VRIJE UNIVERSITEIT AMSTERDAM

Discrete choice for spatial economics

Hans R.A. Koster*

October 8, 2019

PRELIMINARY; PLEASE SEND COMMENTS TO H.KOSTER@VU.NL

1 INTRODUCTION

Many choices individuals make are *discrete.* Consider, for example, the decision how to get to the university this morning. People can take the bicycle, the bus or maybe the underground. Also the route that you have taken can be considered as a discrete decision: there are many routes possible but you can only take one.

Standard micro-economics is mostly concerned with the analysis of continuous choices. For example, it may aim to answer the question what is the impact of an increase in gasoline prices on the the amount of kilometres travelled. Or, it may investigate the impact of environmental taxes on the productivity of firms. The convenient characteristic of continuous choices is that marginal changes in behaviour can be analysed in a straightforward manner. For example, one can estimate that when fuel prices change by 10%, the demand for cars may decrease by 2%. Welfare effects can be analysed assuming that individuals maximise a pre-defined utility function with respect to the chosen quantity of the good.

This syllabus aims to introduce a framework for *discrete* consumer choices. Hence, this means that some explanatory variable, either continuous or discrete, influences a *discrete*

*Department of Spatial Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105 1081HV Amsterdam, email: h.koster@vu.nl. Hans is also research fellow at the Higher School of Economics in St. Petersburg, Tinbergen Institute (Gustav Mahlerplein 117 1082MS Amsterdam) and affiliated with the Centre for Economic Policy Research. I thank Thomas de Graaff, Jasper Knockaert, Paul Koster, and Jos van Ommeren for useful comments on previous versions of this syllabus.

dependent variable. For example,

- What are the most important characteristics that determine where people want to live?
- What factors influence the selection of a seaport for a carrier?
- What is the impact of the weather on route choice?
- How do taxes influence the location choices of multinational enterprises?
- What is the role of environmental regulation on bankruptcy?
- What is the impact of eye colour on partner choice?

and many more examples can be given.

Analysing marginal changes in behaviour seems a hard thing to do for discrete choices, because either one chooses a certain alternative, or one does not. Moreover, discrete choices can be both *binary* – where a person chooses between two alternatives – or *multinomial* – where a person chooses between many alternatives (*e.g.* mode of transport).[1] This issue is resolved by assuming that the utility function is random. The so-called random utility framework in Section 2 is able to deal with discrete binary and multinomial choices. Hence, it is possible to analyse marginal changes in behaviour by modelling the *probability* that someone chooses an alternative.

Section 3 proceeds by discussing several ways to estimate *binary choices*: namely, the Linear Probability Model (LPM), Logit and Probit models, and generalisations of those, such as Nested Logit.

Section 4 extends the approach to a *multinomial setting* and explains when Multinomial Logit, Conditional Logit and Poisson Models can be applied.

All models rely on data. Data can be either gathered in a experimental setting, where respondents are confronted with hypothetical choices, or one may rely on data on realised 'revealed' choices. Section 5 then deals with the distinction between *stated preference* and *revealed preference* data and discusses the advantages and disadvantages.

Section 6 provides a summary.

2 THE RANDOM UTILITY FRAMEWORK

2.1 MAKING CHOICES

Ben-Akiva and Lerman (1985) argue that a choice can be seen as an outcome of a sequential decision-making process that includes the following steps:

[1] Instead of binary or multinomial decisions, one sometimes refers to dichotomous and polytomous decisions.

1. Definition of the choice problem;
2. Generation of alternatives;
3. Evaluation of attributes of the alternatives;
4. Choice;
5. Implementation.

Hence, for each *choice* it is important to know the possible *alternatives* and their *attributes*. The agent can be an individual, a firm, a household or a government body.

Let's consider a simple setting where people have to make a discrete decision (the choice) between two routes, A and B (the alternatives). We observe in a hypothetical dataset that only the travel time t (the attribute) between routes A and B is different. Hence, intuitively, the decision to choose a route depends on the difference in time. However, suppose that people have certain (unknown) preferences for certain routes (*e.g.* some like to take a more scenic route); those characteristics are denoted by ε_A and ε_B.

Hence, to summarise the above setting in a model, one could write:

$$\Pr(Y = A) = f\big(\beta(t_B - t_A) + (\varepsilon_A - \varepsilon_B)\big), \tag{2.1}$$

where $\Pr(Y = A)$ denotes the probability that someone will choose route A and $f(\cdot)$ is an increasing function in its attribute(s), $f'(\cdot) > 0$. This probability is impacted by the difference in the travel time between route A and B. Note that in this formulation β is positive, because the higher the travel time of route B, the higher the probability to choose A.

An important feature of the above model is that the probability cannot be larger than one or smaller than zero. Hence, the probability to choose route B is given by: $\Pr(Y = B) = 1 - \Pr(Y = A)$. This also means that standard linear regression techniques (*i.e.* Ordinary Least Squares) may give an unrealistic answer. For example, for certain values of the travel time, it may predict that the probability is higher than 1 or lower than 0.

2.2 MAXIMISING UTILITY

How does equation (2.1) relate to people maximising their utility? Let's consider again two routes and an individual i that has the following utility functions:

$$U_{iA} = V_A(t_A) + \varepsilon_{iA}, \tag{2.2}$$

$$U_{iB} = V_B(t_B) + \varepsilon_{iB}, \tag{2.3}$$

where U_{iA} and U_{iB} are the utilities received from a certain alternative, $V_A(t_A)$ and $V_B(t_B)$ are referred to as the *deterministic utilities*, which a person receives from observed characteristics (in this case; travel time) of the alternatives.[2] In reality one often observes that, while most people will take the shortest route A, still some people will take the longer route B. The reason is that people may have so-called idiosyncratic or unobserved preferences for a certain route, denoted by ε_{iA} and ε_{iB}.[3] For example, some people may particularly like the unobserved scenic features of route B, or prefer the more quiet route B. Hence, while the agent is assumed to know his or her utility function, the researcher does not. The idiosyncratic term ε aims to capture this (Train, 2003).[4] Like in a standard regression framework $\mathrm{E}[\varepsilon_{iA}] = \mathrm{E}[\varepsilon_{iB}] = 0$ and the ε's are assumed to be independently and identically distributed. Moreover, *they are assumed to be uncorrelated to the deterministic utilities.*

To determine the probability to choose route A, (2.2) and (2.3) are combined into (Train, 2003):

$$\begin{aligned} \Pr(Y = A) &= \Pr(U_{iA} > U_{iB}), \\ &= \Pr(V_A(t_A) - V_B(t_B) > \varepsilon_{iB} - \varepsilon_{iA}), \\ &= \int_{\varepsilon_{iA}=-\infty}^{\infty} \int_{\varepsilon_{iB}=-\infty}^{\infty} I(V_A(t_A) - V_B(t_B) > \varepsilon_{iB} - \varepsilon_{iA}) f(\varepsilon_{iA}) f(\varepsilon_{iB}) \mathrm{d}\varepsilon_{iA} \mathrm{d}\varepsilon_{iB}. \end{aligned} \tag{2.4}$$

The second line in the above equation indicates the probability to choose A, which depends directly on deterministic utility and on idiosyncratic preferences. Hence, if one observes that people make a certain choice in the data, this relates to attributes of the alternatives and the idiosyncratic tastes. In the third line of (2.4) the indicator function $I(\cdot)$ equals one when the condition holds true. That is, $I(\cdot) = 1$ if the values of ε and the value of the travel time difference, induces the agent to choose route A; and $I(\cdot) = 0$ if the values of ε, combined with the travel time difference, induces the agent to choose B. $f(\cdot)$ is the (joint) density function of the idiosyncratic preferences. Hence, the second line in equation (2.4) states that the probability is an integral, *i.e.* an integral of an indicator for the outcome of the choice process over all possible values of the unobserved factors (Train, 2003).

Hence, to be able to calculate the probability in (2.4), one has to deal with two issues:

1. What is the density distribution of the ε's?
2. What is the exact functional form of the deterministic utility $V(\cdot)$?

Let's re-emphasise that the ε's are fundamentally unobserved by the researcher. Hence, one has to make an assumption on the distribution of those that cannot be tested empirically. McFadden (1973) proposes to assume that idiosyncratic preferences are Extreme Value Type I distributed. Figure 2.1 displays an example of such a distribution. As one may observe there

[2] It may seem that the above utility function is not directly related to a setting where consumers maximizing utility *subject to a budget constraint*. However, Appendix A.1 shows that this can be easily accommodated within the current framework.

[3] The term 'idiosyncrasy' originates from Greek 'idiosynkrasía' – 'a peculiar temperament or habit'

[4] Note that this Section focuses on the random *utility* framework. However, the application to profits of firms, rather than utility of people, is straightforward. One would then have a *random profit framework*, which implies that profits consist of a part observed by the researcher and an unobserved idiosyncratic part. See Section 4 for an application.

