Chapter 13: Random Utility Models

Prerequisites: Sections 12.1 - 12.4

13.1 *Some Terminology and a Simple Example*

The subject of this chapter is a type of model known as a Random Utility Model, or RUM. RUMs are very widely applied marketing models, especially to the sales of frequently purchased consumer packaged goods; in other words; the kind of stuff you see in a supermarket. All of the models in this chapter logically follow from Thurstone's Law of Comparative Judgment that we covered in Chapter 12.   However, in this chapter we will consider the situation in which consumers pick one brand from a set of more than two brands, and we will also contemplate distributions other than the normal. We can summarize the assumptions of Thurstone's Law, and of the models in this chapter, as follows:

Assumption one is that choice is a *discrete event*. What this means is that choice is all-or-nothing. The consumer, as a rule, cannot leave the supermarket with .3432 cans of Coke and .6568 cans of Pepsi.   They will tend to leave with 1 can of their chosen brand, and 0 cans of their not chosen brand. Thus choice is not a continuous dependent variable.

Assumption two is that the attraction or utility towards a brand varies across individuals as a random variable. In Thurstone's Law, we called this the discriminal dispersion and we assumed it was normal. By using the term utility, we are being consistent with economic theory. We also fequently use the term *attraction*, we are being consistent with the retailing literature. In any case, assumption two is all about the word "random" in the label random utility model.

The last assumption is that the consumer chooses the brand with the highest utility. This makes our consumer an economically rational being. Thank goodness.

In general, we will be concentrating on the class of RUMs known as *logit models*. These are models that make a distributional assumption different than the normal and lead to much simpler calculations. In the next sections we will be introduced to the logit model in all its glory. But before that happens, here is a list of other important terms that will come into play:

Dichotomous dependent variable – any dependent variable capable of taking on exactly two discrete values.

Polytomous dependent variable – any dependent variable capable of taking on exactly $J > 2$ discrete values.

Income type independent variable – a variable that varies over consumers. A logit model incorporating only income type variables is sometimes be called a *polytomous logit model*.

Price type independent variable – a variable that varies over consumers and brands. A logit model with at least one of these is often called a *conditional logit model*. We might note here however, that there is no difference in the way we treat price and income variables if we are looking at a dichotomous dependent variable. The difference only comes into play when there are three or more possible choices.

Aggregate data – data that have been summarized for each unique combination of the independent variables. To keep things simple, let us say we have just one independent variable; coupon value; and that there are exactly four different values. For each coupon value, we might count up how

many people buy the product (that is, use the coupon) and how many do not. The choice probabilities for each of the four coupon values constitute the data analyzed as the dependent variable. We would obviously have four data points, each point being two numbers: the choice probability and the value of the coupon. We can estimate this sort of data using either Generalized Least Squares or Maximum Likelihood.

Disaggregate data – raw data consisting of individual choices. It is possible that each observation has a unique combination of values on the independent variables. Maybe there are hundreds of different coupon values and hundreds of different possible prices. Each data point might come from a single individual, with a one signifying that that person bought the product, and a zero signifying that that person did not buy. Disaggregate data can only be analyzed by ML.

We are going to start with a simple example involving retail choice. In the small southern city of Rome, Alabama, there is a hypothetical food store that carries hard to find Italian items. A sample of individuals was asked, "Do you shop at the Negozio?" We define the dependent variable such that

$$y_i = \begin{cases} 1 \text{ if Yes} \\ 0 \text{ if No} \end{cases} \tag{13.1}$$

for person i. We can also define x_i as the distance between person i's residence and the Negozio. Of course, we will need to also define e_i as a random, independent error. We could use the linear model of Chapter 5 to fit this model. In that case we would have

$$y_i = \beta_0 + x_i \beta_1 + e_i \tag{13.2}$$

$$\hat{y}_i \equiv E(y_i) = \beta_0 + x_i \beta_1 \ . \tag{13.3}$$

Now we are going to define the probability that individual i chooses (has chosen) the store and the complement of this probability. For the former, we will use the notation p_{i1} and for the latter p_{i2}. Given this notation, we can say that the predicted choice probabilities are

$$\hat{p}_{i1} = Pr[y_i = 1] = Pr[Yes] \quad \text{and} \tag{13.4}$$

$$\hat{p}_{i2} = Pr[y_i = 0] = Pr[No] \ . \tag{13.5}$$

It should be clear that $\hat{p}_{i2} = 1 - \hat{p}_{i1}$. It must also be the case, given the definition of what we mean by expectation that

$$E(y_i) = (1)\hat{p}_{i1} + (0)\hat{p}_{i2}$$

$$= \hat{p}_{i1} \ .$$

Combining this result with Equation (13.3), we conclude that

$$\hat{p}_{i1} = \beta_0 + x_i \beta_1 \ . \tag{13.6}$$

There are two problems with this conclusion. First, a choice probability, really; any probability; has to obey the rule

$$0 \le \hat{p}_{i1} \le 1 \tag{13.7}$$

but there is no requirement that ordinary least squares estimation will produce predicted values between 0 and 1. In other words, OLS may produce *logically inconsistent* choice probabilities. A second important feature of probabilities is that

$$\hat{p}_{i1} + \hat{p}_{i2} = 1 \tag{13.8}$$

but again, we are not guaranteed that regression will produce complementary probabilities that add up to 1. In other words, the predicted values are not *sum constrained*. There is also a third problem. With OLS regression we make the Guass-Markov assumption [Equation (5.16)] in order to perform hypothesis testing. Specifically, in regression we generally assume $e_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = \sigma^2$ for all i, that is; $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. But in the model we are now examining, there are exactly two possible values for e_i –

$$e_i = \begin{cases} 1 - (\beta_0 + x_i\beta_1) & \text{for } y_i = 1 \\ 0 - (\beta_0 + x_i\beta_1) & \text{for } y_i = 0. \end{cases} \tag{13.9}$$

