# Linear projections

- The linear projection is defined as follows

$$\mathrm{L}\left(y|\,1, x_1, \ldots, x_K\right) = \mathrm{L}\left(y|\,1, \boldsymbol{x}\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_K x_K = \boldsymbol{\beta}_0 + \boldsymbol{x}\boldsymbol{\beta}$$

and $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{\beta} = \left[\mathrm{Var}\left(\boldsymbol{x}\right)\right]^{-1} \mathrm{Cov}\left(\boldsymbol{x}, y\right)$$

$\boldsymbol{\beta}_0 = \mathrm{E}\left(y\right) - \mathrm{E}\left(\boldsymbol{x}\right)\boldsymbol{\beta}$

- We can always write

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_K x_K + u$$

and from definition of linear projection $\mathrm{E}(u) = 0$, $\mathrm{Cov}(x_j, u) = 0$, $j = 1, \ldots, N$

- It can be shown that linear projection is **minimum mean square error linear predictor** of $y$

$$\min_{b_0, \boldsymbol{b}} \mathrm{E}\left[y - b_0 - \boldsymbol{x}\boldsymbol{b}\right]^2$$

- Iteration property

$$\mathrm{L}(y\,|1, \boldsymbol{x}) = \mathrm{L}(\mathrm{L}(y\,|1, \boldsymbol{x}, \boldsymbol{z})\,|\boldsymbol{x})$$

- Also

$$\mathrm{L}(y\,|1, \boldsymbol{x}) = \mathrm{L}(\mathrm{E}(y\,|1, \boldsymbol{x}, \boldsymbol{z})\,|\boldsymbol{x})$$

# Endogeneity problem

- Explanatory variable $x_j$ is said to be endogenous if $\mathrm{E}\left(u|\, x_j\right) \neq 0$
- Endogeneity problem arises because:
  - **Omitted variable problem**

$$\mathrm{E}\left(u|\, \boldsymbol{x}, q\right) \neq \mathrm{E}\left(u|\, \boldsymbol{x}\right)$$

    * In such a case model in error form is:

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_K x_K + \boldsymbol{\beta}_q q + u$$

    * if $q$ is omitted from regression part then it has to be included in error term

$$u^* = \boldsymbol{\beta}_q q + u$$

    * **But**: if $q$ and $x_j$ are correlated than $x_j$ and $u^*$ are correlated and $x_j$ is endogenous
  - **Measurement error**. If instead of $x_j$ we approximate value $x_j^*$, than measurement error will become part of $u$ and $u$ can become correlated with $x_j^*$
  - **Simultaneity.** Simultaneity arises if $y$ influences (e.g. with feedback) one of the explanatory variables

- Sometimes the distinction between this three sources of endogeneity is not quite sharp

# Omitted variable problem

- Model with additive omitted variable

$$\mathrm{E}\left(y|\, x_1, x_2, \ldots, x_K, q\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_K x_K + \boldsymbol{\gamma} q$$

  where $q$ is omitted variable

- We are interested in partial effects of $x_j$ on $y$ holding all other variables *including* $q$ constant

- Model in error form

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_K x_K + \boldsymbol{\gamma} q + v$$

$$\mathrm{E}\left(v \mid x_1, \ldots, x_K, q\right) = 0$$

- $v$ - structural error
- Assume that $\mathrm{E}(q) = 0$, this assumption only influences constant
- Remove $v$ from model in error form

$$\begin{aligned} y &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_K x_K + u \\ u &= \boldsymbol{\gamma} q + v \end{aligned}$$

- From $\mathrm{E}(q) = \mathrm{E}(v) = 0$ we have $\mathrm{E}(u) = 0$
- **But**: $u$ is only uncorrelated with all $x_j$ if $q$ is uncorrelated with all $x_j$

- If $q$ is correlated with one of $x_j$ than we have endogeneity problem and by $OLS$ we cannot estimate consistently *any* of $\boldsymbol{\beta}_j$

- The asymptotic bias resulting from omitted variable problem is called omitted variable inconsistency or omitted variable bias

