Demand Modeling Using Discrete Choice Analysis – Part 2

Motivation

Previously, we presented a method for modeling choices using utility theory, where each alternative j in a set has a utility value u_{ij} to each individual i, 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 iid double exponential distributions, 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, including 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.

In the previous example we used an assumed functional form for v established by experts. In this document we will answer: 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, v, and $\mathbf{z}$?

Functional Forms for the Observable Component of Utility v

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 that provide information about probable choice. We have limited 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 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 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. Previously, we used a model developed by researchers Boyd and Mellman (1977) defining 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

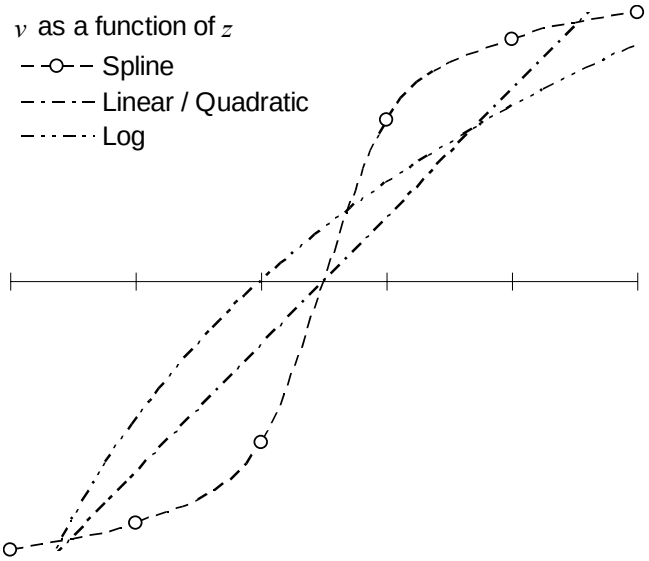
$$v_j = \beta_0 p_j + \beta_1 \left(\frac{1}{z_{j1}} \right) + \beta_2 \left(\frac{1}{z_{j2}} \right), \qquad (1)$$

where β_0, β_1, and β_2 are coefficients obtained using maximum likelihood techniques on observed choice data. However, in general we may not have good intuition about what functional forms to assume for a particular product and set of product characteristics. One method is to simply try different functional forms and see which one results in the highest likelihood value. However, this can be dangerous in the absence of information about the problem because more general forms (say, assuming a quadratic rather than linear relationship) will always yield higher likelihood than more restrictive forms; however, one must be wary of overfitting the data. So, in general, this might be a reasonable technique for testing whether the price relationship is linear or log, it is not a good idea to blindly test arbitrary functional form assumptions and pick the highest likelihood result.

Discretization of the Product Characteristics z and Price p

A more general technique is to divide the relevant range of each product characteristic in $\mathbf{z}$ and price p into discrete *levels*, capture the preference coefficients β at those discrete levels, and then interpolate for intermediate values. This allows the model to capture a wide variety of shapes with respect to the real-valued product characteristics $\mathbf{z}$ and price p. For example, the graph below shows a hypothetical case where the underlying relationship between v and a single product characteristic z is s-shaped. If we descretize z and obtain preference

estimates at the discrete levels (shown as circles), we can interpolate the s-shaped curve. However, if we assume that v is a linear, quadratic, or log function of z, then we obtain a more restrictive estimate that does not capture all of the detail.



This technique of discretizing and interpolating may not be feasible using data from the market, since we may not be able to describe existing market products in terms of a small number of discrete levels of each characteristic. However, if we are collecting choice data using a designed survey, it is feasible and often desirable.

First, we divide each product characteristic **z** into discrete levels that span the relevant domain of characteristic values. If the product characteristics are indexed by ζ, we divide each characteristic z_ζ into levels indexed by ω = {1, 2, 3, ..., Ω_ζ}. For example, characteristic ζ=1 is fuel economy, and if fuel economy z_1 ranges between say 10 mpg and 40 mpg, we might set levels at 10, 20, 30, and 40 mph, so that Ω_1=4, and ω = {1, 2, 3, 4} refers to {10mpg, 20mpg, 30mpg, 40mpg} respectively.

Each product in the choice set must be coded with respect to these characteristic levels using *dummy variables*. Here we notate the dummy variables as $\delta_{j\zeta\omega}$, where $\delta_{j\zeta\omega}$ = 1 if product characteristic ζ of product j is at level ω, and $\delta_{j\zeta\omega}$ = 0 otherwise. We also include price in this set, with price indexed as ζ=0. Thus, any product j with product characteristics and price at the discrete levels can be coded as a set of 1's and 0's in $\delta_{j\zeta\omega}$ $\forall\zeta,\omega$. Assuming that preferences are linear in the discretized set, we have

$$v_j = \sum_\zeta \sum_\omega \beta_{\zeta\omega}\delta_{j\zeta\omega}\,, \qquad (2)$$

where the coefficients $\beta_{\zeta\omega}$ are called *part-worths* because they describe the component of utility derived from characteristic ζ being at level ω. There may be cases where linearity of the characteristics cannot be assumed because of interaction effects, i.e., the shape of preferences for one characteristic may depend on the value of another characteristic. However, we leave these as advanced cases which we do not address here.

