UvA Econometrics

Discussion Paper: 2009/04

The frequency of visiting a doctor: is the decision to go independent of the frequency?

Hans van Ophem

www.feb.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA Universiteit van Amsterdam

The frequency of visiting a doctor: is the decision to go independent of the frequency?

Hans van Ophem[1,2]
Department of Quantitative Economics
of the University of Amsterdam

September 2009

JEL: I11, C35.
Keywords: count models, hurdle-Poisson specification, copulas, correlation.

Abstract
In his analysis of the effects of the reform of the German healthcare system, Winkelmann (2004) investigates the number of doctor visits. He makes a distinction between the decision to go to a physician and the number of times the physician is visited in the observed time period. Winkelmann finds that there is no correlation between both decisions. This results appears to be far from straightforward since the primary driving force in both decision will be the health of the patient. From this perspective a significant correlation is expected. In this paper it is analysed whether the apparent zero correlation is actually true or comes from the way the relation between both decisions is modelled. My empirical analysis confirms the latter, but nevertheless also corroborates Winkelmann's main conclusions on the relevance of the explanatory variables used in his investigation.

1 Department of Quantitative Economics, Faculty of Economics and Econometrics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. Email: j.c.m.vanophem@uva.nl.
2 All ML-routines used in this paper are available on request. All estimations are carried out with R (free software, for information see http://www.r-project.org/).

1. Introduction

In Winkelmann (2004) it is argued that the process governing the frequency of visiting a medical doctor during a certain period of time, can be split up into two parts. In a first step the individual decides to visit a doctor or not. In the second step the decision on the actual positive number of visits, or more general the treatment plan, is made by the doctor in consultation with his patient. The treatment decision can be expected to be primarily made by the physician and the opinion of the individual might not put much weight on the scale. The important thing is that the first decision is made independently by the individual and that the second is not and that it is therefore not a good idea to use a model that does not explicitly take this difference into account. Clearly, the presented structure of the health decision is too simple. All the complexities of real life are impossible to capture in an econometric model. For instance, the decision to visit a doctor or not and consequently the decision on the frequency might have to be made a couple of times during the period considered, because an individual might fall ill or become injured a number of times. Another complexity that can be added is the introduction of medical specialists: often doctors refers their patients to specialists. This brings a third decision maker into the equation. These and other additional complexities will be ignored in this investigation, just like it was done in Winkelmann (2004). The modelling of the first complexity is explicitly taken into account in Santos Silva and Windmeijer (2001), however.

To model the potentially related decisions, Winkelmann (2004) decides to use a hurdle model. He recognizes that both decisions will usually not be made independently, and therefore Winkelmann allows for a correlation between both decision, although in a specific manner. First, let's discuss the reason for a correlation. Following Winkelmann (2004), I model the decision to go to a doctor or not by a simple probit model:

$$z_i = x_i' \gamma + \varepsilon_i \qquad (1)$$

z_i is an observation specific latent indicator variable that exceeds zero if the individual decides not to visit a doctor ($d_i = 1$ in this case) and is smaller or equal than zero if he decides to do so ($d_i = 0$). z_i could be seen as representing the health of the individual. Low values representing bad health and high values good health. This health indicator is dependent on exogenous factors represented in the vector x_i, Winkelmann (2004) uses variables like age,

years of education, labour market status and income, and an error term ε_i. If health gets too bad, passing the threshold 0, the individual decides to see a doctor. During this consultation a decision is made how to improve the condition of the individual, e.g. by prescribing drugs, visiting the doctor a couple of times to monitor the progress or referring the patient to a specialist. The health condition of the patient will be an important factor in both decisions and clearly the initial decision and the follow up decision are likely to be positively correlated.[3] As mentioned this is also the starting point of Winkelmann (2004) but it is abandoned after some time because the empirical results indicate that there does not exist a correlation. The result of a zero correlation strikes me as very odd and the question is whether this result origins from the way the correlation is introduced in the model by Winkelmann (2004) or because it is actually true. Here I will propose an alternative way of modelling the potentially correlated decision based on my previous work (van Ophem, 1999 and van Ophem, 2001). In an empirical analysis using the same data and more or less the same explanatory variables.[4] I will show that the result of zero correlation found by Winkelmann (2004) indeed holds, but that a significant correlation is estimated if a different type of modelling is used.

