Econ 480-3
Introduction to Econometrics

SPRING 2022
Ver. May 23, 2022

Northwestern University

Lecture Notes by

IVAN A. CANAY

Department of Economics
Northwestern University

Contents

Part I

A Primer on Linear Models

Lecture 1

Linear Regression[1]

1.1 Interpretations of the Linear Regression Model

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

The parameter β_0 is sometimes referred to as the *intercept parameter* and the remaining β_j parameters are sometimes referred to as the *slope parameters*. There are several ways to interpret β depending on the assumptions imposed on (Y, X, U). We will study three such ways.

1.1.1 Interpretation 1: Linear Conditional Expectation

Suppose $E[Y|X] = X'\beta$ and define $U = Y - E[Y|X]$. (Note that we've implicitly assumed $E[|Y|] < \infty$, so $E[Y|X]$ exists.) This implies that $E[U|X] = 0$ and therefore that $E[U] = 0$. Moreover, $E[XU] = 0$, so $\mathrm{Cov}[X, U] = 0$. In this case, β is just a convenient way of summarizing a feature of the joint distribution of Y and X, namely, the conditional expectation. It is tempting to interpret the coefficient β_j for $1 \leq j \leq k$ as the *ceteris paribus* (i.e., holding X_{-j} and U constant) effect of a one unit change in X_j on Y, but this is incorrect. Indeed, more generally, it is not appropriate to think of differences in (or derivatives of) conditional expectations causally. After all, Y could be an indicator for rain and X could be an indicator for carrying an umbrella. In this case, it may be the case that $E[Y|X]$ is increasing in X, but one would not want to think of carrying an umbrella as causing rain. What is missing is a model of how Y is determined as a function of X (and possibly other unobserved variables).

[1] This lecture is based on Azeem Shaikh's lecture notes. I want to thank him for kindly sharing them.

1.1.2 Interpretation 2: "Best" Linear Approximation to the Conditional Expectation or "Best" Linear Predictor

In general, one would not expect the conditional expectation to be linear. Suppose $E[Y^2] < \infty$ and $E[XX'] < \infty$ (equivalently, that $E[X_j^2] < \infty$ for $1 \leq j \leq k$). Under these assumptions, one may consider what is the "best" linear approximation (i.e., function of the form $X'b$ for some choice of $b \in \mathbf{R}^{k+1}$) to the conditional expectation. To this end, consider the minimization problem

$$\min_{b \in \mathbf{R}^{k+1}} E[(E[Y|X] - X'b)^2] \ .$$

Denote by β a solution to this minimization problem. In this case, β is simply a convenient way of summarizing another feature of the joint distribution of Y and X, namely, the "best" linear approximation to the conditional expectation. For the same reasons as before, it is not correct to interpret the coefficient β_j for $1 \leq j \leq k$ as the *ceteris paribus* effect of a one unit change in X_j on Y.

Let $V = E[Y|X] - Y$, so $E[XV] = 0$. Note that

$$\begin{aligned} E[(E[Y|X] - X'b)^2] &= E[(E[Y|X] - Y + Y - X'b)^2] \\ &= E[(V + Y - X'b)^2] \\ &= E[V^2 + 2V(Y - X'b) + (Y - X'b)^2] \\ &= E[V^2] + 2E[VY] - 2E[VX']b + E[(Y - X'b)^2] \\ &= \text{constant} + E[(Y - X'b)^2] \ . \end{aligned}$$

Thus, β also solves

$$\min_{b \in \mathbf{R}^{k+1}} E[(Y - X'b)^2] \ .$$

In this sense, β is also a convenient way of summarizing the "best" linear predictor of Y given X. Again, it is tempting to interpret β_j for $1 \leq j \leq k$ causally, but this is not correct.

Consider the second minimization problem. Note $E[(Y - X'b)^2]$ is convex (as a function of b) and

$$D_b E[(Y - X'b)^2] = E[-2X(Y - X'b)] \ .$$

Hence, β must satisfy

$$E[X(Y - X'\beta)] = 0 \ .$$

If we define $U = Y - X'\beta$, then we may rewrite this equation as

$$E[XU] = 0 \ .$$

1.1.3 Interpretation 3: Causal Model

Suppose $Y = g(X, U)$, where X are the observed determinants of Y and U are the unobserved determinants of Y. Such a relationship is a model of how Y is determined and may come from physics, economics, etc. The effect of X_j on Y holding X_{-j} and U constant (i.e., *ceteris paribus*) is determined by g. If g is differentiable, then it is given by $D_{X_j} g(X, U)$. If we assume further that

$$g(X, U) = X'\beta + U \ ,$$

then the *ceteris paribus* effect of X_j on Y is simply β_j. We may normalize U so that $E[U] = 0$ (by replacing U with $U - E[U]$ and β_0 with $\beta_0 + E[U]$ if this is not the case). On the other hand, $E[U|X]$, $E[U|X_j]$ and $E[UX_j]$ for $1 \leq j \leq k$ may or may not equal zero. These are now statements about the relationship between the observed and unobserved determinants of Y.

Probably the easiest way to think about causal relationships is in terms of potential outcomes. As a simple illustration, consider a randomized controlled experiment where individuals are randomly assigned to a treatment (a drug) that is intended to improve their health status. Let Y denote the observed health status and $X \in \{0, 1\}$ denote whether the individual takes the drug or not. The causal relationship between X and Y can be described using the so-called *potential outcomes*:

$Y(0)$	potential outcome in the absence of treatment
$Y(1)$	potential outcome in the presence of treatment

.

In other words, we imagine two potential health status variables $(Y(0), Y(1))$ where $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) X were 0; and $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) X were 1.

The difference $Y(1) - Y(0)$ is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*. Using this notation, we may rewrite the observed outcome as

$$\begin{aligned} Y &= XY(1) + (1 - X)Y(0) \\ &= E[Y(0)] + (Y(1) - Y(0))X + (Y(0) - E[Y(0)]) \\ &= \beta_0 + \beta_1 X + U \ , \end{aligned}$$

where

$$\begin{aligned} \beta_0 &= E[Y(0)] \\ \beta_1 &= Y(1) - Y(0) \\ U &= Y(0) - E[Y(0)] \ . \end{aligned}$$

In order for β_1 to be a constant parameter, we need to assume that $(Y(1) - Y(0)$ is constant across individuals. Under all these assumptions, we end

up with a linear constant effect causal model with $U \perp\!\!\!\perp X$ (from the nature of the randomized experiment), $E[U] = 0$, and so $E[XU] = 0$. Notice that, in order to have a linear causal model a randomized controlled experiment is not enough; we also need a constant treatment effect. Without such an assumption it can be shown that a regression of Y on X identifies the average treatment effect (ATE). The ATE is often interpreted as a causal parameter because it is an *average* of causal effects.

1.2 Linear Regression when $E[XU] = 0$

1.2.1 Solving for β

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[XU] = 0$, $E[XX'] < \infty$, and that there is *no perfect collinearity in* X. The justification of the first assumption varies depending on which of the three preceding interpretations we invoke. The second assumption ensures that $E[XX']$ exists. The third assumption is equivalent to the assumption that the matrix $E[XX']$ is in fact invertible. Since $E[XX']$ is positive semi-definite, invertibility of $E[XX']$ is equivalent to $E[XX']$ being positive definite. We say that there is *perfect collinearity* or *multicollinearity* in X if there exists nonzero $c \in \mathbf{R}^{k+1}$ such that $P\{c'X = 0\} = 1$, i.e., if we can express one component of X as a linear combination of the others.

Lemma 1.1 *Let X be a random vector such that $E[XX'] < \infty$. Then $E[XX']$ is invertible if and only if there is no perfect collinearity in X.*

PROOF: We first argue that if $E[XX']$ is invertible, then there is no perfect collinearity in X. To see this, suppose there is perfect collinearity in X, i.e., that there exists a nonzero $c \in \mathbf{R}^{k+1}$ such that $P\{c'X = 0\} = 1$. Note that $E[XX']c = E[X(X'c)] = 0$. Hence, the columns of $E[XX']$ are linearly dependent, i.e., $E[XX']$ is not invertible.

We now argue that if there is no perfect collinearity in X, then $E[XX']$ is invertible. To see this, suppose $E[XX']$ is not invertible. Then, the columns of $E[XX']$ must be linearly dependent, i.e., there exists nonzero $c \in \mathbf{R}^{k+1}$ such that $E[XX']c = 0$. This implies further that $c'E[XX']c = E[(c'X)^2] = 0$, which in turn implies that $P\{c'X = 0\} = 1.$, i.e., that there is perfect collinearity in X. ■

The first assumption above together with the fact that $U = Y - X'\beta$ implies that $E[X(Y - X'\beta)] = 0$, i.e., $E[XY] = E[XX']\beta$. Since $E[XX']$ is

invertible, we have that there is a unique solution to this system of equations, namely,

$$\beta = E[XX']^{-1}E[XY] \ .$$

If $E[XX']$ is not invertible, i.e., there is perfect collinearity in X, then there will be more than one solution to this system of equations. Importantly, any two solutions β and $\tilde{\beta}$ will necessarily satisfy $P\{X'\beta = X'\tilde{\beta}\} = 1$. Depending on the interpretation, this may be an important distinction or not. For instance, in the second interpretation, each such solution corresponds to the same "best" linear predictor of Y given X, whereas in the third interpretation different values of β could have wildly different implications for how X affects Y holding U constant.

1.3 Estimating β

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, $E[XX'] < \infty$ and that there is no perfect collinearity in X. Above we described three different interpretations and justifications of such a model. We now discuss estimation of β.

1.3.1 Ordinary Least Squares

Let (Y, X, U) be distributed as described above and denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sequence of random vectors with distribution P. By analogy with the expression we derived for β under these assumptions, the natural estimator of β is simply

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \le i \le n} X_i Y_i \right) \ .$$

This estimator is called the *ordinary least squares* (OLS) estimator of β because it can also be derived as the solution to the following minimization problem:

$$\min_{b \in \mathbf{R}^{k+1}} \frac{1}{n} \sum_{1 \le i \le n} (Y_i - X_i'b)^2 \ .$$

To see this, note that $\frac{1}{n}\sum_{1 \le i \le n}(Y_i - X_i'b)^2$ is convex (as a function of b) and

$$D_b \frac{1}{n} \sum_{1 \le i \le n} (Y_i - X_i'b)^2 = \frac{1}{n} \sum_{1 \le i \le n} -2X_i(Y_i - X_i'b) \ .$$

Hence $\hat{\beta}_n$ must satisfy

$$\frac{1}{n}\sum_{1\le i\le n} X_i(Y_i - X_i'\hat{\beta}_n) = 0 \ ,$$

i.e.,

$$\frac{1}{n}\sum_{1\le i\le n} X_iY_i = \frac{1}{n}\sum_{1\le i\le n} X_iX_i'\hat{\beta}_n \ .$$

The matrix

$$\frac{1}{n}\sum_{1\le i\le n} X_iX_i'$$

may not be invertible, but, since $E[XX']$ is invertible, it will be invertible with probability approaching one.

The ith *fitted value* is denoted by $\hat{Y}_i = X_i'\hat{\beta}_n$. The ith *residual* is denoted by $\hat{U}_i = Y_i - \hat{Y}_i$. By definition, we therefore have that

$$\frac{1}{n}\sum_{1\le i\le n} X_i\hat{U}_i = 0 \ .$$

1.3.2 Projection Interpretation

Define

$$\begin{aligned}
\mathbb{Y} &= (Y_1, \ldots, Y_n)' \\
\mathbb{X} &= (X_1, \ldots, X_n)' \\
\hat{\mathbb{Y}} &= (\hat{Y}_1, \ldots, \hat{Y}_n)' \\
&= \mathbb{X}\hat{\beta}_n \\
\mathbb{U} &= (U_1, \ldots, U_n)' \\
\hat{\mathbb{U}} &= (\hat{U}_1, \ldots, \hat{U}_n)' \\
&= \mathbb{Y} - \hat{\mathbb{Y}} \\
&= \mathbb{Y} - \mathbb{X}\hat{\beta}_n \ .
\end{aligned}$$

In this notation,

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

and may be equivalently described as the solution to

$$\min_{b\in\mathbf{R}^{k+1}} |\mathbb{Y} - \mathbb{X}b|^2 \ .$$

Hence, $\mathbb{X}\hat{\beta}_n$ is the vector in the column space of $\mathbb{X}$ that is closest (in terms of Euclidean distance) to $\mathbb{Y}$. From the above, we see that $\mathbb{X}'\hat{\mathbb{U}} = 0$, thus $\hat{\mathbb{U}}$ is orthogonal to all of the columns of $\mathbb{X}$ (and thus orthogonal to all of the vectors in the column space of $\mathbb{X}$). In this sense,

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

is the *orthogonal projection* of $\mathbb{Y}$ onto the ($(k+1)$-dimensional) column space of $\mathbb{X}$. The matrix

$$\mathbb{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

is known as a *projection* matrix. It projects a vector in $\mathbf{R}^n$ (such as $\mathbb{Y}$) onto the column space of $\mathbb{X}$. Note that $\mathbb{P}^2 = \mathbb{P}$, which reflects the fact that projecting something that already lies in the column space of $\mathbb{X}$ onto the column space of $\mathbb{X}$ does nothing. The matrix $\mathbb{P}$ is also symmetric. The matrix

$$\mathbb{M} = \mathbb{I} - \mathbb{P}$$

is also a projection matrix. It projects a vector onto the ($(n-k-1)$-dimensional) vector space orthogonal to the column space of $\mathbb{X}$. Hence, $\mathbb{M}\mathbb{X} = 0$. Note that $\mathbb{M}\mathbb{Y} = \hat{\mathbb{U}}$. For this reason, $\mathbb{M}$ is sometimes called the "residual maker" matrix.

Bibliography

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Lecture 2

More on Linear Regression[1]

2.1 Solving for Sub-vectors of β

Partition X into X_1 and X_2, where X_1 takes values in $\mathbf{R}^{k_1}$ and X_2 takes values in $\mathbf{R}^{k_2}$. Partition β into β_1 and β_2 analogously. In this notation,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U \ .$$

Our preceding results imply that

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} E[X_1X_1'] & E[X_1X_2'] \\ E[X_2X_1'] & E[X_2X_2'] \end{pmatrix}^{-1} \begin{pmatrix} E[X_1Y] \\ E[X_2Y] \end{pmatrix} \ .$$

Using the so called partitioned matrix inverse formula, it would be possible to derive formulae for β_1 and β_2, but such an exercise is not particularly illuminating. We therefore take a different approach to arrive at the same formulae. In doing so, we will make use of the following notation: for a random variable A and a random vector B, denote by $\text{BLP}(A|B)$ the best linear predictor of A given B, i.e. $B'E[BB']^{-1}E[BA]$. If A is a random vector, then define $\text{BLP}(A|B)$ component-wise.

Define $\tilde{Y} = Y - \text{BLP}(Y|X_2)$ and $\tilde{X}_1 = X_1 - \text{BLP}(X_1|X_2)$. Consider the linear regression $\tilde{Y} = \tilde{X}_1'\tilde{\beta}_1 + \tilde{U}$, where $E[\tilde{X}_1\tilde{U}] = 0$ (as, for example, in the second interpretation of the linear regression model described before). It follows that $\tilde{\beta}_1 = \beta_1$. To see this, note that $E[\tilde{X}_1\tilde{X}_1']$ is invertible (because each component of $\tilde{X}_1$ is a linear combination the components of X), so

$$\begin{aligned} \tilde{\beta}_1 &= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1\tilde{Y}] \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1Y] - E[\tilde{X}_1\text{BLP}(Y|X_2)] \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1Y] \ , \end{aligned}$$

[1] This lecture is based on Azeem Shaikh's lecture notes. I want to thank him for kindly sharing them.

where the first equality follows from the formula for $\tilde{\beta}_1$, the second equality follows from the expression for $\tilde{Y}$, and the third equality follows from the fact that $E[\tilde{X}_1 X_2'] = 0$ (because $\tilde{X}_1$ is the error term from a regression of X_1 on X_2). Note that this first part of the derivation shows that $\tilde{\beta}_1$ is also the population coefficient of a linear regression of Y on $\tilde{X}_1$. If we now replace Y by its expression and do some additional steps, we get

$$\begin{aligned}\tilde{\beta}_1 &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1 X_1'\beta_1] + E[\tilde{X}_1 X_2'\beta_2] + E[\tilde{X}_1 U]) \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1 X_1'\beta_1]) \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1\tilde{X}_1'\beta_1] + E[\tilde{X}_1 \text{BLP}(X_1|X_2)'\beta_1]) \\ &= \beta_1 \ ,\end{aligned}$$

where the first equality follows from the expression for Y, the second equality follows from the fact that $E[\tilde{X}_1 X_2'] = 0$ and $E[\tilde{X}_1 U] = 0$ (because $E[XU] = 0$), the third equality follows from the expression for $\tilde{X}_1$, and the final equality follows from the fact that $E[\tilde{X}_1 X_2'] = 0$.

In other words, β_1 in the linear regression of Y on X_1 and X_2 is equal to the coefficient in a linear regression of the error term from a linear regression of Y on X_2 on the error terms from a linear regression of the components of X_1 on X_2. This gives meaning to the common description of β_1 as the "effect" of X_1 on Y after "controlling for X_2."

Notice that if we take X_2 to be just a constant, then $\tilde{Y} = Y - E[Y]$ and $\tilde{X}_1 = X_1 - E[X_1]$. Hence,

$$\begin{aligned}\beta_1 &= E[(X_1 - E[X_1])(X_1 - E[X_1])']^{-1} E[(X_1 - E[X_1])(Y - E[Y])] \\ &= \text{Var}[X_1]^{-1}\text{Cov}[X_1, Y] \ .\end{aligned}$$

Finally, also note that if we use our formula to interpret the coefficient β_j associated with the jth covariate for $1 \leq j \leq k$, we obtain

$$\beta_j = \frac{\text{Cov}[\tilde{X}_j, Y]}{\text{Var}[\tilde{X}_j]} \ , \tag{2.1}$$

which shows that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "partialling out" all the other variables in the model.

2.2 Estimating Sub-Vectors of β

Partition X into X_1 and X_2, where X_1 takes values in $\mathbf{R}^{k_1}$ and X_2 takes values in $\mathbf{R}^{k_2}$. Partition β into β_1 and β_2 analogously. In this notation,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U \ .$$

Using the preceding results, we can derive estimation counterparts to the results above about solving for sub-vectors of β. Again, this may be done using the partitioned matrix inverse formula, but we will use a different approach. Let $\mathbb{X}_1 = (X_{1,1}, \ldots, X_{1,n})'$ and $\mathbb{X}_2 = (X_{2,1}, \ldots, X_{2,n})'$. Denote by $\mathbb{P}_1$ the projection matrix onto the column space of $\mathbb{X}_1$ and $\mathbb{P}_2$ the projection matrix onto the column space of $\mathbb{X}_2$. Define $\mathbb{M}_1 = \mathbb{I} - \mathbb{P}_1$ and $\mathbb{M}_2 = \mathbb{I} - \mathbb{P}_2$.

First note that

$$\mathbb{Y} = \mathbb{X}_1\hat{\beta}_{1,n} + \mathbb{X}_2\hat{\beta}_{2,n} + \hat{\mathbb{U}} \ .$$

This implies that

$$\mathbb{M}_2\mathbb{Y} = \mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n} + \hat{\mathbb{U}}$$

because $\mathbb{M}_2\hat{\mathbb{U}} = \hat{\mathbb{U}}$, as $\hat{\mathbb{U}}$ is orthogonal to the column space of $\mathbb{X}$ (and hence the column space of $\mathbb{X}_2$ as well). This implies further that

$$(\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{Y} = (\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n}$$

because $(\mathbb{M}_2\mathbb{X}_1)'\hat{\mathbb{U}} = \mathbb{X}_1'\hat{\mathbb{U}} = 0$, as $\mathbb{X}'\hat{\mathbb{U}} = 0$. Note that the matrix $(\mathbb{M}_2\mathbb{X}_1)'(\mathbb{M}_2\mathbb{X}_1)$ is invertible provided that $\mathbb{X}'\mathbb{X}$ is invertible. Hence,

$$\hat{\beta}_{1,n} = ((\mathbb{M}_2\mathbb{X}_1)'(\mathbb{M}_2\mathbb{X}_1))^{-1}(\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{Y} \ .$$

In other words, $\hat{\beta}_{1,n}$ can be obtained by estimating via OLS the coefficients from a linear regression of $\mathbb{M}_2\mathbb{Y}$ on $\mathbb{M}_2\mathbb{X}_1$. Upon recognizing that $\mathbb{M}_2\mathbb{Y}$ are the residuals from a regression of $\mathbb{Y}$ on $\mathbb{X}_2$ and that the columns of $\mathbb{M}_2\mathbb{X}_1$ are the residuals from regressions of the columns of $\mathbb{X}_1$ on $\mathbb{X}_2$, we see that this formula exactly parallels the formula we derived earlier for a sub-vector of β. This result is sometimes referred to as the *Frisch-Waugh-Lovell* (FWL) decomposition.

2.3 Properties of LS

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, $E[XX'] < \infty$, and that there is no perfect collinearity in X. Denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sample of random vectors with distribution P. Above we described estimation of β via OLS under these assumptions. We now discuss properties of the resulting estimator, $\hat{\beta}_n$ imposing stronger assumptions as needed.

2.3.1 Bias

Suppose in addition that $E[U|X] = 0$. Equivalently, assume that $E[Y|X] = X'\beta$. Under this stronger assumption,

$$E[\hat{\beta}_n] = \beta \ .$$

In fact,

$$E[\hat{\beta}_n|X_1, \ldots, X_n] = \beta \ .$$

To see this, note that

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

Hence,

$$E[\hat{\beta}_n|X_1, \ldots, X_n] = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}E[\mathbb{U}|X_1, \ldots, X_n] \ .$$

Note for any $1 \leq i \leq n$ that

$$E[U_i|X_1, \ldots, X_n] = E[U_i|X_i] = 0 \ ,$$

where the first equality follows from the fact that X_j is independent of U_i for $i \neq j$. The desired conclusion thus follows.

2.3.2 Gauss-Markov Theorem

Suppose $E[U|X] = 0$ and that $\mathrm{Var}[U|X] = \sigma^2$. When $\mathrm{Var}[U|X]$ is constant (and therefore does not depend on X) we say that U is *homoskedastic*. Otherwise, we say that U is *heteroskedastic*. The *Guass-Markov Theorem* says that under these assumptions the OLS estimator is "best" in the sense that it has the "smallest" value of $\mathrm{Var}[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n]$ among all estimators of the form

$$\mathbb{A}'\mathbb{Y}$$

for some matrix $\mathbb{A} = \mathbb{A}(X_1, \ldots, X_n)$ satisfying

$$E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \beta \ .$$

Here, "smallest" is understood to be in terms of the partial order obtained by defining $B \geq \tilde{B}$ if and only if $B - \tilde{B}$ is positive semi-definite. This class of estimators, of course, includes the OLS estimator as a special case (by setting $\mathbb{A}' = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$). This property is sometimes expressed as saying that OLS is the "best linear unbiased estimator (BLUE)" of β under these assumptions.

To establish this property of OLS, first note that

$$E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \mathbb{A}'\mathbb{X}\beta + \mathbb{A}'E[\mathbb{U}|X_1, \ldots, X_n] \ ,$$

so $E[\mathbb{A}'\mathbb{Y}|X_1,\ldots,X_n] = \beta$ if and only if $\mathbb{A}'\mathbb{X} = \mathbb{I}$. Next, note that

$$\begin{aligned}\mathrm{Var}[\mathbb{A}'\mathbb{Y}|X_1,\ldots,X_n] &= \mathbb{A}'\mathrm{Var}[\mathbb{Y}|X_1,\ldots,X_n]\mathbb{A} \\ &= \mathbb{A}'\mathrm{Var}[\mathbb{U}|X_1,\ldots,X_n]\mathbb{A} \\ &= \mathbb{A}'\mathbb{A}\sigma^2 \ .\end{aligned}$$

When $\mathbb{A}' = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, this last expression is simply $(\mathbb{X}'\mathbb{X})^{-1}\sigma^2$. It therefore suffices to show that

$$\mathbb{A}'\mathbb{A} - (\mathbb{X}'\mathbb{X})^{-1}$$

is positive semi-definite for all matrices $\mathbb{A}$ satisfying $\mathbb{A}'\mathbb{X} = \mathbb{I}$. To this end, define

$$\mathbb{C} = \mathbb{A} - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} \ .$$

Then,

$$\begin{aligned}\mathbb{A}'\mathbb{A} - (\mathbb{X}'\mathbb{X})^{-1} &= (\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1})'(\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}) - (\mathbb{X}'\mathbb{X})^{-1} \\ &= \mathbb{C}'\mathbb{C} + \mathbb{C}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{C} \\ &= \mathbb{C}'\mathbb{C} \ ,\end{aligned}$$

where the last equality follows from the fact that

$$\mathbb{X}'\mathbb{C} = \mathbb{X}'\mathbb{A} - \mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \mathbb{I} - \mathbb{I} = 0 \ .$$

The desired conclusion thus follows from the fact that $\mathbb{C}'\mathbb{C}$ is positive semi-definite by construction.

2.3.3 Consistency

In this case we do not need additional assumptions. Note that $E[XY] < \infty$ since $XY = XX'\beta + XU$, and both $E[XX']$ and $E[XU]$ exist. Under this assumption, the OLS estimator, $\hat{\beta}_n$ is consistent for β, i.e., $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. To see this, simply note that by the WLLN

$$\begin{aligned}\frac{1}{n}\sum_{1\le i\le n} X_iX_i' &\xrightarrow{P} E[XX'] \\ \frac{1}{n}\sum_{1\le i\le n} X_iY_i &\xrightarrow{P} E[XY]\end{aligned}$$

as $n \to \infty$. The desired result therefore follows from the CMT.

2.3.4 Limiting Distribution

Suppose $E[XX'] < \infty$ and that $\mathrm{Var}[XU] = E[XX'U^2] < \infty$. Then,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

as $n \to \infty$, where

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} .$$

To see this, note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} X_i U_i \right) .$$

The WLLN implies that

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \xrightarrow{P} E[XX'] \tag{2.2}$$

as $n \to \infty$. The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} X_i U_i \xrightarrow{d} N(0, \mathrm{Var}[XU])$$

as $n \to \infty$. Thus, the desired result follows from the CMT.

2.4 Estimation of $\mathbb{V}$

In order to make use of the preceding estimators, we will require a consistent estimator of

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} .$$

Note that $\mathbb{V}$ has the so-called sandwich form. As with most sandwich estimators, the interesting object is the "meat" and not the "bread". Indeed, the bread can be consistently estimated by (2.2).

Focusing our attention to the meat, we first consider the case where $E[U|X] = 0$ and $\mathrm{Var}[U|X] = \sigma^2$ (i.e., under homoskedasticity). Under these conditions,

$$\mathrm{Var}[XU] = E[XX'U^2] = E[XX']\sigma^2 .$$

Hence,

$$\mathbb{V} = E[XX']^{-1} \sigma^2 .$$

A natural choice of estimator is therefore

$$\hat{\mathbb{V}}_n = \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \right)^{-1} \hat{\sigma}_n^2 ,$$

where $\hat{\sigma}_n^2$ is a consistent estimator of σ^2. The main difficulty in showing that this estimator is a consistent estimator of $\mathbb{V}$ lies in choosing a consistent estimator of σ^2. A natural choice of such an estimator is

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{1 \leq i \leq n} \hat{U}_i^2 .$$

Note that

$$\hat{U}_i = Y_i - X_i'\hat{\beta}_n = U_i - X_i'(\hat{\beta}_n - \beta) \ ,$$

so

$$\hat{U}_i^2 = (U_i - X_i'(\hat{\beta}_n - \beta))^2 = U_i^2 - 2U_iX_i'(\hat{\beta}_n - \beta) + (X_i'(\hat{\beta}_n - \beta))^2 \ .$$

The WLLN implies that

$$\frac{1}{n}\sum_{1\leq i\leq n} U_i^2 \xrightarrow{P} \sigma^2$$

as $n \to \infty$. Next, note that the WLLN and CMT imply further that

$$\frac{1}{n}\sum_{1\leq i\leq n} U_iX_i'(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)'\frac{1}{n}\sum_{1\leq i\leq n} X_iU_i = o_P(1) \ .$$

Finally, note that

$$\begin{aligned}\left|\frac{1}{n}\sum_{1\leq i\leq n}(X_i'(\hat{\beta}_n - \beta))^2\right| &\leq \frac{1}{n}\sum_{1\leq i\leq n}|(X_i'(\hat{\beta}_n - \beta))^2| \\ &\leq \frac{1}{n}\sum_{1\leq i\leq n}|X_i|^2|\hat{\beta}_n - \beta|^2 \ ,\end{aligned}$$

which tends in probability to zero because of the WLLN, CMT and the fact that $E[|X|^2] < \infty$ (which follows from the fact that $E[XX'] < \infty$). The desired conclusion thus follows.

When we do not assume $\text{Var}[U|X] = \sigma^2$, a natural choice of estimator is

$$\hat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\hat{U}_i^2\right)\left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\right)^{-1} \ . \quad (2.3)$$

Later in the class we will prove that this estimator is consistent, i.e.,

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V} \text{ as } n \to \infty \ ,$$

regardless of the functional form of $\text{Var}[U|X]$. This estimator is called the Heteroskedasticity Consistent (HC) estimator of $\mathbb{V}$. The standard errors used to construct t-statistics are the square roots of the diagonal elements of $\hat{\mathbb{V}}_n$, and this is the topic of the third part of this class. It is important to note that, by default, `Stata` reports homoskedastic-only standard errors.

2.5 Measures of Fit

When reporting the results of estimating a linear regression via OLS, it is common to report a measure of fit known as R^2, defined as follows:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \ ,$$

where

$$\begin{aligned} TSS &= \sum_{1 \leq i \leq n} (Y_i - \bar{Y}_n)^2 \\ ESS &= \sum_{1 \leq i \leq n} (\hat{Y}_i - \bar{Y}_n)^2 \\ SSR &= \sum_{1 \leq i \leq n} \hat{U}_i^2 \ . \end{aligned}$$

Here, TSS is short for *total sum of squares*, ESS is short for *explained sum of squares*, and SSR is short for *sum of squared residuals*. To show that the two expressions for R^2 are the same, and that $0 \leq R^2 \leq 1$, it suffices to show that

$$SSR + ESS = TSS \ .$$

Moreover, $R^2 = 1$ if and only if $SSR = 0$, i.e., $\hat{U}_i = 0$ for all $1 \leq i \leq n$. Similarly, $R^2 = 0$ if and only if $ESS = 0$, i.e., $\hat{Y}_i = \bar{Y}_n$ for all $1 \leq i \leq n$. In this sense, R^2 is a measure of the "fit" of a regression.

Note that $\frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - \bar{Y}_n)^2$ may be viewed as an estimator of $\mathrm{Var}[Y_i]$ and $\frac{1}{n} \sum_{1 \leq i \leq n} \hat{U}_i^2$ may be viewed as an estimator of $\mathrm{Var}[U_i]$. Thus, R^2 may be viewed as an estimator of the quantity

$$1 - \frac{\mathrm{Var}[U_i]}{\mathrm{Var}[Y_i]} \ .$$

Replacing these estimators with their unbiased counterparts yields "adjusted" R^2, defined as

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS} \ .$$

Note that R^2 always increases with the inclusion of an additional regressor, whereas $\bar{R}^2$ may not. Note also that $\bar{R}^2 \leq R^2$, so $\bar{R}^2 \leq 1$, but, unlike R^2, $\bar{R}^2$ may be less than zero.

It is important to understand that a high R^2 does not justify interpreting a linear regression as a causal model, just as a low R^2 does not invalidate interpreting a linear regression as a causal model.

Bibliography

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Lecture 3

Basic Inference and Endogeneity[1]

3.1 Inference

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, that there is no perfect collinearity in X, that $E[XX'] < \infty$, and $\mathrm{Var}[XU] < \infty$. Denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sample of random vectors with distribution P. Under these assumptions, we established the asymptotic normality of the OLS estimator $\hat{\beta}_n$,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V}) \tag{3.1}$$

with

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} \ . \tag{3.2}$$

We also described a consistent estimator $\hat{\mathbb{V}}_n$ of the limiting variance $\mathbb{V}$. We now use these results to develop methods for inference. We will study in particular *Wald tests* for certain hypotheses. Some other testing principles will be covered later in class. Confidence regions will be constructed using the duality between hypothesis testing and the construction of confidence regions.

Below we will assume further that $\mathrm{Var}[XU] = E[XX'U^2]$ is non-singular. This would be implied, for example, by the assumption that $P\{E[U^2|X] > 0\} = 1$. Since $E[XX']$ is non-singular under the assumption of no perfect collinearity in X, this implies that $\mathbb{V}$ is non-singular.

[1] This lecture is based on Azeem Shaikh's lecture notes. I want to thank him for kindly sharing them.

3.1.1 Background

Consider the following somewhat generic version of a testing problem. One observes data $W_i = (Y_i, X_i), i = 1, \ldots, n$, i.i.d. with distribution $P \in \mathbf{P} = \{P_\beta : \beta \in \mathbf{R}^{k+1}\}$ and wishes to test

$$H_0 : \beta \in \mathbf{B}_0 \quad \text{versus} \quad H_1 : \beta \in \mathbf{B}_1 \tag{3.3}$$

where $\mathbf{B}_0$ and $\mathbf{B}_1$ form a partition of $\mathbf{R}^{k+1}$. In our context, β will be the coefficient in a linear regression but in general it could be any other parameter.

A test is simply a function $\phi_n = \phi_n(W_1, \ldots, W_n)$ that returns the probability of rejecting the null hypothesis after observing $W_1, \ldots, W_n$. For the time being, we will only consider non-randomized tests which means that the function ϕ_n will take only two values: it will be equal to 1 for rejection and equal to 0 for non rejection. Most often, ϕ_n is the indicator function of a certain test statistic $T_n = T_n(W_1, \ldots, W_n)$ being greater than some critical value $c_n(1-\alpha)$, this is,

$$\phi_n = I\left\{T_n > c_n(1-\alpha)\right\} \ . \tag{3.4}$$

The test is said to be (pointwise) asymptotically of level α (or consistent in levels) if,

$$\limsup_{n\to\infty} E_{P_\beta}\left[\phi_n\right] = \limsup_{n\to\infty} P_\beta\{\phi_n = 1\} \leq \alpha \ , \quad \forall \beta \in \mathbf{B}_0 \ .$$

Such tests include: Wald tests, quasi-likelihood ratio tests, and Lagrange multiplier tests.

3.1.2 Tests of A Single Linear Restriction

Consider testing

$$H_0 : r'\beta = c \text{ versus } H_1 : r'\beta \neq c \ ,$$

where r is a nonzero $(k+1)$-dimensional vector and c is a scalar, at level α. Probably the most important case in this class happens when r selects the sth component of β, in which case we get

$$H_0 : \beta_s = c \text{ versus } H_1 : \beta_s \neq c \ .$$

The CMT implies that

$$\sqrt{n}(r'\hat{\beta}_n - r'\beta) \xrightarrow{d} N(0, r'\mathbb{V}r)$$

as $n \to \infty$. Since $\mathbb{V}$ is non-singular, $r'\mathbb{V}r > 0$. The CMT implies that $r'\hat{\mathbb{V}}_n r \xrightarrow{P} r'\mathbb{V}r$ as $n \to \infty$. A natural choice of test statistic for this problem is the absolute value of the t-statistic,

$$t_{\text{stat}} = \frac{\sqrt{n}(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\mathbb{V}}_n r}} \ ,$$

so that $T_n = |t_{\text{stat}}|$. Note that when r selects the sth component of β, we get $r'\hat{\mathbb{V}}_n r = \hat{\mathbb{V}}_{n,[s,s]}$, i.e., the sth diagonal element of $\hat{\mathbb{V}}_n$.

Such test statistic has the property that large values of T_n provide evidence against the null hypothesis H_0, and so using rejection rules of the form "reject H_0 if T_n is greater than a certain threshold" makes sense. This threshold value is usually called *critical value*.

A suitable choice of critical value for this test statistic is $z_{1-\frac{\alpha}{2}}$, which exploits the fact that, under the null hypothesis,

$$t_{\text{stat}} = \frac{\sqrt{n}(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\mathbb{V}}_n r}} \xrightarrow{d} N(0,1) \ . \tag{3.5}$$

To see that this test is consistent in level, note that whenever $r'\beta = c$,

$$\begin{aligned} P\{\phi_n = 1\} &= P\{|T_n| > z_{1-\frac{\alpha}{2}}\} \\ &= P\{t_{\text{stat}} > z_{1-\frac{\alpha}{2}}\} + P\{t_{\text{stat}} < -z_{1-\frac{\alpha}{2}}\} \\ &\to 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(-z_{1-\frac{\alpha}{2}}) \\ &= 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(z_{\frac{\alpha}{2}}) \\ &= 1 - (1 - \frac{\alpha}{2}) + \frac{\alpha}{2} \\ &= \alpha \ . \end{aligned}$$

This construction may be modified in a straightforward fashion for testing "one-sided" hypotheses, i.e.,

$$H_0 : r'\beta \leq c \text{ versus } H_1 : r'\beta > c \ .$$

In addition, note that by using the duality between hypothesis testing and the construction of confidence regions, we may construct a confidence region of level α for each component β_s of β as

$$\begin{aligned} C_n &= \left\{ c \in \mathbf{R} : \left| \frac{\sqrt{n}(\hat{\beta}_{n,s} - c)}{\sqrt{\hat{\mathbb{V}}_{n,[s,s]}}} \right| \leq z_{1-\frac{\alpha}{2}} \right\} \\ &= \left\{ \hat{\beta}_{n,s} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\mathbb{V}}_{n,[s,s]}}{n}}, \hat{\beta}_{n,s} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\mathbb{V}}_{n,[s,s]}}{n}} \right\} \ . \end{aligned}$$

This confidence region satisfies

$$P\{\beta_s \in C_n\} \to 1 - \alpha$$

as $n \to \infty$. It is straightforward to modify this construction to construct a confidence region of level α for $r'\beta$.

3.1.3 Tests of Multiple Linear Restrictions

Consider testing

$$H_0 : R\beta = c \text{ versus } H_1 : R\beta \neq c \ ,$$

where R is a $p \times (k+1)$-dimensional matrix and c is a p-dimensional vector, at level α. In order to rule out redundant equations, we require that the rows of R are linearly independent. The CMT implies that

$$\sqrt{n}(R\hat{\beta}_n - R\beta) \xrightarrow{d} N(0, R\mathbb{V}R')$$

as $n \to \infty$. Note that because $\mathbb{V}$ is assumed to be non-singular, $R\mathbb{V}R'$ is also non-singular. To see this, consider $a'R\mathbb{V}R'a$ for a non-zero vector $a \in \mathbf{R}^p$. Next, note that $a'R \neq 0$ because the rows of R are assumed to be linearly independent. Hence, $a'R\mathbb{V}R'a > 0$ because $\mathbb{V}$ is assumed to be non-singular. Hence, from our earlier results, we see that

$$n(R\hat{\beta}_n - R\beta)'(R\hat{\mathbb{V}}_n R')^{-1}(R\hat{\beta}_n - R\beta) \xrightarrow{d} \chi^2_p$$

as $n \to \infty$. Thus, a natural choice of test statistic in this case is therefore

$$T_n = n(R\hat{\beta}_n - c)'(R\hat{\mathbb{V}}_n R')^{-1}(R\hat{\beta}_n - c)$$

and a suitable choice of critical value is $c_{p,1-\alpha}$, the $1-\alpha$ quantile of χ^2_p. The resulting test is consistent in level.

Note that by using the duality between hypothesis testing and the construction of confidence regions, we may construct a confidence region of level α for β as

$$C_n = \{c \in \mathbf{R}^{k+1} : n(\hat{\beta}_n - c)'\hat{\mathbb{V}}_n^{-1}(\hat{\beta}_n - c) \leq c_{k+1,1-\alpha}\} \ .$$

This confidence region satisfies

$$P\{\beta \in C_n\} \to 1 - \alpha$$

as $n \to \infty$. It is straightforward to modify this construction to construct a confidence region of level α for $R\beta$.

3.1.4 Tests of Nonlinear Restrictions

Consider testing

$$H_0 : f(\beta) = 0 \text{ versus } H_1 : f(\beta) \neq 0 \ ,$$

where $f : \mathbf{R}^{k+1} \to \mathbf{R}^p$, at level α. Assume that f is continuously differentiable at β and denote by $D_\beta f(\beta)$ the $p \times (k+1)$-dimensional matrix of

partial derivatives of f evaluated at β. Assume that the rows of $D_\beta f(\beta)$ are linearly independent. The Delta Method implies that

$$\sqrt{n}(f(\hat{\beta}_n) - f(\beta)) \xrightarrow{d} N(0, D_\beta f(\beta) \mathbb{V} D_\beta f(\beta)')$$

as $n \to \infty$. The CMT implies that

$$D_\beta f(\hat{\beta}_n) \hat{\mathbb{V}}_n D_\beta f(\hat{\beta}_n)' \xrightarrow{P} D_\beta f(\beta) \mathbb{V} D_\beta f(\beta)'$$

as $n \to \infty$. It is now straightforward to modify the construction of the test in the preceding section appropriately to develop a test for the present purpose. It is also straightforward to modify the construction of the confidence region in the preceding section to construct a confidence region of level α for $f(\beta)$.

3.2 Linear Regression when $E[XU] \neq 0$

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is constant and equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

In contrast to our earlier discussion, we do not assume that $E[XU] = 0$. Any X_j such that $E[X_j U] = 0$ is said to be *exogenous*; any X_j such that $E[X_j U] \neq 0$ is said to be *endogenous*. By normalizing β_0 if necessary, we assume X_0 is exogenous. Note that it must be the case that we are interpreting this regression as a causal model.

Note that since $E[XU] \neq 0$ we have that

$$E[XY] = E[XX']\beta + E[XU]$$

and so

$$E[XX']^{-1}E[XY] = \beta + E[XX']^{-1}E[XU] \ .$$

The results from the previous class showed that the least squares estimator β_n of β converges to $E[XX']^{-1}E[XY]$. It follows that

$$\hat{\beta}_n \xrightarrow{P} \beta + E[XX']^{-1}E[XU] \ ,$$

and is therefore inconsistent for β under endogeneity.

3.2.1 Motivating Examples

We now briefly review some common ways in which endogeneity may arise.

Omitted Variables

Suppose $k = 2$, so

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \ .$$

We are interpreting this regression as a causal model and are willing to assume that $E[XU] = 0$ (i.e., $E[U] = E[X_1 U] = E[X_2 U] = 0$), but X_2 is unobserved. An example of a situation like this is when Y is wages, X_1 is education, and X_2 is ability. Given unobserved ability, we may rewrite this model as

$$Y = \beta_0^* + \beta_1^* X_1 + U^* \ ,$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_2 E[X_2] \\ \beta_1^* &= \beta_1 \\ U^* &= \beta_2 (X_2 - E[X_2]) + U \ . \end{aligned}$$

Note that we have normalized β_0^* so that $E[U^*] = 0$. In this model,

$$E[X_1 U^*] = \beta_2 \text{Cov}[X_1, X_2] \ ,$$

so X_1 is endogenous whenever $\beta_2 \neq 0$ and $\text{Cov}[X_1, X_2] \neq 0$. Based on the results from the previous class, it follows immediately that running a regression of Y on X_1 produces an estimator with the property that

$$\hat{\beta}_{1,n}^* \xrightarrow{P} \beta_1 + \beta_2 \frac{\text{Cov}[X_1, X_2]}{\text{Var}[X_1]} \ , \tag{3.6}$$

where the term $\beta_2 [\text{Var}[X_1]]^{-1} \text{Cov}[X_1, X_2]$ is usually referred to as *omitted variable bias*.

Measurement Error

Partition X into X_0 and X_1, where $X_0 = 1$ and X_1 takes values in $\mathbf{R}^k$. Partition β analogously. In this notation,

$$Y = \beta_0 + X_1' \beta_1 + U \ .$$

We are interpreting this regression as a causal model and are willing to assume that $E[XU] = 0$, but X_1 is *not* observed. Instead, $\hat{X}_1$ is observed, where

$$\hat{X}_1 = X_1 + V \ .$$

Assume $E[V] = 0$, $\text{Cov}[X_1, V] = 0$, and $\text{Cov}[U, V] = 0$. We may therefore rewrite this model as

$$Y = \beta_0^* + \hat{X}_1' \beta_1^* + U^* \ ,$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 \\ \beta_1^* &= \beta_1 \\ U^* &= -V'\beta_1 + U \ . \end{aligned}$$

In this model,

$$E[\hat{X}_1 U^*] = -E[\hat{X}_1 V'\beta_1] = -E[VV']\beta_1 \ ,$$

so $\hat{X}_1$ is typically endogenous. Note that in the case where X_1 is a scalar random variable, and using results from the previous class, it follows that running a regression of Y on $\hat{X}_1$ produces an estimator with the property that

$$\hat{\beta}_{1,n}^* \xrightarrow{P} \beta_1 + \frac{E[\hat{X}_1 U^*]}{\text{Var}[\hat{X}_1]} = \beta_1 \left(1 - \frac{\text{Var}[V]}{\text{Var}[\hat{X}_1]}\right) < \beta_1 \ , \tag{3.7}$$

so that the regression coefficient is biased towards zero when the regressor of interest is measured with the so-called classical random errors. The last inequality follows from using $\hat{X}_1 = X_1 + V$. Indeed, in the extreme case where $\hat{X}_1 = V$, it follows that $\hat{\beta}_{1,n}^* \xrightarrow{P} 0$.

