New insights into model specification and selection for composite marginal likelihood estimation: Application to car availability

Romain Crastes dit Sourd[*†] David Palma[*‡] Stephane Hess[*§]
Andrew Daly[*¶] Christian Holz-Rau[‖**] Joachim Scheiner[‖††]

February 5, 2018

Abstract

This paper investigates the determinants of car availability in Germany using a unique life course calendar data-set that covers 27 years of data. Modelling car availability over the life course requires to account for spurious state dependence (*i.e.* autocorrelation of the errors). The dimension of integration increases with the number of time periods considered which rules out Maximum Simulated Likelihood (MSL) estimation techniques. An alternative is to use the Composite Marginal Likelihood (CML) inference approach, which replaces high-dimensional integrals by a compounding of bivariate probabilities. The current paper delves into the issue of how to form and select CML functions. Indeed, CML is a flexible tool and different pairs of bivariate margins can be used, leading to different results. The typical approach consists in using the pairing combinations of temporally close choice situations. In this paper, we suggest instead to use randomly selected pairs and semi-randomly selected pairs, which we name smart pairs. We estimate a series of autoregressive random effects ordered probit models built from closest, random and smart pairs, and compare the results obtained using various goodness-of-fit indicators. Our results suggest that smart pairs provide better fit than closest pairs and random pairs and are computationally much less burdensome than closest pairs. These promising results also allow us to unravel the important role played by spurious state dependence in car availability across the life course.

Keywords: car availability, mobility biographies, composite marginal likelihood estimation

[*]Institute for Transport Studies and Choice Modelling Centre, University of Leeds (UK)
[†]r.crastesditsourd@leeds.ac.uk
[‡]d.palma@leeds.ac.uk
[§]s.hess@leeds.ac.uk
[¶]andrew@alogit.com
[‖]TU Dortmund, Germany
[**]christian.holz-rau@tu-dortmund.de
[††]joachim.scheiner@tu-dortmund.de

1 Introduction

Dependence on past travel habits and experiences affect future travel behaviour (Hanly and Dargay, 2000). It is essential to identify the most influential factors of travel behaviour over time in order to contribute to planning practice (Axhausen, 2008). In the past decade the focus of interest therefore shifted towards individual decisions in a life span. A comprehensive review of the theoretical framework and most important studies investigating mobility behaviour and mobility tool ownership over the life course has been recently published by Müggenburg et al. (2015). The authors address open research questions and conclude that studies often investigate mobility decisions with static (panel) models and should not neglect the temporal dimension of decision-making because it may lead to biased estimates. Indeed, a large stream of studies in transport research indicates that past mobility decisions strongly affect current outcomes. State dependence can be "true" or "spurious" (Heckman, 1981). True state dependence simply describes the fact that past observed mobility choices influence future decisions. Spurious state dependence refers to the correlation between model errors across time periods (autocorrelation). While there exist an abundant literature on true state dependence and how to account for it, spurious state dependence is rarely considered in panel choice models, which may be due to the fact that the estimation of such models is usually particularly burdensome.

More precisely, the basic approach for modelling repeated choices in the literature is to introduce random effects. Random effects generate correlations across time for the same individual. This does not only apply to the case where the dependent variable is a binary outcome but also to ordered dependent variables or count data. Random effect models are generally estimated using Maximum Simulated Likelihood (MSL). Indeed, random effects lead to integrals in the likelihood function during estimation, which are approximated using numerical simulation techniques (Hensher and Greene, 2003; Paleti and Bhat, 2013; Train, 2001). However, such an approach is not always feasible. Indeed, numerical simulation methods can be very time consuming and may even not reach a solution, especially when modelling spurious state dependence because accounting for autocorrelation requires to evaluate multiple integrals (one per time period considered). These limitations suggest to use other inference approaches. Recently, the Composite Marginal Likelihood (CML) and its developments such as the Maximum Approximate Composite Marginal Likelihood (MACML) methods have gained popularity in the choice modelling field (Bhat, 2011; Varin, 2008; Varin and Czado, 2009; Varin et al., 2011). Composite Likelihood is an inference function derived by multiplying a collection of component likelihoods, where the collection used is determined by the context (Varin et al., 2011). Each individual component is a conditional or marginal density. As a result the estimating equation obtained from the composite log-likelihood is unbiased. In other words, Paleti and Bhat (2013) simply describe CML as an estimation technique which replaces the multivariate probability of the dependent choices in the likelihood function by a compounding of probabilities of lower dimensions. A growing literature

on CML estimation has proven that this estimation technique can outperform MSL at a fraction of the computational cost. A typical CML likelihood function is the full pairwise likelihood, which features each pair of observed choice situation from the same individual. This approach only requires to evaluate a series of bivariate normal probabilities rather than a single high-dimensional integral.

However, even a full-pairwise likelihood approach may become computationally burdensome in the context of large temporal panel data-sets. A full-pairwise approach requires to evaluate $J \times (J-1)/2$ pairs of bivariate normal probabilities. The literature on CML suggests that it may not be necessary to evaluate all the possible pairs but only the closest ones. More precisely, Varin and Czado (2009), among others, shows that considering pairs which are too distant (*i.e.* more than a given number of time units apart) may not only increase the computational burden but also reduce the model fit. The optimal distance between pairs is usually found by estimating a series of models using increasingly distant pairs and computing the trace of the asymptotic variance covariance matrix for each model. The distance which minimises the trace corresponds to the optimal distance.[1] Another solution which is often mentioned but rarely applied is to select random pairs. More generally, Paleti and Bhat (2013) point out that "the CML approach is flexible, and allows customization based on the problem at hand. The issue then becomes one of balancing between speed gain/convergence improvement and efficiency loss". Moreover, the authors also indicate that how exactly to form a CML function remains a wide open area.

In this paper, we delve into the issue on how to form a CML function. Using a unique mobility biography data-set collected in Dortmund in 2007, we analyse the determinants of car availability for 930 German adults between 1980 and 2006 with a particular focus on spurious state dependence. Indeed, the probability of experiencing a given level of car availability is likely to be measured by unobserved components that are correlated over time. Each observation relates to one individual and one year. The dependent variable in this work is (stated) car availability. In full detail, the concept of car availability can extend to include the holding of a driving licence by the individual and the ownership, renting or leasing of a car by household members, their employers or others (e.g. parents of students). Even more widely, car availability can extend to the availability of a car in which the individual can travel as a passenger. In the data we are using, respondents were asked in which years they had access to a car, with responses for each year being Always, Often or never. It can be seen that these responses are subjective simplifications of a complex concept, but nevertheless they represent states that persist through time, whether because of long-term characteristics like orientation towards cars and the holding of a licence, or medium-term characteristics like car ownership. It is worth noting that modelling state dependence in the context of car availability or car ownership has received a lot of attention in the past (see Cherchi and Manca (2011); Dargay and Hanly (2007); Golob (1990); Hensher (2013); Kitamura and Bunch (1990) and Clark et al. (2009) among

[1] The log-likelihood of a model estimated *via* CML is ill-specified and can not be used for comparing fit across models estimated using a different number of pairs (Varin and Vidoni, 2005).

others). The purpose of the present study is to contribute to this literature by making methodological contributions on how to estimate car availability models using large panel datasets.

We explore the suggestion made by Paleti and Bhat (2013) among others and compare the performances of the CML estimator depending on whether closest pairs or random pairs are selected for building a CML function. Moreover, we introduce a third alternative which consists in using smart pairs, which are semi-randomly selected based on a series of criterion chosen by the researcher and given the data at hand. We estimate a series of autoregressive order probit and compare model performance using the trace of the asymptotic variance-covariance as well as a new indicator which corresponds to the log-likelihood of the underlying MSL model. Our results indicate that smart pairs outperform both the closest pairs and the random pair approaches. Hence, the contribution of this paper is threefold: (i) we provide new insights on the determinants of car availability across time using one of the longest panel data-set available on this topic, (ii) we further explore how to build a CML function and provide suggestions for good practices and (iii) we introduce a new indicator for comparing the goodness-of-fit of competing CML models.

The remainder of this paper is structured as follows: the next section presents the MSL estimation method and the CML estimation method of the autoregressive random-effects ordered probit. The third section discusses different methods for selecting models estimated *via* CML. The fourth Section presents the data collection effort and the model specification. The fifth Section discusses the results and the sixth Section concludes the paper.

2 Model selection and estimation

In this section we describe the equations underlying the MSL and CML estimation of the autoregressive random effects ordered probit model. We build our models step-by-step and we first introduce the existing MSL and CML formulations of the random-effects ordered probit model (REOPROBIT). Second, we present the autoregressive versions of this model (Paleti and Bhat, 2013). In addition, we provide some existing and new insights on model composition and selection for CML estimation.

2.1 The MSL random effects ordered probit model

The REOPROBIT is a simple extension of the ordered probit model (OPROBIT) which accommodates panel data. In this paper, we only provide a short description of this well-known model in order to better introduce the more complex autoregressive models we use in this paper. The model can be specified as follows:

$$y_{ij}^{*} = \beta' X_{ij} + \alpha_i + \epsilon_{ij} \tag{1}$$

where y_{ij}^{*} is the unobserved latent outcome defined as a function of relevant exogenous variables. i is an index for individuals ($i = 1, 2, ..., I$) and j is an index for the j^{th}

observation with $j = 1, 2, ..., J$ the number of periods under study. X_{ij} is a vector of exogenous variables and β' is a vector of coefficients to be estimated. The serially independent error term ϵ_{ij} is assumed to follow a standard normal distribution with zero mean and unit variance.

The discrete outcome observed for individual i at time j corresponds to k_{ij}. k_{ij} may take one value among K at each time period ($k_{ij} = 1, 2, ..., K$). In the context of this paper, k_{ij} refers to a given level of car availability among K (for example, never available, sometimes available, always available). $y_{ij} = k_{ij}$ if $\mu_{ijk-1} < y^*_{ij} < \mu_{ijk}$, where μ_{ijk} is the upper bound threshold corresponding to the discrete level k_{ij} with $\mu_0 = -\infty$ and $\mu_K = +\infty$. $\mu_1, \mu_2, ..., \mu_{K-1}$ are parameters to be estimated with $\mu_1 < \mu_2 < ... < \mu_{K-1}$. Finally, $\alpha_i = \alpha + \eta_i$ where η_i is an individual-specific random term. The role of η_i is to generate an equi-correlation between the repeated choice situations for a given individual. α is normalised to 0 if μ_1 is estimated (and the reverse is also possible). In this paper, we consider that η_i is normally distributed but it is worth noting that other distributional assumptions may be tested.

The REOPROBIT model is easily and rapidly estimated using MSL. The probability of the observed vector k_i of the sequence of ordinal choices ($k_{i1}, k_{i2}, ..., k_{iJ}$) for individual i given the individual specific random term η_i can be written:

$$P(k_i)|\eta_i = \prod_{j=1}^{J} \Phi(\mu_{ijk} - \alpha - \beta' X_{ijk-1} - \eta_i) - \Phi(\mu_{ij} - \alpha - \beta' X_{ij} - \eta_i) \tag{2}$$

where Φ stands for the standard normal cumulative distribution. It is then easy to integrate out the individual specific random-term η_i in order to obtain the unconditional log-likelihood of the observed choice sequence.

$$logL_i(\theta) = log\Big[\int_{-\infty}^{+\infty} \prod_{j=1}^{J} \Phi(\mu_{ijk} - \alpha - \beta' X_{ij} - \sigma v) - \Phi(\mu_{ijk-1} - \alpha - \beta' X_{ij} - \sigma v)\phi(v)dv\Big] \tag{3}$$

$v = \frac{\eta_i}{\sigma}$ with $\eta_q \sim N(0, \sigma^2)$ and θ corresponds to a vector of parameters. The log-likelihood function of the REOPROBIT model entails only a one dimension integral so model estimation is generally fast. Moreover, the model is also not prone to convergence related issues. There are no known reasons for estimating the REOPROB model using CML. We now turn our attention to the autoregressive version of this model, which we refer to as the AREOPROBIT.

