University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2023

REndo: Internal Instrumental Variables to Address Endogeneity

Gui, Raluca Ioana ; Meierer, Markus ; Schilter, Patrik ; Algesheimer, René

Abstract: Endogeneity is a common problem in any causal analysis. It arises when the independence assumption between an explanatory variable and the error in a statistical model is violated. The causes of endogeneity are manifold and include response bias in surveys, omission of important explanatory variables, or simultaneity between explanatory and response variables. Instrumental variable estimation provides a possible solution. However, valid and strong external instruments are difficult to find. Consequently, internal instrumental variable approaches have been proposed to correct for endogeneity without relying on external instruments. The R package REndo implements various internal instrumental variable approaches, i.e., latent instrumental variables estimation (Ebbes, Wedel, Boeckenholt, and Steerneman 2005), higher moments estimation (Lewbel 1997), heteroscedastic error estimation (Lewbel 2012), joint estimation using copula (Park and Gupta 2012) and multilevel generalized method of moments estimation (Kim and Frees 2007). Package usage is illustrated on simulated and real-world data.

DOI: https://doi.org/10.18637/jss.v107.i03

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-237025
Journal Article
Published Version

The following work is licensed under a Creative Commons: Attribution 3.0 Unported (CC BY 3.0) License.

Originally published at:
Gui, Raluca Ioana; Meierer, Markus; Schilter, Patrik; Algesheimer, René (2023). REndo: Internal Instrumental Variables to Address Endogeneity. Journal of Statistical Software, 107(3):1-43.
DOI: https://doi.org/10.18637/jss.v107.i03

Journal of Statistical Software

September 2023, Volume 107, Issue 3. doi: 10.18637/jss.v107.i03

REndo: Internal Instrumental Variables to Address Endogeneity

Raluca Gui University of Zurich
Markus Meierer University of Geneva
Patrik Schilter University of Zurich
René Algesheimer University of Zurich

Abstract

Endogeneity is a common problem in any causal analysis. It arises when the independence assumption between an explanatory variable and the error in a statistical model is violated. The causes of endogeneity are manifold and include response bias in surveys, omission of important explanatory variables, or simultaneity between explanatory and response variables. Instrumental variable estimation provides a possible solution. However, valid and strong external instruments are difficult to find. Consequently, internal instrumental variable approaches have been proposed to correct for endogeneity without relying on external instruments. The R package **REndo** implements various internal instrumental variable approaches, i.e., latent instrumental variables estimation (Ebbes, Wedel, Boeckenholt, and Steerneman 2005), higher moments estimation (Lewbel 1997), heteroscedastic error estimation (Lewbel 2012), joint estimation using copula (Park and Gupta 2012) and multilevel generalized method of moments estimation (Kim and Frees 2007). Package usage is illustrated on simulated and real-world data.

Keywords: endogeneity, internal instrumental variables, multilevel models.

1. Introduction

In the absence of data based on a randomized experiment, endogeneity is a concern in any causal analysis. It implies that the independence assumption between at least one regressor and the error term is not satisfied, leading to biased and inconsistent results. The causes of endogeneity are manifold and include response bias in surveys, omission of important explanatory variables, or simultaneity between explanatory and response variables (Antonakis, Bendahan, Jacquart, and Lalive 2014; Angrist and Pischke 2009). A "devilishly clever" solution to cope with confounders when randomization is not possible is to find additional variables, the so-called "instrumental variable" (Antonakis *et al.* 2014; Theil 1958). These variables are correlated with the suspected endogenous regressor but not correlated with the

structural error. For example, when estimating the demand for bread, the price variable is endogenous since price and demand are jointly determined in the market. Thus, there is a simultaneity problem. A strong instrumental variable would be the number of rainy days in the year, which influences the wheat production and consequently the price of bread, without having any effect on the demand for bread. The difficulty of finding a good instrument is well known. Paradoxically, the stronger the correlation between the instrument and the regressor, the more difficult it is to defend its lack of correlation with the error (Ebbes *et al.* 2005; Lewbel 1997).

Sometimes a suitable instrumental variable can be indicated by the data generating process or by the cause of endogeneity. However, frequently the search for an adequate instrumental variable fails. In this case, an "instrument free" or "internal instrumental variable" (IIV) model can be used. These methods address the endogeneity problem without the need for external instruments.

REndo (Gui, Meierer, Algesheimer, and Schilter 2023) is the first R (R Core Team 2023) package to implement the most recent IIV methods. The package includes implementations of the latent instrumental variable approach (Ebbes *et al.* 2005), the joint estimation using copula (Park and Gupta 2012), the higher moments method (Lewbel 1997), and the heteroscedastic error approach (Lewbel 2012). To model hierarchical data such as students nested within classrooms, nested within schools, **REndo** includes the multilevel generalized method of moments (GMM) estimation proposed by Kim and Frees (2007). All approaches assume a continuous dependent variable. The package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=REndo`.

Section 2 elaborates on the endogeneity problem and its impact on the parameter estimates. Section 3 provides an overview of the intuition and existing software implementations of external instrumental variable approaches. Analogously, Section 4 describes the underlying idea of the internal instrumental variables listed above alongside their specificities. Section 5 describes the implementation of these IIV models in **REndo** and their usage on various simulated datasets. Section 6 outlines the usage of these approaches on real-world data. Thereby, we illustrate how **REndo** facilitates applying multiple approaches to the same dataset and thus, enables researchers to address endogeneity rigorously. The last section presents directions for future development.

2. Endogeneity

Researchers from a wide range of disciplines, from political science, finance, management, marketing or education research are interested in causal relationships. They ask questions such as: Does institutional quality explain the variation in the economic development of different countries? Does CEO's compensation depend on firm size? Does an increase in advertising expenditure lead to an increase in sales? Does repeating a class lead to better test scores?

In order to be able to answer such questions, there are two possible options: (1) run a randomized controlled experiment or (2) use observational data. In many instances, randomized experiments are too expensive, rely on rather small sample sizes, or are seen as unethical. Thus, researchers often turn to observational data. The problem with such data is the threat of obtaining parameter estimates that are not consistent, meaning that they will not converge

to the true population parameter as the sample size increases. This threat occurs when

1. important variables are omitted from the model,
2. one or more explanatory variables are measured with error,
3. there is simultaneity between the response variable and one of the explanatory variables,
4. the sample is biased due to self-selection, or
5. a lagged response variable is included as a covariate.

Ruud (2000) showed that (2)–(5) can be viewed as a special case of (1). Nonetheless, in all these instances the error term of the model is correlated with one of the covariates. This is known as endogeneity. Figure 1 depicts an endogeneity problem: regressor P is endogenous due to its correlation with the error ϵ, while X is an exogenous covariate, since its correlation with the error is zero.

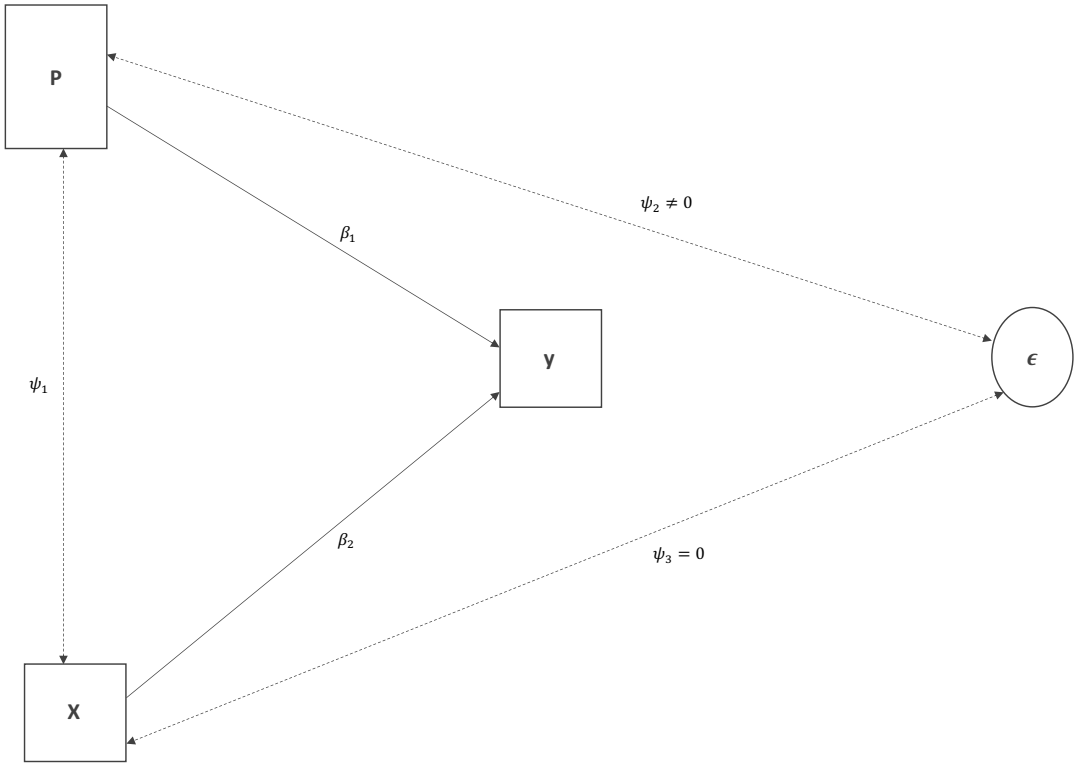
Figure 2 shows the results of a simulation in order to exemplify how the bias of coefficient estimates of the endogenous regressor increases as the correlation between P and the error increases. At low correlation (0.1), the bias is 0.11. But as the correlation between P and the error increases to 0.3 and then to 0.5, so does the bias: it increases from 0.24 to 0.34. The simulation was run over 1000 iterations where each sample had 2500 observations and the error and the omitted variable were normally distributed with a mean 0 and a standard deviation equal to 1.

Sometimes the cause of endogeneity or the data at hand can give clues on how to handle the problem. For example, if endogeneity arises from time-invariant sources, applying fixed-effects estimation on a panel dataset eliminates the omitted variable problem. For endogeneity caused by measurement error, autoregressive models or simultaneous equation models could offer a solution. Structural models that estimate demand-supply models (Draganska and Jain 2004; Berry, Levinsohn, and Pakes 1995; Berry 1994) are yet another alternative to deal with endogeneity.

However, the most frequently used methods in addressing endogeneity are instrumental variables (Theil 1958; Wright 1928). The main idea behind these approaches is to focus on the variations in the endogenous variable that are uncorrelated with the error term and disregard the variations that bias the ordinary-least squares coefficients. This is possible by finding an additional variable, called external instrument, such that the endogenous covariate can be separated into two parts: (1) the instrumental variable that (a) should not be correlated with the structural error and (b) should be correlated with the endogenous regressor; (2) the other part, which is correlated with the structural error of the model. Section 3 provides an overview of the intuition and existing software implementations of external instrumental variable approaches.

3. External IV methods

The concept of instrumental variables was first derived by Wright (1928). He observed that "success with this method depends on success in discovering factors of the type A and B" (page 314), where A and B refer to instrumental variables. Specifically, an "instrumental



Condition	β_1	β_2	Explanation
$\psi_1 = 0$	Inconsistent	Consistent	P correlates with ϵ ($\psi_2 \neq 0$) thus β_1 is inconsistent. β_2 is consistent since X is uncorrelated with both P ($\psi_1 = 0$) and ϵ ($\psi_3 = 0$).
$\psi_1 \neq 0$	Inconsistent	Inconsistent	P correlates with ϵ ($\psi_2 \neq 0$) thus β_1 is inconsistent. Although X is uncorrelated with ϵ ($\psi_3 = 0$), β_2 is inconsistent since it is affected by the bias in P through X's correlation with P ($\psi_1 \neq 0$), although ψ_3 still equals zero.

Figure 1: Endogeneity causes inconsistent estimates.

variable" (IV) is defined as a variable Z (Equation 2) that is correlated with the explanatory variable P and uncorrelated with the structural error, ϵ, in Equation 1:

$$Y = \beta P + \epsilon, \tag{1}$$

$$P = \gamma Z + \nu, \tag{2}$$

where

- The error term ϵ stands for all exogenous factors that affect Y when P is held constant.

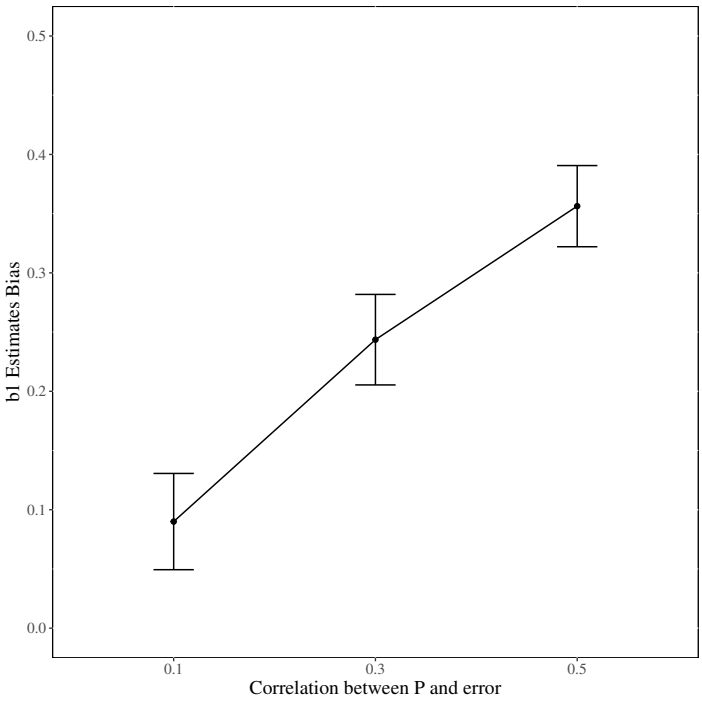


Figure 2: Estimates bias at different correlation levels between endogenous regressor and model error.

- The instrument Z should be independent of ϵ.
- The instrument Z should not affect Y when P is held constant (exclusion restriction).
- The instrument Z should not be independent of P.

Figure 3 gives a graphical representation of the notions above, where ψ_1 and ψ_2 represent the correlation between P and Z and between P and the error, respectively, ψ_3 is the correlation between the dependent variable and the error, while ω_1 represents the correlation between the instrumental variable and the error. The existence of an instrumental variable, Z, identifies the average direct effect (β_1) of the endogenous variable P on the outcome Y, independent of the unobserved sources of variability. The identification is achieved only if the exclusion restriction assumption is met, meaning that the effect of the instrumental variable on Y is solely moderated by its effect on P.

