ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 953–958

Journal of Clinical Epidemiology

Longitudinal tobit regression: A new approach to analyze outcome variables with floor or ceiling effects

Jos Twisk[a,b,*], Frank Rijmen[a]

[a]Department of Clinical Epidemiology and Biostatistics, VU Medical Centre, Amsterdam, The Netherlands

[b]Department of Methodology and Applied Biostatistics, Institute of Health Sciences, Vrije Universiteit, Amsterdam, The Netherlands

Accepted 1 October 2008

Abstract

Background: In many epidemiologic longitudinal studies, the outcome variable has floor or ceiling effects. Although it is not correct, these variables are often treated as normally distributed continuous variables.

Objectives: In this article, the performance of a relatively new statistical technique, longitudinal tobit analysis, is compared with a classical longitudinal data analysis technique (i.e., linear mixed models).

Study Design and Setting: The analyses are performed on an example data set from rehabilitation research in which the outcome variable of interest (the Barthel index measured at on average 16.3 times) has typical floor and ceiling effects. For both the longitudinal tobit analysis and the linear mixed models an analysis with both a random intercept and a random slope were performed.

Results: Based on model fit parameters, plots of the residuals and the mean of the squared residuals, the longitudinal tobit analysis with both a random intercept and a random slope performed best. In the tobit models, the estimation of the development over time revealed a steeper development compared with the linear mixed models.

Conclusion: Although there are some computational difficulties, longitudinal tobit analysis provides a very nice solution for the longitudinal analysis of outcome variables with floor or ceiling effects.

Keywords: Longitudinal studies; tobit analysis; Linear mixed models; Statistical methods; Floor effects; Ceiling effects

1. Introduction

Within epidemiology, there is an increasing interest in performing prospective cohort studies. One of the purposes of these studies is to analyze the longitudinal development over time in a particular outcome variable. In some of these studies, the outcome variable of interest reaches a certain ceiling over time. For instance, in rehabilitation research, most of the patients will recover after a certain amount of time. On the instrument to measure the rehabilitation process, these patients cannot score any higher than the maximum. It is also possible that so-called floor effects occur. For instance, when pain medication is investigated and the outcome variable pain is measured on a visual analog scale, some patients will report ‘‘no pain’’ after a certain amount of time. They cannot score lower than the ‘‘no pain’’ level. Also in studies where there is some detection limit (e.g., for blood parameters or environmental factors, such as pesticides), these floor effects are present. In fact these problems always arise when a measurement instrument that has upper and lower limits is used and when some of the patients in the study reach these upper or lower limits. One can think of an underlying latent variable with an unrestricted range, of which the observed outcome is an on both sides truncated version, so that floor and ceiling effects can be considered to be a kind of interval censoring. Sometimes, floor and ceiling effects are referred to as lower and upper censoring.

In most longitudinal epidemiological studies, these floor and ceiling effects are ignored. The development over time of such outcome variables are analyzed as if they were normally distributed over the whole period of time. This is not the case, because when patients reach the floor or ceiling, the outcome variable is not normally distributed anymore. In cross-sectional studies (especially in econometrics), the problem of upper and lower censoring is solved by using so-called tobit models, after Tobin’s [1] classical example on household expenditures. Within epidemiology, only a few examples are available in which cross-sectional tobit

* Corresponding author. Department of Clinical Epidemiology and Biostatistics, VU Medical Centre, de Boelelaan 1118, 1081 HV Amsterdam, The Netherlands.

E-mail address: jwr.twisk@vumc.nl (J. Twisk).

0895-4356/09/$ – see front matter
doi: 10.1016/j.jclinepi.2008.10.003

What is New?

Longitudinal tobit analysis is suitable for the analysis of longitudinal data with floor and/or ceiling effects and it outperforms the traditional linear mixed models.

Although there could be some computational difficulties, longitudinal tobit analysis can be performed within Stata software.

analysis is used [2–9]. However, for longitudinal epidemiological studies, tobit analysis is (to our knowledge) never used, although it has some nice theoretical advantages above the "classical" longitudinal data analysis.