2. Econometric modelling

Winkelmann (2004) develops a hurdle model that he calls the Probit-Poisson-log-normal model. It consists of two parts: a hurdle has to be taken before the actual frequency can be determined. The hurdle is the decision to go to a doctor during the observed period or not. It is modelled by the probit-equation (1). The frequency of doctor visits during the period under observation, is modelled by assuming that this count has a truncated Poisson distribution with expectation λ_i. This expectation is specified as:

$$\lambda_i | \upsilon_i = \exp(x_i' \beta + \upsilon_i) \qquad (2)$$

A truncated Poisson distribution is used because of the existence of the hurdle. As a result zero counts need not to be modelled by the count process. The introduction of the error term

3 Due to the specific modelling choices, actually a negative correlation is expected between both decisions: bad health (low z, cf eq. (1)) goes together with a higher frequency of doctor visits.

4 I decided not to use exactly the same explanatory variables to circumvent identification of the estimated parameters only because of the nonlinearity of the model (see further). The results reported here do hardly differ if exactly the same specification is used but it is difficult estimate the standard errors precisely due to the large degree of multicollinearity.

υ_i is meant to capture unobserved heterogeneity and is the crux to allow for correlation between the first and second step. Winkelmann (2004) assumes that ε_i and υ_i are bivariate normal distributed with expectation (0, 0), variance (1, σ2) and correlation ρ. Due to the specific nature of modelling the decision to see a doctor or not, a negative correlation is expected. To give an example, a hypochondriac can be expected to visit a doctor frequently because of his bad health perception, so a negative ε_i combined with a large υ_i are likely. An implicit assumption of Winkelmann (2004) is that the not completely observed health indicator is the primary driving force in the determination of the frequency of doctor visits. To make this explicit assume that other factors influencing the frequency are captured in a w_i. This vector can contain variables also influencing health status but also factors that characterize the physician. We then arrive at the follow specification of the expected frequency.

$$\lambda_i = \exp(\alpha z_i + w_i'\theta + \upsilon_i) = \exp(\alpha x_i'\gamma + w_i'\theta + (\upsilon_i + \alpha\varepsilon_i)) \qquad (3)$$

where the substitution is made because the health indicator is not actually observed. This shows two things. First, the error term introduced is very likely to be correlated with the error term in the hurdle equation (1), especially if $\alpha \neq 0$, as Winkelmann (2004) suggests.[5] Second, the parameter vector related to x_i differs from the one in the hurdle equation to the one in the count equation with more than only a factor of proportionality (α) if w_i contains (part of) x_i. One could say that Winkelmann (2004) implicitly assumes that $x_i = w_i$ since he uses x_i as explanatory variables both in the hurdle and count specification. Note that there is no actual need to make this assumption.

Winkelmann (2004) then discusses the practical usefulness of the model that allows for correlation and concludes that it is limited due to identification problems of the correlation. He then proceeds under the assumption of zero correlation, although he claims that the actual estimation of the model that allowed for a correlation yielded an insignificant correlation. Note however, that despite the potentially limited practical usefulness of a model with correlation, it is the only way to proceed if the correlation is indeed nonzero. Otherwise, inconsistent estimation results will be found. A reason for the apparent identification problem

5 The arguments of Winkelmann (2004) on why there exists a correlation are for a large part based on the inclusion of ε_i in his υ_i that equals $\upsilon_i + \alpha\varepsilon_i$ in the present notation. In other words he simply assumes $\alpha \neq 0$ (present notation).

is Winkelmann's choice of specification: he uses exactly the same explanatory variables in both steps of his analysis thereby introducing considerable multicollinearity. Although, given the former arguments (cf. (3)), this is a logical choice, it will be abandoned in the present study. Some insignificant variables will be deleted from the specification to enhance identification.

