An Analysis of Housing Expenditure

Using Semiparametric Cross-Section Models

by E. Charlier, B. Melenberg and A. van Soest[1]

Tilburg University
Department of Econometrics
P.O. Box 90153
5000 LE, Tilburg
The Netherlands
E-mail first author: E.Charlier@kub.nl

February 1997

Abstract

In this paper we model expenditure on housing for owners and renters by means of endogenous switching regression models using cross-section data. We explain the share of housing in total expenditure from family characteristics and total expenditure, where the latter is allowed to be endogenous. We apply various existing parametric and semiparametric techniques for cross-section data. Exogeneity of total expenditure is rejected for the parametric models but not for most semiparametric models. The results are compared on the basis of graphs of the estimated relationship between the budget share spent on housing and the logarithm of total expenditure and on the baisi of budget elasticities.

Keywords: sample selection, Engel curves, semiparametric cross-section models

JEL classification: C14, R21

[1] We thank Marno Verbeek and Roger Klein for helpful comments. Research of the third author was possible due to a fellowship of the Netherlands Royal Academy of Arts and Sciences (KNAW). We are grateful to Statistics Netherlands (CBS) for providing the data.

1. Introduction

Housing is one of the main categories of household expenditure in most industrialized countries. For analyzing household consumption its understanding therefore is crucial. The decision how much to spend on housing is strongly related to the choice between renting and owning. The standard reference is Lee and Trost (1978), who explain annual family expenditure on housing taking the decision to own or to rent explicitly into account, using cross-section data. Their model is a switching regression model with endogenous switching and normally distributed error terms, which is also referred to as Tobit V by Amemiya (1984).

Several authors have focused on different aspects in the demand for housing. Zorn (1993) models the fact that some households cannot obtain a mortgage due to mortgage constraints which results in a kinked budget set. Haurin (1991) investigates the same issue as Zorn (mortgage constraints) and analyzes how the intertemporal variation in income affects tenure choice. Ioannides and Rosental (1994) analyze the choice between renting and owning in relation to consumption and investment demand for housing.

In this paper we focus on housing expenditure and thus not on housing assets, housing equity or mortgage constraints. We will combine the model by Lee and Trost (1978), henceforth referred to as LT model, with the consumer demand literature on expenditure on goods. We will mainly concentrate our attention on two issues. First of all, the (strong) distributional assumptions in the LT model might be violated. Therefore, we investigate the consequences of relaxing these distributional assumptions, applying semiparametric estimators which have recently become available in the literature. Although a variety of such estimators has been developed (see Powell (1994) for an overview), applications are still scarce. Experience with comparing parametric and semiparametric estimates in practical examples should show whether using semiparametric techniques instead of standard parametric methods is worthwhile. The main goal of this paper is to provide such a practical example. Second, when modelling the budget share spent on housing as a function of total expenditure, account has to be taken of the possibility of endogeneity of total expenditure. We test for this and present estimates allowing for it.

In this paper we consider parametric and semiparametric cross-section models, which are tested and estimated using the 1987-wave of the Dutch Socio Economic Panel. The remainder of this paper is organized as follows. In section 2 we describe the data. In section 3 we discuss various parametric and semiparametric cross-section models and their estimates. Section 4 concludes.

2. Data

We use data from the 1987 wave of the Dutch Socio-Economic Panel (SEP). These data are a cleaned subsample with information on family characteristics (including marital status, number of children living with the family, age of the head of household, education level and region of residence), and labour market characteristics (including hours of work, gross and net wage and benefits). The labour market characteristics are used to construct household income which consists of labour earnings, other family income (mainly from letting rooms or child allowances), benefits and pensions. Personal income of children is excluded. Asset income and capital gains are also excluded, since this type of income may depend on the home ownership decision instead of determining it. Wealth data[2] are used to construct savings.[3] For issues on cleaning the savings data we refer to Camphuis (1993). Income and savings are used to construct total expenditure. Expenditure and income are reported in *Dutch guilders per month*.

The budget share spent on housing is defined as the fraction of total expenditure spent on housing. Housing expenditure for renters is the amount of money spent on rent by the family (i.e., excluding gas/water/electricity/heating as well as rental subsidy). For owners expenditure on housing consists of the following components: net interest costs on the mortgage,[4] net rent paid if the land is not owned, taxes on owned housing,[5] costs of insuring the house, opportunity costs of housing equity, maintenance costs, and minus the increase of the value of the house. The latter three costs components are not observed in the data. The opportunity cost reflecting the foregone interest on housing equity, is set equal to 4% of the value of the house minus the mortgage value. Maintenance costs and the increase of the value of the house are set equal to 2% and 1% of the value of the house, respectively. In Appendix A, we shall investigate the sensitivity of the results with respect to these choices. It appears that our main results are hardly affected.

A separate Data Appendix contains some further details on the construction of the sample and the variables of interest. For estimating the expenditure equations, we further excluded from this sample households with a missing observation for expenditure, and a few households with housing

---

[2] Net wealth is constructed using checking accounts, savings and deposits accounts, saving certificates, certificates of deposits, bonds and mortgage bonds, shares, options and other securities, antiques, jewels, coins etc., real estate other than the own residence, own car, claims against private persons, other assets, life-insurance with saving elements, personal loan or revolving credit, hire-purchase and other loans.

[3] We also corrected for donations, bequests, and capital gains.

[4] Mortgage interest payments are tax deductible. See Data Appendix for computation of the marginal tax rate.

[5] This refers to a direct tax on housing property and to extra income tax due to adding the imputed rental value of the house to household income.

budget share larger than 3.[6] This reduces the dataset from 3006 to 2357 observations. Variable definitions and summary statistics are presented in table 1. The average budget share of housing is approximately 0.24 for renters and about 0.22 for owners. Owners are higher educated and have a substantially higher average income than renters. Their average total expenditure is also higher than that of renters. Thus, in absolute terms, owners spend more on housing than renters, in spite of the lower average budget share.

In figure 1, nonparametric density estimates for the budget shares BS0 for renters and BS1 for owners are reported, as well as nonparametric regressions of these budget shares on log(total expenditure). Both budget share distributions are skewed to the right. Some budget shares larger than one are observed (see footnote 6). The nonparametric regression estimates suggest that the housing budget share is nonlinear in log(total expenditure), but can be approximated reasonably well by a quadratic function. This is similar to what Banks et al. (1994) find for many commodity groups.

In figure 2, the result of a nonparametric regression of the probability of owning a house as a function of log(total income) is presented together with the frequency distribution of log(total income). Families with higher total income tend to have a higher probability of owning a house for the main part of the income range.

3. Models

We aim at estimating Engel curves for housing expenditure. Following Banks et al. (1994), we will estimate Engel curves derived from an Integrable Quadratic Almost Ideal Demand System (IQUAIDS), relating the budget share spent on housing to the log of total expenditure and its square, and to household characteristics (taste shifters). This quadratic specification is also suggested by figure 1. It will be used in all our models. Following Lee and Trost (1978), we allow Engel curves of renters and owners to be different, and take account of endogeneity of the decision to own or to rent. This leads to the following cross-section model (assumptions on the distribution of the error terms are given below):

$$d_i = 1(\pi' x_i - u_i \geq 0)$$
$$y_{0i} = \beta_0' x_i + \varepsilon_{0i} \quad \text{if} \quad d_i=0$$
$$y_{1i} = \beta_1' x_i + \varepsilon_{1i} \quad \text{if} \quad d_i=1$$

[6] Some budget shares are larger than one, possibly due to the fact that total expenditure is constructed from income minus savings, which might lead to substantial measurement errors for some households.

Here the index i refers to household i, d_i is a sector selection dummy variable which is 0 for renters and 1 for owners, x_i is a vector of explanatory variables (log total expenditure and its square, and taste shifters), y_{0i} and y_{1i} are the budget shares spent on housing for renters and owners, respectively, β_1, β_0 and π are vectors of unknown parameters, and ε_{0i}, ε_{1i} and u_i are error terms. This model is known as a switching regression model with endogenous switching. It is labelled Tobit V by Amemiya (1984), if the error terms are trivariate normal and independent of the covariates.

Even if the error terms are independent of the regressors, without strong distributional assumptions, identification of the parameters of this model requires that at least one component of both β_1 and β_0 is equal to zero (possibly the same), while the corresponding components of π are not equal to zero. Such exclusion restrictions are not required with distributional assumptions like normality of the errors terms, as in the original Lee and Trost (1978) study. But identification then stems from the normality assumption. Since we also want to estimate semiparametric models using the same explanatory variables, exclusion restrictions on the budget share equation will be imposed throughout. Our main exclusion restriction is that the head of household's education level is not included in the budget share equations. Education level may affect the family's information set and interest in financial matters, and may therefore influence the family's portfolio choice, of which the choice between owning and renting is an important component. It is not clear, however, why education would have a direct impact on housing consumption, given the ownership decision. Another variable which we exclude from the share equations is the number of children. Although there seems no *a priori* reason for this, the number of children was always insignificant in the share equations at any conventional significance level.

