ON HAVING YOUR CAKE AND EATING IT TOO: ECONOMETRIC PROBLEMS IN ESTIMATING THE DEMAND FOR HEALTH SERVICES

PREPARED UNDER A GRANT FROM THE NATIONAL CENTER FOR HEALTH SERVICES RESEARCH AND DEVELOPMENT

JOSEPH P. NEWHOUSE, CHARLES E. PHELPS, M. SUSAN MARQUIS

R-1149-1-NC
OCTOBER 1979

Rand
SANTA MONICA, CA. 90406
This research was supported by PHS Grant 1 RO1 HS00840-01 from the National Center for Health Services Research and Development.

Library of Congress Cataloging in Publication Data

Newhouse, Joseph P.
On having your cake and eating it too.

(Report - Rand Corporation ; R-1149-1-RC)
Bibliography: p.
AS36.R3 R-1149. [RA410.5/3] 061s [362.1'0973 79-15299

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation
ON HAVING YOUR CAKE AND EATING IT TOO: ECONOMETRIC PROBLEMS IN ESTIMATING THE DEMAND FOR HEALTH SERVICES

PREPARED UNDER A GRANT FROM THE NATIONAL CENTER FOR HEALTH SERVICES RESEARCH AND DEVELOPMENT

JOSEPH P. NEWHOUSE, CHARLES E. PHelps, M. SUSAN MARQUIS

R-1149-1-NC
OCTOBER 1979

Rand
SANTA MONICA, CA. 90406
PREFACE

This report revises and expands a Rand report first issued in 1974. The first version critically reviewed the econometric literature on the demand for medical care. The present version revises both theoretical and empirical materials in the first version and considers a few additional studies that have appeared in the literature since 1974.

Some of the theoretical results in the first report were conditional upon the correctness of marginal price (net of insurance) as an explanatory variable in the medical care demand equation; its use is now known to be incorrect when the insurance policy has a deductible. Unfortunately, when a deductible is present, the price consistent with a maximizing model is unobservable; hence, the empirical results are conditional upon a sample of persons who do not have a deductible in their policy. In addition, the empirical results in the first report were based upon a sample of heads of households; the results in this report are based upon a probability sample of the national population.

This report is intended to analyze methodological problems in the field of medical demand studies. Although it explains why the authors believe there are serious problems in estimates of the demand response to health insurance derived by others, it is not their intention to “discredit” anyone else’s work. The problems with other studies stem primarily from the data available rather than from inappropriate theory. The existing literature must be regarded collectively as pioneering attempts to derive estimates where there were none before.

The original work on this report was supported by a grant from the National Center for Health Services Research and Development, 1-R01-HS00840-01. Revision of the report was supported by the Health Insurance Study Grant from the Department of Health, Education, and Welfare.
SUMMARY

This report reviews the econometric literature on the price (insurance) elasticity of demand for medical care services. Estimates of the price elasticity have theoretical interest because they indicate the opportunity cost of risk avoidance; if price elasticities are small, more complete coverage will be optimal. The estimates have policy interest because of the debate over the costs of national health insurance.

The econometric literature unfortunately contains numerous inconsistent estimates of the demand elasticities. The inconsistencies result from limitations inherent in existing data. We consider several different sources of inconsistencies in this report; where possible, we quantify their magnitude.

We first consider systematic dependence between the price variable and the error term. Such a dependence arises when the insurance policy contains a deductible and observations are distributed on both sides of the deductible. (The situation is analogous to demand for electricity with a declining block rate structure.) In our sample, including observations when a deductible is present and using an average price variable (as is common in the literature) raise estimated demand elasticities for office visits some 30 percent relative to estimates when such observations are excluded. Some of this increase may be attributable to selection effects; theory suggests that sicker individuals would have more complete insurance (another type of dependence between the error term and the price variable). Simultaneous equation methods, of course, are consistent, but they yield estimates with quite large standard errors. In the case of hospital length of stay, almost everyone exceeds the deductible; hence, any increase in the estimated elasticity from including individuals with a deductible in the sample should be primarily attributable to selection effects. No increase in elasticity is found, suggesting that for this variable selection effects may not be important.

Three sources of systematic error in the own-price variable are identified; the first arises from the use of marginal or average price as an explanatory variable when a deductible is present. The theoretically correct price is unobservable; the direction of the inconsistency from the use of average or marginal price can be positive or negative. Because we cannot observe the correct price (given existing data), we cannot quantify the magnitude of this inconsistency. Our suggestion is to omit individuals with a deductible from the sample unless selection effects are thought to be large for the given medical service studied (e.g., psychotherapy, dental services).

A second source of error arises in a specification that uses a gross price variable together with a dummy variable for the presence or absence of insurance. The direction of the resulting inconsistency cannot be determined a priori. In the sample we use, the inconsistency appears important; we estimate that the magnitude of the inconsistency in the elasticity of physician office visits could be a factor of three and in the elasticity of hospital length of stay a factor of 50 percent. Both inconsistencies are away from zero. If price must be estimated by dividing expenditure by quantity, and the measurement of quantity is subject to random error, an inconsistency
arises that is away from zero; but this conclusion need not hold if expenditures are also subject to random measurement error.

An important inconsistency arises if data are aggregated across services and an average coinsurance rate is used as the price variable. Medical services that tend to be expensive (e.g., hospitalization) are well-insured (as the theory of insurance would predict), causing a spurious relationship between coinsurance and expenditure. The inconsistency in this case can cause a ten-fold increase in estimated demand elasticities.

Another type of aggregation problem also may be present in the literature. Several researchers have used data aggregated across individuals (e.g., across the residents of a state). Strong, probably unrealistic assumptions are necessary for such estimates to be stable predictors of demand under national health insurance. For example, it is sufficient for stability (given the correctness of a linear functional form) that price changes be proportional across individuals. Yet this condition would not hold if a uniform national health insurance program were enacted; indeed, it would be the intent of the legislation to disproportionately improve the insurance of certain disadvantaged individuals.

Two final sources of inconsistency in the literature were not empirically important in our sample. Omission of a cross price variable did not affect the estimated own-price coefficient, and use of nominal rather than real price and income variables when using cross-sectional data also did not noticeably change estimated price or income elasticities.

Data from a social experiment in health care financing, which will be available in the 1980s, should not be subject to the limitations described above and so should greatly improve our ability to make reliable estimates of the elasticity of demand for medical care.
ACKNOWLEDGMENTS

The authors are most indebted to Ronald Andersen of the Center for Health Administration Studies at the University of Chicago for making available the data upon which this report is based.

The authors would like to thank Rand colleagues Jan Acton, Emmett Keeler, Bridger Mitchell, and Carl Morris for their comments and David Weinschrott for computational assistance. Thanks are also due Robert Evans of the University of British Columbia, Michael Grossman of the City University of New York, Lester Lave of Carnegie-Mellon University, Mark Pauly of Northwestern University, and Richard Rosett of the University of Chicago for their comments.
"If I eat one of these cakes," she thought, "it's sure to make some changes in my size; and, as it can't possibly make me larger, it must make me smaller, I suppose."

So she swallowed one of the cakes, and was delighted to find that she began shrinking directly.

—Lewis Carroll, Alice in Wonderland

I. INTRODUCTION

Health insurance has been characterized as a problem in tradeoffs. Because insurance affects the demand for medical care, an optimal insurance policy reaches some balance between risk spreading and appropriate incentives (Arrow, 1973; Feldstein, 1973; Phelps, 1976; Zeckhauser, 1970). The nature of the tradeoff depends upon how demand for medical care responds to insurance changes; if the response is low, then the costs of more risk avoidance are also low and conversely. In other words, if insurance induces little demand, then one can have the "cake" of risk avoidance with little real cost—eating it too. The ability to have one's cake and eat it too depends upon the magnitude of elasticity of demand for medical care with respect to insurance coverage. Accurate estimates of the price or insurance elasticity of demand are of more than academic interest in view of the policy debate over national health insurance.

In this report we review the econometric literature on the demand for medical care services. The literature unfortunately contains numerous inconsistent estimates of the price elasticity. These arise principally from the lack of appropriate data with which to estimate a demand function and not from inappropriate theory. Before we turn to probability-limit theorems and actual data to estimate demand elasticities for medical care, a word about the extant data limitations seems appropriate. Researchers estimating the magnitude of demand elasticities have been forced to rely upon data that are often ill-suited to econometric analysis. A primary data source for medical demand studies has been cross-section household surveys, the best of which contain data on demographic characteristics of household members, actual medical utilization and expenditure data, and detailed data on health insurance policies held by the families. We use such a data source below, but even it poses many estimation problems. Health insurance policies frequently contain deductibles (fixed amounts per year that must be spent before the insurance coverage affects the price of medical care), coinsurance (fractional payments for care by consumers), internal limits on use (such as cutoffs in the number of hospital days, office visits, numbers of prescriptions allowed, or aggregate expenditure on ancillary services within a hospital), or fee-schedule limits (such as a per-visit limit for office visits) for provider payments. Provisions fre-

1 Pauly (1971) has pointed out that in a public plan externalities can play a role in determining optimality; this is not contemplated by the four papers cited.
quently differ for different medical services. As a result, the price schedule of medical services is multidimensional and nonlinear.

Other researchers rely upon aggregate data, typically showing the average medical expenditure and average health insurance benefit among a group of people (e.g., a state's residents). Given the multidimensionality and nonlinearity of the individual policies, aggregation into a single variable to describe the insurance coverage is very likely to be inappropriate. Still other researchers have used data from insurance company files to study effects of insurance coverage on demand, but these data typically contain little or no evidence on important consumer demographic characteristics. Thus, nearly all data sources will probably give rise to potentially severe problems in estimating demand functions for medical services. Fortunately, some large investments of recent years in improved data will greatly help this situation in the 1980s. In addition to certain general problems discussed in the text, Appendix A details some particular problems of selected studies.

\footnote{The major exception to this is a data set collected by Scitovsky and Snyder (1972) from a natural quasi-experimental change in insurance coverage.}
II. DEPENDENCE BETWEEN THE PRICE VARIABLE AND THE ERROR TERM

There are various misspecifications of the price variable in the econometric literature on the demand for medical care. Where possible, we shall derive the sign of the inconsistency attributable to the misspecification and quantify its magnitude.

We begin by stating the problem formally in the simple regression case (the argument generalizes readily). Suppose the true model is given by:

$$Y = X\beta + \epsilon,$$

where $Y$ and $X$ are each $n \times 1$ vectors of the demand for medical care and price, respectively, $\beta$ is a scalar, and $\epsilon$ is $n \times 1$ and distributed $N(0, \sigma^2)$. The price variable may be subject to stochastic measurement error $u$, such that the estimated equation is:

$$Y = (X + u)\beta + \epsilon' = \bar{X}\beta + \bar{\epsilon},$$

where $\bar{\epsilon} = \epsilon - u\beta$, and $\bar{X} = X + u$.

Then the OLS estimate $b$ of $\beta$ equals

$$b = \Sigma (\bar{X}Y)/\Sigma \bar{X}^2 = \Sigma (\bar{XX}\beta + \bar{X}\epsilon)/\Sigma \bar{X}^2.$$

Suppose $\liminf (1/n)(\Sigma u\epsilon) = 0$, but allow $\liminf (1/n)(\Sigma Xu)$ and $\liminf (1/n)(\Sigma X\epsilon)$ to be nonzero.

