CORRECTING FOR ENDOGENEITY WITHOUT INSTRUMENTS IN DISCRETE CHOICE MODELS: THE MULTIPLE INDICATOR SOLUTION

C. Angelo Guevara (caguevara@miuandes.cl) and Daniel Polanco
Universidad de los Andes, Chile

ABSTRACT

The Control Function (CF) method is the current standard for addressing endogeneity in discrete choice models. One important drawback of the CF is that it requires instrumental variables, which may be very difficult to obtain for various discrete choice applications. In turn, the Multiple Indicator Solution (MIS) method does not require instruments to correct for endogeneity, but it has only been described so far to linear models. In this article we show that MIS can be extended to discrete choice modeling under some mild assumptions and that it can straightforwardly be applied to some relevant discrete choice models in which omitted attributes are measured through indicators in RP and SP surveys. We also use Monte Carlo experiments to illustrate the efficacy and efficiency of MIS and CF methods in Logit Models, and to study the impact of the failure of their respective assumptions. Results suggest that both methods attain similar efficacy and efficiency when their respective assumptions hold, but MIS seems to be more robust to mild violations that may occur in practice.

Key-words: Endogeneity; Discrete Choice; Multiple Indicator, Control-function

April, 2013
1. Introduction

Endogeneity is a practically unavoidable problem that results in inconsistent estimators of the parameters in several types of discrete choice models. It occurs when some explanatory variables are not independent of the error term. The control-function method has been shown to be particularly suitable to address endogeneity for various types of discrete choice models (Ferreira, 2010; Petrin and Train, 2002; Guevara and Ben-Akiva, 2006, 2012). The key aspect of the control-function method is that it requires valid instruments, auxiliary variables that have to be correlated with the endogenous variable and, at the same time, uncorrelated (independent) with the error terms of the model. Finding suitable instruments fulfilling such properties can be problematic, motivating the search for alternative methods.

Wooldridge (2010) proposed the Multiple Indicator Solution (MIS) to address endogeneity without instruments in linear models. The MIS relies on having a couple of suitable indicators, which are measured variables that depend on the latent variable that causes endogeneity, but not on other measured attributes. In this article we extend and assess MIS for Discrete Choice models.

The MIS method is applied in two stages. First, one of the indicators is included as an explanatory variable in the structural equation. By this modification, the endogeneity of other variables is eliminated, and the included indicator becomes the only endogenous variable. Then, in a second stage, the problem is solved by using the second indicator as an instrument for the first one. Wooldridge (2010) applies this method to linear models; where he proposes using the 2SLS method for the second stage. In this article we adapt and apply the MIS method to Logit models by considering the control-function method on the second stage.

For many cases, suitable indicators for applying the MIS method might be easier to obtain in practice than instrumental variables. For example, in a residential choice context, if endogeneity is due to the omission of quality, suitable indicators may be obtained by asking the interviewee about their appreciation of the quality of a dwelling unit. This may be done through a photograph, before knowing other attributes, such as location, price, size or age. Also, in the case of a mode choice model, a question about comfort explicitly differentiated from fare and other modal attributes, can serve as a suitable indicator for applying the MIS method.

We use Monte Carlo experimentation to assess the performance of the MIS relative to the control-function method, and discuss its applicability and limitations for some relevant discrete choice models. Monte Carlo experiments show that both the control-function and the MIS have similar results. This implies that, in cases where obtaining suitable indicators is easier than finding appropriate instruments, the MIS shall indeed be a useful method for addressing endogeneity in discrete choice models.

The article is structured in 5 sections. Following this introduction, Section 2 summarizes the MIS method for linear models. Then, in Section 3 we analyze the requirements needed for the extension of the MIS method to discrete choice models. In addition, the MIS and the CF are compared qualitatively in terms of
their applicability to diverse discrete choice models. Afterwards, in Section 4 we use Monte Carlo experiments to illustrate the efficacy and efficiency of both MIS and CF in small samples, and to study the impact of the failure of their respective assumptions. Finally, Section 5 summarizes the main results and conclusions of the article.

2. The Multiple Indicator Solution (MIS) in Linear Models

The Multiple Indicator Solution (MIS) is a method described by Wooldridge (2010, pp 112) for addressing endogeneity due to omitted variables in linear models. The MIS achieves identification by taking advantage of the correlation between two indicators, correlation that is assumed to be only due to the omitted variable.

