Robert Tibshirani 2001). The diagram gives further intuition as to why the LASSO is more likely, due to the constraint regions having kinks at 0, to set some variables to zero, while Ridge shrinks all coefficients towards 0.

2.3.2 Sparsity

The ℓ^1-penalty of the LASSO provides a way to force sparsity into our model, however the performance of the estimator relies on the underlying true signal being sparse. If the underlying signal is not reasonably sparse, then the LASSO will do a poor job of recovering it (O'Brien 2016). Exact sparsity is a strong assumption in many circumstances, and thankfully, these estimation methods often work under the lesser assumption of approximate sparsity. Belloni, Chernozhukov, and Hansen (2011) formally define an approximately sparse model (ASM) as follows.

ASM

$$y_i = f(z_i) + \varepsilon_i = x_i'\beta_0 + r_i + \varepsilon_i, \;\; \varepsilon \sim N(0, \sigma^2), \;\; i = 1, ..., n \tag{2.10}$$

where z_i are elementary predictors, which through function $f(\cdot)$ determine the dependent variable y_i and where ε is our error term. $x_i = P(z_i)$, where $P(\cdot)$ performs a number of transformations on the elementary predictors. x_i is able to be a lot larger than z_i due to the large number of potential transformations such as interactions and polynomials. Approximate sparsity requires that $f(\cdot)$ can be estimated with a small number of non-zero coefficients (s) where $s \ll n$.

There is no universally agreed-upon definition on sparsity, although many different measures have been proposed (Hurley and S. Rickard 2009)[2]. The authors Dalton (1920), Scott Rickard and Fallon (2004), and Arnold (2012) suggest 6 properties that a good measure of sparsity should capture. These are:

[2] See Figure A.1 for the complete list

- Sparsity is scale-invariant. Multiplying signals by a constant factor does not alter the effective distribution.
- Increasing the strength of a weaker signal while decreasing the strength of a stronger signal by the same amount decreases sparsity.
- Adding a constant to each signal will decrease sparsity. This decreases the relative differences between signals.
- Sparsity is invariant under cloning. If we combine two twin populations with the same signal distributions, then one of the populations' sparsity is equal to the combination's sparsity.
- As one signal becomes infinitely strong, the population becomes sparser.
- Increasing the number of signals that have 0 information increases the sparseness.

ℓ^0 is often used in theoretical settings as a measure of sparsity but is rarely useful in statistical analysis as the gradient contains no information and is unusable in the presence of noise (Karvanen and Cichocki 2003). ℓ^0_ε is often used when there is noise in the signal, where ε is a minimum value which a signal needs to be greater than to be recognised as different to 0 (ibid.). Scott Rickard and Fallon (2004) show that only the Gini index satisfies all six properties that a good sparsity measure should have out of those outlined in Figure A.1. The Gini index is derived from the Lorenz curve, first defined by Lorenz (1905). Given a data-set $x = \{x_1, ..., x_n\}$, it is calculated by firstly ordering the data in order of absolute magnitudes $|x_1| \leq ... \leq |x_n|$. The Lorenz curve is then given by:

$$L\left(\frac{i}{n}\right) = \sum_{j=1}^{n} \frac{|x_i|}{\sum_{k=1}^{n} |x_k|}, \quad \forall i \in \{0, ..., n\} \tag{2.11}$$

The area underneath the Lorenz curve is:

$$A(x) = \frac{1}{2n} \sum_{i=1}^{n} \left(L\left(\frac{i-1}{n}\right) + L\left(\frac{i}{n}\right) \right) \tag{2.12}$$

And the Gini index is given by:

$$G(x) = 1 - 2A(x) \tag{2.13}$$

Finally, we will review how the LASSO can be used in approximately sparse settings to aid in causal estimation.

2.3.3 Post Double Selection

Often researchers are not interested in the causal relationship between the outcome and all potential variables, but instead just the effect of changing one. This section will introduce the concept of the post-double-selection (PDS) method proposed by Belloni, Chernozhukov, and Hansen (2014a) and Belloni, Chernozhukov, and Hansen (2014b).

As discussed earlier, without an RCT, we need to assume unconfounded assignment of the treatment variable to infer causality. However, this is dependent on us choosing the correct controls to make treatment uncorrelated with the outcome ($W_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i$). If too many are included, the model will be over-fitted, too few, and it will suffer from bias, and it can be hard to know beforehand the extent that interactions and polynomials play a part in the relationship (Ahrens, Aitken, and Schaffer 2021). Traditionally researchers have relied on intuition and economic theory to select controls, but this can be especially hard when the number of potential controls is large.

The PDS estimator provides a data-driven method of model selection to estimate ATE. Consider the approach of allowing the LASSO to select controls in the linear model:

$$y_i = \alpha d_i + \beta_y x_i + r_{yi} + \zeta_i,$$

where we are trying to estimate the coefficient α on the treatment variable d_i, and where $E(\zeta_i | d_i, x_i, r_{yi}) = 0$, x_i is a p-dimensional vector of potential controls and where $p \gg n$, and r_{yi} is the approximation error. Allowing the LASSO to select the controls from the equations above, and then regressing y_i on the selected covariates while forcing α into the model, would result in substantial omitted variable bias if variables correlated with d_i were excluded from the regression model (Belloni, Chernozhukov, and Hansen 2014a). This is because LASSO only targets prediction of the outcome variable. Any variable that is correlated with the treatment d_i but does not have substantial predictive power of dependent variable will be omitted. The answer to this problem is to introduce a reduced form equation, modeling the relationship between the treatment and controls,

$$d_i = \beta_d x_i + r_{di} + v_i,$$

where $E(v_i|x_i, r_{yi}) = 0$ (Belloni, Chernozhukov, and Hansen 2014b).

As we are only interested in the causal effect of one variable, the LASSO can be left to do model selection over the ‘nuisance’ parts of the model, the first stages and reduced forms, instead of directly on the structural model (Belloni, Chernozhukov, and Hansen 2014a).

The steps of double selection are as follows:

1. Regress y_i on all of the potential covariates excluding the treatment d

   $$y_i = \beta_1 x_{1i} + ... + \beta_p x_{pi} + \varepsilon_i,$$

   and store the variables with non-zero coefficients in vector x_{yi}.

2. Regress d_i on all the potential covariates

   $$d_i = \beta_1 x_{1i} + ... + \beta_p x_{pi} + \varepsilon_i,$$

   and store the variables with non-zero coefficients in vector x_{di}.

3. The final step to estimating α is to regress y_i on d_i and the union of the variables selected from the two models x_{yi} and x_{di}.

Doing this guards against the naive approach’s result as any excluded variables will only be mildly associated with y_i and d_i. The PDS method approximately finds controls that can orthogonalise y_i and d_i in respect to the error term ε_i (ibid.). Belloni, Chernozhukov, and Hansen (2014b) provide the formal conditions for which the double selection approach provides valid causal inference.

Chapter 3

Research Methodology

3.1 Simulation Design

A Monte Carlo simulation is an algorithm that generates data based on a predefined data generating process (DGP). The researcher has complete control over the DGP and can define the parameters and characteristics of the populations from which the data is sampled as well as the functional form of the model. Monte Carlo simulations can be used for many different purposes; of most interest to this study is their ability to allow researchers to analyse an estimator's properties.

First, the structural and stochastic parts of the DGP are defined. Then at each replicate simulation, the stochastic parts are redrawn as random samples from a given distribution. The structural elements are consistent between replications. The advantage of repeating the experiment many times over a single point estimate is that we can not only observe the outcomes from the experiment but how likely a specific outcome is, as well as analyse the probability density of results graphically.

As noted in chapter 1, the purpose of this study is to assess the effectiveness of the post-double-selection estimator and the Post-LASSO's ability to recover the treatment effect under various levels of sparsity. In the following chapter, I outline both the structural and stochastic parts of the DGPs. I then outline the procedure used to assess the PDS and Post-LASSO methods' ability to recover the average treatment effect α. I then cover the summary statistics used and finally the software and languages used for the study.

3.1.1 The Data Generating Process

In this study, I use three distinct DGPs. DGP1&2 have the same functional form but differ in their construction of coefficients. DGP3 introduces a reduced form that models the relationship between the treatment and specific covariates.

DGP1 & DGP2

The first two DGPs follow the linear model:

$$y_i = \alpha d_i + X_i\beta + \varepsilon_i, \tag{3.1}$$

where α is the treatment effect, $\varepsilon_i \sim N(0,1)$, β is a $(1 \times p-1)$ dimensional vector and X_i is a $(p-1)$ dimensional vector of covariates for observation i. p is the number variables including the treatment effect, and n is the sample size.

The covariates X are drawn from the multivariate normal distribution:

$$X_i \sim N(0,\Sigma),$$

where Σ is a $(p \times p)$ covariance matrix that takes the form:

$$\Sigma = \begin{bmatrix} 1 & 0.5 & \cdots & 0 \\ 0.5 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{bmatrix} \tag{3.2}$$

where $\Sigma_{l,j} = Cov(X_{il}, X_{ij})$ subject to:

$$\Sigma_{l,j} = \begin{cases} 1, & \text{if } l = j \\ 0.5, & \text{if } l \ \& \ j \leq \kappa \\ 0, & \text{otherwise} \end{cases}$$

where κ is the number of correlated variables, such that $0 \leq \kappa \leq p$.

As noted above, the difference between DGP1 and DGP2 is the structure of the β coefficients. DGP1 has a continuously decreasing β. The structure is similar to the simulation design used by Belloni, Chen, et al. (2012) and is based on the following formula:

$$\beta = \{\varphi^1, \varphi^2, \cdots, \varphi^{p-1}\}^T, \quad 0 < \varphi < 1 \tag{3.3}$$

While not exactly sparse, the β values quickly decay towards zero. The rate at which this happens is dictated by the parameter φ. The overall measure of the coefficients' sparsity is therefore set by the value of φ and the number of variables p.

DGP2 differs from DGP1 in that the magnitude of the coefficients fall into three categories. The first are large non-zero coefficients that are of equal magnitude to the treatment effect. The number of coefficients equal to α is dictated by the parameter s. The second group of parameters have small non-zero coefficients that are $\frac{1}{10}$ of the treatment size, the number of which is dictated by the parameter z. The remaining variables have a coefficient equal to zero. For example, if $s = 2$, $z = 3$ and $\alpha = 1$ then:

$$\beta = (1, 1, 0.1, 0.1, 0.1, 0, 0, \cdots, 0)^T$$

Again, we are able to vary the sparsity of the model by chaining the values of s, z and p. DGP2 has a sparse representation in that some variables have coefficients equal to zero.

DGP3

DGP3 has a different structural form to DGP1&2. It follows the linear model:

$$y_i = \alpha d_i + X_i^c \beta^c + X_i^y \beta^y + \varepsilon_i$$

$$d_i = X_i^c \beta^c + X_i^e \beta^e + \zeta_i,$$

where $(\varepsilon_i, \zeta_i) \sim N(0, 1)$, X_i^c is a vector of K_c variables that are 'common' to both y_i and d_i. X_i^y is a vector of variables that only affect y_i of length K_y and X_i^e is a vector of variables that affect d_i

of length K_e. X_i^e is a source of exogenous variation in the variable d_i. If $p > K_y + K_e + K_c$ then the remaining variables will have coefficients equal to zero. The covariates are sampled from a normal distribution $((X_i^c, X_i^y, X_i^e) \sim N(0,1))$. Failing to control for the effect of X_i^c would result in omitted variable bias.

All βs follow the decreasing pattern:

$$\beta^c = \{\varphi_c^1, \varphi_c^2, \cdots, \varphi_c^{K_c}\}^T, \quad 0 < \varphi_c < 1,$$

$$\beta^y = \{\varphi_y^1, \varphi_y^2, \cdots, \varphi_y^{K_y}\}^T, \quad 0 < \varphi_y < 1,$$

$$\beta^e = \{\varphi_e^1, \varphi_e^2, \cdots, \varphi_e^{K_e}\}^T, \quad 0 < \varphi_e < 1,$$

Similarly to DGP1, by varying the values of $\varphi_c,\ \varphi_y, \varphi_e$ & p, we can alter the level of sparsity.

3.2 Procedure

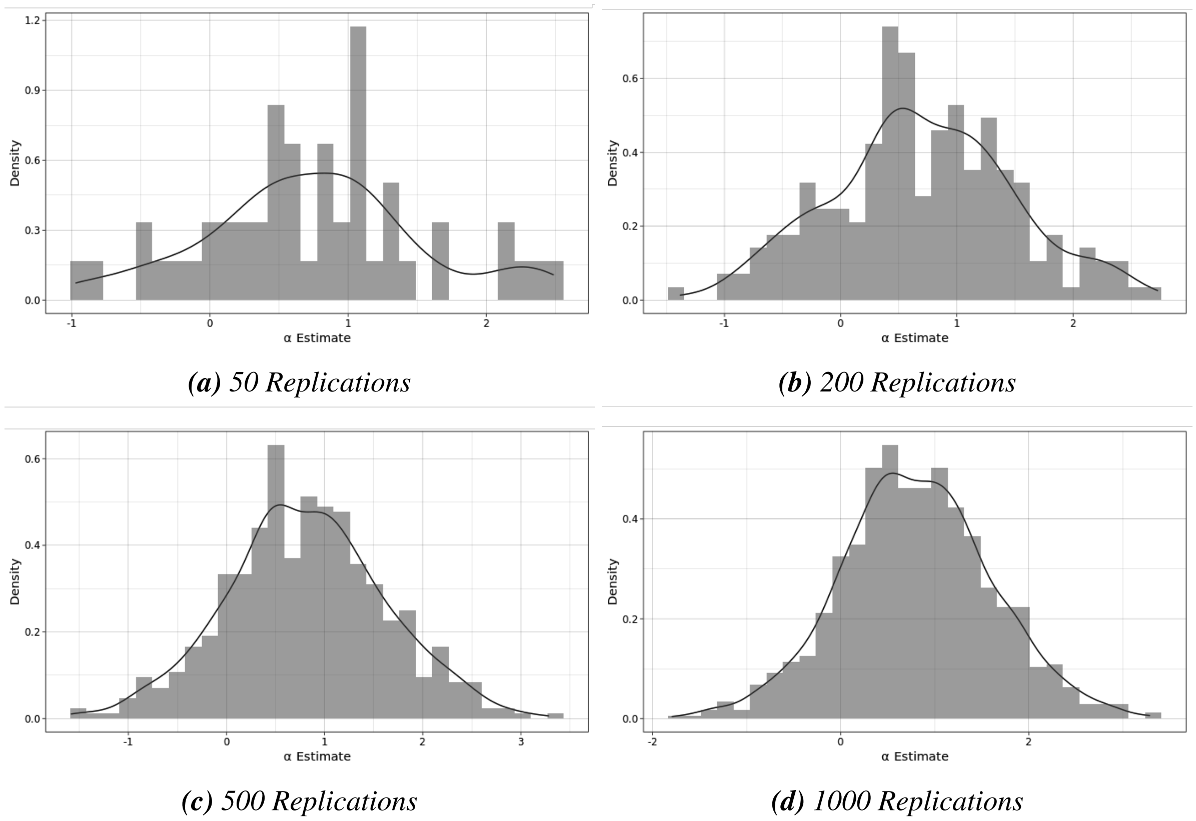
At each replication for DGP1&2, new X_is, including d_is, are drawn from the multivariate normal distribution and ε_i from a normal distribution. For each replication using DGP3, $(d_i, X_i^c, X_i^y, X_i^e, \varepsilon_i$ & $\zeta_i)$ are drawn drawn a normal distribution.

