Contents

1 Introduction 2

2 Model Formulation 3
  2.1 Individual choice 4
  2.2 The analyst’s problem 4
  2.3 Probabilistic choice 4

3 Model Specification 5
  3.1 Binary choice 6
    3.1.1 The binary probit model 6
    3.1.2 The binary logit model 7
  3.2 More than two choices 8
    3.2.1 The multinomial probit model 8
    3.2.2 The multinomial logit model 9
  3.3 Some specification issues 9
    3.3.1 Characteristics of individuals 9
    3.3.2 Choice-specific dummies 10

4 Estimation 10
  4.1 The data 11
  4.2 Maximum likelihood estimation 12
  4.3 Properties of maximum-likelihood 13
  4.4 The likelihoods 14
  4.5 Computer implementations 15
1 Introduction

These notes discuss models of mode choice, with particular attention to the logit model. An Appendix gives an overview of the basic ideas of hypothesis testing, using a simple example.

We shall be studying the determinants of an individual’s mode choice decisions. We shall resolve this into a study of the probability that an individual will make a given choice, and it is important to see at the outset how this arises. Suppose we observe the same individual making the same work-trip mode-choice decision on two successive days. Suppose that it is reasonable to assume that as far as we — the analysts — are concerned, the individual “looks the same” on both days: that is, any individual characteristics that we can measure are the same over the two days. Given this similarity, are we prepared to conclude that the individual must make the same mode-choice decision on both days? Of course not, and the reason
is clear: even if the characteristics that the analyst observes are the same, still, there may be determinants of choice which we do not see which may have changed, and which may affect the individual’s decision. For example on one day the individual’s daughter may be ill: we would not normally expect to know this, but it still may influence his or her modal choice.

The upshot is that because we have only incomplete information (the observable characteristics) we cannot say for certain which mode will be selected. We will need to settle for something less ambitious: the most we will be able to discuss is the probability that the individual will select one more or another (conditional on his/her observable characteristics). Of course, nothing depends on our observing the same individual on two successive days. We can look at any two individuals who appear the same as regards their measured characteristics (they are “observationally identical”). Because of unmeasured individual idiosyncracies, in this case too we will be able to study only the probabilities that they make a given choice. Thus the remainder of this note focusses on the development of models of probabilistic choice.

2 Model Formulation

2.1 Individual choice

We study the determinants of mode choice of an individual $i$. The set of available modes is called individual $i$’s choice set. In general we can assume that individual $i$ has $K_i$ elements in his or her choice set (each individual can have a different set of choices); but in this note we shall assume that it does not depend on $i$, that is, the number of elements in $i$’s choice set is the same $K$, for all individuals.

Individual $i$ bases his/her decision on a vector of modal characteristics $\tilde{x}_{ik} = \tilde{x}_{ik1}, \tilde{x}_{ik2}, \tilde{x}_{ik3}, \ldots$ where $\tilde{x}_{ikl}$ is the quantity of characteristic $l$ of mode $k$ as experienced by (or perceived by) individual $i$. We model choice as involving two steps:

1. Individual $i$ summarizes each mode’s characteristics into a single summary

---

1There’s an interesting unresolved problem here. Suppose that the in-vehicle characteristic of a trip takes 20 minutes, but that the individual believes (tells an interviewer, for example) that it takes 10 minutes. Which to use — actual time or perceived time? It’s not completely clear.
measure according to a function \( u_i \), called a “utility function”, and we refer to the summary value as the “utility” that \( i \) gets from mode \( k \). According to this scheme, individual \( i \) gets utility \( u_i(\tilde{x}_{ik}) = u_i(\tilde{x}_{ik1}, \tilde{x}_{ik2}, \tilde{x}_{ik3}, \ldots) \) from the total characteristics “package” of mode \( k \).

2. It can be shown that the utility function can be constructed to have the property that if \( u_i(\tilde{x}_{ik}) > u_i(\tilde{x}_{im}) \) then individual \( i \) prefers mode \( k \) to mode \( m \). (When this is true we say that the utility function represents \( i \)’s preferences).

We assume that \( i \) selects the most attractive mode. Thus mode \( k \) is selected if

\[
  u_i(\tilde{x}_{ik}) > u_i(\tilde{x}_{im}) \quad \text{for all } m = 1, 2, \ldots, K \text{ and } m \neq k
\]

In the two-mode setting (“binary choice”), mode 1 is selected if \( u_i(\tilde{x}_{i1}) > u_i(\tilde{x}_{i2}) \).

2.2 The analyst’s problem

We now switch our focus to the task of analyzing (understanding, rationalizing) the choices of individual \( i \). Some of the elements of \( \tilde{x}_{ik} \) are observable (to us); others are not. We suppose that the observable ones are the first \( L \) characteristics, and we represent them by \( x_{ik} = (x_{ik1}, x_{ik2}, x_{ik3}, \ldots x_{ikL}) \) (note: no tilde). The unobserved determinants are remaining components of \( \tilde{x}_{ik} \), and we represent their utility by \( \eta_{ik} \). The utility that individual \( i \) gets from mode \( k \) depends on the observed and the unobserved characteristics, and we write:

\[
  u_i(\tilde{x}_{ik}) = v_i(x_{ik}) + \eta_{ik}
\]

where \( v_i(x_{ik}) \) is the systematic or observable component of utility; and \( \eta_{ik} \) is the idiosyncratic or unobservable component.