By the definition of variance [see Equation (4.7)], we have

$$V(e_i) = E[e_i - E(e_i)]^2 \tag{13.10}$$

but since $E(e_i) = 0$, the expression above simplifies to $V(e_i) = E(e_i^2)$. And combining Equations (13.9) and (13.10) we see that

$$E(e_i^2) = \hat{p}_{i1}(1 - \beta_0 - x_i\beta_1)^2 + \hat{p}_{i2}(-\beta_0 - x_i\beta_1)^2 .$$

Note that since e_i is discrete, we use Equation (4.2) for its expectation. Combining the equation above with Equation (13.6) implies

$$E(e_i^2) = \hat{p}_{i1}(1 - \hat{p}_{i1})^2 + (1 - \hat{p}_{i1})(-\hat{p}_{i1})^2$$

$$= \hat{p}_{i1}(1 - \hat{p}_{i1}) \tag{13.11}$$

$$= (\beta_0 + x_i\beta_1)(1 - \beta_0 - x_i\beta_1) .$$

But now we have a problem. The formula for the variance of the error has the independent variable on the right hand side. What's more, that independent variable has the subscript i hanging off it. How can the variance of e_i be the same for all i when it depends on x_i? It cannot – we have heteroskedasticity of error variance. Our OLS parameter estimates might be unbiased and consistent, but they are not efficient. Standard errors and significance tests do not hold. Although by definition, OLS produces the smalleset sum of squared error that can be, we have now uncovered three problems with using it for choice data: logical inconsistency, the lack of the sum constraint, and heteroskedasticity. Some simply find it inelegant to use a procedure capable of predicting a probability of less than zero or more than one.

There are a number of ways to fix these problems. You could at least take care of the logical inconsistency by using the linear probability model. This model simply forces $\hat{y}_i$ to 0 and 1 whenever it shows up outside the range:

$$\hat{p}_{i1} = \begin{cases} 0 & \text{if } \beta_0 + x_i\beta_1 \leq 0 \\ \beta_0 + x_i\beta_1 & \text{if } 0 < \beta_0 + x_i\beta_1 < 1 \\ 1 & \text{if } \beta_0 + x_i\beta_1 \geq 1. \end{cases}$$

A second more theoretically grounded model is the *Probit model*. The probit model uses the same assumptions of the Thurstone model as presented in Chapter 12 namely that the utility of each of the choice options is normally distributed. In that case, we have

$$\hat{p}_{i1} = F_p(\beta_0 + x_i\beta_1) = \Phi(\beta_0 + x_i\beta_1)$$

(13.12)

$$= \int_{-\infty}^{\beta_0 + x_i\beta_1} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right] dz .$$

We could linearize the model by applying the PROBability Inverse Transform, or PROBIT transform (i.e. Φ^{-1}) and see the meaning of the name of this technique, as well as use unweighted least squares on the resulting z scores

$$\Phi^{-1}[\hat{p}_{i1}] = \hat{z}_{i1} = \beta_0 + x_i\beta_1 .$$

Unweighted least squares would not solve the third problem, namely heteroskedasticity. There are a variety of other estimation schemes for probit regression that would deal with this problem, but now we turn our attention to a very widely used model for choice data, the logit model. Note that in Equation (13.12) the appearance of the function F_p. Another version of this F function might be based on a transformation other than the normal or probit. This is illustrated below:

$$\hat{p}_{i1} = F_L(\beta_0 + x_i\beta_1) = \frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}} .$$

(13.13)

F_L is called the logistic function and so the model is sometimes called *logistic regression*. A visual representation is given below:

$$F_L[\beta_0 + x_i\beta_1]$$
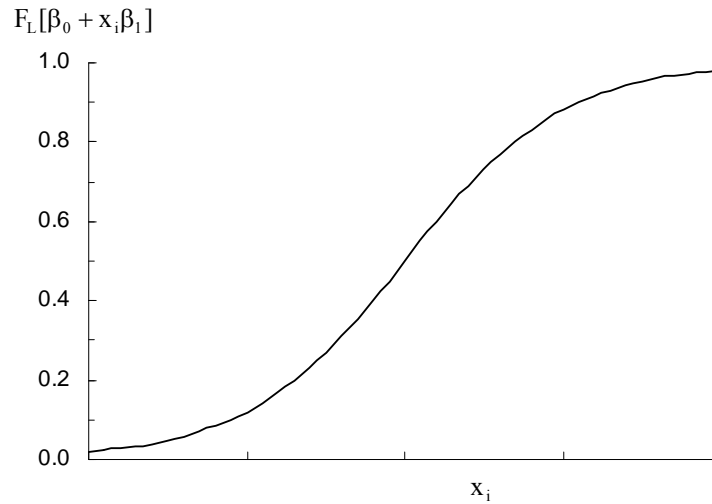
The logistic function is highly similar to the normal ogive. There are some important differences between it and the normal when there are more than two choice objects, but we will get to that topic later. For now, you might note that you can interpret the sign of the β in much the same way that you can in ordinary regression. A positive β implies that the choice probability goes up as x goes up. When dealing with this function, we can make the notation cleaner by defining $u_i = \beta_0 + x_i\beta_1$ so that

$$\hat{p}_{i1} = \frac{e^{u_i}}{1 + e^{u_i}} .$$

Now, multiply both sides by e^{-u_i} / e^{-u_i}

$$\hat{p}_{i1} = \frac{e^{u_i}}{1 + e^{u_i}} \cdot \frac{e^{-u_i}}{e^{-u_i}}$$

$$= \frac{1}{1 + e^{-u_i}}$$

which shows another way to write the model. In general, we will use the previous version,

$$\hat{p}_{i1} = \frac{e^{u_i}}{1 + e^{u_i}} . \tag{13.14}$$

As such, lets look at the probability that the respondent does not go to the store. That is