Then

$$\liminf(b) = \liminf (1/n)(\Sigma XX\beta)/\liminf(1/n)(\Sigma \bar{X}^2)$$

$$+ \liminf(1/n)(\Sigma X\epsilon)/\liminf(1/n)(\Sigma \bar{X}^2)$$

$$= [\sigma_X^2 + \sigma_{Xu}^2][\sigma_X^2 + 2\sigma_{Xu}^2 + \sigma_u^2] + \sigma_X\epsilon/\sigma_X$$

$$= (1 - (\sigma_{Xu} + \sigma_u^2)/(\sigma_X^2 + 2\sigma_{Xu}^2 + \sigma_u^2))\beta + \sigma_X\epsilon/\sigma_X$$

$$= (1 - \lambda)\beta + \gamma,$$

where

$$\sigma_X^2 = \liminf(1/n)(\Sigma X^2), \sigma_{Xu} = \liminf(1/n)(\Sigma Xu), \sigma_X\epsilon = \liminf(1/n)(\Sigma X\epsilon)$$

and $\sigma_u^2 = \liminf(1/n)(\Sigma u^2)$. The first problem we consider arises from specifications in which price is correlated with the error term $\epsilon$, so that $\gamma$ is nonzero. One common problem of this sort arises when the insurance policy contains a deductible and either marginal or

1 The variables are measured as deviations around their means.
average price is used as the explanatory variable $X$. The literature has used both marginal and average price as explanatory variables; obviously they differ if a deductible is exceeded. As we discuss in the next section, neither is in general theoretically correct.

Because individuals with large values of $\epsilon$ (large consumption of medical services) are more likely to exceed a deductible and conversely, $\gamma$ will be negative; the estimated price elasticity will be inconsistent away from zero. (A similar situation obtains if there is a declining block rate for a good such as electricity.) Nearly all the studies described below include individuals with deductibles in the sample and probably overstate elasticities.

To ascertain the potential magnitude of the inconsistency, we have used the 1963 Center for Health Administration Studies (CHAS) national probability sample survey data. These data are particularly useful for our purposes because they have more details about the insurance policy than any other extant survey. The data are described in Andersen and Anderson (1967) in detail.

In Table 1 we present the estimated price elasticity of nonsurgical physician office visits for two different samples, one including individuals with deductibles in their insurance policies and one excluding them. The marginal price is used as an explanatory variable, and the functional form is linear (as it is in all results in this report). As can be seen, including those whose policy has a deductible in the sample increases the estimated price elasticity at the mean of the sample more than 30 percent. (All price elasticities here are reported at the mean of the sample.)

It might be argued that not all this difference should be attributed to the definitional correlation between $X$ and $\epsilon$ with a deductible present; those with deductibles may also be healthier. (See Appendix B for a formal model.) Indeed, the theory of demand for insurance suggests that individuals' expectation of a medical care consumption should be negatively associated with their demand for deductibles, given that premiums are unrelated to utilization. Most health insurance, however, is not selected by individuals but provided through the place of employment with a subsidy from the employer. In fact, 83 percent of private health insurance is sold in this manner (Mitchell and Phelps, 1976). Arguably, insurance might therefore be treated as exogenous. But an unknown number of work groups allow choice among a number of plans, again raising the possibility of endogeneity.

An indirect test of the existence of correlation between $X$ and $\epsilon$ caused by adverse selection is provided by regressions of hospital length of stay upon marginal price, when individuals with deductibles are included and excluded from the sample. Because people admitted to a hospital almost always exceed their deductible, their marginal (but not their average) price should be nearly uncorrelated with the error term. (In this case, marginal price is theoretically correct and average price is not, as discussed in the next section.) Thus, to the degree there is adverse selection, marginal price elasticities for hospital length of stay should be closer to zero when those with deductibles are excluded than when they are included.

---

2 We recognize that inconsistency is an asymptotic property and that our estimates pertain to finite samples. We believe they are still useful in assessing the importance of the problems discussed.

3 Phelps (1976) presents some evidence that less healthy individuals have higher maximums and lower coinsurance rates in their policies; Phelps, however, does not have a good measure of individual health status, and his results are therefore rather preliminary.
Table 1

EFFECT ON THE ELASTICITY OF DEMAND
FOR NONSURGICAL-RHYSICIAN
OFFICE VISITS

<table>
<thead>
<tr>
<th>Condition</th>
<th>Elasticity</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluding persons</td>
<td>-.095</td>
<td>1682</td>
</tr>
<tr>
<td>with deductibles</td>
<td>(t = -2.40)</td>
<td></td>
</tr>
<tr>
<td>Including persons</td>
<td>-.125</td>
<td>2325</td>
</tr>
<tr>
<td>with deductibles</td>
<td>(t = -3.98)</td>
<td></td>
</tr>
</tbody>
</table>

*The survey contained data on 7803 individuals; 2763 individuals in 788 families claimed they owned insurance, but this could not be verified so they were dropped. Seven individuals were excluded because the data about their insurance were incomplete. For computational reasons we also dropped 35 people in families with more than three insurance policies. Among the remaining 4998 individuals, 2817 had at least one physician office visit during the year before the survey, and 368 reported a (nonobstetrical) hospital admission. This group formed the basis of our analyses of physician office visits and hospital length of stay (reported in other tables). We further excluded from the subsamples of users persons with reported wage income or imputed value of time exceeding $500 per week or less than zero (30 users of physician services, 6 hospitalized); persons receiving free care or on welfare (444 users of physician services, 83 hospitalized); persons exceeding the limits of their insurance policy (19 users of physician services, 5 hospitalized), and persons who exceeded $50 per visit for office visits (7 users of physician services, 1 hospitalized). The combined effect of the exclusions (two or more reasons for exclusion may have applied to the same individual) left us with a sample of 2325 persons for estimating the physician office visit equations and 295 persons for estimating the hospital length of stay equations.

The explanatory variables in hospital length of stay and office visit equations include:
- Hospital room and board marginal price
- Physician office visit marginal price
- Wage income/week (0 if no wage income)
- Estimated value of time (0 if wage income > 0)
- Nonwage income if < $3000 (0 if > 3000)
- Nonwage income if > $3000 (0 if < 3000)
- Dummy = 1 if nonwage income > $3000
- 4 dummy variables for education of family head
  - 9–11 years
  - 12 years
  - 13–15 years
  - 16 or more years
  - Omitted = 0–8 years
- 6 dummy variables for age of person
  - 7–17 years
  - 18–24 years
  - 25–34 years
  - 35–54 years
  - 55–64 years
  - 65 or more years
  - Omitted category = 0–6 years
- Family size
- Dummy = 1 if female
- Dummy = 1 if nonwhite
- Number of disability days
- 3 dummy variables for health status
  - Good
  - Fair
  - Poor
  - Omitted: excellent
- Dummy = 1 if married
- M.D.s/100,000 population
- (M.D.s/100,000 population)²
- Beds/1,000 population
- (Beds/1,000 population)²

Money prices and income are in real terms.
As Table 2 shows, this is not the case; when those with deductibles are excluded, marginal price elasticities move away from zero. But precision is sufficiently poor that we are reluctant to infer from these results that selection effects can definitely be ignored; however, it is plausible that in these data (and for this service) they are small.

If selection effects are in fact negligible, a straightforward "solution" to the problem of estimating price elasticities is simply to exclude individuals with deductibles from the sample. This was our procedure in previous work (Newhouse and Phelps, 1976), and we shall use it for physician visit equations here. In the case of hospital length of stay equations we shall include individuals with deductibles, because we obtain approximately a 50 percent increase in sample size, and we believe $X$ and $\epsilon$ are approximately uncorrelated for this particular dependent variable (because the deductible is nearly always exceeded).\(^4\)

A more general method for attacking the possible endogeneity of insurance would be to use a simultaneous equation estimator. Unfortunately we have not found a satisfactory way of identifying the insurance variable.

### Table 2

<table>
<thead>
<tr>
<th>Condition</th>
<th>Elasticity</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Including those with deductibles</td>
<td>$-0.065$ ($t = -1.87$)</td>
<td>295</td>
</tr>
<tr>
<td>Excluding those with deductibles</td>
<td>$-0.099$ ($t = -1.92$)</td>
<td>195</td>
</tr>
</tbody>
</table>

We have estimated all the regressions reported here with insurance treated as endogenous using TSLS (and also using Zellner's Minimum Expected Loss (MELO) estimator), with work group size, industry, and occupation as instrumental variables; we consistently obtained fairly large standard errors. (Newhouse and Phelps, 1976, find similar results.) Because of the large standard errors, we do not report the TSLS (or MELO) results.

In a few years, experimental data from the Health Insurance Study (Newhouse, 1974) will be available; in this controlled experiment insurance is exogenous, so the issue of selection effects is moot. When analyzing other data, one will probably have to choose between OLS and considerably less precise, but asymptotically consistent, estimates. If this is the case, our preference is generally for the OLS estimates, except in certain instances when selection effects are potentially large. Most of the

\(^4\) In this report we analyze demand for hospital and physician services conditional upon some use and do not analyze the use-no use decision. Our analysis points up the general type of inconsistencies in the literature and gives a lower bound on their magnitude (because the inconsistencies would be in a similar direction for the use-no use decision and total price elasticities are additive). The lower bound is sufficiently large to demonstrate that there may be serious problems with the estimates in the literature; therefore, we see little value in also analyzing the magnitude of inconsistencies in the use-no use decision.
literature also uses OLS. Large selection effects are more likely if the service is expensive and the individual has some discretion with respect to timing or election of insurance coverage. Examples of such services include long-term psychotherapy and certain dental services. In these cases OLS results may be seriously biased away from zero.
III. SYSTEMATIC ERRORS IN THE OWN-PRICE VARIABLE

Most econometric studies are plagued by errors of measurement in the explanatory variables. Typically these errors are considered to be random and therefore independent of the explanatory variables. It is well known that in such a case coefficient estimates are inconsistent toward zero.¹

When

$$\text{plim}(1/n)(\Sigma Xu)$$

$$n \to \infty$$

is not zero (measurement error is not random), the plim of the estimated coefficient depends upon the value of $\lambda$ in Eq. (4). Three cases may be distinguished. First, if $1 \geq \lambda \geq 0$, the probability limit is less than the true coefficient. This includes the "standard" errors in the variables case, where $\sigma_{Xu}$ is zero. Second, if $\lambda > 1$, the sign of $b$ in the probability limit will be opposite to the sign of $\beta$. As can be verified by the second line of (4), this will occur only if $-\sigma_{Xu} > \sigma_X^2$—that is, if the error of measurement and the true explanatory variable have a negative covariance larger (in absolute value) than the variance in the true explanatory variable. Third, if $\lambda < 0$, the inconsistency is away from zero.² This occurs if $-\sigma_{Xu} > \sigma_u^2$—that is, if the measurement error and the explanatory variable have a negative covariance larger (in absolute value) than the variance of the measurement error. A graphical representation is given in Fig. 1. It is the possibility of negative covariance in existing empirical studies on which we concentrate in this section.

