Finnish Economic Papers – Volume 1 – Number 2 – Autumn 1988

ESTIMATING ENGEL CURVES: A generalisation of the P-Tobit model

STEPHEN PUDNEY*

London School of Economics and Political Science, London CW2A 2AE, UK

Cross-section demand relationships are usually estimated using data from short-duration expenditure surveys. The interpretation of such observations is not straightforward, since a zero recorded expenditure may understate true demand and a positive expenditure overstates demand. Deaton and Irish have recently proposed and applied the P-Tobit technique to deal with this problem, but with little success. The present paper specifies a generalisation of their model and applies alternative estimators to two different sets of UK survey data.

1. Introduction

A major difficulty in applied cross-section demand analysis lies in the fact that observed data usually come from short-duration expenditure surveys, while the economic theory of the consumer runs in terms of long-term average rates of consumption. The relationship between these two is not simple: if a household is not observed to purchase a particular good in the survey period, it may nevertheless be a consumer of that good *on average* in the longer run; moreover, even if a purchase is observed, the underlying rate of consumption is not necessarily equal to the observed purchase rate, particularly for goods with a storage life greater than the length of the survey period. This difficulty has been well known to practitioners since the earliest work in the field, but there has so far been no convincing theoretical framework for the analysis of consumption behaviour from short-duration expenditure data. It is the aim of this paper to propose, and to examine the estimation problems associated with, some specific econometric models which provide such a framework.

The fundamental distinction we maintain throughout is between consumption and expenditure. The underlying rate of consumption of the good is denoted c_n, where $n = 1 \ldots N$ indexes the households in our sample. The variable c_n is a purely theoretical concept: it is the choice variable in a static utility maximisation problem solved by the consumer. It is to be interpreted as the average (weekly) rate of consumption that we would arrive at by observing the household over a very long period during which all external conditions (prices, income, family composition, etc.) remain unchanged. The form of c_n is irrelevant here: it may be measured in quantity, expenditure or budget share terms. The major problem of cross-section analysis is that we cannot observe households for long periods under unchanging conditions, so c_n, which is the variable we are trying to explain, is not observable.

What we do observe is a variable e_n, de-

* *I am grateful to the participants of seminars at LSE, Bristol, Cambridge and Oxford for helpful comments on earlier versions of this paper. This research was financed by the Leverhulme Trust.*

fined as the total expenditure on the good over some short observation period, divided by the number of weeks in that period. Again, e_n may be expressed in quantity, expenditure or budget share form.

Our aim in this paper is to suggest some simple statistical models for the Engel curve underlying c_n and some mechanisms relating the unobservable c_n to the observable e_n. We then use these to derive estimation techniques for the Engle curve.

The paper is organised as follows. Section 2 considers the problem of specifying statistical Engel curves on a single-equation basis, and suggests a classification of goods into types requiring different statistical models. Section 3 discusses the relationship between e_n and c_n, concentrating particularly on a generalisation of the P-Tobit model of Deaton and Irish (1984). Section 4 proposes a simple nonlinear least squares estimator which does not require the specification of a detailed model of purchasing behaviour; some preliminary results are also discussed. Section 5 introduces the problem of specifying and estimating a joint model of consumption and purchasing behaviour, and Section 6 presents some empirical results.

2. *Modelling the rate of consumption*

Wales and Woodland (1983) and Lee and Pitt (1984) discuss the problem of estimating a full system of demand equations from cross-section data, allowing for the non-consumption of certain goods. The appalling complexity of the statistical models that result is sufficient reason to abandon a full-system approach at this micro level, and to treat demand behaviour on a single-equation basis. As we shall argue below, different goods tend to have distinctive features which require special treatment, and this also makes a full-system specification rather restrictive. Our approach therefore, is a very simple one: we shall concentrate chiefly on the consumption — purchases problem, using convenient Engel curve models.

However, as a pedagogical device to motivate our choice of models for specific goods, it is helpful to interpret alternative statistical models for c_n in terms of a simple two-good utility-maximisation problem. I propose a classification of goods into the following four types.

Type 1: Everyone consumes

For some goods there can be no non-consumption. Everyone wears clothes, and eats food. For goods of this type, we know that all observed zeros must be purely fortuitous: a household may not buy clothes in a particular two-week observation period, but its members do not habitually go naked.

Thus, if we think of the Engel curve for c_n arising from a utility maximisation problem, we have

$$(1) \quad \max u(c, C; \theta_n, \varepsilon_n)$$

subject to

$$(2) \quad p_n c + P_n C = y_n,$$

where C is consumption of a composite 'other goods' category, p_n and P_n are prices, y_n is exogenous total expenditure, θ_n represents observed demographic influences and ε_n represents random preference variation. For convenience, assume that the solution to this problem is an Engel curve linear in some transformations of p_n, P_n, y_n and θ_n:

$$(3) \quad c_n = \beta' x_n + \varepsilon_n$$

where x_n is an observable vector constructed from p_n, P_n, y_n and θ_n and β is a vector of constant parameters.

For this class of goods, u(.) is such that the indifference curves never cut the C-axis, and a corner solution for c_n is impossible. Thus we must choose a distributional form for ε_n (or equivalently for $c_n|x_n$) which ensures that c_n is strictly positive with probability one. There are many possibilities, and I shall investigate two simple specifications.

(a) The truncated normal model

If ε_n has a $N(o, \sigma^2)$ distribution truncated from below at $-\beta' x_n$, then c_n is strictly positive and has a conditional p.f.d.:

$$(4) \quad \text{p.d.f.}\ (c_n|x_n) = \frac{\sigma^{-1}\varphi\left(\frac{c_n - \beta' x_n}{\sigma}\right)}{\Phi\left(\frac{\beta' x_n}{\sigma}\right)},$$

where $\varphi(.)$ and $\Phi(.)$ are the p.d.f. and c.d.f. of the N(0,1) distribution. This has a mean function:

$$(5) \quad E(c_n|x_n) = \beta' x_n + \sigma \lambda^*(\beta' x_n/\sigma)$$

where $\lambda^*(.) = \varphi(.)/\Phi(.)$ is the complement of the inverse Mills' ratio.

(b) The lognormal model

If ε_n has a displaced lognormal distribution, with displacement parameter $-\beta' x_n$, then $c_n|x_n$ has a lognormal distribution, with p.d.f.:

$$(6) \quad \text{p.d.f.}(c_n|x_n) = \sigma^{-1} c_n^{-1} \varphi\left(\frac{\log c_n - \beta' x_n}{\sigma}\right)$$

In this case, c_n has conditional mean function:

$$(7) \quad E(c_n|x_n) = \exp\{\beta' x_n + \sigma^2/2\}$$

$$(8) \quad = \exp\{\beta^{*\prime} x_n\}.$$

In (8), we have absorbed the term $\sigma^2/2$ into the intercept term in the linear form $\beta' x_n$. Thus β^* is identical to β except for its first element. Note that, since:

$$(9) \quad \log c_n|x_n \sim N(\beta' x_n, \sigma^2),$$

our interpretation of the Engel curve is somewhat different here: we have a logarithmic, rather than linear relationship.

It should be observed that neither the truncated normal nor the lognormal distribution is as flexible as we might like: although both are skewed to the left, neither incorporates a separate parameter controlling the degree of skewness. However, identification difficulties prevent the use of more heavily parameterised distributions.

Type 2: Some economic non-consumers

There are some goods which are not consumed by everybody at current prices. However, this non-consumption is economic in nature: if the price were reduced (or income increased) sufficiently, any non-consumer could be induced to become a consumer of the good. Most goods that are generally thought of as luxuries probably fall into this category: consumer durables, various types of entertainment, etc.

