THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

High-dimensional sample selection models

Machine-learning algorithms in the Heckman two-step

Ransi Weerasooriya

Supervised by Dr. Christiern Rose and Dr. Antonio Peyrache

October 26, 2018

Declaration

I hereby declare that the that the work presented in the Honours thesis, to the best of my knowledge and belief, original and my own work, except as acknowledged in the test and that material has not been submitted, either in whole or in part for a degree at this or any other university.

Ransi Weerasooriya

Dedication

To Ammi and Appachchi,
too many of whose calls were never returned while writing this thesis

Acknowledgments

This thesis would never have come into being if it weren't for Dr. Rose. Under his tutelage, my knowledge of econometrics has grown in leaps and bounds and his passion for high-dimensional statistics is simply inspiring.

Dr. Peyrache has been the cause behind my passion for econometric theory and his direction has kept me on track this year. If not for his encouragement of my adventurous ideas, this thesis would not be what it is.

Joshua Mcdonough has been simply amazing in painstakingly proof-reading and editing and his emotional support has been indispensable.

Jasmine and other members of the Miller family have been amazing hosts and have made my experience in Australia immemorial.

Kate Huang, Arabella Bennett and Pham Hoang Hai have been amazing colleagues and there are no better people to share this experience with.

The generous financial support of the School of Economics has made a world of difference to this year's experience.

Abstract

Selection biases contaminate inferences derived from those data that are non-randomly sampled. Many econometric methods have been proposed to salvage such data, principal among them is the Heckman two-step, which combines a predictive first stage with a substantive equation in the second stage (Heckman, 1976). It owes its claim to fame not only to its computational simplicity but also to its use of weaker distributional assumptions, its automatic test for sample selection and the robustness to measurement errors in the dependent variable. Despite this, however, many issues plague its usefulness, especially in high dimensions. Collinearity issues also cause estimates to be unstable. In this study, it is proposed to use a LASSO in a probit framework in the first stage to optimally predict the probability of sample selection without overfitting, and use this as a correction in a second stage which could be an OLS or a LASSO depending on the nature and dimensionality of the second stage equation. Simulation results show the above proposed estimator outperforms the Heckman two step even with only nearly high-dimensional data. These insights are then applied in the estimation of the female labour supply function and compared with the estimates of Mroz (1987).

Contents

List of Tables

List of Figures

Introduction

Economists have traditionally operated in low-data environments. This means that their major method of analysis has been top-down modelling. However, the developments of the age of 'Big Data' have meant that economists now have access to types of information that could previously only be dreamt of. This includes satellite imagery, machine readable images and genetic data (Mullainathan & Spiess, 2017). This is causing a paradigm shift in the way economists approach problems (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015). While economists are benefitting from this unprecedented availability, it has fallen upon econometricians to deal with these new types of often high-dimensional data. In fact, the task of prediction (one of the key functions of econometrics) have for a long time been underrepresented due to the discipline's bias towards structural estimation and causal inference. However, this is rapidly changing with predictive models gaining traction (Varian, 2014).

An interesting development is in how predictive algorithms are used to aid structural estimation. Algorithms that can deal with high dimensions and in the process do model selection are perhaps the most popular in this regard, since it allows for interpretability. These insights are used to improve IV estimation and identification of treatment effects by recognizing that some stages of structural estimation can be posed as a prediction problem (Belloni, Chernozhukov, & Hansen, 2014a).

The central idea of this thesis is to improve the performace of another common estimator. Popularly know as the 'Heckman two-step method' (Heckman, 1977), it is the most widely employed correction of sample selection bias. Despite, and perhaps due to its popularity, many practitioners have been unreserved in their skepticism of this method (Manning, Duan, & Rogers, 1987; Puhani, 2000). They argue that the instability of the estimates of the Heckman two-step makes it inferior to full-information maximum likelihood estimation or even two-part models. A few also have sought to provide justifications for these critiques. Leung and Yu (1996) defend the Heckman two-step, using Monte Carlo experiments to show that its somewhat erratic behaviour arises from collinearity and lack of variability in the regressors. Leung and Yu (2000) provide a detailed explanation as to the cause of

this collinearity. They provide several possible remedies but also acknowledge that these are either infeasible with observational data and at other times introduced a host of other problems.

This thesis proposes a variant of the Heckman two-step by combining a probit-LASSO in the first stage with an OLS or a LASSO in the second stage. This treats the first stage essentially as a classification problem and uses the LASSO in a probit framework to optimally predict the probability of selection. A function of this probability of selection, known as the inverse Mill's ratio, is then used to correct for sample selection bias.

The above estimator is introduced in the context of high-dimensional data (which allows for the number of covariates to exceed the sample size), since such datasets can help address some causes of collinearity in the two-step method. However, the Heckman two step as is traditionally used is expected to perform poorly in the face of high dimensionality due to the overfitting tendency of unpenalized regressions. Predictive algorithms such as LASSOs are specifically designed to keep overfitting at bay and yet uncover generalizeable structure.

High-dimensional methods also combine the utility of parametric and nonparametric methods since they allow many functional forms of variables to be embedded as polynomials and interactions in a parametric setting and yet be estimated without overfitting. As will be shown, the Heckman two step estimates tend to be sensitive to the specification of its first step design matrix. Therefore it might be paramount to find the right specification for the variables to be included in the first stage and functional form with which they are included.

One of the most commonly used remedies to mitigate the effects of collinearity on the Heckman two-step estimator is the use of exclusion restrictions. These are covariates that appear in the first stage but not in the second stage. However, there is no objective or theoretical basis for the choice of these instruments and may result in practitioners arbitrarily excluding a few variables. These arbitrary choices can introduce omitted variable bias, if in fact the variable excluded is relevant in the second stage. This proposed methodology provides a workaround to this situation. By recognizing that in most economic phenomena economic quantities could differ in the way they enter each process, high dimensional data and LASSOs are used to automatically recognize what these functional forms might be.

Chapter 1 introduces sample selection models and the framework used to deal with them. The same chapter introduces the discussion of the major contentions surrounding the Heckman two-step estimator. Chapter 2 introduces high dimensional statistics, its applications in econometrics and particularly structural estimation and

introduces the LASSO and its variants. Chapter 1 and Chapter 2 can be read in any order the reader prefers since they do not assume the knowledge of the other chapter. Chapter 3 continues the discussion of Chapter 1, proposes the new estimator and compares its performance to that of the Heckman two step using a simulation. This new proposed estimator is then applied to some of the specifications of Mroz (1987) in Chapter 4.

Chapter 1

Sample Selection Models

> When the United States military came around to the Statistical Research Group in Manhattan in 1943, wanting the answer to the optimal amount of armour a plane should carry to increase its chance of survival, they presented the mathematicians with the distribution of bullet holes in various sections of surviving planes. They naturally expected that the more bullet holes there were in a section, the more armour that section required. But a brilliant mathematician by the name of Abram Wald (also an econometrician) noticed that their analysis was flawed. The armour, he said, needed to go where the bullet holes were not, because the ones that came back with the bullet holes were the ones that did survive.
>
> Adapted from *How not to be wrong*
> Jordan Ellenberg

Abram Wald and the missing bullet holes is a classic example of biased inference in the face of non-random sampling. Non-random sampling plagues much of economic and sociological data, precisely due to the reliance on observational studies. Insofar as the data is used for causal inference, selection effects can impose upon the internal validity as well as the external validity of a research undertaking (Berk, 1983).

In linear regression models, selection bias can occur when the dependent variables are missing non-randomly, or more precisely, missing non-independently of

the exogenous regressors. In such a context, estimation using first-generation statistical methods can yield biased and inconsistent estimates of the causal impact these estimates have. By extension, sample selection causes a correlation between the regressors and the model error terms, a phenomenon more commonly known as endogeneity (Winship & Mare, 1992). An alternative way to view this is that truncation or censoring of variables introduces a non-linearity in the conditional mean function and any attempts to estimate these using linear regression methods introduces bias (Berk, 1983).

Perhaps the most well-known example of sample selection bias in economics is that of the female labour supply function, where the hours of formal work or the wage rate is only observed for women that are in the workforce. For women who do not work it is truncated at 0. Suppose the empiricist needs to estimate the effect that economic variables such as the number of children have on the labour supply. If any labour supply measure is regressed on economic variables only on the sample of women that are working, these estimates represent actual effects rather than potential effects (Puhani, 2000). For example, the estimate of the impact of an additional young child would be downward biased as only the most work-prone women would remain in work after a child. This would then suggest that had this sample been corrected for this non-random selection, the estimate of the impact of a child would be more pronouncedly negative (Heckman, 1977).

Other examples occur in sociology and health economics. Selection biases are pervasive in the criminal justice system, where those that are sent to prison are only a subsample of those arrested and the severity of the prison sentence therefore would likely be determined jointly with the likelihood of someone being arrested (Bushway, Johnson, & Slocum, 2007; Winship & Mare, 1992). In health economics, estimating the impact of smoking on medical expenditure might be impacted by sample selection, because the heavier smokers that have died from a fatal illness would be excluded from the sample. In this case the exclusion of this nonrandom subset leads to biased inferences on the impact of smoking (Berk, 1983).

In this chapter, the first section lays down the basic symbolic framework exploited in this thesis to describe sample selection models, the next section introduces and contrasts the two primary selection correction techniques including a technical breakdown of the inverse Mills' ratio. The third and final section introduces main points of contention between proponents of the above estimation techniques and explores the major criticisms against the most popular of the above method.

1.1 Sample selection models

Sample selection models are commonly characterised as a two equation process, comprised of a level equation and a selection equation;

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{(level equation)}$$
$$\mathbf{y}^* = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\upsilon} \qquad \text{(selection equation)}$$

$\mathbf{y}^*$ in the selection equation is a latent (unobserved) variable. In most cases therefore it will only be manifested as an indicator variable characterised below, where 1 denotes selection;

$$d_i = \begin{cases} 1 \quad if \quad y_i^* > T \\ \\ 0 \quad if \quad y_i^* \leq T \end{cases}$$

T is the level of truncation, but setting $T = 0$ is an inessential normalization and we follow this convention throughout the thesis.

Non-random sample selection is then introduced to this framework through the error terms, particularly if

$$\text{Covariance}(\epsilon, \upsilon) \neq 0.$$

This framework is also known as 'selection on observables and unobservables' because it is the correlation of the error terms that causes the bias in the estimates of the level equation. Contrast this with when

$$\text{Covariance}(\epsilon, \upsilon) = 0.$$

These models are known as 'two-part models' or 'selection on observables', precisely because the two processes are assumed to be independent.

If the true data generating process is in fact a sample selection framework, estimating just the level equation using linear regression methods can cause bias, since the conditional mean is also further conditional upon whether the observation was selected to the sample or not;

$$\begin{aligned} \mathbb{E}(y_i) &= \mathbb{E}(\mathbf{x}_i\boldsymbol{\beta}|\mathbf{z}_i\boldsymbol{\gamma} + \upsilon_i > 0) + \mathbb{E}(\epsilon_i|\mathbf{z}_i\boldsymbol{\gamma} + \upsilon_i > 0) \\ &= \mathbf{x}_i\boldsymbol{\beta} + \mathbb{E}(\epsilon_i|\mathbf{z}_i\boldsymbol{\gamma} + \upsilon_i > 0) \end{aligned}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are considered non-stochastic.

To estimate the level equation as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using linear regression methods would then omit the second term and create bias in the estimates of $\boldsymbol{\beta}$.

This rather common framework of sample selection can be generalized to include

many contexts. Amemiya (1985) classifies these models into five types based on the likelihood function. The model discussed above is called the 'Type 2 Tobit model', 'Bivariate sample-selection model' or 'incidental truncation'. The 'Type 1 Tobit' model arises when $\mathbf{X} = \mathbf{Z}$. When the selection divides observations into two separate substantive processes (both observed fully), such as wage functions with and without union membership two different substantive equations can be modelled. Analogously in a model such as the p-Tobit more than one selection process occurs before the observation appears in the sample (Deaton & Irish, 1984).

1.2 Correction for sample selection

Similar to the case where valid instrumental variables are required in models that control for endogeneity, sample selection biases can only be corrected if the sample is a censored sample. That is, there must be some observations that are selected to the sample and some that are not and the characteristics (except the dependent variable) of both types must be observed fully.

The first known introduction of sample selection in econometrics and its correction was in Tobin (1958) in what James Heckman describes as a 'justly celebrated' paper. Following this, it was not until 1974 that this issue was discussed again which followed Tobin's recommendation in using the likelihood function (Gronau, 1974; Heckman, 1974). Since then, two main techniques have emerged and have been extensively used in empirical settings, albeit with minor modifications.

These methods are the;

1. Full-Information Maximum Likelihood (hereafter, FIML)
2. Limited- Information Maxmimum Likelihood, more commonly known as the Heckman two-step method (hereafter, HTS)

While both these methods can be expressed generally for any probability density function and cumulative distribution function, the choice is made to do so under the assumption that the errors are distributed Gaussian. This helps to fix ideas and translates better to the later sections of this chapter since it is the most common distribution used.

Full Information Maximum Likelihood

Assume that,

$$\begin{bmatrix} \\ \epsilon_i \\ \\ \upsilon_i \\ \\ \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \\ 0 \\ \\ 0 \\ \\ \end{bmatrix}, \begin{bmatrix} \\ \sigma_\epsilon^2 & \rho \\ \\ \rho & 1 \\ \\ \end{bmatrix} \right)$$

Then the likelihood function of y_i is derived as follows,

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left\{ \mathbb{P}[y_i^* \leq 0] \right\}^{1-d_i} \left\{ f(y_i | y_i^* > 0) \times \mathbb{P}[y_i^* > 0] \right\}^{d_i} \\
&= \prod_{i=1}^{n} \left\{ \mathbb{P}[y_i^* \leq 0] \right\}^{1-d_i} \left\{ \int_0^\infty f(y_i^*, y_i) dy_i^* \right\}^{d_i} \\
&= \prod_{i=1}^{n} \left\{ \mathbb{P}[y_i^* \leq 0] \right\}^{1-d_i} \left\{ \int_0^\infty f(y_i^* | y_i) f(y_i) dy_i^* \right\}^{d_i} \\
&= \prod_{i=1}^{n} \Phi(-\mathbf{z}_i \boldsymbol{\gamma})^{1-d_i} \Phi\left(\left\{ \mathbf{z}_i \boldsymbol{\gamma} + \rho \left(y_i - \mathbf{x}_i \boldsymbol{\beta} \right) \sqrt{1 - \frac{\rho^2}{\sigma_\epsilon^2}} \right\} \times \frac{1}{\sigma_\epsilon} \left\{ \frac{y_i - \mathbf{z}_i \gamma}{\sigma_\epsilon} \right\} \right)^{d_i}
\end{aligned}$$

(Amemiya, 1985).

where ϕ and Φ is the p.d.f. and c.d.f. of the standard normal distribution, respectively.

