Journal of Public Economics 23 (1984) 59–80. North-Holland

STATISTICAL MODELS FOR ZERO EXPENDITURES IN HOUSEHOLD BUDGETS

Angus DEATON and Margaret IRISH*

University of Bristol, Bristol, U.K.

Received September 1982, revised version received February 1983

1. Introduction

Household budget surveys are an important source of data on consumer's expenditure, not just for individual consumers, but also for estimating national aggregates. However, it is a widespread finding that for certain commodities, most notably tobacco and alcohol, the estimate from the household survey falls short of the known consumption total calculated (with some confidence) from data on production, imports, exports and excise duties. For example, in the British Family Expenditure Survey (with which we shall largely be concerned) total tobacco expenditure was underestimated in 1976 by 21 percent [see Kemsley, Redpath and Holmes (1980, p. 51)]. Much of this understatement, together with that on alcohol, is thought to occur because of the design of the survey which excludes many persons amongst whom consumption of such items is thought to be atypically high (e.g. prisoners, hoteliers and their residents, merchant seamen). Even so, the possibility remains that some of the understatement is due to various types of misreporting by households included in the survey. In this paper we consider a model in which the standard tobit specification [Tobin (1958)] is supplemented by the operation of a simple binary censor. The tobit model is essentially a linear regression model in which non-positive observations of the dependent variable are replaced by zero. We take this specification as our starting point but add a second censoring process that randomly replaces a fraction of the observations generated by the tobit model by zeroes. The combined model can serve as a representation of several types of

*We are grateful to the SSRC for financial support under the grant HR7637 'The Economics and Econometrics of Consumer Behaviour'. We received helpful comments from seminar participants at the University of Bristol, at the Netherlands Central Bureau of Statistics, and at the joint SSRC/NBER conference in Oxford in June 1982. We should also like to give particular thanks to Andrew Chesher, Malcolm Pemberton, Peter Schmidt, Christopher Sims, and three anonymous referees.

misreporting. Firstly, the additional zeroes can result from false reporting by either the respondent or the enumerator, the latter being an important consideration in considering surveys from developing countries. Secondly, the additional zeroes may arise because purchases are made infrequently so that over the limited period of the survey, no purchase is recorded for some households, while others record purchases greater than the rate of consumption over the survey period. A variant of this model is one where there are 'beginning of period effects' [see Kemsley et al. (1980, pp. 36, 51)]. Consumers who, prior to the survey, have made a recent purchase of an infrequently purchased item, worry that the expenditure will escape enumeration and in order to 'help' the survey, falsely record the purchase as having taken place during the diary period. Such explanations would not of course generate the known aggregate understatement of expenditures. However, in the simplified models considered here, all the models discussed give rise to the same statistical formulation. In all cases, the essential feature of the model is that an observed zero expenditure can occur either because the household genuinely does not purchase the good, or because, for one reason or another, a zero is incorrectly reported. Which is in fact the case is not known in advance so that the contamination has to be dealt with statistically.

Section 2 states the model formally and links it to other similar models in the econometric literature. A specification test is developed which can be applied to an estimated tobit model to test for the presence of the additional binary censor. Section 3 shows how a first attempt at estimating the model can be constructed from a combination of ordinary least squares and the method of moments in a manner similar to that suggested for tobit in Greene (1981). Section 4 presents results from the Family Expenditure Survey data for the fiscal year 1973/74.[1] The results for tobacco expenditures show no evidence of the operation of the binary censor and the tests suggest no modification to the usual tobit model. There is thus nothing in the model considered here which would cast doubt on the design-based explanation for the underestimation of tobacco expenditures. However, results for alcohol and for durable goods (where frequency of purchase is likely to be a consideration), are quite different. The test statistics indicate a strong rejection of the simple tobit, but in the opposite direction to that predicted by the existence of a binary censor. Instead of there being a *larger* proportion of zero purchases than is predicted from the rest of the distribution of purchases, we typically observe *too few* zeroes. This result, which extends to other commodities and to expenditures from quite a different survey which we have examined (the 1969–70 Socio-economic Survey of Sri Lanka), not only rejects all of the models of misreporting

[1]We are very grateful to Professors Atkinson, King and Stern for permission to use the FES data as processed by them.

considered here, but also implies that the standard tobit model cannot account for an apparently quite general feature of such data. One possible source of the difficulty is the normality assumption embodied in the tobit model. Section 4 investigates this possibility by re-estimating the tobit models without normality using the non--parametric estimation procedure of Buckley and James (1979). Perhaps surprisingly, the results are not seriously altered; parameter estimates are very similar and there are still two few zeroes.

We finish without a convincing explanation of our data. Although our negative results are hopefully of interest in themselves — we find it surprising that the evidence for so many commodities should consistently indicate *under-* rather than *over*-censoring — the solution of the puzzle is left for further work and for other investigators.

2. The model and its relationship to tobit

We work entirely with a single equation model of an individual expenditure; at a later stage it would be desirable to construct a complete system of equations along the same lines, but this is not the topic of this paper. [See Kay, Keen and Morris (1982) for a discussion of some of the issues.] Denote by y_i the observation (i.e. for household i) on the expenditure concerned and $\boldsymbol{x}_i$ a vector of conditioning variables. In the absence of any form of censoring, we assume

$$y_i | \boldsymbol{x}_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2) \tag{1}$$

for parameters $\boldsymbol{\beta}$ and σ^2. The tobit specification can then be written for observations z_i, $i=1,\ldots,n$:

$$z_i \begin{cases} = y_i, & \text{if } y_i > 0, \\ = 0 & \text{otherwise,} \end{cases} \tag{2}$$

or

$$z_i = \max\{y_i, 0\}. \tag{3}$$

In this formulation zeroes arise if and only if the household genuinely does not purchase the good. Consider now a new binary variable w_i which takes on values 1 with probability p_i and 0 with probability $(1-p_i)$. We shall assume that the process generating w_i is independent of y_i conditional on $\boldsymbol{x}_i$. The operation of w_i is not directly observed; instead, (2) and (3) are modified to:

$$z_i \begin{cases} = y_i, & \text{if } y_i > 0 \quad \textit{and} \quad w_i = 1, \\ = 0, & \text{otherwise,} \end{cases} \tag{4}$$

or

$$z_i = w_i \max\{y_i, 0\}. \tag{5}$$

The variable z_i is the recorded expenditure in the survey; it is zero *either* if the consumer genuinely never purchases the good *or* if, for one reason or another, its purchase is not recorded. By contrast, a positive expenditure is always genuine and, in addition, tells us that $w_i = 1$.

The same model already exists in the literature in a number of different contexts. If the binary variable w_i can be modelled by a probit formulation, the model can be written in the general form suggested by Heckman (1980), i.e.

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \\ t_i &= \boldsymbol{s}_i'\boldsymbol{\gamma} + v_i \end{aligned} \begin{pmatrix} u \\ v \end{pmatrix} \sim N(\boldsymbol{0}; \Sigma), \tag{6}$$

with $z_i = y_i$ if *both* $y_i > 0$ and $t_i > 0$ and $z_i = 0$ otherwise, and Σ a diagonal matrix. The model is also essentially identical to the 'double-hurdle' model proposed by Cragg (1971).