For goods of this kind, the utility maximisation problem (1)—(2) must be solved subject to an additional non-negativity constraint:

$$(10) \quad c \geq 0.$$

For this two-good case, the Kuhn-Tucker conditions imply the following demand function:

$$(11) \quad c_n = \max\{\tilde{c}_n, 0\},$$

where $\tilde{c}_n$ is the solution to (1)—(2) with no non-negativity constraint imposed. If we adopt a normal linear regression model for $\tilde{c}_n$, then we have:

$$(12) \quad \tilde{c}_n|x_n \sim N(\beta' x_n, \sigma^2),$$

and the effect of (11) is to censor this normal distribution from below at zero. Thus (11)—(12) constitute the censored regression or Tobit model (Tobin (1958)), which underlies the most widely — used technique for coping with zero expenditures in cross-section work. The Tobit model implies the following mixed discrete-continuous distribution for c_n:

$$(13) \quad \Pr(c_n = 0|x_n) = 1 - \Phi(\beta' x_n/\sigma)$$

$$(14) \quad \text{p.d.f.}(c_n|x_n) = \sigma^{-1} \varphi\left(\frac{c_n - \beta' x_n}{\sigma}\right)$$

for $c_n > 0$.

This distribution has mean function:

$$(15) \quad E(c_n|x_n) = \beta' x_n \, \Phi(\beta' x_n/\sigma) + \sigma\varphi(\beta' x_n/\sigma).$$

Under this interpretation of the Tobit model, use of the maximum likelihood Tobit estimator to estimate an Engel curve is only valid under two special assumptions: e_n and c_n are always identical, and non-consumption is always the result of a strictly economic decision.

Type 3: Conscientious abstention

Most vegetarians do not abstain from meat because it is too expensive, or because they are too poor to afford it. Similarly, a large reduction in the price of tobacco will induce very few non-smokers to adopt the habit. The same

applies to many non-consumers of alcoholic drink.

In all of these cases, non-consumption is the result of a conscientious rather than economic decision: it reflects the fact that the population can be divided into distinct groups of abstainers and non-abstainers, characterised by essentially different preferences. Cragg (1971) proposed a statistical structure, known as the double-hurdle model, which can be interpreted as incorporating this distinction (although Cragg was not very specific about his own interpretation of the double-hurdle model, and as a result there is some ambiguity about the way it is specified in practice: see Atkinson, Gomulka and Stern (1984)).

Suppose that abstainers (non-smokers, say) have no use for the good c: if given a free supply of it, they will simply throw it away. Thus, their preferences are represented by a utility function of the form $u^*(C;\theta_n,\varepsilon_n)$. Non-abstainers, however, have preferences representable by the full utility function (1), and we assume also that all non-abstainers consume a positive amount of the good: smokers cannot do without tobacco, no matter how expensive it becomes.

The simplest way of modelling the distinction between abstainers and non-abstainers is to use a probit mechanism, assumed independent of ε_n. Thus, define an unobservable indicator, v_n, which is such that:

$$(16)\quad v_n|\theta_n \sim N(\gamma'\theta_n,\ 1).$$

Then c_n is generated as follows:

If $v_n > 0$, household n is a consumer, and:
c_n is drawn from a conditional truncated normal (4) or lognormal distribution, (6)

If $v_n \leq 0$, household n is an abstainer, and:
$c_n = 0$.

The truncated normal and lognormal versions of this model are Cragg's (1971) models (9) and (11) respectively.

In each case, c_n has a mixed discrete-continuous distribution:

(a) Truncated normal case

$$(17)\quad \Pr(c_n = 0|x_n) = \Pr(v_n \leq 0|x_n) = 1 - \Phi(\gamma'\theta_n)$$

$$(18)\quad \text{p.d.f.}(c_n|x_n) = \sigma^{-1}\varphi\left(\frac{c_n - \beta'x_n}{\sigma}\right)\frac{\Phi(\gamma'\theta_n)}{\Phi(\beta'x_n/\sigma)} \quad \text{for } c_n > 0.$$

This has mean function:

$$(19)\quad E(c_n|x_n) = \Phi(\gamma'\theta_n)\,[\beta'x_n + \sigma\,\lambda^*(\beta'x_n/\sigma)].$$

This is a rather complicated expression. However, note that if we can discard all households for whom $c_n = 0$, it simplifies to:

$$(20)\quad E(c_n|c_n > 0,\ x_n) = \beta'x_n + \sigma\,\lambda^*(\beta'x_n/\sigma)$$

(b) Lognormal case

$$(21)\quad \Pr(c_n = 0|x_n) = 1 - \Phi(\gamma'\theta_n)$$

$$(22)\quad \text{p.d.f.}(c_n|x_n) = \sigma^{-1}c_n^{-1}\varphi\left(\frac{\log c_n - \beta'x_n}{\sigma}\right)\frac{\Phi(\gamma'\theta_n)}{\Phi(\beta'x_n/\sigma)}$$

This has mean function:

$$(23)\quad E(c_n|x_n) = \Phi(\gamma'\theta_n)\,\exp\{\beta^{*\prime}x_n\},$$

where $\beta^{*\prime}x_n$ is again $\beta'x_n + \sigma^2/2$. If it is possible to discard non-consumers, this becomes:

$$(24)\quad E(c_n|c_n > 0,\ x_n) = \exp\{\beta^{*\prime}x_n\}.$$

Type 4: Conscientious abstention and economic non-consumption

For some goods there may be a mixture of reasons for non-consumption: some households may be conscientious abstainers, and others may not consume because the good is currently too expensive. There is perhaps a case for classifying both tobacco and alcohol as type 4 rather than type 3 goods.

Again, a double-hurdle model is appropriate here, but with a Tobit mechanism generating zeros for some potential consumers. Thus, with $v_n|\theta_n$ distributed as $N(\gamma'\theta_n,1)$:

If $v_n > 0$, household n is potential consumer, and:

$$c_n = 0 \text{ if } \tilde{c}_n \leq 0$$
$$c_n = \tilde{c}_n \text{ if } \tilde{c}_n > 0, \text{ where } \tilde{c}_n | x_n \sim N(\beta' x_n, \sigma^2)$$

If $v_n \leq 0$, household n is an abstainer, and:

$$c_n = 0.$$

This leads to a distribution for c_n of the form:

$$(25) \quad \Pr(c_n = 0|x_n) = 1 - \Pr(c_n > 0|x_n)$$
$$= 1 - \Pr(v_n > 0|z_n) \Pr(\tilde{c}_n > 0|x_n)$$
$$= 1 - \Phi(\gamma'\theta_n)\, \Phi(\beta' x_n/\sigma)$$

$$(26) \quad \text{p.d.f.}\ (c_n|x_n) = \Phi(\gamma'\theta_n)\, \sigma^{-1} \varphi\left(\frac{c_n - \beta' x_n}{\sigma}\right),$$

which has a mean function:

$$(27) \quad E(c_n|x_n) = \Phi(\gamma'\theta_n)\, \Phi(\beta' x_n/\sigma)$$
$$[\beta' x_n + \sigma\lambda^*(\beta' x_n/\sigma)].$$

In this case, there is no simplification to be gained by conditioning on the event $c_n > 0$.

3. *The relationship between purchases and consumption*

The theory of consumer demand is conducted in terms of smooth planned rates of flow of consumption services. What we observe in cross-section data is an aggregate, for each household, of a number of individual purchases of varying amounts over a short observation period. Our task is to provide a statistical mechanism relating these two quantities.