Each experiment is replicated 200 times. This number of simulations was selected when considering the trade-off between precision and computation time. As we can see from table 3.1 and figure 3.1, despite a less smooth density plot, the MSE and Mean Absolute Bias do not change significantly from 200 → 500 → 1000 simulations.

Number of Replicate Simulations	**MSE**	**Mean Absolute Bias**
5	0.511	0.631
50	0.675	0.640
100	0.636	0.646
200	0.698	0.668
500	0.692	0.662
1000	0.685	0.658

Table 3.1: *Comparison of estimate MSE & Mean Absolute Bias with different number of replications*

The selection of the parameter sets for each experiment can be found in Tables B.2 - B.4 in



(a) 50 Replications

(b) 200 Replications

(c) 500 Replications

(d) 1000 Replications

Figure 3.1: *Density Plots of the treatment α estimate using the PDS method with different numbers of replicate simulations*

Appendix B. The values were chosen to provide an approximate spread of sparsity in the true coefficients between experiments, as measured by the Gini coefficient, and to measure the estimators' accuracy at different sample sizes. Under each DGP, I consider two sub-groups of experiments: the first are smaller sample size experiments where $n = (50, 100, 200, 500)$ with the number of non-zero coefficients varying between 10 and 200; the second sub-group have a larger sample size where $n = 1000$ and the number of non-zero coefficients vary between 100 and 1500. The number of correlated variables κ is also varied between experiments from $20 - 100$ in the smaller sample sizes and $20 - 500$ in the larger sample size experiments. For all three DGPs, the treatment effect α is equal to 1.

After generating each data set, the PDS and Post-LASSO estimators are used to estimate the treatment effect, and the Gini coefficient is calculated. These results are then stored, and a new data set is generated. A total of 136 simulation experiments were conducted, with 27,200 simulated data sets.

3.2.1 Post-double-selection Method

As discussed in Section 2.3.3, the steps to implementing PDS is as follows:

1. Use LASSO to select controls that best predict y_i.

2. Use LASSO to select controls that best predict d_i.

3. Regress y_i on d_i and the union of the selected controls with OLS.

Stages 1 and 2 implement the rigorous (or plug-in) LASSO using the theory driven regularisation parameter λ_{PDS} from Belloni and Chernozhukov (2011):

$$\lambda_{PDS} = \sqrt{n}\hat{\sigma}\Phi^{-1}(1 - \frac{\gamma}{2p}), \tag{3.4}$$

where Φ is the cumulative standard normal distribution, $\hat{\sigma}$ is a preliminary estimate of $\sigma = \sqrt{E(\varepsilon^2)}$ found through an iterative algorithm defined in Belloni and Chernozhukov (ibid.), and $\gamma = 0.1$ is the probability level of mistakenly removing X's. This penalty level is chosen so that regularisation is large enough to dominate estimation noise (Belloni, Chen, et al. 2012).

3.2.2 The Post-LASSO Method

The steps to implementing Post-LASSO are as follows:

1. Use LASSO to select controls that best predict y_i.

2. Regress y_i on d_i and the chosen controls with OLS.

The Post-LASSO uses a different method for selecting λ. The regularisation parameter was selected through K-fold cross validation where $K = 10$ and is given by the formula:

$$\lambda_{post} = \arg\min_{\lambda} \frac{1}{K} \sum_{i=1}^{K} MSE_i(\lambda), \tag{3.5}$$

where $MSE(\lambda)$ is the MSE for a given λ value. An example of this process can be seen in the CV plot in Figure B.1.

If both the Post-LASSO and PDS method are to do model selection perfectly, their results will be identical to OLS using only appropriate controls.

3.2.3 Summary Statistics

Summary statistics are created from each experiment. We are exploring the bias and efficiency of the two estimators in different situations. The four reported statistics are the Root Mean Squared Error (RMSE), Mean $\hat{\alpha}$ estimate ($E(\hat{\alpha})$), the Mean Absolute Bias (Bias) and the Mean Variance (Variance).

- The formula for the RMSE is: $RMSE(\hat{\alpha}) = \sqrt{E\left[(\hat{\alpha} - \alpha)^2\right]}$
- The formula for the mean absolute bias is: $Bias(\hat{\alpha}) = E(|\hat{\alpha} - \alpha|)$
- The MSE is a function of the efficiency and bias of the estimator. The variance can therefore be calculated by: $Variance(\hat{\alpha}) = MSE(\hat{\alpha}) - [Bias(\hat{\alpha})]^2$ (Carsey and Harden 2013)

The coefficient sparsity is a measure of the underlying signals' strengths in the true model and is measured by the Gini coefficient. It is calculated from equations 2.11-2.13.

As well as the summary statistics, the parameter values for each experiment are reported alongside.

3.3 Statistical Software

The simulation experiment was coded in R version 3.6.3 inside Rstudio server version 1.4.1104 (R Core Team 2020). The Monte Carlo simulation was run on Google Cloud Platform - Compute Engine Virtual Machine[1] running Ubuntu 18.04.

Normally distributed random values were generated with the *rnorm()* function and multivariate normally distributed values from the *mvrnorm()* function from the *MASS* package, generated with the Mersenne Twister algorithm (Matsumoto and Nishimura 1998). The 'seeds' that governs the pseudo-random number generation are set at the beginning of each experiment and can be found in the accompanying R code. Post-double-selection was done using the *rlasso()* function from the package *hdm*. The first stage of the Post-LASSO method used the *glm()* function from the *glmnet* package. The turning parameter λ was selected through K-fold cross-validation where $K = 10$ with the function *cv.glm()* also from the *glmnet* package. The

[1] Machine Type: c2-standard-8 (8 vCPUs, 32 GB memory), Intel Cascade Lake

second-stage OLS regression used the *lm()* function from base R. Other packages were used to analyse and plot the results. The complete list of packages used in the simulation and the subsequent analysis can be found in Table B.1.

Chapter 4

Results and Analysis

The main results from this Monte Carlo experiment are presented in Tables 4.1 - 4.5. Tables 4.1 and 4.2 present the results for experiments that used DGP1; Table 4.3 presents the 'small coefficient' results and Table 4.4 present the 'large coefficient' results for experiments that used DGP2; and Table 4.5 presents the results for experiments that used DGP3. The results that are reported across all experiments are the RMSE, bias, variance and mean $\hat{\alpha}$.

For the remainder of this section, I will report and analyse the simulation study's findings, looking at where both estimators performed best and worst.

Table 4.1: DGP1 *Estimates from 200 replicate simulations of RMSE, bias, variance and mean estimate of average treatment effect, from the Post-double-selection and Post-LASSO regression models.*

					Post-double-selection				Post-LASSO			
n	p	κ	φ	Gini Coef	RMSE	$E(\hat{\alpha})$	Bias	Variance	RMSE	$E(\hat{\alpha})$	Bias	Variance
50	200	20	0.999	0.0333	2.5297	0.5359	1.9877	2.4193	1.6683	0.2004	0.7996	2.1439
50	200	20	0.99	0.3144	1.4021	0.7811	1.0708	0.8191	1.1928	0.2184	0.7816	0.8119
50	200	20	0.98	0.5408	0.9147	0.9546	0.7187	0.3184	1.0771	0.3641	0.6359	0.7558
50	200	20	0.97	0.6762	0.6604	0.9364	0.5077	0.1761	0.8016	0.4119	0.5881	0.2968
50	200	20	0.95	0.8051	0.4821	0.8730	0.3781	0.0895	0.6996	0.4259	0.5741	0.1598
50	200	20	0.9	0.9050	0.3228	0.8985	0.2565	0.0384	0.6361	0.5260	0.4740	0.1800
50	200	100	0.99	0.3144	24.2524	-1.2896	5.0808	562.0281	75.9481	-7.9213	8.9213	5688.5295
50	200	100	0.999	0.0333	17.2032	0.5049	5.7997	261.8982	5.8723	0.1059	0.8941	33.6842
50	200	100	0.98	0.5408	12.7218	1.2055	2.8216	153.2134	2.5994	0.4067	0.5933	6.4047
50	200	100	0.97	0.6762	2.1360	0.3668	1.2900	2.8963	17.8927	1.2305	0.2305	320.0941
50	200	100	0.95	0.8051	1.1659	0.5038	0.7493	0.7869	1.4098	0.4629	0.5371	1.6991
50	200	100	0.9	0.9050	0.4610	0.8240	0.3439	0.0942	0.6779	0.3926	0.6074	0.0907
100	200	20	0.999	0.0333	1.7287	1.0165	1.3786	1.0837	1.2777	0.1564	0.8436	0.9208
100	200	20	0.99	0.3144	0.9295	0.9271	0.7395	0.3171	0.9389	0.1951	0.8049	0.2336
100	200	20	0.98	0.5408	0.5069	0.9998	0.3943	0.1006	0.6956	0.3924	0.6076	0.1146
100	200	20	0.97	0.6762	0.3619	0.9845	0.2768	0.0542	0.6272	0.4404	0.5596	0.0802
100	200	20	0.95	0.8051	0.2326	0.9816	0.1812	0.0208	0.5003	0.5706	0.4294	0.0660
100	200	20	0.9	0.9050	0.1616	0.9796	0.1257	0.0101	0.3431	0.7498	0.2502	0.0551
100	200	100	0.999	0.0333	1.7229	0.9192	1.4038	0.9959	1.5170	0.2187	0.7813	1.6909
100	200	100	0.99	0.3144	0.6902	0.7268	0.5573	0.1657	0.6466	0.6311	0.3689	0.2820
100	200	100	0.98	0.5408	0.4843	0.6637	0.3965	0.0772	0.4265	0.7063	0.2937	0.0956
100	200	100	0.97	0.6762	0.3826	0.6974	0.3206	0.0436	0.3790	0.6950	0.3050	0.0506
100	200	100	0.95	0.8051	0.2756	0.8636	0.2209	0.0271	0.4247	0.6406	0.3594	0.0512
100	200	100	0.9	0.9050	0.1960	0.9565	0.1580	0.0134	0.3974	0.6812	0.3188	0.0563

Table 4.2: DGP1 Continued *Estimates from 200 replicate simulations of RMSE, bias, variance and mean estimate of average treatment effect, from the Post-double-selection and Post-LASSO regression models.*

					Post-double-selection				Post-LASSO			
n	p	κ	φ	Gini Coef	RMSE	$E(\hat{\alpha})$	Bias	Variance	RMSE	$E(\hat{\alpha})$	Bias	Variance
500	200	20	0.999	0.0333	0.6952	0.9410	0.5576	0.1719	0.0762	1.0019	0.0019	0.0058
500	200	20	0.99	0.3144	0.3159	0.9228	0.2561	0.0338	0.0763	0.9994	0.0006	0.0058
500	200	20	0.98	0.5408	0.1913	0.9115	0.1588	0.0113	0.1024	0.9349	0.0651	0.0062
500	200	20	0.97	0.6762	0.1366	0.9371	0.1164	0.0051	0.1328	0.8913	0.1087	0.0058
500	200	20	0.95	0.8051	0.0998	0.9538	0.0808	0.0034	0.1438	0.8764	0.1236	0.0054
500	200	20	0.9	0.9050	0.0713	0.9890	0.0566	0.0019	0.1123	0.9197	0.0803	0.0062
500	200	100	0.999	0.0333	0.6565	1.0350	0.5322	0.1474	0.0822	0.9971	0.0029	0.0068
500	200	100	0.99	0.3144	0.2001	0.9910	0.1619	0.0138	0.1527	0.8864	0.1136	0.0104
500	200	100	0.98	0.5408	0.0795	1.0103	0.0633	0.0022	0.0998	0.9248	0.0752	0.0043
500	200	100	0.97	0.6762	0.0653	1.0060	0.0519	0.0015	0.0850	0.9543	0.0457	0.0051
500	200	100	0.95	0.8051	0.0633	0.9900	0.0525	0.0012	0.0888	0.9435	0.0565	0.0047
500	200	100	0.9	0.9050	0.0618	0.9976	0.0511	0.0012	0.0904	0.9372	0.0628	0.0042
1000	2000	20	0.999	0.3132	0.9054	0.9131	0.7246	0.2946	0.8944	0.2019	0.7981	0.1630
1000	2000	20	0.99	0.9005	0.2143	0.8936	0.1762	0.0148	0.3378	0.6704	0.3296	0.0055
1000	2000	20	0.98	0.9505	0.1175	0.9375	0.0950	0.0047	0.2529	0.7528	0.2472	0.0028
1000	2000	20	0.97	0.9672	0.0740	0.9652	0.0600	0.0019	0.2377	0.7662	0.2338	0.0018
1000	2000	20	0.95	0.9805	0.0594	0.9713	0.0479	0.0012	0.2803	0.7246	0.2754	0.0027
1000	2000	20	0.9	0.9905	0.0447	0.9949	0.0362	0.0007	0.1407	0.8836	0.1164	0.0063
1000	2000	500	0.999	0.3132	0.7696	1.0076	0.6085	0.2201	0.7291	0.8972	0.1028	0.5210
1000	2000	500	0.99	0.9005	0.0751	0.9554	0.0607	0.0020	0.0920	0.9333	0.0667	0.0040
1000	2000	500	0.9	0.9905	0.0554	1.0246	0.0452	0.0010	0.1158	0.9112	0.0888	0.0055
1000	2000	500	0.95	0.9805	0.0531	1.0157	0.0424	0.0010	0.0769	0.9416	0.0584	0.0025
1000	2000	500	0.97	0.9672	0.0531	1.0064	0.0422	0.0010	0.0579	0.9711	0.0289	0.0025
1000	2000	500	0.98	0.9505	0.0517	0.9926	0.0430	0.0008	0.0674	0.9638	0.0362	0.0032

Table 4.3: DGP2 Small Coefficients *Estimates from 200 replicate simulations of RMSE, bias, variance and mean estimate of average treatment effect, from the Post-double-selection and Post-LASSO regression models.*