2.2 The MSL autoregressive random effects ordered probit model

The simple REPROBIT model presented above assumes that the errors are uncorrelated across time. However, this may be too simplistic. In many cases, the choice model

errors are actually positively correlated, which indicates that past choices proxy for a large random utility draw. This is particularly relevant in the context of car availability where spurious state dependence is expected. The AREOPROBIT can account for correlation between errors across time but costs dearly in terms of computational efforts when estimated using MSL.

We follow Paleti and Bhat (2013) and assume a classic autoregressive structure of order 1 (AR1). We define $corr(\epsilon_{ij}, \epsilon_{ig} = \rho^{|t_{ij}-t_{ig}|})$ with t_{ij} the measurement time for observation y_{ij} ($g \neq j$ and $0 < \rho < 1$). ρ can be easily bounded using a logistic transformation. The latent outcomes $y*_{ij}$ now follow a multivariate normal distribution for the i^th individual. The mean vector of the multivariate normal distribution may be standardised in which case it corresponds to $\frac{\alpha+\beta' X_{i1}}{\tau}, \frac{\alpha+\beta' X_{i2}}{\tau}, ..., \frac{\alpha+\beta' X_{iJ}}{\tau}$ while the correlation matrix Σ has non diagonal entries $\zeta_{ig} = \frac{\sigma^2+\rho^{|t_{ij}-t_{ig}|}}{\tau^2}$, where τ, the standard deviation of the latent outcome $y*_{ij}$ corresponds to $\sqrt{\sigma^2+1}$. As previously stated, the MSL estimation of the AREOPROBIT is much more complicated than its simple random-effect counterpart and much more demanding in terms of computational time. More precisely, while (3) only entails a one-dimension integral, the autoregressive model requires to evaluate an integral of dimension J for individual i. The log-likelihood function becomes:

$$logL_i(\theta) = \left[\int_{w_1=\delta_{m_{i1}-1}}^{\delta_{m_{i1}}}, \int_{w_2=\delta_{m_{i2}-1}}^{\delta_{m_{i2}}}, ..., \int_{w_J=\delta^{m_{iJ}-1}}^{\delta_{m_{iJ}}} \phi_J(w_1, w_2, ..., w_J|\Sigma) dw_1, dw_2, ..., dw_J \right] \tag{4}$$

where $\delta_{m_ij} = \frac{\mu^{m_{ij}}-\alpha-\beta' X_{ij}}{\tau}$ and ϕ_J is the standard multivariate normal distribution of dimension J and $w_1, w_2, ..., w_J$ are the normalised means. The dimensionality of integration may often rule out the use of MSL for estimating the AREOPROBIT. For example, in the context of the application presented in this paper, the full information likelihood estimation has the order of 27 dimensions of integrations. Such a model would take weeks to converge and would be very prone to simulation errors (Paleti and Bhat, 2013). These issues are easily circumvented by the CML estimation approach, which, in the context of this paper, only entails the evaluation of pairs of bivariate normal probabilities.

2.3 The CML autoregressive random effects ordered probit model

As previously introduced, the CML functions presented in this paper are pairwise-likelihood functions formed by the product of likelihood contributions of varying subsets of pairs of observed events. A typical pairwise log-likelihood function for the i^{th} individual corresponds to:

$$logL_i(\theta) = log\left(\prod_{J-1}^{J} \prod_{J}^{g=j+1} \left[Pr(y_{ij} = m_{ij}, y_{ig} = m_{ig}) \right] \right) \tag{5}$$

$$
\begin{aligned}
&Pr(y_{ij} = m_{ij}, y_{ig} = m_{ig}) \\
&= \phi_2(\delta_{mij}, \delta_{mig}, \zeta_{ig}) - \phi_2(\delta_{mij}, \delta_{mig-1}, \zeta_{ig}) \\
&- \phi_2(\delta_{mij-1}, \delta_{mig}, \zeta_{ig}) + \phi_2(\delta_{mij-1}, \delta_{mig-1}, \zeta_{ig})
\end{aligned} \tag{6}
$$

The pairwise estimator obtained by maximizing the logarithm of the function in Equation (4) with respect to the vector

It is worth noting that (6) can be rapidly evaluated using the rectangle properties of the bivariate normal distribution. Varin and Czado (2009) indicate that the CML estimator is consistent and asymptotically normally distributed. The asymptotic variance covariance matrix is given by the Godambe sandwich information matrix (Godambe, 1960; Zhao and Joe, 2005). The CML formulation is remarkably short and simple in comparison to its MSL counterpart. However, Paleti and Bhat (2013) as well as Varin and Czado (2009), among others, have proven that it is as able as the MSL approach to estimate the model parameters while being less prone to convergence issues.

2.4 Model composition and selection for CML estimation

2.4.1 Classical approach - Model composition

As previously discussed in the introduction, the full-pairwise marginal likelihood function also presented in equation (5) requires to evaluate $J \times (J-1)/2$ pairs of bivariate normal probabilities, which we described in the introduction as a full-pairing approach. A full pairing approach is efficient and computationally affordable when the number of time periods is moderate[2]. However, the full pairing approach becomes more computationally intensive as the number of time periods increases. As previously discussed, a full-pairing approach applied to a hypothetical balanced dataset featuring 10 time periods per individual requires to evaluate 45 pairs of bivariate normal probabilities. However, this number goes up to 190 when 20 time periods need to be considered and 1225 for a 50 time periods case, which becomes computationally unaffordable and potentially inefficient. Indeed, a recent stream of studies proved that there may be no need to make use of all the possible pairs, because pairs formed from the closest observations provide more information than distant pairs. This has been found to be true in both temporal and spatial contexts (Bhat et al., 2014; Varin and Vidoni, 2005). Bhat et al. (2014) suggests that the optimal maximum distance between pairs can correspond to the value that minimises the trace (or the determinant) of the asymptotic variance-covariance matrix of the larger model considered (*i.e.* the model which features the largest number of covariates). A similar proposition has been made by Varin and Vidoni (2005). Bhat et al. (2014) and Varin and Vidoni (2005) simply suggest to start with a low value of the distance threshold (which requires to evaluate a small number of pairs in the CML function) and increase the distance threshold up to a point where increasing it does not improve the

[2]The definition of moderate depends on the context, the sample site, etc.

trace, or even reduces it. In the current paper, we will refer to this protocol as the 'close pair approach'.

2.4.2 Classical approach - Model selection

Once the optimal distance between pairs has been determined, Varin and Czado (2009) propose to compare whether some (or all) of the models nested in the most general model exhibit a better goodness-of-fit. As previously mentioned, the log-likelihood value derived from a CML function is ill-specified and should not be used for comparing models. (Varin and Vidoni, 2005) propose to use the Composite Likelihood Information Criterion (CLIC). The CLIC is defined as a direct generalization of the Akaike criterion (Akaike, 1998) for model selection in a CML context and corresponds to:

$$CLIC = log_{CML}(\hat{\theta}) - tr[\hat{J}(\hat{\theta})\hat{J}(\hat{\theta})^{-1}] \tag{7}$$

where $\hat{\theta}$ is a vector of estimated parameters obtained by maximising equation (5), and $\hat{H}$ and $\hat{J}$ are the bread and the meat of the Godambe sandwich matrix. Obviously, it is necessary for model comparison with such approach to fit all the considered models with a pairwise likelihood function based on the same pairs of observations for each model. The model which exhibits the lower CLIC should be the selected one.

The classical model selection approach has great merits when using close pairs, but the fact that it can not be used for comparing models estimated using different pairs limits its use. We now introduce an alternative approach which can be used to evaluate and compare the goodness-of-fit of CML AREOPROBIT models estimated using any arbitrary combination of pairs for each individual and each model. This new model selection approach also allows to use different model composition approaches which are introduced below.

2.4.3 Proposed approach - Model composition

2.4.3.1 Random pairs

An alternative to the close pair approach is to use a given number of randomly selected pairs for each individual. Such approach has already been proposed (although not tested) by (Paleti and Bhat, 2013) for estimating AREOPROBIT models, while other authors have used and tested similar approaches in different context. Engle et al. (2008) have proposed to use random pairs to estimate complex ARCH-type models using CML-like methods. However, the type of models, the data and the field of study are different from the context of our paper.

There are several reasons why using random pairs may be a good alternative to using close pairs. Firstly, although it is true that close pairs are more informative than distant pairs, it may not be systematic in all contexts and for all individuals in a given context. Secondly, using random pairs may allow to cover all the possible combinations of pairs

for a given dataset, providing that the sample size is large enough, which may result in better estimates in some cases. Thirdly, it may be computationally more efficient to use random pairs.

A very simple way of selecting purely random pair would be to generate, for each individual, scores S corresponding to random numbers draws from a uniform distribution with mean 0 and standard-deviation 1. The researcher can then decide that the Pth pairs which received the highest score for each individual should be retained for model estimation.

One limitation of using purely random pairs is that all the possible pairs have the same chances to be selected. As a result, uninformative pairs have as much chances to be selected as informative pairs. Informative pairs may not only be close pairs though, and other criterion may apply depending on the data considered. In order to mitigate the issues associated with selecting uninformative pairs, we propose a new 'smart pairs' approach.

2.4.3.2 Smart pairs

The smart pairs approach can simply be described as a random pairs approach where some pairs have more chances to be randomly selected than others, depending on an arbitrary set of criterion set by the researcher given the data at hand. The first and main criteria which should be considered for a wide range of longitudinal datasets is distance between pairs (in terms of temporal units). Closest pairs should have more chances to be selected than more distant pairs. One can for example decide that the score S receives a penalty based on the distance between the two corresponding pairs, for example $S_{smart1} = S^{distance}$. Moreover, in the context of this paper, we find that some types of pairs are more informative than others. More precisely, we find that introducing a higher proportion of pairs for which the dependent ordered outcome is different between the two periods provides a better model fit (which is measured by a new criteria introduced below). A suggestion is to multiply the random score by 2 for pairs which feature different levels for the dependent ordered outcome. The exact way of selecting the criterion for smart pairs should be based on analysing the results from random pair and close pair models, and depends on the data at hand. As stated by (Bhat et al., 2010), how to exactly build a CML function remains an open research area where it is generally agreed that the CML construction should be based on balancing statistical and computation efficiency.

2.4.4 Proposed approach - Model selection

As previously discussed, the CLIC is not a suitable indicator for comparing CML models formed by a different set of pairs. In this context, we propose to use another criterion which we name the Underlying Log-Likelihood (ULL). More precisely, we note that a CML function is a simplification of an underlying 'true' model which can not be estimated *via* MSL because of computational issues. However, it is possible in most cases to

derive the value of the log-likelihood given by the likelihood function of the underlying model using the vector of parameter values $\hat{\theta}$ derived from simpler CML models. Such approach only requires to evaluate a multidimensional integral once for each competing model estimated *via* CML, which is computationally demanding but feasible even for high dimensional gaussian vectors (Genz and Bretz, 2009). This approach can be easily implemented using any statistical software package. It only requires to specify an MSL function for a given model in addition to a CML function and input the parameter values at convergence for the CML model in the MSL model. Moreover, this approach can be used to compare CML models built using different pairs as well as CML models built using different covariates (or both). indeed, the ULL can also potentially be used to compute additional indicators such as the Underlying Akaike Information Criterion (UAIC) which is simply given by $UAIC = 2C - 2ULL$ with C the number of estimated parameters in the model. In the remainder of the paper, we use both simulation and real-world data to analyse the properties of the ULL and test how it compares to other goodness-of-fit indicators in the context of CML estimation.

