MPRA
Munich Personal RePEc Archive

Bayesian estimation of the infrequency of purchase model with an application to food demand in the UK

Tiffin, R and Arnoult, M
University of Reading

26. August 2008

Online at http://mpra.ub.uni-muenchen.de/18836/
MPRA Paper No. 18836, posted 23. November 2009 / 19:55

Bayesian Estimation of the Infrequency of Purchase Model with an Application to Food Demand in the UK

Richard Tiffin and Matthieu Arnoult

University of Reading

August 26, 2008

Abstract

A Bayesian method of estimating multivariate sample selection models is introduced and applied to the estimation of a demand system for food in the UK to account for censoring arising from infrequency of purchase. We show how it is possible to impose identifying restrictions on the sample selection equations and that, unlike a maximum likelihood framework, the imposition of adding up at both latent and observed levels is straightforward. We find that higher income households consume proportionately more meat and more fresh fruit and vegetables. Regional differences in fruit and vegetable consumption are marked with the highest levels of consumption in London and the South East. The presence of children in a household reduces levels of vegetable consumption. Households employed in the professional or managerial sectors have higher levels of fruit and vegetable consumption. Age

has an influence on the consumption of fats and sugars, with consumption declining amongst older households and on the types of fruit and vegetables consumed with younger households preferring more ready meals and prepared fruit and vegetables.

1 Introduction

It is increasingly recognised that diet related chronic disease represents one of the most significant public health challenges of the twenty first century. For example the prevalence of overweight and obesity has grown rapidly since the 1980s and, according to the Health Survey for England, in 2004 63% of the adult population had a BMI greater than 25 while 24% were obese (BMI greater than 30). In addition to obesity, the roles that can be played by fruit and vegetables in the prevention of cancer also commands attention as do the impacts of dietary fat composition on fat and lipoprotein levels in the blood and associated impacts on heart disease. There is also a recognition that the diet related health problems are not evenly distributed in society: Drewnowski (2004) notes that in the United States obesity and type 2 diabetes follow a socioeconomic gradient with the highest rates of disease observed among groups with the highest poverty rates and the least education. Dowler (2003) considers the concept of "food poverty", noting that it is a term which is gaining currency in the UK. She argues that the concept is moving away from a technical conceptualisation in terms of minimal nutritional standards towards a definition which includes aspects of social and cultural participation. She continues to note however, that regardless of which definition is used, in developed countries a pattern exists whereby those living on low wages, or in areas of deprivation have lower nutrient intakes and worse dietary patterns than those not living in such circumstances. In an economic framework, whether food poverty is a consequence of financial poverty or of preference heterogeneity is an empirical question. The increasing availability and ease of analysing micro-data mean that it has become possible to address this question by estimating models of demand using such data. The first objective of this paper is therefore to disentangle whether poor diets are a consequence of economic poverty or preference heterogeneity between different household types.

Micro-data are in general subject to the econometric problem of censoring. In demand analysis this arises because most households do not purchase all of the commodities available to them. Wales & Woodland (1983) introduce two econometric models for censored demand systems. They refer to the first model as the Kuhn-Tucker approach. As its name implies, it is based on the Kuhn-Tucker conditions for the consumer's optimisation problem. The econometric model is developed by adding a stochastic term to the utility function and as a result to the Kuhn-Tucker conditions. The conditions hold as an equality when an interior solution results and as an inequality when there is a corner solution. As a result the likelihood function is of a mixed discrete-continuous form (Pudney (1989, p163)) and is difficult to maximise for all but relatively small demand systems because of the numerical integration that is required in its evaluation. The intractability of the likelihood function has led to very few examples of the empirical implementation the Kuhn-Tucker approach, one example is Phaneuf, Kling & Herriges (2000). By contrast, the second model proposed by Wales & Woodland (1983), which they refer to as the Amemiya-Tobin approach, has been more widespread in the literature. This second strategy for handling censoring is an application of the

Tobit model (Tobin (1958)) as extended by Amemiya (1974) to the estimation of a system of equations. In this approach the demand model is derived without explicitly incorporating the non-negativity conditions. Instead these are added to the estimated model by truncating the distribution of the stochastic demand choices to allow for a discrete probability mass at zero. A number of strategies have been adopted to the estimation of the Kuhn-Tucker model. The direct estimation of the system by maximum likelihood has been problematic for reasons of computational complexity. Earlier attempts at the estimation of the Amemiya-Tobin model are therefore based on the two stage approach proposed by Heien & Wessells (1990) and developed by Shonkwiler & Yen (1999) which is itself an application of the Heckman (1979) method. The two step approach can be considered a generalisation of the Amemiya-Tobin approach because it comprises two sets of equations: in addition to the censored equations, additional equations are used to model the censoring and this allows the possibility of a difference between the models which determine the censoring rule and the continuous observations. The generalisation of the Tobit model in this way is discussed in the context of demand for a single good by Blundell & Meghir (1987) who refer to the model in which the sample selection rule and the continuous variable models differ as the double hurdle model, a model introduced originally by Cragg (1971). The double hurdle model is adapted by Blundell & Meghir (1987) to form an infrequency of purchase model which addresses the fact that with a truncated survey period, observed purchases may differ from actual demand as stocks are either built up or run down. Yen, Lin & Smallwood (2003) note that two step estimation is consistent but inefficient and they return to maximum likelihood estimation of the original Amemiya-Tobin model using simulated and quasi maximum likelihood methods. These methods are generalised in Stewart & Yen (2004) and Yen (2005) in an analogous way to the generalisation offered by the two step estimators referred to above to account for the differences in processes determining selection and the continuous variable. They recognise that this generalisation is the multivariate equivalent of that proposed by Cragg (1971). Their models are estimated by maximum likelihood and are thus efficient.

The second objective of this paper is to contribute to this literature by applying Bayesian methods to the estimation of multivariate sample selection models. We also extend the range of models that have been estimated by maximum likelihood hitherto to the infrequency of purchase model and we incorporate the Wales & Woodland (1983, p. 273) approach to the imposition of adding-up which, as Pudney (1989, p157) notes, has been problematic in a maximum likelihood context.

2 The Linearised AIDS IPM

The almost ideal demand system (AIDS) is written:

$$s_{it}^{*} = \alpha_i + \sum_{j=1}^{m+1} \gamma_{ij} \ln p_{jt} + \omega_i \ln \left(\frac{e_t}{P_t} \right) + \psi_i' h_t + u_{it} \tag{1}$$

$$i = 1, \ldots, m+1 \text{ and } t = 1, ...T \tag{2}$$

$$(u_{1t}, \ldots u_{mt})' \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{3}$$

where, p_{jt} is the price of the j^{th} good to the t^{th} household e_t is total expenditure, $P_t = \prod_j p_{jt}^{s_{jt}}$ is Stone's price index and h_t is a vector of variables that describes the t^{th} household. Note that the vector $(u_{1t}, \ldots u_{mt})$ excludes the $(m+1)^{th}$ equation so that $\boldsymbol{\Sigma}$ is positive definite. s_{it}^{*} is a latent share defined as follows to account for the fact the observed purchases may differ from actual consumption:

$$s_{it}^{*} = \frac{p_{it}q_{it}^{*}}{\sum_{i=1}^{m} p_{it}q_{it}^{*}} \tag{4}$$

where q_{it}^{*} is a latent variable defined as:

$$q_{it}^{*} = \begin{cases} q_{it}^{*} : & q_{it}^{*} \leq 0 \\ q_{it}\Phi_{it} : & q_{it}^{*} > 0, \end{cases} \tag{5}$$

q_{it} is the quantity of good i purchased by the t^{th} household and Φ_{it} is the probability that a purchase is made in any given survey period. Let us define the following share s_{it}^{**} which is determined by the following censoring rule:

$$s_{it}^{**} = y_{it} max \left(s_{it}^{*}, 0 \right) \tag{6}$$

where y_{it} is a binary variable which has the value one when the i^{th} good is bought by the t^{th} household. Note that the latent shares defined in equation 4 sum to one by construction. The censored shares s_{it}^{**} however will not satisfy this adding up restriction and the commonly adopted practice of treating s_{it}^{**} as the observed share is therefore questionable. In order to address this Wales & Woodland (1983, p270) propose that s_{it}^{it} are treated as latent variables which are related to the “observed” shares s_{it} as follows:[1]

$$s_{it} = \frac{p_{it}q_{it}^{*}}{\sum_{i \in C} p_{it}q_{it}^{*}} = \frac{s_{it}^{**}}{\sum_{i=1}^{m+1} s_{it}^{**}} \tag{7}$$

[1] We refer to these as observed shares for consistency with Wales & Woodland (1983, p270), they are however latent in a sense because they are based on the unobserved consumption levels.

where:

$$C = \{i : s_{it}^{**} > 0\}. \tag{8}$$

Note that the second equality in equation 7 ensures that adding-up is satisfied. The relationship in equation 7 enables us to work back to obtain the unobserved latent shares for the uncensored observations from the "observed" shares computed using equation 4 by applying the following formula:

$$s_i^* = s_i \left(1 - \sum_{i \notin I} s_i^*\right) \quad \forall i \in C \tag{9}$$

