A Trivariate Model for Estimating the Effects of Two Binary Endogenous Explanatory Variables: Application to the Impact of Medical Care Usage on Work Absenteeism

Panagiota Filippou
R&D group
Capita UK

Giampiero Marra
Department of Statistical Science
University College London

Rosalba Radice
Cass Business School
City, University of London

David Zimmer
Department of Economics
Western Kentucky University

April 24, 2019

Abstract

This paper proposes a model that estimates the effects of two endogenous binary regressors. The model specifies three equations connected via a multivariate distribution, which makes it possible to model the correlations between the equations, hence accounting for unobserved heterogeneity. Parameter estimation is based on a trust region algorithm, which exploits first and second order analytical derivatives of trivariate integrals. An empirical application explores the question: Does visiting a medical provider cause an employee to miss work? We find that, observationally, having a curative visit associates with a nearly 80 percent increase in the probability of missing work, while having a preventive visit correlates with a smaller 14 percent increase in the likelihood of missing work. However, after addressing potential endogeneity, neither type of visit appears to significantly relate to missing work. Therefore, we conclude that the observed links between medical usage and absenteeism derive from unobserved heterogeneity, rather than direct causal channels.

1 Introduction

This paper aims to build a statistical model that accommodates two endogenous explanatory variables in an outcome equation. We accomplish that by using a recursive trivariate probit model which can be regarded as an extension of the recursive bivariate model introduced by Marra and Radice (2011) (see also references therein). The adopted modeling framework is based on Filippou, Marra, Radice (2017) and Filippou et al. (2019), which makes it possible to specify and estimate the trivariate model needed for this work.

As a motivating example, we consider the effects of seeking medical care on missing work. Labor economists have long recognized that injury or illness represents one of the most common reasons for worker absenteeism.[1] The U.S. Centers for Disease Control and Prevention (CDC) estimates that health-related worker absenteeism costs employers $225.8 billion annually, or about $1,685 per worker.[2] Those large dollar amounts call for rigorous studies exploring the specific channels through which medical events relate to missing work.

A voluminous body of research, scattered across a wide range of academic disciplines, explores the effects of health-related issues on productivity (Nicholson et al., 2006; Pauly, Nicholson, and Polsky, 2008; Shultz, Chen, and Edington, 2009; Zhang, Bansback, and Anis, 2011; Zhang et al., 2016; Stromberg et al., 2017). Some of the extant literature looks specifically at absenteeism, while other studies investigate the related issue of "presenteeism," defined as ill employees showing up at work. At the risk of over-generalizing such a large body of literature, the consensus appears to be that health-related problems, including absenteeism, hinder worker productivity.

[1] "The Causes And Costs Of Absenteeism In The Workplace," Forbes, 7/10/2013.

[2] "Worker Illness and Injury Costs U.S. Employers $225.8 Billion Annually," CDC Foundation Report, 1/28/2015.

This paper explores a more narrowly-targeted question: Does visiting a medical provider cause an employee to miss work? Providing an answer to that question must confront two complications. First, employees likely possess unobserved (to the researcher) traits that increase their likelihood of visiting a medical provider *while also* increasing their chances of missing work. The most obvious such traits involve unobserved health problems, but other attributes, such as job satisfaction and attitudes toward health care providers, likely also muddle the observed relationship between seeking medical care and missing work. In the jargon of econometrics, seeking medical care likely is endogenous with respect to missing work.

The second complication is that *reasons* for medical visits show substantial heterogeneity, and those reasons likely relate to the probability of missing work. Some visits to medical providers, which we label "curative," involve the diagnosis or treatment of some medical problem, while other visits, which we call "preventive," concern routine checkups and other wellness activities. Curative and preventive medical usage likely affects absenteeism differently for two reasons. First, preventive visits, presumably being more predictable and less urgent, might be easier to schedule around work hours. Second, the recently-passed Affordable Care Act includes provisions that nudge people toward preventive services and away from curative care, in the belief that shifting the mix of care will reduce aggregate medical expenses in the long run. Specifically, most copays for preventive services have been eliminated, while deductibles for curative services have increased (largely in response to the so-called "Cadillac Tax"). Some employers have gone further, by bringing medical professionals onsite, so that workers can obtain preventive services without missing work. Thus, it seems likely that curative and preventive care have different effects on absenteeism, and therefore should be considered separately.