						Post-double-selection				Post-LASSO			
n	p	s	z	κ	Gini Coef	RMSE	$E(\hat{\alpha})$	Bias	Variance	RMSE	$E(\hat{\alpha})$	Bias	Variance
50	200	15	20	20	0.9	0.3314	0.9538	0.2632	0.0406	0.61456016	0.4985	0.5015	0.1262
50	200	15	20	100	0.9	1.8936	0.6946	0.7644	3.0014	1.00280722	0.3643	0.6357	0.6016
50	200	15	100	20	0.6969	0.4774	0.9494	0.3476	0.1071	0.70600057	0.4383	0.5617	0.1829
50	200	15	100	100	0.6969	0.8408	0.4915	0.7031	0.2126	1.52396859	0.5663	0.4337	2.1344
50	200	15	150	20	0.5184	0.5203	0.9904	0.4001	0.1106	0.66364312	0.4737	0.5263	0.1635
50	200	15	150	100	0.5184	2.0171	0.7954	1.0761	2.9107	12.2445686	-0.8159	1.8159	146.6320
50	200	15	180	20	0.4012	0.4932	0.9321	0.3992	0.0839	0.70688143	0.4359	0.5641	0.1815
50	200	15	180	100	0.4012	1.7611	0.4012	0.8922	2.3054	1.50674435	0.5343	0.4657	2.0534
100	200	15	20	20	0.9	0.1507	1.0144	0.1195	0.0084	0.34764571	0.7459	0.2541	0.0563
100	200	15	20	100	0.9	0.1501	0.9827	0.1181	0.0086	0.33850277	0.7337	0.2663	0.0437
100	200	15	100	20	0.6969	0.2395	1.0043	0.1919	0.0205	0.45814149	0.6472	0.3528	0.0854
100	200	15	100	100	0.6969	0.2896	0.7980	0.2380	0.0272	0.40357663	0.6487	0.3513	0.0395
100	200	15	150	20	0.5184	0.2569	1.0015	0.2060	0.0236	0.54442515	0.5344	0.4656	0.0797
100	200	15	150	100	0.5184	0.2938	0.8619	0.2377	0.0298	0.49410608	0.5779	0.4221	0.0660
100	200	15	180	20	0.4012	0.2563	0.9932	0.1960	0.0273	0.50313828	0.5962	0.4038	0.0901
100	200	15	180	100	0.4012	0.2638	0.9005	0.2134	0.0240	0.50687052	0.5769	0.4231	0.0779
500	200	15	20	20	0.9	0.0717	0.9862	0.0564	0.0020	0.12903258	0.8964	0.1036	0.0059
500	200	15	20	100	0.9	0.0559	1.0005	0.0461	0.0010	0.08213064	0.9457	0.0543	0.0038
500	200	15	100	20	0.6969	0.1033	0.9763	0.0822	0.0039	0.15473772	0.8628	0.1372	0.0051
500	200	15	100	100	0.6969	0.0783	0.9828	0.0642	0.0020	0.11991716	0.9078	0.0922	0.0059
500	200	15	150	20	0.5184	0.1131	0.9921	0.0905	0.0046	0.13642447	0.8890	0.1110	0.0063
500	200	15	150	100	0.5184	0.0894	0.9887	0.0718	0.0028	0.10999751	0.9147	0.0853	0.0048
500	200	15	180	20	0.4012	0.1014	1.0205	0.0762	0.0045	0.10433381	0.9251	0.0749	0.0053
500	200	15	180	100	0.4012	0.0957	0.9990	0.0758	0.0034	0.10243331	0.9288	0.0712	0.0054
1000	2000	100	1000	20	0.6756	0.2863	0.8442	0.2337	0.0274	0.55918966	0.4496	0.5504	0.0097
1000	2000	100	1000	100	0.6756	0.1423	0.9928	0.1131	0.0075	0.1420998	0.9923	0.0077	0.0201
1000	2000	100	1899	20	0.2967	0.3332	0.8399	0.2644	0.0411	0.65088565	0.3617	0.6383	0.0163
1000	2000	100	1899	100	0.2967	0.1991	1.0034	0.1619	0.0134	0.19706855	0.9842	0.0158	0.0386

Table 4.4: DGP2 Large Coefficients *Estimates from 200 replicate simulations of RMSE, bias, variance and mean estimate of average treatment effect, from the Post-double-selection and Post-LASSO regression models.*

						Post-double-selection				Post-LASSO			
n	p	s	z	κ	Gini	RMSE	$E(\hat{\alpha})$	Bias	Variance	RMSE	$E(\hat{\alpha})$	Bias	Variance
50	200	20	10	20	0.8880	0.412838112	0.9117	0.3098	0.0744	0.65739998	0.4359	0.5641	0.1140
50	200	20	10	100	0.8880	7.823807243	1.3169	1.1049	59.6991	1.30594249	0.3020	0.6980	1.2183
50	200	100	10	20	0.4896	2.184836718	0.7102	1.6509	2.0480	1.34643686	0.2321	0.7679	1.2232
50	200	100	10	100	0.4896	75.17739246	0.2634	13.9640	5358.8532	112.333519	12.0253	11.0253	12497.2620
50	200	150	10	20	0.2397	2.500051333	1.2057	1.9288	2.4929	1.63930866	0.3382	0.6618	2.2494
50	200	150	10	100	0.2397	12.7460054	0.7240	4.8981	138.1821	6.6551041	0.1343	0.8657	43.5410
50	200	180	10	20	0.0898	2.646597285	0.9681	2.1069	2.5504	3.54479598	0.5415	0.4585	12.3553
50	200	180	10	100	0.0898	38.66386854	0.4764	8.2765	1408.9859	3.84516721	0.0288	0.9712	13.8421
100	200	20	10	20	0.8880	0.194927309	0.9844	0.1447	0.0167	0.35821144	0.7104	0.2896	0.0444
100	200	20	10	100	0.8880	0.170558931	0.9915	0.1376	0.0102	0.3244568	0.7485	0.2515	0.0420
100	200	100	10	20	0.4896	1.180258096	0.9540	0.9193	0.5454	1.06126734	0.3009	0.6991	0.6375
100	200	100	10	100	0.4896	0.893494291	0.5284	0.7071	0.2983	0.79580276	0.5577	0.4423	0.4376
100	200	150	10	20	0.2397	1.457155156	1.1219	1.1795	0.7312	1.30959177	0.1249	0.8751	0.9493
100	200	150	10	100	0.2397	1.456493399	0.9135	1.2111	0.6546	1.05162351	0.4949	0.5051	0.8508
100	200	180	10	20	0.0898	1.982413436	0.9279	1.5252	1.6008	1.46636504	0.1081	0.8919	1.3548
100	200	180	10	100	0.0898	1.817590334	1.1979	1.4546	1.1683	1.30851077	0.3688	0.6312	1.3138
500	200	20	10	20	0.8880	0.072104461	0.9881	0.0575	0.0019	0.1219932	0.9045	0.0955	0.0058
500	200	20	10	100	0.8880	0.061922289	1.0022	0.0499	0.0013	0.08027892	0.9470	0.0530	0.0036
500	200	100	10	20	0.4896	0.481122486	0.8203	0.3798	0.0838	0.09641897	0.9348	0.0652	0.0050
500	200	100	10	100	0.4896	0.073692459	0.9998	0.0592	0.0019	0.10883281	0.9981	0.0019	0.0118
500	200	150	10	20	0.2397	0.682589247	0.8145	0.5347	0.1792	0.08567755	0.9756	0.0244	0.0067
500	200	150	10	100	0.2397	0.48893834	1.0104	0.3967	0.0815	0.07163296	0.9921	0.0079	0.0051
500	200	180	10	20	0.0898	0.667327785	1.0190	0.5336	0.1588	0.07488315	1.0005	0.0005	0.0056
500	200	180	10	100	0.0898	0.634286444	1.0654	0.5154	0.1366	0.08124597	0.9972	0.0028	0.0066
1000	2000	1000	100	20	0.4941	1.352380415	0.9534	1.1177	0.5788	0.96632198	0.2029	0.7971	0.2984
1000	2000	1000	100	500	0.4941	1.342270399	1.3389	1.0264	0.7430	1.22708065	0.8038	0.1962	1.4672
1000	2000	1800	100	20	0.0943	1.781047297	0.8126	1.4529	1.0558	1.32395092	0.0069	0.9931	0.7666
1000	2000	1800	100	500	0.0943	1.962824258	1.6858	1.5718	1.3806	1.69073735	0.2383	0.7617	2.2627

Table 4.5: DGP3 *Estimates from 200 replicate simulations of RMSE, bias, variance and mean estimate of average treatment effect, from the Post-double-selection and Post-LASSO regression models.*

											Post-double-selection				Post-LASSO			
n	z	K_c	K_y	K_e	φ_c	φ_y	φ_e	Gini β^c	Gini β^y	Gini β^e	RMSE	$E(\hat{\alpha})$	Bias	Variance	RMSE	$E(\hat{\alpha})$	Bias	Variance
200	400	100	20	20	0.99	0.9	0.9	0.1647	0.3268	0.3268	0.8401	1.8387	0.8387	0.1367	0.6309	0.6361	0.3639	0.4985
200	400	100	20	20	0.9	0.9	0.9	0.8101	0.3268	0.3268	0.2307	1.2011	0.2046	0.1888	0.6486	0.3819	0.6181	0.2666
200	400	50	20	20	0.99	0.9	0.9	0.0834	0.3268	0.3268	0.7624	1.7601	0.7601	0.1846	0.7943	0.2372	0.7628	0.2124
200	400	50	20	20	0.9	0.9	0.9	0.6304	0.3268	0.3268	0.2212	1.1842	0.1902	0.1850	0.6292	0.4011	0.5989	0.2705
200	400	20	20	20	0.99	0.9	0.9	0.0334	0.3268	0.3268	0.3119	1.2480	0.2703	0.2388	0.6822	0.3389	0.6611	0.2451
200	400	20	20	20	0.9	0.9	0.9	0.3268	0.3268	0.3268	0.2163	1.1785	0.1839	0.1825	0.6600	0.3636	0.6364	0.2550
200	400	20	100	20	0.9	0.99	0.9	0.3268	0.1647	0.3268	0.3226	1.1545	0.2571	0.2565	1.0569	0.2340	0.7660	0.4702
200	400	20	100	20	0.9	0.9	0.9	0.3268	0.8101	0.3268	0.1964	1.1585	0.1659	0.1689	0.6496	0.3775	0.6225	0.2621
200	400	20	50	20	0.9	0.99	0.9	0.3268	0.0834	0.3268	0.2872	1.1469	0.2331	0.2329	0.8024	0.2294	0.7706	0.2086
200	400	20	50	20	0.9	0.9	0.9	0.3268	0.6304	0.3268	0.2116	1.1714	0.1823	0.1783	0.6361	0.3933	0.6067	0.2680
200	400	20	20	20	0.9	0.99	0.9	0.3268	0.0334	0.3268	0.2332	1.1534	0.1923	0.1962	0.6729	0.3585	0.6415	0.2615
200	400	20	20	100	0.9	0.9	0.99	0.3268	0.3268	0.1647	0.0581	1.0382	0.0463	0.0559	0.6777	0.4463	0.5537	0.3711
200	400	20	20	100	0.9	0.9	0.9	0.3268	0.3268	0.8101	0.2003	1.1583	0.1697	0.1715	0.6512	0.3716	0.6284	0.2564
200	400	20	20	50	0.9	0.9	0.99	0.3268	0.3268	0.0834	0.0741	1.0498	0.0624	0.0702	0.7471	0.2864	0.7136	0.2380
200	400	20	20	50	0.9	0.9	0.9	0.3268	0.3268	0.6304	0.1990	1.1598	0.1698	0.1702	0.6357	0.3924	0.6076	0.2665
200	400	20	20	20	0.9	0.9	0.99	0.3268	0.3268	0.0334	0.2113	1.1693	0.1749	0.1807	0.6751	0.3519	0.6481	0.2550
500	400	100	20	20	0.99	0.9	0.9	0.1647	0.3268	0.3268	0.8116	1.8102	0.8102	0.1552	0.2130	0.7994	0.2006	0.1728
500	400	100	20	20	0.9	0.9	0.9	0.8101	0.3268	0.3268	0.1360	1.1153	0.1183	0.1220	0.2196	0.7918	0.2082	0.1762
500	400	50	20	20	0.99	0.9	0.9	0.0834	0.3268	0.3268	0.5023	1.4722	0.4751	0.2765	0.2249	0.7865	0.2135	0.1793
500	400	50	20	20	0.9	0.9	0.9	0.6304	0.3268	0.3268	0.1377	1.1163	0.1196	0.1234	0.2211	0.7906	0.2094	0.1773
500	400	20	20	20	0.99	0.9	0.9	0.0334	0.3268	0.3268	0.0763	0.9767	0.0606	0.0726	0.2174	0.7917	0.2083	0.1740
500	400	20	20	20	0.9	0.9	0.9	0.3268	0.3268	0.3268	0.1030	1.0807	0.0860	0.0956	0.2100	0.7999	0.2001	0.1700
500	400	20	100	20	0.9	0.99	0.9	0.3268	0.1647	0.3268	0.2096	1.0038	0.1610	0.1837	0.2098	0.8029	0.1971	0.1709
500	400	20	100	20	0.9	0.9	0.9	0.3268	0.8101	0.3268	0.0991	1.0737	0.0814	0.0925	0.2262	0.7837	0.2163	0.1794
500	400	20	50	20	0.9	0.99	0.9	0.3268	0.0834	0.3268	0.1529	1.0336	0.1203	0.1384	0.2220	0.7901	0.2099	0.1779
500	400	20	50	20	0.9	0.9	0.9	0.3268	0.6304	0.3268	0.1110	1.0826	0.0901	0.1029	0.2268	0.7846	0.2154	0.1804
500	400	20	20	20	0.9	0.99	0.9	0.3268	0.0334	0.3268	0.1371	1.1182	0.1216	0.1223	0.2137	0.7956	0.2044	0.1719
500	400	20	20	20	0.9	0.9	0.99	0.3268	0.3268	0.0334	0.1842	1.1533	0.1596	0.1587	0.2152	0.7944	0.2056	0.1729
500	400	20	20	50	0.9	0.9	0.9	0.3268	0.3268	0.6304	0.0922	1.0664	0.0754	0.0865	0.2225	0.7876	0.2124	0.1775
500	400	20	20	50	0.9	0.9	0.99	0.3268	0.3268	0.0834	0.1244	1.0962	0.0988	0.1146	0.2226	0.7888	0.2112	0.1780
500	400	20	20	100	0.9	0.9	0.9	0.3268	0.3268	0.8101	0.0984	1.0720	0.0809	0.0919	0.2227	0.7862	0.2138	0.1770
500	400	20	20	100	0.9	0.9	0.99	0.3268	0.3268	0.1647	0.0342	1.0220	0.0275	0.0334	0.2114	0.8012	0.1988	0.1718

4.1 DGP1

The results for DGP1 are reported in Tables 4.1 & 4.2. The Gini coefficients ranged from 0.033, being the densest, to 0.905, being the most sparse. The observed trend of the results across the experiments is captured well in Figure C.2 which plots the RMSE, bias and variance at different sparsity parameter (φ) levels when $n = 100, z = 200$ & $\kappa = 20$. As we can see, both the Post-LASSO and the PDS performed best when $\varphi = 0.9$ and performed worse as φ increases and the Gini coefficient decreases. Both estimators' performance appeared to follow an exponential trend with the greatest increases in RMSE, bias and variance below a Gini coefficient of 0.541. Interestingly, at the lowest levels of sparsity, the Post-LASSO performs better than the PDS method, which was a trend that seen across many of the experiments. Each experiment's bias was higher than the variance but increased more at the highest level of φ.