2.3 Probabilistic choice

Now, \( \eta_{ik} \) is by definition unobservable to us, and in order to make progress we need to say something about it. (We clearly cannot simply ignore it, by the logic sketched in the Introduction). We therefore assume that each \( \eta_{ik} \) is a random variable with a known population distribution. This leads to the so-called “random utility” formulation: the utility that \( i \) gets from mode \( k \) can be viewed, from the perspective of the outside analyst, as the systematic component, \( v_i(x_{ik}) \) plus a drawing from
a random-number generator according to the distribution of $\eta_{ik}$. This implies that the (total) utility $u_i(\tilde{x}_{ik})$ is from our (the analyst’s) perspective, also a random variable.$^2$

Since $u_i(\tilde{x}_{ik})$ is a random variable, we cannot directly analyze an event like $u_i(\tilde{x}_{i1}) > u_i(\tilde{x}_{i2})$ — the event that mode 1 has greater utility than mode 2 for individual $i$, ie that mode 1 will be chosen — and as we saw in the introduction, this corresponds with our intuition about what can be learned in the presence of unobserved determinants of choice. The most we can do is analyze the probability of that event. This is the mode-choice probability, defined as

$$P_{ik} = \text{Prob[ individual } i \text{ selects mode } k] = \text{Prob}[u_i(\tilde{x}_{ik}) > u_i(\tilde{x}_{im}), \text{ for all } m \neq k] = \text{Prob}[v_i(x_{ik}) + \eta_{ik} > v_i(x_{im}) + \eta_{im}, \text{ for all } m \neq k] = \text{Prob}[v_i(x_{ik}) - v_i(x_{im}) > \eta_{im} - \eta_{ik}, \text{ for all } m \neq k] = \text{Prob}[\eta_{im} - \eta_{ik} < v_i(x_{ik}) - v_i(x_{im}), \text{ for all } m \neq k]$$

If we now write $v_{ik} = v_i(x_{ik})$, this is

$$P_{ik} = \text{Prob}[\eta_{im} - \eta_{ik} < v_{ik} - v_{im}, \text{ for all } m \neq k]$$

which we recognize as the value of the cumulative distribution of the random variable $\eta_{im} - \eta_{ik}$ evaluated at the “point” $v_{ik} - v_{im}$.

3 Model Specification

To obtain a concrete model of mode choice (ie to obtain the mode choice probabilities $P_{ik}$) we must specify:

1. How the systematic component of utility $v_{ik} = v_i(x_{ik})$ depends on the observable characteristics vector $x_{ik}$.
2. The distribution of the random variables $\eta_{ik}$.

$^2$Note that this is not to say that there is anything random about the individual’s decision-making. Quite the contrary. The individual sees all the characteristics of the modes; we introduce random variables only to deal with our ignorance/uncertainty, that is, the ignorance of the analyst trying to understand individual behavior.
Note that *any* assumptions here will lead to a concrete model. Nonetheless, there are several standard assumptions in the literature.

As regards the functional form of systematic utility, the usual assumption is it depends linearly on the characteristics, that is,

$$v_i(x_{ik}) = \beta_{ik1} x_{ik1} + \beta_{ik2} x_{ik2} + \cdots + \beta_{ikL} x_{ikL}$$

where we can think of $\beta'$ as a “weighting vector”, a way of weighting the importance of the characteristics $x_{ik}$. (Formally, both $x_{ik}$ and $\beta$ are a $L$-item *column vectors* so $\beta'$ — the transpose of $\beta$ — is an $L$-item *row vector*).

As regards the distribution of the random variables, it is convenient to consider separately the cases of two choices (binary choice) and more than two choices.

### 3.1 Binary choice

In this case we have

$$P_{i1} = \text{Prob}[\eta_{i2} - \eta_{i1} < v_i1 - v_i2]$$

$$P_{i2} = \text{Prob}[\eta_{i1} - \eta_{i2} < v_i2 - v_i1]$$

(and of course $P_{i2} = 1 - P_{i1}$). The literature discusses two important sets of assumptions concerning the $\eta_{ik}$.

#### 3.1.1 The binary probit model

Suppose the random variables $\eta_{i1}$ and $\eta_{i2}$ are independent and identically distributed (iid) random variables with the $N(0, \frac{1}{2})$ distribution. Then both $\eta_{i2} - \eta_{i1}$ and $\eta_{i1} - \eta_{i2}$ have the standard normal $N(0, 1)$ distribution, whose density (frequency) function we write as $\phi(t; 0, 1)$. So from the previous display, the mode-1 choice
probability is:

\[ P_{i1} = \int_{-\infty}^{v_{i1} - v_{i2}} \phi(t; 0, 1) \, dt \]

\[ = \int_{-\infty}^{\beta'(x_{i1} - x_{i2})} \phi(t; 0, 1) \, dt \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta'(x_{i1} - x_{i2})} e^{-t^2/2} \, dt \]

where the second display uses the linearity assumption, and the third just writes out the explicit form of \( \phi \). This is the *binary probit* model of mode choice.

In words, under the binary probit model, \( P_{i1} \) is the value of the cumulative distribution function of the standard normal distribution, evaluated at the point \( \beta'(x_{i1} - x_{i2}) \). This integral cannot be evaluated in “closed form”, but is readily available in tables or as built-in procedures in computer programs.

