UvA-DARE (Digital Academic Repository)

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

van Ophem, H.

Publication date

2009

Document Version

Submitted manuscript

[Link to publication](#)

Citation for published version (APA):

van Ophem, H. (2009). *The frequency of visiting a doctor: is the decision to go independent of the frequency?* (UvA-Econometrics Discussion Paper; No. 2009/04). Faculteit Economie en Bedrijfskunde.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

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

March 2011

JEL: I11, C35.

Keywords: replication, 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 result 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, I first replicate part of Winkelmann's research. I then set off to analyse whether the 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 The comments of three referees are gratefully acknowledged. 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/>). This paper is a shorter version of a paper that is available at <http://www1.fee.uva.nl/pp/jcmvanophem>. This extended version contains additional estimation results, more detailed information on the data etc.

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.³ 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.⁴ Another complexity that can be added is the introduction of medical specialists: often doctors refer 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).

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:

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

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,

³ The idea of splitting up the choice process into two parts can be traced back to Stoddart and Barer (1981).

⁴ This complexity is modelled by Santos Silva and Windmeijer (2001). They also use a hurdle model that allows for correlation. They find little evidence of correlation.

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.⁵ 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 originates 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 that allows for a less restrictive specification of the correlation and that is based on my previous work (van Ophem, 2000).

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, y_i , is modelled by assuming that this count has a truncated Poisson distribution with expectation λ_i . This expectation is specified as:⁶ $\lambda_i | u_i = \exp(x_i' \beta + u_i)$. 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 u_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 u_i are bivariate normal distributed with expectation (0, 0), variance (1, σ^2) and correlation ρ . An implicit assumption of Winkelmann (2004) is that the not completely observed health indicator z_i is an important driving force in the determination of the frequency of doctor visits. To make this explicit, I add this factor to the specification and collect the other explanatory variables in the vector w_i .

5 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.

6 I use a slightly different notation for the error term than Winkelmann (2004). This is done to illustrate that the error structure of Winkelmann (2004) contains an implicit assumption. See below eq. (2).

This vector can contain variables that influence health status but also e.g. factors that characterize the physician. We then arrive at the follow specification of the expected frequency:

$$\lambda_i = \exp(\alpha z_i + w_i' \theta + v_i) = \exp(\alpha x_i' \gamma + w_i' \theta + (v_i + \alpha \varepsilon_i)) \quad (2)$$

where the substitution is made because the health indicator z_i is not actually observed. Winkelmann (2004) specifies $\lambda_i = \exp(x_i' \beta + u_i)$. Equation (2) 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.⁷ 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 . Winkelmann (2004) specifies $x_i = w_i$ since he uses x_i as explanatory variables both in the hurdle and count specification. Since my emphasis is on replication, I will use the specification with $x_i = w_i$, as well. 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.

Equation (2) also shows something else: if v_i and ε_i are uncorrelated, then the correlation, in absolute value, between the error terms in (1) and (2) will be smaller than 1, except if $\text{var}(v_i) = \sigma^2 = 0$. To see this, calculate the correlation between ε_i and $v_i + \alpha \varepsilon_i$:

$$\text{cor}(\varepsilon_i, v_i + \alpha \varepsilon_i) = \frac{\rho \sigma + \alpha}{\sqrt{(\sigma^2 + \alpha^2 + 2 \rho \alpha \sigma)}} \quad (3)$$

Only if the assumption of zero correlation between v_i and ε_i is dropped, i.e. $\rho \neq 0$, a correlation of +1 or -1 between ε_i and $v_i + \alpha \varepsilon_i$ is possible. This overall 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

⁷ 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 u_i that equals $v_i + \alpha \varepsilon_i$ in the present notation (Winkelmann, 2004, p. 463, line 3).