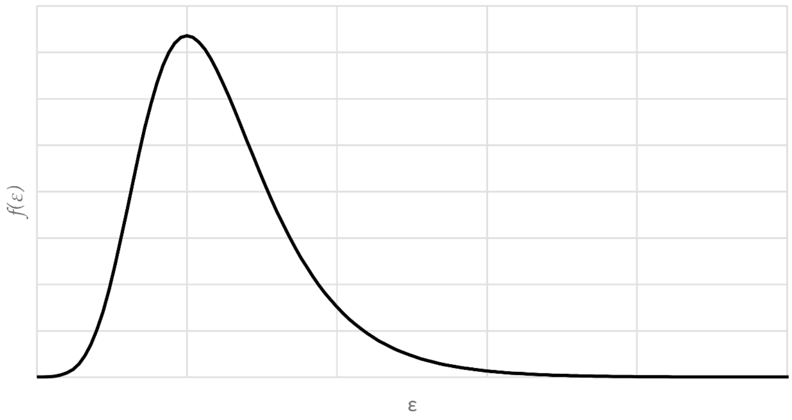

Figure 2.1 – THE EXTREME VALUE TYPE I DISTRIBUTION

is a lot of similarity with a normal distribution, but the distribution is positively skewed.

Why is using this particular distribution so useful? Let's consider again equation (2.4). Train (2003) then shows that:

$$
\begin{aligned}
\Pr(Y = A) &= \Pr(V_A - V_B > \varepsilon_{iB} - \varepsilon_{iA}), \\
&= \int_{\varepsilon_{iA}=-\infty}^{\infty} \int_{\varepsilon_{iB}=-\infty}^{\infty} I\big(\varepsilon_{iA} - \varepsilon_{iB} > -(V_A - V_B)\big) f(\varepsilon_{iA}) f(\varepsilon_{iB}) \mathrm{d}\varepsilon_{iA} \mathrm{d}\varepsilon_{iB}, \\
&= \int_{\varepsilon_{iA}=-\infty}^{\infty} \int_{\varepsilon_{iB}=-\infty}^{\infty} f(\varepsilon_{iA}) f(\varepsilon_{iB}) \mathrm{d}\varepsilon_{iA} \mathrm{d}\varepsilon_{iB} \\
&= 1 - F(-(V_A - V_B)) \\
&= \frac{e^{V_A}}{e^{V_A} + e^{V_B}} \\
&= \frac{1}{1 + e^{V_B - V_A}}.
\end{aligned}
\tag{2.5}
$$

The last line in the above equation is convenient because if one can obtain estimates for the difference between deterministic utilities V_A and V_B, one may simply calculate the probabilities. Moreover, one can easily verify that the predicted probability is always between one and zero, and if V_A and V_B are the same, the predicted choice probability is 0.5. The fact that one just need numbers on the utility differential $V_A - V_B$ implies that there is a so-called *closed form* for the choice probabilities.

Daniel McFadden. In 2000 James Heckman and Daniel McFadden were jointly awarded the Nobel Prize in Economics for their contributions to the analysis of data on individuals and firms. While James Heckman's contribution mainly related to his solution to deal with sample selection, McFadden was lauded for developing models to analyse discrete choice data. He particularly pioneered in developing the theoretical basis for discrete choice.

The second issue if one aims to calculate the probability to choose a certain route is the functional form for the deterministic utility. In principle, one may choose any functional form; however, not every functional form of the deterministic utility is related to a (indirect) utility function with a budget constraint (see Appendix A.1). Let's assume linearity and include more relevant characteristics of a route:

$$U_{iA} = \alpha_A + \beta p_A + \kappa t_A + \varepsilon_{iA}, \tag{2.6}$$

$$U_{iB} = \alpha_B + \beta p_B + \kappa t_B + \varepsilon_{iB}, \tag{2.7}$$

where p_A and p_B are the travel costs to take route A and B respectively; and α_A, α_B, β, and κ are parameters to be estimated. Hence:

$$\Pr(Y = A) = \frac{1}{1 + e^{(\alpha_B - \alpha_A) + \beta(p_B - p_A) + \kappa(t_B - t_a)}}. \tag{2.8}$$

Attentive readers will note that probabilities only depend on the differences in attributes of the two choices. In other words, it is not the absolute time it takes to take route A or B, but only *the difference in travel time is what matters.*

The value of time. An important concept in transport economics is the *value of time* (VOT), which indicates how much a person is willing to pay to reduce travel time with one hour, *holding utility constant.* In many large transport projects, such as the construction of a new road, it is clear what the costs are, but how to determine the benefits? The VOT can be used to predict the benefits, as the benefits of the transport projects would be roughly equal to the number of people using the new road multiplied by their travel time savings and their value of time.

Let denote a change by Δ. Using the utility function (2.6), and holding utility constant, we have:

$$\begin{aligned} \Delta U_{ia} &= \beta \Delta p_A + \kappa \Delta t_A + \Delta \varepsilon_{iA} = 0 \\ -\beta \Delta p_A &= \kappa \Delta t_A, \\ \Delta p_A &= -\frac{\kappa}{\beta} \Delta t_A. \end{aligned} \tag{2.9}$$

Note that $\Delta\varepsilon_{iA} = 0$ in the first line of the above equation, because $\mathrm{E}[\varepsilon_{iA}] = 0$. Suppose that time is measured in hours. Let's evaluate the value of a one hour reduction in travel time, so $\Delta t_A = -1$. Hence:

$$\Delta p_A = \frac{\kappa}{\beta}. \tag{2.10}$$

Equation (2.10) is very convenient because it can be estimated from the data and has direct policy implications. However, the applied utility function does not include income or a budget constraint and seems therefore somewhat odd. Appendix A.1 shows the derivation of the VOT using a more standard utility function.

Many studies have attempted to measure the Value of Time using choice experiments. A study for the Netherlands finds an average VOT for commuters of €9.25 (Kouwenhoven et al., 2014). It is considerably higher for business trips: €26.25, while somewhat lower for social purposes: €7.50. Also income is strongly related to the VOT. Abrantes and Wardman (2011) find that the when income rises by 1%, the VOT increases by 0.9%. Koster and Koster (2015) confirms that VOTs are much higher for higher incomes.

Note that the VOT is just one example of a trade-off between costs/benefits of an alternative and a certain characteristic (*i.e.* time). Other examples are the *Value of Reliability*, or the *Value of a Statistical Life* (see *Application 3*). The latter is an estimate for how much people are willing to pay to reduce their risk of death, which has many applications in the field of transport and environmental economics.

2.3 ALLOWING FOR OBSERVED HETEROGENEITY

In practice, and as discussed in *Application 1*, utility parameters may be heterogeneous. For example, the Value of Time may be higher for high-income households, because the marginal utility of income is lower (see Appendix A.1 for more information). It is pretty straightforward to include this in the utility specification by using interactions between attributes and household characteristics. Let z_i be a household characteristic, *e.g.* income. Then:

$$U_{ij} = \alpha_j + \beta_0 p_j + \beta_1 p_j \cdot z_i + \kappa_0 t_j + \kappa_1 t_j \cdot z_i + \gamma z_i + \varepsilon_{ij}, \tag{2.11}$$

The probability to choose an alternative is now household-specific:

$$\Pr_i(Y = j) = \frac{e^{\alpha_j + \beta_0 p_j + \beta_1 p_j \cdot z_i + \kappa_0 t_j + \kappa_1 t_j \cdot z_i + \gamma z_i}}{\sum_{k=1}^{J} e^{\alpha_k + \beta_0 p_k + \beta_1 p_k \cdot z_i + \kappa_0 t_k + \kappa_1 t_k \cdot z_k + \gamma z_k}}. \tag{2.12}$$

Can we still derive the Value of Time in the above example? Yes, but this will be VOT given the value z_i. Please verify that:

$$\text{VOT} = \frac{\kappa_0 + \kappa_1 z_i}{\beta_0 + \beta_1 z_i}. \tag{2.13}$$

2.4 MULTIPLE ALTERNATIVES

It is straightforward to extend the above framework to a setting with multiple alternatives, which is convenient as many real-life choices have many alternatives. Let's index the alternatives by $j, k = 1, ..., J$. Let's denote attributes of the alternatives by x. It then it is readily verified that:

$$\Pr(Y = j) = \frac{e^{\beta' x_{ij}}}{\sum_{k=1}^{J} e^{\beta' x_{ik}}}, \tag{2.14}$$

where β is a vector of coefficients related to the attributes of the deterministic utility.