### 3.1.2 The binary logit model

Suppose the random variables \( \eta_{i1} \) and \( \eta_{i2} \) are iid random variables with the Type-1-Extreme-Value (T1EV, Weibull, Gnedenko, Gumbel) distribution, whose density function is

\[ f(\eta_{ik}) = e^{-\eta_{ik}} e^{-e^{\eta_{ik}}} . \]

A certain amount of manipulation shows that the mode-1 choice probability is:

\[ P_{i1} = \frac{e^{\eta_{i1}}}{e^{\eta_{i1}} + e^{\eta_{i2}}} \]

\[ = \frac{1}{1 + e^{\eta_{i2} - \eta_{i1}}} \]

\[ = \frac{1}{e^{\beta'(x_{i2} - x_{i1})}} \]

This is the *binary logit* model of mode choice. Note that unlike the probit model, the logit choice probabilities can be easily computed with any calculator with an exponential function built into it.
3.2 More than two choices

We can attempt to make the same kinds of assumptions here as we did in the previous case. There are $K$ modes, and the choice probability for mode 1 is

$$P_{i1} = \text{Prob}[\eta_{i2} - \eta_{i1} < v_{i1} - v_{i2} \& \eta_{i3} - \eta_{i1} < v_{i1} - v_{i3} \& \ldots]$$

and note that there are $K - 1$ separate inequalities in this expression: mode 1 is best if it is better than mode 2 and better than mode 3 and . . . .

3.2.1 The multinomial probit model

If we assume that the random variables $\eta_{im} - \eta_{ik}$ are jointly normally distributed with mean vector $\mathbf{0}$ and an identity covariance matrix $\mathbf{I}$ — this is the multivariate analog of the iid standard normal binary-choice assumption — then, writing the distribution as $\phi(t_2, t_3, \ldots t_K; 0, \mathbf{I})$ we obtain the multinomial probit model as the $K - 1$- dimensional integral

$$P_{i1} = \int_{-\infty}^{v_{i1}-v_{iK}} \int_{-\infty}^{v_{i1}-v_{i3}} \cdots \int_{-\infty}^{v_{i1}-v_{i2}} \phi(t_2, t_3, \ldots t_K; 0, \mathbf{I}) \, dt_2 \, dt_3 \ldots \, dt_K$$

(This model can be generalized to the case of dependent normal random variables, which changes the assumption on the covariance matrix).

The important thing about this equation is that for $K > 3$, it is extremely difficult to compute rapidly and accurately. There are some recent developments here: it turns out that it is possible to use Monte Carlo methods to simulate integrals like this, and with the vastly improved speed of personal computers this is now becoming a real possibility. Still, because these developments are still being assimilated into the literature, one does not see many applications of the multinomial probit model.
3.2.2 The multinomial logit model

If we assume that the random variables \( \eta_{ik} \) are iid T1EV then we get the multinomial logit model whose choice probabilities are

\[
P_{ik} = \frac{e^{\nu_{ik}}}{\sum_m e^{\nu_{im}}} = \frac{e^{\beta x_{ik}}}{\sum_m e^{\beta x_{im}}}
\]

Note that this probability is very easy to compute (once we know \( \beta \) and the \( x_{km} \)). This explains why the logit model is the one most often encountered in real-world applications.

3.3 Some specification issues

A specification of a probabilistic choice model involves making assumptions on the distribution of the random variables \( \eta_{ik} \), on the functional form of the systematic component of utility (\( v_i \)), and finally deciding which characteristics are to be considered as relevant. We’ve said so far that these are characteristics of the modes; but there are other possibilities.

3.3.1 Characteristics of individuals

There is no reason why we cannot include characteristics of individuals in the \( x \)'s: indeed, since we may well believe that different types of individuals may have different preferences (eg, men’s preferences differ from women’s) over the modes, it is almost mandatory to do so. The only difficulty is how to represent them, since for any individual \( i \), personal or individual characteristics will not differ by mode. To see what happens, suppose we write

\[ v_{ik} = \beta' x_{ik} + \gamma h_i \]

where \( h_i \) is some characteristic of individual \( i \) (eg age) independent of mode (\( k \)). Suppose the binary logit model holds, and consider the probability of selecting
mode 1. By previous results this is

\[ P_{11} = \frac{e^{\eta_{11} + \eta_{12} h_i}}{e^{\eta_{11} + \eta_{12}}} = \frac{e^{\beta' x_{11} + \gamma h_i}}{e^{\beta' x_{11} + \gamma h_i}} = \frac{e^{\gamma h_i (e^{\beta' x_{11}} + e^{\beta' x_{12}})}}{e^{\beta' x_{11} + e^{\beta' x_{12}}}} \]

Note that the \( \eta h_i \) term has dropped out: this means that we cannot estimate \( \gamma \). (An even simpler derivation is to note that the choice probability depends on \( v_{12} - v_{11} \), and when we compute this, the term \( \gamma h_i \) drops out). The solution usually adopted here is to “attach” the characteristics to one (or more, but not all) of the modes: that is we construct a variable like \( h_{ik} \) which takes the value “age of individual i” if \( k = \) mode 1 (say) and zero otherwise. If we do this we can consider these individual characteristics just like any \( x_{ik} \).

### 3.3.2 Choice-specific dummies

Suppose we were to assume, by analogy with linear regression, that each mode’s systematic utility \( v_{ik} \) includes a constant, that is

\[ v_{ik} = \beta' x_{ik} + \gamma_k \]

\( \gamma_k \) can be constructed as the coefficient of a variable \( h_{ik} \) which takes on the value 1 for the mode in question, and is zero otherwise. For this reason, it is often called a choice- (or mode-) specific dummy variable. There is only one subtlety to watch out for. Suppose your sample faces \( K \)-dimensional choice sets. Then you can include in your specification at most \( K - 1 \) dummy variables; otherwise you induce a linear dependency in the data, and entire model becomes inestimable. Apart from that, the dummy variables can are just like any other characteristics \( x_{ik} \).

