Regression Notes

1.1 Ordinary Least Squares

Derivation 1.1: Least Squares Estimator

Let $\{(Y_i, X_i)\}_{i=1}^n$ be a random sample from the distribution of (Y, X). The function $S(b) = \mathbb{E}[(Y - X'b)^2]$ is unknown to us, but we can estimate it consistently with the following expression:

$$\hat{S}(b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2 = \frac{1}{n}\text{SSE}(b),$$

where $\text{SSE}(b) = \sum_{i=1}^n (Y_i - X_i'b)^2$ is the sum of the squared errors. Instead of minimizing $S(b)$, we can minimize $\hat{S}(b)$. This is the process that will yield the least squares estimator. More precisely, we have:

$$\hat{\beta} = \text{argmin}_{b\in\mathbb{R}^K}\hat{S}(b) = \text{argmin}_{b\in\mathbb{R}^K}\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2$$

Let us proceed with the minimization.

$$\hat{S}(b) = \left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2\right) - 2b'\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i\right) + b'\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)b$$
$$\frac{\delta}{\delta b}\hat{S}(b) = -2\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i\right) + 2\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)b$$

Setting the above expression to zero, we get:

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i = \left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)\hat{\beta}$$
$$\implies \hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i\right),$$

provided, of course, that $\frac{1}{n}\sum_{i=1}^n X_iX_i'$ is full rank/invertible/positive definite.

Derivation 1.2: Matrix Notation

When deriving these results, matrix notation can significantly simplify things. So here we define:

$$\mathbf{Y} = \begin{bmatrix}Y_1\\ \vdots \\ Y_n\end{bmatrix} \in \mathbb{R}^n, \mathbf{X} = \begin{bmatrix}X_1'\\ \vdots \\ X_n'\end{bmatrix} \in \mathbb{R}^n, \text{ and } \mathbf{U} = \begin{bmatrix}U_1\\ \vdots \\ U_n\end{bmatrix} \in \mathbb{R}^n,$$

We can then condense the n resulting equations into a single system as follows:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$$

The previous expressions from the derivation of the least squares estimators have matrix notation analogues.

The sum of square error function, for example, is:

$$\text{SSE}(b) = (\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b)$$

Derivation 1.3: Linear Projections

The fitted values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ are the projection of the vector $\mathbf{Y}$ on the column space of the matrix $\mathbf{X}$. We can express $\hat{\mathbf{Y}}$ as:

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

$\mathbf{P_X}$ takes the vector $\mathbf{Y}$ and returns the vector $\hat{\mathbf{Y}}$, the element of col($\mathbf{X}$) closest to $\mathbf{Y}$.

The regression residuals can be obtained from:

$$\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y} \equiv \mathbf{M_X Y},$$

where $\mathbf{M_X} = \mathbf{I} - \mathbf{P_X}$ is called the annihilator or residual-maker matrix. $\mathbf{M_X}$ is idempotent and symmetric, like $\mathbf{P_X}$. We can decompose:

$$\mathbf{Y} = \mathbf{P_X Y} + \mathbf{M_X Y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{U}}$$

Definition 1.1: Partitioned Regression

Suppose we have several independent variables but are ultimately only interested in a small subset of the coefficients associated with them. More concretely, suppose we have:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{U},$$

where $\mathbf{X}$ has been partitioned into $[\mathbf{X_1}\ \mathbf{X_2}]$, and we also have $\mathbf{X} \in \mathbb{R}^{n\times K}$, $\mathbf{X_1} \in \mathbb{R}^{b\times K_1}$, $\mathbf{X_2} \in \mathbb{R}^{n\times K_2}$, and $K_1 + K_2 = K$. We are only interested in $\hat{\beta}_1$, the first K_1 coefficients of $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. A naive way to do this would be to directly compute $\hat{\beta}$ and extract its first K_1 entries. Another way is through the Frisch-Waugh-Lovell (FWL) Theorem, which says:

$$\hat{\beta}_1 = (\mathbf{X_1}'\mathbf{M_{X_2}}\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{M_{X_2}}\mathbf{Y},$$

where $\mathbf{M_{X_2}}$ is, by definition, equal to $\mathbf{I} - \mathbf{X_2}(\mathbf{X_2}'\mathbf{X_2})^{-1}\mathbf{X_2}'$. This is called a partitioned regression or residual regression because it is equivalent to a regression involving regression residuals. In other words, it is equivalent to the following two procedures, as I will show:

Procedure 1

1. Regress $\mathbf{Y}$ on $\mathbf{X}_2$ and keep the residuals, $\tilde{\mathbf{U}} = \mathbf{M}_{\mathbf{X}_2}\mathbf{Y}$
2. Regress $\mathbf{X}_1$ on $\mathbf{X}_2$ and keep the residuals, $\tilde{\mathbf{X}}_1 = \mathbf{M}_{\mathbf{X}_2}\mathbf{X}_1$
3. Regress $\tilde{\mathbf{U}}$ on $\tilde{\mathbf{X}}_1$ and obtain $\hat{\beta}_1$
   $\hat{\beta}_1 = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\tilde{\mathbf{U}} =$
   $= (\mathbf{X_1}'\mathbf{M}_{\mathbf{X_2}}'\mathbf{M_{X_2}}\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{M}_{\mathbf{X_2}}'\mathbf{M_{X_2}}\mathbf{Y} =$
   $= (\mathbf{X_1}'\mathbf{M_{X_2}}\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{M_{X_2}}\mathbf{Y}$

Procedure 2

1. Regress $\mathbf{X}_1$ on $\mathbf{X}_2$ and keep the residuals, $\tilde{\mathbf{X}}_1 = \mathbf{M}_{\mathbf{X}_2}\mathbf{X}_1$
2. Regress $\mathbf{Y}$ on $\tilde{\mathbf{X}}_1$ to obtain $\hat{\beta}_1$
   $\hat{\beta}_1 = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\mathbf{Y} =$
   $= (\mathbf{X_1}'\mathbf{M}_{\mathbf{X_2}}'\mathbf{M_{X_2}}\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{M}_{\mathbf{X_2}}'\mathbf{Y} =$
   $= (\mathbf{X_1}'\mathbf{M_{X_2}}\mathbf{X_1})^{-1}\mathbf{X_1}'\mathbf{M_{X_2}}\mathbf{Y}$

Derivation 1.4: Instrumental Variables

Let us return to the old model: $Y = X'\beta + U$. This time, though, suppose we *cannot* make the same assumption that $\mathbb{E}[XU] = 0$, because we have good reason to suspect that the variable U has an effect on X, in addition to its effect on Y. But on the bright side, suppose also that we have some instrument $Z \in \mathbb{R}^{L\times 1}$, which satisfies the following two assumptions:

1. **Exogeneity:** $E[ZU] = 0$
2. **Relevance:** $\text{rank}(E[ZX']) = K$

Z contains all exogenous parts of X, as well as an instrument Z_1 for the endogenous parts of X. In other words, assuming that $L = K$, we have:

$$\mathbf{X} = \begin{bmatrix} 1 \\ X_2 \\ \vdots \\ X_{K-1} \\ X_K \end{bmatrix} \in \mathbb{R}^{K\times 1} \text{ and } \mathbf{Z} = \begin{bmatrix} 1 \\ X_2 \\ \vdots \\ X_{K-1} \\ Z_1 \\ \vdots \\ Z_{K_2} \end{bmatrix} \in \mathbb{R}^{L\times 1}$$

Let us first show that β is identified, meaning it can bewritten as a function of population moments in observable variables. Starting with the base model $Y = X'\beta + U$ and premultiplying with the instrument Z, we get:

$$ZY = ZX'\beta + ZU$$

Taking the expectation of both sides, we get:

$$\mathbb{E}[ZY] = \mathbb{E}[ZX']\beta + \mathbb{E}[ZU]$$

Finally, note that $\mathbb{E}[ZU] = 0$ by assumption. Solving for β, we get:

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

β is clearly identified, as we can estimate it using observed quantities in the data. To see this, let us use the sample analogues for each of the components in the above expression:

$$\hat{\beta}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i Y_i\right) = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}$$

From the above expression, notice that OLS is really "just" a special case of IV when $Z_i = X_i$.

With this newly-derived IV estimator in mind, let us explore some of its finite sample properties. First, its bias. Note that in any applied environment in which we would actually use IV instead of OLS, $\hat{\beta_{IV}}$ is biased. Let me show you what I mean.

