The Importance of Sample Selection Bias in the Estimation of Medical Care Demand Equations

Mary Zimmerman Murphy*

A variety of econometric methods have been used to estimate medical care demand equations. This paper presents medical care demand equations which were estimated using a technique developed by Heckman (1976) that tests for and, if necessary, corrects sample selection bias. Sample selection bias is potentially a problem in the estimation of medical care demand equations because the subsample of those who consumed medical care is used to estimate the demand equation. This subsample was used because information on the price paid for medical care was not always available when medical care was consumed. If there exists any omitted or imperfectly observed variables in the equation predicting the consumption of any medical care and the demand equation, this will lead to dependence between the error terms of these equations. When these error terms are not independent, using ordinary least-squares on the sample of those who consumed medical care yields biased and inefficient estimates of the parameters of the medical care demand equation. In the estimation of medical care demand equations examples of possibly omitted or imperfectly observed variables include the price that the individual pays for medical care since it is difficult to determine because of insurance policies with deductibles and coinsurance rates, the time price which includes traveling and waiting time, the individual’s health status, attitude toward medical care providers and expectations of the future. Since in estimating medical care demand equations interest lies in the potential demand for the population irrespective of whether medical care was actually consumed during the sample period rather than in a conditional demand equation, this examination of the nonrandomness of a sample which includes only those who consumed medical care is important.

In the literature there have been a number of other econometric techniques used to estimate medical care demand equations. Several of the earliest studies used ordinary least-squares on the sample of those who consumed medical care (Feldstein and Severinson, 1965; Andersen and Besham, 1970; Rosenthal, 1970; Phelps, 1975; Goldmann and Grossman, 1975). Later studies used probit on the whole sample to estimate the probability that any medical care was consumed and then ordinary least-squares on the sample of those who consumed medical care to estimate conditional demand equations (Newhouse and Phelps, 1976; Sindelar, 1982; Duan, Manning, Morris and Newhouse, 1983). In another study a logit

---

*St. John Fisher College, Rochester, New York 14618.

This investigation was supported by University of Kansas General Research allocation 9533-20-0038. I wish to thank Robert Moffit, Gregory Duncan, Donald O’Hara, Joel Hay, and an anonymous referee for their suggestions during the preparation of this paper. I am also grateful to many computer consultants at the University of Rochester, Kansas, and Wisconsin-Milwaukee, especially Sue Larsen of the Social Science Research Facility at UW-M.
was used to estimate the probability that any care was consumed with a tobit to estimate the medical care demand equation (Coffee, 1983).

There has been some discussion in the literature as to which econometric technique is appropriate for estimating medical care demand equations. Duan, Manning, Morris and Newhouse (1983) present a comparison of alternative econometric models. Using statistical consistency and mean-squared error as criteria, they concluded that the best model uses separate equations to estimate the probability of positive medical expenditures and the level of medical expenditures. In a subsequent note Hay and Olsen (1984) show that the Duan et al two-part model is nested in the more general sample selection model. Duan et al (1984) show that their two-part model need not be nested in the sample selection model. Yet, their model can be interpreted as a special case of the sample selection model, one that assumes the error terms of the two equations are independent.

The purpose of this paper is to determine the importance of sample selection bias in the estimation of medical care demand equations and to report estimates with the sample selectivity correction. Demand equations for physician office visits, hospital outpatient visits and hospital inpatient days are estimated using a procedure suggested by Heckman (1976). This is reviewed in Section I, which also discusses the data used. In Section II, the medical care demand functions with the sample selectivity correction are reported. The importance of sample selection is discussed and its effects on price and income elasticities, as well as other variables, are analyzed. Conclusions are presented in Section III, Appendix A, which contains the probit regressions that are estimated as part of Heckman's procedure, and Appendix B, which contains the results of the medical care demand equations estimated using ordinary least-squares, may be obtained upon request from the author.

I. THE DATA AND THE ESTIMATION TECHNIQUE

The data used to estimate the medical care demand equations were obtained from the 1970 national health survey conducted by the National Opinion Research Center for Health Administration Studies of the University of Chicago. The sample is an area probability sample from the noninstitutionalized population of the United States with an overrepresentation of inner city, rural, poor and aged persons. These data are particularly good because the information given by individuals was verified by physicians, clinicians, hospitals, insurance companies and employers. The sample used in this empirical investigation consists of 1,926 household heads.

The estimation technique used to test for and correct sample selection bias was developed by Heckman (1976). This procedure is a result of his insight into the relationship between this problem and the omitted variable problem. What follows is a summary of Heckman's estimation procedure as it relates to the problem defined in this paper.