Our interpretation of the fundamental consumption variable, c_n, as a long-term average rate of expenditure, together with our assumption that c_n is predetermined when individual expenditures are made, implies the following fundamental identity

$$(28) \quad E(e_n|c_n, z_n) = c_n,$$

where z_n is a vector of (observable) variables relevant to the determination of purchasing frequency. Define $P(c_n, z_n)$ as follows:

$$(29) \quad P(c_n, z_n) = \Pr(\text{one or more purchases occur during the observation period } |c_n, z_n), \qquad c_n > 0$$

$$(30) \quad P(0, z_n) = 0$$

Using this definition of $P(c_n, z_n)$, we have the following identity:

$$(31) \quad E(e_n|c_n, z_n) = 0 \qquad \text{for } c_n = 0$$

$$(32) \quad = P(c_n, z_n)\, E(e_n|e_n > 0,\ c_n, z_n) \qquad \text{for } c_n > 0.$$

Equations (28), (31) and (32) imply:

$$(33) \quad E(e_n|e_n > 0,\ c_n, z_n) = 0 \qquad \text{for } c_n = 0$$

$$(34) \quad = \frac{c_n}{P(c_n, z_n)} \qquad \text{for } c_n > 0.$$

Equations (33)—(34) constitute an important result, since they give a relationship between the mean function of observed expenditure and the unobserved variable to which our theory relates. All that is required to make this operational is a specification for the purchasing probability, $P(c_n, z_n)$.

A simple example will serve to clarify (33)—(34). Suppose a good is bought regularly once every four weeks, in a quantity q. If the survey lasts one week, there is a probability of observing a purchase, $P(c,z) = 1/4$. The corresponding underlying average rate of consumption is obviously $c = q/4$, and yet any observed positive expenditure will be of size $e = c/P = (q/4)/(1/4) = q$. Thus, the important implications of the distinction between expenditures and consumption are:

(i) observed zeros are not necessarily the result of non-consumption;
(ii) even when expenditure is positive, it is generally a poor measure of consumption.

Deaton and Irish (1984) appear to have been the first to appreciate the importance of identity (33)—(34) and to exploit it in the construction of an applied model. Since then, it has been used quite extensively. Table 1 classifies the existing applications.

Table 1. Applied models based on (33)—(34).

Author	$P(c_n,z_n)$	Model for c_n	Relation between e_n and $c_n/P(c_n,z_n)$
Deaton and Irish (1984)	constant parameter	Tobit	Equality
Kay, Keen and Morris (1984)	constant parameter	Linear regression	Additive error
Keen (1986)	constant parameter	Linear regression	Additive error
Blundell and Meghir (1987)	$\Phi(\alpha' z_n)$ where $z_n = x_n$	Linear regression, lognormal	Additive normal error, multiplicative lognormal error

There are two questions to be decided in specifying a model of expenditures: the nature of the conditional purchase probability, $P(c_n,z_n)$, and the relationship between e_n and $c_n/P(c_n,z_n)$ when e_n is positive.

3.1. *The purchase probability*

How should the function $P(c_n,z_n)$ be specified? As table 1 shows, existing applications have taken a very simple approach. Deaton and Irish (1984), Kay, Keen and Morris (1984) and Keen (1986) all specify $P(c_n,z_n)$ as a constant parameter, to be estimated jointly with the parameters of the consumption model.

This has not proved very successful. Deaton and Irish, working with a Tobit model for c_n, and using 1973—4 Family Expenditure Survey (FES) data, estimate this constant probability, P, as almost exactly unity for tobacco, (implying that their P-Tobit model is identical to a simple Tobit model) and they fail to achieve an estimate at all for two other goods: alcoholic drink and durables. In these latter two cases, the log-likelihood function turns out to be unbounded as P increases above unity, and a Lagrange Multiplier test of the restriction $P = 1$ rejects it in favour of the absurd alternative hypothesis $P > 1$. Thus, Deaton and Irish are left in the unsatisfactory position of having a model that provides significant evidence against the conventional Tobit model, but which is untenable in itself.

There are two obvious shortcomings of the Deaton-Irish model. One is that it is unreasonable to treat $P(c_n,z_n)$ as a fixed parameter, and the other is that the Tobit model for c_n classifies all three goods as type 2: with non-consumption possible, and then interpretated as arising from an »economic» corner solution.

So far, little progress has been made with these shortcomings in applied work. Blundell and Meghir (1987) do present estimates based on alternative models of c_n for a different good, clothing, and they estimate a model with $P(c_n,z_n)$ specified as a simple probit form, with explanatory variables taken to be almost identical to those used in the Engel curve itself, x_n. However, this is still restrictive.

Depending on the good concerned, there are several of explanatory variables that are likely to be important in modelling $P(c_n,z_n)$, including the following:

(i) Consumption It is obvious that the rate of consumption, c_n must play a role in determining the purchase probability. If, as in all previous work, $P(c_n,z_n)$ is taken to be independent of c_n, then this carries the absurd implication that we are just as likely to observe a purchase for a household that consumes hardly any of the good as for a heavy consumer. Given that most goods cannot be bought in arbitrarily small units, this is an impossibility. Note that we cannot simply include the determinants, x_n, of c_n rather than c_n itself, since this fails to take correct account of the random component of c_n.

(ii) Survey duration The longer a household is observed, the greater is our chance of observing a purchase (provided $c_n \neq 0$). For a given survey, duration is usually fixed, and this then only a consideration if we are to make use of evidence from surveys based on observation periods of different lengths. The availability of such surveys would give a valu-

able source of evidence on the nature of purchasing behaviour.

(iii) Household Composition Households with a relatively large number of adult members (particularly members who are unoccupied or retired) can be expected to make more purchases of any given good in a fixed period, simply because it has more collective opportunity to do so.

(iv) Durable ownership Households that own cars and freezers have an opportunity to exploit economics of scale in purchasing, since they are able to transport and store larger quantities of goods (foods in particular). These larger quantities imply less frequent purchases, and thus a smaller purchase probability, provided the survey duration is not larger than the re-stocking interval.

If we aim to build a full structural model of the observable, e_n, it is necessary to choose a convenient functional form for $P(c,z)$. The main restrictions on the choice are:

(i) $0 \leq P(c,z) \leq 1$ all $c \geq 0$, z

(ii) $\lim_{c \to 0} P(c,z) = 0$ all z

(iii) $P(c,z)$ is monotonically increasing in c, for all z

(iv) $\lim_{c \to \infty} P(c,z) = 1$ all z

Properties (i) and (ii) are uncontentious. Strictly speaking, there is no necessity for property (iii) to hold: we can envisage situations in which an increase in consumption leads to a switch to a new mode of behaviour involving storage of goods and less frequent purchasing. However, most such counter-examples are inconvincing or require a concomitant change in other relevant variables, z. Property (iv) is not particularly compelling, since it implies that very large consumers devote all of their time to making purchases, rather than buying in suitable bulk. However, property (iv) is probably innocuous as an approximation, and we retain it for convenience. It could be relaxed without difficulty.

There are many plausible choices for $P(c,z)$. We could, for instance, use any convenient c.d.f. such as the normal, with c entered in log form:

$$(35) \quad P(c,z) = \Phi(\alpha_0 + \alpha_1' z + \alpha_2 \log c).$$

This is similar in spirit, but not in its implications, to the model of Blundell and Meghir (1987), but is rather a complicated form to use, when $\alpha_2 \neq 0$. An alternative specification based on the logistic distribution is:

$$(36) \quad P(c,z) = \frac{c^{\alpha_2}}{\alpha_0 + \alpha_1' z + c^{\alpha_2}},$$

Pudney (1985) discussed a number of more elaborate specifications.

3.2. *The relation between e_n and $c_n/P(c_n,z_n)$*

Behaviour is rarely regular. Even if the underlying rate of consumption is unchanged, successive purchases are likely to be of different sizes, unless there is some technical reason, such as indivisibility, dictating otherwise. Thus, it seems wise to assume that e_n is not exactly equal to $c_n/P(c_n,z_n)$, but fluctuates randomly around this values. Since $c_n/P(c_n,z_n)$ is the mean of e_n conditional on the event $e_n > 0$, we must specify a distribution for $e_n | e_n > 0$, c_n, z_n with support only on the positive real line. The practice of adding a regression-style normal disturbance (see Kay, Keen and Morris (1984) and Keen (1986)), is not therefore strictly valid.

