Effects of Informal Family Care on Formal Health Care: Zero-Inflated Endogenous Count for Censored Response

(June 25, 2011)

Young-sook Kim
Korean Woman's
Development Institute
Seoul 122-707, South Korea

Myoung-jae Lee*
Department of Economics
Korea University;
Research School of Economics
Australian National University

Whether informal family health care is a substitute or complement for formal health care has been debated in the literature. If it is a substitute, then there is a scope to reduce formal health care cost by promoting informal family health care. Using Korean survey data for the elderly of age 65 or higher, this paper estimates the effect of informal family health care on formal health care, where the former is measured by the number of family health care givers and the latter is measured by the formal health care expenditure. But this task poses a number of difficulties. The first is that the number of the family care givers is an endogenous count regressor. The second is that there seem to be too many zeros in the count (85%). The third is that the response variable also has a non-trivial proportion of 0's (14%). This paper overcomes these problems by combining a semiparametric estimator for a censored response with the idea of "zero-inflated" counts. The resulting two-stage procedure avoids strong parametric assumptions and behaves well computationally. Our main empirical finding is that informal family health care has a large substitute effect for diabetics that is statistically significant and large in magnitude, but the other effects are statistically insignificant for our given data size of about 3000.

Key Words: informal health care, formal health care, count variable, zero-inflated, control function, censored model.

* Corresponding Author: Myoung-jae Lee, Department of Economics, Korea University, Anam-dong, Sungbuk-gu, Seoul 136-701, South Korea; myoungjae@korea.ac.kr; 82-2-3290-2229 (phone/fax).

1 Introduction

With low fertility rates prevailing in most developed countries, the populations age fast, and this entails a high demand for health care. If the health care cost is borne only by formal health care, then eventually there may be a point at which the health care system ceases to be sustainable. If formal health care can be replaced to some extent by informal family health care, then this may lead to a considerable reduction on the formal health care cost.

In the literature of health economics, there are studies that examined the effects of informal health care on formal health care, which often find that informal care substitutes for formal care. Although there are studies such as Charles and Sevak (2005) showing that informal care (dummy for any informal care) is a substitute for nursing home care (dummy for ever staying in nursing home), in the following, we briefly review three studies that are the most relevant to our paper: Van Houtven and Norton (2004), Bolin et al. (2008) and Bonsang (2009).

In Van Houtven and Norton (2004), informal care is the care hours provided by all children (their spouse and their children), and formal cares including nursing home care and outpatient care are of eight different types in total (mostly continuously distributed, but home health care and outpatient surgery are binary). About 19% of the respondents received informal care. Van Houtven and Norton used U.S. data: 1998 Health and Retirement Survey (HRS) and 1995 Asset and Health Dynamics Among the Oldest-Old Panel Survey (AHEAD). Van Houtven and Norton found that informal care is mostly a substitute except for outpatient surgery.

In Bolin et al. (2008), nine different formal care variables are used including formal home care, visits to doctors and hospitalization days. For informal care, they used the informal care hours from children and grandchildren, and its non-zero proportion ranged 19-40% across the countries in the data (2004 European data "SHARE"). Bolin et al. found that informal care is a substitute for formal home care, but a complement to doctor and hospital visits, and that the effects vary depending on the region (i.e., informal care interacts with the region dummies).

In Bonsang (2009), informal care is the care hours by children of the respondent (a single-living elderly), and formal cares are paid domestic help (low-skilled) and nursing care (high-skilled); both formal cares are home cares. Using the 2004 European data SHARE,

Bonsang (2009) found that informal care is a substitute for the low-skilled formal home care, but a weak complement for the high-skilled formal home care, and that the substitution effect decreases as the level of disability of the elderly person increases (i.e., informal care interacts with the disability level).

In terms of methods, Van Houtven and Norton (2004), Bolin et al. (2008) and Bonsang (2009) used a 'two-part approach'. But strictly speaking, the methods used there to deal with endogenous regressors apply only when the endogenous regressors are continuously distributed. Probably because of this restriction, least squares estimator (LSE) was used to estimate the reduced form model for informal care that is an endogenous regressor for formal care. But the LSE is problematic as informal care variables include many zeros. As instruments for informal care, distances to children, placement of daughters in the birth order, or the number of (female) children have been used.

One reason for the endogeneity of informal care is that both formal and informal cares may be determined simultaneously. Another reason is that both cares may share common factors—most notably, health status. But controlling for health status is troublesome, as it may be influenced by both cares. Added to this is the aforementioned problem of too many zeros and the fact that the endogenous and response variables are not continuously distributed, but discrete or mixed (discrete and continuous).

While there is no particularly good solution for the endogeneity problem, this paper will show a two-stage procedure to overcome the problems of too many zeros in a non-negative endogenous regressor (informal care) and the non-trivial proportion of zeros in the response variable at zero (formal care). For non-negativity, we will be using 'Quasi Poisson' approach, and for too-many zeros, we will be using the zero-inflated Poisson idea of Lambert (1992). In a nutshell, our two-stage procedure is applicable to censored models with non-negative endogenous regressors including count variables where the endogenous regressors have too many zeros.

The rest of this paper is organized as follows. Section 2 shows the details of the two-stage procedure. Section 3 applies the estimator to Korean data to estimate the effect of informal care on formal care, where informal care is the number of care givers (thus a count). Finally, Section 4 concludes. A word on notation before proceeding further: '$a \amalg b|c$' denotes the independence between a and b given c.

2 Two-Stage Procedure

2.1 Model Assumptions

Suppose that $y_1 \geq 0$ is formal care, $y_2 \geq 0$ is informal care (a count), x_1 is a $k_1 \times 1$ exogenous regressor vector relevant for the y_1 structural form (SF) equation, and x is the $k \times 1$ system exogenous regressor vector for (y_1, y_2) that strictly includes x_1. Observed are

$$(x_i, y_{1i}, y_{2i}),\ i = 1,\ ...,\ N, \quad \text{which are iid across } i.$$

Although we assumed y_2 to be a count, our approach below also applies to a non-negative y_2. In view of the iid assumption, we will often omit the subscript i.