In compact form, the full AIDS is written:

$$\mathbf{s}^* = \mathbf{X}_2 \mathbf{\Lambda} + \mathbf{v} \tag{10}$$

where:

$$\mathbf{X}_2 = \mathbf{I}_m \otimes \mathbf{x}_2, \tag{11}$$

$$\mathbf{x}_2 = (\mathbf{x}_{21} \ldots, \mathbf{x}_{2T})', \tag{12}$$

$$\mathbf{x}_{2t} = \left(1, \ln p_{1,t}, \cdots, \ln p_{m+1,t}, \ln\left(\frac{e_t}{P_t}\right), h_t'\right)', \tag{13}$$

$$\mathbf{s}^* = (s_{1,1}^*, \cdots, s_{1,T}^*, s_{2,1}^*, \ldots, s_{2,T}^*, \ldots, s_{m,1}^*, \ldots s_{m,T}^*)', \tag{14}$$

$$\mathbf{\Lambda} = \left(\alpha_1, \gamma_{11}, \ldots \gamma_{1,m+1}, \omega_1, \psi_{1,}', \ldots, \alpha_m, \gamma_{m1}, \ldots \gamma_{m,m+1}, \omega_m, \psi_{m,}'\right)', \tag{15}$$

and:

$$\mathbf{v} = (v_{1,1}, \cdots, v_{1,T}, v_{2,1}, \ldots, v_{2,T}, \ldots, v_{m,1}, \ldots v_{m,T})' \tag{16}$$

The underlying theory requires that the model satisfies symmetry

$$\gamma_{ij} = \gamma_{ji} \text{ for all } i,j, \tag{17}$$

homogeneity

$$\sum_j \gamma_{ij} = 0 \text{ for all } j \tag{18}$$

and concavity. Concavity implies that the Slutsky matrix ($\mathbf{M}$) which has the elements:

$$M_{ij} = \gamma_{ij} + \omega_i \omega_j \ln\left(\frac{e}{P}\right) - s_i \delta_{ij} + s_i s_j \tag{19}$$

$$\delta_{ii} = 1, \delta_{ij} = 0 : i \neq j \tag{20}$$

is negative semi-definite. All of these restrictions are imposed in our empirical application.[2]

To complete the IPM, the demand equations in 10 are augmented with m probit equations to give the complete model:

$$\mathbf{y}^* = \mathbf{X}_1 \beta_1 + \mathbf{u} \tag{21}$$

$$\mathbf{s}^* = \mathbf{X}_2 \mathbf{\Lambda} + \mathbf{v} \tag{22}$$

where $\mathbf{y}_1^*$ is an $mT \times 1$ vector of latent variables structured in the same way as $\mathbf{s}^*$

[2]Full details of the procedure used to impose concavity are available in an accompanying working paper.

(see equation 14) and based on the binary variable y_{it} defined in equation 23:

$$y_{it}^{*}\begin{cases} > 0 & y_{it} = 1 \\ \leq 0 & y_{it} = 0 \end{cases} \tag{23}$$

and

$$\mathbf{X}_1 = \mathbf{I}_m \otimes \mathbf{x}_1 \tag{24}$$

$$\mathbf{x}_1 = (\mathbf{x}_{11} \ldots , \mathbf{x}_{1T}) \tag{25}$$

is a matrix of variables that describe household specific characteristics which are assumed to determine the probability of the household making a purchase in a given time period. In our application we assume that all households are identical in this respect and stocks are exhausted in a purely random manner and $\mathbf{x}_1$ is a therefore a vector of constants. It is assumed that:

$$\mathbf{e} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \sim N(0, \Sigma), \tag{26}$$

We estimate the model using the Markov chain Monte-Carlo methods. These allow draws to made on the marginal posterior distributions by drawing iteratively on the conditional posterior distributions for each block of parameters in the model. In order to proceed we therefore need to identify the forms of these conditional posterior distributions. If the dependent variables in 21 and 22 were observable, the full system comprising both sets of equation could be treated as a set of seemingly unrelated equations (SUR) and estimation would be straightforward. Writing the complete system in 21 and 22 as:

$$\mathbf{y}^* = \mathbf{X}\beta + \mathbf{e} \tag{27}$$

where:

$$\mathbf{y}^* = (\mathbf{y}^*, \mathbf{s}^*)', \mathbf{X}= \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{W} \end{pmatrix}, \beta= \left(\beta_1', \mathbf{\Theta}'\right)', \mathbf{e}= (\mathbf{u}', \mathbf{v}')' \tag{28}$$

the conditional distributions are:

$$p\,(\beta|\mathbf{y}, \mathbf{X}, \mathbf{\Sigma}) \sim MVN\left(\left(\mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\right)\mathbf{y}^*, \mathbf{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}\right) \tag{29}$$

$$p\,(\mathbf{\Sigma}|\mathbf{y}, \mathbf{X}, \mathbf{\Theta}) \sim IW\,(\tilde{\mathbf{e}}'\tilde{\mathbf{e}}, T) \tag{30}$$

where:

$$\tilde{\mathbf{e}} = \begin{pmatrix} u_{1,1} & \dots & u_{m,1} & e_{1,1} & \dots & e_{m,1} \\ \vdots & & \vdots & \vdots & & \vdots \\ u_{1,T} & \dots & u_{m,T} & e_{1,T} & \dots & e_{m,T} \end{pmatrix} \tag{31}$$

Since the latent data are not observed, we employ data augmentation (Tanner & Wong (1987)) to estimate the model. In this approach the conditional pdfs of the latent data are used to simulate the missing data. The simulated data then replaces the censored observations in all other steps of the Markov chain. The column vector of dependent variables for the t^{th} household is defined as $\mathbf{y}_t$ with its fitted value defined as $\hat{\mathbf{y}}_t$. Defining the precision matrix $\mathbf{H} = \mathbf{\Sigma}^{-1}$, the conditional mean ($\mu_{it}$) and variance ($V_i$) of the latent variables are (Geweke (2005, Theorem

5.3.1)):

$$\mu_{it} = \hat{y}_{it} + \mathbf{\Sigma}_i \mathbf{\Sigma}_{-i}^{-1} \left(\mathbf{y}_{-i,t} - \hat{\mathbf{y}}_{-i,t}\right) = \hat{y}_{it} - H_{ii}^{-1} \mathbf{H}_{-i} \left(\mathbf{y}_{-i,t} - \hat{\mathbf{y}}_{-i,t}\right) \tag{32}$$

$$V_i = \Sigma_{ii} - \mathbf{\Sigma}_i \mathbf{\Sigma}_{-i}^{-1} \mathbf{\Sigma}_i' = \mathbf{H}_{-i}^{-1} \tag{33}$$

where Σ_{ii} is the i^{th} on-diagonal element of $\mathbf{\Sigma}$, $\mathbf{\Sigma}_i$ is the i^{th} row of $\mathbf{\Sigma}$ excluding Σ_{ii}, and $\mathbf{\Sigma}_{-i}$ is the matrix within $\mathbf{\Sigma}$ excluding both the i^{th} column and i^{th} row. H_{ii} and $\mathbf{H}_i$ are similarly defined. $\hat{y}_{it}$ is the fitted value of y_{it} for the t^{th} household and $\hat{\mathbf{y}}_{-i,t}$ and $\mathbf{y}_{-i,t}$ are vectors within $\hat{\mathbf{y}}_t$ and $\mathbf{y}_t$ respectively, with their i^{th} elements removed. The latent data in the probit equations are generated using the rules:

$$y_{it} = 0 : y_{it}^* | \mathbf{y}_{-i,t}^*, \mathbf{\Theta}, \mathbf{X}, \mathbf{\Sigma} \sim \mathbf{N}\left(\mu_{it}, V_i\right) I_{[-\infty,0]} \tag{34}$$

$$y_{it} = 1 : y_{it}^* | \mathbf{y}_{-i,t}^*, \mathbf{\Theta}, \mathbf{X}, \mathbf{\Sigma} \sim \mathbf{N}\left(\mu_{it}, V_i\right) I_{[0,\infty]} \tag{35}$$

and in the share equations by:

$$s_{it} = 0 : s_{it}^* | \mathbf{y}_{-i,t}^*, \mathbf{\Theta}, \mathbf{X}, \mathbf{\Sigma} \sim \mathbf{N}\left(\mu_{it}, V_i\right) \tag{36}$$

where $I_{[-\infty,0]}$ is an indicator variable that is one if $y_{it} \in [-\infty, 0]$ and zero otherwise. Finally, because we employ the Wales & Woodland (1983, p270) approach to ensure that adding up is satisfied by the latent shares we have to obtain latent shares for observations where purchases are made:

$$s_i^* = s_i \left(1 - \sum_{i \notin I} s_i^*\right) \ \forall i \in C \tag{37}$$

where C is defined in equation 8.