The deterministic utility can be used to come up with an estimate for the consumer surplus – or welfare. Small and Rosen (1981) shows that for agent i:

$$\begin{aligned} \mathrm{E}[CS] &= \frac{1}{v}\mathrm{E}\big[\max_{j=1,\dots,J} U_{ij}\big] \\ &= \frac{1}{v}\sum_{j=1}^{J}\Pr(Y=j)U_j \\ &= \frac{1}{v}\log\Big(\sum_{j=1}^{J}\mathrm{e}^{\beta' x_{ik}}\Big)+\mathscr{C}, \end{aligned} \tag{2.15}$$

where v is the marginal utility of income and $\mathscr{C}$ is an unknown constant, representing the fact that the absolute value of utility cannot be measured. Note that the above consumer surplus is the monetised expected direct utility. People often refer to $\log\big(\sum_{j=1}^{J}\mathrm{e}^{\beta' x_{ik}}\big)$ as the *log-sum* – the expected utility from making a choice (De Jong et al., 2005). Note that the log-sum increases when an alternative is added, even if the deterministic utility of that alternative is low. This is what is known as a 'love-for-variety'.

The Multinomial Logit formula also implies a somewhat peculiar property – the so-called *independence of irrelevant alternatives* (IIA). Let's consider three alternatives, j, k and ℓ. Using equation (2.14), the ratio of probabilities for j and k do not depend on whether ℓ is included in the choice set:

$$\frac{\Pr(Y=j)}{\Pr(Y=k)}=\frac{\mathrm{e}^{\beta' x_{ij}}}{\mathrm{e}^{\beta' x_{ik}}}. \tag{2.16}$$

Let's illustrate the IIA-property by introducing the so-called *Red Bus-Blue Bus* problem. Imagine a transportation market with three travel modes: trains, red buses and blue buses. Because red buses and blue buses are identical except for the colour, they have equal market shares of 15%. Let's further suppose that the normalised deterministic utility associated with the train and buses is respectively 2.54 and 1.[5] Now consider that the blue bus is painted red. What are the predicted probabilities? Well, in principle there are now two alternatives – the train and the red bus with deterministic utilities of respectively 2.54 and 1. Hence, the predicted probabilities for the train and the red bus would then be 82.3% and 17.7% so that the *ratio* of probabilities of choosing the train and red bus is unaffected. The bottom line here is that the predicted probabilities are impacted by adding potentially irrelevant alternatives to the choice set (*i.e.* the red bus ánd the blue bus), which is not particularly convenient.

In other words, the IIA implies an unrealistically simple model in this example where an improved product gains share from all other products *in proportion to their original shares*. In the real world, products compete unequally with one another and when an existing product is improved, it usually gains most from a subset of products with which it competes most directly.[6]

[5] Note that $\mathrm{e}^{2.54}/(\mathrm{e}^{2.54}+\mathrm{e}^{1}+\mathrm{e}^{1})=0.7$), so you will arrive at a market share of 70% for the train.

[6] Whether the IIA property is an unrealistic depends on the context. In the context of location choice

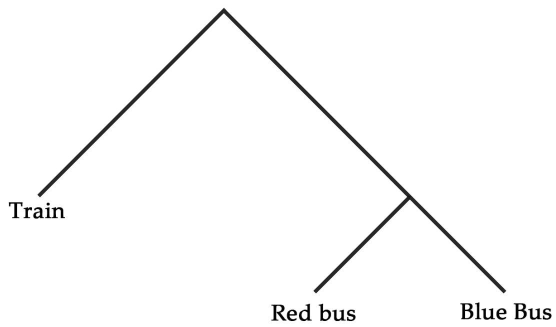

Figure 2.2 – NESTS: AN EXAMPLE

The Multinomial Logit cannot take this into account, but there is an alternative: the so-called Nested Logit model. It allows alternatives to be similar to each other in an unobserved manner. The researcher has to specify a structure that partitions the alternatives into groups (or so-called 'nests') (Heiss, 2002). Figure 2.2 provides an example.

Let's write the following utility function that allows for correlation within nests:

$$U_{jg} = V_j + W_g + \epsilon_{jg} \tag{2.17}$$

where U_{jg} is the utility obtained from choosing a certain alternative j in nest g, V_j is the deterministic utility that only differs within nests between alternatives j, and W_g only differs between nests g.

Train (2003) then shows that:

$$\begin{aligned} \Pr(d_j = 1) &= \Pr(j|g) \times \Pr(g), \\ &= \frac{e^{V_j/\lambda_g}}{\sum_{k\in g} e^{V_k/\lambda_g}} \times \frac{e^{W_g+\lambda_g I_g}}{\sum_{\tilde{g}} e^{W_{\tilde{g}}+\lambda_{\tilde{g}}}}, \end{aligned} \tag{2.18}$$

with $I_g = \log\left(\sum_{j\in g} e^{V_j/\lambda_g}\right)$. I_g is called the inclusive value, which is essentially the log-sum of a given nest. λ_g plays a key role in the above equations and indicates the correlation between nests. If $\lambda_g = 1$ there is no correlation between alternatives in the nest and one is back at the Multinomial Logit model; if $\lambda \to 0$ there is perfect correlation between the alternatives within the nest as in the *Red bus-Blue bus* example.

Equation (2.18) looks quite tedious but is essentially stating that the probability to choose an alternative within the nest multiplied with the probability to choose that nest. Going back to the *Red bus-Blue bus*: the probability to choose the blue bus is the probability to choose the bus (the nest), which is 30%, multiplied by the probability to choose the blue bus as opposed to the red bus, which is 50%. Hence, the probability to choose the blue bus is 15%. When now the blue bus is removed, the probability to choose the bus will remain unaffected, because $\lambda_{\text{bus}} \to 0$.

models (see *Application 2*) it is for example hardly restrictive.

Table 3.1 – SAMPLE DATASET

Respondent identifier	Alternative chosen	Price (in €)	Time (in min)
1	1	14	12
2	0	25	5
3	0	15	15
4	1	15	13
5	1	4	45
6	1	3	40
7	0	20	10
...	...	...	...
I	d_J	p_J	t_J

3 ESTIMATING BINARY DISCRETE CHOICE MODELS

Using the random utility framework and data on choices of people the next step is to estimate the parameters β. Let's first consider *binary* choices. Say that one's data looks something like the data shown in Table 3.1.

Hence, the dependent variable – indicating whether the alternative is chosen – is a dummy variable that depends on two characteristics: price and time. Let's consider three options to estimate such a model: by a Linear Probability Model, a Logit Model or a Probit Model.[7].

3.1 LPM

One may estimate this model by a so-called Linear Probability Model (LPM). The population LPM is given by:

$$\Pr(d_j = 1|x_j) = \beta' x_j, \tag{3.1}$$

where d_j equals one when an alternative is chosen. If $\hat{\beta} = 0.5$, this indicates that the probability to choose j increases by 50 percentage points when x increases by one. Hence, marginal effects are straightforward to obtain and just a function of β.

Equation (3.1) is straightforward to estimate by Ordinary Least Squares by regressing a dummy dependent variable on explanatory variable(s), which is an advantage. In many situations one may want to control for unobserved characteristics of respondents, time trends or location characteristics. This can be done by including fixed effects. Because LPM is estimated by OLS including fixed effects (even multiple fixed effects) is straightforward.

Note that the R^2-statistic is not very useful (Stock and Watson, 2011): imagine a situation in which the R^2 equals one, so that all the data lie on the predicted linear line. This is impossible if the data is binary, except if the covariates x_j are also binary. Note further

[7] One may also estimate Mixed Logit models, where the parameters β have some predefined distribution (see Small et al., 2005). We leave estimation of such models for further study.

that the errors are always heteroscedastic in a LPM.[8] This is easy to accommodate by using heteroscedasticity-robust standard errors (Angrist and Pischke, 2008).

There are essentially three disadvantages related to the Linear Probability Model (Pischke, 2012):

1. The LPM does not estimate the structural parameters of a non-linear model (see Horrace and Oaxaca, 2006). This is particularly a problem when the predictions of d_j often lie outside the unit interval, which is impossible in practice;
2. The LPM does not give consistent estimates of the marginal effects;
3. The LPM does not lend itself towards dealing with measurement error in the dependent variable.[9]

Logit and and Probit regression functions do account for the non-linearity and seem to address the above issues. However, *the exact non-linear function is unknown* because one has to make arbitrary assumptions on the distribution of the ϵ's. According to Angrist and Pischke (2008) and Pischke (2012) there is a lot to be said for sticking to a linear regression function as compared to a fairly arbitrary choice of a non-linear model.

3.2 LOGIT

An alternative to estimating a Linear Probability Model is to estimate a Logit Model. This will entail:

$$\Pr(d_j = 1|x_j) = \frac{1}{1+e^{-\beta' x_j}}. \tag{3.2}$$

The above model is non-linear in its parameters, so it cannot be estimated by Ordinary Least Squares. Instead, one has to rely on *Maximum Likelihood* (ML) to obtain the parameters. The likelihood function is the joint probability distribution of the data as a function of the coefficients to be estimated (Stock and Watson, 2011). The maximum likelihood estimator chooses values of the parameters that maximise the probability of drawing the data that are actually observed. In other words, the ML estimator finds the parameter values that are 'most likely' to have produced the data. The reader is referred to Stock and Watson (2011), pp. 311-312, for more details.