Three types of medical care demand equations are estimated. Assume that each of these is of the following form:

\[ M_i = X_i \beta + e_i = N(0, \sigma^2_{e_i}) \]

where \( M_i \) is the observed quantity of medical care consumed by individual \( i \), \( X_i \) is a \( 1 \times K \) vector of regressors, \( \beta \) is a \( K \times 1 \) vector of parameters, \( e_i \) is the error term, and \( \sigma^2_{e_i} \) is the variance of \( e_i \).

If \( U_i \geq 0 \), the person consumes medical care. When the demand function for medical care is estimated using ordinary least-squares on the sample of those who consumed medical care,

\[ E(\gamma(U_i \geq 0) = \frac{\phi}{\sigma^2} \quad E(A_i | U_i \geq 0) \]

and

\[ V(\gamma | U_i \geq 0) - \frac{\sigma^2}{\sigma^2} [\sigma^2 - V(h_i | U_i \geq 0)] \]

Thus, \( E(\gamma | U_i \geq 0) \) is not necessarily equal to zero, that is, the error terms of the equations that estimate the probability of consuming any medical care and the demand equation are not necessarily independent. This may be caused by omitted and imperfectly observed variables. In addition, \( V(\gamma | U_i \geq 0) \) is not necessarily constant. Thus, two of the assumptions of ordinary least-squares may be violated. In consequence, the problems that result from an omitted variable and heteroscedasticity may result.

The omitted variable interpretation arises because \( E(A_i | U_i \geq 0) \) can be treated as a regressor with coefficient \( \sigma^2_{e_i} / \sigma^2 \). Thus, the medical care demand equation becomes:

\[ M_i = X_i \beta + \lambda U_i + e_i \quad \text{for} \quad U_i \geq 0 \quad \text{and} \quad \sigma^2 = N(0, \sigma^2_{e_i}) \]

Because

\[ E(A_i | U_i \geq 0) = \lambda \sigma^2_{e_i} / \sigma^2 = (f(\phi)/[1 - F(\phi)]) \sigma^2_{e_i} \]

where \( \phi = -X_i \gamma / \sigma^2_{e_i} \),

\( f(\phi) \) is the density function of the standard normal distribution, and \( F(\phi) \) is the cumulative distribution function of the standard normal distribution. This missing variable, \( \lambda \), can be estimated for each individual. First a probit on the use of any of that type of medical care is estimated to obtain \( \gamma / \sigma^2_{e_i} \) for each type of medical care. This regression is used to calculate the \( \phi \) for that type of medical care for each individual. Then the
appropriate transformations are applied to obtain the proxy for \( \lambda \), LAMBDA, for each individual. The second part of the estimation procedure involves adding the appropriate LAMBDA to the list of regressors for each individual and estimating each demand function using a version of ordinary least-squares where appropriate asymptotic errors for the estimators are computed under the assumption that \( e \) and \( \lambda \) are jointly normally distributed. If the coefficient of LAMBDA is statistically significant, sample selection bias occurs when ordinary least-squares is used as the estimation method. The LIMDEP program written by William H. Greene was used to estimate the equations found in this paper.

II. THE EMPIRICAL RESULTS

In this section the sample selectivity corrected parameter estimates for the demand equations for physician office visits, hospital outpatient visits and hospital inpatient days are presented, and the sample selectivity bias is discussed. All of these demand equations are linear in the natural logarithm of the variables. All of the dependent variables exclude obstetric visits. For each demand equation the list of independent variables include the sample selectivity correction variable (LAMBDA), the price paid by the consumer net of insurance (PRICE), the household’s income per person (INCOME), the household head’s hourly wage as a proxy for the value of healthy time and the opportunity cost of time (WAGE), dummy variables that indicate insurance coverage for related medical care services to proxy the prices of related goods (DOCTOR INSURANCE and/or HOSPITAL INSURANCE), demographic variables (RURAL, INNER CITY, AGE, MALE, WHITE, YEARS OF EDUCATION), and health related variables (HEALTH STATUS—GOOD, WORRY). Descriptions of these variables are available upon request from the author. The results of the sample selectivity estimation procedure are reported in Table 1. The probit regressions that were estimated in order to calculate the LAMBDA’s are also available upon request from the author. These probit equations are reduced-form equations because the variable PRICE is not included since it was unobserved when each type of medical care was not consumed. In place of PRICE, dummy variables are included that indicate insurance coverage for that type of medical care (DOCTOR INSURANCE or HOSPITAL INSURANCE), the region in the country in which the individual resides (EAST, MIDWEST and SOUTH) and the family’s regular source of care (CLINIC, GENERAL PRACTITIONER, SPECIALIST, OSTEOPATH and CHIRO-PRACTOR). The ordinary least-squares parameter estimates are also available upon request from the author.

