Forthcoming in Reveiw of Economics and Statistics

Determinants of Healthcare Decisions: Insurance, Utilization, and Expenditures[1]

Chan Shen[2]

Abstract

This paper studies three inter-related healthcare decisions: insurance, utilization, and expenditures. The model treats insurance as an endogenous variable with respect to both utilization and expenditures, addresses potential selection issues, and takes into account that the decisions to utilize healthcare and the level of treatment are determined by different decision makers. We employ semiparametric methods to avoid making distributional assumptions. Using the Medical Expenditure Panel Survey 2005 data, the semiparametric approach predicts insurance to increase the level of expenditures by 48%, a number in accord with an important experimental study and less than half that obtained using parametric methods.

JEL codes: C14,C31,C35,I18.

1 Introduction

A major healthcare policy issue in the U.S. today is the growing population without insurance. The key questions include: How does health insurance coverage affect the likelihood an individual seeks medical care? How does health insurance affect healthcare expenditures?

There are many empirical challenges in studying people's healthcare decisions. An individual's decision about whether to utilize healthcare may depend on her insurance coverage. The level of utilization likely also depends on whether or not the individual has insurance.

However, because insurance is a choice variable for the individual, we must allow for the possibility that this variable is endogenous. For example, people who have greater need for healthcare have more incentive to buy health insurance. Some papers in the literature deal with this endogeneity by using instrumental variables (e.g., Vera-Hernandez, 1999; Holly et al., 2002; Wooldridge, 2002); others use experimental data to avoid this problem (e.g., Manning et al., 1987; Newhouse, 1993; Barros et al., 2008). However, instruments that are correlated with insurance coverage but not with utilization are difficult to find. Experimental data are scarce and often out of date. For example, the RAND Health Insurance Experiment, which remains to be the largest health policy study in U.S. history, started in 1971 and lasted for 15 years (RAND, 1974-1982). The structure, practice, and philosophy of medicine have changed dramatically since the 1980s as has the insurance industry.

Another empirical challenge lies in the expenditure decision, where we only observe positive expenditures from individuals who decide to see a doctor. One standard parametric approach deals with this problem by making distributional assumptions about error terms and then using a Heckman correction for sample selection (Heckman, 1976, 1979). An alternative approach is to use a two-part model (Duan et al., 1983, 1984, 1985). Both of these may be problematic as the Heckman correction approach can be sensitive to the distributional assumptions on error terms, while the two-part model approach also makes implicit distributional assumptions (Puhani, 2000). The literature in health economics or in economics in general does not provide a theoretical foundation or justification for these distributional assumptions. Moreover, if incorrect, they can result in incorrect inferences and policy conclusions with respect to healthcare decisions.

Yet another challenge is the complicated nature of the decision-making process. In healthcare, both the patient and the doctor are involved in making decisions. The patient decides

whether to visit a doctor (or more generally a healthcare provider), and then the patient and doctor jointly decide what treatment the patient will have. These decisions are interrelated. Some papers deal with the two-part decision-making process in healthcare utilization (Newhouse, 1993; Mullahy, 1998), but none address the whole process of insurance choice, utilization, and expenditure level.

This paper contributes to the current literature by taking into account the interrelated nature of healthcare decisions and using a semiparametric approach to address the empirical challenges. We study three healthcare decisions: insurance coverage, utilization, and the level of expenditures. Using the Medical Expenditure Panel Survey (MEPS) 2005 data, we formulate and estimate a model for these three healthcare decisions. As there is not a strong justification for normality assumptions underlying a traditional parametric formulation, we employ a semiparametric approach in which these assumptions are not made. As an additional advantage to a semiparametric approach, it should be remarked that since marginal effects will in general not be constant in nonlinear models, we will report the impact of changing a variable of interest at several different points in its distribution. The semiparametric approach will also allow greater flexibility in the pattern of these effects than in the parametric case. Nevertheless, as a convenient benchmark, we also estimate the model using a standard parametric approach.

While the focus of this paper is on healthcare decisions, the methods used would also apply to other endogenous treatment models. For example, in labor economics, a woman's fertility/marriage decision, the decision to join the work force, and the wage level have a similar structure.

The paper is organized as follows. Section 2 introduces the model and explains the parametric and semiparametric approaches in two subsections; Section 3 describes the dataset;

Section 4 gives the main results; and Section 5 provides conclusions, discussions, and future research directions.

2 The Model

We study a set of three equations to examine the effects of different factors on healthcare decisions: health insurance, utilization, and expenditures. The first equation deals with the health insurance choice. Let 1 be an indicator of whether or not an individual selects private health insurance coverage. In the model below, an individual selects insurance if the net value to so doing, $V_I - \varepsilon_I$, is greater than zero. With V_I determined by a set of exogenous variables X_I, and $1\{\bullet\}$ as an indicator function, the model is as follows:

$$I = 1\{V_I > \varepsilon_I\}, \text{where } V_I = X_I\beta_I,$$

The second equation describes the decision to seek healthcare. Let A be an indicator of whether or not an individual seeks access to healthcare from a doctor or other healthcare providers, and let X_A be a set of exogenous variables that determine the net value of utilizing healthcare. Then:

$$A = 1\{V_A + I\theta_A > \varepsilon_A\}, \text{where } V_A = X_A\beta_A.$$

Notice that the insurance coverage enters this utilization (access) equation. There is a vast literature about the effects of moral hazard and adverse selection (see Arrow, 1963; Rothschild and Stiglitz, 1976; Chiappori and Salanie, 2001; Cardon and Hendel, 2001, for example). On the one hand, people who have insurance are much more likely to utilize healthcare than their uninsured counterparts. On the other hand, people who have greater

demand for healthcare (e.g., those with high comorbidity levels) may have more incentive to obtain insurance coverage. Consequently in our estimations, we will use methods that deal with this endogeneity issue.

The last equation explains the level of expenditures. Denote Y_E as the log of level of expenditures, and X_E as a set of exogenous variables that affects expenditures for those individuals who access healthcare services. Then, the model is given as:

$$Y_E = \quad X_E\beta_E + I\theta_E + u \quad : A = 1 \ \cdot$$

An individual incurs positive expenditures only if a visit is made. The patient decides whether to visit a doctor, and then a joint decision is made by both the doctor and the patient. We address this two-part decision-making process by separating the two equations and allowing them to have different explanatory variables and parameters. Again, insurance, healthcare, and the individual's health status are interrelated. Insurance coverage is included in this model, because it may affect patient and doctor's joint decision about treatment plans. For example, insured people are much more likely to buy brand-name medications instead of their generic counterparts. There could also be an adverse selection problem here, because it is possible that people who are less healthy might have more incentive to purchase insurance. Hence our model will account for the interrelations between the above variables, and we will employ estimation methods that deal with both sample selection and endogeneity issues.