The likelihood function is straightforwardly derived. A positive z_i occurs if and only if *both* y_i and w_i are positive and, because of independence, the contribution to the likelihood is $\sigma^{-1} p_i \phi\{(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma\}$ for the unit normal density function ϕ. A zero z_i occurs *either* if $y_i \leqq 0$, probability $1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$, *or* if $y_i > 0$ and $w_i = 0$, probability $(1 - p_i)\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$. Hence, if there are n_1 positive z_i's out of a total sample of n, the log-likelihood is:

$$\ln L = -\frac{n_1}{2} \ln \sigma^2 + \sum_{+} \ln p_i + \sum_{+} \ln \phi\{(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma\}$$

$$+ \sum_{0} \ln\{1 - p_i \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)\}. \tag{7}$$

A + beneath a $\sum$ denotes summation over the n_1 observations for which $z_i > 0$, a similar 0 denotes summation over the n_2 $(= n - n_1)$ observations for which $z_i = 0$.

Note that this likelihood function simplifies in the two special cases when the source of zero censoring is known. First, if all censoring is known to be caused by the binary censor and none by tobit censoring, e.g. because we know in advance that all households purchase the good at some time (e.g. food), then it is known that $\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$ is effectively unity for all i, so that (7) becomes the sum of two log-likelihood functions:

$$\ln L = -\frac{n_1}{2} \ln \sigma^2 + \sum_{+} \ln \phi\{(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma\} + \sum_{+} \ln p_i + \sum_{0} \ln(1 - p_i). \tag{8}$$

The first two terms are together the standard regression likelihood function for the n_1 positive observations; the last two comprise the likelihood function for the binary process determining whether z_i is zero or positive. Hence, OLS based on the positive observations alone is maximum likelihood, is consistent, and will be fully efficient provided β and σ do not appear as determinants of p_i. Second, if the binary censor is known *not* to be operating, $p_i = 1$ for all i and (7) becomes the tobit log-likelihood function.

In this paper we are mostly concerned with the examination of the tobit model for evidence of misspecification of the kind which would be expected to arise if the binary censor were operating. As a first step, therefore, we examine the case where $p_i = p$ for all i, so that the case where $p = 1$, which is tobit, can be easily tested for. The log-likelihood (7) is then:

$$\ln L = -\frac{n_1}{2}\ln\sigma^2 + n_1 \ln p + \sum_{+} \ln \phi\{(z_i - x_i'\beta)/\sigma\}$$

$$+ \sum_{0} \ln\{1 - p\Phi(x_i'\beta/\sigma)\}. \tag{9}$$

Before looking at a test statistic, consider now a simple frequency of purchase model in which consumers correctly report purchases, but in which purchases themselves are made at intervals which may be longer then the period of the survey. To simplify, assume that *either* the survey period covers a whole number of purchase cycles, in which case purchases equal consumption, the tobit model is correct and there is no additional censoring, *or* that the period of the survey is a fraction p of the purchase period, with $1/p$ an integer. In this latter case, a purchase of y/p is observed with probability p during the survey (the household buying once a month observed for a week will buy four weeks' consumption with probability of one-quarter), while with probability $(1-p)$ no purchases are observed. Formally, if y_i is consumption, and z_i purchases, eq. (4) still holds but (5) becomes:

$$z_i = w_i \max\{y_i/p, 0\}. \tag{10}$$

The new log-likelihood function is easily shown to be:

$$\ln L^* = -\frac{n_1}{2}\ln\sigma^2 + 2n_1 \ln p + \sum_{+} \ln \phi\{(pz_i - x_i'\beta)/\sigma\}$$

$$+ \sum_{0} \ln\{1 - p\Phi(x_i'\beta/\sigma)\}. \tag{11}$$

Define $\zeta = \beta/p$ and $\rho = \sigma/p$, and substitute in (11) to give:

$$\ln L^* = -\frac{n_1}{2}\ln \rho^2 + n_1 \ln p + \sum_{+} \ln \phi\{(z_i - x_i'\zeta)/\rho\}$$

$$+ \sum_{0} \ln\{1 - p\Phi(x_i'\zeta/\rho)\}, \tag{12}$$

which, apart from interpretation, is identical to (9). Hence, on the micro data, the misreporting and purchase frequency models are not distinguishable, at least with constant p. (With non-constant p_i, identification would require some separation of the x variables from those influencing p_i.) Of course, the frequency model does not lead to the under-reporting in aggregate which would be a consequence of straight misreporting.

The foregoing model can also be modified to deal with 'beginning of period' effects. As before, for households that buy the commodity, *actual* purchases are either y_i/p or 0, depending on whether or not stocking-up takes place during the survey. The proportion of households stocking-up is p, but assume that a larger proportion, $\pi > p$, report the purchase y_i/p; $(\pi - p) > 0$ is the proportion of households subject to the beginning of period effect. Eq. (10) still holds for z_i but we now have $\text{prob}\{w_i = 1\} = \pi$ rather than p. The log-likelihood becomes:

$$\ln L^{**} = -\frac{n_1}{2}\ln \sigma^2 + n_1(\ln p + \ln \pi) + \sum_{+} \ln \phi\{(pz_i - x_i'\beta)/\sigma\}$$

$$+ \sum_{0} \ln\{1 - \pi\Phi(x_i'\beta/\sigma)\}. \tag{13}$$

Once again, substitution of $\zeta = \beta/p$ and $\rho = \sigma/p$ shows that, apart from the substitution of π for p, the log-likelihood (13) is identical to (12), and to (9). Hence, at the level of simplification used here, with p (or π) a constant independent of i, the false reporting model and the frequency of purchase model, with or without beginning of period effects, cannot be distinguished on the data. In principle, one way of separating the explanations would be to use the aggregate data to impose the constraint that the predicted average consumption level in the sample should equal the known population mean. However, for the reason discussed in the Introduction, the consumption habits in the sample may not be representative of those in the population so that such a constraint might well be invalid. Of course, there is no problem in separating any of the 'reporting' models from tobit.