To illustrate the MIS method for linear models, consider that the data generation process is described by the following structural equation

\[
y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma q + e_y
\]

where \(y\) is the dependent variable, \(x\) and \(q\) are the independent variables, \(\beta\) and \(\gamma\) are coefficients, and the error term \(e_y\) satisfies \(E(e_y | x, q) = 0\).

When \(q\) is omitted, the usual solution to address endogeneity would be to find instrumental variables (IV) to every element on \(x\) which happens to be correlated with \(q\) and to apply, for example, the Two Stage Least Squares (2SLS) method. Instrumental variables would have to be, at the same time, correlated with the endogenous \(x\) and uncorrelated with the error term \(\gamma q + e_y\), conditions that may be difficult to fulfill in practice (see, e.g., Cameron and Trivedi, 2005, pp. 97; Wooldridge, 2010, pp. 94).

Suppose now that, instead of instrumental variables for the endogenous \(x\), we have two indicators \(q_1\) and \(q_2\) fulfilling the following conditions:

\[
q_1 = \alpha_0 + \alpha_q q + e_{q_1},
\]

\[
q_2 = \delta_0 + \delta_q q + e_{q_2},
\]

where

\[
\text{Cov}(q, e_{q_1}) = \text{Cov}(q, e_{q_2}) = \text{Cov}(x, e_{q_1}) = \text{Cov}(x, e_{q_2}) = \text{Cov}(e_{q_1}, e_{q_2}) = 0,
\]

\(\alpha_q \neq 0\) and \(\delta_q \neq 0\). Then, if \(q\) is replaced by \(q_1\) in Eq.(1), the new error term of the model would be \(v\).

\[
y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma \left( \frac{q_1 - \alpha_0 - e_{q_1}}{\alpha_q} \right) + e_y
\]

\[
y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma \left( \frac{q}{\gamma_1} \right) - \gamma \frac{\alpha_0 - e_{q_1}}{\alpha_q} + e_y,
\]

\(\gamma \neq 1\),
Note that the error term $v$ is not correlated with any $x$ because $\text{Cov}(x, e_q) = 0$ but, in turn, $v$ is correlated by construction with $q_i$ through $e_{q_i}$ as shown in Eq. (2). In other words, $q_i$ becomes the only endogenous variable when $q$ is replaced by $q_i$.

Now, the combination of Eq. (2) and Eq. (3) can be used to show that $q_2$ is a proper instrumental variable for $q_1$. First, $q_2$ is correlated with $q_1$ since both depend on $q$. Second, $q_2$ is not correlated with $v$, because $\text{Cov}(q, e_q) = \text{Cov}(e_{q_i}, e_{q_i}) = 0$, which proves the result.

The MIS method for linear model finishes by considering 2SLS over Eq. (5) and using $q_2$ as instrumental variable for $q_1$. Formally:

Stage 1: Regress $q_i = \theta_0 + \theta_q q_2 + \delta$ to obtain the fitted values $\hat{q}_i = \hat{\theta}_0 + \hat{\theta}_q q_2$

Stage 2: Estimate linear model considering the following explanatory variables

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma_1 \hat{q}_i + \tilde{\epsilon}_y.$$  \hspace{1cm} (6)

Note that only the $\beta$’s, but not $\gamma$ in Eq. (1) are indentified when estimating Eq. (6) for the MIS method. Instead of $\gamma$ for the omitted quality $q$, what it is obtained from Eq. (6) is $\gamma_1 = \gamma / \alpha_q$, and $\alpha_q$ cannot be identified from the first stage of 2SLS. It can be shown that, when estimating the first stage, what is obtained is $\theta_q = \alpha_q / \delta_q$, what again does not allow identification of $\alpha_q$. In other words, despite that all the coefficients of the measured variables are identified when applying the MIS method, only the sign of the omitted variable can be inferred from it.

The MIS is different from the direct application of 2SLS in two ways. First, if all $x$’s happen to be correlated with $q$, for applying 2SLS we would need to gather at least one instrumental variable for each $x$. In turn, MIS would only require having the two proper indicators described in Eq. (2) –(3) to address endogeneity.

