# 5 Continuous/Discrete Models

## 5.1 Motivation

Decisionmakers are often in the situation of making two interrelated choices. If in each choice the decisionmaker faces a finite and exhaustive set of mutually exclusive alternatives, then qualitative choice models can readily be applied to describe the two choices. All that is required is for the choice set facing the decisionmaker to be defined appropriately. For example, suppose a worker had a choice of how many cars to own and which mode of travel to use for the commute to work. To keep the example simple, assume that the alternative modes were auto and bus and that the worker cannot own more than two cars. The two choices that the decisionmaker has can be "collapsed" and considered one choice, with the decisionmaker facing a set of alternatives each of which denotes a particular number of cars **and** a particular mode. That is, the choice set that the decisionmaker faces consists of these alternatives: (1) own no cars and take a bus to work, (2) own one car and take an auto, (3) own one car and take a bus, (4) own two cars and take an auto, (5) own two cars and take a bus. (The alternative of owning no cars and taking an auto to work is not included under the presumption that it is logically impossible.) With alternatives defined in this way, any of the qualitative choice models can be applied. Perhaps the most appealing approach, for this example, is a GEV specification based on the tree diagram in figure 5.1.

In many situations, however, a decisionmaker makes two choices that are not both "qualitative." For example, a household chooses how many cars to own and how many miles to drive each car. The first choice is among a discrete set of alternatives (0, 1, 2, and so on up to some maximum) while the second is among a continuous set of alternatives (any number of miles, and fractions of miles, above zero and below some maximum). The choice of number of cars can be appropriately described by qualitative choice models, but not the number of miles.

Another example is a household's choice of whether or not to obtain air conditioning (with the alternatives being "yes" or "no") and the choice of how much to run the air conditioner each day if it is obtained (the alternatives are any number between 0 and 24 hours). The first choice is among a discrete set of alternatives and can be described by qualitative choice models, but the second choice is among a continuous set of alternatives and cannot appropriately be described by qualitative choice models.

Choice situations such as these are called "continuous/discrete" situations, reflecting the fact that the set of alternatives for one choice is

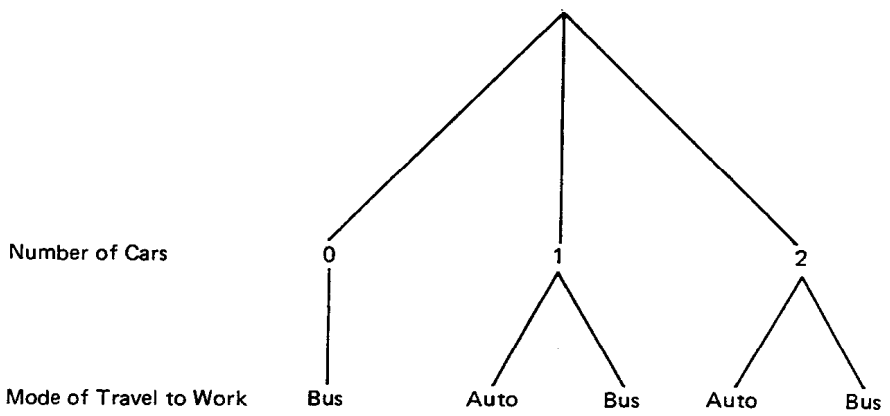


**Figure 5.1**
Tree diagram for choice of number of cars and mode of travel to work.

continuous while that for the other is discrete. Methods have recently been developed (Heckman, 1978, 1979; Dubin and McFadden, 1984) for specifying and estimating models that describe continuous/discrete choice situations. These methods are based on some relatively advanced concepts in microeconomic theory of utility maximization. Since these concepts are not widely known and are crucial to an understanding of continuous/discrete methodologies, they are now presented, prior to the discussion of the models themselves.

## 5.2 Relevant Background on Utility Maximization Theory

Assume (for simplicity) a two-good world and consider a consumer with fixed income $y$ who has a choice of how much to consume of each of the two goods. The quantities of each good are denoted $x_1$ and $x_2$, respectively, and their prices, which are fixed from the consumer's perspective, are denoted $p_1$ and $p_2$, respectively. The consumer's utility function is denoted

$$U = U(x_1, x_2).$$

If the consumer is a utility maximizer, then he will purchase the quantities of the two goods that solves the constrained maximization problem

$$\max_{x_1, x_2} U(x_1, x_2)$$
$$\text{such that} \quad y = p_1 x_1 + p_2 x_2. \tag{5.1}$$

That is, he will choose the $x_1$ and $x_2$ that maximize utility subject to the budget constraint. Label the chosen quantities as $x_1^*$ and $x_2^*$. These quantities will depend, of course, on the price of each good and the consumer's income, and hence can be written as functions of $p_1$, $p_2$, and $y$:

$$x_1^* = g_1(p_1, p_2, y),$$

$$x_2^* = g_2(p_1, p_2, y).$$

The functions $g_1$ and $g_2$ are the consumer's demand functions for $x_1$ and $x_2$.

All of this is standard material in microeconomic courses. These ideas can be extended, however, in the following way. We can substitute the **chosen** quantities of the two goods into the consumer's utility function to determine the utility that he would obtain at these chosen quantities; this gives the actual utility the consumer obtains after he has maximized his utility subject to the budget constraint:

$$U^* = U(x_1^*, x_2^*),$$

where $U^*$ is the actual utility obtained with $x_1^*$ and $x_2^*$. Since $x_1^*$ and $x_2^*$ are functions of $p_1$, $p_2$, and $y$, $U^*$ is also a function of these variables:

$$U^* = U(x_1^*, x_2^*) = U(g_1(p_1, p_2, y), g_2(p_1, p_2, y))$$

$$= Y(p_1, p_2, y).$$

That is, the actual utility that the consumer obtains after he has chosen the quantities that maximize his utility depends on the prices of the goods and his income. The function denoting this relation, $Y$, is called the "indirect utility function."

