The Lasso: Examining the Variable Selection Consistency and an Application -- Comparison of Factor Importance on the Household Carbon Emissions

Xinyue Wang || 3504357 || s6xnwang@uni-bonn.de

Final Project of Computational Statistics | M.Sc. Economics in Summer Semester 2022 | University of Bonn

9 August, 2022

1 Introduction

Lasso is being used as a computationally feasible alternative to model selection. In this project I will focus on the simulation study and an application of Lasso. The structure of this project is as follows:

• Section 2 is the introduction to the properties of the Lasso method.

• Section 3 is the simulation study. I simulate the data to study the Lasso for variable selection purposes, examine the consistency of the Lasso to select the true model under the Strong Irrepresentable Condition.

• Section 4 is the empirical application. I use survey data to calculate the carbon emissions of Chinese households' energy consumption from a micro perspective, and use the Lasso to analyse the importance of a series of factors that affect household carbon emissions in rural and urban regions.

• Section 5 concludes the simulation and application of this project.

2 The Lasso Method

This section is an introduction to the Least Absolute Shrinkage and Selection Operator (LASSO) regression, based on papers and textbooks. I introduce Lasso's formulation, advantages and properties.

2.1 Formulation of the Lasso

There are some ways to improve the simple linear model, by replacing plain least squares fitting with some alternative fitting procedures. *Shrinkage* (or *Regularization*) is an approach that involves fitting a model involving all p predictors, and shrinking the estimated coefficients towards zero. The two best-known techniques for Shrinkage are *Ridge regression* and *the Lasso*.

• Given a linear regression with standardized predictors x_{ij} and centred response values y_i for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, the lasso *(Tibshirani, 1996)* solves the $\ell 1$-penalized regression problem of finding coefficients $\hat{\beta}_\lambda^L$ to

$$\text{minimize} \left\{ \underbrace{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p}|\beta_j|}_{\text{lasso penalty}} \right\} \qquad (1)$$

From the above formulation, the Lasso uses an $\ell 1$ *penalty* ($\ell 1$ norm of a coefficient vector β is $\| \beta \|_1 = \sum |\beta_j|$) , which has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

• Here is another formulation for the Lasso. The Lasso coefficient estimates solve the problem:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \right\} \quad \text{subject to } \sum_{j=1}^{p}|\beta_j| \leq s \qquad (2)$$

For every value of λ, there is some s such that the formulations *(1)* and *(2)* will give the same Lasso coefficient estimates. When $\lambda = 0$ (or $s \to \infty$) , the Lasso simply gives the least squares fit. When $\lambda \to \infty$ (or $s = 0$) , the Lasso shrinks all coefficient estimates to zero. In between these two extremes, depending on the value of λ or s, the Lasso can produce a model involving any number of variables.

2.2 Advantages of the Lasso

Comparing Shrinkage and Least Squares

The Lasso is an improvement of the simple standard linear model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, which is fitted using least squares. The reason for using Shrinkage instead of least squares is that it can yield better prediction accuracy and model interpretability.

• *Prediction Accuracy*

If the true relationship between the response and the predictors is approximately linear, and n (the number of observations) $\gg p$ (the number of variables), then the least squares estimates tend to have low bias and low variance.

However, if n is not much larger than p, the least squares fit will result in overfitting and poor predictions on test data.

And if $p > n$, the least squares fit cannot be used since there is no longer a unique coefficient estimate.

Hence, by shrinking the estimated coefficients, the Lasso can significantly reduce variance, and improve accuracy of predicting the response for observations not used in model training.

• *Model Interpretability*

Some of the variables used in a multiple regression model are in fact not associated with the response. The least squares is unlikely to yield any coefficient estimates that are exactly zero.

Hence, by setting the corresponding coefficient estimates to zero, the Lasso yields a model that is more easily interpreted, and performs *variable selection* (or *feature selection*) for excluding irrelevant variables from a multiple regression model.

Comparing the Lasso and Ridge Regression

• *Model Interpretability*

Ridge regression has one obvious disadvantage: it will include all p predictors in the final model, and the penalty will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero, which means it will not result in exclusion of any of the variables.

This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation when p is quite large. So the Lasso produces simpler and more interpretable models that involve only a subset of the predictors.

• *Prediction Accuracy*

When the least squares estimates have excessively high variance, the Lasso solution can yield a reduction in variance at the expense of a small increase in bias, and consequently can generate more accurate predictions.

Together with the use of the *cross-validation* to determine which approach is better on a particular data set, the Lasso generally performs better when relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or equal to zero.

2.3 Properties of the Lasso

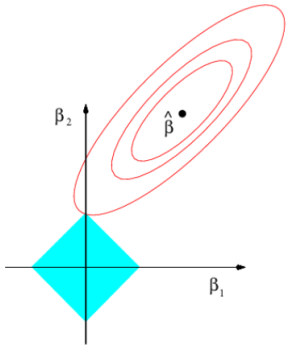
2.3.1 The Variable Selection (Feature Selection) Property of the Lasso

The following figure illustrates the situation where the Lasso results in coefficient estimates which are exactly equal to zero. Consider $p = 2$, $\hat{\beta}$ is the least squares solution, the blue diamond represents the lasso constraints in *(2)*, each of the ellipses centred around $\hat{\beta}$ represents a contour with all of the points on a particular ellipse having the same RSS value.

When s in *(2)* is sufficiently large (corresponding to $\lambda = 0$ in *(1)*), the constraint region will contain $\hat{\beta}$, and the Lasso estimates will be the same as the least squares estimates. When $s = 0$ (equivalently λ is sufficiently large), the Lasso gives the null model in which all coefficient estimates equal zero.

Formulation *(2)* indicates that the Lasso coefficient estimates are given by the first point at which an ellipse contacts the constraint region. The Lasso constraint has corners on each of the axes. When the ellipse intersects the constraint region on an axis, one of the coefficients will equal zero. In higher dimensions, many of the coefficient estimates may equal zero simultaneously.

