

A Rank-Ordered Logit Model with Unobserved Heterogeneity in Ranking Capabilities

Bram van Dijk*
Econometric Institute
Tinbergen Institute
Erasmus University Rotterdam

Dennis Fok
Econometric Institute
Erasmus University Rotterdam

Richard Paap
Econometric Institute
Erasmus University Rotterdam

ECONOMETRIC INSTITUTE REPORT 2007-07

Abstract

In this paper we consider the situation where one wants to study the preferences of individuals over a discrete choice set through a survey. In the classical setup respondents are asked to select their most preferred option out of a (selected) set of alternatives. It is well known that, in theory, more information can be obtained if respondents are asked to rank the set of alternatives instead. In statistical terms, the preferences can then be estimated more efficiently. However, when individuals are unable to perform (part of) this ranking task, using the complete ranking may lead to a substantial bias in parameter estimates. In practice, one usually opts to only use a part of the reported ranking.

In this paper we introduce a latent-class rank-ordered logit model in which we use latent segments to endogenously identify the ranking capabilities of individuals. Each segment corresponds to a different assumption on the ranking capability. Using simulations and an empirical application, we show that using this model for parameter estimation results in a clear efficiency gain over a multinomial logit model in case some individuals are able to rank. At the same time it does not suffer from biases due to ranking inabilities of some of the respondents.

JEL Classification: C25, C52

Keywords: Discrete Choice Models, Random Utility, Stated Preferences

*We thank Philip Hans Franses for useful comments. The calculations in this paper are done in Ox version 4.00 (see Doornik, 2002). Address for correspondence: Bram van Dijk, Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: avandijk@few.eur.nl, phone: +31-10-4088943, fax: +31-10-4089162.

1 Introduction

To determine preferences of individuals, researchers often rely on surveys. In the traditional setup a survey is created in which each respondent is only asked to select the most preferred option out of a set of presented alternatives. To estimate the preferences based on such a survey one can use a standard discrete choice model, like the multinomial logit [MNL] model. It is well known that more information can be obtained from a respondent if (s)he is asked to give a complete ranking of *all* presented alternatives. The rank-ordered logit [ROL] model is now the standard tool to analyze the preferences in case rank data is available.

The ROL model was introduced in the literature by Beggs *et al.* (1981). The model can be used to analyze the preferences of individuals over a set of alternatives, where the preferences are partially observed through surveys or conjoint studies. Empirical applications describing preferences using the ROL model can be found in many fields such as voter preferences (Koop and Poirier, 1994), aging studies (Hsieh, 2005), marketing (Ahn *et al.*, 2006; Dagsvik and Liu, 2006), school choice (Mark *et al.*, 2004), demand for classical music (Van Ophem *et al.*, 1999) and transportation studies (Kockelman *et al.*, 2006; Calfee *et al.*, 2001).

In theory, when individuals are asked to rank the alternatives instead of only choosing the most preferred option, the parameters of the choice model and hence the preferences can be estimated more efficiently. However, in practice respondents may be unable to perform (part of) the ranking task. This may be due to several reasons. First of all, respondents may not be able to perform the task itself. In some cases there may be too many alternatives to rank. Secondly, the respondent may not be able to distinguish between his less-preferred alternatives. In any case, straightforwardly using reported rankings may lead to a substantial bias in the parameter estimates in the ROL model, see Chapman and Staelin (1982). To solve this issue, Chapman and Staelin (1982) suggest to only use the first few ranks in the estimation. They consider several rules to determine the appropriate number of ranks to use, in their words “the explosion depth”. One of these rules is based on a pooling test for the equality of parameter estimates based on

different rank information. Hausman and Ruud (1987) proposed an alternative method to test for the number of ranks to use in the estimation. However, in both approaches this number is assumed to be the same for all respondents. If ranking capabilities differ across individuals, this may lead to an efficiency loss.

In this paper we propose a model in which we use all observed rankings while taking into account the fact that the rankings may not completely reflect the true preferences. To this end we introduce a ROL model which uses latent segments to endogenously identify the unobserved ranking capabilities of respondents. Each segment corresponds to a particular assumption on the unobserved ranking capability. Hence, in contrast to the existing literature, we allow for individual-specific ranking capabilities. We show that in case at least some individuals are able to rank, our model results in a clear efficiency gain relative to a standard MNL model. At the same time it does not suffer from biases due to ranking inabilities of some of the individuals. We refer to our model as the latent-class rank-ordered logit [LCROL] model.

Apart from the efficiency gain, the new model also allows us to learn about the unobserved ranking capabilities of respondents. On itself, this information can also be very valuable. For example, one can use this information to construct more efficient ranking tasks in surveys and for deciding the number of respondents necessary to reach a preferred precision. To investigate the presence of respondents with a particular type of ranking capability, we propose to use a likelihood ratio test. We can, for example, test whether some individuals are able to rank all alternatives. Furthermore, the statistical test can be helpful to remove redundant segments from our model, which may lead to efficiency loss if present.

This study is not the first attempt to maximize preference information from individual ranked ordered data. Van Ophem *et al.* (1999) designed a multichoice logit [MCL] model to use more information, and hence to be more efficient than the MNL model, while avoiding the bias that a ROL model can have due to ranking inability of some respondents. In their setup, respondents need to sequentially divide the items into three preference groups where an item in a higher group is always preferred to an item in a lower group. Finally, the items in the top-ranked group were completely ranked according

to preferences. The final ranking can be seen as an extreme case of dividing the items across groups, where each group just contains one item. The MCL model describes the preference ordering of these groups. By not considering the preferences within all groups, they hope to avoid biases. However, it is still assumed that ranking capabilities are the same across individuals.