The final issue which has to be addressed is the identification of the probit equations. To achieve this it is necessary to restrict the covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}} & \boldsymbol{\Sigma}_{\mathbf{uv}} \\ \boldsymbol{\Sigma}_{\mathbf{vu}} & \boldsymbol{\Sigma}_{\mathbf{vv}} \end{pmatrix}. \tag{38}$$

We impose the restriction that $\Sigma_{uu} = I$. Standard results give:

$$\mathbf{u} \sim (0, \Sigma_{uu}) \tag{39}$$

$$\mathbf{v}|\mathbf{u} \sim N\left(\Sigma_{uv}\Sigma_{uu}^{-1}u,\ \Sigma_{vv} - \Sigma_{uv}\Sigma_{uu}^{-1}\Sigma_{uv}\right). \tag{40}$$

In the regression:

$$\widetilde{\mathbf{v}} = \widetilde{\mathbf{u}}\delta + \varepsilon, \tag{41}$$

where $\widetilde{\mathbf{v}}$, $\widetilde{\mathbf{u}}$ and ε are $T \times m$ matrices:

$$\widetilde{\mathbf{u}} = \begin{pmatrix} v_{11} & v_{21} & \cdots & v_{m1} \\ \vdots & \vdots & & \vdots \\ v_{1T} & v_{2T} & \cdots & v_{mT} \end{pmatrix} \tag{42}$$

$$\widetilde{\mathbf{u}} = \begin{pmatrix} u_{11} & u_{21} & \cdots & u_{m1} \\ \vdots & \vdots & & \vdots \\ u_{1T} & u_{2T} & \cdots & u_{mT} \end{pmatrix} \tag{43}$$

and δ is $M \times M$, we can write:

$$\delta = \left(\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}}\right)^{-1} \widetilde{\mathbf{u}}'\widetilde{\mathbf{v}}. \tag{44}$$

Hence:

$$\delta = \mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uv}, \tag{45}$$

and:

$$\mathbf{\Sigma}_{uv} = \mathbf{\Sigma}_{uu}\delta. \tag{46}$$

Moreover:

$$\mathbf{\Sigma}_{\varepsilon} = \mathrm{cov}(\mathbf{v}|\mathbf{u}) \tag{47}$$

$$= \mathbf{\Sigma}_{vv} - \mathbf{\Sigma}_{uv}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uv}. \tag{48}$$

Hence:

$$\mathbf{\Sigma}_{vv} = \mathbf{\Sigma}_{\varepsilon} + \mathbf{\Sigma}_{uv}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uv}. \tag{49}$$

Therefore, under he assumption that $\mathbf{\Sigma}_{uu} = \mathbf{I}$, we can recover the other parts of Σ as follows:

$$\Sigma_{uv} = \delta, \tag{50}$$

$$\Sigma_{vv} = \Sigma_{\varepsilon} + \Sigma_{uv}\Sigma_{uv}. \tag{51}$$

From the regression in equation (41), it can be seen that the conditional distributions for δ and Σ_{ε} are normal (N) and inverted Wishart (IW) respectively:

$$\delta|\mathbf{\Sigma}_{\varepsilon} \sim N\left[\left(\widetilde{\mathbf{u}}'\mathbf{\Sigma}_{\varepsilon}^{-1}\widetilde{\mathbf{u}}\right)^{-1}\widetilde{\mathbf{u}}'\widetilde{\mathbf{v}}\,,\,\left(\widetilde{\mathbf{u}}'\mathbf{\Sigma}_{\varepsilon}^{-1}\widetilde{\mathbf{u}}\right)^{-1}\right] \tag{52}$$

$$\mathbf{\Sigma}_{\varepsilon}|\delta \sim IW(\varepsilon'\varepsilon, T) \tag{53}$$

In order to identify the probit equations we impose the restriction $\Sigma_{uu} = I$ and replace the inverted Wishart draw on the full covariance matrix $\mathbf{\Sigma}$ with draws on the conditional distributions on the distributions in 52 and 53 and obtain the unrestricted blocks of Σ using 50 and 51.

The estimation algorithm can then be stated as:

1. Draw the parameter vector $\mathbf{\Lambda}$ from the normal distribution in equation 29.

2. Draw the latent data for the probit equations from the truncated normal distributions in equations 34 and 35.

3. Obtain the latent data for the share equations:

   (a) Where the share is censored make a draw on the distribution in equation 36.

   (b) Where a purchase is observed:

      i. compute the probability of a purchase:

      $$\Phi_{it} = p(y_{1it} = 1) = \mathbf{\Phi}\left(v_{it} > -\mathbf{x}_{1t}\beta_1\right) = \mathbf{\Phi}\left(\mathbf{x}_{1t}\beta_{1i}\right)$$

      and use this to compute:

      $$s_{it} = \frac{p_{it}q^*_{it}}{\sum_{i\in I} p_{it}q^*_{it}}$$

      ii. Compute the latent share according to equation 37.

4. Draw the variance-covariance matrix $\mathbf{\Sigma}$:

   (a) Draw δ from the normal distribution in 52.

(b) Draw $\mathbf{\Sigma}_\varepsilon$ from the inverse Wishart distribution in 53.

(c) Construct the complete matrix using equations 50 and 51.

5. Return to step 1.

3 Data and aggregation

We use the UK government's expenditure and food survey (EFS) for 2003-4. Participating households voluntarily record food purchases for consumption at home for a two week period using a food diary. The sample is based on 7,014 households in 672 postcode sectors stratified by Government Office Region, socioeconomic group and car ownership. It is carried out throughout the UK and throughout the year in order to capture seasonal variations.

We estimate three models which are based on subsets of foods aggregated in such a way to be of particular interest from the perspective of dietary health policy. The three groups are respectively: the Balance of Good Health; Fish and Fruit and Vegetables. In all cases observations are excluded where none of the food groups in the model are consumed. This leaves 7,014, 4,914 and 6,800 observations respectively for the Balance of Good Health, Fish and Fruit and Vegetable models respectively. The Balance of Good Health model comprises the following groups: Milk and Dairy; Meat Fish and Alternatives; Bread, Cereals and Potatoes; Fats and Sugar and Fruit and Vegetables. These are chosen because they correspond to groups used by the UK Food Standards Agency (FSA) in recommendations regarding what represents a balanced diet. In this model, levels of censoring vary from 0.54% for the cereals and potatoes group to 3.36% for fruit and vegetables.

In the Fish model we estimate demand equations for: White Fish; Salmon; Blue Fish; Shellfish and Other Fish. This model was chosen oily fish has been shown to have beneficial health impacts and there are therefore concerns about low levels of consumption in some groups. Here the levels of censoring 22.62% for other fish to 85.70% for shellfish. Finally, the Fruit and Vegetable model comprises demand equations for: Fresh Fruit and Vegetable; Frozen Fruit and Vegetable; Tinned Fruit and Vegetable; Prepared Fruit and Vegetable and Fruit and Vegetable based ready meals. The levels of censoring in this model range from 3.6% for fresh fruit and vegetables to 70.69% for frozen fruit and vegetables. This aggregation was chosen because of the objective to increase consumption of fresh fruit and vegetables.[3] Prices are not available in the EFS and we therefore follow what has become common practice (Yen et al. (2003) and Yen & Lin (2006)) in using unit values to represent household prices and by imputing the missing prices for censored observations as regional averages. We recognise that alternatives to this approach exist, for example Deaton (1988) and Deaton (1990) address the problems associated with using unit values as opposed to prices and as Yen et al. (2003) note, Rubin (1996) offers a more robust methods for imputation. We argue however that these methods are beyond the scope of this paper.

Demographic characteristics are included in the demand system by augmenting each of the share equations with a set of dummy variables to represent the characteristics listed in table 1.

