2023 Lima Summer School in Economics
Machine Learning for Economics

Vadim Marmer

Version: March 1, 2023.
Vancouver School of Economics, UBC.

Contents

CHAPTER 1

Recap and limitations of traditional methods

In this chapter, we review the basics of the linear regression model, OLS estimation, the instrumental variables (IVs) regression model, and 2SLS estimation. In the case of OLS, we will focus on the effect of many covariates on the variance of the OLS estimator. We consider a typical scenario where there are some important regressors, which are the main focus of a study, and there are many potential controls. The econometrician is unsure which of the controls should be included. Omitting important controls results in biased estimates for the main parameters when the main regressors and controls are correlated. However, including too many controls may lead to very imprecise estimates for the main coefficients (excessively large standard errors). This motivates the need for automatic or machine learning (ML) methods for selecting of controls.

In the case of IV estimation, we will show that the 2SLS estimator is inconsistent when there are too many IVs. Since dropping important IVs would reduce the efficiency (precision) of the 2SLS estimator, we again face the problem of choosing among many potential problems.

We also discuss the difficulties associated with "traditional" hypothesis-testing-based (i.e. statistical-significance-based) practices of choosing variables, which further motivates the need for machine-learning-based methods.

1.1. Linear Regression and OLS

The researcher observes data on the dependent variable Y_i and the k explanatory variables $X_{i,1}, \ldots, X_{i,k}$. The index i denotes individual observations, and the sample size is n: $i = 1, \ldots, n$. A classical linear regression models the conditional mean of Y_i given the regressors: for all $i = 1, \ldots, n$,

$$E(Y_i \mid X_{i,1}, \ldots, X_{i,k}) = \beta_1 X_{i,1} + \ldots + \beta_k X_{i,k},$$

where $\beta_1, \ldots, \beta_k$ are the unknown regression coefficients to be estimated. The intercept is included by allowing one of the regressors to take the value one for all observations (the corresponding β is the intercept). Equivalently, we can write

$$Y_i = \beta_1 X_{i,1} + \ldots + \beta_k X_{i,k} + U_i,$$

where

$$E(U_i \mid X_{i,1}, \ldots, X_{i,k}) = 0.$$

The residual terms U_i's capture the effect of unobserved factors on Y_i. The classical linear regression model also assumes that for all $i = 1, \ldots, n$,

$$E(U_i^2 \mid X_{i,1}, \ldots, X_{i,k}) = \sigma^2,$$

and that for $i \neq j$,

$$E(U_i U_j \mid X_{i,1}, \ldots, X_{i,k}) = 0.$$

It is convenient to switch to the matrix notation. Define the $n \times k$ matrix of observations on the regressors:

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & \ldots & X_{1,k} \\ X_{2,1} & X_{2,2} & \ldots & X_{2,k} \\ \ldots & \ldots & \ldots & \ldots \\ X_{n,1} & X_{n,2} & \ldots & X_{n,k} \end{pmatrix}.$$

Note that the rows of X represent different observations, and the columns represent different regressors. Thus, the element i, j of X is observation i on the j-th regressor. To rule out multicollinearity, we assume that the $n \times k$ matrix of regressors X has a full column rank:

$$\text{rank}(X) = k.$$

The last condition implies that there is no exact linear combination among the k columns of X (the k regressors).

Similarly, let an $n \times 1$ vector Y collect the n observations on the dependent variable:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},$$

and U collect the n observations on the residuals:

$$U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}.$$

Lastly, let the $k \times 1$ vector β collect the unknown regression coefficients:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Note that $X\beta$ is an $n \times 1$ vector of the predicted values of Y given X. Assuming that observations are independent, the model can now be stated as

$$Y = X\beta + U, \tag{1.1.1}$$

$$E(U \mid X) = 0, \tag{1.1.2}$$

$$Var(U \mid X) = \sigma^2 I_n. \tag{1.1.3}$$

(See Appendix 1.7 for the definition and properties of the variance of a vector.)

For an $n \times 1$ vector y, its Euclidean norm is given by

$$\|y\| = \sqrt{y_1^2 + \ldots + y_n^2}.$$

Hence, the squared distance between two n-vectors y and $\hat{y}$ can be written as

$$\|y - \hat{y}\|^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Note that for a $k \times 1$ vector b, Xb is $n \times 1$.

The ordinary least squares (OLS) estimator of β is obtained by solving

$$\min_{b \in \mathbb{R}^k} \|Y - Xb\|^2, \tag{1.1.4}$$

where

$$\begin{aligned} \|Y - Xb\|^2 &= \sum_{i=1}^{n} (Y_i - X_{i,1} b_1 - \ldots - X_{i,k} b_k)^2 \\ &= \sum_{i=1}^{n} (Y_i - X_i' b)^2, \end{aligned}$$

where X_i is the $k \times 1$ vector collecting the i-th observations on all k regressors:

$$X_i = \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ \vdots \\ X_{i,k} \end{pmatrix}.$$

Let $\hat{\beta}$ denote the OLS estimator, i.e. $\hat{\beta}$ is the solution to the least squares problem in (1.1.4). The first-order condition for the least squares problem is given by

$$\begin{aligned} 0 &= \sum_{i=1}^{n} X_i (Y_i - X_i' \hat{\beta}) \\ &= X'(Y - X\hat{\beta}), \end{aligned}$$

which implies that

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

Define the fitted/estimated residuals

$$\hat{U}_i = Y_i - X_i' \hat{\beta},$$

and collect them into an $n \times 1$ vector

$$\hat{U} = \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_n \end{pmatrix}.$$

We choose the OLS estimator $\hat{\beta}$ so that the resulting fitted residuals are orthogonal to the regressors:

$$\begin{aligned} 0 &= X'\hat{U} \\ &= \sum_{i=1}^{n} X_i \hat{U}_i, \end{aligned}$$

which can be seen from the definition of $\hat{U}$ and the first-order conditions for the least squares problem.

The properties of the OLS estimator are summarized below.

Proposition 1.1.1.

(a) Suppose that (1.1.1) and (1.1.2) hold. Then $\hat{\beta}$ is unbiased:

$$E(\hat{\beta} \mid X) = \beta.$$

(b) Suppose that (1.1.1)-(1.1.3) hold. Then

$$Var(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}.$$

PROOF. For part (a),

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + U) \\ &= \beta + (X'X)^{-1}X'U, \end{aligned} \tag{1.1.5}$$

where the second equality is by (1.1.1). The result follows since

$$\begin{aligned} E(\hat{\beta} \mid X) &= \beta + (X'X)^{-1}X' \cdot E(U \mid X) \\ &= \beta + (X'X)^{-1}X' \cdot 0 \\ &= \beta, \end{aligned} \tag{1.1.6}$$

where the second equality holds since $E(U \mid X) = 0$ by (1.1.2).

For part (b), by (1.1.5):

$$\begin{aligned} Var(\hat{\beta} \mid X) &= Var(\beta + (X'X)^{-1}X'U \mid X) \\ &= (X'X)^{-1}X'Var(U \mid X)X(X'X)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}, \end{aligned}$$

where the first equality holds by (1.1.5), the second equality holds by the properties of the variance,[1] and the third equality holds by (1.1.3). □

The unbiasedness property in Proposition 1.1.1(a) requires that we control for all regressors included in the model (have non-zero β's) unless they are orthogonal. To see this, let us

[1] See Appendix 1.7, Proposition 1.7.1.

partition the model as

$$Y = X_1\beta_1 + X_2\beta_2 + U, \tag{1.1.7}$$

where X_1 is $n \times k_1$ and contains the first k_1 columns of X, X_2 is $n \times k_2$ and contains the last k_2 columns of X, and $k_1 + k_2 = k$:

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}.$$

Similarly, we partition

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where β_1 is $k_1 \times 1$ and β_2 is $k_2 \times 1$. We can view X_1 as the main regressors and X_2 as potential controls or covariates.

Suppose the researcher estimates β_1 by regressing Y only on X_1 without including X_2 (we also say without controlling for X_2). Let $\tilde{\beta}_1$ denote the resulting estimator:

$$\begin{aligned}\tilde{\beta}_1 &= (X_1'X_1)^{-1}X_1'Y \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'U.\end{aligned}$$

We have:

$$E\left(\tilde{\beta}_1 \mid X\right) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2,$$

where $(X_1'X_1)^{-1}X_1'X_2\beta_2$ is the *bias* term. Thus, unless $\beta_2 = 0$ (there is no need to control for X_2) or $X_1'X_2 = 0$ (the regressors in X_1 are orthogonal, i.e. unrelated, to X_2) the estimator $\tilde{\beta}_1$ is *biased*: it is "contaminated" by the effect of X_2 on Y as captured by β_2 and the relationship between X_1 and X_2.

1.2. The effect of covariates on the variance of the OLS estimator

In practical applications, researchers often have long lists of potential controls X_2. See for example the discussion of the cross-country growth regression model in Belloni and Chernozhukov (2011, Example 3). The cross-country growth regression model is concerned with estimating the effect of initial conditions (initial GDP) on future growth rates. There is a long list of additional potential controls related to the initial GDP that may also affect future growth rates. The list includes variables describing institutions and technological factors, and overall there are about 60 potential covariates, while the sample size is about 90 observations. While only a few of the potential covariates may have non-zero coefficients in the true model, unfortunately, we do not know which ones.

To avoid the omitted variables bias, the researcher may attempt to include all potential controls. Unfortunately, that results in large variances and standard errors on the main parameters of interest as we discuss next.

To isolate the variance of one of the elements of the vector $\hat{\beta}$, we need the following result.

Proposition 1.2.1. *Consider the partitioned regression model (1.1.7).*

(a) The OLS estimator of β_1 in the regression of Y against X_1 and X_2 is given by

$$\hat{\beta}_1 = (X_1' M_2 X_1)^{-1} X_1' M_2 Y,$$

where M_2 is an $n \times n$ orthogonal projection matrix:[2]

$$M_2 = I_n - X_2(X_2'X_2)^{-1}X_2'.$$

(b) Suppose that (1.1.1)-(1.1.3) hold. Then,

$$Var(\hat{\beta}_1 \mid X) = \sigma^2 (X_1' M_2 X_1)^{-1}.$$

Remark. The result in part (a) of the proposition is important on its own. Some ML methods that we will be discussing later in the course rely on it. As we discuss below, the result implies that $\hat{\beta}_1$ can be obtained by first regressing X_1 against the other regressors X_2, saving the residuals; then regressing Y against X_2 and saving the residuals, and then lastly regressing the residuals of Y against the first-step residuals of X_1.

PROOF. For part (a), first note that by the properties of the projection matrix M_2,[3]

$$M_2 X_2 = 0.$$

Recall that by construction,

$$X'\hat{U} = 0, \tag{1.2.1}$$

where

$$\hat{U} = Y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2.$$

Re-write the last equation as

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{U}.$$

We now have:

$$\begin{aligned}(X_1' M_2 X_1)^{-1} X_1' M_2 Y &= (X_1' M_2 X_1)^{-1} X_1' M_2 \left(X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{U}\right) \\ &= \hat{\beta}_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 \left(X_2\hat{\beta}_2 + \hat{U}\right) \\ &= \hat{\beta}_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 \hat{U}.\end{aligned}$$

Next,

$$\begin{aligned}M_2\hat{U} &= \hat{U} - X_2(X_2'X_2)^{-1}X_2'\hat{U} \\ &= \hat{U},\end{aligned}$$

where the second equality holds since $X_2'\hat{U} = 0$ by (1.2.1). Similarly,

$$X_1'\hat{U} = 0,$$

and the result in part (a) follows.

[2] See the discussion of projection matrices in Appendix 1.8.
[3] See Proposition 1.8.1(c) in Appendix 1.8.

To prove part (b), note that

$$\hat{\beta}_1 = \beta_1 + (X_1'M_2X_1)^{-1}X_1'M_2U.$$

Note also that the projection matrix M_2 is symmetric and idempotent:[4]

$$M_2' = M_2,$$
$$M_2M_2 = M_2.$$

Using these facts, we obtain:

$$\begin{aligned} Var(\hat{\beta}_1 \mid X) &= (X_1'M_2X_1)^{-1}X_1'M_2Var(U \mid X)M_2'X_1(X_1'M_2X_1)^{-1} \\ &= \sigma^2(X_1'M_2X_1)^{-1}X_1'M_2X_1(X_1'M_2X_1)^{-1} \\ &= \sigma^2(X_1'M_2X_1)^{-1}. \end{aligned}$$

□

Let us partition the regression model as follows:

$$Y = \beta_1X_1 + X_2\beta_2 + U,$$

where X_1 is $n \times 1$ and contains the observations on the main regressor of interest (the initial GDP in the cross-country growth regression model), β_1 is a scalar coefficient on the main regressor, and X_2 is $n \times (k-1)$ and includes observations on the potential controls. In view of Proposition 1.2.1(b), the variance of the OLS estimator of β_1 while controlling for the covariates X_2 is given by[5]

$$Var(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{X_1'M_2X_1}.$$

Let us discuss the effect of adding more controls into X_2, i.e. the expression in the denominator of $Var(\hat{\beta}_1 \mid X)$. Since $M_2 = M_2'$ and $M_2M_2 = M_2$,

$$X_1'M_2X_1 = X_1'M_2M_2X_1 = X_1'M_2'M_2X_1 = \tilde{X}_1'\tilde{X}_1,$$

where

$$\tilde{X}_1 = M_2X_1 = X_1 - X_2(X_2'X_2)^{-1}X_2'X_1 = X_1 - X_2\hat{\gamma},$$

and $\hat{\gamma}$ is the OLS coefficient from the regression of X_1 against X_2. Hence, $\tilde{X}_1$ is the matrix of residuals from the OLS regression of X_1 against X_2, and $\tilde{X}_1'\tilde{X}_1$ is the sum of the squared residuals:

$$\tilde{X}_1'\tilde{X}_1 = \sum_{i=1}^{n}\tilde{X}_{i,1}^2,$$

and we can write

$$Var(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{\sum_{i=1}^n \tilde{X}_{i,1}^2} = \frac{\sigma^2}{\sum_{i=1}^n (X_{i,1} - X_{i,2}'\hat{\gamma})^2}.$$

Thus, the denominator of $Var(\hat{\beta}_1 \mid X)$ contains the residual sample variation of the main regressor of interest X_1 after removing from it everything linearly related (explainable) by

[4] See Proposition 1.8.1(a),(e).

[5] Note that $\hat{\beta}_1$ is an estimator of a scalar parameter, and therefore, its variance is a scalar.

X_2. When we include more controls into X_2, a larger portion of the variation of X_1 is removed resulting in a smaller sample residual variation $\sum_{i=1}^n \tilde{X}_{i,1}^2$.[6]

The above discussion shows that when we include unnecessary controls into X_2 that are correlated with X_1, the variance of the OLS estimator of β_1 increases. As a result, the estimates of the main parameter of interest become noisier. In practice, one would tend to see larger standard errors for $\hat{\beta}_1$, smaller t-statistics and larger p-values, and wider confidence intervals for β_1. When the number of potential controls is large and they are highly correlated with the main variable of interest, the impact of including unnecessary controls on the informativeness and significance of the main estimates can be drastic.

1.3. Including only significant covariates

To address the issue of many potential controls described in the previous section, one might consider including only covariates with statistically significant coefficients. Unfortunately, due to the nature of hypothesis testing, such a practice would typically result in exclusion of relevant controls and the omitted variables bias. The reason for that is that inference procedures do not control the probability of Type II errors (not rejecting a null hypothesis when it is false) and as a result the probability of seeing insignificant coefficients when in fact the true parameters are different from zero can be large. We illustrate this point below using a simple stylized example.

Suppose $\hat{\theta}$ is an estimator for a scalar parameter θ, and

$$\hat{\theta} \sim N(\theta, \omega^2).$$

Suppose further that the variance ω^2 is known. Consider testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. Since

$$\frac{\hat{\theta} - \theta}{\omega} \sim N(0, 1),$$

a size α test rejects H_0 in favor of H_1 when

$$\left| \frac{\hat{\theta}}{\omega} \right| > z_{1-\alpha/2},$$

where z_τ denotes the τ-th quantile of the standard normal distribution. The probability of Type I error (rejecting H_0 when it is true) is controlled as under H_0, $\hat{\theta}/\omega \sim N(0, 1)$ and therefore $P(|\hat{\theta}/\omega| > z_{1-\alpha/2} \mid \theta = 0) = \alpha$.

To consider the probability of Type II error, write

$$\frac{\hat{\theta}}{\omega} = \frac{\hat{\theta} - \theta}{\omega} + \frac{\theta}{\omega} = Z + \frac{\theta}{\omega},$$

where

$$Z \sim N(0, 1).$$

Note that since we are now under H_1, $\theta \neq 0$. The probability of deciding that $\theta \neq 0$ (significant) is given by

$$P\left(|Z + \theta/\omega| > z_{1-\alpha/2}\right).$$

[6] See Proposition 1.8.2 in the Appendix.

Note that this probability converges to $1-\alpha$ as $\theta \to 0$. I.e. in this case, the probability of Type II error can be as large as $1-\alpha$. If the value θ/ω is not very large, the probability of detecting $\theta \neq 0$ can be relatively small.

The above example illustrates that a failure to reject $H_0 : \theta = 0$ cannot be used as reliable evidence that the true coefficient is zero. In the context of regression, dropping insignificant regressors can lead to the omitted variables bias.

1.4. Data snooping

Data snooping (also known as p-hacking) occurs when the researcher repeatedly re-uses the same data in order to produce "statistically significant" estimates with large t-statistics or small p-values (which generated the name "p-hacking"). This is typically done by tweaking the specification of a model numerous times, adjusting the definitions of the variables, excluding some observations, and etc until sufficiently small p-values are obtained for the main parameters of interest. Data snooping/p-hacking destroys the validity of t-statistics and p-values, and leads to false discoveries. We illustrate the issue using a simple example.

Suppose the researcher wants to show that data supports their hypothesis that some scalar parameter θ is different from zero. Suppose that the researcher can construct J *independent* estimators for θ such that

$$\hat{\theta}_j \sim N(\theta, \omega_j^2),$$

where ω_j^2 are known. To demonstrate the "significance" of θ, the researcher conducts J size α tests of $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$, each test based on different θ_j, until they find a test with $|\hat{\theta}_j/\omega_j| > z_{1-\alpha/2}$ or run out of tests. Suppose in fact $\theta = 0$. In this case, the probability of concluding that θ is "significant" is given by

$$\begin{aligned}
P\left(\max_{1\leq j\leq J}\left|\frac{\hat{\theta}_j}{\omega_j}\right| > z_{1-\alpha/2}\right) &= 1 - P\left(\max_{1\leq j\leq J}\left|\frac{\hat{\theta}_j}{\omega_j}\right| \leq z_{1-\alpha/2}\right) \\
&= 1 - P\left(\left|\frac{\hat{\theta}_1}{\omega_1}\right| \leq z_{1-\alpha/2}, \ldots, \left|\frac{\hat{\theta}_J}{\omega_J}\right| \leq z_{1-\alpha/2}\right) \\
&= 1 - \prod_{j=1}^{J} P\left(\left|\frac{\hat{\theta}_j}{\omega_j}\right| \leq z_{1-\alpha/2}\right) \\
&= 1 - (1-\alpha)^J,
\end{aligned}$$

where the second equality holds because the maximum of J statistics is below the critical value if and only if each of the J statistics is below the critical value. The third equality holds by the independence of $\hat{\theta}_j$'s across j, and the last equality holds because each of the J tests has size α:

$$P\left(\left|\frac{\hat{\theta}_j}{\omega_j}\right| > z_{1-\alpha/2}\right) = \alpha.$$

Hence, when the J tests are independent and $\theta = 0$, the probability of falsely concluding that $\theta \neq 0$ is given by

$$1 - (1-\alpha)^J.$$

TABLE 1. The probability of false discovery that $\theta \neq 0$ with J independent size $\alpha = 0.05$ tests

J	1	5	10	15	30	60	85
Prob. of false discovery	0.05	0.23	0.40	0.54	0.79	0.95	0.99

Since $0 < 1 - \alpha < 1$, the probability of falsely concluding that $\theta \neq 0$ quickly grows with J as illustrated in Table 1 for $\alpha = 0.05$. For example, with 10 independent tests, the probability of falsely concluding that $\theta \neq 0$ is 40%, and escalates to almost 80% with 30 tests. With 85 tests, the researcher is almost assured to falsely conclude that $\theta \neq 0$.

