Madden, David

Working Paper

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

UCD Centre for Economic Research Working Paper Series, No. WP06/04

Provided in Cooperation with:
UCD School of Economics, University College Dublin (UCD)

Suggested Citation: Madden, David (2006) : Sample selection versus two-part models revisited: The case of female smoking and drinking, UCD Centre for Economic Research Working Paper Series, No. WP06/04, University College Dublin, UCD School of Economics, Dublin, https://hdl.handle.net/10197/771

This Version is available at:
https://hdl.handle.net/10419/72347

UCD CENTRE FOR ECONOMIC RESEARCH

WORKING PAPER SERIES

2006

Sample Selection Versus Two-Part Models Revisited: The Case of Female Smoking and Drinking

David Madden, University College Dublin

WP06/04

April 2006

UCD SCHOOL OF ECONOMICS
UNIVERSITY COLLEGE DUBLIN
BELFIELD DUBLIN 4

Sample Selection Versus Two-Part Models Revisited: The Case of Female Smoking and Drinking

David Madden

(University College Dublin)

March 2006

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.

Keywords: Selection, Two-part, Smoking, Drinking.

JEL Codes: I12, D12, C24, C25.

Corresponding Author: David Madden
Economics Department,
University College Dublin,
Belfield,
Dublin 4,
Ireland.
Phone: 353-1-7168396
Fax: 353-1-2830068
E-mail: david.madden@ucd.ie

Acknowledgement: 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.

Sample Selection Versus Two-Part Models Revisited: The Case of Female Smoking and 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. Among the more prominent contributions to this debate were Duan et al (1983, 1984, 1985), Hay and Olsen (1984), Maddala (1985) and Leung and Yu (1996). Jones (2000) provides a summary and overview of the debate. 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 area of importance to health economics where discussion of the relative merits of these 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 hence the discussion of the values of the estimated coefficients, *per se*, is somewhat brief.

The remainder of the paper 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, in section 4 we present our results and section 5 presents concluding comments.

2. The Econometric Modelling of Tobacco and Alcohol Consumption

In this section we discuss modelling strategies for goods such as tobacco and alcohol. Since the relevant methodological issues are practically identical for tobacco and alcohol we will confine the discussion to tobacco alone.

When modelling the consumption of tobacco, one important factor which must be taken into account is the high percentage of zeros which can arise in microeconometric 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 paper takes as its starting point the double-hurdle approach to modelling tobacco consumption (see Jones, 1989). In general 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. As we will see below, 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 consumptions 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. Borrowing from Jones (1989) they can be represented as follows:

$$\text{Observed consumption: } y = d.y^{**}$$

$$\text{Participation equation: } w = \alpha'Z + v,\ d = 1 \text{ if } w > 0, =0 \text{ otherwise}$$

$$\text{Consumption equation: } y^{**} = \max[0, y^{*}],\ y^{*} = \beta'X + u.$$

Here Z and X are the regressors influencing participation and consumption and u and v are additive disturbance terms which are randomly distributed with a bivariate normal distribution. 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

$$L0 = \prod_{0}[1 - p(d=1)p(y^{*} > 0 \mid d=1]\prod_{+} p(d=1)p(y^{*} > 0 \mid d=1)g(y^{*} \mid y^{*} > 0, d=1)$$

$$= \prod_{0}[1 - p(v > -\alpha'Z)p(u > -\beta'y \mid v > -\alpha'Z]$$

$$\prod_{+} p(v > -\alpha'Z)p(u > -\beta'X \mid v > -\alpha'Z)g(y \mid u > -\beta'X, v > -\alpha'Z)$$

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

$$L1 = \prod_{0}[1 - p(v > -\alpha'\mathrm{Z})\,p(u > -\beta'X)]\prod_{+} p(v > -\alpha'\mathrm{Z})\,p(u > -\beta'X)\,g(y \mid u > -\beta'X)$$

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 $p(y^* > 0 \mid d = 1) = 1$ and $g(y^* \mid y^* > 0, d = 1) = g(y^* \mid d = 1)$.

In this case if we allow for the possibility of dependence between the disturbance terms the likelihood is

$$L2 = \prod_{0}[1 - p(v > -\alpha'\mathrm{Z})]\prod_{+} p(v > -\alpha'\mathrm{Z})\,g(y \mid v > -\alpha'\mathrm{Z})$$

This 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

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

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 marginal effects of covariates on actual as opposed to potential smoking. Thus we are concerned with $E[y \mid Z, X]$ rather than $E[y^{**} \mid Z, X]$, 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 sample 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 i.e. we identify a variable (or variables) which affect the decision of whether or not to smoke, but do not affect the decision of how much to smoke. 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) which appears in the level equation. The problem here is that the inverse Mills ratio 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 issues.