Our solution to deal with differences in ranking capabilities also bears some similarities with the misreporting literature. If an individual is not able to rank properly, his answer may not be according to his preferences. In other words in this case there may be misreporting, or a classification error. In a non-linear model, measurement errors in the dependent variables in general result in biased estimates, see Hausman (2001) for a small overview. Many studies have considered specifications to account for misclassification, see, for example, Chua and Fuller (1987), Poterba and Summers (1995), Hausman *et al.* (1998), and Abrevaya and Hausman (1999). In our model we account for a more general form of misclassifications, some ranks are correct whereas other ranks are “misclassified”.

The remainder of this paper is set up as follows. In Section 2 we introduce the LCROL model which incorporates the unobserved ranking capabilities of individuals. To illustrate the merits of our model specification we perform a simulation study in Section 3. For the case that not all respondents are able to provide a full ranking, simulations show an estimation bias in the standard ROL model. For the general case that some individuals are not able to give a complete ranking and some individuals are able to give a partial ranking, estimates resulting from our LCROL model show no bias and are more efficient than MNL estimates. Section 4 provides an application of the LCROL model. We apply the model to data obtained from a small survey among 91 students, who are asked to rank six different platforms for computer games. The estimation results show that the LCROL is a useful tool for analyzing these data. The parameter estimates from the LCROL model have the smallest standard errors, even in this small sample situation. Furthermore, the resulting segmentation of the students is plausible. Finally, in Section 5 we conclude and give topics for further research.

2 Model specification

In this section we develop our approach to capture unobserved heterogeneity in the ranking ability of respondents. In Section 2.1 we start the discussion with the standard rank-ordered logit model and in the following sections we will extend this model. Throughout this paper, we consider the simple case that each respondent is confronted with a fixed set of alternatives. The respondent either chooses the most preferred item from this set or is asked to provide a complete preference ordering. Of course more complicated surveys are also possible, for example, in a conjoint setting, see Van Ophem *et al.* (1999). However, in order to focus the discussion we keep the setup of the survey simple.

2.1 Rank-ordered logit model

Suppose that we want to learn the (determinants of) preferences of individuals over a discrete set of items. Classic examples of this problem include preferences over different modes of transportation or different brands in a product category. Preferences can be recovered from historical data but if some items are not available yet, a survey is usually the only option.

Denote the number of alternatives by J . We use the random utility framework, see Manski (1977), to represent the preferences of individuals. The random utilities for individual i are a set of latent variables U_{i1}, \dots, U_{iJ} , defined as

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \tag{1}$$

where $i = 1, \dots, N$ indexes individuals and $j = 1, \dots, J$ indexes the items. The utilities consist of two parts: V_{ij} is the deterministic component of the utility, determined by observed individual characteristics and ε_{ij} is the random component of the utility of alternative j for individual i . In general the deterministic part of the utility is modeled as

$$V_{ij} = x_i' \beta_j, \tag{2}$$

where x_i is an m -dimensional vector with characteristics of individual i and β_j is an m -dimensional parameter vector specific to alternative j .

In the traditional setup, respondents are asked to choose their most preferred option out of the complete set of J alternatives. Let $y_{ij} = 1$ denote that respondent i most prefers alternative j . The information $y_{ij} = 1$ implies that for this respondent the utility of alternative j is larger than all other alternatives, that is,

$$U_{ij} \geq \max\{U_{i1}, \dots, U_{iJ}\}. \quad (3)$$

The probability of this event depends on the distribution of ε_{ij} . If we assume that ε_{ij} has an independent type-I extreme value distribution, we have the setup of a multinomial logit [MNL] model, see McFadden (1973, 1974). This leads to the well-known expression for the probability that item j is most preferred by individual i

$$\begin{aligned} \Pr[y_{ij} = 1; \beta] &= \Pr[U_{ij} \geq \max\{U_{i1}, \dots, U_{iJ}\}] \\ &= \frac{\exp(V_{ij})}{\sum_{l=1}^J \exp(V_{il})}, \end{aligned} \quad (4)$$

where $\beta = \{\beta_1, \dots, \beta_J\}$ and β_J is put equal to zero for identification.

The information on the most preferred item is enough to be able to estimate the model parameters. However, as discussed before, more information per respondent can be obtained if we ask for a ranking of alternatives and this will in general result in an efficiency gain. We will denote the response of respondent i by the vector $y_i = (y_{i1}, \dots, y_{iJ})'$, where y_{ij} now denotes the rank that individual i gives to item j . For example, if $y_{ij} = 2$ this means that the respondent considers alternative j the second most preferred option. For notational convenience we will also use the equivalent notation $r_i = (r_{i1}, \dots, r_{iJ})'$, where r_{ij} denotes the item number that received rank j by individuals i . The relation between r_i and y_i is given by

$$y_{ik} = j \iff r_{ij} = k \quad (5)$$

for $j, k = 1, \dots, J$.

An observed ranking for a respondent implies a complete ordering of the underlying utilities. An individual will prefer an item with a higher utility over an item with a lower utility. If we observe a full ranking r_i , we know that

$$U_{ir_{i1}} > U_{ir_{i2}} > \dots > U_{ir_{iJ}}. \quad (6)$$

It is obvious that (6) provides more information compared to (3). Under the utility assumption in (1) and the assumption of the extreme value distribution, we obtain the rank-ordered logit [ROL] model, see Beggs *et al.* (1981) and Chapman and Staelin (1982). The probability of observing a particular ranking r_i equals

$$\begin{aligned} \Pr[r_i; \beta] &= \Pr[U_{ir_{i1}} > U_{ir_{i2}} > \dots > U_{ir_{iJ}}] \\ &= \prod_{j=1}^{J-1} \frac{\exp(V_{ir_{ij}})}{\sum_{l=j}^J \exp(V_{ir_{il}})}. \end{aligned} \quad (7)$$

The ROL model can be seen as a series of MNL models: an MNL for the most preferred item; another MNL for the second-ranked item to be preferred over all items except the one with rank 1, and so on. Finally, the probability of a complete ranking is made up of the product of these separate MNL probabilities. The product contains only $J - 1$ probabilities, because ranking the least preferred item is done with probability 1. Note that this result holds due to the IIA property of the MNL model, see Beggs *et al.* (1981) for a derivation.