The second difference is regarding the nature of the auxiliary variables needed for 2SLS and MIS. For 2SLS, the instrumental variables have to be correlated with the endogenous variable but uncorrelated with the error terms. In turn, for the MIS, one needs indicators that are only correlated with the omitted variables.

For example, consider the problem of modeling the wage rate as a function of the level of education. There might be endogeneity in this case because of the omission of ability or some family background characteristics that may both explain the level of education attained and the earnings (Wooldridge, 2010, pp. 93–114).

Arguably, first-quarter, a dummy that takes value 1 if the individual was born in the first quarter of the year, might be a proper instrument for endogenous education (Angrist and Krueger, 1992). On the one hand, being born a particular month should be independent of person’s ability and other omitted variables. On the other hand, USA’s compulsory school attendance laws have make some
individuals attend school for longer, an event that happens be correlated with being born in the first quarter of the year.

In turn, for the same problem, possibly proper indicators for omitted ability might be data on an Intelligence Quotient (IQ) test and a Knowledge of the Working World (KWW) test (Wooldridge, 2010, pp. 114).

Two final things have to be highlighted from the MIS. The first is that role of \( q_1 \) and \( q_2 \) can be reversed in Eq (2)-(5) attaining exactly the same result. The second is that the method will not work if some of the \( x \) in Eq. (1) happens to also be in Eq. (2) or (3), basically, because assumption \( \text{Cov}(x, e_{q_1}) \neq 0; \text{Cov}(x, e_{q_2}) \neq 0 \) would be broken. This places a practical problem in the application of the MIS method, as the redundancy assumption cannot be fully verified. For example, in the problem of modeling the wage rate as a function of education, the assumption would be broken if somehow IQ and/or KWW are related with experience, tenure or other variables that might also be explanatory variables for education (Wooldridge, 2010).

3. The Multiple Indicator Solution (MIS) in Discrete Choice Models

We now study the extension of the MIS method to a Discrete Choice framework. In this case there is a choice set \( C_n \) for each individual \( n \) from which only one alternative is chosen. The structural equation in Eq. (1) turns now into a utility \( U \) that is latent. Instead, an indicator \( y_{in} \) is observed, which takes value 1 if the utility of alternative \( i \) is the largest among those in the choice set, and takes value 0 otherwise. This choice model is described in Eq. (7).

\[
U_{in} = \beta_0 + \beta_1 x_{in1} + \cdots + \beta_k x_{inK} + q_j + e_{Uin},
\]

\[
y_{in} = I\left[U_{in} \geq U_{jn} \forall j \in C_n\right],
\]

Consider now that variable \( q \) is omitted. If \( q \) is correlated with any \( x \), the control-function method (Heckman, 1978; Petrin and Train, 2002) could be applied to correct for endogeneity. To apply this method, an instrumental variable is needed for each and every \( x \) that happens to be correlated with the omitted \( q \). As with the 2SLS, the instrumental variables need to be correlated with the endogenous variable, but uncorrelated (formally independent) with the error terms of the model.

For example, Guevara and Ben-Akiva (2006, 2012) used the control-function method to correct for endogeneity due the omission of quality attributes that are correlated with price in a residential location choice model. Following an argumentation used previously by Hausman (1996) and Nevo (2001) in other contexts, Guevara and Ben-Akiva (2006, 2012) used as instrument for dwelling price the average price of other dwellings located not too near (to avoid reflection bias), but also not too far (to avoid being irrelevant) from the incumbent dwelling.

Finding instruments in other circumstances may be difficult or simply impossible. In the same applications of residential location performed by Guevara and Ben-Akiva (2006, 2012), the authors assumed that measured attributes other than price were not correlated with the omitted quality. Although this assumption is
reasonable, if that happens not to be the case, it would not be straightforward to obtain, for example, proper instruments for the size and age of the dwelling.

Another example of the difficulty of gathering proper instruments in discrete choice is Walker et al (2010). In that application the authors studied the problem of endogeneity due to measurement errors in link travel times, case in which it is unclear how to obtain proper instrumental variables for being able to apply the control-function method.