this, consider the special case that $u_i=0$. In this case, the correlation between the error terms in (1) and (2) is 1 if $\alpha > 0$. 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, i.e. we do not know the number of doctor visits (y_i) for sure. y_i remains random although its expectation is fixed. We can even take it one step further. Suppose $\lambda_i = \exp(\alpha x_i' \gamma + w_i' \theta)$ so that there is no random part in the expectation and consequently no correlation between λ_i and ε_i . Still the random count y_i can be correlated with continuous random variable ε_i .⁸ Ignoring this additional correlation, also causes the correlation between the hurdle and the count process to be lower than 1, in absolute value, and that is true even if u_i and ε_i have full correlation, since the variation in y_i does not only result from the variations in the unobserved heterogeneity term. 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 be made of a bivariate copula. The first application of the copula technique in economics on continuous variables is discussed in Lee (1983). 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 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) are transformed to normal distributed random variables and then evaluated using a bivariate normal distribution.⁹ The transformation of an essentially discrete random variable like a count is not straightforward. The way to do this, is discussed in van Ophem (2000).¹⁰

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

8 Cf. e.g. Cox (1974), for a discussion on the correlation between continuous and discrete random variables.

9 For a full discussion of the technique, see Trivedi and Zimmer (2005).

10 The likelihood function is given in Appendix 2 of the extended version of this paper. It needs to be maximized across γ , β , σ^2 and ρ . To understand that all parameters are identified, note that the likelihood function is built from a combination of bivariate and univariate normal probabilities (cf. eq. 11 from van Ophem (2000)). Even in the most simplified case of one bivariate normal probability with linear arguments: $B(x_i' \gamma, x_i' \beta; \rho)$, this probability is relevant in the bivariate probit model, γ , β and ρ are identified. In the present case the arguments of the bivariate normal probability are nonlinear transformations of $x_i' \gamma$ and $x_i' \beta$, which enhances the identification of the model.

(cf. Trivedi and Zimmer, 2005, p. 96). However, it is questionable whether this is actually a disadvantage 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. The unobserved heterogeneity term u_i is assumed to be normally distributed. Correlation between both random processes is allowed by letting u_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 u_i is normally distributed. Even though I will make exactly the same distributional assumptions as Winkelmann (2004), using the copula technique allows for a more general specification of the correlation. Not only u_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. 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 (Winkelmann, 2004, p. 462, eq. (7)). In my replication I will compare Winkelmann's Probit-Poisson-log-normal-model with partial correlation with the same model but allowing for full correlation using copula techniques.

3. Data

Use will be made from the German Socio-Economic Panel (GSOEP).¹¹ The data are the same as used by Winkelmann (2004) except that I only concentrate on the 1999-wave.¹² 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

¹¹ The data are available in the data archive of the Journal of Applied Econometrics.

¹² I concentrate on the 1999-wave because GSOEP is a panel. The use of multiple individual observations will reduce the estimated standard errors, leading to higher significance. To illustrate the main point of this paper, I prefer to be on the safe side and, as a result, to accept higher standard errors.

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.

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 average number of doctor visits is 2.391 and the corresponding standard deviation is 3.943. I will use exactly the same explanatory variables as Winkelmann (2004). The definition of the variables are self-explanatory¹³.

4. Estimation results

To start with, I will replicate the optimal model presented in Winkelmann (2004) but now with partial correlation. 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. Table 1 presents the estimation results.

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. The significance is somewhat reduced if we compare the results in Table 1 with the corresponding ones in Winkelmann, probably due to the reduction in the number of observations. Still, many explanatory variables are strongly significant, but the correlation is not. In fact the estimate of the correlation is very close to 0. Winkelmann (2004) claims to have found the same result.

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 in 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. An alternative explanation is that those participating in sports are more health conscious or are required to have regular physical check-ups. People with higher incomes

¹³ Details can be found in Winkelmann (2004) or in the extended version of this paper.

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.

Table 1: Estimation results of the Probit-Poisson-log-normal model with correlation.