To avoid making strong distributional assumptions that are hard to justify, in this paper we employ a semiparametric method to estimate the three healthcare equations discussed above. Indeed, we will find that standard parametric distributional assumptions (e.g. joint normality) do not hold. Nevertheless, as a convenient benchmark, we also provide the

parametric formulation and results. There are a variety of different methods for estimating the parametric model. To make the role of the parametric assumptions transparent, we estimate the parametric model in a manner that parallels the semiparametric approach.

2.1 Parametric Model

In the parametric model, we assume that the error terms in the above system of three equations follow a trivariate normal distribution. A two-step estimation method is then employed to estimate the three equations. In the first step, the insurance and utilization decisions are jointly estimated by maximum likelihood (bivariate probit).

To identify the parameters without relying on nonlinearities, we require restrictions on the model. The insurance equation will depend on only exogenous variables, X_I, while the access decision will depend on exogenous variables, X_A, and whether or not the individual has insurance. In this triangular system of binary equations, the insurance equation is identified as it is essentially a reduced form. However, to identify the access equation, we impose exclusion restrictions on it. These exclusions will be discussed in the data section.

In addition to the parameters in the joint model for the two decisions, the likelihood depends on the correlation between the errors. A non-zero correlation between the two error terms would indicate the endogeneity of insurance with respect to the utilization decision. As will be described below, we find this correlation to be small in absolute magnitude and not statistically different from zero.

In the second step, we estimate the expenditure equation by employing a Heckman correction (Heckman, 1976; Fishe et al., 1981; Lee, 1982) that controls for both sample selection and endogeneity. To simplify this correction, we employ a form for it that is applicable when, as was found empirically, utilization and insurance errors are not correlated.[3] For individu-

als that utilize healthcare, recall the form of the expenditure model in the previous section. With u as the error term in the log expenditure model, and denoting X_s as the set of all the exogenous variables in the system of three equations, for $d \in \{1, 0\}$, define:

$$\lambda_d G_d \left(V_A, V_I\right) = E(u|X_s, A = 1, I = d).$$

In a parametric model with jointly normal errors, the G-functions above are known and the $\lambda' s$ are parameters whose values are unknown. Typically, the above expectations are not zero and depend on the variables X_s, A, and I. To estimate our model, we seek to remove the dependence of the errors on these conditioning variables. To this end, for $d \in \{1, 0\}$, we can rewrite the expenditure equation as:

$$Y_E = X_E \beta_E + \theta_E d + \lambda_d G_d + u_d^* : A = 1, I = d, u_d^* = u - \lambda_d G_d \left(V_A, V_I\right),$$

By construction the conditional expectation of the recentered error is zero: $E\left(u_d^* | X_s, A = 1, I = d\right) = 0$.

Provided that the above equation is identified and joint normality holds, OLS estimation provides consistent estimates. To identify it without relying on nonlinearities in the G-controls, we impose exclusion restrictions on the exogenous variables X_E that enter this equation. Detailed discussions about these and other restrictions will be provided in the data section.

We conclude this discussion about the parametric model by emphasizing the importance of its restrictive parametric assumptions. Both the bivariate probit specification and the form of the correction term depend on the (joint) normality assumption. If this assumption is

incorrectly imposed, the resulting estimator is typically inconsistent. In the next subsection, we propose a semiparametric approach that does not make distributional assumptions.

2.2 Semiparametric Model

While the semiparametric model generalizes the parametric model, it does retain a parametric (index) restriction to ensure that the estimator "works well" in moderately sized samples. To illustrate this restriction, return to the insurance model. In a commonly employed probit specification:

$$P(I = 1|X) = \Phi\left(X_I\beta_I\right),$$

where the function Φ is the cumulative distribution function for the model's standard normal error component, ε_I. In a semiparametric formulation, this function need not be specified and indeed can be estimated from the data along with parameters of interest. In such a formulation, the model is semiparametric because it makes no parametric assumptions on the error distribution, but does assume a parametric index, $V_I \equiv X_I\beta_I$. This index, V_I, need not be linear, but it is important that it has a parametric form. In a more general, nonparametric formulation, we might write:

$$P(I = 1|X) = F(X_1, X_2, ..., X_K) = E(I|X).$$

However, when the dimension of X is large, it is difficult to "reliably" estimate the above probability (expectation).[4] Index restrictions serve to keep the relevant dimension of the problem small and thereby improve the finite sample behavior of the estimator. In general,

a single index restriction takes the form:

$$E(I|X) = E(I|V_I) \equiv F_1(V_I).$$

In this form, not only is the function F_1 left unspecified, but the model also permits very flexible interactions between errors and the index.

In some problems, a single index may not adequately describe the underlying behavior of interest. Given that the access model is not linear, when insurance is endogenous with respect to access, the access probability depends not only on its own index but also on the exogenous index driving the insurance decision. In this case, a double index model would be appropriate:

$$E(Y|X) = E(Y|V_I, V_A) \equiv F_2(V_I, V_A),$$

where V_I, V_A are now two indices. Again, there are methods for reliably estimating the above expectation under this double index structure. As discussed below, estimators for both single and double index models will be employed here. Throughout, we use the notation $\hat{E}(Y|V)$ to denote an estimated conditional expectation for Y conditioned on V, where V may be a single index or a vector containing two indices. When this estimated expectation is evaluated at an estimate of V, as will be the case below, we will write $\hat{E}(Y|\hat{V})$.

Before continuing, it is important to discuss identification of both index parameters and marginal effects of interest. Recall that in the parametric case, the original parameters are identified under exclusion restrictions. However, as in all nonlinear models, parameters do not translate directly into marginal effects which are of primary interest. Marginal effects are recovered by comparing estimated probabilities based on parametric distributional assumptions. In the semiparametric case, however, it is well known in the literature that the index

parameters can at most be identified up to location and scale. For simplicity, we illustrate the issue for the insurance decision. As will be discussed below, the estimates are based in part on an estimate of the probability:

$$Pr(I = 1|X_I\beta_I) = Pr(I = 1|a + b(X_I\beta_I)) \text{ , where } a \text{ and } b \neq 0 \text{ are constants.}$$

The above probability does not depend on a or b. Therefore, only ratios of index parameters are identified. Nevertheless, the scaled parameters enable us to recover probabilities and hence marginal effects of interest.

Unlike the insurance decision, the utilization decision depends on the endogenous insurance decision with coefficient θ_A. While this parameter is not identified, we can recover the corresponding marginal effect by looking at an appropriate probability change. One possibility is to report the difference in access probabilities conditioned on insurance and no insurance, which are estimable semiparametrically as discussed below. It is easy to justify this calculation if the insurance error does not depend on the access error. As the estimated parametric correlation between access and insurance errors is insignificant and small in absolute magnitude, this calculation may be reasonable and is reported here. It is also possible to estimate the marginal effect of insurance when access and insurance errors are dependent. Vytlacil and Yildiz (2007) discuss identification in this context.[5]

For the expenditure equation, subject to the exclusion restrictions made here, all parameters are directly identified other than the coefficient on insurance, θ_I. This parameter is a direct marginal effect of interest and is a very important parameter in the model. Since there is no evidence that insurance can be treated as exogenous in the expenditure equation, below we discuss an estimation method for recovering this parameter.