In the literature on demand for medical care, three examples of specification error can cause a negative covariance between measurement error and the price variable: (1) ignoring the presence of a deductible; (2) use of a dummy variable to indicate the presence of insurance coverage, together with a variable describing the gross (market transaction) price; and (3) estimation of price by dividing expenditure by quantity.

We have already shown how including those with deductibles could cause inconsistency away from zero because of possible correlation of the price variable with $\epsilon$ (the equation's error term). A second cause of inconsistency when specifying the marginal or average price in the presence of a deductible is that both measure the theoretically correct price with error; $\sigma_\epsilon^2 > 0$. Keeler, Newhouse, and Phelps (1977) show that the theoretically correct price variable for expenditures less than (below) the deductible depends upon the likelihood of exceeding the deductible at the time of use of services; hence, the theoretically correct price depends upon the

¹ It is less well-known, but can also be established, that $t$-statistics are inconsistent toward zero, although the direction of the inconsistency in standard errors cannot be established (Cooper and Newhouse, 1971). Methods of avoiding the problem are (1) to aggregate the data so that the proportion of the variance attributable to random measurement errors is reduced or (2) to use instrumental variable techniques to remove measurement errors or transitory component from the explanatory variables. Studies using aggregate data have the first method built in, whereas many survey data studies rely upon the latter method.

² As a special case, $\lambda$ may equal zero, and there is no inconsistency.
distance remaining to the deductible and the time remaining in the accounting period at the time of use.\(^3\) Because there is a nonzero probability of exceeding the deductible later (the individual's health status may or may not change), the theoretically correct price for illness episodes below the deductible is less than the marginal and average price. Marginal and average price measure the theoretically correct price with systematic error; the sign of \(\sigma_{xu}\), hence the magnitude of \(\lambda\), depend upon the distribution of observations above and below the deductible. Therefore, the direction of the inconsistency from using the incorrect price variable cannot be signed a priori.

With available data it is impossible to measure the theoretically correct price variable; therefore one cannot determine the magnitude of the inconsistency when we analyze a sample that includes people with deductibles in their policies. Data are necessary that show the distance remaining to the deductible and the time remaining in the accounting period at the time of each expenditure decision; exist-

\(^3\) An analogous problem to the deductible exists with respect to upper limits in the insurance policy, but we doubt that it is of much practical importance (because of the low likelihood of exceeding the upper limit) and so we ignore it.
ing studies analyze annual expenditure upon medical services. Again data from the Health Insurance Study should help in remedying this problem.

The second type of specification error that leads to negative covariance between \( X \) and \( u \)—use of a dummy variable to represent insurance—has arisen from lack of information about the terms of the insurance policy. The error occurs in the literature when the percentage insured in an area (such as a state) and the gross price are used as variables to explain demand in an area; the price elasticity is then inferred from the coefficient on the gross price variable. (See Davis and Russell, 1972; M. Feldstein, 1971, 1977; Rosenthal, 1964.) Underlying these aggregated observations is an individual demand curve with a dummy variable that takes the value one if the individual has insurance, together with a measure of the gross price paid.

Such a specification assumes that all those with insurance have the same kind of insurance. Because individuals have different kinds of insurance, this specification involves an error that can be readily seen to have a negative covariance with the true coinsurance rate. We show in Appendix C that this negative covariance between the insurance dummy and the measurement error will in general cause the estimated coefficient of the gross price variable (used to estimate the price elasticity) to be inconsistent. Whether the inconsistency is away from or toward zero cannot be determined a priori.

We have therefore used empirical methods to determine the direction and magnitude of the inconsistency and have estimated two alternative specifications for hospital length of stay and nonsurgical physician visits. The first specification uses the marginal price, which we believe to be the theoretically appropriate specification (given that individuals with deductibles are excluded from the sample). Marginal price is defined as the hospital coinsurance rate times the gross room and board charge in the length of stay equation, and the physician coinsurance rate times physician visit charge in the physician visit equation. The second specification uses the gross price and a dummy variable that takes the value one if the consumer has any insurance. The results are shown in Table 3.

The coefficient of the gross price variable is larger by a factor of three in the case of hospital length of stay and by more than one-third in the case of physician office visits, so the misspecification appears important.

Table 4 lists some examples of estimated price elasticities for hospital length of stay in the literature. The \(-0.23\) value we derive is consistent with estimates in the literature for a similar specification. The estimates in the literature appear sensitive to specification, particularly the Davis-Russell estimates, which exhibit a sign reversal when the specification is changed.

---

4 M. Feldstein (1971, 1977) uses a variant of this specification in which the price variable approximately equals the gross price for state \( i \) in time \( t \) times the percentage with insurance in state \( i \) in time \( t \) times the national average coinsurance rate in time \( t \). Across states in any one year, then, this specification is identical to the specification just outlined. Across time, however, there is additional variation caused by changes in the average national coinsurance rate. Note that our result of inconsistency away from zero in the coefficient of the gross price variable (see below) may explain why Feldstein finds a markedly higher elasticity with respect to gross price than with respect to the coinsurance variable as he defines it, when he permits the elasticities to be unequal.

5 For uninsured individuals there is no error. For insured individuals the error is \( 1 - c \) where \( c \) is the coinsurance rate.

6 We omit the price of ancillary services for lack of data. Because ancillary services are typically few on the marginal day, this omission should not affect our estimates.
Table 3

**PRICE ELASTICITIES AT MEAN FOR HOSPITAL LENGTH OF STAY AND PHYSICIAN VISITS**

<table>
<thead>
<tr>
<th>Specification</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hospital Length of Stay</td>
</tr>
<tr>
<td>Marginal (and average) price</td>
<td>-.065 (t = -1.87)</td>
</tr>
<tr>
<td>Gross price (with insurance dummy included)</td>
<td>-.232 (t = -1.41)</td>
</tr>
</tbody>
</table>

*The mean coinsurance rate for length of stay is 0.3 and for office visits is 0.9.*

Table 4

**EXAMPLES OF ESTIMATES OF LENGTH OF STAY ELASTICITIES IN THE LITERATURE**

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Feldstein (1971)</td>
<td></td>
</tr>
<tr>
<td>Specification with income, population density, Medicaid dummy</td>
<td>-.25</td>
</tr>
<tr>
<td>Specification with income, population density, Medicaid dummy, physician/population, proportion of physician’s GPs</td>
<td>-.39</td>
</tr>
<tr>
<td>Davis-Russell</td>
<td></td>
</tr>
<tr>
<td>Specification using inpatient revenue per patient day as price</td>
<td>-.24</td>
</tr>
<tr>
<td>Specification using room charge as price variable</td>
<td>+.35</td>
</tr>
<tr>
<td>Rosenthal</td>
<td></td>
</tr>
<tr>
<td>1960 data</td>
<td>-.22</td>
</tr>
<tr>
<td>1950 data</td>
<td>-.49</td>
</tr>
</tbody>
</table>

*Estimates are from M. Feldstein (1971), Eqs. (18) and (28); Davis-Russell, Table 3, Eqs. (1) and (3); and Rosenthal, Appendix Table 1. Rosenthal's estimates using a linear rather than a log-arithmetic functional form are -.27 and -.23 in the two years.*
A third type of specification error occurs because price is not observed in some data, although expenditure and quantity data are observed. In this case a number of researchers (Fuchs and Kramer, 1972; M. Feldstein, 1970; P. Feldstein and Severson, 1964) have estimated price by dividing expenditure by quantity, on the grounds that

\[ \frac{E}{x} = \frac{(p \times x)}{x} = p, \]  

(5)

where \( E \) is expenditure, \( x \) is quantity, and \( p \) is price. If there are measurement errors in \( x \) but not \( E \), the estimated price variable will contain errors that are negatively associated with the true price variable.

Suppose the true model (as assumed in the studies cited), was

\[ \ln(x) = \alpha - \beta \ln(p) + \epsilon, \]  

(6)

where \( x \) is quantity and \( p \) is price. If there are errors in observing quantity, and price is estimated from expenditure, we show in Appendix D that there is an inconsistency away from zero in estimates of \( \beta \) as long as \( |\beta| < 1 \), as almost all empirical estimates show.\(^7\)

A solution to this problem exists if \( E \) is measured with negligible error. Then negligible inconsistency results if \( E \) is used as an explanatory variable rather than \( E/x \), as we shall prove.\(^8\)

Because \( \ln(p) = \ln(E) - \ln(x) \), substituting for \( \ln(p) \) in (6) yields:

\[ \ln(x) = [1/(1 - \beta)](\alpha - \beta \ln(E) + \epsilon) = \alpha' - \gamma \ln(E) + \epsilon'. \]  

(7)

If there is negligible error in measuring \( E \), the estimate of \( \gamma \) is approximately consistent, and \( \beta = \gamma/(\gamma + 1) \).\(^9\) Any remaining coefficients must be multiplied by \( (1 - \beta) = 1/(\gamma + 1) \) to obtain consistent estimates of their value in an equation where "price" was the explanatory variable.

If expenditure as well as quantity is measured with error, then the problem is more complicated. From the relationship \( \beta = \gamma/(\gamma + 1) \), we know that \( \partial \beta / \partial \gamma = 1/(\gamma + 1)^2 > 0 \). Hence, an inconsistency in \( \gamma \) (due to measurement error in \( E \)) toward zero will also create an inconsistency in the estimate of \( \beta \) toward zero. The estimates given by this method and by a regression of \( x \) on estimated \( p \) will bracket the true value.

The problem of systematic error in the price variables is ubiquitous in demand studies now in the econometric literature. Although we offer no way to escape these problems given the data used in such studies, we think it is well worth noting the probable effect on estimates of price elasticity.

\(^7\) We also require that total expenditure be measured with less error than quantity (see Appendix D). This appears likely in the cases cited above.

\(^8\) Fuchs and Kramer (1972) use two-stage least squares, which eliminates the problem (asymptotically) if the instruments are correlated with the true variable and not with the error. This assumption is plausible when the measurement error is random, but when the error is correlated with the explanatory variable it is a considerably riskier assumption. Further, their sample contains only 38 observations, so asymptotic properties may be of little benefit. Zellner (1976) has proposed minimum expected loss estimators for simultaneous equation estimation with finite samples, which appear appropriate for such data.

\(^9\) Because \( \beta = \gamma/(\gamma + 1) \), hypothesis tests on \( \gamma \) transform directly into tests on \( \beta \). For example, if \( \gamma_1 \) and \( \gamma_2 \) are 95 percent confidence limits on \( \gamma \), then \( \beta_1 = \gamma_1/(\gamma_1 + 1) \) and \( \beta_2 = \gamma_2/(\gamma_2 + 1) \) are also 95 percent confidence limits for \( \gamma \).
IV. PROBLEMS CAUSED BY AGGREGATION

AGGREGATION ACROSS SERVICES

An empirically very significant problem occurs if expenditures on medical services having different kinds of insurance coverage are aggregated, and a coinsurance rate averaged across the services is used to explain demand.

Define $E$ as total expenditure across a number of services; $B$ as insurance benefits across all services; and $C$ as $1 - (B/E)$, so that $C$ is the average coinsurance rate. Let lower-case subscript letters represent the corresponding variable for the $i$th service, and suppose the true relationships are $e_i = a_i - b_i c_i + v_i$, where $v$ is an error term with the usual properties, and $a_i$ and $b_i$ are constants.

