Practitioner's Corner

Andrea Geraci, Daniele Fabbri and Chiara Monfardini*

Testing Exogeneity of Multinomial Regressors in Count Data Models: Does Two-stage Residual Inclusion Work?

DOI 10.1515/jem-2014-0019
Previously published online November 8, 2016

Abstract: We study a simple exogeneity test in count data models with possibly endogenous multinomial treatment. The test is based on Two Stage Residual Inclusion (2SRI), an estimation method which has been proved to be consistent for a general class of nonlinear parametric models. Results from a broad set of simulation experiments provide novel evidence on important features of this approach. We find differences in the finite sample performance of various likelihood-based tests, analyze their robustness to misspecification arising from neglected over-dispersion or from incorrect specification of the first stage model, and uncover that standardizing the variance of the first stage residuals leads to better results. An original application to testing the endogeneity status of insurance in a model of healthcare demand corroborates our Monte Carlo findings.

Keywords: count data; exogeneity test; healthcare utilization; multinomial endogenous treatment.

1 Introduction

Instrumental variables (IV) methods are the established solution to address endogeneity of regressors in linear models. However, it is well known that IV estimators imply an efficiency loss that might be substantial with respect to OLS estimators. This explains the great attention received by the Hausman (1978) test for endogeneity. Its regression-based form is a computationally simple two-stage procedure: first stage residuals are computed from reduced form estimation and then inserted as additional regressors in the second stage equation for the outcome of interest. The method, known as two-stage residual inclusion (2SRI), tests the null hypothesis of exogeneity of a subset of regressors by way of a variable addition test, i.e. checking whether the coefficients of the first stage residuals are equal to zero in the second stage structural equation.

Accounting for endogeneity in non linear models is more challenging. Terza, Basu, and Rathouz (2008), Wooldridge (2010), and Wooldridge (2014) point out that the application of IV methods in this context is not straightforward. Plugging in the structural equation fitted values of the endogenous variables obtained from first stage regression does not generally lead to consistent estimation for the parameters of interest.[1] On the contrary, 2SRI is recognized to be a consistent procedure for many non linear parametric models. Terza, Basu,

1 The same conclusion is reached by Bhattacharya, Goldman, and McCaffrey (2006), who show by way of a Monte Carlo exercise that in a binary probit with a binary endogenous treatment the two step procedure – based either on probit or linear probability model – is not consistent.

***Corresponding author: Chiara Monfardini,** University of Bologna – Department of Economics, Piazza Scaravilli 2, Bologna 40126, Italy, Phone: 0039 051 2098148, E-mail: chiara.monfardini@unibo.it
Andrea Geraci: INET, University of Oxford, Eagle House, Walton Well Road, Oxford, UK; and ISER, University of Essex, Wivenhoe Park, Colchester, Essex, UK
Daniele Fabbri: University of Bologna, Department of Economics, Via Zamboni, Bologna, Italy; and HEDG, Department of Economics and Related Studies, University of York, Heslington, York, UK

and Rathouz (2008) prove that in a general parametric framework this procedure leads to consistent estimators and valid exogeneity tests, provided the first stage model is consistently estimated.[2] In his textbook Wooldridge (2010) points to addition of first stage estimated residuals as a valid procedure for testing exogeneity of binary regressors when the structural equation is a binary response model, a Tobit model, or an exponential regression model. Conceptual and computational simplicity makes the two step approach extremely appealing for the applied researcher aiming at testing exogeneity in non linear models. Apart from recent contributions suggesting distribution free semiparametric approaches [see Abrevaya, Hausman, and Khan (2010) and the references cited there], the most common practice to handle endogeneity in the nonlinear framework relies upon some parametric distributional assumptions on both the endogenous regressors and the outcome variable. Inference is performed through Maximum Likelihood (ML) methods. In this context, joint maximum likelihood estimation of the two parts of the model is likely to be computationally cumbersome and to require ad hoc routines, while the two-stage approach is available within most statistical/econometric packages.

We consider here the problem of endogeneity in count data models. The empirical microeconometrics literature devoted much attention to this class of non linear models. For instance they are extensively used in applied health economics to represent healthcare demand through the counts of doctor visits where endogeneity is likely to arise due to unobserved individual frailty. Deb and Trivedi (1997), Kenkel and Terza (2001), Mullahy (1997), Windmeijer and Santos Silva (1997), Fabbri and Monfardini (2009), Cheng and Vahid (2011), Bratti and Miranda (2011) are some examples dealing with endogenous binary regressors. Some attention has been recently devoted to count data models with multinomial endogenous regressors. Deb and Trivedi (2006) propose a simulation based full maximum likelihood method for single equation count data models (generalized to the case of multivariate counts by Fabbri and Monfardini 2016). Zimmer (2010) adopts instead a 2SRI procedure, following the suggestion by Terza, Basu, and Rathouz (2008).

In this paper our aim is to establish whether 2SRI represents a viable alternative to detect endogeneity in count data models with multinomial endogenous regressors. This is one of the cases in which 2SRI tests are much easier to implement than maximum likelihood approaches, that require simulation-based inference methods. We develop a broad Monte Carlo Study to assess the finite sample properties of 2SRI exogeneity tests and to compare the performance of alternative testing procedures. We represent endogeneity by alternative specific latent factors entering both the count outcome equation and the multinomial treatment model. This formalization of endogeneity is germane to Terza, Basu, and Rathouz (2008) – based on 2SRI – and Deb and Trivedi (2006) – based on FIML estimation.

Our Monte Carlo experiments produce novel evidence on the 2SRI procedure for count models with multinomial regressors, enhancing existing simulation studies conducted by Terza, Basu, and Rathouz (2008) for duration models with multinomial regressors and Staub (2009) for count models with dichotomous endogenous explanatory variable.[3] We start by evaluating the performance of different likelihood based tests, namely Wald, Lagrange Multiplier and Likelihood Ratio, under correct specification and their robustness to various forms of misspecification, ranging from neglecting over-dispersion to incorrect specification of the first stage model. Then, we compare alternative 2SRI tests, adopting alternative definitions for the first stage residuals. We find that standardizing the variance of the first stage residuals improves the finite sample performance of the tests and reduces the bias of the treatment coefficients estimators. This is a relevant aspect, since the first stage estimation involves multinomial discrete choice models, where no consensus exists on the definition of the error term (see Pagan and Vella 1989). Finally, we bring our findings to real data, and apply the 2SRI procedure to an original case study in health economics. We use data from an important French Survey to model the individual annual number of doctor visits allowing for healthcare insurance status to be endogenously determined. The results of this application are coherent with the main finding of

2 In this paper we focus on 2SRI, i.e. inclusion in the second stage equation of a linear function of the first stage residuals rather than a more flexible function (that would be considered in the so called Control Function approach).

3 Two related works are Kapetanios (2010) and Garrido et al. (2012). The first stydy analyzes through a Monte Carlo study the peformance of some new Hausman-type tests for exogeneity in nonlinear threshold model. The second consider a case study with a linear outcome (healthcare costs) and study how estimation of the effect of a binary treatment varies according to different approaches used to deal with endogeneity, including Control Function and 2SRI.

our Monte Carlo investigation and show that ignoring endogeneity of insurance or accounting for it by way of non-standardized residuals leads to misleading results.

The remainder of the paper is structured as follows. Section 2 provides a general parametric representation of endogeneity in nonlinear models. Section 3 describes the count regression with multinomial endogenous treatment. Section 4 presents the 2SRI estimator/test we study. The design of the Monte Carlo experiment is illustrated in Section 5, together with the simulation results. Section 6 is devoted to the application of the procedure to a model of healthcare demand with endogenous insurance. Section 7 concludes.

2 A parametric Representation of Endogeneity in Nonlinear Models

We consider the non linear conditional mean of the outcome y:

$$E[y_i \mid \mathbf{x}_i, \mathbf{x}_{ei}, \mathbf{q}_i] = M(\mathbf{x}_i\beta + \mathbf{x}_{ei}\beta_e + \mathbf{q}_i\lambda) \tag{1}$$

$$= M\left(\mathbf{x}_i\beta + \sum_{s=1}^{S}\gamma_s x_{eis} + \sum_{s=1}^{S}\lambda_s q_{is}\right) \tag{2}$$

where $M(\cdot)$ is a non-linear function, $\mathbf{x}_i$ is a set of K exogenous regressors, $\mathbf{x}_{ei}$ is a set of S covariates (either discrete or continuous) possibly correlated with a set of S unobservable confounders $\mathbf{q}_i$, hence endogenous. Following Terza, Basu, and Rathouz (2008), we represent endogeneity of regressors $\mathbf{x}_{ei}$ by an idiosyncratic influence of the same latent factors $\mathbf{q}_i$ on both y_i and $\mathbf{x}_{ei}$ in possibly non linear reduced form regressions:

$$x_{esi} = r_s(\mathbf{z}_i\alpha_s) + q_{si} \quad s = 1, \ldots, S \tag{3}$$

where $\mathbf{z}_i = [\mathbf{x}_i\ \mathbf{w}_i]$, and $\mathbf{w}_i$ is a set of at least S instrumental variables satisfying all the necessary assumptions.

