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

the consumption equation estimated over those for whom positive consumption is observed with likelihood function:

$$L3 = \prod_{0}[1 - p(v > -\alpha' Z)]\prod_{+} p(v > -\alpha' Z)g(y).$$

Thus the two crucial factors in terms of modelling strategy are (a) independence of the error terms and (b) the interpretation placed upon the observed zeros which determines whether or not dominance is assumed. For reasons that we explain below we believe that dominance applies to our data and so the crucial choice we face is between a selection model (likelihood $L2$) and a two-part model, likelihood $L3$. How do we choose between these models?

Following the discussion by Dow and Norton (2003), we can think of three criteria which might influence our choice between the two approaches: these are theoretical (what exactly is it we are trying to model), practical (are there valid exclusion restrictions, without which the sample selection model may under-perform) and finally statistical (are there statistical tests which might help discriminate between the models).

Turning first to the theoretical issue of what it is we are trying to model, the choice between a sample selection and a two-part model revolves around whether we wish to model potential or actual outcomes. The sample selection model was first introduced by Heckman (1976, 1979) and its main application was in the context of wage equation estimation (for a general discussion of the sample selection model see Puhani, 2000). In such applications we are often interested in the effect of a variable such as schooling on the wage. Yet we do not observe the wage for people who do not work who in all probability will be those people only able to achieve a relatively low wage, given their schooling. Thus, we may be interested in modelling the potential wage an individual could earn, were they to work. We can then estimate the effect of a covariate such as schooling on both actual and potential workers.

When dealing with smoking, what is the meaning of potential spending on tobacco? For those people with observed zero consumption of tobacco, is there a latent positive expected consumption which might have been incurred under certain circumstances? As we explain below, the nature of the questions regarding tobacco and alcohol consumption in our data leads us to believe that dominance applies and that there is unlikely to be a latent positive expected consumption. Thus, on the first of our three criteria, it seems likely that what we are trying to model is actual smoking, as opposed to potential smoking. It follows that we are interested in the effects of covariates on actual as opposed to potential smoking in which case the two-part model seems more appropriate.

The second issue in terms of choice between sample selection and two-part models concerns the issue of exclusion restrictions. In most cases the vectors Z and X will have many variables in common. In the case of the selection model, in order to separately identify the decision regarding participation (to smoke or not to smoke) from the level decision (how much to smoke) it is necessary that we have variables which enter Z but do not enter X. If such variables (known as exclusion restrictions) cannot be found then separate identification depends upon the non-linearity of the extra term (known as the inverse Mills ratio, IMR) which appears in the level equation. The problem here is that the IMR is frequently an approximately linear function over a wide range of its argument and so estimates from the level equation in the sample selection model may be non-robust owing to collinearity.

This issue was investigated by Leung and Yu (1996) who maintained that (in the absence of plausible exclusion restrictions) collinearity between the regressors in X and the IMR is the decisive criterion in terms of choosing between the selection and two-part models. They also pointed out that the presence of such collinearity problems limits the power of the t-test for sample selectivity on the coefficient of the IMR (a test which is sometimes used as a criterion for model selection). This highlights the need for procedures to analyse the degree of collinearity.

Belsley (1991) provides a comprehensive list of diagnostics for analysing collinearity in general. In the case under study here, where we are concerned precisely with the collinearity between one particular regressor (the IMR) and the others, a less time-consuming procedure may be possible. As explained in Belsley et al., a sufficient condition for the presence of collinearity for any particular regessor is a high value of its variance inflation factor (VIF). The VIF for any regressor X_i is given by $\text{VIF}_i = 1/1 - R_i^2$ where R_i^2 is the multiple correlation coefficient of X_i regressed on the remaining explanatory variates. What precisely defines a "high" value is open to question but Belsely et al. suggest that a VIF in excess of 30 is a cause for concern.[2]

[2] A high VIF for the IMR is a sufficient but not a necessary condition for collinearity. If a high VIF is not found, Belsley et al. recommend a more comnprehensive sequence of tests involving the calculation of the scaled condition indices from the matrix of regressors X, followed by a decomposition of the estimated variance of each regression coefficient into a sum of terms, each of which is associated with a condition index. It is