We now have two utility functions:

1. $U(x_1, x_2)$, which gives the utility that the consumer obtains at given quantities of each good and is called the "direct utility function," and
2. $Y(p_1, p_2, y)$, which gives the utility that the consumer obtains at given prices and income once he has chosen the quantities that maximize his (direct) utility subject to the budget constraint for the given prices and income.

It can be shown (see Varian, 1978) that a consumer's preferences can be equivalently represented by either a direct utility function or an indirect utility function. That is, given a direct utility function that represents the

consumer's preferences, a particular indirect utility function can be derived; and, given a appropriate indirect utility function, the consumer's direct utility function can be derived. Consequently, a researcher can specify an indirect utility function to represent a consumer's preferences[1] and know that a direct utility function is implicit.

Why is this important? A researcher usually examines a consumer's utility function for the purpose of determining the functional form of the consumer's demand function for goods; it is rare that a researcher is interested in the shape of the utility function for its own sake. For deriving demand functions, it is much easier, as will be shown, to work with a consumer's indirect utility function rather than with his direct utility function.

Under the standard analysis of consumer behavior, demand curves are derived from the direct utility function by solving the constrained maximization problem given in (5.1). This involves specifying the Lagrangian, taking derivatives of the Lagrangian with respect to each good and the Lagrangian multiplier, setting these derivatives to zero, and solving for the quantities of each good. Except for every simple direct utility functions, this procedure becomes very complex, and often intractable, so that specific demand functions cannot be derived.

Deriving demand functions from indirect utility functions is much easier, thanks to a result called "Roy's identity." Roy's identity states that the demand for a good is equal to (the negative of) the derivative of the indirect utility function with respect to the good's price divided by the derivative of the indirect utility function with respect to income. That is, using the previous notation,

$$x_1^* = -(\partial Y/\partial p_1)/(\partial Y/\partial y) = g_1(p_1, p_2, y),$$

$$x_2^* = -(\partial Y/\partial p_2)/(\partial Y/\partial y) = g_2(p_1, p_2, y).$$

*Proof of Roy's Identity* The maximum utility the consumer obtains from prices $p_1$, $p_2$ and income $y$ is given by indirect utility function

$$U^* = Y(p_1, p_2, y).$$

By definition

$$Y(p_1, p_2, y) = U(x_1^*, x_2^*), \tag{5.2}$$

where $x_1^*$ and $x_2^*$ are the utility maximizing values of $x_1$ and $x_2$ and are

themselves functions of $p_1$, $p_2$, and $y$. When utility is maximized, two things occur. First, all income is spent:

$$y = p_1 x_1^* + p_2 x_2^*.$$

Consequently, given the utility maximizing amount of good one, we know the utility maximizing amount of good two:

$$x_2^* = (y - p_1 x_1^*)/p_2.$$

Substituting into (5.2) we have

$$Y(p_1, p_2, y) = U(x_1^*, (y - p_1 x_1^*)/p_2).$$

Second, at the utility maximizing quantities $x_1^*$ and $x_2^*$, the ratio of marginal utilities is equal to the ratio of prices:[2]

$$\mathrm{MU}_1/\mathrm{MU}_2 = p_1/p_2,$$

where $\mathrm{MU}_1$ and $\mathrm{MU}_2$ are the marginal utility of goods 1 and 2, respectively. This can be rewritten as $\mathrm{MU}_1 - (p_1/p_2)\,\mathrm{MU}_2 = 0$. Therefore, indirect utility can be written as

$$Y(p_1, p_2, y) = U(x_1, (y - p_1 x_1)/p_2) \qquad \text{(evaluated at the point at which } (\mathrm{MU}_1 - (p_1/p_2)\,\mathrm{MU}_2) = 0).$$

We can now determine the derivatives of $Y$;

$$\partial Y/\partial p_1 = (\partial U(x_1, (y - p_1 x_1)/p_2)/\partial p_1) \qquad \text{(evaulated at the point at which } (\mathrm{MU}_1 - (p_1/p_2)\,\mathrm{MU}_2) = 0)$$

$$= (\partial x_1/\partial p_1)\,\mathrm{MU}_1 + (-x_1/p_2 - (p_1/p_2)(\partial x_1/\partial p_1))\,\mathrm{MU}_2$$

$$= -(x_1/p_2)\,\mathrm{MU}_2 + (\partial x_1/\partial p_1)(\mathrm{MU}_1 - (p_1/p_2)\,\mathrm{MU}_2)$$

$$= -(x_1/p_2)\,\mathrm{MU}_2$$

and

$$\partial Y/\partial y = (\partial U(x_1, (y - p_1 x_1)/p_2)/\partial y) \qquad \text{(evaluated at the point at which } (\mathrm{MU}_1 - (p_1/p_2)\,\mathrm{MU}_2) = 0)$$

$$= (\partial x_1/\partial y)\,\mathrm{MU}_1 + ((1/p_2) - (p_1/p_2)(\partial x_1/\partial y))\,\mathrm{MU}_2$$

$$= (1/p_2)\,\mathrm{MU}_2 + (\partial x_1/\partial y)(\mathrm{MU}_1 - (p_1/p_2)\,\mathrm{MU}_2)$$

$$= (1/p_2)\,\mathrm{MU}_2.$$

Therefore,

$(\partial Y/\partial p_1)/(\partial Y/\partial y) = -x_1$, as required.