The following section outlines the most commonly used approaches which rely on external instrumental variables to model endogeneity. Section 3.2 provides an overview of available functions in R, Stata, and SAS which allow to apply these approaches.

3.1. Overview of estimation approaches

Today, researchers have a variety of instrumental variable models to choose from, depending on the research question, data type (cross-sectional vs. panel, single-level vs. multi-level), and on the number of available external instrumental variables.

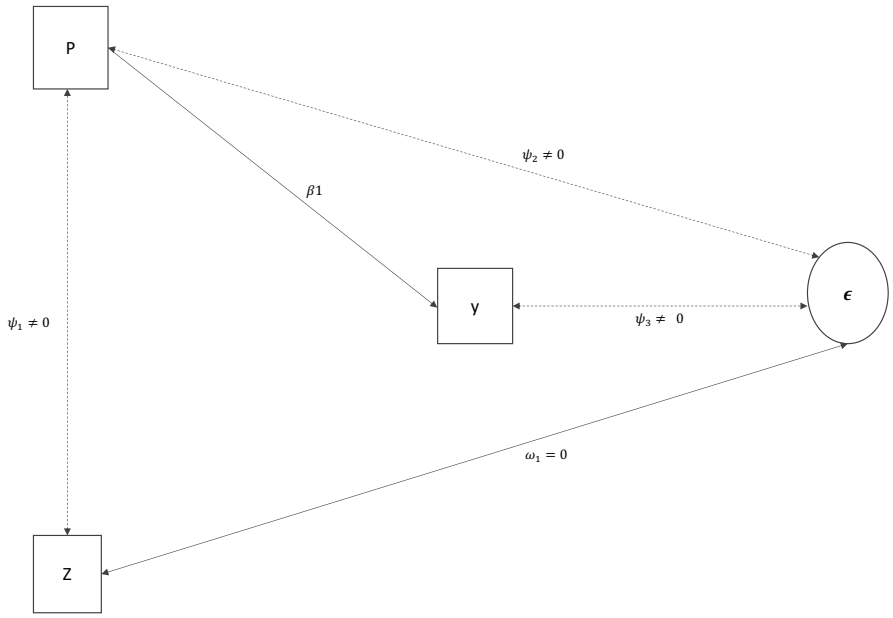


Figure 3: The instrumental variable Z solves the inconsistency of the estimates problem caused by endogeneity.

The simplest and most commonly used method in tackling endogeneity in linear single-level models is the two-stage least squares (2SLS) approach proposed by Theil (1958). In the first stage, each endogenous variable is regressed on all the exogenous variables in the model, both exogenous covariates and the excluded instruments. In the second stage, the regression of interest is estimated as usual via ordinary least squares, except that each endogenous covariate is replaced with the predicted values from the first stage.

Another widely used method for correcting endogeneity is the generalized method of moments (GMM) proposed by Hansen (1982). GMM is a class of estimators that are constructed by exploiting the sample moment counterparts of population moment conditions of the data generating model. When the system is over-identified and the sample size is large, GMM is more efficient than the two-stage least squares (2SLS) method.

However, in real-world applications we do not observe the impact of the endogeneity problem since, in fact, we cannot test how large the correlation between the endogenous regressor and the error is. Thus, for an unbiased and consistent estimate it is important to find a very good, "strong" instrumental variable (see Appendix A). Stock and Yogo (2002) were the first to point to the importance of the quality of the excluded instrumental variable used in the IV estimation. They were the first to differentiate external instrumental variables according to their correlation with the endogenous regressor into weak (low correlation) and strong (high correlation) instruments (Stock and Yogo 2002). Considering Equation 2, Stock and Yogo (2002) constructed the "concentration" parameter, μ^2: $\mu^2 = \gamma^\top Z^\top Z\gamma/\sigma_\nu$. Depending on the number of instruments, the proposed thresholds for the concentration parameter when the instruments are considered weak (Stock and Yogo 2002).

Depending on strength of the instrumental variables and on the sample size, the performance of external instrumental variable methods compared to the ordinary least squares (OLS) can vary significantly (see Appendix A, Table 6 for a comparison of OLS and the two-stage

Software	Approach	Level dep. v.		Level end. v.		Function (Package)
		Cont.	Bin.	Cont.	Dis.	
R	2SLS			✓		`ivreg` (**AER**, Kleiber and Zeileis 2022), `tsls` (**sem**, Fox, Nie, and Byrnes 2022), `systemfit` (**systemfit**, Henningsen and Hamann 2007), `plm` (**plm**, Croissant and Millo 2008)
	3SLS	✓		✓		`systemfit` (**systemfit**, Henningsen and Hamann 2007)
	GMM	✓		✓	✓	`gmm` (**gmm**, Chaussé 2010)
	SEM	✓	✓	✓	✓	`sem` (**sem**, Fox *et al.* 2022), `sem` (**lavaan**, Rosseel 2012)
	Copula splines	✓	✓	✓		`gjrm` (**GJRM**, Marra and Radice 2022)
Stata	2SLS, 3SLS	✓	✓	✓		`ivregress` (StataCorp 2015), `reg3` (StataCorp 2015)
	2SLS		✓	✓	✓	`ivprobit` (StataCorp 2015)
	2SLS		✓	✓		`sspecialreg` (Baum 2012)
	GMM	✓		✓		`ivregress` (StataCorp 2015), `gmm` (StataCorp 2015)
	SEM	✓		✓	✓	`sem` (StataCorp 2015)
	SEM	✓	✓	✓	✓	`gsem` (StataCorp 2015)
SAS	2SLS	✓	✓	✓		`PROC SYSLIN` (SAS Institute Inc. 2020)
	3SLS	✓	✓	✓		`INSTRUMENTS` (SAS Institute Inc. 2020)
	GMM	✓	✓	✓		`INSTRUMENTS` (SAS Institute Inc. 2020)
	SEM	✓	✓	✓	✓	`CALIS` (SAS Institute Inc. 2020)
	FIML	✓	✓	✓	✓	`PROC QLIM` (SAS Institute Inc. 2020)

Table 1: External instrumental variable approaches implemented in R, Stata, and SAS. The table illustrates the various implementations and categorizes them according to the level of the dependent variable (continuous or binary) and the level of the endogenous variable (continuous or discrete).

least squares method with a weak and a strong instrument). Finding a suitable instrumental variable is non-trivial, if ever possible. Therefore, researchers have proposed alternative "instrument-free" models like the ones in **REndo**.

3.2. Software implementing external IV methods

The two-stage least squares and the generalized method of moments are two of the most used approaches for instrumental variable estimation. There are several implementations in R (R Core Team 2023), Stata (StataCorp 2019), and the SAS software (SAS Institute Inc. 2020).

In R, 2SLS can be implemented using the `ivreg()` function in the **AER** (Kleiber and Zeileis

2022) package as well as `tsls()` function in the **sem** package (Fox *et al.* 2022). The `gmm()` function in the **gmm** (Chaussé 2010) performs instrumental variables regression using generalized method of moments estimation. Additionally, in the context of structural equation modeling, one can model the correlation between an explanatory variable and the error using the two available packages, **sem** and **lavaan** (Rosseel 2012).

In Stata, the `ivregress()` function supports estimation either via two-stage least squares, limited-information maximum likelihood (LIML) or generalized method of moments. Furthermore, the `gmm()` function performs instrumental-variables regression using the generalized method of moments approach, while the `sem()` and `gsem()` functions can be used to model the unobserved correlation in multiple equations, for continuous responses and for binary, count or multinomial responses, respectively.

In SAS the 2SLS option in the `PROC SYSLIN` statement implements the two-stage least squares estimation. For instrumental variable estimation using the generalized method of moments approach, one needs to use the `INSTRUMENTS` statement together with the `gmm` option, inside the `MODEL` procedure. In a structural equation modeling approach, one can address the unobserved correlation in SAS using the CALIS procedure.

Additional packages and procedures in R, Stata and SAS that address the endogeneity problem using external instrumental variables are presented in Table 1. With a vivid interest in the endogeneity topic, new packages appear constantly.

Therefore, this table is far from being exhaustive and we underlined only the key packages that treat the endogeneity problem. A detailed look at these implementations illustrates two main points: (1) Controlling for endogeneity should rely on using a diverse set of models carefully considering the identifying assumptions underlying these models (Germann, Ebbes, and Grewal 2015). (2) A harmonized user interface that facilitates a straightforward estimation of alternative models and comparison of results is key.

4. Internal IV methods

Internal instrumental variable methods have been proposed for cases when no observable variable satisfies the properties of a strong instrumental variable. The identification strategy of these methods is conditional on distributional assumptions of the endogenous regressors and of the error term. For example, Ebbes *et al.* (2005) assume the distribution of the endogenous variable to be discrete, Lewbel (1997), Lewbel (2012) and Park and Gupta (2012) consider a skewed distribution, while Rigobon (2003) work with a heteroscedastic distribution.

4.1. Software implementing internal instrumental variable methods

Contrary to the external instrumental variable methods, implementations of methods that address endogeneity with internal instruments are scarce. In R, a notable exception is the implementation of the heteroscedastic errors method (Lewbel 2012), in the **ivlewbel** (Fernihough 2014) package. This technique allows the identification of structural parameters in regression models with endogenous or mismeasured regressors. Identification is achieved by having regressors that are uncorrelated with the product of heteroscedastic errors, which is a feature of many models where error correlations are due to an unobserved common factor. Therefore, instruments may be constructed as simple functions of the model's data. This approach may be applied when no external instruments are available, or, alternatively, used

to supplement external instruments to improve the efficiency of the IV estimator. In Stata, the only IIV approach, implemented by the `ivreg2h()` function (Baum, Schaffer, and Stillman 2002), is also the heteroscedastic errors approach proposed by Lewbel (2012). A useful feature of the Stata implementation is the option to estimate the model using the two-stage least squares approach where the user already has an external instrumental variable, or a model that before augmentation with the generated instruments fails to be identified. In the former case, the Stata implementation provides three sets of estimates: the traditional IV parameter estimates, parameter estimates using only generated instruments and parameter estimates using both, external and internal instruments.

The Stata implementation of Lewbel (2012)'s approach considers all the exogenous regressors when building the internal instruments. However, we find it preferable for the function to allow the user to specify the set of variables from which to construct the instruments. Moreover, a warning message displayed when the model assumptions are not met would also be desirable. We considered these two points in our implementation of Lewbel's heteroscedastic errors approach to endogeneity. Lewbel (2012)'s approach, while becoming more popular due to its availability in both in Stata and R, is not without drawbacks – such as large standard errors and high sensibility to the form of heteroscedasticity assumed (Chau 2015). Therefore, having alternative approaches that address endogeneity through instrument-free methods would help researchers to choose the best method given the assumptions of their model and their data.

4.2. Internal instrumental variable methods implemented by REndo

Extending the rather limited set of IIV implementations across different software packages, **REndo** offers researchers and practitioners the possibility to estimate a variety of IIV approaches which rely on different techniques to control for potentially endogenous regressors. Building up on the findings of our survey of existing implementation for both, external and internal instrumental variable approaches, **REndo** aims to provide a harmonized interface across implementations to facilitate a straightforward estimation and comparison of alternative models (e.g., by using the formula interface and S3 generic methods).

The following sections describe the underlying idea of internal instrumental variables alongside the specificities of the IIV approaches implemented. Among the existing single-level IIV approaches, **REndo** includes implementations of the latent instrumental variables approach proposed by Ebbes *et al.* (2005), the higher moments estimation proposed by Lewbel (1997), the heteroscedastic errors estimation (Lewbel 2012) and the joint estimation using Gaussian copula (Park and Gupta 2012).

There are many instances in which the data have a hierarchical structure, for example students clustered in classes and in schools. Even longitudinal data can be seen as a series of repeated measurements nested within individuals. For such data structures, multilevel models have been developed (Longford 1995; Raudenbush and Bryk 1986). These regression methods recognize the existence of data hierarchies by allowing for residual components at each level in the hierarchy. In these models, endogeneity can have two sources: either the regressors are correlated with the random components, or they are correlated with the structural error of the model at the lowest level (level-one dependence). While theoretically, methods such as two-stage-least squares, weighted two-stage-least squares, or generalized methods of moments could be used in a multilevel setting to deal with both level-one and level-two endogeneity, no such implementations are available in R. **REndo** is the first R package that implements

a method that tackles endogeneity in a multilevel setting by implementing the multilevel generalized method of moments method proposed by Kim and Frees (2007).

Table 2 gives an overview of the IIV approaches implemented in **REndo**, emphasizing the assumptions of each of the approaches. Four of the methods apply to single-level data, while the approach of Kim and Frees (2007) addresses endogeneity in multilevel settings. All approaches allow for just one endogenous regressor. The exception is the copula correction method (Park and Gupta 2012). The response variable is assumed to be continuous in all methods. Regarding possible missing data, it is left to the user to address this point (e.g., apply imputation methods), the execution is stopped if there is any non-finite data in the input.

4.3. Internal instrumental variable methods for non-hierarchical data

The four internal instrumental variable methods presented in this section share the same underlying model presented in Equations 3 and 4. The specific characteristics of each method are discussed in the subsequent sections.

Consider the model:

$$Y_t = \beta_0 + P_t\beta_1 + X_t^\top\beta_2 + \epsilon_t \tag{3}$$

where $t = 1, \ldots, T$ indexes either time or cross-sectional units, Y_t is a 1×1 response variable, X_t is a $k \times 1$ vector of exogenous regressors, where k is the number of exogenous regressors, P_t is a 1 x 1 continuous endogenous regressor, and ϵ_t is the structural error term, assumed to have mean zero and variance σ_ϵ^2. β_0, β_1 are model parameters and β_2 is a $k \times 1$ model parameter vector. The endogeneity problem arises from the correlation of P_t and ϵ_t. As such:

$$P_t = Z_t^\top\gamma + \nu_t \tag{4}$$

where Z_t is a $l \times 1$ vector of internal instrumental variables, γ is a $l \times 1$ vector of parameters and ν_t is the random error with mean zero, variance equal to σ_ν^2 and $\mathsf{E}(\epsilon_t\nu_t) = \sigma_{\epsilon\nu}$. Z_t is assumed to be stochastic with distribution G and ν_t is assumed to have density $h(\cdot)$. The latent instrumental variables and the higher moments models assume Z_t to be uncorrelated with the structural error, which is similar to the "exclusion restriction" assumption for observed instrumental variables methods. Moreover, Z_t is also assumed unobserved.

Internal instrumental variables models require for identification that the distribution of the endogenous regressor, P_t, is distinct from the distribution of the structural error, ϵ_t, which, in most IIV models such as LIV or copula correction is assumed normal. Otherwise, assuming the same distribution for the structural error and the endogenous regressor, would make impossible separating the variation due to the endogenous regressor from that due to the error (Park and Gupta 2012). Therefore, the proposed instruments work best when the sample distribution of the endogenous regressor is "as skewed as possible" (Lewbel 2012, 1997).