$$\hat{p}_{i2} = 1 - \hat{p}_{i1} = \frac{1+e^{u_i}}{1+e^{u_i}} - \frac{e^{u_i}}{1+e^{u_i}}$$

$$\hspace{6cm}(13.15)$$

$$= \frac{1}{1+e^{u_i}}.$$

Now look at the expression for $\hat{p}_{i1}$ in Equation (13.14) and for $\hat{p}_{i2}$ in Equation (13.15). In effect we have a/(a+b) for the one and b/(a+b) for the other, with 1 and e^{u_i} playing the roles of a and b. The logistic model is a special case of Bell, Keeney and Little's (1975) *Market Share Theorem* and what Kotler (1984) once called the *Fundamental Theorem of Market Share*. We can make this theorem more general by using the following notation:

$$\hat{p}_{ij} = \frac{a_{ij}}{\displaystyle\sum_m^J a_{im}}.$$

In our case, there are J = 2 brands, $a_{i1} = e^{u_i}$ and $a_{i2} = e^0 = 1$.

The logit model is not a linear model but it can be linearized. Repeating the model,

$$\hat{p}_{i1} = \frac{e^{u_i}}{1+e^{u_i}}$$

and multiplying both sides by $1/\hat{p}_{i2}$, the reciprocal of Equation (13.15), yields

$$\hat{p}_{i1}/\hat{p}_{i2} = \frac{e^{u_i}}{1+e^{u_i}} \cdot \frac{1+e^{u_i}}{1}$$

$$= e^{u_i}.$$

Now, we can take logs to get

$$\ln(\hat{p}_{i1}/\hat{p}_{i2}) = u_i = \beta_0 + x_i\beta_1.$$

The left hand side is called a *logit*. You can transform your choice probabilities into logits and fit a linear model using unweighted least squares. This, at least, solves both the logical consistency issue and the lack of sum constraint when OLS regression is applied to raw probabilities. It does not deal with the issue of efficiency, however. For that we will need to contemplate weighted least squares or maximum likelihood.

13.2 *Aggregate Data*

Imagine that we have a table of data, a table of different groups really. Our table might look like the one below, which shows N populations and the frequency of Yes's and No's within each population:

Population	Response Yes ($y_i = 1$)	No ($y_i = 0$)	$\mathbf{x}$
1	f_{11}	f_{12}	x_1
2	f_{21}	f_{22}	x_2
…	…	…	…
i	f_{i1}	f_{i2}	x_i
…	…	…	…
N	f_{N1}	f_{N2}	x_N

In the table, f_{i1} is the frequency with which members of group i say Yes, or simply put, the number of people living at distance x_i from the Negozio who go to that store. In what follows, it will be useful to define

$$n_i = f_{i1} + f_{i2}$$

$$p_{i1} = f_{i1} / n_i$$

$$\hat{\ell}_{i,12} = \ln(\hat{p}_{i1}/\hat{p}_{i2})$$

and analogously,

$$\ell_{i,12} = \ln(p_{i1}/p_{i2}) \,.$$

Of course, the distinction between $\ell_{i,12}$ and $\hat{\ell}_{i,12}$ is important. The first one is the observed logit and the second one is the logit as predicted by the model. Even when the model holds in the population studied, sampling error will see to it that they are not identical. To make an analogy to regression, we can say

$$\ell_{i,12} = \beta_0 + x_i\beta_1 + (\ell_{i,12} - \hat{\ell}_{i,12}) \ \,.$$

Without proof, let me claim that

$$E(\ell_{i,12}) = \hat{\ell}_{i,12} \tag{13.16}$$

and that

$$V(\ell_{i,12}) = V(\ell_{i,12} - \hat{\ell}_{i,12})$$

$$= \frac{1}{n_i\hat{p}_{i1}\hat{p}_{i2}} \,. \tag{13.17}$$

In summary, what this means is that in our model,

$$\ell_{i,12} = \beta_0 + x_i\beta_1 + (\ell_{12/i} - \hat{\ell}_{i,12}),$$

the error term in parentheses has

$$E(\ell_{i,12} - \hat{\ell}_{i,12}) = 0 \text{ and}$$

$$V(\ell_{i,12} - \hat{\ell}_{i,12}) = \frac{1}{n_i\hat{p}_{i1}\hat{p}_{i2}}.$$

What's more, it can be shown [this is related to but not the same as Equation (6.2)] as $n_i \to \infty$

$$\ell_{i,12} - \hat{\ell}_{i,12} \sim N\left[0, \ 1/n_i\hat{p}_{i1}\hat{p}_{i2}\right] \tag{13.18}$$

and in fact the approximation to the normal is already quite close by the time $n_i \geq 30$.

13.3 *Weighted Least Squares and Aggregate Data*

Under ordinary circumstances, $E(y_i - \beta_0 + x_i\beta)$ has the constant variance σ^2, and we minimize, as in Equation (5.21),

$$SS_{Error} = \sum_i (y_i - \beta_0 - x_i\beta_1)^2.$$

The residual in our case, that is the term in parentheses above, has variance $1/n_i\hat{p}_{i1}\hat{p}_{i2}$ which is clearly not a constant, since the subscript i appears in the term. In three places! We can, however, use this knowledge to stabilize the variance. We will create a set of weights consisting of the reciprocal of the variance of each observation. Specifically, we define

$$w_i = n_i\hat{p}_{i1}\hat{p}_{i2}$$

as the weights that we will use in the weighted least square (WLS) formula SS_{Error} formula below

$$SS_{Error} = \sum_i w_i(y_i - \beta_0 - x_i\beta_1)^2. \tag{13.19}$$

Here we might note that the weights serve to emphasize or de-emphasize the influence of a particular observation depending on its sampling variance. The higher the variance, the less influence the observation has in the determination of the SS_{Error}.