Simultaneity

A classical example of simultaneity is given by supply and demand. Denote by Q^s the quantity supplied and by Q^d the quantity demanded. As a function of (non-market clearing) price $\tilde{P}$, assume that

$$\begin{aligned} Q^d &= \beta_0^d + \beta_1^d \tilde{P} + U^d \\ Q^s &= \beta_0^s + \beta_1^s \tilde{P} + U^s \ , \end{aligned}$$

where $E[U^s] = E[U^d] = E[U^s U^d] = 0$. We observe (Q, P), where Q and P are such that the market clears, i.e., $Q^s = Q^d$. This implies that

$$\beta_0^d + \beta_1^d P + U^d = \beta_0^s + \beta_1^s P + U^s \ ,$$

so

$$P = \frac{1}{\beta_1^s - \beta_1^d}(\beta_0^d - \beta_0^s + U^d - U^s) \ .$$

It follows that P is endogenous in both of the equations

$$\begin{aligned} Q &= \beta_0^d + \beta_1^d P + U^d \\ Q &= \beta_0^s + \beta_1^s P + U^s \end{aligned}$$

because

$$\begin{aligned} E[PU^d] &= \frac{\text{Var}[U^d]}{\beta_1^s - \beta_1^d} \\ E[PU^s] &= -\frac{\text{Var}[U^s]}{\beta_1^s - \beta_1^d} \ . \end{aligned}$$

As it was the case in the previous two examples, running a regression of Q in P will result in estimators that do not converge to either β_1^d or β_1^s.

Bibliography

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Lecture 4

Endogeneity[1]

4.1 Instrumental Variables

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is constant and equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

In contrast to our earlier discussion, we do not assume that $E[XU] = 0$. Any X_j such that $E[X_jU] = 0$ is said to be *exogenous*; any X_j such that $E[X_jU] \neq 0$ is said to be *endogenous*.

In order to overcome the difficulty associated with $E[XU] \neq 0$, we assume that there is an additional random vector Z taking values in $\mathbf{R}^{\ell+1}$ with $\ell + 1 \geq k + 1$ such that $E[ZU] = 0$. We assume that any exogenous component of X is contained in Z (the so-called included instruments). In particular, we assume that the first component of Z is constant and equal to one, i.e., $Z = (Z_0, Z_1, \ldots, Z_\ell)'$ with $Z_0 = 1$. We also assume that $E[ZX'] < \infty$, $E[ZZ'] < \infty$ and that there is no perfect collinearity in Z. The components of Z are sometimes referred to as *instruments* or *instrumental variables*. The requirement that $E[ZU] = 0$ is termed *instrument exogeneity*. We further assume that the rank of $E[ZX']$ is $k+1$. This condition is termed *instrument relevance* or the *rank condition*. Note that a necessary condition for this to be true is that $\ell \geq k$. This is sometimes referred to as the *order condition*.

Using the fact that $U = Y - X'\beta$ and $E[ZU] = 0$, we see that β solves the system of equations

$$E[ZY] = E[ZX']\beta \ .$$

[1]This lecture is based on Azeem Shaikh's lecture notes. I want to thank him for kindly sharing them.

Since $\ell+1 \geq k+1$, this may be an over-determined system of equations. In order to find an explicit formula for β, the following lemma is useful.

Lemma 4.1 *Suppose there is no perfect collinearity in Z and let Π be such that $BLP(X|Z) = \Pi' Z$. $E[ZX']$ has rank $k+1$ if and only if Π has rank $k+1$. Moreover, the matrix $\Pi' E[ZX']$ is invertible.*

PROOF: Write $X = \Pi' Z + V$ where $E[ZV'] = 0$. It follows that $E[ZX'] = E[ZZ']\Pi$. Recall the rank inequality, which states that

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$

for any conformable matrices A and B. Applying this result, we see that

$$\text{rank}(E[ZZ']\Pi) \leq \text{rank}(\Pi) \ .$$

We have further that

$$\text{rank}(\Pi) = \text{rank}(E[ZZ']^{-1}E[ZZ']\Pi) \leq \text{rank}(E[ZZ']\Pi) \ .$$

Hence,

$$\text{rank}(E[ZX']) = \text{rank}(E[ZZ']\Pi) = \text{rank}(\Pi) \ ,$$

as desired.

To complete the proof, note that $\Pi' E[ZX'] = \Pi' E[ZZ']\Pi$ and argue that $\Pi' E[ZZ']\Pi$ is invertible using arguments given earlier. ■

Since β solves $\Pi' E[ZY] = \Pi' E[ZX']\beta$, we arrive at two formulae for β by applying the lemma:

$$\beta = (\Pi' E[ZX'])^{-1}\Pi' E[ZY] \tag{4.1}$$

$$= (\Pi' E[ZZ']\Pi)^{-1}\Pi' E[ZY] \ . \tag{4.2}$$

Note that if $k = \ell$, then Π is an invertible matrix and therefore

$$\beta = (E[ZX'])^{-1}E[ZY] \ . \tag{4.3}$$

In this case, we say that β is *exactly identified*. Otherwise, we say that β is *over-identified*.

A third formula for β arises by replacing Π with $E[ZZ']^{-1}E[ZX']$,

$$\beta = \left(E[ZX']' E[ZZ']^{-1} E[ZX']\right)^{-1} E[ZX']' E[ZZ']^{-1} E[ZY] \ . \tag{4.4}$$

Before proceeding, it is useful to use the preceding lemma to further examine the rank condition in some simpler settings. To this end, consider the case

where $k = \ell$ and only X_k is endogenous. Let $Z_j = X_j$ for all $0 \leq j \leq k-1$. In this case,

$$\Pi' = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ \pi_0 & \pi_1 & \cdots & \pi_{\ell-1} & \pi_\ell \end{pmatrix}$$

The rank condition therefore requires $\pi_\ell \neq 0$. In other words, the instrument Z_ℓ must be "correlated with X_k after controlling for $X_0, X_1, \ldots, X_{k-1}$."

4.1.1 Partition of β: solve for endogenous components

Partition X into X_1 and X_2, where X_2 is exogenous. Partition Z into Z_1 and Z_2 and β into β_1 and β_2 analogously. Note that $Z_2 = X_2$ are *included* instruments and Z_1 are *excluded* instruments. In this notation,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U \ .$$

Note that

$$\begin{aligned} \text{BLP}(Y|Z_2) &= \text{BLP}(X_1'\beta_1|Z_2) + \text{BLP}(X_2'\beta_2|Z_2) + \text{BLP}(U|Z_2) \\ &= \text{BLP}(X_1|Z_2)'\beta_1 + X_2'\beta_2 \ , \end{aligned}$$

where the second equality uses the fact that $E[Z_2 U] = 0$. It follows that

$$Y^* = X_1^{*\prime}\beta_1 + U \ ,$$

where

$$\begin{aligned} Y^* &= Y - \text{BLP}(Y|Z_2) \\ X_1^* &= X_1 - \text{BLP}(X_1|Z_2) \ . \end{aligned}$$

This calculation shows again the sense in which we may interpret β_1 as summarizing the effect of X_1 on Y "after controlling for X_2." In the exactly identified case, it follows that

$$E[Z_1 Y^*] = E[Z_1 X_1^{*\prime}]\beta_1 \ .$$

Since there must be a unique solution to this system of equations, it must be the case that $E[Z_1 X_1^{*\prime}]$ is invertible. It follows that

$$\beta_1 = E[Z_1 X_1^{*\prime}]^{-1} E[Z_1 Y^*] \ .$$

In the over-identified case, we may repeat this calculation with $\hat{X}_1^* = \text{BLP}(X_1^*|Z_1)$ in place of Z_1. This yields

$$\begin{aligned} \beta_1 &= E[\hat{X}_1^* X_1^{*\prime}]^{-1} E[\hat{X}_1^* Y^*] \\ &= E[\hat{X}_1^* \hat{X}_1^{*\prime}]^{-1} E[\hat{X}_1^* Y^*] \ , \end{aligned}$$

where the second equality uses the fact that $X_1^* = \hat{X}_1^* + V$ with $E[\hat{X}_1^* V'] = 0$.

4.2 Estimating β

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there is no perfect collinearity in Z, and that the rank of $E[ZX']$ is $k+1$. We now discuss estimation of β.

4.2.1 The Instrumental Variables (IV) Estimator

We first consider the case in which $k = \ell$. Let (Y, X, Z, U) be distributed as described above and denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P. By analogy with the expression we derived for β in (4.3) under these assumptions, the natural estimator of β is simply

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Y_i \right) \ .$$

This estimator is called the *instrumental variables* (IV) estimator of β. Note that $\hat{\beta}_n$ satisfies

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i (Y_i - X_i' \hat{\beta}_n) = 0 \ .$$

In particular, $\hat{U}_i = Y_i - X_i' \hat{\beta}_n$ satisfies

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i \hat{U}_i = 0 \ .$$

To gain further insight on the IV estimator, partition X into X_0 and X_1, where $X_0 = 1$ and X_1 is assumed to take values in $\mathbf{R}$. Do the same with Z and β. An interesting interpretation of the IV estimator of β_1 is obtained by multiplying and dividing by $\frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n})^2$, i.e.,

$$\hat{\beta}_{1,n} = \frac{\frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n}) Y_i / \frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n})^2}{\frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n}) X_{1,i} / \frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n})^2} \ . \tag{4.5}$$

The IV estimator of β_1 is simply the ratio of the regression slope of Y on Z_1 (the so-called reduced form) to the regression slope of X_1 on Z_1 (the so-called first stage). To see this in a different way, write the model as

$$Y = \beta_0 + \beta_1 X_1 + U$$

and

$$X_1 = \pi_0 + \pi_1 Z_1 + V \ ,$$

so that replacing the second equation into the first one delivers

$$Y = \beta_0^* + \beta_1 \pi_1 Z_1 + U^*$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_1 \pi_0 \\ U^* &= U + \beta_1 V \ . \end{aligned}$$

Thus, the estimated slope in the reduced form converges in probability to $\beta_1 \pi_1$, while the estimated slope in the first stage converges to π_1. The IV estimator takes the ratio of these two, therefore delivering a consistent estimator of β_1. Note that the IV estimand is predicated on the notion that the first stage slope is not zero ($\pi_1 \neq 0$), which is just another way to state our rank condition in this simple case.

This estimator may be expressed more compactly using matrix notation. Define

$$\begin{aligned} \mathbb{Z} &= (Z_1, \ldots, Z_n)' \\ \mathbb{X} &= (X_1, \ldots, X_n)' \\ \mathbb{Y} &= (Y_1, \ldots, Y_n)' \ . \end{aligned}$$

In this notation, we have

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

4.2.2 The Two-Stage Least Squares (TSLS) Estimator

Now consider the case in which $\ell > k$. The expressions derived for β in this case involved Π, where $\text{BLP}(X|Z) = \Pi' Z$. An estimate of Π can be obtained by OLS. More precisely, since $\Pi = E[ZZ']^{-1}E[ZX']$, a natural estimator of Π is

$$\hat{\Pi}_n = \left(\frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} Z_i X_i' \right) \ .$$

With this estimator of Π, a natural estimator of β is simply

$$\begin{aligned} \hat{\beta}_n &= \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Y_i \right) \\ &= \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Z_i' \hat{\Pi}_n \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Y_i \right) \ . \end{aligned}$$

The first equation above provides an interpretation of the TSLS estimator as an IV estimator with $\hat{\Pi}_n' Z_i$ playing the role of the instrument. Note further that if $k+1 = \ell+1$ and $\hat{\Pi}_n$ is invertible, then the TSLS estimator of β is exactly to the IV estimator of β. The second equality might be expected from our calculations in (4.2). To justify it here, write $X_i = \hat{\Pi}_n' Z_i + \hat{V}_i$ and note from properties of OLS that

$$\frac{1}{n} \sum_{1 \leq i \leq n} Z_i \hat{V}_i' = 0 \ .$$

This estimator of β is called the *two-stage least squares* (TSLS) estimator of β. Note that $\hat{\beta}_n$ satisfies

$$\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i (Y_i - X_i' \hat{\beta}_n) = 0 \ .$$

In particular, $\hat{U}_i = Y_i - X_i' \hat{\beta}_n$ satisfies

$$\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i \hat{U}_i = 0 \ .$$

Notice that this implies that $\hat{U}_i$ is orthogonal to all of the instruments equal to an exogenous regressors, but may not be orthogonal to the other regressors. It is termed the TSLS estimator because it may be obtained in the following way: first, regress (each component of) X_i on Z_i to obtain $\hat{X}_i = \hat{\Pi}_n' Z_i$; second, regress Y_i on $\hat{X}_i$ to obtain $\hat{\beta}_n$. However, in order to obtain proper standard errors, it is recommended to compute the estimator in one step (see the following section).

The estimator may again be expressed more compactly using matrix notation. Define

$$\begin{aligned} \hat{\mathbb{X}} &= (\hat{X}_1, \ldots, \hat{X}_n)' \\ &= \mathbb{P}_Z \mathbb{X} \ , \end{aligned}$$

where

$$\mathbb{P}_Z = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$$

is the projection matrix onto the column space of $\mathbb{Z}$. In this notation, we have

$$\begin{aligned} \hat{\beta}_n &= (\hat{\mathbb{X}}'\mathbb{X})^{-1}(\hat{\mathbb{X}}'\mathbb{Y}) \\ &= (\mathbb{X}'\mathbb{P}_Z\mathbb{X})^{-1}(\mathbb{X}'\mathbb{P}_Z\mathbb{Y}) \ , \end{aligned}$$

which should be expected given our previous derivation in (4.4).

4.3 Properties of the TSLS Estimator

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there there is no perfect collinearity in Z, and that the rank of $E[ZX']$ is $k+1$. Denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P. Above we described estimation of β via TSLS under these assumptions. We now discuss properties of the resulting estimator, $\hat{\beta}_n$, imposing stronger assumptions as needed.

4.3.1 Consistency

Under the assumptions stated above, the TSLS estimator, $\hat{\beta}_n$, is consistent for β, i.e., $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. To see this, first recall from our results on OLS that

$$\hat{\Pi}_n \xrightarrow{P} \Pi$$

as $n \to \infty$. Next, note that the WLLN implies that

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \xrightarrow{P} E[ZZ']$$
$$\frac{1}{n} \sum_{1 \le i \le n} Z_i Y_i \xrightarrow{P} E[ZY]$$

as $n \to \infty$. The desired result therefore follows from the CMT.

4.3.2 Limiting Distribution

In addition to the assumptions above, assume that $\mathrm{Var}[ZU] = E[ZZ'U^2] < \infty$. Then,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

as $n \to \infty$, where

$$\mathbb{V} = E[\Pi' ZZ' \Pi]^{-1} \Pi' \mathrm{Var}[ZU] \Pi E[\Pi' ZZ' \Pi]^{-1} \ .$$

To see this, note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\hat{\Pi}_n' \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \left(\hat{\Pi}_n' \left(\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} Z_i U_i \right) \right) \ .$$

As in the preceding section, we have that

$$\begin{aligned} \hat{\Pi}_n &\xrightarrow{P} \Pi \\ \frac{1}{n}\sum_{1\leq i\leq n} Z_i Z_i' &\xrightarrow{P} E[ZZ'] \end{aligned}$$

as $n \to \infty$. The CLT implies that

$$\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n} Z_i U_i \xrightarrow{d} N(0, \mathrm{Var}[ZU]) \ .$$

The desired result thus follows from the CMT.

4.3.3 Estimation of $\mathbb{V}$

A natural estimator of $\mathbb{V}$ is given by

$$\begin{aligned} \hat{\mathbb{V}}_n &= \left(\hat{\Pi}_n' \left(\frac{1}{n}\sum_{1\leq i\leq n} Z_i Z_i'\right)\hat{\Pi}_n\right)^{-1} \\ &\quad \times \hat{\Pi}_n' \left(\frac{1}{n}\sum_{1\leq i\leq n} Z_i Z_i' \hat{U}_i^2\right)\hat{\Pi}_n \\ &\quad \times \left(\hat{\Pi}_n' \left(\frac{1}{n}\sum_{1\leq i\leq n} Z_i Z_i'\right)\hat{\Pi}_n\right)^{-1} , \end{aligned}$$

where $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$. As in our discussion of OLS, the primary difficulty in establishing the consistency of this estimator lies in showing that

$$\frac{1}{n}\sum_{1\leq i\leq n} Z_i Z_i' \hat{U}_i^2 \xrightarrow{P} \mathrm{Var}[ZU]$$

as $n \to \infty$. Note that part of the complication lies in the fact that we do not observe U_i and therefore have to use $\hat{U}_i$. However, the desired result can be shown by arguing exactly as in the second part of this class.

Note that $\hat{U}_i = Y_i - X_i'\hat{\beta}_n \neq Y_i - \hat{X}_i'\hat{\beta}_n$, so the standard errors from two repeated applications of OLS will be incorrect. Assuming $\mathrm{Var}[ZU]$ is invertible, inference may now be carried out exactly the same way as discussed for the OLS estimator, simply replacing the OLS quantities with their TSLS counterparts.

Bibliography

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Lecture 5

More on Endogeneity[1]

5.1 Efficiency of the TSLS Estimator

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there is no perfect collinearity in Z, and that the rank of $E[ZX']$ is $k+1$. Denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P.

The TSLS estimator identifies β by means of the projection matrix $\Pi = E[ZZ']^{-1}E[ZX']$. However, note that we could have solved for β using any $(\ell+1) \times (k+1)$ dimensional matrix Γ such that $E[\Gamma' ZX']$ has rank $k+1$. For any such matrix,

$$\beta = E[\Gamma' ZX']^{-1} E[\Gamma' ZY] \ ,$$

and we could have estimated β using

$$\tilde{\beta}_n = \left(\frac{1}{n} \sum_{1 \le i \le n} \Gamma' Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \le i \le n} \Gamma' Z_i Y_i \right) \ .$$

Note one could use a consistent estimate of Γ, $\hat{\Gamma}_n$, instead. By arguing as before, it is possible to show under our assumptions that $\tilde{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. If in addition $\mathrm{Var}[ZU] = E[ZZ'U^2] < \infty$, then, by arguing as before, it is also possible to show that

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N(0, \tilde{\mathbb{V}})$$

[1] This lecture is based on Azeem Shaikh's lecture notes. I want to thank him for kindly sharing them.

as $n \to \infty$, where

$$\tilde{\mathbb{V}} = E[\Gamma' Z X']^{-1} \Gamma' \mathrm{Var}[ZU] \Gamma E[\Gamma' Z X']^{-1\prime} .$$

We now argue that under certain assumptions, the "best" choice of Γ is given by Π, i.e., $\tilde{\mathbb{V}} \geq \mathbb{V}$.

In order to establish this claim, we assume that $E[U|Z] = 0$ and $\mathrm{Var}[U|Z] = \sigma^2$. In addition, define $W^* = \Pi' Z$ and $W = \Gamma' Z$, which will simplify the notation below. To see that $\tilde{\mathbb{V}} \geq \mathbb{V}$, first note that under these assumptions

$$\begin{aligned}
\tilde{\mathbb{V}} &= \sigma^2 E[\Gamma' Z X']^{-1} E[\Gamma' Z Z' \Gamma] E[\Gamma' Z X']^{-1\prime} \\
&= \sigma^2 E[\Gamma' Z Z' \Pi]^{-1} E[\Gamma' Z Z' \Gamma] E[\Gamma' Z Z' \Pi]^{-1\prime} \\
&= \sigma^2 E[W W^{*\prime}]^{-1} E[W W'] E[W W^{*\prime}]^{-1\prime}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V} &= \sigma^2 E[\Pi' Z Z' \Pi]^{-1} E[\Pi' Z Z' \Pi] E[\Pi Z Z' \Pi]^{-1} \\
&= \sigma^2 E[\Pi' Z Z' \Pi]^{-1} \\
&= \sigma^2 E[W^* W^{*\prime}]^{-1} ,
\end{aligned}$$

where in both cases the first equality follows from $\mathrm{Var}[ZU] = E[ZZ'U^2] = \sigma^2 E[ZZ']$, and the second equality used the fact that $X = \Pi' Z + V$ with $E[ZV'] = 0$. It suffices to show that $\tilde{\mathbb{V}}^{-1} \leq \mathbb{V}^{-1}$, i.e., to show that

$$E[W^* W^{*\prime}] - E[W W^{*\prime}]' E[W W']^{-1} E[W W^{*\prime}] \geq 0 .$$

Yet this follows upon realizing that the left-hand side of the preceding display is simply $E[\hat{W}^* \hat{W}^{*\prime}]$ with

$$\hat{W}^* = W^* - \mathrm{BLP}(W^*|W) = W^* - E[W W^{*\prime}]' E[W W']^{-1} W .$$

When we do not assume that $E[U|Z] = 0$ and $\mathrm{Var}[U|Z] = \sigma^2$, then better estimators for β exist. Such estimators are most easily treated as a special case of the *generalized method of moments* (GMM), which will be covered later in class.

5.2 "Weak" Instruments

It turns out that the normal approximation justified by the preceding results can be poor in finite samples, especially when the rank of $E[ZX']$ is "close" to being $< k+1$. As a result, hypothesis tests and confidence regions based off of this approximation can behave poorly in finite-samples as well. To gain some insight into this phenomenon in a more elementary way, suppose

$$\begin{aligned}
Y_i &= X_i \beta + U_i \\
X_i &= Z_i \pi + V_i ,
\end{aligned}$$

where $Z_1, \ldots, Z_n$ are non-random, $(U_1, V_1), \ldots, (U_n, V_n)$ is a sequence of i.i.d. $N(0, \Sigma)$ random vectors. Suppose $\pi \neq 0$. Consider the estimator given by

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n Z_i Y_i}{\frac{1}{n}\sum_{i=1}^n Z_i X_i} \ .$$

Note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i U_i}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)\pi + \frac{1}{n}\sum_{i=1}^n Z_i V_i} \ .$$

The finite-sample, joint distribution of the numerator and denominator is simply

$$N\left(\begin{array}{c} 0 \\ \bar{Z}_n^2 \pi \end{array} , \left(\begin{array}{cc} \bar{Z}_n^2 \sigma_U^2 & \frac{1}{\sqrt{n}}\bar{Z}_n^2 \sigma_{U,V} \\ \frac{1}{\sqrt{n}}\bar{Z}_n^2 \sigma_{U,V} & \frac{1}{n}\bar{Z}_n^2 \sigma_V^2 \end{array} \right) \right) \ ,$$

where

$$\bar{Z}_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i^2 \ .$$

This joint distribution completely determines the finite-sample distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$. In particular, it is the ratio of two (correlated) normal random variables. If $\bar{Z}_n^2 \to \bar{Z}^2$ as $n \to \infty$, then it is straightforward to show that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_U^2}{\pi^2 \bar{Z}^2}\right) \ .$$

This approximation effectively treats the denominator like a constant equal to its mean, so we would expect it to be "good" when the mean is "large", i.e.,

$$\bar{Z}_n^2 \pi \gg \frac{1}{\sqrt{n}}\sqrt{\bar{Z}_n^2}\sigma_V \ .$$

When π is "small" , i.e., however, the approximation may be quite poor in finite-samples. Note in particular that $\pi \neq 0$ is not sufficient for the approximation to be good in finite-samples.

A variety of ways of carrying out inference that does not suffer from this problem have been proposed in the literature. We now describe one simple and popular method. Consider the problem of testing the null hypothesis that $H_0 : \beta = c$ versus the alternative hypothesis $H_1 : \beta \neq c$ at level α. Note that under the null hypothesis, one can compute $U_i = Y_i - X_i'\beta$ and $Z_i U_i = Z_i(Y_i - X_i'\beta)$. Since it must be the case that $E[ZU] = 0$, we can simply test whether this is true using $Z_1 U_1, \ldots, Z_n U_n$. To formalize this idea, assume $\mathrm{Var}[ZU]$ is invertible and define $W_i(c) = Z_i(Y_i - X_i'c)$. Note that when $\beta = c$, we have that

$$\sqrt{n}\bar{W}_n(c) = \frac{1}{\sqrt{n}}\sum_{1 \leq i \leq n} W_i(c) \xrightarrow{d} N(0, \Sigma(c)) \ ,$$

where $\Sigma(c) = \text{Var}[W(c)]$. Define

$$\hat{\Sigma}_n(c) = \frac{1}{n} \sum_{1 \leq i \leq n} (W_i(c) - \bar{W}_n(c))(W_i(c) - \bar{W}_n(c))' \ .$$

Using arguments given earlier, we see when $\beta = c$ that

$$T_n = n\bar{W}_n'(c)\hat{\Sigma}_n^{-1}(c)\bar{W}_n(c) \xrightarrow{d} \chi^2_{\ell+1} \ .$$

One may therefore test the null hypothesis by comparing T_n with $c_{\ell+1,1-\alpha}$, the $1-\alpha$ quantile of the $\chi^2_{\ell+1}$ distribution. As we will discuss in the second part of the class, one may now construct a confidence region using the duality between hypothesis testing and the construction of confidence regions. A closely related variant of this idea leads to the *Anderson-Rubin* test, in which one tests whether all of the coefficients in a regression of $Y_i - X_i'c$ on Z_i are zero.

Recent research in econometrics suggests that this method has good power properties when the model is exactly identified, but may be less desirable when the model is over-identified. Other methods for the case in which the model is over-identified and/or one is only interested in some feature of β (e.g., one of the slope parameters) have been proposed and are the subject of current research as well.

Instead of using these "more complicated" methods, researchers may attempt a two-step method as follows. In the first step, they would investigate whether the rank of $E[ZX']$ is "close" to being $< k+1$ or not by carrying out a hypothesis test of the null hypothesis that $H_0 : \text{rank}(E[ZX']) < k+1$ versus the alternative hypothesis that $H_1 : \text{rank}(E[ZX']) = k+1$. In some cases, such a test is relatively easy to carry out given what we have already learned: e.g., when there is a single endogenous regressor, such a test is equivalent to a test of the null hypothesis that certain coefficients in a linear regression are all equal to zero versus not all equal to zero. In the second step, they would only use these "more complicated" methods if they failed to reject this null hypothesis. This two-step method will also behave poorly in finite-samples and should not be used. A deeper discussion of these "uniformity" issues take place in Econ 481.

5.3 Interpretation under Heterogeneity

Despite possible inefficiencies, TSLS remains popular. One reason stems from the following interpretation in the presence of heterogeneous effects of X on Y. To motivate this, note that in the model

$$Y = X'\beta + U \ ,$$

the effect of a change in X (say, from $X = x$ to $X = x'$) is the *same* for everybody. It seems sensible that in many cases the effect of a change in X

on Y may be different for different people. To allow for such heterogeneity, we allow for β to be random. When β is random, we may absorb U into the intercept and simply write

$$Y = X'\beta \ .$$

Note that this means that when we work with a random sample where variables are indexed by i, we would write $Y_i = X_i'\beta_i$, which makes it explicit that every individual has a unique effect β_i.

For the time being, assume that $k = 1$ and write D in place of X_1, which is assumed to take values in $\{0, 1\}$. In this notation,

$$Y = \beta_0 + \beta_1 D \ .$$

In this case, we interpret β_0 as $Y(0)$ and β_1 as $Y(1) - Y(0)$, where $Y(1)$ and $Y(0)$ are *potential* or *counterfactual outcomes*. Using this notation, we may rewrite the equation as

$$Y = DY(1) + (1 - D)Y(0) \ .$$

The potential outcome $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) D were 0; the potential outcome $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) D were 1. The variable D is typically called the *treatment* and $Y(1) - Y(0)$ is called the *treatment effect*. The quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*.

If D were randomly assigned (e.g., by the flip of a coin, as in a randomized controlled trial), then

$$(Y(0), Y(1)) \perp\!\!\!\perp D \ .$$

In this case, under mild assumptions, the slope coefficient from OLS regression of Y on a constant and D yields a consistent estimate of the average treatment effect. To see this, note that the estimand is

$$\begin{aligned} \frac{\text{Cov}[Y, D]}{\text{Var}[D]} &= E[Y|D = 1] - E[Y|D = 0] \\ &= E[Y(1)|D = 1] - E[Y(0)|D = 0] \\ &= E[Y(1) - Y(0)] \ , \end{aligned}$$

where the first equality follows from a homework exercise, the second equality follows from the equation for Y, and the third equality follows from independence of $(Y(0), Y(1))$ and D.

Otherwise, we generally expect D to depend on $(Y(1), Y(0))$. In this case, OLS will not yield a consistent estimate of the average treatment effect. To proceed further, we therefore assume, as usual, that there is an instrument Z that also takes values in $\{0, 1\}$. We may thus consider the

slope coefficient from TSLS regression of Y on D with Z as an instrument. The estimand in this case is

$$\frac{\text{Cov}[Y,Z]}{\text{Cov}[D,Z]} = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]} ,$$

where the equality follows by multiplying and dividing by $\text{Var}[Z]$ and using earlier results. Our goal is to express this quantity in terms of the treatment effect $Y(1)-Y(0)$ somehow. To this end, analogously to our equation for Y above, it is useful to also introduce a similar equation for D:

$$\begin{aligned} D &= ZD(1)+(1-Z)D(0) \\ &= D(0)+(D(1)-D(0))Z \\ &= \pi_0+\pi_1 Z , \end{aligned}$$

where $\pi_0 = D(0)$, $\pi_1 = D(1)-D(0)$, and $D(1)$ and $D(0)$ are *potential* or *counterfactual treatments* (rather than outcomes). We impose the following versions of instrument exogeneity and instrument relevance, respectively:

$$(Y(1),Y(0),D(1),D(0)) \perp\!\!\!\perp Z$$

and

$$P\{D(1)\neq D(0)\} = P\{\pi_1 \neq 0\} > 0 .$$

Note that the first part of the assumption basically states that Z is as good as randomly assigned. In addition, note that we are implicitly assuming that Z does not affect Y directly, i.e., potential outcomes take the form $Y(d)$ as opposed to $Y(d,z)$. This is the exclusion restriction in this setting. In the linear model with constant effects, the exclusion restriction is expressed by the omission of the instruments from the causal equation of interest and by requiring that $E[ZU]=0$.

We further assume the following *monotonicity* (or perhaps better called *uniform monotonicity*) condition:

$$P\{D(1)\geq D(0)\} = P\{\pi_1 \geq 0\} = 1 .$$

Under these assumptions, note that

$$\begin{aligned} E[Y|Z=1]-E[Y|Z=0] &= E[Y(1)D(1)+Y(0)(1-D(1))|Z=1] \\ &\quad - E[Y(1)D(0)+Y(0)(1-D(0))|Z=0] \\ &= E[Y(1)D(1)+Y(0)(1-D(1))] \\ &\quad - E[Y(1)D(0)+Y(0)(1-D(0))] \\ &= E[(Y(1)-Y(0))(D(1)-D(0))] \\ &= E[Y(1)-Y(0)|D(1)>D(0)]P\{D(1)>D(0)\} , \end{aligned}$$

where the first equality follows from the equations for Y and D, the second equality follows from instrument exogeneity, and the fourth equality follows from the monotonicity assumption. Furthermore,

$$E[D|Z=1] - E[D|Z=0] = E[D(1) - D(0)] = P\{D(1) > D(0)\} \ .$$

Hence, the TSLS estimand equals

$$E[Y(1) - Y(0)|D(1) > D(0)] \ ,$$

which is termed the *local average treatment effect* (LATE). It is the average treatment effect among the subpopulation of people for whom a change in the value of the instrument switched them from being non-treated to treated. We often refer to such subpopulation as *compliers*.

A few remarks are in order: First, it is important to understand that this result depends crucially on the monotonicity assumption. Second, it is important to understand that this quantity may or may not be of interest. Third, it is important to understand that a consequence of this calculation is that in a world with heterogeneity "different instruments estimate different parameters." Finally, this result also depends on the simplicity of the model. When covariates are present, the entire calculation breaks down. Some generalizations are available.

5.3.1 Monotonicity in Latent Index Models

The monotonicity assumption states that while the instrument may have no effect on some people, all those who are affected are affected in the same way. Without monotonicity, we would have

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] = {} & E[Y(1) - Y(0)|D(1) > D(0)]P\{D(1) > D(0)\} \\ & - E[Y(1) - Y(0)|D(1) < D(0)]P\{D(1) < D(0)\} \ . \end{aligned}$$

We might therefore have a situation where treatment effects are positive for everyone (i.e., $Y(1) - Y(0) > 0$) yet the reduced form is zero because effects on compliers are canceled out by effects on *defiers*, i.e., those individuals for which the instrument pushes them out of treatment ($D(1) = 0$ and $D(0) = 1$). This doesn't come up in a constant effect model where $\beta = Y(1) - Y(0)$ is constant, as in such case

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] &= \beta\{P\{D(1) > D(0)\} - P\{D(1) < D(0)\}\} \\ &= \beta E[D(1) - D(0)] \ , \end{aligned}$$

and so a zero reduced-form effect means either the first stage is zero or $\beta = 0$.

It is worth noting that monotonicity assumptions are easy to interpret in latent index models. In such models individual choices are determined by

a threshold crossing rule involving observed and unobserved components of the utility. In our context, we could write

$$D = \begin{cases} 1 & \text{if } \gamma_0 + \gamma_1 Z - V > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma_1 > 0$ and V is an unobserved random variable assumed to be independent of Z. This latent index model characterizes potential treatment assignments as

$$D(0) = I\{\gamma_0 > V\} \text{ and } D(1) = I\{\gamma_0 + \gamma_1 > V\} \ .$$

Notably, in this model the monotonicity assumption is automatically satisfied since $\gamma_1 > 0$ is a constant.

5.3.2 IV in Randomized Experiments

The preceding derivation is often relevant in the context of randomized trials, where treatment assignment is independent of potential outcomes. However, in cases where there is non-compliance, one could interpret the treatment assignment as an "offer of treatment" Z (the instrument), and the actual treatment D as the variable that determines whether the subject actually had the intended treatment. This is the case in experiments where participation is voluntary among those randomly assigned to receive treatment. At the same time, it is often the case that no one in the control group has access to the experimental intervention. In other words, $D(0) = 0$ while $D(1) \in \{0, 1\}$. Since the group that receives the assigned treatment (the compliers) is a self-selected subset of those offered treatment, a comparison between those actually treated ($D = 1$) and the control ($D = 0$) group is misleading. Two alternatives are frequently used. The first one is a comparison between those who where *offered* treatment ($Z = 1$) and the control ($Z = 0$) group. This comparison is indeed based on randomly assigned Z and identifies a parameter known as *intention to treat effect*. The second one is to do IV using randomly assigned treatment intended as an instrumental variable for treatment received, which solves the sort of compliance problem previously discuss. In this case, LATE returns the effect of *treatment on the treated*, i.e., $E[Y(1) - Y(0)|D = 1]$.

Bibliography

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Lecture 6

Generalized Method of Moments and Empirical Likelihood

6.1 Generalized Method of Moments

6.1.1 Over-identified Linear Model

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. We assume that $\ell \geq k$, $E[ZU] = 0$, $E[ZX'] < \infty$, and $\text{rank}(E[ZX']) = k+1$. Assume further that the first component of X is constant and equal to one. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Using the fact that $U = Y - X'\beta$ and $E[ZU] = 0$, we see that β solves the system of equations

$$E[Z(Y - X'\beta)] = 0 \ .$$

Since $\ell + 1 \geq k + 1$, this may be an over-determined system of equations and today we will focus on the case where $\ell > k$. This situation is called over-identified. There are $\ell - k = r$ more moment restrictions than parameters to estimate. We usually call r the number of over-identifying restrictions.

The above is a special case of a more general class of moment condition models. Let $m(Y, X, Z, \beta)$ be an $\ell + 1$ dimensional function of a $k + 1$ dimensional parameter β such that

$$E[m(Y, X, Z, \beta)] = 0 \ . \tag{6.1}$$

In the linear model, $m(Y, X, Z, \beta) = Z(Y - X'\beta)$. In econometrics, this class of models are called moment condition models. In the statistics literature, these are known as estimating equations.

6.1.2 The GMM Estimator

Let (Y, X, Z, U) be distributed as described above and denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P. Define the sample analog of $E[m(Y, X, Z, \beta)]$ by

$$\bar{m}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} m_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\beta) = \frac{1}{n}\mathbb{Z}'(\mathbb{Y} - \mathbb{X}\beta) \ , \tag{6.2}$$

where in what follows we will use the notation $m_i(\beta) = m(Y_i, X_i, Z_i, \beta)$. The method of moments estimator for β is defined as the parameter value which sets $\bar{m}_n(\beta) = 0$. This is generally not possible when $\ell > k$ as there are more equations than free parameters. The idea of the generalized method of moments (GMM) is to define an estimator that sets $\bar{m}_n(\beta)$ "close" to zero, given a notion of "distance".

Let Λ_n be an $(\ell+1) \times (\ell+1)$ matrix such that $\Lambda_n \xrightarrow{P} \Lambda$ for a symmetric positive definite matrix Λ and define

$$Q_n(\beta) = n \ \bar{m}_n(\beta)'\Lambda_n\bar{m}_n(\beta) \ .$$

This is a non-negative measure of the "distance" between the vector $\bar{m}_n(\beta)$ and the origin. For example, if $\Lambda_n = \mathbb{I}$, then $Q_n(\beta) = n \ |\bar{m}_n(\beta)|^2$, the square of the Euclidean norm, scaled by the sample size n. The GMM estimator of β is defined as the value that minimizes $Q_n(\beta)$, this is

$$\hat{\beta}_n = \underset{b \in \mathbf{R}^{k+1}}{\operatorname{argmin}} \, Q_n(b) \ . \tag{6.3}$$

Note that if $k = \ell$, then $\bar{m}_n(\hat{\beta}_n) = 0$, and the GMM estimator is the method of moments estimator. The first order conditions for the GMM estimator are

$$\begin{aligned} 0 &= \frac{\partial}{\partial b} Q_n(\hat{\beta}) \\ &= 2\frac{\partial}{\partial b}\bar{m}_n(\hat{\beta}_n)'\Lambda_n\bar{m}_n(\hat{\beta}) \\ &= -2\left(\frac{1}{n}\mathbb{Z}'\mathbb{X}\right)'\Lambda_n\left(\frac{1}{n}\mathbb{Z}'(\mathbb{Y} - \mathbb{X}\hat{\beta}_n)\right) \end{aligned} \tag{6.4}$$

so

$$2\left(\mathbb{Z}'\mathbb{X}\right)'\Lambda_n\left(\mathbb{Z}'\mathbb{X}\right)\hat{\beta}_n = 2\left(\mathbb{Z}'\mathbb{X}\right)'\Lambda_n\left(\mathbb{Z}'\mathbb{Y}\right) \ , \tag{6.5}$$

which establishes a closed-form solution for the GMM estimator in the linear model,

$$\hat{\beta}_n = \left(\left(\mathbb{Z}'\mathbb{X}\right)'\Lambda_n\left(\mathbb{Z}'\mathbb{X}\right)\right)^{-1}\left(\mathbb{Z}'\mathbb{X}\right)'\Lambda_n\left(\mathbb{Z}'\mathbb{Y}\right) \ . \tag{6.6}$$

Without matrix notation, we can write this estimator as

$$\hat{\beta}_n = \left(\left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right)' \Lambda_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right)' \Lambda_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i Y_i \right) .$$

The matrix $(\mathbb{Z}'\mathbb{X})' \Lambda_n (\mathbb{Z}'\mathbb{X})$ may not be invertible for a given n, but, since $E[ZX']'\Lambda E[ZX']$ is invertible, it will be invertible with probability approaching one. Note that using similar arguments to those in Lemma 4.1 we can claim that $E[ZX']'\Lambda E[ZX']$ is invertible provided $E[ZX']$ has rank $k+1$ and Λ has rank $\ell+1$.

6.1.3 Consistency

Let

$$\Sigma = E[ZX'] .$$

Then, by the WLLN and the CMT

$$\left(\frac{1}{n} \mathbb{Z}'\mathbb{X} \right)' \Lambda_n \left(\frac{1}{n} \mathbb{Z}'\mathbb{X} \right) \xrightarrow{P} \Sigma'\Lambda\Sigma$$

and

$$\frac{1}{n} \mathbb{Z}'\mathbb{Y} \xrightarrow{P} E[ZY] = \Sigma\beta$$

as $n \to \infty$. The desired result therefore follows from the CMT.

6.1.4 Asymptotic Normality

In addition to the assumptions above, assume that

$$\Omega = E[ZZ'U^2]$$

is finite and invertible. Write

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - \beta) = & \left(\left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right)' \Lambda_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right) \right)^{-1} \\ & \times \left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right)' \Lambda_n \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i U_i \right) , \end{aligned}$$

and note that by the CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i U_i \xrightarrow{d} N(0, \Omega) .$$

Using this, the results on convergence in probability, and the CMT, we conclude that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

where

$$\mathbb{V} = (\Sigma'\Lambda\Sigma)^{-1}(\Sigma'\Lambda\Omega\Lambda\Sigma)(\Sigma'\Lambda\Sigma)^{-1} .$$

In general, GMM estimators are asymptotically normal with "sandwich form" asymptotic variances. The optimal weigh matrix Λ^* is the one which minimizes $\mathbb{V}$. This turn out to be $\Lambda^* = \Omega^{-1}$. The proof is left as an exercise. This yields the efficient GMM estimator

$$\hat{\beta}_n = \left(\left(\mathbb{Z}'\mathbb{X}\right)' \Omega^{-1} \left(\mathbb{Z}'\mathbb{X}\right) \right)^{-1} \left(\mathbb{Z}'\mathbb{X}\right)' \Omega^{-1} \left(\mathbb{Z}'\mathbb{Y}\right) ,$$

which satisfies

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, (\Sigma'\Omega^{-1}\Sigma)^{-1}) .$$

In practice, Ω is not known but it can be estimated consistently. For any $\hat{\Omega}_n \xrightarrow{P} \Omega$, we still call $\hat{\beta}_n$ the efficient GMM estimator, as it has the same asymptotic distribution. By "efficient", we mean that this estimator has the smallest asymptotic variance in the class of GMM estimators with this set of moment conditions. This is a weak concept of optimality, as we are only considering alternative weight matrices Λ_n. However, it turns out that the GMM estimator is semiparametrically efficient, as shown by Gary Chamberlain (1987). If it is known that $E[m_i(\beta)] = 0$ and this is all that is known, this is a semi-parametric problem, as the distribution of the data P is unknown. Chamberlain showed that in this context, if an estimator has this asymptotic variance, it is semiparametrically efficient. This result shows that no estimator has greater asymptotic efficiency than the efficient GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

6.1.5 Estimation of the Efficient Weighting Matrix

Given any weight matrix Λ_n with the properties previously discussed, the GMM estimator $\hat{\beta}_n$ is consistent yet *inefficient*. For example, we can set $\Lambda_n = \mathbb{I}$. In the linear model, under the additional assumption that Z has no perfect collinearity and that $E[ZZ'] < \infty$, a better choice is $\Lambda_n = (\mathbb{Z}'\mathbb{Z})^{-1}$, which leads to TSLS since

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

As before, $\mathbb{P}_Z = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$ is a projection matrix. Given any such first-step estimator, we can define residuals $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$ and moment equations $\hat{m}_i = Z_i\hat{U}_i$. Construct,

$$\hat{m}_i^* = \hat{m}_i - \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i .$$

Now define

$$\Lambda_n^* = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{m}_i^* \hat{m}_i^{*\prime} \right)^{-1} .$$

By using arguments similar to those in the second part of this class, it can be shown that $\Lambda_n^* \xrightarrow{P} \Omega^{-1}$ and GMM using the weighting matrix above is asymptotically efficient. This is typically referred to as the efficient two-step GMM estimator.

A common alternative choice is to use

$$\Lambda_n = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{m}_i \hat{m}_i' \right)^{-1} , \tag{6.7}$$

which uses the uncentered moment conditions. Since $E[m_i] = 0$ these two estimators are asymptotically equivalent under the hypothesis of correct specification. However, the uncentered estimator may be a poor choice when constructing hypothesis tests, as under the alternative hypothesis the moment conditions are violated, i.e. $E[m_i] \neq 0$.

6.1.6 Overidentification Test

Let

$$Q_n^*(\beta) = n \ \bar{m}_n(\beta)' \Lambda_n^* \bar{m}_n(\beta) \ .$$

If the moment condition model is correctly specified in the sense that there exist $\beta \in \mathbf{R}^{k+1}$ such that (6.1) holds, then it can be shown that

$$Q_n^*(\hat{\beta}_n) \xrightarrow{d} \chi^2_{\ell-k} \ , \tag{6.8}$$

as $n \to \infty$, where $\hat{\beta}_n$ is the efficient two-step GMM estimator. In addition, $Q_n^*(\hat{\beta}_n) \to \infty$ if $E[m(Y, X, Z, \beta)] \neq 0$ for all $\beta \in \mathbf{R}^{k+1}$. The proof of this result is left as an exercise. Note that the degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions. The overidentification test then rejects the null hypothesis "there exist $\beta \in \mathbf{R}^{k+1}$ such that (6.1) holds" when $Q_n^*(\hat{\beta}_n)$ exceeds the $1 - \alpha$ quantile of $\chi^2_{\ell-k}$.