[3]Full details of the foods included in each of these models are available on request

Household Composition	Adults only Single parents Family with children Family with children & more than 2 adults Family without children & more than 2 adults*
Socio-economic Group†	High managerial Low managerial Workers-technical Never work-unemployed Students Other*
Age†	< 30 $30 \leq age < 45$ $45 \leq age < 60$ ≥ 60*
GOR‡	North East – North West & Merseyside – Yorks & Humber – East Midlands – West Midlands – Eastern – London – South East – South West – Wales – Scotland – Northern Ireland*
Ethnic Origin†	White – Mixed race – Asian – Black – Other*
Gender†	Male – Female*

† Relating to the household reference person (HRP)
‡ Government Office Region
* indicates the ommitted dummy variable in each category, thereby defining the reference demographic group for interpretting results

Table 1: Demographic variables included in the share equations

	Price					
Quantity	Dairy	Meat	Fats	Cereals	F and V	Expenditure
Milk and Dairy	-0.202	-0.089	-0.093	-0.172	-0.156	0.718
2.5%	-0.315	-0.173	-0.155	-0.246	-0.218	0.692
97.5%	-0.136	-0.001	-0.036	-0.095	-0.092	0.742
Meat, Fish etc.	-0.092	-0.859	-0.105	-0.093	-0.018	1.163
2.5%	-0.123	-0.918	-0.137	-0.130	-0.050	1.147
97.5%	-0.059	-0.800	-0.074	-0.056	0.013	1.180
Fats	-0.108	-0.161	-0.525	-0.110	-0.026	0.930
2.5%	-0.161	-0.235	-0.597	-0.171	-0.081	0.906
97.5%	-0.059	-0.101	-0.456	-0.050	0.029	0.955
Bread, Cereals, Pots	-0.136	-0.073	-0.080	-0.524	-0.106	0.920
2.5%	-0.183	-0.135	-0.125	-0.699	-0.154	0.901
97.5%	-0.087	-0.021	-0.035	-0.449	-0.057	0.939
Fruit and Veg	-0.155	0.009	-0.038	-0.144	-0.710	1.038
2.5%	-0.199	-0.053	-0.085	-0.200	-0.776	1.014
97.5%	-0.109	0.061	0.008	-0.089	-0.645	1.057

Table 2: Elasticities of Demand for the Balance of Good Health Model

	Price					
Quantity	White	Salmon	Blue	Shell	Other	Expenditure
White	-0.918	0.039	0.011	0.152	0.060	0.873
2.5%	-1.029	-0.061	-0.079	0.057	-0.017	0.825
97.5%	-0.811	0.155	0.100	0.252	0.128	0.924
Salmon	0.016	-0.790	0.147	0.026	-0.194	0.924
2.5%	-0.115	-0.915	0.022	-0.101	-0.308	0.828
97.5%	0.146	-0.663	0.284	0.161	-0.084	0.992
Blue	-0.007	0.174	-0.771	-0.099	-0.060	0.913
2.5%	-0.145	0.042	-0.907	-0.259	-0.162	0.818
97.5%	0.132	0.310	-0.635	0.056	0.045	1.013
Shell	-0.075	-0.168	-0.265	-1.041	-0.324	1.321
2.5%	-0.212	-0.302	-0.407	-1.224	-0.434	1.194
97.5%	0.074	-0.027	-0.130	-0.857	-0.200	1.439
Other	-0.022	-0.163	-0.069	-0.053	-0.673	0.993
2.5%	-0.077	-0.218	-0.115	-0.094	-0.739	0.960
97.5%	0.024	-0.111	-0.021	-0.012	-0.590	1.031

Table 3: Elasticities of Demand for the Fish Model

			Price			
Quantity	Ready	Prepared	Tinned	Fresh	Frozen	Expenditure
Ready	-0.710	-0.017	0.004	-0.125	0.010	0.881
2.5%	-0.779	-0.067	-0.031	-0.199	0.001	0.845
97.5%	-0.643	0.032	0.042	-0.052	0.018	0.922
Prepared	0.000	-0.686	0.020	-0.092	0.022	0.807
2.5%	-0.043	-0.745	-0.014	-0.158	0.015	0.778
97.5%	0.041	-0.628	0.055	-0.025	0.030	0.834
Tinned	0.000	0.077	-0.831	0.115	0.051	0.663
2.5%	-0.018	0.000	-0.926	0.001	0.033	0.616
97.5%	0.122	0.154	-0.738	0.228	0.069	0.708
Fresh	0.049	-0.106	-0.033	-0.963	-0.022	1.153
2.5%	-0.105	-0.126	-0.049	-0.995	-0.025	1.143
97.5%	-0.067	-0.085	-0.017	-0.932	-0.018	1.163
Frozen	0.086	0.003	0.006	0.043	-0.977	0.944
2.5%	-0.007	-0.007	-0.005	0.029	-0.985	0.937
97.5%	0.008	-0.013	0.017	0.060	-0.968	0.951

Table 4: Elasticities of Demand for the Fruit and Vegetables Model

4 Results

Tables 2 to 4 show the price elasticities calculated using the following formula:

$$\epsilon_{ij} \quad = \quad -\delta_{ij} + \frac{\bar{\gamma}_{ij}}{\bar{w}_i} - \bar{\omega}_i \frac{\bar{s}_j}{\bar{s}_i}, \tag{54}$$

where $\bar{\gamma_{ij}}$ and $\bar{\beta_i}$ are the means of the values of the draws in the MCMC sample corresponding to the parameters defined in equation 1. $\bar{s_i}$ is the mean value of the i^{th} share across all observations in the data set and:

$$\begin{cases} \delta_{ii} & = & 1 \\ \delta_{ij} & = & 0 \end{cases} \quad i \neq j. \tag{55}$$

The expenditure elasticities in tables 2 to 4 are calculated as:

$$\epsilon_i = 1 + \frac{\beta_i}{s_i}. \tag{56}$$

We also report the highest posterior density intervals based on the 2.5 and 97.5 centiles in the MCMC sample and these show that a very high proportion of the estimated elasticities are significant in the sense that the interval does not span zero.

The elasticities for the balance of good health model that are reported in table 2 show that all of the foods are own price inelastic with milk and dairy the least responsive and meat and fish the most responsive. All of the significant cross price effects show the goods to be complementary emphasising the importance of the income effect in determining cross price responsiveness. This is a pattern that is repeated in the other two models albeit to a lesser extent and it raises questions about the use of differential pricing through taxation and subsidies in order to induce substitution from healthy to unhealthy foods. For example there is a comparatively strong complementary relationship between the price of fruit and vegetables and the quantity of cereals, bread and potatoes. Thus a subsidy on fruit and vegetables may be expected to have an undesirable impact on the quantity consumed of high calorie cereals, bread and potatoes. The effects may not all be undesirable, there is also a comparatively strong complementarity between the price of fats and sugars, a group which includes butter, jams, biscuits cakes and sweets, and meats, fish etc. This suggests that a "fat tax" my also have a beneficial impact in reducing consumption of read meats.

The expenditure elasticities show the impacts on demand for the individual goods of changes in expenditure on all foods within the system in question. These indicate therefore the relative effects of changes in income on the different food groups although we would expect the magnitude of the true income elasticities of demand to be smaller than these expenditure elasticities. Milk and dairy, fats and sugar and cereals bread and potatoes are income inelastic whilst meat, fish etc and fruit and vegetables are income elastic. This implies that households on higher incomes will consume a relatively higher proportion of meat and of fruit and vegetables.

In table 3 we see that all fish except for shellfish are own price inelastic. The table also shows that all fish except shellfish are income inelastic. Blue fish and Salmon are unusual in so far as they run counter to the general pattern of complementarity that is observed in the majority of cases. The fact that these two are substitutes is perhaps not surprising given that they are both oily fish. The expenditure elasticities suggest that there is likely to be a higher proportion of oily fish in comparison with white fish in the diets of high income households. Table 4 also shows that all of the groups within this category are price inelastic. The most notable feature of these results from the dietary health perspective is that the only expenditure elastic group is fresh fruit and vegetables. Thus not only do higher income households spend more on fruit and vegetables as a whole (c.f. table 2) but within the fruit and vegetable category they spend proportionately more on the fresh products.

Figures 1 to 15 show the effects of the demographic variables on demand for the food groups in each of the three demand systems that are estimated. These are estimated as the coefficients on dummy variables in the share equations and converted so that they measure the marginal effect in natural units at the mean

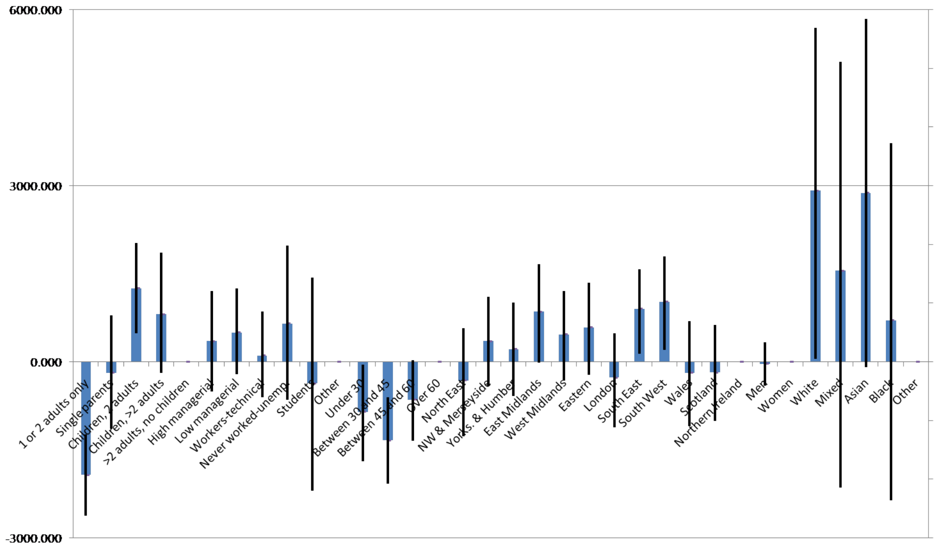

Figure 1: Effects of Demographics on Milk and Dairy Consumption (Balance of Good Health Model, mililitres)