Derivation 1.5: Bias of $\hat{\beta}_{\mathbf{IV}}$

We first calculate $\mathbb{E}[\hat{\beta}_{IV}|\mathbf{Z}]$.

$$\mathbb{E}[\hat{\beta}_{IV}|\mathbf{Z}] = E[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}|\mathbf{Z}] =$$
$$= \mathbb{E}[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\beta + \mathbf{U})|\mathbf{Z}] =$$
$$= \beta + \mathbb{E}[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{U}|\mathbf{Z}]$$

Observe that in the above conditional expression, we cannot pull $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'$ out in front of the expectation, because it contains an $\mathbf{X}$. Naively, we think we can solve this problem by conditioning on both $\mathbf{Z}$ and $\mathbf{X}$. Not quite:

$$\mathbb{E}[\hat{\beta}_{IV}|\mathbf{Z}, \mathbf{X}] = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbb{E}[\mathbf{U}|\mathbf{Z}, \mathbf{X}]$$

Even if $\mathbb{E}[\mathbf{U}|\mathbf{Z}] = 0$, in order to have $\mathbb{E}[\mathbf{U}|\mathbf{Z}, \mathbf{X}] = 0$, we would still need $\mathbb{E}[\mathbf{U}|\mathbf{X}] = 0$. But if we had $\mathbb{E}[\mathbf{U}|\mathbf{X}] = 0$, we could just do OLS; this is the exogeneity assumption! We would not need to bother with instrumental variables if this were the case.

So, in all practical use-cases of IV, $\hat{\beta}_{IV}$ is biased, due to the fact that $\mathbb{E}[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{U}|\mathbf{Z}] \neq 0$.

Proof 1.1: Consistency of $\hat{\beta}_{\mathbf{IV}}$

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\beta + \mathbf{U}) =$$
$$= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{U} = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{U} =$$
$$\underset{(1)}{=} \beta + \left(\frac{1}{N}\sum_{i=1}^{N} Z_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i U_i\right) \underset{(2)}{\xrightarrow{p}} \beta + \mathbb{E}[ZX']^{-1}\mathbb{E}[ZU] \underset{(3)}{=} \beta$$

(1) follows from converting the expression from matrix to scalar notation, (2) follows from the WLLN and CMT, and (3) follows from the relevance assumption, which allows us to invert $\mathbb{E}[ZX']$, and from the exogeneity assumption, which sets $\mathbb{E}[ZU]$ to zero.

Derivation 1.6: Two Stage Least Squares

A related strategy—indeed, we will see in a moment that "related" is an understatement—is to perform this regression in two stages.

1. In the first stage, we regress $\mathbf{X}$ on $\mathbf{Z}$ and get the fitted values, $\hat{\mathbf{X}}$. This is the cleanup stage. We are extracting out the "clean" part of $\mathbf{X}$—that is, the part of it that is not correlated with the error term. The result, $\hat{\mathbf{X}}$, gives us the predicted values of the independent variable, which capture *only* the variation coming from the instrument.

   So we have $\hat{\mathbf{X}} = \mathbf{P_Z X}$, where $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.

2. In the second stage, we regress $\mathbf{Y}$ on $\hat{\mathbf{X}}$ and get $\hat{\beta}_{2SLS}$. Since $\hat{\mathbf{X}}$ only contains variation from the instrument, we have stripped out the endogeneity. So regression $\mathbf{Y}$ on $\hat{\mathbf{X}}$ will give us a consistent estimate of the effect of $\mathbf{X}$.

$$\hat{\beta}_{\mathbf{2SLS}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Y} = ((\mathbf{P_Z X})'\mathbf{P_Z X})^{-1}(\mathbf{P_Z X})'\mathbf{Y} =$$
$$= (\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{P_Z X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{Y} = (\mathbf{X}'\mathbf{P_Z X})^{-1}(\mathbf{X}'\mathbf{P_Z Y})$$

We have performed regression in two stages, hence the name "two stage" regression. If this sounds familiar, it is because

Let us take a look at some of the properties of $\hat{\beta}_{2SLS}$.