The following sections share a general structure: First, we present the underlying idea of each method. Second, we discuss its specific characteristics and finally, present the particular assumptions and weaknesses of the method.

Latent instrumental variables method

Ebbes *et al.* (2005) propose the latent instrumental variables (LIV) model as defined in Equations 3 and 4, with both errors being normally distributed. LIV does not accept additional

Method	Short description	Assumptions
Latent instrumental variable method Ebbes *et al.* (2005)	– one endogenous regressor – no additional exogenous regressors – maximum likelihood (ML) estimation – single-level model	– $P_t \neq N(\cdot, \cdot)$ – $\epsilon_t \sim N(0, \sigma_\epsilon^2)$; $\text{corr}(Z_t,\, \epsilon_t) = 0$ – Z_t discrete with at least 2 groups with different means
Copula correction method Park and Gupta (2012)	– multiple endogenous regressors – additional exogenous regressors allowed – ML estimation – single-level model	– $P_t \neq N(\cdot, \cdot)$, $P_t \neq$ bimodal – P_t can be discrete, but not Bernoulli – $\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$
Higher moments method Lewbel (1997)	– one endogenous regressors – additional exogenous regressors allowed – TSLS estimation – single-level model	– Z_t skewed distribution – $\mathsf{E}(\epsilon_t) = 0$, $\mathsf{E}(\nu_t) = 0$ – Third moment of the data exists
Heteroscedastic errors method Lewbel (2012)	– multiple endogenous regressors – additional exogenous regressors allowed – TSLS estimation – single-level model	– $\mathsf{COV}(Z, \nu^2) \neq 0$ – $\mathsf{E}(X\epsilon_t) = 0$ – $\mathsf{E}(X\nu_t) = 0$ – $\mathsf{E}(XX^\top)$ – non-singular
Multilevel GMM method Kim and Frees (2007)	– multiple endogenous regressors – additional exogenous regressors allowed – GMM estimation – multilevel model	Model: $Y_{ij} = \beta_{0j} + \beta_{ij}X_{ij} + \epsilon_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$ $\beta_{ij} = \gamma_{10}$ – $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $u_{0j} \sim N(0, \sigma_u^2)$ – $\mathsf{COV}(X_{ij}, \epsilon_{ij}, W_j, u_{0j}) = 0$

Table 2: Overview of selected internal instrumental variable models.

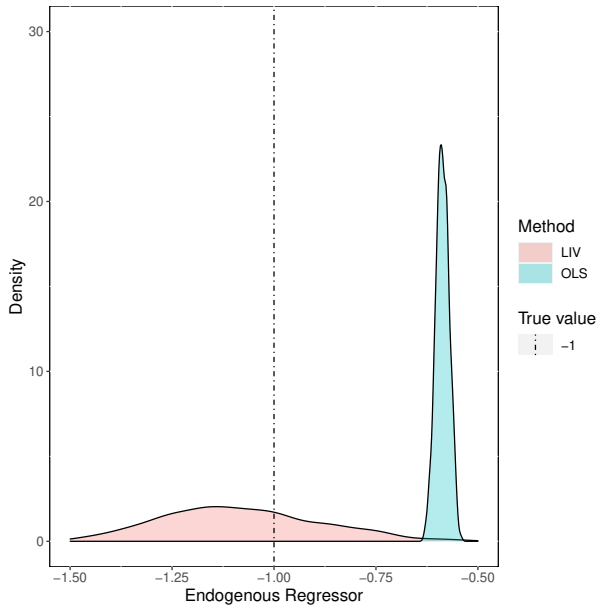


Figure 4: Parameter estimates for latent instrumental variable method vs. OLS. Parameter estimates were obtained over 1000 simulated samples, each of size 2500; the true parameter value is -1.

covariates besides the endogenous regressor, P_t, whose distribution has to be different than normal. A particular characteristic of this approach is that the internal instrumental variables Z_t are assumed unobserved, discrete and exogenous, with an unknown number of groups m, while γ is a vector of group means. The method accepts just one endogenous regressor. Identification of the parameters relies on the distributional assumptions of the latent instruments as well as that of the endogenous regressor, P. Specifically, the endogenous regressor should have a non-normal distribution while the unobserved instruments, should be discrete and have at least two groups with different means (Ebbes, Wedel, and Boeckenholt 2009). A continuous distribution for the instruments leads to an unidentified model, while a normal distribution of the endogenous regressor gives rise to inefficient estimates. The interested reader can find a more detailed description of the identification strategy of the LIV method in Ebbes *et al.* (2005), in the last paragraph of the second chapter (page 369), and a formal proof in Appendix I (page 385).

The LIV model implemented in **REndo** assumes that the latent instrumental variable has two categories. Ebbes *et al.* (2005) showed, using Monte Carlo simulations, that the model estimates are still consistent and the power of the test is good even if the number of categories is misspecified. A simulation study provides evidence of the ability of the LIV model to recover the true parameter value for the endogenous regressor if the assumptions are met. The endogenous regressor has a true parameter value equal to -1. Figure 4 shows the distribution of LIV parameter estimates in comparison with ordinary least squares for 1000 simulated samples. While the OLS estimate is biased, the LIV estimate is unbiased but has a larger standard deviation. The mean and variance of the OLS estimate are -0.59 and 0.02, while the mean and variance of the LIV estimate are -1.08 and 0.22 respectively. For a more detailed overview of the method see Ebbes *et al.* (2005).

Joint estimation using copula method

Park and Gupta (2012) propose a method that allows for the joint estimation of the continuous endogenous regressor and the error term using Gaussian copulas. A copula is a function that maps several conditional distribution functions (CDF) into their joint CDF (see Appendix C, Figure 10).

The underlying idea of the method is that using information contained in the observed data, one selects marginal distributions for the endogenous regressor and the structural error term, respectively. Then, the copula model enables the construction of a flexible multivariate joint distribution allowing a wide range of correlations between the two marginals.

In Equation 3, the error ϵ_t is assumed to have a normal marginal distribution, while the marginal distribution of the endogenous regressor P_t is obtained using the Epanechnikov kernel density estimator (Epanechnikov 1969), as below:

$$\hat{h}(p) = \frac{1}{T \cdot b} \sum_{t=1}^{T} K\left(\frac{p - P_t}{b}\right)$$

where P_t is the endogenous regressor, $K(x) = 0.75 \cdot (1 - x^2) \cdot I(\|x\| <= 1)$ and the bandwidth b is the one proposed by Silverman (1986), and is equal to $b = 0.9 \cdot T^{-1/5} \cdot \min(s, \text{IQR}/1.34)$. IQR is the interquartile range while s is the data sample standard deviation and T is the number of time periods observed in the data. After obtaining the joint distribution of the error term and the continuous endogenous regressor, the model parameters are estimated using maximum likelihood (ML) estimation.

With more than one continuous endogenous regressor or an endogenous discrete regressor, an alternative approach to the estimation using Gaussian copula should be applied. This approach is similar to the control function approach (Petrin and Train 2010). The core idea is to apply ordinary least squares estimation on the original set of explanatory variables in Equation 3 as well as an additional regressor, namely $P_t^* = \Phi^{-1}(H(P_t))$. $H(P_t)$ is the marginal distribution of the endogenous regressor, P. Including this regressor solves the correlation between the endogenous regressor and the structural error in Equation 3, ordinary least squares providing consistent parameter estimates. Due to identification problems, the discrete endogenous regressor cannot have a binomial distribution. For more details on the identification strategy, interested readers are referred to the first section of Chapter 2 (pages 570–573) in Park and Gupta (2012).

A simulation study highlights that the method performs well in recovering the true parameter values even with violations of the normality assumption of the structural error, given that the distribution of the endogenous regressor is different from that of the error. Figure 5 illustrates the performance of the copula correction method compared to the performance of the OLS. The parameter estimates of the copula correction method are less biased than those of OLS when endogeneity is present, even though they are slightly less consistent. The parameter estimate of the endogenous variable has a mean and variance of -0.73 and 0.03 with OLS, while the mean and variance of the estimate obtained with the copula correction method are -1.00 and 0.06 respectively. For further technical details, see Park and Gupta (2012).

Higher moments method

The method proposed by Lewbel (1997) helps identifying structural parameters in regression

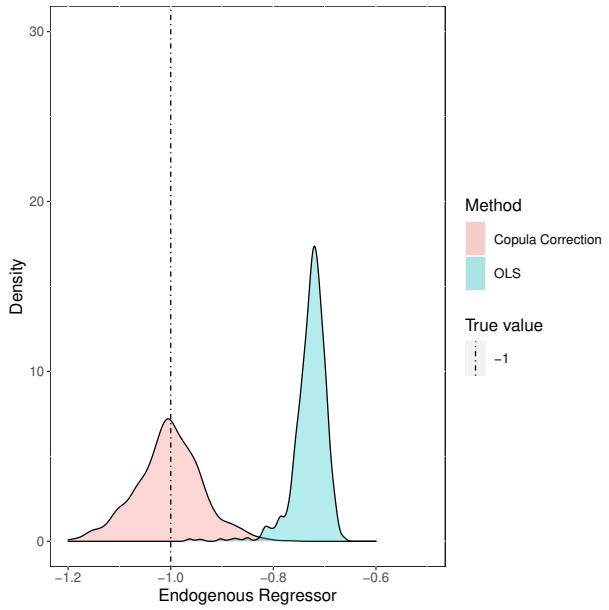


Figure 5: Parameter estimates for copula correction method vs. OLS. Parameter estimates were obtained over 1000 simulated samples, each of size 2500; the true parameter value is −1.

models with endogeneity caused by measurement error, as exposed in Equations 3 and 4. Identification is achieved by exploiting third moments of the data.

Internal instruments, Z_t, alongside the error terms in Equations 3 and 4, ϵ_t, and ν_t, respectively, are assumed unobserved. Unlike previous models, no restriction is imposed on the distribution of the error terms, while their means are set to zero. Lewbel (1997) proves that the following instruments can be constructed and used with two-stage least squares estimation to obtain consistent estimates:

$$q_{1t} = (G_t - \bar{G}) \tag{5}$$
$$q_{2t} = (G_t - \bar{G})(P_t - \bar{P})$$
$$q_{3t} = (G_t - \bar{G})(Y_t - \bar{Y})$$
$$q_{4t} = (Y_t - \bar{Y})(P_t - \bar{P})$$
$$q_{5t} = (P_t - \bar{P})^2 \tag{6}$$
$$q_{6t} = (Y_t - \bar{Y})^2 \tag{7}$$

Here, $G_t = G(X_t)$ for any given function $G(\cdot)$ that has finite third own and cross moments and X_t are all the exogenous in the model. $\bar{G}$ is the sample mean of G_t. The same rule applies to P_t and Y_t.

The instruments in Equations 6 and 7 can be used only when the measurement and the structural errors are symmetrically distributed. Otherwise, the use of the instruments does not require any distributional assumptions for the errors. Given that the regressors $G(X_t)$ are included as instruments, $G()$ should not be linear in X_t in Equation 5 (e.g., $G()$ function could be the square, cubic or logarithmic function).

An important part for identification is the assumption of skewness of the endogenous regressor.

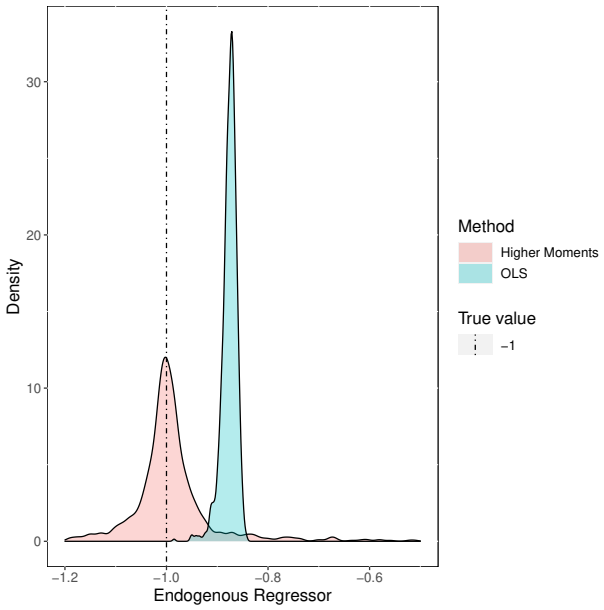


Figure 6: Parameter estimates for higher moments method vs. OLS. Parameter estimates were obtained over 1000 simulated samples, each of size 2500; the true parameter value is -1.

To cite Lewbel (1997, page 1204), "the greater the skewness, the better the quality of the proposed instruments". If this assumption fails, the instruments may be weak, and thus, the parameter estimates will be biased. Since the instruments constructed come along with very strong assumptions, one of their best uses is to provide over-identifying information. The over-identification provided by constructed moments can be used to test the validity of a potential outside instrument, to increase efficiency, and to check for robustness of parameter estimates based on alternative identifying assumptions. Details on the assumptions and proof of the theorems for identification can be found in Chapter 2 (pages 1201–1204) in Lewbel (1997).

A simulation study provides evidence of the ability of the Higher Moments method to recover the true parameter value if its assumptions are met. Figure 6 illustrates the parameter estimates for the endogenous regression of the higher moments method and OLS for 1000 simulated samples. While ordinary least squares are more biased, with an average parameter estimate equal to -0.88 and variance equal to 0.02, the higher moments estimator has an average of -0.99 but it is volatile, with a variance of 0.32. For more details and the proof of the above assumptions, see Lewbel (1997).

Heteroscedastic errors method

The method proposed in Lewbel (2012) identifies structural parameters in regression models with endogenous regressors by means of variables that are uncorrelated with the product of heteroscedastic errors. This feature is encountered in many models in which error correlations are due to an unobserved common factor (Lewbel 2012). The instruments are constructed as simple functions of the underlying data. The method can be applied when no external instruments are available or to supplement external instruments to improve the efficiency of the IV estimator.