Equation (3) also shows something else: if υ_i and ε_i are uncorrelated then the correlation, in absolute value, between the error terms in (1) and (3) will be smaller than 1, except if $\upsilon_i = 0$. Only if the assumption of zero correlation between υ_i and ε_i is dropped a correlation of +1 or -1 between $\upsilon_i + \alpha\varepsilon_i$ and ε_i is possible. This correlation is not equal to the correlation between the hurdle and the count processes, however. For this last correlation, the additional randomness of the count should be taken into account. To illustrate this, consider the special case that $\upsilon_i = 0$. In this case, the correlation between the error terms in (1) and (3) is 1. Despite of this, conditioning on ε_i does not eliminate the randomness of the count process completely. The correlation of 1 can be considered to eliminate only the randomness in the unconditional expectation of the count. To put it differently, even if we know λ_i for sure, there is still randomness in the count process. Ignoring this additional randomness, or actually the correlation it can have with ε_i , causes that the correlation between the hurdle and the count process is lower than 1, in absolute value, even if υ_i and ε_i have full correlation. Winkelmann (2004) ignores this second source of correlation. Lindeboom and van den Berg (1994) report a similar kind of reasoning for duration models.

To model the potential correlation between the hurdle and count fully, use will me made of a bivariate copula, see the appendix for a brief discussion of this technique. The first application of the copula technique in economics is discussed in Lee (1983). Recently, copula's have regained new interest, see for instance, Cameron, Li, Trivedi and Zimmer (2004) or Zimmer and Trivedi (2006). Copula's offer a method to relate two (or more) marginal distributions, if necessary from different families, to each other and allow for the estimation of a correlation between the stochastic processes. I will use the normal or Gaussian copula. The underlying stochastic variables, in the present case the decision to see a doctor or not (eq. (1), where the error already is assumed to be normally distributed) and the number of doctor visits (a Poisson distributed count variable with unobserved heterogeneity, eq. (2)) are transformed to normal distributed random variables and then evaluated using a bivariate normal distribution. For a full discussion of the technique, see Trivedi and Zimmer (2005).

The transformation of a essentially discrete random variable like a count is not straightforward. The way to do this, is discussed in van Ophem (1999).

The use of the copula technique requires the specification of the marginal distributions. This is sometimes seen as an disadvantage since the marginals are rarely known (cf. Trivedi and Zimmer, 2005., p. 96). However, it is questionable whether this is actually a disadvantages compared to alternative ways of modelling the problem. As such it does not really make a difference whether the bivariate distribution is specified or the corresponding marginals since there usually exists a one-to-one relation. To illustrate this for the present case, consider the specification used by Winkelmann (2004). Winkelmann (2004) starts with specifying the decision to go and see a doctor as in (1), where he assumes a normally distributed error (ε_i) . The number of doctor visits is modelled with a Poisson distribution that allows for unobserved heterogeneity as specified in (2). The unobserved heterogeneity term υ_i is assumed to be normally distributed. Correlation between both random processes is allowed by letting υ_i and ε_i have a (correlated) bivariate distribution. To use the copula technique the marginal distributions have to be specified, but they can be imputed directly from the distributional assumption made by Winkelmann (2004). The marginal of the hurdle specification remains exactly the same, whereas for the marginal of the count, the unobserved heterogeneity term has to be integrated out of the specification where again it is assumed that υ_i is normally distributed. Even though I will make exactly the same distributional assumption as Winkelmann (2004), using the copula technique allows for a more general specification of the correlation. Not only υ_i and ε_i are allowed to be correlated but also the count and the error of the decision to visit a doctor are allowed to be correlated, a possibility that is ignored in Winkelmann (2004). Note that although integrating out the unobserved heterogeneity term, looks like complicating things, this is in fact not true. Also in the correlated specification of Winkelmann (2004) the unobserved heterogeneity term has to be integrated out.

Estimation results of the following models will be presented in this paper:

- The Probit-Poisson-log-normal-model with zero correlation. This is the preferred model in Winkelmann (2004) and is constituted by eqs. (1) and (2).
- The Probit-Poisson-log-normal-model with partial correlation. This is the same model as the previous one except that now account is taken of the potential nonzero correlation between the error term in the probit equation (ε_i) and the unobserved heterogeneity term in the Poisson distributed count (υ_i).

- The Probit-Poisson-log-normal-model with full correlation using copula techniques.