The result for good two is obtained analogously. □

In short, a researcher can derive the functional form of demand equations from either direct or indirect utility functions. Since consumers' preferences can be equivalently expressed with either type of utility function, demand curves derived from either are necessarily the same. However, it is much easier to derive demand equations from the indirect utility function (using Roy's identity) than from the direct utility function (which requires solving a constrained maximization problem).

## 5.3 Specification of Continuous /Discrete Models

Consider a person who faces two choices: (1) which alternative to choose from a finite and exhaustive set of mutually exclusive alternatives; and (2) how much of a particular good to obtain, where the amount of the good can be represented by a continuous variable. In general these choices will depend, at least partially, on the same underlying factors, so that the two choices are interrelated.[3] The researcher wishes to describe the situation by specifying both (1) the probability that the person will choose each alternative and (2) the demand function for the continuous good. Label the set of alternatives as $J$, observed characteristics of each alternative $i$ in $J$ as $z_i$, the quantity of the good as $x$, the person's income as $y$, other observed characteristics of the person as $s$, and all unobserved factors as $w_i$. The price of the good can, in the general case, vary depending on which alternative is chosen,[4] and so the price is denoted $p_i$, that is, the price per unit of $x$ given that alternative $i$ is chosen.

Suppose, for the moment, that the person chose alternative $i$ in set $J$ but has not decided how much of the good $x$ to consume. The maximum utility that the person can obtain, given that he has chosen alternative $i$, depends on the price of the good and the person's income (as well as, of course, the characteristics of the person and alternative $i$). This maximum-attainable utility, given alternative $i$, can be written

$$Y_i = Y_i(p_i, y, z_i, s, w_i).$$

This function is an indirect utility function, giving the maximum utility attainable at given price and income. More precisely, it is the indirect utility

function that the person faces **given** that he has chosen alternative $i$. Since it is conditional on the choice of alternative $i$, it is called the "conditional indirect utility function" for alternative $i$. Conditional indirect utility functions can be constructed for each alternative in the set $J$; each of these gives the maximum utility that the person can obtain if he chooses a particular alternative.

We can now specify the demand equation for the good and the choice probabilities for the alternatives. The person will choose alternative $i$ if and only if the conditional indirect utility is higher for alternative $i$ than for any other alternative:

$$Y_i(p_i, y, z_i, s, w_i) > Y_j(p_j, y, z_j, s, w_j) \qquad \text{for all} \quad j \text{ in } J, \qquad j \neq i.$$

Consequently, the probability of alternative $i$ being chosen is

$$P_i = \text{Prob}(Y_i(p_i, y, z_i, s, w_i) > Y_j(p_j, y, z_j, s, w_j) \text{ for all } j \text{ in } J, j \neq i). \tag{5.3}$$

To specify these probabilities, recall that factors $w_i$ entering indirect utility are not observed by the researcher. Therefore, we decompose indirect utility into observed and unobserved parts,

$$Y_i(p_i, y, z_i, s, w_i) = V_i(p_i, y, z_i, s) + e_i,$$

where $e_i$ is a function of unobserved variables $w_i$ and $V_i$ is simply the difference between $e_i$ and $Y_i$. By specifying a distribution for $e_i$ and substituting into (5.3), explicit formulas for the choice probabilities are derived exactly the same as for any qualitative choice model. For example, if each $e_i$ is assumed to be distributed independently, identically extreme value, then the choice probabilities are logit with $V_i$ as representative utility:

$$P_i = \exp(V_i(p_i, y, z_i, s)) / \sum_{v \in J} \exp(V_j(p_j, y, z_j, s)).$$

It is important to note that representative utility in the choice probabilities includes as an explanatory variable the price of the good whose quantity is being chosen simultaneously with the choice of alternative.

The demand for good $x$ is determined from the conditional indirect utility functions using Roy's identity. That is, the demand for $x$, given that alternative $i$ is chosen, is

$$x_i = (\partial Y_i(p_i, y, z_i, s, w_i)/\partial p)/(\partial Y_i(p_i, y, z_i, s, w_i)/\partial y) = g_i(p_i, y, z_i, s, w_i).$$

This is the conditional demand for $x$ (conditional on alternative $i$ being chosen). The marginal demand for $x$, marginal over all alternatives, is the

weighted average of conditional demands with the choice probabilities being weights:

$$x = \sum_{i \in J} P_i g_i(p_i, y, z_i, s, w_i).$$

Note that both the conditional and marginal demands for $x$ depend on unobserved as well as observed factors; the error structure for these equations will depend on how $w_i$ enters $g_i$.

*Example* A simple example will demonstrate how functional forms for choice probabilities and demand functions are derived from indirect utility functions, using the ideas just expressed. Suppose the conditional indirect utility function is of the form

$$Y_i = \ln((\alpha^i + \beta^i p + \theta y + \psi f(z_i, s) + e_i) \cdot e^{-\theta p}),$$

where $f$ is a vector-valued function of observed characteristics of alternative $i$ and the person, $e_i$ is a function of unobserved factors, $\alpha^i$, $\beta^i$, and $\theta$ are scalar parameters, and $\psi$ is a vector of parameters. Note that in this example, the price of good $x$ does not depend on the alternative chosen, and so $p$ is not subscripted by $i$. The demand for $x$ is obtained with Roy's identity. First, take the derivatives of $Y_i$ with respect to $p$ and $y$:

$$\partial Y_i/\partial p = (1/A)(\beta^i e^{-\theta p} - \theta B e^{-\theta p}) = (1/A)(e^{-\theta p}(\beta^i - \theta B)),$$

$$\partial Y_i/\partial y = (1/A)(\theta e^{-\theta p}),$$

where

$$A = (\alpha^i + \beta^i p + \theta y + \psi f(z_i, s) + e_i)e^{-\theta p}$$

and

$$B = \alpha^i + \beta^i p + \theta y + \psi f(z_i, s) + e_i.$$

The conditional demand for $x$ is the negative of the ratio of these two derivatives:

$$\begin{aligned} x_i &= -(\partial Y_i/\partial p)/(\partial Y_i/\partial y) = -(\beta^i - \theta B)/\theta = B - (\beta^i/\theta) \\ &= (\alpha^i - (\beta^i/\theta)) + \beta^i p + \theta y + \psi f(z_i, s) + e_i. \end{aligned} \tag{5.4}$$

That is, the conditional demand equation for good $x$ is linear in price, income, and other explanatory variables, with an intercept term $(\alpha^i - (\beta^i/\theta))$ and an additive error.

The choice probabilities are also simple in form. The conditional indirect utility functions can be rewritten

$$Y_i = \ln(\alpha^i + \beta^i p + \theta y + \psi f(z_i, s) + e_i) - \theta p.$$

Since $\theta p$ does not vary over $i$, the decisionmaker ignores its value in comparing $Y_i$ and $Y_j$ and considers only

$$\tilde{Y}_i = \ln(\alpha^i + \beta^i p + \theta y + \psi f(z_i, s) + e_i).$$

Furthermore, since $\tilde{Y}_i > \tilde{Y}_j$ if and only if $\exp(\tilde{Y}_i) > \exp(\tilde{Y}_j)$, the decisionmaker effectively chooses an alternative on the basis of comparison among

$$\exp(\tilde{Y}_i) = \alpha^i + \beta^i p + \theta y + \psi f(z_i, s) + e_i.$$

Therefore, the probability of choosing alternative $i$ is

$$P_i = \text{Prob}(V_i + e_i > V_j + e_j \text{ for all } j \text{ in } J, j \neq i),$$

where

$$V_i = \alpha^i + \beta^i p + \theta y + \psi f(z_i, s).$$

Specification of the distribution of $e_i$ (e.g., extreme value) provides a functional form (e.g., logit) for the choice probabilities, with representative utility being $V_i$. Note that this representative utility function is linear in price, income, and other explanatory variables, with an alternative-specific constant and an alternative-specific coefficient for price. Other examples of simple continuous/discrete model specifications based on utility theory are given by Dubin and McFadden (1984).

*Remark* A final note is required regarding terminology. It was stated at the beginning of this section that the discrete and continuous choices described by these models are assumed in general to be interrelated, but the form of this interrelation was not described. It is now possible to clarify this point. In the previous specification, the decisionmaker is assumed to choose the discrete alternative and the amount of the continuous good that, in combination, provide the greatest utility. Since the choices are simultaneous, it is not possible for one choice to cause the other, in a strict sense of causation. However, the two choices are caused, or determined, by the same underlying factors, and so there is an observable association between the two. That is, the decisionmaker would (in general) choose a different alternative if, due to a change in an underlying factor, the chosen amount of

the continuous good changed; and the person would consume a different amount of the continuous good if, due to a change in an underlying factor, the person were to choose a different alternative. In these statements, the phrase "due to a change in an underlying factor" is important since the reason each choice changes when the other does is not because of direct causation between the choices but rather because both choices are determined by the same underlying factors.

## 5.4 Estimation

The parameters of both the choice probabilities and the demand equations for the continuous good can conceivably be estimated simultaneously with full information maximum likelihood methods. To do so, it is necessary to (1) specify the probability of each sampled observation (i.e., the probability of observing the alternative that was actually chosen and the amount of the continuous good that was actually consumed), (2) substitute the probability of each observation into the log likelihood function, and (3) maximize the function with respect to the parameters. While feasible, this procedure is difficult, and, to date, no special purpose computer routines have been developed for such estimation.

It is usually the case that researchers estimate the choice probabilities and demand equation sequentially, starting with the choice probabilities. Recall that the choice probabilities in continuous/discrete situations are a function of representative utility ($V_i(p_i, y, z_i, s)$) for all $i$, with the form of the function determined by the distributional assumptions regarding unobserved utility. Since each of the variables entering $V_i$ is exogenous,[5] the parameters of choice probabilities can be estimated the same as if no continuous good were involved. These estimates are consistent, but since (1) some parameters might be common to both the choice probabilities and the demand equation for the continuous good and (2) the unobserved component of utility and the error in the demand equation generally contain some common unobserved factors, the estimates are not as efficient as full information maximum likelihood.

Estimation of the demand equation for the continuous good is considerably less straightforward. The basic difficulty is that some of the explanatory variables in the demand function are, in general, correlated with the error term, causing ordinary least squares estimation to be biased. The precise source of the bias and the methods that are available for eliminating

it are most easily discussed in terms of a specific and simple example. Generalization to more complicated cases is fairly obvious.

Consider a situation in which a household has a choice between a room air conditioner and a central air conditioning system and also chooses how long each day to run the air conditioner. Suppose the conditional demand equations are linear in price and income similar to (5.4). However, in this example, price varies over alternatives since the cost of operating a room air conditioner for a minute is different from that for a central system. In particular, let the conditional demand equations be

$$x_c = \alpha + \beta p_c + \theta y + e_c, \tag{5.5}$$

$$x_q = \beta p_q + \theta y + e_q, \tag{5.6}$$

where $x_c$ is use given that a central air conditioner is chosen and the household faces price $p_c$ per minute of use, and $x_q$ is defined analogously for a room system.