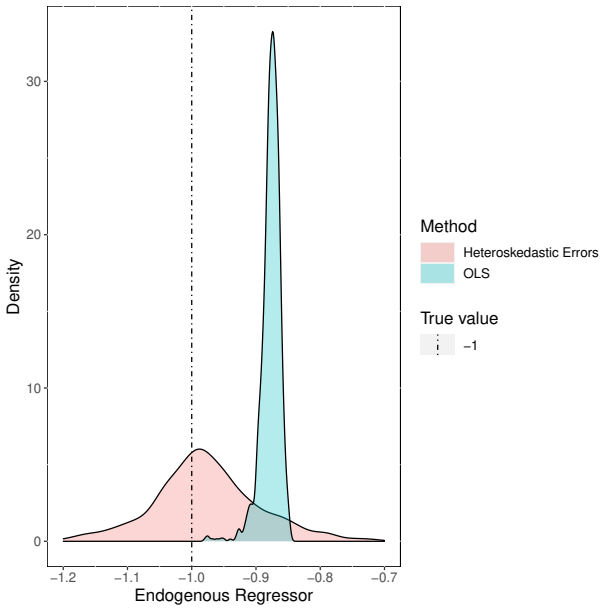


Figure 7: Parameter estimates for heteroscedastic errors method vs. OLS. Parameter estimates were obtained over 1000 simulated samples, each of size 2500; the true parameter value is -1.

Consider the model in Equations 3 and 4, with the exception that in Equation 4 P_t is a function of X_t, while Z_t is a subset of X_t. The model assumes that $\mathsf{E}(X_t\epsilon_t) = 0$, $\mathsf{E}(X_t\nu_t) = 0$ and $\mathsf{COV}(Z_t, \epsilon_t\nu_t) = 0$. The errors, ϵ_t and ν_t, may be correlated with each other. Structural parameters are identified by an ordinary two-stage least squares regression of Y_t on X_t and P_t, using X_t and $[Z_t - \mathsf{E}(Z_t]\nu_t$ as instruments. A vital assumption for identification is that $\mathsf{COV}(Z_t, \nu_t^2) \neq 0$. The strength of the instrument is proportional to the covariance of $(Z_t - \bar{Z}_t)\nu_t$ with ν_t, which corresponds to the degree of heteroscedasticity of ν_t with respect to Z (Lewbel 2012). The assumption that the covariance between Z and the squared error is different from zero can be empirically tested. If it is zero or close to zero, the instrument is weak, producing imprecise estimates, with large standard errors. The interested reader is referred to the first section of Chapter 2 (pages 8–9) in Lewbel (2012) for more details related to identification in a triangular system.

Under homoskedasticity, the parameters of the model are unidentified. But, identification is achieved in the presence of heteroscedasticity related to at least some elements of X_t. This strategy of identification is less reliable than the identification based on coefficient zero restriction, since it relies upon higher moments. But sometimes, it might be the only available strategy. The performance of this method in comparison to OLS is illustrated with a simulation study of over 1000 samples (see Figure 7). As with the previous methods, the heteroscedastic error method is less efficient than the OLS, but the coefficient estimate is unbiased. The mean parameter estimate of the endogenous regressor in the case of OLS is -0.87, compared to -0.98 in the case of the IIV method. The variance of the OLS estimate is 0.015 compared to 0.16 of the heteroscedastic Errors method.

4.4. Internal IV methods for hierarchical data

Many kinds of data have a hierarchical structure. Multilevel modeling is a generalization of regression methods that recognize the existence of such data hierarchies by allowing for residual components at each level in the hierarchy. For example, a three-level multilevel model which allows for grouping of students within classrooms over time, would include time, student, and classroom residuals (Equation 8). Thus, the residual variance is partitioned into four components: between-classroom (variance of the classroom-level residuals), and within-classroom (variance of the student-level residuals), between student (the variance of the student-level residuals) and within-student (variance of the time-level residuals). The classroom residuals represent the unobserved classroom characteristics that affect student's outcomes. These unobserved variables lead to a correlation between outcomes for students from the same classroom. Similarly, the unobserved time residuals lead to a correlation between a student's outcomes over time. A three-level model can be described as below:

$$\begin{aligned} y_{cst} &= Z^1_{cst}\beta^1_{cs} + X^1_{cst}\beta_1 + \epsilon^1_{cst} \\ \beta^1_{cs} &= Z^2_{cs}\beta^2_c + X^2_{cs}\beta_2 + \epsilon^2_{cs} \\ \beta^2_c &= X^3_c\beta_3 + \epsilon^3_c. \end{aligned} \tag{8}$$

Like in single-level regression, in multilevel models endogeneity is also a concern. The additional problem is that in multilevel models there are multiple independent assumptions involving various random components at different levels. Any moderate correlation between some predictors and a random component or error term, can result in a significant bias in the coefficients and of the variance components (Ebbes *et al.* 2005). While panel data models for dealing with endogeneity can be used to address the same problem in two-level multilevel models, their implications to higher levels have not been closely examined (Kim and Frees 2007). Moreover, panel data models allow only the inclusion of one random intercept while multilevel models require a methodology that can handle more general error structures.

Multilevel instrumental variables method

Exploiting the hierarchical structure of multilevel data, Kim and Frees (2007) propose a generalized method of moments technique for addressing endogeneity in multilevel models without the need for external instrumental variables. This approach uses both, the between and within variations of the exogenous variables, but only assumes the within variation of the variables to be endogenous. Another assumption in the multilevel generalized moment of moments model is that the errors at each level are normally distributed and independent of each other. Moreover, the slope variables are assumed to be exogenous. Since the model does not handle "level one dependencies", an additional assumption is that the level-one structural error is uncorrelated with any of the regressors. If this assumption is not met, additional, external instruments are necessary. It is to note that the multilevel data structure considered here excludes the possibility of non-nested clustering. The interested researcher can find more details about the identification strategy in sections 2 and 3 of the second chapter (pages 7–10) in Kim and Frees (2007).

Kim and Frees (2007) apply a three-level multilevel model as in Equation 8. The coefficients of the explanatory variables appear in vectors β_1, β_2, and β_3. The term β^1_{cs} captures latent, unobserved characteristics that are classroom and student-specific while β^2_c captures latent,

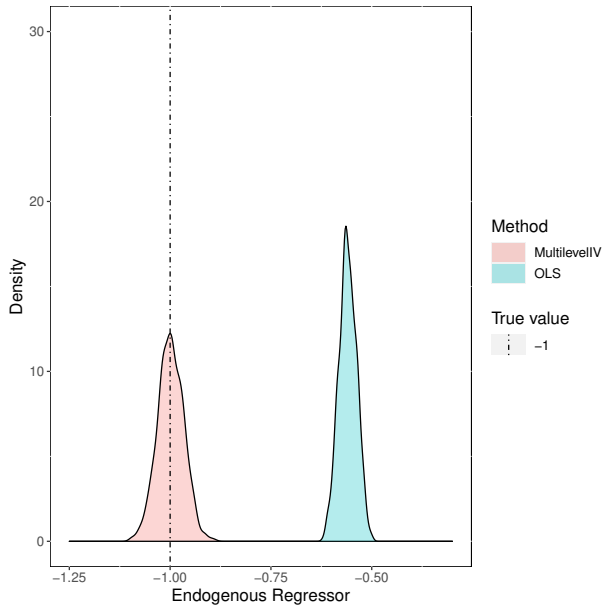


Figure 8: Parameter estimates for multilevel GMM method vs. regular random effects model. Parameter estimates were obtained over 1000 simulated samples, each of size 2500; the true parameter value is -1.

unobserved characteristics that are classroom specific. For identification, the disturbance term ϵ_{cst} is assumed independent of the other variables, Z^1_{cst} and X^1_{cst}.

Given the set of disturbance terms at different levels, there exists a couple of possible correlation patterns that could lead to biased results:

- errors at both levels (ϵ^2_{cs} and ϵ^3_c) are correlated with some of the regressors,
- only third level errors (ϵ^3_c) are correlated with some of the regressors,
- a third case, where there is a concern with errors at both levels, but there is not enough information to estimate level 3 parameters.

When all model variables are assumed exogenous, the GMM estimator is the usual GLS estimator, denoted as b_{RE}. When all variables are assumed endogenous, the fixed-effects estimator is used, b_{FE}. While b_{RE} assumes all explanatory variables are uncorrelated with the random intercepts and slopes in the model, b_{FE} allows for endogeneity of all effects but sweeps out the random components as well as the explanatory variables at the same levels. The more general estimator b_{GMM} proposed by Kim and Frees (2007) allows for some of the explanatory variables to be endogenous and uses this information to build instrumental variables. The multilevel GMM estimator uses both, the between and within variations of the exogenous variables, but only the within variation of the variables assumed endogenous. When all variables are assumed exogenous, b_{GMM} estimator equals b_{RE}. When all covariates are assumed to be endogenous, b_{GMM} equals b_{FE}. In facilitating the choice of the estimator to be used for the given data, Kim and Frees (2007) also propose an omitted variable test. This test is based on the Hausman test (Hausman 1978) for panel data. The omitted variable

test allows the comparison of a robust estimator and an estimator that is efficient under the null hypothesis of no omitted variables, and also the comparison of two robust estimators at different levels.

A simulation study provides evidence on the ability of this model to recover the true parameter value if its assumptions are met. Figure 8 shows the performance of the multilevel generalized method of moments method compared to the regular random effects model over 1000 samples. In this simulation study, each sample represents a three-level dataset with a single endogenous regressor on level-two. The mean random effects parameter estimate is -0.559 and standard error 0.022, while the multilevel GMM parameter estimate has a mean of -0.998 over 1000 simulations, with a standard error of 0.032.

Citing Kim and Frees (2007), "the (multilevel) GMM estimators can help researchers in the direction of exploiting the rich hierarchical data and, at the same time, impeding the improper use of powerful multilevel models" (page 529). **REndo** facilitates the application of this approach and thus, contributes to the growing number of studies relying on multilevel models.

5. Using REndo

REndo encompasses five functions that allow the estimation of linear models with one or more endogenous regressors using internal instrumental variables. Depending on the assumptions of the model and the structure of the data, single or multilevel, the researcher can use one of the following functions:

- `latentIV()` implements the latent instrumental variable estimation as in Ebbes *et al.* (2005). The endogenous variable is assumed to have two components – a latent, discrete, and exogenous component with an unknown number of groups and the error term that is assumed normally distributed and correlated with the structural error. The method supports only one endogenous, continuous regressor and no additional explanatory variables.

- `copulaCorrection()` models the correlation between the endogenous regressor and the structural error with the use of Gaussian copula (Park and Gupta 2012). The endogenous regressor can be continuous or discrete. The method also allows estimating a model with more than one endogenous regressor, either continuous, discrete or a mixture of the two. However, the endogenous regressors cannot have a binomial distribution, due to parameter identification problems.

  In the case of only one continuous endogenous regressor, the method uses maximum likelihood for estimation. In the case of a discrete endogenous regressor, or when several endogenous regressors are suspected, the estimation is carried out using a different model specification, in the style of the control function approach (Petrin and Train 2010), which is nonetheless based on Gaussian copulas.

- `higherMomentsIV()` implements the higher moments approach described in Lewbel (1997) where instruments are constructed by exploiting higher moments of the data, under strong model assumptions. The function allows just one endogenous regressor.

- `hetErrorsIV()` uses the heteroscedasticity of the errors in a linear projection of the endogenous regressor on the other covariates to solve the endogeneity problem induced by measurement error, as proposed by Lewbel (2012). The function allows at the moment only one endogenous regressor.

- `multilevelIV()` implements the instrument free multilevel GMM method proposed by Kim and Frees (2007) where identification is possible due to the different levels of the data. Endogenous regressors at different levels can be present. The function comes along a built-in omitted variable test, which helps in deciding which model is robust to omitted variables at different levels.

The package includes seven simulated datasets. Using just one dataset for exemplifying the functions is not possible due to different assumptions regarding the underlying data generating process for each of the methods. Where possible, the names of the variables were kept consistent across the datasets, with y for the response variable, P for the endogenous variables and X for the exogenous regressors. An intercept is always considered unless otherwise specified (by adding a "-1" in the first right-hand-side of the formula) and the true parameter value of the endogenous regressor is equal to -1 across all simulations. **REndo** builds up on functionalities provided by the following R packages: **Formula** (Zeileis and Croissant 2010), **optimx** (Nash and Varadhan 2011), **mvtnorm** (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2023), **AER** (Kleiber and Zeileis 2022), **lmtest** (Zeileis and Hothorn 2002), **Matrix** (Bates, Mächler, and Jagan 2023), **lme4** (Bates, Mächler, Bolker, and Walker 2015), **data.table** (Dowle and Srinivasan 2023), **corpcor** (Schafer, Opgen-Rhein, Zuber, Ahdesmaki, Silva, and Strimmer 2021), **Rcpp** (Eddelbuettel and François 2011), **RcppEigen** (Bates and Eddelbuettel 2013). Next, the usage of each of the five IIV functions is presented in the sections below.

5.1. Latent instrumental variables method

The syntax of the `latentIV()` function is: `latentIV(y ~ P, data, start.params, optimx.args, verbose)`. The first argument is the formula of the model to be estimated, `y ~ P`, where `y` is the response and `P` is the endogenous regressor, the second argument is the name of the dataset used while the third argument, which is optional, is a vector with the initial parameter values, `start.params`. Additionally, one can specify arguments for the optimization algorithm in the `optimx.args` argument, such as the optimization method or the maximum number of iterations. The argument `verbose` informs about the main steps that the method is performing (such as start parameters or the intermediary models). `TRUE` is the default for all methods. When set to the value `FALSE`, the `verbose` argument still returns warnings.

To illustrate the LIV approach, the dataset `dataLatentIV` was constructed.

```
R> data("dataLatentIV", package = "REndo")
R> resultsLIV <- latentIV(y ~ P, data = dataLatentIV)
R> summary(resultsLIV)

Call:
latentIV(formula = y ~ P, data = dataLatentIV)
Coefficients:
```

```
            Estimate Std. Error z-score Pr(>|z|)
(Intercept)  3.32351    0.37383   8.89   <2e-16 ***
P           -1.04580    0.04894  -21.37  <2e-16 ***
Further parameters estimated during model fitting:
   pi1    pi2 theta5 theta6 theta7 theta8
3.7652  8.5354  0.1905  1.1277  1.5754 13.7507
Initial parameter values:
(Intercept)=2.6275 P=-0.9545 pi1=7.6273 pi2=11.7843 theta5=0.5
theta6=1 theta7=0.5 theta8=1
The value of the log-likelihood function: 10627.18
AIC: -21238.35 , BIC: -21191.76
KKT1: TRUE  KKT2: TRUE Optimx Convergence Code: 0
```

The `summary()` function returns the parameter estimates (Intercept: 3.323, P: -1.045). As starting values for optimizing the log-likelihood function were not provided, the ordinary least squares parameter estimates were used, together with the default values for the correlation matrix and the probability of belonging to group 1: $(2.627, -0.954, 7.627, 111.784, 0.5, 1, 0.5, 1)$. Using maximum likelihood for estimation, the function also returns the log-likelihood value (10627.18), Akaike (AIC) and Bayesian (BIC) information criteria (AIC : -21238.35 and BIC : -21191.76), and a convergence code (here 0) that indicates whether the model converged (`0`) or, an issue occurred during estimation, which is indicated by a set of other codes, as seen with the help of the `optimx()` function. The coefficients, variance-covariance matrix, residuals and fitted values, as well as the Akaike and Bayesian information criteria, can be requested with the respective generic functions, e.g. `coef(resultsLIV)`, `vcov(resultsLIV)`, `residuals(resultsLIV)`, `fitted(resultsLIV)`, `AIC(resultsLIV)`, `BIC(resultsLIV)`, `logLik(resultsLIV)`.