The existence of sample selection bias in these demand equations, is indicated by the statistical significance of the coefficient of the created variable, LAMBDA. If this coefficient is significantly different from zero, sample selection bias occurs when ordinary least-squares is used as the estimation procedure. As reported in Table 1, the coefficients of the LAMBDA’s in the physician office visits and hospital outpatient visits demand equations are significantly different from zero. Both coefficients have significance levels greater than the usual 5% level. The coefficient of LAMBDA in the hospital inpatient days demand equation is significant at the 12% level. To determine the robustness of these results, a number of demand equations with different specifications were estimated. For all of the demand equations, the significance level of LAMBDA’s coefficient varied quite a bit depending on the variables included in the probit and sample selection regressions. The estimates of LAMBDA’s coefficient became more insignificant the greater the correlation between LAMBDA and the other independent variables; that is, when the same or almost the same independent variables were included in the

<table>
<thead>
<tr>
<th>Variable Names</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Physician Visit</td>
</tr>
<tr>
<td></td>
<td>(Hospital Outpatient Visit)</td>
</tr>
<tr>
<td></td>
<td>(Hospital Inpatient Days)</td>
</tr>
<tr>
<td>Lambda</td>
<td>-0.94 (-3.00)</td>
</tr>
<tr>
<td>Price</td>
<td>-0.73 (-3.46)</td>
</tr>
<tr>
<td>Income</td>
<td>-0.66 (1.59)</td>
</tr>
<tr>
<td>Wage</td>
<td>-0.09 (-0.65)</td>
</tr>
<tr>
<td>Doctor Insurance</td>
<td>-0.09 (-0.50)</td>
</tr>
<tr>
<td>Hospital Insurance</td>
<td>-1.25 (-1.93)</td>
</tr>
<tr>
<td>Rural</td>
<td>-0.09 (-0.26)</td>
</tr>
<tr>
<td>Inner City</td>
<td>0.18 (1.56)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.18 (2.35)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.03 (-0.44)</td>
</tr>
<tr>
<td>White</td>
<td>-0.31 (1.60)</td>
</tr>
<tr>
<td>Years of Education</td>
<td>-0.02 (-1.63)</td>
</tr>
<tr>
<td>Health Status—Good</td>
<td>-0.04 (-2.1)</td>
</tr>
<tr>
<td>Worry</td>
<td>-0.14 (2.12)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.027 (2.37)</td>
</tr>
<tr>
<td>R-Squared</td>
<td>0.102</td>
</tr>
<tr>
<td>Chi-Squared</td>
<td>122.8</td>
</tr>
<tr>
<td>N</td>
<td>101.3</td>
</tr>
</tbody>
</table>

*Statistics are in parentheses.*

probit and sample selection correction models. For the physician office visits and hospital outpatient visits demand equations, the estimated coefficients of LAMBDA were often significantly different from zero. For the hospital inpatient days demand equation, the estimated coefficients of LAMBDA were never more significant than is reported in Table 1. Thus, sample selection bias is clearly a problem in the estimation of primary medical care demand equations; that is, the demand equations for physician office visits and hospital outpatient visits. This may be because the consumer has more choice in the consumption of these services, so there is more of a problem due to omitted or imperfectly observed variables. In the estimation of the hospital inpatient days demand equation, sample selection bias seems to be only somewhat of a problem.

The sign of the coefficient of the sample selectivity correction variable, \( \lambda \), is difficult to predict because it depends on the sign of \( \tau_0 \). A priori the medical care demand equations might seem to be from a tobit model. The tobit model is appropriate if the observed terms of the equation predicting the consumption of any medical care and the demand equation are derived from the same normal distribution. In that case, the coefficient of \( \lambda \) is \( \tau_0 \), which is positive. Since the estimated coefficients of the LAMBDA’s in all of the demand equations are negative and less than one in absolute value, this indicates that the numerator of the coefficient of \( \lambda \) is not \( \tau_0 \), so the tobit model is inappropriate. In addition, the estimated sample selectivity corrected and probit coefficients for each type of medical care are not proportional and in many cases are of opposite sign.