These equations can be estimated simultaneously on the entire (i.e., pooled) sample, or separately on the subsample of households that chose each alternative (i.e., estimate (5.5) on those households that chose a central system and (5.6) on those that chose a room system). In either case, ordinary least squares is biased and alternative estimation methods are required. The source of the bias and methods for eliminating it are now described.

**Two Stage Estimation on Pooled Sample**

Since the parameters $\beta$ and $\theta$ are common to both equations, estimation of (5.5) and (5.6) on the pooled sample is equivalent to estimating the single equation

$$x = \alpha d^c + \beta p + \theta y + e,$$

where $x$ is the observed use level of the household, $d^c$ is a dummy variable that equals one if the household chose a central system and zero otherwise, $p$ is the price that the household is observed to face given its chosen system (i.e., $p$ is the price of using a room system if the household chose a room system and the price associated with a central system if it chose a central system), $y$ is income, and $e$ is an error term. Note that this equation is simply a more concise way of writing (5.5) and (5.6) and does not entail any change in specification.

The basic difficulty in estimating this equation is that the dummy variable $d^c$ and the price $p$ are, in general, correlated with the error term.

Consider first the dummy variable. A household whose dwelling, for some unobserved reasons (e.g., poor insulation, large picture windows in unshaded areas), tends to become unusually hot, will tend to purchase a central system since it provides greater cooling capacity than a room system. Thus, for this household, $d^c$ would probably be one, indicating a central system. This household would also, for the very same unobserved reasons, tend to use the air conditioner more than average: since the dwelling becomes unusually hot, the household would run the system for an unusually long time to reduce the heat. That is, $e$ would be high for this household. In this case, $d^c$ is one when $e$ is high. In other cases (e.g., little need for air conditioning because all the household members are away from the house during the hot part of the day), a low $e$ would be associated with a $d^c$ of zero.

Similarly, the price variable is correlated with $e$. The cost per minute of operation is generally higher for a central system than for a room system. Households that, due to unobserved factors such as poor insulation, tend to choose a central system will also tend to have above average use of the system; consequently, $p$ will tend to be high when $e$ is high. For similar, but reversed reasons, a low $p$ will be associated with a low $e$.

The basic problem here is one of endogeneity. The household determines the values of $d^c$ and $p$ in choosing which air conditioner to purchase. Since the choice of air conditioner is endogenous with the use of the air conditioner, $d^c$ and $p$ are necessarily endogenous. Treating them as exogenous in the estimation of the demand equation results in standard endogeneity bias.

The bias is shown visually in figure 5.2. The true relation between price $p$ and use $x$ is depicted by the solid line. The observed data points are the asterisks. Recall that $p$ is correlated with $e$, so that use tends to be below average when $p$ is low and above average when $p$ is high. This correlation is represented in the placement of the asterisks: for low $p$, most of the observed data points are below the true line (i.e., below the true average), while for high $p$, most are above the true line. As can be seen from this graph, the line that best fits these data is the dashed line. This estimated relation is necessarily less steep than the true line, indicating that the estimated effect of price on use is biased toward zero when price is correlated with the error term.

The solution to this problem is a two stage procedure, analogous to that for eliminating endogeneity bias in standard simultaneous equation sys-

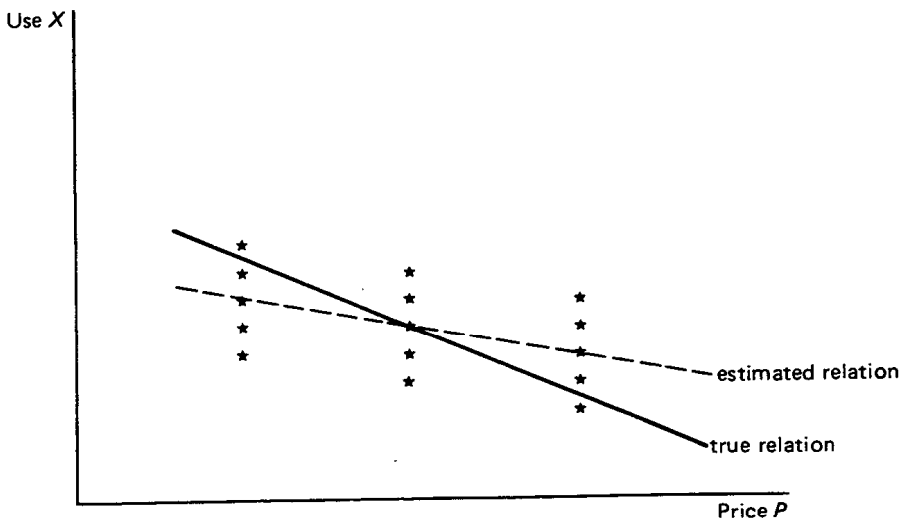


**Figure 5.2**
Bias due to endogeneity.

tems. In particular, the solution is to replace the endogenous explanatory variables by consistently estimated functions of exogenous variables. In the context of our example, the following two steps are required. First, an equation is estimated for each of the two variables $d^c$ and $p$ with only exogenous variables entering on the right-hand side:

$$d^c = f(w);$$

$$p = g(w);$$

where $w$ is a vector of exogenous variables and $f$ and $g$ are parametric functions whose parameters are estimated. Using the estimated parameters in $f$ and $g$, predicted values of $d^c$ and $p$ are obtained denoted $\hat{d}^c$ and $\hat{p}$, respectively. Second, the demand equation for air conditioner use is estimated with the estimated values of $d^c$ and $p$ replacing the observed values:

$$x_i = \alpha\hat{d}^c + \beta\hat{p} + \theta y + e_i.$$

Ordinary least squares is a consistent estimator for this equation; since $\hat{d}^c$ and $\hat{p}$ are functions of exogenous variables, they are necessarily uncorrelated with $e$.