2.2 Ranking ability

In the standard ROL model we implicitly assume that respondents are able to rank each item according to the random utility model. However, it has already been noted by Chapman and Staelin (1982) that for the less preferred items, this assumption does not always hold. One of the possible reasons for this is that the respondent perhaps has no experience with some of the items, and hence is not able to indicate the proper ranking order. It is also possible that respondents tend to find the least preferred items less important and rank those randomly. In practice this means that the observed rank order of the least preferred items may not be according to the model.

If the least preferred items are not ranked according to the underlying utility model, the use of those ranks in the estimation will lead to a bias in the parameter estimates towards zero, see Chapman and Staelin (1982, p. 292). Furthermore, Hausman and Ruud (1987, p. 89) notice in an application on mobile phones that including more ranks in the estimation procedure leads to a decline in the absolute magnitude of the parameters.

The common solution to this problem is rather simple. If the ranks beyond k are biased, then do not use these lower rankings in the estimation procedure¹. This can be done very easily by letting the product in (7) only go up to k instead of $J - 1$. The probability that individual i only ranks the k most preferred items according to utility values and ranks the remaining items ($J - k$) randomly is given by

$$\Pr[y_i|k; \beta] = \prod_{j=1}^k \frac{\exp(V_{ir_{ij}})}{\sum_{l=j}^J \exp(V_{ir_{il}})} \frac{1}{(J - k)!}, \quad (8)$$

where we condition on the ranking capabilities k . The second term in (8), $1/(J - k)!$, is usually not present in many studies. The reason for this is that one usually considers the probability of observing only the first k rankings, that is, $\Pr[r_{i1}, r_{i2}, \dots, r_{ik}|k; \beta]$ instead of considering the full ranking y_i . For reasons that will become clear below, we also have to take into account the probability of observing the least preferred items. We assume that the least preferred $J - k$ items are ordered randomly, therefore all $(J - k)!$ possible orderings are equally likely. Hence, the last term in (8) contains the probability of observing one particular ordering of the last $J - k$ items.

If one assumes k to be the same for the whole sample, this last term can be ignored in the estimation as it becomes a constant in the log likelihood, see Chapman and Staelin (1982) and Hausman and Ruud (1987). These two papers propose estimating different ROL models, each using a different number of ranks, that is a different value for k . Then they have different methods of choosing a model from this set. One can, for example, use a Hausman (1978) test to test for differences in the β parameter for different values of k . The tradeoff is that using more ranks gives more efficient parameter estimates, but it can also introduce a bias in the results.

In the present paper we relax the assumption that k is equal for all individuals. In the next subsection we will allow k to vary over the individuals using a latent-class approach.

¹Hausman and Ruud (1987) also estimate a model where each rank in the estimation receives a weight. This makes it possible that the most preferred rank contain more information than lower ranks. These weights are estimated alongside the model parameters.

2.3 Latent-class rank-ordered logit model

In the applications of the ROL model in the literature one either assumes that k is known a-priori, where $k = J$ is also possible, or one determines k by comparing the results for several values of k . The underlying assumption here is that there is no heterogeneity in the population concerning the capability to order the alternatives. In this section we will introduce such heterogeneity. Hence, determining the value of k is also a part of the model.

The introduction of heterogeneity of ranking abilities solves many practical issues. Firstly, one usually does not know beforehand how many ranks should be used for estimating a ROL model to balance the efficiency against a possible bias. Secondly, by allowing for the heterogeneity we make efficient use of the available data. For example, assume that 10% of the respondents can only give the most preferred item, and give a random ordering for the other items, and the rest (90%) is able to provide a complete ordering. In the standard model we would then be forced to only consider the first rank in the estimation to avoid a bias in the estimated parameters. The additional information available in the responses of the 90% will not be used to make the estimates more efficient. Using the additional information is only possible, when heterogeneity in the ranking ability is incorporated in the model.

To allow for such heterogeneity, we divide our individuals into J latent classes, see, for example, Wedel and Kamakura (2000). For class $k = 0, 1, \dots, J - 1$ we impose that the individual can rank k most preferred items, more formally that they are based on (6). The individual ranks the remaining $(J - k)$ items randomly. The probability of observing a particular ranking for individual i , now becomes

$$\Pr[y_i; \beta, p] = \sum_{k=0}^{J-1} p_k \Pr[y_i|k; \beta], \quad (9)$$

where p_k is the probability that individual i belongs to segment k with $0 \leq p_k \leq 1$ and $\sum_{k=0}^{J-1} p_k = 1$, and where the probability of observing ranking y_i when only the k most preferred items are ranked according to the random utility model $\Pr[y_i|k; \beta]$ is given in (8). Equation (9) explains the reason for including the second term in (8). In order to be able to compare the segments, they must all contain the probability of observing the full

ranking, not just the probability of observing the k most-preferred ranks.