A third example occurs in mode choice modeling. Safety, comfort and reliability are attributes that are very relevant in this type of models, but are hard to measure, and are also correlated with travel cost and time. This may cause endogeneity, but the CF is difficult to apply because it is unclear how to obtain proper instrumental variables for such omitted modal attributes.

Even if plausible instruments can be found and their application results in a coherent correction of the estimated parameters, the validity of the instruments is debatable. Newey (1985) shows that over-identification tests needed to confirm the exogeneity of the instruments are inconsistent, that is, even when the sample size goes to infinity, the power of the test does not go to 1. Although De Blander (2008) shows that the alternate hypothesis that cannot be detected by such tests is very atypical, the fact is that exogeneity of the instrumental variables is as questionable as the redundancy of the indicators for applying the MIS.

Furthermore, there is the problem of weak instruments. When the instruments are only weakly correlated with the endogenous variable, the IV estimation may be as poor as the uncorrected model (Staiger & Stock, 1997). It has been shown that and F test larger than 10 in the first stage of 2SLS is a sign of having sufficiently strong instrumental variables. Recent research by Guevara and Navarro (2013) suggest that similar thresholds are applicable in the case of the CF in Logit models.

Given the possible difficulties of gathering proper instruments, an extension of the MIS to discrete choice seems attractive. The MIS may make possible to correct for endogeneity using potentially milder assumptions or by procedures that may be more feasible to apply in practice in discrete choice models.

The proposed MIS method for discrete choice models is a direct extension of the MIS method for linear models described in Section 2. Consider the choice model described in Eq. (7) but where indicators are available. Then, if the conditions shown in Eq. (4) are fulfilled, consistent estimators of the model parameters would be obtained by applying the following tow stage procedure:

Stage 1: Regress \( \hat{q}_1 = \hat{\theta}_0 + \hat{\theta}_q q_2 + \hat{\delta} \) to obtain the residuals \( \hat{\delta} = q_1 - \hat{\theta}_0 - \hat{\theta}_q q_2 \)

Stage 2: Estimate choice model considering the following variables in the systematic utility:

\[
V_{in} = \beta_0 + \beta_1 x_{1n} + \cdots + \beta_k x_{Kn} + \beta_q q_{1n} + \beta_s \hat{\delta}_m, \tag{8}
\]
Formally, the application of the control-function in this context requires \( (e_{q2}, v) \) to be independent of \( q \); (Wooldridge, 2010, pp.585). This mild assumption is only slightly stronger than Eq.(4), but equally plausible and, furthermore, just alike any other application of the CF method.

As with the traditional CF method, the MIS procedure can be applied in two steps or simultaneously by full information maximum likelihood (FIML). Since only one instrumental variable is used in this case, the two-stage estimator will be as efficient as the FIML version (Rivers and Vuong, 1988). However, the standard errors of the two stage procedure cannot be directly obtained from the information matrix, but would have to be calculated using, for example, the bootstrap approach proposed by Karaca-Mandic and Train (2003).

Just as it occurs with the CF, the direct application of the MIS method described in Eq. (6-7) requires assuming that the error term \( e_{vin} \) to be distributed Normal, what leads to the Probit model. However, accepting that the approximation of a Normal by an Extreme Value distribution causes negligible discrepancies (Lee, 1982; Ruud 1983), the resulting model becomes a Logit, which is much easier to apply in practice.

Finding proper indicators fulfilling the conditions required by Eq. (2-3) is very plausible in practical discrete choice applications. For example, in the case of residential location endogeneity is expected to occur because of the omission of quality attributes that are correlated with price. In that case, proper indicators can be obtained in practice by, for example, asking RP or SP the interviewee to give an opinion or to rank the dwelling, taking into consideration all but measurable attributes such as price, size or location. This final requirement may be the more difficult to fulfill in practice.

On another example, omitted comfort in modal choice is currently been addressed by asking for indicators of it and by modeling it as a latent variable, such as in Glerum et al (2012). However, using latent variables in this context has two major challenges. The first is computational. If the omitted quality is by alternative, a multifold integral would have to be evaluated, making the model impractical. This is why, to the best of our knowledge, all applied latent variables models have considered so far only latent characteristics of the individuals, but not latent attributes of the alternatives. The second challenge of using the latent variable approach in this context is causal. When the latent variable is a person's characteristic, such as the appreciation of the comfort of a given mode, it may make sense to consider socioeconomic characteristics as causal variables. In turn, finding causal variables for modeling the omission of comfort, safety or other modal attributes as latent variables, is a problem that has not been addressed so far in empirical literature. Instead, the MIS approach may be used directly in that case using modal indicators, such as the ones collected by Glerum et al (2012).