Heckman two-step estimator

Assume that

$$\upsilon_i \sim \mathcal{N}(0, 1)$$

and

$$\epsilon_i = \rho \upsilon_i + \xi_i.$$

where $\mathbb{E}(\xi_i | \upsilon_i) = 0$.

$$\begin{aligned}
\mathbb{E}(y_i) &= \mathbb{E}(\mathbf{x}_i \boldsymbol{\beta} | \mathbf{z}_i \boldsymbol{\gamma} + \upsilon_i > 0) + \mathbb{E}(\epsilon_i | \mathbf{z}_i \boldsymbol{\gamma} + \upsilon_i > 0) \\
&= \mathbf{x}_i \boldsymbol{\beta} + \mathbb{E}(\epsilon_i | \mathbf{z}_i \boldsymbol{\gamma} + \upsilon_i > 0) \\
&= \mathbf{x}_i \boldsymbol{\beta} + \mathbb{E}(\rho \upsilon_i + \xi_i | \upsilon_i > -\mathbf{z}_i \boldsymbol{\gamma}) \\
&= \mathbf{x}_i \boldsymbol{\beta} + \rho \mathbb{E}(\upsilon_i | \upsilon_i > -\mathbf{z}_i \boldsymbol{\gamma}) \\
&= \mathbf{x}_i \boldsymbol{\beta} + \rho \phi(-\mathbf{z}_i \boldsymbol{\gamma}) / [1 - \Phi(-\mathbf{z}_i \boldsymbol{\gamma})] \\
&= \mathbf{x}_i \boldsymbol{\beta} + \rho \phi(\mathbf{z}_i \boldsymbol{\gamma}) / [\Phi(-\mathbf{z}_i \boldsymbol{\gamma})] \\
&= \mathbf{x}_i \boldsymbol{\beta} + \rho \lambda \left(\mathbf{z}_i \boldsymbol{\gamma} \right)
\end{aligned}$$

with

$$\lambda(\mathbf{z}_i\widehat{\gamma}) = \frac{\phi(\mathbf{z}_i\widehat{\gamma})}{\Phi(\mathbf{z}_i\widehat{\gamma})}$$

where the first and second equalities follow from section 1.1 and the third inequality exploits the linear relationship between the error terms. The fourth inequality uses the fact that ξ and υ are uncorrelated error terms, while the fifth inequality uses the truncated moments of the standard normal distribution. The next inequality is simplified by appealing to the symmetry of the standard normal distribution (Cameron & Trivedi, 2005).

Then the two step method is estimated using the following procedure,

1. Estimate $\mathbf{y}^* = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\upsilon}$ using maximum likelihood (standard normal errors implying this is a probit estimation).

1a. Use these estimates to calculate the inverse Mills' ratio as,

$$\lambda(\mathbf{z}_i\widehat{\gamma}) = \frac{\phi(\mathbf{z}_i\widehat{\gamma})}{\Phi(\mathbf{z}_i\widehat{\gamma})}$$

2. Then use this as an additional regressor in the estimation of the level equation for y, such that the new level equation is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\boldsymbol{\lambda}(\mathbf{Z}\widehat{\gamma}) + \boldsymbol{\eta}$$

Obtaining standard errors in this method is a bit more complicated, but the full derivation is provided in the Appendix of Chapter 16 of Cameron and Trivedi (2005). A test of whether sample selection exists ($H_0 : \ \rho = 0$) can be automatically performed using the regression output. The computational simplicity and immediate hypothesis test has led to wide popularity of the HTS among practitioners. However, there has been continued debate among econometricians around this method; some prefer the FIML, some contend that no correction is better than any while others have valiantly defended the use of the HTS. Much of the debate, however, is centered around the inverse Mills' ratio.

1.2.1 Inverse Mills Ratio: the devil in detail

There are several key features of the Inverse Mill's ratio (hereafter, IMR)

1. The denominator $\Phi(\mathbf{z_i}\hat{\boldsymbol{\gamma}})$ is the probability of sample inclusion for the i-th observation given the characteristics $\mathbf{z}_i$.

2. The inverse Mills' ratio is a monotone decreasing function of $\mathbf{z}_i\hat{\boldsymbol{\gamma}}$ and is a

monotone increasing function of the probability of sample selection $\Phi(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$.

$$\lim_{\mathbf{z}_i\hat{\boldsymbol{\gamma}}\to-\infty}\lambda_i=\infty \qquad \lim_{\mathbf{z}_i\hat{\boldsymbol{\gamma}}\to\infty}\lambda_i=0$$

(Heckman, 1977).

If $\mathbf{X}$ and $\mathbf{Z}$ are linearly dependent such as when $\mathbf{X}=\mathbf{Z}$, the IMR is identified in the second step because of its non-linearity. However, when the IMR is plotted on a grid from -10 to 10 it looks fairly linear across the segment $\mathbf{z}_i\hat{\boldsymbol{\gamma}}\in[-10,1]$, to the left of the kink. In fact, if one ran a regression of the IMR in that segment against the corresponding value of the domain, it would give an $R^2=0.9975$ suggesting almost perfect linearity along that segment. The non-linearity arises from the interval of values to the right of the kink. For example, $\lambda(2)=5.52\times10^{-2}$, $\lambda(4)=1.34\times10^{-4}$, $\lambda(6)=6.08\times10^{-9}$, $\lambda(8)=5.05\times10^{-15}$ (Leung & Yu, 2000).

Much of the literature in sample selection raises the concern that when the collinearity between $\mathbf{X}$ and $\boldsymbol{\lambda}(\mathbf{Z}\widehat{\boldsymbol{\gamma}})$ is high, the estimates $\boldsymbol{\beta}$ will become unstable and in fact the non-significance of the IMR in many empirical applications could be a result of this. Leung and Yu (2000) propounds that for the collinearity issues to be caused two conditions must be met,

1. all (or almost all) the data points $(\mathbf{z}_i\hat{\boldsymbol{\gamma}},\lambda_i(\mathbf{z}_i\hat{\boldsymbol{\gamma}}))$ in the selected sample fall on the left side of the kink,

2. $\mathbf{x}_i$ and $\mathbf{z}_i\hat{\boldsymbol{\gamma}}$ are highly collinear.

None of these criteria are sufficient by themselves for the collinearity issues to arise. A few remedies exist currently to deal with these issues and will be discussed here in Chapter 3. A plot of the IMR in the interval $[-10,10]$ is reproduced below.

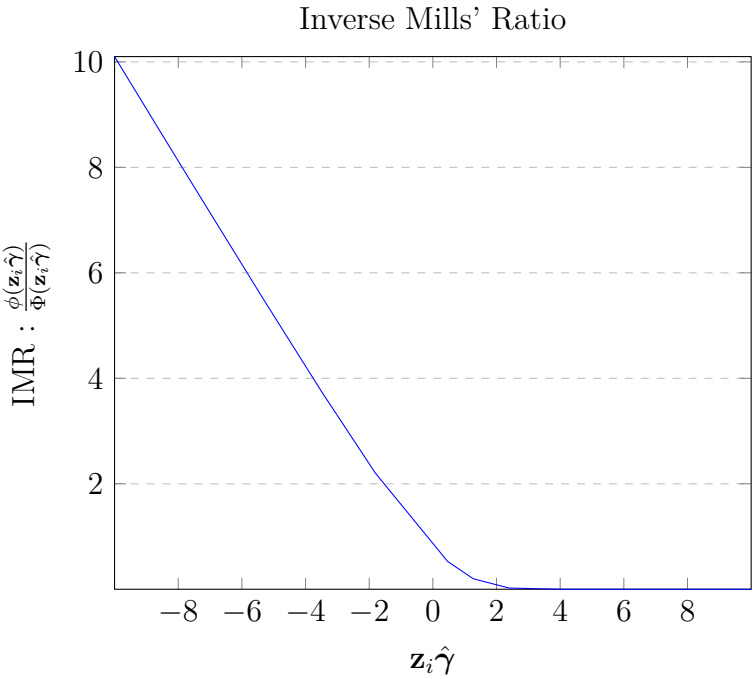

Figure 1.1: A plot of the inverse Mills' ratio

1.3 Heckman two-step method : the debate

Over the years this method has been subject to much criticism. Despite this, however, its computational simplicity and its automatic test for sample selection has meant that the HTS has prevailed over others. The primary goal of this thesis is to explore the role of machine learning algorithms in addressing some of these shortfalls. However, before doing that it is necessary to provide reasons why the HTS wins over its major contenders in aspects other than practicality.

HTS vs. Two-part models

'Two-part models' arise when the error terms of the selection and level equations are assumed to be uncorrelated. In such a case regressing $\mathbf{y}$ on $\mathbf{X}\boldsymbol{\beta}$ will not bias the estimates of $\boldsymbol{\beta}$. Two-part models are simply running OLS without the inverse Mills' ratio and are therefore also known as subsample OLS.

While two-part models are theoretically very different, some econometricians have proposed that the when collinearity exists between $\mathbf{x}_i$ and $\mathbf{z}_i$ two-part models are preferred to any of the other sample selection models. While it seems as if the two-part model is a nested case of sample selection models, i.e. where $\rho = 0$, this is a bit more complicated in practical estimation. In the LIML case, this could be caused by the behaviour of the inverse Mills' ratio, but the FIML suffers from similar deficiencies. Of course, the use of exclusion restrictions is still a workaround of this problem.

A series of Monte-Carlo experiments by Manning et al. (1987) concluded that two-part models were superior to sample selection models even when the true model was a sample selection model. Leung and Yu (1996) refute their claims noting that their design was substantially biased towards the two-part models and run their own Monte-Carlo experiment where the regressors were drawn from a distribution with more variability. Their principal findings are that sample selection correction methods (FIML and LIML) do perform well when the true model is sample selection. A low degree of censoring (number of observations that are not-selected) and exclusion restrictions also improve the performance of the method.

HTS vs. FIML

Here the main points of contention between the LIML and the FIML are explored in as much detail as necessary.

Computational simplicity

Since the introduction of the HTS estimator in Heckman (1976), it has been very popular among empiricists over and above its FIML cousin. While its computational simplicity and flexibility would have helped, the most compelling reasons may have been this footnoted observation in Heckman (1976);

> *An example of the potential cost saving maybe useful. It cost $700 to produce the estimates of the likelihood function reported in Table 3 and $15 to produce the initial consistent estimates and GLS estimates.*

Since then, however, computational capabilities have advanced greatly that has rendered the maximum likelihood estimation much cheaper. Despite this, however, the log-likelihood function above is not typically globally concave, and the estimates are sensitive to the choice of the starting values. In fact, the estimates of the HTS have proven to be effective starting values for the FIML (Leung & Yu, 2000). This issue is exacerbated in high-dimensional settings where computation is more expensive, since even though effective optimization tools exists for concave optimization (such as probit), these are not very useful if the objective function is not globally concave.

Computational flexibility is another issue. The many refinements that can be done to improve the performance of sample selection corrections are only possible in a two-step type framework where each estimation stage is treated separately.

Collinearity issues

Despite the assertions by Heckman (1976) that the HTS estimator produced estimates very similar to its FIML counterpart, this theoretical observation does

not translate fully in empirical settings. The main reasons being the collinearity introduced by the inverse Mills' ratio introduced above (Davidson, MacKinnon, et al., 1993; Puhani, 2000).

This, however, does not imply that the FIML method performs better when such collinearity exists. Leung and Yu (1996) does a series of Monte Carlo experiments and find that collinearity affects both estimators, but perhaps to a lesser degree in FIML.

Many corrections to this have been advanced. Primary among them is the use of exclusion restrictions. These are variables relevant in $\mathbf{Z}$ (selection equation) that are not relevant in $\mathbf{X}$. No objective means exist to isolate such exclusion restrictions.

Efficiency vs. Distributional assumptions

The FIML is more efficient than the HTS since it uses more information such as the higher-order moments of the joint distribution. Leung and Yu (2000) point out that this efficiency is only asymptotic. Some of their previous results have shown that in small samples HTS estimates have smaller bias than FIML estimates (Leung & Yu, 1996).

Furthermore, note that the FIML uses stricter distributional assumptions (joint normality of error terms) than the HTS. Non-parametric estimators of sample selection models have attempted to circumvent this issue.

Measurement errors

Stapleton and Young (1984) defends the HTS for its robustness when faced with a dependent variable that is measured with error. In such a case, the HTS absorbs the measurement error in their error term and the estimates will be consistent and in general more efficient than the FIML.

HTS vs. Non-parametric models

The Heckman two-step estimator relies on the assumption that the error term of the selection equation, υ_i, is normally distributed. If this is not the case, the estimates will be biased and inconsistent. A huge body of literature in semi-parametric and non-parametric statistics have emerged to account for this issue. In fact, much of the work done in sample selection literature at the turn of the century has been in non-parametric models (Newey, Powell, & Walker, 1990; Das, Newey, & Vella, 2003).

Interestingly, Newey et al. (1990) concluded that rather than the assumed distribution of the error terms, it was the assumed specification of the substantive and selection equations that resulted in the unstable estimates of the HTS. This con-

clusion echoed that of Mroz (1987) where his estimates were very similar across the normal, lognormal and logit distribution of the error terms.

There are many questions to be answered before putting this debate to rest. Even though the HTS dominates empirical work, it has many inherent weaknesses that it shares with the other methods as well as weaknesses that has left it open to criticism from practitioners of other methods. Some work has been done to alleviate these issues, the discussion of which is deferred to Chapter 3 as it is best introduced after a brief outline of high-dimensional statistics.

Chapter 2

High-dimensional methods

> For empiricists, theory and data driven modes of analysis have always coexisted. Many estimation approaches have been (often by necessity) based on top-down, theory-driven, deductive reasoning. At the same time, other [inductive] approaches have aimed to simply let the data speak. Machine learning provides a powerful tool to hear more clearly than ever, what the data have to say.
>
> *(Mullainathan & Spiess, 2017)*

The age of 'Big Data' signifies a time where advances in data capture and data storage have revolutionised our conception of 'data', with datasets not just increasing in length, but also breadth. While these changes were occurring, algorithms that crunched the numbers needed to keep up. The discipline that was born as a result is termed 'statistical learning' (more broadly, 'machine learning'), and exists in the intersection of computer science and statistics. The appeal of machine learning is in its predictive capabilities, to find generalized structure when it was not specified in advance. Examples of such algorithms enrich our everyday lives; neural networks recognize characters, random forests recognize faces and ensemble methods decide which websites we must first see when we google.

This thesis explicitly deals with 'supervised learning', whereby prediction is the primary goal rather than understanding how data points relate to each other as in 'unsupervised learning' (Trevor, Tibshirani, & Friedman, 2009). The term 'high-dimensional statistics' refer to the various refinements that must be done to algorithms when there are more covariates (k), than observations (N), which arises commonly in genomics and computational biology.

Statistical learning methods are effective predictive tools, since they can fit varied data structures by sorting through a large set of high-dimensional flexible functional

forms by looking for 'true' predictive power that is simply not overfitting. These algorithms achieve this through regularization and empirical tuning (Mullainathan & Spiess, 2017).

In this chapter an overview of high-dimensional methods are provided before introducing the LASSO and its many variants. Following from that, the many ways high-dimensional statistics has infiltrated econometrics is explored.