Proof 1.2: Consistency of $\hat{\beta}_{2SLS}$

$$\hat{\beta}_{2SLS} = (\mathbf{X'P_Z X})^{-1}(\mathbf{X'P_Z Y}) = (\mathbf{X'P_Z X})^{-1}(\mathbf{X'P_Z}(\mathbf{X}\beta + \mathbf{U})) =$$
$$= (\mathbf{X'P_Z X})^{-1}(\mathbf{X'P_Z X})\beta + (\mathbf{X'P_Z X})^{-1}(\mathbf{X'P_Z U}) =$$
$$= \beta + (\mathbf{X'P_Z X})^{-1}(\mathbf{X'P_Z U}) = \beta + (\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X})^{-1}(\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'U})$$
$$\text{plim}(\hat{\beta}_{2SLS}) = \beta + \text{plim}[(\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X})^{-1}(\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'U})] =$$
$$= \beta + \left[\text{plim}\sum_{i=1}^{N} X_i Z_i' \left(\text{plim}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1} \text{plim}\sum_{i=1}^{N} Z_i X_i'\right]^{-1} \text{plim}\sum_{i=1}^{N} X_i Z_i' \left(\sum_{i=1}^{N} Z_i Z_i'\right)^{-1} \sum_{i=1}^{N} Z_i U_i'$$

We want to invoke the WLLN (and the CMT) here, but in order to do so, we need sample means, not sums. So we can multiply each of the above sums by $\frac{1}{N}$. It turns out that these $\frac{1}{N}$s end up canceling out, because we'll have equally as many N terms in the numerator and denominator.

$$\beta + \left[\text{plim}\frac{1}{N}\sum_{i=1}^{N} X_i Z_i' \left(\text{plim}\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1} \text{plim}\frac{1}{N}\sum_{i=1}^{N} Z_i X_i'\right]^{-1} \text{plim}\frac{1}{N}\sum_{i=1}^{N} X_i Z_i' \left(\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} Z_i U_i'$$
$$\xrightarrow{p} \beta + \left[\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX']\right]^{-1}\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZU]$$

$\mathbb{E}[ZU] = 0$ by the assumption of exogeneity of our instrument. $\mathbb{E}[XZ']$ and $\mathbb{E}[ZZ']$ exist and are invertible by the relevance assumption. So the second term above goes to zero, and we get $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$. Therefore, $\hat{\beta}_{2SLS}$ is consistent.

Proof 1.3: Asymptotic Normality of $\hat{\beta}_{\mathbf{2SLS}}$

Let us define $C = \mathbb{E}[ZX']$ and $W = \mathbb{E}[ZZ']$. Both are invertible by assumption. We have:

$$\sqrt{N}(\hat{\beta}_{2SLS} - \beta) = \sqrt{N}(\mathbf{X'P_Z X})^{-1}\mathbf{X'P_Z U} = \sqrt{N}(\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X})^{-1}\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'U}$$

Once again, we would like to invoke the WLLN and CMT, so we need to convert the above expression into sample means.

$$\sqrt{N}(\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X})^{-1}\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'U} = (\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X})^{-1}\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\sqrt{N}(\mathbf{Z'U}) =$$
$$= \left[\frac{1}{N}\sum_{i=1}^{N} X_i Z_i' \left(\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} Z_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Z_i' \left(\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1}\right] \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i U_i\right) \quad (*)$$

Let us decompose the above expression (*) into subparts and analyze them separately.

$$\left[\frac{1}{N}\sum_{i=1}^{N} X_i Z_i' \left(\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} Z_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Z_i' \left(\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1}\right]$$
$$\xrightarrow{p} (\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX'])^{-1}\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1} = (C'W^{-1}C)^{-1}C'W^{-1}$$

For the remainder of the expression in (*), since $\mathbb{E}[ZU] = 0$ by assumption, we have:

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i U_i\right)\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i U_i - 0\right) = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i U_i - \mathbb{E}[ZU]\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(ZU)) =$$
$$= \mathcal{N}(0, \mathbb{E}[ZU(ZU)'] - \mathbb{E}[ZU]\mathbb{E}[ZU]') = \mathcal{N}(0, \mathbb{E}[ZUU'Z']) = \mathcal{N}(0, \mathbb{E}[U^2 ZZ']) \underset{(1)}{=} \mathcal{N}(0, \sigma^2\mathbb{E}[ZZ']) = \mathcal{N}(0, \sigma^2 W)$$

