*Public Health Nutrition*: 4(5B), 1149–1151

DOI: 10.1079/PHN2001206

# Estimation of age- and gender-specific food availability from household budget survey data

VGS Vasdekis[1,]*, S Stylianou[2] and A Naska[2]

[1]Department of Statistics, Athens University of Economics and Business, 76 Patission Street, 10434 Athens, Greece: [2]Department of Hygiene and Epidemiology, Medical School, University of Athens, 75 Mikras Asias Street, Athens, Greece

## Abstract

*Objective:* To derive estimates of age–gender specific food availability, based on data collected at household level.

*Design:* Two alternative modelling approaches are described leading to linear and non-linear optimisation, respectively. The idea of penalised least squares is used for estimation of model parameters. The effect of household characteristics can be incorporated into both modelling approaches.

*Setting:* Household budget survey data from four European countries (Belgium, Greece, Norway and the United Kingdom), *circa* 1990.

**Keywords**
Household budget survey data
Non-parametric models
Semi-parametric models
Penalised least squares

Household budget surveys (HBSs) and individual nutrition surveys (INSs) are different sources of dietary information, both of interest to nutritionists. Their difference lies in the fact that household budget survey data are aggregated, representing food availability for the whole household, which is not the case for individual-based dietary survey data. Therefore, the estimates of individual food availability from HBS data represent a useful endeavour. In this undertaking, information could be retrieved from a large pool of data, such as the HBS. Moreover, the compatibility of HBS and INS needs to be assessed.

In this respect the problem is similar to that studied by Engle *et al.*[1], in which electricity demand over a billing period is modelled as a sum of independent daily demands each determined by temperature on the day concerned, the demand temperatures relationship being of main interest. If we assume that the days and temperatures of the study of Engle *et al.*[1] are, respectively, the household members and their genders and ages, then we have the appropriate link between the two analyses.

The objective of this paper is to describe two modelling approaches. The first, due to Chesher[2], leads to a semi-parametric model requiring non-linear estimation, while the other leads to a non-parametric model requiring linear weighted estimation. Both approaches use the idea of roughness penalty for the estimation of model parameters. This quantifies the notion of a rapidly fluctuating curve and then poses the estimation problem in a way that takes it into account.

## Models and estimation

Let us consider household $i$ ($i = 1, \ldots, n$) containing $m_i$ members, each with availability $y_{ij}$ ($j = 1, \ldots, m_i$) for some food of interest during a recording period. The personal characteristics such as age and gender of member $j$ are denoted by vector $c_{ij}$, where $j = 1, \ldots, m_i$. In this way, $C_i = (c_{i1}, c_{i2}, \ldots, c_{imi})^{\mathrm{T}}$ is a matrix of individual characteristics of all members of the household and $z_i$ can be a vector of household characteristics such as location of household, income of household or household composition. The average availability for person $m_i$ conditional on household composition and other personal characteristics is

$$E(y_{ij}|C_i, z_i) = f(c_{ij}, z_i), \tag{1}$$

where $f(c_{ij}, z_i)$ is an individual-specific availability function we wish to estimate. $f(c_{ij}, z_i)$ tells us how expected food consumption varies by age, gender and household characteristics. Since the food consumption within the household is the sum of consumptions of the household members during the recording period, the expected household food consumption is:

$$y_i = \sum_{i=1}^{m_i} f(c_{ij}, z_i) + \epsilon_i, \tag{2}$$

where quantities $\epsilon_i$ are assumed to be independent with mean 0 and variance matrix $\Sigma = \sigma^2 I$.

Chesher[2] presented a multiplicative model for the individual availability functions as

$$f(c_{ij}, z_i) = h(c_{ij})g(z_i). \tag{3}$$

*Corresponding author: Email vasdekis@aueb.gr

An alternative model was proposed by Vasdekis and Trichopoulou[3] as

$$f(c_{ij}, z_i) = h(c_{ij}) + g(z_i). \tag{4}$$

We shall refer to these models as the multiplicative and the additive one, respectively. Model (3) has a better interpretation from the nutritional point of view, since a change in a household characteristic causes a proportional effect to availability at different ages. This effect is the same for all ages with model (4). However, the latter leads to computationally simpler solutions and therefore is suitable for massive analyses.