2.1 High-dimensional statistics

In cases where $k \gg N$, first generation statistical methods tend to work poorly. For instance, when $k = N$, an ordinary least squares regression will fit perfectly and when $k > N$, the estimates will not be unique. This reveals that even when $k < N$, but $k \approx N$, ordinary least squares regression will overfit and therefore would be a poor predictive model since the models cannot distinguish between noise covariates and those with true predictive power. The remedies for this issue is regularization and cross-validation, which will be discussed shortly.

Many algorithms exist for dealing with high-dimensional data. Some of the leading classes of methods are;

1. *Subset selection*. Identify the subset of the k covariates that best explain the response variable, either by forward or backward stepwise selection or best subset selection and run least squares on these limited set of covariates.

2. *Dimension reduction*. This involves projecting the k regressors to a lower dimensional (m) subspace, i.e. $k > N > m$. Principal components regression is an example.

3. *Shrinkage*. This method fits all k predictors but manages the overfit by shrinking the coefficients of the variables towards zero, where irrelevant variables are disproportionately shrunken. Least Squares Shrinkage and Selection Operators (hereafter, LASSOs) and ridge regressions are examples.

(James, Witten, Hastie, & Tibshirani, 2013).

A more general treatment of high-dimensional methods is given in the last chapter of Trevor et al. (2009).

Regularization

Regularization is the term used to describe the way the predictive algorithm manages overfit. Every such algorithm therefore must have a regularizer attached to it. Mullainathan and Spiess (2017) tabulates many of them. For the methods listed above regularizers are as follows;

1. *Subset selection* is regularized by the number of covariates that need to be chosen, also known as the $\mathcal{L}_0$ analogue.

2. *Dimension reduction* is regularized with the dimension of the lower-dimensional subspace, m.

3. *Shrinkage* methods are regularized using shrinkage hyperparameters, which are factored to a norm of the coefficients. For the LASSO it is the $\mathcal{L}_1$ norm and for the ridge regression it is the $\mathcal{L}_2$ norm, also known as the quadratic-regularizer.

Cross-validation

One might wonder how the practitioner knows what the 'right' number of covariates, the best m or what the test error-minimizing shrinkage factors might be. This is precisely the task of cross-validation, i.e. finding the best level of regularization.

Cross-validation creates a training-test experiment within the training sample. Here the data is split into many equally-sized parts (usually 5 or 10) and the algorithm is trained leaving one part out. Subsequently, the algorithm is tested on the part left out. This is done for each level of regularization (or an interval with as many points as computationally feasible) and the level that minimizes the loss function is chosen.

Details

Trevor et al. (2009) provide a more rigorous treatment of cross-validation. In notation, if L is the loss function, λ is the tuning parameter, T is the number of folds that data is split into and therefore $\hat{f}^{-\tau}(x, \lambda)$ is the model fit with the entire training set except the τ^{th} fold given λ (note that τ_i is the fold to which the i^{th} observation belongs), then the out-of-sample test error for each observation i is

$$L(y_i, \hat{f}^{-\tau_i}(x_i, \lambda))$$

Then for each λ, the estimate of the test error curve is

$$\mathrm{CV}(\hat{f}, \lambda) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\tau_i}(x_i, \lambda))$$

The $\hat{\lambda}$ is chosen so that the above function $\mathrm{CV}(\hat{f}, \lambda)$ is minimized.

The choice of T is a trade-off between bias and variance and would depend on the sample size. If $T = N$, the bias will be low but the variance will be very high and if $T = 2$ the converse applies. T is usually set to 5 or 10, depending on the

sample size. Trevor et al. (2009) provide a plot of the CV error as the size of the training sample changes.

2.2 LASSO and its cousins

Shrinkage methods are powerful methods to estimate high-dimensional models insofar as the underlying model is sparse in some sense. Sparsity can be defined using the $\mathcal{L}_0$ analogue. If s is a number less than K, sparsity implies that

$$\mathcal{L}_0 \leq s < K.$$

It might be interesting to note that depending on the choice of the regularizer shrinkage methods approximate the other high-dimensional techniques introduced above. For example, a quadratic regularizer (ridge regression) gives results closer to a dimension reduction method and if the $\mathcal{L}_1$ norm (LASSO) is chosen as the regularizer, it would be performing a variant of best subset selection. Shrinkage methods, therefore have the advantage of being very generalizeable and flexible. They also have the added advantage of being able to be solved as convex optimization problems for which powerful computer solvers exist.

The Least Squares Shrinkage and Selection Operator (LASSO) was introduced by Tibshirani (1996). As the name suggests, it performs the dual function of shrinkage and selection, whereby some of the coefficients are set to be exactly zero. This means that the LASSO is generally preferred over the ridge regression when model interpretability is valued. In economics, for example, an empiricist may be interested in knowing which variables were more relevant than others. Despite this ability for the LASSO to perform automatic variable selection, it would not be prudent to think that the coefficient estimates it produces have any meaning in the real world except as weights in a predictive model. This complication is addressed in the next section.

The LASSO

The LASSO estimates are obtained by solving the following minimization problem,

$$\widehat{\boldsymbol{\beta}}_l(\lambda) = \arg\min_{\beta} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2/n + \lambda||\boldsymbol{\beta}||_1 \right)$$

where $||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2$ and $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{k}|\beta_j|$ and $\lambda > 0$ is the penalty parameter (Tibshirani, 1996).

A kink in the objective function above sets the weights of the irrelevant covariate to zero, if the underlying model is sparse. The number of estimates chosen will always by less than N even when sparsity is not met, which is why it is necessary

that this assumption is met. The resulting non-zero parameter estimates $\boldsymbol{\beta}_\lambda$ are biased towards zero. Many methods exist to reduce this shrinkage and will be detailed below.

LASSOs can also provide the flexibility of allowing the researcher to leave out some covariates out of the penalization.

The parameter, λ, is generally chosen by cross-validation and is obtained to minimize cross-validation squared error risk or any other measure of quality of prediction. A more exhaustive treatment of cross-validation and associated heuristics is found in Trevor et al. (2009).

LASSO with penalty loadings

This penalized objective function of the LASSO introduces another complication that is best understood in the context of the ordinary least squares (OLS) regression. The coefficients that result from OLS, $\widehat{\beta}_{j,ls}$, are scale invariant, in that, $\widehat{\beta}_{j,ls} \cdot x_j$ do not depend on the scaling of the associated regressor, x_j, or the scaling of the other regressors x_{-j}. However, in a penalized regression, the coefficients depend on the penalty parameter, λ, and therefore tthe scaling of the other regressors. Most practitioners will circumvent this obstacle by, standardizing the regressors before running the LASSO (James et al., 2013).

Belloni, Chen, Chernozhukov, and Hansen (2012) provides an alternative characterization of the LASSO objective function whereby rather than standardizing coefficients *a priori*, penalty loadings are fed into the regularizer.

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\beta} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2/n + \lambda||\hat{\boldsymbol{\Gamma}}_{\boldsymbol{l}}\boldsymbol{\beta}||_1 \right)$$

where $\hat{\boldsymbol{\Gamma}}_l = \text{diag}(\hat{\gamma}_{1l}, \cdots, \hat{\gamma}_{kl})$. These penalty loadings can be chosen to ensure that coefficients are not affected by their scaling, and can be a useful tool to address heteroskedasticity, clustering and non-normality of model errors within the LASSO framework. Belloni et al. (2012) also provide asymptotically valid formulas for the penalty loadings and penalty parameters.

Post-LASSO

As observed above, even the non-zero coefficients selected by the LASSO have a downward bias. To alleviate the bias Belloni et al. (2012) proposes the Post-LASSO estimator which involves first running the LASSO and subsequently identifying the subset of relevant covariates and running OLS on those.

Let $\hat{T}_l$ be the set of all covariates with non-zero coefficients after running the

LASSO, then the Post-LASSO estimator is the

$$\widehat{\boldsymbol{\beta}}_{pl}(\lambda) = \arg\min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{\widehat{T}_l^c}} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2\right)$$

If model selection is perfect, the Post-LASSO estimator is the standard oracle estimator. However, even in the face of less-than-perfect selection the Post-LASSO estimator has less bias (Belloni et al., 2012).

Adaptive Lasso

While cross-validation is an effective method to reduce overfit, empirical tuning tends to select too many covariates. To see this formally, suppose $\mathcal{S}_0$ is the active set of covariates that must be selected, then the set of covariates selected by LASSO through cross-validation, $\widehat{\mathcal{S}}(\hat{\lambda}_{CV})$, contains $\mathcal{S}_0$ (Bühlmann & Van De Geer, 2011).

Zou (2006) proposes the adaptive LASSO to overcome the overestimation problem of the LASSO. He proposes a two-stage estimator where the penalty is re-weighted.

$$\widehat{\boldsymbol{\beta}}_{adapt}(\lambda) = \arg\min_{\beta} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2/n + \lambda \sum_{j=1}^{k} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|} \right)$$

where $\hat{\beta}_{init,j}$ can be obtained by running a LASSO, with $\lambda_{init,CV}$ chosen by cross-validation. The penalty parameter, λ, is also chosen by cross-validation in the second stage. This procedure has the following special features,

- $\hat{\beta}_{init,j} = 0 \implies \hat{\beta}_{adapt,j} = 0$
- If $|\hat{\beta}_{init,j}|$ is large, the adaptive LASSO employs a small penalty (i.e. little shrinkage) for the j^{th} coefficient.

(Bühlmann & Van De Geer, 2011).

These properties imply that the adaptive LASSO tends to yield a sparser solution but induces less bias. Although computationally more demanding, the adaptive LASSO might be more powerful than the Post-LASSO due to its dual purpose of reducing over-selection while managing bias.

LASSOs for Generalized Linear Models

For generalized linear models, the LASSO can be estimated by penalizing the negative log-likelihood function.

$$\widehat{\mu}(\lambda)\,,\, \widehat{\boldsymbol{\beta}}_l(\lambda) = \arg\min_{\mu,\beta} \left(-\frac{1}{n} \sum_{i=1}^{n} \log(p_{\mu,\beta}(Y_i|X_i)) + \lambda||\boldsymbol{\beta}||_1 \right)$$

where $p_{\mu,\beta}(Y_i|X_i)$ is the conditional density of $Y|X = x$ where the μ is the intercept term which is left out of the penalization. All other refinements discussed above apply in the case of a generalized model, and asymptotic theory applies in the same manner of the linear models (Bühlmann & Van De Geer, 2011).

2.3 Machine Learning in Econometrics

Most work in microeconometrics is targeted at parameter inference, in that empiricists would prefer to think their parameter estimates are causally predictive. This focus on recovering causal parameters, has ignored the other key function of econometrics: prediction. Kleinberg et al. (2015) note that many policy issues are prediction problems. However, even when the end goal is parameter estimation, some estimation techniques involve a prediction stage. Even when this is not the case, dimension reduction and automatic model selection are some key challenges modern econometrics face, for which high-dimensional statistical techniques have much to contribute.

2.3.1 High-dimensional statistics in Econometrics

While it comes as no surprise that in disciplines such as genomics and computational biology high-dimensional datasets may be commonplace, its usefulness in econometrics is not immediately obvious. Belloni et al. (2014) observes that high-dimensional settings that arise in economics, are either *truly high-dimensional*, such as with large administrative datasets particularly those merged across various collection centers, or have been *constructed to be high-dimensional* by taking polynomials and interactions between covariates. Much of the attention in this thesis is given to this second type of high-dimensional dataset, but these insights are generalizeable to any setting where $k \gg N$.

Most applied work in economics takes a top-down theoretical approach to deciding what variables must enter an equation. But economic theory does not extend far enough to inform the practitioner about the nature of all the controls that must enter the equation or the form in which it must enter (i.e. linearly, logarithmically, or as a polynomial). Therefore, the practitioner usually makes an *ad hoc* decision about what these variables might be and its functional form. Non-parametric statistics exists exactly due to this uncertainty in the functional form. However, non-parametric methods again require the researcher to make a decision as to what limited subset of variables must be relevant in explaining the dependent variable. They are affected by the 'curse of dimensionality' and interpretation of such models are more computationally demanding than in parametric models.

However, non-parametric models can be estimated in a parametric framework

using well chosen basis functions (polynomials are an example of such basis functions), but to do so would be to end up with over-parameterized model. However, if the assumption of sparsity is plausible, an approximately sparse high-dimensional linear model arises. Then, the estimators discussed in the above section could be used to derive interpretable results from such a model. High-dimensional statistical methods therefore present a powerful way to combine the linear and non-linear parametric models with their non-parametric counterparts.

2.3.2 Prediction vs. Causal Prediction

The LASSO leads to interpretable models due to its ability to perform variable selection by shrinkage. If the downward bias can be rid from these estimates, such as by using an adaptive LASSO, it might be tempting to think that these variables could have more meaning than simply being weights in a predictive model. Unfortunately, the estimates produced or the variables thus selected (due to having non-zero coefficients) are not informative as revealing underlying structure, especially if the penalty parameter, λ, is produced using cross-validation. The LASSO works well when predictive quality is observable, but parametric estimation requires knowledge of the data generating process which we usually do not possess (Mullainathan & Spiess, 2017). These limitations are detailed below.

Instability of model selection

When covariates are correlated they are substitutes in a prediction algorithm. This means very different sets of covariates can lead to similar prediction quality creating an additional form of uncertainty, i.e. model selection uncertainty. Regularizers contribute to this by preferring less complex models. Therefore, even when the LASSO selects variables, these variables cannot be taken to be relevant in the sense of being indicative of underlying structure, since a different set of data from the same data-generating process could select a very different set of variables as being relevant. See Mullainathan and Spiess (2017) for an example of this model selection instability in the context of a house price regression.

This non selection of correlated variables means that the error terms would be highly correlated to the regressors, bringing with it omitted variable bias, a classic cause of biased parameter estimates (Mullainathan & Spiess, 2017). Zhao and Yu (2006) explores the model selection consistency of the LASSO. They find that structure can be recovered asymptotically, as long as sparsity and the 'irrepresentable condition' is met. This condition is the property that the relevant and irrelevant covariates have very little correlation, a condition unlikely to be met in an empirical setting, particularly if the data was generated as a polynomial expansion of a smaller core set of regressors.

True structural parameter is close to zero

While shrinkage is an essential function of the LASSO, this property of the LASSO makes it likely that variables with small but non-zero coefficients will be selected out due to the shrinkage. This again leaves the parameter estimates vulnerable to omitted variable bias.

Complexity of obtaining standard errors

Standard errors of any high-dimensional model needs to account for model selection uncertainty. Obtaining these is no easy feat.

2.3.3 Predictive algorithms for better estimation

The intention of the subsection above was by no means to invalidate the use of high-dimensional methods in causal inference, but to illustrate that it must be done with great care and creativity to do justice to predictive algorithms. Summarised below are some such applications.

Inference with selection among many instruments

Some estimators used commonly in econometrics proceed in several estimation stages. But what is viewed as an estimation stage might actually be a prediction stage, where only the fitted value obtained from that stage enters the subsequent stages. The classic example is the two-stage least squares estimator, where the first stage is aimed at finding the fitted value of the endogenous regressor using instrumental variables. Mullainathan and Spiess (2017) point out that small sample biases of the IV estimator is the result of the ordinary least squares estimator's tendency to overfit. This results in the two-stage least squares estimator is biased towards ordinary least squares estimator.