Finally, there may be statistical criteria which might enable us to discriminate between the two models. The Monte Carlo study of Leung and Yu (1996) used the criterion of the mean square error (M.S.E.) of the parameter of interest. The M.S.E. is the variance plus the square of the bias, but crucially, knowledge of the true parameter is needed to compute the bias. And thus, this M.S.E. criterion cannot be used in empirical applications where the true parameter values are unknown. In this situation Dow and Norton (2003) recommend the test proposed by Toro-Vizcarrondo and Wallace (1968) which they label an empirical M.S.E. test. This involves calculating the empirical M.S.E. of both estimators under the assumption that one model, e.g. the selection model, is consistent and correct. The M.S.E. for the selection model will then involve only the variance component while that for the two-part model will involve its variance and its "bias" relative to the selection model (by assumption the selection model has zero "bias"). We also calculate the empirical M.S.E. under the assumption that the two-part model is the "true" model. In the next section we describe our data and present results for the collinearity test and the empirical M.S.E.

3. Data

The data set used in this paper is known as the Saffron Survey which was carried out in 1998 by the Centre for Health Economics at University College Dublin. The Saffron Survey's aim was to survey women's knowledge, understanding and awareness of their lifetime health needs. For our purposes in this paper the relevant questions regarding smoking and drinking were as follows: "Do you currently smoke?". For those who answer yes to this question there is a follow-up question: "Approximately how many cigarettes do you smoke per day?". For alcohol consumption the relevant questions are: "In general how often would you say that you take a drink?" and respondents are given a range of seven different replies ranging from "every day" to "never". Those who answer that they take a drink are then asked how much they usually drink.

Note that the questions are phrased in terms of what typical consumption patterns are, as opposed to what recorded consumption is. While there is a danger that this might give rise to under-reporting (particularly since the goods in question are tobacco and alcohol) it nevertheless suggests that recorded zero consumption of tobacco or alcohol represents a discrete choice, and does not arise from either infrequency of purchase or as a corner solution. Thus, we believe it is reasonable to assume that first-hurdle dominance applies. [3] Thus, if we assume dependence between the disturbance terms in the participation and consumption equations we estimate the selection model. If we do not assume such dependence then a two-part model should be estimated. Since the focus of this paper is to examine the relative performance of the two approaches we will present results for both models.

In total the sample consisted of 1260 women. However, of that 1260 relevant information was missing for some women, leaving us with a sample of 1257 women in the case of smoking and 1259 in the case of drinking. The data provides detailed information on individual characteristics involving health, lifestyle choices and demographics. However, in the case of a number of these variables there are clear issues of potential endogeneity. Hence, even though information is provided on self-assessed health, exercise and weight, these variables are not included in the analysis.[4] For similar reasons we also choose not to include drinking status as an explanatory variable in the smoking equation and vice versa. It is likely that some unobserved third variable such as time preference might be influencing both drinking and smoking simultaneously. Ideally we would like to implement an instrumental variables strategy to deal with this, but in the absence of an appropriate instrument it seems best to adopt the relatively parsimonious approach used here.[5] It is also worth pointing out we lack price variation (since we have a single cross-section of time) and hence it not possible to examine cross-price effects to investigate issues of complementarity and substitutability. However

then possible to determine the extent to which each near linear dependency (and high condition index) contributes to each variance. The detection of a high VIF for the IMR eliminates the need for such a sequence of tests. I am grateful to an anonymous referee for this suggestion.

[3] In the case of alcohol we should bear in mind that someone who classifies themselves as an abstainer may have had heavy alcohol consumption in the past and we might wish to regard them as different from someone who has never consumed alcohol. Unlike the case with tobacco however, we do not have sufficient information to distinguish between these two categories of non-drinkers.

[4] An earlier version of this paper included such variables and the qualitative results concerning modelling were unchanged. These results are available on request.