In this setting, the hypothesis of exogeneity of regressors x_{esi}, $s = 1, \ldots, S$ can be formulated as:

$$H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_S = 0$$

Taking a fully parametric approach to inference, let the density of the outcome conditionally to endogenous regressors, exogenous covariates and latent factors be:

$$f(y_i \mid \mathbf{x}_i, \mathbf{x}_{ei}, \mathbf{q}_i)$$

and the marginal density of endogenous regressors conditionally to exogenous covariates $\mathbf{x}_i$, identifying instruments $\mathbf{w}_i$ and latent factors $\mathbf{q}_i$ be denoted as:

$$g(\mathbf{x}_{ei} \mid \mathbf{x}_i, \mathbf{w}_i, \mathbf{q}_i)$$

The two above distributions can be combined into a joint distribution of the type:

$$Pr(y_i, \mathbf{x}_{ei} \mid \mathbf{x}_i, \mathbf{w}_i, \mathbf{q}_i) = f(y_i \mid \mathbf{x}_i, \mathbf{x}_{ei}, \mathbf{q}_i) \times g(\mathbf{x}_{ei} \mid, \mathbf{x}_i, \mathbf{w}_i, \mathbf{q}_i) \tag{4}$$

Unobservability of $\mathbf{q}_i$ can be handled by way of some parametric distributional assumptions, taking them as *i.i.d* draws from density $h(\mathbf{q}_i)$. Their distribution is integrated out via simulation, obtaining the joint density of the observable variables:

$$Pr(y_i, \mathbf{x}_{ei} \mid \mathbf{x}_i, \mathbf{w}_i) = \int f(y_i, \mathbf{x}_{ei} \mid \mathbf{x}_i, \mathbf{w}_i, \mathbf{q}_i) h(\mathbf{q}_i) d\mathbf{q}_i \tag{5}$$

Estimation is performed maximizing a simulated log-likelihood function. Provided that the factor loading parameters $\lambda_1, \lambda_2, \ldots, \lambda_S$ are identified, a test for exogeneity of the vector $\mathbf{x}_{ei}$ can be carried out with the usual maximum likelihood-based tests. This is a full information maximum likelihood (FIML) procedure since all equations are jointly estimated, and it is well known to achieve asymptotic efficiency properties under correct distributional assumptions. Notice, however, that efficiency comes with an heavy computational cost.

An easier test for exogeneity of $\mathbf{x}_{ei}$ can be carried out resorting to the so called two-stage residual inclusion procedure (2SRI). Two-stage residual inclusion and two-stage prediction substitution (2SPS) are the non-linear counterparts of the linear Two-Stage Least Squares (2SLS) approach. While 2SPS substitutes the endogenous regressors in the structural equation with their consistent estimates obtained in the first stage (mimicking 2SLS in the nonlinear case), 2SRI keeps the endogenous regressors in the outcome equation and substitutes the unobservable confounders with residuals obtained from the reduced equation. Wooldridge (2010) and Terza, Basu, and Rathouz (2008) emphasize that, when the conditional expectation is nonlinear, 2SPS is generally an inconsistent procedure, while 2SRI allows one to obtain consistent estimates of the structural equation parameters.[4] In our case of full parametric assumptions the 2SRI approach involves separate maximum likelihood estimation of both the first and the second stage equations (and it corresponds to limited information maximum likelihood, cf. Wooldridge 2014).

After estimation of the reduced form equations, predictors of the endogenous regressors are obtained as:

$$\hat{x}_{esi} = r_s(\mathbf{z}_i \hat{\alpha}_s) \tag{6}$$

Residuals, which estimate q_{si}, are computed as follows:

$$\hat{q}_{si} = x_{esi} - r_s(\mathbf{z}_i \hat{\alpha}_s) \tag{7}$$

and are plugged inside the structural equation. The parameters λ_s $s = 1, \ldots, S$ are the coefficients associated to the estimated residuals $\hat{q}_{si}$, so that the exogeneity test amounts to a variable addition test in the second stage equation, which can be easily performed with likelihood-based tests.

When the functions $M(\cdot)$and $r_s(\cdot)$ in (1) and (3) are linear, 2SRI coincides with the regression-based exogeneity test proposed by Hausman (1978), so that the 2SRI procedure can be seen as an extension of the Hausman test to the non-linear framework. Notice that the non linear function $r_s(\cdot)$ allows for endogenous regressors of different nature in $\mathbf{x}_{ei}$, including multivariate or multinomial treatments. In this case (3) will describe the generation of a set of binary dummies and their relationship with the unobservable confounders. We focus hereafter on the case of a multinomial process, since there are results in the literature showing that a multinomial model is more general than a multivariate one.[5]

Testing exogeneity of a multinomial treatment with 2SRI is a more practical alternative with respect to FIML approaches. However, little is known about its properties in finite samples. Is the two step procedure reliable, so that practitioners can exploit its computational advantages? Are there ways to conduct 2SRI outperforming alternative possibilities? In the following sections we answer to these questions. Despite the fact that we analyze the specific case of a count outcome, some results of ours might also be informative for other nonlinear models with potentially endogenous multinomial regressors.

3 A Count Data Model with Endogenous Multinomial Treatment

In the data generating processes we will specify in the next section, the multinomial treatment affects the count outcome equation trough $\mathbf{d}_i$, a set of J dummies for the $J+1$ mutually exclusive alternatives in the choice set. Endogeneity is pinned down using the same alternative specific latent factors q_{ij}^*, in the treatment and the outcome equations. This formulation adapts the proposal of Deb and Trivedi (2006) to the general representation of endogeneity presented in the previous section.

Let the density for the count outcome conditionally to exogenous variables, treatment and latent factors be written as:

$$f(y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{q}_i) = f(\mu_i, \theta) \tag{8}$$

4 The proof of consistency is carried out by Terza, Basu, and Rathouz (2008) using the theory of two-stage optimization estimators, of which 2SRI can be seen as a special case.

5 Weeks and Orne (1999) show that the multivariate probit model can be obtained as a special case of the multinomial probit one, after imposing specific restrictions on the covariance pattern of the error terms.

In order to accomodate for overdispersion in the data, $f(\cdot)$ is assumed as a Negative Binomial type-2 density function with rate parameter μ. Thus, (8) can be re-written as:

$$f(y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{q}_i) = \frac{\Gamma(y_i + \psi)}{\Gamma(\psi)\Gamma(y_i + 1)} \left(\frac{\psi}{\mu_i + \psi} \right)^{\psi} \left(\frac{\mu_i}{\mu_i + \psi} \right)^{y_i} \tag{9}$$

where ψ is the overdispersion parameter, and the conditional mean for the outcome takes the usual exponential form:

$$\mu_i = E(y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{q}_i) = exp\left(\mathbf{x_i}\beta + \sum_{j=1}^{J} \gamma_j d_{ij} + \sum_{j=1}^{J} \lambda_j q_{ij}^{*} \right) \tag{10}$$

In the above equation $\mathbf{x_i}$ is a vector of exogenous observable characteristics of individual *i* which do not vary among alternatives, and β is the conformable vector of coefficients. The multinomial treatment enters the model trough $\mathbf{d_i}$, the set of *J* dummies, d_{ij} indicating the treatment alternatives. $\mathbf{q_i}$ is a vector of *J* unobservable latent factors q_{ij}^{*}, with associated factor loadings λ_j, which potentially affect both the outcome and the treatment, generating endogeneity in the outcome model.

The equation for the multinomial treatment is derived from a random utility model, according to which each individual chooses the treatment which maximizes her indirect utility. Indirect utility for individual *i* from alternative *j* can be expressed as follows:

$$V_{ij}^{*} = \mathbf{z_i}\alpha_j + q_{ij}^{*} \tag{11}$$

where $\mathbf{z}_i$ is a vector including the exogenous covariates in $\mathbf{x}_i$ in (10) plus a set of instruments, and α_j is the vector of associated parameters for the alternative *j*. Alternative *j* is chosen by individual *i* iff $V_{ij}^{*} \geq V_{ik}^{*}$, $\forall k \neq j$. The dummy d_{ij} in (10) takes value 1 if alternative *j* is chosen, 0 otherwise. Utility from alternative $j = 0$ is normalized so that $V_{i0}^{*} = 0$.