To address this topic, we propose a recursive-style trivariate probit setup. Parameter estimation is based on a trust region algorithm, which exploits first and second order analytical derivatives of trivariate integrals. The method is quick and easy to use, and leads to easy-to-interpret treatment effect type calculations.

We find that, observationally, having a curative visit associates with a nearly 80 percent increase in the probability of missing work, while having a preventive visit correlates with a smaller 14 percent increase in the likelihood of missing work. Those numbers support the supposition that preventive visits, being more predictable and less urgent, are easier to schedule around work hours. However, after addressing potential endogeneity, neither type of visit appears to significantly relate to missing work. Therefore, we conclude that the observed links between medical usage and absenteeism derive from unobserved heterogeneity, rather than direct causal channels.

2 Data

Our proposed statistical model draws inspiration from, and is informed by, our empirical case study. Therefore, we discuss the data before turning to statistical and estimation details.

Data used in this study come from Medical Expenditure Panel Survey (MEPS), collected and published by the Agency for Healthcare Research and Quality, a unit of the U.S. Department of Health and Human Services. The MEPS enjoys a reputation as the most detailed and complete source of information on individual-level medical spending and usage in the U.S. Of particular importance for this study, the MEPS also includes rich information on individual-level employment-related details.

We focus on data from the 2012, 2013, 2014, and 2015 waves, which, at the time of this writing, are the most recent public releases that include details on office-related health

care usage. We extract individual-level socioeconomic information from the main "Full Year Consolidated Data" files, focusing on all males 20 years of age or older who report working full time (at least 35 hours per week) for the full calendar year. The main variable of interest is a binary indicator for whether the person missed any work for health-related reasons.

We then link that individual-level information to medical usage event-level details from the "Office-Based Medical Provider Visits" files. Crucially, those event-level files record the *reason* for visiting a medical provider, allowing us to ascertain whether the visit was for "diagnosis or treatment," which we label as *curative*, or for a "general checkup," which we label as *preventive*.

Table 1 presents sample means. The most important numbers, appearing near the top of the table, show that, among subjects who had neither curative nor preventive office-based visits, only 23 percent reported missing any work, compared to 55 percent who had *both* types of visits. The middle two panels, which focus on subjects who had one type of visit but not the other, suggest that curative visits correlate with far larger probabilities of missing work.

The table also reveals socioeconomic differences across the four medical usage categories. Having any type of office visit appears to positively correlate with age, education, and marital status. Subjects who work for employers that offer paid sick leave, or who work for government organizations, perhaps not surprisingly, appear more likely to report having either type of office-base visit.

Finally, the bottom of the table reports the proportion of subjects who claim that it is "not at all difficult" to contact their usual source of care (USC) by phone. We argue below, both economically and statistically, that this variable represents an appropriate identifying instrument, for two reasons. First, in the U.S., office-based visits with medical providers typi-

cally require an appointment, with those appointments usually arranged by phone. Therefore, the ease of contacting one's USC should predict the likelihood of having an office-based visit, which the numbers in Table 1 seem to confirm. Second, the ease of contacting one's USC should *not* relate to missing work, aside from indirectly through its link to office visits. Though not formally testable in our setting, we offer suggestive evidence of that lack of correlation below.

3 Trivariate probit model with endogenous binary regressors

The endogeneity issue can be understood in terms of a regression model from which important covariates have been omitted (since not readily available) and hence become part of the model's error term. To control for this form of unmeasured heterogeneity in the empirical context of this paper, the proposed model builds on a first equation modeling the first endogenous dummy variable, a second equation for the second endogenous dummy variable, and an outcome equation which determines the response variable and that depends on the endogenous binary regressors. The three equations are then connected via a multivariate distribution which makes it possible to model the correlations between the equations, hence accounting for unobserved heterogeneity. Below, we provide details on the model specification employed here, and briefly discuss estimation and inference.