While in practice tests are rarely independent, the same relationship holds qualitatively. Since by design, each test comes with the probability α of making a false discovery (or the Type I error), when the researcher performs many of such tests the probabilities of Type I error quickly accumulate. Thus by the design of statistical tests, if the researcher searches long enough, with a high probability they would find something that is not actually there.

The nature of empirical research is such that one cannot know the right specification until they start working with data. Since the search for correct specifications is unavoidable and in view of the dangers of data snooping, there is a great need for ML procedures that automatically detect correct specifications in a data-driven way.

1.5. Instrumental variables (IV) regression

In many economic applications the assumption that regressors are exogenous is implausible:

$$Y = X\beta + U, \quad \beta \in \mathbb{R}^k, \tag{1.5.1}$$

but

$$E(U \mid X) \neq 0.$$

The last equation immediately implies that the OLS estimator of β is biased,[7] and therefore one should use an alternative estimation strategy.

One example of such a model is the Mincer earnings regression, where the dependent variable is the log wage, the main regressor of interest in X is years of schooling, while other regressors included in X are a gender dummy, years of experience, etc. The residual term U is often interpreted as the unobserved ability. Since individuals with higher ability typically self-select to obtain more education, the education variable predicts an individual's ability:

$$E(\text{ability} \mid \text{education}) \neq 0.$$

Suppose that in addition, the researcher observes an $n \times l$ matrix of IVs such that

$$E(U \mid Z) = 0.$$

[7] See equation (1.1.6) in the proof of Proposition 1.1.1(a).

Thus, Z contains data on l exogenous variables (in the sense that they are not related to U). We also assume that Z are related to X through the so-called first-stage equation:

$$X = Z\Pi + V, \tag{1.5.2}$$
$$E(V \mid Z) = 0.$$

In the first stage equation, the matrix of coefficients Π is $l \times k$. We assume that the l IVs in Z are sufficiently informative about the k regressors in X in the sense that

$$\operatorname{rank}(\Pi) = k.$$

The last condition implies that $l \geq k$. Note also that the regressors in X that are exogenous can and should be included in Z.

Given the first-stage equation and since $E(U \mid Z) = 0$, the regressors in X can be endogenous only because of the correlation between U and V.

For the Mincer earning regression, Angrist and Krueger (1991) proposed to use the quarter of birth dummy variables as instruments for schooling. To justify the choice, they argue that while the quarter of birth is exogenously assigned, it predicts education due to compulsory schooling laws. In the US, education is mandatory until the age of 16 in most states. Since people born in the first quarter reach 16 before those born later in the same year, they tend to have slightly less education.

Combining the first-stage equation in (1.5.2) with the main regression equation in (1.5.1), we obtain:

$$\begin{aligned} Y &= (Z\Pi + V)\beta + U \\ &= (Z\Pi)\beta + (U + V\beta) \\ &= (Z\Pi)\beta + \epsilon, \end{aligned}$$

where

$$\epsilon = U + V\beta.$$

Note that

$$E(\epsilon \mid Z) = 0,$$

and therefore one can obtain an unbiased estimator of β from the OLS regression of Y against $Z\Pi$. Since Π is unknown, in practice it is replaced by its estimator from the first stage:

$$\hat{\Pi} = (Z'Z)^{-1}Z'X. \tag{1.5.3}$$

The IV or two-stage least squares (2SLS) estimator of β is given by

$$\begin{aligned} \hat{\beta} &= \left((Z\hat{\Pi})'(Z\hat{\Pi})\right)^{-1}(Z\hat{\Pi})'Y \\ &= \left(\hat{\Pi}'Z'Z\hat{\Pi}\right)^{-1}\hat{\Pi}'Z'Y \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}Z'Y, \end{aligned} \tag{1.5.4}$$

where the last equality holds by substituting the expression for $\hat{\Pi}$ from (1.5.3) into the second line.

Define an $n \times n$ projection matrix[8]

$$P_Z = Z(Z'Z)^{-1}Z'.$$

The 2SLS estimator can be written as

$$\hat{\beta} = (X'P_Z X)^{-1}X'P_Z Y.$$

When the number of IVs is the same as the number of regressors ($l = k$), the 2SLS estimator simplifies to

$$\hat{\beta} = (Z'X)^{-1}Z'Y, \tag{1.5.5}$$

which is often referred to as the IV estimator. The result holds by (1.5.4) and because the matrix $Z'X$ is now square ($k \times k$) and $(Z'X)^{-1}$ exists:

$$\begin{aligned}\hat{\beta} &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (Z'X)^{-1}Z'Y.\end{aligned}$$

We can denote

$$\hat{X} = P_Z X = Z\hat{\Pi},$$

i.e. $\hat{X}$ contains the exogenous variation in X due to Z. Note that by construction, $\hat{X}$ is $n \times k$. Thus, whether $l = k$ or $l > k$, the 2SLS estimator can be re-written as

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

Hence, $\hat{X}$ can be viewed as k IVs for X: the 2SLS estimator is an IV estimator that uses the first-stage predicted values of X as IVs. To make the 2SLS estimator more efficient, one would like to include all relevant IV variables into Z to capture all available exogenous variation in X.

1.6. IV regression with many IVs

In this section, we discuss the bias of the 2SLS estimator. Since the first-stage matrix Π must be replaced with its estimates, the 2SLS estimator is typically biased. However, its biased disappears as the sample size $n \to \infty$. I.e. in large enough samples, the bias of the 2SLS estimator is negligible. However, the situation changes drastically when the number of IVs is very large and of a similar order as n. In such cases, the bias of the 2SLS estimator does not disappear even in large samples, which makes it inconsistent.

The many IVs scenario can easily arise in practice when $E(U \mid Z) = 0$, and the econometrician suspects a non-linear relationship between X and Z's. In that case, the econometrician may try to gain more efficiency by including the polynomial and interaction terms on the right-hand side of the first-stage equation. By including polynomial terms of higher orders

[8]See a discussion of projection matrices in Appendix 19.

and many interactions, the number of the right-hand side variables can grow fast. For example, Angrist and Krueger (1991) generated many IVs by interacting the quarter of birth dummies with exogenous regressors in X.

To simplify the presentation, suppose $k = 1$, i.e. there is only on regressor. In that case, the 2SLS estimator satisfies

$$\hat{\beta} = \frac{X'P_Z Y}{X'P_Z X} = \beta + \frac{n^{-1}X'P_Z U}{n^{-1}X'P_Z X}, \tag{1.6.1}$$

and, hence, the bias of the 2SLS estimator depends on $n^{-1}E[X'P_Z U]$.

Proposition 1.6.1.

(a) Suppose that the covariance between U and V is given by

$$E\left(UV' \mid Z\right) = \sigma_{UV} I_n,$$

for some scalar σ_{UV}. Then

$$E\left(\frac{1}{n}X'P_Z U \mid Z\right) = \sigma_{UV}\frac{l}{n}.$$

(b) Suppose in addition that the variance of V is given by

$$E\left(VV' \mid Z\right) = \sigma_V^2 I_n,$$

for some scalar $\sigma_V^2 > 0$. Then

$$E\left(\frac{1}{n}X'P_Z X \mid Z\right) = \frac{1}{n}\Pi' Z'Z\Pi + \sigma_V^2\frac{l}{n}.$$

Proof. For part (a), since $X'P_Z'U$ is a scalar (as X and U are n vectors),

$$\begin{aligned} X'P_Z'U &= \operatorname{tr}\left(X'P_Z'U\right) \\ &= \operatorname{tr}(P_Z UX'), \end{aligned}$$

where the second equality holds by the properties of the trace: $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. Next,

$$E\left(\operatorname{tr}(P_Z UX') \mid Z\right) = \operatorname{tr}\left(P_Z E\left(UX' \mid Z\right)\right),$$

where

$$\begin{aligned} E\left(UX' \mid Z\right) &= E\left(U\Pi' Z' + UV' \mid Z\right) \\ &= E\left(U \mid Z\right)\Pi' Z' + E\left(UV' \mid Z\right) \\ &= \sigma_{UV} I_n, \end{aligned}$$

where the last equality follows because $E(U \mid Z) = 0$. Combining these results, we have:

$$\begin{aligned} E\left(\frac{1}{n}X'P_Z'U \mid Z\right) &= \frac{1}{n}\text{tr}(P_Z\sigma_{UV}I_n) \\ &= \sigma_{UV}\frac{\text{tr}(P_Z)}{n} \\ &= \sigma_{UV}\frac{\text{tr}(Z(Z'Z)^{-1}Z')}{n} \\ &= \sigma_{UV}\frac{\text{tr}((Z'Z)^{-1}Z'Z)}{n} \\ &= \sigma_{UV}\frac{\text{tr}(I_l)}{n} \\ &= \sigma_{UV}\frac{l}{n}. \end{aligned}$$

The result in part (b) holds by the same arguments. □

Note that

$$\frac{1}{n}\Pi'Z'Z\Pi = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{l}(Z_{i,j}\Pi_j)^2, \text{ and}$$

$$E(X_i^2 \mid Z_i) = \sum_{j=1}^{l}(Z_{i,j}\Pi_j)^2.$$

Thus, the first term in the expression for the denominator (in Part (b) of the proposition) measures the exogenous variation of the endogenous regressor X_i, i.e. the variation due to $Z_i'\Pi$. For the variance of X_i to remain finite as $l \to \infty$, we need to assume that for all values z of Z_i,

$$\lim_{l\to\infty}\sum_{j=1}^{l}(z_j\Pi_j)^2 \leq K < \infty. \tag{1.6.2}$$

The condition holds, for example, if $Z_{i,j}$'s are bounded and only a few $\Pi_j \neq 0$. Such models are called sparse: many of $Z_{i,j}$'s are irrelevant. Alternatively, the condition holds when $\Pi_j \to 0$ sufficiently fast. In that case, most of the IVs $Z_{i,j}$ can be viewed as weak.

The result of Proposition 1.6.1 shows that the bias of the 2SLS estimator depends on the covariance between the second-stage errors U and first-stage errors V, and the ratio of the number of IVs to the sample size. When X is endogenous,

$$\sigma_{UV} \neq 0.$$

However, if the number of IVs is fixed (small)

$$\sigma_{UV}\frac{l}{n} \to 0$$

as $n \to \infty$. On the other hand, when the number of IVs is large and comparable to the sample size, it is more appropriate to model it as

$$l = l_n,$$

where

$$l_n \to \infty,$$
$$\frac{l_n}{n} \to c > 0,$$

as $n \to \infty$. In such cases,

$$\sigma_{UV} \frac{l_n}{n} \to \sigma_{UV} c,$$

and, therefore, the bias term is non-negligible even in very large samples (we assume that (1.6.2 holds).

When l is small, adding a few extra IVs can improve the performance of the 2SLS estimator. The situation is drastically different when l/n is large. In such cases, the researcher needs to be able to pick a small subset of the best IVs out of a long list of potential instruments. Hence, there is a need for a procedure that automatically selects the "best" IVs in a data-driven manner.

1.7. Appendix: The variance of a random vector

Let X be an $n \times 1$ vector of random variables:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

Its expectation is defined as a vector (matrix) composed of expected values of its corresponding elements:

$$E(X) = E \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

The variance-covariance matrix of a random n-vector is a $n \times n$ matrix defined as

$$\begin{aligned}
Var(X) &= E\left(X - EX\right)\left(X - EX\right)' \\
&= E\begin{pmatrix} X_1 - EX_1 \\ \vdots \\ X_n - EX_n \end{pmatrix}\begin{pmatrix} X_1 - EX_1 & \dots & X_n - EX_n \end{pmatrix} \\
&= \begin{pmatrix} E\left(X_1 - EX_1\right)\left(X_1 - EX_1\right) & \dots & E\left(X_1 - EX_1\right)\left(X_n - EX_n\right) \\ \dots & \dots & \dots \\ E\left(X_n - EX_n\right)\left(X_1 - EX_1\right) & \dots & E\left(X_n - EX_n\right)\left(X_n - EX_n\right) \end{pmatrix} \\
&= \begin{pmatrix} Var\left(X_1\right) & \dots & Cov\left(X_1, X_n\right) \\ \dots & \dots & \dots \\ Cov\left(X_n, X_1\right) & \dots & Var\left(X_n\right) \end{pmatrix}.
\end{aligned}$$

It is a symmetric matrix with variances on the main diagonal and covariances off the main diagonal. The symmetry follows from the fact that for two random variables X_i and X_j, $Cov(X_i, X_j) = Cov(X_j, X_i)$. The variance-covariance matrix is positive semi-definite (denoted by $Var(X) \geq 0$), since for any n-vector of constants a, we have that $a'Var(X)a \geq 0$:

$$\begin{aligned}
a'Var(X)a &= a'E\left(X - EX\right)\left(X - EX\right)'a \\
&= Ea'\left(X - EX\right)\left(X - EX\right)'a \\
&= E\left((X - EX)'a\right)^2 \\
&\geq 0.
\end{aligned}$$

Proposition 1.7.1. *Suppose $Y = \alpha + \Gamma X$, where $\alpha \in R^k$ is a fixed (non-random) vector and Γ is a $k \times n$ fixed matrix, then $Var(Y) = \Gamma(Var(X))\Gamma'$.*

Proof. First, note that

$$Y - EY = \Gamma(X - EX)$$

By the definition of the variance-covariance matrix,

$$\begin{aligned}
Var(Y) &= E(Y - EY)(Y - EY)' \\
&= \Gamma(X - EX)(X - EX)'\Gamma' \\
&= \Gamma Var(X)\Gamma'.
\end{aligned}$$

□

1.8. Appendix: Projection matrices

Let X be $n \times k$ with $\text{rank}(X) = k$. The first projection matrix we consider is

$$P_X = X(X'X)^{-1}X'.$$

Note that P is $n \times n$ by construction. Let Y be an $n \times 1$ vector, and consider the projection of Y using P:

$$\hat{Y} = P_X Y = X(X'X)^{-1}X'Y = X\hat{\beta},$$

where $\hat{\beta} = (X'X)^{-1}X'Y$. Thus, P_XY can be interpreted as the fitted values of Y from a regression of Y against X. Thus, P_X projects $n \times 1$ vectors onto a linear subspace of $\mathbb{R}^n$ spanned by the columns of the matrix X:

$$S(X) = \{y \in \mathbb{R}^n : y = Xb, b \in \mathbb{R}^k\}.$$

However, the projection is *orthogonal* in the following sense.

Consider the difference between Y and its projection P_XY:

$$Y - P_XY = (I_n - P)Y = (I_n - X(X'X)^{-1}X)Y = M_XY,$$

where

$$M_X = I_n - P = I_n - X(X'X)^{-1}X$$

is the second projection matrix we are interested in. Note that

$$M_XY = Y - \hat{Y} = Y - X\hat{\beta} = \hat{U}.$$

Thus, a projection by P_X generates fitted/predicted values, and a projection by M_X generates fitted/sample residuals. Moreover, since

$$I_n = P_X + M_X,$$

we have

$$Y = P_XY + M_XY = \hat{Y} + \hat{U},$$

with

$$\hat{Y}'\hat{U} = 0,$$

which holds by $X'\hat{U} = 0$. Thus, M_X projects n-vectors onto

$$S^{\perp}(X) = \{y \in \mathbb{R}^n : y'X = 0\}.$$

Additional properties of the projection matrices are given below.

Proposition 1.8.1.

(a) P_X and M_X are symmetric: $P_X' = P_X$ and $M_X' = M_X$.
(b) $P_XX = X$.
(c) $M_XX = 0$.
(d) P_X and M_X are orthogonal: $M_XP_X = 0$ and $P_XM_X = 0$.
(e) P_X and M_X are idempotent: $P_XP_X = P_X$ and $M_XM_X = M_X$.
(f) rank$(P_X) = k$ and rank$(M_X) = n - k$.

Proof. The results in (a) follow immediately from the definitions of P_X and M_X.

For part (b),

$$P_XX = X(X'X)^{-1}X'X = XI_k = X.$$

For part (c),

$$M_XX = (I_n - P_X)X = X - P_XX = X - X = 0.$$

For part (d),

$$M_XP_X = M_XX(X'X)^{-1}X' = 0,$$

and

$$P_X M_X = (M_X P_X)' = 0.$$

For part (e),

$$P_X P_X = P_X X(X'X)^{-1}X' = X(X'X)^{-1}X' = P_X,$$

and

$$M_X M_X = M_X - M_X P_X = M_X.$$

For part (f), using the results that P_X and M_X are symmetric and idempotent, one can show that their ranks are equal to their traces. Then,

$$\mathrm{tr}(P_X) = \mathrm{tr}(X(X'X)^{-1}X') = \mathrm{tr}((X'X)^{-1}X'X) = \mathrm{tr}(I_k) = k.$$

□

In an OLS regression, where Y is the vector of observations on the dependent variable and X is the matrix of regressors, the sum of squared residuals $\hat{U} = M_X Y$ is given by

$$\hat{U}'\hat{U} = Y' M_X Y.$$

Proposition 1.8.2. *The sum of squared residuals cannot decrease when adding more regressors.*

Proof. Consider partitioned regression matrix $X = (Z \;\; W)$. Let us study the effect of adding extra regressors W on the sum of squared residuals. Let

$P_X = X\left(X'X\right)^{-1}X'$ be the projection matrix corresponding to the full regression,

$P_Z = Z\left(Z'Z\right)^{-1}Z'$ be the projection matrix corresponding to the regression without W.