At this time we are going to shift into matrix notation so as to come up with a more general expression for WLS. Lets say that we have one independent variable, as before, consisting of travel distance to our shop in Rome, AL. Call that variable $x_{.1}$. A second independent variable might be the family income of each respondent, $x_{.2}$. Then

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \cdots & \cdots & \cdots \\ 1 & x_{i1} & x_{i2} \\ \cdots & \cdots & \cdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_{1\cdot} \\ \mathbf{x}'_{2\cdot} \\ \cdots \\ \mathbf{x}'_{i\cdot} \\ \cdots \\ \mathbf{x}'_{N\cdot} \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Note that the **X** matrix has N rows, with an upper case N used to maintain a distinction between the number of populations, and n_i, the number of observations within each population i. Now we can write our model as

$$\hat{p}_{i1} = \frac{e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2}}{1 + e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2}}$$

$$= \frac{e^{\mathbf{x}'_{i\cdot}\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'_{i\cdot}\boldsymbol{\beta}}}$$

(13.21)

or using the logit expression,

$$\hat{\ell}_{i,12} = \ln\frac{\hat{p}_{i1}}{\hat{p}_{i2}} = \mathbf{x}'_{i\cdot}\boldsymbol{\beta} .$$

(13.22)

This second way of expressing the model is convenient for estimation using the linear model. To do so, we begin by stacking the predicted and observed logits from each of the N populations into the vectors

$$\hat{\boldsymbol{\ell}}' = \begin{bmatrix} \hat{\ell}_{1,12} & \hat{\ell}_{2,12} & \cdots & \hat{\ell}_{N,12} \end{bmatrix}$$

$$\boldsymbol{\ell}' = \begin{bmatrix} \ell_{1,12} & \ell_{2,12} & \cdots & \ell_{N,12} \end{bmatrix}.$$

The model is then

$$\hat{\boldsymbol{\ell}} = \mathbf{X}\boldsymbol{\beta}$$

(13.23)

Now, we take the variances for each term $\ell_{i,12} - \hat{\ell}_{i,12}$ and place them into the covariance matrix $\mathbf{V}$ as diagonal elements:

$$\mathbf{V} = \begin{bmatrix} 1/n_1\hat{p}_{11}\hat{p}_{12} & 0 & \cdots & 0 \\ 0 & 1/n_2\hat{p}_{21}\hat{p}_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/n_N\hat{p}_{N1}\hat{p}_{N2} \end{bmatrix}.$$

Also note that we can relate the elements of this matrix to the previous scalar notation in Equation (13.19) because

$$\mathbf{V}^{-1} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_N \end{bmatrix}.$$

In matrix terms, our objective function is

$$f = (\ell - \hat{\ell})'\mathbf{V}^{-1}(\ell - \hat{\ell})$$

$$\hspace{4cm}(13.24)$$

$$= (\ell - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\ell - \mathbf{X}\boldsymbol{\beta}).$$

The f function, at its minimum, is distributed as χ^2 when the model holds in the population. Thus, it serves as a test of the null hypothesis that the model is correct. This is basically the same approach we used in Equation (12.19), with Minimum Pearson χ^2. If we were to replace the $\hat{p}$'s in the $\mathbf{V}$ matrix with p's, we would have Modified Minimum χ^2. When we set $\partial f/\partial\boldsymbol{\beta} = \mathbf{0}$ we find

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\ell$$

as the GLS parameter estimates. Since $V(\ell) = V(\ell - \hat{\ell}) = \mathbf{V}$ we further find that

$$V(\hat{\boldsymbol{\beta}}) = \left\{[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\right\}\mathbf{V}\left\{[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\right\}'$$

$$= [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}].$$

Following the same line of reasoning that we used in Section 6.8 (and also Section 17.4), we can use the above matrix for confidence intervals or to test hypotheses of the form

$$H_0: \beta_j = 0$$

or more generally

$$H_0: \mathbf{a}'\boldsymbol{\beta} - c = 0$$

and create the usual *t*-statistic with the denominator being formed by the scalar

$$\mathbf{a}' \, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{a}, \; \text{i.e.}$$

$$\hat{t} = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - c}{\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{a}}}.$$

For multiple degree of freedom hypotheses of the form

$$H_0: \mathbf{A}\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$$

we use

$$SS_H = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{A}]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}),$$

and for error,

$$SS_{Error} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

13.4 *Maximum Likelihood and Disaggregate Data*

With disaggregate data, we have household level observations. For the time being we return to the relatively simple case of a single independent variable, our distance measure from household i to the store. It is quite possible that each household has a unique value on this variable, especially if it is measured as a continuous variable. In addition, for each household we have a 1 if that household goes to the store and we have a 0 otherwise. Modifying our sample size notation once again, lets say we have N households altogether, with N_1 of them having said "Yes" and being scored with a '1' on the dependent variable, and N_2 of them having said "No." The model for the choice probability stays the same as before, we have just returned to the situation of a single independent variable for now,

$$\hat{p}_{i1} = \frac{e^{\beta_0 + x_{i1}\beta_1}}{1 + e^{\beta_0 + x_{i1}\beta_1}}$$

$$= \frac{e^{\mathbf{x}'_{i\cdot}\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'_{i\cdot}\boldsymbol{\beta}}},$$

but our objective function will be quite different. To begin creating the objective function, we might consider sorting the data into two piles: in the first pile we place the N_1 households saying "Yes" and in the second, the remainder who have said "No." We note that under the model, the probability of observing our N_1 Yes's and the rest of the sample with its No's, is

$$\ell_0 = \prod_{i=1}^{N_1} \hat{p}_{i1i} \prod_{i=N_1+1}^{N} \hat{p}_{i2} \qquad (13.25)$$

assuming that each observation is independent of all the others. This notation emphasizes the fact that there are two piles of observations: the first which consists of households going to the store, and the second consisting of those who do not frequent the place. Another way to write the likelihood is perhaps more clever, and relies on the fact that we have decided to score $y_i = 1$ if household i buys from the store and $y_i = 0$ if it does not. Rewriting ℓ_0 we have