The resulting model is a finite mixture model, see, for example, Titterton *et al.* (1985) and Everitt and Hand (1981). To estimate the model parameters β and the mixing proportions $p = (p_0, \dots, p_{J-1})$ we rely on Maximum Likelihood. The likelihood function is given by,

$$\begin{aligned} \mathcal{L}(\beta, p) &= \prod_{i=1}^N \Pr[y_i; \beta, p] \\ &= \prod_{i=1}^N \sum_{k=0}^{J-1} p_k \Pr[y_i|k; \beta] \\ &= \prod_{i=1}^N \sum_{k=0}^{J-1} \frac{p_k}{(J-k)!} \left[\prod_{l=1}^k \frac{\exp(x'_i \beta_{r_{il}})}{\sum_{m=l}^J \exp(x'_i \beta_{r_{im}})} \right]. \end{aligned} \quad (10)$$

The log likelihood is given by

$$\log \mathcal{L}(\beta, p) = \sum_{i=1}^N \log \left\{ \sum_{k=0}^{J-1} p_k \exp \left[-\log((J-k)!) + \sum_{l=1}^k \left(x'_i \beta_{r_{il}} - \log \left(\sum_{m=l}^J \exp(x'_i \beta_{r_{im}}) \right) \right) \right] \right\}, \quad (11)$$

where we have rewritten the product of probabilities as the exponent of a sum for numerical stability. The likelihood function can be maximized using numerical optimization algorithms. In our simulations and in the empirical section, we use a constrained optimization procedure to make sure that the restrictions on the p parameters hold. Standard errors for the parameters can straightforwardly be obtained using the second-order derivative of the log likelihood.

Testing for empty classes

The total number of potential latent classes is equal to the number of items. If some of these classes are not present in the data, this may lead to an efficiency loss in the estimation of the β parameters. Therefore it is useful to test whether some of the classes can be removed from the model. It is especially interesting to test whether $p_0 = 0$, that is whether each individual can at least provide his/her most preferred item. If this is not

the case, even the standard MNL model will provide parameter estimates that are biased towards zero, see Section 4 for an illustration.

To test for the redundance of a segment, we propose to use a standard likelihood ratio [LR] test. Because the β parameters are the same in each latent class, we do not suffer from the Davies (1977) problem, which is usually the case when testing for the number of latent classes. To test for the restriction of the absence of class j we consider the hypothesis $p_j = 0$. As the alternative is $p_j > 0$, we have a test for a parameter on the boundary of the parameter space. The asymptotic distribution of the LR test statistic is then a mixture of χ^2 distributions, for the case of testing one parameter the appropriate distribution is $\frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1)$, see Wolak (1989a,b). Hence, if we want to test at a 5% level of significance we have to use the 90% percentile of a $\chi^2(1)$ distribution, which is equal to 2.705. In case we want to perform a joint test for the absence of two or more segments the asymptotic distribution will be a weighted average of χ^2 distribution where the weights follow from the covariance matrix of the estimated p parameters, see Wolak (1989a,b) for details.

2.4 Extension

So far we have assumed that respondents are only able to rank the first few items correctly. However, it may be possible that in some particular applications respondents are also able to indicate which alternatives they least prefer. For example, it could be that they have tried something and were dissatisfied with it. This allows us to take advantage of the information in the lowest ranks. The efficiency in the parameter estimates will then increase even more relative to a standard MNL model.

We now denote the latent segments by two indices (k, l) , where k denotes the number of most preferred items and l the number of least preferred ranks, that can be ranked correctly. The case $l = 0$ corresponds to the previous discussion. We focus the discussion below on the situation $l = 1$, where the lowest ranked alternative is also consistent with the utility model. We exclude the combinations $(J - 2, 1)$ and $(J - 1, 1)$. These cases actually correspond to an individual who is able to rank all alternatives. These individuals are classified in segment $(J - 1, 0)$. Hence, we have to add $J - 2$ additional latent classes

to the model, with $l = 1$ and k ranging from 0 to $J - 3$. The rank probabilities in the new segments $(k, 1)$ are given by

$$\Pr[y_i|(k, 1); \beta] = \Pr[y_i|k; \beta] \Pr[U_{ir_{iJ}} \leq U_{ir_{im}} \forall m > k; \beta](J - k) \quad (12)$$

for $k = 0, \dots, J - 3$. The probability that alternative r_{iJ} is preferred least in the set of items $(k + 1, k + 2, \dots, J)$, $\Pr[U_{ir_{iJ}} \leq U_{ir_{im}} \forall m > k; \beta]$, is derived in Appendix A. Due to the IIA property of the logit specification this probability is independent of the k top ranked alternatives. The factor $(J - k)$ is added in (12) because the number of possible random combinations of the $(J - k)$ least preferred items is reduced by the factor $(J - 1)$ as we now know the least preferred item.

A similar exercise can be performed for $l > 1$. The number of classes will increase quickly, and one has to consider the tradeoff between the number of new latent classes and the potential efficiency gain.

3 Monte Carlo simulation

To illustrate the merits of our latent-class rank-ordered logit [LCROL] model, we perform a small simulation study. In this simulation we assume the presence of four alternatives ($J = 4$). To facilitate matters, we assume that respondents are able to rank the most preferred items in a correct way and have problems ranking the remaining items. However, we do allow for all four potential latent segments, where in each segment the individuals are able to rank 0, 1, 2 or all most-preferred items correctly.

In the data generating process we put the probability of the first segment (ranking 0 items correctly) to 0 for two reasons. Firstly, this restriction implies that the parameters of a multinomial logit [MNL] model can be estimated consistently, which allows us to make a fair comparison with our latent-segment model. The second reason is that it allows us to analyze the size of our proposed LR test for the redundancy of this segment. The size of the other segments are $p_1 = p_2 = 0.30$ and $p_4 = 0.40$.

The latent utilities are generated according to $U_{ij} = \beta_{0j} + x_{i1}\beta_{1j} + x_{i2}\beta_{2j} + \varepsilon_{ij}$, for $j = 1, \dots, 4$ and $i = 1, \dots, N$. The variable x_{i1} is generated from a standard normal distribution, x_{i2} is a 0/1 dummy variable with probability 0.5 of being 1. The fourth

category is considered to be a base category and hence $\beta_{k4} = 0$ for $k = 0, 1, 2$. The disturbances ε_{ij} are independently drawn from a type-I extreme value distribution. The values of the β parameters are displayed in the second column of Table 1.