3.1 Model definition

The model can be expressed in terms of latent responses as

$$y^*_{1i} = \mathbf{v}^\top_{1i}\boldsymbol{\gamma}_1 + \varepsilon_{1i}, \quad (1)$$

$$y^*_{2i} = \mathbf{v}^\top_{2i}\boldsymbol{\gamma}_2 + \varepsilon_{2i}, \quad (2)$$

$$y^*_{3i} = \psi_1 y_{1i} + \psi_2 y_{2i} + \mathbf{v}^\top_{3i}\boldsymbol{\gamma}_3 + \varepsilon_{3i}, \quad (3)$$

for $i = 1, \ldots, n$, where n is the sample size, y^*_{1i}, y^*_{2i} and y^*_{3i} are the latent continuous variables related to the endogenous and outcome variables such that $y_{mi} = 1$ if $y^*_{mi} > 0$ and 0 otherwise, for $m = 1, \ldots, 3$, $\mathbf{v}_{mi}$ contains (binary, categorical and continuous) covariates, vector $\boldsymbol{\gamma}_m$ represents the effects of the variables in $\mathbf{v}_{mi}$, and the error terms follow the Gaussian distribution $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})^\top \overset{iid}{\sim} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \vartheta_{12} & \vartheta_{13} \\ \vartheta_{21} & 1 & \vartheta_{23} \\ \vartheta_{31} & \vartheta_{32} & 1 \end{pmatrix}.$$

The error variances in $\boldsymbol{\Sigma}$ are as usual normalized to unity, while the off-diagonal elements represent the correlations between the error terms and $\vartheta_{kz} = \vartheta_{zk}$ for $z \neq k$. Each model's parameter takes values in $\mathbb{R}$, whereas the correlations take values in the range $[-1, 1]$.

Since the model includes only unidirectional effects (the endogenous variables affect the outcome but the outcome does not affect them), this system can be regarded as a special case of the recursive model discussed by Wilde (2000). As suggested in the introduction, this model can also be viewed as a simplified version of the trivariate model introduced by Filippou, Marra and Radice (2017). Although the model is theoretically identified (Wilde, 2000), consistent estimation of ψ_1 and ψ_2 is more reliably achieved in the presence of an instrument (an extra covariate in the model that is associated with the endogenous variables, is not directly related to the outcome, and is independent of the unobserved confounders) (e.g.,

Little, 1985). We therefore elect to include an instrumental variable in the set of regressors of (1) and (2) that is not included in (3).

Note that the adopted modeling framework allowed us to explore the use of more general versions of the model defined by equations (1), (2) and (3). For instance, following Filippou, Marra, Radice (2017) and Filippou et al. (2019), we considered different structures for the error terms and included in the model smooth functions of the continuous regressors in order to estimate their effects flexibly and in a data-driven manner. However, these developments did not alter the conclusions of our study, hence we refrained from presenting and discussing a more complicated statistical approach in this paper.

3.2 Parameter estimation

Given an observed random sample $(y_{1i}, y_{2i}, y_{3i})_{i=1}^{n}$, the log-likelihood of the model can be written as

$$\begin{aligned}\ell(\boldsymbol{\delta}) \;=\; & \sum_{i=1}^{n} \{y_{1i}y_{2i}y_{3i}\log(p_{111i}) + y_{1i}y_{2i}(1-y_{3i})\log(p_{110i}) + y_{1i}(1-y_{2i})y_{3i}\log(p_{101i}) + \\ & (1-y_{1i})y_{2i}y_{3i}\log(p_{011i}) + (1-y_{1i})(1-y_{2i})(1-y_{3i})\log(p_{000i}) + \\ & (1-y_{1i})(1-y_{2i})y_{3i}\log(p_{001i}) + (1-y_{1i})y_{2i}(1-y_{3i})\log(p_{010i}) + \\ & y_{1i}(1-y_{2i})(1-y_{3i})\log(p_{100i})\}\,, \end{aligned} \tag{4}$$

where $\boldsymbol{\delta} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top, \psi_1, \psi_2, \boldsymbol{\gamma}_3^\top, \vartheta_{12}, \vartheta_{13}, \vartheta_{23})^\top$, and the joint distributions of the three responses conditional on the model's covariates are denoted as $p_{\bar{e}_1\bar{e}_2\bar{e}_3 i} = \mathbb{P}(y_{1i} = \bar{e}_1, y_{2i} = \bar{e}_2, y_{3i} = \bar{e}_3)$ with $\bar{e}_m \in \{0, 1\}$, $\forall m$ (see Filippou, Marra, Radice (2017) for their mathematical definitions). Since the ϑ_{zk} can only take values in $[-1, 1]$, to facilitate estimation correlation coefficients were transformed according to $\vartheta^*_{zk} = \tanh^{-1}(\vartheta_{zk}) \in \mathbb{R}$. Positive-definiteness of $\boldsymbol{\Sigma}$ was achieved by including range restrictions; fixing ϑ_{13} and ϑ_{23} then ϑ_{12} was restricted to take values in