Using the logit model, the probability of an individual choosing product j is then:

$$P_j = \frac{\exp\left(v_j\right)}{\sum_{j'}\exp\left(v_{j'}\right)} = \frac{\exp\left(\sum_\zeta\sum_\omega \beta_{\zeta\omega}\delta_{j\zeta\omega}\right)}{\sum_{j'}\exp\left(\sum_\zeta\sum_\omega \beta_{\zeta\omega}\delta_{j\zeta\omega}\right)} \qquad (3)$$

and the log likelihood that a model with part-worth coefficients $\beta_{\zeta\omega}$ will reproduce the observed data Φ_{ij}, where Φ_{ij}=1 if individual i chooses product j, and Φ_{ij}=0 otherwise, is

$$LL = \sum_j \sum_i \Phi_{ij} \ln P_j \qquad (4)$$

as derived before, where P_j is given in Eq.(3). Given a set of observed choice data Φ_{ij} we can find the coefficients $\beta_{\zeta\omega}$ that maximize Eq.(4).

Example

In the vehicle example from the previous lesson, we had

	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

Where levels are defined as

ζ	symbol	level ω=1	level ω=2
0	p	\$15,000	\$20,000
1	z_1	25	35
2	z_2	6	8

The corresponding dummy variables $\delta_{j\zeta\omega}$ for these products are

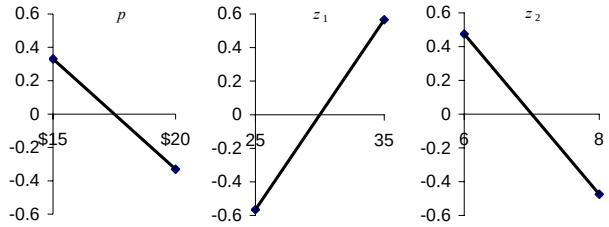
		j=A	j=B	j=C	j=D
ζ=0	ω=1	1	1	0	0
ζ=0	ω=2	0	0	1	1
ζ=1	ω=1	1	0	1	0
ζ=1	ω=2	0	1	0	1
ζ=2	ω=1	1	0	0	1
ζ=2	ω=2	0	1	1	0

As before, given this choice set suppose that 25 respondents choose product A, 30 choose product B, 5 choose product C, and 40 choose product D. If the log likelihood in Eq.(4) is maximized using Excel Solver, the resulting $\beta_{\zeta\omega}$ part-worths are:

$\beta_{\zeta\omega}$	ζ=0	ζ=1	ζ=2
ω=1	0.3304	-0.56544	0.47428
ω=2	-0.3304	0.56544	-0.47428

MODEL IDENTIFICATION: Actually, there are infinitely many sets of part worth coefficients that predict equivalent choice probabilities, and the results shown above are just one such set. This is because our model for v has extra degrees of freedom: i.e., there are more variables than equations in the system of equations. Any of the sets of betas that yield equivalent choice probabilities and log likelihood values are equivalent with respect to the choice model, and any can be used. If we wish to restrict the model to a single answer (this is called *model identification*), we can code Eq.(2) in terms of fewer variables ($1 + \Sigma_\zeta (\Omega_\zeta - 1)$ variables are needed), or we can add extra constraints to restrict the solution to a particular set of beta values from the infinite set of equivalent values for easier interpretation. The solution shown above is the particular beta solution maximizing Eq.(4) where the average β value of each characteristic ζ across all of its levels ω is zero.

The resulting beta values are plotted below for each characteristic and price. Each ζ is divided into only two levels, so we can use linear interpolation to estimate β values for intermediate levels, for example, a price of \$18,000.



By including only two levels per ζ, the resulting interpolation shown in the graphs is linear with respect to the real-valued characteristics, and we have essentially assumed a linear relationship. The final interpolated relationship for intermediate values of v, using linear interpolation, is

$$\hat{v}_j = -0.132\,p_j + 0.113 z_{1j} - 0.474 z_{2j} \qquad (5)$$

We see that the slope of the part worth for price (0.3304 – (-0.3304))/(\$20-\$15) = 0.132 is the same value we obtained in the previous lesson when we had assumed a linear functional form of p. The slopes of z_1 and z_2 are different than the previous lesson because here we have only two levels, which implies a linear relationship, whereas the functional form assumed previously was inversely proportional to each. So, using only two levels for each ζ is not recommended unless the modeler is relatively certain that the relationship is nearly linear, or that a linear representation will suffice. Use of more than two levels allows more general spline interpolation, and can represent more complex relationships.