Define also

$$M_X = I_n - P_X,$$
$$M_Z = I_n - P_Z.$$

Note that since Z is a part of X,

$$P_X Z = Z,$$

and

$$\begin{aligned} P_X P_Z &= P_X Z\left(Z'Z\right)^{-1}Z' \\ &= Z\left(Z'Z\right)^{-1}Z' \\ &= P_Z. \end{aligned}$$

Consequently,

$$\begin{aligned} M_X M_Z &= \left(I_n - P_X\right)\left(I_n - P_Z\right) \\ &= I_n - P_X - P_Z + P_X P_Z \\ &= I_n - P_X - P_Z + P_Z \\ &= M_X. \end{aligned}$$

Define

$$\widehat{U}_X = M_X Y,$$
$$\widehat{U}_Z = M_Z Y,$$

and write

$$\begin{aligned} 0 &\leq \left(\widehat{U}_X - \widehat{U}_Z\right)' \left(\widehat{U}_X - \widehat{U}_Z\right) \\ &= \widehat{U}_X'\widehat{U}_X + \widehat{U}_Z'\widehat{U}_Z - 2\widehat{U}_X'\widehat{U}_Z, \end{aligned}$$

where the inequality in the first line holds by the fact that $x'x = \sum_i x_i^2 \geq 0$. Next,

$$\begin{aligned} \widehat{U}_X'\widehat{U}_Z &= Y' M_X M_Z Y \\ &= Y' M_X Y \\ &= \widehat{U}_X'\widehat{U}_X. \end{aligned}$$

Hence,

$$\widehat{U}_Z'\widehat{U}_Z \geq \widehat{U}_X'\widehat{U}_X.$$

□

1.9. Appendix: Matrix square root

Let A be a symmetric and positive definite $k \times k$ matrix. One can show that there exists a unique symmetric $k \times k$ matrix B such that

$$BB = A.$$

We therefore denote

$$A^{1/2} = B.$$

Moreover, the inverse of the matrix B exists and

$$B^{-1}B^{-1} = A^{-1}.$$

Hence,

$$A^{-1/2} = B^{-1}.$$

Note that

$$A^{-1/2}AA^{-1/2} = A^{-1/2}A^{1/2}A^{1/2}A^{-1/2} = I_k.$$

CHAPTER 2

Selecting regressors using the Bayesian Information Criterion (BIC)

In the context of linear regression and OLS, we discuss information-criteria-based approaches for selecting relevant regressors among many potential controls. We discuss consistency, the oracle properties, and post-selection inference. While the results are presented for the linear regression model, the same approach can be applied to nonlinear models such as probit, logit, and etc.

2.1. Selecting regressors

Consider a linear regression model with k *potential regressors*:

$$Y_i = \sum_{j=1}^{k} \beta_j X_{i,j} + U_i, \tag{2.1.1}$$
$$EX_{i,j}U_i = 0, \quad j = 1, \ldots, k.$$

For now, we assume that the number of potential regressors is small: k is fixed and does not depend on n.

Let $\mathcal{A}$ denote the set (list) of regressors with non-zero coefficients:

$$\mathcal{A} = \{j : \beta_j \neq 0\}.$$

For example, $\mathcal{A} = \{1, 3, 7\}$ implies that only the regressors $X_{i,1}$, $X_{i,3}$, and $X_{i,7}$ have non-zero coefficients, and that the remaining regressors have coefficients equal to zero. We use $\mathcal{A}_0$ to denote the *true set of relevant regressors*: i.e. the true data generating process (DGP) for Y_i only includes the regressors in $\mathcal{A}_0$:

$$Y_i = \sum_{j \in \mathcal{A}_0} \beta_j X_{i,j} + U_i.$$

Our goal is to estimate $\mathcal{A}_0$ using the data $\{(Y_i, X_i')', i = 1, \ldots, n\}$. We use $\hat{\mathcal{A}}_n$ to denote an estimated set of relevant regressors produced by a selection procedure. We say that the selection procedure is consistent if

$$P\left(\hat{\mathcal{A}}_n = \mathcal{A}_0\right) \to 1 \tag{2.1.2}$$

as $n \to \infty$.

Let $\beta = (\beta_1, \ldots, \beta_k)'$, and by $\beta_{\mathcal{A}}$ we denote the subvector of β that includes only the coefficients in $\mathcal{A}$:

$$\beta_{\mathcal{A}} = (\beta_j : j \in \mathcal{A}).$$

We use $|\mathcal{A}|$ to denote the number of elements in $\mathcal{A}$, and hence $\beta_{\mathcal{A}}$ is a $|\mathcal{A}|$-subvector of the k-vector β.

Suppose a procedure produced the set of selected regressors $\hat{\mathcal{A}}_n$ and the vector of estimates $\hat{\beta}_n = (\hat{\beta}_{n,1}, \ldots, \hat{\beta}_{n,k})'$. Here we set $\hat{\beta}_{n,j} = 0$ for $j \notin \hat{\mathcal{A}}_n$. We say that the procedure is *oracle* if, in addition to the consistency property in (2.1.2),

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}_0} - \beta_{\mathcal{A}_0}) \to_d N(0, V(\mathcal{A}_0)),$$

where $V(\mathcal{A}_0)$ is the best asymptotic variance one can obtain when the true model $\mathcal{A}_0$ is known. The oracle property means that not only the econometrician consistently selects the true regressors, but also the coefficients on the relevant regressors are estimated as precisely as when the set of the true relevant regressors in the DGP is known.

2.2. BIC

Recall that if the econometrician tries to select the regressors by minimizing the sample sum of squared residuals (SSR), or equivalently maximizing R^2, the procedure would result in overfitting: the SSR is monotone non-increasing in the number of included regressors. The idea behind BIC is to penalize the SSR for the model complexity.

Let $X_i = (X_{i,1}, \ldots, X_{i,k})'$, and define $X_{i,\mathcal{A}}$ as the subvector of X_i that includes only the regressors in $\mathcal{A}$:

$$X_{i,\mathcal{A}} = (X_{i,j} : j \in \mathcal{A}).$$

Again, $X_{i,\mathcal{A}}$ is a $|\mathcal{A}|$-subvector of the k-vector X_i. The true DGP can now be written as

$$\begin{aligned} Y_i &= \sum_{j \in \mathcal{A}_0} \beta_j X_{i,j} + U_i \\ &= X_{i,\mathcal{A}_0}' \beta_{\mathcal{A}_0} + U_i. \end{aligned}$$

Let $\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})$ denote the OLS estimator of $\beta_{\mathcal{A}}$ that only uses the regressors in $\mathcal{A}$:

$$\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) = \left(\sum_{i=1}^n X_{i,\mathcal{A}} X_{i,\mathcal{A}}' \right)^{-1} \sum_{i=1}^n X_{i,\mathcal{A}} Y_i.$$

We can set

$$\hat{\beta}_{n,\mathcal{A}^c}(\mathcal{A}) = 0,$$

and view $\hat{\beta}_n(\mathcal{A}) = (\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})', \hat{\beta}_{n,\mathcal{A}^c}(\mathcal{A})')'$ as the estimator *of* $\beta = (\beta_{\mathcal{A}}', \beta_{\mathcal{A}^c}')'$ *under model* $\mathcal{A}$. The corresponding SSR is given by

$$SSR_n(\mathcal{A}) = \sum_{i=1}^n \left(Y_i - X_{i,\mathcal{A}}' \hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) \right)^2.$$

The complexity of model $\mathcal{A}$ can be measured by the number of included regressors, i.e. the number of elements in $\mathcal{A}$. BIC for model $\mathcal{A}$ is defined as

$$BIC_n(\mathcal{A}) = SSR_n(\mathcal{A}) + |\mathcal{A}| \log n,$$

where the second term is a *penalty*. A model with more included regressors receives a larger penalty. A BIC-based selection procedure selects the regressors by minimizing BIC across all

possible models:

$$\hat{\mathcal{A}}_n^{BIC} = \arg\min_{\mathcal{A}} BIC_n(\mathcal{A}).$$

We show below that BIC selects the relevant regressors consistently.

Proposition 2.2.1. *Suppose that data are iid, EX_iX_i' and $EU_i^2X_iX_i'$ are finite and positive definite, and $EU_i^2 < \infty$. Then $P\left(\hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) \to 1$ as $n \to \infty$.*

We will see in the proof below that the penalty term only plays a role when $\mathcal{A}_0 \subset \mathcal{A}$: omitted regressors are detected by the SSR term in the BIC definition. The penalty term is only needed to deselect irrelevant controls.

Proof. It suffices to show that for all $\mathcal{A} \neq \mathcal{A}_0$

$$P\left(BIC_n(\mathcal{A}) > BIC_n(\mathcal{A}_0)\right) \to 1, \tag{2.2.1}$$

i.e. the true model $\mathcal{A}_0$ minimizes BIC with probability approaching one.

We say that the random sequence $V_n = o_p(1)$ if V_n converges in probability to zero, see the discussion in Appendix 2.6. For example, $V_n = o_p(1)$ when $EV_n = 0$ and $Var(V_n) \to 0$. First, consider the *average* SSR for the true model:

$$\begin{aligned}
n^{-1}SSR_n(\mathcal{A}_0) &= n^{-1}\sum_{i=1}^n \left(Y_i - X_{i,\mathcal{A}_0}'\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0)\right)^2 \\
&= n^{-1}\sum_{i=1}^n \left(U_i - X_{i,\mathcal{A}_0}'(\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0})\right)^2 \\
&= n^{-1}\sum_{i=1}^n U_i^2 + (\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0})'\left(n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}X_{i,\mathcal{A}_0}'\right)(\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0}) \\
&\quad - 2(\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0})'\left(n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}U_i\right) \\
&= EU_i^2 + o_p(1),
\end{aligned}$$

where the $o_p(1)$ term in the last line is by the LLN and consistency of the OLS estimator under the true model:

$$\begin{aligned}
n^{-1}\sum_{i=1}^n U_i^2 &= EU_i^2 + o_p(1), \\
\hat{\beta}_{n,\mathcal{A}_0} &= \beta_{\mathcal{A}_0} + o_p(1), \\
n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}X_{i,\mathcal{A}_0}' &= EX_{i,\mathcal{A}_0}X_{i,\mathcal{A}_0}' + o_p(1), \\
n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}U_i &= o_p(1).
\end{aligned}$$

Suppose that model $\mathcal{A}$ omits some relevant regressors:

$$(\mathcal{A} \cap \mathcal{A}_0) \neq \mathcal{A}_0.$$

Since the OLS estimator is inconsistent in general when there are omitted relevant regressors,

$$\hat{\beta}_n(\mathcal{A}) - \beta \to_p \delta \neq 0,$$

where $\hat{\beta}_{n,j}(\mathcal{A})$ is the corresponding element of $\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})$ for $j \in \mathcal{A}$, and $\hat{\beta}_n(\mathcal{A}) = 0$ for $j \notin \mathcal{A}$. We have:

$$\begin{aligned}
n^{-1}SSR_n(\mathcal{A}) &= n^{-1}\sum_{i=1}^{n}\left(Y_i - X_i'\hat{\beta}_n(\mathcal{A})\right)^2 \\
&= n^{-1}\sum_{i=1}^{n}\left(U_i - X_i'\left(\hat{\beta}_n(\mathcal{A}) - \beta\right)\right)^2 \\
&= n^{-1}\sum_{i=1}^{n}U_i^2 + (\hat{\beta}_n(\mathcal{A}) - \beta)'\left(n^{-1}\sum_{i=1}^{n}X_iX_i'\right)(\hat{\beta}_n(\mathcal{A}) - \beta) \\
&\quad - 2(\hat{\beta}_n(\mathcal{A}) - \beta)'\left(n^{-1}\sum_{i=1}^{n}X_iU_i\right) \\
&= EU_i^2 + \delta' EX_iX_i'\delta + o_p(1).
\end{aligned}$$

Note also that

$$|\mathcal{A}|\frac{\log n}{n} = o(1).$$

Therefore, for such a model $\mathcal{A}$,

$$\begin{aligned}
P\left(BIC_n(\mathcal{A}) > BIC_n(\mathcal{A}_0)\right) &= P\left(n^{-1}BIC_n(\mathcal{A}) > n^{-1}BIC_n(\mathcal{A}_0)\right) \\
&= P\left(n^{-1}SSR_n(\mathcal{A}) + |\mathcal{A}|\frac{\log n}{n} > n^{-1}SSR_n(\mathcal{A}_0) + |\mathcal{A}_0|\frac{\log n}{n}\right) \\
&= P\left(\delta' EX_iX_i'\delta + o_p(1) + o(1) > 0\right) \\
&\to 1,
\end{aligned}$$

where convergence in the last line holds because $\delta \neq 0$ and EX_iX_i' is positive definite.

Next, consider model $\mathcal{A}$ such that

$$\mathcal{A}_0 \subset \mathcal{A}.$$

In this case, $\mathcal{A}$ contains all the relevant regressors as well as some irrelevant ones. The OLS estimator $\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})$ is consistent and asymptotically normal:

$$n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}}) \to_d \Psi_{\mathcal{A}},$$

where

$$\begin{aligned}
\Psi_{\mathcal{A}} &\sim N\left(0, V(A)\right), \\
V(\mathcal{A}) &= \sigma^2\left(EX_{i,\mathcal{A}}X_{i,\mathcal{A}}'\right)^{-1}.
\end{aligned}$$

The result follows from

$$n^{-1/2}\sum_{i=1}^{n}X_iU_i \to_d \Phi_{\mathcal{A}},$$

where

$$\Phi_{\mathcal{A}} \sim N\left(0, \sigma^2 X_{i,\mathcal{A}} X_{i,\mathcal{A}}'\right).$$

We use $V_n = O_p(1)$ to say that V_n is bounded in probability. For example, the sequence of random variables $V_n = O_p(1)$ when $Var(V_n) \leq K < \infty$ for all n. Convergence in distribution and convergence in probability to a constant both imply $O_p(1)$:

$$n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}}) = O_p(1),$$

$$n^{-1/2}\sum_{i=1}^{n} X_i U_i = O_p(1),$$

$$n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}} X_{i,\mathcal{A}}' = O_p(1).$$

In all the above cases, the variances become bounded as $n \to \infty$ (zero in the latter case).

We have:

$$\begin{aligned} SSR_n(\mathcal{A}) - \sum_{i=1}^{n} U_i^2 &= \sum_{i=1}^{n}\left(U_i - X_{i,\mathcal{A}}'(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}})\right)^2 - \sum_{i=1}^{n} U_i^2 \\ &= n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}})'\left(n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}} X_{i,\mathcal{A}}'\right) n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}}) \\ &\quad - 2n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}})'\left(n^{-1/2}\sum_{i=1}^{n} X_{i,\mathcal{A}} U_i\right) \\ &= O_p(1). \end{aligned}$$

By the same arguments,

$$\begin{aligned} SSR_n(\mathcal{A}_0) - \sum_{i=1}^{n} U_i^2 &\to_d \Psi_{\mathcal{A}_0}'\left(E X_{i,\mathcal{A}_0} X_{i,\mathcal{A}_0}'\right)\Psi_{\mathcal{A}_0} - 2\Psi_{\mathcal{A}_0}'\Phi_{\mathcal{A}_0} \\ &= O_p(1). \end{aligned}$$

Lastly, when $\mathcal{A}_0 \subset \mathcal{A}$,

$$\begin{aligned} P\left(BIC_n(\mathcal{A}) > BIC_n(\mathcal{A}_0)\right) &= P\left(SSR_n(\mathcal{A}) - SSR_n(\mathcal{A}_0) > (|\mathcal{A}_0| - |\mathcal{A}|)\log n\right) \\ &= P\left(O_p(1) > (|\mathcal{A}_0| - |\mathcal{A}|)\log n\right) \\ &\to 1, \end{aligned}$$

where convergence in the last line holds since $|\mathcal{A}_0| < |\mathcal{A}|$, and therefore

$$(|\mathcal{A}_0| - |\mathcal{A}|)\log n \to -\infty.$$

□

2.3. Post BIC inference

Suppose the econometrician selects the true model using $\hat{\mathcal{A}}_n^{BIC}$ and conducts inference using $\hat{\beta}_n(\hat{\mathcal{A}}_n^{BIC})$. For $j \in \hat{\mathcal{A}}_n^{BIC}$, the distribution of the estimator of the j-the coefficient is

given by

$$\begin{aligned}
&P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u\right) \\
&= P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u, \hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) \\
&\quad + P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u, \hat{\mathcal{A}}_n^{BIC} \neq \mathcal{A}_0\right) \\
&= P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u, \hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) + o(1) \\
&= P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u \mid \hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) P\left(\hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) + o(1) \\
&= P\left(n^{1/2}(\hat{\beta}_{n,j}(\mathcal{A}_0) - \beta_j) \leq u\right)(1 + o(1)) + o(1) \\
&= P\left(n^{1/2}(\hat{\beta}_{n,j}(\mathcal{A}_0) - \beta_j) \leq u\right) + o(1).
\end{aligned}$$

where the second equality holds by

$$P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u, \hat{\mathcal{A}}_n^{BIC} \neq \mathcal{A}_0\right) \leq P\left(\hat{\mathcal{A}}_n^{BIC} \neq \mathcal{A}_0\right) = o(1).$$

Hence, the BIC-based selection and estimation procedure is oracle.

2.4. Akaike Information Criterion (AIC)

AIC is another popular criterion for model selection (and actually precedes BIC). AIC for a model $\mathcal{A}$ is defined as

$$AIC_n(\mathcal{A}) = SSR_n(\mathcal{A}) + 2|\mathcal{A}|.$$

In comparison with BIC, AIC penalizes the model complexity less heavily and, therefore, tends to select a bigger model with more regressors than BIC.

By the same arguments as in the proof of Proposition 2.2.1, for a model $\mathcal{A}$ that omits some relevant regressors, i.e. $(\mathcal{A} \cap \mathcal{A}_0) \neq \mathcal{A}_0$,

$$P\left(AIC_n(\mathcal{A}) > AIC_n(\mathcal{A}_0)\right) \to 1.$$

However, because AIC penalty is not sufficiently strong, if $\mathcal{A}_0 \subset \mathcal{A}$,

$$P\left(AIC_n(\mathcal{A}) > AIC_n(\mathcal{A}_0)\right) \nrightarrow 1.$$

Hence, while AIC detects omitted regressors with probability approaching one, it is more likely to overfit by also including some irrelevant regressors than BIC.

2.5. Limitations

One should note several limitations of our arguments. First, we assumed that k is small (fixed). Some of our arguments would breakdown when the number of potential regressors is large (comparable to the sample size). However, this technical issue can often be addressed with somewhat different arguments.

More importantly, our analysis ignores the situation where some β_j are very close but different from zero. One cannot expect that the BIC (or any other procedure) would detect small coefficients with a probability approaching one. Even in the limit, regressors with very

small coefficients are likely to be omitted from the model, which can potentially create the omitted variable bias. This shortcoming can be addressed using a double selection procedure, which will be discussed later in the context of Lasso.

Lastly, while the BIC procedure delivers an automatic selection of regressors, it may be infeasible in practice if the number of potential regressors is very large. With k regressors, there are 2^k possible models $\mathcal{A}$. For example, if $k = 30$ one has to run and compare over 1 billion potential regressions. For $k = 40$, one has to run over 1 trillion (10^{12}) models. Suppose the CPU time for one regression is 10^{-3} (This is a typical CPU time for estimating a regression with $n = 10,000$ and $k = 40$ on a high-end modern laptop using the "lm()" function in R.) In that case, it would take about 2.6 years to run all 1 trillion possible regressions using 12 cores in parallel.

The many regressors situation can easily arise in practice even when there is a relatively small number of explanatory variables. Suppose that the econometrician considers flexible specifications that include quadratic terms as well as pairwise interaction terms of all right-hand side variables. In that case, 10 potential right-hand side variables generate 65 potential regressors.