This issue has been investigated in some detail by Leung and Yu (1996). They maintain that the collinearity between the regressors in x and the inverse Mills ratio is the decisive criterion in terms of choosing between the sample selection and two-part models. They also point out that the presence of such collinearity problems limits the power of the t-test for sample selectivity on the coefficient of the inverse Mills ratio (a test which is sometimes used as a criterion for model selection).[1] They recommend testing for collinearity by calculating the condition number for the regressors in the level equation. If this exceeds

[1] The test that the coefficient is zero can be used to test the null that the two-part is correct against the alternative that the selection model is correct. However, as Duan et al (1984) show, the converse is not necessarily true.

20, then the two-part model is more robust. Otherwise the selection model may be used. However the choice of 20 as a "critical" value appears to be somewhat arbitrary and in addition there are other, related, diagnostics which can provide a more thorough investigation of collinearity (see Belsley, 1991, and Besley et al., 1980). In section 4 below we provide a thorough analysis of our data for collinearity.

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 (MSE) of the parameter of interest. The MSE 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 MSE 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-Vizcarronndo and Wallace (1968) which they label an *empirical* MSE test. This involves calculating the empirical MSE of both estimators under the assumption that one model e.g. the selection model, is consistent and correct. The MSE 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 MSE under the assumption that the two-part model is the "true" model. In the next section we describe our data.

3. **Data**

We now describe our data source and explain the reasons behind our chosen estimation strategies. 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. Much of the focus of the survey was on the issue of hormone replacement therapy[2] but other information regarding health, lifestyle choices and demographics was also collected. 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. In the case of alcohol however, 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. Nevertheless, we still believe it is reasonable to assume that first-hurdle dominance applies. 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

[2] See Thompson, 2000.

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 sample was also reweighted to take account of a number of features including the fact that originally women over the age of 45 were oversampled to ensure that there would be an adequate sample of women currently taking Hormone Replacement Therapy.

The Saffron survey 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.[3]

Table 1 summarises the relevant variables for the total sample of 1260 women and for smokers and drinkers also. Amongst the features worth noting are that smokers and to a lesser extent drinkers tend to be younger (hence a lower proportion of widowed). Smokers tend to have below average educational attainments while drinkers have above average attainments. [4] The higher proportions of drinkers (amongst those who smoke) and smokers (amongst those who drink) also suggest that smoking and drinking may be complementary activities. The last two rows of the table lend some support to this idea as they indicate that smokers tend to drink more than non-smokers. However, it also appears

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

[4] 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.

that drinkers smoke less than non-drinkers, though the proportional difference is quite small.[5]

We now present the results and give a detailed discussion of our procedure to choose between the different models.

4. Estimation of Selection and Two-Part Models

In this section we present results for the estimation of selection and two-part models of smoking and drinking evaluating the two models using the criteria outlined in section 2. We first briefly discuss the estimated coefficients and then turn to the methodological issues involved in choosing between the two models.

Estimated Coefficients

Tables 2 and 3 provide estimates of both selection and two-part models for tobacco and alcohol. 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. Higher levels of education are associated with a lower likelihood of being a smoker, while compared to non-drinkers, drinkers are more likely to smoke. What is perhaps slightly surprising is that the extent of drinking has no effect on the likelihood of being a smoker. The only substantive difference between the two models lies in the role of marital status. For the selection model, it has no effect whereas it exerts a negative and significant for the two-part model.

Turning now to the level equation for tobacco, results for the two models are practically identical. Of the education variables only third level education affects the level of

[5] A more detailed discussion of variable definition etc is provided in the data appendix.

smoking. The role of drinking on the level of smoking is somewhat curious. Compared to the default of not being a drinker, being a frequent or moderate drinker has no effect on the level of smoking, but being a light drinker has a negative and significant effect, suggesting a non-monotonic effect of drinking on the level of smoking.