[5] In a general discussion of dealing with endogeneity in the absence of appropriate instruments Bound et al. (1995) suggest that in some cases the cure may be worse than the disease, i.e. the misspecification arising from an inappropriate instrument may cause greater problems than endogeneity bias.

Table 1
Summary statistics for total sample, smokers and drinkers (standard deviations in italics)

Variable	Mean (total sample)	Mean (smokers only)	Mean (drinkers only)
Age	47.36616 (*17.67702*)	41.91413 (*15.89779*)	42.97564 (*15.91299*)
Single	.234127 (*.4236201*)	.3047091 (*.460923*)	.2632743 (*.4406538*)
Married	.5904762 (*.4919412*)	.565097 (*.4964323*)	.6128319 (*.4873723*)
Widowed	.1269841 (*.3330874*)	.0692521 (*.2542347*)	.0685841 (*.2528854*)
Divorced/separated	.0484127 (*.2147219*)	.0609418 (*.2395556*)	.0553097 (*.2287104*)
No formal education	.0293651 (*.1688947*)	.0415512 (*.1998383*)	.0221239 (*.1471679*)
Primary education	.2825397 (*.4504132*)	.2825485 (*.4508635*)	.1980088 (*.3987195*)
Junior/Inter Cert	.2492063 (*.4327253*)	.3545706 (*.479047*)	.2577434 (*.4376341*)
Leaving Cert	.3126984 (*.4637767*)	.232687 (*.4231308*)	.3595133 (*.4801234*)
Third level	.1555556 (*.3625774*)	.1301939 (*.3369837*)	.1847345 (*.3882969*)
Working	.3785714 (*.4852236*)	.3739612 (*.4845251*)	.4469027 (*.4974479*)
Smoker	.2865079 (*.4523091*)		.3340708 (*.4719257*)
Drinker	.7174603 (*.4504132*)	.8365651 (*.3702752*)	
Cigarettes per day (if smoker)	15.19777 (*9.063858*)		14.82333 (*8.772799*)
Units alcohol per month (if drinker)	10.94685 (*13.78244*)	12.78891 (*14.5134*)	
Number of children	2.630952 (*2.409505*)	2.515235 (*2.425654*)	2.378319 (*2.15373*)
Medical friend	.415873 (*.4930675*)	.3573407 (*.4798815*)	.4457965 (*.4973284*)

Table 2
Max likelihood estimates of Heckman selection and two-part model for tobacco (N = 1257, 898 censored, standard errors in italics)

Variable	Heckman		Two-part	
	Selection	Level	Selection	Level
Age	−0.022 (*0.022*)	0.675 (*0.261*)[a]	−0.013 (*0.017*)	0.669 (*0.201*)[a]
Age Squared	−0.000 (*0.000*)	−0.007 (*0.002*)[a]	−0.000 (*0.000*)	−0.007 (*0.002*)[a]
Married	−0.058 (*0.154*)	−1.190 (*1.364*)	−0.243 (*0.130*)[b]	−1.206 (*1.385*)
Widowed	−0.226 (*0.212*)	−1.828 (*2.681*)	−0.328 (*0.195*)[b]	−1.895 (*2.696*)
Divorced/Separated	0.066 (*0.232*)	−0.528 (*2.563*)	−0.131 (*0.212*)	−0.507 (*2.409*)
Junior/Inter Cert	−0.071 (*0.144*)	−1.254 (*1.475*)	−0.036 (*0.114*)	−1.280 (*1.338*)
Leaving Cert	−0.658 (*0.144*)[a]	−2.232 (*1.602*)	−0.643 (*0.128*)[a]	−2.432 (*1.560*)
Third Level	−0.581 (*0.170*)[a]	−3.454 (*1.705*)[c]	−0.668 (*0.159*)[a]	−3.630 (*1.771*)[c]
Working	−0.107 (*0.130*)	−0.593 (*1.122*)	−0.172 (*0.093*)[b]	−0.623 (*1.081*)
Medical Friend	−0.007 (*0.101*)	−0.706 (*0.913*)	0.015 (*0.086*)	−0.708 (*1.021*)
Number of Children	0.035 (*0.023*)	0.748 (*0.344*)[c]	0.035 (*0.022*)	0.759 (*0.258*)[a]
ρ	−0.051 (*0.157*)			
σ	2.116 (*0.056*)[a]			
λ	−0.419 (*1.301*)			

[a] Significant at 99%.
[b] Significant at 90%.
[c] Significant at 95%.

this issue is not of primary concern to us in this note since our principal aim is to discuss ways of comparing and choosing between the two models.

