University of Zurich[UZH]

SocialNetworks
UNIVERSITY RESEARCH PRIORITY PROGRAM

R PACKAGE TO ADDRESS ENDOGENEITY WITHOUT EXTERNAL INSTRUMENTAL VARIABLES

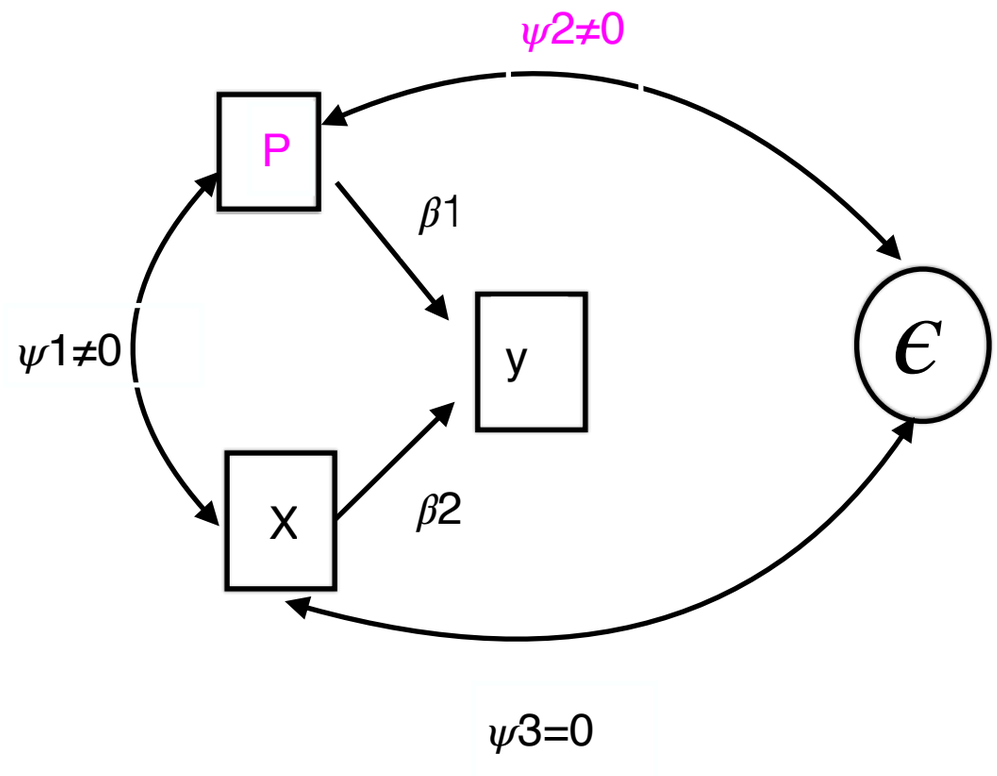
Raluca Gui, Markus Meierer, Patrik Schilter, René Algesheimer