$\left(\vartheta_{13}\vartheta_{23} - \sqrt{(1-\vartheta_{13}^2)(1-\vartheta_{23}^2)}, \vartheta_{13}\vartheta_{23} + \sqrt{(1-\vartheta_{13}^2)(1-\vartheta_{23}^2)}\right)$. In practice, such a restriction was imposed using the eigenvalue method.

Estimation of $\boldsymbol{\delta}$ was achieved using a carefully constructed trust region algorithm. This required working with first and second order analytical derivatives of trivariate integrals which were tediously derived. The availability of analytical derivative information was essential for developing a computationally stable and efficient algorithm which accurately estimates the parameters of the model in practical situations. In fact, some experimentation using simpler optimization schemes or estimation approaches available in the literature (such as `mvprobit()` available in `STATA`) revealed that algorithmic convergence and estimation performance were generally problematic; see Filippou, Marra, Radice (2017) for full details. Finally, confidence intervals are obtained using $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}^{-1}\right)$, where the arguments of the multivariate Gaussian denote the estimated parameter vector and the inverse of minus the Hessian matrix. The reader is referred to Filippou, Marra, Radice (2017) and Filippou et al. (2019) for more technical details and discussions.

3.3 Average treatment effects

The effect of the endogenous variables y_{1i} and y_{2i} on the probability that y_{3i} is equal to 1, given covariate information, is of interest. For the case of y_{1i} and y_{3i}, this can be calculated using the following expression

$$P(y_{3i}|y_{1i}=1, y_{2i}, \mathbf{v}_{3i}^\top) - P(y_{3i}|y_{1i}=0, y_{2i}, \mathbf{v}_{3i}^\top),$$

where $P(y_{3i}=1|y_{1i}=1, y_{2i}, \mathbf{v}_{3i}^\top) = \Phi(\eta_{3i}^{(y_{1i}=1)})$, $P(y_{3i}=1|y_{1i}=0, y_{2i}, \mathbf{v}_{3i}^\top) = \Phi(\eta_{3i}^{(y_{1i}=0)})$, $\eta_{3i}^{(y_{1i}=\bar{e}_1)}$ denotes the predictor in the outcome equation evaluated at $y_{1i}=\bar{e}_1$, $\forall \bar{e}_1 = \{0,1\}$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal. Similarly, for the impact of y_{2i} on y_{3i}.

This is known as the causal treatment effect (TE, Angrist, 1996), and it measures the causal difference in outcomes between individuals that receive the treatment ($y_{1i} = 1$ or $y_{2i} = 1$) and individuals who do not receive it ($y_{1i} = 0$ or $y_{2i} = 0$). For each individual only one of the two potential outcomes can be observed; the other outcome is the counter-factual. The average TE (ATE) in a specific sample is given by $1/n \sum_{i=1}^{n} \text{TE}_i$ where TE_i denotes the TE of individual i (e.g., Abadie, 2004).

4 Simulation study

The aim of this section is to assess the empirical effectiveness of the recursive trivariate probit model employed in this paper. In the following, two binary endogenous variables, a binary response, an instrument, two observed confounders, and correlated error terms are denoted as y_1, y_2, y_3, x_3, x_1, x_2, u_1, u_2 and u_3, respectively. We simulated y_1, y_2 and y_3 using several trivariate distributions for the error terms (i.e., normal, Student's t and χ^2). The error correlations were set to 0.3 between u_1 and u_2, 0.6 between u_1 and u_3 and -0.2 between u_2 and u_3; several other combinations were tried out which did not alter the conclusions. Variables x_1, x_2 and x_3 were generated from uniform distributions over [0,1]. Non-linear covariate effects between y_1 and x_2 and between y_2 and x_2 were also introduced. The coefficients that relate y_1 and y_2 to y_3 were set to 0.5 and -0.5, respectively. The sample sizes were 1000 and 4000, respectively, and each scenario was replicated 500 times. Full details on the data generating process, using `R` syntax, are reported in the Appendix.