6.2 Empirical Likelihood

Empirical Likelihood (EL) is a data-driven nonparametric method of estimation and inference for moment restriction models, which does not require weight matrix estimation like GMM and is invariant to nonsingular linear transformations of the moment conditions. It was introduced by Art B. Owen and later studied in depth by Qin and Lawless (1994), Imbens, Spady and Johnson (1998) and Kitamura (2001), among others. It is basically a non-parametric analog of Maximum Likelihood.

Consider the same setting as before, where

$$E[m(Y, X, Z, \beta)] = 0 \ . \tag{6.9}$$

Again, in the linear model $m(Y, X, Z, \beta) = Z(Y - X'\beta)$ and we still use the notation $m_i(\beta) = m(Y_i, X_i, Z_i, \beta)$.

Empirical likelihood may be viewed as parametric inference in moment condition models, using a data-determined parametric family of distributions. The parametric family is a multinomial distribution on the observed values $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$. This parametric family will have $n-1$ parameters. Having the number of parameters grow as quickly as the sample size makes empirical likelihood very different than parametric likelihood.

The multinomial distribution which places probability p_i at each observation of the data will satisfy the above moment condition if and only if

$$\sum_{i=1}^{n} p_i m_i(\beta) = 0 \ . \tag{6.10}$$

The empirical likelihood estimator is the value of β which maximizes the multinomial log-likelihood subject to the above restriction. This is, the empirical likelihood function is

$$\mathcal{R}_n(b) \equiv \max_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^{n} n p_i \,\middle|\, p_i > 0; \sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i m_i(b) = 0 \right\} \ . \tag{6.11}$$

The Lagrangian for the empirical log-likelihood is

$$\mathcal{L}(b, p_1, \ldots, p_n, \lambda, \kappa) = \sum_{i=1}^{n} \log(n p_i) - \kappa \left(\sum_{i=1}^{n} p_i - 1 \right) - n\lambda' \sum_{i=1}^{n} p_i m_i(b)$$

where κ and λ are Lagrange multipliers. For a given value $b \in \mathbf{R}^{k+1}$, the first order conditions with respect to p_i, κ and λ are:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_i} &= \frac{1}{p_i} - \kappa - n\lambda' m_i(b) = 0 \\
\frac{\partial \mathcal{L}}{\partial \kappa} &= \sum_{i=1}^{n} p_i - 1 = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= n \sum_{i=1}^{n} p_i m_i(b) = 0 \ .
\end{aligned}$$

Multiplying the first equation by p_i, summing over i; and using the second and third equations, we find $\kappa = n$ and

$$p_i(b) = \frac{1}{n} \frac{1}{1 + \lambda(b)' m_i(b)} \ , \tag{6.12}$$

where $\lambda(b)$ solves $g(\lambda) = 0$ and

$$g(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{m_i(b)}{1 + \lambda' m_i(b)} = 0 \ . \tag{6.13}$$

It follows that we can write $\log(\mathcal{R}_n(b))$ as

$$\log(\mathcal{R}_n(b)) = -\sum_{i=1}^{n} \log(1 + \lambda(b)' m_i(b)) \ .$$

The EL estimator of β is the value that maximizes $\log(\mathcal{R}_n(b))$,

$$\tilde{\beta}_n = \underset{b \in \mathbf{R}^{k+1}}{\text{argmax}} \ \log(\mathcal{R}_n(b)) \ .$$

The EL estimator of the lagranage multiplier λ is $\tilde{\lambda}_n = \lambda(\tilde{\beta}_n)$ and the EL probabilities are $\tilde{p}_i = p_i(\tilde{\beta}_n)$.

6.2.1 Asymptotic Properties and First Order Conditions

It turns out that the limit distribution of the EL estimator is the same as that of efficient GMM. This is,

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N(0, (\Sigma' \Omega^{-1} \Sigma)^{-1}) \ . \tag{6.14}$$

We are going to skip the proof of this result in this class. What is more interesting is to compare the first order conditions of the EL estimator to those of two-step GMM. In order to do so, let's re-write the first order condition of the EL estimator, paying specific attention to the linear model where $m_i(\beta) = Z_i(Y_i - X_i'\beta)$. Define

$$M_i(\beta) = -\frac{\partial}{\partial \beta'} m_i(\beta) = Z_i X_i' \ ,$$

and let

$$\begin{aligned} \Sigma(\beta) &= E\left[M_i(\beta)\right] = E[Z_i X_i'] \\ \Omega(\beta) &= E\left[m_i(\beta) m_i(\beta)'\right] = E\left[Z_i Z_i' U_i^2\right] . \end{aligned}$$

Note that $\Sigma(\beta)$ does not depend on β in the linear model. However, in non-linear models it does and so we keep the dependence on β throughout the remainder of the section. Denote the sample analogs of $\Sigma(\beta)$ and $\Omega(\beta)$ by

$$\begin{aligned} \hat{\Sigma}_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} M_i(\beta) \\ \hat{\Omega}_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} m_i(\beta) m_i(\beta)' \ . \end{aligned}$$

Using this notation, the EL estimators $(\tilde{\beta}_n, \tilde{\lambda}_n)$ jointly solve:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{m_i(\tilde{\beta}_n)}{1 + \tilde{\lambda}_n' m_i(\tilde{\beta}_n)} \tag{6.15}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{M_i(\tilde{\beta}_n)'\tilde{\lambda}_n}{1 + \tilde{\lambda}_n' m_i(\tilde{\beta}_n)} \ . \tag{6.16}$$

Note that since $1/(1+a) = 1 - a/(1+a)$, we can re-write (6.15) and solve for $\tilde{\lambda}_n$,

$$\tilde{\lambda}_n = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{m_i(\tilde{\beta}_n) m_i(\tilde{\beta}_n)'}{1 + \tilde{\lambda}_n' m_i(\tilde{\beta}_n)}\right]^{-1} \bar{m}_n(\tilde{\beta}_n) \ . \tag{6.17}$$

where $\bar{m}_n(\beta) = n^{-1}\sum_{1\le i\le n} m_i(\beta)$. By (6.16) and (6.17),

$$\left[\frac{1}{n}\sum_{i=1}^{n} \frac{M_i(\tilde{\beta}_n)}{1 + \tilde{\lambda}_n' m_i(\tilde{\beta}_n)}\right]' \left[\frac{1}{n}\sum_{i=1}^{n} \frac{m_i(\tilde{\beta}_n) m_i(\tilde{\beta}_n)'}{1 + \tilde{\lambda}_n' m_i(\tilde{\beta}_n)}\right]^{-1} \bar{m}_n(\tilde{\beta}_n) = 0 \ .$$

The equation above can be written as

$$\tilde{\Sigma}_n(\tilde{\beta}_n)'\tilde{\Omega}_n^{-1}(\tilde{\beta}_n)\bar{m}_n(\tilde{\beta}_n) = 0 \ ,$$

where by (6.12),

$$\tilde{\Sigma}_n(\beta) = \sum_{i=1}^{n} \tilde{p}_i M_i(\beta)$$

$$\tilde{\Omega}_n(\beta) = \sum_{i=1}^{n} \tilde{p}_i m_i(\beta) m_i(\beta)' \ .$$

We can now see that EL and GMM have very similar first order conditions. Recall from (6.4) that the first order condition for a GMM estimator is given by

$$\hat{\Sigma}_n(\hat{\beta}_n)'\hat{\Omega}_n^{-1}\bar{m}_n(\hat{\beta}_n) = 0 \ ,$$

where $\hat{\Omega}_n$ is a consistent estimator of Ω based on a preliminary estimator of β. It is not surprising then that these two estimators are first order equivalent. However, using a different estimator of the Jacobian matrix Σ and the absence of a preliminary estimator for Ω gives EL some favorable second order properties relative to GMM. The cost of this is some additional computational complexity. This is a topic we cover in Econ 481.

Bibliography

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Lecture 7

Panel Data[1]

Let (Y, X, η, U) be a random vector where Y, η, and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^k$. Note that here we are *not* assuming that the first component of X is a constant equal to one. Let $\beta = (\beta_1, \ldots, \beta_k)' \in \mathbf{R}^k$ be such that

$$Y = X'\beta + \eta + U \ ,$$

where we assume both η and U are unobserved. In addition, we want to allow for the possibility that X and η are correlated, so that $E[X\eta] \neq 0$. Given this, combining $\eta + U$ into a single unobservable would require an IV to get an estimator of β, even if we assume $E[XU] = 0$. Today we will see that when we observe the same units (individuals, firms, families, etc) multiple times (across time, regions, etc) we may identify and consistently estimate β without an IV, at least under certain restrictions on η and U.

Suppose that we observe the same unit at two different points in time, and that the unobservable η captures unobserved heterogeneity that is unit specific *and* constant over time. This is, consider the model

$$\begin{aligned} Y_1 &= X_1'\beta + \eta + U_1 \\ Y_2 &= X_2'\beta + \eta + U_2 \ . \end{aligned}$$

Note that we are also assuming that β is a constant parameter that does not change over time. If this is the case, we could simply take first differences, i.e.,

$$\begin{aligned} Y_2 - Y_1 &= (X_2 - X_1)'\beta + U_2 - U_1 \\ \Delta Y &= \Delta X'\beta + \Delta U \ , \end{aligned}$$

and remove the unobserved individual effect η in the process. Notice that

$$E[\Delta X \Delta U] = E[X_2U_2] + E[X_1U_1] - E[X_2U_1] - E[X_1U_2] \ . \tag{7.1}$$

[1]This lecture is based on Alex Torgovitsky's lecture notes. I want to thank him for kindly sharing them.

For the expression above to be equal to zero it is not enough to assume that $E[X_2U_2] = E[X_1U_1] = 0$, which would be the standard orthogonality assumption. We also need that $E[X_2U_1] = E[X_1U_2] = 0$, i.e., that the covariates in a given time period are uncorrelated with the unobservables in other time periods. This is called strict exogeneity. If this is the case, running least squares of ΔY on ΔX would deliver a consistent estimator of

$$\beta = E[\Delta X \Delta X']^{-1} E[\Delta X \Delta Y] \ , \tag{7.2}$$

provided that $E[\Delta X \Delta X']$ is invertible.

Before we proceed to formalize and extend some of these ideas, there are a few aspects that are worth keeping in mind. First, observing the same units over multiple time periods (the so-called *panel data*) allow us to control for unobserved factors that are constant over time (the η). The trick we just used would not work if η was allowed to *change over time*. Second, the requirement that $E[\Delta X \Delta X']$ is invertible means that we need X to *change over time*, so the trick we just used does not allow us to estimate coefficients of variables that are constant over time. Indeed, such variables are removed by the transformation in the same way η is removed. Finally, strict exogeneity is arguably stronger than simply assuming $E[X_tU_t] = 0$ for all t. Cases where X_2 is a decision variable of an agent in a context where U_1 is known at $t = 2$ may seriously question the validity of $E[X_2U_1] = 0$. Note that this type of dynamic argument is distinct from omitted variables bias in the sense that it could occur even if we were to argue that $E[X_tU_t] = 0$ or even $E[U_t|X_t] = 0$.

7.1 Fixed Effects

7.1.1 First Differences

Let (Y, X, η, U) be distributed as described above and denote by P the distribution of

$$(Y_{i,1}, \ldots, Y_{i,T}, X_{i,1}, \ldots, X_{i,T}) \ . \tag{7.3}$$

We assume that we have a random sample of size n, so that the observed data is given by $\{(Y_{i,t}, X_{i,t}) : 1 \leq i \leq n, 1 \leq t \leq T\}$. Note that while the sampling process is i.i.d. across i, we are being completely agnostic about the dependence across time for a given unit i. We then consider

$$Y_{i,t} = X_{i,t}'\beta + \eta_i + U_{i,t}, \ i = 1, \ldots, n \quad t = 1, \ldots, T \ , \tag{7.4}$$

under the assumptions on $X_{i,t}$ and $U_{i,t}$ that we formalize below. Now define

$$\Delta X_{i,t} = X_{i,t} - X_{i,t-1}$$

for $t \geq 2$, and proceed analogously with the other random variables. Note again that $\Delta\eta_i = 0$. Applying this transformation to (7.4), we get

$$\Delta Y_{i,t} = \Delta X_{i,t}'\beta + \Delta U_{i,t}, \ i = 1, \ldots, n \quad t = 2, \ldots, T \ . \tag{7.5}$$

It follows that a regression of $\Delta Y_{i,t}$ on $\Delta X_{i,t}$ provides a consistent estimator of β if the following two assumptions hold,

FD1. $E[U_{i,t}|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$,

FD2. $\sum_{t=2}^{T} E[\Delta X_{i,t}\Delta X_{i,t}'] < \infty$ is invertible.

FD1 is a sufficient condition for $E[\Delta U_{i,t}|\Delta X_{i,t}] = 0$. FD2 fails if some component of $X_{i,t}$ does not vary over time. The first-difference estimator then takes the form

$$\hat{\beta}_n^{\text{fd}} = \left(\sum_{1 \leq i \leq n} \sum_{2 \leq t \leq T} \Delta X_{i,t}\Delta X_{i,t}' \right)^{-1} \left(\sum_{1 \leq i \leq n} \sum_{2 \leq t \leq T} \Delta X_{i,t}\Delta Y_{i,t} \right) \ . \tag{7.6}$$

Under the assumption that $\text{Var}[U_{i,t}|X_{i,1}, \ldots, X_{i,T}]$ is constant (homoskedasticity), together with the assumption of no serial correlation in $U_{i,t}$, it is possible to show that $\hat{\beta}_n^{\text{fd}}$ is not asymptotically efficient and that a different transformation of the data delivers an estimator with a lower asymptotic variance under those assumption. We will discuss this further after describing this alternative transformation.

7.1.2 Deviations from Means

An alternative transformation to remove the individual effects η_i from (7.4) is the so-called de-meaning technique. In order to define this formally, let

$$\dot{X}_{i,t} = X_{i,t} - \bar{X}_i \quad \text{where} \quad \bar{X}_i = \frac{1}{T} \sum_{1 \leq t \leq T} X_{i,t} \ ,$$

and define $\dot{Y}_{i,t}$ and $\dot{U}_{i,t}$ analogously. Note that $\dot{\eta}_i = 0$ for all $i = 1, \ldots, n$. Applying this transformation to (7.4), we get

$$\dot{Y}_{i,t} = \dot{X}_{i,t}'\beta + \dot{U}_{i,t}, \quad i = 1, \ldots, n \quad t = 1, \ldots, T \ . \tag{7.7}$$

It follows that a regression of $\dot{Y}_{i,t}$ on $\dot{X}_{i,t}$ provides a consistent estimator of β if the following two assumptions hold,

FE1. $E[U_{i,t}|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$,

FE2. $\sum_{t=1}^{T} E[\dot{X}_{i,t}\dot{X}_{i,t}'] < \infty$ is invertible.

FE1, which is the same strict exogeneity condition in FD1, is a sufficient condition for $E[\dot{U}_{i,t}|\dot{X}_{i,t}] = 0$. As before, FE2 fails if some component of $X_{i,t}$ does not vary over time. The de-meaning estimator (commonly known as the fixed effect estimator) or dummy variable estimator takes the form

$$\hat{\beta}_n^{\text{fe}} = \left(\sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \right)^{-1} \left(\sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{Y}_{i,t} \right) . \tag{7.8}$$

Under the assumption that $\text{Var}[U_{i,t}|X_{i,1}, \ldots, X_{i,T}]$ is constant (homoskedasticity), together with the assumption of no serial correlation in $U_{i,t}$, it is possible to show that $\hat{\beta}_n^{\text{fe}}$ is asymptotically efficient. We discuss this in the next section.

7.1.3 Asymptotic Properties

Deriving an asymptotic approximation for estimators in panel data models involves two elements that were not present with cross-sectional data. First, the data is i.i.d. across i but may be dependent across time. This is, we may suspect that $X_{i,t}$ and $X_{i,s}$ for $t \neq s$ may not be independent. Second, the data has two indices now: the number of units (denoted by n) and the number of time periods (denoted by T). We will definitely need $nT \to \infty$ to get a useful asymptotic approximation, but we may achieve this by all sort of different assumptions about how n and/or T grow. The two standard approximations are $n \to \infty$ and T fixed (the so-called short panels) and $n \to \infty$ and $T \to \infty$ (the so-called large panels). Many commonly used panels in applied research include thousands of units (n large) and few time periods (T small) so we will focus on short panels first and discuss large panels later in class.

Under asymptotics where $n \to \infty$ and fixed T, we can show that $\hat{\beta}_n^{\text{fe}}$ and $\hat{\beta}_n^{\text{fd}}$ are asymptotically normal using similar arguments to those we use before, provided we assume

$$(Y_{i,1}, \ldots, Y_{i,T}, X_{i,1}, \ldots, X_{i,T}, U_{i,1}, \ldots, U_{i,T})$$

are i.i.d. across $i = 1, \ldots, n$. Start by writing

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}} - \beta) = \left(\frac{1}{n} \sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{U}_{i,t} \right) .$$

In order to make this expression more tractable, we use two tricks. First, note that

$$\sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{U}_{i,t} = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} - \bar{U}_i \sum_{1 \leq t \leq T} \dot{X}_{i,t} = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} , \tag{7.9}$$

where the last step follows from $\sum_{1\leq t\leq T}\dot{X}_{i,t}=0$. We can therefore replace $\dot{U}_{i,t}$ with $U_{i,t}$. Second, let $\dot{X}_i=(\dot{X}_{i,1},\ldots,\dot{X}_{i,T})'$ be a $T\times k$ vector of stacked observations for unit i, and define U_i in the same way. Using this notation, we can write

$$\dot{X}_i'\dot{X}_i=\sum_{1\leq t\leq T}\dot{X}_{i,t}\dot{X}_{i,t}' \quad \text{and} \quad \dot{X}_i'U_i=\sum_{1\leq t\leq T}\dot{X}_{i,t}U_i\ . \tag{7.10}$$

Combining (7.9) and (7.10), we obtain

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}}-\beta)=\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\dot{X}_i'U_i\right)\ .$$

By the law of large numbers and FE2,

$$\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\xrightarrow{P}\Sigma_{\dot{X}}\equiv E[\dot{X}_i'\dot{X}_i]=\sum_{1\leq t\leq T}E[\dot{X}_{i,t}\dot{X}_{i,t}']\ .$$

In addition, by the central limit theorem and FE1,

$$\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\dot{X}_i'U_i\xrightarrow{d}N(0,\Omega),\quad \text{where}\quad \Omega=\text{Var}[\dot{X}_i'U_i]=E[\dot{X}_i'U_iU_i'\dot{X}_i]\ .$$

Combining these results with the CMT we get

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}}-\beta)\xrightarrow{d}N(0,\mathbb{V}^{\text{fe}}) \tag{7.11}$$

where

$$\mathbb{V}^{\text{fe}}=\Sigma_{\dot{X}}^{-1}\Omega\Sigma_{\dot{X}}^{-1}\ . \tag{7.12}$$

Historically, researchers often assumed that $U_{i,t}$ was serially uncorrelated with variance independent of $X_{i,t}$ (i.e. homoskedastic). The default standard errors in Stata are still based on these assumptions. However, these assumptions are difficult to justify for most economic data, which is often strongly autocorrelated and heteroskedastic. One faces basically the same trade-off as with heteroskedasticity in the cross-sectional case. The most common strategy is to use the fully robust consistent estimator of the asymptotic variance,

$$\hat{\mathbb{V}}^{\text{fe}}=\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\hat{U}_i\hat{U}_i'\dot{X}_i\right)\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1}\ ,$$

where $\hat{U}_i=\dot{Y}_i-\dot{X}_i\hat{\beta}_n^{\text{fe}}$. This is what Stata computes when one uses the `luster(unit) } option to \verb+xtreg+ where unit is the variable that indexes $i$. This` consistent covariance matrix estimator that allows for arbitrary inter-temporal correlation patterns and heteroskedasticity across individuals. As we will see later in class, this estimator is generally known as a cluster covariance estimator (CCE) and is consistent as $n\to\infty$, i.e., $\hat{\mathbb{V}}^{\text{fe}}\xrightarrow{P}\mathbb{V}^{\text{fe}}$.

A comment on efficiency. Traditional arguments in favor of the fixed effects (or within-group) estimator $\hat{\beta}_n^{\text{fe}}$ over the first-difference estimator $\hat{\beta}_n^{\text{fd}}$ rely on the fact that under homoskedasticity and no-serial correlation of $U_{i,t}$, $\hat{\beta}_n^{\text{fe}}$ has a lower asymptotic variance than $\hat{\beta}_n^{\text{fd}}$. Intuitively, taking first differences introduces correlation in $\Delta U_{i,t}$ as

$$\begin{aligned} E[\Delta U_{i,t}\Delta U_{i,t-1}] &= E[U_{i,t}U_{i,t-1} - U_{i,t-1}U_{i,t-1} - U_{i,t}U_{i,t-2} + U_{i,t-1}U_{i,t-2}] \\ &= -\operatorname{Var}(U_{i,t-1}) \ . \end{aligned}$$

However, in the other extreme where $U_{i,t}$ follows a random walk, i.e., $U_{i,t} = U_{i,t-1} + V_{i,t}$ for some i.i.d. sequence $V_{i,t}$, then $\Delta U_{i,t} = V_{i,t}$. These results, at the end of the day, rely on homoskedasticity and so it is advised to simply use a robust standard error as above and forget about efficiency considerations. Note that when $T = 2$, these two estimators are numerically the same. In addition, first differences are used in dynamic panels and difference in differences, as we will discuss later.

Remark 7.1 Panel data traditionally deals with units over time. However, we can think about other cases where the data has a two-dimensional index and where we believe that one of the indices may exhibit within group dependence. For example, it could be that we observe "employees" within "firms", or "students" within "schools", or "families" in metropolitan statistical areas (MSA), etc. Cases like these are similar but not identical to panel data. To start, units are not "repeated" in the sense that each unit is potentially observed only once in the sample. In addition, these are cases where "T" is usually large and "n" is small. For example, we typically observe many students (which may be dependent within a school) and few schools. We will study these cases further in the second part of this class.

7.2 Random Effects

Fixed effects approaches are attractive to economists because they provide a way of addressing omitted variables bias and related forms of endogeneity, as long as the omitted factors are time constant. An alternative way to exploit the time dimension of the panel is to model the evolution of the unobservable term over time within a unit, and use this model to increase efficiency relative to ordinary pooled linear regressions. This is known as a random effects approach. Random effects are not as widely used as fixed effects in economics because they focus on efficiency rather than bias and robustness. Nevertheless, random effects approaches are occasionally used and also have some interesting connections to fixed effects and other types of panel data models.

The standard random effects model adds the following assumption to (7.4),

RE1. $E[\eta_i|X_{i,1},\ldots,X_{i,T}] = 0$.

Hence all of the unobservable time-invariant factors that were being controlled for in the fixed effects approach are now assumed to be mean independent (ergo, uncorrelated) with the explanatory variables at all time periods. The strict exogeneity condition of the fixed effects approach (i.e. FE1) is still maintained, so that the aggregate error term $V_{it} = \eta_i + U_{i,t}$ now satisfies $E[V_{it}|X_{i1},\ldots,X_{iT}] = 0$ for all $t = 1,\ldots,T$. The idea behind the random effects approach is to exploit the serial correlation in V_{it} that is generated by having a common η_i component in each time period. Specifically, the baseline approach maintains the following.

RE2. (i) $\text{Var}[U_{i,t}|X_{i,1},\ldots,X_{i,T}] = \sigma_U^2$, (ii) $\text{Var}[\eta_i|X_{i,1},\ldots,X_{i,T}] = \sigma_\eta^2$, (iii) $E[U_{i,t}U_{i,s}|X_{i,1},\ldots,X_{i,T}] = 0$ for all $t \neq s$, (iv) $E[U_{i,t}\eta_i|X_{i,1},\ldots,X_{i,T}] = 0$ for all $t = 1,\ldots,T$.

Under these assumptions,

$$\text{Var}[V_{i,t}|X_{i,1},\ldots,X_{i,T}] = E[\eta_i^2 + U_{i,t}^2 + \eta_i U_{i,t}|X_{i,1},\ldots,X_{i,T}] = \sigma_\eta^2 + \sigma_U^2 \ ,$$

and

$$E[V_{i,t}V_{i,s}|X_{i,1},\ldots,X_{i,T}] = E[\eta_i^2 + U_{i,t}U_{i,s} + \eta_i U_{i,t} + \eta_i U_{i,s}|X_{i,1},\ldots,X_{i,T}] = \sigma_\eta^2 \, .$$

Combining these results and stacking the observations for unit i, we get that

$$E[V_i V_i'|X_i] = \Omega = \sigma_U^2 \mathbb{I}_T + \sigma_\eta^2 \iota_T \iota_T' \ , \tag{7.13}$$

where $\mathbb{I}_T$ is the $T \times T$ identity matrix and ι_T is a T-dimensional vector of ones. Under these assumption, the estimator with the lowest asymptotic variance is

$$\hat{\beta}_n^{\text{re}} = \left(\sum_{1 \leq i \leq n} X_i' \Omega^{-1} X_i \right)^{-1} \left(\sum_{1 \leq i \leq n} X_i' \Omega^{-1} Y_i \right) \ , \tag{7.14}$$

where $X_i = (X_{i,1},\ldots,X_{i,T})'$ is the $T \times k$ vector of stacked observations for unit i, and similarly for Y_i. Note this is just a generalized least squares (GLS) estimator of β. This GLS estimator is, nevertheless, unfeasible, since Ω depends on the unknown parameters σ_U^2 and σ_η^2. However, these two can be easily estimated to form $\hat{\Omega}$ and deliver a feasible GLS estimator of β.

A few aspects are worth discussing. First, the efficiency gains hold under the additional structure imposed by RE1 and RE2. In particular, we are possibly gaining efficiency in a context where the unobserved heterogeneity η_i is assumed to be mean independent of X_i. In other words, unobserved time-invariant factors must be uncorrelated with observed covariates. This was precisely what made the fixed effects approach attractive in the first

place. Second, the efficiency gains hold under the homoskedasticity and independence assumptions in RE2 and do not hold more generally. These are undoubtedly strong assumptions. Third, unlike the fixed effects estimator, the random effects approach allows to estimate regression coefficients associated with time-invariant covariates (this is, some of the $X_{i,t}$ may be constant across time - i.e., gender of the individual). So if the analysis is primarily concerned with the effect of a time-invariant regressor and panel data is available, it makes sense to consider some sort of random effects type of approach. Fourth, under RE1 and RE2 β is identified in a single cross-section. The parameters that require panel data for identification in this model are the variances of the components of the error σ_η^2 and σ_U^2, which are needed for the GLS approach. Finally, note that the terminology "fixed effects" and "random effects" is arguably confusing as η_i is random in both approaches.

A last word of caution should be made about the use of Hausman specification tests. These are test that compare $\hat{\beta}_n^{\text{fe}}$ with $\hat{\beta}_n^{\text{re}}$ in order to test the validity of RE1 (assuming RE2 holds). Under the null hypothesis that RE1 holds, both estimators are consistent but $\hat{\beta}_n^{\text{re}}$ is efficient. Under the alternative hypothesis, $\hat{\beta}_n^{\text{fe}}$ is consistent while $\hat{\beta}_n^{\text{re}}$ is not. Now, suppose we were to define a new estimator $\hat{\beta}_n^*$ as follows

$$\hat{\beta}_n^* = \hat{\beta}_n^{\text{fe}} I\{\text{Hausman test rejects}\} + \hat{\beta}_n^{\text{re}} I\{\text{Hausman test accepts}\} \ . \quad (7.15)$$

The problem with this new estimator is that its finite sample distribution looks very different from the usual normal approximations. This is generally the case when there is pre-testing, understood as a situation where we conduct a test in a first step, and then depending on the outcome of this test, we do A or B in a second step. A formal analysis of these *uniformity* issues are covered in 481 and are beyond the scope of this class.

7.3 Dynamic Models

One benefit of panel data is that it allows us to analyze economic relationships that are inherently dynamic. Specifically, we may be interested in the effect of lagged outcomes on future outcomes. Let $\{Y_{i,t} : 1 \leq i \leq n, 1 \leq t \leq T\}$ be a sequence of random variables and consider the model

$$Y_{i,t} = \rho Y_{i,t-1} + \eta_i + U_{i,t}, \quad i = 1, \ldots, n \quad t = 2, \ldots, T \ , \quad (7.16)$$

where η_i and $U_{i,t}$ are the same as before but now $Y_{i,t-1}$ is allowed to have a direct effect on $Y_{i,t}$, a feature sometimes referred to as state dependence. We assume that $|\rho| < 1$. As is common in dynamic panel data (and time series) contexts, we will assume that the model is dynamically complete in the sense that all appropriate lags of $Y_{i,t}$ have been removed from the time-varying

error U_{it}, i.e.,

$$E[U_{i,t}|Y_{i,t-1}, Y_{i,t-2}, \dots] = 0 \text{ for all } t = 1, \dots, T. \tag{7.17}$$

Consider now taking first differences to (7.16) to obtain,

$$\Delta Y_{i,t} = \rho \Delta Y_{i,t-1} + \Delta U_{i,t}, \quad i = 1, \dots, n \quad t = 2, \dots, T \ ,$$

where, as before, $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$ and similarly for $U_{i,t}$. In general we will have $\text{Cov}(\Delta Y_{i,t-1}, \Delta U_{i,t}) \neq 0$ since (7.16) implies

$$\text{Cov}(Y_{i,t-1}, U_{i,t-1}) \neq 0 \ . \tag{7.18}$$

A similar conclusion would arise if we tried to use the de-meaning transformation. This inherent endogeneity is a generic feature of models that have both state dependence and time-invariant heterogeneity. In order to get rid of the fixed effects we have to compare outcomes over time, but if past outcomes have effects on future outcomes then differenced error terms will still be correlated with differenced outcomes.

The most commonly proposed solution to this problem is to use other lagged outcomes as instruments. Given (7.17), we know that $Y_{i,t-2}$ is uncorrelated with both $U_{i,t}$ and $U_{i,t-1}$, hence

$$\text{Cov}(Y_{i,t-2}, \Delta U_{i,t}) = 0 \ .$$

At the same time, we also know that

$$\begin{aligned} \text{Cov}(Y_{i,t-2}, \Delta Y_{i,t-1}) &= \text{Cov}(Y_{i,t-2}, Y_{i,t-1}) - \text{Cov}(Y_{i,t-2}, Y_{i,t-2}) \\ &= \text{Cov}(Y_{i,t-2}, \rho Y_{i,t-2} + \eta_i + U_{i,t-1}) - \text{Cov}(Y_{i,t-2}, Y_{i,t-2}) \\ &= -(1-\rho)\,\text{Var}[Y_{i,t-2}] + \text{Cov}(Y_{i,t-2}, \eta_i) \ , \end{aligned}$$

which makes $Y_{i,t-2}$ a valid instrument for $\Delta Y_{i,t-1}$ since we assumed $|\rho| < 1$ and $\text{Cov}(Y_{i,t-2}, \eta_i) \neq 0$. An actual expression for this last covariance can be obtained under additional assumptions. For example, under the assumption that the initial condition, $Y_{i,0}$, is independent of η_i (and $\eta_i \perp U_{i,t}$), then

$$\text{Cov}(Y_{i,t-2}, \eta_i) = \sigma_\eta^2 \sum_{j=0}^{t-3} \rho^j \ .$$

This strategy requires $T \geq 3$, since otherwise we would not have data on $Y_{i,t-2}$. For larger T we could include additional lags such as $Y_{i,t-3}, Y_{i,t-4}$, etc. Following such an approach delivers $(T-2)(T-1)/2$ linear moment restrictions of the form

$$E[Y_{i,t-k}(\Delta Y_{i,t} - \rho \Delta Y_{i,t-1})] = 0, \quad t = 3, \dots, T, \quad k = 2, \dots, t-1 \ . \tag{7.19}$$

The predictive power of these lags for $\Delta Y_{i,t-1}$ is likely to get progressively weaker as the lag distance gets larger. Weak instrument problems may arise as a consequence. If $T \geq 4$ then one could consider using the differenced term $\Delta Y_{i,t-2}$ (instead or in addition to the level $Y_{i,t-2}$) as an instrument for $\Delta Y_{i,t-1}$. In the literature, these approaches are frequently referred to as Arellano-Bond or Anderson-Hsiao estimators; see Arellano (2003).

Bibliography

Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Lecture 8

Difference in Differences

Today we will focus again on the problem of evaluating the impact of a program or treatment on a population outcome Y. As before, we will use potential outcomes to describe the problem,

$$\begin{array}{ll} Y(0) & \text{potential outcome in the absence of treatment} \\ Y(1) & \text{potential outcome in the presence of treatment} \end{array} . \tag{8.1}$$

The treatment effect is the difference $Y(1) - Y(0)$ and the usual quantity of interest is $E[Y(1) - Y(0)]$, typically referred to as the *average treatment effect*.

Suppose that we observe a random sample of n individuals from this population, and that for each individual i we observe both $Y_i(1)$ and $Y_i(0)$. Clearly, for each i we can compute the treatment effect $Y_i(1) - Y_i(0)$ and estimate the average treatment effect as

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0) \ .$$

This is, as we know, infeasible. Indeed, a large fraction of the work in econometric theory precisely deals with deriving methods that may recover the average treatment affect (or similar quantities) from observing $Y_i(1)$ for individuals receiving treatment and $Y_i(0)$ for individuals without treatment. The difference in differences (DD) approach is a popular method in this class that exploits grouped-level treatment assignments that vary over time. We start describing this method in the context of a simple two-groups two-periods example below.

8.1 A Simple Two by Two Case

To simplest setup to describe the DD approach is one where outcomes are observed for two groups for two time periods. One of the groups is exposed

to a treatment in the second period but not in the first period. The second groups is not exposed to the treatment during either period. To be specific, let

$$\{(Y_{j,t}, D_{j,t}) : j \in \{1,2\} \text{ and } t \in \{1,2\}\} \tag{8.2}$$

denote the observed data, where $Y_{j,t}$ and $D_{j,t} \in \{0,1\}$ denote the outcome and treatment status of group j at time t. Note that in our setup, $D_{j,t} = 1$ if and only if $j = 1$ and $t = 2$ (assuming the first group is the one receiving treatment in the second period). The parameter we will be able to identify is

$$\theta = E[Y_{1,2}(1) - Y_{1,2}(0)] \ , \tag{8.3}$$

which is simply the *average treatment effect on the treated*: the average effect of the treatment that occurs in group 1 in period 2. In order to interpret θ as an *average treatment effect*, one would need to make the additional assumption that

$$\theta = E[Y_{j,t}(1) - Y_{j,t}(0)] \tag{8.4}$$

is constant across j and t. This is a strong assumption and, in principle, not fundamental for the DD approach. The assumption in (8.4) has particular bite when we consider multiple treated groups. Consider the following example as an illustration.

Example 8.1 *On April 1, 1992, New Jersey raised the state minimum wage from* \$4.25 *to* \$5.05*. Card and Krueger (1994) collected data on employment at fast food restaurants in New Jersey in February 1992* ($t = 1$) *and again in November 1992* ($t = 2$) *to study the effect of increasing the minimum wage on employment. They also collected data from the same type of restaurants in eastern Pennsylvania, just across the river. The minimum wage in Pennsylvania stayed at* \$4.25 *throughout this period. In our notation, New Jersey would be the first group, $Y_{j,t}$ would be the employment rate in group j at time t, and $D_{j,t}$ denotes an increase in the minimum wage (the treatment) in group j at time t.*

The identification strategy of DD relies on the following assumption,

$$E[Y_{2,2}(0) - Y_{2,1}(0)] = E[Y_{1,2}(0) - Y_{1,1}(0)] \ , \tag{8.5}$$

i.e., both groups have "common trends" in the absence of a treatment. One way to parametrize this assumption is to assume that

$$Y_{j,t}(0) = \eta_j + \gamma_t + U_{j,t} \ , \tag{8.6}$$

where $E[U_{j,t}] = 0$, and η_j and γ_t are (non-random) group and time effects. This additive structure for non-treated potential outcomes implies that $E[Y_{j,2}(0) - Y_{j,1}(0)] = \gamma_2 - \gamma_1 \equiv \gamma$, which is constant across groups. Note that this assumption, together with (8.3) imply that

$$E[Y_{1,2}(1)] = \theta + \eta_1 + \gamma_2 \ . \tag{8.7}$$

In the context of the previous example, this assumption says that in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect, a year effect that is common across states, and a zero mean shock. Before we discuss the identifying power of this structure, we discuss two natural (but unsuccessful) approaches that may come to mind.

8.1.1 Pre and post comparison

A natural approach to identify θ in (8.4) would be to compare $Y_{1,2}$ and $Y_{1,1}$, this is, using outcomes before and after the policy change for the treated group alone. This approach delivers,

$$E[\Delta Y_{1,2}] = E[Y_{1,2}(1) - Y_{1,1}(0)] = \theta + \gamma \ ,$$

where $\Delta Y_{1,2} = Y_{1,2} - Y_{1,1}$ and $\gamma = \gamma_2 - \gamma_1$. Clearly, this approach does not identify θ in the presence of time trends, i.e., $\gamma \neq 0$. In the context of Example 8.1, the employment rate in New Jersey may have been going up (or down) in the absence of a policy change (the treatment), and so before and after comparisons confound the time trend as being part of the treatment effect. Unless one is willing to assume $\gamma = 0$, this approach does not identify θ.

8.1.2 Treatment and control comparison

A second natural approach to identify θ in (8.4) would be to compare $Y_{1,2}$ and $Y_{2,2}$, that is, using outcomes from both groups in the second time period. This approach delivers,

$$E[Y_{1,2} - Y_{2,2}] = E[Y_{1,2}(1) - Y_{2,2}(0)] = \theta - \eta \ ,$$

where $\eta = \eta_2 - \eta_1$. Clearly, this approach does not identify θ in the presence of persistent group differences, i.e., $\eta \neq 0$. In the context of Example 8.1, the employment rate in New Jersey and Pennsylvania may be idiosyncratically different in the absence of a policy change and so comparing these two states confound these permanent differences as being part of the treatment effect. Unless one is willing to assume $\eta = 0$, this approach does not identify θ.

8.1.3 Taking both differences

The DD approach exploits the common trends assumption in (8.5) to identify θ. The idea is to consider a second "difference" to remove γ (the time trend) from the difference that arises from comparing pre and post outcomes. In other words, the idea is to take the "difference" of the "differences", $\Delta Y_{1,2}$

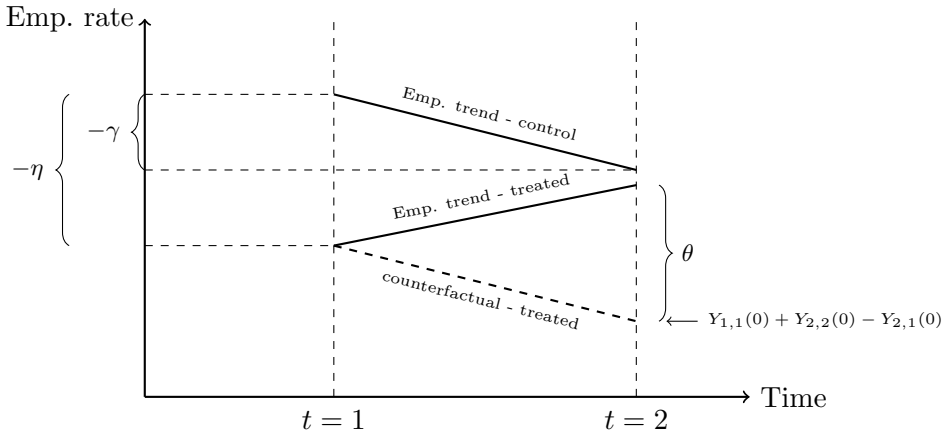

Figure 8.1: Causal effects in the DD model

and $\Delta Y_{2,2}$, to obtain

$$\begin{aligned} E[\Delta Y_{1,2} - \Delta Y_{2,2}] &= E[Y_{1,2}(1) - Y_{1,1}(0)] - E[Y_{2,2}(0) - Y_{2,1}(0)] \\ &= \theta + \gamma - \gamma = \theta \ . \end{aligned}$$

Thus, the approach identifies the treatment effect by taking the differences between pre-versus-post comparisons in the two groups, and exploiting the fact that the time trend γ is "common" in the two groups.

Note that an alternative interpretation to the same idea is to compare $(Y_{1,2} - Y_{2,2})$ and $(Y_{1,1} - Y_{2,1})$, this is, the treatment and control comparison before and after the policy change. This is because

$$\begin{aligned} E[(Y_{1,2} - Y_{2,2}) - (Y_{1,1} - Y_{2,1})] &= E[Y_{1,2}(1) - Y_{2,2}(0)] - E[Y_{1,1}(0) - Y_{2,1}(0)] \\ &= \theta - \eta + \eta = \theta \ . \end{aligned}$$

Using this representation, the difference for the pre-period is used to identify the persistent group difference η, a strategy that again works under the common trends assumption in (8.5).

A final interpretation of the same idea is that the DD approach construct a counterfactual potential outcome $Y_{1,2}(0)$ (which is unobserved) by combining $Y_{1,1}(0)$, $Y_{2,2}(0)$, and $Y_{2,1}(0)$, which are all observed. The "constructed" potential outcome is simply

$$\begin{aligned} \tilde{Y}_{1,2}(0) &= Y_{1,1}(0) + Y_{2,2}(0) - Y_{2,1}(0) \\ &= \eta_1 + \gamma_1 + \eta_2 + \gamma_2 - (\eta_2 + \gamma_1) + U_{1,1} + U_{2,2} - U_{2,1} \\ &= \eta_1 + \gamma_2 + \tilde{U}_{1,2} \ , \end{aligned}$$

where $\tilde{U}_{1,1} = U_{1,1} + U_{2,2} - U_{2,1}$. Computing $E[Y_{1,2} - \tilde{Y}_{1,2}(0)] = \theta$ therefore delivers a valid identification strategy. Figure 8.1 illustrates this idea.

8.1.4 A linear regression representation with individual data

Suppose that we observe

$$\{(Y_{i,j,t}, D_{j,t}) : i \in \mathcal{I}_{j,t}, j \in \{1,2\} \text{ and } t \in \{1,2\}\} \ , \tag{8.8}$$

where $\mathcal{I}_{j,t}$ is the set of individual in group j at time t. For simplicity, take the treatment indicator $D_{j,t} = I\{j = 1\}I\{t = 2\}$ to be non-random and note that the observed outcome is

$$Y_{i,j,t} = Y_{i,j,t}(1)D_{j,t} + (1 - D_{j,t})Y_{i,j,t}(0) = (Y_{i,j,t}(1) - Y_{i,j,t}(0))D_{j,t} + Y_{i,j,t}(0) \ ,$$

so that if we define $U_{i,j,t} = Y_{i,j,t} - E[Y_{i,j,t}]$, we can write

$$\begin{aligned} Y_{i,j,t} &= \theta D_{j,t} + \eta_j + \gamma_t + U_{i,j,t} \\ &= \eta_1 + \gamma_1 + \theta D_{j,t} + \eta_j - \eta_1 + \gamma_t - \gamma_1 + U_{i,j,t} \\ &= \delta + \theta D_{j,t} + \eta I\{j = 2\} + \gamma I\{t = 2\} + U_{i,j,t} \ , \end{aligned} \tag{8.9}$$

where $\delta = \eta_1 + \gamma_1$. Thus, we can estimate θ by running a regression of $Y_{i,j,t}$ on $(1, D_{j,t}, I\{j = 2\}, I\{t = 2\})$ and extracting the coefficient on $D_{j,t}$. The regression formulation of the DD model offers a convenient way to construct DD estimates and standard errors. It also makes it easy to add additional groups and time periods to the regression setup. We might, for example, add additional control groups and pre-treatment periods. The resulting generalization thus includes a dummy for each state and period but is otherwise unchanged.

8.2 A More General Case

Now consider the case with many groups and many time periods (and no individual data for now). The derivation in (8.9) suggests that the natural regression to consider would be

$$Y_{j,t} = \theta D_{j,t} + \eta_j + \gamma_t + U_{j,t} \ \text{ with } \ E[U_{j,t}] = 0 \ . \tag{8.10}$$

Here, the observed data is given by $\{(Y_{j,t}, D_{j,t}) : j \in \mathcal{J}_0 \cup \mathcal{J}_1, t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$, where $Y_{j,t}$ is the outcome of unit j at time t, $D_{j,t}$ is the (non-random) treatment status of unit j at time t, $\mathcal{T}_0$ is the set of pre-treatment time periods, $\mathcal{T}_1$ is the set of post-treatment time periods, $\mathcal{J}_0$ is the set of controls units, and $\mathcal{J}_1$ is the set of treatment units. The scalar random variables η_j, γ_t and $U_{j,t}$ are unobserved and $\theta \in \Theta \subseteq \mathbf{R}$ is the parameter of interest. The regression in (8.10) is known as the two-way fixed effect regression.