Again, there are many convenient possibilities for the distribution of $e_n | e_n > 0$, c_n, z_n; we might use a suitably truncated normal, or alternatively a lognormal distribution parameterised to have mean $c_n/P(c_n,z_n)$, implying that $\log e_n$ has a conditional $N(\mu_n, \sigma_e^2)$ distribution, where $\mu_n = \log\{c_n/P(c_n,z_n)\} - \sigma_e^2/2$.

An example of a fully-specified model, involving both a consumption-dependent purchase probability and a random element in observed expenditure, is examined in section 5 of the paper.

4. *Nonlinear least-squares estimation*

Before specifying a particular generalisation of the Deaton-Irish model in section 5, we consider the possibility of constructing a consistent (albeit inefficient) estimator which is limited information in the sense that it does not require any specific assumption about the nature of $P_n = P(c_n,z_n)$. That this is possible

is a remarkable feature of the P-Tobit framework.

Consider the expected value of e_n, conditional on c_n and z_n:

$$(37)\quad E(e_n|c_n,z_n) = P(c_n,z_n)\,E(e_n|e_n>0,\ c_n,z_n) + [1-P(c_n,z_n)] \times 0 = c_n,$$

using (34). Thus the conditional expectation of e_n is independent of the nature of P_n, and hence any technique founded on the expected value of e_n requires no assumptions about the purchase probability.

Since c_n is unobserved, take the further expectation with respect to c_n:

$$(38)\quad E(e_n|x_n,z_n) = E(c_n|x_n,z_n) = E(c_n|x_n),$$

since z_n is not involved in the determination of consumption. Equality (38) can be written as a regression equation:

$$(39)\quad e_n = g(\beta' x_n,\ \sigma) + \upsilon_n,$$

where $\upsilon_n = e_n - E(e_n|x_n)$ is a disturbance term with zero mean conditional on x_n, and where $g(\beta' x_n,\ \sigma)$ is the mean function of c_n. Depending on the good concerned, we might use expressions (5), (8), (15), (19), (23) or (27) as this mean function.

Equation (39) is nonlinear in the parameters β and σ; any further parameters that might be involved in the function $P(c_n,z_n)$ or the p.d.f. of $e_n|c_n/P(c_n,z_n)$ do not appear in (39) and hence cannot be estimated. However, β and σ determine the Engel curve completely, and we are often not particularly interested in purchasing patterns, so this may not be a serious drawback. More serious is the possibly severe nonlinearity of $g(\beta' x_n,\ \sigma)$ and the fact that nonlinear least-squares estimation is generally inefficient. There may also be difficulty in identifying both β and σ from the mean of e_n alone.

A further potential difficulty is that the random errors, υ_n, are heteroscedastic, with the precise form of the heteroscedasticity dependent upon the nature of $P(c_n,z_n)$. Therefore, to preserve the limited-information spirit of this approach, we must ignore the heteroscedasticity problem during estimation, but take proper account of it when computing the asymptotic covariance matrix.

Thus, the nonlinear least-squares estimator we propose solves the following problem:

$$(40)\quad \min_{\beta,\sigma} \sum_{n=1}^{N} [e_n - g(\beta' x_n,\ \sigma)]^2$$

where N is the sample size. The asymptotic covariance matrix of this estimator can be consistently estimated in the presence of heteroscedasticity of unknown form by the following expression, due to White (1980):

$$(41)\quad \text{a. cov}(\hat{\beta},\ \hat{\sigma}) = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where:

$$(42)\quad \hat{A} = \left[\sum_{n=1}^{N} \hat{g}_n^{\delta}\hat{g}_n^{\delta'}\right]$$

$$(43)\quad \hat{B} = \left[\sum_{n=1}^{N} \hat{\upsilon}_n^2\hat{g}_n^{\delta}\hat{g}_n^{\delta'}\right],$$

where $\hat{g}_n^{\delta}$ is the vector of derivatives of $g(\hat{\beta}' x_n,\ \hat{\sigma})$ with respect to $\hat{\delta}' = (\hat{\beta}',\ \hat{\sigma})$ and $\hat{\upsilon}_n$ is $e_n - g(\hat{\beta}' x_n,\ \hat{\sigma})$.

Although this nonlinear least squares estimator is remarkably simple, it is often beset by identification difficulties. In the lognormal model (expressions (7), (23) or (24)), the intercept and variance parameter are not separately identifiable. In the censored and truncated models (expressions (5), (15), (19), (20) and (27)), residual sum of squares function is often found to have a very poorly defined minimum (see Pudney (1987) for some examples), and this results in severe computational difficulties and imprecise estimates. Wales and Woodland (1980) report similar difficulties in a rather different context.

5. *Fully-specified models*

Although the nonlinear least-squares approach of section 4 is simple and undemanding in terms of model specification, it is inefficient, and frequently suffers from severe identification problems. It is therefore of some interest to develop the efficient maximum likelihood (ML) estimator for a fully-specified

model, and to determine whether or not it has any practical advantages.

In the general case, c_n will have a distribution (conditional on x_n) that has a discrete probability mass at zero, $\pi(x_n)$ say, and a continuous density component defined on $(0, \infty)$, $f_c(c_n|x_n)$ say. Conditional on $c_n/P(c_n,z_n)$, and the event $e_n>0$, e_n will have a truncated density function $f_e(e_n|c_n/P(c_n,z_n))$. The resulting mixed discrete-continuous distribution for e_n is:

$$(44)\quad \Pr(e_n=0|x_n,z_n)=\pi(x_n)+\int_0^\infty [1-P(c,z_n)]\, f_c(c|x_n)\, dc$$

$$(45)\quad \text{pdf}(e_n|x_n,z_n)=\int_0^\infty P(c,z_n)\, f_e(e_n|c/P(c,z_n))\, f_c(c|x_n)\, dc.$$

A full model is then specified by choosing suitable functional forms for $\pi(\cdot)$, $f_c(\cdot)$, $f_e(\cdot)$ and $P(\cdot)$, and maximising numerically the log-likelihood function:

$$(46)\quad L(\theta)=\sum_{e_n=0}\log \Pr(e_n=0|x_n,z_n)+\sum_{e_n>0}\log \text{pdf}(e_n|x_n,z_n),$$

where θ represents all the underlying parameters.

The problem facing the applied worker is that expressions (44) and (45) are difficult from the computational point of view, unless great care is taken with the specification of functional forms.

In our applied examples, relating to clothing and three foods, we shall adopt a simple model that assumes that there is no non-consumption, so that:

$$(47)\quad \pi(x_n)=0$$

We use the lognormal model for consumption:

$$(48)\quad f_c(c_n|x_n)=c_n^{-1}\sigma^{-1}\varphi\left[\frac{\log c_n-\beta' x_n}{\sigma}\right]$$

and a lognormal distribution for e_n:

$$(49)\quad f_e(e_n|c_n/P(c_n,z_n))=e_n^{-1}\sigma_e^{-1}\varphi\left[\frac{\log e_n-\log c_n+\log P(c_n,z_n)+\sigma_e^2/2}{\sigma_e}\right]$$

This form satisfies $E(e_n|e_n>0,\ c_n/P(c_n,z_n))=c_n/P(c_n,z_n)$. For the purchase probability, we use a probit formulation:

$$(50)\quad P(c_n,z_n)=\Phi(\alpha_0+\alpha_1' z_n+\alpha_2\log c_n).$$

Expression (50) is particularly convenient because it permits the discrete probability (44) to be expressed in terms of $\Phi(\cdot)$, rather than as an integral:

$$(51)\quad \Pr(e_n=0|x_n,z_n)=1-\Phi\left[\frac{\alpha_0+\alpha_1' z_n+\alpha_2\beta' x_n}{\sqrt{1+\alpha_2^2\sigma^2}}\right]$$

Note that this allows β to be estimated up to scale (except for the coefficients of any variables that appear in z_n) by probit analysis.