useR2019, Toulouse, 9-12th July

RENDO

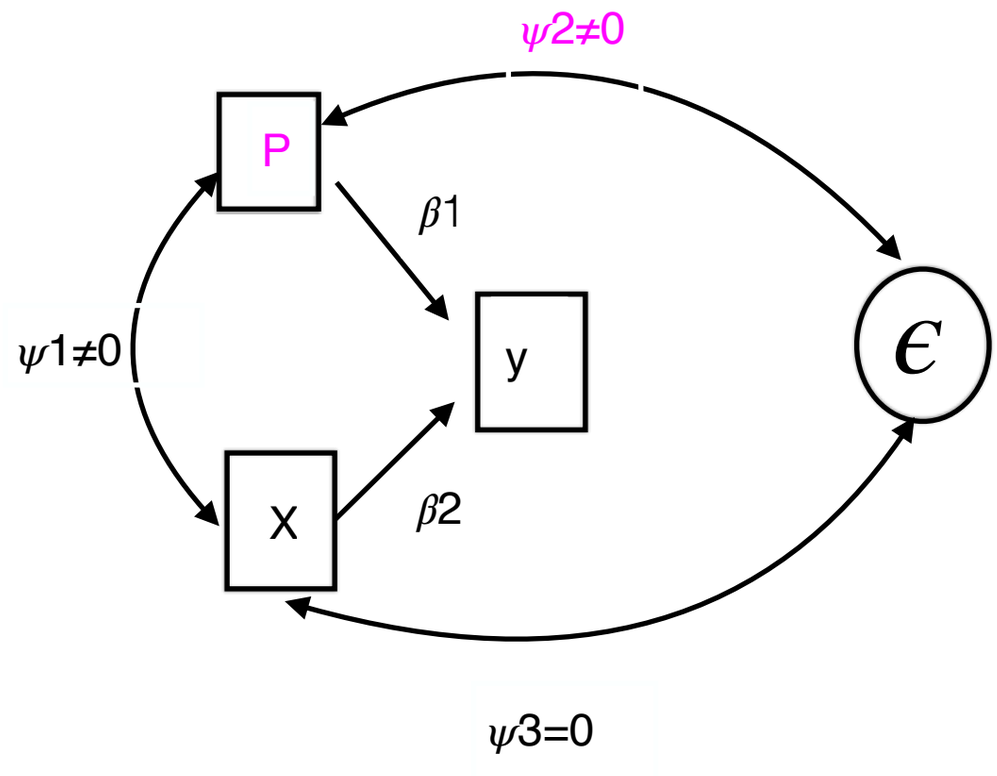
WHAT IS ENDOGENEITY

Fancy word for a **simple** problem



WHY CARE ABOUT ENDOGENEITY ?