Hence, the Lasso can perform variable selection because of the $\ell 1$ penalty. Some Lasso coefficients are shrunken entirely to zero, and generate *sparse models* involving only a subset of the variables, which are more interpretable and can provide a good forecast.

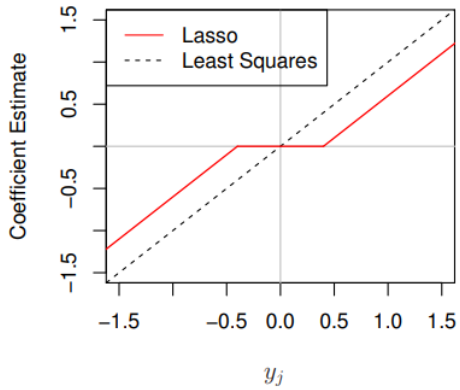


2.3.2 The Shrinkage Behaviour of the Lasso

Consider the special case with $n = p$, and $X = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$ is a diagonal matrix with 1's on the diagonal. and assume we perform regression without an intercept.

With these assumptions, the Lasso amounts to finding the coefficients $\beta_1, \ldots, \beta_p$ such that $\Sigma_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\Sigma_{j=1}^{p}|\beta_j|$ is minimized, and the Lasso estimates take the form $\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & , \text{if } y_j > \frac{\lambda}{2} \\ 0 & , \text{if } |y_j| \leq \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2} & , \text{if } y_j < -\frac{\lambda}{2} \end{cases}$.

The following figure displays the *soft-thresholding* situation: the Lasso shrinks each least squares coefficient towards zero by the constant amount $\frac{\lambda}{2}$, and the least squares coefficients that are less than $\frac{\lambda}{2}$ in absolute value are shrunken entirely to zero.



2.3.3 The Tuning Parameter λ of the Lasso

Implementing the Lasso requires selecting a value for the tuning parameter λ (equivalently the value of the constraint s). *Cross-validation* can tackle this problem:

First, choose a grid of λ values, and compute the cross-validation error for each value of λ.

Then, select the λ value for which the cross-validation error is smallest.

Finally, re-fit the model using all of the available observations and the selected value of λ.

3 Simulation

The simulation study in this project will be based on the paper *On the Consistency of Model Selection in Lasso.* (Zhao and Yu, 2006). In order to use Lasso for variable selection, it is necessary to assess how well the sparse model given by Lasso relates to the true model. Hence this simulation study investigates Lasso's variable selection consistency, uses R to do the Data Generation Process, and it shows that the Lasso will consistently select the true model when Strong Irrepresentable Condition holds.

3.1 Variable Selection Consistency and Strong Irrepresentable Conditions

Before the simulation, here are some definitions and notations on the design in the paper.

• **Notation:** Without loss of generality,
Assume $\beta^n = (\beta_1^n, \ldots, \beta_q^n, \beta_{q+1}^n, \ldots \beta_p^n)^T$, where $\beta_j^n \neq 0$ for $j = 1, .., q$ and $\beta_j^n = 0$ for $j = q+1, \ldots, p$.
Let $\beta_{(1)}^n = (\beta_1^n, \ldots, \beta_q^n)^T$ and $\beta_{(2)}^n = (\beta_{q+1}^n, \ldots, \beta_p^n)$.
Write $\mathbf{X_n}(1)$ and $\mathbf{X_n}(2)$ for the first q and last $p-q$ columns of X_n respectively.
Let $C^n = \frac{1}{n}{\mathbf{X_n}}^T\mathbf{X_n} = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}$.
Set $C_{11}^n = \frac{1}{n}\mathbf{X_n}(1)^T\mathbf{X_n}(1)$, $C_{21}^n = \frac{1}{n}\mathbf{X_n}(2)^T\mathbf{X_n}(1)$, $C_{12}^n = \frac{1}{n}\mathbf{X_n}(1)^T\mathbf{X_n}(2)$ and $C_{22}^n = \frac{1}{n}\mathbf{X_n}(2)^T\mathbf{X_n}(2)$.

• **Sign Consistency:** Written as $\hat{\beta}^n =_s \beta^n$, an estimate $\hat{\beta}^n$ is equal in sign with the true model β^n if and only if $sign(\hat{\beta}^n) = sign(\beta^n)$, where $sign(\cdot)$ maps positive entries to 1, negative entries to -1 and zero to zero. In the simulation part, the variable selection consistency is *Sign Consistency* (Zhao and Yu, 2006).

• **Strong Irrepresentable Condition:** Assuming C_{11}^n is invertible, there exists a positive constant vector η, such that $|C_{21}^n(C_{11}^n)^{-1}sign(\beta_{(1)}^n)| \leq \mathbf{1} - \eta$, where $\mathbf{1}$ is a $p-q$ by 1 vector of 1's and the inequality holds element-wise.

The Irrepresentable Condition closely resembles a regularization constraint on the regression coefficients of the irrelevant covariates ($\mathbf{X_n}(2)$) on the relevant covariates ($\mathbf{X_n}(1)$). In particular, when signs of the true β are unknown, for the Irrepresentable Condition to hold for all possible signs, we need the $\ell 1$ norms of the regression coefficients to be smaller than 1, i.e. $|((\mathbf{X_n}(1)^T\,\mathbf{X_n}(1))^{-1}\,\mathbf{X_n}(1)^T\,\mathbf{X_n}(2)| = |(C_{11}^n)^{-1}\,C_{12}^n| < 1 - \eta$, the total amount of an irrelevant covariate represented by the covariates in the true model is not to reach 1.

When Strong Irrepresentable Condition holds with a larger constant η, it is easier for Lasso to pick up the true model, hence Strong Irrepresentable Condition allows for consistent variable selection and parameter estimation simultaneously. Next, I will demonstrate Lasso's variable selection consistency and the Strong Irrepresentable Conditions using two simulation cases.

3.2 Simulation Case 1: Consistency and Inconsistency with 3 Variables

In this case, my aim is to see the Lasso's behavior when Strong Irrepresentable Condition holds and fails.