The continuous part of the distribution is more complicated. It is defined by the following integral, which is not expressible in closed form:

$$(52)\quad \text{pdf}(e_n|x_n,z_n)=\frac{1}{\sigma\sigma_e e_n}\int_0^\infty \frac{P(c,z_n)}{c}\varphi\left[\frac{\log e_n-\log c+\log P(c,z_n)+\sigma_e^2/2}{\sigma_e}\right]\varphi\left[\frac{\log c-\beta' x_n}{\sigma}\right]dc$$

Numerical integration must be used to evaluate these density components. Experience suggests that this is best done by means of a Gauss-Hermite quadrature algorithm (see Waldeman (1985) for an evaluation of Gaussian quadrature in a different context), and this requires (52) to be written in the form:

$$(53)\quad \text{pdf}(e_n|x_n,z_n)=\frac{1}{\sigma_e e_n}\int_{-\infty}^{\infty}\psi_n(\xi)\, e^{-\xi^2/2}d\xi$$

where:

$$(54)\quad \psi_n(\xi) = P_n(\xi)\,\varphi\left[\frac{\log e_n - \log c_n(\xi) + \log P_n(\xi) + \sigma_e^2/2}{\sigma_e}\right],$$

$$(55)\quad c_n(\xi) = \exp\{\sigma\xi + \beta' x_n\}$$

and

$$(56)\quad P_n(\xi) = \Phi(\alpha_0 + \alpha_1' z_n + \alpha_2 \log c_n(\xi)).$$

6. *Some empirical results*

A previous paper (Pudney (1987)) described the preliminary results of an application of the least squares (LS) and maximum likelihood (ML) estimators to four commodities, using a variety of specifications for the underlying Engel curve model. Considerable identification and computational difficulties were encountered with the LS estimator for all except the lognormal model, and the paper presented only a very limited experiement with ML estimation.

In this section, we confine attention to the lognormal model, and apply it to four commodities drawn from two UK household surveys conducted in 1983. The first of these is the Family Expenditure Survey (FES), based on a two-week observation period. From the FES we estimate an Engel curve for clothing expenditure. Two alternative forms of the dependent variable are used: (i) the ratio of household clothing expenditure to household expenditure on all goods; (ii) the ratio of household expenditure on clothing to the household's normal income. Tables 2 and 3 contain summary information on these vari-

Table 2. Dependent variables.

	Clothing	Beef	Lamb	Pork
Survey source	Family Expenditure Survey 1983		National Food Survey 1983	
Survey duration	2 weeks		1 week	
Number of observations	6971		7162	
Percentage of zero expenditures	22.75	44.53	70.22	67.13
Mean budget share (full sample)	Total expenditure definition: 0.061 Normal income definition: 0.066	0.064	0.029	0.027
Mean budget share (truncated sample)	Total expenditure definition: 0.078 Normal income definition: 0.085	0.117	0.096	0.081
Maximum budget share	Total expenditure definition: 0.562 Normal income definition: 1.814	0.714	0.678	0.782

ables and the variables used as explanatory factors. Note that the assumed form for the Engel curve is quadratic in logarithms (although the squared term is found insignificant and omitted for the model based on the income definition).

Our second data source is the National Food Survey (NFS). This records purchases of foods bought for consumption in the home, and is based on a one-week survey period. From this, we estimate Engel curves for the three main categories of carcase meat: beef

Table 3. Explanatory variables.

Variable	Symbol	Mean in FES	Mean in NFS
log per capita household expenditure	LTE	10.829	—
LTE squared	LTE2	117.582	—
log per capita normal income	LNI	10.912	—
log per capita food expenditure	LFE	—	7.388
LFE squared	LFE2	—	54.879
dummy: household has no workers	NWRK	0.271	—
number of adults	AD	1.922	2.024
number of children under 2 years	KIDS<2	0.080	0.035
number of children aged 2—5	KIDS2—5	0.123	0.161
number of children over 5 years	KIDS>5	0.533	0.546
dummy: head of household aged 25—35	AGE25	0.185	0.180
head of household aged 35—45	AGE35	0.192	0.191
head of household aged 45—60	AGE45	0.234	0.164
head of household aged 60—65	AGE60	0.088	0.169
head of household aged over 65	AGE65	0.251	0.256
dummy: female head of household	FHOH	0.235	—
dummy: household has no children	NOKIDS	0.608	0.600
dummy: unmarried head of household	SINGLE	0.226	0.188
dummy: occupation is professional	PROF	0.223	—
occupation is clerical	CLERK	0.062	0.112
occupation is shop worker	SHOP	0.009	—
occupation is skilled	SKILL	—	0.290
occupation is semi-skilled	SEMI	—	0.072
occupation is unskilled	UNSKILLED	—	0.021
occupation is armed forces	FORCES	0.005	0.061
retired or unoccupied	UNOCC	0.378	0.380
dummy: household has a car	CAR	0.621	—
household has 2 or more cars	2CARS	0.167	—
household has a telephone	TEL	0.773	—
household is owner-occupier	OWNER	0.591	0.598
dummy: household is in Yorkshire	YORKS	0.096	0.102
household is in East Midlands	EMID	0.071	0.075
household is in East Anglia	ANGLIA	0.036	0.021
household is in Greater London	LONDON	0.110	0.124
household is in South East	SE	0.182	0.183
household is in South West	SW	0.071	0.095
household is in Wales	WALES	0.053	0.042
household is in West Midlands	WMID	0.098	0.081
household is in North West	NW	0.119	0.109
household is in Scotland	SCOT	0.087	0.089
household is in N. Ireland	ULSTER	0.019	—
dummy: pre-Christmas period	XMAS1	0.073	—
post-Christmas period	XMAS2	0.087	—
early summer	SUMMER1	0.059	—
late summer	SUMMER2	0.079	—
dummy: household owns a freezer	FREEZER	—	0.602

and veal, mutton and lamb, and pork. In this case, the dependent variables are defined as shares in total food expenditure, and the Engel curve is log-quadratic in total food expenditure per head. Summary statistics are presented in tables 2 and 3.

The lognormal model is certainly appropriate for clothing, since everyone uses clothes. However, it is harder to justify its application to the three meat categories from the NFS. Some people are vegetarians, and it is also conceivable that there might be economic non-consumption, with some poor households unable to afford the more expensive carcase meats. There is certainly scope for non-consumption: in the NFS sample, 26 % of households recorded no purchases of beef and veal, mutton and lamb or pork during the sample week. However, we take the view that neither form of non-consumption is likely to be important for more than a small minority of households, and that fortuitous non-purchasing is a more plausible explanation of most of these cases. If this assumption is true, the lognormal model should provide an adequate approximation to the true consumption relation.

6.1. Clothing

Consider first the model of clothing demand. There is an obvious problem here with the use of total expenditure as the denominator of the budget share and as the principal explanatory variable. Clothing expenditures can be very large in exceptional cases. An expensive coat, for instance, may cost much more than a normal fortnight's entire budget. Indeed, the most extreme case in our sample is of an expenditure 181 % of the household's normal fortnightly income. Purchases of this size cannot be absorbed within the household's usual budget, and must be financed by borrowing or by past savings. This causes observed total expenditure to depart substantially from normal planned expenditure per fortnight, and generates a spurious correlation between expenditure on clothing and total expenditure. Unless one is prepared to accept linear Engel curves (as Keen (1986) does) it is difficult to correct for this, since one is faced with a particularly complicated nonlinear measurement error problem.

In this work, we have attempted to deal with the problem in a rather superficial way, by investigating alternative specifications with normal income replacing total expenditure as the household resources variable. This is not entirely satisfactory, since the model that results is a hybrid Engel curve/consumption function, but it does allow us to give some indication of the seriousness of this problem.

