Available online at www.sciencedirect.com
ScienceDirect
Journal of Health Economics 27 (2008) 300–307
JOURNAL OF HEALTH ECONOMICS
www.elsevier.com/locate/econbase

Sample selection versus two-part models revisited: The case of female smoking and drinking

David Madden *

School of Economics, University College Dublin, Belfield, Dublin 4, Ireland

Received 21 January 2003; received in revised form 26 July 2007; accepted 26 July 2007
Available online 3 January 2008

Abstract

There is a well-established debate between Heckman sample selection and two-part models in health econometrics, particularly when no obvious exclusion restrictions are available. Most of this debate has focussed on the application of these models to health care expenditure. This paper revisits the debate in the context of female smoking and drinking, and evaluates the two approaches on three grounds: theoretical, practical and statistical. The two-part model is generally favoured but it is stressed that this comparison should be carried out on a case-by-case basis.

JEL classification: I12; D12; C24; C25

Keywords: Selection; Two-part; Smoking; Drinking

1. Introduction

There is a well-established debate in health econometrics over the merits of Heckman sample selection models versus two-part models. This debate originally arose in the context of health care expenditure.[1] More recently, the debate has re-surfaced in the context of modelling ageing and health care expenditure with contributions by Zweifel et al. (1999), Salas and Raftery (2001) and Seshamini and Gray (2004).

One important area of health economics where discussion of the relative merits of these approaches is more sparse is in the analysis of smoking and drinking. The importance of the issue for these behaviours arises from the fact that in a population at any given point in time a substantial proportion of people will be observed with zero consumption of tobacco and/or alcohol. As we will discuss in more detail below this may arise for a number of reasons and hence great care must be taken in model selection. This paper presents evidence on the issue in the context of smoking and drinking using data from a sample of Irish women. Our focus in this paper is on the issue of model selection and the criteria which should be used, and a comparison of the two models on the basis of these criteria.

* Tel.: +353 1 7168396; fax: +353 1 2830068.
E-mail address: david.madden@ucd.ie.
[1] See Jones (2000) for a summary and overview of the debate.

0167-6296/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2007.07.001

The remainder of this note is structured as follows: in Section 2 we discuss the modelling issues involved, including the crucial matter of what criteria should be considered in terms of choosing between the different approaches. In Section 3 we discuss our data and present results while Section 4 presents concluding comments.

2. The econometric modelling of tobacco and alcohol consumption

In this section we briefly discuss modelling strategies for goods such as tobacco and alcohol. When modelling the consumption of, say, tobacco, a crucial factor which must be taken into account is the high percentage of zeros which can arise in micro-data sets with highly disaggregated information. Such zero observations may occur for three main reasons: firstly, in survey data with short recording periods infrequency of purchase may generate a large percentage of observations with zero consumption (for example in the case of semi-durable goods such as clothing). Second, tobacco may not be a good for some individuals because they are non-smokers. Thirdly, even though a person may be a potential smoker they may not be able to afford the good at current prices and income. Thus, the corner solution of zero consumption is the utility-maximising decision for these individuals, given current prices and income. The particular interpretation given to zero observations can have a crucial bearing on the estimation approach adopted.

This note takes as its starting point the double-hurdle approach to modelling tobacco consumption (see Jones, 1989). This approach assumes that individuals must pass two hurdles before being observed with a positive level of consumption. Both hurdles are the outcome of individual choices: a participation decision and a consumption decision. The precise form of the double-hurdle approach adopted will depend upon crucial assumptions in two areas: the degree of independence between the error terms in the participation and consumption equations and secondly the issue of dominance, i.e. whether the participation decision dominates the consumption decision.

There are three constituents to the double-hurdle approach: observed consumption, the participation equation and the consumption equation. Suppose observed consumption is given by $y = dy^{**}$, and we have a participation equation, $w = \alpha' Z + v$, $d = 1$ if $w > 0$, $=0$ otherwise, and a consumption equation, $y^{**} = \max[0, y^*]$, $y^* = \beta' X + u$. If we allow for the possibility of dependence between the disturbance terms, then if the sample is divided into those with zero consumption (denoted 0) and those with positive consumption (denoted +) the likelihood for the full double-hurdle model is

$$L0 = \begin{cases} \prod_0 [1 - p(d = 1)p(y^* > 0|d = 1)] \prod_+ p(d = 1)p(y^* > 0|d = 1)g(y^*|y^* > 0, d = 1) \\ = \prod_0 [1 - p(v > -\alpha' Z)p(u > -\beta' y|v > -\alpha' Z)] \prod_+ p(v > -\alpha' Z) \\ \quad p(u > -\beta' X|v > -\alpha' Z)g(y|u > -\beta' X, v > -\alpha' Z) \end{cases}$$

where Z and X are the regressors influencing participation, α and β are vectors of estimated coeffficients and u and v are additive disturbance terms which are randomly distributed with a bivariate normal distribution.

If we assume that the disturbance terms u and v are independent then the model reduces to the Cragg model (Cragg, 1971) with likelihood

$$L1 = \prod_0 [1 - p(v > -\alpha' Z)p(u > -\beta' X)] \prod_+ p(v > -\alpha' Z)p(u > -\beta' X)g(y|u > -\beta' X)$$

An alternative simplifying assumption to independence is what is known as first-hurdle dominance, i.e. that the participation decision dominates the consumption decision. This implies that zero consumption does not arise from a standard corner solution but instead represents a separate discrete choice. Thus, once the first hurdle has been passed, then standard Tobit type censoring (whereby zero, or even negative consumption, could be a utility-maximising choice by someone who has "passed" the participation hurdle) is not relevant. First-hurdle dominance implies that $p(y^* > 0|d = 1) = 1$ and $g(y^*|y^* > 0, d = 1) = g(y^*|d = 1)$.

In this case with dependence between the disturbance terms the likelihood is

$$L2 = \prod_0 [1 - p(v > -\alpha' Z)] \prod_+ p(v > -\alpha' Z)g(y|v > -\alpha' Z)$$

which corresponds to Heckman's sample selection model (henceforth referred to as the selection model). If independence is also assumed the double-hurdle approach reduces to a probit for participation and ordinary least squares for