Table 1 summarises the relevant variables for the total sample of 1260 women and for smokers and drinkers also. [6] Tables 2 and 3 provide estimates of both selection and two-part models for tobacco and alcohol. Since our primary focus is on a comparison between the two models rather than the actual estimated coefficients, we will confine our discussion of the results to such a comparison. Dealing with the selection equation for tobacco first, the estimated coefficients for the two models are quite similar and the sign of the coefficients are in line with intuition. Significance

[6] The default category for education is a combination of the categories "no formal education" and "Primary Cert" indicating that formal schooling ended at approximately the age of 12. "Junior Cert" indicates formal schooling ceased at approximately 16, while "Leaving Cert" indicates schooling ended at approximately 18. A more detailed discussion of variable definition, etc. is provided in the data Appendix A.

Table 3
Max likelihood estimates of Heckman selection and two-part model for alcohol (N = 1259, 380 censored, standard errors in italics)

Variable	Heckman		Two-part	
	Selection	Level	Selection	Level
Age	−0.054 (*0.021*)[a]	−0.323 (*0.203*)	−0.047 (*0.017*)[a]	−0.335 (*0.175*)[b]
Age Squared	0.000 (*0.000*)	0.002 (*0.002*)	0.000 (*0.000*)	0.002 (*0.002*)
Married	0.238 (*0.153*)	−3.250 (*1.304*)[c]	0.297 (*0.141*)[c]	−3.178 (*1.286*)[c]
Widowed	−0.007 (*0.192*)	−2.658 (*1.974*)	0.014 (*0.178*)	−2.707 (*2.392*)
Divorced/separated	0.547 (*0.237*)[c]	−3.126 (*1.757*)[b]	0.651 (*0.227*)[a]	−2.950 (*2.234*)
Junior/Inter Cert	0.238 (*0.120*)[c]	1.855 (*1.313*)	0.242 (*0.113*)[c]	1.959 (*1.353*)
Leaving Cert	0.190 (*0.137*)	1.413 (*1.115*)	0.387 (*0.119*)[a]	1.495 (*1.412*)
Third level	0.154 (*0.171*)	2.184 (*1.505*)	0.360 (*0.159*)[c]	2.257 (*1.574*)
Working	0.145 (*0.123*)	1.446 (*1.056*)	0.129 (*0.101*)	1.492 (*0.929*)
Medical friend	0.268 (*0.098*)[a]	−1.304 (*0.980*)	0.148 (*0.091*)	−1.217 (*0.889*)
Number of children	−0.001 (*0.020*)	0.103 (*0.216*)	0.000 (*0.019*)	0.104 (*0.270*)
ρ	−0.066 (*0.022*)[a]			
σ	2.469 (*0.084*)[a]			
λ	−0.777 (*0.278*)[a]			

[a] Significant at 99%.
[b] Significant at 90%.
[c] Significant at 95%.

levels appear to be higher in the two-part model however. Perhaps the only substantive difference between the two models lies in the role of marital status (being married and being widowed). For the selection model, it has no effect whereas it exerts a negative and significant for the two-part model. For the level equation results for the two models are practically identical.

Regarding alcohol, once again there is broad agreement between the two models. In the selection equation, the two-part model shows greater effects for education, while the selection model shows a greater effect for the presence of a medical friend or relative. Once again, results for the level equation are very similar.

Thus, in summary, the values and sizes of the estimated coefficients in Tables 2 and 3 are generally quite plausible and, perhaps more interestingly, there is comparatively little difference between the selection and two-part models. We now turn to the issue of collinearity.