Assume that the observed y_1 and y_2 are generated from its latent versions y_1^* and y_2^* as follows: for unknown parameters γ_y, γ_x, α and β, an error term u_i and a binary variable q_i,

$$\begin{aligned} &y_{1i} = \max(0, y_{1i}^*) \quad \text{with} \quad y_{1i}^* = \gamma_y y_{2i} + x_{1i}'\gamma_x + u_i \quad \text{and } u|x \text{ is symmetric around 0;} \\ &y_{2i} = q_i y_{2i}^*, \quad P(q=1|x_i) = \frac{\exp(x_i'\alpha)}{1+\exp(x_i'\alpha)} \quad \text{and} \quad E(y_2^*|q=1, x_i) = \exp(x_i'\beta). \end{aligned}$$

In this model, y_1^* is modelled as censored from below at zero with its error term symmetric around zero; this symmetry assumption is to use symmetrically censored least squares estimator (SCL) of Powell (1986) for y_1, and may be replaced by another semiparametric assumption if a different semiparametric estimator as in Powell (1984) or Lee (1992) for the zero-censored model is used.

Some remarks about the model are in order. First, a sample selection model holds for y_2^* because y_2^* is observed only when $q = 1$; the binary 'selection variable' q is assumed to follow the logit model whereas y_2^* given $q = 1$ is posited to have an exponential regression function. Second, a key implication of the selection model for y_2 is

$$E(y_2|x) = P(q=1|x)E(y_2^*|q=1, x) = \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta).$$

Third, it may be better to model y_1 also as a sample selection model rather than the censored model (the censored model is a special case of selection model), but the censored model is adopted for simplicity because dealing with a sample selection model is difficult—this may not matter much though as the proportion of zeros is low for y_1 in our data (14%). Fourth, since the system regressor x appears for q and y_2^*, the q and y_2^* equations should be regarded as 'reduced forms (RF)'. This RF view is necessary because y_1 does not appear for the q and

y_2^* equations (as if y_1 has been substituted out), and also because $E(y_2^*|q=1,x)=\exp(x'\beta)$ is adopted, not the more "structural" $E(y_2^*|x)=\exp(x'\beta)$.

Define $1[A]=1$ if A holds and 0 otherwise, and call $y_2^*=0$ 'participation zero'. As done in Lee (2011), it is helpful to compare three different models for q in relation to the participation zero possibility:

Model 1 : $q=1[y_2^*>0]$ where $y_2\ (=qy_2^*)=0$ implies $y_2^*=0$;

Model 2 : q determined by some variables (and y_2^*) with participation 0 possible;

Model 3 : q determined by some variables (and y_2^*) with no participation 0 ($y_2^*>0$ always).

Model 1 is the 'corner solution model' in which case y_2 becomes also a zero-censored model as y_1 is. Model 2 is relevant if $q=1$ is only an "attempt/try" for an activity and y_2^* is a "performance" in the activity following the attempt/try. Model 3 is relevant if $q=1$ is having the actual activity and y_2^* is the degree of the activity with zero ruled out.

For instance, $q=1$ may be an attempt/try to export, where $y_2^*=0$ is possible even if one tries ($q=1$). Instead of attempt/try, one may define $q=1$ as actually exporting and y_2^* as the actual export volume that cannot be zero. Which one between Models 2 and 3 to adopt may depend on what is available in the data. If a variable for 'whether one desires to export or not' is available in the data along with the export volume including zero, then $y_2=qy_2^*$ is the observed export volume with $y_2^*=0$ possible. If only the actual export volume including zero without such a variable for q is available in the data, then one has no choice but to set $q=1[y_2>0]$ ($\neq 1[y^*>0]$), in which case $q=1 \iff y_2>0$ with no participation zero possible. In our data, since there is no separate variable for q, we will set $q=1[y_2>0]$ to adopt Model 3

One may wonder 'why not adopt Model 1 that looks simpler than Model 3'. The answer is that there is really no difference between Model 1 and Model 3 for our empirical analysis. Suppose $y_2^*=x'\alpha+v_2$ with v_2 being logistic independently of x and Model 1 holds. Then

$$q=1[y_2^*>0]=1[x'\alpha+v_2>0] \implies E(q|x)=\frac{\exp(x'\alpha)}{1+\exp(x'\alpha)} \quad \text{and}$$
$$E(y_2^*|q=1,x)=E(y_2^*|y_2^*>0,x)=x'\alpha+E(v_2|v_2>-x'\alpha,\ x)\neq\exp(x'\alpha).$$

In this case, the exponential model is only an approximation for $x'\alpha+E(v_2|v_2>-x'\alpha,\ x)$, and consequently we need to allow different parameters α for $E(q|x)$ and β for $E(y_2^*|q=1,x)$ as when Model 3 is adopted.

2.2 First Stage To Obtain Control Function

In our two-stage procedure, the first stage consists of two parts: estimating α in the logit model for $E(q|x)$ and estimating β in the exponential model for $E(y_2^*|q=1,x)$. For the latter, one can use Quasi-Poisson (QPOI) maximum likelihood estimator (MLE): maximize the usual Poisson likelihood function with $q=1$ attached to use the "sandwich-form" asymptotic variance. That is, the QPOI maximand is

$$\frac{1}{N}\sum_i q_i\{y_{2i}x_i'b-\exp(x_ib)\}\ (\ =\frac{1}{N}\sum_i q_i\{y_{2i}^*x_i'b-\exp(x_ib)\}\ )$$

and the asymptotic variance matrix is

$$E^{-1}\{qxx'\exp(x'\beta)\}\cdot E[qxx'\{y-\exp(x'\beta)\}^2]\cdot E^{-1}\{qxx'\exp(x'\beta)\}.$$