For ML estimation, it is necessary to specify a form for the conditional purchase probability. For clothing, we use the following

$$(57)\quad P(c_n,z_n)=\Phi(\alpha_0+\alpha_{11}AD_n+\alpha_{12}XMAS1_n+\alpha_{13}XMAS2_n+\alpha_{14}SUMMER1_n+\alpha_{15}SUMMER2_n+\alpha_2(\log c_n+\log y_n))$$

where AD, XMAS1, XMAS2, SUMMER1 and SUMMER2 are described in table 3, and y_n is either total expenditure or normal income, (*not* divided by household size). The variable AD is included to reflect the fact that the more (adult) members a household has, the more chance we have of observing one of them making a purchase. Variables XMAS1, XMAS2, SUMMER1 and SUMMER2 pick up seasonal variation in purchasing behaviour due to sales promotions and the Christmas peak. In our model, c_n is defined as a budget share, yet purchasing frequency presumably depends on the gross volume of purchasing, not its share in the overall budget. Hence the last term in (57) is of this form.

Final estimates for the two versions of the clothing model are presented in tables 4 and 5. For the income definition, the quadratic term in LNI proved insignificant for both LS and ML estimates, and the resulting model is therefore of simple double-log form. In both ML cases, the optimisation algorithm attempted to produce a negative value for σ. Thus, when the constraint $\sigma \geq 0$ was imposed, the ML estimate was $\sigma = 0$. The remaining parameter estimates for this constrained model appear in table 4, and the accompanying asymptotic standard errors are calculated under the (incorrect) assumption that σ is set to zero a priori. If the usual ML parameter covariance matrix is calculated for all of the parameters at $\sigma = 0$, the implied standard errors for the unconstrained parameters are very close to those quoted in table 4, and the standard error for σ is very large. Thus, we should

Table 4. Maximum-likelihood and least-squares estimates of two versions of an Engel curve for clothing; 6971 observations from the 1983 UK Family Expenditure Survey.

	Total Expenditure		Normal Income	
	ML	LS	ML	LS
CONSTANT	—50.12774	—47.63000	1.32760	1.09640
	(4.455)	(4.815)	(0.552)	(1.166)
LTE	8.05486	7.62160	—	—
	(0.796)	(0.869)		
LTE2	—0.33862	—0.31980	—	—
	(0.036)	(0.039)		
LNI	—	—	—0.37348	—0.34730
			(0.050)	(0.108)
NWRK	0.03169	—0.04193	0.02682*	—0.22660
	(0.073)	(0.066)	(0.074)	(0.139)
AD	0.23596	0.12793	0.17708*	0.09408
	(0.031)	(0.023)	(0.033)	(0.040)
KIDS<2	0.24774	0.15577	—0.07412	—0.21441
	(0.077)	(0.052)	(0.080)	(0.078)
KIDS2—5	0.25125	0.16167	—0.04308	—0.04058
	(0.059)	(0.043)	(0.059)	(0.096)
KIDS>5	0.26977	0.19186	0.06420	0.04330
	(0.035)	(0.021)	(0.036)	(0.033)
AGE25	—0.14145	—0.12370	—0.23536	—0.43766
	(0.090)	(0.073)	(0.090)	(0.142)
AGE35	—0.21283	—0.17166	—0.28395*	—0.50468
	(0.094)	(0.076)	(0.094)	(0.148)
AGE45	—0.27164	—0.20020	—0.32364*	—0.64858
	(0.089)	(0.073)	(0.090)	(0.143)
AGE60	—0.32600	—0.22200	—0.37767*	—0.59222
	(0.102)	(0.087)	(0.103)	(0.140)
AGE65	—0.36894	—0.23797	—0.63270*	—0.82005
	(0.100)	(0.086)	(0.100)	(0.131)
FHOH	0.41982	0.25394	0.44891	0.33592
	(0.055)	(0.047)	(0.054)	(0.101)
NOKIDS	—0.10073	—0.07953	—0.05907	0.02521
	(0.073)	(0.053)	(0.075)	(0.084)
SINGLE	—0.37958	—0.28369	—0.15288*	0.19952
	(0.071)	(0.062)	(0.071)	(0.137)
PROF	—0.04903	—0.04525	0.01777	0.00827
	(0.049)	(0.038)	(0.050)	(0.063)
CLERK	0.06266	0.06304	0.14300	0.17464
	(0.075)	(0.060)	(0.076)	(0.093)
SHOP	0.29612	0.16794	0.28654	0.15130
	(0.190)	(0.147)	(0.190)	(0.185)
FORCES	0.15460	0.11765	0.24524*	1.11930
	(0.206)	(0.225)	(0.223)	(0.491)
UNOCC	—0.10585	—0.07103	—0.09659	0.05329
	(0.068)	(0.056)	(0.069)	(0.118)
CAR	—0.24542	—0.21624	0.05276*	0.24717
	(0.049)	(0.042)	(0.048)	(0.107)
2CARS	—0.05431	—0.09508	0.14580	0.15216
	(0.050)	(0.039)	(0.052)	(0.083)

Table 4 continued

	Total Expenditure		Normal Income	
	ML	LS	ML	LS
TEL	0.03484 (0.051)	0.01376 (0.047)	0.21864 (0.051)	0.19918 (0.099)
OWNER	—0.11128 (0.043)	—0.09265 (0.037)	—0.05671 (0.044)	—0.04008 (0.076)
YORKS	—0.00527 (0.091)	—0.06466 (0.074)	0.00048* (0.092)	—0.27520 (0.129)
EMID	—0.08094 (0.095)	—0.05759 (0.080)	—0.06009* (0.096)	—0.17341 (0.112)
ANGLIA	—0.10708 (0.115)	—0.08997 (0.096)	—0.11704 (0.117)	—0.34679 (0.172)
LONDON	—0.14140 (0.089)	—0.09965 (0.073)	—0.00487 (0.087)	0.09899 (0.140)
SE	—0.17334 (0.082)	—0.20908 (0.067)	—0.09603 (0.081)	—0.15389 (0.122)
SW	—0.21075 (0.097)	—0.13965 (0.084)	—0.19540 (0.098)	—0.38269 (0.153)
WALES	0.06564 (0.103)	0.06657 (0.087)	0.02629 (0.103)	—0.07280 (0.128)
WMID	0.03069 (0.090)	—0.03911 (0.073)	0.03296* (0.090)	—0.16144 (0.114)
NW	—0.14318 (0.086)	—0.10397 (0.070)	—0.11129 (0.086)	—0.18170 (0.110)
SCOT	0.08079 (0.093)	0.05667 (0.074)	0.10762 (0.092)	0.07393 (0.121)
ULSTER	0.19905 (0.144)	0.15722 (0.110)	0.17224* (0.145)	0.02555 (0.141)
Parameters of the conditional purchase probability				
α_0	—4.49187 (0.157)	—	—5.54456 (0.248)	—
AD	0.22007 (0.030)	—	0.11615 (0.037)	—
XMAS1	0.13656 (0.072)	—	0.17448 (0.073)	—
XMAS2	0.13203 (0.067)	—	0.07840 (0.065)	—
SUMMER1	—0.17532 (0.066)	—	—0.17419 (0.066)	—
SUMMER2	0.01472 (0.065)	—	0.01469 (0.065)	—
α_2	0.55331 (0.020)	—	0.67508 (0.032)	—
σ_ϵ	0.81086 (0.008)	—	0.75109 (0.008)	—
log L	—11794.51	—	—12327.67	—

* indicates a difference from the corresponding LS estimate significant at the 5 % level

not draw the conclusion that the underlying Engel curve is an exact relation, but rather that σ is not well identified in this example. However, note that $\sigma = 0$ is a very convenient restriction, since the need for numerical integration is removed, and the model collapses to that of Blundell and Meghir (1987) (although with specific restrictions imposed on the coefficients in the purchase probability expression).

The two versions of the model differ considerably in their implied elasticities of demand with respect to household resources. At the sample mean, these are in the region of 1.7 for the total expenditure definition and 0.6 for the normal income definition. Even allowing for an elasticity of total expenditure with respect to normal income of less than unity, this discrepancy is very large, and suggests that the simultaneity problem may be severe.