Regarding alcohol, once again there is broad agreement between the two models. For the selection equation, education tends to increase the likelihood of being a drinker, though there is a suggestion of a non-monotonic relationship in the selection model, with third level education having no effect compared to the positive effect of higher second level education. Being married and or divorced/separated increases the likelihood of being a drinker but in the case of marriage this is only significant for the two-part model while the presence of a medical friend or relative increases the likelihood of being a drinker but is only significant for the selection model. Regarding the level of drinking the size of the coefficients are very similar for both models though significance levels vary slightly. Perhaps, curiously while being married divorced or separated increases the likelihood of being a drinker, it tends to reduce the amount of drinking.

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. Thus the policy conclusions to be drawn from the two approaches would be very similar. We now turn to more formal analysis of the issues raised in section 2.

Collinearity

As outlined above, when comparing the performance of selection and two-part models, a crucial issue is the degree of collinearity in the data. This issue has been

investigated in depth by Belsley (1991) and Belsley et al (1980). The former outlines a sequence of tests to detect collinearity and, perhaps more importantly, whether such collinearity is likely to influence estimated regression coefficients. Belsley lists a series of steps which should be followed and which involve investigation of the *X* matrix of regressors (recall that collinearity is especially important in the case where exclusion restrictions do not apply and the *Z* and *X* matrices of regressors influencing participation and consumption respectively are identical). Belsley first recommends obtaining the scaled condition indexes from this matrix. For the $n \times p$ matrix *X* the condition indexes are $\eta_k \equiv \frac{\mu_{\max}}{\mu_k}$, *k=1,…,p*. The μ_k are obtained from the singular-value-decomposition of the matrix *X*. Any $n \times p$ matrix, *X*, may be decomposed as $X = UDV'$ where $U'U = V'V = I_p$ and *D* is diagonal with non-negative diagonal elements $\mu_1,...,\mu_p$, known as the singular values of *X*. A high value for any η_k is an indication of a near linear dependency (the precise definition of "high" is open to debate but Belsley suggests that any condition index in excess of 30 merits attention) and the largest value of η_k is known as the condition number of *X*. The number of high condition indexes will indicate how many near linear dependencies exist in the data. The next step is to determine which variates are involved in them.

Belsley suggests employing a decomposition of the estimated variance of each regression coefficient into a sum of terms, each of which is associated with a condition index. Thus it is possible to determine the extent to which each near linear dependency (and high condition index) contributes to each variance. Given the least squares estimator $b = (X'X)^{-1}X'y$ the variance-covariance matrix is $V(b) = \sigma^2(X'X)^{-1}$ where σ^2 is the

variance of the components of the error term in the linear model. Using the single-value-decomposition outlined above the variance-covariance matrix may be written as $V(b) = \sigma^2 (X'X)^{-1} = \sigma^2 VD^{-2}V'$. Then the variance of the *k*th regression coefficient, b_k, the *k*th diagonal element of $V(b)$ is $\text{var}(b_k) = \sigma^2 \sum_j \frac{v_{kj}^2}{\mu_j^2}$ where the μ_j's are the singular values of X and $V \equiv (v_{ij})$.

Thus the variance of b_k is decomposed into a sum of components each of them associated with one of the *p* singular values, μ_j, of the matrix X. Since these μ_j^2 appear in the denominator, those components associated with near linear dependencies i.e. with small μ_j will be large relative to other components. Thus if we observe that an unusually high proportion of the variance of two or more coefficients are concentrated in components associated with the same small singular value, we may conclude that the variates corresponding to those coefficients are involved in the near dependency corresponding to that small singular value.

Thus the (*k, j*)th variance-decomposition proportion is defined as the proportion of the variance of the *k*th regression coefficient associated with the *j*th component of the decomposition above. These proportions can be calculated as $\phi_{kj} \equiv \frac{v_{kj}^2}{\mu_j^2}$ with $\phi_k = \sum_{j=1}^{p} \phi_{kj}$ and the variance-decomposition proportions are $\pi_{jk} \equiv \frac{\phi_{kj}}{\phi_k}$.

Tables 4 to 7 calculates the scaled condition indices and variance decomposition proportions for the selection and two-part models for tobacco and alcohol respectively. Dealing with the selection model for tobacco first we observe two condition indices which

should be of concern, those with values of 67.73 and 206.42. We then check for variate involvement with the rough rule of thumb that we only include those whose values of π_{jk} exceed a threshold value of about 0.5. Dealing with the condition index of 67.73, we note that only age squared has a π_{jk} in excess of 0.5.