We specify q_{ij}^{*} as *i.i.d* logistic errors after normalization. This amounts to assume a Multinomial Logit (MNL) representation for the probability of the treatment, which can be written as follows:

$$Pr(d_{ij} \mid \mathbf{z}_i) = \frac{exp(\mathbf{z_i}\alpha_j)}{1 + \sum_{k=1}^{J} exp(\mathbf{z_i}\alpha_k)} \text{ for } j = 0,\ 1,\ \ldots,\ J \tag{12}$$

[Correction added after online publication 8 November 2016: For consistency, q_{ij} was changed to q_{ij}^{*} on p. 4, line 35 and p. 5, line 10. Also, the formatting of $\mathrm{q_i}$ was corrected from italics to bold in eqs. 9 and 10.]

4 The 2SRI Estimator

4.1 First Stage

In multinomial discrete choice models there is no consensus about the definition of errors and residuals. With our notation, let $\widehat{Pr}(d_{ij} \mid \mathbf{z}_i)$ be the predicted probability of choosing alternative *j*, obtained after estimation of a multinomial response model. The most obvious definition of residuals is what we name, following Cameron and Windmeijer (1996), *raw residuals* (adopted for example by Terza, Basu, and Rathouz 2008, and Staub 2009):

$$\hat{q}_{ij}^{R} = d_{ij} - \widehat{Pr}(d_{ij} \mid \mathbf{z}_i) \quad \text{for} \quad j = 0,\ 1,\ \ldots,\ J \tag{13}$$

In a multinomial logit model such as the one we consider in the first stage, it can be easily shown that the raw residuals above coincides with the *generalized residuals* of Gourieroux et al. (1987).

Alternatively, Pagan and Vella (1989) suggest a standardized version of the residuals, with unit variance, we call *standardized residuals*, also named *Pearson residuals* in the literature.

$$\hat{q}_{ij}^{S} = \widehat{Pr}(d_{ij} \mid \mathbf{z}_i)^{-1/2}(1-\widehat{Pr}(d_{ij} \mid \mathbf{z}_i))^{-1/2}(d_{ij} - \widehat{Pr}(d_{ij} \mid \mathbf{z}_i)) \quad \text{for} \quad j=0,\ 1,\ \ldots,\ J \tag{14}$$

In the absence of any guidance on the choice between these two alternatives, in our Monte Carlo study we will compare them, to find out whether standardization improves the performance of the exogeneity test.

4.2 Second Stage

Once the MNL for treatment is estimated, we have available two types of residuals for each alternative (raw, *R*, and standardized, *S*) based on expressions (13) and (14), say:

$$\hat{q}_{ij}^{r} = d_{ij} - \widehat{Pr}(d_{ij} \mid \mathbf{z}_i) \quad \text{for} \quad j=0,\ 1,\ \ldots,\ J \quad r=R,\ S$$

4.2.1 Correct Specification

Our main analysis will consider a correctly specified model, based on the conditional NB2 distribution (9). The first stage residuals are added to the structural equation for the outcome, substituting for the unobservable latent factors, so that equation (10), describing the conditional mean for *y*, can be re-written as follows

$$E(y_i \mid \mathbf{x}_i,\ \mathbf{d}_i,\ \hat{q}_i^{r}) = exp\left(\mathbf{x}_i'\beta + \sum_{j=1}^{J}\gamma_j d_{ij} + \sum_{j=1}^{J}\lambda_j \hat{q}_{ij}^{r}\right) \quad r=R,\ S \tag{15}$$

This second stage model is estimated via maximum likelihood and the exogeneity test for the hypothesis $H_0{:}\lambda_1 = \ldots = \lambda_J = 0$ is carried out through Wald, Likelihood ratio and Lagrange Multiplier tests.

4.2.2 Misspecification

After assessing the finite sample properties of the 2SRI test under correct specification, we will allow for two different kinds of misspecification: in the second stage and in the first stage model, respectively.

In the first misspecified scenario we neglect over-dispersion in the second stage count data model. The conditional density used for estimation is a Poisson distribution, which incorrectly sets $\psi = 0$ in the NB2 distribution generating the data (9), while maintaining the same formulation (10) for the conditional mean. Despite the fact that this specification does not allow for the existing overdispersion of the data, it is well known that the Poisson PML estimator is still consistent for the conditional mean of the outcome, which remains the same. Therefore, it is interesting to investigate if the Poisson 2SRI exhibits some robustness properties.

In the second misspecified scenario we look for the potential consequences of adopting a wrong functional form for the first stage discrete choice model: the treatment dummies are now generated by a multinomial probit process, while the residuals are still evaluated after estimation of a multinomial logit model.

5 The Monte Carlo Study

In order to investigate the finite sample properties of the 2SRI exogeneity tests, we run simulations under different data generating processes (dgp) described below.[6]

6 The study has been conducted using STATA 14. Programming code and user-written routines are available on request.

5.1 Experimental Design

Random utilities are computed by way of a discrete choice model; the j-th status dummy assumes value 1 if its utility has the highest value among the $J+1$ alternatives, 0 otherwise. After having generated the dummies representing the multinomial treatment, the conditional expectation of the count dependent variable, y, is obtained by random sampling from a Negative Binomial type-2 distribution. This is obtained as a Poisson-Gamma mixture with parameter $\lambda_i = \mu_i \nu_i$, where μ_i is the conditional mean of a Poisson random variable, taking the usual exponential form, and ν_i is a random draw from a Gamma distribution. The number of alternatives in the multinomial treatment model, $J+1$, is set to three: $j = 0, 1, 2$, so that only two dummies are included inside the conditional mean for the outcome.

In order to evaluate size and power properties of the exogeneity tests, we build two different dgps under endogeneity and exogeneity of the multinomial treatment. Both dgps we analyze include logistic latent factors, but they differ for the degree of overdispersion. Under dgp2 the count variable is set to be much less overdispersed – i.e. its variance is closer to its mean, compared to the count variable generated under dgp1.[7] The sample size, N, is set to 5.000 observations, which is a realistic size for application of count data models to microeconometric data. The size and power properties of the 2SRI exogeneity tests are evaluated on 5.000 replications of the test statistics.

The following table describes the distribution of the pseudo random variables and the parameter values in our experimental setting.

Description of the Experimental Design.

$d_{1i} = 1$ if $V_{1i}^* = \max(V_{0i}^*, V_{1i}^*, V_{2i}^*)$; 0 otherwise
$d_{2i} = 1$ if $V_{2i}^* = \max(V_{0i}^*, V_{1i}^*, V_{2i}^*)$; 0 otherwise
$V_{0i}^* = 0$
$V_{1i}^* = 0.025 + 0.5 obs_i + 0.5 inst1_i - 0.25 inst2_i + q_{1i}^*$
$V_{2i}^* = 0.25 + 0.1 obs_i + 0.5 inst1_i + 0.5 inst2_i + q_{2i}^*$
$inst1_i = 1[N(0, 1) < 0.5]$ – first instrument
$inst2_i = N(0, 1)$ – second instrument
q_{1i}^*, q_{2i}^* i.i.d draws from a Bivariate Logistic distribution in dgp1 and dgp2[a]
i.i.d draws from a Bivariate Normal Distribution in dgp3[b]
$f(y_i, \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{q}_i) = \frac{\Gamma(y_i + \psi)}{\Gamma(\psi)\Gamma(y_i + 1)} \left(\frac{\psi}{\mu_i + \psi} \right)^{\psi} \left(\frac{\mu_i}{\mu_i + \psi} \right)^{y_i}$
Mixing distribution $\nu_i \Gamma\left(\psi, \frac{1}{\psi}\right)$, with $\psi = 1$ in dgp1, $\psi = 3$ in dgp2
$\mu_i = \exp(k + \beta_{obs} obs_i + \gamma_{d1} d_{1i} + \gamma_{d2} d_{2i} + \lambda_1 q_{1i}^* + \lambda_2 q_{2i}^*)$
$k = 1$ in dgp1; $k = -1$ in dgp2
$\beta_{obs} = 0.5$
$\gamma_{d1} = 0.4$
$\gamma_{d2} = 0.8$
$\lambda_1 = -0.1$ if treatment is endogenous; 0 otherwise
$\lambda_2 = -0.5$ if treatment is endogenous; 0 otherwise

[a]The bivariate logistic distribution of differenced errors $(q_{1i} - q_{oi})$, $(q_{2i} - q_{oi})$ is obtained by postulating an i.i.d. type 1 extreme values distribution for the original errors (q_{oi}, q_{1i}, q_{2i}).