$$\ell_0 = \prod_{i=1}^{N_1} (\hat{p}_{i1})^{y_i} \, (\hat{p}_{i2})^{1-y_i} \tag{13.26}$$

which takes advantage of the fact that for any value a, $a^1 = a$ while $a^0 = 1$. This second form avoids having to sort the observations. However, returning to Equation (13.25), we can flesh out the predicted choice probabilities. When we do that, the likelihood is seen as

$$\ell_0 = \prod_{i=1}^{N_1} \left(\frac{e^{\beta_0 + x_{i1}\beta_1}}{1 + e^{\beta_0 + x_{i1}\beta_1}} \right) \prod_{i=N_1+1}^{N} \left(\frac{1}{1 + e^{\beta_0 + x_{i1}\beta_1}} \right). \tag{13.27}$$

Since the maximum of the likelihood occurs at the same place as the maximum of the log likelihood, we will take logs and get

$$L_0 = \ln(\ell_0)$$

$$= \sum_{i=1}^{N_1} (\beta_0 + x_i\beta_1) - \sum_{i=1}^{N} \ln\left(1 + e^{\beta_0 + x_i\beta_1} \right). \tag{13.28}$$

To go from the previous expression for ℓ_0 in Equation (13.27) to the second line of Equation (13.28) for L_0 immediately above requires that you notice the denominator is identical for both $\hat{p}_{i1}$ and $\hat{p}_{i2}$, and that $\ln(1) = 0$. That explains why the first summation in Equation (13.28) goes to N_1 and the second goes all the way to N. We now wish to set

$$\frac{\partial L_0}{\partial \beta_0} = \frac{\partial L_0}{\partial \beta_1} = 0$$

and solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ using nonlinear optimization as is discussed in Section 3.9. To that end, the second order derivatives are quite useful. These provide additional information about the search direction. But what's more, they can be used to figure out the covariances and variances of the ML parameter estimates, which allows us to do hypothesis testing. For example, lets start with $\partial L_0 / \partial \beta_0$, and think of it as a function of the value of β_1. How does $\partial L_0 / \partial \beta_0$ change as β_1 changes? The limit of the slope of $\partial L_0 / \partial \beta_0$ (treated as a "dependent variable" in the calculus sense) on β_0 (treated as the "independent variable") is the second order derivative and it may be written

$$\frac{\partial}{\partial \beta_0} \left(\frac{\partial L_0}{\partial \beta_1} \right) \equiv \frac{\partial^2 L_0}{\partial \beta_0 \partial \beta_1} = \frac{\partial^2 L_0}{\partial \beta_1 \partial \beta_0} \,.$$

Think of this as element h_{12} and h_{21} in the symmetric **H** matrix, called *the Hessian*. Element 1, 1 is

$$h_{11} = \frac{\partial}{\partial \beta_0}\left(\frac{\partial L_0}{\partial \beta_0}\right) \equiv \frac{\partial^2 L_0}{\partial \beta_0 \partial \beta_0} = \frac{\partial L_0}{\partial \beta_0 \partial \beta_0}$$

and of course element 2, 2 would be defined analogously. Minus the expectation of the Hessian is called the *Information Matrix*, i. e. $-E(\mathbf{H})$. Finally, the inverse of the information matrix gives us the variance-covariance matrix of the unknowns, which is to say

$$V(\hat{\boldsymbol{\beta}}) = [-E(\mathbf{H})]^{-1}.$$

We can now test hypotheses using this matrix to provide the denominator of the *t*-statistic. Note that the Hessian is square and symmetric, and it will have one row (and one column) for each unknown parameter.

A final observation, before we start thinking about what happens if we have three choice options as opposed to only two, is that we can create an R^2 like statistic by comparing the log likelihood of the model, with the log likelihood of a model that consists only of β_0, that is, it has no real independent variables. This is illustrated below:

$$\rho^2 = 1 - \frac{L_0}{L_0^*}$$

where L_0^* is the likelihood under the model with just an intercept.

13.5 *Three or More Choice Options*

The situation with three or more brands, or three or more store choices, or Web links, etc., is rather more complicated than the two option case. Of course, we can say that

$$p_{i1} + p_{i2} + p_{i3} = 1$$

so at least we know something about the situation. However, there are now three potential logits: $\ln(p_{i1}/p_{i3})$, $\ln(p_{i2}/p_{i3})$ and $\ln(p_{i1}/p_{i2})$. But

$$\ln\frac{p_{i1}}{p_{i2}} = \ln\frac{p_{i1}}{p_{i3}} - \ln\frac{p_{i2}}{p_{i3}}$$

so one logit is redundant in the same sense that one of the three choice probabilities is not independent of the other two: if you know two of the probabilities you can figure out the third by subtracting the total of the other two from 1. With J brands, we will create $J - 1$ generalized logits. It is traditional to use the last brand, often a store or generic brand, as the denominator. The full model, called the *Multinomal Logit model* or MNL model is given below:

$$\hat{p}_{ij} = \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}_j}}{\sum_{m}^{J} e^{\mathbf{x}_{mi}'\boldsymbol{\beta}_m}} \tag{13.29}$$

where p_{ij} is the choice probability for brand j (j = 1, 2, $\cdots$, J) for case i. In this context i could either index populations, as would be the case with aggregate data, or individuals, which would be

the case with disaggregate data. The vector $\mathbf{x}'_{ij}$ provides values for the independent variables for brand j, observation i, while the vector $\boldsymbol{\beta}_j$ contains the unknown parameters for each independent variable for brand j. We can express the model as a special case of the Fundamental Theorem of Market Share,

$$\hat{p}_{ij} = \frac{a_{ij}}{\displaystyle\sum_m^J a_{im}} \quad \text{with}$$

$$a_{ij} = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}_j).$$

By tradition, we set the attraction for the last brand, brand J, equal to 1, i. e. $a_{iJ} = 1$ for all i, and thus $\boldsymbol{\beta}'_J = [0 \quad 0 \quad \cdots \quad 0]$. For ML estimation we pick elements of the $\boldsymbol{\beta}_r$ vectors to maximize

$$\ell_0 = \prod_{i=1}^{N_1} \hat{p}_{i1} \prod_{i=N_1+1}^{N_2} \hat{p}_{i2} \cdots \prod_{i=N_{J-1}+1}^{N} \hat{p}_{iJ}$$

where we have sorted the cases into J piles corresponding to the choice option picked by that individual.

