ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)
69, *Part* 4, *pp.* 953–971

Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models

Giampiero Marra,

University College London, UK

Rosalba Radice

City University of London, UK

and David M. Zimmer

Western Kentucky University, Bowling Green, USA

[Received April 2019. Final revision April 2020]

Summary. The paper estimates the causal effect of having health insurance on healthcare utilization, while accounting for potential endogeneity bias. The topic has important policy implications, because health insurance reforms implemented in the USA in recent decades have focused on extending coverage to the previously uninsured. Consequently, understanding the effects of those reforms requires an accurate estimate of the causal effect of insurance on utilization. However, obtaining such an estimate is complicated by the discreteness inherent in common measures of healthcare usage. The paper presents a flexible estimation approach, based on copula functions, that consistently estimates the coefficient of a binary endogenous regressor in count data settings. The relevant numerical computations can be easily carried out by using the freely available GJRM R package. The empirical results find significant evidence of favourable selection into insurance. Ignoring such selection, insurance appears to increase doctor visit usage by 62% but, adjusting for it, the effect increases to 134%.

Keywords: Binary endogenous regressor; Copula; Count data; Moral hazard; Penalized regression spline; Simultaneous estimation

1. Introduction

The Affordable Care Act, which was passed by the US Congress and signed into law by President Obama in 2010, represented one of the largest expansions of the US social safety net since the 1960s. Through a combination of mandates, subsidies and regulations, the law's primary goal was to extend health insurance coverage to the approximately 44 million Americans who, before the law's passage, lacked coverage. Such a large expected increase in insurance coverage, in turn, raised questions about whether the newly insured would respond by increasing their usage of medical services, possibly straining existing health services infrastructure. Concerns about medical infrastructure have recently reached new heights in the midst of the 2020 coronavirus pandemic.

Estimating the effect of insurance on healthcare usage is surprisingly difficult, because of

Address for correspondence: Giampiero Marra, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail: giampiero.marra@ucl.ac.uk

the possible simultaneous relationship between a person's demand for health insurance and his utilization of healthcare services (Arrow, 1963). This paper presents a case-study which estimates the causal effect of having health insurance (a binary variable) on healthcare utilization (a count variable), while accounting for potential simultaneity bias. Health services researchers have historically ignored such simultaneity, effectively treating insurance status as predetermined, similarly to age or race. During the mid-to-late 1980s, however, scholars began to address the possibility that some people seek insurance *in anticipation of* future healthcare needs.

Ideally, the effect of insurance on healthcare usage should be assessed via a randomized controlled experiment, with subjects randomly assigned to insurance states. The famous RAND experiment (Manning *et al.*, 1987) represented an ambitious effort for this, but its findings are more than 40 years old. Alternatively, one could wait for a 'natural' experiment to emerge, as sometimes happens in the USA when individual states change insurance laws, which, in turn, might create an environment that mimics random insurance assignment (Finkelstein *et al.*, 2012).

More often, however, researchers must rely on observational data, which opens up concerns about simultaneity of insurance and healthcare usage. Cameron *et al.* (1988) formalized that simultaneity in an economic model that we present here in abbreviated form. Consider a person who chooses health insurance and healthcare usage in two stages. In the first stage, before a potential future health event is known, the person decides whether to acquire insurance. In the second stage, after knowing whether the health event has occurred, the person finds his optimal level of healthcare usage.

Focus first on the second stage, in which the person treats as given his individual characteristics T, insurance status (with $j=1$ denoting insured and $j=0$ denoting uninsured) and a potential health event s. Here, the person seeks to maximize utility subject to his budget constraint,

$$\max_{C(s),e(s)} U_j[C(s), H\{e(s), s\}|T] \text{ subject to } C(s) + P_j + p_j e(s) \leqslant I,$$

where $C(s)$ represents the person's consumption of goods that directly contribute to utility (e.g. pizza and vacations) in health event s, and $H(e, s)$ is the person's health level (in income equivalent), which itself is a function of health event s and healthcare utilization $e(s)$, which itself also depends on s. In the budget constraint, P_j represents the price (to the person) of insurance, which is usually unobserved in most sources of data, unless the person is uninsured, in which case it equals 0. Term p_j, which depends on insurance status and is usually also unobserved, denotes the unit price of healthcare services. Term I represents disposable income.

Rewinding to the first stage when the possible health event has yet to materialize, the person chooses his insurance status to maximize the expected utility

$$\int_s U_j[C^*(s), H\{e^*(s), s|T\}]\pi(s|T)\,\mathrm{d}s,$$

where C^* and e^* are optimal consumption and healthcare usage obtained from the second-stage problem, and π denotes the probability density of future health events. If the integral is larger while insured ($j=1$) than not ($j=0$), then the person insures.

Cameron *et al.* (1988) showed that, with properly specified functional forms, one can derive a demand function for medical care, with insurance status as a key argument. However, the main point of their model is that insurance status cannot, in general, be treated as predetermined. The reason is that insurance status, which is determined in the first stage, depends, in part, on the person's expected healthcare usage in the second stage, $e^*(s)$. Consequently, such selection into insurance is likely to muddle the observed link between insurance and healthcare usage.

Researchers have attempted to address this concern by using various techniques, most of

which can be regarded as variants of the modelling framework that is discussed in Section 2 (e.g. Goldman (1995), Cardon and Hendel (2001), Mello *et al.* (2002), Deb and Trivedi (2006) and Zheng and Zimmer (2008)). For the most part, existing studies find that insurance increases healthcare usage, but there seems to be widespread disagreement about the extent to which the simultaneity of insurance and usage affects those findings.

When unobserved information—in this case, knowledge of future health needs—simultaneously affects both the treatment (insurance) and the outcome (healthcare usage), statisticians refer to the treatment as 'endogenous' (Cameron and Trivedi (2013), pages 386–388). Ignoring endogeneity might produce incorrect estimates of the causal effect of insurance. For example, if relatively unhealthy people, knowing that they will need to consume healthcare services, seek insurance coverage to help to pay for those services, then insurance will appear positively linked to healthcare usage. But such relationships cannot be interpreted as causal, because (at least part of) the positive link owes to unhealthy people selecting into insurance coverage. Because healthcare policy in the USA in recent decades has focused on extending coverage to the previously uninsured, obtaining an accurate estimate of the causal effect of insurance on usage is crucial to determine whether such reforms will stretch existing healthcare infrastructure.

From a more general statistical modelling perspective, many microeconometric models, especially those that rely on observational data, are plagued by unobserved heterogeneity that simultaneously correlates with the outcome variable and a right-hand-side explanatory variable of interest. When the outcome variable is continuous with a distributional shape that lends itself to linear regression modelling, the standard approach involves finding 'instruments' that correlate with the endogenous right-hand-side regressor, but not with the outcome variable. In the healthcare usage case-study, a possible instrument might be firm size, which is likely to correlate with a person's probability of having private insurance, but should not (directly) alter healthcare usage. When such instruments are available, the method of instrumental variables, which is also known as two-stage least squares, can be employed; see, for example, Greene (2008), chapter 12, for a textbook treatment of instrumental variables methods. But when the outcome variable has discreteness or shows distributional patterns that call for non-linear models (such as generalized linear and additive models), the two-stage least squares method no longer yields consistent estimates. A number of studies that have sought to quantity the effect of insurance on healthcare usage, some of which are mentioned below, have introduced methods that attempt to import the logic of two-stage least squares to more general settings. Unfortunately, as discussed in the following section, those approaches either lack general applicability and/or may require substantial computational resources to implement.

This paper emphasizes cross-sectional settings, where the main focus centres on consistent estimation of the coefficient of a binary endogenous explanatory variable. Throughout, we assume that the researcher has access to valid instruments. As such, the paper does not consider panel data methods for addressing endogeneity, such as fixed effects or difference in differences; nor does it explore instrument-free methods, such as matching (Rosenbaum and Rubin, 1983), synthetic control (Kreif *et al.*, 2016) or approaches that exploit quirks in the higher order moments of the distribution of the outcome variable (Lewbel, 2012).