Turning to the estimation method, in the insurance and utilization decisions, we estimate the model by a method that is analogous to that for the parametric case. For that case, the form of the likelihood is known and the model is estimated by maximum likelihood. In contrast, here we do not make any distributional assumptions on error components, implying that the form of the likelihood is unknown. Nevertheless, it is possible to employ index assumptions above to develop an estimator for the likelihood.

We employ an estimator based on an extension of the approach in Klein and Shen (2010), where a bias correction mechanism was proposed to overcome finite sample performance issues of common semiparametric estimators in the literature. Monte Carlo studies in that paper show that this estimator dominates the others in terms of mean squared error. One component of the model below contains a triangular system of binary response equations. Klein, Shen, and Vella (2010, hereafter KSV) extend the bias-control mechanism discussed above to establish desirable large-sample properties for the estimator of this component. The estimator for these components of the model is then based on maximizing an "estimated log-likelihood". To define this function, for individual i and $r, s \in \{0, 1\}$, let:

$$Y_{rs}(i) = 1\{A(i) = r, I(i) = s\},$$

with the corresponding probabilities:

$$P_{rs}(i) = \Pr(Y_{rs}(i) = 1 | V_A(i), V_I(i)).$$

Suppressing individual subscripts for notational simplicity, for $r = 1$ and $s = 1$ (other cases

are analogous), notice that

$$\begin{aligned} P_{11} &\equiv \Pr\left(Y_{11}=1|V_A,V_I\right)=\Pr\left(A=1,I=1|V_A,V_I\right) \\ &= \Pr\left(A=1|I=1,V_A,V_I\right)\Pr\left(I=1|V_A,V_I\right) \\ &= \Pr\left(A=1|I=1,V_A,V_I\right)\Pr\left(I=1|V_I\right) \\ &= E\left(A|I=1,V_A,V_I\right)E\left(I|V_I\right). \end{aligned}$$

Hence, P_{11} can be estimated by estimating each of the above two expectations semiparametrically. The first expectation over A has a double index form, while the second one has a single index form. The product of the above expectations (probabilities) then provides the joint probability of interest. In general double index models, identification requires that each index contains a continuous variable that is excluded from the other. Since one component (insurance) of the model has a single index form, it is not required here. We do require, however, that the insurance equation contains a continuous variable excluded from the access equation and that the access equation depends on at least one continuous variable. [6] In addition to these continuity restrictions, we require and impose the same exclusion restrictions discussed in the previous section for the parametric model.

Given the estimated probabilities, we can now proceed as in Klein and Spady (1993) to estimate the model by maximizing the following estimated log likelihood:

$$Log\hat{L}=\sum_i\sum_{r,s}Y_{rs}(i)Ln\left[\hat{P}_{rs}(i)\right].$$

When we assume that the above probabilities are known and have a bivariate normal structure, the estimator becomes bivariate probit. By estimating the probabilities using index

assumptions as discussed above, we avoid assuming parametric functional forms.[7] Employing bias controls and regular kernels, KSV show that this estimator is consistent and asymptotically distributed as normal (see Appendix).

Turning to the expenditure equation, we again need a correction term that will enable us to deal with the sample selection and endogeneity problems. As above, with V_o referring to (V_A, V_I), consider the control function:

$$G_d(V_o) \equiv E(u|A=1, I=d, X_s) = E(u|A=1, I=d, V_o)$$

where $d \in \{1, 0\}$. Notice that this adjustment is similar to that in the parametric case, but now we do not make any assumptions on its functional form here in the semiparametric formulation. With c as a constant, let $X_E\beta_E = X_c\beta_c + c$, and rewrite the expenditure equation as:

$$Y_E = X_c\beta_c + c + \theta_E d + G_d + u_d^* : A=1, I=d, u_d^* = u - G_d(V_o)$$

$$E\left[u_d^*|A=1, I=d, X_s\right] = E\left[u_d^*|A=1, I=d, V_o\right] = 0$$

Since the control functions are unknown, we extend Peter Robinson's differencing method (Robinson, 1988) to eliminate the unknown control functions:

$$Y_E - E(Y_E|A=1, I=d, V_o) = [X_c - E(X_c|A=1, I=d, V_o)]\beta_c + u_d^*$$

With "$*$" denoting a differenced variable, we can rewrite the above equation as:

$$Y^* = X^*\beta_c + u^*.$$

Before proceeding to estimate the above differenced model, there are several identification issues that need to be discussed. First, it is clear that the constant term and the insurance variable disappear from the model. Second, as in the parametric model, we require additional identifying restrictions. To this end, we impose the same exclusion restrictions as in the parametric model discussed above. To see that these restrictions are needed, suppose that there are no variables excluded from X_c that appear in the indices V_I and V_A. Without such restrictions, it will be possible to take linear combinations of the X_c variables and reproduce one of the indices.

Replacing true expectations and index parameter values with their estimates, we first use OLS to estimate the expenditure equation and get consistent estimates and residuals. Second, employing squared residuals, in a semiparametric regression, we estimate the variance for the error conditioned on the X-variables through the two indices. We then employ these conditional variances in a GLS approach to obtain the final results. Notice that the GLS estimator deals with the heteroscedasticity but not the first-stage estimation uncertainty. This uncertainty comes from the fact that estimated expectations are employed in place of true expectations and estimated index parameters are substituted for the true ones. It can be shown that the estimated expectations may be taken as known and do not affect standard errors for the estimated expenditure parameters; while the uncertainty from estimated index parameters must be taken into account. In particular, as in standard parametric sample selection models, the covariance matrix for these second-stage expenditure estimates will depend on the covariance matrix for the first-stage, joint-binary estimates. The reported standard errors here appropriately reflect this dependence (see Appendix).

Notice that in the above approach we can not directly estimate the impact of insurance coverage on expenditures (θ_E). Therefore, we next describe a strategy for indirectly obtaining

this marginal effect. Having described an estimator for the coefficient on X_c above, define residual expenditures:

$$R = Y_E - X_c\beta_c = c + \theta_E I + u$$

Heckman (1990) developed a method for estimating the constant term in a semiparametric sample selection model, which can be applied if we did not have the endogenous insurance variable. Andrews and Schafgans (1998), hereafter A&S, subsequently established the large-sample properties of a variant of this estimator. We extend this method to estimate both the constant term and the marginal effect of the endogenous insurance.

To set the intuition for the proposed estimator, notice that if there were no selection issues, we could proceed to develop an IV estimator for these parameters. To deal with selection, with $P_A \equiv \Pr(A = 1|V_o)$, for the error in the expenditure equation:

$$E(u|V_o) = 0 = P_A E(u|A = 1, V_o) + (1 - P_A)E(u|A = 0, V_o).$$

For such individuals with an access probability of 1 ($P_A = 1$), there would not be a selection problem in that from above:

$$E(u|A = 1, V_o) = E(u|V_o) = 0,$$

and we could proceed with IV estimation, employing

$$E(I|A = 1, V_o) = \Pr(I = 1|A = 1, V_o)$$

as an instrument for I. There are two implementation issues that now need to be solved.