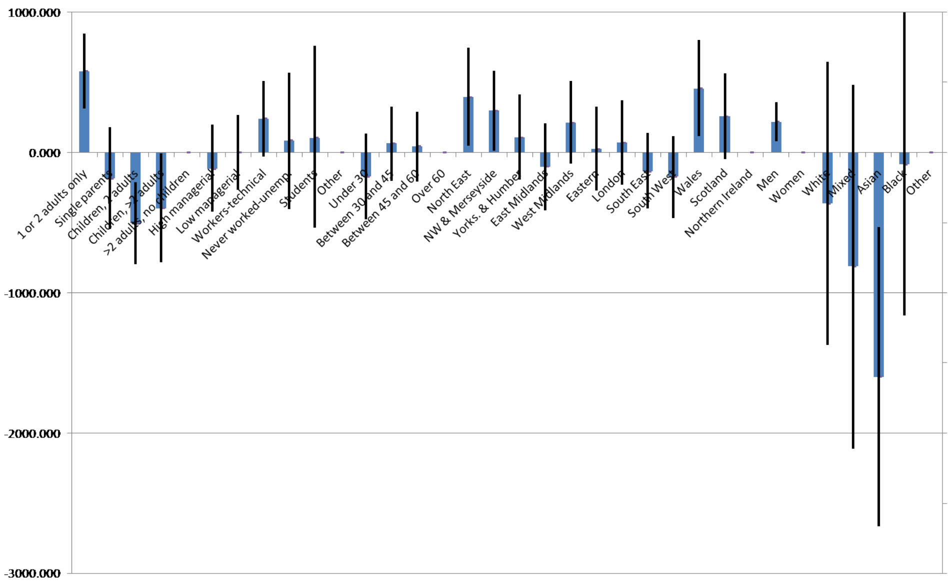

Figure 2: Effects of Demographics on Meat Consumption (Balance of Good Health Model, grammes)

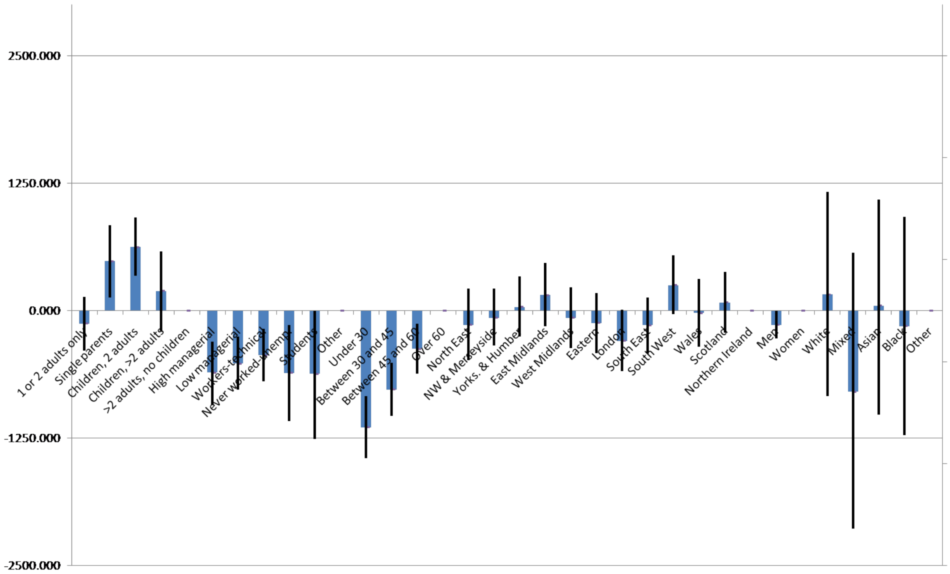

Figure 3: Effects of Demographics on Fats and Sugar Consumption (Balance of Good Health Model, grammes)

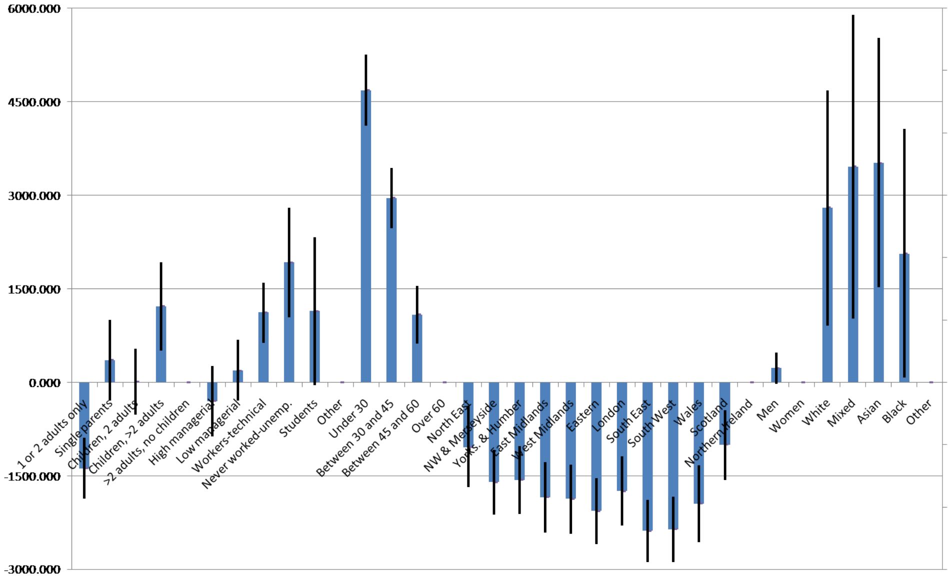

Figure 4: Effects of Demographics on Cereals and Potato Consumption (Balance of Good Health Model, grammes)

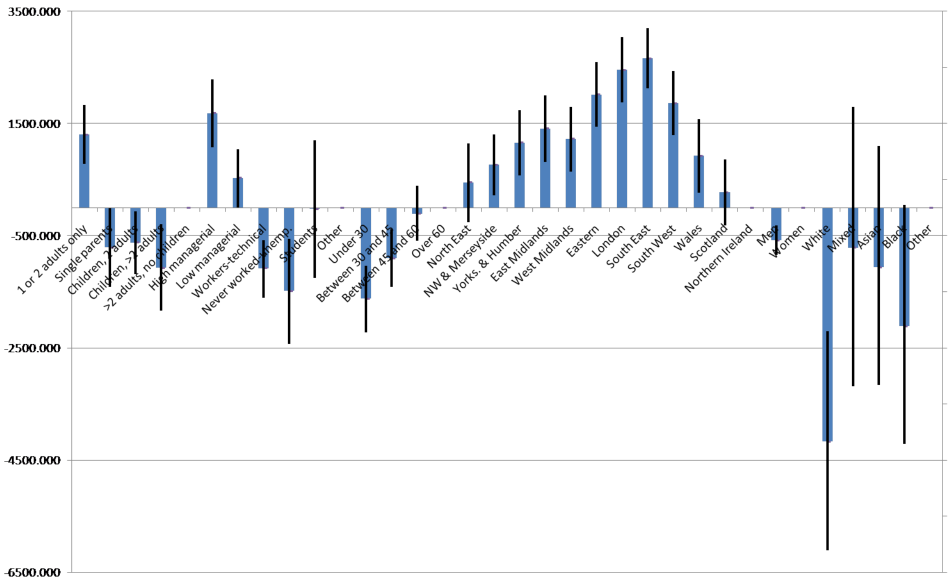

Figure 5: Effects of Demographics on Fruit and Vegetable Consumption (Balance of Good Health Model, grammes)

shares and prices across all households. All of the results show the effect relative to the reference group defined in table 1. The values are based on the mean values of the parameters in the Gibbs sample. We also show highest posterior density intervals for the coeficients based on the 2.5 and 97.5 centiles in the sample. In discussing these results we focus in particular on those which are significant in the sense that these intervals do not span zero.

Figures 1 to 5 show the results for the balance of good health model. In figure 1 we see that families without children and 2 adults or less consume significantly less milk and dairy products whilst families with children and 2 adults consume significantly more. Other family types with children consume more but the effect is not significant. There is a regional effect in South East and South West England where more milk and dairy products are consumed. The largest estimated impacts are the effects of ethnicity, in particular white and Asian families consume a lot more of this group but in both cases the effect is barely significant. Figure 2 shows the effects on meat consumption and we see that households with one or two adults only, households in Wales and households where a male is responsible for purchasing food have significantly elevated levels of meat consumption. Households which have children and two adults, more than two adults and of Asian ethnicity have significantly lower levels of meat consumption. Figure 3 shows the impacts on fats and sugars. All socio-economic groups have depressed levels of consumption in comparison with the reference "other" group. Age has a marked and significant effect with consumption of this group increasing with age. Families with children also have significantly elevated levels of consumption. Figure 4 shows the results for the Bread Cereals and Potatoes group. It can be seen that adults only households and all regions have significantly reduced levels of consumption of this group. By contrast, the workers/technical and never worked unemployed socioeconomic group and all ethnic groups apart from the reference "other" category all have elevated levels of consumption. There is also a clear effect of age, with consumption of this group decreasing with age. The final group in the balance of good health model is fruit and vegetables the results for which are shown in figure 5. Here we see that adult only households have increased levels of consumption in comparison with those of households with children. We also see that the managerial socioeconomic groups have higher levels of consumption than the blue collar group and the unemployed. There is also a regional effect with increased consumption in London and the South in particular compared with the North East, Scotland and Northern Ireland. Consumption is also reduced in households of white ethnicity.