The primary goal of the paper is to investigate the aforementioned case-study by using a general and flexible estimation approach, based on copula functions, for consistently estimating the coefficient of a binary endogenous regressor in count data settings. Some existing studies have attempted different versions of the method that we discuss (e.g. Han and Vytlacil (2017), Park and Gupta (2012), Radice *et al.* (2016), Tran and Tsionas (2015), Winkelmann (2012) and Zimmer (2018)). The work by Zimmer (2018) is the closest to ours since it introduces a copula

model with binary and Poisson margins, and with an endogenous binary variable. However, our proposal is far more general in that it allows for a wider range of discrete distributions for the count variable, for several link functions for the binary margin, for the specification of flexible covariate effects, for the use of a wider set of copula functions and for the dependence parameter to be specified as a function of covariates. Using our method, we find statistically significant evidence that insurance is endogenous with respect to usage of doctors' services, and that, when endogeneity is taken into account, the effect of insurance is larger than when endogeneity is ignored. Health economists refer to such a pattern as 'favourable selection', with relatively light users predisposed towards being insured.

The paper also highlights the newly revised software package `GJRM` (Marra and Radice, 2020), written for the programming language R (R Core Team, 2020), which greatly eases the implementation of our model. To the best of our knowledge, there are no alternative copula regression models, nor respective software implementations, of the type that is discussed in this paper. Although the construction and estimation of the model proposed rely on many of the modular functions and routines that are already available in `GJRM`, extending the software to accommodate the model, the functional forms and nuances that are needed to address the aforementioned case-study required a large amount of programming work. Those developments made it possible to estimate the flexible class of models that are discussed in this paper and hence allow for more flexible specifications than are possible by using extant methods.

1.1. Existing methods

Instrument-based approaches for addressing endogeneity in non-standard settings fall into two main categories: two-stage techniques and simultaneous estimation methods. The simplest two-stage procedure, which also most closely resembles linear two-stage least squares, is the 'control function' approach (Heckman and Robb, 1985; Terza *et al.*, 2008). The first stage involves regressing the endogenous variable on all explanatory variables, including instruments. This regression is then used to calculate residuals, which, in the second stage, appear alongside the endogenous variable as an additional regressor in the main regression of interest. The control function method is simple and quite general, but it encounters problems when the endogenous variable is not continuous (Wooldridge (2010), page 746). In our case-study the endogenous variable is binary and this muddles what exactly constitutes a 'residual' from the first stage, with different types of definition potentially leading to conflicting results.

A second category of approaches, labelled 'simultaneous estimation methods', attempt to assemble the full joint distribution of the outcome variable and the endogenous regressor. The joint distribution is then typically used for likelihood-based estimation; see, for example, Terza (1998) and Cameron and Trivedi (2013), pages 385–412. Such approaches usually decompose the (unknown) joint distribution into the marginal distribution of the outcome conditionally on the endogenous regressor and the marginal distribution of the endogenous regressor. These methods, however, can be quite computationally intensive if, as is often assumed, the two marginals share an unobserved random effect, which then must be integrated out. For example, Deb and Trivedi (2006) reported that one variant of this approach, based on simulated maximum likelihood, shows 'quite slow' convergence, which led them to suggest several simulation acceleration tricks, including alternative methods for drawing random numbers. Indeed, the simulated maximum likelihood based code that was generously provided by Deb and Trivedi (2006)—complete with their 'accelerators'—applied to the case-study that is explored in this paper required almost 30 min to converge on a desktop computer, compared with mere seconds for our proposed copula-based method.

The approach that is proposed in this paper falls into the 'simultaneous estimation methods' category, but it side-steps the numerical integration obstacle by joining (via copulas) the two marginal distributions, yielding a closed form expression for the likelihood function. Consequently, parameter estimation and inference are far less computationally taxing.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`.

2. Methodology

This section discusses a recursive copula additive model to estimate the effect of a binary endogenous variable on a count outcome. Details on identification, parameter estimation and software implementation are also provided.

2.1. The model

To simplify the notation, and without loss of generality, we drop observation index i, while noting that n observations are available for modelling. We begin by assuming that the joint cumulative distribution function (CDF) of a binary (endogenous) variable and a discrete outcome variable, $Y_1 \in \{0, 1\}$ and $Y_2 \in \mathbb{N}_0$ respectively can be expressed as

$$F_{12}(y_1, y_2|\boldsymbol{\vartheta}) = C\{F_1(y_1|\pi), F_2(y_2|\mu, \sigma); \theta\}, \tag{1}$$

where $\boldsymbol{\vartheta} = (\pi, \mu, \sigma, \theta)'$. Terms $F_1(y_1|\pi)$ and $F_2(y_2|\mu, \sigma)$ denote marginal CDFs of Y_1 and Y_2 taking values in $(0, 1)$, whereas the symbols π, μ and σ represent marginal distributional parameters. Function $C : (0, 1)^2 \to (0, 1)$ is a two-place copula function which does not depend on the marginals, and θ is an association copula parameter measuring the dependence between the two random variables.

A substantial advantage of the copula approach is that a joint CDF can be conveniently formed by utilizing two (in this case) arbitrary univariate marginal CDFs and a function C that binds them together. As opposed to what is found in classical copula regression settings, in this work the binary variable y_1 appears as an explanatory variable inside μ in the marginal $F_2(y_2|\mu, \sigma)$. Thus, the copula has a recursive structure. This recursive structure implies that y_1 is endogenous with respect to y_2 if dependence between the two marginals, as captured by θ, is statistically significant. See Han and Vytlacil (2017) and references therein for some works which have adopted the same logic in a copula regression context.

The copulas that were implemented in `GJRM` are reported in Table 1. Table 1 also shows the relationship between θ and Kendall's τ-coefficient, which is a measure of association that lies in the customary range $\tau \in [-1, 1]$. Counterclockwise rotated versions of the Clayton, Gumbel and Joe copulas are obtained by using the formulae in Brechmann and Schepsmeier (2013). For more details on copulas see, for example, Nelsen (2006). In the current setting, the result of Sklar (1973) can only guarantee that the copula is unique over the range of the outcomes. In a regression context, however, this potential issue is less likely to be a concern as noted by several researchers including Joe (2014), Nikoloulopoulos and Karlis (2010) and Trivedi and Zimmer (2017), primarily because regression structures in the marginals generate additional variation in the outcomes and thus more completely cover the outcome domains.

In equation (1), the marginals for Y_1 and Y_2 are assumed to be specified via one- and two-

Table 1. Definition of the copulas that are implemented in R package GJRM, with corresponding parameter range of association parameter θ, one-to-one transformation function of θ and relationship between Kendall's τ and θ†

Copula	*$C(p_1, p_2; \theta)$*	*Range of θ*	*Transformation*	*Kendall's τ*
AMH (`"AMH"`)	$p_1 p_2/\{1-\theta(1-p_1)(1-p_2)\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$-\{2/(3\theta^2)\}\{\theta + (1-\theta)^2 \times \log(1-\theta)\} + 1$
Clayton (`"C0"`)	$(p_1^{-\theta} + p_2^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$	$\log(\theta)$	$\theta/(\theta+2)$
FGM (`"FGM"`)	$p_1 p_2\{1+\theta(1-p_1)(1-p_2)\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{9}\theta$
Frank (`"F"`)	$-\theta^{-1}\log[1 + \{\exp(-\theta p_1) - 1\}]$ $\{\exp(-\theta p_2) - 1\}/\{\exp(-\theta) - 1\}$	$\theta \in \mathbb{R}\backslash\{0\}$	—	$1 - (4/\theta)\{1 - D_1(\theta)\}$
Gaussian (`"N"`)	$\Phi_2\{\Phi^{-1}(p_1), \Phi^{-1}(p_2); \theta\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$(2/\pi)\sin^{-1}(\theta)$
Gumbel (`"G0"`)	$\exp(-[\{-\log(p_1)\}^\theta$ $+\{-\log(p_2)\}^\theta]^{1/\theta})$	$\theta \in [1, \infty)$	$\log(\theta - 1)$	$1 - 1/\theta$
Joe (`"J0"`)	$1 - \{(1-p_1)^\theta + (1-p_2)^\theta$ $-(1-p_1)^\theta(1-p_2)^\theta\}^{1/\theta}$	$\theta \in (1, \infty)$	$\log(\theta - 1)$	$1 + (4/\theta^2) D_2(\theta)$
Plackett (`"PL"`)	$(Q - \sqrt{R})/\{2(\theta - 1)\}$	$\theta \in (0, \infty)$	$\log(\theta)$	—
Student t (`"T"`)	$t_{2\zeta}\{t_\zeta^{-1}(p_1), t_\zeta^{-1}(p_2); \zeta, \theta\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$(2/\pi)\sin^{-1}(\theta)$