First, as the above probability is unknown, we require a semiparametric estimate of this function as a feasible instrument. Second, we need an appropriate definition of a high access probability.

With $a > 0$, define a high probability set as one for which $P_A > 1 - N^{-a}$. In implementing this rule, we use estimated semiparametric probabilities described in Appendix.[8] In setting a, as in A&S, there is a bias-variance trade-off that guides its selection. Namely, if a is set very high, then the bias will be very low. However, the sample size available for IV estimation on the high probability set will then effectively be very small, resulting in a high variance. Similarly, if a is set too low, the variance can be made small, but the bias will not vanish sufficiently fast. To set a, let S be a smoothed indicator of the form in A&S that is zero unless observations are in the high probability set. Then, that paper shows that the bias will vanish appropriately fast if:

$$B = N^{1/2} |E\left[uAS\right] / \sqrt{E\left(AS^2\right)}| \to 0$$

The value of a must be set large enough so that this bias factor converges to 0, but small enough to keep the variance of the estimator low. Letting Z^* be the instrument with its mean removed, Klein, Shen and Vella (2011) show that the following similar bound holds:

$$B = N^{1/2} |E\left[uASZ^*\right] / \sqrt{E\left(AS^2Z^{*2}\right)}| \to 0$$

The choice of a is dictated by the same considerations as in A&S. To set a in this application and in the monte-carlo experiment described below, we employ an upper bound for B that can be estimated.[9] To balance bias and variance, we then select the smallest value of a such

that this bound tends to 0. (a is approximately .4 in this application.)

To get some sense as to how well the method described above works in practice, we conduct a small scale Monte Carlo experiment, where we find that this method performs very well. We generate data from the following design, which has the same structure as our model:

$$I = 1\{V_I > \varepsilon_I\} : V_I = X_1 + X_2 + X_3 + 1,$$

$$A = 1\{V_A + 2I > \varepsilon_A\} : V_A = X_1 + X_2,$$

$$Y_E = \ 4I + 2X_1 + 1 + u \ \ : A = 1$$

where the $X\prime s$ are all distributed as normal, and the errors are jointly normal with non-zero correlations between them. The sample size we use is n=2000, and the number of Monte Carlo replications is 1000. As we can see, the true $\theta_E = 4$. The average $\hat{\theta}_E$ from the Monte Carlo is 4.02, and the standard deviation is 0.14. In other words, the percentage bias is almost zero, and the variance is also small, taking into account that the truth is 4.

3 Data

The Medical Expenditure Panel Survey (MEPS) is an on-going nationally representative survey of U.S. civilian non-institutionalized population started in 1996 by U.S. Department of Health and Human Services. Surveys of households, employers, and medical providers are conducted to collect information on healthcare expenditures and health insurance coverage as well as demographic and socioeconomic characteristics.[10]

We consider the subsample of obese adults between the ages of 22 and 64, who are employed. People who have body mass index (BMI) greater than 30 are considered obese (CDC, 1985-2007). We focus on the obese population, because this is a growing population

that might have different healthcare needs and patterns than other groups. We also focus on individuals who are employed, because in the United States, insurance is often linked with employment. In fact, health insurance plans are often offered by employers. We exclude individuals who have public insurance, because having public insurance is not expected to be a consumer's choice for working adults between the ages of 22 and 64. The final sample consists of 2,771 individuals.[11]

The key endogenous variables that we seek to explain are insurance coverage, utilization of the healthcare system, and the level of expenditures. The insurance variable here is an indicator of whether the individual has private health insurance coverage. It would be important to take the generosity of insurance plans into account in terms of copays and deductibles. However, such information is not available in the MEPS data used here. The expenditures are the total amount paid for healthcare services, including both out-of-pocket payments and payments by insurance; but not including payments for over-the-counter drugs. Note that the expenditures are derived from the MEPS Household and Medical Provider Components. Since both the healthcare providers and the consumers are surveyed, it is more reliable than typical surveys. We define utilization of the healthcare system as having positive healthcare expenditures.[12]

The explanatory variables include demographics, socioeconomic status, and health related characteristics. The demographics include age, gender, race/ethnicity (white,non-white), marital status (married,other), family size, and region (northeast,midwest,south,west). Years of education, income, occupation class, and industry insurance rates are included as socioeconomic characteristics. We use an indicator for white-collar jobs (professional, management, business and financial operations) to reflect the impact of occupation, and the percentage of people having insurance in each industry in the Kaiser study as a variable

to reflect the impact of industry (Kaiser Family Foundation, 2006). The health related characteristics include number of comorbidities, presence of mental illnesses, and whether they are current smokers. Each individual was asked whether or not they had any of a number of conditions. The comorbidity variable then counts the following health problems: Alzheimer's disease, asthma, arthritis, cancer, emphysema, diabetes, heart disease, high blood pressure, osteoarthritis, and stroke. This variable is included to capture differences in people's physical health status and is often employed in health studies (e.g., Klabunde, 2000). Presence of mental illnesses is an indictor of whether an individual has depression, anxiety, or schizophrenia.

Recalling the exclusion restrictions discussed in the previous section, we use the following restrictions in this paper. The industry insurance rate and occupation are excluded from both utilization and expenditure equations; while marital status and region are excluded from the expenditure equation. As is known in the literature, occupation and industry have important effects on people's insurance (Kaiser Family Foundation, 2006). In the United States, insurance plans for working adults often come as a part of the compensation package. Different jobs might offer varied choices of insurance packages at different prices. Hence it affects the insurance decision by affecting the cost of buying insurance. However, once the insurance coverage decision is made, it is plausible to assume that the industry insurance rate and occupation class would not affect the benefit or the cost of utilization and expenditures after controlling for income and education. Married people might have more incentive to obtain insurance coverage. Different regions might have different healthcare policies and plans as well as different availabilities of healthcare services.[13] Consequently region may also have an impact on insurance. However, recall that the patient makes decisions about insurance and utilization, while the doctor and patient jointly decide on the level of treatment, with the

doctor being the main decision maker. Once a patient decides to visit a healthcare provider, we assume that the prescribed treatment does not depend on marriage or region. Hence the level of expenditures may not depend on these variables. We recognize the difficulty in finding appropriate restrictions for the type of model that we estimate, but view the exclusion restrictions discussed above as being plausible.

Some summary statistics of the data are provided in Table 1. Note that the continuous variables are categorized into groups to show the distribution of those variables. However, they remain continuous in estimating the model. Of the 2,771 individuals in our dataset, 488 (18%) are uninsured and 262 (10%) have no utilization. The level of expenditures for those that utilize healthcare is very skewed. About 40% of them have expenditures of less than $1,000, while 8% of them incur more than $10,000 in healthcare expenditures.