But how to calculate *marginal effects*, *i.e.* what happens with the probability of choosing j when x increases by one unit. Let's rewrite (3.2) as $\Pr(d_j = 1|x_j) = (1+e^{-\beta' x_j})^{-1}$. Then, using

[8] The variance of d_j is given by $\Pr(d_j = 1)(1-\Pr(d_j = 1)) = \beta' x_j(1-\beta' x_j)$, which is varying for different levels of x_j.

[9] A regression equation with measurement error in the dependent variable can be written as $Y_j = Y_j^* + \omega_j = \beta' x_j + \xi_j$. If ω_j is uncorrelated to Y_j^*, the parameters to be estimated β are consistent. However in LPM the dependent variable is binary, so we cannot write $Y_j = Y_j^* + \omega_j$.

the chain rule of differentiation:

$$\begin{aligned}\frac{\partial \Pr(d_j = 1|x_j)}{\partial x_j} &= -(1 + e^{-\beta' x_j})^{-2} \times e^{-\beta' x_j} \times -\beta, \\ &= \beta \cdot \Pr(d_j = 1) \cdot (1 - \Pr(d_j = 1)) \\ &= \beta \frac{e^{-\beta' x_j}}{(1 + e^{-\beta' x_j})^2}.\end{aligned} \tag{3.3}$$

The above formula suggests that there is a non-linear marginal effect that not only depends on β (as in the LPM), but also on the specific values of x_j. Hence, the marginal effect will vary across the sample, dependent on the values of x_j.

The β can be interpreted as an odds ratio – the ratio of the odds of j in the presence of k. Because the estimated β is otherwise hard to interpret, people often report average marginal effects. Standard statistical packages such as `Stata` and `R` can calculate those for you.

The main advantage of a Logit model as compared to the LPM is that the predicted probability will always be within the unit interval. Furthermore, it is easier to include fixed effects in Logit models than in Probit models, because the fixed effects can be conditioned out.[10]

The disadvantages are the following:

1. The assumption of an error term that is Extreme Value Type I distributed seems particular and therefore arbitrary. On the other hand, this assumption is consistent with the Random Utility framework from Section 2.

2. A Logit model is clearly a non-linear model, so it will be slower to estimate. For datasets up to a million observations, this will usually not be a main issue, but beyond that, the LPM may be preferred.

3.3 PROBIT

The final common option to estimate models with a binary dependent variable is to estimate Probit Models, which implies that the unobserved idiosyncratic term ϵ is normally distributed. The Probit model is given by:

$$\Pr(d_j = 1|x_j) = \Phi(\beta' x_j), \tag{3.4}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function. The predicted probability that $d_j = 1$ given values of x_j is calculated by computing the z-score, which is the number of standard deviations from the mean a data point is (see Table 1 in the Appendix of Stock and Watson, 2011, for a table with z-scores).

[10]More specifically, the objective function of the fixed effects Logit estimator (confusingly also referred to as the Conditional Logit estimator) is derived by conditioning the density of d_{ij} on $\sum_i d_{ij}$. Thereby, the fixed effects are eliminated from the log-likelihood (Chamberlain, 1980; Stammann et al., 2016).

The marginal effect is given by:

$$\frac{\partial \Pr(d_j = 1|x_j)}{\partial x_j} = \beta\phi(\beta' x_j). \tag{3.5}$$

where $\phi(\cdot)$ is the density of the normal distribution. Note that there is no closed-form (*i.e.* an explicit formula) describing the marginal effect for the Probit model. Hence, the marginal effect is simply obtained by *(i)* computing the predicted probability for one value of the regressors, *(ii)* computing the predicted probability for a new value, and *(iii)* take the difference. Again, statistical software can do this for you when you indicate to report average marginal effects, or evaluate marginal effects at certain values of x_j.

An advantage of the Probit is that, again, the predicted probabilities are always between zero and one. Furthermore, the assumption of normally distributed ϵ's may be a bit less arbitrary.[11]

The disadvantages are similar to the Logit model, but in contrast to the Logit model, including many fixed effects in Probit model is hard because they cannot be conditioned out and parameters therefore need to estimated for each of the included dummy variables.

3.4 LPM, Logit or Probit

What estimation method is then preferred when estimating a binary choice model? The answer is 'it depends on the situation'. But here are some suggestions:

- Look at the goodness of fit $d_j - \bar{d}_j$ when testing for different models; investigate the fraction correctly predicted (*i.e.* if the probability exceeds 50% and $d_j = 1$, then it is said to be correctly predicted); or use the Pseudo-R^2, which uses the likelihood function to investigate the model performance (Stock and Watson, 2011).
- Test for robustness by calculating marginal effects for all three models. Probit models give usually almost the same results as Logit models (Stock and Watson, 2011).
- If you have a very large sample and mostly interested in marginal effects, then the LPM is preferred. Also if you have multiple fixed effects (*e.g.* year and region fixed effects) the LPM may be preferred because linear models with multiple fixed effects can be readily estimated.
- For smaller samples Probit and Logit are usually more efficient (*i.e.* have smaller standard errors) (say less than 1,000 observations), and usually give almost identical results.

However, there is still an ongoing debate among economists with proponents and opponents of the LPM. Hence, showing that your results are robust to the choice of estimation method is the best you can do!

[11] For example, because of the Central Limit Theorem, which establishes that, in some situations, when independent random terms are added, the distribution of the average are in the limit normally distributed.

Application 1: Estimating the value of time and reliability. In Koster and Koster (2015) we obtain data to estimate commuters' value of travel time and value of arriving at the preferred arrival time at work from a stated choice experiment held among participants of a real-world rewarding experiment to combat congestion. We estimate the willingness to pay values (WTP) for reductions in travel time and schedule delay early and late using a linear specification of schedule delay (Vickrey, 1969).

In Figure 3.1 we show an example of choices an agent is confronted with. Each respondent has to make 10 of these choices. From Figure 3.1 we observe that not only travel time matters, but also how much too early or too late an agent will be at her work relative to her *preferred arrival time*.

We assume that the deterministic part of the utility (V_{nti}) of individual i facing choice c and choosing alternative j from a set of J alternatives, is given by thee types of variables: the expected reward $E[R_{icj}]$, the expected travel time $E[T_{icj}]$, expected schedule delay early $E[SDE_{icj}]$ and expected schedule delay late $E[SDL_{icj}]$. Schedule delay is defined as the deviation of the preferred arrival time PAT_i of respondent i, where arrivals different from PAT_i result in a dis-utility. Following Noland and Small (1995), we assume a linear-additive specification of the systematic utility. The random utility of choosing alternative j is then given by:

$$U_{icj} = \beta^{ER} E[R_{icj}] + \beta^{ET} E[T_{icj}] + \beta^{ESDE} E[SDE_{icj}] + \beta^{ESDL} E[SDL_{icj}] + \epsilon_{icj}, \tag{3.6}$$

The expected values of the variables are given by the probability weighted averages of M masspoints, so $E[X_{icj}] = \frac{1}{M}\sum_{m=1}^{M} p_{injm} X_{icj}$, for attribute $X \subset \{ER, ET, ESDE, ESDL\}$. As Figure 3.1 shows, the choice experiment has 2 possible travel times and therefore $M = 2$. We are interested in willingness to pays. These are given by the following ratios:

$$VOT = -\frac{\beta^{ET}}{\beta^{ER}}, \tag{3.7}$$

$$VSDE = -\frac{\beta^{ESDE}}{\beta^{ER}}, \tag{3.8}$$

$$VSDL = -\frac{\beta^{ESDL}}{\beta^{ER}}, \tag{3.9}$$

where VOT is the value of travel time, $VSDE$ is the value of schedule delay early and $VSDL$ the value of schedule delay late.

The parameters in (3.6) are obtained by estimating a Binary Logit model. We report the results in Table 3.2. We find a Value of Time of €35 per hour. This is higher than found in the literature (see for example Brownstone and Small, 2005; Li et al., 2010). There may be two reasons for this. First, on average this study has a high share of high-income travellers in the sample and since these have a lower marginal utility of income

Your preferred arrival time if there is no delay is: 8:40.

	Alternative 1		Alternative 2	
Departure time from home	6:05		6:50	
Probability	***80%***	***20%***	***90%***	***10%***
Total travel time	**30 min**	**40 min**	**20 min**	**35 min**
Travel time from home to camera A	15 min	15 min	10 min	10 min
Travel time from camera A to camera B	5 min	10 min	5 min	15 min
Travel time from camera B to work	10 min	15 min	5 min	10 min
Arrival time at work	**6:35**	**6:45**	**7:10**	**7:25**
Reward	**4 euro**	**4 euro**	**0 euro**	**0 euro**

Figure 3.1 – A CHOICE WITH TWO ALTEANTIVES

Table 3.2 – BASELINE RESULTS

	(1)
Value of time	€35.05
	(€4.158)
Value of schedule delay early	€23.22
	(€2.211)
Value of schedule delay late	€17.16
	(€1.621)
Number of choices	4,870
Number of individuals	487
Log-likelihood	2,719

Note: The bootstrapped standard errors (500 replications) are in parentheses.

they are less sensitive to rewards than average commuters. Second, it is very likely that travellers are less sensitive to rewarding incentives than to the payment of a congestion toll. This difference in valuation of gains and losses is a common finding in prospect theory studies.