The true parameter value of the endogenous regressor is -1. The coefficient for P produced by the latent internal instrument method is -1.045 with standard error 0.049, while OLS returns a coefficient equal to -0.954 and a standard error (SE) 0.004. LIV produces unbiased estimates for the endogenous regressor but standard errors are large. Therefore, when using the latent instrumental variable approach caution has to be taken when interpreting the results.

When optionally specifying the starting values, the following has to be considered: In any model there are seven or eight parameters, depending on whether an intercept is considered or not, where the first parameter is the intercept, if considered, then the coefficient of the endogenous variable followed by the means of the two groups of the underlying model. The means of the groups need to be different, otherwise, the model is not identified. These are returned by the parameters `pi1` and `pi2`. The three parameters, `theta5`, `theta6`, and `theta7`, return the estimates that compose the variance-covariance matrix. The parameter `theta8` returns the probability of being in group 1. When not provided, the initial parameter values are taken to be the OLS parameter estimates of regressing `y` on `P`. For the groups' means, the mean and the mean plus one standard deviation of the endogenous regressor are used as initial values. The variance-covariance parameters are taken to be 1, while the probability of being part of the first group is 0.

5.2. Joint estimation using Gaussian copula

In handling endogeneity by copula correction, either with the maximum likelihood estimation

Name	Exogenous var.	Endogenous var.	True parameter values
`dataCopCont`	$X1$, $X2$	P continuous, t distributed	Exog. var: $\beta_0 = 2$, $\beta_1 = 1.5$, $\beta_3 = -3$ Endog. var.: $\alpha_1 = -1$ $\text{corr}(P, \epsilon) = 0.33$
`dataCopDis`	$X1$, $X2$	P discrete, Poisson distributed	$\beta_0 = 2$, $\beta_1 = 1.5$, $\beta_3 = -3$, Endog. var.: $\alpha_1 = -1$, $\text{corr}(P, \epsilon) = 0.33$
`dataCopDisCont`	$X1$, $X2$	P_1: Poisson distributed, P_2: t distributed	Exog. var: $\beta_0 = 2$, $\beta_1 = 1.5$, $\beta_3 = -3$, Endog. var.: $\alpha_1 = -1$, $\alpha_2 = 0.8$, $\text{corr}(P1, \epsilon) = 0.33$, $\text{corr}(P_1, P_2) = 0.25$, $\text{corr}(P_2, \epsilon) = 0.25$

Table 3: Simulated datasets to exemplify the `copulaCorrection()` function.

or with the augmented OLS, the inference occurs in two steps. Therefore, the standard errors reported by the summary function are the bootstrapped standard errors. Due to the non-normality of the bootstrapped parameters, we report the upper and lower bounds of the 95% bootstrapped confidence interval. It is worth mentioning that the standard errors and the confidence intervals might slightly vary, depending on the platform, but keeping the sign and magnitude.

Three possible cases are discussed: The first case considers one continuous endogenous regressor, the second considers one discrete endogenous regressor and the last case assumes one continuous and one discrete endogenous regressors. To exemplifying the use of each of the functions, due to different model assumptions, three datasets have been simulated (see Table 3).

Case 1: Single continuous endogenous regressor

The `copulaCorrection()` can be used in the case of a single endogenous regressor with a continuous, skewed distribution. The syntax is as follows: `copulaCorrection(y ~ X1 + X2 + P | continuous(P), data, num.boots, start.params, verbose)` where the first argument is a two-part formula of the model to be estimated, with the second part of the RHS (right-hand side) defining the endogenous regressor as continuous, here `P`; the second argument is the name of the data, the third argument of the function is optional and represents the number of bootstraps to be performed (the default is 1000). The argument `start.params` is optional and represents the initial parameter values supplied by the user (when missing, the OLS estimates are considered). The function estimates the model using maximum likelihood and returns two additional estimates `rho` and `sigma` besides the estimates of the explanatory variables. Since these two variables are estimated from the data and not given ex-ante, the standard errors of all the estimates need to be obtained using bootstrapping. The output of the `copulaCorrection()` function, in this case, estimated on the `dataCopCont` dataset with only 50 bootstraps is as follows:

```
R> data("dataCopCont", package = "REndo")
R> set.seed(1002)
R> resultsCC1 <- copulaCorrection(formula = y ~ X1 + X2 + P | continuous(P),
+    data = dataCopCont, num.boots = 50, verbose = FALSE)
```

```
R> summary(resultsCC1)

Call:
copulaCorrection(formula = y ~ X1 + X2 + P | continuous(P),
    data = dataCopCont, num.boots = 50, verbose = FALSE)
Coefficients:
                  Point        Boots    Lower Boots     Upper Boots
                  Estimate     SE       CI (95%)        CI (95%)
(Intercept)        2.0360      0.0384    1.9571          2.1065
X1                 1.4920      0.0081    1.4790          1.5068
X2                -3.0024      0.0099   -3.0209         -2.9832
P                 -1.0122      0.0237   -1.0669         -0.9671
Number of bootstraps: 50
Further parameters estimated during model fitting:
   rho  sigma
0.3522   0.9855
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: 3345.747
AIC: -6679.495 , BIC: -6644.551
KKT1: TRUE  KKT2: TRUE Optimx Convergence Code: 0
```

The `summary()` function returns the estimates of the model parameters, together with the bootstrapped standard errors and the upper and lower bounds of the 95% bootstrapped confidence interval. Next, the estimates of the correlation between the error and the endogenous regressor, `rho` (here 0.352), and of the standard deviation of the structural error, `sigma` (here 0.985) are reported. The value of `rho` confirms the endogeneity of P, while the sign of `rho` indicates the direction of the correlation, positive or negative (here positive). If initial parameter values are not supplied by the user, the parameter estimates obtained running OLS are used as input, while for `rho` and `sigma` the default initial values are set to 0 and 1. Next, the log-likelihood value is returned (here 3345.747) alongside the Akaike and Bayesian information criteria (here AIC : -6679.495 and BIC : -6644.551 respectively) which can be used for model comparison. Last, the convergence code returned by the optimization routine is reported, which can be `0` if the estimation converged without error, or another code, different from zero, if the maximum iteration number has been reached or another error occurred (see `optimx()` for more details).

As seen in Figure 5, copula correction produces unbiased estimates for the continuous endogenous regressor. In the example above, the coefficient of the endogenous regressor P is -1.01 (SE $= 0.023$), a value very close to the true value, -1. In comparison, the OLS coefficient estimate for P is -0.821 (SE $= 0.011$).

Case 2: Single or more discrete endogenous regressor

In the case of one or more discrete endogenous regressors, an alternative model specification also based on Gaussian copula is implemented. In this scenario, the suitable method is ordinary least squares augmented with P^*. In order to be able to identify the model's coefficients, the discrete endogenous variables cannot have a binomial distribution. The syntax

of the `copulaCorrection()` in this case is: `copulaCorrection(y ~ X1 + X2 + P1 + P2 | discrete(P1) + discrete(P2), data, num.boots, verbose)`, where the first argument is a two-part formula, with the second part of the RHS specifying the endogenous regressors, `discrete(P1)` and `discrete(P2)`, the second argument is the name of the dataset, while the third argument lets the user specify the number of bootstraps to be used.
The output of the function, estimated on the `dataCopDis` dataset which has only one endogenous discrete variable, is provided below:

```
R> data("dataCopDis", package = "REndo")
R> set.seed(1003)
R> resultsCC2 <- copulaCorrection(formula = y ~ X1 + X2 + P | discrete(P),
+    data = dataCopDis)
R> summary(resultsCC2)

Call:
copulaCorrection(formula = y ~ X1 + X2 + P | discrete(P), data = dataCopDis)
Coefficients:
                Point       Boots     Lower Boots  Upper Boots
                Estimates   SE        CI (95%)     CI (95%)
(Intercept)     1.8666     0.2506        1.0444      2.007
X1              1.5216     0.0256        1.4738      1.5757
X2             -3.0147     0.0249       -3.0671     -2.9679
P              -0.9750     0.0491       -1.0044     -0.8108
PStar.P         0.2851     0.1104       -0.0948      0.3497
Number of bootstraps: 1000
Discrete endogenous variables:  P
```

We see that copula correction method produces an estimate for the coefficient of P equal to -0.975 (SE $= 0.05$), while the OLS coefficient for P is -0.852 (SE $= 0.008$). The additional coefficient estimate, `PStar.P`, tells the direction of the correlation and whether endogeneity exists or not. In the discrete case, the variable P^* lies between two points of the inverse univariate normal distribution (see Park and Gupta (2012), page 573). Therefore, the coefficients of `PStar.P` will slightly vary at each run of the model (for reproducible results, use a random seed). However, the bootstrapped confidence intervals reported by the summary function give the upper and lower bounds of the 95% bootstrapped interval, and thus a way to assess the variation in the parameter estimates. In this case, the estimate of the endogenous regressor, P, varies between -0.843 and -1.019. Here, the coefficient of `PStar.P` is positive, implying a positive correlation between P and the error. According to the upper and lower bounds of the bootstrapped confidence interval, the value of `PStar.P` can be in more than 5% of the cases negative as well, thus not statistically significant.

Case 3: Two or more, continuous or discrete, endogenous regressors

In the case of two or more endogenous regressors, either discrete or continuous, one should use the `copulaCorrection` function with the following syntax: `copulaCorrection(y ~ X1 + X2 + P1 + P2 | discrete(P1) + continuous(P2), data, num.boots, verbose)`, where the first argument is a two-part formula, with the second part of the RHS specifying the

endogenous regressors and their distribution type, `discrete(P1)` and `continuous(P2)`. The second argument is the name of the data, while the user can also specify, using the `num.boots` argument, the number of replications to be done. The output of the function estimated on the `dataCopDisCont` dataset is:

```
R> data("dataCopDisCont", package = "REndo")
R> set.seed(1004)
R> resultsCC3 <- copulaCorrection(formula = y ~ X1 + X2 + P1 + P2 |
+    discrete(P1) + continuous(P2), data = dataCopDisCont)
R> summary(resultsCC3)


Call:
copulaCorrection(formula = y ~ X1 + X2 + P1 + P2 | discrete(P1) +
    continuous(P2), data = dataCopDisCont, verbose = FALSE)


Coefficients:
                  Point        Boots         Lower Boots     Upper Boots
                  Estimate     SE            CI (95%)        CI (95%)
(Intercept)      1.9081        0.1397         1.4853         2.0270
X1               1.5119        0.02476          1.4641         1.5614
X2              -3.0007        0.0233        -3.0460        -2.9578
P1              -0.9799        0.0452        -1.0193        -0.8435
P2               0.7617        0.0290         0.7043         0.8294
PStar.P2         0.2363        0.0526         0.1179         0.3377
PStar.P1         0.2643        0.0802        -0.01862         0.3326
Number of bootstraps: 1000
Continuous endogenous variables:  P2
Discrete endogenous variables  :  P1
```

In the `dataCopDisCont` dataset, there are two endogenous regressors, one discrete, $P1$, and one continuous, $P2$. The additional coefficients, `PStar.P1` and `PStar.P2`, which are the estimates of P_1^* and P_2^*, tell us whether there exists endogeneity or not. Also, they tell the direction of the correlation between the endogenous regressors and the error.

In the example above, the copula correction method returned a coefficient estimate for $P1$ equal to -0.979 (SE $= 0.045$), while the OLS coefficient estimate is -0.826 (SE $= 0.011$). `PStar.P1` is positive, implying that $P1$ is positively correlated with the error. For $P2$, the method returned a `PStar.P2` which is statistically significant and positive. The true parameter value for $P2$ is 0.8. The OLS estimation returned a coefficient equal to 0.884 (SE $= 0.011$) while copula correction returned an estimate equal to 0.761 (SE $= 0.029$).

5.3. Higher moments method

The `higherMomentsIV()` function has a four-part formula, with the following specification: `higherMomentsIV(y ~ X1 + X2 + P | P | IIV (iiv = gp , g = x2, X1, X2) + IIV (iiv = yp) | Z1, data)`, where `y` is the response; the first RHS of the formula, `X1 + X2 + P`, is the model to be estimated; the second part, `P`, specifies the endogenous regressors;

the third part, `IIV()`, specifies the format of the internal instruments; and the fourth part, `Z1`, is optional, allowing the user to add any external instruments available.

The special function `IIV` in the third part of the formula has a set of three arguments: `iiv` specifies the form of the instrument, `g` specifies the transformation to be applied on the exogenous regressors, and the last argument is the set of exogenous variables from which the internal instruments should be built (they can be multiple). Six different instruments can be constructed which should be specified in the `iiv` argument of `IIV`:

- `g` for $(G_t - \bar{G})$,
- `gp` for $(G_t - \bar{G})(P_t - \bar{P})$,
- `gy` for $(G_t - \bar{G})(Y_t - \bar{Y})$,
- `yp` for $(Y_t - \bar{Y})(P_t - \bar{P})$,
- `p2` for $(P_t - \bar{P})^2$,
- `y2` for $(Y_t - \bar{Y})^2$.

where $G_t = G(X_t)$ is implemented in **REndo** only for the following non-linear elementwise functions of the vector X_t: X^2, X^3, $\ln(X)$ and $1/X$. This can be specified in the parameter `g` of the third RHS of the formula, as `x2, x3, lnx` or `1/x`. In the case of internal instruments built only from the endogenous regressor, e.g., `p2`, or from the response and the endogenous regressor, like for example in `yp`, there is no need to specify the `g` or the set of exogenous regressors in the `IIV` part of the formula. The function additionally reports a set of three diagnostic tests (as returned by the `AER::ivreg()` function) to assess the validity of the instruments and the endogeneity assumption. Lewbel (1997)'s higher moments approach to endogeneity is illustrated on the synthetic dataset `dataHigherMoments`. In the example below the internal instrument was chosen to be *yp*:

```
R> data("dataHigherMoments", package = "REndo")
R> resultsHM <- higherMomentsIV(
+    formula = y ~ X1 + X2 + P | P | IIV(iiv = yp),
+    data = dataHigherMoments)
R> summary(resultsHM)

Call:
higherMomentsIV(formula = y ~ X1 + X2 + P | P | IIV (iiv = yp),
    data = dataHigherMoments )
Residuals:
     Min       1Q   Median       3Q      Max
-6.56248 -1.17414  0.02446  1.20624  5.91371
Coefficients:
               Estimate    Std.Error  t value Pr(>|t|)
(Intercept)       1.6194       0.7094    2.283   0.0225 *
X1                1.5861       0.3609    4.395 1.16e-05 ***
X2                3.0618       0.0783   39.118  < 2e-16 ***
P                -1.0138       0.0819  -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
```

For this dataset, the OLS coefficient of the endogenous regressor is -0.854 with a standard error of 0.005. The higher moments estimate, -1.013 is very close to the true value, with a standard error of 0.081. In the `Diagnostic tests` section, the three diagnostic tests returned by the `ivreg()` function are printed. The first test displayed is the partial F test of the first stage regression for weak instruments (Staiger and Stock 1997). The test rejects the null hypothesis of weak instruments (p value of 0.00020). The second test is the Wu-Hausman endogeneity test proposed by Wu (1973) and Hausman (1978). The test confirms that there is an endogeneity problem, rejecting the null hypothesis at 5% confidence level. The third diagnostic test is the Sargan test of overidentifying restrictions (Sargan 1958). Because the model is just identified, the Sargarn test does not report any results. Interpreting the parameter estimates has to be made with caution since the higher moments approach comes with a set of very strict assumptions (see Section 4.3). And, as seen in the comparison with the OLS in the previous section, the higher moments estimator displays high variance.