The model with likelihood function (9), (12), or (13), we call p-tobit and we

wish to test it against tobit. The score for p, i.e. $\partial \ln L/\partial p$ is given by:

$$d=\sum_{i=1}^{n}\frac{h(z_i)-p\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)}{p\{1-p\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)\}}, \tag{14}$$

where

$$\begin{aligned} h(z_i)&=1, \quad \text{if } z_i>0,\\ h(z_i)&=0, \quad \text{if } z_i=0. \end{aligned} \tag{15}$$

Hence, if $\tilde{\Phi}_i\equiv\Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}/\tilde{\sigma})$, where $\tilde{\beta}$ and $\tilde{\sigma}$ are the MLEs of β and σ when tobit is true, a score test of $p=1$ may then be based on:

$$\tilde{d}=\sum_{i=1}^{n}\left\{\frac{h(z_i)-\tilde{\Phi}_i}{1-\tilde{\Phi}_i}\right\}. \tag{16}$$

Under the null (tobit), $\mathrm{E}\{h(z_i)\}=\Phi_i(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$ which is consistently estimated by $\tilde{\Phi}_i$ so that $\tilde{d}$, being a score, has an expectation of zero under the null. If, however, the binary censor is operating and $p<1$, $\mathrm{E}\{h(z_i)\}=p\Phi_i(\boldsymbol{z}_i'\boldsymbol{\beta}/\sigma)$, which is less than $\Phi_i(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$ so that we should expect $\tilde{d}<0$ if the alternative is true. [Note that the comparison of $h(z_i)$ with Φ_i is also the basis of Nelson's (1981) test of the tobit, though Nelson's test is not the same and is a Hausman (1978) rather than a score test.]

To test the significance of departures of $\tilde{d}$ from zero requires an estimate of variance. This can be straightforwardly, if tediously, obtained from the information matrix of the likelihood (9) evaluated under the null, i.e. at $p=1$. It is simplest to reparameterize the model by defining $\boldsymbol{\xi}=\boldsymbol{\beta}/\sigma$ and $\tau=1/\sigma$. The lower triangle of the (symmetric) information matrix, I say, is obtained by taking the conditional expectation of the Hessian of (9) with respect to $(p,\boldsymbol{\xi}',\tau)'$ to yield:

$$\begin{bmatrix} \sum\Phi_i/(1-\Phi_i) & \cdots & \cdots \\ \sum\phi_i\boldsymbol{x}_i/(1-\Phi_i) & \sum\boldsymbol{x}_i\boldsymbol{x}_i'(\Phi_i+\phi_i'+\phi_i^2/(1-\Phi_i)) & \\ 0 & -\sum\boldsymbol{x}_i'(\Phi_i\boldsymbol{x}_i'\boldsymbol{\xi}+\phi_i)/\tau & \tau^{-2}\sum\{2\Phi_i+(\boldsymbol{x}_i'\boldsymbol{\xi})^2\Phi_i+(\boldsymbol{\chi}_i'\boldsymbol{\xi})\phi_i\} \end{bmatrix}, \tag{17}$$

where $\Phi_i=\Phi(\boldsymbol{x}_i'\boldsymbol{\xi})$, $\phi_i=\phi(\boldsymbol{x}_i'\boldsymbol{\xi})$, $\phi_i'=\phi_i'(\boldsymbol{x}_i'\boldsymbol{\xi})$ and all $\sum$ are over *all* observations. The matrix I in (17) is the variance–covariance matrix of the $(k+2)$ vector of scores, $\boldsymbol{s}$ say, so that, under the null $\boldsymbol{s}'I^{-1}\boldsymbol{s}$ has a χ^2-distribution. In this case, all elements of $\boldsymbol{s}$ are zero except the first which is $\tilde{d}$

in (16). Hence, the score test for tobit is $\tilde{d}^2$ divided by the top left-hand element of the inverse of (17) and the resulting χ^2 has one degree of freedom. Since we are here interested in one-tailed tests, we shall also note the sign of $\tilde{d}$ itself.

3. Short-cut estimation techniques

Greene (1981) has recently suggested an OLS estimation technique for the tobit model which is consistent under appropriate but non-standard (and implausible) assumptions. In this section we show that the procedures can straightforwardlly be applied to the p-tobit model discussed above. Since the assumptions guaranteeing consistency are stringent and unlikely to be met in practice, the OLS procedure should only be used to provide starting values for maximum likelihood. However, the procedure is also useful in theoretical work since it generates estimators which are consistent under at least some circumstances and which, because they have explicit formulae, can be manipulated and analysed with some ease.

As before, let y be a realization of a (latent) uncensored variable and $\boldsymbol{x}$ a realization of a vector of conditioning variables. y and $\boldsymbol{x}$ are to be thought of as being drawn from some parent joint distribution. We are interested in the regression of y on $\boldsymbol{x}$, i.e. in

$$\mathrm{E}(y|\boldsymbol{x})=\beta_0+\boldsymbol{x}'\boldsymbol{\beta}, \tag{18}$$

where y and $\boldsymbol{x}$ are now normalized to have means of zero. Consider now some censoring process on y which generates a new random variable Z, say, such that $Z=z$ if and only if $y\in Y(z)$, where the sets $Y(z)$ have zero intersection and cover the range of y. The realizations of Z, unlike those of y, are observable. Assume that the observations on z are also normalized to have zero mean and let $\hat{\boldsymbol{\beta}}$ be the OLS regression of z on $\boldsymbol{x}$, i.e.

$$\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'z, \tag{19}$$

so that $\hat{\boldsymbol{\beta}}$ is the estimate of the *slopes* using the z realizations in place of the unobservable y. The following proposition is then easily established and is close that that given by Goldberger (1981) for the more difficult case of truncation. See also Chung and Goldberger (1982) for a fuller discussion of the current case.

Proposition. If the joint distribution of y and $\boldsymbol{x}$ is such that the conditional expectation of $\boldsymbol{x}$ given y is linear in y, i.e. if

$$\mathrm{E}(\boldsymbol{x}|y)=\boldsymbol{\alpha}_0+\boldsymbol{\alpha}_1 y \tag{20}$$

for some constant vectors α_0 *and* α_1, *then* $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}/\theta$, *where* $\theta = \text{cov}(z, y)/\text{var}(y)$, *is consistent for* $\boldsymbol{\beta}$.

Note that the proposition does *not* require multivariate normality of y and $\boldsymbol{x}$, although joint normality is clearly sufficient for the linearity (20). However, it is only linearity *plus* homoscedasticity that implies normality, and homoscedasticity is not required for the result; there are many other distributions for which (20) holds. For all of these, without further specification, the result implies that the *ratios* of the slopes are consistently estimated by the *ratios* of the OLS slope estimators. If one wants to go further, the procedure for consistently estimating θ is required and this generally requires knowledge of *both* the precise censoring process linking y and z *and* the marginal distribution of y. The former is a matter of model specification while some information on the latter can usually be obtained by examination of z, a point to which we return below.

If y is assumed to be marginally normal, (20) implies a multivariate normal structure for y and $\boldsymbol{x}$ together, the case explicitly analysed by Greene. Under normality and the censoring structure of p-tobit, we have:

$$\theta = \frac{\text{cov}(z, y)}{\text{var}(y)} = p\Phi\left(\frac{\mu_0}{\sigma_0}\right), \tag{21}$$

where μ_0 and σ_0^2 are the mean and variance of (the marginal distribution of) y. It is clearly also true that the p-tobit model implies that $\pi \equiv n_1/n$, the proportion of the sample for which $z > 0$, has an expectation of $p\ \Phi(\mu_0/\sigma_0)$ so that θ is unbiasedly and consistently estimated by π. Hence,

$$\tilde{\beta} = \hat{\beta}/\pi = n\hat{\beta}/n_1 \tag{22}$$

is consistent for the slope coefficient β. Note that this is exactly the same estimator proposed by Greene for tobit. Of course, if $p < 1$, one would expect $\hat{\beta}$ from the p-tobit to suffer greater attenuation than $\hat{\beta}$ when tobit is true. But this is corrected for by division by π, a quantity which has a smaller expectation the greater is p.