One of the most challenging areas of applied microeconometrics is finding strong instrumental variables. As pointed above, in the search for such instruments practitioners usually do not consider that perhaps the strongest instruments might enter the first-stage equation in a non-linear fashion. Even when this is done in low-dimensional cases fitting too many instruments can lead to overfitting. This issue is exacerbated when the sample size is small. These issues severely limit the performance of such an extensively used econometric technique. High-dimensional statistical methods offer a solution to both these issues; they allow automatic model selection and manages overfitting even when the sample size is small.

Belloni et al. (2012) use exactly this insight and use a LASSO in the first stage to find the best prediction of the endogenous regressor, since it immunizes the final stage results to model selection mistakes such as the non-selection of a variable due

to having a small coefficient. They apply their method to a hedonic regression in the context of eminent domain (takings law) and find the best instrument for the endogenous regressor (the judicial decision), out of a possible 147, was the 'number of judging panels with one or more members with a Juris Doctor from a public university *squared*', a variable many practitioners would not include *a priori* in a regression.

Double-selection LASSO

Belloni, Chernozhukov, and Hansen (2014) propose a two-stage LASSO, in attempting to identify the treatment effect conditional on several, possibly $k > N$ covariates. Naively, running a LASSO on just the substantive equation leaving the treatment variable out of the $\mathcal{L}_1$-penalization leaves the coefficient vulnerable to omitted variable bias, since the LASSO could drop out variables that are highly correlated to the treatment variable. These variables that are highly correlated to the treatment variable are also the most relevant controls and leaving them out causes bias in the coefficient associated with the treatment variable. Their proposed solution involves running two LASSO regressions, one that explains the treatment variable using the controls and the other which explains the outcome variable in terms of all the candidate controls. Consequently, an ordinary least squares regression is run on the union of the two sets of variables thus selected and of course the treatment variable (unpenalized). Belloni et al. (2014) replicate two well-known econometric papers using this methodology and tests the robustness of their assumptions and conclude that high-dimensional methods can complement the conclusions of the original authors.

Chapter 3

Methodology and Simulations

> Econometrics recognizes that social behaviour is exceedingly complex and that a limited number of variables related together in fairly simple and elegant equations cannot explain the whole of such behaviour
>
> *A Textbook for Econometrics*
> (Klein, 1953)

This chapter begins with a section that motivates the proposed methodology. It then proceeds to describe the methodology in as much detail as necessary, pointing the reader to the coding which is attached in Appendix C. Following this description, simulation results are presented that compare the performance of the penalized regressions with that of the Heckman two-step method (HTS), its Full-Information Maximum Likelihood (FIML) counterpart and the Two-part Models (TPM).

3.1 Motivation

In Chapter 1 the ongoing debate surrounding the HTS as a correction for sample selection was explored, which revealed many caveats with the use of the HTS. The goal of this section is to reinforce these caveats, albeit briefly, and proceed to explain what has been done to alleviate these and discuss the inadequacy of these remedies, especially when the data are high dimensional.

3.1.1 Summary of issues

The goal of this thesis is to enhance the performance of the Heckman two-step method. This choice was made due to its computational flexibility and widespread adoption in empirical work. The above discussion, however, also outlined several other econometric advantages of the HTS. Despite this, it is undeniable that there

are many deficiencies of this method that cannot be addressed by FIML, two-part models or non-parametric estimators. Two of these major issues are;

Collinearity caused by the IMR

Collinearity of the IMR with the design matrix $\mathbf{X}$ affects not only estimation in the LIML (HTS) setting but also in the FIML setting (Leung & Yu, 1996). Many remedies to this problem are discussed in the subsection 3.2.1.

Sensitivity of β to the choice of $\mathbf{Z}$

Many empiricists have noted that estimates of the level equation $\boldsymbol{\beta}$ tend to be wildly sensitive to the specification of $\mathbf{Z}$ (Powers & Rock, 1999; Briggs, 2004; Mroz, 1987). This could be down to the lack of exclusion restrictions used as in Mroz (1987), but even when many specifications were tested with such exclusion restrictions in other studies this lack of robustness of the coefficients prevailed. In the next chapter, this issue is demonstrated with real data for married women in 1976.

3.1.2 What has been done?

Leung and Yu (2000) summarizes many recent remedies advanced to cure collinearity. These are succinctly summarized below;

Ridge regression in the second stage

Ridge regression (penalized regression with a quadratic regularizer) has been used to stabilize estimates when the variables are highly collinear. Ridge regressions like LASSOs tend to introduce bias to the estimates. However, this has not been a popular method among frequentist econometricians due to the complications with standard errors.

Get an increased sample size

While this is a more commonly recommended procedure to deal with collinearity, it is not feasible in many instances since most researchers would already be using the full sample size available to them. Simulation results of Leung and Yu (2000) show that the improvements to collinearity that can be done by increasing the sample size is limited.

Choose more data points from the right side of kink

Recall the discussion in the subsection on IMR. For the collinearity issues to arise two conditions need to be met. By allowing many of the observations to be in

the non-linear part of the curve collinearity can be resolved. However, since most data in economics is observational, this is not feasible in many instances.

Exclusion restrictions and where to find them

The most obvious solution to correcting the collinearity problems seems to be to find exclusion restrictions. However, it is unclear where such restrictions can be found *a priori*. One could appeal to economic theory, but most economic theories would contend that anything that causes participation (selection) must also impact the amount of engagement (level). Failing that, most researchers place *ad hoc* restrictions, but the arguments advanced in defense of such choices are tenuous.

While the HTS can be identified without exclusion restrictions, the semi-parametric estimators introduced above require the use of them. It seems though that having to impose such restrictions without an objective basis in economic theory has potential to distort inference and does not give proper due to what the data has to say.

3.1.3 How can LASSOs help

Both the issues associated with the HTS discussed above were to do with the specification of $\mathbf{Z}$. Note, however, that first step in the HTS is a predictive stage. What is required in this stage is to find the probability that a certain observation given the characteristics $\mathbf{z}$ is selected into the sample. In machine learning literature, this type of problem is called a classification problem, since what is required in this step is to give the best classification of observation into or out of the sample without overfitting. Many practitioners tend to include too many variables which causes these probabilities to be overfit. Regularization and empirical tuning helps to deal with this overfit, analogous to the case with instrumental variables.

LASSOs in a generalized linear framework have many features that make them plausible classification algorithms. As well as providing the best prediction probabilities, LASSOs also set some coefficients exactly to zero enabling variable selection. This property becomes quite useful in isolating exclusion restrictions. Below, the motivation for using LASSOs is defended in the context of the issues identified.

Issue 1 : True exclusion restrictions

As noted above in a situation where $\mathbf{X} = \mathbf{Z}$ are the same, the collinearity problems are likely to arise if condition 1 (all or most points lie on the linear part of the IMR) is also met at the same time even when the selection process and substantive process are caused by different phenomena. This is since OLS/probit do not have a variable selection property and tends to overfit, so automatic variable selection could help when one does not want to impose exclusion restrictions without good reason.

Furthermore, exclusion restrictions need not be a difference in what covariates enter the two equations but in what functional form they enter. To illustrate using the example of the female labour supply function, rather than assuming that the number of children under 6 is an exclusion restriction, one can think of the number of children being relevant in both equations but entering in different functional forms in each equation. But to identify such different functional forms is to run overparametrized models, for which OLS/probit is inherently unsuitable.

It seems that what is required is a statistical method that can select from many possible candidate variables automatically; a task that a LASSO is designed for.

***Issue 2 : Specification of* Z**

The second issue of the sensitivity of $\boldsymbol{\beta}$ to the specification of $\mathbf{Z}$ has not been explicitly addressed in the literature. This paucity of treatment is likely to have been overlooked—not because it is inconsequential—but because testing many specifications may have been cumbersome. However, selection methods such as the LASSO provide practitioners exactly this convenience because regularization allows to test many different specifications and choose the specification that gives the best prediction.

Furthermore, this ability to deal with overparameterized models allows many different functional forms of the variables to be entered into the equation. For instance, many researchers do not include more than cubic polynomials of a select few variables, but it is unlikely that the covariates enter the selection equation in such a simple way. Methods like the LASSO allow researchers to deal with these issues with more ease.

3.2 Methodology

The proposed methodology is described below. The exact implementation details are provided in Appendix C.

1. Run a LASSO in a probit framework
   Notes: The variables should be standardized if penalty loadings are not imposed and the constant must be left out of the penalization. The λ_p must be cross-validated.

$$\widehat{\mu}(\lambda_p)\,,\,\widehat{\boldsymbol{\gamma}}(\lambda_p) = \arg\min_{\mu,\gamma}\left(\frac{1}{n}\sum_{i=1}^{n} -\left[d_i \log\Phi(\mathbf{z}_i\gamma) + (1-d_i)\log\Phi(-\mathbf{z}_i\gamma)\right] + \lambda_p||\boldsymbol{\gamma}||_1\right)$$

1a. (Optional) Run an adaptive LASSO or a post LASSO to reduce the over-selection or bias.

1b. Calculate the IMR based on the $\widehat{\boldsymbol{\gamma}}$ thus obtained.

2. Run a LASSO on the level equation along with the IMR as a regressor.
   Notes: The notes for (1) apply.

$$\widehat{\boldsymbol{\beta}}_l(\lambda_l) = \arg\min_{\beta} \left(||\mathbf{y} - [\mathbf{X}\ IMR][\boldsymbol{\beta}^\top \rho]^\top ||_2^2/n + \lambda_l ||\boldsymbol{\beta}||_1 \right)$$

2a. (Optional) Again an adaptive LASSO or a post- LASSO can be run over the coefficients to address the over-selection and downward bias.

A few qualifications

It is not necessary to run a LASSO in the second stage if the practitioner decides to *a priori* define its specification and this specification is low dimensional. As noted in subsection 2.3.2, LASSOs are not designed for causal prediction. However, in a simple data generating process as below when the covariates are not correlated, LASSOs can consistently distinguish the relevant covariates. Furthermore, as in the empirical exercise in the next chapter, this entire procedure above is used as a predictive stage, for which the LASSO is ideal.

Standard errors are not reported. The process of obtaining them is complicated by the inclusion of a generated regressor and the model selection uncertainty. The usual formula for obtaining them would give standard errors that are too small and would be misleading if reported.

Note in the step (2) the IMR is left in the penalization, since its selection provides an informal test of whether sample selection is a relevant phenomenon, particularly in the face of no standard errors. If one wishes, this could be left out of the penalization.

3.3 Simulations

A simulation study was conducted to compare the performance of the proposed procedure with that of the HTS. The design of the experiment is detailed in Figure 3.1 below.

Choice of values

Note that N was chosen so as not to be a true high-dimensional setting to enable comparison with the HTS, which cannot be used when $k > N$. N was chosen to be $N \approx 2k$ in the first stage so that after censoring (which is about 50% in each sample) we would still have $N > k$ in the second step.

$$N = 1650 \qquad K = 755 \qquad \#\text{ of sim.s=1083}$$

$$\begin{bmatrix} \upsilon_i \\ \epsilon_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 1 \end{bmatrix} \right)$$

$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$$

$$y_i^* = 0.2x_{1i} + 0.3x_{2i} + 0.4x_{3i} + 0.3x_{4i} - 0.4x_{5i} + 0 \cdot x_{6i} + \cdots + 0 \cdot x_{755i} + \epsilon_i$$

$$y_i = 0.5x_{1i} + 0.6x_{2i} + 0.2x_{3i} + 0 \cdot x_{4i} + \cdots + 0 \cdot x_{755i} + \upsilon_i$$

Figure 3.1: **Simulation design**. *The second line gives the distribution of the error terms. The first equation is the selection equation and the last equation is the substantive equation*

Also the regressors were generated as standard normals to simplify the algorithm, since otherwise penalty loadings need to be imposed in the penalization. The constant was not included for the same reason.

The regressors $\mathbf{x}_4$ and $\mathbf{x}_5$ are generated to be exclusion restrictions since they have a zero coefficient in the substantive equation. But since it is assumed that it is not known *a priori* what these restrictions are all covariates are included in the regression.

Computational details

Note that the LASSOs in both stages can be reformulated to have globally convex optimization functions. The CVX package was used in MATLAB for this, since MATLAB's inbuilt optimizers took too long to converge to the minimum.

Cross-validation details

Recall from section 2.1 of last chapter that the number of folds of cross-validation depends on the sample size. Since the sample size is 1650, the number of cross-validation folds T was set to 5 which leaves a training sample of 1320 in the first stage and 640 in the second stage within the cross-validation exercise.

Ideally, the shrinkage parameter in each stage needs to be cross-validated in each sample. However, this proved to be computationally expensive for 755 variables therefore it was decided to cross-validate for the first 100 samples in a pilot. Then the median of $\widehat{\lambda_p}$ and $\widehat{\lambda_l}$ were used for all of the 1083 samples. Although not ideal, the results for the first 100 were comparable to when they were cross-validated for each sample.

3.3.1 Results

Here simulation results for each step of the HTS are followed by the results of the corresponding step in the proposed penalized estimator. Each of the four sets of results include a table reporting the mean, median and standard deviation of the coefficient associated with the regressors that are relevant and the coefficients associated with a few noise variables $\mathbf{x}_6$, $\mathbf{x}_7$ and every 100^{th} up to $\mathbf{x}_{400}$. In the second stage results, the coefficient of the inverse Milll's ratio is reported as well. Each table is then followed by the sampling distributions of a few relevant and noise covariates. These results are then followed by heatmaps of the coefficients. Since pictures are worth a thousand words the discussion in this section will be kept concise, only mentioning the key points and directing the reader to the relevant images when needed.

Three main conclusions can be gleaned from these results;

Selection of noise covariates

Contrast figure 3.6 with that of 3.7 and notice the stark difference in variable selection. First generation methods tend to spuriously select noise covariates, while the LASSO almost always selects these out as seen from the large white space in the heatmap. The sampling distribution is yet another way this information can be gleaned. Contrast the plot (c) in Figures 3.4 and 3.5 which depicts the distribution of a noise covariate. In the LASSO this sampling distribution approximates a degenerate distribution around a mean of -0.0003 which means its likelihood of selection is extremely low. Information in Figure 3.4 and 3.6 imply that with OLS, there is very high likelihood that noise covariates are selected.

In terms of statistical significance, taking the standard deviations in Table 3.3 as approximations of the standard error of the coefficient it seems likely that if the coefficient of $\mathbf{x}_k$ is greater than twice the standard deviation it will be significant. For example, the standard deviation of β_4 is 0.1157, which means that any coefficient whose absolute value is above 0.2314 will be deemed significant. In fact, in 50 of the simulations that absolute value of the coefficient was above 0.2314.

The asymptotic standard errors of the HTS estimates were derived for each β_k (results not reported). The mean of these standard errors were about 0.0774 quite

similar to that of the Monte-Carlo standard error (i.e. standard deviation of β_4) and here in 345 counts the variable was deemed significant.