3 Empirical work

Car availability over the life course is a well suited topic to study the properties of AREOPROBIT CML models built from close or random pairs. Indeed, as previously stated, there are strong assumptions that car availability implies a strong degree of spurious state dependency. Car availability may be driven by a lot of unobserved factors which are correlated over time. For example, Dargay (2001) reports that car ownership is clearly associated with habit and resistance to change and that it is difficult to abandon even if the economic consequences of having a car available may not evolve favourably for the owner. These factors may be difficult to measure, especially in a panel setting. It is also worth underlining that the use of autoregressive models for investigating spurious state dependence in car availability or car ownership status is not a novelty. However this is the first time to our knowledge that a AREOPROBIT model is used to model mobility biography data as described in the next paragraph. Previous attempt to use autoregressive models for analysing car availability or car ownership have either focused on microeconomic panel data, merging different data registries or pseudo-panel data obtained by the mean of consumption surveys (see Dargay (2001)). On the other hand, life course calendar data have been often analysed by the mean of static random-effects models, thus ignoring the potential effects of state dependency. As previously stated, the empirical aim of this paper is to reconciliate the richness of the insights provided by life course calendar data, as argued in the previous section, together with the behavioural realism brought by accounting for spurious state dependency in discrete choice models. In this section, we describe the data collection effort and provide details on the data sample we use for modelling car availability status over the life course.

3.1 Survey design

The data originates from a retrospective survey which is carried out since 2007 at the Department of Transport Planning of the TU Dortmund as an annual homework. Since 2012 it is part of the collaborative project "Mobility Biographies: A Life-Course Approach to Travel Behaviour and Residential Choice" and additional data is collected in Frankfurt and Zurich. The survey addresses the students of the seminar, their parents and grandparents. The students are asked to give the questionnaire to both their parents and two of their grandparents - who are randomly chosen, one from the maternal and one from the paternal side. If one of the family members is not available for any reason the students can alternatively recruit another person preferably of the same generation. The questionnaire, which is the same for every generation, contains a series of retrospective questions about residential and employment biography, travel behaviour and holiday trips as well as socio-economic characteristics.

As the sample has a unique structure it is not possible to appraise representativeness (see Erickson (1979) for problems with representativeness in similar surveys). The majority of the respondents lives in Dortmund (North Rhine-Westphalia), one of the most densely populated regions of Germany. Furthermore within the grandparent generation a bias to female participants who live longer on the one hand and are also often younger, more popular and communicative can be recognized (Scheiner et al., 2014). Finally retrospective data especially collected for a long period as the life course always bears the risk of the so called memory bias which means a unintended or voluntary bias of the autobiographic memory (Manzoni et al., 2010). However the whole study focusses on mobility behaviour in the life-course thus the results are not expected to be significantly affected by the differences between the sample and the population. A more detailed documentation of the data set can be found in Scheiner et al. (2014).

3.2 Data sample

In this paper, we focus on the parents' generation. Our study sample features 930 individuals and 27 observations per individual Each observation related to one individual and one year. Hence, the total number of observations is 25,110. We include all the individuals from the parents generations who were at least 18 in 1980. We chose to focus on the parents' generations because it provides the longest series of observations in a contemporary setting. For each year, participants were asked to report their car availability status and had to choose between four propositions: "Never", "Sometimes", "Often" or "Always"[3]. Hence, car availability can be considered as an ordered outcome. The evolution of car availability over the life course for our study sample is reported in Figure 1.

[3]In this paper, we have decided to merge "Often" and "Sometimes" in the same category ("Sometimes") because the number given for these categories have been found to be much smaller. Preliminary models have been estimated in order to assess the impact of this decision on model estimates. We found that this does not substantially affect conclusions, especially given the purpose of our paper.

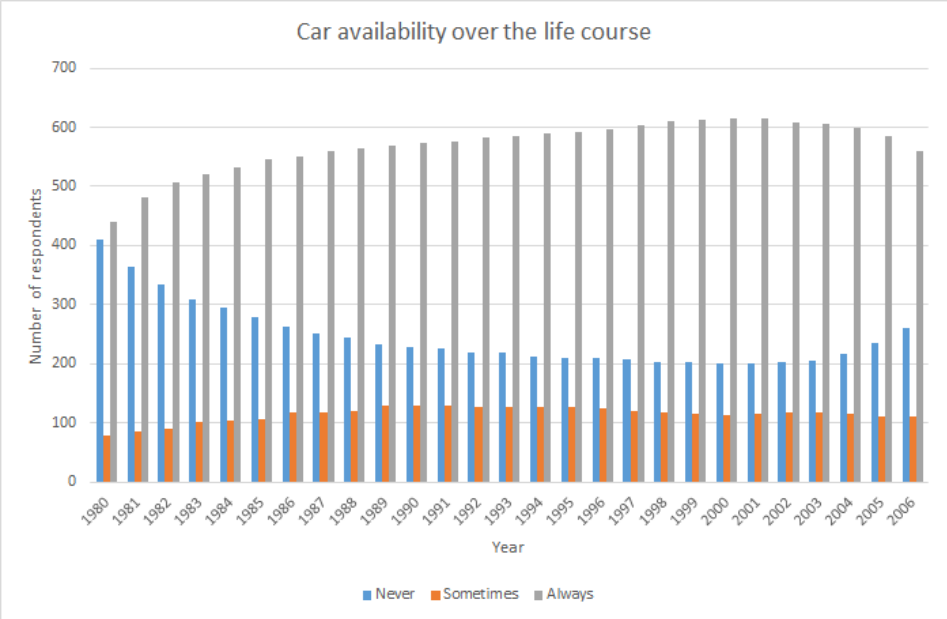

Figure 1: Evolution of the car availability status over the life course

Figure 1 shows that car availability tends to change over the life course. The first category, "Never", significantly decreases between 1980 and 1990 while "Sometimes" and most importantly "Always" increase. There is little evolution between 1990 and 2006, which suggests the presence of state dependence and justifies the need to account for autocorrelation. Our study sample also features annual measures on distance from work location, education level, number of children, residential location, marital status, driving license ownership and home ownership status. These variables will be used as independent variables in the AREOPROBIT models. TABLE 1 below provides descriptives statistics and clarifies how these variable have been coded.

3.3 Modelling strategy

The variables introduced in Table 1 are used as independent variables in a series of AREOPROBIT models estimated *via* CML. In addition, we also estimate two threshold parameters μ_1 and μ_2 described in (2). Since the dependent variable, car availability status, car take one value among three for each individual and each year, it means that we normalise the constant to zero. Finally, we estimate an autocorrelation parameter, ρ, which we bound between 0 and 1 by the means of a logistic transformation.

We estimate a series of CML models which can be divided in three groups. The first group includes the models that are estimated using increasingly distant pairs of bivariate normal probabilities. We estimate 26 models with all the possible maximum distances between pairs, which also includes a full-pairwise likelihood approach featuring 351 pairs.

Table 1: Descriptive statistics

Variable	**Description**	**Mean**	**Std. Dev.**	**Min**	**Max**
car availability	= 0 if car is a car is "never available", 1 if "sometimes available" and 2 if "always available"	1.348	0.869	0	2
driving license	= 1 if respondent has a driving license	0.772	0.419	0	1
distance work	Distance from work, in Km	8.265	14.938	0	150
female	= 1 if respondent is female	0.500	0.500	0	1
age	Age, in years	30.607	14.718	0	69
own home	= 1 if respondent owns her home	0.383	0.486	0	1
children	Number of children	1.293	1.173	0	9
married	= 1 if respondent is married	0.697	0.460	0	1
higher education	=1 if respondent has a university degree	0.249	0.432	0	1
urban	= 1 if respondent lives in a city (population >= 300,000 inhabitants)	0.388	0.487	0	1

The second group includes 30 models estimated using an increasing number of randomly selected pairs for each individual. The first 10 models use 26 random pairs, the second 10 models use 120 pairs, and the last 10 models use 215 pairs. Finally, we estimate 30 models using smart pairs. As previously introduced, we propose 3 different types of smart pairs in this approach and estimate 10 model for each type of smart pairs:

1. Smart1: we generate a score S which receives a malus based on the distance between the two corresponding pairs so that $S_{smart1} = S^{distance}$. For each individual, we only consider the 120 pairs reporting the highest score S_{smart1}. This approach is a good compromise between a random pairs approach and a closest pairs approach in the sense that it recognises that closest pairs are more informative but also that distant pairs should not simply be excluded from the analysis.

2. Smart2: we generate a score S which receives a malus based on the distance between the two corresponding pairs and a bonus if the level of stated car availability is different between the two elements of the pairs so that $S_{smart2} = S^{distance} \times 2 \times YD$ with YD an indicator function which takes the value 1 if the level of stated car availability is different across the two time periods for a given individual and 0 else. The idea behind this approach is that although Smart1 is an improvement over a purely random pairs approach, it also reduces the chances of selecting pairs which feature a different level of stated car availability because close pairs tend to be similar. Hence, the model may report a poorer fit for the pairs which feature different levels of stated car availability unless an arbitrary bonus is given to the score S for these pairs.

3. Smart3: this approach is simply the same as Smart2 but the bonus given to pairs which feature different levels of stated car availability varies depending on the magnitude of the difference between the two pairs. In this study, stated car availability

is coded in such a way that it can take one value among three: 0, 1 or 2. We propose $S_{smart3} = S^{distance} \times 2 \times YD2 \times 3 \times YD3$ with $YD2$ an indicator function which takes the value 1 if the absolute value of the difference between the stated car availability levels of the two elements of the pair considered is equal to 1 and $YD3$ another indicator function which takes the value 1 if this difference is equal to 2.

4 Results

In this section, we first comment on the results from the CML models estimated using varying distances between pairs. Secondly, we discuss the results from the models estimated using an increasing number of random pairs for each individual. Thirdly, we introduce the smart pairs approach. Finally, we give detailed results for the best model for each approach and compare parameter estimates. Detailed model results for all the models are available in the Appendix section.

4.1 Closest pairs models

Model results for the closest pairs models are reported in Table 2 and Figure 2 below. As previously discussed, we follow Bhat et al. (2010) and Varin and Czado (2009) and report the trace of the robust variance-covariance matrix for each model. The optimal maximum distance between pairs in this context corresponds to the model which minimises the trace. Moreover, we also report the ULL for each model. The simplest model only uses pairs which are not distant by more than one time unit (year). The trace of the robust variance-covariance matrix is high (0.899), which suggests that it is necessary to move toward a model specification which makes use of more pairs. We find that increasing the maximum distance between pairs decreases the trace and increases the ULL. Unsurprisingly, we find mediocre values for the trace and the ULL when the maximum distance between pairs is severely restricted. The model which exhibits the best trace, 0.430, allows a maximum distance between pairs equals to 19. Interestingly, this is also the model which reports the best ULL (-4397.621). Finally, we find that the full-pairwise approach is not the most efficient model specification because it uses all the pairs, but reports the worst trace (1.107) and a mediocre ULL (-4410.292).This is a common result in the literature as indicated by Varin and Czado (2009), among others, and it confirms that a full-pairwise likelihood approach is not necessarily the best modelling strategy.

Looking at both the trace and the ULL, we find that these measures follow the same trend overall and, more importantly, lead to the same conclusions. The ULL tends to increase when the trace decreases and the ULL starts decreasing again when the trace increases. Although we find that the model for which the maximum distance between pairs is the best in terms of both trace and ULL, we also report that the worst model in terms of ULL (Model1, -5312.566) does not necessarily correspond to the worst model in terms of trace (Model26, 1.107).

Looking at the computational effort required to estimated the models, we note that the best model requires to evaluate 323 bivariate normal cumulative density functions, which is only 28 less than a full-pairwise likelihood approach. We note that some simpler models perform almost as well as model19. For example, model15 provides both a reasonably high ULL (-4402.696) and a low trace (0.447) although it requires to evaluate only 285 pairs instead of 323 (38 pairs less). Altogether, the results of the closest pair models indicate that achieving the best model given the goodness-of-fit indicators considered is computationally intensive. Models built under more flexible assumptions in terms of which pairs should be considered may provide better performance while being less computationally demanding. We now turn our head to the models estimated using a varying number of random pairs.

