Joint Decisions in Residential Choice: Masters Thesis by Edward Johnsen

Edward Johnsen, Yiğit Sağlam and Toby Daglish*

Abstract

Economic agents frequently make joint decisions, which often require a compromise by some or all of the participants. We propose an econometric model in which groups of agents make a joint decision; each agent has preferences modelled using a combination of multi-nominal logit and conditional logit parts. We combine these marginal preferences to create a joint set of probabilities of the group making a particular choice, which enables parameter estimation by maximum likelihood. We can also make the weight applied to an individual agents preferences depend on characteristics of the agent or group. To demonstrate the use of the model, data is obtained from the New Zealand Household Travel Survey. We estimate our model to show how households might make the joint decision of where to live, given that different household members have different work locations.

*I would like to acknowledge Yiğit Sağlam and Toby Daglish for their guidance and ongoing help throughout this project.

1 Introduction

Joint decision theory has been a widely discussed topic with numerous papers from fields such as economics and psychology contributing papers that explore different aspects of the proposed topic (Abraham and Hunt (1997);Gliebe and Koppelman (2002);Bhat and Guo (2004)). Joint decision theory is about the analysis of how people within a group make decisions together and how they compromise for the overall benefit of the group.

This paper will discuss how individuals determine their residential location, by estimating their individual preferences and identifying to what extent they compromise for the welfare of the group, in this case the household. Joint decisions are often a part of important life choices and can be applicable to a range of areas including the private and public sector. For example, suppose there are two partners deciding on investment opportunities with differing opinions and interest, how do they come to a decision and what impact do choice-dependent and choice-independent information have on this decision. This paper focuses on residential choice as there is a large amount of previous literature (Pinjari et al. (2011);Timmermans et al. (1992)) and substantial data available for the estimation of the model.

This paper presents an econometric model where each household agent has preferences modelled through both conditional and multi-nominal logit parts. Using multi-nominal logit (MNL) models to represent these preferences is prevalent in this area of research, so we continue this method of representation. McFadden (1978) proposed that the rational consumer will choose their particular residential location by weighting different attributes for each alternative and then selecting the location that maximises their utility. In this paper, he discusses the strengths and weaknesses of using this model in estimation. He states that aspects of the MNL can be helpful in the handling of non-linear constraints for full maximum likelihood estimation. McFadden and Train (2000) further develop this proposal, discussing the effectiveness of

a mixed multi-nominal logit (MMNL) approach in a discrete response model, an extension of the standard MNL model. The results of the study find that these models provide a computationally practical approach to the analysis of discrete choice models. Furthermore, that MMNL models are able to estimate any discrete model based on random utility maximisation effectively. In this paper we use MNL because it is appropriate for maximum likelihood estimation as suggested by McFadden. We expand on McFadden (1978) paper by investigating how a compromise is made between individuals when there are different weightings on potential locations.

Bhat and Guo (2004) compare the use of this multi-nominal logit model with their proposed mixed spatially correlated logit (MSCL) model using 1996 data from the Dallas Fort Worth area. The argument behind the MSCL is the ability of the model to take into account that the responsiveness to exogenous determinants of residential choice will vary across individuals. The study finds that both models have similar results, with certain variables having larger effects in the MSCL model. Findings in previous papers show that the MNL models provide computationally practical and accurate estimations for discrete choice models. We use this as a base for our model. However, as discussed in Ben-Akiva et al. (1985), the use of a MNL model requires the assumption of independence of irrelevant alternatives. This assumption requires that the relative probabilities for an individual choosing between two options, is independent of any additional alternatives in the choice set. Another common model used in the estimation of residential choice is the nested logit model. Lee and Waddell (2010), present a two-tier nested logit model in which they examine residential choice and household mobility. This model allowed the researchers to use simple random sampling of the alternative locations along with a process to account for sampling bias. This meant they were able to estimate using a full-information maximum likelihood.

The model presented in this study includes both multi-nominal and conditional logit parts. Hoffman and Duncan (1988) discuss the use of these models

for discrete choice models in demography, finding that conditional logit models are appropriate when the decision is dependent on the differing choice characteristics rather than solely the individual characteristics. Davies et al. (2001), discuss a conditional logit approach to migration between states within the United States. They find that through this approach, they are able to calculate marginal effects and trade-off values, which could also be potentially relevant for residential choice decisions.

The main focus of this paper is to investigate the effects of different variables on residential choice, but more importantly analyse the decision making process itself. We look at how the differing preferences of household members are reflected through the final decision.de Palma et al. (2005), use a residential location choice model within the Paris region in which households are treated as single units. Pinjari et al. (2011) use a similar approach in their estimation of residential choice within the San Francisco Bay Area. Their sample includes 5147 adult commuters, each acting as a single decision making unit, each randomly sampled from a household. This approach however does not accurately capture the influence of each individual within the decision making process. Further, the model used is unable to capture this level of choices within households containing multiple commuting adults. Browning et al. (1994) discuss the limitations and the unrealistic nature of using an individual to represent the household. Browning and Chiappori (1998) expand on this idea in their paper where they propose a collective model which relies on the assumption that the household decisions are efficient. They conclude that the idea of using both household members and analysing the intra-household decision is a plausible next step in this type of research. These papers suggest that the idea of modelling the household, where members have individual utilities, are consistent with random utility maximisation models and are likely to be more realistic than using a single individual to represent a household.

Los (1979) and Pinjari et al. (2011) model households as a single decision making unit and disregard the intra-household decision making process. This

has been a common theme in prior research, as it reduces the computational complexity of the model. However, in a real world setting it is highly unlikely that these substantial decisions are made by a single representative household member, rather they will be the result of a decision making process between multiple individuals. Therefore, in this paper we include all members in the decision making process, and model this process using a weight parameter to measure each individuals input. To estimate this parameter, a process of distinguishing the two household members needed to be established. In order to differentiate the individuals, the household head was selected based on the higher income.

Previous papers have taken different approaches to represent the weight and influence of each household member. Gliebe and Koppelman (2002) provide a psychological based framework to analyse joint activity participation between household members. The motivators behind these joint activities are efficiency, altruism and companionship. While the paper is not examining residential choice specifically, the analysis of the decision making process is relevant to our research. They propose a proportional share model of time allocated to a range of home activity types, in which many are joint activities. Here, time is the dependent variable allocated to the different activity types. This share model is analogous to a MNL model such as the one described in Ben-Akiva et al. (1985). A strength of the model used by Gliebe and Koppelman (2002), is the acknowledgment that individuals may not have equal influence in the decision, a point that our model aims to investigate and estimate further.

In contrast to the MNL models, Abraham and Hunt (1997) provide a modified form of a nested logit model to investigate household joint decisions in residential choice, work place location and commuting mode. The model allows for heterogeneity across household nesting structures determined by individuals' gender and age to provide a system for weight. This previous research provides evidence that treating all households as homogeneous and estimating

a single weighting parameter may initially be an oversimplification.

Timmermans et al. (1992) propose a circumstance in which both adults in a household are working, providing further complication to the decision making process. They provide an experimental model in which they use a sample of recent graduates from the School of Transportation in Tilburg, Netherlands. The experiment aims to test the theoretical hierarchical information integration approach. This is the idea that individual preferences are able to be represented through a hierarchical process. To summarise, the model assumes that individuals form preferences for higher level constructs, such as housing characteristics and the environment, then trade-off between these constructs to arrive at an overall preference. These were then combined and the partners choose jointly the combination of preferences which provides the most efficient outcome. Timmermans et al. (1992) do not include an accurate estimation of the individuals' weights in their decisions, rather they examine how factors such as age, income, and the number of children change the individual preferences and thus their residential location decisions. In contrast, Zhang et al. (2009) propose a different approach to modelling these individual preferences is provided. Rather than creating each preference separately and combining them to form a decision, they create a household utility which is defined as a function of all members utilities. This multi-linear household utility function is able to accurately represent the intra-household interactions as it takes the weighting of each individual into account. The household choice is then derived through a utility maximisation process that is in effect, maximising the utilities of the individual members.

Based on previous research, in this papers estimation we estimate each individuals marginal preferences based on a combination of multi-nominal and conditional logit parts. This is an improvement on previous papers as we are able to include a large range of alternative locations. These are then combined to create a joint set of probabilities which represent the probability of the group making particular choices. The approach allows us to estimate through

the maximum likelihood method. These utilities are combined through an estimated weighting parameter allowing us to examine further the intra-household decision making process. This is an improvement on previous papers as we are able to include a large number of alternative locations.