13.6 *A Transportation Example of the MNL Model*

The following example is inspired by, but not identical to, an actual dataset reported in Currim (1985), who considered the choice faced by household i between getting to work by car (1), bus (2), or using the metro (3). Our explanatory variables are

I_i	Income of household i
C_i^j	Cost (price) of alternative j for household i
CAV_i	Cars per driver for household i
BTR_i	Bus transfers required for member of household i to get to work via the bus

One possible model for this situation might be

$$\ln \frac{\hat{p}_{i1}}{\hat{p}_{i3}} = \alpha_1 + I_i\beta_1 + (C_i^1 - C_i^3)\beta_3 + CAV_i\beta_4$$

$$\ln \frac{\hat{p}_{i2}}{\hat{p}_{i3}} = \alpha_2 + I_i\beta_2 + (C_i^2 - C_i^3)\beta_3 + BTR_i\beta_5.$$

Now we will have an opportunity to put into play some of the terminology we looked at in the beginning of the chapter but have not used up to now. Lets look at the role of income in this model. Income is constant across the choices that a family can make, but in the two logits, income has a different coefficient (β_1 and β_2). The quality of the choice option might vary, and so income may well contribute to families preferring choice 1 over choice 3, but it may lead to families preferring choice 3 over choice 2.

The price variable, C_i^j, varies across choice options as well as households. For one particular family, a car trip may be \$4.00 (including depreciation), a bus trip might be \$1.00 and a trip on the Metro could be \$1.50. But while C_i^j varies, the coefficient β_3 is constant. Such a variable is sometimes called *generic*. McFadden calls this sort of structure the *conditional logit* model. It is also known as the *simple effects model*.

The variables CAV_i and BTR_i only apply to one alternative. Thus they are called *Alternative Specific Variables* (ASVs). The α_j are also alternative specific variables. To be specific, they are alternative specific constants (ASCs). You might imagine a MNL model with only alternative specific constants . This would be quite similar to a Thurstone model, such as the Comparative Judgment model discussed in Section 12.3, only in this case we have a distribution other than the normal. In fact, in current context the ASCs function as a sort of error term. They represent the attraction towards the brand that exists independently of any measured assets such as its price, etc.

For GLS estimation it makes sense to create a single linear equation for the logits. That equation would look like this:

$$
\begin{bmatrix}
\hat{\ell}_{1,13} \\
\hat{\ell}_{2,13} \\
\cdots \\
\hat{\ell}_{N,13} \\
\hat{\ell}_{1,23} \\
\hat{\ell}_{2,23} \\
\cdots \\
\hat{\ell}_{N,23}
\end{bmatrix}
=
\begin{bmatrix}
1 & 0 & I_1 & 0 & C_1^1 - C_1^3 & CAV_1 & 0 \\
1 & 0 & I_2 & 0 & C_2^1 - C_2^3 & CAV_2 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
1 & 0 & I_N & 0 & C_N^1 - C_N^3 & CAV_N & 0 \\
0 & 1 & 0 & I_1 & C_1^2 - C_1^3 & 0 & BTR_1 \\
0 & 1 & 0 & I_2 & C_2^2 - C_2^3 & 0 & BTR_2 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 1 & 0 & I_N & C_N^2 - C_N^3 & 0 & BTR_N
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5
\end{bmatrix}
$$

$$\hat{\boldsymbol{\ell}} = \mathbf{X\beta} . \tag{13.30}$$

On the other hand, the best way to represent the model if we were going to do ML estimation is to show it as the nonlinear equation for the choice probabilities as

$$\hat{p}_{ij} = \frac{e^{\mathbf{x}_{ij}' \boldsymbol{\beta}_j}}{\sum_m^J e^{\mathbf{x}_{im}' \boldsymbol{\beta}_j}} . \tag{13.31}$$

13.7 *Other Choice Models*

There are a variety of related alternative forms for choice models, but for each model discussed in this section, and more generally in this chapter with the clear exception of the probit model of Section 13.10, we will be assuming the Fundamental Theorem,

$$\hat{p}_{ij} = \frac{a_{ij}}{\sum_m^J a_{im}} .$$

We will be assuming that we have k marketing instruments, meaning that we have a set of marketing variables perhaps including price, advertising effort, distribution effort, or some product attributes. The conditional or simple effects MNL is

$$a_{ij} = \exp(\sum_k x_{ijk}\beta_k) \ . \tag{13.32}$$

This model assumes that each marketing instrument has its own β coefficient, but each brand's marketing efforts have the same result for marketing instrument k. In other words, there is a β coefficient for each marketing instrument (price, place, etc.), but these are constant across the J brands.

Another type of simple effects model has been championed by Cooper and Nakanishi (1988). It is called the *Multiplicative Competitive Interaction* (MCI) model and looks like

$$a_{ij} = \prod_k x_{ijk}^{\beta_k} \ . \tag{13.33}$$

The MCI model follows in the footsteps of the economic Cobb-Douglas function of Equation (16.3), often used for demand equations for continuous dependent variables.