[b]The bivariate normal distribution of differenced errors $(q_{1i} - q_{oi})$, $(q_{2i} - q_{oi})$ is obtained by postulating an i.i.d. trivariate normal distribution with zero mean and variance covariance matrix equal to the identity matrix for the original errors (q_{oi}, q_{1i}, q_{2i}).

In Table A1 of Appendix 1 we report basic descriptives of the treatment dummy variables and of the count variables under the different dgps. Notice that the marginal probability distribution of the dummies is kept

7 We obtain these pattern by increasing the scale parameter of the mixing gamma distribution and by lowering the constant inside the conditional mean of the count, as detailed in the table describing the experimental design.

constant over the different data generation processes. dgp2 involves a much lower degree of overdispersion than in dgp1, as it can be observed comparing the variance and the mean of the two count variables. Concomitantly, under dgp2 the count variable displays the "excess of zeros" pattern wich is often encountered in applications.

5.2 Results

5.2.1 Exogeneity Test

In Table 1 we report rejection frequencies over the 5000 Monte Carlo replications of the three asymptotycally equivalent Wald, Likelihood Ratio (LR) and Lagrange Multplier (LM) tests under correct specification of the estimated model, i.e. when the estimated model is NB2. We present two versions of the Wald test, the first of which is based on a Murphy-Topel corrected Variance-Covariance Matrix. This is evaluated taking into account that the model estimated in the second stage involves generated regressors, namely the first stage estimated residuals. We have derived the correction when the first stage model is multinomial adapting the procedure suggested by Hole (2006).[8] The Table reveals a general very good performance of all 2SRI tests, in terms of both size and power for both analyzed dgps. If we compare the test performance among the two definitions of residuals we notice an improvement of power properties of Wald corrected test when switching from "raw" to "standardized" version. A very slight power gain is associated to the use of standardized residuals also for the non corrected version of the Wald test and for the LM test. The improved general performance with standardized residuals is an interesting pattern we discover here and in the following results, and to which we will devote further attention later.

Table 1: dgp1/dgp2 – NB2 Estimator: Rejection Frequencies of Exogeneity Tests.

Nom. size	dgp1				dgp2			
	Raw residuals		Standardized residuals		Raw residuals		Standardized residuals	
	Emp. size	Emp. power	Emp. size	Emp. power	Emp. size	Emp. power	Emp. size	Emp. power
Wald test (Murphy Topel correction)								
0.01	0.0070	0.9164	0.0058	0.9602	0.0082	0.9084	0.0094	0.9540
0.05	0.0404	0.9496	0.0408	0.9774	0.0490	0.9504	0.0496	0.9716
0.10	0.0922	0.9618	0.0904	0.9838	0.0912	0.9642	0.1028	0.9792
Wald test (no correction)								
0.01	0.0088	0.9952	0.0074	0.9974	0.0110	0.9862	0.0112	0.9918
0.05	0.0462	0.9976	0.0426	0.9992	0.0554	0.9926	0.0544	0.9966
0.10	0.1010	0.9980	0.0942	0.9996	0.0986	0.9950	0.1080	0.9972
Likelihood ratio test								
0.01	0.0090	0.9952	0.0074	0.9974	0.0110	0.9862	0.0118	0.9920
0.05	0.0462	0.9976	0.0434	0.9992	0.0552	0.9926	0.0554	0.9966
0.10	0.1010	0.9980	0.0956	0.9996	0.0986	0.9950	0.1072	0.9972
Lagrange multiplier test								
0.01	0.0094	0.9850	0.0086	0.9948	0.0112	0.9666	0.0124	0.9868
0.05	0.0482	0.9928	0.0474	0.9982	0.0568	0.9870	0.0568	0.9954
0.10	0.1046	0.9956	0.0968	0.9990	0.1004	0.9916	0.1102	0.9974

N. of replications of the Monte Carlo experiment (R) = 5.000; Saple size for each replication (N) = 5.000. Raw residuals and Standardized residuals are computed after estimation of the first stage equations using, respectively: $\hat{q}_{ij} = (d_{ij} - \hat{q}_{ij})$, for $j = 0, 1, 2$ and $\hat{q}_{ij} = \hat{p}_{ij}^{-1/2}(1-\hat{p}_{ij})^{-1/2}(d_{ij} - \hat{p}_{ij})$, for $j = 0, 1, 2$.

8 Terza (2016) shows how the computation can be simplified in the general case of two-stage optimization estimators. We experimented his procedure and found very similar corrected standard errors, as expected. We are grateful to an anonymous referee for pointing this new paper to us.

Notice that our approach is similar to the control function one in that the computed residuals, regardless they are variance standardized or not, are meant to be an approximation of the latent errors of the first stage multinomial process, which have type 1 extreme value distribution [cf Wooldridge (2014) and Garrido et al. 2012]. In other words, the residuals are mis-specified by construction, being evaluated as a function of the observed binary dummies. In our intention, this is mimicking the situation faced by the applied econometrician when she postulates a discrete choice model underlying the generation of the multinomial treatment. We also run a variation of this experiment, and generated the errors of the discrete choice model with the population counterpart of the raw residual definition given in Section 4.1: $q_{ij}^{R} = d_{ij} - \Pr(d_{ij} \mid z_i)$, as specified in Terza, Basu, and Rathouz (2008). The results, contained in Table A2 of Appendix 3, show that even in this setting, where the raw residuals are correctly specified by construction, the tests based on standardized version perform generally well, with empirical size close to nominal and limeted loss of empirical power with respect to the tests based on the correctly specified raw residuals.[9]

Table 2 presents the results concerning the effects on exogeneity tests of the first misspecification we study, which affects the second stage count data model. The tests are here obtained estimating a Poisson regression model while both dgps involve a NB2 process. As the count variable exhibits greater overdispersion under dgp1 than under dgp2, the Poisson estimator, which assumes equidispersion, is "more misspecified" under dgp1. In this latter scheme, the bad consequences of misspecifying overdispersion are serious for all tests but for Lagrange Multiplier. More precisely, the empirical size of the Wald test evaluated without Murphy-Topel correction, and that of the Likelihood Ratio test are dramatically affected by misspecification of overdispersion, with rejection frequencies that imply huge probabilities of first type error (reject exogeneity when this is true). Moreover, the Wald test with corrected variance is found to loose any power of spotting true endogeneity. The robustness of LM test means that the quantities involved in its computation (score function of the unrestricted model, restricted estimator) are less affected by overdispersion parameter than the quantities involved in LR (restricted and unrestricted loglikelihood functions) or non-corrected Wald tests

Table 2: dgp1/dgp2 – Poisson Estimator: Rejection Frequencies of Exogeneity Tests.

Nom. size	dgp1				dgp2			
	Raw residuals		Standardized residuals		Raw residuals		Standardized residuals	
	Emp. size	Emp. power	Emp. size	Emp. power	Emp. size	Emp. power	Emp. size	Emp. power
Wald test (Murphy Topel correction)								
0.01	0.0006	0.0000	0.0000	0.0000	0.0204	0.2842	0.0252	0.6040
0.05	0.0352	0.0000	0.0236	0.0000	0.0846	0.4620	0.0988	0.7200
0.10	0.1008	0.0000	0.1356	0.0000	0.1500	0.5564	0.1724	0.7742
Wald test (no correction)								
0.01	0.5466	0.9990	0.5572	0.9988	0.0346	0.9944	0.0324	0.9962
0.05	0.6696	0.9992	0.6820	0.9996	0.1000	0.9964	0.1114	0.9976
0.10	0.7360	0.9992	0.7448	0.9998	0.1696	0.9972	0.1816	0.9984
Likelihood ratio test								
0.01	0.5460	0.9990	0.5460	0.9990	0.0338	0.9944	0.0338	0.9944
0.05	0.6694	0.9992	0.6694	0.9992	0.1000	0.9964	0.1000	0.9964
0.10	0.7358	0.9992	0.7358	0.9992	0.1708	0.9972	0.1708	0.9972
Lagrange multiplier test								
0.01	0.0118	0.8536	0.0106	0.9264	0.0116	0.8926	0.0120	0.9508
0.05	0.0516	0.9286	0.0520	0.9724	0.0574	0.9514	0.0572	0.9794
0.10	0.1010	0.9552	0.1012	0.9826	0.1010	0.9694	0.1078	0.9874

N. of replications of the Monte Carlo experiment (R) = 5.000; Saple size for each replication (N) = 5.000. Raw residuals and Standardized residuals are computed after estimation of the first stage equations using, respectively: $\hat{q}_{ij} = (d_{ij} - \hat{q}_{ij})$, for $j = 0, 1, 2$ and $\hat{q}_{ij} = \hat{p}_{ij}^{-1/2}(1-\hat{p}_{ij})^{-1/2}(d_{ij} - \hat{p}_{ij})$, for $j = 0, 1, 2$.

