Econometrics: Multiple Regression and Applications

ECON4004: LAB 1

Duong Trinh

University of Glasgow

February 7, 2024

Intro

- Duong Trinh
  - PhD Student in Economics (Bayesian Microeconometrics)
  - Email: Duong.Trinh@glasgow.ac.uk

- ECON4004-LB01
  - Wednesday 10am -12 pm
  - 5 sessions (7-Feb, 14-Feb, 21-Feb, 28-Feb, 6-March)
  - ST ANDREWS:357
- ECON4004-LB02
  - Wednesday 12-2 pm
  - 5 sessions (7-Feb, 14-Feb, 21-Feb, 28-Feb, 6-March)
  - ST ANDREWS:357

Record Attendance

Picture the Scenario

- **Objective:** Investigate the effect of fertility on women labor supply behaviour (*with a focus on Instrumental Variable approach*).

- **Dataset:** `fertility.dta`
  - from 1980 U.S. Census.
  - contains information on $254,654$ married women aged 21–35 with two or more children.

- **Key variables:**
  - `weeksm1`: weeks worked (labor supply)
  - `morekids`: the indicator variable denoting having more than 2 children (fertility)
  - `samesex`: equals to 1 if the first two children are of the same sex (boy–boy or girl–girl) and equal to 0 otherwise.

Questions (S&W Exercise E12.1)

Linear regression

(»review)

(a) Regress `weeksm1` on `morekids` using OLS. On average, do women with more than two children work less than women with two children? How much less?

(b) Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (`morekids`) on labor supply (`weeksm1`).

Questions (S&W Exercise E12.1)

IV regression with a single regressor and a single instrument

(»review)

(c) Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?

(d) Explain why `samesex` is a valid instrument for the IV regression of weeks worked on `morekids`.

(e) Is `samesex` a weak instrument?

(f) Estimate the IV regression of weeks worked on `morekids`, using `samesex` as an instrument. How large is the fertility effect on labor supply? How can we test whether `morekids` is endogenous?

Questions (S&W Exercise E12.1)

IV regression with additional control variables

(»review)

(g) Include the variables `agem1`, `black`, `hispan`, and `othrace` in the labor supply regression (treating these variables as exogenous).

- ◇ Do the results change? Explain why or why not.
- ◇ Does the instrumental variable remain relevant? Why?
- ◇ Does the test of endogeneity give different results than in (f)? What can we conclude about the endogeneity of `morekids`?

(a) Regress `weeksm1` on `morekids` using OLS.

- Linear regression model

$$weeksm1_i = \beta_0 + \beta_1 \cdot morekids_i + u_i$$

- OLS estimation results (»stata)

$$\underset{(se)}{\widehat{weeksm1}} = \underset{(0.0561)}{21.0684} - \underset{(0.0871)}{5.3867} \cdot morekids \qquad R^2 = 0.0143$$

- The slope estimate $\hat{\beta}_1^{OLS} \approx -5.39$ indicates that women with more than two children work 5.39 fewer weeks per year than women with two or fewer children *on average*.

(b) Explain why this result is inappropriate for estimating the causal effect of fertility on labor supply.

- Both fertility and labor supply are choice variables which are endogenously determined. Women's age, education, wage, partner's income, or unobservable characteristics related to tastes for children and working might affect both desired fertility and employment decisions simultaneously.

- Ignoring these factors and using the simple linear regression only may distort (overestimate or underestimate) the true causal effect. (»review)

(c) Are couples whose first two children are of the same sex more likely to have a third child?

- Linear regression model

$$morekids_i = \delta_0 + \delta_1 \cdot samesex_i + v_i$$

- Estimation results (»stata)

$$\underset{(se)}{\widehat{morekids}} = \underset{(0.0013)}{0.3464} + \underset{(0.0019)}{0.0675} \cdot samesex \qquad R^2 = 0.0048$$

- $\hat{\delta}_1 \approx 0.0675$ suggests that couples with $\texttt{samesex} = 1$ are 6.75% more likely to have an additional child than couples with $\texttt{samesex} = 0$ *on average*.
- The effect is highly significant (t-statistic $= 35.2$).