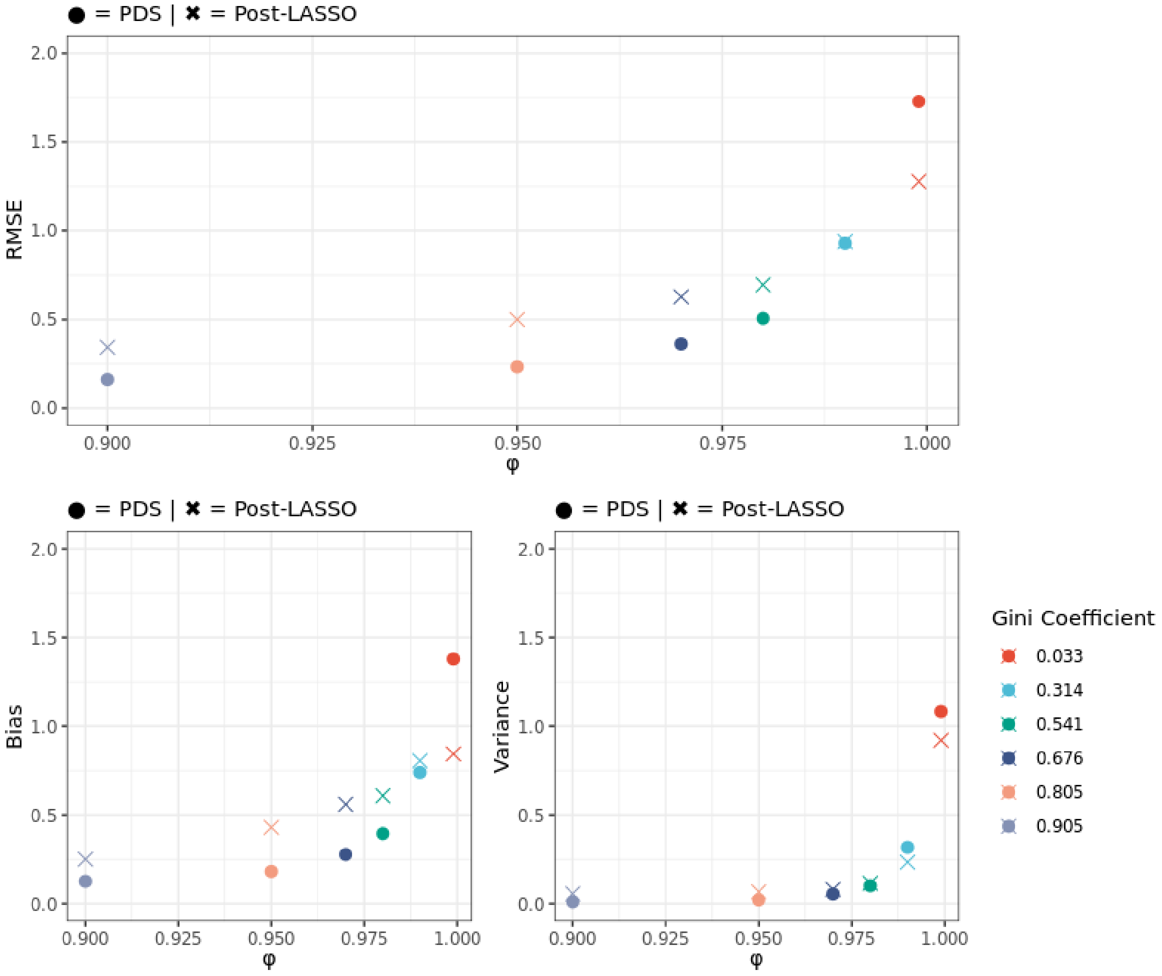


Figure 4.1: DGP1 RMSE, Bias and Variance Plotted against the Gini coefficient where $n = 100, z = 200$ & $\kappa = 20$

The least sparse models were where $\varphi = 0.999$ with a corresponding Gini coefficient of 0.0333 when $n = (20, 100, 500)$ and 0.212 when $n = 1000$. Despite the coefficients' size decaying, these experiments did not have a sparse representation as the 1000th coefficient would still have had a value of 0.37. The performance of both estimators was poor across all the parameter sets apart from where $n = 500$. However, in this case, the number of parameters is no longer greater than n, and both estimators are less constrained in selecting variables. Due to differences in the calculation of the penalty term λ between the two methods, the Post-LASSO likely selected more variables as controls and therefore had better estimates. This is seen in Figure 4.2 with the Post-LASSO estimates being closely centred around 1 (true α) while the PDS estimates are distributed around 1 with substantial variance. The estimators performed worse with fewer observations and performed worse when $n = 50$. In the larger samples size experiments where $n = 200$ & $p = 2000$, the estimators performed better most likely as the larger number of variables allowed the β coefficients to decay to a point closer to zero resulting in a higher Gini coefficient.

When $\varphi = 0.99$ the Gini coefficient is 0.314 when $n = (20, 100, 500)$ and 0.9 when $n = 1000$. The PDS continued to perform poorly when $n = 50$ and performed worst at higher levels of correlation ($\kappa = 100$). The RMSE was better at every simulation compared with $\varphi = 0.999$ experiments, with the estimators performing best when $n = (500, 1000)$. The Post-LASSO and PDS both had a better RMSE than the other in four of the eight experiments. Again, the Post-LASSO performed best when $n = 500$, with estimates centered around the true α.

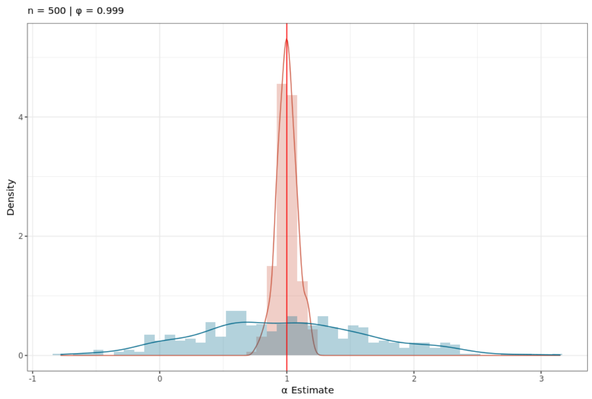


Figure 4.2: *DGP1 Density Plot when $n = 500$ and $\varphi = 0.999$. The distribution of Post-LASSO estimates is red and the distribution of PDS estimates is blue.*

We observe a similar pattern when $\varphi = 0.98$, with a corresponding Gini coefficient of 0.541 when $n = (50, 100, 500)$ and 0.85 when $n = 1000$. Both the PDS and Post-LASSO estimators performed worse when $n = 50$ and $\kappa = 100$ with similar bias and variance levels. However, at this value of φ, we see a considerable performance increases when $n = 100$. This improvement can be seen in Figure 4.1. Furthermore, the number of correlated variables κ had less of an impact. The performance between the two estimators is comparable, but the PDS performed better in five of the eight experiments and had slightly lower bias overall as seen in 4.3.

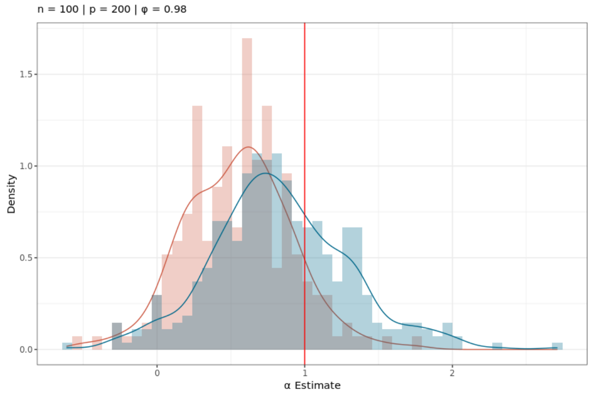


Figure 4.3: *DGP1 Density Plot when $n = 100$, $p = 200$ and $\varphi = 0.98$. The distribution of Post-LASSO estimates is red and the distribution of PDS estimates is blue.*

When $\varphi = 0.97$, the Gini coefficient is 0.68 when $n = (50, 100, 500)$ and 0.96 when $n = 1000$. Again, the PDS estimator performed best in the larger sample size experiment where $n = (500, 1000)$; however, now, the PDS performed well when $n = 50$ and with lower covariance levels ($\kappa = 20$), with a mean estimate of 0.94. The Post-LASSO still struggled at this level with a mean estimate of 0.41. PDS performed better than the Post-LASSO in six of the eight experiments and had lower variance and bias in five. When the number of correlated variables was high, the PDS and Post-LASSO did similarly well in their estimates, but the PDS had lower bias and variance when the number of correlated variables was lower, as seen in Figure 4.4. This could indicate that the CV regularisation parameter used in the Post-LASSO is over-fitting the model.

When $\varphi = 0.95$, the Gini coefficient is 0.91 when $n = (50, 100, 500)$ and 0.98 when $n = 1000$. The levels of sparsity are similar now between the smaller and larger sample size experiments. The PDS performed best when n is larger, with an improvement in the overall accuracy from the previous φ. When $n = 100$, there is no longer a large difference between the experiments with different covariance levels for both Post-LASSO and PDS; however, PDS did a better job at selecting controls resulting in lower bias as seen in Figure 4.5. In all eight experiments, PDS had a lower RMSE and a lower bias in seven. Only when $n = 50$ &$\kappa = 100$ did either estimator not perform well.

Finally, there is a similar pattern when $\varphi = 0.9$, with an overall improvement in the three summary statistics. The results were still worse with smaller sample sizes and more correlated variables; however, PDS did better than the Post-LASSO in terms of RMSE and bias in all

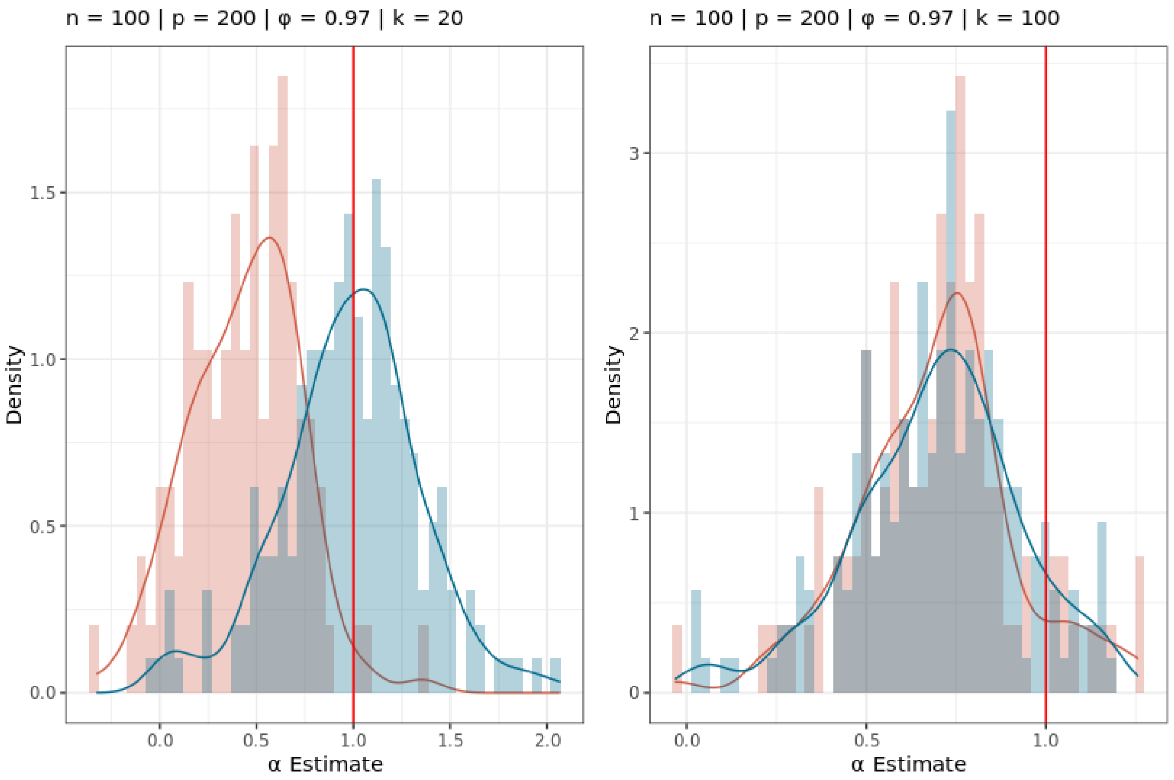


__Figure 4.4:__ DGP1 Density Plots when $n = 100$*,* $p = 200$*,* $\varphi = 0.97$ *and* $\kappa = (20, 100)$*. The distribution of Post-LASSO estimates is red and the distribution of PDS estimates is blue.*

eight experiments, even when $n = 500$.

Overall, PDS had a lower RMSE in 31 of the 48 experiments compared to Post-LASSO, a lower bias in 25 experiments and a lower variance in 32 experiments.

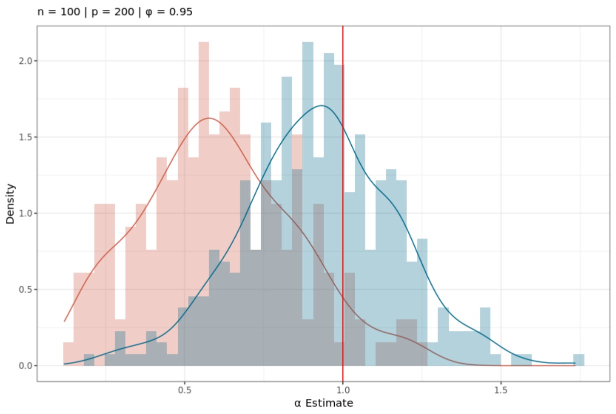


__Figure 4.5:__ DGP1 Density Plot when $n = 100$*,* $p = 200$ *and* $\varphi = 0.95$*. The distribution of Post-LASSO estimates is red and the distribution of PDS estimates is blue.*

4.2 DGP2

With DGP2, the experiments can be classified into two further sub-groups. The first sub-group of experiments varied the total number of small coefficients (z) with the results are reported in Table 4.3, while the second sub-group varied the number of large coefficients (s) with the results reported in

Table 4.4. The purpose of this was to examine further how the coefficients' structure impacted the estimators' ability to recover the treatment effect.

Small Coefficients

Compared with DGP1, there was less of a gradual shift in the performance and more of a clear difference in both the PDS and Post-LASSO estimators' performance. As described in Ch 3, z is the number of small β coefficients that are $\frac{1}{10}$ of the treatment effect where s is the number of coefficients of equal magnitude to the treatment effect α. Firstly, looking at the results in Table 4.3with many small non-zero coefficients. In this sub-group of experiments, the Gini coefficients ranged from 0.9 to 0.4.