In each replication of the simulation experiment we simulate utilities for $N = 1,000$ individuals who provide the ranking of the four alternatives. The 1,000 individuals are divided in one of the three segments according to the mixing proportions. The ranking of each individual is adjusted according to the imposed abilities in each segment, that is, a random ranking is imposed if individuals are assumed not being able to rank. For example, if an individual belongs to segment 2, we keep the preference order of the first two most preferred items but replace the preference order of the final two items by a random order.

We estimate four different models, that is, a standard MNL model, a rank-ordered logit [ROL] model, a LCROL model with 4 segments (including segment 0) and a LCROL model with 3 segments (without segment 0). Table 1 provides the means and the root mean squared errors [RMSE] of the parameter estimates over 10,000 replications.

The third column of Table 1 displays the results for the MNL model. The mean of the parameters is almost the same as the DGP parameters and hence one can estimate the β parameters in the utilities without a bias as expected. The fourth column shows the results for the ROL model. Note that this model is misspecified as about 60% of the individuals are not able to rank all four alternatives properly. This results in a clear bias towards zero in almost all of the β parameters ranging from 4% to about 56%, which is clearly substantial. This confirms the findings of Chapman and Staelin (1982) and Hausman and Ruud (1987).

The fifth and sixth column of Table 1 display the results for our LCROL model. If we consider the case where we do not restrict p_0 to be zero, we see that the difference between the mean of the estimated parameters and the true parameters is at most 0.04. The RMSE of the estimator is however smaller than for the estimator of the same parameters in the MNL model. If we impose that $p_0 = 0$ the results even improve. The mean of the estimated parameters and the true parameters differs at most 0.01. The RMSE of the estimator is even smaller.

Table 1: Mean and RMSE of the distribution of the parameters for $N=1000$ based on 10,000 Monte Carlo replications

Parameter	True	MNL ^a	ROL ^b	LCROL ^c	LCROL ^d
β_{01}	1.00	1.00 (0.14)	0.80 (0.22)	1.03 (0.12)	1.00 (0.11)
β_{11}	0.75	0.76 (0.12)	0.54 (0.22)	0.78 (0.11)	0.76 (0.09)
β_{21}	-0.30	-0.30 (0.21)	-0.32 (0.12)	-0.30 (0.17)	-0.30 (0.16)
β_{02}	0.25	0.25 (0.16)	0.24 (0.08)	0.26 (0.12)	0.25 (0.12)
β_{12}	-0.50	-0.50 (0.12)	-0.42 (0.11)	-0.51 (0.10)	-0.50 (0.09)
β_{22}	0.45	0.45 (0.21)	0.31 (0.18)	0.47 (0.17)	0.45 (0.17)
β_{03}	-0.25	-0.25 (0.18)	-0.11 (0.17)	-0.26 (0.14)	-0.25 (0.13)
β_{13}	1.00	1.01 (0.14)	0.62 (0.39)	1.04 (0.13)	1.00 (0.11)
β_{23}	0.80	0.80 (0.23)	0.50 (0.32)	0.83 (0.19)	0.80 (0.18)
p_0	0.00	-	-	0.02 (0.04)	-
p_1	0.30	-	-	0.29 (0.06)	0.30 (0.06)
p_2	0.30	-	-	0.29 (0.10)	0.30 (0.10)
p_4	0.40	-	-	0.40 (0.09)	0.40 (0.08)

^a Standard multinomial logit model

^b Rank-ordered logit model

^c Latent-class rank-ordered logit model.

^d Latent-class rank-ordered logit model with $p_0 = 0$

Table 2: Theoretical and empirical size of the LR test for $p_0 = 0$ (sample size is 1000)

Theoretical size	Empirical size	Critical value
1%	1.1%	5.41
5%	5.2%	2.71
10%	9.5%	1.64
20%	19.2%	0.71

The simulation results clearly show the advantage of our approach. With our LCROL model we use as much information as possible, while still obtaining unbiased results. An efficiency gain can be obtained if we remove redundant segments from our LCROL model. To illustrate the applicability of the LR test for redundant segments, we compute in each replication of the simulation exercise the LR test for $p_0 = 0$ in the LCROL model. In Table 2 we report the empirical size, based on the asymptotic critical values (quantiles of the $\frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1)$ distribution). It can be seen that the empirical size of the LR test is close to the theoretical size. Also the mean (0.49) and the standard deviation (1.15) of the test statistic are very close to the theoretical mean and standard deviation of the asymptotic distribution of the LR test.

In sum, we can conclude that the LCROL model provides consistent estimates in situations where not all individuals can fulfil a rank task properly. Furthermore, the RMSE of the estimator of the LCROL is smaller than for the MNL model where we only use the most preferred rank. An LR test for the absence of a segment is correctly sized, and hence it can be used to determine which latent segments are present in the data.

4 Application

To illustrate the practical usefulness of our latent-class rank-ordered logit model, we consider the results of a survey among 91 Dutch students. The students were asked to consider buying a new platform to play computer games (assuming they did not have one). They had to rank 6 different platforms suitable to play computer games. The 6 platforms are the X-box (360), the PlayStation (2 or 3), the Gamecube (or Wii), the PlayStation Portable, the Gameboy (color/advance/DS/...) or just a regular PC. Note

that the survey did not distinguish between different generations of the same platform. At the time of the survey, the Nintendo Wii and the Playstation 3 were not available. In addition we know which of the 6 platforms the student owns and the average number of hours that each student spends on gaming each week.