### 4 Estimation

Let’s take stock. To arrive at a concrete mode-choice model we assumed a specific distribution for the random variables; a form of the systematic component of utility
which we took to be linear in the modal characteristics; and finally we decided which characteristics (modal, or individual or dummy) would be included in the specification. We also assumed that we could observe the $x_{km}$. What’s left? The only unknown is the weighting vector $\beta$, and our task is therefore to use data to estimate this.\footnote{Depending on the choice of distribution for the $x_{ik}$ there may also be some distributional parameters (e.g., mean or variance) to estimate; the following discussion applies equally well to those.}

### 4.1 The data

We now describe that data in more detail. We assume that we have a random sample of $I$ individuals. For each individual $i$ we know the complete $L$-vector of the $x_{ik}$. Note that, according to our previous discussion, some of what we are calling (modal) characteristics could be individual characteristics and/or mode-specific dummies. In addition, we assume that we also know the actual choice that each individual made — which mode was chosen. We represent the choices by an indicator variable defined as

$$y_{ik} = \begin{cases} 
1 & \text{if individual } i \text{ selects mode } k \\
0 & \text{otherwise}
\end{cases}$$

Note that for each individual $i$, exactly one of the $y_{ik}$ will be 1 (this corresponds to the $k$ actually chosen) and all the rest will be 0.

We can now collect all our data together into the following dataset representation:
<table>
<thead>
<tr>
<th>Individual $i$</th>
<th>Mode $k$</th>
<th>Char. 1</th>
<th>Char. 2</th>
<th>Char. 3</th>
<th>⋯</th>
<th>Char. $L$</th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$x_{111}$</td>
<td>$x_{112}$</td>
<td>$x_{113}$</td>
<td>⋯</td>
<td>$x_{11L}$</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$x_{121}$</td>
<td>$x_{122}$</td>
<td>$x_{123}$</td>
<td>⋯</td>
<td>$x_{12L}$</td>
<td>$y_{12}$</td>
</tr>
<tr>
<td></td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>$x_{1K1}$</td>
<td>$x_{1K2}$</td>
<td>$x_{1K3}$</td>
<td>⋯</td>
<td>$x_{1KL}$</td>
<td>$y_{1K}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$x_{211}$</td>
<td>$x_{212}$</td>
<td>$x_{213}$</td>
<td>⋯</td>
<td>$x_{21L}$</td>
<td>$y_{21}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$x_{221}$</td>
<td>$x_{222}$</td>
<td>$x_{223}$</td>
<td>⋯</td>
<td>$x_{22L}$</td>
<td>$y_{22}$</td>
</tr>
<tr>
<td></td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>$x_{2K1}$</td>
<td>$x_{2K2}$</td>
<td>$x_{2K3}$</td>
<td>⋯</td>
<td>$x_{2KL}$</td>
<td>$y_{2K}$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>$x_{311}$</td>
<td>$x_{312}$</td>
<td>$x_{313}$</td>
<td>⋯</td>
<td>$x_{31L}$</td>
<td>$y_{31}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$x_{321}$</td>
<td>$x_{322}$</td>
<td>$x_{323}$</td>
<td>⋯</td>
<td>$x_{32L}$</td>
<td>$y_{32}$</td>
</tr>
<tr>
<td></td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>$x_{3K1}$</td>
<td>$x_{3K2}$</td>
<td>$x_{3K3}$</td>
<td>⋯</td>
<td>$x_{3KL}$</td>
<td>$y_{3K}$</td>
</tr>
</tbody>
</table>

Note that the symmetry here is due to the fact that each individual has the same $K$-element choice set, so that each “individual block” has the same number of rows. If choice sets differed, then the blocks would have different numbers of rows.

4.2 Maximum likelihood estimation

Our task at this point is to arrive at an estimate $\hat{\beta}$ of the unknown weighting vector $\beta$. Here’s one strategy for doing it.

Suppose we pick a value for $\hat{\beta}$ (note that this is to pick $L$ individual numbers since $\hat{\beta}$ is a $L$-vector). Given our assumptions (linearity of utility and a joint distribution for the $\eta_{ik}$) and the availability of the data (the $x_{ik}$ and $y_{ik}$), we can now calculate all the choice probabilities. In particular, if individual $i$ is observed to select mode $m$ (ie, $y_{im} = 1$), we can compute $P_{im}(\hat{\beta})$ the probability (with this particular $\hat{\beta}$), that $i$ made the choice he/she actually did make. We can do this for each individual in the sample, and we can then compute the probability, or sample likelihood, that (with this particular $\hat{\beta}$) everyone in the sample made the choices they are actually observed to make.

Now suppose that the sample likelihood is small. There are two ways to inter-
pret this. One is that the observed choices are the result of something very unlikely having happened. (After all, it’s possible to throw ten 6’s in a row with unloaded dice — it’s just very unlikely). But when we remember that the low probability was the result of our (arbitrary) pick of the values for $\hat{\beta}$, an alternative explanation seems more reasonable — namely that our pick was wrong.

This immediately suggests an estimation strategy: take as our estimate $\hat{\beta}$ that value which makes the real-world outcome — the choices that people are actually observed to have made — as likely as possible. This is the principle of maximum likelihood estimation, a standard approach to estimating $\beta$. The value $\hat{\beta}$ which maximizes the probability of observing the actual outcome is called the maximum-likelihood estimator (MLE) of $\beta$.

