University of Zurich[UZH]

SocialNetworks
UNIVERSITY RESEARCH PRIORITY PROGRAM

REndo: An R Package To Address Endogeneity

Raluca Gui[1], Markus Meierer[1], Patrik Schilter[1], René Algesheimer[1]
[1] Department of Business Administration, URPP Social Networks, University of Zurich

1. Background

- The independence between the explanatory variables and the error term is a crucial assumption in regression models. The **endogeneity** problem arises when this condition is not satisfied.
- **Most common forms of endogeneity:**
  - omitted variables
  - measurement error
  - simultaneity.
- Instrumental variables (IV) estimation is a common solution to the endogeneity problem. But, **good IV are rare.**
- **Internal instrumental variable models (IIV)** are an **alternative** to the IV approach.
- IIV rely on the characteristics of the statistical distribution of the endogenous regressor. **No need for external instruments**.

2. Research question

- How to enable researchers and practitioners to do causal inference with observational data when instrumental variables are lacking?

Research Objective

Implement in **R** five of the newest internal instrumental variables methods.

3. Data and Methods

Consider the model

$$Y_t = \beta_0 + \beta_1 W_t + \alpha P_t + \epsilon_t$$

$$P_t = \pi' Z_t + \nu_t$$

- Y_t dependent variable
- W_t vector of exogenous variables
- P_t endogenous regressor
- Z_t vector of internal instruments $E(\epsilon_t) = 0 \quad E(\epsilon_t^2) = \sigma_\epsilon^2$
- ϵ_t structural error, $E(\nu_t) = 0, E(\nu_t^2) = \sigma_{\nu^2}^2, E(\epsilon_t \nu_t) = \sigma_{\epsilon\nu}$
- ν_t random error,

1. Latent Instrumental Variables

(Ebbes et al., 2005)

- Uses a latent variable model to account for regressor-error dependencies.
- The endogenous regressor is assumed to have two categories with distinct means.
- Allows only one endogenous regressor and no additional explanatory variables.

2. Joint estimation using copulas

(Park and Gupta, 2012)

- Unlike other IIV-methods (e.g. LIV), the copula method does not require the "exclusion restriction" assumption.
- No assumptions imposed on the relationship between Z_t and ϵ_t
- The correlation between the structural and the random error modelled using Gaussian copula.

3. Higher moments

(Lewbel, 1997)

- Build to address measurement error problem, but also other more general regressor-error dependency models.
- Internal instrumental variables can be constructed from the existing variables, using any function G=G(W) with finite third own and cross moments, non-linear in W, as below:
- $G_t - \overline{G_t}, (G_t - \overline{G_t})(P_t - \overline{P_t}), G_t - \overline{G_t})(Y_t - \overline{Y_t})$
- $(Y_t - \overline{Y_t})(P_t - \overline{P_t}), (P_t - \overline{P_t})^2, (Y_t - \overline{Y_t})^2$

4. Heteroscedastic errors

(Lewbel, 2012)

- Uses the heteroscedasticity of the errors in a linear projection of the endogenous regressor on the other covariates to solve the endogeneity produced by measurement error.
- Instruments are constructed as simple functions of the model's data: $[Z_t - E(Z_t)]\nu_t$, where Z_t are a subset of W_t, assumed exogenous. Simple two-stage least squares is used as estimation method.

5. Multilevel GMM

(Kim and Frees, 2007)

- The internal instruments are built exploiting the hierarchical structure of the data.
- Uses the between and within variation of the exogenous variables but only the within variation of the endogenous ones.
- Includes the random and fixed effects estimators as special cases and provides additional estimators between these two extremes.