Define

$$\Delta_{n,j} = \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} Y_{j,t} - \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} Y_{j,t} \ , \tag{8.11}$$

and

$$\hat{\theta}_n = \frac{1}{|\mathcal{J}_1|}\sum_{j\in\mathcal{J}_1}\Delta_{n,j} - \frac{1}{|\mathcal{J}_0|}\sum_{j\in\mathcal{J}_0}\Delta_{n,j} \ . \tag{8.12}$$

It is easy to show that $\hat{\theta}_n$ is the LS estimator of a regression of $Y_{j,t}$ on $D_{j,t}$ with groups fixed effects (η_j) and time fixed effects (γ_t). i.e., the regression in (8.10). Simple algebra shows that

$$\begin{aligned}\hat{\theta}_n - \theta \ &= \ \frac{1}{|\mathcal{J}_1|}\sum_{j\in\mathcal{J}_1}\left(\frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1}U_{j,t} - \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}U_{j,t}\right) \\ &\qquad -\frac{1}{|\mathcal{J}_0|}\sum_{j\in\mathcal{J}_0}\left(\frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1}U_{j,t} - \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}U_{j,t}\right) \ .\end{aligned}$$

It follows immediately from $E[U_{j,t}] = 0$ that $E[\hat{\theta}_n] = \theta$. This estimator is also consistent and asymptotically normal in an asymptotic framework with a large number of treated and untreated groups, i.e., $|\mathcal{J}_1| \to \infty$ and $|\mathcal{J}_0| \to \infty$. Other asymptotic approximations may lead to substantially different results and we will discuss some of these in the second part of this class.

The parameter θ could be interpreted as the ATE under assumption (8.4), or as the ATT under the assumption that

$$E[Y_{j,t}(1) - Y_{j,t}(0)] \ , \tag{8.13}$$

is constant for all $j \in \mathcal{J}_1$ and $t \in \mathcal{T}_1$. Alternatively, one could estimate a different θ_j for each $j \in \mathcal{J}_1$. In general, it has recently been show that the regression in (8.10) may produce misleading estimates (i.e., $\hat{\theta}_n$ inconsistent for the ATT), if the policy's effect is heterogeneous between groups or over time, as is often the case in empirical settings. A special case where it would be consistent for the ATT under the parallel trends assumption alone is when (i) the design is staggered, meaning that groups' treatment can only increase over time and can change at most once; (ii) the treatment is binary; (iii) there is no variation in treatment timing: all treated groups start receiving the treatment at the same date. However, conditions (i)-(iii) are seldom met in practice. See De Chaisemartin and D'Haultfoeuille (2022), and references therein, for details on these issues.

8.2.1 Thinking ahead: inference and few treated groups

Inference in DD could be tricky and requires thinking. Two issues are of particular importance. First, what exactly is assumed to be "large"? Are groups going to infinity? Say, $|\mathcal{J}_1| \to \infty$ and $|\mathcal{J}_0| \to \infty$. What happens if we have a few treated groups but many controls? Say, $|\mathcal{J}_1|$ fixed and $|\mathcal{J}_0| \to \infty$.

What happens if we have few treated and control groups but many time periods? Say, $|\mathcal{J}_1|$ and $|\mathcal{J}_0|$ fixed, but $|\mathcal{T}_1| \to \infty$ and $|\mathcal{T}_0| \to \infty$. Second, what are the assumptions on $U_{j,t}$? It is typically common to assume that $U_{j,t} \perp U_{j',s}$ for all $j' \neq j$ and (t,s). However, one would expect $U_{j,t}$ and $U_{j,s}$ to be correlated, at least for t and s being "close" to each other. On top of this, in the context of individual data one would expect $U_{i,j,t}$ to be correlated with $U_{i',j,s}$ - i.e., units in the same group may be dependent to each other even if they are in different time periods. Each of these aspects have tremendous impact on which inference tools end up being valid or not. We will discuss some of these in the second part of this class.

As a way to illustrate how important these assumptions may be, let's consider the case where $\mathcal{J}_1 = \{1\}$ but $|\mathcal{J}_0| \to \infty$ - we also assume that $|\mathcal{T}_0|$ and $|\mathcal{T}_1|$ are finite. This is, only the first group is treated, while there are many control groups. This is common in empirical applications with US state level data, where often a few states exhibit a policy change while all the other states do not. The DD estimator in this case reduced to

$$\begin{aligned}
\hat{\theta}_n &= \Delta_{n,1} - \frac{1}{|\mathcal{J}_0|}\sum_{j\in\mathcal{J}_0}\Delta_{n,j} \ , \\
&= \theta + \frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1}U_{1,t} - \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}U_{1,t} - \frac{1}{|\mathcal{J}_0|}\sum_{j\in\mathcal{J}_0}\left(\frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1}U_{j,t} - \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}U_{j,t}\right) \\
&\xrightarrow{P} \theta + \frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1}U_{1,t} - \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0}U_{1,t} \ ,
\end{aligned}$$

as $|\mathcal{J}_0| \to 0$, assuming $\{U_{j,t} : t \in \mathcal{T}_0 \cup \mathcal{T}_1\}$ is i.i.d. across $j \in \mathcal{J}_0$. We conclude that the DD estimator is not even consistent for θ. Interestingly enough, it is still possible to do inference on θ using the approach proposed by Conley and Taber (2011) or, more recently, the randomization approach in Canay et al. (2017).

8.3 Synthetic Controls

Empirical applications with one or few treated groups and many control groups are ubiquitous in economics. The DD approach as described above in essence treats all control groups as being of equal quality as a control group. This may not be true and so the researcher may want to somehow weight the controls in order to give more importance to those controls that seem "better" for the given treated group. This is basically the idea of the synthetic control method, originally proposed by Abadie et al. (2010). The application in their paper is the effect of California's tobacco control program on state-wide smoking rates. During the time period in question,

there were 38 states in the US that did not implement such programs. Rather than just using a standard DD analysis - which effectively treats each state as being of equal quality as a control group - ADH propose choosing a weighted average of the potential controls. Of course, choosing a suitable control group or groups is often done informally, including matching on pre-treatment predictors. ADH formalize the procedure by optimally choosing weights, and they propose methods for inference.

Consider the simple case in Section 8.1.2, except that we now assume there are $\mathcal{J}_0$ possible controls and that $\mathcal{J}_1 = \{1\}$. Synthetic controls also allow the model for potential outcomes to be more flexible, specially when it comes to the parallel trends assumption required for DD. To be concrete, in what follows assume that

$$Y_{j,t} = \eta_j \gamma_t + U_{j,t} \ , \tag{8.14}$$

so that now the time effect and the group effect interact with each other (note that common trends does not hold in this case). Comparing $Y_{1,2}$ and $Y_{j,2}$ for any $j \in \mathcal{J}_0$ delivers

$$E[Y_{1,2} - Y_{j,2}] = E[Y_{1,2}(1) - Y_{j,2}(0)] = \theta + \gamma_2(\eta_1 - \eta_j) \ ,$$

and so this approach does not identify θ in the presence of persistent group differences. The idea behind synthetic controls is to construct the so-called *synthetic control*

$$\tilde{Y}_{1,2}(0) = \sum_{j \in \mathcal{J}_0} w_j Y_{j,2} \ ,$$

by appropriately choosing the weights $\{w_j : j \in \mathcal{J}_0, w_j \geq 0, \sum_{j \in \mathcal{J}_0} w_j = 1\}$. In order for this idea to work, it must be the case that $E[Y_{1,2}(0)] = E[\tilde{Y}_{1,2}(0)]$ so that $E\left[Y_{1,2} - \tilde{Y}_{1,2}(0)\right] = \theta$. Now, for a given set of weights, this approach delivers

$$E\left[Y_{1,2} - \tilde{Y}_{1,2}(0)\right] = E\left[Y_{1,2} - \sum_{j \in \mathcal{J}_0} w_j Y_{j,2}\right] = \theta + \gamma_2 \left(\eta_1 - \sum_{j \in \mathcal{J}_0} w_j \eta_j\right) \ .$$

It follows this approach identifies θ if we could choose the weights in a way such that

$$\eta_1 = \sum_{j \in \mathcal{J}_0} w_j \eta_j \ . \tag{8.15}$$

This is, however, not feasible as we do not observe the group effects η_j. The main result in Abadie et al. (2010) can be stated for the example in this section as follows: suppose that there exists weights $\{w_j^* : j \in \mathcal{J}_0, w_j^* \geq 0, \sum_{j \in \mathcal{J}_0} w_j^* = 1\}$ such that

$$Y_{1,1} = \sum_{j \in \mathcal{J}_0} w_j^* Y_{j,1} \ . \tag{8.16}$$

If we construct the synthetic control using these optimal weights w_j^*,

$$\tilde{Y}_{1,2}(0) = \sum_{j \in \mathcal{J}_0} w_j^* Y_{j,2} \ ,$$

then it follows that $E\left[Y_{1,2} - \tilde{Y}_{1,2}(0)\right] = \theta$.

Proving this result in the context of our example is straightforward. First, note that by (8.16) we get that

$$\eta_1 \gamma_1 + U_{1,1} = \sum_{j \in \mathcal{J}_0} w_j^* \eta_j \gamma_1 + \sum_{j \in \mathcal{J}_0} w_j^* U_{j,1} \ ,$$

so that

$$\gamma_1 \left(\eta_1 - \sum_{j \in \mathcal{J}_0} w_j^* \eta_j \right) = \sum_{j \in \mathcal{J}_0} w_j^* (U_{1,1} - U_{j,1}) \ . \tag{8.17}$$

Next note that

$$\begin{aligned} Y_{1,2} - \tilde{Y}_{1,2}(0) &= \theta + \eta_1 \gamma_2 + U_{1,2} - \sum_{j \in \mathcal{J}_0} w_j^* (\eta_j \gamma_2 + U_{j,2}) \\ &= \theta + \gamma_2 \left(\eta_1 - \sum_{j \in \mathcal{J}_0} w_j^* \eta_j \right) + \sum_{j \in \mathcal{J}_0} w_j^* (U_{1,2} - U_{j,2}) \\ &= \theta + \frac{\gamma_2}{\gamma_1} \sum_{j \in \mathcal{J}_0} w_j^* (U_{1,1} - U_{j,1}) + \sum_{j \in \mathcal{J}_0} w_j^* (U_{1,2} - U_{j,2}) \ , \end{aligned}$$

where we used (8.17) in the third equality. The result follows from $E[U_{j,t}] = 0$ for all (j,t).

We then get the weights by "matching" the observed outcomes of the treated group and the control groups in the period before the policy change. In practice, $Y_{1,1}$ may not lie in the convex hull of $\{Y_{j,1} : j \in \mathcal{J}_0\}$ and so the method relies on minimizing the distance between $Y_{1,1}$ and $\sum_{j \in \mathcal{J}_0} w_j Y_{j,1}$. Abadie et al. (2010) provide some formal arguments around these issues, and in particular require that $|\mathcal{T}_0| \to \infty$ and that $U_{j,t}$ is independent across j and t. However, the model they consider is slightly more general than the standard DD model, as it does not require the "common trends" assumption.

The basic idea can be extended in the presence of covariates X_j that are not (or would not be) affected by the policy change. In this case, the weights would be chosen to minimize the distance between

$$(Y_{1,1}, X_1) \text{ and } \sum_{j \in \mathcal{J}_0} w_j (Y_{j,1}, X_j) \ .$$

The optimal weights - which differ depending on how we define distance - produce the synthetic control whose pre-intervention outcome and predictors

of post-intervention outcome are "closest". Abadie et al. (2010) propose permutation methods for inference. We will discuss permutation tests in the second part of this class. This method has become popular in recent years and you will probably see it used in applied papers.

8.4 Discussion

To keep the exposition simple we have ignored covariates. However, it is straightforward to incorporate additional covariates under the assumption that potential outcomes are linear in those covariates, i.e.,

$$E[Y_{j,t}(0)|X_{j,t}] = \eta_j + \gamma_t + X'_{j,t}\beta \ .$$

This would simply entail adding $X'_{j,t}\beta$ to the regression in (8.10).

It is important to keep in mind that all the results on DD follow from the assumption that

$$E[Y_{j,t}(0)] = \eta_j + \gamma_t \ ,$$

which is a way to model the "common trends" assumption in (8.5). Where there are multiple time periods, people will often look at the pre (and post) treatment trends and compare them between treatment and control as a way to "eye-ball" verify this assumption. An unpleasant feature of this assumption is that is not robust to nonlinear transformations of the outcome variables. In other words, the assumption that

$$E[Y_{2,2}(0) - Y_{2,1}(0)] = E[Y_{1,2}(0) - Y_{1,1}(0)] \ ,$$

does not imply, for example, that

$$E[\log Y_{2,2}(0) - \log Y_{2,1}(0)] = E[\log Y_{1,2}(0) - \log Y_{1,1}(0)] \ .$$

Indeed, the two assumptions are non-nested and one would typically suspect that both cannot hold at the same time.

Bibliography

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, 105, 493–505.

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017): "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 85, 1013–1030.

Card, D. and A. B. Krueger (1994): "Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 772–793.

Conley, T. G. and C. R. Taber (2011): "Inference with "difference in differences" with a small number of policy changes," *The Review of Economics and Statistics*, 93, 113–125.

De Chaisemartin, C. and X. D'Haultfoeuille (2022): "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," Tech. rep., National Bureau of Economic Research.

Part II

Some Topics

Lecture 9

Non-Parametric Regression

9.1 Setup

Let (Y, X) be a random vector where Y and X take values in $\mathbf{R}$ and let P be the distribution of (Y, X). The case where $X \in \mathbf{R}^k$ will be discussed later. We are interested in the conditional mean of Y given X:

$$m(x) = E[Y|X = x] \ .$$

Let $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ be an i.i.d. sample from P. We first consider the discrete case. If X takes ℓ values $\{x_1, x_2, \ldots, x_\ell\}$, then

$$\hat{m}(x) = \frac{\sum_{i=1}^n I\{X_i = x\}Y_i}{\sum_{i=1}^n I\{X_i = x\}}$$

is a natural estimator of $m(x)$ for $x \in \{x_1, x_2, \ldots, x_\ell\}$. It is straightforward to show that $\hat{m}(x)$ is consistent and asymptotically normal if $E[Y^2] < \infty$.

9.2 Nearest Neighbor vs. Binned Estimator

Suppose now that X is a continuous random variable. In this case, the event $\{X_i = x\}$ has zero probability, so that the previous estimator will be undefined for almost every value of x. However, if $m(x)$ is continuous, we can use observations that are "close" to x to estimate $m(x)$. This motivates the following estimator:

Definition 9.1 (q-Nearest Neighbor Estimator) *Let $J_q(x)$ be the set of indices in $\{1, \ldots, n\}$ associated with q closest-to-x values of $\{X_1, \ldots, X_n\}$. The q-nearest neighbor estimator is defined as*

$$\hat{m}_{\text{nn}}(x) = \frac{1}{q} \sum_{i \in J_q(x)} Y_i \ .$$

Note that if $q = n$, we are using all of the observations in the estimation of each point. Then $\hat{m}_{\text{nn}}(x)$ just becomes $\bar{Y}_n$, producing a perfectly flat estimated function. The variance is very low but unless $m_{\text{nn}}(x)$ is truly flat, bias will be high for many values of x. Alternatively, we can use the X_i that is closest to x. In this case, bias should be relatively small, but since so few observations are used, variance is high. Generally, picking q is a problem. One way to do this is via cross validation – to be discussed later.

The q-NN estimator takes an average of the q observations closest to x, and so the number of "local" observations is always q. However, this means that the distance between these observations and x is random. In particular,

$$h = \max_{i \in J_q(x)} |X_i - x|$$

is random. As an alternative to the q-NN method, we can fix an h and consider all observations with $|X_i - x| \leq h$. Now, it is the number of local observations that is random. This gives rise to the binned estimator:

Definition 9.2 (Binned Estimator) *Let $h > 0$ be given. The binned estimator is defined as*

$$\hat{m}_{\text{b}}(x) = \frac{\sum_{i=1}^n I\{|X_i - x| \leq h\} Y_i}{\sum_{i=1}^n I\{|X_i - x| \leq h\}} \ . \tag{9.1}$$

The above formula can be interpreted as a weighted average,

$$\hat{m}_{\text{b}}(x) = \sum_{i=1}^n w_i(x) Y_i \quad \text{with} \quad w_i(x) = \frac{I\{|X_i - x| \leq h\}}{\sum_{i=1}^n I\{|X_i - x| \leq h\}}$$

and $\sum_{i=1}^n w_i(x) = 1$. Just as the choice of q mattered for the q-NN estimator, the choice of h will be important for the binned estimator.

9.3 Nadaraya-Watson Kernel Estimator

One deficiency of the binned estimator is that it is discontinuous at $x = X_i \pm h$. This occurs because the weights used are based on indicator functions but in principle one could use some other (continuous) weights. The family of weights typically used in non-parametric estimation are called "kernels". Our goal is to obtain continuous estimates $\hat{m}(x)$ by using continuous kernels.

Definition 9.3 (2nd order, Non-negative, Symmetric Kernel) *A second-order kernel function $k(u) : \mathbf{R} \to \mathbf{R}$ satisfies*

1. $\int_{-\infty}^{\infty} k(u) du = 1$

2. $0 \leq k(u) < \infty$

3. $k(u) = k(-u)$

4. $\kappa_2 = \int_{-\infty}^{\infty} u^2 k(u) du \in (0, \infty)$

Note that the definition of the kernel does not involve continuity. Indeed, the binned estimator can be written in terms of a kernel function. To see this, let

$$k_0(u) = \frac{1}{2} I\{|u| \leq 1\}$$

be the uniform density on $[-1, 1]$. Observe that

$$I\{|X_i - x| \leq h\} = I\left\{\frac{|X_i - x|}{h} \leq 1\right\} = 2k_0\left(\frac{X_i - x}{h}\right)$$

so that we can write $\hat{m}_{\mathrm{b}}(x)$ in (9.1) as

$$\hat{m}(x) = \frac{\sum_{i=1}^n k_0\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n k_0\left(\frac{X_i - x}{h}\right)} .$$

This is a special case of the so-called Nadaraya-Watson estimator.

Definition 9.4 (Nadaraya-Watson Kernel Estimator) *Let $k(u)$ be a second-order kernel and $h > 0$ be a bandwidth. Then, the Nadaraya-Watson estimator is defined as*

$$\hat{m}(x) = \frac{\sum_{i=1}^n k\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n k\left(\frac{X_i - x}{h}\right)} .$$

The Nadaraya-Watson estimator is also known as the kernel regression estimator or the local constant estimator. The bandwidth $h > 0$ plays the same role as before. In particular, the larger the h, the smoother the estimates (but the higher the bias):

$$h \to \infty \Rightarrow \hat{m}(x) \to \bar{Y}_n .$$

The smaller the h, the more erratic the estimates (but the lower the bias):

$$h \to 0 \Rightarrow \hat{m}(X_i) \to Y_i .$$

Some popular continuous kernels include the Gaussian kernel,

$$k_g(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) ,$$

and the Epanechnikov kernel,

$$k_e(u) = \frac{3}{4}(1 - u^2) I\{|u| \leq 1\} .$$

9.3.1 Asymptotic Properties

We will use the asymptotic framework in which $n \to \infty$, $h \to 0$, $nh \to \infty$, and $h = O(n^{-1/5})$. We wish to show that for each x,

$$\sqrt{nh}(\hat{m}(x) - m(x)) = \sqrt{nh}\Delta_1(x) + \sqrt{nh}\Delta_2(x)$$

where $\sqrt{nh}\Delta_2(x)$ converges to a limit that is asymptotically normal and centered at zero, and $\sqrt{nh}\Delta_1(x)$ converges to an asymptotic bias term. The rate of convergence is $\sqrt{nh}$ since this reflects the "effective" number of observations that we are using.

Start by writing $Y_i = m(X_i) + U_i$ so that $E[U_i|X_i] = 0$, and let

$$\sigma^2(x) = \mathrm{Var}[U_i|X_i = x] \ .$$

Fix $x \in \mathbf{R}$ and write

$$Y_i = m(x) + (m(X_i) - m(x)) + U_i \ .$$

Then we can rewrite the numerator of $\hat{m}(x)$ as

$$\begin{aligned}
\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) Y_i &= \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) m(x) \\
&+ \quad \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) (m(X_i) - m(x)) \\
&+ \quad \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) U_i \\
&= \hat{f}(x)m(x) + \hat{\Delta}_1(x) + \hat{\Delta}_2(x) \ ,
\end{aligned}$$

where $\hat{f}(x)$ is the non-parametric density estimator of the pdf of X, $f(x)$. It follows that

$$\hat{m}(x) - m(x) = \frac{1}{\hat{f}(x)}\left(\hat{\Delta}_1(x) + \hat{\Delta}_2(x)\right) \ .$$

First consider $\hat{\Delta}_2(x)$ and derive its mean and variance. Since $E[U_i|X_i] = 0$,

$$E[\hat{\Delta}_2(x)] = \frac{1}{h}E\left[k\left(\frac{X_i - x}{h}\right) U_i\right] = 0 \ .$$

The variance can be expressed as

$$\begin{aligned}
\mathrm{Var}\left[\hat{\Delta}_2(x)\right] &= \frac{1}{nh^2}E\left[\left(k\left(\frac{X_i - x}{h}\right) U_i\right)^2\right] \\
&= \frac{1}{nh^2}E\left[k\left(\frac{X_i - x}{h}\right)^2 \sigma^2(X_i)\right] \\
&= \frac{1}{nh^2}\int_{-\infty}^{\infty} k\left(\frac{z - x}{h}\right)^2 \sigma^2(z)f(z)dz
\end{aligned}$$

We now simplify the expression by the change of variables $u = \frac{1}{h}(z - x)$. This will require z to be in the interior of the support of X. Using the proposed substitution yields

$$\frac{1}{nh}\int_{-\infty}^{\infty} k\left(u\right)^2 \sigma^2(x+hu)f(x+hu)du = \frac{\sigma^2(x)f(x)}{nh}\int_{-\infty}^{\infty} k\left(u\right)^2 du + o\left(\frac{1}{nh}\right) ,$$

assuming $f(x)$ is continuously differentiable at x and $\sigma^2(x)$ is continuous at x. Let

$$R(k) = \int_{-\infty}^{\infty} k\left(u\right)^2 du$$

denote the so-called roughness of the kernel. By the above derivation,

$$\mathrm{Var}\left[\hat{\Delta}_2(x)\right] = \frac{\sigma^2(x)f(x)R(k)}{nh} + o\left(\frac{1}{nh}\right) ,$$

and so by a triangular array CLT,

$$\sqrt{nh}\hat{\Delta}_2(x) \xrightarrow{d} N(0, \sigma^2(x)f(x)R(k)) .$$

Second consider $\hat{\Delta}_1(x)$ and derive its mean and variance. Start with the mean,

$$\begin{aligned} E[\hat{\Delta}_1(x)] &= \frac{1}{h}E\left[k\left(\frac{X_i - x}{h}\right)\left(m(X_i) - m(x)\right)\right] \\ &= \frac{1}{h}\int_{-\infty}^{\infty} k\left(\frac{z-x}{h}\right)\left(m(z) - m(x)\right)f(z)dz \\ &= \int_{-\infty}^{\infty} k\left(u\right)\left(m(x+hu) - m(x)\right)f(x+hu)du . \end{aligned}$$

Assuming twice continuous differentiability of $m(x)$ (together with $f(x)$ continuously differentiable), expand $m(x)$ and $f(x)$ up to $o(h^2)$, i.e.,

$$\begin{aligned} m(x+hu) - m(x) &= m'(x)hu + \frac{1}{2}m''(x)h^2u^2 + o(h^2) \\ f(x+hu) &= f(x) + f'(x)hu + o(h) . \end{aligned}$$

Plug into the previous integral to obtain,

$$\begin{aligned} &\int_{-\infty}^{\infty} k\left(u\right)\left(m'(x)hu + \frac{h^2u^2}{2}m''(x)\right)\left(f(x) + uhf'(x)\right)du + o(h^2) \\ &\quad = \left(\int_{-\infty}^{\infty} uk\left(u\right)du\right)m'(x)f(x)h \\ &\quad + \left(\int_{-\infty}^{\infty} u^2k\left(u\right)du\right)h^2\left(\frac{1}{2}m''(x)f(x) + m'(x)f'(x)\right) + o(h^2) . \end{aligned}$$

Let κ_2 be defined as

$$\kappa_2 = \int_{-\infty}^{\infty} u^2 k(u)\, du$$

and

$$B(x) = \left(\frac{1}{2} m''(x) + f^{-1}(x) m'(x) f'(x) \right) .$$

Using this notation and the symmetry of the kernel, we can write

$$E[\hat{\Delta}_1(x)] = \kappa_2 h^2 f(x) B(x) + o(h^2) .$$

A similar expansion shows that

$$\text{Var}\left[\hat{\Delta}_1(x)\right] = O\left(\frac{h^2}{nh}\right) = o\left(\frac{1}{nh}\right) .$$

Again, by a triangular array CLT,

$$\sqrt{nh}(\hat{\Delta}_1(x) - h^2 \kappa_2 f(x) B(x)) \xrightarrow{d} 0 .$$

Putting all the pieces together and using the fact that $\hat{f}(x) \xrightarrow{P} f(x)$, we have our theorem

Theorem 9.1 (Asymptotic Normality) *Suppose that*

1. *$f(x)$ is continuously differentiable at the interior point x with $f(x) > 0$.*
2. *$m(x)$ is twice continuously differentiable at x.*
3. *$\sigma^2(x) > 0$ is continuous at x.*
4. *$k(x)$ is a non-negative, symmetric, 2nd order kernel.*
5. *$E[|Y|^{2+\delta}] < \infty$ for some $\delta > 0$.*
6. *$n \to \infty$, $h \to 0$, $nh \to \infty$, and $h = O(n^{-1/5})$.*

It follows that

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2 \kappa_2 B(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x) R(k)}{f(x)}\right) .$$

From the theorem, we also have that the asymptotic mean squared error of the NW estimator is

$$MSE(x) = h^4 \kappa_2^2 B^2(x) + \frac{\sigma^2(x) R(k)}{nh f(x)} .$$

The optimal rate of h which minimises the asymptotic MSE is therefore $Cn^{-1/5}$, where C is a function of $(\kappa_2, B(x), \sigma^2(x), R(k), f(x))$. In this case, the MSE converges at rate $O(n^{-4/5})$, which is the same as the rate obtained in density estimation. It is possible to estimate C, for instance by plug-in approaches. However, this is cumbersome and other methods, such as cross validation may be easier.

Kernel Choice. The asymptotic distribution of our estimator depends on the kernel through $R(k)$ and κ_2. An optimal kernel would therefore minimize $R(k)$. It turns out that the Epanechnikov family is optimal for regression, as with density estimation.

Bandwidth Choice. The constant C for the optimal bandwidth depends on the first and second derivatives of the mean function $m(x)$. When the derivative function $B(x)$ is large, the optimal bandwidth is small. When the derivative is small, the optimal bandwidth is large. There exists reference bandwidths for nonparametric density estimation (like Silverman's rule-of-thumb) but in nonparametric regression these are less natural.

Bias and Undersmoothing. Note that the bias term needs to be estimated to obtain valid confidence intervals. However, $B(x)$ depends on $m'(x), m''(x), f'(x)$ and $f(x)$. Estimating these objects is arguably more complicated than the problem we started out with. A (proper) residual bootstrap could be used to obtain valid confidence interval.

Alternatively, we can undersmooth. Undersmoothing is about choosing h such that

$$\sqrt{nh}h^2 \to 0 \ ,$$

which makes the bias small, i.e.,

$$\sqrt{nh}h^2\kappa_2 B(x) \approx 0 \ .$$

This eliminates the asymptotic bias but requires h to be smaller than optimal, since optimality requires that

$$nhh^4 \to C > 0 \ .$$

Such an h will also be incompatible with bandwidth choice methods like cross validation. Further, undersmoothing does not work well in finite samples. Better methods exist, though they are outside the scope of the course.

Curse of Dimensionality. Now consider the problem of estimating $m : \mathbf{R}^{d_x} \to \mathbf{R}$. where $d_x > 1$ is the dimension of X. The multivariate NW estimator is implemented in a way similar to the one we just described, except that we now require a multivariate kernel and d_x bandwidths. However, the rate of convergence of the NW estimator becomes

$$\sqrt{nh_1 \dots h_{d_X}} \quad \text{or} \quad \sqrt{nh^{d_x}}$$

depending on whether or not we use the same bandwidth for each component of X. This is the curse of dimensionality: the higher d_x, the slower the rate of convergence. Intuitively, with higher dimensions, it becomes harder to find "effective" observations. In this case, optimal bandwidths and MSE are

$$h = O(n^{\frac{-1}{4+d_x}}) \text{ and } MSE = O(n^{\frac{-4}{4+d_x}}) \ .$$

Linear conditional mean. The NW estimator may not perform well when $m(x)$ is linear, that is when $m(x) = \beta_0 + \beta_1 x$. In particular, it performs poorly if the marginal distribution of X_i is not roughly uniform. Suppose $Y_i = \beta_0 + \beta_1 X_i$ so that there is no error in the model. The NW estimator, when applied to this data generated by this purely linear model, yields a nonlinear output.

Boundaries of the Support. The NW estimator performs poorly on the boundaries of the support of X. For points on the boundary, bias is of order $O(h)$. For x s.t. $x \leq \min\{X_1, \ldots, X_n\}$, the NW estimator is an average only of Y_i values for observations to the right of x. If $m(x)$ is positively sloped, the NW estimator will be upward biased. Our change of variable argument no longer applies and the estimator is inconsistent at the boundary.

9.4 Local Linear Estimator

The Nadaraya-Watson estimator is often called a local constant estimator because it locally (about x) approximates the CEF $m(x)$ as a constant. To see this, note that $\hat{m}(x)$ solves the minimization problem:

$$\hat{m}(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)(Y_i - c)^2 \ , \tag{9.2}$$

which is a weighted regression Y_i on an intercept only. Without the weights, the estimation problem will have the sample mean as the solution. The NW estimator generalizes this to a "local" mean. This suggests that we can construct alternative nonparametric estimators of $m(x)$ by using other local approximations.

A popular choice is the local linear (LL) approximation. Instead of approximating $m(x)$ locally as a constant, the LL approximation approximates $m(x)$ locally by a linear function. We will do this by locally weighted least squares.

Definition 9.5 (Local Linear (LL) Estimator) *For each x, solve the following minimization problem,*

$$\{\hat{\beta}_0(x), \hat{\beta}_1(x)\} = \underset{(b_0, b_1)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)(Y_i - b_0 - b_1(X_i - x))^2 \ . \tag{9.3}$$

The local linear estimator of $m(x)$ is the local intercept: $\hat{\beta}_0(x)$.

The LL estimator of the derivative of $m(x)$ is the estimated slope coefficient:

$$\hat{m}'(x) = \hat{\beta}_1(x) \ .$$

If we write the local model

$$Y_i = \beta_0 + \beta_1(X_i - x) + U_i \quad \text{with} \quad E[U|X = x] = 0 \ ,$$

then taking conditional expectations, we see how using the regressor $X_i - x$ rather than X_i makes the intercept equal to $m(x) = E[Y|X = x]$.

To obtain the least squares formula, set, for each x,

$$Z_i(x) = (1, X_i - x)'$$

and

$$k_i(x) = k\left(\frac{X_i - x}{h}\right) \ .$$

Then

$$\begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix} = \left(\sum_{i=1}^{n} k_i(x) Z_i(x) Z_i(x)'\right)^{-1} \sum_{i=1}^{n} k_i(x) Z_i(x) Y_i \ , \tag{9.4}$$

so that for each x, the estimator is just weighted least squares of Y in $Z(x)$. In fact, as $h \to \infty$, the LL estimator approaches the full-sample linear least-squares estimator

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \ .$$

This is because as $h \to \infty$, all observations receive equal weight regardless of x. The LL estimator is thus a flexible generalization of least squares.

Deriving the asymptotic distribution of the LL estimator is similar to that of the NW estimator, but much more involved. We will skip that here.

Theorem 9.2 (Asymptotic Normality) *Let $\hat{m}(x)$ be the LL estimator as previously defined. Under conditions 1-6 in the NW theorem,*

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2\kappa_2 \frac{1}{2} m''(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x) R(k)}{f(x)}\right) \ .$$

Relative to the bias of the NW estimator,

$$B(x) = \left(\frac{1}{2} m''(x) + f^{-1}(x) m'(x) f'(x)\right) \ ,$$

the second term is no longer present. This simplified expression suggests reduced bias, though in theory, bias could be larger as opposing terms could cancel out. Because the bias of LL estimator does not depend of $f(x)$, we say that it is design adaptive. Furthermore, for the LL estimator to be consistent and asymptotically normal only continuity of $f(x)$ is required, not differentiability. As such, relative to the NW estimator, we can relax condition 1.

9.4.1 Nadaraya-Watson vs Local Linear Estimator

In contrast to the NW estimator, the LL estimator preserves linear data. In particular, if $Y_i = \beta_0 + \beta_1 X_i$, then for any sub-sample, a local linear regression fits exactly, so that $\hat{m}(x) = m(x)$.

Furthermore, the distribution of the LL estimator is invariant to the first derivative of m. As such, it has zero bias when the true regression is linear.

In addition, the LL estimator has better properties at the boundary than the NW estimator. Intuitively, the local linear estimator fits a (weighted) least-squares line through data near the boundary. As such, even if x is at the boundary of the regression support, this estimator will be unbiased as long as the true relationship is linear. More generally, the LL estimator has bias of order $O(h^2)$ at all x.

Extensions that allow for discontinuities in $m(x)$, $f(x)$ and $\sigma(x)$ exist.

9.5 Related Methods

There are several other non-parametric methods we did not cover. For example, the *Splines Estimator* is the unique minimizer of

$$\sum_{i=1}^{n} (Y_i - \hat{m}(X_i))^2 + \lambda \int (\hat{m}''(u))^2 du \ .$$

One advantage of spline estimators over kernels is that global inequality and equality constraints can be imposed more conveniently. *Series Estimators* are of the form

$$\hat{m}(x) = \sum_{j=0}^{\tau_n} \hat{\beta}_j \varphi_j(x) \ .$$

They are typically very easy to compute. However, there is relatively little theory about how to select the basis functions $\varphi(x)$ and the smoothing parameters τ_n.

Bibliography

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Lecture 10

Regression Discontinuity and Matching

(These notes will be revised before this class takes place)

Today we study evaluation methods to measure treatment effects of a policy or intervention on an outcome Y. To do this, we define potential outcomes as usual where

$$Y(0) \text{ potential outcome in the absence of treatment}$$
$$Y(1) \text{ potential outcome in the presence of treatment.}$$

Let $D \in \{0, 1\}$ denote treatment assignment status. The treatment effect is $Y(1) - Y(0)$ and a parameter of interest could be the average treatment effect (ATE) $E[Y(1) - Y(0)]$ or the average treatment on the treated (ATT) $E[Y(1) - Y(0)|D = 1]$, to list a few.

The fundamental difficulty of evaluation arises from the fact that we cannot simultaneously observe $Y(1)$ and $Y(0)$, but only observe either of these per each individual. Identification of the treatment effect may fail if treated individuals from whom we observe $Y(1)$ are systemically different from non-treated individuals from whom we observe $Y(0)$. In order to circumvent this, evaluation methods construct counterfactuals in a convincing way, dealing with endogenous selection. Popular approaches include (1) randomized controlled experiments (or RCTs) which exploits controlled/randomized assignment rules, (2) natural experiments which takes advantage of some "natural" randomization as difference-in-differences does and (3) instrumental variables and control function methods which rely on exclusion restrictions or models for an assignment rule. Today we introduce two additional methods: (4) discontinuity design methods which exploit discreteness in the treatment assignment rule and (5) matching methods which attempt to reproduce the treatment group among the non-treated using information of observed covariates.

10.1 Regression Discontinuity Design

The regression discontinuity designs (RDD) are characterized by a triplet: score, cutoff, and treatment. Suppose that units receive a score and a treatment is assigned based on the score and a known cutoff. Specifically, the treatment is given to units whose score is above the cutoff and it is withheld from units whose score is below the cutoff. For example, we can think of a situation where a scholarship is given to students whose grade in the SAT exceeds 2100. The abrupt change in the probability of treatment assignment allows us to learn something about the effect of treatment.

To be more precise, let us introduce some notation. An observed random variable $Z \in \mathbf{R}$ denotes the score (it is so-called a running variable). A known constant c denotes the cutoff. We normalize the cutoff c to 0 without loss of generality. Then we can represent the treatment assignment by

$$D_i = I\{Z_i \geq 0\}$$

and the observed outcome Y_i by

$$Y_i = \begin{cases} Y_i(0) & \text{if } Z_i < 0 \\ Y_i(1) & \text{if } Z_i \geq 0 \end{cases} .$$

Given this notation, the conditional expected outcome is

$$E[Y \mid Z = z] = \begin{cases} E[Y(0) \mid Z = z] & \text{if } z < 0 \\ E[Y(1) \mid Z = z] & \text{if } z \geqslant 0 \end{cases} \tag{10.1}$$

and the idea would be to exploit the discontinuity in $E[Y \mid Z = z]$ at the cutoff to identify some type of treatment effect.

10.1.1 Identification

The RDD allows us to estimate a specific type of treatment effect known as the ATE at the cutoff, $E[Y(1) - Y(0) \mid Z = 0]$. Note that this is an average effect for those individuals with scores exactly at the cutoff of the running variable. However, $Y(0)$ at the cutoff is not observed by design, so we need some assumptions.

Note that a special situation occurs at the cutoff $Z = 0$, as illustrated by Figure 10.1 where we plot the conditional mean function in (10.1). Consider two groups of units: one with score equal to 0, and the other with score barely below 0, say $Z = -\varepsilon$. If the value of $E[Y(0)|Z = -\varepsilon]$ are not abruptly different from $E[Y(0)|Z = 0]$, then units with $Z = -\varepsilon$ would be a valid counterfactural to units with $Z = 0$. Putting it formally, if the conditional mean function $E[Y(0)|Z = z]$ is continuous at $z = 0$, then the ATE at the cutoff, denoted by θ_{srd}, can be identified as follows

$$\theta_{\text{srd}} = E[Y(1) - Y(0) \mid Z = 0] = E[Y \mid Z = 0] - \lim_{z \uparrow 0} E[Y \mid Z = z] .$$

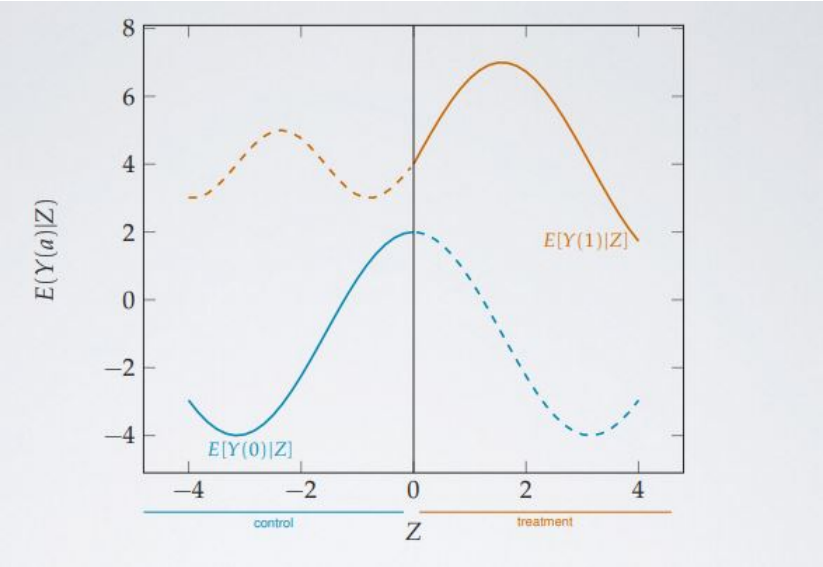

Figure 10.1: Graphs of means of potential outcomes conditional on the score Z

It is worth highlighting that the parameter θ_{srd} is a "local" ATE in the sense that it measures an ATE at a specific point of Z, cutoff. It is similar to the LATE parameter discussed under IV, but not necessarily the same.

Essentially, RDD exploit the discontinuous dependence of D on Z such that $P\{D = 1 \mid Z = z\}$ is discontinuous at $z = 0$. In the so-called *sharp design*, there is a perfect compliance in the sense that every unit with score above 0 receives treatment and every unit with score below 0 is in the control group. This creates a discontinuity in $P\{D = 1 \mid Z = z\}$ as the cutoff,

$$P\{D = 1 \mid Z = z\} = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } z \geqslant 0 \end{cases}.$$

In the so-called fuzzy design, there may be imperfect compliance.

10.1.2 Estimation via Local Linear Regression

In order to estimate θ_{srd}, we can construct nonparametric estimators of $E[Y|Z = 0]$ and $\lim_{z\to 0-} E[Y|Z = z]$ using data to the right and left of 0. Local linear (LL) estimators are predominantly used because the point of interest $E[Y(1) - Y(0) \mid Z = 0]$ is always on the boundary. As addressed in the last class, LL regression estimators performs better at the boundary than Nadaraya-Watson estimators in that the bias of LL regression estimators at the boundary is of order h^2 whereas that of Nadaraya-Watson estimators is of order h given a bandwidth h.

In addition, LL regression is simple to implement in RDD: we can obtain the LL regression estimator first by computing kernel weights based on $c = 0$ and then by running a weighted least squares regression on observations either above or below zero. Especially with an uniform kernel, LL estimators are the same as two unweighted linear regressions on observations with $Z_i \in [-h, 0)$ and $Z_i \in [0, h]$. Specifically, the LL regression estimator of $E[Y|Z = 0]$ is given as

$$\left\{\hat{\beta}_0^+, \hat{\beta}_1^+\right\} = \underset{\left(b_0^+, b_1^+\right)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{Z_i}{h}\right) I\left\{Z_i \geqslant 0\right\} \left(Y_i - b_0^+ - b_1^+ Z_i\right)^2$$

and the LL regression estimator of $\lim_{z \to 0-} E[Y|Z = z]$ is given as

$$\left\{\hat{\beta}_0^-, \hat{\beta}_1^-\right\} = \underset{\left(b_0^-, b_1^-\right)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{Z_i}{h}\right) I\left\{Z_i < 0\right\} \left(Y_i - b_0^- - b_1^- Z_i\right)^2.$$

Note that the regressor is $(Z_i - c)$ but we are assuming $c = 0$. Givne this, we can estimate θ_{srd} by

$$\hat{\theta}_{\text{srd}} = \hat{\beta}_0^+ - \hat{\beta}_0^-.$$

Figure 10.2 depicts estimating θ_{srd} using LL regression estimators. The blue and red curves are conditional mean functions in (10.1) on $(z, y) \in \mathbf{R}^2$. The dots around the curves represent a sample $\{(Z_i, Y_i) : i = 1, \cdots, n\}$. LL regression estimators use only the observations around the cutoff c within the window $[-h, h]$. We obtain the black line on the left by running a weighted least squares regression using observations below the cutoff and in the window (blue dots). Its intercept $\hat{\beta}_0^-$ is our estimate of $E[Y(0)|Z = 0]$. Similarly, we we obtain $\hat{\beta}_0^+$, an estimate of $E[Y(1)|Z = 0]$.

10.1.3 Bandwidth Choice

Running LL regression requires to pick a bandwidth h. Choosing the bandwidth is not straightforward due to a trade-off: heuristically, the bias increases and the variance decreases as the bandwidth h increases. Figure 10.3 shows that the bias, indicated as the gap between $E[Y(0)|Z = z]$ and the intercept $\hat{\beta}_0^-$ (or between $E[Y(1)|Z = z]$ and the intercept $\hat{\beta}_0^+$), increases as h changes from h_1 to h_2.

Taking this trade-off into account, Imbens and Kalyanaraman (2012) propose an "optimal" plug-in bandwidth

$$\hat{h}_{\text{IK}} = \hat{\text{C}}_{\text{IK}} \cdot n^{-1/5}.$$

Calonico et al. (2014) improve this result and suggest

$$\hat{h}_{\text{CCT}} = \hat{C}_{\text{CCT}} \cdot n^{-1/5}.$$

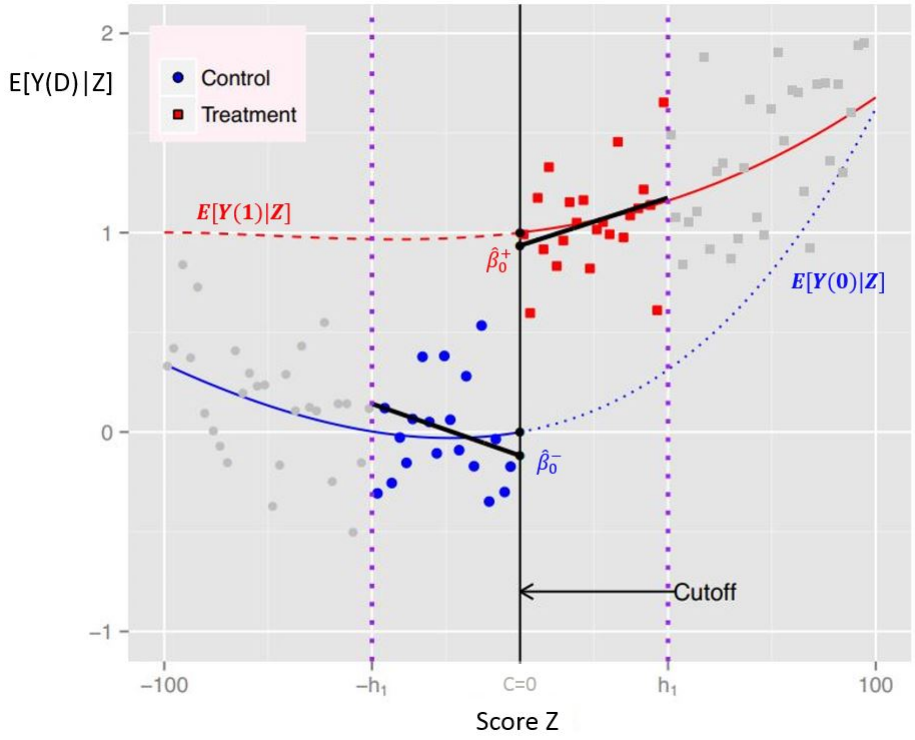

Figure 10.2: Estimation of θ_{srd} via LL regression estimators

In addition to $\hat{h}_{\text{CCT}}$, they also propose bias correction methods and new variance estimators that account for the additional noise introduced by estimating bias. While it is common to see papers based on undersmoothing, i.e., use $uh^5 \to 0$ and ignore asymptotic bias, yet using $\hat{h}_{\text{CCT}}$ is a better approach.