Turning now to the scaled condition index of 206.42, using our threshold of 0.5, the variates age, married, widowed, Junior/Inter Cert, Leaving Cert, 3rd level, working, frequent drinker, moderate drinker, light drinker, number of children and the inverse Mills ratio exceed the threshold level, though in the case of "married" the excess is marginal. The variate with the greatest π_{jk} is the inverse Mills ratio, reflecting the fact that when no valid exclusion restrictions are available, we frequently observe a high degree of collinearity between this variate and others. A value for π_{jk} of 1.0 is a clear warning signal that the selection model may not be appropriate in this case.

Turning now to the value of the condition indices for the two-part model, here we have one condition index which gives cause for concern with a value of 63.62. In this case age and age squared are involved. However, by construction age squared is a (non-linear) function of age, hence there should be no concern regarding this dependency. Thus it seems fair to say that on the basis of the diagnostics using scaled condition indices that for the case of tobacco, the two-part model appears to be the more reliable.

Turning now to alcohol, the pattern of near linear dependencies between the selection and two-part models is similar to that for tobacco although the degree of dependency, particularly for the selection model does not appear to be as severe. This similarity is not entirely unexpected as there is a high degree of overlap in the sets of covariates, although bear in mind that the sample for the level regression in the two-part

model will not be the same as that for tobacco, as those with positive alcohol consumption will not necessarily be the same those with positive tobacco consumption. In the case of the selection model, there are two condition indices clearly in excess of 30, one with a value of 39.64 and the other with a value of 101.53. For the condition index of 39.64 only the constant has a value of π_{jk} in excess of 0.5. For the condition index of 101.53 we observe six values of π_{jk} in excess of 0.5, those for age, age squared, married, divorced/separated, presence of a medical friend and the inverse Mills ratio. The collinearity involving the inverse Mills ratio once again suggests that the selection model may not be appropriate in this case, although the degree of collinearity is considerably less than in the case of tobacco. For the two-part model the pattern is almost identical to that for tobacco.

In summary, comparing the two models for both tobacco and alcohol, detailed analysis of collinearity in the data reveals that this is more of a problem for the selection model than for the two-part model. Since the inverse Mills ratio is implicated for both tobacco and alcohol (especially for tobacco) this raises serious question marks over the suitability of the selection model in the case of tobacco and suggests that the two-part model may be a better bet. For the case of alcohol, the situation is not quite as clearcut, though the evidence still seems to favour the two-part model.

The Empirical MSE Test

Tables 8 and 9 show the results of the Empirical MSE 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 evidence appears to be clearly in favour of the two-part model. For all covariates, with the exception of "Medical Friend", the MSE for the two-part model is smaller than for the selection model, even when the selection model is assumed to be the "true" model. The situation for the case of alcohol is far less clearcut, with results for the covariates split pretty much 50-50 between the two models, though perhaps marginally in favour of the selection model. In general the value of the coefficients for the two models are very close, hence the MSE is determined to a very large degree by the variance, rather than the bias, component.

Overall, the results from this section are in agreement with those concerning the collinearity. The estimates from the selection model for tobacco are unreliable and it would appear that the two-part model is to be preferred here. For the case of alcohol, the performance of the two models is much closer and there appears to be relatively little to choose between them. It is worth noting that were we to assess the two models solely on the basis of the significance of the coefficient for the inverse Mills ratio, then the inclination for the case of tobacco, at least, would be to prefer the selection model.

5. 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

restrictions can be found) and statistical (via implantation of the empirical MSE test). We also took the pragmatic approach of simply looking at the extent to which policy conclusions would differ depending upon the model chosen.

Our conclusion is that on the first three grounds the two-part model is to be preferred to the selection model and this preference is stronger in the case of tobacco rather than alcohol. However, on the more pragmatic grounds of policy conclusions to be drawn, 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 battery of tests outlined in this paper should be applied. This would be particularly the case where no plausible exclusion restrictions can be found for the selection model.

References

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

Belsley, D., E. Kuh and R. Welsch (1980): Regression Diagnostics, Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons.

Cragg, J., (1971): "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods", Econometrica, Vol. 39, pp. 829-844.

Decker, S., and A. Schwartz (2000): "Cigarettes and Alcohol: Substitutes or Complements?". NBER Working Paper 7535.

Dow, W., and E. Norton (2003): "Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions", Health Services and Outcomes Research Methodology, Vol. 4, No. 1, pp. 5-18.