The findings are summarized in Table 2, which compares the results from the traditional univariate probit (ignoring endogeneity) and the recursive trivariate probit model. As expected, the results confirm that the recursive trivariate probit model is appropriate for correcting for endogeneity (bias = 2.44% and 1.50%, and RMSE = 0.27 and 0.29, for ψ_1 and ψ_2,

respectively, when the sample size is 1000) and that the traditional univariate probit model performs poorly (bias = 237.85%, 120.85%, RMSE = 1.19, 0.61, respectively). Table 2 also shows the results under misspecification of the model's distribution (Student's t and χ^2). As compared to the previous case, the performance of the trivariate model worsens, although it still has lower bias and RMSE as compared to those of the univariate probit model and the mean estimates are not far from the true values.

5 Case study

This section first explores the validity of the instrument. It then skips straight to the paper's main punchline. Finally, it presents and discusses some of the more nuanced findings from our trivariate model.

5.1 Instrument validity

Although our model is technically identified via nonlinear functional forms, more robust identification of our model hinges on our instrument – the ease of contacting one's usual source of care by phone – significantly affecting the likelihood of office visits, while also *not* affecting the likelihood of missing work, other than indirectly through its effect on having office visits.

To investigate the first condition, Table 3 reports probit estimates for each type of visit, with coefficients converted to marginal effects. Estimates for the instrument, appearing near the top of the table, reveal that being able to easily phone one's USC correlates increased probabilities of having a curative visit and a preventive visits by 11 and 13 percentage points, respectively. Those percentage point boosts translate to approximate 32 percent and 41 percent respective increases in probabilities, relative to sample means. Consequently, the

instrument appears to significantly and nontrivially affect the likelihood of each type of office visit.

To investigate the second condition, Table 4 reports probit marginal effects for the probability of missing work. Appearing near the top of the table, being able to easily phone one's USC exerts a small, but more importantly, insignificant influence on the likelihood of missing work. Though not a formal test of instrument excludability, the lack of significance of the instrument in Table 4 suggests its plausible exogeneity with respect to missing work.

In addition to the two aforementioned conditions, instrument validity also relies on a third, untestable, requirement that the instrument is independent of unobserved confounders. To afford protection against this concern, we include a set of detailed control variables, both person- and job-specific. For person-specific, we include age, race/ethnicity, education, marital status and family size. For job-specific, we include dummy indicators for whether the job offers paid sick leave, and whether the employer is some sort of state or federal government agency. We also include a set of occupation dummies. Our hope is that those control measures absorb most remaining variation stemming from unobserved confounders.

5.2 Main finding

Skipping ahead to our main punchline, Table 5 shows average treatment effects of office-based visits on the likelihood of missing work. The left-hand panel, under the header "ignoring endogeneity," shows estimates derived from simple probit models that do not correct for potential endogeneity of office visits. Those estimates suggest that having a curative visit increases the probability of missing work by almost 27 percentage points. Compared to the sample mean of missing work (0.34), that 27 point increase corresponds to an approximate 79 percent increase in the likelihood of missing work. Meanwhile, having a preventive visit leads

to a more modest 4.9 percentage point increase (14 percent relative to the mean) of missing work.

However, those estimates should be interpreted with caution, since unobserved attributes might simultaneously correlate with medical care usage *as well as* one's propensity to miss work. Shown in the right-hand panel of Table 5, our trivariate model, which attempts to accommodate such endogeneity bias, finds that neither type of office visit appears to significantly affect missed work. Instead, the observed link between medical care usage and missed work seems to derive almost entirely from unobserved heterogeneity that simultaneously drives both.

5.3 Full presentation of estimates

Table 6 presents estimation results from the full trivariate model. The top row shows that being able to easily phone one's USC increases the likelihood of having both types of visits. As for other control variables, blacks and Hispanics report fewer curative visits than their nonblack/nonHispanic counterparts, while Hispanics also report fewer preventive visits. Marriage positively correlates with visits, while family size negatively correlates. Working for an employer than offers paid sick leave increases the likelihood of both types of visits, as does working for a government agency.

The right hand panel of Table 6 reports estimates from the outcome equation. Blacks and Hispanics are less likely to miss work than their counterparts, family size negatively correlate with missing work, and paid sick leave and government employment positively correlate.