5.4. Heteroscedastic errors method

The `hetErrorsIV()` function specifies the model in a four-part formula:

```
hetErrorsIV(y ~ X1 + X2 + X3 + P | P | IIV(X1, X2) | Z1, data)
```

where `y` is the response variable, `X1 + X2 + X3 + P`, represents the model to be estimated, the second part, `P`, specifies the endogenous regressor; the third part, `IIV(X1, X2)`, specifies the exogenous heteroscedastic variables from which the instruments are derived, and the fourth part, `Z1`, is optional, allowing the user to include additional external instrumental variables. Like in the higher moments approach, allowing the inclusion of additional external variables is a convenient feature of the function, since it increases the efficiency of the estimates. To show the heteroscedastic errors approach, the `dataHetIV` was simulated, where only `X2` satisfies the heteroscedasticity assumption, and thus used as instrument. The use of the function is presented below:

```
R> data("dataHetIV", package = "REndo")
R> resultsHetIV <- hetErrorsIV(y ~ X1 + X2 + P | P | IIV(X2),
+    data = dataHetIV)
R> summary(resultsHetIV)

Call:
hetErrorsIV(formula = y ~ X1 + X2 + P | P | IIV(X2), data = dataHetIV)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-4.83374 -0.95558 -0.03449  0.97774  5.53488
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)   2.1613      0.1241   17.41   <2e-16 ***
X1            1.3609      0.1001   13.60   <2e-16 ***
X2            2.8289      0.1516   18.66   <2e-16 ***
P            -1.0081      0.0386  -26.10   <2e-16 ***
Diagnostic tests:
                  df1   df2    statistic  p-value
Weak instruments    1   2496   6447.04    < 2e-16 ***
Wu-Hausman          1   2495     56.51   7.76e-14 ***
Sargan              0     NA        NA         NA
Residual standard error: 4.85 on 2496 degrees of freedom
Multiple R-Squared: 0.2492, Adjusted R-squared: 0.2483
Wald test: 282.7 on 3 and 2496 DF,  p-value: < 2.2e-16
```

For this example, the OLS coefficient is -0.857 with a 0.033 standard error, while the heterogeneous errors approach produces an estimate for P equal to -1.008 and a standard error equal to 0.038. The result returned by the internal instrumental method is closer to the true parameter value of the endogenous regressor. The F test of the first stage regression for weak instruments signals that the instruments are good. The Wu-Hausman test rejects the null hypothesis, indicating that there is an endogeneity problem, while the Sargan test does not return any value since the model is just identified. In the background, the function performs the Koenker's studentized version of the Breusch-Pagan (BP, Breusch and Pagan 1979) test against heteroscedasticity (i.e., the `bptest()` from the **lmtest** package is applied). This is done to check the assumption of non-zero covariance between the exogenous regressors used for building the internal instruments and the squared residuals from regressing `P` on all the exogenous regressors. In case the test fails to reject the null hypothesis at a 5% significance level, a warning for a possible weak instrument is printed indicating also the p value of the BP test.

5.5. Multilevel internal instrumental variable method

The `multilevelIV()` function allows the estimation of a multilevel model with up to three levels, in the presence of endogeneity. Specifically, Kim and Frees (2007) designed an approach that controls for endogeneity at higher levels in the data hierarchy, e.g., for a three-level model, endogeneity can be present either at level two, level three, or at both, level two and three. The function takes as input the formula, the set of endogenous variables and the name of the dataset. It returns the parameter estimates obtained with fixed effects, random effects and the GMM estimator proposed by Kim and Frees (2007), such that a comparison across models can be done. Asymptotically, the multilevel GMM estimators share the same properties of corresponding fixed effects estimators, but they allow the estimation of all the variables in the model, unlike the fixed effects counterpart. To illustrate the use of the function, `dataMultilevelIV` has a total of 2950 observations clustered into 1370 observations at level 2, which are further clustered into 40 level-three units. Additional details are presented

Name	Variables	Type	True param. values
dataMultilevelIV	level-1: $X11$, $X12$, $X13$, $X14$	exogenous	$\beta_{11} = 3$, $\beta_{12} = 9$, $\beta_{13} = -2$, $\beta_{14} = 2$
	level-1: $X15$	endogenous corr$(X15, \epsilon_{cs}) = 0.7$;	$\beta_{15} = -1$
	level-2: $X21$, $X22$, $X23$, $X24$	exogenous	$\beta_{21} = -1.5$, $\beta_{22} = -4$, $\beta_{23} = -3$, $\beta_{24} = 6$
	level-3: $X31$, $X32$, $X33$	exogenous	$\beta_{31} = 0.5$, $\beta_{32} = 0.1$, $\beta_{33} = -0.5$

Table 4: Simulated dataset to exemplify `multilevelIV()` function. ϵ_{cs} is the level-two error.

in Table 4.

The dataset has five level-one regressors, `X11, X12, X13, X14`, and `X15`, where `X15` is correlated with the level two error, thus endogenous. There are four level-two regressors, `X21, X22, X23`, and `X24`, and three level-three regressors, `X31, X32, X33`, all exogenous. We estimate a three-level model with `X15` assumed endogenous. Having a three-level hierarchy, `multilevelIV()` returns five estimators, from the most robust to omitted variables (`FE_L2`), to the most efficient (`REF`), i.e., lowest mean squared error:

- level-2 fixed effects (`FE_L2`);
- level-2 multilevel GMM (`GMM_L2`);
- level-3 fixed effects (`FE_L3`);
- level-3 multilevel GMM (`GMM_L3`);
- random effects estimator (`REF`).

The random-effects estimator is efficient assuming no omitted variables, whereas the fixed effects estimator is unbiased and asymptotically normal even in the presence of omitted variables. Because of the efficiency, one would choose the random effects estimator if confident that no important variables were omitted. On the contrary, the robust estimator would be preferable if there was a concern that important variables were likely to be omitted. The estimation result is presented below:

```
R> set.seed(1005)
R> data("dataMultilevelIV", package = "REndo")
R> formula1 <- y ~ X11 + X12 + X13 + X14 + X15 + X21 + X22 + X23 + X24 +
+    X31 + X32 + X33 +  (1 | CID) + (1 | SID) | endo(X15)
R> resultsMIV <- multilevelIV(formula = formula1, data = dataMultilevelIV)
R> coef(resultsMIV)

             REF      FE_L2   FE_L3  GMM_L2  GMM_L3
(Intercept) 64.364   0.000   0.000  64.664  64.364
X11          3.036   3.048   3.035   3.036   3.036
```

```
X12          9.000   8.997   9.000   8.997   9.000
X13         -2.008  -2.000  -2.009  -2.022  -2.008
X14          1.981   2.002   1.980   1.985   1.981
X15         -0.574  -1.037  -0.575  -1.034  -0.574
X21         -2.242   0.000  -2.232  -2.217  -2.242
X22         -3.256   0.000  -2.935  -3.315  -3.266
X23         -2.833   0.000  -2.806  -2.858  -2.833
X24          5.070   0.000   5.090   5.018   5.070
X31          2.077   0.000   0.000   2.071   2.077
X32          0.454   0.000   0.000   0.457   0.454
X33          0.099   0.000   0.000   0.098   0.099
```

The `coef()` function was used here, unlike in the other methods, since it provides an overview of the estimates across all estimated models. The `summary()` function returns, for each model, the estimates together with the standard errors and z scores.

As we have simulated the data, we know that the true parameter value of the endogenous regressor (`X15`) is -1. Looking at the coefficients of `X15` returned by the five models, we see that they form two clusters: one cluster is composed of the level-two fixed effects estimator and the level-two GMM estimator (both return -1.03), while the other cluster is composed of the other three estimators, `FE_L3, GMM_L3, REF`, all three having a value of -0.57. The bias of the last three estimators is to be expected since we have simulated the data such that `X15` is correlated with the level-two error, to which only `FE_L2` and `GMM_L2` are robust. After the initial estimation of all applicable methods, the most appropriate estimator has to be identified in the next steps.

To provide guidance for selecting the appropriate estimator, `multilevelIV()` performs an omitted variable test. The results are returned by the `summary()` function. For example, in a three-level setting, different estimator comparisons are possible:

1. Fixed effects versus random effects estimators: To test for omitted level-two and level-three effects, simultaneously, one compares `FE_L2` to `REF`. The test does not indicate the level at which omitted variables might exist.

2. Fixed effects versus GMM estimators: Once it was established that there exist omitted effects but not certain at which level (see 1), we test for level-two omitted effects by comparing `FE_L2` versus `GMM_L3`. A rejection of the null hypothesis will imply omitted variables at level-two. The same is accomplished by testing `FE_L2` versus `GMM_L2`, since the latter is consistent only if there are no omitted effects at level-two.

3. Fixed effects versus fixed effects estimators: We can test for omitted level-two effects, while allowing for omitted level-three effects. This can be done by comparing `FE_L2` versus `FE_L3`, since `FE_L2` is robust against both level-two and level-three omitted effects while `FE_L3` is only robust to level-three omitted variables.

In general, in testing for higher level endogeneity in multilevel settings, one would start by looking at the results of the omitted variable test comparing `REF` and `FE_L2`. If the null hypothesis is rejected, the model suffers from omitted variables, either at level two or level three. Next, test whether there are level-two omitted effects, since testing for omitted level-three effects relies on the assumption there are no level-two omitted effects. To this end, rely

on one of the following model comparisons: FE_L2 versus FE_L3 or FE_L2 versus GMM_L2. If no omitted variables at level-two are found, proceed with testing for omitted level-three effects by comparing FE_L3 versus GMM_L3 or GMM_L2 versus GMM_L3.

The summary() function that returns the results of the omitted variable test takes two arguments: the fitted model object (here resultsMIV) and the name of the estimation method (here REF). Without a second argument, summary() displays by default the random effects coefficients. The second parameter, model, can take the following values, depending on the model estimated (two or three levels): REF, GMM_L2, GMM_L3, FE_L2, FE_L3. It returns the estimated coefficients under the model specified in the second argument, together with their standard errors and z scores. Further, it returns the chi-squared statistic, degrees of freedom, and p value of the omitted variable test between the focal model (here REF) and all the other possible options (here FE_L3, GMM_L2, and GMM_L3).

```
R> summary(resultsMIV, model = "REF")

Call:
multilevelIV(formula = formula1, data = dataMultilevelIV)

Number of levels: 3
Number of observations: 2767
Number of groups: L2(CID): 1347  L3(SID): 40

Coefficients for model REF:
            Estimate Std. Error z-score Pr(>|z|)
(Intercept)   64.364      6.459   9.964   <2e-16 ***
X11            3.035      0.027 109.863   <2e-16 ***
X12            9.000      0.026 345.152   <2e-16 ***
X13           -2.008      0.025 -79.668   <2e-16 ***
X14            1.982      0.026  75.079   <2e-16 ***
X15           -0.574      0.019 -28.987   <2e-16 ***
X21           -2.242      0.186 -12.016   <2e-16 ***
X22           -3.265      0.387  -8.438   <2e-16 ***
X23           -2.833      0.103 -27.427   <2e-16 ***
X24            5.069      0.073  69.240   <2e-16 ***
X31            2.077      0.089  23.246   <2e-16 ***
X32            0.454      0.191   2.375   0.0175 *
X33            0.099      0.041   2.388   0.0169 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .
Omitted variable tests for model REF:
              df    Chisq   p-value
GMM_L2_vs_REF  7    30.19 8.77e-05 ***
GMM_L3_vs_REF 13 -1886.54   1.00000
FE_L2_vs_REF  13     40.0  0.000138 ***
FE_L3_vs_REF  13     39.9  0.000139 ***
```

In the example above, we compare the random effects (REF) with all the other estimators.

Testing `REF`, the most efficient estimator, against the level-two fixed effects estimator, `FE_L2`, which is the most robust estimator, we are actually testing simultaneously for level-2 and level-3 omitted effects. Since the null hypothesis is rejected with a p value of `0.000138`, the test indicates severe bias in the random effects estimator. In order to test for level-two omitted effects regardless of the presence of level-three omitted effects, we have to compare the two fixed effects estimators, `FE_L2` versus `FE_L3`:

```
R> summary(resultsMIV, model = "FE_L2")

Call:
multilevelIV(formula = formula1, data = dataMultilevelIV)

Number of levels: 3
Number of observations: 2767
Number of groups: L2(CID): 1347  L3(SID): 40

Coefficients for model FE_L2:
             Estimate Std. Error z-score Pr(>|z|)
Intercept       0.000  1.438e-18    0.00        1
X11             3.048  3.193e-02   95.47   <2e-16 ***
X12             8.997  3.377e-02  266.43   <2e-16 ***
X13            -2.000  3.211e-02  -62.29   <2e-16 ***
X14             2.002  3.437e-02   58.24   <2e-16 ***
X15            -1.037  3.301e-02  -31.41   <2e-16 ***
X21                    0.000   8.540e-19    0.00        1
X22            0.000   9.154e-19    0.00        1
X23            0.000   1.727e-18    0.00        1
X24            0.000   2.951e-18    0.00        1
X31            0.000   1.427e-17    0.00        1
X32            0.000   1.436e-17    0.00        1
X33            0.000   5.946e-17    0.00        1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .
Omitted variable tests for model FE_L2:
               df Chisq  p-value
FE_L2_vs_REF   13 40.00 0.000138 ***
FE_L2_vs_FE_L3  9 38.95 1.18e-05 ***
FE_L2_vs_GMM_L2 12 40.00 7.20e-05 ***
FE_L2_vs_GMM_L3 13 40.00 0.000138 ***
```

The null hypothesis of no omitted level-two effects is rejected (p value is equal to `1.18e-05`). Therefore, it is possible to conclude that there are omitted effects at level-two. This finding is no surprise as we simulated the dataset with the level-two error correlated with `X15`, which leads to biased `FE_L3` coefficients. Hence, the result of the omitted variable test between level-two fixed effects and level-two generalized method of moments should not come as a surprise: The null hypothesis of no omitted level-two effects is rejected (p value is 0). In case of wrongly assuming that an endogenous variable is exogenous, the random effects as

well as the generalized method of moments estimators will be biased, since the former will be constructed using the wrong set of internal instrumental variables. To conclude this example, the test results provide support that the `FE_L2` should be used.