The purpose of this study is to compare longitudinal tobit analysis with "classical" longitudinal analysis to investigate the longitudinal development over time in outcome variables with floor and/or ceiling effects. The example used in the present article is taken from rehabilitation research.

2. Methods

2.1. Study population

The population used in the present study was taken from a longitudinal rehabilitation study among stroke patients [10]. The main purpose of the study was to analyze the development of the Barthel index. An outcome variable that represents a patient's ability to carry out 10 everyday tasks (i.e., bladder and bowel control, toilet use, dressing, feeding, walking, personal toilet, transfer activities, bathing, and stair climbing) [11]. The lowest score for the Barthel index is 0 and the highest possible score is 20. The Barthel index was assessed weekly during the first 10 weeks after stroke onset, then every 2 weeks until week 20 and finally the Barthel index was assessed at week 26, week 38, and week 52. The study population consisted of 101 patients with on average 16.3 measurements (range, 2–18) per patient. Forty-seven patients had a full data set, whereas 33 patients only missed the first measurement. Furthermore, there were seven patients who dropped-out (varying after the second measurement to the measurement after 26 weeks), there were six patients with intermittent missing data with more than three missing observations, and eight patients with intermittent missing data with only one or two missing observations.

Besides the outcome variable, several covariates were measured at baseline: sit-balance, incontinence, type of stroke, and age. For detailed information see Kwakkel et al. [10].

In the example, two research questions will be addressed. First, the longitudinal development over time will be analyzed, and second, the influence of the covariates measured at baseline will be analyzed. The development over time was modeled as a second order polynomial function.

2.2. Statistical analysis

2.2.1. Classical longitudinal analysis

The classical statistical methods to be used to answer the above research questions are either linear mixed models [12], which are also known as multilevel models, hierarchical models, or random coefficient models or generalized estimating equations (GEE analysis) [13]. In the present example, only linear mixed models will be used because for continuous outcome variables, linear mixed models are, in general, a bit more flexible compared with GEE analysis [14]. Two different analyses will be presented. First, an analysis with only a random intercept and second, an analysis with both a random intercept and a random slope for time.

2.2.2. Tobit longitudinal analysis

The general idea of tobit regression is that it models both the probability of reaching either the floor or ceiling and the development over time between the floor and ceiling. The tobit model originated in the context of linear regression analysis (cross-sectional data), and can be formulated mathematically as follows. Let y^* be a random latent variable that is not censored, and assume a linear regression model for it:

$$y_i^* = x_i'\beta + e_i, \quad e_i \sim N(0, \sigma^2)$$

where i refers to subject i.

Furthermore, it is assumed that we can observe the realizations of y^* for a given range $[l,u]$ only, and that values of y^* smaller than l or larger than u are censored at, respectively, l and u. Hence, the observed limited dependent variable y is obtained from y^* as

$$\begin{aligned} y_i &= l \quad \text{for } y_i^* \leqslant l \\ y_i &= y_i^* \quad \text{for } l < y_i^* < u \\ y_i &= u \quad \text{for } y_i^* \geqslant u \end{aligned}$$

If a dependent variable is limited at one side, only a lower (or upper) limit is needed ($l = -\text{inf}$ or $u = +\text{inf}$).

Because of the censoring mechanism, $E(y)$ is not equal to $E(y^*)$. Because the distribution for y is not the same as the distribution for y^*, the expected values will be different. Therefore, parameter estimates may become inconsistent.

For longitudinal data, a tobit model can be defined in a similar way. As in classical longitudinal analysis, a natural choice for the underlying model for y^* is the linear mixed model [12]:

$$\begin{aligned} & y_{ij}^* | b_i = \mathbf{x}_{ij}'\boldsymbol{\beta} + z_{ij}'\mathbf{b}_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2) \\ & \mathbf{b}_i \sim N(0, \mathbf{D}) \end{aligned}$$

where i refers to case i and j to the jth measurement. That is, conditional on the case-specific parameters $\mathbf{b}_i$, a linear