- Linear projection of $q$ onto explanatory variables

$$q = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 x_1 + \ldots + \boldsymbol{\delta}_K x_K + r$$

- Substituting this equation into our model we obtain:

$$y = (\boldsymbol{\beta}_0 + \boldsymbol{\gamma}\boldsymbol{\delta}_0) + (\boldsymbol{\beta}_1 + \boldsymbol{\gamma}\boldsymbol{\delta}_1)\, x_1 + \ldots + (\boldsymbol{\beta}_K + \boldsymbol{\gamma}\boldsymbol{\delta}_K)\, x_K + v + \boldsymbol{\gamma} r$$

- By definition of linear projection $\mathrm{E}\,(r) = 0$, $\mathrm{Cov}\,(x_j, r) = 0$ for $j = 1, \ldots, N$

- The error $v + \boldsymbol{\gamma} r$ has zero mean and is uncorrelated with all the regressors

- Therefore $OLS$ estimate $\text{plim}\left(\widehat{\boldsymbol{\beta}}_j\right) = \boldsymbol{\beta}_j + \boldsymbol{\gamma}\boldsymbol{\delta}_j$

- If we assume that $q$ is correlated with only one variable, say $x_K$ so that $\boldsymbol{\delta}_j = 0$ for $j \neq K$ and from definition of linear projection

$$\boldsymbol{\delta}_K = \frac{\text{Cov}\,(x_K, q)}{\text{Var}\,(x_K)}$$

  So

$$\text{plim}\,\widehat{\boldsymbol{\beta}}_K = \boldsymbol{\beta}_K + \boldsymbol{\gamma}\frac{\text{Cov}\,(x_K, q)}{\text{Var}\,(x_K)}$$

- This means that:
  - $\boldsymbol{\gamma}\,\text{Cov}\,(x_K, q) > 0$ than bias of $\widehat{\boldsymbol{\beta}}_K$ is positive

– $\boldsymbol{\gamma}\operatorname{Cov}(x_K, q) < 0$ than bias of $\widehat{\boldsymbol{\beta}}_K$ is negative

**Example 1.** *(Wooldridge) Wage equation with unobserved ability*

$$\log(wage) = \boldsymbol{\beta}_1 \mathit{exper} + \boldsymbol{\beta}_2 \mathit{exper}^2 + \boldsymbol{\beta}_3 \mathit{educ} + \boldsymbol{\gamma} \mathit{abil} + v$$

*if abil is only correlated with educ:*

$$\mathit{abil} = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_3 \mathit{educ} + r$$

*and abil is omitted from the model, then estimated coefficient for educ is equal in the limit* $\operatorname{plim}(\boldsymbol{\beta}_3) = \boldsymbol{\beta}_3 + \boldsymbol{\gamma}\boldsymbol{\delta}_3$*. If* $\boldsymbol{\delta}_3 > 0$ *then the influence of education is overestimated.*

## Proxy variable solution for omitted variable

- The proxy variable should be
  - redundant (ignorable)

$$\mathrm{E}\left(y|\,\boldsymbol{x},q,z\right)=\mathrm{E}\left(y|\,\boldsymbol{x},q\right)$$

  - correlation between $q$ and $x_1,\ldots,x_K$ should be zero once the effect of $z$ is removed out

$$L\left(q|\,1,x_1,\ldots,x_K,z\right)=L\left(q|\,z\right)$$

This condition can also be expressed in error form

$$q=\boldsymbol{\theta}_0+\boldsymbol{\theta}_1 z+r$$

by definition of linear projection $\mathrm{E}\left(r\right)=0$, $\mathrm{Cov}\left(x_j,r\right)=0$

- Now we can write

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\gamma}\boldsymbol{\theta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ldots + \boldsymbol{\beta}_k x_K + \boldsymbol{\gamma}\boldsymbol{\theta}_1 z + (\boldsymbol{\gamma} r + v)$$

- Under two assumptions made $(\boldsymbol{\gamma} r + v)$ is uncorrelated with $x_j$ for $j = 1, \ldots, N$