6. Case study: Explaining academic performance

In this section, we apply the instrument-free methods implemented in the **REndo** package to the California Test Score data (`CASchools`) that comes with the **AER** package. Besides providing a comparison across the implemented instrument-free methods, we also compare the results with the ones obtained using frequently used external IV methods: two-stage least squares (TSLS) and control function (Ctrl Function).

The data contain information on test performance, school characteristics, and student demographic backgrounds for schools in different districts in California. The data are aggregated at the district level, across different California counties. In trying to answer the question of how does student/teacher ratio affect the average reading score, we use as covariates the following variables: student/teacher ratio (students/teachers), lunch (percent qualifying for reduced-price lunch), english (percent of English learners), calworks (percent qualifying for income assistance), income (district average income in USD 1000), grades (a dummy variable if the grade is equal to KK-8) and county (dummy for county).

In this model, the student/teacher ratio might be endogenous since it could be correlated with unobserved factors such as teacher salaries or teacher working conditions, both unobserved in the data, that can affect the reading score of the students. However, we have an additional variable, expenditure, representing the expenditure per student (aggregated at district level). This variable can be used as an external instrumental variable since it is correlated with the student/teacher ratio (a correlation of -0.61), but does not directly explain the reading score tests of the students (Hanushek 1997). Therefore, we can apply both external and internal instrumental variable techniques to estimate the model and compare their performance.

Both, the two-stage least squares and the control function approach return an estimate of the student/teacher ratio of around -1, very different from the OLS estimate of -0.30. The instrument-free methods, as underlined in the previous section, come with very many assumptions and are not as efficient. Therefore, we observe a large variation in the estimates returned. The higher moments and the copula correction approaches return estimates closer to the ones returned by the external IV methods: -1.30 for the higher moments and -0.35 for the copula method. The latent instrumental variable method, having only one regressor, returns a value for the student/teacher ratio coefficient equal to -2.27, while the heteroscedastic errors method fails, returning a positive estimate equal to 0.71. (see Table 5 for a comparison of results across the different methods).

The `CASchools` dataset has information at the district level, where the districts are clustered into counties. One could be tempted to apply the multilevel generalized method of moments method to these data, as implemented in the `multilevelIV()` function. However, the endogeneity problem solved by the multilevel GMM approach considers only correlations between level-one variables and level-two errors, while the endogeneity presented in the example above deals with endogeneity between a level-one variable and the level-one error. Therefore, we expect that the `multilevelIV()` function will indicate the use of the fixed effects method. In other words, the results should be similar to the ones returned by OLS since we applied

Dependent variable: *read*
Model: *read = stratio + english + lunch + calworks + income + gr08 + county_dummy*

	OLS	TSLS	Ctrl Function	copulaCor.	latentIV	hetErrorsIV	higherMomentsIV	multilevelIV
stratio	−0.300	−1.136*	−1.047***	−0.357	−2.273	0.714	−1.307	−0.300
	(0.257)	(0.535)	(0.103)	(0.092)	(13.678)	(1.310)	(2.730)	(0.261)
english	−0.205***	−0.213***	−0.036**	−0.217		−0.195***	−0.215***	−0.205***
	(0.037)	(0.038)	(0.015)	(0.0373)		(0.040)	(0.047)	(0.038)
lunch	−0.386***	−0.393***	−0.010	−0.356***		−0.378***	−0.395***	−0.386***
	(0.037)	(0.037)	(0.016)	(0.0380)		(0.039)	(0.044)	(0.037)
calworks	−0.052	−0.049	0.034	−0.069		−0.056	−0.048	−0.052
	(0.061)	(0.062)	(0.024)	(0.080)		(0.063)	(0.063)	(0.062)
grades	−1.912**	−1.892*	−1.177*	−2.021		−1.937	−1.888**	−1.911
	(1.358)	(1.377)	(0.537)	(0.968)		(1.387)	(1.388)	(1.379)
income	0.716***	0.624***	0.248***	0.801***		0.826***	0.606*	0.716***
	(0.098)	(0.111)	(0.040)	(0.074)		(0.172)	(0.313)	(0.099)
Constant	683.453***	700.478***	680.205***	682.254***	699.601*	662.787***	703.956***	–
	(9.562)	(13.580)	(3.778)	(1.968)	(268.618)	(27.901)	(56.182)	–
res_cf			0.878***					
			(0.019)					
Obs.	420	420	420	420	420	420	420	420
R^2	0.88	0.87	0.85	–	–	0.87	0.87	
Adj. R^2	0.86	0.86	0.83	–	–	0.85	0.85	

Table 5: IIV model comparison for the California Test Score data (p value levels: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; Further notes: [1] see Section 5.1, [2] see Section 5.2, [3] see Section 5.4, [4] see Section 5.3, [5] see Section 5.5).

county dummy variables. Indeed, the omitted variable test between the fixed effects and the random effects model rejects the null hypothesis of no omitted variables (p value = 8.52e-7), therefore, indicating an endogeneity problem and the use of fixed effects.
Before concluding, a word of caution has to be given when applying instrument-free methods: First, one has to be careful with the assumptions of each method and second, the efficiency of these methods improves with an increased sample size. The code for reproducing the results in this section is presented in Appendix D. The same model was estimated using the `ivreg2h()` function in Stata. Given that the Stata implementation of Lewbel (2012) builds internal instruments from all exogenous variables, we had to remove three counties from the model indicators before being able to estimate the model. No reasonable error message was given for this. After eliminating three county dummy variables, the coefficient returned by Stata was statistically insignificant and equal to −0.31, similar to the OLS coefficient estimate. In the light of this experience, we find that giving the user the possibility of choosing the variables from which to construct the internal instruments (with warnings on whether they constitute weak instruments or not) is an important feature and thus was implemented in the **REndo** package. Estimating the model without the county dummies, we obtain the same results using **REndo**'s `hetErrorsIV()` function and Stata's `ivreg2h()` module.

7. Conclusion

Endogeneity, one of the challenges of empirical research, can be addressed using different approaches. The use of external instrumental variables is one of the most popular solutions to the endogeneity problem. However, there exist applications in which neither theory nor intuition helps in deriving an adequate external instrumental variable. Therefore, methods that treat endogeneity without the need for external regressors have been proposed. These estimation techniques are called "instrument-free" or "internal instrumental variable" (IIV) methods.

REndo implements various IIV models. Thereby, **REndo** supports researchers from a broad spectrum of disciplines such as sociology, political science, economics, and marketing, interested in the causal analysis of observational data. Moreover, we facilitate a straightforward estimation and comparison across different IIV methods.

Future developments could include a Bayesian approach to the latent instrumental variable method which would allow incorporating additional explanatory variables and even additional endogenous regressors (Ebbes *et al.* 2009). Instrument-free methods that address endogeneity with a binary or categorical dependent variable would be a very useful extension of the package's functionality. In that regard, the recent paper of Kim, Lee, Kim, and Paik (2019) that proposes an IIV method for multilevel models with a binary outcome is a possible future method to be implemented.

A recently published paper (Bun and Harrison 2018) has proposed yet another internal instrumental variable approach that helps recovering the true parameter estimate of solely the interaction term between an endogenous and an exogenous regressor. A future version of **REndo** should consider integrating this approach as well. Another useful development of the package would be the integration of diagnostic tests for the copula correction and the latent internal variable approaches. These would allow the users to assess whether endogeneity is indeed a problem and whether the internal instruments considered are good. Since in the context of these methods no adaptation of these tests has been proposed yet, we could not implement such tests in our package at this stage.

Acknowledgments

This work was supported by the University Research Priority Program Social Networks of the University of Zurich. All authors contributed equally.

References

Angrist J, Pischke JS (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. doi:10.2307/j.ctvcm4j72.

Antonakis J, Bendahan S, Jacquart P, Lalive R (2014). "Causality and Endogeneity: Problems and Solutions." In *The Oxford Handbook of Leadership and Organizations*, pp. 93–117. Oxford University Press. doi:10.1093/oxfordhb/9780199755615.013.007.

Bates D, Eddelbuettel D (2013). "Fast and Elegant Numerical Linear Algebra Using the

RcppEigen Package." *Journal of Statistical Software*, **52**(5), 1–24. doi:10.18637/jss.v052.i05.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Bates D, Mächler M, Jagan M (2023). ***Matrix**: Sparse and Dense Matrix Classes and Methods*. R package version 1.5-4, URL https://CRAN.R-project.org/package=Matrix.

Baum C (2012). "**sspecialreg**: Stata Module to Estimate Binary Choice Model with Discrete Endogenous Regressor via Special Regressor Method." Statistical Software Components, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457546.html.

Baum C, Schaffer M, Stillman S (2002). "**ivreg2**: Stata Module for Extended Instrumental Variables/2SLS and GMM Estimation." Statistical Software Components, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s425401.html.

Berry S, Levinsohn J, Pakes A (1995). "Automobile Prices in Market Equilibrium." *Econometrica*, **63**(4), 841–890. doi:10.2307/2171802.

Berry ST (1994). "Estimating Discrete-Choice Models of Product Differentiation." *The RAND Journal of Economics*, **25**(2), 242–262. doi:10.2307/2555829.

Breusch TS, Pagan AR (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation." *Econometrica*, **47**(5), 1287–1294. doi:10.2307/1911963.

Bun MJ, Harrison TD (2018). "OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms." *Econometric Reviews*, pp. 1–14. doi:10.1080/07474938.2018.1427486.

Chau TW (2015). "Identification through Heteroskedasticity: What If We Have the Wrong Form of Heteroskedasticity?" *Technical Report 70333*, Munich Personal RePEc Archive. URL https://mpra.ub.uni-muenchen.de/70333/.

Chaussé P (2010). "**gmm**: Computing Generalized Method of Moments and Generalized Empirical Likelihood with R." *Journal of Statistical Software*, **34**(11), 1–35. doi:10.18637/jss.v034.i11.

Croissant Y, Millo G (2008). "Panel Data Econometrics in R: The **plm** Package." *Journal of Statistical Software*, **27**(2), 1–43. doi:10.18637/jss.v027.i02.

Dowle M, Srinivasan A (2023). ***data.table**: Extension of* `data.frame`. R package version 1.14.8, URL https://CRAN.R-project.org/package=data.table.

Draganska M, Jain D (2004). "A Likelihood Approach to Estimating Market Equilibrium Models." *Management Science*, **50**(5), 605–616. doi:10.1287/mnsc.1040.0227.

Ebbes P, Wedel M, Boeckenholt U (2009). "Frugal IV Alternatives to Identify the Parameter for an Endogeneous Regressor." *Journal of Applied Econometrics*, **24**(3), 446–468. doi:10.1002/jae.1058.

Ebbes P, Wedel M, Boeckenholt U, Steerneman AGM (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**(4), 365–392. doi:10.1007/s11129-005-1177-6.

Eddelbuettel D, François R (2011). "**Rcpp**: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08.

Epanechnikov VA (1969). "Nonparametric Estimation of a Multidimensional Probability Density." *Theory of Probability & Its Applications*, **14**(1), 156–161. doi:10.1137/1114019.

Fernihough A (2014). ***ivlewbel**: Uses Heteroscedasticity to Estimate Mismeasured and Endogenous Regressor Models*. R package version 1.1, URL https://CRAN.R-project.org/src/contrib/Archive/ivlewbel/.

Fox J, Nie Z, Byrnes J (2022). ***sem**: Structural Equation Models*. R package version 3.1-15, URL https://CRAN.R-project.org/package=sem.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2023). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 1.2-1, URL https://CRAN.R-project.org/package=mvtnorm.

Germann F, Ebbes P, Grewal R (2015). "The Chief Marketing Officer Matters!" *Journal of Marketing*, **79**(3), 1–22. doi:10.1509/jm.14.0244.

Gui R, Meierer M, Algesheimer R, Schilter P (2023). ***REndo**: Fitting Linear Models with Endogenous Regressors Using Latent Instrumental Variables*. R package version 2.4.9, URL https://CRAN.R-project.org/package=REndo.

Hansen LP (1982). "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, **50**(4), 1029–1054. doi:10.2307/1912775.

Hanushek EA (1997). "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis*, **19**(2), 141–164. doi:10.3102/01623737019002141.

Hausman JA (1978). "Specification Tests in Econometrics." *Econometrica*, **46**(6), 1251–1271. doi:10.2307/1913827.

Henningsen A, Hamann JD (2007). "**systemfit**: A Package for Estimating Systems of Simultaneous Equations in R." *Journal of Statistical Software*, **23**(4), 1–40. doi:10.18637/jss.v023.i04.

Kim GS, Lee Y, Kim H, Paik MC (2019). "Cluster-Specific Nonignorably Missing, Endogenous, and Continuous Regressors in Multilevel Model for Binary Outcome." *Statistical Methods in Medical Research*, **29**(7). doi:10.1177/0962280219876959.

Kim S, Frees F (2007). "Multilevel Modeling with Correlated Effects." *Psychometrika*, **72**(4), 505–533. doi:10.1007/s11336-007-9008-1.

Kleiber C, Zeileis A (2022). ***AER**: Applied Econometrics with R*. R package version 1.2-10, URL https://CRAN.R-project.org/package=AER.

Lewbel A (1997). "Constructing Instruments for Regressions with Measurement Error When No Additional Data Are Available, With an Application to Patents and R&D." *Econometrica*, **65**(5), 1201–1213. doi:10.2307/2171884.

Lewbel A (2012). "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models." *Journal of Business and Economic Statistics*, **30**(1), 67–80. doi:10.1080/07350015.2012.643126.