(d) Explain why `samesex` is a valid instrument for the IV regression of `weeksm1` on `morekids`.

$$weeksm1_i = \beta_0 + \beta_1 \cdot morekids_i + u_i$$

- Two conditions for a valid instrument:
  1. *Relevant*? $\text{corr}(samesex_i, morekids_i) \neq 0$?
     Plausibly: The effect of `samesex` on `morekids` is statistically significant, as discussed in (c). The first stage $F - statistic = 1238.17$ is large.
  2. *Exogenous*? $\text{corr}(samesex_i, u_i) = 0$?
     Plausibly: `samesex` is random and is unrelated to any of the other variables in the model including the error term in the labor supply equation.

$\Rightarrow$ Together, these imply that `samesex` is a valid instrument.

(e) Is `samesex` a weak instrument?

- This is related to the first condition - *Instrument Relevance* in (d).
- First-stage regression

$$morekids_i = \delta_0 + \delta_1 \cdot samesex_i + v_i$$

- The instrument is *weak* if δ_1 is either zero or nearly zero, i.e. it explains very little of the variation in `morekids`. From (c), this is not the case of `samesex`. (»stata)

(f) Estimate the IV regression of `weeksm1` on `morekids`, using `samesex` as an instrument.

TSLS has two stages - two regressions:

1. Regress $morekids_i$ on $samesex_i$ to isolate the part of `morekids` that is uncorrelated with u

$$morekids_i = \delta_0 + \delta_1 \cdot samesex_i + v_i$$

   Then, compute the predicted values $\widehat{morekids}_i = \hat{\delta}_0 + \hat{\delta}_1 \cdot samesex_i$ for $i = 1, \ldots, n$.

2. Regress $weeksm1_i$ on $\widehat{morekids}_i$

$$weeksm1_i = \beta_0 + \beta_1 \cdot \widehat{morekids}_i + u_i$$

   We eventually obtain $\hat{\beta}_1^{TSLS}$, which is the TSLS estimator.

[SN] Stata command for IV regression of Y on a single endogenous X instrumented by Z

OLS standard errors from the second stage regression are not correct as they do not take into account the estimation in the first stage when $\hat{X}$ is estimated. `ivregress` command in Stata adjusts for this 2 stage process.

```
* ivregress 2sls yvar (xvar = IV), r
// report result of intrinsic interest with 2SLS estimate
```

```
* ivregress 2sls yvar (xvar = IV), r first
// report additional result from first-stage regression
```

```
* ivregress 2sls yvar (xvar = IV), r
* estat firststage
// report first-stage regression statistics
```

```
* ivregress 2sls yvar (xvar = IV), r
* estat endog
// perform tests of the endogeneity of xvar
```

[SN] IV regression with Y : *weeksm1*, X : *morekids*, and intrument Z : *samesex*

```
* ivregress 2sls yvar (xvar = IV), r
```

```
. ivregress 2sls weeksm1 (morekids = samesex), r

Instrumental variables 2SLS regression          Number of obs   =    254,654
                                                Wald chi2(1)    =      24.53
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.0139
                                                Root MSE        =     21.715
```

weeksm1	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
morekids	-6.313685	1.274681	-4.95	0.000	-8.812013	-3.815357
_cons	21.42109	.4872487	43.96	0.000	20.4661	22.37608

```
Endogenous: morekids
Exogenous:  samesex
```