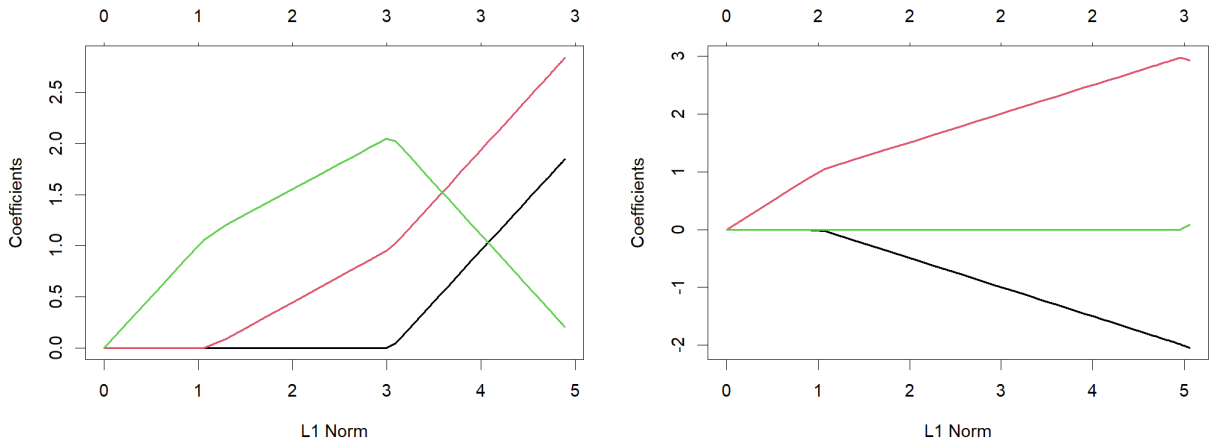
First, the **Data Generation Process** is as follows:

- Generate *i.i.d.* variables x_{i1}, x_{i2}, e_i and ε_i, with variance 1 and mean 0 for $i = 1, \ldots, n$ and $n = 1000$
- Generate a third *i.i.d.* predictor x_{i3}, by $x_{i3} = \frac{2}{3}x_{i1} + \frac{2}{3}x_{i2} + \frac{1}{3}e_i$
- The response is generated by $Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$
- Lasso is applied (through `glmnet` in R) on Y, X_1, X_2 and X_3 in two settings:

  (setting a) : $\beta_1 = 2, \beta_2 = 3$; (setting b) :$\beta_1 = -2, \beta_2 = 3$.

Then, in both settings, I verify whether the Strong Irrepresentable Condition holds or fails. According to $\mathbf{X}(1) = (X_1, X_2)$, $\mathbf{X}(2) = X_3$ and $|((\mathbf{X_n}(1)^T\,\mathbf{X_n}(1))^{-1}\,\mathbf{X_n}(1)^T\,\mathbf{X_n}(2)| = |(C_{11}^n)^{-1}\,C_{12}^n| < 1 - \eta$, I calculate $C_{21}\,C_{11}^{-1} = \mathbf{X}(2)^T\mathbf{X}(1)\,(\mathbf{X}(1)^T\mathbf{X}(1))^{-1} = (\frac{2}{3}, \frac{2}{3})$. In R the result of the calculation is (0.66, 0.68). Therefore, the Strong Irrepresentable Condition fails for *setting a*, and holds for *setting b*.

Then, I investigate how these two different settings lead to Lasso's sign consistency and inconsistency respectively. As the amount of regularization (controlled by λ) varies, I observe that different Lasso solutions form *the Lasso path* (illustrated by following figure). In the following figures, $\beta_1, \beta_2, \beta_3$ are represented by black, red and green lines, respectively.



The **results** of case 1 are as follows:

• For *setting a* (left figure), the Strong Irrepresentable Condition fails and Lasso is sign inconsistent. Since it does not shrink $\hat{\beta}_3$ to 0, but picks up X_3 first and never shrinks it back to zero, the $\ell 1$ regularization prefers X_3.

• For *setting b* (right figure), the Strong Irrepresentable Condition holds and Lasso is sign consistent. With a proper amount of regularization, the Lasso correctly shrinks $\hat{\beta}_3$ to 0.

The R **code** for case 1 is following:

```
################### Simulation Case 1: (In)Consistency with 3 Variables ###################
rm(list=ls())
library(MASS)
library(glmnet) # package for the Lasso

####=== data generating process ===####
mydata1 <- function(n, miu, sigma, beta){
    x.1    <- rnorm(n, miu, sigma)
    x.2    <- rnorm(n, miu, sigma)
```

```
    x.3    <- 2/3*x.1 + 2/3*x.2 + 1/3*rnorm(n,miu,sigma)
    x      <- cbind(x.1, x.2, x.3)
    error  <- rnorm(n, miu, sigma)
    y      <- x.1*beta[1] + x.2*beta[2]+ error
    data   <- data.frame(y, x)
    return(data)
}

####=== setting a: β=(2,3) ===####
set.seed(123)
traindata.a <- mydata1(n = 1000, miu = 0, sigma = 1, beta = c(2, 3))
x           <- cbind(traindata.a$x.1, traindata.a$x.2, traindata.a$x.3)
#=== verify Strong Irrepresentable Condition (fails for setting a) ===#
X1 <- cbind(traindata.a$x.1, traindata.a$x.2)
X2 <- traindata.a$x.3
solve(t(X1) %*% X1) %*% t(X1) %*% X2
#=== the Lasso paths for setting a ===#
lasso.sim.a <- glmnet(x, traindata.a$y, family = 'gaussian', alpha = 1,
                      nlambda = 100, standardize = T, intercept = F)
plot(lasso.sim.a, lwd = 2)
cv.sim.a <- cv.glmnet(x, traindata.a$y, family = 'gaussian', nfolds = 10, alpha = 1,
                      nlambda = 100, standardize = T, intercept = F)

####=== setting b: β=(-2,3) ===####
set.seed(123)
traindata.b <- mydata1(n = 1000, miu = 0, sigma = 1, beta = c(-2, 3))
x           <- cbind(traindata.b$x.1, traindata.b$x.2, traindata.b$x.3)
#=== verify Strong Irrepresentable Condition (holds for setting b) ===#
X1 <- cbind(traindata.b$x.1, traindata.b$x.2)
X2 <- traindata.b$x.3
solve(t(X1) %*% X1) %*% t(X1) %*% X2
#=== the Lasso paths for setting 2 ===#
lasso.sim.b <- glmnet(x, traindata.b$y, family = 'gaussian', alpha = 1,
                      nlambda = 100, standardize = T, intercept = F)
plot(lasso.sim.b, lwd = 2)
cv.sim.b <- cv.glmnet(x, traindata.b$y, family = 'gaussian', nfolds = 10, alpha = 1,
                      nlambda = 100, standardize = T, intercept = F)
```