In practice some services are covered by insurance and others are not. One would expect that services with higher mean expenditure and variance of expenditure would have greater coverage (Phelps, 1976). In fact, covered services tend to increase as a proportion of total expenditure as total expenditure rises (see Table 5). Therefore, estimates of the relationship between $E$ and $C$ will be spuriously large (relative to the true relationships between the $e_i$ and the $c_i$).

To illustrate the problem consider the simple case where all expenditures are random ($b_i = 0$ for all $i$), and all individuals have complete insurance against hospital expenditures but no insurance against ambulatory expenditures. Suppose a hospital expenditure, if it occurs, costs on average $1000 and that the 10 percent of the population who are hospitalized use on average $100 worth of ambulatory services. The other 90 percent of the population on average use $20 worth of ambulatory services. If $E$ is regressed on $C$ (linearly), the expected elasticity will be about $-8.5$ at the means of the two variables, when in fact the true elasticity is zero.\(^1\)

To estimate the magnitude of the problem caused by aggregation across services, we used the same 1963 survey data on which our previous tables are based. We defined $E$ to be total hospital and physician expenditure and regressed $E$ on $C$. The results are shown in Table 6. The estimated values of $-2.1$ in Table 6 should be compared with expenditure-weighted averages of the elasticities for the individual services. With the elasticities estimated in Table 3 as well as those we have estimated elsewhere (Phelps and Newhouse, 1974), such a weighted average would be an order of magnitude smaller than the elasticities in Table 6.\(^2\) Thus, aggregation of hospital and physician expenditures is likely to produce highly misleading estimates.

The problem of aggregation across services occurs in the study by Rosett and Huang (1973). Although their methodology was complicated because of data limitations, in principle they attempted to regress $E$ on $C$, where $E$ included both hospital

\(^1\) 10 percent of the observations cluster around $C$ of 0.1 and $E$ of 1100, and the other 90 percent cluster around $C$ of 1.0 and $E$ of 20. The expected value of the coefficient of $C$ will be approximately 1200. The mean of $C$ is 0.91 and of $E$, 128.

\(^2\) Phelps and Newhouse (1974) present evidence that the elasticity of total hospital and physician expenditure is about 0.1, as measured by premiums charged for plans with coinsurance rates ranging from 0.25 to 0.10.
Table 5  
PERCENT OF EXPENSES COVERED BY HEALTH INSURANCE BENEFITS, COMPARED WITH AVERAGE EXPENSE PER PERSON, 1963

<table>
<thead>
<tr>
<th>Service</th>
<th>Average Expenditure for Those with Positive Expenditure</th>
<th>Benefit as Percentage of Expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital</td>
<td>$223</td>
<td>69</td>
</tr>
<tr>
<td>Surgery</td>
<td>200</td>
<td>58</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>142</td>
<td>32</td>
</tr>
<tr>
<td>MD office visits</td>
<td>35</td>
<td>13</td>
</tr>
<tr>
<td>Prescription drugs</td>
<td>36</td>
<td>1</td>
</tr>
</tbody>
</table>

SOURCE: Andersen and Anderson (1967).

and physician expenditures (see Appendix A for further discussion of the Rosett and Huang study). Rosett and Huang’s estimated elasticities at the mean of C and E are quite large. Their specification differed from ours in that it excluded supply variables; such an exclusion affects the estimated elasticities negligibly (Table 6). Note also that the apparent precision of the estimates is high, although this precision is around a biased mean.

Although we have discussed the aggregation problem in the context of hospital and physician expenditures, the problem of aggregation across services occurs in other instances as well. For example, as Table 5 shows, surgeons’ services are covered more extensively by insurance than physician office visits; surgical expenditures are also considerably larger, on the average. Aggregation of surgical and

Table 6  
COINSURANCE ELASTICITIES OF TOTAL PHYSICIAN AND HOSPITAL EXPENDITURE AT MEAN USING A COINSURANCE RATE AVERAGED ACROSS SERVICES⁹

<table>
<thead>
<tr>
<th>Condition</th>
<th>Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Including supply variables</td>
<td>-2.13</td>
</tr>
<tr>
<td></td>
<td>(t = -15.63)</td>
</tr>
<tr>
<td>Excluding supply variables</td>
<td>-2.14</td>
</tr>
<tr>
<td></td>
<td>(t = -15.73)</td>
</tr>
</tbody>
</table>

⁹ There were 1760 observations. Exclusions were similar to those described in the footnote to Table 1; 78 additional observations were added to the sample of 1682. These 78 individuals had hospital expenditure but no ambulatory expenditure (and so were excluded from the ambulatory care analysis in Table 1). Those with no expenditure were excluded because no meaningful average coinsurance rate could be computed. The Rosett and Huang study did not exclude them, because Rosett and Huang estimated the coinsurance rate from the premium paid. However, there should be no effect from excluding these individuals if the Rosett and Huang specification were correct.

A problem analogous to computing price by dividing expenditure by quantity arises in our calculation if benefits are measured with error. However, the problem, if any, is likely to be of little quantitative importance (see Appendix D, Eq. (D.6) ff.).

Supply variables are physician/population and bed/population ratios.
nonsurgical physician expenditures, therefore, will cause a problem similar to aggregation of hospital and physician services. However, services of psychiatrists tend to be less well covered by insurance but associated with large expenditures; therefore if psychiatric services were aggregated with other physician services, the measured elasticity could be too small.

THE PROBLEM OF INSTABILITY

Coefficient estimates are likely to be unstable when we aggregate across either different types of medical services or different individuals.

Suppose demand curves for two services or two individuals are:

\[ q_1 = a_0 + a_1 p_1 + u_1 \]  \hspace{1cm} (8a)

\[ q_2 = b_0 + b_1 p_2 + u_2 \]  \hspace{1cm} (8b)

where \( q \) is expenditure on separate services or by different people and the \( p_i \) are their prices. These demand curves are aggregated as follows:

\[ Q = A_0 + A_1 (p_1 + p_2) + U, \]  \hspace{1cm} (9)

where \( Q = q_1 + q_2, A_0 = a_0 + b_0, A_1 = (a_1 p_1 + b_1 p_1)/(p_1 + p_2) \), and \( U = u_1 + u_2 \). The coefficient \( A_1 \) is said to represent the average response of demand to price. As is well known, \( A_1 \) will be stable if \( a_1 \) equals \( b_1 \), or if \( p_1/(p_1 + p_2) \) is constant.3

When one is aggregating medical services these conditions do not appear to exist. Price elasticities of services differ in general (Phelps and Newhouse, 1974; Manning and Phelps, 1978). The stability of the price share is also doubtful if health insurance legislation is enacted. The appropriate price is net of insurance, and insurance changes caused by a public program are likely to be very different across services, because components of the medical bill are currently insured at vastly different levels. For example, hospital services are well insured, but prescription drugs are insured for only a small fraction of the population (B. Cooper and Worthington, 1973; see also Table 5). Much of the importance of estimating elasticities is to obtain accurate predictions of demand under alternative health insurance legislation, yet it is precisely such legislation that will cause changes in relative prices.

Similar problems arise when demand for the same service or commodity is aggregated across individuals—for example, if data from states or other geographic areas are used as units of observation. Predictions from such equations concerning the effect of a universal plan are not likely to be stable, because the distribution of insurance across people currently varies considerably (Phelps, 1976). A universal plan would introduce little, if any, added insurance for many parts of the population (those with "complete" private insurance, and those on Medicare and Medicaid), and others would receive substantial increases in the coverage (most notably, those poor now ineligible for Medicaid). Thus, the distribution of coverage across in-

3 Theil (1954). By "stable" we mean the structure will be identical in all future periods. Zellner (1969) has shown that these conditions can be relaxed to permit \( a_1 \) and \( b_1 \) to be drawn from the same distribution. This relaxation is of no importance for this problem. The generalization to a more than two-service case is straightforward.
individuals would certainly shift under universal insurance. As a result, estimated elasticities will be stable only if individual responses are uniform. Little evidence is available about uniformity of response across persons, but the assumption appears to be a heroic one. Manning and Phelps (1978) show that price elasticities for dental services differ substantially across demographic groups.

An example of instability possibly caused by aggregation can be observed in Feldstein (1977), who uses states as units of observation; his estimated price elasticities for two subperiods differ markedly, as shown in Table 7. Although it is possible that the true coefficients could differ by as much as shown in the table, another plausible explanation is instability from aggregation. Thus, elasticities estimated from data aggregated across individuals could well be unstable predictions of demand under a national health insurance program.

<table>
<thead>
<tr>
<th>Period</th>
<th>Price Elasticity of Hospital Length of Stay</th>
<th>Price Elasticity of Hospital Admissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1959-1965</td>
<td>-.080</td>
<td>-.042</td>
</tr>
<tr>
<td></td>
<td>(t = -2.00)</td>
<td>(t = -0.95)</td>
</tr>
<tr>
<td>1966-1973</td>
<td>+.005</td>
<td>-.236</td>
</tr>
<tr>
<td></td>
<td>(t = 0.17)</td>
<td>(t = -3.11)</td>
</tr>
<tr>
<td>1959-1973</td>
<td>+.028</td>
<td>-.099</td>
</tr>
<tr>
<td></td>
<td>(t = 1.22)</td>
<td>(t = -4.13)</td>
</tr>
</tbody>
</table>

SOURCE: Feldstein (1977), Table 1, Eqs. (1)-(6).

ARE DATA AGGREGATED ACROSS INDIVIDUALS APPROPRIATE?

We have just argued that aggregated data pose problems in estimating stable demand elasticities. However, there is an argument to the contrary that aggregation is "not necessarily bad" (Ginsburg and Manheim, 1973). This argument goes as follows: Physicians rarely know the insurance coverage of an individual patient. Rather, they know the average insurance coverage of their patients (the norm). Therefore, they plan their treatment according to the average coverage level. If the average coverage level changes, treatments will change accordingly, but the variation in demand one observes in a cross-section of patients with different insurance coverages understates the variation that would occur if the average coverage changed. Because health insurance legislation would change the average insurance coverage, estimates based on variation across individuals underestimate the change in demand such legislation would cause. We refer to this as the "norms" argument; two of us have shown elsewhere that it lacks convincing empirical support (Newhouse and Marquis, 1978).
One basis for the norms argument is that higher elasticities are estimated from aggregated data than from individual data (Ginsburg and Manheim, 1973). However, the estimates in the literature using aggregate data are based on important misspecifications, so there is an alternative explanation for the discrepancy.
V. OMISSION OF THE CROSS-PRICE VARIABLE AND USE OF NOMINAL PRICE VARIABLES

Two specification errors commonly made in the literature have little effect on the estimates made from the 1963 survey data. The first error is the omission of a cross-price variable. With two exceptions (Davis and Russell, 1972; Newhouse and Phelps, 1976) all studies cited in Section II use only an own-price variable in the demand function. To explain the demand for hospital days, for example, no account is taken of the price of physician services. This is a straightforward case of an omitted variable; as is well-known, the inconsistency from omitting such a variable equals the true coefficient of the omitted variable times the coefficient estimated in an auxiliary regression in which the omitted variable is regressed on the included variables (Theil, 1961, pp. 326–327).