Model specification

First we estimate a multinomial logit [MNL] model for the most preferred platform. As explanatory variables we include platform intercepts, time (hours gaming) and a 0/1 dummy to indicate whether the student owns the platform, where 1 corresponds to ownership. The base alternative is a personal computer. The second column of Table 3 displays the parameter estimates. We notice that platform ownership has a positive effect on preference and that individuals who spend more time gaming seem to prefer a personal computer over a real game computer. However, this effect is not significant for any platform.

The third column of Table 3 displays the parameter estimates for a standard rank-ordered logit [ROL] model. Hence, we implicitly assume that each student is capable of performing the complete ordering task. The parameter values differ substantially from the MNL estimates which suggests that this assumption is not valid. The Hausman (1978) test statistic for equal parameters equals 34.3, which is significant at the 5% level of significance.

The estimation results suggests that it seems necessary to include latent ordering abilities in the model. We include 6 classes indicating that the individuals cannot rank at all (segment 0), rank only the most preferred item (segment 1), the first 2, 3, 4 most preferred items (segment 2, 3, and 4) and all items (segment 5) with corresponding mixing proportions p_j for $j = 0, \dots, 5$. The fourth column in Table 3 displays the parameter estimates of the latent-class rank-ordered logit [LCROL] model. The parameters seem to be different from the MNL estimates. This difference can be explained by the fact that about 23% (p_0) of the students are not able to rank the platforms at all. Indeed, the likelihood ratio test for $p_0 = 0$ equals 14.42 and hence this segment cannot be neglected. As additional check we compute the LR statistic for the restriction $p_1 = 1$ which leads to

Table 3: Parameters estimates results for the MNL and ROL models with standard errors in parenthesis.

Variable	MNL	ROL	LCROL ^a	LCROL ^b	LCROL ^c
			<i>intercept</i>		
XBox	0.92 (0.49)	1.41 (0.29)	1.53 (0.51)	1.47 (0.42)	1.51 (0.53)
Playstation	0.58 (0.45)	0.93 (0.27)	1.11 (0.47)	1.05 (0.40)	1.04 (0.47)
Playstation portable	-0.03 (0.59)	0.80 (0.28)	0.44 (0.52)	0.79 (0.51)	0.53 (0.56)
GameCube	0.49 (0.59)	-0.00 (0.30)	-3.50 (1.61)	-0.65 (0.55)	-2.18 (1.16)
GameBoy	-1.47 (0.99)	0.08 (0.29)	-2.71 (1.41)	-0.51 (0.70)	-1.59 (0.99)
			<i>hours spent on gaming</i>		
XBox	-0.10 (0.06)	-0.17 (0.05)	-0.14 (0.06)	-0.14 (0.06)	-0.14 (0.06)
Playstation	-0.11 (0.07)	-0.13 (0.04)	-0.11 (0.06)	-0.11 (0.05)	-0.11 (0.06)
Playstation portable	-0.10 (0.11)	-0.23 (0.05)	-0.36 (0.12)	-0.35 (0.12)	-0.40 (0.13)
GameCube	-0.39 (0.24)	-0.18 (0.05)	-0.01 (0.15)	-0.22 (0.11)	-0.15 (0.16)
GameBoy	-0.05 (0.18)	-0.23 (0.05)	-0.23 (0.15)	-0.33 (0.11)	-0.33 (0.15)
Platform ownership	1.78 (0.38)	0.97 (0.19)	1.72 (0.37)	1.45 (0.29)	1.71 (0.35)
p_0	-	-	0.23 (0.07)	-	0.21 (0.07)
p_1	-	-	0.20 (0.09)	0.34 (0.11)	0.27 (0.08)
p_2	-	-	0.07 (0.08)	0.01 (0.09)	-
p_3	-	-	0.07 (0.09)	0.08 (0.12)	-
p_4	-	-	0.00 (0.16)	0.00 (0.62)	-
p_5	-	-	0.43	0.57	0.52
LR statistic ^d	110.20	32.73	-	14.42	1.85

^a Latent-class rank ordered logit model

^b Latent-class rank ordered logit model with $p_0 = 0$

^c Latent-class rank ordered logit model with $p_2 = p_3 = p_4 = 0$

^d LR statistic to test against the LCROL model in fourth column.

the MNL model. The value of the statistic is 110.20 and hence this restriction is clearly rejected. The ROL model is obtained when we impose the restriction $p_0 = p_1 = p_2 = p_3 = p_4 = 0$. Again, this restriction is rejected.

The estimates of p_2 , p_3 and p_4 are relatively small. The LR statistic for the restriction $p_2 = p_3 = p_4 = 0$ equals 1.85 and hence we cannot reject the redundancy of these 3 segments. After imposing the restriction we obtain the model presented in the final column of Table 3. The median individual prefers the PC for playing games.² The PC has a probability of 0.45 of being the most preferred platform. The Xbox and the Playstation come second and third with 0.28 and 0.19, respectively.

Students who spend more time playing games, have more preference for the personal computer. Owning a platform has a positive effect on the preference for that platform. This can be due to two effects. The first is because of reverse causality, a student owns the platform, because he/she likes it. But since we consider multiple generations of the same platform this can also be interpreted as a backwards compatibility effect. Someone who owns a lot of PS2 games would rather have a PS3 than a Nintendo Wii, because the PS3 can, for example, still play the PS2 games.

Interpretation of the segments

In the final model, the mixing proportions divide the students in three segments. About 52% of the students know enough about the different platforms to give a complete ranking. The remaining 48% of the students can either not provide a clear ranking at all (21%), or only know which of the platforms they prefer most (27%).

We expect that the first segment of students consists of heavy gamers, who spend a lot of time on gaming. The other 2 segments contain students who game less. To segment the individual students we can compute the conditional segment membership probabilities defined by

$$\pi_{ij} = \frac{p_j \Pr[y_i|j; \beta]}{\sum_{k \in \mathcal{K}} p_k \Pr[y_i|k; \beta]}, \quad (13)$$

where π_{ij} is the conditional probability that individual i belongs in latent class j given the observed ranking, p_j are the estimated mixing proportions of the segments and $\Pr[y_i|j; \beta]$

²The median individual spends 2 hours per week on gaming, and only owns a PC.