[SN] Another way to check Instrument Relevance

```
* ivregress 2sls yvar (xvar = IV), r
* estat firststage
```

```
. quiet ivregress 2sls weeksm1 (morekids = samesex), r

. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,254652)	Prob > F
morekids	**0.0048**	**0.0048**	**0.0048**	**1238.17**	**0.0000**

⇒ The instrument `samesex` remains relevant!

(»compare)

(f) How large is the fertility effect on labor supply?

- Estimation result

$$\hat{\beta}_1^{TSLS} \approx -6.31, \quad \text{robust s.e} \approx 1.27$$

  suggests that that women with more than two children work 6.31 fewer weeks per year than women with two or fewer children *on average*.

- Homogeneous effects: If the causal effect is the same for every individual, the result above has *causal interpretation*.

- Heterogeneous effects: *we will discuss later*.

(f) Test endogeneity of regressor X : *morekids*

```
* ivregress 2sls yvar (xvar = IV), r
* estat endog
```

```
. quiet ivregress 2sls weeksm1 (morekids = samesex), r

. estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Robust score chi2(1)            =    .53116  (p = 0.4661)
  Robust regression F(1,254651)   =   .531155  (p = 0.4661)
```

$\Rightarrow$ Result of robustified DWH test suggests variable `morekids` is unlikely to be endogenous, as the $p - values$ are well above 0.1.

(g) Include (exogenous) variables `agem1`, `black`, `hispan`, and `othrace` in the labor supply regression.

1. Regress $morekids_i$ on $samesex_i$ and all exogenous variables

$$morekids_i = \delta_0 + \delta_1 \cdot samesex_i + \delta_2 \cdot agem1_i + \delta_3 \cdot black_i + \\ + \delta_4 \cdot hispan_i + \delta_5 \cdot othrace_i + v_i$$

   Then, compute the predicted values $\widehat{morekids}_i$ for $i = 1, \ldots, n$.

2. Regress $weeksm1_i$ on $\widehat{morekids}_i$

$$weeksm1_i = \beta_0 + \beta_1 \cdot \widehat{morekids}_i + \beta_2 \cdot agem1_i + \beta_3 \cdot black_i + \\ + \beta_4 \cdot hispan_i + \beta_5 \cdot othrace_i + u_i$$

   We eventually obtain $\hat{\beta}_1^{TSLS}$, which is the TSLS estimator.

[SN] Stata command for IV regression of Y on a single endogenous X instrumented by Z, and several exogenous W

Implement `ivregress` command with additional exogenous variables

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
// report result of intrinsic interest with 2SLS estimate
```

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r first
// report additional result from first-stage regression
```

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
* estat firststage
// report first-stage regression statistics
```

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
* estat endog
// perform tests of the endogeneity of xvar
```

(g) IV regression with Y : *weeksm*1, X : *morekids*, Z : *samesex*, and additional exogenous regressors

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
```

```
. ivregress 2sls weeksm1 agem1 black hispan othrace (morekids = samesex), r

Instrumental variables 2SLS regression          Number of obs   =    254,654
                                                Wald chi2(5)    =    6954.98
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.0437
                                                Root MSE        =     21.384

------------------------------------------------------------------------------
             |               Robust
    weeksm1  | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   morekids  |  -5.821051   1.246386    -4.67   0.000    -8.263923   -3.378179
      agem1  |   .8315975   .0226406    36.73   0.000     .7872228    .8759722
      black  |   11.62327   .2317953    50.14   0.000     11.16896    12.07758
     hispan  |   .4041802   .2607962     1.55   0.121     -.106971    .9153314
    othrace  |   2.130962   .2109857    10.10   0.000     1.717438    2.544486
      _cons  |  -4.791894   .3897868   -12.29   0.000    -5.555862   -4.027925
------------------------------------------------------------------------------
Endogenous: morekids
Exogenous:  agem1 black hispan othrace samesex
```

(g) How large is the fertility effect on labor supply?

- Estimation result

$$\hat{\beta}_1^{TSLS} \approx -5.82, \quad \text{robust s.e} \approx 1.25$$

- The results do not change in an important way. The reason is that `samesex` is unrelated to `agem1`, `black`, `hispan`, `othrace`. Thus its covariance with these variables is zero, and thus `samesex` is likely to be uncorrelated with the error term, even when the latter includes those variables.