The foregoing implies that if the short-cut estimation method is a good guide to the full MLEs, then allowing for additional binary censoring within tobit is unlikely to make much difference to the estimated slope coefficients. The main differences will lie in the intercept, in the variance, and in the estimate of p. These remaining parameters can conveniently be estimated by the method of moments. The first two moments, together with the formula for $\text{E}(\pi)$, are the most natural to use; under normality, these are:

$$\text{E}(\pi) = \text{E}(n_1/n) = p\Phi(\mu_0/\sigma_0), \tag{23}$$

$$E(z) = \sigma_0 p \Phi(\mu_0/\sigma_0)\{\mu_0/\sigma_0 + \lambda(\mu_0/\sigma_0)\}, \tag{24}$$

$$E(z^2) = \sigma_0^2 p \Phi(\mu_0/\sigma_0)\{1 + \mu_0/\sigma_0(\mu_0/\sigma_0 + \lambda(\mu_0/\sigma_0))\}, \tag{25}$$

where $\lambda(\mu_0/\sigma_0) = \phi(\mu_0/\sigma_0)/\Phi(\mu_0/\sigma_0)$, and μ_0 and σ_0^2 are the (marginal) mean and variance of y from which β_0 and the residual variance can readily be calculated. Substitution of the sample estimates on the left-hand side allows solution for estimates of p, μ_0 and σ_0. Denoting μ_0/σ_0 by ψ, then $\tilde{\psi}$ is the solution to

$$\{1 + (\lambda(\tilde{\psi}) + \tilde{\psi})\tilde{\psi}\} = \pi(1 + v^2)\{\lambda(\tilde{\psi}) + \tilde{\psi}\}^2, \tag{26}$$

where v is the coefficient of variation of z in the sample. A simple search procedure yields $\tilde{\psi}$ from (26) and $\tilde{p}$ and $\tilde{\sigma}$ are immediately given by (23) and (24).

4. Results

The survey considered here is the 1973–74 fiscal year data from the 1973 and 1974 Family Expenditure Surveys. We use data from all 6837 households, although in some experiments a 10 percent random sample was drawn in order to keep down unnecessary computer costs. The basic model takes the *share* of total expenditure devoted to a particular good as the dependent variable; this is conditioned on a fairly extensive set of 'independent' variables. Clearly, zero expenditures become zero shares. We consider three household expenditure categories each with a substantial proportion of households recording no expenditure over the two-week period of the survey. These are: tobacco, with 66.3 percent of households showing some purchase; alcohol, with 71.4 percent purchasing; and durables, with 90.8 percent purchasing. All expenditures are coded as weekly flows in tenths of pence per week. The variables and summary statistics are listed in table 1. The independent variables are divided into five groups: (1) economic: per capita total household expenditure, its square, and a dummy for households with no working members; (2) demographic: the composition of the household and ages of its members; (3) occupational dummies for six standard classifications; (4) indicators for the presence of certain stocks (car, telephone, house) that are to be regarded as commitments to certain types of expenditure (rates, mortgage repayment, telephone rental charges, car licence fees) and so should negatively influence expenditure on normal goods; and (5) regional dummies giving the standard region in which the household is located. Note that the use of total expenditure as an exogenous variable is theoretically inconsistent with our formulations of 'reporting' bias which imply that total expenditure is a random variable determined by the sum of

Table 1
Variables in the analysis.

			Symbol	Mean	Standard deviation
Dependent variables:					
tobacco share				0.040	0.045
alcohol share				0.046	0.062
durables share				0.053	0.090
Independent variables:			*Symbol*		
1. (i)	log per capita household expenditure		LPCE	9.519	0.538
(ii)	log per capita household expenditure squared		$LPCE^2$	90.901	10.413
(iii)	household with no workers	(d)	DW	0.215	0.411
2. (i)	number of adults		#Ad	1.980	0.702
(ii)	number of children under 2 years		#Ch(<2)	0.090	0.299
(iii)	number of children over 2 and under 5		#Ch(2–5)	0.158	0.424
(iv)	number of children 5 years and over		#Ch(>5)	0.627	1.087
(v)	head of household aged <25	(d)	AGP1	0.046	0.210
(vi)	head of household aged 25–35	(d)	AGP2	0.180	0.384
(vii)	head of household aged 35–45	(d)	AGP3	0.167	0.373
(viii)	head of household aged 45–60	(d)	AGP4	0.268	0.443
(ix)	head of household aged 60–65	(d)	AGP5	0.097	0.296
(x)	head of household aged ≥65	(d)	AGP6	0.242	0.428
(xi)	Sex of head of household	(d1)	Sex	1.199	0.399
(xii)	Household has zero children	(d)	DC	0.583	0.493
(xiii)	Household consists of a single adult	(d)	DAC	0.187	0.390
3. (i)	Occupation is professional, administrative, teacher, etc.	(d)	OC1	0.200	0.400
(ii)	Occupation is clerical	(d)	OC2	0.064	0.244
(iii)	Occupation is shop assistant	(d)	OC3	0.008	0.091
(iv)	Occupation is armed forces	(d)	OC4	0.006	0.074
(v)	Occupation is retired, unoccupied	(d)	OC5	0.266	0.442
(vi)	Occupation is manual worker	(d)	OC6	0.456	0.498
4. (i)	Household has one or more cars	(d)	CD	0.546	0.498
(ii)	Household has two or more cars	(d)	CCD	0.098	0.298
(iii)	Household has telephone	(d)	TEL	0.452	0.498
(iv)	Household is an owner occupier	(d)	OOC	0.493	0.500
5. (i)	Household lives in Northern region	(d)	R1	0.066	0.248
(ii)	Household lives in Yorkshire region	(d)	R2	0.090	0.286
(iii)	Household lives in E. Midlands region	(d)	R3	0.062	0.241
(iv)	Household lives in E. Anglia region	(d)	R4	0.037	0.189
(v)	Household lives in Greater London region	(d)	R5	0.128	0.334
(vi)	Household lives in South East (excl. London) region	(d)	R6	0.170	0.376
(vii)	Household lives in South West region	(d)	R7	0.068	0.251
(viii)	Household lives in Wales region	(d)	R8	0.049	0.216
(ix)	Household lives in W. Midlands region	(d)	R9	0.094	0.292
(x)	Household lives in N. West region	(d)	R10	0.122	0.327
(xi)	Household lives in Scotland region	(d)	R11	0.091	0.288
(xii)	Household lives in N. Ireland region	(d)	R12	0.023	0.149

Notes.
(d) indicates a dummy variable which is 1 if the descriptor is true, it is otherwise zero.
(d1) sex = 1 if head is male, 2 if female.

all the reporting effects over all goods. This issue can only really be dealt with in the context of a system of demand equations and is central in Kay, Keen and Morris's (1982) paper. For the time being we ignore it.