†$\Phi_2(\cdot, \cdot; \theta)$ denotes the CDF of a standard bivariate normal distribution with correlation coefficient θ, and $\Phi(\cdot)$ the CDF of a univariate standard normal distribution. $t_{2\zeta}(\cdot, \cdot; \zeta, \theta)$ indicates the CDF of a standard bivariate Student t-distribution with correlation θ and fixed $\zeta \in (2, \infty)$ degrees of freedom, and $t_\zeta(\cdot)$ denotes the CDF of a univariate Student t-distribution with ζ degrees of freedom. $D_1(\theta) = (1/\theta)\int_0^\theta [t/\{\exp(t) - 1\}]\,\mathrm{d}t$ is the Debye function and $D_2(\theta) = \int_0^1 t\log(t)(1-t)^{2(1-\theta)/\theta}\mathrm{d}t$. Quantities Q and R are given by $1 + (\theta - 1)(p_1 + p_2)$ and $Q^2 - 4\theta(\theta - 1)p_1 p_2$ respectively. Kendall's τ for `"PL"` is computed numerically as no analytical expression is available. The argument `BivD` of `gjrm()` in `GJRM` enables the user to employ the desired copula function and can be set to any of the values within parentheses next to the copula names in the first column; for example, `BivD = "J0"`. For the Clayton, Gumbel and Joe copulas, the number after the capital letter indicates the degree of rotation required: the possible values are `0`, `90`, `180` and `270`. AMH, Ali–Mikhail–Haq; FGM, Farlie–Gumbel–Morgenstern.

parameter distributions respectively: hence the notation that is adopted. However, the computational framework can be conceptually easily extended to distributions with more parameters. For the binary endogenous variable Y_1, we have considered the Bernoulli distribution with parameter $\pi \in [0, 1]$ (representing the probability of 'success'). For the outcome variable, Y_2, `GJRM` has been extended to include four possible choices for discrete distributions; see Table 2 for the expressions of their probability mass functions (PMFs) f, expectations and variances.

Finally, each model's parameter can be related to covariates and regression coefficients via an additive predictor $\eta \in \mathbb{R}$ (defined in generic terms in the next paragraph) and a known monotonic one-to-one transformation function (which ensures that the restriction on the respective parameter space is not violated). For example, if we wish to employ a Gumbel copula model with Bernoulli and Poisson margins and would like to express π, μ and θ as functions of additive predictors then $g_\pi(\pi) = \eta_\pi$, $g_\mu(\mu) = \eta_\mu$ and $g_\theta(\theta) = \eta_\theta$, where $g_\pi(\cdot) = \mathrm{logit}(\cdot)$, $g_\mu(\cdot) = \log(\cdot)$ and $g_\theta(\cdot) = \log(\cdot - 1)$. For the binary margin, in this example, we have assumed a logit link function; however the probit and cloglog-functions have also been implemented.

Note that specifying the dependence parameter as a function of covariates allows for the strength of the (upper tail, in the above example) dependence between the marginals to vary across observations (e.g. Marra and Radice (2017), and references therein). Furthermore, because the dependence parameter in our proposed set-up captures the magnitude (and direction) of endogeneity, specifying the dependence term as a function of covariates allows endogeneity to vary across observations, which is something that has not been previously explored. In

Table 2. Definition and some properties of the discrete distributions implemented in GJRM†

	$f(y\|\mu,\sigma)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$
Poisson (`"PO"`)	$\frac{\exp(-\mu)\mu^y}{y!}$	μ	μ
Negative binomial type I (`"NBI"`)	$\frac{\Gamma(y+1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$	μ	$\mu+\sigma\mu^2$
Negative binomial type II (`"NBII"`)	$\frac{\Gamma(y+\mu/\sigma)\sigma^y}{\Gamma(\mu/\sigma)\Gamma(y+1)(1+\sigma)^{y+\mu/\sigma}}$	μ	$(1+\sigma)\mu$
Poisson–inverse Gaussian (`"PIG"`)	$\left(\frac{2\alpha}{\pi}\right)^{0.5}\frac{\mu^y\exp(1/\sigma)K_{y-0.5}(\alpha)}{(\alpha\sigma)^y y!}$	μ	$\mu+\sigma\mu^2$

†These have been parameterized according to Rigby and Stasinopoulos (2005) and are defined in terms of μ and σ. In all cases, $y\in\mathbb{N}_0$ and $\mu,\sigma\in(0,\infty)$. Since the distributional parameters can take only positive values, the transformation function $\log(\cdot)$ is employed in estimation. $\alpha=\sqrt{(1/\sigma^2+2\mu/\sigma)}$ and $K_\lambda(t)=\frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\{-0.5t(x+x^{-1})\}\mathrm{d}x$ is the modified Bessel function of the third kind. Argument `margins` of `gjrm()` in `GJRM` enables the user to employ the desired binary and discrete marginals. For the discrete margin, the possible values are indicated within parentheses next to the names of the distributions.

our case-study—which seeks to estimate the effect of insurance on healthcare use—we present evidence that the magnitude of endogeneity is larger among females than among males.

Predictor η_i (where the subscript denoting which parameter the predictor belongs to has been dropped for simplicity) takes the additive form

$$\eta_i=\beta_0+\sum_{k=1}^{K}s_k(\mathbf{z}_{ki}),\qquad i=1,\ldots,n, \tag{2}$$

where $\beta_0\in\mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the kth subvector of the complete covariate vector $\mathbf{z}_i$ (containing, for example binary, categorical, continuous and spatial variables) and the K functions $s_k(\mathbf{z}_{ki})$ represent generic effects which are chosen according to the type of covariate(s) that is considered. Each $s_k(\mathbf{z}_{ki})$ can be approximated as a linear combination of J_k basis functions $b_{kj_k}(\mathbf{z}_{ki})$ and regression coefficients $\beta_{kj_k}\in\mathbb{R}$, i.e. (e.g. Wood (2017))

$$\sum_{j_k=1}^{J_k}\beta_{kj_k}b_{kj_k}(\mathbf{z}_{ki}). \tag{3}$$

This formulation implies that the vector of evaluations $\{s_k(\mathbf{z}_{k1}),\ldots,s_k(\mathbf{z}_{kn})\}'$ can be written as $\mathbf{Z}_k\boldsymbol{\beta}_k$ with $\boldsymbol{\beta}_k=(\beta_{k1},\ldots,\beta_{kJ_k})'$ and design matrix $Z_k(i,j_k)=b_{kj_k}(\mathbf{z}_{ki})$. This enables the predictor in equation (2) to be written as

$$\boldsymbol{\eta}=\beta_0\mathbf{1}_n+\mathbf{Z}_1\boldsymbol{\beta}_1+\ldots+\mathbf{Z}_K\boldsymbol{\beta}_K, \tag{4}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of 1s. Equation (4) can also be written in a more compact way as $\boldsymbol{\eta}=\mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z}=(\mathbf{1}_n,\mathbf{Z}_1,\ldots,\mathbf{Z}_K)$ and $\boldsymbol{\beta}=(\beta_0,\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_K')'$.

Each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k\boldsymbol{\beta}_k'\mathbf{D}_k\boldsymbol{\beta}_k$ whose role is to enforce specific properties on the kth function, such as smoothness. Note that $\mathbf{D}_k$ depends only on the choice of basis functions, but not on $\boldsymbol{\beta}_k$. Smoothing parameter $\lambda_k\in[0,\infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{s}_k(\mathbf{z}_{ki})$. The overall penalty can be defined as $\boldsymbol{\beta}'\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$, where $\mathbf{D}_{\boldsymbol{\lambda}}=\mathrm{diag}(0,\lambda_1\mathbf{D}_1,\ldots,\lambda_K\mathbf{D}_K)$. Finally, the smooth functions are subject to centring (identifiability) constraints (Wood, 2017). The above smooth

function representation enables us to specify a rich variety of covariate effects (such as linear, non-linear and geographic effects) and we refer the reader to (Wood, 2017) for full details.