We can also have *differential effects* models in which the impact of each brands varies. Perhaps one of the brands is better than some of the others at leveraging its marketing efforts so it receives more benefit per dollar spent on advertising, to use that instrument as an example. There is a version of the differential effects model for the MNL,

$$a_{ij} = \exp(\sum_k x_{ijk}\beta_{jk}) \tag{13.34}$$

and for the MCI:

$$a_{ij} = \prod_k x_{ijk}^{\beta_{jk}} \ . \tag{13.35}$$

You will note that the beta coefficients have a subscript for the brand in the above two models. Finally, there is the *full extended model*. In the case of the MNL, this has been called the *universal logit model*:

$$a_{ij} = \exp(\sum_m^J \sum_k x_{imk}\beta_{mjk}) \ . \tag{13.36}$$

There is also a fully extended MCI model,

$$a_{ij} = \prod_m \prod_k x_{imk}^{\beta_{mjk}} \ . \tag{13.37}$$

These include asymmetric cross effects of one brand on another.

13.8 *Elasticities and the MNL Model*

How does our share change when we change the value of a marketing instrument? Lets assume that we have a model with only one marketing instrument; price. In line with Section 16.1, we define the *price elasticity of market share* for brand j as

$$e_{ij} = \frac{\partial \hat{p}_{ij}}{\partial x_{ij}} \cdot \frac{x_{ij}}{\hat{p}_{ij}}$$

for observation i. According to the generic or simple-effects model,

$$a_{ij} = \exp\left[\alpha_j + (x_{ij} - x_{iJ})\beta\right] \text{ or}$$

$$a_{ij} = \exp(\alpha_j + x_{ij}\beta).$$

In order to figure out the elasticity, we must start with the derivative,

$$\frac{\partial \hat{p}_{ij}}{\partial x_{ij}} = \frac{\partial}{\partial x_{ij}}\left[a_{ij}\left(\sum_m a_{im}\right)^{-1}\right].$$

In addition to the power rule and chain rule of the calculus (see Section 3.3), we need to note that

$$de^a/da = e^a$$

and

$$\frac{de^{f(x)}}{dx} = e^{f(x)} \cdot f'(x).$$

In that case

$$\frac{d\hat{p}_{ij}}{dx_{ij}} = -a_{ij}\left(\sum_m a_{im}\right)^{-2} a_{ij}\beta + a_{ij}\beta\left(\sum_m a_{im}\right)^{-1}$$

$$= -\hat{p}_{ij}^2\beta + \hat{p}_{ij}\beta$$

$$= \beta\hat{p}_{ij}(1 - \hat{p}_{ij}),$$

so that the elasticity is then

$$e_{ij} = \beta x_{ij}(1 - \hat{p}_{ij}). \tag{13.38}$$

Since x_{ij} appears in the expression for the elasticity, the elasticity is not constant and instead changes along the price-share curve. The elasticity is also inversely proportional to the share

which makes sense – the higher your share already, the harder it is to drive it closer to one by dropping prices even more.

For the simple effects MCI where $a_{ij} = x_{ij}^\beta$, we have

$$e_{ij} = \beta(1 - \hat{P}_{ij}).$$

In contrast to the MNL model, the MCI model produces constant elasticities much like the Cobb-Douglas function does for continuous dependent variables.

Marketing scientists are often interested in the cross elasticity for brand j with respect to some other brand j′. This quantity summarizes the extent to which the share of j depends on the prices set by the brand management of j′. This reveals the nature and amount of competition among the brands in the choice set. By definition, the price *cross elasticity of share* for brand j with respect to brand j′ is

$$e_{i,jj'} = \frac{\partial \hat{p}_{ij}}{\partial x_{ij'}} \cdot \frac{x_{ij'}}{\hat{p}_{ij}} .$$ (13.39)

The derivative for the simple effects MNL is

$$\frac{\partial \hat{p}_{ij}}{\partial x_{ij'}} = -\hat{p}_{ij}\hat{p}_{ij'}\beta$$

which makes the cross elasticity

$$e_{i,jj'} = -\hat{p}_{ij'}x_{ij'}\beta .$$ (13.40)

Since no j subscript appears on the right hand side, only j′, brand j′ has the same impact on all other brands. This impact is proportional to the share of j. For the differential effects model, i. e. $a_{ij} = \exp(\alpha_j + x_{ij}\beta_j)$,

$$e_{i,jj'} = -\hat{p}_{ij'}x_{ij'}\beta_{j'} ,$$

each brand exerts a different pressure, but that pressure is the same on all the other brands.

That brand j′ should exert the same pressure on all brands flies in the face of common sense. We often think that some brands compete more with certain other brands and less with others. This common sense notion is part of what is known as *Independence of Irrelevant Alternatives*.

13.9 *Independence of Irrelevant Alternatives*

Independence of Irrelevant Alternatives, or *IIA* as it is lovingly known, refers to the tendency of the Fundamental Theorem to model competition in a very symmetric way. We will now discuss the issue of asymmetric competition. Imagine that the transportation needs of a certain city are served by two companies: The Blue Bus Company and the Yellow Cab Company. Imagine further that these two companies split the market in half with each getting a market share of 50%.

What would happen if a new competitor arrives, namely, the Red Bus Company. The Fundamental Theorem would have us believe that the new market shares will be $1/3^{rd}$ each. Does this seem realistic to you?

The universal logit model can handle asymmetric competition. Technically speaking, however, it is actually not a RUM! The only other model in this chapter to be able to deal with the problem of IIA is presented next.