As mentioned above, x_i will include the log of total expenditure and its square, which might be endogenous. For example, in the two-stage budgeting literature[7] a household first decides how much to spend in total in each period and, given this decision, it decides how much of this to spend on food, clothing, housing, etc. Thus, total expenditure per period is a decision variable. In the standard model where error terms arise due to future uncertainty only, total expenditure is exogenous to the share equations. However, introducing random preferences in a life-cycle consistent way will lead to a model in which the resulting error term is correlated with total expenditure and hence total expenditure is endogenous.

For simplicity and because we are mainly interested in the share equations we do not include the log of total expenditure and its square in the selection equation. Instead, this equation includes the log of household income and its square, which can be seen as instruments for the total

[7] See Blundell and Walker (1986), for example.

expenditure variables. If one is particularly interested in the coefficients related to the log of total expenditure and its square one could employ the two-step estimation procedure described in Lee (1996).

We write $x_i=(x_{ai}',x_{bi}',x_{di}')'$, with x_{ai} containing the log of total expenditure and its square, x_{bi} containing the log of household income and its square, and x_{di} containing the remaining variables. Define a subvector x_{ci} of x_{di} so that the model can be written as

$$d_i = 1(\pi_b'x_{bi} + \pi_d'x_{di} - u_i \geq 0)$$
$$y_{pi} = \beta_{pa}'x_{ai} + \beta_{pc}'x_{ci} + \varepsilon_{pi} \text{ if } d_i=p, \quad p=0,1$$

Throughout the paper, we assume that $(\varepsilon_{0i},\varepsilon_{1i},u_i)$ is independent of (x_{bi},x_{di}). x_{ai} will be allowed to be correlated with the error terms and x_{bi} will be used as a vector of instruments for x_{ai}. Together with the identification discussion given above, this implies that x_{di} should contain elements that are not in x_{ci}.

To estimate this model, note that

$$E\{y_{pi} - \beta_{pa}'x_{ai} | x_{bi}, x_{di}, d_i=p\} = \beta_{pc}'x_{ci} + E\{\varepsilon_{pi} | x_{bi}, x_{di}, d_i=p\}, \quad p=0,1.$$

This can be rewritten as

$$y_{pi} = \beta_{pa}'x_{ai} + \beta_{pc}'x_{ci} + g_p(x_{bi},x_{di}) + \tilde{\varepsilon}_{pi}, \quad \text{with}$$
$$g_p(x_{bi},x_{di})=E\{\varepsilon_{pi} | x_{bi}, x_{di}, d_i=p\} \text{ and } E\{\tilde{\varepsilon}_{pi} | x_{bi}, x_{di}, d_i=p\} = 0, \quad p=0,1.$$

The distributional assumptions with respect to ε_{0i}, ε_{1i} and u_i determine the functions g_0 and g_1 and, as a consequence, the way to estimate the parameters. Various estimation procedures exist in the literature. We discuss those which we will apply.

(i) Parametric model

The parametric model imposes multivariate normality of $(\varepsilon_{0i},\varepsilon_{1i},u_i)$ with a zero mean vector. This implies that g_0 and g_1 are given by $\sigma_{0u}\lambda_0(\pi_b'x_{bi}+\pi_d'x_{di})$ and $\sigma_{1u}\lambda_1(\pi_b'x_{bi}+\pi_d'x_{di})$, respectively, where λ_0 and λ_1 are the inverse Mill's ratios and $\sigma_{pu}=Cov\{\varepsilon_{pi},u_i\}$, p=0,1. To estimate the parameters, first the probit selection equation is estimated by Maximum Likelihood (ML). Then λ_0 and λ_1 are evaluated at the probit estimates. If x_a is exogenous, Ordinary Least Squares (OLS) can then be applied to the budget share equations including the estimated λ_0 and λ_1 as additional regressors. This results

in two stage estimates for β_0 and β_1. These are consistent but not asymptotically efficient. Efficiency can be obtained by applying ML, using the two stage estimates as starting values. Significance of σ_{0u} and σ_{1u} would imply that selection influences the budget shares and that correcting for selectivity bias is necessary.

To allow for endogeneity of x_{ai}, we assume that, in case of the parametric specification, the two endogenous variables in x_{ai} depend linearly on x_{bi} and x_{di} and an error term independent of x_{bi} and x_{di}. Moreover, we assume joint normality of all error terms. ML will then be performed on the complete system of five equations. Endogeneity of x_{ai} will be tested using a Lagrange Multiplier test, for the null hypothesis of zero correlation between the errors in the two auxiliary equations and the errors in the two budget share equations.

ML estimates for the parametric models for 1987 are presented in table 2. In the third and fourth column we present the results for the model in which LEXP and L2EXP are assumed to be exogenous. The following findings are significant at the 5% level. The probability of owning a house is, *ceteris paribus*:

- increasing with log(income);
- increasing with education level;
- increasing with age up to age 46, decreasing thereafter. This corresponds to CBS figures which do not control for other characteristics;[8]
- higher for married than for unmarried people;
- increasing with the number of children living with the family;
- lower in the west of the Netherlands than in other regions. The west is the region where population density and industrial concentration is largest, and where house prices are higher than in other regions.

We conclude that the budget share for owners is

- increasing up to age 44, decreasing after that age;
- decreasing in log(total expenditure) for all relevant values of log(total expenditure);
- higher for households with married head than for others;
- lower in the north of the Netherlands than in other regions.

The budget share for renters is:

- decreasing in log(total expenditure) for nearly all relevant values of log(total expenditure);
- lower for married than for unmarried people.

[8] According to CBS (1987), the fraction of house owners increases from 0.13 when the head of the household is below 25 years of age, up to 0.57 for the age category 40-45, and then decreases to 0.27 for heads of households aged 75 and over.

The covariances of the error terms are highly significant, and imply estimated correlation coefficients $\hat{\rho}_{1u}=\hat{\sigma}_{1u}/\sqrt{\hat{\sigma}_{11}}=0.26$, and $\hat{\rho}_{0u}=0.96$. This implies that selection matters in this model.

The model might be misspecified due to endogeneity of LEXP and L2EXP. The value of the LM test statistic (described above) was 88.6, exceeding the critical value of the χ^2_4 distribution at any conventional significance level. Thus, the null of exogeneity of LEXP and L2EXP is strongly rejected.

The ML estimates allowing for endogeneity of LEXP and L2EXP are presented in the fifth and sixth column of table 2. In the budget share equations, the constant term and the coefficients of LEXP and L2EXP change significantly compared to the third column. For renters, the coefficients related to AGE, AGE2 and DREG2 are now significant. In the selection equation the coefficients related to the constant term, LINC and L2INC changed significantly. The results for the additional equations for LEXP and L2EXP are not reported and are available upon request. The coefficients related to the variables LINC, L2INC and DMAR are strongly significant in these equations.

(ii) Semiparametric model

We first consider the computationally easy approach of Newey (1988). He uses the fact that the independence assumption implies that the distribution of $(\varepsilon_{0i},\varepsilon_{1i},u_i)'$ depends on (x_{bi},x_{di}) only through the index $\pi_b'x_{bi}+\pi_d'x_{di}$. The functions g_0 and g_1 can then be written as

$$g_p(x_{bi},x_{di}) = \tilde{g}_p(\pi_b'x_{bi}+\pi_d'x_{di}), \quad p=0,1.$$

To estimate the budget equations, $\tilde{g}_0$ and $\tilde{g}_1$ are approximated by $\Sigma_{k=0}^{K}\alpha_{pk}(\pi_b'x_{bi}+\pi_d'x_{di})^k$, p=0,1, with K=K(p,n) (p=0,1, n the number of observations). The following regression equations can now be used for the subsamples of renters and owners separately

$$y_{pi} = \beta_{pa}'x_{ai} + \beta_{pc}'x_{ci} + \Sigma_{k=0}^{K}\alpha_{pk}(\hat{\pi}_b'x_{bi}+\hat{\pi}_d'x_{di})^k + \hat{\varepsilon}_{pi}, \qquad (1)$$

where $\hat{\pi}_b$ and $\hat{\pi}_d$ denote estimates of π_b and π_d, respectively (to be discussed later). If x_a is exogenous, consistent and asymptotically normal estimates for $(\beta_{0a}',\beta_{0c}')$ and $(\beta_{1a}',\beta_{1c}')$ can be obtained by applying OLS to equation (1) for each subsample. This was shown by Newey (1988), who also derives an estimator for the asymptotic covariance matrices of the estimators.