4.3 Properties of maximum-likelihood

Assuming we can implement this strategy, does it have anything to recommend it? In particular how is the MLE $\hat{\beta}$ related to the unknown (true) vector $\beta$, and how is the MLE related to other possible ways of estimating $\beta$? One can show that in large samples (strictly, as the sample size tends to infinity), the MLE has the following properties:

1. Although the MLE may be biased in small (finite) samples (that is, $E(\hat{\beta})$ may not equal $\beta$), the bias disappears as the sample size gets large, and in fact the density of $\hat{\beta}$ converges to a spike at the true (but unknown) value $\beta$. (In the statistics jargon: $\hat{\beta}$ is consistent for $\beta$).

2. In a large class of possible ways to estimate $\beta$, the MLE has the lowest variance (jargon: $\hat{\beta}$ attains the Cramér-Rao lower-bound on the variances of these possible estimators).

In this sense, the MLE is an “optimal” or “best” estimator of the unknown weighting vector $\beta$. In addition, the distribution of the MLE converges to a normal distribution (jargon: $\hat{\beta}$ is asymptotically normal) so that the usual tests of hypotheses which are based on the normal distribution are (asymptotically) valid.
4.4 The likelihoods

In order to implement the maximum likelihood idea we need to obtain the probabilities of observing the choices actually made by the individuals in our sample, as a function of the unknown weighting vector $\beta$.

Focus now on individual $i$. We know that this person made (say) choice $m$. What is the probability of this? Obviously, nothing more than $P_{im}$. Consider another individual $j$: Suppose his/her choice is mode $n$. Then the probability of observing this is just $P_{jn}$ and by independence the probability of observing both that $i$ chose $m$ and that $j$ chose $n$ is $P_{im}P_{jn}$. And so on for all the individuals in our sample.

Very easy. The only practical difficulty is computational. We needed to identify which choice each individual actually made before writing down the product of the probabilities. Can we avoid that, or figure out a way for a computer to do it for us automatically? Consider the expression:

$$L_i(\beta) = P_{i1}^{y_{i1}} P_{i2}^{y_{i2}} P_{i3}^{y_{i3}} \cdots P_{iK}^{y_{iK}}$$

for individual $i$. What’s going on here? Well, if alternative $m$ was not selected, then $y_{im} = 0$ and the term $P_{im}^{y_{im}}$ is just $P_{im}^0 = 1$, and it adds nothing to the product. On the other hand, if alternative $m$ was selected, $P_{im}^{y_{im}} = P_{im}^1 = P_{im}$. In other words, the expression $P_{i1}^{y_{i1}} P_{i2}^{y_{i2}} P_{i3}^{y_{i3}} \cdots P_{iK}^{y_{iK}}$ is just another way of writing down the probability (depending on $\beta$) that individual $i$ chose the alternative that he/she actually selected.

The upshot is that, because of random sampling, the sample likelihood is the product of the individual likelihoods, or

$$L(\beta) = L_1(\beta) L_2(\beta) \cdots L_I(\beta)$$

$$= \prod_{i=1}^I P_{i1}^{y_{i1}} P_{i2}^{y_{i2}} P_{i3}^{y_{i3}} \cdots P_{iK}^{y_{iK}}$$

$$= \prod_{i=1}^I \prod_{k=1}^K P_{ik}^{y_{ik}}$$

As a practical matter, it is often computationally easier to maximize, not $L(\beta)$ but rather its logarithm, the log-likelihood function, which, recalling that taking
logs converts products to sums, is

\[ \ln L(\beta) = \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ik} \ln P_{ik} \]

### 4.5 Computer implementations

There are many computer programs which will do the maximum-likelihood estimation for you, given the data. One of the better ones (in my view) is Bill Greene’s LIMDEP program, which is available on a site-licence basis for students in both City Planning and Civil Engineering. I have produced an online introduction to LIMDEP, available at

??

Note that in LIMDEP, the procedure corresponding to the logit mode-choice model called the Discrete Choice model. LIMDEP also contains a procedure to estimate what it calls the “logit” model, but that is a bit different from the model we’re discussing here.

### 5 Prediction

Once we have obtained our estimate \( \hat{\beta} \) of the unknown parameter vector \( \beta \), we can predict the mode choice probabilities for any individual \( i \) by plugging in to the functional form of the choice probability equation. For example, if we are using a logit model, then we can compute the predicted choice probabilities for individual \( i \) as

\[
\hat{P}_{im} = \frac{e^{\hat{\beta}_1 x_{im1} + \hat{\beta}_2 x_{im2} + \cdots + \hat{\beta}_L x_{imL}}}{\sum_k e^{\hat{\beta}_1 x_{ik1} + \hat{\beta}_2 x_{ik2} + \cdots + \hat{\beta}_L x_{ikL}}}
\]

\[
= \frac{e^{\hat{\beta}_{x_{im}}}}{\sum_k e^{\hat{\beta}_{x_{ik}}}}
\]

---

4The logarithm is an increasing function, so whatever maximizes the log likelihood also maximizes the likelihood itself.
that is, we just replace $\beta$ in the formula for the logit choice probabilities by its maximum-likelihood estimator $\hat{\beta}$.

6 Goodness of fit

How well have we done in explaining or understanding the choices made by the individuals in our sample? There are several ways to address this.