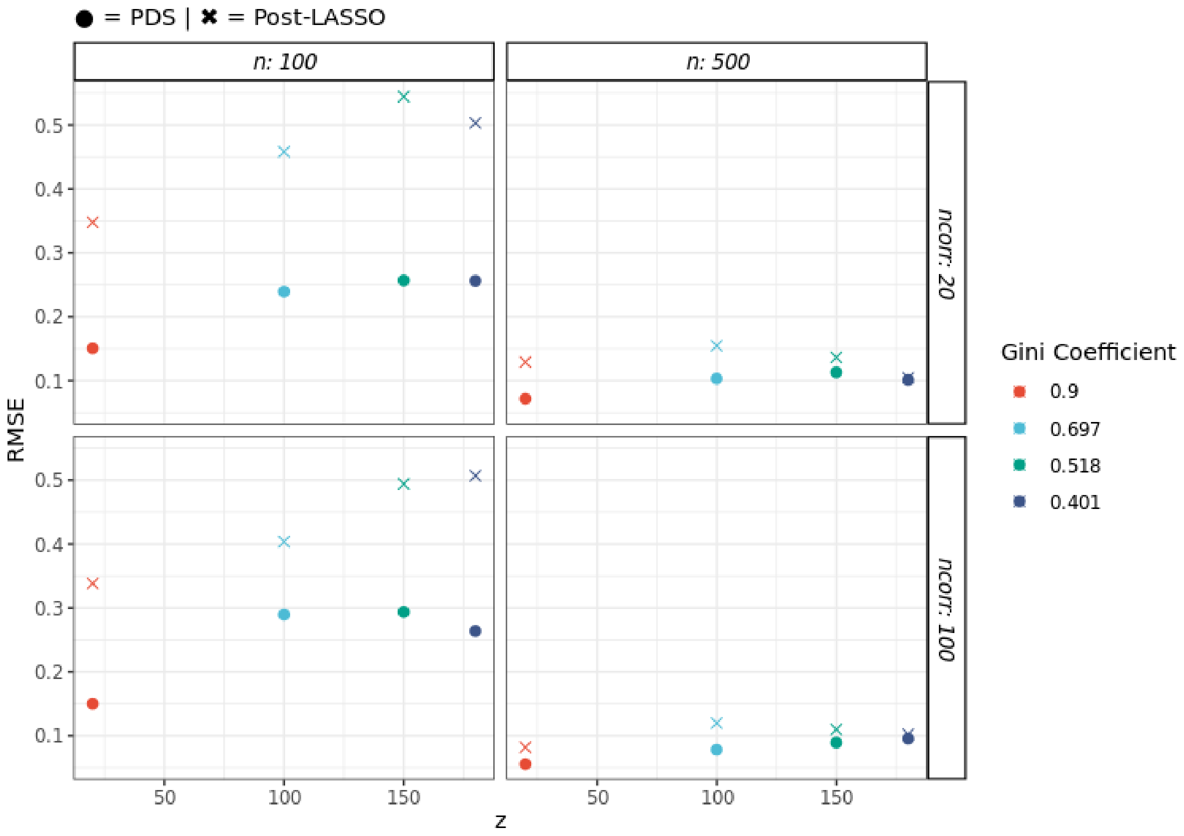


Figure 4.6: *DGP2 RMSE against z, where* $n = (100, 500)$ *and ncorr is the number of correlated variables* $\kappa = (20, 100)$

As we can see in Figure 4.6, PDS estimates had a lower RMSE than the Post-LASSO in all of the experiments with the most significant difference with the smaller sample size ($n = 100$). The trend also shows that estimators become worse as the sparsity of the true coefficients decrease. This trend is perhaps less pronounced than seen in Figure 4.1 as there is less range in the Gini coefficient. The estimators also performed better with a larger sample size, with more minor differences between the PDS and Post-LASSO estimates. Surprisingly, the number of correlated variables did not significantly impact the RMSE for these parameter sets.

Both estimators performed best when $z = 20$ with a Gini coefficient of 0.9. PDS had a lower RMSE, bias-variance and closer $E(\hat{\alpha})$ in five of the six experiments compared to Post-LASSO. As shown in Figure 4.7, PDS had lower bias and variance overall. Post-LASSO only performed better than PDS when $n = 20$ & $\kappa = 100$ although neither produced accurate estimates.

When $z = 100$ with a Gini coefficient of 0.70, PDS had an RMSE less than 0.5 in five of the six experiments, performing best when the sample sizes were larger and with lower levels of covariance. PDS had a lower RMSE and variance in all experiments and lower bias in five of the six compared to Post-LASSO. Again, Post-LASSO only performed better when $n = 20$ & $\kappa = 100$.

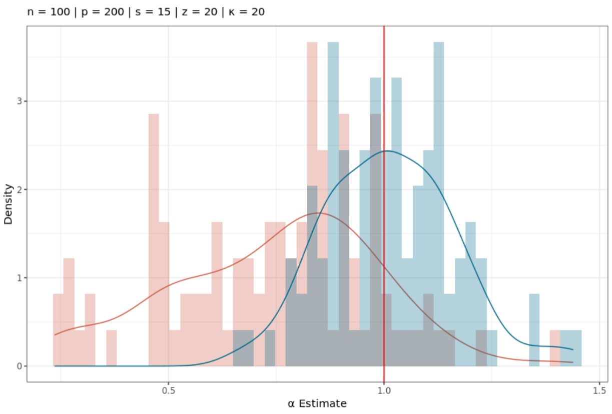


Figure 4.7: *DGP2 Density Plot when $n = 100$, $p = 200$ and $z = 20$. The distribution of Post-LASSO estimates is red and the distribution of PDS estimates is blue.*

As z increases to 150 with a Gini coefficient of 0.52, the estimates continue to worsen. The most notable decrease in both estimators' performance is with the smaller sample size ($n = 50$). For the Post-LASSO, performance is worse, only having an RMSE less than 0.5 in three of the six experiments. PDS had lower RMSE, bias and variance in all six experiments. When $z = 180$ with a Gini coefficient of 0.401, PDS continued to perform well except when $n = 20$ & $\kappa = 100$. PDS had a lower RMSE and variance in five of the six experiments; however, both estimators experienced higher levels of bias, each performing better in three of the six experiments.

In the sub-group of experiments with larger samples sizes ($n = 1000$), PDS had a low RMSE, bias and variance in all experiments, while Post-LASSO only performed better when the number of correlated variables was high. This could be due to the Post-LASSO selecting more controls which were only of benefit when κ was large.

The trend across these simulations has been for PDS to perform well apart from experiments where there were fewer observations and more covariance. Post-LASSO was more sensitive to changes in the Gini coefficient, seeing a significant increase in RMSE and bias.

Large Coefficients

Table 4.4 presents the simulations that varied the number of large coefficients (s) in DGP2. Overall, the PDS method did not perform as well as under DGP1 or the first sub-group under DGP2. Furthermore, in many of the simulations, the Post-LASSO was able to outperform the PDS method. This is seen in Figure 4.8 where Post-LASSO has a lower RMSE than PDS for every value of s apart from where $s = 20$. We still see a strong link between the coefficients' sparsity and the performance of the estimators when $n = 100$. However, when $n = 500$, the Post-LASSO is not impacted by the sparsity and has a low RMSE across all experiments.

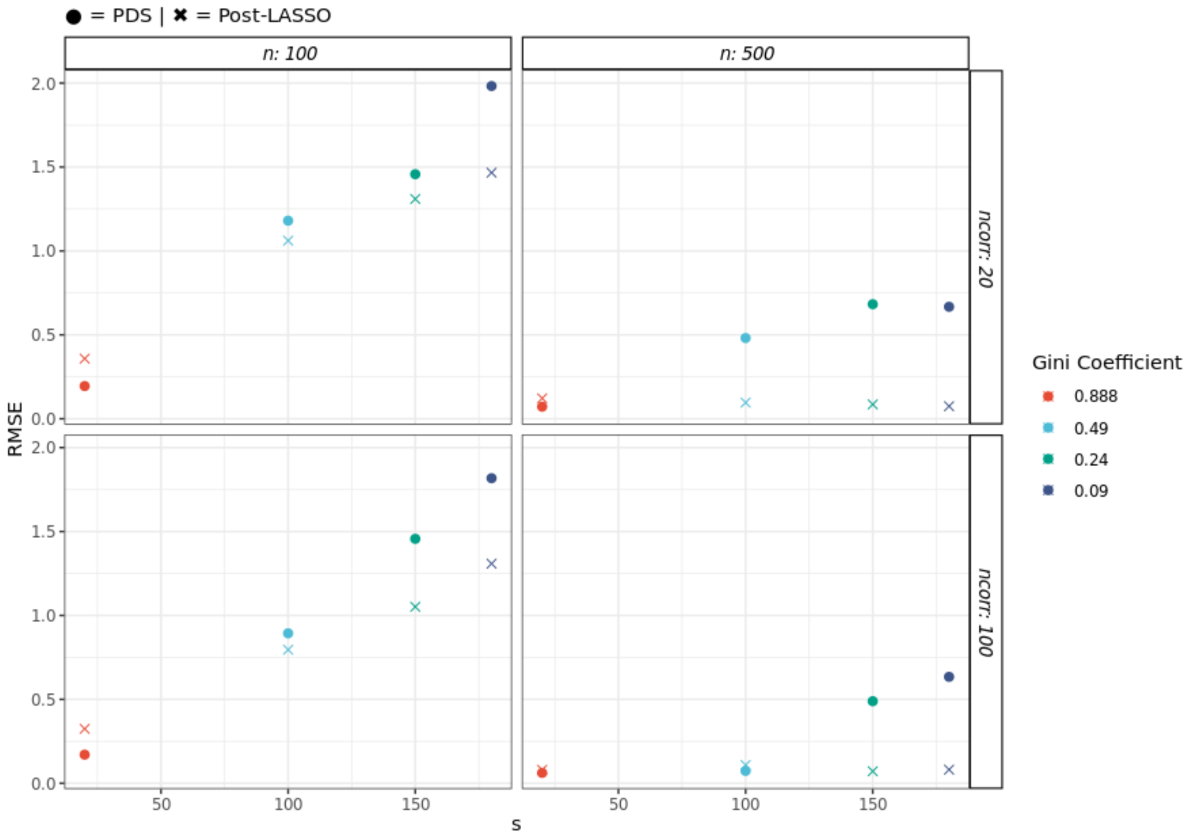


Figure 4.8: *DGP2 RMSE against s, where* $n = (100, 500)$ *and ncorr is the number of correlated variables* $\kappa = (20, 100)$

Both estimators performed best when $s = 20$ with a corresponding Gini coefficient of 0.88. In five of the six simulations, PDS had an RMSE less than 0.5. This was the only group of experiments where PDS performed better than Post-LASSO; however, as with the other experiments using DGP1&2, it did not perform well when $n = 20$ & $\kappa = 100$.

Apart from experiments with a low number of large coefficients ($s = 20$), the Post-LASSO nearly always performed better than PDS. Both estimators performed well when $n = 500$; however, Post-LASSO had a lower RMSE, bias and variance in nearly all of these experiments.

Both the PDS and Post-LASSO performed similarly when $n = 100$ & $p = 200$ and $n = 1000$ & $p = 2000$, with results closest to similar Gini coefficient values regardless of sample size.

The results from this group of experiments show that the coefficients' size had a significant impact on the estimator's performance and a greater impact than the sparsity of the coefficients. The sparsity level, given by the Gini index, still had an impact but less than the other DGPs. Post-LASSO had a lower RMSE in 20 of the 28 simulations.

4.3 DGP3

The results for the experiments using DGP3 are presented in Table 4.5. In these experiments, PDS performed consistently better than under the other 2 DGPs. Overall, PDS performed better than the Post-LASSO, having a lower RMSE, variance and bias in 25 of the 28 experiments. As discussed in section 3.1.1, three types of coefficients can be varied. The first being coefficients of variables that affect both the dependant variable and the treatment (β^c); the second being coefficients of variables that only affect the dependant variable (β^y); the third being coefficients of variables that only affect the treatment (β^e). I, therefore, group the experiments based on which coefficients are varied. The remainder of the section will be looking at the estimators' performance as the structure of the three coefficient groups change.

Figure 4.9 shows that there is not a clear link between the Gini coefficient and the RMSE for exogenous coefficients (β^e) or y coefficients (β^y). There is a weak link between the common coefficients' sparsity (β^c) and the RMSE. Across the experiments, the biggest negative impacts on PDS's performance were changes in the number of common variables (K_c) or the density of common coefficients (φ_c).

PDS performed best in experiments where exogenous coefficients β^e were varied. These estimates had the lowest RMSE, bias and variance of the three sub-groups of experiments. PDS performed similarly when varying coefficients that only impacted y_i. It might have been ex-