Finally, Table 7 shows estimates of the copula dependence parameters. First, the link between curative and preventive visits is positive, and precisely estimated, 0.271. The interpretation is that unobserved attributes that increase a person's likelihood of having a curative

visit also, perhaps not surprisingly, increase his chances of having a preventive visit. The next two rows reveal disparate patterns for endogeneity bias for the two types of visits. The positive dependence term in the second row of the table suggests that unobserved attributes that increase the likelihood of obtaining curative care also increase a person's chances of missing work. Such a pattern would be evident if unobserved (to the researcher) health problems simultaneously increase the probabilities of obtaining curative services *and* missing work. In the simple treatment effect reported in Table 5 that ignores endogeneity, that positive dependence becomes absorbed into the treatment effect, creating an upward bias. The third row, however, indicates much smaller endogeneity bias with respect to preventive care. Overall, the finding of such stark differences between those latter two dependence terms offers further evidence that curative and preventive services are different types of care, with distinctly different links to work absenteeism.

6 Conclusion

This paper proposes a trivariate model, where one equation models an outcome of interest, which depends, among other things, on two endogeneous variables. The other two equations model those two endogenous explanatory variables. The three equations are connected via a multivariate distribution, which makes it possible to model the correlations between the equations, hence accounting for unobserved heterogeneity.

An empirical application explores whether visiting a medical provider causes an employee to miss work. Ignoring endogeneity, we find large, and statistically significant, effects. Specifically, having a curative visit associates with a nearly 80 percent increase in the probability of missing work, while having a preventive visit correlate with a smaller 14 percent increase in the likelihood of missing work. However, after addressing potential endogeneity, neither

type of visit appears to significantly relate to missing work.

The proposed model should prove useful for empirical problems with two endogeneous explanatory variables. For example, husband and wife employment decisions might endogenously affect childcare decisions. Or the votes of two senators from the same state might endogenously affect economic conditions in that state. The model proposed in this paper offers an intuitive, and easy-to-estimate, route to explore such topics.

References

Abadie, A., Drukker, D., Herr, J. L., Imbens, G. W. (2004). "Implementing matching estimators for average treatment effects in stata", Stata Journal, 4, 290-311.

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). "Identification of causal effects using instrumental variables", Journal of the American Statistical Association, 91(434), 444-455.

Filippou, P., Kneib, T., Marra, G., Radice, R. (2019). "A Trivariate Additive Regression Model with Arbitrary Link Functions and Varying Correlation Matrix", Journal of Statistical Planning and Inference, 199, 236-248.

Filippou, P., Marra, G., Radice, R. (2017). "Penalized Likelihood Estimation of a Trivariate Additive Probit Model", Biostatistics, 18(3), 569-585.

Little, R.J.A. (1985). "A note about models for selectivity bias", Econometrica, 53(6), 1469-1474.

Marra, G., Radice, R. (2011). "Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity", Canadian Journal of Statistics, 39, 259279.

Nicholson, S., Pauly, M., Polsky, D., et al. (2006). "Measuring the effects of workloss on productivity with team production", Health Economics, 15, 111-123.

Pauly, M., Nicholson, S., Polsky, D., et al. (2008). "Valuing reductions in on-the-job illness: 'presenteeism' from managerial and economic perspectives", Health Economics, 17, 469-485.

Schultz, A., Chen, C., Edington, D. (2009). "The cost and impact of health conditions on presenteeism to employers a review of the literature", Pharmacoeconomics, 27, 365-378.

Strömberg,C.Aboagye, E., Hagberg, J., Bergström, G., Lohela-Karlsson, M. (2017). "Estimating the Effect and Economic Impact of Absenteeism, Presenteeism, and Work Environment–Related Problems on Reductions in Productivity from a Managerial Perspective", Value in Health, 20, 1058-1064.

Wilde, J. (2000). "Identification of multiple equation probit models with endogenous dummy regressors", Economics Letters, 69(3), 309-312.

Zhang, W., Bansback, N., Anis, A. (2011). "Measuring and valuing productivity loss due to poor health: a critical review", Social Science and Medicine, 72, 185-192.

Zhang, W., Sun, H., Woodcock, S., Anis, A. (2015). "Illness related wage and productivity losses: valuing 'presenteeism", Social Science and Medicine, 147, 62-71.