The results of the analysis of the tobacco data are given in table 2. Columns 1–3 relate to the straightforward tobit specification. Column 1 lists

Table 2
Tobacco share. Full sample, 1973–74.

	Tobit			*p*-tobit		
Variable	OLS (β/σ)	MLE (β/σ)	(S.E.)	OLS (β/σ)	MLE (β/σ)	(S.E.)
Constant	−0.828	−3.890	3.032	−1.154	−3.932	3.021
LPCE	0.726	1.240	0.628	0.691	1.242	0.626
$LPCE^2$	−0.050	−0.070	0.032	−0.048	−0.070	0.032
#Ad	0.073	0.085	0.032	0.069	0.086	0.031
#Ch (<2)	0.024	0.027	0.057	0.023	0.027	0.056
#Ch (2–5)	−0.049	−0.036	0.042	−0.047	−0.036	0.042
#Ch (>5)	−0.095	−0.058	0.024	−0.090	−0.058	0.024
Sex	−0.525	−0.552	0.038	−0.500	−0.548	0.039
DC	0.111	0.076	0.058	0.106	0.073	0.057
DAC	−0.000	−0.125	0.051	−0.000	−0.125	0.051
DW	−0.571	−0.537	0.065	−0.543	−0.532	0.066
OC1	−0.307	−0.325	0.042	−0.292	−0.324	0.042
OC2	−0.122	−0.113	0.054	−0.116	−0.114	0.053
OC3	0.086	0.097	0.131	0.082	0.095	0.130
OC4	−0.459	−0.502	0.183	−0.437	−0.499	0.182
OC5	0.303	0.297	0.068	0.288	0.294	0.068
AGP2	0.079	0.074	0.069	0.075	0.074	0.068
AGP3	0.141	0.116	0.073	0.134	0.115	0.073
AGP4	0.213	0.181	0.067	0.203	0.180	0.066
AGP5	0.128	0.096	0.074	0.122	0.095	0.074
AGP6	−0.235	−0.263	0.074	−0.224	−0.262	0.074
CD	−0.319	−0.262	0.033	−0.304	−0.259	0.034
CCD	−0.066	−0.036	0.057	−0.063	−0.037	0.057
TEL	−0.151	−0.136	0.032	−0.144	−0.135	0.032
OOC	−0.330	−0.333	0.030	−0.314	−0.331	0.030
R2	−0.051	−0.046	0.062	−0.049	−0.044	0.061
R3	−0.156	−0.116	0.072	−0.148	−0.114	0.072
R4	−0.310	−0.306	0.085	−0.295	−0.304	0.085
R5	−0.189	−0.158	0.060	−0.180	−0.156	0.061
R6	−0.259	−0.252	0.058	−0.246	−0.250	0.058
R7	−0.224	−0.212	0.069	−0.213	−0.210	0.069
R8	0.017	0.029	0.074	0.016	0.029	0.074
R9	−0.087	−0.065	0.064	−0.083	−0.064	0.064
R10	−0.042	0.001	0.061	−0.040	0.003	0.061
R11	0.171	0.107	0.061	0.163	0.106	0.060
R12	0.163	0.109	0.086	0.155	0.109	0.086
p	—	($\chi^2_1 = 0.019$)		1.069	1.004	0.015
$1/\sigma$	17.796	17.701	0.163	16.917	17.634	0.238
2 log L	9505.0	9590.4			9590.4	

the starting values for ML estimation obtained by Greene's procedure applied to the OLS estimates; obtaining these requires no iterative procedures. Column 2 gives the corresponding MLEs and column 3 the estimates of the asymptotic standard errors. These require iterative methods and, given the size of the sample, the calculations are not inexpensive. Although the likelihood changes considerably, in this particular case the differences in parameter estimates are typically not large (even though independent variables are clearly not multinormally distributed). In substantive terms, the results make a good deal of sense. Tobacco is an inferior good for all those with per capita household expenditure greater than about £10 per week in 1973–74; in addition, households with no working members buy less even at the same level of PCE. The use of tobacco is positively related to the number of adults in the household, but is lower if there are older children, if the household is headed by a woman, or if the household is composed of a single adult (perhaps because such individuals are either relatively young or relatively old). As the age of the head of household increases up to 60, so do tobacco purchases and the consumption of the 45–60 group is significantly higher than of the under 25s. This presumably reflects the relative success of anti-smoking propaganda amongst the young. After age 60, the positive effects diminish and become negative after 65; heads of household who reach the age of retirement are rather less likely to be smokers! The presence of the various stocks all decrease consumption, although these effects may reflect omitted wealth or social status variables as well as the income effects of the stocks. Several regional effects are important, with tobacco purchases generally lower in the south and east of the country.

The comparison between tobit and p-tobit suggests no misspecification of the former in the direction predicted by the latter. The score test has a value of 0.019 which is not significant and the short-cut p-tobit estimation procedure (column 4) yields a value of p of 1.069. Just to double-check, the MLEs were calculated and are given in column 5 and 6. These are very close to the tobit MLEs, yield an estimate of p at 1.004, and a log-likelihood value essentially identical to that of the tobit, thus confirming the results of the score test. There is therefore no evidence from these tests of the operation of a binary censor in addition to the censoring explicable by tobit. And although one would perhaps not expect the frequency of purchase model to be useful for tobacco (British consumers, unlike Americans, rarely buy cigarettes in large quantities — or did not do so in 1973–74), the deliberate suppression of information might reasonably have been expected to exist. But there is no evidence of it here and this might be regarded as (rather weak, see below) evidence in favour of the design-based explanation of under-recording advanced in the *Family Expenditure Survey Handbook.*

The calculations were next carried out for durable goods and for alcoholic

beverages. Once again, and using the same set of independent variables, but now a 10 percent subsample, the Greene procedure gave good starting values for the MLEs, and the MLEs themselves give perfectly reasonable tobit estimates, see table 3. For both of the data sets the two PCE terms are

Table 3
Durables and alcohol shares, 10 percent sample 1973–74, tobit estimates.

