Demand Modeling Using Discrete Choice Analysis – Part 1

Motivation

As designers, whether focused on satisfying user wants or on making profit for a firm, we are interested in the preferences that people have and the choices that they make. Formal mathematical models of preference and choice structures built upon empirical data help designers make predictions about the appeal of new products or changes to existing products. These models can help inform intuition and assist the designer in understanding and designing for the market, avoiding the tendency for designers to design only with respect to the preferences of the people they know best - themselves. In this class, we will consider one class of consumer choice models built on utility theory.

Utility Theory

Utility is a ubiquitous concept in economics as an abstract measurement of the degree of goal-attainment or want-satisfaction provided by a product or service.[1] We cannot measure directly how much utility a person may gain from a product; however, we can make inferences about utility based on the person's behavior, if we presume that people act *rationally*. In computer science, a *rational* agent is defined as one that acts to attain its goals. Likewise, in economics we assume that a rational person acts to increase her utility.

All else being equal, if a rational consumer is given a choice between product A, with utility $u_{\mathrm{A}} = 1$, and product B, with utility $u_{\mathrm{B}} = 2$, she will choose product B because it provides more utility. In general, given a set of alternatives $j = \{1,2,...,J\}$, a rational person will choose the alternative that provides the highest utility, so that alternative j is chosen if $u_j > \{u_{j'}\}_{\forall j' \neq j}$. This model does not take into account the degree to which the utility of one product exceeds the utility of another. For instance, if $u_{\mathrm{A}} = 1$ then product B will be chosen if $u_{\mathrm{B}} > 1$, regardless of weather $u_{\mathrm{B}} = 1.0001$ or $u_{\mathrm{B}} = 1000$. In reality, uncertainty in utility estimates would lead one to be more confident in predicting choice B if $u_{\mathrm{B}} = 1000$ and less confident if $u_{\mathrm{B}} = 1.0001$

Random Utility Discrete Choice Models

In general, we cannot measure utility (predict choices) exactly because, for example, we may not be able to observe or measure every characteristic of the individual, product, or choice situation that affects choice behavior; however, if we can observe some information about the individual, the product, or the choice situation, we can use that information to help predict choice. So, in random utility models we presume that the utility u_{ij} provided to individual I by product j is composed of a deterministic component v_{ij}, which can be calculated based on observed characteristics, and a stochastic error component ε_{ij}, which is unobserved, so that

$$u_{ij} = v_{ij} + \varepsilon_{ij}\,. \tag{1}$$

Later we will discuss how to estimate the observable component of utility v_{ij} for individual i and product j using data, but for now we take it as given. Because we never observe the error component ε_{ij}, we do not have enough information to predict a specific individual's choice on a specific choice occasion, but, as in regression, we can make predictions about the patterns of choices over many individuals and many choice occasions. The probability P_{ij} of individual i choosing product j from a set of products is

$$\begin{aligned} P_{ij} &= \Pr\left[u_{ij} > \left\{u_{ij'}\right\}_{\forall j' \neq j}\right] \\ &= \Pr\left[v_{ij} + \varepsilon_{ij} > \left\{v_{ij'} + \varepsilon_{ij'}\right\}_{\forall j' \neq j}\right] \end{aligned} \tag{2}$$

Distributions for the ε Error Terms

The ε error terms are unobserved random variables that are described by a probability distribution. In general, this may be a joint distribution of all the error terms, so we use the vector $\boldsymbol{\varepsilon}_i = [\varepsilon_{i1}\ \varepsilon_{i2}\ ...\ \varepsilon_{in}]^T$, which aggregates the error terms for all products, and describe it's probability distribution by the cumulative distribution function (CDF) $F_{\varepsilon}(\boldsymbol{\varepsilon})$ and its corresponding probability density function (PDF) $f_{\varepsilon}(\boldsymbol{\varepsilon})$.

Let us examine a simple case where the choice set is composed of only two products, j and j', and we can generalize later. In this case

[1] For a basic introduction, see http://ingrimayne.saintjoe.edu/econ/LogicOfChoice/Overview7i.html

$$
\begin{aligned}
P_{ij} &= \Pr\left[v_{ij} + \varepsilon_{ij} > v_{ij'} + \varepsilon_{ij'} \right] \\
&= \Pr\left[\varepsilon_{ij'} < v_{ij} - v_{ij'} + \varepsilon_{ij} \right]
\end{aligned} \tag{3}
$$