Appendix

Code to simulate data and estimate the models.

```
# Load library, set seed and sample size
library(GJRM)
set.seed(100)
n <- 1000

# Set correlation matrix of error terms
Sigma      <- matrix(0.6, 3, 3); diag(Sigma) <- 1
Sigma[1,2] <- Sigma[2,1] <- 0.3
Sigma[2,3] <- Sigma[3,2] <- -0.2

# Set correlation matrix of covariates
SigmaC <- matrix(0.5, 3, 3); diag(SigmaC) <- 1

# Function to generate non-linear covariate effects
f1 <- function(x) cos(pi*2*x) + sin(pi*x)
f2 <- function(x) x + exp(-30*(x - 0.5)^2)

# Generate errors from trivariate normal distribution
u <- rMVN(n, rep(0,3), Sigma)

# Generate covariates from multivariate uniform distribution
cov <- rMVN(n, rep(0,3), SigmaC)
cov <- pnorm(cov)
x1  <- cov[, 1]
x2  <- cov[, 2]
x3  <- cov[, 3]

# Generate the endogenous and response variables
y1 <- ifelse(-1 +       2*x1 - f1(x2) +      x3      + u[,1] > 0, 1, 0)
y2 <- ifelse(0.25 -  1.25*x1 + f2(x2) - 1.25*x3      + u[,2] > 0, 1, 0)
y3 <- ifelse(-0.75 + 0.25*x1 + x2 + 0.5*y1 - 0.5*y2  + u[,3] > 0, 1, 0)

# Construct a dataframe
dataSim <- data.frame(y1, y2, y3, x1, x2, x3, x4)

# Fit the recursive trivariate probit model
f.l <- list(y1 ~ x1 + s(x2) + x3,
            y2 ~ x1 + s(x2) + x3,
            y3 ~ x1 + x2 + y1 + y2)
out  <- gjrm(f.l, data = dataSim, Model = "T",
             margins = c("probit", "probit", "probit"))
```

To allow the error terms to be Student's t (with two degrees of freedom) or χ^2 (with two degrees of freedom) distributed, respectively, the above `R` code was be easily modified by replacing

```
u <- rMVN(n, rep(0,3), Sigma)
```

with

```
library(mvtnorm)
u <- rmvt(n, rep(0,3), sigma = Sigma, df = 2)
```

or with

```
library(copula)
norm.cop <- ellipCopula("normal", param = c(0.3, -0.6, 0.2),
                        dim = 3, dispstr = "un")
myMvd <- mvdc(copula = norm.cop, margins = c("chisq", "chisq", "chisq"),
              paramMargins = list(list(df = 2), list(df = 2),  list(df = 2)) )
u <- rMvdc(mvdc = myMvd, n = n)
```

Tables

Table 1: Sample means

	Curative visit = NO Preventive visit = NO n = 10,591	Curative visit = YES Preventive visit = NO n = 3,656	Curative visit = NO Preventive visit = YES n = 3,199	Curative visit = YES Preventive visit = YES n = 3,323
Miss any work	0.23	0.53	0.29	0.55
Age	39.4	42.5	46.3	49.5
Black	0.16	0.10	0.19	0.12
Hispanic	0.40	0.26	0.23	0.17
College degree	0.22	0.33	0.34	0.42
Married	0.54	0.65	0.68	0.71
Family size	3.49	3.15	3.04	2.81
Paid sick leave	0.47	0.62	0.64	0.67
Government job	0.09	0.16	0.17	0.20
Occupation	10 dummies	10 dummies	10 dummies	10 dummies
Easy to phone USC	0.21	0.39	0.41	0.45

Table 2: Mean, % bias and root mean squared error (RMSE) of the estimates obtained when fitting the traditional probit (univariate) and the recursive trivariate probit model (trivariate) to 500 data-sets generated using the trivariate normal ($\mathcal{N}$), Student's t and χ^2 distributions for the error terms. The sample sizes considered were 1000 and 4000. True values for ψ_1 and ψ_2 are 0.5 and -0.5, respectively