2.6. Appendix: Law of Large Numbers (LLN); Little–o notation

We say that $\hat{\theta}_n$ converges in probability to θ if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\|\hat{\theta}_n - \theta\| > \epsilon\right) = 0.$$

Convergence in probability implies that the probability of $\hat{\theta}_n$ deviating from θ by any amount $\epsilon > 0$ becomes negligible as $n \to \infty$. We use the notation

$$\hat{\theta}_n - \theta \to_p 0$$

and

$$\hat{\theta}_n - \theta = o_p(1).$$

The main device for establishing convergence in probability is the law of large numbers. Let $X_1, \ldots, X_n$ be uncorrelated random variables with $EX_i = \mu$ and $Var(X_i) = \sigma^2$, and consider the average

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i.$$

Note that

$$E\bar{X}_n = \mu,$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty. \tag{2.6.1}$$

Hence, as $n \to \infty$ the distribution of the average $\bar{X}_n$ becomes concentrated around the mean μ. More formally, by Markov's inequality

$$P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \leq \frac{1}{\epsilon^2} E\left|\bar{X}_n - \mu\right|^2 = \frac{\sigma^2}{\epsilon^2 n} \to 0. \tag{2.6.2}$$

To show (2.6.1), which also implies the equality in (2.6.2),

$$\begin{aligned} Var(\bar{X}_n) &= Var\left(n^{-1}\sum_{i=1}^{n} X_i\right) \\ &= n^{-2}Var\left(\sum_{i=1}^{n} X_i\right) \\ &= n^{-2}\left(\sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n}\sum_{j\neq i} Cov(X_i, X_j)\right) \\ &= n^{-2}\sum_{i=1}^{n} Var(X_i) \\ &= n^{-1}\sigma^2, \end{aligned}$$

where the equality in the fourth line holds because we assume that $Cov(X_i, X_j) = 0$ for $i \neq j$, and the equality in the last line holds because $Var(X_i) = \sigma^2$.

In the case of iid data, the following result can be used. Let $X_1, \ldots, X_n$ be iid random variables such that $E|X_i| < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_p 0,$$

or equivalently

$$\bar{X}_n = EX_i + o_p(1).$$

2.7. Appendix: Consistency of OLS

We say that the OLS estimator $\hat{\beta}$ is consistent for the true β if

$$\hat{\beta} = (X'X)^{-1}X'Y = \left(n^{-1}\sum_{i=1}^{n} X_iX_i'\right)^{-1} n^{-1}\sum_{i=1}^{n} X_iY_i \to_p \beta$$

Proposition 2.7.1. *Suppose that data $\{(Y_i, X_i) : i = 1, \ldots, n\}$ are iid,*

$$Y_i = X_i'\beta + U_i, \tag{2.7.1}$$

$$EU_iX_i = 0,$$

$$EX_iX_i' \text{ is finite and positive definite.} \tag{2.7.2}$$

Then,

$$\hat{\beta} \to_p \beta.$$

PROOF. Write

$$\begin{aligned}
\hat{\beta} &= \left(n^{-1} \sum_{i=1}^{n} X_i X_i' \right)^{-1} n^{-1} \sum_{i=1}^{n} X_i Y_i \\
&= \beta + \left(n^{-1} \sum_{i=1}^{n} X_i X_i' \right)^{-1} n^{-1} \sum_{i=1}^{n} X_i U_i \\
&= \beta + \left(EX_i X_i' + o_p(1) \right)^{-1} \left(EX_i U_i + o_p(1) \right) \\
&\to_p \beta + \left(EX_i X_i' \right)^{-1} \cdot 0 \\
&= \beta.
\end{aligned}$$

□

The OLS estimator is inconsistent when

$$EX_i U_i \neq 0.$$

In this case,

$$\begin{aligned}
\hat{\beta} &= \left(n^{-1} \sum_{i=1}^{n} X_i X_i' \right)^{-1} n^{-1} \sum_{i=1}^{n} X_i Y_i \\
&= \beta + \left(n^{-1} \sum_{i=1}^{n} X_i X_i' \right)^{-1} n^{-1} \sum_{i=1}^{n} X_i U_i \\
&\to_p \beta + \left(EX_i X_i' \right)^{-1} EX_i U_i \\
&\neq \beta,
\end{aligned}$$

where

$$\left(EX_i X_i' \right)^{-1} EX_i U_i \neq 0$$

can be viewed as asymptotic bias. For example, suppose the true model is given by

$$\begin{aligned}
Y_i &= X_{i,1}' \beta_1 + X_{i,2}' \beta_2 + \epsilon_i, \\
EX_{i,1} \epsilon_i &= 0, \\
EX_{i,2} \epsilon_i &= 0,
\end{aligned}$$

but the econometrician omits $X_{i,2}$ from the model:

$$\begin{aligned}
Y_i &= X_{i,1}' \beta_1 + U_i, \\
U_i &= X_{i,2}' \beta_2 + \epsilon_i.
\end{aligned}$$

Then

$$EX_{i,1} U_i = EX_{i,1} X_{i,2}' \beta_2 \neq 0,$$

and

$$\tilde{\beta}_1 = \left(n^{-1} \sum_{i=1}^{n} X_{i,1} X_{i,1}' \right)^{-1} n^{-1} \sum_{i=1}^{n} X_{i,1} Y_i \to_p \beta_1 + \left(EX_{i,1} X_{i,1}' \right)^{-1} EX_{i,1} X_{i,2}' \beta_2.$$

2.8. Appendix: Convergence in distribution and asymptotic normality/Central Limit Theorem; Big–O notation

Let $\hat{\theta}_n$ denote an estimator of a scalar parameter θ. To perform hypothesis testing about θ (or construct a confidence interval for θ) using the estimator $\hat{\theta}_n$, one needs to know the distribution of the latter. Unfortunately, in many circumstances, it is impossible to derive the exact distribution of $\hat{\theta}_n$ either because the expression is too complicated, or because the derivation of the exact finite sample distribution requires very restrictive assumptions. In such cases, we rely on asymptotic approximations that are usually applied to $\sqrt{n}(\hat{\theta}_n - \theta)$, i.e. we approximate the distribution of the scaled estimation error. The scaling is necessary when $\hat{\theta}_n - \theta \to_p 0$. While the probability

$$P(\sqrt{n}(\hat{\theta}_n - \theta) \leq x)$$

is unknown for finite n, suppose we can establish that for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P(\sqrt{n}(\hat{\theta}_n - \theta) \leq x) = P\left(X \leq x\right),$$

where $X \sim N(0, \omega^2)$. In such cases, we say that $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to a normal random variable, denoted as

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \omega^2). \tag{2.8.1}$$

We use the $N(0, \omega^2)$ distribution to approximate that of $\sqrt{n}(\hat{\theta}_n - \theta)$. Suppose that (2.8.1) holds. Then for any $M > 0$,

$$P\left(|\sqrt{n}(\hat{\theta}_n - \theta)| \geq M\right) \to P\left(|X| \geq M\right) \quad \text{where } X \sim N(0, \omega^2).$$

Thus, in large samples, the probability of $\sqrt{n}(\hat{\theta}_n - \theta)$ taking on a large value greater than M is approximately the same as that of a normal random variable.

We say $V_n = O_p(1)$ if it is bounded in probability: for all $\epsilon > 0$ there is $M_\epsilon > 0$ such that $P(\|V_n\| > M_\epsilon) < \epsilon$ for all n large enough. Since $\lim_{M\to\infty} P(|X| \geq M) = 0$ for any proper random variable (that does not take infinite values), the convergence in distribution result $\sqrt{n}(\hat{\theta}_n - \theta) \to_d X$ implies that.

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_p(1).$$

We can also write

$$\hat{\theta}_n = \theta + \frac{1}{\sqrt{n}} O_p(1) = \theta + O_p\left(\frac{1}{\sqrt{n}}\right),$$

i.e. $\hat{\theta}_n$ converges to θ at the rate $1/\sqrt{n}$.

The concept can be extended to random vectors by considering the joint distribution of its elements. Suppose now that the random k-vector $\hat{\theta}_n$ is an estimator of $\theta \in \mathbb{R}^k$. Suppose further that for all $x = (x_1, \ldots, x_k)' \in \mathbb{R}^k$,

$$\lim_{n\to\infty} P\left(\sqrt{n}(\hat{\theta}_{n,1} - \theta_1) \leq x_1, \ldots, \sqrt{n}(\hat{\theta}_{n,k} - \theta_k) \leq x_k\right) = P\left(X_1 \leq x_1, \ldots, X_k \leq x_k\right),$$

where for some positive definite and symmetric $k \times k$ matrix Ω,

$$\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \sim N(0, \Omega),$$

where $N(0, \Omega)$ denotes the multivariate normal distribution with zero means and a variance-covariance matrix given by Ω. Then we say that $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to the $N(0, \Omega)$ random vector, denoted as

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \Omega).$$

We use the $N(0, \Omega)$ distribution to approximate the joint distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$.

The main device for establishing convergence in distribution is the Central Limit Theorem (CLT).

Proposition 2.8.1. *Suppose that $X_1, \ldots, X_n$ are iid random k-vectors such that $EX_i = 0$ and $Var(X_i) = EX_iX_i' = \Omega$, where Ω is a positive definite matrix. Then,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to_d N(0, \Omega).$$

Note that

$$E\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right) = 0,$$
$$Var\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right) = Var(X_i) = \Omega,$$

however, the CLT also says that the distribution of $n^{-1/2} \sum_{i=1}^{n} X_i$ can be approximated by that of a $N(0, \Omega)$ vector.

2.9. Appendix: Asymptotic normality of the OLS estimator

Consider the model defined by equations (2.7.1)–(2.7.2). The following result establishes the asymptotic normality of the OLS estimator.

Proposition 2.9.1. *Suppose that data are iid, (2.7.1)–(2.7.2) hold, and*

$$E\left(U_i^2 \mid X_i\right) = \sigma^2. \tag{2.9.1}$$

Then,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d N(0, \sigma^2 (EX_iX_i')^{-1}).$$

Proof. By (2.7.1),

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \left(\frac{1}{n} \sum_{i=1}^{n} X_iX_i'\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_iU_i.$$

By the LLN,

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \to_p EX_i X_i',$$

and the matrix in the limit is positive definite and therefore invertible. By the law of iterated expectation,

$$Var(X_i U_i) = E(U_i^2 X_i X_i') = E\left(E(U_i^2 \mid X_i) X_i X_i'\right) = E\left(\sigma^2 EX_i X_i'\right) = \sigma^2 EX_i X_i',$$

where the third equality holds by (2.9.1). Hence,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i U_i \to_d N\left(0, \sigma^2 EX_i X_i'\right),$$

and

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d (EX_i X_i)^{-1} N\left(0, \sigma^2 EX_i X_i'\right) = N\left(0, \sigma^2 \left(EX_i X_i'\right)^{-1}\right).$$

□

CHAPTER 3

Ridge and Least Absolute Shrinkage and Selection Operator (Lasso)

We start by discussing Ridge regression and its coefficients-dependent penalty. We then cover the basics of Lasso using results for convex optimization. The special case of orthonormal regressors (for which Lasso has an analytical solution) is analyzed before moving to the general case. Using the Lasso first-order conditions, we discuss its consistency properties. We also cover weighted and adaptive Lasso and conclude with some adjustments needed for high-dimensional data.

3.1. Ridge Regression

Ridge regression is based on a similar idea to that of BIC and AIC: penalize a measure of fit by taking into account the size/complexity of the model. However, the Ridge penalty term depends on the estimates of the regression coefficients.

For a vector x, recall that the p-norm is defined as

$$\|x\|_p = \left(\sum_i |x_i|^p \right)^{1/p},$$

with $\|\cdot\|_2$ being the usual Euclidean norm. The Ridge criterion function for a linear regression model with k potential regressors is given by

$$n^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{k} b_j X_{i,j})^2 + \lambda \sum_{j=1}^{k} b_j^2 = n^{-1} \|Y - Xb\|_2^2 + \lambda \|b\|_2^2.$$

where Y is the $n \times 1$ vector of observations on the dependent variable, and X is the $n \times k$ matrix of observations on the potential regressors. Here $\lambda > 0$ is a tuning penalty parameter: larger values of λ imply heavier penalization.

The Ridge estimator is given by

$$\hat{\beta}^{Ridge} = \arg\min_{b \in \mathbb{R}^k} \left\{ n^{-1} \|Y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}.$$

The first-order condition for the Ridge problem is

$$-X' \left(Y - X\hat{\beta}^{Ridge} \right) + \lambda \hat{\beta}^{Ridge} = 0,$$

which implies that

$$\hat{\beta}^{Ridge} = \left(X'X + \lambda I_k \right)^{-1} X'Y.$$

One can see that, relative to the OLS estimator, Ridge *shrinks* the estimates of all coefficients toward zero, with more shrinkage applied for larger values of λ.

Suppose that the classical regression assumptions hold:

$$\begin{aligned} Y &= X\beta + U, \\ E(U \mid X) &= 0, \\ \operatorname{rank}(X) &= k. \end{aligned}$$

In that case, the Ridge estimator is biased:

$$E\left(\hat{\beta}^{Ridge} \mid X\right) = \left(X'X + \lambda I_k\right)^{-1} X'X\beta \neq \beta.$$

However, when there are many potential regressors and k is close to n and, as a result, $X'X$ is close to being singular, Ridge provides a *regularization* as the eigenvalues of

$$X'X + \lambda I_k$$

are pushed away from zero. Even when the number of potential regressors is very large and $k > n$ (so that $X'X$ is singular and the OLS estimator cannot be computed), the Ridge estimator is still well-defined.

The regularization property of Ridge makes it useful in prediction problems where the number of potential predictors is very large. While Ridge regularization introduces a bias, it also reduces the variance, which may produce better (in the mean squared error sense) forecasts for Y.

The Ridge problem can be also viewed as a constrained optimization problem:

$$\min_{b \in \mathbb{R}^k} \|Y - Xb\|_2^2 \quad \text{s.t. } \|b\|_2^2 \leq M,$$

where $M > 0$ is some constant. Hence, Ridge imposes a "budget" constrained on the coefficients: $\sum_{j=1}^k b_j^2 \leq M$. Note that the budget constraint defines a sphere of radius M with the center at zero. Next, consider the contours $C_{SSR} = \{b \in \mathbb{R}^k : \|Y - Xb\|_2^2 - \|Y - X\hat{\beta}_{OLS}\|_2^2 = SSR\}$. These contours have a stretched ellipsoid form with the center at the OLS estimates $\hat{\beta}_{OLS}$. Note that larger contours correspond to larger SSRs. Due to the quadratic shape of the budget constraint and of the contours, the solution to the Ridge problem is always in the interior in the sense that $\hat{\beta}_j^{Ridge} \neq 0$ for all $j = 1, \ldots, k$. Hence, Ridge cannot provide a selection of regressors.

3.2. Lasso criterion function

As discussed in the previous section, Ridge is not useful for selecting regressors because of the shape of the constraint of its constrained minimization problem. Lasso addresses that issue by replacing the 2-norm in the penalty with the 1-norm. Thus, the Lasso criterion function is given by

$$n^{-1} \sum_{i=1}^n (Y_i - \sum_{j=1}^k b_j X_{i,j})^2/2 + \lambda \sum_{j=1}^k |b_j| = n^{-1}\|Y - Xb\|_2^2/2 + \lambda\|b\|_1.$$

The corresponding constrained optimization problem is now

$$\min_{b\in\mathbb{R}^k} \|Y - Xb\|_2^2 \quad \text{s.t. } \|b\|_1 \le M,$$

i.e. the "budget" constrain is given by

$$\sum_{j=1}^{k} |b_j| \le M.$$

The "budget" constrain now has sharp corners at zero coordinates ($b_j = 0$ for some $j = 1, \ldots, k$), and for sufficiently large λ, we will see corner solutions with exactly zero estimates of some of the coefficients. The Lasso estimator is defined as

$$\hat{\beta} = \arg\min_{b\in\mathbb{R}^k} \left\{ \frac{1}{2n}\|Y - Xb\|_2^2 + \lambda\|b\|_1 \right\},$$

and for sufficiently large penalty parameter λ, $\hat{\beta}_j = 0$ for some j's. Moreover, for larger values of λ, more estimates tend to be exactly zero.

Note that $\|Y - Xb\|_2^2$ and $\|b\|_1$ are convex functions of b and, therefore, the Lasso criterion function is convex. Moreover, it is differentiable every except $b = 0$. To discuss the solution to the Lasso problem, we need some results from convex optimization.

3.3. Convex minimization and subgradients

Recall that if a real-valued function $f(x), x \in \mathbb{R}^k$, is convex and differentiable at x, then for all $y \in \mathbb{R}^k$

$$f(y) - f(x) \ge \nabla f(x)'(y - x),$$

where

$$\nabla f(x) = \frac{\partial f(x)}{\partial x}$$

is the gradient, i.e. the k-vector of the partial derivatives of $f(x)$. Suppose that $f(x)$ is convex but not necessarily differentiable. *Subgradient* generalizes the notion of the gradient to such cases.

Definition. A k-vector g is a subgradient of f at x if for all $y \in \mathbb{R}^k$

$$f(y) - f(x) \ge g'(y - x).$$

The set $\partial f(x)$ of all subgradients at x is called subdifferential of f at x.

$$\partial f(x) = \{g \in \mathbb{R}^k : f(y) - f(x) \ge g'(y - x) \text{ for all } y \in \mathbb{R}^k\}.$$

One can show that if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. Moreover, if the subdifferential of f at x is a singleton, i.e. $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$.

For example, the absolute value function $f(x) = |x|$ is convex, continuous, and differentiable everywhere except at $x = 0$. Hence, for $x > 0$, $\partial f(x) = \{1\}$. Similarly, for $x < 0$, $\partial f(x) = \{-1\}$. Next, consider $x = 0$. The condition

$$|y| - |0| \ge g(y - 0) \text{ for all } y \in R,$$

implies that

$$-1 \leq g \leq 1.$$

Thus,

$$\partial|x| = \begin{cases} -1, & x < 0, \\ [-1,1], & x = 0, \\ 1, & x > 0. \end{cases} \tag{3.3.1}$$

Recall that if a convex function f is differentiable and minimized at x^*, then $\nabla f(x^*) = 0$. Subdifferentials can be used to generalize the property to non-differentiable convex functions.

Proposition 3.3.1. *Let $\mathcal{M}$ be the set of minima points of a convex function f:*

$$\mathcal{M} = \left\{ x \in \mathbb{R}^k : f(x) \leq f(y) \text{ for all } y \in \mathbb{R}^k \right\}.$$

Then $x \in \mathcal{M}$ if and only if $0 \in \partial f(x)$.

Proof. For sufficiency, suppose that $0 \in \partial f(x)$. By the definition of the subgradient, for all $y \in \mathbb{R}^k$

$$f(y) - f(x) \geq 0'(y - x) = 0,$$

or $f(x) \leq f(y)$ for $y \in \mathbb{R}^k$. Hence, $x \in \mathcal{M}$.

For necessity, suppose that $x \in \mathcal{M}$: for all $y \in \mathbb{R}^k$,

$$f(y) - f(x) \geq 0 = 0'(y - x).$$

Hence, by the definition of the subgradient, $0 \in \partial f(x)$. □

For example, $|x|$ is minimized at $x = 0$ as $0 \in [-1, 1] = \partial|0|$.