3. Data

Use will be made from the German Socio-Economic Panel (GSOEP).[6] The data are the same as used by Winkelmann (2004) except that I only concentrate on the 1999-wave. As I will show, the essential result of zero correlation in the Winkelmann-set up will be retained. Winkelmann's primary goal is to estimate the effect of reforms in German health care, and therefore a comparison in time is essential. Here, I only want to show that the result of zero correlation is due to modelling choice and not due to the actual absence of correlation or to multicollinearity. To avoid identification only because of the nonlinearities in the model, I will also use a somewhat different set of explanatory variables than Winkelmann. I will not use exactly the same explanatory variables in the choice to go to a doctor and the choice on the frequency of the visits. Several explanatory variables appeared to be not relevant for one, or even both, of these choice processes.

The 1999-wave of GSOEP consists of 6231 observations. Respondents were asked to report their number of visits to a physician during the three months prior to the interview. Unfortunately visits to general practitioners and dentists and other specialists are not distinguished and included in this definition. 2156 individuals (34.6%) did not visit a doctor during this period. The frequency of doctor visits is given it Table 1.

- insert Table 1 -

The average number of doctor visits is 2.391 and the corresponding standard deviation is 3.943.

The definitions of the explanatory variables are as follows:[7]

- active sport: dummy variable equal to one if the respondent participates in sports at least once a week
- age: age of the respondents in years

6 The data are available in the data archive of the Journal of Applied Econometrics (http://qed.econ.queensu.ca/jae/).

7 The description is based on the information supplied in Winkelmann (2004). More information on the data can be found there as well.

- age2: (age*age/100) - mean(age/10))^2[8]
- bad health: dummy variable equal to one of the respondent classifies his own health as either 'very bad' or 'bad' ('fair' is the reference category)
- education in years: years of education of the respondents
- fall: dummy variable equal to one if the respondent was interviewed in the fall of 1999 (summer is the reference category)
- full-time: a dummy variable equal to 1 if the respondent is full-time employed at the time of the interview (self-employed is the reference category)
- good health: dummy variable equal to one of the respondent classifies his own health as either 'very good' or 'good' ('fair' is the reference category)
- household size: the number of persons living in the household the respondent belongs to
- log(income): the logarithm of household equivalent income (OECD-scale)
- male: dummy variable equal to one if the respondent is a male
- married: dummy variable equal to one if the respondent is a married
- part-time: a dummy variable equal to 1 if the respondent is part-time employed at the time of the interview (self-employed is the reference category)
- social assistance: a dummy variable equal to one if the respondent receives welfare payments (self-employed is the reference category)
- spring: dummy variable equal to one if the respondent was interviewed in the spring of 1999 (summer is the reference category)
- unemployed: a dummy variable equal to 1 if the respondent was unemployed at the time of the interview (self-employed is the reference category)
- winter: dummy variable equal to one if the respondent was interviewed in the winter of 1999 (summer is the reference category)

A constant is included in all estimations. The means, standard deviations, maxima and minima of the explanatory variables are presented in Table 2.

- insert Table 2 -

8 This specification of 'age squared' is chosen to diminish multicollinearity.

4. Estimation results

To start with, I will replicate the optimal model presented in Winkelmann (2004). After an elaborate comparison with the standard Poisson model, several count models with unobserved heterogeneity of different forms, hurdle models, a finite mixture model and a multi-episode model, Winkelmann concludes that the Probit-Poisson-log-normal model performs best. The Probit-Poisson-log-normal model is a specific form of a hurdle model where the the decision to go and see a doctor or not is modelled by a probit model and the decision on the frequency is modelled by a truncated Poisson-model with a normal distributed unobserved heterogeneity term. Both decisions, or actually the error terms, are uncorrelated. Table 3 presents the estimation results of this model.

- insert Table 3 -

The estimation results show a large resemblance with the ones presented in Table IV of Winkelmann (2004). Recall that the results are not exactly replicated due to using only the 1999-wave and using different sets of explanatory variables for both the probit and the count model. Variables left out of either the probit or the count were not significantly different from zero. Insignificant explanatory variables were included if they belong to a category from which one of the other explanatory variables is significant. Also, a constant, age and its square were always included because they are considered to be important determinants of doctor visits (see e.g. Deb and Trivedi, 1997, 2002, Prieger, 2002, Riphahn, Wambach, and Million, 2003). The significance is somewhat reduced if we compare the results in Table 3 with the corresponding ones in Winkelmann, probably due to the reduction in the number of observations. Still, many explanatory variables are strongly significant.