10.1.4 Other RD Designs

So far we have focused on RDD with a single running variable and a single cutoff where $P\{D = 1|Z = z\}$ is either 0 or 1. There are other designs generalizing these simple RDD. Their inference methods use similar tools (LL regresion, etc) but are different.

Sharp RD (SRD) and Fuzzy RD (FRD) While sharp RDD are characterized by perfect compliance, fuzzy RDD allow partial compliance which arises if some units with running variable above c decide not to receive treatment. For example, people may not cast a vote even if they are older than 18 and eligible for voting. Such partial compliance induces a discontinuity in $P\{D = 1|Z = z\}$ at c, but it does not necessarily change from 0 to 1.

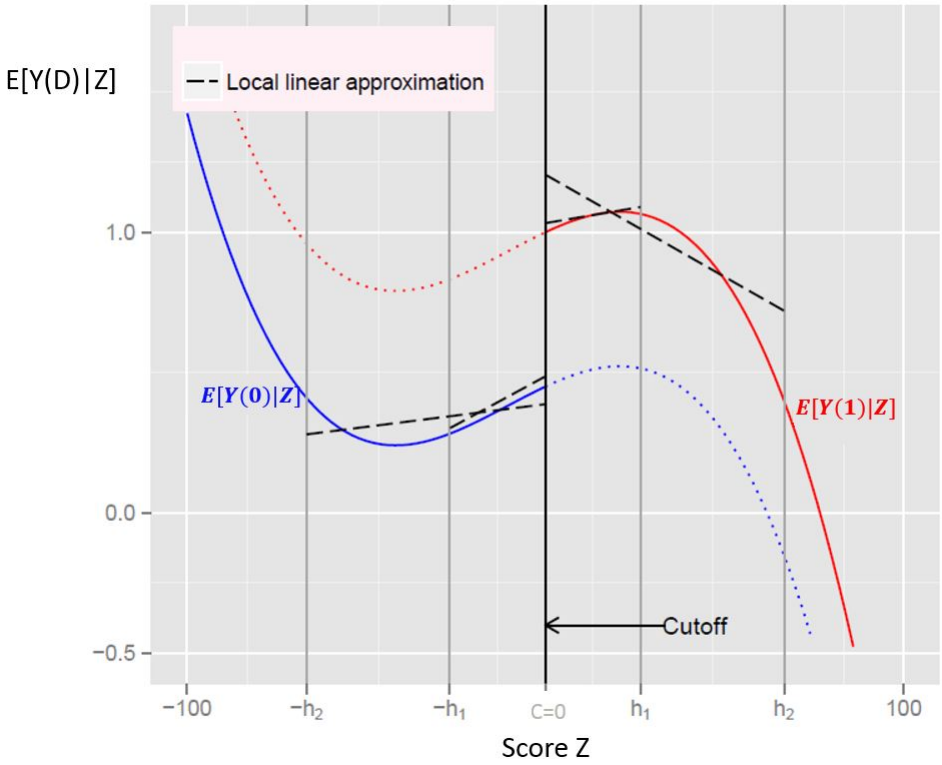

Figure 10.3: Biases of $E[Y(d)|Z=0]$ in Local Linear Regression for Varying Bandwidth

Kink RD (KRD) and Kink Fuzzy RD (KFRD) KRD/KFRD designs assume that $P\{D=1|Z=z\}$ is continuous but have a kink at the cutoff c which in turn introduces kinks of $E[Y|Z=z]$ at c. Conceptually, KRD and KFRD are similar to SRD and FRD except that they exploit discontinuity in the first derivatives of $P\{D=1|Z=z\}$ and $E[Y|Z=z]$.

Multiple scores RD and Geographic RD These deigns involve at least two running variables and discontinuities arise in $\mathbf{R}^2$ (or higher dimensional space). As an example of multiple scores RD, we can think of a scholarship which is awarded to a student whose math as well as English scores are above certain numbers. Similarly, geographic RD uses location information represented by latitude and longitude as their running variables and boundary between two different regions as a cutoff.

Multiple cutoff RD This involves multiple treatments that are given at multiple cutoffs.

10.1.5 Extension to Fuzzy RD

Imperfect compliance often occurs in real applications. Some units with score above c may decide not to take up treatment. Other units with score below c sometimes manage to receive treatment. For instance, having a score Z larger than c makes the application "strong" but may not guarantee a scholarship.

Under imperfect compliance, the probability of receiving treatment changes at c, but not necessarily from 0 to 1. This allows for identification of another local treatment effect. The argument is similar to LATE, but a little subtler due to limits. To formalize, define potential treatments as usual:

$D(0)$ potential receipt of treatment when the treatment is assigned

$D(1)$ potential receipt of treatment when the treatment is not assigned.

The treatment status is $D = D(1)I\{Z \geq c\} + D(0)I\{Z < c\}$. If $E[Y(d)|Z = z]$ and $E[D(d)|Z = z]$ for $d = 0, 1$ are continuous in z at c, then the canonical parameter is identified,

$$\begin{aligned}\theta_{\text{frd}} &= \frac{E\left[Y_i(1) \mid Z_i = c\right] - E\left[Y_i(0) \mid Z_i = c\right]}{E\left[D_i(1) \mid Z_i = c\right] - E\left[D_i(0) \mid Z_i = c\right]} \\ &= \frac{\lim_{z \downarrow c} E\left[Y_i \mid Z_i = z\right] - \lim_{z \uparrow c} E\left[Y_i \mid Z_i = z\right]}{\lim_{z \downarrow c} E\left[D_i \mid Z_i = z\right] - \lim_{z \uparrow c} E\left[D_i \mid Z_i = z\right]}.\end{aligned}$$

The parameter θ_{frd} can be interpreted as the ATE for units with $Z_i = c$ and only for compliers who are affected by the cutoff and satisfy $D_i(1) > D_i(0)$.

We need to estimate four different conditional mean functions to estimate θ_{frd}. As in SRD, we can use local linear regression estimators. Let us define the following estimators:

$$\begin{aligned}\left\{\hat{\beta}_0^+, \hat{\beta}_1^+\right\} &= \underset{\left(b_0^+, b_1^+\right)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{Z_i - c}{h}\right) I\left\{Z_i \geqslant c\right\} \left(Y_i - b_0^+ - b_1^+ \left(Z_i - c\right)\right)^2 \\ \left\{\hat{\beta}_0^-, \hat{\beta}_1^-\right\} &= \underset{\left(b_0^-, b_1^-\right)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{Z_i - c}{h}\right) I\left\{Z_i < c\right\} \left(Y_i - b_0^- - b_1^- \left(Z_i - c\right)\right)^2 \\ \left\{\hat{\gamma}_0^+, \hat{\gamma}_1^+\right\} &= \underset{\left(g_0^+, g_1^+\right)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{Z_i - c}{h}\right) I\left\{Z_i \geqslant c\right\} \left(D_i - g_0^+ - g_1^+ \left(Z_i - c\right)\right)^2 \\ \left\{\hat{\gamma}_0^-, \hat{\gamma}_1^-\right\} &= \underset{\left(g_0^-, g_1^-\right)}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{Z_i - c}{h}\right) I\left\{Z_i < c\right\} \left(D_i - g_0^- - g_1^- \left(Z_i - c\right)\right)^2.\end{aligned}$$

Then we estimate θ_{frd} by

$$\hat{\theta}_{\text{frd}} = \frac{\hat{\beta}_0^+ - \hat{\beta}_0^-}{\hat{\gamma}_0^+ - \hat{\gamma}_0^-}.$$

Alternatively, we can obtain the estimator $\hat{\theta}_{\text{frd}}$ using two stage least squares. Define intention to treat by $T = I\{Z \geq c\}$. Note T is a valid instrument for D in that T is exogenous conditional on Z. It can be shown that the LL approach with uniform kernels and same bandwidths is numerically equivalent to a TSLS regression:

$$Y_i = \delta_0 + \theta_{\text{frd}} D_i + \delta_1 (Z_i - c) + \delta_2 T_i (Z_i - c) + U_i$$

with T_i as the excluded instrument for D_i on the sample $\{i : c - h_n \leq Z_i \leq c + h_n\}$.

10.1.6 Validity of RD

RD imposes relatively weak assumptions and identifies a very specific and local parameter. The identification hinges on the continuity of $E[Y(d)|Z = z]$ at the cutoff, yet this assumption is fundamentally untestable and can be violated in the following situation. Suppose that the running variable is a test score. Individuals know the cutoff and have an option to re-take the test, and may do so if their scores are just below the cutoff. This leads to a discontinuity of the density $f_Z(z)$ of Z at the cutoff c, and possibly a discontinuity of $E[Y(d)|Z = z]$ as well because it is a functional of $f_Z(z)$,

$$E[Y(d)|Z = z] = \int y f_{Y|Z}(y|z) dy \text{ where } f_{Y|Z}(y|z) = \frac{f_{Y,Z}(y,z)}{f_Z(z)}.$$

This may invalidate the design. This problem is called "manipulation" of the running variable.

As a way to detect such manipulation, **?** proposes a test for continuity of the density of $f_Z(z)$ at the cutoff. In principle, one does not need continuity of the density of Z at c, but a discontinuity is suggestive of violations of the no-manipulation assumption. **?** also propose a new test based on order statistics that does not require smoothness assumptions.

In addition to manipulation, the continuity assumption may fail to hold due to discontinuity in the distribution of covariates. To see this, suppose there are an observed factor X and an unobserved factor U that affect potential outcomes, say

$$Y(d) = m_d(Z, X) + U$$

for some function m_d. Suppose that the distribution of X is discontinuous at $z = 0$. Then the discontinuity in X at 0 may affect the outcome because

$$E[Y(d)|Z = z] = E[E[Y(d)|Z = z, X]|Z = z] = \int \int y f_{Y|Z,X}(y|z, x) f_{X|Z}(x|z) dx dy$$

where $f_{Y|Z,X}(y|z, x) = \frac{f_{Y,Z,X}(y,z,x)}{f_{Z,X}(z,x)}$ and $f_{X|Z}(x|z) = \frac{f_{X,Z}(x,z)}{f_Z(z)}$. These effects may be attributed erroneously to the treatment of interest.

A common practice to test discontinuity of covariates is to test the null hypothesis that

$$H_0 : \lim_{z \uparrow 0} E(X \mid Z = z) = \lim_{z \downarrow 0} E(X \mid Z = z).$$

The rejection of this null suggests that $E(Y(d) \mid Z = z)$ may not be continuous either. However, $E(X \mid Z = z)$ could be still continuous and H_0 holds true even if the distribution of X is discontinuous at the cutoff. The intuition on how discontinuity in X may confound the effect of the treatment is about the entire distribution of X. **?** propose a test for continuity of $F_{X|Z}(x|z)$ at the cutoff. The test is easy to implement and based on permutation tests and it involves novel asymptotic arguments.

10.1.7 RD Packages

The statistical packages to compute LL RD estimators and run RDD validity tests are available online. Below we introduce four packages. [1]

rdrobust package It provides estimation, inference and graphical presentation using local polynomials, partitioning and spacings estimators. `rdrobust` implements local polynomial RD point estimators with classic and robust bias-corrected confidence intervals. `rdbwselect`, which is called by `rdrobust`, provides different data-driven bandwidth selectors based on **?**, cross-validation, and Calonico et al. (2014). `rdplot` plots data with "optimal" block length.

rddensity package It implements automatic manipulation tests based on density discontinuity at the cutoff using polynomial density estimator. `rddensity` runs manipulation testing using local polynomial density estimation and `rdbwdensity` selects a bandwidth or window.

rdperm packages It implements the approximate permutation test for RDD, developed in **?**.

rdcont packages It implements the approximate sign-test for RDD, developed in **?**.

10.2 Matching Estimators

10.2.1 Identification through Unconfoundedness

We change gear and study another way of estimating average treatment effect using matching estimators.

[1] You can download `rdrobust` and `rddensity` at https://rdpackages.github.io, and `rdperm` and `rdcont` at http://sites.northwestern.edu/iac879/software.

Suppose we observe (Y, D, X) and consider the following unconfoundedness assumption

$$(Y(0), Y(1)) \perp\!\!\!\perp D \mid X \tag{10.2}$$

which is often alternatively called selection on observables, conditional independence and so on. Unconfoundedness assumes that for subgroups of agents with the same X there are no unobservable differences between the treatment and control groups. This provides a way to identify the "conditional" average treatment effect (CATE) because

$$\begin{aligned} E[Y(1) - Y(0) \mid X = x] &= E[Y(1) \mid D = 1, X = x] - E[Y(0) \mid D = 0, X = x] \\ &= E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x] \end{aligned}$$

where the first line follows by unconfoundedness. Note that by integrating over X we can identify the ATE as well.

The idea behind the matching estimators is to find (or "match") units in the treatment group ($D = 1$) and control group ($D = 0$) with the same value of X, i.e., $X = x$. To be able to match, we need the overlap assumption

$$0 < P\{D = 1 \mid X = x\} < 1 \text{ for all } x$$

meaning that, within each subgroup of agents with the same X, there should be both treated and control units. Complication arises when X is continuously distributed.

Identification through the unconfoundedness assumption is inherently different from RDD. In sharp RDD, the unconfoundedness assumption holds trivially because if we define $D = I\{Z \geq c\}$ then

$$(Y(0), Y(1)) \perp\!\!\!\perp D \mid Z.$$

Moreover, the overlap assumption never holds in sharp RDD because the probabilities to receive treatments given the running variable is either 1 or 0, i.e.,

$$P\{D = 1 \mid Z < c\} = 0 \quad \text{and } P\{D = 0 \mid Z \geqslant c\} = 0.$$

10.2.2 Matching Metrics

If $X \in \mathbf{R}^k$ has continuous components, the event $\{X = x\}$ has measure zero, and so previous matching strategy is not feasible. To get around this, we match X's that are close according to some matching metric.

A common matching metric Mahalanobis distance is given by

$$M_{ij} = (X_i - X_j)' \, \Sigma^{-1} \, (X_i - X_j)$$

where $\Sigma = \text{Var}[X]$. Given this, j is the qth closest to X_i if

$$\sum_{s=1}^{n} I\{M_{is} \leqslant M_{ij}\} = q.$$

For alternative choices, there are Euclidean distance

$$M_{ij} = |X_i - X_j|$$

and the diagonal version of the Mahalanobis distance

$$M_{ij} = (X_i - X_j)' \operatorname{diag}\left[\Sigma^{-1}\right](X_i - X_j).$$

10.2.3 Matching Estimator

In order to define the matching estimator, fix q. Let $j_q(i)$ be the index $j \in \{1, \cdots, n\}$ that solves the following two conditions:

$$\begin{array}{cc} \text{Opposing treatment:} & D_j = 1 - D_i \\ \text{Opposing } q\text{th closest to } i\text{:} & \sum_{s:D_s=1-D_i} I\{M_{is} \leqslant M_{ij}\} = q. \end{array}$$

That is, $j_q(i)$ is the index of the unit that is the qth closest to unit i in terms of the covariate values, among the units with the treatment opposite to that of unit i. Let $\mathcal{J}_q(i)$ denote the set of indices for the first q matches for unit i :

$$\mathcal{J}_q(i) = \{j_1(i), \ldots, j_q(i)\}.$$

Then the matching estimator of $\theta_{ate} = E[Y(1) - Y(0)]$ is given by

$$\hat{\theta}_{\text{ate}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i(1) - \hat{Y}_i(0)\right) \text{ where } \hat{Y}(d) = \begin{cases} Y_i & \text{if } D_i = d \\ \frac{1}{q}\sum_{j\in\mathcal{Z}_q(i)} Y_j & \text{if } D_i \neq d \end{cases}.$$

This is a type of nearest neighbor (NN) estimator and thus as q increases the variance goes down while the bias increases. **?** study asymptotic properties of $\hat{\theta}_{\text{ate}}$ under a fixed number of matches as $n \to \infty$. Here we summarize some of noteworthy properties of $\hat{\theta}_{\text{ate}}$. First, $\hat{\theta}_{\text{ate}}$ is consistent as $n \to \infty$ for fixed q. Second, the bias of order $O(n^{-1/k_c})$ where k_c is the dimension of the continuous covariates and the variance is of order $O(1/n)$. Since $\sqrt{n}$Bias converges to 0, some constant, or ∞ for $k_c = 1$, $k_c = 2$, and $k_c > 2$ respectively, the estimator is not $\sqrt{n}$-asymptotically normal if $k_c > 2$. Third, $\hat{\theta}_{\text{ate}}$ is generally not efficient, and even if the bias is low enough, the estimators are not efficient given a fixed number of matches. Lastly, regarding the resampling methods which we will cover later in this course, **?** show that the bootstrap is generally invalid for the matching estimators due to non-smoothness in the matching process. However, subsampling is valid for $k_c \leq 2$.

10.2.4 Propensity Score Matching and Weighting

Aforementioned properties of the matching estimators suggest that we can make inference with the matching estimator only in limited cases where the number of continuous covariates does not exceed 2. Propensity score matching provides an alternative way to match.

Let $p(X) = P\{D = 1 \mid X = x\}$ denote the propensity score. **?** make an observation of paramount importance that unconfoundedness implies that

$$(Y(0), Y(1)) \perp\!\!\!\perp D \mid p(X).$$

This means that we no longer need to condition on the entire X but only on one-dimensional propensity score $p(X)$ in order to achieve independence between the potential outcome $(Y(0), Y(1))$ and the treatment status D. To see how it holds, note that

$$\begin{aligned} P\{D = 1 \mid Y(0), Y(1), p(X)\} &= E[E[D \mid Y(0), Y(1), P(X), X] \mid Y(0), Y(1), p(X)] \\ &= E[E[D \mid Y(0), Y(1), X] \mid Y(0), Y(1), p(X)] \\ &= E[E[D \mid X] \mid Y(0), Y(1), p(X)] \\ &= E[p(X) \mid Y(0), Y(1), p(X)] \\ &= p(X), \end{aligned}$$

which is the same as $P\{D = 1 \mid p(X)\}$. Interestingly, all the biases due to observable covariates can be removed by conditioning solely on the propensity score.

The Rosenbaum-Rubin result implies that

$$\theta_{ate} = E[E[Y \mid D = 1, p(X)] - E[Y \mid D = 0, p(X)]]$$

and thus we can use the matching estimator matching on the propensity score only. This can be formulated by nothing that

$$\begin{aligned} E\left[\frac{DY}{p(X)}\right] &= E\left[\frac{1}{p(X)} E[DY(1) \mid p(X)]\right] \\ &= E\left[\frac{1}{p(X)} E[D \mid p(X)] E[Y(1) \mid p(X)]\right] = E[Y(1)] \end{aligned}$$

and similarly

$$E\left[\frac{(1-D)Y}{1-p(X)}\right] = E[Y(0)]$$

which in turn allows us to write

$$\theta_{ate} = E\left[\frac{[D_i - p(X_i)]Y_i}{p(X_i)(1-p(X_i))}\right] = E\left[\frac{DY}{p(X)}\right] - E\left[\frac{(1-D)Y}{1-p(X)}\right].$$

We define the estimator for θ_{ate} by the sample analog of θ_{ate}:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{[D_i - p(X_i)] Y_i}{p(X_i)(1 - p(X_i))} \right).$$

$\hat{\theta}_n$ does not explicitly match observations but puts weights induced by the propensity score to outcome Y_i, while it is still based on the unconfoundedness assumption.

The propensity score is a scalar. **?** imply that the bias term is of lower order than the variance term and matching leads to a $\sqrt{n}$-consistent, asymptotically normal estimator. Given the data, we cannot compute $\hat{\theta}_n$ because it depends on the unknown propensity score function $p(\cdot)$. The estimator based on the true propensity score has the same asymptotic variance in **?**. With estimated propensity scores, the asymptotic variance of matching estimators is more involved due to the "generated regressor". The topic is beyond our scope. Those who are interested can consult **?**.

Bibliography

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

Imbens, G. and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of Economic Studies*, 933–959.

Lecture 11

Random Forests

11.1 Coming soon

The lecture notes for this lecture will be updated before class takes place.

Lecture 12

LASSO

12.1 High Dimensionality and Sparsity

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^k$. Let $\beta = (\beta_1, \ldots, \beta_k)' \in \mathbf{R}^k$ be such that

$$Y = X'\beta + U \ .$$

We observe a random sample $\{(Y_i, X_i) : 1 \leq i \leq n\}$ from the distribution of (Y, X) and without loss of generality, we further assume that

$$\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i = 0 \qquad \text{and} \qquad \hat{\sigma}^2_{n,j} \equiv \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - \bar{X}_j)^2 = 1 \ , \qquad (12.1)$$

where $X_{i,j}$ denotes the j^{th} component of X_i. In other words, we assume the model does not have a constant term and that all variables are on the same scale (something that will be important later). Our goal today is to study estimation of β when k is large relative to n. That could mean that $k < n$, but not by much, or simply that $k > n$. For simplicity, we assume X and U are independent.

When $k > n$, the ordinary least squares estimator is not well-behaved since the $\mathbb{X}'\mathbb{X}$ matrix does not have full rank and is not invertible. In particular, the estimator is not unique and will overfit the data. If all explanatory variables are important in determining the outcome, it is not possible to tease out their individual effects. However, if the model is *sparse* – that is, only a few components of X have an influence on Y – then it might be possible to discriminate between the relevant and irrelevant components of X. The following definition formalizes this notion.

Definition 12.1 (Sparsity) *Let $S = \{j : \beta_j \neq 0\}$ be the identity of the relevant regressors. A model is said to be sparse if $s = |S|$ is fixed as $n \to \infty$.*

If we knew the identity of the relevant regressors S then we could simply do least squares in the usual manner. Since this would represent a sort of ideal situation, we will call such a strategy the "oracle".

Definition 12.2 (Oracle Estimator) *The oracle estimator $\hat{\beta}_n^{\text{o}}$ is the infeasible estimator that is estimated by least squares using only the variables in S.*

In practice, we do not know the set S and so our goal is to estimate β, and possibly S, exploiting the fact that the model is known to be sparse. In particular, we would like our estimator $\hat{\beta}_n$ to satisfy three properties: estimation consistency, model selection consistency, and oracle efficiency.

Definition 12.3 (Estimation Consistency) *An estimator $\hat{\beta}_n$ is estimation consistent if*

$$\hat{\beta}_n \xrightarrow{P} \beta \ .$$

Definition 12.4 (Model-Selection Consistency) *Let*

$$\hat{S}_n = \{j : \hat{\beta}_{n,j} \neq 0\}$$

be the set of relevant covariates selected by an estimator $\hat{\beta}_n$. Then, $\hat{\beta}_n$ is model-selection consistent if

$$P\{\hat{S}_n = S\} \to 1 \text{ as } n \to \infty \ .$$

Definition 12.5 (Oracle Efficiency) *An estimator $\hat{\beta}_n$ is oracle efficient if it achieves the same asymptotic variance as the oracle estimator $\hat{\beta}_n^{\text{o}}$.*

Achieving Oracle efficiency requires stronger conditions than achieving model selection consistency, which in turn requires stronger assumptions than estimation consistency. The last statement is rather straightforward if we are able to consistently select variables, we would then be able to run least squares on the selected variables. On the other hand, it is possible for $\|\hat{\beta}_n - \beta\|_2^2$ to be small even when $\hat{\beta}_n$ is non-zero at every component, so that selection never occurs.

12.2 LASSO

LASSO is short for Least Absolute Shrinkage and Selection Operator and is one of the well known estimators for sparse models. The LASSO estimator $\hat{\beta}_n$ is defined as the solution to the following minimization problem

$$\hat{\beta}_n = \arg\min_b \left(\sum_{i=1}^n (Y_i - X_i' b)^2 + \lambda_n \sum_{j=1}^k |b_j| \right) , \tag{12.2}$$

where λ_n is a scalar tuning parameter. For a fixed $\lambda_n > 0$, LASSO corresponds to OLS with an additional term that imposes a penalty for non-zero coefficients. The penalty term shrinks the estimated coefficients towards zero and this gives us model selection, albeit at the cost of introducing bias in the estimated coefficients. The LASSO estimator can be alternatively described as the solution to

$$\min_b \sum_{i=1}^n (Y_i - X_i'b)^2 \quad \text{subject to} \quad \sum_{j=1}^k |b_j| \leq t \ , \tag{12.3}$$

where now t is a scalar tuning parameter.

LASSO has the feature of delivering estimated coefficients that can be exactly 0 for a given sample size n. The form of the penalty function is important for selection, which does not occur under OLS or other penalty functions (e.g., ridge regression). For intuition, consider penalty functions of the form

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

If $\gamma > 1$, the objective function is continuously differentiable at all points. The first order condition with respect to $\beta_{n,j}$ would then be

$$2\sum_{i=1}^n (Y_i - X_i'\beta)X_{i,j} = \lambda_n \gamma |\beta_j|^{\gamma-1} \text{sign}(\beta_j) \ .$$

Suppose $\beta_j = 0$. Then, $\hat{\beta}_{n,j} = 0$ if and only if

$$0 = \sum_{i=1}^n (Y_i - X_i'\hat{\beta}_n)X_{i,j} = \sum_{i=1}^n (U_i - X_i'(\hat{\beta}_n - \beta))X_{i,j} \ . \tag{12.4}$$

Whenever U is continuously distributed, the above equation holds with probability 0 and model selection does not occur.

On the other hand, if $\gamma \leq 1$, the penalty function is not differentiable at 0. In this case, Karush-Kuhn-Tucker conditions are expressed in terms of the subgradient.

Definition 12.6 (Sub-gradient & Sub-differential) *We say $g(\cdot) \in \mathbf{R}$ is a sub-gradient of $f(x) : \mathbf{R} \to \mathbf{R}$ at point x if $f(z) \geq f(x) + g(z - x)$ for all $z \in \mathbf{R}$. The set of sub-gradients of $f(\cdot)$ at x, denoted by $\partial f(x)$, is the sub-differential of $f(\cdot)$ at x.*

In the case of LASSO, we need the sub-differential of the absolute value $f(x) = |x|$. For $x < 0$ the sub-gradient is uniquely given by $\partial f(x) = \{-1\}$. For $x > 0$ the sub-gradient is uniquely given by $\partial f(x) = \{1\}$. At $x = 0$ the sub-differential is defined by the inequality $|z| \geq gz$ for all z, which is

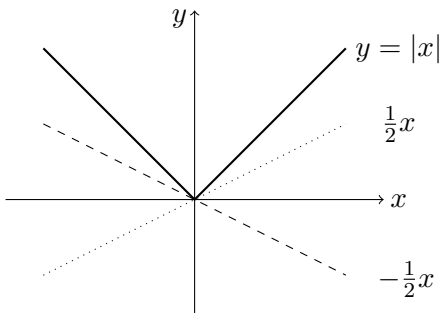

Figure 12.1: Two sub-gradients of $f(x) = |x|$ at $x = 0$

satisfied if and only if $g \in [-1, 1]$. We therefore have $\partial f(x) = [-1, 1]$. This is illustrated in Figure 12.1

For non-differentiable functions, the Karush-Kuhn-Tucker theorem states that a point minimizes the objective function of interest if and only if 0 is in the sub-differential. Applying this to the problem in (12.2) implies that the first order conditions are given by

$$2\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)X_{i,j} = \lambda_n \operatorname{sign}(\hat{\beta}_{n,j}) \quad \text{if} \quad \hat{\beta}_{n,j} \neq 0 \tag{12.5}$$

and

$$-\lambda_n \leq 2\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)X_{i,j} \leq \lambda_n \quad \text{if} \quad \hat{\beta}_{n,j} = 0\ . \tag{12.6}$$

Compared to our previous result in (12.4), this inequality is attained with positive probability even when U is continuously distributed. Model selection is therefore possible when the penalty function has a cusp at 0. The difference between using a penalty with $\gamma = 1$ (LASSO) and $\gamma = 2$ (Ridge) in the constraint problem in (12.3) is illustrated in Figure 12.2 for the simple case where $k = 2$.

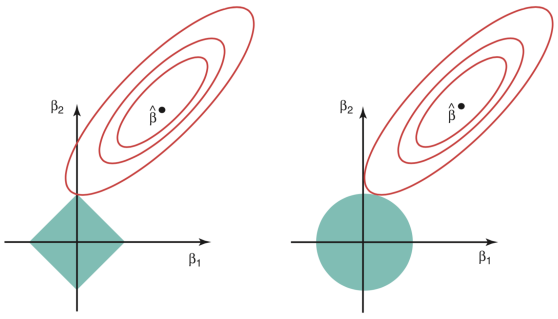

Figure 12.2: Constrained problem in (12.3) when $k = 2$: $\gamma = 1$ (left panel) and $\gamma = 2$ (right panel).

12.2.1 Theoretical Properties of the LASSO

For ease of exposition, we only discuss the case where k as fixed as $n \to \infty$. Assume without loss of generality that S consists of the first s variables. We partition X into $X = (X_1', X_2')'$ where X_1 are the first s explanatory variables, and X_2 are the $k - s$ remaining variables. Partition the variance-covariance matrix of X accordingly,

$$\Sigma = E[XX'] = \begin{pmatrix} E[X_1X_1'] & E[X_1X_2'] \\ E[X_2X_1'] & E[X_2X_2'] \end{pmatrix} .$$

Assumption 12.1 (Irrepresentable Condition) $\exists \eta > 0$ *s.t.*

$$\|E[X_2X_1']E[X_1X_1']^{-1} \cdot \text{sign}(\beta_1, \ldots, \beta_s)\|_\infty \leq 1 - \eta .$$

To understand the condition, note that when the sign of β is unknown, we basically require the condition to hold for all possible signs. That is,

$$\|E[X_1X_1']^{-1}E[X_1X_2']\|_\infty \leq 1 - \eta .$$

This means that the regression coefficients of the irrelevant variables on the relevant variables must all be less than 1. In that sense, the former are irrepresentable by the latter. Under this condition, the following holds.

Theorem 12.1 (Zhao and Yu (2006)) *Suppose k and s are fixed and that $\{X_i : 1 \leq i \leq n\}$ and $\{U_i : 1 \leq i \leq n\}$ are i.i.d. and mutually independent. Let X have finite second moments, and U have mean 0 and variance σ^2. Suppose also that the irrepresentable condition holds and that*

$$\frac{\lambda_n}{n} \to 0 \quad \textit{and} \quad \frac{\lambda_n}{n^{\frac{1+c}{2}}} \to \infty \quad \textit{for} \quad 0 \leq c < 1 . \tag{12.7}$$

Then LASSO is model-selection consistent.

The irrepresentable condition is a restrictive condition. When this condition fails and $\lambda_n/\sqrt{n} \to \lambda^* > 0$, it can be shown that LASSO selects *too many* variables (i.e., it selects a model of bounded size that contains all variables in S). Intuitively, if the relevant variables and irrelevant variables are highly correlated, we will not be able discriminate between them.

Knight and Fu (2000) showed that the LASSO estimator is asymptotically normal when $\lambda_n/\sqrt{n} \to \lambda^* \geq 0$, but that the nonzero parameters are estimated with some asymptotic bias if $\lambda^* > 0$. If $\lambda^* = 0$, LASSO has the same limiting distribution as the LS estimator and so even with $\lambda^* = 0$, LASSO is not oracle efficient. In addition, the requirement for asymptotic normality is at conflict with $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ and so it follows that LASSO cannot be both model selection consistent and asymptotically normal (hence oracle efficient) at the same time. Oracle efficient penalization methods work by penalizing small coefficients a lot and large coefficients very little or not at all. This could be done by using weights (as in the Adaptive LASSO below) or by changing the penalty function (which we discuss later).

12.3 Adaptive LASSO

Definition 12.7 (Adaptive LASSO) *The adaptive LASSO is the estimator $\tilde{\beta}_n$ that arises from the following two steps.*

1. *Estimate β using ordinary LASSO,*

$$\hat{\beta}_n = \arg\min_b \left(\sum_{i=1}^n (Y_i - X_i'b)^2 + \lambda_{1,n} \sum_{j=1}^k |b_j| \right) ,$$

   where $\lambda_{1,n}/\sqrt{n} \to \lambda^* > 0$.

2. *Let $\hat{S}_1 = \{j : \hat{\beta}_n \neq 0\}$ be the set of selected covariates from the first step. Estimate β by*

$$\tilde{\beta}_n = \arg\min_b \left(\sum_{i=1}^n (Y_i - \sum_{j \in \hat{S}_1} X_{i,j} b_j)^2 + \lambda_{2,n} \sum_{j \in \hat{S}_1} |\hat{\beta}_{n,j}|^{-1} |b_j| \right) ,$$

   where $\lambda_{2,n}/\sqrt{n} \to 0$ and $\lambda_{2,n} \to \infty$.

Adaptive LASSO imposes a penalty in the second step that is inversely proportional to the magnitude of the estimated coefficient in the first step. This adaptive weights allows us to eliminate small, irrelevant covariates while retaining the relevant ones without introducing asymptotic bias.

Theorem 12.2 (Zou, 2006) *Suppose $\{X_i : 1 \leq i \leq n\}$ and $\{U_i : 1 \leq i \leq n\}$ are i.i.d. and mutually independent. Let X have finite second moments, and U have mean 0 and variance σ^2. The adaptive LASSO is model selection consistent and oracle efficient, i.e.,*

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 E(X_1 X_1')^{-1}) .$$

To see that adaptive LASSO is oracle efficient, note that the asymptotic variance of the estimator is the same we would have achieved had we known the set S and performed OLS on it. The rates at which $\lambda_{1,n}$ and $\lambda_{2,n}$ grow are important for this result.

To see why the adaptive LASSO is model selection consistent and oracle efficient, consider the following. Recall that $\beta_1, \ldots, \beta_s \neq 0$ and $\beta_{s+1}, ..., \beta_k = 0$. Suppose that $\hat{\beta}_n$ has r non-zero components asymptotically. Without the irrepresentable condition, the LASSO includes too many variables, so that $s \leq r \leq k$. Without loss of generality, suppose $\hat{\beta}_n$ is non-zero in its first r components. Let b be any $r \times 1$ vector, and let $\tilde{\beta}_n$ denote the adaptive LASSO estimator. Define $u = \sqrt{n}(b - \beta)$. Some algebra shows,

$$\sqrt{n}(\tilde{\beta}_n - \beta) = \arg\min_u \sum_{i=1}^n \left(U_i - \frac{1}{\sqrt{n}} \sum_{j=1}^r X_{i,j} u_j \right)^2 + \lambda_{2,n} \sum_{j=1}^r |\hat{\beta}_{n,j}|^{-1} (|\beta_j + \frac{1}{\sqrt{n}} u_j| - |\beta_j|) .$$

By Knight and Fu (2000), $\hat{\beta}_n$ converges at rate $\sqrt{n}$, so $|\hat{\beta}_n - \beta| = O_P(n^{-1/2})$. We then split the analysis according to whether β_j is zero or not.

Case $\beta_j = 0$: here $|\hat{\beta}_{n,j}| = O_P(n^{-1/2})$. Then,

$$\lambda_{2,n}|\hat{\beta}_{n,j}|^{-1}(|\beta_j + \frac{1}{\sqrt{n}}u_j| - |\beta_j|) \approx \lambda_{2,n}|u_j|$$

where we have "canceled" the $1/\sqrt{n}$ term using $|\hat{\beta}_{n,j}|$. Now suppose $u_j \neq 0$ and note that

$$\lambda_{2,n}|u_j| \to \infty \text{ since } \lambda_{2,n} \to \infty \ .$$

The penalty effectively tends to infinity, so that $b_j \neq 0$ ($u_j \neq 0$) cannot be the minimizer. It must be that $u_j = 0$, i.e., $b_j = \beta_j = 0$.

Case $\beta_j \neq 0$: here $|\hat{\beta}_{n,j}| = O_P(1)$. It follows that,

$$\lambda_{2,n}|\hat{\beta}_{n,j}|^{-1}(|\beta_j + \frac{1}{\sqrt{n}}u_j| - |\beta_j|) \approx \lambda_{2,n}\frac{1}{\sqrt{n}}|u_j| \ .$$

It follows that $\lambda_{2,n}\frac{1}{\sqrt{n}}|u_j| \xrightarrow{P} 0$ since $\frac{\lambda_{2,n}}{\sqrt{n}} \to 0$ and $u_j = O_P(1)$. That is, asymptotically there is no penalty on non-zero terms, and the adaptive LASSO becomes asymptotically equivalent to OLS estimation on S. This gives rise to model selection consistency and oracle efficiency.

12.4 Penalties for Model Selection Consistency

Another way to achieve a model-selection consistent estimator is to use a penalty function that is strictly concave (as a function of $|b_j|$) and has a cusp at the origin. As previously mentioned, LASSO is essentially OLS with an L^1 penalty term. As such, it belongs to the larger class of Penalized Least Squares estimators:

$$\hat{\beta}_n^{PLS}(\lambda) = \arg\min_b \left(\sum_{i=1}^n (Y_i - X_i'b)^2 + \sum_{j=1}^k p_\lambda(|b_j|) \right) .$$

Clearly, ordinary LASSO corresponds to the case where $p_\lambda(|\nu|) = \lambda|\nu|$, but such a penalty is not strictly concave and so model selection consistency generally does not occur. Some alternative penalty functions include that have the desire property are

1. Bridge: $p_\lambda(|\nu|) = \lambda|\nu|^\gamma$ for $0 < \gamma < 1$
2. Smoothly Clipped Absolute Deviation (SCAD): for $a > 2$,

$$p'_\lambda(|\nu|) = \lambda\left[I\left\{ |\nu| \leq \frac{\lambda}{n} \right\} + \frac{(a\lambda/n - |\nu|)_+}{(a-1)\lambda/n} I\left\{ |\nu| > \frac{\lambda}{n} \right\} \right] .$$

Note that this function is defined by its derivative.

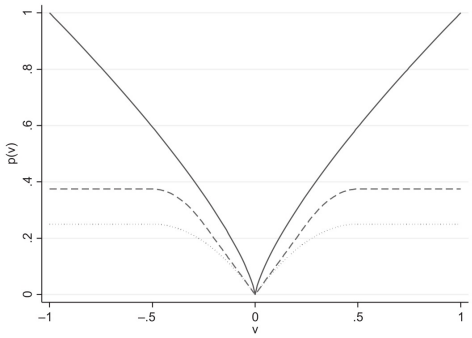

Figure 12.3: Bridge penalty (solid line), SCAD penalty (dashed line) and minimax concave penalty (dotted line)

3. Minimax Concave: for $a > 0$,

$$p_\lambda(|\nu|) = \lambda \int_0^{|\nu|} \left(1 - \frac{nx}{a\lambda}\right)_+ dx$$

where $(x)_+ = \max\{0, x\}$. These penalty functions are plotted in Figure 12.3. Note that they are all steeply sloped near $\nu = 0$. Bridge penalty, like the LASSO, continues to increase far away from $\nu = 0$, whereas SCAD and minimax concave penalties flatten out. For this reason, the latter penalties exhibit lower bias.

12.5 Choosing lambda

The need for model selection consistency imposes constraints on the growth rate of λ_n, but does not pin down their specific values. In practice, λ_n for the ordinary LASSO is often chosen by Q-fold cross validation.

Let Q be some integer, and suppose for ease of exposition that $n = Qn_q$. We partition the sample into the sets $I_1, \ldots, I_Q$ each with n_q members. For each $1 \leq q \leq Q$, perform LASSO on all but the observations in I_q to obtain $\hat{\beta}_{n,-q}(\lambda)$. Then, calculate the squared prediction error of $\hat{\beta}_{n,-q}(\lambda)$ on the set I_q:

$$\Gamma_q(\lambda) = \sum_{i \in I_q} (Y_i - X_i' \hat{\beta}_{n,-q}(\lambda))^2 \ .$$

Doing so for each q, we are able to find total error for each λ: $\Gamma(\lambda) = \sum_{q=1}^{Q} \Gamma_q(\lambda)$. Then we define the cross validated λ as:

$$\hat{\lambda}_n^{CV} = \arg\min_\lambda \Gamma(\lambda) \ .$$

For the adaptive LASSO, we need to choose both $\lambda_{1,n}$ and $\lambda_{2,n}$. A computationally efficient way of doing so is to choose $\lambda_{1,n}$ via the above

cross-validation procedure, and then having fixed this $\lambda_{1,n}$, choose $\lambda_{2,n}$ by a second round of cross-validation.

Arguably, there exist few results about the properties of the LASSO when λ_n is chosen via cross-validation. In a recent working paper, Chetverikov et al. (2016) show that in a model with random design, in which k is allowed to depend on n, and assuming $U_i|X_i$ is Gaussian, it follows that

$$\|\hat{\beta}_n - \beta\|_{2,n} \leq Q \cdot ((|S| \log k)/n)^{1/2} \log^{7/8}(kn)$$

holds with high probability, where $\|b-\beta\|_{2,n} = (\frac{1}{n}\sum_{i=1}^n (X_i'b)^2)^{1/2}$ is the prediction norm. It turns out that $((|S| \log k)/n)^{1/2}$ is the fastest convergence rate possible so that cross-validated LASSO is nearly optimal. However, it is not known if the $\log^{7/8}(kn)$ term can be dropped.

Finally, we mention one alternative approach to choosing λ_n. This is done by minimizing the Bayesian Information Criterion. Define:

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

and

$$BIC(\lambda) = \log(\hat{\sigma}^2(\lambda)) + |\hat{S}_n(\lambda)| C_n \frac{\log(n)}{n}$$

where C_n is an arbitrary sequence that tends to ∞. Wang et al. (2009) show that under some technical conditions, choosing λ_n to minimize $BIC(\lambda)$ leads to model selection consistency when U is normally distributed.

12.6 Concluding Remarks

Today we focused on the framework that keeps k fixed even as $n \to \infty$. There exist many extensions to the stated theorems that are valid in cases where $k_n = O(n^a)$ or even $k_n = O(e^n)$. Sources such as Fan et al. (2011) and Horowitz (2015).

Finally, we note that many packages are available for LASSO estimation. A few starting points are `lassopack` in Stata, and `glmnet` or `parcor` in R.

Bibliography

Chetverikov, D., Z. Liao, and V. Chernozhukov (2016): "On Cross-Validated LASSO," *arXiv preprint arXiv:1605.02214*.

Fan, J., J. Lv, and L. Qi (2011): "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, 291–317.

Horowitz, J. L. (2015): "Variable selection and estimation in high-dimensional models," *Canadian Journal of Economics*, 48, 389–407.

Knight, K. and W. Fu (2000): "Asymptotics for lasso-type estimators," *The Annals of statistics*, 28, 1356–1378.

Wang, H., B. Li, and C. Leng (2009): "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71, 671–683.

Zhao, P. and B. Yu (2006): "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Lecture 13

Binary Choice

Let (Y, X) be a random vector where Y takes values in $\{0, 1\}$ and X takes values in $\mathbf{R}^{k+1}$. Let us consider the problem of estimating

$$P\{Y = 1 \mid X\} \ . \tag{13.1}$$

This problem has two interpretations that can deliver different approaches. The first interpretation consists in predicting the outcome variable Y for a given value of the covariate X. This problem can be solved estimating the probability (13.1) - which is also called propensity score - by different methods, for instance local linear regression or classification trees.

The second interpretation of the problem consist in viewing (13.1) as a model with structure, where we are interested in the partial effects or the causal effects of X. This is traditionally the approach that is often used in the Industrial Organization, where (13.1) models the behavior of the decision makers and the estimated model is used to do counterfactual analysis. Sometimes, this second interpretation is called a *structural form* for (13.1), while the first one is a *reduced form*.

In this lecture, we consider the second interpretation. We restrict our attention to parametric and semiparametric models using the linear index model. This model assumes the existence of $\beta \in \mathbf{R}^{k+1}$ such that

$$P\{Y = 1 \mid X\} = P\left\{Y = 1 \mid X'\beta\right\} \ . \tag{13.2}$$

This condition reduces the dimension of the problem. To see this, note that the left hand side in (13.2) is a function of $X \in \mathbf{R}^{k+1}$. While the right hand side in (13.2) is a function of $X'\beta \in \mathbf{R}$, which is known as *linear index*.

13.1 Linear Index Model

Let (Y, X, U) be a random vector where Y takes values in $\{0, 1\}$, X takes values in $\mathbf{R}^{k+1}$ with $X_0 = 1$ and U take values in $\mathbf{R}$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in$

$\mathbf{R}^{k+1}$ be such that

$$Y = I\{X'\beta - U \geq 0\} \ .$$

This model is known as *Threshold crossing model* or *Single index model* or *Linear index model*.

In this setup, the binary outcome Y often indicates the observable choice between two alternatives of a decision maker. This choice is modeled by utility maximization. For instance, let us consider two alternatives A and B that gives the following utility levels

$$X'\beta_A + U_A \quad \text{and} \quad X'\beta_B + U_B$$

to the decision maker, respectively. If the decision maker maximizes utility, she will choose B over A if

$$X'\beta_B + U_B \geq X'\beta_A + U_A \ ,$$

which is equivalent to

$$X'(\beta_B - \beta_A) + U_B - U_A \geq 0 \ .$$

This implies that $X'\beta - U$ can be interpreted as the difference in the utility level between two choices. Often times, one of the options is normalize to zero, which leads to the linear index model described above.