For the previous example (of a Gumbel copula model with Bernoulli and Poisson margins), the predictors for π_i, μ_i and θ_i can be written in a compact way as

$$\begin{aligned}\boldsymbol{\eta}_\pi &= \beta_{10}\mathbf{1}_n + \mathbf{Z}_{11}\boldsymbol{\beta}_{11} + \ldots + \mathbf{Z}_{1K}\boldsymbol{\beta}_{1K},\\ \boldsymbol{\eta}_\mu &= \beta_{20}\mathbf{1}_n + \beta_{\text{end}}\mathbf{y}_1 + \mathbf{Z}_{21}\boldsymbol{\beta}_{21} + \ldots + \mathbf{Z}_{2K}\boldsymbol{\beta}_{2K},\\ \boldsymbol{\eta}_\theta &= \beta_{30}\mathbf{1}_n + \mathbf{Z}_{31}\boldsymbol{\beta}_{31} + \ldots + \mathbf{Z}_{3K}\boldsymbol{\beta}_{3K},\end{aligned}$$

where $\mathbf{y}_1$ appears as an (endogenous) predictor in $\boldsymbol{\eta}_\mu$, thus giving the set-up a recursive structure. Our main interest is the parameter β_{end}, which represents the effect of the binary endogenous regressor on the predictor of the outcome of interest. In the particular context of a recursive model with an endogenous regressor the set of covariates might be common across the predictors, except for the endogenous equation which must include at least one instrument that is not included in the other predictors (e.g. Han and Vytlacil (2017) and Meango and Mourifie (2014)).

2.2. Identification

Han and Vytlacil (2017) considered a version of this model, but where the outcome variable is binary instead of a non-negative integer count. Borrowing from the familiar binary probit set-up, they provided proofs that establish situations where recursive copula constructions have a full rank Jacobian. Their result requires two conditions. The first is that the copula must show first-order stochastic dominance with respect to θ, which says that, as F_1 increases, so also does F_2, and that such a relationship becomes stronger as θ increases. The second condition is the presence of an instrument that affects the endogenous regressor but not the outcome variable. Without such an instrument, it remains possible to write down copula expressions with recursive structures, but there is no guarantee that specific parameter values yield a unique maximum to the likelihood function that is formed from the copula expression.

Zimmer (2018) extended Han and Vytlacil's argument to settings in which the outcome variable follows a Poisson distribution. The extension hinges on the famous 'law of rare events', which states that a Poisson outcome may be viewed as the sum of 'successes' from many independent Bernoulli trials, so long as the number of trials is large, and the probability of a success in any individual trial is small (Cameron and Trivedi (2013), page 5). And, because Han and Vytlacil's result holds for each individual 'trial', their result should also hold for the sum of many trials, so long as those trials remain independent.

More generally, copula specifications should support recursive structures beyond just binary and Poisson settings. The reason is that parametric copula functions can be generated by 'mixing' marginals distributions that share a common random effect (Trivedi and Zimmer (2007), pages 36–38), and such a mixing method has long been recognized as a way to combine marginals of disparate forms, beyond just binary and Poisson (Fridman and Harris, 1998). For example, in trying to form the (unknown) joint distribution for the pair (Y_1, Y_2), we can start by decomposing the joint distribution into a product of marginals:

$$f_2(y_2|y_1, u)\, f_1(y_1|u),$$

where y_1 appears as a conditioning variable in the marginal for y_2. Then, by assuming the presence of a shared random effect u, with an assumed probability density function $f_u(u)$, and then numerically integrating out the random effect,

$$\int f_2(y_2|y_1, u)\, f_1(y_1|u)\, f_u(u)\mathrm{d}u,$$

we arrive at a valid joint distribution, albeit one without an analytical expression.

However, as argued in Section 1.1, the numerical integration step can be very computationally taxing, especially for applications with large estimation samples and many explanatory variables. To side-step such a computational headache, the recursive copula approach replaces the assumption of a particular distribution for the random effect, $f_u(u)$, with an assumption about the final form of the joint distribution itself, which, of course, implies some (unknown) distribution for the random effect.

It is not obvious, neither *a priori* nor *ex post*, which assumption is stronger: assuming a distribution for the random effect, or assuming an analytical form for the final joint distribution. But the latter offers two advantages. First, it makes estimation more manageable. Second, the availability of many off-the-shelf copula functions enables a researcher to test the robustness of the distributional assumption by easily changing the form of the copula. Such checks of robustness are not as straightforward in shared random-effects settings, because the distribution of the random effect, $f_u(u)$, must be symmetric about zero for the signs of the coefficient estimates to have natural interpretations, which strongly restricts available choices for the form of $f_u(u)$.

2.3. Estimation

For notational convenience, let us suppress for the moment the conditioning on parameters and observation index i and recall that F_1 and F_2 indicate the marginal CDFs and f_2 the PMF of y_2. The joint distribution in equation (1) is a CDF, whereas for maximum likelihood estimation we need the joint probability mass function f_{12} of a binary random variable Y_1 and a discrete random variable Y_2. Using the fact that $f_2(y_2) = F_2(y_2) - F_2(y_2 - 1)$, this can be expressed as

$$\begin{aligned} f_{12}(y_1, y_2) = &[C\{F_1(0), F_2(y_2)\} - C\{F_1(0), F_2(y_2) - f_2(y_2)\}]^{1-y_1}[f_2(y_2) - C\{F_1(0), F_2(y_2)\} \\ &+ C\{F_1(0), F_2(y_2) - f_2(y_2)\}]^{y_1}, \end{aligned} \tag{5}$$

where $F_1(0)$ represents $\Pr(Y_1 = 0)$. The following shows how equation (5) is derived. If $Y_1 = 0$ then $\Pr(Y_1 = 0, Y_2 = y_2) = \Pr(Y_1 = 0, Y_2 \leqslant y_2) - \Pr(Y_1 = 0, Y_2 \leqslant y_2 - 1)$, which can be calculated by using $C\{F_1(0), F_2(y_2)\} - C\{F_1(0), F_2(y_2 - 1)\}$. If $Y_1 = 1$ then $\Pr(Y_1 = 1, Y_2 = y_2) = \Pr(Y_1 = 1, Y_2 \leqslant y_2) - \Pr(Y_1 = 1, Y_2 \leqslant y_2 - 1)$, which, exploiting the result for the $Y_1 = 0$ case, can be obtained as $F_2(y_2) - C\{F_1(0), F_2(y_2)\} - [F_2(y_2 - 1) - C\{F_1(0), F_2(y_2 - 1)\}]$. Expression (5) is convenient computationally, first because it involves only the evaluation of two joint CDFs (instead of four; the case in which $\Pr(Y_1 = 1, Y_2 = y_2)$ would be calculated naively) and second because it avoids the evaluation of F_2 for negative arguments. Note that the presence of a binary endogenous variable in η_μ (see, for example, the last paragraph of Section 2.1) does not alter the construction of the joint PMF; $f_{12}(y_1, y_2)$ can be written as $f_{2|1}(y_2|y_1) f(y_1)$; hence its form does not change if η_μ includes y_1.

Assuming that a random sample $\{(y_{1i}, y_{2i}, \mathbf{z}_i)\}_{i=1}^n$ is available, the log-likelihood function can be written as

$$\begin{aligned} l(\boldsymbol{\delta}) = \sum_{i=1}^{n} ((1 - y_{1i}) \log[&C\{F_1(0|\pi_i), F_2(y_{2i}|\mu_i, \sigma_i); \theta_i\} - C\{F_1(0|\pi_i), F_2(y_{2i}|\mu_i, \sigma_i) \\ &- f_2(y_{2i}|\mu_i, \sigma_i); \theta_i\}] y_{1i} \log[f_2(y_{2i}|\mu_i, \sigma_i) - C\{F_1(0|\pi_i), F_2(y_{2i}|\mu_i, \sigma_i); \theta_i\} \\ &+ C\{F_1(0|\pi_i), F_2(y_{2i}|\mu_i, \sigma_i) - f_2(y_{2i}|\mu_i, \sigma_i); \theta_i\}]), \end{aligned}$$

where $\pi_i = g_\pi^{-1}(\eta_{\pi_i})$, $\mu_i = g_\mu^{-1}(\eta_{\mu_i})$, $\sigma_i = g_\sigma^{-1}(\eta_{\sigma_i})$ and $\theta_i = g_\theta^{-1}(\eta_{\theta_i})$. Parameter vector $\boldsymbol{\delta}$ is given as $(\boldsymbol{\beta}_\pi^{\mathrm{T}}, \boldsymbol{\beta}_\mu', \boldsymbol{\beta}_\sigma', \boldsymbol{\beta}_\theta')'$ which contains the coefficient vectors of additive predictors $\eta_{\pi_i}, \eta_{\mu_i}, \eta_{\sigma_i}$ and η_{θ_i}.