Figures 6 to 10 show the effects of the demographic characteristics on demand for the five categories in the fish model. Figure 6 shows that the only significant

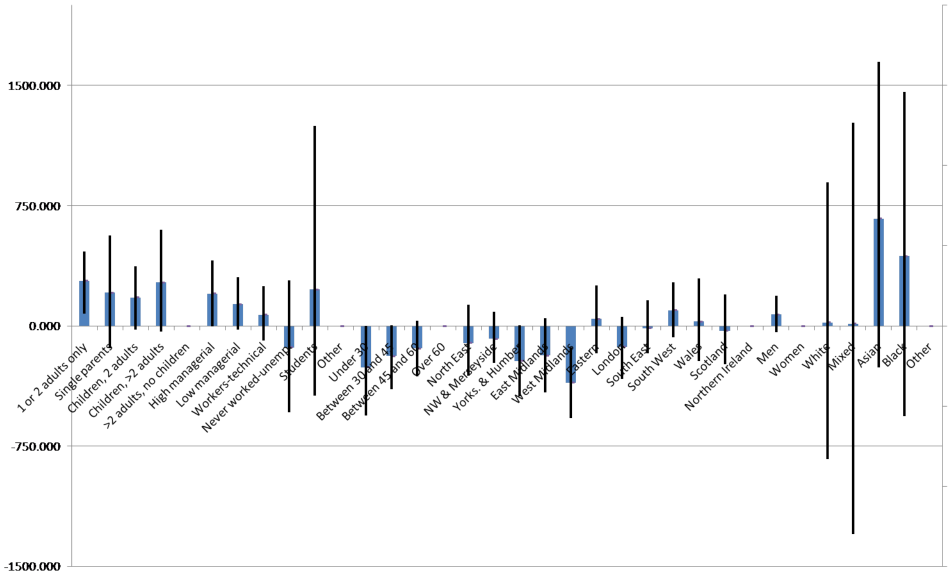

Figure 6: Effects of Demographics on White Fish Consumption (Fish Model, grammes)

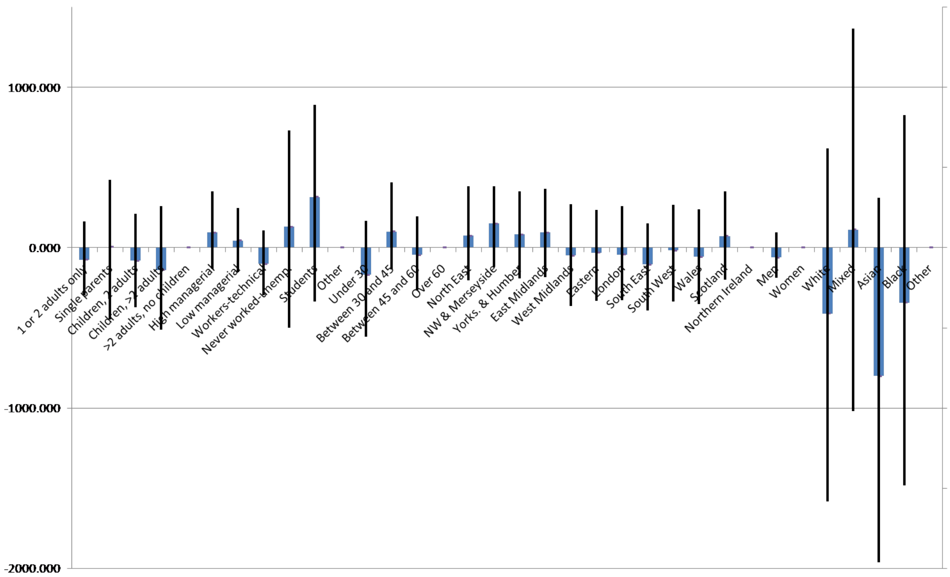

Figure 7: Effects of Demographics on Salmon Consumption (Fish Model, grammes)

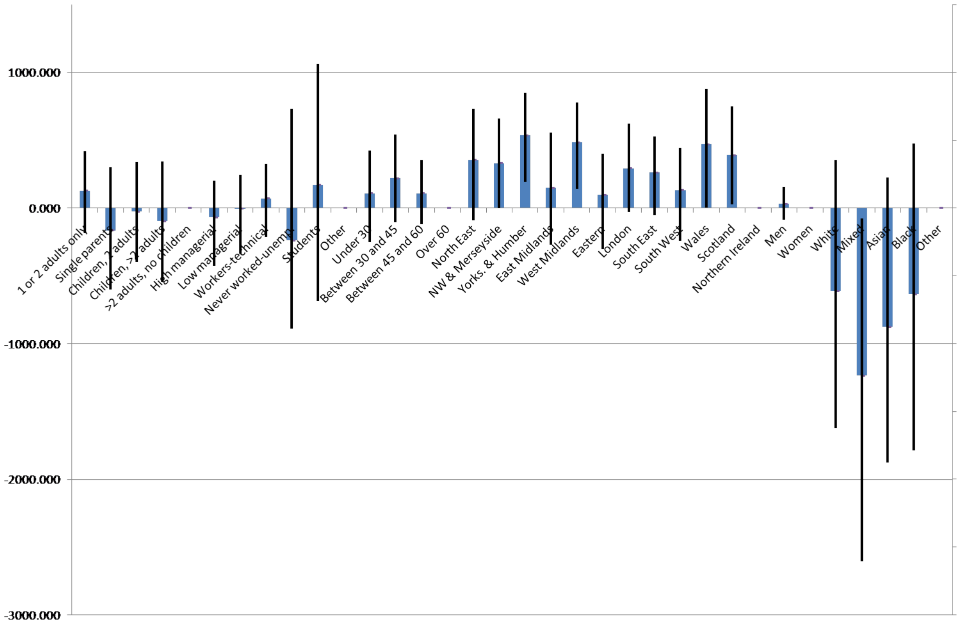

Figure 8: Effects of Demographics on Blue Fish Consumption (Fish Model, grammes)

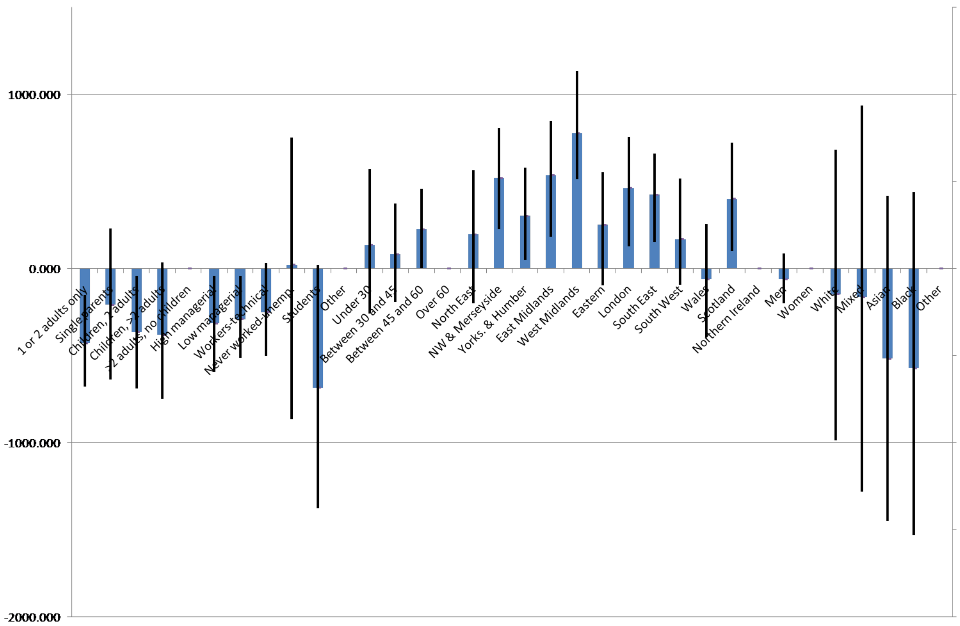

Figure 9: Effects of Demographics on Shellfish Consumption (Fish Model, grammes)