For a given value of ε_{ij} Eq.(3) is $F_\varepsilon(\varepsilon_{ij}, v_{ij}–v_{ij'}+\varepsilon_{ij})$: the CDF of the joint random variable distribution evaluated at the point $(\varepsilon_{ij}, v_{ij}–v_{ij'}+\varepsilon_{ij})$, i.e., the probability that the random variable $\varepsilon_{ij'}$ is less than the value $(v_{ij}–v_{ij'}+\varepsilon_{ij})$, given ε_{ij}. However, ε_{ij} is a not deterministic fixed value, but instead is itself described by a probability density function $f_\varepsilon(\varepsilon_{ij})$. Therefore, the probability can be calculated by integrating over all values of ε_{ij}

$$
\begin{aligned}
P_{ij} &= \int_{\varepsilon_{ij}=-\infty}^{\infty} F_\varepsilon\left(\varepsilon_{ij}, v_{ij} - v_{ij'} + \varepsilon_{ij}\right) d\varepsilon_{ij} \\
&= \int_{\varepsilon_{ij}=-\infty}^{\infty} \int_{\varepsilon_{ij'}=-\infty}^{v_{ij}-v_{ij'}+\varepsilon_{ij}} f_\varepsilon\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) d\varepsilon_{ij'} d\varepsilon_{ij}
\end{aligned} \tag{4}
$$

In general, for a set of products

$$
\begin{aligned}
P_{ij} &= \Pr\left[v_{ij} + \varepsilon_{ij} > \left\{ v_{ij'} + \varepsilon_{ij'} \right\}_{\forall j' \neq j} \right] \\
&= \Pr\left[\left\{ \varepsilon_{ij'} < v_{ij} - v_{ij'} + \varepsilon_{ij} \right\}_{\forall j' \neq j} \right] \\
&= \int_{\varepsilon_{ij}=-\infty}^{\infty} \left(\int_{\varepsilon_{i1}=-\infty}^{v_{ij}-v_{i1}+\varepsilon_{ij}} \int_{\varepsilon_{i2}=-\infty}^{v_{ij}-v_{i2}+\varepsilon_{ij}} \cdots \int_{\varepsilon_{iJ}=-\infty}^{v_{ij}-v_{iJ}+\varepsilon_{ij}} f_\varepsilon\left(\boldsymbol{\varepsilon}\right) d\tilde{\boldsymbol{\varepsilon}}_{ij} \right) d\varepsilon_{ij} \\
&\text{where } d\tilde{\boldsymbol{\varepsilon}}_{ij} = d\varepsilon_{iJ} \ldots d\varepsilon_{i(j+1)} d\varepsilon_{i(j-1)} \ldots d\varepsilon_{i2} d\varepsilon_{i1}
\end{aligned} \tag{5}
$$

The Probit Model

Most commonly in statistics, unobserved random error terms are taken to be normally distributed (e.g., least squares, etc). The central limit theorem provides a theoretical justification for this choice in the absence of other information about distributional forms. If $f_\varepsilon(\varepsilon)$ in Eq. (5) is assumed to be a multivariate joint normal distribution with mean vector θ and covariance matrix Λ, this is called the *probit model*. The probit model allows for quite a general model; however, it does not yield a closed form solution and requires multidimensional integration.

Some econometricians have alternatively used a restricted form of the probit model where error terms are taken to be independently and identically distributed: i.e., the covariance matrix Λ is assumed to be diagonal. In this case, Eq.(5) reduces to a single dimensional integral:

$$
\begin{aligned}
P_{ij} &= \Pr\left[\left\{ \varepsilon_{ij'} < v_{ij} - v_{ij'} + \varepsilon_{ij} \right\}_{\forall j' \neq j} \right] \\
&= \prod_{j' \neq j} \Pr\left[\varepsilon_{ij'} < v_{ij} - v_{ij'} + \varepsilon_{ij} \right] \\
&= \int_{\varepsilon_{ij}=-\infty}^{\infty} \left(\prod_{j' \neq j} F_\varepsilon\left(v_{ij} - v_{ij'} + \varepsilon_{ij} \right) \right) dF_\varepsilon\left(\varepsilon_{ij} \right)
\end{aligned} \tag{6}
$$

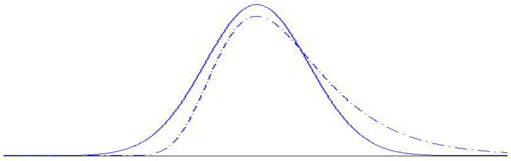
This simplified form is desirable; however, the assumption of independence of the error terms is a restriction that leads to specific implications, which we will discuss later.