Denoting the first-stage estimators as $\hat{\alpha}$ and $\hat{\beta}$, the second-stage is estimating γ_y and γ_x for the y_1 SF allowing for the endogeneity of y_2 in the y_1 SF. As reviewed in Lee (2012), there are several different methods to deal with an endogenous regressor in a limited dependent variable (LDV) model—the LDV model is the zero-censored model for y_1 in our case. Among those methods, the most convenient for our empirical analysis is 'control function (CF)' approach, because many interaction terms between y_2 and elements of x will be used. With the endogeneity of y_2 removed by a CF, we can freely allow such interaction terms, which is complicated in the other approaches for the y_2 endogeneity. Specifically, a residual $\hat{v}_2$ for y_2 is obtained from the first stage, and it is used as an extra regressor in the y_1 SF. Not just $\hat{v}_2$, but also $\hat{v}_2^2$ and $\hat{v}_2^3$ (or higher-order terms) can be used if including those terms removes the y_2 endogeneity better by accounting for the additive part of u that depends on v_2. Then $(\hat{v}_2,\hat{v}_2^2,\hat{v}_2^3)$ becomes the CF, and the y_2 endogeneity can be tested by looking at whether their coefficients are all zero or not.

For an LDV regressor such as y_2, it is not obvious which form of residual will be the best choice for CF. For a count regressor, there is no "natural" residual. To motivate our approach to this, consider generating a Poisson regressor y with the parameter $\exp(x'\xi+\varepsilon)$ where ε is related to u so that y becomes endogenous for y_1; e.g., u consists of ε and an additive error. For such y, many exponential random variables with the same parameter $\exp(x'\xi+\varepsilon)$ should be generated first. Then the number of the exponential durations that can be fit into the unitary time interval is the desired y—after this, y_1 can be generated using (x and) y and u

that depend on ε. For the endogenous y, at least the following two types of residuals can be thought of.

The 'additive residual' for y is $y - \exp(x'\xi)$, from which it follows that

$$E\{y - \exp(x'\xi) \mid x\} = E[\, E\{y - \exp(x'\xi)|\varepsilon, x\} \mid x] = E[\, \exp(x'\xi)e^{\varepsilon} - \exp(x'\xi) \mid x]$$
$$= E[\exp(x'\xi) \cdot (e^{\varepsilon} - 1) \mid x] = 0$$

which holds by rescaling ε such that $e^{\varepsilon} = 1$ and including the constant scale factor in the intercept of $x'\xi$. That is, using $y - \exp(x'\xi)$ amounts to using $\exp(x'\xi)(e^{\varepsilon} - 1)$ as a CF in the y_1 SF. If ε is small, then

$$\exp(x'\xi)(e^{\varepsilon} - 1) \simeq \exp(x'\xi) \cdot \varepsilon.$$

A better choice than the additive residual might be the multiplicative residual $y \exp(-x'\xi) - 1$, which leads to

$$E\{y \exp(-x'\xi) - 1 \mid x\} = E[\, E\{y_2 \exp(-x'\xi) - 1|\varepsilon, x\} \mid x] = E(e^{\varepsilon} - 1|x) = 0.$$

Hence, using $y_2 \exp(-x'\xi) - 1$ is analogous to using $e^{\varepsilon} - 1$ as a CF in the y_1 SF. If ε is small, then $e^{\varepsilon} - 1 \simeq \varepsilon$.

The main difference between the two residuals is that the additive residual carries the heteroskedasticity factor $\exp(x'\xi)$ while the multiplicative residual does not. For $y_2 = qy_2^*$, the two residuals are, respectively,

$$y_2 - \frac{\exp(x'\alpha)}{1 + \exp(x'\alpha)} \exp(x'\beta) \quad \text{and} \quad y_2\{\frac{\exp(x'\alpha)}{1 + \exp(x'\alpha)} \exp(x'\beta)\}^{-1} - 1.$$

For our empirical analysis, we will try both residuals, because which is better will be determined ultimately by how much endogeneity can be picked up by each type of residual; the more the better.

Since SCL in the second stage needs only the symmetry of $u|x$, the only parametric assumption invoked in our two-stage procedure is the logit in the first-stage. But since there is no practical semiparametric estimator for binary responses, assuming logit does not seem so restrictive. If we desire to avoid even the logit assumption, then we may assume simply

$$E(y_2|x) = \exp(x'\beta).$$

This will be also applied to our data later, and as it turn outs, its performance is inferior to the two-stage procedure.

2.3 Second Stage with Symmetrically Censored LSE (SCL)

In our two-stage procedure, the second-stage is SCL with a CF used as an extra regressor to remove the y_2 endogeneity. Here we explain SCL first, pretending that y_2 is exogenous for a while. To simplify notations, define

$$w \equiv (y_2, x_1')' \quad \text{and} \quad \gamma \equiv (\gamma_y, \gamma_x')'.$$

to get $y_{1i} = \max(0, w_i'\gamma + u_i)$.

Observe

$$w'\gamma + u \geq 0 \iff u \geq -w'\gamma.$$

If $w'\gamma > 0$, then the censoring of y_1 at zero replaces the lower tail of u with a "mass" $-w'\gamma$. The idea of SCL is to replace the upper tail with $w'\gamma$ to restore the symmetry of u. This leads to a moment condition:

$$E\{\ 1[w'\gamma > 0] \cdot (1[|u| < w'\gamma]u + w'\gamma 1[|u| \geq w'\gamma]) \cdot w\} = 0.$$

A minimand with the moment condition as its asymptotic first order condition is

$$\frac{1}{N}\sum_i [\ \{y_{1i} - \max(0.5y_{1i}, w_i'\gamma)\}^2 \ + \ 1[y_{1i} > 2w_i'\gamma] \cdot \{(0.5y_{1i})^2 - (\max(0, w_i'\gamma))^2\}\ ]$$

and SCL is obtained by minimizing this for γ.