Renders **biased** and **inefficient** estimates



FORMS OF ENDOGENEITY

Omitted variable

- Ability
- Income
- Teacher dedication

Measurement error

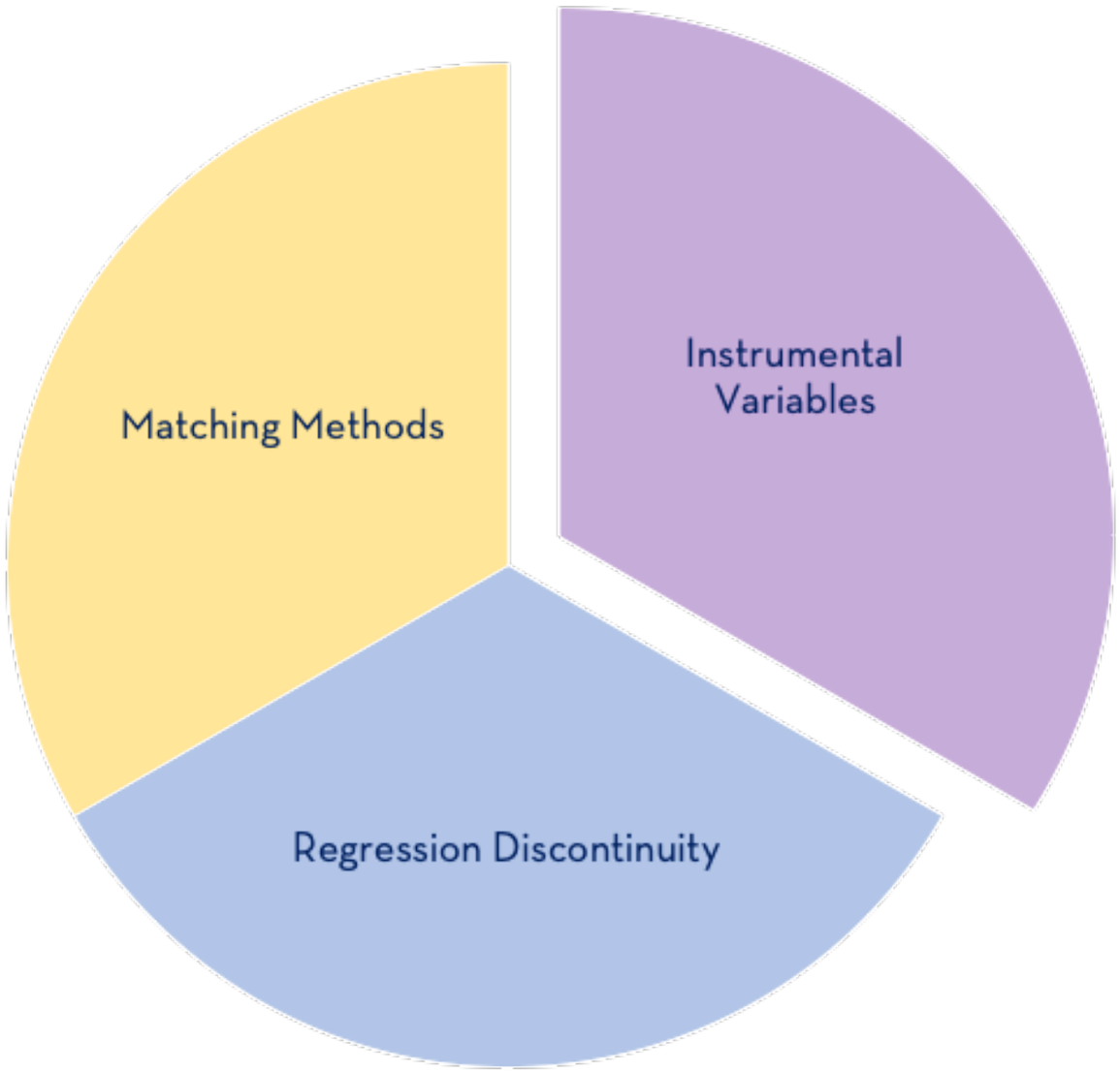
- Income
- Drug consumption

Simultaneity

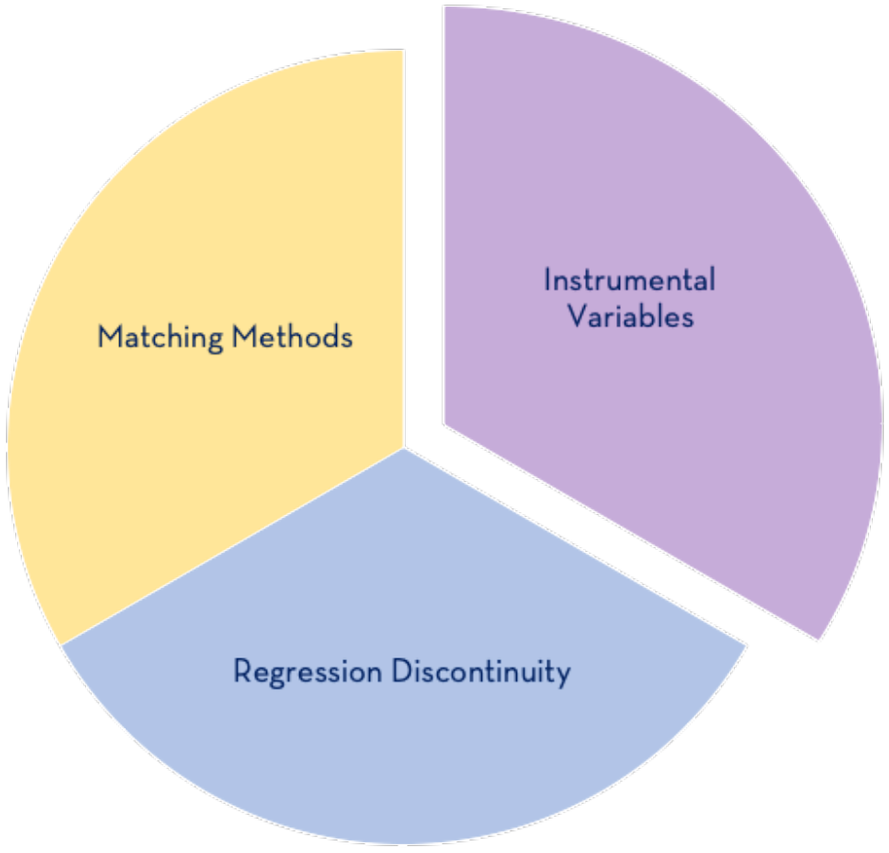
- Supply and demand
- Marketing expenditure and sales