(g) Check Instrument Relevance

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
* estat firststage
```

```
. quiet ivregress 2sls weeksm1 agem1 black hispan othrace (morekids = samesex), r

. estat firststage

  First-stage regression summary statistics
```

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,254648)	Prob > F
morekids	**0.0242**	**0.0242**	**0.0050**	**1280.94**	**0.0000**

$\Rightarrow$ The instrument `samesex` remains relevant. The reason is that the addition of the new variables does not affect the strength of the relationship between `morekids` and `samesex`.

(g) Test endogeneity of regressor X : *morekids*

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
* estat endog
```

```
. quiet ivregress 2sls weeksm1 agem1 black hispan othrace (morekids = samesex), r

. estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Robust score chi2(1)            =   .108388  (p = 0.7420)
  Robust regression F(1,254647)   =   .108385  (p = 0.7420)
```

$\Rightarrow$ The endogeneity test gives the same result as without the additional regressors. However, we should still be very skeptical about the possibility of `morekids` being exogenous. (»review)

Table of Results

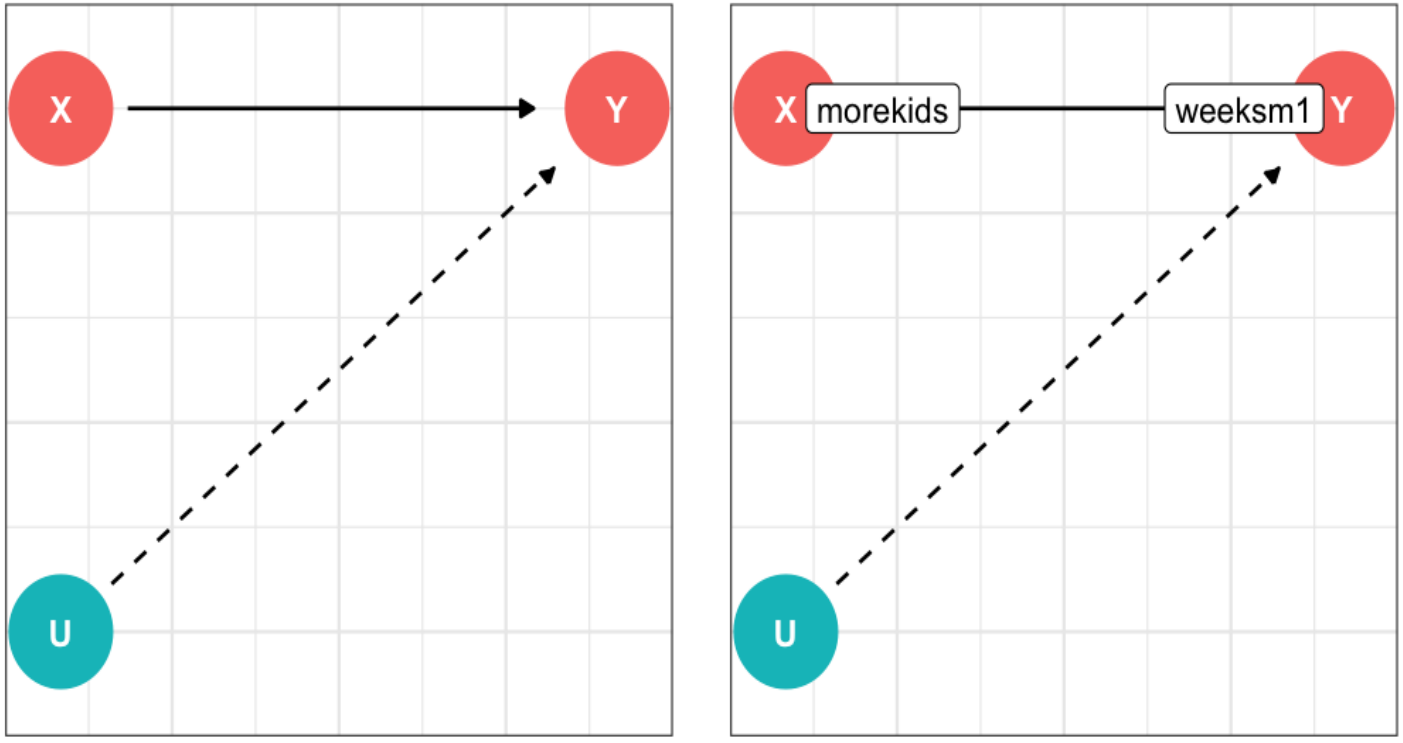
Regressor	**Estimation method**		
	OLS	**TSLS**	**TSLS**
morekids	-5.39 (0.09) [-5.56, -5.22]	-6.31 (1.27) [-8.81, -3.81]	-5.82 (1.25) [-8.26, -3.38]
Additional regressors	*Intercept*	*Intercept*	*Intercept, agem1, black, hispan, othrace*
First Stage F-Statistic		1238.2	1280.9