	Probit	Truncated Poisson-log-normal
age	0.002 (0.002)	-0.001 (0.002)
age ²	-0.046 (0.016)**	0.013 (0.016)
male	0.442 (0.039)**	-0.100 (0.060)
education	-0.014 (0.008)	-0.018 (0.008)*
married	-0.115 (0.044)**	0.036 (0.045)
household size	0.040 (0.015)**	-0.042 (0.016)**
active sport	-0.131 (0.040)**	0.029 (0.043)
good health	0.498 (0.040)**	-0.465 (0.069)**
bad health	-0.567 (0.068)**	0.657 (0.065)**
social assistance	-0.060 (0.101)	0.143 (0.094)
log(income)	-0.109 (0.047)*	0.036 (0.048)
variance unobserved heterogeneity		0.781 (0.016)**
correlation	0.001 (0.280)	

Log-likelihood = -11895.94. Absolute asymptotic standard errors between parentheses. **/* = significant at 1%/5% (two-sided test). The model also contains the variables: a constant, full-time, part-time, unemployed, spring, fall, winter in both the Probit and the Truncated Poisson-log-normal.

Also with respect tot the number of doctor visits, given that an individual goes at least once, it can be concluded that 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. 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.

In Table 2 the estimation results of the copula-based model that takes full account of the correlation is presented. The estimated correlation is about -0.47 and strongly significant with a t-value of about 6. 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 1 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. 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).

Table 2: Estimation results of the Probit-Poisson-log-normal model with correlation using a copula.

	Probit	Truncated Poisson-log-normal
age	0.001 (0.002)	-0.002 (0.002)
age ²	-0.041 (0.015)**	0.024 (0.017)
male	0.424 (0.040)**	-0.209 (0.046)**
education	-0.016 (0.008)*	-0.015 (0.008)
married	-0.108 (0.044)*	0.063 (0.046)
household size	0.039 (0.014)**	-0.051 (0.015)**
active sport	-0.127 (0.041)**	0.062 (0.043)
good health	0.487 (0.040)**	-0.612 (0.049)**
bad health	-0.564 (0.067)**	0.768 (0.055)**
social assistance	-0.048 (0.105)	0.167 (0.100)
log(income)	-0.108 (0.047)*	0.063 (0.050)
variance unobs. heterogeneity (σ^2)		0.866 (0.028)**
correlation	-0.467 (0.078)**	

Log-likelihood = -11890.94. Absolute asymptotic standard errors between parentheses. **/* = significant at 1%/5% (two-sided test). The model also contains the variables: a constant, full-time, part-time, unemployed, spring, fall, winter in both the Probit and the Truncated Poisson-log-normal.

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 3. This table presents estimates of the conditional mean and other sample statistics for the complete sample and the sample with positive doctor visits.

Table 3: 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 $\rho \neq 0$				
Complete sample	3.352	2.166	10.102	1.876
Subsample count > 0	3.619	2.650	10.102	1.889
Copula				
Complete sample	3.375	2.407	11.683	1.907
Subsample count > 0	3.653	2.975	11.683	1.909

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 two methods is much smaller than the maximum observed count. There does exist a notable difference between the estimated 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 -0.47, 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 -0.47 if an alternative estimation method 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

- Cox NR. 1974. Estimation of the correlation between a continuous and a discrete variable. *Biometrics* **30**: 171-178.
- Lee L. 1983. Generalized econometric models with selectivity. *Econometrica* **51**: 507-512.
- Lindeboom M, van den Berg G. 1994. Heterogeneity in models for bivariate survival: the importance of the mixing distribution. *Journal of the Royal Statistical Society, Series B* **56**: 49-60.
- Santos Silva JMC, Windmeijer F. 2001. Two-part multiple spell models for health care and demand. *Journal of Econometrics* **104**: 67-89.
- Stoddart GL, Barer ML. 1981. Analyses of the demand and utilization through episodes of medical care. In *Health, Economics and Health Economics*, van der Gaag J, Perlman M (eds.), North Holland: Amsterdam.
- Trivedi PK, Zimmer DM. 2005. Copula modeling: An introduction for practitioners, *Foundations and Trends in Econometrics* **1**: 1-111.
- van Ophem H. 2000. Modeling selectivity in count data models. *Journal of Business & Economic Statistics* **18**: 503-511.
- Winkelmann R. 2004. Health care reform and the number of doctor visits - an econometric analysis. *Journal of Applied Econometrics* **19**: 455-472.