- Conditions of consistency of $OLS$ imply that $\beta$ can be consistently estimated in this case

- Inperfect proxy

$$q = \boldsymbol{\theta}_0 + \boldsymbol{\rho}_1 x_1 + \ldots + \boldsymbol{\rho}_K x_K + \boldsymbol{\theta}_1 z + r$$

- This gives $\operatorname{plim}\left(\widehat{\boldsymbol{\beta}}_j\right) = \boldsymbol{\beta}_j + \boldsymbol{\gamma}\boldsymbol{\rho}_j$

- We may hope that the bias is much smaller than $\boldsymbol{\beta}_j$

- Models with interactions with unobservables

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ldots + \boldsymbol{\beta}_K x_K + \boldsymbol{\gamma}_1 q + \boldsymbol{\gamma}_2 x_K q + v$$

- Partial effect of $x_K$

$$\frac{\partial \operatorname{E}(y|\, \boldsymbol{x}, q)}{\partial x_K} = \boldsymbol{\beta}_K + \boldsymbol{\gamma}_2 q$$

- We cannot estimate the partial effect as $q$ is not observable

- Assuming that $\operatorname{E}(q) = 0$ we can however estimate the *average partial effect*

$$\operatorname{E}(\boldsymbol{\beta}_K + \boldsymbol{\gamma}_2 q) = \boldsymbol{\beta}_K$$

or for binary variable

$$\mathrm{E}\left(y|\, x_1,\ldots,x_{K-1},1,q\right)-\mathrm{E}\left(y|\, x_1,\ldots,x_{K-1},0,q\right)=\boldsymbol{\beta}_K$$

- If $\mathrm{E}\left(q|\, \boldsymbol{x}\right)=0$ than we can estimate this with $OLS$
- In the case of using proxy

$$\mathrm{E}\left(q|\, \boldsymbol{x,}z\right)=\boldsymbol{\theta}_1 z+r$$

- Estimated equation

$$\begin{aligned}\mathrm{E}\left(y|\, \boldsymbol{x},z\right) &= \boldsymbol{\beta}_0+\boldsymbol{\beta}_1 x_1+\boldsymbol{\beta}_2 x_2+\ldots+\boldsymbol{\beta}_K x_K \\ &\quad +\boldsymbol{\gamma}_1\boldsymbol{\theta}_1 z+\boldsymbol{\gamma}_2 x_K\boldsymbol{\theta}_1 z+\boldsymbol{\gamma}_1 r+\boldsymbol{\gamma}_2 x_K r+v\end{aligned}$$

- The result of estimation will be correct if proxy variable has expected value equal to zero

- If proxy variable has mean significantly different from zero then we may demean it to have the zero mean condition fulfilled

**Example 2.** *(Wooldridge) Using IQ as proxy for ability (Blackburn and Neumark 1992)*

$$\begin{aligned}\log(wage) &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \textit{exper} + \boldsymbol{\beta}_2 \textit{tenure} + \boldsymbol{\beta}_3 \textit{married} \\ &\quad + \boldsymbol{\beta}_4 \textit{south} + \boldsymbol{\beta}_5 \textit{urban} + \boldsymbol{\beta}_6 \textit{black} + \boldsymbol{\beta}_7 \textit{educ} + \boldsymbol{\gamma} \textit{IQ} + v\end{aligned}$$

*Estimated coefficient for education:* $\boldsymbol{\beta}_7 = .065$ *and with IQ included* $\boldsymbol{\beta}_7 = .054$*. Indeed it seems that coefficient for education overestimated.*

# Instrumental Variable Estimator (IV)

- Linear model

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ldots + \boldsymbol{\beta}_K x_K + u \tag{1}$$

$$\mathrm{E}\,(u) = 0,\ \ \mathrm{Cov}\,(x_j, u) = 0,\ \ j = 1, 2, \ldots, K-1$$

- $x_1, \ldots, x_{K-1}$ are uncorrelated with $u$ <u>but $x_K$ is correlated with $u$</u> ($x_K$ is endogenous)