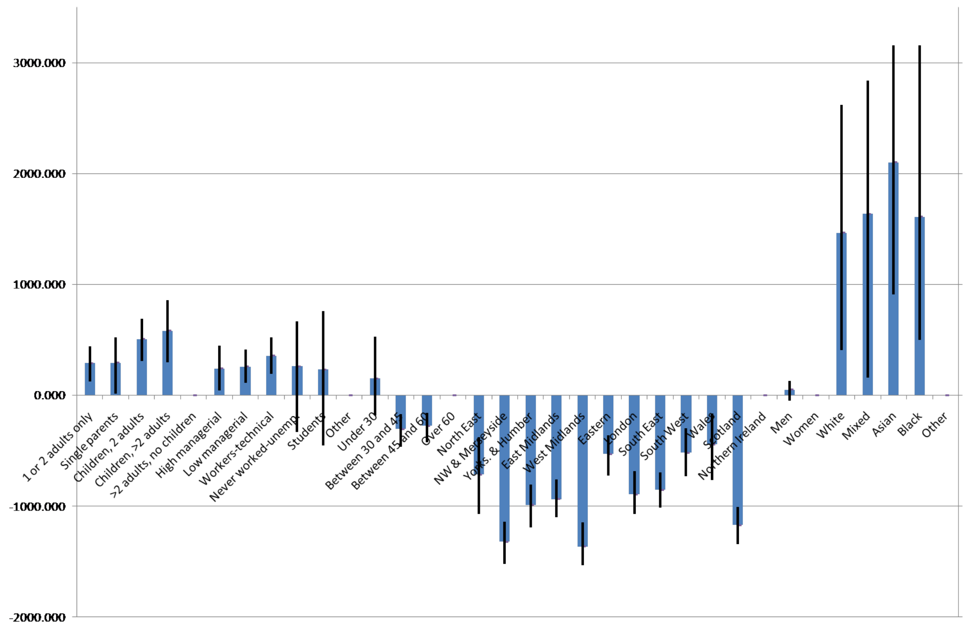

Figure 10: Effects of Demographics on Other Fish Consumption (Fish Model, grammes)

effects for white fish are elevated consumption by adults only households and depressed levels of consumption in the West Midlands, which is the English region that is most remote from the coast. Figure 7 shows that the differing demographic characteristics of households have no significant impact on the consumption of Salmon. In figure 8 we see that the only significant effects on blue fish consumption are regional with a consumers in the Yorkshire and Humberside and in the West Midlands favouring this category. Turning to the effects on demand for shellfish as depicted in figure 9 we see that adult only and households with 2 adults and children have significantly depressed demand for this category as do households in the managerial socio-economic groups. There is also a regional impact with households in the North West and Merseyside, the East and West Midlands, London and the South East and Scotland having an elevated preference for shell-

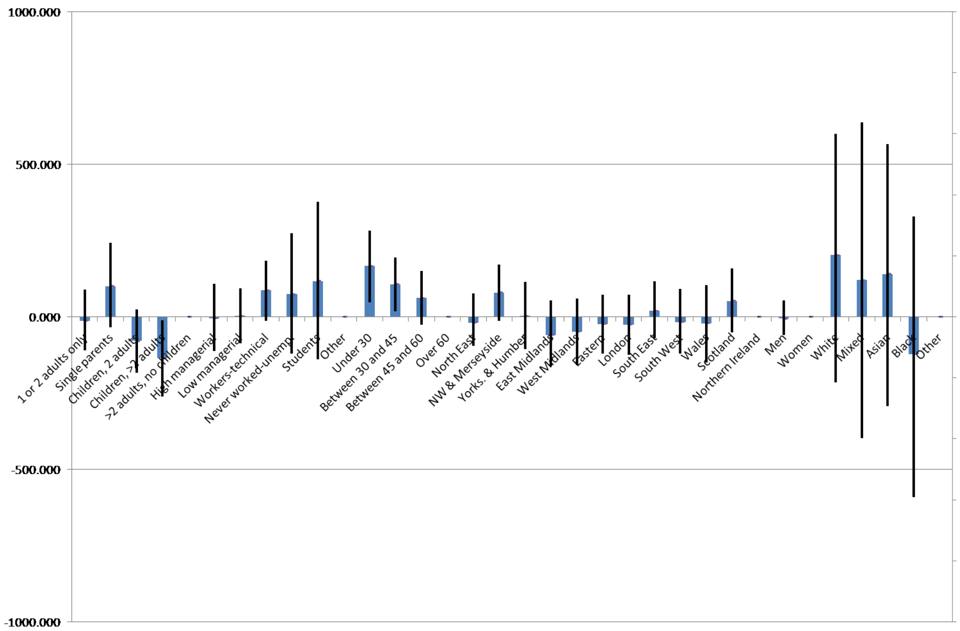

Figure 11: Effects of Demographics on Vegetable Based Ready Meal Consumption (Fruit and Vegetable Model, grammes)

fish. Figure 10 shows that the other fish category is the one which shows the most significant demographic impacts on demand. This category includes take-away fish, tinned fish and ready meals. The most marked demographic influence on the demand for white fish is regional where all regions have depressed demand. The largest effects are seen in the North West and Merseyside, the West Midlands and Scotland. There are also significant impacts for household composition, age, where those between 30 and 60 have a depressed demand and for ethnicity with Asian households having the strongest preference for this category.

Figures 11 to 15 show the demographic impacts on demand for foods in the fruit and vegetable category. Figure 11 shows that the only significant impact on the demand for vegetable based ready meals is age where households under the age of 45 have an elevated demand. Similarly, as can be seen in figure 12 age is the only

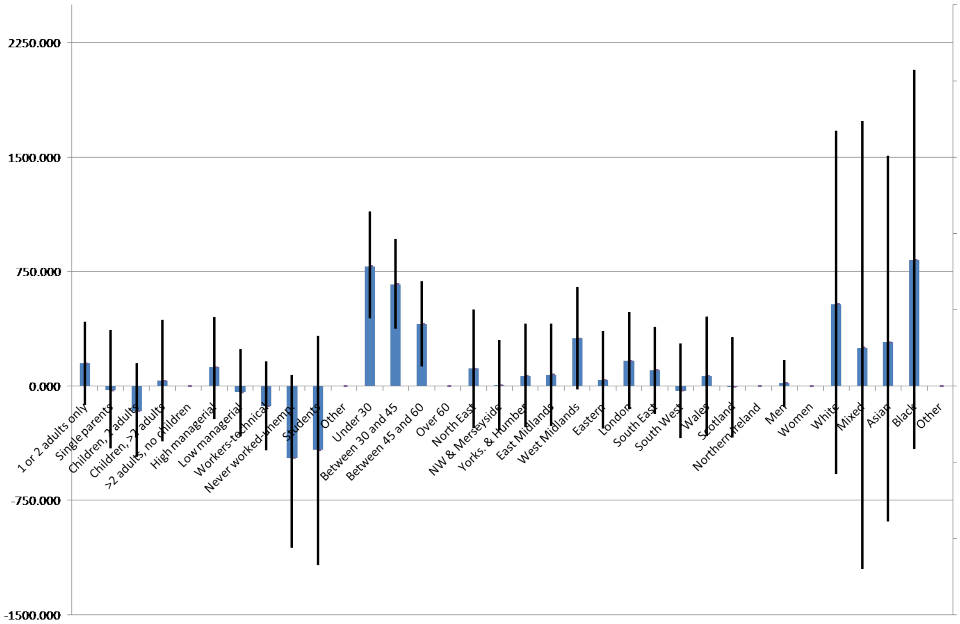

Figure 12: Effects of Demographics on Prepared Fruit and Vegetable Consumption (Fruit and Vegetable Model, grammes)

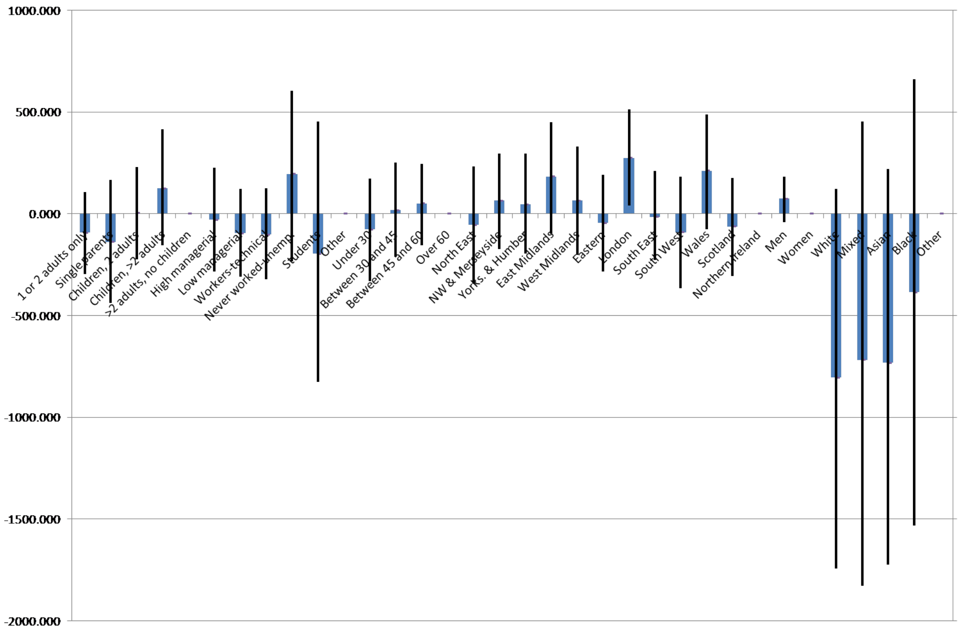

Figure 13: Effects of Demographics on Tinned Fruit and Vegetable Consumption (Fruit and Vegetable Model, grammes)