4. Monte Carlo Experiment

With the purpose of illustrating the application and to assess the relative performance of the MIS and CF methods in finite samples we develop a Monte Carlo experiment...
Carlo experiment of a binary choice model. We create endogeneity by omitting a variable that is correlated with another observed variable. Then, we create proper and improper instruments and indicators to apply both the CF and the MIS to the problem. We repeat each experiment 100 times and analyze the sampling distribution of the estimators by plotting the results and comparing various statistics.

The systematic utility of the binary choice model considers four attributes \( p, a, b, \) and \( q \) for each alternative \( (i) \) and an error term distributed Extreme Value \((0,1)\) for 1000 observations \((n)\).

Because the error term of the utility is distributed Extreme Value \((0,1)\), the choice model is a binary Logit, which was simulated considering the following set of population coefficients \( \beta_p = -2, \beta_a = 1, \beta_b = 1, \beta_q = 1 \). Formally, an indicator taking value \( i \) if the alternative was chosen was simulated using the probability model shown in Eq. (9).

\[
P_p(i) = \frac{e^{-2p_u + a_u + b_u + q_u}}{e^{-2p_u + a_u + b_u + q_u} + e^{-2p_u + a_u + b_u + q_u}},
\]

Variables \( a, b, q, z_1 \) and \( z_2 \) were generated as independent random uniform \((-3,3)\). The attribute \( p \) was generated as function of a variable \( z \) and attribute \( q \), plus an error term \( \varepsilon \) distributed random uniform \((-1,1)\) with the coefficients show in Eq. (10).

\[
p_{in} = 5 + 1z_{in} + 0.03wz_{in} + 1q_{in} + \varepsilon_{in}
\]

By virtue of Eq. (10), \( q \) is correlated with \( p \). Therefore, if \( q \) is omitted in the estimation, the choice model will suffer of endogeneity. Furthermore, also by virtue of Eq. (10), \( z \) makes a valid instrument for \( p \) because it is correlated with \( p \) and is exogenous to the model. In turn, \( w_z \) is a weak instrument, because it will be only slightly correlated with \( p \). Additionally, an invalid instrumental variable that is correlated with \( q \) was created as shown in Eq. (11) to study the impact of the failure of instruments' exogeneity assumption.

\[
x_{z_{in}} = 1z_{in} + 1q_{in}
\]

Finally, to apply the MIS method, two proper indicators \((q_1, q_2)\), two endogenous indicators \((xq_1, xq_2)\) and two weak indicators \((wq_1, wq_2)\) were generated for the omitted quality \( q \). The expression used to generate the indicators are shown in Eq. (12), where the error terms \( \varepsilon_{in} \) are distributed random uniform \((-1,1)\).
Under this setting, the following nine models were estimated using the same samples generated 100 times:

**True model**: Estimated considering all the explanatory variables described in Eq. (9). This model is used as a benchmark as the best possible outcome that could be obtained by the models that correct for endogeneity.

**Endogenous model**: Estimated excluding variable \( q \) in the systematic utilities in Eq. (9). This model is used as a benchmark for the doing-nothing alternative.

**CF**: Control-function method using the proper instruments \( z \) described in Eq. (10) to correct the endogenous model. Since \( z \) are exogenous and also sufficiently correlated with the endogenous \( p \), the results show fully functional application of the CF method. The strength of the instruments was verified by checking that the F test of the first stage of the CF method was by far larger than 10 (Guevara and Navarro, 2013).

**CF-weak**: Control function method using the weak instruments \( wz \) described in Eq. (10). In this case the F test of the first stage of the CF method was smaller than 10 (Guevara and Navarro, 2013).

**CF-endog**: Control function method using the endogenous instruments \( xz \) described in Eq. (11). Since \( xz \) is correlated with \( q \) by construction the estimators of this model are inconsistent.

**MIS**: Multiple Indicator Solution using the proper indicators \( q_1 \) and \( q_2 \) described in Eq. (12). This is an example of a fully functional version of the MIS method.