- $OLS$ estimator of the parameters is inconsistent for *all* the parameters $\boldsymbol{\beta}_j$ for $j = 1, \ldots, K$

- The method of instrumental variables gives the estimation technique which solves this problem

- **Instrumental variable estimator** $(IV)$

- We need an observable variable $z_1$, not included in regression (redundant) which satisfies:

1. In uncorrelated with the error therm $u$

$$\mathrm{Cov}\,(z_1, u) = 0$$

2. Coefficient on $z_1$ in linear projection of $x_K$ on $x_1, \ldots, x_K, z_1$ is not equal to zero

$$x_K = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 x_1 + \ldots + \boldsymbol{\delta}_{K-1} x_{K-1} + \boldsymbol{\theta}_1 z_1 + r_K \tag{2}$$

$$\boldsymbol{\theta}_1 \neq 0$$

- The second condition can also be formulated as the requirement that $z_1$ is partially correlated with $x_K$

- $z_1$ satisfying these two conditions is called *instrumental variable* (instrument) for $x_K$

- Equation (2) is called *reduced form equation* for endogenous variable $x_K$

- Equation (1) can be rewritten as

$$y = \boldsymbol{x}\boldsymbol{\beta} + u$$

- The vector of all exogenous (uncorrelated with $u$) variables is

$$z = (1, x_1, \ldots, x_{K-1}, z_1)$$

- From assumptions

$$\mathrm{E}\left(\boldsymbol{z}'u\right) = 0$$

- Multiply (2), by $\boldsymbol{z}'$ and take expectations

$$\mathrm{E}\left(\boldsymbol{z}'y\right) = \mathrm{E}\left(\boldsymbol{z}'\boldsymbol{x}\boldsymbol{\beta}\right) + \underbrace{\mathrm{E}\left(\boldsymbol{z}'u\right)}_{0}$$

- If $\mathrm{Rank}\left[\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{z}\right)\right] = K$ (this condition is satisfied if $\boldsymbol{\theta}_1 \neq 0$) then

$$\boldsymbol{\beta} = \left[\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{x}\right)\right]^{-1} \mathrm{E}\left(\boldsymbol{z}'y\right)$$

- Condition that $\boldsymbol{\theta}_1 \neq 0$ can be tested by checking whether $z_1$ is significant in reduced form equation

- Parameter $\boldsymbol{\beta}$ is said to be identified if it can be expressed as a function of expectations of the data

- Using analogy principle (replacing expectations with data means) we obtain:

$$\widehat{\boldsymbol{\beta}} = \left( N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i' \boldsymbol{x}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' y_i \right) = \left( \boldsymbol{Z}' \boldsymbol{X} \right)^{-1} \boldsymbol{Z}' \boldsymbol{y}$$

**Example 3.** *(Wooldridge) Instrumental Variables for education in wage equation*

$$\log\left(\textit{wage}\right) = \boldsymbol{\beta}_0 + \textit{exper} + \boldsymbol{\beta}_2 \textit{exper}^2 + \boldsymbol{\beta}_3 \textit{educ} + u$$

*As an instrument for child education mother education can be used: it is correlated with child education but it should not directly influence wages.*

*Problem: mother education can be correlated with other omitted factors in wage education.*

- Angrist and Kruger (1991) instrument for education: quarter of birth
- Card (1995): instrument for education: college proximity
- *Natural experiment instruments:* Angrist (1990) effect of veteran status on wages - instrument: draft lottery number

## 2 Step Least Squares (2SLS)

- Model:

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 +, \ldots, \boldsymbol{\beta}_{K-1} x_{K-1} + \boldsymbol{\beta}_K x_K + u$$

- For simplicity we assume that only $x_K$ is endogenous
- More then one instrument for endogenous variable $x_K$
- $M$ instruments, $z_1, \ldots, z_K$ such that forn any $h$

$$\mathrm{Cov}\left(z_h, u\right) = 0$$

- Vector of exogenous variables contains exogenous explanatory and instrumental variables:

$$\boldsymbol{z} = (x_1, \ldots, x_{K-1}, z_1, \ldots, z_M)$$

- It is possible to find $M$ instrumental variable ($IV$) estimators of parameter $\boldsymbol{\beta}$ which are consistent

- Which estimator we should use?