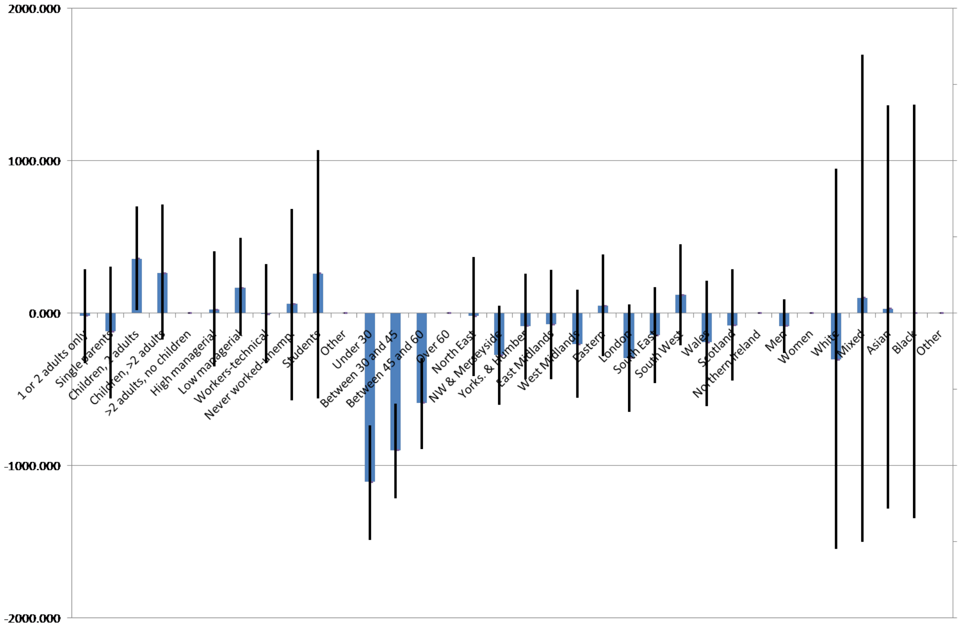

Figure 14: Effects of Demographics on Fresh Fruit and Vegetable Consumption (Fruit and Vegetable Model, grammes)

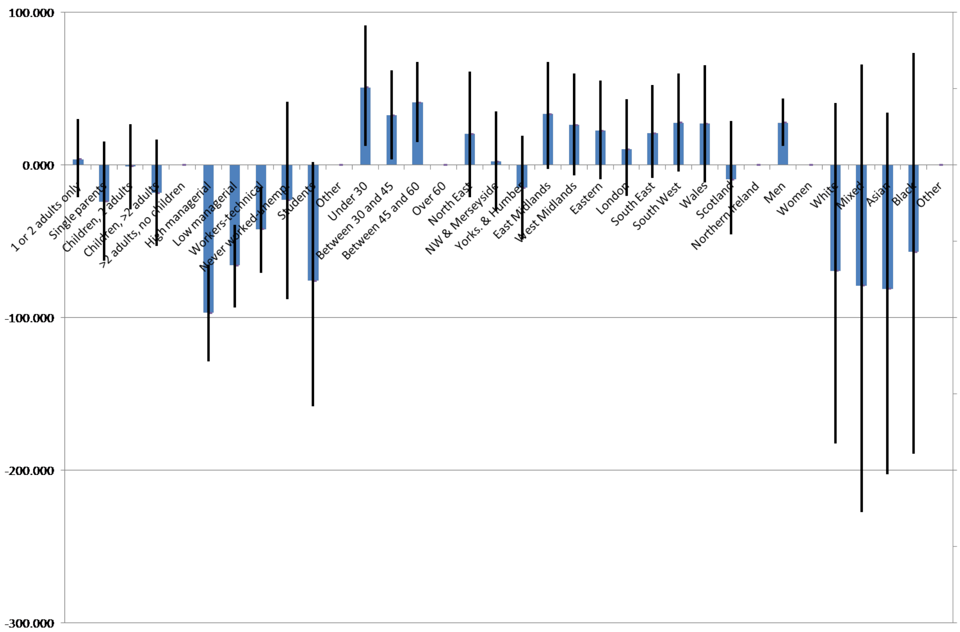

Figure 15: Effects of Demographics on Frozen Fruit and Vegetable Consumption (Fruit and Vegetable Model, grammes)

significant factor affecting the demand for prepared vegetables with households under 60 all have an elevated preference for this category. The only significant effect on the demand for tinned vegetables shown in figure 13 is in London where demand is elevated. The corollary of the age effects in ready meals and prepared vegetables is seen in the case of fresh vegetables (figure 14) where households under the age of 60 have a depressed preference. Figure 15 shows that households in the socioeconomic groupings in employment have significantly depressed demand for frozen fruit and vegetables whilst those in which a male responsible for purchasing food have an elevated level of demand. It is interesting to note that the presence of children in a household has no significant impact in any of the categories within the fruit and vegetable model. This implies that, whilst the results of the balance of good health model suggest that families with children spend less overall on fruit and vegetables, the allocation of spending between the different categories within the group is no different when there are children in the household.

5 Conclusion

We have demonstrated how the infrequency of purchase model can be estimated for a system of equations using Monte Carlo Markov chain methods. The method has been illustrated by estimating a model which is designed to disentangle the impacts of economic factors from preference heterogeneity resulting from differing demographic conditions in influencing the healthiness of diets in England and Wales.

Our results imply that households which have a higher level of income will tend to consume more meat and more fresh fruit and vegetables. Households in

London and the South East have higher levels of vegetable consumption whilst it is reduced by the presence of children. Households employed in the professional or managerial sectors have higher levels of fruit and vegetable consumption. Age has an influence on the consumption of fats and sugars with consumption declining amongst older households. Age also has an impact on the types of fruit and vegetables consumed with younger households preferring more ready meals and prepared fruit and vegetables.

References

Amemiya, T. (1974). Multivariate regression and simultaneous equation models when dependent variables are truncated normal, *Econometrica* **42**(6): 999 – 1012.

Blundell, R. & Meghir, C. (1987). Bivariate alternatives to the tobit-model, *Journal of Econometrics* **34**(1-2): 179 – 200.

Cragg, J. (1971). Some statistical models for limited dependent variables with application to demand for durable goods, **39**(5): 829 – 844.

Deaton, A. (1988). Quality, quantity and spatial variation of price, *American Economic Review* **78**(3): 418–430.

Deaton, A. (1990). Price elasticities from survey data: Extensions and indonesian results, *Journal of Econometrics* **44**: 281–309.

Dowler, E. (2003). Food and poverty, *Development Policy Review* **21**(5-6): 569–580.

Drewnowski, A. (2004). Obesity and the food environment - dietary energy density and diet costs, *American Journal of Preventive Medicine* **27**(3): 154 – 162.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, Wiley, Hoboken, New Jersey.

Heckman, J. (1979). Sample selection bias as a specification error, *Econometrica* **47**(1): 153–161.

Heien, D. & Wessells, C. (1990). Demand systems estimation with microdata - a censored regression approach, *Journal Of Business And Economic Statistics* **8**(3): 365 – 371.

Phaneuf, D., Kling, C. & Herriges, J. (2000). Estimation and welfare calculations in a generalized corner solution model with an application to recreation demand, *Review of Economics and Statistics* **82**(1): 83–92.

Pudney, S. (1989). *Modelling Individual Choice: The Econometrics of Corners Kinks and Holes*, Blackwell, Oxford.

Rubin, D. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**: 473–89.

Shonkwiler, J. & Yen, S. (1999). Two-step estimation of a censored system of equations, *American Journal Of Agricultural Economics* **81**(4): 972 – 982.

Stewart, H. & Yen, S. (2004). Changing household characteristics and the away-from-home food market: a censored equation system approach, *FOOD POLICY* **29**(6): 643 – 658.

Tanner, M. & Wong, W. (1987). The calculation of the posterior distribution by data augmentation (with discussion), *Journal of the American Statistical Association* **82**: 528–550.

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

Wales, T. & Woodland, A. (1983). Estimation of consumer demand systems with binding non-negativity constraints, *Journal Of Econometrics* **21**(3): 263 – 285.

Yen, S. (2005). A multivariate sample-selection model: Estimating cigarette and alcohol demands with zero observations, *American Journal Of Agricultural Economics* **87**(2): 453 – 466.

Yen, S. & Lin, B. (2006). A sample selection approach to censored demand systems, *American Journal of Agricultural Economics* **88**(3): 742 – 749.

Yen, S., Lin, B. & Smallwood, D. (2003). Quasi- and simulated-likelihood approaches to censored demand systems: Food consumption by food stamp recipients in the united states, *American Journal of Agricultural Economics* **85**(2): 458 – 478.