REndo: A Package to Address Endogeneity Without External Instrumental Variables

Raluca Gui

2020-05-17

1. What does REndo do

REndo is the first **R** package to implement the most recent *internal instrumental variable* methods to address endogeneity. 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 heteroskedastic error approach (Lewbel, 2012). To model hierarchical data (not cross-classified) such as students nested within classrooms, nested within schools, **REndo** includes the multilevel GMM estimation proposed by Kim and Frees (2007). All approaches assume a **continuous dependent variable**.

Internal instrumental variable approaches, also called **instrument free methods**, have been proposed as alternative to external instrumental variable approaches (like IV regression) to address endogeneity concerns, when valid, strong instruments are difficult to find.

The only alternative to **REndo** we could find in **R** is the **ivlewbel** package that implements the heteroskedastic errors method proposed by Lewbel (2012).

2. Short Description of REndo's Functions

2.1 Instrument Free Methods for Non-hierarchical Data

The four instrument free methods presented in this section share the same underlying model presented in equations (1) and (2) below. The specific characteristics of each method are discussed in the subsequent sections.

Consider the model:

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 X_t + \epsilon_t \quad (1) \tag{1}$$

where $t = 1, .., T$ indexes either time or cross-sectional units, Y_t is a k x 1 response variable, X_t is a k x n exogenous regressor, P_t is a k x 1 continuous *endogenous* regressor, ϵ_t is a structural error term with mean zero and $E(\epsilon^2) = \sigma_\epsilon^2$, α and β are model parameters. The endogeneity problem arises from the correlation of P_t and ϵ_t. As such:

$$P_t = \gamma Z_t + \nu_t \quad (2) \tag{2}$$

where Z_t is a l x 1 vector of internal instrumental variables, and ν_t is a random error with mean zero, $E(\nu^2) = \sigma_\nu^2$ and $E(\epsilon\nu) = \sigma_{\epsilon\nu}$. 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**. Therefore, Z_t and ν_t cannot be identified without distributional assumptions.

The distributions of Z_t and ν_t should be specified such that two conditions are met: **(1)** endogeneity of P_t is corrected, and **(2)** the distribution of P_t is empirically close to the integral that expresses the amount of overlap of Z as it is shifted over ν (= the convolution between Z_t and ν_t). When the density $h(\cdot)$ is chosen to be normal, then G cannot be normal because the parameters would not be identified (Ebbes et al., 2005). Consequently, in the LIV model the distribution of Z_t is discrete while in the higher moments and joint estimation with copulas methods, the distribution of the internal instruments is taken to be skewed.

Latent Instrumental Variable Approach Ebbes et al. (2005) propose the latent instrumental variables approach whose model is described in equations (1) and (2) above. 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.

Identification of the parameters relies on the distributional assumptions of the latent instruments, Z_t, as well as that of the endogenous regressor, P_t. Specifically:

- P_t should have a non-Gaussian distribution.
- Z_t should be discrete and have at least two groups with different means.

A continuous distribution for the instruments leads to an unidentified model, while a normal distribution of the endogenous regressor gives rise to inefficient estimates.

Gaussian Copula Correction Approach 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).

The underlying idea 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.

The method allows both continuous and discrete endogenous regressors. In the case of **one continuous endogenous regressor**, the model is estimated using maximum likelihood. Otherwise, an alternative approach, still based on Gaussian copulas, but using an augmented OLS estimation is being used. The assumption of a skewed endogenous regressor is maintained here as well for the recovery of the correct parameter estimates.

The structural error ϵ_t is assumed to have a normal marginal distribution. The marginal distribution of the endogenous regressor P_t is obtained using the Epanechnikov kernel density estimator, as below:

$$\hat{h}(p) = \frac{1}{T \cdot b} \sum_{t=1}^{T} K\left(\frac{p - P_t}{b}\right) \tag{3}$$

where P_t is the endogenous regressor, $K(x) = 0.75 \cdot (1 - x^2) \cdot I(\|x\| <= 1)$ and the bandwidth b is equal to $b = 0.9 \cdot T^{-1/5} \cdot min(s, IQR/1.34)$, as proposed by Silvermann (1969). 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. In both cases, augmented OLS and maximum likelihood, the inference procedure occurs in two stages (first the empirical distribution of the endogenous regressor is computed and then used in constructing the likelihood function), the standard errors are not correct. Therefore, in both cases, the standard errors and the confidence intervals are obtained based on the sampling distributions resulted from bootstrapping. Since the distribution of the bootstraped parameters is highly skewed, we report the percentile confidence intervals. The variance-covariance matrix is also computed based on the bootstraped parameters, and not based on the Hessian.