The Logit Model

To simplify matters more, econometricians often use an alternative assumption for the distribution of the error terms: Instead of normal, error terms are assumed to be independently and identically distributed (iid) following the double exponential (Gumbel Type II extreme value) distribution:

$$
\begin{aligned}
F_\varepsilon\left(\varepsilon_{ij}\right) &= \exp\left(-e^{-\varepsilon_{ij}}\right) \\
f_\varepsilon\left(\varepsilon_{ij}\right) &= e^{-\varepsilon_{ij}} \cdot \exp\left(-e^{-\varepsilon_{ij}}\right)
\end{aligned} \tag{7}
$$

This assumption yields the *logit model*. Unlike the normal distribution, there is no theoretical reason to believe that the double exponential is a good assumption for the error terms; however, under this assumption P_{ij} in Eq.(5) reduces to a simple, explicit, usable form, and studies have shown that results obtained under this logit assumption are nearly indistinguishable from those produced by the probit model, except when large amounts of data are available. So, the logit assumption is a useful “engineering approximation”. The standard normal and double exponential PDFs are shown below:



Using Eq.(7) in Eq.(6), we have

$$P_{ij} = \int_{\varepsilon_{ij}=-\infty}^{\infty} e^{-\varepsilon_{ij}} \exp\left(e^{-\varepsilon_{ij}}\right) \prod_{j' \neq j} \exp\left(-e^{v_{ij}-v_{ij'}+\varepsilon_{ij}}\right) d\varepsilon_{ij} \quad (8)$$

since v_{ij}–v_{ij}=0, the exponential term can be brought inside the product, so that the expression is rewritten as

$$\begin{aligned} P_{ij} &= \int_{\varepsilon_{ij}=-\infty}^{\infty} e^{-\varepsilon_{ij}} \prod_{j'} \exp\left(-e^{-\left(v_{ij}-v_{ij'}+\varepsilon_{ij}\right)}\right) d\varepsilon_{ij} \\ &= \int_{\varepsilon_{ij}=-\infty}^{\infty} e^{-\varepsilon_{ij}} \exp\left(-\sum_{j'} e^{\left(v_{ij}-v_{ij'}+\varepsilon_{ij}\right)}\right) d\varepsilon_{ij} \qquad (9) \\ &= \int_{\varepsilon_{ij}=-\infty}^{\infty} e^{-\varepsilon_{ij}} \exp\left(-e^{-\varepsilon_{ij}} \sum_{j'} e^{v_{ij'}-v_{ij}}\right) d\varepsilon_{ij} \end{aligned}$$

We can solve this integral with a change of variables. Let t = exp(-ε_{ij}). Then dt = -exp(-ε_{ij})$d\varepsilon_{ij}$ and $d\varepsilon_{ij}$ = -dt/t. For the integration limits: as ε_{ij} approaches infinity, t approaches zero, and as ε_{ij} approaches negative infinity, t approaches infinity. Rewriting Eq.(9) in terms of t:

$$\begin{aligned} P_{ij} &= -\int_{t=\infty}^{0} \exp\left(-t\sum_{j'} e^{v_{ij'}-v_{ij}}\right) dt \\ &= \left(-\sum_{j'} e^{v_{ij'}-v_{ij}}\right)^{-1} \exp\left(-t\sum_{j'} e^{v_{ij'}-v_{ij}}\right)\Bigg|_{0}^{\infty} \qquad (10) \\ &= \left(\sum_{j'} e^{v_{ij'}-v_{ij}}\right)^{-1} = \left(e^{-v_{ij}} \sum_{j'} e^{v_{ij'}}\right)^{-1} \end{aligned}$$

$$P_{ij} = \frac{e^{v_{ij}}}{\sum_{j'} e^{v_{ij'}}} \qquad (11)$$

The iid double exponential error term assumption has led to a very simple formula for choice probabilities with appropriate properties: choice probabilities range from zero to one and sum to one over all alternatives in the choice set.

Independence of Irrelevant Alternatives

It is important to be aware that assuming independence of the error terms (in both the logit and the restricted probit models) gives rise to a property called *independence from irrelevant alternatives*, or IIA. We know that if a new alternative product is added to the choice set, some individuals who would otherwise have chosen a product in the initial choice set will instead choose the new product. The IIA property means the *ratio* of choice probabilities between any two alternatives is unaffected by the presence of a third alternative, and any new alternative introduced to a choice set will take its choice share proportionally from all other alternatives in the choice set. For the logit model, this is easy to show:

$$\frac{P_{i\mathrm{A}}}{P_{i\mathrm{B}}} = \frac{e^{v_{i\mathrm{A}}} \Big/ \sum_{j'} e^{v_{ij'}}}{e^{v_{iB}} \Big/ \sum_{j'} e^{v_{ij'}}} = \frac{e^{v_{i\mathrm{A}}}}{e^{v_{iB}}} \qquad (12)$$