13.1.1 Identification

Let us denote by P the distribution of the observed data. And denote by $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ a (statistical) model for P. These probability distributions are indexed by the parameter θ, where this parameter could have infinite dimensional components, e.g. a nonparametric distribution of the unobservable component.

Using this notation, the model $\mathbf{P}$ is correctly specified if the distribution of the observable data belong to the model, i.e. $P \in \mathbf{P}$. In this case, there is a parameter θ such that $P = P_\theta$. Now our interest interest might be in θ or a function of $\lambda(\theta)$.

Definition 13.1 *Let $\Theta_0(P)$ be the collection of θ such that $P = P_\theta$, i.e.*

$$\Theta_0(P) = \{\theta \in \Theta : P_\theta = P\} \ .$$

We say that θ is identified if $\Theta_0(P)$ is a singleton for all $P \in \boldsymbol{P}$.

Remark 13.1 $\lambda(\theta)$ may be identified even if θ is not. For instance, in the linear model, we can identified the coefficient using the assumptions described in Lecture 1, but we cannot identified the distribution of the errors. In this example, θ is defined by the coefficients of the linear model, the distribution of the covariates, and the conditional distribution of the errors.

13.1.2 Identification of the parametric binary model

In the binary model described by the linear index model, the parameter is $\theta = (\beta, P_X, P_{U|X})$. Denote by Θ the set of all the possible values of θ. And let us consider the following parametric assumption.

Assumption 13.1 *Consider the parametric assumptions:*

P1. $P_{U|X} = N(0, \sigma^2)$.

P2. *There exists no $A \subseteq \boldsymbol{R}^{k+1}$ such that A has probability one under P_X and A is a proper linear subspace of $\boldsymbol{R}^{k+1}$.*

The parametric assumption **P1** allow us to replace $P_{U|X}$ with σ, since that parameter characterizes the parametric distribution. Now we can write $\theta = (\beta, P_X, \sigma)$. Using this notation, let us study the identification of θ. We prove the result by contradiction: assume there are two values $\theta = (\beta, P_X, \sigma)$ and $\theta^* = (\beta^*, P_X^*, \sigma^*)$ such that $\theta \neq \theta^*$ and $P_\theta = P_{\theta^*}$ and the reach a contradiction.

First, notice that the marginal distribution of X is identified from the joint distribution of (Y, X). This implies that $P_X = P_X^*$. Second, we can use assumption **P1** to compute the probability of $Y = 1$ given X using both models. That is

$$P_\theta\{Y = 1|X\} = \Phi\left(\frac{X'\beta}{\sigma}\right) \quad \text{and} \quad P_{\theta^*}\{Y = 1|X\} = \Phi\left(\frac{X'\beta^*}{\sigma^*}\right) ,$$

which by assumption on P_θ and P_{θ^*} deliver the same probability. Since $\Phi(\cdot)$ is an increasing function, we obtain

$$X'\left(\frac{\beta}{\sigma} - \frac{\beta^*}{\sigma^*}\right) = 0 .$$

By assumption **P2**, we conclude

$$\frac{\beta}{\sigma} = \frac{\beta^*}{\sigma^*} . \tag{13.3}$$

Otherwise, we can define a proper linear subspace $A = \{x \in \mathbf{R}^{k+1} | x'(\beta/\sigma - \beta^*/\sigma^*) = 0\}$ such that A has probability one under P_X.

Note that we cannot conclude that $\beta = \beta^*$ or $\sigma = \sigma^*$. Indeed, our analysis shows that any θ and θ^* such that (13.3) holds and $P_X = P_X^*$ satisfies $P_\theta = P_{\theta^*}$. This implies we cannot identify $\theta = (\beta, P_X, \sigma)$ but we can identify $\lambda(\theta) = (P_X, \beta/\sigma)$.

A few remarks are in order. First, researchers typically assume further that $|\beta| = 1$ or $\beta_0 = 1$ or $\sigma = 1$, which is a *normalization* argument to conclude that now we can identify the parameter θ. Second, the model with $\sigma = 1$ is called the **Probit** model. Let us verify the identification of θ.

Under this additional assumption, we have $\sigma^* = 1$. By equation (13.3), we conclude $\beta = \beta^*$. This implies that $\theta = (\beta, P_X, 1)$ and $\theta^* = (\beta^*, P_X, 1)$ are the same. Finally, other parametric assumptions on the conditional distribution $P_{U|X}$, which is used to compute the probability $Y = 1$ given X, deliver other models. In particular, if this parametric distribution is the logistic distribution, we obtain the **Logit** model.

A natural question that motivate the next section is the identification of θ without parametric assumptions on $P_{U|X}$.

13.1.3 Identification via median independence

Let us recall that in the linear model that we studied in Lecture 1, the identification assumption that we need from $P_{U|X}$ was $E[U|X] = 0$. If we assume $E[U|X] = 0$ instead of Assumption **P1**, it can be prove that nothing is learned about $(\beta, P_{U|X})$. This result was shown by Manski (1988), which point it out that this conditional mean assumption is not even useful to identify $\lambda(\theta) = \beta$.

In general, the mean independence assumptions are rather useless in non-linear models. This is not the case of the median independence assumption, $\text{Med}(U|X) = 0$, which can be used to identify $\lambda(\theta) = \beta$. Let us describe in more detail the assumptions necessary for the identification result.

Assumption 13.2 *Consider the following semi-parametric assumptions that include the existence of a special covariate:*

S1. *Med$(U|X) = 0$ with probability 1 under P_X.*

S2. *There exists no $A \subseteq \boldsymbol{R}^{k+1}$ such that A has probability one under P_X and A is a proper linear subspace of $\boldsymbol{R}^{k+1}$.*

S3. $|\beta| = 1$.

S4. *P_X is such that at least one component of X has support equal to $\boldsymbol{R}$* conditional on the other components *with probability 1 under P_X. Moreover, the corresponding component of β is non-zero.*

Note that **S1** is a weaker assumption than **P1**, **S2** is the same as assumption **P2** and **S3** is a normalization assumption similar to $\sigma = 1$ discussed in the Probit case. What is new is assumption **S4**, which is a stronger assumption on P_X and also on β. This assumption is also known as the special regressor assumption and will be fundamental for the identification of β.

Before to present the main result of this section, let us present the following lemma. This provides an additional insight that is useful for the proof of the identification result.

Lemma 13.1 *Let $\theta = (\beta, P_X, P_{U|X})$ be a parameter that satisfies **S1**. Consider any β^*. If*

$$P_\theta \left\{ X'\beta^* < 0 \leq X'\beta \cup X'\beta < 0 \leq X'\beta^* \right\} > 0 \ , \tag{13.4}$$

*then there exists no $\theta^* = (\beta^*, P_X^*, P_{U|X}^*)$ satisfying **S1** and also having $P_\theta = P_{\theta^*}$.*

PROOF. Suppose by contradiction that (13.4) holds yet there exists such θ^*. Because $P_\theta = P_{\theta^*}$, we conclude that $P_X = P_{X^*}$. Let us recall that $Y = I\{X'\beta - U \geq 0\}$. Note that

$$P_\theta \{Y = 1|X\} \geq \frac{1}{2} \iff P_\theta \left\{ X'\beta \geq U|X \right\} \geq \frac{1}{2} \ ,$$

by Assumption **S1**, the median independence assumption, this happens

$$\iff X'\beta \geq 0 \ .$$

In a similar way, we have

$$P_{\theta^*} \{Y = 1|X\} \geq \frac{1}{2} \iff P_{\theta^*} \left\{ X'\beta^* \geq U|X \right\} \geq \frac{1}{2} \iff X'\beta^* \geq 0 \ ,$$

where the last equivalence follows by Assumption **S1** as before. Now, our condition (13.4) implies that with positive probabilty, either

$$X'\beta^* < 0 \leq X'\beta \quad \text{or} \quad X'\beta < 0 \leq X'\beta^* \ ,$$

which implies that either

$$P_{\theta^*} \{Y = 1|X\} < \frac{1}{2} \leq P_\theta \{Y = 1|X\}$$

or

$$P_\theta \{Y = 1|X\} < \frac{1}{2} \leq P_{\theta^*} \{Y = 1|X\} \ .$$

This contradicts the fact that $P_\theta = P_{\theta^*}$. This complete the proof. ■

Now we are ready to state our mail result of this section.

Theorem 13.1 *Under assumption **S1**-**S4**, $\lambda(\theta) = \beta$ is identified.*

PROOF. Assume without loss of generality that the (special) component of X specified in **S4** is the **kth component** and that $\beta_k > 0$.

Let $\theta = (\beta, P_X, P_{U|X})$ be a parameter that satisfies **S1**-**S2**. Consider any $\beta^* \neq \beta$. We want to show that there is no $\theta^* = (\beta^*, P_X^*, P_{U|X}^*)$ that satisfies **S1**-**S4** such that $P_\theta = P_{\theta^*}$.

From the previous Lemma it is sufficient to show that:

$$P_\theta \left\{ X'\beta^* < 0 \leq X'\beta \cup X'\beta < 0 \leq X'\beta^* \right\} > 0 \ ,$$

which is equivalent to prove that

$$P_\theta \left\{ X'\beta^* < 0 \leq X'\beta \right\} > 0$$

or

$$P_\theta \left\{ X'\beta < 0 \leq X'\beta^* \right\} > 0 \ .$$

To prove this condition (or one of the equivalence), we consider three cases according to the sign of β_k^*. In what follows, we denote by X_{-k} as the vector X without its **kth** component. In a similar way, β_{-k} and β_{-k}^* are the vector of parameters without their kth component.

Case 1. Suppose $\beta_k^* < 0$. Then,

$$P_\theta \left\{ X'\beta^* < 0 \leq X'\beta \right\} = P_\theta \left\{ X_k > -\frac{X'_{-k}\beta^*_{-k}}{\beta_k^*},\ X_k > -\frac{X'_{-k}\beta_{-k}}{\beta_k} \right\} \ .$$

By Assumption **S4**, the above probability is positive.

Case 2. Suppose $\beta_k^* = 0$. Then,

$$P_\theta \left\{ X'\beta^* < 0 \leq X'\beta \right\} = P_\theta \left\{ X'_{-k}\beta^*_{-k} < 0,\ X_k > -\frac{X'_{-k}\beta_{-k}}{\beta_k} \right\} \ , \quad (13.5)$$

and

$$P_\theta \left\{ X'\beta < 0 \leq X'\beta^* \right\} = P_\theta \left\{ X'_{-k}\beta^*_{-k} \geq 0,\ X_k < -\frac{X'_{-k}\beta_{-k}}{\beta_k} \right\} \ . \quad (13.6)$$

If $P_\theta\{X'_{-k}\beta^*_{-k} < 0\} > 0$, we can use Assumption **S4** to conclude that (13.5) is positive. If $P_\theta\{X'_{-k}\beta^*_{-k} \leq 0\} > 0$, as before, we conclude (13.6) is positive using Assumption **S4**.

Case 3. Suppose $\beta_k^* > 0$. Then,

$$P_\theta \left\{ X'\beta^* < 0 \leq X'\beta \right\} = P_\theta \left\{ -\frac{X'_{-k}\beta_{-k}}{\beta_k} \leq X_k < -\frac{X'_{-k}\beta^*_{-k}}{\beta_k^*} \right\} \quad (13.7)$$

and

$$P_\theta \left\{ X'\beta < 0 \leq X'\beta^* \right\} = P_\theta \left\{ -\frac{X'_{-k}\beta^*_{-k}}{\beta_k^*} \leq X_k < -\frac{X'_{-k}\beta_{-k}}{\beta_k} \right\} \ . \quad (13.8)$$

As we did in Case 2, we can prove that (13.7) or (13.8) is positive using Assumption **S4**. Thus, we only need to prove that

$$P_\theta \left\{ \frac{X'_{-k}\beta_{-k}}{\beta_k} > \frac{X'_{-k}\beta^*_{-k}}{\beta_k^*} \right\} > 0$$

or

$$P_\theta \left\{ \frac{X'_{-k}\beta_{-k}}{\beta_k} < \frac{X'_{-k}\beta^*_{-k}}{\beta^*_k} \right\} > 0 \ .$$

Let us assume by contradiction that the probabilities above are equal to zero. This implies

$$P_\theta \left\{ \frac{X'_{-k}\beta_{-k}}{\beta_k} = \frac{X'_{-k}\beta^*_{-k}}{\beta^*_k} \right\} = 1 \ ,$$

which is equivalent to

$$P_\theta \left\{ X'_{-k} \left(\frac{\beta_{-k}}{\beta_k} - \frac{\beta^*_{-k}}{\beta^*_k} \right) = 0 \right\} = 1 \ .$$

By Assumption **S2**, that said that there is no proper linear subspace that contains X, we conclude

$$\frac{\beta_{-k}}{\beta_k} = \frac{\beta^*_{-k}}{\beta^*_k} \ .$$

This implies that β^* is a scalar multiple of β. By Assumption **S3**, we conclude $\beta = \beta^*$, but this is a contradiction. This complete the proof of this case. ■

13.2 Estimation of the Linear Index Model

13.2.1 Estimation of parametric binary model

The semi-parametric assumptions **S1**-**S4** and Theorem 13.1 implies that the parameter β is identified. However, this result is not enough to identify the marginal effects. Now let us explore how to estimate β in the parametric case and how can we interpret it.

Let us consider the parametric case. That is

$$P\{Y = 1|X\} = F(X'\beta) \ ,$$

where $F(\cdot)$ can be the Probit function, $F(x) = \Phi(x)$, or the Logit function, $F(x) = \frac{\exp(x)}{1+\exp(x)}$. Suppose we have a random sample of size n from the distribution (Y, X); this is $(Y_1, X_1), \ldots, (Y_n, X_n)$. Since we have a parametric model, we can use the maximum likelihood estimator. Let us write the likelihood of the observation Y_i:

$$f_\beta(Y_i|X_i) = F(X'_i\beta)^{Y_i}(1 - F(X'_i\beta))^{1-Y_i} \ .$$

We can use this expression to write the log-likelihood of the random sample:

$$\begin{aligned}\ell_n(b) =& \frac{1}{n}\sum_{i=1}^{n} \ln\left(f_b(Y_i|X_i)\right) \\ =& \frac{1}{n}\sum_{i=1}^{n} \left\{Y_i \ln\left(F(X_i'\beta)\right) + (1-Y_i)\ln\left(1-F(X_i'\beta)\right)\right\} \ .\end{aligned}$$

It can be shown that β is the unique maximizer of $Q(b) = E[\ell_n(b)]$. Let us denote by $\hat{\beta}_n$ the maximum likelihood estimator (MLE). The asymptotic normality of the MLE implies

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} N(0, \mathbb{V}) \ ,$$

where $\mathbb{V} = \mathbb{I}_\beta^{-1}$ and

$$\mathbb{I}_\beta = -E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\ln\left(f_\beta(Y_i|X_i)\right)\right]$$

is the Fisher information matrix. By the information equality, we can rewrite the expression above as follows

$$\mathbb{I}_\beta = E\left[\frac{\partial}{\partial\beta}\ln\left(f_\beta(Y_i|X_i)\right)\frac{\partial}{\partial\beta'}\ln\left(f_\beta(Y_i|X_i)\right)\right] \ .$$

Since

$$\frac{\partial}{\partial\beta}\ln\left(f_\beta(Y_i|X_i)\right) = \left[\frac{Y_i - F(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))}\right]F'(X_i'\beta)X_i \ ,$$

we can rewrite one more time the information matrix as follows

$$\begin{aligned}\mathbb{I}_\beta =& E\left[\left[\frac{Y_i - F(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))}\right]^2 F'(X_i'\beta)^2 X_i X_i'\right] \\ =& E\left[\frac{F'(X_i'\beta)^2}{F(X_i'\beta)(1-F(X_i'\beta))}X_i X_i'\right] \ ,\end{aligned}$$

where the second equality above follows from the law of iterated expectations and law of total variance.

This final expression implies that we can estimate the asymptotic variance, $\mathbb{I}_\beta^{-1}$. This estimation can be done using the MLE and the sample analogue to compute the expected value. Note that this implies that we can do inference on β, but nothing yet about the inference on the marginal effects.

How can we interpret β?

Let us assume that X_j is continuously distributed. In the linear regression with $E[U|X] = 0$, we had

$$\frac{\partial E[Y|X]}{\partial X_j} = \beta_j \ .$$

In this case, β_j was capturing the marginal effect of X_j on Y. In the Binary models we rather have that the marginal effect is non-linear and depends on X:

$$\begin{aligned} \frac{\partial E[Y|X]}{\partial X_j} &= \frac{\partial P\{Y = 1|X\}}{\partial X_j} \\ &= F'(X'\beta)\beta_j \ , \end{aligned}$$

where $F'(\cdot)$ is the derivative of F. In the case of the Probit, we obtain

$$\frac{\partial P\{Y = 1|X\}}{\partial X_j} = \phi(X'\beta)\beta_j \ ,$$

and in the case of the Logit, we have

$$\frac{\partial P\{Y = 1|X\}}{\partial X_j} = F(X'\beta)(1 - F(X'\beta))\beta_j \ .$$

Note that the marginal effect of X_j on $E[Y|X]$ depends on the linear index $X'\beta$ and β_j. However, we can still extract information by simply inspecting β. For instance, we can use the ratio between β_j and β_k to obtain the ratio of the partial effects, since we have

$$\frac{\frac{\partial P\{Y=1|X\}}{\partial X_j}}{\frac{\partial P\{Y=1|X\}}{\partial X_k}} = \frac{\beta_j}{\beta_k} \ .$$

Also, because $F(\cdot)$ is an increasing function, we can conclude that the sign β_j identifies the sign of the marginal effect of X_j on $E[Y|X]$. Finally, it is possible to obtain upper bounds on the marginal effects from β using that $F'(\cdot)$ is bounded. In the case of the Probit model, we obtain

$$\frac{\partial P\{Y = 1|X\}}{\partial X_j} \leq 0.4\beta_j \quad \text{since } \phi(x) \leq \phi(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4 \ ,$$

and in the case of the Logit model,

$$\frac{\partial P\{Y = 1|X\}}{\partial X_j} \leq \frac{1}{4}\beta_j \quad \text{since } F(x)(1 - F(x)) \leq \frac{1}{4} \ .$$

13.2.2 Estimation of marginal effects

Let us define and compute the average/mean marginal effect as follows

$$E\left[\frac{\partial P\{Y=1|X\}}{\partial X_j}\right] = E\left[F'(X'\beta)\right]\beta_j \ .$$

We can estimate this quantity using the sample analogue and the MLE:

$$\frac{1}{n}\sum_{i=1}^{n} F'\left(X_i'\hat{\beta}_n\right)\hat{\beta}_{n,j} \ .$$

We can also compute the marginal effects "at the average", this is defined by

$$F'(E[X]'\beta)\beta_j \ ,$$

and can be estimated by

$$F'(\bar{X}_n'\hat{\beta}_n)\hat{\beta}_{n,j} \ .$$

Stata offers both options with the option margins. Note that these two quantities, average marginal effect and the marginal effect at the average, are different. The second one could make sense if there is meaning behind the evaluation of the effect at the average of the sample, but often this is not the case. For instance, if X_j is a binary variable (e.g. gender).

It is important to note that we computed the average marginal effect of X_j assuming this variable was continuously distributed. Now, let us focus to the case in which this variable is binary. In particular, let us consider the partition $X = (X_1, D)$, where $X_1 \in \mathbf{R}^k$ and $D \in \{0,1\}$. Also, let us consider the partition for $\beta = (\beta_1, \beta_2)$ accordingly. In this case, the following expression

$$E\left[\frac{\partial P\{Y=1|X\}}{\partial D}\right] = E\left[F'(X'\beta)\right]\beta_2$$

does not make a lot of sense since D take only two values. Instead, we can consider the following marginal effect of D,

$$P\{Y=1|X_1, D=1\} - P\{Y=1|X_1, D=0\} = F(X_1'\beta_1 + \beta_2) - F(X_1'\beta_1) \ .$$

Then, we can define the average marginal effect of D equal to

$$E\left[F(X_1'\beta_1 + \beta_2) - F(X_1'\beta_1)\right] \ ,$$

which can be estimated by the sample analogue and the MLE:

$$\frac{1}{n}\sum_{i=1}^{n} F(X_1'\hat{\beta}_{n,1} + \hat{\beta}_{n,2}) - F(X_1'\hat{\beta}_{n,1}) \ .$$

We can compute and report the standard errors for those estimated marginal effects. Let us remember that for the continuous case, we derived

$$\frac{\partial P\{Y=1|X\}}{\partial X_j} \;=\; F'(X'\beta)\beta_j \;,$$

which is a known function of β. For the discrete case, we obtained $F(X_1'\beta_1 + \beta_2) - F(X_1'\beta_1)$ which is also a known function of β. This implies that we can compute standard errors via the Delta Method. We can also compute the marginal effect on the treated by only conditioning on $D = 1$. *Stata* has options for this; see *margins* for more details.

Logit and the odds ratio interpretation

In statistic and Biostatistic, the logit model has particular appeal. Let $p_i = P\{Y_i = 1|X_i\}$. Using the parametric form of the Logit,

$$p_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \;,$$

we can conclude an equation for the odds ratio or relative risk,

$$\frac{p_i}{1-p_i} = \exp(X_i'\beta) \;,$$

which, after taking logs, is equivalent to

$$\ln\left(\frac{p_i}{1-p_i}\right) = X_i'\beta \;.$$

In this case, we can interpret β_j as the marginal effect of X_j on the log odds ratio. For example, suppose a clinical trial, denote by $Y = 1$ if you live and $Y = 0$ if you die. An odds ratio of 2 means that the odds of survival are twice those of death. Now, if $\beta_j = 0.1$, it means the relative probability of survival increases by 10% (roughly) if X_j increase in one unit.

13.3 Linear probability model

Some people still advocate the use of the Linear Probability Model (LPM) where

$$Y = X'\beta + U$$

and $E[U|X] = 0$. The main reason for this stance is that β in the linear case is a well-studied model and directly delivers "marginal effects". Moreover, the linear model easily accommodates the analysis of instrumental variables, panel with fixed effects, etc. Finally, and as we discussed in Lecture 5, if Y is binary and there are heterogeneous effects, the TSLS estimator admits

a LATE interpretation. All these possible extensions discussed above are hard to implement together in the Probit/Logit model.

However, it is hard to interpret the linear probability model causally as $E[Y|X]$ cannot be linear in most cases (e.g. Probit/Logit model). As we discuss in Lecture 1, the causal interpretation requires to believe in the existence of a model and this is not the case of the LPM. This is usually recognized by some of their supporters who claim:

> The true $E[Y|X]$ may arise from a causal model, but the regression is only providing a linear approximation to the true $E[Y|X]$.

This suggests that the LPM follows the second interpretation of the linear regression presented in Lecture 1. This means that LPM is a descriptive tool that approximate $E[Y|X]$ rather than a model that admit a causal interpretation.

The linear probability model delivers predicted probabilities outside $[0, 1]$, which makes it internally inconsistent as a model. A well-known textbook that support this approach recognize this issue. In Angrist and Pischke (2008, p. 103) appears textually

> ...[linear regression] may generate fitted values outside the limited dependent variable boundaries. This fact bothers some researchers and has generated a lot of bad press for the linear probability model.

Angrist and Pischke (2008) acknowledge that there are available approaches for the binary choice model, which admit a causal interpretation and are different than the LPM. However, in page 197, they add about this point the following

> "Yet we saw that the added complexity and extra work required to interpret the results from latent index models may not be worth the trouble"

At the very least, this statement may be controversial.

Remark 13.2 It is expected that Logit, Probit, and LPM yield quite different estimates $\hat{\beta}_n$. For instance, if we use the upper bounds for marginal effects, we get

$$\begin{aligned}\hat{\beta}_{\text{logit}} &\approx 4\hat{\beta}_{\text{ols}} \\ \hat{\beta}_{\text{probit}} &\approx 2.5\hat{\beta}_{\text{ols}} \\ \hat{\beta}_{\text{logit}} &\approx 1.6\hat{\beta}_{\text{probit}}\end{aligned}$$

However, average marginal effects from Logit, Probit, and even LPM are often "close", partly due because there is averaging going on.

The binary choice model discussed here using the linear index model is an idea that is applied to other settings. For instance, ordered choice models, where individual decides how many units to buy from the same item, or unordered choice models, where individual decides to buy one of many different alternatives. In these kind of models, it is common to find conditional Logit and multinomial Logit. The most popular example in Industrial Organization (IO) is the random coefficient logit model introduced by Berry et al. (1995), which is also known as BLP and is useful to estimate demand. These topics are covered in second year IO classes.

Bibliography

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Manski, C. F. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729–738.

Part III

A Primer on Inference and Standard Errors

Lecture 14

Heteroskedastic-Consistent Variance Estimation

14.1 Setup and notation

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ . \tag{14.1}$$

Suppose that $E[XU] = 0$, that there is no perfect collinearity in X, that $E[XX'] < \infty$, and that $\text{Var}[XU] < \infty$. Denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sample of random vectors with distribution P. Under these assumptions, we established the asymptotic normality of the OLS estimator, $\hat{\beta}_n$, this is

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

for

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} \ .$$

We wish to test

$$H_0 : \beta \in \mathbf{B}_0 \quad \text{versus} \quad H_1 : \beta \in \mathbf{B}_1$$

where $\mathbf{B}_0$ and $\mathbf{B}_1$ form a partition of $\mathbf{R}^{k+1}$, paying particular attention to hypotheses for one of the components of β. Without loss of generality, assume we are interested in the first slope component of β so that,

$$H_0 : \beta_1 = c \quad \text{versus} \quad H_1 : \beta_1 \neq c \ . \tag{14.2}$$

The CMT implies that

$$\sqrt{n}(\hat{\beta}_{1,n} - \beta_1) \xrightarrow{d} N(0, V_1)$$

as $n \to \infty$ where $V_1 = \mathbb{V}_{[2,2]}$ is the element of $\mathbb{V}$ corresponding to β_1. A natural choice of test statistic for this problem is the absolute value of the t-statistic,

$$t_{\text{stat}} = \frac{\sqrt{n}(\hat{\beta}_{1,n} - c)}{\sqrt{\hat{V}_{1,n}}} ,$$

so that $T_n = |t_{\text{stat}}|$. In order for t_{stat} to be asymptotically standard normal under the null hypothesis, we need a consistent estimator $\hat{\mathbb{V}}_n$ of the limiting variance $\mathbb{V}$. In this part of the course we will cover consistent estimators of $\mathbb{V}$ under different assumptions on the dependence and heterogeneity in the data. We will, however, start with the usual i.i.d. setting, where one of such estimators is

$$\hat{\mathbb{V}}_n = \left(\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \hat{U}_i^2 \right) \left(\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \right)^{-1} ,$$

where $\hat{U}_i = Y_i - X_i' \hat{\beta}_n$. This is the most widely used form of the robust, heteroskedasticity-consistent standard errors and it is associated with the work of White (1980) (see also Eicker, 1967; Huber, 1967). We will refer to these as robust EHW (or HC) standard errors.

14.2 Consistency of HC standard errors

We now prove that $\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V}$. The main difficulty lies in showing that

$$\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \hat{U}_i^2 \xrightarrow{P} \text{Var}[XU]$$

as $n \to \infty$.

Note that

$$\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \hat{U}_i^2 = \frac{1}{n} \sum_{1 \le i \le n} X_i X_i' U_i^2 + \frac{1}{n} \sum_{1 \le i \le n} X_i X_i' (\hat{U}_i^2 - U_i^2) .$$

Under the assumption that $\text{Var}[XU] < \infty$, the first term on the righthand side of the preceding display converges in probability to $\text{Var}[XU]$. It therefore suffices to show that the second term on the righthand side of the preceding display converges in probability to zero. We argue this separately for each of the $(k+1)^2$ terms. To this end, note for any $0 \le j \le k$ and $0 \le j' \le k$ that

$$\begin{aligned} \left| \frac{1}{n} \sum_{1 \le i \le n} X_{i,j} X_{i,j'} (\hat{U}_i^2 - U_i^2) \right| &\le \frac{1}{n} \sum_{1 \le i \le n} |X_{i,j} X_{i,j'}| |\hat{U}_i^2 - U_i^2| \\ &\le \frac{1}{n} \sum_{1 \le i \le n} |X_{i,j} X_{i,j'}| \max_{1 \le i \le n} |\hat{U}_i^2 - U_i^2| . \end{aligned}$$

Because $E[XX'] < \infty$, we have that $E[|X_j X_{j'}|] < \infty$. Hence,

$$\frac{1}{n} \sum_{1 \leq i \leq n} |X_{i,j} X_{i,j'}| = O_P(1) \ ,$$

so it suffices to show that

$$\max_{1 \leq i \leq n} |\hat{U}_i^2 - U_i^2| = o_P(1) \ .$$

For this purpose, the following lemma will be useful:

Lemma 14.1 *Let $Z_1, \ldots, Z_n$ be an i.i.d. sequence of random vectors such that $E[|Z_i|^r] < \infty$. Then $\max_{1 \leq i \leq n} |Z_i| = o_P(n^{\frac{1}{r}})$, i.e.,*

$$n^{-\frac{1}{r}} \max_{1 \leq i \leq n} |Z_i| \xrightarrow{P} 0 \ .$$

PROOF: Let $\epsilon > 0$ be given. Note that

$$\begin{aligned} P\{n^{-\frac{1}{r}} \max_{1 \leq i \leq n} |Z_i| > \epsilon\} &= P\{ \bigcup_{1 \leq i \leq n} \{|Z_i|^r > \epsilon^r n\}\} \\ &\leq \sum_{1 \leq i \leq n} P\{|Z_i|^r > \epsilon^r n\} \\ &\leq \frac{1}{n\epsilon^r} \sum_{1 \leq i \leq n} E[|Z_i|^r I\{|Z_i|^r > \epsilon^r n\}] \\ &= \frac{1}{\epsilon^r} E[|Z_i|^r I\{|Z_i|^r > \epsilon^r n\}] \\ &\to 0 \end{aligned}$$

as $n \to \infty$, where the first equality follows by inspection, the first inequality follows from Bonferonni's inequality, the second inequality follows from Markov's inequality, the final equality follows from the i.i.d. assumption, and the convergence to zero follows from the assumption that $E[|Z_i|^r] < \infty$. ■

We now use Lemma 14.1 to establish the desired convergence in probability to zero. Note that $E[|X|^2] < \infty$ (which follows from the fact that $E[XX'] < \infty$) and $E[|UX|^2] < \infty$ (which follows from the fact that $\mathrm{Var}[XU] < \infty$). Recall that $\hat{U}_i = U_i - X_i'(\hat{\beta}_n - \beta)$, so that

$$|\hat{U}_i^2 - U_i^2| \leq 2|U_i||X_i||\hat{\beta}_n - \beta| + |X_i|^2 |\hat{\beta}_n - \beta|^2 \ .$$

Next, note that Lemma 14.1 and the fact that $\sqrt{n}(\hat{\beta}_n - \beta) = O_P(1)$ imply that

$$\begin{aligned} |\hat{\beta}_n - \beta| \max_{1 \leq i \leq n} |U_i||X_i| &= o_P(1) \\ |\hat{\beta}_n - \beta|^2 \max_{1 \leq i \leq n} |X_i|^2 &= o_P(1) \ . \end{aligned}$$

The desired conclusion thus follows. If we combine this result with

$$\frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' \xrightarrow{P} E[XX'] \ ,$$

it follows immediately that

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V} \ .$$

Let $\hat{V}_{1,n}$ denote the $(2,2)$-diagonal element of $\hat{\mathbb{V}}_n$ - i.e., the entry corresponding to β_1. It follows that the test that rejects H_0 in (14.2) when

$$T_n = |t_{\text{stat}}| = \left| \frac{\sqrt{n}(\hat{\beta}_{1,n} - c)}{\sqrt{\hat{V}_{1,n}}} \right|$$

exceeds $z_{1-\frac{\alpha}{2}}$, is consistent in levels. As before, using the duality between hypothesis testing and the construction of confidence regions, we may construct a confidence region of level α for β_1 as

$$\begin{aligned} C_n &= \left\{ c \in \mathbf{R} : \left| \frac{\sqrt{n}(\hat{\beta}_{1,n} - c)}{\sqrt{\hat{V}_{1,n}}} \right| \leq z_{1-\frac{\alpha}{2}} \right\} \\ &= \left\{ \hat{\beta}_{1,n} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{V}_{1,n}}{n}}, \hat{\beta}_{1,n} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{V}_{1,n}}{n}} \right\} \ . \end{aligned}$$

This confidence region satisfies

$$P\{\beta_1 \in C_n\} \to 1 - \alpha$$

as $n \to \infty$.

It is worth noting that `Stata` does not compute $\hat{\mathbb{V}}_n$ in the default "robust" option, but rather a version of this estimator that includes a finite sample adjustment to "inflate" the estimated residuals (known to be too small in finite samples). This version of the HC estimator is commonly known as $HC1$ and given by

$$\hat{\mathbb{V}}_{\text{hc1,n}} = \left(\frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' \right)^{-1} \left(\frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' \hat{U}_i^{*2} \right) \left(\frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' \right)^{-1} \ ,$$

where $\hat{U}_i^{*2} = \frac{n}{n-k-1}\hat{U}_i^2$. It is immediate to see that this estimator is also consistent for $\mathbb{V}$. With the obvious modification for the components β_j, $1 \leq j \leq k$, and using HC1 standard errors, these are the "robust" confidence intervals reported by `Stata`. Other versions, including the one discussed in the next section are also available as an option.

The consistency of the standard errors does not necessarily translate into accurante finite sample inference on β in general, something that lead to a number of finite sample adjustments that are sometimes used in practice. The simplest one is the HC1 correction, although better alternatives are available. Below we discuss some of these adjustments.

14.3 Improving finite sample performance: HC2

An alternative to $\hat{\mathbb{V}}_n$ and $\hat{\mathbb{V}}_{\text{hc1,n}}$ is what MacKinnon and White (1985) call the HC2 variance estimator, here denoted by $\hat{\mathbb{V}}_{\text{hc2,n}}$. In order to define this estimator, we need additional notation. Let

$$\mathbb{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

be the $n \times n$ projection matrix, with i-th column denoted by

$$P_i = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}X_i$$

and (i,i)-th element denoted by

$$P_{ii} = X_i'(\mathbb{X}'\mathbb{X})^{-1}X_i \ .$$

Let Ω be the $n \times n$ diagonal matrix with i-th diagonal element equal to $\sigma^2(X_i) = \text{Var}[U_i|X_i]$, and let $e_{n,i}$ be the n-vector with i-th element equal to one and all other elements equal to zero. Let $\mathbb{I}$ be the $n \times n$ identity matrix and $\mathbb{M} = \mathbb{I} - \mathbb{P}$ be the residual maker matrix. The residuals $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$ can be written as

$$\hat{U}_i = e_{n,i}'\mathbb{M}\mathbb{U} \ , \text{ or, in vector form, } \hat{\mathbb{U}} = \mathbb{M}\mathbb{U} \ . \tag{14.3}$$

The (conditional) expected value of the square of the residual is

$$\begin{aligned} E[\hat{U}_i^2|X_1,\ldots,X_n] &= E[(e_{n,i}'\mathbb{M}\mathbb{U})^2|X_1,\ldots,X_n] \\ &= (e_{n,i} - P_i)'\Omega(e_{n,i} - P_i) \ . \end{aligned}$$

If we further assume homoskedasticity (i.e., $\text{Var}[U|X] = \sigma^2$), the last expression reduces to

$$E[\hat{U}_i^2|X_1,\ldots,X_n] = \sigma^2(1 - P_{ii}) \ ,$$

by exploiting that $\mathbb{P}$ is an idempotent matrix. In other words, even when the error term U is homoskedastic, the LS residual $\hat{U}$ is heteroskedastic (due to the presence of P_{ii}). Moreover, since it can be shown that $\frac{1}{n} \leq P_{ii} \leq 1$, it follows that $\text{Var}[\hat{U}_i]$ underestimates σ^2 under homoskedasticity. This discussion makes it natural to consider

$$\tilde{U}_i^2 \equiv \frac{\hat{U}_i^2}{1 - P_{ii}} \ , \tag{14.4}$$

as the squared residual to use in variance estimation as $\tilde{U}_i^2$ is unbiased for $E[U_i^2|X_1,\ldots,X_n]$ under homoskedasticity. This is the motivation for the variance estimator MacKinnon and White (1985) introduce as HC2,

$$\hat{\mathbb{V}}_{\text{hc2,n}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \tilde{U}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} , \quad (14.5)$$

where $\tilde{U}_i^2$ is as in (14.4). Under heteroskedasticity this estimator is unbiased only in some simple examples (we will cover one of these next class), but it is biased in general. However, it is expected to have lower bias relative to HC/HC1 - a statement supported by simulations.

There are other finite sample adjustments that give place to HC3, HC4, and even HC5. For example, HC3 is equivalent to HC2 with

$$\tilde{U}_i^{*2} \equiv \frac{\hat{U}_i^2}{(1-P_{ii})^2} , \quad (14.6)$$

replacing $\tilde{U}_i^2$, and its justification is related to the Jackknife estimator of the variance of $\hat{\beta}_n$. However, we will not consider these in class as these adjustments do not deliver noticeable additional benefits relative to HC2 (at least for the purpose of this class). It is worth noting that HC2 and HC3 are available as an option in `Stata`.

14.4 The Behrens-Fisher Problem

The Behrens-Fisher problem is that of comparing the means of two populations when the ratio of their variances is unknown and the distributions are assumed normal, i.e.,

$$Y(0) \sim N(\mu_0, \sigma^2(0)) \text{ and } Y(1) \sim N(\mu_1, \sigma^2(1)) . \quad (14.7)$$

We know from previous results that this problem can be viewed as a special case of linear regression with a binary regressor, i.e. $X = (1, D)$ and $D \in \{0, 1\}$. In this case, the coefficient on D identifies the average treatment effect, which in this case equals precisely $\mu_1 - \mu_0$. To be specific, consider the linear model

$$Y = X'\beta + U = \beta_0 + \beta_1 D + U$$

where

$$Y = Y(1)D + (1-D)Y(0) ,$$

and U is assumed to be normally distributed conditional on D, with zero conditional mean and

$$\text{Var}[U|D=d] = \sigma^2(d) \text{ for } d \in \{0,1\} .$$

We are interested in

$$\beta_1 = \frac{\text{Cov}(Y, D)}{\text{Var}(D)} = E[Y|D = 1] - E[Y|D = 0] \ .$$

Because D is binary, the least squares estimator of β_1 can be written as

$$\hat{\beta}_{1,n} = \bar{Y}_1 - \bar{Y}_0 \ , \tag{14.8}$$

where for $d \in \{0, 1\}$,

$$\bar{Y}_d = \frac{1}{n_d} \sum_{i=1}^{n} Y_i I\{D_i = d\} \text{ and } n_d = \sum_{i=1}^{n} I\{D_i = d\} \ .$$

Conditional on $D^{(n)} = (D_1, \ldots, D_n)$, the exact finite sample variance of $\hat{\beta}_{1,n}$ is

$$V_1^* = \text{Var}[\hat{\beta}_{1,n}|D^{(n)}] = \frac{\sigma^2(0)}{n_0} + \frac{\sigma^2(1)}{n_1} \ ,$$

so that, under normality, it follows that

$$\hat{\beta}_{1,n}|D^{(n)} \sim N\left(\beta_1, \frac{\sigma^2(0)}{n_0} + \frac{\sigma^2(1)}{n_1}\right) \ .$$

The problem of how to do inference on β_1 in the absence of knowledge of $\sigma^2(d)$ in this context is old, and known as the Behrens-Fisher problem. In particular, the question is whether there exists $\kappa \in \mathbf{R}$ such that for some estimator $\hat{V}_{1,n}^*$ we get

$$\frac{\hat{\beta}_{1,n} - \beta_1}{\sqrt{\hat{V}_{1,n}^*}} \sim t(\kappa) \ , \tag{14.9}$$

where $t(\kappa)$ denotes a t-distribution with κ degrees of freedom (dof). We explore this question below under different assumptions.

Comment on notation. Today we are framing the discussion around the "actual" conditional variance of $\hat{\beta}_{1,n}$ as opposed to the asymptotic variance. This means that the estimator $\hat{V}_{1,n}^*$ above is an estimator of such variance (which also explains why there is no $\sqrt{n}$ in the numerator of (14.9)). Of course, if $\hat{V}_{1,n}$ is a consistent estimator of the asymptotic variance of $\hat{\beta}_{1,n}$, then $\hat{V}_{1,n}^* = \frac{1}{n}\hat{V}_{1,n}$ is an estimator of the variance of $\hat{\beta}_{1,n}$. I will use $*$ to denote finite sample variances.

14.4.1 The homoskedastic case

Suppose the errors are homoskedastic: $\sigma^2 = \sigma^2(0) = \sigma^2(1)$, so that the exact conditional variance of $\hat{\beta}_{1,n}$ is

$$V_1^* = \sigma^2 \left(\frac{1}{n_0} + \frac{1}{n_1}\right) \ .$$

In this case, we can estimate σ^2 by

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

and let

$$\hat{V}^*_{1,\text{ho}} = \hat{\sigma}^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right) \ , \tag{14.10}$$

be the estimator of V_1^*. This estimator has two important features.

(a) **Unbiased**. Since $\hat{\sigma}^2$ is unbiased for σ^2, it follows that $\hat{V}^*_{1,\text{ho}}$ is *unbiased* for the true variance V_1^*.

(b) **Chi-square**. Under normality of U given D, the scaled distribution of $\hat{V}^*_{1,\text{ho}}$ is chi-square with $n-2$ dof,

$$(n-2)\frac{\hat{V}^*_{1,\text{ho}}}{V_1^*} \sim \chi^2(n-2) \ .$$

It follows that, under normality of U given D, the t-stat has an exact t-distribution under the null hypothesis in (14.2),

$$t_{\text{ho}} = \frac{\hat{\beta}_{1,n} - c}{\sqrt{\hat{V}^*_{1,\text{ho}}}} \sim t(n-2) \ . \tag{14.11}$$

This t-distribution with dof equal to $n-2$ can be used to test (14.2) and, by duality, for the construction of exact confidence intervals, i.e.,

$$CS^{1-\alpha}_{\text{ho}} = \left\{\hat{\beta}_{1,n} - t^{n-2}_{1-\frac{\alpha}{2}}\sqrt{\hat{V}^*_{1,\text{ho}}}, \hat{\beta}_{1,n} + t^{n-2}_{1-\frac{\alpha}{2}}\sqrt{\hat{V}^*_{1,\text{ho}}}\right\} \ . \tag{14.12}$$

Here $t^{n-2}_{1-\frac{\alpha}{2}}$ denotes the $1-\frac{\alpha}{2}$ quantile of a t distributed random variable with $n-2$ dof. Such confidence interval is *exact* under these two assumptions, normality and homoskedasticity, and we can conclude that (14.9) holds with $\kappa = n-2$.

14.4.2 The robust EHW variance estimator

In the Behrens-Fisher example, the component of the EHW variance estimator $\frac{1}{n}\hat{\mathbb{V}}_n$ corresponding to β_1 simplifies to

$$\hat{V}^*_{1,\text{hc}} = \frac{\hat{\sigma}^2(0)}{n_0} + \frac{\hat{\sigma}^2(1)}{n_1}$$

where

$$\hat{\sigma}^2(d) = \frac{1}{n_d}\sum_{i=1}^{n}(Y_i - \bar{Y}_d)^2 I\{D_i = d\} \quad \text{for} \quad d \in \{0,1\} \ .$$

Unfortunately, however, there are no assumptions under which there exists a value of κ such that (14.9) holds, even when U is normally distributed conditional on D.

The standard, normal-distribution-based, $1-\alpha$ confidence interval based on the robust variance estimator is

$$CS_{\rm hc}^{1-\alpha} = \left\{ \hat{\beta}_{1,n} - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{1,\rm hc}^*}, \hat{\beta}_{1,n} + z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{1,\rm hc}^*} \right\} . \tag{14.13}$$

In small samples the properties of these standard errors are not always attractive: $\hat{V}_{1,\rm hc}^*$ is biased downward, i.e.,

$$E[\hat{V}_{1,\rm hc}^*] = \frac{n_0 - 1}{n_0} \frac{\sigma^2(0)}{n_0} + \frac{n_1 - 1}{n_1} \frac{\sigma^2(1)}{n_1} < V_1^* ,$$

and $CS_{\rm hc}^{1-\alpha}$ can have coverage substantially below $1-\alpha$. A common "correction" for this problem is to replace $z_{1-\frac{\alpha}{2}}$ with $t_{1-\frac{\alpha}{2}}^{n-2}$. However, as we will illustrate in the next section, such correction if often ineffective.