Table 2: Model results: Closest pairs

Max. distance	**Nb. of pairs**	**ULL**	**Trace**
1	26	-5312.566	0.899
2	51	-4998.658	0.871
3	75	-4794.307	0.734
4	98	-4655.958	0.647
5	120	-4587.666	0.593
6	141	-4541.178	0.551
7	161	-4509.928	0.519
8	180	-4478.874	0.497
9	198	-4452.856	0.479
10	215	-4435.852	0.464
11	231	-4422.257	0.461
12	246	-4415.112	0.456
13	260	-4409.818	0.453
14	273	-4405.660	0.450
15	285	-4402.696	0.447
16	296	-4400.296	0.445
17	306	-4398.429	0.439
18	315	-4398.026	0.435
19	**323**	**-4397.621**	**0.430**
20	330	-4399.111	0.431
21	336	-4401.220	0.478
22	341	-4402.943	0.520
23	345	-4402.543	0.681
24	348	-4405.338	0.785
25	350	-4409.651	0.831
26	351	-4410.292	1.107

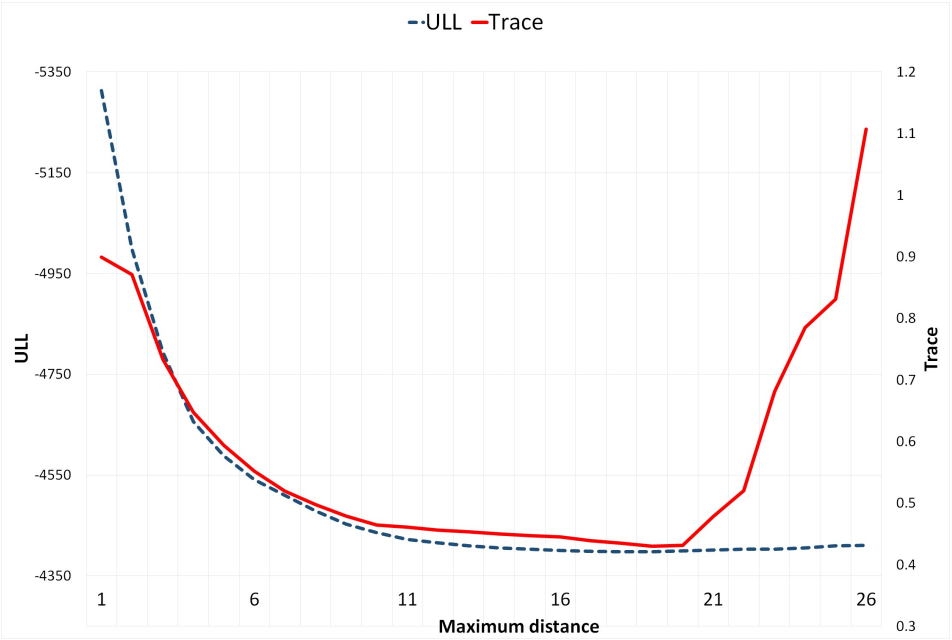

Figure 2: Close pair models: ULL *versus* fit

4.2 Random pairs

As previously introduced, 3 different numbers of random pairs are considered: 26 pairs (which corresponds to a closest pairs model with a maximum distance of 1), 120 pairs (which corresponds to a closest pairs model with a maximum distance of 5) and 215 pairs (which corresponds to a closest pairs model with a maximum distance of 10). Results are reported in Table 3. below.

Table 3: Random pair model results

Model	1	2	3	4	5	6	7	8	9	10
ULL										
26 random	-4364.190	-4401.491	-4395.057	-4481.351	-4416.579	-4412.082	-4466.592	-4425.905	-4438.316	-4409.920
120 random	-4387.262	-4387.608	-4423.530	-4401.674	-4405.906	-4401.775	-4411.283	-4397.174	-4410.141	-4414.419
215 random	-4412.567	-4410.702	-4405.611	-4403.064	-4425.659	-4407.589	-4399.523	-4406.376	-4421.490	-4405.170
Trace										
26 random	0.899	0.883	2.950	1.566	1.294	1.141	0.153	0.531	0.860	0.557
120 random	0.586	0.344	0.253	0.599	0.607	0.423	0.607	0.356	0.519	0.282
215 random	0.780	1.074	0.519	0.774	0.548	0.513	0.769	0.546	0.713	0.592

One of the main result we observe is that the best model estimated using 26 random pairs outperforms all the model estimated using closest pairs in terms of ULL. However, we also note that the ULL for the model estimated using 26 random pairs varies a lot across models. The best ULL is equal to -4364.191 while the worst is -4481.354 which is a result close to a model estimated using 8 closest pairs (180 pairs in total). The trace values are generally high and vary following an erratic pattern, which is a result one

would expect when using random pairs. We note that the trace is an indicator of limited value when using random pairs.

3 of the 10 models estimated using 120 random pairs are found to provide a better ULL than the best model estimated using closest pairs (-4387.262, -4387.608 and -4397.174 *versus* -4397.621). Moreover, the results are found to be more stable than in the 26 random pairs case, with the worst ULL being -4423.530. We note that the trace greatly decreases when using 120 random pairs and detailed parameter estimated reported in the Appendix show that the parameters' standard deviations improve greatly when using 120 random pairs instead of 26.

Finally, none of the 10 models estimated using 215 random pairs outperform the best model estimated using closest pairs. However, we note that the standard deviation for the sample of ULL values derived from this model is the lowest (8.18, *versus* 11.49 for the 120 random pair models and 34.23 for the 26 random pairs models), which is expectable. The higher the number of random pairs one use, the less the results will vary across models.

The average ULL for the models estimated using 26 random pairs is -4421.148, while it is -4404.077 for the models estimated using 120 random pairs and -4409.775 for the models estimated using 215 random pairs. On average, these results are worst than the best model estimated using closest pairs. However, some of the results reported above are better and model estimation is much less computationally demanding, especially in the 26 and 120 random pair cases. These results suggests that, in some cases, random pair models outperform closest pair models and that there are some combinations of pairs which perform better than others. Given the results of the random pair models, we propose to build smart pairs models using 120 pairs. Indeed, 120 random pairs appears to be a good compromise between goodness-of-fit, computational burden and stability of the parameters across models built from different random pairs.

4.3 Smart pairs

Results for the 30 smart pairs models are given in Table 4. below. As previously described, each model features 120 pairs.

Table 4: Smart pairs model results

Model	1	2	3	4	5	6	7	8	9	10
ULL										
120 smart 1	-4363.697	-4352.991	-4361.896	-4372.898	-4374.696	-4358.071	-4365.463	-4367.166	-4362.807	-4365.965
120 smart 2	-4307.253	-4310.092	-4303.398	-4312.946	-4301.914	-4306.100	-4306.074	-4312.558	-4310.585	-4310.323
120 smart 3	-4290.452	-4290.308	-4283.146	-4287.199	-4295.456	-4292.902	-4294.826	-4294.826	-4288.981	-4288.401
Trace										
120 smart 1	0.449	0.498	0.557	0.435	0.425	0.383	0.422	0.452	0.450	0.582
120 smart 2	0.613	0.747	0.515	0.613	0.495	0.620	0.507	0.624	0.498	0.601
120 smart 3	0.594	0.636	0.051	1.247	0.716	0.938	0.721	0.860	0.401	0.677

Table 4 reveals that most of the 10 Smart1 models outperform the best model estimated using random pairs in terms of ULL (-4364.191). Moreover, the detailed model results available in Appendix (Table 12) show that parameter estimates and T-ratios are

more stable across model estimated using different pairs than models built on purely random pairs. Such result is also confirmed by the fact that the ULL of Smart1 models is more stable, with the maximum corresponding to -4352.991 and the minimum to -4374.696.

The results above are improved by introducing different weights depending on whether pairs feature different levels of stated car availability or not. Indeed, each one of the Smart2 models largely outperform the best closest, random and smart pair models previously introduced. The average ULL is -4308.124, which is an improvement of about 56.441 ULL points in comparison to the Smart1 models. Moreover, the ULL is more stable across models and T-ratios appear slightly higher overall.

Finally, the results from the Smart3 models, where more weight is given to pairs which feature more different levels of stated car availability, show that the ULL has further improved for all the 10 models estimated. More precisely, the best Smart3 model outperforms all the models previously discussed in this paper with a ULL value equal to -4283.146. The average ULL is -4290.650, which is 17.475 points better than what is found for the Smart2 models. Paramter estimates are about equally stable than previous smart pairs models.

Overall, these results indicate that smart pairs outperform the closest pairs approach as well as the random pairs approach. Moreover, different selection processes for the smart pairs lead to different outputs, with the best output provided by smart pairs where the semi-random selection puts more weight on closest pairs and pairs which feature more different levels of stated car availability.

4.4 Empirical results

We now analyse in details the results obtained from the best closest and smart pair models and compare them to the results obtained from a simple random effects probit where autocorrelation is set to 0. Results are reported in Table 5 below.

The results obtained from the two CML models justify the use of an AREOPROBIT approach because the autocorrelation coefficient, ρ, is positive and significant. The goodness-of-fit of the two CML models is also much better than the fit of the simple random effects ordered probit. Moreover, the ULL of the Smart3 model is more than a 100 points better than the ULL of the Closest19 model, although the estimation of the Smart3 model only requires to evaluate 120 pairs while estimating the Closest19 model requires to evaluate 323 pairs. The parameters vary a lot across models. All the parameters apart from the autocorrelation coefficient, ρ, are lower for the Smart3 model. T-ratios (*i.e.* standard deviations of the parameters) are stable across CML models, while the MSL reoprobit model leads to different conclusions. We now comment on the results of the Smart3 model, which report the best goodness-of-fit.

The value of the autocorrelation coefficient (3.290, which corresponds to 0.961 for the transformed parameter) is very high, which indicates strong autocorrelation between errors across years. This result suggests that spurious state dependence has a strong influence on car availability status. In other words, ρ indicates that there are strong,

Table 5: Best model results

Model	**MSL reoprobit**		**CML - Closest 19**		**CML - Smart 3**	
Nb. of pairs	NA		323		120	
ULL	-6763.324		-4397.62		-4283.146	
Trace	0.621		0.430		0.051	
	Coeff.	**T-ratio**	**Coeff.**	**T-ratio**	**Coeff.**	**T-ratio**
$female$	-2.035	-7.572	-1.288	-5.32	-0.981	-5.02
age	0.349	7.961	0.273	6.69	0.220	6.08
age^2	-0.405	-7.239	-0.316	-6.53	-0.253	-5.91
$children$	-0.127	-1.866	-0.186	-3.90	-0.137	-3.47
$degree$	0.022	0.081	0.265	1.68	0.220	1.78
$distw$	0.035	4.689	0.024	3.92	0.020	3.86
$distw^2$	-0.023	-2.723	-0.017	-3.14	-0.014	-3.23
$car\_license$	3.678	5.875	2.200	5.34	1.697	5.30
big_city	0.004	0.021	-0.126	-1.21	-0.103	-1.27
own_home	0.020	0.161	-0.091	-1.18	-0.056	-0.91
$married$	0.077	0.605	-0.037	-0.45	-0.009	-0.14
μ_1	1.732	18.326	1.475	9.82	1.293	8.05
μ_2	0.513	8.281	0.017	0.1	-0.279	-1.48
σ	1.063	20.362	0.323	1.41	-0.161	-0.43
ρ	0.000	NA	2.899	8.68	3.290	9.67

constant unobserved factors which persistently influence our dependent variables across years. In the context of car availability, this means that previous car availability status, attitudes, lifestyles and habits have a strong influence on mobility behaviour. This result also highlights one of the limits of the life course calendar approach. The life course calendar approach allows to gather a large amount of longitudinal data at a fraction of the cost of a repeated panel survey such as the British Household Panel Survey (Taylor et al., 1993). However, it doest not allow to collect accurate data on attitudes and other latent factors.

Our other results indicate that women are significantly less likely than men to report a higher level of car availability. Older people are more likely to report a higher level of car availability although this effect decreases over time as indicated by the variable age^2. Respondents with children are less likely to report a higher level of car availability. Having a higher education degree has not been found to be significant while having a driving license is a strong predictor (which is of course a very common result in the car ownership/car availability literature). The distance from work has a positive impact on the dependent variables although this effect decreases after a certain point. Living in a big city, being a home owner or being married do not significantly influence car availability. Altogether, these results suggest that more research is needed to understand which are the latent factors that influence car availability and how do they evolve over time. As

previously mentioned, many studies such as Van Acker et al. (2014) have investigated the effect of latent factors on car availability. However, many of these research do not provide information on how car availability evolves over time based on these factors.