In our data set, cross-price is nearly uncorrelated with own-price; its omission does not affect the estimated own-price coefficient, as may be seen by comparing the first two lines in Table 8. This lack of correlation may not be found in other data sets, but other results suggest that the true coefficient of the cross-price variable may be near zero, so that omission of a cross-price variable may not seriously affect the coefficient of the own-price variables (Lewis and Keairnes, 1970; Hill and Veney, 1970).

Davis and Russell (1972) have argued that cross-price is important for hospital admissions, although not for length of stay. Although they have not deduced this

<p>| Table 8 |
|-----------------|-----------------|-----------------|
| Price and Income Elasticities with Alternative Specifications |</p>
<table>
<thead>
<tr>
<th>Hospital Length of Stay</th>
<th>Annual Physician Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price elasticity with cross price, price in real terms</td>
<td>-.065</td>
</tr>
<tr>
<td>(t = -1.87)</td>
<td>(t = -2.40)</td>
</tr>
<tr>
<td>Price elasticity without cross price</td>
<td>-.066</td>
</tr>
<tr>
<td>(t = -1.90)</td>
<td>(t = -2.60)</td>
</tr>
<tr>
<td>Price elasticity, price in nominal terms</td>
<td>-.065</td>
</tr>
<tr>
<td>(t = -1.85)</td>
<td>(t = -2.34)</td>
</tr>
<tr>
<td>Wage income elasticity, income in real terms</td>
<td>.013</td>
</tr>
<tr>
<td>(t = .25)</td>
<td>(t = 1.01)</td>
</tr>
<tr>
<td>Wage income elasticity, income in nominal terms</td>
<td>.014</td>
</tr>
<tr>
<td>(t = .26)</td>
<td>(t = 1.02)</td>
</tr>
<tr>
<td>Nonwage income &lt;$3000, elasticity in real terms</td>
<td>-.026</td>
</tr>
<tr>
<td>(t = -.80)</td>
<td>(t = -.51)</td>
</tr>
<tr>
<td>Nonwage income &lt;$3000, elasticity in nominal terms</td>
<td>-.025</td>
</tr>
<tr>
<td>(t = -.76)</td>
<td>(t = -.52)</td>
</tr>
</tbody>
</table>
argument from maximizing behavior, their results strongly support it. If true, the result is important because it implies that extending coverage for ambulatory services (which many individuals do not now have) would reduce the demand for hospital services. The result, however, is contained in an equation that misspecifies the insurance variable. Because the argument made in Appendix C about inconsistency in the own-price coefficient from misspecification applies to the cross-price coefficient as well, the Davis-Russell result is suspect.

The experimental data of Hill and Veney (1970) and Lewis and Keairnes (1970) do not support the Davis-Russell argument; there is no significant effect of additional outpatient coverage on either admissions or length of stay.1 Because these data must be taken as more reliable on this particular point, the large cross-price elasticity for hospital admissions found by Davis and Russell is probably an artifact caused by the misspecification of the insurance variable.

Another kind of error is made by studies using cross-section data that specify price in nominal terms; that is, they assume that the price of other goods and services is similar across areas. Manning and Phelps (1978) derive expressions for the bias in price and income elasticities from using nominal prices rather than real prices. They show that expenditure elasticities will be biased away from zero if the sum of income and price elasticities is less than one, as appears to be the case empirically (see Table 8). Quantity elasticities will be biased away from zero if the price elasticity is greater than the income elasticity (in absolute value); this also appears to be the case (see Table 8). But in the 1963 survey data, the amount of the bias was very small; estimates using real price and real income are indistinguishable from those using nominal price and income variables.

---

1 Helms, Newhouse, and Phelps (1978), however, find that copayments among a Medicaid population may have raised hospital days, consistent with the Davis-Russell argument. This result may indicate the Medicaid population differs from the general population or may represent a statistical artifact for reasons discussed in that paper.
VI. CONCLUSION

We have presented elsewhere (Phelps and Newhouse, 1974) what we regard as approximate demand elasticities in the range of coinsurance from zero to 25 percent. These estimates use various data sources and different techniques from the studies cited here, and they show an aggregate demand elasticity (a weighted average of component-service elasticities) of \(-.1\) in the coinsurance range of 25 percent to zero. We believe these estimates are more accurate for this range than the ones reviewed here; specifically the reviewed estimates seem too large. A nibble or two of Alice's cake might be in order for them.

The estimates in the studies cited above generally came from a different range of coinsurance than 25 percent to zero. Although most authors of those studies do not give the distribution of their insurance variable, it is almost surely from a higher level of coinsurance.\(^1\) However, it is unlikely that a different range is the main reason for the discrepancy, because the distribution of coinsurance in the data used in this report is probably quite similar to the distribution in the data used by the studies cited.

The problems encountered in the studies we have reviewed are due to the restrictive nature of available data and cannot be remedied short of gathering more appropriate data. Toward that end, we have undertaken a social experiment in health care financing (Newhouse, 1974). We believe that this experiment will generate the necessary data to compute more precise estimates of the effects of insurance on demand for medical care.

\(^1\) Such a distribution would be helpful in appraising these estimates.
APPENDIX A

SPECIFIC PROBLEMS IN THREE STUDIES

ROSETT AND HUANG (1973)

We have already alluded to the large inconsistency introduced into the Rosett and Huang study by aggregation across services. In addition, a number of other factors in Rosett and Huang's methodology lead to overestimation of the price elasticity. Because they exercise considerable ingenuity to overcome severe data restrictions, it is helpful to reconstruct their methodology.

Available in their survey data are measures of out-of-pocket insurance premiums and out-of-pocket medical expenses. They attempt to construct proxies for k (coverage ratio) using these variables. In brief form, their procedure involves the following:

1. List observations by size of out-of-pocket premiums and divide into small groups in the ordered list. Families within a given group will have similar reported out-of-pocket premiums.
2. Estimate average medical expenses within each group (of 100 families) using the formula $M = D + R/1.17$, where $M$ is estimated expenses, $D$ is direct (out-of-pocket) medical expenses, $R$ is out-of-pocket premiums, and .17 is the average loading fee.
3. Estimate a preliminary k (the complement of the coinsurance rate) for each family using observed $R_i$, so that $k' = R_i/(1.17M)$, where $M$ is taken from the group of 100 families.
4. Correct preliminary estimates by an adjustment factor so that (summed over all groups) the insurance companies paid out, on average, benefits B, such that $R = 1.17B$. This gives $k^*$, for each family.
5. Use a curve-fitting technique to smooth out extreme points, by regressing $k^*$ on $R$, $R^2$, ..., $R^8$. (This equation had $R^2 = .93$.)

This procedure has a variety of defects that cause serious systematic errors in their estimates of k. First, it ignores any employer contributions toward health insurance premiums, so families with large employer contributions will be observed to have little coverage. Their demand will be high, because they have "good" insurance coverage, but those high demands will be attributed to a small k. The effects of k on demand will be overstated for them. In other words, there is a systematic negative covariance between the measurement error in k and the value of k; families with high k (from employer contributions) will be observed with low k. Hence, there is negative covariance established in the data set, which, as we have seen, may lead to upward inconsistencies in price elasticity estimates.

Some evidence is available on the extent of employer payment for insurance. In a 1963 household survey conducted by the Center for Health Administration

---

1 U.S. Department of Labor, Consumer Expenditures and Income: Survey Guidelines, Bulletin 1684, Washington, D.C., 1971. Although data were collected on employer contributions to insurance (see question 8, p. 153 of Survey Guidelines), those data were not included in any calculations of medical insurance amounts (pp. 209, 214).
Studies of the University of Chicago, detailed information on insurance policies was obtained from the insurers, so that premiums and coverage amounts were verified and explicit. These data show that (in 1963) over 27 percent of the families had some employer contributions toward their insurance payments; and, among those, the average employer contribution was $113. The average out-of-pocket premium in the entire sample was $71 ($140 for those families with positive verified purchases); the average total premium, including employer subsidy, was $102 ($173 for those families with positive verified purchases). Rosett and Huang’s data come from approximately the same time, 1960. This suggests that Rosett and Huang’s procedure measures a substantial proportion (roughly one-third) of the insurance with a systematic error. Note that those whose entire insurance policy is paid for by employers simply are treated as having no insurance and enter a separate estimation (which is then presumably plagued by an omitted variable problem, leading to overestimates of the income elasticity in that no-insurance equation, because the income elasticity of demand for insurance is positive).

A second problem in the Rosett and Huang study is their assumption that all insurance policies in the population had the average loading fee (17 percent). This supposition leads to inconsistency in the estimation of k across groups. Rosett and Huang themselves recognize the crucial nature of this assumption when they state, “The fundamental assumption on which our estimates of M rest is that if two households pay the same premium, they have the same level of coverage, k." They are forced to make this assumption because their data source has no evidence on the type or source of insurance.

But there are well-known differences in the loading fee. Group insurance, for example, is sold at a substantially lower loading fee than is individual insurance, and group policy loadings vary systematically by size of group. Insurance companies file their automatic loading fee reduction schedules with insurance commissioners in many states, and these are published in many texts on health insurance. This in itself is a major problem (those who had large loadings would have smaller ks than Rosett and Huang believe and those with very small loadings would have larger ks, violating what Rosett and Huang regard as a “fundamental assumption” of their work). But the problem is even worse: It has been shown empirically that demand for insurance is such that premiums rise as the loading fee falls, both in household survey cross-section and in the time series estimates (Phelps, 1973), so that those in Rosett and Huang’s low-premium groupings (recall that they group families by premium size) systematically paid a higher load than those in high premium groups. Assuming that all persons faced the same loading fee tends to overstate the coverage of those with small amounts of premiums and to understate the coverage of those with high amounts of insurance premiums. This is another systematic negative covariance between the error in k (their key explanatory variable) and the size of k itself. As was shown in the text, this may create an inconsistency in their coefficient away from zero.

---

2 These data are reported in Phelps (1973), Appendix F.
2 Rosett and Huang, p. 284.
4 Note that although Rosett and Huang find a high price elasticity of demand for medical care, they are willing to operate with the implicit assumption that demand for insurance has zero price elasticity. See Phelps (1973) for some typical loading-fee schedules for different work group sizes. Phelps estimates an elasticity of premiums with respect to the loading fee of approximately — .3.
Rosett and Huang carry out a simulation to verify their methodology. However, their simulation does not deal with any of these problems, so their simulation results (showing that their coefficients might be inconsistent toward zero, if anything) have little or no bearing on the true inconsistencies in their coefficients. Their simulation data set is constructed under the assumptions of their original data: that the loading fee is uniformly .17 for all insurance, that the price elasticity of demand for insurance is zero, and that there are no unobserved insurance premiums omitted for some elements of the population. Their procedure for generating the simulation data set is straightforward, and one could analytically derive their results (that the coefficient is inconsistent toward zero) because they have built into the system a random error in the explanatory variable. That they constructed a data set with a random error and found the analytically predictable result provides no evidence on the effects of the systematic errors of measurement that lead to the inconsistencies we have discussed.