Duan, N., W. Manning, C. Morris and J. Newhouse (1983): "A Comparison of Alternative Models for the Demand for Medical Care" Journal of Business Economics and Statistics, Vol. 1, pp. 115-126.

----------, ----------, ----------and----------(1984): "Choosing Between the Sample Selection Model and the Multi-Part Model", Journal of Business Economics and Statistics, Vol. 2, pp. 283-289.

----------, ----------, ----------and----------(1985): "Comments on Selectivity Bias" in Advances in Health Economics and Health Services Research Vol. 6 (R. M. Scheffler and L.F. Rossiter, eds).

Hay, J., and R. Olsen (1984): "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care", Journal of Business Economics and Statistics, Vol. 2, pp. 279-282.

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, Vol. 5, pp. 475-492.

----------(1979): "Sample Selection Bias as a Specification Error", Econometrica, Vol. 47, pp. 53-161.

Jones, A.M. (1989): "A Double-Hurdle Model of Cigarette Consumption", Journal of Applied Econometrics, Vol. 4, pp23-39.

----------(2000): "Health Econometrics" in Handbook of Health Economics Vol. 1A (ed. A.. Culyer and J. Newhouse), North-Holland.

Leung, S.F., and S. Yu (1996): "On the Choice between Sample Selection and Two-Part Models", Journal of Econometrics, Vol. 72, pp. 197-229.

Maddala, G., (1985): "A Survey of the Literature on Selectivity Bias as it Pertains to Health Care Markets", in Advances in Health Economics and Health Services Research Vol. 6 (R. M. Scheffler and L.F. Rossiter, eds).

Puhani, P., (2000): "The Heckman Correction for Sample Selection and its Critique", Journal of Economic Surveys, Vol. 14, pp. 53-67.

Salas, C., and J. Raftery (2001): "Econometric Issues in Testing the Age Neutrality of Health Care Expenditure", Health Economics Letters, Vol. 5, pp. 12-15.

Seshamini, M., and A. Gray (2004): "Ageing and Health Care Expenditure : the Red Herring Argument Revisited", Health Economics, Vol. 13, pp. 303-314.

Thompson, J., (2000): Economic Aspects of Women's Health with regard to Hormone Replacement Therapy in Ireland, unpublished MA thesis, Economics Department, University College Dublin.

Toro-Vizcarrondo, C., and T. Wallace (1968): A Test of the Mean Square Error Criterion for Restrictions in Linear Regression, Journal of the American Statistical Association, Vol. , pp. 558-572.

Zweifel, P., S. Felder and M. Meiers (1999): "Ageing of the Population and Health Care Expenditure: A Red Herring?", Health Economics, Vol. 8, pp. 485-496.

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*	
No. 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)

Variable	Heckman		Two-Part	
	Selection	**Level**	**Selection**	**Level**
Age	-0.019	0.684	-0.009	0.680
	(0.022)	*(0.258)***	*(0.017)*	*(0.201)***
Age²	-0.000	-0.007	-0.000	-0.007
	(0.000)	*(0.002)***	*(0.000)*	*(0.002)***
Married	-0.102	-1.094	-0.279	-1.118
	(0.160)	*(1.416)*	*(0.132)***	*(1.392)*
Widowed	-0.226	-2.466	-0.321	-2.522
	(0.220)	*(2.712)*	*(0.200)*	*(2.711)*
Divorced/Separated	-0.002	-0.723	-0.185	-0.722
	(0.241)	*(2.588)*	*(0.216)*	*(2.408)*
Junior/Inter Cert	-0.110	-1.221	-0.068	-1.252
	(0.142)	*(1.437)*	*(0.115)*	*(1.344)*
Leaving Cert	-0.698	-2.252	-0.687	-2.428
	*(0.147)***	*(1.860)*	*(0.131)***	*(1.569)*
Third Level	-0.634	-3.271	-0.730	-3.431
	*(0.167)***	*(1.858)**	*(0.161)***	*(1.770)**
Working	-0.124	-0.648	-0.187	-0.677
	(0.129)	*(1.121)*	*(0.094)***	*(1.102)*
Medical Friend	-0.039	-0.634	-0.006	-0.642
	(0.101)	*(0.923)*	*(0.086)*	*(1.023)*
Frequent Drinker	0.538	-2.627	0.498	-2.487
	*(0.140)***	*(1.776)*	*(0.129)***	*(1.719)*
Moderate Drinker	0.440	-2.399	0.386	-2.283
	*(0.135)***	*(1.807)*	*(0.115)***	*(1.525)*
Light Drinker	0.603	-3.642	0.474	-3.485
	*(0.131)***	*(1.945)**	*(0.124)***	*(1.621)***
Number of Children	0.034	0.758	0.035	0.767
	(0.023)	*(0.332)***	*(0.022)*	*(0.257)***
ϱ	-0.043			
	(0.239)			
σ	2.109			
	*(0.056)***			
λ	-0.353			
	(1.965)			