The estimation results with respect to the decision to go to a doctor or not are straightforward. Younger individuals and males are less likely to visit a doctor. People in good (bad) health are less (more) likely to do so. Married people and people living is small households are likely to visit a doctor more often. A somewhat counter-intuitive result is that those participating actively in sports are more bound to see a doctor than those who do not. This is likely to be due to the increased probability of suffering injuries while engaging in sports. People with higher incomes have a higher probability of going to see a doctor once or more. Full-timers are less likely to visit doctors. That the same conclusion holds for the

unemployed is somewhat surprising. Also with respect tot the number of doctor visits, given that an individual goes at least once, it can be concluded than the estimation results are in accordance with expectations and the results presented in Winkelmann (2004). A result that is at odds is the effect of age: there appears to be no effect in my estimations, whereas Winkelmann found a significant impact. However, note that Winkelmann's estimated impact is close to being not significant and given the reduction of the number of observations, my findings are actually not a real surprise. Still, the conclusion itself is meaningful. The elderly have a higher probability to visit a doctor in a certain period but do not have a higher frequency of visits, once the first effect is taken into account. Higher educated individuals have a lower frequency of visits, just like those working full-or part-time or being unemployed. Healthy individuals also have a lower frequency of doctor visits. Participating actively in sports does not appear to have an impact on the frequency. Note the strong significance of the variance of the unobserved heterogeneity component. This confirms, Winkelmann's conclusion that unobserved heterogeneity should be taken into account.

- insert Table 4 -

Table 4 presents the estimation results of Winkelmann's Probit-Poisson-log-normal model with allowance for a nonzero correlation. As discussed in section 2, the correlation is between the error term of the probit hurdle and the unobserved heterogeneity term. As Winkelmann found, although he did not present the estimation results, the correlation is not significant. It has the correct negative sign, indicating that those more likely to see a doctor at least once (ε_i relatively small) have a higher chance of having a relatively large unobserved heterogeneity v_i and therefore to visit a doctor more often. Due to the insignificance of the correlation, the other estimates and the corresponding standard errors are very similar to those presented in Table 3.

- insert Table 5 -

In Table 5 the estimation results of the copula-based model that takes full account of the correlation is presented. The estimated correlation is about -0.4 and strongly significant wit a t-value of 4.9. This result indicates that there exists a much more direct correlation between

the two decisions distinguished than that can be captured through the unobserved heterogeneity component. The estimates on the decision to see a doctor or not are very similar to the ones presented earlier both in size and significance. Some larger deviations are found for the estimated count distribution but there are no sign changes. As a result, the conclusions drawn with respect to Table 3 remain valid. Due to the significant correlation estimated the loglikelihood-value based on the copula technique is smaller than that of the model with $\rho = 0$. As a result, the model based on copula's should perform even better than the Probit-Poisson-log-normal model advocated by Winkelmann (2004). Note, that the model with no correlation is a special case of the model based on copula's.[9] Despite of the significance of the correlation, the Schwartz Information Criterium (SIC) indicates that we should actually prefer the Probit-Poisson-log-normal model with zero correlation (SIC = 24070.97 and 24073.30). This indicates that the SIC is not always a good measure to use.

To explore the consequences of the different estimation methods with respect to the estimated mean of the count distribution given that the individual goes to a physician at least once, consider Table 6. This table presents estimates of the conditional mean for the complete sample, the sample with positive doctor visits and the sample with zero doctor visits. It also gives information on the variance of the mean across the entire sample and its maximum and minimum and presents information on the observed count of doctor visits.

- insert Table 6 -

The results indicate that despite the significance of the correlation the estimated means and variances are very similar. The relatively large difference between the observed mean count of the complete sample and the mean of the estimated counts points at a well known problem in the estimation of count data: there appear to be excess zero observations. The corresponding estimates for the subsample of positive counts are very close to the actual observation, where the estimates based on the copula estimates are closest. The estimates of the variance of the count differ considerably from the variances of the observations. This comes as no surprise since the maximum of the estimated means of each of the three methods is much smaller than the maximum observed count. There does exist a notable difference between the estimated

9 Indeed, if I impose $\rho = 0$ on the model based on the copula, exactly the same likelihood value is found. This indicates that the procedures needed for the copula, i.e. the inverse of the univariate standard normal distribution and the cumulative bivariate standard normal distribution, are very precise.