9 We are grateful to an anonymous referee for suggesting us to analyze this case.

(unrestricted estimator). Under dgp2, where overdispersion is "less misspecified," the empirical sizes of the Wald and LR test are higher than nominal ones, but their magnitudes are more reasonable than their counterparts under dgp1. Remarkably, the LM test is still found to be a robust testing procedure. Our results under dgp1 and dgp2 convey two main messages for the practitioners: first, the LM test is robust to misspecification of overdispersion in the count data model; second, the LM test displays higher power when evaluated with standardized residuals rather than raw residuals, confirming the pattern already spotted for all tests under correct specification of the count data model.

Table 3 reports the results obtained under misspecification of the first stage multinomial model. Here, the treatment is generated under dgp3, corresponding to a multinomial probit model, while the residuals are evaluated after estimation of a multinomial logit model. In this experiment we mantain correct specification of the count data model to isolate the potential consequences of misspecifying the discrete choice model functional form. We find evidence of robustness of all the considered 2SRI exogeneity tests, which show empirical size very close to nominal and very high empirical power.

5.2.2 Comparing Raw and Standardized First Stage Residuals

In this section we carefully analyze the different finite sample performance of the exogeneity tests obtained including in the second stage two alternative definitions of residuals, i.e. raw versus standardized. In Figures A1 and A2 of Appendix 2 the empirical power is plotted against the nominal size for different values of the latter. The interesting part of the plot is for small values of nominal size that will be chosen in practice (usually nominal size is set below 0.10). The higher power of the test obtained standardizing the first stage residuals is a clear pattern for all test statistics under both dgps, confirming that there is a gain in using the standardized residuals.

To get some insights on the source of this gain we exploit our simulation setting to compare the generated logistic errors of the discrete choice model and their two alternative estimates represented by raw and standardized residuals, respectively. To this purpose, we perform the following elaborations. First, we regress the

Table 3: dgp3: Rejection Frequencies of Exogeneity Tests.

Nom. size	Raw residuals		Standardized residuals	
	Emp. size	Emp. power	Emp. size	Emp. power
Wald test (Murphy Topel correction)				
0.01	0.0066	0.9998	0.0086	1.0000
0.05	0.0452	1.0000	0.0492	1.0000
0.10	0.0934	1.0000	0.0984	1.0000
Wald test (no correction)				
0.01	0.0104	1.0000	0.0098	1.0000
0.05	0.0498	1.0000	0.0514	1.0000
0.10	0.0974	1.0000	0.1016	1.0000
Likelihood ratio test				
0.01	0.0104	1.0000	0.0104	1.0000
0.05	0.0498	1.0000	0.0518	1.0000
0.10	0.0972	1.0000	0.1016	1.0000
Lagrange multiplier test				
0.01	0.0088	0.9998	0.0092	1.0000
0.05	0.0516	1.0000	0.0554	1.0000
0.10	0.0994	1.0000	0.1066	1.0000

N. of replications of the Monte Carlo experiment (R) = 5.000; Saple size for each replication (N) = 5.000. Raw residuals and Standardized residuals are computed after estimation of the first stage equations using, respectively: $\hat{q}_{ij} = (d_{ij} - \hat{q}_{ij})$, for $j = 0, 1, 2$ and $\hat{q}_{ij} = \hat{p}_{ij}^{-1/2}(1-\hat{p}_{ij})^{-1/2}(d_{ij} - \hat{p}_{ij})$, for $j = 0, 1, 2$.

generated errors used to simulate the two random utilities attached to alternatives 1 and 2, named *q1* and *q2* in Figure A3 of Appendix 2, against the two alternative definitions of residuals. In the upper part of Figure A3 we plot the true latent errors versus the fitted latent errors obtained using as regressor the two alternative definitions of residuals. The slope of the estimated regression line is found to be very close to one in the standardized residual case, revealing that the latter is approximately equal to the conditional expectation of the latent error. In the bottom part of Figure A3 we plot together the densities of the true latent errors, the raw residuals and the standardized residuals. These plots reveal that standardizing the variance of the residuals allows for a better coverage of the range of possible values assumed by the latent error component. Indeed, standardization will always increase the variance, since the variance of raw residuals is equal to $p(1-p)$, and therefore lower than one. Taken together, these results reveal that standardized residuals are better predictors of the latent errors, and we inpute to this fact their better performance with respect to their raw counterpart when testing for exogeneity.

As a by-product of our analysis, we look at the performance of the endogenous dummy coefficients estimators. We find that -across all estimated models and dgps- their overall finite sample bias is always lower when estimation involves the standardized version of the first sage residuals, and this corroborates the inference gain already spotted for the exogeneity test.[10]

6 Supplemental Insurance and Healthcare Demand: New Evidence from a French Case Study

Count data models with possibly multinomial endogenous treatment arise quite frequently in the Health Economics literature. Here demand naturally comes as a count (for physician visits or hospital admissions) with polytomous health insurance status being the potentially endogenous treatment of interest. To provide a vivid example of how the two alternative 2SRI strategies perform in the applied econometrics practice we revisit and update a case study on the effect of complementary health insurance on health care utilization in France originally explored by Buchmueller et al. (2004).

6.1 Motivation and Background

In France, Sècuritè Sociale, the social security program financed out of personal income tax, covers most of individual healthcare expenditure for legal residents, i.e. about 78% as reported by Grignon and Kambia-Chopin (2009). Copayments for visits for general practitioner or specialist visits, hospital stay, or prescription drugs, are customarily levied to moderate moral hazard. To get rid of most of these copayments French citizens obtain complementary health insurance (CHI) on a voluntary basis (individually purchase) or within an employer-sponsored set-up (employer insured). CHI plans enroll 85% of the French population and fund about 13% of total health care expenditure, 9% of health care expenditure rests on individuals as out-of-pocket payments. Similarly to Buchmueller et al. (2004) our exercise aims at assessing the impact of the CHI status on physician visits' utilization in France. We differ in that we allow for heterogeneous impacts of the individually purchased and the employer provided CHI with respect to the benchmark case of no complementary coverage. Accounting for endogeneity of the trichotomous treatment and modeling the dependent variable as a count represent the biggest improvements on the previous study.

We use data from the 2006 wave of the Enquète sur la Santè et la Protection Sociale (ESPS), a national household survey conducted by IRDES. The full sample contains data for 22,725 individuals. We restrict our analysis to individuals aged between 25 and 75. After excluding those who have not completed the "health status" section of the questionnaire, have missing data on key regressors or declare their value unknown,

10 Detailed results are available in the working paper version (Geraci, Fabbri, and Monfardini 2014).

we end up with a final estimation sample of 5989 observations. Table S1 in the online Appendix presents the distributions of our utilization measure, i.e. counts for visits to any physician in the 12 months before the interview,[11] and the treatment variable, i.e. availability of a Complementary Health Insurance. On average, individuals consume 5.3 visits per year, sample variance being fivefold larger, a clear sign of overdispersion. 56% of the sample receives CHI as part of work total compensation while 33% purchases it deliberately in the market. 10% of the sample does not obtain complementary insurance either as employer provided or as individually purchased. However, more than half of these individuals (6% of our sample) is covered by CMU-C (Couverture Maladie Universelle Complementaire), a plan introduced in 2000 by the French Government to improve the non-elderly poor access to health care. CMU-C beneficiaries are asked no co-pay at the point of use. Eligible individuals are those with a household income below a given threshold (€587 per adult equivalent per month in 2005, see Grignon, Perronnin, and Lavis 2008). Upon control for income per adult equivalent we assume this complementary insurance status (being CMU-C beneficiary) as conditionally exogenous.

Descriptive statistics for the regressors we control for in our models are provided in Table S2 in the online Appendix. Their extended name is self-explanatory on their definition.

6.2 Estimation Results

The Complementary Health Insurance status may vary according to employer-based or individually purchased vis-a-vis being either covered by CMU-C or not covered. We model it as a Multinomial Logit. Following Buchmueller et al. (2004), professional occupational variables and labor market status are used as instruments – and therefore excluded from the visit equation, to avoid identification being based only on non-linearity. The usual argument here is that different employment sectors offer different opportunities to enroll into complementary health insurance schemes and also attract individuals with different degrees of risk aversion (Fabbri and Monfardini 2016).