3.3 Simulation Case 2: Impact of Strong Irrepresentable Condition on Model Selection

In this case I simulate data to show the relationship between the probability of Lasso selecting the correct model and how well the Strong Irrepresentable Condition holds or fails.

The **Data Generation Process** and the idea for the code are as follows:

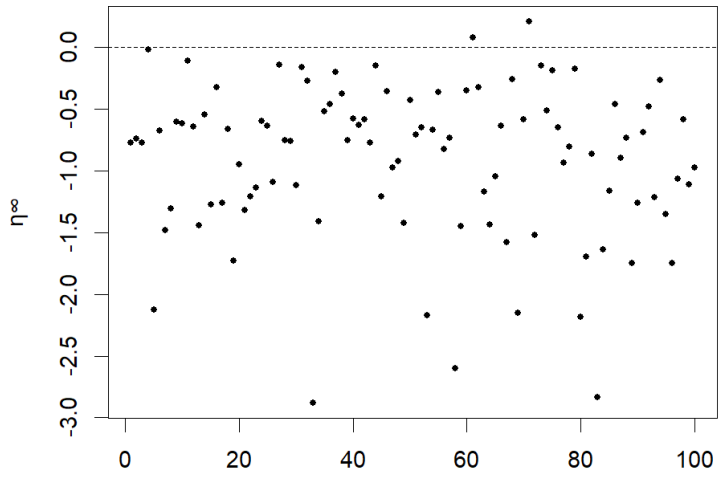
• Take $n = 100, p = 32, q = 5, \beta_1 = (7, 4, 2, 1, 1)^T$, choose a small $\sigma^2 = 0.1$ to go to asymptotic quickly.

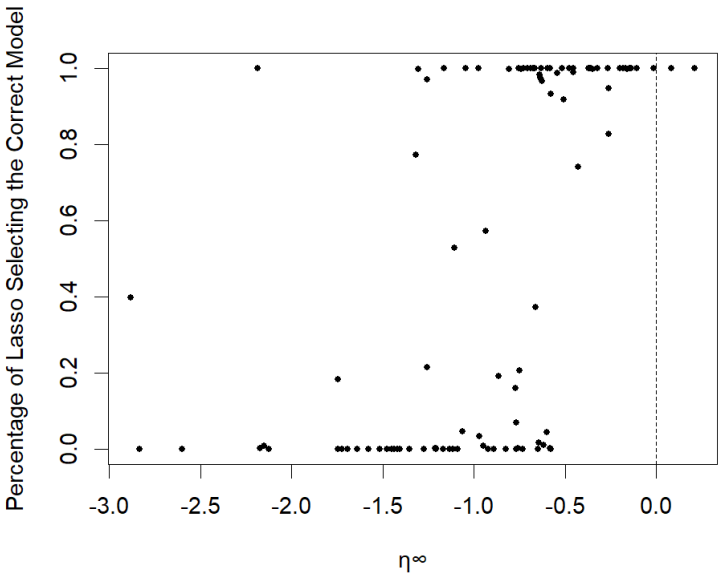
• Generate $i = 100$ designs of X as follows. First sample a covariance matrix S from Wishart(p, p) (through `rwishart` in R), take the Wishart(p, p) measure which centers but does not concentrate around the identity matrix. Then take n samples of $X_i \sim \mathcal{N}(0, S)$. Such generated samples represent a variety of designs: some satisfy the Strong Irrepresentable Condition with a large η, while others fail the condition badly.

• Calculate $\eta_\infty = 1 - \|C_{21}^n (C_{11}^n)^{-1} sign(\beta_{(1)}^n)\|_\infty$, to evaluate how well the Strong Irrepresentable condition holds, if $\eta_\infty > 0$, the Strong Irrepresentable holds, otherwise it fails.

• Run the simulation $T = 1000$ times for each design, and examine general sign consistencies. Each time, n samples of ε are generated from $\mathcal{N}(0, \sigma^2)$ and $Y = X\beta + \varepsilon$ are calculated. Next, I run Lasso to calculate the Lasso path, which is examined to see if there exists a model estimate that matches the signs of the true model. Then, I compute the percentage of $T = 1000$ runs that generated matched models for each design.

• Compare the percentage of correct selections (y-axis) to η_∞ (x-axis) for $i = 100$ designs of X, shown in the following figures.





The **results** of case 2 are as follows:

• In terms of the Strong Irrepresentable Condition: according to the calculation, η_∞ of the 100 simulated designs are within [-2.88, 0.21], with 2 of them bigger than 0.

• In terms of Variable Selection: when η_∞ gets large, the percentage of Lasso selecting the correct model goes up and there is a steep increase. For η_∞ larger than 0, the percentage is close to 1. On the other hand, for η_∞ considerably smaller than 0 ($\eta_\infty < -1.5$), there is little chance for Lasso to select the true model.

This quantitatively illustrates the importance of the Strong Irrepresentable Condition for Lasso's variable selection performance.