6.1 Percent correctly predicted

Suppose individual $i$ chose mode $m$. Then $y_{im} = 1$, and the rest of the $y_{ik} = 0$. Now pick out the maximum of the computed $\hat{P}_{ik}$; suppose it is the $n$-th. In other words, mode $n$ is the mode which has the highest computed choice probability, and is, in a sense, the mode that we predict that individual $i$ will select. Define a “prediction success indicator” for individual $i$ as

\[ s_i = \begin{cases} 1 & \text{if } y_{in} = 1 \\ 0 & \text{otherwise} \end{cases} \]

That is, $s_i = 1$ if individual $i$ actually chose the mode that our model predicts will be chosen (the mode with the highest predicted choice probability); and it will be zero otherwise. Recalling that we have a sample of $I$ individuals, we define

\[ \text{Percent Correctly Predicted} = \frac{\sum_i s_i}{I} \]

6.2 Likelihood ratio index

A second approach asks how well we’ve done relative to a situation in which we know nothing about $\beta$. We can calculate the maximized value of the log likelihood function at $\hat{\beta}$ for individual $i$ as

\[ \ln \mathcal{L}_i(\hat{\beta}) = \sum_k y_{ik} \ln \hat{P}_{ik} \]
and similarly for the whole sample, giving the log likelihood value at convergence, \( \ln \mathcal{L}(\hat{\beta}) \).\(^5\) To represent the "no information" case, the standard approach is to look at the value of the log likelihood function when all the elements of \( \hat{\beta} \) equal zero. If we are using the logit model, then

\[
\hat{P}_{ik}(\hat{\beta} = 0) = \frac{e^0}{\sum_k e^0} = 1/K
\]

where \( K \) is the number of modes; the individual likelihood is also \( 1 / K \) also (since we will be raising just one of these to the power 1, and the rest to the power 0) and the sample log likelihood is \( \ln \mathcal{L}(0) = I \ln (1 / K) = -I \ln K \). Then we define the likelihood ratio index, LRI, or McFadden's \( \rho^2 \), as the percentage improvement at convergence over the no-information situation, or

\[
\text{LRI} = \rho^2 = \frac{\ln \mathcal{L}(0) - \ln \mathcal{L}(\hat{\beta})}{\ln \mathcal{L}(0)} = 1 - \frac{\ln \mathcal{L}(\hat{\beta})}{\ln \mathcal{L}(0)}
\]

This goodness-of-fit statistic is analogous to the \( R^2 \) of linear regression analysis, except that values of \( \rho^2 \) around .3 or better are considered as acceptable (even, good), which they would often not be in the \( R^2 \) case.

### 7 Hypothesis testing

We now want to use the computed MLE \( \hat{\beta} \) to answer questions about \( \beta \). (Note: at this point you may want to skim the appendix for a review of the basic ideas and terminology of hypothesis testing).

#### 7.1 Is the \( l \)-th observable characteristic irrelevant?

The hypothesis that the \( l \)-th characteristic is ignored by decision-makers when making their mode-choice decisions is the null hypothesis

\[ H_0 : \beta_l = 0. \]

\(^5\)The “at convergence” just refers to the fact that numerical methods of maximizing the likelihood are iterative methods; we are working with the value of the log likelihood based on the value of \( \hat{\beta} \) that the machine thinks is “close enough” to a maximizing value.
Under this null hypothesis, the test statistic $T$, called a t-statistic, has a $t$-distribution on $N - K$ degrees of freedom:

$$T = \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} \sim t_{N-K}$$

where $\text{se}(\hat{\beta}_l)$ is the estimated standard error of $\hat{\beta}_l$, $N$ is the number of rows in the data matrix and $L$ is the number of observable characteristics being considered. As $N - K$ gets large, the $t$-distribution approaches the standard normal distribution (this is why tables of the $t$-distribution don’t report values for degrees of freedom much above 30).

7.2 Does the $l$-th observable characteristic have weight $w_l$?

This is the null hypothesis

$$H_0 : \beta_l = w_l$$

where $w_l$ is some given numeric value. If this null hypothesis is true,

$$T = \frac{\hat{\beta}_l - w_l}{\text{se}(\hat{\beta}_l)} \sim t_{N-K}$$

Note that the previous case is a special case of this one, with $w_l = 0$.

7.3 Are all the observable characteristics irrelevant?

If we wonder if none of the observable characteristics is relevant to individual decision-making, we are asserting the truth of the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_L = 0$$

Under this null hypothesis, the statistic

$$X = -2[\ln \mathcal{L}(0) - \ln \mathcal{L}(\hat{\beta})] \sim \chi^2_L$$

That is, $X$ has a chi-squared distribution on $L$ degrees of freedom.
8 Comparison of models

Suppose we estimate a model with $L$ observable characteristics. In an attempt to better explain the behavior of the individuals in the sample, we add additional observable characteristics (explanatory variables) to the specification. Have the new characteristics really contributed to our understanding? It is easy to see that the (log)-likelihood must go up when more characteristics are added, so we cannot just compare log-likelihoods. Suppose we add $\ell$ additional characteristics, so that in our expanded model we have $L^* = L + \ell$ characteristics (variables, columns in the data matrix). Our null hypothesis is that the new variables contribute nothing, that is,

$$H_0 : \beta_{L+1} = \beta_{L+2} = \cdots = \beta_L = 0$$

To test this, we proceed as follows:

1. Estimate the original model, and call the likelihood at convergence $\mathcal{L}(\hat{\beta}; L)$ to remind us that it is based on $L$ observable characteristics.

2. Estimate the model with the additional characteristics: call the likelihood of this model $\mathcal{L}(\hat{\beta}; L^*)$. Note that $\mathcal{L}(\hat{\beta}; L^*) > \mathcal{L}(\hat{\beta}; L)$.

Then, if the null hypothesis is true, the statistic

$$X = -2[\ln \mathcal{L}(\hat{\beta}; L) - \ln \mathcal{L}(\hat{\beta}; L^*)] \sim \chi^2_\ell$$

That is, $X$ has a chi-squared distribution, with the degrees of freedom given by the number of variables added. Exactly the same idea works for removing some variables from the specification.