**MIS-two endog**: Multiple Indicator Solution using the endogenous indicators \( xq \) described in Eq. (12). In this case both indicators depend not only on the omitted quality, but also to the measured variable \( p \).

**MIS-one endog**: Multiple Indicator Solution using one endogenous indicators \( xq \) and one proper indicator \( q \) described in Eq. (12). In this case only one of the indicators depend also on the measured variable \( p \). This could be seen as a case of a mild violation of the MIS assumptions.

**MIS-weak**: Multiple Indicator Solution using the weak indicators \( wq \) described in Eq. (12). In this case both indicators are only weakly correlated with the omitted quality \( q \).

\[
\begin{align*}
q_{1in} &= 2q_{in} + e_{q1in} \\
q_{2in} &= 2q_{in} + e_{q2in} \\
xq_{1in} &= 0.5q_{in} + 1p_{in} + e_{q1in} \\
xq_{2in} &= 0.5q_{in} + 1p_{in} + e_{q2in} \\
wq_{1in} &= 0.03q_{in} + e_{wq1in} \\
wq_{2in} &= 0.03q_{in} + e_{wq2in}
\end{align*}
\]
These nine model were compared in terms of efficacy an efficiency. Because the correction of endogeneity in discrete choice models produces a change in the scale of the estimators (Guevara and Ben-Akiva, 2012), we analyze the ratio of the estimators instead of the estimators themselves. Table 1 summarizes the result of the ratio $\hat{\beta}_p / \hat{\beta}_a$ for the following four statistics:

**Bias**: Difference between the average of the respective estimators true value of the ratio $\beta_p / \beta_a = -2$ for the 100 repetitions. A smaller Bias implies better small sample efficacy for recovering the true values of the model.

**Root Mean Squared Error (RMSE)**: Square root of the sum of the sampling variance and the square of the bias. A smaller RMSE implies better small sample efficiency of the method.

**t-test**: Ratio between the bias and the sampling standard deviation of the estimators. This statistic can be used to test the null hypothesis that the mean of the sampling distribution is equal to its respective true value.

**Count**: Number of times the estimators of the 100 repetitions are within a 75% confidence interval of the respective true value. The interval is constructed using the sampling variance from all the repetitions. This statistic is usually termed the empirical coverage. The larger this statistic is, the better the performance of the method. The closer to 75% this statistic is, the closer its empirical distribution is to its theoretical sampling distribution.

**Table 1: Monte Carlo Analysis. Comparison of CF and MIS Methods**

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>RMSE</th>
<th>t-test</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.007084</td>
<td>0.1086</td>
<td>0.06538</td>
<td>75</td>
</tr>
<tr>
<td>Endogenous</td>
<td>0.3924</td>
<td>0.4057</td>
<td>3.801</td>
<td>2</td>
</tr>
<tr>
<td>CF</td>
<td>0.01371</td>
<td>0.1093</td>
<td>0.1265</td>
<td>79</td>
</tr>
<tr>
<td>CF-weak</td>
<td>0.2381</td>
<td>3.808</td>
<td>0.06266</td>
<td>87</td>
</tr>
<tr>
<td>CF-endog</td>
<td>0.4132</td>
<td>0.4256</td>
<td>4.050</td>
<td>2</td>
</tr>
<tr>
<td>MIS</td>
<td>0.01616</td>
<td>0.1095</td>
<td>0.1492</td>
<td>74</td>
</tr>
<tr>
<td>MIS-both endog</td>
<td>-1.238</td>
<td>1.259</td>
<td>-5.416</td>
<td>0</td>
</tr>
<tr>
<td>MIS-one endog</td>
<td>-0.06471</td>
<td>0.2020</td>
<td>-0.3382</td>
<td>72</td>
</tr>
<tr>
<td>MIS-weak</td>
<td>0.3892</td>
<td>0.4028</td>
<td>3.744</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 1 replicates the same results, but by plotting the whole sampling distribution of the 100 repetitions. The vertical axis corresponds to the ratio $\hat{\beta}_p / \hat{\beta}_a$ and the methods are ordered in the horizontal axis. The grey dots correspond to the values of obtained for the ratio $\hat{\beta}_p / \hat{\beta}_a$ for each repetition and the black larger dot represents the respective sample average. The horizontal line in the middle marks the true value of the ratio $\beta_p / \beta_a = -2$. 