	Alcohol			Durables		
Variables	OLS (β/σ)	MLE (β/σ)	(S.E.)	OLS (β/σ)	MLE (β/σ)	(S.E.)
Constant	−21.621	−40.499	12.884	25.494	25.425	10.213
LPCE	4.297	8.168	2.650	−5.725	−5.725	2.040
LPCE^2	−0.203	−0.398	0.136	0.332	0.332	0.102
#Ad	0.395	0.382	0.103	−0.000	0.032	0.127
#Ch(<2)	0.161	0.175	0.220	−0.183	−0.192	0.203
#Ch(2–5)	0.014	0.064	0.160	−0.026	−0.047	0.142
#Ch(≧5)	0.051	0.090	0.080	−0.076	−0.078	0.079
Sex	−0.520	−0.688	0.136	−0.018	0.005	0.162
DC	0.148	0.147	0.191	−0.533	−0.535	0.170
DAC	0.476	0.267	0.174	−0.134	−0.219	0.204
DW	−0.625	−0.692	0.205	−0.021	0.072	0.256
OC1	−0.097	−0.182	0.131	−0.044	−0.022	0.141
OC2	0.011	−0.098	0.187	0.172	0.153	0.157
OC3	−0.116	0.025	0.740	−0.105	−0.006	0.889
OC4	−0.049	−0.061	1.402	−0.256	−0.220	1.868
OC5	0.410	0.456	0.226	−0.107	−0.153	0.274
AGP2	0.013	0.028	0.207	0.155	0.224	0.206
AGP3	−0.229	−0.226	0.209	0.157	0.219	0.227
AGP4	−0.271	−0.271	0.194	−0.028	0.022	0.220
AGP5	−0.330	−0.395	0.224	0.081	0.147	0.244
AGP6	−0.541	−0.592	0.229	0.125	0.146	0.256
CD	−0.419	−0.377	0.116	−0.170	−0.162	0.115
CCD	0.065	0.005	0.176	−0.099	−0.088	0.192
TEL	−0.166	−0.110	0.113	−0.149	−0.122	0.109
OOC	−0.110	−0.042	0.108	0.072	0.104	0.110
R2	0.233	0.191	0.196	−0.245	−0.251	0.221
R3	−0.334	−0.409	0.275	−0.341	−0.312	0.299
R4	−0.337	−0.175	0.356	0.120	0.129	0.240
R5	−0.263	−0.269	0.220	−0.197	−0.221	0.207
R6	−0.390	−0.378	0.216	−0.307	−0.302	0.196
R7	−0.068	−0.083	0.233	−0.234	−0.198	0.266
R8	−0.003	0.008	0.244	−0.438	−0.457	0.334
R9	−0.028	−0.014	0.213	−0.004	−0.059	0.203
R10	−0.025	−0.101	0.198	−0.210	−0.253	0.203
R11	−0.264	−0.232	0.221	−0.095	−0.101	0.190
R12	0.007	−0.146	0.272	−0.433	−0.496	0.525
$1/\sigma$	13.745	14.067	0.377	10.931	11.140	0.260
Normalized score statistic		$\chi^2_1 = 51.55$			$\chi^2_1 = 214.24$	
2log L	933.84	958.76		1137.8		1144.8

significant and expenditure elasticities are in the region of 2. Alcohol shares, however, are decelerating with rising PCE, while durable shares are accelerating over the range in question. Alcohol shares are positively related to the number of adults in the household but are significantly reduced if the head of household is a woman or if there is no worker. The share also falls with age and with car ownership. Apart from the PCE terms the only explanatory variable of any significance in determining the share of expenditure on durables is the indicator of a household with no children, which reduces the share.

However, when we come to score tests, the tobit model is rejected. The normalized scores [to be compared with a $N(0, 1)$ under the null] are $+14.6$ for durables and $+7.2$ for alcohol. The positive signs indicate that at the tobit MLEs, the likelihoods would be locally increased by *increasing* the value of p beyond unity, a result which makes no sense in terms of the alternatives discussed above. Even with beginning of period effects, with purchases reported that do not exist, we should still expect the likelihood to be maximized at a value of p less than unity [see (13)]. These results can also be cross-checked by other methods. The short-cut estimates of p-tobit suggest an essentially infinite estimate for p for both models, a result which was not contradicted by our (essentially unsuccessful) attempts to estimate p-tobit directly by maximum likelihood. Table 4 reports the results of crude grid searches over values of p for 10 percent subsamples of the three commodities considered. The tobacco MLE of 0.91 differs from that shown in table 2 because the latter is based on the full sample.

So, whatever value of p indeed maximizes the likelihood, it is very much larger than unity. Such a finding is of course inconsistent *both* with the null and with the alternative. If tobit is correct, p should be estimated as insignificantly different from unity and the scores should be insignificantly different from zero; neither is the case and indeed the scores are large and positive. The alternative, p-tobit, predicts negative scores and estimates of p less than unity, so that it, like tobit, offers no explanation of what is actually observed. Nor are durables and alcohol atypical in this respect and qualitatively identical results have been obtained for five other broad group

Table 4
Likelihood values for various values of p.

Tobacco		Alcohol		Durables	
p	$2\ln L$	p	$2\ln L$	p	$2\ln L$
1	926.9	1	958.7	1	1144.8
0.91	928.7	0.9	913.2	0.9	1041.4
0.8	920.5	0.8	851.7	0.8	913.8
0.7	897.6	0.7	771.8	0.7	763.3

expenditures from the FES. Hence, taking *all* broad groups (we distinguish ten in all which together comprise total household expenditure) which leave some households showing zero purchases, for only tobacco is the tobit not rejected, and in all cases the rejection is in the opposite direction to that predicted by p-tobit. We have also looked at some quite different data in which zero expenditures are even more common. These are the household consumption figures from the 1969–70 Sri Lanka Socio-economic Survey and once again, for all categories showing zero expenditures, tobit is rejected in the direction opposite to that predicted by p-tobit. The rejection of the tobit is therefore widespread, and tobacco in the FES is so far the only good that we have found (and the first we tried) for which the model appears satisfactory. In all the other cases, the fitted distributions predict fewer households making zero purchases than is actually the case.

5. Non-parametric estimation

One possible cause of our results is a failure of the normality assumption which is embodied in both tobit and p-tobit. There is no particular reason to assume normality (other than analytical convenience and familiarity), and the basic features of the models do not depend upon it. However, the empirical results are unlikely to be robust against failures of normality. In the usual linear regression model, unbiasedness, efficiency, and consistency do not depend on normality. This is no longer the case for models such as tobit, where normality is necessary even for consistency and where its failure can have serious consequences [see the examples in Goldberger (1980) and Arabmazar and Schmidt (1982)]. In this section we re-estimate the simple tobit without the normality assumption. This can be done in several ways. One possibility is to work with more general distributions, for example members of the Pearson system, and test the specialization to normality. Alternatively, we can break altogether with parametric forms, and use one of the currently available techniques for non-parametric estimation of censored regression models. Several of these are discussed by Miller (1981) and, following the recommendations of Miller and Halpern (1982), we have used the method proposed by Buckley and James (1979). Amemiya (1982) suggests the use of a least absolute deviation estimator as proposed by Powell (1981), but we became aware of this work too late to consider using it. A brief outline of the basis of the Buckley–James technique is given in the appendix; fuller discussion of the calculations for this paper is given in Irish (1982).