In other respects, the two models generate very similar results. The estimated family composition effects imply that clothing expenditure per head falls as household size increases with resources held constant, but that if resources are increased in proportion, there is a substantial increase in spending: roughly 10 %—20 % for an additional family member. There are considerable negative age effects, with expenditure declining monotonically as the age of the head of household — increases: Households with a female head are estimated to spend approximately 30 %—50 % more on clothes than an identical household with a male head, and there is some evidence that households with the commitments of onwer-occupation and car ownership spend up to 20 % less. There are major differences in behaviour between occupational classes, but less pronounced regional differences.

The ML estimates of the parameters of the conditional purchase probability are well determined and seem plausible. The typical probability is around 77 %. This increases to approximately 80 % in the Christmas period, but falls to 72 % in early summer. An additional adult in the household increases the probability to 83 % and 80 % respectively in the total expenditure and normal income versions of the model.

Is there any evidence of misspecification here? One approach is to look for significant differences between the LS and ML coefficient estimates, since these can be expected to differ if, for instance, the purchase frequency model is incorrectly specified. However, we encountered problems in trying to implement this Hausman specification test. Despite the fact that the ML estimator is asymptotically efficient, the difference between the computed LS and ML covariance matrices is not positive definite (or even positive semi-definite), and so the test statistic is undefined. This appears to arise from the different types of finite-sample approximation used for the two covariance matrices. Instead of attempting to compute the statistic using a generalised inverse, we have merely indicated coefficients for which a simple asymptotic t-test indicates a significant difference between the ML and LS coefficients. These significant discrepancies are confined to the normal income definition, where they are quite numerous.

A second source of diagnostic information concerns the ability of the model to explain the frequency of zeros and mean budget share in the sample. The model's predictions of these are:

Table 5. Properties of the estimated clothing models.

	Total Expenditure		Normal Income	
	ML	LS	ML	LS
Elasticity at y_{min}*	3.03	6.30	0.63	0.65
Elasticity at mean of log y*	1.72	1.64	0.63	0.65
Elasticity at y_{max}*	—0.34	—0.18	0.63	0.65
Predicted probability of zeros (actual proportion in parentheses)	0.2249 (0.2275)	—	0.2268 (0.2275)	—
Predicted mean consumption (mean budget share in parentheses)	0.07728 (0.06062)	—	0.08325 (0.06578)	—

* y represents total expenditure or normal income; y_{min} and y_{max} are the smallest and greatest values of y in the sample.

Table 6. Maximum likelihood and least-squares estimates of Engel curves for beef, mutton and lamb and pork; 7162 observations from the 1983 UK National Food Survey.

	Beef		Mutton and lamb		Pork	
	ML	LS	ML	LS	ML	LS
CONSTANT	—13.53021	—12.57500	—23.67681	—18.27800	—16.67365	—13.82900
	(2.886)	(3.531)	(4.847)	(5.609)	(3.771)	(5.026)
LFE	2.31792	2.07460	4.62412	3.15820	3.12296	2.38380
	(0.755)	(0.951)	(1.263)	(1.500)	(1.002)	(1.330)
LFE2	—0.12209	—0.10533	—0.27596	—0.17786	—0.19065	—0.14174
	(0.050)	(0.064)	(0.083)	(0.100)	(0.067)	(0.088)
AD	0.04232	0.02967	0.02408	0.07206	0.05228	0.06395
	(0.026)	(0.024)	(0.038)	(0.048)	(0.035)	(0.034)
KIDS < 2	—0.08367	—0.05084	0.16395	0.17388	—0.10209	—0.07782
	(0.089)	(0.093)	(0.144)	(0.149)	(0.131)	(0.125)
KIDS2—5	—0.00140	0.01967	0.07228	0.14285	—0.15428	—0.16176
	(0.046)	(0.044)	(0.067)	(0.101)	(0.062)	(0.069)
KIDS > 5	0.01624*	0.00344	0.04770	0.08132	—0.02801	—0.03775
	(0.027)	(0.027)	(0.040)	(0.054)	(0.036)	(0.042)
AGE25	0.08407	0.07998	0.04637	0.09342	—0.17114	—0.18465
	(0.061)	(0.059)	(0.098)	(0.114)	(0.083)	(0.089)
AGE35	—0.37028	—0.39151	—0.14254	—0.08562	—0.04424	—0.02767
	(0.062)	(0.061)	(0.090)	(0.094)	(0.086)	(0.083)
AGE45	0.10089	0.10015	0.27080	0.25926	0.17245	0.24112
	(0.100)	(0.102)	(0.168)	(0.180)	(0.141)	(0.140)
AGE60	0.07407	0.10459	0.36998	0.36786	0.17350	0.21372
	(0.102)	(0.106)	(0.172)	(0.184)	(0.145)	(0.143)
AGE65	0.14827	0.16423	0.56925	0.56149	0.23327	0.30592
	(0.101)	(0.106)	(0.173)	(0.184)	(0.145)	(0.143)
NOKIDS	0.25952	0.27743	0.64188*	0.59245	0.25055*	0.34600
	(0.101)	(0.105)	(0.177)	(0.177)	(0.145)	(0.145)
SINGLE	0.31048	0.32389	0.98176	0.90969	0.20078	0.26705
	(0.107)	(0.109)	(0.183)	(0.180)	(0.149)	(0.155)
CLERK	0.11578*	0.17131	0.04741	0.07520	0.02568	0.07468
	(0.071)	(0.076)	(0.111)	(0.132)	(0.099)	(0.106)
SKILL	0.12355	0.15629	0.05672	0.03788	0.08376	0.10068
	(0.065)	(0.068)	(0.100)	(0.124)	(0.091)	(0.094)
SEMI	0.05648*	0.12137	0.12009	0.15504	0.07055	0.10820
	(0.081)	(0.083)	(0.124)	(0.144)	(0.112)	(0.115)
UNSKILLED	—0.15890	—0.09780	—0.00853*	0.19308	0.08222	0.21158
	(0.130)	(0.138)	(0.188)	(0.204)	(0.163)	(0.154)
FORCES	—0.13521*	—0.04453	—0.10643	0.02171	0.00014	0.02239
	(0.091)	(0.094)	(0.136)	(0.207)	(0.121)	(0.125)
UNOCC	0.00069	0.05690	0.06613	0.14881	0.02301	0.08718
	(0.077)	(0.076)	(0.118)	(0.130)	(0.102)	(0.112)
OWNER	0.03785	0.05080	—0.00784	—0.00195	—0.00389*	—0.02918
	(0.034)	(0.036)	(0.053)	(0.055)	(0.049)	(0.049)
YORKS	0.09309	0.10600	0.12361	0.14095	0.08666	0.10071
	(0.071)	(0.065)	(0.115)	(0.117)	(0.099)	(0.102)
EMID	—0.17859	—0.17150	—0.11043	0.01457	0.17410	0.16044
	(0.078)	(0.078)	(0.125)	(0.158)	(0.105)	(0.108)