Table 4: Platform ownership and average time spend per week on gaming in hours for each latent segment

	Segment 0	Segment 1	Segment 5
Hours spent on gaming	2.54	3.47	4.64
Ownership XBox	11%	17%	12%
Ownership PlayStation	30%	27%	35%
Ownership PSP	7%	13%	10%
Ownership Gamecube	11%	14%	5%
Ownership Gameboy	21%	12%	11%
Ownership PC	83%	85%	91%

is given in (8). The set \mathcal{K} contains the segments included in the model. In our final model we had the segments 0, 1 and 5.

The segment membership probabilities (13) allow us to assign the students to the three groups. The average value of the largest conditional segment membership probability over the students is 0.80, which indicates that the model is capable to make a clear distinction between the segments of students. In Table 4 we display the average value of the explanatory variables in the 3 segments. If we consider the number of hours spent on gaming, the results are as we expect. Students who spend more time on gaming are better able to rank the platforms than students who game less. The differences are less clear with ownership, except for the PC where we find that respondents who are not able to rank are less likely to own a PC, although differences are small.

5 Conclusions

The respondents inability to accurately provide a full ranking of all presented alternatives in a survey leads to a bias in the application of the well-known rank-ordered logit model. To remove this bias, while still taking maximum advantage of the information in the ranked data, we propose in this paper to augment the rank-ordered logit model with latent segments. Each latent segment is associated with a particular ranking ability. More specifically, given J alternatives we define J segments. In the k -th segment we assume that the respondent is only able to rank the k most preferred items correctly. We also allow k to be zero, in this case the respondent does not even report the most

preferred item correctly. Under this situation even the multinomial logit model would provide biased estimates.

Using simulation and an empirical application, we show that our new model is indeed robust against inabilities of individuals to give proper ranks. Moreover, it is more efficient than a standard multinomial logit approach. All this taken together, our model makes it very attractive to ask respondents in a survey to rank all options instead of asking them to select their preferred option. The application also shows that it is not unlikely that one may encounter respondents who cannot rank the alternatives at all. The direct application of the multinomial logit model is therefore not always appropriate.

Our analysis is based on the logit framework. Therefore we have to assume that the IIA property holds. If one wants to relax this assumption, our proposed strategy can be incorporated in rank-ordered probit models (Hajivassiliou and Ruud, 1994) or mixed rank-ordered logit model of Calfee *et al.* (2001), which do not suffer from the IIA assumption.

A Derivation of probability of least preferred item

In this appendix we derive the probability that a particular item is preferred least in a set of J items, see Van Ophem *et al.* (1999) for an alternative proof. The probability of interest is given by

$$\Pr[U_{ir_{iJ}} \leq U_{ir_{im}} \quad \forall m > k]. \quad (14)$$

The IIA property of the logit models implies that this probability does not depend on the utilities of the k most preferred items. In fact, this probability is the same as the probability that item r_{iJ} has rank $J - k$ in the set of items $(r_{i,k+1}, \dots, r_{i,J})$. To simplify the discussion, we derive without loss of generality the probability that item 1 is ranked last in a full set of alternatives, which can easily be generalized to the probability in (14).

$$\begin{aligned} \Pr[y_1 = J] &= \Pr[U_1 \leq U_2, U_1 \leq U_3, \dots, U_1 \leq U_J] \\ &= \Pr[V_1 + \varepsilon_1 \leq V_2 + \varepsilon_2, \dots, V_1 + \varepsilon_1 \leq V_J + \varepsilon_J] \\ &= \Pr[\varepsilon_2 > V_1 - V_2 + \varepsilon_1, \dots, \varepsilon_J > V_1 - V_J + \varepsilon_1] \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) \int_{V_1 - V_2 + \varepsilon_1}^{\infty} f(\varepsilon_2) \cdots \int_{V_1 - V_J + \varepsilon_1}^{\infty} f(\varepsilon_J) d\varepsilon_J \dots d\varepsilon_2 d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) \int_{V_1 - V_2 + \varepsilon_1}^{\infty} f(\varepsilon_2) \cdots \int_{V_1 - V_{J-1} + \varepsilon_1}^{\infty} f(\varepsilon_{J-1}) \times \\ &\quad [\exp(-e^{-\varepsilon_J})]_{V_1 - V_J + \varepsilon_1}^{\infty} d\varepsilon_{J-1} \dots d\varepsilon_2 d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) \int_{V_1 - V_2 + \varepsilon_1}^{\infty} f(\varepsilon_2) \cdots \int_{V_1 - V_{J-1} + \varepsilon_1}^{\infty} f(\varepsilon_{J-1}) \times \\ &\quad [1 - \exp(-e^{V_J - V_1 - \varepsilon_1})] d\varepsilon_{J-1} \dots d\varepsilon_2 d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) [1 - \exp(-e^{V_2 - V_1 - \varepsilon_1})] \cdots [1 - \exp(-e^{V_J - V_1 - \varepsilon_1})] d\varepsilon_1, \end{aligned} \quad (15)$$

where we suppress the subscript i for notational convenience. Expanding the terms in brackets leads to