In both cases, maximum likelihood estimation and augmented OLS, the reported standard errors are the bootstrapped standard errors, due to the inference being done in two steps. The confidence intervals are also the bootstrapped confidence intervals, due to the non-normality of the bootstrapped parameters.

Higher Moments Approach The higher moments approach proposed by Lewbel (1997) helps identify structural parameters in regression models with endogeneity caused by *measurement error*. Identification is achieved by exploiting third moments of the data, with no restrictions imposed on the distribution of the error terms.

The following instruments are constructed and can be used with two-stage least squares estimation to obtain consistent estimates:

$$\begin{aligned} q_{1t} &= (G_t - \bar{G}) && (3a) \\ q_{2t} &= (G_t - \bar{G})(P_t - \bar{P}) && (3b) \\ q_{3t} &= (G_t - \bar{G})(Y_t - \bar{Y}) && (3c) \\ q_{4t} &= (Y_t - \bar{Y})(P_t - \bar{P}) && (3d) \\ q_{5t} &= (P_t - \bar{P})^2 && (3e) \\ q_{6t} &= (Y_t - \bar{Y})^2 && (3f) \end{aligned} \tag{4}$$

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

The instruments in equations (3e) and (3f) 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 $G(X) = X$ are included as instruments, $G(X)$ should not be linear in X in equation (3a) above.

Since the constructed instruments come along with very strong assumptions, one of their best uses is to provide over-identifying information. The over-identification can be used to test validity of a potential outside instrument, to increase efficiency, and to check for robustness of parameter estimates based on alternative identifying assumptions (Lewbel 1997).

Heteroskedastic Errors Approach The heteroskedastic errors method identifies structural parameters in regression models with endogenous regressors by means of variables that are uncorrelated with the product of heteroskedastic errors. The instruments are constructed as simple functions of the model's 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 (Lewbel, 2012).

Consider the model in equations (1) and (2). This approach assumes that:

- $E(X\epsilon) = 0$
- $E(X\nu) = 0$
- $cov(Z, \epsilon\nu) = 0$.
- the errors, ϵ and ν, may be correlated with each other.

Structural parameters are identified by an ordinary two stage least squares regression of Y on X and P, using X and $[Z - E(Z)]\nu$ as instruments. A vital assumption for identification is that $cov(Z, \nu^2) \neq 0$.

The strength of the instrument is proportional to the covariance between $(Z - \bar{Z})\nu$ and ν, which corresponds to the degree of heteroskedasticity of ν with respect to Z (Lewbel, 2012). This assumption can be empirically tested. If it is zero or close to zero, the instrument is weak, producing imprecise estimates, with large standard errors. Under homoskedasticity, the parameters of the model are unidentified. But, identification is achieved in the presence of heteroskedasticity related to at least some elements of X.

2.2 Instrument Free Methods for Hierarchical Data (not cross-classified)

Like in single-level regression, also in multilevel models endogeneity is 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 of the coefficients and of the variance components.

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 of 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.

The model comes with a set of assumptions such as:

- the errors at each level are normally distributed and independent of each other.
- the slope variables are exogenous.
- the level-1 structural error is uncorrelated with any of the regressors.

If the last assumption is not met, additional, external instruments are necessary.

Consider a hierarchical model with three levels like below:

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

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

- errors at the higher two 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,
- an intermediate case, where there is concern with the higher level errors, but there is not enough information to estimate level 3 parameters.

The ingenious approach proposed by Kim and Frees (2007) lies in the fact that when all variables are assumed exogenous, the proposed estimator equals the random effects estimator. When all covariates are assumed endogenous, it equals the fixed effects estimator.

In facilitating the choice of the estimator to be used for the given data, Kim and Frees (2007) also propose an omitted variable test (which is reported by the summary function after the estimation using multilevelIV() function in **REndo**). 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.