3.4. Analytical solution to the Lasso problem: a special case

A closed-form analytical solution to the Lasso problem exists only in a special case where the regressors are orthogonal to each other and normalized to have a unit sample second moment:

$$n^{-1} \sum_{i=1}^{n} X_{i,j} X_{i,l} = \begin{cases} 0, & j \neq l, \\ 1, & j = l. \end{cases} \tag{3.4.1}$$

Note that in this case,

$$n^{-1} X'X = I_k,$$
$$\hat{\beta}^{OLS} = n^{-1} X'Y,$$

or

$$\hat{\beta}_j^{OLS} = \frac{\sum_{i=1}^{n} X_{i,j} Y_i}{\sum_{i=1}^{n} X_{i,j}^2} = n^{-1} \sum_{i=1}^{n} X_{i,j} Y_i.$$

Define

$$(x)^+ = \max\{x, 0\},$$

$$\operatorname{sign}(x) = \begin{cases} -1 & x < 0, \\ 0 & x = 0, \\ 1 & x > 0. \end{cases}$$

Proposition 3.4.1. *Suppose that $n^{-1}X'X = I_k$. Then the Lasso estimator, i.e. the minimizer of*

$$Q_{n,\lambda}^{Lasso}(b) = \left\{ \frac{1}{2n}\|Y - Xb\|_2^2 + \lambda\|b\|_1 \right\},$$

satisfies

$$\hat{\beta}_j = \operatorname{sign}(\hat{\beta}_j^{OLS}) \left(|\hat{\beta}_j^{OLS}| - \lambda \right)^+,$$

$j = 1, \ldots, k$.

Proof. We will introduce some notation first for handling subdifferentials. Let $c \in \mathbb{R}$,v be a k-vector, and $\mathcal{S}$ be a set of k-vectors. We define

$$v + c\mathcal{S} = \{v + cg : g \in \mathcal{S}\}.$$

Note that $v + c\mathcal{S}$ is a set.

Using the above notation, the subdifferential of the Lasso criterion function can be written as

$$\begin{aligned} \partial Q_{n,\lambda}^{Lasso}(b) &= n^{-1}X'(Y - Xb) + \lambda\partial\|b\|_1 \\ &= -(\hat{\beta}^{OLS} - b) + \lambda\partial\|b\|_1 \\ &= -\begin{pmatrix} \hat{\beta}_1^{OLS} - b_1 - \lambda\partial|b_1| \\ \vdots \\ \hat{\beta}_k^{OLS} - b_k - \lambda\partial|b_k| \end{pmatrix}. \end{aligned} \tag{3.4.2}$$

Hence, the first-order condition for the Lasso estimator of the j-th coefficient is independent of the other Lasso coefficients: we can solve for $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)'$ element-by-element. By Proposition 3.3.1, the Lasso estimator for the j-th coefficient satisfies

$$0 \in \hat{\beta}_j^{OLS} - \hat{\beta}_j - \lambda\partial|\hat{\beta}_j|. \tag{3.4.3}$$

By (3.3.1), $\hat{\beta}_j = 0$ if and only if

$$\begin{aligned} 0 &\in \hat{\beta}_j^{OLS} - \lambda\partial|0| \\ &= \hat{\beta}_j^{OLS} - \lambda \cdot [-1, 1] \\ &= [\hat{\beta}_j^{OLS} - \lambda, \hat{\beta}^{OLS} + \lambda]. \end{aligned}$$

Equivalently,

$$\hat{\beta}_j^{OLS} - \lambda \leq 0 \leq \hat{\beta}_j^{OLS} + \lambda,$$

or by subtracting $\hat{\beta}_j^{OLS}$ from each side of the inequalities and multiplying by -1,

$$-\lambda \leq \hat{\beta}_j^{OLS} \leq \lambda.$$

Hence, $\hat{\beta}_j = 0$ if and only if

$$|\hat{\beta}_j^{OLS}| < \lambda,$$

or, equivalently,

$$\left(|\hat{\beta}_j^{OLS}| - \lambda\right)^+ = 0.$$

The Lasso estimator $\hat{\beta}_j$ is nonzero, if and only if $|\hat{\beta}_j^{OLS}| > \lambda$. This occurs when

$$\hat{\beta}_j^{OLS} > \lambda \text{ or } \hat{\beta}_j^{OLS} < -\lambda.$$

In this case, the first-order condition in (3.4.3) becomes

$$\hat{\beta}_j = \hat{\beta}_j^{OLS} - \lambda \partial|\hat{\beta}_j|, \tag{3.4.4}$$

with $\partial|\hat{\beta}_j| = 1$ if $\hat{\beta}_j > 0$, and $\partial|\hat{\beta}_j| = -1$ if $\hat{\beta}_j < 0$.

Suppose $\hat{\beta}_j^{OLS} > \lambda > 0$. One can see that a negative $\hat{\beta}_j$ cannot satisfy (3.4.4) as for $\hat{\beta}_j < 0$,

$$\hat{\beta}_j^{OLS} - \lambda \partial|\hat{\beta}_j| = \hat{\beta}_j^{OLS} + \lambda \cdot (-1) = \hat{\beta}_j^{OLS} + \lambda > 0.$$

Hence, $\hat{\beta}_j > 0$ and $\partial|\hat{\beta}_j| = 1$, and we have:

$$\hat{\beta}_j = \hat{\beta}_j^{OLS} - \lambda = + \left(|\hat{\beta}_j^{OLS}| - \lambda\right)^+.$$

Suppose that $\hat{\beta}_j^{OLS} < -\lambda < 0$. One can see that a positive $\hat{\beta}_j$ cannot satisfy (3.4.4) as for $\hat{\beta}_j > 0$,

$$\hat{\beta}_j^{OLS} - \lambda \partial|\hat{\beta}_j| = \hat{\beta}_j^{OLS} - \lambda \cdot (1) = \hat{\beta}_j^{OLS} - \lambda < 0.$$

Hence, $\hat{\beta}_j < 0$ and we have

$$\hat{\beta}_j = \hat{\beta}_j^{OLS} + \lambda = - \left(|\hat{\beta}_j^{OLS}| - \lambda\right)^+.$$

□

This special case (when the regressors are orthogonal and normalized to have a unit sample second moment) clearly illustrates the "shrinkage" and "selection" operations performed by Lasso. First, Lasso shrinks estimates toward zero relatively to the OLS estimates, where the amount of shrinkage is given exactly by λ. Moreover, if the amount of shrinkage exceeds the magnitude of a coefficient, it would be set to zero exactly.

Note that similarly to the Ridge regression, Lasso estimates are biased due to the shrinkage. However, unlike Ridge, Lasso can detect near-zero coefficients and "automatically" shrink them to zero, which is equivalent to dropping such regressors from the model. If Lasso keeps only the relevant regressors, one can consider post-Lasso OLS estimation to avoid the bias: after Lasso, use OLS to regress the dependent variable only on the regressors that survived the Lasso selection procedure.

Consider the case of a small number of potential regressors. Since OLS is consistent and asymptotically normal,

$$\hat{\beta}_j^{OLS} = \beta_j + O_p\left(\frac{1}{\sqrt{n}}\right).$$

Thus, for the irrelevant regressors with $\beta_j = 0$,

$$\hat{\beta}_j^{OLS} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

To shrink the corresponding Lasso estimates to zero, the penalty term λ has to dominate the noise component $O_p(1/\sqrt{n})$. However, to keep the relevant regressors with $\beta_j \neq 0$, the penalty term λ has to be smaller then $\min_j |\beta_j|$.

Suppose that for some $\delta > 0$,

$$\min_j |\beta_j| \geq \delta > 0. \tag{3.4.5}$$

Then Lasso can consistently select only the relevant regressors if, for example, we set

$$\lambda = \lambda_n = \sqrt{\frac{\log n}{n}}.$$

For the irrelevant regressors with $\beta_j = 0$,

$$\begin{aligned} P\left(\hat{\beta}_j = 0 \mid \beta_j = 0\right) &= P\left(|\hat{\beta}_j^{OLS}| < \lambda_n \mid \beta_j = 0\right) \\ &= P\left(O_p\left(\frac{1}{\sqrt{n}}\right) < \sqrt{\frac{\log n}{n}}\right) \\ &\to 1. \end{aligned}$$

For the relevant regressors with $\beta_j \neq 0$,

$$\begin{aligned} P\left(\hat{\beta}_j \neq 0 \mid \beta_j \neq 0\right) &= P\left(|\hat{\beta}_j^{OLS}| > \lambda_n \mid \beta_j \neq 0\right) \\ &= P\left(\left|\beta_j + O_p\left(\frac{1}{\sqrt{n}}\right)\right| > \sqrt{\frac{\log n}{n}} \mid \beta_j \neq 0\right) \\ &> P\left(\delta + O_p\left(\frac{1}{\sqrt{n}}\right) > \sqrt{\frac{\log n}{n}}\right) \\ &\to 1. \end{aligned}$$

However, note that the condition in (3.4.5) rules out small coefficients near zero: $\beta_j = c/\sqrt{n}$. Consistent detection of such coefficients is not possible as they are of the same order as the $O_p(1/\sqrt{n})$ noise component.

3.5. Lasso: the general case

There is no closed-form solution for the Lasso estimator in the general case. However, we can discuss its properties using the first-order conditions for the Lasso problem. By

Proposition 3.3.1 and (3.3.1), the Lasso estimator $\hat{\beta}$ solves

$$-\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i'\hat{\beta}\right) + \lambda_n \hat{g} = 0,$$

where $\hat{g} \in \partial\|\hat{\beta}\|_1$ is a subgradient at $\hat{\beta}$:

$$\hat{g}_j = \begin{cases} \text{sign}(\hat{\beta}_j) & \hat{\beta}_j \neq 0, \\ \in [-1,1] & \hat{\beta}_j = 0. \end{cases}$$

Let $\hat{\mathcal{A}}_n$ be the set of Lasso-selected regressors:

$$\hat{\mathcal{A}}_n = \{j : \hat{\beta}_j \neq 0\}.$$

The first-order conditions for $\hat{\beta}_{\hat{\mathcal{A}}_n}$ (the estimated coefficients on the selected regressors) are given by

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,\hat{\mathcal{A}}_n}\left(Y_i - X_{i,\hat{\mathcal{A}}_n}'\hat{\beta}_{\hat{\mathcal{A}}_n}\right) - \lambda_n \,\text{sign}\left(\hat{\beta}_{\hat{\mathcal{A}}_n}\right) = 0, \tag{3.5.1}$$

where for a k-vector x, $\text{sign}(x) = (\text{sign}(x_1), \ldots, \text{sign}(x_k))'$. The set of Lasso-excluded regressors is given by $\hat{\mathcal{A}}_n^c$. We have $j \in \hat{\mathcal{A}}_n^c$ or equivalently $\hat{\beta}_j = 0$, if and only if

$$0 = -\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\left(Y_i - X_{i,\hat{\mathcal{A}}_n}'\hat{\beta}_{\hat{\mathcal{A}}_n}\right) + \lambda_n \cdot \hat{g}_j,$$

$$\text{where } |\hat{g}_j| \leq 1,$$

or

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\left(Y_i - X_{i,\hat{\mathcal{A}}_n}'\hat{\beta}_{\hat{\mathcal{A}}_n}\right)\right| \leq \lambda_n. \tag{3.5.2}$$

Let $\mathcal{A}_0 = \{j : \beta_j \neq 0\}$ denote the set of relevant regressors, so the model can be written as

$$Y_i = X_{i,\mathcal{A}_0}'\beta_{\mathcal{A}_0} + U_i.$$

We can now describe the conditions for Lasso selecting the true regressors correctly. Consider the *sign equality* condition:

$$\text{sign}(\hat{\beta}) = \text{sign}(\beta). \tag{3.5.3}$$

The condition implies that Lasso correctly selects the relevant regressors. This is because for $j \in \mathcal{A}_0^c$, $\beta_j = 0$, and the condition implies $\hat{\beta}_j = 0$. Similarly, for $j \in \mathcal{A}_0$, $\text{sign}(\beta_j) = \pm 1$, and therefore $\hat{\beta}_j \neq 0$. Hence, the sign equality in (3.5.3) implies that $\hat{\mathcal{A}}_n = \mathcal{A}_0$. We have the following result (Wainwright, 2009).

Proposition 3.5.1. *Suppose that* $EX_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}$ *is positive definite. Then* $\operatorname{sign}(\hat{\beta}) = \operatorname{sign}(\beta)$ *if and only if*

(3.5.4)
$$\operatorname{sign}(\beta_{\mathcal{A}_0}) = \operatorname{sign}\left(\beta_{\mathcal{A}_0} + \left(n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}U_i - \lambda_n \operatorname{sign}\left(\beta_{\mathcal{A}_0}\right)\right)\right),$$

and for all $j \in \mathcal{A}_0^c$,

(3.5.5)
$$\left|n^{-1}\sum_{i=1}^{n} X_{i,j}X'_{i,\mathcal{A}_0}\left(n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}U_i - \lambda_n \operatorname{sign}\left(\beta_{\mathcal{A}_0}\right)\right) - \frac{1}{n}\sum_{i=1}^{n} X_{i,j}U_i\right| \leq \lambda_n.$$

Proof. Suppose $\operatorname{sign}(\hat{\beta}) = \operatorname{sign}(\beta)$, so that only the correct regressors are selected. By (3.5.1),

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}\left(Y_i - X'_{i,\mathcal{A}_0}\hat{\beta}_{\mathcal{A}_0}\right) - \lambda_n \operatorname{sign}\left(\beta_{\mathcal{A}_0}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}\left(U_i - X'_{i,\mathcal{A}_0}(\hat{\beta}_{\mathcal{A}_0} - \beta_{\mathcal{A}_0})\right) - \lambda_n \operatorname{sign}\left(\beta_{\mathcal{A}_0}\right), \text{ or} \end{aligned}$$

(3.5.6)
$$\hat{\beta}_{\mathcal{A}_0} = \beta_{\mathcal{A}_0} + \left(n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}U_i - \lambda_n \operatorname{sign}\left(\beta_{\mathcal{A}_0}\right)\right).$$

This implies that

$$\begin{aligned} \operatorname{sign}(\beta_{\mathcal{A}_0}) &= \operatorname{sign}(\hat{\beta}_{\mathcal{A}_0}) \\ &= \operatorname{sign}\left(\beta_{\mathcal{A}_0} + \left(n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}U_i - \lambda_n \operatorname{sign}\left(\beta_{\mathcal{A}_0}\right)\right)\right), \end{aligned}$$

and the condition in (3.5.4) holds. The result in (3.5.5) follows from the fact that the sign equality in (3.5.3) implies that $\hat{\mathcal{A}}_n = \mathcal{A}_0$, and by (3.5.2) and (3.5.6).

Next, suppose that (3.5.4) holds. Consider first the restricted problem that includes only the relevant regressors:

$$\tilde{\beta}_{\mathcal{A}_0} = \arg\min_b \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - X'_{i,\mathcal{A}_0}b\right)^2 + \lambda\sum_{j\in\mathcal{A}_0}|b_j|.$$

The corresponding first-order condition is

$$-\frac{1}{n}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}\left(Y_i - X'_{i,\mathcal{A}_0}\tilde{\beta}_{\mathcal{A}_0}\right) + \lambda_n\tilde{g} = 0,$$

where

$$\tilde{g} \in \partial\left(\sum_{j\in\mathcal{A}_0}|\tilde{\beta}_j|\right).$$

Hence, the restricted estimator satisfies

$$\tilde{\beta}_{\mathcal{A}_0} = \beta_{\mathcal{A}_0} + \left(n^{-1} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} X'_{i,\mathcal{A}_0} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} U_i - \lambda_n \tilde{g} \right). \tag{3.5.7}$$

The restricted estimator $\tilde{\beta}_{\mathcal{A}_0}$ together with $\tilde{\beta}_{\mathcal{A}_0^c} = 0$ is also the solution to the unrestricted original problem if

$$\tilde{g} = \text{sign}(\tilde{\beta}_{\mathcal{A}_0}), \tag{3.5.8}$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_{i,j} \left(Y_i - X'_{i,\mathcal{A}_0} \tilde{\beta}_{\mathcal{A}_0} \right) \right| \leq \lambda_n \quad \text{for all } j \in \mathcal{A}_0^c. \tag{3.5.9}$$

The result in (3.5.4) implies that

$$\text{sign}\left(\tilde{\beta}_{\mathcal{A}_0} \right) = \text{sign}\left(\beta_{\mathcal{A}_0} + \left(n^{-1} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} X'_{i,\mathcal{A}_0} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} U_i - \lambda_n \tilde{g} \right) \right). \tag{3.5.10}$$

By comparing (3.5.10) with (3.5.4), one can see that

$$\tilde{g} = \text{sign}\left(\tilde{\beta}_{\mathcal{A}_0} \right) = \text{sign}\left(\beta_{\mathcal{A}_0} \right)$$

is a feasible solution. The condition in (3.5.9) is also satisfied because of (3.5.5). Hence, $\tilde{\beta}_{\mathcal{A}_0} = \hat{\beta}_{\mathcal{A}_0}$, and the sign equality $\text{sign}(\beta_{\mathcal{A}_0}) = \text{sign}(\hat{\beta}_{\mathcal{A}_0})$ holds. □

For a vector x, define $[x]_j = x_j$. Note that $\text{sign}(x_j) = \pm 1$ if and only if $|x_j| > 0$. Hence, we can restate the result of Proposition 3.5.1 as follows (Belloni and Chernozhukov, 2011, Lemma 4): $\hat{\mathcal{A}}_n = \mathcal{A}_0$, i.e. Lasso correctly selects only the relevant regressors, if and only if for all $j \in \mathcal{A}_0$

$$\left| \beta_j + \left[\left(n^{-1} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} X'_{i,\mathcal{A}_0} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} U_i - \lambda_n \text{sign}\left(\beta_{\mathcal{A}_0} \right) \right) \right]_j \right| > 0, \tag{3.5.11}$$

and for all $j \notin \mathcal{A}_0$

$$\left| n^{-1} \sum_{i=1}^{n} X_{i,j} X'_{i,\mathcal{A}_0} \left(n^{-1} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} X'_{i,\mathcal{A}_0} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i,\mathcal{A}_0} U_i - \lambda_n \text{sign}\left(\beta_{\mathcal{A}_0} \right) \right) - \frac{1}{n} \sum_{i=1}^{n} X_{i,j} U_i \right| \leq \lambda_n. \tag{3.5.12}$$

Suppose that

$$\lambda_n = \sqrt{\frac{\log n}{n}}.$$

The condition in (3.5.11) becomes for all $j \in \mathcal{A}_0$

$$\left| \beta_j + O_p(1) \left(O_p\left(\frac{1}{\sqrt{n}} \right) - O\left(\sqrt{\frac{\log n}{n}} \right) \right) \right| = |\beta_j + o_p(1)| > 0.$$

It is satisfied with probability approaching one as long as $\min_{j\in\mathcal{A}_0}|\beta_j| \geq \delta > 0$ for some δ. Next consider the condition in (3.5.12): for all $j \notin \mathcal{A}_0$

$$\left| O_p\left(\frac{1}{\sqrt{n}}\right) + \lambda_n \cdot n^{-1}\sum_{i=1}^n X_{i,j}X'_{i,\mathcal{A}_0}\left(n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\operatorname{sign}(\beta_{\mathcal{A}_0})\right| \leq \lambda_n.$$

The condition may fail with a positive probability even asymptotically when for some $j \notin \mathcal{A}_0$ the following term is large:

$$\begin{aligned}&n^{-1}\sum_{i=1}^n X_{i,j}X'_{i,\mathcal{A}_0}\left(n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\operatorname{sign}(\beta_{\mathcal{A}_0})\\ =&EX_{i,j}X'_{i,\mathcal{A}_0}\left(EX_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\operatorname{sign}(\beta_{\mathcal{A}_0}) + o_p(1).\end{aligned}$$

Thus, to ensure that Lasso correctly excludes irrelevant regressors

$$EX_{i,j}X'_{i,\mathcal{A}_0}\left(EX_{i,\mathcal{A}_0}X'_{i,\mathcal{A}_0}\right)^{-1}\operatorname{sign}(\beta_{\mathcal{A}_0})$$

has to be below one in absolute value for all $j \notin \mathcal{A}_0$. Such conditions are known in the literature as the *irrepresentability* property (Zhao and Yu, 2006) and imply that the coefficients in the regressions of irrelevant regressors against the relevant ones are small. Thus, in the general case, Lasso may include irrelevant regressors if they are highly correlated with the true regressors in model $\mathcal{A}_0$.