4 Results

Before we discuss the results, we recall that parameters are only identified up to location and scale in the semiparametric case. After estimation, we normalize the parameter of education to the corresponding parametric estimate for presentation purposes.[14] Below we examine both parametric and semiparametric results for the three decisions. We compare not only the normalized estimates and average marginal effects but also patterns of marginal effects calculated at different levels of certain continuous variables of interest. Most of the normalized estimates and average marginal effects are close between the two approaches for insurance and utilization decisions. However, the two estimation methods yield very different estimated effects of insurance on expenditures. Furthermore, the semiparametric approach gives richer patterns of marginal effects. Detailed results are provided in Tables 2-5.

As shown in Table 2, for the insurance decision, both the normalized estimates and the average marginal effects are similar for parametric and semiparametric approaches. The biggest marginal effect on the probability of having insurance comes from marital status, with the p-values of the coefficient on married in both approaches being less than 0.01. Marriage increases the probability of having private insurance coverage by more than 7% points. Region also has a significant effect on the insurance coverage, with the northeast indicator having coefficient p-values of 0.06 and 0.02 in parametric and semiparametric models respectively. People in the northeast region are 4-5% points more likely to have insurance compared to people living in the west. White people are 4% points more likely to have insurance than non-whites (coefficient p-value < 0.01). Education and income level both have significant positive impacts on insurance coverage. Industry insurance rate, which is one of the exclusions, has a substantial impact on the insurance decision. Increasing the industry insurance rate by 5% increases the probability of having insurance by more than 2% points on average. Occupation class is marginally significant. The number of comorbidities also has a significant positive effect on insurance. When the number of comorbidities increases by one, the average increase in the probability of having insurance is 2% points.

With respect to the normalized parameter estimates and averaged marginal effects, the parametric and semiparametric results for the utilization decision are also similar. These results are presented in Table 3. One of the most important questions here is how insurance coverage affects utilization, and parametric and semiparametric estimations provide very similar results. The average marginal effect of insurance coverage is 14-15% points (the probabilities move from 78-80% to 93-94%,) meaning if we move everyone in the sample from uninsured to insured, the average gain in the probability of visiting a doctor is 14-15% points, a large number. Both the number of comorbidities and the presence of mental

illnesses have very significant positive impacts on utilization (coefficient p-value < 0.05). For the number of comorbidities, parametric estimation gives a higher marginal effect of 6% points compared to the 3% points of the semiparametric approach; while for the presence of mental illnesses, both approaches give an average marginal effect of 4% points. One interesting finding here is that females are much more likely to visit a doctor. Parametric and semiparametric estimations yield average marginal effects of 6% points and 4% points respectively. Another interesting finding is that income does not have a significant impact on utilization once the insurance decision is fixed. Marital status, which is one of the exclusions, has a highly significant impact on utilization. Married people are 2-3% points more likely to utilize healthcare. Region, as an additional exclusion, is marginally significant. Another important finding here is that the correlation factor in the parametric estimation is very small in absolute magnitude (-0.09), and it is statistically insignificant with a p-value of 0.61. As discussed earlier, this finding has implications for the form of an adjustment factor in estimating the expenditure equation.

The final equation deals with the level of healthcare expenditures. Note that most of the marginal effects are the same as the coefficient estimates here. With the exception of the impact of insurance, estimates in the two approaches are similar. The number of comorbidities and the presence of mental illnesses both have very significant effects in this equation. Both have p-values of less than 0.01. Having one more physical disease can increase the level of expenditures by about 35% on average; while having a mental illness increases it by more than 45%.

The semiparametric approach estimates the marginal impact of insurance on expenditures to be 48%.[15] This impact would seem to be credible as it is close to the number in a previous study by Newhouse and the Insurance Experiment Group (Newhouse et al., 1993).

Their study based on the RAND Health Insurance Experiment shows that mean predicted expenditure in the 0% coinsurance (free-care) plan is 46 percent higher than in the 95% coinsurance plan. We want to keep in mind that the relevance of the study may be lowered by the fact that it was done more than a decade ago and not all the insured people in the sample get free care.[16] In contrast, parametric estimation gives a marginal effect of 125%. We note that there are many other parametric studies that have also found an insurance impact of this magnitude (e.g., Hadley and Holahan, 2003; Miller et al., 2004). These studies treat insurance as exogenous and state that in so doing the marginal impact of insurance has an upward bias. However, none of these studies have quantified the extent of this bias.

To understand the large difference between semiparametric and parametric results, we performed several different checks. First, we examined the normality assumption in the insurance equation by using semiparametric methods to estimate the density of the error. In particular, we obtained the semiparametric estimate of the expectation of the insurance dummy conditioned on the index. In a traditional threshold-crossing model, this estimated expectation is the estimate of the distribution function for the error term. Taking a numerical derivative then produces its density. As shown in Figure 1, the density estimator, which we re-centered to have median zero, is remarkably non-normal for the insurance error. It should be noted that other components of the model (access and expenditures) depend on the insurance decision. Therefore, misspecification errors in the insurance equation will be transmitted to these other components of the model.

To evaluate the implications of parametric distributional assumptions not holding, we performed the following experiment. Recall that in the parametric model, the G-functions that control for selection and endogeneity are known under normality. Given the failure of parametric distributional assumptions to hold, it would seem that these parametric G-functions

are incorrect. Accordingly, we semiparametrically estimated these functions without making any assumptions on their functional forms. Recall that the semiparametric estimates of the expenditure equation were obtained by differencing out the G-functions as their form was not known. However, once all of the parameters of the expenditure model have been estimated, it is possible to obtain the semiparametric estimates of these functions. With the subscript s indicating a semiparametric estimator:

$$\hat{G}_{ds} = \hat{E}\left[Y_E - X_c\hat{\beta}_s - \hat{c}_s - \hat{\theta}_s I \mid A = 1, I = d, \hat{V}_I, \hat{V}_A\right].$$

Replacing the parametric G-functions with the flexibly estimated semiparametric functions above, we then re-estimated the parametric expenditure equation. The marginal impact of insurance was found to be .50, which confirms the finding that the parametric marginal effect has an upward bias by a factor of more than two.