maximum count across the estimation methods: for the copula technique it is clearly higher than for the models proposed by Winkelmann (2004). All in all, we have to conclude that the differences between the characteristics of the estimates Poisson distribution are very small, despite of the strongly significant correlation between the choice processes distinguished. It appears that neglecting the correlation does not alter the predictions of the model. I investigated this conclusion further by calculating the root mean squared errors (between estimated mean and actual observation), root absolute errors, correlations etc., but this only gave a confirmation. All these measures differ only in the third or higher digits across the alternative measures. The correlation between the estimated means using the copula technique and the model with correlation 0, is extremely high: 0.998. The overall conclusion has to be that, although he ignored significant correlation, Winkelmann's conclusions are correct. The estimated correlation of about -0.4, has only a very marginal effect on the estimated probabilities.

5. Conclusion

To investigate the surprising result of no correlation found in Winkelmann (2004), I reestimated the models proposed by Winkelmann and estimated an alternative specification allowing for more general correlation. My analysis confirms Winkelmann's finding of no correlation in his setup, but does find a significant correlation of -0.4 if an alternative estimation methods is employed. Despite of the strong significance of the correlation, it hardly influences important characteristics of the underlying distribution resulting in a confirmation of the empirical conclusions drawn in Winkelmann (2004).

References

Cameron, A.C., T. Li, P.K. Trivedi, and D.M. Zimmer, 2004, Modelling the differences in counted outcomes using bivariate copula model with application to mismeasured counts, *Econometrics Journal*, vol. 7, pp. 566-584.

Deb, P., and P.K. Trivedi, 1997, Demand for medical care by the eldery: a finite mixture approach, *Journal of Applied Econometrics*, vol. 12, pp. 313-336.

Deb, P., and P.K. Trivedi, 2002, The structure of demand for health care: latent class versus two-part models, *Journal of Health Economics*, vol. 21, pp. 601-625.

Lee, L., 1983, Generalized econometric models with selectivity, *Econometrica*, vol. 51, pp. 507-512.

Lindeboom, M., and G. van den Berg, 1994, Heterogeneity in models for bivariate survival: the importance of the mixing distribution, *Journal of the Royal Statistical Society, Series B*, vol. 56, pp. 49-60.

Prieger, J., 2002, A flexible parametric selection model for non-normal data with application to health care usage, *Journal of Applied Econometrics*, vol. 17, pp. 367-392.

Riphahn, R.T., A. Wambach, and A. Million, 2003, Incentive effects in the demand for health care: a bivariate panel count data estimation, *Journal of Applied Econometrics*, vol. 18, pp. 387-405.

Santos Silva, J.M.C., and F. Windmeijer, 2001, Two-part multiple spell models for health care and demand, *Journal of Econometrics*, vol. 104, pp. 67-89.

Sklar, A., 1959, Fonctions de repartitiona n-dimensions et leurs marges, *Publication de l'Institute de Statistique de l'Universite de Paris*, vol. 8, pp. 229-231.

Trivedi, P.K., and D.M. Zimmer, 2005, Copula modeling: An introduction for practitioners, *Foundations and Trends in Econometrics*, vol. 1, pp. 1-111.

van Ophem, H., 1999, A general method to estimate correlated discrete variables, *Econometric Theory*, vol. 15, pp. 228-237.

van Ophem, H., 2000, Modeling selectivity in count data models, *Journal of Business & Economic Statistics*, vol. 18, pp. 503-511.

Winkelmann, R., 2004, Health care reform and the number of doctor visits - an econometric analysis, *Journal of Applied Econometrics*, vol. 19, pp. 455-472.

Zimmer, D.M., and P.K. Trivedi, 2006, Using trivariate copulas to model sample selection and treatment effects: application to family health care demand, *Journal of Business & Economics Statistics*, vol. 24, pp. 63-76.

Appendix: Copula's

The copula-technique was first introduced in econometrics by Lee (1983), although he did not use the term 'copula'. The idea originates from Sklar (1959). The copula approach is a useful method for deriving joint distributions given the marginal distributions, especially when the random variables are not normally distributed. I will concentrate here on the Gaussian or normal copula. Alternatives are discussed in e.g. Trivedi and Zimmer (2005).