3.6. Weighted and adaptive Lasso

A nice property of regression (and OLS) is that the coefficients automatically adjust when regressors are re-scaled (for example, when the units of measurement change). If we re-scale a regressor by a constant c, the coefficient is adjusted accordingly:

$$Y_i = \beta X_i + U_i = \frac{\beta}{c}(X_i \cdot c) + U_i = \beta^* X_i^* + U_i,$$

where $\beta^* = \beta/c$ and $X_i^* = X_i \cdot c$. The re-scaling would also automatically be applied to the standard errors with no effect on statistical significance. Hence, when estimating a model by OLS, one does not have to worry about the units of measurement of X_i.

Unfortunately, the presence of penalty terms such as $\|b\|_1$ makes estimation sensitive to the units of measurement. Consider

$$\sum_{i=1}^n (Y_i - b_1X_{i,1} - \ldots - b_kX_{i,k})^2 + \lambda\sum_{j=1}^k |b_j|,$$

and suppose we scale $X_{i,1}$ by a large constant $c > 1$. To balance the equation, the coefficient b_1 is now expected to be of a smaller magnitude by $1/c$. It is now more likely that the corresponding Lasso estimator would shrink to zero. Unfortunately, by simply manipulating the units of measurements of different regressors, we can make some of them more (or less) likely to be shrunk to zero by Lasso.

A solution often used in practice (and automatically implemented as the default in some software packages, such as "glmnet()" in R) is to standardize all regressors so that they all

have the same scale. Instead of $X_{i,j}$ one can use

$$\frac{X_{i,j} - \bar{X}_j}{\hat{\sigma}_j},$$

where

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j},$$
$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_{i,j} - \bar{X}_j\right)^2.$$

Since after standardization, all regressors have the same scale (equal to one), the same shrinkage/penalty parameter can be applied to all of them. However, this solution is not ideal as it changes the interpretation of the coefficients.

Since the source of the problem is potentially different scaling of different coefficients, an alternative solution is to apply different shrinkage to different coefficients. This idea is implemented in so-called weighted Lasso. Let $w_1, \ldots, w_k$ be some known (non-negative) weights. Weighted Lasso solves

$$\min_{b_1,\ldots,b_k} \left\{ \frac{1}{2n}\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} b_j X_{i,j} \right)^2 + \lambda \sum_{j=1}^{k} w_j |b_j| \right\}.$$

For example, to adjust for the different scales of different regressors, one may use

$$w_j = \sqrt{n^{-1}\sum_{i=1}^{n} X_{i,j}^2}.$$

Alternatively, *adaptive* Lasso adjusts the weights for individual coefficients in a data-dependent manner based on preliminary estimates of the coefficients. Let $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ denote some preliminary estimates of the true coefficients $\beta_1, \ldots, \beta_k$. For example, they can be the OLS estimates or the Ridge estimates when OLS cannot be computed. Adaptive Lasso sets the weights as

$$w_j = \frac{1}{|\tilde{\beta}_j|}, \quad j = 1, \ldots, k.$$

The main idea of adaptive Lasso is that if $\tilde{\beta}_j$ is a good initial guess for β_j, e.g. because $\tilde{\beta}_j \to_p \beta_j$, it can be used to find the amount of shrinkage to be applied to different parameters. We can expect $|\tilde{\beta}_j|$ to be large for β_j that is further away from zero, resulting in a smaller weight in the penalty term. Similarly, larger weights are given to coefficients closer to zero.

Suppose that $\tilde{\beta}_j = \beta_j + O_p(n^{-1/2})$, in which case

$$w_j = \frac{1}{|\tilde{\beta}_j|} = \frac{1}{\left|\beta_j + O_p(n^{-1/2})\right|},$$

it follows that

$$w_j = \begin{cases} O_p(1) & j \in \mathcal{A}_0, \\ O_p(\sqrt{n}) & j \notin \mathcal{A}_0. \end{cases}$$

Suppose that we set

$$\lambda_n = \frac{\sqrt{\log n}}{n}.$$

Correct regressors are selected if for all $j \in \mathcal{A}_0$

$$\begin{aligned} &\left| \beta_j + \left[\left(n^{-1} \sum_{i=1}^n X_{i,\mathcal{A}_0} X_{i,\mathcal{A}_0}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_{i,\mathcal{A}_0} U_i - O\left(\frac{\sqrt{\log n}}{n} \right) \operatorname{sign}\left(\beta_{\mathcal{A}_0} \right) \right) \right]_j \right| \\ =& \left| \beta_j + O_p\left(\frac{1}{\sqrt{n}} \right) + O_p\left(\frac{\sqrt{\log n}}{n} \right) \right| \\ >& 0. \end{aligned}$$

The condition holds provided that the true non-zero coefficients are bounded away from zero in absolute value.

By (3.5.12), adaptive Lasso eliminates the irrelevant regressors if for all $j \notin \mathcal{A}_0$

$$\begin{aligned} \left| n^{-1} \sum_{i=1}^n X_{i,j} X_{i,\mathcal{A}_0}' \left(n^{-1} \sum_{i=1}^n X_{i,\mathcal{A}_0} X_{i,\mathcal{A}_0}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_{i,\mathcal{A}_0} U_i - O_p\left(\frac{\sqrt{\log n}}{n} \right) \right) \right. \\ \left. - \frac{1}{n} \sum_{i=1}^n X_{i,j} U_i \right| \leq O_p\left(\sqrt{\frac{\log n}{n}} \right). \end{aligned}$$

Note that the rates from the penalty term are now different inside and outside the absolute value, with the outside term converging to zero at a slower rate. We now have that adaptive Lasso excludes regressors $j \notin \mathcal{A}_0$ when

$$\left| O_p\left(\sqrt{\frac{1}{n}} \right) + O_p\left(\frac{\sqrt{\log n}}{n} \right) \right| \leq O_p\left(\sqrt{\frac{\log n}{n}} \right),$$

which holds with probability approaching one. Thus, adaptive Lasso does not require the irrepresentability condition for consistency.

3.7. Sparse high-dimensional models

Suppose that the number of potential regressors is large

$$k \to \infty \text{ as } n \to \infty.$$

It is often assumed in such cases that the true model $\mathcal{A}_0$ is small, and $|\mathcal{A}_0|$ is fixed: only a small (fixed) number of the potential k regressors have non-zero coefficients. In such cases, we say that $\mathcal{A}_0$ is *sparse*. In order to eliminate many irrelevant regressors, we need that

$$\sup_{j \notin \mathcal{A}_0} \left| \frac{1}{n} \sum_{i=1}^n X_{i,j} U_i + \lambda_n \cdot E X_{i,j} X_{i,\mathcal{A}_0}' \left(E X_{i,\mathcal{A}_0} X_{i,\mathcal{A}_0}' \right)^{-1} \operatorname{sign}\left(\beta_{\mathcal{A}_0} \right) + o_p(1) \right| \leq \lambda_n.$$

Provided that the irrepresentability condition holds, we have to consider

$$\sup_{j \notin \mathcal{A}_0} \left| \frac{1}{n} \sum_{i=1}^{n} X_{i,j} U_i \right|,$$

where the supremum is over a large number of terms of order k with $k \to \infty$.

Recall that $n^{-1/2} \sum_{i=1}^{n} X_{i,j} U_i$ is asymptotically normal:

$$n^{-1/2} \sum_{i=1}^{n} X_{i,j} U_i \to_d \xi_k \sim N(0, \sigma^2),$$

where for simplicity we assume that $E(U_i^2 \mid X_{i,j}) = \sigma^2$ and $n^{-1} \sum_{i=1}^{n} X_{i,j}^2 = 1$. One can show that

$$E\left(\max_{1 \le j \le k} |\xi_j| \right) \le \sqrt{2\sigma^2 \log k} + O\left(\frac{1}{\sqrt{\log k}} \right).$$

Hence,

$$\sup_{j \notin \mathcal{A}_0} \left| \frac{1}{n} \sum_{i=1}^{n} X_{i,j} U_i \right| = O_p\left(\sqrt{\frac{2\sigma^2 \log k}{n}} \right).$$

Thus, to account for a large number of potential regressors, we can set the penalty parameter as

$$\lambda_n \sim \sqrt{\frac{2\sigma^2 \log(kn)}{n}}.$$

Note that σ^2 in the above expression measures the noise level and is used to adjust the penalty parameter for the scale of the residual term U_i. Estimation of σ^2 is non-trivial in high dimensional models, and Belloni and Chernozhukov (2011) suggest an iterative approach. First, use $\hat{\sigma}_Y^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, where $\bar{Y}$ is the average of Y's in the sample. The estimator is conservative as it also contains the portion of the variation of Y due to X's. However, $\hat{\sigma}_Y^2$ can be used for the first pass of Lasso. Given the corresponding Lasso estimates $\check{\beta}$, one can compute

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - X_i' \check{\beta})^2,$$

which can be used to determine the penalty parameter and a subsequent application of Lasso.

CHAPTER 4

Post- and double- Lasso

Post-Lasso is an OLS regression that includes only the controls that survive the Lasso selection step. Since Lasso is unlikely to detect regressors with small coefficients, the naive application of post-Lasso can result in substantial bias. The bias in post-Lasso can be eliminated using the double-Lasso or the partialling out approaches covered in this section. The double-Lasso or partialling out steps are needed when the goal is the consistent estimation of some important coefficients instead of the prediction of outcome variables.

4.1. Post-Lasso

Consider the regression model

$$Y_i = \alpha D_i + X_i'\beta + U_i, \tag{4.1.1}$$
$$E(U_i \mid D_i, X_i) = 0,$$

where D_i is the main regressor of interest, and X_i includes k potential covariates or controls: $\beta = (\beta_1, \ldots, \beta_k)'$. The researcher wants to always include D_i in the regression but needs to select a list of relevant controls from X_i. Thus, when estimating the model by Lasso, the coefficient on D_i is excluded from the penalty term.

Let $\mathcal{A}_0$ denote the set of relevant controls:

$$\mathcal{A}_0 = \{j \in \{1, \ldots, k\} : \beta_j \neq 0\}.$$

Let $\hat{\beta}_\lambda = (\hat{\beta}_{1,\lambda}, \ldots, \hat{\beta}_{k,\lambda})'$ denote a Lasso estimator of $\beta = (\beta_1, \ldots, \beta_k)'$. Such estimators are constructed, for example, using a Lasso regression of Y_i against D_i and X_i with no penalty applied to the coefficient on D_i. The estimated set of selected controls is given by

$$\hat{\mathcal{A}} = \left\{j \in \{1, \ldots, k\} : \hat{\beta}_j \neq 0\right\}.$$

Let $X_{i,\hat{\mathcal{A}}}$ denote the sub-vector of X_i that only the controls in $\hat{\mathcal{A}}$. A post-Lasso estimator of α can be constructed using the OLS regression of Y_i against D_i and $X_{i,\hat{\mathcal{A}}}$:

$$Y_i = \hat{\alpha}(\hat{\mathcal{A}}) \cdot D_i + X_{i,\hat{\mathcal{A}}}'\hat{\beta}_{\hat{\mathcal{A}}} + \hat{U}_i.$$

Here the notation $\hat{\alpha}(\hat{\mathcal{A}})$ is used to indicate that the estimator is constructed using the Lasso-selected set of controls $\hat{\mathcal{A}}$. Note that the coefficients on the vector of included controls $X_{i,\hat{\mathcal{A}}}$ are re-estimated in post-Lasso, since the Lasso estimates $\hat{\beta}_\lambda$ are biased.

The main concern about the post-Lasso estimator $\hat{\alpha}(\hat{\mathcal{A}})$ is if its properties are affected by the Lasso-selection first stage. Let $\hat{\alpha}(\mathcal{A}_0)$ denote the infeasible OLS estimator of α when the

true set of controls $\mathcal{A}_0$ is known:

$$Y_i = \hat{\alpha}(\mathcal{A}_0) \cdot D_i + X'_{i,\mathcal{A}_0}\tilde{\beta}_{\mathcal{A}_0} + \tilde{U}_i,$$

where $X_{i,\mathcal{A}_0}$ denotes the true vector of controls defined by $\mathcal{A}_0$.

The following proposition exploits the consistent selection property of Lasso.

Proposition 4.1.1. *Suppose that* $P\left(\hat{\mathcal{A}} = \mathcal{A}_0\right) \to 1$ *as* $n \to \infty$. *Suppose further that* $\sqrt{n}(\hat{\alpha}(\mathcal{A}_0) - \alpha) \to_d N(0, \omega^2(\mathcal{A}_0))$, *where* $\omega^2(\mathcal{A}_0) > 0$ *denotes the asymptotic variance of the infeasible estimator* $\hat{\alpha}(\mathcal{A}_0)$ *when the true model is known. Then,*

$$\sqrt{n}(\hat{\alpha}(\hat{\mathcal{A}}) - \alpha) \to_d N(0, \omega^2(\mathcal{A}_0)).$$

PROOF. Let $\Phi(\cdot)$ denote the standard normal CDF. Note that if $\xi \sim N(0, \omega^2)$, then $Z = \xi/\omega \sim N(0,1)$, and $P(\xi \leq x) = P(\xi/\omega \leq x/\omega) = P(Z \leq x/\omega) = \Phi(x/\omega)$. For $x \in \mathbb{R}$,

$$\begin{aligned}
&P\left(\sqrt{n}(\hat{\alpha}(\hat{\mathcal{A}}) - \alpha) \leq x\right) \\
&= P\left(\sqrt{n}(\hat{\alpha}(\hat{\mathcal{A}}) - \alpha) \leq x, \hat{\mathcal{A}} = \mathcal{A}_0\right) + P\left(\sqrt{n}(\hat{\alpha}(\hat{\mathcal{A}}) - \alpha) \leq x, \hat{\mathcal{A}} \neq \mathcal{A}_0\right) \\
&= P\left(\sqrt{n}(\hat{\alpha}(\mathcal{A}_0) - \alpha) \leq x\right) + o(1) \\
&\to \Phi\left(\frac{x}{\omega(\mathcal{A}_0)}\right),
\end{aligned}$$

where $o(1)$ denotes terms converging to zero as $n \to \infty$. The second equality holds by the following argument:

$$0 \leq P\left(\sqrt{n}(\hat{\alpha}(\hat{\mathcal{A}}) - \alpha) \leq x, \hat{\mathcal{A}} \neq \mathcal{A}_0\right) \leq P\left(\hat{\mathcal{A}} \neq \mathcal{A}_0\right) \to 0.$$

□

Proposition 4.1.1 suggests that the post-Lasso estimator can be as good as the OLS estimator under a known $\mathcal{A}_0$. However, the results require Lasso to detect true controls with probability approaching one. Unfortunately, that does not hold for controls in $\mathcal{A}_0$ with small non-zero coefficients.

Suppose that

$$\frac{X'X}{n} = I_k,$$

and recall that, in this case, the Lasso estimator satisfies

$$\hat{\beta}_{j,\lambda} = \text{sign}(\tilde{\beta}_j)\left(|\tilde{\beta}_j| - \lambda\right)^+,$$

where $\tilde{\beta}_j$ is the corresponding OLS estimator, and

$$\tilde{\beta}_j = \beta_j + O_p\left(\frac{1}{\sqrt{n}}\right).$$

Recall that we set the penalty parameter λ to over-rule the noise

$$\lambda = 2\sigma\sqrt{\frac{2\log(kn)}{n}}.$$

The Lasso coefficient $\hat{\beta}_{j,\lambda}$ is shrunk all the way to zero when $|\tilde{\beta}_j| < \lambda$.

Suppose that β_j is small. Since the magnitude of statistics is measured relative to the $1/\sqrt{n}$ rate of convergence of the noise part, small coefficients can be modeled as

$$\beta_j = \frac{c}{\sqrt{n}}$$

for some constant c. I.e. the coefficient β_j is of the same magnitude as estimation noise. Such coefficients will be eliminated by Lasso with probability approaching one:

$$P\left(\left|\frac{c}{\sqrt{n}} + O_p\left(\frac{1}{\sqrt{n}}\right)\right| < 2\sigma\sqrt{\frac{2\log(kn)}{n}}\right) \to 1.$$

This simple model illustrates that Lasso cannot distinguish small regression coefficients from the noise, and the corresponding controls will be selected by Lasso with a high probability.

4.2. Bias of a naive post-Lasso estimator

One may wonder that since Lasso eliminates only controls with very small coefficients, it might not have a substantial impact on the post-Lasso estimator of the main parameter of interest. However, that depends on the relationship between the omitted controls and the main regressor, as we illustrate in the following simple example.

Consider the following model with the main regressor D_i and a single control variable X_i:

$$Y_i = \alpha D_i + \beta X_i + U_i, \tag{4.2.1}$$

i.e. $\beta \in \mathbb{R}$. Suppose that the coefficient β is small in the sense discussed above

$$\beta = \frac{c}{\sqrt{n}}.$$

Suppose further that Lasso has eliminated X_i from the model and $\hat{\mathcal{A}} = \emptyset$. The post-Lasso estimator of α is estimated by a simple regression of Y_i against D_i:

$$\hat{\alpha}(\emptyset) = \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i^2} = \alpha + \beta\frac{\sum_{i=1}^n D_i X_i}{\sum_{i=1}^n D_i^2} + \frac{\sum_{i=1}^n D_i U_i}{\sum_{i=1}^n D_i^2},$$

or

$$\sqrt{n}(\hat{\alpha}(\emptyset) - \alpha) = c\frac{\sum_{i=1}^n D_i X_i}{\sum_{i=1}^n D_i^2} + \frac{n^{-1/2}\sum_{i=1}^n D_i U_i}{n^{-1}\sum_{i=1}^n D_i^2}.$$

The second term on the right-hand side of the above equation is the usual asymptotically normal component with a zero mean. The asymptotic bias of the post-Lasso estimator is determined by the first term. Note that

$$\hat{\gamma} = \frac{\sum_{i=1}^n D_i X_i}{\sum_{i=1}^n D_i^2}$$

is the OLS estimator in the simple regression of the control X_i against the main regressor D_i:

$$X_i = \gamma D_i + \eta_i.$$

Hence,

$$\hat{\gamma} = \gamma + O_p\left(\frac{1}{\sqrt{n}}\right),$$

and unless $\gamma = o(1)$, the naive post-Lasso estimator $\hat{\alpha}(\emptyset)$ suffers from an asymptotic bias.

We conclude that the bias of the naive post-Lasso estimator can be substantial if there is a substantial correlation between the main regressor D_i and controls with small non-zero coefficients.

4.3. Double Lasso

Belloni et al. (2014) developed a procedure that addresses the shortcoming of the naive Lasso approach. Since the bias of a post-Lasso estimator depends on the magnitude of the correlation between the main regressor D_i and the control variables in X_i, one can use a Lasso regression of D_i against X_i to detect correlated controls.