The R **code** for case 2 is as follows:

```
################## Simulation Example 2: Quantitative Evaluation of Impact #################
library(matrixsampling) # package for Wishart distribution
library(mvtnorm)        # package for X distribution
library(glmnet)         # package for the Lasso

#######=== data generating process ===#######
mydata2 <- function(n, p, q, beta, S){
    miu   <- 0
    sigma <- sqrt(0.1)
    error <- rnorm(n, miu, sigma)
    X     <- rmvnorm(n, rep(miu,p), S)
    Y     <- X %*% beta + error
    data  <- data.frame(Y, X)
    return(data)
}
n    <- 100    # number of observations
p    <- 32     # number of variables
q    <- 5      # number of non-zero beta
beta <- c(7, 4, 2, 1, 1, rep(0,p-q))

#######=== Impact of Strong Irrepresentable Condition on Variable Selection ===#######
correct    <- matrix(NA, 1000, 1)
percentage <- matrix(NA, 100, 1) # creat store for y-axis
eta        <- matrix(NA, 100, 1) # creat store for x-axis

for (i in 1:100){ # generate 100 design of X #
    set.seed(123)
    S <- as.matrix(rwishart(100, nu = 13, Sigma = diag(p)) [, , i])

    # verify Strong Irrepresentable Condition by η #
    sim.data <- mydata2(n, p, q, beta, S)
    X1  <- as.matrix( sim.data[ , 2:(q+1)] )     # X with non-zero β
    X2  <- as.matrix( sim.data[ , (q+2):(p+1)] ) # X with zero β
    C11 <- 1/n * t(X1) %*% X1
    C21 <- 1/n * t(X2) %*% X1
    beta.sign <- as.matrix(sign(beta[1:q]))
    eta[i,1]  <- 1 - norm(C21 %*% solve(C11) %*% beta.sign, type = 'I')

    # compute percentage of generating matched models #
    set.seed(123)
    for (T in 1:1000){ # generate 1000 simulations for each design #
        data <- mydata2(n, p, q, beta, S)
        # calculate the Lasso path #
        lasso.sim <- glmnet(as.matrix(data[,2:(p+1)]), data$Y,
                            alpha = 1, intercept = F)
        cv.sim <- cv.glmnet(as.matrix(data[,2:(p+1)]), data$Y,
                            alpha = 1, intercept = F)
        beta.sim <- predict(lasso.sim, type = 'coefficients',
                            s = cv.sim$lambda.min) [2:(q+1), ]
        # examine sign of estimate #
        ifelse(sum(sign(beta.sim)) == q, correct[T,1] <- 1, correct[T,1] <- 0)
    }
    percentage[i,1] <- sum(correct)/1000
}
```

```
# plot #
plot(eta, type = 'p', pch = 20, xlab=' ', ylab='η∞')
abline(h = 0, lty = 2)
plot(eta,percentage, type = 'p', pch = 20,
     xlab = 'η∞', ylab = 'Percentage of Lasso Selecting the Correct Model')
abline(v = 0, lty = 2)
```

4 Empirical Application

4.1 Research Background

According to the World Energy Statistics Yearbook, China became the world's largest emitter of carbon dioxide in 2006, with 35% of the carbon emissions coming from household consumption (Tian et al., 2014). As an important component of global carbon emissions, Household Carbon Emissions (HCEs) cannot be ignored. At the same time, as the urbanisation develops in China, urban and rural households differ in terms of lifestyle, household size and income, which in turn influence household consumption behaviour and the carbon emissions they generate. Studying the differences in carbon emissions between urban and rural households in China is conducive to exploring energy-efficient and environmentally sustainable household consumption patterns from the perspective of coordinated urban and rural development.

Therefore, my empirical application for this project will be based on the paper *Prioritizing driving factors of household carbon emissions: An application of the LASSO model with survey data* (Shi et al., 2020). I will use the CHFS survey data and CLA to measure the direct carbon emissions of urban and rural households in China in 2015, then use the Lasso model to explore the impact of a range of economic and social factors on the direct carbon emissions of households.

4.2 Data and Regression

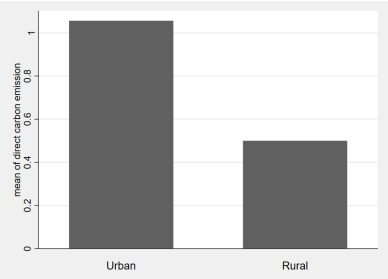
4.2.1 Direct Carbon Emissions

This application is concerned with the Household Carbon Emissions (HCEs). Qu et al. (2013) gave the definition of HCEs as direct and indirect emissions emitted by individuals or their households to meet their requirements for survival and development under certain socio-economic conditions. In the empirical application part, I will focus on the *Direct Carbon Emissions* (DCEs) for household consumption, which are the emissions from the household's final consumption of fossil fuels, i.e., coal , oil and gas.

First, I calculate the household direct carbon emissions based on the Consumer Lifestyle Approach (CLA) (Bin and Dowlatabadi, 2005). The CLA measures carbon emissions associated with household's consumption activities, and combines the input-output approach with the carbon emission factor, which can better reflect household-level carbon emissions from a micro perspective (Ding Q et al, 2017).

The direct carbon emissions for household i are $E_{direct} = \sum f_{ij} * y_{ij}$, where E_{direct} is the direct carbon emissions of household i from energy consumption, f_{ij} is the carbon emissions factor of energy source j, i.e. the carbon emissions of each sector divided by the total output of that sector, and y_{ij} represents the expenditure of household i on the direct energy source j.

The following figure shows the difference between mean direct carbon emissions per capital for *Urban* and *Rural* households. The DCEs per capita for rural households is about 0.5 tonnes, while this variable exceeds 1 tonne for urban households, more than twice as much as for rural households. Urban households emit much more carbon per capita from direct energy consumption than rural households. The reason might be that compared to urban households, rural households prefer biomass such as fuelwood and use less energy to heat their homes in winter.



The survey data on household consumption expenditure is from the 2015 China Household Finance Survey, in which the sample is distributed across 29 provinces, and the data are nationally and provincially representative. Based on the 2015 input-output tables and the 2015 China Carbon Emissions Database, I calculate the carbon emission coefficients of each household consumption behaviour, then get the final household direct carbon emissions.