CURRENT SOLUTIONS TO ENDOGENEITY

with secondary data



RENDO



INTERNAL INSTRUMENTAL VARIABLE METHODS

- **Latent instrumental variables** (Ebbes et al., 2005)
- **Gaussian copula correction** (Park and Gupta, 2012)
- **Heteroskedastic errors** (Lewbel 2012)
- **Higher moments** (Lewbel 1997)
- **Multilevel GMM** (Kim and Frees, 2007)

INTERNAL INSTRUMENTAL VARIABLES

Consider the model:

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 X_t + \epsilon_t$$

where $cor(P_t, \epsilon_t) \neq 0$ => endogeneity

$$P_t = \gamma Z_t + \nu_t$$

Vector of **internal instrumental variables** (observed or unobserved) => distributional assumptions needed for identification

LATENT INSTRUMENTAL VARIABLES

EBBES, WEDEL, BÖCKENHLOT, STEERNEMAN, 2005

ASSUMPTIONS

- Only **one** endogenous regressor, P_t
- Only **one** exogenous regressor.
- $P_t \neq N(\cdot, \cdot)$
- Z_t - **discrete** latent instrument, with at least 2 groups with different means
- $\epsilon_t \sim N(0, \sigma_\epsilon^2), \ corr(Z_t, \epsilon_t) = 0$
- Estimation - MLE

R CODE

```
latentIV(y ~ P, data,  start.params)
```

Initial parameter values = user provided or OLS estimates

R OUTPUT

```
# latent instrumental variable
latentIV(formula = y ~ P, data = dataLatentIV)

Coefficients:

             Estimate Std.Error z-score Pr(>|z|)

(Intercept)  2.9863    0.0303     98.53   <2e-16 ***
P           -0.9753    0.0384    -25.38   <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Further parameters estimated during model fitting:
pi1 pi2 theta5 theta6 theta7 theta8
0.01744 2.21266 0.74306 1.01343 0.26144 1.09164
(see help file for details)

Initial parameter values:
(Intercept)=2.9152 P=-0.8519 pi1=0.589 pi2=2.0011
theta5=0.5 theta6=1 theta7=0.5 theta8=1

The value of the log-likelihood function: 7873.344
AIC: -15730.69 , BIC: -15684.1
KKT1: FALSE KKT2: TRUE Optimx Convergence Code: 0
```

COPULA CORRECTION METHOD(1/3)

PARK & GUPTA, 2012

ASSUMPTIONS

- Both **continuous** and **discrete** endogenous regressors allowed.
- $P_t \neq N(\cdot, \cdot)$ - if continuous
- $P_t \neq$ Bernoulli - if discrete
- $\epsilon_t \sim N(0, \sigma_\epsilon^2)$
- Estimation - MLE

R CODE

```
# 1 continuous endogenous regressor, P
copulaCorrection(y ~ P + X1 + X2 | continuous (P),
data, start.params, num.boots)

# discrete endogenous regressors, P1, P2
copulaCorrection(y ~ P1 + P2 + X1 + X2 |
discrete(P1) + discrete(P2), data)

# 1 discrete,P1, 1 continuous (P2) end. regressors
copulaCorrection(y ~ P1 + P2 + X1 + X2 |
discrete(P1) + continuous(P2), data)
```