If $w_i'\gamma \simeq \infty\ \forall i$, then the SCL minimand becomes the LSE minimand $N^{-1}\sum_i (y_{1i} - w_i'\gamma)^2$; in fact, what is needed is only $u > -w'\gamma$ (i.e., $w'\gamma$ being large relative to the lower support boundary of $u|w$) for which $w_i'\gamma \simeq \infty$ is sufficient. The second-order (Hessian) matrix of SCL is

$$H \equiv E(1[|u| < w'\gamma]ww')$$

which becomes $E(ww')$ that is the second-order matrix of LSE when $|u| < w'\gamma$ (implied by $w'\gamma \simeq \infty$). If the censoring proportion becomes small, then SCL becomes close to LSE, and in this sense, SCL is a natural estimator to use for censored responses with a small censoring proportion. The main advantage of SCL over MLE's for the censored model is that SCL does not specify the distribution of $u|w$ and allows an unknown form of heteroskedasticity because the above moment condition does not require $u \amalg w$.

Powell (1986) suggested an iterative scheme to get $\hat{\gamma}$. Start with an initial estimate $\hat{\gamma}_0$, say LSE, and then iterate the following until convergence:

$$\hat{\gamma} = (\sum_i 1[w_i'\hat{\gamma}_0 > 0] \cdot w_i w_i')^{-1} \sum_i \{1[w_i'\hat{\gamma}_0 > 0] \min(y_{1i}, 2w_i'\hat{\gamma}_0) \cdot w_i\}.$$

This does not guarantee global convergence. Also the matrix to be inverted may not be invertible. If this problem occurs, then removing $1[w_i'\hat{\gamma}_0 > 0]$ in the inverted matrix may help. From our experience, however, this algorithm works well.

Going back to the case with endogenous y_2, let v_2 be either the additive or multiplicative residual from the y_2 RF. Then the second-stage in our two-stage procedure is SCL with w augmented by the CF $\hat{v}_2$ ($\hat{v}_2^2$ and $\hat{v}_2^3$). With the endogeneity of y_2 removed by the presence of the CF, SCL can be implement as above. The only modification needed is the asymptotic variance of SCL because the first stage estimation errors $\hat{\alpha}-\alpha$ and $\hat{\beta}-\beta$ affect the asymptotic variance through $\hat{v}_2$, which is to be examined in detail in the following subsection.

Our two-stage procedure works well computationally, because all estimators involved (logit, QPOI and SCL) converge well. This computational advantage should not be downplayed as it matters greatly in practice. Initially a generalized method of moment with $E(v_2|x) = 0$ was tried to estimate α and β but then scrapped later, as its convergence property was not so good.

2.4 Asymptotic Distribution

With w exogenous for y_1, the first- and second-order derivatives of the SCL minimand gives the following asymptotic linear expansion of SCL:

$$\begin{aligned}\sqrt{N}(\hat{\gamma}-\gamma) &= \frac{1}{\sqrt{N}}\sum_i H^{-1}\cdot 1[w_i'\gamma>0](1[|u_i|<w_i'\gamma]u_i + w_i'\gamma 1[|u_i|\geq w_i'\gamma])w_i + o_p(1)\\ &= \frac{1}{\sqrt{N}}\sum_i H^{-1}\zeta_i + o_p(1), \quad \text{where } \zeta_i \equiv 1[w_i'\gamma>0](1[|u_i|<w_i'\gamma]u_i + w_i'\gamma 1[|u_i|\geq w_i'\gamma])w_i.\end{aligned}$$

From this, it follows that

$$\sqrt{N}(\hat{\gamma}-\gamma) = N(0, H^{-1}E(\zeta\zeta')H^{-1}) \quad \text{where } E(\zeta\zeta') = E\{1[w'\gamma>0]\min(u^2,(w'\gamma)^2)\cdot ww'\}.$$

As already mentioned, in the two-stage procedure, the first-stage estimation errors $\hat{\alpha}-\alpha$ and $\hat{\beta}-\beta$ affect the SCL asymptotic variance through $\hat{v}_2$, which is discussed now.

Redefine w and γ as

$$w = (y_2, x_1', \hat{v}_2, \hat{v}_2^2, \hat{v}_2^3)' \quad \text{and} \quad \gamma = (\gamma_y, \gamma_x', \gamma_1, \gamma_2, \gamma_3)'$$

where $\hat{v}_2 = \hat{v}_2(\hat{\alpha},\hat{\beta})$ that depends on $\hat{\alpha}$ and $\hat{\beta}$ is either the additive or multiplicative residual, and $(\gamma_1,\gamma_2,\gamma_3)$ is the coefficient vector for $(\hat{v}_2,\hat{v}_2^2,\hat{v}_2^3)$.

Recalling the above asymptotic linear expansion of SCL, the presence of the first-stage estimators $\hat{\alpha}$ and $\hat{\beta}$ matters for its 'gradient vector' ζ, but not for the second-order matrix H. Hence write the asymptotic linear expansion as

$$\begin{aligned}\sqrt{N}(\hat{\gamma}-\gamma) &= \frac{1}{\sqrt{N}}\sum_i H^{-1}\zeta_i(\hat{\alpha},\hat{\beta}) + o_p(1) \\ &= \frac{1}{\sqrt{N}}\sum_i H^{-1}\{\zeta_i(\alpha,\beta) + E(\zeta_{\alpha'})\eta_{\alpha i} + E(\zeta_{\beta'})\eta_{\beta i}\} + o_p(1)\end{aligned}$$

where $\zeta_{\alpha'}$ and $\zeta_{\beta'}$ denote the derivatives of $\zeta(\alpha,\beta)$ for α and β, respectively, and $\eta_{\alpha i}$ and $\eta_{\beta i}$ are 'influence functions' for $\hat{\alpha}$ and $\hat{\beta}$:

$$\begin{aligned}\eta_{\alpha i} &= \{E(ss')\}^{-1}s_i \quad \text{for logit score function } s_i = \{y_{2i} - \frac{\exp(x_i'\alpha)}{1+\exp(x_i'\alpha)}\}x_i, \\ \eta_{\beta i} &= [E\{qxx'\exp(x'\beta)\}]^{-1}q_i x_i\{y_{2i} - \exp(x_i'\beta)\}.\end{aligned}$$

Since the dimension of γ is $(k_1+4)\times 1$ and the dimension of α and β are both $k\times 1$, $\zeta_{\alpha'}$ and $\zeta_{\beta'}$ are $(k_1+4)\times k$ matrices, which can be obtained by numerical differentiation. See, e.g., Lee (2010) for more details on this way of accounting for the first-stage estimation errors.