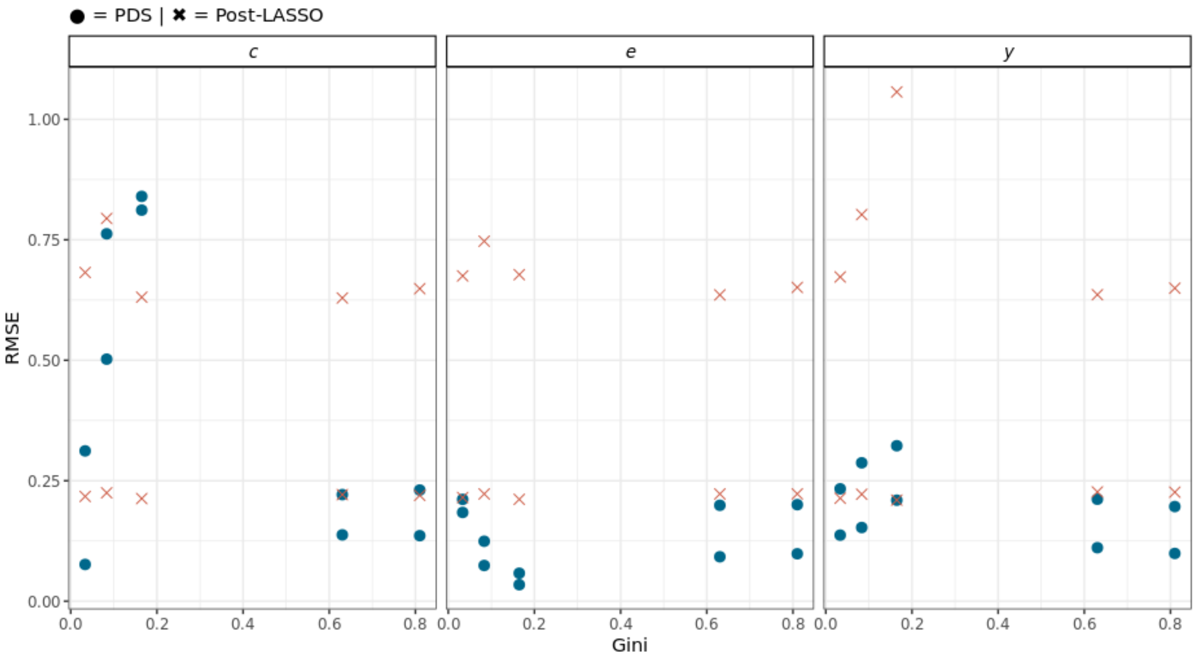


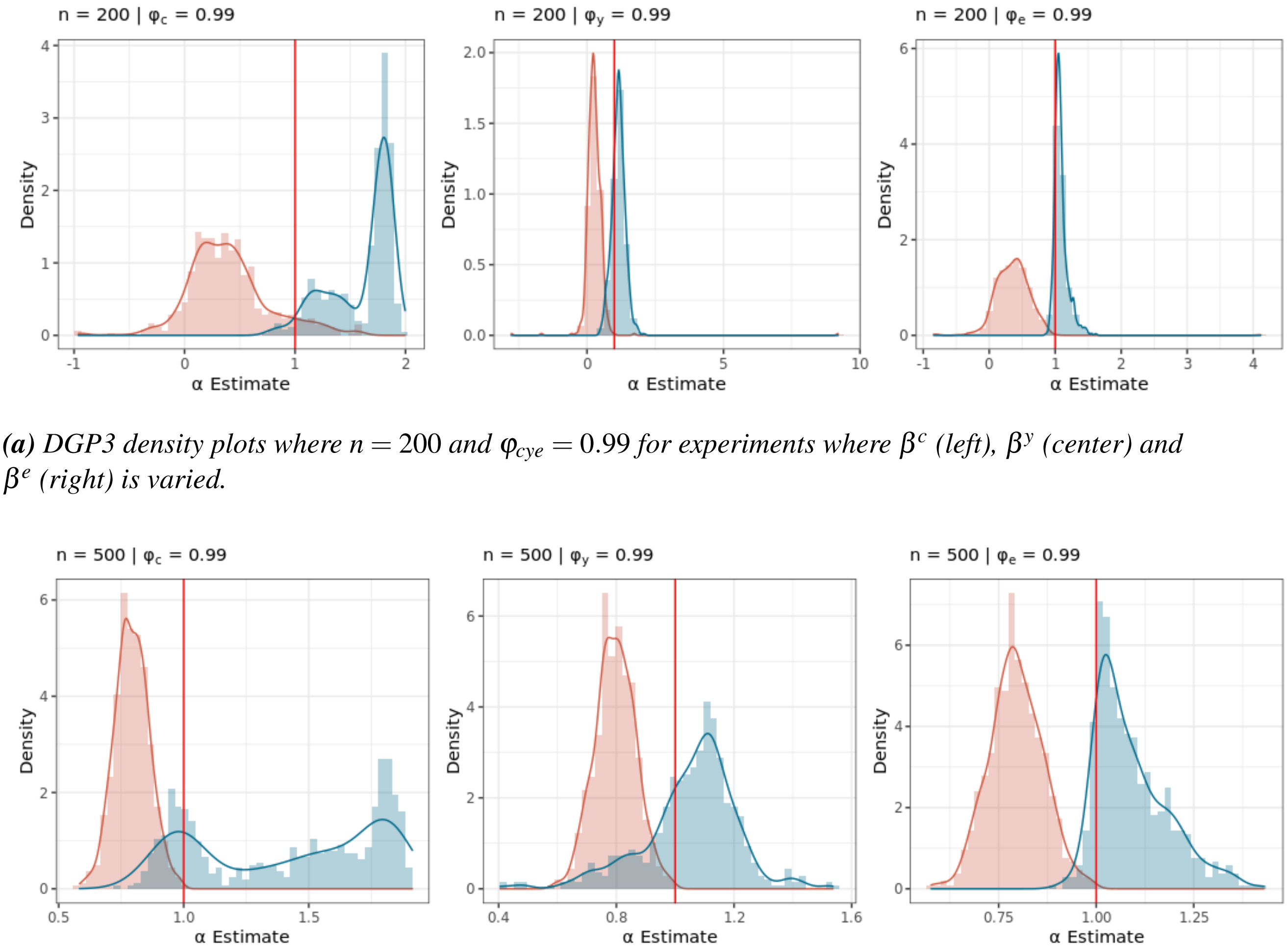
Figure 4.9: *DGP3 RMSE and Gini coefficient for PDS (blue circles) and Post-LASSO (red cross) when varying β^c (left), β^e (center) and β^y (right)*

pected that more / denser exogenous coefficients would have worsened the PDS estimates relative to the Post-LASSO's as the method would have been more likely to select irrelevant controls, introducing bias. This may not have happened due to the decaying nature and the smaller number of correlated variables relative to p, meaning that the impact of the exogenous or y variables was large enough to crowd out the effect of the common variables. The number of common coefficients had the largest impact on the RMSE, bias and variance with PDS performing worse when $K_c = 100$ and $\varphi_c = 0.99$, for both $n = 200\&500$. The next worse two performing estimates were when $K_c = 50$ and $\varphi_c = 0.99, n = 200\&500$.

The Post-LASSO performed nearly identically across most of the experiments. When $n = 500$ the RMSE, bias and variance remained consistency between 0.21-0.22, 0.20-0.22 and 0.17-0.18 respectively. This indicates that the Post-LASSO was selecting the same model for every simulation resulting in consistent, all be it biased, estimates. When $n = 200$, the RMSE, bias and variance had more variation but remained mostly within a narrow range. The estimator performed worse when varying the density of y coefficients, most likely as they were crowding out the common variables that were needed as controls. There were not any experiments that stood out where the performance of the Post-LASSO was improved.

There were only two experiments where the Post-LASSO had a lower RMSE than the PDS.

These were both experiments where K_c and φ_c were higher than their baseline and where the PDS performed poorly. This result was due more to the PDS performing worse than the Post-LASSO performing better. We can see this Figure 4.10. In both 4.10a and 4.10b, Post-LASSO was consistently biased across all of the parameter sets. When $n = 200$, we can see that only when $\varphi = 0.99$ does the PDS fail to give reliable estimates. When $n = 500$, Post-LASSO estimates were more focused around 0.8 and PDS continued to perform poorly when $\varphi_c = 0.99$. Overall, when $\varphi = 0.99$, there was also a slight increase in bias for PDS estimates and slightly larger variance. There appears to be a bi-modal distribution in the PDS estimates when varying common variables. This is due to the greater sensitivity to the number of β^c coefficients and is clearly illustrated in Figure C.1 in the Appendix.



***(a)** DGP3 density plots where $n = 200$ and $\varphi_{cye} = 0.99$ for experiments where β^c (left), β^y (center) and β^e (right) is varied.*

***(b)** DGP3 density plots where $n = 500$ and $\varphi_{cye} = 0.99$ for experiments where β^c (left), β^y (center) and β^e (right) is varied.*

Figure 4.10

Chapter 5

Discussion

The purpose of this study has been to investigate the performance and statistical properties of the post-double-selection estimator and the post-single-selection estimator (Post-LASSO) in their ability to recover the average treatment effect. The study also considered the impact of various sparsity levels in the coefficients on these estimators. This has been done through replicate simulations in the form of a Monte Carlo study, allowing for a precise understanding of the estimators' characteristics. In general, the observed properties of the estimators are in line with their respective theoretical foundations. Overall, the results are favourable to the PDS estimator as a tool for estimating average treatment effects. The results are less favourable for the Post-LASSO, where estimates are often more biased. That being said, there are plausible situations where both estimators fail or where Post-LASSO performs better than PDS.

This section will first discuss the Gini coefficient as a measure of sparsity and an indicator of the estimators' performance. I will then discuss unexpected outcomes in the results as well as the study's relation to the existing literature on post-selection estimators. The implications of the research are discussed throughout.

5.1 Analysis Summary

The Gini Coefficient

Considering the relationship between the Gini coefficient and the estimators' performance, we observe that the performance is an increasing function of the Gini coefficient. However, this relationship is dependent on various factors, including the functional form of the DGP. Across the three DGPs, the smallest sample size of $n = 50$ will be considered separately from other sample sizes as the impact of sparsity was often different.

For DGP1 experiments (Table 4.1 and Table 4.2), when $n = 50$ and at low covariance levels, the PDS estimator produced results with low bias when the Gini coefficient was greater than 0.65. The impact of increasing covariance was most pronounced in smaller sample sizes, where the estimator failed to produce reliable estimates of the treatment effect until the Gini coefficient was in the range of 0.85-0.9. The implication here for applied research is that in data settings where there are few observations but many potential controls, such as cross-country growth regressions, 'ultra sparsity' will likely be needed for the estimators to produce reliable estimates. The Post-LASSO required a higher Gini coefficient of around 0.8 to produce more reliable estimates but still experienced bias. Both estimators performed considerably better in experiments with larger sample sizes. When $n = 100$, PDS estimates became approximately unbiased with a low RMSE when the Gini coefficient was in the range of 0.5, while this was the case for Post-LASSO when the Gini coefficient was approximately 0.65. However, the bias of the Post-LASSO estimates was consistently higher than that of the PDS estimates. In the larger sample size experiments where $n = 500$, the Post-LASSO was able to select the model perfectly and therefore performed better than PDS. Because of this, the Gini coefficient did not relate to the Post-LASSO's performance. It did, however, for PDS, where the RMSE constantly decreased as the Gini coefficient increased. There was a significant increase in performance when the Gini coefficient rose above 0.3. We can also compare the performance of the smaller ($n < 1000$) and larger ($n = 1000$) sample size experiments while keeping the ratio of the number of observations to the number of parameters the same. The estimators' performance between the two experiments is similar at approximately the same Gini coefficients, although both estimators performed better in the large sample size experiments at the highest Gini values.

Considering DGP2 experiments (Table 4.3 and Table 4.4), the first group of small coefficient experiments are perhaps closest to what one might expect in an applied research setting with the structure of few variables with large coefficient and many variables with small coefficients. Similarly to DGP1, the point at which the estimator became reliable depends mainly on the sample size, with the impact of sparsity being more pronounced in smaller sample sizes. When $n = 50$, the results are more sensitive to changes in the number of correlated variables but improve as the Gini coefficient increases. PDS produced reliable estimates when the Gini coefficient was greater than approximately 0.7 when covariance is low. In the medium sample size experiments where $n = (100, 500)$, the PDS and Post-LASSO performance improved with lower RMSE and better mean estimates as the Gini coefficient increased. The PDS provides approximately unbiased estimates for all Gini coefficient values; however, interestingly, while the RMSE of Post-LASSO estimates improved as the Gini coefficient increased, the bias stayed approximately the same. Furthermore, the estimates are impacted less by the number of correlated variables. Overall, there was a weaker link between the Gini coefficient and the estimators' accuracy. This may be due to the lower spread of the Gini coefficient as it only ranged between 0.4 and 0.9. Again, performance was similar when comparing estimates from larger sample size experiments at similar Gini coefficient values.

In the second group of experiments under DGP2 with large coefficients, the Gini coefficients ranged from 0.01 to 0.9. As before, the accuracy of both estimators worsened as the density increased. When $n = 50$, neither estimator performed well and were more sensitive to changes in the number of correlated variables. However, the strongest relationship in these experiments was between the Gini coefficient and the variance. In experiments where the sample size was equal to 100, both the PDS and Post-LASSO were only able to produce unbiased estimates when the Gini coefficient reached 0.88. In the larger sample size experiments, PDS produced unbiased estimates when the Gini coefficient was greater than 0.45. For PDS, the relationship between the Gini coefficient and the summary statistics is similar; however, for Post-LASSO, the variance and Gini coefficient are more strongly related.

In the third DGP, the overall relationship between the Gini coefficient and the estimators' performance is least strong. Only for the PDS method, when varying variables affecting both the treatment and dependent variables does the Gini coefficient indicate performance, with the strongest effect being on the bias. As discussed in the results chapter, Post-LASSO produced near-identical results regardless of the parameter set and therefore was not impacted by the sparsity.

Chernozhukov, Hansen, and Spindler (2016) find that the sparsity conditions $x_i\beta$ for the PDS method needs to satisfy $s \ll \sqrt{n}$ and Jankova, Van De Geer, et al. (2018) point to $z \gg n/logp$ being insufficiently sparse. Interestingly, this study did not find these conditions to indicate the performance of the PDS estimator. For example, under DGP2, experiments that satisfied this conditions did perform well, but as did other experiment where $s > \sqrt{n}$ or $s \gg n/logp$. In the case of approximate sparsity, it is difficult to define s as it would depend on which point is defined as being sufficiently sparse.

Overall, both the bias and RMSE are increasing functions of the Gini coefficient. The effect of decreasing sparsity degraded the performance of the estimators as expected. That said, it was far weaker an indicator of performance under certain conditions such as experiments using DGP3. Other factors often had a more considerable impact, such as the number of correlated variables or the functional form of the data generating process. Ultimately, whether the Gini or the sparsity conditions proposed above, a single sparsity measure will be unlikely to capture the complexities dictating the estimators' performance. However, the Gini coefficient captures the sparseness better of approximately sparse models and sparse models with varying coefficient sizes.

Sample Size and Covariance

As mentioned already, there was a strong link between the sample size and the estimators' performance. The most varied estimates for both estimators was when $n = 50$, with a significant difference between high and low Gini coefficient values. Furthermore, the impact of the number of correlated variables was most pronounced in smaller sample sizes. This is unsurprising as the smaller sample size limits the degrees of freedom, reducing the potential complexity of the LASSO model (Zou, Hastie, Robert Tibshirani, et al. 2007). The effect of covariance in larger sample sizes for both PDS and Post-LASSO were modest in most circumstances and non-existent in a few.

Unexpected Outcomes

Both the impact of the number of the variables and the sparsity of the coefficients was largely predictable. However, there were a few simulations that produced unexpected results.

Firstly, there were several simulations where Post-LASSO had consistently better estimates than PDS. As has already been noted, when $n = 500$, Post-LASSO was often able to select the true model and estimate the average treatment effect precisely. These were experiments where the number of variables was no longer greater than the number of observations ($p < n$). The Post-LASSO was likely able to produce better estimates due to the type of selection procedure for the regularisation parameter λ. The CV method often selected more controls, leading to nearly all variables being included in the model in $p < n$ situations. However, this likely becomes a downside in situations with few correlated variables as the theory-driven penalty term of the PDS method adjusts better when fewer controls are needed or estimating structural parameters (Wang 2019).

Furthermore, the Post-LASSO most likely benefited from the construction of the covariance matrix as variables with the greatest explanatory power were often the variables needed as controls. Using a different covariance matrix structure, for example, where covariance with the treatment is assigned randomly throughout the variables, would likely result in the Post-LASSO producing substantially more biased results. The results from DGP3 are consistent with this hypothesis as there was a different covariance structure which led to consistently biased estimates. There were a few other notable examples where Post-LASSO outperformed PDS. These were mainly experiments where there was high density and many correlated variables. Again, this suggests that the Post-LASSO selected more controls due to the regularisation parameter.

Secondly, in DGP3, there were some unexpected results. Intuitively, we might have assumed that PDS would have performed worse when increasing the number of variables impacting only the treatment as the estimator would select more exogenous variables in favour of common variables, which were needed as controls. Furthermore, we might have expected the Post-LASSO have performed better in this setting as it would have avoided selecting these variables due to the first stage only selecting variables to predict the independent variable. This was not observed possibly due to the effect of the exogenous variables being relatively small compared to the total number of variables.

Overall, in nearly every simulation apart from those highlighted above, in experiments where both the Post-LASSO and PDS gave reasonable estimates, the PDS method nearly always performed better, having a lower RMSE and bias. The Post-LASSO performance indicates that it cannot reliably control for bias in many circumstances or to the same extent as PDS. As a tool for model selection, the Post-LASSO cannot be recommended unless relatively strong assumptions about the data are made.