4.5 Further insights and results validation

Finally, we propose to complete our analysis by the means of a test which seeks to compare the closest pair approach and the random approach using simulated data. We do not include the smart pair approach in this comparison because our preliminary tests indicate that it is not substantially different from a random pairs approach when using simulated data. We use the results obtained from one of the best models (in terms of ULL) estimated using 120 random pairs and we simulate 10 datasets based on these results (in other words, the true value of the simulated parameters corresponds to the value of the estimated parameters of the model aforementioned). Moreover, we repeat this protocol using the results from the closest pairs model with a maximum distance of 5 (120 pairs in total).Each dataset features 500 simulated individuals over 27 time periods. Details about the method we used to generate the simulated datasets can be found in Paleti and Bhat (2013). We then estimate 110 models:

1. Random Recover Closest (RRC): We estimate 10 random pairs models for each one of the 10 datasets simulated using the closest pairs model results. We use 120 random pairs for each simulated individual.

2. Closest Recover Random (CRR): We estimate 1 closest pairs model for each one of the 10 datasets simulated using the 120 random pairs model results. We use all the pairs for which the distance is inferior or equal to 5 (120 pairs in total).

The results of the parameter recovery tests are reported in Table 5 below. We report the mean value of the estimated parameters across model as well as the mean Root Mean Square (RMS). Other indicators such as the Average Percentage Bias (APB) have also been used and led to the same conclusions.

Several results are worth noting. Firstly, the RRC models are found to perform well for most of the parameters. The mean RMS is only found to be high for the parameters for which the true value is close to zero. More importantly, the average RMS is found to be 0.146 for the autocorrelation parameter ρ, which is acceptable given the context of the test. The minimum of the RMS across parameters and models is found to be 0 or close to 0 for all the parameters. The maximum varies a lot across parameters. Unsurprisingly, it is found to be the highest for the autocorrelation parameter. The CRR models do not perform as well as the RCC models. The mean RMS values are found to be much higher than for the RRC models for 11 parameters out of 15. Moreover, the CRR approach seem to only perform better for the parameters for which the value is either low or non-significant. The minimum RMS measures are found to be close to 0 for only 4 parameters while the maximum RMS measures are above the ones derived from the RRC models in

Table 6: Test 2 - Simulation results

MODELS RRC - Recover parameters from 5 closest pairs models (120 pairs in total) using 120 random pairs					
	True value	Mean value across models	RMS mean	Min RMS	Max RMS
$female$	-2.427	-2.485	0.024	0.000	0.739
age	0.532	0.546	0.027	0.000	0.192
age^2	-0.617	-0.616	0.003	0.001	0.252
$children$	-0.313	-0.341	0.089	0.000	0.208
$degree$	0.519	0.544	0.048	0.000	0.226
$distw$	0.041	0.064	0.560	0.000	0.161
$distw^2$	-0.028	-0.011	0.588	0.001	0.146
$car\_license$	3.950	4.048	0.025	0.002	1.091
big_city	-0.197	-0.160	0.188	0.000	0.243
own_home	-0.130	-0.116	0.109	0.001	0.094
$married$	-0.060	-0.070	0.157	0.000	0.118
μ_1	2.150	2.169	0.009	0.000	0.249
μ_2	0.598	0.623	0.043	0.000	0.250
σ	1.101	1.119	0.017	0.000	0.328
ρ	1.336	1.142	0.146	0.003	1.774
MODELS CRR - Recover parameters from 120 random pairs models using 5 closest pairs (120 pairs in total)					
	True value	Mean value across models	RMS mean	Min RMS	Max RMS
$female$	-1.079	-1.918	0.839	0.439	1.392
age	0.2295	0.408	0.179	0.091	0.313
age^2	-0.2657	-0.443	0.177	0.073	0.323
$children$	-0.1598	-0.259	0.099	0.020	0.156
$degree$	0.2052	0.378	0.173	0.055	0.322
$distw$	0.0212	0.037	0.028	0.000	0.055
$distw^2$	-0.0143	-0.003	0.035	0.002	0.088
$car\_license$	1.8826	3.300	1.418	0.768	2.065
big_city	-0.1063	-0.173	0.074	0.013	0.199
own_home	-0.0737	-0.136	0.062	0.008	0.135
$married$	-0.0328	-0.072	0.039	0.000	0.100
μ_1	1.3102	1.868	0.558	0.342	0.819
μ_2	-0.1595	0.400	0.560	0.306	0.780
σ	0.0322	0.829	0.797	0.547	1.107
ρ	3.2765	1.897	1.379	0.861	1.924

10 cases out of 15. Finally, the mean RMS for ρ is about 10 times higher in the CRR case in comparison to the RRC case.

Overall, these results indicate that the RRC models performed better than the CRR models, which may be due to the flexibility gains introduced by the use of random pairs. While the closest pairs approach only gives weights to what is considered to be the most informative pairs, the random pairs approach seeks to distribute weights equally across all the different combinations of pairs derived from the data at hand. As a result, the random pairs approach is better suited for accommodating data obtained from a wider range of generation processes while the close pairs approach can only perform well in a narrower set of contexts.

In addition to this simulation experiment, we also conducted a series of simulations which are not reported in this paper but available from the authors upon request. More

precisely, we used simulated data to compare the performance of closest pair, random pairs and smart pair models. Using the same simulated datasets, we tested whether these 3 approaches can adequately recover the parameter values. Our results showed marginal differences between approaches when using simulated data.

5 Conclusions

This paper focuses on model specification and selection in the context of CML estimation. Using a unique data-set which covers 27 years of car availability data in Germany, we provide new insights on how to build a pairwise likelihood function for estimating AREOPROBIT models. In addition, we suggest the use of a new goodness-of-fit indicator, the ULL, based on the intractable MSL model which underlines CML specifications. Finally, we provide new results on spurious state dependence in car availability levels using one of the longest panel data-set found on this topic in the literature.

Our results suggest that CML functions built using a random selection of bivariate normal probabilities for each individual in the sample are more efficient than functions built using closest pairs. CML functions composed of random pairs provide better results in terms of ULL and are computationally much less burdensome. This important result reinforces the appeal of CML estimation methods in comparison to MSL estimation methods.

Our results also suggest that a purely random approach to pairs selection may not always be desirable. Indeed, the performances of a purely random pairs approach greatly vary depending on which pairs are randomly selected. As a result, we introduce smart pairs, which is a semi-random pair selection process where some pairs have more chances to be randomly selected than other pairs, based on criterion set up by the researcher given the data at hand. The smart pair models introduced in this paper provide better results in terms of ULL and are more stable than purely random pair models. In other words, the smart pair models reduce the risks associated with using random pairs, that is to select uninformative pairs, while also improving model fit.

Finally, we explore the determinants of car availability over the life course with a specific focus on spurious state dependence. Our results strongly suggest that there exist unobserved factors which influence car availability and persist over years. We note that the life course calendar approach, which is the data collection method used in this paper, has many advantages over other survey design but can not recover essential information such as attitudes toward motorised mobility. As a result, we conclude that future efforts on this topic should consider new data collection approaches which can more adequately capture the information identified as spurious state dependence in our model.

The present study has some limits which should be acknowledged. Firstly, it only addresses the case of balanced panel datasets without missing values. Some adjustments may be necessary to apply the presented methodology in an unbalanced panel data context. More precisely, the selection of the smart pairs as well as the ULL measure may require to be modified to be used in such context. Moreover, the bivariate normal

probabilities may need to be weighted when using CML on an unbalanced panel (Joe and Lee, 2009; Kuk and Nott, 2000; Yi et al., 2011).

Future research on this topic should consider new ways of building CML functions. New criterion for selecting smart pairs and new weights should be tested. Random coefficient models built using smart pairs or random pairs should are also worth investigating. Moreover, CML estimation allows to use more complex autocorrelation structures which could provide more insights on how spurious state dependence evolves over time. For example, one could use a Toeplitz error structure instead of the classic AR(1) structure used in this paper and adapt the distance between pairs at different points in time based on how autocorrelation evolves over time. The CML approach is flexible and can be enriched in many ways as demonstrated by the recent stream of literature on this topic. Once again, how to build a CML function remains an open research area (Bhat et al., 2010).

6 Acknowledgements

The authors acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Selected papers of hirotugu akaike. Springer, pp. 199–213.

Axhausen, K. W., 2008. Social networks, mobility biographies, and travel: survey challenges. Environment and Planning B: Planning and design 35 (6), 981–996.

Bhat, C. R., 2011. The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models. Transportation Research Part B: Methodological 45 (7), 923–939.

Bhat, C. R., Sener, I. N., Eluru, N., 2010. A flexible spatially dependent discrete choice model: formulation and application to teenagers weekday recreational activity participation. Transportation research part B: methodological 44 (8), 903–921.

Bhat, C. R., et al., 2014. The composite marginal likelihood (cml) inference approach with applications to discrete and mixed dependent variable models. Foundations and Trends® in Econometrics 7 (1), 1–117.

Cherchi, E., Manca, F., 2011. Accounting for inertia in modal choices: some new evidence using a rp/sp dataset. Transportation 38 (4), 679.

Clark, B., Lyons, G., Chatterjee, K., 2009. Understanding the dynamics of car ownership: Some unanswered questions.

Dargay, J., Hanly, M., 2007. Volatility of car ownership, commuting mode and time in the uk. Transportation Research Part A: Policy and Practice 41 (10), 934–948.

Dargay, J. M., 2001. The effect of income on car ownership: evidence of asymmetry. Transportation Research Part A: Policy and Practice 35 (9), 807–821.

Engle, R. F., Shephard, N., Sheppard, K., 2008. Fitting vast dimensional time-varying covariance models.

Erickson, B. H., 1979. Some problems of inference from chain data. Sociological methodology 10, 276–302.

Genz, A., Bretz, F., 2009. Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Godambe, V. P., 1960. An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31 (4), 1208–1211.

Golob, T. F., 1990. The dynamics of household travel time expenditures and car ownership decisions. Transportation Research Part A: General 24 (6), 443–463.

Hanly, M., Dargay, J., 2000. Car ownership in great britain: Panel data analysis. Transportation Research Record: Journal of the Transportation Research Board (1718), 83–89.

Heckman, J. J., 1981. Heterogeneity and state dependence. In: Studies in labor markets. University of Chicago Press, pp. 91–140.

Hensher, D. A., 2013. Dimensions of automobile demand: a longitudinal study of household automobile ownership and use. Elsevier.

Hensher, D. A., Greene, W. H., 2003. The mixed logit model: the state of practice. Transportation 30 (2), 133–176.

Joe, H., Lee, Y., 2009. On weighting of bivariate margins in pairwise likelihood. Journal of Multivariate Analysis 100 (4), 670–685.

Kitamura, R., Bunch, D. S., 1990. Heterogeneity and state dependence in household car ownership: A panel analysis using ordered-response probit models with error components. University of California Transportation Center.

Kuk, A. Y., Nott, D. J., 2000. A pairwise likelihood approach to analyzing correlated binary data. Statistics & Probability Letters 47 (4), 329–335.

Manzoni, A., Vermunt, J. K., Luijkx, R., Muffels, R., 2010. Memory bias in retrospectively collected employment careers: A model-based approach to correct for measurement error. Sociological methodology 40 (1), 39–73.

Müggenburg, H., Busch-Geertsema, A., Lanzendorf, M., 2015. Mobility biographies: A review of achievements and challenges of the mobility biographies approach and a framework for further research. Journal of Transport Geography 46, 151–163.

Paleti, R., Bhat, C. R., 2013. The composite marginal likelihood (cml) estimation of panel ordered-response models. Journal of choice modelling 7, 24–43.