Consider two random variables u and v with known marginal distributions $F_u(u)$ and $F_v(v)$. The transformed random variables $u^* = \Phi^{-1}(u)$ and $v^* = \Phi^{-1}(v)$ are standard normal distributed, where $\Phi^{-1}(.)$ is the inverse of the standard normal univariate cumulative distribution function. These transformed random variables can be related to each other by using the (standard) normal bivariate distribution. To accommodate a discrete or count random variable, use can be made of van Ophem (1999). The basic idea is to use the following identity:

$$Pr(u \leq k) = \Phi(\eta_k) = \Phi(\Phi^{-1}(\sum_{j=0}^{k} Pr(u = j)))$$

$$Pr(v \leq p) = \Phi(\lambda_p) = \Phi(\Phi^{-1}(\sum_{j=0}^{p} Pr(v = j)))$$

The bivariate probability (with nonzero correlation) $u \leq k$ and $v \leq p$ can now be written as:

$$Pr(u \leq k, v \leq p) = B(\eta_k, \lambda_p; \rho)$$

where B(.,.;.) denotes the bivariate normal cumulative distribution with mean (0,0), variance (1,1) and correlation ρ. η_k and λ_p depend on the parameters of the original marginal distributions. Maximization of the likelihood function is done across the original parameters and ρ.

Using the Gaussian copula has the advantage that ρ can take any value between -1 and 1. Computer routines to calculate the inverse of the standard normal cumulative distribution and the bivariate normal distribution are readily available in many software packages and usually yield high precision results.

Table 1: The observed number of doctor visits.

Number of visits	Frequency (%)	Number of visits	Frequency (%)	Number of visits	Frequency (%)	Number of visits	Frequency (%)
0	2156 (34.6%)	10	148 (2.4%)	21	1 (0.0%)	40	5 (0.1%)
1	1215 (19.5%)	11	3 (0.0%)	22	2 (0.0%)	45	1 (0.0%)
2	921 (14.8%)	12	51 (0.8%)	24	9 (0.1%)	48	1 (0.0%)
3	684 (11.0%)	13	7 (0.1%)	25	1 (0.0%)	50	2 (0.0%)
4	352 (5.7%)	14	7 (0.1%)	26	1 (0.0%)	60	2 (0.0%)
5	240 (3.9%)	15	38 (0.6%)	27	1 (0.0%)		
6	190 (3.1%)	16	6 (0.1%)	30	5 (0.1%)		
7	48 (0.8%)	17	3 (0.0%)	33	1 (0.0%)		
8	61 (1.0%)	18	2 (0.0%)	35	2 (0.0%)		
9	28 (0.5%)	20	35 (0.6%)	36	2 (0.0%)		

N = 6231. Empty cells are deleted.

Table 2: Descriptive statistics of the explanatory variables

Variable	Mean	Standard dev.	Maximum	Minimum
active sport	0,264	0,442	1	0
age	38,920	11,230	60	20
age2	1,261	1,253	0,000	4,444
bad health	0,129	0,335	1	0
education in years	11,330	2,364	7	18
fall	0,013	0,115	1	0
full-time	0,537	0,499	1	0
good health	0,580	0,494	1	0
household size	3,087	1,329	1	11
log(income)	7,524	0,432	5,636	9,420
male	0,466	0,499	1	0
married	0,649	0,477	1	0
part-time	0,114	0,318	1	0
social assistance	0,033	0,178	1	0
spring	0,531	0,499	1	0
unemployed	0,075	0,263	1	0
winter	0,317	0,465	1	0

Table 3: Estimation results of the Probit-Poisson-log-normal model with zero correlation.

	Probit	Truncated Poisson-log-normal
constant	0,025 (0.359)	1.367 (0.132)**
age	0,002 (0.002)	-0.000 (0.002)
age2	-0,042 (0.015)**	0.011 (0.015)
male	0,437 (0.039)**	-0.106 (0.038)**
education in years		-0.017 (0.008)*
married	-0,111 (0.043)**	
household size	0,041 (0.015)**	-0.040 (0.013)**
active sport	-0,141 (0.040)**	0.028 (0.040)
good health	0,496 (0.040)**	-0.461 (0.040)**
bad health	-0,562 (0.068)**	0.656 (0.046)**
full-time	0,087 (0.048)*	-0.174 (0.043)**
part-time	0,065 (0.063)	-0.269 (0.059)**
unemployed	0,204 (0.071)**	-0.149 (0.070)*
social assistance	-0,057 (0.068)	0.123 (0.091)
log(income)	-0,129 (0.046)**	
spring	-0,025 (0.051)	
fall	-0,317 (0.164)*	
winter	0,015 (0.055)	
variance unobserved heterogeneity		0.782 (0.016)**

Log-likelihood = -11900.06. Absolute asymptotic standard errors between parentheses. **/* = significant at 1%/5% (two-sided test).