The value of schedule delay early and late are €23 and €17 for being an hour too early or late, respectively. The finding that VSDE >VSDL is remarkable, since usually the opposite is found (people prefer being too early over being too late) (see Lam and Small, 2001; Brownstone and Small, 2005; Li et al., 2010). An explanation may be that this is due to a selection effect of participants who have lower values of schedule delay late and are employed in *e.g.* flexible highly-educated jobs.

4 ESTIMATING MULTINOMIAL DISCRETE CHOICE MODELS

Many choices in real-life are not binary as respondents can choose between many alternatives. Examples are mode choice, route choice, location choice, etc. Let's first consider estimation of Multinomial Logit models, where the number of alternatives is not too large, then Nested Logit, which relaxes the assumption of independence of irrelevant alternatives as discussed in Section 2. This Section concludes by discussing estimation of Conditional Logit models with many alternatives.[12]

4.1 MULTINOMIAL LOGIT WITH ALTERNATIVE-SPECIFIC PARAMETERS

Let's first consider the Multinomial Logit model where parameters are alternative-specific. From here, we just refer to this as the Multinomial Logit model. Hence:

$$\Pr(d_j = 1|x_j) = \frac{e^{\beta'_j x_j}}{\sum_{k=1}^{J} e^{\beta'_k x_k}}. \tag{4.1}$$

Having alternative-specific β's is convenient in certain circumstances. For example, in case of mode choice the travel time parameter may be different in the train compared to the car, because in the train people may spend time working or reading.

In the Multinomial Logit model all coefficients are compared to one base category, which is omitted. Recall that also in the Binary Logit one compares the probability of alternative A, as compared to B. Moreover, if one know the probability of $J-1$ alternatives, one knows that the probability of choosing the last alternative J is simply $1-\sum_{j=1}^{J-1}\Pr(d_j = 1)$.

While having alternative-specific β's in one case might be useful, it becomes less useful when the number of alternatives grow large: the set of resulting coefficients will be too large to be able to meaningfully interpret the coefficients. Then, the Nested Logit or Conditional Logit models are probably more useful.

4.2 NESTED LOGIT

Nested Logit overcomes the particular Independence of Irrelevant Alternatives property of Multinomial Logit models (see Section 2.4) – adding an alternative does not affect the relative odds between two other options considered. Nested Logit models can be estimated in standard packages such as `Stata` or more specialised ones such as `Biogeme`. The researcher has to define the nests, but the software will calculate the correlation parameter λ_g for each nest. If λ_g is close to one for all nests, one might as well estimate a standard Multinomial or Conditional Logit model.

[12] It may also be that one's dependent variable has an ordering (*e.g.* a likert-scale of satisfaction). In cases where Ordinary Least Squares are inappropriate one may consider the estimation of Ordered Logit or Ordered Probit models. The exact estimation of these models reaches beyond the scope of this syllabus.

4.3 CONDITIONAL LOGIT

Many real-life choices are characterised by many alternatives. For example, when households choose where to reside, they have many alternatives. Also when choosing to buy a certain car, an agent faces many alternatives. The Multinomial Logit with alternative-specific parameters is not very suitable because the many coefficients would be hard to interpret. Also the Nested Logit may not be useful, because it my be hard to define nests for many alternatives (nevertheless, see for an example Head and Mayer, 2004). Moreover, there is no reason why, say, a park or water body have a different impact when choosing location i or location j. Let's quickly recall the Conditional Logit specification:

$$\Pr(d_j = 1|x_j) = \frac{e^{\beta' x_{ij}}}{\sum_{k=1}^{J} e^{\beta' x_{ik}}}, \tag{4.2}$$

The issue is that with many alternatives and many agents, Logit models may be infeasible to estimate. Consider, for example, the case where you would observe a thousand agents that choose their location in The Netherlands. The Netherlands has about half a million detailed zip-code locations. Hence, this would entail the inclusion of $1,000 \times 500,000 = 0.5$ billion observations. Few software packages can handle such large number of observations. However, note that with the current surge in the availability of Big Data it is likely that the number of agents and alternatives is even larger.

There are essentially three ways to overcome the problem of a too large number of observations:

1. The first way is to aggregate alternatives in a meaningful way. For example, instead of modelling location choices of households at the zip-code level, one may model this at the municipality level. Rather than having half a billion observations, one now has $1,000 \times 415 = 415,000$ observations, which is perfectly possible to estimate. Dependent on the application, this is a useful way to deal with the large number of observations. For example, if one is interested in the effects of municipality-specific taxes on location choices of people, this is a fine approach. On the other hand, if one is interested in the effects of parks on location choices, this may not be appropriate because the effects of parks tend to be very local.

2. It may therefore be preferred to consider an alternative way. McFadden (1978) proposes to include the chosen alternative in the dataset, as well as *a random subset of J alternatives*. For example, one includes the chosen car type and randomly selects 10 cars that are not chosen. McFadden (1978) shows that this does not affect the *consistency* of the estimated properties. So with many agents and a not too small number of alternatives (a large $n \times J$), this approach should give the correct parameters. This approach has been applied in many good papers (see *e.g.* Bayer et al., 2007). The question is, however, how many alternatives to include; and what is the minimum number of agents, to get approximately the right parameters? The answer to this question is unknown, because the so-called small sample properties of this approach are yet unclear.

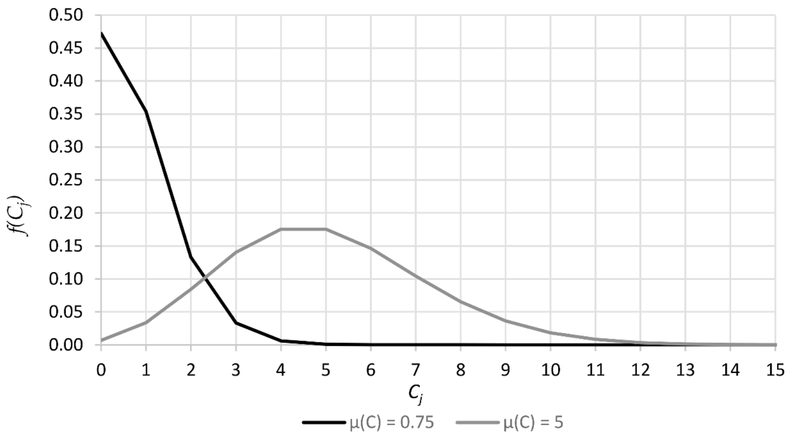

Figure 4.1 – THE POISSON DISTRIBUTION

3. A third promising alternative is to estimate the Conditional Logit model (4.2) by means of a Poisson model. A Poisson model is a count data model, which has a dependent variable that is a count variable (0, 1, 2, 3, and so on). Rather than including the chosen alternative and the unchosen alternative for each agent, one now *counts the times an alternative is chosen.* Let's estimate the following regression:

$$C_j = \mathrm{e}^{\beta' x_j}, \tag{4.3}$$

where C_j is the number of times an alternative is chosen, which is assumed to be *Poisson distributed.* Examples of Poisson distributions are given in Figure 4.1.[13] In line with Guimarães et al. (2003), Appendix A.2 shows that the Likelihood function of a Conditional Logit model and a Poisson model are identical up to a constant, implying that they deliver the same parameters β.

A big advantage is that the number of observations is now equal to the number of alternatives, which almost always implies that the number of observations is not too large. Another advantage is that the coefficients can be interpreted in a straightforward manner, as they are equal to (semi-)elasticities.

A disadvantage of using Poisson models is that individual characteristics, such as income, age or household composition, cannot be take into account, *i.e.* x_{ij} cannot be i-specific. The reason is that for the Conditional Logit model to be equivalent to a Poisson model, one has to aggregate the choices of all agents.[14] Hence, dependent on the application, this is an issue.

[13] A Poisson model implies equidispersion, which means that the mean is equal to the variance of the dependent variable. When this is (approximately) the case, or when the number of alternatives is very large, this is fine. If this is not the case, one may consider to estimate Negative Binomial regressions or Zero-Inflated Poisson regressions (see Guimarães et al., 2004). Those models reach beyond the scope of this syllabus.

[14] Guimarães et al. (2003) and Koster et al. (2014) derive some special cases in which agent's characteristics can be included.