Initially, we estimate a weighting parameter that allows us to assign the influence of each household member in the decision, estimated based on who has the higher income. Through this approach, we are estimating a single weight parameter for all households, effectively assuming that households are homogeneous and that for all households the individuals within have the same weighting. In the latter stages of our estimation, we extend out model to include a semi-parametric form for the weighting parameter allowing us to examine the effects of variables such as income and gender while also allowing the weight structure to vary between households. This is a valuable contribution to the literature, as the findings from this estimation could allow further researchers who wish to model the household as a single decision making unit to more accurately choose which member within the household to use as the decision maker.

In existing literature there is minimal research completed with models that include households both single and dual adult households. The inclusion of the single adult households is essential as it allows us to more accurately examine the marginal preferences for the individuals and therefore the effects of factors influencing residential choice.

Another contribution of this paper is the ability of the model to use a large number of potential residential locations. Previously, it has been regarded as almost impossible to assess all potential residential locations for a household. Salon (2009) addresses this by using 10 alternative Census Tracts for each household's location. This small selection of alternatives may not be large enough to analyse the effect of different variables, while also running the risk that a household may be allocated 10 locations randomly that do not repre-

sent the range of options adequately. In our study, each household is allocated one meshblock, with a population of around 50-100, within each different area unit, defined as around 3000-8000 residents. In this paper, the data we are using covers the entire Greater Wellington Region, contrary to previous papers that have examined a particular city such as Salon (2009) who used New York City (NYC). By focusing on NYC alone, Salon (2009) is unable to include a large number of residential locations, as individuals are likely to travel between areas outside of the city such as New Jersey. Our data set allows us to have a large variation in the types of commuters and residential areas, for example inner-city apartments, suburban and more rural areas are included in this model. By including this large variation in residential areas, we can be confident that we have captured all viable residential locations within the region.

The results of our empirical estimation suggest that longer commute times as well as environmental factors such as amount of sunlight, coastal attributes and greenspace were observed to have significant effects on household residential decisions. Distance to schools and local school quality are also found to have a substantial effect in these decisions. Further, the results for the weight find that the individuals who have higher incomes are likely to have less impact in the decision with only 25%. However after extending to allow for a semi-parametric model, we find that income is likely to have a positive effect on an individuals weighting, while being female is likely to have a negative effect.

The rest of this paper is laid out as follows. Section 2 a detailed description of the logit model used in estimation is provided. Section 3 then provides a discussion of the variables included within the estimation and the data used. In section 4 the details of each stage of the estimation process are discussed. Section 5 then presents the results of each stage of estimation. These results are further discussed in section 7. Finally, section 8 concludes.

2 Model

We estimate the residential location decisions of households in which each household is jointly choosing between $j = 1, ..., J$ alternatives. The joint decision j is determined at the household level as a combination of both individual's separate preferences. For individual i in household h, we define the utility function in scalar form is given as follows:

$$U_{hij} = \sum_{l=1}^{L} X_{hil}\alpha_{jl} + \sum_{k=1}^{K} Z_{hijk}\beta_k = \boldsymbol{X}_{hi}\boldsymbol{\alpha}_j + \boldsymbol{Z}_{hij}\boldsymbol{\beta}$$

where X is a set of L choice-independent variables and Z is a set of K exogenous choice-dependent variables, with α and β being their respective coefficients. We define the first term as the multi-nominal logit component, made up of choice-independent information, while the second term is defined as the conditional logit component, made up of choice-dependent information. As Z forms the choice dependent information, there will be k coefficients, one for the effect of each variable as it does not change based on the decision j. However, as the effect of each choice-independent variable, indexed by l, will be different for each alternative location option j, there will be a separate α coefficient for each variable and location.

The multi-nominal logit (MNL) part of the model includes variables that will not change with a residential decision. These are variables such as gender, age and ethnicity. The conditional logit (CL) model represents any factors that are conditional on the decision, meaning that the variable values will change for each different residential location option. Examples of such variables will be commuting time, house prices, schools and school quality within the immediate vicinity. Through the inclusion of both the MNL and CL models, we are able to model a larger number of potential choices, ensuring we are able to cover the entire choice region.

Using this combination, this utility function is able to combine a large amount of variables which could impact either individual's or the households

utility.

We represent these in matrix form as follows.

$$[X]_{HI\times L} = \begin{bmatrix} X_{111} & X_{112} & \dots & X_{11L} \\ X_{121} & X_{122} & \dots & X_{12L} \\ \vdots & \vdots & \ddots & \vdots \\ X_{HI1} & X_{HI2} & \dots & X_{HIL} \end{bmatrix} [\alpha]_{L\times J} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1J} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{L1} & \alpha_{L2} & \dots & \alpha_{LJ} \end{bmatrix}$$

We define I as the number of individuals i in household h. I has value one or two to represent the differing household types. The X matrix is $HI \times L$, where there are $H \times I$ rows to represent each individual and the columns are value of X for each variable L, allowing for both single and dual adult households. The α matrix is $L \times J$ as there is a coefficient α_{lj} for each of the different choice-dependent variables for each residential location J. This will increase exponentially in size with the inclusion of more choices j.

$$[Z_k]_{HI\times J} = \begin{bmatrix} Z^k_{111} & Z^k_{112} & \dots & Z^k_{11J} \\ Z^k_{121} & Z^k_{122} & \dots & Z^k_{12J} \\ \vdots & \vdots & \ddots & \vdots \\ Z^k_{HI1} & Z^k_{HI2} & \dots & Z^k_{HIJ} \end{bmatrix} [\beta_k]_{J\times J} = \begin{bmatrix} \beta_k & 0 & 0 & \dots \\ 0 & \beta_k & 0 & \dots \\ 0 & 0 & \beta_k & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

There is a Z matrix for each of the choice-dependent variables. It has dimensions which are $HI \times J$, where rows are the individuals in each household HI and the J columns are the values of Z for each residential location j. There are K number of matrices for β, each having dimensions $J \times J$, as the coefficient for each choice-dependent variable is the same for each location j. For example, the amount of importance you place on your commute time will affect your decision, but how important it is will not change for each location you consider.

2.1 Individual Probability of Choosing a Particular Location

Using the utility function defined in the previous section, we use a logistic model to represent the probability of individual i choosing option j. We define the action taken by individual i in household h is given by y_{hi}, therefore, the probability of an individual choosing option j is given by the following function.

$$Prob(y_{hi} = j) = P_{hij} = \frac{e^{U_{hij}}}{\sum_{j'=1}^{J} e^{U_{hij'}}}$$

P_{hij} is defined as the probability that individual i in household h makes choice j. We assume that our multi-nominal logit model satisfies the IIA property, implying that the relative probability of individual i choosing between options j and j' is independent of any additional alternatives in the choice set.

The P_{hij} matrix has dimensions $HI \times J$. The rows are the individual level probabilities for each member of the household, which sum to $H \times I$, allowing inclusion of both single and dual adult households. When single adult households are included, there are $(H_2 \times 2) + (H_1 \times 1)$ rows, where H_2 and H_1 are the number of dual and single adult households respectively. The Jcolumns are the value of each individuals probability for choosing each alternative residential location j.

$$[P]_{HI \times J} = \frac{e^{U_{hij}}}{\sum_{j'=1}^{J} e^{U_{hij'}}} = \begin{bmatrix} P_{111} & P_{112} & \dots & P_{11J} \\ P_{121} & P_{122} & \dots & P_{12J} \\ \vdots & \vdots & \ddots & \vdots \\ P_{HI1} & P_{HI2} & \dots & P_{HIJ} \end{bmatrix}$$

2.2 Probability of Household Deciding on Alternative j

The individual level probabilities P_{hij} are aggregated to the household level probability, P_{hj}, defined as the probability of household h choosing option j. However, it is not as simple as combining these two probabilities together, as a major aim of this research is to estimate the potential weight in the decision

for each individual of the household. This weight is included as in the model as the parameter ω_i, where $\sum_{i=1}^{I} \omega_i = 1$. To estimate this, the first individual in each household, $i = 1$, is given the value ω, then the other individual is given the weighting $1 - \omega$. In order to get a reliable estimate of ω, the individuals need to be structured in a way to ensure that we are able to distinguish the two individuals. To accomplish this, the individual with the higher income is selected to be individual one.