Table 5 shows that parametric and semiparametric approaches also give very different marginal effects for different population groups. Here, the parametric approach restricts the marginal effects of the groups to be monotonic, while the semiparametric approach does not have this restriction and hence can provide more accurate results. In the parametric estimation, the marginal effects of education on insurance for the three groups (less than high school, high school, and some college or more) are 3.22% points, 2.00% points, and 1.05% points, which shows a strong monotonic relation; in the semiparametric estimation, the largest marginal effect is also in the "less than high school" population, and it is 1.37% points. However, the marginal effects in the other two groups are at a similar level of 1.0-1.1% points. The same pattern happens for the marginal effects of education on utilization. The parametric estimation yields marginal effects of 1.21% points, 0.58% points, and 0.38%

points respectively for the three groups, while semiparametric estimation suggests again that the marginal effects are close in the three groups (3.25% points, 2.81% points, and 2.96% points). This result suggests that it is important to improve health literacy in all groups, with probably more of the effort placed on people having less than high school education. Another interesting observation concerns the industry insurance rate. In the semiparametric case, the biggest marginal effect (1.71% points compared to 1.55% points and 1.37% points) is in the middle group where the industry insurance rate is 75-90%. The people in those industries have the greatest marginal benefit of getting into a more insured industry. In contrast, in the parametric case, marginal effects are again monotonic.

The above results are based on the sample including people who have only outpatient use and those with inpatient use. It is noted in the RAND experimental study that the distribution of medical expenditures differs for these two groups. To address the issue of whether or not the model differs for these groups, we re-estimated the model for individuals with only outpatient use and found that the results are similar. Detailed results are available upon request.

5 Conclusions

This paper studies the determinants of three healthcare decisions: insurance, utilization, and expenditures. We study the above interrelated healthcare decisions by analyzing a system of three simultaneous equations. Both parametric and semiparametric methods are employed to estimate the model. The merit of our semiparametric approach compared to a parametric approach is that it avoids distributional and functional form assumptions, which are not well justified. Indeed, while there are many similarities, parametric and semiparametric

approaches yield some very different results, which can lead to different policy implications.

In summary, we find that insurance has a substantial effect on both utilization and expenditures. Both methods suggest that having private insurance coverage increases the likelihood of seeking healthcare by about 15% points. However, the estimated magnitude of the effect on expenditure diverge. The parametric estimation predicts the level of expenditures to increase by 125% if universal insurance is given; while semiparametric estimation predicts an increase of 48%, a number close to that found in a RAND experimental study (Newhouse et al., 1993). Because the parametric assumptions are incorrect, the parametrically estimated impact of insurance on expenditures has an upward bias on the order of 100%. The policy relevance of this finding is that the cost of extending universal healthcare is much lower than predicted by traditional parametric methods.

Other marginal effects are also worth noting. Education is an important factor in every healthcare decision, and hence improving health literacy is an important issue in the obese population. Given the pattern in the marginal effects, parametric results suggest that most, if not all, of the emphasis be placed on improving health literacy of the low education group (below high school). In contrast, results from the semiparametric case suggest that it is important to improve health literacy among all education groups (with the low group somewhat favored). Finally both physical and mental illnesses increase expenditures dramatically. Physical illnesses increase the level of expenditures by about 35%, and mental illnesses increase it even more (45%+). This suggests that the obese population with physical and mental illnesses is a very challenging population. More prevention and treatment of physical and mental illnesses should be provided to this population.

There are some limitations and consequently some future research directions that we want to point out. First, this study is based on the obese (BMI>30) population. It would

be interesting to investigate the magnitude of marginal effects for different BMI categories. Second, we do not have information to distinguish the type of healthcare encounters, for example, whether it is a preventive checkup with a physician or an acute episode of some disease. It would be useful to distinguish different types of healthcare use, so that we can study the effects on different types of healthcare. Third, since this is a cross-sectional dataset, we do not know the temporal effects. It would be interesting to know, for example, how the use of preventive care in the previous periods affect inpatient care use in subsequent time periods.

Notes

[1]I want to thank Roger Klein, Carolyn Moehling, John Landon-Lane, Francis Vella, the editor and the referees for all their helpful comments and suggestions. I would also like to thank Louise Russell and Usha Sambamoorthi for helpful discussions. I have also benefitted from comments at various seminars. All mistakes are mine.

[2]Department of Economics, Georgetown University, 37th and O Streets, Washington, DC 20057. Email: cs589@georgetown.edu.

[3]As discussed in the next section, in a semiparametric formulation we will not need to make any assumptions on the functional form of this correction factor.

[4]If X is continuous, then the convergence rate of the estimated expectation to the truth becomes slower as the dimension of X increases. If X is discrete, there may be few observations to estimate $E(Y|X)$ at each value of X.

[5]Klein, Shen, and Vella (2010b) develop an adaptive estimator for this impact.

[6]In the insurance model, we treat the following variables: age, age^2, number of comorbidities, years of education, family size, and industry insurance rate as being approximately continuous; while in the access decision, these variables are: age, age^2, number of comorbidities, years of education, and family size.

[7]For technical reasons, and as is standard in this literature, we trim out certain observations for which the probabilities are poorly estimated (see Appendix A).

[8]Tail assumptions similar to A&S enable us to keep index density denominators from being too small while remaining in a high probability set. Appendix A develops an appropriate trimming strategy.

[9]Assume that the error, u, has finite r absolute moments. Then, Klein, Shen and

Vella(2010b) show that an upper bound on the bias is given by:

$$\bar{B} = N^{1/2}|N^{-a(r-1)/r} E\left[ASZ^*\right] / \sqrt{E\left(AS^2Z^{*2}\right)}|$$

In our application, we replace the expectations with sample averages.

[10]We note that the semiparametric model can be less sensitive to reporting errors than parametric models (see, for example, Hausman et al., 1998).

[11]Other exclusion criteria include: individuals who died during the year, missing values on the exogenous variables used. Various robustness checks indicate that there are no selection issues in this sample.

[12]We use this indicator instead of the self-reported healthcare utilization, because the self-reported utilization may suffer from recall errors, whereas the expenditure data were collected by both sides and hence are more reliable.

[13]No detailed information about state of residence is available in the MEPS dataset.

[14]The choice of variable on which to normalize does not affect estimation results (provided that the variable belongs in the model).

[15]The 90 percent confidence interval for the marginal effect is approximately [.12, .84], which is based on the asymptotic distribution of the estimator as given in Klein, Shen, and Vella (2011).

[16]The impact would be somewhat lower than 46% in comparison to other coinsurance levels plans (not totally free), further supporting the finding that there is severe upward bias in the parametric estimate.

References

Andrews, Donald W. K., and Schafgans, Marcia. M. A, "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies* 65:3 (1998), 497-517.

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review* 53 (1963), 941-973.

Blundell, Richard. W., and Powell, James. L., "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies* 71:3 (2004), 655-679.

Cardon, James. H., and Hendel Igal, "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey," *RAND Journal of Economics* 32:3 (2001), 408-27.

Centers for Disease Control and Prevention (CDC), Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 1985-2007.

Chiappori, Pierre-André, and Salanie Bernard, "Testing for Asymmetric Information in Insurance Markets," *Journal of Political Economy* 108:1 (2001), 56-78.

Duan, Naihua, Manning, Willard G., Morris, Carl N., and Newhouse Joseph P., "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business & Economic Statistics* 1:2 (1983), 115-126.

Duan, Naihua, Manning, Willard G., Morris, Carl N., and Newhouse Joseph P., "Choosing Between the Sample-Selection Model and the Multi-Part Model," *Journal of Business &Economic Statistics* 2(1984), 283-289.