Because of the flexible predictors' structures that are employed here, the use of a classic

(unpenalized) optimization algorithm is likely to result in smooth function estimates which may not reflect the true underlying trends in the data (e.g. Wood (2017)). Therefore, we maximize

$$l_p(\boldsymbol{\delta}) = l(\boldsymbol{\delta}) - \tfrac{1}{2}\boldsymbol{\delta}'\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\delta}, \tag{6}$$

where $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(\boldsymbol{\lambda}_\pi \mathbf{D}_\pi, \boldsymbol{\lambda}_\mu \mathbf{D}_\mu, \boldsymbol{\lambda}_\sigma \mathbf{D}_\sigma, \boldsymbol{\lambda}_\theta \mathbf{D}_\theta)$, with each smoothing parameter vector containing all the smoothing parameters that are related to the corresponding $\mathbf{D}$-component, and the overall $\boldsymbol{\lambda}$ is defined as $(\boldsymbol{\lambda}_\pi', \boldsymbol{\lambda}_\mu', \boldsymbol{\lambda}_\sigma', \boldsymbol{\lambda}_\theta')'$. To estimate $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$, we have extended the efficient and stable trust region algorithm with integrated automatic multiple-smoothing-parameter selection that was proposed by Marra *et al.* (2017) to the context of copula models with binary and discrete margins. For all the copulas and margins in Tables 1 and 2, the analytical score and Hessian of $l_p(\boldsymbol{\delta})$ that is required for estimation have been tediously derived and verified by using numerical derivatives. It is worth noting that these quantities have been implemented in a modular fashion; hence it will be in principle easy to extend our algorithm to other parametric copulas and marginal distributions that are not considered in this work. As stressed in Section 1, although the implementation of the models proposed exploited many of the functions and routines that are already available in `GJRM`, extending the software to accommodate the new developments required a large amount of programming work, which resulted in a set of newly introduced functions.

At convergence, confidence intervals for any linear and non-linear function of $\boldsymbol{\delta}$ can be reliably obtained by using the Bayesian large sample approximation $\boldsymbol{\delta} \sim^{\text{a}} N\{\hat{\boldsymbol{\delta}}, -\mathbf{H}_p(\hat{\boldsymbol{\delta}})^{-1}\}$, where $\mathbf{H}_p(\hat{\boldsymbol{\delta}})^{-1}$ is the model's penalized Hessian and superscript 'a' denotes 'asymptotically distributed as'. Furthermore, it can be proved that $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0 = O_P(n^{-1/2})$ as $n \to \infty$, where $\boldsymbol{\delta}^0$ is the 'true' parameter vector. Appendices A and B in the on-line supplementary material provide details on the estimation algorithm and confidence intervals, as well as some asymptotic results.

2.4. The R GJRM package

The models can be employed via the `gjrm()` function in the R package `GJRM`. An example of the syntax is

```
fl <- list(y1 ~ z1 + s(z2) + s(z3),
           y2 ~ y1 + z1 + s(z2),
              ~ z1 + s(z3),
              ~ z1 + s(z2))
md <- gjrm(fl, margins = c("probit", "NBI"), BivD = "PL", Model = "B")
```

where `fl` is a list containing four equations. The first equation is for parameter π of the Bernoulli distribution of the binary endogenous regressor y1 with probit link function (logit and cloglog are also allowed for). The second and third equations are for parameters μ and σ of the discrete distribution of the outcome of interest y2, which in this example is negative binomial (NB) type I (`"NBI"`). Finally, the fourth equation is for the copula dependence parameter θ. Argument `BivD` specifies the copula function and `Model="B"` means that a bivariate model is employed. Symbol `s()` refers to the smooth function that was mentioned in Section 2.1. The default is `bs="tp"` (penalized low rank thin plate spline) with `k=10` (the number of basis functions) and `m=2` (the order of derivatives). However, argument `bs` can also be set to, for example, `cr` (penalized cubic regression spline), `ps` (*P*-spline) and `mrf` (Markov random field), to name but a few. Functions such as `AIC()`, `summary()` and `predict()` can be employed in the usual manner. Function `post.check()` will produce, for the discrete margin, a histogram and normal Q–Q-plot of randomized normalized quantile residuals (Dunn and Smyth, 1996). Building on the results of Kalliovirta (2008), we have been looking into implementing diagnostics based on bivariate randomized normalized quantile residuals; more theoretical work is required here and these may be available in future releases of `GJRM`. Appendix C in the on-line

supplementary material presents the results of a simulation study and supports the empirical effectiveness of the approach proposed.

3. Case-study

The Affordable Care Act, which aimed to extend health insurance coverage to the previously uninsured, might have led to increased usage. For this, assessing the effect of the law requires an accurate estimate of the effect of insurance on the usage of medical services.

This section applies the proposed method to estimate the effect of insurance status (a binary variable) on doctor visits (a count variable). The method finds statistically significant evidence that insurance is endogenous with respect to usage of doctor services. When endogeneity is taken into account, the effect of insurance is larger than when endogeneity is ignored.

3.1. Data

The estimation sample, drawn from the 2010 wave of the Medical Expenditure Panel Survey, was originally used by Zimmer (2018). The sample considers all respondents in the 2–64 years age range who are not covered by any form of federal or state public health insurance programme. The outcome variable records a person's number of visits in January 2010 to a family or general practice physician. (Most private insurance plans in the USA reset deductibles at the start of the calendar year, so focusing on January usage should capture a stronger insurance effect.) Following standard practice in health economics research, we opt to study discrete count measures of usage, rather than total spending, for three reasons. First, count measures of usage link better to economic theory, where 'demand' reflects the number of units that are consumed, not the total spending on those units. Second, discrete measures of usage avoid confusion about 'charges' *versus* 'spending', which usually differ. Third, with nearly 90% of medical spending in the USA channelled through third-party payers, discrete count measures are likely to suffer from less recollection error. The final estimation sample contains $n = 13137$ unique individuals.

Sample means appear in Table 3. Insurance appears to correlate with a person's number of visits to a doctor, but that relationship cannot be interpreted as causal, because insured and uninsured people appear to differ along several other dimensions. Specifically, insured subjects are older, on average, than their uninsured counterparts and are more likely to be female, employed and healthy.

As discussed above, the model requires that at least one variable appears in the insurance

Table 3. Sample means for the 2010 Medical Expenditure Panel Survey data†

	Results for insured, $n = 9302$	*Results for uninsured, $n = 3835$*
Doctor visits in January	0.08	0.05
Age	44.1	41.4
Female	0.53	0.48
Fair or poor health	0.09	0.18
Employed	0.85	0.61
Firm size > 50	0.43	0.11
Firm has multiple locations	0.56	0.20

†All respondents are in the age range 25–64 years, and are not covered by public insurance.

marginal, but not in the utilization equation. The bottom two rows of Table 3 show two candidates:

(a) firm size and
(b) an indicator of whether the firm has multiple locations.

(These variables equal 0 for non-employed subjects.) Firm size should influence insurance status because of economies of scale that make it cheaper for larger firms to offer health coverage. Literature on human resource information systems (Chae *et al.*, 2011) argued that larger firms with dedicated human resource staffs should also have more developed infrastructures for enrolling employees into insurance plans (Hendrickson, 2003). Furthermore, because of differences in state level mandates, firms with multiple locations are likely to harmonize their insurance offerings in accordance with the strictest jurisdiction in which they operate. However, these two variables are unlikely to affect usage (directly), especially after controlling for employment status. It also is worth noting that health economics research offers many examples of employer-specific information serving as identifying instruments in contexts that are similar to the one that is explored in this case-study (Dowd *et al.*, 1991; Meer and Rosen, 2004; Deb and Trivedi, 2006; Selden and Hudson, 2006).