3. 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:

1. **latentIV()** - implements the latent instrumental variable estimation as in Ebbes (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. The **latent instrumental variable** function has the following syntax:

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

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 and the last one, **start.params=c()**, which is optional, is a vector with the initial parameter values. When not indicated, the initial parameter values are taken to be the coefficients returned by the OLS estimator of **y** on **P**.

2. **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 estimation. In the case of a discrete endogenous regressor, or when several endogenous regressors are suspected, the estimation is carried out using an augmented OLS estimation which is nonetheless based on Gaussian copulas.

The **copula correction** function has the following syntax:

```
copulaCorrection( y ~ X1 + X2 + P1 + P2 | continuous(P1) + discrete(P2),
data, start.params=c(), num.boots=10, optimx.args=list())
```

The first argument is a two-part formula of the model to be estimated, with the second part of the RHS defining the endogenous regressor, here **continuous(P1) + discrete(P2)**. The second argument is the name of the data, the third argument of the function, **start.params**, is optional and represents the initial parameter values supplied by the user (when missing, the OLS estimates are considered); the fourth argument, **num.boots**, also optional, is the number of bootstraps to be performed (the default is 1000). The fifth argument,**optimx.args**, is used in order to choose the optimization algorithm and the maximum number of iterations for the selected algorithm. The default is the Nelder-Mead algorithm with 100.000 iterations. Transformation of explanatory variables, such as I(X), ln(X) are supported. The standard errors reported are obtained through bootstrapping, sice in both cases, the inference is done in two stages. Due to the skewness of the bootstrapped parameters, the confidence intervals reported are the percentile confidence intervals. The variance-covariance matrix is also based on the boostrapped values.

3. **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.

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; the fourth part, **Z1**, is optional, allowing the user to add any external instruments available.

Regarding the third part of the formula, **IIV()**, it has a set of three arguments:

- **iiv** - specifies the form of the instrument,
- **g** - specifies the transformation to be done on the exogenous regressors,
- the set of exogenous variables from which the internal instruments should be built (any subset of the exogenous variables).

A set of six 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 $G = G(X_t)$ can be either x^2, x^3, $ln(x)$ or $1/x$ and should be specified in the **g** argument of the third RHD of the formula, as **x2, x3, lnx** or **1/x**. In 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 **g** or the set of exogenous regressors in the **IIV()** part of the formula. The function returns a set of tests for checking the validity of the instruments and the endogeneity assumption. Here as well, transformation of explanatory variables, such as I(X), ln(X), are supported.

4. **hetErrorsIV()** - uses the heteroskedasticity 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 more than one endogenous regressors.

The function **hetErrorsIV()** has a four-part formula specification:

```
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 regressors, the third part, **IIV(X1, X2)**, specifies the exogenous heteroskedastic variables from which the instruments are derived, while the final 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. Transformation of the explanatory variables, such as I(X), ln(X) are possible both in the model specification as well as in the IIV() specification.

5. **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 **multilevelIV()** function allows the estimation of a multilevel model with up to three levels, and it has a syntax in the spirit of the **lmer()** function:

```
multilevelIV(y ~ X11 + X12 + X21 + X22 + X23 + X31 + X33 + X34 +
(1|CID) + (1|SID) | endo(X12), data, lmer.control = lmerControl(list()))
```

The call has a two-part formula and an argument for data specification. In the formula, the first part is the model specification, with fixed and random parameter specification, and the second part which specifies the regressors assumed endogenous, here **X12**. The function 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. The user has the possibility to choose the optimization algorithm by specifying it in the **lmer.control** argument. The default is the Nelder Mead algorithm.

4. Examples using Real Data

Using the publicly available dataset CASchools which comes with the **AER** package, the results of implementing the instrument-free methods are presented.

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 affects 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 1.000),
- grades (a dummy variable if the grade is equal to KK-08)

- county (dummy for county).

The student/teacher ratio might be endogenous here since it could be correlated with unobserved factors such as teacher salaries or teacher working conditions, which are both unobserved, but can affect the reading score of the students. Having access to an additional variable, namely **expenditure** (the expenditure per student aggregated at district level), we can use it as external instrumental variable. This is possible 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. Therefore, we can apply both external(two-stage least squares) and internal instrumental variables techniques to estimate the model and compare their performance.

In orde to have a reference point, we apply OLS on the above data:

```
library(AER)
library(REndo)
set.seed(421)
data("CASchools")
school <- CASchools
school$stratio <- with(CASchools, students/teachers)

m1.ols <- lm(read ~ stratio +  english + lunch + grades + income + calworks + county,
             data=school)

summary(m1.ols)$coefficients[1:7,]
#>                 Estimate Std. Error     t value      Pr(>|t|)
#> (Intercept) 683.45305948 9.56214469  71.4748711 3.011667e-218
#> stratio      -0.30035544 0.25797023  -1.1643027  2.450536e-01
#> english      -0.20550107 0.03765408  -5.4576041  8.871666e-08
#> lunch        -0.38684059 0.03700982 -10.4523759  1.427370e-22
#> gradesKK-08  -1.91291321 1.35865394  -1.4079474  1.599886e-01
#> income        0.71615378 0.09832843   7.2832829  1.986712e-12
#> calworks     -0.05273312 0.06154758  -0.8567863  3.921191e-01
```

The OLS coefficient estimate for the student/teacher ratio is **-0.30**. Now, using **expenditure** as external IV, we can estimate a two-stage least squares model, using **ivreg()**:

```
m2.2sls <- ivreg(read ~ stratio +  english + lunch + grades + income + calworks +
           county| expenditure +  english + lunch + grades + income + calworks +
           county , data=school)

summary(m2.2sls)$coefficients[1:7,]
#>                 Estimate  Std. Error     t value      Pr(>|t|)
#> (Intercept) 700.47891593 13.58064436  51.5792106 8.950497e-171
#> stratio      -1.13674002  0.53533638  -2.1234126  3.438427e-02
#> english      -0.21396934  0.03847833  -5.5607753  5.162571e-08
#> lunch        -0.39384225  0.03773637 -10.4366757  1.621794e-22
#> gradesKK-08  -1.89227865  1.37791820  -1.3732881  1.704966e-01
#> income        0.62487986  0.11199008   5.5797785  4.668490e-08
#> calworks     -0.04950501  0.06244410  -0.7927892  4.284101e-01
```

The external IV method returns an estimate for the assumed endogenous regressor equal to **-1.13**, very different from the OLS estimate.

Next, we estimate the same model using the instrument-free methods from **REndo**. The **latent instrumental variables** approach will probably return a coefficient very different from the other methods, given that the only regressor allowed is the endogenous one. Let's see:

```
m3.liv <- latentIV(read ~ stratio, data=school)
#> No start parameters were given. The linear model read ~ stratio is fitted to derive them.
#> The start parameters c((Intercept)=706.449, stratio=-2.621, pi1=19.64, pi2=21.532, theta5=0.5, theta
summary(m3.liv)$coefficients[1:7,]
#>                   Estimate    Std. Error       z-score     Pr(>|z|)
#> (Intercept)   6.996014e+02  2.686186e+02  2.604441e+00 9.529597e-03
#> stratio      -2.272673e+00  1.367757e+01 -1.661605e-01 8.681108e-01
#> pi1          -4.896363e+01  5.526907e-08 -8.859139e+08 0.000000e+00
#> pi2           1.963920e+01  9.225351e-02  2.128830e+02 0.000000e+00
#> theta5       6.939432e-152 3.354672e-160  2.068587e+08 0.000000e+00
#> theta6        3.787512e+02  4.249457e+01  8.912932e+00 1.541524e-17
#> theta7       -1.227543e+00  4.885276e+01 -2.512741e-02 9.799653e-01
```

Indeed, the value returned is equal to **-2.273**. The **latentIV()** function returns, besides the coefficient estimates, also the initial parameter values used in the maximum likelihood optimization and the AIC and BIC. The latter two can also be accessed calling AIC(m3.liv) and BIC(m3.liv). The function also returns the fitted values and the residuals, as well as the confidence interval for the coefficients (the bootstrapped confidence intervals will not be reported here, since we used only 2 bootstraps, and 1000 are needed for reporting standard errors).

Next, we call the **copulaCorrection()** function:

```
set.seed(110)
m4.cc <-  copulaCorrection(read ~  stratio + english + lunch + calworks +
           grades + income + county | continuous(stratio), data= school,
           optimx.args = list(method=c("Nelder-Mead"), itnmax= 60000),
            num.boots=2, verbose = FALSE)
#> Warning: It is recommended to run 1000 or more bootstraps.