Noise in the selection of relevant covariates

Notice that OLS seems to be unable to consistently select $\mathbf{x}_3$ in Figure 3.6. The 3^{rd} row of the figure, does not look very dissimilar to many rows below it. Light patches populate even the 1^{st} and 2^{nd} rows quite frequently. In contrast, the LASSO has correctly identified the relevant covariates consistently despite the shrinkage (a low coefficient was associated with $\mathbf{x}_{3i}$ in the simulation, i.e. $\beta_3^0 = 0.2$, yet it is selected almost always).

The mean of the standard errors of β_3 was 0.0774, and this variable was deemed insignificant in 328 of the simulations.

The behaviour of the IMR

Perhaps the most interesting result is that of the calculation of the IMR. The true coefficient of the IMR is $\beta_{756}^0 = 0.4$. The mean of HTS estimates of this coefficient is 3.0890 a severe positive bias and an associated large standard deviation. In contrast, in the penalized regression this same coefficient is estimated as 0.2929 where the downward bias is as expected. The sampling distributions are given in the plot (d) in Figures 3.4 and 3.5. In the unpenalized OLS, this distribution is highly right skewed and does not approximate a z-distribution. Its large standard deviation of 4.4113, if taken as a proxy for the standard error would mean that the IMR will be deemed insignificant in over 90% of the simulation samples. Less strikingly, the mean of the standard error of β_{756} is 1.824 and it was deemed insignificant on 317 occasions.

It is also interesting to note that in the case of the penalized regression the coefficient of the IMR was not adversely affected by the inclusion of $\mathbf{x}_4$ and $\mathbf{x}_5$ in the second stage. Due to time constraints, it was not explored what the effect of an *a priori* imposition of exclusion restrictions might have been on the estimate associated with the IMR.

This final result of the IMR coefficient is significant, since it provides a plausible explanation as to why most empirical studies in sample selection find the IMR to be insignificant. The overfitting behaviour of the probit model feeds into the estimation of the coefficient associated with the IMR and its associated standard error resulting in an insignificant conclusion.

First-stage - HTS

$$\widehat{\boldsymbol{\gamma}} = \arg\max_{\gamma} \left(\frac{1}{n} \sum_{i=1}^{n} d_i \log \Phi(\mathbf{z}_i \gamma) + (1 - d_i) \log \Phi(-\mathbf{z}_i \gamma) \right)$$

$\boldsymbol{\gamma}$	Mean	Median	Std. Dev.
γ_1	2.275	1.452	1.697
γ_2	3.384	2.047	2.332
γ_3	4.529	2.669	3.049
γ_4	3.429	2.063	2.375
γ_5	-4.561	-2.697	3.093
γ_6	0.013	0.014	0.788
γ_7	-0.003	0.007	0.750
γ_{100}	-0.001	0.016	0.779
γ_{200}	0.013	0.007	0.737
γ_{300}	0.035	0.024	0.742
γ_{400}	0.014	0.006	0.749
γ_{500}	-0.005	-0.024	0.727

Table 3.1: **Probit model coefficients**. *The first 5 regressors were given a non-zero coefficient and all others are irrelevant. The mean, median and standard deviation is reported for all these variables*

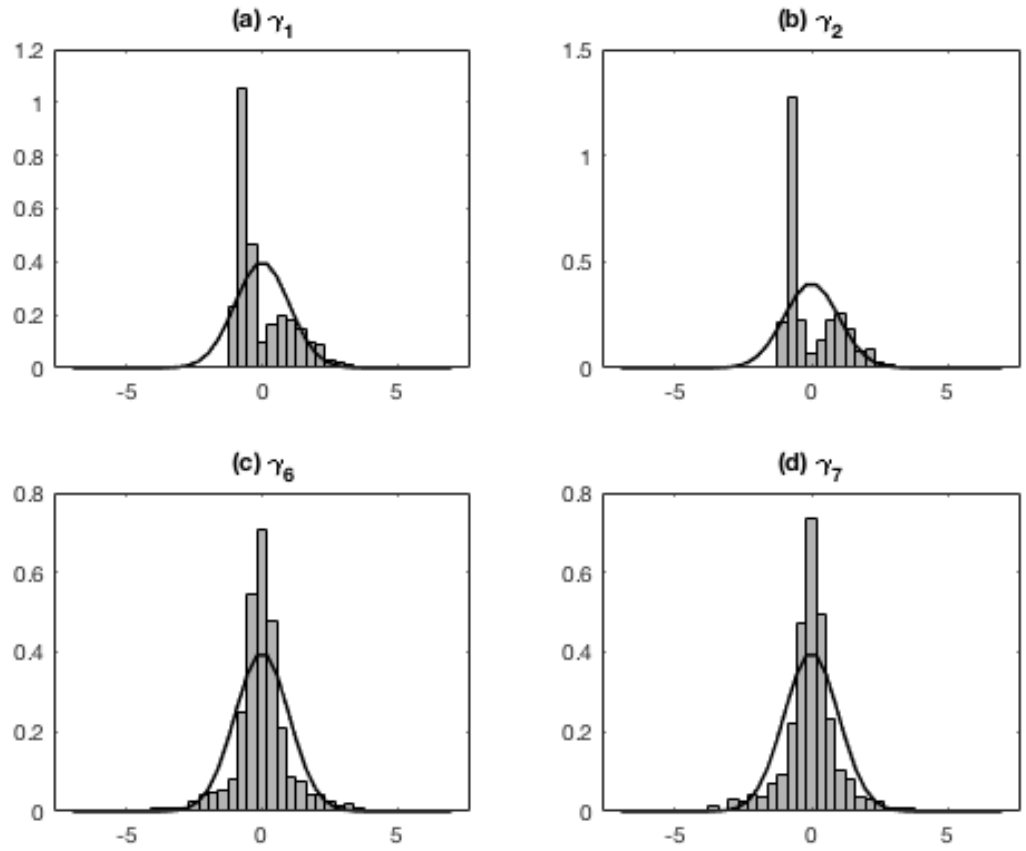

Figure 3.2: **Probit - Sampling distribution of γ's.** *All coefficients were studentized and centered on the mean and a standard normal distribution was superimposed as a comparison. The top panel (a) and (b) are coefficients initially generated to be non-zero and the bottom panel (c) and (d) were generated as irrelevant.*

First-stage penalized probit

$$\widehat{\boldsymbol{\gamma}}(\lambda_p) = \arg\min_{\gamma} \left(\frac{1}{n} \sum_{i=1}^{n} - [d_i \log \Phi(\mathbf{z}_i \gamma) + (1 - d_i) \log \Phi(-\mathbf{z}_i \gamma)] + \lambda_p ||\boldsymbol{\gamma}||_1 \right)$$

$\boldsymbol{\gamma}$	Mean	Median	Std. Dev.
γ_1	0.0923	0.0919	0.0332
γ_2	0.1881	0.1890	0.0314
γ_3	0.2836	0.2823	0.0335
γ_4	0.1895	0.1900	0.0319
γ_5	-0.2849	-0.2853	0.0324
γ_6	0.0000	0.0000	0.0009
γ_7	0.0000	0.0000	0.0006
γ_{100}	0.0000	0.0000	0.0011
γ_{200}	0.0000	0.0000	0.0004
γ_{300}	0.0000	0.0000	0.0009
γ_{400}	0.0000	0.0000	0.0000
γ_{500}	0.0000	0.0000	0.0003

Table 3.2: **Penalized Probit coefficients**. *The first 5 regressors were given a non-zero coefficient and all others are irrelevant. The mean, median and standard deviation is reported for all these variables*

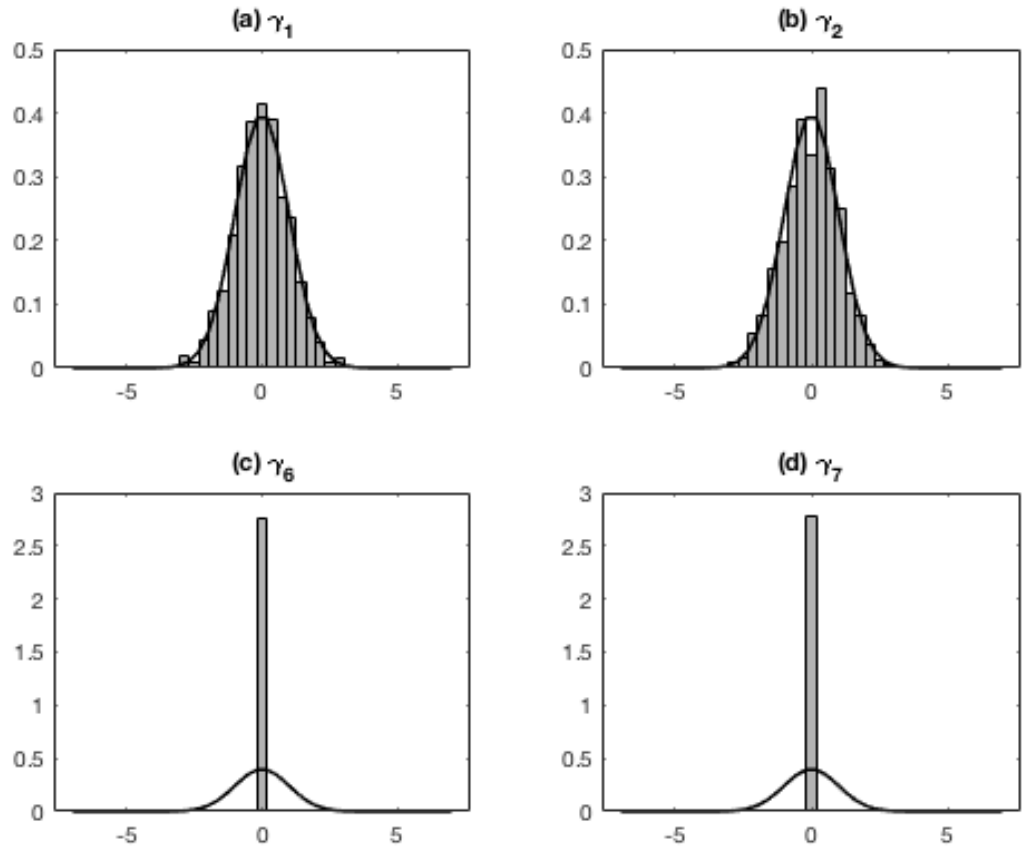

Figure 3.3: **Penalized probit- Sampling distribution of γ's.** *All coefficients were studentized and centered around the mean and a standard normal distribution was superimposed as a comparison. The top panel (a) and (b) are coefficients initially generated to be non-zero and the bottom panel (c) and (d) were generated as irrelevant.*

Second stage - HTS

$$\widehat{\boldsymbol{\beta}}_l = \arg\min_{\beta} \left(||\mathbf{y} - [\mathbf{X}\ IMR][\boldsymbol{\beta}^\top \rho]^\top ||_2^2 / n \right)$$

$\boldsymbol{\beta}$	Mean	Median	Std. Dev.
β_1	0.5017	0.5076	0.1165
β_2	0.5997	0.6005	0.1146
β_3	0.2035	0.2060	0.1115
β_4	-0.0017	-0.0039	0.1157
β_5	-0.0017	0.0002	0.1140
β_6	0.0010	0.0034	0.1129
β_7	0.0021	0.0047	0.1191
β_{100}	0.0063	0.0059	0.1127
β_{200}	0.0039	0.0041	0.1141
β_{300}	0.0052	0.0035	0.1147
β_{400}	0.0015	0.0054	0.1098
β_{756}	3.0890	0.7292	4.4113

Table 3.3: **Unpenalized second step OLS coefficients**. *The first 3 regressors were given a non-zero coefficient and all others are irrelevant.* $\mathbf{x}_4$ *and* $\mathbf{x}_5$ *were generated as exclusion restrictions. The variable* $\mathbf{x}_{756}$ *is the Inverse Mills' Ratio.*

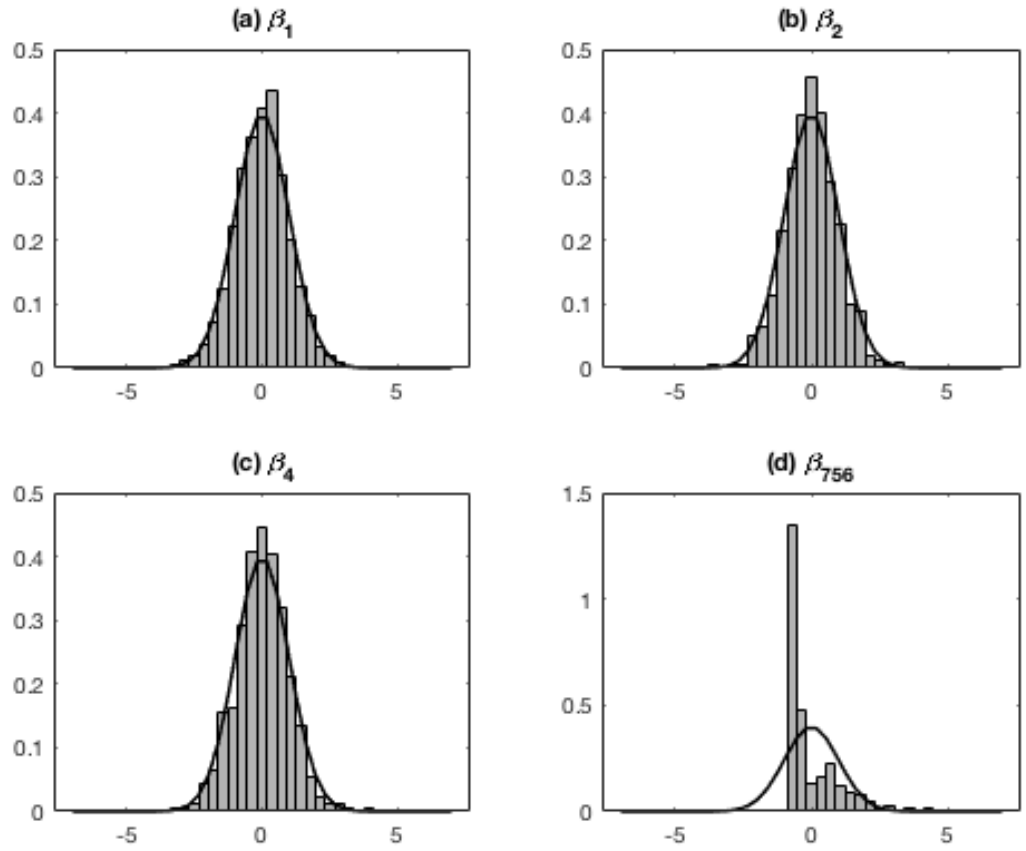

Figure 3.4: **Unpenalized OLS- Sampling distribution of β's.** *All coefficients were studentized and centered around the mean and a standard normal distribution was superimposed as a comparison. The top panel (a) and (b) are coefficients initially generated to be non-zero and the bottom panel (c) was generated as being irrelevant and (d) is the sampling distribution of the inverse Mills' ratio.*

Second stage - Penalized

$$\widehat{\boldsymbol{\beta}}_l(\lambda_l) = \arg\min_{\beta} \left(||\mathbf{y} - [\mathbf{X}\ IMR][\boldsymbol{\beta}^\top \rho]^\top ||_2^2 / n + \lambda_l ||\boldsymbol{\beta}||_1 \right)$$