A second class of problems arises when Rosett and Huang attempt to reconstruct the dependent variable from their estimated explanatory variable. That is, they compute

\[ \hat{M}_i = D_i/(1 - \hat{k}_i). \]

(We have adopted the convention of “hatting” all variables estimated or computed by Rosett and Huang.) We have demonstrated that there are multiple opportunities to introduce measurement error into Rosett and Huang’s estimate of \( \hat{k} \). Error in \( \hat{k} \) introduces further error into their estimates of the effects of \( k \) on \( M \). To demonstrate this, we compute a simple elasticity of change of \( \hat{M} \) to \( k \), which is not a behavioral elasticity (not a demand elasticity) but simply a transformation elasticity (because \( \hat{M} \) is simply a transformation of \( \hat{k} \), given \( D \)).

\[ \frac{\partial \hat{M}_i}{\partial \hat{k}_i} \times \hat{k}_i \times \hat{M}_i = \frac{D_i}{(1 - \hat{k})^2} \times \hat{k}_i/\left[D_i/(1 - \hat{k}_i)\right] = \hat{k}_i/(1 - \hat{k}_i). \]

Any decrease in the estimated \( \hat{k} \) will reduce the estimate of \( \hat{M} \). But if the “transformation elasticity” were uniformly less than +1.0, there would be less inconsistancy in their estimate of the effects of \( k \) on \( M \) than we have implied, because any (proportional) decrease in \( \hat{k} \) would similarly decrease \( \hat{M} \). For values of \( \hat{k} \) below 0.5, the transformation elasticity is less than unity, approaching zero as \( \hat{k} \) approaches zero. And for \( \hat{k} \) above .5, the elasticity is greater than unity (approaching \( \infty \) as \( \hat{k} \) approaches 1.0) so that when there is an upward error in \( \hat{k} \) (placing it too high by measurement error), it tends to disproportionately raise the estimated value of \( \hat{M} \). Therefore, unless all of their estimated \( \hat{k} \) were identically .5 (which would make demand curve estimation a trifle difficult), or unless there are no errors of measurement in \( \hat{k} \) (which we can reject conclusively), there still remains an upward inconsistency in their estimation technique. Apparently, the majority of Rosett and Huang’s estimated \( \hat{k} \)s fall into a range below .5, so that (although the estimated \( \hat{k} \)s are too small systematically) the proportional reduction in the \( \hat{M} \)s is less than in the \( \hat{k} \)s. Although this will tend to reduce their inconsistency somewhat, it cannot compensate completely for the inconsistencies.

* For no resultant inconsistency in the coefficient estimate, they would have to be using a constant elasticity form, which they cannot, with observed expenses equal to zero for some observations.
generated by the errors in the estimated \( \hat{k} \), because the transformation elasticity is less than unity for these "low \( \hat{k} \)" observations.\(^6\)

Finally, two other specific problems in their estimation affect their results.

1. They have used observed income rather than the theoretically desirable permanent income, leading them to understate the income elasticity. Because insurance demand is positively correlated with permanent income, the true \( k \) is positively correlated with permanent income. But because Rosett and Huang’s estimated income elasticity is inconsistent downward, their estimate of the effects of \( k \) on \( M \) will have a further upward inconsistency.

2. They chose to estimate two separate demand functions (for those with and those without insurance) without giving their reasoning behind this split. One could easily combine these two equations, simply assigning a value of \( k = 0 \) for those without insurance. The effects of this split are not clear \emph{a priori}.

**FELDSTEIN (1970)**

This study contains a number of the problems discussed in our text, notably use of the average coinsurance rate and aggregation across individuals. The study is of importance because of the significance that Feldstein himself (1974) attaches to it. Briefly, Feldstein attempted to estimate supply and demand curves for physician services for 19 consecutive years. He found "wrong" signs on price and insurance variables in his demand equations (and a backward bending supply curve) and concluded that a demand equation could not be estimated because permanent excess demand had prevailed in the market for physician services. It is always somewhat dangerous to draw inferences from negative results, and this is particularly true when there are only 19 observations.

An alternative and simpler explanation of Feldstein’s result is that his demand equation is underidentified. The variables included in the demand equation are: average price, insurance, consumer price index, median income, government medical services per capita, and time.\(^7\) Feldstein treats insurance as exogenous, but it is properly considered endogenous (see Appendix E). Compare this specification with the specification of the supply equation, which includes average price, consumer price index, “reference” income, inputs, time, and government medical services per physician.\(^8\) If we ignore the insurance variable because it should be endoge-

\(^6\) From Rosett and Huang’s Table 2, we see that 1357 estimates of \( \hat{k} \) fall into categories above .6, 4341 fall between .4 and .6, and 2589 fall below .4. This distribution is skewed toward low estimated values of \( \hat{k} \), so their estimated coefficient for \( \hat{k} \) in the demand equation appears to be inconsistent upward. Indeed, Rosett and Huang have only 13 cases where the estimated \( \hat{k} \) is above .80, but the average expenditure for those cases is $2257. (In practice there are hardly any marginal coverage ratios less than .75 and greater than zero.) There are four cases of high income (above $11,500) that have estimated \( \hat{k} \) above .80, and the average expenditure for those four families was $6008, with a standard deviation of $7339. This means that a family of two with an imputed \( k \) above .8 received an imputed expenditure above $10,000 or $15,000. It appears that their estimates of expenditures for this high \( k \) group are abnormal and are sensitive to these one or two observations. Those observations tend to greatly overstate the interaction between \( k \) and response to \( k \) (their quadratic term) and the interaction between \( k \) and income (which they estimate to be very high).

\(^7\) Equations D3 and D4, Table 1.

\(^8\) Equations S1 and S3, Table 3.
nous, there are three differences. Reference income (the 95th percentile of the income distribution) is used rather than median income, but the two variables have a correlation of .994 over the period spanned by Feldstein's data. The government medical services per capita variable is almost perfectly correlated with government medical services per physician, because the physician to population ratio for this period was nearly constant (the coefficient of variation is .014). That leaves a measure of physician inputs as the one excluded "exogenous" variable in the demand equation. It should not be considered exogenous in the supply equation; moreover, it was generally insignificant in the supply equation and excluded from two specifications.

Thus, the demand equation is underidentified. Feldstein's footnote 30 seems to indicate that he also believes "the demand parameters are underidentified." If so, nothing can be inferred about demand from his results. A number of other criticisms of the Feldstein methodology are made by Fuchs and Kramer (1972).

FUCHS AND KRAMER (1972)

Fuchs and Kramer use two different approaches to the problem of measuring the responsiveness to price of physician services. The first is to divide the 1948–1968 period into three subperiods and look at rates of growth of a number of variables in these subperiods. Although their analysis is somewhat speculative in light of the approximations necessary to derive the rate of growth of the variables, it is a plausible interpretation of time series data. It does not, however, shed much light on the size of insurance elasticities. That subject is more explicitly treated in their second approach, which is to explain variation in physician visits across 33 states in 1966. When they do so, their estimated elasticities are in the 

- .15 to 

.35 range.

We have already commented on the use of an average coinsurance rate, state aggregate data, and expenditure divided by quantity to obtain a price variable. Use of the average coinsurance rate may lead to inconsistent estimates. Another aspect of the Fuchs and Kramer methodology creates an inconsistency toward zero in the estimated elasticity, and the net direction of inconsistency in their estimate cannot be ascertained.

Fuchs and Kramer were unable to obtain data on physician visit by state. Rather they obtained data on home and office visits per capita for the four census divisions in 1966–1967, which they distributed among nine census regions according to the region's share of the division's visits in 1957–1959. They then attribute the estimated mean level of visits in the region to each state within the region. Finally, they estimate a price for each state by dividing the expenditure for that state by the estimated quantity for that state.

Their procedure can be approximated as follows: Let the true model be:

\[ q_{jk} = \beta p_{jk} + \epsilon_{jk}, \quad j=1\ldots R, \quad k=1\ldots m, \]  

(A.1)

For example, the distribution of the quantity and price of visits among inpatient surgical, inpatient nonsurgical, and outpatient is based on Medicare data. One could reasonably ask for some support for the assumption that the distribution of physician visits among these categories is the same in the under 65 as in the over 65 population.

\[ \text{We owe this proof to Carl Morris.} \]
where \( j \) indexes the region and \( k \) indexes the state. Let \( q \) and \( p \) be logarithms of the quantity and price variables respectively and \( \varepsilon \) be a randomly distributed error term. Assume there are \( m \) states per region.

Define

\[
\varepsilon_{jk} = p_{jk} + q_{jk} + \delta_{jk},
\]

where \( \varepsilon \) is the logarithm of expenditure and \( \delta \) is a randomly distributed error term. If expenditures are measured without error, \( \delta \) is zero.

Define

\[
\tilde{q}_{jk} = \frac{1}{m} \sum_{k=1}^{m} q_{jk} \quad \text{(and } \tilde{p}_{j}, \tilde{\delta}_{j} \text{ and } \tilde{\varepsilon}_{j} \text{ analogously});
\]

\( q \) is given a \( k \) subscript because it is attributed to each state. Hence, \( \tilde{q}_{jk} \) is a geometric mean. Fuchs and Kramer work with the arithmetic mean, but the geometric mean is considerably more convenient and should not alter the general result.

Suppose \( q \) is regressed on \( e - \tilde{q} \). Call the resulting estimator \( b \). We proceed to find the plim of \( b \) as \( R \), the number of regions, tends to infinity, and \( m \), the number of states in a region, is fixed.

Let unbiased estimators of the within-region variances be defined as follows:

\[
s_{j}(p) = \frac{1}{m-1} \sum_{k=1}^{m} (p_{jk} - \tilde{p}_{j})^2
\]

\[
s_j^2(p) = \frac{1}{m-1} \sum_{k=1}^{m} (p_{jk} - \tilde{p}_{j})^2
\]

Define

\[
\text{plim}_{R \to \infty} \left[ \frac{1}{R} \sum_{j=1}^{R} s_{j}^2(p) \right] = s^2(p)
\]

\[
\text{plim}_{R \to \infty} \left[ \frac{1}{R} \sum_{j=1}^{R} a_{j}^2 \right] = a^2
\]

\[
\text{plim}_{R \to \infty} \left[ \frac{1}{mR} \sum_{j=1}^{R} \sum_{k=1}^{m} \delta_{jk}^2 \right] = m\tau
\]

(A.2)

(A.3)

(A.4)

(A.5)
\[
\plim_{R \to \infty} \left[ \frac{1}{R} \sum_{j=1}^{R} \frac{1}{p_j} \right] = \mu^2.
\]

Then (dropping the \(R \to \infty\) notation)

\[
\plim b = \frac{\plim \left[ \frac{1}{R} \sum_{j} \sum_{k} q_{jk} \left( p_{jk} + q_{jk} - \bar{q}_{jk} + \delta_{jk} \right) \right]}{\plim \left[ \frac{1}{R} \sum_{j} \sum_{k} \left( p_{jk} + q_{jk} - \bar{q}_{jk} + \delta_{jk} \right)^2 \right]}.
\]

(A.6)