The estimation results of the multinomial model for insurance choice are displayed in Table S3 in the online Appendix. The Pseudo R square, the percent of correctly classified observations and the Wald test statistic for the significance of the regression support the overall goodness of fit of the estimated discrete choice model. We also obtain evidence on the relevance of our instruments, strongly rejecting the null hypothesis that their coefficients are jointly null. In this setting exclusion restrictions, hence exogeneity of instruments, can be questioned. The signs of the Average Marginal Effects (AME) are coherent with theoretical predictions and previous empirical research in the field. Being unemployed and self-employed significantly reduces the probability of receiving employer CHI by 31 and 16 percentage points, respectively, while increasing the probability of individually purchase a CHI by 20 and 14.5 percentage points, respectively. Occupation categories are almost all significant determinants of employer provided CHI, being less relevant for individually purchased ones. The effect of aging on the probability of receiving employer provided CHI is positive and declining until the individuals reaches his 50 years, being negative then after. The opposite pattern is found for individually purchased CHI. Income is a good predictor for the CHI status. The higher is individual's income and the higher (lower) is the probability of receiving CHI coverage by employers (by individually purchased plans).

Coming to the count regression, we adopt the Negative Binomial type 2 (NB2) model which was the object of our Monte Carlo investigation and has been shown to provide good fit of overdispersed count data. Table 4 reports the main estimation results. The significance of the overdispersion parameter $ln(\alpha)$ displayed in the bottom part of the table supports the NB2 specification against the Poisson. The values of observed versus estimated probabilities attached to the first ten count outcomes witness the good fit of both the exogenous and endogenous versions of the model (see Table S5 of the Appendix).

We first look at the outcome of the different exogeneity tests reported in Table 5. Opposite conclusions are implied by the alternative definitions of residuals. While exogeneity of health insurance is not rejected

11 One year recall data guarantees to have enough variation in the dependent of interest. It allows to span the full range of count values from low to high users.

Table 4: Supplemental Insurance and Healthcare Demand: II Stage NB2 Regressions for the Tot. Number of Doctor Visits (GP + Specialists).

Variables	Exogenous	Raw residuals			Std. residuals		
		(NC s.e.)	(MT s.e.)	(Boot. s.e.)	(NC s.e.)	(MT s.e.)	(Boot. s.e.)
Employer insured	0.2048***	0.3243	0.3243	0.3243	0.5630***	0.5630**	0.5630***
	[0.054]	[0.342]	[0.374]	[0.349]	[0.184]	[0.224]	[0.207]
Individually purchased	0.2398***	0.3018	0.3018	0.3018	0.4952**	0.4952	0.4952*
	[0.055]	[0.425]	[0.460]	[0.438]	[0.211]	[0.322]	[0.256]
Raw Res. – Emp. Ins.		−0.1296	−0.1296	−0.1296			
		[0.337]	[0.365]	[0.343]			
Raw Res. – Ind. Purch.		−0.0694	−0.0694	−0.0694			
		[0.421]	[0.452]	[0.434]			
Std Res. – Emp. Ins.					−0.1719**	−0.1719**	−0.1719**
					[0.072]	[0.086]	[0.083]
Std Res. – Ind. Purch.					−0.1224	−0.1224	−0.1224
					[0.089]	[0.154]	[0.109]
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	1.7910***	1.7755***	1.7755***	1.7755***	1.6659***	1.6659***	1.6659***
	[0.156]	[0.417]	[0.443]	[0.434]	[0.263]	[0.366]	[0.302]
ln(α)	−0.7211***	−0.7213***	−0.7213***	−0.7213***	−0.7228***	−0.7228***	−0.7228***
	[0.027]	[0.043]	[0.050]	[0.041]	[0.043]	[0.091]	[0.041]
Observations	5989	5989	5989	5989	5989	5989	5989
Log-Pseudologlik	−15428.080	−15427.820	−15427.820	−15427.820	−15424.609	−15427.820	−15427.820
LR test (all coeff.)	2022.460	2209.310	2209.310	2209.310	2213.075	2213.075	2213.075
(*p*-Value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Dependent variable is the total number of consultations (GP and specialists) in the year before the interview. The coefficients are estimated using a 2SRI procedure. The table reports the second stage NB2 regressions. All regressions include the following set of controls: Gender dummy; Age; Age squared; Married dummy; Children dummy; Family size; Education dummies (3); HH's income quintiles dummies (4); Self-assessed health status dummies (2); Smoking habits dummies (3); Chronic condition dummy; Limitation with daily activities dummy; Exemptions dummies (2); Regime Genarale dummy; CMU complementaire dummy. Estimates of the first stage MNL model are available in the online appendix. Exclusion restrictions used in the first stage MNL: labor status dummies (2); labor sector dummies (2); type of occupation dummies (6); Joint significance of exclusion restrictions: Wald test statistic = 237.92 (0.000).

Table 5: Supplemental Insurance and Healthcare Demand: Exogeneity Tests for Insurance Dummies.

	Raw residuals			Std. residuals		
	(NC s.e.)	(MT s.e.)	(Boot. s.e.)	(NC s.e.)	(MT s.e.)	(Boot. s.e.)
Wald test	0.341	0.311	0.365	5.730*	4.082	4.471
(*p*-Value)	(0.843)	(0.856)	(0.833)	(0.057)	(0.130)	(0.107)
LR test	0.520	0.520	0.520	6.942**	6.942**	6.942**
(*p*-Value)	(0.771)	(0.771)	(0.771)	(0.031)	(0.031)	(0.031)
LM test	0.406	0.406	0.406	6.516**	6.516**	6.942**
(*p*-Value)	(0.816)	(0.816)	(0.816)	(0.038)	(0.038)	(0.031)

$H_0: \lambda_1 = \lambda_2 = 0$, where λ_1 and λ_2 are the coefficients of the residuals.

according to raw residuals inclusion, irrespective of the test approach used, the opposite conclusion is reached by 7 out of 9 test statistics evaluated with the standardized version of the residuals.[12] Since we have reasons to believe that health insurance is endogenous (on both theoretical grounds and existing empirical

12 The exceptions are represented by the Wald test evaluated with corrected Murphy Topel and boostrapped variance-covariance matrix.

evidence), we reconcile this result with the higher power displayed in the Monte Carlo study by the 2SRI tests using standardized residuals.

Our main coefficients of interests are the treatment dummies coefficients (moral hazard effects) and the residuals's coefficients (selection effect).[13] The first column of Table 4 displays estimates obtained under exogeneity of health insurance, the following three columns account for endogeneity by inclusion of raw residuals, using variance matrix non corrected, corrected and bootstrapped, respectively, while the last three columns corresponds to inclusion of standardized residuals. Under exogeneity, significant moral hazard effects arise from either employer provided or individually purchased CHI. Both types of complementary coverage are associated to a a 22–27% increase in the conditional mean number of visits with respect to the baseline case of no complementary insurance. Once we allow for endogeneity of the insurance status, results on the treatment dummies coefficients and the residuals' coefficients are largely affected by the definition of residuals adopted. If we rely upon 2SRI with raw residuals we obtain evidence of both CHI status dummies being exogenous and exerting no effect on healthcare consumption. Adopting standardized residuals, the procedure favored by our Monte Carlo Study, we find larger moral hazard effects, especially for employer provided CHI. Moreover, we find a statistically significant negative coefficient of standardized residual corresponding to employer provided CHI status. This evidence supports the view that individuals are favorably selected (see Fang, Keane, and Silverman 2008) into employer provided plans and engage in more moral hazard than expected according to observable characteristics only. On the contrary, we do not find evidence of statistically significant selection into individually purchased plans. Overall, the results of our application confirm that inference based upon raw residuals can be severely misleading on both the effects of insurance and its endogeneity status.[14]

7 Concluding Remarks

We study two-stage residuals inclusion (2SRI) approach exogeneity testing of multinomial treatment in count data models. The procedure involves estimating the residuals from a discrete choice model for the endogenous treatment status, and plugging them as additional variables in the structural count regression, where their joint significance can be tested with likelihood based inference.

The results of our Monte Carlo study show that 2SRI exogeneity tests using Wald, LR and LM approaches have good finite sample properties when the distribution of the outcome is correctly specified. In this case, all tests display proper empirical size and power. We then analyze the performance of 2SRI under misspecification of the first and second stage models and find that LM test is the only robust procedure when overdispersion in the data is ignored.

We investigate the properties of the testing procedure as for two alternative definitions of residuals: raw and standardized. We observe that the power of the test is generally higher using standardized residuals, which provide a better fit of the discrete choice model errors. Furthermore, resorting to standardized residuals leads to a smaller bias of the endogenous treatment dummies coefficients.