We model ω using the following matrix:

$$[\omega]_{H \times HI} = \begin{bmatrix} \omega & 1-\omega & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & \omega & 1-\omega & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

Where the H rows represent each household and the $H \times I$ columns represent each household. ω and $1 - \omega$ run diagonally starting in columns 1 and 2, while all off-diagonal elements are zeros. Using this weight, we are able to combine our individual probabilities to create our household probability as below.

$$[P]_{H \times J} = [\omega]_{H \times HI} \times [P]_{HI \times J}$$

This is modelled at the household level by the following equation.

$$P_{hj} = \sum_{i=1}^{I} \omega_i P_{hij}$$

This means in a two adult household the probability of household h choosing option j is given by.

$$P_{hj} = \omega P_{h1j} + (1 - \omega) P_{h2j}$$

Where it is the combination of persons one and two's probabilities, multiplied by their respective weighting.

Proof. P_{hj} is a probability mass function

$$P_{hj} \in [0,1] \tag{1}$$

$$\sum_{j=1}^{J} P_{hj} = 1 = \sum_{j=1}^{J} \sum_{i=1}^{I} \omega_i P_{hij} \tag{2}$$

$$\sum_{j=1}^{J} \sum_{i=1}^{I} \omega_i P_{hij} = \sum_{i=1}^{I} \omega_i \left(\sum_{j=1}^{J} P_{hij} \right) = \sum_{i=1}^{I} \omega_i = 1 \tag{3}$$

$$\text{where } \sum_{i=1}^{I} \omega_i = 1$$

□

There are potential cases in which there are more than two adults in a household who are making a joint decision, such as in flats. Situations such as this are unable to estimated using this model. It is also highly likely that there will be households where there is only one adult who is making a decision. This is simple to include in the model as due to there only being one adult, that individual gets assigned an omega value of one.

2.3 Estimation via MLE

Estimation of this logistic model will be done using the maximum likelihood method. The log likelihood is given below. We employ σ_{hj} as an indicator function, which is equal to 1 if the alternative was chosen by the household in the data, providing us an indication of the actual choice. This gives a the function as below.

$$L(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega) = \prod_{h=1}^{H} \prod_{j=1}^{J} \left[\sum_{i} \omega_i P_{hij} \right]^{\sigma_{hj}}$$

$$\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega) = \log L(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega))$$

$$\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega) = \sum_{h} \sum_{j} \sigma_{hj} \ \log \left[\sum_{i} \omega_i P_{hij} \right]$$

$$\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega) = \sum_{h} \sum_{j} \sigma_{hj} \ \log \left[\sum_{i} \omega_i \frac{e^{U_{hij}}}{\sum_{j'=1}^{J} e^{U_{hij'}}} \right]$$

where $j \neq j'$

The maximum likelihood estimation procedure is then conducted in order to estimate the effect of the variables, shown through α_{lj} and β_k, as well as the weighting parameter to give further insight on how these decisions are made.

2.4 Formulae for the Score Vector

One of the major motivations for the use of MLE was that the first order conditions are relatively easy to take. Therefore, we take the first order conditions of $\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega)$ with respect to α, β and ω to formulate the score vector. These are represented in matrix form below.

The FOC with respect to α:

$$\frac{\partial \mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega)}{\partial \alpha_{lj"}} = \mathbf{1}_{1 \times H} \times \left[\boldsymbol{\sigma} \left[\frac{\boldsymbol{\omega} \frac{\partial \boldsymbol{P}}{\alpha_{lj"}}}{\boldsymbol{\omega P}} \right] \right] \times \mathbf{1}_{J \times 1}$$

where $\frac{\partial \boldsymbol{P}}{\alpha_{lj"}} = X_{il} P_{ij} (\sigma^{j}_{j"} - P_{ij"})$ and $\sigma^{j}_{j"} = 1$ if $j = j$" and 0 otherwise.

The FOC with respect to β:

$$\frac{\partial \mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega)}{\partial \beta_k} = \mathbf{1}_{1\times H} \times \left[\boldsymbol{\sigma}\left[\frac{\boldsymbol{\omega}\frac{\partial \boldsymbol{P}}{\beta_k}}{\boldsymbol{\omega P}}\right]\right] \times \mathbf{1}_{J\times 1}$$

where $\frac{\partial \boldsymbol{P}}{\beta_k} = \boldsymbol{P} \times [\boldsymbol{Z}_k - (\boldsymbol{P}\boldsymbol{Z}_k) \times \mathbf{1}_{J\times J}]$

The FOC with respect to ω:

$$\frac{\partial \mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta, \omega)}{\partial \beta_k} = \mathbf{1}_{1\times H} \times \left[\boldsymbol{\sigma}\left[\frac{\boldsymbol{DP}}{\boldsymbol{\omega P}}\right]\right] \times \mathbf{1}_{J\times 1}$$

where $\boldsymbol{D}_{H\times HI}$ is the derivative of $\boldsymbol{\omega}$ with respect to ω.

Using these, we are able to create a score vector to allow estimation using analytical derivatives.

3 Data

The data used within this study is that from Daglish et al. (2018). This is obtained through participants from the New Zealand Ministry of Transport's Household Travel Survey from within the Greater Wellington Region. This region stretches from the city itself north to Masterton and the Kapiti Coast. The ability to include this entire region allows us to be confident that we have accounted for each potential residential location. This is in contrast to papers such as Salon (2009) which were only able to capture a handful of locations. The survey documents a household over a period of two days and includes information such as where the household is located and how they commute to work. It also provides household demographic information such as age, income, gender and number of children. Within this study, only households in which all adults work full time and are commuting were included. The data covers an eight year span from 2003-2010.

To conduct the empirical estimation, we make use of variables created using GIS software, allowing us to get accurate distance and time attributes. This proves important when looking at commuting times as an impacting factor to residential choice. It also allows us to get accurate information on the closest schools, neighbouring house prices, green space, sunlight levels and coastal attributes, all of which have a significant effect on residential location decisions.

The HTS data is used along side GIS data, which gives us the information on the number and location of the area units within the greater Wellington region, as well as the meshblocks within each area unit. Both meshblocks and area units are based on population, with meshblocks containing 50-100 individuals while area units contain 3000-8000 individuals. It also gives information about the environmental factors of that meshblock, such as the percentage of north facing houses etc. Another seemingly obvious consideration for households when choosing residential location is house prices. In our study, the stratified sampling technique is employed, where the total population is broken down into groups and a random sample is drawn from each. Each household within our study is offered a random meshblock from each area unit. However, this caused certain meshblocks to be sampled more than others, potentially having a negative effect on the results. Ideally samples would be taken from each meshblock, however this would not be feasible.

For each household, one of the potential alternative locations, $j = 1, ...J$, is the actual location for that household, therefore we are able to index this decision using the indicator function. However, this raises some issues when we found that we have missing values for certain locations for a range of reasons. For example, if no houses were sold within a certain meshblock in a particular year, then the house price data is not available. Due to this, within our data set we technically have households who by our model, have made an invalid choice. These choices are not actually invalid, however using the house price example, a family may just live in an area where no houses have been sold recently, meaning they appear to have no prices. These households that have

Table 1: Variable descriptions.

Variables	Description
Time_Drive_Alt	Commute time in minutes for an average one way trip.
Price_Alt	Average house price at the mesh block level.
ResPrice_Alt	Residuals from the auxiliary house price regression.
UE_Coed_Alt	Average university entrance rate for a COED secondary school in the zone.
DT_Prim_Alt	Average driving time to the nearest primary school in the zone.
DT_Coed_Alt	Average driving time to the nearest coed secondary school in the zone.
DT_Boys_Alt	Average driving time to the nearest boys only secondary school in the zone.
DT_Girls_Alt	Average driving time to the nearest boys only secondary school in the zone.
PercMBNoVeg_Alt	Average percentage of area with no vegetation at the mesh block level.
PercMBDenseVeg_Alt	Average percentage of area with dense vegetation at the mesh block level.
Shape_Area_Alt	Size of the mesh block in square meters.
MBMeanBed_Alt	Average number of beds in a dwelling at the mesh block level.
PercNorth_Alt	Percentage of the mesh block that is north-west facing.
ln_dist_coast_Alt	Natural log of the travel distance in meters to the nearest coast.
ln_visi_coast_Alt	Natural log of the coastal visibility in the zone.
ln_Income	Natural log of personal income.
Gender	Dummy variable for gender, equals one if they are female.

missing values for their actual residential location were identified before the estimation and removed from the data set.