13.10 *The Polytomous Probit Model*

Again we will be concerned with the market share of brand j out of J different brands. The utility of each brand is normally distributed over the consumers in the market. Each individual picks the utility that is largest. We will define our model as

$$\mathbf{y} = \mathbf{Bx} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is the J by 1 random vector of utilities described above, $\mathbf{B}$ is J by k and $\mathbf{x}$ is k by 1. This model can include income or price type variables in $\mathbf{x}$. Their presence determines the appearance of $\mathbf{B}$ which, like in covariance structure models discussed in Chapter 10, or the ML MNL models discussed earlier in this chapter, can have zeroes in various positions. The random input vector can be characterized by noting that

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) .$$

The share for brand j is

$$\hat{p}_j = Pr[y_j > y_{j'}]$$

$$= Pr[y_j - y_{j'} > 0] \quad \text{for all } j' \neq j.$$

Now define $v_{j'}^{(j)} = y_j - y_{j'}$. Now we simply rewrite the expression for the share of brand j as

$$\hat{p}_j = Pr[v_{j'}^{(j)} > 0] \quad \text{for all } j' \neq j.$$

For the next step, we will place all of the $v_{j'}^{(j)}$ for each brand $j' \neq j$ into the vector $\mathbf{v}^{(j)}$. The action of subtracting all of the other brands from brand j is obviously a linear operation. We will illustrate this operation using brand 1 as our example. Define the J - 1 by J matrix

$$\mathbf{M}^{(1)} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

for brand j = 1. As we can see, this $\mathbf{M}$ matrix differences all of the other rows from the first row of any postmultiplying matrix. So in particular,

$$\mathbf{v}^{(1)} = \mathbf{M}^{(1)}\mathbf{y}$$

and in general for brand j

$$\mathbf{v}^{(j)} = \mathbf{M}^{(j)}\mathbf{y}.$$

Of course, Theorem (4.5) and Theorem (4.9) show us that

$$E[\mathbf{v}^{(j)}] \equiv \hat{\mathbf{v}}^{(j)} = \mathbf{M}^{(j)}\hat{\mathbf{y}} = \mathbf{M}^{(j)}\mathbf{B}\mathbf{x} \text{ and}$$

$$V[\mathbf{v}^{(j)}] \equiv \mathbf{\Sigma}^{(j)} = \mathbf{M}^{(j)}\mathbf{\Sigma}\mathbf{M}^{(j)'}.$$

Therefore according to the multivariate normal distribution, presented in Equation (4.17), the share for brand j is

$$\hat{P}_j = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}^{(j)}|^{1/2}} \int_0^\infty \int_0^\infty \cdots \int_0^\infty \exp\left[-(\mathbf{v}^{(j)} - \hat{\mathbf{v}}^{(j)})'\mathbf{\Sigma}^{(j)^{-1}}(\mathbf{v}^{(j)} - \hat{\mathbf{v}}^{(j)})/2\right]dv_{J-1}^{(j)} dv_{J-2}^{(j)} \cdots dv_1^{(j)}.$$

Share is equal to the probability that the utility for brand j exceeds the utility for all other brands, $j' \neq j$.

References

Cooper, Lee G and Masao Nakanishi (1988) *Market-Share Analysis*. Boston: Kluwer.

Kotler, Philip (1984) *Marketing Management Fifth Edition*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

McFadden, Daniel (1986) "The Choice Theory Approach to Market Research," *Marketing Science* 5 (Fall), 275-97.

Logit Models

Green, Paul E., Frank J. Carmone and David P. Wachspress (1977) "On the Analysis of Qualitative Data in Marketing Research," *Journal of Marketing Research*, 14 (February), 52-9.

Malhotra, Noresh K. (1984) "The Use of Linear Logit Models in Marketing Research," *Journal of Marketing Research,* 21 (February), 20-31.

Gensch, Dennis H. and Wilfred W. Recker (1979) "The Multinomial Multiattribute Logit Choice Model," *Journal of Marketing Research*, 16 (February), 124-32.

Flath, David and E.W. Leonard (1979) "A comparison of Two Logit Models in the Analysis of Qualitative Marketing Data." *Journal of Marketing Research* 16, (November), 533-8.

Bunche, David S. and Richard R. Batsell (1989) " A Monte Carlo Comparison of Estimators for the Multinomial Logit Model," *Journal of Marketing Research,* 26 (February), 56-68.

Probit Models

Malhotra, Naresh K. (1983) "A Threshold Model of Store Choice," *Journal of Retailing,* 59, (Summer), 3-21.

Hofacker, Charles F. (1990) "Derivation of Covariance Probit Elasticities," *Management Science,* 36 (April), 500-4.

Currim, Imran S., (1982) "Predictive Testing of Consumer Choice Models Not Subject to Independence of Irrelevant Alternatives, "*Journal of Marketing Research,* 21 (November), 20-31.

Kamakura, Wagner A. and Rajendra Srivastava (1984) "Predicting Choice Shares under Conditions of Brand Interdependence," *Journal of Marketing Research,* 21 (November), 20-31.

Nested Logit Models and Extensions of the MNL Mode

Louviere, Jordan J. and George Woodworth (1983) "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20 (November), 350-67.

Moore, William L. and Donald R. Lehman (1989) "A Paired Comparison Nested Logit Model of Individual Preference Structures," *Journal of Marketing Research¸ 26 (November), 420-8.*

Elasticities and Competition

Mahajan, Vijay, Paul E. Green and Stephen M. Goldberg (1982) " A Conjoint Model for Measuring Self-and Cross-Price/Demand Relationships," 19 (August), 334-42.

Gupta, Sunil (1988) "Impact of Sales Promotions on When, What and How Much to Buy," *Journal of Marketing Research*, 25 (November), 342-55.

Noncompensatory Choice

Manrai, Ajay K. and Prabhakant Sinha (1989) "Elimination by Cutoffs," *Marketing Science,* 8 (Spring), 133-52.

Rotondo, John (1986) "Price as an Aspect of Choice in EBA," *Marketing Science* , 5 (Fall), 391-402.

Johnson, Eric J. and Robert J. Meyer (1984) "Compensatory Choice Models of Noncompensatory Processes: The Effect of Varying Context," *Journal of Consumer Research* 11 (June), 528-41.

Grether, David and Louis Wilde (1984) "An Analysis of Conjunctive Choice: Theory and Experiments," *Journal of Consumer Research*, 10 (March) 373-85