3.2. Model specification

Each marginal includes as control variables age, gender, health status (fair or poor health) and a proxy for income (employment status). The presence of these socio-economic measures—called 'demand shifters' in economics—in models of healthcare and insurance demand follows from economic theory (Cameron *et al.*, 1988), and, as a consequence, they have become standard in such settings. US law precludes insurance companies from using other potential demand shifters in setting premiums, implying that, economically, such information remains 'unobserved' when people purchase insurance. Consequently, we do not include other demand shifters in either marginal. Rather, the copula dependence parameter will pick up the influence of those factors.

Table 4. AIC-values for various copulas and discrete marginal distributions for doctor visits†

Copula	*AIC-values for the following distributions:*			
	Poisson	*Poisson–inverse Gaussian*	*NB I*	*NB II*
N	20018.42	19879.48	19890.38	19864.83
Clayton 90	19998.39	19881.22	19891.79	19862.93
Clayton 270	20023.20	19882.13	19892.63	19870.38
Joe 90	20023.62	19883.87	19894.03	19871.87
Joe 270	19998.09	19881.34	19891.91	19863.03
Gumbel 90	20023.47	19883.53	19893.76	19871.46
Gumbel 270	20008.84	19879.91	19890.69	19862.21
AMH	20004.93	19880.38	19890.94	19862.65
FGM	20001.53	19880.64	19891.21	19862.57
Student t	20056.04	19909.04	19920.38	19890.51
Plackett	19997.74	19880.64	19891.22	*19861.99*
Frank	19998.24	19880.62	19891.20	19862.04

†For the insurance status variable, the conventional probit link function is employed. (Using the Bayesian information criterion led to the same choice of marginal distribution and copula.) The lowest AIC-value is in italics.

The marginal distribution for visits to a doctor and the copula are chosen by using the Akaike information criterion AIC evaluated for all possible combinations of discrete distributions and copulas that are available in `GJRM`. For the insurance status variable, all the three link functions (probit, logit and cloglog) led to the same conclusions in our case-study; hence we present only the results that were obtained by using the classical probit link. Note that we used only the rotated 90° and 270° versions of the Clayton, Gumbel and Joe copulas because the data support the presence of negative dependence between the responses, and therefore it would not make sense to consider rotations allowing for positive dependence. From Table 4, we can see that the Plackett copula with NB II margin appears to offer the best fit. This corroborates other evidence that this distribution often outperforms alternative count data distributions, especially in settings, such as that considered here, where the count variable shows large probability mass at zero (Cameron and Trivedi (2013), pages 84–85). The fact that the Plackett, Gaussian, Frank, Farlie–Gumbel–Morgenstern and Ali–Mikhail–Haq copulas produce similar AIC-values suggests a lack of asymmetric dependence since these copulas' shapes show relatively symmetric dependence patterns in each tail (as opposed to the Clayton, Gumbel and Joe copulas, which exhibit asymmetric dependence).

3.3. Results

The main results appear in Table 5. The left-hand panel shows estimates from a simple NB II

Table 5. Estimates (with standard errors in parentheses) obtained from the univariate NB II model, the recursive Plackett copula model with Bernoulli distribution (with probit link) for insurance and NB II distribution for doctor visits, and the control function approach

	Results for simple NB II, doctor visits	*Results for Plackett copula*		*Results for control function*	
		Doctor visits	*Insured*	*Doctor visits*	*Insured*†
Insured	0.62	1.34	—	1.56	—
	(0.09)	(0.22)	—	(0.29)	—
Age	0.02	0.02	Spline	0.02	Spline
	(0.003)	(0.003)		(0.003)	
Female	0.27	0.21	0.18	0.20	0.05
	(0.07)	(0.07)	(0.03)	(0.07)	(0.01)
Fair or poor health	0.94	1.02	−0.45	1.08	−0.14
	(0.08)	(0.08)	(0.04)	(0.09)	(0.01)
Employed	−0.16	−0.40	0.31	−0.46	0.13
	(0.08)	(0.11)	(0.03)	(0.13)	(0.01)
Firm size > 50	—	—	0.68	—	0.16
	—	—	(0.03)	—	(0.01)
Firm has multiple locations	—	—	0.62	—	0.18
	—	—	(0.03)	—	(0.01)
Constant	−4.31	−4.35	−0.12	−4.49	0.46
	(0.19)	(0.18)	(0.03)	(0.19)	(0.01)
Overdispersion σ (95% confidence interval)	0.22 (0.18, 0.29)	0.21 (0.16, 0.28)		0.22 (0.16, 0.28)	
First-stage residuals	—	—	—	−1.01	—
	—	—	—	(0.30)	—
Dependence τ (95% confidence interval)	—	−0.19 (−0.28, −0.09)			

†Ordinary least squares.

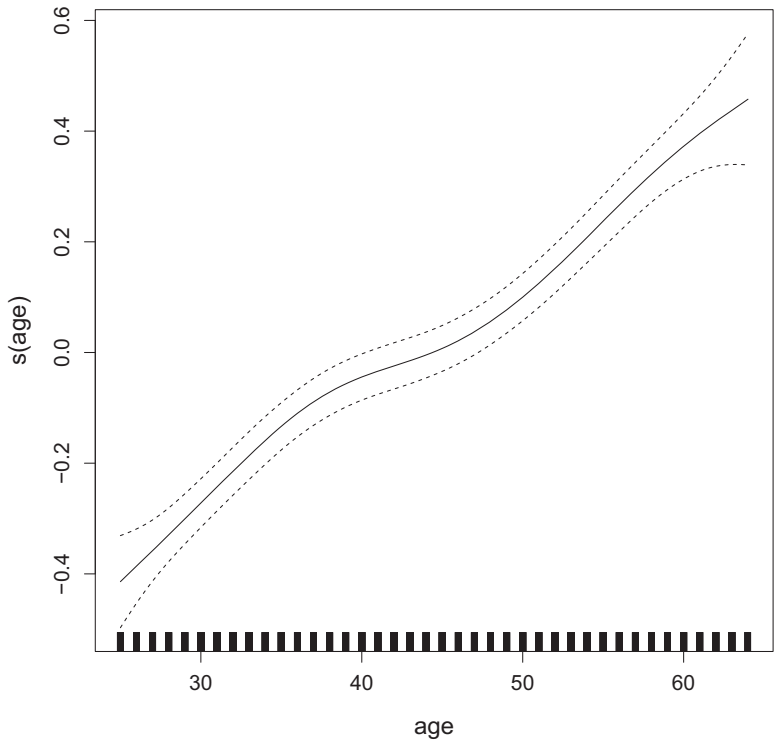

Fig. 1. Estimated smooth effect of age on insurance status on the scale of the predictor, and associated 95% pointwise intervals: the jittered rug plot, at the bottom of the graph, shows the covariate values

specification, with no attempt to address the endogeneity of insurance. Focusing first on the control variables, usage appears to increase with age and health problems. Females have more visits to a doctor than their male counterparts, and employed subjects report fewer visits to a doctor than those not working. Finally, turning to the main result of interest, the simple NB II specification shows that insurance correlates with an approximate 62% increase in visits to a doctor.

Attempting to correct for possible endogeneity bias, the second panel shows the results for the preferred Plackett copula specification. Most of the control variables exhibit similar estimates to those of the simple NB II set-up, although the effect of being employed becomes larger (more negative). As for insurance, the coefficient estimate suggests that, after correcting for endogeneity, insurance leads to an approximate 134% increase in visits to a doctor, which is substantially larger than the effect that was reported by the simple NB II model. Results for the dependence parameter (converted to Kendall's τ) appear at the bottom of Table 5. The finding of negative dependence suggests that unobserved traits that induce certain people to enrol in insurance also tend to reduce the usage of doctor services. Such a pattern, which is often called 'favourable selection' by health economists, is similar to what appears in other studies of insurance and healthcare usage (e.g. Finkelstein and McGarry (2005), Pauly (2005) and Cameron and Trivedi (2013)).

3.4. Spline estimate

The model proposed uses a smooth function for the age variable in the insurance equation.