Two data sets are used within the empirical estimation. Firstly, the previously discussed data set is used for the first 4 stages of estimation, with stages 1 and 2 including the dual adult households only, while stages 3 and 4 include both single and dual adult households. The final stage of estimation aims to use a larger data set covering a longer period using HTS and GIS data to further test the model, however this data set lacks some variables. Section 4.5 discusses this in more detail.

Table 1 provides a detailed explanation of each variable used in the estimation. Comber et al. (2008) and Conway et al. (2010) show that the amount of green space nearby has a positive effect on house prices, this provides evidence that it may play a role in residential choice decisions.

Helbich et al. (2013) show that the amount of sunlight a household gets has a positive effect on its price. This suggests that it is an important consid-

eration in residential choice. The Wellington region is a very hilly area, and the most expensive suburbs are generally on the coast, on hill tops and slopes that face north, as this gives maximum sunlight. Slopes that face North East, North and North West were used as a proxy for sunlight.

Jin et al. (2015) discuss how the affect of coastal attributes on house prices, finding they are positively related with house prices, suggesting that households that are close to the coast and/or have good views will be favourable. However, we take the natural logarithm of both these variables. For a household living within a coastal area unit, then the difference between being beach front versus being two roads back from the beach is substantial, however if they live in an area such as Masterton which is an hour from the nearest coast, being a further two roads from the coast will have little effect.

Another important consideration for households when considering residential location is the proximity and the quality of local schools. For each residential location given by the area unit, the closest primary, co-educated secondary, boys secondary and girls secondary schools are calculated, however most schools in the Wellington region are co-educated. To assess the quality of the secondary schools, we use each school's University Entrance rate. The University Entrance rate is the percentage of students that complete the requirements during their final year at secondary school. This is completed through a range of classes and exams.

The average size of houses within an area and the number of people within the households can be representative of larger family friendly suburbs and other potentially favourable characteristics. To investigate the effect, the average number of beds in a dwelling is calculated at a meshblock level. Furthermore, the size of the meshblock can be viewed as a representation of how densely populated it is. As meshblocks are defined by population, larger meshblocks will be less densely populated. Bhat and Guo (2004) show that population density and the size of the residential zone are important factors in

Table 2: Gender Distribution of Household Heads.

Model	Female Household Heads	Male Household Heads
Base model	60	80
Base model w/ single adult HHs.	138	141
Semi-parametric model w/ single adult HHs.	138	141
Semi-parametric model w/ extended data set.	203	285

residential choice decision.

When moving on to the later stages of estimation, the weight parameter is no longer treated as a purely parametric variable, rather we define a functional form for the weight, ω. Abraham and Hunt (1997) employ a weighting system which is dependent on income and gender. We employ a similar idea here. We let the weight depend on is the natural logarithm of income. It is likely the amount of money you earn in the household will have an effect on your weight in the decision. We use the natural logarithm as for example, if there are a couple who are earning 50,000 and 60,000, the 10,000 difference is relatively substantial. However, if we have a couple who are earning 120,000 and 130,000, this difference is less substantial.

Based on Abraham and Hunt (1997), another variable that will impact the weight is gender, whether being female meant you had more or less impact on the decision making process. Table 2 provides a breakdown on the gender distribution of the household heads over each of the estimation stages. While conducting estimation using the original data set, we can see that the split of male and female household heads is close to 50:50. As the condition for being the household head was having the higher income in the household, this distribution implies that there is an even spread of females and males having higher incomes. During the final stage of estimation however we can see that this household head distribution shifts to being almost 60 percent male.

Table 3 provides detailed summary statistics of the variables used within

Table 3: Summary Statistics of Variables used in Estimation.

Variables	Stage 1 and 2		Stage 3 and 4		Stage 5	
	Mean	se	Mean	se	Mean	se
Time_Drive_Alt	26.9809	1.5592	27.7526	1.2657	28.0980	0.9791
Price_Alt	203.7692	12.1775	361.1938	11.5794	387.4184	6.6510
ResPrice_Alt	-	-	-	-	-	-
UE_Coed_Alt	0.3260	0.0084	0.3206	0.0067	0.2921	0.0042
DT_Prim_Alt	2.3878	0.3546	2.3793	0.2777	2.3576	0.2690
DT_Coed_Alt	5.5823	0.5519	5.5738	0.4341	5.4487	0.3747
DT_Boys_Alt	28.1414	1.5842	28.1391	1.2543	12.8470	0.5228
DT_Girls_Alt	22.7933	1.5040	22.7897	1.1907	12.5431	0.5260
PercMBNoVeg_Alt	28.9159	1.8344	28.9974	1.4539	72.5264	1.4114
PercMBDenseVeg_Alt	27.0962	1.8539	26.9614	1.4699	11.2523	0.8815
Shape_Area_Alt	2.6048	0.8023	2.6256	0.6382	-	-
MBMeanBed_Alt	3.0464	0.0263	3.0464	0.0209	-	-
PercNorth_Alt	33.9767	1.8865	33.9595	1.4983	33.7524	1.1388
Dist_Coast_Alt	10.7732	0.9452	10.7890	0.7496	-	-
Visi_Coast_Alt	9.1080	0.0012	9.0961	0.0009	-	-

the estimation at each stage. We can see that as the sample sizes increase, the standard error of the variables decreases. House prices are calculated in thousands of dollars and the distance to coast is calculated in kilometres. The shape and area of the meshblock is calculated in square kilometres.

4 Estimation

In order for this model to be estimated, mathematical programming was necessary. The model was estimated in MATLAB as this program would allow us to take advantage of its strong matrix algebra functionality. The use of mathematical programming for estimation has been widely documented such as in Abraham and Hunt (1997). As previously discussed, the model is estimated through the maximum likelihood method in a series of different stages. This method was particularly attractive due to its nature of being robust to reparameterization and as the first order conditions are relatively simple to take as shown in section 2.

4.1 Stage 1 - Two Adult Households Only

Initial the model only uses data on households that have two working adults. In this version of the model, the individual weighting parameter *omega* is treated as purely parametric, which we estimate through the maximum likelihood process.

Furthermore, during the first stage of estimation, all variables that are classified as choice-independent, which make up the multi-nominal part of the model, are excluded from the estimation. In terms of the model, this means removing $\boldsymbol{X}_{hi}\boldsymbol{\alpha}_j$ from the estimation entirely. The reason for this is that due to the way the model is specified, if we were to include variables such as gender or ethnicity, the effect of these variables, shown through the coefficient value, would be different for each potential residential location. In other words, rather than a single coefficient to represent the effect as we have in the conditional logit model, there would be a different coefficient value for each individual area unit, causing the estimation process to become too computationally demanding and infeasible. This parameter problem shows the benefit of the conditional logit model, as we are able to include a larger amount of residential locations without the β coefficient being affected.

4.2 Stage 2 - Addressing House Price Endogeneity

In this model, households form preferences based on the favourable characteristics of potential residential locations. These areas are likely to have higher prices due to being more favourable, suggesting that higher house prices will have a positive effect on location decision. However, households are constrained by what they can afford, therefore areas that are more expensive will be less favourable. This leads to an endogeneity problem for the house prices, an issue we aim to eliminate in this second stage of estimation.

To address this endogeneity, we run a regression of the meshblock house prices on the surrounding meshblock prices. If there are favourable amenities

nearby that make an area more attractive, it is likely that these amenities will have the same affect on the neighbouring house prices. House prices are likely to fluctuate, therefore the year in which these house prices were sampled will be very important. We leave out the dummy variable for 2003 as to avoid the dummy variable trap.

$$House_Price_{jt} = \beta_0 + \beta_1 Neigh_Price_{jt} + \beta_2 D_2004 + \beta_3 D_2005+$$

$$\beta_4 D_2006 + \beta_4 D_2006 + \beta_5 D_2007 + \beta_6 D_2008 + \beta_7 D_2009 + \epsilon_t$$

This is a variation of a two stage least squares estimation, where in the first stage we estimate the residuals and use these in the second stage (MLE) estimation. These residuals strip out any collinearity, by effectively including the effect of the unobserved error, ϵ_t in the second stage. This allows us to strip out this unobserved effect, removing any endogeneity problem.

4.3 Stage 3 - Single and Two Adult Households

Following on from stages 1 and 2, the next step of estimation is to include as many households as possible from the data set available. We extract all single and two adult households where all adults in the household are working. This requires the structure of the weighting matrix to be changed, as for a single adult household, the respective weighting will be one, as there is no compromise between two people.

As the value of omega is one for these individuals, mathematically in their respective likelihood functions, the omega parameter would simply not appear. This meant that they will have no contribution to the score vector, so in terms of the results, we would expect the inclusion of single adult households to have little effect on the size of the weight parameter. Furthermore, as with the previous stages of estimation, we only include choice dependent variables.