9 Values of time

An individual’s value of time measures the rate at which he or she is willing to substitute time and money (cost), holding utility (behavior) constant. In the economics literature, it is the marginal rate of substitution between time and cost. If $x_{iC}$ is the variable describing the cost of trip-making (that is, the $C$-th observable variable measures cost) and $x_{iT}$ describes (a particular component of) time, the value of
this time component for individual $i$, $v_{iT}$, is defined as

$$v_{iT} = \frac{\partial v_{ik}/\partial x_{iT}}{\partial v_{ik}/\partial x_{iC}}$$

In the case of a linear-in-parameters specification, this is just

$$v_{iT} = \frac{\beta_T}{\beta_C}$$

the ratio of the estimated coefficients of time and cost.

Values of time are a standard way of valuing the benefits of transportation improvements: if an individual $i$ has a value of time of $v_{iT}$ and an reduction in time results in a change in this time component of $\Delta T$, then an estimate of the benefits of the improvement is $v_{iT} \Delta T$; if there are $N_i$ individuals “like” individual $i$ in the sense of having the same value of time, an estimate of the benefits to the group is $N_i v_{iT} \Delta T$.

10 The IIA property

10.1 Definition

Consider the logit model. Define the odds of individual $i$‘s choosing mode $k$ over mode $m$ as $P_{ik}/P_{im}$. It is easy to see that for this model one has

$$\frac{P_{ik}}{P_{im}} = \frac{e^{v_{ik}}}{e^{v_{im}}}$$

— that is, the odds of choosing mode $k$ over mode $m$ depend only on the characteristics of modes $k$ and $m$ — and not on the characteristics of any other modes in

---

6One small point needs to be addressed here. In the economics literature, the marginal rate of substitution is the slope of a level curve of a utility function; and hence it is minus the ratio of the partial derivatives. Including the minus sign has the advantage that it tells us whether the characteristic in question is perceived as a “good” or a “bad” (“good” characteristics will have positive signs). In the transport literature, however, since we know that time is a “bad” — people prefer not to spend more time commuting — the convention is to drop the minus sign, and speak of “values of time” as positive quantities.

7Note that from the previous footnote, in order to get the benefit signs right, we need to take $\Delta T$ to be positive if time is reduced, and negative if time increases.
the choice set. This is the Independence of Irrelevant Alternatives (IIA) property. Though it is most easily seen for the logit model, it turns out that IIA holds in any model whose specification involves iid random variables.

The problem with this is that it can lead to the logit model’s delivering counter-intuitive (ie wrong) predictions when modes are perceived as identical, witness the famous Red Bus–Blue Bus counterexample.

What this means is that it is important to know whether the choices made by the individuals in our sample do or do not satisfy IIA. If they do not, then we will be making a mistake if we use the logit model to model their decision-making, since IIA is “built-in” to the logit model. On the other hand, if the observed choices are consistent with IIA, then the logit model is appropriate.

10.2 Tests of IIA

There are many tests of IIA in the literature, but all are based on the following idea: suppose IIA holds in the observed choices of a particular sample. Now suppose we eliminate some non-selected alternatives from each individual’s choice set (ie, delete some non-chosen rows from the data matrix). If IIA really does hold, then we have deleted “irrelevant alternatives”, and “nothing much” should change in the estimated results.

The formal difficulty is to define a test statistic and derive its distribution under the null hypothesis that the sample satisfies IIA. There are a number of such tests in the literature: here we describe one of them, the test of K. Small and C. Hsiao, *International Economic Review* 26:3 (1985), which also contains a review of some other tests, notably ones by Hausman-McFadden and McFadden-Train-Tye. To simplify the exposition, we assume that all elements of β are estimable in all cases below (the paper deals with the general case).

1. Divide the sample *randomly* into two (asymptotically) equal sub-samples by individual. (That is, so that each sub-sample contains about the same number of people). Call the parts A and B. Estimate the logit model separately for each sub-sample. Call the resulting parameter estimates $\hat{\beta}_A$ and $\hat{\beta}_B$ respectively.

2. Compute the following quantity:

$$\hat{\beta}_{AB}^0 = (1/\sqrt{2}) \hat{\beta}_A^0 + [1 - (1/\sqrt{2})] \hat{\beta}_B^0$$
Note that this involves no new statistical estimation: it’s just a weighted average of the estimates you’ve just computed. (For Small-Hsiao, this is an indirect way of computing an estimate of \( \beta \) for the whole sample).

3. Compute the sample likelihood for sub-sample B, based on \( \hat{\beta}^{AB}_{0} \). Call the result \( L^{B}(\hat{\beta}^{AB}_{0}) \). Note that this is just a matter of enumerating through sub-sample B, and involves no new estimation.

4. Delete some non-chosen elements from the original choice set (this need not be done randomly). Re-estimate the logit model using sub-sample B, based on the (restricted) choice set. Call the parameter estimates \( \hat{\beta}^{B}_{1} \), and the likelihood at convergence of this model \( L^{B}(\hat{\beta}^{B}_{1}) \).

Small and Hsiao then show that under the null hypothesis

\[ H_{0} : \text{IIA holds in the sample} \]

(equivalently: the logit model is appropriate for the individuals in the sample) the statistic

\[ \Delta = -2[\ln L^{B}(\hat{\beta}^{AB}_{0}) - \ln L^{B}(\hat{\beta}^{B}_{1})] \sim \chi^2_{L} \]

where \( L \) is the number of observed characteristics on which choices are assumed to depend.