R OUTPUT

```
# 1 continuous endogenous regressor
copulaCorrection(formula = y ~ X1 + X2 + P | continuous(P), data = dataCopCont,num.boots = 50)

Coefficients:

            Point      Boots     Lower Boots  Upper Boots
            Estimate   SE        CI (95%)     CI (95%)
(Intercept)  2.0625    0.0475     1.9591       2.1105
X1           1.4887    0.0091     1.4763       1.5074
X2          -3.0065    0.0109    -3.0227      -2.9849
P           -0.9652    0.0175    -0.9879      -0.9220

Number of bootstraps: 50
Continuous endogenous variables: P
Further parameters estimated during model fitting:
rho    sigma
0.3715 1.3621
(see help file for details)
Initial parameter values:
(Intercept)=2.0347 X1=1.492 X2=-3.0008 P=-0.8209 rho=0 sigma=0
The value of the log-likelihood function: 5473.473
AIC: -10934.95 , BIC: -10900
```

R OUTPUT

```
# 1 discrete + 1 continuous endogenous regressors
copulaCorrection(formula = y ~ X1 + X2 + P | discrete(P1) + continuous(P2), data = dataCopDisCont)

Coefficients:
```

	Point Estimate	Boots SE	Lower Boots CI (95%)	Upper Boots CI (95%)
(Intercept)	1.8467	0.1524	1.3994	1.9938
X1	1.5018	0.0087	1.4857	1.5189
X2	-2.9979	0.0114	-3.0200	-2.9743
P1	-0.9608	0.0490	-1.0090	-0.8129
P2	0.7673	0.0283	0.7109	0.8218
PStar.P2	**0.1358**	**0.0490**	**0.0432**	**0.2282**
PStar.P1	**0.2376**	**0.0850**	**-0.0155**	**0.3189**

HIGHER MOMENTS METHOD

LEWBEL, 1997

ASSUMPTIONS

- Only **one** endogenous regressor allowed.
- Z_t - **skewed** distribution
- $E(\epsilon_t) = 0$
- 3rd moment of the data exists.
- Estimation - 2SLS

R CODE

```
higherMomentsIV(y ~ P + X1 + X2 | P | IIV(iiv, g)|
W, data)
```

P = the endogenous regressor

IIV(iiv, g) - iiv = internal instrument to be computed;

g = transformation to the exogenous reg.

W - additional exogenous regressors

```
higherMomentsIV(y ~ P + X1 + X2 | P | IIV(iiv =
gp, g = x2, X1, X2), data)
```

```
higherMomentsIV(y ~ P + X1 + X2 | P | IIV(iiv =
gp, g = x2, X1, X2) + IIV(iiv = yp), data)
```

R OUTPUT

```
# higher moments method
higherMomentsIV(formula = y ~ X1 + X2 + P | P | IIV(iiv = yp), data = dataHigherMoments)

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.61941   0.70937    2.283   0.0225 *
X1           1.58613  0.36092    4.395   1.16e-05 ***
X2           3.06182  0.07827   39.118   < 2e-16 ***
P           -1.01376  0.08187  -12.383   < 2e-16 ***

Diagnostic tests:
                df1 df2  statistic p-value
Weak instruments 1  2496 13.838    0.000204 ***
Wu-Hausman       1  2495 5.271     0.021763 *
Sargan 0 NA NA NA
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .
Residual standard error: 1.77 on 2496 degrees of freedom
Multiple R-Squared: 0.8925, Adjusted R-squared: 0.8924
Wald test: 605 on 3 and 2496 DF, p-value: < 2.2e-16
```

HETEROSKEDASTIC ERRORS METHOD

LEWBEL, 2012

ASSUMPTIONS

- Multiple endogenous regressors allowed.
- $cov(Z, \nu^2) \neq 0$ - **testable**.
- $cov(Z, \nu\epsilon) = 0$
- 2SLS estimation using as instruments X and $(Z - E(Z))\nu$

R CODE

```
hetErrorsIV(y ~ P + X1 + X2 + X3 | P | IIV(X1,X2) |
W, data)
```

P = the endogenous regressor

IIV() - specifies which exogenous variables should be used for the construction of the internal instruments