For data processing, I first truncated 5% of the outlier data in the questionnaire data. Before performing the Lasso regression, all variables were standardized to mean 0 and standard deviation 1. To simplify the data, I kept the main variables in the data file that would be used in the Lasso regression, as follows:

Variable	**Description**
hhid	household identifier
rural	location of household i , rural = 1 , urban = 0
E_direct_energy	household i's direct carbon emissions by energy consumption
pop	number of household i's members (including interviewee)
per_direct	average direct carbon emissions of household i's members
per_income	average weighted income of household i's members
edu_mean	average education attainment of household i's members
health_mean	average healthy condition of household i's members
age_mean	average age of household i's members
house_area	living area of household i

4.2.2 The Linear Regression Model

My purpose in this empirical application is to analyse the importance of six different influencing factors on the direct carbon emissions of consumption by urban and rural households, therefore the regression model is developed as follows:

$$perE_{direct} = \beta_0 + \beta_{1i}\, income_i + \beta_{2i}\, population_i + \beta_{3i}\, education_i + \beta_{4i}\, area_i + \beta_{5i}\, age_i + \beta_{6i}\, health_i + \varepsilon_i$$

For the independent variables, six influencing factors were selected: household income per capita ($income_i$), household size ($population_i$), household average years of schooling ($education_i$), household living area ($area_i$), household average age (age_i) and household average health status ($health_i$). The dependent variable $perE_{direct}$ is the per capita direct carbon emissions of households consumption in urban and rural areas. Descriptive statistics were obtained by taking logarithms of all variables as follows:

Variable	Mean	Std. Dev.	Min	Max
ln (*per_direct*)	-0.61	1.02	-4.22	2.21
ln (*per income*)	9.48	1.51	2.18	14.93
ln (*pop*)	0.99	0.50	0.00	2.71
ln (*edu_mean*)	1.17	0.43	0.00	2.20
ln (*health_mean*)	1.09	0.37	0.00	1.61
ln (*age_mean*)	3.72	0.35	0.70	4.62
ln (*house_area*)	3.90	0.73	0.00	7.09

4.2.3 The Lasso

The decomposition method and the regression model are the two widely used paradigms for investigating the driving factors of HCEs, but the decomposition method faces some difficulties in decomposing per capita HCEs based on survey data and including more influencing factors, while the regression model will lead to the problems of overfitting and multicollinearity when the total number of relevant variables increases. (Shi et al., 2020)

Hence, using Lasso in this empirical application can have two practical benefits:

• Firstly, in the formulation of carbon reduction policies, accuracy, simplicity, effectiveness and cost optimization must be taken into consideration, this requires that the aim of the research should identify the most influential factors. The Lasso model can set the regression coefficients of relatively unimportant factors to zero by imposing the $\ell 1$ penalty, thereby minimizing the issue of too many variables in the policy-making process.

• Secondly, the importance of the variables in terms of the change of parameters of the Lasso model can be ranked, this gives policy-makers more flexibility in determining policy interventions.

Based on the previous analysis of the characteristics of Lasso, I chose Lasso as the fit procedure for the regression model to analyse the importance of six different influences on the direct carbon emissions by urban and rural households. The Lasso objective function is:

$$\underset{\beta_1,\beta_2,\cdots,\beta_p}{\arg\,\text{minimize}} \left\{(y-\beta X)^2\right\} \qquad s.t.\ \sum_j^p |\beta_j| \leq \lambda\,,$$

where λ is a nonnegative regularization parameter, y is the dependent variable, p is the number of independent variables X , and β are the parameters. If the value of λ is exceptionally large, then all the parameters β would have a value of zero. By reducing the value of λ gradually, some parameters will turn from a value of zero to non-zero. In this case, the larger corresponding λ value means that the variable is more important for prediction.

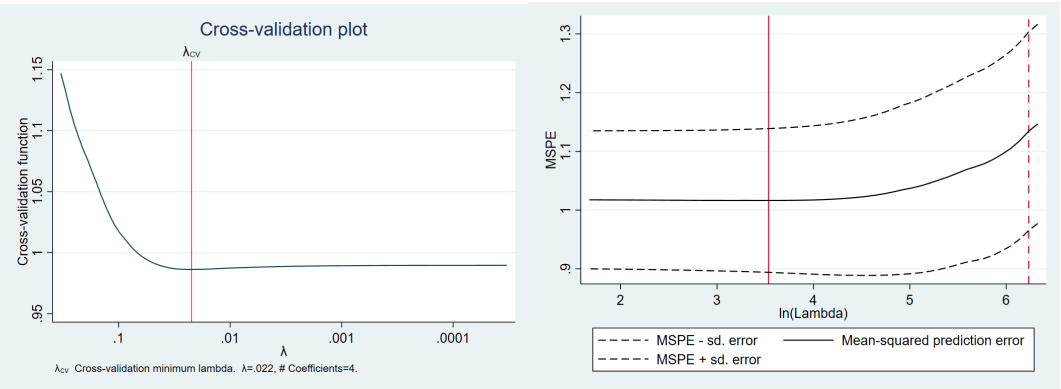
4.3 Discussion

I will analyse the Lasso regression model in three steps based on cross-validation, to explain the impact of family demographics on DCEs.

4.3.1 Filtering the household demographics that were most important to DCEs

First, I compare the Mean Square Percentage Error (MSPE) of different λ, to select the most important influencing variables on DCEs of urban and rural household separately. MSEP is a reliable judgment standard for variable selection through the Lasso regression model, since a smaller value of MSEP means a smaller deviation between the estimated value and the actual value, which also means better fit of the model. This standard can lead to the most accurate model with the smallest degree of freedom, namely the minimum number of independent variables.

Following figures show the results of MSEP of Lasso in terms of DCEs of **urban households**. The optimal λ which leads to the minimal MSEP shows that there are four family demographics that affect DCEs of urban households significantly.



Following tables show the result of Lasso for urban households. There are more coefficients of independent variables changed to non-zero with the increase of ℓ1-Norm. The model prioritizes the variables $population_i$, $income_i$, $education_i$ and $area_i$ respectively before reaching the optimal λ (ID = 30).