The patterns emerging from the Monte Carlo investigation are quite revealing when we bring the 2SRI method to real data on an important case study in health economics: the modeling of visits' count with

13 Full estimation results are contained in Table S4 of the online Appendix. Coefficients of other regressors exhibit the expected signs across all specifications. The most prominent are those related to health status: self-assessed health, suffering from chronic conditions or from some limitation in daily activities. All of them testify that worse health positively correlates with healthcare consumption. Notice, however, that these effects are likely biased by self-reporting (see Bago d'Uva et al. 2011). Consumption rises, as expected, as the individual ages. Moreover being highly educated is positively correlated with healthcare consumption, a common finding in the literature.

14 As a robustness check, we re-run the whole set of estimates dropping the regressors whose exogeneity is more questionable, such as self-assessed health and smoking habits dummies. Table S6 in the Online Appendix show that our main findings about exogeneity of health insurance are generally unaffected: exogeneity is still rejected only using standardized residuals, despite with a lower strength.

endogenous health insurance choice. In our empirical analysis, the use of raw, non-standardized residuals leads to the – most likely wrong – conclusion that health insurance choice is exogenously determined, and it has limited to no effect on healthcare consumption. On the contrary, the specification based on standardized residuals is able to detect insurance endogeneity, i.e. favorable selection into employer provided complementary insurance, and to identify positive and significant treatment effects.

Appendix 1

Table A1: Summary Statistics of Dependent Variables Generated in Monte Carlo Study.

	dgp1		dgp2		dgp3	
	Endog.	**Exog.**	**Endog.**	**Exog.**	**Endog.**	**Exog.**
	%	**%**	**%**	**%**	**%**	**%**
Multinomial treatment dummies						
$d0_i$	24.63	24.63	24.68	24.68	21.94	21.94
$d1_i$	34.18	34.18	34.38	34.38	34.12	34.12
$d2_i$	41.19	41.19	40.94	40.94	43.94	43.94
Count variable – y_i						
Mean	7.622	5.224	1.021	0.678	5.756	5.058
Variance	456.329	50.768	11.704	1.087	82.778	47.227
Value	%	%	%	%	%	%
0	22.30	20.48	55.60	58.28	21.84	20.90
1	14.72	14.86	24.04	25.80	16.26	15.22
2	10.98	11.16	10.54	9.88	10.38	11.28
3	8.16	9.00	3.80	3.72	8.80	9.60
4	6.28	7.28	2.32	1.32	5.74	7.60
5	4.84	5.60	1.20	0.64	5.32	5.64
6	4.20	4.54	0.60	0.14	4.30	4.52
7	3.66	4.18	0.40	0.12	3.60	3.32
8	2.68	3.14	0.38	0.04	3.48	3.02
9	2.36	2.78	0.30	0.04	2.26	2.64
10	1.94	2.32	0.18	0.02	2.00	2.12
>10	17.88	14.46	0.64	0	16.02	14.14

Summary statistics are computed on the 5000 observations of the first replication of the experiment.

Appendix 2

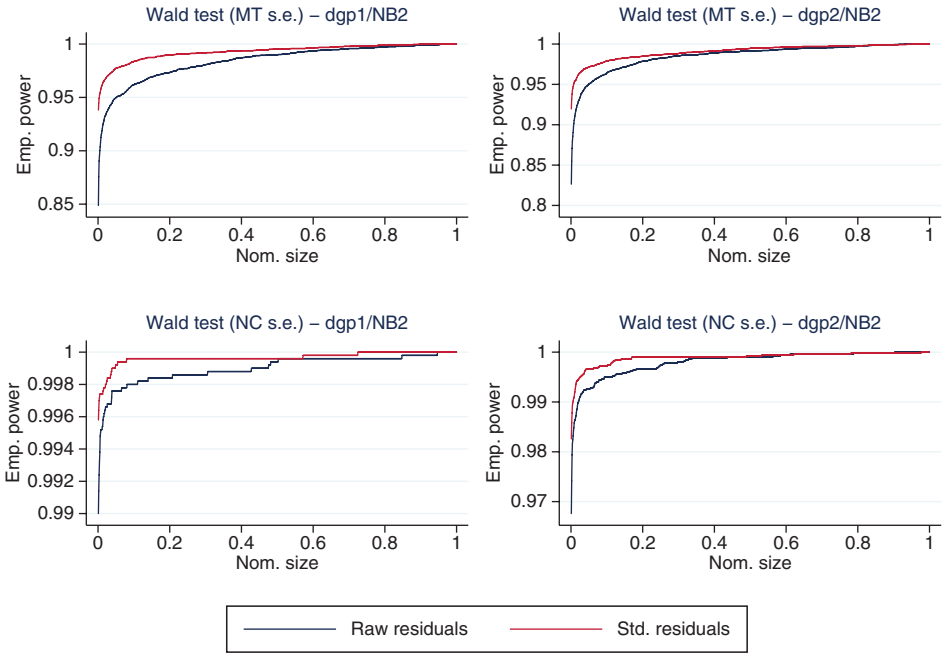

Figure A1: Empirical Power Plot of Wald Tests using Raw and Standardized Residuals.

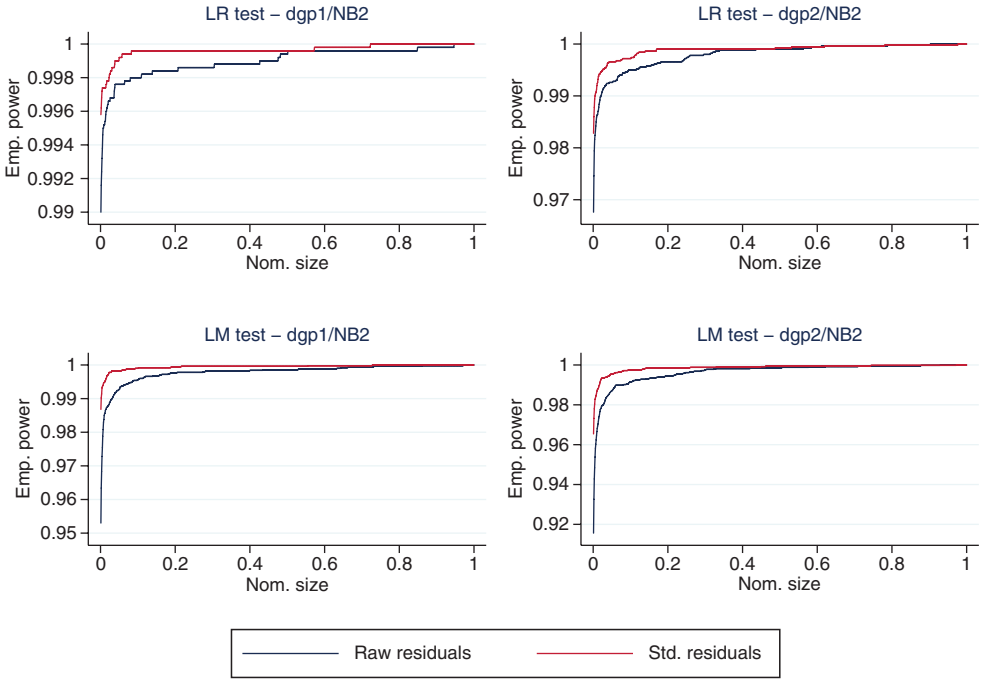

Figure A2: Empirical Power Plot of LR and LM Tests using Raw and Standardized Residuals.