Duan, Naihua, Manning, Willard G., Morris, Carl N., and Newhouse Joseph P., "Comments on Selectivity Bias," *Advances in Health Economics and Health Services Research* 6(1985), 19-24.

Hadley, Jack, and Holahan, John, "Covering the Uninsured: How Much Would It Cost?" *Health Affairs* (2003), W3-250-65.

Hausman, Jerry A., Abrevaya, Jason, and Scott-Morton, Fiona M., "Misclassification of the Dependent Variable in a Discrete Response Setting," *Journal of Econometrics* 87(1998), 239–269.

Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of*

Economic Social Measurement 5:4(1976), 475-492.

Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica* 47:1(1979), 53-161.

Heckman, James J., "Varieties of Selection Bias," *American Economic Review* 80(1990), 313–318.

Holly, Alberto, Lucien, Gardiol and Jacques, Huguenin, "Hospital Services Utilization in Switzerland: The Role of Supplementary Insurance," Institute of Health Economics and Management, University of Lausanne, manuscript (2002).

The Kaiser Family Foundation, Kaiser Fast Facts, Health Insurance Coverage in America (2006).

Klabunde Carrie N. Potosky Arnold L., Legler Julie M., and Warren Joan L., "Development of a Comorbidity Index Using Physician Claims," *Journal of Clinical Epidemiology* 53(2000), 1258-1267.

Klein, Roger W. and Shen, Chan, "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory* 26:6(2010), 1683-1718.

Klein, Roger W. and Spady, Richard H., "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61(1993), 387-421.

Klein, Roger W., Shen, Chan, and Vella, Francis G., "Triangular Semiparametric Models Featuring Two Dependent Endogenous Binary Outcomes," unpublished manuscript(2010).

Klein, Roger W., Shen, Chan, and Vella, Francis G., "Semiparametric Selection Models with Binary Outcomes." unpublished manuscript(2011).

Lee, Lung-Fei, "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies* 49:3(1982), 355-372.

Lindau Stacy T., Basu, Anirban, and Leitsch, Sara A., "Health Literacy as a Predictor of Follow-up After an Abnormal Pap smear: a Prospective Study," *Journal of General Internal Medicine* 21:8(2006), 829-834.

Maddala, Gangadharrao S., "A Survey of the Literature on Selectivity Bias as it Pertains to Health Care Markets," *Advances in Health Economics and Health Services Research* 6(1985), 3-18.

Manning, Willard G., Newhouse, Joseph P., Duan, Naihua, Keeler,Emmett B., and Leibowitz, Arleen, "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review* 77(1987), 251-277.

Miller, Edward, Banthin, Jessica S., and Moeller, John F., "Covering the Uninsured: Estimates of the Impact on Total Health Expenditures for 2002," Agency for Healthcare Research and Quality Working Paper No. 04007(2004).

Mullahy, John, "Much Ado about Two: Reconsidering Retransformation and Two-part Model in Health Econometrics," *Journal of Health Economics* 17(1998), 247–281.

Newhouse Joseph P., and the Insurance Experiment Group, *Free for all? Lessons from the RAND Health Insurance Experiment* (Cambridge: Harvard University Press 1993).

Powell, James L., Stock, James H., and Stoker, Thomas M., "Semiparametric Estimation of Weighted Average Derivatives," *Econometrica* 57(1989), 1403-1430.

Puhani, Patrick A., "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys* 14:1(2000),53–68.

RAND Health Insurance Experiment [in Metropolitan and Non-Metropolitan Areas of the United States], 1974-1982.

Robinson, Peter M., "Root-n-consistent Semiparametric Regression," *Econometrica* 56(1988), 931-954.

Rothschild, Michael, and Stiglitz, Joseph, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *The Quarterly Journal of Economics* 90:4(1976), 630-49.

Serfling, Robert J., *Approximation Theorems of Mathematical Statistics* (New York: John Wiley & Sons, 1980), .

Vera-Hernandez, Angel M., "Duplicate Coverage and Demand for Healthcare. The case of Catalonia," *Health Economics* 8(1999), 579-598.

Vytlacil, Edward, and Yildiz, Nese, "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica* 75(2007), 757-779.

Wooldridge, Jeffery M, *Econometric Analysis of Cross Section and Panel Data* (Cambridge: MIT Press, 2002).

Table 1.----Description of Study Population

		N	%
All		2771	100.0
Insurance Coverage			
	Insured	2283	82.4
	Uninsured	488	17.6
Utilization			
	yes	2509	90.5
	no	262	9.5
Expenditures			
	no expenditures	262	9.5
	<1,000	990	35.7

	1,000-2,000	443	16.0
	2,000-5,000	607	21.9
	5,000-10,000	265	9.6
	10,000+	204	7.4
Education			
	Less than high school	471	17.0
	High school	946	34.1
	College or higher	1354	48.9
Age			
	<40	1022	36.9
	40-49	856	30.9
	50+	893	32.2
Income			
	<20,000	781	28.2
	20,000-30,000	569	20.5

	30,000-50,000	794	28.7
	50,000+	627	22.6
Gender			
	Female	1460	52.7
	Male	1311	47.3
Race			
	White	1551	56.0
	Non-white	1220	44.0
Number of Comorbidities			
	Zero	1352	48.8
	One	881	31.8
	Two plus	538	19.4
Mental Illnesses			
	yes	540	19.5
	no	2231	80.5

Current Smoker			
	yes	542	19.6
	no	2229	80.4
Marital Status			
	Married	1714	61.9
	Other	1057	38.1
Family Size			
	One-Two	1219	44.0
	Three-Four	1069	38.6
	Five plus	483	17.4
Region			
	Northeast	381	13.7
	Midwest	611	22.0
	South	1206	43.5
	West	573	20.7

Industry Insurance Rate			
	<75% insured	519	18.7
	75-90% insured	1326	47.9
	90%+ insured	926	33.4
Occupation			
	White-collar	830	30.0
	Other	1941	70.0

Table 2.----Parametric and Semiparametric Estimation Results -- Insurance Coverage

	Parametric Estimation				*Semiparametric Estimation*			
	Estimate	**(SE)**	**p-value**	**ME (% pts.)**	**Estimate**	**(SE)**	**p-value**	**ME (% pts.)**
Intercept	-6.47	(0.57)	<.01					
Age	0.04	(0.02)	0.09	0.10	-0.01	(0.02)	0.56	0.00
Age²	-3.86E-04	(2.62E-04)	0.14		1.65E-04	(2.91E-04)	0.57	
# of Comorbidities	0.10	(0.04)	0.01	1.90	0.12	(0.05)	0.01	1.84
Mental Illnesses	0.14	(0.09)	0.12	2.68	0.07	(0.09)	0.47	1.00
Female	-0.01	(0.07)	0.83	-0.30	0.09	(0.08)	0.25	1.42
White	0.28	(0.07)	<.01	5.74	0.27	(0.09)	<.01	4.32
Income	0.29	(0.03)	<.01	0.58	0.74	(0.13)	<.01	3.31
Current Smoker	-0.07	(0.08)	0.36	-1.45	-0.07	(0.09)	0.45	-1.03
Years of Education	0.09	(0.01)	<.01	1.85	0.09			1.40
Married	0.36	(0.07)	<.01	7.54	0.45	(0.10)	<.01	7.22
Family Size	0.01	(0.02)	0.72	0.16	0.00	(0.02)	0.85	0.06