We apply Newey's procedure to the case that x_{ai} is allowed to be endogenous by replacing OLS with IV. Denote the regressors in equation (1) corresponding to the case p=1 by $\hat{x}_i^s$, i.e. $\hat{x}_i^s=(x_{ai}',x_{ci}',1,(\hat{\pi}_b'x_{bi}+\hat{\pi}_d'x_{di})^1,..,(\hat{\pi}_b'x_{bi}+\hat{\pi}_d'x_{di})^K)'$ (with now K=K(1,n)) and let $\hat{X}^s=(\hat{x}_1^s,..,\hat{x}_{n1}^s)'$ where n1 is

the number of observations with $d_i=1$. Furthermore, let $\hat{w}_i^s$ be the vector of instruments, i.e. $\hat{x}_i^s$ with x_{ai} replaced by x_{bi} (hence $\hat{w}_i^s$ is of the same dimension as $\hat{x}_i^s$), and let $\hat{W}^s=[\hat{w}_1^s,..,\hat{w}_{n1}^s]'$. The parameters β_{1a}, β_{1c}, and α_{11} to α_{1K} can now be estimated by applying IV to equation (1). Under appropriate regularity conditions[9] the IV-estimates for β_{1a} and β_{1b} will be consistent and asymptotically normal: $\sqrt{n}[(\hat{\beta}_{1a}',\hat{\beta}_{1c}')'-(\beta_{1a}',\beta_{1c}')'] \rightarrow^d N(0,V)$. Notice, however, that the constant term in the regression equation cannot be estimated separately, since the series approximation also includes a constant term.[10] The asymptotic covariance matrix V can be estimated consistently by

$$[I,0]\,(\hat{W}^{s\prime}\hat{X}^s)^{-1}\left\{\sum_{i=1}^{n1}\hat{w}_i^s\hat{w}_i^{s\prime}\tilde{e}_i^2 + \hat{H}_W\hat{V}(\hat{\pi}_b,\hat{\pi}_d)\hat{H}_W'\right\}(\hat{X}^{s\prime}\hat{W}^s)^{-1}[I,0]'$$

where $\tilde{e}_i$ is the IV residual and

$$\hat{H}_W = \sum_{i=1}^{n1}\left\{\hat{w}_i^s(x_{bi}',x_{di}')\left(\sum_{k=1}^{K}k\hat{\alpha}_{1k}(\hat{\pi}_b'x_{bi}+\hat{\pi}_d'x_{di})^{k-1}\right)\right\}$$

where $\hat{\alpha}_{1k}$, k=1,..,K, are the IV estimates of the α_{1k}-s. The expressions in Newey (1988) are a special case with $\hat{W}^s$ replaced by $\hat{X}^s$, $\tilde{e}_i$ by the OLS residuals and $\hat{\alpha}_{1k}$, k=1,..,K, by the OLS estimates. The parameters in the other equation (p=0) can be estimated analogously.

The smoothing parameter in the estimation procedure is the number of terms in the series approximation, which is chosen such that adding more terms no longer affects the estimates of the regression coefficients. In practice, often only a few terms in the series approximation turn out to be required.

An alternative semiparametric approach is given by Ahn and Powell (1993). They assume that

$$g_p(x_{bi},x_{di}) = \Theta_p(f_{pi}), \text{ with } f_{pi}=P\{d_i=p\,|\,x_{bi},x_{di}\},\ p=0,1,$$

for continuous functions Θ_p, p=0,1. To estimate the share equation for the subsample of owners ($d_i=1$, p=1), consider two observations i and j with $d_i=d_j=1$ and $f_{1i}\approx f_{1j}$. Then, using the continuity of

---

[9] Appropriate regularity conditions should include conditions guaranteeing consistency of the IV estimates of β_{1a} and β_{1c} and conditions that allow one to derive the presented limit distribution. The former conditions will be different from Newey's, since identification should now be based on moment restrictions. Given identification (and consistency) the latter conditions will be comparable to Newey's conditions.

[10] Andrews and Schafgans (1995) show how the constant term can be estimated if observations with selection probability close to one are available. Since, however, we do not have many observations with probability of ownership close to zero or one, this approach is practically infeasible for both renters and owners.

Θ_1,

$$\begin{aligned} y_{1i}-y_{1j} &= \beta_{1a}'(x_{ai}-x_{aj}) + \beta_{1c}'(x_{ci}-x_{cj}) + (\Theta_1(f_{1i})-\Theta_1(f_{1j})) + (\tilde{\varepsilon}_{1i}-\tilde{\varepsilon}_{1j}) \\ &\approx \beta_{1a}'(x_{ai}-x_{aj}) + \beta_{1c}'(x_{ci}-x_{cj}) + (\tilde{\varepsilon}_{1i}-\tilde{\varepsilon}_{1j}). \end{aligned}$$

This leads to the IV-estimator proposed by Ahn and Powell: let $x_{aci}'=(x_{ai}',x_{ci}')$, $w_i'=(x_{bi}',x_{ci}')$ and let

$$S_{wx} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \omega_{ij} (w_i - w_j) (x_{aci} - x_{acj})'$$

$$S_{wy1} = \binom{n}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \omega_{ij} (w_i - w_j) (y_{1i} - y_{1j})$$

$$\omega_{ij} = \frac{1}{s_{1n}} K\left(\frac{f_i - f_j}{s_{1n}}\right) d_i d_j$$

where s_{1n} is a smoothing parameter and K is a kernel. If f_i were observed, the IV-estimator would be $(\hat{\beta}_{1a}',\hat{\beta}_{1c}')'=(S_{wx})^{-1}S_{wy1}$. Since f_i is unobserved, it is replaced by some consistent estimate. As shown by Ahn and Powell (1993), the resulting estimator is consistent and asymptotically normal and converges at the rate $n^{½}$. Notice that, similar to the Newey estimator, the constant term cannot be estimated.

When applying Ahn and Powell's estimator we have to choose a kernel and the smoothing parameter s_{1n} (and the corresponding one for renters, p=0) to determine which observations are 'close', i.e., for which $|\hat{f}_i-\hat{f}_j|$ is small. For the kernel, we choose the standard normal density function. A convenient choice for s_{1n} is to use $c_d\hat{\sigma}_h$ where $\hat{\sigma}_h$ is the sample standard deviation of the fitted first round values $\hat{f}_i$, and c_d equals 0.1 or 0.2, see Ahn and Powell (1993). In both semiparametric models (Newey and Ahn & Powell), exogeneity of x_{ai} will be tested with a Hausman (type) test, comparing the estimates which do and do not allow for endogeneity of x_{ai}.[11]

All approaches for estimating β_p, p=0,1, require estimation of a single index binary choice model[12] to obtain estimates for $(\pi_b',\pi_d')'$. Klein and Spady (1993) have proposed an estimator which is semiparametrically efficient under weak regularity assumptions. This estimator, however,

---

[11] The limit distribution of the difference of both estimators can easily be calculated under the null hypothesis of no endogeneity. This allows us to construct a test for endogeneity based on the difference of the estimators. However, since neither of the estimators is efficient under the null, and since the behaviour of the estimators is not known under the alternative, the test is not a 'pure' Hausman test. Therefore, we call it a Hausman type test.

[12] Ahn and Powell (1993) allow for a more general model, in which the probability of ownership is estimated completely nonparametrically. Due to the large number of explanatory variables in the selection equation, such an approach is practically infeasible for our purposes.

is difficult to compute. Instead, we started with the probit ML estimates for (π_b',π_d'). We tested for normality and heteroskedasticity of exponential form using tests described in Chesher and Irish (1987). Both normality and homoskedasticity were rejected. Therefore, we experimented with the following specification, in which the single index assumption is retained:

$$P\{d_i=1 \mid x_{bi},x_{di}\} = \Phi(m(\tau,\pi_b'x_{bi}+\pi_d'x_{di})/\exp\{\sigma(\gamma,\pi_b'x_{bi}+\pi_d'x_{di})\})$$

Here m and σ are power series in $\pi_b'x_{bi}+\pi_d'x_{di}$ with coefficients τ and γ, respectively. This can be seen as a series approximation to an arbitrary single index model. Let τ_j and γ_j denote the coefficients related to $(\pi_b'x_{bi}+\pi_d'x_{di})^j$. The normalizations imposed are $\tau_0=0$, $\tau_1=1$ and $\gamma_0=0$. We estimated this model for several lengths of the two power series, and found one significant term: $(\pi_b'x_{bi}+\pi_d'x_{di})^2$ in m.