Application 2: The location choice of business services. Jacobs et al. (2013) studies the location choice of start-ups of business services as a function of the presence of multinational enterprises (MNEs) and other business services. This means that there is a situation where there are many agents (start-ups) and many alternatives (locations).

The spatial behaviour of (knowledge-intensive) business services and their contribution to regional growth and innovation systems has gained growing scholarly attention over the last decade (Den Hertog and ., 2000; Muller and Zenker, 2001; Keeble and Nachum, 2002; Wood, 2002; Koch and Stahlecker, 2006). In empirical case studies, scholars have suggested that business services cluster in large metropolitan areas due to the agglomeration benefits they enjoy with their (global) clients (Keeble and Nachum, 2002; Shearmur and Alvergne, 2002; Shearmur and Doloreux, 2008; Müller and Doloreux, 2009). This is related to agglomeration economies, which include, most notably, input sharing, knowledge spillovers and a specialised labour force, which result in increasing returns to scale.

The micro-data on the location of business services and other firms in each zip-code in the Northwing of the Randstad are derived from LISA. See Van Oort (2004) for a detailed description of these data and Figure 4.2 for an overview of the study area. The data on MNEs are derived from the Achilles database for foreign-owned multi-nationals. The Achilles database is compiled by the Dutch national agency for foreign direct investment (NFIA) and consists of MNE establishments at locations in the Netherlands, including their year of establishment in the Netherlands and their number of employees.

We then consider the location choice of business services start-ups as a function of the density of multi-national employment, business services employment and other employment. Hence, we assume the following profit function for a start-up:

$$\pi_{ij} = \alpha + \beta e_j^{MNE} + \gamma e_j^{BS} + \delta e_j^{OF} + \zeta X_j + \eta_{j\in M} + \epsilon_{ij}, \tag{4.4}$$

where π_{ij} is the profit of firm i locating in j, e_j^{MNE} is the spatially weighted density of employment in multi-national enterprises, e_j^{BS} the density in business services firms, e_j^{OF} the density of employment in other firms, X_j are control variables related to infrastructure and planning (*e.g.* the distance to the nearest highway and station), $\eta_{j\in M}$ are municipality fixed effects and ϵ_{ij} is an idiosyncratic shock to profits that is assumed to be Extreme Value Type I distributed. Hence, the probability that firm i will choose location j is given by:

$$\Pr(d_j = 1) = \frac{\mathrm{e}^{\alpha + \beta e_j^{MNE} + \gamma e_j^{BS} + \delta e_j^{OF} + \zeta X_j + \eta_{j\in M}}}{\sum_{k=1}^{J} \mathrm{e}^{\alpha + \beta e_k^{MNE} + \gamma e_k^{BS} + \delta e_k^{OF} + \zeta X_k + \eta_{k\in M}}}. \tag{4.5}$$

Because there are no firm-specific characteristics included in the profit function, one can estimate equation (4.5) by a Poisson model where the number of business services

Table 4.1 – A POISSON MODEL
(Dependent variable: The number of business services start-ups per location)

	(1)	(2)	(3)
Multi-national employment density *(log)*	0.0709***	0.0422***	0.0772***
	(0.0151)	(0.0092)	(0.0121)
Business services employment density *(log)*	0.4304***	0.4374***	0.3821***
	(0.0240)	(0.0162)	(0.0214)
Other employment density *(log)*	-0.2242***	-0.2203***	-0.1352***
	(0.0162)	0.0071	(0.0178)
Control variables (9)	No	Yes	Yes
Municipality fixed effects (61)	No	No	Yes
Number of locations	13,655	13,655	13,655
Log-likelihood	-13,146.903	-13,051.163	-12,709.249

Notes: We include locations with at least 10 employees in 2000. The coefficients can be interpreted as elasticities and differ from Jacobs et al. (2013) because of a slightly different set of controls and because we estimate Poisson models instead of Negative-binomial regressions. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

start-ups in zip-code is the dependent variable.

We report results in Table 4.1. In the first column we only include the variables of interest. As stressed earlier, the coefficients in a Poisson model can be readily interpreted as elasticities. Hence, the coefficient indicates that a 1% increase in the density of employment in multi-national firms is associated with a 0.071% increase in business services start-ups. The impact of business services employment is much more important: a 1% increase in density of business services employment is associated with a 0.43% increase in business services start-ups. Hence, localisation – within-sector – agglomeration economies seem to be a much more important determinant of location choices than the presence of multi-national enterprises. Having said this, the impact of multi-national enterprises is not negligible. The final coefficient relates to the effect of density of employment in other firms, which we find to be negative. The reason may be that other employment does not generate much agglomeration economies for business services, while at the same time raising land rents, so that it is less attractive to locate in areas with just a higher concentration of (unrelated) employment.

We investigate whether omitted variable bias is an issue by including a set of control variables (*i.e.* distance to the highway, railway station, etc.) in column (2) and control variables and municipality fixed effects in column (3). One may see that the coefficients of interest do not change substantially, which suggests that omitted variable bias is not a big issue.

Figure 4.2 – THE STUDY AREA IN JACOBS ET AL. (2013)

5 STATED AND REVEALED PREFERENCE DATA

Quantitative empirical research relies on data. Data can either be obtained from existing sources on observed or reported actual behaviour. Those data are referred to as *Revealed Preference* (RP) data. Alternatively, one may undertake a so-called stated choice experiment, where respondents are confronted with hypothetical choice sets. Those are so-called *Stated Preference* (SP) data. Although research designs are imaginable where one combines RP and SP data, usually one chooses between the use of RP and SP data.

5.1 REVEALED PREFERENCE DATA

Revealed preference data is often readily available. Given the rise of so-called *Big Data* more and more individual-level data is becoming available. For example, large datasets on housing transactions may be used to model residential location choices; or micro-data from Statistics Netherlands on the workplace and residential location may be used to model commuting decisions.

The advantages of RP data are the following:

1. Probably the biggest advantage of RP data is that it is based on actual behaviour. Hence, biases arising from the discrepancy between stated choices and actual choices are irrelevant. Those biases are discussed in the next subsection.

2. Usually, RP data is relatively cheaper to obtain, as those have been collected for other purposes. By using RP data one avoids the high costs associated with designing and undertaking stated choice experiments.

3. RP data is often collected over a longer time, enabling the researcher to track individuals/agents over a (long) time. This provides the researcher with sufficient variation to control for unobserved factors, *e.g.* by including fixed effects at the level of the agent.

However, there are also a couple of disadvantages associated with the use of RP:

1. In certain contexts, lack of variability may be an issue. For example, if one considers the effects of gasoline taxes on car ownership decisions, one faces the issue that gasoline taxes usually only vary at the country level.

2. There may be (multi-)collinearity between attributes of alternatives. For example, if one considers the trade-off between travel costs and travel time (*e.g.* in order to be able to calculate the value of time), one faces the issue that longer trips almost always imply higher costs (*e.g.* because of higher gasoline costs). Hence, it may be hard to distinguish between the two.[15]

3. RP data is of course hard to use when analysing new choice alternatives. The hydrogen-powered car, for example, is yet hardly used. Using existing data it will be hard to predict what will happen when this type of car becomes much more affordable.

4. Actual behaviour observed in RP data is not necessarily in accordance with utility or profit maximisation. University students may for example end up in the study of their second choice because of *numerus fixus*. To the extent these additional constraints are unknown, this may imply biases in the estimated parameters if it is assumed that decision makers freely choose.

5. Agents may also be subject to perception errors and imperfect information. For example, prices of airline tickets depend on where and when the person buys his ticket. Hence, because of imperfect information, the person may end up making a suboptimal choice.

5.2 STATED PREFERENCE DATA

In stated choice experiments agents are confronted with alternatives based on usually two alternatives. An example of such a choice is given in Figure 3.1.

The advantages of SP data are the following:

[15] Nevertheless, there are exceptions where there is enough variability in travel times and costs, see Peer et al. (2015).

1. In contrast to RP data, it is relatively easy to introduce new alternatives with new attributes. This would help in predicting market shares of currently non-existing products.

2. Large variability between the attributes is also easy to achieve, because the values are to be determined by the researcher. Hence, large differences between, say, prices and travel times can be included in the different attributes. To the extent collinearity between attributes is an issue one may solve this by implementing an *orthogonal design*, as to minimise the correlation between attributes (see Koster and Tseng, 2010, for a discussion).

3. Because stated choice experiments occur in a 'laboratory' setting, the choice set is clearly defined and it is clear what trade-off people make.

Using SP is not without its problems. The main disadvantages are summarised as follows:

1. There are several biases associated with the use of SP data (Brown, 2019).[16]

    - There is *information bias*. This arises when the respondent has incorrect information on the research context, which may lead to misunderstanding. To reduce information bias, one should make the experiment as realistic as possible.

    - There may be a *starting point bias*. There is a large literature on behavioural economics showing that respondents are influenced by the set of available responses to the experiment. Hence, agents may make different trade-offs when setting a relatively high or low price in the attributes. It is therefore important to thoroughly test your design and choose attribute values that are in accordance with reality.