14.4.3 An unbiased estimator of the variance

An alternative to $\frac{1}{n}\hat{\mathbb{V}}_n$ is what MacKinnon and White (1985) call the HC2 variance estimator, here denoted by $\frac{1}{n}\hat{\mathbb{V}}_{\rm hc2,n}$. We learned that this estimator is unbiased under homoskedasticity and that, in general, it removes only part of the bias under heteroskedasticity. However, in the single binary regressor (Behrens-Fisher) case the MacKinnon-White HC2 correction removes the entire bias. Its form in this case is

$$\hat{V}_{1,\rm hc2}^* = \frac{\tilde{\sigma}^2(0)}{n_0} + \frac{\tilde{\sigma}^2(1)}{n_1} ,$$

where

$$\tilde{\sigma}^2(d) = \frac{1}{n_d - 1} \sum_{i=1}^{n} (Y_i - \bar{Y}_d)^2 I\{D_i = d\} .$$

These conditional variance estimators differ from $\hat{\sigma}^2(d)$ by a factor $n_d/(n_d - 1)$. In combination with the normal approximation to the distribution of the t-statistic, this variance estimator leads the following $1-\alpha$ confidence interval

$$CS_{\rm hc2}^{1-\alpha} = \left\{ \hat{\beta}_{1,n} - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{1,\rm hc2}^*}, \hat{\beta}_{1,n} + z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{1,\rm hc2}^*} \right\} . \tag{14.14}$$

The estimator $\hat{V}_{1,\rm hc2}^*$ is *unbiased* for V_1^*, but it does not satisfy the chi-square property in (b) above. As a result, the associated confidence interval is still not exact. Just as in the previous case, there are no assumptions under which there exists a value of κ such that (14.9) holds, even when U

	dof	$\sigma^2(0)=0$	$\sigma^2(0)=1$	$\sigma^2(0)=2$
$\hat{V}^*_{\text{ho}}$	∞	72.5	94.0	99.8
	$n-2$	74.5	95.0	99.8
$\hat{V}^*_{1,\text{hc}}$	∞	76.8	80.5	86.6
	$n-2$	78.3	82.0	88.1
$\hat{V}^*_{1,\text{hc2}}$	∞	82.5	85.2	89.8
	$n-2$	83.8	86.5	91.0

Table 14.1: Angrist-Pischke design. $n_1 = 3$, $n_0 = 27$.

is normally distributed conditional on D. In fact, in small samples these standard errors do not work very well.

Consider the following simple simulation, borrowed from Imbens and Kolesar (2012) and Angrist and Pischke (2008), where $n_1 = 3$, $n_0 = 27$,

$$U_i|D_i \sim N(0, \sigma^2(D_i)) \ ,$$

$\sigma^2(1) = 1$, $\sigma^2(0) \in \{0, 1, 2\}$, and $1 - \alpha = 0.95$. The results are reported in Table 14.1. From the table it is visible that the confidence intervals typically undercover. Note that the table also reports the usual finite-sample adjustment, i.e. replacing $z_{1-\frac{\alpha}{2}}$ with $t^{n-2}_{1-\frac{\alpha}{2}}$. However, in this single-binary-covariate case it is easy to see why $n-2$ may be a poor choice for the degrees of freedom for the approximating t-distribution. Suppose that there are many units with $D_i = 0$ and few units with $D_i = 1$ (say $n_1 = 3$ and $n_0 = 1,000,000$). In that case $E[Y_i|D_i = 0]$ is estimated relatively precisely, with variance $\sigma^2(0)/n_0 \approx 0$. As a result the distribution of the t-statistic is approximately equal to that of

$$\frac{\bar{Y}_1 - E[Y_i|D_i = 1]}{\sqrt{\tilde{\sigma}^2(1)/n_1}} \ .$$

The latter has, under normality, an exact t-distribution with dof equal to $n_1 - 1 = 2$, substantially different from the t-distribution with $n - 2 \approx \infty$ dof. The question is whether we can figure out the appropriate dof in an automatic data dependent way, and this leads to topic of "degrees of freedom adjustment".

Bibliography

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Imbens, G. W. and M. Kolesar (2012): "Robust standard errors in small samples: some practical advice," Tech. rep., National Bureau of Economic Research.

MacKinnon, J. G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 817–838.

Lecture 15

Heteroskedasticity Autocorrelation Consistent Covariance Estimation

15.1 Setup and notation

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ . \tag{15.1}$$

Suppose that $E[U|X] = 0$, that there is no perfect collinearity in X, that $E[XX'] < \infty$, and that $\text{Var}[XU] < \infty$. We have already discussed Heteroskedasticity Consistent (HC) covariance matrix estimation, together with a number of adjustments intended to improve the performance in finite samples. Today we will consider the case where the sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ is not necessarily i.i.d. due to the presence of dependence across observations. In particular, we will introduce the term "autocorrelation", to denote the case where X_i and $X_{i'}$ may not be independent for $i \neq i'$. To derive an estimator of the variance covariance matrix that is consistent we need two tools: (a) appropriate LLNs and CLTs for dependent processes, and (b) and description of the object we intend to estimate. For the sake of simplicity, we will start assuming that $X_i = X_{1,i}$ is a scalar random variable that is naturally ordered (e.g., a time series) to then move to the general case.

15.2 Limit theorems for dependent data

This section will informally discuss the issues that arise when extending law of large numbers and central limit theorems to dependent data. The discussion here follows Mikusheva (2007) and Billingsley (1995).

Let's start with the LLN. When $\{X_i : 1 \leq i \leq n\}$ is a sequence of i.i.d. random variables with mean μ and variance σ_X^2, it follows that

$$\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] = \frac{\sigma_X^2}{n} \to 0 \ , \tag{15.2}$$

and so convergence in probability follows by a simple application of Chebyshev's inequality. However, without the independence, we need additional assumptions to control the variance of the average. We will start by assuming that the process we are dealing with are "stationary" as follows,

Definition 15.1 *A process $\{X_i : 1 \leq i \leq n\}$ is strictly stationarity if for each j, the distribution of $\{X_i, \ldots, X_{i+j}\}$ is the same for all i.*

Definition 15.2 *A process $\{X_i : 1 \leq i \leq n\}$ is weakly stationary if $E[X_i]$, $E[X_i^2]$, and, for each j, $\gamma_j \equiv \text{Cov}[X_i, X_{i+j}]$, do not depend on i.*

Assuming stationarity, the unique mean μ is well defined, and we can consider the variance of the sample average again,

$$\begin{aligned}
\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\text{Cov}[X_i, X_k] \\
&= \frac{1}{n^2}(n\gamma_0 + 2(n-1)\gamma_1 + 2(n-2)\gamma_2 + \cdots) \\
&= \frac{1}{n}\left(\gamma_0 + 2\sum_{j=1}^{n}\gamma_j\left(1 - \frac{j}{n}\right)\right) \ ,
\end{aligned}$$

where we have used the notation $\gamma_j = \text{Cov}[X_i, X_{i+j}]$, so that $\gamma_0 = \sigma_X^2$. We can immediately see that, for the variance to vanish, we need to make sure the last summation does not explode. A sufficient condition for this is absolute summability,

$$\sum_{j=-\infty}^{\infty} |\gamma_j| < \infty \ , \tag{15.3}$$

in which case a law of large numbers follows one more time from an application of Chebyshev's inequality.

Lemma 15.1 *If $\{X_i : 1 \leq i \leq n\}$ is a weakly stationary time series (with mean μ) with absolutely summable auto-covariances, then a law of large numbers holds (in probability and L2).*

Remark 15.1 Stationarity is not enough. Let $\zeta \sim N(0, \sigma_\zeta^2)$. Suppose $X_i = \zeta \ \forall i$. Then $\text{Cov}[X_i, X_{i'}] = \sigma_\zeta^2 \ \forall i, i'$, so we do not have absolute summability, and clearly we do not have a LLN since the average equals ζ, which is random.

Remark 15.2 Absolutely summability follows from mixing assumptions, i.e., assuming the sequence $\{X_i : 1 \leq i \leq n\}$ is α-mixing, see Billingsley (1995, Lemma 3, p. 365). The notion of α-mixing captures the dependence in the data as follows. Let α_n be a number such that

$$|P(A \cap B) - P(A)P(B)| \leq \alpha_n \ , \tag{15.4}$$

for any $A \in \sigma(X_1, \ldots, X_j)$, $B \in \sigma(X_{j+n}, X_{j+n+1}, \ldots)$, where $\sigma(X)$ is the σ-field generated by X, and $j \geq 1$, $n \geq 1$. If $\alpha_n \to 0$ as $n \to \infty$, the sequence is then said to be α-mixing, the idea being that X_j and X_{j+n} are then approximately independent for large n.

From the new proof of LLN one can guess that the variance in a central limit theorem should change. Remember that we wish to normalize the sum in such a way that the limit variance would be 1. Note that

$$\begin{aligned} \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right] &= \gamma_0 + 2\sum_{j=1}^{n}\gamma_j\left(1 - \frac{j}{n}\right) \\ &\to \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j = \Omega \ , \end{aligned} \tag{15.5}$$

where Ω is called the long-run variance. There are many central limit theorems for serially correlated observations. Below we provide a commonly used version, see Billingsley (1995, Theorem 27.4) for a proof under slightly stronger assumptions.

Theorem 15.1 *Suppose that $\{X_i : 1 \leq i \leq n\}$ is a strictly stationary α_n-mixing stochastic process with $E[|X|^{2+\delta}] < \infty$, $E[X] = 0$, and*

$$\sum_{n=1}^{\infty} \alpha_n^{\delta/(2+\delta)} < \infty \ . \tag{15.6}$$

Then Ω in (15.5) *is finite (i.e.* (15.3) *holds) and, provided $\Omega > 0$,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{d} N(0, \Omega) \ . \tag{15.7}$$

15.3 Estimating long-run variances

Let's go back to the regression model in (15.1). In the case of i.i.d. data, one of the exclusion restrictions is formulated as $E[U_i|X_i] = 0$. However, whenever the data is potentially dependent (time series, panel data, clustered data), we have to describe the conditional mean relative to all variables that may be important. In particular, we say X_i is weakly exogenous if

$E(U_i|X_i, X_{i-1}, \dots) = 0$, where we are implicitly assuming the observations have a natural ordering (as it is the case for time series).

Now consider the LS estimator of β,

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i U_i \ . \tag{15.8}$$

Invoking Lemma 15.1 under appropriate assumption on $\{X_i : 1 \leq i \leq n\}$ gives us $\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{P} \Sigma_X = E[XX']$. In addition, assuming the conditions in Theorem 15.1 for $\eta_i \equiv X_i U_i$ gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_i \xrightarrow{d} N(0, \Omega) \ , \tag{15.9}$$

and thus

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1}) \ . \tag{15.10}$$

The only thing that is different from the usual sandwich formula is the "meat", i.e., $\Omega = \sum_{j=-\infty}^{\infty} \gamma_j$ where γ_j are now the autocovariances of η_i. This long-run variance is significantly harder to estimate than the usual variance-covariance matrices that arise under i.i.d. assumptions. We need to figure out how to estimate Ω.

Below we discuss the HAC approach. For simplicity, however, let's ignore the fact that in practice U_i will be replaced by a regression residual $\hat{U}_i$ (since such modification is easy to incorporate and follows similar steps to those in previous lectures).

15.3.1 A naive approach

We know Ω is the sum of all auto-covariances (an infinite number of them). However, we can only estimate $n-1$ of them with a sample of size n. What if we just use the ones we can estimate? This gives the following estimator,

$$\tilde{\Omega} \equiv \sum_{j=-(n-1)}^{n-1} \hat{\gamma}_j \ , \quad \hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j} \eta_i \eta_{i+j} \ . \tag{15.11}$$

Unfortunately, this does not result in a consistent estimator. To see this, note that

$$\begin{aligned}
\tilde{\Omega} &= \sum_{j=-(n-1)}^{n-1} \hat{\gamma}_j \\
&= \frac{1}{n}\sum_{j=-(n-1)}^{n-1}\sum_{i=1}^{n-j} \eta_i \eta_{i+j} \\
&= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_i\right)^2 \xrightarrow{d} (N(0,\Omega))^2 \ .
\end{aligned}$$

The problem is that we are summing too many imprecisely estimated covariances. So, the noise does not die out. For example, to estimate γ_{n-1} we use only one observation.

15.3.2 Simple truncation

Given the problem with the naive estimator, a natural question would be: what if we do not use all the covariances? This gives us a truncated estimator,

$$\bar{\Omega} \equiv \sum_{j=-m_n}^{m_n} \hat{\gamma}_j = \hat{\gamma}_0 + 2\sum_{j=1}^{m_n} \hat{\gamma}_j \ . \tag{15.12}$$

where $m_n < n$, $m_n \to \infty$, and $m_n/n \to 0$ as $n \to \infty$. First, we have to notice that due to truncation there will be a finite sample bias. As m_n increases, the bias due to truncation should be smaller and smaller. But we don't want to increase m_n too fast for the reason stated above (we don't want to sum up noises). Assume that we can choose m_n in such a way that this estimator is consistent. Then we might face another small sample problem: this estimator may be negative: $\bar{\Omega} < 0$ (or in vector case, $\bar{\Omega}$ not positive definite). To see this, take $m_n = 1$, so that $\bar{\Omega} = \hat{\gamma}_0 + 2\hat{\gamma}_1$. In small samples, we may find $\hat{\gamma}_1 < -\frac{1}{2}\hat{\gamma}_0$, then $\bar{\Omega}$ will be negative.

15.3.3 Weighting and truncation: the HAC estimator

Following Newey and West (1987), the renewed suggestion is to create a weighted sum of sample auto-covariances with weights guaranteeing positive-definiteness:

$$\hat{\Omega}_n \equiv \sum_{j=-(n-1)}^{n-1} k\left(\frac{j}{m_n}\right)\hat{\gamma}_j \ . \tag{15.13}$$

We need conditions on m_n and $k(\cdot)$ to give us consistency and positive-definiteness. First, $m_n \to \infty$ as $n \to \infty$, although not too fast. For the proof below we will assume that $m_n^3/n \to 0$, but the result can be proved under $m_n^2/n \to 0$ (see Andrews, 1991). On the other hand, $k(\cdot)$ needs to be such that it guarantees positive-definiteness by down-weighting high lag covariances, but we also need $k(j/m_n) \to 1$ as $n \to \infty$ for consistency. As with non-parametric density estimation, there exist a variety of kernels that satisfy all the properties needed for consistency and positive-definiteness. The first one proposed by Newey-West (1987), was the Barlett kernel, which is defined as follows.

Barlett Kernel (Newey and West, 1987)

$$k(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Parzen kernel (Gallant, 1987)

$$k(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{if } |x| \leq 1/2 \\ 2(1-|x|)^3 & \text{if } 1/2 \leq |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Quadratic spectral kernel (Andrews, 1991)

$$k(x) = \frac{25}{12\pi^2 x^2} \left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(\sin(6\pi x/5)) \right) .$$

These kernels are all symmetric at 0, where the first two kernels have a bounded support $[-1, 1]$, and the QS has unbounded support. For the Bartlett and Parzen kernels, the weight assigned to $\hat{\gamma}_j$ decreases with $|j|$ and becomes zero for $|j| \geq m_n$. Hence, m_n in these functions is also known as a truncation lag parameter. For the quadratic spectral, m_n does not have this interpretation because the weight decreases to zero at $|j| = 1.2m_n$, but then exhibits damped sine waves afterwards. Note that for the first two kernels, we can write

$$\hat{\Omega}_n \equiv \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right) \hat{\gamma}_j . \tag{15.14}$$

so that the truncation at m_n is explicit. In the results that follow below we will focus on this representation to simplify the intuition behind the formal arguments.

Theorem 15.2 *Assume that $\{\eta_i : 1 \leq i \leq n\}$ is a weakly stationary sequence with mean zero and autocovariances $\gamma_j = \mathrm{Cov}[\eta_i, \eta_{i+j}]$ that satisfy* (15.3)*. Assume that*

1. *$m_n \to \infty$ as $n \to \infty$ and $m_n^3/n \to 0$.*
2. *$k(x) : \mathbf{R} \to [-1, 1]$, $k(0) = 1$, $k(x)$ is continuous at 0, and $k(-x) = k(x)$.*
3. *For all j the sequence $\xi_{i,j} = \eta_i \eta_{i+j} - \gamma_j$ is stationary and*

$$\sup_j \sum_{k=1}^{\infty} |\mathrm{Cov}(\xi_{i,j}, \xi_{i+k,j})| < C$$

for some constant C (limited dependence).

Then, $\hat{\Omega}_n \xrightarrow{P} \Omega$.

I will provide a sketch of the proof below under the assumptions of the theorem above. Start by writing the difference between our estimator and the object of interest,

$$\hat{\Omega}_n - \Omega = -\sum_{|j|>m_n} \gamma_j + \sum_{j=-m_n}^{m_n} \left(k\left(\frac{j}{m_n}\right) - 1 \right) \gamma_j + \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right) (\hat{\gamma}_j - \gamma_j) \ .$$

The first terms represents a truncation error, is non-stochastic, and goes to zero as $m_n \to \infty$ by (15.3).

The second term is also non-stochastic and it represents an error from using a kernel (as opposed to uniform weights). If we let

$$f_n(j) \equiv \left| k\left(\frac{j}{m_n}\right) - 1 \right| |\gamma_j| \ ,$$

under condition 2 it follows that $f_n(j) \leq g(j) \equiv 2|\gamma_j|$, which, by (15.3), is summable. By the same condition, $f_n(j) \to f(j) = 0$ for all j and, invoking the dominated convergence theorem,

$$\left| \sum_{j=-m_n}^{m_n} \left(k\left(\frac{j}{m_n}\right) - 1 \right) \gamma_j \right| \leq \sum_{j=-m_n}^{m_n} \left| k\left(\frac{j}{m_n}\right) - 1 \right| |\gamma_j| \to 0 \ , \qquad (15.15)$$

so that the second term vanishes asymptotically.

Now consider the third term. This is the error from estimating the covariances and it is stochastic. Notice that for the first term we want m_n big enough to eliminate it. For this last term, we want m_n to be small enough.

Start by noting that $\hat{\gamma}_j = \frac{1}{n} \sum_{i=1}^{n-j} \eta_i \eta_{i+j}$ is biased for γ_j and

$$\gamma_j^* \equiv E[\hat{\gamma}_j] = \frac{n-j}{n} \gamma_j \ .$$

This bias disappears as $n \to \infty$, so we can split the last term in two parts,

$$\sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right) (\hat{\gamma}_j - \gamma_j^*) + \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right) (\gamma_j^* - \gamma_j) \qquad (15.16)$$

and conclude that the last term (which is non-stochastic) goes to zero by similar arguments to those in (15.15). It then follows that it suffices to show that

$$\left| \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right) (\hat{\gamma}_j - \gamma_j^*) \right| \leq \sum_{j=-m_n}^{m_n} |\hat{\gamma}_j - \gamma_j^*| \xrightarrow{P} 0 \ .$$

In order to do this, let $\xi_{i,j} = \eta_i \eta_{i+j} - \gamma_j$ so that

$$\frac{1}{n} \sum_{i=1}^{n-j} \xi_{i,j} = \hat{\gamma}_j - \gamma_j^* \ .$$

Simple albegra shows that

$$\begin{aligned} E[(\hat{\gamma}_j - \gamma_j^*)^2] &= \frac{1}{n^2} \sum_{i=1}^{n-j} \sum_{k=1}^{n-j} \mathrm{Cov}[\xi_{i,j}, \xi_{k,j}] \\ &\leq \frac{1}{n^2} \sum_{i=1}^{n-j} \sum_{k=1}^{n-j} |\mathrm{Cov}[\xi_{i,j}, \xi_{k,j}]| \\ &\leq \frac{1}{n^2} \sum_{i=1}^{n-j} C \\ &\leq \frac{C}{n} \ , \end{aligned}$$

where in the second inequality we used

$$\sup_{1 \leq j < \infty} \sum_{k=1}^{\infty} |\mathrm{Cov}(\xi_{i,j}, \xi_{i+k,j}| < C \ .$$

By Chebyshev's inequality,

$$P\left\{|\hat{\gamma}_j - \gamma_j^*| > \epsilon\right\} \leq \frac{E[(\hat{\gamma}_j - \gamma_j^*)^2]}{\epsilon^2} \leq \frac{C}{n\epsilon^2} \ , \tag{15.17}$$

where, importantly, the bound holds uniformly in $1 \leq j < \infty$. This characterize the accuracy with which we estimate each covariance. Now we need to assess how many auto-covariances we can estimate well simultaneously:

$$\begin{aligned} P\left\{\sum_{j=-m_n}^{m_n} |\hat{\gamma}_j - \gamma_j^*| > \epsilon\right\} &\leq P\left\{\bigcup_{j=-m_n}^{m_n} \{|\hat{\gamma}_j - \gamma_j^*| > \frac{\epsilon}{2m_n+1}\}\right\} \\ &\leq \sum_{j=-m_n}^{m_n} P\left\{|\hat{\gamma}_j - \gamma_j^*| > \frac{\epsilon}{2m_n+1}\right\} \\ &\leq \sum_{j=-m_n}^{m_n} \frac{E[(\hat{\gamma}_j - \gamma_j^*)^2](2m_n+1)^2}{\epsilon^2} \\ &\leq (2m_n+1)\frac{C(2m_n+1)^2}{n\epsilon^2} \\ &\leq C^* \frac{m_n^3}{n\epsilon^2} \to 0 \ , \end{aligned}$$

where the last step uses $m_n^3/n \to 0$. This completes the proof.

We have proved consistency but we have not addressed the question of positive definiteness of our HAC estimator. To do this, it is convenient to characterize positive definiteness using the Fourier transformation of $\hat{\Omega}$. We will skip this in class, but the interested reader should see Newey and West (1987).

Bandwidth choice. After the original paper by Newey-West (1987), a series of papers addressed the issue of bandwidth choice (notably, Andrews (1991)). The general idea here is that we are facing a bias-variance trade-off in the choice of bandwidth m_n (also called truncation lag). Namely, a bigger m_n reduces the cut-off bias, however, it increases the number of estimated covariances used (and hence the variance of the estimate). Andrews (1991) proposed to choose m_n by minimizing the mean squared error (MSE) of the HAC estimator,

$$MSE(\hat{\Omega}_n) = \text{bias}(\hat{\Omega}_n)^2 + \text{Var}(\hat{\Omega}_n) \ . \tag{15.18}$$

Andrews (1991) did this minimization and showed that the optimal bandwidth is $m_n = C^* n^{1/r}$, where $r = 3$ for the Barlett kernel and $r = 5$ for other kernels. He also provided values for the optimal constant C^*, that depends on the kernel used, among other things.

Bibliography

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Billingsley, P. (1995): *Probability and Measure*, Wiley-Interscience.

Mikusheva, A. (2007): "Course materials for Time Series Analysis," MIT OpenCourseWare, Massachusetts Institute of Technology.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–08.

Lecture 16

Cluster Covariance Estimation

Let (Y, X, U) be a random vector where Y and U takes values in $\mathbf{R}$ and X take values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ . \tag{16.1}$$

and assume that $E[XU] = 0$, that there is no perfect collinearity in X, that $E[XX'] < \infty$, and that $\text{Var}[XU] < \infty$.

To describe the sampling process, we use two indices as we did when we covered panel data. The first index, $j = 1, \ldots, q$, denotes what are known as "clusters". By clusters, we mean a group of observations that may be related to each other. For example, a cluster could be a family, a school, an industry, or a city. The second index, $i = 1, \ldots, n_j$, denotes units within a cluster. For example, family members, students, firms, or individuals. If we let

$$X_j = (X_{1,j}, \ldots, X_{n_j,j})'$$

be a $n_j \times (k+1)$ matrix of stacked observations for cluster j, and define Y_j and U_j analogously, we can write (16.1) as

$$Y_j = X_j\beta + U_j \ , \quad j = 1, \ldots, q \ ,$$

where $E[X_j'U_j] = 0$. We will assume that (Y_j, X_j) are independent across $j \leq q$ but wish to allow for the possibility that $(Y_{i,j}, X_{i,j})$ and $(Y_{i',j}, X_{i',j})$ may be dependent for $i \neq i'$ and same $j \leq q$. In terms of constructing valid tests for hypotheses on the parameter β, this problem translates into constructing standard errors that account for the fact that $(X_{i,j}, U_{i,j})$ and $(X_{i',j}, U_{i',j})$ may be correlated within a cluster. In order to do this, we start by presenting appropriate versions of the law of large numbers and the central limit theorem, following Hansen and Lee (2019).

16.1 Law of Large Numbers

We start by focusing on the sample mean of $X_{i,j}$. Since the least squares estimator is a function of sample means, doing so will prove useful to analysis the properties of the LS estimator of β later on. Note that we can write

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{q} X_j' \mathbf{1}_j = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j} X_{i,j} \ ,$$

where $\mathbf{1}_j$ is an n_j-dimensional vector of ones. Our first result is:

Theorem 16.1 *Suppose that as $n \to \infty$,*

$$\max_{j \geq q} \frac{n_j}{n} \to 0 \tag{16.2}$$

and that

$$\lim_{M \to \infty} \sup_{i,j} E\left[||X_{i,j}|| I\{||X_{i,j}||\} > M\right] = 0 \ . \tag{16.3}$$

Then, as $n \to \infty$

$$||\bar{X}_n - E\left[\bar{X}_n\right]|| \xrightarrow{P} 0 \ .$$

A few comments about this theorem are worth highlighting. First, the condition in (16.3) states that $X_{i,j}$ is uniformly integrable and this is a technical requirement that is usually assumed outside the i.i.d. setting. Second, the condition in (16.2) states that each cluster size n_j is asymptotically negligible. This automatically holds when n_j is fixed as $q \to \infty$, which is the traditional framework we discussed with panel data. It also implies that $q \to \infty$, so we do not explicitly list this as a condition.

Remark 16.1 The condition in (16.2) allows for considerable heterogeneity in cluster sizes and it allows the cluster sizes to grow with sample size, so long as the growth is not proportional. An example of this we will use multiple times below is $n_j = n^a$ for $0 \leq a < 1$, which leads to

$$n = \sum_{j=1}^{q} n_j = \sum_{j=1}^{q} n^a = q n^a \implies q = n^{1-a} \ .$$

Assumption (16.2) is necessary for parameter estimation consistency while allowing arbitrary within-cluster dependence. Otherwise a single cluster could dominate the sample average.

16.2 Rates of Convergence

Under i.i.d. sampling the rate of convergence of the sample mean is $n^{-1/2}$. That is

$$\sqrt{n}\left(\bar{X}_n - E\left[\bar{X}_n\right]\right) \xrightarrow{d} N(0, V) \ .$$

In the case of clustering data, the rate of convergence may or may not be affected. We will see in the example below that often is affected. For instance, if the dependence within the cluster is strong, the rate of convergence is determined by the number of clusters: $q^{-1/2}$. If the dependence within clusters is weak —in a precise sense that we illustrate later in the examples—, the rate of convergence is $n^{-1/2}$. However, if the dependence is in-between weak and strong, the rate of convergence can be in-between or even slower than the rates mentioned above.

To analyze the convergence rate, we can compute the standard deviation of the sample mean. That is

$$\begin{aligned}\text{sd}(\bar{X}_n) &= \left(\text{Var}[\bar{X}_n]\right)^{1/2} \\ &= \frac{1}{n}\left(\sum_{j=1}^{q} \text{Var}[X_j' \mathbf{1}_j]\right)^{1/2} \ ,\end{aligned}$$

where the last equation uses that X_j are independent across clusters j.

Example 16.1 *Consider the case where $n_j = n^a$ and $q = n^{1-a}$ for some $a \in (0,1)$. Let $X_{i,j} \in \boldsymbol{R}$ be a random variable such that $Var[X_{i,j}] = 1$ and $Cov[X_{i,j}, X_{i',j}] = 0$. In this case it follows that*

$$Var[X_j' \boldsymbol{1}_j] = \sum_{i=1}^{n_j} Var[X_{i,j}] = n_j \ ,$$

and then

$$sd(\bar{X}_n) = \frac{1}{n}\left(q n_j\right)^{1/2} = n^{-1/2} \ ,$$

where the last equality follows from $q = n^{1-a}$ and $n_j = n^a$.

Example 16.2 *Consider the case where $n_j = n^a$ and $q = n^{1-a}$ for some $a \in (0,1)$. Let $X_{i,j} \in \boldsymbol{R}$ be a random variable such that $Var[X_{i,j}] = 1$ and $Cov[X_{i,j}, X_{i',j}] = 1$. In this case it follows that*

$$Var[X_j' \boldsymbol{1}_j] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} Cov[X_{i,j}, X_{i',j}] = n_j^2 = n^{2a} \ ,$$

and then

$$sd(\bar{X}_n) = \frac{1}{n}\left(q n^{2a}\right)^{1/2} = q^{-1/2} \ ,$$

where the last equality follows because $q = n^{1-a}$.

Example 16.3 *Consider the case where $n_j = n^a$ and $q = n^{1-a}$ for some $a \in (0,1)$. Let $X_{i,j} \in \boldsymbol{R}$ be a random variable such that $Var[X_{i,j}] = 1$ and $Cov[X_{i,j}, X_{i',j}] = 1/|i - i'|$ for $i \neq i'$. These conditions implies that*

$$Var[X_j' \mathbf{1}_j] = \sum_{i=1}^{n_j} \sum_{i'=1}^{n_j} Cov[X_{i,j}, X_{i',j}] = \sum_{i=1}^{n_j} \left(1 + \sum_{i' \neq i}^{n_j} \frac{1}{|i - i'|} \right) \ .$$

Now, let us rely on the following asymptotic proportional approximation

$$\sum_{i'=1}^{n_j} \frac{1}{|i - i'|} \propto \log(n_j) \propto \log(n) \ ,$$

where the last proportion approximation follows because $\log(n_j) = a \log(n)$. We can use this to conclude that

$$Var[X_j' \mathbf{1}_j] \propto \sum_{i=1}^{n_j} (1 + \log(n)) \propto n_j \log(n) \ .$$

This last expression can be use to approximate the standard deviation of the sample mean,

$$sd(\bar{X}_n) \propto \frac{1}{n} \left(\sum_{j=1}^{q} n_j \log(n) \right)^{1/2} = \frac{1}{n} \left(q n^a \log(n) \right)^{1/2} = \sqrt{\frac{\log(n)}{n}} \ ,$$

where the last equality uses $q = n^{1-a}$. It follows from here that the convergence rate is slower than $n^{-1/2}$ since

$$\sqrt{n}\, sd(\bar{X}_n) \propto \sqrt{\log(n)} \to \infty \quad as \quad n \to \infty \ .$$

At the same time, the convergence rate is faster than $q^{-1/2}$ since

$$\sqrt{q}\, sd(\bar{X}_n) \propto \sqrt{n^{1-a}} \sqrt{\frac{\log(n)}{n}} = \sqrt{\frac{\log(n)}{n^a}} \to 0 \quad as \quad n \to \infty \ .$$

Example 16.4 *Consider the case where there are two type of clusters. In the first group, the number of cluster is $q_1 = n/2$ and $n_j = 1$ for $j = 1, \ldots, q_1$. In the second group, the number of cluster is $q_2 = n^{1-a}/2$ and $n_j = n^a$ for $j = q_1 + 1, \ldots, q_1 + q_2$ and $a \in (0,1)$. The number of cluster is denoted by $q = q_1 + q_2$. Let $X_{i,j} \in \boldsymbol{R}$ be a random variable such that $Var[X_{i,j}] = 1$ and $Cov[X_{i,j}, X_{i',j}] = 1$. These conditions implies that*

$$Var[X_j' \mathbf{1}_j] = \sum_{i=1}^{n_j} \sum_{i'=1}^{n_j} Cov[X_{i,j}, X_{i',j}] = n_j^2 \ ,$$

which we can use to compute the standard deviation of the sample mean,

$$sd(\bar{X}_n) = \frac{1}{n}\left(\sum_{j=1}^{q} n_j^2\right)^{1/2} = \frac{1}{n}\left(q_1 + q_2 n^{2a}\right)^{1/2} ,$$

where the last equality follows because $Var[X_j' \mathbf{1}_j] = 1$ *if* $j = 1, \ldots, q_1$, *and* $Var[X_j' \mathbf{1}_j] = n^{2a}$ *if* $j = q_1 + 1, \ldots, q_1 + q_2$. *Now, we can use that* $q_1 = n/2$ *and* $q_2 = n^{1-a}/2$ *to conclude that*

$$sd(\bar{X}_n) = \frac{1}{n}\left(\frac{n}{2} + \frac{n^{1-a}}{2} n^{2a}\right)^{1/2} = \left(\frac{1 + n^a}{2n}\right)^{1/2} \propto n^{-(1-a)/2} .$$

It follows that the convergence rate is slower than $n^{-1/2}$ *since*

$$\sqrt{n}\, sd(\bar{X}_n) \propto n^{1/2} n^{-1/2} n^{a/2} \to \infty \quad as \quad n \to \infty .$$

In addition, the convergence rate is slower than $q^{-1/2}$ *since* $q = q_1 + q_2 \propto n$ *and*

$$\sqrt{q}\, sd(\bar{X}_n) \propto n^{1/2} n^{-1/2} n^{a/2} \to \infty \quad as \quad n \to \infty .$$

This means that $sd(\bar{X}_n)$ *goes to zero at a slower rate than* $n^{-1/2}$ *and* $q^{-1/2}$.

The examples above provide insight about the convergence rate $\text{sd}(\bar{X}_n)$. This rate can be equal to $\sqrt{n}$ (square root of the sample size). It also can be equal to $\sqrt{q}$ (the square root of the number of clusters). It can be between both of these rates and, under heterogeneity, can be even slower than both. When $\text{sd}(\bar{X}_n)$ is a vector, it is possible that each of its elements converge at different rates.

The last example illustrates the importance of considering heterogeneous cluster sizes. In this case, the reason why the convergence rate is slower than both $n^{-1/2}$ and $q^{-1/2}$ is because the number of clusters is determined by the large number of small clusters (q_1), but the convergence rate is determined by the (relatively) small number of large clusters ($q_2 n^{2a}$).

16.3 Central Limit Theorem

Under i.i.d. sampling the standard deviation of the sample mean, $\bar{X}_n$, is of order $O(n^{-1/2})$, so $\sqrt{n}$ to be the natural scaling to obtain the central limit theorem (CLT). However, clustering can alter the rate of convergence, as we saw in the examples above. Thus, it is essential to standardize the sample mean by the actual variance rather than an assumed rate. Now, define the variance-covariance matrix of $\sqrt{n}\bar{X}_n$ by

$$\begin{aligned}\Omega_n &= E\left[n\left(\bar{X}_n - E[\bar{X}_n]\right)\left(\bar{X}_n - E[\bar{X}_n]\right)'\right] \\ &= \frac{1}{n}\sum_{j=1}^{q} E\left[\left(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j]\right)\left(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j]\right)'\right] ,\end{aligned}$$

where $\mathbf{1}_j$ is an n_j-dimensional vector of ones. We also denote by $\lambda_n = \lambda_{\min}(\Omega_n)$ the minimum eigenvalue of Ω_n.

The next theorem presents a central limit theorem for the sample mean considering the correct scaling, $\Omega_n^{-1/2}\sqrt{n}$, so that

$$\Omega_n^{-1/2}\sqrt{n}(\bar{X}_n - E[\bar{X}_n])$$

is a random variable with mean zero and covariance matrix equal to the identity matrix $\mathbb{I}_{k+1}$ by construction.

Theorem 16.2 (Central Limit Theorem) *Suppose that for some* $2 \leq r < +\infty$,

$$\lim_{M\to\infty} \sup_{i,j} E\left[||X_{i,j}||^r I\{||X_{i,j}|| > M\}\right] = 0 \ , \tag{16.4}$$

and

$$\frac{\left(\sum_{j=1}^q n_j^r\right)^{2/r}}{n} \leq C < \infty \ , \tag{16.5}$$

for some positive $C > 0$. *Assume further that as* $n \to \infty$,

$$\max_{j\leq q} \frac{n_j^2}{n} \to 0 \tag{16.6}$$

and

$$\lambda_n \geq \lambda > 0 \ , \tag{16.7}$$

for some positive $\lambda > 0$. *Then, as* $n \to \infty$

$$\Omega_n^{-1/2}\sqrt{n}\left(\bar{X}_n - E[\bar{X}_n]\right) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ . \tag{16.8}$$

Let's discuss the conditions under which this theorem holds. Assumption (16.4) states that $||X_{i,j}||^r$ is uniformly integrable. When $r = 2$, this condition is similar to the Lindeberg condition for the CLT under independent heterogeneous sampling. Assumption (16.5) involves a trade-off between the cluster sizes and the number of moments r. It is least restrictive for large r, and more restrictive for small r. Note that as $r \to \infty$, we can conclude $\max_{j\leq q} n_j^2/n = O(1)$, which is implied by (16.6).

Assumption (16.6) allows for growing and heterogeneous cluster sizes. It allows clusters to grow uniformly at the rate $n_j = n^a$ for any $0 \leq a \leq (r-2)/(2r-2)$. Note that this requires the cluster sizes to be bounded if $r = 2$. It also allows for only a small number of clusters to grow. For example, $n_j = \bar{n}$ (bounded clusters) for $q - k$ clusters and $n_j = q^{a/2}$ for k clusters, with k fixed. In this case the assumption holds for any $a < 1$ and $r = 2$.

Finally, Assumption (16.7) specifies that $\text{Var}[\sqrt{n}c'\bar{X}_n]$ does not vanish for any vector $c \neq 0$, since the condition implies that the the minimum eigenvalue of the variance-covariance matrix is positive.

16.4 Cluster Covariance Estimation

Let us now consider the estimation of Ω_n. Suppose that $E[X_j'\mathbf{1}_j] = 0$ for all $j = 1, \ldots, q$. This implies that

$$\Omega_n = \frac{1}{n}\sum_{j=1}^{q} E\left[(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j])(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j])'\right]$$

is equal to

$$\frac{1}{n}\sum_{j=1}^{q} E[X_j'\mathbf{1}_j\mathbf{1}_j'X_j] \ .$$

The natural estimator of Ω_n would be

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{j=1}^{q} X_j'\mathbf{1}_j\mathbf{1}_j'X_j = \frac{1}{n}\sum_{j=1}^{q}\left(\sum_{i=1}^{n_j} X_{i,j}\right)\left(\sum_{i=1}^{n_j} X_{i,j}\right)' \ .$$

It is worth to mention that this estimator is robust to dependence within clusters. It allows for arbitrary within-cluster correlation patterns. Also, it allows for heterogeneity since $E[X_j'\mathbf{1}_j\mathbf{1}_j'X_j]$ can vary across j. The following theorem present the relevance of this estimator to obtain the asymptotic normality of the sample mean $\bar{X}_n$ after consider the right scaling.

Theorem 16.3 (Consistency of CCE) *Under the same assumptions of Theorem 16.2 and assuming that $E[X_j'\mathbf{1}_j] = 0$, we obtain as $n \to \infty$ that*

$$||\widehat{\Omega}_n - \Omega_n|| \xrightarrow{P} 0$$

and

$$\widehat{\Omega}_n^{-1/2}\sqrt{n}\bar{X}_n \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

This theorem shows that the cluster covariance estimator is consistent. Moreover, replacing the covariance matrix in the central limit theorem described in Theorem 16.2 with the estimated covariance matrix does not affect the asymptotic distribution. This implies that cluster-robust t-statistics are asymptotically standard normal. It is worth mentioning that we do not need to know the actual rate of convergence of $\bar{X}_n$ as the cluster covariance estimator capture this rate of convergence. For the proof of these results, see Hansen and Lee (2019).

16.4.1 Application to Linear Regression

Let us recall our initial setup. For each cluster j, let us denote by $X_j = (X_{1,j}, \ldots, X_{n_j,j})' \in \mathbf{R}^{n_j} \times \mathbf{R}^{k+1}$ the matrix of stacked observations. Define

$Y_j \in \mathbf{R}^{n_j} \times 1$ and $U_j \in \mathbf{R}^{n_j} \times 1$ in a similar way. Using this notation, we have

$$Y_j = X_j\beta + U_j, \quad j = 1, \ldots, q, \quad \text{where} \quad E[X_j'U] = 0 \ ,$$

and we assume that (Y_j, X_j) are independent across clusters but remain agnostic about the dependence within clusters.

The least square (LS) estimator of β is given by

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{j=1}^{q} X_j'X_j \right)^{-1} \frac{1}{n} \sum_{j=1}^{q} X_j'Y_j \ .$$

Using this expression and the model for Y_j, we can derive the following

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n} \sum_{j=1}^{q} X_j'X_j \right)^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{q} X_j'U_j \ .$$

Now, let us introduce notation before we discuss the consistency and the asymptotic normality properties of the LS estimator.

$$\Sigma_n = \frac{1}{n} \sum_{j=1}^{q} E[X_j'X_j] \quad \text{and} \quad \Omega_n = \frac{1}{n} \sum_{j=1}^{q} E[X_j'U_jU_j'X_j] \ .$$

Consistency of LS

If the condition (16.2) in Theorem 16.1 holds, Σ_n has full rank, $\lambda_{min}(\Sigma_n) \geq C > 0$, and the uniform integrability condition in (16.3) holds for $X_{i,j}X_{i,j}'$ and $X_{i,j}'U_{i,j}$, then

$$\hat{\beta}_n \to \beta \ .$$

Asymptotic Normality of LS

To properly normalize $\sqrt{n}(\hat{\beta}_n - \beta)$ we define

$$\mathbb{V}_n = \Sigma_n^{-1}\Omega_n\Sigma_n^{-1}$$

as the rate of convergence may not be $\sqrt{n}$. Using this notation, we assume that the conditions in Theorem 16.2 hold for some r, Σ_n has full rank, $\lambda_{\min}(\Sigma_n) \geq C > 0$, $\lambda_{\min}(\Omega_n) \geq C > 0$, and the uniform integrability condition in (16.4) holds for $X_{i,j}X_{i,j}'$ and $X_{i,j}'U_{i,j}$. It follows that as $n \to \infty$:

$$\mathbb{V}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

Note that to conduct inference, all we need is a consistent estimator $\widehat{\mathbb{V}}_n$ such that

$$\widehat{\mathbb{V}}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

This is what we develop next.

Cluster Covariance Estimator

Building on earlier results, we can immediately derive a consistent estimator of $\mathbb{V}_n$ as follows.

Definition 16.1 (CCE) *The CCE estimator of $\mathbb{V}_n$ is given by*

$$\widehat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{j=1}^{q} X_j'X_j\right)^{-1} \frac{1}{n}\sum_{j=1}^{q} X_j'\hat{U}_j\hat{U}_j'X_j \left(\frac{1}{n}\sum_{j=1}^{q} X_j'X_j\right)^{-1} ,$$

where $\hat{U}_j = Y_j - X_j\hat{\beta}_n$ are the LS residuals.

Under the same conditions listed for the asymptotic normality of the LS estimator, we can conclude

$$||\widehat{\mathbb{V}}_n - \mathbb{V}_n|| \xrightarrow{P} 0 \quad \text{and} \quad \widehat{\mathbb{V}}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) .$$

Note that in the special case with $n_j = 1$ for all $j = 1, \ldots, q$, this estimator becomes the HC estimator presented in Lecture 14. It is worth mentioning that `Stata` uses a multiplicative adjustment to reduce the bias,

$$\widehat{\mathbb{V}}_{\text{stata}} = \frac{n-1}{n-k-1}\frac{q}{q-1}\widehat{\mathbb{V}}_n .$$

This estimator allows for arbitrary within-cluster correlation patterns and heteroskedasticity across clusters. Unlike HAC estimators, it does not require the selection of a kernel or bandwidth parameter.

Inference

For $s \in \{0, 1, \ldots, k\}$, let β_s be the s-th element of β and let $\hat{V}_{n,s}$ be the $(s+1)$-th diagonal element of $\hat{V}_n$. Using this notation, consider testing

$$H_0 : \beta_s = c \quad \text{versus} \quad H_1 : \beta_s \neq c$$

at level α. Using the results we just derived, it follows that under the null hypothesis, the t-statistic is asymptotically standard normal,

$$t_{\text{stat}} = \frac{\sqrt{n}(\hat{\beta}_{n,s} - c)}{\sqrt{\hat{V}_{n,s}}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty .$$

This implies that the test that rejects H_0 when $|t_{\text{stat}}| > z_{1-\alpha/2}$ is consistent in levels, were $z_{1-\alpha/2}$ is a critical value defined by the $(1-\alpha/2)$-quantile of the standard normal distribution.

Remark 16.2 The previous results show that inference based on the t-statistic for the LS estimator of β and the cluster robust covariance estimator $\hat{\mathbb{V}}_n$ is valid as $q \to \infty$, regardless of whether $n \to \infty$ as well or not, and regardless of whether the dependence within each cluster is weak or strong. That is, even though the LS estimator may converge at different rates depending on the data structure, the studentization by the CCE captures such a rate of convergences and makes the t-statistic adaptive.

16.4.2 Small q ad-hoc adjustments

As we mentioned before, the cluster-robust inference asymptotics are based on many clusters, this means $q \to \infty$. However, often in empirical settings, there are few clusters (few regions, few schools, few states, etc). Following the ideas discussed with HC standard errors, Lecture 14, there are some finite-sample adjustments that people use in practice. For instance, Bell and McCaffrey (2002) proposes a bias-reduction modification analogous to that of HC2. That is

$$\hat{V}_{\text{bm}} = \left(\frac{1}{n}\sum_{j=1}^{q} X_j'X_j\right)^{-1} \frac{1}{n}\sum_{j=1}^{q} X_j'\tilde{U}_j\tilde{U}_j'X_j \left(\frac{1}{n}\sum_{j=1}^{q} X_j'X_j\right)^{-1} ,$$

where

$$\tilde{U}_j = (\mathbb{I}_{n_j} - \mathbb{P}_{jj})^{-1/2}\hat{U}_j ,$$

$\mathbb{I}_{n_j}$ is the $n_j \times n_j$ identity matrix, $\mathbb{P}_{jj}$ is the $n_j \times n_j$ matrix defined as

$$\mathbb{P}_{jj} = X_j(\mathbb{X}'\mathbb{X})^{-1}X_j' ,$$

and $\mathbb{X}$ is the $n \times (k+1)$ matrix constructed by stacking X_1 through X_q. Bell and McCaffrey (2002) also proposes a t critical value with degree of freedom adjustment, following the same intuition we discussed in the context of the Behrens-Fisher problem.