Consider the following regression estimated by Lasso:

$$D_i = \sum_{j=1}^{k} \rho_j X_{i,j} + \eta_i. \tag{4.3.1}$$

Controls with large enough coefficients ρ_j will be selected by Lasso with a probability approaching one. Controls with small coefficients ρ_j will be dropped by Lass. However, in view of the discussion in the previous section, omitting such controls does not result in a substantial bias for the post-Lasso estimator of α.

By combining equations (4.1.1) and (4.3.1), we can write a reduced-form equation that connects Y_i with controls in X_i:

$$\begin{aligned} Y_i &= \alpha \left(\sum_{j=1}^{k} \rho_j X_{i,j} + \eta_i \right) + \sum_{j=1}^{k} \beta_j X_{i,j} + U_i \\ &= \sum_{j=1}^{k} \left(\alpha \rho_j + \beta_j \right) X_{i,j} + \left(\alpha \eta_i + U_i \right) \\ &= \sum_{j=1}^{k} \pi_j X_{i,j} + \epsilon_i, \end{aligned} \tag{4.3.2}$$

where

$$\begin{aligned} \pi_j &= \alpha \rho_j + \beta_j, \\ \epsilon_i &= \alpha \eta_i + U_i. \end{aligned}$$

Hence, a control $X_{i,j}$ is useful for predicting Y_i if it directly affects Y_i through β_j, or affects D_i through ρ_j , or both.

Let $\hat{\rho}_\lambda = (\hat{\rho}_{1,\lambda}, \ldots, \hat{\rho}_{k,\lambda})'$ denote the Lasso estimator for equation (4.3.1). Let $\hat{\pi}_\lambda = (\hat{\pi}_{1,\lambda}, \ldots, \hat{\pi}_{k,\lambda})'$ be the Lasso estimator for equation (4.3.2). The double Lasso algorithm proposed in Belloni et al. (2014) works as follows.

Step 1 Use Lasso on equation (4.3.1) to select controls that are useful for predicting D_i. Let $\hat{\mathcal{A}}_D$ denote the corresponding set of selected controls:

$$\hat{\mathcal{A}}_D = \left\{ j \in \{1, \ldots, k\} : \hat{\rho}_{j,\lambda} \neq 0 \right\}.$$

Step 2 Use Lasso on equation (4.3.2) to select controls that are useful for predicting Y_i. Let $\hat{\mathcal{A}}_Y$ denote the corresponding set of selected controls:

$$\hat{\mathcal{A}}_Y = \{j \in \{1, \ldots, k\} : \hat{\pi}_{j,\lambda} \neq 0\}.$$

Step 3 Estimate α using the OLS regression of Y_i against D_i and controls in $\hat{\mathcal{A}}_D \cup \hat{\mathcal{A}}_Y$.

Note that post-Lasso Step 3 in the above algorithm includes controls selected either for predicting D_i or for predicting Y_i. A control is excluded from the post-Lasso step only if it was dropped in steps 1 and 2 of the algorithm. I.e., a control $X_{i,j}$ is excluded from the post-Lasso step when both ρ_j and β_j are small, which provides protection against the post-Lasso bias.

The double Lasso procedure with post-Lasso is implemented in R package "hdm" (see Chernozhukov et al., 2016a,b).

4.4. A partialling out approach

Another post-Lasso approach that avoids the bias of the naive post-Lasso estimator is based on the partialling out idea or the orthogonality principle. Recall that the OLS estimator of α in (4.1.1) can be written as

$$\tilde{\alpha}_{OLS} = \frac{D' M_X Y}{D' M_X D},$$

where

$$M_X = I_n - X(X'X)^{-1}X',$$

and

$$\tilde{Y} = M_X Y, \tag{4.4.1}$$

$$\tilde{D} = M_X D, \tag{4.4.2}$$

denote the residuals from the respective OLS regressions of Y against X, and of D against X. Hence, the OLS estimator of α can be written as

$$\tilde{\alpha}_{OLS} = \frac{\tilde{D}'\tilde{Y}}{\tilde{D}'\tilde{D}},$$

where the equality holds because M_X is symmetric and idempotent. In other words, $\tilde{\alpha}_{OLS}$ can be constructed by regressing the residuals $\tilde{Y}$ of the dependent variable against the residuals $\tilde{D}$ of the main regressor.

When there are many potential covariates in X, $\tilde{\alpha}_{OLS}$ can have a very large variance as we discussed in Chapter 1. The partialling out approach proposes to replace (4.4.1)–(4.4.2) with the residuals from the corresponding post-Lasso regressions, see Chernozhukov et al. (2016b).

Step 1 Use Lasso on equation (4.3.1) to select controls that are useful for predicting D_i. Let $\hat{\mathcal{A}}_D$ denote the corresponding set of selected controls:

$$\hat{\mathcal{A}}_D = \{j \in \{1, \ldots, k\} : \hat{\rho}_{j,\lambda} \neq 0\}.$$

Regress D_i on the controls in $\hat{\mathcal{A}}_D$, and save the residuals as $\tilde{D}_i^{PL}$.

Step 2 Use Lasso on equation (4.3.2) to select controls that are useful for predicting Y_i. Let $\hat{\mathcal{A}}_Y$ denote the corresponding set of selected controls:

$$\hat{\mathcal{A}}_Y = \{j \in \{1, \ldots, k\} : \hat{\pi}_{j,\lambda} \neq 0\}.$$

Regress Y_i on the controls in $\hat{\mathcal{A}}_Y$, and save the residual as $\tilde{Y}_i^{PL}$.

Step 3 Estimate α using the OLS regression of $\tilde{Y}_i^{PL}$ against $\tilde{D}_i^{PL}$.

We compare the partialling out approach with a naive post-Lasso estimator $\hat{\alpha}_{Naive}$. The latter is constructed by first using a Lasso regression of Y_i against D_i and all X_i's, selecting controls from X_i, and then regressing Y_i against D_i and the selected controls. Let $\hat{\beta}_{Naive}$ denote the estimated coefficients on X_i in the second stage with zero values for the controls dropped by Lasso. The naive post-Lasso estimator $\hat{\alpha}_{Naive}$ satisfies

$$\hat{\alpha}_{Naive} = \frac{\sum_{i=1}^n D_i(Y_i - X_i'\hat{\beta}_{Naive})}{\sum_{i=1}^n D_i^2} = \alpha + \frac{\sum_{i=1}^n D_i(U_i - X_i'(\hat{\beta}_{Naive} - \beta))}{\sum_{i=1}^n D_i^2},$$

or

$$\sqrt{n}(\hat{\alpha}_{Naive} - \alpha) = \frac{n^{-1/2}\sum_{i=1}^n D_i U_i}{n^{-1}\sum_{i=1}^n D_i^2} + \frac{1}{n^{-1}\sum_{i=1}^n D_i^2}\sum_{j=1}^k \sqrt{n}(\hat{\beta}_{j,Naive} - \beta_j)\frac{1}{n}\sum_{i=1}^n D_i X_{i,j}. \quad (4.4.3)$$

Suppose that $\beta_j = c/\sqrt{n}$ for some j and therefore $\hat{\beta}_{j,Naive} = 0$. If D_i and $X_{i,j}$ are correlated,

$$\frac{1}{n}\sum_{i=1}^n D_i X_{i,j} \to_p ED_i X_{i,j} \neq 0,$$

and as a result, the naive post-Lasso estimator $\hat{\alpha}_{Naive}$ suffers from an asymptotic bias.

For the post-Lasso estimator with partialling out, D_i in (4.4.3) is replaced with $\tilde{D}_i^{PL}$. If the correlation between D_i and $X_{i,j}$ is sufficiently strong, $X_{i,j}$ will be selected in Step 1 of the algorithm with a probability approaching one. In this case, by construction,

$$\frac{1}{n}\sum_{i=1}^n \tilde{D}_i^{PL} X_{i,j} = 0.$$

Hence, the partialling out step protects post-Lasso against the bias due to small β_j's.

The partialling out procedure is implemented in the R package "hdm" (see Chernozhukov et al., 2016a,b).

CHAPTER 5

Lasso and instrumental variables estimation

This chapter discusses Lasso and post-Lasso-based methods for instrumental variable (IV) estimation. We discuss several scenarios: many potential IVs and few controls, many potential controls and few IVs, and many potential IVs and many controls.

5.1. Instrumental variables

Consider the model:

$$Y_i = \alpha D_i + U_i, \tag{5.1.1}$$

where D_i is the main endogenous regressor of interest:

$$E(U_i \mid D_i) \neq 0.$$

and recall that the OLS estimator of α is inconsistent. We assume that there is an $l \times 1$ vector of potential instruments Z_i such that

$$E(U_i \mid Z_i) = 0, \tag{5.1.2}$$

and

$$D_i = Z_i'\pi + V_i, \tag{5.1.3}$$

$$E(V_i \mid Z_i) = 0, \tag{5.1.4}$$

where $\pi = (\pi_1, \ldots, \pi_l)'$.

The equations (5.1.1) and (5.1.3) can be written in the matrix form as

$$Y = \alpha D + U,$$
$$D = Z\pi + V,$$

where Y is the $n \times 1$ vector of observations on the dependent variable, D is the $n \times 1$ vector of observations on the regressor, and Z is the $n \times l$ matrix of the instruments. The $n \times 1$ vectors U and V are defined similarly. The 2SLS estimator of α is given by

$$\hat{\alpha} = \frac{D'P_Z Y}{D'P_Z D},$$

where

$$P_Z = Z(Z'Z)^{-1}Z'.$$

Note that

$$\hat{D} = P_Z D = Z(Z'Z)^{-1}Z'D = Z\hat{\pi},$$

where

$$\hat{\pi} = (Z'Z)^{-1}Z'D$$

is the OLS estimator of π in (5.1.3). The 2SLS estimator can be equivalently written as

$$\hat{\alpha} = \frac{\hat{D}'Y}{\hat{D}'D} = \frac{\sum_{i=1}^n \hat{D}_i Y_i}{\sum_{i=1}^n \hat{D}_i D_i}.$$

We say that

$$\hat{D}_i = \sum_{j=1}^{l} \hat{\pi}_j Z_{i,j}$$

is the IV for D_i.

As we discuss in Chapter 1, the 2SLS estimator is inconsistent when there are many IVs: $l/n \to c > 0$. However, suppose that the first-stage equation (5.1.3) is sparse. Define

$$\mathcal{A}_0 = \{j \in \{1, \ldots, l\} : \pi_j \neq 0\}.$$

Let l^* denote the number of elements in $\mathcal{A}_0$:

$$l^* = |\mathcal{A}_0|.$$

While it is possible that there are many potential IVs, $l \to \infty$ as $n \to \infty$, we assume that l^* is small and keep l^* as fixed.

Let $Z_{i,\mathcal{A}_0}$ denote the l^*-sub-vector of Z_i that consists only of the IVs in $\mathcal{A}_0$. Let $\pi_{\mathcal{A}_0}$ denote the corresponding sub-vector of π. One can show that when the residuals in (5.1.1) are homoskedastic, i.e.

$$E\left(U_i^2 \mid D_i\right) = \sigma^2, \tag{5.1.5}$$

the efficient IV estimator of α is given by

$$\hat{\alpha}^* = \frac{\sum_{i=1}^n (Z'_{i,\mathcal{A}_0}\pi_{\mathcal{A}_0})Y_i}{\sum_{i=1}^n (Z'_{i,\mathcal{A}_0}\pi_{\mathcal{A}_0})D_i}.$$

In other words, the best IV is given by

$$E\left(D_i \mid Z_i\right) = Z'_{i,\mathcal{A}_0}\pi_{\mathcal{A}_0}.$$

Note that the above equation follows from (5.1.3) and (5.1.4).

Proposition 5.1.1. *Suppose that iid data are generated according to (5.1.1), (5.1.2), (5.1.3), (5.1.4), and (5.1.5). Then,*

$$\sqrt{n}\left(\hat{\alpha}^* - \alpha\right) \to_d N\left(0, \frac{\sigma^2}{E(Z'_{i,\mathcal{A}_0}\pi_{\mathcal{A}_0})^2}\right).$$

Proof. To simplify the notation, define

$$\zeta_i^* = Z'_{i.\mathcal{A}_0}\pi_{\mathcal{A}_0}, \tag{5.1.6}$$

and write the first stage as

$$D_i = \zeta_i^* + V_i.$$

The estimator $\hat{\alpha}^*$ satisfies

$$\sqrt{n}\left(\hat{\alpha}^* - \alpha\right) = \frac{n^{-1/2}\sum_{i=1}^n \zeta_i^* U_i}{n^{-1}\sum_{i=1}^n \zeta_i^*\left(\zeta_i^* + V_i\right)}.$$

Note that

$$E\zeta_i^* U_i = E\left(\zeta_i^* E(U_i \mid Z_i)\right) = 0. \tag{5.1.7}$$

By the CLT,

$$n^{-1/2}\sum_{i=1}^n \zeta_i^* U_i \to_d N\left(0, E\left(\zeta_i^* U_i\right)^2\right),$$

and

$$E\left(\zeta_i^* U_i\right)^2 = E\left(\zeta_i^{*2} E(U_i^2 \mid Z_i)\right) = \left(E\zeta_i^{*2}\right)\sigma^2.$$

Similarly to (5.1.7),

$$E\zeta_i^* V_i = 0.$$

Hence, by LLN,

$$n^{-1}\sum_{i=1}^n \zeta_i^*\left(\zeta_i^* + V_i\right) \to_p E\zeta_i^{*2}.$$

We conclude

$$\sqrt{n}\left(\hat{\alpha}^* - \alpha\right) \to_d \frac{N\left(0, \left(E\zeta_i^{*2}\right)\sigma^2\right)}{E\zeta_i^{*2}} = N\left(0, \frac{\left(E\zeta_i^{*2}\right)\sigma^2}{\left(E\zeta_i^{*2}\right)^2}\right) = N\left(0, \frac{\sigma^2}{E\zeta_i^{*2}}\right),$$

and the result follows by (5.1.6). □

While other functions of Z_i can be used to instrument D_i ,

$$\zeta_i^* = Z'_{i.\mathcal{A}_0}\pi_{\mathcal{A}_0}$$

is the most efficient IV as we show below. Define a function of IVs

$$\zeta_i = f(Z_i),$$

where

$$f : \mathbb{R}^l \to \mathbb{R}.$$

Note that

$$E\zeta_i U_i = 0,$$

which holds by exactly the same arguments as in (5.1.7) in the proof of Proposition 5.1.1. Hence, $\zeta_i = f(Z_i)$ is an IV if

$$E\zeta_i D_i = E\zeta_i\zeta_i^* \neq 0.$$

The IV estimator corresponding to f is given by

$$\hat{\alpha}_f = \frac{\sum_{i=1}^n \zeta_i Y_i}{\sum_{i=1}^n \zeta_i D_i}.$$

We have the following result.

Proposition 5.1.2. *Suppose that the assumptions of Proposition 5.1.1 hold and*

$$E\zeta_i\zeta_i^* \neq 0.$$

Then,

$$\sqrt{n}\left(\hat{\alpha}_f - \alpha\right) \to_d N\left(0, \frac{\sigma^2 E\zeta_i^2}{\left(E\zeta_i\zeta_i^*\right)^2}\right).$$

Moreover,

$$\frac{\sigma^2 E\zeta_i^2}{\left(E\zeta_i\zeta_i^*\right)^2} \geq \frac{\sigma^2}{E\zeta_i^{*2}}.$$

Proof. Write

$$\begin{aligned}
\sqrt{n}\left(\hat{\alpha}_f - \alpha\right) &= \frac{n^{-1/2}\sum_{i=1}^n \zeta_i U_i}{n^{-1}\sum_{i=1}^n \zeta_i D_i} \\
&= \frac{n^{-1/2}\sum_{i=1}^n \zeta_i U_i}{n^{-1}\sum_{i=1}^n \zeta_i(\zeta_i^* + V_i)} \\
&\to_d \frac{N(0, \sigma^2 E\zeta_i^2)}{E\zeta_i\zeta_i^*} \\
&= N\left(0, \frac{\sigma^2 E\zeta_i^2}{\left(E\zeta_i\zeta_i^*\right)^2}\right),
\end{aligned}$$

which establishes the first claim.

For the second claim,

$$E\zeta_i^{*2} - \frac{\left(E\zeta_i\zeta_i^*\right)^2}{E\zeta_i^2} = \frac{E\zeta_i^{*2}E\zeta_i^2 - \left(E\zeta_i\zeta_i^*\right)^2}{E\zeta_i^2},$$

and the result follows by Cauchy-Schwartz inequality

$$\left(E\zeta_i\zeta_i^*\right)^2 \leq E\zeta_i^{*2}E\zeta_i^2.$$

□

Note that in practice, the first-stage equation can be a non-linear function $f^*(\cdot)$ of a small number of some "primitive" IVs $W_i = (W_{i,1}, \ldots, W_{i,p})'$:

$$\begin{aligned}
D_i &= f^*(W_i) + V_i, \\
E(U_i \mid W_i) &= 0, \\
E(V_i \mid W_i) &= 0,
\end{aligned}$$

where the function $f^* : \mathbb{R}^p \to \mathbb{R}$ is unknown. In that case, the efficient IV is given by

$$\zeta_i^* = f^*(W_i).$$

Since the function $f^*(W_i)$ is unknown, we can try to approximate it using a polynomial approximation with interaction terms between the primitive IVs:

$$\begin{aligned} f^*(W_i) = \delta_1 W_{i,1} + \delta_2 W_{i,2} + \ldots + \delta_p W_{i,p} \\ + \delta_{p+1} W_{i,1}^2 + \delta_{p+2} W_{i,1} W_{i,2} \ldots + \delta_{2p} W_{i,1} W_{i,p} \\ + \delta_{2p+1} W_{i,2}^2 + \ldots . \end{aligned}$$

The vector of IVs Z_i then will be given by

$$Z_i' = \left(W_{i,1}, W_{i,2}, \ldots, W_{i,p}, W_{i,1}^2, (W_{i,1} W_{i,2}), \ldots, (W_{i,1} W_{i,p}), W_{i,2}^2, \ldots\right).$$

If we try to approximate $f^*(\cdot)$ as closely as possible, we can end up with a long list of polynomial and interaction terms between the primitive IVs with potentially only a few of them playing a significant role in approximating $f^*(\cdot)$. In such cases, we can write

$$f^*(W_i) = Z_{i,\mathcal{A}_0}' \pi_{\mathcal{A}_0} + r_i,$$

where $\mathcal{A}_0$ is now the list of a small number of important approximation terms in Z_i, and r_i is a small (remainder) approximation error. As discussed in Belloni et al. (2012, p. 2378), there is an approximating set of "effective" IVs such that

$$l^* = |\mathcal{A}_0| = o(n),$$

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} r_i^2} = O_p\left(\sqrt{\frac{s}{n}}\right).$$

To avoid the bias of many IVs and to estimate α as precisely as possible, we need to be able to select the effective IVs $Z_{i,\mathcal{A}_0}$.

Belloni et al. (2012) proposed the following algorithm for Lasso/post-Lasso selection of IVs and estimation of the IV regression model defined by (5.1.1) and (5.1.3).