The ML results (taking the number of terms in the series expansions as fixed) with this additional term included, are presented in the lower part of table 3. τ_2, the coefficient of $(\pi_b'x_{bi}+\pi_d'x_{di})^2$, is significantly negative, but the probability $P\{d_i=1 \mid x_{bi},x_{di}\}$ increases with the index $\pi_b'x_{bi}+\pi_d'x_{di}$ over the sample range.[13]

The semiparametric estimates based upon Newey (1988) for the case that LEXP and L2EXP are assumed to be exogenous, are presented in the second and third column of (the upper part of) table 3. In the series approximation of the correction term six terms were used for owners and four for renters. These choices resulted from estimating models with up to nine terms included; the estimates did not change much after including more than six and four terms, respectively.

The estimated standard errors, which take into account the first stage estimation error in the parameters of the selection equation, appear to differ substantially from the standard OLS standard error estimates, but are similar to the Eicker-White standard errors. This indicates that the first stage errors hardly affect the standard errors of the second stage estimates.

Comparing the estimates of the slope coefficients in the share equations to their parametric counterparts in table 2, we find that for owners the coefficients related to AGE, AGE2 and DMAR changed somewhat whereas for renters the coefficients related to LEXP and L2EXP also changed somewhat. Particularly the marital status dummy in the equation for renters changed: its effect is now virtually zero, while it was significant at the 1% level in table 2.. The shape of the Engel curves as a function of total expenditure remains the same, and so do the effects of educational and

[13] Using LM tests similar to those in Chesher and Irish (1987), normality in this extended probit model could not be rejected. Homoskedasticity, however, is still rejected, suggesting that the single index specification might be inadequate. Due to the lack of feasible alternatives, however, we have to retain this assumption (see also previous footnote).

regional dummies in the selection equation.

The Ahn-Powell estimates for the same model, using (for $\hat{f}_i$) the predictions from the single index probit model in the kernel weights, are in the sixth and seventh column of table 3. The smoothing parameter in the kernel weights was set to 0.2 times the standard deviation of the first stage predictions. The Ahn-Powell estimates are similar to the Newey estimates in column two. The standard errors are not corrected for the errors in the first stage estimates, since this would be computationally too demanding. These standard errors will therefore underestimate the true standard errors. The findings for the Newey estimates, however, suggest that this problem may be negligible.

We present Newey and Ahn-Powell instrumental variables (IV) estimates, that allow for endogeneity of LEXP and L2EXP in the budget share equation, in the fourth and fifth, and eighth and ninth column, respectively, of table 3.[14] For the Newey estimates, we used series approximations of six terms for owners and five terms for renters. The smoothing parameter for the Ahn-Powell estimator are the same as for the OLS case. Using IV mainly affects the parameter estimates related to LEXP and L2EXP. Newey and Ahn-Powell estimates are again similar and a Hausman type test on the difference between them leads to the conclusion that the model cannot be rejected.

A Hausman type test on exogeneity of LEXP and L2EXP is based on the difference between the share equation estimates in columns two and four (Newey) and columns six and eight (Ahn-Powell). For the Newey estimates, the realization of the test statistic is 1.2 for owners and 12.9 for renters. This is below the critical value of a χ^2_8 distribution for any conventional level, so that exogeneity of LEXP and L2EXP can no longer be rejected. For the Ahn-Powell estimates, the realizations of the Hausman type test statistic are 1.2 and 21.4. The latter is significant at the 1% level.[15]

To check whether the semiparametric Newey estimates are significantly different from the parametric ML estimates, we compared the results for the case that endogeneity of LEXP and L2EXP is taken into account. A Hausman test was based on the difference between the two sets of estimates for $(\beta_0',\beta_1')'$. The difference between the estimated covariance matrices was not positive semidefinite, but an alternative estimator which is guaranteed to be positive definite is easily computed, see, for example, Newey (1985). The resulting test statistic is 205.12 which implies that the parametric model was strongly rejected at any conventional significance level.

---

[14] Results in Appendix A show that the results of the Newey (1988) estimates are not sensitive with respect to the definition of the expenditure measure for owners.

[15] Like the standard errors, these test statistics take no account of the first stage estimation error.

The estimated shares spent on housing as a function of LEXP are presented in figure 3. The other explanatory variables are set equal to their sample means. The bottom figures show the distribution of expenditure. The two top figures refer to the parametric estimates. They also include OLS and IV estimates not allowing for selectivity ("no sel." and "no sel end.", respectively, not presented in table 2). For owners, they also contain estimates based upon alternative definitions of housing expenditure (BS12, BS10, see Appendix A). For owners, the main difference between the shapes of the curves is that allowing for endogeneity of total expenditure leads to higher elasticities for those in the highest total expenditure quantiles. For renters, there are more substantial differences between the various curves. In particular, if selectivity and endogeneity of LEXP and L2EXP are allowed for (BS0 end), the curve is almost linear, while it is U-shaped in all other cases.

The second set of two figures refer to the semiparametric models. The constant terms are not estimated (see discussion above). Instead, we have chosen them such that the means of the predicted budget shares equal the mean of the observed budget shares. Thus, only the shapes of the curves can be compared, and not their level. For both owners and renters, we again find that allowing for endogeneity of total expenditure makes a big difference for high levels of total expenditure.

In the third set of two figures, all results for the parametric and semiparametric models, taking into account selectivity and endogeneity, are compared. The curves for the Ahn-Powell estimates are very similar to those based upon the Newey estimates. For owners, the curve for the parametric model is similar to these two. For renters, however, the difference is much larger.

Another way to evaluate and compare the results is to look at implied elasticities of housing expenditure with respect to total expenditure. In table 4 we present means of these elasticities for owners and renters separately, weighted with total household expenditure. These can be interpreted as aggregate elasticities (cf. Banks et al. (1994)). We present the means and their standard errors, and the fraction of households for which the elasticity estimate is larger than zero.[16] In most cases, the elasticities are much smaller than one, suggesting that housing is a necessity. We find large variation in the outcomes across the various models, however, including implausible negative signs in some cases. In the semiparametric models the standard errors are often quite large, so that the means are insignificantly different from zero. An exception is the significantly positive estimate according to the Ahn-Powell model ignoring endogeneity of LEXP and L2EXP. To see whether the negative sign for the elasticity in the Newey IV model is caused by an inappropriate choice of the instruments, we also replaced the instruments by the lagged values of log(household income)

[16] The median elasticities (not reported), were very close to zero in most cases.

and its square. This, however, led to similar parameter estimates as before and the elasticities for renters increased only slightly.

4. Conclusions

We have modelled expenditure on housing for both owners and renters using endogenous switching regression models, applied to cross section data. Attention is paid to the construction of the variables needed in the econometric model, especially to the definition of housing expenditure for owners. In choosing the model assumptions we are guided by both economic theory and by econometric models for which suitable estimators are available. We focused on estimation techniques which allow some of the explanatory variables in the budget share equations to be endogenous, and on application of semiparametric estimation techniques.

We have presented estimation results using both parametric and semiparametric models. We also present results taking into account the endogeneity of the variables related to total expenditure. Taking into account the endogeneity of LEXP and L2EXP mainly affects the parameter estimates related to the endogenous explanatory variables. The economic conclusions from the parameter estimates from both the parametric and the semiparametric cross-section model are similar. In terms of budget share spent on housing as a function of LEXP the results for the parametric and semiparametric models for owners are similar, but for renters the results for the parametric model differ from the results for the semiparametric models. This suggests that in the current practical example, using parametric techniques can lead to misleading outcomes if model assumptions are violated. It makes it worthwhile to use semiparametric estimation techniques instead, particularly since the extra computational effort required is limited.

This study shows that semiparametric estimation of the endogenous switching regression is practically feasible and useful. Still, this is not necessarilty the case for models with a richer economic structure. For example, the more structural model of Zorn (1993) cannot be estimated semiparametrically given the current techniques and the limitations of the data. On the other hand, a promising direction of extending semiparametric applications is to use panel data. The models estimated here are consistent with random individual effects panel data models, but not with fixed effects specifications. Application of the latter type of models to housing expenditure is considered in a companion paper, Charlier et al. (1996).

References

Ahn, H. and J.L. Powell (1993), "Semiparametric Estimation of Censored Selection Models with a

Nonparametric Selection Mechanism," *Journal of Econometrics*, 58 (1/2), 3-29.