The IIA property is also known as the “red bus, blue bus problem” because of a famous illustration of this property: Let’s say commuters have the two options {car, blue bus} available to them and gain equal utility from each (v_{CAR} = v_{BLUEBUS}), therefore choosing each with probability 0.5. If a new product is added to the choice set that is very similar to one of the existing products {car, blue bus, red bus} with equal utility, the IIA property implies that the new product will draw choice proportionally from all other alternatives, so that P_{CAR} = P_{BLUEBUS} = P_{REDBUS} = 0.333. In reality we would expect the red bus to draw far more commuters from the blue bus than from car travel since the two busses are very similar. Choice probabilities will likely be closer to P_{CAR} = 0.5, P_{BLUEBUS} = P_{REDBUS} = 0.25. The IIA also would imply, for instance, that the ratio of votes for Democratic and Republican candidates is unaffected by the presence of a third party candidate. Thus there are limitations to the applicability of models that possess the IIA property; however, a number of extensions exist to mitigate or eliminate this problem, and in many practical applications the IIA property is not problematic. For the remainder of this course we will use the simple logit model; however, interested students are welcome to research more advanced models.

Functional Forms for the Observable Component of Utility v

The preceding discussion presumes that the observable component of utility v_{ij} is known for each individual i and each product j. We said v_{ij} is observable in that it is a function of the observable characteristics of the product, the individual, and the purchase situation. For now, we will limit our discussion so that v_j depends only on the characteristics of the product, i.e., all individuals have the same *observable* component of utility, individual differences are described only by the random error term, and the index i is dropped. The term *product characteristics* is used specifically to describe objective, measurable aspects of the product that are observed by and relevant to the consumer during the choice process. For example, fuel economy of a vehicle may be considered a product characteristic, but "sportyness" is not a characteristic because it is perceived subjectively, and transmission ratio is probably not a characteristic since it is generally not observed directly by customers (except for special cases), but rather by engineering designers. The value of the product characteristics of product j are written as the real-valued vector $\mathbf{z}_j$, and v_j is a function of $\mathbf{z}_j$ as well as the product's price p_j, which, by convention, is not included in $\mathbf{z}_j$.

Just as in regression, we do not know, in general, the functional form relating $\mathbf{z}_j$ and p_j to v_j; however, if we have experience with choice models and experience in the problem domain, we may be able to posit reasonable functional relationships that produce good predictions. For example, researchers Boyd and Mellman (1977) proposed a functional relationship for vehicles including price p_j, gas mileage z_{j1}, and performance measured as time to accelerate from 0-60 mph z_{j2}, among other characteristics. Their model proposed that

$$v_j = \beta_0 p_j + \beta_1\left(\frac{1}{z_{j1}}\right) + \beta_2\left(\frac{1}{z_{j2}}\right), \tag{13}$$

where β_0, β_1, and β_2 are coefficients. If we could observe v_j directly, then we could collect data for various values of p_j and $\mathbf{z}_j$ and perform an ordinary regression to find the best values for the β coefficients; however, v_j is not observed: Only choice is observed. Given past data on choices among vehicles with various values for p_j and $\mathbf{z}_j$, it is possible to find values for the β coefficients that result in choice predictions that best match the observed choice data, as we would do in simple regression, using a technique called *maximum likelihood*.

Maximum Likelihood

In this case, we have 1) assumed the distribution of the error terms (double exponential for logit), and 2) assumed the functional form of v_j with respect to observed characteristics. Now we want to find the best model parameters (β coefficients) to match observed data, given the model form. To do this we search for the coefficients that maximize the likelihood that the choice model (with coefficients β) would generate the data we observed: i.e., the model predicts choices probabilistically, and we want to maximize the likelihood that choices predicted by the model would be exactly those observed. On a specific choice occasion, the probability of the model predicting the same choice as the one observed for individual i is

$$\prod_j P_{ij}^{\Phi_{ij}} , \tag{14}$$

where $\Phi_{ij} = 1$ if individual i chooses product j, and $\Phi_{ij} = 0$ otherwise. If this process is repeated for many individuals, the total number of individuals choosing product j is given by $\Sigma_i \Phi_{ij}$, and the probability of the model generating the observed choices is

$$\prod_j P_j^{\sum_i \Phi_{ij}} . \tag{15}$$

We are searching for the values of β that maximize this quantity. To simplify calculations and avoid numerical difficulties, it is common practice to maximize the log of the Eq.(15), which has the same maximum, rather than maximizing Eq.(15) directly. This is called the *log-likelihood*, often written *LL*. The maximum (log) likelihood β terms are therefore:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left(\sum_j \sum_i \Phi_{ij} \log P_j \right). \tag{16}$$

where P_{ij} is given by Eq.(11).