Notes: Standard errors shown in parentheses and 95% confidence intervals are shown in brackets.

BRIEF REVIEW

Causal Graph (I)

Linear regression

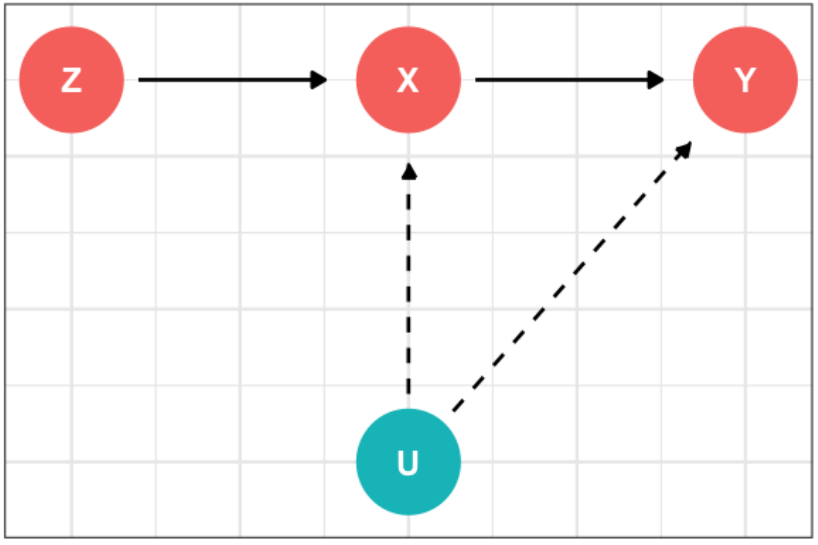


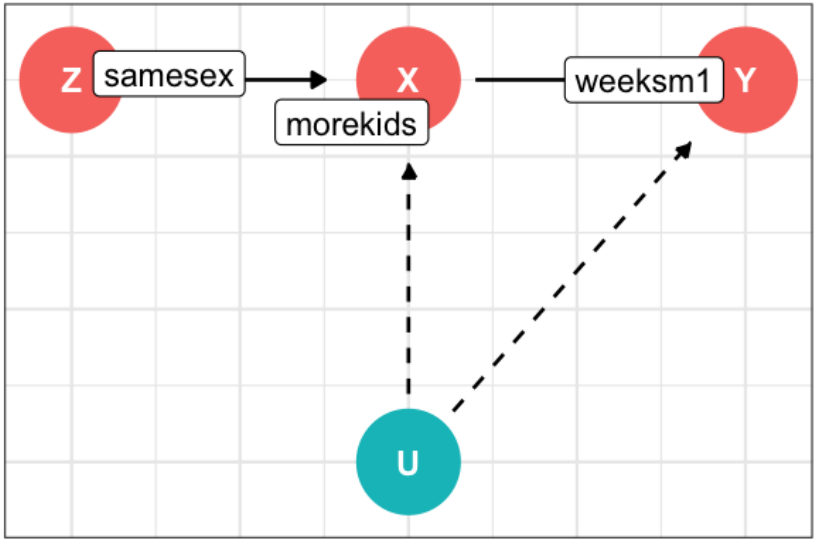
$$E(u_i \mid X_i) = 0$$

(»back)