The only question with this approach is what functions $f$ and $g$ to use in

replacing $d^c$ and $p$. Any function of exogenous variables will allow consistency; three particular functions have traditionally been used.

METHOD I The most obvious and, in some sense, straightforward method is to specify $f$ and $g$ as regression equations of all observed exogenous variables: $d^c = \omega w + u_1$ and $p = \phi w + u_2$, where $w$ is a vector of observed exogenous variables, $\omega$ and $\phi$ are vectors of parameters, and $u_1$ and $u_2$ are error terms. Ordinary least squares applied to these equations provides consistent estimates of $\omega$ and $\phi$.

METHOD II The choice probabilities are functions of exogenous variables. Since these have previously been estimated, $d^c$ and $p$ can be expressed in terms of the choice probabilities, thus avoiding the estimation of additional regression equations. That is, let the function replacing $d^c$ be the estimated probability of choosing a central system, and let the function replacing $p$ be the expected price given that either of the two systems could be chosen:

$$\hat{d}^c = \hat{P}_c,$$

$$\hat{p} = p_c\hat{P}_c + p_q\hat{P}_q,$$

where $\hat{P}_c$ and $\hat{P}_q$ are the estimated probabilities of choosing a central and a room system, respectively.

METHOD III Methods I and II can be combined for greater efficiency. Let $f$ and $g$ be regression equations, but include the estimated choice probabilities as explanatory variables in addition to exogenous variables. That is, estimate by ordinary least squares:

$$d^c = \alpha_1\hat{P}_c + \omega w + u_1,$$

$$p = \alpha_2(\hat{P}_c p_c + \hat{P}_q p_q) + \phi w + u_2,$$

where $\alpha_1$ and $\alpha_2$ are scalar parameters. Since method I is obtained when $\alpha_1 = \alpha_2 = 0$ and method II results from $\alpha_1 = \alpha_2 = 1$ and $\omega = \phi = 0$, this third method is a generalization of the other two and is consequently more efficient.

It is important to note that methods I and III are equivalent to instrumental variables estimation. For method I, the instruments are all exogenous variables available prior to estimation of the choice probabilities, while for method III the instruments are all of these exogenous variables plus the estimated choice probabilities and variables created from these

choice probabilities. Furthermore, just as two stage least squares can be equivalently performed in one stage as instrumental variables estimation, methods I and III can also be estimated in only one stage using instrumental variables routines.

**Parameters Varying over Equations** In the previous example, there are parameters common to both conditional demand equations. While this specification simplifies the notation, it is important to realize that the two stage estimation procedure is not restricted to cases with common parameters. To show this fact, suppose that all the parameters in the air conditioner use equations are different for room and central systems:

$$x_c = \alpha + \beta^c p_c + \theta^c y + e_c,$$

$$x_q = \beta^q p_q + \theta^q y + e_q.$$

This specification is actually quite reasonable. A central air system produces much more cooling within a given period of time than does a room system. A household would consequently be more willing to use a central system than it would a room system if the household (somehow) faced the same price per minute of operation for each type of system. That is, $\beta^c$ is less negative than $\beta^q$. For similar reasons, income might have a larger effect on the use of a room system than on a central one, implying that $\theta^q$ is larger than $\theta^c$.

With parameters varying over alternatives, the approach just described is applied by rewriting the two use equations as one:

$$x = \alpha d^c + \beta^c p_c d^c + \beta^q p_q d^q + \theta^c y d^c + \theta^q y d^q + e,$$

where $d^c$ and $d^q$ are dummies indicating that central and room systems were chosen, respectively. The parameters are estimated by first replacing $d^c$ and $d^q$ by their predicted values based on estimated functions of exogenous variables (using any of the three methods described), and then applying ordinary least squares.

### Selectivity Correction Approach

It is most natural to discuss this approach in the context of parameters that are not equal over equations, since the additional complication of incorporating this equality is avoided. Therefore, consider for now the specification

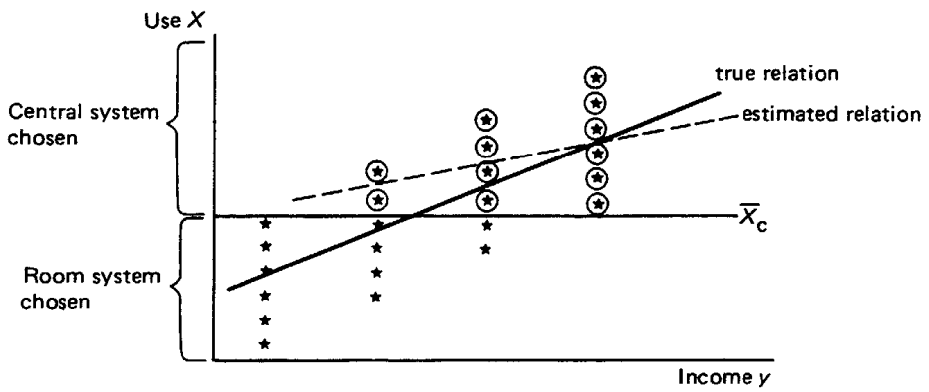


**Figure 5.3**
Bias due to self-selection.

$$x_c = \alpha + \beta^c p_c + \theta^c y + e_c, \tag{5.7}$$

$$x_q = \beta^q p_q + \theta^q y + e_q, \tag{5.8}$$

where $\beta^c \neq \beta^q$ and $\theta^c \neq \theta^q$ for the reasons discussed.