    - There could be a *hypothetical bias*. It appears that respondents tend to respond differently to hypothetical scenarios than they do to the same scenarios in the real world, *i.e.* there is cognitive incongruity with actual behaviour. To potentially reduce hypothetical bias it is paramount to make the experiment as realistic as possible. However, the hypothetical bias is otherwise hard to mitigate and is probably the main disadvantage of stated choice experiments.

    - The final bias is a *strategic bias* and may arise when respondents wants a specific outcome. Strategic biases are only relevant in specific contexts in which respondents have a direct or indirect interest in the outcome of the choice experiment.

    Note that information bias, starting point bias and hypothetical bias are *unintentional* biases (Brown, 2019), as the agent is not aware of the mistakes he or she makes. Strategic bias, on the other hand, is an *intentional* bias, as the respondent is aware of the fact that his stated behaviour may be fundamentally different from his actual behaviour.

2. A second issue in SP data is the fact that choices may be subject to artificial errors.

[16] See for these and other psychological biases, Tversky and Kahneman (1981).

> These errors may arise due to boredom, due to the fact that respondents do not carefully read instructions or due to respondents not understanding the questions properly. It is important that stated choice experiments do not take too long and are not too complicated so that all respondents can partake meaningfully in the experiment.

The question remains whether to rely on RP or SP data. The outcome of this decision is of course highly context-specific. As a general rule-of-thumb, economists prefer RP data because they consider the biases associated with RP data less severe than the ones associated with SP data.

Application 3. The value of a statistical life. The value of a statistical life (VSL) is an economic value used to quantify the benefits of avoiding a fatality. VSL is more of an estimate of willingness to pay for small reductions in mortality risks rather than how much a human life is worth. In the evaluation of fatal accident costs the availability of an estimate of the economic value of a statistical life is paramount, as in Europe, for example, approximately 40,000 fatalities occur in traffic accidents every year. The number of additional non-fatal accidents amounts to a multiple of this figure.

De Blaeij et al. (2003) undertake a meta-analysis based on 30 studies to study the different values obtained in the literature regarding the value of a statistical life in road safety. The estimates of the VSL in different studies is vastly different, and they range from less than $ 200,000 to almost $ 30 million (in 1997 prices).

Using regression techniques De Blaeij et al. (2003) then aim to explain these huge differences. An interesting finding is that there are large differences between studies using SP and RP data. More specifically, as expected, stated preference studies lead to about 132% *higher* estimates than revealed preference studies, so the difference is non-negligible. This may be explained by RP studies referring to policy measures that are actually implemented, while policy measures in SP studies are often purely hypothetical – a hypothetical bias. Furthermore, they also show that differences in survey design (particularly regarding payment vehicle and elicitation format) also have a strong impact on the outcome. Hence, the study of De Blaeij et al. (2003) illustrates that the different biases discussed earlier are very important and should not be ignored.

6 SUMMARY

This syllabus discussed discrete choice models. First, the basic theory underlying discrete choice is introduced, which is the *random utility framework*. This framework can be readily extended to multiple, or *multinomial*, choices. A particular property of the Multinomial Logit setting is considered: the *Independence of Irrelevant Alternatives*. Defining nests within the alternatives and estimating Nested Logit models may circumvent this issue.

Second, estimation of binary choice models is discussed – models where the dependent variable is dichotomous. One may estimate *Linear Probability models*, *Logit models* and *Probit models*. Each estimation method has its advantages and disadvantages, but because of its straightforward implementation, the Linear Probability model is probably preferred when one has panel data and many fixed effects.

Finally, the estimation of multinomial choice models is considered – models where the dependent variable has more than two possible discrete outcomes. *Multinomial Logit models* with alternative-specific parameters are discussed first. This model relevant in the context of mode choice. Then, estimation of *Nested Logit* and *Conditional Logit* models are discussed. The latter are models where there are many alternatives. To deal with potentially large choice sets, one may estimate Conditional Logit models by a *Poisson* model.

REFERENCES

Abrantes, P. and M. Wardman (2011). "Meta-analysis of UK Values of Travel Time: An Update". In: *Transportation Research Part A: Policy and Practice* 45.1, pp. 1–17 (cit. on p. 7).

Angrist, J. and J. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press (cit. on p. 11).

Bayer, P., F. Ferreira, and R. McMillan (2007). "A Unified Framework for Measuring Preferences for Schools and Neighborhoods". In: *Journal of Political Economy* 115.4, pp. 588–638 (cit. on p. 17).

Ben-Akiva, M. and S. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* Cambridge, MA: MIT press (cit. on p. 2).

Brown, J. L. (2019). "Environmental Economics," in: *Encyclopaedia Britannica* https://www.britannica.com/topic/environmental-economics (cit. on p. 23).

Brownstone, D. and K. A. Small (2005). "Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations." In: *Transportation Research Part A: Policy and Practice* 39.4, pp. 279–293 (cit. on pp. 14, 15).

Chamberlain, G. (1980). "Analysis of Covariance with Qualitative Data". In: *Review of Economic Studies* 47, pp. 225–238 (cit. on p. 12).

De Blaeij, A. et al. (2003). "The Value of Statistical Life in Road Safety: A Meta-analysis." In: *Accident Analysis and Prevention* 35.6, pp. 973–986 (cit. on p. 24).

De Jong, G. et al. (2005). "Using the Logsum as an Evaluation Measure". In: *RAND working paper WR-275-AVV* (cit. on p. 8).

Den Hertog, P. D. and . (2000). "Knowledge-intensive Business Services as Co-producers of Innovation". In: *International Journal of Innovation and Management* 4, pp. 491–528 (cit. on p. 19).

Guimarães, P., O. Figueirdo, and D. Woodward (2003). "A Tractable Approach to the Firm Location Decision Problem." In: *Review of Economics and Statistics* 85.1, pp. 201–204 (cit. on pp. 18, A1).

Guimarães, P., O. Figueirdo, and W. Woodward (2004). "Industrial Location Modeling: Extending the Random Utility Framework". In: *Journal of Regional Science* 44.1, pp. 1–20 (cit. on p. 18).

Head, K. and T. Mayer (2004). "Market Potential and the Location of Japanese Investment in the European Union". In: *Review of Economics and Statistics* 86.4, pp. 959–972 (cit. on p. 17).

Heiss, F. (2002). "Specification(s) of Nested Logit Models". In: *University of Mannheim, Working paper* (cit. on p. 9).

Horrace, W. C. and R. L. Oaxaca (2006). "Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model". In: *Economics Letters* 90.3, pp. 321–327 (cit. on p. 11).

Jacobs, W. A. A., H. R. A. Koster, and F. G. Van Oort (2013). "Co-agglomeration of Knowledge-intensive Business Services and Multinational Enterprises". In: *Journal of Economic Geograp* 14.2, pp. 443–475 (cit. on pp. 19–21).

Keeble, D. and L. Nachum (2002). “Why do Business Service Firms Cluster? Small Consultancies, Clustering and Decentralization in London and Southern England”. In: *Transactions of the Institute of British Geographers* 27.1, pp. 67–90 (cit. on p. 19).

Koch, A. and T. Stahlecker (2006). “Regional Innovation Systems and the Foundation of Knowledge Intensive Business Services. A Comparative Study in Bremen, Munich, and Stuttgart, Germany.” In: *European Planning Studies,* 14.2, pp. 123–146 (cit. on p. 19).

Koster, H. R. A., J. N. van Ommeren, and P. Rietveld (2014). “Estimation of Semiparametric Sorting Models: Explaining Geographical Concentration of Business Services”. In: *Regional Science and Urban Economics* 44, pp. 14–28 (cit. on p. 18).

Koster, P. R. and H. R. A. Koster (2015). “Commuters’ Preferences for Fast and Reliable Travel: A Semi-Parametric Approach”. In: *Transportation Research Part B: Methodological* 81, pp. 289–301 (cit. on pp. 7, 14).

Koster, P. R. and Y. Y. Tseng (2010). “Stated Choice Experimental Designs for Scheduling Models”. In: *Choice Modelling: The State-of-the-art and The State-of-practice*. Ed. by S. Hess and A. J. Daly, pp. 217–235 (cit. on p. 23).

Kouwenhoven, M. et al. (2014). “New Values of Time and Reliability in Passenger Transport in The Netherlands”. In: *Research in Transportation Economics* 47.37-49 (cit. on p. 7).

Lam, T. C. and K. A. Small (2001). “The Value of Time and Reliability: Measurement from a Value Pricing Experiment.” In: *Transportation Research Part E: Logistics and Transportation Review* 37.2-3, pp. 231–251 (cit. on p. 15).