summary(m4.cc)$coefficients[1:7,]
#>              Point Estimate    Boots SE Lower Boots CI (95%)
#> (Intercept)   682.56562449 2.892541893                    NA
#> stratio        -0.40322352 0.188960137                    NA
#> english        -0.20380742 0.010970833                    NA
#> lunch          -0.36426215 0.031529433                    NA
#> calworks       -0.07186591 0.003223237                    NA
#> gradesKK-08    -0.72045639 0.195490165                    NA
#> income          0.78911122 0.040396762                    NA
#>              Upper Boots CI (95%)
#> (Intercept)                    NA
#> stratio                        NA
#> english                        NA
#> lunch                          NA
#> calworks                       NA
#> gradesKK-08                    NA
#> income                         NA
```

The copula correction with one endogenous continuous regressor, estimates the model using maximum likelihood. The optimization algorithm used is the Nelder-Mead, which it is known to converge slowly, so it might happen that sometimes your code will not converge (Converge Code = 1). Therefore, the copulaCorrection() allows the user to specify the desired optimization algorithm (see the optimx() function for a list of vailable options) and also the maximum number of iterations for the optimization algorithm.

In the current case, the algorithm converged, and we see that the coefficient of the student/teacher ratio returned is equal to **-0.38**.

The **heteroskedastic errors** approach returns an estimate of the student/teacher ratio equal to **0.71**, far away from the coefficients returned by the external instrumental variables or even OLS. As Lewbel (2012) underlined, it is often better to use this approach in order to create additional instruments, which together with external ones, could lead to improved efficiency.

```
set.seed(111)
m5.hetEr <- hetErrorsIV(read ~ stratio + english + lunch + calworks  + income  +
        grades+ county | stratio | IIV(income, english), data=school)
#> Warning: A studentized Breusch-Pagan test (stratio ~ english) indicates at a 95%
#> confidence level that the assumption of heteroscedasticity for the variable is
#> not satisfied (p-value: 0.2428). The instrument built from it therefore is weak.

summary(m5.hetEr)$coefficients[1:7,]
#>                 Estimate  Std. Error    t value     Pr(>|t|)
#> (Intercept) 662.78791557 27.90173069 23.7543657 2.380436e-76
#> stratio       0.71480686  1.31077325  0.5453322 5.858545e-01
#> english      -0.19522271  0.04057527 -4.8113717 2.188618e-06
#> lunch        -0.37834232  0.03927793 -9.6324402 9.760809e-20
#> calworks     -0.05665126  0.06302095 -0.8989273 3.692776e-01
#> income        0.82693755  0.17236557  4.7975797 2.335271e-06
#> gradesKK-08  -1.93795843  1.38723186 -1.3969968 1.632541e-01
```

Last, but not least, **higher moments approach** returns an estimate in the range of the estimate produced by the two-stage least squares and control function methods, namely **-1.30**:

```
set.seed(112)
m6.highMoment <- higherMomentsIV(read ~ stratio + english + lunch + calworks + income +
                 grades + county| stratio | IIV(g = x3,iiv = gp, income), data=school)

summary(m6.highMoment)$coefficients[1:7,]
#>                 Estimate  Std. Error    t value     Pr(>|t|)
#> (Intercept) 703.95605932 56.18284961 12.5297322 2.974075e-30
#> stratio      -1.30755252  2.73072188 -0.4788304 6.323429e-01
#> english      -0.21569879  0.04726222 -4.5638738 6.848861e-06
#> lunch        -0.39527218  0.04409111 -8.9648953 1.576520e-17
#> calworks     -0.04884574  0.06367608 -0.7670971 4.435143e-01
#> income        0.60623924  0.31312518  1.9360923 5.361980e-02
#> gradesKK-08  -1.88806451  1.38805414 -1.3602240 1.745894e-01
```

The **CASchools** dataset has information at the district level, where the districts are clustered into counties. One could be tempted to apply the multilevel GMM 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-1 variables and level-2 errors, while the endogeneity presented in the example above deals with endogeneity between a level-1 variable and the level-1 error. Therefore, we expect that the **multilevelIIV()** function will indicate the use of *fixed effects* method. In other words, the results should be similar with the ones returned by OLS since we included county dummy variables. Indeed, the omitted variable test between the fixed effects and the GMM model rejects the null hypothesis, therefore indicating an endogeneity problem at level one and the use of fixed effects.