Scheiner, J., Sicks, K., Holz-Rau, C., 2014. Generationsübergreifende mobilitätsbiografien–dokumentation der datengrundlage: Eine befragung unter studierenden, ihren eltern und großeltern. Dortmund, Deutschland: Technische Universität Dortmund.

Taylor, M. F., Brice, J., Buck, N., Prentice-Lane, E., 1993. British household panel survey user manual. University of Essex.

Train, K., 2001. A comparison of hierarchical bayes and maximum simulated likelihood for mixed logit. University of California, Berkeley, 1–13.

Van Acker, V., Mokhtarian, P. L., Witlox, F., 2014. Car availability explained by the structural relationships between lifestyles, residential location, and underlying residential and travel attitudes. Transport Policy 35, 88–99.

Varin, C., 2008. On composite marginal likelihoods. AStA Advances in Statistical Analysis 92 (1), 1–28.

Varin, C., Czado, C., 2009. A mixed autoregressive probit model for ordinal longitudinal data. Biostatistics 11 (1), 127–138.

Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statistica Sinica, 5–42.

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. Biometrika 92 (3), 519–528.

Yi, G. Y., Zeng, L., Cook, R. J., 2011. A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. Canadian Journal of Statistics 39 (1), 34–51.

Zhao, Y., Joe, H., 2005. Composite likelihood estimation in multivariate data analysis. Canadian Journal of Statistics 33 (3), 335–356.

7 Appendix

Table 7: Details outputs - Closest pairs models

Model results													
Max. distance	1	2	3	4	5	6	7	8	9	10	11	12	13
Nb. Of pairs	26	51	75	98	120	141	161	180	198	215	231	246	260
ULL	-5312.566	-4998.658	-4794.307	-4655.958	-4587.666	-4541.178	-4509.928	-4478.874	-4452.856	-4435.852	-4422.257	-4415.112	-4409.818
Trace	0.899	0.871	0.734	0.647	0.593	0.551	0.519	0.497	0.479	0.464	0.461	0.456	0.453
Coefficients													
female	-3.346	-3.062	-2.808	-2.577	-2.427	-2.307	-2.205	-2.088	-1.976	-1.877	-1.785	-1.721	-1.659
age	0.769	0.692	0.626	0.569	0.532	0.503	0.478	0.451	0.425	0.403	0.382	0.367	0.353
age^2	-0.898	-0.806	-0.728	-0.661	-0.617	-0.583	-0.554	-0.522	-0.492	-0.466	-0.442	-0.425	-0.409
children	-0.438	-0.392	-0.357	-0.328	-0.313	-0.303	-0.295	-0.285	-0.274	-0.264	-0.253	-0.246	-0.238
degree	0.801	0.701	0.624	0.560	0.519	0.483	0.457	0.430	0.407	0.386	0.367	0.353	0.340
distw	0.063	0.053	0.047	0.043	0.041	0.039	0.038	0.037	0.035	0.034	0.032	0.031	0.030
$distw^2$	-0.044	-0.037	-0.032	-0.029	-0.028	-0.027	-0.026	-0.025	-0.024	-0.023	-0.022	-0.021	-0.020
car_license	5.499	4.998	4.565	4.190	3.950	3.765	3.615	3.436	3.262	3.109	2.965	2.866	2.770
big_city	-0.300	-0.259	-0.229	-0.209	-0.197	-0.188	-0.184	-0.179	-0.174	-0.168	-0.162	-0.158	-0.154
own_home	-0.245	-0.177	-0.145	-0.131	-0.130	-0.131	-0.132	-0.130	-0.127	-0.123	-0.119	-0.116	-0.112
married	-0.024	-0.027	-0.049	-0.060	-0.060	-0.060	-0.059	-0.055	-0.051	-0.047	-0.043	-0.041	-0.039
μ_1	2.528	2.421	2.319	2.220	2.150	2.090	2.038	1.977	1.917	1.861	1.808	1.768	1.729
μ_2	0.887	0.800	0.722	0.647	0.598	0.556	0.520	0.473	0.424	0.378	0.332	0.298	0.264
σ	1.463	1.359	1.261	1.167	1.101	1.045	0.995	0.934	0.870	0.811	0.751	0.706	0.661
ρ	0.420	0.705	0.955	1.185	1.336	1.461	1.569	1.706	1.846	1.975	2.102	2.193	2.285
T-ratios													
female	-7.72	-7.27	-7.26	-7.12	-7.07	-7.06	-7.04	-6.92	-6.79	-6.67	-6.61	-6.48	-6.34
age	9.07	11.34	11.38	11.13	11.01	11	10.98	10.58	10.16	9.81	9.62	9.25	8.84
age^2	-8.44	-10.15	-10.25	-10.09	-10.05	-10.09	-10.11	-9.83	-9.51	-9.23	-9.08	-8.78	-8.43
children	-3.42	-3.64	-3.77	-3.87	-3.98	-4.1	-4.21	-4.27	-4.3	-4.32	-4.33	-4.31	-4.27
degree	2.21	2.01	1.96	1.92	1.89	1.85	1.82	1.8	1.79	1.77	1.76	1.75	1.73
distw	4.29	4.8	4.66	4.61	4.6	4.58	4.59	4.58	4.55	4.51	4.47	4.41	4.34
$distw^2$	-3.51	-3.56	-3.4	-3.31	-3.26	-3.24	-3.27	-3.31	-3.33	-3.32	-3.3	-3.28	-3.25
car_license	9.7	8.07	8.06	7.88	7.76	7.68	7.61	7.41	7.2	7.01	6.89	6.69	6.48
big_city	-1.15	-1.1	-1.09	-1.09	-1.08	-1.09	-1.1	-1.12	-1.14	-1.15	-1.17	-1.18	-1.18
own_home	-1.13	-0.98	-0.91	-0.91	-0.96	-1.01	-1.07	-1.1	-1.13	-1.15	-1.15	-1.16	-1.16
married	-0.12	-0.15	-0.31	-0.41	-0.44	-0.45	-0.46	-0.45	-0.43	-0.41	-0.4	-0.38	-0.38
μ_1	26.11	29.91	28.79	26.93	25.69	24.79	23.97	22.21	20.47	19.03	18	16.79	15.56
μ_2	12.02	8.98	7.73	6.49	5.82	5.32	4.89	4.25	3.63	3.1	2.65	2.29	1.95
σ	16.06	17.31	15.66	13.57	12.32	11.44	10.66	9.38	8.12	7.07	6.21	5.46	4.75
ρ	2.01	4.29	5.66	6.76	7.4	7.9	8.3	8.52	8.72	8.89	9.18	9.1	8.97

Table 8: Detailed outputs - Closest pairs models (cont.)

Model results													
Max. distance	14	15	16	17	18	19	20	21	22	23	24	25	26
Nb. Of pairs	273	285	296	306	315	323	330	336	341	345	348	350	351
ULL	-4405.66	-4402.7	-4400.3	-4398.43	-4398.03	-4397.62	-4399.11	-4401.22	-4402.94	-4402.54	-4405.34	-4409.65	-4410.29
Trace	0.45	0.447	0.445	0.439	0.435	0.43	0.431	0.478	0.519807	0.681	0.785	0.831	1.107
Coefficients													
female	-1.5986	-1.5384	-1.4778	-1.416	-1.3569	-1.2879	-1.236	-1.1594	-1.1094	-1.0814	-1.0268	-0.9585	-0.9589
age	0.3398	0.3266	0.3134	0.2998	0.2874	0.2729	0.2618	0.246	0.2356	0.23	0.2186	0.2043	0.2044
age^2	-0.3939	-0.3787	-0.3633	-0.3477	-0.333	-0.3162	-0.3032	-0.2847	-0.2726	-0.266	-0.2527	-0.2361	-0.2362
children	-0.2297	-0.221	-0.2124	-0.2023	-0.1959	-0.1857	-0.1786	-0.1685	-0.1615	-0.1578	-0.1501	-0.1403	-0.1402
degree	0.3271	0.3142	0.3021	0.2888	0.2789	0.2646	0.2546	0.2388	0.2298	0.2241	0.2136	0.1994	0.1995
distw	0.0293	0.0284	0.0274	0.0265	0.0256	0.0244	0.0237	0.0223	0.0216	0.0211	0.02	0.0188	0.0188
$distw^2$	-0.0198	-0.0192	-0.0187	-0.0181	-0.0175	-0.0168	-0.0164	-0.0155	-0.015	-0.0147	-0.0139	-0.0131	-0.0131
car_license	2.6776	2.5853	2.4925	2.3987	2.307	2.1995	2.1197	1.9966	1.9177	1.8768	1.7875	1.6739	1.6733
big_city	-0.1494	-0.1449	-0.1398	-0.135	-0.1303	-0.1255	-0.1212	-0.1144	-0.11	-0.108	-0.1027	-0.096	-0.0962
own_home	-0.1093	-0.1059	-0.1019	-0.0999	-0.0969	-0.0909	-0.0884	-0.0831	-0.0801	-0.0781	-0.074	-0.0687	-0.0689
married	-0.0379	-0.0376	-0.0377	-0.0395	-0.0373	-0.0371	-0.0372	-0.0375	-0.0371	-0.0371	-0.0364	-0.0347	-0.0344
μ_1	1.6898	1.6504	1.6101	1.5666	1.5252	1.4753	1.435	1.374	1.3322	1.3093	1.26	1.1938	1.1936
μ_2	0.2285	0.1916	0.1525	0.1098	0.0692	0.0171	-0.0235	-0.0872	-0.1302	-0.1552	-0.2068	-0.2763	-0.2755
σ	0.614	0.5653	0.513	0.4553	0.3977	0.3226	0.2609	0.1584	0.0837	0.0393	-0.0622	-0.218	-0.2176
ρ	2.3773	2.4723	2.5716	2.6757	2.7739	2.8986	2.9952	3.1438	3.2418	3.2962	3.4144	3.5708	3.5711
T-ratios													
female	-6.22	-6.09	-6.05	-5.66	-5.86	-5.32	-5.53	-5.03	-4.98	-5.13	-4.54	-4.18	-3.71
age	8.53	8.24	8.25	7.39	7.84	6.69	7.16	6.18	6.08	6.35	5.37	4.78	4.13
age^2	-8.17	-7.92	-7.94	-7.15	-7.58	-6.53	-6.97	-6.06	-5.97	-6.22	-5.29	-4.73	-4.09
children	-4.23	-4.18	-4.17	-4.04	-4.08	-3.9	-3.98	-3.78	-3.76	-3.82	-3.58	-3.38	-3.14
degree	1.72	1.71	1.71	1.69	1.7	1.68	1.69	1.66	1.66	1.67	1.64	1.62	1.57
distw	4.29	4.23	4.21	4.07	4.13	3.92	4	3.79	3.74	3.8	3.53	3.33	3.06
$distw^2$	-3.22	-3.21	-3.21	-3.18	-3.22	-3.14	-3.2	-3.11	-3.09	-3.13	-2.99	-2.87	-2.7
car_license	6.32	6.18	6.17	5.69	5.98	5.34	5.6	5.06	5.03	5.19	4.57	4.21	3.69
big_city	-1.19	-1.19	-1.19	-1.19	-1.2	-1.21	-1.22	-1.21	-1.21	-1.22	-1.21	-1.2	-1.19
own_home	-1.17	-1.17	-1.17	-1.19	-1.2	-1.18	-1.19	-1.19	-1.19	-1.18	-1.17	-1.16	-1.15
married	-0.38	-0.39	-0.4	-0.44	-0.43	-0.45	-0.47	-0.51	-0.52	-0.54	-0.55	-0.56	-0.56
μ_1	14.61	13.73	13.39	11.57	12.02	9.82	10.26	8.45	8.08	8.28	6.73	5.69	4.9
μ_2	1.63	1.32	1.05	0.69	0.45	0.1	-0.14	-0.46	-0.68	-0.85	-0.97	-1.17	-1.02
σ	4.14	3.58	3.15	2.41	2.16	1.41	1.17	0.57	0.28	0.13	-0.16	-0.41	-0.35
ρ	8.97	9	9.46	8.74	9.87	8.68	9.85	8.86	9.08	9.77	8.43	7.9	6.72