Lasso linear mode for Urban

Selection: Cross-validation

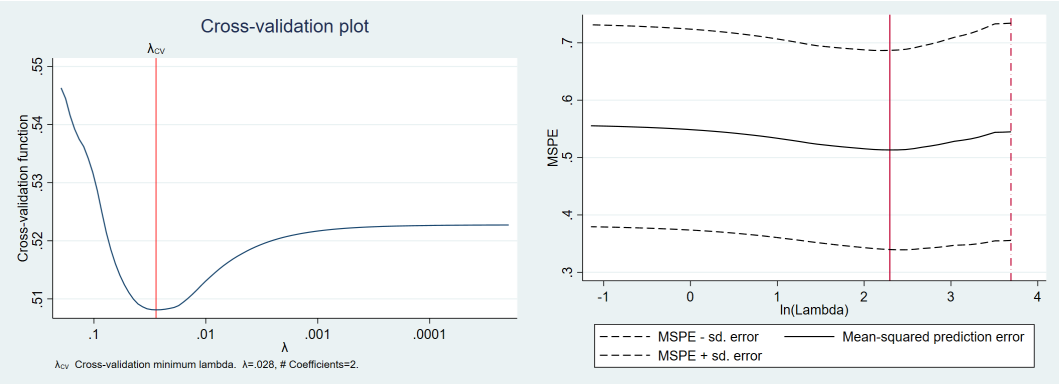
No. of observations	= 847
No. of covariates	= 6
No. of CV folds	= 10

ID of λ	Description	No. of nonzero β	Out-of-sample R-squared	CV mean prediction error
1	first lambda	0	-0.001	1.147
29	lambda before	4	0.140	0.986
* 30	selected lambda	4	0.140	0.986
31	lambda after	5	0.140	0.986
100	last lambda	6	0.137	0.990

* lambda selected by cross-validation.

ID of λ	ℓ1-Norm	R-squared	MSPE	Action
1	0.00	0.00	1.1468	Added _cons
2	0.03	0.02	1.1345	Added population
6	0.15	0.06	1.0860	Added income
9	0.27	0.09	1.0688	Added education
10	0.32	0.10	1.0626	Added area
31	0.80	0.17	1.0164	Added health

Similarly, the following figures and tables show the results of Lasso in terms of DCEs of **rural households**. The optimal λ (ID = 22) means there are only two important family demographics, the variables $population_i$ and $income_i$, which are prioritized in this order.



Lasso linear mode for Rural

		No. of observations	= 102
		No. of covariates	= 6
Selection: Cross-validation		No. of CV folds	= 10

ID of λ	Description	No. of nonzero β	Out-of-sample R-squared	CV mean prediction error
1	first lambda	0	-0.038	0.546
21	lambda before	2	0.034	0.508
* 22	selected lambda	2	0.034	0.508
23	lambda after	2	0.034	0.508
100	last lambda	6	0.007	0.523

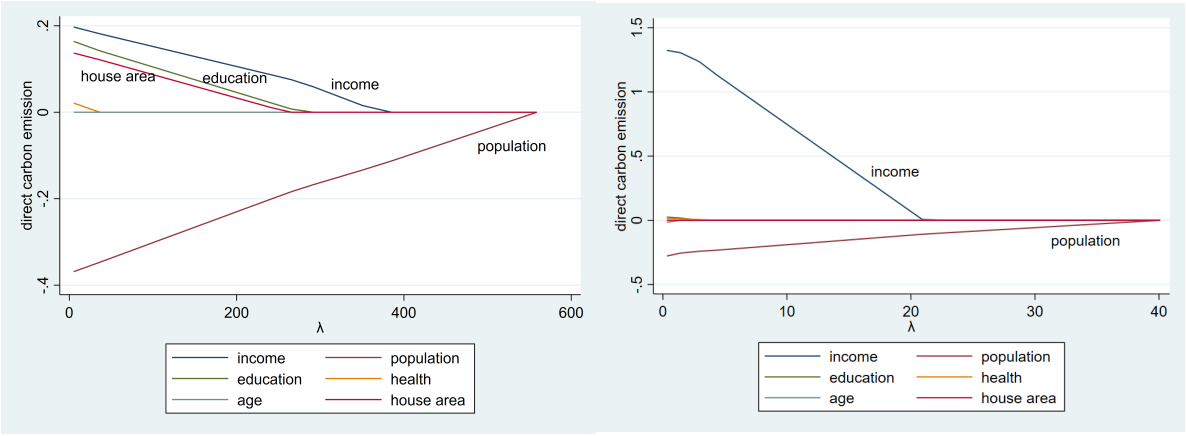
* lambda selected by cross-validation.

ID of λ	ℓ1-Norm	R-squared	MSPE	Action
1	0.00	0.00	0.5449	Added _cons
2	0.02	0.01	0.5443	Added population
8	0.12	0.05	0.5284	Added income
25	1.37	0.11	0.5234	Added health
30	1.50	0.11	0.5338	Added education
38	1.60	0.11	0.5461	Added age

4.3.2 Sorting the importance of household demographics

In the Lasso regression model, when the value of λ decreases, the constraint strength gradually decreases as well. This means that the sparse degree of the optimal solution matrix of the objective function also gradually decreases. Some coefficients of variables start to change from zero to non-zero, the variable with the first non-zero coefficient has the greatest impact on DCEs and the corresponding λ value is also the biggest.

The importance of each influencing factor to DCEs of urban (left) and rural (right) households are shown in the following figures. For **urban households**, the coefficients of the variable $population_i$ first become non-zero, meaning that the number of household members has the largest impact on DCEs. As the value of λ decreases, the coefficients of the variables $income_i$, $education_i$ and $area_i$ become non-zero. For **rural households**, the variable $population_i$ is the most important influencing factor for DCEs, and then the variables $income_i$. The results are in line with the tables above.



4.3.3 Determining regression coefficients

From the above analysis, observing the change in MSPE can identify the most important household driving factors of DCEs, and the λ value corresponding to the demographics gives the importance rank. Next, I apply Lasso regression to obtain the estimated coefficients under optimal λ.