(1) follows from the homoskedasticity assumption. So by Slutsky's theorem, we have:

$$(*) \xrightarrow{d} (C'W^{-1}C)^{-1}C'W^{-1}\mathcal{N}(0, \sigma^2 W) = \mathcal{N}(0, (C'W^{-1}C)^{-1}C'W^{-1}(\sigma^2 W)W^{-1}C(C'W^{-1}C)^{-1}) =$$
$$= \mathcal{N}(0, \sigma^2(C'W^{-1}C)^{-1}) = \mathcal{N}(0, \sigma^2(\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX'])^{-1}$$

Proof 1.4:

From the expression $\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}$, we see rather clearly that our calculation of $\hat{\beta}_{IV}$ depends on the invertibility of $\mathbf{Z}'\mathbf{X}$. Since $\mathbf{Z} \in \mathbb{R}^{n\times L}$ and $\mathbf{X} \in n \times K$, we need $L = K$ to have $\mathbf{Z}'\mathbf{X}$ be square (and invertible). What happens, then, when $L > K$?

First of all, what would $L > K$ correspond to in the "real world?" Here, we are looking at a setting where we have more instruments than regressors (wishful thinking, really, since most of the time we are lucky to even have one good instrument). But for the sake of the argument and then the math, let us assume that have $L > K$ and therefore that $\mathbb{E}[ZX'] \in \mathbb{R}^{L\times K}$ is not square, and therefore not invertible. This is called the over-identified case, as opposed to the just-identified case when $L = K$.

One solution would simply be to select any subset of Z containing K instruments, and these instruments will be exogenous. Alternatively, we can consider a new set of instruments $\Pi'Z$ where $\Pi \in \mathbb{R}^{L\times K}$ such that $\Pi'Z \in \mathbb{R}^K$ (this is a generalization of the first solution). These new instruments $\Pi'Z$ are still exogenous because:

$$\mathbb{E}[(\Pi'Z)U] = \Pi'\mathbb{E}[ZU] = \Pi'(\mathbf{0}) = \mathbf{0}$$

At this point, we can identify β:

$$\mathbb{E}[Z(Y - X'\beta)] = \mathbb{E}[ZY] - \mathbb{E}[ZX']\beta = \mathbf{0}$$
$$\Pi'\mathbb{E}[ZY] - \Pi'\mathbb{E}[ZX']\beta = \mathbf{0}$$

As you can see, we are getting ready to invert $\Pi'\mathbb{E}[ZX']$, which is square and invertible as long as Π' and $\mathbb{E}[ZX']$ both have full rank and $L \geq K$. So with these conditions, we get:

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

for any full rank matrix $\Pi \in \mathbb{R}^{L\times K}$. Thus, as shown, β is identified. To estimate β, we can replace the components of the above expression with their sample counterparts. This will result in the Generalized IV

estimator.

$$\hat{\beta}_{GIV}(\hat{\Pi}) = \left(\hat{\Pi}' \frac{1}{N} \sum_{i=1}^{N} Z_i X_i'\right)^{-1} \left(\hat{\Pi}' \frac{1}{N} \sum_{i=1}^{N} Z_i Y_i\right) = (\hat{\Pi}' \mathbf{Z}' \mathbf{X})^{-1} (\hat{\Pi}' \mathbf{Z}' \mathbf{Y})$$

The question now is how exactly to choose Π (and $\hat{\Pi}$). One option is to let Π be the coefficients in the projection of X on Z.

$$\mathbb{L}(X|Z) = Z'\Pi$$
$$\Pi = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX']$$

We can estimate Π using the least squares estimator of $\mathbf{X}$ on $\mathbf{Z}$.

$$\hat{\Pi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

Let us plug in this expression for $\hat{\Pi}$ into our result for the Generalized IV estimator above:

$$\hat{\beta}_{GIV}(\hat{\Pi}) = (\hat{\Pi}\mathbf{Z}'\mathbf{X})^{-1}(\hat{\Pi}\mathbf{Z}'\mathbf{Y}) = ([(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]'\mathbf{Z}'\mathbf{X})^{-1}([(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]'\mathbf{Z}'\mathbf{Y}) =$$
$$= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}) = (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{Y})$$

This is the same as the 2SLS estimator.