Our first check is on the value of the VIF for the IMR for the selection model. A regression of the IMR on the other covariates reveal R^2 of 0.9975 and 0.9925 for tobacco and alcohol, respectively. This indicates values of the VIF for the IMR of 400 and 133, well in excess of the recommended threshold and a clear indication of collinearity between the IMR and the other regressors.[7] Thus, our collinearity analysis clearly indicates problems with the selection model and suggests that estimates from such a model should be treated with caution. We now turn to the empirical M.S.E.

Tables 4 and 5 show the results of the empirical M.S.E. test for smoking and drinking, respectively. We show the results under two different null hypotheses: first on the basis that the true model is the selection model and secondly, that the true model is the two-part model.

Dealing with tobacco first of all, the overall evidence is in favour of the two-part model. When the selection model is assumed to be the true model, the M.S.E. for the two-part model is still lower for half the covariates. When the two-part model is assumed to be the true model then the M.S.E. for the two-part model is lower for the majority of the covariates. In the case of alcohol however, the empirical M.S.E. suggests that the selection model is to be favoured as its M.S.E. is lower for the majority of covariates, even when the two-part model is assumed to be the true model.

Overall, the results from this test are to some degree consistent with those from the collinearity tests, at least in terms of the ranking of the models. The extremely high VIF for the IMR in the case of tobacco raised questions regarding the reliability of the selection model and this is consistent with the results from the M.S.E. which favoured the two-part

[7] Details of the regressions underlying these R^2 are available on request. The problem of collinearity regarding the IMR was confirmed by the more comprehensive battery of tests suggested by Belsley et al.

Table 4
Empirical M.S.E., tobacco

	H_0: Heckman "True" Model			H_0: 2PM "True" Model		
Variable	M.S.E. (Heckman)	M.S.E. (2PM)	Choice	M.S.E. (Heckman)	M.S.E. (2PM)	Choice
Age	0.068	0.041	2PM	0.068	0.041	2PM
Age^2	0	0	–	0	0	–
Married	1.862	1.919	H	1.862	1.919	H
Widowed	7.189	7.273	H	7.194	7.269	H
Divorced/separated	6.569	5.803	2PM	6.569	5.803	2PM
Junior/Inter Cert	2.176	1.79	2PM	2.177	1.79	2PM
Leaving Cert	2.566	2.475	2PM	2.606	2.435	2PM
Third level	2.906	3.166	H	2.937	3.135	H
Working	1.26	1.169	H	1.261	1.168	2PM
Medical friend	0.834	1.043	H	0.834	1.043	H
Number of children	0.118	0.067	2PM	0.118	0.066	2PM

Table 5
Empirical M.S.E., alcohol

	H_0: Heckman "True" Model			H_0: 2PM "True" Model		
Variable	M.S.E. (Heckman)	M.S.E. (2PM)	Choice	M.S.E. (Heckman)	M.S.E. (2PM)	Choice
Age	0.041	0.031	2PM	0.041	0.031	2PM
Age^2	0	0	–	0	0	–
Married	1.701	1.659	2PM	1.706	1.654	2PM
Widowed	3.895	5.723	H	3.8975	5.7209	H
Divorced/separated	3.086	5.019	H	3.1165	4.9886	H
Junior/Inter Cert	1.724	1.84	H	1.735	1.8297	H
Leaving Cert	1.244	1.999	H	1.2511	1.9923	H
Third level	2.265	2.484	H	2.27	2.4785	H
Working	1.114	0.865	2PM	1.1162	0.8628	2PM
Medical friend	0.961	0.799	2PM	0.9684	0.791	2PM
Number of children	0.047	0.073	H	0.0468	0.0731	H

model. In the case of alcohol the VIF indicated that collinearity was not as great a problem as in the case of tobacco, though still a cause for concern. The M.S.E. however came down in favour of the selection model.