Suppose the researcher segmented the sample on the basis of the type of air conditioner chosen, and estimated (5.7) on the subsample that chose central air and (5.8) on the subsample that chose room air conditioner. Both equations would be estimated with bias. Consider income $y$ in (5.7). A household that had low income would have a low probability of choosing a central system since it costs more than a room system. As a result, if a low income household chose a central system, then there must be unobserved factors (such as poor insulation in the house necessitating a powerful air conditioner) that induced the household to do so. These same factors would also tend to induce the household to use the system more than expected. Thus, when $y$ is low for a household that chose a central system, we would expect $e_c$ to be high; the household must have higher than expected use to induce it to purchase the central air system.

The bias is shown graphically in figure 5.3. Suppose the household chooses a central system if its use will exceed $\bar{x}_c$ but will choose a room air conditioner if its use of a central system will be below $\bar{x}_c$. Suppose further that use increases with income. If the amount that a household would use a central air system were somehow observed for all households, whether or not they actually chose a central system, then the data points indicated by asterisks would be obtained; for any given income level, there would be a distribution of use around the "true" line. However, if the equation is

estimated **only** on those people who actually chose a central system, then the only data points used in the regression are those above $\bar{x}_c$, that is, the circled asterisks. As this graph shows, the line that best fits the data points used in the regression has a downwardly biased slope. This bias is called "selectivity bias," or "self-selection bias," because the estimation is performed on a subsample of households that, through their choice of alternative, essentially selected themselves to be included in estimation.

The correlation between income and the error term in the demand equation can also be seen in figure 5.3. For low levels of income, the only observed data points are those above the true line, i.e., those with positive errors; however, as income rises, negative errors become likely and larger in magnitude. Hence the negative correlation between $y$ and $e_c$.

The bias is not limited to income. The choice probabilities are really the underlying issue; when the probability of choosing a central air system is low and it is purchased anyway, we expect use to be higher than average. Thus, any variable that affects the choice probability $P_c$ is correlated, in that portion of the sample that chose central air, with the error term $e_c$ in the use equation.

Stated more formally, the problem is that the expectation of $e_c$ is not zero for each observation as required for ordinary least squares, but rather a function of the choice probability $P_c$. Therefore, to solve this problem, we decompose $e_c$ into its expectation and a deviation from its expectation:

$$e_c = E(e_c) + \eta,$$

where $E(e_c)$ is a function of the probability of choosing a central system. The deviation $\eta$ is due to factors that are unnrelated to the choice between central and room systems and so is independent of $P_c$. The use equation for a central system becomes

$$x_c = \alpha + \beta^c p_c + \theta^c y + E(e_c) + \eta. \tag{5.9}$$

Since $E(\eta) = 0$ and $p_c$, $y$, and $E(e_c)$ are independent of $\eta$, this equation can be estimated by ordinary least squares if a consistent estimate for $E(e_c)$ can be obtained. The term $E(e_c)$ is called the "selectivity correction" since its inclusion corrects for selectivity bias.

Heckman (1978, 1979) has derived expressions for $E(e_c)$ under various sets of distributional assumptions. Using these techniques, Dubin and McFadden have shown that, if the choice probabilities are logit and $e_c$ and $e_q$ are normally distributed, then the selectivity correction is

$$E(e_c) = (\sqrt{6\sigma^2}/\pi)[(\rho_q P_q \ln P_q/(1 - P_q)) - \rho_c \ln P_c], \tag{5.10}$$

where $\sigma^2$ is the variance in $e$ in the entire population (not conditional on the choice of system) and $\rho_q$ and $\rho_c$ are the correlation of $e$ with the unobserved utility associated with room and central air systems, respectively.[6]

Generally $\sigma^2$, $\rho_q$, and $\rho_c$ are unknown to the researcher. Furthermore, $\rho_q$ and $\rho_c$ are not independent; in fact, $\rho_c$ necessarily equals the negative of $\rho_q$. This fact is explained as follows. Recall that only differences in utility matter, not the absolute level (see section 2.3). Therefore, any factor either (1) increases the utility of a central system relative to a room system (and thereby decreases the relative utility of a room system) or (2) decreases the relative utility of a central system (and increases that of a room system). A factor cannot increase or decrease the relative utility of **both** alternatives. (If the utility of a central system increased by $u_c$ and that of a room system increased by $u_q$, then the relative utility of a central system increases and the relative utility of a room system decreases if $u_c - u_q$ exceeds zero, and vice versa if $u_c - u_q$ is less than zero.) Consider now an unobserved factor that increases a household's use of an air conditioner (for example, poor insultation). If this factor increases the relative utility of a central system, then it necessarily decreases the relative utility of a room system by the same amount. Thus, if $\rho_c$ is positive, then $\rho_q$ is necessarily negative by the same amount.