Causal Graph (II)

IV regression with a single regressor and a single instrument



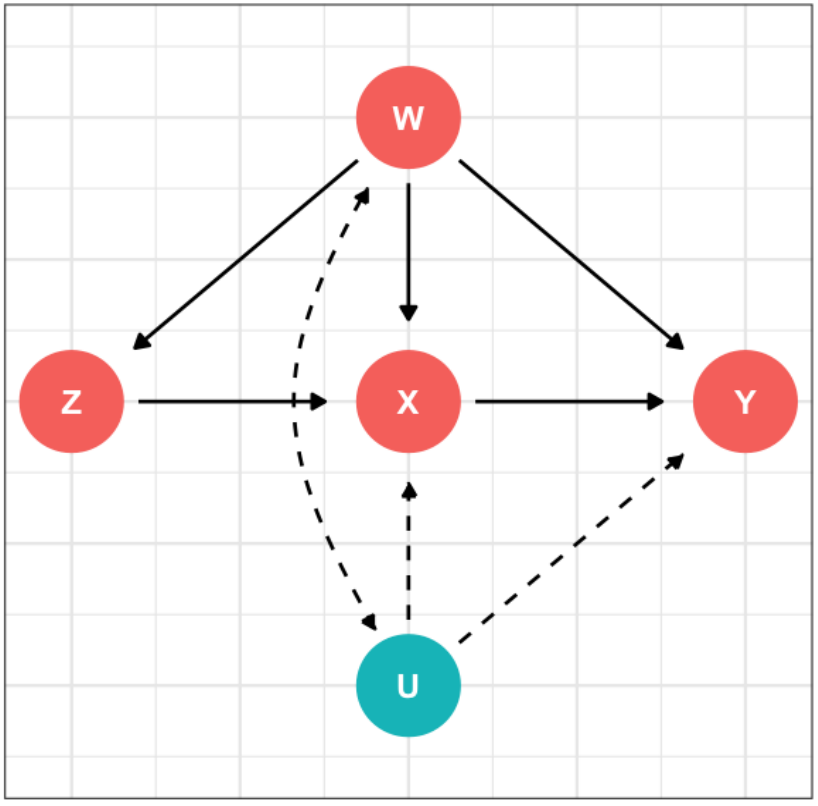


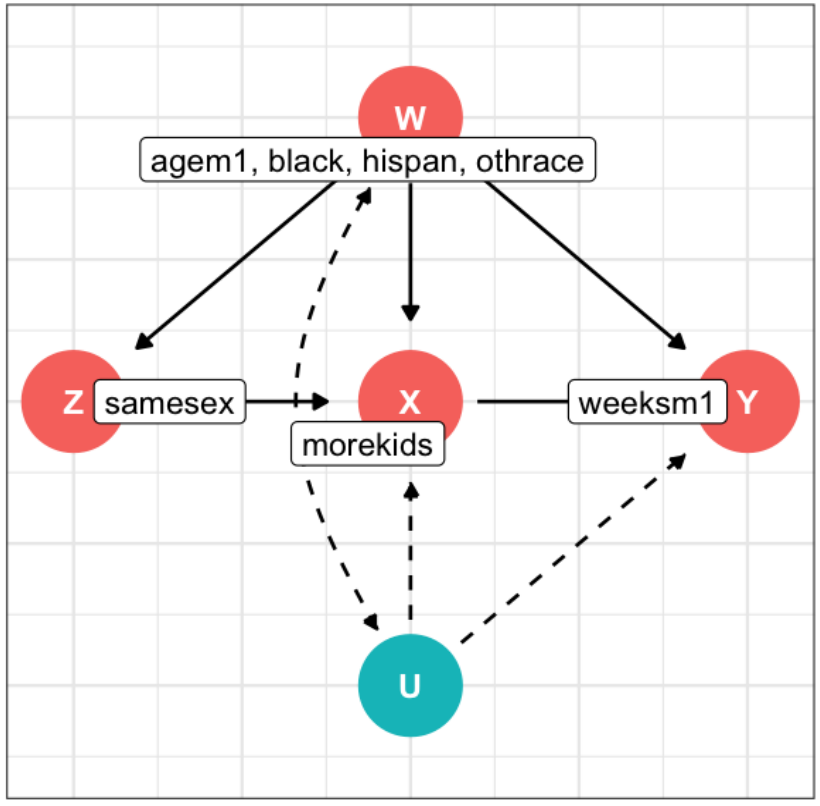
$$\underset{\textit{relevance}}{\text{corr}(Z_i, X_i) \neq 0} \text{ and } \underset{\textit{exogeneity}}{\text{corr}(Z_i, u_i) = 0}$$

(»back)

Causal Graph (III)

IV regression with additional control variables





$$E(u_i \mid Z_i, W_i) = E(u_i \mid W_i) = 0$$

(»back)

Omitted Variable Bias

- If there is another factor F that is a determinant of Y and correlated with X which makes X and Y be associated, ignoring F will cause omitted variable bias

$$plim\hat{\beta}_1 = \beta_1 + \beta_2 \frac{Cov(X,F)}{Var(X)}$$

- $\hat{\beta}_1$ is biased upward $\Leftrightarrow plim\hat{\beta}_1 > \beta_1$
- $\hat{\beta}_1$ is biased downward $\Leftrightarrow plim\hat{\beta}_1 < \beta_1$

	$Cov(X,F) > 0$	$Cov(X,F) < 0$
$\beta_2 > 0$	$plim\hat{\beta}_1 > \beta_1$	$plim\hat{\beta}_1 < \beta_1$
$\beta_2 < 0$	$plim\hat{\beta}_1 < \beta_1$	$plim\hat{\beta}_1 > \beta_1$

(»backB)

Testing for regressor endogeneity (I)

1. Hausman test

- If there is little difference between OLS and IV estimators, then there is no need to instrument, and we conclude that the regressor was exogenous.
- If instead there is considerable difference, then we needed to instrument and the regressor is endogenous.
- Idea: Compares just the coefficients of the endogenous variables, with the use of the Hausman test statistic

$$T_H = \frac{(\hat{\beta}_{IV} - \hat{\beta}_{OLS})^2}{\hat{V}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})} \sim \chi^2(1)$$

- It relies on a very strong assumption that model errors are independent and homoskedastic.

Testing for regressor endogeneity (II)

2. Durbin-Wu-Hausman (DWH) test

◇ Idea: Use augmented regressors to produce a robust test statistic. Specifically, rewrite the structural equation with an additional variable

$$Y_i = \beta X_i + \mathbf{W}_i\gamma + \rho v_i + u_i$$

Under Null hypothesis that D_i is exogenous, $E(v_i u_i \mid D_i \mathbf{X}_i) = 0$.

◇ Null hypothesis becomes:

$$H_0 : \rho = 0$$

◇ Valid even in the case of heteroskedastic errors provided that we use robust variance estimates.