Table 6. Estimates (with standard errors in parentheses) obtained from the recursive Plackett copula model with NB II margin and dependence parameter allowed to vary by gender

	Results for doctor visits	*Results for insured*	*Results for dependence*
Insured	1.38	—	—
	(0.22)	—	—
Age	0.02	Spline	—
	(0.004)		—
Female	0.29	0.18	−0.45
	(0.08)	(0.03)	(0.21)
Fair or poor health	1.02	−0.45	—
	(0.08)	(0.04)	—
Employed	−0.42	0.31	—
	(0.11)	(0.03)	—
Firm size > 50	—	0.68	—
	—	(0.03)	—
Firm has multiple locations	—	0.62	—
	—	(0.03)	—
Constant	−4.38	−0.12	−0.62
	(0.18)	(0.03)	(0.26)
Overdispersion σ (95% confidence interval)	0.21	—	—
	(0.16, 0.27)	—	—

The graph of the estimated effect is shown in Fig. 1. There appears to be some non-linearity in the estimated shape, although it could be reasonably argued that the functional form looks roughly linear and hence that a parametric effect would adequately describe the effect of age. We would indeed agree with this course of action. Generally, the main point about using flexible specifications is to avoid making *a priori*, potentially questionable, assumptions. In this case, the relationship turned out to be roughly linear, suggesting that a simpler model specification is acceptable. The plot suggests that, as subjects age, their probabilities of having insurance also increase, and that such an increase seems less marked for individuals in the 39–48 years age range. This finding, which has been long recognized in health economics, is likely to reflect that medical risks increase with age, leading to increases in demand for insurance to mitigate financial uncertainty that is associated with those risks (Arrow, 1963).

3.5. Alternative specifications

For comparison, Table 5 reports estimates from a simple control function approach; the first stage uses a linear probability model for insurance; then residuals from the first stage are included in the second stage NB II specification for visits to a doctor. Following other applications of the control function method for binary endogenous variables (Terza *et al.*, 2008), we use the simplest definition of a residual (i.e. the difference between the observed binary outcome and the predicted probability of the outcome) despite aforementioned potential problems with such practice (Wooldridge (2010), page 746). The coefficient of insurance is 1.56, which is slightly larger than, but nonetheless comparable with, the estimate from the copula approach. Further, the coefficient of the first-stage residuals is negative and statistically significant, which confirms the finding of favourable selection.

Table 6 also shows estimates from a specification in which the copula dependence parameter is given a regression structure. Such a set-up makes sense if endogeneity, as captured by the

Table 7. Estimates (with standard errors in parentheses) obtained from the recursive Plackett copula model with marginal for doctor visits specified as Poisson

	Results for doctor visits	*Results for insured*
Insured	1.63	—
	(0.17)	—
Age	0.01	Spline
	(0.003)	
Female	0.16	0.18
	(0.07)	(0.03)
Fair or poor health	1.04	−0.45
	(0.07)	(0.04)
Employed	−0.55	0.32
	(0.09)	(0.03)
Firm size > 50	—	0.67
	—	(0.03)
Firm has multiple locations	—	0.62
	—	(0.03)
Constant	−4.20	−0.13
	(0.17)	(0.03)
Overdispersion σ (95% confidence interval)	—	—
Dependence τ (95% confidence interval)	−0.27 (−0.36, −0.19)	

dependence term, differs with respect to observable characteristics. This specification does not alter the main substantive conclusions about the link between insurance and physician usage, but, as shown in the rightmost column of Table 6, gender appears to correlate significantly with the magnitude of dependence. (Other explanatory variables did not appear to alter dependence significantly.) Recalling that *overall* dependence, as reported at the bottom of Table 5, is −0.19, the numbers in Table 6 imply that, once converted to Kendall's τ, dependence among females is −0.24, with 95% interval (−0.34, −0.14), whereas dependence among males is −0.14, with 95% interval (−0.25, −0.02). Thus, despite the slight overlap in confidence intervals, the results suggest that females exhibit stronger favourable selection than do their male counterparts. This is likely to reflect higher levels of risk aversion among females: a finding that is widely reported in both economics and psychology research (Hartog *et al.*, 2002; Agnew *et al.*, 2008; Eckel and Grossman, 2008; Borghans *et al.*, 2009). Medical research has attempted to link such gender disparities in risk aversion to differences in levels of testosterone (Sapienza *et al.*, 2009).

Finally, along the lines of Zimmer (2018), Table 7 shows results from a model in which the marginal for visits to a doctor is specified as Poisson, which Table 4 suggests offers the worst fit of the count distributions under consideration. Estimates from this set-up find a larger insurance effect and a larger (negative) dependence estimate, suggesting that some of the unaccounted-for overdispersion in visits to a doctor is infecting those parameter estimates.

4. Conclusions

In trying to estimate the causal effect of insurance on healthcare utilization, researchers must confront the likelihood that people seek insurance in anticipation of future healthcare needs.

Ignoring such endogeneity bias might produce misleading estimates. The topic has important policy implications, because health insurance reforms that have been implemented in the USA in recent decades have focused on extending coverage to the previously uninsured. Consequently, understanding the effects of those reforms requires an accurate estimate of the causal effect of insurance on utilization. However, obtaining such an estimate is complicated by the discreteness that is inherent in common measures of healthcare usage.

This paper presents an estimation approach, based on copula functions, that consistently estimates the coefficient of an endogenous regressor in count data settings. The method is general in the types of non-linear data patterns that it can accommodate. Moreover, the statistical significance (or lack thereof) of the copula dependence parameter provides a convenient means to assess the presence of endogeneity. The results of our case-study point to evidence of favourable selection into insurance and, once favourable selection has been taken into account, the effect of insurance is larger than when it is ignored. Specifically, ignoring favourable selection, insurance appears to increase doctor visit usage by 62% but, adjusting for favourable selection, the effect increases to 134%. The results also suggest that females exhibit larger favourable selection than do males.

When health insurance is not required, health economists often worry about *adverse* selection, which is characterized by relatively unhealthy people being disproportionately drawn towards insurance coverage. The sicker risk pool might require insurance companies to increase premiums, potentially driving out the least sick members of the risk pool, resulting in an even sicker pool. In the extreme, this cycle might keep repeating, resulting in a 'death spiral' in which the insurance market ceases to exist. However, the finding of *favourable* selection suggests that concerns about death spirals might be misplaced. Many commentators warned that the Affordable Care Act, with its relatively weak mandate that everyone obtain coverage, might encourage death spirals in private insurance markets. The finding of favourable selection into such markets might partially explain the evident lack of such death spirals to date.

As presented in this paper, the marginals and copula are parametrically specified with several choices possible. Furthermore, all the parameters of the joint distribution assumed can be specified as flexible functions of covariates which can help to uncover interesting patterns in the data as highlighted in our empirical application. The numerical computations can be easily and efficiently carried out by using the newly revised R package `GJRM`. The estimation framework proposed does not require computationally taxing simulation-based estimators as is the case with other simultaneous estimation approaches. Specifically, as elaborated in Section 1.1, the copula approach is computationally more efficient than an approach that is based on shared unobserved random effects since it side-steps the problem of integrating out such effects. However, it could be argued that, because equation (5) requires the evaluation of CDFs, when using elliptical copulas (i.e. the Gaussian and Student t-distributions) integration is still required. We found that this was not problematic for this paper since we deal with the bivariate case, and because there are not redundant calculations. Increasing the number of margins will surely increase the computational cost of the approach based on elliptical copulas. In such a case, a proper exploration of the practical advantages and disadvantages of various alternatives is warranted.

It is worth noting that the methodology that was developed in this paper, although flexible, is fundamentally parametric, and as such it may suffer from the usual potential drawbacks resulting from model misspecifications. Developments where the margins and/or copula function are estimated by using non-parametric techniques (e.g. Kauermann *et al.* (2013)) were explored. However, we found that these were limited with respect to the inclusion of covariate effects and required large sample sizes to produce reliable results in a regression context.

Despite its parametric flavour, the approach proposed enables a large amount of model exploration via the many functional forms that have been included in the newly revised R package `GJRM`. Specifically, the `GJRM` package offers a wide menu of marginal distributions and copula functions—far more than are emphasized in this paper. Moreover, the spline capabilities permit a large degree of flexibility in how regressors relate to outcome variables and model parameters. Thus, a researcher can explore a vast number of permutations of functional forms and regressor effects using nothing more than simple changes in computer syntax. Such exploration, although certainly not non-parametric, captures the spirit of non-parametric approaches in that it enables the data to point to meaningful model structures. Moreover, such ease of exploration has the potential to reveal interesting economic patterns that might otherwise remain hidden. As a key example presented here, the apparent difference in selection effects between males and females would have remained undetected if we had not had access to such an easy-to-employ framework for model assembly.