Li, Z., D. A. Hensher, and J. M. Rose (2010). “Willingness to Pay for Travel Time Reliability in Passenger Transport: A Review and Some New Empirical Evidence.” In: *Transportation research part E: logistics and transportation review,* 46.3, pp. 384–403 (cit. on pp. 14, 15).

McFadden, D. (1973). “Conditional Logit Analysis of Qualitative Choice Behavior”. In: *Frontiers in Econometrics,* pp. 105–142 (cit. on p. 4).

— (1978). “Modeling the Choice of Residential Location”. In: *Transportation Research Record* 673, pp. 72–77 (cit. on p. 17).

Müller, E. and D. Doloreux (2009). “What We Should Know about Knowledge-Intensive Business Services, .” In: *Technology in Society* 31, pp. 64–72 (cit. on p. 19).

Muller, E. and A. Zenker (2001). “Business Services as Actors of Knowledge Transformation: The role of KIBS in Regional and National Innovation Systems.” In: *Research policy,* 30.9, pp. 1501–1516 (cit. on p. 19).

Noland, R. and K. A. Small (1995). “Travel-time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes”. In: *Transportation Research Record* 1493, pp. 150–158 (cit. on p. 14).

Peer, S. et al. (2015). “Long-Run Versus Short-Run Perspectives On Consumer Scheduling: Evidence From A Revealed-Preference Experiment Among Peak-Hour Road Commuters”. In: *International Economic Review* 56.1, pp. 303–323 (cit. on p. 22).

Pischke, J. S. (2012). *Probit better than LPM?* (Cit. on p. 11).

Shearmur, R. and C. Alvergne (2002). “Intrametropolitan Patterns of High-order Business Service Location: A Comparative Study of Seventeen Sectors in Ile-de-France,” in: *Urban Studies* 39, pp. 1143–1163 (cit. on p. 19).

Shearmur, R. and D. Doloreux (2008). "Urban Hierarchy or Local Buzz? High-Order Producer Services and (or) Knowledge-Intensive Business Service Location in Canada, 1991-2001". In: *The Professional Geographer* 60, pp. 333–355 (cit. on p. 19).

Small, K. A. and H. S. Rosen (1981). "Applied Welfare Economics with Discrete Choice Models". In: *Econometrica* 49.1, pp. 105–131 (cit. on p. 8).

Small, K. A., C. Winston, and J. Yan (2005). "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability". In: *Econometrica* 73.4, pp. 1367–1382 (cit. on p. 10).

Stammann, A., F. Heiß, and D. McFadden (2016). "Estimating Fixed Effects Logit Models with Large Panel Data," in: *Beiträge zur Jahrestagung des Vereinsfür Socialpolitik 2016* (cit. on p. 12).

Stock, J. and M. Watson (2011). *Introduction to Econometrics*. 3rd editio. Cambridge, MA: Pearson (cit. on pp. 10–13).

Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press (cit. on pp. 4, 5, 9).

Tversky, A. and D. Kahneman (1981). "The Framing of Decisions and the Psychology of Choice". In: *Science* 211.4481, pp. 453–458 (cit. on p. 23).

Van Oort, F. G. (2004). *Urban growth and Innovation. Spatially Bounded Externalities in the Netherlands*. Aldershot: Ashgate. (cit. on p. 19).

Vickrey, W. (1969). "Congestion Theory and Transport Investment". In: *American Economic Review* 59.2, pp. 251–260 (cit. on p. 14).

Wood, P. (2002). "Knowledge-intensive services and urban innovativeness." In: *Urban Studies* 39.5-6, pp. 993–1002 (cit. on p. 19).

APPENDIX

A.1 UTILITY MAXIMISATION AND THE VALUE OF TIME

Here the value of time from a standard utility framework where people maximise utility is derived. Let's assume the following utility function:

$$U_i = \sum_{j=1}^{J} \psi_{ij} d_{ij} + \beta q_0, \tag{A.1}$$

where U_i is the utility of individual i and $j = 1, \ldots, J$ are the number of alternatives. d_{ij} equals one when the alternative is chosen by agent i and is zero otherwise. q_0 denotes the consumption of a composite good.

Consumers maximise the above utility function subject to a budget constraint:

$$\bar{y} = \sum_{j=1}^{J} p_j d_{ij} + q_0, \tag{A.2}$$

where $\bar{y}$ is the income, which is spend either on the alternative chosen or on the composite good. The consumption of the composite good can be written as $q_0 = \bar{y} - \sum_{j=1}^{J} p_j d_{ij}$ and plug this in (A.1):

$$\begin{aligned} U_i &= \sum_{j=1}^{J} \psi_{ij} d_{ij} + \beta \bar{y} - \beta \sum_{j=1}^{J} p_j d_{ij}, \\ &= \beta \bar{y} + \sum_{j=1}^{J} (\psi_{ij} - \beta p_j) d_{ij}, \end{aligned} \tag{A.3}$$

Let's assume that $\beta \bar{y}$ is deterministic. Hence, a utility maximising individual will maximise $\max_j (\psi_{ij} - \beta p_j)$. Let's now write $\psi_{ij} = V_j + \epsilon_{ij} = \kappa t_j + \epsilon_{ij}$. Hence, the utility received from an alternative is a function of an deterministic part, say the travel time, and a idiosyncratic part. Hence, without loss of generality, one may write:

$$U_i = \beta p_j + \kappa t_j + \epsilon_{ij}, \tag{A.4}$$

which is the familiar specification of utility used in Section 2.

The above is straightforward to derive with linear income. However, with non-linear income effects, it becomes (much) harder, if possible, to derive the value of time directly from a utility function.

A.2 FROM LOGIT TO POISSON

Here the proof Guimarães et al. (2003) is replicated showing why a Conditional Logit model should deliver the same coefficients as a Poisson model. Let's consider a Conditional Logit setting where agents have to choose between many locations (say location choice). Let $i = 1,...,N$ be the number of agents and $j,...,J$ the number of alternatives. d_{ij} is a dummy variable that equals one when the alternative is chosen by agent i and is zero otherwise. The log-likelihood function of a Conditional Logit Model can then be written as:

$$\log \mathscr{L}_{CL} = \sum_{i=1}^{N}\sum_{j=1}^{J} d_{ij} \log \frac{\mathrm{e}^{\alpha+\beta' x_{ij}}}{\sum_{k=1}^{J} \mathrm{e}^{\alpha+\beta' x_{ik}}}. \tag{A.5}$$

When characteristics of alternatives do not differ between agents (*e.g.* households and firms have identical preferences up to the idiosyncratic term):

$$\log \mathscr{L}_{CL} = \sum_{j=1}^{J} n_j \log \frac{\mathrm{e}^{\alpha+\beta' x_{j}}}{\sum_{k=1}^{J} \mathrm{e}^{\alpha+\beta' x_{k}}}, \tag{A.6}$$

where n_j is the number of agents that choose an alternative j.

Let's now assume that n_j is Poisson distributed with:

$$\mathrm{E}[n_j] = \lambda_j = \mathrm{e}^{\alpha+\beta' x_j}, \tag{A.7}$$

then one may write the log-likelihood of the Poisson model as follows:[17]

$$\begin{aligned}\log \mathscr{L}_P &= \sum_{j=1}^{J} \Big(-\lambda_j + n_j \log \lambda_j - \log n_j! \Big), \\ &= \Big(-\mathrm{e}^{\alpha+\beta' x_j} + n_j(\alpha + \beta' x_j) - \log n_j! \Big).\end{aligned} \tag{A.8}$$

Let's consider the first-order condition with respect to α:

$$\frac{\partial \log \mathscr{L}_P}{\partial \alpha} = \sum_{j=1}^{J} \Big(n_j - \mathrm{e}^{\alpha+\beta' x_j} \Big) = 0, \tag{A.9}$$

from which we can derive:

$$\mathrm{e}^{\alpha} = \frac{N}{\sum_{j=1}^{J} \mathrm{e}^{\alpha+\beta' x_j}}. \tag{A.10}$$

Let's substitute α back into the log-likelihood function (A.8) to obtain:

$$\begin{aligned}\log \mathscr{L}_P &= -N + N - \sum_{j=1}^{J} n_j \log\Big(\mathrm{e}^{\alpha+\beta' x_j}\Big) + \sum_{j=1}^{J} n_j \beta' x_j - \sum_{j=1}^{J} \log n_j!, \\ &= \sum_{j=1}^{J} n_j \log \frac{\mathrm{e}^{\alpha+\beta' x_j}}{\sum_{k=1}^{J} \mathrm{e}^{\alpha+\beta' x_k}} - N + N \log N - \sum_{j=1}^{J} \log n_j!.\end{aligned} \tag{A.11}$$

It can be seen that the first part on the second line of the above log-likelihood function is

[17] Note that $n! = n \times (n-1) \times (n-2) \times (n-3) \times ... \times 3 \times 2 \times 1$.

identical to the log-likelihood function of the Conditional Logit model, while the remaining terms are constants. Hence, the resulting estimates β should also be the same.