Amemiya, T. (1984), "Tobit models: a survey," *Journal of Econometrics*, 24, 3-63.

Andrews, D. and M. Schafgans (1995), "Semiparametric estimation of a sample selection model," mimeo, Yale University.

Banks, J., R. Blundell and A. Lewbel (1994), "Quadratic Engel Curves, Indirect tax Reform and Welfare Measurement," University College London Discussion Paper 94-04.

Blundell, R. and I. Walker (1986), "A Life-Cycle Consistent Empirical Model of Labour Supply Using Cross-Section Data," *Review of Economic Studies, 53, 539-558.*

Budgethandboek NIBUD (1987), "*Gegevens Omtrent Inkomsten, Uitgaven en Bestedingspatronen van Particuliere Huishoudens*," NIBUD.

Camphuis, H. (1993), "Checking, Editing and Imputation of Wealth Data of the Netherlands Socio-Economic Panel for the period '87-'89," VSB-CentER Savings Project Discussion paper.

CBS (1987), "*Woningbehoeftenonderzoek* 1985/1986," CBS, The Hague.

Charlier, E., B. Melenberg and A. van Soest (1996), "An Analysis of Housing Expenditure using Semiparametric Models and Panel Data," Mimeo, Tilburg University.

Chesher, A. and M. Irish (1987), "Residual Analysis in the Grouped and Censored Normal Linear Model," *Journal of Econometrics*, 34, 33-61.

Euwals, R. and A. van Soest (1995), "Desired and Actual Family Labour Supply in the Netherlands", CentER Discussion Paper no. 9623, Tilburg University.

Haurin, D.R. (1991), "Income Variability, Homeownership, and Housing Demand," *Journal of Housing Economics*, 1, 60-74.

Ioannides, Y.M. and S.S. Rosental (1994), "Estimating the consumption and investment demands for housing and their effect on housing tenure status," *The Review of Economics and Statistics*, 76, 127-141.

Klein, R.W. and R.S. Spady, (1993), "An Efficient Semiparametric Estimator of the Binary Response Model," *Econometrica*, 61, 387-423.

Lee, L.F. and R.P. Trost (1978), "Estimation of Some Limited Dependent Variable Models with Application to Housing Demand," *Journal of Econometrics*, 8, 357-382.

Lee, M-J (1996), "Nonparametric Two-Stage Estimation of Simultaneous Equations with Limited Endogenous Regressors," *Econometric Theory*, 12, 305-330.

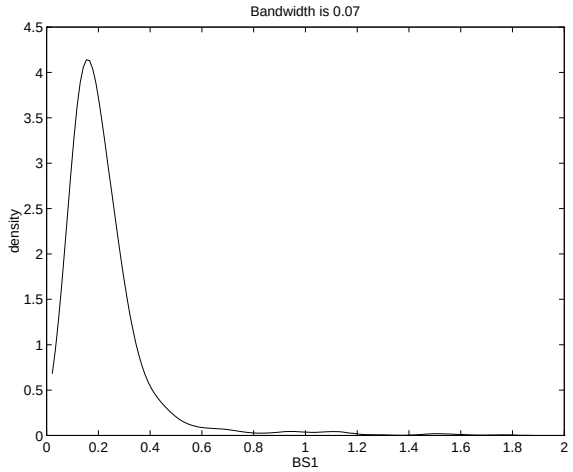
Newey, W.K. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica*, 53 (5), 1047-1070.

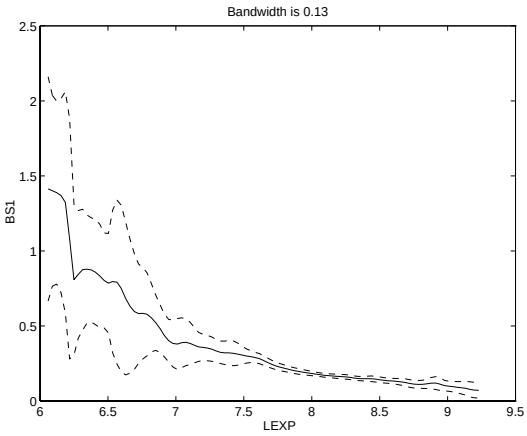
Newey, W.K. (1988), "Two Step Series Estimation of Sample Selection Models," mimeo, MIT (revised version October 1991).

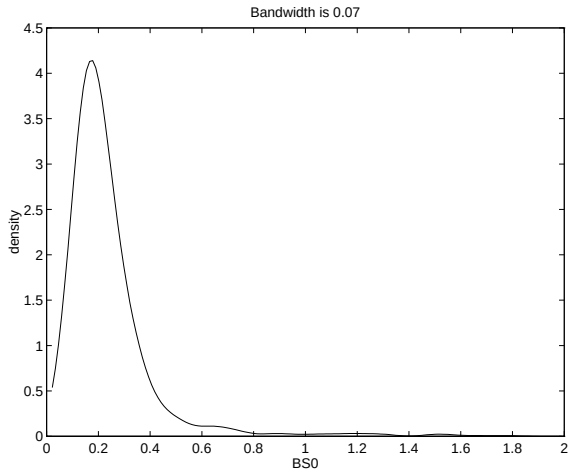
Powell, J.L. (1994), "Estimation of semiparametric models," in *Handbook of Econometrics, Volume 4*, R.F. Engle and D.L. McFadden (eds.), North-Holland, Amsterdam, 2444-2523.

Zorn, P.M. (1993), "The Impact of Mortgage Qualification Criteria on Households' Housing Decisions: An Empirical Analysis Using Microeconomic Data," *Journal of Housing Economics*, 3, 51-75.

Figure 1: *Nonparametric density estimates for BS1 and BS0 and nonparametric regression estimates of the same variables on log total expenditure (LEXP), together with 95% uniform confidence bands*







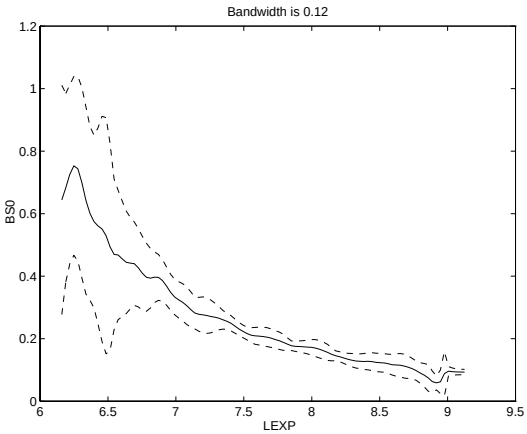
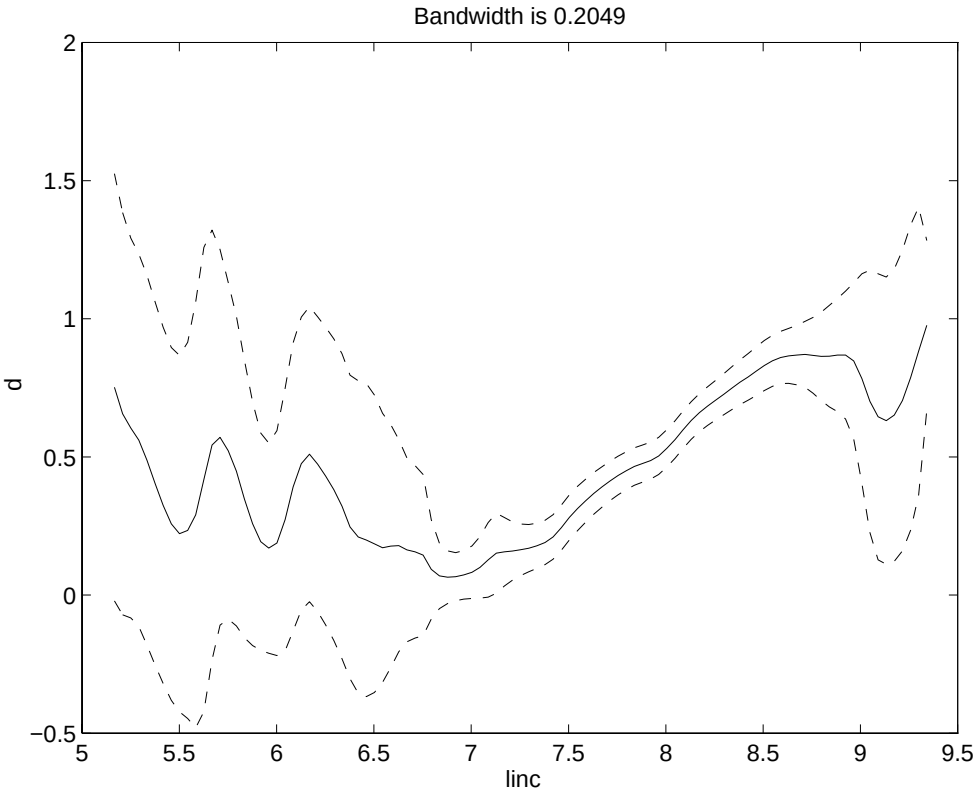