Other Results

A number of other papers have used Monte Carlo simulations to analyse the statistical performance of various post-selection estimators. Closest to this paper is a study undertaken by Belloni, Chernozhukov, and Hansen (2014b) which compared the performance of PDS and Post-LASSO. Their simulations similarly included three DGPs, the first two with a similar functional form to DGP1. Their third DGP is closer to DGP2 but adds more uncertainty by including both deterministic and random coefficients. Their conclusions are similarly favourable to the PDS method over the Post-LASSO when recovering average treatment effects. They find that PDS performed consistently well overall combinations of R^2 and similar to the Oracle estimator in many simulations. The Post-LASSO performs similarly to PDS at low values of R^2 but becomes significantly worse at higher values.

Furthermore, the authors find that introducing random coefficients into the model does impact the PDS RMSE at low R^2 values but decreases performance at higher R^2 levels. Their study also tested for the effect of heteroskedasticity and found that this increases the bias and standard deviation for the PDS procedure. The Post-LASSO was more sensitive to both heteroskedasticity and random coefficients. From this, we could understand that introducing these elements into this simulation would decrease the Post-LASSO performance and make the PDS method relatively more attractive.

In the case of many small variables, as in DGP2, Belloni, Chernozhukov, and Hansen (ibid.) find that the PDS estimator's robustness can be improved by incorporating Ridge regression. However, they use K-fold CV to tune the regularisation parameter. In their discussion, it is not clear to what extent this difference in the regularisation parameter added to the robustness. As the results in this study indicate, in certain situations, there may be scope for the CV regularisation parameter to improve the PDS's robustness when used alongside the plug-in LASSO. A possible avenue of future research could explore how PDS estimates differ when using the plug-in LASSO and CV LASSO and if this indicates anything about the true data generating process.

Since the introduction of the PDS method, many other papers have constructed asymptotically normal estimators for use in high-dimensional settings. In more recent years, we have seen the rise of alternative one-stage and two-stage estimators such as the R-Split and de-sparsified LASSO and methods such as data splitting or jackknife re-sampling, which have been shown to improve variance and bias (Wang 2019). This paper adds to this rapidly growing literature at the intersection of machine learning and econometrics. Understanding the limitations of an estimator is of great importance when doing applied work, and this paper hopefully gives researchers greater insight and a better intuitive understanding of post-double-selection and Post-LASSO methodologies.

Chapter 6

Conclusion

The purpose of this study was to assess the performance and properties of the post-double-selection method and the Post-LASSO in their ability to recover the average treatment effects. Monte Carlo Simulation experiment showed that although neither estimator was found to outperform the other across all experiments, the post-double-selection estimator was favourable in most circumstances. The Post-LASSO estimator showed that it could be used in certain situations and showed promise in settings where p was no longer greater than n.

As social scientists are increasingly presented with new and complex data sets that pose challenges to traditional statistic and econometric methods, new methods to assess research will need to be used. The PDS method appears to be fairly robust in high to medium sparse data sets. However, as these methods have been shown to be sensitive to the underlying data generating process, researchers need to be cautious when extrapolating the results. Extra consideration should be given to the sparsity assumption when using data sets with fewer observations. That said, the results presented here present a promising tool for future empirical research.

References

Ahrens, Achim, Christopher Aitken, and Mark E. Schaffer (2021). “Using Machine Learning Methods to Support Causal Inference in Econometrics.” en. In: *Behavioral Predictive Modeling in Economics*. Studies in Computational Intelligence. Cham: Springer International Publishing, pp. 23–52. ISBN: 978-3-030-49728-6. DOI: 10.1007/978-3-030-49728-6_2 (cit. on pp. 8, 15).

Arnold, Barry C (2012). *Majorization and the Lorenz order: A brief introduction*. Vol. 43. Springer Science & Business Media (cit. on p. 13).

Athey, Susan (2018). “The impact of machine learning on economics.” In: pp. 507–547 (cit. on p. 8).

Athey, Susan and Guido Imbens (Mar. 2019). “Machine Learning Methods Economists Should Know About.” In: *arXiv:1903.10075 [econ, stat]*. arXiv: 1903.10075 [econ, stat] (cit. on p. 1).

Athey, Susan and Guido W. Imbens (2017). “The State of Applied Econometrics: Causality and Policy Evaluation.” In: *Journal of Economic Perspectives* 31.2, pp. 3–32. DOI: 10.1257/jep.31.2.3 (cit. on pp. 1, 2).

Belloni, Alexandre, Daniel Chen, et al. (2012). “Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain.” en. In: *Econometrica* 80.6, pp. 2369–2429. ISSN: 1468-0262. DOI: 10.3982/ECTA9626 (cit. on pp. 19, 22).

Belloni, Alexandre and Victor Chernozhukov (2011). “High dimensional sparse econometric models: An introduction.” In: *Inverse Problems and High-Dimensional Estimation*. Springer, pp. 121–156 (cit. on pp. 1, 22).

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (Dec. 2011). “Inference for High-Dimensional Sparse Econometric Models.” In: *arXiv:1201.0220 [econ, stat]*. arXiv: `1201.0220 [econ, stat]` (cit. on p. 13).

– (2014a). “High-Dimensional Methods and Inference on Structural and Treatment Effects.” In: *Journal of Economic Perspectives* 28.2, pp. 29–50. DOI: `10.1257/jep.28.2.29` (cit. on pp. 15, 16).

– (2014b). “Inference on treatment effects after selection among high-dimensional controls.” In: *The Review of Economic Studies* 81.2, pp. 608–650 (cit. on pp. 2, 3, 15, 16, 46).

Carsey, Thomas M and Jeffrey J Harden (2013). *Monte Carlo simulation and resampling methods for social science*. Sage Publications (cit. on p. 23).

Chandler, Dana, Steven D Levitt, and John A List (2011). “Predicting and preventing shootings among at-risk youth.” In: *American Economic Review* 101.3, pp. 288–92 (cit. on p. 8).

Chernozhukov, Victor, Christian Hansen, and Martin Spindler (2016). “High-dimensional metrics in R.” In: *arXiv preprint arXiv:1603.01700* (cit. on p. 44).

Dalton, Hugh (1920). “The measurement of the inequality of incomes.” In: *The Economic Journal* 30.119, pp. 348–361 (cit. on p. 13).

Friedman, Jerome, Trevor Hastie, Rob Tibshirani, et al. (Feb. 2021). *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models* (cit. on p. 56).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York (cit. on pp. 9, 10, 13).

Heiberger, Tony Plate and Richard (July 2016). *Abind: Combine Multidimensional Arrays* (cit. on p. 56).

Hlavac, Marek (May 2018). *Stargazer: Well-Formatted Regression and Summary Statistics Tables* (cit. on p. 56).

Hoerl, Arthur E and Robert W Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems.” In: *Technometrics* 12.1, pp. 55–67 (cit. on p. 12).

Holland, Paul W (1986). “Statistics and causal inference.” In: *Journal of the American statistical Association* 81.396, pp. 945–960 (cit. on pp. 5, 6).

Hurley, N. and S. Rickard (Oct. 2009). “Comparing Measures of Sparsity.” In: *IEEE Transactions on Information Theory* 55.10, pp. 4723–4741. ISSN: 1557-9654. DOI: 10.1109/TIT.2009.2027527 (cit. on pp. 13, 55).

Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (cit. on pp. 5, 6, 8).

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer (cit. on pp. 8, 9).

Jankova, Jana, Sara Van De Geer, et al. (2018). “Semiparametric efficiency bounds for high-dimensional models.” In: *Annals of Statistics* 46.5, pp. 2336–2359 (cit. on pp. 11, 44).

Kang, Jun Seok et al. (2013). “Where not to eat? Improving public policy by predicting hygiene inspections using online reviews.” In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1443–1448 (cit. on p. 8).

Karvanen, Juha and Andrzej Cichocki (2003). “Measuring Sparseness Of Noisy Signals.” In: *4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (Ica2003*, pp. 125–130 (cit. on p. 14).

Kleinberg, Jon et al. (May 2015). “Prediction Policy Problems.” In: *The American economic review* 105.5, pp. 491–495. ISSN: 0002-8282. DOI: 10.1257/aer.p20151023 (cit. on p. 8).

Kuhn, Max et al. (Mar. 2020). *Caret: Classification and Regression Training* (cit. on p. 56).

Lorenz, Max O (1905). “Methods of measuring the concentration of wealth.” In: *Publications of the American statistical association* 9.70, pp. 209–219 (cit. on p. 14).

Matsumoto, Makoto and Takuji Nishimura (1998). “Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator.” In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8.1, pp. 3–30 (cit. on p. 23).

Müller, Kirill et al. (Feb. 2021). *Tibble: Simple Data Frames* (cit. on p. 56).

Neyman, Jersey (1923). “On the applications of the theory of probability to agricultural experiments: Essay of the principles.” In: *Roczniki Nauk Rolniczych* 10, pp. 1–51 (cit. on p. 5).

O’Brien, Carl M (2016). *Statistical learning with sparsity: The lasso and generalizations*. Wiley Periodicals, Inc. (cit. on pp. 2, 11–13).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/` (cit. on p. 23).

Rickard, Scott and Maurice Fallon (2004). "The Gini Index of Speech." en. In: *Proceedings of the 38th Conference on Information Science and Systems (CISS'04*, p. 6 (cit. on pp. 13, 14).

Ripley, Brian et al. (Feb. 2021). *MASS: Support Functions and Datasets for Venables and Ripley's MASS* (cit. on p. 56).

Rockoff, Jonah E et al. (2011). "Can you recognize an effective teacher when you recruit one?" In: *Education finance and Policy* 6.1, pp. 43–74 (cit. on p. 8).

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects." In: *Biometrika* 70.1, pp. 41–55 (cit. on p. 7).

Rubin, Donald B (2005). "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." In: *Journal of the American Statistical Association* 100.469, pp. 322–331. ISSN: 0162-1459 (cit. on pp. 5, 6).

Spindler, Martin et al. (Jan. 2019). *Hdm: High-Dimensional Metrics* (cit. on p. 56).

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288 (cit. on pp. 2, 10).

Wang, Jingshen (2019). "Debiased Post Selection Inference." Doctoral dissertation (cit. on pp. 45, 47).

Wickham, Hadley, Winston Chang, et al. (Dec. 2020). *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (cit. on p. 56).

Wickham, Hadley, Romain François, et al. (Feb. 2021). *Dplyr: A Grammar of Data Manipulation* (cit. on p. 56).

Wickham, Hadley, Jim Hester, et al. (Oct. 2020). *Readr: Read Rectangular Text Data* (cit. on p. 56).

Wickham, Hadley and RStudio (Jan. 2021a). *Forcats: Tools for Working with Categorical Variables (Factors)* (cit. on p. 56).

– (Mar. 2021b). *Tidyr: Tidy Messy Data* (cit. on p. 56).

Zeileis, Achim and Christian Kleiber (July 2014). *Ineq: Measuring Inequality, Concentration, and Poverty* (cit. on p. 56).

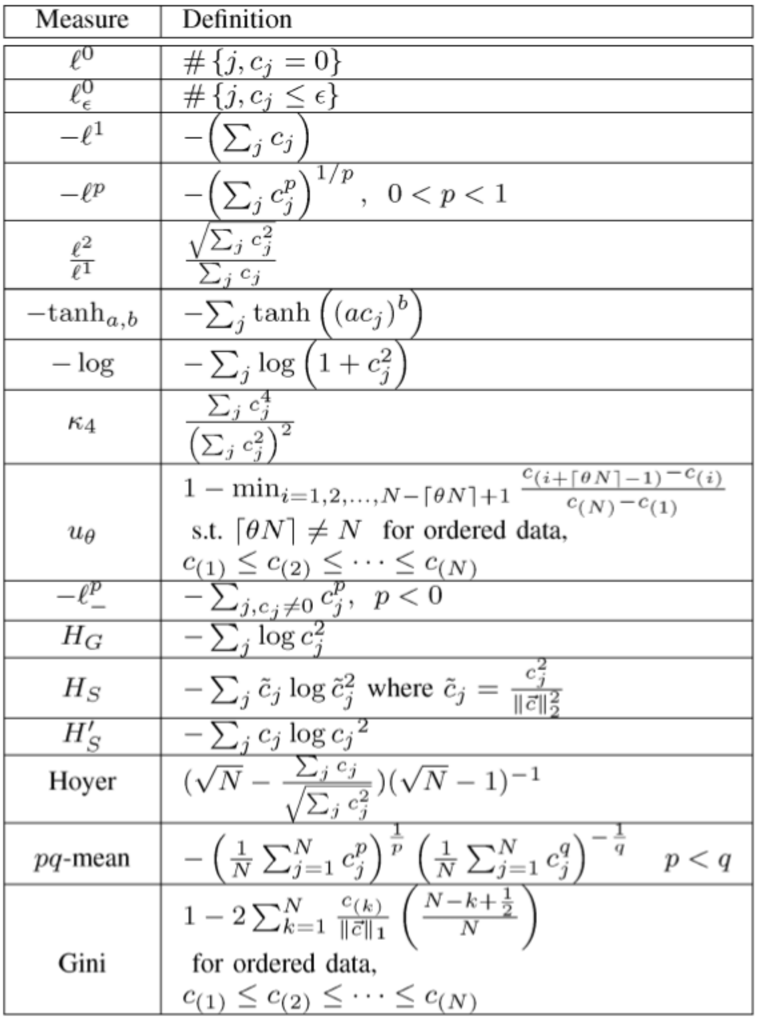
Zhao, Peng and Bin Yu (2006). “On model selection consistency of Lasso.” In: *The Journal of Machine Learning Research* 7, pp. 2541–2563 (cit. on p. 2).

Zou, Hui, Trevor Hastie, Robert Tibshirani, et al. (2007). “On the “degrees of freedom” of the lasso.” In: *The Annals of Statistics* 35.5, pp. 2173–2192 (cit. on p. 44).