- The most efficient one is the one using the instrument being the linear combination of $\boldsymbol{z}$ most highly correlated with $x_K$

- The most highly correlated with $x_K$ combination of $\boldsymbol{z}$ is the linear projection of $x_K$ on $\boldsymbol{z}$.

- Reduced form for $x_K$ is

$$x_K = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 x_1 + \ldots + \boldsymbol{\delta}_{K-1} x_{K-1} + \boldsymbol{\theta}_1 z_1 + \ldots + \boldsymbol{\theta}_M z_M + r_K$$

- New instrument is:

$$x_K^* = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 x_1 + \ldots + \boldsymbol{\delta}_{K-1} x_{K-1} + \boldsymbol{\theta}_1 z_1 + \ldots + \boldsymbol{\theta}_M z_M$$

- As the linear combination of exogenous variables it is uncorrelated with $u$

- $\boldsymbol{\delta}_i$ and $\boldsymbol{\theta}_j$ can easily estimated as $OLS$ gives the consistent estimators of coefficients in linear projection

$$\widehat{x}_{iK}^* = \widehat{\boldsymbol{\delta}}_0 + \widehat{\boldsymbol{\delta}}_1 x_{i1} + \ldots + \widehat{\boldsymbol{\delta}}_{K-1} x_{iK-1} + \widehat{\boldsymbol{\theta}}_1 z_{i1} + \ldots + \widehat{\boldsymbol{\theta}}_M z_{iM}$$

where $\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{X}$.

- Using $\widehat{x}_{iK}^{*}$ as instrument in $IV$ we get

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \widehat{\boldsymbol{x}}_i' \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{N} \widehat{\boldsymbol{x}}_i' y_i\right) = \left(\widehat{\boldsymbol{X}}' \boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{X}}' \boldsymbol{Y}$$

- Notice that $\widehat{\boldsymbol{X}} = \boldsymbol{Z}'\widehat{\boldsymbol{\theta}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{P}_X\boldsymbol{X}$, where projection matrix $\boldsymbol{P}_X = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ is symmetric ($\boldsymbol{P}_X' = \boldsymbol{P}_X$) and idempotent ($\boldsymbol{P}_X\boldsymbol{P}_X = \boldsymbol{P}_X$)

- Then

$$\begin{aligned}\left(\widehat{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y} &= \left(\boldsymbol{X}'\boldsymbol{P}'_X\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y} \\ &= \left(\boldsymbol{X}'\boldsymbol{P}'_X\boldsymbol{P}_X\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y} \\ &= \left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{Y}\end{aligned}$$

- **Two Stage Least Squares**:

  1. Obtain fitted values $\widehat{x}_K$ from regression of $x_K$ on $x_1, \ldots, x_{K-1}$ and $z_1, \ldots, z_K$
  2. Run regression of $y_i$ on $x_1, \ldots, x_{K-1}, \widehat{x}_K$

- Instruments have to be correlated wit endogenous explanatory variable: at least one of $\boldsymbol{\theta}_j \neq 0$

- This assumption can be tested by testing hipothesis

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \ldots = \boldsymbol{\theta}_M$$

- General case:
  - $G$ number of endogenous explanatory variables
  - $L_X$ - number of exogenous explanatory variables
  - $L_Z$ - number of instruments which are not explanatory variables
  - $L = L_Z + L_X$ total number of instruments (including $L_X$ exogenous explanatory variables)
  - $K$ - total number of explanatory variables

- Assumptions needed for consistency of $2SLS$:

1. $\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{u}\right) = 0$
2. Rank conditions:

(a) $\operatorname{Rank}\left[\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{z}\right)\right] = L$
(b) $\operatorname{Rank}\left[\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{x}\right)\right] = K$

- Condition (2a) is technical: no linear dependence between instruments

- Condition (2b) is important.