```
set.seed(113)
school$gr08 <- school$grades=="KK-06"
m7.multilevel <- multilevelIV(read ~ stratio + english + lunch  + income + gr08 +
          calworks + (1|county) | endo(stratio), data=school)
summary(m7.multilevel)$coefficients[1:7,]
#>                Estimate Std. Error     z-score     Pr(>|z|)
#> (Intercept) 675.8228656 5.58008680 121.1133248 0.000000e+00
```

```
#> stratio      -0.4956054 0.23922638  -2.0717005 3.829339e-02
#> english      -0.2599777 0.03413530  -7.6160948 2.614656e-14
#> lunch        -0.3692954 0.03560210 -10.3728537 3.295342e-25
#> income        0.6723141 0.08862012   7.5864728 3.287314e-14
#> gr08TRUE      2.1590333 1.28167222   1.6845440 9.207658e-02
#> calworks     -0.0570633 0.05711701  -0.9990596 3.177658e-01
```

However, we can use the simulated data that comes with the package in order to give an example of the workings of the **multilevelIV()** function.

The dataset has five level-1 regressors, X11, X12, X13, X14 and X15, where X15 is correlated with the level two error, thus endogenous. There are four level-2 regressors, X21, X22, X23 and X24, and three level-3 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. The random effects estimator (REF) is efficient assuming no omitted variables, whereas the fixed effects estimator (FE) 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 below:

```
data(dataMultilevelIV)
set.seed(114)
formula1 <- y ~ X11 + X12 + X13 + X14 + X15 + X21 + X22 + X23 + X24 +
X31 + X32 + X33 +  (1 | CID) + (1 | SID) | endo(X15)
m8.multilevel <- multilevelIV(formula = formula1, data = dataMultilevelIV)
coef(m8.multilevel)
#>                    REF      FE_L2      FE_L3     GMM_L2     GMM_L3
#> (Intercept) 64.3168856  0.0000000  0.0000000 64.3485944 64.3168868
#> X11          3.0213405  3.0459605  3.0214255  3.0146686  3.0213403
#> X12          8.9522160  8.9839088  8.9524723  8.9747533  8.9522169
#> X13         -2.0194178 -2.0145054 -2.0193321 -2.0021426 -2.0194171
#> X14          1.9651420  1.9791437  1.9648317  1.9658681  1.9651421
#> X15         -0.5647915 -0.9777361 -0.5647621 -0.9750309 -0.5648070
#> X21         -2.3316225  0.0000000 -2.2845297 -2.3052516 -2.3316215
#> X22         -3.9564944  0.0000000 -3.9553644 -4.0130975 -3.9564966
#> X23         -2.9779887  0.0000000 -2.9756848 -2.9488487 -2.9779876
#> X24          4.9078293  0.0000000  4.9084694  4.7933756  4.9078250
#> X31          2.1142348  0.0000000  0.0000000  2.1164477  2.1142349
#> X32          0.3934770  0.0000000  0.0000000  0.3799626  0.3934764
#> X33          0.1082086  0.0000000  0.0000000  0.1108386  0.1082087
```

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 -0.975), while the other cluster is composed of the other three estimators, FE_L3, GMM_L3, REF, all three having a value of -0.564. 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.

To provide guidance for selecting the appropriate estimator, **multilevelIV()** function 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:

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

- **Fixed effects versus GMM estimators:** Once the existence of omitted effects is established 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.
- **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_L3since 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, 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 if rejected, this means 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, one can 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.

In order to have a quick overview of the coefficients returned by each of the possible estimation approaches (fixed effects, GMM, random effects), one should use the **coef()** function, with the name of the estimated model as parameter (here m8.multilevel). For a detailed summary of each estimated model, the **summary()** function should be used, which takes two arguments: the name of the model object (here m8.multilevel) and the estimation method (here REF). The second parameter 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).