(»backG)

STATA CODES & RESULTS

```
* regress yvar xvar, r
// add option r to report Hetroskedasticity-robust standard errors
```

```
. reg weeksm1 morekids, r

Linear regression                               Number of obs     =    254,654
                                                F(1, 254652)      =    3820.91
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0143
                                                Root MSE          =      21.71
```

weeksm1	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
morekids	-5.386996	.0871491	-61.81	0.000	-5.557806	-5.216186
_cons	21.06843	.0560681	375.76	0.000	20.95854	21.17832

(»backA)

```
* regress xvar IV, r
// add option r to report Hetroskedasticity-robust standard errors
```

```
. reg morekids samesex, r

Linear regression                               Number of obs     =    254,654
                                                F(1, 254652)      =    1238.17
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0048
                                                Root MSE          =     .48435

------------------------------------------------------------------------------
             |               Robust
    morekids | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    samesex  |   .0675253   .001919    35.19   0.000     .0637641    .0712865
       _cons |   .3464248   .001341   258.34   0.000     .3437965    .3490531
------------------------------------------------------------------------------
```

(»backC) (»backF)

```
* ivregress 2sls yvar (xvar = IV), r
// report result of intrinsic interest with 2SLS estimate
```

```
. ivregress 2sls weeksm1 (morekids = samesex), r

Instrumental variables 2SLS regression          Number of obs   =    254,654
                                                Wald chi2(1)    =      24.53
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.0139
                                                Root MSE        =     21.715

------------------------------------------------------------------------------
             |               Robust
    weeksm1 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   morekids |  -6.313685   1.274681    -4.95   0.000    -8.812013   -3.815357
      _cons |   21.42109   .4872487    43.96   0.000      20.4661    22.37608
------------------------------------------------------------------------------
Endogenous: morekids
Exogenous:  samesex
```

```
* ivregress 2sls yvar (xvar = IV), r
* estat firststage
// report first-stage regression statistics
```

```
. quiet ivregress 2sls weeksm1 (morekids = samesex), r

. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,254652)	Prob > F
morekids	**0.0048**	**0.0048**	**0.0048**	**1238.17**	**0.0000**

```
* ivregress 2sls yvar (xvar = IV), r
* estat endog
// perform tests of the endogeneity of xvar
```

```
. quiet ivregress 2sls weeksm1 (morekids = samesex), r

. estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Robust score chi2(1)              =    .53116  (p = 0.4661)
  Robust regression F(1,254651)     =   .531155  (p = 0.4661)
```

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
// report result of intrinsic interest with 2SLS estimate
```

```
. ivregress 2sls weeksm1 agem1 black hispan othrace (morekids = samesex), r

Instrumental variables 2SLS regression          Number of obs   =    254,654
                                                Wald chi2(5)    =    6954.98
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.0437
                                                Root MSE        =     21.384

------------------------------------------------------------------------------
             |               Robust
    weeksm1 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   morekids |  -5.821051   1.246386    -4.67   0.000    -8.263923   -3.378179
      agem1 |   .8315975   .0226406    36.73   0.000     .7872228    .8759722
      black |   11.62327   .2317953    50.14   0.000     11.16896    12.07758
     hispan |   .4041802   .2607962     1.55   0.121    -.106971    .9153314
    othrace |   2.130962   .2109857    10.10   0.000     1.717438    2.544486
      _cons |  -4.791894   .3897868   -12.29   0.000    -5.555862   -4.027925
------------------------------------------------------------------------------
Endogenous: morekids
Exogenous:  agem1 black hispan othrace samesex
```

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
* estat firststage
// report first-stage regression statistics
```

```
. quiet ivregress 2sls weeksm1 agem1 black hispan othrace (morekids = samesex), r

. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,254648)	Prob > F
morekids	**0.0242**	**0.0242**	**0.0050**	**1280.94**	**0.0000**

```
* ivregress 2sls yvar wvar1 wvar2 wvark (xvar = IV), r
* estat endog
// perform tests of the endogeneity of xvar
```

```
. quiet ivregress 2sls weeksm1 agem1 black hispan othrace (morekids = samesex), r

. estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Robust score chi2(1)            =   .108388  (p = 0.7420)
  Robust regression F(1,254647)   =   .108385  (p = 0.7420)
```