***=significant at 99%, **=significant at 95%, *=significant at 90%

Table 3: Max Likelihood Estimates of Heckman Selection and Two-Part Model for Alcohol (N=1259, 380 Censored)

Variable	Heckman		Two-Part	
	Selection	**Level**	**Selection**	**Level**
Age	-0.058	-0.378	-0.050	-0.390
	*(0.020)***	*(0.199)**	*(0.017)***	*(0.174)***
Age^2	0.000	0.003	0.000	0.003
	*(0.000)**	*(0.002)*	*(0.000)*	*(0.002)*
Married	0.239	-3.167	0.311	-3.100
	(0.153)	*(1.282)***	*(0.142)***	*(1.276)***
Widowed	-0.007	-2.265	0.023	-2.314
	(0.193)	*(1.993)*	*(0.179)*	*(2.376)*
Divorced/Separated	0.525	-3.079	0.643	-2.921
	*(0.238)***	*(1.729)**	*(0.229)****	*(2.216)*
Junior/Inter Cert	0.222	1.561	0.227	1.656
	*(0.121)**	*(1.293)*	*(0.114)***	*(1.345)*
Leaving Cert	0.226	1.834	0.428	1.927
	*(0.137)**	*(1.136)*	*(0.120)****	*(1.405)*
Third Level	0.192	2.726	0.405	2.811
	(0.172)	*(1.520)**	*(0.161)***	*(1.569)**
Working	0.144	1.350	0.132	1.392
	(0.124)	*(1.038)*	*(0.101)*	*(0.922)*
Ciggs per day	0.014	0.158	0.014	0.163
	*(0.005)****	*(0.048)****	*(0.004)****	*(0.043)****
Medical Friend	0.261	-1.290	0.140	-1.211
	*(0.098)****	*(0.982)*	*(0.091)*	*(0.883)*
Number of Children	-0.006	0.016	-0.005	0.015
	(0.021)	*(0.218)*	*(0.020)*	*(0.269)*
ϱ	-0.063			
	*(0.028)***			
σ	2.460			
	*(0.085)****			
λ	-0.735**			
	(0.342)			

***=significant at 99%, **=significant at 95%, *=significant at 90%

Table 4: Scaled Condition Indices and Variance-Decomposition Proportions for Heckman Selection Model for Tobacco

Cond. Ind.	Constant	Age	Age^2	Married	Widowed	Div/Sep	Jun/Int	Leav C
1.00								
2.11								
2.49								
2.54								
2.67								
2.79								
2.92								
3.50								
3.76								
4.87								
5.18								
6.62								
7.04								
12.87								
67.73			0.83					
206.42	0.90	0.68		0.53	0.69		0.70	0.98

Cond. Ind.	3rd lev	Work	Med Friend	Freq. Drink	Mod. Drink	Light. Drink	No.Kids	λ
1.00								
2.11								
2.49								
2.54								
2.67								
2.79								
2.92								
3.50								
3.76			0.52					
4.87								
5.18								
6.62								
7.04								
12.87								
67.73								
206.42	0.97	0.74		0.97	0.97	0.98	0.85	1.00

Table 5: Scaled Condition Indices and Variance-Decomposition Proportions for Two-Part Model for Tobacco

Cond. Ind.	Constant	Age	Age^2	Married	Widowed	Div/Sep	Jun/Int	Leav C
1.00								
1.97								
2.33								
2.39								
2.5								
2.61								
2.72								
3.27								
3.53								
4.57								
4.96								
6.44								
6.61					0.51			
13.07								
63.62	0.83	1.00	0.96					

Cond. Ind.	3rd lev	Work	Med Friend	Freq. Drink	Mod. Drink	Light. Drink	No.Kids
1.00							
1.97							
2.33							
2.39							
2.5							
2.61							
2.72							
3.27							
3.53			0.69				
4.57		0.57					
4.96							0.57
6.44							
6.61							
13.07							
63.62							