```
summary(m8.multilevel, "REF")
#>
#> Call:
#> multilevelIV(formula = formula1, data = dataMultilevelIV)
#>
#> Number of levels: 3
#> Number of observations: 2824
#> Number of groups: L2(CID): 1368  L3(SID): 40
#>
#> Coefficients for model REF:
#>             Estimate Std. Error z-score Pr(>|z|)
#> (Intercept) 64.31689    7.87332   8.169 3.11e-16 ***
#> X11          3.02134    0.02576 117.306  < 2e-16 ***
#> X12          8.95222    0.02572 348.131  < 2e-16 ***
#> X13         -2.01942    0.02409 -83.835  < 2e-16 ***
#> X14          1.96514    0.02521  77.937  < 2e-16 ***
#> X15         -0.56479    0.01950 -28.962  < 2e-16 ***
#> X21         -2.33162    0.16228 -14.368  < 2e-16 ***
#> X22         -3.95649    0.13119 -30.160  < 2e-16 ***
#> X23         -2.97799    0.06611 -45.044  < 2e-16 ***
#> X24          4.90783    0.19796  24.792  < 2e-16 ***
#> X31          2.11423    0.10433  20.264  < 2e-16 ***
#> X32          0.39348    0.30426   1.293   0.1959
#> X33          0.10821    0.05236   2.067   0.0388 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#>
#> Omitted variable tests for model REF:
#>               df     Chisq  p-value
#> GMM_L2_vs_REF  7     18.74 0.009040 **
#> GMM_L3_vs_REF 13 -12872.98 1.000000
#> FE_L2_vs_REF  13     39.99 0.000139 ***
#> FE_L3_vs_REF  13     39.99 0.000138 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 0.000139, 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:

```
summary(m8.multilevel,"FE_L2")
#>
#> Call:
#> multilevelIV(formula = formula1, data = dataMultilevelIV)
#>
#> Number of levels: 3
#> Number of observations: 2824
#> Number of groups: L2(CID): 1368  L3(SID): 40
#>
#> Coefficients for model FE_L2:
#>               Estimate Std. Error z-score Pr(>|z|)
#> (Intercept)  0.000e+00  4.275e-19    0.00        1
#> X11          3.046e+00  2.978e-02  102.30   <2e-16 ***
#> X12          8.984e+00  3.360e-02  267.41   <2e-16 ***
#> X13         -2.015e+00  3.107e-02  -64.83   <2e-16 ***
#> X14          1.979e+00  3.203e-02   61.80   <2e-16 ***
#> X15         -9.777e-01  3.364e-02  -29.06   <2e-16 ***
#> X21          0.000e+00  1.824e-18    0.00        1
#> X22          0.000e+00  1.303e-18    0.00        1
#> X23          0.000e+00  4.389e-18    0.00        1
#> X24          0.000e+00  1.724e-18    0.00        1
#> X31          0.000e+00  1.468e-17    0.00        1
#> X32          0.000e+00  8.265e-18    0.00        1
#> X33          0.000e+00  2.793e-17    0.00        1
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Omitted variable tests for model FE_L2:
#>                 df Chisq  p-value
#> FE_L2_vs_REF    13 39.99 0.000139 ***
#> FE_L2_vs_FE_L3   9 36.02 3.92e-05 ***
#> FE_L2_vs_GMM_L2 12 39.99 7.21e-05 ***
#> FE_L2_vs_GMM_L3 13 39.99 0.000139 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis of no omitted level-two effects is rejected (p-value is equal to $3.92e-05$). Therefore,

we 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. The omitted variable test between level-two fixed effects and level-two GMM should shows that 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 GMM estimators will be biased, since the former will be constructed using the wrong set of internal instrumental variables. Consequently, comparing the results of the omitted variable tests when the variable is considered endogenous versus exogenous can indicate whether the variable is indeed endogenous or not. To conclude this example, the test results provide support that the FE_L2 should be used.

References

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

Epanechnikov V (1969). "Nonparametric Estimation of a Multidimensional Probability Den- sity." Teoriya veroyatnostei i ee primeneniya, 14(1), 156–161.

Kim S, Frees F (2007). "Multilevel Modeling with Correlated Effects." Psychometrika, 72(4), 505–533.

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

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

Park S, Gupta S (2012). "Handling Endogeneous Regressors by Joint Estimation Using Cop- ulas." Marketing Science, 31(4), 567–586.