When we sum over \(k\) in the numerator, because \(q_{jk}\) is fixed, the \(q_{jk} - \bar{q}_{jk}\) term in the product equals zero. By definition of \(\bar{p}_{jk}\) and \(\delta_{jk}\) the numerator becomes

\[
\plim \left[ \frac{1}{R} \sum_{j} \sum_{k} q_{jk} \left( \bar{p}_{jk} + \delta_{jk} \right) m \right].
\]

Substituting \(\beta p_{jk} + \epsilon_{jk}\) for \(q\) and \(\beta p_{jk} + \bar{\epsilon}_{jk}\) for \(\bar{q}\) in the denominator yields

\[
\plim \left[ \frac{1}{R} \sum_{j} \sum_{k} \left( p_{jk} (1 + \beta) + \epsilon_{jk} - \beta \bar{p}_{jk} - \bar{\epsilon}_{jk} + \delta_{jk} \right)^2 \right].
\]

(A.7a)

Adding and subtracting \(p_{jk}\) yields

\[
\plim \frac{1}{R} \sum_{j} \sum_{k} \left[ \left( p_{jk} - \bar{p}_{jk} \right)(1 + \beta) + \left( \epsilon_{jk} - \bar{\epsilon}_{jk} \right) + \bar{p}_{jk} + \delta_{jk} \right]^2.
\]

(A.7b)

\(\delta\) and \(\beta\) are assumed independent of \(p\) and each other. Substituting \(\beta \bar{p}_{jk} + \bar{\epsilon}_{jk}\) for \(\bar{q}\) in the numerator yields

\[
\plim b = \beta \frac{m \mu^2}{\left[ (1 + \beta)^2 (m - 1) \sigma^2(p) + (m - 1) \sigma^2 + m \mu^2 \right]}.\]

(A.7c)

Dividing numerator and denominator by \(m \mu^2\), we have

\[
\plim b = \frac{\beta}{1 + \left( \frac{m - 1}{m} \right) \left[ \frac{\sigma^2}{\mu^2} + \left( \frac{1 + \beta}{\Pi^2} \right)^2 \frac{\sigma^2(p)}{\Pi^2} + \beta^2 \right]}.
\]

(A.8)
Because all terms in the denominator are positive, the inconsistency is toward zero. Two features might be noted. The "usual" case is $m = 1$, or one state per region. There is then inconsistency only from the error in measuring expenditure, the usual errors-in-the-variables problem. But if $m$ is greater than one, there is an inconsistency even if there is no within-region price variation; because there is variation in $\epsilon$, the within region $q$s will still differ.
Appendix B

INCONSISTENCY FROM USING OLS RATHER THAN TSLS

Assume the true model of demand for medical care \((y)\) and insurance coverage \((K)\) is as follows:

\[
y_i = K_i \beta_1 + S_i \beta_2 + \epsilon_i
\]

\[
K_i = \alpha_i \bar{S}_i + u_i
\]

where \(K = \text{coverage} (1 - C_i), S_i = \text{a random level of sickness, and } \bar{S}_i = \text{E}(S_i)\) for the \(i\)th person, so that \(S_i = \bar{S}_i + \delta\) with \(\delta \sim (0, \sigma^2)\). For notational convenience, assume \(E(\bar{S}_i) = E(S_i) = 0\), and that \(y_i\) and \(K_i\) have held constant all other factors affecting them.

Assume that one cannot (perfectly) observe \(S_i\), hence the estimate is

\[
y_i = K_i \beta_1' + \epsilon_i'.
\]

The OLS estimator of \(\beta_1'\) has (omitting the limit notation as the number of observations \(T\) tends to infinity)

\[
\text{plim } \hat{\beta}_1' = \frac{\text{plim } \Sigma (K_i - \bar{K})(K_i \beta_1 + S_i \beta_2 + \epsilon_i - \bar{K} \beta_1 - \bar{S}_i \beta_2)/T}{\text{plim } \Sigma (K_i - \bar{K})^2/T}
\]

\[
= \beta_1 + \text{plim } \frac{\Sigma (K_i - \bar{K})(S_i \beta_2 + \epsilon_i)/T}{\Sigma (K_i - \bar{K})^2/T}
\]

\[
= \beta_1 + \text{plim } \frac{\Sigma (a_i \bar{S}_i + u_i - a_i \bar{K}) (S_i \beta_2 + \epsilon_i)/T}{\Sigma (K_i - \bar{K})^2/T}
\]

On the assumption that \(\text{plim } \Sigma u_i \epsilon_i/T = 0\), then

\[
\text{plim } \hat{\beta}_1' = \beta_1 + \frac{\text{plim } \Sigma a_i \beta_2 \bar{S}_i S_i/T}{\text{plim } (K_i - \bar{K})^2/T}
\]

\[
= \beta_1 + \frac{a_i \beta_2 \sigma_{\bar{S}_i S_i}}{\sigma_K^2}
\]

where \(\sigma_{\bar{S}_i S_i} = \text{plim } \Sigma \bar{S}_i S_i/T\) and \(\sigma_K^2 = \Sigma (K_i - \bar{K})^2/T\).
From economic theory, both $\beta_2$ and $a_1$ are positive, so $\text{plim } \beta_1^* > \beta_1$ if $\sigma_\bar{S} > 0$.

By assumption, $S_i = \bar{S}_i + \delta_i$ and $\text{plim } \Sigma S_i \delta_i = 0$, so $\text{plim } \sigma_{\bar{S}S} = \sigma_{\bar{S}} = \sigma_{\bar{S}}^2 > 0$, which establishes the inconsistency of OLS in this case.

1 It is important to note that, under some restrictive circumstances, demand for $K$ might not be a positive function of $\bar{S}$ (Phelps, 1973). However, it can be shown that whenever there is either group rating of insurance (among subgroups) or imperfect ability by the insurer to measure $S_i$, then standard assumptions would yield a model with $K_i$ a positive function of $S_i$. Further, it is possible under many conditions to have $K_i = f(S_i)$ (with $f$ positive) even with individual insurance pricing and full knowledge of $\bar{S}$ by the insurer.
Appendix C

PROBABILITY LIMIT OF ESTIMATED COEFFICIENT IF COINSURANCE IS MEASURED WITH A DUMMY VARIABLE

Suppose the true model is:

\[ \ln Y = \beta_0 + \beta_1 \ln C + \beta_2 \ln P + \epsilon, \]  \hspace{1cm} (C.1)

or \( \beta_0 + \beta_1 c + \beta_2 p + \epsilon \), where lower case letters denote logarithms. \( Y \) is utilization, \( c \) is coinsurance, and \( p \) is gross price.\(^1\) Suppose \( D \), which takes the value 1 if \( c < 0 \) and 0 if \( c = 0 \), is used in place of \( c \). (If \( c = 0 \), the coinsurance rate equals 1, because \( c \) is a logarithm.) Let the estimated equation be:

\[ y = b_0 + b_1 D + b_2 p + u. \]  \hspace{1cm} (C.2)

We are interested in the plim of \( b_2 \) as the number of observations \( T \) tends to infinity. By definition,

\[ D = c + u, \]  \hspace{1cm} (C.3)

so \( u = 0 \) if \( c = 0 \), and \( u = 1 - c \) if \( c < 0 \).

Define

\[ \text{plim} \left( \frac{1}{T} \Sigma (c - \bar{c})^2 \right) = \sigma_{cc} > 0 \]

\[ \text{plim} \left( \frac{1}{T} \Sigma (u - \bar{u})^2 \right) = \sigma_{uu} > 0 \]

\[ \text{plim} \left( \frac{1}{T} \Sigma (c - \bar{c})(u - \bar{u}) \right) = \sigma_{cu} < 0 \]

\[ \text{plim} \left( \frac{1}{T} \Sigma (D - \bar{D})^2 \right) = \sigma_{DD} = (1 + \lambda) \sigma_{cc}, \lambda > -1 \]

\[ \text{plim} \left( \frac{1}{T} \Sigma (p - \bar{p})^2 \right) = \sigma_{pp} > 0 \]  \hspace{1cm} (C.4)

\[ \text{plim} \left( \frac{1}{T} \Sigma (c - \bar{c})(p - \bar{p}) \right) = \sigma_{cp} < 0 \]

\[ \text{plim} \left( \frac{1}{T} \Sigma (u - \bar{u})(p - \bar{p}) \right) = \mu \sigma_{cp}, \mu < -1 \]

\(^1\) \( \beta_2 \) can equal \( \beta_1 \) as a special case.
\[ \text{plim} \frac{1}{T} \sum (c - \bar{c})(D - \bar{D}) = (1 + \theta) \sigma_{cc}, \quad \theta < -1 \]

\[ \text{plim} \frac{1}{T} \sum (c - \bar{c})(c) = \text{plim} \frac{1}{T} \sum (p - \bar{p}) e = \text{plim} \frac{1}{T} \sum (u - \bar{u}) e = 0, \]

where a bar over a variable indicates its mean.

Before proceeding, we note the pattern of signs \( \sigma_{cc}, \sigma_{uu}, \sigma_{DD}, \text{ and } \sigma_{pp} \) are all positive, because they are sums of squares. \( \sigma_{cc} \text{ and } \sigma_{DD} \) are positive, so \( \lambda \) must exceed \(-1\). \( \sigma_{cu} \) is negative from the definition of \( c \) and \( u \). \( \sigma_{cp} \) was observed empirically to be negative in the 1963 CHAS-NORC survey (Newhouse and Phelps, 1976). Proofs that \( \mu \text{ and } \theta \) are less than \(-1\) can be found at the end of this appendix. Note also that \( \sigma_{cu} = \theta \sigma_{cc} \). We are interested in

\[
\text{plim} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.
\]

\[
\text{plim} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \text{plim} \begin{pmatrix} (D') (D p) \\ (p') \end{pmatrix}^{-1} \begin{pmatrix} D' \\ p' \end{pmatrix} y
\]

\[
= \begin{pmatrix} (1 + \lambda) \sigma_{cc} (1 + \mu) \sigma_{cp} \\ (1 + \mu) \sigma_{cp} \sigma_{pp} \end{pmatrix}^{-1} \begin{pmatrix} (1 + \theta) \sigma_{cc} (1 + \mu) \sigma_{cp} \\ \sigma_{cp} \sigma_{pp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}
\]

\[
= \begin{pmatrix} a & 0 \\ \gamma & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} a \beta_1 \\ \beta_2 + \gamma \beta_1 \end{pmatrix}
\]

where

\[
a = \frac{(1 + \theta) \sigma_{pp} \sigma_{cc} - (1 + \mu) (\sigma_{cp})^2}{(1 + \lambda) \sigma_{pp} \sigma_{cc} - (1 + \mu)^2 (\sigma_{cp})^2}.
\]

\[
\gamma = \frac{[(1 + \lambda) - (1 + \mu)(1 + \theta)] \sigma_{cc} \sigma_{cp}}{(1 + \lambda) \sigma_{pp} \sigma_{cc} - (1 + \mu)^2 (\sigma_{cp})^2}.
\]
Unfortunately, $\gamma$ is unsigned, and so, therefore, is the inconsistency of $b_2$. By definition, the first term in the numerator of $\gamma$,

$$1 + \lambda - (1 + \mu)(1 + \theta) = 1 + \frac{2 \sigma_{u} + \sigma_{uu}}{\sigma_{cc}} - 1 - \mu - \theta - \mu \theta$$

$$= 1 + 2\theta + \frac{\sigma_{uu}}{\sigma_{cc}} - 1 - \mu - \theta - \mu \theta$$

$$= \theta - \mu(1 + \theta) + \frac{\sigma_{uu}}{\sigma_{cc}}. \quad (C.5.2)$$

$\theta$ is less than $-1$, $\mu(1 + \theta)$ is positive, but $\sigma_{uu}/\sigma_{cc}$ can be shown to be greater than $+1$, so the entire term is unsigned. If $1 + \lambda - (1 + \mu)(1 + \theta)$ is negative, $\gamma$ is positive, and plim $b_2$ is more negative than $\beta_2$ (because $\beta_1$ is negative).