			Mean		% Bias		RMSE	
Distribution			Trivariate	Univariate	Trivariate	Univariate	Trivariate	Univariate
$\mathcal{N}$	$n = 1000$	ψ_1	0.49	1.69	2.44	237.85	0.27	1.19
		ψ_2	-0.49	-1.10	1.50	120.85	0.29	0.61
	$n = 4000$	ψ_1	0.50	1.69	0.20	238.09	0.13	1.19
		ψ_2	-0.50	-1.10	1.11	120.91	0.14	0.61
t	$n = 1000$	ψ_1	0.53	1.63	6.22	225.23	0.32	1.13
		ψ_2	-0.57	-1.03	15.04	107.04	0.35	0.54
	$n = 4000$	ψ_1	0.54	1.62	7.89	224.80	0.16	1.12
		ψ_2	-0.57	-1.03	14.75	106.36	0.17	0.53
χ^2	$n = 1000$	ψ_1	0.63	-0.38	26.53	176.13	0.29	0.89
		ψ_2	-0.64	-0.14	27.45	72.28	0.24	0.37
	$n = 4000$	ψ_1	0.65	-0.36	30.66	172.99	0.20	0.87
		ψ_2	-0.65	-0.14	29.94	71.41	0.18	0.36

Table 3: Marginal effects from probit models

	Curative visit		Preventive visit	
	Marg. eff.	St. err.	Marg. eff.	St. err.
Easy to phone USC	0.107	0.007	0.130	0.007
Age	0.005	0.0003	0.009	0.0003
Black	-0.138	0.008	-0.014	0.009
Hispanic	-0.089	0.008	-0.068	0.008
College degree	0.022	0.009	0.042	0.009
Married	0.059	0.008	0.061	0.008
Family size	-0.027	0.002	-0.030	0.002
Paid sick leave	0.066	0.007	0.072	0.007
Government job	0.061	0.011	0.041	0.011
Occupation dummies	yes		yes	

Table 4: Marginal effects from probit model

	Miss any work	
	Marg. eff.	St. err.
Easy to phone USC	−0.007	0.007
Curative visit	0.277	0.008
Preventive visit	0.054	0.008
Age	−0.002	0.0003
Black	−0.046	0.010
Hispanic	−0.057	0.008
College degree	−0.043	0.009
Married	−0.033	0.008
Family size	−0.011	0.002
Paid sick leave	0.088	0.007
Government job	0.090	0.011
Occupation dummies	yes	

Table 5: Treatment effects of office visits on missing work

	Ignoring endogeneity		Trivariate model	
	Estimate	95% interval	Estimate	95% interval
Curative visit	0.270	(0.258, 0.285)	0.117	(−0.046, 0.256)
Preventive visit	0.049	(0.037, 0.063)	0.036	(−0.101, 0.206)

Table 6: Main estimation results

	Curative visit		Preventive visit		Miss any work	
	Coeff.	St. err.	Coeff.	St. err.	Coeff.	St. err.
Easy to phone USC	0.293^{**}	0.020	0.374^{**}	0.020	—	—
Curative visit	—	—	—	—	0.327	0.256
Preventive visit	—	—	—	—	0.101	0.200
Age	0.015^{**}	0.001	0.028^{**}	0.001	-0.003^{**}	0.001
Black	-0.422^{**}	0.029	-0.042	0.029	-0.188^{**}	0.043
Hispanic	-0.257^{**}	0.023	-0.209^{**}	0.024	-0.200^{**}	0.026
College	0.061^{**}	0.025	0.123^{**}	0.026	-0.108^{**}	0.026
Married	0.168^{**}	0.023	0.185^{**}	0.024	-0.065^{**}	0.025
Family size	-0.076^{**}	0.007	-0.088^{**}	0.007	-0.040^{**}	0.007
Paid sick leave	0.187^{**}	0.021	0.215^{**}	0.022	0.270^{**}	0.022
Government job	0.167^{**}	0.029	0.117^{**}	0.030	0.266^{**}	0.030
Occupation dummies	yes		yes		yes	
Intercept	-0.975^{**}	0.052	-1.721^{**}	0.055	-0.370	0.075

** p < .10 * p < .05

Table 7: Estimated dependence parameters and related 95% intervals

	Estimate	95% interval
$\theta_{\text{curative visit, preventive visit}}$	0.271	(0.249, 0.291)
$\theta_{\text{curative visit, missed work}}$	0.252	(0.001, 0.454)
$\theta_{\text{preventive visit, missed work}}$	0.070	$(-0.083, 0.280)$