Although there are various individual characteristics recorded in an HBS, the analysis presented in this paper will deal with age ($a_{ij}$ for member $j$ of household $i$) and gender ($s_{ij} = 1$ if member $j$ of household $i$ is male, and $s_{ij} = 0$ otherwise). Even among very young children, the average food consumption is different for males and females[4], so

$$f(c_{ij}) = s_{ij} f_{\mathrm{M}i}(a_{ij}) + (1 - s_{ij}) f_{\mathrm{F}i}(a_{ij}),$$

where $f_{\mathrm{M}i}(.)$ and $f_{\mathrm{F}i}(.)$ are age–intake functions for household $i$ for males and females, respectively. Estimation is difficult if functions $f$ are left unspecified. If, however, the problem is discretised, by approximating $f_{\mathrm{M}}(.)$ and $f_{\mathrm{F}}(.)$ by step functions with points of increase at integer years of age, then considerable simplification is obtained. Dropping subscript $i$, let, for member $j$, $w_j = (w_{j,0}, \ldots, w_{j,a_t})$ be a vector of binary indicators with $w_{j,a} = 1_{[a \le a_j \le a+1]}$, $a_j$ being this member's age and $t = \mathrm{M}$ or $t = \mathrm{F}$. Note that $a_{\mathrm{M}}$ and $a_{\mathrm{F}}$ represent the highest age at which an analysis is required for males and females, respectively. Then, age–intake relationships for either males or females are approximated by the discrete form $f_{\mathrm{M}}(a_j) = w_j^{\mathrm{T}} \beta_{\mathrm{M}}$ or $f_{\mathrm{F}}(a_j) = w_j^{\mathrm{T}} \beta_{\mathrm{F}}$, where $\beta_{\mathrm{M}} = (\beta_{\mathrm{M},0}, \ldots, \beta_{\mathrm{M},a_{\mathrm{M}}})$ and $\beta_{\mathrm{F}} = (\beta_{\mathrm{F},0}, \ldots, \beta_{\mathrm{F},a_{\mathrm{F}}})$ are vectors of age- and gender-specific average intakes. The final form of the model therefore is

$$y_i = (\beta_0 + \eta_{\mathrm{M}i}^{\mathrm{T}} \beta_{\mathrm{M}} + \eta_{\mathrm{F}i}^{\mathrm{T}} \beta_{\mathrm{F}}) g(z_i) \tag{5}$$

for the multiplicative case, while for the additive case

$$y_i = (\beta_0 + \eta_{\mathrm{M}i}^{\mathrm{T}} \beta_M + \eta_{\mathrm{F}i}^{\mathrm{T}} \beta_{\mathrm{F}}) + m_i g(z_i), \tag{6}$$

where $\eta_{\mathrm{M}j}$ and $\eta_{\mathrm{F}j}$ are $a_{\mathrm{M}} \times 1$ and $a_{\mathrm{F}} \times 1$ vectors of counts of the number of males and females in household $i$ falling into each age category, respectively. Interpretation of $\beta$ values suggests that they should actually resemble a smooth curve. Therefore, they should be smoothed and a natural inference tool can be the penalised least-squares criterion, which in the case of the multiplicative model is

$$\sum_{i=1}^{n} q_i (y_i - x_i^{\mathrm{T}} \beta g(z_i))^2 + \beta^{\mathrm{T}} W^{\mathrm{T}} W \beta, \tag{7}$$

where $x_i = (1, \eta_{\mathrm{M}i}, \eta_{\mathrm{F}i})((a_{\mathrm{M}} + a_{\mathrm{F}} + 1) \times 1)$ contains the number of male and female members of household $i$ at each of the $a_{\mathrm{M}}$ and $a_{\mathrm{F}}$ ages, respectively, plus a constant. Model parameters $\beta^{\mathrm{T}} = (\beta_0, \beta_{\mathrm{M}}^{\mathrm{T}}, \beta_{\mathrm{F}}^{\mathrm{T}})$ with $\beta_{\mathrm{M}}(a_{\mathrm{M}} \times 1)$ and $\beta_{\mathrm{F}}(a_{\mathrm{F}} \times 1)$ represent mean individual availability for males and females, respectively, while $\beta_0$ is the constant of the model allowing for availability not taken into account from individual characteristics and $q_i$ are specified quantities giving different weights to households.