Whilst extracting this data from the full data set, there are cases of house-

holds that have more than two working adults, where two parents are working full time who also have a 18+ year old child living at home working. In such cases we remove the children from the sample, so only the parents are considered.

4.4 Stage 4 - Treating Weight as Semi-Parametric

In this stage of the estimation, we shift away from treating the weight parameter as purely parametric and we introduce a functional form of ω. This can be represented by the following function, which we will define as $\omega(\gamma)$.

$$\omega_j = \gamma_0 + \gamma_1 ln_Income_j + \gamma_2 Gender_j$$

This extension to the model is essential as treating the weight parameter as purely parametric may not be the most accurate and realistic approach. It is likely that individual level factors such as gender and income will affect a persons influence in the decision. Treating weight as semi-parametric, allows us to model the effect of these variables and lets the weighting between the individuals to vary across households.

This extension of the model is highly important as it will give insight into which factors have the largest influence on a person's weight. This is relevant as stated previously, many papers choose to treat households as a single decision making unit, therefore selecting one person from the household as the primary decision maker or household head. By analysing and identifying the most important factors which influence this decision making, this could allow for more accurate choice of household heads in these simpler models.

This change meant that the MATLAB code was reworked in order to include this new functional form of omega. However, whilst it is now a function, it works very similarly as before, where individual one is given $\omega(\gamma)$ and individual two is $1 - \omega(\gamma)$. The omega matrix below is similar to that in section

2.

$$[\omega]_{H\times HI} = \begin{bmatrix} \omega(\gamma) & 1-\omega(\gamma) & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & \omega(\gamma) & 1-\omega(\gamma) & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

This extension to the baseline model used the full data set of both single and two adult households, with only the choice dependent variables. However, we do now include age and gender within the functional form of ω.

4.5 Stage 5 - Including More Data

In this stage of the estimation we further assess the performance of the model through the inclusion of a new data set. To estimate this, we merge the two data sets together to increase the total sample size by a large amount. This is important as a larger sample size will further reduce the standard errors and therefore increase the t-statistics.

The extended data set covered a longer period of 14 years from 2003-2016. However, coastal information was unavailable meaning these variables were excluded from the estimation. Furthermore, information on the shape and area of the meshblock and the average number of bedrooms were also unavailable and therefore excluded.

5 Results

5.1 Stage 1 Results - Two Adult Households

From Table 4, We observe that the effect of driving time is substantial and found to be statistically significant, with a coefficient value of -0.18(2dp) and a t statistic of -6.15 (2dp). This negative coefficient indicates that the farther a residential location is from your work location, the less favourable

Table 4: Base model with 2 adult households only including variables as in Table 1.

Variables	Coef	SE	tstat
Time_Drive_Alt	-0.177	0.029	-6.15
Price_Alt	-0.644	0.993	-0.65
UE_Coed_Alt	2.258	0.865	2.61
DT_Prim_Alt	0.069	0.169	0.41
DT_Coed_Alt	-0.081	0.068	-1.18
DT_Boys_Alt	0.003	0.045	0.08
DT_Girls_Alt	0.131	0.055	2.40
PercMBNoVeg_Alt	0.003	0.006	0.48
PercMBDenseVeg_Alt	0.004	0.007	0.57
Shape_Area_Alt	-0.055	0.248	-0.22
MBMeanBed_Alt	0.461	0.370	1.25
PercNorth_Alt	0.010	0.004	2.76
ln_dist_coast_Alt	-0.105	0.126	-0.84
ln_visi_coast_Alt	0.160	0.080	2.01
Weight	0.248	0.110	2.26

the area and therefore the less likely you are to live in the area. Further, this sizeable effect shows that drive time to work is an important variable with a substantial effect on residential location decisions, meaning that individuals will favour areas that have faster commute times via driving.

We expected that driving times to both primary and secondary schools would have a sizeable effect on household residential choice. From Table 4, driving time to the closest primary school has a positive coefficient, indicating that the longer it takes to drive to the closest primary school the more favourable a residential area becomes. This is counter intuitive as for many households, when purchasing a house they are often planning on being in the area for a relatively long period of time and therefore consider the scenario that they have children, how far away is the closest primary school. However, there is a much larger number of primary schools in the Greater Wellington Region with there likely being one within each area unit. This would mean that the driving time is likely to be less important. Driving time to the closest co-educated secondary school however does have a negative coefficient which is

in line with what was expected. However, we observe that the driving time to both types is not statistically significant in the current version of the model. When we look at the driving time to single sex schools, drive time to both boys and girls only schools have a positive relationship, implying that as this commute increases, it is a more preferable residential location. However as these schools are almost exclusively within the city itself, these variables are likely to be measuring the effect of the commute time to the city.

We further observe the quality of schools in the area is a significant consideration for households when deciding on residential location. We assume that the most common assessment for a school is their academic success, therefore we use the university entrance statistic. UE_Coed_Alt measured the performance of the closest co-educated school. From Table 4, we can see that UE_Coed_Alt, had a coefficient value of 2.26(2dp), and a t-statistic of 2.61(2dp), indicating a significant positive relationship. This would suggest that areas which have higher academically performing schools are more favourable.

We would expect that when households are considering potential residential locations, that for a majority of households the price would be negatively related with residential choice. From Table 4 we observe that Price_Alt has a coefficient value of -0.64435 but with a t-statistic that is well below the critical value. Therefore, we consider the case that there may be an endogeneity problem due to unobserved factors that are not estimated in the model, such as that there are nearby amenities which positively affect peoples decisions. This endogeneity may potentially have a large effect on the results, therefore it will be dealt with in section 5.2.

Amount of vegetation within the meshblock is another factor which we expect to have an effect on residential choice. From Table 4, we observe that both living in areas which have either dense or no vegetation are favourable as both have a positive coefficient. However, both of which have a test statistic

well below the critical value and therefore are not significant in this estimation. Due to this, it is likely that areas which have sparse vegetation are therefore preferable. We can see that the mean number of beds within a meshblock has a positive coefficient suggesting that areas with larger houses are more favourable.

We observe that the amount of sunlight an area receives is an important and statistically significant factor in choosing residential location. This was measured through the 'PercNorth_Alt' variable. From Table 1, we can see the coefficient value is 0.013601 with a t-stat of 2.557, therefore showing a positive, significant relationship.

When looking at the distance to and the visibility of the coast, we would expect for areas that are by the coast, the closer you are and the better view the area has will be favourable. We observe from Table 4 that the distance to coast has a negative coefficient, implying that the further from the coast you are, the less favourable location it is. The visibility of the coast had a statistically significant coefficient value of 0.16 (2dp), implying that when looking at areas that are near the coast, locations with a better view are more favourable.

From Table 4 we can see that the weight parameter is 0.23945, implying that individual one has a weighting of almost 24% and individual two has a weighting of almost 76%. which was much lower than anticipated. We hypothesised that individual one, deemed the household head, was the higher of the two incomes, and therefore it was expected that they would have a larger weighting in the decision. Consider a situation were one of the household members was offered a job in Wellington and the couple decided to find a house and live in the Kapiti Coast (roughly 45-60 minutes north of Wellington). The second member of the household then decides to find a job that is closer to home for convenience. This would have an effect on our estimation in terms of how the couple compromised to choose that particular residential location. In the data it would look like the second member, who is on a lower income, had

a larger weight in the overall decision as the weight is estimated based on the households actual location and the distance to both individuals' workplaces.

Another potential reason for this unexpected result, could be the specification of the weight parameter in the model. In the current estimation, the weighting variable is treated as purely parametric, However, this may not be the most accurate way, as there are likely other factors at play which we have currently been unable to estimate. For example the choice-independent variables are likely to be important when estimating the weight and therefore the process of compromise within the household. Factors such as age and gender may have a substantial effect. The extension of this estimation in section 4 treats the weight as semi-parametric, allowing the effect of these choice-independent variables to be estimated. Furthermore, it is entirely possible that estimating a single value of ω for all two adult households is not going to be the most accurate measure. Even though the current version was significant, the model may not be specified as accurately as it could be. This then provides us with motivation to expand this model to allow the households to be treated as heterogeneous and allow certain variables to influence the weight.

Overall, we observe that many of the included variables have test statistics that are well below the critical value. This provides justification to include more data through the inclusion of single adult households, as well as a different extended data set. This larger sample will provide more accurate results and therefore larger test statistics.