**10.3 Models which do not involve IIA**

It is possible to write down complicated models in which IIA does not hold: in the logit case, such models are known as “nested logit” models. LIMDEP can also estimate the parameters of the nested logit model, though the data setup is fairly complicated, and estimation requires multiple steps (that is, LIMDEP’s estimator in the nested logit case is “sequential” not “full information”).

**Appendix A  Hypothesis testing review**

This Appendix summarizes some of the concept involved in hypothesis testing, and provides a simple illustration of them. It is not meant to be rigorous. For a very clear exposition, see R.V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics*, chapter on Statistical Hypotheses (chapter 6 of the third edition).
Some terminology: A statistic is a quantity that can be computed from data. Thus (in our problems) \( \hat{\beta} \) is a statistic; but \( \beta \) is not. A statistical hypothesis is an assertion about the distribution of one or more random variables. The null hypothesis (conventionally labelled \( H_0 \)) is the particular statistical hypothesis that we are interested in. A test of a statistical hypothesis is a rule, based on the value of a statistic, tells us either to accept of reject the null hypothesis. The set of values which lead us to reject the null is called the critical region of the test.

A.1 The basic idea

The basic idea of classical hypothesis testing is as follows:

1. We assume that the null hypothesis is true, and we prove that this implies a distribution for a statistic, called the test statistic. This is the distribution of the test statistic under the null hypothesis.

2. We compute the value of the test statistic from our data.

3. Now, there is always some probability of obtaining any value of the test statistic even when the null hypothesis is true. But “extreme values” seem to be inconsistent with the truth of the null hypothesis — an extreme value says that if \( H_0 \) is true, something very unlikely has occurred — and thus observing extreme values of the test statistic leads us to reject the null hypothesis.

4. So how do we decide when a value is “extreme”? The key is to note that if \( H_0 \) is true and we get an extreme value, and we reject based on that value, then we have made a mistake. We have rejected the null hypothesis when it is true, called a Type-I error. Given our rejection strategy, we can never eliminate the probability of this error.

5. But what we can do is decide up-front on how willing we are to run the risk of it. We pick a probability \( \alpha \) and mark off an area in the tail of the distribution of the test statistic which is of size (probability) \( \alpha \). We reject the null hypothesis if the value of the test statistic falls within this area: this says that we are willing to take an \( \alpha \) chance of a Type-I error. \( \alpha \) is called the significance level of the test: standard significance levels in the literature are 5% and 1%.

\[ \text{Unfortunately, about half the literature talks about this significance level by using } 1 - \text{ the significance level. This can rarely cause a problem as long as you regard it with a bit of common} \]

23
A.2 An example

Suppose we draw a random sample \( z_1, \ldots, z_5 \) of size 5 from a normal distribution with unknown mean \( \mu \), but known variance 1. We are interested in testing the null hypothesis that \( \mu = 0 \).

Now it can be shown that if we take as our test statistic the sample average \( \bar{Z} = \frac{1}{5} \sum z_i \), then if the null hypothesis is true, it is distributed as \( N(0, .2) \). We decide that the probability of a Type-I error of 5% or less is an acceptable risk, so we shall be testing at a 5% significance level. We note that very large or very small values of the test statistic are evidence against the null hypothesis. So we generate our critical region of size 5% by taking areas of \( 2\frac{1}{2}\% \) in each tail. (This is called a two-tailed test). Here is a picture of the distribution of the sample mean which would hold if the null hypothesis were true (so this is a picture of the \( N(0, .2) \) distribution):

![Distribution of Sample Mean](image)

From tables of the normal distribution, the probability of finding a value larger than \( .8765 \) from the \( N(0, .2) \) distribution is \( .025 \). Since the normal distribution is symmetric around its mean, the probability of a \( N(0, .2) \) variate being larger than \( -.8765 \) is also \( .025 \). So our critical region is the set of points \(|t| > .8765\).

From tables, \( s = 1.96 \). Now solve for \( s \), using the fact that for this example \( \mu = 0 \) and \( \sigma = \sqrt{2} \). So \( s^{*} = 1.96 \times .4472 = 0.87651 \).
Now suppose that we perform the sampling, and we find that $\bar{Z} = 1.0$. Intuitively, if $\mu = 0$, and if the picture correctly showed the distribution of $\bar{Z}$ (that is, if the null hypothesis is correct) then we’d expect computed values for the sample mean to be around zero, and certainly not out in the tails of the distribution. But our value of $\bar{Z} = 1.0$ is out in the right tail, which convinces us that the evidence is against the null hypothesis. Formally, since the test statistic is larger than our critical value we reject $H_0$.

Alternatively, the probability (if $H_0$ is true) of getting a value of 1.0 or larger (from tables) is about .013.$^{10}$ This probability is smaller than .025 (the area in the right tail above the critical value), so we are in the critical region, and we reject $H_0$.

Note that our decision to reject $H_0$ is completely dependent on the chosen significance level. Suppose we wanted to be much more cautious, and incur a risk of just 1% (rather than 5%) rejecting the null hypothesis when it is true. This generates a critical region demarcated by $-1.1519$ and $1.1519$: in this case our test statistic $\bar{Z} = 1.0$ is not inside the critical region, so we do not reject.$^{11}$

---

$^{10}$Prob[$N(0,2) > 1.0] = \text{Prob}[N(0,1) > 1.0/\sqrt{2}] = \text{Prob}[N(0,1) > 2.236] \approx .013.$

$^{11}$From the previous footnote, Prob[$N(0,2) > 1.0] \approx .013$ which in this case is larger than the .01/2 = 0.005 area in the upper tail under our 1% significance level.