Method	Model	Assumptions
Latent IV Ebbes et al. (2005)	▪ One endogenous variable; ▪ No additional regressors; ▪ ML estimation.	▪ $P_t \neq N(\cdot,\cdot)$ ▪ Z_t discrete, with at least 2 groups with different means. ▪ $\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$
Copula correction Park and Gupta (2012)	▪ Multiple endogenous variables; ▪ Additional regressors allowed; ▪ ML estimation.	▪ $P_t \neq N(\cdot,\cdot)$,not bi-modal, ▪ P_t continuous or discrete, but not Bernoulli, ▪ $\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$
Higher moments Lewbel (1997)	▪ One endogenous variable; ▪ Additional regressors allowed; ▪ Two-stage least squares estimation.	▪ Z_t skewed, ▪ ϵ_t , ν_t symmetrically distributed, ▪ Third moment of the data exists.
Heteroscedastic errors Lewbel (2012)	▪ Multiple endogenous variables; ▪ Additional regressors allowed; ▪ Two-stage least squares estimation.	▪ $cov(W_t, v_t^2) \neq 0$ ▪ $E(W_t W_t')$- not singular ▪ $E(W_t \epsilon_t) = E(W_t \nu_t) = 0$
Multilevel GMM Kim and Frees (2007)	▪ The model: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$ $\beta_{1j} = \gamma_{10}$ ▪ GMM estimation	▪ $cov(P_t, \epsilon_t \nu_t) = 0$ ▪ $e_{ij} \sim N(0, \sigma_e^2)$ ▪ $u_{0j} \sim N(0, \sigma_u^2)$ $cov(X_{ij}, e_{ij}, W_j, u_{0j}) = 0$

Table 1: Description of internal instrumental variables methods in REndo

4. Results

- **REndo** offers estimation techniques that *do not require external instrumental variables*.
- Some of the IIV methods come with a set of strong assumptions especially regarding the distribution of the endogenous regressor.
- The IIV methods are in general less efficient than OLS or internal IV methods with good instruments.
- The internal instruments constructed can be used in addition to any existing external instrument to increase efficiency.
- The performance of two of these methods, the higher moments and the heteroskedastic errors approaches, relative to the OLS is presented in Figure 1 below.

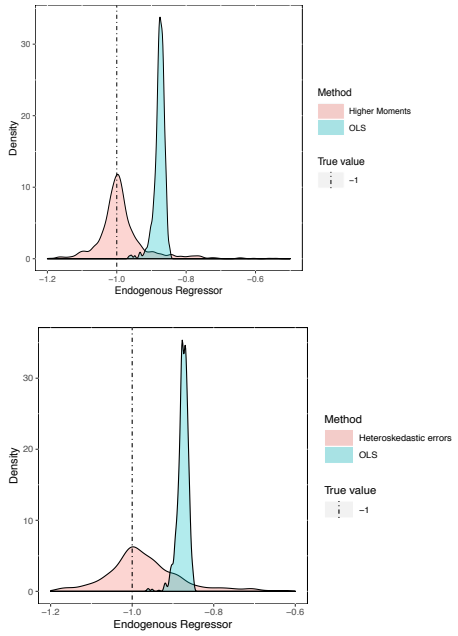

Fig.1: a) Comparison between the higher moments method and OLS, over 1000 simulations b) Comparison between the heteroskedastic errors method and OLS, over 1000 simulations

5. Conclusion

- The existence of a good external instrumental variable is often questionable.
- Statistical models that **exploit the distribution or the hierarchical structure** of the data make it possible to create instruments from the existing data, with **no need of external instruments. REndo**, implements five of these models.

Future research:

- An internal instrumental variable method that addresses endogeneity with a **binary or categorical** dependent variable.
- Integration of **diagnostic tests** for the copula correction and the latent internal variable approaches.

References

1. Ebbes P, Wedel M, Boeckenholt U, Steerneman A (2005). "Solving and testing for regressor- error (in)dependence when no instrumental variables are available: with new evidence for the effect of education on income". Quantitative Marketing and Economics, 3, 365–392.
2. Hogan V, Rigobon R (2003). "Using unobserved supply shocks to estimate the returns to educations". Technical report, University College Dublin.
3. Kim S, Frees F (2007). "Multilevel Modeling with Correlated Effects". Pshychometrika, 72(4), 505-533.
4. 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, 1201–1213.
5. Lewbel A. (2012). "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenoue Regressor Models". Journal of Business & Economic Statistics, 30(1), 67-80.
6. Park S, Gupta S (2012). "Handling Endogeneous Regressors by Joint Estimation Using Copulas". Marketing Science, 31(4), 567–586.