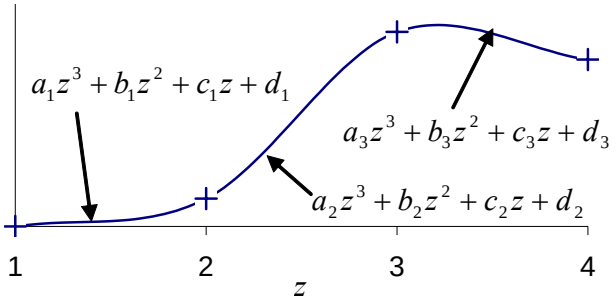
Interpolation of Part-Worth Coefficients Using Splines

In general, a spline can be fit through the part worth values $\beta_{\zeta\omega}$ of all levels ω in each ζ to interpolate intermediate values of ζ. It is possible to use many types of splines to interpolate the points; however, to facilitate optimization over the real-valued product characteristic values, it is desirable to interpolate using a spline function that is smooth and continuous over the domain. In particular, we will focus on natural cubic splines: a set of $(\Omega_\zeta - 1)$ cubic polynomials, each of which has a domain between two

adjacent levels ω (one between ω=1 and ω=2, another between ω=2 and ω=3, etc), that:

1. Match the value $\beta_{\zeta\omega}$ at each of the two domain endpoints ω,
2. Match the first and second derivatives of the adjacent cubic polynomial at each domain endpoint
3. Have second derivatives of zero at the extreme bounds of the spline: ω=1 and ω=Ω_ζ.

An illustration is shown below with Ω_ζ=4 four levels for hypothetical characteristic z:



It is possible to calculate the coefficients of the (Ω_ζ–1) cubic polynomials in a spline for characteristic ζ given $\beta_{\zeta\omega}$ by solving a system of equations representing the three conditions; however, we refrain from this detail here. Instead, software packages such as Excel or Matlab can be used to automatically calculate cubic splines given values for $\beta_{\zeta\omega}$. We will notate the cubic spline function for characteristic (or price) ζ that passes through the levels ω of $\beta_{\zeta\omega}$ as Ψ_ζ. The interpolated observable component of utility then involves the resulting spline function evaluated at the intermediate, real-valued product characteristics and price:

$$\hat{v}_j = \Psi_0\left(p_j\right) + \sum_{\zeta>0} \Psi_\zeta\left(z_{\zeta j}\right) \qquad (6)$$

This interpolated value of v can then be used in the logit model to predict the choice probabilities of new products with intermediate product characteristic and price values.

Example

Suppose that we had included more levels in our earlier example

ζ	symbol	level ω=1	level ω=2	level ω=3
0	p	$15,000	$20,000	$25,000
1	z_1	25	35	45
2	z_2	6	8	10

and the three separate choice sets below were provided to survey respondents, and their choices were recorded for each choice set.

Choice set		A	B	C	None
1	p_j ($1000s)	15	20	25	-
	z_{j1} (mpg)	25	35	45	-
	z_{j2} (sec)	6	10	6	-
2	p_j ($1000s)	15	20	25	-
	z_{j1} (mpg)	35	45	25	-
	z_{j2} (sec)	8	6	10	-
3	p_j ($1000s)	15	20	25	-
	z_{j1} (mpg)	45	25	35	-
	z_{j2} (sec)	10	8	6	-

Suppose 100 people were given this survey and the number of people choosing each option in each set is given by:

Choice set	A	B	C	None	Total
1	45	5	45	5	100
2	40	55	0	5	100
3	30	25	30	15	100

Given these data, the partworths (centered around zero for each characteristic, as before) can be calculated as

ζ	symbol	level ω=1	level ω=2	level ω=3
0	p	0.64	-0.03	-0.61
1	z_1	-0.67	-0.07	0.74
2	z_2	0.74	0.57	-1.32

with the no-choice option utility value of -1.829. Interpolating a spline through the levels of price and each characteristic would enable estimate of the part worth of an intermediate level.

HOMEWORK:

1. Show that the logit model is sensitive only to *relative* values of utility, not *absolute* values of utility. What, if anything, does it mean for a utility value in this model to be negative?
2. Using the model obtained in the original four-car example in Part 1, what is the predicted probability of choosing cars A, B, C, and D? How well do these predictions match the data?
3. Imagine that a fifth vehicle E was added to the mix with characteristics p_E = \$17,000, z_{j1} = 30 mpg, and z_{j2} = 7 seconds. What is the predicted probability of choosing cars A, B, C, D, and E?
4. Suppose that in the original vehicle example from part 1 we had observed that 30 people chose product A, 45 chose product B, 5 chose product C, and 25 chose product D. Estimate β_1, β_2, and β_3 using maximum likelihood. Compare these coefficients to those obtained in the example. What are the differences? What do these differences mean? Are these differences reflective of the data?