Step 1 Estimate the first stage equation in (5.1.3) using Lasso. Let $\hat{\mathcal{A}}$ denote the set of selected IVs:

$$\hat{\mathcal{A}} = \{j \in \{1, \ldots, l\} : \hat{\pi}_{j,\lambda} \neq 0\},$$

where $\hat{\pi}_\lambda = (\hat{\pi}_{1,\lambda}, \ldots, \hat{\pi}_{l,\lambda})'$ is the vector of the corresponding Lasso-estimated coefficients.

Step 2 Estimate the first stage by OLS using only the instruments in $\hat{\mathcal{A}}$. Construct

$$\hat{\zeta}_i(\hat{\mathcal{A}}) = Z_{i,\hat{\mathcal{A}}}' \hat{\pi}_{\hat{\mathcal{A}}},$$

where $Z_{i,\hat{\mathcal{A}}}$ is the sub-vector of the instruments selected in Step 1, and $\hat{\pi}_{\hat{\mathcal{A}}}$ is the vector of post-Lasso OLS estimates for the first stage equation:

$$\hat{\pi}_{\hat{\mathcal{A}}} = \left(\sum_{i=1}^{n} Z_{i,\hat{\mathcal{A}}} Z_{i,\hat{\mathcal{A}}}'\right)^{-1} \sum_{i=1}^{n} Z_{i,\hat{\mathcal{A}}} D_i.$$

Step 3 Estimate (5.1.1) using $\hat{\zeta}_i(\hat{\mathcal{A}})$ as the IV for D_i:

$$\hat{\alpha}(\hat{\mathcal{A}}) = \frac{\sum_{i=1}^n \hat{\zeta}_i(\hat{\mathcal{A}})Y_i}{\sum_{i=1}^n \hat{\zeta}_i(\hat{\mathcal{A}})D_i}.$$

Note that in Step 2, we use the post-Lasso estimates of the coefficients on the selected IVs instead of the Lasso estimates from Step 1 to eliminate the shrinkage bias.

With a probability approaching one, the above algorithm in Step 1 selects the few instruments that have a significant relationship to D_i. The dropped IVs are either unrelated to D_i or have a very small impact not contributing to the efficient variance of the IV estimator in 5.1.1. Hence, the post-Lasso IV estimator $\hat{\alpha}(\hat{\mathcal{A}})$ can achieve the same level of efficiency with a probability approaching one.

5.2. Many potential IVs and few controls

In a typical situation, the second-stage equation also includes a number of exogenous covariates/controls. Thus, in practice we often have to consider the following model:

$$Y_i = \alpha D_i + X_i'\beta + U_i, \tag{5.2.1}$$
$$E(U_i \mid X_i) = 0.$$

For example, the intercept is typically one of the elements of X_i. The exogenous controls with non-zero β's should be included in the first stage as they are typically also correlated with the endogenous regressor D_i.

$$D_i = Z_i'\pi + X_i'\gamma + V_i, \tag{5.2.2}$$
$$E(V_i \mid Z_i, X_i) = 0.$$

Omitting relevant controls, i.e. controls that related to Y_i and D_i, from the first stage in IV estimation can result in inconsistent estimates, see Appendix 5.5.

When the number of controls is small, Chernozhukov et al. (2016b) propose the following algorithm.

Step 1 Estimate the first stage equation in (5.2.2) using Lasso. Force inclusion of X_i's by assigning zero penalty weights to their coefficients. Let $\hat{\mathcal{A}}$ denote the set of selected IVs:

$$\hat{\mathcal{A}} = \{j \in \{1, \ldots, l\} : \hat{\pi}_{j,\lambda} \neq 0\},$$

where $\hat{\pi}_\lambda = (\hat{\pi}_{1,\lambda}, \ldots, \hat{\pi}_{l,\lambda})'$ is the vector of the corresponding Lasso-estimated coefficients.

Step 2 Estimate the first stage by OLS (post-Lasso) using only the instruments in $\hat{\mathcal{A}}$ and the controls X_i. Construct

$$\hat{\zeta}_i(\hat{\mathcal{A}}) = Z_{i,\hat{\mathcal{A}}}'\hat{\pi}_{\hat{\mathcal{A}}} + X_i'\hat{\gamma}(\hat{\mathcal{A}}),$$

where $Z_{i,\hat{\mathcal{A}}}$ is the sub-vector of the instruments selected in Step 1, and $\hat{\pi}_{\hat{\mathcal{A}}}$ and $\hat{\gamma}(\hat{\mathcal{A}})$ are the post-Lasso OLS estimates for the first stage equation.

Step 3 Estimate (5.1.1) using $\hat{\zeta}_i(\hat{\mathcal{A}})$ as the IV for D_i:

$$\begin{pmatrix} \hat{\alpha}(\hat{\mathcal{A}}) \\ \hat{\beta}(\hat{\mathcal{A}}) \end{pmatrix} = \left(\sum_{i=1}^{n} \begin{pmatrix} \hat{\zeta}_i(\hat{\mathcal{A}}) \\ X_i \end{pmatrix} \begin{pmatrix} D_i \\ X_i \end{pmatrix}' \right)^{-1} \sum_{i=1}^{n} \begin{pmatrix} \hat{\zeta}_i(\hat{\mathcal{A}}) \\ X_i \end{pmatrix} Y_i.$$

We can also provide a more convenient expression for the post-Lasso IV estimator $\hat{\alpha}(\hat{\mathcal{A}})$. Write the second- and first-stage equations in the matrix form:

$$Y = \alpha D + X\beta + U,$$
$$D = Z\pi + X\gamma + V.$$

Let M_X be the projection matrix on the space orthogonal to the span of X:

$$M_X = I_n - X(X'X)^{-1}X'.$$

Since

$$M_X X = 0,$$

we have

$$M_X Y = \alpha M_X D + M_X U,$$
$$M_X D = M_X Z\pi + M_X V.$$

Recall that

$$\tilde{Y} = M_X Y,$$
$$\tilde{D} = M_X D,$$
$$\tilde{Z} = M_X Z$$

are the residuals from the OLS regressions against X of Y, D, and Z respectively. Using the partialling out arguments, the post-Lasso IV estimator $\hat{\alpha}(\hat{\mathcal{A}})$ can be computed as follows.

Step 1 Estimate the first stage equation $\tilde{D} = \tilde{Z}\pi + \tilde{V}$ using Lasso. Let $\hat{\mathcal{A}}$ denote the set of selected IVs:

$$\hat{\mathcal{A}} = \{j \in \{1, \ldots, l\} : \hat{\pi}_{j,\lambda} \neq 0\},$$

where $\hat{\pi}_\lambda = (\hat{\pi}_{1,\lambda}, \ldots, \hat{\pi}_{l,\lambda})'$ is the vector of the corresponding Lasso-estimated coefficients.

Step 2 Estimate the first stage by OLS (post-Lasso) using only the instruments in $\hat{\mathcal{A}}$ and the controls X_i. Construct

$$\tilde{\zeta}_i(\hat{\mathcal{A}}) = \tilde{Z}'_{i,\hat{\mathcal{A}}} \hat{\pi}_{\hat{\mathcal{A}}},$$

where $\tilde{Z}_{i,\hat{\mathcal{A}}}$ is the sub-vector of the instruments selected in Step 1, and $\hat{\pi}_{\hat{\mathcal{A}}}$ is the post-Lasso OLS estimator for the first stage equation:

$$\hat{\pi}_{\hat{\mathcal{A}}} = \left(\tilde{Z}'_{\hat{\mathcal{A}}} \tilde{Z}_{\hat{\mathcal{A}}} \right)^{-1} \tilde{Z}'_{\hat{\mathcal{A}}} \tilde{D}.$$

Step 3 Estimate $\tilde{Y} = \alpha\tilde{D} + \tilde{U}_i$ using $\tilde{\zeta}_i(\hat{\mathcal{A}})$ as the IV for D_i:

$$\hat{\alpha}(\hat{\mathcal{A}}) = \frac{\sum_{i=1}^n \tilde{\zeta}_i(\hat{\mathcal{A}})\tilde{Y}_i}{\sum_{i=1}^n \tilde{\zeta}_i(\hat{\mathcal{A}})\tilde{D}_i}.$$

The two algorithms described in this section produce identical estimates of α by the partialling out argument.

5.3. Few IVs and many controls

Consider again the IV regression model

$$Y_i = \alpha D_i + X_i'\beta + U_i,$$
$$D_i = Z_i'\pi + X_i'\gamma + V_i.$$

We assume now that the number of IVs l is small, and all l IVs are used in estimation. On the other hand, the number of potential controls k is large, however, the model is sparse: the set of relevant controls

$$\mathcal{A}_0 = \{j \in \{1, \ldots, k\} : \beta_j \neq 0\}$$

is small. We now need to select the relevant controls from X_i.

The procedure can again be based on the partialling out arguments. Chernozhukov et al. (2016b) describe the following algorithm.

Step 1 Estimate by Lasso $D_i = \sum_{j=1}^k \eta_j^D X_{i,j} + \epsilon_i^D$. Let $\hat{\mathcal{A}}_D$ denote the set of selected controls:

$$\hat{\mathcal{A}}_D = \left\{j \in \{1, \ldots, k\} : \hat{\eta}_{j,\lambda}^D \neq 0\right\},$$

where $\hat{\eta}_{1,\lambda}^D, \ldots, \hat{\eta}_{k,\lambda}^D$ are the corresponding Lasso-estimated coefficients. Post-Lasso: Regress D_i on the controls in $\hat{\mathcal{A}}_D$, and save the residuals as $\tilde{D}_i^{PL}$.

Step 2 Estimate by Lasso $Y_i = \sum_{j=1}^k \eta_j^Y X_{i,j} + \epsilon_i^Y$. Let $\hat{\mathcal{A}}_Y$ denote the set of selected controls:

$$\hat{\mathcal{A}}_Y = \left\{j \in \{1, \ldots, k\} : \hat{\eta}_{j,\lambda}^Y \neq 0\right\},$$

where $\hat{\eta}_{1,\lambda}^Y, \ldots, \hat{\eta}_{k,\lambda}^Y$ are the corresponding Lasso-estimated coefficients. Post-Lasso: Regress Y_i on the controls in $\hat{\mathcal{A}}_Y$, and save the residuals as $\tilde{Y}_i^{PL}$.

Step 3 For $m = 1, \ldots, l$, estimate by Lasso $Z_{i,m} = \sum_{j=1}^k \eta_j^{Z_m} X_{i,j} + \epsilon_i^{Z_m}$. Let $\hat{\mathcal{A}}_{Z_m}$ denote the set of selected controls:

$$\hat{\mathcal{A}}_{Z_m} = \left\{j \in \{1, \ldots, k\} : \hat{\eta}_{j,\lambda}^{Z_m} \neq 0\right\},$$

where $\hat{\eta}_{1,\lambda}^{Z_m}, \ldots, \hat{\eta}_{k,\lambda}^{Z_m}$ are the corresponding Lasso-estimated coefficients. Post-Lasso: Regress $Z_{i,m}$ on the controls in $\hat{\mathcal{A}}_{Z_m}$, and save the residuals as $\tilde{Z}_{i,m}^{PL}$. Note that the procedure must be repeated for all $m = 1, \ldots, l$ IVs.

Step 4 Construct the IV:

$$\tilde{\zeta}_i = \sum_{m=1}^l \hat{\pi}_m \tilde{Z}_{i,m}^{PL},$$

where $\hat{\pi}_1, \ldots, \hat{\pi}_l$ are the OLS estimates of $\pi_1, \ldots, \pi_l$ from the regression of $\tilde{D}_i^{PL}$ against $\tilde{Z}_{i,1}^{PL}, \ldots, \tilde{Z}_{i,l}^{PL}$.

Step 5 Use $\tilde{\zeta}_i$ as the IV for $\tilde{D}_i^{PL}$:

$$\hat{\alpha}\left(\hat{\mathcal{A}}\right) = \frac{\sum_{i=1}^n \tilde{\zeta}_i \tilde{Y}_i^{PL}}{\sum_{i=1}^n \tilde{\zeta}_i \tilde{D}_i^{PL}}.$$

Here $\hat{\mathcal{A}} = \hat{\mathcal{A}}_D \cup \hat{\mathcal{A}}_Y \cup \hat{\mathcal{A}}_{Z_1} \cup \ldots \cup \hat{\mathcal{A}}_{Z_l}$ and includes all controls that are useful for predicting D_i, Y_i, or one of the IVs $Z_{i,m}$, $m = 1, \ldots, l$.

The algorithm is similar to the partialling out algorithm in Chapter 4. The IV estimator $\hat{\alpha}\left(\hat{\mathcal{A}}\right)$ solves the following equation:

$$\sum_{i=1}^n \tilde{\zeta}_i \left(\tilde{Y}_i^{PL} - \hat{\alpha}(\hat{\mathcal{A}})\tilde{D}_i^{PL}\right) = 0.$$

Suppose a control $X_{i,j}$ for some $j = 1, \ldots, k$ has been dropped by Lasso in Step 2 of the algorithm. With a probability approaching one, the coefficient β_j is small. Hence, omitting $X_{i,j}$ would not introduce a bias unless it is strongly related to $\tilde{\zeta}_i$ through one of the IVs or D_i. However, since the effect of X_i's has been partialled out from Z_i's and D_i, with a probability approaching one there will be no significant correlation between $\tilde{\zeta}_i$ and $X_{i,j}$.

5.4. Many IVs and many controls

The approach can be extended to the case when there are many potential IVs and controls. When there are many IVs, it is impractical to partial out the effect of many controls from the IVs. Instead, we can partial out the effect of controls from the efficient post-Lasso-based IV. The following algorithm is proposed in Chernozhukov et al. (2015).

Step 1 Use Lasso and post-Lasso to partial out the effects of the controls X_i's from Y_i. Save the residuals as $\tilde{Y}_i^{PL}$.

Step 2 Use Lasso and post-Lasso to *predict* D_i in the first-stage regression $D_i = Z_i'\pi + X_i'\gamma + V_i$. Save the *predicted* value $\hat{D}_i(\hat{\mathcal{A}})$:

$$\hat{\zeta}_i(\hat{\mathcal{A}}) = Z_i'\hat{\pi}_{\hat{\mathcal{A}}} + X_i'\hat{\gamma}_{\hat{\mathcal{A}}},$$

where $\hat{\mathcal{A}}$ is the set of Lasso selected IVs *and* controls. The coefficients $\hat{\pi}_{\hat{\mathcal{A}}}$ and $\hat{\gamma}_{\hat{\mathcal{A}}}$ are from the post-Lasso OLS regression of D_i against the IVs and controls in $\hat{\gamma}_{\hat{\mathcal{A}}}$.

Step 3 Use Lasso and post-Lasso to partial out the effect of X_i's on $\hat{\zeta}_i(\hat{\mathcal{A}})$. Save the residuals as $\tilde{\zeta}_i^{PL}$.

Step 4 Use $\tilde{\zeta}_i$ as the IV for $\tilde{D}_i^{PL}$:

$$\hat{\alpha} = \frac{\sum_{i=1}^n \tilde{\zeta}_i^{PL} \tilde{Y}_i^{PL}}{\sum_{i=1}^n \tilde{\zeta}_i^{PL} \tilde{D}_i^{PL}}.$$

Note that in Step 3, partialling out will remove $X_i'\hat{\gamma}_{\hat{\mathcal{A}}}$ from $\hat{\zeta}_i(\hat{\mathcal{A}})$ constructed in Step 2. However, we keep X_i in the first-stage equation in Step 2 to obtain consistent estimates of π's.

5.5. Appendix IV estimation and second-stage controls

Consider the IV regression model

$$Y_i = \alpha D_i + X_i'\beta + U_i, \tag{5.5.1}$$

$$D_i = Z_i'\pi + X_i'\gamma + V_i, \tag{5.5.2}$$

$$E(U_i \mid X_i, Z_i) = 0,$$

$$E(V_i \mid X_i, Z_i) = 0.$$

Substituting (5.5.2) into (5.5.1), we obtain

$$Y_i = \alpha\left(Z_i'\pi\right) + X_i'(\beta + \alpha\gamma) + (U_i + \alpha V_i).$$

Since Z_i and X_i are uncorrelated with (U_i, V_i), the OLS regression of Y_i against $(Z_i'\pi)$ and X_i would produce a consistent estimator of α. This regression is infeasible and in practice, π is replaced with its first-stage OLS estimator. However, the result is asymptotically equivalent to that of the infeasible regression.

Suppose that the econometrician omits X_i from the first stage. The resulting first-stage regression is now

$$D_i = Z_i'\pi^* + V_i^*, \tag{5.5.3}$$

where π^* is the coefficient in the population regression of D_i against Z_i only:

$$\begin{aligned}
\pi^* &= \left(EZ_iZ_i'\right)^{-1} EZ_iD_i \\
&= \left(EZ_iZ_i'\right)^{-1} E\left(Z_i(Z_i'\pi + X_i'\gamma + V_i)\right) \\
&= \pi + \left(EZ_iZ_i'\right)^{-1} EZ_iX_i'\gamma \\
&= \pi + \theta\gamma,
\end{aligned}$$

where

$$\theta = \left(EZ_iZ_i'\right)^{-1} EZ_iX_i'$$

is the coefficient in from the population regression of the controls X_i' against the instruments Z_i. The residual V_i^* is given by

$$\begin{aligned}
V_i^* &= D_i - Z_i'\pi^* \\
&= D_i - Z_i'(\pi + \theta\gamma) \\
&= \left(X_i' - Z_i'\theta\right)\gamma + V_i \\
&= \tilde{X}_i'\gamma + V_i,
\end{aligned}$$

where $\tilde{X}_i'$ is the residual in the population the regression of X_i' against Z_i:

$$\tilde{X}_i' = X_i' - Z_i'\theta.$$

Hence, by construction:

$$EZ_i\tilde{X}_i' = 0.$$

Equations (5.5.1) and (5.5.3) imply that

$$Y_i = \alpha(Z_i'\pi^*) + X_i'\beta + \left(U_i + \alpha\tilde{X}_i'\gamma\right). \tag{5.5.4}$$

While Z_i is uncorrelated with $\tilde{X}_i'\gamma$ by construction, the controls X_i are correlated with $\tilde{X}_i'\gamma$. Consequently, OLS estimation of (5.5.4) would produce inconsistent estimates not only for β, but also for α if X_i and Z_i are correlated.

Note that if Z_i and X_i are uncorrelated, $\theta = 0$ and $\tilde{X}_i' = X_i$. In this case, $\alpha\tilde{X}_i'\gamma$ term in (5.5.4) would be replaced by $\alpha X_i'\gamma$ and can be combined with $X_i'\beta$ as before. Hence, it is important to include the second-stage controls X_i into the first stage unless they are uncorrelated with the instruments Z_i or $\gamma = 0$.

Bibliography

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics 106, 979–1014.

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

Belloni, A., Chernozhukov, V., 2011. High dimensional sparse econometric models: An introduction, in: Inverse Problems and High-Dimensional Estimation. Springer, pp. 121–156.

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

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. American Economic Review: Papers and Proceedings 105, 486–90.

Chernozhukov, V., Hansen, C., Spindler, M., 2016a. hdm: High-dimensional metrics. arXiv preprint arXiv:1608.00354 .

Chernozhukov, V., Hansen, C., Spindler, M., 2016b. High-dimensional metrics in R. arXiv preprint arXiv:1603.01700 .

Wainwright, M.J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory 55, 2183–2202.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541–2563.