Table 5 compares results of the parametric and the non-parametric estimation procedures for tobacco, for alcohol, and for durables, again based on the subsample of 700 observations. To save space, only the parameters of the first 12 explanatory variables are included in the tables; all variables were included in the analysis. For tobacco, the MLEs and the Buckley–James

Table 5
Parametric and non-parametric estimates.[a]

	Tobacco		Alcohol		Durables	
Variable	MLE	BJ	MLE	BJ	MLE	BJ
Constant	−93.3 (68.9)	−90.5 (62.4)	−288 (91.6)	−230 (85.4)	228 (91.7)	160 (100.3)
LPCE	20.8 (14.3)	20.3 (12.8)	58.1 (18.8)	46.3 (17.5)	−51.4 (18.3)	−37.1 (20.6)
LPCE^2	−1.08(0.75)	−1.06(0.66)	−2.83(0.97)	−2.23(0.90)	2.98(0.91)	2.24(1.05)
#Ad	1.21(0.60)	1.22(0.44)	2.72(0.73)	2.56(0.65)	0.28(1.14)	0.11(0.85)
#Ch(<2)	−0.38(1.02)	−0.37(0.83)	1.25(1.57)	1.19(1.14)	−1.72(1.82)	−1.36(1.48)
#Ch(2–4)	−0.20(0.71)	−0.21(0.58)	0.45(1.14)	0.37(0.80)	−0.42(1.27)	−0.14(1.06)
#Ch(>4)	0.00(0.38)	−0.01(0.30)	0.64(0.57)	0.50(0.43)	−0.70(0.71)	−0.53(0.55)
Sex	−3.90(0.70)	−3.89(0.74)	−4.89(0.96)	−4.60(1.15)	−0.05(1.45)	−0.04(1.33)
DC	0.11(0.92)	0.18(0.75)	1.04(1.36)	0.93(1.05)	−4.80(1.52)	−4.47(1.35)
DAC	0.42(1.04)	0.43(0.93)	1.90(1.24)	2.20(1.45)	−1.97(1.83)	−1.75(1.67)
DW	−3.98(1.08)	−4.17(1.09)	−4.92(1.46)	−4.86(1.69)	0.20(2.30)	0.13(2.01)
OOC	−1.91(0.54)	−1.92(0.46)	−0.30(0.77)	−0.41(0.06)	0.94(0.98)	0.75(0.82)
$\hat{\sigma}_{MLE}$	5.509	—	7.109	—	8.977	—
$\hat{\sigma}_u$	4.139	4.134	6.412	6.273	8.720	8.708
$\hat{\mu}_u$	2.879	2.800	2.311	1.958	0.960	0.707

[a]The alcohol and durables figures under MLE are the same as those given in table 3, although here estimates for β and σ are given, rather than for β/σ and $1/\sigma$. Standard errors are given in parentheses and all figures are multiplied by 10^2.

estimates are essentially identical; for alcohol and for durables there are a few differences, for example in LPCE and its square, but, once again, the dominant impression is of little change from estimation technique to the other. This evidence suggests little reason to believe that failure of normality is a major source of inconsistency in the previous estimates.

The Buckley–James estimation procedure also produces an empirical estimate of the distribution function of the (uncensored) residuals and this can be used to re-evaluate the score statistic (16). Of course, such a procedure is without a strict formal basis. However, it is easily shown that the score test is correct for *any* distribution provided the Φ_i's in (16) are replaced by the appropriately evaluated distribution functions. Furthermore, as shown in the appendix, the Buckley–James estimator can be thought of as an approximate MLE if the (unknown) density of the residuals belongs to the exponential family of distributions. It thus seems plausible that the value of (16) will once again provide an indication of the presence of a binary censor in addition to the censored regression model. The un-normalized values of d, using the non-parametric distribution function estimator, are 14.5 (tobacco), 82.7 (alcohol) and 217.5 (durables). We have not been able to calculate standard errors in order to normalize these figures, but note that (a) all the scores remain positive and (b) they are similar to (although somewhat smaller than) the un-normalized scores calculated under normality, so that given the similarity in standard errors in table 5, the normalized scores can reasonably be supposed to reject the alcohol and durables models as before. Once again, the uncensored distribution, even without normality, leads us to expect too many zeroes.

6. Conclusions

We have proposed a number of models of misreporting designed to help explain the number of households reporting zero purchases of various goods. All of these have the implication that zero purchases are more likely to occur than would be the case if all zero purchases represented genuine non-consumption. Our tests of this hypothesis, with or without the incidental assumption of a normal distribution for unobserved household heterogeneity, all give the same result, that there are *too few* zero purchases, not too many. For only one good, tobacco, is this not the case; for all the others considered both from the British Family Expenditure Survey (1973–74) and from the Sri Lankan Socio-economic Survey (1969–70), the test statistics reject the censored regression model but in the direction opposite to that predicted by the misreporting models. We cannot at this stage offer any convincing explanation of this phenomenon.

One interesting suggestion that has been offered is that the experiments must be repeated with more finally disaggregated commodities. In the data

as currently used, alcohol expenditure includes purchases of alcohol for home consumption as well as purchases in public houses, and there is no distinction between beers, wines and spirits. The durable category is even more heterogeneous and includes, for example, weekly imputed sums for insurance premia on durable goods. Such considerations suggest that mixtures of distributions be considered. This is a conceivable explanation for a failure of normality but it is unclear to us at this point why the non-parametric estimator would not satisfactorily deal with such mixing. It is also possibly the case that the binary censoring process is not independent of the tobit process, although, once again, it is unclear why this could account for our results.

There are also some incidental points worth making. On the models themselves, it is clear that quite different formulations can give rise to similar or identical statistical structures. This is not surprising; only parts of the data generation processes are observed and we should not expect to be able to distinguish between models which differ only in unobservables. On estimation, it turns out that on these data, lack of normality is not a serious problem. However, when it is (or may be), there exist feasible (although not cheap) non-parametric techniques. These deserve more exploration in econometrics, in spite of the currently somewhat unsatisfactory state of their development in statistics.

Appendix: Non-parametric estimation of censored regression models

This appendix provides a brief, intuitive summary of the Buckley and James estimator in an econometric context. Many issues are not discussed; interested readers should consult Buckley and James (1979) or the excellent exposition in Miller (1981, pp. 39–57 and pp. 150–154). The details of the calculation for the present case are given in Irish (1982).

Begin from the 'complete' data model:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i. \tag{A.1}$$

In the censored regression model, y_i is observed as y_i if y_i is uncensored (i.e. when $y_i > 0$ in this case), otherwise a zero is recorded. If y_i were known, β could be estimated in the usual way, i.e. by $(X'X)^{-1}X'y$. But some y_i's are missing; those represented by zeroes. However, given β, we can calculate the expectation of these missing y_i's using:

$$\mathrm{E}(y_i \mid y_i < 0) = \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(u_i \mid u_i < -\boldsymbol{x}_i'\boldsymbol{\beta}). \tag{A.2}$$

Denote the last-term on the right-hand side by θ_i, where

$$\theta_i \equiv \theta(x_i'\beta) = \int_{-\infty}^{-x_i'\beta} u\,dF(u)/F(-x_i'\beta), \tag{A.3}$$

where $F(\cdot)$ is the distribution function of u_i and we have assumed that the u_i's are i.i.d. Estimation can then proceed using the usual normal equations, but replacing the censored y_i's, i.e. the zeroes, by their expectations as given by (A.2). To write this conveniently, partition the X matrix as $[X_u' \quad X_c']'$ conformably with $y=[y_u'0']'$ for censored (c) and uncensored (u) observations. The new normal equations are then:

$$[X_u'X_u + X_c'X_c]\hat{\beta} = X_u'y_u + X_c'X_c\hat{\beta} + X_c'\theta$$

or

$$(X_u'X_u)\hat{\beta} = X_u'y_u + X_c'\theta. \tag{A.4}$$

Since θ depends on β, (A.4) must be solved iteratively, even if $F(\cdot)$ is known, but an obvious Gauss–Siedel scheme suggests itself, i.e.