Sample selection bias affects the coefficients of the independent variables that are of interest in public policy discussions. The sample selectivity corrected estimate of the own-price
elasticity of physician office visits is -.073. This implies that the demand for physician office visits is very inelastic, but it is not perfectly inelastic since the coefficient is significantly different from zero. Thus, changes in insurance coverage for physician office visits will cause changes in their quantity demanded. The ordinary least-squares estimated own-price elasticity is -.059 which is 9% smaller than the sample selectivity corrected estimate. In this case, the sample selection bias is toward a more inelastic estimate of the own-price elasticity. Both of these estimates are close to those in the literature. For example, the own-price elasticity reported in Feldstein and Severson (1964) is -.19 and the one reported in Newhouse and Phelps (1976) is -.08.

Correcting for sample selection bias makes even more of a difference in the estimation of the income elasticity of demand for physician office visits. The sample selectivity corrected estimate of the income elasticity is -.066 which is not significant at the 10% level, while the ordinary least-squares estimate is .113 which is almost twice as large and is significant at better than the 5% level. Thus, the sample selectivity income elasticity estimate implies that those with higher incomes do not demand more physician office visits while the ordinary least-squares estimate implies that physician office visits are normal goods. These estimates are in the range of estimates reported in the literature. Feldstein and Severson (1964) estimated the income elasticity of physician office expenditures to be .56, Andersen and Benham (1970) estimated it to be .32, Silver (1970) estimated it to be .85, and Phelps (1975) estimated it to be -.03.

Newhouse and Phelps (1976) found that income did not have a significant effect on the demand for physician office visits. Sindelar (1982) estimated the physician expenditure elasticity to be -.00001 for males and .00002 for females.

The importance of the selectivity correction in the estimation of the own-price and income elasticities of hospital outpatient visits is similar to the physician office visits case. The own-price elasticity of outpatient visits is -.018 when sample selectivity is corrected and -.007 when it is not. This is a 60% difference, but neither of these elasticities are significantly different from zero. These perfectly inelastic estimates of the own-price elasticity may occur because 28.8% of the outpatient visits were emergency department visits. The income elasticity of demand for hospital outpatient visits is -.096 when selectivity bias is corrected, but is not significant at the 10% level. In contrast, when selectivity is not corrected, the income elasticity is -.138 and is significant at better than the 5% level. Thus, the sample selectivity corrected estimate implies hospital outpatient visits demand is not affected by income level while the ordinary least-squares estimate implies outpatient visits are inferior goods. Both the estimated own-price elasticity and income elasticity for outpatient visits, which includes outpatient and emergency department visits, are different from the estimates obtained by Davis and Russell (1972) for outpatient department visits. Using aggregate data, they estimated the own-price elasticity to be -.166 and the income elasticity to be in the range between -.07 and .72.

Although the coefficient of LABMBA in the hospital inpatient days demand equation was not as significant as the ones in the primary care demand equations, the sample selectivity correction has a similar, although smaller, effect on the own-price and income elasticities. The sample selectivity corrected own-price elasticity of hospital inpatient days is -.079 and the ordinary least-squares estimate is -.076. Both coefficients are significant at approximately the 10% level. These elasticities are in the range of those reported in the literature. Feldstein and Severson (1964) reported an own-price elasticity of -.04, Rosenthal (1970) reported elasticities ranging from 0 to -.7 depending on the diagnosis, Rosett and Huang (1972) reported elasticities from -.25 to -.5 depending on the coinsurance rate, Davis and Russell (1972) reported elasticities from -.32 to -.66 depending on the measure of price used, and Newhouse and Phelps (1976) reported an elasticity of -.06.

The selectivity corrected income elasticity for hospital inpatient days is -.313 and the ordinary least-squares estimated income elasticity is -.282. Both are significant at better than the 5% level. This result that hospital inpatient care is an inferior good differs from the results reported in the literature. Feldstein and Severson (1964) reported an elasticity of .0005, Rosett and Huang (1972) reported elasticities between .25 and .45 depending on the income level, Davis and Russell (1972) reported income elasticities between .35 and .42, Phelps (1975) reported an income elasticity of .06, and Newhouse and Phelps (1976) reported an income elasticity of .03 when non-wage income was greater than $3,000 per year.

Sample selection bias also affects the magnitude and significance of the other estimated parameters in the medical care demand equations. In theory, the ordinary least-squares coefficients measure the effect of a change in an independent variable on the conditional demand for medical care, while the sample selectivity corrected coefficients measure the effect of a change in an independent variable on the demand for medical care. That is, in addition to measuring the effect of a change in an independent variable on medical care demand given some medical care was consumed, the sample selection model coefficient also measures the effect of a change in an independent variable on the probability any medical care is consumed.