Table 9: Detailed outputs - 26 random pairs

ULL	-4364.190	-4401.491	-4395.057	-4481.351	-4416.579	-4412.082	-4466.592	-4425.905	-4438.316	-4409.920
Trace	0.899	0.883	2.950	1.566	1.294	1.141	0.153	0.531	0.860	0.557
					Coefficients					
female	-1.007	-1.089	-0.922	-0.813	-0.983	-1.051	-0.840	-0.875	-1.071	-1.308
age	0.218	0.232	0.198	0.173	0.207	0.227	0.178	0.193	0.230	0.276
age^2	-0.251	-0.269	-0.228	-0.200	-0.239	-0.263	-0.204	-0.225	-0.264	-0.318
children	-0.156	-0.160	-0.147	-0.108	-0.147	-0.153	-0.128	-0.132	-0.169	-0.206
degree	0.233	0.244	0.162	0.178	0.191	0.213	0.185	0.191	0.201	0.233
distw	0.018	0.021	0.020	0.014	0.019	0.021	0.017	0.019	0.022	0.026
$distw^2$	-0.013	-0.015	-0.014	-0.010	-0.013	-0.016	-0.013	-0.015	-0.015	-0.019
car_license	1.740	1.904	1.650	1.402	1.671	1.767	1.468	1.538	1.839	2.265
big_city	-0.096	-0.106	-0.109	-0.096	-0.134	-0.116	-0.071	-0.095	-0.101	-0.129
own_home	-0.093	-0.091	-0.067	-0.057	-0.079	-0.087	-0.038	-0.067	-0.077	-0.110
married	-0.044	-0.032	0.021	-0.053	-0.037	-0.073	-0.054	-0.031	-0.043	-0.049
μ_1	1.248	1.322	1.175	1.017	1.191	1.275	1.059	1.130	1.309	1.487
μ_2	-0.221	-0.149	-0.302	-0.456	-0.260	-0.195	-0.411	-0.370	-0.159	0.039
σ	-0.111	0.059	-0.271	-0.982	-0.204	-0.041	-0.725	-0.472	0.030	0.347
ρ	3.466	3.277	3.587	4.057	3.529	3.381	3.944	3.781	3.355	2.890
					T-ratios					
female	-3.39	-4.04	-2.41	-5.37	-3.54	-3.58	-7.74	-5.44	-4.1	-5.24
age	3.64	4.5	2.51	6.32	3.84	3.95	13.72	6.66	4.58	6.54
age^2	-3.63	-4.43	-2.49	-6.19	-3.79	-3.93	-12.91	-6.56	-4.52	-6.46
children	-2.86	-3.31	-2.28	-3.35	-3.13	-3.04	-4.53	-3.8	-3.4	-3.74
degree	1.67	1.76	1.19	1.86	1.48	1.56	1.89	1.78	1.47	1.45
distw	2.75	3.25	2.26	3.8	2.86	2.97	4.55	4.03	3.24	3.95
$distw^2$	-2.44	-2.77	-2.12	-2.96	-2.6	-2.76	-3.72	-3.43	-2.79	-3.1
car_license	3.35	4.1	2.42	6.44	3.47	3.58	9.06	5.67	4.09	5.38
big_city	-1.15	-1.16	-1.26	-1.43	-1.58	-1.27	-1.03	-1.32	-1.14	-1.24
own_home	-1.37	-1.3	-1.07	-1.09	-1.22	-1.27	-0.72	-1.18	-1.14	-1.38
married	-0.62	-0.44	0.3	-1	-0.55	-0.98	-0.99	-0.52	-0.59	-0.58
μ_1	4.49	6	2.94	7.03	4.54	5.05	16.02	7.6	6	9.64
μ_2	-0.75	-0.62	-0.72	-3.09	-0.92	-0.72	-4.67	-2.21	-0.66	0.22
σ	-0.18	0.14	-0.25	-0.86	-0.31	-0.08	-2.83	-0.93	0.07	1.49
ρ	5.44	6.31	4.06	11.74	5.87	5.73	31.75	12.11	6.59	7.42

Table 10: Detailed outputs - 120 random pairs

ULL	-4387.262	-4387.608	-4423.530	-4401.674	-4405.906	-4401.775	-4411.283	-4397.174	-4410.141	-4414.419
Trace	0.586	0.344	0.253	0.599	0.607	0.423	0.607	0.356	0.519	0.282
					Coefficients					
female	-1.079	-1.127	-0.899	-1.076	-0.917	-1.308	-1.060	-1.016	-0.962	-0.988
age	0.230	0.240	0.192	0.228	0.194	0.276	0.224	0.216	0.205	0.210
age^2	-0.266	-0.278	-0.222	-0.264	-0.224	-0.318	-0.259	-0.250	-0.237	-0.243
children	-0.160	-0.167	-0.128	-0.155	-0.131	-0.206	-0.150	-0.145	-0.142	-0.146
degree	0.205	0.238	0.186	0.234	0.193	0.233	0.225	0.212	0.204	0.208
distw	0.021	0.023	0.018	0.021	0.018	0.026	0.021	0.021	0.020	0.019
$distw^2$	-0.014	-0.017	-0.013	-0.015	-0.013	-0.019	-0.015	-0.015	-0.014	-0.013
car_license	1.883	1.970	1.568	1.891	1.608	2.265	1.860	1.784	1.694	1.721
big_city	-0.106	-0.119	-0.089	-0.109	-0.091	-0.129	-0.114	-0.117	-0.093	-0.094
own_home	-0.074	-0.076	-0.066	-0.078	-0.064	-0.110	-0.077	-0.084	-0.070	-0.065
married	-0.033	-0.042	-0.026	-0.041	-0.038	-0.049	-0.039	-0.040	-0.042	-0.022
μ_1	1.310	1.356	1.134	1.306	1.144	1.487	1.283	1.252	1.202	1.221
μ_2	-0.160	-0.112	-0.334	-0.160	-0.323	0.039	-0.175	-0.218	-0.276	-0.245
σ	0.032	0.129	-0.387	0.036	-0.344	0.347	-0.003	-0.080	-0.197	-0.151
ρ	3.277	3.166	3.721	3.305	3.667	2.890	3.356	3.412	3.566	3.519
					T-ratios					
female	-4.72	-5.97	-6.61	-4.80	-5.64	-5.24	-4.63	-5.69	-5.07	-6.25
age	5.69	7.95	9.52	5.54	7.05	6.54	5.53	7.46	6.09	8.45
age^2	-5.58	-7.74	-9.22	-5.45	-6.89	-6.46	-5.43	-7.30	-5.98	-8.22
children	-3.78	-4.06	-4.04	-3.52	-3.79	-3.74	-3.61	-3.88	-3.78	-4.23
degree	1.53	1.73	1.75	1.74	1.71	1.45	1.70	1.70	1.70	1.78
distw	3.66	4.15	4.45	3.55	3.96	3.95	3.56	4.08	3.78	4.11
$distw^2$	-2.98	-3.41	-3.45	-3.00	-3.21	-3.1	-3.04	-3.41	-3.14	-3.09
car_license	4.81	6.05	7.08	4.73	5.76	5.38	4.80	6.05	5.16	6.55
big_city	-1.20	-1.28	-1.22	-1.23	-1.22	-1.24	-1.32	-1.41	-1.17	-1.18
own_home	-1.10	-1.07	-1.18	-1.15	-1.12	-1.38	-1.19	-1.32	-1.18	-1.08
married	-0.47	-0.57	-0.45	-0.59	-0.63	-0.58	-0.57	-0.61	-0.68	-0.34
μ_1	7.49	10.69	11.11	7.15	8.11	9.64	7.20	9.49	7.32	10.58
μ_2	-0.80	-0.75	-2.61	-0.79	-1.99	0.22	-0.86	-1.38	-1.46	-1.74
σ	0.09	0.59	-1.28	0.11	-0.85	1.49	-0.01	-0.28	-0.49	-0.58
ρ	8.39	11.64	18.23	8.22	12.81	7.42	8.52	12.05	10.25	14.91

Table 11: Detailed outputs - 215 random pairs

ULL	-4412.567	-4410.702	-4405.611	-4403.064	-4425.659	-4407.589	-4399.523	-4406.376	-4421.490	-4405.170
Trace	0.780	1.074	0.519	0.774	0.548	0.513	0.769	0.546	0.713	0.592
					Coefficients					
female	-0.955	-0.957	-0.976	-0.982	-0.951	-0.983	-0.976	-0.984	-0.920	-0.981
age	0.204	0.203	0.208	0.209	0.203	0.208	0.208	0.210	0.195	0.209
age^2	-0.236	-0.234	-0.241	-0.242	-0.235	-0.240	-0.241	-0.243	-0.226	-0.241
children	-0.141	-0.137	-0.144	-0.144	-0.139	-0.142	-0.143	-0.145	-0.133	-0.143
degree	0.201	0.203	0.205	0.200	0.192	0.200	0.199	0.214	0.191	0.203
distw	0.019	0.019	0.019	0.019	0.018	0.019	0.020	0.019	0.018	0.019
$distw^2$	-0.014	-0.013	-0.014	-0.013	-0.013	-0.013	-0.014	-0.014	-0.012	-0.013
car_license	1.658	1.670	1.712	1.711	1.662	1.719	1.705	1.724	1.604	1.713
big_city	-0.097	-0.100	-0.096	-0.100	-0.101	-0.098	-0.094	-0.100	-0.098	-0.098
own_home	-0.074	-0.072	-0.072	-0.076	-0.075	-0.064	-0.064	-0.078	-0.064	-0.068
married	-0.030	-0.036	-0.032	-0.031	-0.035	-0.031	-0.037	-0.038	-0.038	-0.035
μ_1	1.188	1.185	1.215	1.215	1.185	1.215	1.214	1.223	1.145	1.215
μ_2	-0.281	-0.284	-0.254	-0.256	-0.284	-0.254	-0.254	-0.246	-0.319	-0.255
σ	-0.232	-0.237	-0.165	-0.168	-0.244	-0.162	-0.167	-0.145	-0.339	-0.164
ρ	3.584	3.586	3.516	3.514	3.612	3.525	3.509	3.498	3.682	3.521
					T-ratios					
female	-4.34	-3.75	-5.02	-4.32	-4.99	-4.99	-4.29	-4.92	-4.60	-4.73
age	4.96	4.20	6.12	4.95	6.08	6.20	4.92	6.00	5.40	5.63
age^2	-4.90	-4.17	-5.98	-4.89	-5.99	-6.11	-4.85	-5.88	-5.34	-5.53
children	-3.40	-3.15	-3.80	-3.49	-3.71	-3.73	-3.39	-3.73	-3.53	-3.58
degree	1.64	1.61	1.67	1.59	1.63	1.63	1.60	1.72	1.64	1.65
distw	3.44	3.10	3.76	3.38	3.71	3.72	3.37	3.64	3.49	3.66
$distw^2$	-2.99	-2.66	-3.15	-2.85	-3.10	-3.10	-2.92	-3.04	-2.99	-3.10
car_license	4.33	3.79	5.14	4.29	5.06	5.17	4.30	4.98	4.66	4.81
big_city	-1.22	-1.24	-1.18	-1.23	-1.29	-1.22	-1.15	-1.23	-1.28	-1.21
own_home	-1.24	-1.21	-1.19	-1.24	-1.28	-1.07	-1.05	-1.27	-1.13	-1.11
married	-0.49	-0.57	-0.51	-0.48	-0.57	-0.49	-0.57	-0.59	-0.64	-0.55
μ_1	5.86	4.98	7.46	5.97	7.17	7.57	5.96	7.34	6.16	6.86
μ_2	-1.24	-1.08	-1.34	-1.12	-1.48	-1.34	-1.11	-1.26	-1.52	-1.26
σ	-0.45	-0.38	-0.42	-0.35	-0.57	-0.42	-0.35	-0.37	-0.62	-0.39
ρ	8.25	6.92	10.00	7.93	10.32	10.12	7.91	9.60	9.50	9.39