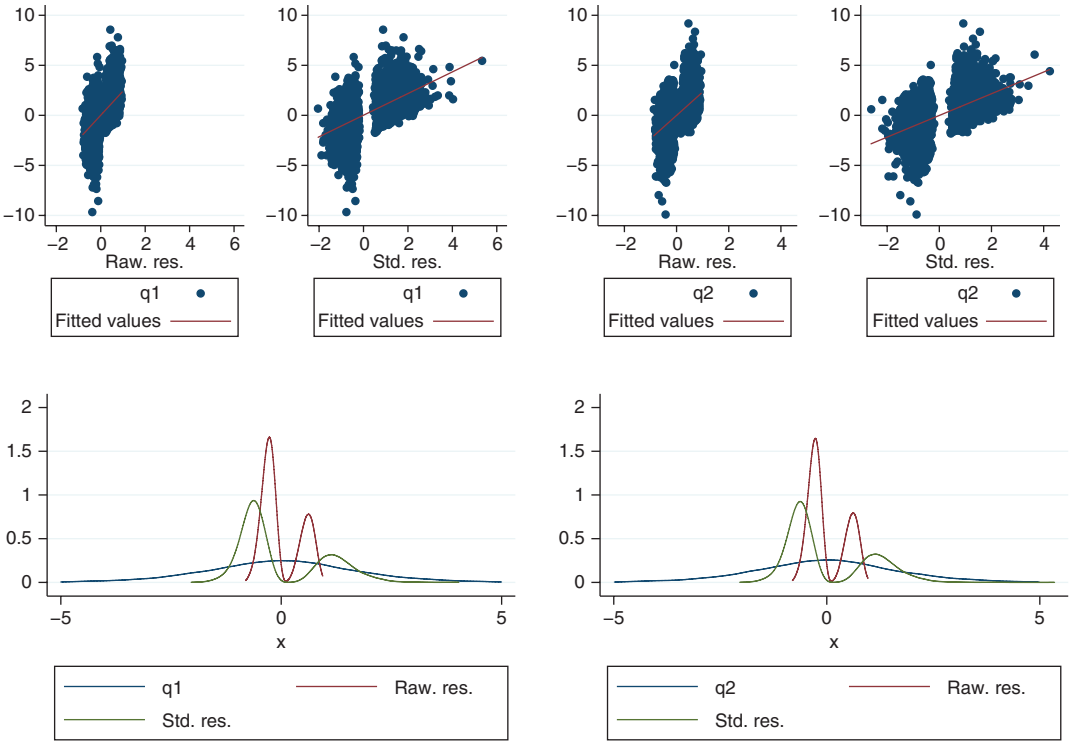

Figure A3: True Latent Factors against Estimated Residuals.

Appendix 3

Table A2: NB2 Estimator with Correctly Specified Residuals: Rejection Frequencies of Exogeneity Tests.

Nom. size	Raw residuals		Standardized residuals	
	Emp. size	Emp. power	Emp. size	Emp. power
Wald test (Murphy Topel correction)				
0.01	0.0068	0.7760	0.0084	0.7344
0.05	0.0468	0.9172	0.0468	0.8942
0.10	0.0978	0.9564	0.0922	0.9410
Wald test (no correction)				
0.01	0.0102	0.7874	0.0110	0.7476
0.05	0.0536	0.9204	0.0486	0.8980
0.10	0.1066	0.9570	0.0954	0.9432
Likelihood ratio test				
0.01	0.0104	0.7876	0.0108	0.7452
0.05	0.0532	0.9200	0.0490	0.8972
0.10	0.1062	0.9572	0.0968	0.9426
Lagrange multiplier test				
0.01	0.0110	0.7902	0.0116	0.7482
0.05	0.0542	0.9204	0.0520	0.8990
0.10	0.1084	0.9572	0.1018	0.9434

No of replications of the Monte Carlo experiment (R) = 5.000; Saple size for each replication (N) = 5.000. Raw residuals and Standardized residuals are computed after estimation of the first stage equations using, respectively: $\hat{q}_{ij} = (d_{ij} - \hat{q}_{ij})$, for $j = 0, 1, 2$ and $\hat{q}_{ij} = \hat{p}_{ij}^{-1/2}(1-\hat{p}_{ij})^{-1/2}(d_{ij} - \hat{p}_{ij})$, for $j = 0, 1, 2$.

References

Abrevaya, J., J. A. Hausman, and S. Khan. 2010. "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors." *Econometrica* 78: 2043–2061.

Bago d'Uva, T., M. Lindeboom, O. O'Donnell, and E. Van Doorslaer. 2011. "Education-related Inequity in Healthcare with Heterogeneous Reporting of Health." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174: 639–664.

Bhattacharya, J., D. Goldman, and D. McCaffrey. 2006. "Estimating Probit Models with Self-selected Treatments." *Statistics in Medicine* 25: 389–413.

Bratti, M., and A. Miranda. 2011. "Endogenous Treatment Effects for Count Data Models with Endogenous Participation or Sample Selection." *Health Economics* 20: 1090–1109.

Buchmueller, T. C., A. Couffinhal, M. Grignon, and M. Perronnin. 2004. "Access to Physician Services: Does Supplemental Insurance Matter? Evidence from France." *Health Economics* 13: 669–687.

Cameron, A. C., and F. A. Windmeijer. 1996. "R-squared Measures for Count Data Regression Models with Applications to Health-care Utilization." *Journal of Business & Economic Statistics* 14: 209–220.

Cheng, T. C., and F. Vahid. 2011. "Demand for Hospital Care and Private Health Insurance in a Mixed Public-private System: Empirical Evidence using a Simultaneous Equation Modeling Approach." Melbourne Institute Working Paper Series, No. 26. Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.

Deb, P., and P. K. Trivedi. 1997. "Demand for Medical Care by the Elderly: A Finite Mixture Approach." *Journal of Applied Econometrics* 12: 313–336.

Deb, P., and P. K. Trivedi. 2006. "Specification and Simulated Likelihood Estimation of a Non-normal Treatment-outcome Model with Selection: Application to Health Care Utilization." *The Econometrics Journal* 9: 307–331.

Fabbri, D., and C. Monfardini. 2009. "Rationing the Public Provision of Healthcare in the Presence of Private Supplements: Evidence from the Italian Nhs." *Journal of Health Economics* 28: 290–304.

Fabbri, D., and C. Monfardini. 2016. "Opt Out or Top Up? Voluntary Health Care Insurance and the Public vs. Private Substitution." *Oxford Bulletin of Economics and Statistics* 78: 75–93.

Fang, H., M. P. Keane, and D. Silverman. 2008. "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market." *Journal of Political Economy* 116: 303–350.

Garrido, M. M., P. Deb, J. F. Burgess, and J. D. Penrod. 2012. "Choosing Models for Health Care Cost Analyses: Issues of Nonlinearity and Endogeneity." *Health Services Research* 47: 2377–2397.

Geraci, A., D. Fabbri, and C. Monfardini. 2014. "Testing exogeneity of multinomial regressors in count data models: does two stage residual inclusion work?" Working Paper DSE 921, DSE – University of Bologna.

Gourieroux, C., A. Monfort, E. Renault, and A. Trognon. 1987. "Generalised Residuals." *Journal of Econometrics* 34: 5–32.

Grignon, M., and B. Kambia-Chopin. 2009. "Income and the demand for complementary health insurance in france," IRDES Document de Travail 24, Institut de recherche et documentation en économie de la santé (IRDES).

Grignon, M., M. Perronnin, and J. N. Lavis. 2008. "Does Free Complementary Health Insurance Help the Poor to Access Health Care? Evidence from France." *Health Economics* 17: 203–219.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica: Journal of the Econometric Society* 46: 1251–1271.

Hole, A. R. 2006. "Calculating Murphy-topel Variance Estimates in Stata: A Simplified Procedure." *Stata Journal* 6: 521–529.

Kapetanios, G. 2010. "Testing for Exogeneity in Threshold Models." *Econometric Theory* 26: 231–259.

Kenkel, D. S., and J. V. Terza. 2001. "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect." *Journal of Applied Econometrics* 16: 165–184.

Mullahy, J. 1997. "Instrumental-variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior." *Review of Economics and Statistics* 79: 586–593.

Pagan, A., and F. Vella. 1989. "Diagnostic Tests for Models based on Individual Data: A Survey." *Journal of Applied Econometrics* 4: S29–S59.

Staub, K. E. 2009. "Simple Tests for Exogeneity of a Binary Explanatory Variable in Count Data Regression Models." *Communications in Statistics-Simulation and Computation* 38: 1834–1855.

Terza, J. V. 2016. "Simpler Standard Errors for Two-stage Optimization Estimators Estimation in Normal Linear Models." *Stata Journal* 16: 368–385.

Terza, J. V., A. Basu, and P. J. Rathouz. 2008. "Two-stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics* 27: 531–543.

Weeks, M. and C. Orne. 1999. "The Statistical Relationship between Bivariate and Multinomial Choice Models." Cambridge Working Papers in Economics, No. 9912. Faculty of Economics, University of Cambridge.

Windmeijer, F. A. and J. M. Santos Silva. 1997. "Endogeneity in Count Data Models: An Application to Demand for Health Care." *Journal of Applied Econometrics* 12: 281–294.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Wooldridge, J. M. 2014. "Quasi-maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables." *Journal of Econometrics* 182: 226–234.
Zimmer, D. M. 2010. "Health Insurance and Health Care Demand among the Self-employed." *Journal of Labor Research* 31: 1–19.

Supplemental Material: The online version of this article (DOI: 10.1515/jem-2014-0019) offers supplementary material, available to authorized users.