Figure 2: *Nonparametric estimates of the probability of owning a house as a function of log household income (LINC), and distribution of LINC*



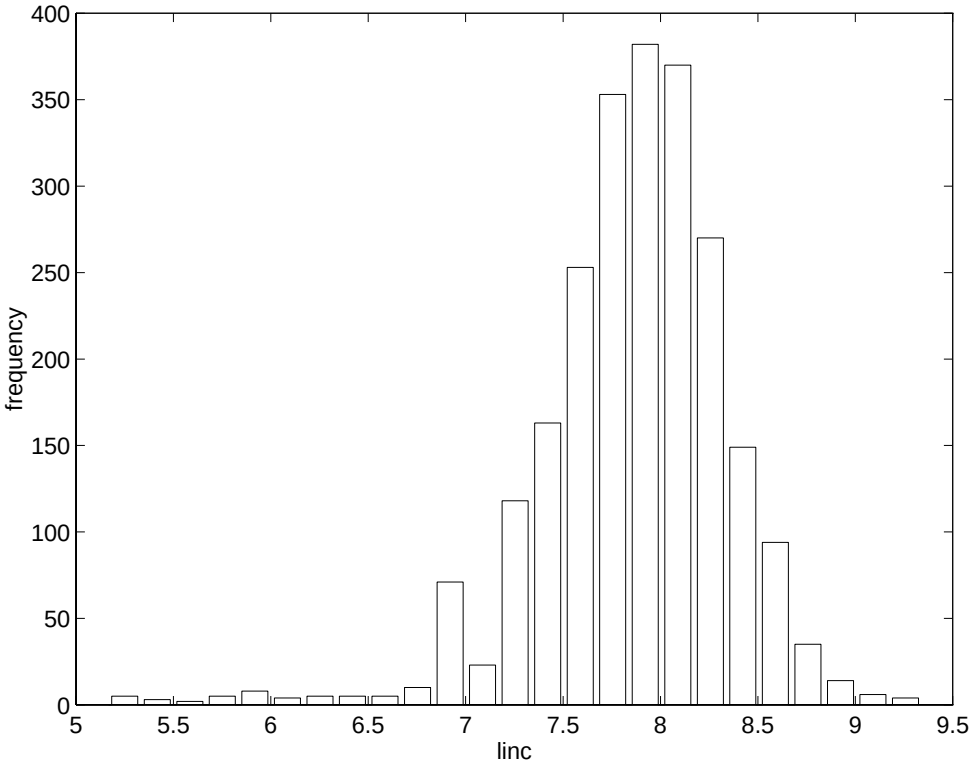
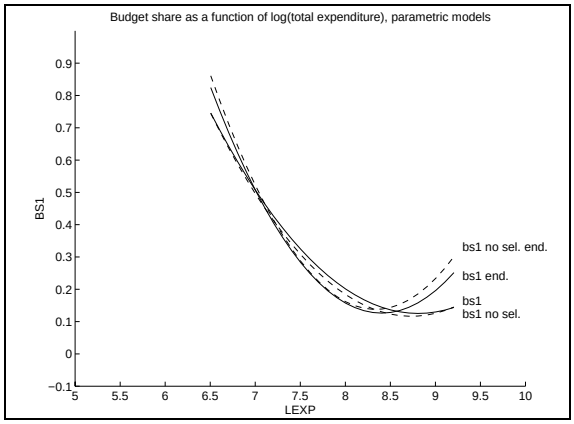
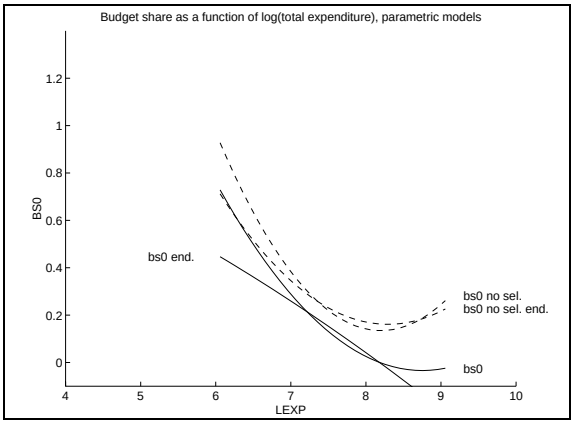
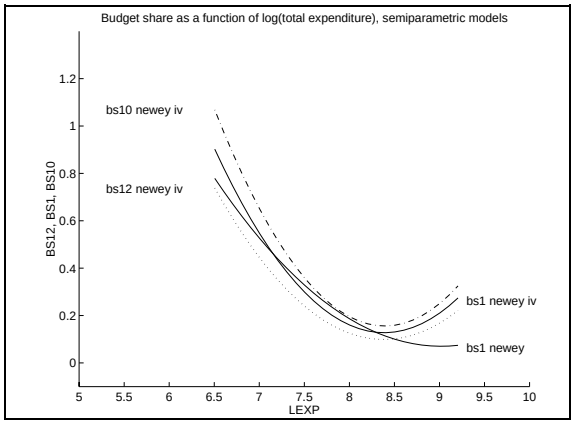


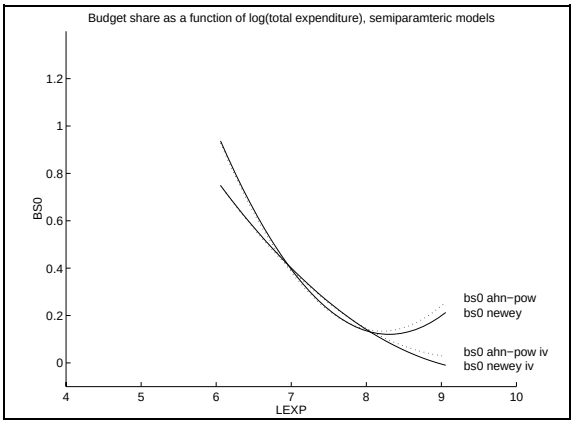
Figure 3: Budget share spent on housing as a function of LEXP

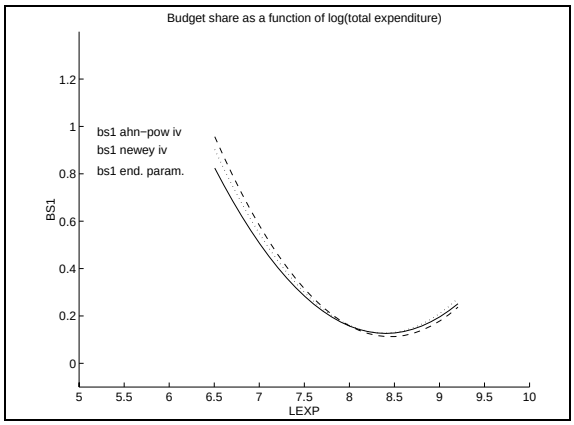
owners: renters:

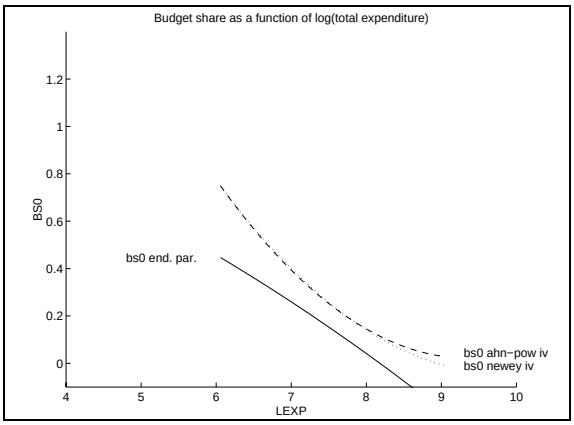


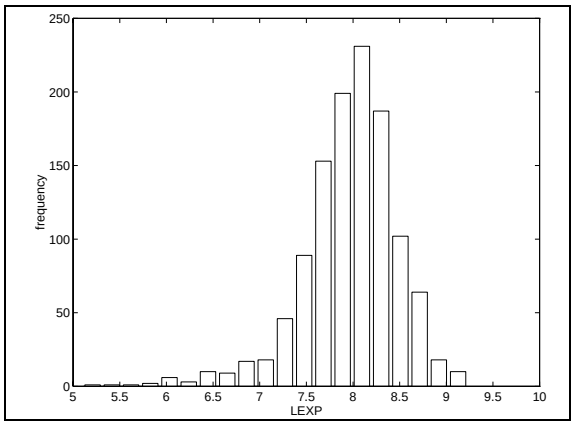












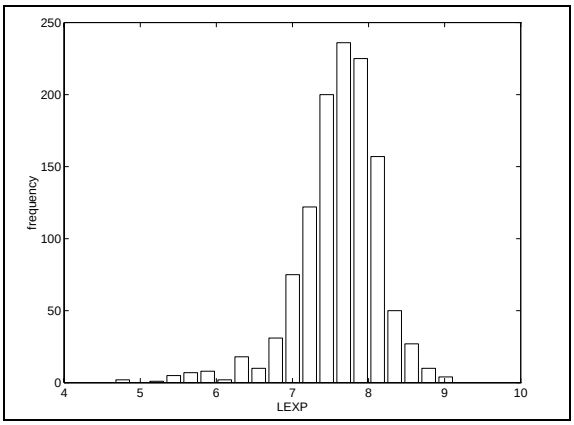


Table 1. Overview of variables and summary statistics (standard errors in parentheses)

Variable	Description	Renters	Owners
	number of obs.	1190	1167
BS0, BS1	Budget share (i.e. monthly expenditure on housing divided by monthly total expenditure)	0.24 (0.23)	0.22 (0.18)
DOP2	dummies for education level	0.24	0.15
DOP3		0.37	0.48
DOP4		0.09	0.19
DOP5		0.03	0.07
AGE	age of the head of the household in decennia	4.03 (1.21)	4.10 (0.98)
AGE2	and its square	17.64	17.78
LINC	logarithm of monthly family income and	7.69 (0.47)	8.04 (0.44)
L2INC	its square (in guilders)	59.34	64.90
EXP	monthly total family expenditure	2304 (1117)	3233 (1483)
LEXP	logarithm of monthly total family expenditure	7.62 (0.54)	7.97 (0.51)
L2EXP	and its square	58.56	63.73
DMAR	dummy for married	0.73	0.94
NCH	number of children living with the family	0.83	1.22
DREG1	region dummies for north, east and south respectively	0.11	0.11
DREG2		0.20	0.22
DREG3		0.23	0.28

Table 2. Estimation results for the cross-section parametric model (standard errors in parentheses)[a]

Equation	Variable	Exogenous LEXP and L2EXP		Endogenous LEXP and L2EXP	
BS1 owners	CONSTANT	9.018**	(0.217)	13.644**	(0.437)
	AGE	0.061*	(0.028)	0.062	(0.036)
	AGE2	−0.007*	(0.003)	−0.008*	(0.004)
	LEXP	−2.061**	(0.058)	−3.260**	(0.133)
	L2EXP	0.117**	(0.004)	0.194**	(0.010)
	DMAR	0.070**	(0.010)	0.076**	(0.022)
	DREG1	−0.041**	(0.013)	−0.037*	(0.015)
	DREG2	−0.002	(0.008)	0.002	(0.011)
	DREG3	−0.010	(0.008)	−0.007	(0.010)
BS0 renters	CONSTANT	8.147**	(0.192)	1.420*	(0.627)
	AGE	−0.053	(0.028)	−0.081**	(0.027)
	AGE2	0.006	(0.003)	0.010**	(0.003)
	LEXP	−1.838**	(0.054)	−0.068	(0.171)
	L2EXP	0.105**	(0.004)	0.010	(0.011)
	DMAR	−0.030**	(0.010)	−0.050**	(0.012)
	DREG1	−0.028	(0.017)	−0.029	(0.016)
	DREG2	−0.017	(0.011)	−0.021	(0.011)
	DREG3	−0.007	(0.011)	−0.005	(0.011)
Selection	CONSTANT	−9.511**	(1.611)	6.254	(3.285)
	DOP2	0.147*	(0.070)	0.164*	(0.078)
	DOP3	0.332**	(0.060)	0.348**	(0.067)
	DOP4	0.490**	(0.074)	0.522**	(0.084)
	DOP5	0.540**	(0.111)	0.592**	(0.134)
	AGE	0.863**	(0.201)	0.871**	(0.211)
	AGE2	−0.093**	(0.023)	−0.094**	(0.025)
	LINC	1.013*	(0.457)	−3.215**	(0.873)
	L2INC	−0.021	(0.032)	0.260**	(0.058)
	DMAR	0.663**	(0.082)	0.687**	(0.092)
	NCH	0.091**	(0.023)	0.093**	(0.025)
	DREG1	0.200*	(0.094)	0.224*	(0.096)
	DREG2	0.160*	(0.072)	0.206**	(0.074)
	DREG3	0.194**	(0.068)	0.228**	(0.070)
error	σ_{11}	0.011**	(0.0003)	0.013**	(0.0005)
distribution	σ_{00}	0.028**	(0.0006)	0.028**	(0.0008)
	σ_{1u}	0.018*	(0.008)	0.013	(0.028)
	σ_{0u}	0.161**	(0.002)	0.149**	(0.002)
	$\sigma_{\text{LEXP, BS owners}}$			0.002	(0.002)
	$\sigma_{\text{L2EXP, BS owners}}$			0.011	(0.031)
	$\sigma_{\text{LEXP, BS renters}}$			−0.006**	(0.002)
	$\sigma_{\text{L2EXP, BS renters}}$			−0.067*	(0.028)

[a] * means significant at the 5% level, ** means significant at the 1% level

Table 3. Estimation results for the cross-section semiparametric models using BS1 for owners and BS0 for renters (standard errors in parentheses)[a]

Variable	Newey[b]		Newey IV[c]		Ahn-Powell[d]		Ahn-Powell IV[c,d]	
BS owners								
CONSTANT	9.234[e]		15.454[e]		.		.	
AGE	−0.027	(0.033)	0.023	(0.066)	−0.015	(0.023)	0.021	(0.044)
AGE2	0.002	(0.004)	−0.005	(0.008)	0.001	(0.003)	−0.002	(0.006)
LEXP	−2.024**	(0.336)	−3.670	(1.939)	−2.430**	(0.252)	−3.756*	(1.398)
L2EXP	0.112**	(0.021)	0.212	(0.121)	0.138**	(0.016)	0.222*	(0.089)
DMAR	0.024	(0.022)	0.041	(0.024)	0.020	(0.018)	0.032	(0.020)
DREG1	−0.050**	(0.011)	−0.037*	(0.016)	−0.045**	(0.007)	−0.036**	(0.010)
DREG2	−0.011	(0.011)	−0.003	(0.012)	−0.018*	(0.007)	−0.014	(0.006)
DREG3	−0.020	(0.010)	−0.009	(0.014)	−0.021**	(0.007)	−0.012	(0.009)
BS renters								
CONSTANT	11.290[e]		5.589[e]		.		.	
AGE	0.017	(0.036)	−0.068	(0.042)	0.014	(0.024)	−0.063**	(0.023)
AGE2	−0.002	(0.004)	0.008	(0.004)	−0.002	(0.002)	0.007**	(0.002)
LEXP	−2.690**	(0.324)	−1.107*	(0.469)	−2.814**	(0.233)	−1.247**	(0.313)
L2EXP	0.162**	(0.021)	0.056	(0.032)	0.171**	(0.015)	0.067**	(0.021)
DMAR	−0.002	(0.016)	−0.049*	(0.018)	−0.000	(0.011)	−0.048**	(0.012)
DREG1	−0.028*	(0.011)	−0.038*	(0.016)	−0.021**	(0.007)	−0.034**	(0.010)
DREG2	−0.015	(0.012)	−0.024	(0.015)	−0.015	(0.008)	−0.023*	(0.010)
DREG3	−0.003	(0.011)	−0.003	(0.013)	−0.001	(0.007)	0.001	(0.008)
Selection								
CONSTANT	23.113**	(4.440)	«		«		«	
DOP2	0.207*	(0.083)	«					
DOP3	0.510**	(0.075)	«					
DOP4	0.599**	(0.121)	«					
DOP5	0.570**	(0.208)	«					
AGE	1.252**	(0.223)	«					
AGE2	−0.134**	(0.026)	«					
LINC	−7.991**	(1.193)	«					
L2INC	0.581**	(0.080)	«					
DMAR	0.481**	(0.088)	«					
NCH	0.051	(0.033)	«					
DREG1	0.303**	(0.094)	«					
DREG2	0.240**	(0.078)	«					
DREG3	0.302**	(0.073)	«					
τ_2	−0.172**	(0.036)	«					