From the asymptotic linear expansion, it follows that

$$\sqrt{N}(\hat{\gamma}-\gamma) \rightsquigarrow N(0, H^{-1}E(\lambda_i\lambda_i')H^{-1}) \quad \text{where } \lambda_i \equiv \zeta_i(\alpha,\beta) + E(\zeta_{\alpha'})\eta_{\alpha i} + E(\zeta_{\beta'})\lambda_i.$$

$E(\lambda\lambda')$ can be estimated consistently by replacing (α,β,γ) with $(\hat{\alpha},\hat{\beta},\hat{\gamma})$ and the expected values in λ by the corresponding sample means. As already noted, if $E(y_2|x) = \exp(x'\beta)$ is adopted, then the only required change is redefining v_2 without the logit probability and then removing $E(\zeta_{\alpha'})\eta_{\alpha i}$ in λ. The endogeneity of y_2 can be tested using $(\hat{\gamma}_1,\hat{\gamma}_2,\hat{\gamma}_3)$, as their coefficients should be all zero under the null. It is convenient to note that, under the null, the first-stage estimation errors $\hat{\alpha}-\alpha$ and $\hat{\beta}-\beta$ can be ignored for SCL.

2.5 Details on Control Function

In practice, it may be enough for a CF to carry a significant estimate, and thus the results under y_2 exogeneity assumption differ much from those allowing y_2 endogeneity. But it would be more desirable to know what the CF looks like underneath and to justify it properly. Here we take a detailed look at the CF under more assumptions.

For an error term ε related to u and a parameter vector $\tilde{\beta}$, assume

$$E(y_2^*|q=1,x,\varepsilon) = \exp(x'\tilde{\beta}+\varepsilon) \quad \text{and} \quad \varepsilon \amalg (x,q).$$

This implies our earlier model assumptions

$$E(q=1|x,\varepsilon) = P(q=1|x) \ (= \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}),$$
$$E(y_2^*|q=1,x) = \int E(y_2^*|q=1,x,\varepsilon) f(\varepsilon|x,q=1)d\varepsilon = \exp(x'\tilde{\beta}) \int e^{\varepsilon} f(\varepsilon)d\varepsilon = \exp(x'\beta)$$

where β differs from $\tilde{\beta}$ in that the intercept in β absorbs $E(e^{\varepsilon}) = \exp\{\ln E(e^{\varepsilon})\}$.

The reason for the extra assumption on $E(y_2^*|q=1,x,\varepsilon)$ can be seen in

$$E\{y_2 - \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta) \ |x\} = E[\ E\{y - \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta)|\varepsilon,x\} \ |x]$$
$$= E[\ \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta)e^{\varepsilon} - \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta) \ |x]$$
$$= E[\frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta)\cdot(e^{\varepsilon}-1) \ |x] = 0.$$

That is, using the additive residual CF amounts to using

$$\frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta)\cdot(e^{\varepsilon}-1) \ \{\simeq \frac{\exp(x'\alpha)}{1+\exp(x'\alpha)}\exp(x'\beta)\varepsilon \text{ if } \varepsilon \text{ is small}\}.$$

Analogously, using the multiplicative residual CF amounts to using $e^{\varepsilon}-1$ ($\simeq \varepsilon$ if ε is small).

In the above extra assumption, since we need to have exogenous y_2, the relation of ε to u should be the only source for the y_2 endogeneity. A natural question to arise is how restrictive the assumption '$\varepsilon \amalg (x,q)$' is. Literally, it is restrictive in requiring that the y_2 endogeneity source ε be independent of the selection equation q as well as of x. But '$\varepsilon \amalg (x,q)$' does not imply '$q \amalg y_2^*|x$' that the selection equation q for y_2^* is independent of y_2 given x. To see a counter example, it is enough to think of generating an uniform random variable to use it along with (x,ε) to generate both q and y_2^*; through the same uniform random variable, q and y_2^* become related.

2.6 Two-Part Approach in the Literature

It seems helpful to compare our two-stage procedure to the two-part approach in the literature. The two-part approach assumed

$$\text{first part} \ : \ 1[y_1>0] = 1[\gamma_y y_2 + x_1'\gamma_x + u > 0] \quad \text{and} \quad y_2 = x'\delta + v$$
$$\text{second part} \ : \ y_1 = \xi_y y_2 + x_1'\xi_x + e_i \text{ given } y_1 > 0$$

where δ and ξ are parameters, and v and e are error terms.

For the first part, substitute $y_{2i} = x_i'\delta + v_i$ to obtain

$$1[y_{1i} > 0] = 1[\gamma_y(x_i'\delta + v_i) + x_i'S\gamma_x + u_i > 0] = 1[x_i'\psi + \gamma_y v_i + u_i > 0]$$
$$\text{where} \quad \psi \equiv \gamma_y\delta + S\gamma_x \quad \text{and} \quad S \text{ consists of 0's and 1's such that } x_1' = x'S.$$

Note that ψ is the RF parameters for $1[y_1 > 0]$ while (γ_y, γ_x) is the SF parameters. For the endogeneity of y_2 in the first part, a CF approach combined with minimum distance estimator (MDE) was used: the LSE residual $\hat{v}$ for the y_2 equation is used along with x to obtain $(\hat{\zeta}, \hat{\gamma}_y)$, and then (δ, γ_x) is estimated by MDE using $\hat{\psi} = \hat{\gamma}_y\delta + S\gamma_x$—simply imagine LSE of $\hat{\psi}$ on $(\hat{\gamma}_y, S)$ to estimate (δ, γ_x).