References

Agnew, J., Anderson, L., Gerlach, J. and Szykman, L. (2008) Who chooses annuities?: an experimental investigation of the role of gender, framing, and defaults. *Am. Econ. Rev.*, **98**, 418–442.

Arrow, K. J. (1963) Uncertainty and the welfare economics of medical care. *Am. Econ. Rev.*, **53**, 941–973.

Borghans, L., Heckman, J., Golsteyn, B. and Meijers, H. (2009) Gender differences in risk aversion and ambiguity aversion. *J. Eur. Econ. Ass.*, **7**, 649–658.

Brechmann, E. C. and Schepsmeier, U. (2013) Modeling dependence with C- and D-vine copulas: the R package CDVine. *J. Statist. Softwr.*, **52**, no. 3, 1–27.

Cameron, A. C. and Trivedi, P. K. (2013) *Regression Analysis of Count Data*, 2nd edn. New York: Cambridge University Press.

Cameron, A. C., Trivedi, P. K., Milne, F. and Piggott, J. (1988) A microeconometric model of the demand for health care and health insurance in Australia. *Rev. Econ. Stud.*, **55**, 85–106.

Cardon, J. H. and Hendel, I. (2001) Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey. *RAND J. Econ.*, **32**, 408–427.

Chae, B., Prince, J. B., Katz, J. and Kabst, R. (2011) An exploratory cross-national study of information sharing and human resource information systems. *J. Globl Inform. Mangmnt*, **19**, no. 4, 18–44.

Deb, P. and Trivedi, P. K. (2006) Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization. *Econmetr. J.*, **9**, 307–331.

Dowd, B., Feldman, R., Cassou, S. and Finch, M. (1991) Health plan choice and the utilization of health care services. *Rev. Econ. Statist.*, **73**, 85–93.

Dunn, P. K. and Smyth, G. K. (1996) Randomized quantile residuals. *J. Computnl Graph. Statist.*, **5**, 236–245.

Eckel, C. C. and Grossman, P. J. (2008) Men, women and risk aversion: experimental evidence. In *Handbook of Experimental Economics Results*, vol. 1 (eds C. Plott and V. Smith), pp. 1061–1073. Amsterdam: North-Holland.

Finkelstein, A. and McGarry, K. (2005) Private information and its effect on market equilibrium: evidence from long-term care insurance market. *Am. Econ. Rev.*, **96**, 938–958.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. and the Oregon Health Study Group (2012) The Oregon health insurance experiment: evidence from the first year. *Q. J. Econ.*, **127**, 1057–1106.

Fridman, M. and Harris, L. (1998) A maximum likelihood approach for non-Gaussian stochastic volatility models. *J. Bus. Econ. Statist.*, **16**, 284–291.

Goldman, D. (1995) Managed care as a public cost-containment mechanism. *RAND J. Econ.*, **26**, 277–295.

Greene, W. (2008) *Econometric Analysis*, 6th edn. Upper Saddle River: Prentice Hall.

Han, S. and Vytlacil, E. J. (2017) Identification in a generalization of bivariate probit models with dummy endogenous regressors. *J. Econmetr.*, **199**, 63–73.

Hartog, J., Ferrer-i-Carbonell, A. and Jonker, N. (2002) Linking measured risk aversion to individual characteristics. *Kyklos*, **55**, 3–26.

Heckman, J. and Robb, R. (1985) Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data* (eds J. Heckman and B. Singer), pp. 156–245. New York: Cambridge University Press.

Hendrickson, A. R. (2003) Human resource information systems: backbone technology of contemporary human resources. *J. Lab. Res.*, **24**, 381–394.

Joe, H. (2014) *Dependence Modeling with Copulas*. Boca Raton: CRC Press.

Kalliovirta, L. (2008) Quantile residuals for multivariate models. *Discussion Paper 247*. University of Helsinki, Helsinki.

Kauermann, G., Schellhase, C. and Ruppert, D. (2013) Flexible copula density estimation with penalized hierarchical B-splines. *Scand. J. Statist.*, **40**, 685–705.
Kreif, N., Grieve, R., Hangartner, D., Turner, A., Nikolova, S. and Sutton, M. (2016) Examination of the synthetic control method for evaluating health policies with multiple treated units. *Hlth Econ.*, **25**, 1514–1528.
Lewbel, A. (2012) Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Rev. Econ. Statist.*, **30**, 67–80.
Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B. and Leibowitz, A. (1987) Health insurance and the demand for medical care: evidence from a randomized experiment. *Am. Econ. Rev.*, **77**, 251–277.
Marra, G. and Radice, R. (2017) Bivariate copula additive models for location, scale and shape. *Computnl Statist. Data Anal.*, **112**, 99–113.
Marra, G. and Radice, R. (2020) GJRM: generalized joint regression modeling. *R Package Version 0.2-2*. (Available from `https://cran.r-project.org/package=GJRM`.)
Marra, G., Radice, R., Bärnighausen, T., Wood, S. N. and McGovern, M. E. (2017) A simultaneous equation approach to estimating HIV prevalence with non-ignorable missing responses. *J. Am. Statist. Ass.*, **112**, 484–496.
Meango, R. and Mourifie, I., (2014) A note on the identification in two equations probit model with dummy endogenous regressor. *Econ. Lett.*, **125**, 360–363.
Meer, J. and Rosen, H. S. (2004) Insurance and the utilization of medical services. *Socl Sci. Med.*, **58**, 1623–1632.
Mello, M., Stearns, S. and Norton, E. (2002) Do Medicare HMOs still reduce health services use after controlling for selection bias? *Hlth Econ.*, **11**, 323–340.
Nelsen, R. (2006) *An Introduction to Copulas*, 2nd edn. New York: Springer.
Nikoloulopoulos, A. K. and Karlis, D. (2010) Regression in a copula model for bivariate count data. *J. Appl. Statist.*, **37**, 1555–1568.
Park, S. and Gupta, S. (2012) Handling endogenous regressors by joint estimation using copulas. *Marktng Sci.*, **31**, 567–586.
Pauly, M. V. (2005) Effects of insurance coverage on use of care and health outcomes for nonpoor young women. *Am. Econ. Rev.*, **95**, 219–233.
Radice, R., Marra, G. and Wojtys, M. (2016) Copula regression spline models for binary outcomes. *Statist. Comput.*, **26**, 981–995.
R Core Team (2020) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Rosenbaum, P. and Rubin, D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Sapienza, P., Zingales, L. and Maestripieri, D. (2009) Gender differences in financial risk aversion and career choices are affected by testosterone. *Proc. Natn. Acad. Sci. USA*, **106**, 15268–15273.
Selden, T. and Hudson, J. (2006) Access to care and utilization among children: estimating the effects of public and private coverage. *Med. Care*, **44**, I19–I26.
Sklar, A. (1973) Random variables, joint distributions, and copulas. *Kybernetica*, **9**, 449–460.
Terza, J. (1998) Estimating count data models with endogenous switching: sample selection and endogenous switching effects. *J. Econmetr.*, **84**, 129–139.
Terza, J., Basu, A. and Rathouz, P. (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Hlth Econ.*, **27**, 531–543.
Tran, K. and Tsionas, E. (2015) Endogeneity in stochastic frontier models: copula approach without external instruments. *Econ. Lett.*, **133**, 85–88.
Trivedi, P. and Zimmer, D. (2007) Copula modeling: an introduction for practitioners. *Found. Trends Econmetr.*, **1**, 1–111.
Trivedi, P. and Zimmer, D. (2017) A note on identification of bivariate copulas for discrete count data. *Econometrics*, **5**, no. 1, 1–11.
Winkelmann, R. (2012) Copula bivariate probit models: with an application to medical expenditures. *Hlth Econ.* **21**, 1444–1455.
Wood, S. N. (2017) *Generalized Additive Models: an Introduction with R*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
Wooldridge, J. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edn. Cambridge: MIT Press.
Zheng, X. and Zimmer, D. (2008) Farmers' health insurance and access to health care. *Am. J. Agric. Econ.*, **90**, 267–279.
Zimmer, D. (2018) Using copulas to estimate the coefficient of a binary endogenous regressor in a Poisson regression: application to the effect of insurance on doctor visits. *Hlth Econ.*, **27**, 545–556.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'On-line supplementary material: Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models'.