It remains to show that $(1 + \mu)$ and $(1 + \theta)$ are negative:

$$\mu = \frac{\sigma_{up}}{\sigma_{cp}} = \frac{\lim_{T \to \infty} \left( \frac{1}{T} \sum_{1}^{T} (u_i - \bar{u})(p_i - \bar{p}) \right)}{\lim_{T \to \infty} \left( \frac{1}{T} \sum_{1}^{T} (c_i - \bar{c})(p_i - \bar{p}) \right)} \quad (C.6)$$

Without restricting the number $m$, let the first $m$ observations be ordered such that $u_i = c_i = 0$, and let $m/T$ be constant as $m$ and $T$ tend to infinity. For the remaining $T - m$ observations, $c_i < 0$. Consider the numerator

$$\sigma_{up} = \lim_{T \to \infty} \frac{1}{T} \left[ \sum_{1}^{m} (u_i - \bar{u})(p_i - \bar{p}) + \sum_{m+1}^{T} (u_i - \bar{u})(p_i - \bar{p}) \right]$$

$$u = \frac{\Sigma u_i}{T} = \frac{1}{T} + \frac{m+1}{T} = \frac{m+1}{T} = \frac{T - m}{T} = T - m - \bar{c}, \quad (C.7)$$

because the first $m$ values of $c_i$ also equal 0.

Then $\sigma_{up}$ equals
\[
\text{plim} \frac{1}{T} \left[ \frac{m}{T} \left( - \frac{m}{T} + \bar{v} \right) (p_i - \bar{p}) + \sum_{m+1}^{T} \left( 1 - c_i - \left( \frac{T-m}{T} - \bar{v} \right) (p_i - \bar{p}) \right) \right] \\
= \text{plim} \frac{1}{T} \left[\frac{T-m}{T} \sum_{1}^{T} (p_i - \bar{p}) + \sum_{1}^{m} \bar{c} (p_i - \bar{p}) \right. \\
\left. + \sum_{m+1}^{T} \left( 1 - c_i + \bar{c} \right) (p_i - \bar{p}) \right] \\
= \text{plim} \frac{1}{T} \left[ \sum_{1}^{m} \bar{c} (p_i - \bar{p}) + \sum_{m+1}^{T} (p_i - \bar{p}) - \sum_{m+1}^{T} (c_i - \bar{c}) (p_i - \bar{p}) \right],
\]

as \( \sum_{1}^{T} (p_i - \bar{p}) = 0 \).

The denominator of (C.8) equals

\[
\sigma_{cp} = \text{plim} \frac{1}{T} \left[ \sum_{1}^{m} (1 - \bar{c}) (p_i - \bar{p}) + \sum_{m+1}^{T} (c_i - \bar{c}) (p_i - \bar{p}) \right],
\]

Note that the terms in the brackets are the negative of the first and third terms of (C.8). The term \( \sum_{m+1}^{T} \bar{c} (p_i - \bar{p}) \) is assumed positive, because it is the gross price selected by those insured, and \( \sigma_{cp} \) is assumed negative. The term in the brackets in (C.8) therefore equals the negative of the negative sum in (C.9), plus a positive term, so that

\[
\sigma_{up} > | \sigma_{cp} | \quad \text{and} \quad \mu < -1.
\]

Next we show \( \theta < -1 \).

\[
\sigma_{cu} = \text{plim} \frac{1}{T} \left[ \sum_{1}^{T} (c_i - \bar{c}) (u_i - \bar{u}) \right] \\
= \text{plim} \frac{1}{T} \left[ \sum_{1}^{m} (- \bar{c}) \left( - \frac{T-m}{T} + \bar{c} \right) + \sum_{m+1}^{T} (c_i - \bar{c}) \left( 1 - c_i - \frac{T-m}{T} + \bar{c} \right) \right]
\]
\[
= \plim \frac{1}{T} \left[ -\frac{m}{1} \bar{c}^2 + \sum_{1}^{m} \bar{c} \left( \frac{T - m}{T} \right) - \sum_{m+1}^{T} (c_i - \bar{c})^2 + \sum_{m+1}^{T} (1 - \frac{T - m}{T})(c_i - \bar{c}) \right].
\]

\[
= \plim \frac{1}{T} \left[ -\frac{T}{1} (c_i - \bar{c})^2 + \sum_{1}^{m} \bar{c} \left( \frac{T - m}{T} \right) + \sum_{m+1}^{T} (1 - \frac{T - m}{T})(c_i - \bar{c}) \right].
\]

The first term in probability limit is \(-\sigma_{cc}\). Because \(c\) is negative, and \(T > m > 0\), the second term is negative. The third term is also negative, as \(\sum_{m+1}^{T} (c_i - \bar{c})\) is negative. This is the case because \(c_i = 0\), \(i = 1 \ldots m\), and \(c_i < 0\), \(i > m\). Hence \(\|\sigma_{cu}\| > \sigma_{cc}\) and \(\theta < -1\).
Appendix D
INCONSISTENCY USING EXPENDITURE DIVIDED BY QUANTITY TO ESTIMATE PRICE

We first show that if expenditure \( E \) is measured without error, but quantity \( x \) is measured with random error \( u \), and if quantity is negatively related to price, then estimating price \( p \) by dividing expenditure by observed quantity will result in a variable whose error is negatively associated with price. Because estimated \( p = (p \times x)/(x + u) = p(1 - \frac{u}{x+u}) \), the error is \(-up/(x + u)\). Because \( x \) is negatively related to \( p \), the plim of the covariance of this expression with \( p \) is negative.

Next we show that this results in an inconsistency away from zero in a commonly used specification. Suppose the true model were, as in Fuchs and Kramer:

\[
\ln (x) = \alpha - \beta \ln (p) + \epsilon, \tag{D.1}
\]

where \( \epsilon \) is a random error term, and all variables are measured as deviations around means.

Let observed quantity be measured with multiplicative random error \( u \), so that observed quantity equals \( \ln(x) + u \). Let total expenditure be measured with multiplicative random error \( e \), so that \( \ln(E) = \ln(x) + \ln(p) + e \). Let

\[
\text{plim}_{n \to \infty} \Sigma \left( \frac{1}{n} \Sigma \epsilon u \right) = \sigma_{\epsilon u} \geq 0, \tag{D.2}
\]

allowing for the possibility that \( e \) and \( u \) may have common elements. Let price be estimated by dividing observed expenditure by observed quantity, and let \( \beta \) be estimated by regressing the \( \ln \) of observed quantity on estimated price. The estimated equation is then

\[
\ln(x) + u = a - b(\ln(E) - \ln(x) - u), \tag{D.3}
\]

where \( a \) and \( b \) are estimated.

Then plim \( b \) equals

\[
\frac{\text{plim}_{n \to \infty} \left( \frac{1}{n} \Sigma (\ln(x) + u) (\ln(E) - \ln(x) - u) \right)}{\text{plim}_{n \to \infty} \left( \frac{1}{n} \Sigma (\ln(E) - \ln(x) - u)^2 \right)}
\]

\[
= \frac{\text{plim}_{n \to \infty} \left( \frac{1}{n} \Sigma (a - \beta \ln(p) + \epsilon + u) (\ln(p) + e - u) \right)}{\text{plim}_{n \to \infty} \left( \frac{1}{n} \Sigma (\ln(p) + e - u)^2 \right)} \tag{D.4}
\]

36
Define

$$\text{plim} \left( \frac{1}{n} \sum \ln(p)^2 \right) = \sigma_{pp} > 0$$

$$\text{plim} \left( \frac{1}{n} \sum u^2 \right) = \sigma_{uu} > 0$$

$$\text{plim} \left( \frac{1}{n} \sum e^2 \right) = \sigma_{ee} > 0$$

$$\text{plim} \left( \frac{1}{n} \sum (u-e)^2 \right) = \sigma_{(u-e)(u-e)} > 0$$

$$\text{plim} \left( \frac{1}{n} \sum \epsilon u \right) = \text{plim} \left( \frac{1}{n} \sum \epsilon e \right) = \text{plim} \left( \frac{1}{n} \sum \epsilon \ln(p) \right)$$

$$= \text{plim} \left( \frac{1}{n} \ln(p)e \right) = \text{plim} \left( \frac{1}{n} \sum \ln(p)u \right) = 0.$$  \hspace{1cm} (D.5)

Then plim b equals

$$\frac{-\beta \sigma_{pp} - \sigma_{uu} + \sigma_{ue}}{\sigma_{pp} + \sigma_{ee} - 2\sigma_{ue} + \sigma_{uu}} = \frac{-\beta \sigma_{pp} - \sigma_{uu} + \sigma_{ue}}{\sigma_{pp} + \sigma_{(u-e)(u-e)}}.$$ \hspace{1cm} (D6.a)

This expression reduces to

$$- \beta - \frac{\sigma_{uu}(1-\beta)}{\sigma_{pp} + \sigma_{(u-e)(u-e)}} + \frac{(1-2\beta)\sigma_{ue} + \beta \sigma_{ee}}{\sigma_{pp} + \sigma_{(u-e)(u-e)}}.$$ \hspace{1cm} (D6.b)

If expenditures are measured without error, \( \sigma_{ue} = \sigma_{ue} = 0 \), and the third term in (D6.b) is zero. In this case plim (b) is more negative than \(-\beta\), so long as \( \beta \) is less than one, as all empirical estimates find. If expenditures are measured with error, such errors are probably positively correlated with the error in measuring quantity. In this case the third term in (D6.b) is positive if \( \beta \) is less than 0.5, as most empirical estimates show. The inconsistency away from zero in the estimate of \( \beta \) is thus reduced if expenditures as well as quantity are measured with error. In fact, the direction of the inconsistency may be toward zero if expenditures are measured with sufficiently large error. If, however, expenditures are measured with substantially less error than quantity, plim (b) will be more negative than \( \beta \). This is likely to be the case with the Fuchs and Kramer data, because quantities must be estimated from regional data.

An analogous problem occurs if one regresses expenditure on a coinsurance rate computed as \( C = (\text{Expenditure} - \text{Benefits})/\text{Expenditure} \). The proof is identical
to that given above, except that quantity (x) is replaced by expenditure (E), and expenditure (E) is replaced by Expenditure − Benefits (E − B).¹

¹ In our computations in Table 6, we made this error (of necessity) while attempting to estimate the inconsistency due to aggregation across services. Hence, not all of the increase in the estimates in Table 6 can be attributed to aggregation across services, although we believe that a large portion is due to this problem.
BIBLIOGRAPHY