5.2 Stage 2 Results - House Price Endogeneity

We observe, in Table 4, that the coefficient value for the house prices was -0.64(2dp), however has a large standard error value and a very small t-stat deeming it insignificant. This is likely the result of an endogeneity problem in which there could be favourable amenities that are in or nearby a residential

Table 5: Baseline model after addressing house price endogeneity with two adult households only and variables as in Table 1.

Vars	Coef	SE	tstat
Time_Drive_Alt	-0.178	0.030	-6.15
Price_Alt	-0.743	1.257	-0.59
ResPrice_Alt	0.667	1.956	0.34
UE_Coed_Alt	2.287	0.922	2.48
DT_Prim_Alt	0.0769	0.169	0.46
DT_Coed_Alt	-0.081	0.069	-1.19
DT_Boys_Alt	0.003	0.046	0.058
DT_Girls_Alt	0.132	0.056	2.36
PercMBNoVeg_Alt	0.003	0.006	0.50
PercMBDenseVeg_Alt	0.004	0.007	0.56
Shape_Area_Alt	-0.033	0.261	-0.13
MBMeanBed_Alt	0.439	0.373	1.18
PercNorth_Alt	0.011	0.004	2.80
ln_dist_coast_Alt	-0.104	0.126	-0.83
ln_visi_coast_Alt	0.162	0.080	2.03
Weight	0.250	0.110	2.28

location that are not currently included (such as a golf course in the particular neighbourhood). These favourable amenities may influence individuals' preferences, however are not sufficiently included in our model. These unobserved positive amenities create higher demand for certain residential locations, however due to this higher demand, houses are more expensive. As these areas may have features that are not measurable, this gives us the result that people may like areas with higher prices. However, realistically less people are able to afford homes in the area and are less likely to live in these residential locations.

As discussed in the estimation section, we run a two stage least squares regression in order to address this endogeneity and strip out any collinearity. The results of the auxiliary regression were as follows.

From the above table we can observe that the β_1 value is 0.99. This is extremely close to one as expected, as if prices change in a neighbouring area they are expected to have the same effect to the specific area being observed. Furthermore, if the neighbouring area house prices are high, it is likely that

Coefficient	Value
β_0	5702.52
β_1	0.99
β_2	-125.11
β_3	-1267.65
β_4	9.37
β_5	-1900.56
β_6	-2196.20
β_7	2345.76

there are favourable amenities nearby.

There is a large variation among the year dummy coefficients. The most notable are beta values 5 and 6, representing years 2007 and 2008 respectively. These large negative values are justified as this is around the time of the Global Financial Crisis where house prices were heavily impacted. Following these years we see a sharp increase as the house prices begun to rise relatively quickly as they recovered from the recession.

Following this auxiliary regression, the residuals are calculated and included as ResPrice_Alt in the maximum likelihood estimation. We would expect to see that these residuals strip out any unobserved factors and collinearity, therefore allowing us to get a more accurate estimation of the effect of prices on residential location. From Table 5, we observe that the price coefficient has become further negative, however is still statistically insignificant. This provides further justification to introduce a larger sample to increase the accuracy of our estimation.

5.3 Stage 3 Results - Single and Two Adult Households

This stage follows on from stage 2 as we are continuing to fix the house price endogeneity. With the inclusion of the single adult households, we see the results as shown in table 5. The inclusion of the single adult households

Table 6: Baseline model with single and two adult households, addressing house price endogeneity, and variables as in Table 1

Variables	Coef	SE	tstat
Time_Drive_Alt	-0.148	0.016	-9.14
Price_Alt	-0.797	0.789	-1.01
ResPrice_Alt	0.195	1.267	0.15
UE_Coed_Alt	1.862	0.617	3.02
DT_Prim_Alt	-0.006	0.102	-0.06
DT_Coed_Alt	-0.084	0.043	-1.95
DT_Boys_Alt	-0.034	0.030	-1.13
DT_Girls_Alt	0.154	0.035	4.36
PercMBNoVeg_Alt	-0.002	0.004	-0.44
PercMBDenseVeg_Alt	0.003	0.005	0.72
Shape_Area_Alt	-0.076	0.190	-0.40
MBMeanBed_Alt	0.260	0.259	1.01
PercNorth_Alt	0.010	0.003	3.93
ln_dist_coast_Alt	-0.193	0.081	-2.39
ln_visi_coast_Alt	0.078	0.056	1.41
Weight	0.257	0.105	2.44

means the sample size has increased from $n = 280$ to $n = 445$, representing 279 households. We expect that the increase in sample size will lead to more accurate results due to the reduction in standard errors. .

Firstly, we observe from Table 6 that the weighting parameter has only slightly changed with this addition, from 0.25 (2dp) to 0.26 (2dp). This result implies that the household head has only a 26% weighting in the decision, while the other individual has 74%. Single adult households by definition have no other adult in which to compromise with, giving them total control over the decision. To address this, we model our estimation by giving these individuals an omega value of one. As expected, the inclusion of these households did not substantially change the weight parameter. We do observe, however, that with inclusion of more data the weight t-statistic has increased from the previous value 2.28 (2dp) to 2.44 (2dp). While it is statistically significant, this version of the model in which households are treated as homogeneous is likely to be an unrealistic assumption. Rather, it is very likely that households will be

heterogeneous and that the weightings between individuals will vary largely between households.

From Table 6, we can see that in this stage of the estimation, the driving time to co-educated schools has become significant with a coefficient value of -0.008. The increase in sample size from including the single adult households has cased the test statistic to increase to a significant level. This negative relationship implies that residential locations which are further in terms of driving time are less favourable locations. This relationship is intuitive as from our previous results we can see that proximity of schools and school quality are favourable factors in terms of residential location.

As with the previous stages of estimation, we can see that the distance and time to commute is still a very significant and important consideration for households when making a residential location decision. The coefficient value is -0.14 with a much larger t-statistic of -9.20, implying a high level of statistical significance, further reinforcing the previous discussion. As the sample used contains purely working individuals, it is expected that commuting time to their work location will be a very important consideration. An extension to this model would be to include a data set in which households where stay at home parents are included, as by only using working adults who all must commute, there could potentially be a bias to the results in terms of the importance of commute times. For example, if a household has a stay at home parent, it is likely that factors that affect their children will have a higher importance, rather than just how long it takes for the parents to commute to work.

Another interesting change in this stage of the estimation is that the driving time to boys and girls have opposite signs. The results from Table 6 suggest that the longer the commute to a single sex boys school, the less favourable the residential location, while also suggesting the opposite for the commute to single sex girls schools. As discussed previously, it is likely that these variables

do not play a large role in the residential location decision as there are far less single sex schools within the Greater Wellington Region and they are all grouped within the city itself.

An interesting difference between the two adult model and this extended version of the model is the difference in effects of the coastal attributes. From Table 5, we observe that only visibility of the coast is significant, however in this stage of estimation, from Table 6, the test statistic has fallen to below the critical level to 1.41. The coefficient value of 0.08 (2dp), implies that within coastal areas, having a view is a favourable characteristic (as we would expect). The distance to coast has had a much larger reduction in the standard error and is now significant with a test statistic of -2.39 (2dp). The coefficient value of -0.19 (2dp) implies that for residential locations that are close to the coast, the further a location is from the coast has a negative impact on that location. These results are interesting as it is a common thought that having a view and being close to the coast have positive relationships with house prices, therefore we would expect to see what our results show that living close to and having a view of the coast are both favourable attributes in residential location choice.

As in the previous estimation, we can see that amount of sunlight is still a largely significant factor in the residential location decision. This is unsurprising as in general, houses that are north facing have higher rates of sunlight and are much less likely to be damp and cold. There could also be an unobserved effect due to the fact that the most expensive suburbs within the Greater Wellington Region are likely on hill top areas, or within areas that are majority north facing. This will mean that the houses within these areas are likely to be of a nicer quality and therefore preferable, explaining why percentage north is an important variable in residential choice decisions.

We can see that the size of houses has a positive coefficient of 0.26 (2dp). While the test statistic is below the critical value, the proposed effect of this variable is still interesting to look at. This positive relationship implies that

areas which have larger sized houses on average are more favourable residential locations. This is intuitive as larger houses are often more valuable, therefore areas which have average higher prices are also likely to have larger houses. This is similar to the percentage north results as these large houses are often in more expensive suburbs which are viewed as more favourable residential locations.