Some remarks on the two-part approach are in order. First, (δ, γ_x) can be estimated in the $1[y_1 > 0]$ SF with $\hat{v}$ controlled; no MDE is needed. Second, the linear model for y_2 is not plausible as y_2 has many zeros. Third, the second part of the two-part approach has been "sold" (relative to sample selection models) for a better prediction of y_1; hence the second part is not suitable to allow for endogenous regressors.

3 Empirical Analysis

Our data was drawn from the elderly of age 65 or above in 'the Korean Longitudinal Study of Ageing' for the year 2008. The information on the variables can be found in Table 1. In Table 1, 'formal' is the annual medical and long-term care expenditure in about $1000—the other amounts in the table are all annual amounts in the same unit. The number of care givers is our informal family care variable, 85% of which are zeros. Table 1 also shows yearly informal care hours ('care hours') of which 85% are zeros again, but this variable will not be used for y_2—the estimation results with care hours as y_2 is mostly insignificant with no endogeneity of y_2 picked up by the CF's.

'fi. asset' is financial asset amount, and 'real est.' is real asset amount. 'own house' is the dummy for owning a house. 'fam.inc.' is household income, and pension is pension and other welfare receipt amount. 'hi.bl. pressure' is the dummy for high blood pressure. 'cancer/tumor' is the dummy for cancer or malign tumor. 'chronic pulmo.' is the dummy for chronic pulmonary disease. 'chronic liver' is the dummy for chronic liver disease. 'cerebral bl.vessel' is the dummy for cerebral blood vessel disease. 'arthritis/rheuma.' is the dummy for arthritis or rheumatism. 'male' is the dummy for being male', 'Seoul' is the dummy for

living in Seoul, and 'work' is the dummy for working. 'kid-par' is the transfer amount from children to the parents. 'nkids' is the number of children and 'nfem.kids' is the number of female children. 'nkids-co' is the number of children cohabiting with the respondent, and 'nkids-act' is the number of children economically active. 'nkids-30' is the number of non-cohabiting children living in 1-30 minutes' distance by public transportation; nkids-60 and nkids-120 are analogously defined for 31-60 minutes and 61-120 minutes, respectively. '# generations' is the number of generations living together.

Table 1: Descriptive Statistics

Variable	Mean (SD)	Min,Max	Variable	Mean (SD)	Min,Max
formal ($1,000)	1.179 (2.34)	0, 48.4	age	74.6 (6.12)	65, 107
# care givers	0.215 (0.58)	0, 4	male	0.425 (0.494)	0, 1
care hours	157 (619)	0, 8760	married	0.636 (0.481)	0, 1
fi. asset ($1,000)	4.88 (21.6)	0, 500	Seoul	0.137 (0.343)	0, 1
real est. ($1,000)	152 (222)	0, 2948	work	0.213 (0.409)	0, 1
own house	0.409 (0.49)	0, 1	kid-par ($1,000)	13.5 (28.2)	0, 866
fam.inc. ($1,000)	16.3 (21.0)	0, 700			
pension ($1,000)	1.42 (4.44)	0, 94.9	nkids	3.99 (1.61)	0, 10
hi.bl. pressure	0.091 (0.288)	0, 1	nfem.kids	1.92 (1.40)	0, 8
diabetes	0.048 (0.215)	0, 1	nkids-co	0.412 (0.56)	0, 3
cancer/tumor	0.013 (0.114)	0, 1	nfem.kids-co	0.092 (0.30)	0, 3
chronic pulmo.	0.016 (0.127)	0, 1	nkids-act	2.61 (1.41)	0, 8
chronic liver	0.005 (0.073)	0, 1	nfem.kids-act	0.765 (0.97)	0, 7
cardio disease	0.035 (0.183)	0, 1	nkids-30	0.597 (0.99)	0, 6
cerebral bl.vessel	0.038 (0.191)	0, 1	nkids-60	0.838 (1.18)	0, 6
mental disease	0.016 (0.125)	0, 1	nkids-120	0.768 (1.22)	0, 9
arthritis/rheuma.	0.195 (0.396)	0, 1	# generations	1.48 (1.06)	0, 4

To avoid extreme values in the amount variables, all amount variables are transformed with $\ln(\cdot+1)$ so that 0 remains 0 after the transformation and positive values remain positive after transformation. Other than the variables in Table 1, self-reported health status is also available in five categories. When health status was used for estimation, its coefficient was significantly positive, implying that health status is likely to be affected by formal/informal

care, and thus it cannot be used as a regressor. Although the children-related variables can be used as instruments (IV) for y_2, there is no good IV for health status. Hence health status is dropped from the regressor list. By omitting health status, the endogeneity of y_2 becomes more likely.

Table 2: Logit and Quasi-Poisson for y_2

Variables	Logit (t-value)	QPOI (t-value)
financial asset	-0.034 (-1.53)	-0.012 (-1.35)
real estate	0.011 (0.26)	-0.007 (-0.41)
own hose	-0.245 (-1.63)	-0.107 (-1.84)
family income	0.057 (1.06)	0.031 (1.50)
pension	0.026 (0.91)	-0.012 (-1.25)
age	-0.068 (-0.45)	0.006 (0.10)
age2	0.109 (1.16)	0.000 (0.00)
male	0.661 (4.01)	0.029 (0.47)
married	0.119 (0.70)	-0.025 (-0.31)
Seoul	-0.707 (-3.68)	0.126 (1.91)
work	-0.820 (-3.80)	-0.109 (-1.40)
kid-par	-0.052 (-2.54)	-0.006 (-0.70)
nkids	0.225 (2.07)	0.024 (0.63)
nfem.kids	-0.180 (-1.63)	0.003 (0.09)
nkids-co	0.097 (0.60)	0.084 (1.36)
nfem.kids-co	0.349 (1.74)	0.057 (0.89)
nkids-act	-0.150 (-1.47)	-0.028 (-0.74)
nfem.kids-act	0.010 (0.08)	-0.127 (-2.75)
nkids-30	0.040 (0.60)	0.046 (2.05)
nkids-60	0.022 (0.41)	0.049 (2.43)
nkids-120	-0.033 (-0.55)	-0.009 (-0.42)
# generations	0.227 (2.92)	0.050 (1.64)

Table 2 'Logit and Quasi-Poisson for y_2' presents the estimates for the first-stage. Since most disease variables are highly significant but of no direct interest, we omit their results in Table 2 and the remaining tables to simplify presentation; also omitted are the intercept

estimates. In Table 2, age^2/100 ('age2') is used. The main variable of interest are the children-related variables as they are the IV's for y_2 and thus should be significant in explaining y_2. 'nkids' and # generations are significant for logit, whereas nfem.kids-act, nkids-30 and nkids-60 are significant for QPOI.