- Total number of explanatory variables $K = G + L_X$

- Necessary condition for condition (2b) is $L \geq K \implies L_Z \geq G$ - number of instruments has to be equal or bigger then the number of endogenous variables in the regression equation.

- Asymptotic distribution of $\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is normal if $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{z}_i u_i$ which can be proven from central limit theorem

- If the homoscedasticity assumption is true: $\mathrm{E}\left(u^2 \boldsymbol{z}'\boldsymbol{z}\right) = \boldsymbol{\sigma}^2\,\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{z}\right)$, where $\boldsymbol{\sigma}^2 = \mathrm{E}\left(u^2\right)$ then asymptotic variance of $\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is

$$Avar\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\sigma}^2 \left\{\mathrm{E}\left(\boldsymbol{x}'\boldsymbol{z}\right)\left[\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{z}\right)\right]^{-1}\mathrm{E}\left(\boldsymbol{z}'\boldsymbol{x}\right)\right\}$$

- If we define $2SLS$ residuals as

$$\widehat{u}_i = y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$$

then the unbiased estimator of $\boldsymbol{\sigma}^2$ is

$$\widehat{\boldsymbol{\sigma}}^2 = \left(N-K\right)^{-1}\sum_{i=1}^{N}\widehat{u}_i^2$$

- The estimator of asymptotic variance is then

$$\widehat{\boldsymbol{\sigma}}^2 \left( \sum_{i=1}^{N} \widehat{\boldsymbol{x}}_i' \widehat{\boldsymbol{x}}_i \right)^{-1} = \widehat{\boldsymbol{\sigma}}^2 \left( \widehat{\boldsymbol{X}}' \widehat{\boldsymbol{X}} \right)^{-1}$$

**Example 4.** *(Wooldridge) Instruments for educ is motheduc, fatheduc, huseeduc. Reduced form equation for educ*

$$\textit{educ} = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 \textit{exper} + \boldsymbol{\delta}_2 \textit{exper}^2 + \boldsymbol{\theta}_1 \textit{motheduc} + \boldsymbol{\theta}_2 \textit{fathed} + \boldsymbol{\theta}_3 \textit{huseduc} + r$$

$t$*-statistics for all* $\boldsymbol{\theta}_i$ *significant*

*Structural equation*

$$\log(wage) = -\underset{(.285)}{.187} + \underset{(.013)}{.043}\,\textit{exper} - \underset{(.00040)}{.00086}\,\textit{exper}^2 + \underset{(.022)}{.080}\,\textit{educ}$$

*$OLS$ coefficient for education $.107$.*

- For testing hypothesis we may use the Wald statistics or use the sum of squares
$$F = \frac{(SSR_R - SSR)/\, g}{SRR/\,(N-K)}$$
where $SSR$ is a sum of $2SLS$ residuals in unrestricted model and $SSR$ is a sum of $2SLS$ residuals in restricted model

- Beware that $SSR = \sum_{i=1}^{N} \widehat{u}_i^2$ and $\widehat{u}_i = y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}$ (not $\widehat{u}_i = y_i - \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{\beta}}$)

- We can also use the $LM$ statistics for testing

- The heteroscedasticity robust variance matrix and test statistics can be computed

## Possible Pitfalls using $2SLS$

- The weak correlation between endogenous variable on instrument ($x$ and $z$)
- It can be shown that for $y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + u$

$$\operatorname{plim} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\boldsymbol{\sigma}_u / \boldsymbol{\sigma}_{x_1}\right) \left(\boldsymbol{\rho}_{z_1 u} / \boldsymbol{\rho}_{z_1 x_1}\right)$$

- If correlation between $z_1$ and $x_1$ is small then even very small correlation between $z_1$ and $u$ can result in large asymptotic bias
- It is also the case that if instruments are poor (weakly partially correlated with endogenous explanatory variable $x_K$) that standard deviation of estimator $\widehat{\boldsymbol{\beta}}_K$ will be large