Region-Northeast	0.23	(	0.12	)	0.06	4.41	0.00	(	0.14	)	0.02	4.83
Region-Midwest	0.12	(	0.10	)	0.23	2.44	0.04	(	0.10	)	0.68	0.64
Region-South	-0.10	(	0.08	)	0.23	-2.03	-0.03	(	0.09	)	0.75	-0.44
Industry Insurance												
Rate	2.61	(	0.30	)	<.01	2.54	2.81	(	0.56	)	<.01	2.06
White-collar	-0.04	(	0.08	)	0.61	-0.88	-0.12	(	0.08	)	0.13	-1.86

Estimate=parameter estimate; SE=standard error; ME (% pts.)=average marginal effect in percentage points.

Expenditure and income are in $1,000 and are logged.

Reference group for region = West.

Marginal effects of continuous variables are calculated by moving everyone in the sample above by 1 unit, except income and industry insurance rate which were moved by 10% and 5% respectively.

Marginal effects of discrete variables are calculated by moving everyone in the sample from zero to one.

Table 3.----Parametric and Semiparametric Estimation Results -- Utilization

	Parametric Estimation				*Semiparametric Estimation*			
	Estimate	**SE**	**p-value**	**ME (% pts.)**	**Estimate**	**SE**	**p-value**	**ME (% pts.)**
Intercept	0.03	(0.67)	0.97					
Age	-0.04	(0.03)	0.20	0.10	0.01	(0.04)	0.86	0.12
Age²	5.64E-04	(3.56E-04)	0.11		1.25E-04	(4.81E-04)	0.79	
# of Comorbidities	0.56	(0.07)	<.01	5.55	0.55	(0.26)	0.03	3.39
Mental Illnesses	0.31	(0.12)	0.01	3.69	0.64	(0.33)	0.05	4.18
Female	0.43	(0.08)	<.01	5.77	0.51	(0.27)	0.05	3.52
White	0.05	(0.09)	0.57	0.66	-0.35	(0.22)	0.10	-2.27
Income	-0.01	(0.04)	0.86	-0.01	-0.28	(0.21)	0.19	-1.40
Current Smoker	-0.10	(0.09)	0.28	-1.36	0.13	(0.12)	0.27	0.88
Years of Education	0.05	(0.02)	<.01	0.70	0.05			0.36
Married	0.25	(0.09)	0.01	3.38	0.36	(0.18)	0.05	2.40
Family Size	-0.04	(0.03)	0.12	-0.55	-0.10	(0.07)	0.15	-0.66

Region-Northeast	0.02	(	0.14	)	0.88	0.28	0.00	(0.19)	0.41	-1.05
Region-Midwest	0.04	(	0.12	)	0.76	0.47	0.19	(0.18)	0.27	1.31
Region-South	0.08	(	0.10	)	0.39	1.11	0.21	(0.18)	0.22	1.44
Insurance Coverage	0.88	(	0.29	)	<.01	15.50				13.70
Correlation Factor	-0.09	(	0.17	)	0.61					

Estimate=parameter estimate; SE=standard error; ME (% pts.)=average marginal effect in percentage points.

Expenditure and income are in $1,000 and are logged.

Reference group for region = West.

Marginal effects of continuous variables are calculated by moving everyone in the sample above by 1 unit, except income and industry insurance rate which were moved by 10% and 5% respectively.

Marginal effects of discrete variables are calculated by moving everyone in the sample from zero to one.

Table 4.----Parametric and Semiparametric Estimation Results -- Level of Expenditures

	Parametric Estimation				*Semiparametric Estimation*			
	Estimate	**(SE)**	**p-value**	**ME (%)**	**Estimate**	**(SE)**	**p-value**	**ME (%)**
Intercept	5.00	(0.52)	<.01					
Age	-1.23E-03	(0.02)	0.95	1.40	1.60E-04	(0.02)	0.99	1.00
Age^2	1.72E-04	(2.23E-04)	0.45		1.15E-04	(2.36E-04)	0.63	
# of Comorbidities	0.40	(0.04)	<.01	40.40	0.35	(0.08)	<.01	34.96
Mental Illnesses	0.55	(0.07)	<.01	54.99	0.45	(0.10)	<.01	45.32
Female	0.15	(0.06)	0.01	15.45	0.08	(0.08)	0.33	8.06
White	0.25	(0.06)	<.01	25.17	0.36	(0.07)	<.01	36.37
Income	-0.01	(0.04)	0.78	-0.10	0.09	(0.05)	0.06	0.93
Current Smoker	-0.18	(0.07)	0.01	-17.73	-0.20	(0.07)	0.01	-19.75
Years of Education	0.03	(0.01)	0.01	3.28	0.03	(0.01)	0.02	3.23
Family Size	-0.04	(0.02)	0.05	-3.56	-0.03	(0.02)	0.13	-3.07

Insurance Coverage	1.25	(	0.32	)	<.01	124.85	47.91
Correction Term wrt Visit	-0.05	(	0.33	)	0.89		
Correction Term wrt Insurance	-0.32	(	0.16	)	0.05		

Estimate=parameter estimate; SE=standard error; ME (%) =average marginal effect in percentages.

Expenditure and income are in $1,000 and are logged.

Reference group for region = West.

Marginal effects of continuous variables are calculated by moving everyone in the sample above by 1 unit, except income and industry insurance rate which were moved by 10% and 5% respectively.

Marginal effects of discrete variables are calculated by moving everyone in the sample from zero to one.

Table 5.----Marginal Effects across the Distribution of Select Variables of Interest

	ME on Insurance (% pts.)		ME on Utilization (% pts.)	
	parametric	semiparametric	parametric	semiparametric
Education				
Less than high school	3.22	1.37	1.21	3.25
High school	2.00	1.00	0.58	2.81
College or higher	1.05	1.11	0.38	2.96
Industry Insurance Rate				
<75% insured	4.28	1.55	--	--
75-90% insured	2.32	1.71	--	--
90%+ insured	1.41	1.37	--	--

ME on Insurance (% pts.)=median marginal effect on insurance in percentage points.

ME on Utilization (% pts.)=median marginal effect on utilization in percentage points.

Marginal effects of education are calculated by moving everyone in the sample above by one year.

Marginal effects of industry insurance rate are calculated by moving everyone in the sample above by 5%.

Figure 1.----Estimated error distribution in the insurance equation