Table 3: SCL, CFE-additive and CFE-multiplicative for y_1

Variables	SCL (tv)	CFEa (tv2, tv1)	CFEm (tv2, tv1)
y_2	2.135 (2.40)	1.172 (0.98, 1.05)	1.757 (0.16, 1.71)
y_2×hi.bl. pressure	-0.275 (-2.08)	-0.162 (-1.11, -1.14)	-0.248 (-0.28, -1.81)
y_2×diabetes	-0.686 (-3.81)	-0.668 (-3.56, -3.68)	-0.673 (-0.52, -3.68)
y_2×mental disease	-0.605 (-1.88)	-0.461 (-1.42, -1.50)	-0.575 (-0.86, -1.77)
y_2×arthritis/rheuma.	0.125 (0.83)	0.123 (0.79, 0.80)	0.133 (0.19, 0.88)
y_2×age	-0.026 (-2.32)	-0.020 (-1.65, -1.70)	-0.022 (-0.19, -1.78)
y_2×male	0.191 (1.21)	0.237 (1.41, 1.45)	0.201 (0.39, 1.28)
financial asset	0.047 (3.42)	0.046 (3.29, 3.31)	0.047 (2.81, 3.40)
real estate	0.159 (4.64)	0.159 (4.71, 4.76)	0.158 (3.81, 4.62)
own hose	-0.029 (-0.30)	-0.046 (-0.44, -0.44)	-0.032 (-0.28, -0.32)
family income	0.001 (0.03)	0.009 (0.27, 0.27)	0.001 (0.03, 0.05)
pension	0.068 (3.48)	0.068 (3.52, 3.52)	0.068 (3.03, 3.50)
age	0.378 (2.29)	0.348 (2.40, 2.43)	0.380 (1.44, 2.31)
age2	-0.262 (-2.43)	-0.239 (-2.49, -2.53)	-0.263 (-1.50, -2.45)
male	-0.136 (-1.15)	-0.115 (-0.93, -0.94)	-0.137 (-1.05, -1.16)
married	0.093 (0.91)	0.088 (0.86, 0.86)	0.091 (0.65, 0.89)
Seoul	-0.006 (-0.05)	-0.031 (-0.24, -0.25)	-0.006 (-0.04, -0.05)
work	-0.184 (-1.63)	-0.213 (-1.83, -1.84)	-0.187 (-1.46, -1.65)
kid-par	0.026 (1.78)	0.023 (1.48, 1.49)	0.026 (1.41, 1.77)
$\hat{v}_2$		0.414 (0.97, 1.10)	0.027 (0.03, 0.74)
$\hat{v}_2^2$		0.230 (1.85, 2.00)	-0.001 (0.00, -0.85)
$\hat{v}_2^3$		-0.069 (-2.06, -2.19)	0.000 (0.00, 1.10)

Table 3 presents the main estimation results where 'tv' stands for t-value, CFEa is the estimator with the additive error for CF, CFEm is the estimator with the multiplicative error for CF, and 'tv2' is the correct t-value taking into account the first-stage estimation errors

whereas 'tv1' is the t-value ignoring the first-stage estimation errors (correct only under the null of no y_2 endogeneity). For the sake of comparison, we show the SCL results ignoring the y_2 endogeneity in the first column, although we will not interpret the results.

Comparing CFEa and CFEm in Table 3, CFEm does not pick up the y_2 endogeneity as the CF $(\hat{v}_2, \hat{v}_2^2, \hat{v}_2^3)$ are all insignificant—the Wald test for $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ is not rejected. In contrast, CFEa does pick up the y_2 endogeneity, which results in appreciable differences between SCL and CFEa in the estimates involving y_2. In the CFEa column, among the terms involving y_2, only the interaction term with diabetes is significant with a large effect estimate; there is also weak evidences that y_2 interacts with mental disease, age and male.

Also notable in the CFEa column of Table 3 is that tv2 and tv1 are not much different: there is no reversal of statistical significance except for $\hat{v}_2^2$ where tv2 is 1.85 while tv1 is 2.00. In contrast, tv2 and tv1 are much different in CFEm, particularly for the variables involving y_2 and $\hat{v}_2$. This might be due to the division of y_2 by the regression function for the multiplicative residual, as this might result in too big numbers and consequently some numerical instability. The poor performance of CFEm relative to CFEa is somewhat surprising, given the intuitive appeal of the multiplicative residual in the exponential model. This might be attributed to two factors: the just mentioned numerical instability, and u containing the heteroskedastic factor present in the additive residual, but not in the multiplicative residual.

Table 4 presents the estimation results under $E(y_2|x) = \exp(x'\beta)$ which does away with logit. In Table 4, neither CFEa nor CFEm pick up the y_2 endogeneity in view of the t-values for the CF's. As the result, the estimates and t-values of CFEa and CFEm are not much different from those of SCL which ignores the y_2 endogeneity. As in Table 3, tv2 and tv1 are little different in CFEa, whereas they are substantially different for CFEm, particularly for the variables involving y_2 and $\hat{v}_2$.

Although not shown, we also tried the 'logit-only first stage' just to see which part between logit and QPOI contributes more. The results for the mean squared error $N^{-1}\sum_i (y_{2i} - \hat{y}_{2i})^2$ where $\hat{y}_{2i}$ is the estimated $E(y_2|x)$ are, respectively, 0.284 (QPOI only), 0.283 (logit only), and 0.271 (both QPOI and logit as in the main two-stage procedure). This shows that most explanatory power for y_2 comes from its binary aspect and the positive values contribute only a little.