Since $\rho_c = -\rho_q$, equation (5.10) can be expressed in a form that does not require the researcher to know $\sigma^2$, $\rho_q$, or $\rho_c$. In particular, substituting $-\rho_c$ for $\rho_q$ in (5.10), we have

$$E(e_c) = -(\sqrt{6\sigma^2}/\pi)\cdot\rho_c\left[\frac{P_q \ln P_q}{1 - P_q} + \ln P_c\right]. \tag{5.11}$$

With estimated choice probabilities $P_q$ and $P_c$, the researcher can calculate the term in brackets. Entering this into the use equation gives

$$X_c = \alpha + \beta^c p_c + \theta^c y + \gamma^c C_c + \eta, \tag{5.12}$$

where $C_c$ is the "selectivity correction term," calculated as $((P_q \ln P_q/(1 - P_q)) + \ln P_c)$ and $\gamma^c$ is the coefficient of the selectivity correction term, which equals $(-\sqrt{6\sigma^2}/\pi)\rho_c$. Estimation of (5.12) provides a consistent estimate of each coefficient, including $\gamma^c$. Note that if the researcher expects $\rho_c$ to be positive (e.g., unobserved factors that cause high use also increase the utility of a central system), then the coefficient of the selectivity correction term is expected to be negative.[7]

By the same arguments, we can show that the conditional demand equation for use of room air conditioners can be estimated by ordinary least squares on the subsample of households that chose a room system, provided a selectivity correction term is added. The estimation equation is

$$x_q = \beta^q p_q + \theta^q y + \gamma^q C_q + \eta,$$

where $C_q$ equals $(P_c \ln P_c/(1 - P_c) + \ln P_q)$ and $\gamma^q$ equals $(-\sqrt{6\sigma^2}/\pi)\rho_q$, which can be estimated by ordinary least squares on the subsample of households that chose a room system.

**The Selectivity Approach with Common Parameters** In the specification used thus far in describing the selectivity correction approach (i.e., equations (5.7) and (5.8)), there are no parameters that are equal across the use equations.[8] However, in the original specification of the example (equations (5.5) and (5.6)), common parameters appeared in the two equations. In fact, in many real world situations, particularly if the number of choice alternatives is large compared with the sample size, the researcher will choose to specify the conditional demand equations with parameters equal over equations.

Common parameters can be handled in two ways with the selectivity correction approach. The demand equations can be estimated as a system of simultaneous equations with each equation estimated on its own subsample (i.e., the subsample that chose that alternative) and with parameters explicitly restricted in the estimation procedure to be equal across equations. Alternatively, and usually more simply, the separate demand equations can be written as one and estimated on the pooled sample. In the example of air conditioner use, equations (5.5) and (5.6) would be rewritten as

$$x = \alpha d^c + \beta(p_c d^c + p_q d^q) + \theta y + \gamma^c(C_c d^c - C_q d^q) + \eta \tag{5.13}$$

(since $\gamma^c = -\gamma_d$). With the selectivity correction term, the explanatory variables are not correlated with $\eta$ and so ordinary least squares is consistent.

Equation (5.13) points out that the selectivity correction approach can be applied on either a pooled sample or choice based subsamples. When applied on a pooled sample, it is an alternative to two stage estimation, while on choice based samples it is the only option (since two stage estimation requires a pooled sample). In fact, even if there are no common

parameters, the selectivity correction approach can be applied on a pooled sample. In the air conditioning example, instead of estimation of

$$x_c = \alpha + \beta^c p_c + \theta^c y + \gamma^c C_c + \eta$$

and

$$x_q = \beta^q p_q + \theta^q y + \gamma^q C_q + \eta$$

separately on the subsample of households that chose room and central systems, respectively, the researcher can estimate

$$x = \alpha d^c + \beta^c p_c d^c + \beta^q p_q d^q + \theta^c y d^c + \theta^q y d^q + \gamma^c (C_c d^c - C_q d^q) + \eta$$

on the pooled sample. In short, the selectivity correction approach is not restricted to the use of choice based subsamples, but is also applicable on pooled samples as an alternative to the two step estimation procedures described.

**The Selectivity Correction Approach When Conditional Demand Is Observed Only for a Subsample** The selectivity correction approach is applicable in situations that cannot be handled with the two stage estimation procedures, namely, when conditional demand is observed only for those sampled decisionmakers that chose a particular alternative. For example, suppose an electric utility has a conservation program that customers are invited to join. The utility records the savings in electricity that each program participant obtained as a result of the program, and wants to relate these savings to characteristics of the customer as well as the price of electricity faced by the customer. The situation is a continuous/discrete one, with the choice of whether to join the program being discrete and the savings from the program being continuous. However, savings are observed only for those customers who joined the program. The savings that **would have** been obtained from participating in the program by those who did not join are not observed, and savings resulting from nonparticipation are necessarily zero. The savings equations in this case are

$$x_p = \theta s + e,$$

$$x_n = 0,$$

where $x_p$ is the savings conditional upon being a participant, $x_n$ is the savings conditional upon choosing not to be a participant, and $s$ is a vector of characteristics of the customer and other explanatory variables.

The selectivity correction approach is perfectly appropriate for this situation (and was in fact developed for this type of situation rather than for ones in which two stage estimation could be used as an alternative). The researcher estimates a qualitative choice model on a sample of participants and nonparticipants; this model allows calculation of the probability of participating as a function of exogenous variables. The researcher then estimates the equation

$$x_p = \theta s + \hat{E}(e/p) + \eta, \tag{5.14}$$

on the subsample of customers that chose to participate in the program. In this equation, $\hat{E}(e/p)$ is a consistent estimate of the expectation of the error given that the customer participated in the program and is calculated as a function of the estimated probability of choosing to be a participant. If the choice model is logit, then $\hat{E}(e/p)$ takes the form of equation (5.11) with an appropriate change in terms. With this value of $\hat{E}(e/p)$, ordinary least squares applied to (5.14) is consistent.