In the previous results, both living in areas where larger percentages of the meshblock had no vegetation such as within Wellington city itself, as well as areas where large percentages of the meshblock are dense vegetation were both viewed as favourable residential attributes. We use these variables to represent the effect of living within dense urban areas as opposed to living in more densely vegetated areas. In this stage, we can see that the coefficient of PercMBNoVeg_Alt has flipped, implying living in areas with little vegetation is not desirable. This negative relationship could be the result of the fact that when looking at purchasing a house, many people move farther out from the city into more suburban areas, with lesser amounts looking for inner city or rural housing.

Overall, with the inclusion of single adult households, we can see that there is a general decrease in standard errors leading to more accurate results and larger test statistics due to this sample size. However, as discussed earlier this approach for the estimation of the weight is not as extensive as it could be. A main objective of this paper was to analyse the intra-household decision making directly, therefore in the next stage of estimation we propose a semi-parametric approach.

5.4 Stage 4 Results - Treating Weight as Semi-Parametric

With the inclusion of the choice-independent variables into the weight function, we no longer have a single weight parameter to represent all homogeneous

Table 7: Full model with semi-parametric weight, single and dual adult households, addressing house price endogeneity and variables as in Table 1.

Variables	Coef	SE	tstat
Time_Drive_Alt	-0.146	0.015	-9.87
Price_Alt	-0.799	0.616	-1.30
ResPrice_Alt	0.198	0.924	0.21
UE_Coed_Alt	1.871	0.590	3.17
DT_Prim_Alt	-0.006	0.071	-0.08
DT_Coed_Alt	-0.084	0.033	-2.58
DT_Boys_Alt	-0.036	0.025	-1.41
DT_Girls_Alt	0.154	0.030	5.11
PercMBNoVeg_Alt	-0.002	0.003	-0.56
PercMBDenseVeg_Alt	0.003	0.004	0.91
Shape_Area_Alt	-0.076	0.104	-0.73
MBMeanBed_Alt	0.260	0.184	1.41
PercNorth_Alt	0.010	0.002	4.86
ln_dist_coast_Alt	-0.194	0.057	-3.38
ln_visi_coast_Alt	0.078	0.051	1.54
Constant	-1.214	0.687	-1.77
ln_Income	0.300	12.895	0.02
Gender	-0.499	0.757	-0.66

households. In this stage of estimation, we allow there to be heterogeneity across the households and estimate the effects of certain choice-independent variables on the weight directly. Using these values from the estimation we are able to calculate the weights for each of the households individually.

We observe from Table 7 that the gender coefficient has a negative value of -0.50 (2dp). This suggests that being female (as the dummy variable = 1 for female), has a negative effect on your weight in the decision. While this result is not statistically significant, these results are still interesting. Furthermore, the income coefficient has a positive value of 0.30, indicating a positive relationship between having the higher income in the household and having more weight in the decision, providing some justification for our procedure of determining the household head in the previous stages of estimation. This result is somewhat surprising after the previous stages of estimation. Previously, the household head was chosen based on whoever had the higher income and we can see that from Table 2, the split between females and males is almost exactly fifty-fifty, meaning that the there are almost equal numbers of females and males who have higher incomes in their respective households.

Another interesting observation from this stage of results, we can see that at lower significance levels the visibility of the coast is now significant, further justifying our previous discussion. These results further suggest the importance of the environmental factors for households in the residential location decision. These results are promising as we enter the final stage of our estimation to include a much larger data set, which has been sampled over a longer period of time.

5.5 Stage 5 Results - Including New Data

With the inclusion of the new data, our sample size has increased from $n = 445$ to $n = 786$. However, during this last stage of estimation, the new

Table 8: Baseline Model with Extended Data set, single and dual adult households, addressing house price endogeneity and variables as in Table 1.

Variables	Coef	SE	tstat
Time_Driving_Alt	0.006	0.006	0.92
Price_Alt	-0.244	0.916	-0.27
ResPrice_Alt	-0.119	1.681	-0.07
Avg_UE_Coed_Alt	2.220	1.475	1.50
DT_Primary_Alt	0.047	0.153	0.31
DT_SeCoed_Alt	-0.213	0.046	-4.57
DT_SeBoys_Alt	0.031	0.023	1.37
DT_SeGirls_Alt	-0.032	0.025	-1.25
PercNoVeg_Alt	-0.007	0.005	-1.24
PercDenseVeg_Alt	-0.020	0.011	-1.96
PercNorth_Alt	0.011	0.003	4.02
Weight	0.771	0.155	4.97

data was only able to be estimated using the baseline model. Furthermore, as this data is more recent, the number of area units has increased from 194 to 204 due to population increase. However the results provided are promising. We see that generally, the results are different than what we would expect. From Table 8, we observe that the driving time has a positive coefficient, suggesting longer commute times are more favourable. However the test statistic in this estimation has fallen from -9.87 (2dp) from Table 7 to 0.48(2dp), implying a insignificant result. This result is surprising as throughout the estimation it had a significant influence on these decisions.

The results showing the influence of the proximity and quality of schools within the area are consistent with our earlier results. The average university acceptance rate has a lower test statistic than in the previous estimation, and is now insignificant. The contradicting effects of driving times to single sex schools is also still present, however these test statistics are lower, providing more evidence for these variables lack of influence in the decision. In contrast, the test statistic for driving time to co-educated schools has increased, implying that this variable has a significant effect on the overall decision.

An interesting result is the change in the effects of the environmental variables. We see that in previous stages that areas with dense levels of vegetation were favoured compared to areas with no vegetation, whereas from Table 8 we observe that the coefficient of dense vegetation levels is negative and significant, implying that areas that have dense vegetation are less favourable. As with the previous stage, the amount of sunlight a household receives has a significant positive effect on residential location choice.

A limitation of this stage of estimation was that many variables had to be excluded and we were unable to estimate this data using the full model, meaning that the weight had to again be treated as parametric. We can see that with the increase in sample size, the weight parameter has a value of 0.77(2dp), with a much higher test statistic value than previously at 4.98 (2dp). This result suggests that the higher income earner within the household has 77% of the weight in the decision, contradicting to what was found in previous stages. This result could be the effect of the skew towards males within the household heads, as we found that being female had a negative impact on your weight in the decision. Due to this limitation, further research is needed to adapt our model to more accurately examine what components impact the weight parameter.

6 Discussion of Results

When looking at these stages of results holistically, we find that the most important factor within the model is commuting time, with the exception of the final stage. Previous research has shown that when examining residential choice decisions, that commuting costs, which increase with the commuting time, have had significant influence on locations in which households choose to live. Vega and Reynolds-Feighan (2009), show that when households are relocating, having higher car travel costs are less favoured, and areas which entail higher commuting times are also less favourable. de Palma et al. (2005)

also find a similar result, that commuting times are negatively related with residential preferences. This result is intuitive, as generally, the less time it takes one to commute to work, the more favourable the location. However, not all individuals commute via the same method, therefore further research should look at the individual preference and observe how the commute time via different modes affects the decision.

Throughout the estimation, distance to schools and the quality of nearby schools have been shown to have significant influence on household residential choice decisions. We see that as the estimation progresses, the test statistics increase for the driving time to, and the university entrance rate of, co-educated secondary schools. This result is intuitive as these residential decisions are often long term, therefore assessing the quality of schools within the area is a natural step in the decision making process for a majority of households. We observe that the driving times to single sex schools have contrasting coefficients. We estimate that these variables play a far lesser role in the residential location decision. This is as within the Greater Wellington Region, there are far fewer single sex schools than there are co-educated, and a large proportion of these schools are all nested within Wellington City. Due to this, in residential areas situated large distances away from the city, these single sex schools are often not a viable option. Also, due to the location of these schools, it is likely that these variables are capturing the effect of commute time to Wellington City rather than the schools themselves.

When looking at primary education. we see that the coefficient value of driving times to primary schools is negative, suggesting that residential areas further away from primary schools are less favourable, however the test statistic deems this insignificant. A potential reason for this could be through the sample used in estimation. As we are only including households in which both adults are working full time, this could mean that we have a large number of households that do not yet have children, or that have children who are older and closer to secondary school age. This would provide further reason for the

importance of the secondary school variables.

Timmermans et al. (1992), show that residential environment is an important factor in residential choice decisions. This is consistent with the results found throughout our estimation. We see that the levels of sunlight that a household receives is significant throughout all stages of estimation. The importance of the coastal attributes for households is also significant. Houses that get more sunlight or are on the coast, are often higher in price, making these homes more desirable. This suggests an interesting idea, that while households favour these areas with better environmental factors, these areas are often the more expensive suburbs. This suggests a trade-off within the household between the price and the environment. While households put large weights on these environmental factors, all households will be constrained by price, which we observe from our results to have a negative relationship with residential choice.