[a] * means significant at the 5% level, ** means significant at the 1% level.
[b] series approximation using single index ML probit in estimating the selection equation.
[c] IV using AGE, AGE2, LINC, L2INC, DMAR, DREG1, DREG2, DREG3 as instruments
[d] standard errors **not** corrected for the first stage ML probit estimates
[e] estimates include the estimate for the constant term in the series approximation

Table 4: Budget elasticities for the cross-section models (standard errors in parentheses)[a]

	owners	fraction > 0	renters	fraction > 0
OLS (no selection)	0.232**	0.61	0.378**	0.63
	(0.049)		(0.055)	
IV (no selection)	0.637**	0.69	0.498**	0.82
	(0.089)		(0.084)	
Parametric (column 3 table 2)	0.526**	0.84	−0.237**	0.30
	(0.058)		(0.0504)	
Parametric (column 5 table 2)	0.531**	0.67	1.461**	1.00
	(0.146)		(0.095)	
Newey (table 3)	−0.036	0.37	0.246	0.56
	(0.626)		(0.572)	
Newey IV (table 3)	0.508	0.62	−0.178	0.28
	(0.592)		(0.226)	
BS12 Newey IV (appendix A)	0.498	0.59		
	(0.622)			
BS10 Newey IV (appendix A)	0.508	0.62		
	(0.574)			
Ahn-Pow. (table 3)	0.014	0.40	0.370**	0.62
	(0.061)		(0.073)	
Ahn-Pow. IV (table 3)	0.332	0.52	−0.042	0.40
	(0.322)		(0.145)	

[a] * means significant at the 5% level, ** means significant at the 1% level.

DATA APPENDIX

In this appendix we give some details on the construction of the variables used in the application. The data come from the 1986/1987 waves of the Dutch Socio Economic Panel. Since we only use wealth data from the 1986 data (to subtract from the wealth in 1987 to get savings for 1987) we do not include information on this wave. Instead we refer to Charlier et al. (1996).

Housing

Initial dataset: 3613 households for 1987.

Dropped from the analysis are:

- families that live for free;
- families with a total income below Dfl. 1,- per month;
- families that receive a so called *huurgewenningsbijdrage* (i.e., a governmental allowance for people who experienced a large rent increase because of renovation of their dwelling or who had to search for a different dwelling after pull down of their previously rented dwelling). The reason for this latter drop is that the amount is a substantial part of the housing expenditure and it is not clear from the data whether this amount is included in the answers on rent payments or not.

Housing consumption for owners:

(1-tax)**erfpacht* + tax**huurwaardeforfait* + (1-tax)*interest payment + foregone interest − increase in the value of the house + maintenance costs + *eigenaarsgedeelte onroerend goedbelasting* + *opstalverzekering*.

Here *erfpacht* is the amount of money you have to pay if you do not own the land on which your dwelling is built (which is partly deductible), tax is the marginal tax rate of the most earning adult in the household, *huurwaardeforfait* is tax levied on the value of the house of owners, *eigenaarsgedeelte onroerend goedbelasting* is municipal tax for house owners and *opstalverzekering* is a house insurance for fire, broken windows etc. Expenditure on gas/water/electricity/heating is excluded.

Computation of the variables in expenditure for owners

Some house owning families are dropped because the value of the house is not known, which is necessary to correct for, among other things, *huurwaardeforfait*. In the data we have either the amount spent on interest payments on the mortgage or the interest rate on the mortgage. If we only have the interest rate on the mortgage we computed the interest payments by multiplying this

percentage with the mortgage value. If the mortgage value is not reported we used 149000 (the average value of a house for 1987). Foregone interest is set equal to 0.04 times the difference in the value of the house and the mortgage value. Maintenance costs are defined as 2 percent of the value of the house. In the main text we investigate the sensitivity of the results with respect to the percentage increase in the value of a house and the percentage used in the maintenance costs. Because the *eigenaarsgedeelte onroerend goedbelasting* can differ per municipal it is calculated as follows: we have data over 1986-1987 on Tilburg and we will consider Tilburg to be representative for its province. Per province we have the amount of tax that was payed to the local government per inhabitant of the municipality (CBS, Statistiek der gemeentebegroting). The *eigenaarsgedeelte onroerend goedbelasting* per province is calculated as the figure for Tilburg times the relative tax per inhabitant of the province. The relative tax for the provinces is approximately constant over time. The *opstalverzekering* is simply 12.95 times the value of the house divided by 100000 (Budgethandboek NIBUD, 1987).

Computation of marginal tax rate

In the SEP we only observe net income like net wages, net unemployment benefits, net pensions etc. To calculate the marginal tax rate we need gross income of the spouse that earns most because he/she will have to report the tax related issues of owning a house (like e.g. *huurwaardeforfait*). From the net income we could try to invert the tax system and infer gross income. However, this is a very cumbersome approach. Therefore we will follow Euwals and Van Soest (1995). Gross income is already available for individuals with a payed job. We now estimate a net wage equation using the households in which at least one individual has a paid job. An important variable to be included is the tax free allowance (TFA). Constructing this for married couples involves the gross income of the other spouse. All the households for whom we could determine the TFA were included in estimation. The equation estimated is the same as in Euwals and Van Soest (1995), i.e. without a constant term. Without making differences between men and women we got an R^2 of .9955 and the parameter estimates are fairly similar. Given the net income we can now estimate gross income by inverting the relationship. By taking derivatives of net income with respect to gross income we can estimate the marginal tax rate.

General remarks concerning the data

The following data cleaning operations have been applied.

- People who got married or divorced are left out in the analysis to avoid dependence between households in the sample;

- households that spend more than 1.5 times their monthly income on housing are also left out.

In general we loose approximately 600 households. If we use only the observations with income budget shares smaller than 1.5 we end up with 3006 observations.

APPENDIX A

In this appendix we will investigate the sensitivity of the cross-section Newey IV results with respect to the maintenance costs and the mortgage costs in housing consumption for owners. Let BS1ab denote the Budget Share spent on housing for owners with a% increase of the value of a house (a=0,1,2,3,4) and b% of the value of the house as the maintenance costs (b=1,2). In the main text a equals 1 and b equals 2. From the definition of housing costs for owners it follows that BS1ab=BS1a+1,b+1 so eg. BS121=BS132. Because the averages for BS142, BS132 (and henceBS131 and BS121) are very low compared to the average for renters we only consider BS122, BS112 and BS102. The last digit is then dropped because it is fixed at 2. Hence we consider BS1a with the maintenance costs fixed at 2 % of the value of the house. BS11 is used throughout the main text. The means for BS12, BS11 and BS10 are respectively 0.18, 0.22 and 0.27 with standard errors of 0.15, 0.18 and 0.22.

In the next table we indicate the sensitivity of the parameter estimates of the Newey IV estimates with respect to the measure for housing expenditure for owners. The coefficients related to LEXP, L2EXP, DMAR and DREG1 tend to change somewhat, but the main conclusions remain the same. The standard errors remain rather large such that we do not find significant differences in the parameter estimates when varying housing expenditure for owners.

Sensitivity of the estimation results with respect to the measure for housing expenditure of owners, cross-section[a]

Variable	BS12 Newey IV[b,c]		BS11 Newey IV[b,c]		BS10 Newey IV[b,c]	
CONSTANT	7.289[d]		15.454[d]		18.108[d]	
AGE	0.028	(0.053)	0.028	(0.066)	0.035	(0.077)
AGE2	−0.004	(0.007)	−0.005	(0.008)	−0.005	(0.009)
LEXP	−3.040	(1.634)	−3.670	(1.939)	−4.300	(2.241)
L2EXP	0.181	(0.105)	0.219	(0.121)	0.256	(0.144)
DMAR	0.032	(0.020)	0.041	(0.024)	0.050	(0.028)
DREG1	−0.029[*]	(0.013)	−0.037[*]	(0.016)	−0.045[**]	(0.018)
DREG2	−0.005	(0.010)	−0.003	(0.012)	−0.001	(0.014)
DREG3	−0.010	(0.011)	−0.009	(0.014)	−0.008	(0.017)

[a] [*] means significant at the 5% level, [**] means significant at the 1% level. The results for renters and for the selection equation are the ones presented in the second and third column of table 3

[b] series approximation using single index ML probit in estimating the selection equation

[c] IV using AGE, AGE2, LINC, L2INC, DMAR, DREG1, DREG2 and DREG3 as instruments

[d] estimates include the estimate for the constant term in the series approximation