16.4.3 Simulations

Table 16.1 reports simulations results for the following five designs. The model in all designs is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i \\ X_i &= V_{C_i} + W_i \\ U_i &= \nu_{C_i} + \eta_i , \end{aligned}$$

where C_i denotes the cluster of i and all variables are $N(0,1)$. Also, in all designs $\beta_0 = \beta_1 = 0$. In the first design there are $q = 10$ clusters, with $n_j = 30$ units in each cluster. In the second design $q = 5$ with $n_j = 30$.

In the third design there there are $q = 10$ clusters, half with $n_j = 10$ and half with $n_j = 50$. The fourth design has heteroskedasticity, with $\eta_i|X_i \sim N(0, 0.9X_i^2)$, and the fifth design, the covariate is fixed within the clusters: $W_i = 0$ and $V_s \sim N(0, 2)$. The last two design have $q = 10$ clusters with $_j = 30$.

The following table reports the coverage probability of the Confidence Intervals

Table 16.1: Design in Imbens and Kolesar/CGM: $1 - \alpha = 95\%$

	dof	I	II	III	IV	V
$\hat{V}_n$	∞	84.7	73.9	79.6	85.7	81.7
	q-1	89.5	86.9	85.2	90.2	86.4
$\hat{V}_{\text{stata}}$	∞	86.7	78.8	81.9	87.6	83.6
	q-1	91.1	90.3	87.2	91.8	88.1
$\hat{V}_{\text{bm}}$	∞	89.2	84.7	87.2	89.1	87.7
	q-1	93.0	93.3	91.3	92.8	91.4
	k_{bm}	94.4	95.3	94.4	94.2	96.6

Final Comments

When q is small, $\hat{\mathbb{V}}_{bm}$ (more so $\hat{\mathbb{V}}_n$) typically leads to confidence sets that under-cover. Bell and McCaffrey (2002) consider more traction from the degree of freedom adjustment to the t-distribution that is used to compute the critical value of the test. This adjustment performs well sometimes, but is ad-hoc (no formal results).

The literature on inference with few clusters (i.e, q fixed) has made significant progress recently and the main alternatives to using CCE are:

- **The Wild Bootstrap:** See Cameron et al. (2008) and Canay et al. (2021).
- **Exact t-approach:** See Ibragimov and Müller (2010).
- **Approximate Randomization Tests:** See Canay et al. (2017).

The cluster consistent estimator discussed above is still a very good option when q is large, but remember that its performance does not improve if n_j gets large while q remains small.

Bibliography

Bell, R. M. and D. F. McCaffrey (2002): "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology*, 28, 169–182.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90, 414–427.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017): "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 85, 1013–1030.

Canay, I. A., A. Santos, and A. M. Shaikh (2021): "The wild bootstrap with a "small" number of "large" clusters," *Review of Economics and Statistics*, 103, 346–363.

Hansen, B. E. and S. Lee (2019): "Asymptotic theory for clustered samples," *Journal of econometrics*, 210, 268–290.

Hansen, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141, 597–620.

Ibragimov, R. and U. K. Müller (2010): "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28, 453–468.

Lecture 17

Bootstrap

17.1 Confidence Sets

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sample of observations with distribution $P \in \mathbf{P}$. The family $\mathbf{P}$ may be a parametric, nonparametric, or semiparametric family of distributions. We are interested in making inferences about some parameter $\theta(P) \in \Theta = \{\theta(P) : P \in \mathbf{P}\}$. Typical examples of $\theta(P)$ are the mean of P or median of P, but, more generally, it could be any function of P. Specifically, we are interested in constructing a confidence set for $\theta(P)$; that is, a random set, $C_n = C_n(X_1, \ldots, X_n)$ such that

$$P\{\theta(P) \in C_n\} \approx 1 - \alpha \ ,$$

at least for n sufficiently large.

The typical way of constructing such sets is based off of approximating the distribution of a *root*, $R_n = R_n(X_1, \ldots, X_n, \theta(P))$. A root is simply any real-valued function depending on both the data, $X_i, i = 1, \ldots, n$, and the parameter of interest, $\theta(P)$. The idea is that if the distribution of the root were known, then one could straightforwardly construct a confidence set for $\theta(P)$. To illustrate this idea, let $J_n(P)$ denote the sampling distribution of R_n and define the corresponding cumulative distribution function as,

$$J_n(x, P) = P\{R_n \leq x\} \ . \tag{17.1}$$

The notation is intended to emphasize the fact that the distribution of the root depends on both the sample size, n, and the distribution of the data, P. Using $J_n(x, P)$, we may choose a constant c such that

$$P\{R_n \leq c\} \approx 1 - \alpha \ .$$

Given such a c, the set

$$C_n = \{\theta \in \Theta : R_n(X_1, \ldots, X_n, \theta) \leq c\}$$

is a confidence set in the sense described above. We may also choose c_1 and c_2 so that

$$P\{c_1 \leq R_n \leq c_2\} \approx 1 - \alpha \ .$$

Given such c_1 and c_2, the set

$$C_n = \{\theta \in \Theta : c_1 \leq R_n(X_1, \ldots, X_n, \theta) \leq c_2\}$$

is a confidence set in the sense described above.

17.1.1 Pivots and Asymptotic Pivots

In some rare instances, $J_n(x, P)$ does not depend on P. In these instances, the root is said to be *pivotal* or a *pivot*. For example, if $\theta(P)$ is the mean of P and $\mathbf{P} = \{N(\theta, 1) : \theta \in \mathbf{R}\}$, then the root

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) \tag{17.2}$$

is a pivot because $R_n \sim N(0, 1)$. In this case, we may construct confidence sets C_n with finite-sample validity; that is,

$$P\{\theta(P) \in C_n\} = 1 - \alpha$$

for all n and $P \in \mathbf{P}$.

Sometimes, the root may not be pivotal in the sense described above, but it may be *asymptotically pivotal* or an *asymptotic pivot* in that $J_n(x, P)$ converges in distribution to a limit distribution $J(x, P)$ that does not depend on P. For example, if $\theta(P)$ is the mean of P and $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance, then

$$R_n = \frac{\sqrt{n}(\bar{X}_n - \theta(P))}{\hat{\sigma}_n} \tag{17.3}$$

is asymptotically pivotal because it converges in distribution to $J(x, P) = \Phi(x)$. In this case, we may construct confidence sets that are asymptotically valid in the sense that

$$\lim_{n \to \infty} P\{\theta(P) \in C_n\} = 1 - \alpha$$

for all $P \in \mathbf{P}$.

17.1.2 Asymptotic Approximations

Typically, the root will be neither a pivot nor an asymptotic pivot. The distribution of the root, $J_n(x, P)$, will typically depend on P, and, when it exists, the limit distribution of the root, $J(x, P)$, will, too. For example,

if $\theta(P)$ is the mean of P and $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance, then

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) \tag{17.4}$$

converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$. In this case, we can approximate this limit distribution with $\Phi(x/\hat{\sigma}_n)$, which will lead to confidence sets that are asymptotically valid in the sense described above.

Note that this third approach depends very heavily on the limit distribution $J(x, P)$ being both known and tractable. Even if it is known, the limit distribution may be difficult to work with (e.g., it could be the supremum of some complicated stochastic process with many nuisance parameters). Moreover, even if it is known and manageable, the method may be poor in finite-samples because it essentially relies on a double approximation: first, $J_n(x, P)$ is approximated by $J(x, P)$, then $J(x, P)$ is approximated in some way by estimating the unknown parameters of the limit distribution.

17.2 The Bootstrap

The bootstrap is a fourth, more general approach to approximating $J_n(x, P)$. The idea is very simple: replace the unknown P with an estimate $\hat{P}_n$. Given $\hat{P}_n$, it is possible to compute (either analytically or using simulation to any desired degree of accuracy) $J_n(x, \hat{P}_n)$. In the case of i.i.d. data, a typical choice is the empirical distribution (though if $P = P(\psi)$ for some finite-dimensional parameter ψ, then one may also use $\hat{P}_n = P(\hat{\psi}_n)$ for some estimate $\hat{\psi}_n$ of ψ). The hope is that whenever $\hat{P}_n$ is "close" to P (which may be ensured, for example, by the Glivenko-Cantelli Theorem), $J_n(x, \hat{P}_n)$ is "close" to $J_n(x, P)$. Essentially, this requires that $J_n(x, P)$, when viewed as a function of P, is continuous in an appropriate neighborhood of P. Often, this turns out to be true, but, unfortunately, it is not true in general.

17.2.1 The Nonparametric Mean

We will now consider the case where P is a distribution on $\mathbf{R}$ and $\theta(P)$ is the mean of P. We will consider first the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$. Let $\hat{P}_n$ denote the empirical distribution of the $X_i, i = 1, \ldots, n$. Under what conditions is $J_n(x, \hat{P}_n)$ "close" to $J_n(x, P)$?

The sequence of distributions $\hat{P}_n$ is a random sequence, so it is more convenient to answer the question first for a nonrandom sequence P_n. The following theorem does exactly that.

Theorem 17.1 Let $\theta(P)$ be the mean of P and let $\mathbf{P}$ denote the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance. Consider the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$. Let $P_n, n \geq 1$ be a nonrandom sequence of distributions such

that P_n converges in distribution to P, $\theta(P_n) \to \theta(P)$ and $\sigma^2(P_n) \to \sigma^2(P)$. Then,

(i) $J_n(x, P_n)$ converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$.

(ii) $J_n^{-1}(1-\alpha, P_n) = \inf\{x \in \mathbf{R} : J_n(x, P_n) \geq 1-\alpha\}$ converges to

$$J^{-1}(1-\alpha, P) = z_{1-\alpha}\sigma(P) \ .$$

PROOF. (i) For each n, let $X_{i,n}, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P_n. We must show that

$$\sqrt{n}(\bar{X}_{n,n} - \theta(P_n))$$

converges in distribution to $N(0, \sigma^2(P))$. To this end, let

$$Z_{n,i} = \frac{X_{n,i} - \theta(P_n)}{\sigma(P_n)}$$

and apply the Lindeberg-Feller central limit theorem. We must show that

$$\lim_{n\to\infty} E[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\}] = 0 \ .$$

Let $\epsilon > 0$ be given. By the assumption that P_n converges in distribution to P and Slutsky's Theorem,

$$Z_{n,i} \xrightarrow{d} Z = \frac{X - \theta(P)}{\sigma(P)} \ ,$$

where X has distribution P. It follows that for any $\lambda > 0$ for which the distribution of $|Z|$ is continuous at λ, we have that

$$E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}] \to E[Z^2 I\{|Z| > \lambda\}] \ .$$

To prove this last claim, we need a couple of results:

1. Lehmann and Romano (2005, Example 11.2.14): Suppose that Y_n and Y are real valued random variables and that $Y_n \xrightarrow{d} Y$. If the Y_n are uniformly bounded, then $E[Y_n] \to E[Y]$. (In general, convergence in distribution *does not* imply convergence of moments!)

2. Continuous Mapping Theorem (Lehmann and Romano, 2005, Theorem 11.2.13): Suppose that $Y_n \xrightarrow{d} Y$. Let g be a measurable map from $\mathbf{R}$ to $\mathbf{R}$. Let C be the set of point in $\mathbf{R}$ for which g is continuous. If $P\{Y \in C\} = 1$, then $g(Y_n) \xrightarrow{d} g(Y)$.

We now use these two results. First, note that for any $\lambda > 0$ for which the distribution of $|Z|$ is continuous at λ, the continuous mapping theorem above implies that

$$g(|Z_{n,i}|) = Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\} \xrightarrow{d} Z^2 I\{|Z| \leq \lambda\} = g(|Z|) \ . \tag{17.5}$$

Note that g is discontinuous at λ but that $P\{|Z| = \lambda\} = 0$, and so the result follows. Second, note that

$$E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}] = E[Z_{n,i}^2] - E[Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\}] \ .$$

The first term on the right-hand side is always equal to one and also equal to $E[Z^2] = 1$. The second term is the expectation of

$$Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\} \in [0, \lambda^2] \ ,$$

which is uniformly bounded. By (17.5) and the first result above,

$$E[Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\}] \to E[Z^2 I\{|Z| \leq \lambda\}] \ .$$

We conclude that

$$\begin{aligned} E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}] &= E[Z^2] - E[Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\}] \\ &\to E[Z^2] - E[Z^2 I\{|Z| \leq \lambda\}] \\ &= E[Z^2 I\{|Z| > \lambda\}] \ . \end{aligned}$$

As $\lambda \to \infty$, $E[Z^2 I\{|Z| > \lambda\}] \to 0$. To complete the proof, note that for any fixed $\lambda > 0$

$$E[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\}] \leq E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}]$$

for n sufficiently large. Thus,

$$\sqrt{n}\bar{Z}_{n,n} \xrightarrow{d} N(0,1)$$

under P_n. The desired result now follows from Slutsky's Theorem and the fact that $\sigma(P_n) \to \sigma(P)$.

(ii) This follows from part (i) and Lemma 17.1 below applied to $F_n(x) = J_n(x, P)$ and $F(x) = J(x, P)$. ■

Lemma 17.1 Let $F_n, n \geq 1$ and F be nonrandom distribution functions on $\mathbf{R}$ such that F_n converges in distribution to F. Suppose F is continuous and strictly increasing at $F^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : F(x) \geq 1-\alpha\}$. Then, $F_n^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : F_n(x) \geq 1-\alpha\} \to F^{-1}(1-\alpha)$.

PROOF: Let $q = F^{-1}(1-\alpha)$. Fix $\delta > 0$ and choose ϵ so that $0 < \epsilon < \delta$ and F is continuous at $q-\epsilon$ and $q+\epsilon$. This is possible because F is continuous at q and therefore continuous in a neighborhood of q. Hence, $F_n(q-\epsilon) \to F(q-\epsilon) < 1-\alpha$ and $F_n(q+\epsilon) \to F(q+\epsilon) > 1-\alpha$, where the inequalities follow from the assumption that F is strictly increasing at q. For n sufficiently large, we thus have that $F_n(q-\epsilon) < 1-\alpha$ and $F_n(q+\epsilon) > 1-\alpha$. It follows that $q-\epsilon \leq F^{-1}(1-\alpha) \leq q+\epsilon$ for such n. ■

We are now ready to pass from the nonrandom sequence P_n to the random sequence $\hat{P}_n$.

Theorem 17.2 Let $\theta(P)$ be the mean of P and let $\mathbf{P}$ denote the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance. Consider the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$. Then,

(i) $J_n(x, \hat{P}_n)$ converges in distribution to $J(x,P) = \Phi(x/\sigma(P))$ a.s.

(ii) $J_n^{-1}(1-\alpha, \hat{P}_n)$ converges to $J^{-1}(1-\alpha, P) = z_{1-\alpha}\sigma(P)$ a.s.

PROOF: By the Glivenko-Cantelli Theorem,

$$\sup_{x \in \mathbf{R}} |\hat{P}_n((-\infty, x]) - P((-\infty, x])| \to 0 \quad a.s.$$

This implies that $\hat{P}_n$ converges in distribution to P a.s. Since $|x| \leq 1 + x^2$ and that $\sigma^2(P) < \infty$, we have that $E[|X|] \leq 1 + E[X^2] < \infty$. Thus, we may apply the Strong Law of Large Numbers to conclude that $\theta(\hat{P}_n) = \bar{X}_n$ converges to $\theta(P)$ a.s. and $\sigma(\hat{P}_n)$ converges to $\sigma(P)$ a.s. Thus, w.p.1, $\hat{P}_n$ satisfies the assumptions of Theorem 17.1. The conclusions of the theorem now follow. ■

Remark 17.1 Similar results hold for the studentized root in (17.1) where $\hat{\sigma}_n$ is a consistent estimator of $\sigma(P)$. Using this root leads to the so-called Bootstrap-t, as the root is just the t-statistic. A key step in the proof of this result is to show that $\hat{\sigma}_n$ converges in probability to $\sigma(P)$ under an appropriate sequence of distributions. We skip this in this class. However, the advantage of working with a studentized root like the one in (17.3) is that the limit distribution of R_n is pivotal, which affects the properties of the bootstrap approximation as discussed in the next section.

It now follows from Slutsky's Theorem that confidence sets of the form

$$C_n = \left\{\theta \in \mathbf{R} : R_n(X_1, \ldots, X_n, \theta) \leq J_n^{-1}(1-\alpha, \hat{P}_n)\right\},$$

which are known as *symmetric* confidence sets, or

$$C_n = \left\{\theta \in \mathbf{R} : J_n^{-1}\left(\frac{\alpha}{2}, \hat{P}_n\right) \leq R_n(X_1, \ldots, X_n, \theta) \leq J_n^{-1}\left(1-\frac{\alpha}{2}, \hat{P}_n\right)\right\},$$

which are known as *equi-tailed* confidence sets, satisfy

$$P\{\theta(P) \in C_n\} \to 1 - \alpha \tag{17.6}$$

for all $P \in \mathbf{P}$.

In general, the consistency of the bootstrap is proved in the following two steps:

1. For some choice of metric (or pseudo-metric) d on the space of probability measures, it must be known that $d(P_n, P) \to 0$ implies that $J_n(P_n)$ converges weakly to $J(P)$. That is, the convergence of $J_n(P)$ to $J(P)$ must hold in a suitably locally uniform in P manner. After all, we are replacing P by $\hat{P}_n$ so $J_n(P)$ must be smooth in P. Note that in Theorem 17.1, the "metric" d that we used involved weak convergence together with convergence of first and second moments, see Remark 15.4.1 in Lehmann and Romano (2005) for details. However, other problems may require a different metric.

2. The estimator $\hat{P}_n$ must then be known to satisfy $d(\hat{P}_n, P) \to 0$ almost surely or in probability under P. This is what we proved in the proof of Theorem 17.2.

17.2.2 Asymptotic Refinements

Note that even a confidence set C_n based off of the asymptotic normality of either root would satisfy (17.6). It can be shown under certain conditions (that ensure the existence of so-called Edgeworth expansions of $J_n(x, P)$) that one-sided confidence sets C_n based off of such an asymptotic approximation satisfy

$$P\{\theta(P) \in C_n\} - (1 - \alpha) = O(n^{-1/2}) \ . \tag{17.7}$$

One-sided confidence sets based off of the bootstrap and the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$ also satisfy (17.7), though there is some evidence to suggest that it does a bit better in the size of $O(n^{-1/2})$ term. On the other hand, one-sided confidence sets based off the bootstrap-t, i.e., using the root

$$R_n = \frac{\sqrt{n}(\bar{X}_n - \theta(P))}{\hat{\sigma}_n}$$

as in Remark 17.1, satisfy

$$P\{\theta(P) \in C_n\} - (1 - \alpha) = O(n^{-1}) \ . \tag{17.8}$$

Thus, the one-sided coverage error of the bootstrap-t interval is $O(n^{-1})$ and is of smaller order than that provided by the normal approximation or the bootstrap based on a nonstudentized root. One-sided confidence sets that

satisfy only (17.7) are said to be first-order accurate, where as one-sided confidence sets that satisfy (17.8) are said to be second-order accurate. See Section 15.5 of Lehmann and Romano (2005) for further details.

A heuristic reason why the bootstrap based on the root (17.3) outperforms the bootstrap based on the root (17.4) is as follows. In the case of (17.4), the bootstrap is estimating a distribution that has mean 0 and unknown variance $\sigma^2(P)$. The main contribution to the estimation error is the implicit estimation of $\sigma^2(P)$ by $\sigma^2(\hat{P}_n)$. On the other hand, the root (17.3) has a distribution that is nearly independent of P since it is an asymptotic pivot.

The bootstrap may also provide a refinement in two-sided tests. For example, symmetric intervals based on the absolute value of the root in (17.3) are $O(n^{-2})$, versus the asymptotic approximation that is of order $O(n^{-1})$. Note that, by construction, such intervals are symmetric about $\hat{\theta}_n$.

17.2.3 Implementation of the Bootstrap

Outside certain exceptional cases, the bootstrap approximation $J_n(x, \hat{P}_n)$ cannot be calculated exactly, i.e., it is often not available in closed form. However, we can approximate this distribution to an arbitrary degree of accuracy by taking samples from $\hat{P}_n$, computing the root for each of these samples, and then using the empirical distribution of these roots as an approximation to $J_n(x, \hat{P}_n)$. The usual algorithm used to implement the bootstrap involves the following steps.

Step 1. Conditional on the data $(X_1, \ldots, X_n)$, draw B samples of size n from $\hat{P}_n$. Denote the jth sample by

$$(X^*_{1,j}, \ldots, X^*_{n,j})$$

for $j = 1, \ldots, B$. When $\hat{P}_n$ is the empirical distribution, this amounts to resampling the original observations in $(X_1, \ldots, X_n)$ with replacement.

Step 2. For each bootstrap sample j, compute the root, i.e.,

$$R^*_{j,n} = R_n(X^*_{1,j}, \ldots, X^*_{n,j}, \hat{\theta}_n) \ .$$

Note that $\theta(\hat{P}_n) = \hat{\theta}_n$, so in the bootstrap distribution the parameter $\theta(P)$ becomes $\hat{\theta}_n$.

Step 3. Compute the empirical cdf of $(R^*_{1,n}, \ldots, R^*_{B,n})$ as

$$L_n(x) = \frac{1}{B} \sum_{j=1}^{B} I\{R^*_{j,n} \leq x\} \ . \tag{17.9}$$

Step 4. Compute the desired function of $L_n(x)$, for example, a quantile,

$$L_n^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : L_n(x) \geq 1-\alpha\} \ ,$$

for a given significance level α.

Remark 17.2 Sampling from $\hat{P}_n$ in Step 1 is easy even when $\hat{P}_n$ is the empirical distribution. In such case $\hat{P}_n$ is a discrete probability distribution that puts probability mass $\frac{1}{n}$ at each sample point $(X_1, \ldots, X_n)$, so sampling from $\hat{P}_n$ is equivalent to drawing observations (with probability $\frac{1}{n}$) from the observed data *with* replacement. In consequence, a bootstrap sample will likely have some ties and multiple values, which is generally not a problem. In parametric problems one would simply get a new sample of size n from $\hat{P}_n = P(\hat{\psi}_n)$.

Because B can be taken to be large (assuming enough computing power), the resulting approximation $L_n(x)$ can be made arbitrarily close to $J_n(x, \hat{P}_n)$. It then follows that the properties of tests and confidence sets based on $J_n^{-1}(1-\alpha, \hat{P}_n)$ and $L_n^{-1}(1-\alpha)$ are the same. In practice, values of B in the order of $1,000$ are frequently enough for the approximation to work well.

Bibliography

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Lehmann, E. and J. P. Romano (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer, New York.

Lecture 18

Subsampling & Randomization Tests

18.1 Subsampling

Suppose $X_i, i = 1, \ldots, n$ is an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$. Let $\theta(P)$ be some real-valued parameter of interest, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be some estimate of $\theta(P)$. Consider the root

$$R_n = \sqrt{n}(\hat{\theta}_n - \theta(P)) \ ,$$

where root stands for a functional depending on both, the data and $\theta(P)$. Let $J_n(P)$ denote the sampling distribution of Rz_n and define the corresponding cumulative distribution function as,

$$J_n(x, P) = P\{R_n \leq x\} \ . \tag{18.1}$$

We wish to estimate $J_n(x, P)$ so we can make inferences about $\theta(P)$. For example, we would like to estimate quantiles of $J_n(x, P)$, so we can construct confidence sets for $\theta(P)$. Unfortunately, we do not know P, and, as a result, we do not know $J_n(x, P)$.

The bootstrap solved this problem simply by replacing the unknown P with an estimate $\hat{P}_n$. In the case of i.i.d. data, a typical choice of $\hat{P}_n$ is the empirical distribution of the $X_i, i = 1, \ldots, n$. For this approach to work, we essentially required that $J_n(x, P)$ when viewed as a function of P was continuous in a certain neighborhood of P. An alternative to the bootstrap known as subsampling, originally due to Politis and Romano (1994), does not impose this requirement but rather the following much weaker condition.

Assumption 18.1 *There exists a limiting law $J(P)$ such that $J_n(P)$ converges weakly to $J(P)$ as $n \to \infty$.*

In order to motivate the idea behind subsampling, consider the following thought experiment. Suppose for the time being that $\theta(P)$ is known. Suppose that, instead of n i.i.d. observations from P, we had a very, very large number of i.i.d. observations from P. For concreteness, suppose $X_i, i = 1, \ldots, m$ is an i.i.d. sequence of random variables with distribution P with $m = nk$ for some very big k. We could then estimate $J_n(x, P)$ by looking at the empirical distribution of

$$\sqrt{n}(\hat{\theta}_n(X_{n(j-1)+1}, \ldots, X_{nj}) - \theta(P)),\ j = 1, \ldots, k\ .$$

This is an i.i.d. sequence of random variables with distribution $J_n(x, P)$. Therefore, by the Glivenko-Cantelli theorem, we know that this empirical distribution is a good estimate of $J_n(x, P)$, at least for large k. In fact, with a simple trick, we could show that it is even possible to improve upon this estimate by using all possible sets of data of size n from the m observations, not just those that are disjoint; that is, estimate $J_n(x, P)$ with the empirical distribution of the

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta(P)),\ j = 1, \ldots, \binom{m}{n}\ .$$

where $\hat{\theta}_{n,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size n from the original m observations.

In practice $m = n$, so, even if we knew $\theta(P)$, this idea won't work. The key idea behind subsampling is the following simple observation: replace n with some smaller number b that is much smaller than n. We would then expect

$$\sqrt{b}(\hat{\theta}_{b,j} - \theta(P)),\ j = 1, \ldots, \binom{n}{b}\ ,$$

where $\hat{\theta}_{b,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size b from the original n observations, to be a good estimate of $J_b(x, P)$, at least if $\binom{n}{b}$ is large. Of course, we are interested in $J_n(x, P)$, not $J_b(x, P)$. We therefore need some way to force $J_n(x, P)$ and $J_b(x, P)$ to be close to one another. To ensure this, it suffices to assume that $J_n(x, P) \to J(x, P)$. Therefore, $J_b(x, P)$ and $J_n(x, P)$ are both close to $J(x, P)$, and thus close to one another as well, at least for large b and n. In order to ensure that both b and $\binom{n}{b}$ are large, at least asymptotically, it suffices to assume that $b \to \infty$, but $b/n \to 0$.

This procedure is still not feasible because in practice we typically do not know $\theta(P)$. But we can replace $\theta(P)$ with $\hat{\theta}_n$. This would cause no problems if

$$\sqrt{b}(\hat{\theta}_n - \theta(P)) = \frac{\sqrt{b}}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta(P))$$

is small, which follows from $b/n \to 0$ in this case. The next theorem formalizes the above discussion.

Theorem 18.1 *Assume Assumption 18.1. Also, let $J_n(P)$ denote the sampling distribution of $\tau_n(\hat{\theta}_n - \theta(P))$ for some normalizing sequence $\tau_n \to \infty$, $N_n = \binom{n}{b}$, and assume that $\tau_b/\tau_n \to 0$, $b \to \infty$, and $b/n \to 0$ as $n \to \infty$.*

i) If x is a continuity point of $J(\cdot, P)$, then $L_{n,b}(x) \to J(x, P)$ in probability, where

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{j=1}^{N_n} I\{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_n) \leq x\} \ . \tag{18.2}$$

ii) Let

$$c_{n,b}(1-\alpha) = \inf\{x : L_{n,b}(x) \geq 1-\alpha\} \ ,$$
$$c(1-\alpha, P) = \inf\{x : J(x, P) \geq 1-\alpha\} \ .$$

If $J(\cdot, P)$ is continuous at $c(1-\alpha, P)$, then

$$P\{\tau_n(\hat{\theta}_n - \theta(P)) \leq c_{n,b}(1-\alpha)\} \to 1-\alpha \text{ as } n \to \infty \ . \tag{18.3}$$

In practice, N_n is too large to actually compute $L_n(x)$, so what one would do is randomly sample B of the N_n possible data sets of size b and just use B in place of N_n when computing $L_n(x)$. Provided $B = B_n \to \infty$, all the conclusions of the theorem remain valid. This approximation step is similar in spirit to approximating the bootstrap distribution $J_n(x, \hat{P}_n)$ using simulations from $\hat{P}_n$. In fact, except for the first step, implementing the bootstrap and subsampling requires the same algorithm. The change in the first step is as follows:

Step 1: Non-parametric Bootstrap. Conditional on the data $(X_1, \ldots, X_n)$, draw B samples of size n from the original observations *with replacement*.

Step 1: Subsampling. Conditional on the data $(X_1, \ldots, X_n)$, draw B samples of size b from the original observations *without replacement*.

Essentially, all we required was that $J_n(x, P)$ converged in distribution to a limit distribution $J(x, P)$, whereas for the bootstrap we required this and additionally that $J_n(x, P)$ was continuous in a certain sense. Showing continuity of $J_n(x, P)$ was very problem specific. There are examples where $J_n(x, P) \to J(x, P)$, but this continuity fails (e.g., the extreme order statistic). Subsampling would have no problems handling the extreme order statistic.

Typically, when both the bootstrap and subsampling are valid, the bootstrap works better in the sense of higher-order asymptotics (see the lecture notes on the bootstrap), but subsampling is more generally valid.

There is a variant of the bootstrap known as the m-out-of-n bootstrap. Instead of using $J_n(x, \hat{P}_n)$ to approximate $J_n(x, P)$, one uses $J_m(x, \hat{P}_n)$ where m is much smaller than n. If one assumes that $m^2/n \to 0$, then all the conclusions of the theorem remain valid with $J_m(x, \hat{P}_n)$ in place of $L_n(x)$. This follows because if $m^2/n \to 0$, then (i) $m/n \to 0$ and (ii) with probability tending to 1, the approximation to $J_m(x, \hat{P}_n)$ is the same as the approximation to $L_n(x)$ because the probability of drawing all distinct observations tends to 1. To see this, note that this probability is simply equal to

$$\frac{n(n-1)(n-2)\cdots(n-b+1)}{n^b} = \prod_{1 \le i \le b-1} (1 - \frac{i}{n}) \ .$$

Since $1 - \frac{i}{n} \ge 1 - \frac{b}{n}$, we have that

$$\prod_{1 \le i \le b-1} (1 - \frac{i}{n}) \ge (1 - \frac{b}{n})^b = (1 - \frac{\frac{b^2}{n}}{b})^b \ .$$

If $b^2/n \to 0$, then for every $\epsilon > 0$ we have that $b^2/n < \epsilon$ for all n sufficiently large. Therefore,

$$(1 - \frac{\frac{b^2}{n}}{b})^b > (1 - \frac{\epsilon}{b})^b \to \exp(-\epsilon) \ .$$

By choosing $\epsilon > 0$ sufficiently small, we see that the desired probability converges to 1.

18.2 Randomization Tests

Before we describe the general construction of randomization tests, we start our discussion in the context of a simple example.

18.2.1 Motivating example: sign changes

Let $X = (X_1, \ldots, X_{10}) \sim P$ be an i.i.d. sample of size 10 where each X_i takes values in $\mathbf{R}$, has a finite mean $\theta \in \mathbf{R}$, and has a distribution that is symmetric about θ. Let $\mathbf{P}$ be the collection of all distributions P satisfying these conditions. Consider testing

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \neq 0 \ .$$

We only have 10 observations, so using an asymptotic approximation does not seem fruitful. At the same time, this is more general than the normal location model where each X_i has distribution $N(\theta, \sigma^2)$, so exploiting normality is not possible.

Suppose we decided to use the absolute value of $\bar{X}_{10}$ to test the above hypothesis. Denote this test statistic by $T(X)$. The question is: how do we

compute a critical value that delivers a valid test? It turns out we can do this by exploiting symmetry.

To do this, let ϵ_i take on either the value 1 or -1 for $i = 1, \ldots, 10$. Note that the distribution of $X = (X_1, \ldots, X_{10})$ is symmetric about 0 under the null hypothesis. Now consider a transformation $g = (\epsilon_1, \ldots, \epsilon_{10})$ of $\mathbf{R}^{10}$ that defines the following mapping

$$(X_1, \ldots, X_{10}) \mapsto gX = (\epsilon_1 X_1, \ldots, \epsilon_{10} X_{10}) \ . \tag{18.4}$$

Finally, let $\mathbf{G}$ be the $M = 2^{10}$ collection of such transformations. It follows that the random variable X and gX have the *same distribution* under the null hypothesis. What this means is that we can get "new samples" from P by simply applying g to X. We can get a total of $M = 1,024$ samples and use these samples to simulate the distribution of $T(X)$. This approach leads to a test that is valid in finite samples as the next section shows.

18.2.2 The main result

In this section X denotes the observed sample and P denotes the distribution of the entire sample X (as in the motivating example). Since all results are finite sample in nature, we do not use an index n to denote the sample size and do not index objects by n.

Based on data X taking values in a sample space $\mathcal{X}$, it is desired to test the null hypothesis $H_0 : P \in \mathbf{P}_0$, where P is the true distribution of X and $\mathbf{P}_0$ is a subset of distributions in the space $\mathbf{P}$. Let $\mathbf{G}$ be a finite group of transformations $g : \mathcal{X} \mapsto \mathcal{X}$. The following assumption allows for a general test construction.

Definition 18.1 (Randomization Hypothesis) *Under the null hypothesis, the distribution of X is invariant under the transformations in $\mathbf{G}$; that is, for every $g \in \mathbf{G}$, gX and X have the same distribution whenever $X \sim P \in \mathbf{P}_0$.*

Note that we do not require the alternative hypothesis parameter space to remain invariant under g in $\mathbf{G}$. Only the space $\mathbf{P}_0$ is assumed invariant.

Let $T(X)$ be any real-valued test statistic for testing H_0. Suppose that the group $\mathbf{G}$ has M elements. Given $X = x$, let

$$T^{(1)}(x) \leq T^{(2)}(x) \leq \cdots \leq T^{(M)}(x) \tag{18.5}$$

be ordered values of $T(gX)$ as g varies in $\mathbf{G}$. Fix a nominal level α, $0 < \alpha < 1$, and let k be defined as

$$k = \lceil (1 - \alpha)M \rceil \tag{18.6}$$

where $\lceil C \rceil$ denotes the smallest integer greater than or equal to C. Let

$$M^+(x) = \sum_{j=1}^{M} I\{T^{(j)}(x) > T^{(k)}(x)\}$$
$$M^0(x) = \sum_{j=1}^{M} I\{T^{(j)}(x) = T^{(k)}(x)\} \ .$$

Now set

$$a(x) = \frac{M\alpha - M^+(x)}{M^0(x)} \ , \tag{18.7}$$

and define the randomization test as

$$\phi(x) = \begin{cases} 1 & T(x) > T^{(k)}(x) \\ a(x) & T(x) = T^{(k)}(x) \\ 0 & T(x) < T^{(k)}(x) \end{cases} . \tag{18.8}$$

This test is randomized, and since $M^+(x) \leq M - k \leq M\alpha$ and $M^+(x) + M^0(x) \geq M - k + 1 > M\alpha$, we have $a(x) \in [0, 1)$.

Under the randomization hypothesis, Hoeffding (1952) shows that this construction results in a test of exact level α, and this is true for *any* choice of test statistic $T(X)$. Note that this is possibly a randomized test if $(1-\alpha)M$ is not an integer and there are ties in the ordered values. Alternatively, if one prefers not to randomize, the slightly conservative but non-randomized test that rejects when $T(X) > T^{(k)}$, i.e.,

$$\phi^{\rm nr}(X) = I\{T(X) > T^{(k)}\} \ , \tag{18.9}$$

is level α.

Theorem 18.2 *Suppose that X has distribution P ion $\mathcal{X}$ and the problem is to test the null hypothesis $P \in \mathbf{P}_0$. Let $\mathbf{G}$ be a finite group of transformations of $\mathcal{X}$ onto itself. Suppose the randomization hypothesis (18.1) holds. Given a test statistic $T(X)$, let ϕ be the randomization test as described above. Then, $\phi(X)$ is a similar α level test, i.e.,*

$$E_P[\phi(X)] = \alpha, \ \textit{for all } P \in \mathbf{P}_0 \ .$$

PROOF. By construction, for every $x \in \mathcal{X}$,

$$\sum_{g \in \mathbf{G}} \phi(gx) = M^+(x) + a(x)M^0(x) = M\alpha \ ,$$

and so

$$M\alpha = E_P \left[\sum_{g \in \mathbf{G}} \phi(gX) \right] = \sum_{g \in \mathbf{G}} E_P[\phi(gX)] \ .$$

By the randomization hypothesis $E_P[\phi(gX)] = E_P[\phi(X)]$, so that

$$M\alpha = \sum_{g\in\mathbf{G}} E_P[\phi(gX)] = \sum_{g\in\mathbf{G}} E_P[\phi(X)] = ME_P[\phi(X)] \ ,$$

and the result follows. ■

Remark 18.1 Note that by construction the randomization test not only is of level α for all n, but also "similar", meaning that $E_P[\phi(X)]$ is never below α for any $P \in \mathbf{P}_0$.

In general, one can define a p-value $\hat{p}$ of a randomization test by

$$\hat{p} = \frac{1}{M} \sum_{g\in\mathbf{G}} I\{T(gX) \geq T(X)\} \ . \tag{18.10}$$

It can be shown that $\hat{p}$ satisfies, under the null hypothesis,

$$P\{\hat{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 \ . \tag{18.11}$$

Therefore, the non-randomized test that rejects when $\hat{p} \leq \alpha$ is level α.

Because $\mathbf{G}$ may be large, one may resort to an approximation to construct the randomization test, for example, by randomly sampling transformations g from $\mathbf{G}$ with or without replacement. In the former case, for example, suppose $g_1, \ldots, g_{B-1}$ are i.i.d. and uniformly distributed on $\mathbf{G}$. Let

$$\tilde{p} = B^{-1} \left[1 + \sum_{i=1}^{B-1} I\{T(g_i X) \geq T(X)\} \right] \ . \tag{18.12}$$

Then, it can be shown that, under the null hypothesis,

$$P\{\tilde{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 \ , \tag{18.13}$$

where this probability reflects variation in both X and the sampling of the g_i.

18.2.3 Special case: Permutation tests

Probably the most popular application of randomization tests in economic applications are the so-called permutation tests, which are just a special case of the general construction we just described.

Two sample problem

Suppose that $Y_1, \ldots, Y_m$ are i.i.d. observations from a distribution P_Y and, independently, $Z_1, \ldots, Z_n$ are i.i.d. observations from a distribution P_Z. In other words, we have two samples that are not paired, i.e., Z_1 and Y_1 do not correspond to the same "unit". Here X is given by

$$X = (Y_1, \ldots, Y_m, Z_1, \ldots, Z_n) \ .$$

Consider testing

$$H_0 : P_Y = P_Z \text{ vs } H_1 : P_Y \neq P_Z \ .$$

To describe an appropriate group of transformations $\mathbf{G}$, let $N = m + n$. For $x = (x_1, \ldots, x_N) \in \mathbf{R}^N$, let $gx \in \mathbf{R}^N$ be defined by

$$(x_1, \ldots, x_N) \mapsto gx = (x_{\pi(1)}, \ldots, x_{\pi(N)}) \ , \tag{18.14}$$

where $(\pi(1), \ldots, \pi(N))$ is a permutation of $\{1, \ldots, N\}$. Let $\mathbf{G}$ be the collection of all such g, so that $M = N!$. It follows that whenever $P_Y = P_Z$, X and gX have the same distribution.

In essence, each transformation g produces a new data set gx, of which the first m elements are used as the Y sample and the remaining n as the Z sample to recompute the test statistic. Note that, if a test statistic is chosen that is invariant under permutations within each of the Y and Z samples, like $\bar{Y}_m - \bar{Z}_n$, it is enough to consider the $\binom{N}{m}$ transformed data sets obtained by taking m observations from all N as the Y observations and the remaining n as the Z observations (which, of course, is equivalent to using a subgroup $\mathbf{G}'$ of $\mathbf{G}$).

Treatment effects

Suppose that we observe a random sample $\{(Y_1, D_1), \ldots, (Y_n, D_n)\}$ from a randomized controlled trial where

$$Y = Y(1)D + (1 - D)Y(0)$$

is the observed outcome and $D \in \{0, 1\}$ is the exogenous treatment assignment. Here, $(Y(0), Y(1))$ are the usual potential outcomes. Suppose that we are interested in testing the hypothesis that the distribution Q_0 of $Y(0)$ is the same as the distribution Q_1 of $Y(1)$. That is,

$$H_0 : Q_0 = Q_1 \text{ vs. } H_1 : Q_0 \neq Q_1 \ . \tag{18.15}$$

Under the null hypothesis in (18.15), it follows that the distribution of $\{(Y_1, D_1), \ldots, (Y_n, D_n)\}$ and $\{(Y_1, D_{\pi(1)}), \ldots, (Y_n, D_{\pi(n)})\}$ are the same for any permutation $(\pi(1), \ldots, \pi(n))$ of $\{1, \ldots, n\}$, and so a permutation test that permutes individual from "treatment" to "control" (or from "control" to "treatment") delivers a test that is valid in finite samples.

However, researchers are often interested in hypotheses about the average treatment effect (ATE) as opposed to those in (18.15). For example, consider

$$H_0 : E[Y(1)] = E[Y(0)] \text{ v.s. } H_1 : E[Y(1)] \neq E[Y(0)] \ . \tag{18.16}$$

In this case, one may still consider the permutation test that results from considering all possible permutations to the vector of treatment assignment $(D_1, \ldots, D_n)$. Unfortunately, such an approach does not lead to a valid test and may over-reject in finite samples. These test may be asymptotically valid though, is one carefully chooses an appropriate test statistic.

The distinction between the null hypothesis in (18.15) and that in (18.16) and their implications on the properties of permutation tests are often ignored in applied research.

Randomization test are often dismissed in applied research due to the belief that the randomization hypothesis is too strong to hold in a real empirical application. For example, the distribution P may not be symmetric in hypotheses about the mean of X. However, it turns out that the randomization test is asymptotically valid (under certain conditions), even when P is not symmetric. See Bugni et al. (2018) for an example in the context of randomized controlled experiments. Moreover, recent developments on the asymptotic properties of randomization tests show that such a construction may be particularly useful in regression models with a fixed and small number of clusters, see Canay et al. (2017). The approach does not require symmetry in the distribution of X, but rather symmetry in the asymptotic distribution of $\hat{\theta}_n$ - which automatically holds when these estimators are asymptotically normal. We cover these topics in Econ 481.

Bibliography

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018): "Inference under Covariate Adaptive Randomization," *Journal of the American Statistical Association*, 113, 1784–1796.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017): "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 85, 1013–1030.

Lehmann, E. and J. P. Romano (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer, New York.

Bibliography

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, 105, 493–505.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.

Bell, R. M. and D. F. McCaffrey (2002): "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology*, 28, 169–182.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Billingsley, P. (1995): *Probability and Measure*, Wiley-Interscience.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018): "Inference under Covariate Adaptive Randomization," *Journal of the American Statistical Association*, 113, 1784–1796.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90, 414–427.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017): "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 85, 1013–1030.

Canay, I. A., A. Santos, and A. M. Shaikh (2021): "The wild bootstrap with a "small" number of "large" clusters," *Review of Economics and Statistics*, 103, 346–363.

Card, D. and A. B. Krueger (1994): "Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 772–793.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2016): "On Cross-Validated LASSO," *arXiv preprint arXiv:1605.02214*.

Conley, T. G. and C. R. Taber (2011): "Inference with "difference in differences" with a small number of policy changes," *The Review of Economics and Statistics*, 93, 113–125.

De Chaisemartin, C. and X. D'Haultfoeuille (2022): "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," Tech. rep., National Bureau of Economic Research.

Fan, J., J. Lv, and L. Qi (2011): "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, 291–317.

Hansen, B. E. (2019): "Econometrics," University of Wisconsin - Madison.

Hansen, B. E. and S. Lee (2019): "Asymptotic theory for clustered samples," *Journal of econometrics*, 210, 268–290.

Hansen, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141, 597–620.

Horowitz, J. L. (2015): "Variable selection and estimation in high-dimensional models," *Canadian Journal of Economics*, 48, 389–407.

Ibragimov, R. and U. K. Müller (2010): "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28, 453–468.

Imbens, G. and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of Economic Studies*, 933–959.

Imbens, G. W. and M. Kolesar (2012): "Robust standard errors in small samples: some practical advice," Tech. rep., National Bureau of Economic Research.

Knight, K. and W. Fu (2000): "Asymptotics for lasso-type estimators," *The Annals of statistics*, 28, 1356–1378.

Lehmann, E. and J. P. Romano (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

MacKinnon, J. G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

Manski, C. F. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729–738.

Mikusheva, A. (2007): "Course materials for Time Series Analysis," MIT OpenCourseWare, Massachusetts Institute of Technology.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–08.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer, New York.

Wang, H., B. Li, and C. Leng (2009): "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71, 671–683.

White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 817–838.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Zhao, P. and B. Yu (2006): "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.