Longford NT (1995). *Random Coefficient Models*, chapter 10, pp. 519–570. Handbook of Statistical Modeling for the Social and Behavioral Sciences. Springer-Verlag. doi:10.1007/978-1-4899-1292-3_10.

Marra G, Radice R (2022). ***GJRM**: Generalised Joint Regression Modelling*. R package version 0.2-6.1, URL https://CRAN.R-project.org/package=GJRM.

Nash J, Varadhan R (2011). "Unifying Optimization Algorithms to Aid Software System Users: **optimx** for R." *Journal of Statistical Software*, **43**(9), 1–14. doi:10.18637/jss.v043.i09.

Park S, Gupta S (2012). "Handling Endogeneous Regressors by Joint Estimation Using Copulas." *Marketing Science*, **31**(4), 567–586. doi:10.1287/mksc.1120.0718.

Petrin A, Train K (2010). "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research*, **47**(1), 3–13. doi:10.1509/jmkr.47.1.3.

Raudenbush SW, Bryk AS (1986). "Hierarchical Model for Studying School Effects." *Sociology of Education*, **59**(1), 1–17. doi:10.2307/2112482.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rigobon R (2003). "Identification through Heteroskedasticity." *Review of Economics and Statistics*, **85**(4), 777–792. doi:10.1162/003465303772815727.

Rosseel Y (2012). "**lavaan**: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36. doi:10.18637/jss.v048.i02.

Ruud PA (2000). *An Introduction to Classical Econometric Theory*. Oxford University Press, New York. doi:10.2307/1061587.

Sargan J (1958). "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica*, **26**(3), 393–415. doi:10.2307/1907619.

SAS Institute Inc (2020). *The SAS System, Version 15.2*. SAS Institute Inc., Cary. URL https://www.sas.com/.

Schafer J, Opgen-Rhein R, Zuber V, Ahdesmaki M, Silva APD, Strimmer K (2021). ***corpcor**: Efficient Estimation of Covariance and (Partial) Correlation*. R package version 1.6.10, URL https://CRAN.R-project.org/package=corpcor.

Silverman BW (1986). *Density Estimation for Statistics and Data Analysis*. CRC Monographs on Statistics and Applied Probability. Chapman & Hall, London. doi:10.1201/9781315140919.

Staiger D, Stock JH (1997). "Instrumental Variables Regression with Weak Instruments." *Econometrica*, **65**(3), 557–586. doi:10.2307/2171753.

StataCorp (2015). *Stata Base Reference Manual Release 14*. A Stata Press, College Station.

StataCorp (2019). *Stata Statistical Software: Release 16*. StataCorp LLC, College Station. URL https://www.stata.com/.

Stock JH, Yogo M (2002). "Testing for Weak Instruments in Linear IV Regression." *Technical Report 284*, National Bureau of Economic Research. doi:10.3386/t0284.

Theil H (1958). *Economic Forecasts and Policy*. Contributions to Economic Analysis No. 15. North-Holland, Amsterdam.

Wright S (1928). *The Tariff on Animal and Vegetable Oils*, chapter Appendix. The MacMillan Company.

Wu DM (1973). "Alternative Tests of Independence between Stochastic Regressors and Disturbances." *Econometrica*, **41**(4), 733–750. doi:10.2307/1914093.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. doi:10.18637/jss.v034.i01.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), 7–10. URL https://CRAN.R-project.org/doc/Rnews/.

A. Weak versus strong instrumental variables

Table 6 illustrates the performance of OLS in comparison with two-stage least squares, one of the most frequently used instrumental variable methods, distinguishing between strong and weak instruments as well as small and large sample sizes. The sample size varies between 500 and 2500 observations respectively, and the correlation (ρ) between the instrument and the endogenous regressor takes two values, 0.10 and 0.40. The results are obtained by running a simulation over 1000 random samples, where the true parameter value is -1 and the correlation between the omitted variable and the endogenous regressor (ϕ) takes two values, 0.1 or 0.3.

ϕ	ρ	True value	OLS	IV	Sample size
0.1	0.1	-1	-1.038	-1.112	500
0.1	0.1	-1	-0.919	-1.011	2500
0.1	0.4	-1	-0.937	-0.935	500
0.1	0.4	-1	-0.918	-1.035	2500
0.3	0.1	-1	-0.769	-0.373	500
0.3	0.1	-1	-0.767	-1.125	2500
0.3	0.4	-1	-0.723	-0.952	500
0.3	0.4	-1	0.800	-1.067	2500

Table 6: Performance of OLS vs. IV. ϕ = correlation between endogenous variable and the error, ρ = correlation between the IV and the endogenous variable.

The bias of the estimates, both of the OLS and of the IV, depends heavily on the sample size but also on how much the regressor correlates with the error and with the IV. In Table 6 we can easily see that, with a relatively low sample size and a low correlation of the regressor with the error, a weak instrument produces a more biased estimate than the OLS. Once the sample size increases, even using a weak instrument, the two-stage least squares produces less biased estimates than the ordinary least squares.

B. Bias under endogeneity

The bias induced by an endogenous regressor on the estimates of the other regressors depends, besides the correlation between the two variables, on the correlation between the structural error and the endogenous variable. To illustrate this fact, Figure 9 as well as Table 7 illustrate the parameter estimates for the model presented in Figure 1. The correlation between the error and P takes three values: 0.1 0.3 and 0.5. For each of these three values, the correlation between the two covariates, P and X, is also taken to be equal to 0.1, 0.3 and 0.5. The parameter estimates are the mean over 1000 simulated samples, each having the same size (2500 observations).

Panel A in Figure 9 presents the parameter estimates for the endogenous regressor, β_1, while panel B the parameter estimates for the X covariate, β_2. In both cases the bias grows as we move from lower to higher levels of the correlation between the error and the endogenous regressor. In panel B, for the same correlation between the error and the endogenous regressor, the increase in bias is steeper than in panel A. For example, when the correlation between

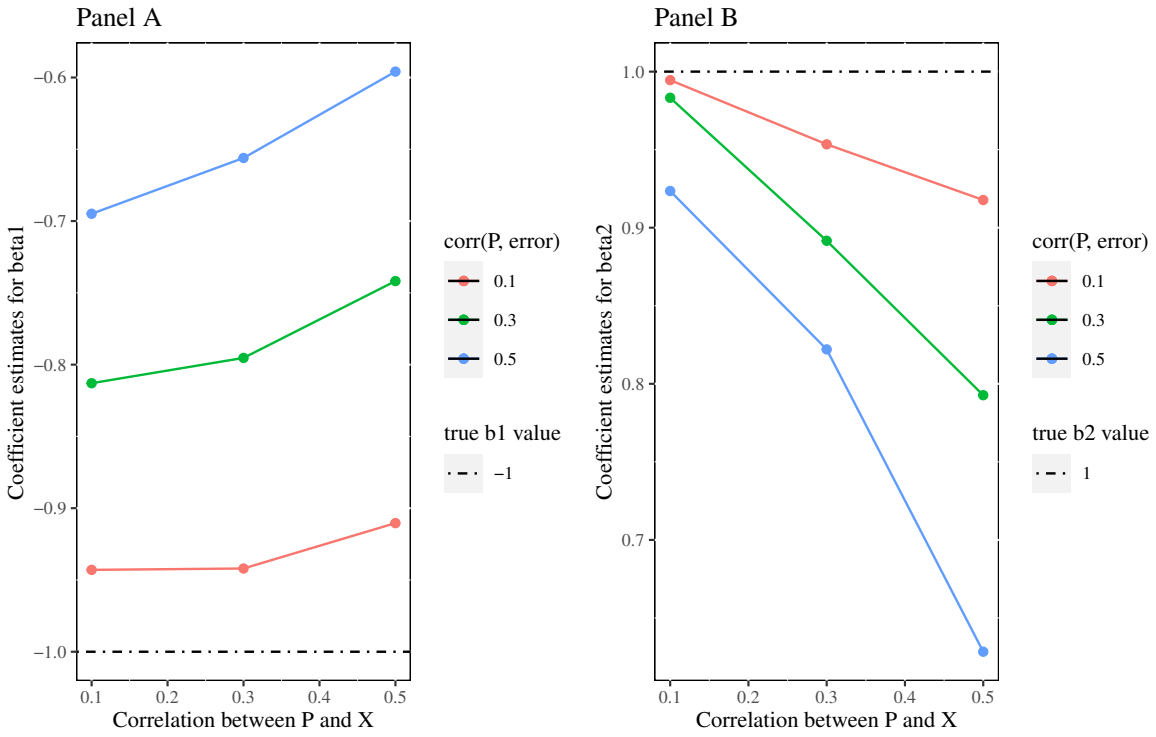


Figure 9: Varying levels of endogeneity and resulting OLS parameter estimates for β_1 and β_2. The lines illustrate the bias at different levels of correlation between the endogenous regressor (P) and the error.

	$\text{corr}(P, \epsilon)$	$\text{corr}(P, X)$		
		0.1	0.3	0.5
β_1	0.1	−0.057	−0.058	−0.090
	0.3	−0.187	−0.205	−0.258
	0.5	−0.305	−0.344	−0.404
β_2	0.1	−0.005	0.046	0.083
	0.3	0.017	0.108	0.207
	0.5	0.076	0.178	0.372

Table 7: Bias for estimates for β_1 (intercept) and β_2 (slope) depending on different levels for the correlation between the error and the endogenous regressor.

the error and P is 0.5, the bias of the covariate estimate increases from 0.07 to 0.17, to 0.37 for different levels of correlation between P and X. Similar patterns occur for different levels of correlation between the error and P, either for β_1 or β_2, as can be observed in Table 7.

C. Gaussian copula

A copula is a function that maps several conditional distribution functions (CDF) into their joint CDF. Here, the CDF of x_1 is a t distribution and the CDF of x_2 is a normal distribution. Function $H()$ is the joint conditional distribution function.

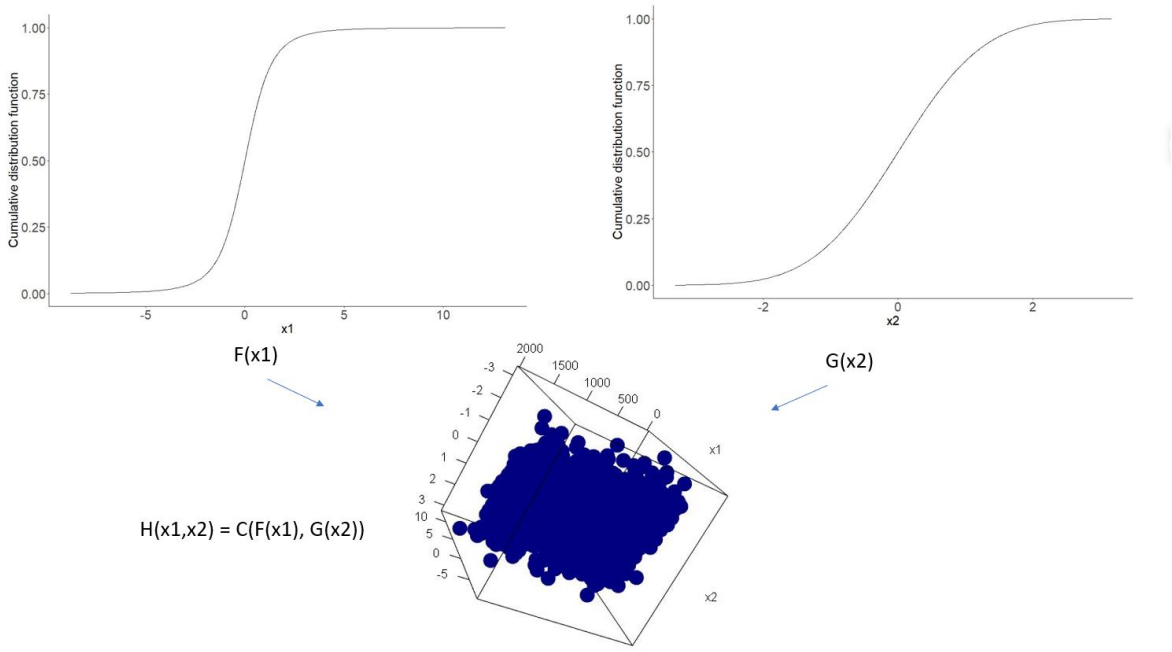


Figure 10: Copula example.

D. R Code: Application to real data

```
R> library("AER")
R> library("REndo")
R> library("sem")
R> data("CASchools", package = "AER")
R> CASchools$stratio <- with(CASchools, students/teachers)
R> school <- CASchools
R> school$gr08 <- 1
R> school$gr08[school$grades == "KK-06"] <- 0
R> school$cc <- as.numeric(school$county)
R> ols <- lm(read ~ stratio + english + lunch + grades + calworks +
+     income + county, data = school)
R> summary(ols)$coefficients[1:7,]
R> extiv <- ivreg(read ~ stratio + english + lunch + grades + income +
R> calworks + county, ~ expenditure + english + lunch + income +
R> grades + county + calworks, data = school)
R> summary(extiv)$coefficients[1:7,]
```

```
R> res_cf <- lm(read ~ expenditure, data = school)$residuals
R> m_cf <- lm(read ~ stratio + res_cf + lunch + english + calworks +
+    income + grades + county , data = school)
R> summary(m_cf)$coefficients[1:7,]
R> set.seed(43223)
R> cop.model <- copulaCorrection(read ~ stratio + english + lunch +
+    calworks + grades + income + county | continuous(stratio),
+    num.boots = 50, data = school, verbose = FALSE)
R> summary(cop.model)$coefficients[1:7,]
R> liv <- latentIV(read ~ stratio, data = school)
R> summary(liv)
R> hetEr <- hetErrorsIV(read ~ stratio + english + lunch + calworks +
+    income +  grades + county | stratio | IIV(income, english),
+    data = school)
R> summary(hetEr)$coefficients[1:7,]
R> highMoment <- higherMomentsIV(read ~ stratio + english + lunch +
+    calworks + income + grades + county | stratio |
+    IIV(g = x3, iiv = gp, income), data = school)
R> summary(highMoment)$coefficients[1:7,]
R> multilev <- multilevelIV(read ~ stratio + english + lunch + gr08 +
+    income + calworks + (1|cc) | endo(stratio), data = school)
R> coef(multilev)
```

Affiliation:

Raluca Gui
Department of Business Administration and URPP Social Networks
University of Zurich
8050, Zurich, Switzerland
E-mail: raluca.gui@business.uzh.ch
URL: https://www.socialnetworks.uzh.ch/

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
September 2023, Volume 107, Issue 3 *Submitted:* 2019-06-25
doi:10.18637/jss.v107.i03 *Accepted:* 2023-02-22