Example

Let's suppose our choice set consists of four vehicles with prices and characteristics shown below

	A	B	C	D
p_j (\$1000s)	15	15	20	20
z_{j1} (mpg)	25	35	25	35
z_{j2} (sec)	6	8	8	6

Suppose we ask 100 people which vehicle each would choose, and we find that 25 choose product A, 30 choose product B, 5 choose product C, and 40 choose product D. Using the logit model in Eq.(11) for choice probabilities P_j

and Eq.(13) as the form of the utility function v_j we would solve for the β terms as:

$$\underset{\beta_0,\beta_1,\beta_2}{\text{maximize}}\left(25\log P_\text{A}+30\log P_\text{B}+5\log P_\text{C}+40\log P_\text{D}\right)$$

$$\text{where } P_j=\frac{\exp\left(\beta_0 p_j+\beta_1\left(\frac{1}{z_{j1}}\right)+\beta_2\left(\frac{1}{z_{j2}}\right)\right)}{\sum_{j'}\exp\left(\beta_0 p_{j'}+\beta_1\left(\frac{1}{z_{j'1}}\right)+\beta_2\left(\frac{1}{z_{j'2}}\right)\right)} \tag{17}$$

.

To find the maximum by hand, we can take the gradient of the function, set it equal to zero, and solve the resulting system of equations. Alternatively, we can use an optimization algorithm such as Excel Solver to find the values for the β terms that maximize Eq.(17). Using either technique, the solution is β_0 = -0.132, β_1 = -99.0, β_2 = 22.8. We see that β_0 is negative, indicating that increasing price will decrease utility, β_1 is negative, indicating that increasing fuel economy (decreasing $1/z_{j1}$) will increase utility, and β_2 is positive, indicating that increasing 0-60 time (decreasing $1/z_{j2}$) will decrease utility. Note that five individuals chose product C, even though it is more expensive, has worse fuel economy, and worse performance. While this goes against the utility trends in a deterministic utility model, random utility choice models, such as the logit model, allow for unobserved characteristics that may affect the decisions of individuals while still capturing the overall trends.

Using these newly obtained beta values, and the corresponding model of choice, we can now make predictions about new products or changes to existing products. Suppose we wanted to lower the price of product C to attract more buyers. How much would we have to lower the price to double market share (attract 10 out of 100 buyers instead of 5)? To make this prediction, we would simply solve

$$P_\text{C}=\frac{\exp\left(\beta_0 p_\text{C}+\beta_1 z_{\text{C}1}+\beta_2 z_{\text{C}2}\right)}{\sum_{j'}\exp\left(\beta_0 p_{j'}+\beta_1 z_{j'1}+\beta_2 z_{j'2}\right)}=0.10\,. \tag{18}$$

for p_C using the beta values and characteristic values from above. In this case the answer is \$14,300. So, vehicle C, with the least desirable characteristics, would have to drop its price below the prices of competitors in order to capture 10% of the market.

Summary

We presented a method for modeling choices using utility theory, where each alternative in a set has a utility value to each individual u_{ij}, and individuals choose the alternative with highest utility. In random utility, u is composed of a deterministic, observable component v, and an unobserved stochastic error component ε. Standard assumptions about the error terms are that they follow a joint normal distribution, the probit model, or an i.i.d. double exponential distribution, the logit model. The logit model is often sufficient, and it is easier to work with; however, it is important to be aware of its limitations, such as the IIA property. The observable component of utility v is taken to be a function of the price p and characteristics $\mathbf{z}$ of the product. The form of this function is assumed, and the parameters (β coefficients) are estimated using maximum likelihood techniques on observed choice data. Once these coefficients have been found, the model can be used to predict choices in new situations, including new products or changes to existing products.

Two remaining questions will be addressed in the following lectures:

1) In our example we used an assumed functional form for v. In general, how does one know what functional form to use, and what kind of functional form for v should be assumed when there is no prior knowledge about the relationship between p, $\mathbf{z}$, and v.

2) In our example we used choice data for four hypothetical vehicles (A,B,C,D) to “fit” the model using maximum likelihood techniques. How were the characteristics of these vehicles chosen? Having too few data points would prevent formation of an accurate model, and the choice of alternatives in the original set could bias the model. In general, how might the choice of characteristics in the choice set affect the model, and how can a choice set be designed to collect data efficiently while minimizing bias?