Table 6: Scaled Condition Indices and Variance-Decomposition Proportions for Heckman Selection Model for Alcohol

Cond. Ind.	Constant	Age	Age^2	Married	Widowed	Div/Sep	Jun/Int
1.00							
2.19							
2.44							
2.63							
2.70							
3.24							
3.71							
4.02							
5.27							
5.87							
7.34							
11.39							
39.64	0.83						
101.53		1.00	0.62	0.61		0.63	

Cond. Ind.	Leav C	3rd lev	Work	Ciggs per day	Med Friend	No. Kids	λ
1.00							
2.19							
2.44							
2.63							
2.70							
3.24							
3.71							
4.02							
5.27							
5.87							
7.34							
11.39							
39.64							
101.53					0.55		0.70

Table 7: Scaled Condition Indices and Variance-Decomposition Proportions for Two-Part Model for Alcohol

Cond. Ind.	Constant	Age	Age^2	Married	Widowed	Div/Sep	Jun/Int
1.00							
2.12							
2.35							
2.48							
2.54							
3.08							
3.52							
3.91							
5.04							
5.81							
7.04				0.66	0.50		
11.58							
61.20	0.83	1.00	0.96				

Cond. Ind.	Leav C	3rd lev	Work	Ciggs per day	Med Friend	No. Kids
1.00						
2.12						
2.35						
2.48						
2.54						
3.08						
3.52						
3.91					0.84	
5.04						
5.81						
7.04						
11.58						
61.20						

Table 8: Empirical MSE, Tobacco

	H_0: Heckman "True" Model			H_0: 2PM "True" Model		
Variable	**MSE (Heckman)**	**MSE (2PM)**	**Choice**	**MSE (Heckman)**	**MSE (2PM)**	**Choice**
Age	0.067	0.04	**2PM**	0.067	0.04	**2PM**
Age^2	0	0	**-**	0	0	**-**
Married	2.005	1.938	**2PM**	2.006	1.938	**2PM**
Widowed	7.355	7.353	**2PM**	7.358	7.35	**2PM**
Divorced/Separated	6.698	5.798	**2PM**	6.698	5.798	**2PM**
Junior/Inter Cert	2.065	1.807	**2PM**	2.066	1.806	**2PM**
Leaving Cert	3.46	2.493	**2PM**	3.491	2.462	**2PM**
Third Level	3.452	3.159	**2PM**	3.478	3.133	**2PM**
Working	1.257	1.215	**2PM**	1.257	1.214	**2PM**
Medical Friend	0.852	1.047	**H**	0.852	1.047	**H**
Frequent Drinker	3.154	2.975	**2PM**	3.174	2.955	**2PM**
Moderate Drinker	3.265	2.339	**2PM**	3.279	2.326	**2PM**
Light Drinker	3.783	2.652	**2PM**	3.808	2.628	**2PM**
Number of Children	0.11	0.066	**2PM**	0.11	0.066	**2PM**

Table 9: Empirical MSE, Alcohol

	H_0: Heckman "True" Model			H_0: 2PM "True" Model		
Variable	**MSE (Heckman)**	**MSE (2PM)**	**Choice**	**MSE (Heckman)**	**MSE (2PM)**	**Choice**
Age	0.04	0.03	**2PM**	0.04	0.03	**2PM**
Age^2	0	0	**-**	0	0	**-**
Married	1.644	1.633	**2PM**	1.648	1.628	**2PM**
Widowed	3.972	5.648	**H**	3.9745	5.6454	**H**
Divorced/Separated	2.989	4.936	**H**	3.0144	4.9107	**H**
Junior/Inter Cert	1.672	1.818	**H**	1.6809	1.809	**H**
Leaving Cert	1.29	1.983	**H**	1.2991	1.974	**H**
Third Level	2.31	2.469	**H**	2.3176	2.4618	**H**
Working	1.077	0.852	**2PM**	1.0792	0.8501	**2PM**
Medical Friend	0.002	0.002	**-**	0.0023	0.0018	**2PM**
No. Ciggs	0.964	0.786	**2PM**	0.9706	0.7797	**2PM**
Number of Children	0.048	0.072	**H**	0.0475	0.0724	**H**

Appendix – 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). The variable "Frequent drinker" corresponds to those answering (1) or (2), "moderate drinker" to those answering (3) or (4) and "light drinker" to those answering (5) or (6). "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.