The last term in (7) corresponds to smoothing parameters $\beta$ with

$$W = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda_{\mathrm{M}} A_{a_{\mathrm{M}}} & 0 \\ 0 & 0 & \lambda_{\mathrm{F}} A_{a_{\mathrm{F}}} \end{bmatrix} \tag{8}$$

and

$$A_s = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}, \tag{9}$$

with $A_s((s-2) \times s)$ being a matrix of second differences. Function $g(z_i)$ is that part of individual availability assigned to household characteristics. Chesher[2] considered

$$g(z_i) = \exp\{z_i^{\mathrm{T}} \gamma_i\} \tag{10}$$

with $\gamma_i(a_\gamma \times 1)$ being a vector of unknown parameters. On the other hand, considering the additive model, the penalised formula (7) becomes

$$\sum_{i=1}^{n} q_i (y_i - x_i^{\mathrm{T}} \beta - m_i g(z_i))^2 + \beta^{\mathrm{T}} W^{\mathrm{T}} W \beta \tag{11}$$

and, by extending the discretisation argument to $g(z_i)$, we can express it as

$$g(z_i) = u_i^{\mathrm{T}} \gamma, \tag{12}$$

where $u_i(a_\gamma \times 1)$ contains dummy variables for the expression of qualitative or quantitative household characteristics. The coefficients $\gamma^{\mathrm{T}} = (\gamma_d^{\mathrm{T}}, \gamma_c^{\mathrm{T}})$ that correspond to the latter can be smoothed by using a quadratic expression similar to the last term in (11) and thus the penalised least-squares criterion is equivalent to

$$\sum_{i=1}^{n} q_i (y_i - x_i^{\mathrm{T}} \beta - m_i g(z_i))^2 + \beta^{\mathrm{T}} W^{\mathrm{T}} W \beta + \lambda_\gamma^2 \gamma_c^{\mathrm{T}} A_{a_{\gamma_c}}^{\mathrm{T}} A_{a_{\gamma_c}} \gamma_c. \tag{13}$$

Both criteria can be written in concise form as

$$R(\theta) = (y - G(\theta))^{\mathrm{T}} Q (y - G(\theta)) + \theta^{\mathrm{T}} V_\lambda^{\mathrm{T}} V_\lambda \theta, \tag{14}$$

where $y$ is an $n \times 1$ vector of observations, $G(\theta)$ is an $n \times 1$ vector representing the expected household availability as a function of parameters $\theta$, $Q = \mathrm{diag}_{i=1,\ldots,n}\{q_i\}$ is an $n \times n$ diagonal matrix containing the quantities $q_i$ and $V_\lambda$ is a block diagonal matrix containing matrices representing second differences as blocks depending on a smoothing parameter vector $\lambda$. For example, in the multiplicative

https://doi.org/10.1079/PHN2001206 Published online by Cambridge University Press

model,

$$G(\theta) = \begin{bmatrix} x_1^{\mathrm{T}}\beta \exp(z_1^{\mathrm{T}}\gamma) \\ \vdots \\ x_n^{\mathrm{T}}\beta \exp(z_n^{\mathrm{T}}\gamma) \end{bmatrix}, \quad V_\lambda = W. \tag{15}$$

In the additive model, on the other hand,

$$G(\theta) = Z\theta, \quad V_\lambda = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_{\mathrm{M}} A_{a_{\mathrm{M}}} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{\mathrm{F}} A_{a_{\mathrm{F}}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_\gamma A_{a_\gamma} \end{bmatrix} \tag{16}$$

with

$$Z = \begin{bmatrix} x_1^{\mathrm{T}} & m_1 u_1^{\mathrm{T}} \\ \vdots & \vdots \\ x_n^{\mathrm{T}} & m_n u_n^{\mathrm{T}} \end{bmatrix}. \tag{17}$$