Table 12: Detailed outputs - 120 smart pairs 1

ULL	-4363.697	-4352.991	-4361.896	-4372.898	-4374.696	-4358.071	-4365.463	-4367.166	-4362.807	-4365.965
Trace	0.449	0.498	0.557	0.435	0.425	0.383	0.422	0.452	0.450	0.582
					Coefficients					
female	-1.071	-1.053	-1.071	-1.061	-0.990	-1.087	-1.003	-1.047	-1.065	-1.099
age	0.238	0.234	0.240	0.234	0.222	0.241	0.222	0.232	0.236	0.243
age^2	-0.274	-0.271	-0.279	-0.270	-0.256	-0.280	-0.255	-0.267	-0.273	-0.280
children	-0.151	-0.147	-0.150	-0.153	-0.141	-0.149	-0.140	-0.149	-0.151	-0.155
degree	0.225	0.232	0.214	0.211	0.213	0.234	0.207	0.227	0.228	0.239
distw	0.021	0.021	0.021	0.020	0.020	0.021	0.020	0.021	0.022	0.022
$distw^2$	-0.014	-0.015	-0.014	-0.013	-0.014	-0.014	-0.014	-0.015	-0.015	-0.015
car_license	1.838	1.817	1.813	1.801	1.685	1.852	1.715	1.794	1.817	1.884
big_city	-0.100	-0.099	-0.109	-0.100	-0.094	-0.098	-0.104	-0.097	-0.098	-0.107
own_home	-0.081	-0.072	-0.073	-0.071	-0.065	-0.066	-0.069	-0.069	-0.074	-0.080
married	-0.019	-0.019	-0.016	-0.035	-0.014	-0.010	-0.024	-0.016	-0.018	-0.018
μ_1	1.362	1.344	1.366	1.336	1.287	1.375	1.290	1.333	1.353	1.380
μ_2	-0.185	-0.195	-0.184	-0.202	-0.264	-0.172	-0.257	-0.201	-0.188	-0.159
σ	0.022	-0.002	0.025	-0.013	-0.145	0.044	-0.119	-0.017	0.011	0.067
ρ	3.154	3.192	3.140	3.212	3.318	3.137	3.313	3.213	3.170	3.100
					T-ratios					
female	-4.57	-4.23	-4.93	-4.66	-5.06	-4.60	-4.97	-4.59	-5.03	-4.68
age	5.41	4.83	5.99	5.44	6.07	5.37	5.99	5.35	6.06	5.48
age^2	-5.27	-4.77	-5.86	-5.33	-5.95	-5.27	-5.86	-5.23	-5.91	-5.38
children	-3.45	-3.14	-3.50	-3.52	-3.55	-3.35	-3.47	-3.42	-3.57	-3.37
degree	1.66	1.70	1.62	1.59	1.72	1.72	1.66	1.71	1.74	1.72
distw	3.63	3.50	3.88	3.60	3.85	3.65	3.79	3.59	3.88	3.67
$distw^2$	-3.01	-3.04	-3.07	-2.99	-3.19	-3.04	-3.11	-3.04	-3.23	-3.03
car_license	4.68	4.32	5.02	4.66	5.17	4.69	5.17	4.62	5.14	4.77
big_city	-1.14	-1.13	-1.24	-1.16	-1.16	-1.10	-1.26	-1.12	-1.13	-1.18
own_home	-1.22	-1.09	-1.09	-1.08	-1.05	-0.99	-1.12	-1.06	-1.14	-1.17
married	-0.28	-0.28	-0.23	-0.52	-0.22	-0.14	-0.38	-0.24	-0.27	-0.25
μ_1	7.44	6.52	8.23	7.29	7.86	7.43	7.83	7.14	8.29	7.64
μ_2	-0.88	-0.85	-0.95	-0.95	-1.40	-0.82	-1.34	-0.94	-0.99	-0.77
σ	0.06	0.00	0.08	-0.03	-0.39	0.13	-0.33	-0.05	0.03	0.20
ρ	7.90	7.06	8.73	8.11	9.73	7.70	9.37	8.01	8.96	7.76

Table 13: Detailed outputs - 120 smart pairs 2

ULL	-4307.253	-4310.092	-4303.398	-4312.946	-4301.914	-4306.100	-4306.074	-4312.558	-4310.585	-4310.323
Trace	0.613	0.747	0.515	0.613	0.495	0.620	0.507	0.624	0.498	0.601
					Coefficients					
female	-1.333	-1.272	-1.252	-1.431	-1.417	-1.386	-1.413	-1.374	-1.333	-1.209
age	0.289	0.274	0.272	0.309	0.306	0.300	0.304	0.293	0.286	0.261
age^2	-0.335	-0.317	-0.315	-0.358	-0.354	-0.348	-0.352	-0.341	-0.330	-0.302
children	-0.187	-0.179	-0.178	-0.205	-0.203	-0.196	-0.195	-0.188	-0.187	-0.172
degree	0.280	0.271	0.264	0.309	0.291	0.315	0.294	0.280	0.290	0.257
distw	0.025	0.024	0.023	0.027	0.027	0.026	0.025	0.025	0.025	0.023
$distw^2$	-0.017	-0.016	-0.015	-0.019	-0.019	-0.018	-0.017	-0.018	-0.018	-0.015
car_license	2.236	2.135	2.108	2.416	2.385	2.366	2.371	2.340	2.236	2.067
big_city	-0.110	-0.119	-0.116	-0.133	-0.124	-0.129	-0.130	-0.104	-0.116	-0.122
own_home	-0.087	-0.082	-0.084	-0.081	-0.089	-0.089	-0.085	-0.095	-0.079	-0.075
married	-0.042	-0.010	-0.014	-0.044	-0.031	-0.043	-0.036	-0.037	-0.051	-0.040
μ_1	1.536	1.482	1.473	1.600	1.591	1.579	1.588	1.554	1.522	1.439
μ_2	0.024	-0.013	-0.033	0.104	0.094	0.072	0.076	0.057	0.031	-0.064
σ	0.369	0.306	0.286	0.480	0.465	0.441	0.442	0.417	0.369	0.231
ρ	2.817	2.882	2.942	2.635	2.665	2.685	2.695	2.748	2.797	3.025
					T-ratios					
female	-5.59	-5.28	-5.06	-5.84	-5.85	-6.11	-5.99	-5.69	-5.51	-4.85
age	7.17	6.46	6.03	7.86	7.82	8.34	7.77	7.36	7.14	5.72
age^2	-6.93	-6.29	-5.88	-7.56	-7.49	-7.95	-7.47	-7.07	-6.88	-5.58
children	-3.85	-3.69	-3.67	-4.06	-4.04	-3.98	-3.92	-3.85	-3.91	-3.58
degree	1.76	1.74	1.71	1.80	1.72	1.90	1.76	1.71	1.81	1.72
distw	4.12	3.90	3.74	4.21	4.33	4.28	4.09	4.15	4.13	3.75
$distw^2$	-3.14	-2.97	-3.02	-3.33	-3.44	-3.31	-3.10	-3.35	-3.32	-3.02
car_license	5.68	5.32	5.03	6.06	6.04	6.29	6.03	5.80	5.69	4.91
big_city	-1.04	-1.17	-1.15	-1.18	-1.11	-1.17	-1.18	-0.97	-1.11	-1.25
own_home	-1.12	-1.09	-1.13	-0.96	-1.07	-1.07	-1.03	-1.17	-1.01	-1.04
married	-0.52	-0.13	-0.18	-0.50	-0.36	-0.52	-0.42	-0.45	-0.62	-0.54
μ_1	11.16	9.63	8.95	12.82	12.72	13.44	12.47	11.64	11.07	8.32
μ_2	0.14	-0.07	-0.17	0.69	0.62	0.50	0.50	0.36	0.19	-0.32
σ	1.83	1.31	1.12	2.78	2.69	2.69	2.51	2.19	1.83	0.83
ρ	9.18	8.50	7.94	9.09	9.39	10.19	9.48	9.01	9.08	7.90

Table 14: Detailed outputs - 120 smart pairs 3

ULL	-4290.452	-4290.308	-4283.146	-4287.199	-4295.456	-4292.902	-4294.826	-4294.826	-4288.981	-4288.401
Trace	0.594	0.636	0.051	1.247	0.716	0.938	0.721	0.860	0.401	0.677
					Coefficients					
female	-0.885	-1.023	-0.981	-0.906	-0.981	-0.941	-0.892	-0.938	-0.912	-0.963
age	0.199	0.231	0.220	0.204	0.220	0.214	0.200	0.210	0.204	0.215
age^2	-0.229	-0.267	-0.253	-0.235	-0.253	-0.247	-0.230	-0.243	-0.236	-0.248
children	-0.128	-0.146	-0.137	-0.124	-0.139	-0.136	-0.126	-0.132	-0.127	-0.135
degree	0.202	0.229	0.220	0.194	0.206	0.202	0.186	0.207	0.206	0.214
distw	0.018	0.021	0.020	0.018	0.020	0.019	0.018	0.019	0.018	0.019
$distw^2$	-0.012	-0.014	-0.014	-0.013	-0.014	-0.013	-0.012	-0.013	-0.013	-0.013
car_license	1.534	1.787	1.697	1.574	1.701	1.632	1.544	1.607	1.571	1.660
big_city	-0.083	-0.093	-0.103	-0.092	-0.094	-0.091	-0.079	-0.096	-0.093	-0.090
own_home	-0.066	-0.068	-0.056	-0.055	-0.073	-0.073	-0.069	-0.063	-0.059	-0.071
married	-0.015	-0.014	-0.009	-0.014	-0.022	-0.017	-0.010	-0.019	-0.008	-0.008
μ_1	1.192	1.343	1.293	1.214	1.290	1.264	1.199	1.239	1.215	1.268
μ_2	-0.365	-0.225	-0.279	-0.348	-0.265	-0.309	-0.365	-0.315	-0.343	-0.290
σ	-0.422	-0.047	-0.161	-0.353	-0.146	-0.243	-0.415	-0.272	-0.345	-0.201
ρ	3.500	3.193	3.290	3.466	3.287	3.380	3.519	3.386	3.445	3.335
					T-ratios					
female	-6.04	-4.55	-5.02	-3.67	-4.35	-3.96	-4.73	-4.09	-5.59	-4.49
age	4.37	5.30	6.08	4.02	5.04	4.48	5.50	4.76	7.16	5.26
age^2	-5.25	-5.20	-5.91	-3.97	-4.92	-4.42	-5.36	-4.69	-6.96	-5.16
children	-3.72	-3.30	-3.47	-2.89	-3.24	-3.11	-3.33	-3.17	-3.62	-3.27
degree	2.09	1.72	1.78	1.59	1.62	1.64	1.64	1.71	1.83	1.73
distw	5.18	3.59	3.86	3.14	3.59	3.30	3.71	3.34	4.11	3.54
$distw^2$	-2.67	-3.04	-3.23	-2.75	-3.05	-2.86	-3.02	-2.77	-3.21	-2.96
car_license	4.57	4.65	5.30	3.69	4.43	4.02	4.81	4.19	5.95	4.60
big_city	-1.73	-1.10	-1.27	-1.17	-1.14	-1.14	-1.06	-1.21	-1.24	-1.11
own_home	-1.71	-1.07	-0.91	-0.95	-1.17	-1.21	-1.21	-1.07	-1.04	-1.16
married	-0.35	-0.20	-0.14	-0.23	-0.35	-0.27	-0.16	-0.31	-0.14	-0.12
μ_1	8.22	7.16	8.05	4.90	6.55	5.68	6.67	5.93	8.91	6.72
μ_2	-2.03	-1.05	-1.48	-1.27	-1.17	-1.23	-1.75	-1.32	-2.10	-1.34
σ	-0.50	-0.12	-0.43	-0.48	-0.32	-0.42	-0.72	-0.48	-0.89	-0.43
ρ	9.25	7.85	9.67	6.50	7.69	7.00	9.35	7.48	12.21	8.15