In the following tables, I list the coefficients of household demographics on DECs of urban and rural households, under Lasso and OLS method, separately.

* Estimate Lasso with the lambda that minimizes MSPE (ID=30) for Urban

Selected	Coefficients of the Lasso	Post-estimation of the OLS
population	-0.3485	-0.3724
income	0.1826	0.1997
education	0.1436	0.1675
area	0.1224	0.1396
health	0	0.0247
age	-	-
_cons	0.0175	0.0136

* Estimate Lasso with the lambda that minimizes MSPE (ID=22) for Rural

Selected	Coefficients of the Lasso	Post-estimation of the OLS
income	0.7524	1.4273
population	-0.1907	-0.2651
education	-	-
area	-	-
health	-	-
age	-	-
_cons	-0.3148	-0.1365

In terms of DCEs, comparing urban and rural households, both $population_i$ and $income_i$ are significant influencing factors, with DCEs being positively correlated with income but negatively correlated with household's size. In addition, $education_i$ and $area_i$ also have a positive effect on DCEs for urban households, while these factors have no significant effect on rural households. A possible explanation is that the household's direct carbon emissions come from direct energy consumption, specifically the fuel consumption in the activities of transportation, cooking and heating. The relatively small household size and low household income limit the adoption of some appliances such as dishwashers and clothes dryers in China.

On the other hand, comparing the coefficients of Lasso and OLS method, the overall Lasso coefficients are smaller than those of OLS, reflecting the shrinkage performance of Lasso.

4.4 The code of application

The STATA **code** for empirical application is as follows:

```
********** the LASSO Regression of Application **********
use "Z:\Desktop\code.dta", clear
ssc install asdoc
ssc install lassopack
ssc install estout

* number of people in household *
drop pop
gen pop = a2000a + 1
* dependent variable *
gen per_direct    = E_direct_energy / pop
gen ln_per_direct = log(per_direct)
* independent variables *
gen per_income     = total_income_w / pop
gen ln_per_income  = log(per_income)
gen ln_pop         = log(pop)
gen ln_edu_mean    = log(edu_mean)
gen ln_health_mean = log(health_mean)
gen ln_age_mean    = log(age_mean)
gen ln_house_area  = log(house_area)

* descriptive statistics *
asdoc sum ln_per_direct ln_per_income ln_pop ln_edu_mean ln_health_mean ln_age_mean
ln_house_area

* plot direct carbon emissions pre person for Urban and Rural *
bysort region rural:egen direct_emission = mean(per_direct)
graph bar direct_emission if region==1, over(rural) scheme(s2mono)

*** LASSO on direct carbon emissions  ***
* standardisation *
egen direct_s = std(per_direct)
egen income_s = std(per_income)
egen pop_s    = std(pop)
egen edu_s    = std(edu_mean)
egen health_s = std(health_mean)
egen age_s    = std(age_mean)
egen area_s   = std(house_area)

* Urban *
lasso linear direct_s income_s pop_s edu_s health_s age_s area_s if rural==0, selection(cv,
alllambdas) stop(0) rseed(12345) nolog
estimates store cv
cvplot
lassocoef, display(coef,penalized) sort(coef,penalized)
lasso2  direct_s income_s pop_s edu_s health_s age_s area_s  if rural==0, plotpath(lambda)
cvlasso direct_s income_s pop_s edu_s health_s age_s area_s  if rural==0, lopt seed(123) plotcv

* Rural *
lasso linear direct_s income_s pop_s edu_s health_s age_s area_s if rural==1, selection(cv,
alllambdas) stop(0) rseed(12345) nolog
estimates store cv
cvplot
lassocoef, display(coef,penalized) sort(coef,penalized)
```

```
lasso2  direct_s income_s pop_s edu_s health_s age_s area_s  if rural==1, plotpath(lambda)
cvlasso direct_s income_s pop_s edu_s health_s age_s area_s  if rural==1 , lopt seed(123) plotcv
```

5 Conclusion

This project aims to study the variable selection property of Lasso by simulation and application.

In the simulation part, I do the Data Generation Process and look at two cases to examine the relation of consistency of variable selection and the Strong Irrepresentable Condition. Case 1 shows that when the Strong Irrepresentable Condition holds, Lasso is sign consistent. Case 2 shows that in 100 X designs and 1000 repetitions of the simulation for each design, when the Strong Irrepresentable Condition holds, the percentage of Lasso selecting the correct model is close to 1. The simulation case 1 in my project is successfully replicating the simulation example 1 of Zhao and Yu (2006), although the result of the simulation case 2 has some minor variations from their example 2. These two cases show the relation of the Strong Irrepresentable Condition to Lasso's variable selection.

In the application part, I use survey data to empirically study the variable selection property of Lasso. By measuring the direct carbon emissions from urban and rural household energy consumptions in China, Lasso is used to analyse the differences in the effects of household income level, population size, average age of households, average years of education of households, health status of households, and living area on the direct carbon emissions of urban and rural households. I obtain the following conclusions: the direct carbon emissions of urban households are higher than rural households. Both urban and rural households' direct carbon emissions are most influenced by the household size, followed by the household income. Direct carbon emissions of urban households are also influenced by the living area and average years of schooling, while rural households are not influenced by other factors.

References

Bin, S. and Dowlatabadi, H., 2005. Consumer lifestyle approach to US energy use and the related CO2 emissions. *Energy policy*, *33*(2), pp.197-208.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. *An introduction to statistical learning*. New York: springer.

Qu, J., Zeng, J., Li, Y., Wang, Q., Maraseni, T., Zhang, L., Zhang, Z. and Clarke-Sather, A., 2013. Household carbon dioxide emissions from peasants and herdsmen in northwestern arid-alpine regions, China. *Energy Policy*, *57*, pp.133-140.

Shi, X., Wang, K., Cheong, T.S. and Zhang, H., 2020. Prioritizing driving factors of household carbon emissions: An application of the LASSO model with survey data. *Energy Economics*, *92*, p.104942.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), pp.267-288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), pp.273-282.

Zhao, P. and Yu, B., 2006. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, *7*, pp.2541-2563.