W - additional exogenous regressors

R OUTPUT

```
# heteroskedastic errors
hetErrorsIV(formula = y ~ X1 + X2 + P | P | IIV(X1,X2), data = dataHetIV)

Coefficients:

            Estimate Std.Error t value Pr(>|t|)

(Intercept) 2.0348    0.4223   4.818    1.54e-06 ***
X1           1.5814    0.3272    4.833    1.43e-06 ***
X2           2.9252    0.2122   13.782    < 2e-16 ***
P           -0.9362    1.8809   -0.498    0.619

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments 2   2495 0.726    0.484
Wu-Hausman       1   2495 0.080    0.778
Sargan           1   NA   0.172    0.678

Residual standard error: 2.294 on 2496 degrees of freedom
Multiple R-Squared: 0.4359, Adjusted R-squared: 0.4353
Wald test: 657.9 on 3 and 2496 DF, p-value: < 2.2e-16
```

Large std. errors - to be expected since identification is based on higher moments of the data.

MULTILEVEL GMM METHOD

KIM & FREES, 2007

ASSUMPTIONS

- Continuous dependent variable.
- Multiple endogenous regressors.
- Multilevel model with at most 3 levels.
- **No level 1 endogeneity.**
- GMM estimation

R CODE

```
multilevelIV(y ~ X11 +… + X15 + X21 +..+ X24 + X31
+ .. + X33 + (1|CID) + (1|SID)| endo(X15), data)
```

X15 assumed correlated with level 2 or level 3 model error, **not** with level 1 error

WHY IS RENDO RELEVANT ?

with REndo we can ...

- address endogeneity without external instrumental variables.
- address endogeneity both, in single and multilevel models.
- increase the efficiency of the estimates by creating internal instruments and include them in the IV regression.
- address endogeneity where the endogenous regressors can be continuous or discrete.
- the only package on CRAN that implements internal instrumental variables methods.

University of Zurich[UZH]

SocialNetworks
UNIVERSITY RESEARCH PRIORITY PROGRAM

THANK YOU

Raluca Gui, raluca.gui@business.uzh.ch

REFERENCES

- Kim S, Frees F (2007). “Multilevel Modeling with Correlated Effects.” Pshychometrika, 72(4), 505–533.

- Ebbes P, Wedel M, Boeckenholt U, Steerneman A (2005). “Solving and Testing for Regressor-Error (In)Dependence When no Instrumental Variables Are Available: With Evidence for the Effect of Education on Income.” Quantitative Marketing and Economics, 3(4), 365–392.

- Lewbel A (1997). “Constructing Instruments for Regressions with Measurement Error When No Additional Data are Available, With an Application to Patents and R and D.” Econometrica, 65(5), 1201–1213.

- Lewbel A (2012). “Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models”, Journal of Business and Economic Statistics, 30(1), 67–80.

- Park S, Gupta S (2012). “Handling Endogeneous Regressors by Joint Estimation Using Copulas.”Marketing Science, 31(4), 567–586.

R OUTPUT

```
# multilevelIV - 3 levels model
res <- multilevelIV(y ~ X11 + X12 + X13 + X14 + X15 + X21 + X22 + X23 + X24 + X31 + X32  + X33 + (1| CID) +
(1| SID) | endo(X15), data = dataMultilevelIV)

coef(res)
```

	REF	FE_L2	FE_L3	GMM_L2	GMM_L3
(Intercept)	64.3168	0.0000	0.0000	64.3485	64.3168
X11	3.0213	3.0459	3.0214	3.0146	3.0213
X12	8.9522	8.9839	8.9524	8.9747	8.9522
X13	-2.0194	-2.0145	-2.0193	-2.0021	-2.0194
X14	1.9651	1.9791	1.9648	1.9658	1.9651
X15	**-0.5647**	**-0.9777**	**-0.5647**	**-0.9750**	**-0.5648**
X21	-2.3316	0.0000	-2.2845	-2.3052	-2.3316
X22	-3.9564	0.0000	-3.9553	-4.0130	-3.9564
X23	-2.9779	0.0000	-2.9756	-2.9488	-2.9779
X24	4.9078	0.0000	4.9084	4.7933	4.9078
X31	2.1142	0.0000	0.0000	2.1164	2.1142
X32	0.3934	0.0000	0.0000	0.3799	0.3934
X33	0.1082	0.0000	0.0000	0.1108	0.1082

true coefficient value for **X15 = -1**

X15 simulated to be correlated with level 2 error.