$\boldsymbol{\beta}$	Mean	Median	Std. Dev.
β_1	0.4154	0.4158	0.0300
β_2	0.5171	0.5160	0.0300
β_3	0.1180	0.1168	0.0323
β_4	-0.0003	0.0000	0.0031
β_5	0.0003	0.0000	0.0028
β_6	0.0000	0.0000	0.0013
β_7	0.0000	0.0000	0.0030
β_{100}	0.0000	0.0000	0.0017
β_{200}	0.0000	0.0000	0.0016
β_{300}	0.0000	0.0000	0.0021
β_{400}	-0.0001	0.0000	0.0012
β_{756}	0.2929	0.2928	0.0410

Table 3.4: **LASSO second stage coefficients**. *The first 3 regressors were given a non-zero coefficient and all others are irrelevant.* $\mathbf{x}_4$ *and* $\mathbf{x}_5$ *were generated as exclusion restrictions. The variable* $\mathbf{x}_{756}$ *is the Inverse Mills' Ratio.*

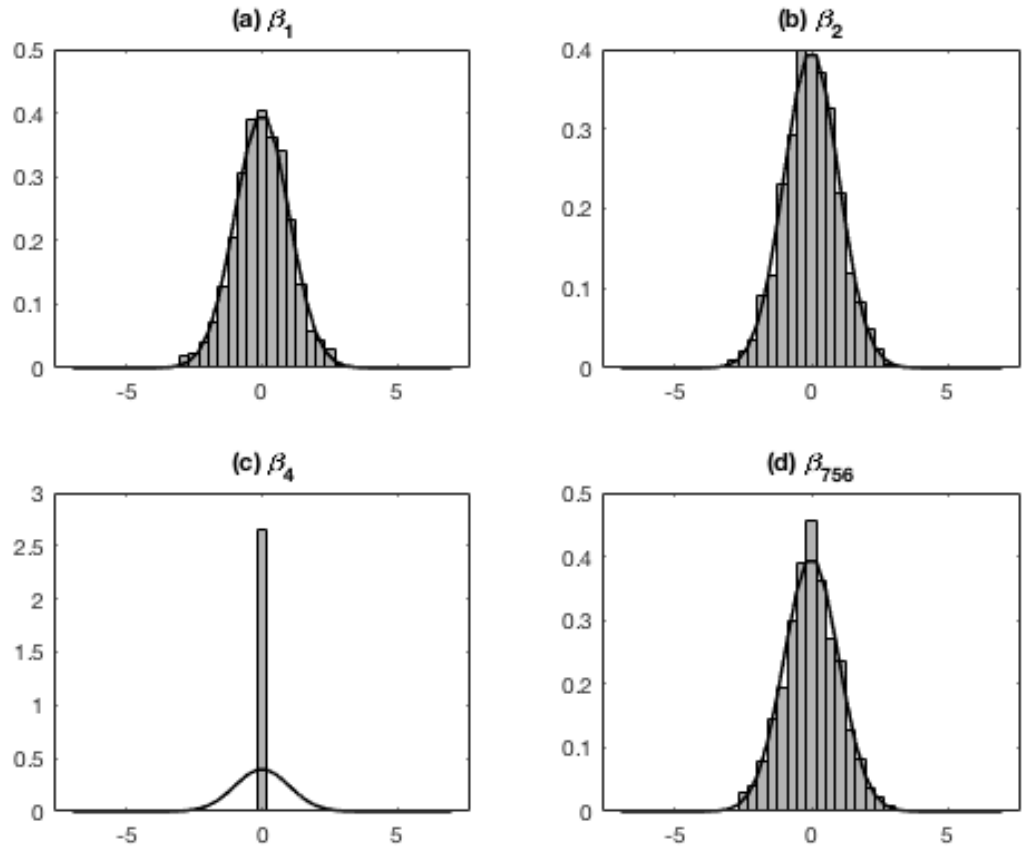

Figure 3.5: **Penalized OLS- Sampling distribution of β's.** *All coefficients were studentized and centered around the mean and a standard normal distribution was superimposed as a comparison. The top panel (a) and (b) are coefficients initially generated to be non-zero and the bottom panel (c) was generated as being irrelevant and (d) is the sampling distribution of the inverse Mills' ratio.*

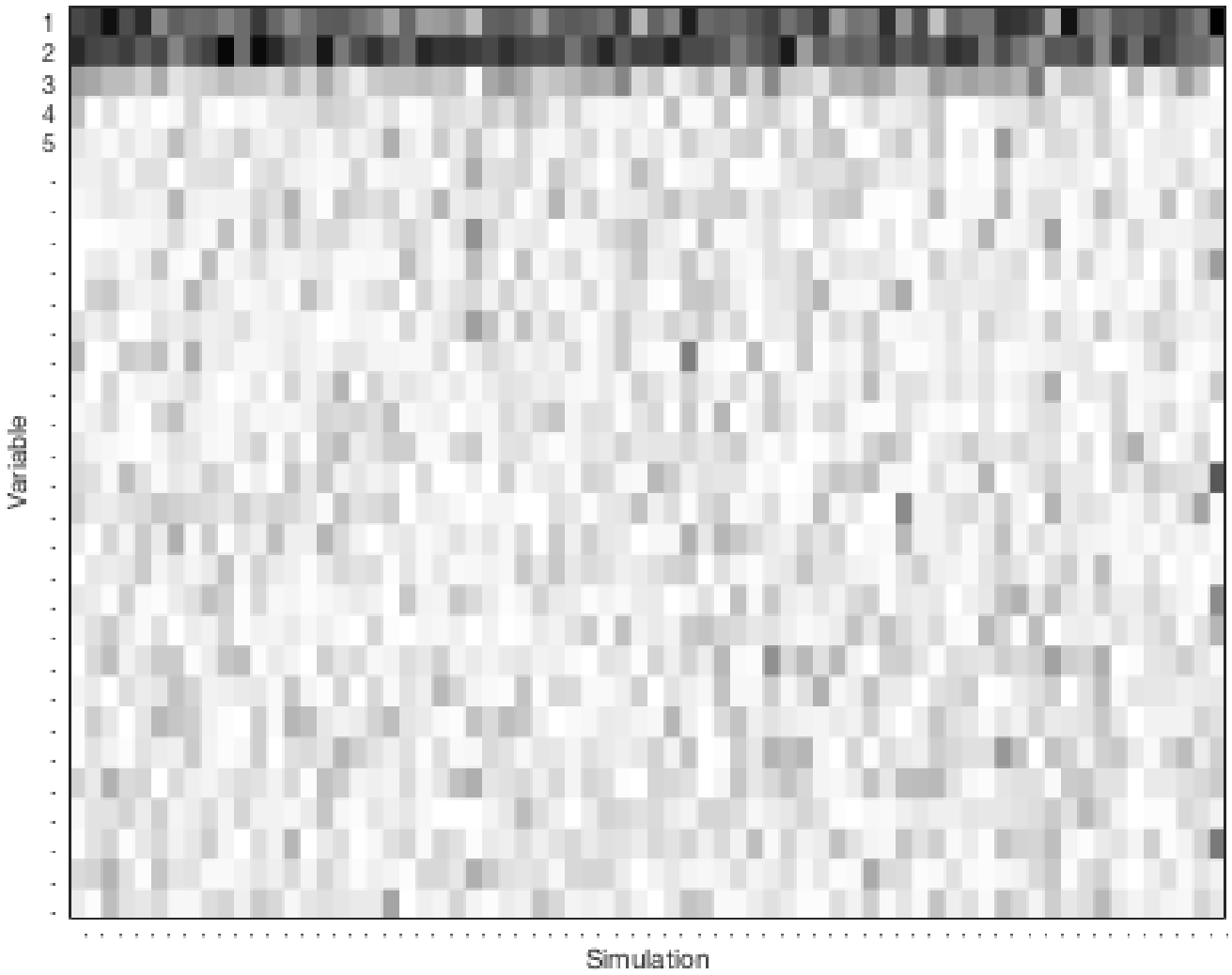

Figure 3.6: **HTS - Second stage variable selection** *Darker shades represent selection, therefore lighter shades represent smaller coefficients. The horizontal axis represents the simulation and the vertical axis represents the variables. Only the first 70 simulations for the first 30 variables are in the heatmap and the tick labels are obscured for easier reading.*

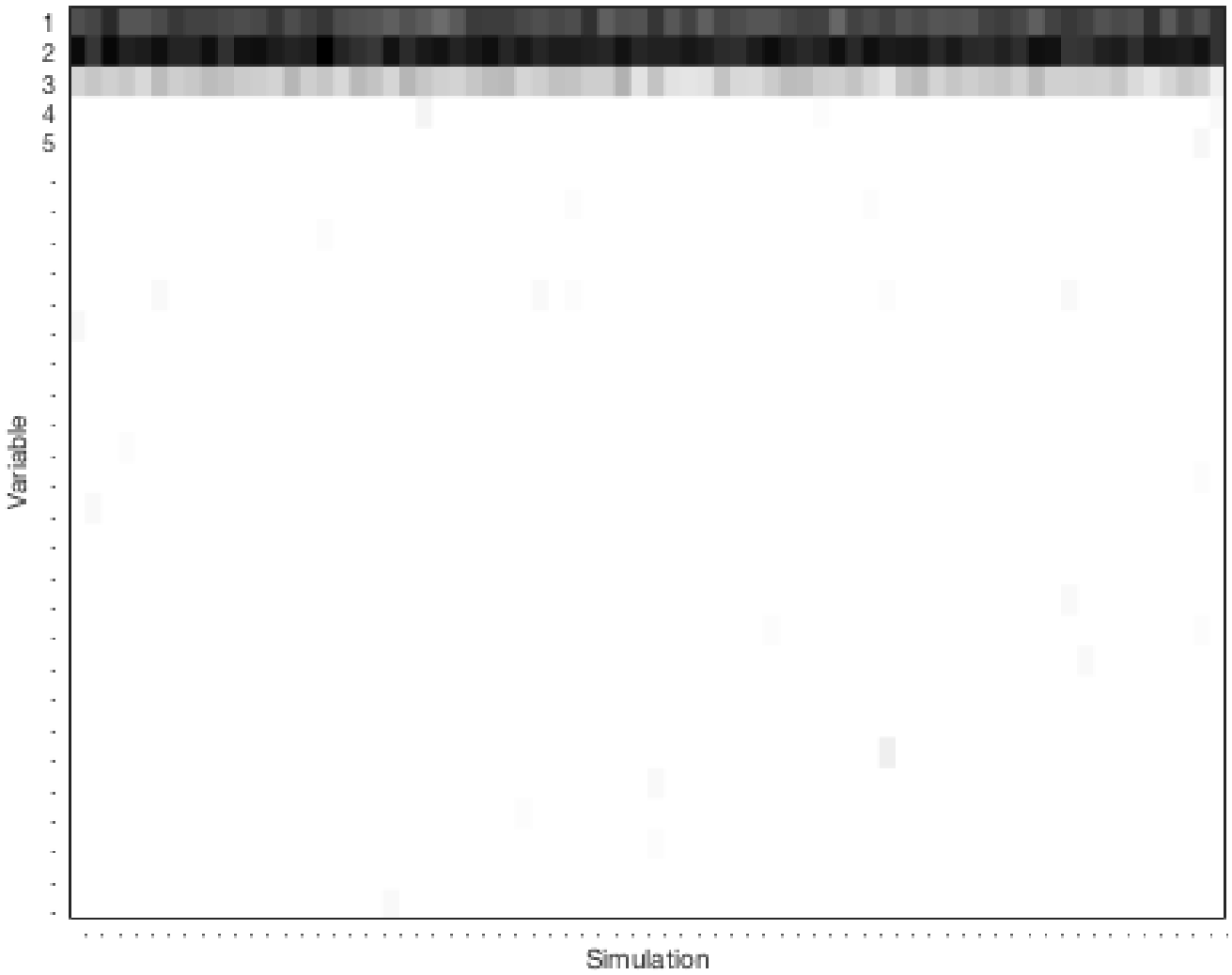

Figure 3.7: **Penalized - Second stage variable selection** *Darker shades represent selection, therefore lighter shades represent smaller coefficients. The horizontal axis represents the simulation and the vertical axis represents the variables. Only the first 70 simulations for the first 30 variables are in the heatmap and the tick labels are obscured for easier reading.*

Chapter 4

Empirical Application

4.1 Mroz (1987)

This chapter applies the methodology developed in the section above to an empirical setting. The empirical setting studied is the labour supply of women in the workforce. The dataset used in Mroz (1987) is used to illustrate the methodology. It has been used extensively in the literature of sample selection models, including semi-parametric and non-parametric applications allowing econometricians to compare the robustness of the estimates to changes in the specification. It is not the intention to review this study in any complete way, but rather to observe how the coefficients change as the methods proposed in this thesis are used. Therefore, Mroz's choice of instruments is not challenged and where possible his specification of the design matrix is retained.

In particular, the focus is to explain a particular confusing outcome of Mroz's study. To test for sample selection, Mroz (1987) defines many specifications of the selection equation, one of which includes experience and one of which does not. However, while the estimates when experience was included in the first stage favoured sample selection, the estimates when experience was not included did not reject the null hypothesis of no sample selection. Mroz (1987) conceded he did not have a plausible explanation. Leung and Yu (1996) conjecture that this was a collinearity issue. The proposed methodology is used to determine if this was correctly hypothesized.

4.1.1 Set-up

Data

The data comes from the University of Michigan Panel Study of Income Dynamics (PSID) for the interview year 1976. This year was the only year the wives were interviewed, in contrast to other years when only the household head was interviewed. There are 753 observations in the sample, and 428 in the truncated sample.

There are 14 covariates that could be used as regressors in the dataset, of which one is a dummy variable.

Model

Mroz (1987) tests a multiplicity of specifications across three broad sections. Here only a select few of estimates of Table VI and X are reproduced and replicated. The sample selection model of Mroz (1987) proceeds as follows.

$$LFP_i = f(\mathbf{z}_i\boldsymbol{\gamma}) + \epsilon_i \tag{4.1}$$

$$\log w_i = \mathbf{x}_{1i}\boldsymbol{\beta} + \upsilon_{1i} \tag{4.2}$$

$$h_i = \mathbf{x}_{2i}\boldsymbol{\beta} + \upsilon_{2i} \tag{4.3}$$

where LFP_i stands for whether the woman is working or not, w_i is the calculated wage rate, and h_i are the number of hours. Mroz (1987) specifies the $\mathbf{z}_{1i}$ to be a vector that includes number of young children, number of older children, the education of the wife's mother and father, polynomials in the wife's age and education, dummy variable for living in the city and the wife's labour market experience and its square. The specification of $\mathbf{x}_{1i}$ are the same variables above plus the inverse Mill's ratio calculated in stage (4.1). The contents of $\mathbf{x}_{2i}$ included the predicted log wage from regression (4.2), the non-wife income, number of older and younger children, the age and education of the wife and the inverse Mills' ratio from (4.1).