For many of the independent variables, the sample selectivity and ordinary least-squares estimated coefficients are of a similar sign and significance level. This is true in the physician office visits equations for the variables WAGE, RURAL, INNER CITY, AGE, and WORRY. The estimated coefficients of WAGE are negative, but insignificant. These results may reflect the dual role of WAGE in medical care demand functions. It is both a proxy for the value of the individual's time and a proxy for the time cost of consuming medical care. For both estimation methods for the demand for physician office visits these effects cancel out. For both estimation methods, the coefficients of RURAL are negative, but insignificant, while the coefficients of RURAL and RURAL are positive, but insignificant. Thus consumers who reside in these areas are no more or less likely to demand physician office visits than those of urban areas. For both estimation methods, the coefficients of AGE and WORRY are significantly positive. Thus the older the individual, or the more an individual worries about his or her health, the more physician office visits are demanded.

The remainder of the independent variables of the physician office visits demand equation have the same sign with both estimation methods, but HOSPITAL INSURANCE and YEARS OF EDUCATION are significant in the sample selectivity model only, while MALE, WHITE and HEALTH STATUS-GOOD are significant in the ordinary least-squares model only. These differences in statistical significance may result from multilinearity in the sample selectivity model caused by the high degree of collinearity between the sample selectivity variable, LABMBA, and the other independent variables. Although there are several identifying variables in the probit regression, the sample selectivity model is basically identified by the functional form of LABMBA. The negative coefficients for HOSPITAL INSURANCE imply hospital inpatient days and physician office visits are substitutes, although the probit regression implies that an individual having hospital insurance coverage is more likely to seek care in a physician's office. The negative coefficients of YEARS OF EDUCATION imply that the more educated demand fewer physician office visits. This may reflect a greater efficiency in the production of health and more use of preventive measures since the probit regression indicates that the more educated are more likely to seek medical
THE IMPORTANCE OF SAMPLE SELECTION BIAS

This paper explores the importance of sample selection bias in the estimation of medical care demand equations. Physician office visits, hospital outpatient visits, and hospital inpatient days demand equations were estimated using an econometric technique developed by Heckman which simultaneously tests for sample selection bias and corrects as necessary. Sample selection bias is determined by testing the significance of the coefficient of the sample selectivity variable, LAMBDA. Based on the coefficients of the LAMBDA's in the demand equations, sample selection bias was found to be a very significant problem in the estimation of the primary care demand equations and a rather significant problem in the estimation of the hospital inpatient days demand equation.

Comparison of the selectivity corrected regression coefficients and the sample selection biased ordinary least-squares coefficients shows that the bias in the estimated own-price elasticity is toward less elastic, but is small. The bias in the income elasticity is more important. In both of the primary care demand equations the sample selectivity bias caused the effect of income on demand to be significantly different from zero (positive in the office visits case and negative in the inpatient case), when the sample selectivity corrected estimates indicated that income had no effect on either of these demands. In the hospital inpatient days equation, the sample selection bias caused the coefficient of income to become more negative. Sample selection bias also affected the estimates of the other coefficients in these equations. In the sample selectivity corrected regressions some of the coefficients of independent variables were insignificant, whereas they were significant in the ordinary least-squares regressions. From this study it appears that the coefficients that result from the Heckman estimation procedure are sensitive to the model's specification. In addition, multicollinearity may be a problem because the sample selectivity variable, LAMBDA, is highly collinear with the independent variables.

Comparing the predicted yearly medical care demand at the sample means of the independent variables reflects the effects of sample selection bias on the entire equation. The individual's demand for physician office visits is 3.250 per year for the selectivity corrected equation compared to 3.067 per year for the ordinary least-squares equation. This is a difference of 5.6%. The individual's predicted demand for outpatient services is 2.077 for the selectivity corrected equation compared to 1.801 per year for the ordinary least-squares equation. This is a difference of 13.3%. The individual's predicted demand for hospital inpatient days is 9.268 per year for the selectivity corrected equation compared to 6.698 for the ordinary least-squares equation. This is a difference of 27.7%. Although these predictions are fairly close, when they are used to predict the effects of co-insurance rate changes for the Medicare or Medicaid populations, for example, or other changes in public policies, the differences become quite substantial. Therefore, sample selection should be investigated when medical care demand equations are estimated.

FOOTNOTES
1. This formulation of the sample selectivity problem is from Heckman (1979), pages 154-157 and
2. The natural logarithm is not taken of the dummy variables.
4. These elasticities from Sindelar (1982) were calculated at the mean values of income and expenditures for the adult population.

5. Newhouse and Phelps (1976) reported that when own-wage income was less than $3,000 per year, income did not have a significant effect on the demand for hospital inpatient days.

REFERENCES