Table 4: SCL, CFE-additive and CFE-multiplicative for y_1: No Logit

Variables	SCL (tv)	CFEa (tv2, tv1)	CFEm (tv2, tv1)
y_2	2.135 (2.40)	1.816 (1.40, 1.57)	1.413 (0.55, 1.31)
y_2×hi.bl. pressure	-0.275 (-2.08)	-0.250 (-1.71, -1.80)	-0.223 (-0.49, -1.62)
y_2×diabetes	-0.686 (-3.81)	-0.674 (-3.67, -3.78)	-0.680 (-0.94, -3.63)
y_2×mental disease	-0.605 (-1.88)	-0.620 (-1.90, -1.88)	-0.550 (-1.03, -1.68)
y_2×arthritis/rheuma.	0.125 (0.83)	0.131 (0.82, 0.85)	0.146 (0.26, 0.96)
y_2×age	-0.026 (-2.32)	-0.025 (-2.00, -2.05)	-0.019 (-0.69, -1.48)
y_2×male	0.191 (1.21)	0.195 (1.15, 1.18)	0.216 (0.53, 1.37)
financial asset	0.047 (3.42)	0.047 (3.35, 3.37)	0.047 (3.13, 3.41)
real estate	0.159 (4.64)	0.159 (4.55, 4.60)	0.158 (4.49, 4.72)
own hose	-0.029 (-0.30)	-0.033 (-0.32, -0.33)	-0.032 (-0.30, -0.33)
family income	0.001 (0.03)	0.004 (0.11, 0.11)	0.002 (0.07, 0.07)
pension	0.068 (3.48)	0.069 (3.52, 3.53)	0.068 (3.23, 3.50)
age	0.378 (2.29)	0.362 (2.03, 2.08)	0.372 (1.77, 2.25)
age2	-0.262 (-2.43)	-0.250 (-2.11, -2.17)	-0.258 (-1.85, -2.38)
male	-0.136 (-1.15)	-0.132 (-1.08, -1.09)	-0.139 (-1.11, -1.18)
married	0.093 (0.91)	0.097 (0.94, 0.94)	0.090 (0.78, 0.88)
Seoul	-0.006 (-0.05)	-0.015 (-0.12, -0.12)	-0.009 (-0.06, -0.07)
work	-0.184 (-1.63)	-0.199 (-1.65, -1.70)	-0.185 (-1.58, -1.62)
kid-par	0.026 (1.78)	0.025 (1.58, 1.62)	0.026 (1.53, 1.75)
$\hat{v}_2$		0.183 (0.37, 0.50)	0.074 (0.10, 1.47)
$\hat{v}_2^2$		0.102 (0.92, 1.50)	-0.005 (0.00, -1.58)
$\hat{v}_2^3$		-0.027 (-0.80, -1.44)	0.000 (0.00, 1.76)

4 Conclusions

This paper examined whether informal health care can reduce formal health care, where the formal care y_1 is medical and long-term care expenditures (14% zeros) and the informal care y_2 is the number of family care givers (85% zeros). This task posed a number of difficulties, because y_2 is an endogenous regressor that is a count variable with too-many zeros, in addition to y_1 having a non-trivial proportion of zeros.

Facing the difficulties, we proposed a two-stage procedure where the first stage is estimating $E(y_2|x)$ as the product of logit (using y_2 being positive or not) and an exponential regression function (using only positive y_2's)—the idea borrowed from 'zero-inflated Poisson'. The second stage is applying a semi-parametric censored model estimator for y_1 with the endogeneity of y_2 removed by a control function (CF). Two types of CF's were considered: one based on the additive residual $y_2 - E(y_2|x)$, and the other based on the multiplicative residual $y_2/E(y_2|x) - 1$; the actual CF used was polynomial functions of these residuals.

Despite the intuitive appeal of the multiplicative residual as an exponential function appears, the additive residual CF approach performed much better than the multiplicative residual CF approach. Also, using only an exponential function for $E(y_2|x)$ (i.e., ignoring the too-many zero problem) was tried, but the outcome was inferior to the procedure with both logit and exponential functions.

Our empirical result using Korean data for the elderly (of age 65 and above) showed that informal care is a substitute only for certain cases such as diabetes. There are weak evidences that informal care effect on formal care interacts also with mental disease, age and male. That is, as noted in the literature of informal and formal care trade-off, the effect of informal care on formal care is heterogeneous.

REFERENCES

Bolin, K., Lindgren, B. and P. Lundborg, 2008, Informal and formal care among single-living elderly in Europe, Health Economics 17, 393-409.

Bonsang, E., 2009, Does informal care from children to their elderly parents substitute for formal care in Europe?, Journal of Health Economics 28, 143-154.

Charles, K. and P. Sevak, 2005, Can family caregiving substitute for nursing home care?, Journal of Health Economics 24, 1174-1190.

Lambert, D., 1992, Zero-inflated Poisson regression, with an application to defects in manufacturing, Technometrics 34, 1–14.

Lee, M.J., 1992, Winsorized mean estimator for censored regression model, Econometric Theory 8, 368-382.

Lee, M.J., 2010, Micro-econometrics: methods of moments and limited dependent variables, Springer.

Lee, M.J., 2011, Treatment effects in sample selection models and their nonparametric estimation, Journal of Econometrics, forthcoming.

Lee, M.J., 2012, Semiparametric estimators for limited dependent variable (LDV) models with endogenous regressors, Econometric Reviews, forthcoming.

Powell, J.L., 1984, Least absolute deviations estimation for the censored regression model, Journal of Econometrics 25, 303-325.

Powell, J.L., 1986, Symmetrically trimmed least squares estimation for Tobit models, Econometrica 54, 1435-1460.

Van Houtven C.H. and E.C. Norton, 2004, Informal care and health care use of old adults, Journal of Health Economics 23, 1159–1180.