**Example 5.** *(Wooldridge) Bound, Jaeger and Baker (1995) have shown that Angrist and Kruger (1991)* $2SLS$ *estimator using instruments for education based on the date of birth behave poorly even for sample size of* $500,000$*. Reason: very weak correlation between the date of birth and the number of years of education.*

## $IV$ solution to omitted variable problem

- Model

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 +, \ldots, \boldsymbol{\beta}_{K-1} x_{K-1} + \boldsymbol{\beta}_K x_K + \boldsymbol{\gamma} q + \boldsymbol{\upsilon}$$

- Explanatory variable $q$ unobserved
- **Instrumental variable solution**:
- Find instruments

1. redundant in structural equation

2. uncorrelated with omitted variable $q$
3. correlated with endogenous variable (with a variable correlated with omitted variable)

- Use $IV$ or $2SLS$ for estimating model

$$y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 +, \ldots, \boldsymbol{\beta}_{K-1} x_{K-1} + \boldsymbol{\beta}_K x_K + u$$

where $u = \boldsymbol{\gamma} q + \boldsymbol{v}$

- **Indicator variables solution**n:

1. Find $2$ variables redundant in structural equation ($\mathrm{Cov}\left(q_1, \boldsymbol{v}\right) = \mathrm{Cov}\left(q_2, \boldsymbol{v}\right) = 0$) but correlated with $q$

$$\begin{aligned} q_1 &= \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 q + a_1 \\ q_2 &= \boldsymbol{\rho}_0 + \boldsymbol{\rho}_1 q + a_2 \end{aligned}$$

where

$$\mathrm{Cov}\,(q, a_1) = \mathrm{Cov}\,(q, a_2) = \mathrm{Cov}\,(\boldsymbol{x}, a_1) = \mathrm{Cov}\,(\boldsymbol{x}, a_2) = \mathrm{Cov}(a_1, a_2) = 0$$

2. Rearranging we get $q = -\frac{\boldsymbol{\delta}_0}{\boldsymbol{\delta}_1} + \frac{1}{\boldsymbol{\delta}_1} q_1 - \frac{a_1}{\boldsymbol{\delta}_1} = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 q_1 - \boldsymbol{\gamma}_1 a_1$ and plugging for $q$ we obtain

$$y = \boldsymbol{\alpha}_0 + \boldsymbol{x\beta} + \boldsymbol{\gamma}_1 q_1 + (\boldsymbol{v} - \boldsymbol{\gamma}_1 a_1)$$

   where $\boldsymbol{\gamma}_1 = \frac{\gamma}{\boldsymbol{\delta}_1}$ but still $q_1$ is correlated with $a_1$
3. From assumptions $q_1$ and $a_1$ are correlated but $q_2$ is not correlated with $a_1$ as $a_1$ is not correlated with $q$ and $a_2$. But $q_1$ and $q_2$ are correlated with each other. Then, indicator $q_2$ may be used as instrument in estimation of the last equation.

**Example 6.** *(Wooldridge) IQ and KKW (Knowledge of the Working World test score) as Indicators of Ability. IQ*$= \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1$*abil*$+a_1$, *KKW*$= \boldsymbol{\rho}_0 + \boldsymbol{\rho}_1$*abil*$+a_2$.

*We add IQ to equation and use KKW as instrument*

$$\begin{aligned}\log(\textit{wage}) \quad = \quad & \underset{(0.33)}{4.59} + \underset{(.003)}{.014}\,\textit{exper} + \underset{(.003)}{.010}\,\textit{tenure} + \underset{(.042)}{.201}\,\textit{married} \\ & + \underset{(.031)}{.051}\,\textit{south} + \underset{(.028)}{.177}\,\textit{urban} + \underset{(.074)}{.023}\,\textit{black} + \underset{(.017)}{.025}\,\textit{educ} + \underset{(.005)}{.013}\,\textit{IQ}\end{aligned}$$

*return to education only* $2.5\%$ *and not significant.*