$$\begin{aligned}
\Pr[y_1 = J] &= \int_{-\infty}^{\infty} f(\varepsilon_1) d\varepsilon_1 - \sum_{i=2}^J \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{V_i - V_1 - \varepsilon_1}) d\varepsilon_1 \\
&\quad + \sum_{i=2}^{J-1} \sum_{j=i+1}^J \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{V_i - V_1 - \varepsilon_1}) \exp(-e^{V_j - V_1 - \varepsilon_1}) d\varepsilon_1 \\
&\quad + \dots + (-1)^{J-1} \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{V_2 - V_1 - \varepsilon_1}) \dots \exp(-e^{V_J - V_1 - \varepsilon_1}) d\varepsilon_1 \\
&= 1 - \sum_{i=2}^J \frac{\exp(V_1)}{\exp(V_1) + \exp(V_i)} + \sum_{i=2}^{J-1} \sum_{j=i+1}^J \frac{\exp(V_1)}{\exp(V_1) + \exp(V_i) + \exp(V_j)} \\
&\quad + \dots + (-1)^{J-1} \frac{\exp(V_1)}{\exp(V_1) + \exp(V_2) + \dots + \exp(V_J)},
\end{aligned} \tag{16}$$

where we use the standard logit characteristic. To simplify notation, we define the set T_i to contain the sums of all possible combinations of i elements from the set $\{\exp(V_2), \dots, \exp(V_J)\}$, that is, T_i contains $\binom{J-1}{i}$ elements. We will denote one specific element from the set T_i as T_{ij} , $j = 1, \dots, \binom{J-1}{i}$. The probability (16) can now be written as

$$\Pr[y_1 = J] = 1 + \sum_{i=1}^{J-1} (-1)^i \sum_{j=1}^{\binom{J-1}{i}} \frac{\exp(V_1)}{\exp(V_1) + T_{ij}}. \tag{17}$$

References

- Abrevaya, J. and J. Hausman (1999), Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells, *Annales D'Economie et de Statistique*, **55-56**, 243–275.
- Ahn, J., J. Lee, J. Lee, and T. Kim (2006), An Analysis of Consumer Preferences among Wireless LAN and Mobile Internet Services, *ETRI Journal*, **28**, 205–215.
- Beggs, S., S. Cardell, and J. Hausman (1981), Assessing the Potential Demand for Electric Cars, *Journal of Econometrics*, **16**, 1–19.
- Calfee, J., C. Winston, and R. Stempiski (2001), Econometric Issues in Estimating Consumer Preferences From Stated Preference Data: A Case Study of the Value of Automobile Travel Time, *The Review of Economics and Statistics*, **83**, 699–707.
- Chapman, R. and R. Staelin (1982), Exploiting Rank Ordered Choice Set Data Within the Stochastic Utility Model, *Journal of Marketing Research*, **19**, 288–301.
- Chua, T. and W. Fuller (1987), A Model for Multinomial Response Error Applied to Labor Flows, *Journal of the American Statistical Association*, **82**, 46–51.
- Dagsvik, J. and G. Liu (2006), A Framework for Analyzing Rank Ordered Panel Data with Application to Automobile Demand, Discussion Papers 480, Research Department of Statistics Norway.
- Davies, R. (1977), Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternative, *Biometrika*, **64**, 247–254.
- Doornik, J. (2002), *Object-Oriented Matrix Programming using Ox*, 3rd edn., Timberlake Consultants Press, London.
- Everitt, B. and D. Hand (1981), *Finite Mixture Distributions*, Chapman and Hall, London.
- Hajivassiliou, V. and P. Ruud (1994), Classical Estimation Methods for LDV Models using Simulation, in R. Enle and D. McFadden (eds.), *Handbook of Econometrics*, vol. 4, Elsevier, Amsterdam.

- Hausman, J. (1978), Specification Tests in Econometrics, *Econometrica*, **46**, 1251–1271.
- Hausman, J. (2001), Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left, *Journal of Economic Perspectives*, **15**, 57–67.
- Hausman, J. and P. Ruud (1987), Specifying and Testing Econometric Models for Rank-Ordered Data, *Journal of Econometrics*, **34**, 83–104.
- Hausman, J., J. Abrevaya, and F. Scott-Morton (1998), Misclassification of the Dependent Variable in a Discrete-Response Setting, *Journal of Econometrics*, **87**, 239–269.
- Hsieh, C. (2005), Age and Relative Importance of Major Life Domains, *Journal of Aging Studies*, **19**, 503–512.
- Kockelman, K., K. Podgorski, M. Bina, and S. Gadda (2006), Public Perceptions of Pricing Existing Roads and Other Transportation Policies: The Texas Perspective, working paper, URL: <http://www.ce.utexas.edu/prof/kockelman/>.
- Koop, G. and D. Poirier (1994), Rank-Ordered Logit Models: an Empirical Analysis of Ontario Voter Preferences, *Journal of Applied Econometrics*, **9**, 369–388.
- Manski, C. (1977), The Structure of Random Utility Models, *Theory and Decision*, **8**, 229–254.
- Mark, D., J. Lusk, and M. Scott Daniel (2004), Recruiting Agricultural Economics Graduate Students: Student Demand for Program Attributes, *American Journal of Agricultural Economics*, **86**, 175–184.
- McFadden, D. (1973), Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York.
- McFadden, D. (1974), The Measurement of Urban Travel Demand, *Journal of Public Economics*, **3**, 303–328.
- Poterba, J. and L. Summers (1995), Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model With Errors in Classification, *The Review of Economics and Statistics*, **77**, 207–216.

- Titterton, D., A. Smith, and U. Makov (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.
- Van Ophem, H., P. Stam, and B. van Praag (1999), Multichoice Logit: Modeling Incomplete Preference Rankings of Classical Concerts, *Journal of Business & Economic Statistics*, **17**, 117–128.
- Wedel, M. and W. Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations*, second edn., Springer, Boston.
- Wolak, F. (1989a), Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models, *Econometric Theory*, **5**, 1–35.
- Wolak, F. (1989b), Testing Inequality Constraints in Linear Econometric Models, *Journal of Econometrics*, **41**, 205–235.