Mroz (1987) instruments the $\log w_i$ since he suspects if it is measured with error it would be negatively correlated with hours, because of the way the wage is constructed (i.e wage is the wife's total earnings divided by wife's total hours). In another specification of the model he treats experience as endogenous and therefore does not include it in the specification of $\mathbf{z}_{1i}i$ or $\mathbf{x}_{1i}$.

Constructed data

High-dimensionality is achieved by taking a 3^{rd} degree polynomial expansion of the 13 non-dummy regressors and in addition interacting these with the dummy. This results in 663 covariates, which is more than the 428 observations in the uncensored sample. For (4.1) and (4.2) all these covariates are used, but Mroz's specification is retained for (4.3) for comparison with other specifications. This design matrix will be referred to as $\mathbf{Z}_2$ to contrast it from Mroz's specification (for the i^{th} individual this will be $\mathbf{z}_{2i}$).

4.1.2 Replication, methods and results

Notice that in the multi-stage set up of Mroz (1987), the first two equations are both predictive stages. In fact, (4.1) together with (4.2) is a complete Heckman

two-step model used only as a means for prediction.

Before applying the full model, each specification to its lead up is scrutinized by applying a high-dimensional counterpart.

LASSO for instruments

Table 4.1 reports the results when a LASSO is implemented in the first stage of the 2SLS with design matrix $\mathbf{Z}_2$, in a set-up similar to that of Belloni et al. (2012). Out of these 664 covariates only five were selected as having a non-zero coefficient. This was the cubed term of education, the squared term of education interacted with previous labour force experience, and a few interactions of wife's education and experience. While it is difficult to draw any inferences about the causal effect these coefficients might have on the wage function of a woman since this is not the purpose of high-dimensional methods, it is not difficult to imagine that some functional form of education and experience will affect wages. The Appendix details the exact form of these variables.

Using the predicted wage from the LASSO in place of the wage in the regression almost doubles the coefficient associated with log wage leaving other coefficients almost unchanged. The associated standard errors are provided in the Appendix C.

Calculation of the Inverse Mills' Ratio

The LASSO in a probit framework chose 52 covariates out of 664 as being non-zero. Since this is symptomatic of the overfitting behaviour of the LASSO, an adaptive LASSO was also run using the $\hat{\gamma}$ as the initial coefficients, which gave 15 covariates a non-zero coefficient. The inverse Mills' ratio thus generated will be referred to as LIMR hereafter.

These results of Table 4.2 reaffirm the observation of Leung and Yu (1996) that the inclusion of experience greatly reduces collinearity and causes most values of the IMR to be in the non-linear area of the curve. Recall that one of the conditions for collinearity as suggested by Leung and Yu is that almost all points of $\mathbf{Z}\widehat{\gamma}$ lie on the right of the kink. The figure 4.1 shows clearly that when experience is not included in the specification of $\mathbf{Z}$ most points lie before the kink. Plot (d) corresponds to the fitted values of the probit-LASSO, where the distribution is much wider and many points lie to the right of the kink.

The inclusion of IMR_h does not greatly improve collinearity and gives similar estimates to column (1). This could be because the specification of $\mathbf{Z}$ that Mroz (1987) used and the covariates selected by the LASSO have similar predictive quality, even when they are entirely different variables. However, it is unlikely that the empiricist could pick the right specification *a priori*. Therefore, formal variable selection methods can always be used to complement this intuition.

Variable	Baseline - IV (1)	2SLS - X (6)	2SLS - LASSO
log wage	-17.41	672.30	1202.25
non-wife income	-4.25	-6.45	-6.75
# kids < 6	-342.50	-284.37	-326.95
# kids 6-18	-115.02	-85.24	-91.35
wife's age	-7.73	-9.08	-13.48
wife's education	-14.44	-86.41	-113.83
constant	2114.70	2254.87	2175.97
Condition number	26.047	28.524	40.551

Table 4.1: **Baseline and 2SLS (assume no sample selection)**. *The first column replicates the first set of estimates of Mroz (1987) without any sample selection correction or instruments. The second column replicates the* 6^{th} *set of estimates from Table X of Mroz (1987)which instruments log wage with* $\mathbf{z}_i$*. The third column provides the estimates when a LASSO is used to predict log wage. The condition number is a measure of collinearity of the design matrix where a high number implies more collinearity. The last row reports the condition numbers for each design matrix.*

Variable	(1)	(2)	(3)
log wage	-71.21127043	-17.766855	-74.65711902
non-wife income	5.267443969	-0.344129111	2.881705836
# kids <6	82.54360952	-171.718656	6.714251185
# kids 6-18	-93.63735896	-108.6229936	-97.1172127
wife's age	5.084425168	-1.768345761	3.445831496
wife's education	-72.20336637	-43.20860473	-66.78392274
constant	2523.813187	2307.597804	2589.345899
IMR/LIMR	-811.0103078	-312.4619385	-839.0068123*
Condition number	28.15190663	53.31401463	28.2464708

Table 4.2: **Sample selection corrections in (4.3)**. *The first column and second column shows what the estimates of equation (4.3) with the IMR calculated with experience and without experience (also its squared term) respectively.* * *means the third column uses* $LIMR$. *The condition number is a measure of collinearity of the design matrix where a high number implies more collinearity. The last row reports the condition numbers for each design matrix.*

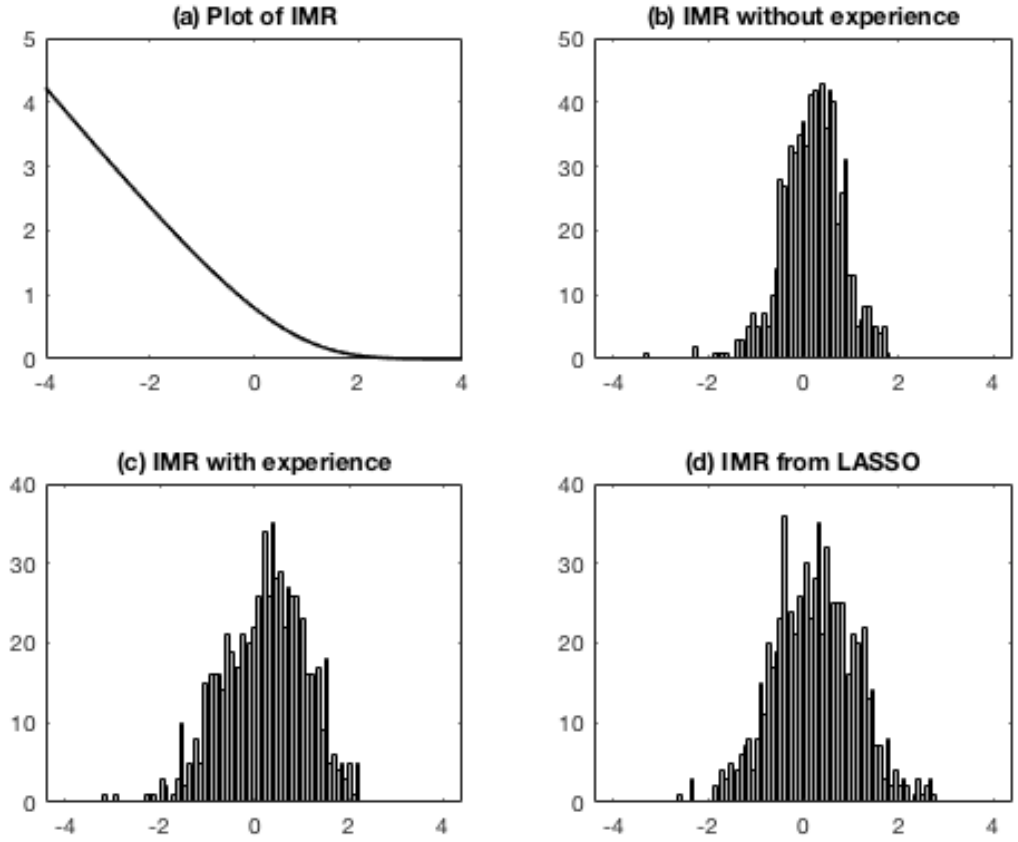

Figure 4.1: **Distribution of $\mathbf{Z}\widehat{\gamma}$** *Plot (a) reproduces the IMR over the range (-4,4) and plots (b) -(d) are histograms of based on various specifications of* $\mathbf{Z}$.

Predicted log wage and inverse mills ratio

Here the entire model detailed above was estimated with LASSOs and compared to the estimates of Mroz (1987). Interestingly in a LASSO regression of log wage on 665 covariates only seven variables had a non-zero coefficient, one of which was the LIMR. The results of this regression is reported in Table 4.3.

The inclusion of log wage predicted from the LASSO changes the coefficient associated in log wage greatly. The small positive coefficient has now become a large negative coefficient. The condition number has increased for this regression and is possibly indicative of multi-collinearity. However, in contrast to the column 2 of the same table where a large condition number is associated with a small coefficient of the IMR, in column 5 the converse happens.

4.1.3 Key Insights and Limitations

It was interesting to note the when the same 664 covariates were used in both (4.1) and (4.2) to estimate a penalized Heckman two-step, their was very little overlap between the covariates selected although they were all certain functions of

Variable	HTS - X (3)	HTS- IX (3)	HTS - $\mathbf{Z}_2$	HTS - $\mathbf{Z}_2$
log wage	126.5141585	63.13602011	102.1567156	-439.7860077
non-wife income	3.627552299	-0.4946964438	1.65778783	5.223809979
# kids <6	52.23439227	-160.4561569	-11.91254026	41.75327326
# kids 6-18	-87.47672669	-104.9451722	-91.41130163	-121.922687
wife's age	3.25850198	-1.755594255	2.052371047	5.851378901
wife's education	-86.06774412	-52.44371211	-79.71041358	-40.22743094
constant	2527.312797	2330.176047	2579.227847	2649.918165
IMR/LIMR	-730.743426	-322.3757792	-759.5495362*	-996.2198729*
Condition number	30.42328419	56.1180136	30.50249056	45.02166735

Table 4.3: **Sample selection corrections and predicted log wage**. *The first column and second column shows what the estimates of equation (4.3) with the IMR calculated with experience and without experience (also its squared term) respectively. * imply the use of LIMR. The third column uses LIMR but the fitted log wage still uses the specification* $\mathbf{Z}$*. The fourth column uses the fitted log wage derived from the LASSO along with LIMR. The condition number is a measure of collinearity of the design matrix where a high number implies more collinearity. The last row reports the condition numbers for each design matrix.*

the 14 covariates initially used as a basis to construct the 664. This is affirmation of the hypothesis of automatic exclusion restrictions detailed in Chapter 3.

The above results reaffirmed the suspicion of Leung and Yu (1996) that the insignificance was indeed caused by collinearity issues. Therefore, it supports the use of high-dimensional methods and/or formal variable selection as a complement in alleviating the specification and collinearity issues.

This exercise also reveals many challenges in the use of high-dimensional methods empirically. The difficulty in obtaining standard errors limits the usefulness of these techniques to any practitioner that wants to perform significance testing. Furthermore, it is computationally challenging. For example, finding the correct shrinkage parameter in a LASSO involves cross-validating on a finer and finer grid which could

take hours to days depending on the complexity of the algorithm. Lastly, while these methods can help the data speak louder, they are useful insofar as the data has much to say.

Chapter 5

Conclusion

Machine learning is here to stay. This implies that practitioners must be able to evolve with these changes. Econometricians working with predictive models have incorporated these algorithms in many innovative ways and evidently these are creeping into work of empirical microeconomists. They provide a different way to view estimation and provides empiricists means to deal with issues they have only acknowledged as caveats, such as functional forms and overfitting. This thesis has taken a small leap in this direction in enhancing the capabilities of the models that empricists had traditionally worked with.

These developments, while opening new doors for econometricians poses new challenges. Principal among them is that of proving consistency of the estimates used and obtaining standard errors. The simulation results of a very simple model has shown that the resulting estimators look very close to standard normal, particularly the sampling distribution of the sampling distribution of the coefficient of the inverse Mills' ratio. While calculating standard errors are difficult, this can certainly be done and is an exciting area of research in econometrics (Belloni et al., 2012).

There were also many different directions in which one could have progressed with sample selection models. Zadrozny (2004) detail an excellent application of machine learning algorithms other than LASSOs that deal with sample selection bias. Bayesian econometricians feed shrinkage into priors and some have recently been working with priors such as the 'double exponential' that perform variable selection (De Mol, Giannone, & Reichlin, 2008).

References

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

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

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 386–398.

Briggs, D. C. (2004). Causal inference and the heckman model. *Journal of Educational and Behavioral Statistics*, *29*(4), 397–420.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bushway, S., Johnson, B. D., & Slocum, L. A. (2007). Is the magic still there? the use of the heckman two-step correction for selection bias in criminology. *Journal of quantitative criminology*, *23*(2), 151–178.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*(1), 33–58.

Davidson, R., MacKinnon, J. G., et al. (1993). Estimation and inference in econometrics. *OUP Catalogue*.

Deaton, A., & Irish, M. (1984). Statistical models for zero expenditures in household budgets. *Journal of Public Economics*, *23*(1-2), 59–80.

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*(2), 318–328.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *Journal of political Economy*, *82*(6), 1119–1143.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica: journal of the econometric society*, 679–694.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475–492). NBER.

Heckman, J. (1977). *Sample selection bias as a specification error (with an application to the estimation of labor supply functions).* National Bureau of Economic Research Cambridge, Mass., USA.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Klein, L. R. (1953). Textbook of econometrics.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–95.

Leung, S. F., & Yu, S. (1996). On the choice between sample selection and two-part models. *Journal of econometrics*, *72*(1-2), 197–229.

Leung, S. F., & Yu, S. (2000). Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Computational Economics*, *15*(3), 173–199.

Manning, W. G., Duan, N., & Rogers, W. H. (1987). Monte carlo evidence on the choice between sample selection and two-part models. *Journal of econometrics*, *35*(1), 59–82.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, 765–799.

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

Newey, W. K., Powell, J. L., & Walker, J. R. (1990). Semiparametric estimation of selection models: some empirical results. *The american economic review*, *80*(2), 324–328.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on sat i: Reasoning test scores. *Journal of Educational Measurement*, *36*(2), 93–118.

Puhani, P. (2000). The heckman correction for sample selection and its critique. *Journal of economic surveys*, *14*(1), 53–68.

Stapleton, D. C., & Young, D. J. (1984). Censored normal regression with measurement error on the dependent variable. *Econometrica: Journal of the Econometric Society*, 737–760.

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

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24–36.

Trevor, H., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* New York, NY: Springer.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual review of sociology*, *18*(1), 327–350.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on machine learning* (p. 114).