Table 4: Estimation results of the Probit-Poisson-log-normal model with partial correlation.

	Probit	Truncated Poisson-log-normal
constant	0,027 (0.359)	1.392 (0.142)**
age	0,002 (0.002)	-0.000 (0.002)
age2	-0,042 (0.015)**	0.010 (0.015)
male	0,437 (0.039)**	-0.090 (0.052)
education in years		-0.017 (0.008)*
married	-0,112 (0.044)**	
household size	0,041 (0.015)**	-0.038 (0.014)**
active sport	-0,141 (0.040)**	0.023 (0.042)
good health	0,496 (0.040)**	-0.443 (0.057)**
bad health	-0,562 (0.068)**	0.640 (0.057)**
full-time	0,087 (0.048)*	-0.171 (0.043)**
part-time	0,065 (0.063)	-0.268 (0.059)**
unemployed	0,205 (0.072)**	-0.141 (0.072)*
social assistance	-0,060 (0.101)	0.124 (0.091)
log(income)	-0,129 (0.046)**	
spring	-0,022 (0.052)	
fall	-0,318 (0.164)	
winter	0,016 (0.055)	
variance unobserved heterogeneity		0.783 (0.018)**
correlation	-0.092 (0.200)	

Log-likelihood = -11899.96. Absolute asymptotic standard errors between parentheses. **/* = significant at 1%/5% (two-sided test).

Table 5: Estimation results of the Probit-Poisson-log-normal model with full correlation using a copula.

	Probit	Truncated Poisson-log-normal
constant	-0,022 (0.354)	1.225 (0.142)**
age	0,001 (0.002)	-0.000 (0.002)
age2	-0,036 (0.015)*	0.019 (0.017)
male	0,423 (0.040)**	-0.203 (0.047)**
education in years		-0.017 (0.008)*
married	-0,094 (0.043)*	
household size	0,040 (0.015)**	-0.047 (0.014)**
active sport	-0,140 (0.040)**	0.061 (0.043)
good health	0,487 (0.040)**	-0.590 (0.049)**
bad health	-0,559 (0.067)**	0.752 (0.055)**
full-time	0,091 (0.048)*	-0.193 (0.046)**
part-time	0,064 (0.062)	-0.286 (0.064)**
unemployed	0,206 (0.072)**	-0.197 (0.076)*
social assistance	-0,039 (0.106)	0.126 (0.094)
log(income)	-0,119 (0.045)**	
spring	-0,048 (0.051)	
fall	-0,283 (0.170)	
winter	-0,003 (0.054)	
variance unobserved heterogeneity		0.851 (0.028)**
correlation	-0.414 (0.084)**	

Log-likelihood = -11896.87. Absolute asymptotic standard errors between parentheses. **/* = significant at 1%/5% (two-sided test).

Table 6: Observed and estimated mean number of doctor visits, its variance, its maximum and, its minimum.

	Mean	Variance	Maximum	Minimum
Observed count				
Complete sample	2,391	15,551	60,000	0,000
Subsample count > 0	3,656	19,150	60,000	1,000
Winkelmann $\rho = 0$				
Complete sample	3,353	2,156	9,590	1,935
Subsample count > 0	3,618	2,633	9,590	1,935
Subsample count = 0	2,852	0,870	8,717	1,936
Winkelmann $\rho \neq 0$				
Complete sample	3,353	2,156	9,528	1,927
Subsample count > 0	3,618	2,630	9,528	1,927
Subsample count = 0	2,850	0,872	8,701	1,927
Copula				
Complete sample	3,373	2,356	11,258	1,974
Subsample count > 0	3,646	2,901	11,258	1,974
Subsample count = 0	2,857	0,917	9,186	1,977