Table 6 continued

	Beef		Mutton and lamb		Pork	
	ML	LS	ML	LS	ML	LS
ANGLIA	—0.05846	—0.08966	—0.41072	—0.17503	0.09274	0.08695
	(0.125)	(0.102)	(0.217)	(0.262)	(0.161)	(0.152)
LONDON	—0.22029	—0.24553	0.35938	0.38974	—0.01017	0.01716
	(0.070)	(0.069)	(0.107)	(0.111)	(0.096)	(0.101)
SE	—0.26848	—0.24719	0.15013	0.16748	—0.01244	0.01606
	(0.067)	(0.063)	(0.104)	(0.108)	(0.092)	(0.093)
SW	—0.19472	—0.20355	0.08148	0.08858	0.05577	0.10092
	(0.075)	(0.072)	(0.121)	(0.121)	(0.102)	(0.110)
WALES	—0.20747	—0.19102	0.29453	0.30892	—0.06165*	0.03067
	(0.093)	(0.093)	(0.145)	(0.137)	(0.129)	(0.131)
WMID	—0.14427	—0.16665	0.37754	0.36023	0.16097	0.23679
	(0.076)	(0.072)	(0.116)	(0.116)	(0.099)	(0.109)
NW	—0.16041	—0.21440	0.39786	0.38828	0.07880	0.09750
	(0.072)	(0.068)	(0.112)	(0.107)	(0.099)	(0.100)
SCOT	0.19417	0.18778	—0.32661	—0.35434	—0.35679	—0.41973
	(0.074)	(0.065)	(0.125)	(0.148)	(0.109)	(0.114)
Parameters of the conditional purchase probability						
α_0	—3.26607	—	—2.78279	—	—2.63515	—
	(0.093)		(0.072)		(0.067)	
AD	—0.05893	—	—0.11252	—	—0.06102	—
	(0.018)		(0.015)		(0.015)	
FREEZER	—0.20909	—	—0.14927	—	—0.14926	—
	(0.025)		(0.023)		(0.020)	
α_2	0.67862	—	0.55229	—	0.52868	—
	(0.020)		(0.018)		(0.018)	
σ	0.49468	—	0.54365	—	0.62127	—
	(0.083)		(0.116)		(0.081)	
σ_ε	0.63000	—	0.63767	—	0.52400	—
	(0.025)		(0.022)		(0.023)	
log L	—8522.3	—	—6240.7	—	—6419.6	—

* indicates a difference from the corresponding LS estimate significant at the 5 % level

$$(58)\quad \hat{\pi} = N^{-1} \sum_{n=1}^{N} \Phi \left(\frac{\hat{\alpha}_0 + \hat{\alpha}_1' z_n + \hat{\alpha}_2 \hat{\beta}' x_n}{\sqrt{1 + \hat{\alpha}_1^2 2\hat{\sigma}^2}} \right)$$

and:

$$(59)\quad \hat{C} = N^{-1} \sum_{n=1}^{N} \exp \{\hat{\beta}' x_n + \hat{\sigma}^2/2\}.$$

Table 5 shows the results. The frequency of zeros is well predicted by both models, but the correspondence between actual and fitted mean budget shares is not very close in either case. It is possible to compute asymptotic t-ratios based on these latter discrepancies, and in both cases, a rejection of the model is indicated. Thus, there remains scope for further development of the clothing model.

6.2. *Expenditure on meats*

The NFS data on meat purchases displays a much higher proportion of zero observations than the FES clothing data. Owing to the rather limited nature of the income information available from the NFS, we use total

Table 7. Properties of the estimated food models.

	Beef and veal		Mutton and lamb		Pork	
	ML	LS	ML	LS	ML	LS
Elasticity at y_{min}*	2.62	2.43	4.02	3.08	3.02	2.52
Elasticity at mean of y*	1.53	1.52	1.56	1.53	1.31	1.29
Elasticity at y_{max}*	1.07	1.13	0.53	0.88	0.60	0.77
Predicted probability of zeros (actual proportion in parentheses)	0.44521 (0.44527)	—	0.70218 (0.70218)	—	0.67089 (0.67132)	—
Predicted mean consumption (mean budget share in parentheses)	0.06632 (0.06511)	—	0.02918 (0.02869)	—	0.02661 (0.02659)	—

* y represents total expenditure or normal income; y_{min} and y_{max} are the smallest and greatest values of y in the sample

recorded food expenditure by the household as the basic resources variable. The other explanatory variables are similar to those used in the clothing model, except that variables NWRK, FHOH, CAR, 2CARS and TEL are not available in the NFS, and the occupational classification is rather different. The model of purchasing frequency is a simple one, with the z_n vector containing two variables reflecting the number of adults in the household and the ownership of a freezer. The unobservable consumption variable is again included in expenditure, rather than budget share, form.

The expenditure elasticities are presented in table 7. For all three meats, these are very high for low-expenditure households, falling considerably as we move to the high end of the expenditure distribution. The family composition effects imply a substantial drop in consumption per head as family size increases, and there are also strong age effects, with consumption being relatively low in the 35—45 age group. Occupational differences do not appear very strong, but there are significant regional shifts in behaviour. The estimated Engel curve thus appears plausible.

For all three goods, the parameters of the purchasing model are well-determined, including both variance parameters, σ^2 and σ_ε^2, which are approximately equal. As we would expect, our estimates imply that the ownership of a freezer reduces the chance of observing a purchase, with the typical purchase probability falling by 3 percentage points for beef and veal and 5 percentage points for mutton and lamb and pork. The one puzzling feature of the results is the negative coefficient of AD in the model of purchasing frequency. These coefficients are all significant, and imply falls of 1—4 percentage points in the purchasing probabilities as an additional adult member is added to the household. A possible explanation of this finding is that it arises from pensioner households that are often too poor to own a car, and which are consequently obliged to make frequent small-scale shopping trips, rather than infrequent motorised expeditions to the supermarket. Further research will clarify this.

There is little evidence of misspecification in tables 6 and 7, and the model appears to be very successful in explaining observed expenditures. There are a few »significant» differences between the ML and LS coefficients, but these appear to arise from differences in the covariance matrix approximations, rather than any major discrepancies in the coefficients themselves. The ML estimates produce a very close fit, both in terms of the mean budget share and the frequency of zero observations.

7. *Conclusion*

In this paper, we have discussed the problem of estimating an Engel curve from short-duration survey data, considering two

estimation methods: a limited-information least-squares estimator, and a maximum likelihood estimator for a fully-specified model.

We have applied these to a simple Engel curve expressed in logarithmic form with an additive normal disturbance, for four expenditure categories drawn from two surveys. Both estimators produce plausible results, and a particularly encouraging finding is that the simple least-squares estimator performs well and appears to be reasonably efficient. Further research is required to determine whether the maximum likelihood estimator is a feasible alternative to least squares in more complex models where the least squares estimator is often found to be unstable.

References

Atkinson, A.B., J. Gomulka and N.H. Stern (1984), »Household expenditure on tobacco 1970—1980: evidence from the Familiy Expenditure Survey», ESRC Programme on Taxation Incentives and the Distribution of Income, discussion paper 57.

Blundell, R.W. and C. Meghir (1987), »Bivariate alternatives to the Tobit model», *Journal of Econometrics* 34 (Annals 1987-1), 179—200.

Cragg, J.G. (1971), »Some statistical models for limited dependent variables with applications to the demand for durable goods», *Econometrica* 39, 829—44.

Deaton, A.S. and M. Irish (1984), »A Statistical model for zero expenditures in household budgets», *Journal of Public Economics* 23, 59—80.

Kay, J.A., M.J. Keen and C.N. Morris (1984), »Estimating Consumption from expenditure data», *Journal of Public Economics* 23, 169—182.

Keen, M. (1986), »Zero expenditures and the estimation of Engel curves», *Journal of Applied Econometrics* 1, 277—286.

Lee, L-F. and M.M. Pitt (1984), »Microeconometric models of consumer and producer demand with limited dependent variables», mimeo.

Marquardt, D.W. (1963), »An algorithm for least-squares estimation of nonlinear parameters», *Journal of the Society for Industrial and Applied Mathematics* 11, 431—441.

Pudney, S.E. (1985), »Frequency of purchase and Engel curve estimation», discussion paper A56, LSE Econometrics Programme.

Pudney, S.E. (1987), »On the estimation of Engel curves», mimeo LSE.

Tobin, J. (1958), »Estimation of relationships for limited dependent variables», *Econometrica* 26, 24—36.

Waldman, D.M. (1985), »Computation in duration models with heterogeneity», *Journal of Econometrics (Annals 1985-1)* 28, 127—134.

Wales, T.J. and A. Woodland (1980), »Sample Selectivity and the estimation of labour supply functions», *International Economic Review* 21, 437—468.

Wales, T.J. and A. Woodland (1983), »Estimation of consumer demand systems with binding non-negativity constraints», *Journal of Econometrics* 21, 263—285.

White, H. (1980), »Nonlinear regression on cross-section data», *Econometrica* 48, 721—746.