After dealing with the house price endogeneity, we observe a steady increase in test statistics following the increase in the sample size (with the exception of the last step). Further research with richer and larger data sets should be conducted to further examine this endogeneity problem and get more accurate results on the effect of house prices on residential location decisions. We observe throughout our results that house prices have a negative impact on the likelihood of you living within a particular residential location. This result is intuitive as residential location can be thought of as a utility maximisation problem in which the household is constrained by what it can afford. Households choose the location they can afford, in which they achieve the highest utility, based on the variables we identify within the estimation.

We expect that areas with higher amounts of greenspace would be viewed as more favourable. Within the early stages of the estimation we find that areas that have no vegetation are viewed less favourably than areas which have larger areas of greenspace. However from the last stage of estimation we can

see that both of these levels of greenspace have a negative influence on residential locations, suggesting that areas that fall into an in-between category of sparse vegetation may potentially be favoured. This would be intuitive as it is a common theme that when households seek to purchase a home, they look within suburban areas, which have a moderate level of vegetation.

When analysing the results for the estimation of the weight parameter overall, it is hard to draw conclusions as when comparing the first stages of the estimation with the final stage, they have opposite results. This could be a result of sampling and the difference between these two data sets. However, we draw conclusions from stages 1-4 and exclude 5 due to the lack of variables and difficulty of estimation. From our results in stages 1,2 and 3, we find that the household member that had a lower level of income had a large influence in the decision. In contrast, when allowing the weight to be semi-parametric, we observed that income had a positive influence on the weight, while being female had a negative coefficient. These contradicting results show that it is essential that further research be undertaken to allow the other choice-independent factors to influence the weight. From our semi-parametric estimation, we observe that the individual with the higher income is likely to have a larger influence in the decision, which is also shown in the final stage of estimation. Further, we observe the impact of gender on the weight, which shows that being female lowers the amount of influence in the decision. These results provide a fundamental level of insight into the effect of gender and income, suggesting that if you were to conduct a study using only one adult from each household as the decision maker, it is a viable option to pick the individual with the higher income as the household head.

7 Model Limitations

While the results of this paper are promising and provide fuel for discussion, there were limitations that hinder the overall success of the model. One limitation is the lack of variables that are able to be included in the esti-

mation. In the theoretical model we discuss the choice-independent variables that comprise the multi-nominal logit model. Only two of these variables were eventually included indirectly through the weight, however this approach shows potential as it gives an indication of which factors have the biggest influence on the weightings between individuals.

Another factor that was excluded in our model was allowing for the individual level commuting preferences of the household members. This would allow analysis on the different commuting modes, and further insight into the intra-household decision making. For example, currently throughout the estimation the only commuting mode we assume is driving. However, for a large percentage of the population, commuting is done via other modes. Further research needs to be conducted into these commuting decisions, allowing for individual level decisions about commuting mode.

Another limitation in this study is the way in which the income data is provided. This data is broken up into brackets with differing step sizes making this non-linear. This limitation is less significant, as in our analysis, we are interested in who has a higher income between the individuals, and the effect that this has on the intra-household decisions. However, by being non-linear, this means that 0.1 units of income is sometimes worth $10000 while in other cases it is only worth $5000. Further research needs to be conducted with a linear income scale to accurately capture the effect of income on individuals weight.

8 Conclusion

This paper seeks to investigate and answer how individuals determine their residential location, by estimating their individual preferences and identifying the extent to which they compromise for the welfare of the group. A logistic model is developed to estimate the residential location and the intra-household

decision making process.

Data for estimation is obtained through GIS and the Ministry of Transport's HTS, giving a sample of single and dual working adult households. The results of our study suggest that longer commute times as well as environmental factors such as amount of sunlight, coastal attributes and greenspace are observed to have significant effects on household residential decisions. Further, we find that the distance to schools and local school quality have a substantial effect in these decisions.

This papers main contributions are as follows. Firstly, as by using the data from Daglish et al. (2018), we are able to cover an entire region, allowing us to account for all potential residential locations. Furthermore, we are able to break the region down at a more granular level, a point which many previous papers have struggled to achieve. Secondly, our extension of treating the weight as a semi-parametric function allows us to gain insight into factors that affect individuals' influence on the decision. Analysis into the intra-household decisions finds that having the relatively higher income within the household has a positive influence on the weighting, while being female has a negative impact. Given the limitations listed, these results should be interpreted as indicative, as further empirical work is needed to apply this semi-parametric model to larger data sets, while allowing the weight to depend on more variables.

The estimation procedure suggests three main extensions for future research. First, as in the last estimation stage we were unable to use the semi-parametric form and certain variables had to be excluded, further research is needed in order to be able to apply this model to new data sets and for different regions. Secondly, the inclusion of choice-independent variables is essential to gain insight into the varying preferences of differing household types. Lastly, the model could be extended to allow for individual level decisions of commute mode, allowing different methods of commuting to be included within the es-

timation.

References

Abraham, J. E. and Hunt, J. D. (1997). Specification and estimation of nested logit model of home, workplaces, and commuter mode choices by multiple-worker households. *Transportation Research Record*, 1606(1):17–24.

Ben-Akiva, M. E., Lerman, S. R., and Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press.

Bhat, C. R. and Guo, J. (2004). A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Transportation Research Part B: Methodological*, 38(2):147–168.

Browning, M., Bourguignon, F., Chiappori, P.-A., and Lechene, V. (1994). Income and outcomes: A structural model of intrahousehold allocation. *Journal of political Economy*, 102(6):1067–1096.

Browning, M. and Chiappori, P.-A. (1998). Efficient intra-household allocations: A general characterization and empirical tests. *Econometrica*, pages 1241–1278.

Comber, A., Brunsdon, C., and Green, E. (2008). Using a gis-based network analysis to determine urban greenspace accessibility for different ethnic and religious groups. *Landscape and Urban Planning*, 86(1):103–114.

Conway, D., Li, C. Q., Wolch, J., Kahle, C., and Jerrett, M. (2010). A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values. *The Journal of Real Estate Finance and Economics*, 41(2):150–169.

Daglish, T., de Riste, M., Sağlam, Y., and Law, R. (2018). Commuting and residential decisions in the greater wellington region. *Forthcoming*.

Davies, P. S., Greenwood, M. J., and Li, H. (2001). A conditional logit approach to us state-to-state migration. *Journal of Regional Science*, 41(2):337–360.

de Palma, A., Motamedi, K., Picard, N., and Waddell, P. (2005). A model of residential location choice with endogenous housing prices and traffic for the paris region.

Gliebe, J. P. and Koppelman, F. S. (2002). A model of joint activity participation between household members. *Transportation*, 29(1):49–72.

Helbich, M., Jochem, A., Mücke, W., and Höfle, B. (2013). Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning. *Computers, environment and urban systems*, 39:81–92.

Hoffman, S. D. and Duncan, G. J. (1988). Multinomial and conditional logit discrete-choice models in demography. *Demography*, 25(3):415–427.

Jin, D., Hoagland, P., Au, D. K., and Qiu, J. (2015). Shoreline change, seawalls, and coastal property values. *Ocean & Coastal Management*, 114:185–193.

Lee, B. H. and Waddell, P. (2010). Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation*, 37(4):587–601.

Los, M. (1979). Combined residential-location and transportation models. *Environment and Planning A*, 11(11):1241–1265.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record*, (673).

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470.

Pinjari, A. R., Pendyala, R. M., Bhat, C. R., and Waddell, P. A. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6):933.

Salon, D. (2009). Neighborhoods, cars, and commuting in new york city: A discrete choice approach. *Transportation Research Part A: Policy and Practice*, 43(2):180–196.

Timmermans, H., Borgers, A., van Dijk, J., and Oppewal, H. (1992). Residential choice behaviour of dual earner households: a decompositional joint choice model. *Environment and Planning A*, 24(4):517–533.

Vega, A. and Reynolds-Feighan, A. (2009). A methodological framework for the study of residential location and travel-to-work mode choice under central and suburban employment destination patterns. *Transportation Research Part A: Policy and Practice*, 43(4):401–419.

Zhang, J., Kuwano, M., Lee, B., and Fujiwara, A. (2009). Modeling household discrete choice behavior incorporating heterogeneous group decision-making mechanisms. *Transportation Research Part B: Methodological*, 43(2):230–250.