4. Conclusions

This paper has revisited the debate between selection and two-part models in the context of smoking and drinking, applications where a large proportion of zero observations are typically found. The comparison was carried out on three grounds: theoretical (which approach was most appropriate for what we were trying to model), practical (the existence of valid exclusion restrictions and the problems which may arise with collinearity if no plausible exclusion restrictions can be found) and statistical (via implementation of the empirical M.S.E. test). Our conclusion is that on the basis of the collinearity tests the two-part model is to be preferred to the selection model and this preference is stronger in the case of tobacco compared to alcohol. The empirical M.S.E. tests favoured the selection model for the case of alcohol. However it is also worth noting that on the more pragmatic grounds of policy conclusions which could be drawn from estimated coefficients, there was relatively little to choose between the two approaches. It is not clear that this would always be the case, so from a practitioners point of view the moral of this exercise would seem to be that when choosing between the two models, ideally the sequence of tests outlined in this paper should be applied. This would particularly be the case where no plausible exclusion restrictions can be found for the selection model.

Acknowledgements

I would like to thank two anonymous referees, Joe Durkan, Brendan Walsh and participants at a Dublin Labour Studies Group for helpful comments. I would also like to thank Joe Durkan for providing the data. The usual disclaimer applies.

Appendix A. Data definition

Variable	Categories
Marital status	Single, married, widowed, divorced/separated. Excluded category was "single"
Education	No formal education, primary education, Junior/Inter Cert, Leaving Cert, 3rd level. No formal education and primary cert were combined and used as excluded category
Labour market status	At work as employee, self employed/employee, assisting relative, unemployed, retired, student, home duties, other. This was converted into 0/1 variable with the first three categories defined as "working" and the others as "non-working".
Smoking status	Smoker was constructed as 0/1 variable on basis of answer to question "Do You Currently Smoke?". Number of cigarettes constructed from question "Approximately how many cigarettes do you smoke per day?".
Drinking status	Drinker was constructed as 0/1 variable from question "In general how often would you say that you take a drink?". Categories were "(1) every day", "(2) 2–3 days per week", "(3) once a week", "(4) 2–3 times a month", "(5) about once a month", "(6) less than once a month", "(7) never" with 0 for those answering (7). "Units of alcohol" was constructed from question "how much would you usually drink per day/week/month?".
Medical friend	0/1 variable constructed from question "Are your spouse/partner, other relatives or friends (a) a doctor/consultant (b) a nurse or paramedic?". (1) was assigned to anyone who answered "yes" to any of these questions.

References

Belsley, D., 1991. Conditioning Diagnostics: Collinearity and Weak Data in Regression. John Wiley and Sons, New York.

Bound, J., Jager, D., Baker, R., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association 90, 443–450.

Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.

Dow, W., Norton, E., 2003. Choosing Between and interpreting the Heckit and Two-part models for corner solutions. Health Services and Outcomes Research Methodology 4 (No. 1), 5–18.

Heckman, J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic Social Measurement 5, 475–492.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47, 53–161.

Jones, A.M., 1989. A double-hurdle model of cigarette consumption. Journal of Applied Econometrics 4, 23–39.

Jones, A.M., 2000. "Health Econometrics" in Handbook of Health Economics, Vol. 1A, Culyer, A., Newhouse J. (Eds.), North Holland.

Leung, S.F., Yu, S., 1996. On the choice between sample selection and two-part models. Journal of Econometrics 72, 197–229.

Puhani, P., 2000. The Heckman correction for sample selection and its critique. Journal of Economic Surveys 14, 53–67.

Salas, C., Raftery, J., 2001. Econometric Issues in testing the age neutrality of health care expenditure. Health Economics Letters 5, 12–15.

Seshamini, M., Gray, A., 2004. Ageing and health care expenditure: the Red Herring argument revisited. Health Economics 13, 303–314.

Toro-Vizcarrondo, C., Wallace, T., 1968. A test of the mean square error criterion for restrictions in linear regression. Journal of the American Statistical Association, Vol., 558–572.

Zweifel, P., Felder, S., Meiers, M., 1999. Ageing of the population and health care expenditure: a Red Herring? Health Economics 8, 485–496.