$$\hat{\beta}_{(r+1)} = (X_u'X_u)^{-1}(X_u'y_u + X_c'\theta(\beta_{(r)})). \tag{A.5}$$

Although this whole procedure is an intuitive one, it can be given a rigorous formal basis. Provided $F(\cdot)$ belongs to the exponential family of distributions, the condition (A.4) together with (A.3) turns out to be the first-order condition for maximum-likelihood [see Dempster, Laird and Rubin (1977)]. And in this case, the iterative scheme (A.5) is the EM algorithm which, both before and after the Dempster, Laird and Rubin article, was and is widely applied to such incomplete data models.

In the current case, we wish to avoid assuming a parametric form for $F(\cdot)$. The usual empirical distribution function estimator is excluded because of the censoring, but Buckley and James suggest estimating $F(\cdot)$ by the product-limit, or Kaplan–Meier (1958), estimator. This is formed from the residuals at each iteration of (A.5) so that, as the procedure advances, not only is $\hat{\beta}$ updated, but so is the estimate of $F(\cdot)$. Given a vector of residuals, the product-limit estimator works as follows. Let e be an n-vector of residuals and order them from smallest to largest (assuming no ties) as:

$$e_{(1)} < e_{(2)} < e_{(3)} < \ldots < e_{(n-1)} < e_{(n)}. \tag{A.6}$$

Note that *all* residuals are included, whether from censored observations or not, but we know which of the residuals are associated with censored y_i's and which with uncensored y_i's. If a given $e_{(i)}$ belongs to a censored y_i, we

know that it is too high in the ordering; the recorded value is $-x_i'\beta$, whereas we know that the true residual is less than or equal to $-x_i'\beta$. Consider the residual $e_{(i)}$, censored or not. Then:

$$\begin{aligned}\Pr\{u \leqq e_{(i)}\} &= \Pr\{u \leqq e_{(i)} \mid u \leqq e_{(i+1)}\}\cdot \\ &\quad \Pr\{u \leqq e_{(i+1)} \mid u \leqq e_{(i+2)}\}\cdot \\ &\quad \cdots \\ &\quad \Pr\{u \leqq e_{(n)}\}\cdot \\ &= p_i \cdot p_{i+1} \cdots p_n \cdot \end{aligned} \tag{A.7}$$

in an obvious notation. The intuitive empirical estimator for p_i is $1-1/(i+1)=i/(i+1)$ if $e_{(i+1)}$ is uncensored, and is 1 if $e_{(i+1)}$ is censored; there are $(i+1)$ observations less than or equal to $e_{(i+1)}$, if $e_{(i+1)}$ is uncensored, one of them is in the interval $(e_{(i)}, e_{(i+1)}]$, while if $e_{(i+1)}$ is censored, its uncensored value must be smaller and there is no valid observation in the interval. This gives the product-limit estimator as:

$$\hat{F}(x) = \hat{\Pr}\{u \leqq x\} = \prod_{e_{(i)} > x}^{u} \left(\frac{i-1}{i}\right), \tag{A.8}$$

where the superscript u indicates that the product is taken only over the uncensored residuals. Clearly $\hat{F}(x)$ has jump points only at these uncensored residuals although, if $e_{(1)}$ is censored, it is conventionally assigned the remaining weight to ensure that $\hat{F}(x)=0$ for $x<e_{(1)}$. $\hat{F}(\cdot)$, once calculated, is used directly in the denominator of (A.3) and its 'jumps' are treated as weights for the uncensored residuals below $-x_i'\beta$ in order to calculate the numerator.

The Kaplan–Meier estimator of $\hat{F}(\cdot)$ is known to be consistent [see Miller (1981, pp. 59–63)], but no proof yet exists of the consistency of the Buckley–James estimator of β. There are also practical difficulties in that the algorithm (A.5) may cycle between several distinct values of β instead of converging to a single estimate. In such cases, the recommended procedure is to average such values. Calculation of standard errors is also somewhat ad hoc; again, see Irish (1982) for the exploration of some alternatives.

References

Amemiya, Takeshi, 1982, Tobit models: A survey (Rhodes Associates, Criminal Justice Research Series, Palo Alto, Ca.).

Arabmazar, Abbas and Peter Schmidt, 1982, An investigation of the robustness of the tobit estimator to non-normality, Econometrica 50, 1055–1063.
Buckley, Jonathan and Ian James, 1979, Linear regression with censored data, Biometrika 66, 429–436.
Chung, C-F. and Arthur S. Goldberger, 1982, Proportional projections in limited dependent variable models, Department of Economics, University of Wisconsin, mimeo.
Cragg, John G., 1971, Some statistical models for limited dependent variables with applications to the demand for durable goods, Econometrica 39, 829–844.
Dempster, A.P., N.M. Laird and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B39, 1–22.
Goldberger, Arthur, S., 1980, Abnormal selection bias, University of Wisconsin, Madison, mimeo.
Goldberger, Arthur S., 1981, Linear regression after selection, Journal of Econometrics 15, 357–366.
Greene, William H., 1981, On the asymptotic bias of the ordinary least squares estimation of the tobit model, Econometrica 49, 505–513.
Hausman, Jerry A., 1978, Specification tests in econometrics, Econometrica 46, 1251–1272.
Heckman, James J., 1980, Sample selection bias as a specification error, in: J.P. Smith, ed., Family labor supply (Princeton University Press).
Irish, Margaret, 1982, Linear regression using censored data, Department of Economics, University of Bristol, mimeo.
Kaplan, E.L. and Paul Meier, 1958, Nonparametric estimation from incomplete observations, Journal of the American Statistical Association 53, 457–481.
Kay, John A., M.J. Keen and C.N. Morris, 1982, Consumption, income and the interpretation of household expenditure data, Institute for Fiscal Studies, London, mimeo.
Kemsley, W.F.F., R.V. Redpath and M. Holmes, 1980, Family Expenditure Survey handbook (HMSO, London).
Miller, Rupert G., 1981, Survival analysis (John Wiley, New York).
Miller, Rupert G. and Jerry Halpern, 1982, Regression with censored data, Division of Biostatistics, Stanford University Medical Center, California, mimeo.
Nelson, Forrest D., 1981, The effect of a test for misspecification in the censored-normal model, Econometrica 49, 1317–1329.
Powell, J.L., 1981, Least absolute deviations estimation for censored and truncated regression models, Technical report no. 356, IMSSS, Stanford University.
Tobin, James, 1958, Estimation of relationships for limited dependent variables, Econometrica 26, 24–36.