Figure 1: Detailed Results of 100 Repetitions of Variations of CF and MIS Methods to Correct for Endogeneity

Results show that both CF and MIS have very similar properties when their respective assumptions hold, with statistics only slightly poorer than those of the true model. The bias is below 5% of the bias obtained with the uncorrected model. RMSE are of the same order of magnitude as that of the true model, and the empirical coverage is very close to its population value of 75. Also, the hypothesis that $\beta_p/\beta_a = -2$ cannot be rejected at any reasonable level of confidence.

In turn, in some cases when the assumptions of each method fail, the results are remarkably wrong. The worse case is for the MIS method estimated with two endogenous indicators. The results in this case are even worse than those obtained for the endogenous model. The bias relative to the uncorrected model is over 300% and the relative RMSE is larger by approximately the same amount. Also, the count is zero, indicating that not even one of the 100 simulations
resulted in an estimator within a 75% confidence interval of its true value. Concordantly, the null hypothesis that $\beta_p/\beta_a = -2$ is erroneously rejected with 99% confidence.

MIS-weak and CF-Endog have similar results, just as bad as those of the uncorrected model. Both the relative Bias and RMSE are approximately 100% of the respective value of the uncorrected model. The empirical coverage is also very similar, with a count just equal to 2. Concordantly, the null hypothesis that $\beta_p/\beta_a = -2$ is erroneously rejected with 99% confidence.

On the other hand, when CF is estimated using weak instruments, the mean is closer to the true value, but the variance is huge. The relative bias is 61% but the relative RMSE is more than 900%. Note that some realizations of this method were not plotted in Figure 1, to allow having a proper scale to compare the results of the various methods. Finally, the t-test is low and the count is high, but just because of the huge variance attained in this case.

Finally, it is interesting to remark what happens with MIS when only one of the indicators was endogenous. The results are only slightly worse than the true model. The count is larger than 70, the Bias relative to the endogenous model is about 15% and the RMSE is even half that of the true model. This suggests that a mild violation of the MIS assumptions may yield only a reduced impact in the quality of the estimators. In turn when the violation is strong, meaning that both indicators are endogenous, the results might be significantly worse than those of the uncorrected model.

4. Conclusion

This article shows that Multiple Indicator Solution (MIS), proposed originally by Wooldridge (2010) for linear models, can be extended to discrete choice modeling under some mild assumptions. This is valuable because, differently from the control-function (CF) method, the MIS has no need of gathering instrumental variables, which may be very difficult to obtain for various discrete choice applications.

Formally, the condition needed for the extension of the MIS method to Discrete Choice modeling is that, instead of assuming that one of the indicators is only uncorrelated with the error terms of the model, it need to be assumed the indicator is independent of them. This assumption is mild in the sense that it is the same formal assumption needed for the application the CF to correct for endogeneity. Furthermore, no empirical differences would be noticed in the estimation if the variables happen to be dependent but are uncorrelated.

The MIS can straightforwardly be applied to some relevant discrete choice models in which omitted attributes are measured through indicators in RP and SP surveys. For example, this method can be applied when modeling the impact of the omission of comfort, safety and reliability in mode choice models. In such a case the MIS method can be applied if two indicators of the omitted attributes are available, such as in the application of Glerum, et. al (2012)
We also use Monte Carlo experiments to illustrate the efficacy and efficiency of MIS and CF methods in Logit Models, and to study the impact of the failure of their respective assumptions. Results suggest that both the MIS and CF methods attain similar efficacy and efficiency when their respective assumptions hold, but MIS seems to be more robust to mild violations that may occur in practice, such as that only one of the indicators are not uncorrelated with other measured attributes.

Finally, it can be expected that the MIS method to correct for endogeneity may have a relevant impact for practitioners for three reasons. First, the MIS can be applied using indicators that can be easily collected in practice in several discrete choice models. Second, the MIS seems to be robust to some mild violations of its assumptions. Third, the MIS method can be easily applied with canned software (at least in its two stage version), with no need for evaluating multifold integrals when addressing endogeneity due to the omission of attributes of the alternatives.

References