How to choose which model is correct? - Look at the **omitted variable test**, returned by **summary()**

R OUTPUT

```
# multilevelIV - test for endogeneity at level 2
OR at level 3
summary(res, "REF")

Coefficients for model REF:

            Estimate Std.Error  z-score  Pr(>|z|)
(Intercept) 64.3168  7.8734     8.169    3.11e-16
***
X11          3.0213  0.0257   117.306    < 2e-16 ***
X12          8.9522  0.0257   348.131    < 2e-16 ***
X13         -2.0194  0.0240   -83.835    < 2e-16 ***
X14          1.9651  0.0252    77.937    < 2e-16 ***
X15         -0.5647  0.0195   -28.962    < 2e-16 ***
X21         -2.3316  0.1622   -14.368    < 2e-16 ***
X22         -3.9564  0.1317   -30.160    < 2e-16 ***
X23         -2.9779  0.0661   -45.044    < 2e-16 ***
X24          4.9078  0.1979    24.792    < 2e-16 ***
X31          2.1142  0.1043    20.264    < 2e-16 ***
X32          0.3934  0.3042     1.293    0.1959
X33          0.1082  0.0523     2.067    0.0388 *
---
Signif. codes: 0*** 0.001** 0.01* 0.05. 0.1
```

Omitted variable tests for model REF:

```
                  df   Chisq     p-value
GMM_L2_vs_REF      7   18.74     0.009040 **
GMM_L3_vs_REF     13 -12872.98 1.000000
FE_L2_vs_REF      13   39.99     0.000139 ***
FE_L3_vs_REF      13   39.99     0.000138 ***
```

Omitted variable test btw. fixed effects estimator at level 2 (more robust when endogeneity) and the random effects (more efficient when no endogeneity) - > null hypothesis rejected => ∃ endogeneity at level 2 or 3

R OUTPUT

```
# multilevelIV - test for endogeneity at level 2
summary(res, "FE_L2")
```

```
Coefficients for model FE_L2:

            Estimate Std. Error z-score Pr(>|z|)
(Intercept) 0.0000   4.580e-19    0.00   1
X11         3.0460   2.978e-02  102.30 <2e-16 ***
X12         8.9840   3.360e-02  267.41 <2e-16 ***
X13        -2.0150   3.107e-02  -64.83 <2e-16 ***
X14         1.9790   3.203e-02   61.80 <2e-16 ***
X15        -0.9777   3.364e-02  -29.06 <2e-16 ***
X21         0        1.340e-18    0.00   1
X22         0        2.136e-18    0.00   1
X23         0        2.971e-18    0.00   1
X24         0        2.313e-18    0.00   1
X31         0        6.526e-18    0.00   1
X32         0        9.436e-18    0.00   1
X33         0        3.673e-17    0.00   1
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

Omitted variable tests for model FE_L2:

```
                    df Chisq p-value
FE_L2_vs_REF        13 39.99 0.000139 ***
FE_L2_vs_FE_L3       9 36.02 3.92e-05 ***
FE_L2_vs_GMM_L2     12 39.99 7.21e-05 ***
FE_L2_vs_GMM_L3     13 39.99 0.000139 ***
```

The **null hypothesis of no omitted level 2 effects is rejected** (p-value = 3.92e-05) => $\exists$ omitted variables at level 2 => **use fixed effects estimator at level 2 or the GMM at level 2** if coefficients for all variables are needed.