Appendices

Appendix A

Literature Review



Measure	Definition
ℓ^0	$\#\{j, c_j = 0\}$
ℓ^0_ϵ	$\#\{j, c_j \leq \epsilon\}$
$-\ell^1$	$-\left(\sum_j c_j\right)$
$-\ell^p$	$-\left(\sum_j c_j^p\right)^{1/p}, \; 0 < p < 1$
$\frac{\ell^2}{\ell^1}$	$\frac{\sqrt{\sum_j c_j^2}}{\sum_j c_j}$
$-\tanh_{a,b}$	$-\sum_j \tanh\left((ac_j)^b\right)$
$-\log$	$-\sum_j \log\left(1 + c_j^2\right)$
κ_4	$\frac{\sum_j c_j^4}{\left(\sum_j c_j^2\right)^2}$
u_θ	$1 - \min_{i=1,2,\ldots,N-\lceil \theta N \rceil + 1} \frac{c_{(i+\lceil \theta N \rceil - 1)} - c_{(i)}}{c_{(N)} - c_{(1)}}$ s.t. $\lceil \theta N \rceil \neq N$ for ordered data, $c_{(1)} \leq c_{(2)} \leq \cdots \leq c_{(N)}$
$-\ell^p_-$	$-\sum_{j, c_j \neq 0} c_j^p, \; p < 0$
H_G	$-\sum_j \log c_j^2$
H_S	$-\sum_j \tilde{c}_j \log \tilde{c}_j^2$ where $\tilde{c}_j = \frac{c_j^2}{\lVert \vec{c} \rVert_2^2}$
H'_S	$-\sum_j c_j \log {c_j}^2$
Hoyer	$(\sqrt{N} - \frac{\sum_j c_j}{\sqrt{\sum_j c_j^2}})(\sqrt{N} - 1)^{-1}$
pq-mean	$-\left(\frac{1}{N}\sum_{j=1}^N c_j^p\right)^{\frac{1}{p}} \left(\frac{1}{N}\sum_{j=1}^N c_j^q\right)^{-\frac{1}{q}} \quad p < q$
Gini	$1 - 2\sum_{k=1}^N \frac{c_{(k)}}{\lVert \vec{c} \rVert_1} \left(\frac{N - k + \frac{1}{2}}{N}\right)$ for ordered data, $c_{(1)} \leq c_{(2)} \leq \cdots \leq c_{(N)}$

Figure A.1: *Sparsity Measures (Hurley and S. Rickard 2009)*

Appendix B

Research Methodology

R Packages		
MASS (Ripley et al. 2021)	dplyr (Wickham, François, et al. 2021)	forcats (Wickham and RStudio 2021a)
ggplot2 (Wickham, Chang, et al. 2020)	readr (Wickham, Hester, et al. 2020)	tibble (Müller et al. 2021)
tidyr (Wickham and RStudio 2021b)	ineq (Zeileis and Kleiber 2014)	hdm (Spindler et al. 2019)
caret (Kuhn et al. 2020)	glmnet (Friedman, Hastie, Rob Tibshirani, et al. 2021)	stargazer (Hlavac 2018)
abind (Heiberger 2016)		

Table B.1: *All packages used in completion of study*

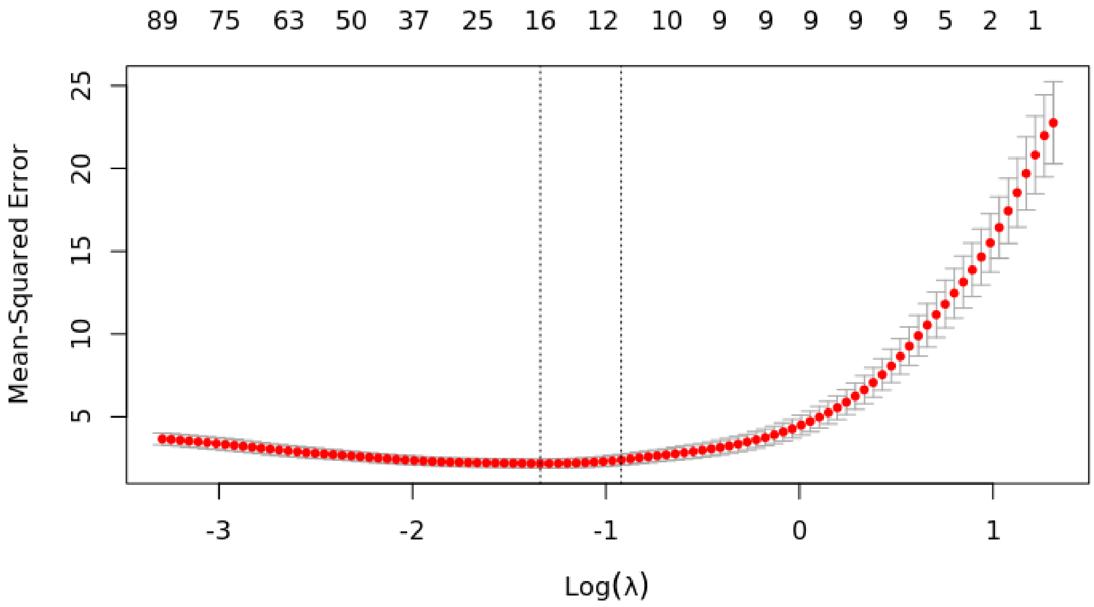


Figure B.1: *10-fold Cross-validation Plot for penalty level*

Table B.2: *DGP1. Description of the parameter sets for each simulations experiment using DGP1*

n	p	φ	κ	n	p	φ	κ
50	200	0.9	20	500	200	0.9	20
			100				100
		0.95	20			0.95	20
			100				100
		0.97	20			0.97	20
			100				100
		0.98	20			0.98	20
			100				100
		0.99	20			0.99	20
			100				100
		0.999	20			0.999	20
			100				100
100	200	0.9	20	1000	2000	0.9	20
			100				500
		0.95	20			0.95	20
			100				500
		0.97	20			0.97	20
			100				500
		0.98	20			0.98	20
			100				500
		0.99	20			0.99	20
			100				500
		0.999	20			0.999	20
			100				500

Table B.3: *DGP2. Description of the parameter sets for each simulations experiment using DGP2*

N	p	s	z	κ	N	p	s	z	κ
50	200	20	10	20	50	200	15	20	20
		20	10	100			15	20	100
		100	10	20			15	100	20
		100	10	100			15	100	100
		150	10	20			15	150	20
		150	10	100			15	150	100
		180	10	20			15	180	20
		180	10	100			15	180	100
100	200	20	10	20	100	200	15	20	20
		20	10	100			15	20	100
		100	10	20			15	100	20
		100	10	100			15	100	100
		150	10	20			15	150	20
		150	10	100			15	150	100
		180	10	20			15	180	20
		180	10	100			15	180	100
500	200	20	10	20	500	200	15	20	20
		20	10	100			15	20	100
		100	10	20			15	100	20
		100	10	100			15	100	100
		150	10	20			15	150	20
		150	10	100			15	150	100
		180	10	20			15	180	20
		180	10	100			15	180	100
1000	2000	1000	100	20	1000	2000	100	1000	20
		1000	100	500			100	1000	100
		1800	100	20			100	1900	20
		1800	100	500			100	1900	100

Table B.4: *DGP3. Description of the parameter sets for each simulations experiment using DGP3*

n	z	K_c	K_y	K_e	φ_c	φ_y	φ_e
200	400	20	20	20	0.9	0.9	0.9
		20			0.99	0.9	0.9
		50			0.9	0.9	0.9
		50			0.99	0.9	0.9
		100			0.9	0.9	0.9
		100			0.99	0.9	0.9
500	400	20	20	20	0.9	0.9	0.9
		20			0.99	0.9	0.9
		50			0.9	0.9	0.9
		50			0.99	0.9	0.9
		100			0.9	0.9	0.9
		100			0.99	0.9	0.9
200	400	20	20	20	0.9	0.99	0.9
			50		0.9	0.9	0.9
			50		0.9	0.99	0.9
			100		0.9	0.9	0.9
			100		0.9	0.99	0.9
500	400	20	20	20	0.9	0.99	0.9
			50		0.9	0.9	0.9
			50		0.9	0.99	0.9
			100		0.9	0.9	0.9
			100		0.9	0.99	0.9
200	400	20	20	20	0.9	0.9	0.99
				50	0.9	0.9	0.9
				50	0.9	0.9	0.99
				100	0.9	0.9	0.9
				100	0.9	0.9	0.99
500	400	20	20	20	0.9	0.9	0.99
				50	0.9	0.9	0.9
				50	0.9	0.9	0.99
				100	0.9	0.9	0.9
				100	0.9	0.9	0.99

Appendix C

Results

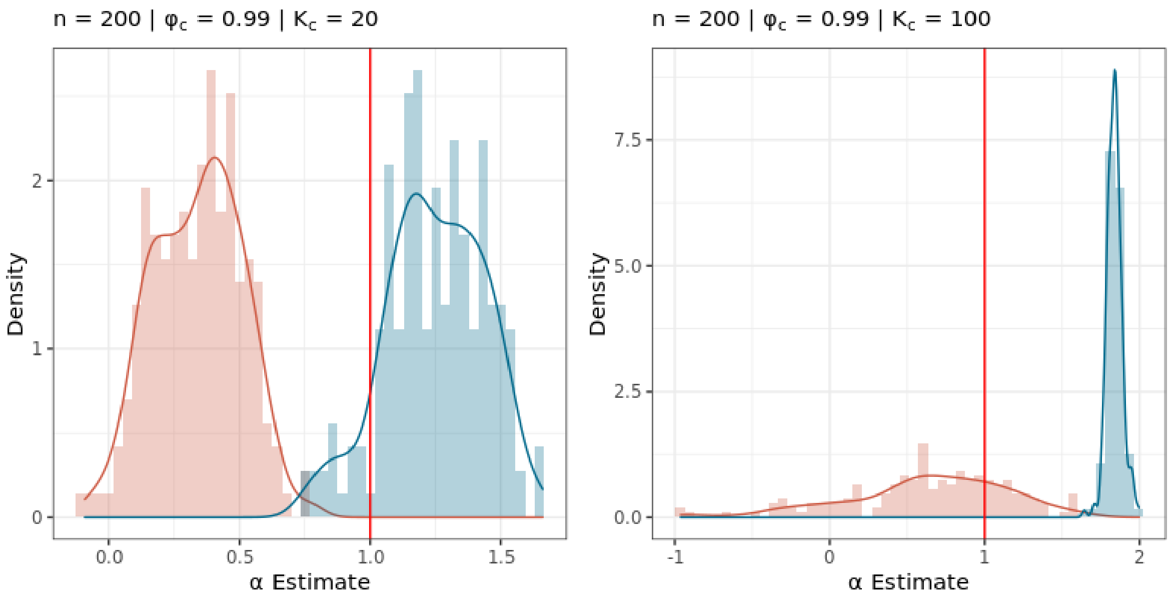
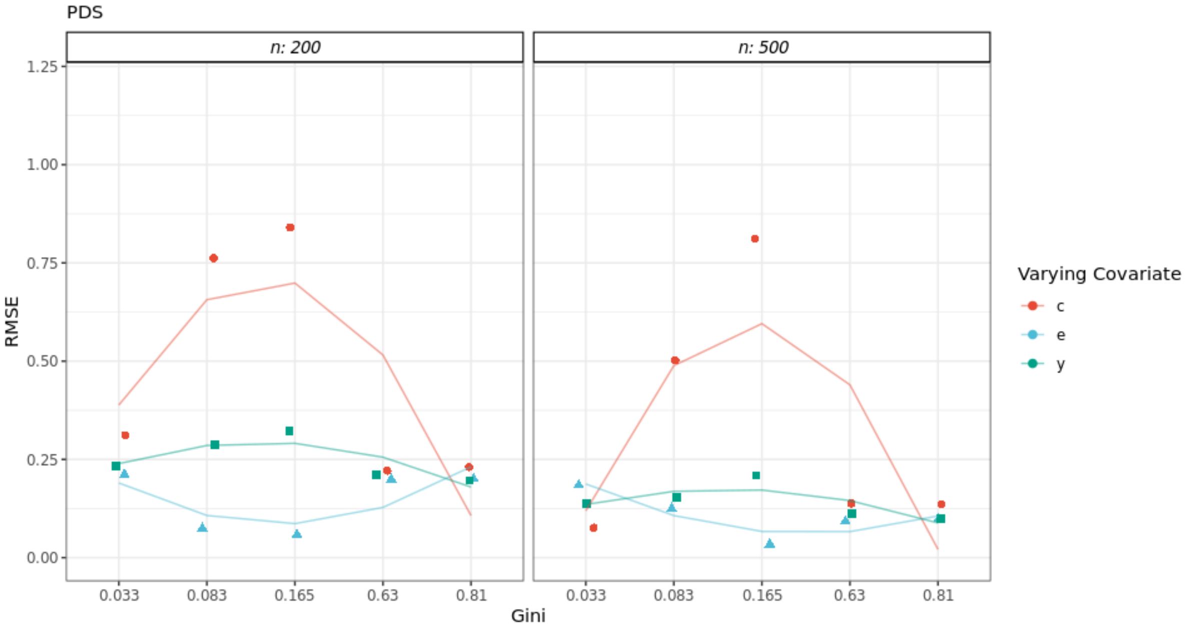
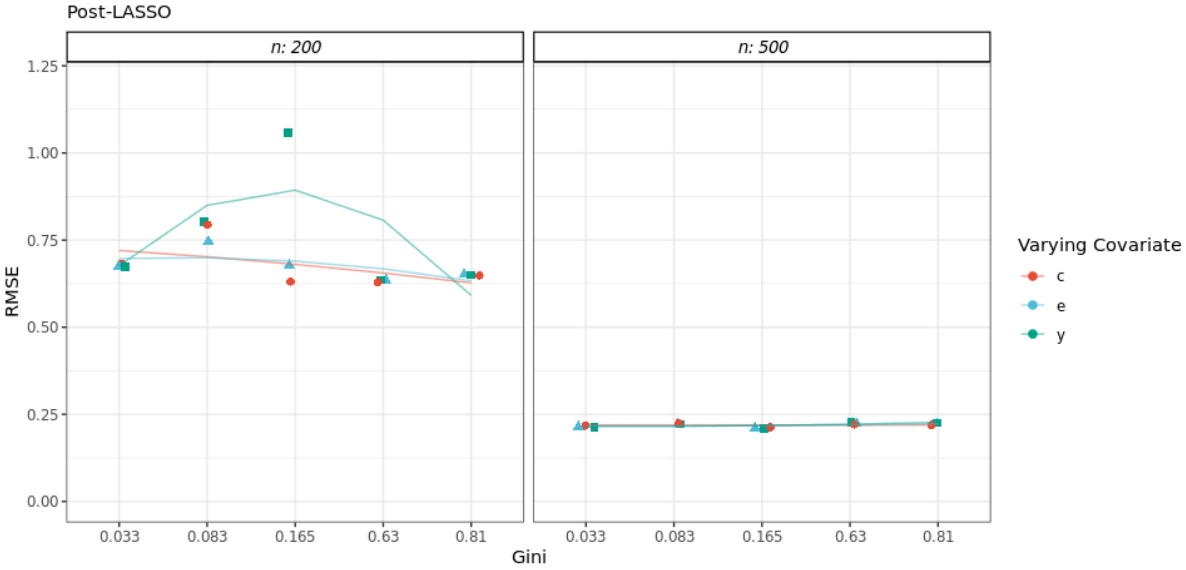


Figure C.1: *Plots display the distribution of the Post-LASSO $\hat{\alpha}$ estimate (red) and the PDS $\hat{\alpha}$ estimate (blue) where $n = 200$, φ_c = 0.99 and $K_c = 20$ (left) and $K_c = 100$ (right)*



(a)



(b)

Figure C.2: *Plots show RMSE against the Gini coefficient where* $n = (200, 500)$*, for DGP3 experiments. The colour and shape of the points indicate the varying coefficient. Plot (a) shows the RMSE produced by PDS. Plot (b) shows the RMSE produced by Post-LASSO*