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, *7*(Nov), 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Appendix A

Truncated Moments of the Standard Normal

Suppose $z \sim \mathcal{N}[0,1]$. Then the left-truncated moments of z are,

1. $\mathbb{E}[z|z > c] = \phi(c)/[1 - \Phi(c)]$
2. $\mathbb{E}[z^2|z > c] = 1 + c\phi(c)//[1 - \Phi(c)]$
3. $V[z|z > c] = 1 + c\phi(c)//[1 - \Phi(c)] - \phi(c)^2/[1 - \Phi(c)]^2$

(Cameron & Trivedi, 2005)

Appendix B

Heckman two-step

```
clear all
S=100;
rng(1234)
s=1;
for s=1:S
R=1575;
mu=[0 0];
sigma=[1 0.4;0.4 0.8];
F=mvnrnd(mu,sigma,R);
e1=F(:,1);
e2=F(:,2);
X=[randn(R,5)];
beta=[0.2 0.3 0.4 0.3 -0.4]';
y1=X*beta+e1;
y=y1>=0;
y=double(y);
X3=[X(:,1:3)];
beta1=[0.5 0.6 0.2]';
y2=X3*beta1+e2;
y2(y<0.5)= NaN;
Xir=[randn(R,750)];
dataM(:,:,s)=[y2 X Xir];
end
clearvars -except dataM S
s=1;
for s=1:S
 clearvars -except s S dataM betaM beta1M se_beta se_H ind1
     ind2
data=dataM(:,:,s);
X=data(:,2:end);
yt=data(:,1);
y=1-isnan(yt);
C=size(X,2); %the number of columns of a matrix
R=size(X,1); %the number of rows of a matrix
%beta=[0;0.2;0.3] %initalize the coeff with 0
k=C;
```

```
cvx_precision('low');
cvx_begin quiet
    variables b(k,1)
    N=size(y,1);
    minimize((1/N)*(-sum(y.*log_normcdf(X*b)+(1-y).*
       log_normcdf(-X*b))));
    cvx_end
beta=b;
betaM(:,s)=beta;
xb=X*beta;
currentPrecision = digits;
digitsOld = digits(2^28);
pxb=normpdf(xb);
ind1(:,s)=pxb;
cxb=normcdf(xb);
ind2(:,s)=cxb;
cxb(cxb==0)=1e-16;
cxb(cxb==1)=1-1e-16;
pxb(pxb==0)=1e-16;
v1=(pxb.^2)./(cxb.*(1-cxb));
v1(v1==0)=1e-16;
V5=X'*diag(v1)*X;
V6=inv(V5);
se_beta(:,s)=sqrt(diag(V6));
% the mills ratio
M=pxb./cxb;

% uncensored data
y3=yt(find(1-isnan(yt)));
X2=[X,M];
X2=X2(find(1-isnan(yt)),:);
beta1=(X2'*X2)^-1*(X2'*y3);
beta1M(:,s)=beta1;

% Heckman standard errors
M1=M(find(1-isnan(yt))); %inverse mill's ratio
Xs=X(find(1-isnan(yt)),:);
W=[Xs M1];
delta=M1.*(M1+Xs*beta);
e=y3-W*beta1;
sigma2=e'*e+beta1(end)^2*sum(delta);
sigma2=sigma2/size(y3,1);
sigma=sqrt(sigma2);
rho=beta1(end)/sigma;

% Censor delta and selection equation covariates

% Heckman variance matrix
```

```
R1=size(Xs,1);
D=delta.*eye(R1);
Q=rho^2.*((X2'*D*Xs)*V6*(Xs'*D*X2));
Rv=(ones(R1,1)-rho^2.*delta).*eye(R1);
V2s=sigma^2.*(inv(X2'*X2)*(X2'*Rv*X2+Q)*inv(X2'*X2));
se_H(:,s)=sqrt(diag(V2s));
digNew=digits;
digits(digitsOld)
s
save 2step12341575.mat betaM beta1M se_beta se_H ind1 ind2
end
%% heatmaps
%tstat=betaM./se_beta;
%tstat1=beta1M./se_H;
index=find(isnan(betaM(1,:)));
betaM(:,index)=[];
beta1M(:,index)=[];
beta1map=abs(betaM)>10^-3;
beta2map=abs(beta1M)>10^-3;
%heat1=heatmap(abs(beta1map(1:40,:)),'CellLabelColor','none
   ', 'GridVisible','off','XLabel','Simulation','YLabel', '
   Variables')
heat2=heatmap(abs(beta2map([1:40 end],:)),'CellLabelColor',
   'none', 'GridVisible','off','XLabel','Simulation','
   YLabel', 'Variables')
%heat1=heatmap(abs(betaM(1:40,200:250)),'CellLabelColor','
   none', 'GridVisible','off')
%heat2=heatmap(abs(beta1M([1:40],200:250)),'CellLabelColor
   ','none', 'GridVisible','off')
%heatt=heatmap(abs(tstat(1:40,:)),'CellLabelColor','none',
   'GridVisible','off')
%heatt1=heatmap(abs(tstat1([1:40 end],:)),'CellLabelColor
   ','none', 'GridVisible','off')
%se_beta(:,index)=[];
%se_H(:,index)=[];
```

The proposed model

```
%clear all
rng(1234);
S=1000;
R=1650;
sp=750;
P=sp+6; %update this and the number of irrelevant variables
data=zeros(R,P,S);
for s=1:S
mu=[0 0];
sigma=[1 0.4;0.4 0.8];
```

```
F=mvnrnd(mu,sigma,R);
e1=F(:,1);
e2=F(:,2);
X=[randn(R,5)];
beta=[0.2 0.3 0.4 0.3 -0.4]';
y1=X*beta+e1;
y=y1>=0;
y=double(y);
X3=[X(:,1:3)];
beta1=[0.5 0.6 0.2]';
y2=X3*beta1+e2;
y2(y<0.5)= NaN;
Xir=[randn(R,sp)];
dataM(:,:,s)=[y2 X Xir];
end
%this is to get the range to cross-validate and then apply
   on the
clearvars -except dataM
S=1000;
s=1;
for s=1:S
tic
clearvars -except S s dataM beta1fx beta2fx
lamax=0.4;
data=dataM(:,:,s);
ymiss=data(:,1);
yt=ymiss;
y=1-isnan(yt);
X=data(:,2:end); %change this
C=size(X,2); %the number of columns of a matrix
R=size(X,1); %the number of rows of a matrix
beta= (X'*X)\(X'*y);
h=1;
j=1;
J=5;
CV5=[];
lamcv=[];
lamc=0;
k=C;
cvx_precision('low');
%PROBIT
while h==1
cvx_begin quiet
    variables b(k,1)
    N=size(y,1);
    minimize((1/N)*(-sum(y.*log_normcdf(X*b)+(1-y).*
       log_normcdf(-X*b)))+lamc*norm(b,1));
    cvx_end
```

```
    h=max(abs(b))>10^-3;
    lamc=lamc+1;
b0=b;
end
toc
tic
lamc=lamc-2;
h=1;
while h==1
cvx_begin quiet
    variables b(k,1)
    N=size(y,1);
    minimize((1/N)*(-sum(y.*log_normcdf(X*b)+(1-y).*
       log_normcdf(-X*b)))+lamc*norm(b,1));
    cvx_end
h=max(abs(b))>10^-3;
lamc=lamc+0.1;
b0=b;
end
lamax(s)=lamc;
toc
tic
K=40;
i=1;
lamcv=linspace(0,lamax(s),K); %this is set to 2 to save on
   time
for i=1:K
    lamc5=lamcv(i);
    j=1;
for j=1:J
    yf=y; %pick the the sample without k
    yf((((R/J)*(j-1))+1):((R/J)*j))=[];
    Xf=X;
    Xf((((R/J)*(j-1))+1):((R/J)*j),:)=[];
    cvx_begin quiet
    variables b(k,1)
    N=size(yf,1);
    minimize((1/N)*(-sum(yf.*log_normcdf(Xf*b)+(1-yf).*
       log_normcdf(-Xf*b)))+lamc5*norm(b,1));
    cvx_end
    pred5(:,j)=normcdf(X((((R/J)*(j-1))+1):((R/J)*j),:)*b);
    predclass5(:,j)=pred5(:,j)>0.5;
    predacc5(:,j)=(predclass5(:,j)==y((((R/J)*(j-1))+1):((R
       /J)*j)));
    misclassr5=1-(sum(predacc5(:,j))/size(predacc5(:,j),1))
       ; %this is a scalar
    misclass5(j)=misclassr5;
end
```

```
CV5=[CV5; mean(misclass5)];
end
toc
tic
[M,I]=min(CV5);
lamf1(s)=lamcv(I);
lamfl(s)=lamf1(s)*0.6;
lamfr(s)=lamf1(s)*1.4;
K=10;
i=1;
lamc=2;
pred5=[];
predclass5=[];
predacc5=[];
misclassr5=[];
misclass5=[];
CV5=[];
lamcv=[];
lamcv=linspace(lamfl(s),lamfr(s),K);
for i=1:K
    lamc5=lamcv(i);
    j=1;
for j=1:J
    yf=y; %pick the the sample without k
    yf((((R/J)*(j-1))+1):((R/J)*j))=[];
    Xf=X;
    Xf((((R/J)*(j-1))+1):((R/J)*j),:)=[];
    cvx_begin quiet
    variables b(k,1)
    N=size(yf,1);
    minimize((1/N)*(-sum(yf.*log_normcdf(Xf*b)+(1-yf).*
       log_normcdf(-Xf*b)))+lamc5*norm(b,1));
    cvx_end
    pred5(:,j)=normcdf(X((((R/J)*(j-1))+1):((R/J)*j),:)*b);
    predclass5(:,j)=pred5(:,j)>0.5;
    predacc5(:,j)=(predclass5(:,j)==y((((R/J)*(j-1))+1):((R
       /J)*j)));
    misclassr5=1-(sum(predacc5(:,j))/size(predacc5(:,j),1))
       ; %this is a scalar
    misclass5(j)=misclassr5;
end
b0=b;
CV5=[CV5; mean(misclass5)];
end
toc
M=[];
Id=[];
[M,Id]=min(CV5);
```

```
lamf(s)=lamcv(Id);
y=1-isnan(yt);
X=data(:,2:end);
cvx_begin quiet
    variables b(k,1)
    N=size(y,1);
    minimize((1/N)*(-sum(y.*log_normcdf(X*b)+(1-y).*
       log_normcdf(-X*b)))+lamf(s)*norm(b,1));
 cvx_end
beta1(:,s)=b;
bseli=find(abs(b)>10^-3); %post lasso here
bsel=b(bseli);
ka=size(bsel,1);
Xsel=X(:,bseli);
lamk=0;
cvx_begin quiet
    variables b(ka,1)
    N=size(y,1);
    minimize((1/N)*(-sum(y.*log_normcdf(Xsel*b)+(1-y).*
       log_normcdf(-Xsel*b)))+lamk*norm(b,1));
 cvx_end
beta1pl=b;
tic
yhat=Xsel*b;
IMR=normpdf(yhat)./normcdf(yhat);
y=ymiss(find(1-isnan(ymiss)));
IMR=IMR(find(1-isnan(ymiss)));
X=[X(find(1-isnan(ymiss)),1:end) IMR];
lamd=0;
h=1;
j=1;
J=5;
CV5d=[];
lamdv=[];
kb=size(X,2);
while h==1
cvx_begin quiet
    variables b(kb,1)
    N=size(y,1);
    minimize((1/N)*(sum_square(y-X*b))+lamd*norm(b,1));
cvx_end
h=max(abs(b))>10^-4;
lamd=lamd+1;
end
lamd=lamd-2;
h=1;
while h==1
cvx_begin quiet
```

```
    variables b(kb,1)
    N=size(y,1);
    minimize((1/N)*(sum_square(y-X*b))+lamd*norm(b,1));
 cvx_end
h=max(abs(b))>10^-4;
lamd=lamd+0.01;
end
lamax2(s)=lamd;
ysz=size(y,1); %get it to the closest multiple of 5
modu=mod(ysz,J); %take this modulus then delete this
   observations from the end of the dataset
ymod=y;
ymod(ysz-modu+1:ysz)=[];
Xmod=X;
Xmod(ysz-modu+1:ysz,:)=[];
% need some censoring here
R=size(ymod,1);
toc
tic
K=40;
i=1;
lamdv=linspace(0,lamax2(s),K); %to save time, I don't know
   if 2 is the best
CV5d=[];
predl5=[];
prederror=[];
rmsev=[];
for i=1:K
    lamd5=lamdv(i);
    j=1;
for j=1:J
    yf=ymod; %pick the the sample without k
    yf((((R/J)*(j-1)))+1:((R/J)*j))=[];
    Xf=Xmod;
    Xf((((R/J)*(j-1))+1):((R/J)*j),:)=[];
    cvx_begin quiet
    variables b(kb,1)
    N=size(yf,1);
    minimize(((1/N)*sum_square(yf-Xf*b))+lamd5*norm(b,1));
    cvx_end %find the beta
    Dummy=Xmod((((R/J)*(j-1))+1):((R/J)*j),:);
    predl5(:,j)=Dummy*b;
    prederror(:,j)=ymod((((R/J)*(j-1))+1):((R/J)*j))-predl5
       (:,j);
    rmsev(:,j)=sqrt(sum(prederror(:,j).^2)/(size(ymod((((R/
       J)*(j-1))+1):((R/J)*j)),1)));
end
CV5d=[CV5d; mean(rmsev)];
```

```
end
toc
[Mm,Im]=min(CV5d);
lamg1(s)=lamdv(Im);
lamgl(s)=lamg1(s)*0.4;
lamgr(s)=lamg1(s)*1.6;
K=10;
i=1;
lamdv=[];
lamdv=linspace(lamgl(s),lamgr(s),K);
CV5d=[];
predl5=[];
prederror=[];
rmsev=[];
tic
for i=1:K
    lamd5=lamdv(i);
    j=1;
for j=1:J
    yf=ymod; %pick the the sample without k
    yf((((R/J)*(j-1)))+1:((R/J)*j))=[];
    Xf=Xmod;
    Xf((((R/J)*(j-1))+1):((R/J)*j),:)=[];
    cvx_begin quiet
    variables b(kb,1)
    N=size(yf,1);
    minimize(((1/N)*sum_square(yf-Xf*b))+lamd5*norm(b,1));
    cvx_end %find the beta
    Dummy=Xmod((((R/J)*(j-1))+1):((R/J)*j),:);
    predl5(:,j)=Dummy*b;
    prederror(:,j)=ymod((((R/J)*(j-1))+1):((R/J)*j))-predl5
       (:,j);
    rmsev(:,j)=sqrt(sum(prederror(:,j).^2)/(size(ymod((((R/
       J)*(j-1))+1):((R/J)*j)),1)));
end
CV5d=[CV5d; mean(rmsev)];
end
toc
Mm=[];
Im=[];
[Mm,Im]=min(CV5d);
lamg(s)=lamdv(Im);
cvx_begin quiet
    variables b(kb,1)
    N=size(y,1);
    minimize(((1/N)*sum_square(y-X*b))+lamg(s)*norm(b,1));
cvx_end
blindx=find(b>10^-3);
```

```
blasso=b(find(b>10^-3));
beta2(:,s)=b;
kc=size(blasso,1);
Xsel2=X(:,blindx);
cvx_begin quiet
    variables b(kc,1)
    N=size(y,1);
    minimize(((1/N)*sum_square(y-Xsel2*b))+lamg(s)*norm(b
       ,1));
cvx_end
beta2pl=b;
s
end
```