In both cases, $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})$. Estimation in the multiplicative model requires non-linear optimisation. Details of the estimation and the variance of the estimated $\theta$ can be found in Chesher[2] and Vasdekis *et al.*[5]. In the additive case, it is straightforward that

$$\vartheta_\lambda = (Z^{\mathrm{T}}QZ + V_\lambda^{\mathrm{T}}V_\lambda)^{-1}Z^{\mathrm{T}}Qy, \tag{18}$$

which is a biased estimator of $\vartheta$ with

$$V(\theta) = (Z^{\mathrm{T}}QZ + V_\lambda^{\mathrm{T}}V_\lambda)^{-1}Z^{\mathrm{T}}QZ^{\mathrm{T}}Q\Sigma QZ(Z^{\mathrm{T}}QZ + V_\lambda^{\mathrm{T}}V_\lambda)^{-1}. \tag{19}$$

This is a robust estimate of the covariance matrix of observations as suggested by White[6] with $\Sigma$ being a diagonal matrix obtained by setting the diagonal entries equal to the squared residual of each household. Estimates of variance can be used for the derivation of confidence intervals.

Chesher[2] obtained 95% pointwise intervals that are not confidence intervals in a strict sense but give an idea of the mean variability of the availability curves. These are given by

$$\vartheta_{\lambda i} \pm 1.96\sqrt{V_{ii}(\vartheta_\lambda)}, \quad i = 1, \ldots, a_{\mathrm{M}} + a_{\mathrm{F}} + a_\gamma + 1. \tag{20}$$

The degree of bias depends on the value of smoothing parameter, which is unknown and has to be estimated. In recent years, two approaches have been used. One is to subjectively choose the value of the smoothing parameter according to possible previous ideas on the degree of smoothing. This choice is encouraged by Green and Silverman[7] and Silverman[8], and it certainly is a reasonable solution provided that a few testing runs have shaped a good idea about the value of the smoothing parameter. The other, to some extent opposing, view is that there is a need for an automatic method whereby the smoothing parameter value is chosen by the data. It is better to use the word 'automatic' rather than 'objective' for such a method. There are a number of different automatic procedures available. The best known is the generalised cross-validation[7], which, for the additive model, is equivalent to minimising

$$\mathrm{GCV}(\lambda) = n^{-1}\frac{\mathrm{RSS}}{(1 - n^{-1}\mathrm{tr}\ Z^{\mathrm{T}}QZ(Z^{\mathrm{T}}QZ + V_\lambda^{\mathrm{T}}V_\lambda)^{-1})^2}, \tag{21}$$

where RSS is the residual sum of squares from the model fit. Values of $\lambda$ equal to zero or around zero are equivalent to no smoothing, the resulting estimates are ordinary weighted least-squares estimates and their graph is very noisy. Moderate values of $\lambda$ correspond to moderate smoothing. The largest value, however, makes model parameters be over-smoothed and for $\lambda \to \infty$ they form a straight line.

## Acknowledgement

The methodological approach summarised in this paper was developed in the context of the FAIR-97-3096 project (Tasks 1–3) of the European Union.

## References

1 Engle RF, Granger CWJ, Rice JA, Weiss A. Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statistical Assoc.* 1986; **81**: 310–20.
2 Chesher A. Diet revealed? Semiparametric estimation of nutrient intake–age relationships. *J. Roy. Statistical Soc. A* 1997; **160**: 389–428.
3 Vasdekis VGS, Trichopoulou A. Nonparametric estimation of individual food availability along with bootstrap confidence intervals in household budget surveys. *Statistics Probability Lett.* 2000; **46**: 337–45.
4 Mills A, Tyler H. *Food and Nutrient Intakes of British Infants Aged 6–12 Months*. London: Her Majesty's Stationery Office, 1992.
5 Vasdekis VGS, Naska A, Trichopoulou A. Modelling household budget survey availability data and model selection. *Statistics and Computing* [submitted for publication].
6 White H. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 1980; **48**: 817–38.
7 Green PJ, Silverman BW. *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall, 1994.
8 Silverman BW. Discussion of the paper by A. Chesher. *J. Roy. Statistical Soc. A* 1997; **160**: 423–4.

https://doi.org/10.1079/PHN2001206 Published online by Cambridge University Press